ADVANCED MODELING OF LONGITUDINAL SPECTROSCOPY DATA

Madan Gopal Kundu

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy
in the Department of Biostatistics,
Indiana University

May 2014

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Jaroslaw Harezlak, Ph.D., Chair

Timothy W. Randolph, Ph.D.

Doctoral Committee

Jyotirmoy Sarkar, Ph.D.

January 24, 2014

Gregory K. Steele, Ph.D.

Constantin T. Yiannoutsos, Ph.D.

DEDICATION

To my family...

ACKNOWLEDGMENTS

I am profoundly thankful to my wife, Jhili, who stands by my side at every step through the frustration and happiness, always willing to lend a hand, an ear, a shoulder, and a cheer. Her support, encouragement, quiet patience and unwavering love were undeniably the bedrock upon which the past three years of my life have been built. Her tolerance of my occasional bad temper is a testament in itself of her unyielding devotion and love. I thank my parents and my brother for their faith in me and allowing me to be as ambitious as I wanted. It was under their watchful eye that I gained so much drive and an ability to tackle challenges head on. I also extend my thank to Jhili's parents.

I would like to express my deep gratitude to my research advisor Prof. Jaroslaw Harezlak for his thorough guidance, unwavering support and extreme patience in supervising me over the last 3 years. His mentorship was paramount in providing a well rounded experience, consistent with my long-term career goals. I have truly enjoyed working with him. I have learnt the art of creativity, critical-thinking, and hard work as a researcher through him. I extend my thanks to my research committee members - Prof. Constantine Yiannoutsos, Prof. Tim Randolph, Prof. Jyoti Sarkar and Prof. Gregory Steele for their advice, cooperation, encouragement and willingness to help. I am also thankful to the program chairs, Prof. Bary Katz and Prof. Ben Boukai. Their efforts were extremely critical for timely completion in the PhD program and smooth transition to next stage of my life.

I also acknowledge the support I received from my friend Sukhdev Mishra. He has helped me to survive through difficult situation. Indeed, he is friend in need. I extend my thanks to Pradipta Sarkar, Manohar Chakroborty and Neha Patel who have helped me at

various capacities. They inspired me to work smarter and harder and I am forever grateful.

Above all, I am grateful to the Almighty God for guiding me all through.

Madan Gopal Kundu

Madan Gopal Kundu

ADVANCED MODELING OF LONGITUDINAL SPECTROSCOPY DATA

Magnetic resonance (MR) spectroscopy is a neuroimaging technique. It is widely used to quantify the concentration of important metabolites in a brain tissue. Imbalance in concentration of brain metabolites has been found to be associated with development of neurological impairment. There has been an increasing trend of using MR spectroscopy as a diagnosis tool for neurological disorders. We established statistical methodology to analyze data obtained from the MR spectroscopy in the context of the HIV associated neurological disorder. First, we have developed novel methodology to study the association of marker of neurological disorder with brain MR spectrum and its evolution over time. This setting fits in the framework of scalar-on-function regression model with individual spectrum as the functional predictor. We have extended one of the existing cross-sectional scalar-on-function regression techniques for longitudinal set-up. Advantages of the proposed method include: 1) ability to model flexible time-varying associations between the scalar response and the functional predictor and (2) ability to incorporate prior information.

In the second part of my research, I studied the influence of the clinical and demographic factors on the progression of the brain metabolites over time. To understand the influence of these factors in a fully non-parametric way, we proposed LongCART algorithm to construct a regression tree with the longitudinal data. Such a regression tree identifies smaller subpopulations (characterized by the baseline factors) with differential longitudinal profiles and hence helps us to identify the influence of the baseline factors. Advantages of the LongCART algorithm include: (1) maintaining type-I error while determining the best split, (2) substantially reducing computation time and (3) applicability in the mistimed observation setting.

Finally, I carried out an in-depth analysis of longitudinal changes in the brain metabolite concentrations in three brain regions, namely, white matter, gray matter and basal ganglia in chronically infected HIV patients enrolled in the HIV Neuroimaging Consortium study. We studied the influence of important baseline factors (clinical and demographic) on the longitudinal brain metabolite profiles using the LongCART algorithm in order to identify subgroups of patients at higher risk of neurological impairment.

Jaroslaw Harezlak, Ph.D., Chair

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Motivation and Objective

The biological basis for several psychiatric disorders are not yet fully understood. The recent advancement in neuroimaging techniques has allowed to relate development of psychiatric disorders to brain mechanisms (Dager et al., 2008). For past two decades, researchers have started to use magnetic resonance imaging (MRI) techniques, e.g., magnetic resonance spectroscopy (MRS), as a diagnostic tool for various psychiatric disorders (Hasler et al., 2010; Horská et al., 2013; Lagopoulos, 2007; O'Neill et al., 2013; Wang et al., 2006). In general, development of a psychiatric disorder is seen as a result of abundance or reduction in certain metabolites in brain. Each of these metabolites contributes to the MR spectrum. Therefore, the changes in concentration of metabolites corresponding to development or progression of psychiatric disorders should be reflected in the observed MR spectrum. This justifies the use of MRS to study psychiatric disorders.

1.1 Magnetic resonance spectroscopy

Magnetic resonance spectroscopy (MRS) is one of the several modalities of neuroimaging. The other well known modalities of neuroimaging includes computerized tomography (CT), magnetic resonance imaging (MRI), functional MRI (fMRI) and diffusion tensor imaging (DTI). In general, neuroimaging is the process of producing images of structure or activity of the brain or other parts of nervous system. It produces potentially useful clinical tools for structural and functional assessment of psychiatric disorders such as dementia (Malhi and Lagopoulos, 2008).

MRS looks at the chemical composition of the tissue of interest and displays it in a spectrum. MR spectra may be obtained from different nuclei e.g., Protons (^{1}H), phosphorus (^{31}P), fluorine (^{19}F), carbon (^{13}C) and sodium (^{23}Na). Protons (^{1}H) are the most used nuclei for clinical applications in the human brain mainly because of its high sensitivity and abundance. Proton MRS is also known as H-MRS.

The MRS technique (or MRI technique in general) is developed based on following principle: when an atom is placed in an external magnetic field, the spin frequency (also known as *resonance*) of its nuclei are changed and aligned with the direction of magnetic field. In MRS technique, an in-vivo tissue is first placed in big magnetic field and then radio-frequency (RF) waves are turned on and off systematically to yield pulses of energy. This pulse of energy is then captured through a detector coil outside the tissue that can be measured. The pulse of energy received in the detector coil is generally in the form of oscillating sinusoidal curve in time domain and then it is Fourier transformed in frequency domain. A Fourier transformed function is represented by the set of amplitudes corresponding to the set of frequencies. An MR spectrum is obtained by plotting the amplitude (y-axis) against the frequency of nucleus in ppm (x-axis) which looks like curve with sharp bumps. A spectrum from a tissue is displayed in Figure 1.1. More detailed information about spectroscopy is available in many literature including Malhi and Lagopoulos (2008) and Bertholdo et al. (2013).

MRS produces spectrum that contains information about concentration of metabolites those are present in tissue. We can obtain the information about the concentration of several metabolites present in the tissue via some dedicated software such as LCModel (Provencher, 2005). Traditionally MRS has been used as a diagnostic tool in the biochemical characteri-

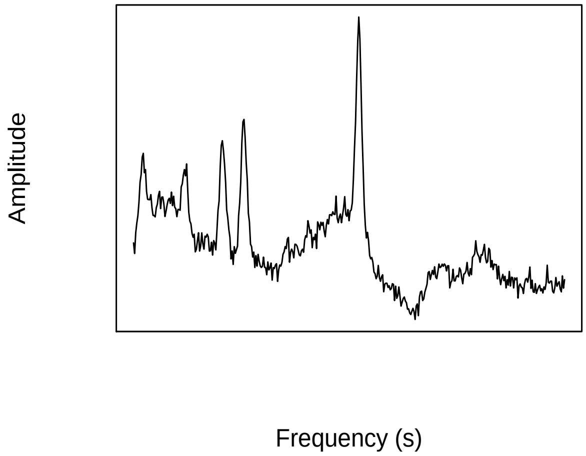

Figure 1.1: An MR spectrum from basal ganglia region of brain.

zation of pathophysiological processes predominantly in the brain such as tumors, abscesses and stroke. However, more recently the researchers have been increasingly applying MRS successfully to several psychiatric disorders.

1.2 External information about observed spectra

A tissue contains a number of metabolites such as Creatine (Cr), Glutamate (Glu), Glucose (Glc), Glycerophosphocholine (GPC), *myo*-Inositol (mI), N-Acetylaspartate (NAA), N-Acetylaspartylglutamate (NAAG), *scyllo*-Inositol (Scyllo) and Taurine (Tau). These metabolites are generated as a bi-product of metabolism in tissue. Each of these metabolites produces a unique spectrum. In other words, two metabolites do not produce spectra of same shape. Spectra associated with some of the metabolites are displayed in Figure 1.2. An observed spectrum from a tissue is mixture of spectra from all of these metabolites present in that tissue, plus spectrum of water and some random noise. The relative contribution of spectrum of individual metabolites on observed spectrum from a particular tissue

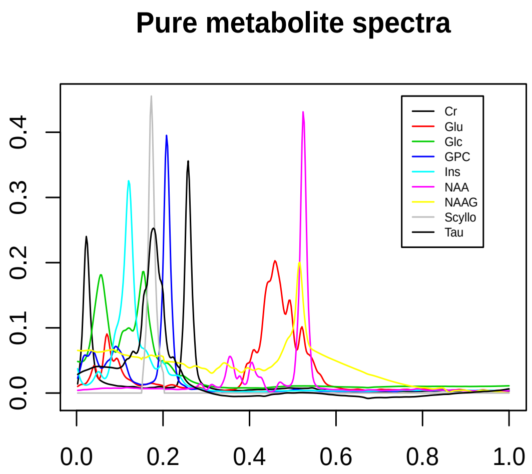

Figure 1.2: MR spectra of the 9 pure metabolite profiles.

depends on its relative concentration in that tissue. A metabolite with greater relative concentration in a tissue is more likely to influence the observed spectrum.

By comparing the two plots in Figure 1.1 and 1.2, we can verify that the observed spectra have their peaks at the locations where at least one of the metabolites has its peak. Since an observed spectrum is mixture of pure metabolites' spectra, we can expect that the observed spectra to lie near a functional subspace spanned by the spectra of pure metabolites. Hence, **we can utilize the spatial information about the spectra of pure metabolites as external information**. This information has been employed in the estimation process in Chapter 2 as discussed in Section 2.5.2.

1.3 Objectives

In this dissertation, we have proposed statistical methods to analyze the data we get from MRS in the context of neurocognitive impairment. Psychiatric disorders are often quantified by some disease marker. Throughout this work, we have considered only continuous and scalar disease marker. For example, often the degree of neuro-psychological impairment is measured by *global deficit score (GDS)* which is scalar and continuous. We are interested in studying a) the progression of psychiatric disorders using entire spectra obtained via MRS and b) the progression of brain metabolites. We are interested in addressing following relevant questions:

Question 1. How does an MR spectrum from brain tissue influence some scalar disease marker longitudinally? Here we have considered entire observed spectrum as our predictor.

Question 2. Is there any baseline factor that influence the longitudinal change of a given metabolite?

The first two objectives (Objective 1 and 2) of this research were formulated to answer these questions. Finally, in objective 3, we we have analyzed the metabolite concentration data in depth from HIV Neuroimaging Consortium (HIVNC) study using the method developed under Objective 3. These objectives are briefly described below and have been detailed in Chapters 2 - 4.

1.3.1 Objective 1: Longitudinal functional regression with structured penalties

Under this objective, we proposed a methodology that enables us to answer the Question 1. In practice, researchers first extract the concentration of the individual metabolites from the observed spectra using dedicated software programs such as LCModel (Provencher, 2005). Then they model some disease marker using the concentration of these metabolites using standard approaches for univariate and multivariate analyses. There are two major problems with this approach: (1) We loose information when we use only concentrations instead of entire spectrum, and (2) The process of extracting concentration of metabolites from spectra is not completely accurate.

Our goal is to use entire MR spectrum to model disease marker. Here response is disease marker which is scalar and continuous. But our predictor is MR spectrum which is a function. This is a scenario where we are interested in modeling a scalar continuous response using predictor function. We cannot apply simple regression technique here; rather, we have to consider functional linear modeling (fLM) that connects a scalar response to a predictor function. Moreover, we had access to data on observed MR spectra and a disease marker, both observed longitudinally. Hence, we are interested in fLM in longitudinal setting. Although fLM have recently been well studied in cross-sectional settings, extensions to longitudinally-collected functions have not received much attention. We have proposed LongPEER method to model scalar outcomes with predictor function via fLM in longitudinal setting. The proposed method is well-suited for the situation where external information is available about the predictor function and can efficiently use that information during estimation. For MR spectra, the external information comes in form of spatial information about pure metabolite spectra as discussed in Section 1.2. The work under this objective is described in Chapter 2.

1.3.2 Objective 2: Construction of regression tree with longitudinal data

Under this objective, we considered both the response and predictor variables as scalar. Here, we seek answer for Question 2: Is there any factor that influence the longitudinal profile of individual metabolite concentrations? Our goal is to analyze the information on the concentration of several metabolites obtained via LCModel. Usually, the change over time is studied via either mixed effects model or marginal models. In practice longitudinal profile of concentrations of metabolites may be influenced by the some demographic and clinical factors such as age, gender, duration of highly active antiretroviral therapy

(HAART), CD4 count. In order to understand the influence of different clinical and demographic variables at baseline on the progression of concentration of a specific metabolites in *fully non-parametric way*, we considered constructing regression tree with longitudinal data.

The major problem with constructing regression tree is the controlling for type I error. This is because typically, in construction of regression tree, the best split is determined by statistical testing at each cut-off point of all the candidate clinical and demographic variables. We call this as a *naive approach.* This naive approach requires large number of statistical tests. Let's assume that we have S partitioning variables as $X_1^{G_1}, \cdots, X_S^{G_S}$ with cut-off points as $G_1, \cdots, G_S$, respectively. In that case, total number of tests would be $\sum_{s=1}^{S}(G_s - 1)$. We have proposed LongCART algorithm for construction of regression tree that involves only single test for each partitioning variable. We call these tests as *test for parameter instability.* Hence, with S partitioning variables, we need to perform only S *tests for parameter instability.* The number of tests with the proposed approach would be much smaller than $\sum_{s=1}^{S}(G_s - 1)$ in presence of continuous partitioning variables and/or categorical partitioning variables with greater than two categories. Consequently, it would be much easier to control type I error for S tests with some adjustment for multiplicity compared to $\sum_{s=1}^{S}(G_s - 1)$ number of tests. Further, the proposed LongCART algorithm is faster than the naive approach and can be extended in more complicated situation. The work under this objective is described in Chapter 3.

1.3.3 Objective 3: Identifying factors influencing longitudinal changes of brain metabolites in HIV-infected subjects enrolled in HIVNC study

We have considered MRS assessments from 243 HIV patients on antiretroviral regimen enrolled in HIV Neuroimaging Consortium (HIVNC) study. HIVNC was formed to examine pattern or extent of brain injury in chronically infected patients on antiretroviral (ARV) treatment. We have studied the influence of different demographic and clinical factors on the longitudinal change of brain metabolite concentrations using the methods developed under Objective 2. Baseline was determined as the date of enrollment and only the observations within 3 years from baselines were included. The individuals included in the study had at least one post-baseline measurement within 3 years. Five metabolites were considered: creatine (Cr), N-acetylaspartate (NAA), choline (Cho), myo-inositol (MI), and glutamate and glutamine (Glx) in the basal ganglia, frontal white matter, and mid-frontal cortex.

1.4 MRS data for the application of proposed methods

Our work is motivated by HIV Neuroimaging Consortium (HIVNC) study where magnetic resonance (MR) spectra have been collected longitudinally from late stage HIV patients (Harezlak et al., 2011). The study cohort was comprised of chronically HIV-infected patients enrolled in a longitudinal study of HIV associated brain injury at the following sites: University of California (San Diego), University of California (Los Angeles), Harbor-UCLA, Stanford University, University of Colorado, University of Pittsburgh, and Rochester University. Details of the inclusion and exclusion criteria and MRS for this cohort have been described elsewhere (Harezlak et al., 2011). The methods developed under Objectives 1 and 2 have been illustrated in this data. The metabolite concentration data collected from this study have been analyzed in depth to study the influence of baseline factors on the longitudinal change of brain metabolites under Objective 3.

Chapter 2

Longitudinal functional regression model with structured penalties

The term functional data analysis (FDA) was first introduced by Ramsay and Dalzell (1991) in the statistical literature. FDA goes one big step further from multivariate data analysis. In multivariate data analysis, we concerned with data in the form of random vectors, whereas in FDA we analyze functions, such as curves, shapes and images (Müller, 2005). FDA is becoming increasingly common than ever before due to technological advancements and increased availability of storage of large datasets, even in longitudinal setting. One of the various interesting aspect of FDA is the extension of the notion of linear model towards functional context. We can incorporate function in linear model in following ways (Müller, 2005; Ramsay and Silverman, 1997)

1. The dependent or response variable is scalar, but the predictor is function
2. The dependent or response variable is function, but the predictor(s) is (are) scalar
3. Both the dependent and predictor variables are function

Our interest is in the first one that relates scalar response to a predictor function. Throughout this dissertation, by functional linear model (fLM), we refer only to the models connecting scalar continuous response with a predictor function. In the cross-sectional setting, there have been many proposed methods to fit fLM. One of the approach is "Partially empirical eigenvectors for regression" (PEER) that allows to exploit the external information about the predictor function (Randolph et al., 2012). In this chapter, we have extended this approach to longitudinal setting.

2.1 Functional data

Functional data is represented by set of measurements obtained via observing the value of a function at several sampling points in its domain. Let $W(s)$ be any function in domain Ω. Suppose we have observed $W(s)$ at the discretized sampling points $s_1 < \cdots < s_p$ and measurements are denoted by $w_1, \cdots, w_p$. Then $w \equiv \{w_1, \cdots, w_p\}$ will be considered as *functional data* representing the function $W(\cdot)$. The functional data (i.e. observed data at discretized samping point from a function) might seem as multivariate data, but there are important distinction. First, functional data should be thought of as sequence of observations from a single entities, rather than merely a sequence of individual observations. Second, the order of the observation in the functional data is important. According to Müller (2005), functional data is multivariate data with an ordering on the dimensions. Third, a multivariate data represents the data which are in finite dimension, but functional data represents a function which is *infinite dimensional.* Finally and most importantly, the term *functional* in reference to functional data refers to the intrinsic structure of the data. Multivariate data lacks such kind of intrinsic structure (Ramsay and Silverman, 1997). We cite here two examples of functional data.

Example 1. Suppose we have measurement of height of an individual at 10 different ages. Then these measurements can be treated as *functional data* because these measurements along with the age of measurements represents the function $height(age)$.

Example 2. The output of MRS from a tissue is represented by the measurements on *amplitude* at several frequency levels. The set of measurements on amplitude along with the frequency level should be considered as *functional data* because these observations are sampled from the function $W(s)$, where s denotes the frequency and $W(s)$ denotes the amplitude at frequency s.

2.2 Functional linear model

The cross-sectional fLM with scalar response y and predictor function $W(\cdot)$ can be stated as follows (see e.g., Yao and Müller, 2010)

$$y_i = \beta + \int_{\Omega} W_i(s)\gamma(s)ds + \epsilon_i \tag{2.2.1}$$

where i is the index for subject, β is the usual intercept term that adjusts for the origin of the y, Ω denotes the domain of the predictor functions $W(s)$, $s \in \Omega$, and $\gamma(s)$ is a square integrable function that models the linear relationship between the predictor function and scalar response. In addition, $\epsilon_i \sim N(0, \sigma_{\epsilon}^2)$. Here, $\int_{\Omega} W_i(s)\gamma(s)ds$ is the subject specific functional effect. The regression function, $\gamma(s)$, can be thought of as the weighting function that weights information within the $W_i(s)$ across the values of s (Ramsay and Silverman, 1997). We assume that $\gamma(\cdot) \in L^2(\Omega)$.

As there is no unique $\gamma(\cdot)$ that solves the equation (2.2.1), additional regularization or constraint is required. Typically, some form of smoothness is imposed on $\gamma(\cdot)$, one approach being to expand both regression function $\gamma(\cdot)$ and predictor function $W(\cdot)$ in terms of a set of spline basis functions such as B-splines and then obtain the regularized estimate of $\gamma(\cdot)$ (Ramsay and Silverman, 1997). Another approach is to express the regression function $\gamma(\cdot)$ in terms of the orthonormal eigenfunctions of covariance of $W(\cdot)$ using Karhunen-Loève (K-L) basis expansion (see e.g., Müller, 2005). A third approach is to combine the above two approaches, known as penalized functional regression (PFR) approach (Goldsmith et al., 2011). In PFR approach, a spline basis is used to represent the regression function and a basis of eigenfunctions from the set of predictors is used to represent each $W(\cdot)$. Another approach is to use wavelet basis, instead of eigenfunctions, to represent the predictor functions (Morris and Carroll, 2006). Here, we have adopted "Partially empirical eigenvectors for regression" (PEER) approach by Randolph et al. (2012) which does not begin by explicitly projecting onto a pre-specified basis of functions. Instead, the estimate is obtained by projecting onto a space determined jointly by the covariance function of $W(\cdot)$ and the *decomposition-based penalty. Decomposition-based penalty* provides a mean to incorporate external information into the estimation process and this is discussed in Section 2.5.2.

2.3 PEER estimation in fLM

PEER approach exploits the familiar framework of penalized least-squares regression by imposing a scientifically-informed quadratic penalty term into the estimation process. The resulting estimate is a function that is represented by a set of "partially empirical" eigenvectors that arise from a joint eigen-basis decomposition of the discretized predictor functions and the penalty term; see (Randolph et al., 2012).

Let consider $\Omega = [0, 1]$, a closed interval in $\mathbb{R}$ and let $W(\cdot)$ denotes a random function in $L^2(\Omega)$. Consider $W(\cdot)$ as a predictor function of interest and $W_i(\cdot)$ is the predictor function associated with the i^{th} subject $(i = 1, \ldots, N)$. Technically, we can observe an idealized predictor function only at finite sampling points. We will assume that each observed predictor function is sampled equally at p sampling points, $s_1, \ldots, s_p \in [0, 1]$, with sampling that is appropriately regular and dense enough to capture informative spatial structure, as seen, for instance, in the MRS data in Section 1.1. Let $w_i := [w_i(s_1), \cdots, w_i(s_p)]^\top$ be the $p \times 1$ vector of values sampled from the realized function $W_i(\cdot)$. Clearly, as explained in Section 2.1, w_i is *functional data* because it represents the function $W_i(\cdot)$. The observed data are of the form $\{y_i; w_i\}$, where y_i is a scalar outcome, and w_i is the vector of values sampled from the realized predictor function from i^{th} subject. When $s_1, \cdots s_p$ are equi-spaced, (2.2.1) can be approximated as

$$y = \beta + W\gamma + \epsilon.$$

where, $y = [y_1, \cdots, y_N]^\top$, $W = [w_1^\top, \cdots, w_N^\top]^\top$, and $\gamma = [\gamma(s_1), \ldots, \gamma(s_p)]^\top$. One of the major advantage of PEER approach is its ability to incorporate external information about predictor function $W(\cdot)$. This is done via using *decomposition-based penalty* as discussed below.

2.3.1 Decomposition-based penalty

Suppose we have prior information available that $W(\cdot) \in \mathcal{Q}$ where $\mathcal{Q} \subset L^2(\Omega)$. Further, assume that $\mathcal{Q}$ is the space spanned by the basis functions $q_1(\cdot), \cdots q_j(\cdot), \cdots, q_J(\cdot)$. For example, in Section 1.2, we have seen that the spectra of pure metabolites provide *external information* about observed MR spectra meaning that we expect observed MR spectra to lie on or near the subspace spanned by the spectra of these pure metabolites. Hence, when our predictor function, $W(\cdot)$, is observed MR spectrum, one can determine $\mathcal{Q}$ considering each $q_j(\cdot)$ as spectrum of individual pure metabolites.

The prior information that $W(\cdot) \in \mathcal{Q}$ may be implemented by encouraging the estimate to be on or near a subspace, $\mathcal{Q}$. In PEER approach, this is done via use of *decomposition-based penalty.* Assuming that $W(\cdot)$ and $q_j(\cdot)$'s are observed only at p sampling points, we represent $\mathcal{Q}$ by the range of a $p \times J$ matrix Q whose columns are $q_1, \ldots, q_J$. Here q_j is the vector of the observations obtained from $q_j(\cdot)$ at the p sampling points. Then *decomposition-based penalty* can be defined as (Randolph et al., 2012)

$$L_Q = \alpha_0 P_Q + \alpha_1 (I - P_Q) \tag{2.3.1}$$

for some scalars α_0 and α_1. $P_Q = QQ^+$ is the orthogonal projection onto the Range(Q), where Q^+ is Moore-Penrose inverse of Q.

The purpose of the *decomposition-based penalty* is to encourage the estimate of regression function to be on or near the space $\mathcal{Q}$. To see how L_Q works, let $\tilde{\gamma}$ be any estimate of γ. When $\tilde{\gamma} \in Sp(Q)$, we have $L_Q\tilde{\gamma} = \alpha_0\tilde{\gamma}$, but when $\tilde{\gamma} \notin Sp(Q)$, we have $L_Q\tilde{\gamma} = \alpha_1\tilde{\gamma}$. The condition $\alpha_1 > \alpha_0$ imposes more penalty for $\tilde{\gamma} \notin P_Q$ compared to when $\tilde{\gamma} \in P_Q$. The weights α_1 and α_0 determine the relative strength of emphasizing $\mathcal{Q}$ in the estimation process. Note, in particular, that taking $\alpha_1 = \alpha_0$ results in a ridge estimate and that L_Q is invertible, provided α_1 and α_0 are non-zero.

2.3.2 Estimation

Let L be any penalty matrix to be used for estimation of $\gamma = [\gamma(s_1), \cdots, \gamma(s_p)]^\top$. It could be either *decomposition-based penalty* or some non-informative penalty such as, smoothness penalty or ridge penalty. The estimate of β and γ are obtained minimizing

$$(y - \beta - W\gamma)^\top (y - \beta - W\gamma) + \lambda^2 (L\gamma)^\top (L\gamma) \tag{2.3.2}$$

where λ^2 is the tuning parameter. A generalized ridge estimate of γ based on minimizing the above expression is obtained as (see e.g., Ruppert et al., 2003, p. 66)

$$\hat{\gamma} = [W^\top W + \lambda^2 L^\top L]^{-1} W^\top y$$

Randolph et al. (2012) also shown that the above estimate of γ can be expressed as the generalized singular vectors determined by the generalized singular vector decomposition (e.g. see Paige and Saunders, 1981; Van-Loan, 1976) of W and L. Therefore, PEER estimates of γ is obtained by focusing on a subspace that is spanned by a basis of generalized singular vectors which arise jointly from the predictors, W and the penalty, L. For details about PEER estimation, please refer to Randolph et al. (2012). The estimation process is implemented in `peer()` in `refund()` package in `R`.

2.4 Longitudinal functional linear model

The problem we address involves repeated observations from each of N subjects. At each observation time, t, we collect data on a scalar response variable, y, and a (idealized) predictor function, $W(\cdot)$. We are interested in longitudinal functional linear models of the following form:

$$y_{it} = x_{it}^{\top}\beta + \int_0^1 W_{it}(s)\gamma(t,s)ds + z_{it}^{\top}b_i + \epsilon_{it} \tag{2.4.1}$$

Here $\gamma(t,\cdot)$ denotes the regression function at time t, x_{it} is a vector of scalar-valued (non-functional) predictors. Further, $\epsilon_{it} \sim N(0,\sigma_\epsilon^2)$ and b_i is the vector of r random effects pertaining to subject i and distributed as $N(0,\Sigma_{b_i})$. We want to estimate $\gamma(t,\cdot)$ along with β in PEER framework.

2.4.1 Literature review

The cross-sectional fLM with scalar response has been a focus of various investigations (Cai and Hall, 2006; Cardot et al., 2007, 1999, 2003; Fan and Zhang, 2000; Faraway, 1997; Ramsay and Silverman, 1997; Reiss and Ogden, 2009). Some of these methods estimate regression functions in two steps. For example, principal components regression (PCR) estimates of the regression function are obtained first and then these PCR estimates are projected onto a B-spline basis (Cardot et al., 2003) or vice-versa; i.e., PCR fitting is performed only after projection onto B-splines (Reiss and Ogden, 2009). Extensions of fLM have been made towards generalized linear model with predictor function (James, 2002;

Müller and Stadtmüller, 2005) and quadratic functional regression (Yao and Müller, 2010). Another class of models, known as Functional Analysis of Variance (FANOVA), decompose repetitively-observed functional predictors into several (fixed and random) groups and subject-specific component functions (Brumback and Rice, 1998; Di et al., 2009; Greven et al., 2011; Guo, 2002). However, it is important to distinguish the proposed work from FANOVA methods which do not relate predictor function(s) to the scalar response in the longitudinal setup, the way we did it in this paper. Although these models have recently been well studied, extensions to longitudinally-collected functions have not received much attention. To our knowledge, LPFR (Goldsmith et al., 2012) and LFPCR (Gertheiss et al., 2013) are the only published methods addressing regression estimation in the longitudinal functional predictor framework.

2.4.2 LPFR and LFPCR approaches

The LPFR approach assumes the regression function in (2.4.1) is independent of time and proceeds in three steps: uses a truncated set of K-L vectors to represent the predictor functions; expresses the regression function using a spline basis; and fits the longitudinal model using an equivalent mixed-model framework that incorporates subject-specific random effects. In LFPCR approach, first, the predictor function is decomposed into visit- and subject- specific functions according via longitudinal functional principle component analysis (LFPCA) (Greven et al., 2011) and then longitudinal analysis is carried out in second step with the outcome of LFPCA. Both LPFR and LFPCR assume that the regression function, $\gamma(t, \cdot)$ remains constant over time. Due to this restrictive assumption, LPFR and LFPCR are not suited for situations in which the association between the predictor function and scalar response may evolve over time.

2.4.3 Proposed method

We have extended the scope of the PEER approach to estimate $\gamma(t, \cdot)$ in Eq. (2.4.1) in a manner that allows the estimated regression function to vary with time. The extension of PEER framework to the longitudinal setting has two major advantages: 1) the regression function is allowed to vary over time; and 2) external or a priori information about the structure of the regression function can be incorporated directly into the estimation process. We have formulated the estimation procedure within a mixed-model framework making the method computationally efficient and easy to implement. We call our proposed approach as LongPEER.

In the following section we have described LongPEER method of estimation for longitudinal fLM. This includes discussion on how external information can be implemented in the LongPEER estimation process via *decomposition-based penalty* in Subsection 2.5.2. In Section 2.6.1, we have presented how these estimates can be obtained as best linear unbiased predictors (BLUP) through mixed model equivalence. Expressions for the precision of the estimates are derived in Section 2.6.2. In the appendix of this chapter, we have presented present how our longitudinal generalized ridge estimate, along with its bias and precision, can be obtained in terms of generalized singular (GS) vectors under weak assumptions.

Numerical illustrations are provided in Section 2.7. In particular, the simulation in Section 2.7.1 compares LPFR with the method proposed in this paper. The simulation in Section 2.7.2 evaluates the influence of sample size and relative contributions of prior spatial information on the proposed method using a decomposition-based penalty. Confidence band's coverage probabilities are explored through a simulation in Section 2.7.3. The performance of PEER estimate when only partial information is available are evaluated in

Section 2.7.4. We analyze the MRS data using our method and summarize our findings in Section 2.8. The methods discussed in this paper have been implemented in the `refund` package (Ciprian Crainiceanu et al., 2012) in `R` in the `lpeer()` functions. Throughout the presentation we consider a single functional predictor.

2.5 LongPEER estimation in longitudinal fLM

As before in Section 2.3, we consider $\Omega = [0, 1]$, a closed interval in $\mathbb{R}$ and let $W(\cdot)$ denotes a random function in $L^2(\Omega)$. In addition, we use t to indicate the time of measurement. Let $W_{it}(\cdot)$ denotes a predictor function from the i^{th} subject $(i = 1, \dots, N)$ at the t^{th} timepoint $(t = t_1, \dots, t_{n_i})$.

Technically, we can observe an idealized predictor function only at finite sampling points. We will assume that each observed predictor function is sampled equally at p sampling points, $s_1, \dots, s_p \in [0, 1]$, with sampling that is appropriately regular and dense enough to capture informative spatial structure, as seen, for instance, in the MRS data in Section 1.1. Let $w_{it} := [w_{it}(s_1), \cdots, w_{it}(s_p)]^\top$ be the $p \times 1$ vector of values sampled from the realized function $W_{it}(\cdot)$ at p sampling points $S = s_1, \cdots, s_p$. Clearly, as explained in Section 2.1, w_{it} is *functional data* because it represents the function $W_{it}(\cdot)$.

The observed data are of the form $\{y_{it}; x_{it}; w_{it}\}$, where y_{it} is a scalar outcome, x_{it} is a $K \times 1$ column vector of measurements on K scalar predictors, and w_{it} is the vector of values sampled from the realized predictor function from i^{th} subject at time t. Denoting the true regression function at time t by $\gamma(t, \cdot)$, the longitudinal functional regression outcome model of interest is

$$y_{it} = x_{it}^\top \beta + \int_0^1 W_{it}(s)\gamma(t, s)ds + z_{it}^\top b_i + \epsilon_{it} \tag{2.5.1}$$

where, $\epsilon_{it} \sim N(0, \sigma_\epsilon^2)$ and b_i is the vector of r random effects pertaining to subject i and distributed as $N(0, \Sigma_{b_i})$. As usual we assume that z_{it} is a subset of x_{it}, ϵ_{it} and b_i are independent, ϵ_{it} and $\epsilon_{i't'}$ are independent whenever $i \neq i'$ or $t \neq t'$ or both, and b_i and $b_{i'}$ are independent if $i \neq i'$. Here $x_{it}^\top \beta$ is the standard fixed effect from K univariate predictors, $z_{it}^\top b_i$ is the standard random effect and $\int_0^1 W_{it}(s)\gamma(t,s)ds$ is the subject/time specific functional effect. We assume that $\gamma(t,\cdot) \in L^2(\Omega)$, for all t.

When the association between predictor function and response changes over time, the regression function $\gamma(t,s)$ varies over both spatial and time domain. For example, $\gamma(t,s)$ may vary linearly with time, $\gamma(t,s) = \gamma_0(s) + t\gamma_1(s)$, or quadratically, $\gamma(t,s) = \gamma_0(s) + t\gamma_1(s) + t^2\gamma_2(s)$. This is in a spirit similar to a linear mixed effects model with linear or quadratic time slope (see e.g., Fitzmaurice and Ware, 2004). In general, we assume that $\gamma(t,s)$ can be decomposed into several time-invariant component functions $\gamma_0(s), \cdots, \gamma_D(s)$ as

$$\gamma(t,s) = \gamma_0(s) + f_1(t)\gamma_1(s) + \cdots + f_D(t)\gamma_D(s)$$

where, $f_1, \ldots, f_D$ are D prescribed linearly independent functions of t and $f_d(0) = 0$ for all d; the time component t enters into $\gamma(t,s)$ through these terms. At $t = 0$, $\gamma(t,s)$ reduces to $\gamma_0(s)$ and has the obvious interpretation of a baseline regression function pertaining to the sampling points s. When $D = 0$, $\gamma(t,s) \equiv \gamma_0(s)$ is independent of t, a situation considered by Goldsmith et al. (2012). In general, each f may be any function of t with $f(0) = 0$, e.g., $f(t) = t$ or $t\exp(t)$. We can rewrite the equation (2.5.1) as

$$y_{it} = x_{it}^\top \beta + \int_0^1 W_{it}(s)\{\gamma_0(s) + f_1(t)\gamma_1(s) + \cdots + f_D(t)\gamma_D(s)\}ds + z_{it}^\top b_i + \epsilon_{it}$$

In a PEER approach, the dependence of y_{it} on $W_{it}(\cdot)$ is seen as a linear dependence on observations at p sampling points, w_{it}; spatial (functional) structure is imposed directly into the estimation of $\gamma_d = [\gamma_d(s_1), \ldots, \gamma_d(s_p)]^\top$, for $d = 0, \ldots, D$. Combining all $n_\bullet = \sum_{i=1}^{N} n_i$ observations from the N subjects obtained across all time points, we express the model as

$$y = X\beta + W\gamma + Zb + \epsilon. \tag{2.5.2}$$

Here, $y = [y_{1t_1}, \cdots, y_{1t_{n_1}}, \ldots, y_{1t_N}, \ldots, y_{Nt_{n_N}}]^\top$ is a $n_\bullet \times 1$ vector of all responses, $X = [x_{1t_1}^\top, \cdots, x_{1t_{n_1}}^\top, \cdots, x_{1t_N}^\top, \cdots, x_{Nt_{n_N}}^\top]^\top$ is an $n_\bullet \times K$ design matrix pertaining to K univariate predictors, β is the associated coefficient vector, $\gamma = [\gamma_0^\top, \gamma_1^\top, \cdots, \gamma_D^\top]^\top$ is a $(D+1)p \times 1$ vector of functional coefficients, W is the corresponding $n_\bullet \times (D+1)p$ design matrix. Further, b is the $rN \times 1$ vector of random effects and Z is the corresponding $n_\bullet \times rN$ design matrix. The matrix W has the structure

$$W = \begin{bmatrix} W_1 \\ \vdots \\ W_N \end{bmatrix} \qquad W_i = \begin{bmatrix} w_{it_1}^\top & f_1(t_1) w_{it_1}^\top & \cdots & f_D(t_1) w_{it_1}^\top \\ \vdots & \vdots & \ddots & \vdots \\ w_{it_{n_i}}^\top & f_1(t_{n_i}) w_{it_{n_i}}^\top & \cdots & f_D(t_{n_i}) w_{it_{n_i}}^\top \end{bmatrix}$$

2.5.1 Generalized ridge estimate

The formal model in (2.5.1) is ill-posed and has no unique solution for γ. Common approaches to estimate a regression function in a fLM involve reducing dimension by projecting onto a subspace defined by a few K-L (empirical) eigenvectors or onto the span of a set of spline basis functions. Alternatively, our use of a generalized ridge penalty constrains the estimation of γ in the spatial (or s) dimension without preliminary smoothing or explicit dimension reduction. The process encourages structure of a particular type via the choice of penalty operator. In the longitudinal (or t) dimension, γ is more explicitly and severely constrained by the choice of $f_1, \ldots, f_D$.

The model of interest described in the previous Section can be written as follows:

$$y = X\beta + W\gamma + \epsilon^* \tag{2.5.3}$$

where $\epsilon^* = Zb + \epsilon \sim N(0, V)$ and $V = Z\Sigma_b Z^\top + \sigma_\epsilon^2 I$. Further, assume L_d be the penalty operator for γ_d and λ_d^2 be the associated tuning parameter, $\forall\ d = 0, \ldots, D$. Then the penalized estimates of β and γ can be obtained by minimizing

$$||y - X\beta - W\gamma||^2_{V^{-1}} + \lambda_0^2 ||\gamma_0||^2_{L_0^\top L_0} + \cdots + \lambda_D^2 ||\gamma_D||^2_{L_D^\top L_D} \tag{2.5.4}$$

Here we have used the notation $||a||^2_B = a^\top B a$, where B is a symmetric, positive definite matrix. A generalized ridge estimate of β and γ based on minimizing the above expression is obtained as (see e.g., Ruppert et al., 2003, p. 66)

$$\begin{bmatrix} \hat{\beta} \\ \\ \hat{\gamma} \end{bmatrix} = (C^\top V^{-1} C + D)^{-1} C^\top V^{-1} y \tag{2.5.5}$$

where, $C = [X \;\; W]$, $D = \text{blockdiag}\{0, L^\top L\}$ and $L = \text{blockdiag}\{\lambda_0 L_0, \cdots, \lambda_D L_D\}$.

In Randolph et al. (2012), a PEER estimate and its mean squared error (MSE) were established in terms of generalized singular (GS) vectors as obtained through generalized singular value decomposition (GSVD); the setting was that of the cross-sectional fLM with a single predictor function and no random effects. For additional background on the GSVD and its computation, see Bjorck (1996); Golub and Van-Loan (1996); Paige and Saunders (1981); Van-Loan (1976). In the Appendix of this chpater, we derive an expression for the generalized ridge estimate $\hat{\gamma}$ explicitly in terms of the GSVD components.

2.5.2 Incorporating prior information via decomposition-based penalty

Our goal is to estimate γ imposing some presumed functional structure. In other words, the aim is to supplement, not necessarily smooth, the predictor functions with knowledge about spatial structure in a mathematically tractable way. A common approach to incorporate spatial structure into the functional regression model is to use the strongest structure from the predictors by considering only first few K-L vectors (Cardot et al., 2003; Hall et al., 2001). However, we will incorporate spatial structure through an informed choice of penalty operator as proposed by Randolph et al. (2012).

Let $\hat{\gamma}_d \equiv \hat{\gamma}_{L_d,\lambda_d}$ be the estimate of γ_d obtained from the penalty operator L_d and tuning parameter λ_d^2, for each $d = 0, \ldots, D$. For example, L_d may denote I_p (a ridge penalty) or a second-order derivative penalty (giving an estimate having continuous second derivative). Alternatively, with prior knowledge about potentially-relevant structure in a regression function, a *decomposition-based penalty* can be defined in terms of a subspace defined by such structure (Randolph et al., 2012). To be precise, if it is appropriate to impose scientifically-informed constraints on the "signal" being estimated by $\gamma_d, d = 0, \cdots, D$, this prior may be implemented by encouraging the estimate to be in or near a subspace, $\mathcal{Q} \subset L^2(\Omega)$. This can be done by using *decomposition-based penalty* P_Q as defined in equation (2.3.1).

As an example, when our predictor function is observed spectra obtained from a tissue via Magnetic Resonance Spectroscopy (MRS), one can use the spatial information of pure metabolites as a prior information as discussed in Section 1.2. In that case, $\mathcal{Q}$ would be the subspace spanned by the spectra of available pure metabolites. If we have J such pure metabolites and all of these metabolites are sampled at p points, then $\mathcal{Q}$ will be represented by $p \times J$ matrix Q. The j^{th} column of Q representing the observations at p sampling points corresponding to spectrum of j^{th} metabolites.

2.5.3 Selection of time-structure in $\gamma(t, \cdot)$

The proposed approach allows us to choose very flexible time structure for $\gamma(t, \cdot)$ during estimation; however, in practice, we don't have the information available about its time structure. As an example, we don't know whether $\gamma_0(t, \cdot) + t\gamma_1(t, \cdot)$ is sufficient or we need $\gamma_0(t, \cdot) + t\gamma_1(t, \cdot) + t^2\gamma_2(t, \cdot)$. The problem of choosing appropriate time-structure in $\gamma(t, \cdot)$ ($\gamma_0(\cdot) + t\ \gamma_1(\cdot)$ or $\gamma_0(\cdot) + t\ \gamma_1(\cdot) + t^2\ \gamma_2(\cdot)$) is similar, in principle, to the problem of choosing the time structure in the linear mixed effects model (e.g., $E(y_{it}|b_i) = \beta_0 + \beta_1\ t$ or $E(y_{it}|b_i) = \beta_0 + \beta_1\ t + \beta_2\ t^2$). We can think of at least two ways to decide about the form of unknown regression function: (a) Use of AIC to compare different structures, and (b) Use of the point-wise confidence band for the $\gamma_0(\cdot)$, $\gamma_1(\cdot)$ and $\gamma_2(\cdot)$. If the confidence band for $\gamma_2(\cdot)$ encloses 0 in the entire region, then probably it is not good idea to consider $\gamma_0(t, \cdot) + t\gamma_1(t, \cdot) + t^2\gamma_2(t, \cdot)$; rather we should continue with $\gamma_0(t, \cdot) + t\gamma_1(t, \cdot)$ only.

2.5.4 Selection of ϕ_a and ϕ_b for *decomposition based penalty*

We view ϕ_a and ϕ_b as weights in a tradeoff between preferred and non-preferred subspaces and assume $\phi_a \cdot \phi_b = constant$. In the current implementation, we use REML to estimate λ_d's for a fixed value of ϕ_a, and do a grid search over the ϕ_a values to jointly select the tuning parameters which maximize the REML criterion such as AIC.

2.6 Mixed model representation

Estimates of β and γ obtained by minimizing the expression in equation (2.5.4) correspond to a generalized ridge estimate. In this section we aim to construct an appropriate mixed model that minimizes the expression in equation (2.5.4). In general, the penalty, L, is not required to be invertible but for simplicity this will be assumed here. The mixed model approach provides an automatic selection of tuning parameters. REML-based estimation of the tuning parameters has been shown to perform as well as the other criteria and under certain conditions it is less variable than GCV-based estimation (Reiss and Ogden, 2009).

2.6.1 Estimation of parameters

Using Henderson's (1950) justification [(Henderson, 1950)], one can show that the model $y = X\beta + W\gamma + \epsilon^*$ where, $\epsilon^* \sim N(0, V)$ and $\gamma_d \sim N(0, \frac{1}{\lambda_d^2}(L_d^\top L_d)^{-1})$, for each $d = 0, \ldots, D$, minimizes the expression in equation (2.5.4) to obtain the BLUP. Thus the generalized ridge estimate of β and γ correspond to the BLUP from the following model:

$$y = X\beta + W^*\gamma^* + \epsilon$$

where, $W^* = [W\ Z]$, $\gamma^* = [\gamma^\top\ b^\top]^\top \sim N[0, \Sigma_{\gamma^*}]$ and $\epsilon \sim N(0, \sigma_\epsilon^2 I)$ with

$$\Sigma_{\gamma^*} = \text{blockdiag}\{(L^\top L)^{-1}, \quad \Sigma_b\} \qquad \Sigma_b = \text{blockdiag}\{\Sigma_{b_1}, \cdots, \Sigma_{b_N}\}$$

This representation allows us to estimate fixed and functional predictors simply by fitting a linear mixed model (e.g., using the `lme()` of the `nlme` package in `R` or `PROC MIXED` in `SAS`).

2.6.2 Precision of estimates

Our ridge estimate is the BLUP from equivalent mixed model, hence the variance of the estimate depends on whether the parameters are random or fixed. Randomness of γ is a device used to obtain the ridge estimate while ϵ and b in our case are truly random. With this argumentation, it can be advocated that variance of estimates to be conditional on γ, but not on b (Ruppert et al., 2003). BLUP of β, γ and b can be expressed as (see e.g., Robinson, 1991; Ruppert et al., 2003):

$$\tilde{\beta} = \left(X^\top V_1^{-1} X\right)^{-1} X^\top V_1^{-1} y \qquad \tilde{\gamma} = (L^\top L)^{-1} W^\top V_1^{-1}(y - X\tilde{\beta})$$

$$\tilde{b} = \Sigma_b Z^\top V_1^{-1}(y - X\tilde{\beta})$$

where, $V_1 = V + W(L^\top L)^{-1}W^\top$. $\tilde{\beta}$ is an unbiased estimator of β, but $\tilde{\gamma}$ is not unbiased. It is trivial to see that $\text{Cov}(y|\gamma) = V$. Thus, the variances of $\tilde{\beta}$ and $\tilde{\gamma}$, conditional on γ, are:

$$\text{Cov}(\tilde{\beta}|\gamma) = \left(X^\top V_1^{-1} X\right)^{-1} X^\top V_1^{-1} V V_1^{-1} X \left(X^\top V_1^{-1} X\right)^{-1}$$

$$\text{Cov}(\tilde{\gamma}|\gamma) = A_\gamma V A_\gamma^\top \qquad A_\gamma = (L^\top L)^{-1} W^\top V_1^{-1}\{V_1 - X(X^\top V_1^{-1} X)^\top\} V_1^{-1} \tag{2.6.1}$$

To obtain the unconditional variance, we need to replace V by V_1 in the above expressions but this is will overestimate the variance of the estimates. The expressions for predicted value of y and its variance are:

$$\tilde{y} = X\tilde{\beta} + W\tilde{\gamma} + Z\tilde{b} \qquad \text{Cov}(\tilde{y}|\gamma) = A_y V A_y^\top$$

where,

$$A_y = [\{V_1 - WL^\top LW - Z\Sigma_b Z^\top\}^{-1} X \left(X^\top V^{-1} X\right)^{-1} X^\top V^{-1} + WL^\top LW^\top + Z\Sigma_b Z^\top] V_1^{-1}$$

Let, $T = [1 \ f_1(t) \ \cdots \ f_d(t)] \otimes I_K$. Then the discretized version of regression function at time t is $\gamma_{(t)} = [\gamma(t, s_1), \cdots, \gamma(t, s_K)] = T\gamma$. Therefore, the estimate of $\gamma_{(t)}$ is $\tilde{\gamma}_{(t)} = T\tilde{\gamma}$ and the estimate of its variance is $T\text{Cov}(\tilde{\gamma}|\gamma)T^\top$.

2.7 Simulation

We pursue simulations to evaluate the properties of the LongPEER method. The first simulation study (Section 2.7.1) compares the performance of the LongPEER method with the LPFR approach. In the remaining simulation studies, only the LongPEER method is considered. The purpose of the second simulation study is to evaluate the influence of sample size and the contribution of prior information about the spatial structure (as determined by two different tuning parameters ϕ_a and ϕ_b in equation 2.3.1) on the LongPEER estimate. In the third simulation study, we evaluate the coverage probabilities of the confidence bands constructed using the formula presented in Section 2.6.2. Finally, we evaluate the performance of LongPEER estimate when information on some features is missing and the results are summarized in Section 2.7.4. In all the simulation studies, the simulated predictor functions resemble the MRS data. All results summarized in this Section are based on 100 simulated datasets.

For each subject and visit, predictor functions were simulated independently. Predictor functions are"bumpy" curves with bumps at some pre-specified locations. White noise was added to the predictor functions to account for the instrumental measurement noise. Bumpy regression functions were generated with bumps at some (but, not all) of the bump locations of the predictor function. For the simulation in Section 2.7.1, the regression function is assumed to be independent of time, whereas it varies with time in simulation in Section 2.7.2. For both the predictor and regression functions, 100 equi-spaced sampling points in [0,1] are used.

For the decomposition based penalty, the matrix L_d is defined as follows: 1) select the discretized functions $q_j, j = 1, \dots, J$ spanning the "non-penalized-subspace" and 2) compute $L_d = QQ^+$, where $Q = \text{col}[q_1, \dots, q_J]$. Vectors q_j are discretized functions defined to be zero except at one bump corresponding to a region in the simulated predictor functions.

Estimation error was summarized in terms of the mean squared error (MSE) of the estimated regression function defined as $||\gamma - \tilde{\gamma}||^2$, where $\tilde{\gamma}$ denotes the estimate of γ. Further, MSE was decomposed into the trace of the variance and squared norm of bias. We also calculated sum of squares of prediction error (SSPE) as $||y - \tilde{y}||^2/N$, where $\tilde{y}$ denotes the estimate of the true (noiseless) y. The estimates based on the proposed methods, including the LongPEER estimate, were obtained as BLUPs from the mixed model formulation described in Section 2.6.1.

2.7.1 Comparison with LPFR

As previously stated, LPFR estimates a regression function that does not vary with time. Therefore, in the first set of simulations, we generated the outcomes using a time-invariant regression function (i.e., $\gamma(t, s) = \gamma_0(s)$, for all t). The following model was used to generate the outcome data for 100 individuals ($i = 1, \cdots, 100$), each at 4 timepoints ($t = 0, 1, 2, 3$):

$$\begin{aligned} y_{it} &= \beta_0 + \int_0^1 W_{it}(s)\gamma_0(s)ds + b_i + \epsilon_{it}, \qquad i = 1, \cdots, 100, \qquad (2.7.1) \\ \text{where} \qquad \gamma_0(s) &= \sum_{h \in H_{\gamma_0}} a_{0h} \exp\left[-2500 * \left(\frac{h-s}{100}\right)^2\right] \end{aligned}$$

The bumpy predictor functions were generated from the following equation

$$\begin{aligned} w_{it}(s) &= \sum_{h \in H_1} (\xi_{1h} + c_{1h})\text{exp}\left[-2500 * \left(\frac{s-h}{100}\right)^2\right] \qquad (2.7.2) \\ &+ \sum_{h \in H_2} (\xi_{2h} + c_{2h})\text{exp}\left[-1000 * \left(\frac{s-h}{100}\right)^2\right] \\ &+ (\xi_{31} + 0.9)\text{exp}\left[-250 * \left(\frac{s-50}{100}\right)^2\right], \end{aligned}$$

where c_{1h}, c_{2h} and a_{0h} are defined in Table 2.1. $\{\xi_{1h}, h \in H_1\}$, $\{\xi_{2h}, h \in H_2\}$, and ξ_{31} were drawn independently from Uniform(0, 0.1). Also, $\beta_0 = 0.06$, $\epsilon_{it} \sim N[0, (0.02)^2]$ and $b_i \sim N[0, (0.05)^2]$.

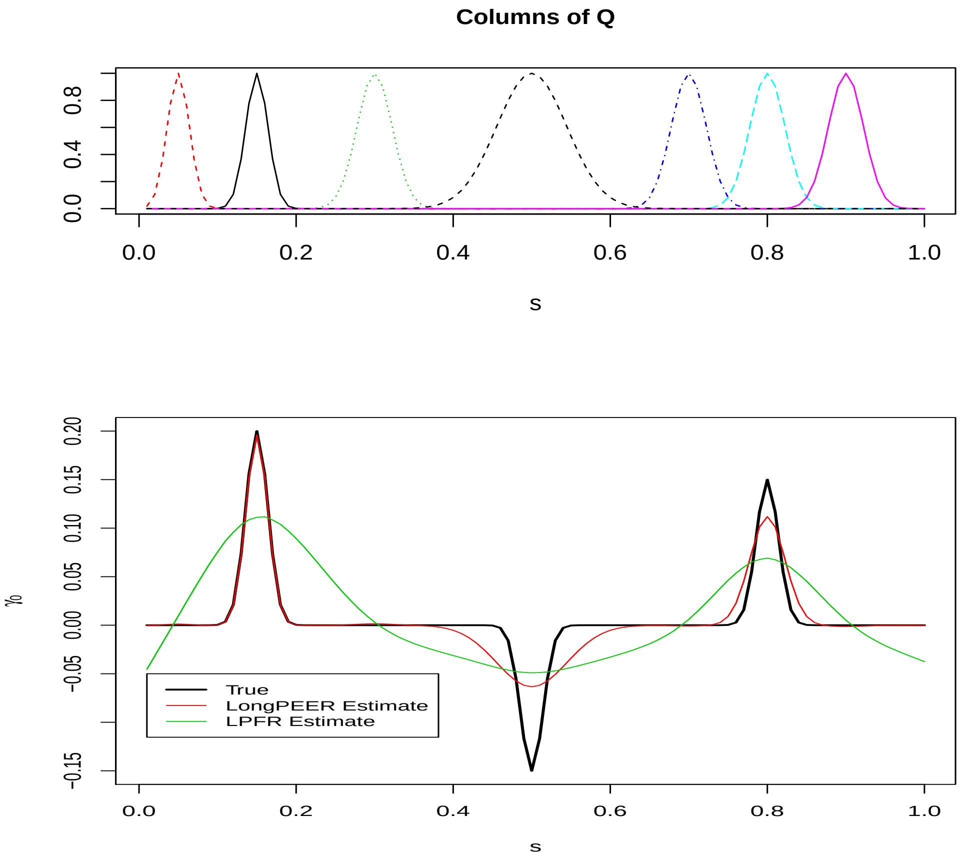

Figure 2.1: Average estimates of γ for the simulation study described in Section 2.7.1 with $\phi_a = 10$ and $\phi_b = 1$. Top panel displays the columns of Q matrix used in defining *decomposition based penalty*. The bottom panels display the true γ and the average estimates in 100 simulations.

Table 2.1: Values of c_{1h}, c_{2h}, a_{0h} and a_{1h} for generating predictor and regression function in simulation studies in Sections 2.7.1, 2.7.2, 2.7.3 and 2.7.4.

$h \in H_1$		$h \in H_2$		$h \in H_{\gamma_0}$		$h \in H_{\gamma_1}$	
h	c_{1h}	h	c_{2h}	h	a_{0h}	h	a_{1h}
15	0.10	30	0.60	15	0.20	30	0.06
5	0.10	70	0.50	50	-0.15	70	-0.06
		80	0.50	80	0.15		
		90	0.40				

Table 2.2: Estimation and prediction errors for LPFR and LongPEER estimates based on 100 simulated datasets. The sample size is set at N=100 and the number of longitudinal observations at $n_i = 4$.

	LongPEER	LPFR
MSE(γ_0)	0.0323	0.2244
Trace of Variance(γ_0)	0.0028	0.0490
$\|\|$Bias(γ_0)$\|\|^2$	0.0295	0.1754
SSPE of Y	1.1566	1.1535

We applied both LPFR (using `lpfr()` available in the `refund` package in `R` (Ciprian Crainiceanu et al., 2012)) and the LongPEER method to the simulated data. To obtain LPFR estimate, the dimension of both principal components for predictor function and truncated power series spline basis for the regression function were set to 60. The columns of Q matrix used for defining decomposition based penalty, L_Q, for LongPEER estimate of γ are plotted in the top panel of Figure 2.1.

Table 2.2 displays the MSE and prediction error obtained for LongPEER and LPFR estimates. The SSPE in both the methods were very close (1.1566 and 1.1535). The LongPEER estimate has much smaller MSE compared to the LPFR estimate. Both the bias and variance are higher for the LPFR estimate and consequently it has the greater MSE. Figure 2.1 displays the estimates of the regression function. The LongPEER estimate is closer to the true regression function compared to the LPFR estimate. The LPFR estimate seems to be over-smoothed with the magnitude of the bumps underestimated. Better performance of the LongPEER estimate is due to its ability to exploit presumed structural information which is ignored by the LPFR method.

2.7.2 Simulation with time varying regression function

In this simulation, we consider a regression function that varies parametrically with time. We are not aware of other longitudinal functional regression methods for estimating a time-varying regression function so we only evaluated the performance of LongPEER. Our primary goal was to assess the effects of sample size, fraction of variance explained by the model and the relative contribution of external information (as determined by ϕ_a and ϕ_b in equation 2.3.1) on the regression function estimate.

Without loss of generality, we set $\phi_b = 1$ and vary ϕ_a on an exponential scale. Larger values of ϕ_a indicate greater emphasis of prior information on the estimation process. The model considered here is similar to that described in Section 2.7.1 with the exception that $\gamma(t, s) = \gamma_0(s) + t\ \gamma_1(s)$. The function $\gamma_0(s)$ is defined in equation (2.7.2) and $\gamma_1(s)$ is of the form

$$\gamma_1(s) = \sum_{h \in H_{\gamma_1}} a_{1h} \exp\left[-2500 * \left(\frac{h-s}{100}\right)^2\right]$$

where the value of h and a_{1h} are listed in Table 2.1 and $\beta_0 = 0.06$. Realizations of functional predictors were generated as described in Section 2.7.1. For each simulation, an appropriate σ_ϵ^2 was chosen to ensure that the squared multiple correlation coefficient $R^2 = s_y^2/[s_y^2 + \sigma_\epsilon^2]$ is 0.6 and 0.9. Here, $s_y^2 = \frac{1}{4}\sum_{t=0}^{3} \frac{1}{N-1}\sum_{i=1}^{N}(y_{it} - \bar{y}_{.t})^2$ denotes the average sample variance in the set $\{y_{it} - \epsilon_{it} : i = 1, \cdots, N; t = 0, \cdots, 3\}$ with $\bar{y}_{.t} = \frac{1}{N}\sum_{i=1}^{N} y_{it}$.

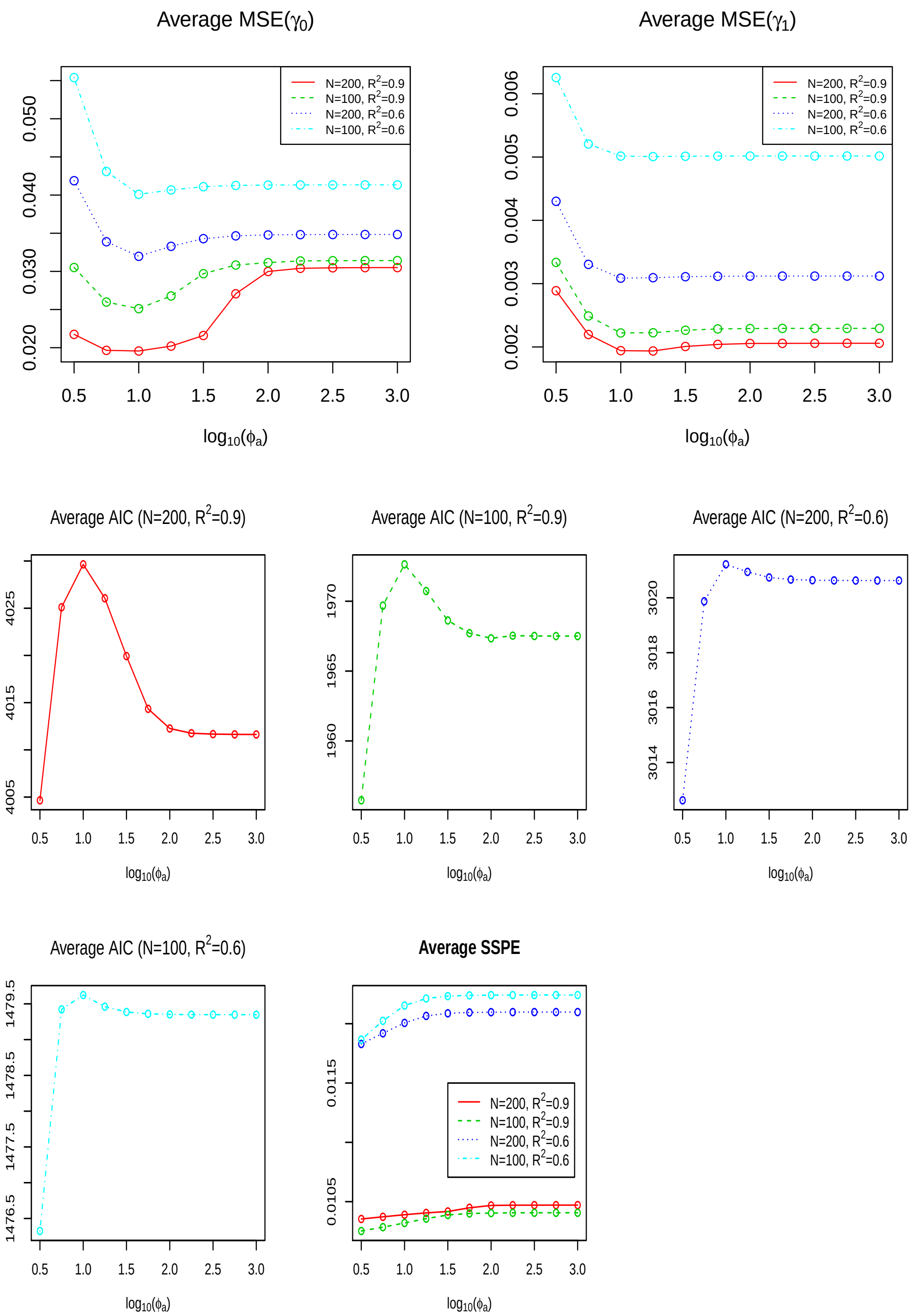

Figure 2.2: Average AIC, SSPE and MSE for simulation study in Section 2.7.2 over 100 simulations. At $\phi_a = 10$, average AIC were maximized and MSE(γ_0) and MSE(γ_1) were minimized. In general, average AIC increased with the increase in sample size and R^2 whereas SSPE, MSE(γ_0) and MSE(γ_1) decreased.

We have repeated the simulation for 4 distinct scenario : (i) $N = 100$, $R^2 = 0.6$, (i) $N = 100$, $R^2 = 0.9$, (i) $N = 200$, $R^2 = 0.6$ and (i) $N = 200$, $R^2 = 0.9$. The estimate of γ_0 and γ_1 were obtained using decomposition based penalty. The columns of Q matrix used for defining decomposition based penalty, L_Q, are plotted in the top panel of Figure 2.4. Results for AIC, MSE and SSPE are displayed graphically in Figure 2.2. The standard deviation of MSE were plotted in Figure 2.3. As the sample size and R^2 increased, both the MSE(γ_0) and MSE (γ_1) were decreased, providing empirical evidence that the LongPEER estimates were consistent. In all the 4 scenario, MSE(γ_0) were minimized at $\phi_a = 10$, then it increased upto $\phi_a = 100$ and became plateau after that. On the contrary, we observed a decrease in MSE(γ_1) as the value of ϕ_a increased up to 10 and after that it became plateau. That is, increase in ϕ_a up to 10 resulted in improvement in estimation of both γ_0 and γ_1. However, increase in ϕ_a beyond 10 resulted in deterioration in performance of estimation for γ_0 and estimation performance for γ_1 remained almost unchanged. In order to understand this result, we need to compare the plots of columns for Q matrix used in defining L_Q with true γ_0 and γ_1 in Figure 2.4. The γ_0 had 3 peaks at $s = 0.2$, 0.5 and 0.8. There were also columns in Q matrix representing peaks at these locations (which serves as a prior information in the estimation process), however, the shape of the peak at $s = 0.5$ were much different from that in γ_0. Due to this difference in shape, as ϕ_a increased from 10 to 100, the feature at $s = 0.5$ in $\tilde{\gamma}_0$ gradually became smaller and smaller leading to gradual increase in MSE(γ_0). On the other hand, γ_1 had two features and the corresponding features in Q matrix were almost similar in shape and that's why MSE(γ_1) was stabilized after $\phi_a = 10$. Therefore, it is important to identify appropriate ϕ_a, especially when we know about possible locations of the features in true functions, but we are not too sure about their shape. Alike MSE(γ_0), AIC were maximized at $\phi_a = 10$ and decreased sharply after that. That is, the value of ϕ_a that maximized AIC also minimized MSE(γ_0) and MSE(γ_1). This suggests

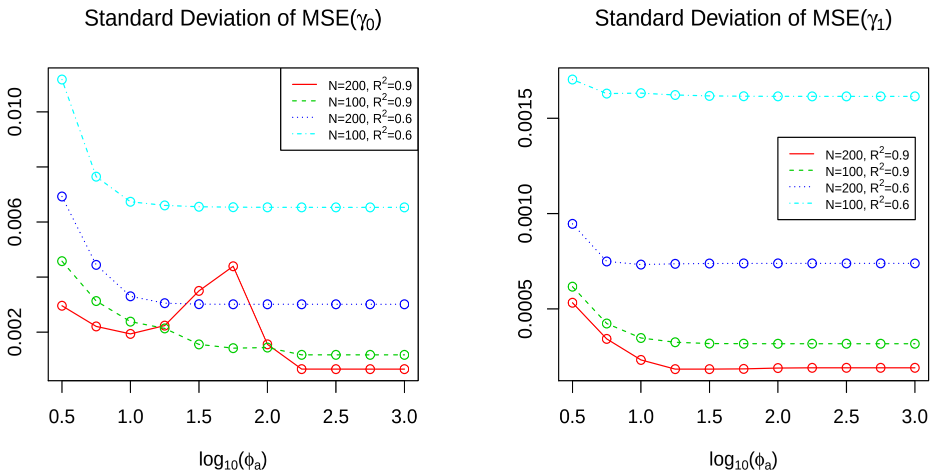

Figure 2.3: Standard deviation of MSE for simulation study in Section 2.7.2 over 100 simulations. In general, standard deviation of MSE(γ_0) and MSE(γ_1) decreased with the increase in sample size and R^2. Also, both MSE(γ_0) and MSE(γ_1) were decreased upto $10^1.25$ and then became a plateau with only exception for MSE(γ_0) in the scenario with $N = 200$ and $R^2 = 0.9$

that AIC can be used to guide the choice of ϕ_a while setting ϕ_b at 1. In general, the choice of ϕ_a should be one that maximizes AIC.

The average LongPEER estimate of γ_0 and γ_1 using decomposition based penalty are displayed in Figure 2.4 with $\phi_a = 10$ and $\phi_b = 1$. For smaller sample size and R^2 the LongPEER estimate (a) seems to over-smooth (i.e. negatively biased) the estimate regression function at the location of the true feature and (b) was positively biased in the locations where we have the prior information about the possible existence of feature, but the true function did not have any feature. However, as we increased the sample size to 200 and/or R^2 to 0.9, we observed that the average LongPEER estimate $\gamma_0(\cdot)$ and $\gamma_1(\cdot)$ approached to the true functions.

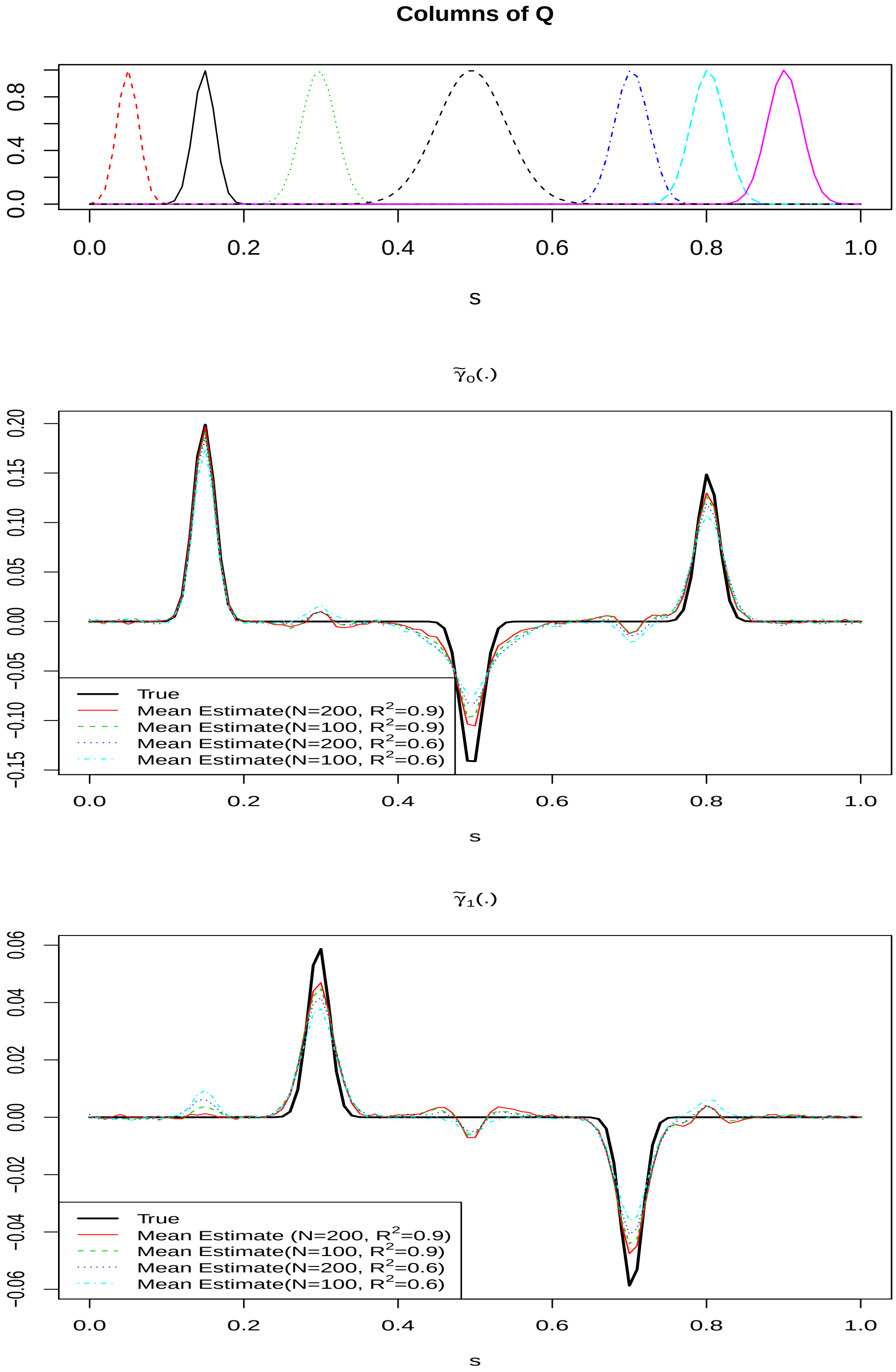

Figure 2.4: Average estimates of the components of regression functions for the simulation study described in Section 2.7.2 with $\phi_a = 10$ and $\phi_b = 1$. Top panel displays the columns of Q matrix used in defining *decomposition based penalty*. The middle and bottom panels display the average estimates of γ_0 and γ_1. Average estimates of γ_0 and γ_1 improved as N and/or R^2 increased.

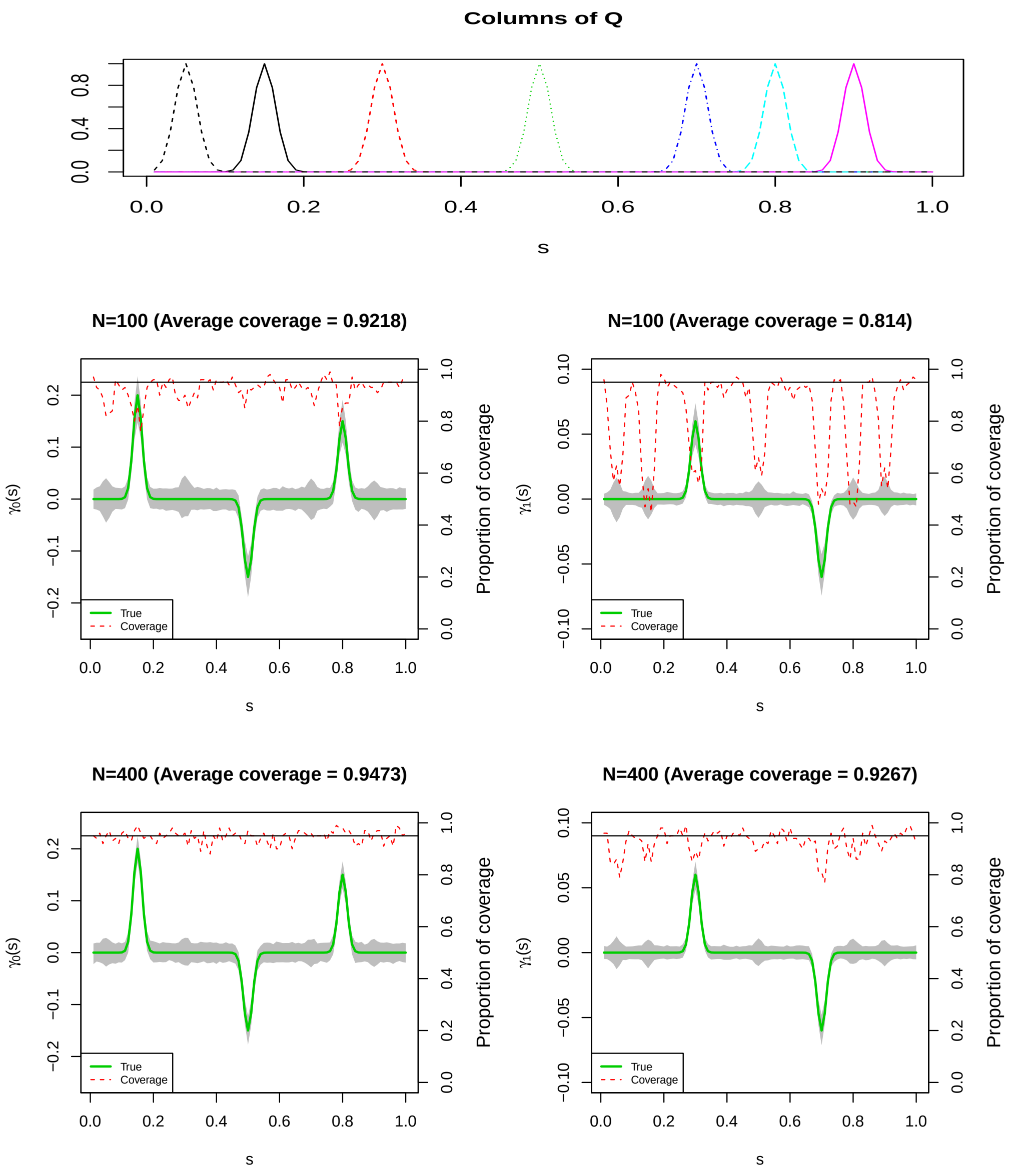

Figure 2.5: Coverage probabilities of LongPEER estimates in 100 simulations with $\phi_a = 10$ and $\phi_b = 1$ as discussed in Section 2.7.3. Top panel displays the columns of Q matrix used in defining *decomposition based penalty*. The plots in middle and bottom panels display the pointwise 95% confidence band (shaded region) and coverage proportions (the dotted line) based on $N = 100$, and $N = 400$ subjects, respectively.. Plots in the left column display the cross-sectional function $\gamma_0(\cdot)$ and the plots in the right column the longitudinal function $\gamma_1(\cdot)$. The horizontal line in each plot marks the nominal coverage of 95%.

2.7.3 Coverage probability

In this section, we used the simulation setup described in Section 2.7.2 with $R^2 = 0.90$. The columns of Q matrix used in defining *decomposition based penalty* are displayed in top panel of Figure 2.5. The middle and bottom panel plots the confidence bands and the coverage probabilities obtained using $\phi_a = 10$. The 95% confidence bands are constructed as $Estimate \pm 1.96\times$ ($Standard\ Error$). When the sample size N increased, there was a notable improvement in coverage of both $\gamma_0(\cdot)$ and $\gamma_1(\cdot)$. For $N = 100$, the coverage of $\gamma_1(\cdot)$ by the confidence bands was only around 81%. This confidence band under-coverage of $\gamma_1(\cdot)$ is caused by the comparatively larger bias in the estimation of $\gamma_1(\cdot)$ with $N = 100$. (see Section 2.7.2 and Figure 2.4). The observed coverage approaches the desired level of 0.95 when N increases and for $N = 400$, the coverage is very close to the 95% mark. We also explored the influence of ϕ_a on the confidence band and coverage probability (not shown here). The higher values of ϕ_a led to the confidence band shrinkage and this in turn resulted in under-coverage of both $\gamma_0(\cdot)$ and $\gamma_1(\cdot)$.

2.7.4 Estimation in the presence of only partial information

Since the LongPEER estimate uses external information in the estimation process, it is of interest to evaluate its estimation performance when only partial information is available. In this Section, we use the similar simulation scenario as described in Section 2.7.3. However, the penalty is defined without regard for information about the feature at the location $s = 0.5$. As displayed in Figure 2.6, the LongPEER estimates of $\gamma_0(s)$ has proper structure at $s = 0.5$ on average. Indeed as with the ordinary ridge penalty this structure is inherited from the empirical eigenvectors of $W(\cdot)$. This highlights the advantage of the PEER estimate which arises from the jointly determined eigenvectors of $W(\cdot)$ and L (see

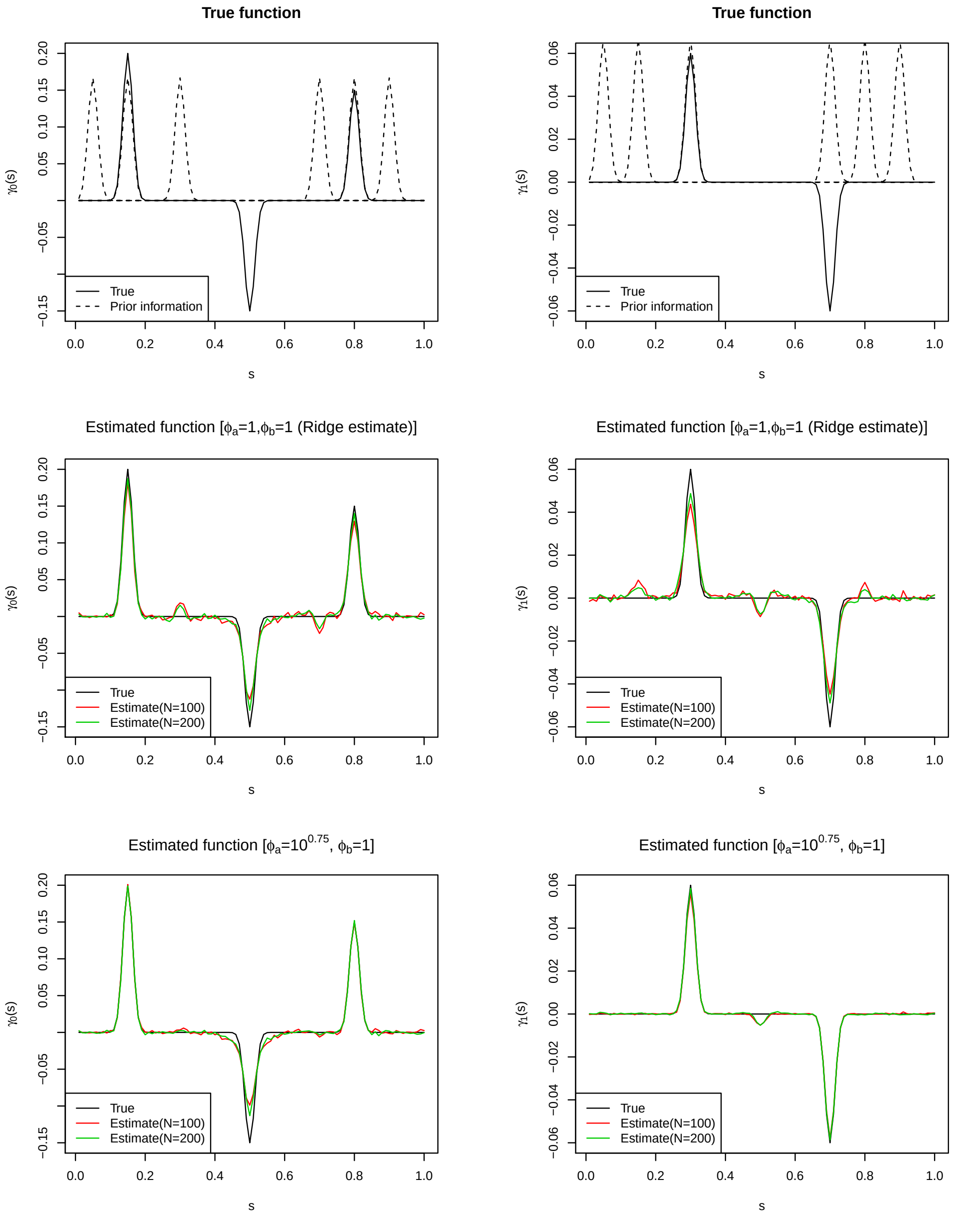

Figure 2.6: Performance of LongPEER estimate when only partial information is available via simulation as described in Section 2.7.4. *Top panel*: true regression functions $\gamma_0(\cdot)$ (left) and $\gamma_1(\cdot)$ (right) plotted using solid lines and 6 vectors spanning P_Q (dashed lines). *Middle panel*: Average LongPEER estimates (over 100 simulations) obtained from the ordinary ridge penalty. *Bottom panel*: Average LongPEER estimates (over 100 simulations) obtained from the penalty using P_Q defined by 6 vectors displayed in the top panel and $\phi_a = 10^{0.75}$.

Appendix). This estimate depends on the relative contribution of the predictors and the penalty controlled by the ratio of ϕ_a to ϕ_b.

The relative increase in the contribution of external information in the estimation process resulted in the estimate shrinkage towards zero at $s = 0.5$. The estimates displayed in Figure 2.6 result from keeping ϕ_b constant at 1 and values of $\phi_a = 1$ (i.e. ridge penalty in the middle row) and $\phi_a = 10^{0.75}$ (bottom row). For ϕ_a values larger than $10^{0.75}$, minimal change in the estimate were observed.

2.8 MRS study application

We applied LongPEER to understand the association of metabolite spectra obtained from basal ganglia and the global deficit score (GDS) in a longitudinal study of late stage HIV patients and how this association evolves over time. The study description is available elsewhere (Harezlak et al., 2011). We treat global deficit score (GDS) as our scalar continuous response variable and MR spectrum (sampled at $K = 399$ distinct frequencies) as functional predictor. GDS is often used as a continuous measure of neurocognitive impairment (e.g., Carey et al., 2004) and a large GDS score is an indicator of high degree of impairment. The collected MRS spectra are composed of the combination of pure metabolite spectra,

Table 2.3: Comparison of AIC for selection of scalar covariates, ϕ_a ($\phi_b = 1$) and time structure in $\gamma(t, \cdot)$ in Section 2.8

	Scalar covariates	Time structure in $\gamma(t, \cdot)$	ϕ_a	AIC
Model 1	t	$\gamma_0(t, \cdot) + t\gamma_1(t, \cdot)$	10	-395.2335
Model 2	Age, t	$\gamma_0(t, \cdot) + t\gamma_1(t, \cdot)$	10	-405.2796
Model 3	Gender, t	$\gamma_0(t, \cdot) + t\gamma_1(t, \cdot)$	10	-395.9040
Model 4	Race, t	$\gamma_0(t, \cdot) + t\gamma_1(t, \cdot)$	10	-398.5607
Model 5	t, t^2	$\gamma_0(t, \cdot) + t\gamma_1(t, \cdot) + t^2\gamma_2(t, \cdot)$	10	-394.5752
Model 6	t	$\gamma_0(t, \cdot) + t\gamma_1(t, \cdot)$	100	-395.3670
Model 7	t	$\gamma_0(t, \cdot) + t\gamma_1(t, \cdot)$	$\sqrt{10}$	-394.9788

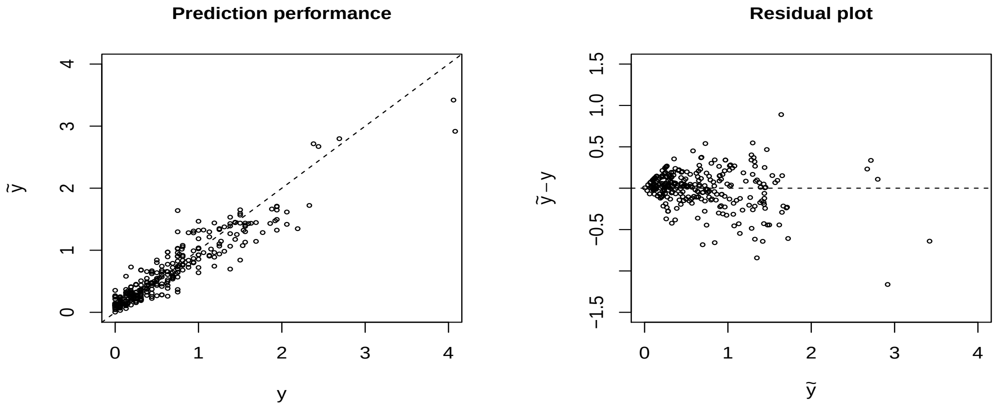

Figure 2.7: Prediction performance of Model in equation (2.8.1) as discussed in Section 2.8. *Left panel* displays plot for observed GDS score (y) and predicted value ($\tilde{y}$). *Right panel* displays plot for observed $\tilde{y}$ and residuals ($y - \tilde{y}$).

instrument noise and baseline profile. We collected a total of $n_\bullet = 306$ observations from $N = 114$ subjects. The longitudinal observations for each subject were within 3 years from baseline. The number of observations per subject ranged from 1 to 5 with a median equal to 3. Information on spectra obtained from nine pure metabolites was available and hence we were able to use this to define a decomposition based penalty L_Q as in equation (2.3.1). The pure metabolite spectra included spectra of Creatine (Cr), Glutamate (Glu), Glucose (Glc), Glycerophosphocholine (GPC), *myo*-Inositol (*m*I), N-Acetylaspartate (NAA), N-Acetylaspartylglutamate (NAAG), *scyllo*-Inositol (Scyllo) and Taurine (Tau). These pure metabolite spectra are displayed in Figure 1.2.

We also had the information available on demographic factors including *age* at baseline, *gender* and *race*. We relied on AIC to choose (a) scalar covariates in the model, (b) ϕ_a (while setting $\phi_b = 1$) for defining decomposition based penalty L_Q and (c) the time structure in $\gamma(t, \cdot)$. Based on the AIC (see Table 2.3), it appeared that the Models 1, 5 and 7 were fairly close and doing better compared to the remaining models. In both the models, ϕ_a was set at 10 and gender was considered as only covariate. Models 1 and 5 were different in terms of

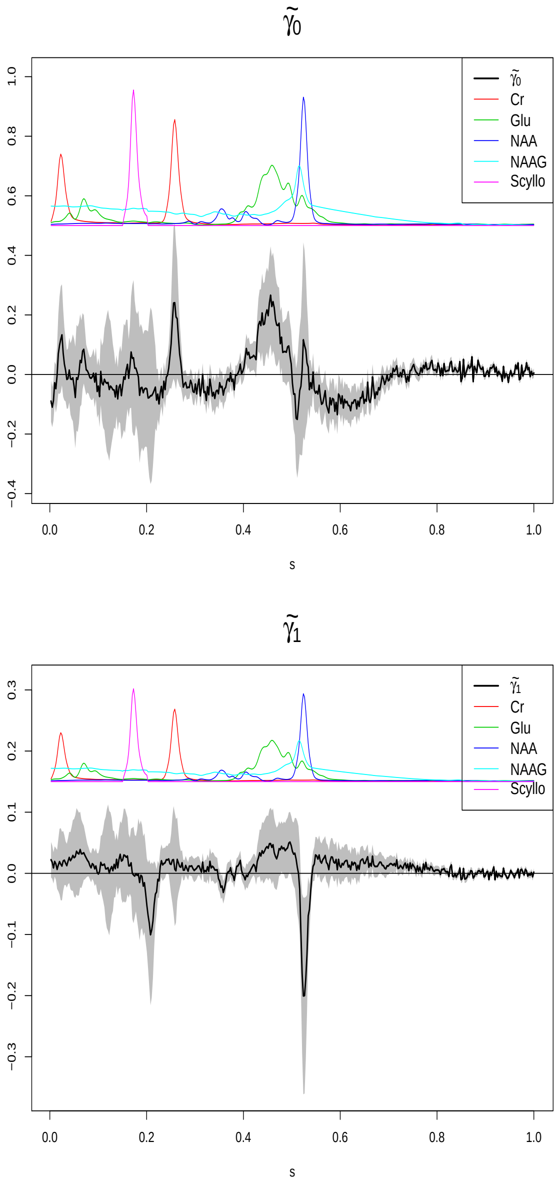

Figure 2.8: Estimates of the regression function (with 95% point-wise confidence band) for the analysis described in Section 2.8. Shaded region in both the plots represent the point-wise confidence band. The top panel shows the estimated $\gamma_0(\cdot)$ and bottom panel the estimate of $\gamma_1(\cdot)$. Selected scaled pure metabolite profiles are also shown on both plots. Estimation was carried out using decomposition based penalty using $\phi_a = 10$, $\phi_b = 1$.

the time structure in $\gamma(t,s)$. Even though inclusion of $\gamma_2(t,\cdot)$ led to only marginal increase in AIC (-394.5752 vs -395.2335), we did not find any region of $\gamma_1(t,\cdot)$ significant using point-wise 95% confidence interval in Model 5. Hence we preferred Model 5 over Model 1. Models 1 and 7 were different in terms of the ϕ_a. Use of smaller ϕ_a led to slight increase in AIC (-394.9788 vs -395.2335), but estimates for $\gamma_0(\cdot)$ and $\gamma_1(\cdot)$ became very wiggly. Hence, we fit Model 1 to fit data (with $\phi_a = 10$, $\phi_b = 1$) as follows:

$$y_{it} = \beta_0 + \beta_1 \, t + \int_{\Omega} W_{it}(s)\gamma(t,s)ds + b_i + \epsilon_{it}, \tag{2.8.1}$$

where, $\gamma(t,s) = \gamma_0(s) + t \; \gamma_1(s)$ and y_{it} and $W_{it}(\cdot)$ are the GDS and basal ganglia spectrum for subject i at time t, respectively. We assume that $\epsilon_{it} \sim N(0,\sigma^2_\epsilon)$ and b_i is the subject-specific random intercept distributed as $N(0,\sigma^2_b)$. The estimates were obtained as the BLUP from the mixed model formulation described in Section 2.6.1 using $L_0 = L_1 = L_Q$.

The estimate of λ (tuning parameter) associated with $\gamma_0(\cdot)$ and $\gamma_1(\cdot)$ were 1.1516 and 2.242, respectively. Estimates of σ^2_ϵ and σ^2_b were 0.0786 and 0.3332, respectively. Plot of observed GDS score and fitted value and residual plot are displayed in Figure 2.7 for the purpose of model checking. It appears that the residuals lie evenly both side of the origin.

Figure 2.8 displays the estimates of $\gamma_0(\cdot)$ and $\gamma_1(\cdot)$ with pointwise 95% confidence bands. To make the interpretation easier, we include the selected pure metabolite spectra. These figures reveal that $\hat{\gamma}_0(\cdot)$ (the 'baseline' part of the regression function) is different from zero at the locations where at least one of the pure metabolites Cr and Gluhas a bump. Similarly, each non-zero part of $\hat{\gamma}_1(\cdot)$ (the 'longitudinal' part of the regression function) coincides with bump locations of one or more pure metabolite profiles of NAA and Scyllo.

Pointwise confidence interval for $\gamma_0(\cdot)$ and $\gamma_1(\cdot)$ contain the"zero" line over most of the intervals of interest. The estimated γ_0 is significant in the region $s \in (0.4, 0.5)$ which corresponds to spectrum of Glu. The estimated γ_0 was also significant in $s \in (0.55, 0.70)$, however, none of the 9 pure metabolite spectra has peak in this region. The estimated γ_1 is significant in the region $s \in (0.50, 0.55)$ which corresponds to spectrum of NAA. The finding of the significant negative 'longitudinal' effect of NAA is worth noticing. This suggests that GDS increases as the NAA concentration decreases in basal ganglia. This finding is consistent with the studies where reduced concentration of NAA has been found to be associated with decrease in neuronal mass (Christiansen et al., 1993; Lim and Spielman, 2005; Soares et al., 2009).

The proposed method also allowed us to investigate other form of $f(t)$ (such as $exp(t)-1$ or $log(t+1)$). When we actually compared $\gamma(t,\cdot) = \gamma_0(t,\cdot) + [exp(t) - 1]\gamma_1(t,\cdot)$ with $\gamma(t,\cdot) = \gamma_0(t,\cdot) + t\gamma_1(t,\cdot)$ the AIC was increased to -394.5601 from -395.2335. However, the estimation with $\gamma(t,\cdot) = \gamma_0(t,\cdot) + log(t+1)\gamma_1(t,\cdot)$ did not show any region in $\gamma_1(\cdot)$ using 95% confidence band. This suggests that it may be good idea to investigate other time structures in $\gamma(t,\cdot)$ if we can collect longitudinal observations for longer period from baseline, say 10 years.

2.9 Discussion

We have proposed a novel estimation method for longitudinal functional regression and derived some properties of the coefficient function estimate. Within this framework, the LongPEER method is the first to allow a coefficient function to *vary with time.* It extends the scope of generalized ridge regression to the realm of longitudinal data. The approach may be viewed as an extension of longitudinal mixed effects models, replacing scalar predictors by functional predictors. Advantages of this method include: 1) a framework that allows the regression function to vary over time; 2) the ability to incorporate structural information into the estimation process; and 3) easy implementation through the linear mixed model equivalence.

The emphasis here is on a general statistical framework for incorporating scientific knowledge into the estimation process when both the scalar response and predictor functions are observed longitudinally. An advantage when we use specific prior information is illustrated in the first simulation study where smoothness constraints or a spline basis representations perform poorly. The simulation in Section 2.7.3 suggests that the coverage probabilities of the confidence bands for the true regression function are close to the nominal level. However, for smaller sample size, the naive confidence bands do not seem to be sufficient and an alternative solution which can take into account the estimation bias might be needed. In the case when only partial information is available, the proposed method can be still very useful if we limit the relative contribution of "informed" space and/or increase the sample size (see Subsection 2.7.4). In total absence of prior information, one may impose more vaguely-defined constraints—such as smoothness or re-weighted empirical subspaces—to estimate the coefficient function.

Solutions to the generalized ridge regression problem can be expressed in many forms. The linear mixed model equivalence provides an easy computational implementation as well as an automatic choice of the tuning parameters using REML criterion. The GSVD provides algebraic insight and a convenient way to derive the bias and variance expressions of the estimates. Another natural way to obtain the regression function estimates is by using Bayesian equivalence (see e.g., Robinson, 1991) with the informative priors quantifying the available scientific knowledge.

One of the natural extensions of our work can incorporate multiple functional predictors. For example, if we observe two functional predictors $W_t^{(1)}(\cdot)$ and $W_t^{(2)}(\cdot)$ with $\gamma^{(1)}(t,\cdot)$ and $\gamma^{(2)}(t,\cdot)$ their associated coefficient functions, respectively. Furthermore, we can express $\gamma^{(1)}(t,s) = \gamma_0^{(1)}(s) + f_1^{(1)}(t)\gamma_1^{(1)}(s) + \cdots + f_d^{(1)}(t)\gamma_d^{(1)}(s)$ and $\gamma^{(2)}(t,s) = \gamma_0^{(2)}(s) + f_1^{(2)}(t) + \gamma_1^{(2)}(s) + \cdots + f_d^{(2)}(t)\gamma_d^{(2)}(s)$. If $W^{(1)}$ and $W^{(2)}$ represent design matrices for the two functional predictors, then we can estimate $\gamma^{(1)}(t,\cdot)$ and $\gamma^{(2)}(t,\cdot)$ by finding the BLUP estimate of $\gamma^{(1)}$ and $\gamma^{(2)}$ from the mixed model, $y = X\beta + W^{(1)}\gamma^{(1)} + W^{(2)}\gamma^{(2)} + Zb + \epsilon$. The simplified formula for bias and variance derived in Appendix still holds with an additional assumption $(W^{(1)})^\top V^{-1} W^{(2)} = 0$.

As presented here, the method addresses models having a continuous scalar outcome, but allowing for either binary or count responses is of interest. Indeed, an important problem that arises in MRS data is that of understanding the neurocognitive impairment status of HIV patients, defined as a binary variable, based on functional predictors collected over time. Extending our approach to these general settings is possible and currently being pursued.

Appendix

Connection with the GSVD

We provide the derivation of the estimates using the GSVD. After some algebra, the generalized ridge estimate in Eq. (2.5.5) for γ can be expressed as

$$\hat{\gamma} = -A_1 X^\top V^{-1} y + A_2 W^\top V^{-1} y$$

where

$$A_1^\top = (X^\top V^{-1}X)^{-1}X^\top V^{-1}W[W^\top V^{-1}W + L^\top L - W^\top V^{-1}X(X^\top V^{-1}X)^{-1}X^\top V^{-1}W]^{-1}$$
$$A_2 = W^\top V^{-1}W + L^\top L - W^\top V^{-1}X(X^\top V^{-1}X)^{-1}X^\top V^{-1}W$$

When $X = 0$ (a situation without any scalar predictors) or $X^\top V^{-1}W = 0$ the generalized ridge estimation of γ can be put into a PEER estimation framework in terms of GS vectors, as discussed below.

With $X = 0$ or $X^\top V^{-1}W = 0$, the $\hat{\gamma}$ reduces to $[W^\top V^{-1}W + L^\top L]^{-1}W^\top V^{-1}y$. Moreover, in this case generalized ridge estimate of β becomes $[X^\top V^{-1}X]^{-1}X^\top V^{-1}y$. Now, if we transform $\tilde{W} := V^{-1/2}W$ and $\tilde{y} := V^{-1/2}y$, we can rewrite L as

$$L = \lambda_0 \text{ blockdiag}\left\{L_0, \frac{\lambda_1}{\lambda_0}L_1, \cdots, \frac{\lambda_D}{\lambda_0}L_D\right\} = \lambda_0 L^s$$

Here, L^s can be interpreted as a scaled L where scaling is done for all the tuning parameters associated with the 'longitudinal' part of the regression function with respect to the 'baseline' tuning parameter.

Set $\tilde{p} = (D + 1)p$, let m denote the number of rows in L and set $c = \dim[\text{Null}(L)]$. Further, assume that $n_\bullet \leq m \leq \tilde{p} \leq m + n_\bullet$ and the rank of the $(n_\bullet + m) \times \tilde{p}$ matrix $[\tilde{W}^\top \quad (L^s)^\top]^\top$ is $\tilde{p}$. The following describes the *GSVD* of the pair $(\tilde{W}, L^s)$: there exist orthogonal matrices $\mathcal{U}$ and $\mathcal{V}$, a nonsingular $\mathcal{G}$ and diagonal matrices S and M such that Randolph et al. (2012)

$$\tilde{W} = \mathcal{U}\mathcal{S}\mathcal{G}^{-1} \qquad \mathcal{S} = [0 \;\; S] \qquad S = \text{blockdiag}\{S_1, \;\; I_{\tilde{p}-m}\}$$

$$L^s = \mathcal{V}\mathcal{M}\mathcal{G}^{-1} \qquad \mathcal{M} = [M \;\; 0] \qquad M = \text{blockdiag}\{I_{\tilde{p}-n_\bullet}, \;\; M_1\}$$

Submatrices S_1 and M_1 have $\ell = n_\bullet + m - \tilde{p}$ diagonal entries ordered as

$$\begin{array}{l} 0 < \sigma_1 \leq \sigma_2 \leq \cdots \leq \sigma_\ell < 1 \\ \\ 0 > \mu_1 \geq \mu_2 \geq \cdots \geq \mu_\ell > 1 \end{array} \qquad \text{where,} \qquad \sigma_k^2 + \mu_k^2 = 1, \qquad k = 1, \ldots, \ell$$

Here, the columns $\{g_k\}$ of $\mathcal{G}$ are the GS vectors determined by the GSVD of the pair $(\tilde{W}, L^s)$. Denote the columns of $\mathcal{U}$ and $\mathcal{V}$ by u_k and v_k, respectively. Now, it can be shown that $[W^\top V^{-1} W + L^\top L]^{-1} W^\top V^{-1} = [W^\top V^{-1} W + \lambda_0^2 (L^s)^\top L^s]^{-1} W^\top V^{-1} = \mathcal{G}(\mathcal{S}^\top \mathcal{S} + \lambda_0^2 \mathcal{M}^\top \mathcal{M})^{-1} \mathcal{G}^\top \; \tilde{W}^\top V^{-1/2}$ and consequently, $\hat{\gamma}$ can be expressed as

$$\hat{\gamma} = \mathcal{G}(\mathcal{S}^\top \mathcal{S} + \lambda_0^2 \mathcal{M}^\top \mathcal{M})^{-1} \mathcal{S}^\top \mathcal{U}^\top \tilde{y} = \sum_{k=\tilde{p}-n_\bullet+1}^{\tilde{p}-c} \frac{\sigma_k^2}{\sigma_k^2 + \lambda_0^2 \mu_k^2} \frac{1}{\sigma_k} u_k^\top \tilde{y} g_k + \sum_{k=\tilde{p}-c+1}^{\tilde{p}} u_k^\top \tilde{y} g_k$$

Further, the bias and variance can be expressed as

$$\begin{aligned} Bias[\hat{\gamma}] = (I - W^{\#} W)\gamma \qquad & = \mathcal{G}(\mathcal{S}^\top \mathcal{S} + \lambda_0^2 \mathcal{M}^\top \mathcal{M})^{-1} (\lambda_0^2 \mathcal{M}^\top \mathcal{M}) \mathcal{G}^{-1} \\ & = \textstyle\sum_{k=1}^{\tilde{p}-n_\bullet} g_k \tilde{g}_k^\top \gamma + \sum_{k=\tilde{p}-n_\bullet+1}^{\tilde{p}-c} \frac{\lambda_0^2 \mu_k^2}{\sigma_k^2 + \lambda_0^2 \mu_k^2} g_k \tilde{g}_k^\top \gamma \end{aligned}$$

$$\begin{aligned} Var[\hat{\gamma}] = W^{\#} V (W^{\#})^\top \quad & = \mathcal{G}(\mathcal{S}^\top \mathcal{S} + \lambda_0^2 \mathcal{M}^\top \mathcal{M})^{-1} \mathcal{S}^\top \mathcal{S} (\mathcal{S}^\top \mathcal{S} + \lambda_0^2 \mathcal{M}^\top \mathcal{M})^{-1} \mathcal{G}^\top \\ & = \textstyle\sum_{k=\tilde{p}-n_\bullet+1}^{\tilde{p}-c} \frac{\sigma_k^2}{(\sigma_k^2 + \lambda_0^2 \mu_k^2)^2} g_k g_k^\top + \sum_{k=\tilde{p}-c+1}^{\tilde{p}} g_k g_k^\top \end{aligned}$$

where, $W^{\#} = [W^\top V^{-1} W + L^\top L]^{-1} W^\top V^{-1}$ and $\tilde{g}_k$ denotes the kth column of $\mathcal{G}^{-T} = (\mathcal{G}^{-1})^\top = (\mathcal{G}^\top)^{-1}$. Further, we can express bias as $[W^\top V^{-1} W + L^\top L]^{-1} L^\top L \gamma$ which means $\hat{\gamma}$ will be unbiased only when $\gamma \in \text{Null}(L)$.

For estimates obtained using this technique, the bias and variance can be expressed in terms of generalized singular vectors, provided the assumption of $X^\top V^{-1}W = 0$ applies. In this case, one can show that $\hat{\beta}$ is simply the generalized least squares estimate from the linear model $y = X\beta + \epsilon^*$, and $\hat{\gamma}$ is the generalized ridge estimate from $y = W\gamma + \epsilon^*$ with penalty L. That is, β is estimated as if $W\gamma$ were not present, and γ is estimated as if $X\beta$ were not present. The PEER estimate discussed in this Section can be thought of as an extension of the estimation discussed in Randolph et al. (2012) in two ways: we allow for a general covariance matrix V (for y) and we also extend the penalty operator to apply across multiply-defined domains, $L_0, \ldots, L_D$.

Chapter 3

Regression tree with longitudinal data

In longitudinal studies, repeated measurements of the outcome variable are often collected at irregular and possibly subject-specific time points. Parametric regression methods for analyzing such data have been developed by Laird and Ware (1982) and Liang and Zeger (1986) among others, and have been summarized by Diggle et al. (2002). Often the population under consideration is *heterogeneous* in terms of trend and covariate effect. Under such situation traditional mixed effect models (such as, linear mixed effect model) assuming a “common parametric form” for covariates and time might not be an appropriate option. If the population under consideration is diverse and there exists several distinct subgroups within it, the true parameter value(s) for longitudinal mixed effect model may vary between these subgroups. For example, Raudenbush (2001) used a longitudinal depression study as an example to argue that it is incorrect to assume that all the people in a given population will be experiencing either increasing or decreasing levels of depression. In such instances, an assumption of “common parametric form” will mask important subgroup differences and will lead to erroneous conclusions. In this chapter, we present an regression tree construction algorithm to identify meaningful and interpretable subgroups with differential longitudinal trajectories and/or differential covariate effect(s) on the response variable from such a heterogeneous population. We propose a regression tree construction technique (LongCART algorithm) with longitudinal data that (1) takes the decision about further splitting at each node controlling type I error, and (2) is applicable in cases when measurements are taken at subject specific time-points.

3.1 Regression tree in cross-sectional setting

Tree based methods, unlike classical regression techniques, do not require any pre-specified relationship between the response and predictors. In principle, tree based methods tries to partition the model space (defined by the predictors) into smaller subspaces in 'best possible' way and then fit simple model (like a constant) within each of these subspaces. Among the tree based methods, classification and regression tree (CART) method (Breiman et al., 1984; Clark and Pregibon, 1992; Verbyla, 1987) is probably the most popular one. When the response variable of interest is categorical, then this problem is known as *classification tree*. We are interested in *regression tree*, where the response variable is continuous. In the context of CART, the predictor variables available to construct the tree are often known as *partitioning variables.*

Let y be the response variable and $X_1, \cdots, X_S$ are the candidate partitioning variables. We want to construct regression tree via recursive binary partitioning with the data of N individuals. The main challenge in constructing regression tree is to find out the partitioning variable $X_s \in \{X_1, \cdots, X_S\}$ and split point $g \in \text{Range}\{X_s\}$ that solves (Hastie et al., 2001)

$$\min_{s,g} \left[\min_{c_1} \sum_i I(x_{is} \leq g)(y_i - c_1)^2 + \min_{c_2} \sum_i I(x_{is} > g)(y_i - c_2)^2 \right]$$

where, y_i and x_{is} are the respective values of y and X_s for i^{th} individual and $I(\cdot)$ is the indicator function. For any choice of X_s and g, the inner minimization is solved by

$$\hat{c}_1 = \frac{1}{\sum_i I(x_{is} \leq g)} \sum_i I(x_{is} \leq g) y_i \qquad \hat{c}_2 = \frac{1}{\sum_i I(x_{is} > g)} \sum_i I(x_{is} > g) y_i$$

The pair (X_s, g) is chosen by scanning through all the cut-off points of all the partitioning variables. Having found the best split, we partition the data into two resulting regions and repeat the splitting process on each of the two regions. Then this process is repeated on all of the resulting regions. We continue growing tree until we hit some threshold point such as minimum number of observations in a region.

3.2 Longitudinal tree

The thrust of any tree techniques is the extraction of meaningful subgroups characterized by common covariate values and homogeneous outcome. The idea of constructing "tree" can be generalized to longitudinal setting with linear mixed effect model in order to find relatively more homogeneous subgroups. For longitudinal data, this homogeneity can pertain to the mean and/or covariance structure (Segal, 1992).

We refer to the regression tree for longitudinal data as '*Longitudinal tree*'. Figure 3.1 displays a toy example for a longitudinal tree. This regression tree represents a heterogeneous population with three distinct subgroups in terms of their longitudinal profiles. These subgroups can be characterized by *gender* and *age*. Here, *gender* and *age* are the baseline attributes. We consider these baseline attributes as *partitioning variables* in construction of *Longitudinal tree*. In each of the three subgroups, the longitudinal trajectory of y depends

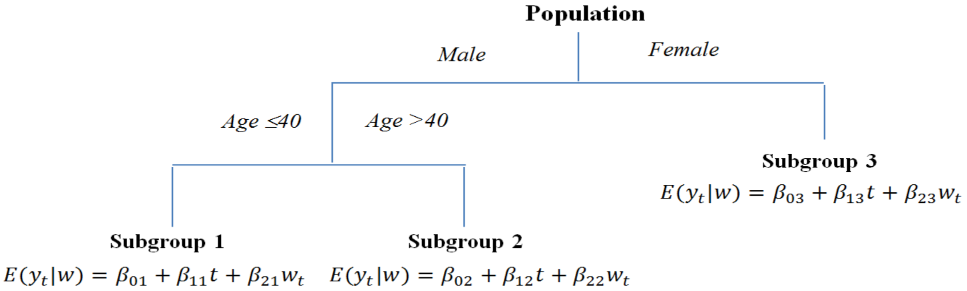
Population
Male
Female
Age ≤40
Age >40
Subgroup 3
$E(y_t|w) = \beta_{03} + \beta_{13}t + \beta_{23}w_t$
Subgroup 1
$E(y_t|w) = \beta_{01} + \beta_{11}t + \beta_{21}w_t$
Subgroup 2
$E(y_t|w) = \beta_{02} + \beta_{12}t + \beta_{22}w_t$

Figure 3.1: Sample longitudinal tree. The population consists of 3 subgroups and they differ in their longitudinal profiles (To be precise, intercept and the coefficients associated with time, t and covariate, w, are not all same). These subgroups are defined by the partitioning variables gender and age.

on the covariates $w_1, \cdots, w_q$, but these subgroups are heterogeneous in terms of the true coefficients associated with their longitudinal profiles. That is, all $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_3$ are different where $\boldsymbol{\beta}_j = [\beta_{0j}, \beta_{1j}, \beta_{2j}]^\top$ is the vector of true coefficients in j^{th} subgroup $(j = 1, 2, 3)$.

3.2.1 Why longitudinal tree?

The overall goal of 'Longitudinal Tree' technique is to identify subpopulations with distinct longitudinal trajectories within a population. The identification of subpopulation is done in a way that the individuals within a subpopulation a) are relatively more comparable in terms of their longitudinal profile with the other individuals in that subgroup compared to other individuals in the population b) have common realized values for one or few baseline attributes. When the longitudinal profile in a population depends on some baseline attributes, the most common strategy is to include these attributes (and their interaction terms) as covariates in the model. However, this strategy has some inherent drawbacks: (a) it can lead to overfitting due to inclusion of all possible interaction terms, especially when the number of potential baseline attributes is large, (b) functional form of the association with baseline attributes need to be known and correctly specified, and (c) it cannot capture nonlinear (not intrinsically linear) effect of baseline attributes. Our goal is to determine the most parsimonious model consisting of a number of homogeneous subgroups from a heterogeneous population profile without strict parametric restrictions or prior information.

One of the popular technique to construct homogeneous subgroups is latent class modeling (LCM) (Muthén and Shedden, 1999). LCM is a statistical method used to identify a set of discrete, mutually exclusive latent classes of individuals based on their responses. An alternative approach is to construct regression tree with longitudinal data (Segal, 1992).

Advantages of regression tree technique over LCM includes: (1) it characterizes the subgroups in terms of partitioning variables and (2) number of subgroups need not to be known a-priori.

3.2.2 *Homogeneity* and *heterogeneity*

Consider the following form of a linear longitudinal mixed effect model

$$y_{it} = \beta_0^x + \beta_1^x t + \mathbf{w}_{it}^\top \boldsymbol{\beta}^x + \mathbf{z}_{it}^\top \mathbf{b}_i + \epsilon_{it} \tag{3.2.1}$$

where i is the subject index and y, t and $\mathbf{w}$ denote the outcome variable, time and the vector of measurements on scalar covariates $w_1, \cdots, w_q$, respectively. Let $X_1^{G_1}, \cdots, X_S^{G_S}$ include all potential baseline attributes that might influence the longitudinal trajectory in (3.2.1). The superscript x is added to the coefficients β_0, β_1 and $\boldsymbol{\beta}$ to reflect their possible dependence on these baseline attributes. Denote $\boldsymbol{\theta}^{x\top} = (\beta_0^x, \beta_1^x, \boldsymbol{\beta}^{x\top})$. With such a model, '*homogeneity*' refers to the situation when the coefficients' true values remain the same for all the individuals in the entire population, *i.e.* $\boldsymbol{\theta}^x = \boldsymbol{\theta}$. When the longitudinal changes in the population of interest are *heterogeneous* there exists distinct subgroups differing in terms of the true values of the coefficients, *i.e.* $\boldsymbol{\theta}^x \neq \boldsymbol{\theta}$. $X_1^{G_1}, \cdots, X_S^{G_S}$ are the partitioning variables used in the regression tree construction. The superscripts to these partitioning variables indicate the number of cut-off points these partitioning variables have. That is, X_s has G_s number of cut-off points ($s = 1, \cdots, S$).

3.2.3 Challenges

Generally, in construction of regression tree with cross-sectional data, the best split is determined by examining the each cut-off point of all partitioning variables as explained in Section 3.1. We can carry that idea in longitudinal setting by performing a test for improvement due to binary partitioning at each cut-off point of all partitioning variables. This idea have been pursued by Abdolell et al. (2002) considering deviance as goodness of fit criterion. They performed test for deviance at each split of a given partitioning variable and selected the partition with maximum (and statistically significant) reduction in deviance for the binary splitting. However, repetitive evaluation of goodness-of-fit criterion at each cut-off point of all partitioning variable leads to the multiple testing problem. With S partitioning variables as $X_1^{G_1}, \cdots, X_S^{G_S}$ (with cut-off points as $G_1, \cdots, G_S$, respectively), total number of tests would be $\sum_{s=1}^{S}(G_s - 1)$. Clearly, it would be very challenging to control type I error with so many tests, especially when one or more partitioning variables are continuous.

3.2.4 Proposed approach

To minimize the problem of multiplicity, we have proposed LongCART algorithm for construction of regression tree that involves only single test for each partitioning variable. We call these tests as *test for parameter instability.* Hence, with S partitioning variables, we need to perform only S *tests for parameter instability.* The number of tests with the proposed approach would be much smaller than $\sum_{s=1}^{S}(G_s - 1)$ in presence of continuous partitioning variables and/or categorical partitioning variables with greater than two categories. Consequently, LongCART algorithm seems more promising to put a better check on the type I error.

We want to construct the regression with certain level of confidence. In general, the controlling type I error rate in the entire tree construction process would be very difficult firstly, because the number of branches are unknown a-priori, and secondly, there are large number of possibilities in choosing a split. To address the issue of type I error, at each node, we divide task of finding best split at a given node into two sub-tasks: (a) First, identify if there is any need for further splitting and (b) Second, given there is need for further splitting, choose the optimum splitting point. Our proposed LongCART algorithm controls type I error while performing the first task, that is, to decide whether there any need for further splitting. Once this decision is taken in favor of splitting, the optimum split can be chosen adopting some model selection process. In order to offer an better overview of LongCART, let's assume there is only a single partitioning variable, say X^G, with G cut-off points. In such case, LongCART algorithm identifies the best split at a given node in a two-step process as follows:

- **Step 1.** Perform an overall *parameter instability test* to detect any evidence of heterogeneity of longitudinal model parameters across G cut-off points of X^G.
- **Step 2.** Given that there is a 'significant' evidence for heterogeneity, the split that provides maximal improvement in goodness of fit criterion is chosen as a partitioning point for the tree construction.

We adapt the LongCART algorithm in situations with multiple partitioning variables via repeating the *parameter instability test* for each partitioning variable controlling type I error at a given level (with some adjustment for multiple testing in step 1). We continue to the second step using the 'most significant' partitioning variable. Details of this algorithm are presented in Section 3.5. The key idea here is that we are combining the multiple testing procedure (step 1) with model selection (step 2) in order to control the type I error while taking the decision on splitting at each node.

In order to construct a test for *parameter instability*, we borrow an idea from the time-series literature. In time-series context often the goal is to evaluate whether the parameter of a regression model is stable across different time points. This is often known as a *test for structural change* or *constancy of parameters* (e.g., Brown et al., 1975; Hjort and Koning, 2002; Nyblom, 1989). We apply very similar idea to evaluate whether the true values of the parameter remains the same across the cut-off values of a partitioning variable in a mixed effects longitudinal model of interest.

3.2.5 Literature review

Binary partitioning for longitudinal data has been proposed first by Segal (1992). However, Segal's implementation is restricted to longitudinal data with a regular structure, that is all the subjects have an equal number of repeated observations at the same time points (Zhang and Singer, 1999). Zhang (1997) proposes multivariate adaptive splines to analyze longitudinal data. Their method, multivariate adaptive splines for the analysis of longitudinal data (MASAL), can be used to generate regression trees for longitudinal data. Abdolell et al. (2002) used deviance as a goodness-of-fit criterion for binary partitioning. They controlled the level of Type I error via permutation test taking into account testing multiplicity. However, permutation tests are computer intensive and the time taken to fit the models is intimidatingly high even for medium-sized data. Sela and Simonoff (2012) as well as Galimberti and Montanari (2002) merged the subgroup differences with the random individual differences. They constructed the regression tree through an iterative two-step process. In the first step, they obtained the random effects' estimates and in the second step, they constructed the regression tree ignoring the longitudinal structure. They repeat these two steps until the estimates of the random effect converge in the first step. The LongCART algorithm provides an improvement over the existing methods in the following aspects: (1) the decision about further splitting at each node is type I error controlled, (2) it is applicable to the when measurements are taken at subject-specific time points, (3) it does not merge group differences with the random subject effect components and (4) it reduces computational time.

In this paper we utilize the parameter instability test in multiple ways. First, in the case of continuous partitioning variables, the proposed test uses the results on score process derived by Hjort and Koning (2002) in conjunction with the properties of Brownian motion and Brownian Bridge. Second, for categorical partitioning variables with a small number of cut-off points, a test for parameter instability is derived in a straightforward way by employing asymptotic normality of the score functions. We derive the asymptotic properties of the instability test and explore its size and power through an extensive simulation study. Finally, we use these instability tests to construct an algorithm for regression trees with longitudinal data.

The remainder of this paper is organized as follows. In Section 3.3 the longitudinal mixed effects model of interest have been summarized. Tests for parameter instability for continuous and categorical partitioning variable cases are discussed separately in Section 3.4. Algorithm for constructing regression trees along with measures of improvement and a pruning technique have been presented in Section 3.5. Results from the simulation studies examining the performance of the instability test and the regression tree as a whole are reported in Section 3.6. An application of the longitudinal regression tree method was illustrated on the metabolite data collected from the chronically HIV-infected patients in Section 3.7.

3.3 Notations and assumptions

Let $\{y_{it}, \mathbf{w}_{it}\}$ be a set of measurements recorded on the i^{th} subject $(i = 1, \ldots, N)$ at time $t = (t_1, \ldots, t_{n_i})$, where y is a continuous scalar outcome; and $\mathbf{w}$ is the vector of measurements on scalar covariates $w_1, \cdots, w_q$. We assume that these covariates are linearly associated with y. In addition, for each individual, we observe a vector of attributes

$(X_{1i}^{G_1}, \cdots, X_{Si}^{G_S})$ measured at baseline. We assume that $X_1^{G_1}, \cdots, X_S^{G_S}$ includes all the potential baseline attributes that might influence the longitudinal trajectory of y and its association with covariates $w_1, \cdots, w_q$. Further, we don't have any idea about the functional form of influence of these baseline attributes. We use the variables $X_1^{G_1}, \cdots, X_S^{G_S}$ as the candidate partitioning variables to construct a longitudinal regression tree to discover meaningful and interpretable subgroups with differential changes in y characterized by the $X_1^{G_1}, \cdots, X_S^{G_S}$.

When the longitudinal profile is homogeneous in the entire population, we can fit the following traditional linear mixed effects model for all N individuals (Laird and Ware, 1982)

$$y_{it} = \beta_0 + \beta_1 t + \mathbf{w}_{it}^\top \boldsymbol{\beta} + \mathbf{z}_{it}^\top \mathbf{b}_i + \epsilon_{it}, \tag{3.3.1}$$

where $\epsilon_{it} \sim N(0, \sigma^2)$ and $\mathbf{b}_i$ is the vector of random effects pertaining to subject i and distributed as $N(0, \sigma^2 \mathbf{D})$. By '*homogeneity*' we mean that the true value of $\boldsymbol{\theta}^\top = (\beta_0, \beta_1, \boldsymbol{\beta}^\top)$ remains the same for all the individuals. In fact, (3.3.1) is the simplified version of model in (3.2.1) under homogeneity.

We follow the common assumptions made in longitudinal modeling that $\mathbf{z}_{it}$ is a subset of $[\mathbf{w}_{it}^\top \;\; t]^\top$; ϵ_{it} and $\mathbf{b}_i$ are independent; ϵ_{it} and $\epsilon_{i't'}$ are independent whenever $i \neq i'$ or $t \neq t'$ or both, and $\mathbf{b}_i$ and $\mathbf{b}_{i'}$ are independent if $i \neq i'$. Here, $\mathbf{w}_{it}^\top \beta$ is the fixed effect term and $\mathbf{z}_{it}^\top \mathbf{b}_i$ is the standard random effects term. For the i^{th} subject, we rewrite the Eq. (3.3.1) as follows

$$\mathbf{y}_i = \mathbf{w}_i \boldsymbol{\theta} + \mathbf{z}_i \mathbf{b}_i + \epsilon_i, \tag{3.3.2}$$

where $\mathbf{y}_i^\top = (y_{i1}, \cdots, y_{in_i})$, $\mathbf{w}_i$ is the design matrix consisting of the intercept, time (t) and covariates ($\mathbf{w}$). n_i is the number of observations obtained from the i^{th} individual. The score function for estimating $\boldsymbol{\theta}$ under (3.3.2) is (see e.g., Demidenko, 2004)

$$\mathbf{u}(\mathbf{y}_i, \boldsymbol{\theta}) = \frac{d}{d\boldsymbol{\theta}} l(\mathbf{y}_i, \boldsymbol{\theta}) = \frac{1}{\sigma^2} \mathbf{w}_i^\top \mathbf{V}_i^{-1} (\mathbf{y}_i - \mathbf{w}_i \boldsymbol{\theta})$$

where $\mathbf{V}_i = \mathbf{I} + \mathbf{z}_i \mathbf{D} \mathbf{z}_i^\top$ and $\mathbf{e}_i = \mathbf{y}_i - \mathbf{w}_i \boldsymbol{\theta}$. Further, its variance is

$$\text{Var}\left[\mathbf{u}(\mathbf{y}_i, \boldsymbol{\theta})\right] = \mathbf{J}(\boldsymbol{\theta}) = -E\left[\frac{d}{d\boldsymbol{\theta}} \mathbf{u}(\mathbf{y}_i, \boldsymbol{\theta})\right] = \frac{1}{\sigma^2} \mathbf{w}_i^\top \mathbf{V}_i^{-1} \mathbf{w}_i$$

Likelihood estimate of $\boldsymbol{\theta}$ obtained using all the observation from N subjects is valid only if all the individuals under considerations are homogeneous. If the individuals are not homogeneous in terms of $\boldsymbol{\theta}$ then the likelihood estimate obtained considering all the subjects together are misleading; the extent and direction of ambiguity in the estimate will depend on the nature and proportion of heterogeneity in the sampled individuals. Therefore, it is important to decide first whether the true value of $\boldsymbol{\theta}$ remains the same for all the subjects or not. In the next section, we describe a way to test whether the true value of $\boldsymbol{\theta}$ remains the same across all the values of a given partitioning variable.

3.4 Test for parameter instability

In this section, we utilize the ideas introduced by Hjort and Koning (2002) to test for the constancy of model parameters over time in time-series context. Our goal is to test whether the true value of $\boldsymbol{\theta}$ remains the same across all distinct values of a given partitioning variable. We refer to this test as a *test for parameter instability.* The testing strategy is described in this section with a single partitioning variable. For multiple partitioning variables, the test needs to be repeated for each of them with an adjustment for multiple testing.

Let $X^G \in \{X_1^{G_1}, \cdots, X_S^{G_S}\}$ be any partitioning variable with G ordered cut-off points $c_{(g)}, g = 1, \cdots, G; c_{(1)} \leq \cdots \leq c_{(G)}$ and $\boldsymbol{\theta}_{(g)}$ be the true value of $\boldsymbol{\theta}$ when $X^G = c_{(g)}$. Assume that there are m_g subject with $X^G = c_{(g)}$. We denote the cumulative number of subjects with $X^G \leq c_{(g)}$ by M_g. That is, $M_g = \sum_{j=1}^{g} m_j$ and $M_G = \sum_{j=1}^{G} m_j = N$. We want to conduct an omnibus test,

$$H_0 : \boldsymbol{\theta}_{(g)} = \boldsymbol{\theta}_0 \quad H_1 : \boldsymbol{\theta}_{(g)} \neq \boldsymbol{\theta}_0.$$

Here, H_0 indicates the situation when parameter $\boldsymbol{\theta}$ remains constant (that is, *homogeneity*) and H_1 corresponds to the situation of parameter instability (that is, *heterogeneity*) . The two tests described in this section utilize the following properties of score function under H_0:

- A1: $E_{H_0}[\mathbf{u}(\mathbf{y}_i, \boldsymbol{\theta}_0)] = 0$;
- A2: $\text{Var}_{H_0}[\mathbf{u}(\mathbf{y}_i, \boldsymbol{\theta}_0)] = \mathbf{J}(\boldsymbol{\theta}_0) = \mathbf{J}$;
- A3: $\mathbf{u}(\mathbf{y}_i, \hat{\boldsymbol{\theta}})|_{H_0} \rightarrow^d N[0, \hat{\mathbf{J}}]$,

where $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimate of $\boldsymbol{\theta}$ and $\hat{\mathbf{J}} = \mathbf{J}(\hat{\boldsymbol{\theta}})$. We discuss the instability test separately for the categorical and continuous variables X^G.

3.4.1 Instability test with a categorical partitioning variable

It is straightforward to obtain a test for parameter instability using the properties A1–A3 when the partitioning variable, X^G, is categorical with a small number of categories (that is, $G \ll N$). Since the score functions $\mathbf{u}(\mathbf{y}_i, \hat{\boldsymbol{\theta}})$ are independent, we have under H_0, the following quantity

$$\chi^2_{cat} = \sum_{g=1}^{G} \left[\sum_{i=1}^{N} I(X_i^G = c_{(g)}) \mathbf{u}(\mathbf{y}_i, \hat{\boldsymbol{\theta}}) \right]^\top \left[m_g \hat{\mathbf{J}} \right]^{-1} \left[\sum_{i=1}^{N} I(X_i^G = c_{(g)}) \mathbf{u}(\mathbf{y}_i, \hat{\boldsymbol{\theta}}) \right]$$

is asymptotically distributed as χ^2 with $(G-1)p$ degrees of freedom where p is the dimension of $\boldsymbol{\theta}$. Here, $I(\cdot)$ is the indicator function. The reduction in p degrees of freedom is due to the estimation of p dimensional $\boldsymbol{\theta}$ from the data.

3.4.2 Instability test with continuous partitioning variable

Here, we first review the results obtained by Hjort and Koning (2002) and then we propose the test for instability with a single continuous partitioning variable. We begin by defining the following *score process*

$$\mathbf{W}_N(t, \boldsymbol{\theta}_0) = N^{-1/2} \sum_{i=1}^{M_g} \mathbf{u}(\mathbf{y}_i, \boldsymbol{\theta}_0) \qquad t \in [t_g, t_{g+1})$$

where $t_g = \dfrac{M_g}{N}$. Using multivariate version of Donsker's theorem and Cramér-Wold theorem (see e.g. Billingsley, 2009) it can be shown that

$$\mathbf{W}_N(t, \boldsymbol{\theta}_0) \to_d \mathbf{Z}(t)$$

where $\mathbf{Z}(t)$ is the zero-mean Gaussian process with $\text{cov}[\mathbf{Z}(t), \mathbf{Z}(s)] = \min(t, s)\mathbf{J}(\boldsymbol{\theta}_0)$. Note that Z is a linear transformation of p independent Brownian motions. Since, $\boldsymbol{\theta}_0$ is unknown in practice, we define the following *estimated score process* replacing $\boldsymbol{\theta}_0$ by $\hat{\boldsymbol{\theta}}$

$$\mathbf{W}_N(t, \hat{\boldsymbol{\theta}}) = N^{-1/2} \sum_{i=1}^{M_g} \mathbf{u}(\mathbf{y}_i, \hat{\boldsymbol{\theta}})$$

By applying Taylor series expansion it is straightforward to show that

$$\mathbf{W}_N(t, \hat{\boldsymbol{\theta}}) \doteq \mathbf{W}_N(t, \boldsymbol{\theta}_0) - t\, \mathbf{W}_N(1, \boldsymbol{\theta}_0)$$

where $A_n \doteq B_n$ means that $A_n - B_n$ tends to zero in probability. In the case of linear mixed effects models, this relationship is exact as the second derivative of the score function is equal to 0. That is, $\mathbf{W}_N(t, \hat{\boldsymbol{\theta}}) = \mathbf{W}_N(t, \boldsymbol{\theta}_0) - t\,\mathbf{W}_N(1, \boldsymbol{\theta}_0)$. Consequently,

$$\mathbf{W}_N(t, \hat{\boldsymbol{\theta}}) \rightarrow_d \mathbf{Z}^0(t) = \mathbf{Z}(t) - t \cdot \mathbf{Z}(1)$$

The limit process $\mathbf{Z}^0(t)$ is a p-dimensional process with covariance function $\text{cov}[\mathbf{Z}^0(t), \mathbf{Z}^0(s)] = s(1-t)J(\boldsymbol{\theta}_0)$ for $s < t$. We can go on to the construction of *canonical monitoring process* $\mathbf{M}_N(t, \hat{\boldsymbol{\theta}})$, and under H_0,

$$\mathbf{M}_N(t, \hat{\boldsymbol{\theta}}) = \hat{\mathbf{J}}^{-1/2}\mathbf{W}_N(t, \hat{\boldsymbol{\theta}}) \rightarrow_d \mathbf{W}^0(t)$$

where $\mathbf{W}^0(t) = (W_1^0(t), \cdots, W_p^0(t))$ is a vector with p independent standard Brownian Bridges as component processes. In other words, kth component of $\mathbf{M}_N(t, \hat{\boldsymbol{\theta}})$ is distributed as a standard Brownian Bridge, $W^0(t)$. That is,

$$M_N(t, \hat{\theta}_k) \rightarrow_d W^0(t)$$

The above weak convergence continues to hold for any 'reasonable' functionals (including supremum) of $M_N(t, \hat{\theta}_k)$ (see e.g. **?**, pp 509, Theorem 1). At this point, Hjort and Koning (2002) proposed several functionals of $M_N(t, \hat{\theta}_k)$ as possible test statistics and suggested to approximate their distribution functions through simulation for comparison purpose. For example, they stated

$$\max_{0 \le t \le 1} ||M_N(t, \hat{\theta}_k)||^2 \rightarrow_d \max_{0 \le t \le 1} ||W^0(t)||^2$$

and suggested to use $\max_{0\leq t\leq 1} ||M_N(t, \hat{\theta}_k)||^2$ as test statistic. Instead we propose to use D_k as defined below as a test statistic

$$D_k \equiv \max_{0\leq t\leq 1} |M_N(t, \hat{\theta}_k)| = \max_{1\leq j\leq N-1} |M_N(t, \hat{\theta}_k)| \to_d \max_{0\leq t\leq 1} |W^0(t)| \equiv D \tag{3.4.1}$$

The primary reason for preferring $\max_{0\leq t\leq 1} |M_N(t, \hat{\theta}_k)|$ over the $\max_{0\leq t\leq 1} ||M_N(t, \hat{\theta}_k)||^2$ is that the limiting distribution of the former is known. The use of D_k as a test statistic eliminates the additional simulation work approximating the limiting distribution and thus making the testing process much more computationally efficient. The resulting reduction in the computation time is significant in the context of regression tree construction with the longitudinal data. D has distribution function (Billingsley, 2009)

$$F_D(x) = 1 + 2\sum_{l=1}^{\infty} (-1)^l \exp\left(-2\, l^2 x^2\right).$$

Although this expression involves an infinite series, this series converges very rapidly. Usually a few terms suffice for very high accuracy. This result can be used to formulate a test for instability of parameters at α level of significance as follows: (1) Calculate the value of the process D_k for each parameter $k = 1, \cdots, p$ and obtain the raw p-values. (2) Adjust the p-values according to a chosen multiple testing procedure. (3) Reject H_0 if the adjusted p-value for any of the processes, D_k, is less than α.

3.4.3 Instability test for multiple partitioning variables

The testing strategy discussed in Sections 3.4.1 and 3.4.2 for a single partitioning variable depends only on the predictor variable type, either categorical or continuous. However, in practice, we expect to have more than one partitioning variable. Let there be S partitioning variables: $\{X_1^{G_1}, \cdots, X_S^{G_S}\}$. In that case we need to perform the instability test for each of the partitioning variables $X_1^{G_1}, \cdots, X_S^{G_S}$ subject to adjustment for multiplicity of type I errors. Let the p-values after multiplicity adjustment be $p_1, \cdots, p_S$, respectively and $p_{min} = \min\{p_1, \cdots, p_S\}$. Candidate partitioning variable with the smallest p-value (p_{min}) is chosen as a partitioning variable if p_{min} is smaller than the nominal significance level. For further discussion please see Section 3.5.

3.4.4 Power under the alternative hypothesis

We consider the following form of Pitman's local alternatives in the vicinity of H_0

$$\boldsymbol{\theta}_{(g)} = \boldsymbol{\theta}_0 + \boldsymbol{\delta} \circ \mathbf{h}\Big(\frac{c_{(g)}}{c_{(G)}}\Big)\frac{1}{\sqrt{N}} + O\left(\frac{1}{N}\right) \tag{3.4.2}$$

where $\boldsymbol{\delta} = (\delta_1, \cdots, \delta_p)^\top$ is the vector containing degrees of departure from the null hypothesis and $\mathbf{h} = (h_1, \cdots, h_p)^\top$ is the vector containing magnitudes of departure. The operation $\circ$ denotes the point-wise multiplication, i.e.,

$$\boldsymbol{\delta} \circ \mathbf{h}\Big(\frac{c_{(g)}}{c_{(G)}}\Big) = \left[\delta_1 h_1\Big(\frac{c_{(g)}}{c_{(G)}}\Big), \cdots, \delta_p h_p\Big(\frac{c_{(g)}}{c_{(G)}}\Big)\right]^\top$$

Theorem 3.4.1. *Under* (3.4.2), *the limiting distribution for the χ^2_{cat} is a non-central chi-square distribution*

$$\chi^2_{cat} \longrightarrow_d \chi'^2\left[(G-1)p,\ \sum_{g=1}^{G}\lambda_g^2\right]$$

where

$$\lambda_g = \mathbf{J} \cdot m_g \mathbf{h}\Big(\frac{c_{(g)}}{c_{(G)}}\Big) \cdot \frac{1}{\sqrt{N}}$$

Theorem 3.4.2. *Under* (3.4.2)*, the limiting distribution for the canonical monitoring process is as follows*

$$\mathbf{M}_N(t, \hat{\boldsymbol{\theta}}) \longrightarrow_d \mathbf{J}^{1/2} \cdot t_g \cdot \boldsymbol{\delta} \circ (\bar{\mathbf{h}}_g - \bar{\mathbf{h}}) + \mathbf{W}^0(t) \qquad t \in [t_g, t_{g+1})$$

where,

$$\bar{\mathbf{h}}_g = \frac{1}{M_g} \sum_{j=1}^{g} m_j \mathbf{h}\Big(\frac{c_{(j)}}{c_{(G)}}\Big) \qquad \bar{\mathbf{h}} = \bar{\mathbf{h}}_G$$

Proofs of these theorems are provided in the Appendix. Briefly, we first approximate the density function using Taylor series expansion and then proceed in the way analogous to the one discussed in Section 3.4.2.

3.5 Longitudinal regression tree

3.5.1 LongCART algorithm

Smaller p-values from the instability test indicate greater evidence towards instability. Intuitively, splits in the tree should be based on the partitioning variable that shows higher evidence towards instability of the parameters. Therefore, we propose the following algorithm in order to construct a regression tree for longitudinal data.

Step 1. Perform the instability test for each partitioning variable separately at a prespecified level of significance α. The level of significance for performing instability test is subject to adjustment for multiple comparisons in order to maintain the level of type I error.

Step 2. Stop if no partitioning variable is significant at level α. Otherwise, choose the partitioning variable with the smallest p-value and proceed to step 3.

Step 3. Consider all cut-off points of the chosen partitioning variable. At each cut-off point, calculate the improvement in the goodness of fit criterion (e.g., deviance). With X^G as the chosen partitioning variable, the improvement in goodness of fit criterion can be obtained at the cut-off point $c_{(g)}$ in the following steps:

a. Split the data in two parts. One group will include the observations from subjects with $X^G \leq c_{(g)}$ and the other group will have the observations from subjects with $X^G > c_{(g)}$.

b. Fit the longitudinal model on (i) all the individuals in the node, (ii) the individuals with $X^G \leq c_{(g)}$ and (iii) the individuals with $X^G > c_{(g)}$. Calculate the goodness of fit criterion from each of these three models. Call them as GOF_{all}, GOF_I and GOF_{II}, respectively.

c. Calculate the improvement in goodness of fit criterion as $\text{GOF}_I + \text{GOF}_{II} - \text{GOF}_{all}$.

Step 4. Choose the cut-off value that provides the maximum improvement in goodness of fit criterion and use this cut-off for binary splitting.

Step 5. Follow the Steps 1-4 for each non-terminal node.

The above strategy for construction of regression tree with longitudinal data has two major advantages over the currently existing algorithms. First, the decision about further splitting at each node is taken controlling type I error. Second, there are huge savings in computation time as we are evaluating the improvement in selected goodness of fit criterion at the cut-off points of the chosen partitioning variable only.

3.5.2 Improvement

A measure of improvement due to regression tree can be provided in terms of likelihood function based criterion. For example, Akaike Information criterion (AIC) for a tree T can be obtained as

$$\mathrm{AIC}_T = 2\sum_{k=1}^{|T|} l_k - 2 \cdot |T| \cdot p$$

where $|T|$ denotes the number of terminal nodes in T, l_k is the log-likelihood in kth terminal node and p is the number of estimated parameters in each node. If we denote the AIC obtained from the traditional linear mixed effects model at root node (that is, common parametric form for covariates and time for the entire population) by AIC_0, the improvement due to regression tree can be measured as

$$\text{Improvement } (T) = \mathrm{AIC}_T - \mathrm{AIC}_0$$

Since the overall model fitted to all the data is nested within the regression tree based model, a likelihood ratio test or test for deviance can be constructed as well to evaluate the overall significance of a given regression tree.

3.5.3 Pruning

The improvement in regression tree comes at a cost of adding complexity to the model. If we can summarize complexity of a tree by number of terminal nodes, the cost adjusted AIC of a regression tree T can be defined as follows

$$\mathrm{AIC}_T(\gamma) = \mathrm{AIC}_T - \gamma(|T| - 1), \quad \gamma > 0$$

where γ be the *average complexity* for each terminal node. As a result, the tree T will be selected if

$$\mathrm{AIC}_T - \gamma(|T| - 1) > \mathrm{AIC}_0$$

or

$$\gamma < \frac{\mathrm{AIC}_T - \mathrm{AIC}_0}{|T| - 1} \equiv \gamma_T \tag{3.5.1}$$

That is, the tree T will be chosen as long as γ_T does not exceed some pre-set level of *average complexity*, γ_0; otherwise, we have to prune the tree T to bring γ_T below γ_0.

3.6 Simulation

We have explored the performance of instability test for continuous partitioning variables and the performance of proposed LongCART algorithm as a whole through simulation studies. The first two simulation studies evaluate the performance of instability test with continuous partitioning variable (as discussed in Section 3.4.2). The third simulation study is aimed to explore the performance of the LongCART algorithm in Section 3.5.1.

3.6.1 Performance of instability test with continuous partitioning variable

Let X^G be continuous partitioning variable with ordered cut-off points as $c_{(1)} \leq \cdots \leq c_{(G)}$. We first investigated the size of the test and then obtained the size-corrected power.

Size of the test

In order to examine the size of the test we have considered a longitudinal model with single mean parameter. We generated observations for N subjects at $t = 0, 1, 2, 3$ from the following model

$$X^G = c_{(g)} : \quad y_{it} = \beta_0 + b_i + \epsilon_{it} \tag{3.6.1}$$

with $\beta_0 = 2$, $b_i \sim N(0, 0.5^2)$ and $\epsilon_{it} \sim N(0, 0.2^2)$. The observations for X^G were generated for each simulation separately from uniform(0,300). For each N, $10,000$ Monte-Carlo samples were generated and the test statistic D_k (see Eq. (3.4.1)) was calculated for each sample separately. The null hypothesis of parameter stability is rejected at $\alpha\%$ level of significance when D_k exceeds the $(1-\alpha) \times 100$th percentile of the limiting distribution.

Table 3.1: Size of proposed parameter instability test for continuous partitioning variable via simulation as discussed in Section 3.6.1. The results were summarized based on $10,000$ simulations for various nominal levels of type I error (α) and sample size (N). The critical values (D_α) from the true limiting distribution of test statistic D_k (see Eq. 3.4.1) is also provided for each α. For each N and α, the simulation results have been summarized by (a) percentage of rejection (to be compared with α) and (b) observed $(1-\alpha)100^{th}$ percentile of D_k (to be compared with D_α). The propose parameter instability test seems to be conservative; however, the size of the test approaches to nominal level with the increase in N.

(a) Percentage of rejection

	N				
$\alpha(\%)$	50	100	200	500	1000
1.25	0.54	0.56	0.89	1.02	0.95
1.67	0.75	0.85	1.10	1.33	1.29
2.50	1.20	1.46	1.77	2.04	1.94
5.00	2.78	3.35	3.48	4.07	4.19
10.00	5.66	7.14	7.19	8.37	8.53
20.00	13.05	14.73	15.83	16.97	17.14

(b) Observed $(1-\alpha)100^{th}$ percentile of D_k

		N				
$\alpha(\%)$	D_α	50	100	200	500	1000
1.25	**1.5930**	1.4760	1.4938	1.5366	1.5643	1.5504
1.67	**1.5472**	1.4447	1.4532	1.4891	1.5147	1.4986
2.50	**1.4802**	1.3722	1.3998	1.4180	1.4392	1.4412
5.00	**1.3581**	1.2497	1.2924	1.2934	1.3154	1.3287
10.00	**1.2238**	1.1236	1.1585	1.1629	1.1901	1.1857
20.00	**1.0728**	0.9859	1.0045	1.0194	1.0350	1.0373

The observed percentiles and the percentage of rejected null hypotheses are summarized in Table 3.1. We can make following observations: 1) the type I error of test does not exceed the nominal level, 2) the size of the test approaches to the desired significance level α with the increase in the sample size N, and 3) the test is under-sized for smaller sample sizes. The severe problem with the size of the test for smaller sample size can be explained as follows. Calculation of test statistic, D_k, involves σ^2 and $\mathbf{V}_i$. However, in practice, the true values of σ^2 and $\mathbf{V}_i$ are unknown and we replace them by their estimates. A consistent estimator (e.g. ML- or REML-based) approaches the true value with an increasing sample size. However, the estimates might be biased for smaller sample sizes. To be precise, for smaller sample size, σ^2 and $\mathbf{V}_i$ may remain underestimated and this leads to smaller value of D_k which in turn results in a smaller size of the test. However, bias in estimation of σ^2 and $\mathbf{V}_i$ fades away with the increase in N and this increases the size of the test. We observe this trend in Table 3.1 as the size of test approaches the nominal level of type I error with the increase in sample size. However, the size of test remains smaller than nominal level even for the reasonably large N. The reduced size has been also reported in other tests based on the Brownian Bridge process. For example, Kolmogorov Smirnov test for normality (which also uses the Brownian Bridge as limiting distribution) is conservative (Birnbaum, 1952; Lilliefors, 1967; Massey Jr, 1951). As N exceeds 500, the size of the test is close to the nominal level of significance. As a remedy for smaller sample sizes, one might consider using a liberal α level or small sample distribution for D_k obtained through simulation.

Power

We generated observations for N subjects at $t = 0, 1, 2, 3$ from the following model to evaluate performance of instability test for X^G

$$X^G = c_{(g)}: \quad y_{it} = \beta_{0(g)} + \beta_{1(g)}t + b_i + \epsilon_{it},$$

Table 3.2: Power (%) of parameter instability test with continuous partitioning variable obtained in the simulation described in Section 3.6.1. Numbers corresponding to β_1 and β_0 represent the percentages of rejection associated with parameter instability for β_1 and β_0, respectively. The 'Overall' figures represent the percentage of at least one rejection out of the two.

		% of rejection					
		δ					
N	Parameter instability test	0	.25(−.25)	.50(−.50)	.75(−.75)	1.00(−1.00)	1.2(−1.2)
50	β_1	1.4	1.4(1.4)	1.6(1.6)	1.9(1.9)	2.3(2.3)	2.4(2.3)
	β_0	1.6	4.4(4.3)	16.9(16.6)	41.9(42.0)	70.2(70.6)	86.9(87.0)
	Overall	2.9	5.6(5.5)	17.9(17.6)	42.6(42.5)	70.5(70.8)	87.0(87.1)
100	β_1	1.5	1.6(1.6)	2.0(2.1)	2.5(2.6)	3.0(3.0)	3.2(3.2)
	β_0	1.7	(5.2(5.3)	18.7(19.7)	44.4(46.0)	72.9(73.9)	88.9(89.0)
	Overall	3.1	6.6(6.7)	19.8(20.8)	45.0(46.6)	73.1(74.2)	89.0(89.1)
200	β_1	1.8	1.9(1.8)	2.2(2.2)	2.7(2.7)	3.3(3.3)	3.5(3.4)
	β_0	1.9	5.6(5.3)	20.7(19.8)	47.5(46.8)	75.7(75.2)	90.1(89.8)
	Overall	3.6	7.4(6.8)	21.9(21.0)	48.2(47.4)	76.0(75.4)	90.6(89.9)
500	β_1	2.1	2.1(2.2)	2.7(2.5)	3.2(3.2)	3.6(3.7)	3.9(4.0)
	β_0	1.8	6.1(6.0)	21.4(20.1)	48.1(48.2)	76.6(76.6)	91.1(91.1)
	Overall	3.7	7.8(7.8)	22.8(22.2)	48.8(49.1)	77.0(77.0)	91.3(91.2)

$$\beta_{0(g)} = \beta_0 \qquad \beta_{1(g)} = \beta_1 + \delta \cdot \frac{c_{(g)}}{c_{(G)}}$$

We set $\beta_0 = 1$ and $\beta_1 = 2$. b_i, ϵ_{it} and X^G were generated similarly as before in Section 3.6.1. In this simulation, the parameter β_1 is not stable unless $\delta = 0$. We dealt with two parameters: β_0 and β_1, thus we will have two Brownian bridge processes. We adjusted the p-values according to the Hochberg's step-up procedure (Hochberg, 1988). We chose Hochberg's step-up procedure because it is relatively less conservative than the Bonferroni procedure (Hochberg and Tamhane, 1987). However, in principle, any multiple comparison procedure can be applied here.

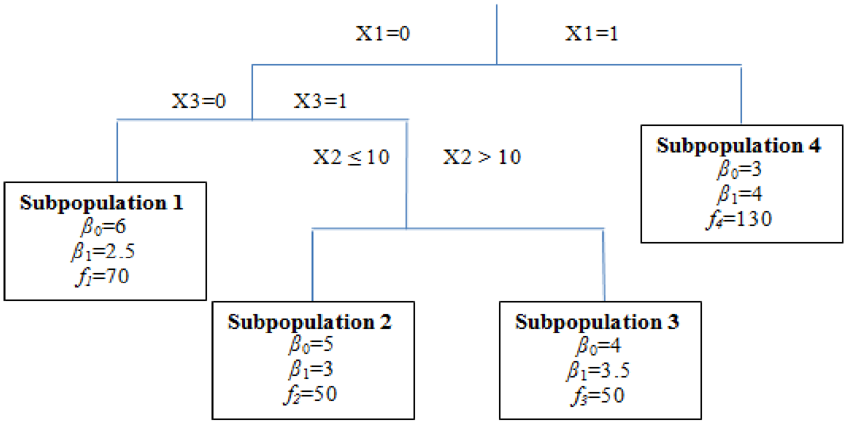

Figure 3.2: True tree structure for the simulation described in Section 3.6.2. In r^{th} subgroup, f_r observations were generated according to Eq. (3.6.2) with specified β_0 and β_1.

The results based on 10,000 simulation are displayed in Table 3.2. As the absolute value of δ deviates from zero, the power increases. The power of test is close to 80% and approaching the 90% mark as $|\delta| > 1$. The sign of δ does not influence the power of the test. Sizes of the test are very much in agreement with the first simulation study. As observed previously, the test is mildly conservative in the current simulation scenario as the observed level of type I error is consistently slightly below the nominal value $\alpha = 0.05$.

3.6.2 Performance of regression tree for longitudinal data

In this simulation, our goal is to assess the improvement in estimation due to LongCART algorithm when the population under consideration is truly heterogeneous. We have simulated observations for $N = 300$ subjects and these subjects come from one of the four different subgroups. Description of these subgroups is displayed in the form of a tree structure in Figure 3.2. The subgroups can be defined in terms of the partitioning variables X_1, X_2 and X_3. In rth subgroup ($r = 1, \cdots, 4$), the values for continuous response variable y were generated at $t = 0, 1, 2, 3$ according to following model:

$$y_{it} = \beta_{0r} + \beta_{1r}t + b_i + \epsilon_{it}; \qquad i = 1, \cdots, f_r \tag{3.6.2}$$

Table 3.3: Description of the mixed models used in Section 3.6.2 for the comparison with LongCART algorithm (Model 1). All models include random intercepts to account for the subject-specific effects.

	Predictors
Model 2	t
Model 3	t, X_1, X_2, X_3
Model 4	t, X_1, X_2, X_3, X_1X_2, X_1X_3, X_2X_3
Model 5	t, X_1, X_2, X_3, X_1X_2, X_1X_3, X_2X_3, $X_1X_2X_3$
Model 6	t, X_1, X_2, X_3, tX_1, tX_2, tX_3
Model 7	t, X_1, X_2, X_3, X_1X_2, X_1X_3, X_2X_3, tX_1, tX_2, tX_3, tX_1X_2, tX_1X_3, tX_2X_3
Model 8	t, X_1, X_2, X_3, X_1X_2, X_1X_3, X_2X_3, $X_1X_2X_3$, tX_1, tX_2, tX_3, tX_1X_2, tX_1X_3, tX_2X_3, $tX_1X_2X_3$

where $b_i \sim N(0,4)$ and $\epsilon_{it} \sim N(0,1)$. As displayed in Figure 3.2, the true values of β_1 were set at 2.5, 3.0, 3.5 and 4.0 and for β_0, the true values were set at 6, 5, 4 and 3, for the four subgroups, respectively. Further, observations were generated for $f_1 = 70$ individuals in subgroup 1, $f_2 = 50$ individuals in subgroup 2, $f_3 = 50$ individuals in subgroup 3, and $f_4 = 130$ individuals in subgroup 4. In order to study the performance of our algorithm constructing the longitudinal regression tree, we calculated the mean absolute deviation (MAD) in β_0 and β_1 in rth subgroup for each simulation as defined below

$$\text{MAD}(\hat{\beta}_{0r}), \beta^d_{0r} = \frac{1}{f_r}\sum_{j\in S_r} |\beta_{0r} - \hat{\beta}_{0j}| \qquad \text{MAD}(\hat{\beta}_{1r}), \beta^d_{1r} = \frac{1}{f_r}\sum_{j\in S_r} |\beta_{1r} - \hat{\beta}_{1j}| \tag{3.6.3}$$

where β_{0r} and β_{1r} are the true values of β_0 and β_1 in the rth subgroup and $\hat{\beta}_{0j}$ and $\hat{\beta}_{1j}$ are the corresponding estimates for the jth individual applyig longitudinal tree and then fitting mixed model in each subgroup. S_r is the set of indices for all individuals in the rth subgroup while f_r denotes their number.

Table 3.4: Summary of the results for the simulation described in Section 3.6.2

		Subpop 1		Subpop 2		Subpop 3		Subpop 4		Overall	
	ψ	$\bar{\beta}_0^d$	$\bar{\beta}_1^d$	$\bar{\beta}_0^d$	$\bar{\beta}_1^d$	$\bar{\beta}_0^d$	$\bar{\beta}_1^d$	$\bar{\beta}_0^d$	$\bar{\beta}_1^d$	$\bar{\beta}_0^d$	$\bar{\beta}_1^d$
Model 1	$8^\star$	0.29	0.08	0.31	0.09	0.32	0.09	0.16	0.03	0.241	0.064
Model 2	2	1.80	0.90	0.80	0.40	0.20	0.10	1.20	0.60	1.107	0.554
Model 3	5	1.44	0.90	0.62	0.40	0.32	0.10	0.89	0.60	0.879	0.554
Model 4	8	1.35	0.90	0.64	0.40	0.28	0.10	0.90	0.60	0.855	0.554
Model 5	9	1.35	0.90	0.63	0.40	0.28	0.10	0.90	0.60	0.854	0.554
Model 6	8	0.40	0.19	0.20	0.06	0.59	0.29	0.43	0.20	0.413	0.189
Model 7	14	0.29	0.07	0.31	0.11	0.30	0.11	0.29	0.07	0.294	0.085
Model 8	16	0.29	1.82	0.31	1.23	0.30	2.18	7.09	2.26	3.242	1.970

ψ: No.of parameters

$\bar{\beta}_0^d$=Average β_0^d; $\bar{\beta}_1^d$=Average β_1^d; β_0^d and β_1^d are defined in Eq. (3.6.3)

Model 1: Subgroups are extracted using LongCART algorithm and mixed model with time slope and random intercept fitted separately in each subgroup.

Models 2 - 8: Description is given in Table 3.3

* - In Model 1, 81% of time regression tree with 4 subgroups were extracted.

The simulation results are summarized in Table 3.4 based on 1000 simulations in each case. In each simulation, regression tree was constructed with the following specifications: (1) the overall significance level of instability test was set at 5%, (2) minimum node size for further split was set at 40, and (3) minimum terminal node size was set at 20. Recall that we are considering four subgroups in the current simulation. The LongCART algorithm extracted exactly four subgroups in 81% of the cases. Five subgroups were extracted in 16% of the cases and in these trees we observed a split in subgroup 4 which was not present in the true tree (see Figure 3.2). There were only 1.3% [1.6%] instances when three [six] subgroups were extracted.

For the comparison purposes, we considered seven linear mixed models (Models 2 - 8). These models are described in Table 3.3. The application of the LongCART algorithm (Model 1) shows comparatively larger improvements in the estimation of the coefficients in all four subgroups. Both the MAD($\hat{\beta}_0$) and MAD($\hat{\beta}_1$) were considerably smaller in Model 1 compared to the Models 2 - 8. The improvement in estimation of coefficients in regression tree was attributed to its ability to extract homogeneous subgroups and then fitting mixed model separately within each group. On the contrary, Models 2 - 8 assume either additive (Models 2 - 3) or an interaction (Models 4 - 8) mixed effects model for the entire population assuming parametric form for both covariates and time. These models do not capture the complexity for the heterogeneous subgroups and overestimate it for the homogeneous subgroups.

Inclusion of the interaction terms in the model does not necessarily take into account subgroup heterogeneity in the presence of continuous partitioning variable. For example, in Models 4 and 5 common slope is assumed for the entire population, but include interaction terms in the baseline effect; still, the absolute deviation in estimating β_0 is almost 2.5 times higher compared to that of in regression tree. Similarly, the Models 6 – 8 include interaction terms for both baseline and longitudinal effects, but again the absolute deviations in estimating β_0 and β_1 are higher compared to what we have obtained with the longitudinal regression tree.

Model 6 including the interaction terms with t and the partitioning variables is probably the most commonly used model in practice. However, the application of the LongCART algorithm offers a considerable improvement in the estimation compared to Model 6. Models 6 – 7 provide some improvement over regression tree in some of the subgroups. However, these improvements are comparatively rare and largely influenced by the fact how the subgroups are defined. We would close this section pointing out, apart from providing improvement in estimation, the LongCART algorithm also identifies the meaningful subgroups defined by the partitioning variables which would remain unidentified otherwise.

3.7 Application

We applied the LongCART algorithm to study the changes in concentration of choline in gray matter region of brain among HIV patients. These patients were enrolled in HIV Neuroimaging Consortium (HIVNC) which was formed to examine pattern or extent of brain injury in chronically infected patients on ARV treatment (Harezlak et al., 2011). Concentrations of choline were obtained via proton MRS. Choline is considered as marker of inflammation. An elevated concentration of choline is an indicator of increased cellular turnover. In general, the concentration of choline is increased in tumors and inflammatory processes. It has been found in previous studies that the concentrations of choline were elevated in all three brain regions among HIV patients (Chang et al., 2002). We considered a total of $\sum_{i=1}^{N} n_i = 780$ observations from $N = 239$ subjects. The longitudinal observations for each subject were within 3 years from baseline. The number of observations per subject ranged from 2 to 6 with median number of observations equal to 3. We observed overall significant decrease of 0.077 AU per year (p-value=0.003) in concentration of choline.

For the construction of regression tree we used baseline measurements of several clinical and demographic variables including sex, race, education, age, CD4 count, nadir CD4 count, duration of HIV, duration of antiretroviral (ARV) treatment, duration of highly active antiretroviral therapy (HAART), plasma HIV RNA count, antiretroviral CNS penetration-effectiveness (CPE) score and AIDS dementia complex (ADC) stage as partitioning vari-

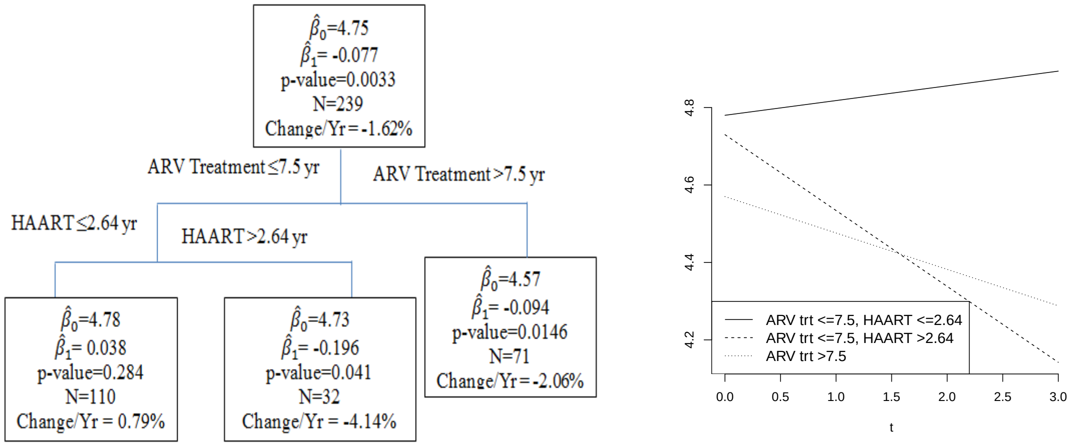

Figure 3.3: Longitudinal regression tree results for progression of concentration of *choline* as discussed in Section 3.7. *Left panel.* Longitudinal regression tree obtained via LongCART algorithm. The p-value in each node corresponds to the estimate of the slope β_1. *Right panel.* Estimated linear trajectory for longitudinal change within each sub-group obtained via fitting mixed effect model of form Eq. (3.7.1). This regression tree suggests duration of ARV treatment and HAART are significant determinants for longitudinal change of choline.

ables. In each node we consider fitting the following model separately

$$y_{it} = \beta_0 + \beta_1 t + b_i + \epsilon_{it} \tag{3.7.1}$$

where y_{it} indicates the measurement of the concentration of choline from the ith individual at time t (in years) and b_i is the subject specific intercept. It was assumed that b_i and ϵ_{it} are independently and normally distributed with mean equal to origin. LongCART algorithm was applied with the following specifications: (1) the significance level for individual instability test was set to 5%, (2) the minimum node size for further split was set to 50, and (3) the minimum terminal node size was set to 25. Figure 3.3 displays the estimated longitudinal regression tree with the estimates of β_0 and β_1 for each terminal node or subpopulation and the plot of estimated linear trajectories within each subgroup.

Duration of ARV treatment (p-value=0.004) and HAART (p-value=0.004) seem to influence the change in concentration of choline over time. Improvement in deviance due to application of LongCART algorithm was 519 (log-likelihoods were -1427 vs. -1687; with 4 degrees of freedom). ARV treatment for over 7.5 years not only helped to reduce baseline concentration of choline, but also resulted in a significant decrease of 0.094 per year (p-value=0.015). A higher baseline value of choline concentration was observed among those who received ARV treatment for at most 7.5 years; however, a longer period of HAART therapy in them led to significant decrease of 0.196 per year (p-value=0.041) in concentration over time. We did not observe any decrease among those who received ARV treatment for less than 7.5 years and HAART therapy for 2.64 years.

In summary, both the longer duration of ARV treatment and HAART resulted in reduction of concentration in choline. However, the rate of reduction is almost double (4.14% vs 2.06%) when patients were on HAART compared to only ARV treatment (see Figure 3.3). This suggests that both ARV treatment and HAART are effective in controlling brain inflammation via reducing choline concentration, however, HAART should be preferred whenever possible with appropriate advice from medical doctor. Finally, all these interpretable subgroups along with a significant improvement in overall model fit suggests underlying heterogeneity in the population in terms of longitudinal change in concentration of choline. Thus considering a traditional linear mixed effects model for the entire population is not defensible.

3.8 Discussion

The longitudinal profile in a population may be influenced by several baseline characteristics. This may be true both in observational and controlled studies (e.g., clinical trials). The most common strategy to incorporate the effect of baseline attributes in a traditional linear mixed effects model is to include these baseline characteristics (and probably their interaction) by including them as covariates in the model. However, this approach has its own limitation as discussed in Introduction section. Longitudinal trees (regression trees for longitudinal data) are extremely useful to identify the heterogeneity in longitudinal trajectories in a given population in a non-parametric way. We proposed LongCART algorithm for the construction of longitudinal tree which controls type I error at the time of taking decision about splitting at each node. Secondly, LongCART algorithm reduces the computation time substantially as we first choose the partitioning variable and then evaluate the goodness of fit criterion at all cut-off points of the selected partitioning variable only. Both the instability test and the LongCART algorithm discussed in this paper are based on the score process. We can apply a similar algorithm in other scenarios as long as we can obtain (or approximate) an expression for the score function and the Hessian matrix in a tractable form. For example, there is a scope to extend LongCART algorithm in the generalized linear mixed effects model (GLMM) or multiple response variables settings. There is a plethora of evidence for the heterogeneity of longitudinal profiles; for example Leuchter et al. (2002) reported heterogeneity in progression of depression in a double-blind randomized trial. Other reported examples include heterogeneous trend in aggressive behavior among different classes of students (Ialongo et al., 1999; Kellam et al., 1994), differential math achievement among different dropout groups (Muthén, 2004), and varying age-crime curve among different birth cohorts (Loughran and Nagin, 2006).

Both the instability test and the LongCART algorithm discussed in this paper are based on the score process. This increases the utility of the proposed method beyond the application to the mixed effects longitudinal models studied in this paper. We can apply a similar algorithm in other scenarios as long as we can obtain (or approximate) an expression for the score function and the Hessian matrix in a tractable form. For example, we can apply our method in the generalized linear mixed effects model (GLMM) where score function is difficult to obtain, but can be approximated. With the binary response it would be analogous to the construction of a classification tree with the longitudinal data. Another extension, we currently work on is in the context of regression tree construction with multiple response variables, both in cross-sectional and longitudinal setting.

One of the drawbacks of the proposed method is an underestimation of the nominal test size, especially for the small sample sizes. As already mentioned in Section 3.6.1, this finding is consistent with other score type tests that use Brownian Bridge as limiting process. One way to address this issue is by increasing the nominal type I error level. A more principled approach to address this problem would be to find the exact distribution through a simulation study. As an follow-up work, it would be interesting to compare the results of the parameter instability test for continuous partitioning variable (and, regression tree in general) between the exact and the limiting distributions. We end our conclusions by discussing the possibility of sup-Wald type test (e.g. see Andrews, 1993) as an alternative to the score test. In general, Wald test has higher power compared to score test (**?**), however, the former is often criticised for not maintaining the type I error. Further, we are not aware of any result on the convergence of the test statistic distribution used in sup-Wald type tests. Unavailability of limiting distribution for sup-Wald type test makes it infeasible to use in construction of a longitudinal tree.

Appendix

Proofs of expressions for power in parameter instability tests

Proof of Theorem 3.4.1

Proof. Using Taylor series expansion we can write

$$f(\mathbf{y},\boldsymbol{\theta}_{(g)}) \doteq f(\mathbf{y},\boldsymbol{\theta}_0)\left\{1+\mathbf{u}(\mathbf{y},\boldsymbol{\theta}_0)^\top\boldsymbol{\delta}\circ\mathbf{h}\Big(\frac{c_{(g)}}{c_{(G)}}\Big)\frac{1}{\sqrt{N}}\right\}$$

Consequently,

$$\begin{aligned} E_{\boldsymbol{\theta}_g}[\mathbf{u}(\mathbf{y},\boldsymbol{\theta}_0)] &= \int u(\mathbf{y},\boldsymbol{\theta}_0)f(\mathbf{y},\boldsymbol{\theta}_{(g)})dy = E_{\boldsymbol{\theta}_0}[\mathbf{u}(\mathbf{y},\boldsymbol{\theta}_0)]+\mathbf{J}\cdot\boldsymbol{\delta}\circ\mathbf{h}\Big(\frac{c_{(g)}}{c_{(G)}}\Big)\frac{1}{\sqrt{N}} \\ &= \mathbf{J}\cdot\boldsymbol{\delta}\circ\mathbf{h}\Big(\frac{c_{(g)}}{c_{(G)}}\Big)\frac{1}{\sqrt{N}} \end{aligned} \tag{3.8.1}$$

It can be shown that

$$\mathrm{cov}_{H_1}[\mathbf{W}_N(t,\boldsymbol{\theta}_0)] = \mathrm{cov}_{H_0}[\mathbf{W}_N(t,\boldsymbol{\theta}_0)]+O\left(\frac{1}{N}\right) \doteq \mathbf{J} \tag{3.8.2}$$

Proof of Theorem 3.4.1 follows from the definition of non-central chi-square distribution.

□

Proof of Theorem 3.4.2

Proof. Using (3.8.1) and (3.8.2),

$$E_{H_1}[\mathbf{W}_N(t,\boldsymbol{\theta}_0)] = \mathbf{J}\frac{1}{N}\sum_{i=1}^{M_g}\delta\circ\mathbf{h}\Big(\frac{c_{(g)}}{c_{(G)}}\Big) = \mathbf{J}\cdot t_g\cdot\boldsymbol{\delta}\circ\bar{\mathbf{h}}_g \qquad t\in[t_g,t_{g+1})$$

This time using the FCLT along with Cramer-Wold device we can show that

$$\mathbf{W}_N(t, \boldsymbol{\theta}_0) \longrightarrow_d \mathbf{J} \cdot t_g \cdot \delta \circ \bar{\mathbf{h}}_g + \mathbf{Z}(t) \qquad t \in [t_g, t_{g+1})$$

Therefore, for $t \in [t_g, t_{g+1})$,

$$\mathbf{W}_N(t, \hat{\boldsymbol{\theta}}) = \mathbf{W}_N(t, \boldsymbol{\theta}_0) - t_g \, \mathbf{W}_N(1, \boldsymbol{\theta}_0) + o_p(1) \longrightarrow_d \mathbf{J} \cdot t_g \cdot \delta \circ (\bar{\mathbf{h}}_g - \bar{\mathbf{h}}) + \{\mathbf{Z}(t) - t \cdot \mathbf{Z}(1)\}$$

Thus under H_1,

$$\mathbf{M}_N(t, \hat{\boldsymbol{\theta}}) = \hat{\mathbf{h}}^{-1/2} \mathbf{W}_N(t, \hat{\boldsymbol{\theta}}) \longrightarrow_d \mathbf{J}^{1/2} \cdot t_g \cdot \boldsymbol{\delta} \circ (\bar{\mathbf{h}}_g - \bar{\mathbf{h}}) + \mathbf{W}^0(t) \qquad t \in [t_g, t_{g+1})$$

□

Chapter 4

Identifying factors influencing longitudinal changes of brain metabolites in HIV-infected subjects enrolled in HIVNC study

4.1 Introduction

The introduction of highly active antiretroviral therapy (HAART) or antiretroviral (ARV) treatment has resulted in marked improvement in survival with a substantial increase in the number of asymptomatic HIV infected patients with improved immunological status (Antinori et al., 2007; Palella Jr et al., 1998). However, HIV may continue to affect the brain even in the presence of HAART (Cysique et al., 2004; Dore et al., 1999; Robertson et al., 2004, 2007; Tozzi et al., 2007; Valcour et al., 2006). The basal ganglia are affected early in the course of the disease and carry the heaviest HIV load of all brain structures; however, white matter and gray matter are also typically involved (Navia et al., 1986, Pumarola-Sune et al., 1987, Meyerhoff et al., 1995, Kumar et al., 2009). HIV infection in brain results in imbalance in brain metabolites leading to cognitive, motor and behavioral impairments and other neurological complications. Therefore, coupled with increased survival and an aging patient population, the HIV infection in brain could result in an increase in the prevalence of impairment in the chronically infected and treated population.

Proton magnetic resonance spectroscopy (1H-MRS) provides a sensitive and noninvasive in-vivo method to detect changes in levels of specific cerebral metabolites, including N-acetylaspartate (NAA), choline (Cho), myo-inositol (MI), glutamate and glutamine (Glx) and creatine (Cr). A number of MRS-derived abnormalities were described including reduced levels of NAA as a marker of neuronal metabolism, whereas elevations in the Cho considered as markers of cell membrane damage, MI as a marker of neuroinflammation in the context of HIV, Glx as a major excitatory neurotransmitter and Cr as a marker of metabolism and cellular energy. MRS may provide a useful marker for early detection of brain injury associated with HIV infection Tracey et al. (1996). For example, a common finding in most of the MRS studies in HIV patients is the reduced level of NAA and NAA to Cr ratio (Laubenberger et al., 1996; López-Villegas et al., 1997). However, there are evidences that HAART reverses brain metabolite abnormalities (e.g., see Chang et al., 1999).

We are interested in studying the progression of concentration of brain metabolites among HIV patients. There are evidences (e.g. see Chang et al., 1999) that this longitudinal change might be affected by the clinical and demographic factors. Despite considerable evidence of HIV associated brain disease, there have been no published longitudinal studies to date of metabolite abnormalities in the setting of chronic infection and treatment. The HIV Neuroimaging Consortium (HIVNC) was formed to examine pattern or extent of brain injury in chronically infected patients on ARV treatment. It is a prospective multicenter study of chronically HIV-infected individuals. Recently, Gongvatana et al. (2013) studied the progressive changes in cerebral metabolites in this multicenter MRS study. We have hypothesized that one or more demographic and clinical factors influence the progression of brain metabolites. This would help us to identify high risk subgroups early and will allow us to design more sophisticated treatment for them.

4.2 Patients and methods

4.2.1 Participants

Our study cohort was comprised of 243 chronically HIV-infected patients enrolled in HIV Neuroimaging Consortium (HIVNC), a longitudinal study of HIV associated brain injury, at the following sites: University of California (San Diego), University of California (Los

Angeles), Harbor-UCLA, Stanford University, University of Colorado, University of Pittsburgh, and Rochester University. At the time of enrollment, patients were on stable ARV treatment with any Food and Drug Administration (FDA)-approved therapy. Details of the inclusion and exclusion criteria for this cohort have been described elsewhere (Harezlak et al., 2011). For the purpose of longitudinal analysis, baseline was defined as the time of enrollment. We considered only observations within 3 years from baseline. Subjects with at least one post-baseline measurements within 3 years from baseline were only included and this criteria was evaluated for each metabolites separately.

4.2.2 Brain metabolites

We studied longitudinal changes of concentration of N-acetylaspartate (NAA), choline (Cho), myo-inositol (MI), glutamine and glutamate (Glx) and creatine (Cr) collected from three different brain regions, namely, mid-frontal cortex (gray matter), mid-frontal centrum semiovale (white matter), and basal ganglia. NAA, Cho, MI, Glx and Cr are commonly seen on MR spectrum (Hesselink, 2013). The concentrations of individual metabolites were determined using the LC Model spectral analysis software (Provencher, 2005) from single-voxel ^{1}H spectra. Please see Harezlak et al. (2011) for a more detailed description of the imaging study.

4.2.3 Baseline factors

We considered several clinical and demographic variables at baseline as candidate baseline factors that might influence the longitudinal change of metabolites. This included age, gender, race (White vs Non-white), education (college vs. no college), duration of HIV infection, ARV treatment and HAART, and laboratory measures such as baseline CD4 count, nadir CD4 count, plasma HIV RNA levels (dichotomized as detectable, if $>$ 400 copies/ml – versus undetectable), AIDS dementia complex (ADC) stage (no impairment, subclinical impairment and clinical impairment), and CNS penetration effectiveness (CPE) score of the antiretroviral regimen received at baseline. The CPE score is a measure of the relative effectiveness of antiretroviral regimens to cross the blood–brain barrier with 0 as the lowest penetration and increases as the degree of penetration increases. ADC is one of the most common and clinically important CNS complications of late HIV-1 infection leading to mental and motor impairment (0: No impairment, 0.5: subclinical impairment and ≥ 1: impaired).

4.2.4 Statistical methods

We summarized all the baseline factors and concentration of metabolites at baseline using descriptive statistics including number of observations, median and inter-quartile range. Our primary goal was to study the influence of different baseline factors on the longitudinal change of metabolite concentrations non-parametrically. We applied LongCART algorithm of constructing regression tree for longitudinal data (proposed and discussed in Chapter 3) for this purpose. We considered 12 baseline factors (see Section 4.2.3) as partitioning variables to construct regression tree. LongCART algorithm, at each non-terminal node, identifies the best splitting point for binary partitioning in two steps: 1) First, a statistical test (also referred as parameter instability test) is carried out for each of the partitioning variables separately, and the partitioning variable with smallest p-value is chosen. 2) In second step, at each cut-off point of the partitioning variable (chosen in first step), the improvement in deviance due to binary partitioning were obtained. Then, the best split was identified as the cut-off point that provided maximum improvement in deviance. The overall error rate at each split was controlled at 10% level of significance according to Hommel's step-up procedure (Hommel, 1988).

We applied following additional criteria in construction of regression tree: 1) minimum number of individuals in a terminal node as 15 and b) minimum number of individuals in a terminal node for further splitting as 30. Within each node of regression we considered fitting regression model with time slope and random subject intercept as follows:

$$y_{it} = \beta_0 + \beta_1 \, t + b_i + \epsilon_{it} \qquad 0 \leq t \leq 3, \;\; i = 1, \cdots, N$$

where, y_{it} is the concentration of a specific metabolite from subject i at time t from baseline. β_0 and β_1 are population level intercept and time slope, b_i is the subject specific random intercept and ϵ_{it} is the error term. In addition, we assumed b_i and ϵ_{it} were independently and normally distributed. All analyses were carried out in `R-2.15.1` (R Core Development System: `http//www.r-project.org`).

4.3 Results

4.3.1 Participant characteristics

Summary characteristics are provided for the 243 participants having at least one post-baseline observations within 3 years from baseline in Tables 4.1 - 4.2. The median age of patients was 47 years. 84.4% of subjects were male and 70% of subjects were white. The median duration of HIV was 12 year. The patients received ARV treatment and HAART for

Table 4.1: Descriptive summary of continuous baseline factors

	N	Min	Q1	Median	Q3	Max
Age [yr]	243	23.00	42.00	47.00	53.00	70.00
Duration of HIV [yr]	242	0.00	7.00	12.00	17.00	26.00
Duration of ARV treatment [yr]	242	0.00	1.80	3.90	8.20	19.20
Duration of HAART [yr]	210	0.00	0.73	1.55	3.07	10.87
CD4 count	239	10.00	208.00	326.00	473.00	1445.00
Nadir CD4 count	242	0.00	12.00	36.00	96.00	811.00

Min: Minimum; Q1: First quartile; Q3: Third quartile; Max: Maximum

Table 4.2: Summary of categorical baseline factors

	N [%]
Gender	
Female	38 [15.6]
Male	205 [84.4]
Race	
White	170 [70.0]
Non-white	73 [30.0]
Education	
High school or less	95 [39.1]
College or higher	148 [60.1]
CPE score	
0.0	5 [2.2]
0.5	20 [8.7]
1.0	53 [23.1]
1.5	51 [22.3]
2.0	41 [17.9]
2.5	34 [14.9]
3.0	9 [3.9]
3.5	11 [4.8]
4.0	5 [2.2]
ADC stage	
No impairment	143 [58.8]
Sub-clinical impairment	54 [22.2]
Clinical impairment	41 [16.9]
Missing	5 [2.1]
Plasma HIV RNA	
Detectable*	190 [78.2]
Not detectable	49 [20.2]
Missing	4 [1.6]

Percentages were calculated using total number of subjects of 243.

$^* > 400$ copies/ml

Table 4.3: Descriptive summary of MRS metabolites at baseline

	N	Minimum	Q1	Median	Q3	Maximum
White matter						
Cr	230	3.247	6.905	8.192	9.148	14.700
NAA	241	4.058	7.408	8.200	8.939	11.840
Cho	241	2.894	4.437	4.806	5.272	7.589
MI	241	0.834	1.459	1.674	1.896	4.403
Glx	240	1.324	4.447	5.034	5.909	10.200
Basal ganglia						
Cr	202	5.752	8.653	10.170	11.98	15.590
NAA	230	2.629	7.156	7.940	8.585	11.060
Cho	230	1.737	4.814	5.308	5.886	8.552
MI	229	0.717	1.291	1.485	1.642	4.543
Glx	227	1.221	3.518	4.169	4.790	7.512
Gray matter						
Cr	226	3.497	8.830	10.180	11.080	15.080
NAA	239	2.678	6.174	6.898	7.41	10.200
Cho	239	1.577	4.25	4.723	5.204	7.570
MI	238	0.383	1.083	1.249	1.401	2.063
Glx	238	1.937	3.734	4.355	5.052	8.146

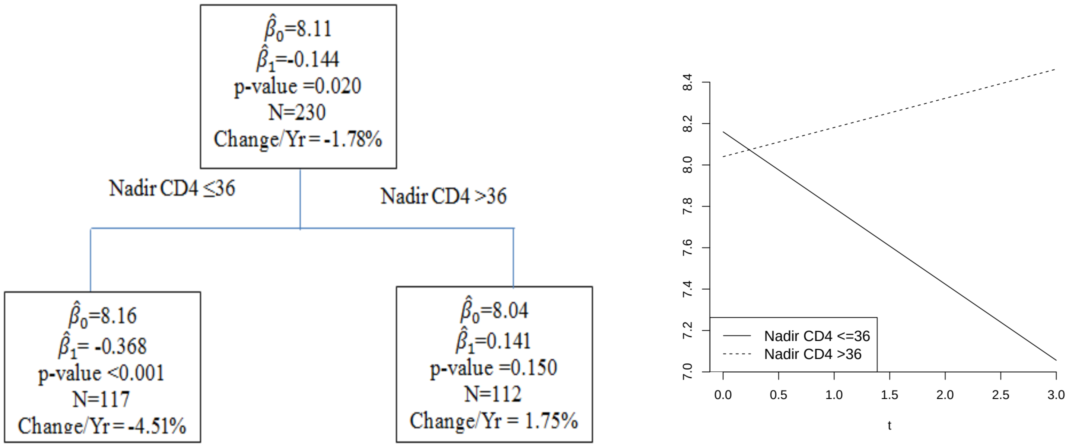

Figure 4.1: Influence of baseline factors in longitudinal progression of Cr in white matter. Regression tree for longitudinal progression is displayed on the left. The figure on right displays the estimated longitudinal profiles for the subgroups extracted by the regression tree.

a median period of 3.90 and 1.55 years, respectively. 78.2% of the individuals had detectable plasma HIV RNA levels (i.e. $>$ 400 copies/ml). For 58.8% of subjects no CNS impairment was observed. Subclinical and clinical CNS impairment were observed among 22.2% and 16.9% subjects, respectively. The median values of current and nadir CD4 counts were 326 and 36, respectively. The summary of the baseline values of MRS metabolites are displayed in Table 4.3.

4.3.2 Longitudinal change of concentration of metabolites

We observed overall decrease in most of the metabolites under consideration in the three brain regions except for Cho in basal ganglia. The estimate of intercept and slope for each of the metabolites in the overall population are provided in Table 4.4.

Metabolite change in white matter

Creatine: The concentration of Cr was significantly (p-value=0.020) decreased at 0.144 per year among 230 individuals. Nadir CD4 was found to be significant (p-value<0.001) determinant of longitudinal progression of Cr (see Figure 4.1). A significant (p-value<0.001)

Table 4.4: Overall trend of MRS metabolites and factor influencing the longitudinal change

	No. of subjects	Total obs	$\hat{\beta}_0$	$\hat{\beta}_1$ (p-value)	Factors (p-value) influencing longitudinal change
White matter					
Cr	230	721	8.11	-0.144 (0.020)	Nadir CD4 (< 0.001)
NAA	241	802	8.13	-0.144 (< 0.001)	ARV trt dur.(< 0.001)
Cho	241	802	4.86	-0.039 (0.085)	ARV trt dur. (0.005)
MI	241	802	1.70	-0.030 (0.022)	-
Glx	240	795	5.17	-0.092 (0.013)	-
Basal ganglia					
Cr	202	628	10.11	-0.291 (< 0.001)	ARV trt dur. (< 0.001)
NAA	230	740	7.82	-0.010 (0.807)	-
Cho	230	737	5.33	0.034 (0.246)	-
MI	229	735	1.49	-0.011 (0.301)	ARV trt dur.(< 0.001)
					ARV trt dur. (< 0.001)
Glx	227	722	4.20	-0.011 (0.760)	-
Gray matter					
Cr	226	702	9.92	-0.510 (< 0.001)	Age (0.009)
NAA	239	780	6.78	-0.111 (< 0.001)	ARV trt dur. (< 0.001)
Cho	239	780	4.75	-0.077 (0.003)	ARV trt dur. (0.004)
					HAART dur. (0.004)
MI	238	777	1.24	-0.025 (0.003)	CD4 count (0.002)
Glx	238	776	4.44	-0.101 (0.002)	ARV trt dur. (0.004)

obs: observation; dur: duration; trt: treatment
Estimate of β_0 and β_1 were obtained from all the subjects.

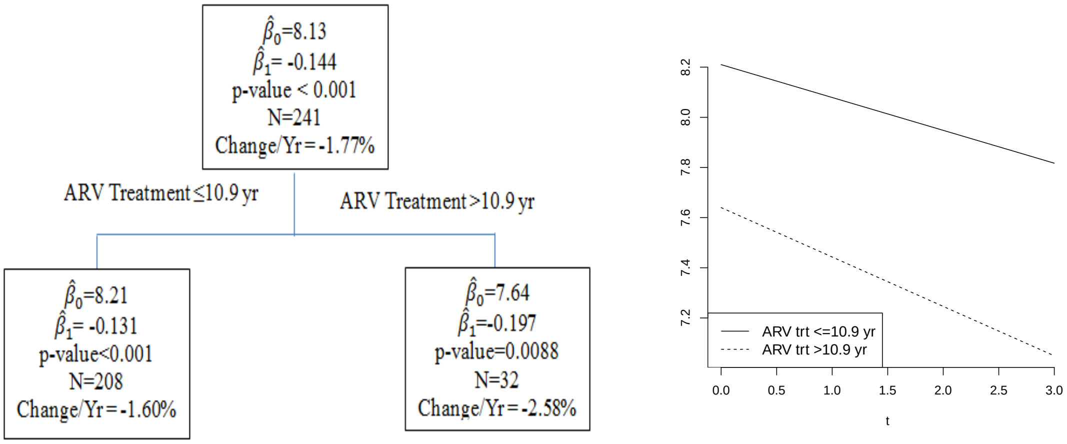

Figure 4.2: Influence of baseline factors in longitudinal progression of NAA in white matter. Regression tree for longitudinal progression is displayed on the left. The figure on right displays the estimated longitudinal profiles for the subgroups extracted by the regression tree.

sharp decrease of 0.368 per year in concentration of Cr was observed among those who had nadir CD4 count smaller than 36. On the contrary, the individuals with higher CD4 count experienced increase in concentration of Cr at 0.141 per year.

NAA: The concentration of NAA was significantly (p-value<0.001) decreased at 0.144 per year among 241 individuals. Duration of ARV treatment was found to be significant determinant of longitudinal progression of NAA (see Figure 4.2). The baseline concentration of NAA was decreased with longer period of ARV treatment. Further, the prolonged period of ARV treatment was associated with greater rate of decrease in NAA. We observed decrease of 0.131 and 0.197 per year among those who had ARV treatment for less than 10.9 year, and greater than 10.9 years, respectively.

Choline: The concentration of Cho was decreased at 0.039 per year among 241 individuals. Duration of ARV treatment was found to be significant (p-value=0.005) determinant of longitudinal progression of Cho (see Figure 4.3). Those who received ARV treatment for

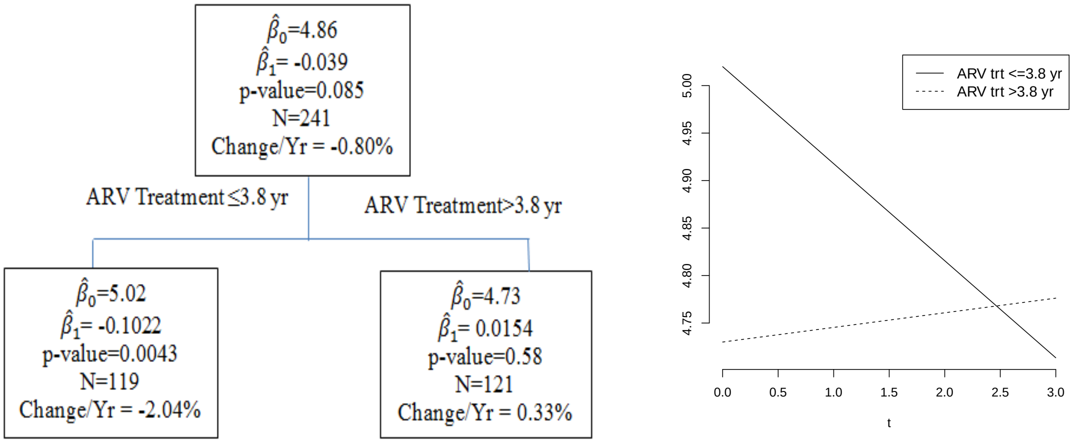

Figure 4.3: Influence of baseline factors in longitudinal progression of Cho in white matter. Regression tree for longitudinal progression is displayed on the left. The figure on right displays the estimated longitudinal profiles for the subgroups extracted by the regression tree.

greater than 3.8 years had smaller baseline concentration of Cho (4.73 vs 5.02) and also experienced significant decrease (p-value=0.0043) at the rate of 0.102 per year.

In overall, the concentration of MI (0.030 per year, p-value=0.022) and Glx (0.092 per year, p-value=0.013) were decreased in white matter region; however, we did not find any factor influencing their change over time.

Metabolite change in basal ganglia

Creatine: The concentration of Cr was significantly (p-value<0.001) decreased at 0.291 per year among 202 individuals. Duration of ARV treatment was found to be significant (p-value<0.001) determinant of longitudinal progression of Cr (see Figure 4.4). A significant (p-value<0.001) sharp decrease of 1.018 per year in concentration of Cr was observed among those who were receiving ARV treatment for at least 11.3 years at baseline. However, the rate of decrease was only about one-nine-th among the individuals with shorter duration of ARV treatment.

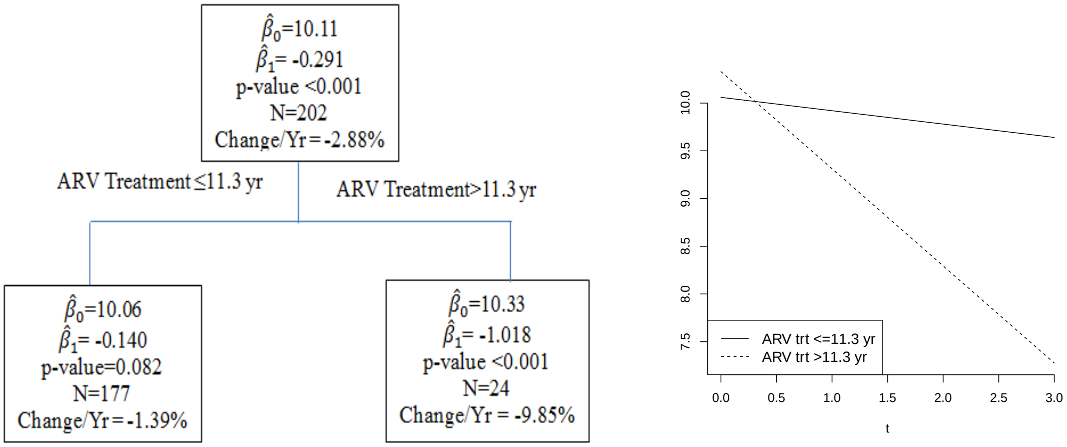

Figure 4.4: Influence of baseline factors in longitudinal progression of Cr in basal ganglia. Regression tree for longitudinal progression is displayed on the left. The figure on right displays the estimated longitudinal profiles for the subgroups extracted by the regression tree.

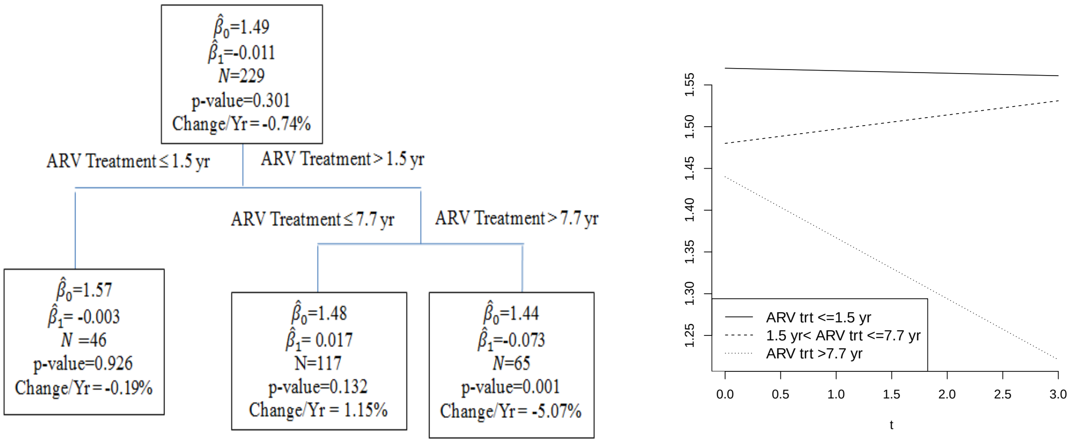

Figure 4.5: Influence of baseline factors in longitudinal progression of MI in basal ganglia. Regression tree for longitudinal progression is displayed on the left. The figure on right displays the estimated longitudinal profiles for the subgroups extracted by the regression tree.

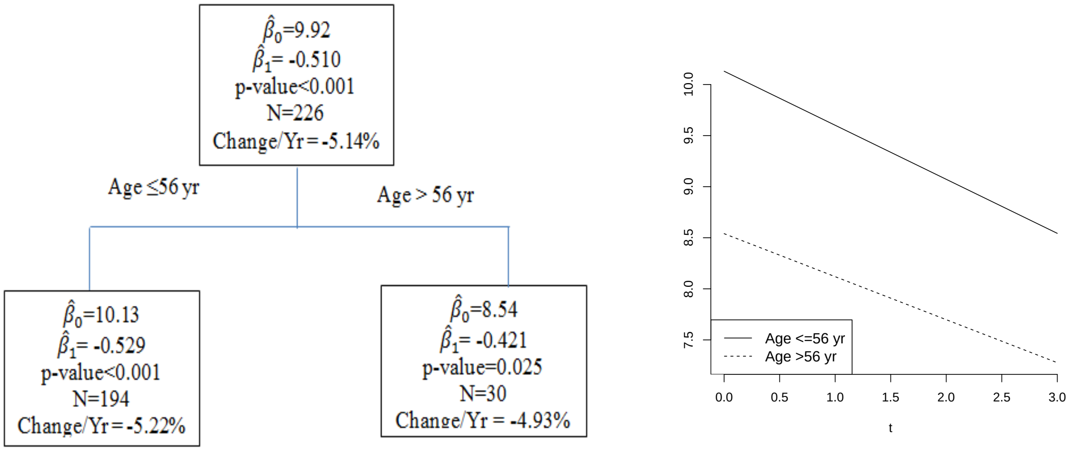

Figure 4.6: Influence of baseline factors in longitudinal progression of Cr in gray matter. Regression tree for longitudinal progression is displayed on the left. The figure on right displays the estimated longitudinal profiles for the subgroups extracted by the regression tree.

Myo-inositol: The concentration of MI decreased at 0.011 per year (p-value=0.301) among 229 individuals. Duration of ARV treatment was found to be significant determinant (p-value< 0.001) of longitudinal progression of MI (see Figure 4.5). The baseline concentration of MI decreased with longer period of ARV treatment. We observed MI concentration of 1.57, 1.48 and 1.44 at baseline, among those who had ARV treatment for less than 1.5 year, between 1.5 and 7.7 year and greater than 7.7 years, respectively. Significant decrease in MI concentration (0.073 per year, p-value= 0.001) was observed among those who received ARV treatment for over 7.7 years.

Metabolite change in gray matter

Creatine: The concentration of Cr was significantly (p-value<0.001) decreased at 0.510 per year among 226 individuals. Age was found to be significant determinant (p-value= 0.009) of longitudinal progression of MI (see Figure 4.6). Increase in age was associate with smaller baseline concentration of Cr and smaller decrease.

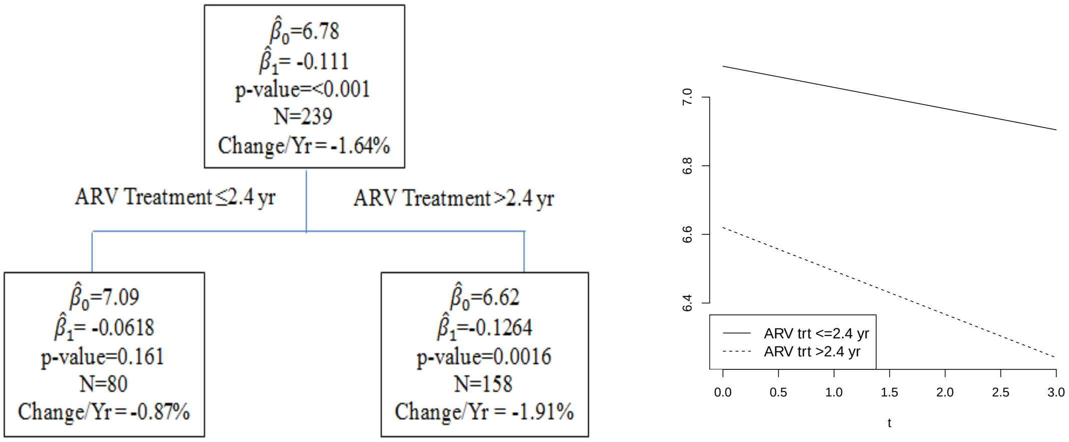

Figure 4.7: Influence of baseline factors in longitudinal progression of NAA in gray matter. Regression tree for longitudinal progression is displayed on the left. The figure on right displays the estimated longitudinal profiles for the subgroups extracted by the regression tree.

NAA: The concentration of NAA was significantly (p-value<0.001) decreased at 0.111 per year among 239 individuals. Duration of ARV treatment was found to be significant determinant (p-value< 0.001) of longitudinal progression of NAA (see Figure 4.7). The baseline concentration of NAA was smaller among those who received ARV treatment for greater than 2.4 years compared to those who did not (6.62 vs 7.09). The rate of decrease in concentration of NAA was also almost double (0.126 vs 0.062) among those who received the ARV treatment for longer period.

Choline: We observed overall significant decrease of 0.077 per year (p-value=0.003) in concentration of Cho. Duration of ARV treatment and HAART period seem to influence the change in concentration of Cho over time. ARV treatment for over 7.5 years was associated with decreased baseline concentration of Cho and significant decrease of 0.094 per year (p-value=0.015). A higher baseline value of Cho concentration was observed among those who received ARV treatment for at most 7.5 years; however, a longer period of HAART therapy in them was found to be associated with significant decrease (0.196 per year; p-

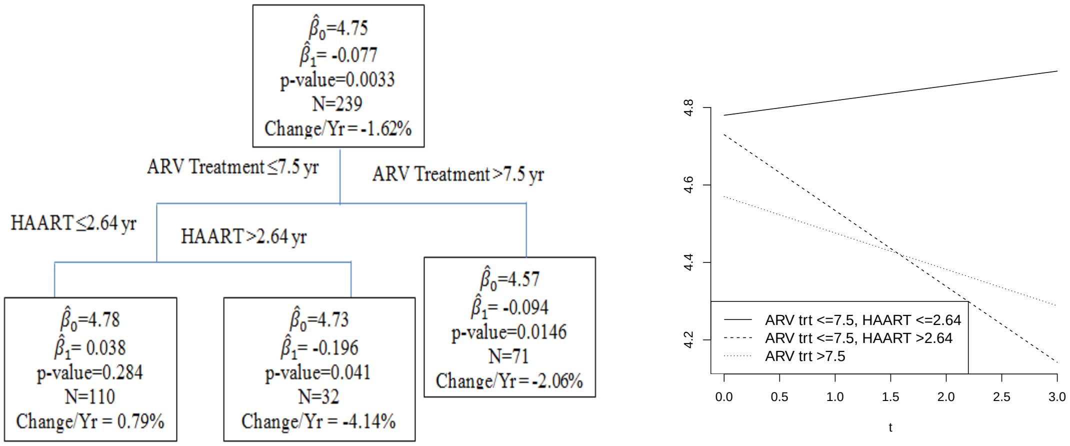

Figure 4.8: Influence of baseline factors in longitudinal progression of Cho in gray matter. Regression tree for longitudinal progression is displayed on the left. The figure on right displays the estimated longitudinal profiles for the subgroups extracted by the regression tree.

value=0.041) in concentration over time. We did not observe any decrease among those who received ARV treatment for less than 7.5 years and HAART therapy for 2.64 years.

Myo-inositol: The concentration of MI was significantly (p-value=0.003) decreased at 0.025 per year among 238 individuals. CD4 count appeared to be significant determinant (p-value= 0.002) of longitudinal progression of MI (see Figure 4.9). The concentration of MI was decreased significantly (p-value<0.001) at 0.057 per year in those who had CD4 count greater than 359.

Glutamine-Glutamate: The concentration of Glx was significantly (p-value=0.002) decreased at 0.101 per year among 238 individuals. Duration of ARV treatment was found to be significant determinant (p-value= 0.004) of longitudinal progression of Glx (see Figure 4.10).The baseline value of concentration of Glx was smaller (4.22 vs 4.68) among those who received ARV treatment for greater than 3.7 years; however, the rate of decrease was smaller (0.075 vs 0.135 per year).

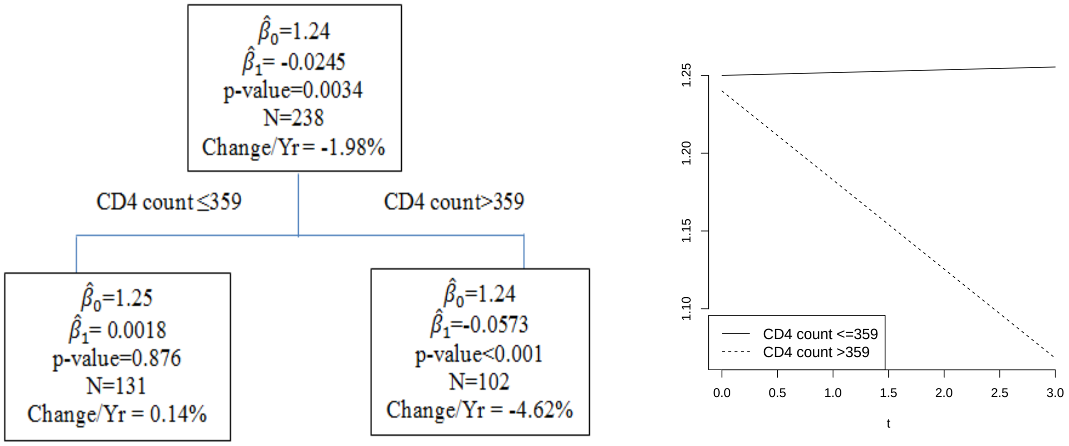

Figure 4.9: Influence of baseline factors in longitudinal progression of MI in gray matter. Regression tree for longitudinal progression is displayed on the left. The figure on right displays the estimated longitudinal profiles for the subgroups extracted by the regression tree.

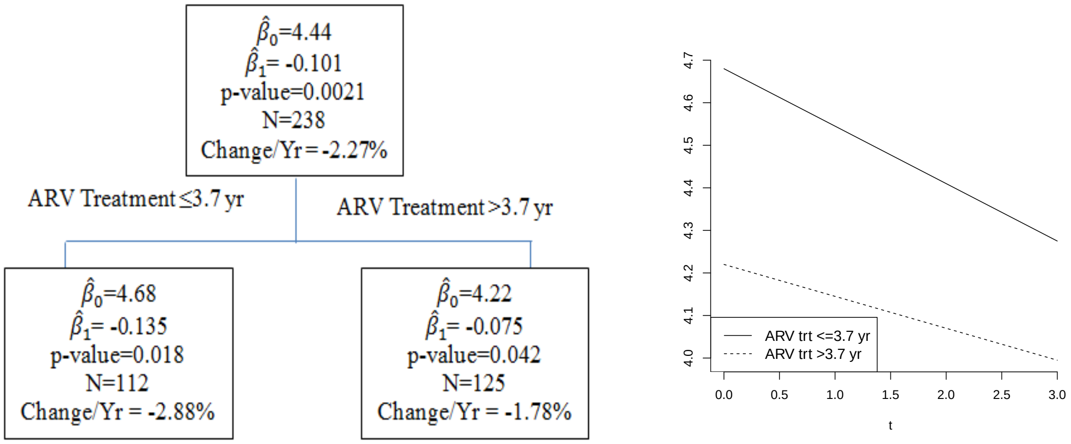

Figure 4.10: Influence of baseline factors in longitudinal progression of Glx in gray matter. Regression tree for longitudinal progression is displayed on the left. The figure on right displays the estimated longitudinal profiles for the subgroups extracted by the regression tree.

4.4 Discussion

In this analysis we have studied the influence of baseline factors on the longitudinal change in concentration of brain metabolites non-parametrically via constructing regression tree. Usually the influence of baseline factors is adjusted via including them (and possibly their interactions) as covariate in the longitudinal model. By using regression tree methodology, we were able to explore more complex relationships among baseline covariates than what could be discovered including them as covariates in the model. For example, the observed influence of duration of HAART and duration of (non HAART) ARV therapy on the progression of Cho in gray matter region of brain would be difficult to explore via traditional mixed effects model without applying regression tree technique (see Figure 4.8). The important feature of the regression tree technique is that we can identify distinct subpopulations with differential longitudinal change. The subpopulations are characterized by baseline factors and the entire process is data-driven.

In general, we found that almost all the metabolites under consideration were decreased with time in the three brain regions with only exception in choline in basal ganglia. The longitudinal change in brain metabolites were influenced by baseline factors, namely, duration on ARV treatment and HAART, CD4 counts and age. These metabolites changes are often found to be associated to the course of the disease. Hence, the study of longitudinal progression of these metabolites may help us to understand the progression of disease due to imbalance of these metabolites and consequently, helps to identify the high risk group in advance.

Creatine (Cr). Creatine is considered as marker for metabolism and cellular energy. In general, sick people tend to have reduced concentration of creatine, however, there is inconsistency in the relationship between Cr and cognitive function (Ross and Sachdev, 2004).

A reduced level of Cr has been observed in previous studies in gray matter region of HIV patients; however elevation was observed in white matter and basal ganglia (Chang et al., 2002; Meyerhoff et al., 1999). We observed overall significant reduction in concentration of Cr with time in all the three regions of brain, however, the reduction was sharp in gray matter followed by basal ganglia. In frontal white matter region, we observed a decreasing trend in the concentration of Cr with time among individuals with nadir CD4 count lower than 36 (see Figure 4.1). Cr has been found positively correlated with IQ in the frontal white matter (Yeo et al., 2000). This suggests that individuals with smaller CD4 count are at greater risk of developing impairment reflected by smaller IQ. It has been shown in multiple studies that HAART helps to improve CD4 counts in HIV patients. Therefore, HAART actually helps to improve brain performance via elevating the concentration of Cr.

In basal ganglia, duration of ARV treatment seems to influence the progression of Cr over time (see Figure 4.4). Longer duration of ARV treatment was associated with increased rate of decrease in concentration of Cr. Decrease in creatine level in patients with longer duration of ARV treatment may be caused by the progression of the disease or by the direct action of ARV treatment in basal ganglia region. Elevation of Cr does not necessarily indicate always better neuronal activity. Higher concentration of Cr is not good especially when increase in metabolism is due to active gliosis or inflammation (not due to neuronal cell) (Ratai et al., 2011). This is likely case when entry of HIV in brain leads to neuronal damage and glial cell activation and proliferation. In this case high concentration of Cr would be reflective of high metabolic demand of glial cells (not of the neuronal cells). Under such scenario, our finding about the observed association of duration of ARV treatment with progression of Cr in basal ganglia suggests that longer duration of ARV treatment is actually helpful in reversing the neuronal damage.

In gray matter, the concentration of Cr was reduced with increase in age, but the rate of decrease slowed down with time. Probably this is true for any population that rate of metabolism decreases with age, so the concentration of Cr.

N-acetylaspartate (NAA). NAA is a marker of neuronal integrity and viability. It decreases when the neuronal integrity is adversely affected. Laubenberger et al. (1996) and Chang et al. (2002) reported a decreased concentration of NAA in white matter region and increased concentration in basal ganglia among HIV patients in comparison to normal population. We observed that the longer duration of ARV treatment not only reduced the baseline concentration of NAA, it also increased the rate of the decrease in frontal white matter and gray matter region (see Figures 4.2 and 4.7). There are two possible explanations for this. First, ARV treatment may be neurotoxic as suspected by others (e.g., see Maschke et al., 2000). Secondly, patients with longer duration of ARV treatments were also those who had longer duration of HIV (i.e., advanced stages of HIV infection). The longer duration of HIV might result in accelerated decay in NAA in frontal white matter and gray matter and ARV treatment was not effective in reversing or decelerating this decay. The decrease in concentration of NAA in basal ganglia region over time was very minimal and we did not find any factor influencing this change. This is good thing because basal ganglia control several functions including voluntary motor control, learning ability, eye movements, cognitive, and emotional functions.

Choline (Cho). Choline is marker of inflammation. It is measure of increased cellular turnover and is elevated in tumors and inflammatory processes. It has been found in previous studies that the concentrations of choline were elevated in all three brain regions among HIV patients (Chang et al., 2002). We observed decreasing trend in the concentration of choline in white matter region, but the change was not significant when duration of ARV treatment at baseline exceeded 3.8 year (see Figure 4.3). This suggests that ARV treatment reduced choline only upto a certain level and after that the concentration seems to be stabilized. Also, in gray matter, we observed decreasing trend in concentration of choline with longer period of ARV treatment (see Figure 4.8). However, those who did not receive ARV treatment for long time, but received HAART therapy for more than 2.64 year, also experienced sharp decrease in concentration of choline. This suggests that both long-term ARV treatment and HAART are effective in reducing concentration of Cho in gray matter region which might be reflective of reduced inflammation.

Myo-inositol (MI). MI is considered a glial marker and is seen in higher concentration during development of several diseases (e.g., pre-dementia phase of Alzheimer's disease). Concentration of MI has been found to be elevated in all three brain regions among HIV patients (Chang et al., 2002). We observed decreasing trend in concentration of MI in the patients with longer period of ARV treatment in basal ganglia and higher CD4 count in gray matter (see Figures 4.5 and 4.9). Therefore, both the ARV treatment and grerater CD4 count help to control the concentration of MI in brain. It is well known that longer duration of HAART leads to increase in CD4 count. These findings are very much consistent with previous finding that treatment of HIV patients with HAART results in decrease in concentration of MI in brain (Chang et al., 1999). This again suggests that longer duration of both ARV treatment and HAART help to reduce the brain inflammation.

Glutamine-Glutamate (Glx). Glial cells and neurons are believed to be the primary sources of Glx. Brain uses Glx for chemical communication between cells. However, when Glx is present in excessive amount, the extra Glx is taken up by surrounding glial cells and transported to the neuron where it acts as potential excitotoxin. Excitotoxins cause particular brain cells to become excessively excited, to the point they will quickly die. Excitotoxins can also cause a loss of brain synapses and connecting fibers. Increase in glutamate resulting from glial cell dysfunction may lead to neuronal injury (Kaul et al., 2001; Kaul and Lipton, 1999). Chang et al. (2002) has reported elevation in concentration of Glx in basal ganglia region, but decrease in gray and white matter regions among HIV patients. We observed overall decreasing trend in concentration of Glx in all three brain regions. The baseline concentration of Glx in gray matter was reduced with increase in duration of ARV treatments, but the rate of decrease was relatively slow among those who received ARV treatment for more than 3.7 years (see Figure 4.10). This suggests that long term ARV treatment may be effective in reducing the concentration of Glx, but the rate get decelerated with longer duration of ARV treatment.

In summary, we observed that longer periods of ARV treatment and HAART were associated with reduced concentration of Cho, MI and Glx. Therefore, our findings suggest that individuals with lower duration of ARV treatment or HAART are at higher risk of developing cognitive dysfunction and neuronal injury. However, since longer duration of ARV is also likely to increase the rate of the decrease in the concentration of NAA, ARV treatment may be neurotoxic or at the minimum cannot stop neuronal injury among the patients with relatively higher duration of HIV. We also observed the influence of nadir CD4 count and duration of ARV treatment on the longitudinal progression of concentration of creatine in white matter and basal ganglia, respectively. This finding need to be explained depending on the role of Cr in each of the brain regions. Further knowledge on the development of disorders due to imbalance in brain metabolites in conjunction with our study findings would be helpful to identify the subgroup of individuals at higher risk.

The present study is a first attempt to identify the factors influencing the progression of cerebral metabolites in a data-driven manner. Along with the strength of this study, the work has important limitations. It is not possible to discern how stable the splits (i.e. thresholds) and branches (the location of baseline factors). This problem is only partially ameliorated by comparing Akaike information criterion (AIC) of the tree based model. Thus, the observations described in these analyses, must be replicated by others in order to ensure that the results are not data specific. We have used only baseline factors available to us in regression tree construction.

While these limitations are significant, the primary focus of the study was to identify the factors influencing the longitudinal progression of metabolite. Our findings are, by and large, interpretable and most of the time is in c with the findings reported by other researchers. In conclusion, we are able to identify the subpopulation of individuals, characterized by clinical and demographic, factors who are at greater risk of developing disorders due to imbalances in brain metabolites early. The practical implication of these findings are to identify the high risk group much prior to the development of the brain disorders. Several reports suggested to use MRS to monitor the effect of HIV treatments (Chang et al., 1999; Stankoff et al., 2001; Wilkinson et al., 1997). Prior and accurate knowledge about the progression of brain metabolites in the subgroups of HIV patients would minimize the need of repeated MRS in HIV patients and thus a more accurate and affordable therapy could be made available to HIV patients. However, there is need for further studies to confirm our results.

BIBLIOGRAPHY

Abdolell, M., M. LeBlanc, D. Stephens, and R. Harrison (2002). Binary partitioning for continuous longitudinal data: categorizing a prognostic variable. *Statistics in medicine 21*(22), 3395–3409.

Andrews, D. W. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrica: Journal of the Econometric Society*, 821–856.

Antinori, A., G. Arendt, J. Becker, B. Brew, D. Byrd, M. Cherner, D. Clifford, P. Cinque, L. Epstein, K. Goodkin, et al. (2007). Updated research nosology for hiv-associated neurocognitive disorders. *Neurology 69*(18), 1789–1799.

Bertholdo, D., A. Watcharakorn, and M. Castillo (2013). Brain proton magnetic resonance spectroscopy: Introduction and overview. *Neuroimaging Clinics of North America*.

Billingsley, P. (2009). *Convergence of probability measures*, Volume 493. Wiley-Interscience.

Birnbaum, Z. (1952). Numerical tabulation of the distribution of kolmogorov's statistic for finite sample size. *Journal of the American Statistical Association 47*(259), 425–441.

Bjorck, A. (1996). *Numerical methods for least square problems.* SIAM, Philadelphia.

Breiman, L., J. Friedman, C. Stone, and R. Olshen (1984). *Classification and regression trees.* Chapman & Hall/CRC.

Brown, R., J. Durbin, and J. Evans (1975). Techniques for testing the constancy of regression relationships over time. *Journal of the Royal Statistical Society. Series B (Methodological)*, 149–192.

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

Cai, T. and P. Hall (2006). Prediction in functional linear regression. *The Annals of Statistics 34*(5), 2159–2179.

Cardot, H., C. Crambes, A. Kneip, and P. Sarda (2007). Smoothing splines estimators in functional linear regression with errors-in-variables. *Computational Statistics Data Analysis 51*(10), 4832 – 4848.

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

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

Carey, C., S. Woods, R. Gonzalez, E. Conover, T. Marcotte, I. Grant, and R. Heaton (2004). Predictive validity of global deficit scores in detecting neuropsychological impairment in hiv infection. *Journal of Clinical and Experimental Neuropsychology 26*(3), 307–319.

Chang, L., T. Ernst, M. Leonido-Yee, M. Witt, O. Speck, I. Walot, and E. Miller (1999). Highly active antiretroviral therapy reverses brain metabolite abnormalities in mild hiv dementia. *Neurology 53*(4), 782–782.

Chang, L., T. Ernst, M. D. Witt, N. Ames, M. Gaiefsky, and E. Miller (2002). Relationships among brain metabolites, cognitive function, and viral loads in antiretroviral-naıve hiv patients. *Neuroimage 17*(3), 1638–1648.

Christiansen, P., P. Toft, H. Larsson, M. Stubgaard, and O. Henriksen (1993). The concentration of¡ i¿ n¡/i¿-acetyl aspartate, creatine+ phosphocreatine, and choline in different parts of the brain in adulthood and senium. *Magnetic resonance imaging 11*(6), 799–806.

Ciprian Crainiceanu, Philip Reiss, Jeff Goldsmith, Lei Huang, Lan Huo, and Fabian Scheipl (2012). *refund: Regression with Functional Data.* R package version 0.1-6.

Clark, L. and D. Pregibon (1992). Tree-based models. chambers jm and hastie tj, editors. statistical models in s.

Cysique, L. A., P. Maruff, and B. J. Brew (2004). Prevalence and pattern of neuropsychological impairment in human immunodeficiency virus-infected/acquired immunodeficiency syndrome (hiv/aids) patients across pre-and post-highly active antiretroviral therapy eras: A combined study of two cohorts clinical report. *Journal of neurovirology 10*(6), 350–357.

Dager, S. R., N. Oskin, T. L. Richards, and S. Posse (2008). Research applications of magnetic resonance spectroscopy (mrs) to investigate psychiatric disorders. *Topics in magnetic resonance imaging: TMRI 19*(2), 81.

Demidenko, E. (2004). *Mixed models: theory and applications*, Volume 518. Wiley-Interscience.

Di, C., C. Crainiceanu, B. Caffo, and N. Punjabi (2009). Multilevel functional principal component analysis. *Annals of Applied Statistics 4*, 458–288.

Diggle, P., P. Heagerty, K. Liang, and S. Zeger (2002). *Analysis of longitudinal data*, Volume 25. Oxford University Press, USA.

Dore, G. J., P. K. Correll, Y. Li, J. M. Kaldor, D. A. Cooper, and B. J. Brew (1999). Changes to aids dementia complex in the era of highly active antiretroviral therapy. *Aids 13*(10), 1249–1253.

Fan, J. and J. Zhang (2000). Two-step estimation of functional linear models with applications to longitudinal data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 62*(2), 303–322.

Faraway, J. (1997). Regression analysis for a functional response. *Technometrics 39*(3), 254–261.

Fitzmaurice, G. Laird, M. and J. Ware (2004). *Applied Longitudinal Analysis.* Wiley series in probability and statistics.

Galimberti, G. and A. Montanari (2002). Regression trees for longitudinal data with time-dependent covariates. *Classification, clustering and data analysis*, 391–398.

Gertheiss, J., J. Goldsmith, C. Crainiceanu, and S. Greven (2013). Longitudinal scalar-on-functions regression with application to tractography data. *Biostatistics 14*(3), 447–461.

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

Goldsmith, J., C. Crainiceanu, B. Caffo, and D. Reich (2012). Longitudinal penalized functional regression for cognitive outcomes on neuronal tract measurements. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 61*(3), 453–469.

Golub, G. and C. Van-Loan (1996). *Matrix computations.* John Hopkins University Press, Baltimore.

Greven, S., C. Crainiceanu, B. Caffo, and D. Reich (2011). Longitudinal functional principal component analysis. *Recent Advances in Functional Data Analysis and Related Topics*, 149–154.

Guo, W. (2002). Functional mixed effects models. *Biometrics 58*(1), 121–128.

Hall, P., D. Poskitt, and B. Presnell (2001). A functional data-analytic approach to signal discrimination. *Technometrics 43*(1), 1–9.

Harezlak, J., S. Buchthal, M. Taylor, G. Schifitto, J. Zhong, E. Daar, J. Alger, E. Singer, T. Campbell, C. Yiannoutsos, et al. (2011). Persistence of hiv-associated cognitive impairment, inflammation, and neuronal injury in era of highly active antiretroviral treatment. *Aids 25*(5), 625.

Hasler, G., J. W. van der Veen, C. Grillon, W. C. Drevets, and J. Shen (2010). Effect of acute psychological stress on prefrontal gaba concentration determined by proton magnetic resonance spectroscopy. *The American journal of psychiatry 167*(10), 1226.

Hastie, T., R. Tibshirani, and J. J. H. Friedman (2001). *The elements of statistical learning*, Volume 1. Springer New York.

Henderson, C. (1950). Estimation of genetic parameters (abstract). *Annals of Mathematical Statistic 21*(1), 309–310.

Hesselink, J. R. (2013). Fundamentals of mr spectroscopy. *Online, accessed on*, 14–05.

Hjort, N. and A. Koning (2002). Tests for constancy of model parameters over time. *Journal of Nonparametric Statistics 14*(1-2), 113–132.

Hochberg, Y. (1988). A sharper bonferroni procedure for multiple tests of significance. *Biometrika 75*(4), 800–802.

Hochberg, Y. and A. C. Tamhane (1987). *Multiple comparison procedures.* John Wiley & Sons, Inc.

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified bonferroni test. *Biometrika 75*(2), 383–386.

Horská, A. et al. (2013). ^{1}h magnetic resonance spectroscopy of the brain during adolescence: Normal brain development and neuropsychiatric disorders. In *MR Spectroscopy of Pediatric Brain Disorders*, pp. 193–212. Springer.

Ialongo, N. S., L. Werthamer, S. G. Kellam, C. H. Brown, S. Wang, and Y. Lin (1999). Proximal impact of two first-grade preventive interventions on the early risk behaviors for later substance abuse, depression, and antisocial behavior. *American journal of community psychology 27*(5), 599–641.

James, G. (2002). Generalized linear models with functional predictors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64*(3), 411–432.

Kaul, M., G. A. Garden, and S. A. Lipton (2001). Pathways to neuronal injury and apoptosis in hiv-associated dementia. *Nature 410*(6831), 988–994.

Kaul, M. and S. A. Lipton (1999). Chemokines and activated macrophages in hiv gp120-induced neuronal apoptosis. *Proceedings of the National Academy of Sciences 96*(14), 8212–8216.

Kellam, S. G., G. W. Rebok, N. Ialongo, and L. S. Mayer (1994). The course and malleability of aggressive behavior from early first grade into middle school: Results of a developmental epidemiologically-based preventive trial. *Journal of Child Psychology and Psychiatry 35*(2), 259–281.

Lagopoulos, J. (2007). Spectroscopy. *Acta Neuropsychiatrica 19*(6), 382–383.

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

Laubenberger, J., D. Häussinger, S. Bayer, S. Thielemann, B. Schneider, A. Mundinger, J. Hennig, and M. Langer (1996). Hiv-related metabolic abnormalities in the brain:

depiction with proton mr spectroscopy with short echo times. *Radiology 199*(3), 805–810.

Leuchter, A. F., I. A. Cook, E. A. Witte, M. Morgan, and M. Abrams (2002). Changes in brain function of depressed subjects during treatment with placebo. *American Journal of Psychiatry 159*(1), 122–129.

Liang, K. and S. Zeger (1986). Longitudinal data analysis using generalized linear models. *Biometrika 73*(1), 13–22.

Lilliefors, H. (1967). On the kolmogorov-smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association 62*(318), 399–402.

Lim, K. and D. Spielman (2005). Estimating naa in cortical gray matter with applications for measuring changes due to aging. *Magnetic resonance in medicine 37*(3), 372–377.

López-Villegas, D., R. E. Lenkinski, and I. Frank (1997). Biochemical changes in the frontal lobe of hiv-infected individuals detected by magnetic resonance spectroscopy. *Proceedings of the National Academy of Sciences 94*(18), 9854–9859.

Loughran, T. and D. S. Nagin (2006). Finite sample effects in group-based trajectory models. *Sociological methods & research 35*(2), 250–278.

Malhi, G. and J. Lagopoulos (2008). Making sense of neuroimaging in psychiatry. *Acta Psychiatrica Scandinavica 117*(2), 100–117.

Maschke, M., O. Kastrup, S. Esser, B. Ross, U. Hengge, and A. Hufnagel (2000). Incidence and prevalence of neurological disorders associated with hiv since the introduction of highly active antiretroviral therapy (haart). *Journal of Neurology, Neurosurgery & Psychiatry 69*(3), 376–380.

Massey Jr, F. (1951). The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association 46*(253), 68–78.

Meyerhoff, D., C. Bloomer, V. Cardenas, D. Norman, M. Weiner, and G. Fein (1999). Elevated subcortical choline metabolites in cognitively and clinically asymptomatic hiv patients. *Neurology 52*(5), 995–995.

Morris, J. and R. Carroll (2006). Wavelet-based functional mixed models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68*(2), 179–199.

Müller, H. (2005). Functional modelling and classification of longitudinal data. *Scandinavian Journal of Scandinavian Journal of 32.*

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

Muthén, B. (2004). Latent variable analysis. *The Sage handbook of quantitative methodology for the social sciences. Thousand Oaks, CA: Sage Publications*, 345–68.

Muthén, B. and K. Shedden (1999). Finite mixture modeling with mixture outcomes using the em algorithm. *Biometrics 55*(2), 463–469.

Nyblom, J. (1989). Testing for the constancy of parameters over time. *Journal of the American Statistical Association 84*(405), 223–230.

O'Neill, J., J. G. Levitt, and J. R. Alger (2013). Magnetic resonance spectroscopy studies of attention deficit hyperactivity disorder. In *MR Spectroscopy of Pediatric Brain Disorders*, pp. 229–275. Springer.

Paige, C. and M. Saunders (1981). Towards a generalized singular value decomposition. *SIAM Journal on Numerical Analysis 18*(3), 398–405.

Palella Jr, F. J., K. M. Delaney, A. C. Moorman, M. O. Loveless, J. Fuhrer, G. A. Satten, D. J. Aschman, and S. D. Holmberg (1998). Declining morbidity and mortality among patients with advanced human immunodeficiency virus infection. *New England Journal of Medicine 338*(13), 853–860.

Provencher, S. (2005). Estimation of metabolite concentrations from localized in vivo proton nmr spectra. *Magnetic Resonance in Medicine 30*(6), 672–679.

Ramsay, J. and C. Dalzell (1991). Some tools for functional data analysis. *Journal of the Royal Statistical Society. Series B (Methodological) 53*(3), 539–572.

Ramsay, J. and B. Silverman (1997). *Functional Data Analysis*. Springer-Verlag, Berlin.

Randolph, T., J. Harezlak, and Z. Feng (2012). Structured penalties for functional linear models—partially empirical eigenvectors for regression. *Electronic Journal of Statistic 6*, 323–353.

Ratai, E.-M., L. Annamalai, T. Burdo, C.-G. Joo, J. P. Bombardier, R. Fell, R. Hakimelahi, J. He, M. R. Lentz, J. Campbell, et al. (2011). Brain creatine elevation and n-acetylaspartate reduction indicates neuronal dysfunction in the setting of enhanced glial energy metabolism in a macaque model of neuroaids. *Magnetic Resonance in Medicine 66*(3), 625–634.

Raudenbush, S. (2001). Comparing personal trajectories and drawing causal inferences from longitudinal data. *Annual review of psychology 52*(1), 501–525.

Reiss, P. and R. Ogden (2009). Smoothing parameter selection for a class of semiparametric linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71*(2), 505–523.

Robertson, K. R., W. T. Robertson, S. Ford, D. Watson, S. Fiscus, A. G. Harp, and C. D. Hall (2004). Highly active antiretroviral therapy improves neurocognitive functioning. *JAIDS Journal of Acquired Immune Deficiency Syndromes 36*(1), 562–566.

Robertson, K. R., M. Smurzynski, T. D. Parsons, K. Wu, R. J. Bosch, J. Wu, J. C. McArthur, A. C. Collier, S. R. Evans, and R. J. Ellis (2007). The prevalence and incidence of neurocognitive impairment in the haart era. *Aids 21*(14), 1915–1921.

Robinson, G. (1991). That blup is a good thing: the estimation of random effects. *Statistical Science 6*(1), 15–32.

Ross, A. J. and P. S. Sachdev (2004). Magnetic resonance spectroscopy in cognitive research. *Brain Research Reviews 44*(2), 83–102.

Ruppert, D., M. Wand, and R. Carroll (2003). *Semiparametric Regression.* Cambridge Series in Statistical and Probabilistic Mathematics.

Segal, M. (1992). Tree-structured methods for longitudinal data. *Journal of the American Statistical Association 87*(418), 407–418.

Sela, R. J. and J. S. Simonoff (2012). Re-em trees: a data mining approach for longitudinal and clustered data. *Machine learning 86*(2), 169–207.

Soares, D., M. Law, et al. (2009). Magnetic resonance spectroscopy of the brain: review of metabolites and clinical applications. *Clinical radiology 64*(1), 12.

Stankoff, B., A. Tourbah, S. Suarez, E. Turell, J. Stievenart, C. Payan, A. Coutellier, S. Herson, L. Baril, F. Bricaire, et al. (2001). Clinical and spectroscopic improvement in hiv-associated cognitive impairment. *Neurology 56*(1), 112–115.

Tozzi, V., P. Balestra, R. Bellagamba, A. Corpolongo, M. F. Salvatori, U. Visco-Comandini, C. Vlassi, M. Giulianelli, S. Galgani, A. Antinori, et al. (2007). Persistence of neuropsychologic deficits despite long-term highly active antiretroviral therapy in patients with hiv-related neurocognitive impairment: prevalence and risk factors. *JAIDS Journal of Acquired Immune Deficiency Syndromes 45*(2), 174–182.

Tracey, I. D., C. Carr, A. Guimaraes, J. Worth, B. Navia, and R. Gonzalez (1996). Brain choline-containing compounds are elevated in hiv-positive patients before the onset of aids dementia complex a proton magnetic resonance spectroscopic study. *Neurology 46*(3), 783–788.

Valcour, V. G., N. C. Sacktor, R. H. Paul, M. R. Watters, O. A. Selnes, B. T. Shiramizu, A. E. Williams, and C. M. Shikuma (2006). Insulin resistance is associated with cognition among hiv-1-infected patients: the hawaii aging with hiv cohort. *JAIDS Journal of Acquired Immune Deficiency Syndromes 43*(4), 405–410.

Van-Loan, C. (1976). Generalizing the singular value decomposition. *SIAM Journal on Numerical Analysis 13*(1), 76–83.

Verbyla, D. (1987). Classification trees: a new discrimination tool. *Canadian Journal of Forest Research 17*(9), 1150–1152.

Wang, P. W., N. Sailasuta, R. A. Chandler, and T. A. Ketter (2006). Magnetic resonance spectroscopic measurement of cerebral gamma-aminobutyric acid concentrations in patients with bipolar disorders. *Acta Neuropsychiatrica 18*(2), 120–126.

Wilkinson, I. D., S. Lunn, K. A. Miszkiel, R. F. Miller, M. N. Paley, I. Williams, R. J. Chinn, M. A. Hall-Craggs, S. P. Newman, B. E. Kendall, et al. (1997). Proton mrs and quantitative mri assessment of the short term neurological response to antiretroviral therapy in aids. *Journal of Neurology, Neurosurgery & Psychiatry 63*(4), 477–482.

Yao, F. and H. Müller (2010). Functional quadratic regression. *Biometrika 97*(1), 49–64.

Yeo, R. A., D. Hill, R. Campbell, J. Vigil, and W. M. Brooks (2000). Developmental instability and working memory ability in children: a magnetic resonance spectroscopy investigation. *Developmental Neuropsychology 17*(2), 143–159.

Zhang, H. (1997). Multivariate adaptive splines for analysis of longitudinal data. *Journal of Computational and Graphical Statistics 6*(1), 74–91.

Zhang, H. and B. Singer (1999). *Recursive partitioning in the health sciences.* Springer Verlag.

CURRICULUM VITAE

Madan Gopal Kundu

EDUCATION

- Ph.D. in Biostatistics, Indiana University, Indianapolis, IN, 2014.
- M.Sc. in Agricultural Statistics, Indian Agricultural Statistics Research Institute (IASRI), New Delhi, India, 2005.
- B.Sc. in Agriculture, Uttar Banga Krishi Viswavidyalaya (UBKV), Coochbehar, India, 2003.

WORKING EXPERIENCE

- Aug 2013 - Jan 2014: Research Assistant, Indiana University, Indianapolis, IN
- May 2013 - Aug 2013: Summer Intern, Boehringer Ingelheim, Ridgefield, CT
- Aug 2012 - May 2013: Teachng Assistant, IUPUI, Indianapolis, IN
- May 2012 - Aug 2012: Summer Intern, Cytel Inc., Cambridge, MA
- Jul 2010 - May 2012: Research Assistant, Indiana University, Indianapolis, IN
- Apr 2008 - Aug 2009: Research Scientist (Biostatistician), Ranbaxy Labs. Ltd, Gurgaon, India
- April 2007 - Apr 2008: Biostatistician I, i3 Statprobe, Gurgaon, India
- Mar 2006 - Apr 2007: Research Associate (Biostatistician), Ranbaxy Labs. Ltd, Gurgaon, India

SELECTED PUBLICATIONS

- Kundu M G, Harezlak J and Randolph T W (in revision), Longitudinal functional regression models with structured penalties, `arXiv:1211.4763 [stat.AP]`

- Kundu M G and Harezlak J (submitted), Regression tree with longitudinal data, `arXiv:1309.7733 [stat.ME]`