Multivariate Functional Regression and Selection

by

Joseph Naiman

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2020

Doctoral Committee:

Professor Peter Song, Chair
Assistant Professor Walter Dempsey
Assistant Professor Peisong Han
Professor Kerby Shedden

Joseph Naiman

jnaiman@umich.edu

ORCID id: 0000-0002-9027-7569

This thesis is dedicated to my wife, Jessica, who has stood by me throughout the PhD program with love and support, and to my children, Dovi, Shira, Sari, Talya and Meir.

ACKNOWLEDGEMENTS

First and foremost I would like to thank God for providing me with the insight, strength and guidance to complete this dissertation.

I am grateful to the Department of Biostatistics for providing me with the essential tools to complete this dissertation. The outstanding faculty and diversity in research helped guide me throughout my PhD studies.

My deepest gratitude goes out to my advisor, Professor Peter Song, who patiently guided me through my thesis work. Professor Song always made himself available despite his busy schedule and the multiple PhD students he is currently mentoring. Many thanks to the members of Song Lab who provided me with the data used in this dissertation. I am grateful for the insightful weekly discussions presented at Song Lab, and feel fortunate to have been part of the innovation and growth that happens there on a regular basis.

Thank you to my other committee members- Professor Walter Dempsey, Professor Peisong Han, and Professor Kerby Shedden- for their many contributions. Their feedback of my proposal helped shape the direction of my thesis, and for that I am grateful.

Thank you Professor Shedden and Professor Brenda Gillespie for allowing me to work at the Consulting for Statistics and Analytics Research (CSCAR) throughout the duration of my PhD program. At CSCAR I learned invaluable skills about applying statistics to real life studies being conducted both within and outside the University. At weekly CSCAR meetings we tackled difficult statistical problems and

became better statisticians through the collaborative efforts of the staff and graduate students.

My endless gratitude to the many family members who helped make this dissertation possible. I thank my parents, Mr. Amiel and Dr. Channah Naiman for their support. My mother has provided me with the motivation to pursue a PhD by serving as a role model of teaching excellence, and her invaluable advice and guidance has helped me throughout the program. My in-laws, Dr. Ephraim and Mrs. Rose Zinberg, I thank for their ongoing support and availability to help with the kids. My father-in-law Mr. Jared Cohen, I thank for providing me with a job that allowed our family to manage financially as I pursued this PhD over many years and for being a role model of professional success.

As a father of four when I started this program, I was able to complete it only with the support of my wife and children. To my wife, Jessica, I thank for lovingly and patiently standing by me throughout the program. I am extremely grateful for the help of my wife and my mother in editing this dissertation.

TABLE OF CONTENTS

DEDICATION ii

ACKNOWLEDGEMENTS iii

LIST OF FIGURES viii

LIST OF TABLES x

LIST OF APPENDICES xii

ABSTRACT xiii

CHAPTER

I. Introduction 1

1.0.1 Motivation 1
1.0.2 Accelerometer Data 1
1.0.3 Functional Regression 4

II. Multivariate Functional Regression and Selection (MFRS) Framework 8

2.1 Introduction 8
2.1.1 Least Squares Kernel Machine (LSKM) 8
2.1.2 Feature Selection 10
2.2 Proposed Model 11
2.3 Algorithm 16
2.4 Theoretical Analysis 20
2.5 Identifiability 23
2.6 Simulations 26
2.7 Discussion 35

III. Accelerometer Modeling Application 38

3.1 Introduction 38
3.2 ELEMENT Dataset 39
3.3 Accelerometer Preprocessing 42
3.4 Review of Statistical Methods 45
3.4.1 FPCA 45
3.4.2 Least Squares Kernel Machine 45
3.4.3 MFRS for additive LSKM 50
3.5 Proposed Statistical Models 51
3.5.1 Results from ELEMENT dataset 54
3.5.2 7-day vs 1-day averaged 57
3.5.3 Tri-axis AC vs VM vs AI 57
3.6 Simulation 58
3.7 Discussion 63

IV. Accelerometer Modeling with Multilevel Functional Principal Component Analysis 65

4.1 Introduction 65
4.2 Functional Anova Model 67
4.3 MFRS Framework For Decomposed Functional 70
4.4 Results using the X(t) process 71
4.5 Joint Modeling with both the X(t) and U(t) processes 74
4.5.1 Setup 74
4.5.2 Results from the joint modeling 75
4.6 Discussion 77

V. Functional Logistic Regression 79

5.1 Introduction 79
5.2 Background for selection of Import points for KLR 80
5.2.1 KLR 81
5.2.2 Tikhonov regularization 82
5.2.3 Logistic regression with lasso (L1) penalty 83
5.2.4 Elastic Net 84
5.3 KLR with Import Selection 85
5.4 MFRS Logistic Regression 88

VI. Summary and Future Work 92

APPENDICES 94
A.1 Technical assumptions and proofs 95
A.1.1 Proof of Theorem **??** 95
A.1.2 Proof of Corollary **??** 99
A.2 Gauss-Newton Algorithm 103
A.3 Additional Simulation Results in Scenario 2 104
B.1 Additional Graphs from ELEMENT dataset from Chapter 3 . 106
C.1 Additional Graphs from ELEMENT dataset from Chapter 4 . 109

BIBLIOGRAPHY 112

LIST OF FIGURES

Figure

2.1 Estimated marginal functions with 95 percent shaded confidence bands of the function h evaluated at 100 grid points for each component while holding all other components equal to 0.5 in Scenario 2 35

3.1 TriAxis Activity Count for the 7 Days of accelerometer wear 40

3.2 VM Activity Count for the 7 Days of accelerometer wear 41

3.3 TriAxis Activity Count averaged minute-by-minute for the 7 Days of accelerometer wear 41

3.4 VM Activity Count averaged minute-by-minute for the 7 Days of accelerometer wear 42

3.5 AI for the 7 Days of accelerometer wear 43

3.6 AI averaged minute-by-minute for the 7 Days of accelerometer wear 44

3.7 60

B.1 Leading Eigenfunction extracted for Tri-axis 7-day functional data . 106

B.2 Leading Eigenfunction extracted for VM 7-day functional data . . . 107

B.3 Leading Eigenfunction extracted for Tri-axis 1-day averaged functional data 107

B.4 Leading Eigenfunction extracted for VM 1-day averaged functional data 108

C.1 Leading Eigenfunction extracted from X(t) process for Tri-axis data 109

C.2 Leading Eigenfunction extracted from U(t) process for Tri-axis data 110

C.3 Leading Eigenfunction extracted from X(t) and U(t) process for VM 110

C.4 Leading Eigenfunction extracted from X(t) and U(t) process for AI 111

LIST OF TABLES

Table

2.1 Goodness of Fit for Scenario 1 29

2.2 Model Size for Scenario 1 30

2.3 FPC Selection for Scenario 1 30

2.4 Goodness of Fit via the concordance regression for Scenario 2 33

2.5 Sensitivity and Specificity of Functional Selection for Scenario 2 . . 33

2.6 FPC Selection for Scenario 2 Functional Z^1 34

2.7 FPC Selection for Scenario 2 Functional Z^2 34

3.1 #FPC Scores that explain $\geq 50\%$ 55

3.2 FPCA Functional selection of 3-D Activity Count for X,Y and Z axis for the 7 day functional 55

3.3 FPCA Functional selection of 3-D Activity Count for X,Y and Z axis for the 1 day averaged functional 55

3.4 R^2_{AQ} using the 7-day functional of 3-D Activity Count 56

3.5 R^2_{AQ} using the 7-day functional of VM Activity Count 56

3.6 R^2_{AQ} using the 7-day functional of AI 56

3.7 R^2_{AQ} using 1-day averaged functional of 3-D Activity Count 56

3.8 R^2_{AQ} using 1-day averaged functional of VM Activity Count 57

3.9 R^2_{AQ} using 1-day averaged functional of AI 57

3.10 R^2_{AQ} for Simulated acceleromter data Scenarios 1 and 2 62

3.11 Sensitivity and Specificity of Functional Selection 62

3.12 Feature Selection for Scenario 1 63

3.13 FPC Selection for Scenario 2 63

4.1 SFPCA R^2_{AQ} using the 7-day functional of 3-D Activity Count . . . 72

4.2 SFPCA R^2_{AQ} using the 7-day functional of VM Activity Count . . . 73

4.3 SFPCA R^2_{AQ} using the 7-day functional of AI 73

4.4 #FPC Scores that explain $\geq 50\%$ 73

4.5 % of variance explained from the decomposition $Z(t) = X(t) + U(t)$ 73

4.6 SFPCA Functional selection of 3-D Activity Count for X,Y and Z axis 74

4.7 SFPCA R^2_{AQ} using the 7-day functional of 3-D Activity Count using $X(t)$ and $U(t)$ process. 75

4.8 SFPCA R^2_{AQ} using the 7-day functional of VM Activity Count. . . . 75

4.9 SFPCA R^2_{AQ} using the 7-day functional of AI using $U(t)$ and $X(t)$, 76

4.10 SFPCA Functional selection of 3-D Activity Count for X,Y and Z axis for $U(t)$ process. 76

A.1 Model Size for Scenario 2 105

LIST OF APPENDICES

Appendix

A. Proofs and additional Tables from Chapters 2 95

B. Additional Graphs for Chapters 3 106

C. Additional Graphs for Chapters 4 109

ABSTRACT

With the pervasiveness of sensor data, real-time physiological signals and behavioral data are often collected in many biomedical studies. This thesis is motivated by data collected from a tri-axis accelerometer ActiGraph GT3X, a device that measures acceleration in the 3-D directions with a sampling frequency of 30-100 Hz. The central task is to relate this multivariate functional quantity with various scalar health outcomes of interest in the presence of other scalar covariates.

In the first project, we propose a new methodological framework of semi-parametric regression models that allow the study of a non-linear relationship between a scalar response and multiple functional predictors in the presence of scalar covariates. The proposed methodology is termed as MFRS (Multivariate Functional Regression and Selection). Utilizing functional principal components analysis (FPCA) and least-squares kernel machine methods (LSKM), we substantially extend the classical semi-parametric regression model of scalar responses on scalar predictors, in which multiple functional predictors are included in the non-linear model. Regularization is established for feature selection in the setting of reproducing kernel Hilbert spaces. The proposed method enables us to perform simultaneous model fitting and variable selection on functional features. For implementation, we propose an effective algorithm to solve related optimization problems, in that iterations take place between both linear mixed models and a variable selection procedure (e.g. sparse group lasso). We show algorithmic convergence results and theoretical guarantees for the proposed methodology. We illustrate its performance through extensive simulation experiments.

In the second project we apply our MFRS framework developed in project I to perform a comprehensive mobile health application. This is a study conducted in

Mexico City where participants wore an ActiGraph (a tri-axis accelerometer) for seven days with no interruption. We investigated various ways of preprocessing the raw accelerometer data and focused on an important comparative analysis. This comparison concerns methods that treat either the full accelerometer data of seven days as one functional or average the seven days of data into a one day functional. We extend the LSKM framework developed in project I to handle an additive model for multiple functional covariates and compare the extension with our MFRS method given in project I.

In the third project we adopt structural principal component analysis (SFPCA) for an alternative analysis of the accelerometer data to that done in project II. SFPCA allows us to treat the functional data of seven days into seven repeats of one day functional. Utilizing the MFRS framework, we demonstrate the benefits of allowing a non-linear and non-additive relationship between health outcomes and repeated functional predictors. Taken together, the second and third projects collectively provide some useful approaches to preprocessing functional data from a mobile device and performing non-linear and non-additive regression with functional covariates.

In the fourth project we briefly describe how to extend the MFRS framework to the case where the outcome of interest is binary. In addition, we present a method on how to select import points in the context of kernel logistic regression (KLR) by extending the elastic net via Tikhonov regularization. This project should demonstrate the general approach on how to extend the MFRS framework to other outcomes in the GLM family as well.

CHAPTER I

Introduction

1.0.1 Motivation

With the pervasiveness of sensor data, real-time physiological signals and behavioral data are often collected in many biomedical studies [49]. Data are often collected at a high frequency through these mobile devices. Trying to relate these high frequency data with health outcomes poses a challenge and standard regression techniques is often inadequate to handle such a relationship. With the advent of new technologies that create devices that can bring critical care levels of monitoring to the population at large [24], the need to extract useful information and avoid data overload is crucial. The high frequency data must be summarized in a way that does not throw out important signals, while at the same time avoids the "noise" that is sure to come about through this type of data. While the motivation for this dissertation is using data collected from a tri-axis accelerometer, the framework that is presented in this dissertation can apply to any type of sensor data sampled at high frequency.

1.0.2 Accelerometer Data

There are several different devices of accelerometers available such as the ActiGraph GT3X+ (ActiGraph, Pensacola, FL) and Actical (Phillips Respironics, Bend, OR), among others. Raw accelerometer data is often collected in high-resolution via

high-frequency signals sampling over the range of 30-100 Hz. Placing the accelerometer on the hip or wrist as a means of monitoring physical activity is becoming increasingly common; see for example, [10, 11, 2, 27]. The existing commercial software on these devices provides activity counts (AC), or steps, which are calculated from the raw tri-axis accelerometer measurements using proprietary algorithms. A known caveat with such data is that the exact meaning of the AC is not always clear. Different devices provide various types of AC measures making it difficult to compare across devices [2, 11]. There are general approaches to calculating the AC. One way is with a counter that is used to add up the number of times a signal crosses a preset threshold. Since the range of the raw accelerometer data is between -2g to 2g, the value of 0 could be used (which is known as the zero-crossing method). See below Figure1; $ActiGraph, LLC^{©}$)

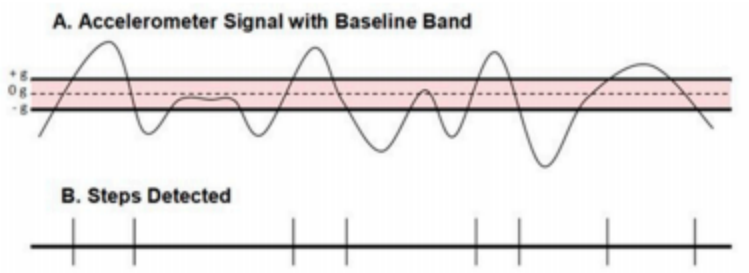


Figure 1. Crossing Threshold Mode using the accelerometer raw signal (**A**) for counting steps; (**B**) A step is detected when the signal passes beyond the baseline band (——) and crossed the midline (- - - -) of the band (ActiGraph, LLC.).

Bai et al. (2016) [2] provides a measure called the Activity Index (AI), which is calculated based on user-defined epochs from the raw tri-axis accelerometer data. The AI is calculated with the focus on the variability of raw acceleration signals that is then converted to a single time-domain functional. Let $\sigma^2_{im}(t; H)$ denote the variability of the raw accelerometer signal for individual i at time t for the epoch of length H along axis m. Typically, H will be a minute in length. The Activity

Index for a specified epoch, H for individual i at time t is denoted by $AI_i(t:H) = \sqrt{max(\frac{1}{3}\{\sum_{m=1}^{3}\sigma_{im}^2(t;H) - \bar{\sigma^2}_i\}, 0)}$ where $\bar{\sigma^2}_i$ is the systematic noise variance when the device is at rest. This is calculated by taking the sums of the variances of three axis during the time points that the device is at rest. This new functional measure has been shown to correctly classify certain physical activity levels better than the AC and to allow a comparison of the AI across different devices [2, 1]. A tri-axis accelerometer generates AC on three axes. When the device is worn at the hip, often just the vertical axis/ axis 1, which is the dominant plane of movement, is used [10]. Another common summary done with tri-axis AC data is the calculation of vector magnitude (VM). VM is calculated as the Euclidean norm involving the three axes of AC [10, 12]. As there is no dominant plane of movement when the device is worn at the wrist a single axis alone would not provide sufficient information of physical activity [10].

Typical goals of using accelerometer data is to categorize the physical activity into various categories such as heavy or light physical activity [2, 10], or to measure metabolic equivalents (METS) which is typically calculated by converting VO2 by dividing the oxygen intake by 3.5 ml / (kg.min). These studies focus on identifying specific cut-points for the physical activities and using classification methods such as receiver operating characteristic curve, (ROC) curves to identify the sensitivity and specificity of specific cut-points. Often, only a summary of the total daily AC count is used for the above analysis instead of using the entire functional curve [37]. These summaries include extracted time-domain features or frequency domain features [13, 46] from the AC. Those features are used for prediction in a regression equation or to classify physical activity types [41].

More recently, to relax the excessive data compression researchers consider using the entire functional AC curve through functional data analysis techniques [18, 3, 29, 46]. The accelerometer data can be viewed as a functional data analysis (FDA)

problem treating a person's captured acceleration (or AC) over time as a function of physical activities. Further details on current methods being used to retrieve and interpret accelerometer data can be found in [56]. One of the main questions of interest that we have is whether various health outcomes such as blood pressure and obesity are related to a person's movement throughout the day which we will explore in detail in Chapters (III) and (IV).

1.0.3 Functional Regression

There has been much attention in recent years to functional data analysis (FDA) where either predictors or covariates, response, or both, are functional as opposed to scalar in nature [39, 9, 8, 57, 16, 34]. In this dissertation, we focus on the methodology that allows us to relate multiple functional covariates to a scalar outcome in a non-linear way in the presence of other scalar covariates.

To proceed, let us introduce some notation. Let $L^2(\mathcal{T})$ be the class of square-integrable functions on a compact set $\mathcal{T}$. This is a separable Hilbert space with inner product $< f\ ,\ g >:= \int_{\mathcal{T}} fg$ for $f,\ g \in L^2(\mathcal{T})$. Consider a probability space $(\Omega, \mathcal{F}, P)$, where Z denotes a functional random variable that maps into $L^2(\mathcal{T})$, namely $Z : \Omega \mapsto L^2(\mathcal{T})$. Define $L^2(\Omega) := \{Z : (\int_{\Omega} \|Z\|^2\, dP)^{\frac{1}{2}} < \infty\}$, the L^2-norm $\|Z\|^2 = < Z\ ,\ Z >$ and assume $Z \in L^2(\Omega)$ in the rest of this paper.

For convenience, we also assume that Z is mean centered, namely $E(Z) = 0$. Historically, Functional Linear Models (FLM) (e.g. [8, 9, 57]) are proposed to relate a functional covariate Z with a mean-centered scalar outcome y, in which the optimal solution of the unknown functional parameter $b \in L^2(\mathcal{T})$ is typically obtained by minimizing the following goodness-of-fit criterion: $\inf_{b \in L^2(\mathcal{T})} E(y - < b, Z >)^2$. Consequently, such solution satisfies functional model $y = < b, Z > + \epsilon$. Here the error term ϵ is a mean zero random variable uncorrelated with Z.

Equivalently, we may write $E(y|Z) = \int_{\mathcal{T}} Z(t)b(t)dt$ for the mean centered scalar

y. As suggested in the literature, we may solve the above least-squared optimization by expanding Z in terms of certain basis functions. In this paper, we focus on the utility of functional principal component analysis (FPCA) to perform decomposition of the functional Z. By the Karhunen-Loève expansion (e.g. [5, 22, 21]) we can write $Z(t) = \sum_{k=1}^{\infty} \sqrt{\varsigma_k}\xi_k\phi_k(t)$, where $\varsigma_k > 0$ are the eigenvalues, and loadings $\xi_k := \frac{1}{\sqrt{\varsigma_k}} < Z, \phi_k >$ satisfy (i) mean zero, $E(\xi_k) = 0$; (ii) variance one, $E(\xi_k\xi_j) = 1$ for $k = j$; and (ii) uncorrelated, $E(\xi_k\xi_j) = 0$ for $k \neq j$. Then, the mean model may be rewritten as follows,

$$E(y|Z) = \sum_{k=1}^{\infty} \beta_k\xi_k, \tag{1.0.1}$$

where coefficients $\beta_k =< b, \sqrt{\varsigma_k}\phi_k >$, $k = 1, \cdots$, which are unknown due to the unknown b. Equation (1.0.1) presents a linear dynamic system between the standardized principal components (PCs) ξ_k of functional predictor Z and scalar outcome y. On these lines of research, Müller and Yao (2008) proposed the seminal Functional Additive Model (FAM) that extends (1.0.1) by allowing a nonparametric form of the conditional mean model with respect to FPCA coefficients (or features), which takes the following form:

$$E(y|Z) = \sum_{k=1}^{\infty} f_k(\xi_k), \tag{1.0.2}$$

where f_k is a fully unspecified non-linear function. It is obvious that in Müller and Yao's FAM (1.0.2), the relationship between Z and y is assured to be additive in the individual coefficient (or feature) components ξ_k's. Regularization is often needed for both (1.0.1) and (1.0.2) in order to deal with these infinite-dimensional unknowns. One of the challenges concerning regularization for (1.0.2) lies in the technical treatment on the function space. Müller and Yao (2008) [36] proposed truncation (or hard-threshold) of the eigenspace to retain only the leading components that explain

the majority of the total variation in Z. Zhu, Yao and Zhang (2012) [57] proposed a regularization of the functions f_k using the powerful COSSO method [32]. One advantage for this kind of regularization method is that sums of higher order functional principal components are allowed to be potentially included in the fitted model, if they make stronger contributions to the functional relationship than the leading functional principal components. Zhu et al.'s method [57] begins with an additive model $E(y|Z) = \sum_{k=1}^{s} f_k(\xi_k)$, where s represents some initial degrees of truncation to specify the total number of additive components to be considered. Then the use of COSSO helps simultaneously regularize and select important functional components among the s functions f_k. Although the above discussion was based on a single functional predictor Z in mind, it is appealing to extend such framework with multiple functional predictors. However, when multiple functional predictors are considered, it is not clear if the above additive model specification remains suitable to handle the complexity, especially a non-additive relationship may be of interest to understand the association between a scalar outcome and multiple functional predictors. In effect, from both perspectives of theoretical advances and application needs, relaxing the additive relationship is an important task in the functional data analysis.

Alternatively, there are some methods (e.g. [34, 16]) in the literature that do not use the strategy of decomposing Z into its functional components. Ferraty, Mas and Vieu [16] took a different approach. Instead they modeled $y = r(Z) + \epsilon$ where the model only assumed smoothness of the operator r. They called this a doubly infinite dimensional problem [15] where the functional data and unknown function r are both infinitely dimensional. Instead of using a basis expansion on Z, they used kernel methods and estimated r with the classical Nadaray-Watson estimate where

$$\hat{r}(Z) = \frac{\sum_{k=1}^{n} y_k K(h^{-1}||Z_k - Z||)}{\sum_{k=1}^{n} K(h^{-1}||Z_k - Z||)}. \tag{1.0.3}$$

The norm $\|\cdot\|$ can be defined on $L(\Omega)^2$ as discussed above, and K is some posi-

tive semi-definite Kernel. The advantage to this approach is that there is no need to decompose the functional Z into its functional principal components. All of the above models were made for a single functional predictor. However, there is not much literature on functional regression in the presence of multiple functional predictors. Fan, James and Radchenko (2015) [14] proposed functional additive regression for multiple functional predictors using a model $E(y|Z^1, \cdots, Z^p) = \sum_{j=1}^{p} f_j(Z^j)$ for p functional predictors Z^j's with p unknown functions f_j's. In addition, sparsity is allowed in the functional predictors by minimizing the penalized L_2-norm: $\frac{1}{2}\left\|\mathbf{Y} - \sum_{j=1}^{p} \mathbf{f}_j\right\|_2^2 + \sum_{j=1}^{p} \rho_{\lambda_n}(\frac{1}{\sqrt{n}} \|f_j\|_2)$ where $\mathbf{Y}$ is the $n \times 1$ vector of outcomes, $\mathbf{f}_j$ is the $n \times 1$ vector with entry's consisting of the jth functional covariate evaluated at f_j for each subject, $\rho_{\lambda_n(\cdot)}$ is a penalty function and $\lambda_n > 0$ being a regularization tuning parameter. In [23], they proposed a Karhunen-Loève for multivariate functional data. Letting the p multivariate functional data, $Z(\mathbf{t}) = (Z^1(t_1), \cdots, Z^p(t_p)) \in \mathcal{R}^p$, they proposed decomposing $Z(\mathbf{t})$ as $Z(\mathbf{t}) = \sum_{k=1}^{\infty} \rho_k \phi_k(\mathbf{t})$ where the mean zero random variable ρ_k is the projection or inner product of Z onto the function ϕ_k using a specially defined inner product (see [23] for details).

This dissertation extends the methodologies presented above under a more useful yet challenging modeling framework with non-additive relationships between multiple functional predictors and the scalar outcome.

CHAPTER II

Multivariate Functional Regression and Selection (MFRS) Framework

2.1 Introduction

2.1.1 Least Squares Kernel Machine (LSKM)

Liu, Lin and Ghosh (2007) [33] proposed a semi-parametric regression model $y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + h(\mathbf{z}_i) + \epsilon_i$, in that we use least-squares kernel machine to analyze multi-dimensional genetic pathways denoted by $\mathbf{z}_i$. In their model, parameter $\boldsymbol{\beta}$ needs to be estimated for $\mathbf{x}$, some vector of clinical covariates, with $\mathbf{z}$ being a vector of gene expressions within a pathway that is potentially related to the outcome via a non-parametric function h. Function h is assumed to lie in a reproducing kernel Hilbert space (RKHS), $\mathcal{H}_\mathcal{K}$, generated by a positive definite kernel function $\mathcal{K}(\cdot,\cdot)$. For ease of exposition, we suppress the bandwidth for the kernel $\mathcal{K}$ in the following discussion. One can estimate $\boldsymbol{\beta}$ and h by maximizing the scaled penalized likelihood function:

$$J(h,\boldsymbol{\beta}) = -\frac{1}{2}\sum_{i=1}^{n}\{y_i - \mathbf{x}_i^\top\boldsymbol{\beta} - h(\mathbf{z}_i)\}^2 - \frac{1}{2}\lambda_1 \|h\|^2_{\mathcal{H}_\mathcal{K}}\,, \tag{2.1.1}$$

where $\lambda_1 > 0$ is the tuning parameter and $\|\cdot\|_{\mathcal{H}_\mathcal{K}}$ is the norm of the RKHS.

Solving (2.1.1) turns out to be mathematically equivalent to solving the normal

equations [53, 33] from the following linear mixed-effects model (LMM):

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{h} + \boldsymbol{\epsilon}, \tag{2.1.2}$$

where $\mathbf{h}$ is an $n \times 1$ vector of random effects with distribution $N(\mathbf{0}, \tau\mathbf{K})$, and n-dimensional vector error term $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$, with $\tau = \lambda_1^{-1}\sigma^2 > 0$, and $\mathbf{K}$ being an $n \times n$ matrix whose (i, j)th element is $\mathcal{K}(\mathbf{z}_i, \mathbf{z}_j)$. Although there is a closed form solution for a fixed λ_1 in maximizing (2.1.1), one remarkable advantage of solving (2.1.1) through the numerical procedure of LMM is most advocated in the literature [30] where we can estimate λ_1 easily as part of the estimation of the variance components of the LMM. So, instead of using cross-validation or other information-based tuning methods, we can solve simultaneously for all the parameters in (2.1.1) as pointed out in [33]. This kernel machine regression model allows us to consider a non-linear relationship for multiple covariates in a non-additive way in a similar fashion. We will extend this framework by incorporating FPCA to handle multiple functional covariates. Assuming that function h belongs to an RKHS, we can use existing software packages for solving LMMs to estimate h and $\boldsymbol{\beta}$ and λ_1 simultaneously.

In addition, Liu, Lin and Ghosh [33] develops variable selection procedures on the feature vector $\mathbf{z}$ by defining kernel machine types of AIC and BIC. Furthermore, testing the variance component $\tau = 0$ is useful to test the global effect of the feature vector $\mathbf{z}$. It is worth noticing that for high dimensional features associated with Z, feature selection based on AIC and BIC (i.e. L_0 penalty approach) can be time consuming or even computationally prohibited. Thus, a computationally effective feature selection procedure is appealing in real-world application. We adopt the sparse regularization approach to this analytic purpose.

2.1.2 Feature Selection

For motivation of our proposed model, we now present a brief review on the group lasso [55], sparse group lasso [47] and non-negative garrote [6]. Note that for both mean models (1.0.1) and (1.0.2) one needs to truncate the series from the Karhunen-Loève expansion. Regularization helps reduce from an infinite number of terms to a sum of finite terms. Yuan and Lin (2007) [55] proposed the group lasso which solves the convex optimization problem:

$$\min_{\boldsymbol{\beta}\in\mathcal{R}^p} \left\| \mathbf{Y} - \sum_{\ell=1}^{L} \mathbf{X}^{\ell}\boldsymbol{\beta}^{\ell} \right\|_2^2 + \lambda \sum_{\ell=1}^{L} \left\|\boldsymbol{\beta}^{\ell}\right\|_2, \tag{2.1.3}$$

where L is the total number of groups of covariates and $\mathbf{X}^{\ell}$ refers to a subset of covariates associated with group ℓ. Friedman, Hastie and Tibshirani (2013) [47] extended the group lasso to allow within-group sparsity, the so-called sparse group lasso (SGL), given as

$$\min_{\boldsymbol{\beta}\in\mathcal{R}^p} \left\| \mathbf{Y} - \sum_{\ell=1}^{L} \mathbf{X}^{\ell}\boldsymbol{\beta}^{\ell} \right\|_2^2 + \lambda(1-\delta) \sum_{\ell=1}^{L} \left\|\boldsymbol{\beta}^{\ell}\right\|_2 + \lambda\delta \left\|\boldsymbol{\beta}\right\|_1, \tag{2.1.4}$$

where $\delta \in [0,1]$ and the additional ℓ_1-norm penalty term on $\boldsymbol{\beta}$ encourages individual sparsity, while the first penalty targets sparsity at the group level. It is easy to see that group lasso is a special case of the SGL when $\delta = 0$.

The non-negative garrote proposed by Breiman (1995) [6] is useful for variable selection, which invokes a scaled version of least squares estimation given by:

$$\arg\min_{\mathbf{d}} \frac{1}{2} \left\| \mathbf{Y} - \tilde{\mathbf{X}}\mathbf{d} \right\|^2 + \lambda \sum_{j=1}^{p} d_j, \text{ subject to } d_j \geq 0, \forall j, \tag{2.1.5}$$

where $\tilde{\mathbf{X}} = (\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_p)$ is a matrix of size $n \times p$ with columns $\tilde{\mathbf{x}}_j = \mathbf{x}_j \hat{\beta}_j^{OLS}$, where $\hat{\beta}_j^{OLS}$ is the least squares estimate from the unconstrained optimization $\text{argmin}_{\boldsymbol{\beta}} \frac{1}{2} \left\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\right\|^2$.

For covariates unrelated to y the corresponding scaling factor d_j helps shrink estimates $\hat{\beta}_j^{OLS}$ towards 0.

In the presence of multiple functional covariates considered in this dissertation, if we turn each functional into its principal feature components as done by FPCA, then we would end up with a similar setting where each functional, Z, forms naturally its own group consisting of functional principal components, and within such a group sparsity may be enforced. Thus, for FLM models, we can use SGL with the FPC features to perform model selection on functional covariates. Any group that is knocked out in SGL would correspond to a functional that is not needed in the model. However, for a nonlinear relationship between an outcome and a functional covariate, more work is needed. That is precisely where LSKM comes in. We propose a model that uses functional data in the LSKM framework while simultaneously performing feature selection in a manner similar to the non-negative garrote.

2.2 Proposed Model

Let $\mathbf{z}_i^\ell = (\xi_1^\ell, \ldots, \xi_{s_\ell}^\ell)_i^\top$ be the vector of FPC features from the ith observation of the functional covariate Z^ℓ and let $\vec{\mathbf{z}}_i = [(\mathbf{z}_i^1)^\top, \ldots, (\mathbf{z}_i^p)^\top]^\top$ be the grand vector of all FPC features from all p functional covariates. In total there are p groups with $s = \sum_{\ell=1}^p s^\ell$ many FPC features, and $\vec{\mathbf{z}}_i \in \mathcal{R}^s$. We consider the following functional kernel regression model:

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + h(\vec{\mathbf{z}}_i) + \epsilon_i,\ i = 1, \cdots, n, \tag{2.2.1}$$

where $\boldsymbol{\beta} \in \mathcal{R}^q$, $h \in \mathcal{H}_\mathcal{K}$, with $\mathcal{H}_\mathcal{K}$ being the functional space generated by a Mercer kernel $\mathcal{K}$, and $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. Model (2.2.1) allows for not only non-linear but also non-additive relationships with multiple functional covariates $Z^\ell, \ell = 1, \ldots, p$, and a scalar response, y. We aim to estimate and select important functional covariates that

are related to the outcome of interest, while regularize the FPC features within each functional covariate, simultaneously. To proceed, we introduce a new s-dimensional scaling vector $\boldsymbol{\gamma} \in \mathcal{R}^s$, $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{s_1}, \ldots, \gamma_s)^\top$; similar to Beiman's [6] non-negative garrote method, we set $\boldsymbol{\gamma} \circ \vec{\mathbf{z}}_i = (\gamma_1 \xi_1^1, \ldots, \gamma_{s_1} \xi_{s_1}^1, \ldots, \gamma_s \xi_{s_p}^p)_i^\top$ a new vector of weighted FPC features by $\boldsymbol{\gamma}$ via the Hadamard product (i.e. elementwise product). Obviously, when element, say γ_j, is equal to zero, the corresponding FPC feature ξ_j will not be selected into the set of important FPCs.

We estimate the unknowns in (2.2.1) as well as $\boldsymbol{\gamma}$ by minimizing the following penalized likelihood function:

$$\begin{aligned} \min_{h,\boldsymbol{\beta},\boldsymbol{\gamma}} J_1(h, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \min_{h,\boldsymbol{\beta},\boldsymbol{\gamma}} \frac{1}{2n} \sum_{i=1}^{n} \{y_i - \mathbf{x}_i^\top \boldsymbol{\beta} - h(\boldsymbol{\gamma} \circ \mathbf{z}_i)\}^2 \\ + \frac{1}{2} \lambda_1 \left\| h \right\|_{\mathcal{H}_\mathcal{K}}^2 + \lambda_2 \rho(\boldsymbol{\gamma}; \delta), \end{aligned} \tag{2.2.2}$$

where $\lambda_1 > 0$, and $\lambda_2 > 0$ are tuning parameters, $\boldsymbol{\gamma} = ((\boldsymbol{\gamma}^1)^\top, \ldots, (\boldsymbol{\gamma}^p)^\top)^\top$ with $\boldsymbol{\gamma}^\ell$ being an $s^\ell \times 1$ vector associated with the ℓth functional covariate FPC features $\mathbf{z}^\ell$, and penalty $\rho(\boldsymbol{\gamma}; \delta)$ may be specified according to a certain regularized method. For example, in the case of sparse group lasso we take $p(\boldsymbol{\gamma}; \delta) = (1 - \delta) \sum_{\ell=1}^{p} \left\| \boldsymbol{\gamma}^\ell \right\|_2 + \delta \left\| \boldsymbol{\gamma} \right\|_1$, $\delta \in [0, 1]$. Typically, δ is predetermined and set to 0.95 or 0.05 depending on the trade off between group and within group sparsity. Here the factor $(1 - \delta)$ controls relative group sparsity to individual sparsity of each functional predictor Z^ℓ. In the meanwhile, a large tuning parameter for λ_2 would set certain groups of FPC features $\boldsymbol{\gamma}^\ell$ entirely equal to zero with by the corresponding $\boldsymbol{\gamma}^\ell = 0$. An equivalent formulation of (2.2.2) results in minimizing the following objective function:

$$\begin{aligned} \min_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}} J_2(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \min_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}} \frac{1}{2n} \sum_{i=1}^{n} \left\{ y_i - \mathbf{x}_i^\top \boldsymbol{\beta} - \sum_{k=1}^{n} \alpha_k \mathcal{K}(\boldsymbol{\gamma} \circ \vec{\mathbf{z}}_i, \boldsymbol{\gamma} \circ \vec{\mathbf{z}}_k) \right\}^2 \\ + \frac{1}{2} \lambda_1 \boldsymbol{\alpha}^\top \mathbf{K}(\boldsymbol{\gamma}; \mathbf{Z}) \boldsymbol{\alpha} + \lambda_2 \rho(\boldsymbol{\gamma}; \delta), \end{aligned} \tag{2.2.3}$$

where $\mathbf{K}(\boldsymbol{\gamma}; \mathbf{Z})$ is an $n \times n$ matrix whose (i,k)th element is $[\mathbf{K}(\boldsymbol{\gamma}; \mathbf{Z})]_{ik} = \mathcal{K}(\boldsymbol{\gamma} \circ \vec{\mathbf{z}}_i, \boldsymbol{\gamma} \circ \vec{\mathbf{z}}_k)$. Lemma 1 below establishes the equivalence between (2.2.2) and (2.2.3), which is crucial in our estimation procedure.

.

Lemma 1. *A solution $(\hat{h}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$ is a minimizer of (2.2.2) if and only if $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$ is a minimizer of (2.2.3), where $\hat{h}(\hat{\boldsymbol{\gamma}} \circ \vec{\mathbf{z}}) = \sum_{k=1}^{n} \hat{\alpha}_k \mathcal{K}(\hat{\boldsymbol{\gamma}} \circ \vec{\mathbf{z}}, \hat{\boldsymbol{\gamma}} \circ \vec{\mathbf{z}}_k)$.*

Proof. It suffices to show that for any $J_1(h, \boldsymbol{\beta}, \boldsymbol{\gamma})$ in (2.2.2) we can always find $\boldsymbol{\alpha} \in \mathcal{R}^n$ such that $J_1(\tilde{h} = \sum_{i=1}^{n} \alpha_i \mathcal{K}(\cdot, \boldsymbol{\gamma} \circ \vec{\mathbf{z}}_i), \boldsymbol{\gamma}, \boldsymbol{\beta}) \leq J_1(h, \boldsymbol{\beta}, \boldsymbol{\gamma})$ where $\tilde{h}$ is the projection of h onto the linear spanned space given by $span\{\mathcal{K}(\cdot, \boldsymbol{\gamma} \circ \vec{\mathbf{z}}_i), \cdots, \mathcal{K}(\cdot, \boldsymbol{\gamma} \circ \vec{\mathbf{z}}_n)\}$. For any h we can write $h = h^{\perp} + \tilde{h}$ where $h^{\perp} \in span\{\mathcal{K}(\cdot, \boldsymbol{\gamma} \circ \vec{\mathbf{z}}_1), \cdots, \mathcal{K}(\cdot, \boldsymbol{\gamma} \circ \vec{\mathbf{z}}_n)\}^{\perp}$. Since $\mathcal{H}_k$ is a reproducing kernel Hilbert space we can rewrite (2.2.2) as follows:

$$
\begin{aligned}
J_1(h, \boldsymbol{\gamma}, \boldsymbol{\beta}) = \frac{1}{2n} \sum_{i=1}^{n} \{ y_i - \mathbf{x}_i^{\top} \boldsymbol{\beta} - < h, \mathcal{K}(\cdot, \boldsymbol{\gamma} \circ \vec{\mathbf{z}}_i) > \}^2 \\
+ \frac{1}{2} \lambda_1 \left\| h \right\|_{\mathcal{H}_k}^2 + \lambda_2 \rho(\boldsymbol{\gamma}; \delta).
\end{aligned}
$$

Since $< h^{\perp}, \mathcal{K}(\cdot, \boldsymbol{\gamma} \circ \vec{\mathbf{z}}_i) > = 0$ for every i, we get

$$
\begin{aligned}
J_1(h, \boldsymbol{\gamma}, \boldsymbol{\beta}) &= \frac{1}{2n} \sum_{i=1}^{n} \left\{ y_i - \mathbf{x}_i^{\top} \boldsymbol{\beta} - \sum_{k=1}^{n} \alpha_k \mathcal{K}(\boldsymbol{\gamma} \circ \vec{\mathbf{z}}_i, \boldsymbol{\gamma} \circ \vec{\mathbf{z}}_k)) \right\}^2 \\
&\quad + \frac{1}{2} \lambda_1 \left\| h^{\perp} + \tilde{h} \right\|_{\mathcal{H}_k}^2 + \lambda_2 \rho(\boldsymbol{\gamma}; \delta) \\
&\geq \frac{1}{2n} \sum_{i=1}^{n} \left\{ y_i - \mathbf{x}_i^{\top} \boldsymbol{\beta} - \sum_{k=1}^{n} \alpha_k \mathcal{K}(\boldsymbol{\gamma} \circ \vec{\mathbf{z}}_i, \boldsymbol{\gamma} \circ \vec{\mathbf{z}}_k)) \right\}^2 \\
&\quad + \frac{1}{2} \lambda_1 \left\| \tilde{h} \right\|_{\mathcal{H}_k}^2 + \lambda_2 \rho(\boldsymbol{\gamma}; \delta) \\
&= J_1(\tilde{h}, \boldsymbol{\gamma}, \boldsymbol{\beta}).
\end{aligned}
$$

□

Theorem 2. *(Existence of optimizers) If the kernel $\mathcal{K}(\cdot, \boldsymbol{\gamma} \circ \vec{\mathbf{z}})$ is continuous with respect to $\boldsymbol{\gamma} \in \mathcal{R}^s$, then there exists a global minimizer $(\hat{h}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$ for the optimization problem* (2.2.2).

Proof. We will assume we are using the penalty function for sparse group lasso but this proof can easily be modified for other convex penalty functions. We will fix $\lambda_1 = \lambda_2 = \delta = 1$. We will assume $\beta \in \mathcal{R}$ and that the design matrix $\mathbf{X}$ (or vector in this case) is scaled to have norm 1. The case of $\boldsymbol{\beta} \in \mathcal{R}^q$ will follow along similar lines. Let $\boldsymbol{\gamma} \in D_3$ where $D_3 = \{\boldsymbol{\gamma} : \|\boldsymbol{\gamma}\|_1 \leq \frac{1}{2n} \|\mathbf{Y}\|_2^2\}$. Define $f(\boldsymbol{\gamma}) = \|\mathbf{K}(\boldsymbol{\gamma}; Z)\| = \eta_{max}(\mathbf{K}(\boldsymbol{\gamma}; Z)) \geq 0$ where $\eta_{max}(\mathbf{K}(\boldsymbol{\gamma}; Z))$ is the largest eigenvalue of $\mathbf{K}(\boldsymbol{\gamma}; Z)$ with the operator norm (the norm of $\mathbf{K}(\boldsymbol{\gamma}; Z)$) defined in its usual way $\|\mathbf{K}(\boldsymbol{\gamma}; Z)\| = sup\{\|\mathbf{K}(\boldsymbol{\gamma}; Z)\mathbf{x}\|_2^2 : \|\mathbf{x}\|_2^2 = 1\}$. Since D_3 is compact and $\mathbf{K}(\boldsymbol{\gamma}; Z)$ is continuous with respect to $\boldsymbol{\gamma}$ it achieves its maximum over D_3 so we can define $\eta^\star = sup_{\boldsymbol{\gamma} \in D_3} f(\boldsymbol{\gamma}) \geq 0$. Define D_2 as

$$D_2 = \{\beta :| \beta |\leq (1 + \eta^\star) \|\mathbf{Y}\|_2\}.$$

Let

$$b^\star = (1 + \eta^\star) \|\mathbf{Y}\|_2 \geq 0$$

Define D_1 as

$$D_1 = \{\boldsymbol{\alpha} : \|\boldsymbol{\alpha}\|_2 \leq \sqrt{n}(\|\mathbf{Y}\|_2 + b^\star)\}$$

Since D_1, D_2 and D_3 are compact there exists a $(\boldsymbol{\alpha}^\star, \beta^\star, \boldsymbol{\gamma}^\star)$ such that $J_2(\boldsymbol{\alpha}^\star, \beta^\star, \boldsymbol{\gamma}^\star) \leq J_2(\boldsymbol{\alpha}, \beta, \boldsymbol{\gamma})$ for all $(\boldsymbol{\alpha}, \beta, \boldsymbol{\gamma}) \in D_1 \times D_2 \times D_3$. Remark: we have $J_2(\mathbf{0}, 0, \mathbf{0}) = \frac{1}{2n} \|\mathbf{Y}\|_2^2$ and $(\mathbf{0}, 0, \mathbf{0}) \in D_1 \times D_2 \times D_3$. We claim that $(\boldsymbol{\alpha}^\star, \beta^\star, \boldsymbol{\gamma}^\star)$ is a global minimizer. This is a proof by contradiction. Suppose that there exists $(\tilde{\boldsymbol{\alpha}}, \tilde{\beta}, \tilde{\boldsymbol{\gamma}}) \notin D_1 \times D_2 \times D_3$ where

$J_2(\tilde{\boldsymbol{\alpha}}, \tilde{\beta}, \tilde{\boldsymbol{\gamma}}) < J_2(\boldsymbol{\alpha}^\star, \beta^\star, \boldsymbol{\gamma}^\star)$. We must have that $\tilde{\boldsymbol{\gamma}} \in D_3$ for if not, $J_2(\tilde{\boldsymbol{\alpha}}, \tilde{\beta}, \tilde{\boldsymbol{\gamma}}) \geq \|\tilde{\boldsymbol{\gamma}}\|_1 \geq J_2(\mathbf{0}, 0, \mathbf{0}) \geq J_2(\boldsymbol{\alpha}^\star, \beta^\star, \boldsymbol{\gamma}^\star)$. Let $q_1, \cdots, q_n$ be the orthonormal vectors of $\mathbf{K}(\tilde{\boldsymbol{\gamma}}; Z)$ with its associated eigenvalues $\eta_1 \geq \cdots, \eta_n \geq 0$. We can write out $\tilde{\boldsymbol{\alpha}}, \mathbf{X}, \mathbf{Y}$ in terms of these basis functions where $\tilde{\boldsymbol{\alpha}} = \sum_{i=1}^n < \tilde{\boldsymbol{\alpha}}, q_i > q_i$, $\mathbf{Y} = \sum_{i=1}^n < \mathbf{Y}, q_i > q_i$ and $\mathbf{X} = \sum_{i=1}^n < \mathbf{X}, q_i > q_i$. Let $C_i^{\tilde{\boldsymbol{\alpha}}} =< \tilde{\boldsymbol{\alpha}}, q_i >$, $C_i^{\mathbf{Y}} =< \mathbf{Y}, q_i >$ and $C_i^{\mathbf{X}} =< \mathbf{X}, q_i >$. We have that

$$J_2(\tilde{\boldsymbol{\alpha}}, \tilde{\beta}, \tilde{\boldsymbol{\gamma}}) \geq \frac{1}{2n} \left\| \sum_{i=1}^n C_i^{\mathbf{Y}} q_i - \sum_{i=1}^n C_i^{\mathbf{X}} \tilde{\beta} q_i - \sum_{i=1}^n C_i^{\tilde{\boldsymbol{\alpha}}} \eta_i q_i \right\|_2^2 + \frac{1}{2} \sum_{i=1}^n (C_i^{\tilde{\boldsymbol{\alpha}}})^2 \eta_i,$$

which is equal to $\frac{1}{2n} \sum_{i=1}^n (C_i^{\mathbf{Y}} - C_i^{\mathbf{X}} \tilde{\beta} - C_i^{\tilde{\boldsymbol{\alpha}}} \eta_i)^2 + \frac{1}{2} \sum_{i=1}^n (C_i^{\tilde{\boldsymbol{\alpha}}})^2 \eta_i$. We can minimize the above with respect to $C_i^{\tilde{\boldsymbol{\alpha}}}$ and $\tilde{\beta}$. First, note that for any $\eta_i = 0$ we can let $C_i^{\tilde{\boldsymbol{\alpha}}} = 0$ and it will not effect the expression above. We will then only consider $\eta_i > 0$. Taking the first derivative and setting it equal to zero we get the score equations the minimizer must satisfy (for our minimum $\tilde{\beta}$ and $C_i^{\tilde{\boldsymbol{\alpha}}}$) as

$$\beta = \sum_{i=1}^n C_i^{\mathbf{X}} (C_i^{\mathbf{Y}} - C_i^{\tilde{\boldsymbol{\alpha}}} \eta_i) \tag{2.2.4}$$

$$C_i^{\tilde{\boldsymbol{\alpha}}} = \frac{1}{n + \eta_i} (C_i^{\mathbf{Y}} - C_i^{\mathbf{X}} \tilde{\beta}). \tag{2.2.5}$$

Remark: for the above derivation we used the fact that $1 = \|\mathbf{X}\|_2^2 = \sum_{i=1}^n (C_i^{\mathbf{X}})^2$. Plugging (2.2.5) into (2.2.4) we get that

$$\beta = \frac{\sum_{i=1}^n C_i^X C_i^Y (1 - \frac{\eta_i}{n+\eta_i})}{1 - \sum_{i=1}^n (C_i^X)^2 \frac{\eta_i}{n+\eta_i}} \tag{2.2.6}$$

From (2.2.6) we see that

$$\beta \leq \frac{\sum_{i=1}^n \mid C_i^{\mathbf{X}} C_i^{\mathbf{Y}} \mid}{1 - \sum_{i=1}^n (C_i^{\mathbf{X}})^2 \frac{\eta^\star}{n+\eta^\star}} \leq \frac{\|\mathbf{X}\|_2 \|\mathbf{Y}\|_2}{\|\mathbf{X}\|_2^2 (1 - \frac{\eta^\star}{n+\eta^\star})} \leq \frac{\|\mathbf{Y}\|_2}{(1 - \frac{\eta^\star}{1+\eta^\star})} = b^\star$$

This shows that the β that minimizes J_2 for a given $\boldsymbol{\gamma} \in D_3$ is in D_2. By (2.2.5) we see that $\mid C_i^{\tilde{\boldsymbol{\alpha}}} \mid \leq (\|\mathbf{Y}\|_2 + \|\mathbf{X}\|_2 \|\beta\|_2)$ which implies that the optimal $\boldsymbol{\alpha}$ for the given $\tilde{\boldsymbol{\gamma}} \in D_3$ and $\beta \in D_2$ that minimizes J_2 satisfies $\|\boldsymbol{\alpha}\|_2 \leq \sqrt{n}(\|\mathbf{Y}\|_2 + b^\star)$ which implies that $\alpha \in D_2$. This shows that for any $(\tilde{\boldsymbol{\alpha}}, \tilde{\beta}, \tilde{\boldsymbol{\gamma}}) \notin D_1 \times D_2 \times D_3$ we can find an $(\boldsymbol{\alpha}, \beta, \boldsymbol{\gamma}) \in D_1 \times D_2 \times D_3$ such that $J_2(\tilde{\boldsymbol{\alpha}}, \tilde{\beta}, \tilde{\boldsymbol{\gamma}}) \geq J_2(\boldsymbol{\alpha}, \beta, \boldsymbol{\gamma})$. □

Note that there may exist multiple optimal global minimizers for (2.2.2); Theorem 2 ensures only the existence of optimal solutions but no guarantees for uniqueness due to the fact that (2.2.2) or (2.2.3) is a non-linear and non-convex optimization problem. Remarks: Previously we suppressed the bandwidth parameter of the kernel for the ease of exposition. In both (2.2.2) and (2.2.3) we fix the bandwidth parameter for the kernel to a constant due to identifiability issues with respect to the $\boldsymbol{\gamma}$ parameters. We will provide more details concerning the parameter identifiability later in this chapter.

2.3 Algorithm

To implement our proposed estimation procedure, we require differentiability of the kernel with respect to the scaling factor $\boldsymbol{\gamma}$, and some additional assumptions presented below in order to ensure algorithmic convergence. The first step to solving (2.2.2) is to notice that with fixed $\boldsymbol{\gamma}$, this minimization problem reduces to the equivalent maximization problem in the least squares kernel machine (2.1.1) where the FPC features, $\vec{\mathbf{z}}_i$, are replaced by $\boldsymbol{\gamma} \circ \vec{\mathbf{z}}_i$. As pointed out in Section 1.2, the numerical solution can be obtained in the same fashion as the solution from the linear mixed model (2.1.2). The solution to (2.1.2) includes the optimal tuning parameter λ_1 directly from the REML estimation part of the variance components. In this way there is no need to tune λ_1. Alternatively, you can use cross validation to tune λ_1. In turn with $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and λ_1 being given, we then solve the non-linear and non-convex optimization problem to determine the optimal $\boldsymbol{\gamma}$. Lemma 3 below helps us solve for

$\boldsymbol{\gamma}$.

Lemma 3. *For fixed ($\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, λ_1), minimizing* (2.2.3) *over $\boldsymbol{\gamma}$ is equivalent to minimizing over $\boldsymbol{\gamma}$ the following objective function:*

$$\frac{1}{2n}\left\|\mathbf{F}(\boldsymbol{\gamma})-\tilde{\mathbf{Y}}\right\|_2^2+\lambda_2\rho(\boldsymbol{\gamma};\delta),\ \textit{for each } \lambda_2>0, \tag{2.3.1}$$

where $\mathbf{F}(\boldsymbol{\gamma})=\mathbf{K}(\boldsymbol{\gamma};\mathbf{Z})\boldsymbol{\alpha}$ *and* $\tilde{\mathbf{Y}}=\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}-\frac{n}{2}\lambda_1\boldsymbol{\alpha}$.

Proof. The equivalence of forms become clear once we rewrite (2.2.3) in matrix notation. Equation (2.2.3) can be written as:

$$\min_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}} J_2(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma})=\min_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}}\frac{1}{2n}\left\|\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{K}(\boldsymbol{\gamma};\mathbf{Z})\boldsymbol{\alpha}\right\|_2^2+\frac{1}{2}\lambda_1\boldsymbol{\alpha}^\top\mathbf{K}(\boldsymbol{\gamma};Z)\boldsymbol{\alpha}+\lambda_2\rho(\boldsymbol{\gamma};\delta). \tag{2.3.2}$$

For fixed $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and λ_1, minimizing the function in (2.3.2) with respect to $\boldsymbol{\gamma}$ is equivalent to:

$$\min_{\boldsymbol{\gamma}}\left\{\frac{1}{2n}\left\|\left(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}-\frac{n}{2}\lambda_1\boldsymbol{\alpha}\right)-\mathbf{K}(\boldsymbol{\gamma};Z)\boldsymbol{\alpha}\right\|_2^2+\lambda_2\rho(\boldsymbol{\gamma};\delta)\right\}. \tag{2.3.3}$$

□

Linearizing the function $\mathbf{F}(\boldsymbol{\gamma})$ in (2.3.1) leads to minimizing the following:

$$\min_{\boldsymbol{\gamma}}\frac{1}{2n}\left\|\tilde{\mathbf{Y}}-\sum_{\ell=1}^{p}\nabla_{\boldsymbol{\gamma}}\mathbf{F}^{(\ell)}(\tilde{\boldsymbol{\gamma}})\boldsymbol{\gamma}^\ell\right\|_2^2+\lambda_2\rho(\boldsymbol{\gamma};\delta), \tag{2.3.4}$$

where $\tilde{\mathbf{Y}}=\left(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}-\frac{n}{2}\lambda_1\boldsymbol{\alpha}\right)-\mathbf{F}(\tilde{\boldsymbol{\gamma}})+\nabla_{\boldsymbol{\gamma}}\mathbf{F}(\tilde{\boldsymbol{\gamma}})\tilde{\boldsymbol{\gamma}}$, $\nabla_{\boldsymbol{\gamma}}\mathbf{F}(\tilde{\boldsymbol{\gamma}})$ is the gradient of the function F with respect to $\boldsymbol{\gamma}$ evaluated at $\tilde{\boldsymbol{\gamma}}$ for some $\tilde{\boldsymbol{\gamma}}$, and $\nabla_{\boldsymbol{\gamma}}\mathbf{F}^{(\ell)}(\tilde{\boldsymbol{\gamma}})$ are the columns of $\nabla_{\boldsymbol{\gamma}}\mathbf{F}(\tilde{\boldsymbol{\gamma}})$ associated with the ℓth group of $\boldsymbol{\gamma}^\ell$. This is precisely the form of

the standard sparse group regularization problem

$$\min_{\boldsymbol{\beta}\in\mathcal{R}^p} \frac{1}{2n} \left\| \mathbf{Y} - \sum_{\ell=1}^{p} \mathbf{X}^{\ell}\boldsymbol{\beta}^{\ell} \right\|_2^2 + \lambda_2 \rho(\boldsymbol{\gamma}; \delta).$$

This implies that (2.3.4) presents a standard sparse group regularization problem with a specific choice of penalty function $\rho(\boldsymbol{\gamma}; \delta)$. The convergence of the above iterative search algorithm for updating $\tilde{\boldsymbol{\gamma}}$ for fixed ($\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, λ_1) can be justified by the proximal Gauss-Newton method [40]. In the Appendix we provide some details on the proximal Gauss-Newton method. One of the key assumptions of the proximal Gauss-Newton method is the existence of a local minimizer. This condition is satisfied in the above (2.3.4). This is because according to Theorem 2 there exists a global minimizer. It is easy to show that given ($\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, λ_1), a global minimizer exists for (2.3.4) when minimizing with respect to $\boldsymbol{\gamma}$. If we start our algorithm with a value $\tilde{\boldsymbol{\gamma}}$ within a ball of a certain radius of the global minimizer, we are guaranteed to stay within that ball and converge monotonically to the minimizer under suitable Lipschitz condition of $\nabla_{\boldsymbol{\gamma}} F$. See [40] for more details on the technical conditions on $\nabla_{\boldsymbol{\gamma}} F$ and the radius of the ball.

In summary, we propose the following descent algorithm to search for the optimal solution to the problem given in (2.2.3).

Algorithm 1:

(i) Step 1.1: Perform FPCA (e.g. R package `fdapace`) to extract the functional component scores for the p functional predictors and store them in a grand vector for each individual subject $\vec{\mathbf{z}}_i = [(\mathbf{z}_i^1)^\top, \ldots, (\mathbf{z}_i^p)^\top)]^\top$, $i = 1, \cdots, n$;

(ii) Step 1.2: Initialize $\boldsymbol{\gamma}$ to be a vector of 1's which translates to mapping the original component scores to itself. Set up a grid of possible tuning parameters for λ_1 and λ_2, respectively. Set the kernel bandwidth parameter which may depend on λ_1. For each pair of (λ_1, λ_2) from our grid perform steps Steps 2-4

below.

(iii) Step 2.1: At the $(r+1)$-th step in the algorithm, first solve the LSKM problem with fixed $(\boldsymbol{\gamma}^{(r)}, \lambda_1)$ (based on a closed-form solution) to update $\boldsymbol{\beta}^{(r+1)}$ and $\boldsymbol{\alpha}^{(r+1)}$.

(iv) Step 2.2: Solve the group regularity problem (2.3.4) with fixed $\tilde{\boldsymbol{\gamma}} = \boldsymbol{\gamma}^{(r)}$ and fixed $(\boldsymbol{\alpha}^{(r+1)}, \boldsymbol{\beta}^{(r+1)}, \lambda_1, \lambda_2)$ using the $r+1$ updates from the previous iteration. At this step the proximal Gauss-Newton algorithm produces an update $\boldsymbol{\gamma}^{(r+1)}$ at convergence.

(v) Step 2.3: Repeat steps 2.1-2.2 until convergence.

(vi) Step 3: Perform cross-validation over all pairs of (λ_1, λ_2) to determine the final $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$

It is easy to show that we get a descent method where $J_2(\boldsymbol{\alpha}^{(r+1)}, \boldsymbol{\beta}^{(r+1)}, \boldsymbol{\gamma}^{(r+1)}) \leq J_2(\boldsymbol{\alpha}^{(r)}, \boldsymbol{\beta}^{(r)}, \boldsymbol{\gamma}^{(r)})$. This assumes the convergence of the proximal Gauss-Newton algorithm for Step 2.2. It should be noted that although we proposed a possible starting value for $\boldsymbol{\gamma}$ as a vector of 1's, when there is sparsity within a large number of FPC features, you may consider trying out different starting values that downplay the effect of the many features at hand. To speed up the above algorithm, we propose the following operartional schemes that eliminate setting up the pairs of (λ_1,λ_2) and performing Step 3:

Algorithm 2:

(i) Step 2.1 is done by running the linear mixed model with our initial fixed $\boldsymbol{\gamma}$ from step 1.2 to get λ_1, $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$.

(ii) Step 2.2 is done with solving the group regularity problem (2.3.4) with λ_1, $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ from the previous step using cross-validation (e.g. R package `oem`). At this step the Gauss-Newton algorithm produces an update for $\boldsymbol{\gamma}$ at convergence. We

are running the group regularity problem multiple times. The main difference in the ideal algorithm and the proposed implementation of the algorithm for step 2.3 is that λ_2 is fixed in the descent algorithm, while λ_2 is changing through cross-validation in our proposed implementation algorithm. We see similar algorithms with changing tuning parameters using single index model demonstrated in [38].

(iii) Rerun Step (ii) using the updated $\boldsymbol{\gamma}$ from Step (iii) to get the final estimates for $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$.

Remark: There is no guarantee that the above algorithm will converge to a global minimizer, and the proximal Gauss-Newton method in step 2.2 can only find stationary point. This requires good starting values to begin the search. This indeed is an open problem in the field of nonlinear and nonconvex optimization.

2.4 Theoretical Analysis

Our theoretical analysis focuses on the finite-sample L_2 error bounds for the estimators $(\hat{h}, \hat{\boldsymbol{\gamma}})$ obtained by (2.2.2) or (2.2.3). Consequently, we are able to establish the estimation consistency. We will consider random vectors $\mathbf{z}_1, \ldots, \mathbf{z}_n$ for the purpose of this section which may or may not correspond to the FPC features $\vec{\mathbf{z}_1}, \ldots, \vec{\mathbf{z}_n}$. This work follows along similar lines as those of [57] and [17]. Specifically, we choose the sparse-group-lasso penalty function to establish the estimation consistency. These theoretical analysis may hold for other penalties by slight modifications. For the ease of exposition, we set $\boldsymbol{\beta} = \mathbf{0}$ in this section. For the issue of identifiability, readers refer to the next section for more discussion, including some additional intuition on the behavior of the proposed estimator. For a measurable function $f : L^2(\mathcal{T}) \mapsto \mathcal{R}$, its empirical norm is defined as $\|f\|_n := \sqrt{\frac{1}{n}\sum_{i=1}^n f(Z_i)^2}$. This is a random quantity as being sample dependent. Let Γ be a map from $\mathcal{R}^s \mapsto \mathcal{R}^s$ such that $\Gamma(\mathbf{z}) = \boldsymbol{\gamma} \circ \mathbf{z}$, where operator "$\circ$" denotes the elementwise product between vector $\boldsymbol{\gamma} \in \mathcal{R}^s$ and

$\mathbf{z} \in \mathcal{R}^s$. Sometimes operation $\circ$ may be regarded as the Hadamard product while at other times this notation may be referred to as the composition of two functions. It should be clear from the context which one we are referring to. Each map Γ is clearly defined with a unique $\boldsymbol{\gamma} \in \mathcal{R}^s$. Consider a collection of all scaling map functions $\mathcal{A} = \{\Gamma : \mathcal{R}^s \mapsto \mathcal{R}^s \mid \Gamma(\mathbf{z}) = \boldsymbol{\gamma} \circ \mathbf{z}, \mathbf{z} \in \mathcal{R} \text{ for a fixed } \boldsymbol{\gamma} \in \mathcal{R}^s\}$. Since Γ is a linear (and bounded) operator, $\mathcal{A}$ is a real vector space where $(c_1\Gamma_1 + c_2\Gamma_2)(\mathbf{z}) = c_1\Gamma_1(\mathbf{z}) + c_2\Gamma_2(\mathbf{z})$ with any $c_1, c_2 \in \mathcal{R}$ and $\Gamma_1, \Gamma_2 \in \mathcal{A}$. To perform a group regularity estimation, we define a Sparse Group Lasso penalty which can also be viewed as a norm on $\mathcal{A}$ for a fixed $\delta \in [0, 1]$ as follows:

$$\|\Gamma\|_{SGL} = \delta \sum_{\ell=1}^{p} \left\|\boldsymbol{\gamma}^{\ell}\right\|_2 + (1-\delta) \|\boldsymbol{\gamma}\|_1 . \tag{2.4.1}$$

Then, we need to perform the following constrained optimization:

$$\min_{\Gamma \in \mathcal{A}, h \in \mathcal{H}_{\mathcal{K}}} \|\mathbf{Y} - h \circ \Gamma\|_n^2 + \lambda_1 \|h\|_{H_K}^2 + \lambda_2 \|\Gamma\|_{SGL} \tag{2.4.2}$$

where $\|\mathbf{Y} - h \circ \Gamma\|_n^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - (h \circ \Gamma)(\mathbf{z}_i))^2$. Let $\hat{h} \circ \hat{\Gamma}$ be the minimizer of (2.4.2). Let $h_0 \circ \Gamma_0$ be the true function for the model below,

$$y_i = (h_0 \circ \Gamma_0)(\mathbf{z_i}) + \epsilon_i, i = 1, \ldots, n. \tag{2.4.3}$$

Above we have abused notation slightly by considering $h \circ \Gamma$ as an $n \times 1$ vector with ith entry $h(\Gamma(\mathbf{z}_i))$ in (2.4.2) as well as considering it as a function composition from $\mathcal{R}^s \mapsto \mathcal{R}$ in (2.4.3). It should be clear from the context which notation we are referring to in the following presentation. Lemma 3 below provides the essential finite-sample inequalities that lead us to the estimation consistency.

Lemma 4. *(Basic Inequality)*

$$\begin{aligned}\left\|\hat{h}\circ\hat{\Gamma}-h_0\circ\Gamma_0\right\|_n^2+\lambda_1\left\|\hat{h}\right\|_{\mathcal{H}_\mathcal{K}}^2+\lambda_2\left\|\hat{\Gamma}\right\|_{SGL}\leq\\ 2(\boldsymbol{\epsilon},\hat{h}\circ\hat{\Gamma}-h_0\circ\Gamma_0)_n+\lambda_1\left\|h_0\right\|_{\mathcal{H}_\mathcal{K}}^2+\lambda_2\left\|\Gamma_0\right\|_{SGL},\end{aligned}\tag{2.4.4}$$

where $2(\boldsymbol{\epsilon},\hat{h}\circ\hat{\Gamma}-h_0\circ\Gamma_0)_n=\frac{2}{n}\sum_{i=1}^n\epsilon_i\left((\hat{h}\circ\hat{\Gamma})(\mathbf{z}_i)-(h_0\circ\Gamma_0)(\mathbf{z}_i)\right)$.

Proof. This is made obvious by noticing that
$\left\|\mathbf{Y}-\hat{h}\circ\hat{\Gamma}\right\|_n^2+\lambda_1\left\|\hat{h}\right\|_{\mathcal{H}_\mathcal{K}}^2+\lambda_2\left\|\hat{\Gamma}\right\|_{SGL}\leq$
$\left\|\mathbf{Y}-h_0\circ\Gamma_0\right\|_n^2+\lambda_1\left\|h_0\right\|_{\mathcal{H}_\mathcal{K}}^2+\lambda_2\left\|\Gamma_0\right\|_{SGL}$.

Substitute (2.4.3) in for $\mathbf{Y}$ and we have the inequality. □

We need the following notation before presenting our theoretical guarantees. We let $\mathcal{N}(\delta,M,P_n)$ denote the minimal δ covering number of the function set $\mathcal{M}$ under the empirical metric P_n based on the random vectors $\mathbf{z}_1,\cdots,\mathbf{z}_n$. Let $N=\mathcal{N}(\delta,M,P_n)$. This means that there exist functions $m_1,\cdots,m_N$ (not necessarily in the set $\mathcal{M}$) such that for every function $m\in\mathcal{M}$ there exists a $j\in\{1,\cdots,N\}$ such that $\left\|m-m_j\right\|_{P_n}\leq\delta$ where $\left\|m-m_j\right\|_{P_n}=\sqrt{\frac{1}{n}\sum_{i=1}^n\{m(\mathbf{z}_i)-m_j(\mathbf{z}_i)\}^2}$. We define the δ-entropy of $\mathcal{M}$ for the empirical metric, P_n, as $H(\delta,\mathcal{M},P_n):=log(\mathcal{N}(\delta,\mathcal{M},P_n))$. Let $\mathcal{B}=\left\{b:=b(h,\Gamma)=\frac{h\circ\Gamma-h_0\circ\Gamma_0}{\|h\|_{\mathcal{H}_\mathcal{K}}^2+\|h_0\|_{\mathcal{H}_\mathcal{K}}^2+\|\Gamma\|_{SGL}^2+\|\Gamma_0\|_{SGL}^2}|h\in\mathcal{H}_\mathcal{K},\Gamma\in\mathcal{A}\right\}$. We need the following assumptions:

Assumption 1. *The error term ϵ is uniformly sub-Gaussian; that is for constants C_1, and C_2*

$$\sup_n\max_{i=1,\cdots,n}C_1^2\left\{E\left(\exp^{\frac{\epsilon_i^2}{C_1^2}}\right)-1\right\}\leq C_2.$$

Assumption 2. $\left\|\Gamma_0\right\|_{SGL}^2+\left\|h_0\right\|_{\mathcal{H}_\mathcal{K}}^2>0$, *and the entropy of $\mathcal{B}$ with respect to the*

empirical metric P_n is bounded as follows:

$$H(\delta, \mathcal{B}, P_n) \leq C_3 \delta^{-2\psi},$$

where C_3 is some constant and $\psi \in (0, 1)$. See the Appendix for more details about this constant ψ.

Assumption 3. $\sup_{b \in \mathcal{B}} \|b\|_{P_n} \leq C_4$ *for some constant C_4.*

Theorem 5. *(Consistency) Under Assumptions 1-3 above, if the tuning parameters λ_1 and λ_2 satisfy*

$$\lambda_2^{-1} = n^{\frac{1}{1+\psi}} \left(\|h_0\|_{\mathcal{H}_\mathcal{K}}^2 + \|\Gamma_0\|_{SGL} \right)^{\frac{1-\psi}{1+\psi}} \text{ and } \lambda_1 = O_p(1)\lambda_2,$$

then we have

(i) $\left\| \hat{h} \circ \hat{\Gamma} - h_0 \circ \Gamma_0 \right\|_n = O_p(n^{-\frac{1}{2+2\psi}}) \left(\|h\|_{\mathcal{H}_\mathcal{K}}^2 + \|\Gamma\|_{SGL} \right)^{\frac{\psi}{1+\psi}}$, *and*

(ii) $\left\| \hat{h} \right\|_{\mathcal{H}_\mathcal{K}}^2 + \left\| \hat{\Gamma} \right\|_{SGL} = O_p(1) \left(\|h_0\|_{\mathcal{H}_\mathcal{K}}^2 + \|\Gamma_0\|_{SGL} \right)$.

Theorem 5 suggests that for the right λ_1 and λ_2 we can establish estimation consistency. Due to the potential identifiability issues we will explain in the next section, although the estimator $(\hat{h}, \hat{\Gamma})$ may not be unique, the sum of $\hat{h}$ and $\hat{\Gamma}$ is not too far away from the sum of the original h_0 and Γ_0 in terms of the norms or distances we defined above.

Corollary 6. *If the RKHS, $\mathcal{H}_\mathcal{K}$, contains functions that are differentiable, and $< \nabla h(\mathbf{z}), \nabla h(\mathbf{z}) >$ is uniformly bounded for all functions $h \in \mathcal{H}_\mathcal{K}$ and $\mathbf{z} \in \mathcal{R}^s$, then Assumption 2 holds when Theorem 5 is replaced by $H(\delta, \mathcal{H}_\mathcal{K}, P_n) \leq C_1 \delta^{-2\psi}$, for all $\delta \geq 0$.*

The proof of Theorem 5 and Corollary 6 are given in the Appendix. Often, when we are only interested in a subset of functions in the RKHS (e.g. functions less than

norm 1) we can substitute the full space $\mathcal{H}_{\mathcal{K}}$ in Corollary 6 with the subset of interest. Refer to [57] or [17] where both consider an RKHS (i.e. Sobolev space) with functions less than or equal to norm 1.

2.5 Identifiability

We introduce $\boldsymbol{\gamma}$ as a way of performing variable selection on our vector of FPC features. We wanted to illustrate this with some concrete examples and discuss identifiability issues with the estimator. There are two ways of looking at the estimation of the unknown functions h_0 and Γ_0. The first way is to view our feature vector, $\mathbf{z}$ as being related to the dependent variable y through the composite function $h \circ \Gamma$ as explained in Section 4. The second and equivalent way is to view our features as unknown. The true features are $\boldsymbol{\gamma} \circ \mathbf{z}$, where in this case the $\circ$ is used as the Hadamard product. We are given $\mathbf{z}$ and need to estimate the "true" features $\boldsymbol{\gamma} \circ \mathbf{z}$. In addition, we need to estimate the relationship between $\boldsymbol{\gamma} \circ \mathbf{z}$ and y, which is done through the function $h \in \mathcal{H}_{\mathcal{K}}$.

The first way of looking at the problem is to try and estimate the function $h_0 \circ \Gamma_0$. The function will belong to the RKHS $\mathcal{H}_{\mathcal{K} \circ \Gamma}$ where $\mathcal{K}$ is the kernel generating the RKHS that h belongs to. We are essentially looking at many different function spaces to find our estimator. The intersection between the function spaces do not have to be empty, which means our estimator does not have to be unique. We will now proceed to build this concept more formally. Let $\mathcal{K} : \mathcal{R}^s \times \mathcal{R}^s \mapsto \mathcal{R}$ be a positive definite function. Let $\Gamma : \mathcal{R}^s \mapsto \mathcal{R}^s$. We define $\mathcal{K} \circ \Gamma : \mathcal{R}^s \times \mathcal{R}^s \mapsto \mathcal{R}$ as the function given by $\mathcal{K} \circ \Gamma(\mathbf{s}, \mathbf{t}) = \mathcal{K}(\Gamma(\mathbf{s}), \Gamma(\mathbf{t}))$. This new function, $\mathcal{K} \circ \Gamma$ is positive definite. There is a relationship between the original RKHS, $\mathcal{H}_{\mathcal{K}}$ and the new RKHS, $\mathcal{H}_{\mathcal{K} \circ \Gamma}$. The result is that $\mathcal{H}_{\mathcal{K} \circ \Gamma} = \{h \circ \Gamma : h \in \mathcal{H}_{\mathcal{K}}\}$ and for any vector $u \in H_{K \circ \Gamma}$ we have that $\|u\|_{\mathcal{H}_{\mathcal{K} \circ \Gamma}} = inf\{\|h\|_{\mathcal{H}_{\mathcal{K}}} : u = h \circ \Gamma\}$. In general, $\mathcal{H}_{\mathcal{K} \circ \Gamma} \not\subset \mathcal{H}_{\mathcal{K}}$. In (2.2.2) we are taking the norm with respect to the original space $\mathcal{H}_{\mathcal{K}}$. Our iterative

procedure essentially allows us to view our problem the second way which is that the true features are unknown while our theoretical arguments view the problem the first way. Given the knowledge of the features (which translates to fixing a $\boldsymbol{\gamma}$), we are confined to just one RKHS, $\mathcal{H}_{\mathcal{K}}$. Lets take the linear kernel, $\mathcal{K}(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{x}_1^\top \mathbf{x}_2$ as an example. Suppose the truth is that y is related to a one dimensional feature z_0 through the following formulation: $y = h_0(z_0) + error$ where $h_0 \in \mathcal{H}_{\mathcal{K}_1}$, where $\mathcal{K}_1$ is the kernel that maps from $\mathcal{R} \times \mathcal{R} \mapsto \mathcal{R}$. So, if we knew the feature z_1, we would proceed to optimize (2.2.3) using the standard LSKM. However, suppose we have associated with each y a two dimensional vector $\mathbf{z} = (z_1, z_2)$. z_2 is just a "noisy" feature and unrelated to y. However, apriori we don't know that. So we assume the formulation is $y = h((z_1, z_2)) + error$ where $h \in \mathcal{H}_{\mathcal{K}}$, where now, $\mathcal{K}$ is the kernel that maps from $\mathcal{R}^2 \times \mathcal{R}^2 \mapsto \mathcal{R}$. We introduce our $\boldsymbol{\gamma}$ vector (γ_1, γ_2) and look at $y = h((\gamma_1 z_1, \gamma_2 z_2)) + error$. All functions, h in the space $\mathcal{H}_{\mathcal{K}}$ is of the form $h(\mathbf{z}) = \mathbf{x}^\top \mathbf{z}$ for some two dimensional vector $\mathbf{x} = (x_1, x_2)$. There is a one-to-one relationship between h and $\mathbf{x}$. The true function, h_0 has an associated real number c where $h_1(z_1) = cz_1$. We can recover $h_1 \in \mathcal{H}_{\mathcal{K}_1}$ from our estimation of h and γ if we set $\boldsymbol{\gamma} = (1, 0)$ and $\mathbf{x} = (c, \star)$ where "$\star$" is any real number. Equivalently, we can recover h_1 by looking at $\boldsymbol{\gamma} = (1, 1)$ where $\mathbf{x} = (c, 0)$. There are many functions that will recover the original function in the RKHS corresponding to the linear space kernel. Looking at our problem the first way, through function composition, we can estimate Γ_0 with the associated $\boldsymbol{\gamma}$ as the vector $(1, 0)$ or $(1, 1)$.

We can then see that in the intersection between $\mathcal{H}_{\mathcal{K} \circ \Gamma_1}$ and $\mathcal{H}_{\mathcal{K} \circ \Gamma_2}$ where Γ_1 has associated $\boldsymbol{\gamma_1} = (1, 0)$ and Γ_2 has associated $\boldsymbol{\gamma_2} = (1, 1)$ lies our estimate of h_1. In truth, for the linear space RKHS, there is no need to apply our method since $h_0 \in \mathcal{H}_{\mathcal{K}_1}$ can be estimated directly from the larger space $\mathcal{H}_{\mathcal{K}}$ where we set $h(\mathbf{z}) = \mathbf{x}^\top \mathbf{z}$ where $\mathbf{x} = (c, 0)$. We can never hope to have variable selection consistency nor can we hope to have identifiability of our estimator for these types of spaces. However, from a

goodness of fit standpoint, we are able to do just as good a job with many types of function compositions. Our hope is that we can glean some variable selection by penalizing the $\boldsymbol{\gamma}$ vector with the $\rho(\boldsymbol{\gamma}; \delta)$ term which, going back to the above scenario, should give preference to $\boldsymbol{\gamma} = (1, 0)$ over $\boldsymbol{\gamma} = (1, 1)$. For the RKHS associated with the Gaussian Kernel, the "larger dimensional space", a Gaussian Kernel mapping from higher dimensions, does not necessarily contain the functions from a "lower dimensional space", a Gaussian Kernel mapping from lower dimensions. However through the introduction of the $\boldsymbol{\gamma}$ transformation of the features, we can recover the equivalent functions of the "lower dimensional space".

2.6 Simulations

In this chapter we performed three simulation experiments to investigate the performance of our proposed procedure, including the performance of variable selection and its overall accuracy. For performance accuracy, we used both quasi-R^2 and adjusted quasi-R^2 defined as follows:

$$\begin{aligned} R^2_Q &:= 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y}_i)^2}, \\ R^2_{AQ} &:= 1 - \left(1 - R^2_Q\right) \left(\frac{n-1}{n-(k+1)}\right). \end{aligned}$$

The latter is a similar criterion used in the FAM paper [57], which was appealing for the comparison on the estimation sparsity. There is another performance of interest in addition to model accuracy. Performance in variable selection is summarized in terms of the stability measured by sensitivity and specificity for both functional and variable selections under these three simulation experiments. Specifically, we designed the following two simulation settings:

Scenario 1: A single functional predictor with sparsity in the FPC features;

Scenario 2: Multiple functional predictors with sparsity in the functional predictors and with sparsity in the FPC features.

Each of these scenarios would be handled using certain suitable penalty functions to address the designed sparsity; for example, in Scenario 3 we will use a two-level variable selection penalty (e.g sparse group lasso) to deal with two types of sparsity in the true model.

In all analyses, we used the Gaussian Kernel $\mathcal{K}(u,v) = \exp^{-\frac{1}{p}\|u-v\|^2}$ in our estimation where p was set as the number of features, which is equivalent to dividing the $\boldsymbol{\gamma}$ vector by $\sqrt{p}$. Typically, in LSKM this scaling parameter is either estimated or set to the number of features due to the consideration of the identifiability issue. See [20] for a theoretical argument to use the number of features for the bandwidth parameter, p, when using the Gaussian Kernel.

To run Steps 2-3 in our algorithm we used existing R packages; they are, the `EMMREML`,`KSPM` and `OEM` packages respectively available at:
https://cran.r-project.org/web/packages/oem/index.html,
https://cran.r-project.org/web/packages/KSPM/index.html, and
https://cran.r-project.org/web/packages/EMMREML/index.html.

Following the LSKM paper [33], due to the difficulty to graphically display the fitted value of the estimated function $h(\cdot)$ as a function of $\mathbf{z}$, we summarized the goodness of fit by regressing the true h on the estimated $\hat{h}$, with both being evaluated at the design points. From this concordance regression analysis, we may measure the goodness of fit on $\hat{h}$ through the average intercepts, slopes and R^2's obtained over the number of replications. Clearly, a high-quality fit is reflected by (i) the intercept is close to zero, (ii) the slope is close to one, and (iii) the R^2 is also close to one. In the meanwhile, we also graphically display the estimated function $\hat{h}$ by setting all variables equal to 0.5 except the one of interest, which is graphed over a grid of 100 equally spaced points form the interval $[0, 1]$. Such graphs provide

supplementary visualization of the estimation in addition to the table results derived from the concordance regression analyses.

In all three scenarios we generated 1000 IID functional paths of which 750 paths were assigned to the training set and 250 paths were assigned to the test set. It is the test set that we used to display the performance accuracy for. We used a one-dimensional fixed effect x_i to show the flexibility of our model in a semi-parametric setting, with $x_i \sim N(0,1)$. Following the LSKM paper [33], we chose similar true coefficients in the model with relatively strong signals.

Setting of Scenario 1: In this simple scenario, we simulated data from a model with a single functional predictor with sparsity in its FPC features. To do so, we generated a single functional predictor Z_i for each individual i by using the first 15 eigenbasis of the Fourier basis functions over the interval $[0,1]$: $Z(t) = \sum_{j=1}^{15} \varsigma_j \xi_j \phi_j(t)$. In other words, each functional predictor was created as a linear combination of the 15 basis functions, where $\phi_j(\cdot)$ is the j^{th} Fourier basis function, ς_j is the jth eigenvalue of Z, and ξ_j is the jth FPC feature.

There were 100 sampled points, t, equally spaced in the interval $[0,1]$ with very small deviations governed by the corresponding independent measurement errors drawn from $\nu \sim N(0, 0.001)$. Set $\varsigma_j = 45 \times 0.64^j$, and $\xi_j \sim N(0,1)$. As was done in [34], instead of directly using ξ_j, we used $\zeta_j = \Phi(\xi_j)$, where Φ is the CDF of the standard normal. This resulted in $\vec{\mathbf{z}} = (\zeta_1, \ldots, \zeta_{15})^\top$. We chose the second, ζ_2, and ninth, ζ_9, features as important features in the following true nonlinear non-additive model:

$$y_i = 2x_i + 20\ \cos(2\pi\zeta_{i2}) - 10\ \sin(2\pi\zeta_{i9}) + \zeta_{i2}\zeta_{i9} + \epsilon_i,$$

with $\epsilon_i \overset{iid}{\sim} N(0,1)$. FPCA was performed by the R package `PACE` available at https://cran.r-project.org/web/packages/fdapace/index.html [54]. This allowed us to extract the estimated FPC scores, $\hat{\xi}_j$, as well as the estimated eigenvalues, $\hat{\varsigma}_j$, which in turn enabled us to compute $\hat{\zeta}_j$.

In the first scenario, we used both LASSO and MCP penalty functions in our implementation, termed as $MFRS_{Lasso}$ and $MFRS_{MCP}$, respectively. We compared the results of our method with the standard linear approach with both LASSO and MCP under the assumption of linear functional relationships as well as the COSSO method for functional additive regression [57]. Functional additive regression via COSSO was performed by the R package `COSSO` available at https://cran.r-project.org/web/packages/cosso/index.html [54, 57]. Since the COSSO package is built for nonparametric regression (and not partial linear models) we regressed the residuals from the linear model with our fixed effect x_i on the extracted FPC scores.

In addition, we compared our method with an oracle LSKM estimator, called $LSKM^{oracle}$, that assumed the full knowledge of the true ζ's and two true signals, namely, ζ_2 and ζ_9. We also considered two oracle versions of our proposed algorithm, $MFRS^{oracle}_{Lasso}$ and $MFRS^{oracle}_{MCP}$, both of which used the true ζ's. This allows us to evaluate the performance of the FPCA procedure. This evaluation is important as our proposed procedure can be in principle used in simpler cases that do not involve functional covariates. This is because once we use FPCA to obtain our $\hat{\zeta}_i$ features we are in a standard regression setting with sparsity of covariates. In Scenario 1, due to the highly nonlinear relationships between the FPC features and the outcome, as expected the linear model performed poorly in terms of both model selection and model consistency. The results for Scenario 1 can be found in the second section of the supplemental materials. It is easy to see that our proposed method worked well. COSSO also did well in this Scenario in terms of model fit. COSSO tended to select noisy features more frequently then our proposed method. Simulation results for Scenario 1 based on the average of 100 simulations.

Table 2.1: Goodness of Fit for Scenario 1

			Reg of $\boldsymbol{h}$ on $\hat{\boldsymbol{h}}$		
Model	R^2_{AQ}	β	Intercept	Slope	R^2
$MFRS_{Lasso}$	0.948	2.00	0.006	1.00	0.953
$MFRS_{MCP}$	0.948	2.00	0.006	1.00	0.953
$MFRS^{oracle}_{Lasso}$	0.996	2.00	0.005	1.00	1.00
$MFRS^{oracle}_{MCP}$	0.996	2.00	0.005	1.00	1.00
$LSKM^{oracle}$	0.996	2.00	0.005	1.00	1.00
$COSSO$	0.946				
$Lasso$	0.101				
MCP	0.109				

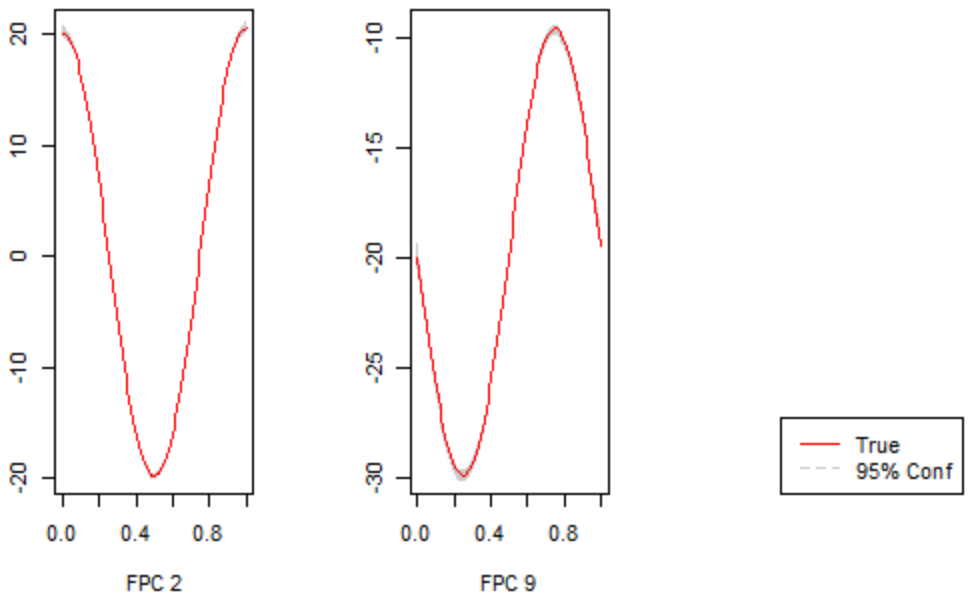
Table 2.2: Model Size for Scenario 1

	Model Size					
Model	**1**	**2**	**3**	**4**	**5**	**>5**
$MFRS_{Lasso}$	0	92	8	0	0	0
$MFRS_{MCP}$	0	95	4	1	0	0
$MFRS^{oracle}_{Lasso}$	0	99	1	0	0	0
$MFRS^{oracle}_{MCP}$	0	99	1	0	0	0
$COSSO$	0	63	23	11	2	1
$Lasso$	20	17	18	2	12	31
MCP	74	9	5	2	3	7

Table 2.3: FPC Selection for Scenario 1

	Selection Frequency														
Model	$\hat{\zeta}_1$	$\hat{\zeta}_2$	$\hat{\zeta}_3$	$\hat{\zeta}_4$	$\hat{\zeta}_5$	$\hat{\zeta}_6$	$\hat{\zeta}_7$	$\hat{\zeta}_8$	$\hat{\zeta}_9$	$\hat{\zeta}_{10}$	$\hat{\zeta}_{11}$	$\hat{\zeta}_{12}$	$\hat{\zeta}_{13}$	$\hat{\zeta}_{14}$	$\hat{\zeta}_{15}$
$MFRS_{Lasso}$	2	100	1	0	0	1	0	0	100	2	0	0	0	2	0
$MFRS_{MCP}$	2	100	1	1	0	1	0	0	100	2	0	0	0	2	0
$MFRS^{oracle}_{Lasso}$	1	100	0	0	0	0	0	0	100	0	0	0	0	0	0
$MFRS^{oracle}_{MCP}$	1	100	0	0	0	0	0	0	100	0	0	0	0	0	0
$COSSO$	6	100	6	3	1	2	9	15	100	8	2	2	0	1	0
$Lasso$	23	31	21	17	28	19	18	32	100	23	21	20	23	25	21
MCP	10	11	5	3	5	6	4	6	100	7	6	3	3	3	5

Estimated marginal plot with 95 percent shaded confidence bands of the function h evaluated at 100 grid points for each component while holding all other components equal to 0.5 in Scenario 1.



Setting of Scenario 2: In this scenario, the objective was to assess the performance of our method on both functional sparsity and within functional sparsity. Because of this complexity, we reported detailed numerical results in the main text of this paper. Here for each subject i, we generated 4 functional predictors $\{Z_i^1, \cdots, Z_i^4\}$ of the form: $Z^\ell(t) = \sum_{j=1}^{9} \sqrt{\varsigma_j}\xi_j\phi_j(t), \ell = 1, \ldots, 4$, where ϕ_j, ς_j, and ξ_j are set at the same values as those given in Scenario 1. It follows that $\vec{\mathbf{z}} = (\zeta_1^1, \ldots, \zeta_9^1, \ldots, \zeta_1^4, \ldots, \zeta_9^4)^\top$ where ζ_j^ℓ is the jth transformed feature for the ℓ'th functional covariates. To specify sparsity, we chose the first and second functional covariates, Z^1 and Z^2, by relating the transformed FPC features, $\{\zeta_1^1, \zeta_3^1, \zeta_4^1, \zeta_2^2, \zeta_7^2\}$; they are the first, third and fourth features from the first functional and the second and seventh from the second functional covariate, which will be related to the outcome in a non-linear and non-additive way:

$$
\begin{aligned}
y_i = & 2x_i + \zeta_{i1}^1 + \zeta_{i3}^1 + \zeta_{i4}^1 + \zeta_{i2}^2 + \zeta_{i7}^2 + \\
& 10\,\cos(2\pi\zeta_{i1}^1) - 10\left(\zeta_{i2}^2\right)^2 + 10\left(\zeta_{i7}^2\right)^2 - 10\left(\zeta_{i3}^1\right)^2 + \\
& 10\,\exp(-\zeta_{i3}^1)\zeta_{i4}^1 - 8\,\sin(2\pi\zeta_{i7}^2)\,\cos(2\pi\zeta_{i3}^1) + 20\zeta_{i1}^1\zeta_{i7}^2 + \epsilon_i
\end{aligned}
$$

where $\epsilon_i \overset{iid}{\sim} N(0,1)$ and ζ_{ij}^{ℓ} is the jth transformed score for the ℓth functional predictor for subject i. In this scenario, we set up both group sparsity (with only 2 of the 4 functional predictors being used) and within-group sparsity (with less then 9 of FPC features being used). In addition, we designed a non-additive structure in the true model across multiple functional covariates. We considered linear models, COSSO method for functional additive regression and oracle methods in the comparison. From Table 2.4 regarding the goodness of fit, we see that all of our MFRS estimators outperformed the standard linear estimators in terms of R_{AQ}^2 among all of our penalty functions and it outperformed COSSO for penalties that account for group sparsity. COSSO tended to perform on par for penalties that do not account for group sparsity (LASSO and MCP). It is evident that using a group sparsity penalty function (SGL, GLasso, and GMCP) clearly outperformed the methods that did not regularize grouping of covariates (Lasso and MCP). In addition, our estimators performed as well as the oracle LSKM estimator both in terms of R_{AQ}^2 and in terms of our estimate of h. The results also indicate that there were little differences between using a concave (MCP or GMCP) penalty function or using a convex (Lasso, GLasso or SGL) penalty function. In regard to the group sparsity, Table 2.5 indicates that the all methods had high sensitivity of detecting functional signals, while the proposed MFRS methods had better specificity than the linear models and the COSSO. Concerning the within group sparsity, it is interesting to note that a bigger difference is seen in terms of what type of penalty function is being used in model selection. Using

a general penalty (e.g. Lasso and MCP) that does not take the grouping structure into account tends to under-select specific members within a group as shown in tables 2.6 and 2.7. COSSO tended to perform well within group sparsity. Figure 2.1 shows that the MFRS method estimated the signal functions (Z^1 and Z^2) well.

Table 2.4: Goodness of Fit via the concordance regression for Scenario 2

			Reg of $\boldsymbol{h}$ on $\hat{\boldsymbol{h}}$		
Model	R^2_{AQ}	β	Intercept	Slope	R^2
$MFRS_{Lasso}$	0.830	2.00	-0.062	1.01	0.848
$MFRS_{GLasso}$	0.937	1.99	-0.055	1.01	0.972
$MFRS_{SGL}$	0.928	2.00	-0.051	1.01	0.955
$MFRS_{MCP}$	0.835	2.01	-0.062	1.01	0.856
$MFRS_{GMCP}$	0.935	1.99	-0.056	1.01	0.970
$LSKM^{oracle}$	0.911	1.99	-0.049	1.01	0.937
$COSSO$	0.832				
$Lasso$	0.453				
$GLasso$	0.324				
SGL	0.450				
MCP	0.513				
$GMCP$	0.307				

Table 2.5: Sensitivity and Specificity of Functional Selection for Scenario 2

	Selection Frequency			
Model	$\hat{Z}^1$	$\hat{Z}^2$	$\hat{Z}^3$	$\hat{Z}^4$
$MFRS_{Lasso}$	100	100	0	0
$MFRS_{GLasso}$	100	100	4	4
$MFRS_{SGL}$	100	100	0	0
$MFRS_{MCP}$	100	100	0	0
$MFRS_{GMCP}$	100	100	3	4
$COSSO$	100	100	5	6
$Lasso$	100	100	19	21
$GLasso$	94	99	7	8
SGL	100	100	19	18
MCP	100	100	20	19
$GMCP$	93	99	7	8

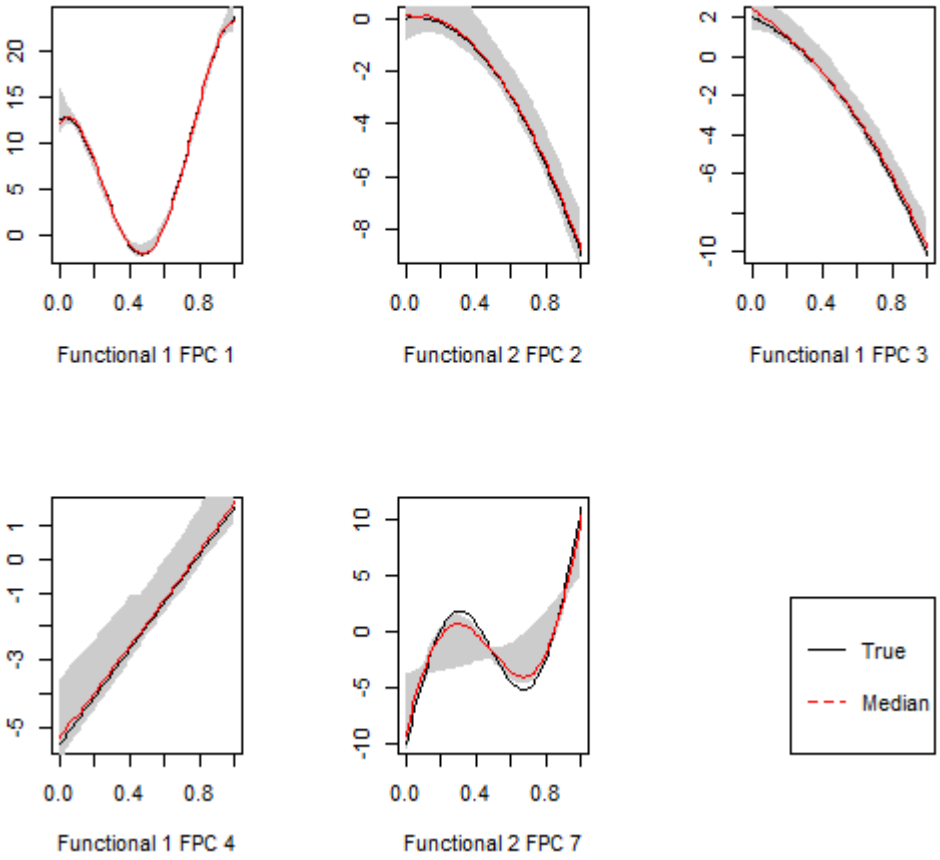
Table 2.6: FPC Selection for Scenario 2 Functional Z^1

	Selection Frequency								
Model	$\hat{\zeta}_1^1$	$\hat{\zeta}_2^1$	$\hat{\zeta}_3^1$	$\hat{\zeta}_4^1$	$\hat{\zeta}_5^1$	$\hat{\zeta}_6^1$	$\hat{\zeta}_7^1$	$\hat{\zeta}_8^1$	$\hat{\zeta}_9^1$
$MFRS_{Lasso}$	100	1	97	0	0	0	0	0	0
$MFRS_{GLasso}$	100	100	100	100	100	100	100	100	100
$MFRS_{SGL}$	100	21	100	71	26	20	17	16	15
$MFRS_{MCP}$	100	1	99	1	0	0	0	0	0
$COSSO$	100	2	100	93	1	0	0	1	0
$MFRS_{GMCP}$	100	100	100	100	100	100	100	100	100
$Lasso$	100	10	100	100	10	8	7	10	5
$GLasso$	94	94	94	94	94	94	94	94	94
SGL	100	12	100	100	10	8	8	11	5
MCP	100	10	100	100	9	8	9	7	5
$GMCP$	93	93	93	93	93	93	93	93	93

Table 2.7: FPC Selection for Scenario 2 Functional Z^2

	Selection Frequency								
Model	$\hat{\zeta}_1^2$	$\hat{\zeta}_2^2$	$\hat{\zeta}_3^2$	$\hat{\zeta}_4^2$	$\hat{\zeta}_5^2$	$\hat{\zeta}_6^2$	$\hat{\zeta}_7^2$	$\hat{\zeta}_8^2$	$\hat{\zeta}_9^2$
$MFRS_{Lasso}$	0	3	0	0	0	0	100	0	0
$MFRS_{GLasso}$	100	100	100	100	100	100	100	100	100
$MFRS_{SGL}$	16	100	14	7	16	23	100	15	7
$MFRS_{MCP}$	0	11	0	0	0	1	100	0	0
$MFRS_{GMCP}$	100	100	100	100	100	100	100	100	100
$COSSO$	8	97	5	5	5	15	100	3	3
$Lasso$	17	100	14	7	16	23	100	15	6
$GLasso$	99	99	99	99	99	99	99	99	99
SGL	17	100	14	7	16	23	100	15	7
MCP	17	100	13	6	16	23	100	15	8
$GMCP$	99	99	99	99	99	99	99	99	99

Figure 2.1: Estimated marginal functions with 95 percent shaded confidence bands of the function h evaluated at 100 grid points for each component while holding all other components equal to 0.5 in Scenario 2



2.7 Discussion

In this chapter we proposed a method to model the non-linear relationship between multiple functional predictors and a scalar outcome in the presence of other scalar confounders. We used the FPCA to decompose the functional predictors for feature extraction, and used the LSKM framework to model the functional relationship between the outcome and components. We developed a simultaneous procedure to select the important functional predictors and important features within selected functionals. We proposed a computationally efficient algorithm to implement the pro-

posed regularization method, which has been easily programmed in R with the utility of multiple existing R packages. It should be noted that although we focused on functional regression in this paper, the method proposed can be applied to non functional predictors that are related non-linearly and non-additively with an scalar outcome. In effect, by using functional principal components we essentially bypassed the infinite-dimensional problem and worked effectively in a non-functional framework with the FPC features. Through simulation and using data from the ELEMENT dataset, we demonstrated how the MFRS estimator outperformed existing methods both in terms of variable selection and model fit. It should be noted that the existing non-linear additive model, COSSO, did perform well in terms of variable selection as shown in the simulation.

As noted in the paper, there were identifiability limitations with regard to the bandwidth parameter and to the RKHS estimator. To overcome this issue, we have suggested fixing the bandwidth parameter; see the detailed discussion in the Identifiability section of this Chapter. We established key theoretical guarantees for our proposed estimator. In the case where there are multiple proposed estimators (and thus the identifiability issues arise), the established theoretical properties we established apply to any of those estimators.

Variable section on functional predictors presents many technical challenges, and there are many methodological problems remain unsolved. This Chapter demonstrated a possible framework to regularize estimation with bi-level sparsity of functional group sparsity and within-group sparsity. In the LSKM paper [33], it is briefly mentioned that if the relationship between the scalar outcome and p genetic pathways are additive, we can tweak the model to look like $y_i = \boldsymbol{x}_i^\top \beta + h_1(\boldsymbol{z}_i^1) + \cdots + h_p(\boldsymbol{z}_i^p) + \epsilon_i$ where each h_j belongs to its own RKHS. It is conceptually straightforward to extend our method and algorithm to handle this case, however, it is computationally expensive. We will explore this further in the next Chapter. For future re-

search, an extension to our framework would be to look at correlated data and model $y_{ij} = \boldsymbol{x}_i^\top \beta + h(\boldsymbol{z}_{ij}) + \boldsymbol{u}_{ij}^\top v_i + \epsilon_{ij}$ where $\boldsymbol{u}_{ij}^\top v_i$ are the random effects for subject i. Future research can extend the proposed paradigm to discrete outcome variables (e.g. binary) in the frameworks of generalized linear models and Cox regression models. We will detail how to extend the MFRS framework in the case where the outcome of interest is binary in Chapter 5.

CHAPTER III

Accelerometer Modeling Application

3.1 Introduction

In the introduction chapter to this dissertation, we discuss historical methods that have been used with tri-axis accelerometer data and review the terminology that we will use in this chapter. This chapter will focus on Functional Data Analysis modeling techniques for all three axes of accelerometer data and relate it to three health outcomes: BMI, weight and pulse pressure. The scientific goal was to assess see if there is a association between the physical activity data and these three health outcomes. The Centers for Disease Control and Prevention (CDC) https://www.cdc.gov/obesity/childhood/defining.html defines Body mass index (BMI) as a measure used to determine childhood obesity ($\geq$ 95th percentile for children and teens of the same age and sex). Subsequently, we control for age and sex in our models. To better define the amount of physical activity (PA) necessary to prevent overweight and obesity in children, studies try to find associations with physical activity (PA) and obesity [45, 35]. Elevated blood pressure (BP) during childhood and adolescence increases the risk of hypertension and cardiovascular disease in adulthood [50]. Physical activity is recommended for preventing and treating elevated BP and hypertension in children and adults [19, 50].

The aims of this chapter are as follows:

1. Demonstrate the benefits of using FDA; specifically, how utilizing the complete functional curves of the AC have higher explained variability of the outcomes when utilizing the fucnitonal predictors.

2. Provide methods on how we can use all three dimensions of the accelerometer data (which has not been done previously), and demonstrate its superior performance over the one-dimensional VM.

3. Analyze different ways to model 7 days of accelerometer data and show that viewing the entire curve which we will denote as a "7-day functional" can at times outperform averaging the 7 days of data into a 1-day function over 7 days we will denote as a "1-day averaged functional". These terms will be defined more clearly in the chapter.

4. Compare using the Activity Index (AI) vs the AC for the above aims.

3.2 ELEMENT Dataset

The study population includes adolescent participants from two enrolled cohorts of the Early Life Exposure in Mexico to Environmental Toxicants (ELEMENTS) study. A subset of 550 adolescent were recruited from the ELEMENTS study [28] beginning in 2015 to participate in a follow-up study. Among them, 539 adolescents aged 9-17 consented to wear an actigraph (ActiGraph GT3X+; ActiGraph LLC. Pensacola, FL), which was placed on their non-dominant wrist for five to seven days. The actigraph measured tri-axis accelerometer data sampled at 30Hz capturing three different directions of a person's movement. In addition to the accelerometer data, BMI, weight, blood pressure, sociodemographic, socioeconomic and nutritional variables were collected on the adolescents as well.

The actigraph did not register all of the adolescents for the full seven days and the software provides missing timestamps for the duration of wear. Our analysis only

included those cases where the actigraph recorded 85 percent or more of the full seven days. Other studies [44] have excluded days of accelerometer data with more than five percent missing. For the purpose of our study, we will focus on a 395-participant subset of participants who wore the actigraph for 85 percent of the seven days and had the three outcomes of interest (BMI, weight and pulse pressure) as well as the confounders of interest (age and sex). See the methods section in [26] for more details about this study and how the outcomes of interest were calculated.

Figure 3.1: TriAxis Activity Count for the 7 Days of accelerometer wear

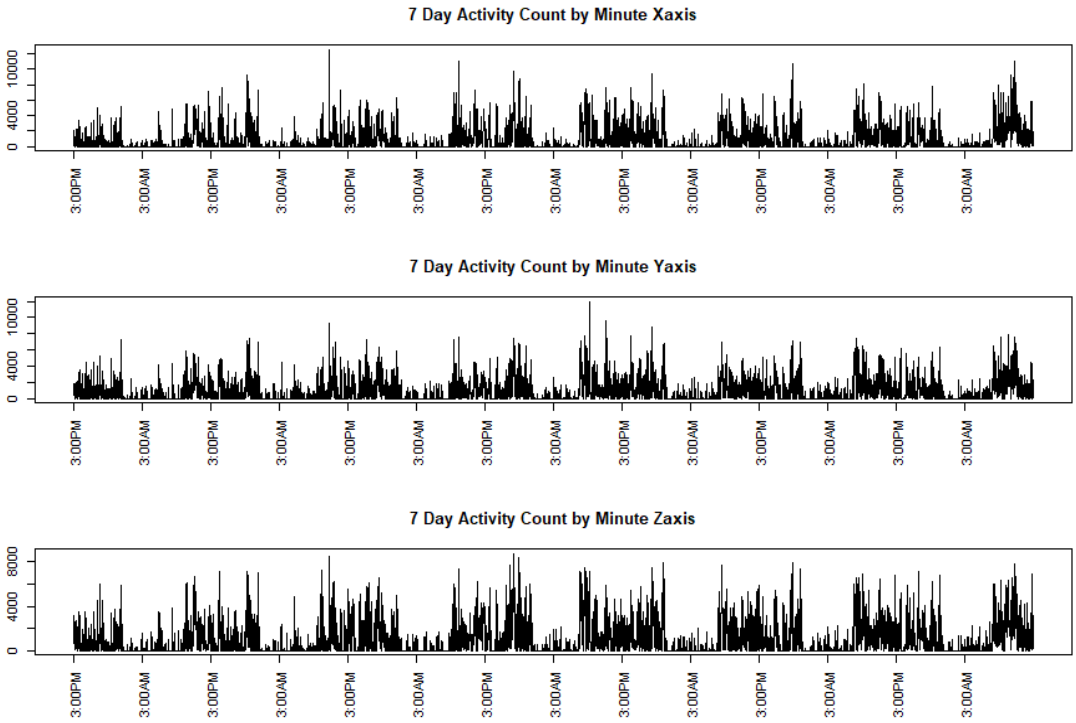


Figure 3.2: VM Activity Count for the 7 Days of accelerometer wear

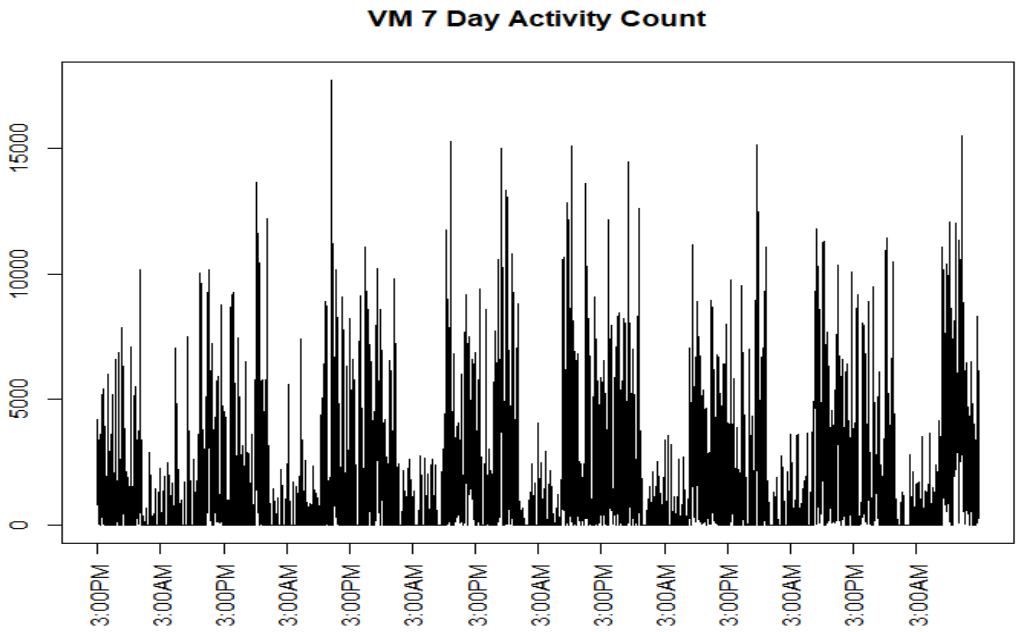


Figure 3.3: TriAxis Activity Count averaged minute-by-minute for the 7 Days of accelerometer wear

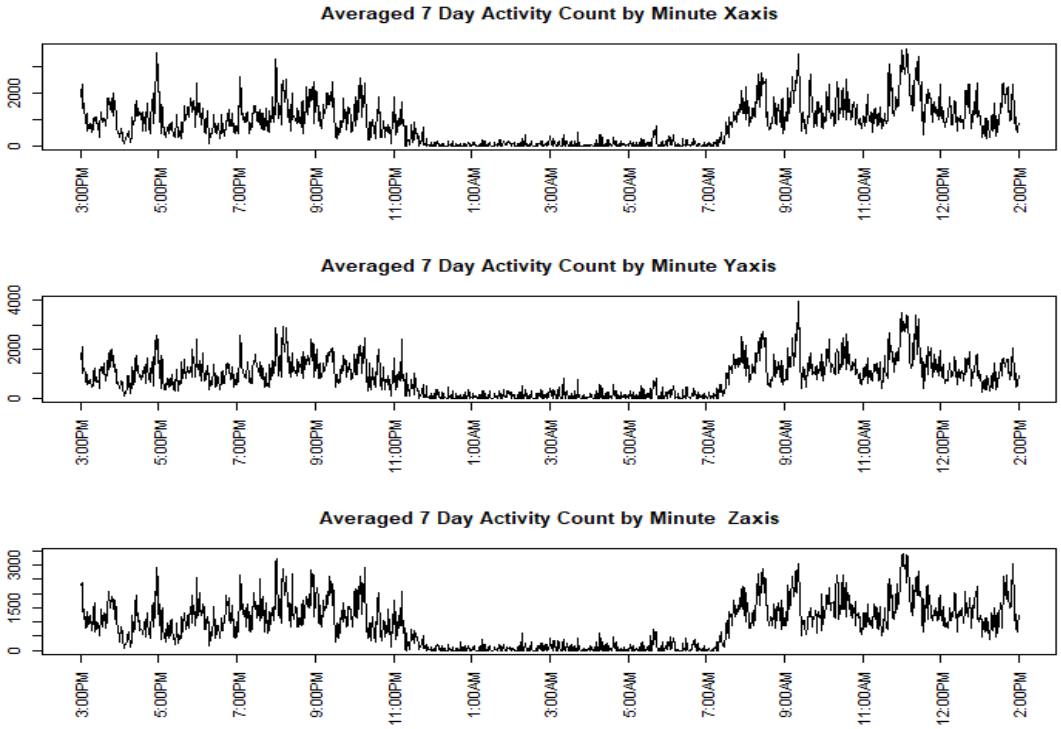
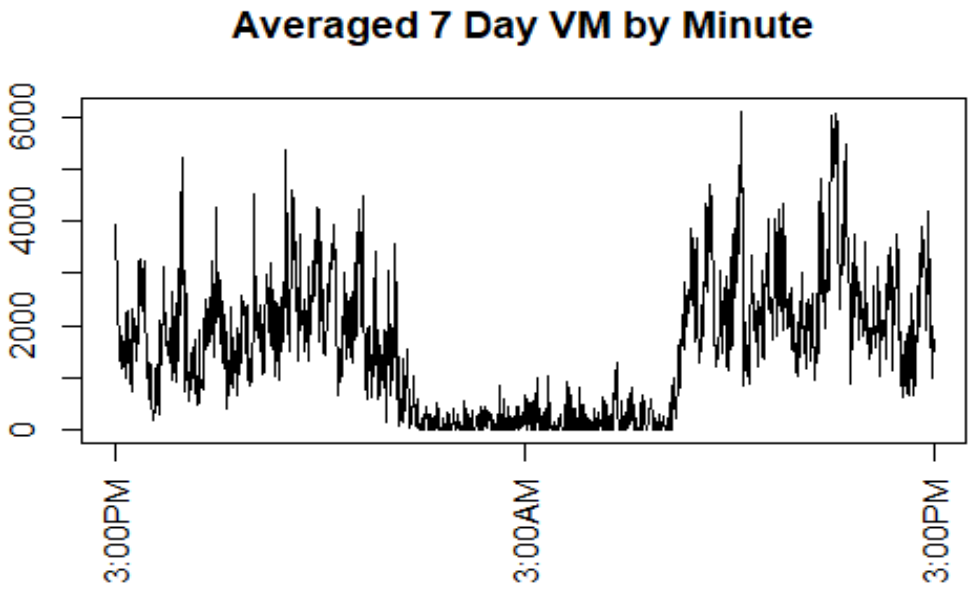


Figure 3.4: VM Activity Count averaged minute-by-minute for the 7 Days of accelerometer wear



3.3 Accelerometer Preprocessing

The 30 Hz raw accelerometer tri-axis data was summarized by the ActiGraph GT3X+ software to minute-by-minute activity counts for each axis over the 7 days of wear. In the literature there are multiple ways to handle multiple days of accelerometer data. For example, [44], used minute-by-minute medians across several days, while [3] included only the most active day (in terms of daily average AC) in the analysis. We can also treat the entire seven days as one long functional. In all three of the above ways, VM can be calculated from the three axes. Figure 3.1 shows an example of the curves from one subject when using all seven days. Figure 3.3 shows the curves when averaged across all seven days of data. The ActiGraph GT3X+ software provides a .GT3X file of raw data that can be read in R studio using the read.gt3x function which can be found at https://github.com/THLfi/read.gt3x.

The AI, was calculated from the raw acceleromter data file (.GT3X) using the

computeActivityIndex function found at https://rdrr.io/cran/ActivityIndex/man/computeActivityIndex.html. Part of the AI calculation requires the systematic noise variance $\bar{\sigma}_i^2$, which is supposed to be estimated from when the device is not moving (at rest). However, we did not have values from when the device was at rest (not moving) so we used the missing time points that were displayed in the .GT3X file which essentially puts the systematic noise variance at 0. These time points for the missingness were fed into the computeActivityIndex function. See the Introduction chapter for a more detailed description of the AI formula.

Figure 3.5: AI for the 7 Days of accelerometer wear

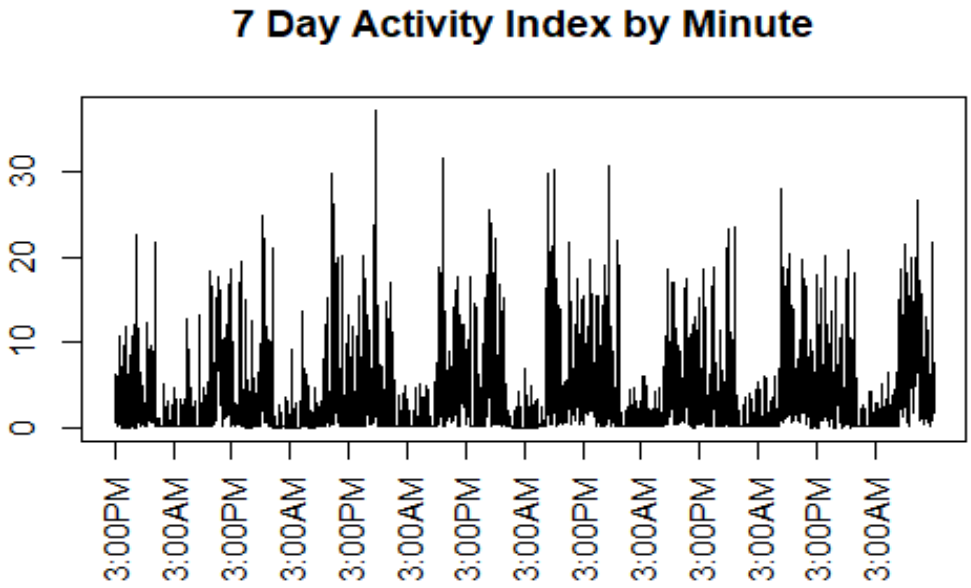
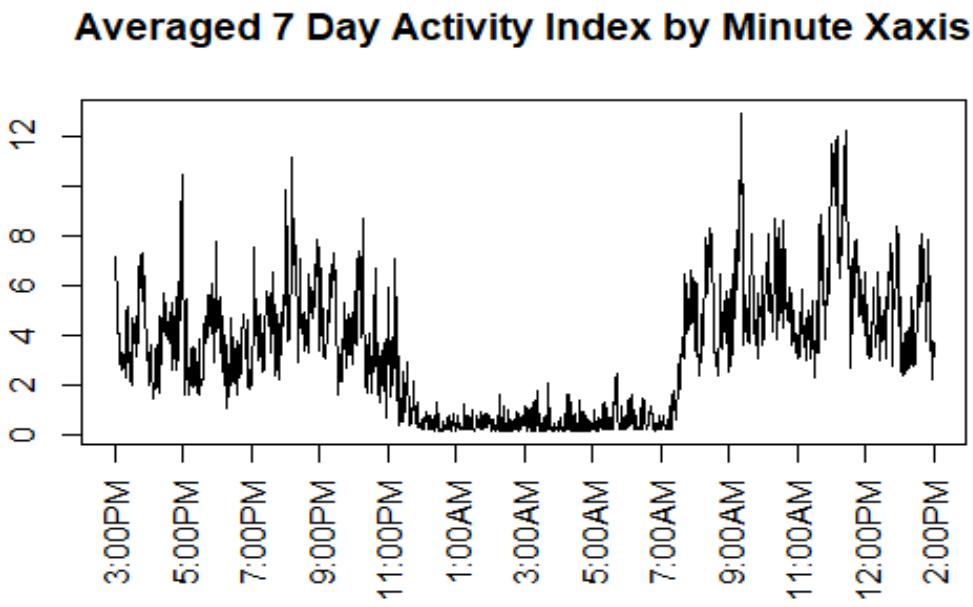


Figure 3.6: AI averaged minute-by-minute for the 7 Days of accelerometer wear



We preprocessed the accelerometer data to create six versions of functional variables which will be used in our subsequent analyses:

1. Time domain (as opposed to using the frequency domain) for the AC of the three axes were used covering the entire seven days of wear.

2. VM was calculated from the above which was the square root of the sum of squares of the three axes activity counts minute-by-minute.

3. Time domain for the AC that was averaged minute-by-minute was covering the entire seven days of wear. For example, since all the participants started wearing the actigraph at 3pm, the first data point for each individual is an average of 7 days of AC at 3pm.

4. VM was calculated from the above.

5. AI was calculated for the 7-day functional.

6. AI was calculated for the 1-day averaged functional

3.4 Review of Statistical Methods

3.4.1 FPCA

In this section, we focus on the utility of functional principal component analysis (FPCA) to perform decomposition of the three functionals from the tri-axis data described in the previous section. Let $Z(t)$ denote the functional we wish to decompose. For example, this might be the first axis activity counts. As explained in the previous chapter, by the Karhunen-Loève expansion we can write $Z(t) = \sum_{k=1}^{\infty} \sqrt{\varsigma_k}\xi_k\phi_k(t)$, where $\varsigma_k > 0$ are the eigenvalues, and loadings $\xi_k := \frac{1}{\sqrt{\varsigma_k}} < Z, \phi_k >$ satisfies mean zero, $E(\xi_k) = 0$, and variance one, $E(\xi_k\xi_j) = 1$ for $k = j$, and uncorrelated, $E(\xi_k\xi_j) = 0$ for $k \neq j$. We will use the set of estimates $\{\hat{\xi}_j\}$ as features for the accelerometer functional, $Z(t)$. In the cases of all three axes being used we obtain three sets of features, one from each axis respectively.

3.4.2 Least Squares Kernel Machine

We extended the LSKM proposed by Liu, Lin and Ghosh (2007) [33]. To review, the model they proposed was $y_i = \mathbf{x}_i^\top\boldsymbol{\beta} + h(\mathbf{z}_i) + \epsilon_i$, where they used LSKM to analyze a multidimensional predictor $\mathbf{z}_i$. In their model, parameter $\boldsymbol{\beta}$ needs to be estimated for some $\mathbf{x}$ vector of clinical covariates and $\mathbf{z}$ is a p-element vector of gene expressions that is potentially related to the outcome y via a non-parametric function h. The p-variate function $h(\cdot)$ is assumed to lie in a reproducing kernel Hilbert space (RKHS), $\mathcal{H}_\mathcal{K}$, generated by a positive definite kernel function $\mathcal{K}(\cdot,\cdot)$. They briefly mention an extension of that model to modeling multiple genetic pathway effects where one could consider a semiparametric additive model

$$y = \mathbf{X}^\top\boldsymbol{\beta} + h_1(\mathbf{z}^1) + \cdots + h_1(\mathbf{z}^m) + \epsilon, \tag{3.4.1}$$

where $\mathbf{z}^\ell$ ($\ell = 1, \ldots, m$) denotes a $p_\ell \times 1$ vector of genes in the ℓth pathway for m number of pathways and $h_\ell(\cdot) \in \mathcal{H}_{\mathcal{K}_\ell}$ denotes a non-parametric function.

We apply the numerical recipe from the linear mixed-effects model (LMM) to solve (3.4.1). In our case, $m = 3$ corresponds to the three axes from the accelerometer, with, $\mathbf{z}^\ell$ corresponds to the FPC scores extracted for the ℓth axes.

Note that $\boldsymbol{\beta}$ and h_ℓ may be estimated by maximizing the scaled penalized likelihood function:

$$\begin{aligned} J(h_1, h_2, h_3, \boldsymbol{\beta}) = &- \frac{1}{2}\sum_{i=1}^{n}\{y_i - \mathbf{x}_i^\top\boldsymbol{\beta} - h_1(\mathbf{z}_i^1) - h_2(\mathbf{z}_i^2) - h_3(\mathbf{z}_i^3)\}^2 \\ &- \frac{1}{2}\lambda_1 \left\|h_1\right\|^2_{\mathcal{H}_{\mathcal{K}_1}} - \frac{1}{2}\lambda_2 \left\|h_2\right\|^2_{\mathcal{H}_{\mathcal{K}_2}} - \frac{1}{2}\lambda_3 \left\|h_3\right\|^2_{\mathcal{H}_{\mathcal{K}_3}}, \end{aligned} \tag{3.4.2}$$

where for each $\ell = 1, 2, 3, \lambda_\ell > 0$ is the tuning parameter and $\|\cdot\|_{\mathcal{H}_{\mathcal{K}_\ell}}$ is the norm of the RKHS generated by the kernel $\mathcal{K}_\ell$. Following a similar procedure described in paper for a single genetic pathway, we can show by the representer theorem (Kimeldorf and Wahba 1970) that the solution to (3.4.2) for the nonparametric function h_ℓ can be expressed as $h_\ell(\cdot) = \sum_{i=1}^n \alpha_i^\ell \mathcal{K}_\ell(\cdot, \mathbf{z}_i^\ell)$ where $\alpha_i^\ell \in \mathcal{R}$. Then an equivalent optimization problem is to maximize the following objective function with respect to $(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \boldsymbol{\alpha}_3, \boldsymbol{\beta})$:

$$\begin{aligned} &J(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \boldsymbol{\alpha}_3, \boldsymbol{\beta}) = \\ &-\frac{1}{2}\sum_{i=1}^{n}\{y_i - \mathbf{x}_i^\top\boldsymbol{\beta} - \sum_{j=1}^{n}\alpha_j^1\mathcal{K}_1(\mathbf{z}_i^1, \mathbf{z}_j^1) - \sum_{j=1}^{n}\alpha_j^2\mathcal{K}_2(\mathbf{z}_i^2, \mathbf{z}_j^2) - \sum_{j=1}^{n}\alpha_j^3\mathcal{K}_3(\mathbf{z}_i^3, \mathbf{z}_j^3)\}^2 \\ &\qquad - \frac{1}{2}\lambda_1\boldsymbol{\alpha}_1^\top\mathbf{K}_1\boldsymbol{\alpha}_1 - \frac{1}{2}\lambda_2\boldsymbol{\alpha}_2^\top\mathbf{K}_2\boldsymbol{\alpha}_2 - \frac{1}{2}\lambda_3\boldsymbol{\alpha}_3^\top\mathbf{K}_3\boldsymbol{\alpha}_3, \end{aligned} \tag{3.4.3}$$

where $\mathbf{K}_\ell$ is an $n \times n$ matrix with ijth entry equal to $\mathcal{K}_\ell(\mathbf{z}_j^\ell, \mathbf{z}_i^\ell)$. The resulting

equations for estimating (3.4.3) are:

$$\begin{aligned}\frac{\partial J(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \boldsymbol{\alpha}_3, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= 0 \\ &\Longrightarrow \mathbf{X}^\top \mathbf{Y} - \mathbf{X}^\top \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}^\top \hat{\mathbf{h}}_1 - \mathbf{X}^\top \hat{\mathbf{h}}_2 - \mathbf{X}^\top \hat{\mathbf{h}}_3 = 0\end{aligned} \tag{3.4.4}$$

$$\begin{aligned}\frac{\partial J(\boldsymbol{\alpha}^1, \boldsymbol{\alpha}^2, \boldsymbol{\alpha}^3, \boldsymbol{\beta})}{\partial \boldsymbol{\alpha}_j} &= 0 \\ &\Longrightarrow \mathbf{K}_j^\top \mathbf{Y} - \mathbf{K}_j^\top \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{K}_j^\top \hat{\mathbf{h}}_1 - \mathbf{K}_j^\top \hat{\mathbf{h}}_2 - \mathbf{K}_j^\top \hat{\mathbf{h}}_3 - \lambda_j \mathbf{K}_\ell^\top \boldsymbol{\alpha}_j = 0 \\ &\Longrightarrow \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} - \hat{\mathbf{h}}_1 - \hat{\mathbf{h}}_2 - \hat{\mathbf{h}}_3 - \lambda_j \mathbf{K}_\ell^{-1} \hat{\mathbf{h}}_\ell = 0\end{aligned} \tag{3.4.5}$$

where $\hat{\mathbf{h}}_\ell$ is a vector with the ith element corresponding to the estimated function $\hat{h}_j(\mathbf{z}_i^j)$. In other words, $\hat{\mathbf{h}}_j = \mathbf{K}_j \hat{\boldsymbol{\alpha}}_j$. Let $\mathbf{R} = \mathbf{I}\sigma^2$ for some constant $\sigma^2 \geq 0$ (so its a diagonal matrix with σ^2 on its diagonals). Rewriting the above estimating equations in matrix notation, we obtain

$$\begin{bmatrix} \mathbf{X}^\top \mathbf{R}^{-1} \mathbf{X} & \mathbf{X}^\top \mathbf{R}^{-1} & \mathbf{X}^\top \mathbf{R}^{-1} & \mathbf{X}^\top \mathbf{R}^{-1} \\ \mathbf{R}^{-1}\mathbf{X} & \mathbf{R}^{-1} + (\tau_1 \mathbf{K})^{-1} & \mathbf{R}^{-1} & \mathbf{R}^{-1} \\ \mathbf{R}^{-1}\mathbf{X} & \mathbf{R}^{-1} & \mathbf{R}^{-1} + (\tau_2 \mathbf{K})^{-1} & \mathbf{R}^{-1} \\ \mathbf{R}^{-1}\mathbf{X} & \mathbf{R}^{-1} & \mathbf{R}^{-1} & \mathbf{R}^{-1} + (\tau_3 \mathbf{K})^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{h}}_1 \\ \hat{\boldsymbol{h}}_2 \\ \hat{\boldsymbol{h}}_3 \end{bmatrix} = \begin{bmatrix} \mathbf{R}^{-1}\mathbf{X}^\top \mathbf{y} \\ \mathbf{R}^{-1}\mathbf{y} \\ \mathbf{R}^{-1}\mathbf{y} \\ \mathbf{R}^{-1}\mathbf{y} \end{bmatrix}.$$

The above formulation is mathematically equivalent to solving the normal equations from the linear mixed-effects model (LMM) of the form:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{h}_1 + \mathbf{h}_2 + \mathbf{h}_3 + \boldsymbol{\epsilon}, \tag{3.4.6}$$

where $\mathbf{h}_j$ is an $n \times 1$ vector of random effects with distribution $N(\mathbf{0}, \tau_\ell \mathbf{K}_\ell)$, n-dimensional vector error terms $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$,and $\tau_\ell = \lambda_\ell^{-1}\sigma^2 > 0$. To see the connection, calculating the best linear unbiased estimators (BLUPS) for the random effects $(\hat{h}_1, \hat{h}_2, \hat{h}_3)$ is accomplished through the log likelihood function:

$$\ell(\boldsymbol{\beta}, \mathbf{h}_1, \mathbf{h}_2, \mathbf{h}_3) = \ell(\mathbf{Y} \mid \mathbf{h}_1, \mathbf{h}_2, \mathbf{h}_3) + \ell(\mathbf{h}_1) + \ell(\mathbf{h}_2) + \ell(\mathbf{h}_3) \tag{3.4.7}$$

since we are assuming that $\mathbf{h}_1 \perp \mathbf{h}_2 \perp \mathbf{h}_3$. We have $\ell(\mathbf{Y} \mid \mathbf{h}_1, \mathbf{h}_2, \mathbf{h}_3) \propto -\frac{1}{2\sigma^2} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{h}_1 - \mathbf{h}_2 - \mathbf{h}_3\|_2^2$ and $\ell(h_j) \propto \frac{-\mathbf{h}_j^\top \mathbf{K}_j^{-1} \mathbf{h}_j}{2\tau_j}$ the score equations for (3.4.7) are:

$$\begin{aligned} &\frac{\ell(\boldsymbol{\beta}, \mathbf{h}_1, \mathbf{h}_2, \mathbf{h}_3)}{\partial \boldsymbol{\beta}} = 0 \\ &\qquad \Longrightarrow \frac{1}{\sigma^2}(\mathbf{X}^\top \mathbf{Y} - \mathbf{X}^\top \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}^\top \hat{\mathbf{h}}_1 - \mathbf{X}^\top \hat{\mathbf{h}}_2 - \mathbf{X}^\top \hat{\mathbf{h}}_3) = 0 \end{aligned} \tag{3.4.8}$$

$$\begin{aligned} &\frac{\ell(\boldsymbol{\beta}, \mathbf{h}_1, \mathbf{h}_2, \mathbf{h}_3)}{\partial \mathbf{h}_j} = 0 \\ &\qquad \Longrightarrow \frac{1}{\sigma^2}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} - \hat{\mathbf{h}}_1 - \hat{\mathbf{h}}_2 - \hat{\mathbf{h}}_3) - \frac{\mathbf{K}^{-1}\hat{h}_j}{\tau_j} = 0 \end{aligned} \tag{3.4.9}$$

which is the same equations as (3.4.5) (setting $\lambda_j = \frac{\sigma^2}{\tau_j}$).

As we discussed previously in Chapter 2, although there is a closed form solution for fixed $\lambda_1, \lambda_2, \lambda_3$ in maximizing (3.4.2), through the numerical procedure of LMM we can get estimate λ_j easily as part of the estimation of the variance components with the REML equations. In this case, the LSKM regression model allows us to consider a non-linear relationship for each axis of the accelerometer by considering a non-linear and non-additive relationship with the FPC corresponding to that axis. Furthermore, the above framework allows us to simultaneously model the multiple axes from the accelerometer data in an additive way. If we don't want to assume an additive relationship between the health outcome and the 3 axes, we can always stack the vectors from the three FPC axes into one vector and use the standard LSKM framework with one function, h. The above extension may not make sense for the accelerometer data, since movement in one direction will logically be correlated with movement in another directions. However, we will compare the two ways of modeling the FPC features from the axes's with our dataset. Specifically, we will compare using

the additive LSKM (3.4.1) for modeling the three axis of the accelerometer data with the original LSKM model (2.1.2).

The score test developed by Liu, Lin and Ghosh [33] on the feature vector $\mathbf{z}$ can be extended to the additive LSKM model as well to test the overall effect of $\mathbf{z}^j$. Thus, you can use the score tests for functional variable selection. This is similar to using the group lasso as a penalty in the MFRS framework. However, in the non-additive model where the vectors of FPC scores for all the functionals are stacked, they would resort to the AIC or BIC method they proposed to perform functional variable selection. Furthermore, since there are many FPC scores associated with each functional, in both the original LSKM or additive LSKM you would have to resort to using AIC or BIC method to determine within a functional which scores to select. It is important to stress that the LSKM framework was not designed with the intention of handling functional (and certainty multiple functional) predictors. The MFRS framework will be used to compare with the additive model. To implement the additive LSKM framework we used an R package called KSPM found: https://cran.r-project.org/web/packages/KSPM/index.html that performs the model fit as well as the score tests. However, this package uses cross validation for the three λ_j's (for the three axis) and does not use the solution to the REML equations. Computationally, this takes several hours to run. If in addition you estimate the bandwidth parameter on each kernel running the model takes several days. We fixed the bandwidth parameter to the number of FPC scores to save time. As in the previous Chapter, this approach is justified in [20]. Furthermore, in that paper, they argue that unlike in the standard kernel regression of kernel density estimation, only the smoothing parameter is the kernel bandwidth; however, in the case where there is a smoothing term, λ, the role of the kernel bandwidth is fundamentally different. It acts as a way of determining how vectors that are different need to be in the standardized before calculating how "close" or similar those vectors are. They further show that a

range of values for the kernel bandwidth tends to perform in a similar fashion after optimizing over λ.

3.4.3 MFRS for additive LSKM

We will provide the details as to how one could extend the MFRS framework to the additive LSKM. However, as we will explain below, we do not implement this in the analysis for the acceleromter data. The importance of this subsection will be discussed further in the concluding chapter of this dissertation. Following the same notation as in the previous subsections and chapters we wish to model

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + (h_{01} \circ \Gamma_{01})(\mathbf{z}_i^1) + (h_{02} \circ \Gamma_{02})(\mathbf{z}_i^2) + (h_{03} \circ \Gamma_{03})(\mathbf{z}_i^3) + \epsilon_i, i = 1, \ldots, n. \quad (3.4.10)$$

and estimate the unknown functions $h_{01} \circ \Gamma_{01}, h_{02} \circ \Gamma_{02}, h_{03} \circ \Gamma_{03}$, and unknown vector, $\boldsymbol{\beta}$. We accomplish this by minimizing:

$$\min_{\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \boldsymbol{\alpha}_3, \boldsymbol{\beta}, \boldsymbol{\gamma}} J_3(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \boldsymbol{\alpha}_3, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \frac{1}{2n} \sum_{i=1}^{n} \left\{ y_i - \mathbf{x}_i^\top \boldsymbol{\beta} - \sum_{j=1}^{3} \sum_{k=1}^{n} \alpha_k^j \mathcal{K}_j(\boldsymbol{\gamma}_j \circ \mathbf{z}_i^j, \boldsymbol{\gamma}_j \circ \mathbf{z}_k^j) \right\}^2$$
$$+ \frac{1}{2} \lambda_1 \boldsymbol{\alpha}_1^\top \mathbf{K}(\boldsymbol{\gamma}_1; Z^1)_1 \boldsymbol{\alpha}_1 + \frac{1}{2} \lambda_2 \boldsymbol{\alpha}_2^\top \mathbf{K}(\boldsymbol{\gamma}_2; Z^2)_2 \boldsymbol{\alpha}_2 + \frac{1}{2} \lambda_3 \boldsymbol{\alpha}_3^\top \mathbf{K}(\boldsymbol{\gamma}_1; Z^3)_3 \boldsymbol{\alpha}_3 + \lambda_4 \rho(\boldsymbol{\gamma}; \delta), \quad (3.4.11)$$

where there is an association between functions Γ_{0j} and γ_j as explained in chapter 2.

We proceed in a similar fashion as described in chapter 2 (Algorithm 1 and Algorithm 2). Given $\hat{\boldsymbol{\gamma}}$, we estimate $(\hat{\boldsymbol{\alpha}}_1, \hat{\boldsymbol{\alpha}}_2, \hat{\boldsymbol{\alpha}}_3, \hat{\boldsymbol{\beta}}, \lambda_1, \lambda_2, \lambda_3)$ by solving the additive LSKM problem given in (3.4.1). Given, $(\hat{\boldsymbol{\alpha}}_1, \hat{\boldsymbol{\alpha}}_2, \hat{\boldsymbol{\alpha}}_3, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}_2, \hat{\boldsymbol{\gamma}}_3)$ we can estimate $\hat{\boldsymbol{\gamma}}_1$ using the proximal Gauss-Newton algorithm (Step 2.2) of the algorithm mentioned in chapter 2. We can then use the updated $\hat{\boldsymbol{\gamma}}_1$ and the old $\hat{\boldsymbol{\gamma}}_3$ to find $\hat{\boldsymbol{\gamma}}_2$. In essence we are just using Step 2.2 in the algorithm but instead of solving for the updated $\hat{\boldsymbol{\gamma}}$ directly we cycle through each groups. Unlike in the previous chapter where we we used the proximal Gauss-Newton method in (2.3.4) to estimate (2.3.1), we do not

have such an equivalent form for (3.4.11) to minimize over the entire $\boldsymbol{\gamma}$ for a fixed $(\hat{\boldsymbol{\alpha}}_1, \hat{\boldsymbol{\alpha}}_2, \hat{\boldsymbol{\alpha}}_3, \hat{\boldsymbol{\beta}}, \lambda_1, \lambda_2, \lambda_3)$. Another possible approach for solving for $\boldsymbol{\gamma}$ for a fixed $(\hat{\boldsymbol{\alpha}}_1, \hat{\boldsymbol{\alpha}}_2, \hat{\boldsymbol{\alpha}}_3, \hat{\boldsymbol{\beta}}, \lambda_1, \lambda_2, \lambda_3)$ is described in chapter 5 using an inexact linear search with backtracking. For the following three reasons we do not implement the MFRS in the additive setting for the accelerometer data:

1. It makes more sense that movement in one direction is informative about movement in another direction and therefore, a non-linear additive model for the three axis would not be as accurate as a non-linear non-additive model.

2. The main objective of the MFRS framework is to perform functional variable selection in the LSKM setting since no other method besides AIC/ BIC existed. However, as pointed out for the additive LSKM setting, the score tests provide a way of performing functional variable selection (although it can not differentiate the FPC scores within groups).

3. Performing the additional cycling through groups for Step 2.2 in the algorithm can be computationally challenging.

There are settings where the MFRS algorithm can make sense in the LSKM framework as will be pointed out in the concluding chapter where the first two points above will not ne an issue. However, the third issue raised regarding the computational costs remains an issue.

3.5 Proposed Statistical Models

In this section we list the models that we use to analyze accelerometer data based on the previous sections. We use these models for each of the 6 settings we listed in the ELEMENT dataset section. Let $\mathbf{z}_i^\ell = (\xi_1^\ell, \cdots, \xi_{s_\ell}^\ell)_i^\top$ be the vector of FPC features from the ith observation of the functional covariate Z^ℓ. In total we will

have $\ell \in \{1,2,3\}$ where $1,2,3$ correspond to the first, second and third axis of the accelerometer functional. Let $\mathbf{z}_i = [(\mathbf{z}_i^1)^\top, (\mathbf{z}_i^2)^\top, (\mathbf{z}_i^3)^\top]^\top$ be the grand vector of FPC features from all 3 functional covariates and $s = \sum_{\ell=1}^{3} s_\ell$ number of FPC features, and $\mathbf{z}_i \in R^s$. Let $\mathbf{z}_i^{VM}$ be the vector of FPC features form the ith observation of the vector magnitude curve.

First, consider the following models we will reference as M0-M5:

1. M0: Linear model with only baseline covariates and with no predictors of physical activities:

$$E(BMI_i|Age_i, Sex_i) = \beta_0 + \beta_1 Age_i + \beta_2 Sex_i$$

2. M1: Linear model with fixed and functional features

$$E(BMI_i|Age_i, Sex_i, \mathbf{z}_i) = \beta_0 + \beta_1 Age_i + \beta_2 Sex_i + \sum_{j=1}^{3}\sum_{k=1}^{s_k} \beta_j^k \xi_{ij}^k.$$

3. M2: LSKM non-additive

$$E(BMI_i|Age_i, Sex_i, \mathbf{z}_i) = \beta_0 + \beta_1 Age_i + \beta_2 Sex_i + h(\mathbf{z}_i).$$

4. M3: LSKM additive

$$E(BMI_i|Age_i, Sex_i, \mathbf{z}_i) = \beta_0 + \beta_1 Age_i + \beta_2 Sex_i + h_1(\mathbf{z}_i^1) + h_2(\mathbf{z}_i^2) + h_3(\mathbf{z}_i^3).$$

5. M4: FAM+COSSO

$$E(BMI_i|Age_i, Sex_i, \mathbf{z}_i) = \beta_0 + \beta_1 Age_i + \beta_2 Sex_i + \sum_{k=1}^{s} f_k((\mathbf{z}_i)_k).$$

6. M5: the MFRS+SGL

$$E(BMI_i|Age_i, Sex_i, \mathbf{z}_i) = \beta_0 + \beta_1 Age_i + \beta_2 Sex_i + h(\boldsymbol{\gamma} \circ \mathbf{z}_i).$$

Where ξ_{ij}^k is the ith persons kth FPC score for functional j and $(\mathbf{z}_i)_k$ is the kth index of the vector $\mathbf{z}_i$. When modeling with the VM we used the following models:

1. VM1: Linear model with fixed and functional features

$$E(BMI_i|Age_i, Sex_i, \mathbf{z}_i^{VM}) = \beta_0 + \beta_1 Age_i + \beta_2 Sex_i + \sum_{k=1}^{s} \beta^k \xi_i^{VM,k}.$$

2. VM2: LSKM non-additive

$$E(BMI_i|Age_i, Sex_i, \mathbf{z}_i) = \beta_0 + \beta_1 Age_i + \beta_2 Sex_i + h(\mathbf{z}_i^{VM}).$$

3. VM4: FAM+COSSO estimator

$$E(BMI_i|Age_i, Sex_i, \mathbf{z}_i) = \beta_0 + \beta_1 Age_i + \beta_2 Sex_i + \sum_{k=1}^{s} f_k((\mathbf{z}_i^{VM})_k).$$

4. VM5: the MFRS+Lasso estimator

$$E(BMI_i|Age_i, Sex_i, \mathbf{z}_i) = \beta_0 + \beta_1 Age_i + \beta_2 Sex_i + h(\boldsymbol{\gamma} \circ \mathbf{z}_i^{VM}).$$

Where $\xi_i^{VM,k}$ is the kth FPC feature extracted for the VM curve for person i. For the AI, we simply substitute AI for VM in the above models to get the AI1-AI4 models. In addition to model accuracy, R^2_{AQ}, we also compared different variable selection techniques as well. As mentioned in Section 4, in the LSKM method there is a proposed score test which we can use to test whether the h or h_j is needed. However, within a functional it does not select important features. The MFRS framework

allows for both the functional and within functional selection for the SGL penalty function. We report both the results of the score tests and the MFRS framework for variable selections. In addition, we applied the sparse group lasso to the model M1 and the lasso to the models VM1 and AI1 for additional comparison. Models M1-M4, VM1-VM5 and AI1-AI5 were repeated with Weight and Pulse Pressure in place of BMI as well. We include M0 without the accelerometer covariate, to assess the importance of using the accelerometer data in terms of model accuracy. In addition, since we took the Z-scores of our outcome variable (mean centered and scaled to variance 1), $\beta_0 = 0$ in the above models. Statistical analysis was conducted in R. For part of our analysis we used the existing R packages `EMMREML` and `OEM` respectively available at:

https://cran.r-project.org/web/packages/oem/index.html, and

https://cran.r-project.org/web/packages/KSPM/index.html.

3.5.1 Results from ELEMENT dataset

The mean ± SD weight of the study was 54.7 ± 13.2 kg. The mean age of the study was 14.3 ± 2.1 years. The mean BMI was 21.5 ± 4.1 where BMI was calculated as $\frac{weight(kg)}{height(m^2)}$. The mean Pulse Pressure was 73.9 ± 12.1.

From Table 3.1, we see a large difference in the number of FPC scores (17-18) extracted from using the 7-day functional data, as opposed to the 1-day averaged functional data to produce our FPC scores (4-5). Due to the small number of components extracted from VM and AI in the case of 1-day averaged functional data, we did not employ methods that perform variable selection (MFRS, COSSO or Lasso). However, we still used the sparse group lasso as the penalty function for the 1-day averaged functional data when using the Tri-Axis AC data. From the tables below we see a clear need to regularize the many FPC scores extracted, as the MFRS model in general did better than the other models. We can also note two interesting points,

detailed below. We will go through this in detail below.

Table 3.1: #FPC Scores that explain $\geq 50\%$

Functional	**7-day functional**	**1-day averaged functional**
ACX	20	6
ACY	19	5
ACZ	18	4
VM	19	5
AI	18	4

Table 3.2: FPCA Functional selection of 3-D Activity Count for X,Y and Z axis for the 7 day functional

	BMI			**WEIGHT**			**BPP**		
Model	X	Y	Z	X	Y	Z	X	Y	Z
Linear Model+SGL (M1)	✓	✓	✓	✓	✓		✓	✓	✓
LSKM non-additive (M2)							✓	✓	✓
LSKM additive (M3)									
FAM+COSSO (M4)	✓		✓	✓		✓	✓	✓	✓
MFRS+SGL (M5)	✓	✓	✓		✓		✓	✓	✓

Table 3.3: FPCA Functional selection of 3-D Activity Count for X,Y and Z axis for the 1 day averaged functional

	BMI			**WEIGHT**			**BPP**		
Model	X	Y	Z	X	Y	Z	X	Y	Z
Linear Model+SGL (M1)	✓	✓		✓	✓		✓		✓
LSKM non-additive (M2)				✓	✓	✓	✓	✓	✓
LSKM additive (M3)		✓			✓				
FAM+COSSO (M4)	✓	✓			✓		✓	✓	✓
MFRS+SGL (M5)	✓	✓		✓	✓		✓	✓	✓

Table 3.4: R^2_{AQ} using the 7-day functional of 3-D Activity Count

Model	BMI	WEIGHT	BPP
Linear Model (M0)	0.07	0.24	0.13
Linear Model (M1)	0.11	0.25	0.13
LSKM non-additive (M2)	0.1	0.25	0.13
LSKM additive (M3)	0.25	0.25	0.14
FAM+COSSO (M4)	0.11	0.26	0.16
MFRS (M5)	0.55	0.28	0.17

Table 3.5: R^2_{AQ} using the 7-day functional of VM Activity Count

Model	BMI	WEIGHT	BPP
Linear Model (M0)	0.07	0.24	0.13
Linear Model (VM1)	0.10	0.25	0.11
LSKM non-additive (VM2)	0.10	0.25	0.13
FAM+COSSO (VM4)	0.12	0.26	0.15
MFRS+LASSO (VM5)	0.21	0.31	0.14

Table 3.6: R^2_{AQ} using the 7-day functional of AI

Model	BMI	WEIGHT	BPP
Linear Model (M0)	0.07	0.24	0.13
Linear Model (AI1)	0.09	0.25	0.12
LSKM non-additive (AI2)	0.10	0.25	0.13
FAM+COSSO (AI4)	0.10	0.26	0.14
MFRS+LASSO (AI5)	0.11	0.27	0.15

Table 3.7: R^2_{AQ} using 1-day averaged functional of 3-D Activity Count

Model	BMI	WEIGHT	BPP
Linear Model (M0)	0.07	0.24	0.13
Linear Model (M1)	0.12	0.27	0.14
LSKM non-additive (M2)	0.18	0.31	0.13
LSKM additive (M3)	0.19	0.31	0.14
FAM+COSSO (M4)	0.13	0.27	0.14
MFRS+SGL (M5)	0.26	0.35	0.21

Table 3.8: R^2_{AQ} using 1-day averaged functional of VM Activity Count

Model	**BMI**	**WEIGHT**	**BPP**
Linear Model (M0)	0.07	0.24	0.14
Linear Model (VM1)	0.11	0.26	0.13
LSKM non-additive (VM2)	0.12	0.27	0.14

Table 3.9: R^2_{AQ} using 1-day averaged functional of AI

Model	**BMI**	**WEIGHT**	**BPP**
Linear Model (M0)	0.07	0.24	0.14
Linear Model (AI1)	0.10	0.25	0.13
LSKM non-additive (AI2)	0.11	0.25	0.13

3.5.2 7-day vs 1-day averaged

From Tables 3.4 and 3.7, we see that when using all three axes of functional data, 7 days of accelerometer data tended to do better for BMI while for Weight, the 1-day averaged functional did better. Without the MFRS framework, treating the 7 days of data as one long functional does not seem to be justified as we can see by comparing the 7-day functional tables to the 1-day averaged functional tables for all other models. This is due to the increase in noise that is accompanied when using the full 7 days. This could be due to the high number of FPC scores extracted when treating the 7 days as one long functional. We saw that the MFRS+SGL (M5) model had the highest R^2_{AQ} of 0.55. Comparing Tables 3.9 with 3.8 and Tables 3.6 with 3.5 we do not see a significant difference between using the VM or AI.

3.5.3 Tri-axis AC vs VM vs AI

Comparing Tables 3.4 and 3.5 shows that using all three axes of the accelerometer data tend to perform in similarly to the VM summary with the exception to the MFRS estimator.

The results illustrate that, using the acceleromter data as a functional covariate

seemed to help explain the variance in the three outcomes, particularly with BMI. However, with the exception of the MFRS+SGL estimator, a functional linear model (FLM) seemed to suffice. There were some notable differences in terms of functional selection between the COSSO and the MFRS estimator as show in table 3.2 particularly with using the Weight as a potential health outcome in the 7 day functional setting. The AI did not perform well across all models and health outcomes.

3.6 Simulation

We performed a simulation experiment mimicking acceleromter data to investigate the performance of our proposed MFRS procedure, including the performance of variable selection and its overall accuracy for accelerometer type of data. Specifically, we wanted to investigate the ability of the models listed in the above section to correctly specify and model which axes are associated with a potential health outcome. Performance in variable selection is summarized in terms of the stability measured by sensitivity and specificity for both functional and variable selections under these this simulation as well. In all analyses, we used the Gaussian Kernel $\mathcal{K}(u,v) = \exp^{-\frac{1}{s}\|u-v\|^2}$ in our estimation where s was set as the number of features, which is equivalent to dividing the $\boldsymbol{\gamma}$ vector by $\sqrt{s}$. See [20] for a theoretical justification to using the number of features for the bandwidth parameter, s, for the Gaussian Kernel.

To simulate acceleromter data, we first generated 9 eigenfunctions using FPCA on 400 curves each representing a functional of one day. We drew the 400 curves from a multivariate normal distribution with mean 0 and $\Sigma_{1400\times1400}$ matrix according to the following scheme:

1. From 8:00AM-12:00PM we assumed the variance at each minute was 1 (light variability)

2. From 12:01PM-6:00PM we assumed the variance at each minute was 2 (moder-

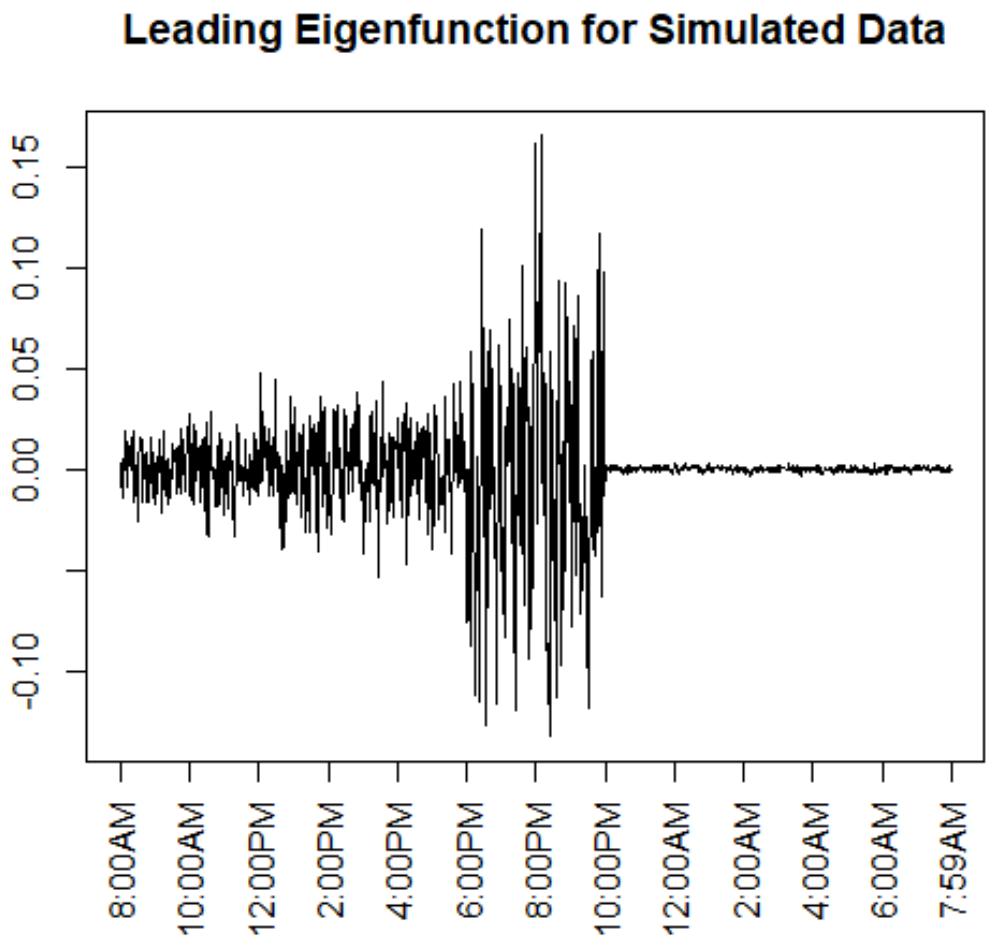
ate variability)

3. From 6:01PM-10:00PM we assumed the variance at each minute was 5 (heavy variability)

4. From 10:01PM-7:59AM we assumed the variance at each minute was .01 (sleeping)

Let V_k denote the variance at time $k \in \{1, \cdots, 1440\}$.
We set $\Sigma_{ij} = \sqrt{V_j V_i}\exp(-(i-j)^2)$ to corrolate our time points. We then performed FPCA and extracted 9 eigenfunctions, $\phi_p, p \in \{1 \cdots 9\}$. Figure 3.7 below is what the leading eigenfunction looked like. Here for each subject i, we generated 3 functional predictors $\{Z_i^1, \cdots, Z_i^3\}$ of the form: $Z^\ell(t) = \sum_{j=1}^{9} \sqrt{\varsigma_j}\xi_j\phi_j(t), \ell = 1, \ldots, 3$, where $\varsigma_j = 45 \times 0.64^j$ and $\xi_j \sim N(0,1)$ which is comparable to what was done in the [34]. There were 1440 sampled points, t, equally spaced in the interval $[1, 1440]$ corresponding to the minutes of the day. As was done in [34], instead of directly using ξ_j, we used $\zeta_j = \Phi(\xi_j)$, where Φ is the CDF of the standard normal. It follows that $\vec{\mathbf{z}} = (\zeta_1^1, \ldots, \zeta_9^1, \ldots, \zeta_1^4, \ldots, \zeta_9^3)^\top$ where ζ_j^ℓ is the jth transformed feature for the ℓth functional covariates.

Figure 3.7:



To specify sparsity, we chose the first and second functional covariates, $Z^1(t)$ and $Z^2(t)$, by relating the following transformed features, $\{\zeta_1^1, \zeta_3^1, \zeta_4^2, \zeta_7^2\}$; $\{\zeta_1^1, \zeta_3^1\}$ are the first and third features from the first functional, $Z^1(t)$, and $\{\zeta_4^2, \zeta_7^2\}$ are the fourth and seventh feature from the second functional $Z^2(t)$. We related the above FPC scores to the outcome in the following two models:

1. Scenario 1: Non-Linear and Non-additive model:

$$y_i = 0.5x_i + 3\log\zeta_{i3}^1 \log\zeta_{i7}^2 + 10\zeta_{i4}^2 \cos(2\pi\zeta_{i1}^1) + \epsilon_i, i = 1, \cdots, n$$

2. Scenario 2: Non-Linear and additive:

$$y_i = 0.5x_i + 3\log\zeta_{i3}^1 + \log\zeta_{i7}^2 + 10\zeta_{i4}^2 + \cos(2\pi\zeta_{i1}^1) + \epsilon_i, i = 1, \cdots, n$$

where $\epsilon_i \overset{iid}{\sim} N(0,1)$, $x_i \overset{iid}{\sim} N(0,1)$, and ζ_{ij}^{ℓ} is the jth transformed feature for the ℓth functional predictor $Z^{\ell}(t)$ for subject i. In both scenarios we set up both group sparsity (with only 2 of the 3 functional predictors being used) and within-group sparsity (with less then 9 of FPC features being used). We generated 400 IID functional paths of which 300 paths were assigned to the training set and 100 paths were assigned to the test set. It is the test set that we used to demonstrate the performance. The purpose of presenting both the additive and non-additive scenarios is twofold. The first objective is to demonstrate the power of using the MFRS framework over other additive methods like the COSSO in Scenario 1. The second objective is to illustrate the difference in performance between an additive and non-additive assumption. It is this second objective that could explain why the MFRS estimator outperformed COSSO using the ELEMNENT dataset.

We used SGL as the penalty function in our implementation, termed as $MFRS_{SGL}$. It should be noted that the penalty function for the sparse group lasso was defined by taking $p(\boldsymbol{\gamma};\delta) = (1-\delta)\sum_{\ell=1}^{p}\|\gamma_\ell\|_2 + \delta\|\gamma\|_1$, $\delta \in [0,1]$ and in our simulation, δ was set to 0.95. This was similar to what was done in the original paper for the sparse group lasso [47], where they mentioned that for the simulated data, δ was set to 0.95. If strong group sparsity in anticipated but only mild within group sparistiy they recommended setting δ to 0.05. We compared the results of our method with using a linear model with SGL, and with the COSSO method for functional additive regression [57]. In addition, we compared our method with LSKM and an oracle LSKM estimator, called $LSKM^{oracle}$, that assumed the full knowledge of the true ζ's and the true signals, namely, $\{\zeta_1^1, \zeta_3^1, \zeta_4^2, \zeta_7^2\}$.

From Tables 3.4 we see that the $MFRS_{SGL}$ outperformed all of the non-oracle estimators by a large margin in terms of model consistency in Scenario 1. We believe this is owing to the fact that only the $MFRS_{SGL}$ has the ability to model non-additive (and non-linear) models and select important features. Only the $LSKM^{oracle}$ model outperformed the $MFRS_{SGL}$ estimator which assumed full knowledge of which signals were important and the true values of those signals.

In terms of functional selection, COSSO did well for both scenarios as shown in Tables 3.11 and 3.12. As we can see from Table 3.12, COSSO is not able to pick up the signal of $\hat{\zeta}_4^2$, in Scenario 1, however, we see COSSO improve dramatically in Scenario 2 where the true model is additive. COSSO was designed specifically for this framework and is now able to pick up the correct FPC signals as shown in 3.13. In Scenario 2, $MFRS_{SGL}$ performed comparably to COSSO in terms of model consistency but COSSO overall did better in terms of variable selection specifically for Z^1.

Table 3.10: R^2_{AQ} for Simulated acceleromter data Scenarios 1 and 2

Model	**Scenario 1**	**Scenario 2**
Linear Model + SGL	0.19	0.73
LSKM	0.21	0.76
LSKM$^{\text{oracle}}$	0.87	0.89
COSSO	0.58	0.86
MFRS$_{\text{SGL}}$	0.70	0.84

Table 3.11: Sensitivity and Specificity of Functional Selection

	Selection Frequency					
	Scenario 1			**Scenario 2**		
Model	$\hat{Z}^1$	$\hat{Z}^2$	$\hat{Z}^3$	$\hat{Z}^1$	$\hat{Z}^2$	$\hat{Z}^3$
Linear Model + SGL	99	100	17	100	100	34
MFRS$_{\text{SGL}}$	98	96	17	100	100	14
COSSO	100	100	12	100	100	11

Table 3.12: Feature Selection for Scenario 1

	Functional Z^1								
Model	$\hat{\zeta}_1^2$	$\hat{\zeta}_2^2$	$\hat{\zeta}_3^2$	$\hat{\zeta}_4^2$	$\hat{\zeta}_5^2$	$\hat{\zeta}_6^2$	$\hat{\zeta}_7^2$	$\hat{\zeta}_8^2$	$\hat{\zeta}_9^2$
Linear Model + SGL	11	15	99	7	5	5	12	8	5
MFRS$_{\mathbf{SGL}}$	96	30	98	14	24	29	28	23	36
COSSO	99	8	100	3	0	1	1	5	1
	Functional Z^2								
	$\hat{\zeta}_1^2$	$\hat{\zeta}_2^2$	$\hat{\zeta}_3^2$	$\hat{\zeta}_4^2$	$\hat{\zeta}_5^2$	$\hat{\zeta}_6^2$	$\hat{\zeta}_7^2$	$\hat{\zeta}_8^2$	$\hat{\zeta}_9^2$
Linear Model + SGL	5	11	10	13	10	13	100	9	8
MFRS$_{\mathbf{SGL}}$	21	21	28	71	31	29	95	20	24
COSSO	2	1	4	2	2	12	100	4	2

Table 3.13: FPC Selection for Scenario 2

	Functional Z^1								
Model	$\hat{\zeta}_1^2$	$\hat{\zeta}_2^2$	$\hat{\zeta}_3^2$	$\hat{\zeta}_4^2$	$\hat{\zeta}_5^2$	$\hat{\zeta}_6^2$	$\hat{\zeta}_7^2$	$\hat{\zeta}_8^2$	$\hat{\zeta}_9^2$
Linear Model + SGL	36	49	100	38	19	11	15	15	13
MFRS$_{\mathbf{SGL}}$	81	44	100	24	23	27	24	21	23
COSSO	92	43	100	27	6	3	2	1	1
	Functional Z^2								
	$\hat{\zeta}_1^2$	$\hat{\zeta}_2^2$	$\hat{\zeta}_3^2$	$\hat{\zeta}_4^2$	$\hat{\zeta}_5^2$	$\hat{\zeta}_6^2$	$\hat{\zeta}_7^2$	$\hat{\zeta}_8^2$	$\hat{\zeta}_9^2$
Linear Model + SGL	30	47	58	100	41	25	98	16	10
MFRS$_{\mathbf{SGL}}$	20	28	37	100	33	28	91	33	27
COSSO	12	36	45	100	32	12	100	3	2

3.7 Discussion

In this chapter we applied at new techniques to both preprocess and model how accelerometer data relates to different health outcomes. We investigated particularly nonparametric modeling of the accelerometer data with a health outcome adjusting for baseline confounders. We compared two different approaches to processing multiple days of accelerometer wear and found that depending on the health outcome of interest, using the full 7 days of accelerometer data could achieve better results in terms of R^2_{AQ} when using the MFRS method. An explanation for this is that a

person's physical activity varies across days with strong heterogeneity, and therefore averaging across multiple days in epochs of minute does not fully capture the physical activity performed during the duration of the 7 days of wear. It is always important to perform variable selection on FPC features to identify signals of physical activities. We also analyzed the VM and AI and found that in the MFRS, the tri-axis data outperformed the VM and AI summary to predict the outcome. The MFRS framework also outperformed the additive LSKM model which confirms that movement in one directions is related to movement in another direction. As discussed, there is no consensus on how to treat multiple days of accelerometer wear. In the next chapter we will look at yet another way on how to decompose multiple days of accelerometer data.

CHAPTER IV

Accelerometer Modeling with Multilevel Functional Principal Component Analysis

4.1 Introduction

This chapter of the dissertation proposes an alternative way to handle multiple days of accelerometer functional data within the MFRS framework. In Chapter III, we analyzed the 7 days of accelerometer data from the ELEMENT cohort study in two ways:

1. Viewing the full 7 days of accelerometer data as one long functional;
2. Averaging minute-by-minute of the 7 days of data within a 24 hour time window, and viewing the aggregated data as a one-day functional.

Both ways listed above do not fully capture the person's inter-day variability. When faced with many days of data, it becomes difficult to view the data as one long functional as there naturally exist day-to-day heterogeneity of personal activities. Consequently, the extracted FPCA components becomes less informative as the nature of a person's physical activity is not entirely periodic. Averaging the minute-by-minute throughout multiple days also poses a potential loss of inter-day variability as mentioned in Chapter III; there is an underlying assumption that personal daily

patterns are repeated over a 24 hour time window across multiple days. To address these issues, we take a different data preprocessing route in this chapter. We decompose the accelerometer functional data by the functional equivalent of analysis of variance (ANOVA). The following idea for decomposing a functional into different components is proposed in [46]. Let $(Z_{ij}^1(t), Z_{ij}^2(t), Z_{ij}^3(t))$ denote the three-axes functional for individual i at day j for time t where the random process Z_{ij}^ℓ is assumed to have mean zero for $\ell \in \{1, 2, 3\}$. Similarly consider $Z_{ij}^{VM}(t)$ and $Z_{ij}^{AI}(t)$ the VM and AI functionals respectively, defined in Chapter I and Chapter III. We decompose each axis of functional data (representing multiple days) into repeated daily measurements of functional data:

$$Z_{ij}^\ell(t) = X_i^\ell(t) + U_{ij}^\ell(t), \ell = 1, 2, 3, \tag{4.1.1}$$

where $X_i^\ell(t)$ is the subject variability, and $U_{ij}^\ell(t)$ is within subject variability. As shown in [46], one may consider three-way factors, which nest hours within days within subject. In this chapter, we will focus on the two-factor decomposition, namely days and subject. The reason that we ignore hours as a factor is that the activity of a person seems to be highly variable, so that it is difficult to yield a stable signal at the hourly level. Equation (4.1.1) is referred to as the "noise-free" model. However in reality, as suggested in [46] a decomposition model should take account "noise". That is,

$$Z_{ij}^\ell(t) = X_i^\ell(t) + U_{ij}^\ell(t) + \epsilon_{ij}(t), \tag{4.1.2}$$

where $\epsilon_{ij}(t)$ is the white noise $\overset{iid}{\sim} N(0, \sigma^2)$ (for convenience we will assume it is normally distributed). In the next subsection we discuss the detail on the assumptions and the algorithm for decomposing a functional in (4.1.1) and then proceed to model the level-specific stochastic process $X_i^\ell(t)$ with the health outcomes from our ELEMENT dataset within the MFRS framework.

4.2 Functional Anova Model

Define the covariance functions K_Z^ℓ, K_X^ℓ and K_U^ℓ as $K_Z^\ell(t,s) := Cov(Z_i^\ell(t), Z_i^\ell(s))$, $K_X^\ell(t,s) := Cov(X_i^\ell(t), X_i^\ell(s))$ for all i and $K_U^\ell(t,s) := Cov(U_{ij}^\ell(t), U_{ij}^\ell(s))$ for all i,j. For the purpose of identifiability, we assume that the random process $X_i^\ell(t)$ and $U_{ij}^\ell(t)$ are mean zero and uncorrelated with each other.

Lemma 7. *Under the assumption that $X_i^\ell(t)$ and $U_{ij}^\ell(t)$ are mean zero and uncorrelated, we have the following equality:*

$$K_Z^\ell = K_X^\ell + K_U^\ell. \tag{4.2.1}$$

Proof.

$$\begin{aligned} K_Z^\ell(t,s) &= Cov(Z_i^\ell(t), Z_i^\ell(s)) \\ &= Cov(X_i^\ell(t) + U_{ij}^\ell(t), X_i^\ell(s) + U_{ij}^\ell(s)) \text{ for all } i,j \text{ by (4.1.1)} \\ &= Cov(X_i^\ell(t), X_i^\ell(s)) + Cov(U_{ij}^\ell(t), U_{ij}^\ell(s)) + Cov(X_i^\ell(t), U_{ij}^\ell(s)) + Cov(U_{ij}^\ell(t), X_i^\ell(s)) \\ &= K_X^\ell + K_U^\ell \text{ (since } X_i^\ell \text{ and } U_{ij}^\ell \text{ are uncorrelated).} \end{aligned} \tag{4.2.2}$$

□

As in Chapter II, we use FPCA to decompose X_i^ℓ and U_{ij}^ℓ to yield their respective FPC features. By the truncated Karhunen-Loève expansion we can rewrite (4.1.1) as

$$Z_{ij}^\ell(t) = \sum_{m=1}^{N_1} \phi_{X_m}^\ell(t)\xi_{Xim}^\ell + \sum_{n=1}^{N_2} \phi_{U_n}^\ell(t)\xi_{Uijn}^\ell \tag{4.2.3}$$

where the eigenfunctions of X^ℓ and U^ℓ are $\{\phi_{X_m}^\ell; m = 1, \cdots, N_1\}$ and $\{\phi_{U_n}^\ell; n = 1, \cdots, N_2\}$, respectively. Once again, we truncated the infinite sums of the Karhunen-Loève expansion with finite numbers N_1 and N_2 as given in (4.2.3). We assume that the variances are homogeneous in that $var(\xi_{Xim}^\ell) = \varsigma_m^X$ and $var(\xi_{Uijn}^\ell) = \varsigma_n^U$. Notice

that the above variances do not depend on the subject, i nor the day, j. This is an important assumption that is used in structured functional principal component analysis (SFPCA). It is similar to the sphericity assumption in repeated measures Anova and homogeneity of variance assumption in Anova. We will discuss the ramifications of this assumption of homogeneity of variance in more detail later in this chapter.

In application to our accelerometer functional data, we truncated the expansion based on the eigenvalues associated with the eigenfunctions that explain the majority (greater than 50%) of the variance. Following the procedure in [46] we estimated the FPC scores $\xi^{\ell}_{X_{im}}$ and $\xi^{\ell}_{U_{ijn}}$ using the best linear unbiased predictor (BLUP) of the mixed effects model, leading to the following sample counterpart of the decomposition (4.2.3):

$$\mathbf{Z}^{\ell}_{ij} = \hat{\Phi}^{\ell}_{X}\boldsymbol{\xi}^{\ell}_{X_i} + \hat{\Phi}^{\ell}_{U}\boldsymbol{\xi}^{\ell}_{U_{ij}}, \tag{4.2.4}$$

where $\mathbf{Z}^{\ell}_{ij} := (Z^{\ell}_{ij}(t_1), \cdots, Z^{\ell}_{ij}(t_w))$ for w time points (in our case, w will be 1440 corroding to the minutes of the day), $\hat{\Phi}^{\ell}_{X}$ is the estimate of $\Phi^{\ell}_{X} = (\phi^{\ell}_{X_1}, \cdots, \phi^{\ell}_{X_{N_1}})$, $\hat{\Phi}^{\ell}_{U}$ is the estimate of $\Phi^{\ell}_{U} = (\phi^{\ell}_{U_1}, \cdots, \phi^{\ell}_{U_{N_2}})$, $\boldsymbol{\xi}^{\ell}_{X_i} \sim N(\mathbf{0}, \hat{\Lambda}^{\ell}_{X})$ and $\boldsymbol{\epsilon}^{\ell}_{U_{ij}} \sim N(\mathbf{0}, \hat{\Lambda}^{\ell}_{U})$ (normality is not necessary but assumed for convenience) for estimated eigenvalue matrices associated with the first N_1 and N_2 eigenvalues for $\hat{\Phi}^{\ell}_{X}$, $\hat{\Phi}^{\ell}_{U}$. The steps needed to obtain (4.2.4) are as follows:

1. Estimate the covariance matrices $\hat{K^{\ell}}_{X}$ and $\hat{K^{\ell}}_{U}$.

2. Extract the eigenfunctions $\hat{\Phi}^{\ell}_{X}$, $\hat{\Phi}^{\ell}_{U}$, and eigenvalue matrices, $\hat{\Lambda}^{\ell}_{X}$ and $\hat{\Lambda}^{\ell}_{U}$ from Step 1.

3. Solve for the FPC scores, $\boldsymbol{\xi}^{\ell}_{X_i}$ and $\boldsymbol{\xi}^{\ell}_{U_{ij}}$ using the means of BLUP in (4.2.4).

Our goal is to use the estimated FPC scores, $\boldsymbol{\xi}^{\ell}_{X_i}$, of the subject level process, $X^{\ell}_{i}(t)$ to replace the scores used in Chapter III to perform the analysis. For Step 1, we use

the method of moment estimators (MoM) in a similar way to what was proposed in [46]. Specifically, in the case of our model, we have these MoM equations:

$$
\begin{aligned}
&E\{\left(Z_{ij}^{\ell}(t)-Z_{mn}^{\ell}(t)\right)\left(Z_{ij}^{\ell}(s)-Z_{mn}^{\ell}(s)\right)\} \\
&= E\{Z_{ij}^{\ell}(t)Z_{ij}^{\ell}(s)\}+E\{Z_{mn}^{\ell}(t)Z_{mn}^{\ell}(s)\}-E\{Z_{ij}^{\ell}(t)Z_{mn}^{\ell}(s)\}-E\{Z_{mn}^{\ell}(t)Z_{ij}^{\ell}(t)\} \\
&= \begin{cases} 2K_U^{\ell}(t,s), & \text{when } i=m, j\neq n; \\ 2K_X^{\ell}(t,s)+2K_U^{\ell}(t,s), & \text{when } i\neq m. \end{cases}
\end{aligned}
\tag{4.2.5}
$$

Note that our dataset has 395 participants with accelerometer data for 7 days. Let $H_U = \frac{1}{395*7*6}\sum_{i=1}^{395}\sum_{j=1}^{7}\sum_{n\neq j}(\mathbf{Z}_{ij}^{\ell}-\mathbf{Z}_{in}^{\ell})(\mathbf{Z}_{ij}^{\ell}-\mathbf{Z}_{in}^{\ell})^{\top}$, for the 1440×1 vector $\mathbf{Z}_{ij}^{\ell}$ (representing the 1440 minutes in 24 hours). The denominator, $395*7*6$, corresponds to the sum with $i=m, j\neq n$, where each subject contributes $7*6$ times in the summation (each day with all other days within subject). Our MoM estimator, $\hat{K}_U^{\ell}(t,s)$ would then be obtained as $\frac{H_U}{2}$. Similarly, we let $H_X = \frac{1}{395*7^2*394}\sum_{i=1}^{395}\sum_{m\neq i}\sum_{j=1}^{7}\sum_{n=1}^{7}(\mathbf{Z}_{ij}^{\ell}-\mathbf{Z}_{mn}^{\ell})(\mathbf{Z}_{ij}^{\ell}-\mathbf{Z}_{mn}^{\ell})^{\top}$. The denominator, $395*7^2*394$ corresponds to the fact we are summing up when $i\neq m$ so each of the 395 subjects 7 days ($395*7$) will be used against all other days for all subject ($394*7$). By MoM, then estimate $\hat{K}_X^{\ell}(t,s)$ as $\frac{H_U-H_X}{2}$. For Step 2, estimating the eigenfunctions and eigenvalues from the above matrices is straightforward (e.g. SVD). Step 3 is a special case of a three-way nested design provided in the appendix of [46]. Following the same logic, the solution for the two-way nested design is:

$$
\begin{bmatrix} \hat{\boldsymbol{\xi}}_{X_i}^{\ell} \\ \hat{\boldsymbol{\xi}}_{U_i}^{\ell} \end{bmatrix} = \begin{bmatrix} 7*I_{N_1\times N_1} & \mathbf{1}^{\top}\otimes(\hat{\boldsymbol{\Phi}}_X^{\ell})^{\top}\hat{\boldsymbol{\Phi}}_U^{\ell} \\ & I_{7\times 7}\otimes I_{N_2\times N_2} \end{bmatrix} \begin{bmatrix} (\hat{\boldsymbol{\Phi}}_X^{\ell})^{\top}\mathbf{Z}_i^{\ell}\mathbf{1}_7 \\ vec((\hat{\boldsymbol{\Phi}}_U^{\ell})^{\top}\mathbf{Z}_i^{\ell}) \end{bmatrix},
$$

where $\mathbf{Z}_i^{\ell}$ is a matrix of dimension 1440×7 whose (m,j)th element is $Z_{ij}^{\ell}(t_m)$, $\hat{\boldsymbol{\xi}}_{X_i}^{\ell} = (\hat{\xi}_{X_{i1}}^{\ell},\cdots,\hat{\xi}_{X_{iN_1}}^{\ell})^{\top}$, $\hat{\boldsymbol{\xi}}_{U_i}^{\ell} = (\hat{\xi}_{U_{i11}}^{\ell},\cdots,\hat{\xi}_{U_{i1N_2}}^{\ell},\cdots,\hat{\xi}_{U_{i7N_2}}^{\ell})^{\top}$, $I_{c\times c}$ is a $c\times c$ identity matrix, $\otimes$ is the Kronecker product, and vec is an operation of vectorizing a matrix. In total,

there are $395 \times N_1$ FPC estimated scores from the X^ℓ process and $395 * 7 * N_2$ FPC estimated scores from the U^ℓ process. It should be noted that there is another way proposed in [23] to estimate the FPC scores through Markov Chain Monte Carlo (MCMC) methods. However, we focused on using the BLUP to estimate the FPC scores in this dissertation

4.3 MFRS Framework For Decomposed Functional

In Chapter II and Chapter III we defined the vector $\mathbf{z}_i^\ell = (\xi_1^\ell, \ldots, \xi_{s_\ell}^\ell)_i^\top$ as the vector of FPC features from the ith observation for the functional covariate Z^ℓ where $\ell \in \{1, 2, 3\}$ for our accelerometer data. We set $\vec{\mathbf{z}}_i = [(\mathbf{z}_i^1)^\top, (\mathbf{z}_i^2)^\top, (\mathbf{z}_i^3)^\top]^\top$ as the grand vector of all FPC features from all 3 functional covariates. We consider the model

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + h(\vec{\mathbf{z}}_i) + \epsilon_i,\ i = 1, \cdots, n.$$

Modeling the functional data through the truncated FPC scores is a proxy for the true underlying relationship which follows:

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + f(Z_i^\ell) + \epsilon_i,\ i = 1, \cdots, n, \tag{4.3.1}$$

where the unknown function f maps the functional, Z^ℓ into $\mathcal{R}$. This formulation applies in the cases of AI or VM. When considering the three axes, f is mapping (Z^1, Z^2, Z^3) into $\mathcal{R}$. We approximate the function $f(Z_i^\ell) = f(\sum_{j=1}^\infty \phi^\ell \xi_j^\ell)$, with $h(\xi_{i1}^\ell, \cdots, \xi_{is_\ell}^\ell)$ where h is mapping from $\mathcal{R}^{s_\ell}$ into $\mathcal{R}$. Decomposing the functional, Z^ℓ by (4.1.1) we look at approximating the function $\tilde{f}(Z_{i1}^\ell, \cdots, Z_{i7}^\ell) = \tilde{f}(\sum_{m=1}^\infty \phi_{X_m}^\ell \xi_{X_{im}}^\ell + \sum_{n=1}^\infty \phi_{U_n}^\ell \xi_{U_{i1n}}^\ell, \cdots, \sum_{m=1}^\infty \phi_{X_m}^\ell \xi_{X_{im}}^\ell + \sum_{n=1}^\infty \phi_{U_n}^\ell \xi_{U_{i7n}}^\ell)$ with
$\tilde{h}_1(\xi_{X_{i1}}^\ell, \cdots \xi_{X_{iN_1}}^\ell) + \tilde{h}_2(\xi_{U_{i11}}^\ell, \cdots \xi_{U_{i1N_2}}^\ell, \cdots, \xi_{U_{i7N_2}}^\ell)$, where we assume an additive model for the effects of between and within subject FPC scores through two functions, $\tilde{h}_1$

and $\tilde{h}_2$. We primarily focus on modeling $\tilde{h}_1$ as an approximation to the function $\tilde{f}(Z_{i1}^{\ell}, \cdots, Z_{i7}^{\ell})$. The downside of only using the FPC scores from the between individual process, X^{ℓ}, is that two individuals may have the same averaged physical activity as evidenced by the X^{ℓ} process yet many vary greatly with the inter-day variation. For example, an individual who is physically very active one day and physically very dormant the next day would possibly be treated the same as an individual who is moderately active on all days. Due to the high number of components that are extracted for the within subject variation, it becomes difficult to model all of the FPC scores extracted both from the X^{ℓ} process and the U^{ℓ} process. The results in the next section will be of the same format as the results in Chapter III where instead of using the standard FPCA on the entire functional, we use the FPC scores extracted from the X^{ℓ} process. For obvious reasons we only decompose the 7-day functional into the two processes X^{ℓ} and U^{ℓ} and not the 1-day averaged functional.

4.4 Results using the X(t) process

Tables 4.4 and 4.5 provide the breakdown of the variance of the $Z(t)$ process when it is decomposed into the between individual process, $X(t)$, and within individual process, $U(t)$. The contribution from the $X(t)$ process is given by $\frac{\sum_{i=1}^{N_1} \varsigma_i^X}{\sum_{i=1}^{N_1} \varsigma_i^X + \sum_{j=1}^{N_2} \varsigma_j^U}$ and the contribution of the $U(t)$ process is equal to $\frac{\sum_{j=1}^{N_2} \varsigma_j^U}{\sum_{i=1}^{N_1} \varsigma_i^X + \sum_{j=1}^{N_2} \varsigma_j^U}$. In section 4.5 we will briefly look at a possible way of using the information contained in $U(t)$. When relating the accelerometer data to a health outcome, the $\tilde{h}_2$ function does not seem to provide additional information that is not already contained in the $\tilde{h}_1$ function with respect to the model fit . When looking at the performance of using the $X(t)$ process in Tables 4.1 and 4.3 it is more appropriate to compare this with the 1-day averaged functional tables from Chapter III. This is because both the FPC scores extracted from the 1-day averaged functional and the FPC scores extracted

from the $X(t)$ functional are trying to extract the general diurnal patterns of an individual. The eigenfunctions extracted from the $X(t)$ process are similar in pattern to that of the 1-day averaged functionals. See Appendix C for graphs that display the leading eigenfunctions. The FPC scores in the $X(t)$ process not only outperformed the FPC scores from the 1-day averaged functional for the MFRS estimator but outperformed using the full 7-day functional using all three axis (which performed the best in Chapter III) in the case of Weight and Blood Pulse Pressure (BPP). This is an indication that decomposing the $Z(t)$ with functional ANOVA might be a way to proceed with modeling multiple days of wear from an accelerometer. What was also interesting was that for the health outcome, Weight, the SFPCA procedure with the VM curve tended to do better than using the tri-axes curves. However, in all other instances, implementing SFPCA using all tri-axes curves outperformed AI and the VM functionals when using the best estimator, MFRS. From Table 4.6 we see that the functional associated with the Y-axis is consistently selected across all three health outcomes. With an accelerometer worn at the wrist, this is difficult to explain as there is no dominant plane of motion [10]. However, there is no distinct pattern for which functionals are selected or discarded among for the COSSO. MFRS was able to extract information from all three axes for all of the health outcomes. The results and models are displayed in a similar format to Chapter III with figures in Appendix C.

Table 4.1: SFPCA R^2_{AQ} using the 7-day functional of 3-D Activity Count

Model	BMI	WEIGHT	BPP
Linear Model (M0)	0.07	0.24	0.13
Linear Model+SGL (M1)	0.07	0.24	0.12
LSKM non-additive (M2)	0.10	0.24	0.13
LSKM additive (M3)	0.10	0.25	0.15
FAM+COSSO (M4)	0.10	0.25	0.15
MFRS+SGL (M5)	0.33	0.41	0.54

Table 4.2: SFPCA R^2_{AQ} using the 7-day functional of VM Activity Count

Model	BMI	WEIGHT	BPP
Linear Model (M0)	0.07	0.24	0.13
Linear Model+LASSO (VM1)	0.10	0.24	0.12
LSKM non-additive (VM2)	0.15	0.24	0.13
FAM+COSSO (VM4)	0.09	0.26	0.15
MFRS+LASSO (VM5)	0.28	0.44	0.13

Table 4.3: SFPCA R^2_{AQ} using the 7-day functional of AI

Model	BMI	WEIGHT	BPP
Linear Model (M0)	0.07	0.24	0.13
Linear Model+LASSO (AI1)	0.09	0.24	0.13
LSKM non-additive (AI2)	0.10	0.24	0.13
FAM+COSSO (AI4)	0.10	0.25	0.15
MFRS+LASSO (AI5)	0.16	0.26	0.16

Table 4.4: #FPC Scores that explain $\geq 50\%$

Functional	X process	U process (multiply by 7)
ACX	30	50
ACY	31	57
ACZ	25	55
VM	25	49
AI	17	42

Table 4.5: % of variance explained from the decomposition $Z(t) = X(t) + U(t)$

	X process		U process	
Model	**Sum of Eigenvalues**	**% of Total**	**Sum of Eigenvalues**	**% of Total**
ACX	811280553	18 %	3729411108	82 %
ACY	712722053	19 %	3119611012	81 %
ACZ	726101911	19 %	3092941823	81 %
VM	1081393401	19 %	4737383139	81 %
AI	4344.073	19 %	18629.79	81 %

Table 4.6: SFPCA Functional selection of 3-D Activity Count for X,Y and Z axis

	BMI			WEIGHT			BPP		
Model	X	Y	Z	X	Y	Z	X	Y	Z
Linear Model+SGL (M1)	✓	✓	✓	✓	✓				
LSKM non-additive (M2)							✓	✓	✓
LSKM additive (M3)					✓				
FAM+COSSO (M4)		✓	✓	✓	✓		✓	✓	✓
MFRS+SGL (M5)	✓	✓	✓	✓	✓	✓	✓	✓	✓

4.5 Joint Modeling with both the X(t) and U(t) processes

4.5.1 Setup

In the previous section we modeled the diurnal physical activity through the FPC scores extracted from the $X(t)$ process. There were about 50 FPC scores extracted from the $U(t)$ process for each function as displayed in Table 4.4 *for each day* (so 7 times that amount was extracted for the 7 days). To distinguish between two individuals that have the same mean physical activity pattern as presented with the $X(t)$ process, we decided to take the variance of the FPC scores extracted from the $U(t)$ process. For example, there were $50 * 7$ scores extracted for each individual i for the functional ACX, $\{\hat{\xi}^1_{U_{ijk}}\}$, $j \in \{1, \cdots, 7\}, k \in \{1, \cdots, 50\}$. We made new scores for each individual, i, $\hat{\xi}^1_{U_{ik}}$ where $\hat{\xi}^1_{U_{ik}}$ is the sample variance from $\{\hat{\xi}^1_{U_{i1k}}, \cdots, \hat{\xi}^1_{U_{i7k}}\}$. As explained earlier in this chapter, the assumption SFPCA makes, similar to the homogeneity of variance assumptions in the ANOVA context, is that the variance $var(\xi^1_{U_{ijk}}) = \varsigma^U_k$ for all i, j. In the context of physical activity, this may not be correct as peoples diurnal routines do not have to follow the same level of variability. By looking at sample variance summaries, $\hat{\xi}^1_{U_{ik}}$ we are allowing for such variability. Unlike in the previous section where we were concerned with estimating $\tilde{f}$ with $\tilde{h}_1$, we estimate $\tilde{f}(\sum_{m=1}^{\infty} \phi^{\ell}_{X_m} \xi^{\ell}_{X_{im}} + \sum_{n=1}^{\infty} \phi^{\ell}_{U_n} \xi^{\ell}_{U_{i1n}}, \cdots, \sum_{m=1}^{\infty} \phi^{\ell}_{X_m} \xi^{\ell}_{X_{im}} + \sum_{n=1}^{\infty} \phi^{\ell}_{U_n} \xi^{\ell}_{U_{i7n}})$

jointly with

$\tilde{h}(\vec{\mathbf{z}}_i)$ where $\vec{\mathbf{z}}_i = (\hat{\xi}^{\ell}_{X_{i1}}, \cdots, \hat{\xi}^{\ell}_{X_{iN_1}}, \hat{\xi}^{\ell}_{U_{i1}}, \cdots, \hat{\xi}^{\ell}_{U_{iN_2}})$. As before, when considering multiple functional predictors (unlike the AI or VM) you extend the $\vec{\mathbf{z}}_i$ vector to accommodate all the FPC scores from the three axis $\ell \in \{1, 2, 3\}$.

4.5.2 Results from the joint modeling

From Table 4.10 we see that only the COSSO chose some of the scores from those of the $U(t)$ process. None of the other estimators selected scores from the $U(t)$ process for any of the health outcomes. This may indicate that most of the signal from $Z(t)$ is contained in $X(t)$. For the most part we did not see an improvement over using the $X(t)$ process alone. This confirms that $U(t)$ may be a "noisy" part of the decomposition.

Table 4.7: SFPCA R^2_{AQ} using the 7-day functional of 3-D Activity Count using $X(t)$ and $U(t)$ process.

Model	BMI	WEIGHT	BPP
Linear Model (M0)	0.07	0.24	0.13
Linear Model+SGL (M1)	0.06	0.24	0.15
LSKM non-additive (M2)	0.10	0.24	0.13
FAM+COSSO (M4)	0.06	0.27	0.13
MFRS+SGL (M5)	0.33	0.41	0.61

Table 4.8: SFPCA R^2_{AQ} using the 7-day functional of VM Activity Count.

Model	BMI	WEIGHT	BPP
Linear Model (M0)	0.07	0.24	0.13
Linear Model+LASSO (VM1)	0.09	0.23	0.12
LSKM non-additive (VM2)	0.09	0.24	0.13
FAM+COSSO (VM4)	0.10	0.24	0.14
MFRS+LASSO (VM5)	0.28	0.48	0.18

Table 4.9: SFPCA R^2_{AQ} using the 7-day functional of AI using $U(t)$ and $X(t)$,

Model	BMI	WEIGHT	BPP
Linear Model (M0)	0.07	0.24	0.13
Linear Model+LASSO (AI1)	0.09	0.24	0.12
LSKM+LASSO (AI2)	0.10	0.24	0.13
FAM+COSSO (AI4)	0.10	0.25	0.13
MFRS+LASSO (AI5)	0.21	0.23	0.12

Table 4.10: SFPCA Functional selection of 3-D Activity Count for X,Y and Z axis for $U(t)$ process.

	BMI			WEIGHT			BPP		
Model	X	Y	Z	X	Y	Z	X	Y	Z
Linear Model+SGL (M1)									
LSKM non-additive (M2)									
FAM+COSSO (M4)			✓	✓		✓	✓		
MFRS+SGL (M5)									

4.6 Discussion

In this chapter we presented another method that allows us to handle multiple days of accelerometer wear through SFPCA. The results from decomposing the accelerometer data into the between and within subjects processes, $X(t)$ and $U(t)$ outperformed the previous two methods presented in Chapter III of viewing the functional as one long 7 day functional or 1 day-averaged functional. This could be because we removed the "noisy" process, $U(t)$, from $Z(t)$. We did not notice much of a difference when using the $U(t)$ process as well. It seems that using the scores from $X(t)$ extracted by SFPCA produces the best results overall. However, we did not fully use the information in the scores from the $U(t)$ process, just its sample variance. Furthermore, the MFRS framework assumes a non-additive framework. It might make more sense to model the sample variance from the $U(t)$ in a separate function from that of the FPC scores from $X(t)$ as described in Chapter III which leads to an extension of MFRS to the LSKM additive framework. A key assumption in the functional decomposition is the homogeneous variance assumption. In reality, this assumption may be violated as discussed and demonstrated in section 4.5. As mobile devices are worn for longer periods on end, the ability to estimating separate variance parameters per individual become more realistic and more appealing. With only 7 days of data, estimating a separate eigenvalue, ς_{ik}^{U} for each individual, i, is challenging, and worth further exploration.

We can also consider change points in the individual time series of functional data to allow multiple $X_i^{\ell}(t)$ processes for the same individual, i. This could happen suddenly, for example, if someone breaks a limb and becomes immobile, or if someone becomes more active. In both situations there is a break or a "change" from the previous physical activity movements as modeled with $X_i^{\ell}(t)$.

Built on the methodology framework described in Chapter II, this chapter com-

bined with the modeling framework of Chapter III demonstrates that the MFRS framework seems to outperform other existing models when relating potential health outcomes with accelerometer data. Through data analysis we have demonstrated the superior performance of the MFRS framework over other existing models. Furthermore, using all three axes from the accelerometer is possible with this framework; in general, the results from our MFRS indicate that this 3-D approach outperforms the VM and AI summaries. As more people wear accelerometer data for longer periods of time, the need to extract the signal from the noisy functional data becomes necessary and the MFRS framework has the ability to handle the complex, non-linear and non-additive potential relationship that other existing models to not have.

CHAPTER V

Functional Logistic Regression

5.1 Introduction

In the first half of this chapter we will briefly discuss a novel approach to selecting import points in the context of Kernel Logistic Regression (KLR). In the second half, we will extend the MFRS framework for a binary outcome. By the representer theorem presented in Chapter II, the estimate of $h \in \mathcal{H}_{\mathcal{K}}$ is of the form $\hat{h(\cdot)} = \sum_{k=1}^{n} \alpha_k \mathcal{K}(\cdot, \mathbf{z}_k)$ for training data $\{\mathbf{z}_1, \cdots, \mathbf{z}_n\}$. We will follow the notation described in [58]. Let $\mathcal{S} \subset \{\mathbf{z}_1, \cdots, \mathbf{z}_n\}$. We call the elements of $\mathcal{S}$ import points. We wish to estimate our function h as $\hat{h(\cdot)} = \sum_{\mathbf{z}_k \in \mathcal{S}} \alpha_k \mathcal{K}(\cdot, \mathbf{z}_k)$ using only the subset $\mathcal{S}$ of the training data. The goal is to retain the quality of our estimator (in the case of a binary outcome our objective is classification) while simultaneously reduce the computing cost that is associate with large datasets. In the past, the subset $\mathcal{S}$ was chosen independently of the outcome of interest, y [31, 48, 51]. Zhu and Hastie (2005) proposed a novel approach called Import Vector Machine (IVM) that selects $\mathcal{S}$ and estimates $\hat{h}$ while taking into account the outcome of interest through a greedy algorithm approach. We propose an alternative method that transforms the problem into a problem that iteratively solves the lasso for ordinary least squares.

In the second half of this chapter we will focus on techniques how to implement the MFRS framework in the context of a binary outcome. The objective is to use the KLR

method while simultaneously performing variable selection on the FPC scores with the MFRS framework. For ease of exposition, we will not consider the kernel bandwidth parameter in this chapter for reasons mentioned in earlier chapters. In addition we will also not consider a partially linear model but rather a fully parametric model for a single functional covariate that FPCA has been performed on. The extension to multiple functional covariates will be immediate obvious by stacking our vector of interest, $\mathbf{z}$, with the FPC scores from all of the functional covariates and changing our penalty function from the lasso (or mcp) to the sparse group lasso as described in detail in Chapter II.

5.2 Background for selection of Import points for KLR

In this section we will provide the necessarily background before presenting the proposed solution to the selection of import points that we described in the introduction to this chapter. The idea comes from four different existing statistical methods that we will briefly review. In previous chapters, the outcome that we considered, y, was continuous and our MFRS estimator simultaneously performed LSKM and variable reduction on the FPC scores. In this chapter we will consider a binary outcome $y \in \{0, 1\}$. Standard parametric logistic regression looks at the following relationship between y and our FPC scores $\mathbf{z}$:

$$E(y_i|\mathbf{z}_i) = p_i = \frac{\exp^{\mathbf{z}_i^\top \boldsymbol{\beta}}}{1 + \exp^{\mathbf{z}_i^\top \boldsymbol{\beta}}}. \tag{5.2.1}$$

From 5.2.1 we see that the linear relationship holds: $logit(p_i) = \mathbf{z}_i^\top \boldsymbol{\beta}$.

5.2.1 KLR

For non-parametric logistic regression, we are looking at the following equation:

$$logit(p_i) = h(\mathbf{z}_i) \tag{5.2.2}$$

where $h : \mathcal{R}^s \mapsto \mathcal{R}$ is an unknown function. Kernel logistic regression (KLR), assumes (5.2.2) holds where $h \in \mathcal{H}_\mathcal{K}$ for some RKHS with kernel $\mathcal{K}$. To solve the KLR for unknown function h, we look to minimize the following penalized loss function:

$$\min_{h \in \mathcal{H}_\mathcal{K}} \sum_{i=1}^{n} \left(-y_i h(\mathbf{z}_i) + ln\left(1 + \exp(h(\mathbf{z}_i))\right)\right) + \frac{\lambda_1}{2} \left\|h\right\|^2_{\mathcal{H}_\mathcal{K}}. \tag{5.2.3}$$

By the representer theorem described in Chapter II, 5.2.3 is equivalent to the following:

$$\min_{\boldsymbol{\alpha} \in \mathcal{R}^n} \sum_{i=1}^{n} \left(-y_i K_i \boldsymbol{\alpha} + ln\left(1 + \exp(K_i \boldsymbol{\alpha})\right)\right) + \frac{\lambda_1}{2} \boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha} \tag{5.2.4}$$

where K_i is the ith row of the matrix $\mathbf{K}$ whose ijth entry is $K(\mathbf{z}_i, \mathbf{z}_j)$. In matrix notation, we can write (5.2.4) as

$$\min_{\boldsymbol{\alpha} \in \mathcal{R}^n} L(\boldsymbol{\alpha}) = -\mathbf{y}^\top \mathbf{K} \boldsymbol{\alpha} + \mathbf{1}^\top ln\left(1 + exp(\mathbf{K}\boldsymbol{\alpha})\right) + \frac{\lambda_1}{2} \boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha}.$$

Using the Newton-Raphson algorithm (N-R) to solve (5.2.4) for $\boldsymbol{\alpha}$ we iteratively get:

$$\boldsymbol{\alpha}^{(n)} = \boldsymbol{\alpha}^{n-1} - H^{-1} \nabla L \big|_{\boldsymbol{\alpha}^{(n-1)}}$$

where $\nabla L = -\mathbf{K}^\top (\mathbf{y} - \mathbf{p}) + \lambda_1 \mathbf{K} \boldsymbol{\alpha}$ and the hessian, $H = \mathbf{K}^\top \mathbf{P} \mathbf{K} + \lambda_1 \mathbf{K}$ where the ith entry of $n \times 1$ vector $\mathbf{p}$ is p_i and $n \times n$ matrix $\mathbf{P} = diag\{(p_1(1-p_1), \cdots, p_n(1-p_n))\}$. We can solve for $\boldsymbol{\alpha}^{(n)}$ as:

$$\boldsymbol{\alpha}^{(n)} = \left[\mathbf{K}^\top \mathbf{P}^{(n-1)} \mathbf{K} + \lambda_1 \mathbf{K}\right]^{-1} \mathbf{K}^\top \mathbf{P}^{(n-1)} \mathbf{m}^{(n-1)} \tag{5.2.5}$$

where $\mathbf{m}^{(n-1)} = \mathbf{P}^{(n-1)^{-1}}\left(\mathbf{y} - \mathbf{p}^{(n-1)}\right) + \mathbf{K}\boldsymbol{\alpha}^{(n-1)}$. The algorithm for solving functional KLR would be as follows:

KLR Algorithm:

(i) Step1.1: Perform FPCA (e.g. R package `fdapace`) to extract the functional component scores for the functional predictor and store it in a vector for each individual subject $\mathbf{z}_i$;

(ii) Step 1.2: Set up a grid of possible tuning parameters for λ_1 and initialize $\boldsymbol{\alpha}^{(0)}$ to be a vector of ones. Perform steps Steps 2-3 below.

(iii) Step 2: At the n-th step in the algorithm, solve for $\boldsymbol{\alpha}^{(n)}$ from equation 5.2.5. Stop at convergence.

(iv) Step 3: Perform cross-validation over the specified grid of λ_1 to determine the final $\boldsymbol{\alpha}$

5.2.2 Tikhonov regularization

Consider the standard ordinary least squares problem

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathcal{R}^p}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 . \tag{5.2.6}$$

This problem may be ill-poised if $\hat{\boldsymbol{\beta}}$ is not unique for example, $p > n$. Tikhonov regularization addresses this [7] by adding a regularization term to 5.2.6 as follows:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathcal{R}^p}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \|\mathbf{T}\boldsymbol{\beta}\|_2^2 . \tag{5.2.7}$$

Where $\mathbf{T}$ is known as the Tikhonov matrix. When $\mathbf{T} = \sqrt{\lambda}\mathbf{I}_{p\times p}$ where $\lambda \in \mathcal{R}$ we get what is commonly called ridge regression [25].

The solution to (5.2.7) is:

$$\hat{\boldsymbol{\beta}} = \left[\mathbf{X}^\top \mathbf{X} + \mathbf{T}^\top \mathbf{T}\right]^{-1} \mathbf{X}^\top \mathbf{y}. \tag{5.2.8}$$

We see an equivalence between (5.2.5) and (5.2.8). If we let $\mathbf{T} = \sqrt{\frac{\lambda}{2}}\mathbf{K}^{\frac{1}{2}}$, $\mathbf{X} = \mathbf{P}^{\frac{1}{2}}\mathbf{K}$ and $\mathbf{y} = \mathbf{P}^{\frac{1}{2}}\mathbf{m}^{(n-1)}$ we get that (5.2.5) is the solution to the Tikhonov regularization problem (5.2.7).

5.2.3 Logistic regression with lasso (L1) penalty

We formulate out loss function for the regularized parametric logistic regression problem (not KLR) as follows:

$$\min_{\boldsymbol{\beta}\in\mathcal{R}^p} \sum_{i=1}^{n} \left(-y_i \mathbf{x}_i^\top \boldsymbol{\beta} + ln\left(1 + \exp(\mathbf{x}_i^\top \boldsymbol{\beta})\right)\right) + \lambda_1 \left\|\boldsymbol{\beta}\right\|_1 = \min_{\boldsymbol{\beta}\in\mathcal{R}^p} L(\boldsymbol{\beta}) + \lambda_1 \left\|\boldsymbol{\beta}\right\|_1 \tag{5.2.9}$$

Using the (N-R) algorithm to solve $\text{argmin}_{\boldsymbol{\beta}}\, L(\boldsymbol{\beta})$ (similarly to what was done for KLR in the previous subsections) we get the well known updated step as the iteratively reweighted least squares (IRLS) solution:

$$\left[\mathbf{X}^\top \mathbf{P} \mathbf{X}\right]^{-1} \mathbf{X}^\top \mathbf{P} \mathbf{m}^{(n-1)} \tag{5.2.10}$$

where $\mathbf{m}^{(n-1)} = \mathbf{P}^{-1}(\mathbf{y} - \mathbf{p}) + \mathbf{X}\boldsymbol{\alpha}^{(n-1)}$. This is the solution to the unpenalized parametric logistic regression and equivalent to:

$$\underset{\boldsymbol{\beta}\in\mathcal{R}^p}{\text{argmin}} \left\| \mathbf{P}^{\frac{1}{2}}\mathbf{m}^{(n-1)} - \mathbf{P}^{\frac{1}{2}}\mathbf{X}\boldsymbol{\beta} \right\|_2^2.$$

Combining the lasso penalty at each stage of the (N-R) algorithm we get each stage of the IRLS solution for the lasso penalized logistic regression as:

$$\boldsymbol{\beta}^{(n)} = \underset{\boldsymbol{\beta}\in\mathcal{R}^p}{\operatorname{argmin}} \left\| \mathbf{P}^{\frac{1}{2}}\mathbf{m}^{(n-1)} - \mathbf{P}^{\frac{1}{2}}\mathbf{X}\boldsymbol{\beta} \right\|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|_1 \tag{5.2.11}$$

which is simply the lasso solution to the standard ordinary least squares problem with design matrix $\mathbf{P}^{\frac{1}{2}}\mathbf{X}$ and outcome vector $\mathbf{P}^{\frac{1}{2}}\mathbf{m}^{(n-1)}$.

5.2.4 Elastic Net

The elastic net [59] (or more precisely the naive elastic net) aims to regularize ordinary least squares using the following regularization criteria:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathcal{R}^p}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_2 \|\boldsymbol{\beta}\|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|_1 \tag{5.2.12}$$

$$\Longleftrightarrow$$

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathcal{R}^p}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2, \text{ subject to} (1-\alpha)\|\boldsymbol{\beta}\|_1 + \alpha\|\boldsymbol{\beta}\|_2^2 \leq t \text{ for some } t, \alpha \in \mathcal{R}$$

In the elastic net paper, *Lemma* 1 provides a way to solve the elastic net problem by turning it into a simple lasso problem. We will make use of the idea they present in *Lemma* 1 which says that when you augment your design matrix $\mathbf{X}^\star{}_{(n+p)\times p} = (1+\lambda_2)^{\frac{-1}{2}} \begin{bmatrix} \mathbf{X} \\ \sqrt{\lambda_2}\mathbf{I} \end{bmatrix}$, and outcome vector $\mathbf{y}^\star = \begin{bmatrix} \mathbf{y} \\ \mathbf{0} \end{bmatrix}$, then the naive elastic net can be written as

$$\underset{\boldsymbol{\beta}^\star\in\mathcal{R}^p}{\operatorname{argmin}} \|\mathbf{y}^\star - \mathbf{X}^\star\boldsymbol{\beta}^\star\| + \frac{\lambda_1}{\sqrt{1+\lambda_2}} \|\boldsymbol{\beta}^\star\|_1$$

where the solution to the original elastic net problem (5.2.12) is $\hat{\boldsymbol{\beta}} = \frac{1}{1+\lambda_2}\hat{\boldsymbol{\beta}}^\star$.

5.3 KLR with Import Selection

Combining the four ideas listed in the previous section we will propose a way of reducing the dimensions of the columns of the matrix, K in the KLR problem. First, some notation. We wish to find a subset $\mathcal{S}$ of the data $\{\mathbf{z}_1, \cdots, \mathbf{z}_n\}$ such that $\tilde{h}(\mathbf{z}) = \sum_{z_i \in \mathcal{S}} \tilde{\alpha}_i k(\mathbf{z}, \mathbf{z}_i) \approx \sum_{i=1}^{n} \alpha_i k(\mathbf{z}, \mathbf{z}_i) = \hat{h}(\mathbf{z})$. Zhu and Hastie (2005) [58] proposed the Import Vector Machine (IVM) where they use a greedy algorithm to choose which subset, $\mathcal{S}$, should be used. A desirable goal, once S is determined, is that $\tilde{h}$ satisfies

$$\tilde{h} = \underset{h \in span(\{K(\cdot, \mathbf{z}_i)\}; \mathbf{z}_i \in S)}{\operatorname{argmin}} \sum_{i=1}^{n} \left(-y_i h(\mathbf{z}_i) + ln\left(1 + \exp(h(\mathbf{z}_i))\right)\right) + \frac{\lambda_1}{2} \left\|h\right\|_{\mathcal{H}_\mathcal{K}}^2 .$$

We propose the following loss function (minimization of the loss function) with lasso penalty to accomplish this goal:

$$\min_{\boldsymbol{\alpha} \in \mathcal{R}^n} L(\boldsymbol{\alpha}) = -\mathbf{y}^\top \mathbf{K}\boldsymbol{\alpha} + \mathbf{1}^\top ln\left(1 + exp(\mathbf{K}\boldsymbol{\alpha})\right) + \frac{\lambda_1}{2} \boldsymbol{\alpha}^\top \mathbf{K}\boldsymbol{\alpha} + \lambda_2 \left\|\boldsymbol{\alpha}\right\|_1 . \qquad (5.3.1)$$

Before describing the algorithm to solve (5.3.1), lets discuss the objective of the penalty term $\lambda_2 \left\|\boldsymbol{\alpha}\right\|_1$. The goal of lasso penalty is to impose sparsity on the $\boldsymbol{\alpha}$ vector. Let $\hat{\boldsymbol{\alpha}} = \begin{bmatrix} \tilde{\boldsymbol{\alpha}} \\ \mathbf{0} \end{bmatrix}$ be the solution to (5.3.1) where we ordered the elements so that $\tilde{\boldsymbol{\alpha}}$ are all of the non-zero elements first. We can similarity order the columns of the matrix $\mathbf{K}$. Let $\hat{\mathcal{S}} \subset \{\mathbf{z}_1, \cdots, \mathbf{z}_n\}$ where the indices of $\mathbf{z}$ that are in the set $\hat{\mathcal{S}}$ correspond to the indices of all non-zero $\boldsymbol{\alpha}$ (the $\tilde{\boldsymbol{\alpha}}$ vector). We can order $\mathbf{K}$ to correspond to the non-zero elements of $\boldsymbol{\alpha}$ first and then the remaining columns correspond to the zero elements of $\hat{\boldsymbol{\alpha}}$ so $\mathbf{K} = [\tilde{\mathbf{K}} \star]$. We have the following equality that we state as a lemma:

Lemma 8.

$$L(\hat{\boldsymbol{\alpha}}) = -\mathbf{y}^\top \tilde{\mathbf{K}}\tilde{\boldsymbol{\alpha}} + \mathbf{1}^\top ln\left(1 + exp(\tilde{\mathbf{K}}\tilde{\boldsymbol{\alpha}})\right) + \frac{\lambda_1}{2}\tilde{\boldsymbol{\alpha}}^\top \tilde{\mathbf{K}}_{(card(\tilde{S})\times card(\tilde{S}))}\tilde{\boldsymbol{\alpha}} + \lambda_2 \left\|\tilde{\boldsymbol{\alpha}}\right\|_1$$
$$= \min_{\boldsymbol{\alpha}\in\mathcal{R}^{card(\hat{S})}} \tilde{L}(\alpha) = -\mathbf{y}^\top \tilde{\mathbf{K}}\boldsymbol{\alpha} + \mathbf{1}^\top ln\left(1 + exp(\tilde{\mathbf{K}}\boldsymbol{\alpha})\right) + \frac{\lambda_1}{2}\boldsymbol{\alpha}^\top \tilde{\mathbf{K}}_{(card(\tilde{S})\times card(\tilde{S}))}\boldsymbol{\alpha} + \lambda_2 \left\|\boldsymbol{\alpha}\right\|_1 .$$

Proof. $\min_{\boldsymbol{\alpha}\in\mathcal{R}^{card(\hat{S})}} \tilde{L}(\boldsymbol{\alpha}) \leq \tilde{L}(\tilde{\boldsymbol{\alpha}}) = L(\hat{\boldsymbol{\alpha}})$ since $\tilde{\boldsymbol{\alpha}} \in \mathcal{R}^{card(\hat{S})}$. It suffices to show that $\tilde{\boldsymbol{\alpha}} = \text{argmin}_{\boldsymbol{\alpha}\in\mathcal{R}^{card(\hat{S})}} \tilde{L}(\boldsymbol{\alpha})$. Let $\boldsymbol{\alpha}^\star = \text{argmin}_{\boldsymbol{\alpha}\in\mathcal{R}^{card(\hat{S})}} \tilde{L}(\boldsymbol{\alpha})$. Then we have the following inequalities:

$$L(\hat{\boldsymbol{\alpha}}) \leq L\left(\begin{bmatrix}\boldsymbol{\alpha}^\star \\ \\ \mathbf{0}\end{bmatrix}\right) = \tilde{L}(\boldsymbol{\alpha}^\star) = \min_{\boldsymbol{\alpha}\in\mathcal{R}^{card(\hat{S})}} \tilde{L}(\boldsymbol{\alpha}).$$

□

We set $\tilde{h}(\mathbf{z}) = \sum_{z_i\in\hat{\mathcal{S}}} \tilde{\alpha}_i K(\mathbf{z}, \mathbf{z}_i) \implies \tilde{h} \in span\left(\{K(\cdot, \mathbf{z}_i\}; \mathbf{z}_i \in \hat{S}\right)$. From Lemma 8 and the representer theorem, it is easy to show that we have accomplished the goal of:

$$\tilde{h} = \underset{h\in span\left(\{k(\cdot,\mathbf{z}_i\};\mathbf{z}_i\in\hat{S}\right)}{\text{argmin}} \sum_{i=1}^{n} \left(-y_i h(\mathbf{z}_i) + ln\left(1 + \exp(h(\mathbf{z}_i))\right)\right) + \frac{\lambda_1}{2}\left\|h\right\|^2_{\mathcal{H}_\mathcal{K}} + \lambda_2 \left\|\boldsymbol{K}^{-1}\mathbf{h}\right\|_1 ,$$

where the $n \times 1$ vector, $\mathbf{h}$ has ith entry as $h(\mathbf{z}_i)$ (as a reminder, $\boldsymbol{\alpha} = \boldsymbol{K}^{-1}\mathbf{h}$). This is not quite identical to the original goal of:

$$\tilde{h} = \underset{h\in span\left(\{k(\cdot,\mathbf{z}_i\};\mathbf{z}_i\in\hat{S}\right)}{\text{argmin}} \sum_{i=1}^{n} \left(-y_i h(\mathbf{z}_i) + ln\left(1 + \exp(h(\mathbf{z}_i))\right)\right) + \frac{\lambda_1}{2}\left\|h\right\|^2_{\mathcal{H}_\mathcal{K}} .$$

To accomplish the above goal we can first estimate the set $\hat{S}$ using the algorithm we describe below to solve (5.3.1) and then subsequently use the (N-R) algorithm in (5.2.5) to solve for $\boldsymbol{\alpha} \in \mathcal{R}^{card(\hat{S})}$. This procedure would be similar in idea to running OLS after running a Lasso regression using the features selected by the Lasso

estimator which is, in fact, advocated in [4].

To solve (5.3.1) we notice that without the penalty, we have the (N-R) algorithm (5.2.5) which corresponds to the Tikhonov regularization problem (5.2.7). Adding on the L1 penalty we get the following Newton-Raphson like estimation procedure for $\boldsymbol{\alpha}$:

$$\boldsymbol{\alpha}^n = \underset{\boldsymbol{\alpha}\in\mathcal{R}^n}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\alpha}\|_2^2 + \lambda_1 \|\mathbf{T}\boldsymbol{\alpha}\|_2^2 + \lambda_2 \|\boldsymbol{\alpha}\|_1 . \tag{5.3.2}$$

where similar to before we have $\mathbf{T} = \mathbf{K}^{\frac{1}{2}}$, $\mathbf{X} = \mathbf{P}^{\frac{1}{2}}\mathbf{K}$ and $\mathbf{y} = \mathbf{P}^{\frac{1}{2}}\mathbf{m}^{(n-1)}$. Following along the same lines as *lemma* 1 in the elastic net paper we define $\mathbf{X^\star}_{2n\times n} = \begin{bmatrix} \mathbf{X} \\ \sqrt{\lambda_1}\mathbf{T} \end{bmatrix}, \mathbf{y}^\star = \begin{bmatrix} \mathbf{y} \\ \mathbf{0} \end{bmatrix}$ then (5.3.2) is equivalent to

$$\boldsymbol{\alpha}^n = \underset{\boldsymbol{\alpha}\in\mathcal{R}^n}{\operatorname{argmin}} \|\mathbf{y}^\star - \mathbf{X}^\star\boldsymbol{\alpha}\|_2^2 + \lambda_2 \|\boldsymbol{\alpha}\|_1 \tag{5.3.3}$$

which is the lasso estimator. Thus our proposed algorithm to solving (5.3.1) for a functional covariate is as follows:

Algorithm Import Selection

(i) Step1.1: Perform FPCA (e.g. R package `fdapace`) to extract the functional component scores for the functional predictor and store it in a vector for each individual subject $\mathbf{z}_i$;

(ii) Step 1.2: Set up a grid of possible tuning parameters for (λ_1, λ_2) and initialize $\boldsymbol{\alpha}^{(0)}$ to be a vector of ones. Perform steps Steps 2-3 below.

(iii) Step 2: At the n-th step in the algorithm, solve for $\boldsymbol{\alpha}^{(n)}$ equation (5.3.2) which is equivalent to running the lasso solution (e.g. R package `oem` or `glmnet`). Stop at convergence.

(iv) Step 3: Perform cross-validation over the specified grid of (λ_1, λ_2) to determine

the final $\boldsymbol{\alpha}$

As described in the elastic net paper [59], you can pick a small grid of values for λ_1, $(0.01, 0.1, 1, 10, 100)$ to help ease the computationally burden.

5.4 MFRS Logistic Regression

In this section we will outline how to apply the MFRS framework when the dependent variable of interest is binary. The goal, is to fit the KLR problem while simultaneously performing functional selection (via the FPC scores). We will use the same notation as defined in Chapter II. For the purpose of this discussion we will consider a fully non-parametric model (not a partially linear one) and assume that λ_1 and λ_2 are fixed. We look to minimize the following objective function:

$$\min_{h\in\mathcal{H}_\mathcal{K},\boldsymbol{\gamma}\in\mathcal{R}^s} \sum_{i=1}^{n} \left(-y_i h(\boldsymbol{\gamma}\circ\mathbf{z}_i) + ln\left(1+\exp(h(\boldsymbol{\gamma}\circ\mathbf{z}_i))\right)\right) + \frac{\lambda_1}{2}\left\|h\right\|^2_{\mathcal{H}_\mathcal{K}} + \lambda_2\left\|\boldsymbol{\gamma}\right\|_1 \iff$$
$$\min_{\boldsymbol{\alpha}\in\mathcal{R}^n,\boldsymbol{\gamma}\in\mathcal{R}^s} -\mathbf{y}^\top\mathbf{K}(\boldsymbol{\gamma};\mathbf{Z})\boldsymbol{\alpha} + \mathbf{1}^\top ln\left(1+exp(\mathbf{K}(\boldsymbol{\gamma};\mathbf{Z})\boldsymbol{\alpha})\right) + \frac{\lambda_1}{2}\boldsymbol{\alpha}^\top\mathbf{K}(\boldsymbol{\gamma};\mathbf{Z})\boldsymbol{\alpha} + \lambda_2\left\|\boldsymbol{\gamma}\right\|_1. \tag{5.4.1}$$

Given $\boldsymbol{\gamma}$, minimizing (5.4.1) with respect to $\boldsymbol{\alpha}$ reduces to the KLR problem with FPC scores, $\mathbf{z}_i \mapsto \boldsymbol{\gamma}\circ\mathbf{z}_i$. Given $\boldsymbol{\alpha}$ we aim to minimize (5.4.1) with respect to $\boldsymbol{\gamma}$ which is of the form

$$\min_{\boldsymbol{\gamma}} L(\boldsymbol{\gamma}) + \lambda_2\left\|\boldsymbol{\gamma}\right\|_1$$

a continuously differentiable (but non-convex) function plus a convex loss. In [43] the authors review current algorithms that solve the above problem and propose two new approaches as well. We will adopt a coordinate sub-gradient descent algorithm as described in [42]. Adopting the same notation let $\boldsymbol{\gamma}^j$ be the estimate of $\boldsymbol{\gamma}$ in the jth iteration and define $\boldsymbol{\gamma}^{(j,j-1,\gamma_k)} := \left(\gamma_1^{(j)},\cdots,\gamma_{k-1}^{(j)},\gamma_k,\gamma_{k+1}^{(j-1)},\cdots,\gamma_s^{(j-1)}\right)^\top$ where γ_q^j is the estimate of γ_q in the jth iteration and let $\boldsymbol{\gamma}^{(j,j-1;k)} := \left(\gamma_1^{(j)},\cdots,\gamma_{k-1}^{(j)},\gamma_k^{j-1},\gamma_{k+1}^{(j-1)},\cdots,\gamma_s^{(j-1)}\right)^\top$.

We use the following algorithm which is described in detail in [42] with slight modifications.

Gamma Step Algorithm:

(i) Step 0: Repeat for $j = 1, \ldots,$ until convergence (which can be defined as $\left\|\boldsymbol{\gamma}^{(j)} - \boldsymbol{\gamma}^{(j-1)}\right\|_2^2 \leq \epsilon_1$ or $|L(\boldsymbol{\gamma}^{(j)}) + \lambda_2 \left\|\boldsymbol{\gamma}^{(j)}\right\|_1 - L(\boldsymbol{\gamma}^{(j-1)}) + \lambda_2 \left\|\boldsymbol{\gamma}^{(j-1)}\right\|_1| \leq \epsilon_2$ for predefined $\epsilon_1, \epsilon_2 \geq 0$) Steps 1-4

(ii) Step 1: For $k = 1, \ldots, s$ in the jth step of the algorithm perform steps 2-4

(iii) Step 2: Calculate $g_k^{(j)} := \frac{\partial}{\partial \gamma_k} L(\boldsymbol{\gamma}^{(j,j-1,\gamma_k)})\big|_{\gamma_k = \gamma_k^{j-1}}$, $L_k^{(j)} := L(\boldsymbol{\gamma}^{(j,j-1;\gamma_k)})$

(iv) Step 3: Calculate the descent direction $d_k^{(j)}$ where $d_k^{(j)} = \text{argmin}_{d \in \mathcal{R}}\, L_k^{(j)} + g_k^{(j)} d + \lambda_2 \left\|\boldsymbol{\gamma}^{(j,j-1;k)} + d\mathbf{e}_k\right\|_1$ where $\mathbf{e}_k$ is the standard unit vector.

(v) Step 4: Perform an inexact line search algorithm with backtracking that satisfies the Armijo−Goldstein condition; to find $\alpha_k^{(j)} \geq 0$ where we then set $\boldsymbol{\gamma}^{(j,j-1;k+1)} = \boldsymbol{\gamma}^{(j,j-1;k)} + \alpha_k^{(j)} d_k^j \mathbf{e}_k$

To solve for $d_k^{(j)}$ in step 3 we take the sub-gradient of $L_k^{(j)} + g_k^{(j)} d + \lambda_2 \left\|\boldsymbol{\gamma}^{(j,j-1;k)} + d\right\|_1$ (with respect to d).

$$\partial\left(L_k^{(j)} + g_k^{(j)} d + \lambda_2 \left\|\boldsymbol{\gamma}^{(j,j-1;k)} + d\mathbf{e}_k\right\|_1\right) =$$
$$\begin{cases} g_k^{(j)} + \lambda_2, \text{if } d > -\gamma^{(j,j-1;k)} \\ g_k^{(j)} - \lambda_2, \text{if } d < -\gamma^{(j,j-1;k)} \\ g_k^{(j)} + \lambda_2 \left[-1, 1\right], \text{if } d = -\gamma^{(j,j-1;k)}. \end{cases}$$

Since $0 \in \partial\left(L_k^{(j)} + g_k^{(j)}d + \lambda_2 \left\|\boldsymbol{\gamma}^{(j,j-1;k)} + d\mathbf{e}_k\right\|_1\right)$ for optimality to be satisfied we get:

$$d_k^{(j)} = \begin{cases} -\lambda_2 - g_k^{(j)}, \text{if } -(\lambda_2 + g_k^{(j)}) > -\gamma^{(j,j-1;k)} \\ \lambda_2 - g_k^{(j)}, \text{if } \lambda_2 - g_k^{(j)} < -\gamma^{(j,j-1;k)} \\ -\gamma^{(j,j-1;k)}, \text{if } 0 \in \left[-\lambda_2 + g_k^{(j)}, \lambda_2 + g_k^{(j)}\right]. \end{cases}$$

which is equivalent to setting $d_k^{(j)} = median(-\lambda_2 - g_k^{(j)}, \lambda_2 - g_k^{(j)}, -\gamma^{(j,j-1;k)})$.

To solve for $\alpha_k^{(j)}$ in Step 4, we can start with an initial $\alpha_{int} = 1$ and set $\alpha_k^{(j)}$ as $\max_{i=0,1...}\{0.5^i\alpha_{int}\}$ where the following criteria is satisfied (this is equivalent to the Armijo−Goldstein conditions):

$$L(\boldsymbol{\gamma}^{(j,j-1;k)} + \alpha_k^{(j)}d_k^{(j)}\mathbf{e}_k) + \lambda_2 \left\|\boldsymbol{\gamma}^{(j,j-1;k)} + \alpha_k^{(j)}d_k^{(j)}\mathbf{e}_k\right\|_1 \leq$$
$$L(\boldsymbol{\gamma}^{(j,j-1;k)}) + \lambda_2 \left\|\boldsymbol{\gamma}^{(j,j-1;k)}\right\|_1 + 0.1\alpha_k^j \left(g_k^j d_k^{(j)} + \lambda_2 \left\|\boldsymbol{\gamma}^{(j,j-1;k)} + d_k^{(j)}\mathbf{e}_k\right\|_1 - \lambda_2 \left\|\boldsymbol{\gamma}^{(j,j-1;k)}\right\|_1\right).$$

where we used the same choice of constants $(0.1, 0.5)$ as mentioned in [42]. There are other options that can be used to solve the inexact line search. The Wolf conditions [52] are an alternative to the Armijo−Goldstein condition. An exact line search might also be possible, however, we wanted to keep this algorithm as conceptually simple as possible.

We now arrive at the MFRS-Logistic algorithm to solve (5.4.1):

MFRS-Logistic Algorithm

(i) Step 1.1: Perform FPCA (e.g. R package `fdapace`) to extract the functional component scores for the functional predictor and store it in a vector for each individual subject $\mathbf{z}_i$;

(ii) Step 1.2: Set up a grid of possible tuning parameters for (λ_1, λ_2) and initialize $\hat{\boldsymbol{\gamma}}$ to be a vector of ones and $\hat{\boldsymbol{\alpha}}$ to be a vector of ones. Perform steps Steps 2-4 below.

(iii) Step 2: Perform step 2 in the KLR algorithm with transformed scores $\hat{\boldsymbol{\gamma}} \circ \mathbf{z}_i$ to get an updated estimate of $\hat{\boldsymbol{\alpha}}$.

(iv) Step 3: Perform the Gamma Step Algorithm with the current estimates of $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\alpha}}$ to get an updated estimate of $\hat{\boldsymbol{\gamma}}$.

(v) Step 4: Repeat steps 2-3 until convergence.

(vi) Step 5: Perform cross-validation over the specified grid of (λ_1, λ_2) to determine the final estimate of $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})$.

CHAPTER VI

Summary and Future Work

In this dissertation we have proposed a novel framework, MFRS, to simultaneously perform functional variable selection and model fit in the presence of scalar counfounders. The framework was tested both via simulation and using real-world accelerometer data. Before applying the MFRS framework to the accelerometer data, there were many issues that came up on how to properly pre-process the accelerometer data and handle multiple days of accelerometer wear. We provided several techniques on how to deal with both issues. While researching how to extend the MFRS to the case of a binary outcome we discovered a possible solution to the import selection problem.

The contribution of the MFRS framework in Chapter II is an important addition to the literature on both non-linear and non-additive modeling and on functional variable selection. Currently, the literature does not discuss this relationship. There is still future work to be done with the MFRS algorithm. The main focus of the MFRS algorithm was for a continuous outcome for a cross-sectional study. As mentioned at the end of Chapter II, the MFRS can be extended to the generalized linear mixed model (GLMM) setting. This is a very natural setting for mobile health devices. Through constant wear of a mobile health device and subsequent visits to the doctor, various health outcomes not accounted for in the mobile devices are repeated for

the individual. It is precisely this setting that future work needs to be devoted too and that this dissertation scratched the surface of work. As mentioned in Chapter II and III, it is not clear how to handle multiple days of accelerometer wear. As society moves toward individuals monitoring their health activity via mobile devices, it becomes important to ascertain when to detect change points in the corresponding time series and associate the important functional covariates with the correct health outcome. For example, if someone wears an accelerometer constantly and visits the doctor every 6 months for a checkup, it is important to identity which device data or which functional should be associated with the visit? With non-functional data, the timing of the measurements usually match between the repeated health outcomes and covariates of interest. This is indeed a deep question. In a similar vein, while the MFRS algorithm assumes that the functional is well defined for modeling, in practice, as we found with the accelerometer data this is not the case. It is not clear how to properly match up individuals who are wearing the accelerometer over multiple days. This is because the activity of the individuals do not necessarily correspond with one another. The days starts and ends for one individual at different times, so if we want to start and end the activity functional when someone wakes up and goes to sleep, we might find that one functional is defined on a different time domain (i.e. one individual might be awake for more time then another).

We hope that this dissertation will inspire future work in the direction of mobile health data and extend the contributions that we have made towards that goal

APPENDICES

APPENDIX A

Proofs and additional Tables from Chapters 2

A.1 Technical assumptions and proofs

A.1.1 Proof of Theorem 5

By Lemma 8.4 on page 129 in [17] assumptions 1,2 and 3 imply

$$P(\sup_{b\in\mathcal{B}} \frac{\frac{1}{\sqrt{n}}|\sum_{i=1}^{n}\epsilon_i b(\mathbf{z}_i)|}{\|b\|_{P_n}^{1-\psi}} \geq T) \leq c\exp(-\frac{T^2}{c^2}), \tag{A.1}$$

where the constant c is dependent on C_1, C_2, C_3, C_4, and ψ. (A.1) holds for all $T \geq c$. This implies that,

$$\sup_{b\in\mathcal{B}} \frac{\frac{1}{\sqrt{n}}|\sum_{i=1}^{n}\epsilon_i b(\mathbf{z}_i)|}{\|b\|_{P_n}^{1-\psi}} = O_p(1). \tag{A.2}$$

Therefore, for any $h \in \mathcal{H}_\mathcal{K}$ and $\Gamma \in \mathcal{A}$ we get

$$\frac{\sqrt{n}(\epsilon, h\circ\Gamma - h_0\circ\Gamma_0)_n \left(\|h\|_{\mathcal{H}_\mathcal{K}}^2 + \|h_0\|_{\mathcal{H}_\mathcal{K}}^2 + \|\Gamma\|_{SGL}^2 + \|\Gamma_0\|_{SGL}^2\right)^{-\psi}}{\|h\circ\Gamma - h_0\circ\Gamma_0\|_{P_n}^{1-\psi}} = O_p(1). \tag{A.3}$$

For our estimator, $\hat{h}$ and $\hat{\Gamma}$ we then have

$$\begin{aligned}
&(\epsilon, \hat{h} \circ \hat{\Gamma} - h_0 \circ \Gamma_0)_n = \\
&O_p(n^{-\frac{1}{2}}) \left\| \hat{h} \circ \hat{\Gamma} - h_0 \circ \Gamma_0 \right\|_n^{1-\psi} \left(\left\| \hat{h} \right\|_{\mathcal{H}_\mathcal{K}}^2 + \|h_0\|_{\mathcal{H}_\mathcal{K}}^2 + \left\| \hat{\Gamma} \right\|_{SGL}^2 + \|\Gamma_0\|_{SGL}^2 \right)^{\psi}.
\end{aligned} \tag{A.4}$$

From Lemma 3 and (A.4) we get the following inequality:

$$\begin{aligned}
&\left\| \hat{h} \circ \hat{\Gamma} - h_0 \circ \Gamma_0 \right\|_n^2 + \lambda_1 \left\| (\hat{h}) \right\|_{\mathcal{H}_\mathcal{K}}^2 + \lambda_2 \left\| (\hat{\Gamma}) \right\|_{SGL}^2 \leq \\
&O_p(n^{-\frac{1}{2}}) \left\| \hat{h} \circ \hat{\Gamma} - h_0 \circ \Gamma_0 \right\|_n^{1-\psi} \left(\left\| \hat{h} \right\|_{\mathcal{H}_\mathcal{K}}^2 + \|h_0\|_{\mathcal{H}_\mathcal{K}}^2 + \left\| \hat{\Gamma} \right\|_{SGL}^2 + \|\Gamma_0\|_{SGL}^2 \right)^{\psi} \\
&+ \lambda_1 \|h_0\|_{\mathcal{H}_\mathcal{K}}^2 + \lambda_2 \|\Gamma_0\|_{SGL}^2 .
\end{aligned} \tag{A.5}$$

We need $\lambda_1 = O_p(1)\lambda_2$ which implies that λ_2 and λ_1 go to zero at the same rate. We will show at the end of the proof what happens if they are not of the same order. Therefore, without loss of generality, assume $\lambda_1 = \lambda_2$. We will call it λ. We can divide (A.5) into two cases.

Case 1: If

$$\begin{aligned}
O_p(n^{-\frac{1}{2}}) \left\| \hat{h} \circ \hat{\Gamma} - h_0 \circ \Gamma_0 \right\|_n^{1-\psi} \left(\left\| \hat{h} \right\|_{\mathcal{H}_\mathcal{K}}^2 + \|h_0\|_{\mathcal{H}_\mathcal{K}}^2 + \left\| \hat{\Gamma} \right\|_{SGL}^2 + \|\Gamma_0\|_{SGL}^2 \right)^{\psi} \\
\geq \lambda \left(\|h_0\|_{\mathcal{H}_\mathcal{K}}^2 + \|\Gamma_0\|_{SGL}^2 \right)
\end{aligned}$$

we have

$$\begin{aligned}
&\left\| \hat{h} \circ \hat{\Gamma} - h_0 \circ \Gamma_0 \right\|_n^2 + \lambda \left(\left\| \hat{h} \right\|_{\mathcal{H}_\mathcal{K}}^2 + \left\| \hat{\Gamma} \right\|_{SGL}^2 \right) \leq \\
&O_p(n^{-\frac{1}{2}}) \left\| \hat{h} \circ \hat{\Gamma} - h_0 \circ \Gamma_0 \right\|_n^{1-\psi} \left(\left\| \hat{h} \right\|_{\mathcal{H}_\mathcal{K}}^2 + \|h_0\|_{\mathcal{H}_\mathcal{K}}^2 + \left\| \hat{\Gamma} \right\|_{SGL}^2 + \|\Gamma_0\|_{SGL}^2 \right)^{\psi}.
\end{aligned} \tag{A.6}$$

We have two subdivided cases to consider for (A.6):
Case 1a: If $\|h_0\|^2_{\mathcal{H}_\mathcal{K}} + \|\Gamma_0\|^2_{SGL} \leq \left\|\hat{h}\right\|^2_{\mathcal{H}_\mathcal{K}} + \left\|\hat{\Gamma}\right\|^2_{SGL}$ then

$$\begin{aligned} \left\|\hat{h}\circ\hat{\Gamma} - h_0\circ\Gamma_0\right\|_n^2 + \lambda\left(\left\|\hat{h}\right\|^2_{\mathcal{H}_\mathcal{K}} + \left\|\hat{\Gamma}\right\|^2_{SGL}\right) \leq \\ O_p(n^{-\frac{1}{2}})\left\|\hat{h}\circ\hat{\Gamma} - h_0\circ\Gamma_0\right\|_n^{1-\psi}\left(\left\|\hat{h}\right\|^2_{\mathcal{H}_\mathcal{K}} + \left\|\hat{\Gamma}\right\|^2_{SGL}\right)^{\psi}. \end{aligned} \tag{A.7}$$

Therefore,

$$\left(\left\|\hat{h}\right\|^2_{H_K} + \left\|\hat{\Gamma}\right\|^2_{SGL}\right)^{\psi} \leq O_p(n^{-\frac{\psi}{2(1-\psi)}})\left\|\hat{h}\circ\hat{\Gamma} - h_0\circ\Gamma_0\right\|_n^{\psi}\lambda^{-\frac{\psi}{1-\psi}} \tag{A.8}$$

and we get

$$\begin{aligned} \left\|\hat{h}\circ\hat{\Gamma} - h_0\circ\Gamma_0\right\|_n &= O_p(n^{-\frac{1}{2(1-\psi)}})O_p(\lambda^{-\frac{\psi}{1-\psi}}) \\ \left\|\hat{h}\right\|^2_{H_K} + \left\|\hat{\Gamma}\right\|^2_{SGL} &= O_p(n^{-\frac{1}{1-\psi}})O_p(\lambda^{-\frac{1+\psi}{1-\psi}}) \end{aligned} \tag{A.9}$$

Case 1b: If $\|h_0\|^2_{\mathcal{H}_\mathcal{K}} + \|\Gamma_0\|^2_{SGL} \geq \left\|\hat{h}\right\|^2_{H_K} + \left\|\hat{\Gamma}\right\|^2_{SGL}$ then
$\left\|\hat{h}\right\|^2_{\mathcal{H}_\mathcal{K}} + \left\|\hat{\Gamma}\right\|^2_{SGL} = O_p(\|h_0\|^2_{\mathcal{H}_\mathcal{K}} + \|\Gamma_0\|^2_{SGL})O_p(1)$.
Therefore,

$$\left\|\hat{h}\circ\hat{\Gamma} - h_0\circ\Gamma_0\right\|_n = O_p(n^{-\frac{1}{2(1+\psi)}})\left(\|h_0\|^2_{\mathcal{H}_\mathcal{K}} + \|\Gamma\|^2_{SGL}]\right)^{\frac{\psi}{1+\psi}},$$

and we get

$$\begin{aligned} \left\|\hat{h}\circ\hat{\Gamma} - h_0\circ\Gamma_0\right\|_n &= O_p(n^{-\frac{1}{2(1-\psi)}})O_p(\lambda^{-\frac{\psi}{1-\psi}}), \\ \left\|\hat{h}\right\|^2_{\mathcal{H}_\mathcal{K}} + \left\|\hat{\Gamma}\right\|^2_{SGL} &= O_p(n^{-\frac{1}{1-\psi}})O_p(\lambda^{-\frac{1+\psi}{1-\psi}}). \end{aligned} \tag{A.10}$$

These are the same rates as (A.9).

Case 2: If

$$O_p(n^{-\frac{1}{2}}) \left\| \hat{h} \circ \hat{\Gamma} - h_0 \circ \Gamma_0 \right\|_n^{1-\psi} \left(\left\| \hat{h} \right\|_{\mathcal{H}_\mathcal{K}}^2 + \|h_0\|_{\mathcal{H}_\mathcal{K}}^2 + \left\| \hat{\Gamma} \right\|_{SGL}^2 + \|\Gamma_0\|_{SGL}^2 \right)^{\psi} \leq \lambda(\|h_0\|_{\mathcal{H}_\mathcal{K}}^2 + \|\Gamma_0\|_{SGL}^2),$$

then,

$$\left\| \hat{h} \circ \hat{\Gamma} - h_0 \circ \Gamma_0 \right\|_n^2 + \lambda \left(\left\| \hat{h} \right\|_{\mathcal{H}_\mathcal{K}}^2 + \left\| \hat{\Gamma} \right\|_{SGL}^2 \right) \leq 2\lambda \left(\|h_0\|_{\mathcal{H}_\mathcal{K}}^2 + \|\Gamma_0\|_{SGL}^2 \right).$$

This implies that

$$\begin{aligned} \left\| \hat{h} \circ \hat{\Gamma} - h_0 \circ \Gamma_0 \right\|_n &= O_p(\lambda^{\frac{1}{2}}) \left(\|h_0\|_{\mathcal{H}_\mathcal{K}}^2 + \|\Gamma_0\|_{SGL}^2 \right)^{\frac{1}{2}}, \\ \left\| \hat{h} \right\|_{\mathcal{H}_\mathcal{K}}^2 + \left\| \hat{\Gamma} \right\|_{SGL}^2 &= O_p(1) \left(\|h_0\|_{\mathcal{H}_\mathcal{K}}^2 + \|\Gamma_0\|_{SGL}^2 \right). \end{aligned} \tag{A.11}$$

To make (A.11) and (A.9) the same rates we want to equate the terms $O_p(\lambda^{\frac{1}{2}}) \left(\|h\|_{\mathcal{H}_\mathcal{K}}^2 + \|\Gamma\|_{SGL}^2 \right)^{\frac{1}{2}}$ with $O_p(n^{-\frac{1}{2(1-\psi)}}) O_p(\lambda^{-\frac{\psi}{1-\psi}})$ and solve for a common λ. So for

$$\lambda^{-1} = n^{\frac{1}{1+\psi}} \left(\|h\|_{\mathcal{H}_\mathcal{K}}^2 + \|\Gamma\|_{SGL}^2 \right)^{\frac{1-\psi}{1+\psi}}$$

we get that (A.9), (A.10), (A.11) are

$$\left\| \hat{h} \circ \hat{\Gamma} - h_0 \circ \Gamma_0 \right\|_n = O_p(n^{-\frac{1}{2(1+\psi)}}) \left(\|h_0\|_{\mathcal{H}_\mathcal{K}}^2 + \|\Gamma_0\|_{SGL}^2 \right)^{\frac{\psi}{1+\psi}}, \tag{A.12}$$

$$\left\| \hat{h} \right\|_{\mathcal{H}_\mathcal{K}}^2 + \left\| \hat{\Gamma} \right\|_{SGL}^2 = O_p(1) \left(\|h_0\|_{\mathcal{H}_\mathcal{K}}^2 + \|\Gamma_0\|_{SGL}^2 \right) \tag{A.13}$$

which completes the proof.

If we try to separate out λ_1 and λ_2 we would run into the following issue. Taking Case 2 as an example we see:

Case 2: If

$$O_p(n^{-\frac{1}{2}}) \left\| \hat{h} \circ \hat{\Gamma} - h_0 \circ \Gamma_0 \right\|_n^{1-\psi} \left(\left\| \hat{h} \right\|_{\mathcal{H}_\mathcal{K}}^2 + \|h_0\|_{\mathcal{H}_\mathcal{K}}^2 + \left\| \hat{\Gamma} \right\|_{SGL}^2 + \|\Gamma_0\|_{SGL}^2 \right)^{\psi} \leq \lambda_1 \|h_0\|_{\mathcal{H}_\mathcal{K}}^2 + \lambda_2 \|\Gamma_0\|_{SGL}^2 ,$$

we now need to subdivide this into two cases:

Case 2a: If $\lambda_1 \|h_0\|_{\mathcal{H}_\mathcal{K}}^2 \leq \lambda_2 \|\Gamma_0\|_{SGL}^2$ then, following the same logic as before:

$$\begin{aligned} \left\| \hat{h} \circ \hat{\Gamma} - h_0 \circ \Gamma_0 \right\|_n &= O_p(\lambda_2^{\frac{1}{2}}) \|\Gamma_0\|_{SGL}), \\ \left\| \hat{h} \right\|_{\mathcal{H}_\mathcal{K}}^2 &= O_p(\frac{\lambda_2}{\lambda_1}) \|\Gamma_0\|_{SGL}^2 , \\ \left\| \hat{\Gamma} \right\|_{SGL}^2 &= O_p(1) \|\Gamma_0\|_{SGL}^2 . \end{aligned} \tag{A.14}$$

Case 2b: If $\lambda_1 \|h_0\|_{\mathcal{H}_\mathcal{K}}^2 \geq \lambda_2 \|\Gamma_0\|_{SGL}^2$ then following the same logic as before:

$$\begin{aligned} \left\| \hat{h} \circ \hat{\Gamma} - h_0 \circ \Gamma_0 \right\|_n &= O_p(\lambda_1^{\frac{1}{2}}) \|h_0\|_{\mathcal{H}_\mathcal{K}}), \\ \left\| \hat{\Gamma} \right\|_{SGL}^2 &= O_p(\frac{\lambda_1}{\lambda_2}) \|h_0\|_{\mathcal{H}_\mathcal{K}}^2 , \\ \left\| \hat{h} \right\|_{\mathcal{H}_\mathcal{K}}^2 &= O_p(1) \|h_0\|_{\mathcal{H}_\mathcal{K}}^2 . \end{aligned} \tag{A.15}$$

We see that we have terms $O_p(\frac{\lambda_1}{\lambda_2})$ and $O_p(\frac{\lambda_2}{\lambda_1})$. We therefore need λ_1 and λ_2 to go to zero at the same rates.

We can think of our estimator $\hat{h} \circ \hat{\Gamma}$ as one object. See Appendix B for more details on this, which can explain why we have one rate for the two penaltys.

A.1.2 Proof of Corollary 6

We will use the following lemma from page 20 in [17].

Lemma 9. *A d dimensional ball of radius R, $B_d(R)$, in $\mathcal{R}^d$ with Euclidean metric can be covered by $(\frac{4R+\delta}{\delta})^d$ balls of radius δ.*

We have shown in the proof for Theorem 1 that the optimal $\boldsymbol{\gamma}$ vector is restricted to be within a ball of a radius that depends on the norm of $\mathbf{Y}$. For the sake of simplicity

let us confine our $\boldsymbol{\gamma}$ to be within a norm ball of radius 1, $\boldsymbol{\gamma} \in \{\|\boldsymbol{\gamma}\|_2^2 \leq 1\}$. We then confine our set which we called $\mathcal{A}$ to be restricted to those $\boldsymbol{\gamma}$, that is $\mathcal{A} = \{\Gamma : \Gamma(\mathbf{z}) = \boldsymbol{\gamma} \circ \mathbf{z}$ where $\boldsymbol{\gamma} \in \{\|\boldsymbol{\gamma}\|_2^2 \leq 1\}$. Since our $\boldsymbol{\gamma} \in R^s$, we can use Lemma 4 and cover our set $\mathcal{A}$ with $N_1 = \left(\frac{4+\delta}{\delta}\right)^s$ number of functions in the following sense. The ball of radius 1 in R^s can be covered (using the euclidean metric) by $\{\boldsymbol{\gamma}_1, \cdots \boldsymbol{\gamma}_{N_1}\}$. Since there is a one to one relationship between the functions Γ and $\boldsymbol{\gamma}$, take the set $\{\Gamma_1, ...\Gamma_{N_1}\}$ and define the metric between some Γ_j and Γ_k in the set $\mathcal{A}$ as $d(\Gamma_j, \Gamma_k) = \left\|\boldsymbol{\gamma}_j - \boldsymbol{\gamma}_k\right\|_2$. Then, the set of functions $\{\Gamma_1, ...\Gamma_{N_1}\}$ is a δ covering for $\mathcal{A}$ under this metric with entropy $s\ log(\frac{4+\delta}{\delta})$. For each Γ_j we have an induced RKHS, $\mathcal{H}_{\mathcal{K}\circ\Gamma_j} = \{h \circ \Gamma_j : h \in \mathcal{H}_\mathcal{K}\}$ with entropy no larger then that of $\mathcal{H}_\mathcal{K}$ which we are assuming has entropy $\leq A\delta^{-2\psi}$ for some $\psi \in (0, 1)$ and $A \in \mathcal{R}$. Therefore, the covering number $N_2 = N(\delta, \mathcal{H}_{\mathcal{K}\circ\Gamma_j}, P_n) \leq exp^{A\delta^{-2\psi}}$ which implies that for every Γ_j there exists a set $\{h_{j_1} \circ \Gamma_j, \cdots, h_{j_{N_2}} \circ \Gamma_j\}$ where for every $h \circ \Gamma_j \in \mathcal{H}_{\mathcal{K}\circ\Gamma_j}$ there exists an integer $i \in \{1, ...N_2\}$ where $\|h \circ \Gamma_j - h_{j_i} \circ \Gamma_j\|_{P_n} \leq \delta$. Our set $\mathcal{B}$ is essentially looking at the union of the different Hilbert spaces of the form $\mathcal{H}_{\mathcal{K}\circ\Gamma}$. Based on our setup, a natural guess of the delta covering number of this set would be roughly of size $N_1 \times N_2$ where we consider functions of the form $\{h_{1_1} \circ \Gamma_1, \cdots, h_{1_{N_2}} \circ \Gamma_1, \cdots, h_{N_{11}} \circ \Gamma_{N_1}, \cdots, h_{N_{1N_2}} \circ \Gamma_{N_1}\}$. In addition, we add N_2 functions from the set $\{h_1 \circ \Gamma_0, \cdots, h_{N_2} \circ \Gamma_0\}$ where Γ_0 is the true Γ_0 (or one of the true Γ_0) we are trying to estimate. Since $\mathcal{H}_{\mathcal{K}\circ\Gamma_j}$ is a Hilbert space for every j, if $h \circ \Gamma_j \in \mathcal{H}_{\mathcal{K}\circ\Gamma_j}$ so is $\frac{h\circ\Gamma_j}{\|h\|^2_{\mathcal{H}_\mathcal{K}}+\|h_0\|^2_{\mathcal{H}_\mathcal{K}}+\|\Gamma_j\|^2_{SGL}+\|\Gamma_0\|^2_{SGL}}$. We can simply ignore the denominator and substitute $\frac{h\circ\Gamma_j}{\|h\|^2_{\mathcal{H}_\mathcal{K}}+\|h_0\|^2_{\mathcal{H}_\mathcal{K}}+\|\Gamma_j\|^2_{SGL}+\|\Gamma_0\|^2_{SGL}}$ with $\tilde{h} \circ \Gamma_j \in H_{K\circ\Gamma_j}$ where $\tilde{h} = \frac{h}{\|h\|^2_{\mathcal{H}_\mathcal{K}}+\|h_0\|^2_{\mathcal{H}_\mathcal{K}}+\|\Gamma_j\|^2_{SGL}+\|\Gamma_0\|^2_{SGL}}$.
We will now formally prove Corollary (**??**).

Proof. Set $M = \sup_h < \nabla h(\mathbf{z}), \nabla h(\mathbf{z}) >$ where the inner product is the standard euclidean inner product (this is for a fixed $\mathbf{z}$, or we can take the $\sup_{h\in H_k, \mathbf{z}\in R^s} < \nabla h(\mathbf{z}), \nabla h(\mathbf{z}) >$ since we are assuming the gradient is uniformly bounded). Let

$N_1 = \frac{4+\left(\frac{\delta}{3M^{\frac{1}{2}}}\right)^s}{\left(\frac{\delta}{3M^{\frac{1}{2}}}\right)}$ which is the number of balls needed to provide a $\left(\frac{\delta}{3M^{\frac{1}{2}}}\right)$ covering for a norm 1 ball in $\mathcal{R}^s$. Let $N_2 = exp^{\left(A(\frac{\delta}{3})^{-2\psi}\right)}$ which is the covering number needed to provide a $\frac{\delta}{3}$ cover of our space $\mathcal{H}_\mathcal{K}$.

Let

$$\tilde{\hat{h}} \circ \hat{\Gamma} - \tilde{h}_0 \circ \Gamma_0 =$$
$$\frac{\hat{h} \circ \hat{\Gamma}}{\left\|\hat{h}\right\|^2_{\mathcal{H}_\mathcal{K}} + \|h_0\|^2_{\mathcal{H}_\mathcal{K}} + \left\|\hat{\Gamma}\right\|^2_{SGL} + \|\Gamma_0\|^2_{SGL}} - \frac{h_0 \circ \Gamma_0}{\left\|\hat{h}\right\|^2_{\mathcal{H}_\mathcal{K}} + \|h_0\|^2_{\mathcal{H}_\mathcal{K}} + \left\|\hat{\Gamma}\right\|^2_{SGL} + \|\Gamma_0\|^2_{SGL}}$$

be an arbitrary function in the set $\mathcal{B}$. There exists an Γ_j where $j \in \{1, ...N_1\}$ such that $d(\Gamma_j, \hat{\Gamma}) \leq \frac{\delta}{3 \max_{i=1,\cdots,n} \|\mathbf{z}_i\|_2 \sqrt{M}}$, and there exists a i where $i \in \{1, ...N_2\}$ such that $\left\|\tilde{\hat{h}} \circ \Gamma_j - h_{j_i} \circ \Gamma_j\right\|_{P_n} \leq \frac{\delta}{3}$. Similarly, there exists a $t \in \{1, ...N_2\}$ such that $\left\|\tilde{h}_0 \circ \Gamma_0 - h_t \circ \Gamma_0\right\|_{P_n} \leq \frac{\delta}{3}$. We construct our approximating function of $\tilde{\hat{h}} \circ \hat{\Gamma} - \tilde{h}_0 \circ \Gamma_0$ as $h_{j_i} \circ \Gamma_j - h_t \circ \Gamma_0$. We will show that this function is within δ of our arbitrary function $\tilde{\hat{h}} \circ \hat{\Gamma} - \tilde{h}_0 \circ \Gamma_0$. We have:

$$\left\|(\tilde{\hat{h}} \circ \hat{\Gamma} - \tilde{h}_0 \circ \Gamma_0) - (h_{j_i} \circ \Gamma_j - h_t \circ \Gamma_0)\right\|_{P_n} \leq$$
$$\left\|\tilde{\hat{h}} \circ \hat{\Gamma} - h_{j_i} \circ \Gamma_j\right\|_{P_n} + \left\|\tilde{h}_0 \circ \Gamma_0 - h_t \circ \Gamma_0\right\|_{P_n} \leq$$
$$\left\|\tilde{\hat{h}} \circ \hat{\Gamma} - h_{j_i} \circ \Gamma_j\right\|_{P_n} + \frac{\delta}{3} =$$
$$\left\|\tilde{\hat{h}} \circ \Gamma_j - h_{j_i} \circ \Gamma_j + \nabla \tilde{\hat{h}}(C(\cdot))(\hat{\Gamma} - \Gamma_j)\right\|_{P_n} + \frac{\delta}{3}$$

where we used the mean value theorem for multivariate functions:

$\tilde{\hat{h}} \circ \hat{\Gamma}(\mathbf{z}) = \tilde{\hat{h}} \circ \Gamma_j(\mathbf{z}) + \nabla \tilde{\hat{h}}(C(\mathbf{z}))(\hat{\Gamma}(\mathbf{z}) - \Gamma_j(\mathbf{z}))$ for some vector $\mathbf{z} \in \mathcal{R}^s$ that lies in the segment from $\gamma_j \circ \mathbf{z}$ and $\hat{\gamma} \circ \mathbf{z}$. $C(\cdot)$ is an unknown function that maps from $\mathcal{R}^s$ into

$\mathcal{R}^s$ that allows for the formula to hold. Continuing our chain of inequalities we get

$$\left\|\tilde{\hat{h}} \circ \Gamma_j - h_{j_i} \circ \Gamma_j + \nabla\tilde{\hat{h}}(C(\cdot))(\hat{\Gamma} - \Gamma_j)\right\|_{P_n} + \frac{\delta}{3} \leq$$
$$\left\|\nabla\tilde{\hat{h}}(C(\cdot))(\hat{\Gamma} - \Gamma_j)\right\|_{P_n} + \frac{\delta}{3} + \frac{\delta}{3} =$$
$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\nabla\tilde{\hat{h}}(C(\mathbf{z}_i))(\hat{\Gamma}(\mathbf{z}_i) - \Gamma_j(\mathbf{z}_i))\right)^2} + \frac{\delta}{3} + \frac{\delta}{3} \leq$$
$$\sqrt{\frac{1}{n}\sum_{i=1}^{n} M\left\|\hat{\boldsymbol{\gamma}} \circ \mathbf{z_i} - \boldsymbol{\gamma_j} \circ \mathbf{z_i}\right\|_2^2} + \frac{\delta}{3} + \frac{\delta}{3} \leq$$
$$\sqrt{M\left(\frac{\delta}{3\max\limits_{i=1,\cdots,n}\|\mathbf{z}_i\|_2\sqrt{M}}\right)^2 \max_{i=1,\cdots,n}\|\mathbf{z}_i\|_2^2} + \frac{\delta}{3} + \frac{\delta}{3} =$$
$$\frac{\delta}{3} + \frac{\delta}{3} + \frac{\delta}{3} = \delta.$$

So to provide a δ cover we need $N_1 \times N_2 + N_2$ number of functions or

$$exp^{\left(A(\frac{\delta}{3})^{-2\psi}\right)}\left(\frac{4 + \left(\frac{\delta}{3M^{\frac{1}{2}}}\right)}{\left(\frac{\delta}{3M^{\frac{1}{2}}}\right)}\right)^s + exp^{\left(A(\frac{\delta}{3})^{-2\psi}\right)} =$$
$$exp^{\tilde{A}(\delta)^{-2\psi}}\left(\frac{C+\delta}{\delta}\right)^s + exp^{\tilde{A}(\delta)^{-2\psi}},$$

where $\tilde{A} = \frac{A}{3^{-2\psi}}$ and $C = 12M^{\frac{1}{2}}$. Taking the log we see the entropy is $\leq \tilde{A}\delta^{-2\psi} + log((\frac{C+\delta}{\delta})^s + 1$ which is of the same order as $\leq \tilde{A}\delta^{-2\psi}$ (the log term is dominated by the first term). Therefore a sufficient (but not necessary) condition for our set $\mathcal{B}$ to have the same entropy as that of the original RKHS H_K is for the $\sup_h < \nabla h(\mathbf{z}), \nabla h(\mathbf{z}) >$ to be bounded. Having bounded derivatives is reasonable for any RKHS since every

RKHS satisfies a "sort of" Lipschitz condition where

$$|h(X) - h(Y)| = |< h, \mathcal{K}_X > - < h, \mathcal{K}_Y > | \leq$$
$$\|h\|_{\mathcal{H}_\mathcal{K}} < \mathcal{K}_X, \mathcal{K}_Y >^{\frac{1}{2}} = \|h\|_{\mathcal{H}_\mathcal{K}} d(X, Y)$$

Where the distance metric in $\mathcal{R}^s$ is defined as $d(X,Y)^2 = \mathcal{K}(X,X) - 2\mathcal{K}(X,Y) + \mathcal{K}(Y,Y)$ If we restrict our functions in the RKHS of norm $\leq C$ for some constant C then we have a universal Lipschitz constant C which implies bounded derivatives.

□

A.2 Gauss-Newton Algorithm

The Gauss-Newton method for non linear optimization looks at the class of minimization problems of the form

$$\min_{\upsilon} \frac{1}{2} \|F(\upsilon) - y\|^2_{\mathcal{H}_2} \tag{A.1}$$

for a differentiable (Fréchet or Gateaux) operator $F : \mathcal{H}_1 \mapsto \mathcal{H}_2$ where $\mathcal{H}_1$ and $\mathcal{H}_2$ are Hilbert spaces, and solves for υ by iterating

$$\upsilon_{n+1} = \upsilon_n - [F^{'}(\upsilon_n)^\star F^{'}(\upsilon_n)]^{-1} F^{'}(\upsilon_n)^\star (F(\upsilon_n) - y) \tag{A.2}$$

where $F^{'}(\upsilon_n)^\star$ is the adjoint of $F^{'}(\upsilon_n)$. For our purposes, we are mapping from $R^s \mapsto R^n$ so the adjoint is just the transpose of the matrix associated with the linear operator $F^{'}(\upsilon)$. At each iteration of the Gauss-Newton method, this is equivalent to linearizing the function F and solving for υ_{n+1} where υ_{n+1} is the solution to the following minimization problem:

$$\min_{\upsilon} \frac{1}{2} \left\| F(\upsilon_n) + F^{'}(\upsilon_n)(\upsilon - \upsilon_n) - y \right\|^2_{\mathcal{H}_2} \tag{A.3}$$

Salzo and Villa (2012) [40] extend that method and proposed the proximal Gauss-Newton method which looks at the class of minimization problems:

$$\min_{\upsilon} \frac{1}{2} \|F(\upsilon) - y\|^2_{\mathcal{H}_2} + J(\upsilon), \tag{A.4}$$

where F is smooth with respect to υ and J is a convex but possibly non-smooth function with respect to υ. They proposed a proximal Gauss-Newton algorithm to solve problems of the form in equation (A.4) by linearizing the functional in (A.4) and solving for υ_{n+1} by iterating

$$\min_{\upsilon} \frac{1}{2} \left\| F(\upsilon_n) + F^{'}(\upsilon_n)(\upsilon - \upsilon_n) - y \right\|^2 + J(\upsilon_n) \tag{A.5}$$

which is equivalent to setting

$$\upsilon_{n+1} = prox_J^{H(x)}(\upsilon_n - [F^{'}(\upsilon_n)^{\star} F^{'}(\upsilon_n)]^{-1} F^{'}(\upsilon_n)^{\star}(F(\upsilon_n) - y)) \tag{A.6}$$

where the proximal operator $prox_J^{H(x)}(y) = \arg\min_{x \in \chi}(\frac{1}{2} \|x - y\|^2_H + J(x))$ and where $H(x) := F^{'}(x)^{\star} F^{'}(x)$. This H induces a new inner product and norm denoted by $\|\cdot\|_H$ where $< x, z >_H \quad = \quad < x, Hz >$. We see an equivalence between the proximal Gauss-Newton algorithm and the algorithm we propose in Section 3 step (iii) of our paper.

A.3 Additional Simulation Results in Scenario 2

Table A.1: Model Size for Scenario 2

	Model Size																			
Model	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**>18**
$MFRS_{Lasso}$	0	0	3	93	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
$MFRS_{GLasso}$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	94	6
$MFRS_{SGL}$	0	0	0	0	1	11	24	41	13	5	4	1	0	0	0	0	0	0	0	0
$MFRS_{MCP}$	0	0	1	88	8	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
$MFRS_{GMCP}$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	93	7
$Lasso$	0	0	0	0	0	49	17	8	8	1	2	3	1	1	3	0	0	0	0	7
$GLasso$	1	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	86	8
SGL	0	0	0	0	0	49	17	10	6	1	1	3	1	1	2	1	1	0	0	7
MCP	0	0	0	0	0	50	18	10	4	1	2	2	1	3	1	1	0	0	1	6
$GMCP$	1	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	85	8

APPENDIX B

Additional Graphs for Chapters 3

B.1 Additional Graphs from ELEMENT dataset from Chapter 3

Figure B.1: Leading Eigenfunction extracted for Tri-axis 7-day functional data

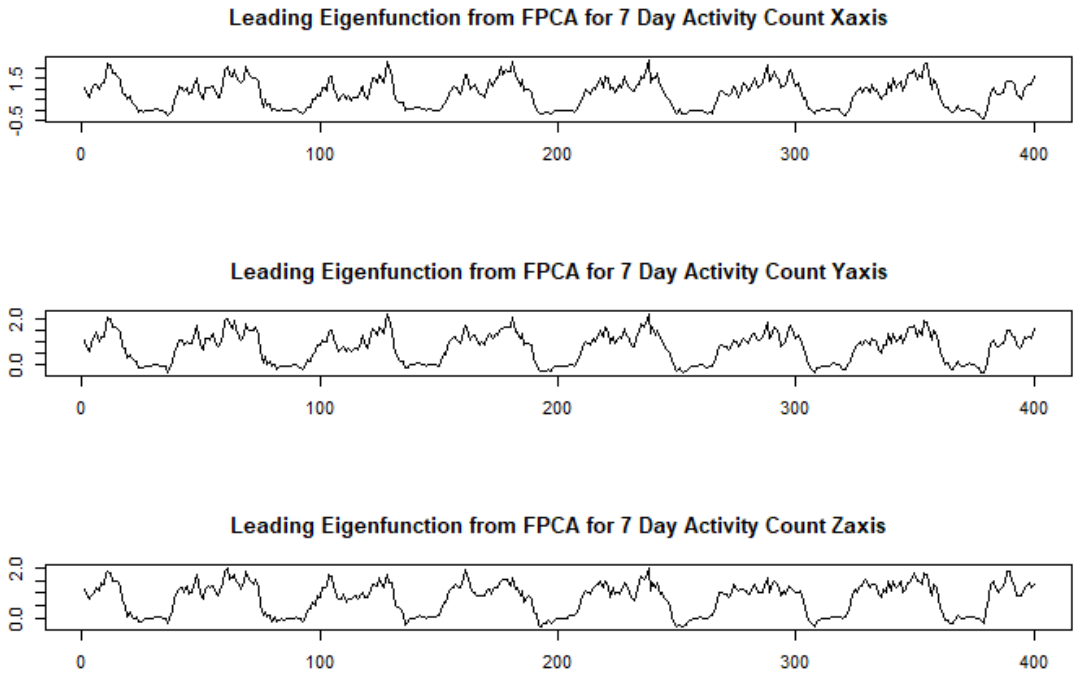


Figure B.2: Leading Eigenfunction extracted for VM 7-day functional data

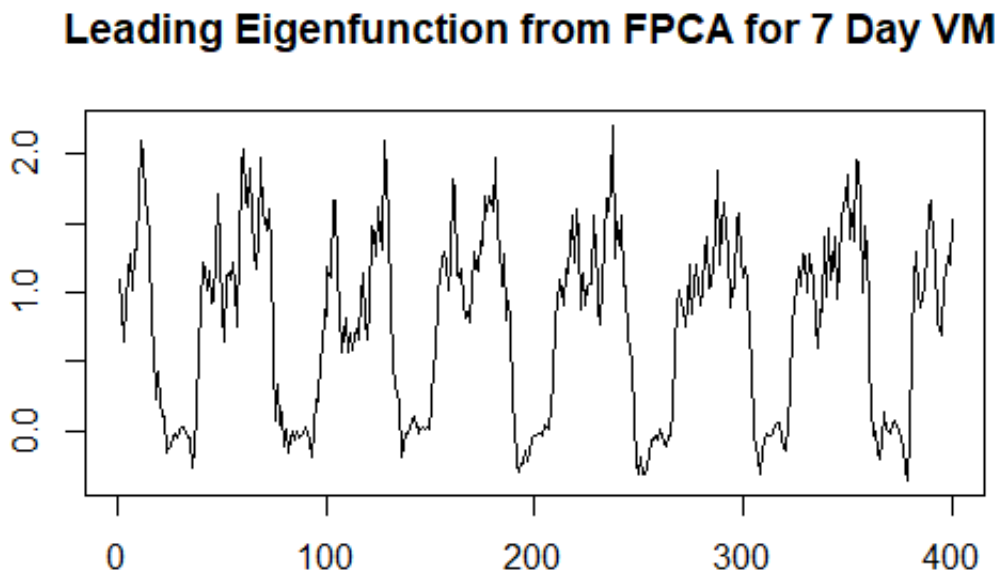


Figure B.3: Leading Eigenfunction extracted for Tri-axis 1-day averaged functional data

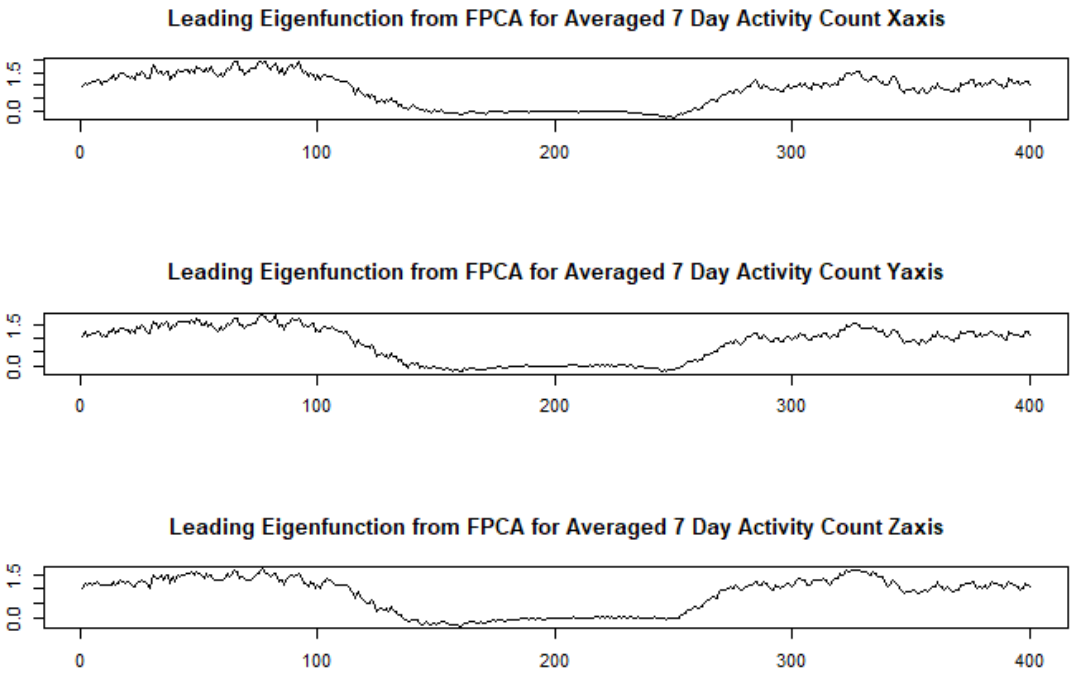
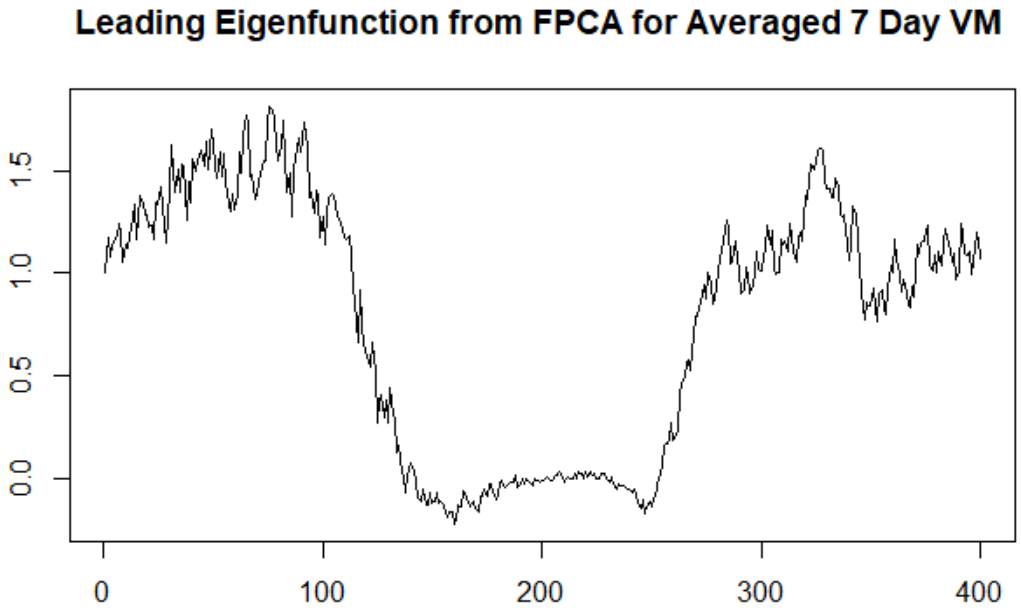


Figure B.4: Leading Eigenfunction extracted for VM 1-day averaged functional data



APPENDIX C

Additional Graphs for Chapters 4

C.1 Additional Graphs from ELEMENT dataset from Chapter 4

Figure C.1: Leading Eigenfunction extracted from X(t) process for Tri-axis data

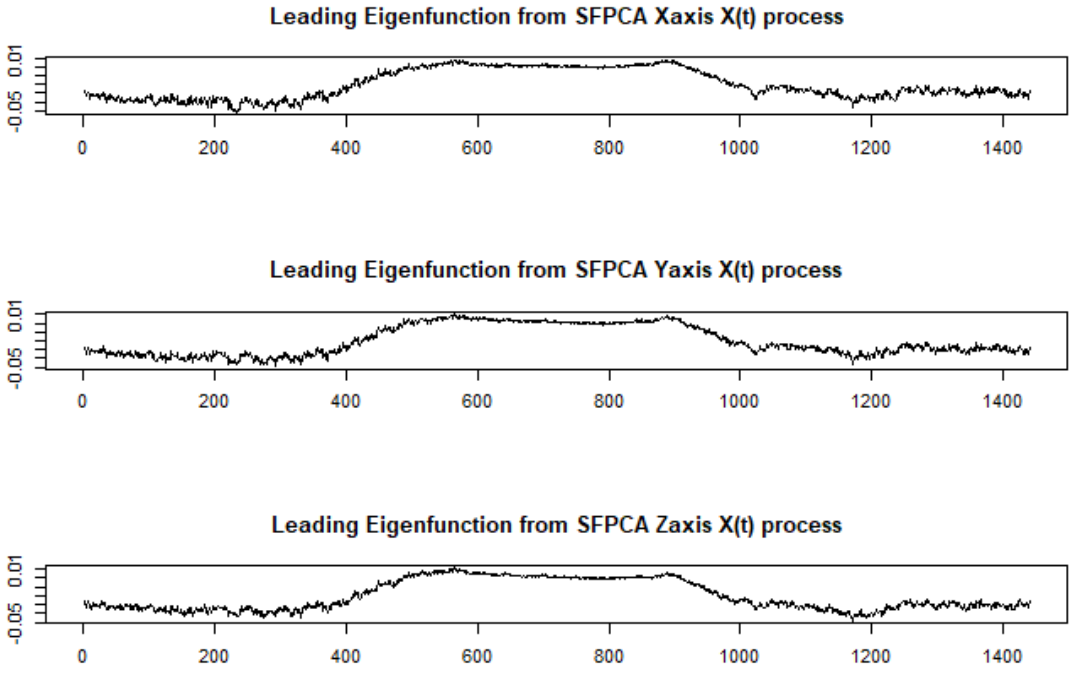


Figure C.2: Leading Eigenfunction extracted from U(t) process for Tri-axis data

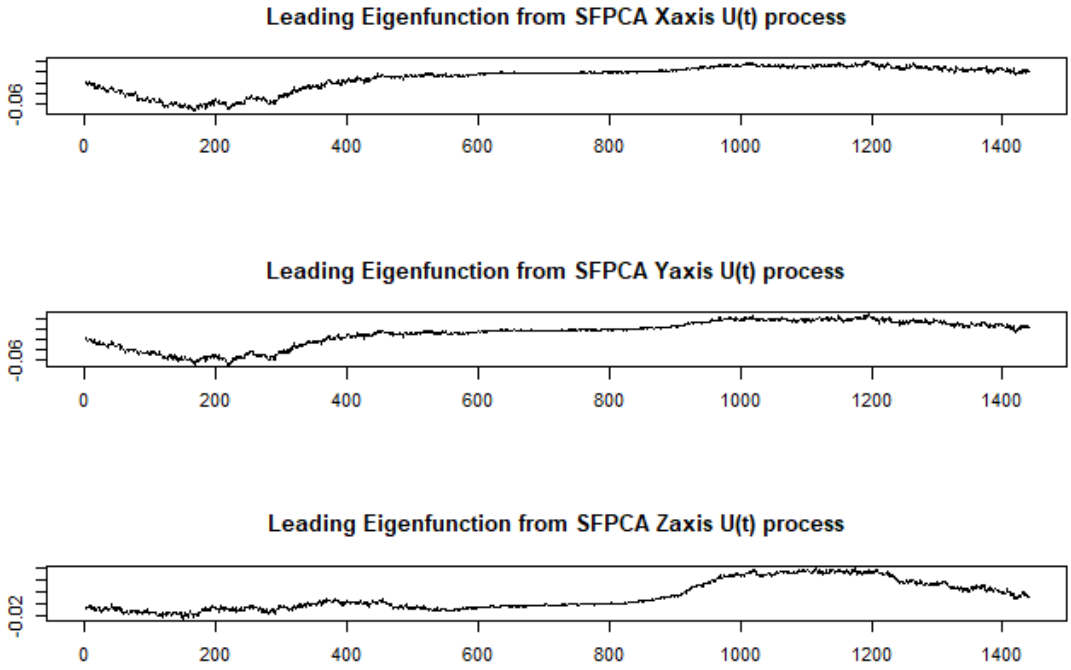


Figure C.3: Leading Eigenfunction extracted from X(t) and U(t) process for VM

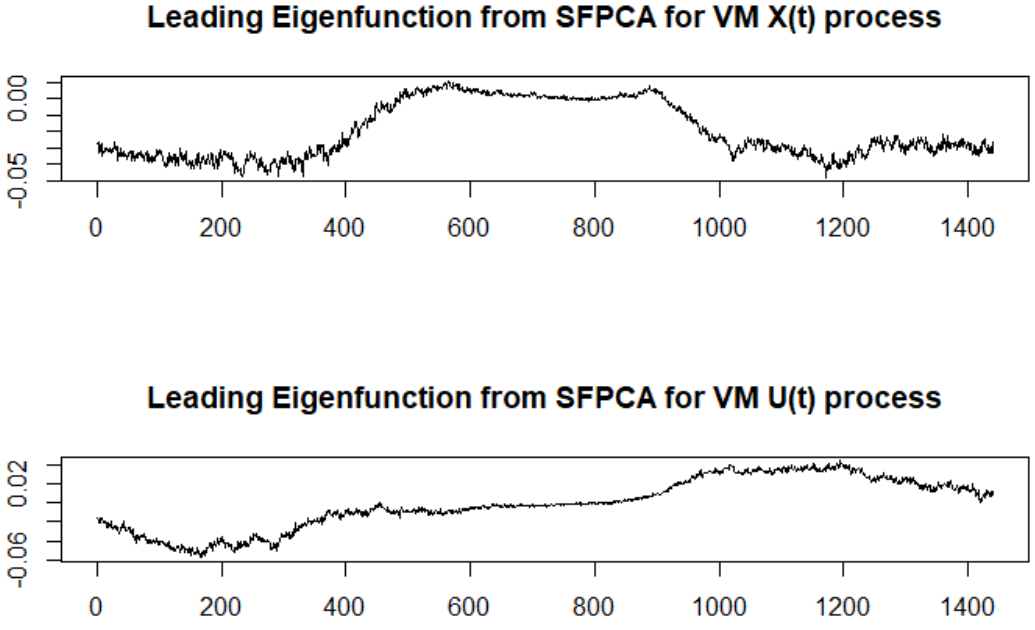
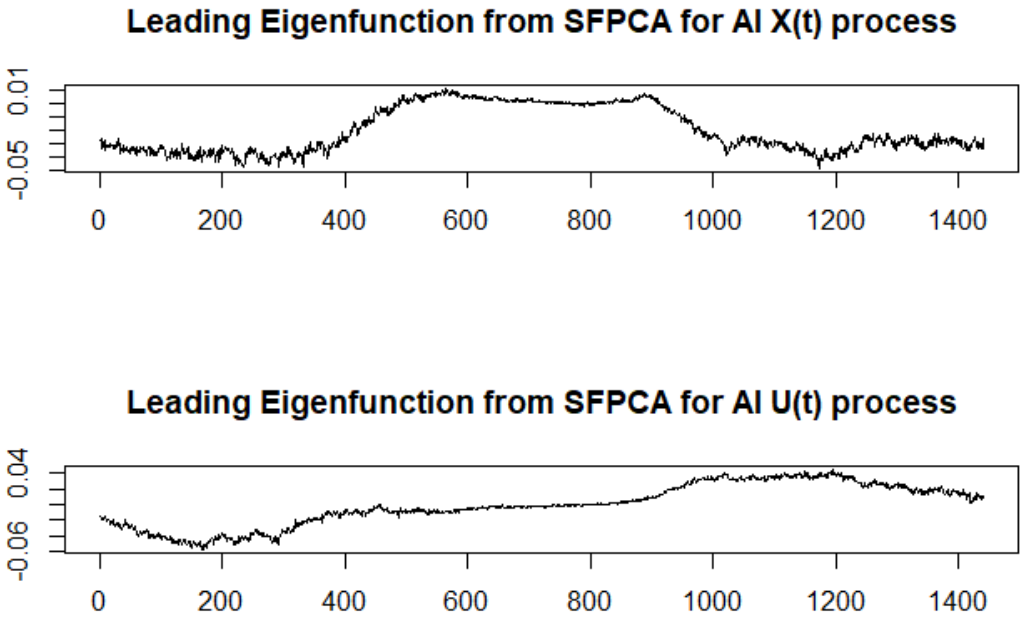


Figure C.4: Leading Eigenfunction extracted from X(t) and U(t) process for AI



BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Alhassan, S., K. Lyden, C. A. Howe, S. K. Keadle, O. Nwaokelemeh, and P. S. Freedson (2012), Accuracy of accelerometer regression models in predicting energy expenditure and mets in children and youth., *Pediatric exercise science*, *24 4*, 519–36.

[2] Bai, J., C. Di, L. Xiao, K. R. Evenson, A. Z. LaCroix, C. M. Crainiceanu, and D. M. Buchner (2016), An activity index for raw accelerometry data and its comparison with other activity metrics, *PLoS ONE*, *11*(8), e0160,644, doi: 10.1371/journal.pone.0160644.

[3] Bai, J., Y. Sun, J. A. Schrack, C. M. Crainiceanu, and M.-C. Wang (2018), A two-stage model for wearable device data, *Biometrics*, *74*(2), 744752, doi: 10.1111/biom.12781.

[4] Belloni, A., and V. Chernozhukov (2013), Least squares after model selection in high-dimensional sparse models, *Bernoulli*, *19*(2), 521547, doi:10.3150/11-BEJ410.

[5] Bosq, D. (2000), Linear processes in function spaces, *149*, doi:10.1007/978-1-4612-1154-9.

[6] Breiman, L. (1995), Better subset regression using the nonnegative garrote, *Technometrics*, *37*(4), 373384, doi:10.1080/00401706.1995.10484371.

[7] Buccini, A. (2017), Regularizing preconditioners by non-stationary iterated tikhonov with general penalty term, *Applied Numerical Mathematics*, *116*, 6481, doi:10.1016/j.apnum.2016.07.009.

[8] Cardot, H., F. Ferraty, and P. Sarda (1999), Functional linear model, *Statistics & Probability Letters*, *45*, 11–22, doi:10.1016/S0167-7152(99)00036-X.

[9] Cardot, H., F. Ferraty, and P. Sarda (2003), Spline estimators for the functional linear model, *Statistica Sinica*, *13*(3), 571–591.

[10] Chandler, J. L., K. Brazendale, M. W. Beets, and B. A. Mealing (2016), Classification of physical activity intensities using a wristworn accelerometer in 812yearold children, *Pediatric Obesity*, *11*(2), 120127, doi:10.1111/ijpo.12033.

[11] Chen, K. Y., and D. R. Bassett (2005), The technology of accelerometry-based activity monitors: Current and future, *Medicine and Science in Sports and Exercise*, *37*(11 Suppl), S490S500, doi:10.1249/01.mss.0000185571.49104.82.

[12] CROUTER, S. E., J. I. FLYNN, and D. R. BASSETT (2015), Estimating physical activity in youth using a wrist accelerometer, *Medicine and Science in Sports and Exercise*, *47*(5), 944951, doi:10.1249/MSS.0000000000000502.

[13] Di, C.-Z., C. M. Crainiceanu, B. S. Caffo, and N. M. Punjabi (2009), Multilevel functional principal component analysis, *Annals of Applied Statistics*, *3*(1), 458488, doi:10.1214/08-AOAS206.

[14] Fan, Y., G. M. James, and P. Radchenko (2015), Functional additive regression, *The Annals of Statistics*, *43*(5), 22962325, doi:10.1214/15-AOS1346.

[15] Ferraty, F., and P. Vieu (2003), Functional nonparametric statistics: A double infinite dimensional framework, *Recent Advances and Trends in Nonparametric Statistics*, *45*, doi:10.1016/B978-044451378-6/50005-3.

[16] Ferraty, F., A. Mas, and P. Vieu (2007), Nonparametric regression on functional data: Inference and practical aspects, *Australian & New Zealand Journal of Statistics*, *49*(3), 267286, doi:10.1111/j.1467-842X.2007.00480.x.

[17] Geer, S. (2000), *Empirical Processes in M-Estimation*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press.

[18] Goldsmith, J., X. Liu, A. Rundle, and J. Jacobson (2016), New insights into activity patterns in children, found using functional data analyses, *Medicine and Science in Sports and Exercise*, *48*(9), 17231729, doi:10.1249/MSS.0000000000000968.

[19] Gopinath, B., L. L. Hardy, E. Teber, and P. Mitchell (2011), Association between physical activity and blood pressure in prepubertal children, *Hypertension Research*, *34*(7), 851855, doi:10.1038/hr.2011.46.

[20] Hainmueller, J., and C. Hazlett (2014), Kernel regularized least squares: Reducing misspecification bias with a flexible and interpretable machine learning approach, *Political Analysis*, *22*(2), 143168, doi:10.1093/pan/mpt019.

[21] Hall, P., and M. Hosseini-Nasab (2006), On properties of functional principal components analysis, *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *68*(1), 109126, doi:10.1111/j.1467-9868.2005.00535.x.

[22] Hall, P., H.-G. Mller, and J.-L. Wang (2006), Properties of principal component methods for functional and longitudinal data analysis, *The Annals of Statistics*, *34*(3), 14931517, doi:10.1214/009053606000000272.

[23] Happ, C., and S. Greven (3/4/2018), Multivariate functional principal component analysis for data observed on different (dimensional) domains, *Journal of the American Statistical Association*, *113*(522), 649659, doi: 10.1080/01621459.2016.1273115.

[24] Henry, I., D. Bernstein, M. Banet, J. Mulligan, S. Moulton, G. Grudic, and V. Convertino (2011), Body-worn, non-invasive sensor for monitoring stroke volume, cardiac output and cardiovascular reserve, in *Proceedings of the 2nd Conference on wireless health*, WH 11, p. 12, ACM, doi:10.1145/2077546.2077575.

[25] Hoerl, A. E., and R. W. Kennard (1970), Ridge regression: Biased estimation for nonorthogonal problems, *Technometrics*, *12*(1), 55–67, doi: 10.1080/00401706.1970.10488634.

[26] Jansen, E. C., et al. (2018), Adiposity in adolescents: The interplay of sleep duration and sleep variability, *The Journal of Pediatrics*, *203*, 309316, doi: 10.1016/j.jpeds.2018.07.087.

[27] John, D., and P. Freedson (2012), Actigraph and actical physical activity monitors: A peek under the hood, *Medicine and Science in Sports and Exercise*, *44*(1), S86S89, doi:10.1249/MSS.0b013e3182399f5e.

[28] Lewis, R. C., J. D. Meeker, K. E. Peterson, J. M. Lee, G. G. Pace, A. Cantoral, and M. M. Tllez-Rojo (2013), Predictors of urinary bisphenol a and phthalate metabolite concentrations in mexican children, *Chemosphere*, *93*(10), 23902398, doi:10.1016/j.chemosphere.2013.08.038.

[29] Li, H., S. Kozey Keadle, J. Staudenmayer, H. Assaad, J. Z. Huang, and R. J. Carroll (2015), Methods to assess an exercise intervention trial based on 3-level functional data, *Biostatistics*, *16*(4), 754771, doi:10.1093/biostatistics/kxv015.

[30] Lin, X., and D. Zhang (1999), Inference in generalized additive mixed models by using smoothing splines, *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *61*(2), 381400, doi:10.1111/1467-9868.00183.

[31] Lin, X., G. Wahba, D. Xiang, F. Gao, R. Klein, and B. Klein (2000), Smoothing spline anova models for large data sets with bernoulli observations and the randomized gacv, *The Annals of Statistics*, *28*(6), 15701600.

[32] Lin, Y., and H. H. Zhang (2006), Component selection and smoothing in multivariate nonparametric regression, *The Annals of Statistics*, *34*(5), 22722297, doi:10.1214/009053606000000722.

[33] Liu, D., X. Lin, and D. Ghosh (2007), Semiparametric regression of multidimensional genetic pathway data: Least-squares kernel machines and linear mixed models, *Biometrics*, *63*(4), 10791088, doi:10.1111/j.1541-0420.2007.00799.x.

[34] McLean, M. W., G. Hooker, A.-M. Staicu, F. Scheipl, and D. Ruppert (2014), Functional generalized additive models, *Journal of Computational and Graphical Statistics*, *23*(1), 249269, doi:10.1080/10618600.2012.729985.

[35] Mitchell, J. A., R. R. Pate, V. EspaaRomero, J. R. ONeill, M. Dowda, and P. R. Nader (2013), Moderatetovigorous physical activity is associated with decreases in body mass index from ages 9 to 15 years, *Obesity*, *21*(3), E280E286, doi: 10.1002/oby.20118.

[36] Mller, H.-G., and F. Yao (2008), Functional additive models, *Journal of the American Statistical Association*, *103*(484), 15341544, doi: 10.1198/016214508000000751.

[37] Montoye, A. H., M. Begum, Z. Henning, and K. A. Pfeiffer (2017), Comparison of linear and non-linear models for predicting energy expenditure from raw accelerometer data, *Physiological Measurement*, *38*(2), 343357, doi:10.1088/1361-6579/38/2/343.

[38] Peng, H., and T. Huang (2011), Penalized least squares for single index models, *Journal of Statistical Planning and Inference*, *141*(4), 13621379, doi: 10.1016/j.jspi.2010.10.003.

[39] Ramsay, J. O., and B. W. Silverman (2005), Functional data analysis, doi: 10.1007/978-1-4757-7107-7.

[40] Salzo, S., and S. Villa (2012), Convergence analysis of a proximal gauss-newton method, *Computational Optimization and Applications*, *53*(2), 557589, doi: 10.1007/s10589-012-9476-9.

[41] Sasaki, J. E., A. M. Hickey, J. W. Staudenmayer, D. John, J. A. Kent, and P. S. Freedson (2016), Performance of activity classification algorithms in free-living older adults, *Medicine and Science in Sports and Exercise*, *48*(5), 941950, doi:10.1249/MSS.0000000000000844.

[42] Schelldorfer, J., L. Meier, and P. Bhlmann (2014), Glmmlasso: An algorithm for high-dimensional generalized linear mixed models using 1-penalization, *Journal of Computational and Graphical Statistics*, *23*(2), 460477, doi: 10.1080/10618600.2013.773239.

[43] Schmidt, M., G. Fung, and R. Rosales (2007), Fast optimization methods for l1 regularization: A comparative study and two new approaches, in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 4701, p. 286297.

[44] Schrack, J. A., V. Zipunnikov, J. Goldsmith, J. Bai, E. M. Simonsick, C. Crainiceanu, and L. Ferrucci (2014), Assessing the physical cliff: Detailed quantification of age-related differences in daily patterns of physical activity, *Journals of Gerontology - Series A Biological Sciences and Medical Sciences*, *69*(8), 973979, doi:10.1093/gerona/glt199.

[45] Schwarzfischer, P., et al. (2017), Bmi and recommended levels of physical activity in school children, *BMC Public Health*, *17*(1), 5959, doi:10.1186/s12889-017-4492-4.

[46] Shou, H., V. Zipunnikov, C. M. Crainiceanu, and S. Greven (2015), Structured functional principal component analysis, *Biometrics*, *71*(1), 247257, doi:10.1111/biom.12236.

[47] Simon, N., J. Friedman, T. Hastie, and R. Tibshirani (2013), A sparse-group lasso, *Journal of Computational and Graphical Statistics*, *22*(2), 231245, doi:10.1080/10618600.2012.681250.

[48] Smola, A., and B. Schoelkopf (2000), *Sparse Greedy Matrix Approximation for Machine Learning*, Morgan Kauffman Publishers.

[49] Steinhubl, S. R., E. D. Muse, and E. J. Topol (2015), The emerging field of mobile health, *Science Translational Medicine*, *7*(283), 283rv3, doi:10.1126/scitranslmed.aaa3487.

[50] Tsioufis, C., et al. (2011), Relation between physical activity and blood pressure levels in young greek adolescents: the leontio lyceum study, *European journal of public health*, *21*(1), 6368, doi:10.1093/eurpub/ckq006.

[51] Williams, C., and M. Seeger (2001), Using the nystrm method to speed up kernel machines, in *Advances in Neural Information Processing Systems 13*, pp. 682–688, MIT Press.

[52] Wolfe, P. (1/4/1969), Convergence conditions for ascent methods, *SIAM Review*, *11*(2), 226235, doi:10.1137/1011036.

[53] Wood, S. N. (2006), Generalized additive models: an introduction with r.

[54] Yao, F., H.-G. Mller, and J.-L. Wang (2005), Functional data analysis for sparse longitudinal data, *Journal of the American Statistical Association*, *100*(470), 577590, doi:10.1198/016214504000001745.

[55] Yuan, M., and Y. Lin (2006), Model selection and estimation in regression with grouped variables, *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *68*(1), 4967, doi:10.1111/j.1467-9868.2005.00532.x.

[56] Zhang, Y., H. Li, S. K. Keadle, C. E. Matthews, and R. J. Carroll (2019), A review of statistical analyses on physical activity data collected from accelerometers, *Statistics in Biosciences*, *11*(2), 465476, doi:10.1007/s12561-019-09250-6.

[57] Zhu, H., F. Yao, and H. H. Zhang (2014), Structured functional additive regression in reproducing kernel hilbert spaces, *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *76*(3), 581603, doi:10.1111/rssb.12036.

[58] Zhu, J., and T. Hastie (2005), Kernel logistic regression and the import vector machine, *Journal of Computational and Graphical Statistics*, *14*(1), 185205, doi: 10.1198/106186005X25619.

[59] Zou, H., and T. Hastie (2005), Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *67*(2), 301320, doi:10.1111/j.1467-9868.2005.00503.x.