ABSTRACT

MCKENDRY, COLLEEN MARIE. Association Study, Risk Assessment, and Prediction of Children's Growth Trajectories through Methylation Profiles using Functional Mixed Models. (Under the direction of Arnab Maity and Jung-Ying Tzeng.)

This dissertation is motivated by the Newborn Epigenetic STudy (NEST) data. The overall goal is to investigate the relationship between growth trajectories and gene methylation profiles in children. Each child in the NEST data set has up to thirty weight measurements (in kilograms), recorded on days irregularly spaced over a five year time frame. The methylation profiles contain numerous vectors of related, highly correlated, covariates. Although there have been many developments in the functional data literature, standard methods are not suitable to adequately model this type of data, as explained in Chapter 2.

To model this type of data, a functional semiparametric regression modeling framework is introduced, where the response is a function (children growth trajectories measured over time) and the covariates are both vector and scalar (gene methylation profiles and other confounders). The model framework combines standard functional methods, such as functional principal components analysis, with Gaussian process regression and is flexible enough to handle sparse, irregularly spaced data. Using this modeling framework, we first consider the problem of determining if there is an association between the growth trajectories and gene methylation profiles, while accounting for other confounders. A hypothesis testing procedure is developed to test the joint effect of a specific vector of methylation values on the functional response. This is done by modeling the effect of the vector as a random effect and utilizing a linear score test for variance components.

The initial model is then reparameterized and fit as a penalized regression model. By using a different model fitting procedure and making some alterations to the testing procedure, we have a comparable method in terms of Type I error, power, and hypothesis testing results but with significantly reduced computation time. The penalized regression model is expanded to develop a prediction mode. This enables us to determine how the functional response changes based on

the vector of related covariates of interest and to predict the full growth trajectory for each individual. Changes in the functional response due to the covariates of interest are captured through what is referred to as individual risk profiles. Using the assumed distribution of the functional random effect and the prediction model, the risk profile and full growth trajectory for a new subject can be predicted using their methylation profiles and other baseline covariate information. The use of prediction variances is introduced to construct prediction bands around the curves, which improves the coverage of the intervals as compared to methods found in current functional data literature.

Association Study, Risk Assessment, and Prediction of Children's Growth Trajectories through Methylation Profiles using Functional Mixed Models

by
Colleen Marie McKendry

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Statistics

Raleigh, North Carolina

2018

APPROVED BY:

Cathrine Hoyo	Luo Xiao
Arnab Maity Co-chair of Advisory Committee	Jung-Ying Tzeng Co-chair of Advisory Committee

DEDICATION

To my amazing family and friends for all of their support throughout this crazy thing called graduate school.

BIOGRAPHY

Colleen Marie McKendry was born in November 1989 to William and Darlene McKendry. She grew up in West Deptford, New Jersey and graduated from West Deptford High School in 2008. After high school, Colleen attended Monmouth University, a small liberal arts college in central NJ. Colleen discovered statistics the summer between her sophomore and junior year when she participated in a statistical consulting program with Dr. Richard Bastian. She continued consulting throughout her time at Monmouth. Although ultimately obtaining a dual degree in Mathematics and Secondary Education in 2012, Colleen decided that teaching was not for her and enrolled in the statistics PhD program at North Carolina State University.

While at NCSU, Colleen took advantage of a variety of funding opportunities that helped to shape her as a statistician. She was on an NIH grant for two years that gave her the opportunity to work with a statistical group at Duke Clinical Research institute. She was also a statistical consultant for the NCSU College of Veterinary medicine for a year. In the summer of 2016, she joined the JMP Division of SAS as a statistical writing intern and held that position for the remainder of her time as a student. Under the direction of Dr. Arnab Maity and Dr. Jung-Ying Tzeng, Colleen is scheduled to graduate in May 2018. After graduation she will become a full-time employee at JMP and will continue working on the documentation team.

Outside of statistics, Colleen likes to stay active. She enjoys working out and going on hiking adventures, especially in western NC and the areas surrounding Asheville. She is also very interested in food and nutrition and loves to cook at home as much as possible.

ACKNOWLEDGEMENTS

I would like to thank my co-advisors, Arnab Maity and Jung-Ying Tzeng, for their guidance, patience, and support; but mostly, I would like to thank them for seeing me as a person and always treating me as an equal. Prior to walking into his office three years ago to discuss research opportunities, Dr. Maity and I had never actually met. I was a little lost at the time and honestly had no idea what I was looking for in a research advisor. I had already met with many professors and none of it felt right. When I first met with Dr. Maity, he was one of the only professors to ask me questions about myself and show an interest in me as a person. He also took the time to explain his research, sitting next to me and writing stuff out as he went. These two characteristics have remained constant and I have been grateful for the opportunity to work with him ever since. The addition of Jung-Ying as my co-advisor only added to my gratefulness. Our meetings were a great combination of lightness and productivity, and I always felt comfortable asking questions and brainstorming.

There are numerous other professors and staff at NCSU that have helped me on my journey through graduate school. I would like to thank my other committee members, Dr. Luo Xiao and Dr. Cathrine Hoyo, for their time and input. I would like to thank Dr. Marie Davidian for giving me a position on her grant that enabled me to work with the statistical team at DCRI. I would also like to thank Dr. Emily Griffith for the opportunity to work under her as the statistical consultant for the veterinary school. Not only did she provide support and advice on consulting projects, but also served as a mentor and friend. A very special thank you goes to Reneé Moore, who is sadly no longer a professor at NCSU, but was a true mentor for me in my early years in the program. When I did not pass my qualification exam the first time around and had to retake it, she encouraged me and helped to get my head in the right place. It was through her that I was able to believe in myself again. When I was searching for an advisor, she gave me advice and talked with me after meetings to help me determine the best fit. When things got tough, which let's be honest, was a lot of the time, she was always there.

Without her support, I am not sure that I would have stayed in the program and for that, I am forever grateful. Another professor that served as a mentor was Dr. Alyson Wilson. She was also vital in helping with my search for an advisor and I have always appreciated her honesty, realness, and perspective. A big thanks to Chris Waddell for taking the time to look at my code that suddenly stopped working and helping to debug an `R` package. His computing help truly saved my first project and started a snowball effect of things falling into place to complete my dissertation. Finally, I would like to thank my graduate school mom, Alison McCoy, for all of the post-meeting chats, the hugs (both comforting and celebratory), the encouragement, the love, and for convincing (ie, forcing) me to get computing help from Chris!

Dr. Richard Bastian and Dr. Bonnie Gold, professors from my undergraduate years at Monmouth, also deserve recognition and thanks. Dr. Gold was the first professor to "teach" me how to teach myself and to stress how important that skill would be in graduate school. Turns out she was right! If there is a single person that deserves credit for me even attending graduate school, it is Dr. Bastian. He saw my potential early on at Monmouth and got my statistical career started by asking me to join his statistical consulting group. Since then, he has been a constant source of encouragement and support - from student teaching struggles to graduate school applications to qualifying exams and even to my dissertation defense! Thank you, Dr. B., for understanding my aspirations to succeed academically and the need to have balance in my life. I feel lucky to have had you as a mentor and even luckier to now call you a friend - but you'll still always be Dr. B. to me!

I would also like to thank my co-workers at JMP, particularly my managers and immediate team. Although it took me about six months longer than expected to finish my degree, they were very understanding, patient, and supportive of my graduate school experience. In particular, I was very appreciative of the time I was able to take off around important milestones, such as my oral preliminary exam and most recently, to prepare for my dissertation defense.

Last, but farthest from least, I would like to thank my family and friends for the crazy amount of support I have recieved. I am thankful for the people I have met in NC, from the

friends I made in the statistics program to the friends I made at the gym; it's always nice to be surrounded by good people. Thank you to my friends back home (actually, all over the country at this point), particularly Stacie Leporati, Allison Douglass, and Derrick Alcott. Your friendship and support is so important to me! And now to my other half and partner-in-PhD, Brian. Thank you, Brian, for supporting me and always being able to make me laugh. Thank you for talking things out with me on Friday nights or whatever other random times I had research ideas. At times it seemed like maybe it wouldn't happen, but we made it!! Finally, thank you to my mom and dad for more things than I could possibly write. You have always been my biggest cheerleaders, which at times, probably hasn't been the easiest. Thank you for being on this rollercoaster of a journey with me, celebrating the ups and offering words of encouragements during the downs. Thank you for being understanding whenever I was home, but had to work...for driving down to NC for a 24 hour trip when I passed my qualifying exam just so we could celebrate together...for taking the time to understand the process of my program...and for happily sitting through an hour plus presentation that you knew nothing about...thank you, thank you, thank you. I appreciate and love you both so much!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

1.1 Motivation

This dissertation is motivated by the data set from the Newborn Epigenetics STudy (NEST). Participants in NEST were selected between 2009 and 2011 from pregnant women who visited one of six prenatal clinics in Durham County, NC. Prior to giving birth, a variety of demographic and lifestyle information was collected from the mothers. Some of the demographic information included age, race, education history, and socioeconomic status. Information on each mother's lifestyle was mostly concentrated on lifestyle during pregnancy. This included whether or not the mother smoked during pregnancy, the type of nutrients consumed, and any exposure to toxic metals. In addition, some lifestyle information was collected after birth, such as whether or not the mother breastfed her baby.

At birth, cord blood that contained methylation profile information was collected and processed from each child. Developmental exposures are associated with alterations to the epigenome in early life, although determining how these alterations affect an individual has proved challenging. Genomically imprinted genes are a class of genes characterized by monoallelic expression, and use allele-specific differential CpG methylation by silencing one of the two parentally-derived alleles (Skaar et al., 2012; Woodfine et al., 2011; Ollikainen & Craig, 2011).

There are several known differentially methylated regions (DMRs), where multiple adjacent CpG sites show this type of parent-of-origin-specific methylation. Literature has suggested that certain gene methylation profile characteristics within a DMR can control the expression of genes to impact the weight of a child (Soubry et al., 2013, 2015; Wang et al., 2015b).

Within each DMR, there are a different number of CpG sites. In the NEST data set, there is information for nine different DMRs, with the number of CpG sites within a particular DMR ranging from four to ten. The individual methylation values at CpG sites within a particular DMR vary, but the set of methylation values are highly correlated and therefore interact with each other in a complex way. Figure 1.1 provides an example of the relationship among the methylation values in females within a particular DMR, PEG3. The methylation values are percentages, where a normal methylation value is considered to be 50%. However, this has been shown to vary within different DMRs. The boxplot shows that the means of the CpG sites fall between 31 and 41 and the standard deviations are between 3 and 4. All CpG sites appear to have a few outliers on the upper end, with the maximum values ranging from 50 to 55. The correlation matrix shows that all of the sites have strong positive correlations with each other, which is to be expected since they are from the same DMR.

Once a child was born, the subject was followed for up to five years in order to track the child's growth. Measurements for weight, in kilograms, were recorded at follow-up visits. Therefore, each child in the NEST data set has up to thirty weight measurements, recorded on days irregularly spaced over a five year time frame. Using these data points, one can get an approximate idea of the child's growth trajectory over time, as can be viewed in Figure 1.2.

1.1.1 Clinical Questions

The clinical questions resulting from this data set center around investigating the relationship between the methylation profiles and a child's growth trajectory. It is of interest to test whether there is an association between methylation profiles and a child's growth trajectory. More specifically, we wish to determine if there are certain DMRs that are associated with a child's

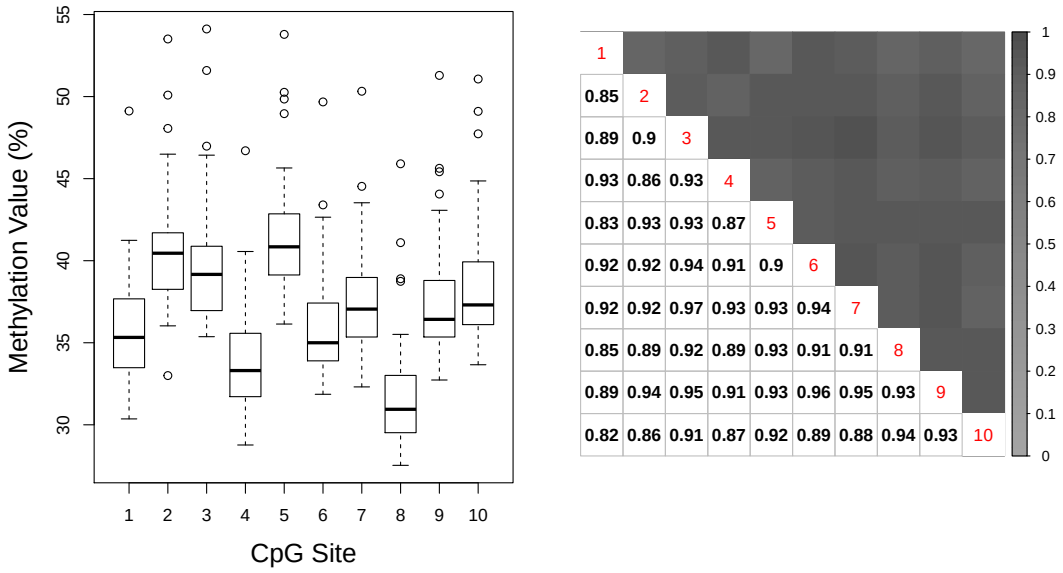


Figure 1.1: Example of the nature of the relationship among methylation values within a DMR. This is the female data from the PEG3 DMR.

growth trajectory. Assuming that there is an association, it is also of interest to determine how a methylation profile affects a child's growth curve. Additionally, can the methylation values within a certain DMR be used to predict a child's growth trajectory at birth, or to assess his/her risk for obesity? While the NEST data contains information for only nine DMRs, there are many more and it would also be useful to be able to screen a large number of DMRs for association.

Translating this into a statistical problem, we have a functional response with sparse, irregularly spaced observations. The covariates of interest are the correlated methylation values within a DMR. The goal is to first test if the vector of related covariates is significantly associ-

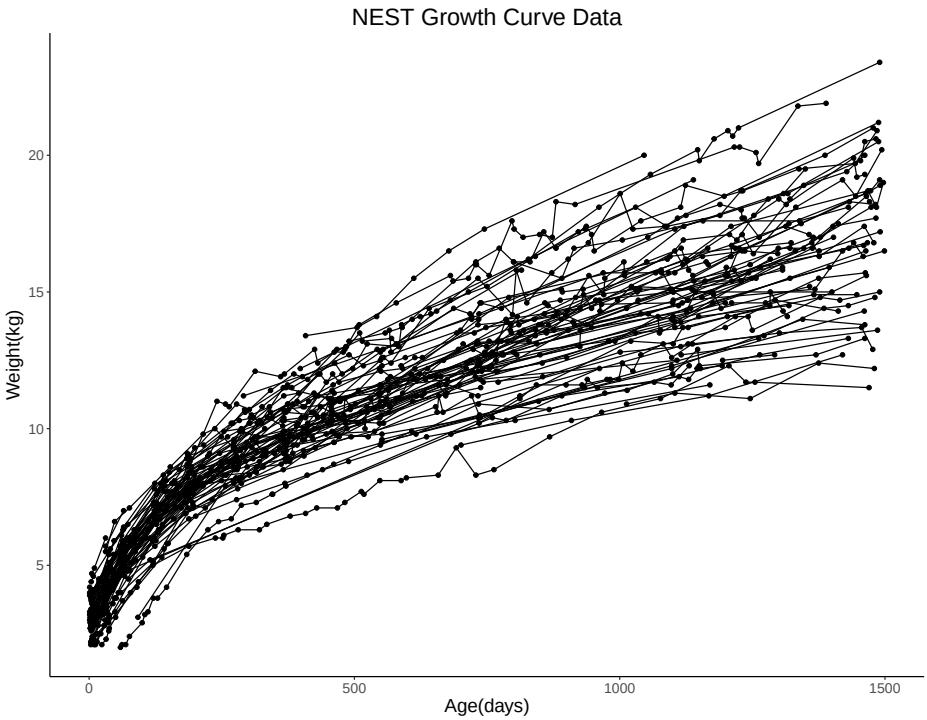


Figure 1.2: Growth curve data for NEST subjects.

ated with the growth trajectory. Then, we wish to determine exactly how the values contribute to the growth trajectory and use this information to create individual risk profiles. Finally, we wish to create a statistical model to predict a new subject's growth trajectory and assess his/her risk profile at birth.

1.2 Summary of Contributions

Although there have been developments in the functional data literature regarding testing and prediction, the methods available are not suitable for this type of problem. The aim of this dissertation is to present novel research and methods that can be utilized to answer the motivating questions. First, we develop a method to test the joint effect of a vector of related, and possibly correlated, covariates of interest on a functional response curve. The procedure utilizes a nonlinear model that combines standard functional methods, such as functional prin-

cipal component analysis, with Gaussian process regression. It is also flexible enough to handle sparse, irregularly spaced data, as is found in the NEST data set. Second, we reparameterize the initial model into a penalized regression model. This allows us to use the same type of testing procedure, but with significantly reduced computation time due to the different model fitting procedure. Third, we expand on the penalized regression model to determine how the functional response changes based on the vector of related covariates of interest. We quantify this by creating individual risk profiles. The model can be used to predict the individual risk over different stages of growth, as well as the full response trajectory. The subject-specific functional random effects included in the model help to better capture the proper covariance structure, and therefore lead to better predictions. Additionally, the assumed distribution of the random effects, along with the prediction model, can be used to predict the risk profile and response curve for a new subject using the methylation information collected at birth. We introduce the use of prediction variances to construct prediction bands around these curves to improve the coverage of the interval as compared to current standard methods.

Chapter 2

Functional Data Analysis Literature Review

2.1 Introduction

Data that are recorded over a continuous domain, with multiple measurements per subject, and present as some form of curve, are considered to be functional data. Applications of functional data are often measured over time, such as growth curves, hormone profiles, biomarkers, tracking data, and weather data (Guo, 2002; Ramsay, 2006; Morris, 2015; Faraway, 1997). Methods for how to handle this type of data have grown and developed into their own field, known as functional data analysis. Functional data analysis began with techniques to explore, smooth, and understand the covariance structure of the curves over time. For a complete review of these techniques, see Ramsay (2006). In many scenarios, it is also of interest to determine the relationships between functional data and other covariates.

2.2 Testing and Prediction in Functional Data Analysis

There has been a lot of work that focuses on relating functional covariates to a scalar response, also known as scalar-on-function regression. However, it can be argued that the more common

type of functional data problem in practice deals with a functional repsonse and scalar covariates (Ramsay, 2006). For categorical covariates, functional analysis of variance (FANOVA) has been studied by many, including Staniswalis & Lee (1998), Ramsay et al. (1996), Brumback & Rice (1998), and Ramsay (2006). Several of these methods utilize conventional design matrices and are simply extensions of ordinary ANOVA models. The key is to reinterpret the residual sum of squares in the appropriate way and minimize them subject to parameter constraints. There are several techniques to minimize the sum of squares, but QR decomposition is a common choice (Ramsay, 2006). The obvious limitation is that this model can test only categorical covariates.

2.2.1 Functional Linear Models

Extensions of multiple linear regression also exist for continuous covariates and this is known as functional response or function-on-scalar regression. Early work on functional response regression by Faraway (1997) involved a fairly straightforward, two-step procedure. The first step smooths the observed curves. Then, for each time point, t, point-wise regression is performed for the covariates of interest. Once the regression values are obtained for each time point, one can interpolate between the values to determine if there is any shape or function to the regression coefficients that can provide some interpretability. The prominent lacking feature of this approach is the fact that it ignores the correlation across time points, often referred to as the within-function correlation. Additionally, it requires that the observations be observed on the same grid of time points.

A more general linear functional response regression model can be written as

$$Y_i(t_j) = \sum_{a=1}^{p} X_{ia}\beta_a(t_j) + E_i(t_j), \tag{2.1}$$

for $a = 1, \ldots, p$ scalar predictors, $i = 1, \ldots, N$ functions, and t_j on some continuous domain. The partial effect of predictor X_a on the response Y at time t_j is represented through $\beta_a(t_j)$, which is the functional coefficient. In the model, $E_i(t_j)$ represents the curve-to-curve residual error.

The covariance structure of the errors accounts for the within-function correlation. It is possible to split the errors such that $E_i(t_j) = U_i(t_j) + \epsilon_{ij}$, where $U_i(t_j)$ is a curve-level random effect function and ϵ_{ij} are white noise residual errors, assumed to be independently and identically distributed random variables with mean 0 and variance σ^2. Splitting the residuals in this fashion enables one to calculate estimates that are denoised. Numerous methods are based off of this model, with the methods differing in how $\beta_a(t)$ is smoothed and what assumptions are made about the within-function covariance. According to a review paper by Wang et al. (2015a), a common theme among methods is to expand the responses and regression coefficients in the same functional basis (B-spline or eigenbasis) and treat as a traditional linear model. A specific method proposed by Wu & Chiang (2000) utilized kernel smoothing on the functional coefficients and assumed independent and identically distributed random errors. For a comprehensive review of functional regression methods, see review articles by Wang et al. (2015a) and Morris (2015). It should be noted that the model in (2.1) is similar to the varying coefficient model first introduced by Hastie & Tibshirani (1993). In fact, it is possible to reduce any functional linear model to some form of a varying coefficient model (Ramsay, 2006). One group that studied the model in the context of a varying coefficient model was Huang et al. (2002). Huang et al. (2002) estimated the regression coefficients by smoothing the regression functions using basis function approximations. A choice of polynomial, Fourier, polynomial splines, and B-splines was given for the expansions. Hypothesis testing and confidence intervals for the coefficients were constructed using a resampling subject bootstrap method.

A lot of the methods for general linear functional response regression models assume that the residuals are independently and identically distributed. However, by the nature of most functional data, this assumption is not realistic and does not properly capture the within-function correlation. While these methods give decent point-wise estimates, any hypothesis tests or confidence intervals based around them are not reliable. Not properly accounting for the within-function correlation also leads to imprecise predictions. In addition, most of the methods discussed are not sufficient for a large number of predictors and/or predictors with

complex interactions. This is a problem in the motivating example due to the fact that our covariates of interest are vectors of highly correlated, related values.

2.2.2 Functional Mixed Effects Models

The methods described above all deal with fixed effects covariates, but random effects may also be of interest in functional modeling. A variety of techniques have been developed in recent years. One of the first to introduce the functional mixed effects model was Guo (2002). The general set up for this model is

$$y(t_{ij}) = X_{ij}\boldsymbol{\beta}(t_{ij}) + Z_{ij}\boldsymbol{\alpha}_i(t_{ij}) + e_{ij}$$

where $y(t_{ij})$ is the response of the i^{th} curve at t_{ij}. In this paper, both $\boldsymbol{\beta}(\cdot)$ and $\boldsymbol{\alpha}_i(\cdot)$ are modeled in the same functional space to ensure that they have the same smoothness properties. Specifically, cubic smoothing splines are used and combined with linear mixed model extentions to estimate both the fixed and random effects. A likelihood-ratio test was developed to test the fixed effects. The purpose of the random effects was to account for between-subject variations. There was no inference done on the random effects and they were included mainly to improve the fit of the model.

Another method to handle random effects was proposed by Aston et al. (2010). This functional mixed model is a two step procedure based on the principal components. It can handle a large number of covariates, but requires the observations to be observed on a common grid. The model is as follows:

$$\mathrm{E}\{Y_i(t)|X_i, Z_i\} = \mu(t) + \sum_{j=1}^{K} \mathrm{E}\{A_{ij}|X_i, Z_i\}\phi_j(t) \tag{2.2}$$

where X_i is a set of scalar effects, Z_i is a set of random effects, $\phi_j(t)$ is the j^{th} basis function, and K is the total number of basis functions. The eigenfunctions are used as the basis functions in (2.2). Therefore, the first step is to estimate the eigenfunctions based on the empirical

covariance matrix using functional principal components analysis (FPCA). The set of scores associated with each eigenfunction is computed as

$$\hat{A}_{ij} = \sum_{k=1}^{m} \{Y_i(t_k) - \hat{\mu}(t_k)\}\hat{\phi}_j(t_k)\Delta_k,$$

where $\Delta_k = t_k - t_{k-1}$ and m is the number of time points for each subject. The mean function, $\hat{\mu}(t_j)$, is calculated from the mean of the data. It is assumed that the covariates effect the data through the principal component scores. In other words, they are modeled as $\mathrm{E}(A_{ij}|X_i, Z_i) = X_i\beta^{(j)} + Z_i\gamma^{(j)}$, where $\gamma^{(j)} \sim \mathrm{N}(0, \Sigma_{\gamma^{(j)}})$. Using the scores as the dependent variables, independent linear mixed models are used to estimate the covariates for each set of scores, $j = 1, \ldots, K$. Standard linear mixed effects model analysis can be used to estimate and test the fixed effects and estimate the random effects. Therefore, this model includes random effects that account for the within-function correlation, but no testing is done. Additionally, the method does not estimate the curve-to-curve (or between subject) variability.

One of the first methods to introduce inference on random effects was proposed by Morris & Carroll (2006). This fully Bayesian model is extremely flexible and centered around wavelet regression. Wavelets are used to represent other functions through groups of orthonormal basis functions. Morris & Carroll (2006) combine wavelet expansion and the standard functional mixed model, $\mathbf{Y}(t) = \boldsymbol{X}\mathbf{B}(t) + \boldsymbol{Z}\boldsymbol{U}(t) + \boldsymbol{E}(t)$. Since they assume that the functions are observed on a common equally spaced grid, $\mathbf{t} = (t_1, \ldots, t_T)^T$, their model is written in matrix notation as

$$\mathbf{Y} = \boldsymbol{X}\mathbf{B} + \boldsymbol{Z}\boldsymbol{U} + \boldsymbol{E} \tag{2.3}$$

where $\mathbf{B}$ is a $p \times T$ matrix of fixed effects and $\boldsymbol{U}$ is an $m \times T$ matrix of random effects. The response matrix is constructed such that each row is a function (or curve) and the columns are the grid $\mathbf{t}$, meaing that $\mathbf{Y}$ is an $N \times T$ matrix. In the standard functional mixed model, it is assumed that $\boldsymbol{U}(t)$ and $\boldsymbol{E}(t)$ are independent and realizations from mutivariate Gaussian processes with mean zero. Thus, for the model in (2.3), they follow matrix normal distributions

such that $\boldsymbol{U} \sim \text{MN}(\mathbf{P}, \mathbf{Q})$ and $\boldsymbol{E} \sim \text{MN}(\boldsymbol{R}, \boldsymbol{S})$. The between-function covariance is represented by the $m \times m$ matrix $\mathbf{P}$, the within-function covariance surface is represented by the $T \times T$ matrix $\mathbf{Q}$, and $\boldsymbol{E}$ is the matrix of residual errors.

The curve data is projected onto the wavelet space by applying the discrete wavelet transform method to each row of $\mathbf{Y}$. This leads to the transformed model $\boldsymbol{D} = \mathbf{Y}\mathbf{W}^T$. Once the appropriate priors are put on the parameters, MCMC is performed to obtain posterior samples for the parameters in the wavelet-space version of the functional mixed model. This wavelet-based method produces nonparametric estimates of the random effects functions, as well as covariance matrices for both between-subject and within-subject effects. Since the method is fully Bayesian, inference can be done on any of the parameter estimates using posterior samples from the model, including the variance parameter estimates. Similarly, the posterior predictive distributions can be calculated in order to make predictions for a new observation.

Antoniadis & Sapatinas (2007) used a similar wavelet based decomposition in their functional model, but instead focused on developing a frequentist hypothesis testing procedure. Using profile and restricted profile maximum likelihood estimators, they developed a hypothesis test to determine the significiance of the curve-to-curve (or between subject) variability. A limitation of the wavelet-based functional mixed effects models described above is that they require the response data to be observed on a fine grid of points. That is, each subject must have measurements recorded at the same equally spaced time points. In larger studies, across a wide range of years, this requirement is often not feasible. Additionally, the methods focus on accounting for subject specific variation only, in both estimation and testing, and do not discuss the addition of a random covariate.

Another class of models in this field is functional additive mixed models. The specific class of models proposed by Scheipl et al. (2015, 2016) are designed for correlated functional responses and support a variety of correlation structures. The models are also able to handle either dense

or sparse observations. The structured additive regression models have the form

$$y_i(t) = \sum_{r=1}^{R} f_r(X_{ri}, t) + \epsilon_i(t).$$

For each r term in the model, $f_r(\cdot, \cdot)$ is a function of t and some subset of the full set of covariates. Types of covariate subsets include scalar and functional covariates, as well as nested or crossed grouping factors. These can either be constant or vary across t. The form of the function depends on the type of covariate set and the assumptions made about the smoothness of the covariates. It should be noted that it is not necessary to assume the same smoothness property for all functions. Random effects, either scalar or functional, can be included for a grouping variable g through the terms b_g or $b_g(t)$, respectively. Each function is represented by a linear combination of basis functions defined by tensor products. Using the tensor product representation of the functions, the model is rewritten in the form of a mixed model and the parameters are estimated using penalized regression.

Testing is not explicitly discussed for this method, as the main focus is on estimation and prediction. While some ad hoc testing can be done using this model, the type I error and power are unknown. For the purposes of our motivating example, this is a limitation. Additionally, the authors state that the method has a tendency to overfit complex effects, which can lead to inaccurate predictions for new observations. A false sense of confidence is also given with the standard errors provided for the new predictions. Using the provided standard errors in these types of models results in very narrow confidence bands with poor coverage. Therefore, this model is not suitable for testing, or even estimating, the joint effect of a large number of related covariates, as are present in the motivating example.

2.2.3 Gaussian Process Functional Regression

An area of study somewhat separate from the traditional functional data analysis literature is Gaussian process functional regression (GPFR). Shi & Choi (2011) developed GPFR models

mainly for prediction purposes. The proposed model is defined as,

$$y_m(t) = \mu_m(t) + \tau_m(\boldsymbol{x}_m) + \epsilon_m(t)$$

for $m = 1, \ldots, M$ subjects across time points $\{t_{mi}, i = 1, \ldots, n_m\}$. In this context, is it possible that a subject, m, has multiple curves. The common mean structure across curves is modeled by $\mu_m(t)$ and depends on the set of scalar covariates, $\mathbf{u}_m$. The covariance structure of $y_m(t)$ is determined by $\tau_m(\boldsymbol{x}_m)$ and depends on the set of functional covariates $\boldsymbol{x}_m$. It is assumed that $\tau_m(\boldsymbol{x}_m)$ follows a Gaussian process distribution, such that $\tau_m(\boldsymbol{x}_m) \sim GP_m(0, \boldsymbol{\Psi}_m)$ where each entry of the covariance structure is defined through a kernel, k_m, as $\Psi_{m_{i,j}} = k_m(\boldsymbol{x}_{mi}, \boldsymbol{x}_{mj}; \boldsymbol{\theta}_m)$. Thus, there is a Gaussian process model for each subject where the covariance structure is based on realization of functional covariates for a specific subject. This type of model is good for prediction, but because of the structure, can only determine how a covariate effects a response within a particular curve. The model lacks the ability to determine how a covariate, or set of covariates, effects the response trajectory in general. This type of information is essential in a testing procedure, which is essential to one of the clinical questions in our motivating example.

2.3 Data Analysis in Epidemiology

The association between DMRs and other factors has been previously studied in the epidemiologic literature. Parental obesity (both maternal and paternal) has been found to be associated with altered methylation patterns of children in specific DMRs (Soubry et al., 2013, 2015). Lead exposure has also been associated with methylation differences (Li et al., 2016). Additionally, a study by Wang et al. (2015b) found that HIF3A DNA methylation is associated with childhood obesity in a cohort of Chinese Han children.

From a statistical stand point, the analysis in the studies on DMRs have, for the most part, been fairly naive. In most of the studies involving DMRs, the CpG sites are generally analyzed separately or the mean of the DMR is used. Analyzing the CpG sites separately incorrectly

assumes independence and using the mean across sites loses information. Common statistical techniques found in these studies are chi-square tests of association, t-tests, and linear regression models. The aim of the obesity study by Wang et al. (2015b) is most similar to those in our motivating example. However, the study did not investigate how methylation patterns affect growth or obesity across time. In the study, children between the ages of 7 and 17 were classified as obsese or not obese, therefore, association was tested at a fixed time point only. Methylation patterns across time were analyzed and discussed in the study on lead exposure, but the mean of the DMR and point-wise regression were used (Li et al., 2016).

More sophisticated statistical methods are needed to appropriately explore the relationships between methylation values and growth trajectories. The methods discussed in this dissertation provide these types of advanced techniques and also fill in some gaps in the functional data literature.

Chapter 3

Association Study of Children's Growth and Methylation

3.1 Introduction

The goal of this chapter is to introduce a new method that enables the testing of the joint effect of a set of related, and possibly correlated, covariates on irreguarly spaced functional data. As described in Chapter 1, this method is motivated by the NEST growth curve data. Literature has suggested that certain gene methylation profile characteristics can impact the trajectory of a child's weight trajectory. Each DMR contains a different number of sites, with values that are highly correlated. Due to the correlation of sites within a DMR and the desire to assess the joint effect of the CpG sites, an additive functional model is not appropriate. It is also possible that, jointly, the CpG sites within a DMR have a nonlinear effect on the growth trajectory. In the current functional data literature, there is not an adequate testing procedure for a set of related covariates such as this. Our procedure combines functional methods with Gaussian process regression to provide a nonlinear model for the joint effect of a set of covariates. This joint effect is modeled as a random effect and we utilize variance component testing to determine significance. Additionally, the method is flexible enough to allow the data to be sparse

and irregularly observed across time, as is the case in the NEST data.

3.2 Model Framework

Suppose for $i = 1, \ldots, N$, $Y_i(t)$ denotes the functional response of the i^{th} subject at some point t in a continuous interval, $t \in [0, T]$. For simplicity, assume that each subject is observed at n equally spaced time points, $t_1, \ldots, t_n$. Let $\boldsymbol{m}_i = (m_{i1}, \ldots, m_{iP})^T$ be a vector of related covariates of interest. Let $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iC})^T$ denote other secondary covariates, such as demographic or clinical information. Let $\boldsymbol{M} = (\boldsymbol{m}_1, \ldots, \boldsymbol{m}_N)^T$ be the full matrix of related covariates for all subjects, where each row is a subject and each column is a covariate.

In our scenario, the $Y_i(\cdot)$ values are the growth trajectories of children measured over 1500 days. Specifically, $Y_i(t_j)$ is the weight, in kilograms, of child i on the j^{th} day. The covariates of interest are gene methylation values within a specific DMR. In this chapter, we look at the PEG3 DMR, which contains ten CpG sites. For each child, the vector of methylation values for PEG3 is the $\boldsymbol{m}_i$ vector in the model, with $P = 10$ being the number of sites. We are interested in testing whether the methylation values of the PEG3 DMR affect the trajectory of a child's growth curve. Possible secondary covariates in this data set include race, smoking status of the mother during pregnancy, age of the mother at delivery, information about education and income status, and breast feeding information. These covariates are described in further detail in Section 3.4.

Consider the model

$$Y_i(t) = \mu(t) + \sum_{c=1}^{C} x_{ic}\beta_c(t) + \alpha_i(t) + h(\boldsymbol{m}_i, t) + \epsilon_i(t) \tag{3.1}$$

where $\mu(\cdot)$ is the overall mean growth trajectory, x_{ic} is the value of the c^{th} baseline covariate for subject i, $\beta_c(\cdot)$ are the unknown coefficient functions, and $\epsilon_i(\cdot)$ is a random error process with mean zero and finite variance. Both $\alpha_i(t)$ and $h(\boldsymbol{m}_i, t)$ follow time dependent, latent processes that are subject specific and can be thought of as functional random components. The inclusion

of $\alpha_i(t)$ and $h(\boldsymbol{m}_i, t)$ are the main contributions of this method. The $\boldsymbol{m}_i$ vector is modeled non-parametrically through $h(\cdot, \cdot)$, which also allows the effect of $\boldsymbol{m}_i$ to vary over time. The $\alpha_i(\cdot)$ terms are assumed to be independent realizations from a latent process. They capture the within subject correlation across time and are essential to a suitable model fit and proper inference.

It is assumed that even after accounting for the baseline covariates and independent random effects, the response trajectories of two subjects can still be similar if their levels of $\boldsymbol{m}_i$ are similar as well. The goal is to test if there is an association between the covariate vector $\boldsymbol{m}_i$ and the response curve. This can be measured through $h(\boldsymbol{m}_i, t)$. Therefore, we can determine the significance of the covariate vector of interest by testing the null hypothesis

$$\mathrm{H}_0 \colon h(\boldsymbol{m}_i, t) = 0 \text{ for all values of } \boldsymbol{m}_i \text{ and } t. \tag{3.2}$$

In (3.2), H_0 is an infinite dimensional hypothesis test and is therefore difficult to test. We propose a novel, two-step fitting procedure based on Gaussian process (GP) regression to reduce the dimension and perform the hypothesis test. The functional mixed model is first reformulated as a mixed model. The first step uses a common technique in the functional literature, functional PCA, to handle the time component of the model. The second step fits the model and tests the random effects. Since the model is formulated in terms of mixed model notation, we utilize computational techniques that are already established for estimating and testing variance components in the mixed model setting.

3.2.1 Mixed Model Formulation

The components of the model in (3.1) can be represented using basis expansions. The mean function, $\mu(\cdot)$, is modeled using truncated power basis splines, as in Wand (2003). Notationally,

$$\mu(t) = \mu_0 + \mu_1 t + \sum_{k=1}^{K} \mu_{k+1}(t - \kappa_k)_+,$$

where $(t-\kappa_k)_+ = 0$ if $t \leq \kappa_k$ and $(t-\kappa_k)_+ = (t-\kappa_k)$ if $t > \kappa_k$. K is the total number of knots and is chosen using a simple formula described in Wand (2003) such that $K = \min\{(N \times n)/4, 35\}$. The placement of the knots is such that κ_k is the $\frac{k+1}{K+2}$th quantile of $\{t_1, \ldots, t_n\}$. The parameters μ_0 and μ_1 are the linear components of the mean function and are modeled as fixed effects. Each parameter in the vector $\mathbf{u}_\mu = (\mu_2, \ldots, \mu_{K+1})^T$ is modeled as a random effect that follows a normal distribution with mean 0 and variance σ^2_μ.

The remaining components, $\beta_c(\cdot)$, $\alpha_i(\cdot)$, and $h(\cdot, \cdot)$, are expanded as follows:

$$\begin{aligned}
\beta_c(t) &= \phi_1(t)\beta_{1c} + \cdots + \phi_B(t)\beta_{Bc}, \\
\alpha_i(t) &= \phi_1(t)\alpha_{1i} + \cdots + \phi_B(t)\alpha_{Bi}, \\
h(\boldsymbol{m}_i, t) &= \phi_1(t)h_1(\boldsymbol{m}_i) + \cdots + \phi_B(t)h_B(\boldsymbol{m}_i),
\end{aligned}$$

where $\phi_1(\cdot), \ldots, \phi_B(\cdot)$ are basis functions. In our proposed method, the eigenfunctions of $Y(\cdot)$ are used as the $\phi(\cdot)$'s to provide the basis for representing the functional data. These eigenfunctions are obtained by performing functional PCA on the $Y(\cdot)$'s. We utilize FPCA because it is a data driven approach that does not assume any type of parametric model and it can be performed on sparsely observed data. There are several methods available for performing FPCA in `R` including `fpca.mle` (Peng & Paul, 2009), `face.sparse` (Xiao et al., 2016), and `fpca.sc` (Goldsmith et al., 2013). To estimate this set, we use the restricted MLE procedure proposed by Peng & Paul (2009). When the FPCA is performed, the eigenfunctions are evaluated over a grid of 500 equally spaced timepoints for $t \in [0, T]$. This vector of grid points is denoted as $\boldsymbol{t}_{\tilde{g}}$. A total of B eigenvalues and eigenfunction vectors are selected based on the percent of variance explained (PVE), such that PVE $\geq 95\%$. Optionally, B can be set manually as well. The set of eigenvalues, eigenfunctions, and scores are denoted $\left\{\hat{\lambda}_b, \hat{\phi}_b(\cdot), \hat{\zeta}_{ib}\right\}_{b=1\ldots,B}$. For the selected eigenfunctions, the eigenfunction vectors are denoted as $\hat{\boldsymbol{\phi}}_b(\boldsymbol{t}_{\tilde{g}})$. The values are then interpolated to evaluate $\hat{\phi}_b(t)$ for a specific timepoint t corresponding to a time point in the given data set. Then, for each

individual, the model to be fit at a specific timepoint t is

$$\begin{aligned} Y_i(t) = \mu_0 + \mu_1 t + \sum_{k=1}^{K} [\mu_{k+1}(t-\kappa_k)_+] \\ + \sum_{c=1}^{C} [x_{ic}\{\hat{\phi}_1(t)\beta_{1c} + \cdots + \hat{\phi}_B(t)\beta_{Bc}\}] \\ + \sum_{b=1}^{B} [\hat{\phi}_b(t)\alpha_{bi} + \hat{\phi}_b(t)h_b(\boldsymbol{m}_i)] + \epsilon_i(t). \end{aligned} \tag{3.3}$$

In (3.3), μ_0, μ_1, and $\beta_{1c}, \ldots, \beta_{Bc}$, $c = 1, \ldots, C$, are modeled as fixed effects. The remaining parameters, $\mu_2, \ldots, \mu_K$, $\alpha_{1i}, \ldots, \alpha_{Bi}$, and $h_1(\cdot), \ldots, h_B(\cdot)$ are modeled as random effects. As described before, $\mathbf{u}_\mu \sim \mathrm{N}(\mathbf{0}, \sigma_\mu^2 \boldsymbol{I}_K)$. For $b = 1, \ldots, B$, each $\boldsymbol{\alpha}_b = (\alpha_{b1}, \ldots, \alpha_{bN})^T$ follows a normal distribution, $\boldsymbol{\alpha}_b \sim \mathrm{N}(\mathbf{0}, \delta_b \boldsymbol{I}_N)$. We assume $\boldsymbol{h}_b(\cdot)$ follows a Gaussian process. As a result, $\boldsymbol{h}_b(\boldsymbol{M}) = \{h_b(\boldsymbol{m}_1), \ldots, h_b(\boldsymbol{m}_N)\}^T$ follows a multivariate normal distribution for $b = 1, \ldots B$.

3.2.2 Gaussian Process Modeling

To model the functional random components of $\boldsymbol{h}_b(\boldsymbol{M})$, we utilize a Gaussian process modeling procedure. As defined in Rasmussen (2006), a Gaussian process is a collection of random variables. Gaussian process regression is a flexible extension of a linear model and defines a distribution over an infinite number of functions. Each observation is associated with a random variable and the joint distribution of any finite set of these random variables is a Gaussian distribution (Rasmussen, 2006). Let $f(\boldsymbol{x})$ define a real process. Then, the Gaussian process is written as

$$f(\boldsymbol{x}) \sim GP\{m(\boldsymbol{x}), R(x_i, x_j)\},$$

where $m(\boldsymbol{x})$ is the mean function and $R(x_i, x_j)$ is the covariance function. It is common practice to assume that the mean vector of the Gaussian process distribution is zero. In this case, the Gaussian process is completely defined by its covariance function. An important property of the Gaussian process is the marginalization property. This means that if some Gaussian process is

defined such that $(y_1, y_2)^T \sim \mathrm{N}(\boldsymbol{\mu}, \Sigma)$, then it also must be true that $y_1 \sim \mathrm{N}(\mu_1, \Sigma_{11})$, where Σ_{11} is the appropriate submatrix of Σ (Rasmussen, 2006).

Observations x_i and x_j are related through some kernel function $R(x_i, x_j)$, which defines the covariance matrix of the Gaussian process. It is also possible to pass vector observations through the kernel function. The form of the kernel function depends on how the observations are related (Ebden, 2015). Our proposed method offers three main kernel options for the construction of the covariance matrix. For any given input vectors, $\boldsymbol{m}_i$ and $\boldsymbol{m}_j$, the choices are as follows:

- Linear kernel: $R(\boldsymbol{m}_i, \boldsymbol{m}_j) = \boldsymbol{m}_i^T \boldsymbol{m}_j$
- Polynomial kernel, of power d: $R(\boldsymbol{m}_i, \boldsymbol{m}_j) = (1 + \boldsymbol{m}_i^T \boldsymbol{m}_j)^d$
- Gaussian kernel: $R(\boldsymbol{m}_i, \boldsymbol{m}_j) = \exp(\frac{||\boldsymbol{m}_i - \boldsymbol{m}_j||^2}{2\sigma^2})$

For simplicity, we focus on a specific polynomial kernel, the quadratic kernel, in which $d = 2$. Additionally, we let $\sigma^2 = 1$ in the Gaussian kernel.

In the motivating example, the $\boldsymbol{m}_i$ vectors are used as inputs in the kernel function to construct the covariance matrix. We assume that the effects of $\boldsymbol{m}_i$ and $\boldsymbol{m}_j$ are correlated and that the magnitude of the correlation depends on their similarity. Therefore, we can capture the effect that the similarity of the $\boldsymbol{m}_i$ vectors may have on the response through the kernel function. In the construction of the covariance matrix, some type of standardization or normalization of the inputs to the kernel function is common in the Gaussian process literature. If the inputs are not standardized, it is often the case that the variance grows at too large of a rate and does not make sense (Rasmussen, 2006). Simulation studies showed that this was the case in our method. Thus, we standardize the columns of $\boldsymbol{M}$ and denote the standardized matrix $\widetilde{\boldsymbol{M}}$. By standardizing the columns of $\boldsymbol{M}$, the resulting covariance matrix entries did not become too large and the matrix was more stable, resolving various numerical issues. In the NEST data application, $\boldsymbol{M}$ is the matrix of CpG sites for a specific DMR, where each column is a site and each row is a subject. Thus, to form $\widetilde{\boldsymbol{M}}$, each site is standardized across subjects to have mean 0 and variance 1.

Then, each $\boldsymbol{h}_b(\boldsymbol{M})$ follow a Gaussian process, such that $\boldsymbol{h}_b(\boldsymbol{M}) \sim \mathrm{N}(\mathbf{0}, \gamma_b \boldsymbol{R})$. The covariance matrix, $\boldsymbol{R}$, depends on $\widetilde{\boldsymbol{M}}$ through the specified kernel, such that $\boldsymbol{R}_{ij} = R(\widetilde{\boldsymbol{m}}_i, \widetilde{\boldsymbol{m}}_j)$, and γ_b is an unknown variance component. Recall the null hypothesis from (3.2). In order for $h(\boldsymbol{m}_i, t)$ to be equal to zero for all $\boldsymbol{m}_i$, $\boldsymbol{h}_b(\boldsymbol{M})$ must equal zero for all b. Based on the definition of the Gaussian process, $\boldsymbol{h}_b(\boldsymbol{M})$ is defined completely by its covariance. Since $\boldsymbol{R}$ is known, then the γ_b variance components are the parameters that must be tested. Therefore, we rewrite H_0 as

$$\mathrm{H}_0 \colon \gamma_b = 0, \text{ for all } b = 1, \ldots, B. \tag{3.4}$$

3.2.3 Full Model

We formulate the functional model as a mixed model in matrix notation, considering the general case where it is possible for each individual to have a different number of observations. That is, the observations for the i^{th} individual are $j = 1, \ldots, n_i$. Then, the full model for all observations can be written as

$$\mathbf{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}_\mu \mathbf{u}_\mu + \boldsymbol{Z}_\alpha \mathbf{u}_\alpha + \boldsymbol{Z}_h \mathbf{u}_h + \boldsymbol{\epsilon}. \tag{3.5}$$

Define $\tilde{n} = n_1 + n_2 + \cdots + n_N$. Then, $\mathbf{Y}_{\tilde{n}\times 1} = (\mathbf{Y}^T_{1_{n_1}\times 1}, \mathbf{Y}^T_{2_{n_2}\times 1}, \ldots, \mathbf{Y}^T_{N_{n_N}\times 1})^T$ is the stacked response vector and $\boldsymbol{\beta} = (\mu_0, \mu_1, \beta_{11}, \ldots, \beta_{B1}, \ldots, \beta_{1C}, \ldots, \beta_{BC})^T$ is the vector of fixed effects. The vector of random effects is $\mathbf{u} = (\mathbf{u}_\mu^T, \mathbf{u}_\alpha^T, \mathbf{u}_h^T)^T$, where $\mathbf{u}_\mu$ is as defined in Section 3.2.1, $\mathbf{u}_\alpha = (\boldsymbol{\alpha}_1^T, \ldots, \boldsymbol{\alpha}_B^T)^T$ and $\mathbf{u}_h = \{\boldsymbol{h}_1(\boldsymbol{M})^T, \ldots, \boldsymbol{h}_B(\boldsymbol{M})^T\}^T$.The design matrix for the covariates is $\boldsymbol{X}_{\tilde{n}\times(2+BC)} = [\tilde{\boldsymbol{X}}_0 \ \tilde{\boldsymbol{X}}_1 \ \cdots \ \tilde{\boldsymbol{X}}_C]$. The matrix $\tilde{\boldsymbol{X}}_0$ represents the intercept and linear effect of the mean and is as follows:

$$\tilde{\boldsymbol{X}}_0 = \begin{bmatrix} 1 & \ldots & 1 & \ldots & \ldots & 1 & \ldots & 1 \\ t_{11} & \ldots & t_{1n_1} & \ldots & \ldots & t_{N1} & \ldots & t_{Nn_N} \end{bmatrix}^T$$

For any $c = 1, \ldots, C$ the design matrix for a fixed effect is as follows:

$$\tilde{\boldsymbol{X}}_c = \begin{bmatrix} x_{1c}\phi_1(t_{11}) & \dots & x_{1c}\phi_B(t_{11}) \\ \vdots & & \vdots \\ x_{1c}\phi_1(t_{1n_1}) & \dots & x_{1c}\phi_B(t_{1n_1}) \\ \vdots & & \vdots \\ \vdots & & \vdots \\ x_{Nc}\phi_1(t_{N1}) & \dots & x_{Nc}\phi_B(t_{N1}) \\ \vdots & & \vdots \\ x_{Nc}\phi_1(t_{Nn_N}) & \dots & x_{Nc}\phi_B(t_{Nn_N}) \end{bmatrix}$$

Additionally, $\boldsymbol{Z}_\mu$, $\boldsymbol{Z}_\alpha$, and $\boldsymbol{Z}_h$ are known matrices associated with the random effects vectors. They are defined as follows:

$$\boldsymbol{Z}_\mu = \begin{bmatrix} (t_{11} - \kappa_1)_+ & \dots & (t_{11} - \kappa_K)_+ \\ \vdots & & \vdots \\ (t_{1n_1} - \kappa_1)_+ & \dots & (t_{1n_1} - \kappa_K)_+ \\ (t_{21} - \kappa_1)_+ & \dots & (t_{21} - \kappa_K)_+ \\ \vdots & & \vdots \\ (t_{2n_2} - \kappa_1)_+ & \dots & (t_{2n_2} - \kappa_K)_+ \\ \vdots & & \vdots \\ \vdots & & \vdots \\ (t_{N1} - \kappa_1)_+ & \dots & (t_{N1} - \kappa_K)_+ \\ \vdots & & \vdots \\ (t_{Nn_N} - \kappa_1)_+ & \dots & (t_{Nn_N} - \kappa_K)_+ \end{bmatrix}$$

and

$$\boldsymbol{Z}_\alpha = \boldsymbol{Z}_h = \begin{bmatrix} \boldsymbol{\phi}_1(\mathbf{t}_1) & \dots & \mathbf{0} & \dots & \dots & \boldsymbol{\phi}_B(\mathbf{t}_1) & \dots & \mathbf{0} \\ \vdots & \ddots & \vdots & \dots & \dots & \vdots & \ddots & \vdots \\ \mathbf{0} & \dots & \boldsymbol{\phi}_1(\mathbf{t}_N) & \dots & \dots & \mathbf{0} & \dots & \boldsymbol{\phi}_B(\mathbf{t}_N) \end{bmatrix}$$

The full random effect vector is distributed as

$$\begin{pmatrix} \mathbf{u}_\mu \\ \mathbf{u}_\alpha \\ \mathbf{u}_h \\ \boldsymbol{\epsilon} \end{pmatrix} \sim \mathrm{N}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{G}_\mu & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_\alpha & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{G}_h & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \sigma_\epsilon^2 \boldsymbol{I}_{\tilde{n}} \end{bmatrix} \right)$$

where $\mathbf{G}_\mu = \sigma_\mu^2 \boldsymbol{I}_K$, $\mathbf{G}_\alpha = \mathrm{diag}(\boldsymbol{\delta}) \otimes \boldsymbol{I}_N$, $\mathbf{G}_h = \mathrm{diag}(\boldsymbol{\gamma}) \otimes \boldsymbol{R}_{N\times N}$, $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_B)^T$ and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_B)^T$. Therefore, the full response vector $\mathbf{Y}$ is distributed as $\mathbf{Y} \sim \mathrm{N}(\boldsymbol{X}\boldsymbol{\beta}, \mathbf{V})$, where $\mathbf{V} = \boldsymbol{Z}_\mu \mathbf{G}_\mu \boldsymbol{Z}_\mu^T + \boldsymbol{Z}_\alpha \mathbf{G}_\alpha \boldsymbol{Z}_\alpha^T + \boldsymbol{Z}_h \mathbf{G}_h \boldsymbol{Z}_h^T + \sigma_\epsilon^2 \boldsymbol{I}_{\tilde{n}}$.

Again, recall the null hypothesis that $\gamma_b = 0$ for all $b = 1, \ldots, B$. Because of the distribution of $\boldsymbol{h}_1(\boldsymbol{M}), \ldots, \boldsymbol{h}_B(\boldsymbol{M})$, this is equivalent to testing the vector of the corresponding variance parameters. Thus, we treat $(\sigma_\mu^2, \delta_1, \ldots, \delta_B)^T$ as nuisance variance parameters and test the hypothesis

$$\mathrm{H}_0 : \boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_B)^T = \mathbf{0}.$$

Model Example

To ensure the clarity of the full model written in matrix notation, we provide an example. For simplicity, we assume that there are two subjects, one with four observed time points and one with three. We also assume there is only one covariate and that two eigenfunctions are used in the expansion. Therefore, $N = 2$, $n_1 = 4$, $n_2 = 3$, $C = 1$, and $B = 2$. Then,

$$
\begin{pmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{14} \\ Y_{21} \\ Y_{22} \\ Y_{23} \end{pmatrix} = \begin{pmatrix} 1 & t_{11} & x_{11}\phi_1(t_{11}) & x_{11}\phi_2(t_{11}) \\ 1 & t_{12} & x_{11}\phi_1(t_{12}) & x_{11}\phi_2(t_{12}) \\ 1 & t_{13} & x_{11}\phi_1(t_{13}) & x_{11}\phi_2(t_{13}) \\ 1 & t_{14} & x_{11}\phi_1(t_{14}) & x_{11}\phi_2(t_{14}) \\ 1 & t_{21} & x_{21}\phi_1(t_{21}) & x_{21}\phi_2(t_{21}) \\ 1 & t_{22} & x_{21}\phi_1(t_{22}) & x_{22}\phi_2(t_{22}) \\ 1 & t_{23} & x_{21}\phi_1(t_{23}) & x_{23}\phi_2(t_{23}) \end{pmatrix} \begin{pmatrix} \mu_0 \\ \mu_1 \\ \beta_{11} \\ \beta_{12} \end{pmatrix} + \begin{pmatrix} (t_{11}-\kappa_1)_+ & \dots & (t_{11}-\kappa_K)_+ \\ (t_{12}-\kappa_1)_+ & \dots & (t_{12}-\kappa_K)_+ \\ (t_{13}-\kappa_1)_+ & \dots & (t_{13}-\kappa_K)_+ \\ (t_{14}-\kappa_1)_+ & \dots & (t_{14}-\kappa_K)_+ \\ (t_{21}-\kappa_1)_+ & \dots & (t_{21}-\kappa_K)_+ \\ (t_{22}-\kappa_1)_+ & \dots & (t_{22}-\kappa_K)_+ \\ (t_{23}-\kappa_1)_+ & \dots & (t_{23}-\kappa_K)_+ \end{pmatrix} \begin{pmatrix} \mu_2 \\ \vdots \\ \mu_K \end{pmatrix}
$$

$$
+ \begin{pmatrix} \phi_1(t_{11}) & 0 & \phi_2(t_{11}) & 0 \\ \phi_1(t_{12}) & 0 & \phi_2(t_{12}) & 0 \\ \phi_1(t_{13}) & 0 & \phi_2(t_{13}) & 0 \\ \phi_1(t_{14}) & 0 & \phi_2(t_{14}) & 0 \\ 0 & \phi_1(t_{21}) & 0 & \phi_2(t_{21}) \\ 0 & \phi_1(t_{23}) & 0 & \phi_2(t_{22}) \\ 0 & \phi_1(t_{23}) & 0 & \phi_2(t_{23}) \end{pmatrix} \begin{pmatrix} \alpha_{11} \\ \alpha_{12} \\ \alpha_{21} \\ \alpha_{22} \end{pmatrix} + \begin{pmatrix} \phi_1(t_{11}) & 0 & \phi_2(t_{11}) & 0 \\ \phi_1(t_{12}) & 0 & \phi_2(t_{12}) & 0 \\ \phi_1(t_{13}) & 0 & \phi_2(t_{13}) & 0 \\ \phi_1(t_{14}) & 0 & \phi_2(t_{14}) & 0 \\ 0 & \phi_1(t_{21}) & 0 & \phi_2(t_{21}) \\ 0 & \phi_1(t_{23}) & 0 & \phi_2(t_{22}) \\ 0 & \phi_1(t_{23}) & 0 & \phi_2(t_{23}) \end{pmatrix} \begin{pmatrix} h_1(M_1) \\ h_1(M_2) \\ h_2(M_1) \\ h_2(M_2) \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \\ \epsilon_{14} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{23} \end{pmatrix}
$$

3.2.4 Generalization to Time Varying Covariates

For simplicity, the model was described as having baseline covariates, or covariates that do not change value over time. However, the model can also be extended to handle time-varying covariates as well, as long as they correspond to the same time points. Since the response vector is stacked, the design matrix for the covariates, $\boldsymbol{X}$, is stacked as well. Therefore, the baseline covariate values are repeated across all rows of the subject. For covariates that change value over time, one would simply enter the corresponding value associated with the correct time row in the design matrix. Therefore, the design matrix for a time-varying covariate, in a model where subjects have unequal observations, would be as follows:

$$\tilde{\boldsymbol{X}}_c = \begin{bmatrix} x_{1c}(t_{11})\phi_1(t_{11}) & \dots & x_{1c}(t_{11})\phi_B(t_{11}) \\ \vdots & & \vdots \\ x_{1c}(t_{1n_1})\phi_1(t_{1n_1}) & \dots & x_{1c}(t_{1n_1})\phi_B(t_{1n_1}) \\ \vdots & & \vdots \\ \vdots & & \vdots \\ x_{Nc}(t_{N1})\phi_1(t_{N1}) & \dots & x_{Nc}(t_{N1})\phi_B(t_{N1}) \\ \vdots & & \vdots \\ x_{Nc}(t_{Nn_N})\phi_1(t_{Nn_N}) & \dots & x_{Nc}(t_{Nn_N})\phi_B(t_{Nn_N}) \end{bmatrix}$$

3.3 Linear Score Test for Variance Components

The distribution of $\mathbf{Y}$ using the full model matrix notation in (3.5) follows the linear mixed-effects model outlined in Qu et al. (2013). The model in their article has the form

$$\mathbf{Y}_{\tilde{n}} \sim \mathrm{N}_{\tilde{n}}\{\boldsymbol{X}\boldsymbol{\beta},\, \sigma_\epsilon^2 \mathbf{V}(\boldsymbol{\tau})\} \tag{3.6}$$

with

$$\mathbf{V}(\boldsymbol{\tau}) = \boldsymbol{I}_{\tilde{n}} + \sum_{i=1}^{I} \tau_i \mathbf{K}_i + \sum_{i=I+1}^{I+J} \tau_i \mathbf{K}_i \tag{3.7}$$

In (3.7), there are $(I + J)$ total variance components. The first I are treated as nuisance parameters, denoted as $\boldsymbol{\tau}_{(I)} = (\tau_1, \ldots, \tau_I)^T$ and the remaining J are the variance components of interest, denoted as $\boldsymbol{\tau}_{(J)} = (\tau_{I+1}, \ldots, \tau_{I+J})^T$. The $\mathbf{K}$'s are positive semidefinite, fixed matrices and σ^2_ϵ is the error variance parameter. Note that the τ's are variance component parameters that are scaled by a factor of σ^2_ϵ.

In our proposed model, $I = B + 1$ and $J = B$, for a total of $(2B + 1)$ variance components. The first I known positive semidefinite matrices are $\mathbf{K}_1 = \boldsymbol{Z}_\mu \boldsymbol{Z}_\mu^T$ and $\mathbf{K}_{b+1} = \boldsymbol{Z}_{\alpha,b} \boldsymbol{Z}_{\alpha,b}^T$ for $b = 1, \ldots B$. The vector of variance components that are considered to be nuisance parameters is $\boldsymbol{\tau}_{(I)} = (\sigma^2_\mu/\sigma^2_\epsilon,\, \delta_1/\sigma^2_\epsilon,\, \ldots,\, \delta_B/\sigma^2_\epsilon)^T$. For $b = 1, \ldots, B$, $\mathbf{K}_{I+b} = \boldsymbol{Z}_{h,b} \boldsymbol{R} \boldsymbol{Z}_{h,b}^T$ and the corresponding vector of variance components is $\boldsymbol{\tau}_{(J)} = (\gamma_1/\sigma^2_\epsilon, \ldots, \gamma_B/\sigma^2_\epsilon)^T$. In this case, we wish to test $H_0 \colon \boldsymbol{\tau}_{(J)} = \mathbf{0}$ against $H_1 \colon \boldsymbol{\tau}_{(J)} \geq \mathbf{0}$

The model fitting process in Qu et al. (2013) uses a profiled restricted maximum likelihood (PREML) approach based on a set of residual contrasts. In all cases, the error variance is profiled out of the restricted maximum likelihood. In cases where $I > 0$, as in our method, the additional nuisance parameters are also removed through profiling by choosing a residual contrast matrix $\mathbf{A}$, such that

$$\widetilde{\mathbf{Y}}_{\tilde{n}_2} = \mathbf{A}\mathbf{Y}, \quad \widetilde{\mathbf{K}}_i = \mathbf{A}\mathbf{K}_i, \quad \mathbf{A}\boldsymbol{X} = \mathbf{0}, \quad \mathbf{A}\mathbf{K}_i^{1/2} = \mathbf{0}, \quad \mathbf{A}\mathbf{A}^T = \boldsymbol{I},$$

where $\tilde{n}_2 = \tilde{n} - \mathrm{rk}(\boldsymbol{X})$. The constraint involving $\mathbf{K}_i$ holds for all $i = 1, \ldots, I$ and forces the nuisance random effect parameters to be treated as fixed effect factors under the null. Additionally, $\widetilde{\mathbf{V}}(\boldsymbol{\tau}) = \boldsymbol{I}_{\tilde{n}_2} + \sum_{i=1}^{I} \tau_i \widetilde{\mathbf{K}}_i + \sum_{i=I+1}^{I+J} \tau_i \widetilde{\mathbf{K}}_i$. To test $\boldsymbol{\tau}_{(J)}$, the score function is approximated using moment matching and the linear score test statistic is calculated. As derived in Qu et al. (2013), the score function for τ_i is

$$S_i(\boldsymbol{\tau}) = \frac{\tilde{n}_2}{2} \frac{\widetilde{\mathbf{Y}}^T \widetilde{\mathbf{V}}(\boldsymbol{\tau})^{-1} \widetilde{\mathbf{K}}_i \widetilde{\mathbf{V}}(\boldsymbol{\tau})^{-1} \widetilde{\mathbf{Y}}}{\widetilde{\mathbf{Y}}^T \widetilde{\mathbf{V}}(\boldsymbol{\tau}) \widetilde{\mathbf{Y}}} - \frac{1}{2} \mathrm{tr}\{\widetilde{\mathbf{V}}(\boldsymbol{\tau})^{-1} \widetilde{\mathbf{K}}_i\}$$

and the form of the expected information matrix is

$$\mathcal{I}_{ij}(\boldsymbol{\tau}) = \frac{\tilde{n}_2}{2(\tilde{n}_2+2)}[\text{tr}\{\widetilde{\mathbf{V}}(\boldsymbol{\tau})^{-1}\widetilde{\mathbf{K}}_i\widetilde{\mathbf{V}}(\boldsymbol{\tau})^{-1}\widetilde{\mathbf{K}}_j\} - \tilde{n}_2^{-1}\text{tr}\{\widetilde{\mathbf{V}}(\boldsymbol{\tau})^{-1}\widetilde{\mathbf{K}}_i\}\text{tr}\{\widetilde{\mathbf{V}}(\boldsymbol{\tau})^{-1}\widetilde{\mathbf{K}}_j\}].$$

Let $\boldsymbol{S}_{(I)}(\boldsymbol{\tau})$ and $\boldsymbol{S}_{(J)}(\boldsymbol{\tau})$ denote the first I and remaining J score statistics, respectively. As a reminder, this corresponds to the score statistics associated with the I nuisance parameters and those associated with the J variance components of interest. The information matrix is notated similarly and partitioned into the following four blocks: $\mathcal{I}_{(II)}(\boldsymbol{\tau})$, $\mathcal{I}_{(IJ)}(\boldsymbol{\tau})$, $\mathcal{I}_{(JI)}(\boldsymbol{\tau})$, and $\mathcal{I}_{(JJ)}(\boldsymbol{\tau})$. Additionally, let the parameter under the null be $\boldsymbol{\tau}_0 = (\boldsymbol{\tau}_{(I)}^T, \mathbf{0}_{(J)}^T)^T$. It is assumed that under the null hypothesis, $\boldsymbol{S}_{(J)}(\boldsymbol{\tau}_0)$ is approximated by a multivariate normal distribution with mean $\mu(\boldsymbol{\tau}_0) = \mathcal{I}_{(JI)}(\boldsymbol{\tau}_0)\mathcal{I}_{(II)}^{-1}(\boldsymbol{\tau}_0)\boldsymbol{S}_{(I)}(\boldsymbol{\tau}_0)$ and variance-covariance matrix $\vartheta(\boldsymbol{\tau}_0) = \mathcal{I}_{(JJ)}(\boldsymbol{\tau}_0) - \mathcal{I}_{(JI)}(\boldsymbol{\tau}_0)\mathcal{I}_{(II)}^{-1}(\boldsymbol{\tau}_0)\mathcal{I}_{(IJ)}(\boldsymbol{\tau}_0)$.

The test statistic is a convex combination of score statistics,

$$U(\hat{\boldsymbol{\tau}}_0) = \sum_{i=I+1}^{I+1+J} w_i(\hat{\boldsymbol{\tau}}_0)S_i(\hat{\boldsymbol{\tau}}_0), \tag{3.8}$$

where $w_i \geq 0$ are non-stochastic weights subject to $\sum_{i=I+1}^{I+1+J} w_i = 1$. By default, each w_i is proportional to the inverse standard deviation of the corresponding variance component under the null. These are obtained from the diagonal elements of the expected (or efficient) information matrix. Other weighting schemes include equal weights for each variance component of interest (in this case, B^{-1}), weights proportional to the inverse square root of the covariance matrix, or weights that minimize the variance of the test statistic, $U(\hat{\boldsymbol{\tau}}_0)$. For a detailed discussion of weight choices, see Qu et al. (2013). The first two moments are matched to the asymptotically theoretical moments and the distributional shape is a ratio of quadratic forms. However, instead of the usual scaled χ^2 approximation, the null distribution of $U(\hat{\boldsymbol{\tau}}_0)$ is a linear combination of independent χ^2 variables. The p-value is evaluated using the Applied Statistics Algorithm 155 (AS155) (Davies, 1980), which is available in the `CompQuadForm R` package (Duchesne & de Micheaux, 2010).

In our method, all of this is performed through the `varComp` package (Qu, 2017) using the following functions: `varComp`, `varComp.test`, and `anova.varComp`. For details and information about arguments that control the model fitting and testing, see Qu et al. (2013) and Qu & Qu (2015).

3.4 Real Data Application

We apply the proposed method to the NEST data discussed in Section 3.1. Participants in NEST were selected between 2009 and 2011 from pregnant women who visited one of six prenatal clinics in Durham County, NC. Prior to giving birth, a variety of demographic and lifestyle information was collected on the mother. Some of the demographic information included age, race, education history, and socioeconomic status. The information on the lifestyle of each mother was mostly concentrated on lifestyle during pregnancy. This included whether or not the mother smoked during pregnancy, the type of nutrients they consumed, and any toxic metals they were exposed to. In addition, some lifestyle information was collected after the birth, such as whether the mother breastfed her baby.

At birth, cord blood that contained methylation profile information for nine DMRs was collected and processed from each child. In this chapter, we focus on Paternally-expressed gene 3 (PEG3), which is expressed from the paternal allele. In mice and humans, disrupted PEG3 causes low birth weight (Chiavegatto et al., 2012; Vidal et al., 2015) frequently followed by steeper growth trajectories in early life, so-called catch-up growth (Kim et al., 2013). With or without frank obesity, such accelerated growth in early life is a consistent risk factor for cardiometabolic impairment in adulthood (Barker, 2004; Whincup et al., 2008; Ezzahir et al., 2005; Meas et al., 2008; Howe et al., 2010; Anderson et al., 2014; De Kroon et al., 2010). This is why we are interested in determining if there is a relationship between PEG3 and the growth trajectories of children in the NEST data.

The PEG3 DMR contains ten CpG sites. The methylation values at different CpG sites within PEG3 are highly correlated and may interact with each other in a complex way. Figure 3.1

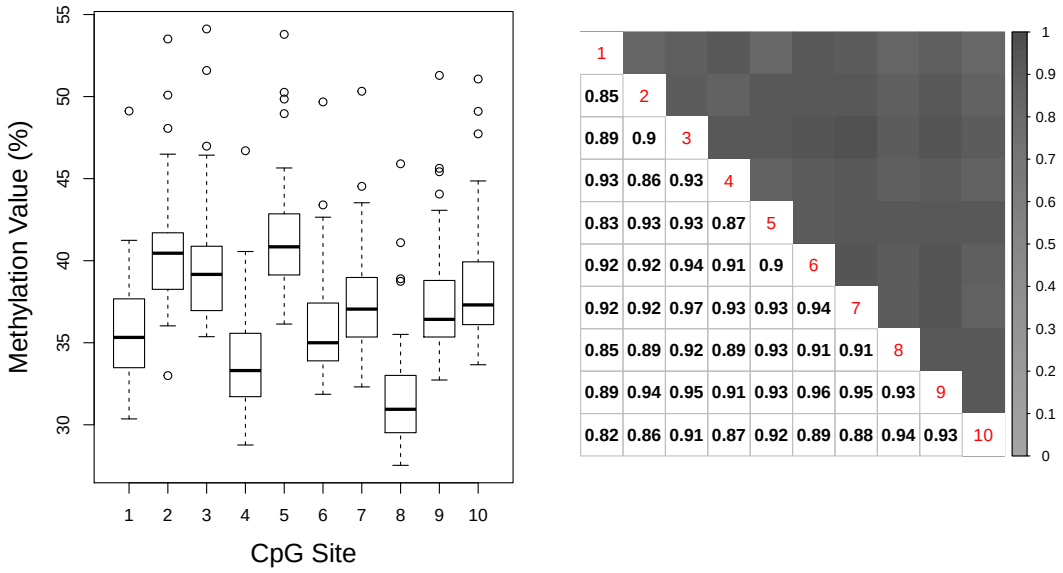


Figure 3.1: Boxplot and correlation matrix for female PEG3 DMR Values.

shows the relationship among the methylation values of females within the PEG3 DMR. The methylation values are percentages and a normal methylation value is considered to be 50%. However, this has been shown to vary within different DMRs. The boxplot shows that the means of the CpG sites fall between 31 and 41 and the standard deviations are between 3 and 4. All CpG sites appear to have a few outliers on the upper end, with the maximum values ranging from 50 to 55. Additionally, the sites are positively correlated with each other, which is to be expected since they are from the same DMR.

3.4.1 Data Cleaning

Since we were working with the raw NEST data, a decent amount of data cleaning was necessary prior to analysis. The full data set was a combination of data from various sources, thus there were some duplicate entries. To be consistent, we used data from only one of the sources (DEDUCE). Additionally, only subjects with complete methylation profiles were included in the analysis. The time variable used was age in days. The observations were cut off at 1500 days, which is a little over four years. A cut off was imposed on the data because there were very few time points in the four to five year range and the few time points that were present caused problems at the boundary when performing the FPCA. The response variable was the subject's weight, in kilograms (kg). If there happened to be more than one measurement in a day, the average of the measurements was used. Measurements greater than 50kg (110 pounds) were considered outliers and were removed, since this is not a reasonable weight for a child under five years of age. Additional outliers that caused extreme jumps in an individual's growth curve were manually removed. Individuals with less than two data points were removed from the analysis.

The covariates included (a) maternal smoking status, where 0 indicates "never smoked during pregnancy" and 1 otherwise; (b) breastfeeding status, where 1 indicates "ever breastfed the child" and 0 otherwise; and (c) education/income status, where 1 indicates both high education and high income and 0 otherwise. Only subjects with full covariate information were included in the analysis. After data cleaning, there were 44 female subjects and 45 male subjects. The cleaned data is presented in Figure 3.2.

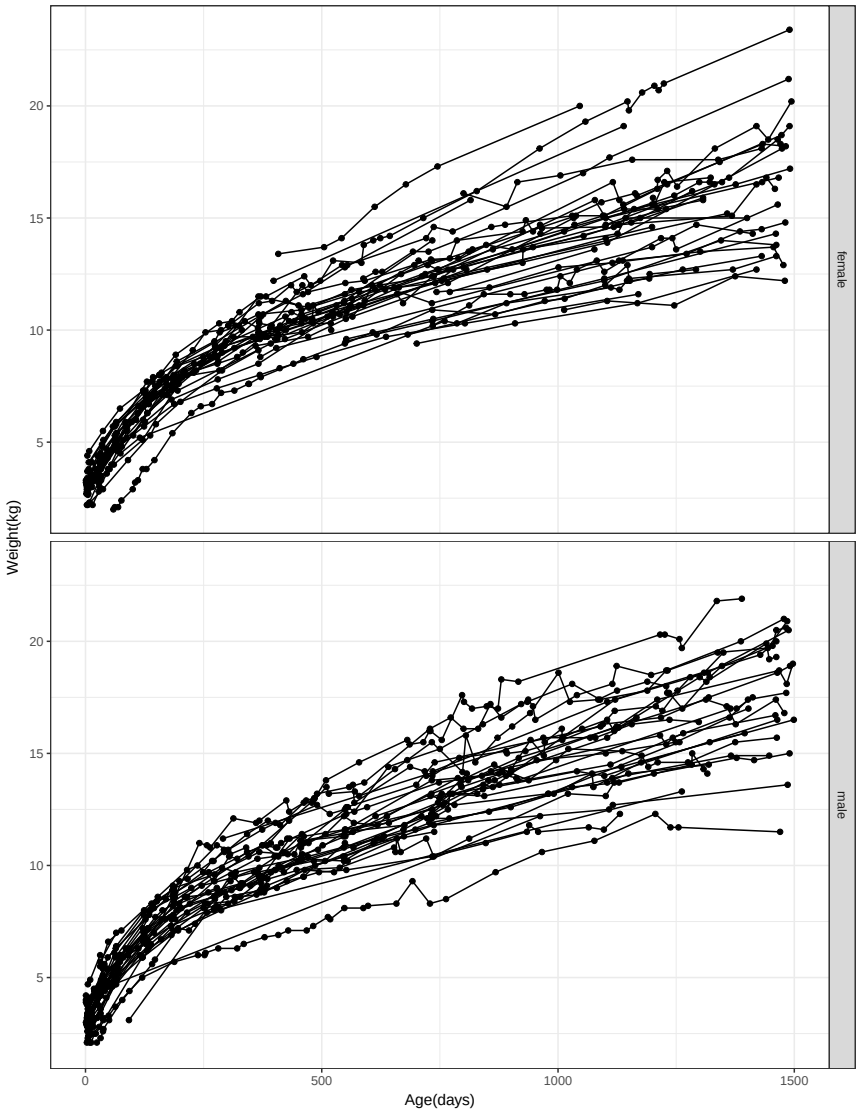


Figure 3.2: Cleaned data, stratified by gender, from DEDUCE source only. The top plot is the female data and the bottom plot is the male data.

It is common practice in the literature to perform growth-curve analyses stratified by gender. To further investigate whether this was a valid approach or not, we performed an eigenfunction analysis on the data. To combine the male and female data for a single analysis, the eigenfunction decomposition for males and females would need to be the same. This is because the eigenfunctions are used in the model expansion. However, we see in Figure 3.3 that although the

first eigenfunction is essentially the same for both males and females, the second eigenfunction behaves differently. Additionally, the male data needed three eigenfunctions to achieve a PVE above 95%, while the female data only needed two. This is further justification for performing a stratified analysis and treating the males and females as separate populations.

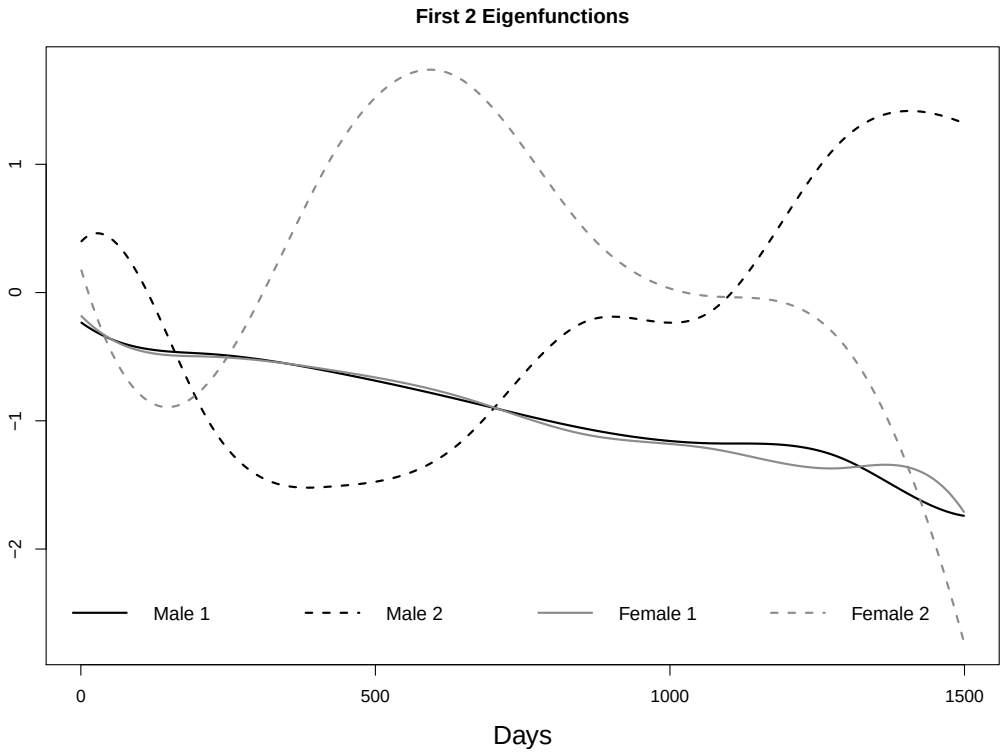


Figure 3.3: Plot of the first two eigenfunctions, for males and females.

3.4.2 Results

In addition to performing a stratified analysis on PEG3, three kernels were used to see if one detected a signal that another did not. As a conservative approach, we use a Bonferroni corrected significance level, α, within gender, such that $\alpha_0 = 0.05/3 = 0.0167$. For the male data, the

Table 3.1: Testing results (p-values) from PEG3 by Gender and Kernel.

	Linear	Quadratic	Gaussian
Male	0.7101	0.9487	0.9769
Female	0.0703	0.0132	0.4932

FPCA used in the first step of the procedure required three principal components to achieve PVE $\geq 95\%$. Therefore, three eigenfunctions were used in the expansion of the parameters. The parameter vector $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \gamma_3)^T$ was tested against the null $\boldsymbol{\gamma} = (0, 0, 0)^T$. As shown in Table 3.1, there were no significant results for the male data in any of the kernels. The high p-values for all kernels indicate that PEG3 does not significantly affect the growth trajectories of the male subjects.

For the female data, the FPCA required two principal components and therefore there were two variance components of interest. The parameter vector $\boldsymbol{\gamma} = (\gamma_1, \gamma_2)^T$ was tested against the null $\boldsymbol{\gamma} = (0, 0)^T$. The p-value results in Table 3.1 show that the quadratic kernel detected a significant signal, with a p-value of 0.013. This indicates that PEG3 is associated with the growth trajectories of female subjects and is worth further investigation.

Although we were specifically interested in PEG3, the remaining eight DMRs were tested as well. There were no other significant DMRs and we leave these results to Appendix A.

3.5 Simulation Experiment

3.5.1 Design

To test the performance of our proposed method, a simulation study was designed to assess the Type I error and power. The simulation study is modeled after the real data used in our application of the method. Thus, for all settings, the sample size was fixed at $N = 44$ and the range of t was $t \in [0, 1500]$. This is the sample size of the female data set and the age range of the children in days. There are three settings corresponding to the three kernel choices (linear, quadratic, and Gaussian). We use 1000 replications for the Type I error analysis, with

$\alpha_0 = 0.01$, 0.05, and 0.10. We use 250 replications for the power analysis, with $\alpha_0 = 0.05$.

The simulation uses female data from the PEG3 DMR to generate parameter values. This location has ten CpG sites, meaning that $P = 10$ and the methylation vector for each subject, $\boldsymbol{m}_i$, is 10×1. For each simulation, the rows of the methylation matrix, $\boldsymbol{M}$, are sampled with replacement from the female PEG3 methylation matrix. This sampled matrix is referred to as $\boldsymbol{M}_{\text{sim}}$. To create growth curves similar to those found in the real data application, the response vector, $\mathbf{Y}$, was generated using the logistic function,

$$Y(t) = \frac{C}{(1 + ae^{-bt})}.$$

The logistic function is a growth curve model used in Panik (2014). The parameters C, a, and b control the maximum value, shape, and scale of the simulated growth curve, which is different for each individual. Therefore, the set of parameters (C_i, a_i, b_i) is generated for $i = 1, \ldots, N$. The parameter a_i is randomly generated from the uniform distribution, $\text{U}(3, 5)$. The other parameters, C_i and b_i, are generated based on $\boldsymbol{M}_{\text{sim}}$. To do so, we compute two summary statistics for each subject, i: the mean methylation, denoted by $m^{*^1}_{i,\text{sim}} = \frac{1}{P}\sum_{p=1}^{P} m_{i,p}$ and the square of the methylation sums, denoted by $m^{*^2}_{i,\text{sim}} = \frac{1}{P}(\sum_{p=1}^{P} m_{i,p})^2$. The vector of a particular methylation summary for all subjects is denoted as $\boldsymbol{m}^*_{\text{sim}}$. A set of growth curves dependent on $\boldsymbol{m}^{*^1}_{\text{sim}}$ implies a linear methylation effect, while a set of growth curves dependent on $\boldsymbol{m}^{*^2}_{\text{sim}}$ implies a quadratic methylation effect.

The strength of the relationship between the response curves and $\boldsymbol{m}^*_{\text{sim}}$ is determined by $k \in [0, 1]$. When $k = 0$, $\boldsymbol{m}^*_{\text{sim}}$ has no effect on the generation of $\mathbf{b} = (b_1, \ldots, b_N)^T$ and thus each b_i is generated from $\text{U}(3, 6)$. In other words, this is the model under the null hypothesis. As k gets larger, $\boldsymbol{m}^*_{\text{sim}}$ has a greater effect on how b_i is generated for each subject, therefore having a greater effect on the shape of the response curves. Specifically, b_i is generated for each subject as

$$b_i = k^* m^*_{i,\text{sim}} + \text{Uniform}\{3 - \min(k^* \boldsymbol{m}^*_{\text{sim}}),\ 6 - \max(k^* \boldsymbol{m}^*_{\text{sim}})\},$$

where $k^* = \frac{k}{w}$ and w is a constant used to scale $\boldsymbol{m}^*_{\text{sim}}$ to ensure that the values in $\mathbf{b}$ are in the correct range. For $\boldsymbol{m}^{*^1}_{\text{sim}}$, $w = 13$ and for $\boldsymbol{m}^{*^2}_{\text{sim}}$, $w = 1500$.

The C_i parameter also depends on the methylation through $\boldsymbol{m}^*_{\text{sim}}$. The range of possible values for C_i is c_1 to c_2. Assign $r_1, r_2, r_3,$ and r_4 to be the intervals for subjects within each quartile of $\boldsymbol{m}^*_{\text{sim}}$.

$$r_1 = [c_1, c_1 + x]$$
$$r_2 = [c_1 + kx, c_1 + kx + x]$$
$$r_3 = [c_1 + 2kx, c_1 + 2kx + x]$$
$$r_4 = [c_1 + 3kx, c_1 + 3kx + x]$$

Here, x is the interval length and k controls the amount of overlap between each interval. It can be shown that $x = (c_2 - c_1)/(1 + 3k)$ and the percentage of overlap between the two intervals is $1 - k$. In our simulations, $c_1 = 15$ and $c_2 = 25$. As an example, $k = 0.5$ implies that the percentage of overlap is 0.5. Since the length of the full range for C_i is 10, this means that the interval lengths are $x = 10/2.5$. The ranges in this case are

$$r_1 = \big[15, 19\big] \quad r_2 = \big[17, 21\big] \quad r_3 = \big[19, 23\big] \quad r_4 = \big[21, 25\big].$$

As can be seen, there is some overlap between the four ranges. If $k = 0$, there is only one range and C_i is sampled between 15 and 25 for all subjects, regardless of their quartile. If $k = 1$, then there are 4 distinct ranges, one for each of the quartiles. The intuition behind this is that the stronger the impact of $\boldsymbol{m}^*_{\text{sim}}$, the more separation there would be between the maximum values of subjects with different $m^*_{i,\text{sim}}$ values. Once the ranges are set, a C_i is sampled for each

subject in the following manner:

$$C_i = \begin{cases} \mathrm{U}(r_1), & \text{if } m^*_{i,\text{sim}} \leq m^*_{q_2,\text{sim}} \\ \mathrm{U}(r_2), & \text{if } m^*_{q_2,\text{sim}} < m^*_{i,\text{sim}} \leq m^*_{q_3,\text{sim}} \\ \mathrm{U}(r_3), & \text{if } m^*_{q_3,\text{sim}} < m^*_{i,\text{sim}} \leq m^*_{q_4,\text{sim}} \\ \mathrm{U}(r_4), & \text{if } m^*_{i,\text{sim}} > m^*_{q_4,\text{sim}} \end{cases}$$

where $\mathrm{U}(r)$ is a uniform sample from the given range and $m^*_{q,\text{sim}}$ represents the q^{th} quantile of $\boldsymbol{m}^*_{\text{sim}}$.

Once the full growth curves were generated, a sparse response matrix was constructed by randomly selecting 5-20 data points to keep for each child's growth curve. Additionally, three covariates were generated as $x_1 \sim \text{Binomial}(1, 0.5)$, $x_2 \sim \text{Binomial}(1, 0.1)$, and $x_3 \sim \text{Binomial}(1, 0.85)$. The covariates were generated from binomial distributions to match the binary covariates in the real data analysis.

3.5.2 Simulation Results

The Type I error simulation results are presented in Table 3.2. For all kernels and nominal levels of α_0, the empirical Type I errors were around α_0. This shows that the proposed method is fairly conservative, with a low rate of false positives.

Table 3.2: Type I Error Results.

α_0-level \ Kernel	Linear	Quadratic	Gaussian
0.01	0.006	0.008	0.006
0.05	0.056	0.049	0.044
0.10	0.101	0.108	0.097

The power simulation results are presented in Figure 3.4. The power was evaluated at $k = (0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1)$ for all kernels. The linear and Gaussian

kernels perform similarly regardless of $\boldsymbol{m}^{*}_{\text{sim}}$. The quadratic kernel is more powerful for $\boldsymbol{m}^{*2}_{\text{sim}}$ than $\boldsymbol{m}^{*1}_{\text{sim}}$. Recall that $\boldsymbol{m}^{1^{*}}_{\text{sim}}$ represents a linear relationship between the methylation and the response curve, while $\boldsymbol{m}^{2^{*}}_{\text{sim}}$ represents a quadratic relationship. The power results show that the quadratic kernel does indeed do a better job at picking up the quadratic relationship than the linear relationship. It is currently unclear why the linear and Gaussian kernels pick up the linear and quadratic relationships equally.

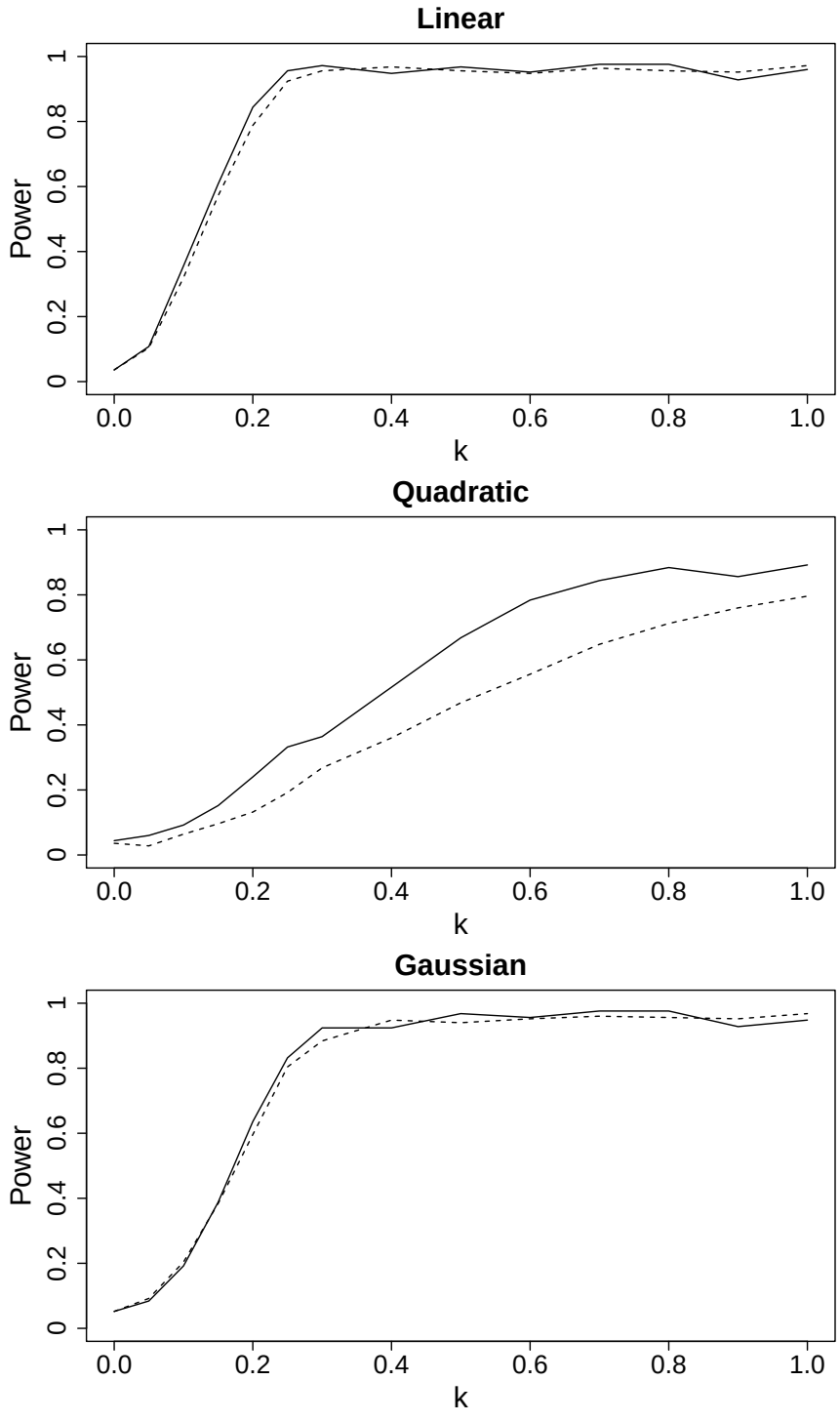


Figure 3.4: Power curves for each combination of covariance kernel and methylation summary setting. The dotted line represents $\boldsymbol{m}_{\text{sim}}^{1^*}$, the linear relationship, while the solid line represents $\boldsymbol{m}_{\text{sim}}^{2^*}$, the quadratic relationship.

3.6 Discussion

In this chapter, we introduced a novel, two-step procedure for testing the joint effect of a set of related, and possibly correlated, covariates on irregularly spaced functional response data. This method reformulates a functional mixed model as a generalized mixed model. The functional terms are expanded using the eigenfunctions from an FPCA performed on the $Y(\cdot)$'s. The model is then fit and tested using a combination of Gaussian process regression and standard mixed model variance component testing.

We successfully performed data analysis on the motivating NEST data using the proposed method. The results indicate that there is an association between PEG3 and the growth trajectories of females. Future work involves further investigating the nature of this association through the $h(\boldsymbol{m}, t)$ functions. We also showed that the proposed model has appropriate Type I error across the three kernel choices (linear, quadratic, Gaussian) through simulation studies. Additional simulation studies also showed that the method has desirable power curves.

The random effects in the model can be divided into three categories: random effects involving the mean function ($\mathbf{u}_\mu$), independent subject-specific random effects ($\mathbf{u}_\alpha$), and subject-specific random effects determined by the covariate vector of interest ($\mathbf{u}_h$). Although the main focus was on testing the variance components associated with $\mathbf{u}_h$, the independent subject-specific random effects across time are a vital inclusion in the model. In addition to negatively affecting the predictions, failure to include the independent subject-specific random effects also caused problems with the Type I error. A separate simulation study was performed to assess the Type I error for the model without the inclusion of the subject-specific random effects. The set up was exactly the same as the initial Type I error simulation study with the only difference being the removal of the subject-specific random effects in the model. This model resulted in extremely inflated empirical Type I error for all kernels and nominal levels of α_0, as seen in Table 3.3. To avoid these problems, we recommend always including the $\mathbf{u}_\alpha$ terms in the model.

This chapter addresses the first clinical question in our motivating example by developing a method to test the joint effect of a set of related covariates (methylation profiles) on irregularly

Table 3.3: Type I Error results without the subject-specific random effects.

α_0-level \ Kernel	Linear	Quadratic	Gaussian
0.01	0.540	0.549	0.962
0.05	0.680	0.681	0.965
0.10	0.761	0.760	0.970

spaced functional response data (growth curve data). However, it does not determine exactly how the growth trajectory is affected. In Chapter 5, we address this by determining what it means when a DMR is associated with the growth trajectory. Specifically, we investigate how the growth trajectory changes based on the DMR values.

Chapter 4

Fast Variance Component Testing in Functional Mixed Models

4.1 Introduction

In Chapter 3, a method was developed to test the joint effect of a vector of related, and possibly correlated, covariates of interest on a functional response curve. The method reformulates a functional mixed model as a generalized mixed model using basis expansions. The functional mean is expanded using a truncated power basis spline and the remaining functional effects are expanded using the eigenfunctions from an FPCA performed on the functional responses. Through a combinaton of Gaussian process regression (GP) and standard mixed model techniques, an association with the response is ultimately determined by testing the significance of a subset of variance components in the model. We refer to this method as VarCompGP. In VarCompGP, the mixed model is fit using a profiled restricted maximum likelihood (PREML) approach based on a set of residual contrasts. A linear score test proposed by Qu et al. (2013) is used to test the variance components of interest. Both the model fitting and variance component testing in VarCompGP is implemented through the `varComp` package (Qu, 2017).

VarCompGP performed well in simulation studies, showing proper Type I error and ade-

quate power curves. It was also used to analyze the NEST data. The data analysis focused on determining if there was an association between childrens' growth curves (over 1500 days) and the Paternally expressed gene 3 (PEG3) methylation profile. However, during the course of the development of VarCompGP, there was a problem with the `varComp` `R` package that was utilized in the method. Upon an update to a new version of `R`, there was a compatibility issue between the `varComp` package and the newer version of `R`. This caused the code that implemented VarCompGP to stop working, which obviously proved to be problematic. Eventually, it was determined that the problem was with the function that supplied the starting values to the optimizer that estimated the variance components. The solution was to manually set the starting values to be a vector of zeros. However, before this solution was found, many attempts were made to find a different way to estimate the functional mixed model discussed in Chapter 3. The model presented in this chapter is a result of these attempts.

A fortunate result of this new model was improved computation time. Although not directly discussed, one drawback to the VarCompGP approach was computation time, which was further increased by manually setting the starting values to zero. Though not a problem for simple data analyses, this could be a potential problem in studies that wish to screen a large number of DMRs. In this chapter, we aim to address the timing issue seen in the VarCompGP method. As stated, we do this by using a different model fitting approach and making some alterations to the linear score test, both of which reduce computation time.

4.2 Review of the Model

As in Chapter 3, $Y_i(t)$ denotes the functional response of the i^{th} subject at some point t in a continuous interval, $t \in [0, T]$ for $i = 1, \ldots, N$. The vector of related covariates of interest is denoted as $\boldsymbol{m}_i = (m_{i1}, \ldots, m_{iP})^T$ and $\boldsymbol{M} = (\boldsymbol{m}_1, \ldots, \boldsymbol{m}_N)^T$ is the full matrix of related covariates for all subjects, where each row is a subject and each column is a covariate. The secondary covariates are denoted as $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iC})^T$ and can include demographic or clinical

information. Then we have the functional model

$$Y_i(t) = \mu(t) + \sum_{c=1}^{C} x_{ic}\beta_c(t) + \alpha_i(t) + h(\boldsymbol{m}_i, t) + \epsilon_i(t) \tag{4.1}$$

where $\mu(\cdot)$ is the overall mean trajectory, x_{ic} is the value of the c^{th} baseline covariate for subject i, $\beta_c(\cdot)$ are the unknown coefficient functions, and $\epsilon_i(\cdot)$ is a random error process with mean zero and finite variance. Both $\alpha_i(t)$ and $h(\boldsymbol{m}_i, t)$ follow time dependent, latent processes that are subject specific and can be thought of as functional random components.

After the expansions metioned in Section 4.1 and detailed in Chapter 3, the functional model can be rewritten as a mixed model, where for each indivudal, the model to be fit at a specific timepoint t is

$$\begin{aligned} Y_i(t) = \mu_0 + \mu_1 t + \sum_{k=1}^{K} [\mu_{k+1}(t-\kappa_k)_+] \\ + \sum_{c=1}^{C} [x_{ic}\{\hat{\phi}_1(t)\beta_{1c} + \cdots + \hat{\phi}_B(t)\beta_{Bc}\}] \\ + \sum_{b=1}^{B} [\hat{\phi}_b(t)\alpha_{bi} + \hat{\phi}_b(t)h_b(\boldsymbol{m}_i)] + \epsilon_i(t). \end{aligned} \tag{4.2}$$

In (4.2), μ_0, μ_1, and $\beta_{1c}, \ldots, \beta_{Bc}$, $c = 1, \ldots, C$, are modeled as fixed effects. The remaining parameters, $\mu_2, \ldots, \mu_K$, $\alpha_{1i}, \ldots, \alpha_{Bi}$, and $h_1(\cdot), \ldots, h_B(\cdot)$ are modeled as random effects. The random effects vector associated with the mean, $\mathbf{u}_\mu = (\mu_2, \ldots, \mu_K)^T$, is distributed as $\mathbf{u}_\mu \sim \text{N}(\mathbf{0}, \sigma_\mu^2 \boldsymbol{I})$. For $b = 1, \ldots, B$, each $\boldsymbol{\alpha}_b = (\alpha_{b1}, \ldots, \alpha_{bN})^T$ follows a normal distribution, $\boldsymbol{\alpha}_b \sim \text{N}(\mathbf{0}, \delta_b \boldsymbol{I})$. We assume $\boldsymbol{h}_b(\cdot)$ follows a Gaussian process, such that $\boldsymbol{h}_b(\boldsymbol{M}) \sim \text{N}(\mathbf{0}, \gamma_b \boldsymbol{R})$. The covariance matrix, $\boldsymbol{R}$, depends on the covariate vectors of interest through a specified kernel matrix (linear, quadratic, or Gaussian) and γ_b is an unknown variance component.

4.2.1 Matrix Notation

For the general case where it is possible for each individual to have a different number of observations, the full model is written in matrix notation as

$$\mathbf{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}_\mu \mathbf{u}_\mu + \boldsymbol{Z}_\alpha \mathbf{u}_\alpha + \boldsymbol{Z}_h \mathbf{u}_h + \boldsymbol{\epsilon}, \tag{4.3}$$

where $\tilde{n} = n_1 + n_2 + \cdots + n_N$, $\mathbf{Y}_{\tilde{n}\times 1} = (\mathbf{Y}^T_{1_{n_1\times 1}}, \mathbf{Y}^T_{2_{n_2\times 1}}, \ldots, \mathbf{Y}^T_{N_{n_N\times 1}})^T$ is the stacked response vector and $\boldsymbol{\beta} = (\mu_0, \mu_1, \beta_{11}, \ldots, \beta_{B1}, \ldots, \beta_{1C}, \ldots, \beta_{BC})^T$ is the vector of fixed effects. The vector of random effects is $\mathbf{u} = (\mathbf{u}^T_\mu, \mathbf{u}^T_\alpha, \mathbf{u}^T_h)^T$, where $\mathbf{u}_\mu = (\mu_1, \ldots, \mu_K)^T$, $\mathbf{u}_\alpha = (\boldsymbol{\alpha}^T_1, \ldots, \boldsymbol{\alpha}^T_B)^T$ and $\mathbf{u}_h = \{\boldsymbol{h}_1(\boldsymbol{M})^T, \ldots, \boldsymbol{h}_B(\boldsymbol{M})^T\}^T$. Recall that B is the number of eigenfunction vectors used in the expansion of the functional model that are selected based on the percent of variance explained (PVE) in the FPCA, such that PVE $\geq 95\%$. The associated design matrix for the covariates is $\boldsymbol{X} = [\tilde{\boldsymbol{X}}_0\ \tilde{\boldsymbol{X}}_1\ \cdots\ \tilde{\boldsymbol{X}}_C]$. Additionally, $\boldsymbol{Z}_\mu$, $\boldsymbol{Z}_\alpha$, and $\boldsymbol{Z}_h$ are known matrices associated with the random effects vectors. The full random effects vector is distributed as

$$(\mathbf{u}^T_\mu, \mathbf{u}^T_\alpha, \mathbf{u}^T_h, \boldsymbol{\epsilon}^T)^T \sim \mathrm{N}\{\mathbf{0}, bdiag(\mathbf{G}_\mu, \mathbf{G}_\alpha, \mathbf{G}_h, \sigma^2_\epsilon \boldsymbol{I}_{\tilde{n}})\} \tag{4.4}$$

where $bdiag(\cdot)$ denotes a block diagonal matrix of the listed matrices. In (4.4), $\mathbf{G}_\mu = \sigma^2_\mu \boldsymbol{I}_K$, $\mathbf{G}_\alpha = \mathrm{diag}(\boldsymbol{\delta}) \otimes \boldsymbol{I}_N$, $\mathbf{G}_h = \mathrm{diag}(\boldsymbol{\gamma}) \otimes \boldsymbol{R}$, $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_b)^T$ and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_B)^T$. In VarCompGP, the variance components $(\sigma^2_\mu, \delta_1, \ldots, \delta_B)^T$ are treated as nuisance variance parameters and the hypothesis test is performed on the variance components of interest is

$$\mathrm{H}_0 : \boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_B)^T = \mathbf{0} \text{ vs. } \mathrm{H}_1 : \boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_B)^T \geq \mathbf{0}.$$

4.2.2 Connection Between Linear Mixed Models and Penalized Regression

There are many references that draw the connection between mixed models and penalized regression models. Here, we refer to a simple example in Wand (2003). In the example, the true

function is $y_i = f(x_i) + 0.4\epsilon_i$. The linear model

$$y_i = \beta_0 + \beta_1 x_i + \sum_{k=1}^{K} u_k (x_i - \kappa_k)_+ + \epsilon_i$$

is introduced, where $(x - \kappa_k)_+ = (x - \kappa_k)$ if $x > \kappa_k$ and is 0 otherwise and $u_k \overset{iid}{\sim} \mathrm{N}(0, \sigma_u^2)$. Defining the design matrices as $\boldsymbol{X} = [1\ x_i]_{1 \le i \le n}$ and $\boldsymbol{Z} = [(x_i - \kappa_k)_+]_{1 \le i \le n,\ 1 \le k \le K}$ and the effect vectors as $\boldsymbol{\beta} = [\beta_0, \beta_1]^T$ and $\mathbf{u} = [u_1, \ldots, u_K]^T$, then the linear mixed model is written in matrix notation as

$$\mathbf{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\mathbf{u} + \boldsymbol{\epsilon},\ \begin{pmatrix} \mathbf{u} \\ \boldsymbol{\epsilon} \end{pmatrix} \sim \mathrm{N}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \sigma_u^2 \boldsymbol{I} & \mathbf{0} \\ \mathbf{0} & \sigma_\epsilon^2 \boldsymbol{I} \end{bmatrix} \right).$$

This type of mixed model can be solved using restricted maximum likelihood (REML) to estimate $\boldsymbol{\beta}, \sigma_u^2$ and σ_ϵ^2 and best linear unbiased prediction (BLUP) to estimate $\mathbf{u}$. This is similar to the approach used to solve the model in VarCompGP. As in Wand (2003), this approach is equivalent to solving the penalized least squares regression model

$$\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} = \underset{\boldsymbol{\beta}, \mathbf{u}}{\operatorname{argmin}} (||\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\mathbf{u}||^2 + \theta ||\mathbf{u}||^2),$$

where θ is a smoothing parameter and for any vector $\boldsymbol{x}$, $||\boldsymbol{x}||^2 = \sqrt{\boldsymbol{x}^T \boldsymbol{x}}$. In the penalized regression framework, both $\boldsymbol{\beta}$ and $\mathbf{u}$ are estimated as fixed effects. To obtain the variance component from the penalized regression model, one simply computes $\sigma_u^2 = \sigma_\epsilon^2 / \theta$.

4.2.3 Penalized Regression Framework

To speed up computation time, we used a penalized regression model to estimate the model parameters in (4.3). However, we must first reparameterize so that the covariance structures fit into the penalized regression framework. To be in the proper form, each block of the block diagonal full covariance matrix must be a variance component multiplied by the identity matrix.

For simplicity, assume that $B = C = 1$. This means that $\boldsymbol{\beta} = (\mu_0, \mu_1, \beta_1)^T$, $\mathbf{u}_\alpha = \boldsymbol{\alpha}$ where $\boldsymbol{\alpha} \sim \mathrm{N}(\mathbf{0}, \delta \boldsymbol{I}_N)$, and $\mathbf{u}_h = \boldsymbol{h}(\boldsymbol{M})$ where $\boldsymbol{h}(\boldsymbol{M}) \sim \mathrm{N}(\mathbf{0}, \gamma \boldsymbol{R})$. We assume that there exists some matrix $\boldsymbol{D}$ such that $\boldsymbol{R} = \boldsymbol{D}\boldsymbol{D}^T$. Additionally, let $\mathbf{v}_h$ be some random effects vector such that $\mathbf{v}_h \sim \mathrm{N}(\mathbf{0}, \gamma \boldsymbol{I}_N)$. Then, $\boldsymbol{h}(\boldsymbol{M}) \equiv \boldsymbol{D}\mathbf{v}_h \sim \mathrm{N}(\mathbf{0}, \gamma \boldsymbol{D}\boldsymbol{D}^T) \equiv \mathrm{N}(\mathbf{0}, \gamma \boldsymbol{R})$. We can rewrite (4.3) as,

$$\mathbf{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}_\mu \mathbf{u}_\mu + \boldsymbol{Z}_\alpha \boldsymbol{\alpha} + \boldsymbol{Z}_h^* \mathbf{v}_h + \boldsymbol{\epsilon}$$

where $\boldsymbol{Z}_h^* = \boldsymbol{Z}_h \boldsymbol{D}$ and the full random effects vector is

$$(\mathbf{u}_\mu^T, \boldsymbol{\alpha}^T, \mathbf{v}_h^T, \boldsymbol{\epsilon}^T)^T \sim \mathrm{N}\{\mathbf{0}, bdiag(\sigma_\mu^2 \boldsymbol{I}_K, \delta \boldsymbol{I}_N, \gamma \boldsymbol{I}_N, \sigma_\epsilon^2 \boldsymbol{I}_{\tilde{n}})\}.$$

Once we have the correct form, we can then obtain estimates for the vector $(\boldsymbol{\beta}^T, \mathbf{u}_\mu^T, \boldsymbol{\alpha}^T, \mathbf{v}_h^T)^T$ by solving the penalized regression problem

$$\underset{\boldsymbol{\beta}, \mathbf{u}_\mu, \boldsymbol{\alpha}, \mathbf{v}_h}{\operatorname{argmin}} \{||\mathbf{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}_\mu \mathbf{u}_\mu - \boldsymbol{Z}_\alpha \boldsymbol{\alpha} - \boldsymbol{Z}_h^* \mathbf{v}_h)||^2 + \theta_\mu ||\mathbf{u}_\mu||^2 + \theta_\alpha ||\boldsymbol{\alpha}||^2 + \theta_h ||\mathbf{v}_h||^2)\}. \quad (4.5)$$

We can also obtain estimates of the variance components through the smoothing parameters as

$$\begin{pmatrix} \hat{\sigma}_\mu^2 \\ \hat{\delta} \\ \hat{\gamma} \end{pmatrix} = \begin{pmatrix} \hat{\sigma}_\epsilon^2 / \hat{\theta}_\mu \\ \hat{\sigma}_\epsilon^2 / \hat{\theta}_\alpha \\ \hat{\sigma}_\epsilon^2 / \hat{\theta}_h \end{pmatrix}.$$

This process is easily generalized to the cases where $B > 1$ and/or $C > 1$. To solve the penalized regression problem in (4.5), we use the method proposed in Wood (2011). This is implemented through the `gam` function in the `mgcv R` package, with REML as the smoothing parameter estimation method.

4.3 Linear Score Test for Variance Components

Although the model fitting procedure is different, we utilize the same type of linear score test for variance component testing as in Chapter 3. The method proposed in this chapter makes some changes that further reduce computation time. The differences between the linear score test implemented in VarCompGP and the test implemented in the proposed method are as follows. First, we utilize a true score test, in the sense that for testing purposes, we fit only the null model,

$$\underset{\boldsymbol{\beta},\mathbf{u}_\mu,\boldsymbol{\alpha}}{\text{argmin}}\{||\mathbf{Y}-\boldsymbol{X\beta}-\boldsymbol{Z}_\mu\mathbf{u}_\mu-\boldsymbol{Z}_\alpha\boldsymbol{\alpha})||^2+\theta_\mu||\mathbf{u}_\mu||^2+\theta_\alpha||\boldsymbol{\alpha}||^2\}.$$

In the `varComp` package used in VarCompGP, the default procedure is to fit both the null and alternative models when testing. This takes time and is unecessary, especially in scanning scenarios where there are only a few significant DMRs that we would need to further investigate and estimate variance parameters for. Additionally, we use simpler weights when constructing the test statistic, $U(\hat{\boldsymbol{\tau}}_0)=\sum_{i=I+1}^{I+1+J} w_i(\hat{\boldsymbol{\tau}}_0)S_i(\hat{\boldsymbol{\tau}}_0)$. In VarCompGP, the default weights are proportional to the inverse standard deviations, obtained from the diagonal of the efficient information matrix. The proposed method in this chapter uses equal weights for each variance component of interest, such that $w_i = 1/B$ for all i.

4.4 Results Comparison from Data Application

We apply the new method to the NEST data and compare the results and computation time to the VarCompGP method. We present the data from PEG3 as was done in the previous chapter. Tables 4.1 and 4.2 show the variance component estimates, p-values, and computation time (in seconds) for each available kernel, for both the VarCompGP (VC) method and the new proposed method, GamGP (GAM). It can be seen that the computation time for GamGP is significantly less than the computation time for VarCompGP. The VarCompGP method takes

approximately 7 to 15 times longer than the GamGP method, depending on the size of the model (ie, the value of B) and the kernel. All of the male models take longer than the female models because $B = 3$ for the male models and $B = 2$ for the female models.

Table 4.1: Female PEG3 Results Comparison.

	Linear		Quadratic		Gaussian	
	VC	GAM	VC	GAM	VC	GAM
$\hat{\sigma}_\mu$	25.9328	6e-6	28.2687	7e-6	24.5411	6e-6
$\hat{\delta}_1$	2.1965	2.9178	2.2036	2.9273	2.2964	3.0505
$\hat{\delta}_2$	0.0528	0.0582	0.0447	0.0494	0.0469	0.0518
$\hat{\gamma}_1$	0.0596	0.0089	0.0878	0.0002	0.0000	0.0002
$\hat{\gamma}_2$	0.0038	0.0005	0.0071	0.0000	0.0131	0.0145
p-value	0.0703	0.0714	0.0132	0.0097	0.4932	0.2704
$\hat{\sigma}^2_\epsilon$	0.2081	0.2081	0.2076	0.2076	0.2081	0.2081
Time (sec)	246.3400	16.8600	130.8300	17.7900	205.8100	15.8000

Table 4.2: Male PEG3 Results Comparison.

	Linear		Quadratic		Gaussian	
	VC	GAM	VC	GAM	VC	GAM
$\hat{\sigma}_\mu$	42.5552	1.1e-5	42.3814	1.1e-5	42.3814	1.1e-5
$\hat{\delta}_1$	1.5348	2.1174	1.5445	2.1308	1.5445	2.1308
$\hat{\delta}_2$	0.0922	0.0999	0.0922	0.0998	0.0922	0.0998
$\hat{\delta}_3$	0.1488	0.0945	0.1491	0.0947	0.1491	0.0947
$\hat{\gamma}_1$	0.0108	0.0012	0.0000	0.0000	0.0000	0.0000
$\hat{\gamma}_2$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
$\hat{\gamma}_3$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Pvalue	0.7101	0.9650	0.9487	0.9962	0.9769	0.9491
$\hat{\sigma}^2_\epsilon$	0.1383	0.1383	0.1383	0.1383	0.1383	0.1383
Time (sec)	428.8300	56.4100	417.7000	51.8400	407.0300	50.8200

It should be noted that both methods produce adequate and sensible models, determined through residual and predicted value plots. Despite this, there is a difference between the variance component estimates of the two methods, particularly that of $\hat{\sigma}^2_\mu$. It has been determined

that the difference in variance component estimates is caused by the fact that the `varComp` package utilized in VarCompGP normalizes the individual variance-covariance matrices so that the variance components are on the same scale. If one sets the `normalizeTrace` option in `varComp` to `FALSE`, the variance component estimates are equivalent to those produced from `gam`. An option to normalize the trace in the GamGP method is also available. However, this does add some computation time and therefore is not the default setting.

While the p-values for the respective kernels are similar across methods, they are not exactly the same. These differences are due to the weights that are used in constructing the test statistic, as described in Section 4.3. Setting the `LinScore.wt` option in `varCompTest.control` to `'EqWt'` produces p-values that are essentially equivalent to those produced from GamGP.

4.5 Simulations

We performed numerous simulation studies to compare GamGP to VarCompGP, as well as to ensure that GamGP had appropriate Type I error and power. In all studies, the simulated data is modeled after the NEST data, exactly as was done in Chapter 3. Thus, for all iterations, the sample size was fixed at $N = 44$ and the range of t was $t \in [0, 1500]$. There are three settings corresponding to the three kernel choices (linear, quadratic, and Gaussian). Additionally, there are two summary settings for the methylation information that is used to generate parameter values for the simulated data; the mean of the CpG sites and the square of the sums. Recall that the strength of the relationship between the response curve and the methylation information is determined by $k \in [0, 1]$. When $k = 0$, the methylation has no effect on the response curve and this is the null model. As k gets larger, the methylation vectors have a greater effect on the shape of the response curve.

The first simulation study assessed the Type I error and power of GamGP and compared it to VarCompGP. We use 1000 replications for the Type I error analysis, with $\alpha_0 = 0.01$, 0.05, and 0.10. Since this is evaluated for the null model, $k = 0$. For the power analysis, power was evaluated at $k = (0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1)$. We use

250 replications for each combination of settings at each value of k, with $\alpha_0 = 0.05$.

The second simulation study compared computation time and analysis results between GamGP and VarCompGP. As in the power analysis, there were 250 replications for each combination of kernel, summary setting, and method for a total of 12 different settings. Within each setting, k was randomly generated from a Uniform$(0, 1)$ distribution for each iteration. This was to ensure that there were a variety of signal strengths and therefore, a variety of resulting p-values.

4.5.1 Type I Error and Power Results

The empirical Type I error was found to be almost identical to that of VarCompGP, as shown in Table 4.3. This shows that GamGP behaves the same as VarCompGP under the null model.

Table 4.3: Comparison of Type I error results between VarCompGP (VC) and GamGP (GAM).

	Linear		Quadratic		Gaussian	
α_0	VC	GAM	VC	GAM	VC	GAM
0.01	0.006	0.007	0.008	0.009	0.006	0.006
0.05	0.056	0.057	0.049	0.046	0.044	0.046
0.10	0.101	0.100	0.108	0.113	0.097	0.104

The power simulation results from VarCompGP and GamGP are compared in Figure 4.1. GamGP performs similarly to VarCompGP, in the sense that the linear and Gaussian kernels perform similarly regardless of the summary setting, $\boldsymbol{m}^*_{\text{sim}}$, while the quadratic kernel is more powerful for the true quadratic relationship ($\boldsymbol{m}^{*^2}_{\text{sim}}$) than the true linear relationship ($\boldsymbol{m}^{*1}_{\text{sim}}$). However, it does appear that the GamGP method is slightly less powerful than VarCompGP, particularly for the linear and Gaussian kernels. Although the power curves eventually approach one, this happens more gradually. This indicates that GamGP does not have as much power to detect small to medium signals as VarCompGP. This is due to the differences in weights when constructing the test statistic, as discussed in Section 4.3. It appears that the simpler weights

help to speed up computation time, but slightly reduce the power of the test in GamGP. This conclusion was verified by performing an additional power simulation for the VarCompGP model, but with equal weights. The power curves in this simulation are identical to the GamGP power curves. These results are shown in Appendix B.1.

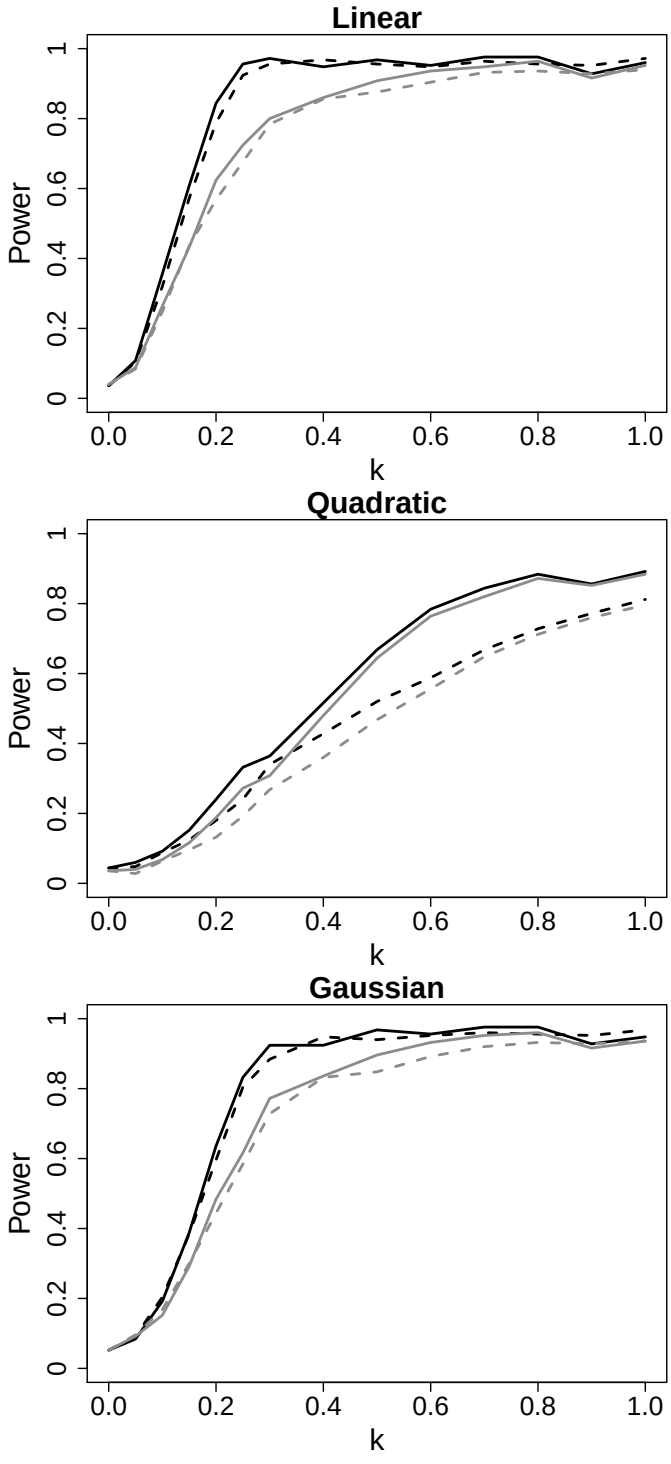


Figure 4.1: Power curves for each combination of covariance kernel, methylation summary setting, and method. The dotted lines represent $\boldsymbol{m}_{\text{sim}}^{1^*}$, the linear relationship, while the solid lines represent $\boldsymbol{m}_{\text{sim}}^{2^*}$, the quadratic relationship. The black lines represent the VarCompGP method, while the gray lines represent the GamGP method.

4.5.2 Timing Results

The main goal of this chapter is to introduce a comparable method to VarCompGP that takes significantly less computation time. While this was true in the real data analysis, a timing simulation study was performed to get a better idea of the differences in computation time across methods. As described in Section 4.5, there were a total of 12 different settings for the timing simulations, based on kernel, summary setting, and method, with 250 iterations in each setting. After the simulations were run, the results were split further according to the value of B that was selected for the simulated data. This is because the size of the model is proportional to B and therefore the methods take longer as B gets larger.

We found that differences in timing were due mainly to the value of B and the method used, while the kernel and summary settings did not appear to matter. Therefore, the timing results were aggregated over kernel and summary settings. The pre-summarized results are presented in Appendix B.2. The aggregated timing results are presented in Figure 4.2. As expected, the computation time for each method increases as B increases. It is also clear that GamGP significantly out performs VarCompGP across the board. In fact, the average computation time for GamGP when $B = 3$ is less than the average computation for VarCompGP when $B = 1$. There is also much less variability in computation time for the GamGP method.

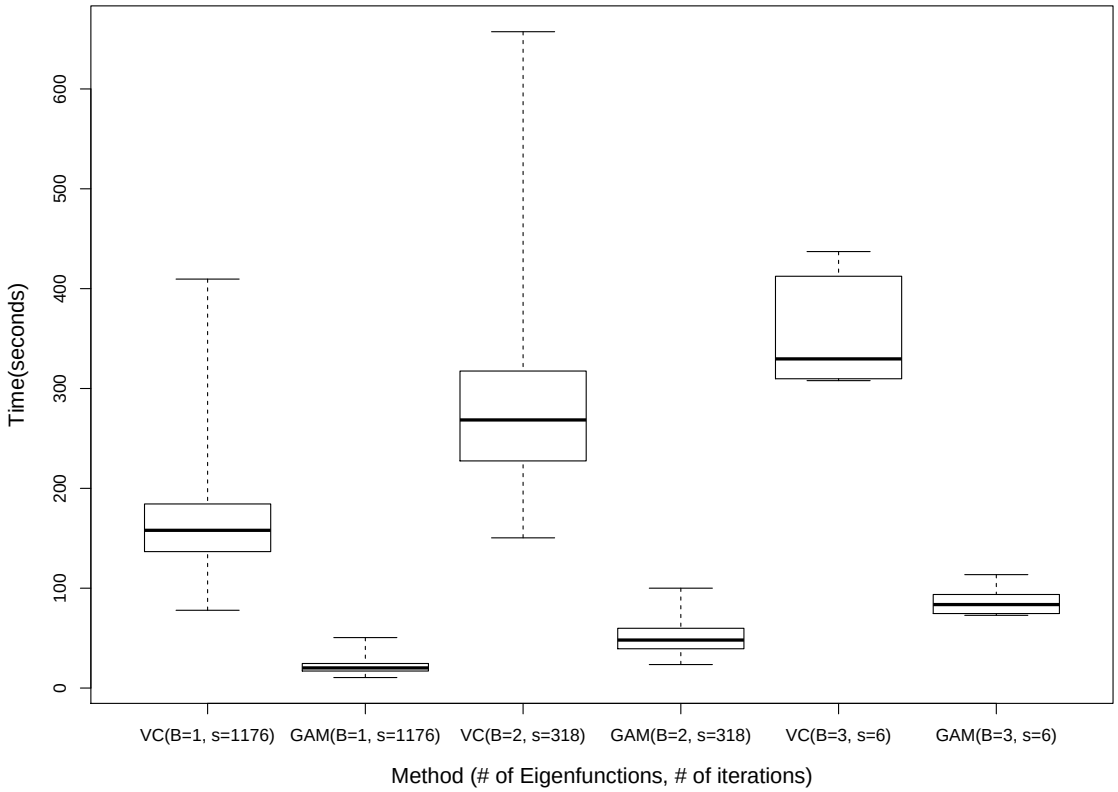


Figure 4.2: Computation Times.

To further investigate the speed up gained by using GamGP instead of VarCompGP, we looked at how much faster GamGP was compared to VarCompGP for each iteration that, other than method, had the same settings. These results were then combined across kernel and summary setting and compared for values of B. This information is presented in Figure 4.3. For one simulated data set, VarCompGP takes on average about 8 times longer to run when $B = 1$, about 6 times longer when $B = 2$, and about 4.5 times longer when $B = 3$.

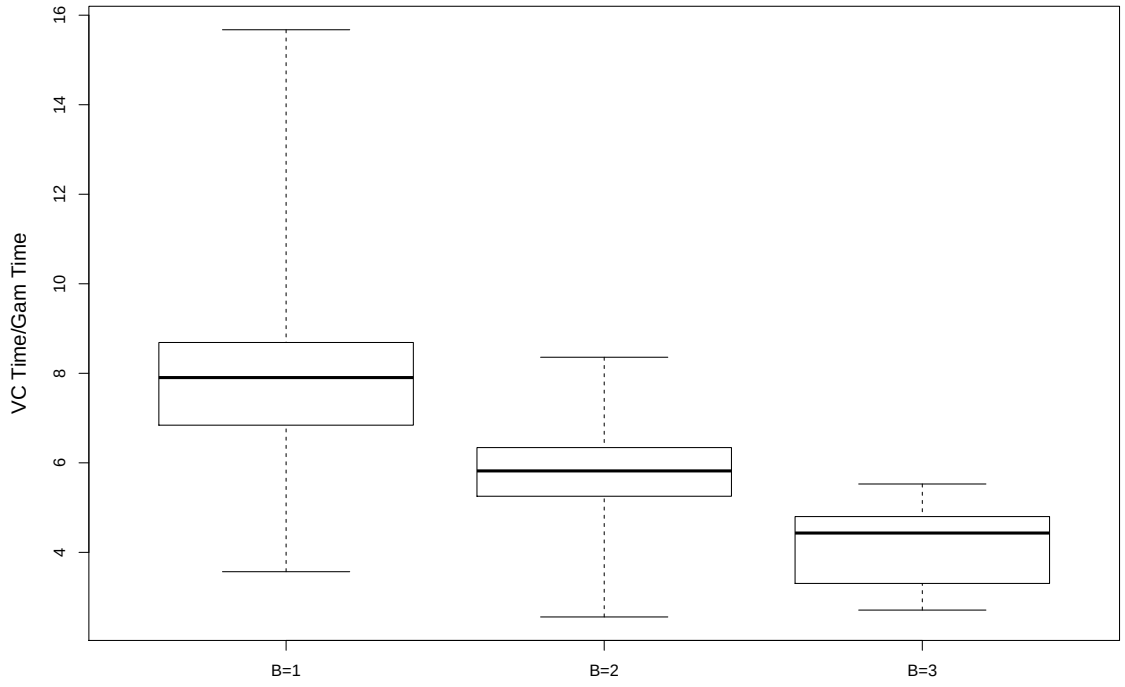


Figure 4.3: Speed up gained from using GamGP vs. VarCompGP on the same set of simulated data.

4.5.3 Disagreement in P-values and Rejection Status

Reducing computation time is not useful if the new method is not comparable to VarCompGP in terms of testing results. Therefore, the testing results from the timing simulations in Section 4.5.2 were examined to see how the methods compared to each other within each combination of kernel, summary setting, and value of B. First, we looked at what will be referred to as the percentage of disagreement (PoD) between the two methods. In other words, we did not care about the actual value of the p-value, only the result of the hypothesis test at $\alpha_0 = 0.05$. Within each setting, the percentage of iterations in which the tests did not agree on whether to reject or not reject was calculated. When $B = 1$, VarCompGP and GamGP agree almost completely for every setting, with the highest PoD at 0.5%. It was determined that this is because when

$B = 1$, there is only one variance component of interest being tested so all weighting choices are equivalent as $w = 1$. Therefore, the fact that different weighting schemes are used in the two methods does not matter in this scenario. Since there was only one iteration per setting for $B = 3$, the results were combined to create a category of $B > 1$. When $B > 1$, the PoD is about 25% for the quadratic and Gaussian kernels, regardless of summary setting (exact values were 22.03%, 28.1%, 26.53%, 26.53%). The worst PoD is found in the linear kernel when $B > 1$. The percentages across summary settings are 40.68% and 46.94%. It is unclear why the percentage of disagreement is almost double for the linear kernel and that is a potential topic for future investigation.

To determine the cases in which the two methods disagreed on rejection status, the exact values of the p-values were also compared. Plots of the p-values for each combination of kernel and summary statistic are seen in Figure 4.4. As expected, when $B = 1$ (ie, when the weights are the same), the p-values for GamGP and VarCompGP fall directly on the diagonal line, indicating that they match exactly. Each plot can be split in to four quadrants. The small square at the bottom is the scenario in which both methods reject the null hypothesis, whereas the larger square at the top is the scenario in which both methods fail to reject the null. The bottom right rectangle represents the scenario where VarCompGP rejects, but GamGP does not. In all settings, especially for the linear kernel, there is a fairly significant cluster of points here. The upper left rectangle, where GamGP rejects, but VarCompGP does not, contains almost no points. Therefore, whenever the methods disagree on rejection status, it is because VarCompGP is detecting a signal that GamGP cannot. This reflects the lower power of GamGP that was found in Section 4.5.1.

4.6 Discussion

In this chapter, we introduced a method that is comparable to VarCompGP in terms of Type I error, power, and testing results, but is much faster in terms of computation time. The new method is based on a model fitting procedure that utilizes the connection between mixed models

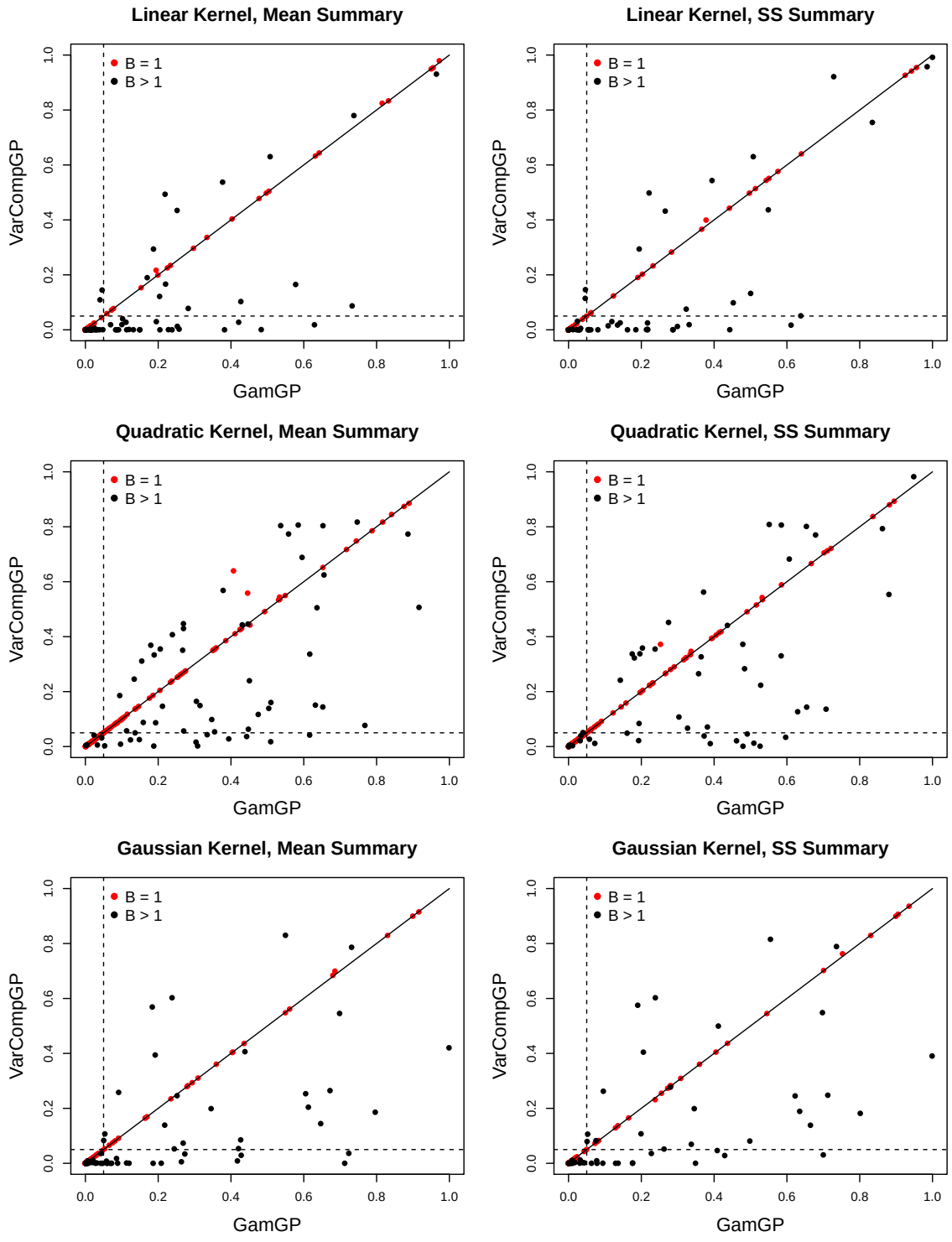


Figure 4.4: Comparing p-values between GamGP and VarCompGP for same set of simulated data.

and penalized regression. The revised fitting prodecure, as well as some changes to the testing procedure, significantly reduce computation time. The reduction in computation time does come with a cost. Through simulation studies, it was shown that GamGP has slightly less power, particularly when using the linear or Gaussian kernel choices. This is also expressed through the differences in p-values and testing results between VarCompGP and GamGP. However, we believe that as long as a user is aware of the trade-offs between VarCompGP and GamGP, both methods can be extremely useful in practice. GamGP is recommended for scanning procedures, where one may have numerous covariates of interest to test separately for an association with a response. In this type of scenario, computation time is important, while you may not care as much about rejecting or not rejecting, but are rather looking for a few tests that pop out as different from the rest. GamGP would also be the better choice if $B = 1$ for the given data set, since the tests are equivalent in this case. On the other hand, VarCompGP is recommended for more powerful testing of only a few separate covariate vectors of interest. Additionally, if there is prior knowledge that the signal is small, VarCompGP would be a more appropriate choice.

Chapter 5

Risk Profile Analysis and Prediction Model

5.1 Introduction

In this chapter, we look to expand on the model introduced in Chapters 3 and 4. In the previous two chapters, we developed methods to test the joint effect of a vector of related covariates on a functional response. Specifically, we tested the association between the methylation values within a specific DMR and the growth trajectories of children in the NEST data set. However, although this test determines if the methylation values within a specific DMR jointly affect the growth trajectory, it does not determine how the growth trajectory is affected. One question we wish to answer is what it means when a DMR is jointly significant to the growth trajectory. Recall the original null hypothesis,

$$\mathrm{H}_0 : h(\boldsymbol{m}_i, t) = 0, \text{ for all values of } \boldsymbol{m} \text{ and } t.$$

Rejecting the null hypothesis means that $h(\boldsymbol{m}_i, t) \neq 0$ for some values of $\boldsymbol{m}_i$ at some values of t. Therefore, there is some non-zero function that describes how the response varies across

time depending on $\boldsymbol{m}_i$. We refer to these functions as individual risk profiles. One goal of this chapter is to estimate individual risk profiles across different stages of growth based on methylation values. These risk profiles can be used to explain the contribution of $\boldsymbol{m}_i$ to an individual's full growth trajectory.

Another goal of this chapter is to use the model for predictive purposes. As discussed in Chapter 2, many current prediction models for functional data focus on linear additive effects for a limited number of predictors. Even if interaction terms are included in these types of models, they have a tendency to overfit and generally perform poorly when used to predict for new subjects. Therefore, they are not suited for the type of complex predictors that are found in the motivating example. Additional methods that allow for non-linearity, including Gaussian process functional regression, are also not suitable because they do not account for the proper covariance structure. To appropriately capture the complexity of the data, our model is nonlinear and combines standard functional methods with Gaussian process regression. Functional random effects are included in the form of subject-specific time-varying random effects and time-varying random effects dependent on the methylation profiles. These novel inclusions of functional random effects help to better capture the within-subject variability and the appopriate covariance structure of the data in general. This leads to a prediction model that is better suited to predicting the full growth curve of a subject.

Finally, the assumed distribution of the methylation dependent random effect, along with the prediction model are used to predict the risk profile and growth trajectory of a new subject based on his/her methylation profile and baseline covariates. We introduce the use of prediction variances to construct prediction bands around these curves, which is not typically done in the current functional data literature. Using the prediction variance to construct the prediction bands drastically improves the coverage compared to competing methods.

5.2 Model Framework

Recall the functional model first introduced in Chapter 3,

$$Y_i(t) = \mu(t) + \sum_{c=1}^{C} x_{ic}\beta_c(t) + \alpha_i(t) + h(\boldsymbol{m}_i, t) + \epsilon_i(t) \tag{5.1}$$

where $\mu(\cdot)$ is the overall mean trajectory, x_{ic} is the value of the c^{th} baseline covariate for subject i, $\beta_c(\cdot)$ are the unknown coefficient functions, and $\epsilon_i(\cdot)$ is a random error process with mean zero and finite variance. The vector of related covariates of interest is denoted as $\boldsymbol{m}_i = (m_{i1}, \ldots, m_{iP})^T$ and $\boldsymbol{M} = (\boldsymbol{m}_1, \ldots, \boldsymbol{m}_N)^T$ is the full matrix of related covariates for all subjects, where each row is a subject and each column is a covariate. The secondary covariates are denoted as $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iC})^T$ and can include demographic or clinical information. Both $\alpha_i(t)$ and $h(\boldsymbol{m}_i, t)$ follow time dependent, latent processes that are subject specific and can be thought of as functional random effects. The functional model can be rewritten as a mixed model, where for each individual, the model to be fit at a specific timepoint t is

$$\begin{aligned} Y_i(t) = \mu_0 + \mu_1 t &+ \sum_{k=1}^{K}[\mu_{k+1}(t-\kappa_k)_+] \\ &+ \sum_{c=1}^{C}[x_{ic}\{\hat{\phi}_1(t)\beta_{1c} + \cdots + \hat{\phi}_B(t)\beta_{Bc}\}] \\ &+ \sum_{b=1}^{B}[\hat{\phi}_b(t)\alpha_{bi} + \hat{\phi}_b(t)h_b(\boldsymbol{m}_i)] + \epsilon_i(t). \end{aligned} \tag{5.2}$$

Recall that C is the number of baseline covariates and B is the number of eigenfunction vectors used in the expansion that are selected based on the percent of variance explained (PVE) in the FPCA, such that PVE $\geq 95\%$.

For the general case where it is possible for each individual to have a different number of observations, the full model is written in matrix notation as

$$\mathbf{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}_\mu \mathbf{u}_\mu + \boldsymbol{Z}_\alpha \mathbf{u}_\alpha + \boldsymbol{Z}_h \mathbf{u}_h + \boldsymbol{\epsilon}, \tag{5.3}$$

where $\tilde{n} = n_1 + n_2 + \cdots + n_N$, $\mathbf{Y}_{\tilde{n}\times 1} = (\mathbf{Y}^T_{1_{n_1\times 1}}, \mathbf{Y}^T_{2_{n_2\times 1}}, \ldots, \mathbf{Y}^T_{N_{n_N\times 1}})^T$ is the stacked response vector and $\boldsymbol{\beta} = (\mu_0, \mu_1, \beta_{11}, \ldots, \beta_{B1}, \ldots, \beta_{1C}, \ldots, \beta_{BC})^T$ is the vector of fixed effects. The vector of random effects is $\mathbf{u} = (\mathbf{u}_\mu^T, \mathbf{u}_\alpha^T, \mathbf{u}_h^T)^T$, where $\mathbf{u}_\mu = (\mu_1, \ldots, \mu_K)^T$, $\mathbf{u}_\alpha = (\boldsymbol{\alpha}_1^T, \ldots, \boldsymbol{\alpha}_B^T)^T$ and $\mathbf{u}_h = \{\boldsymbol{h}_1(\boldsymbol{M})^T, \ldots, \boldsymbol{h}_B(\boldsymbol{M})^T\}^T$. The random effects vector associated with the mean, $\mathbf{u}_\mu$, is distributed as $\mathbf{u}_\mu \sim \mathrm{N}(\mathbf{0}, \sigma_\mu^2 \boldsymbol{I})$. For $b = 1, \ldots, B$, each $\boldsymbol{\alpha}_b = (\alpha_{b1}, \ldots, \alpha_{bN})^T$ follows a normal distribution, $\boldsymbol{\alpha}_b \sim \mathrm{N}(\mathbf{0}, \delta_b \boldsymbol{I})$, and each $\boldsymbol{h}_b(\boldsymbol{M}) = \{h_b(\boldsymbol{m}_1), \ldots, h_b(\boldsymbol{m}_N)\}^T$ follows a Gaussian process, such that $\boldsymbol{h}_b(\boldsymbol{M}) \sim \mathrm{N}(\mathbf{0}, \gamma_b \boldsymbol{R})$. The covariance matrix, $\boldsymbol{R}$, depends on the covariate vectors of interest through a specified kernel matrix (linear, quadratic, or Gaussian) and γ_b is an unknown variance component.

The associated design matrix for the covariates is $\boldsymbol{X} = [\tilde{\boldsymbol{X}}_0 \ \tilde{\boldsymbol{X}}_1 \ \cdots \ \tilde{\boldsymbol{X}}_C]$. Additionally, $\boldsymbol{Z}_\mu$, $\boldsymbol{Z}_\alpha$, and $\boldsymbol{Z}_h$ are known positive semi-definite matrices associated with the random effects vectors. The full random effect vector is distributed as

$$\begin{pmatrix} \mathbf{u}_\mu \\ \mathbf{u}_\alpha \\ \mathbf{u}_h \\ \boldsymbol{\epsilon} \end{pmatrix} \sim \mathrm{N}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{G}_\mu & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_\alpha & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{G}_h & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \sigma_\epsilon^2 \boldsymbol{I}_{\tilde{n}} \end{bmatrix} \right),$$

where $\mathbf{G}_\mu = \sigma_\mu^2 \boldsymbol{I}_K$, $\mathbf{G}_\alpha = \mathrm{diag}(\boldsymbol{\delta}) \otimes \boldsymbol{I}_N$, $\mathbf{G}_h = \mathrm{diag}(\boldsymbol{\gamma}) \otimes \boldsymbol{R}$, $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_B)^T$, and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_B)^T$. Therefore, the full response vector $\mathbf{Y}$ is distributed as

$$\mathbf{Y} \sim \mathrm{N}(\boldsymbol{X}\boldsymbol{\beta},\ \boldsymbol{Z}_\mu \mathbf{G}_\mu \boldsymbol{Z}_\mu^T + \boldsymbol{Z}_\alpha \mathbf{G}_\alpha \boldsymbol{Z}_\alpha^T + \boldsymbol{Z}_h \mathbf{G}_h \boldsymbol{Z}_h^T + \sigma_\epsilon^2 \boldsymbol{I}_{\tilde{n}}). \tag{5.4}$$

5.2.1 Penalized Regression Framework

If we appropriately reparameterize (5.3), a penalized regression model can be used to estimate the model parameters. The connection between mixed models and penalized regression was introduced in Chapter 4. To reparameterize, we assume that there exists some matrix $\boldsymbol{D}$ such that $\boldsymbol{R} = \boldsymbol{D}\boldsymbol{D}^T$. Let $\widetilde{\boldsymbol{D}} = bdiag(\boldsymbol{D}, \ldots, \boldsymbol{D})$, a block diagonal matrix of B $\boldsymbol{D}$ matrices. Ad-

ditionally, let $\mathbf{v}_h = (\boldsymbol{\nu}_1^T, \ldots, \boldsymbol{\nu}_B^T)^T$ be some random effects vector such that $\boldsymbol{\nu}_b \sim \mathrm{N}(\mathbf{0}, \gamma_b \boldsymbol{I}_N)$. Then, $\mathbf{u}_h = \{\boldsymbol{h}_1(\boldsymbol{M})^T, \ldots, \boldsymbol{h}_B(\boldsymbol{M})^T\}^T \equiv (\boldsymbol{D}\boldsymbol{\nu}_1^T, \ldots, \boldsymbol{D}\boldsymbol{\nu}_B^T)^T$ and we can rewrite the model and corresponding random effects vector as,

$$\mathbf{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}_\mu \mathbf{u}_\mu + \boldsymbol{Z}_\alpha \mathbf{u}_\alpha + \boldsymbol{Z}_h^* \mathbf{v}_h + \boldsymbol{\epsilon}, \quad \begin{pmatrix} \mathbf{u}_\mu \\ \mathbf{u}_\alpha \\ \mathbf{v}_h \\ \boldsymbol{\epsilon} \end{pmatrix} \sim \mathrm{N}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{G}_\mu & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_\alpha & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{G}_h^* & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \sigma_\epsilon^2 \boldsymbol{I}_{\tilde{n}} \end{bmatrix} \right),$$

where $\boldsymbol{Z}_h^* = \boldsymbol{Z}\widetilde{\boldsymbol{D}}$ and $\mathbf{G}_h^* = \mathrm{diag}(\boldsymbol{\gamma}) \otimes \boldsymbol{I}_N$. Then, the equivalent penalized regression problem is

$$\begin{aligned} \underset{\boldsymbol{\beta}, \mathbf{u}_\mu, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_B, \boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_B}{\operatorname{argmin}} & \big\{ ||\mathbf{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}_\mu \mathbf{u}_\mu - \boldsymbol{Z}_\alpha \mathbf{u}_\alpha - \boldsymbol{Z}_h^* \mathbf{v}_h)||^2 \\ & + \theta_\mu ||\mathbf{u}_\mu||^2 + \sum_{b=1}^{B} \theta_{\alpha_b} ||\boldsymbol{\alpha}_b||^2 + \sum_{b=1}^{B} \theta_{h_b} ||\boldsymbol{\nu}_b||^2) \big\}. \end{aligned} \tag{5.5}$$

To solve the penalized regression problem in (5.5), we use a stable, restricted maximum likelihood estimation method to select the smoothing parameters, as proposed in Wood (2011). This is implemented through the `gam` function in the `mgcv` `R` package, with REML as the smoothing parameter estimation method. In penalized regression, the effects are assumed to be fixed effects and the estimated fixed effects vector is denoted as $\hat{\mathbf{C}}_{\text{coeff}} = (\hat{\boldsymbol{\beta}}^T, \hat{\mathbf{u}}_\mu^T, \hat{\mathbf{u}}_\alpha^T, \hat{\mathbf{v}}_h^T)^T$. The variance matrix for the coefficients is also estimated as $\mathbf{V}_c = \mathrm{Var}(\hat{\mathbf{C}}_{\text{coeff}})$. There are several choices for how to construct $\mathbf{V}_c$, but we choose to implement the Bayesian posterior covariance matrix for the parameters, corrected for the uncertainty of the smoothing parameter estimates (Wood, 2011). As noted in the `mgcv` documentation and confirmed by our simulations, this variance matrix ultimately provides standard errors that are better suited for confidence interval construction. The estimates for the variance components can also be obtained through the smoothing parameters as $(\hat{\sigma}_\mu^2, \hat{\delta}_1, \ldots, \hat{\delta}_B, \hat{\gamma}_1, \ldots, \hat{\gamma_B})^T = (\hat{\sigma}_\epsilon^2/\hat{\theta}_\mu, \hat{\sigma}_\epsilon^2/\hat{\theta}_{\alpha_1}, \ldots, \hat{\sigma}_\epsilon^2/\hat{\theta}_{\alpha_B}, \hat{\sigma}_\epsilon^2/\hat{\theta}_{h_1}, \ldots, \hat{\sigma}_\epsilon^2/\hat{\theta}_{h_B})^T$.

5.3 Prediction of Risk Profiles and Full Growth Curves

Recall that the eigenfunctions used in the expansion of the functional model are evaluated over a grid of $\tilde{g}$ equally spaced timepoints for $t \in [0, T]$, where the vector of grid points is denoted as $\boldsymbol{t}_{\tilde{g}}$. In order to construct the risk profiles and full growth curves, we must use the full grid of time points, rather than the observed points. Therefore, the full design matrices must be defined. The full design matrix for the covariates is $\boldsymbol{X}_{\text{full}} = [\tilde{\boldsymbol{X}}_{0_{\text{full}}}\ \tilde{\boldsymbol{X}}_{1_{\text{full}}}\ \cdots\ \tilde{\boldsymbol{X}}_{C_{\text{full}}}]$. The matrix $\tilde{\boldsymbol{X}}_{0_{\text{full}}}$ is $(N\tilde{g} \times 2)$ and represents the intercept and linear effect of the mean. For any $c = 1, \ldots, C$ the full design matrix for a fixed effect is $(N\tilde{g} \times B)$. They are defined as

$$\tilde{\boldsymbol{X}}_{0_{\text{full}}} = \begin{bmatrix} \mathbf{1}_{\tilde{g}} & \boldsymbol{t}_{\tilde{g}} \\ \vdots & \vdots \\ \mathbf{1}_{\tilde{g}} & \boldsymbol{t}_{\tilde{g}} \end{bmatrix} \text{ and } \tilde{\boldsymbol{X}}_{c_{\text{full}}} = \begin{bmatrix} x_{1c}\phi_1(\boldsymbol{t}_{\tilde{g}}) & \ldots & x_{1c}\phi_B(\boldsymbol{t}_{\tilde{g}}) \\ \vdots & & \vdots \\ x_{Nc}\phi_1(\boldsymbol{t}_{\tilde{g}}) & \ldots & x_{Nc}\phi_B(\boldsymbol{t}_{\tilde{g}}) \end{bmatrix}.$$

Additionally, $\boldsymbol{Z}_{\mu_{\text{full}}}$ is $(N\tilde{g} \times K)$ and $\boldsymbol{Z}_{\alpha_{\text{full}}} = \boldsymbol{Z}_{h_{\text{full}}}$ are $(N\tilde{g} \times NB)$. They are defined as

$$\boldsymbol{Z}_{\mu_{\text{full}}} = \begin{bmatrix} (t_{11} - \kappa_1)_+ & \ldots & (t_{11} - \kappa_K)_+ \\ \vdots & & \vdots \\ (t_{1\tilde{g}} - \kappa_1)_+ & \ldots & (t_{1\tilde{g}} - \kappa_K)_+ \\ \vdots & & \vdots \\ (t_{N1} - \kappa_1)_+ & \ldots & (t_{N1} - \kappa_K)_+ \\ \vdots & & \vdots \\ (t_{N\tilde{g}} - \kappa_1)_+ & \ldots & (t_{N\tilde{g}} - \kappa_K)_+ \end{bmatrix},$$

$$\boldsymbol{Z}_{\alpha_{\text{full}}} = \boldsymbol{Z}_{h_{\text{full}}} = \begin{bmatrix} \boldsymbol{\phi}_1(\boldsymbol{t}_{\tilde{g}}) & \ldots & \mathbf{0} & \ldots & \ldots & \boldsymbol{\phi}_B(\boldsymbol{t}_{\tilde{g}}) & \ldots & \mathbf{0} \\ \vdots & \ddots & \vdots & \ldots & \ldots & \vdots & \ddots & \vdots \\ \mathbf{0} & \ldots & \boldsymbol{\phi}_1(\boldsymbol{t}_{\tilde{g}}) & \ldots & \ldots & \mathbf{0} & \ldots & \boldsymbol{\phi}_B(\boldsymbol{t}_{\tilde{g}}) \end{bmatrix}.$$

Let $\mathbf{P}_{\text{v}} = [\mathbf{0}_{NB \times (2+BC+K+NB)}\ \boldsymbol{I}_{NB}]$, such that $\hat{\mathbf{v}}_h = \mathbf{P}_{\text{v}}\hat{\mathbf{C}}_{\text{coeff}}$. Then, $\hat{\mathbf{u}}_h = \widetilde{\boldsymbol{D}}\mathbf{P}_{\text{v}}\hat{\mathbf{C}}_{\text{coeff}}$ and $\mathbf{V}_{\text{u}_h} = \text{Var}(\hat{\mathbf{u}}_h) = (\widetilde{\boldsymbol{D}}\mathbf{P}_{\text{v}})\mathbf{V}_c(\widetilde{\boldsymbol{D}}\mathbf{P}_{\text{v}})^T$. The full risk profiles for all subjects across the grid of

points is estimated as $\hat{\mathbf{h}}(\boldsymbol{M}, \boldsymbol{t}_{\tilde{g}}) = \boldsymbol{Z}_{h_{\text{full}}} \hat{\mathbf{u}}_h$ and the full covariance matrix for all risk profiles is $\boldsymbol{Z}_{h_{\text{full}}} \mathbf{V}_{\mathrm{u}_h} \boldsymbol{Z}_{h_{\text{full}}}^T$. The square root of the diagonal from this covariance matrix can be used to estimate the standard errors for risk profiles, which can be used to construct confidence bands. Along with the risk profile estimates, we can also estimate the full growth trajectories for all individuals as

$$\hat{\mathbf{Y}}_{\text{full}} = [\boldsymbol{X}_{\text{full}} \ \boldsymbol{Z}_{\mu_{\text{full}}} \ \boldsymbol{Z}_{\alpha_{\text{full}}} \ \boldsymbol{Z}^*_{h_{\text{full}}}] \hat{\mathbf{C}}_{\text{coeff}}$$

where $\boldsymbol{Z}^*_{h_{\text{full}}} = \boldsymbol{Z}_{h_{\text{full}}} \widetilde{\boldsymbol{D}}$. The variance for $\hat{\mathbf{Y}}_{\text{full}}$ is

$$\text{Var}(\hat{\mathbf{Y}}_{\text{full}}) = [\boldsymbol{X}_{\text{full}} \ \boldsymbol{Z}_{\mu_{\text{full}}} \ \boldsymbol{Z}_{\alpha_{\text{full}}} \ \boldsymbol{Z}^*_{h_{\text{full}}}] \mathbf{V}_c [\boldsymbol{X}_{\text{full}} \ \boldsymbol{Z}_{\mu_{\text{full}}} \ \boldsymbol{Z}_{\alpha_{\text{full}}} \ \boldsymbol{Z}^*_{h_{\text{full}}}]^T$$

5.3.1 Prediction of New Observations

For a new subject with covariate profile $\boldsymbol{m}_{\text{new}}$ and baseline covariate vector $\boldsymbol{x}_{\text{new}}$, the risk profile and full growth trajectory are defined as

$$\boldsymbol{h}(\boldsymbol{m}_{\text{new}}, \boldsymbol{t}_{\tilde{g}}) = \hat{\boldsymbol{\phi}}_1(\boldsymbol{t}_{\tilde{g}}) h_1(\boldsymbol{m}_{\text{new}}) + \cdots + \hat{\boldsymbol{\phi}}_B(\boldsymbol{t}_{\tilde{g}}) h_B(\boldsymbol{m}_{\text{new}}) = \boldsymbol{Z}_{h_{\text{full}}} \mathbf{u}_{h_{\text{new}}}$$

$$\mathbf{Y}_{\text{new}} = \boldsymbol{X}_{\text{new}_{\text{full}}} \hat{\boldsymbol{\beta}} + \boldsymbol{Z}_{\mu_{\text{full}}} \hat{\mathbf{u}}_\mu + \boldsymbol{Z}_{\alpha_{\text{full}}} \mathbf{u}_{\alpha_{\text{new}}} + \boldsymbol{Z}_{h_{\text{full}}} \mathbf{u}_{h_{\text{new}}} + \boldsymbol{\epsilon}_{\text{new}}$$

where $\mathbf{u}_{h_{\text{new}}} = \{h_1(\boldsymbol{m}_{\text{new}}), \ldots, h_B(\boldsymbol{m}_{\text{new}})\}^T$ and the full design matrices are constructed as in Section 5.3, with $N = 1$. Recall that each $\boldsymbol{h}_b(\boldsymbol{M})$ follows a Gaussian process, such that $\boldsymbol{h}_b(\boldsymbol{M}) \sim \text{N}(\mathbf{0}, \gamma_b \boldsymbol{R})$. The covariance matrix, $\boldsymbol{R}$, depends on the standardized methylation matrix, $\widetilde{\boldsymbol{M}}$, through the specified kernel, such that $\boldsymbol{R}_{ij} = R(\widetilde{\boldsymbol{m}}_i, \widetilde{\boldsymbol{m}}_j)$. Therefore, the first step in predicting $\boldsymbol{h}(\boldsymbol{m}_{\text{new}}, \boldsymbol{t}_{\tilde{g}})$ is to standardize $\boldsymbol{m}_{\text{new}}$ using the same means and standard errors used to standardize the columns of $\boldsymbol{M}$ in the original analysis. This is denoted as $\widetilde{\boldsymbol{m}}_{\text{new}}$. Then for each $b = 1, \ldots, B$,

$$\begin{bmatrix} \boldsymbol{h}_b(\boldsymbol{M}) \\ h_b(\boldsymbol{m}_{\text{new}}) \end{bmatrix} \sim \text{N}\left(\begin{bmatrix} \mathbf{0} \\ 0 \end{bmatrix}, \gamma_b \begin{bmatrix} \boldsymbol{R} & \boldsymbol{R}_* \\ \boldsymbol{R}_*^T & \boldsymbol{R}_{**} \end{bmatrix} \right),$$

where $\boldsymbol{R}_* = [R(\widetilde{\boldsymbol{m}}_{\text{new}}, \widetilde{\boldsymbol{m}}_1), \cdots, R(\widetilde{\boldsymbol{m}}_{\text{new}}, \widetilde{\boldsymbol{m}}_N)]^T$ and $\boldsymbol{R}_{**} = R(\widetilde{\boldsymbol{m}}_{\text{new}}, \widetilde{\boldsymbol{m}}_{\text{new}})$ for the designated covariance kernel $R(\cdot, \cdot)$. The distribution of $h_b(\boldsymbol{m}_{\text{new}})|\boldsymbol{h}_b(\boldsymbol{M})$ is determined by the laws of conditional normal distributions, such that

$$h_b(\boldsymbol{m}_{\text{new}})|\boldsymbol{h}_b(\boldsymbol{M}) \sim \text{N}\{\boldsymbol{R}_*^T \boldsymbol{R}^{-1} \boldsymbol{h}_b(\boldsymbol{M}), \gamma_b(\boldsymbol{R}_{**} - \boldsymbol{R}_*^T \boldsymbol{R}^{-1} \boldsymbol{R}_*)\}.$$

The best estimate of $h_b(\boldsymbol{m}_{\text{new}})$ is the expected value of the mean of the conditional distribution, $\hat{h}_b(\boldsymbol{m}_{\text{new}}) = \boldsymbol{R}_*^T \boldsymbol{R}^{-1} \hat{\boldsymbol{h}}_b(\boldsymbol{M})$. Then, the estimated risk profile for the new observation is

$$\hat{\boldsymbol{h}}(\boldsymbol{m}_{\text{new}}, \boldsymbol{t}_{\tilde{g}}) = \hat{\boldsymbol{\phi}}_1(\boldsymbol{t}_{\tilde{g}}) \hat{h}_1(\boldsymbol{m}_{\text{new}}) + \cdots + \hat{\boldsymbol{\phi}}_B(\boldsymbol{t}_{\tilde{g}}) \hat{h}_B(\boldsymbol{m}_{\text{new}}) = \boldsymbol{Z}_{h_{\text{full}}} \hat{\mathbf{u}}_{h_{\text{new}}}.$$

Once we have an estimate of $\hat{\mathbf{u}}_{h_{\text{new}}}$, it can be used to predict $\hat{\mathbf{Y}}_{\text{new}}$. Since all α's are assumed to be independent and identically distributed with mean zero, it is assumed that $\hat{\mathbf{u}}_{\alpha_{\text{new}}} = \mathbf{0}$. Therefore,

$$\hat{\mathbf{Y}}_{\text{new}} = \boldsymbol{X}_{\text{new}_{\text{full}}} \hat{\boldsymbol{\beta}} + \boldsymbol{Z}_{\mu_{\text{full}}} \hat{\mathbf{u}}_{\mu} + \boldsymbol{Z}_{h_{\text{full}}} \hat{\mathbf{u}}_{h_{\text{new}}}.$$

Since the independent subject-specific random effects are assumed to be zero, $\hat{\mathbf{Y}}_{\text{new}}$ can be thought of as the mean growth curve of all subjects with methylation profile $\boldsymbol{m}_{\text{new}}$ and baseline covariates $\boldsymbol{x}_{\text{new}}$.

5.3.2 Prediction Variances and Bands

In addition to the point-wise predictions of the risk profiles and growth trajectories, we would also like to construct prediction bands. This involves calculating the prediction variances for $\boldsymbol{Z}_{h_{\text{full}}} \hat{\mathbf{u}}_{h_{\text{new}}}$ and $\hat{\mathbf{Y}}_{\text{new}}$. As discussed in Section 5.1, this is important to provide proper coverage, which will be shown through simulations in Section 5.5.

The key term in both prediction variances is $\hat{\mathbf{u}}_{h_{\text{new}}}$ and so we start here. For simplicity, assume that $B = 1$. There are many ways to calculate variance estimates for $\mathbf{u}_{h_{\text{new}}}$ and $\hat{\mathbf{u}}_{h_{\text{new}}}$ and we list some below.

- Using the variance of $\hat{\mathbf{u}}_h$ from the original model, $\text{Var}(\hat{\mathbf{u}}_{h_{\text{new}}}) = \boldsymbol{R}_*^T \boldsymbol{R}^{-1} \mathbf{V}_h \boldsymbol{R}^{-1} \boldsymbol{R}_*$.
- Using the definition of the conditional distribution, $\text{Var}(\mathbf{u}_{h_{\text{new}}} | \mathbf{u}_h) = \hat{\gamma}(\boldsymbol{R}_{**} - \boldsymbol{R}_*^T \boldsymbol{R}^{-1} \boldsymbol{R}_*)$.
- Using the definition of the marginal distribution, $\text{Var}(\mathbf{u}_{h_{\text{new}}}) = \hat{\gamma} \boldsymbol{R}_{**}$.

To account for the extra prediction variability, the prediction variance is used instead, which involves two of the variance estimates above. The prediction variance for $\hat{\mathbf{u}}_{h_{\text{new}}}$ is

$$\text{Var}(\hat{\mathbf{u}}_{h_{\text{new}}} - \mathbf{u}_{h_{\text{new}}}) = \text{Var}(\hat{\mathbf{u}}_{h_{\text{new}}}) + \text{Var}(\mathbf{u}_{h_{\text{new}}}) - 2\text{Cov}(\hat{\mathbf{u}}_{h_{\text{new}}}, \mathbf{u}_{h_{\text{new}}}). \tag{5.6}$$

The first two variance terms are defined above, but the key to getting the correct prediction variance is the covariance term (which is not zero). Since $\hat{\mathbf{u}}_{h_{\text{new}}} = \boldsymbol{R}_*^T \boldsymbol{R}^{-1} \hat{\mathbf{u}}_h$, the covariance of interest is really $\text{Cov}(\hat{\mathbf{u}}_h, \mathbf{u}_{h_{\text{new}}})$. In order to calculate this covariance, we go back to thinking about the model in the mixed models context. We can rewrite (5.3) as $\mathbf{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\mathbf{u} + \boldsymbol{\epsilon}$, where $\boldsymbol{Z} = [\boldsymbol{Z}_\mu \ \boldsymbol{Z}_\alpha \ \boldsymbol{Z}_h]$ and $\mathbf{G} = bdiag(\mathbf{G}_\mu, \mathbf{G}_\alpha, \mathbf{G}_h)$. Then $\mathbf{Y} \sim \text{N}(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{Z}\mathbf{G}\boldsymbol{Z}^T + \sigma_\epsilon^2 \boldsymbol{I}_{\tilde{n}})$, and the variance of $\mathbf{Y}$ is $\mathbf{V} = \boldsymbol{Z}\mathbf{G}\boldsymbol{Z}^T + \sigma_\epsilon^2 \boldsymbol{I}_{\tilde{n}}$. The best linear unbiased estimator (BLUE) of the fixed effects vector is $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \mathbf{V}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T \mathbf{V}^{-1} \mathbf{Y}$ and the best linear unbiased predictor (BLUP) estimate of the random effects vector is $\hat{\mathbf{u}} = \hat{\mathbf{G}} \boldsymbol{Z}^T \hat{\mathbf{V}}^{-1} (\mathbf{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$. By substituting in the estimate for $\hat{\boldsymbol{\beta}}$, $\hat{\mathbf{u}}$ can be written as a linear combination of $\mathbf{Y}$ such that $\hat{\mathbf{u}} = (\hat{\mathbf{u}}_\mu^T, \hat{\mathbf{u}}_\alpha^T, \hat{\mathbf{u}}_h^T)^T = \mathbf{H}\mathbf{Y}$, for $\mathbf{H} = \hat{\mathbf{G}} \boldsymbol{Z}^T \hat{\mathbf{V}}^{-1} \big\{ \boldsymbol{I}_{\tilde{n}} - \boldsymbol{X}(\boldsymbol{X}^T \mathbf{V}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T \mathbf{V}^{-1} \big\}$. Then, when assuming mixed model notation, $\hat{\mathbf{u}}_h = \mathbf{P}_h \mathbf{H} \mathbf{Y}$, where $\mathbf{P}_h = [\mathbf{0}_{NB \times (K+NB)} \ \boldsymbol{I}_{NB}]$. The prediction variance for $\hat{\mathbf{u}}_{h_{\text{new}}}$ is derived as follows:

$$
\begin{aligned}
\text{Var}(\hat{\mathbf{u}}_{h_{\text{new}}} - \mathbf{u}_{h_{\text{new}}}) &= \text{Var}(\hat{\mathbf{u}}_{h_{\text{new}}}) + \text{Var}(\mathbf{u}_{h_{\text{new}}}) - 2\text{Cov}(\hat{\mathbf{u}}_{h_{\text{new}}}, \mathbf{u}_{h_{\text{new}}}) \\
&= \boldsymbol{R}_*^T \boldsymbol{R}^{-1} \mathbf{V}_h \boldsymbol{R}^{-1} \boldsymbol{R}_* + \hat{\gamma} \boldsymbol{R}_{**} - 2\text{Cov}(\boldsymbol{R}_*^T \boldsymbol{R}^{-1} \hat{\mathbf{u}}_h, \mathbf{u}_{h_{\text{new}}}) \\
&= \boldsymbol{R}_*^T \boldsymbol{R}^{-1} \mathbf{V}_h \boldsymbol{R}^{-1} \boldsymbol{R}_* + \hat{\gamma} \boldsymbol{R}_{**} - 2\boldsymbol{R}_*^T \boldsymbol{R}^{-1} \text{Cov}(\hat{\mathbf{u}}_h, \mathbf{u}_{h_{\text{new}}}) \\
&= \boldsymbol{R}_*^T \boldsymbol{R}^{-1} \mathbf{V}_h \boldsymbol{R}^{-1} \boldsymbol{R}_* + \hat{\gamma} \boldsymbol{R}_{**} - 2\boldsymbol{R}_*^T \boldsymbol{R}^{-1} \text{Cov}(\mathbf{P}_h \mathbf{H} \mathbf{Y}, \mathbf{u}_{h_{\text{new}}}) \\
&= \boldsymbol{R}_*^T \boldsymbol{R}^{-1} \mathbf{V}_h \boldsymbol{R}^{-1} \boldsymbol{R}_* + \hat{\gamma} \boldsymbol{R}_{**} - 2\boldsymbol{R}_*^T \boldsymbol{R}^{-1} \mathbf{P}_h \mathbf{H} \text{Cov}(\mathbf{Y}, \mathbf{u}_{h_{\text{new}}}) \\
&= \boldsymbol{R}_*^T \boldsymbol{R}^{-1} \mathbf{V}_h \boldsymbol{R}^{-1} \boldsymbol{R}_* + \hat{\gamma} \boldsymbol{R}_{**} \\
&\quad - 2\boldsymbol{R}_*^T \boldsymbol{R}^{-1} \mathbf{P}_h \mathbf{H} \text{Cov}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}_\mu \mathbf{u}_\mu + \boldsymbol{Z}_\alpha \mathbf{u}_\alpha + \boldsymbol{Z}_h \mathbf{u}_h + \boldsymbol{\epsilon}, \mathbf{u}_{h_{\text{new}}}) \\
&= \boldsymbol{R}_*^T \boldsymbol{R}^{-1} \mathbf{V}_h \boldsymbol{R}^{-1} \boldsymbol{R}_* + \hat{\gamma} \boldsymbol{R}_{**} - 2\boldsymbol{R}_*^T \boldsymbol{R}^{-1} \mathbf{P}_h \mathbf{H} \text{Cov}(\boldsymbol{Z}_h \mathbf{u}_h, \mathbf{u}_{h_{\text{new}}}) \\
&= \boldsymbol{R}_*^T \boldsymbol{R}^{-1} \mathbf{V}_h \boldsymbol{R}^{-1} \boldsymbol{R}_* + \hat{\gamma} \boldsymbol{R}_{**} - 2\boldsymbol{R}_*^T \boldsymbol{R}^{-1} \mathbf{P}_h \mathbf{H} \boldsymbol{Z}_h \text{Cov}(\mathbf{u}_h, \mathbf{u}_{h_{\text{new}}}) \\
&= \boldsymbol{R}_*^T \boldsymbol{R}^{-1} \mathbf{V}_h \boldsymbol{R}^{-1} \boldsymbol{R}_* + \hat{\gamma} \boldsymbol{R}_{**} - 2\boldsymbol{R}_*^T \boldsymbol{R}^{-1} \mathbf{P}_h \mathbf{H} \boldsymbol{Z}_h \hat{\gamma} \boldsymbol{R}_* \\
&= \boldsymbol{R}_*^T \boldsymbol{R}^{-1} \mathbf{V}_h \boldsymbol{R}^{-1} \boldsymbol{R}_* + \hat{\gamma} \boldsymbol{R}_{**} - 2\hat{\gamma} \boldsymbol{R}_*^T \boldsymbol{R}^{-1} \mathbf{P}_h \mathbf{H} \boldsymbol{Z}_h \boldsymbol{R}_*
\end{aligned}
$$

Then, the prediction variance for $\boldsymbol{Z}_{h_{\text{full}}} \hat{\mathbf{u}}_{h_{\text{new}}}$ is

$$
\boldsymbol{Z}_{h_{\text{full}}} (\boldsymbol{R}_*^T \boldsymbol{R}^{-1} \mathbf{V}_h \boldsymbol{R}^{-1} \boldsymbol{R}_* + \hat{\gamma} \boldsymbol{R}_{**} - 2\hat{\gamma} \boldsymbol{R}_*^T \boldsymbol{R}^{-1} \mathbf{P}_h \mathbf{H} \boldsymbol{Z}_h \boldsymbol{R}_*) \boldsymbol{Z}_{h_{\text{full}}}^T.
$$

We perform similar computations to obtain the prediction variance for $\hat{\mathbf{Y}}_{\text{new}}$ as $\text{Var}(\hat{\mathbf{Y}}_{\text{new}} - \mathbf{Y}_{\text{new}}) = \text{Var}(\hat{\mathbf{Y}}_{\text{new}}) + \text{Var}(\mathbf{Y}_{\text{new}}) - 2\text{Cov}(\hat{\mathbf{Y}}_{\text{new}}, \mathbf{Y}_{\text{new}})$. Utilizing the prediction variance is important because although $\hat{\mathbf{Y}}_{\text{new}}$ does not include the independent subject-specific random effects, they are accounted for in the prediction variance. This is important to providing proper coverage. The derivations for the prediction variance of $\hat{\mathbf{Y}}_{\text{new}}$ are below.

$$
\begin{aligned}
\mathrm{Var}(\hat{\mathbf{Y}}_{\mathrm{new}}) &= \mathrm{Var}(\boldsymbol{X}_{\mathrm{new}}\hat{\boldsymbol{\beta}} + \boldsymbol{Z}_{\mu}\hat{\mathbf{u}}_{\mu} + \boldsymbol{Z}_{h_{\mathrm{new}}}\hat{\mathbf{u}}_{h_{\mathrm{new}}}) \\
&= [\boldsymbol{X}_{\mathrm{new}}\ \boldsymbol{Z}_{\mu}]\mathrm{Var}\{(\hat{\boldsymbol{\beta}}^{T}, \hat{\mathbf{u}}_{\mu}^{T})^{T}\}[\boldsymbol{X}_{\mathrm{new}}\ \boldsymbol{Z}_{\mu}]^{T} + \boldsymbol{Z}_{h_{\mathrm{new}}}\mathrm{Var}(\hat{\mathbf{u}}_{h_{\mathrm{new}}})\boldsymbol{Z}_{h_{\mathrm{new}}}^{T} \\
&= [\boldsymbol{X}_{\mathrm{new}}\ \boldsymbol{Z}_{\mu}\ \mathbf{P}_{\beta,\mu}]\mathbf{V}_{c}[\boldsymbol{X}_{\mathrm{new}}\ \boldsymbol{Z}_{\mu}\ \mathbf{P}_{\beta,\mu}]^{T} + \boldsymbol{Z}_{h_{\mathrm{new}}}\mathrm{Var}(\hat{\mathbf{u}}_{h_{\mathrm{new}}})\boldsymbol{Z}_{h_{\mathrm{new}}}^{T} \\
&= [\boldsymbol{X}_{\mathrm{new}}\ \boldsymbol{Z}_{\mu}\ \mathbf{P}_{\beta,\mu}]\mathbf{V}_{c}[\boldsymbol{X}_{\mathrm{new}}\ \boldsymbol{Z}_{\mu}\ \mathbf{P}_{\beta,\mu}]^{T} + \boldsymbol{Z}_{h_{\mathrm{new}}}\boldsymbol{R}_{*}^{T}\boldsymbol{R}^{-1}\mathbf{V}_{h}\boldsymbol{R}^{-1}\boldsymbol{R}_{*}\boldsymbol{Z}_{h_{\mathrm{new}}}^{T}
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{Var}(\mathbf{Y}_{\mathrm{new}}) &= \mathrm{Var}(\boldsymbol{X}_{\mathrm{new}}\boldsymbol{\beta} + \boldsymbol{Z}_{\mu_{\mathrm{new}}}\mathbf{u}_{\mu} + \boldsymbol{Z}_{\alpha_{\mathrm{new}}}\mathbf{u}_{\alpha_{\mathrm{new}}} + \boldsymbol{Z}_{h_{\mathrm{new}}}\mathbf{u}_{h_{\mathrm{new}}} + \boldsymbol{\epsilon}) \\
&= \hat{\delta}\boldsymbol{Z}_{\alpha_{\mathrm{new}}}\boldsymbol{Z}_{\alpha_{\mathrm{new}}}^{T} + \hat{\gamma}\boldsymbol{Z}_{h_{\mathrm{new}}}\boldsymbol{R}_{**}\boldsymbol{Z}_{h_{\mathrm{new}}}^{T} + \hat{\sigma}_{\epsilon}^{2}\boldsymbol{I}
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{Cov}(\hat{\mathbf{Y}}_{\mathrm{new}}, \mathbf{Y}_{\mathrm{new}}) &= \mathrm{Cov}(\boldsymbol{X}_{\mathrm{new}}\hat{\boldsymbol{\beta}} + \boldsymbol{Z}_{\mu}\hat{\mathbf{u}}_{\mu} + \boldsymbol{Z}_{h_{\mathrm{new}}}\hat{\mathbf{u}}_{h_{\mathrm{new}}}, \\
&\qquad \boldsymbol{X}_{\mathrm{new}}\boldsymbol{\beta} + \boldsymbol{Z}_{\mu_{\mathrm{new}}}\mathbf{u}_{\mu} + \boldsymbol{Z}_{\alpha_{\mathrm{new}}}\mathbf{u}_{\alpha_{\mathrm{new}}} + \boldsymbol{Z}_{h_{\mathrm{new}}}\mathbf{u}_{h_{\mathrm{new}}} + \boldsymbol{\epsilon}) \\
&= \mathrm{Cov}(\boldsymbol{Z}_{h_{\mathrm{new}}}\hat{\mathbf{u}}_{h_{\mathrm{new}}}, \boldsymbol{Z}_{h_{\mathrm{new}}}\mathbf{u}_{h_{\mathrm{new}}}) \\
&= \boldsymbol{Z}_{h_{\mathrm{new}}}\mathrm{Cov}(\hat{\mathbf{u}}_{h_{\mathrm{new}}}, \mathbf{u}_{h_{\mathrm{new}}})\boldsymbol{Z}_{h_{\mathrm{new}}}^{T} \\
&= \boldsymbol{Z}_{h_{\mathrm{new}}}(\hat{\gamma}\boldsymbol{R}_{*}^{T}\boldsymbol{R}^{-1}\mathbf{P}_{h}\mathbf{H}\boldsymbol{Z}_{h}\boldsymbol{R}_{*})\boldsymbol{Z}_{h_{\mathrm{new}}}^{T}
\end{aligned}
$$

Combining these, we have the prediction variance of $\hat{\mathbf{Y}}_{\mathrm{new}}$ as

$$
\begin{gathered}
[\boldsymbol{X}_{\mathrm{new}}\ \boldsymbol{Z}_{\mu}\ \mathbf{P}_{\beta,\mu}]\mathbf{V}_{c}[\boldsymbol{X}_{\mathrm{new}}\ \boldsymbol{Z}_{\mu}\ \mathbf{P}_{\beta,\mu}]^{T} + \boldsymbol{Z}_{h_{\mathrm{new}}}\boldsymbol{R}_{*}^{T}\boldsymbol{R}^{-1}\mathbf{V}_{h}\boldsymbol{R}^{-1}\boldsymbol{R}_{*}\boldsymbol{Z}_{h_{\mathrm{new}}}^{T} + \\
\hat{\delta}\boldsymbol{Z}_{\alpha_{\mathrm{new}}}\boldsymbol{Z}_{\alpha_{\mathrm{new}}}^{T} + \hat{\gamma}\boldsymbol{Z}_{h_{\mathrm{new}}}\boldsymbol{R}_{**}\boldsymbol{Z}_{h_{\mathrm{new}}}^{T} + \hat{\sigma}_{\epsilon}^{2}\boldsymbol{I} - \\
2(\boldsymbol{Z}_{h_{\mathrm{new}}}(\hat{\gamma}\boldsymbol{R}_{*}^{T}\boldsymbol{R}^{-1}\mathbf{P}_{h}\mathbf{H}\boldsymbol{Z}_{h}\boldsymbol{R}_{*})\boldsymbol{Z}_{h_{\mathrm{new}}}^{T})
\end{gathered}
$$

Once we have the prediciton variances, the prediction bands can be calcuated for the desired

confidence level, α_0. The prediction bands are calculated as

$$\boldsymbol{Z}_{h_{\text{full}}}\hat{\mathbf{u}}_{h_{\text{new}}} \pm z_{(\alpha_0/2)}\sqrt{diag(\text{Prediction variance of } \boldsymbol{Z}_{h_{\text{full}}}\hat{\mathbf{u}}_{h_{\text{new}}})}$$
$$\hat{\mathbf{Y}}_{\text{new}} \pm z_{(\alpha_0/2)}\sqrt{diag(\text{Prediction variance of } \hat{\mathbf{Y}}_{\text{new}})}$$

5.4 Data Application

The data application for this chapter furthers the study of the NEST data, which is the motivating example for the dissertation. In Chapter 3, it was shown that the PEG3 DMR is significantly associated with female growth trajectories. This signal is strongest when using the quadratic kernel to model the similarity between methylation profiles. Therefore, to further investigate the relationship between PEG3 and female growth trajectories, the data is modeled using the model proposed in this chapter with the quadratic kernel. First, we look at whether the model is appropriate or not. Then, using this method, additional insight is gained on how PEG3 affects the female growth curves by estimating and analyzing the $\boldsymbol{h}(\boldsymbol{m}, \boldsymbol{t}_{\tilde{g}})$ functions (risk profiles).

Using standard, existing methods, a common approach to modeling this data would be to use a functional model with the methylation modeled as fixed effects covariates. A penalized flexible functional regression model is used as the standard method to compare the proposed GamGP model to (Scheipl et al., 2015, 2016). This is implemented through the `pffr` function in the `R refund` package (Goldsmith et al., 2016). In one model, the mean of the CpG sites within the PEG3 DMR is modeled as a single fixed effect. We refer to this as the pffr mean model. In a second model, the pffr all sites model, each CpG site of the PEG3 DMR is modeled as its own independent fixed effect. Once all three models were fit, model diagnostics were used to assess the model fits. Residual plots, actual versus predicted plots, and plots of the full predicted curves with observed values on top are shown in Figure 5.1. The residual plots for both pffr models show a distinct fanning over time, indicating that these approaches do not adequately model the variability in the data. In contrast, the random effects in the GamGP model appropriately capture the variability in the data, as shown by the fact that the residual

plot has constant variability across time. This difference in model fits is seen in the actual versus predicted plots as well. The predicted curves for the pffr mean model are clustered around the mean, resulting in poor predictions as the data spreads out. Although the pffr all sites model does a little better, many of the predicted curves do not capture the observed points as time increases. The GamGP model performs best at estimating the full predicted curves. The diagnostics in Figure 5.1 indicate that the GamGP model provides a more adequate fit to the NEST data than either of the pffr models.

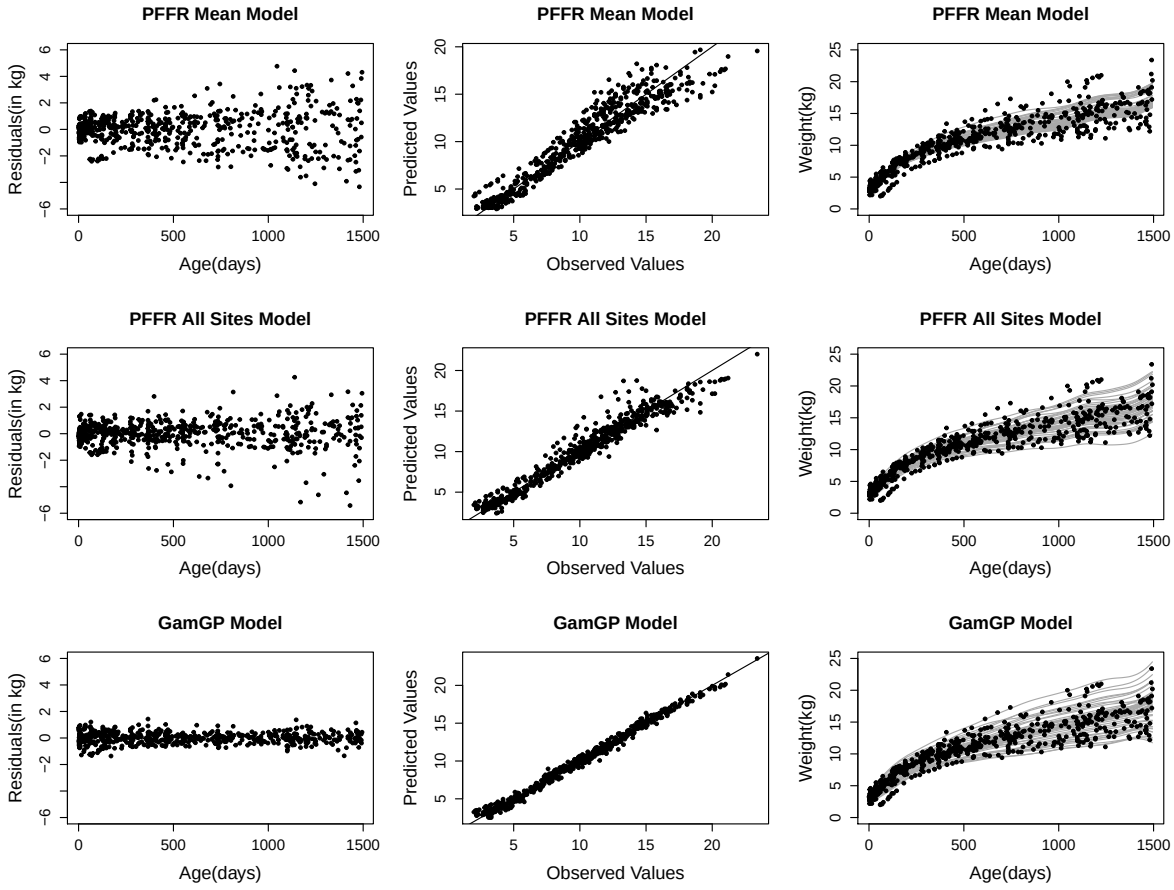


Figure 5.1: Plots of the residuals, observed versus predicted values, and full predicted curves for the two `pffr` models and the GamGP model. Observed values are also included in the predicted curve plots.

Since the GamGP model adequately fits the NEST data for the PEG3 DMR, we can confidently analyze the $\boldsymbol{h}(\boldsymbol{m}, \boldsymbol{t}_{\tilde{g}})$ functions, also known as the risk profiles. The risk profiles represent the contribution to the growth curve and are in the same units as the response, in this case kilograms. Therefore, a curve around the zero line indicates that the PEG3 methylation profile does not contribute anything to that individual's growth curve. The estimated risk profiles for this data example are shown in Figure 5.2. There are many risk profiles that are clustered around the zero line, but some show a positive contribution to the growth curves.

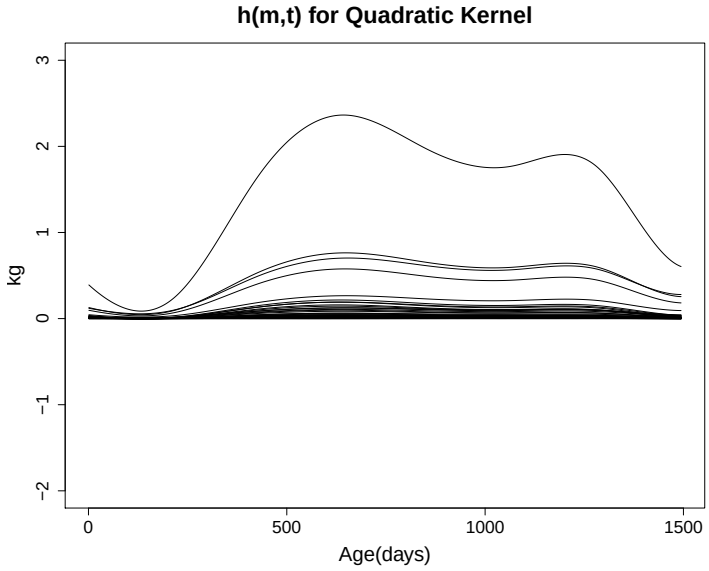


Figure 5.2: Plot of individual risk profiles, estimated through $\boldsymbol{h}(\boldsymbol{m}, \mathbf{t})$ using the quadratic kernel.

Once the risk profiles are estimated, there are many exploratory analyses that can be done to get a better idea of how the methylation affects growth. First, we can more definitively look at which risk profiles are different from the mean by constructing a point-wise confidence interval around the mean curve. The mean risk profile curve is estimated by first finding the mean of each $\boldsymbol{h}_b(\boldsymbol{M})$. Using the full effect vector, we assume $\mathbf{P}$ is some matrix such that $\bar{\hat{\mathbf{u}}}_h = \{\bar{\hat{\boldsymbol{h}}}_1(\boldsymbol{M})^T, \ldots, \bar{\hat{\boldsymbol{h}}}_B(\boldsymbol{M})^T\}^T = \mathbf{P}\hat{\mathbf{u}}_h$. Then, using $\boldsymbol{Z}_{h_{\text{full}}}$ for one curve ($N = 1$), the mean

risk profile is $\bar{\boldsymbol{h}}(\boldsymbol{m}, \mathbf{t}_{\tilde{g}}) = \boldsymbol{Z}_{h_{\text{full}}} \mathbf{P} \hat{\mathbf{u}}_h$. To construct the point-wise confidence interval, we use the square root of the diagonal of the covariance matrix of the mean curve, calculated as $\text{Var}\{\bar{\boldsymbol{h}}(\boldsymbol{m}, \mathbf{t}_{\tilde{g}})\} = (\boldsymbol{Z}_{h_{\text{full}}} \mathbf{P}) \mathbf{V}_{\text{u}_h} (\boldsymbol{Z}_{h_{\text{full}}} \mathbf{P})^T$. By doing so, we can determine which risk profiles fall outside of the "norm" and see what the corresponding PEG3 methylation profiles look like. In Figure 5.3, the PEG3 methylation values are shown in a line plot, with the subjects color coordinated as those whose risk profiles are different than the mean. Through this, it appears that subjects with generally higher methylation values across all CpG sites have a higher risk profile. The full predicted curves show that just because a subject has a positive contribution in their risk profile, this does not mean that he/she has the highest growth trajectory. The subjects with non-zero risk profiles cover the full spectrum of predicted growth curves.

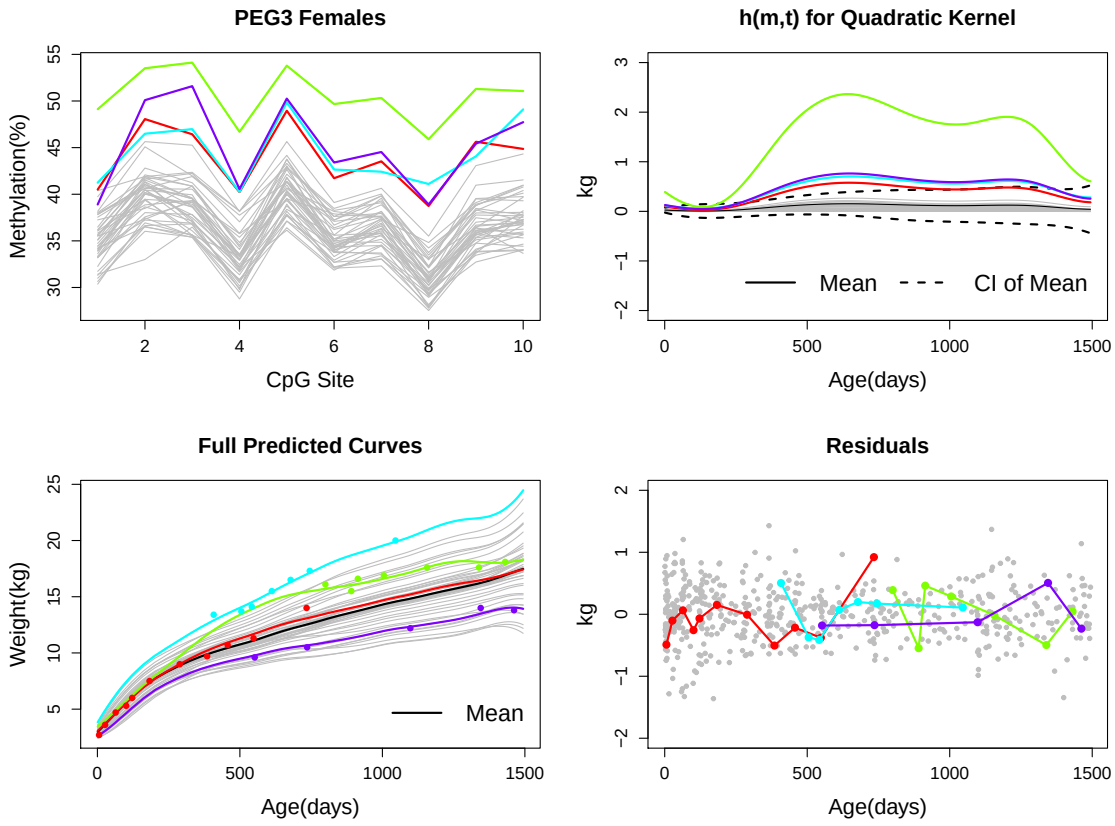


Figure 5.3: Methylation values for PEG3 CpG sites, predicted risk profiles, predicted curves, and residuals. Colored according to individual subjects.

To investigate further, the subjects are clustered into three groups by their raw PEG3 methylation profiles, using hierarchical clustering. Within each clustered group, the mean methylation for the CpG sites within PEG3 is calculated. The group means are then used as "new" observations to predict $\boldsymbol{h}(\boldsymbol{m}_{\text{new}}, \boldsymbol{t}_{\tilde{g}})$ for each of the three groups. A full growth curve, $\mathbf{Y}_{\text{new}}$, is also predicted for each group. This represents the mean growth curve for individuals with that methylation profile, as specific growth curves could be different by indivdual subject (does not account for α_i). The results are shown in Figure 5.4. All of the graphs in Figure 5.4 are colored according to the clustering groups. The first graph shows that the subjects were clustered into

groups with generally high, medium, and low methylation values. Consistent with the pattern observed in Figure 5.3, the group with the higher PEG3 values also has a positive risk profile and a slightly higher mean growth trajectory. There is no difference in the estimated risk profiles or growth curves for the two groups with medium and low PEG3 methylation values.

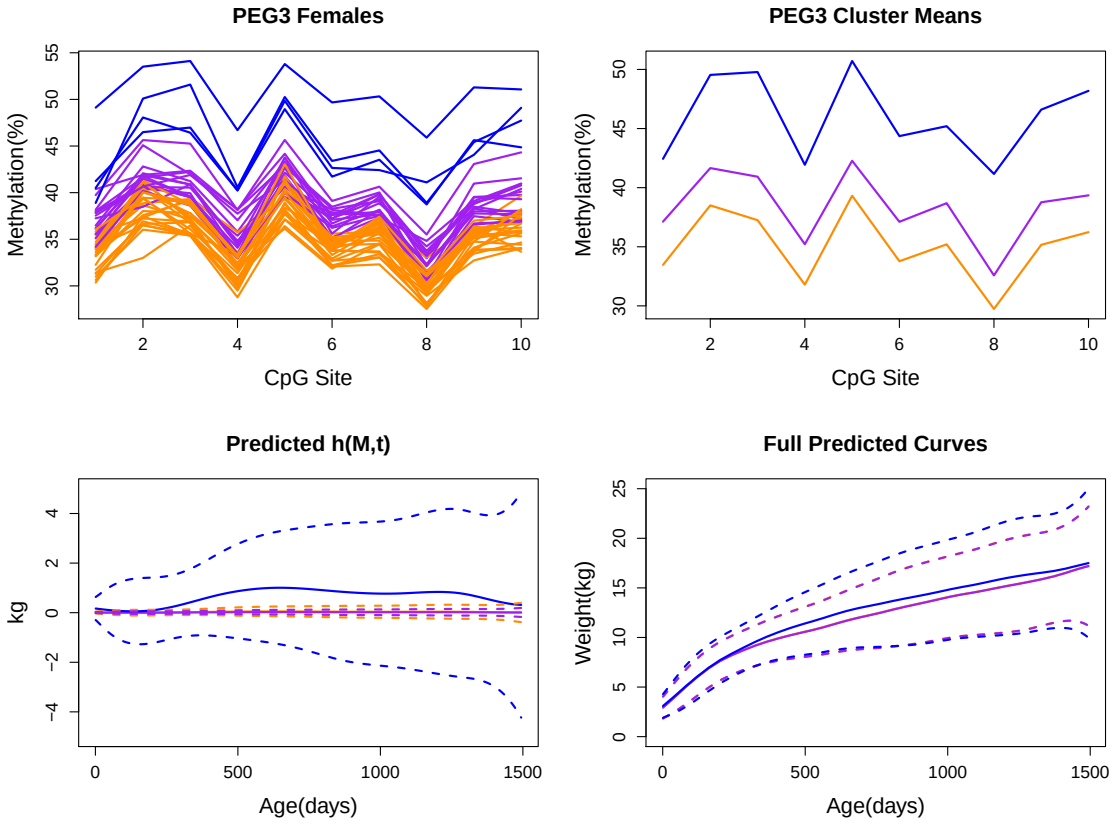


Figure 5.4: Raw methylation values for PEG3 CpG sites, methylation values for group means, predicted risk profiles with corresponding prediction intervals for each group, and predicted curves with corresponding prediction intervals for each group. Colored according to groups clustered by the raw methylation values.

We perform a second, similar analysis, but instead cluster the subjects according to their estimated $h_1(\boldsymbol{m}_i)$ and $h_2(\boldsymbol{m}_i)$ values. These results are presented in Figure 5.5. When clustered

this way, the one subject with the highest methylation values is clustered into its own group. Then there is a group with generally higher methylation and a group with generally medium to low methylation. This creates an increase in both the risk profile and predicted growth trajectory for the single subject. However, this also increases the width of the prediction bands, indicating that there is increased variability for this high level of methylation. In this analysis, there is bit more separation between the high and medium to low groups.

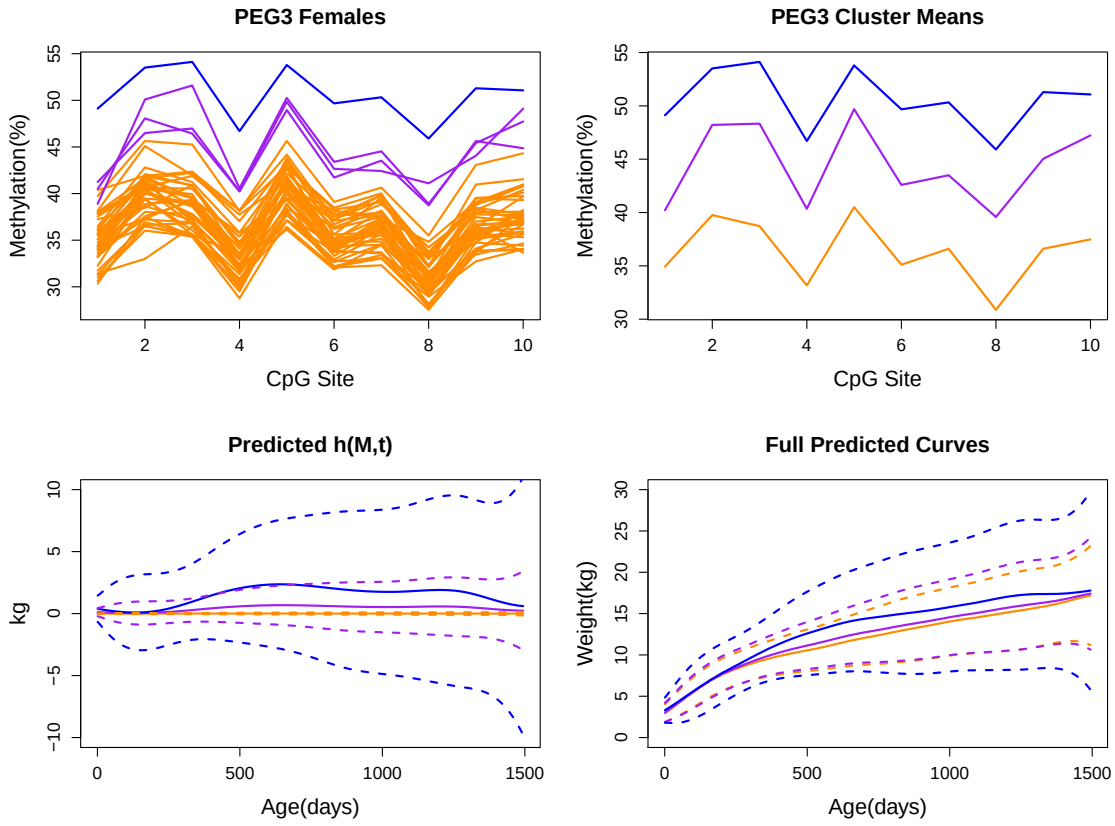


Figure 5.5: Raw methylation values for PEG3 CpG sites, methylation values for group means, predicted risk profiles with corresponding prediction intervals for each group, and predicted curves with corresponding prediction intervals for each group. Colored according to groups clustered by the estimated $\mathbf{u}_h$ values.

5.5 Simulation Study to Validate Prediction

A simulation study was performed to assess the prediction capabilities of the proposed model. Prior to any simulation iterations, an FPCA is performed on the full data to obtain an approximated full curve for each subject. This is done solely to have the most accurate full curve to compare any predicted full curves to in the simulation. For the simulation study, the female data is split into training and testing sets, with 39 and 5 subjects, respectively. This is done 100 times. To assess prediction ability, the mean square error (MSE) and median square error (MedSE) are calcuated using the observed points from the test data. Additionally, the MSE and MedSE are calculated across time using the approximated curves of the test subjects from the full FPCA. Time is grouped into less than 1 year, 1-2 years, 2-3 years, 3-4 years, and greater than 4 years. Using the observed points from the test set, the coverage of the confidence and prediction bands is also examined. We again compare GamGP to the two pffr models discussed in Section 5.4, the pffr mean model and the pffr all sites model. The folowing notation is used in the discussion of the simulation study.

- pffr(A) - The model is fit using `pffr`, with each CpG site in PEG3 modeled as its own independent, fixed effect. The new estimates and standard errors used to construct the confidence bands are obtained from this function as well.

- pffr(M) - The model is fit using `pffr`, with the mean of PEG3 modeled as a fixed effect. The new estimates and standard errors used to construct the confidence bands are obtained from this function as well.

- GamGP - The model is fit using the proposed model, GamGP, and the new predictions are obtained by estimating $\mathbf{u}_{h_{\text{new}}}$. When discussing coverage and standard errors, this model produces both traditional confidence bands (through regular variance) and prediction bands (through prediction variance). This model is analyzed for all three kernels.

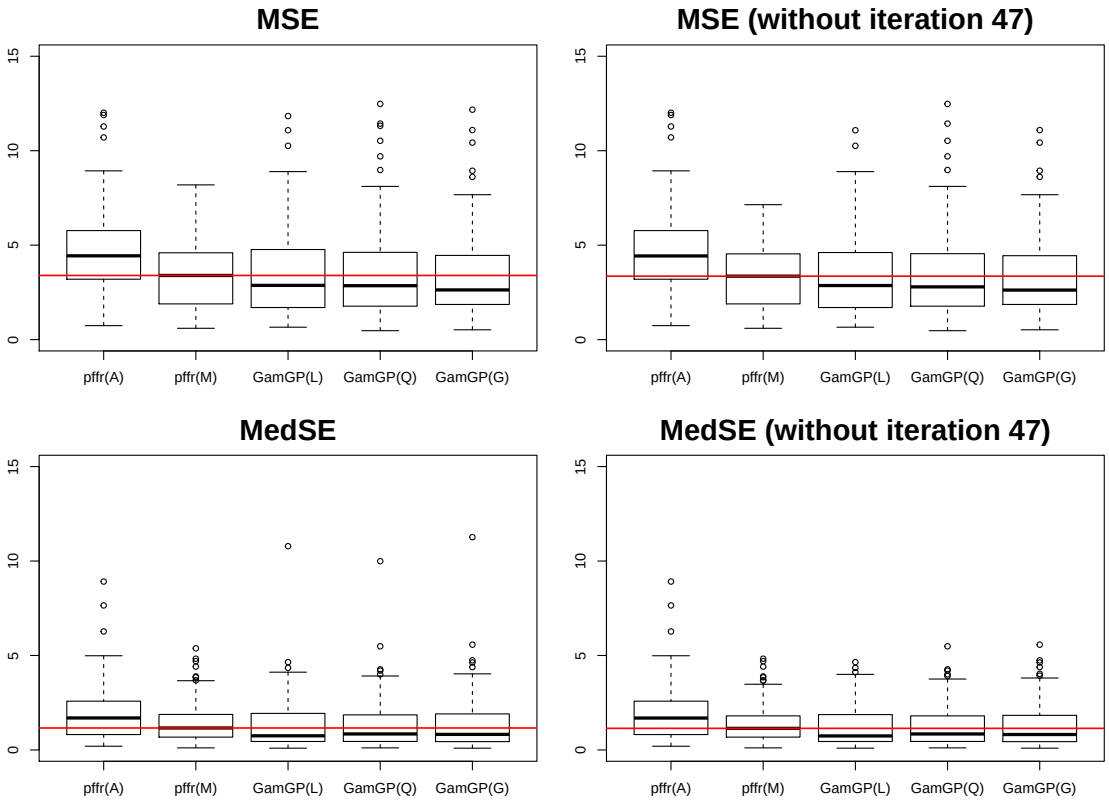


Figure 5.6: Boxplots of MSE and MedSE for the pffr and GamGP models. In the GamGP models, L, Q, and G correspond to the linear, quadratic, and Gaussian kernels, respectively.

5.5.1 Simulation Results

Figure 5.6 shows boxplots of the overall MSE and MedSE for the observed points across the 100 iterations. An additional graph of boxplots is shown for each with the removal of iteration 47, which proved to be an outlier in the GamGP models. The removal of the outlier did not impact the results. The pffr all sites model performs the worst, followed by the pffr mean model. The GamGP models for all kernels perform better in terms of both MSE and MedSE than either of the pffr models. These results are due to the fact that the pffr all sites model is using ten fixed effects and ultimately overfitting the data. While the predicted values for the observed data are

not that bad, the model does not extend well to predicting new observations. The pffr mean model loses information compared to the GamGP model. It is possible that two subjects could have the same mean value, but very different looking PEG3 methylation vectors. The GamGP model takes these differences into account, while not overfitting the data.

The boxplots of MSE and MedSE across time are shown in Figure 5.7. In these boxplots, only the pffr mean model is used since it performed better than the pffr all sites model. From the boxplots, we can see that for the most part, GamGP performs the same or better than pffr across time. The exception to this is for the timepoints greater than five years. This may be due to the fact that there were very few timepoints in this range.

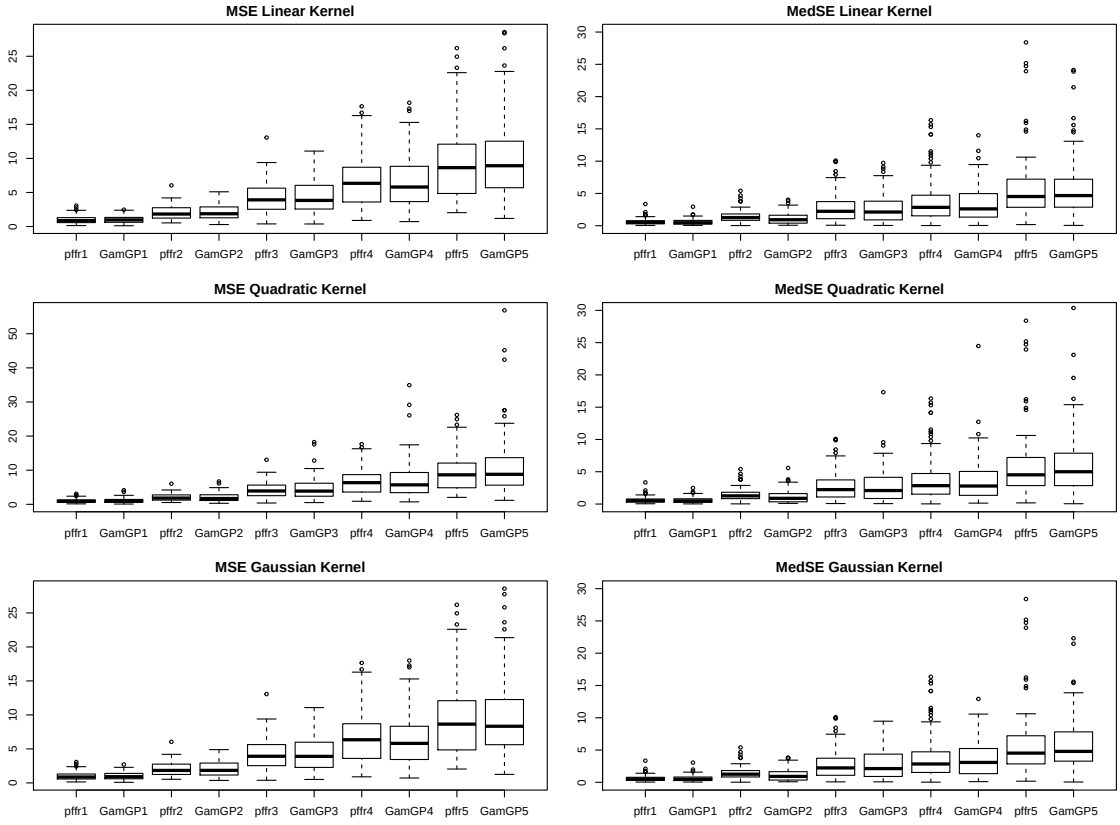


Figure 5.7: Boxplots of MSE and MedSE over time (grouped by year) for the pffr and GamGP models. In the GamGP models, L, Q, and G corresponde to the linear, quadratic, and Gaussian kernels, respectively.

The empirical coverages of the confidence and prediction bands are shown in Figure 5.8. The GamGP.pred models, which use the prediction bands, significantly outperform the other models. For all kernels, the prediction bands in the GamGP models achieve nominal coverage, at about 95%. The empirical coverage using the standard variance estimates was about 50% and the coverage by both pffr model falls below 40%. This is due to the fact that the pffr models to not properly capture the covariance structure of the data. The coverage of the confidence bands in the competing methods is problematic because the give a false sense of confidence about the predictions. Using the prediction intervals in the GamGP method, however, provides

much more useful information.

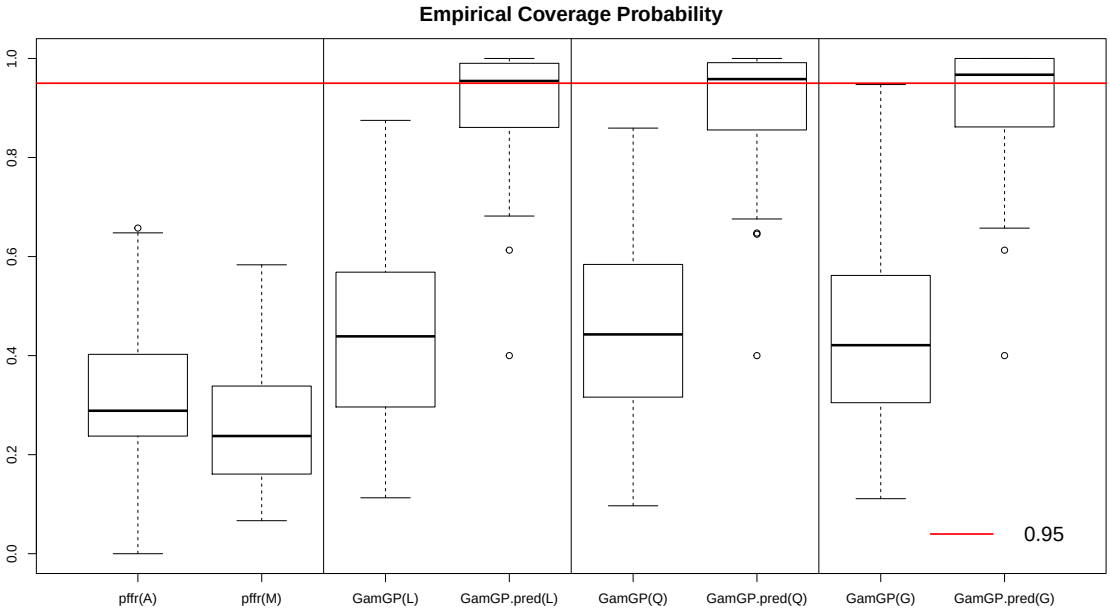


Figure 5.8: Boxplots of the percent of coverage achieved by each model's confidence or prediction bands. In the GamGP models, L, Q, and G correspond to the linear, quadratic, and Gaussian kernels, respectively. GamGP denotes the standard confidence intervals and GamGP.pred denotes the prediction intervals.

5.5.2 Example in Predicting New Growth Curves

The "best case" is using the real data as the training set and then treating the same data as new observations for the test set. The growth trajectories produced by GamGP in this example are not the same as those produced in the regular data analysis section. Since the data are treated as new observations, the subject-specific random effects, $\mathbf{u}_{\alpha_{\text{new}}}$, are assumed to be zero. Therefore, the predicted growth trajectories represent the mean growth curve for individuals with those methylation profiles and baseline covariates, as specific growth curves could be different by indivdual subject. The results for the pffr all sites model, the pffr mean

model, and the GamGP model with quadratic kernel are shown in Appendix C. The pffr models have corresponding confidence bands, while the GamGP models have corresponding confidence bands in black and prediction bands in red. The confidence bands for both pffr models are extremely narrow, while the prediction bands from the GamGP model appropriately capture the majority of the actual observations.

5.6 Discussion

In this chapter we round out the investigation of the relationship between PEG3 and the female growth trajectories. This is done by first estimating the individual risk profile for each subject through the $\boldsymbol{h}(\boldsymbol{m}, \mathbf{t}_{\tilde{g}})$ functions, which quantify the contribution that the PEG3 methylation values have on the growth trajectory across time. Analyzing the individual risk profiles enables us to determine which subjects have non-zero $\boldsymbol{h}(\boldsymbol{m}, \mathbf{t}_{\tilde{g}})$ functions and then see what their PEG3 methlyation values look like. We introduce some exploratory analyses that can be performed to try and find a pattern in the methylation values that lead to non-zero risk profiles. It is possible that there are better data visualization techniques to determine what types of methylation profile characteristics result in a positive, negative, or zero contribution to the growth trajectory. This is an area of future research.

We also develop a prediciton model to estimate the full growth trajectories for all subjects. This is especially important because this type of analysis focuses on sparsely observed functional responses. Therefore, since only a few points are actually observed, it is important to be able to provide an estimate of the full growth trajectories.

Since $\boldsymbol{h}_b(\boldsymbol{M})$ follows a Gaussian process, the properties of the assumed normal distribution can be used to predict $h_b(\boldsymbol{m}_{\text{new}})$ for a new subject given the estimated $\boldsymbol{h}_b(\boldsymbol{M})$ from the original model. Not only does this enable the prediction of a risk profile for a new subject, but also provides the parameters needed to predict the full growth trajectory. It should be noted that the curve estimates should not be taken as the estimate for that specific individual. Rather, these are the mean curves for subjects with the given methylation profile vector and baseline covariates.

This is because, for new predictions, there is not a way to account for the subject specific random effects ($\mathbf{u}_{\alpha_{\text{new}}}$) and thus they are assumed to be $\mathbf{0}$. We also construct prediction bands around the curves that offer near nominal coverage. We could improve upon new predictions by finding a way to estimate $\hat{\mathbf{u}}_{\alpha_{\text{new}}}$ and this is an area of future research.

Throughout the chapter, the results of the proposed GamGP model were compared to the current standard, a penalized flexible functional regression (pffr) model. In terms of model fit and coverage of confidence/prediction bands around new predicted curves, the GamGP model outperforms two versions of the pffr model. This is partially attributed to the inclusion of subject-specfic time-varying random effects in the GamGP model. The inclusion of these random effects allow the model to more adequately capture the correlation structure of the functional data, which results in a better performing model. The improved coverage is also due to the calculation of prediction variances, which is not the current standard in functional models. In terms of point-wise prediction accuracy, the GamGP model performs better or the same as the pffr models in simulation studies. Through the combination of comparable prediction accuracy and significantly improved coverage of prediction bands, we conclude that the GamGP model is an improvement on the current pffr models available.

This completes the general workflow for analyzing the effect of a vector of related covariates on a sparse functional response. One can first determine significance using either the VarCompGP method introduced in Chapter 3 or the GamGP method introduced in Chapter 4. Testing all three kernel options can determine which kernel best captures the significant relationship. Then, for the chosen significant vectors of related covariates, the prediction model and risk profiles discussed in this chapter can be used to determine how the covariates ultimately affect the response. Alternatively, if one is not interested in association testing, the best prediction model can be determined by fitting all three kernels and chosing the model based on some model fitting criterion, such as MSE.

REFERENCES

Anderson, E. L., Howe, L. D., Fraser, A., Callaway, M. P., Sattar, N., Day, C., Tilling, K., & Lawlor, D. A. (2014). Weight trajectories through infancy and childhood and risk of non-alcoholic fatty liver disease in adolescence: the alspac study. *Journal of hepatology*, *61*(3), 626–632.

Antoniadis, A., & Sapatinas, T. (2007). Estimation and inference in functional mixed-effects models. *Computational Statistics & Data Analysis*, *51*(10), 4793–4813.

Aston, J. A., Chiou, J.-M., & Evans, J. P. (2010). Linguistic pitch analysis using functional principal component mixed effect models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *59*(2), 297–317.

Barker, D. (2004). Developmental origins of adult health and disease. *Journal of Epidemiology & Community Health*, *58*(2), 114–115.

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

Chiavegatto, S., Sauce, B., Ambar, G., Cheverud, J. M., & Peripato, A. C. (2012). Hypothalamic expression of peg3 gene is associated with maternal care differences between sm/j and lg/j mouse strains. *Brain and behavior*, *2*(4), 365–376.

Davies, R. B. (1980). Algorithm as 155: The distribution of a linear combination of χ 2 random variables. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *29*(3), 323–333.

De Kroon, M. L., Renders, C. M., Van Wouwe, J. P., Van Buuren, S., & Hirasing, R. A. (2010). The terneuzen birth cohort: Bmi change between 2 and 6 years is most predictive of adult cardiometabolic risk. *PloS one*, *5*(11), e13966.

Duchesne, P., & de Micheaux, P. L. (2010). Computing the distribution of quadratic forms: Further comparisons between the liu-tang-zhang approximation and exact methods. *Computational Statistics and Data Analysis*, *54*, 858–862.

Ebden, M. (2015). Gaussian processes: A quick introduction. *arXiv preprint arXiv:1505.02965*.

Ezzahir, N., Alberti, C., Deghmoun, S., Zaccaria, I., Czernichow, P., Lévy-Marchal, C., & Jaquet, D. (2005). Time course of catch-up in adiposity influences adult anthropometry in individuals who were born small for gestational age. *Pediatric research*, *58*(2), 243.

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

Goldsmith, J., Greven, S., & Crainiceanu, C. (2013). Corrected confidence bands for functional data using principal components. *Biometrics*, *69*(1), 41–51.

Goldsmith, J., Scheipl, F., Huang, L., Wrobel, J., Gellar, J., Harezlak, J., McLean, M. W., Swihart, B., Xiao, L., Crainiceanu, C., & Reiss, P. T. (2016). *refund: Regression with Functional Data*. R package version 0.1-16.
URL `https://CRAN.R-project.org/package=refund`

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

Hastie, T., & Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological)*, (pp. 757–796).

Howe, L. D., Tilling, K., Benfield, L., Logue, J., Sattar, N., Ness, A. R., Smith, G. D., & Lawlor, D. A. (2010). Changes in ponderal index and body mass index across childhood and their associations with fat mass and cardiovascular risk factors at age 15. *PloS one*, *5*(12), e15186.

Huang, J. Z., Wu, C. O., & Zhou, L. (2002). Varying-coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika*, (pp. 111–128).

Kim, J., Frey, W. D., He, H., Kim, H., Ekram, M. B., Bakshi, A., Faisal, M., Perera, B. P., Ye, A., & Teruyama, R. (2013). Peg3 mutational effects on reproduction and placenta-specific gene families. *PloS one*, *8*(12), e83359.

Li, Y., Xie, C., Murphy, S. K., Skaar, D., Nye, M., Vidal, A. C., Cecil, K. M., Dietrich, K. N., Puga, A., Jirtle, R. L., et al. (2016). Lead exposure during early human development and dna methylation of imprinted gene regulatory elements in adulthood. *Environmental health perspectives*, *124*(5), 666.

Meas, T., Deghmoun, S., Armoogum, P., Alberti, C., & Levy-Marchal, C. (2008). Consequences of being born small for gestational age on body composition: an 8-year follow-up study. *The Journal of Clinical Endocrinology & Metabolism*, *93*(10), 3804–3809.

Morris, J. S. (2015). Functional regression. *Annual Review of Statistics and Its Application*, *2*, 321–359.

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

Ollikainen, M., & Craig, J. M. (2011). Epigenetic discordance at imprinting control regions in twins. *Epigenomics*, *3*(3), 295–306.

Panik, M. J. (2014). *Growth curve modeling: theory and applications*. John Wiley & Sons.

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

Qu, L. (2017). *varComp: Variance Component Models*. R package version 0.2-0.
URL `https://CRAN.R-project.org/package=varComp`

Qu, L., Guennel, T., & Marshall, S. L. (2013). Linear score tests for variance components

in linear mixed models and applications to genetic association studies. *Biometrics*, *69*(4), 883–892.

Qu, L., & Qu, M. L. (2015). Package varcomp.

Ramsay, J., Munhall, K. G., Gracco, V. L., & Ostry, D. J. (1996). Functional data analyses of lip motion. *The Journal of the Acoustical Society of America*, *99*(6), 3718–3727.

Ramsay, J. O. (2006). *Functional data analysis*. Wiley Online Library.

Rasmussen, C. E. (2006). Gaussian processes for machine learning.

Scheipl, F., Gertheiss, J., Greven, S., et al. (2016). Generalized functional additive mixed models. *Electronic Journal of Statistics*, *10*(1), 1455–1492.

Scheipl, F., Staicu, A.-M., & Greven, S. (2015). Functional additive mixed models. *Journal of Computational and Graphical Statistics*, *24*(2), 477–501.

Shi, J. Q., & Choi, T. (2011). *Gaussian process regression analysis for functional data*. CRC Press Boca Raton, FL:.

Skaar, D. A., Li, Y., Bernal, A. J., Hoyo, C., Murphy, S. K., & Jirtle, R. L. (2012). The human imprintome: Regulatory mechanisms, methods of ascertainment, and roles in disease susceptibility. *ILAR Journal*, *53*(3-4), 341–358.
URL + http://dx.doi.org/10.1093/ilar.53.3-4.341

Soubry, A., Murphy, S., Wang, F., Huang, Z., Vidal, A., Fuemmeler, B., Kurtzberg, J., Murtha, A., Jirtle, R., Schildkraut, J., et al. (2015). Newborns of obese parents have altered dna methylation patterns at imprinted genes. *International journal of obesity*, *39*(4), 650–657.

Soubry, A., Schildkraut, J. M., Murtha, A., Wang, F., Huang, Z., Bernal, A., Kurtzberg, J., Jirtle, R. L., Murphy, S. K., & Hoyo, C. (2013). Paternal obesity is associated with igf2 hypomethylation in newborns: results from a newborn epigenetics study (nest) cohort. *BMC medicine*, *11*(1), 29.

Staniswalis, J. G., & Lee, J. J. (1998). Nonparametric regression analysis of longitudinal data. *Journal of the American Statistical Association*, *93*(444), 1403–1418.

Vidal, A. C., Semenova, V., Darrah, T., Vengosh, A., Huang, Z., King, K., Nye, M. D., Fry, R., Skaar, D., Maguire, R., et al. (2015). Maternal cadmium, iron and zinc levels, dna methylation and birth weight. *BMC Pharmacology and Toxicology*, *16*(1), 20.

Wand, M. P. (2003). Smoothing and mixed models. *Computational statistics*, *18*(2), 223–249.

Wang, J.-L., Chiou, J.-M., & Müller, H.-G. (2015a). Review of functional data analysis.

Wang, S., Song, J., Yang, Y., Zhang, Y., Wang, H., & Ma, J. (2015b). Hif3a dna methylation is associated with childhood obesity and alt. *PloS one*, *10*(12), e0145944.

Whincup, P. H., Kaye, S. J., Owen, C. G., Huxley, R., Cook, D. G., Anazawa, S., Barrett-Connor, E., Bhargava, S. K., Birgisdottir, B. E., Carlsson, S., et al. (2008). Birth weight and risk of type 2 diabetes: a systematic review. *Jama*, *300*(24), 2886–2897.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73*(1), 3–36.

Woodfine, K., Huddleston, J. E., & Murrell, A. (2011). Quantitative analysis of dna methylation at all human imprinted regions reveals preservation of epigenetic stability in adult somatic tissue. *Epigenetics & chromatin*, *4*(1), 1.

Wu, C. O., & Chiang, C.-T. (2000). Kernel smoothing on varying coefficient models with longitudinal dependent variable. *Statistica Sinica*, (pp. 433–456).

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

APPENDICES

Appendix A

Additional Data Analysis Results

Additional testing results from Chapter 3. All 9 DMRs are tested for each gender.

Table A.1: Testing results (p-values) from all DMRs by gender and kernel.

DMR	Gender	Linear	Quadratic	Gaussian
IGF2 CBS1	Male	0.1684	0.5112	0.1546
	Female	0.6595	0.2640	0.4031
IGF2 DMR	Male	0.9869	0.8814	0.8123
	Female	0.0825	0.1275	0.1724
MEG3 CBS	Male	0.9753	0.2700	0.5120
	Female	0.9692	0.5135	0.1837
MEG3 IG	Male	0.6010	0.6655	0.0197
	Female	0.6311	0.3486	0.7307
MESTIT1	Male	0.2250	0.1494	0.3598
	Female	0.0469	0.2823	0.2862
NNAT	Male	0.6198	0.7832	0.7075
	Female	0.8632	0.3609	0.5585
PEG3	Male	0.7101	0.9487	0.9769
	Female	0.0703	0.0132	0.4932
SGCE	Male	0.2821	0.4470	0.5135
	Female	0.4662	0.7700	0.7204
ZAC	Male	0.3341	0.1607	0.4160
	Female	0.2534	0.4694	0.4071

Appendix B

B.1 Additional Power Simulations

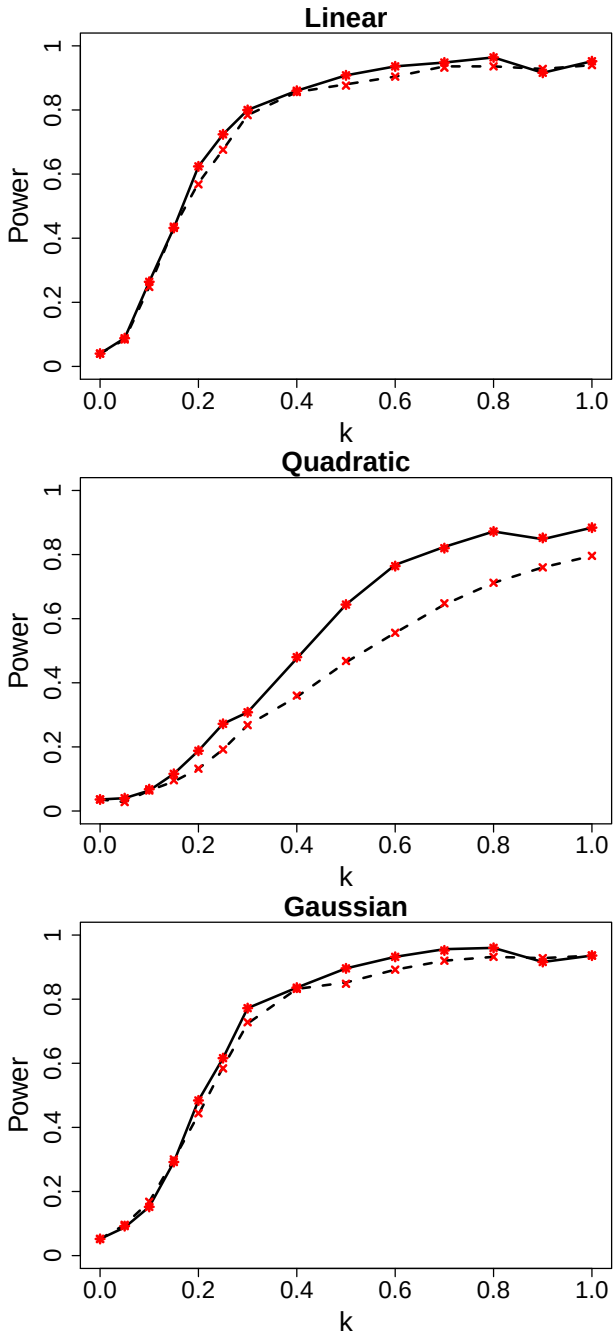


Figure B.1: Power curves for each combination of covariance kernel, methylation summary setting, and method. The dotted lines and crosses represent $\boldsymbol{m}_{\text{sim}}^{1^*}$, the linear relationship, while the solid lines and stars represent $\boldsymbol{m}_{\text{sim}}^{2^*}$, the quadratic relationship. The black lines represent the VarCompGP method with equal weights, while the red points represent the GamGP method.

B.2 Additonal Timing Results

The timing results from Chapter 4 prior to aggregating over kernel and methylation summary.

Table B.1: Timing Results when B=1 [Mean(SD, N)].

Methylation Summary	Kernel	GamGP	VarCompGP
Mean Summary	Linear	25.6(8.5, 191)	160.82(31.47, 191)
	Quadratic	21.04(5.98, 191)	173.48(52.15, 191)
	Gaussian	20.97(6.06, 191)	161.05(46.54, 191)
SS Summary	Linear	20.73(5.59, 201)	170.06(51.75, 201)
	Quadratic	21(6.15, 201)	175.16(52.75, 201)
	Gaussian	21.05(6.03, 201)	160.19(47.13, 201)

Table B.2: Timing Results when B=2 [Mean(SD, N)].

Methylation Summary	Kernel	GamGP	VarCompGP
Mean Summary	Linear	55.05(18.64, 58)	274.18(59.87, 58)
	Quadratic	49.37(13.66, 58)	290.39(90.23, 58)
	Gaussian	49.08(13.83, 58)	276.17(81.23, 58)
SS Summary	Linear	47.65(12.92, 48)	289.88(89.3, 48)
	Quadratic	49.35(14.5, 48)	293.29(92.16, 48)
	Gaussian	49.06(14.29, 48)	277.89(88.14, 48)

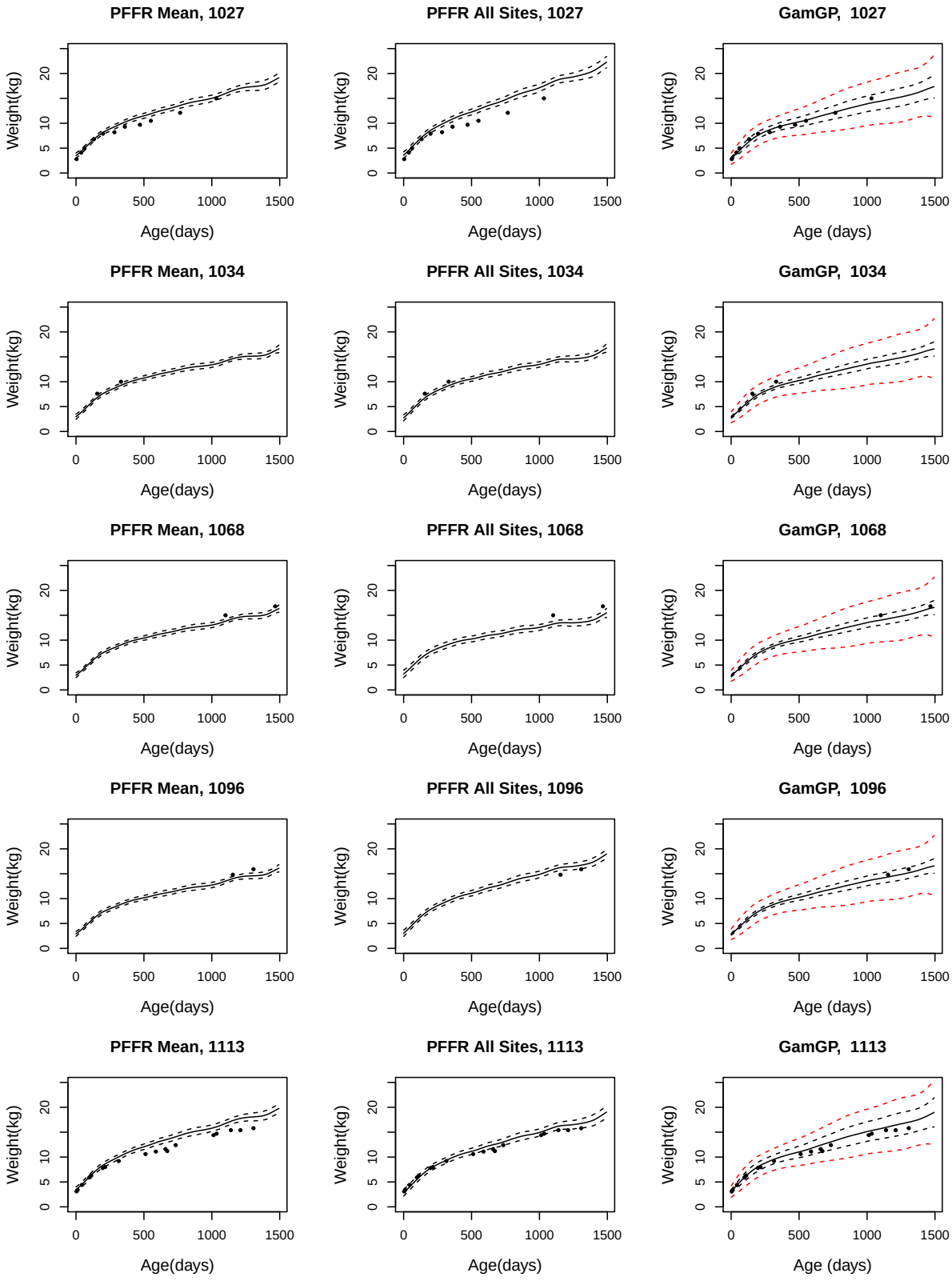
Table B.3: Timing Results when B=3 [Mean(SD, N)].

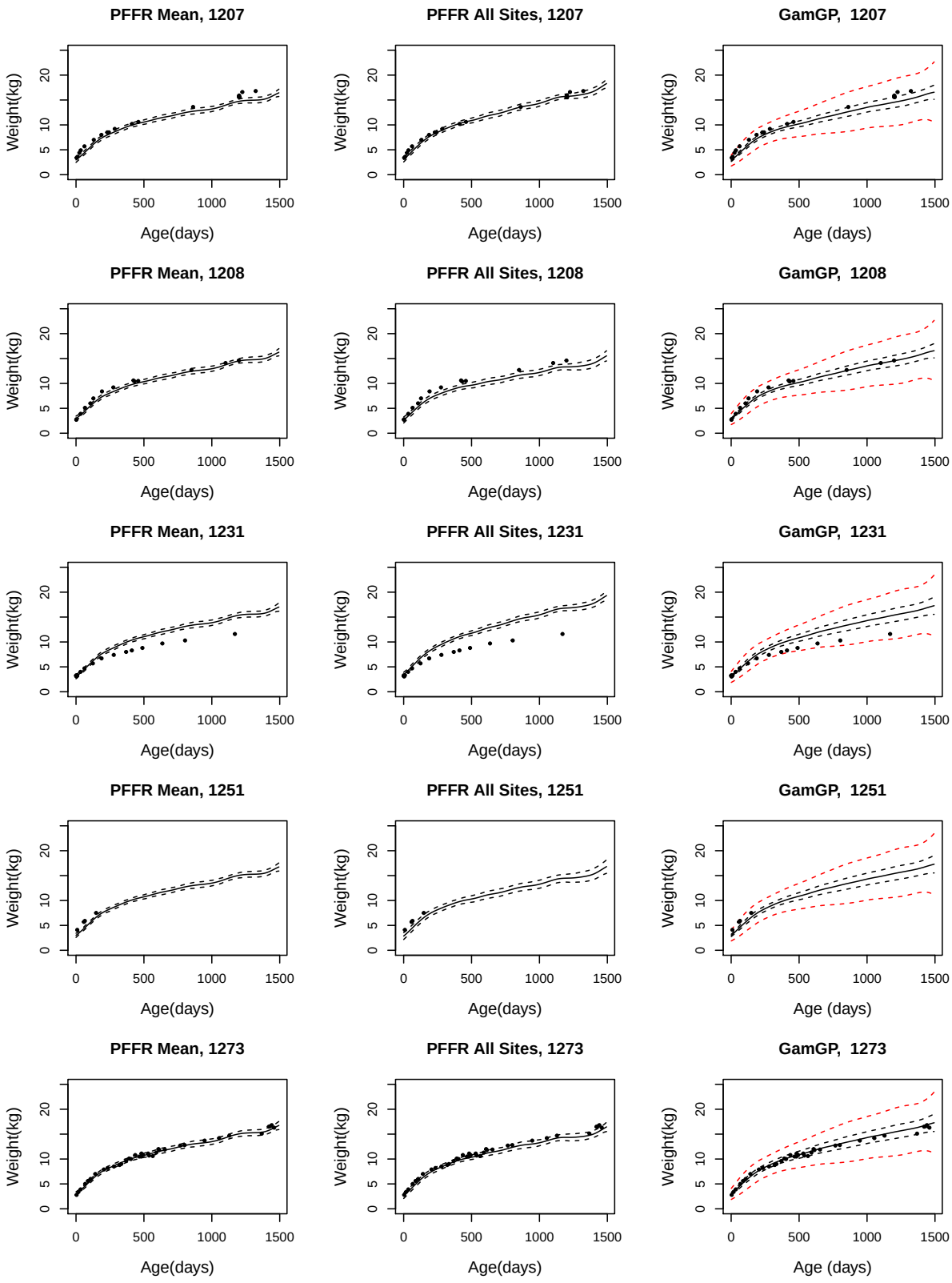
Methylation Summary	Kernel	GamGP	VarCompGP
Mean Summary	Linear	113.59(NA, 1)	307.87(NA, 1)
	Quadratic	91.08(NA, 1)	437.14(NA, 1)
	Gaussian	93.7(NA, 1)	309.76(NA, 1)
SS Summary	Linear	76.11(NA, 1)	320.77(NA, 1)
	Quadratic	74.61(NA, 1)	412.41(NA, 1)
	Gaussian	72.82(NA, 1)	338.54(NA, 1)

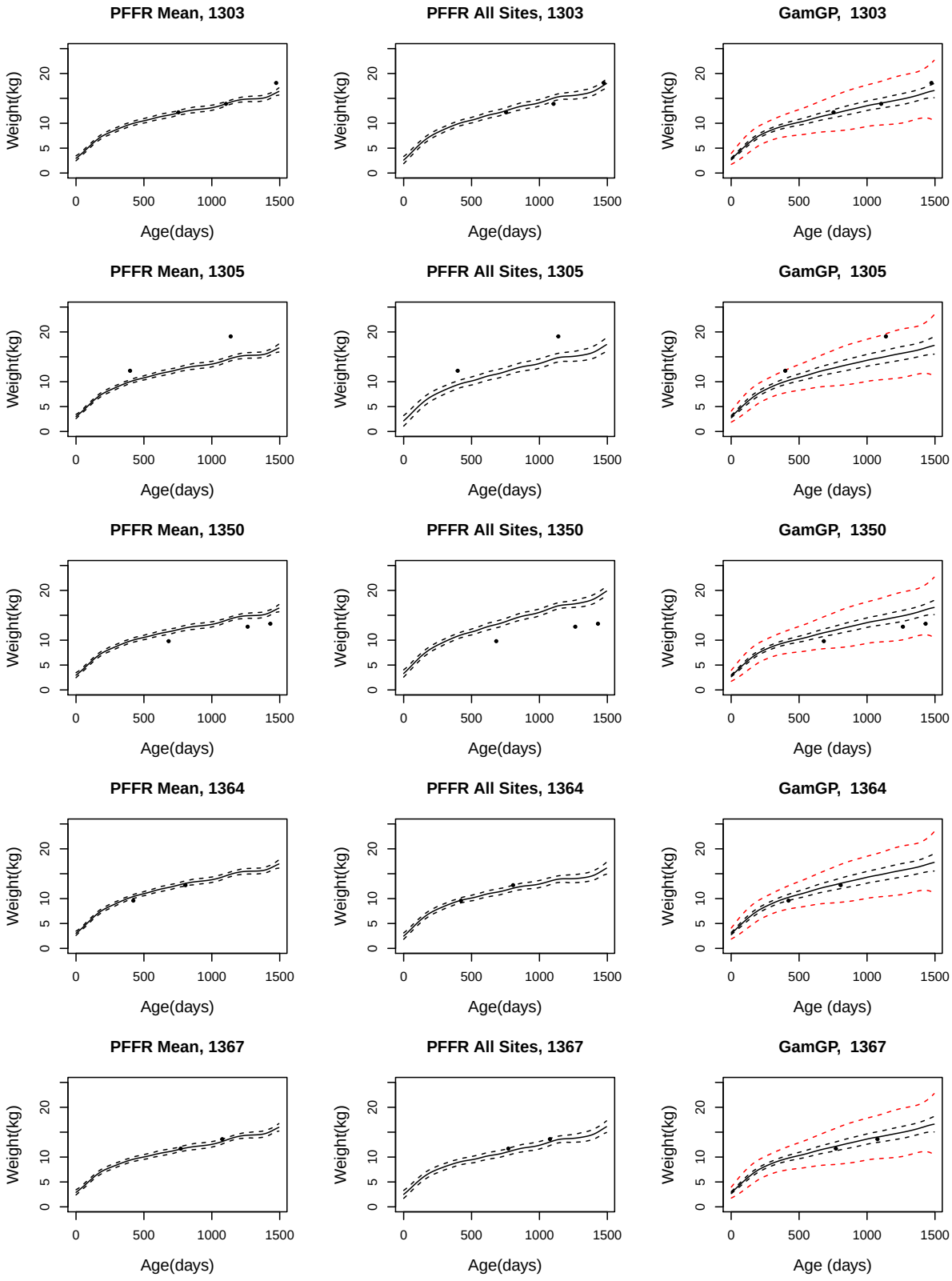
Appendix C

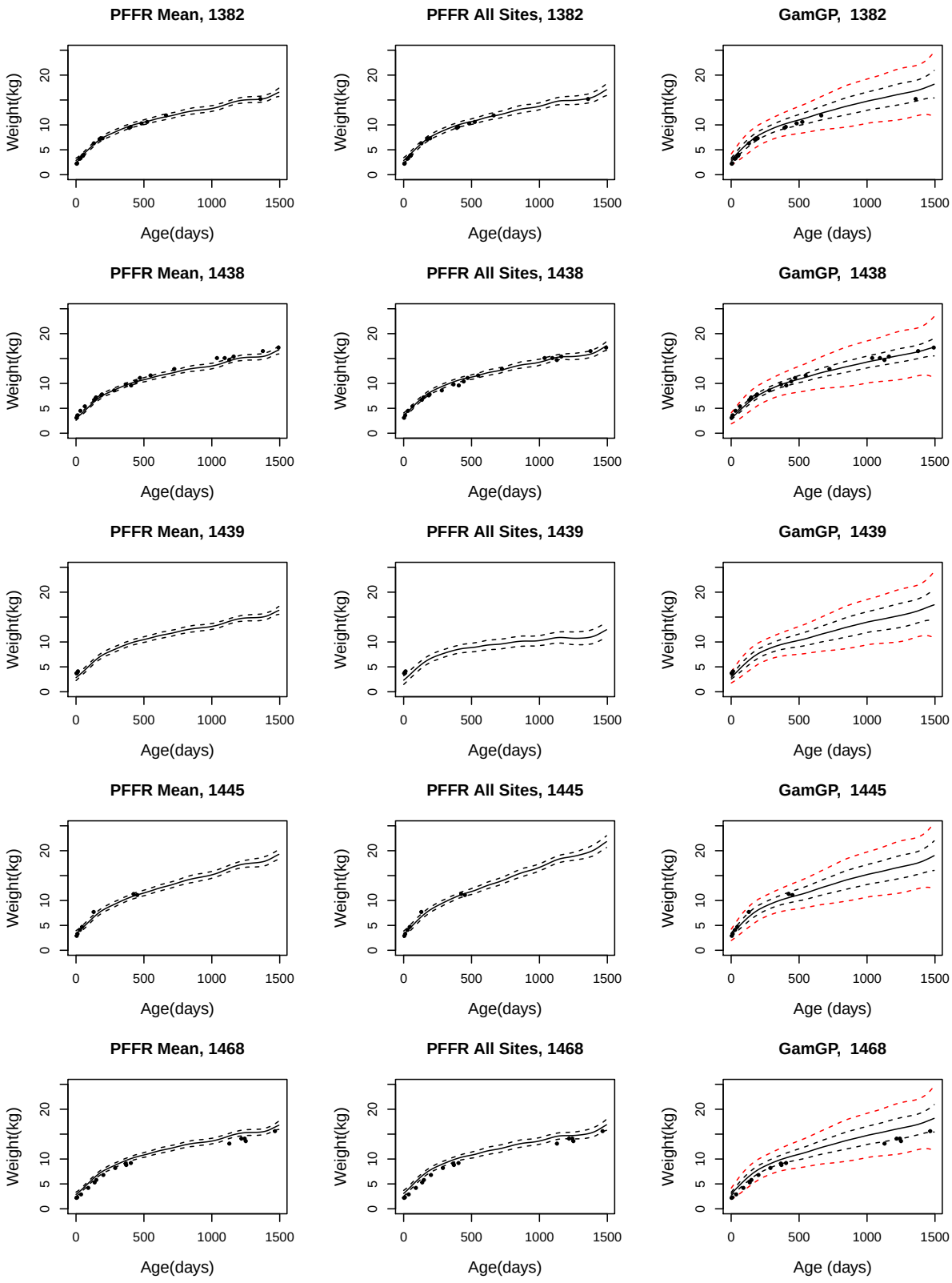
Prediction Results

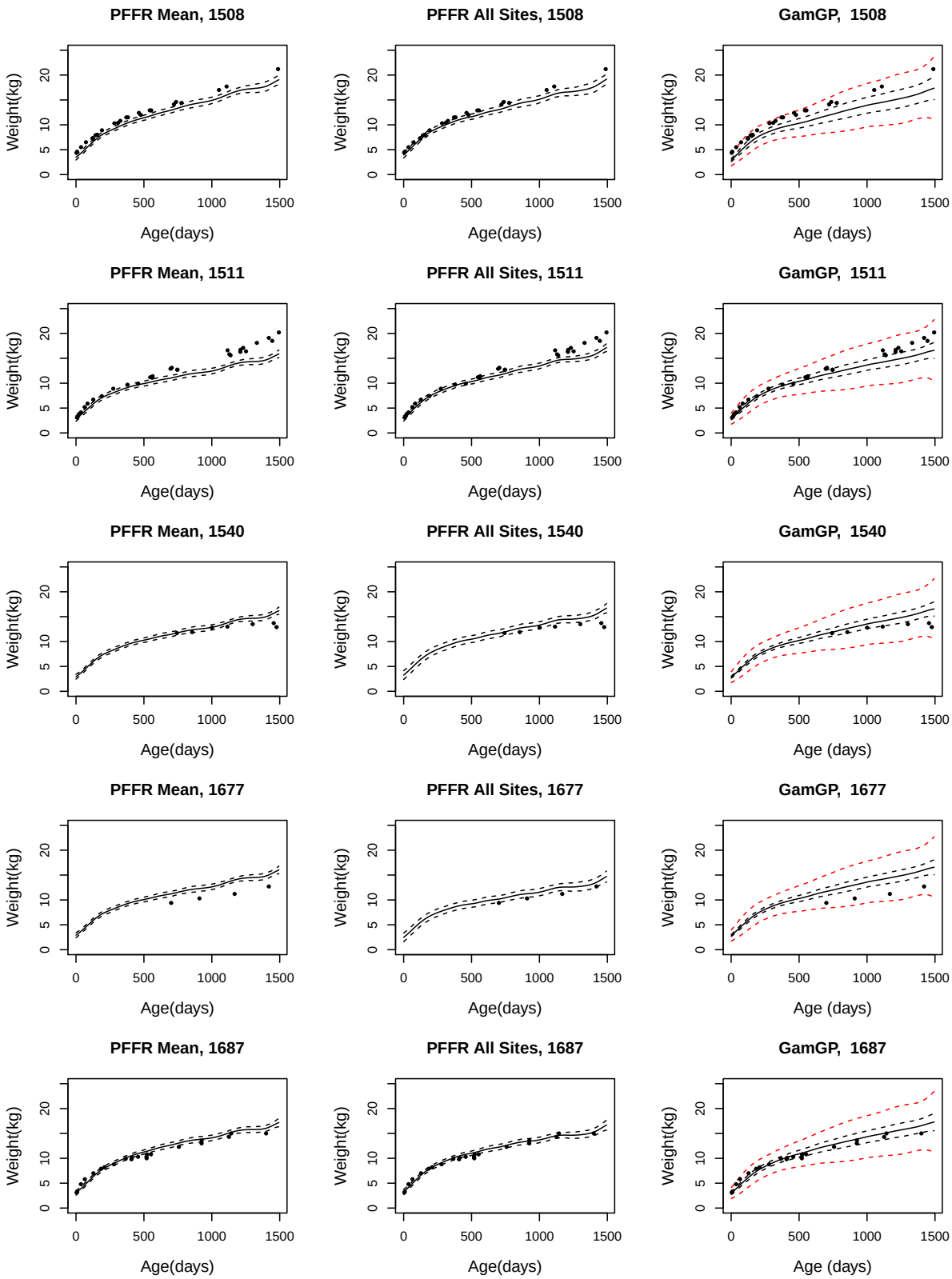
Prediction results from Chapter 5. In the GamGP plots, the black dotted lines represent confidence bands constructed from the standard variance estimates, while the red dotted lines represent the prediction bands constructed from the prediction variance estimates.

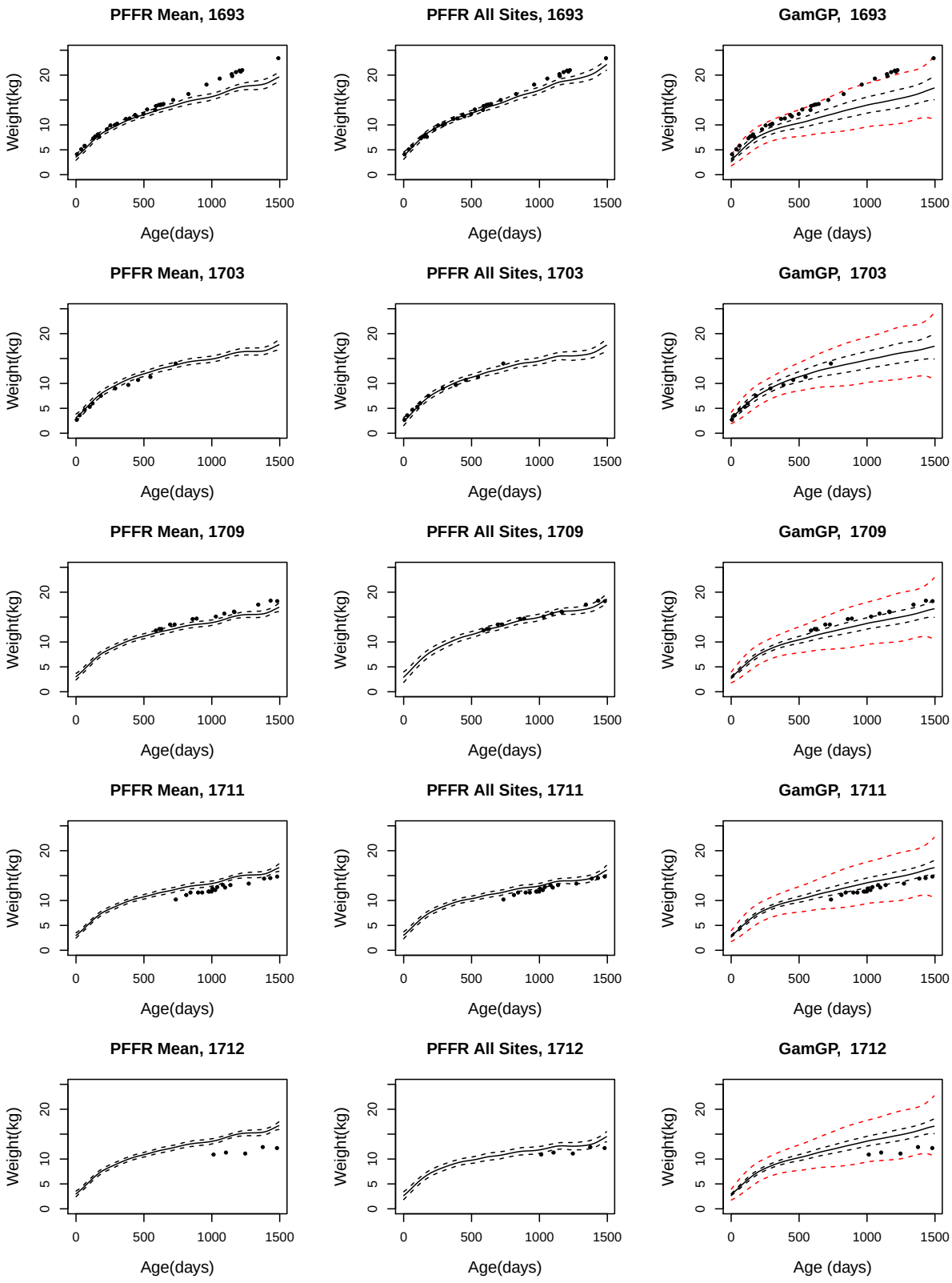

PFFR Mean, 1027
PFFR All Sites, 1027
GamGP, 1027
PFFR Mean, 1034
PFFR All Sites, 1034
GamGP, 1034
PFFR Mean, 1068
PFFR All Sites, 1068
GamGP, 1068
PFFR Mean, 1096
PFFR All Sites, 1096
GamGP, 1096
PFFR Mean, 1113
PFFR All Sites, 1113
GamGP, 1113
Weight(kg)
0
5
10
20
0
500
1000
1500
Age(days)
Age (days)

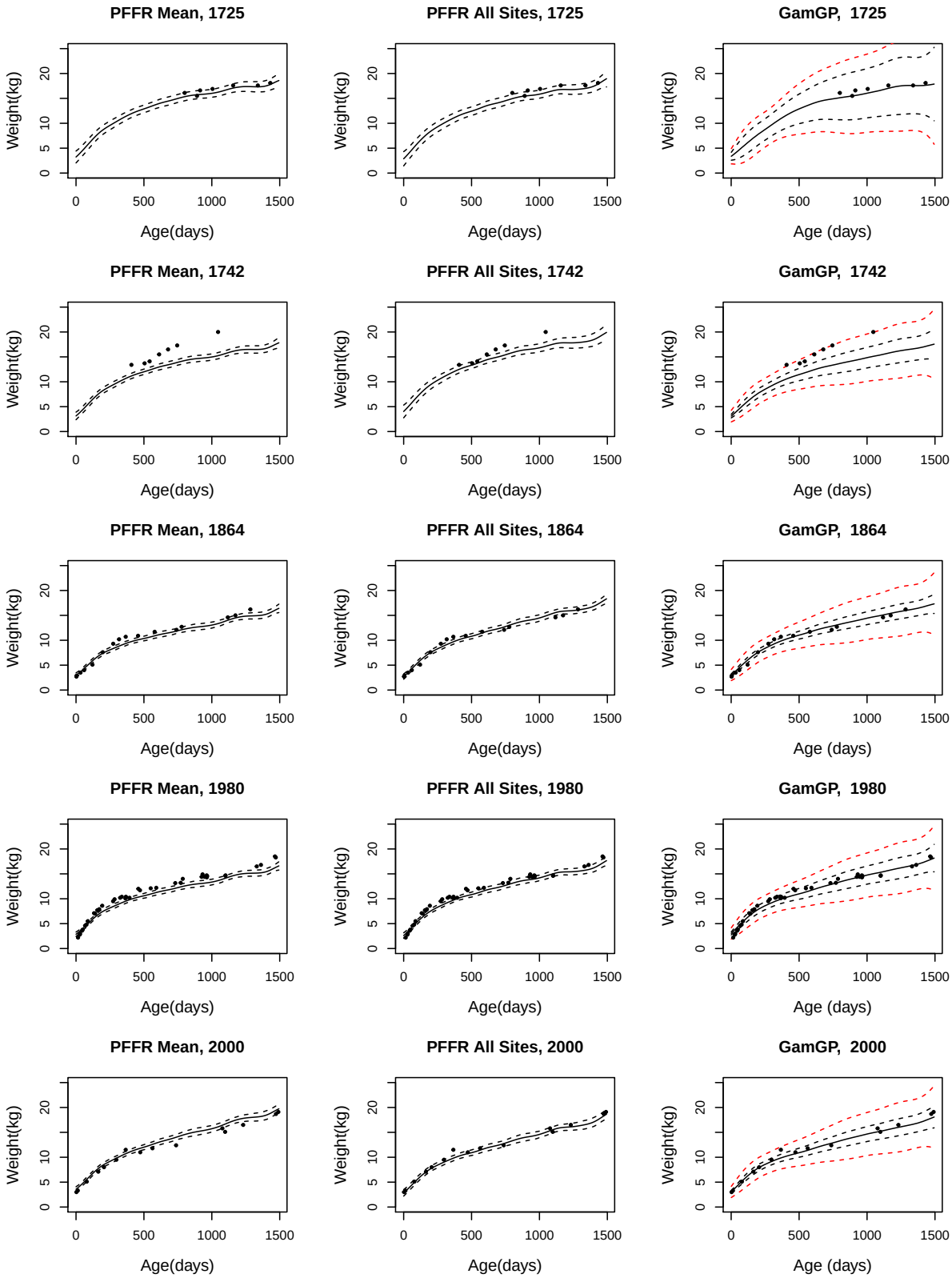

PFFR Mean, 1207
PFFR All Sites, 1207
GamGP, 1207
PFFR Mean, 1208
PFFR All Sites, 1208
GamGP, 1208
PFFR Mean, 1231
PFFR All Sites, 1231
GamGP, 1231
PFFR Mean, 1251
PFFR All Sites, 1251
GamGP, 1251
PFFR Mean, 1273
PFFR All Sites, 1273
GamGP, 1273
Weight(kg)
0
5
10
20
0
500
1000
1500
Age(days)
Age (days)

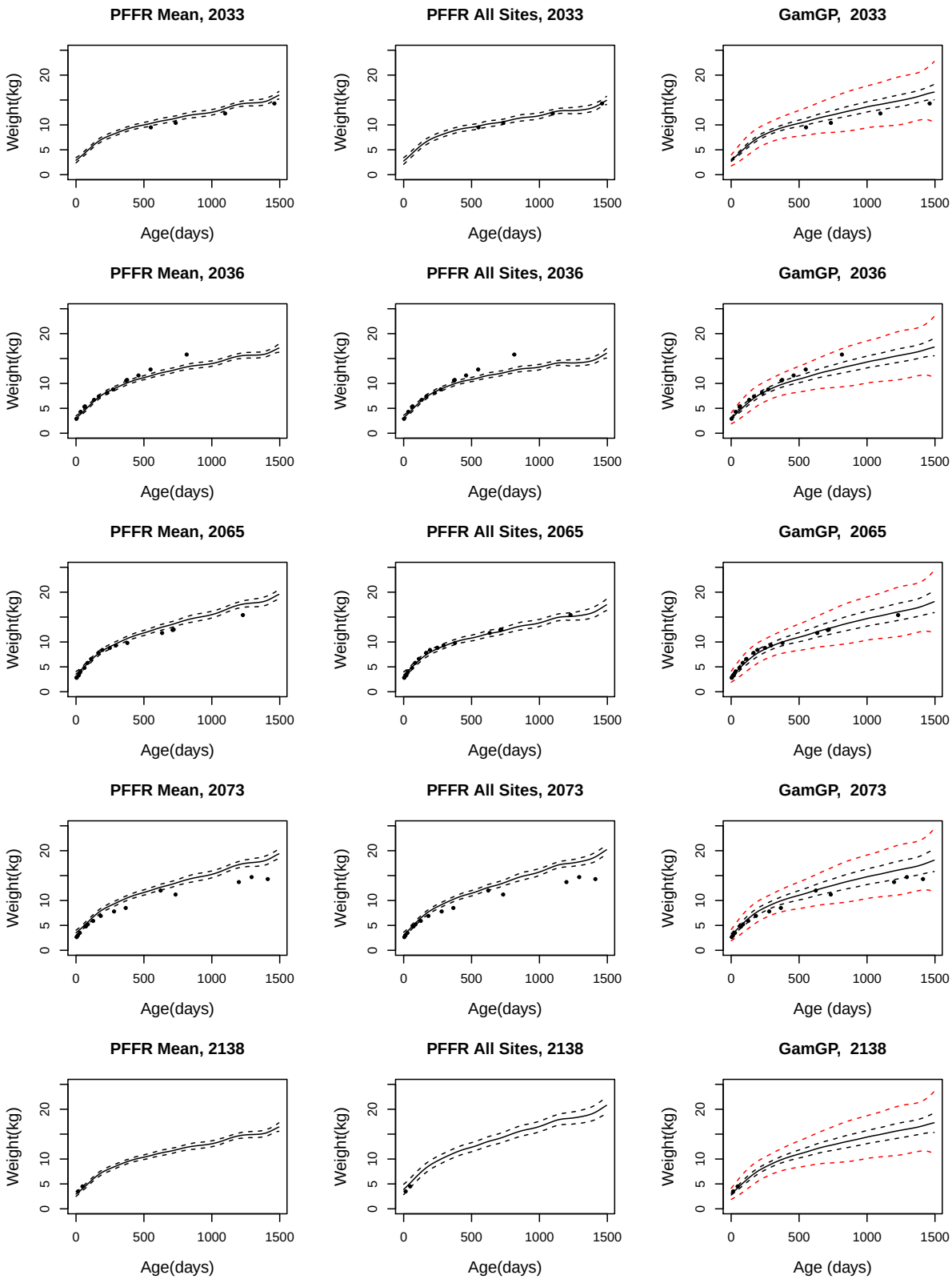

PFFR Mean, 1303
PFFR All Sites, 1303
GamGP, 1303
PFFR Mean, 1305
PFFR All Sites, 1305
GamGP, 1305
PFFR Mean, 1350
PFFR All Sites, 1350
GamGP, 1350
PFFR Mean, 1364
PFFR All Sites, 1364
GamGP, 1364
PFFR Mean, 1367
PFFR All Sites, 1367
GamGP, 1367
Weight(kg)
0
5
10
20
0
500
1000
1500
Age(days)
Age (days)

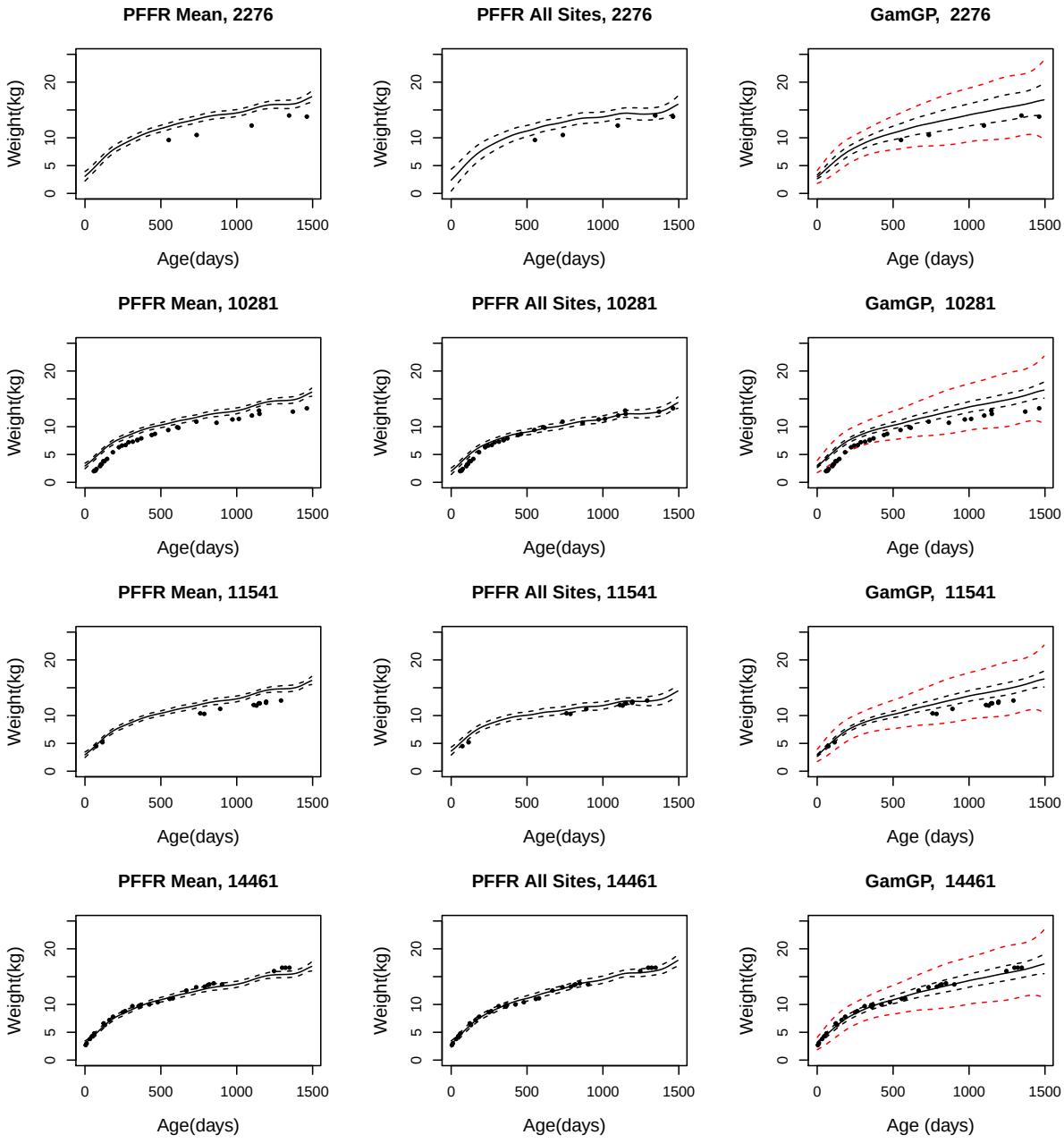

PFFR Mean, 1382
PFFR All Sites, 1382
GamGP, 1382
PFFR Mean, 1438
PFFR All Sites, 1438
GamGP, 1438
PFFR Mean, 1439
PFFR All Sites, 1439
GamGP, 1439
PFFR Mean, 1445
PFFR All Sites, 1445
GamGP, 1445
PFFR Mean, 1468
PFFR All Sites, 1468
GamGP, 1468
Weight(kg)
0
5
10
20
0
500
1000
1500
Age(days)
Age (days)


PFFR Mean, 1508
PFFR All Sites, 1508
GamGP, 1508
PFFR Mean, 1511
PFFR All Sites, 1511
GamGP, 1511
PFFR Mean, 1540
PFFR All Sites, 1540
GamGP, 1540
PFFR Mean, 1677
PFFR All Sites, 1677
GamGP, 1677
PFFR Mean, 1687
PFFR All Sites, 1687
GamGP, 1687
Weight(kg)
0
5
10
20
0
500
1000
1500
Age(days)
Age (days)


PFFR Mean, 1693
PFFR All Sites, 1693
GamGP, 1693
PFFR Mean, 1703
PFFR All Sites, 1703
GamGP, 1703
PFFR Mean, 1709
PFFR All Sites, 1709
GamGP, 1709
PFFR Mean, 1711
PFFR All Sites, 1711
GamGP, 1711
PFFR Mean, 1712
PFFR All Sites, 1712
GamGP, 1712
Weight(kg)
0
5
10
20
0
500
1000
1500
Age(days)
Age (days)


PFFR Mean, 1725
PFFR All Sites, 1725
GamGP, 1725
PFFR Mean, 1742
PFFR All Sites, 1742
GamGP, 1742
PFFR Mean, 1864
PFFR All Sites, 1864
GamGP, 1864
PFFR Mean, 1980
PFFR All Sites, 1980
GamGP, 1980
PFFR Mean, 2000
PFFR All Sites, 2000
GamGP, 2000
Weight(kg)
0
5
10
20
0
500
1000
1500
Age(days)
Age (days)


PFFR Mean, 2033
PFFR All Sites, 2033
GamGP, 2033
PFFR Mean, 2036
PFFR All Sites, 2036
GamGP, 2036
PFFR Mean, 2065
PFFR All Sites, 2065
GamGP, 2065
PFFR Mean, 2073
PFFR All Sites, 2073
GamGP, 2073
PFFR Mean, 2138
PFFR All Sites, 2138
GamGP, 2138
Weight(kg)
0
5
10
20
0
500
1000
1500
Age(days)
Age (days)


PFFR Mean, 2276
PFFR All Sites, 2276
GamGP, 2276
PFFR Mean, 10281
PFFR All Sites, 10281
GamGP, 10281
PFFR Mean, 11541
PFFR All Sites, 11541
GamGP, 11541
PFFR Mean, 14461
PFFR All Sites, 14461
GamGP, 14461
Weight(kg)
0
5
10
20
0
500
1000
1500
Age(days)
Age (days)