ABSTRACT

COLES, ADRIAN LAMONT. New Approaches to Conducting Inference in Nonlinear Functional Regression Models with Novel Applications to Copy Number Data. (Under the direction of Arnab Maity.)

In recent years, researchers in many areas of science have embraced the assumption that several types of observed data are realizations of an underlying smooth process. Given this paradigm shift, functional data analysis techniques have grown in popularity. In particular, many researchers elect to frame their scientific problems as functional regression models, where often times the interest is in regressing scalar responses onto functional covariates. Estimating the functional effect in these types of regression models are well-developed in the literature. However, procedures that test whether the functional covariate is necessary in the regression model have received far less attention–especially when the relationship between the random process and the outcome are believed to be nonlinear. In this dissertation, we propose three new approaches to testing for the effect of nonlinear functional covariates on scalar outcomes. Each approach shares a common framework that connects the complex functional model to its corresponding mixed model framework.

Chapter 2 presents the nonlinear functional regression model. This model regresses scalar and continuous outcomes onto a functional covariate while adjusting for scalar covariates such as height and weight. The scalar covariates are modeled parametrically and the functional covariate is modeled nonparametrically. We develop testing procedures to investigate nonlinear functional effects as well as estimation procedures to explore the magnitude and direction of the functional effect. We investigate the finite-sample performance of the proposed procedures via simulations, and we apply the procedures to conduct a genome-wide copy number association analysis in multiple myeloma patients. The response in our analysis is the level of a prognostic marker used to determine disease progression.

Chapter 3 presents the generalized nonlinear functional regression model. This model extends the nonlinear functional regression model to the case of non-normal outcomes. We develop testing procedures to explore the relationship between the functional covariate and the outcome. To help create a comprehensive testing procedure, we develop an adaptive composite kernel which serves as the covariance structure in a mixed model representation of the functional model. Simulation results show that the proposed testing procedure performs well when the functional effect is linear and when the functional effect is nonlinear. The model is used to determine which local regions of copy number alteration are related to the progression of multiple myeloma. In this chapter, disease progression is determined by dichotomizing cancer

stages.

Chapter 4 presents the functional nonlinear Cox proportional hazards model. Again, the core methodology in Chapter 2 is extended to the case of censored survival times. We develop a main effects model that includes scalar covariates and a functional covariate, and we also develop an interaction model that captures the interaction between a single scalar covariate and a functional covariate. Both models are reduced to simple random effects models whereby testing procedures for both models are proposed. Our numerical experiments show that both testing procedures perform well on finite samples. We apply both models to investigate the effect of eight genes in the glioblastoma multiforme pathway on survival times in patients that have been diagnosed with this aggressive form of brain cancer.

New Approaches to Conducting Inference in Nonlinear Functional Regression Models with Novel Applications to Copy Number Data

by
Adrian Lamont Coles

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

2014

APPROVED BY:

Jung-Ying Tzeng	Donald Martin
Eric Laber	Arnab Maity Chair of Advisory Committee

DEDICATION

To Sylvia, Abena and Tameka.

BIOGRAPHY

Adrian Coles was born in Danville, VA to Sylvia and Lansing Coles. He graduated from George Washington High School in 1996, and he spent the first eight and half years of his adult life in the United States Marine Corps. After his time in the military, he attended the University of North Carolina Wilmington and graduated cum laude and with honors in 2010 with a B.A. in Mathematics. He joined the Department of Statistics at North Carolina State University in 2010 and completed a Master's Degree in Statistics in 2012.

ACKNOWLEDGEMENTS

First, I want to thank my advisor Arnab Maity. I came into his office during the second year of my program and asked to "read" with him in this exciting area of functional data analysis. Who knew that those few initial papers that he gave me to read would turn into the three research projects contained in my dissertation, a separate joint work with one of his other students, and a fifth work with our collaborators from The University of Texas M.D. Anderson Cancer Center? I'm truly grateful for the opportunity that he provided me. Second, I want to acknowledge one of my collaborators/mentors Veera Baladandayuthapani. The conversations that we had when I spent the summer of 2013 at M.D. Anderson really helped to mold me as a young researcher. He provided me with my first graduate level research experience outside of my advisor's safety net which helped me to establish my confidence as an "application driven methodologist".

Next, I want to acknowledge my committee members Jung-Ying Tzeng, Donald Martin and Eric Laber. The conversations that I have had with each of these professors have been very helpful along the way. In addition, I must thank Renee Moore for being an excellent mentor. From undergrad, I want to thank my research advisor Michael Freeze and my very first statistics professor Susan Simmons. Dr. Freeze guided me through my first academic research experience, and the work that we did on my honor's project paved the way for all that I've accomplished during my four years in graduate school. Dr. Simmons encouraged me to pursue statistics on the graduate level. Her actions helped me to find my passion and have undoubtedly changed my life for the better.

Last but not least, I want to thank my family and friends. Your love, support and patience enabled me to reach this goal. During the times that I was weary, your encouragement powered me through. Any success that has been achieved is equally shared among each of you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

The statistical methodologies contained in this dissertation represent a point of synergy between two bodies of statistical literature: (1) functional data analysis and (2) kernel machine regression. While we develop models that have a broad appeal, we are motivated by a desire to solve three challenges within the area of cancer genomics. Thus throughout this work, we discuss our models in the context of these problems. The aim of this chapter is to provide brief introductions into both bodies of statistical literature, as well as a brief introduction to copy number aberrations, which is at the core of the cancer genomics problems for which we offer three novel solutions.

1.1 Functional Regression

The observable data from several types of scientific phenomena often arise from unknown smooth processes. Data such as these can be referred to as functional data. One approach to analyzing such data is to consider the finite collection of observations from an underlying smooth process as multidimensional data. However, this approach can discard valuable information about the actual process that generated the observed data. This can make it difficult to gain deep insight into the true nature of the scientific phenomena. In general, functional data analysis is a sub-discipline of statistics that extends multivariate statistical analysis techniques to allow investigators to extract more information about underlying smooth processes than what is often available when considering a finite set of realizations as vector-valued data.

There are several types of functional data methodologies that target a wide range of goals. For example, some methods simply seek to describe central tendencies and variation across a set of random curves, while other methods seek to estimate the smooth curves that yield finite collections of observations. Another common goal within functional data analysis is to

explain the variation in a response variable (functional or non-functional) through independent functional and/or non-functional covariates. The research presented here falls into this latter category, where we propose three new functional regression models. Our primary goal is to determine if a single functional covariate is necessary to explain the variation in scalar outcomes when the effect of the functional covariate is assumed to be complex and nonlinear.

The nonlinear functional regression models developed in this work share common challenges and strategies as many previously developed functional regression models. For instance, the functional covariate is assumed to be observed on an infinite-dimensional domain. In practice, this induces sparseness in observed data of any finite size. Thus, one common theme throughout Chapters 2, 3, and 4 is to reduce the functional regression model to a parsimonious finite form while minimizing loss of information about the true curves. For heuristic motivation, consider the functional linear model (FLM) introduced by Ramsay and Dalzell [1991] and popularized by Ramsay and Silverman [2005],

$$Y_i = \alpha + \int_{\mathcal{T}} X_i(t)\beta(t)\, dt + \epsilon_i,\ t \in \mathcal{T}.$$

In this equation, $X_i(\cdot)$ is a smooth function, $\beta(\cdot)$ models the effect of $X_i(\cdot)$ on Y_i, and ϵ_i are independent mean zero random errors with finite variance. This model assumes that $X_i(\cdot)$ and $\beta(\cdot)$ are square integrable, and that the relationship between the response and the function are linear. The basic idea is to use an orthonormal basis of $L^2[\mathcal{T}]$ to yield the following basis expansions, $X_i(t) = \sum_{j=1}^{\infty} \xi_{ij}\phi_j(t)$, $\beta(t) = \sum_{j=1}^{\infty} \beta_j^{\star}\phi_j(t)$, where $\{\phi_j(\cdot), j \geq 1\}$ is an orthonormal basis. Truncating to J appropriately selected basis functions yields the following truncated model,

$$Y_i = \alpha + \sum_{j=1}^{J} \xi_{ij}\beta_j^{\star} + \epsilon_i. \tag{1.1}$$

Conditional on J and $\{\xi_{ij}\}_{j=1}^{J}$ and assuming normality for the errors, Eq. (1.1) is a multiple linear regression model. This reduced dimension model is used to investigate the relationship between Y_i and $X(\cdot)$. We emphasis here that connecting complex functional models to well-known and simpler frameworks is a common strategy within the literature, and it's a strategy that we employ in the methods developed in this work. We offer more details about such connections in subsequent chapters.

1.2 Kernel Machine Regression

Kernel machine regression has recently emerged as a powerful solution to several challenges that exist within traditional regression models such as modeling high-dimensional data. It is also able to reduce the computational burden that stems from constructing models where the covariates have complex relationships between themselves and the outcome. In general, kernel machine regression is a subset of the larger machine learning literature, where the focus is on regressing outcomes onto the features of a complex space through simple terms in the model space. The features of the complex space are obtained via a kernel function, which implicitly determines (1) the relationship between the multidimensional covariates across subjects, and (2) the relationship between the multidimensional covariates and the outcome. The selection of an appropriate kernel function is important due to the multiple roles that it serves in the regression model. A poor selection of kernel function can cause a reduction in power for associated tests.

To formulate these ideas, we present a comparison between a standard quadratic regression model and its kernel machine equivalent. This comparison is offered only as a learning aid, thus we omit the technical details in our introduction. Consider the following observed data $\{Y_i, \mathbf{x}_i\}_{i=1}^n$, where Y_i is a continuous outcome and $\mathbf{x}_i$ is a p-dimensional vector of covariates. Assume that one desires to fit the following regression model that contains linear terms, quadratic terms and all two way interactions,

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \sum_{l=1}^{p} \sum_{k=1}^{l-1} \theta_{lk} x_l x_k + \sum_{j=1}^{p} \gamma_j x_{ij}^2 + \epsilon_i. \tag{1.2}$$

If the total number of terms is much greater than the size of the sample, then sparseness is induced in the number of available observations making it difficult to fit this model using standard techniques. In contrast, the equivalent kernel machine regression model has the simple form

$$Y_i = \beta_0 + \boldsymbol{\alpha}^{\mathrm{T}} \mathbf{K}_i + \epsilon_i, \tag{1.3}$$

where $\boldsymbol{\alpha}$ is an n-dimensional vector of effects, and $\mathbf{K}_i = [K_{i1}, \ldots, K_{in}]^{\mathrm{T}}$ with $K_{lk} = K(\mathbf{x}_l, \mathbf{x}_k) = (\mathbf{x}_l^{\mathrm{T}} \mathbf{x}_k + 1)^2$. It's easy to see the advantage of such a framework when considering that the number of parameters in Eq. (1.2) can easily exceed n. In addition, the kernel machine framework makes it easy to model complex effects, such as the fraction of alleles shared by any two individuals in the sample which is applicable to genetic studies. In fact, the kernel used for this application is the IBS kernel $K(\mathbf{x}_l, \mathbf{x}_k) = (2p)^{-1} \sum_{j=1}^{p} (2 - |x_{lj} - x_{kj}|)$. We refer the reader to Wessel and Schork [2006] and Wu et al. [2013] for more details about this kernel.

These methods have broad appeal; but most notably, they have experienced a great deal

of recent success in statistical genetics and genomics applications. As suggested above, this is because they are well-adapted to model the complex relationships between genetic markers and disease outcomes as well as they are able to handle the high-dimensionality of genetic problems [Liu et al., 2007, Kwee et al., 2008, Wu et al., 2010, 2011]. Given the nature of our motivating problem in cancer genomics, extending such methods into the functional data literature only adds to the utility of kernel-based methods in these areas of scientific research. We provide more of the inner workings of these methods in later chapters as we build upon them to develop powerful tests for nonlinear functional effects on scalar responses.

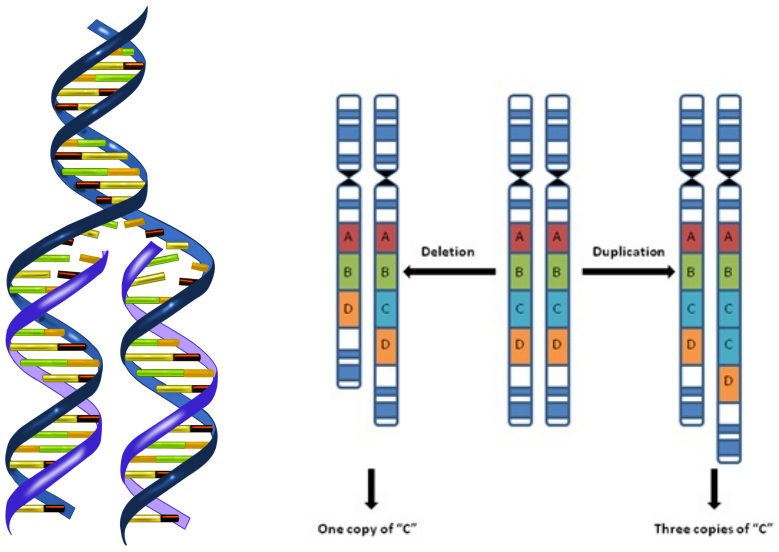
1.3 Copy Number Aberration

In this section, we briefly introduce the form of genetic variation referred to as copy number aberration. In general, there are two classes of variation in the human genome, sequence variation and structural variation. There are several subclasses of structural variation, but we focus our attention to copy number aberrations (CNAs). CNAs are a type of structural variation that occur when a section of a DNA molecule has more (or fewer) copies of genetic material when compared to a reference sample [Sun et al., 2009]. We note that the biological definition of CNA includes more phenomena than quantitative structural variations; however, for the purposes of this work, we consider only gains and deletions of genetic material that can be quantified.

To aid in understanding, consider Figure 1.1. Panel (a) illustrates the process of synthesizing two DNA molecules from one original DNA molecule. During a normal DNA replication process, two identical copies are produced as illustrated by the center pair of chromosomes in panel (b). However, it's possible for aberrations to occur during the replication process where multiple copies of a section are produced as illustrated by the duplication of section C in the right pair of chromosomes in panel (b). Similarly, it's possible for fewer copies to be produced as illustrated in the left pair of chromosomes in panel (b).

The subject specific copy number profile consists of thousands of continuous measurements along the genome. The resolution of these observations is determined by the technology used to measure the copy number aberrations. Figure 1.2 provides a toy illustration of a subject-specific copy number profile across the human genome. The individual dots that are parallel to each chromosome represent a single measurement at the corresponding genomic location. We will discuss these measurements in more detail in our data illustration in Chapter 2.

In the genetics community, it is well-known that the differences between the genomes of multiple individuals are largely attributable to structural variation as opposed to sequence variation. Thus, researchers are increasingly interested in investigating the role that structural variation plays in the evolution and progression of complex diseases. As noted by Alkan et al.

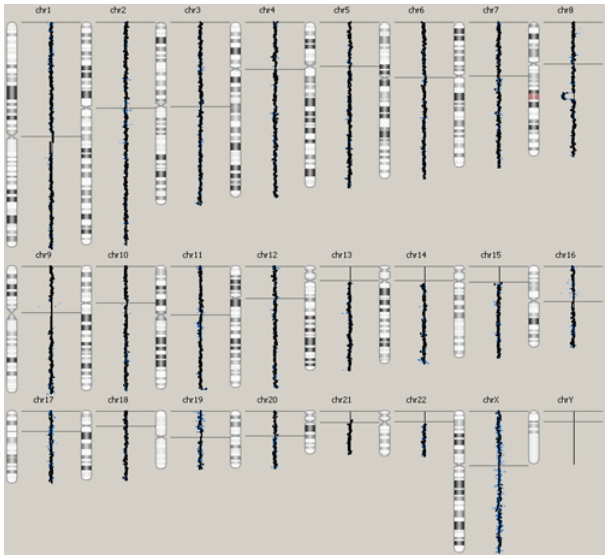


(Source: amazingbiotech.blogspot.com) (Source: readingroom.mindspec.org)

Figure 1.1: Panel (a) provides an illustration of the DNA replication process. Panel (b) provides a simple example of quantitative copy number aberrations.

[2011], the primary approach to investigating the association between disease and CNAs consists of inferring regions of copy number gains and losses (i.e. CNV calling or CNV segmentation) on individual samples with the same diagnosis. Next, recurrent regions of aberration are determined by assessing whether the frequency of samples experiencing a gain or loss exceeds a predetermined threshold [Diskin et al., 2006]. These recurrent regions are flagged for follow-up biological studies. This approach has two key limitations: (1) it fails to consider the genetic similarity across subjects and does not borrow strength across subjects to detect association with disease outcome, and (2) projecting continuous measurements into loss/gain categories discards information about the true biological process that generates the observed copy number intensities and makes it difficult to detect a signal in observations with low to moderate changes [Shah et al., 2007]. We propose novel model-based approaches to investigating the association between CNAs and disease outcomes that overcome the aforementioned limitations. Thus, we believe that our solution to detecting recurrent CNA regions better prioritizes genomic locations for follow-up biological assessments.

In Chapter 2, we propose the nonlinear functional regression model to investigate the relationship between local regions of CNA and continuous prognostic markers that are used as surrogate markers for cancer progression. This model regresses a scalar and continuous outcome onto a nonlinear functional covariate while adjusting for potentially confounding covariates. Our



(Source: http://www.german-mrnet.de/e80/e219/index_eng.html)

Figure 1.2: This figure is an example of a copy number profile for an individual. Each dot represents a continuous measurement of copy number aberration across a local genomic region.

primary focus is testing whether the functional covariate is necessary to model the outcome. We also propose procedures to estimate the magnitude and direction of the functional effect. We use our model to investigate the association between copy number aberrations and continuous prognostic markers for Multiple Myeloma.

In Chapter 3, we propose the generalized nonlinear functional regression model to investigate the relationship between copy number aberrations and binary disease status or categorical disease stages. This models extends the nonlinear functional regression model by considering outcomes that do not follow a normal distribution. In fact, the method is suitable for any outcomes whose distribution is a member of the exponential family. Similar to Chapter 2, the primary focus is testing whether the functional covariate is necessary to model the outcome. In addition to considering non-normal outcomes, we propose a flexible adaptive kernel function that yields a rich feature space capable of mitigating a poor choice of kernel function. The adaptive kernel function appeals to the kernel machine regression literature in addition to being a critical component of the model proposed in the chapter.

In Chapter 4, we propose the functional nonlinear Cox proportional hazards model. This models extends the nonlinear functional regression model into the realm of censored survival outcomes. As one may imagine, determining the variables that are associated with an increased life expectancy is an important component to improving the quality of life for those that have

developed complex diseases such as cancer. As before, we focus our attention on testing for an effect of a single functional covariate on survival; however, we also develop a novel approach to investigating the interaction between a single covariate and the functional covariate. In terms of our cancer genomics problem, this model allows us to investigate the interaction between copy number aberrations over a gene region and the level of expression of the gene housed in the region.

Chapter 2

Nonlinear Functional Regression Models with Application to Copy Number Data

2.1 Introduction

An increasing amount of research suggests that genomic abnormalities in the number of copies of DNA are associated with the development and progression of cancer [Cappuzzo et al., 2005, Diskin et al., 2006]. For example, a genomic aberration associated with the progression of some types of cancers is the rearrangement of the regulatory gene *v-myc myelocytomatosis viral oncogene homolog* (MYC). Such an altered regulation by MYC is believed to promote the progression of multiple myeloma [Sawyer, 2011]. Another genomic aberration associated with cancer involves the deletion of a genomic marker that negatively impacts the well-known tumor suppressor gene *tumor protein p53* (TP53) [Sawyer, 2011]. Recent research has shown that common genomic aberrations exist in a large percentage of patients with the same cancer diagnosis [Rueda and Diaz-Uriarte, 2009]. Thus, it is believed that copy number alterations can have a negative impact on particular genomic regions that harbor genes, which suggests that copy number polymorphisms are critically related to disease progression [Pinkel and Albertson, 2005, Misra et al., 2005]. Detection of these significant regions of CNA remains an important problem in the field of cancer genomics. Such regions contain key information that can be used to determine the evolution of disease as well as to design personalized therapies based on molecular targets. In this article, we propose methods for detecting CNA regions in terms of their association with relevant clinical biomarkers of disease progression across patients with the same diagnosis, as well as for quantifying the effect of such regions on the disease outcome.

Our motivating application arises from multiple myeloma (MM) data collected by the Multiple Myeloma Research Consortium. These data are a comprehensive collection of various genomics measurements obtained from individuals with multiple myeloma through the Multiple Myeloma Genomics Portal (`http://www.broadinstitute.org/mmgp`). A detailed description of the data is provided in Section 2.4. Briefly, the data we consider consist of array-based comparative genomic hybridization (aCGH) copy number profiles and clinical, demographic, and biomarker data on 235 patients with MM. aCGH is a high-throughput technique for measuring the chromosomal DNA copy number along the genome [Pinkel and Albertson, 2005]. The resulting copy number profile consists of approximately 244,000 serially related measurements (along the genome) for each patient. In addition, the data contain measurements on various clinical outcomes that include β_2-microglobulin (β_2M), a protein that is recognized as a prognostic biomarker of MM disease progression with high (lower) values indicative of bad (good) prognosis [Greipp et al., 2005]. It is also well-established that β_2M is related to the frequency of cytogenetic abnormalities–such as CNA–in patients with the same diagnosis [Pignone et al., 2004, Munshi et al., 2011]. In this chapter, we are interested in the following scientific question: what CNA regions are associated with β_2M measurements and what is the effect of such regions on the biomarker?

There are several statistical challenges involved in answering these questions. First, the presence of serial correlation in the copy number data makes it difficult to use standard parametric regression models. An illustration of this feature of the data is provided in Figure 2.1. Notice in this figure that significant magnitudes of autocorrelation persist throughout the duration of the location lag, which suggests that copy number alterations across wide neighborhoods are related. Therefore, the classical approach to genetic association testing, where each site is tested individually, is not well-suited for copy number data. It ignores the possible interaction between sites and the serial correlation among nearby sites. This may result in reduced of power to detect a signal. Second, if one attempts to capture relationships more complex than the linear main effects, then the high dimensionality of the copy number data will cause typical multivariate regression techniques, which attempt to accommodate multiple sites at once, to suffer from the "curse of dimensionality." Third, the observed copy number intensities are inherently contaminated with measurement error by the aCGH technology. Thus, direct usage of the observed intensities in such models may lead to biased conclusions.

To address these challenges, we propose to view the aCGH copy number profiles as functional data, i.e., the serial dependence along the genomic locations suggest that there is an underlying random and smooth process that produces the realized aCGH measurements for each patient with additional measurement error. This functional data view of copy number data was first proposed by Baladandayuthapani et al. [2010]. Their work models the copy number profile as

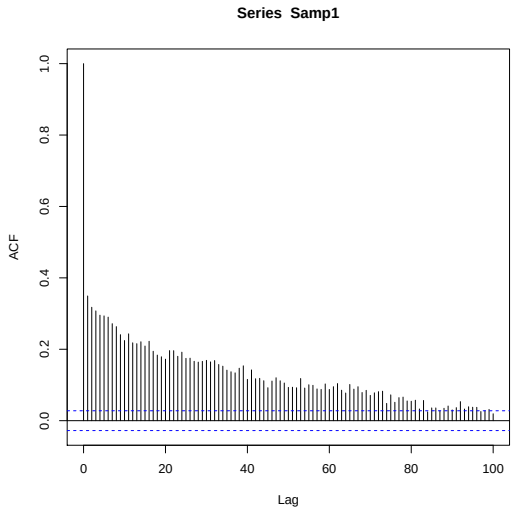


Figure 2.1: Serial correlation in the copy number profile of the chromosome 1 p-arm for a random sample.

a functional mixed-effects model, which has been popularized by Morris and Carroll [2006]. However, the primary aim of their work is to first detect shared patterns of copy number change along the genome, and then to determine if the shared segments can be categorized as neutral, gains, or losses. This approach is similar in spirit to the traditional approach discussed in Chapter 1.

We propose an alternative approach to investigating the association between CNA and disease progression that avoids CNV calling in patients with similar characteristics, such as cancer stages. We cast the problem as a functional regression model, where-in we treat the aCGH profiles as functional covariates and the prognostic biomarker (β_2M) as a continuous response. This enables us to specify a unified modeling framework that accounts for both serial dependence and measurement error, thus allowing us to directly assess the effect of the copy number profile on the clinical biomarker via functional data analysis techniques. The primary goal is to test for the functional effect, which allows us to determine if a region of copy number alteration is associated with the prognostic marker. The secondary goal is to estimate the functional effects (by genomic location) that characterize the relationship between significant regions of copy number alteration and the prognostic marker.

The most common model for continuous scalar on function regression is the functional linear model (FLM) [Ramsay and Dalzell, 1991]. There has been many approaches developed to estimate the magnitude and direction of the functional effect in the FLM. An early approach reduces the dimension of the functional covariates via functional principal component analysis

and uses the principal scores in a standard regression model [Ramsay and Silverman, 2005]. This basic principal component approach has been extended to include roughness penalties on the regression function [Reiss and Ogden, 2007]. Other authors have proposed spline based approaches to estimate the effect of the functional covariate [Marx and Eilers, 1999, Cardot et al., 2003b, Crambes et al., 2009, Goldsmith et al., 2011].

Although estimation procedures are well-developed for the FLM, inferential procedures have received far less attention in the literature, especially procedures designed to test whether a functional covariate should be included in a regression model. In an early work, Cardot et al. [2003a] developed a testing procedure based on the norm of the cross covariance operator of the functional predictor and the scalar response, and subsequently Cardot et al. [2004] proposed a permutation-based test and F tests. A Wald test was proposed by Müller and Stadtmüller [2005] for a generalized extension of the FLM. More recently, Kong et al. [2013] developed an F test and Swihart et al. [2013] proposed restricted likelihood ratio tests to test whether the more complex functional effects are an improvement over standard linear models.

All of the aforementioned testing procedures are based on the FLM. The assumed linear relationship within the FLM may be too restrictive in some scientific settings such as ours. A few authors have proposed functional regression models that model nonlinear functional relationships, which are more appropriate for our copy number association problem. A nonparametric kernel based approach was proposed by Ferraty and Vieu [2006]. Yao and Müller [2010] proposed the functional quadratic regression model (FQRM), which extends the FLM by adding a quadratic term. The functional generalized additive model (FGAM) is another recent approach to model nonlinear relationships between a scalar response and a functional covariate [McLean et al., 2012]. This model extends generalized additive models to functional data. While the estimation procedures for each of these models have been shown to perform well via numerical studies, each model lacks the development of a testing procedure.

In this chapter, our primary interest is in determining whether the functional predictor (i.e. the copy number profile) is related to the continuous response (i.e. β_2M) when the relationship is nonlinear. We formulate a score-like testing procedure by extending the kernel machine framework into the functional data literature. The kernel machine framework projects the model's original design space into a feature space that is specified by a choice of kernel function. The effects of the covariates are then modeled in the feature space as opposed to the model space, which makes these methods well-suited for high-dimensional data. These methods also provide a unified framework for modeling linear and nonlinear effects of various complexities including fully nonparametric models. In addition, this framework has been shown to be computationally equivalent to the linear mixed model framework, which aids in the development of our testing and estimation procedures. These features have led to the kernel machine framework gaining

in popularity in many areas, such as genetic association studies. Our work demonstrates an increased utility of the kernel machine framework as we extend it into the functional data literature.

The new procedures presented in this chapter makes two key contributions in the functional data literature and a key contribution to the statistical genetics literature. First, we develop a testing procedure to investigate whether a functional covariate is needed in a regression model when the effect is modeled nonlinearly. We reduce our functional model to a working linear mixed model, whereby we propose a variance component score test based on the working linear mixed model. Second, we develop procedures to estimate the magnitude and direction of the nonlinear functional effect based on the working linear mixed model. Lastly, we take a model based approach to investigating the relationship between copy number alterations and disease progression as opposed to the approach demonstrated in Avet-Loiseau et al. [2009], which consists of investigating the frequency of patients with known copy number losses or gains across various genomic regions. Our approach has the potential to help genetic researchers retain much more information about the true underlying genetic processes that contribute to the development and progression of complex diseases.

We investigate the finite sample performance of our proposed testing and estimation procedures via simulations. The results show that our proposed testing procedures control type I error rate well. They also yield a large increase in power when the true functional effect is nonlinear. In addition, simulation results show that our estimation procedures have similar performance as other popular nonlinear functional regression approaches.

The remainder of this chapter is organized as follows. We detail our nonlinear functional regression model and the working linear mixed model in Section 2.2. In addition, we provide the details for our proposed testing and estimating procedures in Section 2.2. In Section 2.3, we discuss our numerical studies, and in Section 2.4 we apply our proposed testing procedure to investigate the association between genomic copy number and β_2M in multiple myeloma patients.

2.2 Nonlinear Functional Regression Model

Suppose for $i = 1, \ldots, n$, we observe a continuous clinical biomarker, Y_i, a q-dimensional vector of demographic covariates (confounders), $\mathbf{z}_i$, and copy number intensities, $W_i(t_j)$, where t_j is the jth probe location, $j = 1, \ldots, p$, along a region of interest within a chromosome. We assume that the observed copy number intensities are observed with measurement error. In particular, $W_i(t_j) = X_i(t_j) + \delta_{ij}$, where $X_i(\cdot)$ is the true underlying process that generates the copy number realizations $W_i(t_j)$, and δ_{ij} is mean zero white noise with finite variance. In the application to

MM data, Y_i is β_2M and $\mathbf{z}_i = (z_{i1}, z_{i2})^T$ is the vector of the age and gender of the ith subject, respectively.

Classical functional regression models often make restrictive assumptions about how the effect of the functional covariate is modeled. To provide more details to our discussion in Section 2, consider the FLM: $E(Y_i|X_i) = \alpha + \int X_i(t)\beta(t)\, dt$, where $\beta(\cdot)$ is an unknown coefficient function that linearly models the effect of $X(\cdot)$ on Y_i [Ramsay and Silverman, 2005]. As previously suggested, such an assumption is restrictive in reality and may fail to capture the relationships in the data. The FQRM by Yao and Müller [2010] extends the FLM by adding an extra quadratic term, $\int X_i(s)X_i(t)\gamma(s,t)\, dsdt$. The FGAM proposed by McLean et al. [2012] has the form $E(Y_i|X_i) = \alpha + \int F\{t, X_i(t)\}\, dt$, where $F\{t, X_i(t)\}$ is an unknown function.

We propose a general framework that models the effect of the functional covariate in a nonlinear fashion. We assume that the effect of $X_i(\cdot)$ is modeled by an unknown functional operator $\mathcal{L}\{X_i(\cdot)\}$. We further assume that the continuous response, Y_i, is related to $\mathbf{z}_i$ and $X_i(\cdot)$ through the following functional model

$$Y_i = \mathbf{z}_i^T\boldsymbol{\beta} + \mathcal{L}\{X_i(\cdot)\} + \epsilon_i, \tag{2.1}$$

where ϵ_i are mean zero errors with finite variance and $\mathcal{L}(\cdot) : L^2[\mathcal{T}] \mapsto \mathbb{R}$. To properly identify $\mathcal{L}(\cdot)$, we assume that E$[\mathcal{L}\{X(\cdot)\}]$= 0 and E$[X(\cdot)]$= 0 without any loss of generality and with the understanding that $\mathbf{z}_i$ contains an intercept. The FLM is a special case of model (2.1), where $\mathcal{L}\{X_i(\cdot)\} = \int X_i(t)\beta(t)\, dt$. Similar expressions exist for the FQRM and the FGAM. By not expressing a functional form, we assert that $\mathcal{L}(\cdot)$ may represent any functional that maps a continuous quadratically integrable function to $\mathbb{R}$. Thus in summary, confounding covariate effects are modeled parametrically and the functional covariate is modeled nonparametrically. In this framework, we develop statistical methodology for testing the effect of the functional covariate on the response and estimating the model components utilizing a procedure based on a Gaussian model.

2.2.1 Estimation under the Gaussian Process Framework

Smoothing and Dimension Reduction

Our first task for model fitting is to transition $\mathcal{L}(\cdot)$ from an infinite-dimensional function space to a lower-dimensional space. To this end, we employ functional principal components analysis.

Assume that $X_i(\cdot)$ is a real valued smooth process defined on $\mathcal{T}$ with continuous covariance function $V(s,t) = \text{Cov}\{X_i(s), X_i(t)\}$. By Mercer's theorem, there exists an orthonormal set of continuous eigenfunctions in $L^2[\mathcal{T}]$, $\{\phi_j(\cdot), j \geq 1\}$, and a non-increasing set of eigenvalues,

$\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$, such that $V(s,t) = \sum_{j=1}^{\infty} \lambda_j \phi_j(s)\phi_j(t)$, for $t, s \in \mathcal{T}$ [Cristianini and Shawe-Taylor, 2000]. The eigenbasis that corresponds to $V(s,t)$ yields the Karhunen-Loéve expansion of the random process, $X_i(t) = \sum_{j=1}^{\infty} \xi_{ij}\phi_j(t)$, where $\phi_j(t)$ is the jth eigenfunction and $\xi_{ij} = \int X_i(t)\phi_j(t)dt$ is the random functional principal component score with $E(\xi_{ij}) = 0$ and $\text{Var}(\xi_{ij}) = \lambda_j$. In practice, the infinite sum is truncated such that $X_i(t) \approx \sum_{j=1}^{J} \xi_{ij}\phi_j(t)$, where J determines the closeness of the approximation.

While there are several types of orthogonal expansions of the process $X_i(\cdot)$, the Karhunen-Loéve expansion provides the most parsimonious approximation of $X_i(\cdot)$ for fixed J. In other words, the use of the eigenbasis minimizes the mean integrated square error fitting criterion, $\text{MISE} = \sum_{j=1}^{J} ||X_i(\cdot) - \hat{X}_i(\cdot)||^2$, with respect to all other sets of orthogonal basis functions [Ramsay and Silverman, 2005].

There are several ways to choose the truncation point J. One popular approach fixes J by selecting the number of eigenfunctions, $\{\phi_1, \ldots, \phi_J\}$, that account for a predetermined percentage of functional variation explained (FVE) by the truncated expansion. Specifically, $\text{FVE} = \sum_{j=1}^{J} \lambda_j / \sum_{j=1}^{\infty} \lambda_j$, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_J$ are the ordered eigenvalues that correspond to the eigenfunctions. This approach has a low computational cost compared to other approaches to selecting J, such as cross validation.

Recall that we do not observe $X(\cdot)$; instead, we observe $W(\cdot)$, a proxy version of the true functions with measurement errors. To account for such measurement error in the observed functional observations, we use principal analysis by conditional expectation (PACE) to estimate the FPC scores, which is available under programs at `http://www.stat.ucdavis.edu/~mueller/` [Yao et al., 2005]. PACE accounts for measurement error by estimating $V(s,t)$ using local linear kernel-smoothers and subsequently estimating ξ_{ij} under the mixed model framework. For more technical details about PACE, see Appendix A.1.

Working Mixed Model

Each function defined on $\mathcal{T}$ has a unique basis expansion for every set of basis functions that span the function space. The uniqueness of each function is determined by the coefficients within the basis expansion. Thus, when using the eigenbasis, the FPC scores, ξ_{ij}, capture most of the information in the true functional covariates. Drawing from the information retained in ξ_{ij}, we approximate $\mathcal{L}\{X_i(\cdot)\}$ with a multivariate function $\mathcal{L}^*(\boldsymbol{\xi}_i) : \mathbb{R}^J \mapsto \mathbb{R}$, where $\boldsymbol{\xi}_i = \{\xi_{i1}, ..., \xi_{iJ}\}^T$. To fix the ideas, consider the FLM where $\mathcal{L}\{X_i(\cdot)\} = \int X_i(t)\beta(t)dt$. Expanding $\beta(\cdot)$ with the same eigenbasis used to expand $X_i(\cdot)$ gives $\mathcal{L}\{X_i(\cdot)\} \approx \sum_{j=1}^{J} \eta_j \xi_{ij}$. In general, we assume that

$\mathcal{L}\{X_i(\cdot)\} \approx \mathcal{L}^*(\boldsymbol{\xi}_i)$, where J is chosen appropriately. Thus, we write an approximate model,

$$Y_i = \mathbf{z}_i^T \boldsymbol{\beta} + \mathcal{L}^*(\boldsymbol{\xi}_i) + \epsilon_i, \tag{2.2}$$

where $\mathcal{L}^*(\boldsymbol{\xi}_i)$ is a smooth function with a finite-dimensional argument.

To estimate the function $\mathcal{L}^*(\boldsymbol{\xi}_i)$ in model (2.2), we assume that $\mathcal{L}^*(\cdot)$ is a mean zero Gaussian process with covariance $\tau K(\cdot,\cdot)$. Here, τ is an unknown variance component and $K(\cdot,\cdot)$ is a kernel function, such that $K(\boldsymbol{\xi}_l, \boldsymbol{\xi}_k) = \text{Cov}\{\mathcal{L}^*(\boldsymbol{\xi}_l), \mathcal{L}^*(\boldsymbol{\xi}_k)\}$, $l, k = 1, \ldots, n$. Thus, we can express model (2.2) as the following working mixed model,

$$\mathbf{Y} = \mathbf{Z}\boldsymbol{\beta} + \mathcal{L}^* + \boldsymbol{\epsilon}, \tag{2.3}$$

where $\mathbf{Y}$, Z and $\boldsymbol{\beta}$ follow from model (2.2), $\mathcal{L}^* = [\mathcal{L}^*(\boldsymbol{\xi}_1), \ldots, \mathcal{L}^*(\boldsymbol{\xi}_n)]^{\text{T}}$ is a n-dimensional vector of random variables such that $\mathcal{L}^* \sim N(\mathbf{0}, \tau\mathbf{K})$, and $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$. Here, $\mathbf{K}$ is an $n \times n$ covariance matrix that captures the variation of the functional effects across subjects such that $\mathbf{K}_{lk} = K(\boldsymbol{\xi}_l, \boldsymbol{\xi}_k)$.

The kernel function performs two tasks in our model: (1) it models the copy number similarity between subjects, which allows us to borrow information across subjects to quantify the effect of a recurrent CNA region as opposed to taking the traditional approach of observing the frequency of patients with particular copy number alterations, and (2) it expresses the functional relationship between the copy number profile and the clinical biomarker. The choice of kernel function determines the function space used to approximate $\mathcal{L}^\star$. Two popular choices of kernel functions are the *dth* polynomial kernel function $K(\boldsymbol{\xi}_l, \boldsymbol{\xi}_k) = (\boldsymbol{\xi}_l^T \boldsymbol{\xi}_k + 1)^d$ and the Gaussian kernel function $K(\boldsymbol{\xi}_l, \boldsymbol{\xi}_k) = \exp\{(\sum_{j=1}^J (\xi_{lj} - \xi_{kj})^2/\kappa^2\}$, where κ is an unknown tuning parameter. The linear kernel function ($d = 1$) assumes that the relationship between Y_i and $X_i(\cdot)$ is linear, while the quadratic kernel function ($d = 2$) assumes that the relationship between Y_i and $X_i(\cdot)$ is quadratic. The Gaussian kernel function assumes the space of $\mathcal{L}^\star$ is spanned by a radial basis and the tuning parameter κ determines the relationship between Y_i and $X_i(\cdot)$.

In principle, we can choose any kernel for implementation. Choosing an optimal $K(\cdot,\cdot)$ is an open problem in the kernel machine literature. Liu et al. [2007] suggests using AIC/BIC methods to select kernels and Wu et al. [2013] proposes to use composite kernels that average over several candidate kernels. Regardless of the approach, the choice of kernel function for $\mathcal{L}^\star$ should determine a function space that has a set of basis functions that is capable of capturing a wide range of complex and nonlinear functions. For computational efficiency, which worked well in practice for both the simulated and real data, we use the quadratic kernel, $K(\boldsymbol{\xi}_l, \boldsymbol{\xi}_k) =$

$(\boldsymbol{\xi}_l^T \boldsymbol{\xi}_k + 1)^2$, to demonstrate the novelty of our approach. Simulation results suggest that the quadratic kernel is robust in the sense that when the true functional effect is clearly more complex than a quadratic form, usage of the quadratic kernel provides reasonable power and estimation accuracy for $\boldsymbol{\beta}$ and $\mathcal{L}^*$.

We estimate the components of model (2.3) using standard procedures for linear mixed effects models (LMMs). In particular, we construct the best linear unbiased estimate (BLUE) of $\boldsymbol{\beta}$ and the best linear unbiased prediction (BLUP) of $\mathcal{L}^*$ as $\widehat{\boldsymbol{\beta}} = (Z^{\mathrm{T}} V^{-1} Z)^{-1} Z^{\mathrm{T}} V^{-1} \mathbf{Y}$ and $\widehat{\mathcal{L}}^* = (\lambda^{-1}\sigma^2)\mathbf{K} V^{-1}(Y - Z\widehat{\boldsymbol{\beta}})$. Here $\lambda = \tau^{-1}\sigma^2$ and $V = \sigma^2(I + \lambda^{-1}\mathbf{K})$. We estimate the standard errors of our estimates as $\mathrm{Cov}(\widehat{\boldsymbol{\beta}}) = (Z^{\mathrm{T}} V^{-1} Z)^{-1}$ and $\mathrm{Cov}(\widehat{\mathcal{L}}^*) = (\sigma^2/\lambda)\mathbf{K}(I - P\mathbf{K})$, where $P = V^{-1} - V^{-1}Z(Z^{\mathrm{T}} V^{-1} Z)^{-1} Z^{\mathrm{T}} V^{-1}$. It remains to estimate σ^2 and τ.

Given the LMM framework, σ^2 and τ can be estimated by maximizing the restricted maximum likelihood (REML) under model (2.3). We proceed by profiling the REML below:

$$l_R(\sigma^2, \lambda) = -\frac{1}{2}\log|\mathbf{V}(\theta)| - \frac{1}{2}|Z^{\mathrm{T}} V^{-1}(\theta) Z| - \frac{1}{2}(\mathbf{Y} - Z\boldsymbol{\beta})\mathrm{T}\mathbf{V}^{-1}(\theta)(\mathbf{Y} - Z\boldsymbol{\beta}),$$

where $\theta = (\sigma^2, \lambda)^{\mathrm{T}}$ and $V(\theta) = \sigma^2(I + \lambda^{-1}\mathbf{K})$. Note that σ^2 cancels out in $\hat{\boldsymbol{\beta}}$ and $\widehat{\mathcal{L}}^\star$. This suggests that we can find an optimal tuning parameter λ without considering the residual variance σ^2. Thus, we fix $\sigma^2 = 1$ and $\hat{\lambda}$ is obtained by performing a grid search over $l_R(\lambda; \sigma^2 = 1)$. Having found the optimal tuning parameter, $\hat{\lambda}$, that yields $\hat{\boldsymbol{\beta}}$ and $\hat{\mathcal{L}}^\star$, we maximize $l_R(\sigma^2; \hat{\lambda})$. The score equation for σ^2 is $U(\sigma^2) = -\frac{1}{2}tr(P) + \frac{1}{2}(\mathbf{Y} - Z\hat{\boldsymbol{\beta}})^{\mathrm{T}} V^{-1} V^{-1}(\mathbf{Y} - Z\hat{\boldsymbol{\beta}})$. Given this score equation, σ^2 can be estimated as

$$\hat{\sigma}^2 = (n - \mathrm{tr}(A))^{-1} \sum_{i=1}^{n} (y_i - \mathbf{z}_i^{\mathrm{T}} \hat{\boldsymbol{\beta}} - \hat{\mathcal{L}}_i^\star)^2,$$

where $A = Z^{\mathrm{T}}(Z^{\mathrm{T}} \tilde{V}^{-1} Z)^{-1} Z^{\mathrm{T}} \tilde{V}^{-1} + \lambda^{-1} \tilde{V}^{-1}[I_n - Z^{\mathrm{T}}(Z^{\mathrm{T}} \tilde{V}^{-1} Z)^{-1} Z^{\mathrm{T}} \tilde{V}^{-1}]$ [Liu et al., 2007]. Here $\tilde{\mathbf{V}} = (I + \hat{\lambda}^{-1}\mathbf{K})$ and A is the hat matrix such that $\hat{\mathbf{Y}} = A\mathbf{Y}$. The term $\mathrm{tr}(A)$ represents the loss in degrees of freedom from estimating $\boldsymbol{\beta}$ and $\mathcal{L}^\star$. Estimating λ and σ^2 in this manner simplifies the numerical optimization problem. The computation time required for such a grid search and calculation of the closed form expression is very low. Choosing a wide and dense grid ensures that we obtain a local maxima or that we are close enough to the local maxima that the efficiency of our estimates is relatively high. The performance of the estimation procedure is evaluated using simulations and illustrated on an MM data set.

2.2.2 Testing for the Effect of the Functional Covariate

A problem of particular interest in many real data situations such as ours, is to explicitly test whether the functional covariate is necessary in the regression model for an outcome. Thus from model (2.1), we are interested in testing the hypothesis $H_0 : \mathcal{L}\{X(\cdot)\} = 0$. Using the mixed model formulation in (2.3), an equivalent hypothesis is $H_0 : \tau = 0$.

Here we adopt the approach proposed by Liu et al. [2007]; that is, we use an REML-based variance component score test under the LMM framework. The score statistic of τ under the null hypothesis is $\mathcal{Q}_\tau(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2) = \frac{1}{2\widehat{\sigma}^2}(Y - Z\widehat{\beta})^{\mathrm{T}}\mathbf{K}(Y - Z\widehat{\beta})$, where $\widehat{\beta}$ and $\widehat{\sigma}^2$ are the maximum likelihood estimates of β and σ^2 under the null model $\mathbf{Y} = Z\boldsymbol{\beta} + \boldsymbol{\epsilon}$.

We use the linear mixed model formulation in (2.3) to derive the distribution of $\mathcal{Q}_\tau(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2)$. From this model, we have $(\mathbf{Y} - Z\widehat{\boldsymbol{\beta}}) \sim N(0, V)$, where $V = \Sigma + (Z^{\mathrm{T}}\Sigma^{-1}Z)^{-1}$, $\Sigma = \sigma^2 I + \tau\mathbf{K}$. We rewrite the score statistic such that $\mathcal{Q}_\tau(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2) = (\mathbf{Y} - Z\widehat{\boldsymbol{\beta}})^{\mathrm{T}}V^{-1/2}V^{1/2}\mathbf{K}V^{1/2}V^{-1/2}(\mathbf{Y} - Z\widehat{\boldsymbol{\beta}})$. Defining a matrix $M = V^{1/2}\mathbf{K}V^{1/2}$, we have $\mathcal{Q}_\tau(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2) = (\mathbf{Y} - Z\widehat{\boldsymbol{\beta}})^{\mathrm{T}}V^{-1/2}MV^{-1/2}(\mathbf{Y} - Z\widehat{\boldsymbol{\beta}})$. Given that M is a real symmetric matrix, we have the following spectral decomposition, $M = UDU^T$. This implies that $\mathcal{Q}_\tau(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2) = \sum_{i=1}^{n} d_i r_i^2$, where d_i is the ith eigenvalue obtained from the spectral decomposition of M and $r_i = (Y_i - Z_i^T\widehat{\boldsymbol{\beta}})(\Sigma_{ii} + \{Z_i^T\Sigma^{-1}Z_i\}^{-1})^{-1/2}$. Clearly, $r_i \sim N(0, 1)$ which implies that $r_i^2 \sim \chi_1^2$. Thus, $\mathcal{Q}_\tau(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2)$ follows a mixture of chi-squares under the H_0.

Following Zhang and Lin [2003], the Satterthwaite method is used to approximate the null distribution of $\mathcal{Q}_\tau(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2)$ by a scaled chi-squared distribution, $k\chi_v^2$. The scale parameter k and the degrees of freedom v are estimated by matching the moments of $\mathcal{Q}_\tau(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2)$ and $k\chi_v^2$. Let $\mu(\mathcal{Q}) = \frac{1}{2}\mathrm{tr}(P_0\mathbf{K})$, where $P_0 = I - Z(Z^{\mathrm{T}}Z)^{-1}Z^{\mathrm{T}}$. It can be shown that $\tilde{k} = \tilde{I}_{\tau\tau}/2\mu(\mathcal{Q})$ and $\tilde{v} = 2\mu(\mathcal{Q})^2/\tilde{I}_{\tau\tau}$, where $I_{\tau\tau} = \mathrm{tr}((P_0\mathbf{K})(P_0\mathbf{K}))/2$, $I_{\tau\sigma^2} = \mathrm{tr}(P_0\mathbf{K}P_0)/2$, $I_{\sigma^2\sigma^2} = \mathrm{tr}(P_0P_0)/2$, and $\tilde{I}_{\tau\tau} = I_{\tau\tau} - I_{\tau\sigma^2}I_{\sigma^2\sigma^2}^{-1}I_{\tau\sigma^2}$. We use simulations to evaluate the performance of the score test and apply it to our MM data set for illustration.

2.3 Simulation

We conduct simulation studies to evaluate NFRM's estimation and test procedures. For each simulation, we evaluate NFRM using several functional operators that vary in the degree of complexity. The first simulation is designed to evaluate NFRM's type I error control and power. To evaluate our testing procedure, we compare our NFRM to the FLM to exemplify the benefits of the nonlinear approach. The second simulation is designed to evaluate how well NFRM estimates $\boldsymbol{\beta}$ and $\mathcal{L}\{X_i(\cdot)\}$ from model (2.1). For these investigations, we compare NFRM to the FLM and FQRM.

2.3.1 Type I Error and Power

We design our first simulation study to evaluate how well NFRM and FLM control the following type I errors, $\alpha = \{0.05, 0.01, 0.005, 0.001, 0.0005, 0.0001\}$. We benchmark our simulations to mimic the copy number data in terms of sample size, variation and matched covariates. We generate data from the model, $Y_i = \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\beta} + h\{X_i(t)\} + \epsilon_i$, where $\mathbf{z}_i = (z_{i,1},\ z_{i,2})^{\mathrm{T}}$, $z_{i,1} \sim N(0,1)$ and $z_{i,2} \sim Bin(1, 0.66)$, represent the appropriately standardized age and gender, respectively, of the ith subject, and the error is $\epsilon_i \sim N(0,1)$ and independent. The frequency of males in the multiple myeloma data set is approximately 0.66.

The true underlying trajectories, $X_i(t)$, are generated from a Fourier basis with five basis functions. The coefficients for each basis function are independent realizations of an N(0,0.5) distribution. We generate the observed copy number intensity profiles as $W_i(t_j) = X_i(t_j) + \delta_{ij}$, where $\delta_{ij} \sim N(0, 0.16)$. We set the true values of $\boldsymbol{\beta}$ as $\boldsymbol{\beta} = (1,1)^{\mathrm{T}}$.

To retain the identifiability of the functional effect, we fit the model $Y_i = \alpha + \mathbf{z}_i^T\boldsymbol{\beta} + \mathcal{L}\{X_i(t)\} + \epsilon_i$ where we define $\mathcal{L}\{X_i(\cdot)\} = h\{X_i(\cdot)\} - \mathrm{E}[h\{X(\cdot)\}]$ and $\alpha = \mathrm{E}[h\{X(\cdot)\}]$. Our final estimate of the functional effect is $\widehat{h}\{X_i(\cdot)\} = \widehat{\alpha} + \widehat{\mathcal{L}}\{X_i(\cdot)\}$. We explore five functional operators that vary in complexity:

1. Linear functional: $h(f) = \int f(t)\gamma(t)\ dt$;
2. Quadratic functional: $h(f) = (\int f(t)\gamma(t)\ dt)^2$;
3. Absolute value of the 1^{st} derivative: $h(f) = |\int f'(t)\gamma(t)\ dt|$;
4. Signed square root of the 2^{nd} derivative: $h(f) = sgn(\int f''(t)\gamma(t)\ dt) * \sqrt{|\int f''(t)\gamma(t)\ dt|}$;
5. Linear functional of the squared 1^{st} derivative: $h(f) = (0.9)\int f'(t)^2\ dt$,

where $\gamma(t)$ is generated from a Fourier basis with five basis functions, with 0.9 as the coefficients for each of the functions. The linear functional is the relationship assumed by the FLM. The quadratic functional is the relationship assumed by the quadratic kernel function used in model building and the FQRM. The remaining functionals are arbitrary nonlinear functionals of a derivative of the functional covariate. In the functional data analysis literature, it is common that the relationship between the response and the functional covariate is determined by a derivative of the functional covariate [Ramsay and Silverman, 2005]; thus, a robust method should be able to capture those types of effects.

We consider two sample sizes, $n = 100, 200$ and we set $\mathcal{L}\{X_i(t)\} = 0$ and the level of functional variation explained (FVE) = 0.95. We generate 1,000,000 data sets. Following the procedure described in Section 2.2.2, we compare $S = \mathcal{Q}_\tau(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2, \rho)/k$ to a χ^2_v distribution.

We follow the approach of Kong et al. [2013] for testing in the FLM. They propose the following test: $T = \frac{(SSE_R - SSE_F)/J}{SSE_F/(n-p)}$, where J is determined by the number of FPC scores needed to account for the selected FVE level and $p = J + 2$. Under the H_0, $T \sim F_{(J,n-p)}$.

For NFRM and FLM, the empirical rejection probability is recorded as the average number of p-values less than the significance level α. Table 2.1 shows that for $n = 100$, both NFRM and FLM perform well with respect to controlling a wide range of nominal type I error rates. Similar results hold for $n = 200$ and may be viewed in Appendix B.

Table 2.1: Simulation results for type I error based on 1,000,000 generated datasets and $n = 100$. Values are displayed in percentages.

	Type I Error	NFRM	FLM
n =100	$5 \times (10^{-}2)$	5.0	5.5
	$1 \times (10^{-}2)$	1.1	1.2
	$5 \times (10^{-}3)$	5.5	5.9
	$1 \times (10^{-}3)$	1.3	1.2
	$5 \times (10^{-}4)$	7.0	6.1
	$1 \times (10^{-}4)$	1.8	1.2

Next, we conduct a simulation study to assess the empirical power of NFRM and the FLM. The process is similar to that used to evaluate the type I error control, except that we use eight equally spaced functional effect levels $b_l \in [0, 1)$, $l = 1, \ldots, 8$. We consider four levels of FVE, $\{0.85, 0.90, 0.95, 0.99\}$. For each setting, we generate 1,000 data sets. The empirical rejection probability is recorded as the average number of p-values less than $\alpha = 0.05$.

Figure 2.2 shows that for the quadratic functional, NFRM has a smooth monotone increasing power curve, while FLM fails to detect any signal. The results across the sample sizes are similar. Although the covariance structure is misspecified, the results for Functional 3 to functional 5 are similar to those of the quadratic functional. The figures corresponding to these functional effects may be viewed in Appendix B. Figure 2.2 also shows that the FLM outperforms the NFRM in the case of the linear functional. However, the linear model is nested within the implicit quadratic model. Thus, this performance gain is expected to disappear with larger sample sizes. To summarize in the context of our multiple myeloma data, if the copy number profiles are nonlinearly related to the clinical biomarker, we expect the NFRM to be far superior to the FLM. In contrast, if the relationship is linear, then the FLM is expected to be slightly superior to the NFRM. Thus, it may be advantageous to use the testing procedures from NFRM and FLM as companion tests, illustrating different aspects of dependence in the data.

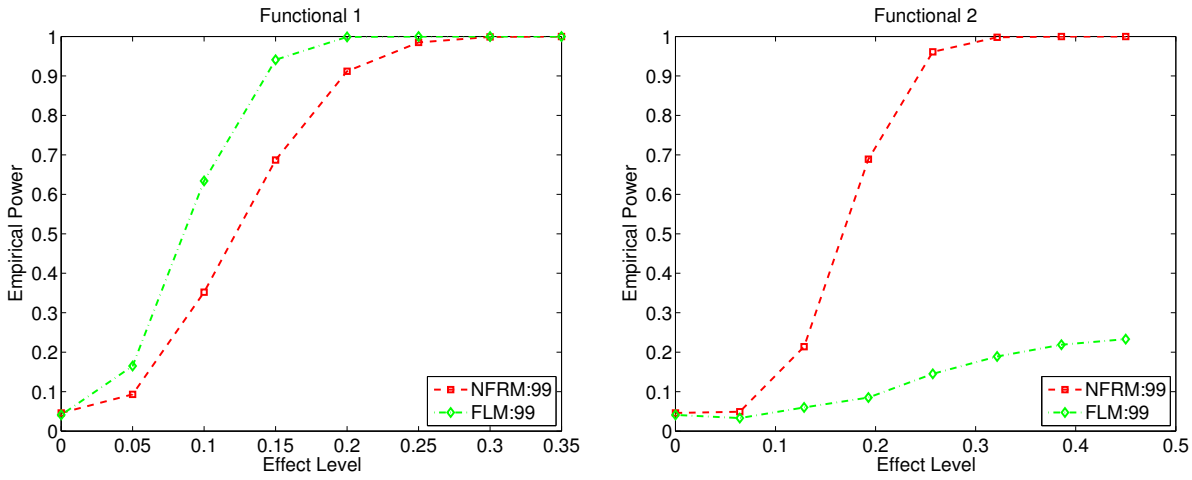


Figure 2.2: Simulation results of the rejection probability as a function of b as outlined in Section 2.3.1. Both panels display the results of NFRM and FLM where the number of principal components, J, are determined by a FVE level of 99%. The left panel shows the results for the linear functional at sample size $n = 100$. The right panel shows the results for the quadratic functional at sample size $n = 100$.

2.3.2 Estimation

In this section, we evaluate the performance of our estimation procedure. The data generation process is identical to that described in Section 2.3.1 which benchmarks this numerical investigation to our motivating multiple myeloma data.

For this study, we also use four levels of FVE to choose J, FVE $= 0.85, 0.90, 0.95, 0.99$ and we generate 1000 simulated data sets for each setting. To maximize the REML with respect to λ, we perform a grid search over a regular grid of 91 λ values equally spaced on the log-scale, such that $\lambda_1 = 10^{-5}$ and $\lambda_{91} = 10^5$.

We compare our procedure to the FLM and the FQRM using principal components regression which is the most fair comparison [Ramsay and Silverman, 2005, Yao and Müller, 2010]. Estimation under both the FLM and the FQRM is done modeling both $X_i(\cdot)$ and $\gamma(t)$ using the eigenbasis. For the FQRM, $\zeta(s,t)$ is also estimated using the eigenbasis. This implies that the FLM is approximated with the following parametric model: $Y_i = \alpha + \mathbf{z}_i\boldsymbol{\beta} + \sum_{j=1}^{J} \widehat{\xi}_{ij}\eta_j + \epsilon_i$. The FQRM is approximated with a similar model with the following terms added for the quadratic effect: $\sum_{j=1}^{J}\sum_{l=1}^{j} \widehat{\xi}_{ij}\widehat{\xi}_{il}\omega_{jl}$.

Given this representation, $\widehat{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{\eta}}$, $\widehat{\boldsymbol{\omega}}$ and their standard errors are obtained via least squares. Thus, for the FLM we estimate $\widehat{\mathcal{L}}\{X_i(t)\} = \widehat{\alpha} + \sum_{j=1}^{J} \widehat{\xi}_{ij}\widehat{\eta}_j$ and for the FQRM, we add the additional terms $\sum_{j=1}^{J}\sum_{l=1}^{j} \widehat{\xi}_{ij}\widehat{\xi}_{il}\widehat{\omega}_{jl}$.

To evaluate how well NFRM estimates the parameter $\boldsymbol{\beta}$, we compare the following:

- $\text{Bias}(\widehat{\beta}_1) = M^{-1}\sum_{m=1}^{M}(\widehat{\beta}_{1m} - \beta_1)$
- $\text{RMSE}(\widehat{\beta}_1) = M^{-1}\{\sum_{m=1}^{M}(\widehat{\beta}_{1m} - \beta_1)^2\}^{1/2}$
- $\text{SE}(\widehat{\beta}_1) = M^{-1}\sum_{m=1}^{M}\text{SE}(\widehat{\beta}_{1m})$
- $\text{Coverage}(\widehat{\beta}_1) = M^{-1}\sum_{m=1}^{M}\mathbb{1}\{\beta_1 \in [\widehat{\beta}_{1m} - \iota(\widehat{\beta}_{1m}), \widehat{\beta}_{1m} + \iota(\widehat{\beta}_{1m})]\}$

Here, $M = 1000$ and $\iota(\widehat{\beta}_{1m}) = 1.96\text{SE}(\widehat{\beta}_{1m})$. The same comparisons are made for $\widehat{\beta}_2$.

To evaluate how well NFRM estimates $\mathcal{L}\{X_i(t)\}$, we estimate the following:

- $\text{RMSE}(\widehat{\mathcal{L}}) = \{M^{-1}\sum_{m=1}^{M}(n^{-1}\sum_{i=1}^{n}[\mathcal{L}_{im} - \widehat{\mathcal{L}}_{im}]^2)\}^{1/2}$
- $\text{ABSE}(\widehat{\mathcal{L}}) = M^{-1}\sum_{m=1}^{M}(n^{-1}\sum_{i=1}^{n}|\mathcal{L}_{im} - \widehat{\mathcal{L}}_{im}|)$
- $\text{Corr}(\widehat{\mathcal{L}}, \mathcal{L}) = M^{-1}\sum_{m=1}^{M}\text{Corr}(\widehat{\mathcal{L}}_m, \mathcal{L}_m)$

Here, $\mathcal{L}_{im}$ is the functional effect of the ith subject for the mth generated data set, and $\mathcal{L}_m$ is the n-dimensional vector of the functional effects for the mth generated data set. For each generated data set, we also use a linear model to regress $\mathcal{L}$ on $\widehat{\mathcal{L}}$ such that $\mathcal{L}_m = \alpha_m\mathbf{1}_n + \widehat{\mathcal{L}}_m\gamma_m + \boldsymbol{\epsilon}_m$. We summarize these regressions across iterations such that $\widehat{\alpha} = M^{-1}\sum_{m=1}^{M}\alpha_m$ and $\widehat{\gamma} = M^{-1}\sum_{m=1}^{M}\gamma_m$. If $\widehat{\mathcal{L}}$ is close to the true $\mathcal{L}$, then we expect $\widehat{\alpha} \approx 0$ and $\widehat{\gamma} \approx 1$.

With respect to estimating $\mathcal{L}\{X_i(\cdot)\}$, Table 2.2 highlights the NFRM's advantages over the FLM in the case of nonlinear functionals. For functional 2, the covariance structure is correctly specified. The higher correlation values and lower mean squared error values suggest that NFRM performs well. The results for functional 3 to functional 5 further highlight the novelty of our approach. In each case, the same measures suggest that NFRM performs reasonably well when the covariance function is clearly misspecified. For functional 2 to functional 5, the same measures for the FLM reveal its inability to capture nonlinear effects. In the case of the linear functional, FLM slightly outperforms NFRM. These results are provided in Appendix B. We also note that Table 2.2 also shows that the NFRM and the FQRM have similar performance under all settings.

With respect to estimating $\boldsymbol{\beta}$, the performance of the NFRM, the FQRM and the FLM are essentially identical. This is because the NFRM provides the BLUEs from the LMM framework whereas the FQRM and the FLM provide the BLUEs from the linear model framework. Thus, these results are trivial and may be viewed in Appendix B.

In summary, our method outperforms the FLM with respect to estimating nonlinear functional effects and performs similar to the FQRM. In addition, the NFRM performs similar to

both the FQRM and the FLM with respect to estimating the effect of the demographic covariates. This makes our NFRM (with a quadratic kernel) attractive for modeling the effect of copy number data, as we show in Section 2.4.

Table 2.2: Estimation results for the functional effect, $\mathcal{L}\{X_i(\cdot)\}$, in the model $Y_i = \beta_1 z_{1i} + \beta_2 z_{2i} + \mathcal{L}\{X_i(\cdot)\} + \epsilon_i$ based on 1000 generated data sets.

Definitions: NFRM, nonlinear functional regression model; FLM, functional linear model; FQRM, functional quadratic regression model; FVE, functional variation; MSE, mean squared error; ABSE, absolute error.

		NFRM					FQRM					FLM				
n	FVE	MSE	ABSE	Int	Slope	Corr	MSE	ABSE	Int	Slope	Corr	MSE	ABSE	Int	Slope	Corr
Functional 2																
100	0.85	1.380	0.941	-0.098	1.041	0.847	1.361	0.936	0.063	0.971	0.850	2.780	1.943	0.207	0.917	0.235
	0.99	0.835	0.626	-0.034	1.014	0.956	0.836	0.629	0.056	0.972	0.957	2.764	1.937	0.205	0.919	0.259
200	0.85	1.178	0.761	-0.039	1.018	0.895	1.173	0.759	0.028	0.988	0.895	2.821	1.945	0.216	0.925	0.177
	0.99	0.805	0.576	-0.017	1.010	0.959	0.804	0.576	0.025	0.989	0.959	2.816	1.943	0.201	0.928	0.186
Functional 3																
100	0.85	0.827	0.654	-0.179	1.103	0.729	0.821	0.642	0.236	0.867	0.740	1.216	0.968	0.563	0.658	0.182
	0.99	0.670	0.540	-0.095	1.054	0.852	0.673	0.532	0.217	0.863	0.860	1.212	0.965	0.578	0.650	0.202
200	0.85	0.714	0.559	-0.079	1.042	0.794	0.713	0.553	0.130	0.920	0.797	1.212	0.969	0.566	0.654	0.134
	0.99	0.596	0.481	-0.047	1.023	0.878	0.599	0.475	0.114	0.924	0.879	1.211	0.968	0.575	0.648	0.141
Functional 4																
100	0.85	0.775	0.607	-0.198	1.090	0.771	0.772	0.601	0.293	0.866	0.778	1.176	0.937	0.791	0.639	0.179
	0.99	0.636	0.503	-0.108	1.047	0.860	0.645	0.506	0.308	0.857	0.866	1.173	0.935	0.798	0.632	0.201
200	0.85	0.663	0.509	-0.091	1.039	0.829	0.663	0.506	0.158	0.921	0.830	1.173	0.935	0.806	0.630	0.131
	0.99	0.560	0.439	-0.058	1.026	0.887	0.562	0.438	0.158	0.922	0.888	1.173	0.934	0.790	0.634	0.139
Functional 5																
100	0.85	1.696	1.219	-0.191	1.041	0.886	1.686	1.205	0.071	0.986	0.887	3.616	2.718	0.279	0.945	0.244
	0.99	1.003	0.770	-0.080	1.015	0.963	1.001	0.768	0.059	0.984	0.964	3.589	2.708	0.280	0.944	0.271
200	0.85	1.467	1.009	-0.077	1.017	0.914	1.465	1.005	0.030	0.994	0.914	3.651	2.736	0.250	0.950	0.180
	0.99	0.997	0.749	-0.043	1.009	0.963	0.996	0.748	0.031	0.993	0.963	3.645	2.733	0.225	0.949	0.189

2.4 Analysis of Multiple Myeloma Data

To demonstrate the practical usefulness of our proposed NFRM methods, we apply our estimation and testing procedures to relate copy number profiles to clinical biomarkers of disease progression in MM as introduced in Section 2.1.

2.4.1 Data Description and Analysis

MM is a cancer that begins in the bone marrow and which most commonly develops in older adults. The data set is a comprehensive collection of gene expression levels, DNA copy numbers, sequencing, and RNA interference data from patients who were newly diagnosed with or previously treated for MM. In addition to demographic information such as gender and age, the data have detailed clinical information on commonly used prognostic markers, β_2M (measured in ug/dL) and serum albumin (measured in g/dL) [Bataille et al., 1983, Greipp et al., 2005]. The copy number profiles were measured using Agilent 244K aCGH arrays and consist of log2 ratios indexed by genomic location across all chromosomes (approximately 244,000 measurements). In this article, we are interested in studying the relationship between the clinical biomarkers and the copy number profile of each chromosome while controlling for the demographic covariates. In addition to the natural ordering, high dimensionality and resolution of these observations, there exists serial dependence between neighboring locations within the copy number profile. Figure 2.1 provides an illustration of the serial correlation in the copy number profiles. Along with the aforementioned features, the slow decay of the autocorrelation demonstrated in this figure suggests that the copy number profile is measured along a functional axis. These features help to justify our application of NFRM.

Using model (2.1), we conduct a genome-wide association analysis. We use β_2M as the continuous outcome and age and gender as the demographic covariates $\mathbf{z}$. We assume that $X_i(\cdot)$ is the random process that produces the observed copy number profiles. Modeling the entire copy number profile is numerically difficult due to the large number of observations that could potentially drown out any local (chromosome-specific) signals in the data. Therefore, we use a moving window approach to test and estimate the effect of local regions within the copy number profile on β_2M. Since each log2 ratio is measured across a range of probes, we first find the midprobe location for each observation and apply a sequential index to the midprobe locations. Each window consists of the observed copy number profile for 100 midprobe locations. Each adjacent window has an overlap of 50 midprobe locations. We analyze the p-arm and q-arm of each chromosome separately. The physical boundary set by the centromere of the chromosome makes this a reasonable approach. However, our goal is to test and conduct inference for the effect of the copy number profile for the entire chromosome. Thus, we use a Benjamini-Hochberg

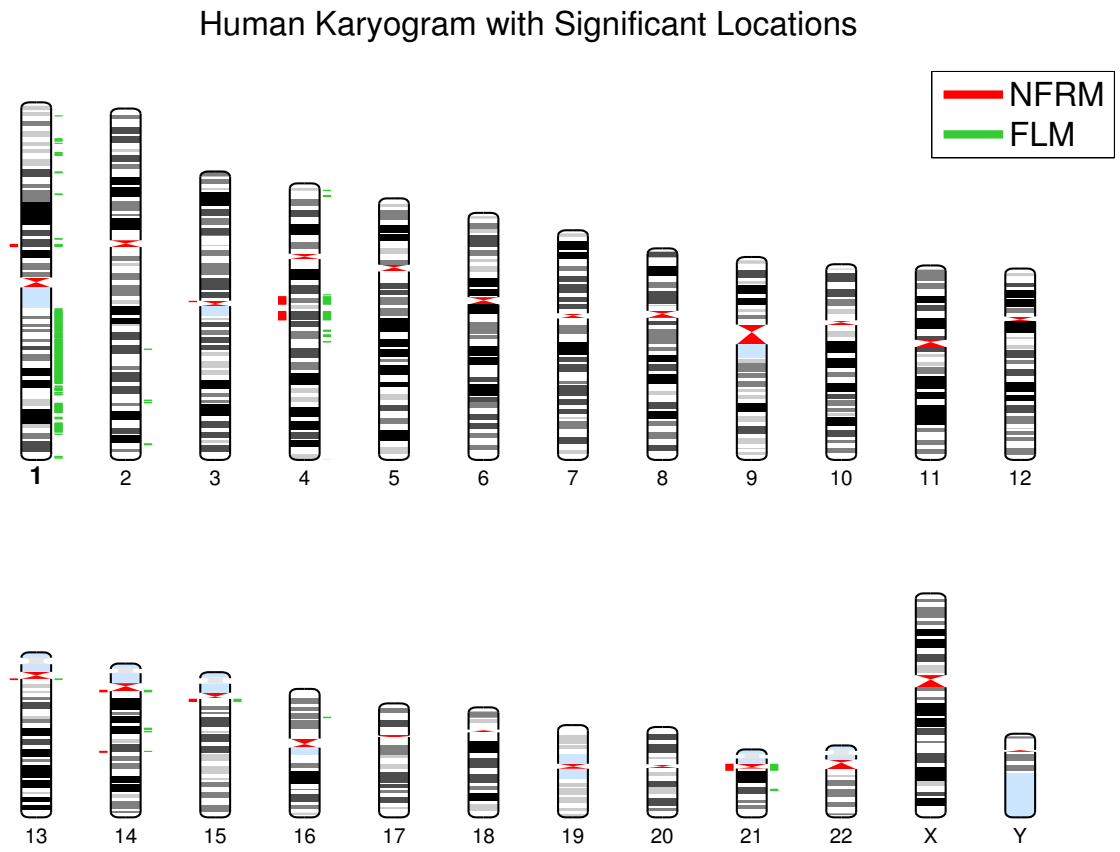


Figure 2.3: Test results for the MM application. The figure is a karyogram that depicts the test results for NFRM and FLM across the genome. Red regions to the left of each chromosome were identified by NFRM and green regions to the right were identified by FLM.

correction to adjust for the multiple tests along each arm. The moving window index is mapped back to the chromosome probe location to isolate regions that are found to be significant.

Significant locations. We conducted an genome-wide analysis across all chromosomes and found significant locations along chromosomes 1, 2, 3, 4, 14, 16, and 21. Figure 2.3 displays these significant regions on a human karyogram. These results suggest that there is an advantage to using NFRM and FLM as companion tests. There are significant regions found by NFRM but not by the FLM, and vice versa. This suggests that some regions of the genome are nonlinearly related to β_2M, whereby NFRM has higher power to detect them. Other regions have a simpler relationship and may be best detected by FLM in small to moderate sized samples.

Having tested for the effect of the copy number profile on β_2M, we estimate the size and direction of the copy number profile effect for all significant regions. To demonstrate our esti-

mation procedure, we focus our attention on window 175 of chromosome 1. This is the most significant window of chromosome 1. The results for this region suggest that it is negatively associated with β_2M. Figure 2.4 displays the estimated functional effects for each subject along with pointwise 95% confidence bands. The estimates are ordered to better distinguish the confidence intervals that do not contain zero. Recall that our methodology employs functional

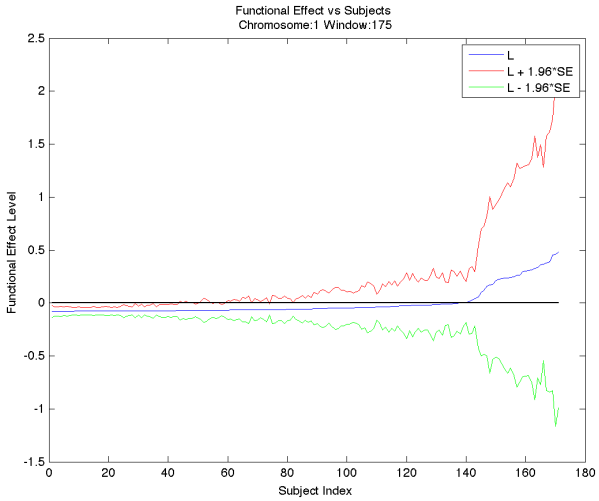


Figure 2.4: This figure displays the ordered estimated copy number profile effect (along with pointwise 95% confidence bands) by subject for window 175 of chromosome 1 p-arm.

principal components analysis by conditional expectation. It is reasonable to assume that the first and second principal components determine the size and direction of the estimated effect of the copy number profile. We are interested in visualizing the dominant functional principal components and their impact on estimation. Thus, we project the data in the positive and negative direction of the first and second principal components. We then estimate the size of the copy number profile effect along the principal component directions. This allows us to determine how the estimated effect of the copy number profile is affected by the magnitude and direction of the principal components. Figure 2.5 shows a nonlinear trend along the first principal component for window 175 of chromosome 1 p-arm, which was found to be significant by NFRM and FLM. Figure 2.6 suggests a linear trend along the first principal component for window 61 of chromosome 1 q-arm, which was solely found to be significant by FLM.

Goodness of fit. To assess the goodness of fit of NFRM and FLM, we use the quasi-R^2 measure proposed by [Yao and Müller, 2010], $R^2_Q = 1 - \sum_{i=1}^{n}(Y_i - \hat{Y})^2 / \sum_{i=1}^{n}(Y_i - \bar{Y})^2$. This measure provides a comparison of the prediction error using the sample mean of the β_2M

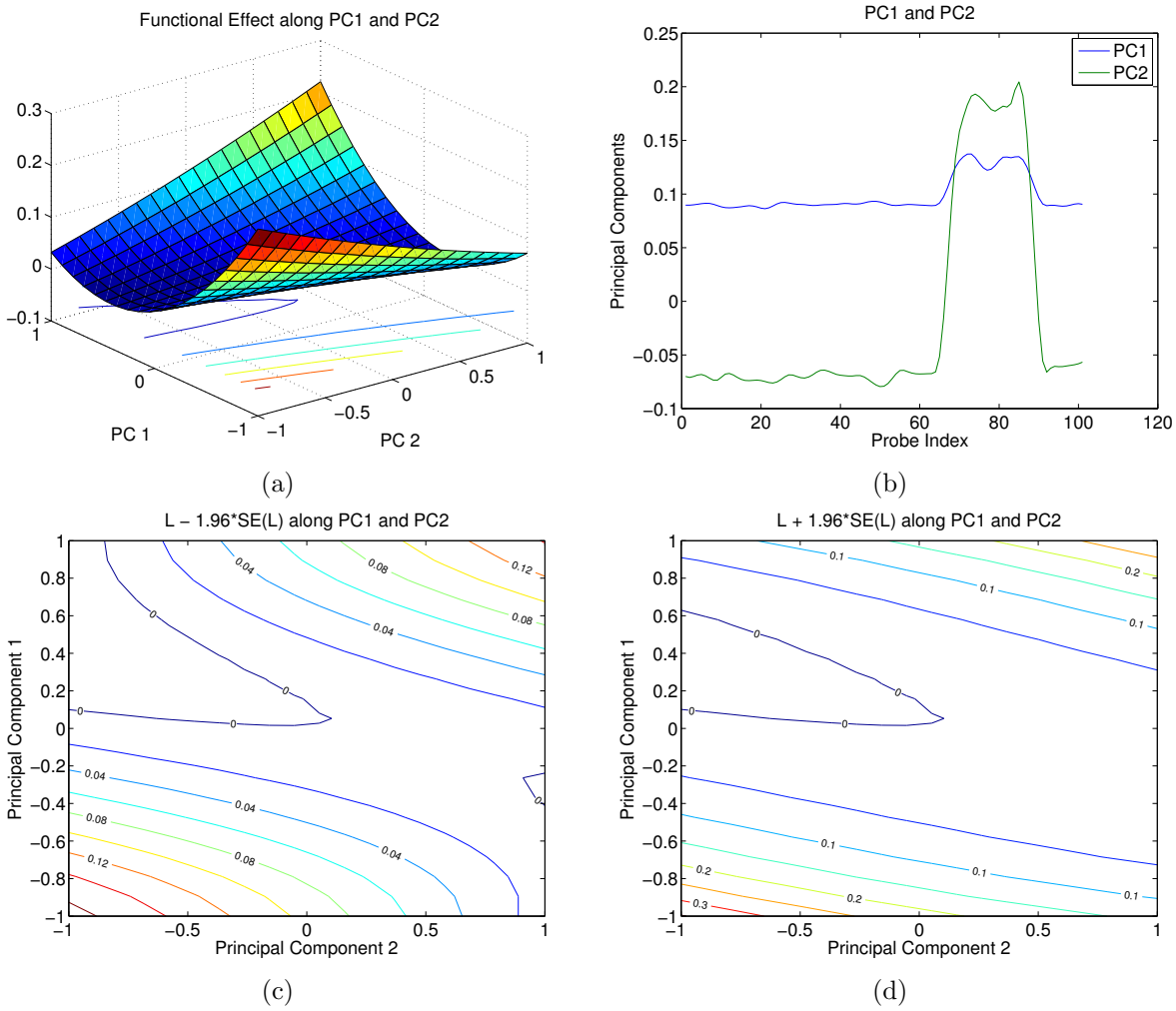


Figure 2.5: Estimation results for window 175 of the chromosome 1 p-arm from the multiple myeloma data. The top panels show the estimation results of $\mathcal{L}\{X_i(\cdot)\}$ along the direction of the 1st and 2nd principal components. The bottom panels show the upper and lower pointwise confidence bounds for $\widehat{\mathcal{L}}\{X_i(\cdot)\}$ along the direction of the 1st and 2nd principal components.

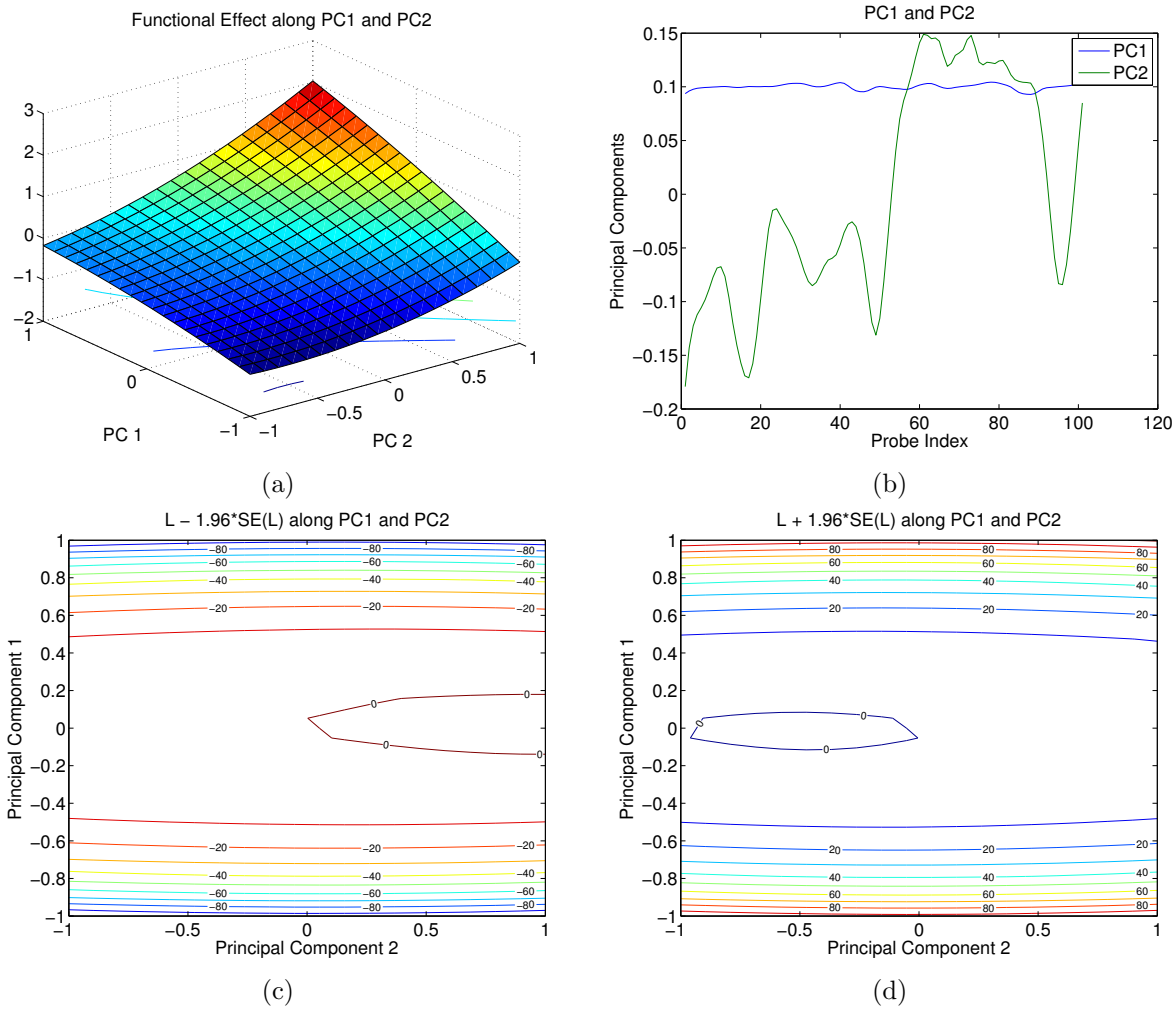


Figure 2.6: Estimation results for window 61 of the chromosome 1 q-arm from the multiple myeloma data. The top panels show the estimation results of $\mathcal{L}\{X_i(\cdot)\}$ along the direction of the 1st and 2nd principal components. The bottom panels show the upper and lower pointwise confidence bounds for $\widehat{\mathcal{L}}\{X_i(\cdot)\}$ along the direction of the 1st and 2nd principal components.

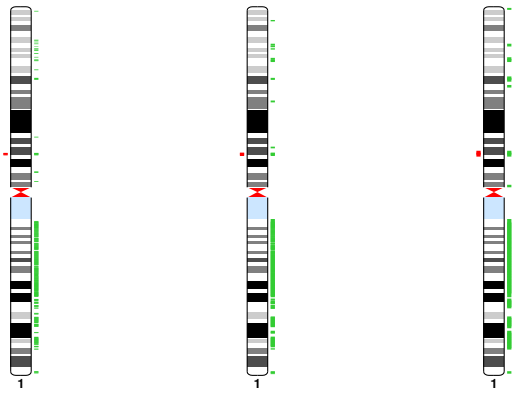


Figure 2.7: Chromosome 1 karyograms using different probe windows and overlap sizes. The left, center, and right panels show test results using window sizes of 50 probes (25 probes overlap), 100 probes (50 probes overlap) and 200 probes (100 probes overlap), respectively.

for prediction versus the proposed predictor. For window 175 of chromosome 1, NFRM yields $R_Q^2 = 0.1451$ and the FLM yields $R_Q^2 = 0.1007$. This suggests that a small deviation from the average copy number across subjects can result in a statistically significant region, which justifies the need to have tests more powerful than the test provided for the FLM.

Sensitivity to window size. We investigate the sensitivity of our moving window approach by conducting the tests using different window sizes. Specifically, we use the following window size and overlap size combinations: (50, 25), (100, 50), and (200, 100). Figure 2.7 shows similar results for each window size. Ideally, the window size should be determined by the biology of the problem, but as a practical guideline, we recommend conducting the combined test procedure using multiple window sizes and focusing on any regions detected by multiple window sizes. Alternatively, window size can be determined by estimating the auto correlation within the functional covariate using a lag equal to the desired window size. Note the serial correlation in the MM data in Figure 2.1. In this figure, we observe a steady decay of the autocorrelation over a lag of 100 probes which justifies using a window size of 100 probes for our data analysis.

2.4.2 Biological Ramifications

Using β_2M as an outcome, NFRM identifies five significant contiguous genomic locations in chromosomes 1, 3, 4 and 14. These genomic locations are shown in Table B.5. The FLM also identifies the locations in chromosomes 4 and 14 as significant. The probes in these locations appear to show a complex (linear and non-linear) relationship with the β_2M values. Locations in chromosomes 1 and 3 that are exclusively identified by NFRM demonstrate the importance of exploring nonlinear relationships. Using the FLM, we identify significant genomic locations

related to β_2M in chromosomes 1, 2, 4, 14, 16 and 21 (see Table B.5 in Appendix B).

The probes on the locations in chromosome 1 are all positively associated with β_2M, while the probes in chromosome 4 are antagonistically related to B2M, i.e. in samples with high β_2M, there is a high probability of amplification in this location or sub-location of chromosome 1 and a high probability of deletion in the region of chromosome 4.

Next, we verify the significant genes related to β_2M for their known roles in cancer progression and development. As high β_2M has been correlated with poor survival, genes that are negatively associated with β_2M are expected to be tumor suppressors and vice versa. A significant locus on chromosome 4 contains the gene GRID2 (ionotropic glutamate receptor delta 2), which is known to be located in a common fragile site where chromosomal deletions occur frequently in multiple types of cancer [Rozier et al., 2004, McAvoy et al., 2008]. The gene PAQR3 (progestin and adipoQ receptor member 3), located in chromosome 4, is a known tumor suppressor in colorectal cancer and is also negatively related to β_2M [Wang et al., 2012]. Expression of annexin A3 (ANXA3), located in chromosome 4, has been correlated negatively with tumor progression in papillary thyroid cancer [Jung et al., 2010]. Genes that have been positively related with β_2M are known to be associated with the progression of multiple myeloma. WNT3A, located in chromosome 1, is a member of the Wnt signaling pathway and has been positively correlated with β_2M. WNT3A is associated with morphological changes and rearrangement of the actin cytoskeleton in myeloma cells [Qiang et al., 2003]. Down regulation of surface molecule CD229 reduces the number of viable myeloma cells, thus CD229 is a novel target gene for the treatment of multiple myeloma [Atanackovic et al., 2011]. An upstream regulator of β_2M, USF1 (upstream stimulatory factor 1), located in chromosome 1, also has been positively associated with β_2M [Gobin et al., 2003]. In essence, the whole spectrum of genomic copy number profiles in multiple myeloma show that there are oncogenic transformations significantly associated with β_2M in locations identified by NFRM and FLM.

In addition, we use both NFRM and FLM to analyze the relation between the DNA copy number profiles in the multiple myeloma data set and the concentration of serum albumin (SA), another known clinical marker. Using NFRM, we identify multiple genomic locations as significant in chromosomes 7, 9, 14, 19 , X and Y. Using the FLM, we identify significant genomic locations related to SA in chromosomes 1, 6, 14 and Y. The regions in chromosomes 1, 6, and 14 are significantly related to SA. The probes in these regions are negatively related to SA, indicating that all the associated genes are oncogenic, as low SA concentrations indicate higher disease level in multiple myeloma [Kim et al., 2010].

Chapter 3

Testing in Generalized Nonlinear Functional Regression Models

3.1 Introduction

When prognosing cancer patients, oncologists typically assess the progression of disease via a process known as staging [Can, 2013]. For example, multiple myeloma patients are categorized as Stage I, II, or III based on the concentration levels of the proteins β_2M (measured in ug/dL) and serum albumin (measured in g/dL) [Greipp et al., 2005]. In such classifications, higher stages typically correspond to worsening prognosis. As is the case in multiple myeloma, staging algorithms often involve multiple prognostic markers and/or other clinical factors. The use of several types of markers in classification rules suggest that the underlying genetic architecture of cancer is complex and contributes to several observable phenotypes.

In Chapter 2, our goal was to investigate the association between copy number alterations and disease progression. We used a single continuous biomarker as a surrogate marker for progression. However, in our efforts to investigate the association between copy number alterations and disease progression, it may be advantageous to consider stages as an outcome. This approach has the potential to provide more in depth insight into the complex biological relationship between copy number aberrations and disease progression as explained in multiple observable phenotypes.

Again, we cast this problem as a functional regression model, except that in this chapter, we now consider cancer stage as a discrete outcome. While there are several available functional regression approaches that regress continuous responses onto functional covariates [Ramsay and Silverman, 2005, Shin, 2009, Yao and Müller, 2010], models that regress discrete responses onto functional covariates have not received far less attention in the literature. James [2002] extended

the generalized linear model framework of Nelder and Wedderburn [1972] into the functional data analysis literature. They model the random trajectory via a natural cubic splines basis and estimate the resulting parameters via the EM algorithm. Müller and Stadtmüller [2005] proposed the generalized functional linear model (GFLM). They approximated the underlying random trajectory using a Karhunen-Loéve expansion and establish a connection between their functional model and the finite-dimensional generalized linear model framework. Asymptotic tests are developed to investigate the effect of the functional effect in the reduced-dimensional setting. Other authors have also developed approaches for estimation and prediction in functional regression models with non-normal responses, such as the functional principal component logistic regression model proposed by Escabias et al. [2005] (and extended by Aguilera-Morillo et al. [2013]) as well as the penalized regression approach of Goldsmith et al. [2011].

With respect to problems in cancer genomics, the assumption of linearity remains a limitation of these previous works. Recall that it is believed that the genetic architecture underlying such diseases is thought to be very complex [Morgan et al., 2012]. Thus, for our motivating problem (and similar problems), there is a need to develop new functional regression models that, (1) model the relationship between a random curve and a discrete response nonparametrically, which lends itself to complex nonlinear relationships, and (2) test whether the functional predictor is necessary to model the discrete outcome when the effect is allowed to be nonlinear.

The NFRM developed in Chapter 2 models the effect of the random curve on the continuous response nonparametrically. Under the framework, we proposed testing procedures to test for the effect of the functional covariate. However, the approach is not suitable for discrete responses. In addition, the proposed procedures rely on the selection of a kernel function to model the nonlinear relationship between the random curve and the response. We demonstrated a "robustness" to detecting complex nonlinear forms when using the quadratic kernel to construct the model; however, we also showed that using this kernel resulted in decreased power to detect simple linear dependencies. Thus, the resulting testing procedure requires a companion test for maximum effectiveness.

In this chapter, we propose the generalized nonlinear functional regression model (GNFRM), as an extension of the NFRM. Within the proposed model framework, we regress discrete (or continuous) responses onto functional covariates. The work presented in this chapter makes contributions to three bodies of statistical literature. First, under the generalized framework, we propose testing procedures to investigate the relationship between the random curve and a discrete (or continuous) response when the effects are believed to be nonlinear. This is important because the functional data testing literature is thin, especially with respect to nonparametric functionals as defined by Ferraty and Vieu [2006]. Second, we explore the benefits of using adaptive composite kernels where weights are allowed to adapt with respect to a tuning pa-

rameter. The adapting weights allow the composite kernel to model both linear and nonlinear complex forms. This is important because it helps to mitigate the negative effects that result from poorly selecting a kernel function to model the unknown complex relationship between the functional covariate and the scalar response. We note that the adaptive composite kernel also addresses general challenges in the kernel machine literature. The resulting testing procedure is omnibus—capable of capturing linear and a wide range of nonlinear functional effects. In addition, this work continues to build the case for employing our model-based approach to investigating the association between genomic copy number and clinical outcomes.

The remainder of this chapter is organized as follows. We detail our generalized nonlinear functional regression model and the working linear mixed model in Section 3.2 and Section 3.2.1, respectively. Our proposed testing procedures are discussed in Section 3.2.2. We discuss our numerical studies in Section 3.3, and we apply our proposed testing procedure to investigate the association between genomic copy number and cancer progression in multiple myeloma patients in Section 3.4.

3.2 Generalized Nonlinear Functional Regression Model

For each subject, $i = 1, \ldots, n$, we observe a discrete (or continuous) response Y_i such as cancer stage, a q-dimensional vector of covariates $\mathbf{z}_i$ such as age and gender, and p realizations of a predictor process observed with error, $\mathbf{W}_i = (W_{i1}, \ldots, W_{ip})^{\mathrm{T}}$, where $W_{ij} = X_i(t_j) + \delta_{ij}$, $X_i(\cdot)$ is an underlying square integrable process, $t_j \in \mathcal{T}$ and δ_{ij} is mean zero white noise with finite variance. As in Chapter 2, we assume that W_{ij} is the observed copy number intensity at the jth probe location along the genome.

We assume that each Y_i is independent with conditional mean $\mathrm{E}[Y_i|\mathbf{z}_i, X_i(\cdot)] = \mu_i$ and conditional variance $\mathrm{Var}[Y_i|\mathbf{z}_i, X_i(\cdot)] = a_i(\psi)\nu(\mu_i)$ where $a_i(\psi)$ is a known function of the dispersion parameter ψ, and $\nu(\cdot)$ is a known variance function. We assume that the conditional mean is related to the covariates and a functional covariate through a known link function $g(\cdot)$,

$$g(\mu_i) = \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathcal{L}\{X_i(\cdot)\}, \tag{3.1}$$

where $\boldsymbol{\beta}$ is a q-dimensional vector of regression coefficients and $\mathcal{L}(\cdot)$ is an unknown functional that maps from $L^2[\mathcal{T}]$ to $\mathbb{R}$. To avoid identifiability issues, we assume that $\mathbf{z}_i$ contains an intercept and that $\mathrm{E}[\mathcal{L}\{X_i(\cdot)\}] = 0$ and $\mathrm{E}[X_i(\cdot)] = 0$.

The generalized functional linear model of Müller and Stadtmüller (2005) is a special case of model (3.1) where $\mathcal{L}\{X_i(\cdot)\} = \int_{\mathcal{T}} X_i(t)\beta(t)dt$. Similar to Chapter 2, model (3.1) allows $\mathcal{L}(\cdot)$ to be any functional that maps a square integrable function defined on $\mathcal{T}$ to $\mathbb{R}$. To determine

if the copy number profile is associated with a discrete disease outcome, we develop procedures to test for the effect of the functional covariate on the conditional mean of the response, where in particular we are interested in the following hypothesis:

$$H_0 : \mathcal{L}(\cdot) = 0 \text{ vs. } H_1 : \mathcal{L}(\cdot) \neq 0 \tag{3.2}$$

3.2.1 Connection to Generalized Linear Mixed Models

In this section, we establish a connection with the generalized linear mixed model framework, whereby we propose procedures to test (3.2). We begin by establishing a finite-dimensional representation of $X(\cdot)$. Given this representation, we define an approximate model that is subsequently cast into the generalized linear mixed model framework.

Working Mixed Model

The first step in developing our testing procedure is to reduce the dimension of the problem. Following Section 2.2.1, we use the the Karhunen-Loéve expansion to obtain the following finite representation of the underlying smooth function, $X_i(t) = \sum_{j=1}^{J} \xi_{ij}\phi_j(t)$ for $i = 1, \ldots, n$. Again, a key step is to perform a functional principal component analysis to estimate the eigenfunctions $\{\phi_1, \ldots, \phi_J\}$, and we obtain consistent estimates of the principal scores via PACE, which serve as proxies for the true scores in practice. More technical details on PACE are provided in Appendix A.1.

Goldsmith et al. [2012] notes that the asymptotics developed in Yao et al. [2005] may be viewed as justification for the common approach of *overlooking* the uncertainty surrounding the FPCA process in large samples. However, we note that in small to moderate samples, the increased variability from the FPCA process may be non-negligible. Failing to account for this uncertainty can lead to inflated type I error rates in our proposed testing procedure. We examine the effect of this common practice in our numerical studies.

Conditional on $X_i(\cdot)$, J, and $\boldsymbol{\xi}_i = (\xi_{i1}, \ldots, \xi_{iJ})^{\mathrm{T}}$, we leverage the information in these scores to approximate $\mathcal{L}\{X(\cdot)\}$ with a smooth function $\mathcal{L}^*(\boldsymbol{\xi}_i) : \mathbb{R}^J \mapsto \mathbb{R}$. Similar to Chapter 2, we can justify the approximation by considering the generalized functional linear model where $\mathcal{L}\{X_i(\cdot)\} = \int_{\mathcal{T}} X_i(t)\beta(t)dt$. Given $\beta(t) \in L^2[\mathcal{T}]$, we have that $\beta(t) \approx \sum_{j=1}^{J} \eta_j\phi_j(t)$, where η_j is the unknown coefficient that corresponds to ϕ_j. Given the orthonormality of each eigenfunction, we have that $\mathcal{L}\{X_i(\cdot)\} \approx \sum_{j=1}^{J} \eta_j\xi_{ij}$.

Given this heuristic justification, we propose an approximate model for the conditional mean,

$$g(\mu_i) = \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathcal{L}^*(\boldsymbol{\xi}_i), \tag{3.3}$$

where $\mathcal{L}^*(\boldsymbol{\xi}_i)$ is a smooth function with finite-dimensional argument. Following Chapter 2, we further assume that $\mathcal{L}^*(\cdot)$ is a mean zero Gaussian process where the behavior of the process is completely determined by covariance function $\tau K(\cdot,\cdot)$. Here, τ is an unknown variance component and $K(\cdot,\cdot)$ is a kernel function such that $K(\boldsymbol{\xi}_l,\boldsymbol{\xi}_k) = \text{Cov}\{\mathcal{L}^*(\boldsymbol{\xi}_l),\mathcal{L}^*(\boldsymbol{\xi}_k)\}$ for $l,k = 1,\ldots,n$. Given $\boldsymbol{\xi}_1\ldots,\boldsymbol{\xi}_n$, we can express (3.3) as a working generalized linear mixed model

$$g(\boldsymbol{\mu}) = \mathbf{Z}\boldsymbol{\beta} + \mathcal{L}^*, \tag{3.4}$$

where $\mathbf{Z} = [\mathbf{z}_1,\ldots,\mathbf{z}_n]^{\mathrm{T}}$ is a matrix of fixed covariates and $\mathcal{L}^* = [\mathcal{L}^*(\boldsymbol{\xi}_1),\ldots,\mathcal{L}^*(\boldsymbol{\xi}_n)]^{\mathrm{T}}$ is an n-dimensional vector of random variables such that $\mathcal{L}^* \sim N(0,\tau\mathbf{K})$. Here, $\mathbf{K}$ is the positive definite gram matrix with $\mathbf{K}_{lk} = K(\boldsymbol{\xi}_l,\boldsymbol{\xi}_k)$.

Similar to NFRM, the kernel function $K(\cdot,\cdot)$ performs several roles within this framework. However, in this chapter we separate the kernel functions into two classes to support the development of a new type of composite kernel in Section 3.2.2.

Class 1 kernel functions are those with no tuning parameters such as the quadratic kernel function discussed in Section 2.2.1, $K(\boldsymbol{\xi}_l,\boldsymbol{\xi}_k) = (\boldsymbol{\xi}_l^T\boldsymbol{\xi}_k + 1)^2$. Recall that this kernel assumes that the relationship between the conditional mean and random process is quadratic.

Class 2 kernel functions are those with a tuning parameter such as the Gaussian kernel function which was also discussed in Section 2.2.1, $K(\boldsymbol{\xi}_l,\boldsymbol{\xi}_k) = \exp\{(\sum_{j=1}^{J}(\xi_{lj} - \xi_{kj})^2/\rho^2\}$, where ρ is an unknown tuning parameter. Recall that this kernel assumes that the relationship between conditional mean and the random process is nonlinear where the tuning parameter ρ determines the nonlinearity. We emphasize here that as ρ increases toward infinity, the Gaussian kernel models a simple linear relationship between the conditional mean and the random process [Keerthi and Lin, 2003]. This feature of the Gaussian kernel is a critical component of the adaptive composite kernel that we present in Section 3.2.2.

3.2.2 Testing for the Effect of the Functional Covariate

Under model (3.4), the hypothesis expressed in (3.2) is equivalent to

$$H_0 : \tau = 0 \text{ vs. } H_1 : \tau > 0. \tag{3.5}$$

To test this hypothesis, we follow the profile quasi-likelihood approach of Breslow and Clayton [1993]. Denote a working dependent vector $\tilde{\mathbf{Y}} = \mathbf{Z}\boldsymbol{\beta} + \mathcal{L}^* + (\mathbf{Y} - \boldsymbol{\mu})\text{diag}\{g'(\boldsymbol{\mu})\}$. This yields the working linear mixed model $\tilde{\mathbf{Y}} = \mathbf{Z}\boldsymbol{\beta} + \mathcal{L}^* + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim N(\mathbf{0},\boldsymbol{\Sigma}_\psi^{-1})$, and $\boldsymbol{\Sigma}_\psi$ is a diagonal matrix with diagonal terms $w_i = \{a_i(\psi)\nu(\mu_i)[g'(\mu_i)]^2\}^{-1}$. Our inference on τ is based on the

REML version of the profiled quasi-likelihood that is displayed below,

$$l_R(\tau|\hat{\boldsymbol{\beta}}_0, \hat{\psi}) = -\frac{1}{2}\log|\mathbf{V}_{\hat{\psi}}| - \frac{1}{2}\log|\mathbf{Z}^{\mathrm{T}}\mathbf{V}_{\hat{\psi}}^{-\mathbf{1}}\mathbf{Z}| - \frac{1}{2}(\tilde{\mathbf{Y}} - \mathbf{Z}\hat{\boldsymbol{\beta}}_0)^{\mathrm{T}}\mathbf{V}_{\hat{\psi}}^{-1}(\tilde{\mathbf{Y}} - \mathbf{Z}\hat{\boldsymbol{\beta}}_0), \tag{3.6}$$

where $\mathbf{V}_\psi = \tau\mathbf{K} + \boldsymbol{\Sigma}_\psi^{-1}$ and $\hat{\boldsymbol{\beta}}_0$ and $\hat{\psi}$ are estimated under the null model $g(\boldsymbol{\mu}) = \mathbf{Z}\boldsymbol{\beta}$. Differentiating (3.6) with respect to τ and evaluating the resulting equation at the null yields the following score equation,

$$\mathcal{U}_\tau = \frac{1}{2}(\tilde{\mathbf{Y}} - \mathbf{Z}\hat{\boldsymbol{\beta}}_0)^{\mathrm{T}}\boldsymbol{\Sigma}_{\hat{\psi}}\mathbf{K}\boldsymbol{\Sigma}_{\hat{\psi}}(\tilde{\mathbf{Y}} - \mathbf{Z}\hat{\boldsymbol{\beta}}_0) - \frac{1}{2}\mathrm{tr}(\mathbf{P}_0\mathbf{K}), \tag{3.7}$$

where, $\mathbf{P}_0 = \boldsymbol{\Sigma}_{\hat{\psi}} - \boldsymbol{\Sigma}_{\hat{\psi}}\mathbf{Z}(\mathbf{Z}^{\mathrm{T}}\boldsymbol{\Sigma}_{\hat{\psi}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}}\boldsymbol{\Sigma}_{\hat{\psi}}$. Conditioned on $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n$ and $\hat{\psi}$, we have $\mathrm{E}(\mathcal{U}_\tau) = \frac{1}{2}\mathrm{tr}(\mathbf{P}_0\mathbf{K})$ and $\mathrm{VaR}(\mathcal{U}_\tau) = \frac{1}{2}\mathrm{tr}(\mathbf{P}_0\mathbf{K}\mathbf{P}_0\mathbf{K})$ (which is the Fisher Information).

The testing procedures that we propose are based on (3.7), which is a function of $\mathbf{K}$. Recall that class 2 kernel functions depend on an unknown tuning parameter ρ. In such cases, it's clear to see that ρ is not estimable under $H_0 : \tau = 0$. This prevents us from using standard large sample testing procedures such as the score test. Thus, we propose separate testing procedures for each class of kernel function.

Testing: Kernels with no Tuning Parameters

In Section 3.2.1, we briefly introduced two classes of kernel functions. In this section, we develop a testing procedure for class 1 kernel functions, which we refer to as *Test 1*.

Conditioned on $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n$ and $\hat{\psi}$, the second term of (3.7) is a constant. Thus, we define the following test statistic based on the first term of (3.7),

$$\mathcal{Q}_\tau = \frac{1}{2}(\tilde{\mathbf{Y}} - \mathbf{Z}\hat{\boldsymbol{\beta}}_0)^{\mathrm{T}}\boldsymbol{\Sigma}_{\hat{\psi}}\mathbf{K}\boldsymbol{\Sigma}_{\hat{\psi}}(\tilde{\mathbf{Y}} - \mathbf{Z}\hat{\boldsymbol{\beta}}_0) = \frac{1}{2}(\mathbf{Y} - \hat{\boldsymbol{\mu}}_0)^{\mathrm{T}}\mathbf{K}(\mathbf{Y} - \hat{\boldsymbol{\mu}}_0),$$

where $\hat{\boldsymbol{\mu}}_0 = \mathrm{logit}^{-1}(\mathbf{Z}\hat{\boldsymbol{\beta}}_0)$.

It can be shown that the distribution of $\mathcal{Q}_\tau$ follows a mixture of χ_1^2 distributions under $H_0 : \tau = 0$ [Zhang and Lin, 2003]. Critical values for this distribution are difficult to ascertain, thus we approximate the distribution of $\mathcal{Q}_\tau$ with a scaled chi-square distribution $\kappa\chi_\nu^2$. Here, the Satterthwaite method is used to estimate the scale parameter and the degrees of freedom by matching the first two moments of $\mathcal{Q}_\tau$ to those of a $\kappa\chi_\nu^2$ distribution. Calculations show that $\hat{\kappa} = \sigma_{\mathcal{Q}}^2/2\mu_{\mathcal{Q}}$ and $\hat{\nu} = 2\mu_{\mathcal{Q}}^2/\sigma_{\mathcal{Q}}^2$. However, to account for the use of $\hat{\boldsymbol{\beta}}_0$ and $\hat{\psi}$, we replace $\sigma_{\mathcal{Q}}^2$ with the efficient information, $\tilde{\sigma}_{\mathcal{Q}}^2 = \sigma_{\mathcal{Q}}^2 - I_{\tau,\psi}I_{\psi,\psi}^{-1}$, $I_{\tau,\psi}$, where $I_{\tau,\psi} = \frac{1}{2}\mathrm{tr}(\mathbf{P}_0\mathbf{K}\mathbf{P}_0)$ and $I_{\psi,\psi} = \frac{1}{2}\mathrm{tr}(\mathbf{P}_0\mathbf{P}_0)$. Thus under the null, the final test statistic $\mathcal{S}_\tau = \mathcal{Q}_\tau/\hat{\kappa}$ has an approximate $\chi_{\hat{\nu}}^2$ distribution.

Testing: Kernels with Tuning Parameters

We now consider class 2 kernel functions which depend on some unknown tuning parameter ρ. Given that ρ is not estimable under the null, the test statistic derived in Section 3.2.2 is not applicable. Thus, we propose an alternative testing procedure for this class of kernels which we refer to as *Test 2*.

Davies [1987] develops testing procedures in cases where nuisance parameters, such as ρ, are present only under the alternative hypothesis. Based on this work, we propose a score based test statistic for $H_0 : \tau = 0$. We define our test statistic as

$$\mathcal{S}_\tau(\rho) = [\mathcal{Q}_\tau(\rho) - \mu_\mathcal{Q}]/\sigma_\mathcal{Q}.$$

Under suitable regularity conditions, we assume that $\mathcal{S}_\tau(\rho)$ is approximately a Gaussian process indexed by ρ. Following Davies [1987], we compute a sharp upper bound for the p-value based on the expected number of upcrossings of $\mathcal{S}_\tau(\rho)$ over an appropriate range of ρ values. The sharp bound for the p-value is computed as follows,

$$\mathrm{P}(\mathcal{S}_\tau(\rho) \geq M : L \leq \rho \leq U) = \Phi(-M) + H\exp(-M^2/2)/\sqrt{8\pi}, \tag{3.8}$$

where $\Phi(\cdot)$ is the standard normal distribution function, M is the maximum of $\mathcal{S}_\tau(\rho)$ over the range of ρ, $H = |\mathcal{S}_\tau(\rho_1) - \mathcal{S}_\tau(L)| + |\mathcal{S}_\tau(\rho_2) - \mathcal{S}_\tau(\rho_1)| + \cdots + |\mathcal{S}_\tau(U) - \mathcal{S}_\tau(\rho_m)|$, where $\rho_1, \ldots, \rho_m$ are the grid of ρ values between L and U.

Testing: Composite Kernels

Given the many roles of the kernel function, it's easy to see that the choice of kernel function can impact the power the determine if the functional covariate is significantly related to the condition mean of the response. While choosing an optimal kernel remains an open problem in the kernel machine regression literature, there has been recent work to mitigate the negative effects of choosing a *less than optimal* kernel from a set of candidate kernels. Wu et al. [2013] proposes a simple kernel averaging technique to accomplish this goal.

Consider a set of L candidate kernel functions, $K_1(\cdot,\cdot), \ldots, K_L(\cdot,\cdot)$. We define a composite kernel function $K_c(\boldsymbol{\xi}_l, \boldsymbol{\xi}_k) = \sum_{j=1}^{L} w_j K_j(\boldsymbol{\xi}_l, \boldsymbol{\xi}_k)$, where $w_1, \ldots, w_L$ are appropriately defined weights. In particular, we define $w_j = \mathrm{tr}\{\mathbf{P}_0^{1/2}\mathbf{K}_j\mathbf{P}_0^{1/2}\}^{-1}$ where $\mathbf{P}_0$ is defined as above [Wu et al., 2013]. The composite kernel function produces the composite kernel matrix $\mathbf{K}_c = \sum_{j=1}^{L} w_j\mathbf{K}_j$. The key idea is that if we are willing to assume that the optimal kernel is one of the L candidate kernels, then we expect the performance of the composite kernel to be close to the optimal kernel.

When no candidate kernel is a function of an unknown tuning parameter, we treat $K_c(\cdot,\cdot)$ as a class 1 kernel. This is the case investigated by Wu et al. [2013]. However, when a single candidate kernel contains a tuning parameter, we treat $K_c(\cdot,\cdot;\rho)$ as a class 2 kernel. To illustrate the ideas, we use the following composite kernel function: $K_c(\cdot,\cdot;\rho) = w_q K_q(\cdot,\cdot) + w_g K_g(\cdot,\cdot;\rho)$, where K_q is the quadratic kernel function and K_g is the Gaussian kernel function with respective weights w_q and w_g. For each $\rho \in [L,U]$, we compute w_g, $K_g(\cdot,\cdot;\rho)$ and $K_c(\cdot,\cdot;\rho)$, where the standardized test statistic $\mathcal{S}_\tau(\rho)$ is computed using $K_c(\cdot,\cdot;\rho)$.

Recall that the performance of Test 2 relies on establishing an appropriate range of ρ values. As $\rho \to \infty$, $\mathbf{K}_g \to \mathbf{J}_{n\times n}$, where $\mathbf{J}_{n\times n}$ is a matrix of all ones. This implies that $\text{tr}\{\mathbf{P}_0^{1/2}\mathbf{K}_g\mathbf{P}_0^{1/2}\} = \text{tr}\{\mathbf{P}_0\mathbf{K}_g\} \to \text{tr}\{\mathbf{P}_0\mathbf{J}_{n\times n}\} = 0$ as $\rho \to \infty$, which further implies that $w_g \to \infty$ as $\rho \to \infty$. Thus, the Gaussian kernel dominates the composition for suitably large values of ρ. This is a desirable property, because it is well-known that as $\rho \to \infty$, the Gaussian kernel converges to the linear kernel [Keerthi and Lin, 2003]. This enables the class 2 composite kernel to capture simple linear relationships. Similarly, as $\rho \to 0$, $w_g \to \text{tr}\{\mathbf{P}_0\}$. This implies that the Gaussian kernel has a reduced impact on the composition for suitably small ρ. Herein lies the advantage of considering composite kernels where a single kernel is a function of a tuning parameter. They are capable of adapting to simple linear relationships as well as a wide range of nonlinear relationships determined by the combined feature spaces of the kernels within the composition.

We now consider selecting L and U. In the case of a single Gaussian kernel, this range has been studied by Liu et al. [2008]. They propose setting $L = 0.1 * \min_{i\neq k}\sum_{j=1}^{J}(t_{ij} - t'_{kj})^2$ and $U = 100 * \max_{i\neq k}\sum_{j=1}^{J}(t_{ij} - t'_{kj})^2$. Given that ρ affects $K_c(\cdot,\cdot;\rho)$ only through $K_g(\cdot,\cdot;\rho)$, we believe that these boundaries are also appropriate for $K_c(\cdot,\cdot;\rho)$.

3.3 Simulation Study

We conduct simulation studies to evaluate the finite sample performance of the GNFRM using both class 1 and class 2 kernels. Our experiments are designed to evaluate how well the GNFRM controls type I error rate and power. We focus our attention to the special case of functional binary regression where $g(\pi_i) = \text{logit}(\pi_i)$, $\nu(\pi_i) = \pi_i(1-\pi_i)$, and $a_i(\psi) = 1$.

Similar to 2.3.1, we generate data from the following model,

$$\text{logit}(\pi_i) = \beta_1 z_{i,1} + \beta_2 z_{i,2} + h\{X_i(\cdot)\}, \tag{3.9}$$

where $z_{i,1} \sim \text{N}(0,1)$ for standardized age and $z_{i,2} \sim \text{Bin}(1, 0.66)$ for gender. Recall that the proportion of males in the Multiple Myeloma dataset is 0.66.

Again, we mimic the copy number profiles by generating the true underlying trajectories,

$X_i(t)$, from an orthonormal Fourier basis with five basis functions, where each coefficient is an independent realization of a N(0,0.5) distribution. The observed copy number intensities are generated as $W_{ij} = X_i(t_j) + \delta_{ij}$, where $\delta_{ij} \sim \text{N}(0, 0.16)$. We set $\beta_1 = 0.7$ and $\beta_2 = -0.7$.

From this model, $\mathcal{L}\{X_i(\cdot)\} = h\{X_i(\cdot)\} - \alpha$, where $\alpha = \frac{1}{n}\sum_{i=1}^{n} h\{X_i(\cdot)\}$, which enforces the aforementioned assumptions $\text{E}[\mathcal{L}\{X_i(\cdot)\}] = 0$ and $\text{E}[X_i(\cdot)] = 0$. To compare the performance of this approach with NFRM, we explore four of the functional effects that were explored in that Chapter 2:

1. Linear Functional: $h(f) = \int f(t)\gamma(t)\ dt$;
2. Quadratic Functional: $h(f) = \{\int f(t)\gamma(t)\ dt\}^2$;
3. Linear Functional of Squared 1^{st} Derivative: $h(f) = \int f'(t)^2\gamma(t)\ dt$;
4. Signed Square Root of 2^{nd} Derivative: $h(f) = sgn\{\int f''(t)\gamma(t)\ dt\} * \sqrt{|\int f''(t)\gamma(t)\ dt|}$.

In these functionals, $\gamma(t) = c\ \sqrt{2}\ \sin(2\pi t)$, where c is the effect level. For each functional, we use 8 increasing and equally spaced effect levels from the interval $[0, 1]$. The null model occurs when $c = 0$. In addition, we consider two sample sizes, $n = 200$ and 300, we set FVE = 0.99, and we generate 1,000 data sets for each setting.

Following Section 3.2.2, we compute S_τ and $S_\tau(\rho)$ using the quadratic and Gaussian kernels, respectively. When using the quadratic kernel, S_τ is compared to a χ^2_v distribution, where k and v are estimated using the Satterthwaite approach. When using the Gaussian kernel, the sharp upper-bound for the p-value is obtained via the approach of Davies [1987]. Furthermore, we explore the advantages of using an adaptive composite kernel function that averages across the quadratic and Gaussian kernel functions.

We compare the GNFRM to the generalized functional linear regression model (GFLM). In particular, both testing procedures of the GNFRM are compared to the GFLM's Wald test which was proposed by Müller and Stadtmüller [2005]. The Wald test is based on the following functional logistic model which assumes that $X_i(\cdot)$ is linearly related to the conditional mean of Y_i via the regression function $\gamma(\cdot)$:

$$\text{logit}(\pi_i) = \alpha + \mathbf{z}_i^{\text{T}}\boldsymbol{\beta} + \int_{\mathcal{T}} X_i(t)\gamma(t)\ dt. \tag{3.10}$$

This model assumes that $X_i(\cdot)$ is observed without error, whereby a classical approach to functional principal component analysis is used to reduce (3.10) to the following model:

$$\text{logit}(\pi_i) = \alpha + \mathbf{z}_i^{\text{T}}\boldsymbol{\beta} + \widehat{\boldsymbol{\xi}}_i^{\text{T}}\boldsymbol{\gamma}. \tag{3.11}$$

Table 3.1: Simulation results for type I error rate based on 1,000 generated datasets and $n = 300$. Standard errors for each estimate < 0.001.

Type I Error	GNFRM_G	GNFRM_Q	GNFRM_CK	GFLM
0.01	0.016	0.008	0.013	0.020
0.05	0.040	0.033	0.046	0.061
0.10	0.069	0.081	0.085	0.103

Definitions: GNFRM_G, GNFRM with Gaussian kernel; GNFRM_Q, GNFRM with quadratic kernel; GNFRM_CK, GNFRM with composite kernel

In this model, $\widehat{\boldsymbol{\xi}}_i$ is a J-dimensional vector of consistent estimates of the functional principal component scores for the ith subject. We generalize this model such that we relax the assumption that $X(\cdot)$ is observed without error. Under the relaxed assumption, we use PACE to obtain the consistent estimate $\widehat{\boldsymbol{\xi}}_i$. Note that this model is conditional on $\widehat{\boldsymbol{\xi}}_i$, thus the asymptotic properties of the proposed Wald test are unaffected by our use of PACE, and we assume that the added variability from our FPCA process is negligible.

The proposed Wald test is $Z = (\widehat{\boldsymbol{\gamma}}^{\mathrm{T}}\hat{\Gamma}^{-1}\widehat{\boldsymbol{\gamma}} - J)/\sqrt{2J}$, where $\widehat{\boldsymbol{\gamma}}$ is the MLE of Eq. (3.11) found using iteratively reweighted least squares and J and $\hat{\Gamma}$ are the dimension and the estimated covariance of $\widehat{\boldsymbol{\gamma}}$, respectively. Under the null hypothesis, $Z \sim N(0, 1)$. We estimate the empirical size and power of each approach using the empirical rejection probability which is the average number of p-values less than a significance level of 5%.

Table 3.1 provides the results of our type I error rate investigations for $n = 300$. The results show that the empirical sizes of the Wald test are inflated, especially in the case of small type I error rates. GNFRM with a class 2 composite kernel controls type I error rates of 5% and 10% well, however at the smaller 1%, the size of the test is slightly inflated. Similar performance holds for GNFRM with a Gaussian kernel, while GNFRM with a quadratic kernel performs well at each type I error rate investigated. The results for $n = 200$ are similar. They are provided in Appendix C.

Figure 3.1 displays the empirical power results for $n = 300$. The results suggests that in the case of the linear functional effect, the Wald test outperforms all the GNFRM tests. However, we observe that the performance of the GNFRM with a Gaussian kernel is close to the performance of the Wald test which provides empirical evidence of the linearity of the Gaussian kernel feature space as $\rho \to \infty$. In the case of the quadratic functional effect, the GNFRM with a quadratic kernel outperforms all other tests. While the Gaussian kernel is able to detect this type of nonlinearity, its performance is not close to the quadratic kernel, and the Wald test fails to

detect the quadratic relationship. We also observe that the quadratic kernel performs best in the case of our arbitrary complex functional effects 3 and 4. Again the Gaussian kernel detects the nonlinearity but not as well as the quadratic kernel; while the Wald test fails to detect the nonlinear relationship.

Figure 3.1 also highlights the advantage of using an appropriately defined adaptive kernel. Here we see that in the case of the linear functional effect, the composite kernel performs equally as well as the Gaussian kernel. This can be explained by the adapting of the weight w_g to the magnitude of ρ as discussed in Section 3.2.2. In the case of the quadratic functional effect, the composite kernel's performance is close to the performance of the quadratic kernel (which is the optimal kernel for this functional effect) while exhibiting a large performance gain over the Gaussian kernel. Again, the weight w_g adapts and allows the quadratic kernel to dominate the composition. Similar results hold for functionals 3 and 4.

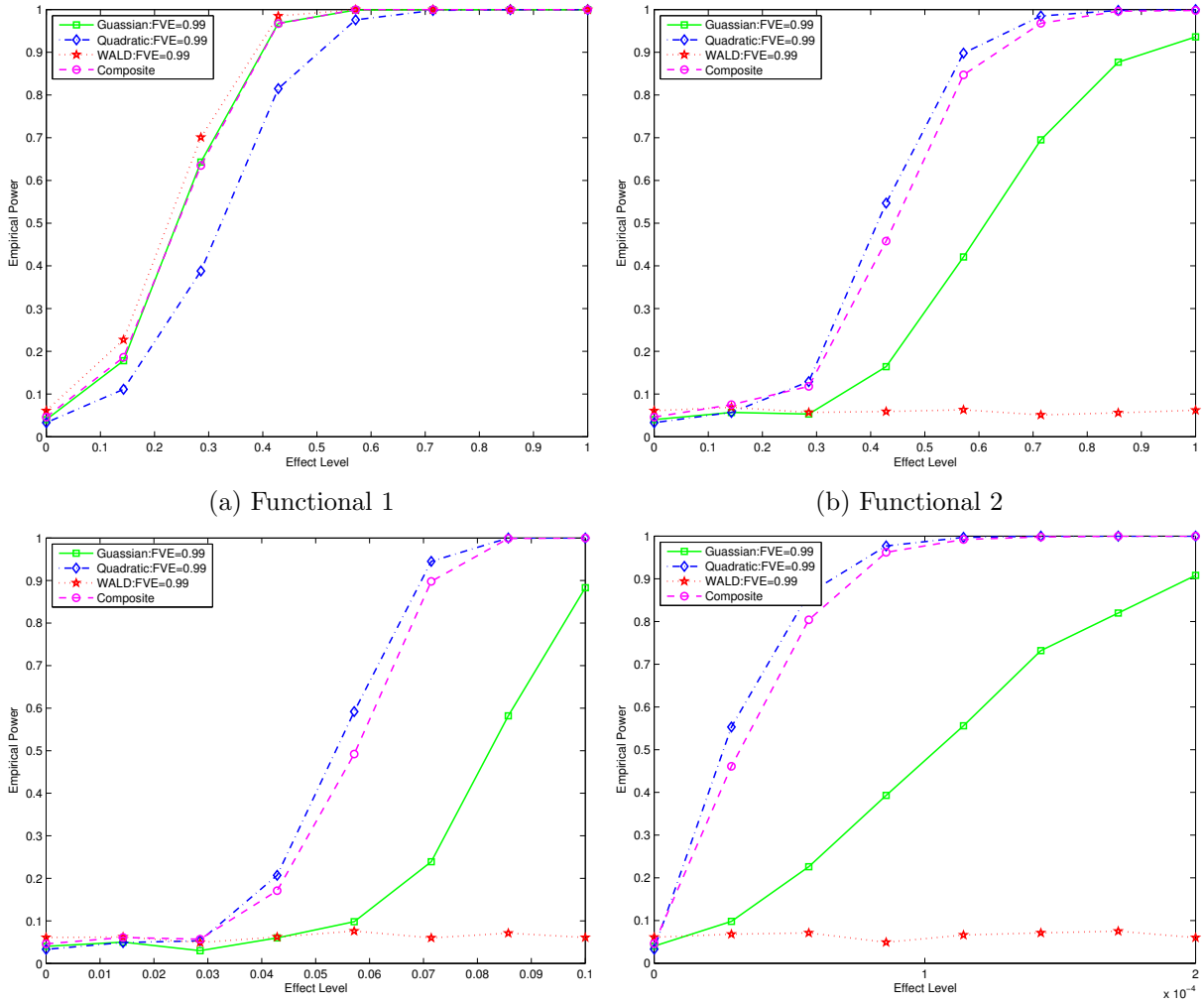
In general, these results suggest that using the GNFRM with an appropriately defined adaptive kernel can be nearly as powerful as the test designed for linear functional effects while also maintaining reasonable power to detect a wide range of nonlinear functional effects. Thus, GNFRM with a class 2 adaptive kernel can be viewed as an omnibus test capable of detecting any type of functional relationship between the random trajectory and the conditional mean of the response with moderate sample sizes.

3.4 Analysis of Multiple Myeloma Data

In this section, we apply the GNFRM to the same multiple myeloma data set investigated in Chapter 2. Recall that this data consists of approximately 244,000 copy number observations per subject indexed across the genome. Following Greipp et al. [2005], we stage each subject as Stage I, II, or III using the recorded values of the prognostic markers beta2 microglobulin (measured in ug/dL) and serum albumin (measured in g/dL). We are interested in studying the effect of local regions of copy number alterations on the progression of multiple myeloma.

We consider only the 162 complete cases. Four patients are categorized as stage I, five patients are categorized as stage II, and 153 patients are categorized in the advanced stage III. Furthermore, we consider disease progression as an advancement from the less severe stages (I and II) to stage III. Thus, we dichotomize the stages into a binary outcome where $Y_i = 1$ if the ith subject has stage III cancer and $Y_i = 0$ otherwise.

Using Eq. (3.1) with the class 2 adaptive kernel described in Section 3.2.2, we conduct a genome-wide analysis using a logit link function to relate the conditional mean of the dichotomized outcome to the copy number profile. We control for age and gender ($\mathbf{z}_i$) in our analysis. We note that we assume that $X_i(\cdot)$ is the random process that produces the observed



(a) Functional 1

(b) Functional 2

(c) Functional 3

(d) Functional 4

Figure 3.1: This figure displays the Type I Error and power for each functional at FVE = 0.99 and $n = 300$. Here the composite kernel is included in the analysis. The solid line represents GNFRM with the Gaussian kernel; the dotted-dashed line represents GNFRM with the quadratic kernel;the dashed line represents GNFRM with a composite kernel; the dotted line represents the WALD test for the FLM.

copy number profiles. Following our analysis in Chapter 2, we use a similar moving window approach to isolate local effects of the copy number profile on disease progression, and we use the Benjamini-Hochberg correction to adjust for multiple comparisons.

Figure 3.2 displays the results of our investigations. GNFRM found statistically significant regions on chromosomes 4, 5, 6, 12, 13, 14, 15, 19, and 20. We compared our results to the work of Avet-Loiseau et al. [2009], where they identified three genomic regions that were associated with survival among 192 patients who were newly diagnosed with MM. Their work identified amplifications at cytobands 1q23.3 and 5q31.3 and deletions at 12p13.31. Although, our investigations focused on a different clinical endpoint, we also detected the reported copy number alterations within 5q31.3 and 12p13.31. The details of our analysis can be viewed in Figure 3.3. We stress the detection of the significant genomic region 12p13.31, as this is the sole location that the GNFRM testing procedure found to be significant over this entire chromosome. In addition, our exploratory analysis suggests that there are several additional regions of copy number alteration that may help to explain the progression of MM to more severe disease states.

To further illustrate the potential of our approach in investigating copy number association using functional data methods, we draw our attention to Chromosome 5. Figure 3.4 provides a Manhattan plot of Chromosome 5, which corresponds to Figure 3.3. The purpose of this figure is to illustrate the level of significance across our moving window analysis. The p-values are displayed on the $-\log_{10}$ scale to help distinguish between significance levels. In this figure, window 202 is located in cytoband 5q31.3 and is found to be mildly significant. However, there are several genomic regions along the p-arm that contain a much stronger signal. This suggests other interesting genomic regions to explore to help understand the biological mechanisms that contribute to the progression of MM.

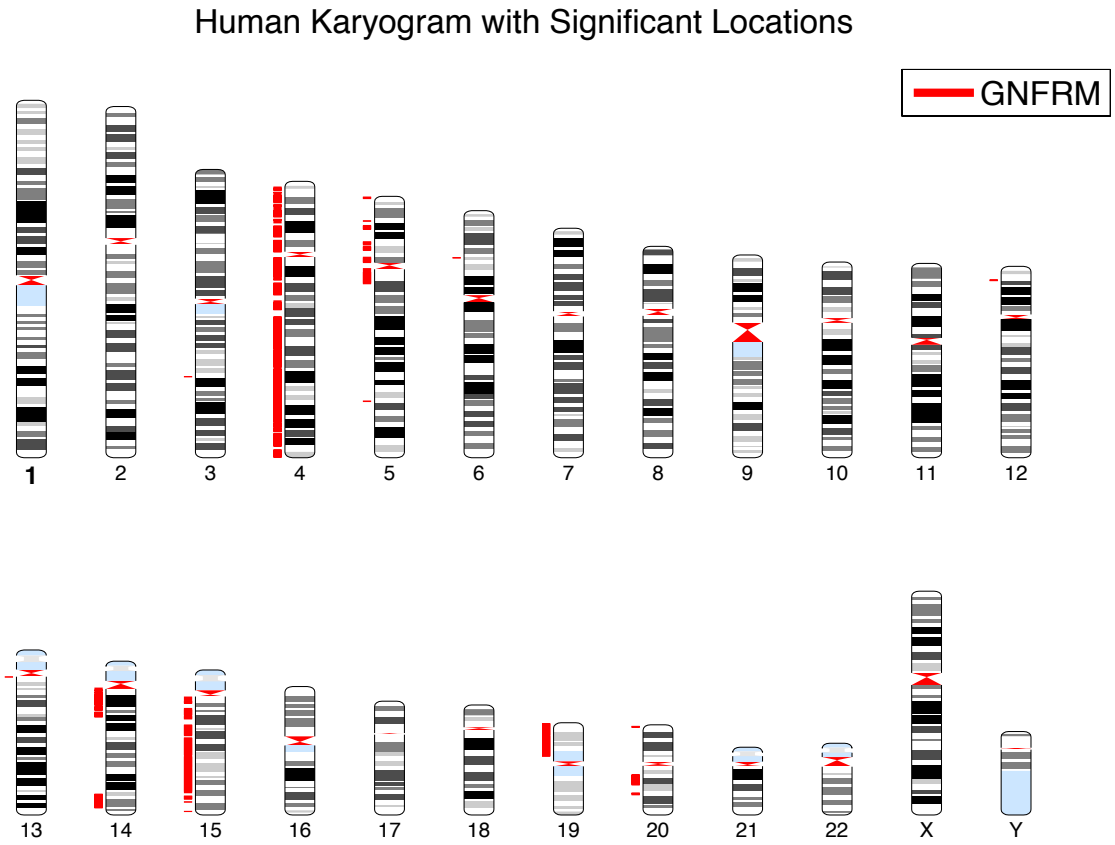
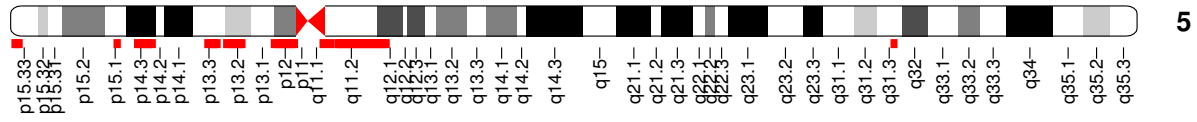
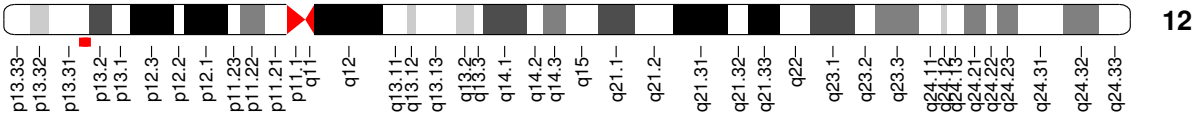


Figure 3.2: Test results for the MM application. The figure is a karyogram that depicts the test results for GNFRM and GFLM across the genome using a Benjamini-Hochberg correction for multiple tests. Red regions to the left of each chromosome were identified by GNFRM and green regions to the right were identified by GFLM.



(a) Chromosome 5



(b) Chromosome 12

Figure 3.3: Chromosome 5 ideogram of our analysis of association between Multiple Myeloma progression and quantitative copy number alterations. Significant genomic regions of copy number alteration are mapped according to their cytoband location.

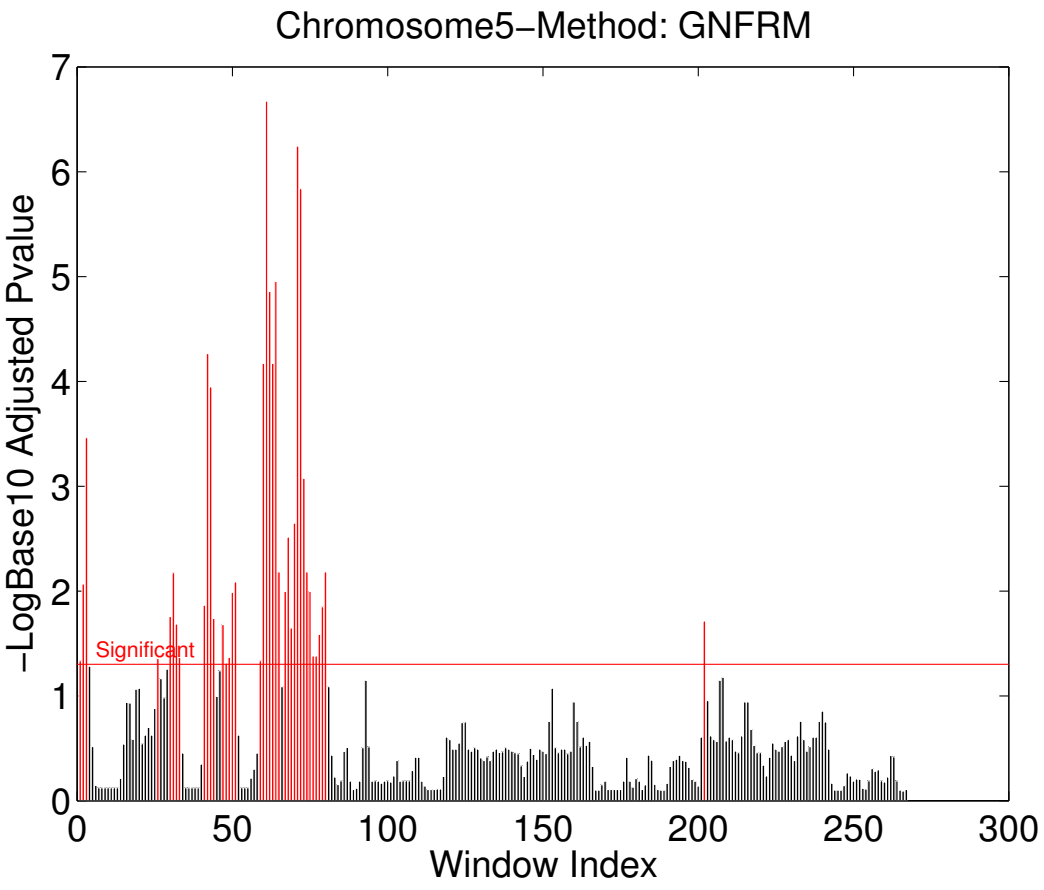


Figure 3.4: Manhattan plots of Multiple Myeloma copy number association analysis for chromosome 5. Significance levels are plotted along the moving window index.

Chapter 4

Functional Nonlinear Cox Proportional Hazards Model

4.1 Introduction

Another important goal in oncology research is modeling the time to death for patients diagnosed with cancer. This is especially true in the case of aggressive cancers such as Glioblastoma Multiforme [Krishnan et al., 2013, Johnson et al., 2012]. Predicting the survivability of patients with cancer in early stages helps physicians to tailor personalized treatments [Win et al., 2014]. Recently, there has been an interest in predicting survival based on molecular targets. As technology continues to lend itself to better understanding of the genetic etiology of complex diseases, cancer researchers seek to understand the relationship between survival in cancer patients and genetic markers [Liu et al., 2010]. In fact, this interest in understanding the genetic mechanisms that extend (or shorten) survival is demonstrated in Chapter 2 where we related the findings of our multiple myeloma copy number association analysis to the work of Avet-Loiseau et al. [2009]. Recall that the authors investigated the relationship between survival time and copy number aberration using traditional techniques. Given the growing interest in understanding the cytogenetic contributions to survival in cancer patients, we are motivated to extend the NFRM to consider censored survival outcomes.

Our motivating application arises from Glioblastoma Multiforme (GBM) data which is publicly available at the Cancer Genome Atlas (TCGA) domain `http://cancergenome.nih.gov/`. GBM is the most common form of malignant brain cancer in adults, and the prognosis is generally very poor [Jain et al., 2013]. This data consist of aCGH copy number profiles for 233 patients that have been diagnosed with GBM. In addition to the log2 copy number intensities, the data contains gene expression, miRNA expression, DNA methylation, clinical data such as

age and gender, as well as censored survival times for each patient. The copy number ratios are obtained using the Agilent 244A microarray.

Recently, researchers associated with the Cancer Genome Atlas identified several recurrent regions of copy number alterations among patients diagnosed with GBM [Network, 2008]. Some of these regions were previously detected by other researchers, and some of these regions are new locations not previously identified in the literature. In addition to detecting these regions of copy number alteration, their research suggested that about 76% of the genes in the GBM pathway exhibited a correlation between expression levels and copy number over their gene region. Verhaak et al. [2010] extended their work to establish molecular subclasses of GBM patients based on copy number alterations. They showed that the response to therapies differed by subclass. In addition, Jain et al. [2013] suggested that some of these molecular subclasses are good predictors of survival, especially when integrated with radiological imaging data.

Given these reported relationships between copy number and gene expression among GBM patients, we propose novel statistical models as alternative approaches to investigate the following scientific questions:

1. What regions of copy number alteration are associated with survival in GBM patients?

2. Can we identify significant interactions between copy number alteration and gene expression on survival in GBM patients?

Based on the arguments presented in Section 2.1, we translate both of these scientific questions into functional regression models, whereby we develop a model to relate a functional covariate to censored survival time nonlinearly. In addition, we develop a second model that considers the interaction between a functional covariate and a scalar response. In Section 2.1 and Section 3.1, we briefly discussed several approaches to functional regression when the outcome is continuous or discrete. However, to the best of our knowledge, the models developed in this chapter represent a first attempt to test for the effect of functional covariates on survival outcomes—representing a key advancement in the functional data analysis testing literature.

We propose the functional nonlinear Cox proportional hazards model (FNCPH) as a solution to these cancer genomics problems. Two versions of the FNCPH model are developed: a main effects model and an interaction model. Following the modeling approach utilized in both the NFRM and the GNFRM, we establish a connection between our complex functional models and the random effects Cox model. Testing procedures are developed under the random effects framework.

The remainder of this chapter is organized as follows. We detail the main effects version of the FNCPH model in Section 4.2.1, and we discuss its extension which includes an interaction

term in Section 4.2.2. We establish a connection with the random effects Cox model through a kernel machine representation in Section 4.2.1. We propose testing procedures in Section 4.2.3 and investigate the finite sample performance of the proposed procedures in Section 4.3. In Section 4.4, we use the procedures developed in this chapter to conduct an integrated genomic analysis of our motivating GBM dataset.

4.2 Functional Nonlinear Cox Proportional Hazards Model

4.2.1 Main Effects Model

Assume that for each subject, $i = 1, \ldots, n$, we observe the random vector $(Y_i, \Delta_i, \mathbf{z}_i, \mathbf{W}_i)$, where $Y_i = \min(T_i, C_i)$, $\Delta_i = I(T_i \leq C_i)$, T_i is the survival time, C_i is censoring time, $\mathbf{z}_i$ is a q-dimensional vector of covariates, and $\mathbf{W}_i = (W_{i1}, \ldots, W_{ip})$ is a p dimensional vector of realizations of the underlying mean-zero process with measurement error, i.e. $W_{ij} = X_i(t_j) + \delta_{ij}$ where $X_i(\cdot)$ is a square-integrable zero mean process defined on $\mathcal{T}$ and δ_{ij} is mean-zero white noise with finite variance. We also assume that T_i is independent of C_i conditional on $\mathbf{z}_i$ and $X_i(\cdot)$, and we assume that the random vectors are independent and identically distributed across subjects.

We relate the survival time T_i to $\mathbf{z}_i$ and $X_i(\cdot)$ through the following proportional hazards model [Cox, 1972]:

$$\lambda[t|\mathbf{z}_i, X_i(\cdot)] = \lambda_0(t)\exp[\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathcal{L}\{X_i(\cdot)\}]. \tag{4.1}$$

Here $\lambda[t|\mathbf{z}_i, X_i(\cdot)]$ is the conditional hazard, $\lambda_0(t)$ is the baseline hazard, $\boldsymbol{\beta}$ models the effect of the covariates and $\mathcal{L}(\cdot)$ is an unknown functional (linear or nonlinear) that maps from L^2 to $\mathbb{R}$. Furthermore, we assume that $\mathrm{E}[\mathcal{L}\{X_i(\cdot)\}] = 0$ and $\mathrm{E}[X_i(\cdot)] = 0$. Our primary interest is in testing $H_0 : \mathcal{L}\{X(\cdot)\} = 0$, which implies that conditional on $\mathbf{z}_i$, the subject specific curves are not associated with survival time, i.e. there is no local effect of copy number alterations on survival time.

Dimension Reduction

Following the dimension reduction techniques discussed in Section 2.2.1 (FPCA), we posit the following finite-dimensional model,

$$\lambda(t|\mathbf{z}_i, \boldsymbol{\xi}_i) = \lambda_0(t)\exp[\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\beta} + h(\boldsymbol{\xi}_i)], \tag{4.2}$$

where $\boldsymbol{\xi}_i = [\xi_{i1}, \ldots, \xi_{iJ}]^{\mathrm{T}}$ and $h(\cdot) : \mathbb{R}^J \mapsto \mathbb{R}$ is a smooth and centered function. In practice, we implement a smooth covariance technique to conduct the FPCA [Di et al., 2009]. The

smooth covariance approach is similar in nature to the PACE approach of Yao et al. [2005]. The covariance function $V(s,t) = \text{Cov}\{X_i(s), X_i(t)\}$ is smoothed via penalized thin plate splines as opposed to local kernel-smoothers. The scores $\xi_j = \int X(t)\phi_j(t)\ dt$ are estimated as the best linear unbiased predictors from a linear mixed model framework. This method can be implemented via the fpca.sc function in the refund R package. More technical details about this smooth covariance approach are provided in Appendix A.3. Having reduced the dimension of our problem, we next focus our attention on testing the effect of $\boldsymbol{\xi}_i$ and T_i via the hypothesis H_0: $h(\cdot) = 0$. This hypothesis is equivalent to the hypothesis expressed in Section 4.2.1.

Given the use of FPCA to reduce the dimension of the underlying trajectories, it's reasonable to consider extending developed testing procedures for the functional linear model (FLM) to censored survival times, such as the Wald test considered in Chapter 3 [Müller and Stadtmüller, 2005]. This implies that $h(\boldsymbol{\xi}_i) = \sum_{j=1}^{J} \gamma_j \xi_{ij}$, which is a result of using the same orthonormal eigenbasis to approximate $X_i(\cdot)$ and $\beta(\cdot)$. See Section 2.2.1 for more details on this reduced form. Conditional on J and $\boldsymbol{\xi}_i$, this approach yields a fully parametric model, whereby it's reasonable to consider using classical procedures and asymptotic theory to perform large sample tests such as the likelihood ratio test. We refer the reader to Klein and Moeschberger [2003] for details on such tests.

In the case of the functional quadratic regression model (FQRM) proposed by Yao and Müller [2010], we can assume that $h(\boldsymbol{\xi}_i) = \sum_{j=1}^{J} \gamma_j \xi_{ij} + \sum_{j=1}^{J} \sum_{l=1}^{j} \vartheta_{jl} \xi_{ij} \xi_{il}$. While this model reduces to a linear model, the performance of large sample testing procedures in this finite-dimensional framework have not been previously investigated in the literature, and we do not take on the task here. We mention it only as a point of future research consideration.

Unfortunately, our numerical studies suggest that likelihood ratio test based on the simple extension of the FLM does not perform well. See Section 4.3 for more details. Furthermore, we do not expect this approach to detect nonlinear dependencies well. These considerations further motivate the development of a kernel machine based working model to investigate the relationship between the random curve and survival time when the effect of the curve on survival time is expected to be nonlinear.

Kernel Machine Working Model

From Eq. (4.2) we assume that $h(\cdot) \in \mathcal{H}_K$, where the function space $\mathcal{H}_K$ is generated by a positive definite kernel function $K(\cdot,\cdot)$. Let $\{\zeta_j(\cdot), j \geq 1\}$ be an orthogonal basis of $\mathcal{H}_K$ such that $h(\boldsymbol{\xi}_i)$ has the primal representation $h(\boldsymbol{\xi}_i) = \sum_{j=1}^{\infty} a_{ij} \zeta_j(\boldsymbol{\xi}_i)$ [Mercer, 1909]. Then by the representer theorem, $h(\boldsymbol{\xi}_i)$ also has a dual representation $h(\boldsymbol{\xi}_i) = \sum_{j=1}^{n} \alpha_j K(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j)$ [Kimeldorf

and Wahba, 1971]. We use the dual representation to construct the following working model,

$$\lambda(t|\mathbf{z}_i, \boldsymbol{\xi}_i) = \lambda_0(t)\exp[\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{K}_i]. \tag{4.3}$$

Here $\mathbf{K}_i = [K_{i1}, \ldots, K_{in}]^{\mathrm{T}}$ and $K_{ij} = K(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j)$.

There are several common kernel functions used in the kernel machine regression literature [Suykens and Vandewalle, 1999, Schölkopf and Smola, 2002, Liu et al., 2007]. As in Chapter 2, we focus our attention on (1) the dth polynomial kernel: $K(\boldsymbol{\xi}_l, \boldsymbol{\xi}_k) = (\boldsymbol{\xi}_l^{\mathrm{T}}\boldsymbol{\xi}_k + 1)^d$, where d determines the degree of the polynomial and (2) the Gaussian kernel: $K(\boldsymbol{\xi}_l, \boldsymbol{\xi}_k) = \exp(\sum_{j=1}^{J} \xi_{lj} - \xi_{kj})^2/\rho$, where ρ is an unknown tuning parameter in the context of the kernel function.

The roles of the kernel function that are stated in Section 2.2.1 are especially important in the context of investigating survival times. The work of Verhaak et al. [2010] and Jain et al. [2013] demonstrates that molecular subclasses determined by copy number alterations over specific gene regions are good predictors of survival. In some cases, these subclasses are determined by considering the frequency of patients exhibiting copy number variations, whereas in our model, the kernel function borrows information across subjects to quantify the effect of recurrent regions of copy number change. Thus, in addition to being able to capture the frequency of recurrent regions of copy number alteration, our approach also directly considers the effect of such a region on the outcome.

Using the working model (4.3), we construct the following penalized partial likelihood,

$$\ell(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{i=1}^{n} \sum_{\{\mathrm{u}\}} dN_i(u) \left[(\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{K}_i) - \log\left(\sum_{l=1}^{n} \exp[\mathbf{z}_l^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{K}_l] R_l(u) \right) \right] - \frac{\varphi}{2}\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{K}\boldsymbol{\alpha}, \tag{4.4}$$

where φ is a penalty term, $\{u\}$ denotes all grid points over time, $dN_i(u) = I(Y_i \in [u, u + \Delta u), \Delta_i = 1)$ which is the indicator for the ith subject being observed to have the event in the interval $[u, u + \Delta u)$, $R_l(u) = I(Y_l \geq u)$ representing whether or not the ith subject is at risk at time u, and $\mathbf{K}$ is the $n \times n$ kernel matrix where the (l, k)th element is $K(\boldsymbol{\xi}_l, \boldsymbol{\xi}_k)$. Similarly, we can view $dN_i(u)$ to be the change in the counting process $N_i(u) = I(Y_i \leq u)\Delta_i$ over a short interval $[u, u + \Delta u)$. This view equates (4.4) to the penalized partial likelihood function presented in Cai et al. [2011], which allows us to extend their kernel machine score test for censored survival outcomes to the functional setting. This test is discussed in more detail in Section 4.2.3.

4.2.2 Interaction Model

In many applications such as ours, there is an interest in modeling the interaction between a single covariate and the random trajectory. For example, consider our motivating scientific problem. Let $X_i(\cdot)$ represent the random process that generates the observed copy number observations over some gene region and let z_i represent the level of expression of the gene. In this scenario, it may be of interest to determine if the interaction between gene expression and the copy number *process* is associated with survival time. To investigate such scientific questions, Eq. (4.1) is extended to yield the following interaction model,

$$\lambda[t|z_i, X_i(\cdot)] = \lambda_0(t)\exp[z_i\beta + \mathcal{L}_1\{X_i(\cdot)\} + \mathcal{L}_2\{X_i(\cdot), z_i\}], \tag{4.5}$$

where the assumptions on $\mathcal{L}_1$ follow directly from the main effects model. To capture linear or nonlinear interaction effects, we assume only that $\mathcal{L}_2$ is a an unknown functional that maps from $(L^2 \times \mathbb{R})$ to $\mathbb{R}$. In this model, the primary interest is in testing $H_0 : \mathcal{L}_2\{X_i(\cdot), z_i\} = 0$, which implies that there is no interaction between the random curve and the scalar covariate. For ease of exposition, we present only the covariate that we are interested in testing for interaction with $X_i(\cdot)$. However, the model can be easily extended to include additional covariates that do not interact with z_i or $X_i(\cdot)$.

Following the dimension reduction ideas presented in section Section 2.2.1, we approximate Eq. (4.5) with the following truncated model,

$$\lambda(t|z_i, \boldsymbol{\xi}_i) = \lambda_0(t)\exp[z_i\beta + h_1(\boldsymbol{\xi}_i) + h_2(\boldsymbol{\xi}_i, z_i)]. \tag{4.6}$$

To balance the tradeoff between constructing a parsimonious model while allowing for flexibility in modeling the interaction effects, we posit a parametric quadratic form for the main effect of the random curve, and we model the interaction via a kernel-based varying coefficient approach. Specifically, we set $h_1(\boldsymbol{\xi}_i) = \sum_{j=1}^{J} \gamma_j \xi_{ij} + \sum_{j=1}^{J} \sum_{l=1}^{j} \vartheta_{jl} \xi_{ij} \xi_{il}$ [Yao and Müller, 2010]. Recall that in Chapter 2 we showed that such a parametric quadratic model performed well with respect to estimating linear functional effects as well as a wide range of complex nonlinear effects on continuous outcomes. Since our primary focus in on testing the interaction, the use of the parametric form provides great flexibility while maintaining model simplicity.

The interaction term is modeled via a kernel-based varying-coefficient like approach [Hastie and Tibshirani, 1993]. Specifically, we model the interaction term as $h_2(\boldsymbol{\xi}_i, z_i) = z_i g(\boldsymbol{\xi}_i)$, where $g(\cdot) \in \mathcal{H}_K$ is a smooth function and $\mathcal{H}_K$ is generated by a positive definite kernel $K(\cdot, \cdot)$. This approach lends itself to several useful interpretations. In the case that z_i is binary, such as smoking status, $z_i = 1$ represents the additive effect of the copy number profile on survival

time due to being in the smoking subgroup. In general, the function $g(\cdot)$ models the additional effect of z_i on survival time that results from some unknown relationship between z_i and $\boldsymbol{\xi}_i$. In terms of notation, we can see that $z_i\beta + z_i g(\boldsymbol{\xi}_i) = z_i[\beta + g(\boldsymbol{\xi}_i)]$. Thus, if $g(\boldsymbol{\xi}_i)$ is a constant, then the copy number adjusted relationship between gene expression and time to event is still linear although the direction and the magnitude of the effect may differ. However, if $g(\boldsymbol{\xi}_i)$ is an unspecified function, then the copy number adjusted relationship between gene expression and time to event is nonlinear in nature.

As in section 4.2.1, $g(\boldsymbol{\xi}_i) = \sum_{j=1}^{n} \alpha_j K(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j)$ is the dual representation of this functional effect. We use this representation to posit a working model,

$$\lambda(t|z_i, \boldsymbol{\xi}_i) = \lambda_0(t)\exp\Big[z_i\beta + \sum_{j=1}^{J}\gamma_j\xi_{ij} + \sum_{j=1}^{J}\sum_{l=1}^{j}\vartheta_{jl}\xi_{ij}\xi_{il} + z_i\sum_{j=1}^{n}\alpha_j K_{ij}\Big]. \tag{4.7}$$

Define $\varsigma_i = z_i\beta + \sum_{j=1}^{J}\gamma_j\xi_{ij} + \sum_{j=1}^{J}\sum_{l=1}^{j}\vartheta_{jl}\xi_{ij}\xi_{il} + z_i\sum_{j=1}^{n}\alpha_j K_{ij}$. Then $(\boldsymbol{\alpha}, \beta, \boldsymbol{\gamma}, \boldsymbol{\vartheta})$ can be estimated via the following penalized likelihood,

$$\ell(\boldsymbol{\alpha}, \beta, \boldsymbol{\gamma}, \boldsymbol{\vartheta}) = \sum_{i=1}^{n}\sum_{\{\mathrm{u}\}} dN_i(u)\Big[\varsigma_i - \log\Big(\sum_{l=1}^{n}\exp(\varsigma_l)R_l(u)\Big)\Big] - \frac{\varphi}{2}\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{K}^{\star}\boldsymbol{\alpha}, \tag{4.8}$$

where $\mathbf{Z} = [z_1, \ldots, z_n]^{\mathrm{T}}$ and $\mathbf{K}^{\star} = \mathbf{K} * \mathbf{Z}\mathbf{Z}^{\mathrm{T}}$. Note that $\mathbf{Z}\mathbf{Z}^{\mathrm{T}}$ is the centered linear kernel. Thus, $\mathbf{K}^{\star}$ is the element-wise product of two kernels, which is in itself a kernel [Schölkopf and Smola, 2002]. We refer to this kernel as the *interaction kernel*.

Note that Eq. (4.8) is similar to Eq. (4.4) with the exception of the additional model terms and the structure of the interaction kernel. Thus, we are also able to extend the kernel machine score test to test for an interaction between a scalar covariate and a functional covariate on censored survival time.

4.2.3 Variance Component Score Tests

Recall that the primary interest in our main effects model (4.2) is in determining if the random process is necessary to model survival time, i.e. $H_0 : h(\cdot) = 0$. From the working model (4.3), this is equivalent to the hypothesis $H_0 : \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{K} = 0$. If we are further willing to assume that $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)^{\mathrm{T}} \sim \mathrm{N}(\mathbf{0}, \tau\mathbf{K}^{-})$, where $\mathbf{K}^{-}$ is the generalized inverse of $\mathbf{K}$, then the null hypothesis of interest is equivalent to the following hypothesis about the variance component, $H_0 : \tau = 0$.

Under this framework, we adopt the approach of Cai et al. [2011] and Lin et al. [2011] and

propose the following test statistic for the variance component,

$$Q = \widehat{\mathbf{M}}^{\mathrm{T}}\mathbf{K}\widehat{\mathbf{M}} - \widehat{q}, \tag{4.9}$$

where $\widehat{\mathbf{M}} = (\widehat{M}_1, \ldots, \widehat{M}_n)^{\mathrm{T}}$, $\widehat{M}_i = \Delta_i - \int_0^\infty R_i(t)\exp(\mathbf{z}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}})\, d\widehat{\Lambda}_0(t)$ is the estimated martingale residual under the null hypothesis, and $\widehat{\beta}$ is the maximum partial likelihood estimate of β under the null model $\lambda(t|\mathbf{z}_i) = \lambda_0(t)\mathrm{e}^{\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\beta}}$. Furthermore, $\widehat{\Lambda}_0(t) = \sum_{i=1}^n \frac{N_i(t)}{\sum_{i=1}^n R_i(t)\mathrm{e}^{\mathbf{z}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}}}$ is Breslow's estimator of the baseline cumulative hazard under the null model and

$$\widehat{q} = \sum_{i=1}^n \int K(\boldsymbol{\xi}_i, \boldsymbol{\xi}_i) R_i(t)\mathrm{e}^{\mathbf{z}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}}\, d\widehat{\Lambda}_0(t) - \sum_{i=1}^n \sum_{j=1}^n \int \frac{R_i(t)R_j(t)\mathrm{e}^{\mathbf{z}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}}\mathrm{e}^{\mathbf{z}_j^{\mathrm{T}}\widehat{\boldsymbol{\beta}}}K(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j)}{\sum_{i=1}^n R_i(t)\mathrm{e}^{\mathbf{z}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}}}\, d\widehat{\Lambda}_0(t).$$

We construct a similar statistic to test for the effect of the interaction term. Recall that we model the interaction term via a kernel-based varying-coefficient like approach. Based on the working interaction model (4.7), the null hypothesis of interest expressed in Section 4.2.2 can be represented as $H_0 : \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{K}^\star = 0$. Using the same argument as above, an equivalent hypothesis is $H_0 : \tau = 0$. Therefore, Q can be adapted to test for an interaction by replacing $\mathbf{K}$ with $\mathbf{K}^\star$ in the expression above under the null model $\lambda(t|z_i, \boldsymbol{\xi}_i) = \lambda_0(t)\exp(z_i\beta + \sum_{j=1}^J \gamma_j\xi_{ij} + \sum_{j=1}^J \sum_{l=1}^j \vartheta_{jl}\xi_{ij}\xi_{il})$.

Cai et al. [2011] expands Q as a double integrated martingale process. The limiting distribution of the martingale process under the null is a mixture of chi-squares, and thus can be approximated via a scaled chi-square distribution $k\chi_v^2$, where k and v are estimated via the Satterthwaite approach. Specifically, B realizations of the martingale process are generated, and the first two moments of the process are estimated numerically. The final p-value is based on $\hat{k}\chi_{\hat{v}}^2$. We note that the scaled chi-square approximation has been shown to perform well in the non-functional case [Liu et al., 2007] and also in the functional case as highlighted in Chapter 2 and Chapter 3. We investigate the performance of this score test numerically and apply it to our Glioblastoma Multiforme dataset for illustration.

4.3 Simulation Study

This section details the design of our simulation study, which investigates the finite sample performance of the FNCPH model (main effects version and interaction effects extension) via simulation studies. We compare the empirical size of both approaches to several nominal type I error rates under various settings. We also examine the behavior of the empirical rejection probability as effects sizes increase from the null.

Our goal is to mimic our motivating cancer genomics problem where the copy number profile

appears to be periodic in nature. Thus for the main effects model (4.1), we generated data from the following model,

$$\log(T_i) = z_{i1}\beta_1 + z_{i2}\beta_2 + h[X_i(\cdot)] + \epsilon_i, \tag{4.10}$$

where $z_{i1} \sim \mathrm{N}(0,1)$, $z_{i2} \sim \mathrm{Bin}(1, 0.5)$, $\beta_1 = 0.4$, and $\beta_2 = -0.4$. We generate ϵ_i from a standard extreme value distribution. The random curves, $X_i(\cdot)$, are generated from a orthonormal Fourier basis with five basis functions, where the coefficients for each basis function are independent draws from an N(0,0.5) distribution. We generate realizations of the underlying process as $W_i(t_j) = X_i(t_j) + \delta_{ij}$, where $\delta_{ij} \sim N(0, 0.16)$.

We consider two types of functional effects: linear and quadratic. For the linear functional effects, we model $h[X_i(\cdot)] = \int X_i(t)\gamma(t)\ dt$. For the quadratic effects, we simply square the integral. The coefficient regression function, $\gamma(\cdot)$, is generated from the same Fourier basis with the exception that the coefficients are determined by the imposed effect level (detailed in the corresponding tables and figures). Censoring was generated independent of T_i from a exponential distribution.

For the interaction model (4.5), we generated data from the following model,

$$\log(T_i) = z_{i1}\beta_1 + z_{i2}\beta_2 + h_1[X_i(\cdot)] + z_{i1}g[X_i(\cdot)] + \epsilon_i,$$

where $g[X_i(\cdot)]$ is generated in the same manner as $h(\cdot)$ in model (4.10). For simplicity, we generate $h_1[X_i(\cdot)] = \int X_i(t)\gamma_1(t)\ dt$, where the function $\gamma_1(\cdot)$ is generated as before with the exception that each coefficient is set to a common nonzero constant.

We consider two sample sizes, $n = 200$ and 300, and we set the number of realizations of the underlying process to $m = 100$. In addition, we evaluate the performance of the score test in each setting using both the linear and quadratic kernel functions. For each setting, the mean of the exponential distribution is chosen to obtain censoring proportions of $\sim 25\%$ and $\sim 50\%$. We generate 50,000 datasets to estimate the empirical size of the test, and 1000 datasets are generated to estimate the empirical power. Recall that we approximate the null distribution of our statistic Q as a scaled chi-square distribution $k\chi^2_v$, where the parameters k and v are estimated numerically. We set $B = 1000$ to obtain $\hat{k}$ and $\hat{v}$.

We compare our proposed approach to the naive extension discussed in Section 4.2.1. The FLM reduces to standard a Cox model when conditioned on J and $\{\boldsymbol{\xi}_i\}_{i=1}^n$. Thus, we construct the following likelihood ratio test for the approach:

$$T = 2[\ell(\hat{\boldsymbol{\theta}}) - \ell(\hat{\boldsymbol{\theta}}_0)],$$

where $\ell(\hat{\boldsymbol{\theta}})$ is the maximized value of log partial likelihood of the alternative model and $\ell(\hat{\boldsymbol{\theta}}_0)$

is the maximized value of the log partial likelihood of the null model [Klein and Moeschberger, 2003].

Table 4.1 provides the results of our type I error rate investigations for the main effects model when $n = 200$. The score test with a quadratic kernel (Q_{quad}) performs well with respect to maintaining all sizes explored, albeit a bit conservative. Similarly, the size of the score test when using the linear kernel (Q_{lin}) is close to the nominal levels in all cases. In contrast, the naive approach discussed in Section 4.2.1 performs poorly as demonstrated by the inflated sizes of the test. Results for $n = 300$ are similar. They can be viewed in Table D.1 in Appendix D.

Table 4.1: Empirical size of the tests for main effects at $n = 200$ at multiple type I error rates when censoring is moderate (25%) and high (50%). Testing was performed using (1) FNCPH with a linear kernel (Q_{lin}), (2) FNCPH with a quadratic kernel, (Q_{quad}) (3) naive linear approach (LRT_L), and (4) the naive quadratic approach (LRT_Q). The results are based on $50,000$ generated datasets.

Censoring	Type I Error Rate	Q_{lin}	Q_{quad}	LRT_L
25%	0.05	0.0569	0.0373	0.0787
	0.01	0.0113	0.0049	0.0198
	0.005	0.0057	0.0020	0.0108
	0.001	0.0011	0.0003	0.0024
50%	0.05	0.0547	0.0466	0.0726
	0.01	0.0109	0.0073	0.0182
	0.005	0.0054	0.0037	0.0097
	0.001	0.0012	0.0004	0.0025

Table 4.3 provides the empirical sizes for the testing procedures developed for the interaction model when $n = 200$. In general, the testing procedures associated with the FNCPH model maintain the size of the test, but they are conservative. The naive approach again performs very poorly. In fairness, we note that the literature does not contain any methods that extend the FLM to consider this type of interaction when the outcome is continuous; thus, there is no point of reference to compare this result in other data situations. For the FNCPH testing procedures, the tests became more conservative as censoring increased.

Given the inflated empirical sizes of the naive approach for both the main effects model and the interaction model, we limit our power investigations to only the FNCPH model with both a linear kernel and a quadratic kernel, Q_{lin} and Q_{quad}, respectively. Figure 4.1 and Figure 4.2 display the results of our power analysis for the main effects model and the interaction model,

Table 4.2: Empirical size of the tests for interaction effects at $n = 200$ at multiple type I error rates when censoring is moderate (25%) and high (50%). Testing was performed using (1) FNCPH with a linear kernel (Q_{lin}), (2) FNCPH with a quadratic kernel, (Q_{quad}) (3) naive linear approach (LRT_L), and (4) the naive quadratic approach (LRT_Q). The results are based on $50,000$ generated datasets.

Censoring	Type I Error Rate	Q_{lin}	Q_{quad}	LRT_L
25%	0.05	0.0379	0.0292	0.1289
	0.01	0.0088	0.0057	0.0387
	0.005	0.0052	0.0027	0.0227
	0.001	0.0011	0.0006	0.0072
50%	0.05	0.0288	0.0198	0.1192
	0.01	0.0060	0.0034	0.0355
	0.005	0.0029	0.0015	0.0207
	0.001	0.0006	0.0002	0.0058

respectively, when $n = 200$. For the main effects model, (4.1), we note that these results are similar to those observed for the NFRM in Chapter 2 and the GNFRM in Chapter 3 for both moderate and high levels of censoring. Thus, both score tests maintain power to detect a signal even in the presence of a significant level of censoring (50%). We also note that Q_{quad} performs well with respect to detecting simple linear dependencies of reasonable effect sizes, while Q_{lin} is not capable of detecting nonlinear effects.

Table 4.3 quantifies the reduction in power that results from using a quadratic kernel in the testing procedure when the true main effect is linear. We can see that when the signal in the data is small, the misspecification can result in a power loss as large as about 40% across all experimental settings. However, we note that this loss decreases as we increase our sample size. Therefore, we expect the Q_{quad} to perform reasonably well for large sample sizes.

For the interaction model, Figure 4.2 suggests the FNCPH model performs well with respect to detecting interaction effects. Recall that the interaction model was generated under a varying coefficient assumption, i.e. $\mathcal{L}_2\{z_i, X_i(\cdot)\} = z_{i1}g[X_i(\cdot)]$. When modeling $g[X_i(\cdot)]$ with the simple linear form, panels (a) and (b) are similar to those of Figure 4.1. This suggests that in the presence of simple linear dependencies in the interaction term, censoring has little effect on the performance of the score test with either the linear kernel or the quadratic kernel. In contrast, when the interaction term has a nonlinear form, censoring has a substantial impact on the performance of the score test. At 25% censoring, the power achieved by Q with a quadratic kernel is 92%, whereas at 50% censoring, the power achieved by Q with a quadratic kernel is

Table 4.3: Percent power loss that results from using the FNCPH model with a quadratic kernel when the true main effect is linear. These results for 25% censoring correspond to panel (a) of Figure 4.1 and panel (a) of Figure D.1. These results for 50% censoring correspond to panel (c) of Figure 4.1 and panel (c) of Figure D.1.

		Effect Levels						
Censoring	n	0.09	0.17	0.25	0.34	0.43	0.51	0.60
25%	200	0.406	0.297	0.175	0.050	0.012	0.000	0.000
	300	0.303	0.207	0.068	0.013	0.000	0.000	0.000
50%	200	0.303	0.224	0.143	0.045	0.005	0.000	0.000
	300	0.208	0.174	0.054	0.007	0.000	0.000	0.000

71%.

Table 4.4 quantifies the reduction in power that results from using a quadratic kernel in the testing procedure when the true interaction effect is linear. The results yield the same conclusion as the main effects model. We conclude that for problems such as our cancer genomics problem, using Q_{quad} and Q_{lin} as companions may be a more comprehensive approach.

Table 4.4: Percent power loss that results from using the FNCPH model with a quadratic kernel when the true interaction effect is linear. These results for 25% censoring correspond to panel (a) of Figure 4.2 and panel (a) of Figure D.2. These results for 50% censoring correspond to panel (c) of Figure 4.2 and panel (c) of Figure D.2.

		Effect Levels						
Censoring	n	0.14	0.29	0.43	0.57	0.71	0.86	1.00
25%	200	0.335	0.213	0.050	0.010	0.000	0.000	0.000
	300	0.271	0.097	0.003	0.001	0.000	0.000	0.000
50%	200	0.377	0.228	0.106	0.031	0.002	0.000	0.000
	300	0.304	0.121	0.013	0.002	0.000	0.000	0.000

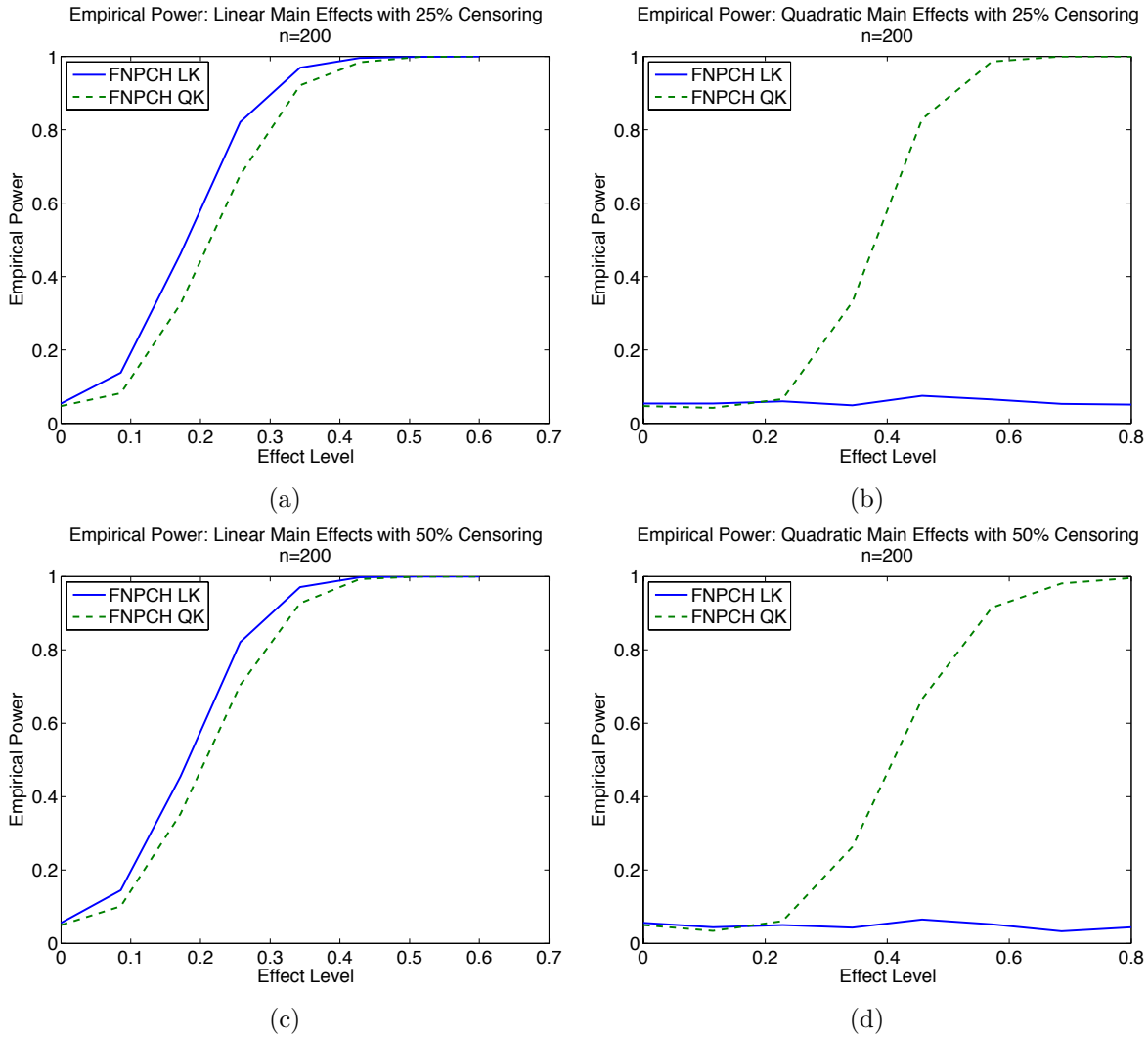


Figure 4.1: Empirical power for the model $\lambda[t|\mathbf{z}_i, X_i(\cdot)] = \lambda_0(t)\exp[\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathcal{L}\{X_i(\cdot)\}]$ at $n = 200$. In panels (a) and (c), $\mathcal{L}\{X_i(\cdot)\} = \int_{\mathcal{T}} X_i(t)\beta(t)\ dt$. In panels (b) and (d), $\mathcal{L}\{X_i(\cdot)\} = [\int_{\mathcal{T}} X_i(t)\beta(t)\ dt]^2$. All panels display the results for (1) the FNCPH model constructed with a linear kernel and (2) the FNCPH model constructed with a quadratic kernel and are based on $1,000$ generated data sets.

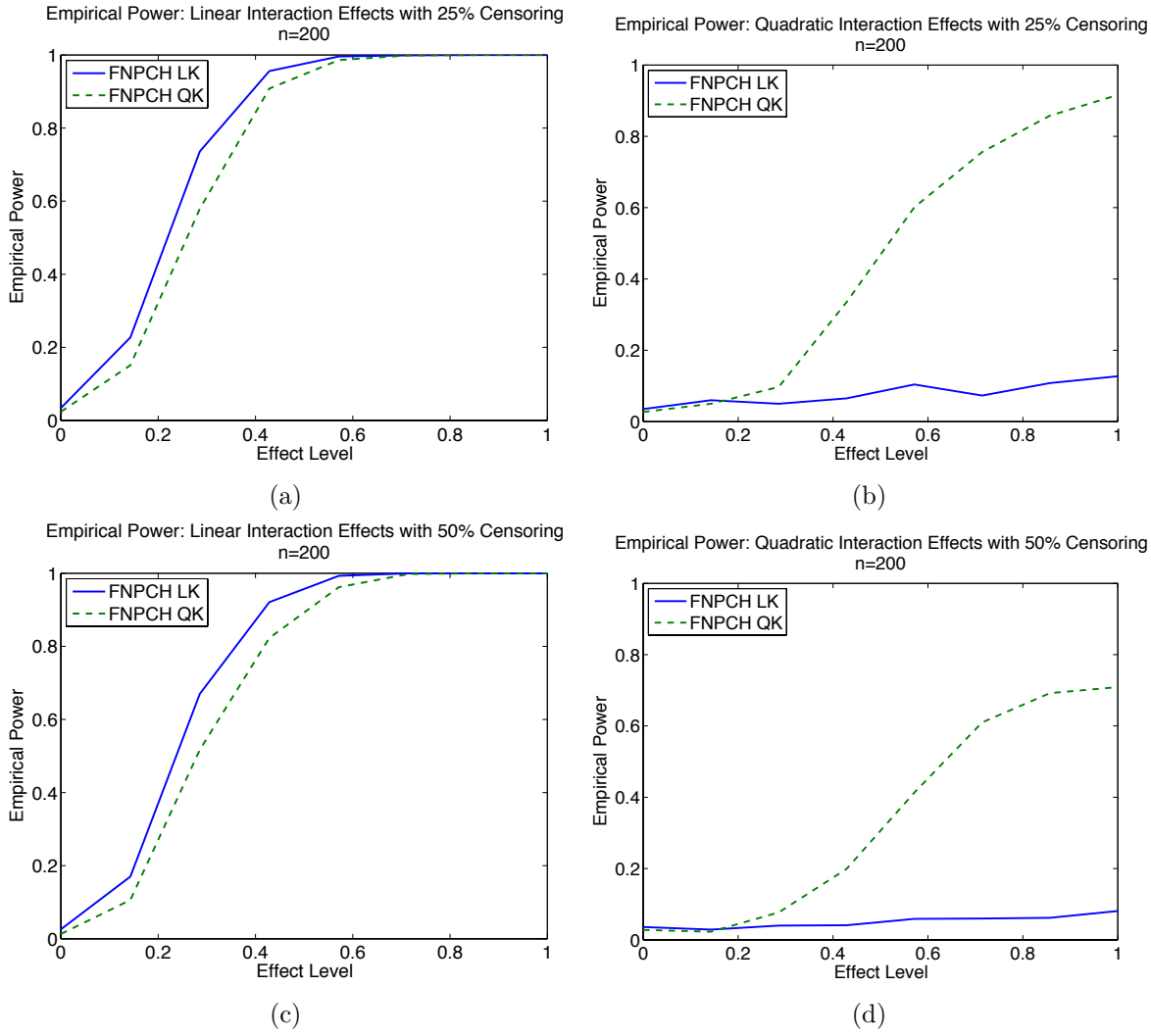


Figure 4.2: Empirical power for the model $\lambda[t|z_i, X_i(\cdot)] = \lambda_0(t)\exp[z_i\beta + \mathcal{L}_1\{X_i(\cdot)\} + \mathcal{L}_2\{X_i(\cdot), z_i\}]$ at $n = 200$. In panels (a) and (c), $\mathcal{L}_2\{X_i(\cdot)\} = z_i \int_{\mathcal{T}} X_i(t)\beta(t)\ dt$. In panels (b) and (d), $\mathcal{L}_2\{X_i(\cdot)\} = z_i[\int_{\mathcal{T}} X_i(t)\beta(t)\ dt]^2$. All panels display the results for (1) the FNCPH model constructed with a linear kernel and (2) the FNCPH model constructed with a quadratic kernel and are based on $1,000$ generated data sets.

4.4 Integrated Analysis of Glioblastoma Multiforme

We use both FNCPH models to investigate the association between copy number alterations over local gene regions and survival time for patients diagnosed with GBM. The data consists of gene expression levels, methylation levels, DNA copy numbers, and demographic data such as age and gender. In addition, the data contains information on survival times, time to tumor progression. The gene expression data was measured for 12,042 genes using Affymetrix Human Genome U133A arrays. The copy number data was measured using Agilent 244A microarrays which yields approximately 235,000 copy number intensities along the genome. We focus on the copy number intensities covering the 52 genes within the GBM pathway. Only 47 genes in the GBM pathway have both copy number data and gene expression data.

We apply to proposed testing procedures to investigate the effect of copy number and gene expression interaction on survival time and to investigate the main effect of copy number alterations on survival time. There is about 30% censoring in the data. In addition, several of the genes within the GBM pathway have only a few probes spanning their genomic regions. This makes it difficult to employ our functional approach on the probe values for these loci. Thus for illustration purposes, we narrow our focus to the eight genes in the GBM pathway that have observed expression levels and at least 20 copy number probes covering their genomic regions.

In many studies that consider gene expression, genes are prioritized by considering the variability of the expression data across patients, such as Chen et al. [2014]. A common measure of variability used in this context is the median absolute deviation (MAD). Chen et al. [2014] considers only genes with MAD > 0.5. In Table 4.5, we provide the name of each of the eight genes investigated as well as the number of copy number probes spanning each locus, the mean expression level, and the median absolute deviation. Note that several of the genes exhibit low variation in their expression levels. This is an interesting result when considering the level of variation in the copy number profiles covering these gene regions as displayed in Figure 4.3. Although there is a substantial increase in the variation of the copy number intensities between the tumor samples for each patient and their corresponding normal tissue samples, the MAD of the gene expression level suggests that this phenomena does not necessarily lead to increased variability of expression across patients. For example, the EGFR gene has the largest variation in copy number among the eight genes examined; however, it has nearly the smallest amount of variation in expression level. In contrast, the BRAF gene has expression levels above the 0.5 MAD threshold used by Chen et al. [2014]; however, the variation in the copy number intensities over its genomic region is relatively low.

Although this summary suggests that there may not be a relationship between copy number

Table 4.5: Summary of the expression levels for eight genes within the GBM pathway. Expression levels are measured using Affymetrix Human Genome U133A arrays.

CN = copy number; MAD = median absolute deviation

		Gene Expression	
Gene	Number of CN Probes	Mean	MAD
AKT3	37	4.2	0.51
BRAF	21	4.7	0.59
CDK6	27	4.1	0.19
EGFR	21	7.5	0.13
IGF1R	39	6.4	0.34
NF1	34	5.4	0.36
PIK3C2G	43	4.1	0.11
RB1	24	6.6	0.17

alterations and gene expression measured on Affymetrix Human Genome U133A arrays, we note that Network [2008] illustrated a connection between copy number variation (measured via the Affymetrix SNP 6.0 array) over the EGFR locus and relative exon expression levels across known EGFR exons (measured via the Affymetrix Exon array) for three patients. Thus, we are motivated to formally investigate the interaction between copy number alterations and gene expression as measured by the array technologies that produced our genomics data.

We applied the combined testing procedure developed for the interaction model (4.5) to test for the interaction between gene expression and copy number alteration on survival time. We limit our investigations to the EGFR gene and the NF1 gene, because these genes were reported to be related to the biology of GBM by the Cancer Genome Atlas. We adjusted for age and gender in our model, and we corrected for multiple comparisons using a simple Bonferroni correction. The score test did not detect a significant interaction for the NF1 gene or the EGFR gene. However, we note that the unadjusted p-value for the NF1 gene was 0.052. So while we failed to reach statistical significance, we believe that the borderline result suggests that there is indeed a biological phenomenon occuring. As mentioned in our numerical power experiments, it's possible that the level of censoring in this data, (30%), has weakened the signal.

Next, we investigated the main effects of copy number alterations on survival in GBM patients using the testing procedure proposed in Section 4.2.3. Although the main effects model is not directly nested within the interaction model, the use of the parametric quadratic form in the interaction model relates to the quadratic kernel used in the main effects model. Thus,

in terms of testing, the main effects model is equivalent to the interaction model without the interaction term. In the first model, we included age, gender and gene expression as covariates. In another model, we included only the copy number profile across the gene region. From the first model, the unadjusted p-value for the NF1 gene was 0.04. When testing within the simple model, the unadjusted p-value for the EGFR and NF1 genes were 0.01 and 0.05 respectively. Clearly, gene expression, age and gender have a confounding effect with the copy number profile over the EGFR genomic region. However, we again assert that the lack of statistical significance in these analyses may be attributed to censoring, thus the level of signal that we detected in our unadjusted p-values may suggest that there are some interesting biological mechanisms relating these genes to survival time in GBM patients. It may be advantageous to extend the models in this chapter to consider multiple regions of copy number alteration simultaneously to reduce the impact of multiple testing corrections on power.

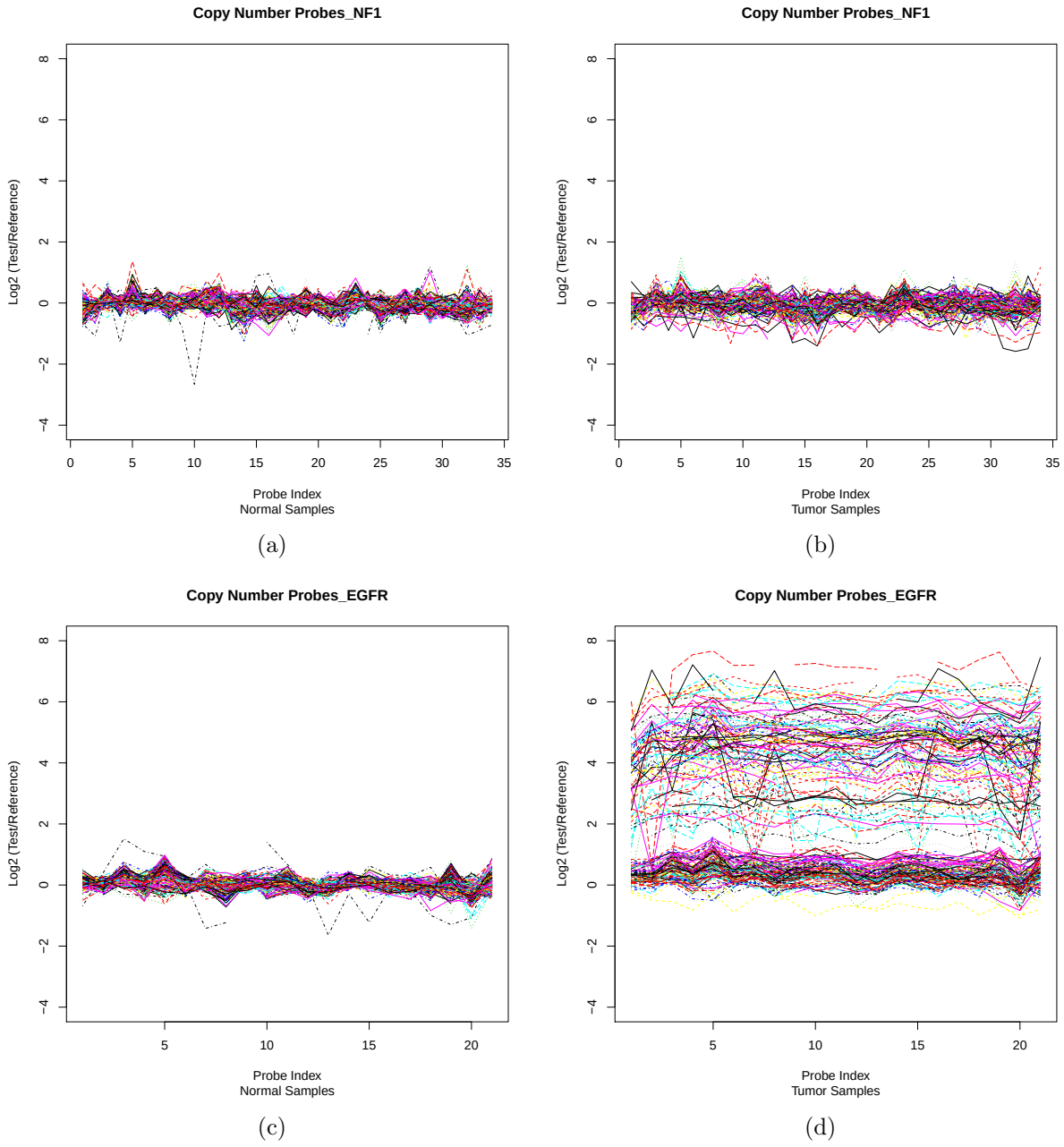


Figure 4.3: Copy number intensities over two genes in the GBM pathway (BRAF and EGFR). The left panels display the copy number intensities measured in each patient's normal tissue sample. The right panels display the copy number intensities measured in each patient's diseased tissue sample.

Chapter 5

Conclusion

The statistical methods presented in this work model nonlinear relationships between a single random curve and various types of scalar outcomes. A core component of each model is the extension of kernel machine ideas into the functional regression literature; whereby, the kernel machine framework serves as a bridge between our complex functional models and their corresponding mixed model representations. The advantage of using the kernel machine framework is that it provides a flexible approach to modeling nonlinear relationships. Our use of this framework also allows us to model the similarity in curves for every pair of subjects which is very useful in our motivating application and similar problems. Another key advantage of reducing our complex functional models to simple mixed model representations is that our proposed testing procedures are easy to implement in most available statistical software packages such as R and Matlab®.

In Chapter 2, we develop the nonlinear functional regression model (NFRM) as an approach to test and estimate the effect of a nonlinear functional covariate on a continuous response. Our approach adjusts for the effects of scalar covariates, such as age and gender. We test for a nonlinear functional effect using a variance component score test, and we estimate the size and direction of the effect under the mixed model representation. Our approach is able to capture a wide range of complex nonlinear relationships. Thus, it is attractive for many functional regression applications when the relationship between the random trajectory and the continuous response is believed to be complex and nonlinear. When used in conjunction with the functional linear model, the combined testing procedure provides a comprehensive approach to determining whether the functional covariate is necessary in the regression model.

We demonstrate the NFRM using a quadratic kernel to model the covariance structure in the linear mixed model representation. Empirically, we show that the quadratic kernel is somewhat robust to misspecification of the true covariance structure that expresses the functional

relationship. Many common covariance functions used in the mixed model and kernel machine literature are functions of unknown tuning parameters. Although these tuning parameters lend themselves to more flexible models, estimating the parameters adds to the complexity of the problem and greatly increases the time needed to estimate the copy number effect. The robustness of the quadratic kernel reduces the tradeoff between using a simple covariance structure and a more complicated covariance structure. It is of further interest to study the performance of the NFRM using more complicated covariance functions and numerical optimization schemes. In addition, further research is needed to develop an approach to identify optimal covariance structures.

In Chapter 3, we develop the generalized nonlinear functional regression model (GNFRM) as an extension of the NFRM to non-normal responses. However, the focus of this work is only on testing. Similar to the NFRM, we propose a variance component score test to test for an effect of the functional covariate. In Chapter 2, a companion testing procedure was suggested to detect both linear and nonlinear relationships for small to moderate samples sizes. The multiple comparison induced by this approach may be viewed as a limitation in some situations. As a solution, we develop an adaptive composite kernel that is nonparametric. This feature is especially useful because it allows the data to speak for themselves. The resulting testing procedure is able to detect linear relationships as well as nonlinear relationships with a minimal impact on power.

We demonstrate the performance of the method using a binary outcome, but the model can be easily adapted for any outcome whose distribution is in the exponential family. Our simulation results show that the GNFRM constructed with the adaptive composite kernel performs well for binary outcomes. In particular, we note that the model performs as well as or close to the optimal kernel across a wide range of complex relationships. The reported benefits of the adaptive composite kernel are also applicable to the standard kernel machine regression framework, which further extends the utility of the framework in many settings. While the use of this kernel enhances our testing procedures, estimation of the functional effect is not straightforward due to the unspecified tuning parameter. One can employ cross validation to accomplish estimation in the GNFRM; however, cross validation is known to be time consuming which is a limitation in applications such as ours where the interest is in conducting a genome wide scan. Thus, it is of further research interest to develop estimation procedures that are efficient with respect to time.

In Chapter 4, we develop the functional nonlinear Cox proportional hazards model (FNCPH) to investigate the effect of a functional covariate on censored survival outcomes. To the best of our knowledge, this model is the first of its kind. We reduce the functional model to a kernel machine representation which is subsequently reframed as a mixed effects Cox model.

We first develop a main effects model which is a straightforward extension of the NFRM and the GNFRM. To add to the novelty of the method, we extended the main effects model to include a term that models the nonlinear interaction between a single scalar covariate and a single functional covariate. The flexibility in the interaction term is also a product of the kernel machine framework.

We construct the model with a linear kernel and a quadratic kernel. Simulation results show that the performance is similar to the NFRM for both the main effects model and the interaction model. Thus, we suggest using the FNCPH model with both kernels as a companion test–similar to the recommendation for the NFRM. Although we did not investigate estimation procedures in this work, it's easy to see that we can estimate the effect of the functional covariate on survival times via currently existing mixed model approaches such as the frailty models in the R package coxme [Ripatti and Palmgren, 2000, Therneau et al., 2003]. In the interaction model, the main effects of the functional covariate are modeled using a parametric quadratic form. Although the numerical studies that were conducted for the NFRM showed that this quadratic representation is capable of detecting a wide range of nonlinear relationships, it is of further research interest to develop an approach to model both the main and interaction effects with kernel machine representations.

Each model developed in this work is motivated by unique problems in cancer genomics. In each chapter, we discuss the proposed model in context of the motivating science. Our goal is to build a case for casting copy number association as functional regression models. Recall that the level of serial correlation in the copy number profile justifies the novel approach to analyzing copy number data. In Chapter 2, we regressed a continuous prognostic marker for multiple myeloma onto local regions of copy number alterations. We conducted a genome wide association analysis and detected several regions that were previously reported to be related to the prognostic marker. In Chapter 3, we dichotomized stages of multiple myeloma and investigated the association between local regions of copy number alteration on the binary response. Again we detected genomic regions that were reported to be associated with survival in multiple myeloma patients. In Chapter 4, we switched our focus to the aggressive brain cancer glioblastoma multiforme where our goal was to investigate the effect of local regions of copy number alteration on survival time. We defined local regions by considering the copy number probes that covered genes in the glioblastoma multiforme pathway. Our analyses suggested that copy number alterations were not significantly related to survival times in pediatric and adult patients for a small selection of genes in the glioblastoma multiforme pathway. In general, all of the data analyses contained in this work, (Chapters 2-4), highlight the ability of our models to be used to prioritize genomic regions and genes for further biological investigation. Although our work is motivated by the area of cancer genomics, we again underscore the fact that our

models can be broadly applied to other areas where there is an interest in testing for a nonlinear effect of a functional covariate on a scalar response.

REFERENCES

National cancer institute: Cancer staging, May 2013. URL `http://www.cancer.gov/cancertopics/factsheet/detection/staging`.

M. C. Aguilera-Morillo, A. M. Aguilera, M. Escabias, and M. J. Valderrama. Penalized spline approaches for functional logit regression. *Test*, 22(2):251–277, September 2013.

C. Alkan, B. P. Coe, and E. E. Eichler. Genome structural variation discovery and genotyping. *Nature Reviews. Genetics*, 12:363–376, 2011.

D. Atanackovic, J. Panse, Y. Hildebrandt, A. Jadczak, S. Kobold, Y. Cao, J. Templin, S. Meyer, H. Reinhard, K. Bartels, N. Lajmi, A. Zander, A. Marx, C. Bokemeyer, and N. Kröger. Surface molecule cd229 as a novel target for the diagnosis and treatment of multiple myeloma. *Haematologica*, 96(10):1512–1520, October 2011.

H. Avet-Loiseau, C. Li, F. Magrangeas, W. Gouraud, C. Charbonnel, J. Harousseau, M. Attal, G. Marit, C. Mathiot, T. Facon, P. Moreau, K. C. Anderson, L. Campion, N. C. Munshi, and S. Minvielle. Prognostic significance of copy-number alterations in multiple myeloma. *Journal of Clinical Oncology*, 27(27):4585–4590, September 2009.

V. Baladandayuthapani, Y. Ji, R. Talluri, L. E. Nieto-Barajas, and J. S. Morris. Bayesian random segmentation models to identify shared copy number aberrations for array cgh data. *Journal of the American Statistical Association*, 105(492):1358–1375, 2010.

R. Bataille, B. G. M. Durie, and J. Grenierj. Serum beta2 microglobulin and survival duration in multiple myeloma: a simple reliable marker for staging. *British Journal of Haematology*, 55(3):439–447, 1983.

N. E. Breslow and D. G. Clayton. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421):9–25, March 1993.

T. Cai, G. Tonini, and X. Lin. Kernel machine approach to testing the significance of multiple genetic markers for risk prediction. *Biometrics*, 67(3):975–986, September 2011. doi: DOI: 10.1111/j.1541-0420.2010.01544.x.

F. Cappuzzo, F. R. Hirsch, E. Rossi, S. Bartolini, G. L. Ceresoli, L. Bemis, J. Haney, S. Witta, K. Danenberg, I. Domenichini, V. Ludovini, E. Magrini, V. Gregorc, C. Doglioni, A. Sidoni, M. Tonato, W. A. Franklin, L. Crino, P. A. Bunn Jr., and M. Varella-Garcia. Epidermal growth factor receptor gene and protein and gefitinib sensitivity in non–small-cell lung cancer. *Journal of the National Cancer Institute*, 97(5):643–655, 2005.

H. Cardot, F. Ferraty, A. Mas, and P. Sarda. Testing hypotheses in the func- tional linear model. *Scandinavian Journal of Statistics*, 30:241–255, 2003a.

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

H. Cardot, A. Goia, and P. Sarda. Testing for no effect in functional linear regression models, some computational approaches. *Communications in Statistics - Simulation and Computation*, 30, 2004.

Q.-R. Chen, Y. Hu, C. Yan, K. Buetow, and D. Meerzaman. Systematic genetic analysis identifies cis-eqtl target genes associated with glioblastoma patient survival. *PLoS One*, 9(8), 2014. doi: 10.1371/journal.pone.0105393.

D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

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

N. Cristianini and J. Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press, 2000.

R. B. Davies. Hypothesis testing when a nuisance parameter is present only under the alternatives. *Biometrika*, 74(1):33–43, March 1987.

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

S. J. Diskin, T. Eck, J. Greshock, Y. P. Mosse, T. Naylor, C. J. Stoeckert Jr., B. L. Weber, J. M. Maris, and G. R. Grant. Stac: A method for testing the significance of dna copy number aberrations across multiple array-cgh experiments. *Genome Research*, 16:1149–1158, 2006.

M. Escabias, A. M. Aguilera, and M. J. Valderrama. Modeling environmental data by functional principal component logistic regression. *Environmetrics*, 16:95–107, 2005.

F. Ferraty and P. Vieu. *Nonparametric Functional Data Analysis Theory and Practice*. Springer, 2006.

S. Gobin, P. Biesta, and P. Van den Elsen. Regulation of human beta 2-microglobulin transactivation in hematopoietic cells. *Blood*, 101(8):3058–3064, April 2003.

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

J. Goldsmith, S. Greven, and C. Crainiceanu. Corrected confidence bands for functional data using principal components. *Biometrics*, doi: 10.1111/j.1541-0420.2012.01808.x, 2012.

P. R. Greipp, J. S. Miguel, B. G. Durie, J. J. Crowley, B. Barlogie, J. Bladé, M. Boccadoro, J. A. Child, H. Avet-Loiseau, R. A. Kyle, J. J. Lahuerta, H. Ludwig, G. Morgan, R. Powles, K. Shimizu, C. Shustik, P. Sonneveld, P. Tosi, I. Turesson, and J. Westin. International staging system for multiple myeloma. *Journal of Clinical Oncology*, 23(15):3412–3421, 2005.

T. Hastie and R. Tibshirani. Varying-coefficient models. *Journal of the Royal Statistical Society. Series B*, 55(4):757–796, 1993.

R. Jain, L. Poisson, J. Narang, D. Gutman, L. Scarpace, S. N. Hwang, C. Holder, M. Wintermark, R. R. Colen, J. Kirby, J. Freymann, D. J. Brat, C. Jaffe, and T. Mikkelsen. Genomic mapping and survival prediction in glioblastoma: Molecular subclassification strengthened by hemodynamic imaging biomarkers. *Radiology*, 267(1):212–220, April 2013.

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

D. R. Johnson, D. J. Ma, J. C. Buckner, and J. E. Hammack. Conditional probability of long-term survival in glioblastoma. *Cancer*, 118(22):5608–5613, 2012.

E. Jung, H. Moon, S. Park, B. Cho, S. Lee, C. Jeong, Y. Ju, S. Jeong, Y. Lee, S. Choi, W. Ha, J. Lee, K. Kang, and S. Hong. Decreased annexin a3 expression correlates with tumor progression in papillary thyroid cancer. *Protemics. Clinical Applications.*, 4(5):528–537, May 2010.

S. S. Keerthi and C. Lin. Asymptotic behaviors of support vector machines with gaussian kernel. *Neural Computation*, 15(7), 1667-1689 2003.

J. Kim, C. Yoo, D. Lee, S. Kim, J. Lee, and C. Suh. Serum albumin level is a significant prognostic factor reflecting disease severity in symptomatic multiple myeloma. *Annals of Hematology*, 89(4):391–397, April 2010.

G. Kimeldorf and G. Wahba. Some results on tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, 33(1):82–95, January 1971. doi: DOI:10.1016/0022-247X(71)90184-3.

J. P. Klein and M. L. Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data (Statistics for Biology and Health)*. Springer, second edition, 2003.

D. Kong, A.-M. Staicu, and A. Maity. Classical testing in functional linear models. Technical report, North Carolina State University, 2013.

M. Krishnan, J. S. Temel, A. A. Wright, R. Bernacki, K. Selvaggi, and T. Balboni. Predicting life expectancy in patients with advanced incurable cancer: a review. *The Journal of Supportive Oncology*, 11(2):68–74, June 2013.

L. Kwee, D. Liu, X. Lin, D. Ghosh, and M. Epstein. A powerful and flexible multilocus association test for quantitative traits. *American Journal of Human Genetics*, 82(386-397), 2008.

X. Lin, T. Cai, M. C. Wu, Q. Zhou, G. Liu, D. C. Christiani, and X. Lin. Kernel machine snp-set analysis for censored survival outcomes in genome-wide association studies. *Genetic Epidemiology*, 35(7):620–631, November 2011.

D. Liu, X. Lin, and D. Ghosh. Semiparametric regression of multidimensional genetic pathway data: Least-squares kernel machines and linear mixed models. *Biometrics*, 63(4):1079 – 1088, December 2007.

D. Liu, D. Ghosh, and X. Lin. Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. *BMC Bioinformatics*, 9(1):292–303, June 2008.

Y. Liu, S. Shete, C. J. Etzel, M. Scheurer, G. Alexiou, G. Armstrong, S. Tsavachidis, F. Liang, M. Gilbert, K. Aldape, T. Armstrong, R. Houlston, F. Hosking, L. Robertson, Y. Xiao, J. Wiencke, M. Wrensch, U. Andersson, B. S. Melin, and M. Bondy. Polymorphisms of lig4, btbd2, hmga2, and rtel1 genes involved in the double-strand break repair pathway predict glioblastoma survival. *Journal of Clinical Oncology*, 28(14):2467–2474, 2010.

B. D. Marx and P. H. Eilers. Generalized linear regression on sampled signals and curves: a p-spline approach. *Technometrics*, 41:1–13, 1999.

S. McAvoy, Y. Zhu, D. Perez, C. James, and D. Smith. Disabled-1 is a large common fragile site gene, inactivated in multiple cancers. *Genes, Chromosomes and Cancer*, 47(2):165–174, February 2008.

M. W. McLean, G. Hooker, A.-M. Staicu, F. Scheipl, and D. Ruppert. Functional generalized additive models. *Journal of Computational and Graphical Statistics*, DOI:10.1080/10618600.2012.729985, 2012.

J. Mercer. Functions of positive and negative type, and their connection with the theory of integral equations. *The Royal Society*, Series A 83(559):69–70, November 1909.

A. Misra, M. Pellarin, J. Nigro, I. Smirnov, D. Moore, K. R. Lamborn, D. Pinkel, D. G. Albertson, and B. G. Feuerstein. Array comparative genomic hybridization identifies genetic subgroups in grade 4 human astrocytoma. *Clinical Cancer Research*, 11:2907–2918, 2005.

G. J. Morgan, B. A. Walker, and F. E. Davies. The genetic architecture of multiple myeloma. *Nature Reviews. Cancer*, 12:335–348, May 2012.

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

H.-G. Müller and U. Stadtmüller. Generalized functional linear models. *Annals of Statistics*, 33(22):774–805, 2005.

N. Munshi, D. Longo, and K. Anderson. *Plasma Cell Disorders*, chapter 111. New York, McGraw-Hill Professional, 18th edition, 2011.

J. A. Nelder and W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society A.*, 135(3), 1972.

T. C. G. A. T. R. Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455:1061–1068, 2008.

M. Pignone, D. Nicoll, and S. McPhee. *Pocket Guide to Diagnostic Tests*. New York, McGraw-Hill, 4th edition, 2004.

D. Pinkel and D. G. Albertson. Array comparative genomic hybridization and its applications in cancer. *Nature Genetics*, 37, 2005.

Y. Qiang, Y. Endo, J. Rubin, and S. Rudikoff. Wnt signaling in b-cell neoplasia. *Oncogene*, 22 (10):1536–1545, March 2003.

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

J. O. Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer, second edition, 2005.

P. Reiss and R. Ogden. Functional principal component regression and functional partial least squares. *Journal of the American Statistical Association*, 102:984–996, 2007.

S. Ripatti and J. Palmgren. Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics*, 56(4):1016–1022, December 2000.

L. Rozier, E. El-Achkar, F. Apiou, and M. Debatisse. Characterization of a conserved aphidicolin-sensitive common fragile site at human 4q22 and mouse 6c1: Possible association with an inherited disease and cancer. *Oncogene*, 23(41):6872–6880, September 2004.

O. M. Rueda and R. Diaz-Uriarte. Detection of recurrent copy number alterations in the genome: Taking among-subject heterogeneity seriously. *BMC Bioinformatics*, 10(308), 2009.

J. R. Sawyer. The prognostic significance of cytogenetics and molecular profiling in multiple myeloma. *Cancer Genetics*, 204:3–12, 2011.

B. Schölkopf and A. J. Smola. *Learning with Kernels : Support Vector Machines, Regularization, Optimization, and Beyond.* Adaptive Computation and Machine Learning. MIT Press, 2002.

S. P. Shah, W. L. Lam, R. T. Ng, and K. P. Murphy. Modeling recur- rent dna copy number alterations in array cgh data. *Bioinformatics*, 23(13):450–458, 2007.

H. Shin. Partial functional linear regression. *Journal of Statistical Planning and Inference*, 139: 3405–3418, 2009.

W. Sun, F. A. Wright, Z. Tang, S. H. Nordgard, P. V. Loo, T. Yu, V. N. Kristensen, and C. M. Perou. Integrated study of copy number states and genotype calls using high-density snp arrays. *Nucleic Acids Research*, 37(16):5365–5377, 2009.

J. Suykens and J. Vandewalle. Least squares support vector machine classifiers. *Neural Processing Letters*, 9:293–300, 1999.

B. J. Swihart, J. Goldsmith, and C. M. Crainiceanu. Restricted likelihood ratio tests for functional effects in the functional linear model. *Technometrics*, 2013. doi: 10.1080/00401706.2013.863163.

T. Therneau, P. Grambsch, and V. Pankratz. Penalized survival models and frailty. *Journal of Computational and Graphical Statistics*, 12(1):156–175, March 2003.

R. G. Verhaak, K. A. Hoadley, E. Purdom, V. Wang, Y. Qi, M. D. Wilkerson, C. R. Miller, L. Ding, T. Golub, J. P. Mesirov, G. Alexe, M. Lawrence, M. O'Kelly, P. Tamayo, B. A. Weir, S. Gabriel, W. Winckler, S. Gupta, L. Jakkula, H. S. Feiler, J. G. Hodgson, C. D. James, J. N. Sarkaria, C. Brennan, A. Kahn, P. T. Spellman, R. K. Wilson, T. P. Speed, J. W. Gray, M. Meyerson, G. Getz, C. M. Perou, D. N. Hayes, and T. C. G. A. R. Network. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in pdgfra, idh1, egfr, and nf1. *Cancer Cell*, 17:98–110, 2010.

X. Wang, X. Li, F. Fan, S. Jiao, L. Wang, L. Zhu, Y. Pan, G. Wu, Z. Ling, J. Fang, and Y. Chen. Paqr3 plays a suppressive role in the tumorigenesis of colorectal cancers. *Carcinogenesis*, 33 (11):2228–2235, November 2012.

J. Wessel and N. J. Schork. Generalized genomic distance–based regression methodology for multilocus association analysis. *The American Journal of Human Genetics*, 79:792–806, 2006.

S. L. Win, Z. Z. Htike, F. Yusof, and I. A. Noorbatcha. Gene expression mining for predicting survivability of patients in early stages of lung cancer. *International Journal on Bioinformatics and Biosciences*, 4(2), June 2014.

M. C. Wu, P. Kraft, M. P. Epstien, D. M. Taylor, S. J. Chanock, D. J. Hunter, and X. Lin. Powerful snp-set analysis for case-control genome-wide association studies. *The American Journal of Human Genetics*, 86(6):929 942, June 2010.

M. C. Wu, S. Lee, T. Cai, Y. Li, M. Boehnke, and X. Lin. Rare-variant association testing for sequencing data with the sequence kernel association test. *American Journal of Human Genetics*, 89(1):82–93, 2011.

M. C. Wu, A. Maity, S. Lee, E. M. Simmons, Q. E. Harmon, X. Lin, S. M. Engel, J. J. Molldrem, and P. M. Armistead. Kernel machine snp-set testing under multiple candidate kernels. *Genetic Epidemiology*, 37(3):267–275, 2013.

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

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

D. Zhang and X. Lin. Hypothesis testing in semiparametric additive mixed models. *Biostatistics*, 4(1):57–74, 2003.

APPENDICES

Appendix A

Functional Principal Component Analysis

A.1 Basic Concepts

Functional principal component analysis (FPCA) is a common tool used in functional regression models to circumvent the phenomena known as "the curse of dimensionality." It captures the dominate modes of variation within random curves. Through this process, it greatly reduces the dimension of the data while capturing the important features.

Let $X(t) \in L^2(\mathcal{T})$, where $t \in \mathcal{T}$ and $\mathcal{T}$ is a compact interval. Without a loss of generality, assume that $X(\cdot)$ is a mean-zero random process. Define the covariance of the process as $V(s,t) = \text{Cov}\{X(s), X(t)\}$. Given that $V(\cdot,\cdot)$ is a continuous symmetric and positive definite kernel function, Mercer's theorem states that the covariance of the process can be represented as $V(s,t) = \sum_{j=1}^{\infty} \lambda_j \phi_j(s)\phi_j(t)\ dt$, where $\{\phi_j : j \geq 1\}$ are the eigenfunctions and $\{\rho_j : j \geq 1\}$ are the corresponding eigenvalues associated with the spectral decomposition of the covariance operator $\rho_j \phi_j(s) = \int_{\mathcal{T}} V(s,t)\phi(t)\ dt$. The eigenfunctions yield an orthonormal basis for $L^2(\mathcal{T})$ such that the inner product is represented as follows, $\int_{\mathcal{T}} \phi_j(s)\phi_k(s) = \varsigma_{jk}$. Here $\varsigma_{jk} = 1$ if $j = k$ and $\varsigma_{jk} = 0$ if $j \neq k$.

In FPCA, the eigenfunctions are referred to as functional principal components (FPCs). The principal components are viewed as weight functions such that $\xi_{ij} = \int \phi_j(t) X_i(t) dt$ is the random functional principal component score for the ith subject that corresponds to the jth mode of variation in the covariance of the process. The first FPC is chosen so that $\phi_1(t)$ maximizes $N^{-1} \sum_{i=1}^{n} \xi_{i1}^2$ subject to the constraint $||\phi_1||^2 = 1$. Thus, the FPC score ξ_{i1} captures the largest mode of variation in the covariance of the process. A second FPC, $\phi_2(t)$, is selected so that it maximizes $N^{-1} \sum_{i=1}^{n} \xi_{i2}^2$ subject to the constraints $||\phi_2||^2 = 1$ and $\int \phi_1(t)\phi_2(t) dt = 0$.

The additional constraint ensures that the second principal component captures the second largest mode of variation remaining in $V(s,t)$. This process is then repeated to obtain the desired number of FPCs. Each successive weight function must be orthogonal to all previous weight functions and will capture successively smaller modes of variation within $V(s,t)$.

Based on the results of Mercer's theorem, The Karhenun-Loéve expansion is used to represent the random process as $X_i(t) = \sum_{j=1}^{\infty} \xi_{ij}\phi_j(t)$. In practice, the infinite sum is reduced to a finite set of J FPCs, and the random process is approximated as $X_i(t) = \sum_{j=1}^{J} \xi_{ij}\phi_j(t)$. In our analyses, we use a common approach selecting J which consists of choosing the number of FPCs that account for a predetermined level of functional variation explained (FVE) within $V(s,t)$. The eigenvalue that is associated with each functional principal component is the variance of its associated score. In other words, $\text{Var}(\xi_{ij}) = \rho_j$. Thus, the FVE of the first J principal components is defined as $\text{FVE} = \sum_{j=1}^{J} \rho_j / \sum_{l=1}^{\infty} \rho_l$. To see more clearly how to choose J, let's consider a brief example. Assume that we want to conduct an FPCA that accounts for 95% of the variation in some data. Now let's assume that $\sum_{j=1}^{3} \rho_j / \sum_{l=1}^{\infty} \rho_l = 0.94$ and $\sum_{j=1}^{4} \rho_j / \sum_{l=1}^{\infty} \rho_l = 0.99$. Then the truncation point for this analysis is $J = 4$. It is common to set the FVE to 95% or 99%.

In reality, the functional principal components and their associated scores are not known. There are several methods to estimate $\{\phi_j\}_{j=1}^{J}$ and corresponding $\{\xi_{ij}\}_{j=1}^{J}$ for $i = 1, \ldots, n$. We do not detail them all here, but it's important to note that most of the available methods have a similar theme:

1. Estimate the covariance of the process $V(s,t)$.
2. Estimate the mean of the process, which we denote as $\mu(t) = \text{E}X_i(t)$.
3. Estimate the FPCs.
4. Estimate the FPC scores.

The various methods may differ in how they estimate these objects; however, the final result is always $X_i(t) = \sum_{j=1}^{J} \hat{\xi}_{ij}\hat{\phi}_j(t)$. Note that we did not put a "hat ($\hat{}$)" on $X_i(t)$. This is because the common practice in the FDA literature is to treat the estimates of the FPCs and the FPC scores as proxies for truth.

Recall that Chapter 2 and Chapter 3 used the PACE method of Yao et al. [2005] to estimate the FPC scores, and Chapter 4 used the more recent smooth covariance approach of Di et al. [2009]. We discuss these two approaches in more detail in Section A.2 and Section A.3, respectively.

A.2 Principal Analysis by Conditional Expectation

The principal analysis by conditional expectation (PACE) approach to FPCA was developed to handle sparse and irregular longitudinal data observed with error. The basic assumptions for this approach are that the data has been observed on a sparse and irregular grid across subjects, and the data has been polluted with measurement error. This creates two key problems: (1) numerical integration may not provide a good estimate of the FPC scores when the data is sparse, and (2) using the observations that are polluted with measurement error may lead to biased estimates of the scores. In addition, the assumption is that the grid on which observations are observed as well as the number of observations for each subject are independently and identically distributed random variables that are independent of all other random elements in the model. Thus, the observed data are represented by a modified Karhenun-Loéve expansion, $X_i^\star(t_{ik}) = X_i(t_{ik}) + \epsilon_{ik} = \mu(t_{ik}) + \sum_{j=1}^{\infty} \xi_{ij}\phi_j(t_{ik}) + \epsilon_{ik}$, $t_{ij} \in \mathcal{T}$ for $i = 1, \ldots, n$ and $k = 1, \ldots, N_i$.

As with classical approaches to FPCA, PACE approaches functional principal component analysis in four steps. The first step consist of estimating $\mu(t)$ via local linear-kernel smoothers, and a "raw" estimate of the resulting covariance matrix is estimated via the method of moments using $\hat{\mu}(t)$. The next goal is to estimate $V(s,t)$. This is not straight forward due to the measurement error. Consider the model for the observed functional data, $W_{ij} = \mu(t) + X_i(t_{ij}) + \epsilon_{ij}$ for $i = 1, \ldots, n$, $j = 1, \ldots, N_i$, and $t_{ij} \in \mathcal{T}$. Thus, $G(s,t) = \text{Cov}\{X_{ij}(s), X_{ij}(t)\} + \sigma^2\mathbf{I}(s = t) = V(s,t) + \sigma^2\mathbf{I}(s = t)$. Note that the diagonal elements are impacted by the measurement error. Thus, a two step approach is taken to obtain estimate $V(s,t)$. First, the off-diagonal elements ($t \neq s$) corresponding to $V(s,t)$ are estimated by smoothing the off-diagonal elements of $\hat{G}(s,t)$ using local linear-kernel smoothers. Next the diagonal elements are separately estimated via local linear-kernel smoothers, and the estimate of the measurement error variance is estimated by $\hat{\sigma}^2 = \frac{2}{|\mathcal{T}|}\int\{\hat{G}(s,t) - \hat{V}(s,t)\}\ dt$, where $|\mathcal{T}|$ is the length of the domain $\mathcal{T}$. The kernel-smoothers borrow information across subjects to help negate the effects of the sparseness and irregularity in the data. Next the eigenfunctions and eigenvalues are estimated by conducting an eigenanalysis on the discretized smoothed covariance generated by $\hat{V}(s,t)$.

To further address the issue of measurement error, the FPC scores are computed via conditional expectation under a mixed model framework. Again, consider the model for the functional data, $W_{ij} = X_i(t_{ij}) + \epsilon_{ij} = \mu(t_{ij}) + \sum_{k=1}^{\infty} \xi_{ik}\phi_k(t_{ij}) + \epsilon_{ij}$, where $t_{ij} \in \mathcal{T}$, $E(\epsilon_{ij}) = 0$ and $Var(\epsilon_{ij}) = \sigma^2$. Now assume that ξ_{ik} and ϵ_{ij} are jointly Gaussian. Given this Gaussian assumption, the BLUP of the principal component scores is $\tilde{\xi}_{ik} = E[\xi_{ik}|\widetilde{\mathbf{W}}_i] = \rho_k\phi_{ik}^T\Sigma_{W_i}^{-1}(\widetilde{\mathbf{W}}_i - \boldsymbol{\mu}_i)$, where $\Sigma_{W_i} = cov(\mathbf{W}_i, \mathbf{W}_i)$. The *estimated* BLUP is thus $\hat{\xi}_{ik} = \hat{\rho}_k\hat{\phi}_{ik}^T\widehat{\Sigma}_{W_i}^{-1}(\widetilde{\mathbf{W}}_i - \boldsymbol{\mu}_i)$.

A.3 Smooth Covariance Approach

The smooth covariance approach used in Chapter 4 was developed as part of a larger solution to perform FPCA on multilevel functional data [Di et al., 2009]. The term "multilevel" refers to functional data with subject specific effects as well as cluster effects. The approach is also applicable to single level problems where there is no clustering in the data, for example modeling copy number ratios across the genome. The smooth covariance approach considers data that has been densely or sparsely observed, as well as data with and without measurement error. It differs from PACE in that the grid upon which the data are observed is assumed to be the same for all subjects, and sparseness is induced as a result of missing observations. In the PACE approach, each subject is allowed to have data observed on irregular grids with the number of observations (N_i) being a random variable that is *iid* across subjects.

The basic process to conducting FPCA via this method is similar to that laid out in Section A.2. First, lets consider the model for the observed curve for the i*th* subject, $W_{ij} = \mu(t) + X_{ij}(t) + \epsilon_{ij}$ for $i = 1, \ldots, n$, $j = 1, \ldots, n_i$, and $t \in \mathcal{T}$. Note the subtle difference in this model as opposed to that presented in Section A.2. Again, the primary interest is in estimating $V(s,t)$, which is complicated by the measurement error in the functional data. Similar to PACE, the mean function ($\mu(t)$) is first estimated with the exception that this approach uses penalized spline smoothing as opposed to local linear-kernel smoothers. Next, $G(s,t)$ is estimated via the method of moments and $V(s,t)$ is estimated by smoothing $\hat{G}(s,t)$ for $t \neq s$ via penalized thin plate spline smoothing. Again, the diagonal elements are estimated separately, and the variance of the measurement error is estimated as $\hat{\sigma}^2 = \int \{\hat{G}(s,t) - \hat{V}(s,t)\}\, dt$. Lastly, the resulting covariance matrix is discretized to obtain estimates of the eigenvalues and eigenfunctions. The estimation of the FPC scores follow directly from the PACE method discussed in Section A.2.

Appendix B

NFRM: Additional Results

B.1 Simulation Results

B.1.1 Estimation Results

Table B.1: Simulation results for the estimation of the functional effect, $\mathcal{L}\{X_i(\cdot)\}$, in the model $Y_i = \beta_1 z_{1i} + \beta_2 z_{2i} + \mathcal{L}\{X_i(\cdot)\} + \epsilon_i$ based on 1000 generated data sets.

Definitions: NFRM, nonlinear functional regression model; FLM, functional linear model; FQRM, functional quadratic regression model; FVE, functional variation; MSE, mean squared error; ABSE, absolute error.

		NFRM					FQRM					FLM				
n	FVE	MSE	ABSE	Int	Slope	Corr	MSE	ABSE	Int	Slope	Corr	MSE	ABSE	Int	Slope	Corr
							Functional 1									
100	0.85	0.669	0.511	-0.010	1.029	0.883	0.690	0.528	-0.005	0.916	0.877	0.626	0.469	-0.004	0.979	0.894
	0.99	0.506	0.396	-0.009	1.009	0.945	0.554	0.431	-0.011	0.905	0.937	0.398	0.315	-0.010	0.978	0.970
200	0.85	0.544	0.405	-0.007	1.013	0.922	0.554	0.413	-0.009	0.955	0.920	0.501	0.364	-0.006	0.989	0.933
	0.99	0.419	0.328	-0.007	1.005	0.960	0.434	0.339	-0.008	0.951	0.958	0.342	0.273	-0.008	0.989	0.975

Table B.2: Simulation results of the estimation of β_1 in the model $Y_i = \beta_1 z_{1i} + \beta_2 z_{2i} + \mathcal{L}\{X_i(\cdot)\} + \epsilon_i$ based on 1000 generated data sets. True $\beta_1 = 1$.

Definitions: NFRM, nonlinear functional regression model; FLM, functional linear model; FQRM, functional quadratic regression model; FVE, functional variation; RMSE, root mean squared error; SE, standard error

		NFRM				FQRM				FLM			
n	FVE	Bias	RMSE	SE	Coverage(%)	Bias	RMSE	SE	Coverage(%)	Bias	RMSE	SE	Coverage(%)
						Functional 1							
100	0.85	-0.006	0.122	0.123	0.950	-0.005	0.118	0.118	0.948	-0.006	0.126	0.126	0.948
	0.99	-0.004	0.114	0.114	0.945	-0.004	0.108	0.108	0.946	-0.003	0.118	0.118	0.950
200	0.85	-0.005	0.082	0.081	0.936	-0.005	0.079	0.079	0.935	-0.005	0.082	0.082	0.939
	0.99	-0.003	0.077	0.077	0.939	-0.002	0.074	0.075	0.936	-0.003	0.078	0.078	0.942
						Functional 2							
100	0.85	-0.010	0.194	0.179	0.941	-0.001	0.315	0.306	0.939	-0.011	0.195	0.181	0.938
	0.99	-0.005	0.144	0.140	0.952	-0.001	0.316	0.307	0.942	-0.005	0.144	0.141	0.951
200	0.85	-0.003	0.116	0.110	0.949	0.007	0.217	0.215	0.940	-0.003	0.116	0.111	0.948
	0.99	-0.004	0.095	0.094	0.954	0.007	0.218	0.215	0.940	-0.004	0.095	0.094	0.955
						Functional 3							
100	0.85	-0.004	0.133	0.133	0.949	-0.002	0.156	0.162	0.953	-0.005	0.135	0.136	0.948
	0.99	-0.002	0.124	0.124	0.938	-0.002	0.157	0.162	0.953	-0.002	0.127	0.127	0.943
200	0.85	-0.002	0.087	0.088	0.948	-0.003	0.114	0.113	0.944	-0.002	0.087	0.089	0.950
	0.99	-0.003	0.083	0.084	0.949	-0.002	0.114	0.113	0.945	-0.003	0.084	0.084	0.950
						Functional 4							
100	0.85	-0.004	0.132	0.130	0.949	0.001	0.157	0.159	0.953	-0.006	0.134	0.133	0.949
	0.99	-0.004	0.124	0.122	0.946	0.001	0.159	0.159	0.949	-0.005	0.125	0.125	0.954
200	0.85	-0.003	0.086	0.087	0.946	-0.002	0.114	0.111	0.936	-0.003	0.087	0.087	0.946
	0.99	-0.004	0.084	0.083	0.943	-0.002	0.114	0.111	0.937	-0.004	0.084	0.083	0.944
						Functional 5							
100	0.85	-0.011	0.226	0.217	0.942	0.009	0.389	0.392	0.961	-0.013	0.226	0.218	0.945
	0.99	-0.005	0.159	0.157	0.946	0.009	0.393	0.392	0.957	-0.006	0.159	0.157	0.944
200	0.85	-0.000	0.129	0.129	0.950	-0.004	0.277	0.274	0.947	-0.000	0.129	0.129	0.953
	0.99	-0.003	0.104	0.105	0.950	-0.004	0.278	0.274	0.946	-0.003	0.104	0.106	0.950

Table B.3: Simulation results of the estimation of β_2 in the model $Y_i = \beta_1 z_{1i} + \beta_2 z_{2i} + \mathcal{L}\{X_i(\cdot)\} + \epsilon_i$ based on 1000 generated data sets. True $\beta_2 = 1$.

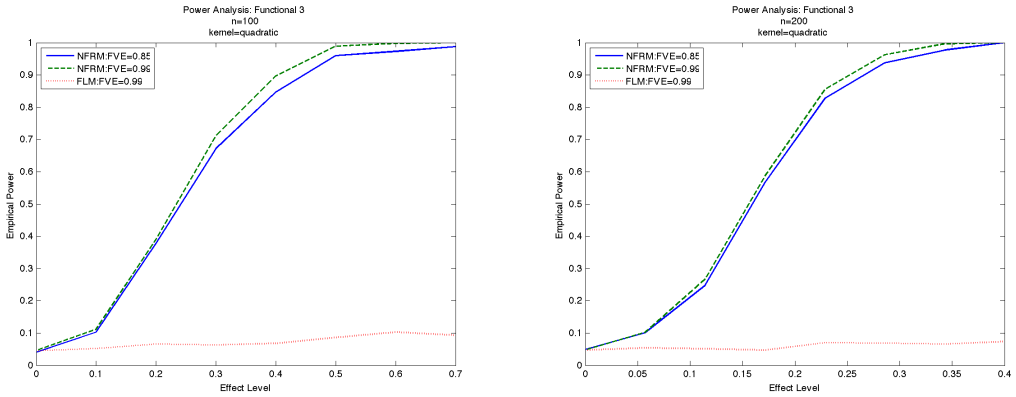
Definitions: NFRM, nonlinear functional regression model; FLM, functional linear model; FQRM, functional quadratic regression model; FVE, functional variation; RMSE, root mean squared error; SE, standard error

		NFRM				FQRM				FLM			
n	FVE	Bias	RMSE	SE	Coverage(%)	Bias	RMSE	SE	Coverage(%)	Bias	RMSE	SE	Coverage(%)
						Functional 1							
100	0.85	-0.004	0.273	0.259	0.937	-0.005	0.263	0.249	0.940	-0.003	0.281	0.265	0.938
	0.99	-0.006	0.255	0.239	0.933	-0.006	0.239	0.227	0.937	-0.005	0.266	0.247	0.930
200	0.85	-0.006	0.176	0.171	0.949	-0.005	0.170	0.166	0.948	-0.006	0.178	0.172	0.950
	0.99	-0.006	0.169	0.163	0.948	-0.005	0.162	0.158	0.948	-0.006	0.170	0.164	0.952
						Functional 2							
100	0.85	-0.003	0.408	0.377	0.940	0.007	0.668	0.644	0.941	-0.003	0.414	0.381	0.936
	0.99	0.002	0.319	0.295	0.929	0.006	0.670	0.646	0.934	0.002	0.322	0.296	0.932
200	0.85	0.002	0.247	0.232	0.953	0.016	0.461	0.453	0.944	0.003	0.247	0.232	0.953
	0.99	-0.001	0.199	0.199	0.951	0.015	0.461	0.453	0.941	-0.001	0.199	0.198	0.948
						Functional 3							
100	0.85	0.010	0.301	0.281	0.935	0.005	0.348	0.341	0.944	0.013	0.307	0.287	0.927
	0.99	0.004	0.285	0.262	0.936	0.002	0.348	0.342	0.942	0.004	0.288	0.267	0.932
200	0.85	-0.003	0.188	0.186	0.947	-0.005	0.235	0.237	0.950	-0.003	0.189	0.187	0.945
	0.99	-0.006	0.178	0.176	0.945	-0.005	0.235	0.238	0.951	-0.007	0.180	0.177	0.944
						Functional 4							
100	0.85	0.007	0.291	0.274	0.924	0.003	0.343	0.334	0.940	0.008	0.297	0.279	0.930
	0.99	0.001	0.276	0.256	0.926	0.001	0.343	0.335	0.941	-0.001	0.282	0.262	0.933
200	0.85	-0.005	0.188	0.181	0.944	-0.010	0.239	0.232	0.943	-0.004	0.188	0.182	0.943
	0.99	-0.007	0.181	0.173	0.937	-0.011	0.239	0.232	0.944	-0.007	0.181	0.174	0.933
						Functional 5							
100	0.85	0.020	0.466	0.454	0.936	0.018	0.805	0.824	0.958	0.020	0.472	0.456	0.934
	0.99	-0.002	0.353	0.329	0.922	0.013	0.805	0.825	0.960	-0.003	0.355	0.330	0.920
200	0.85	-0.006	0.284	0.271	0.941	-0.031	0.602	0.576	0.934	-0.006	0.284	0.271	0.940
	0.99	-0.007	0.232	0.221	0.932	-0.032	0.601	0.576	0.934	-0.007	0.232	0.221	0.933

B.1.2 Testing Results

Table B.4: Simulation results to evaluate Type I Error Rate based on 1,000,000 generated data sets and $n = 200$. Values are displayed in percentages.

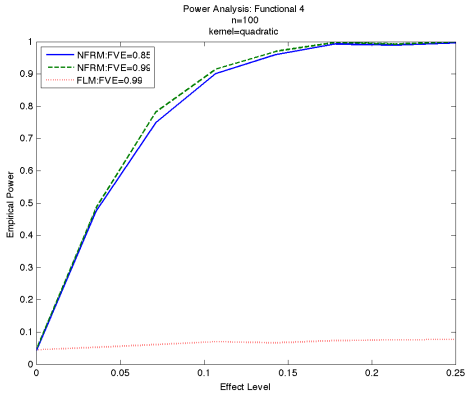
	Type I Error	NFRM	FLM
n =200	$5 \times (10^{-}2)$	5.1	5.2
	$1 \times (10^{-}2)$	1.1	1.1
	$5 \times (10^{-}3)$	5.8	5.4
	$1 \times (10^{-}3)$	1.4	1.1
	$5 \times (10^{-}4)$	7.7	5.6
	$1 \times (10^{-}4)$	1.8	1.3



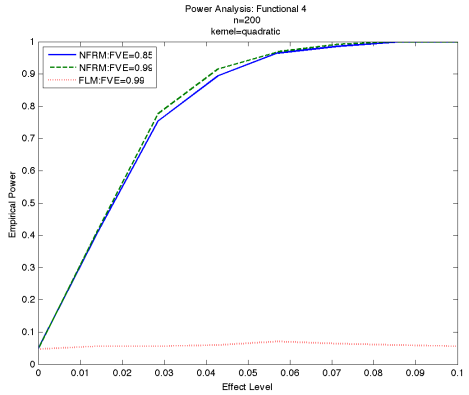
(a) Sample Size = 100

(b) Sample Size = 200

Figure B.1: Simulation results of the rejection probability as a function of b for functional 3. The left and right panel shows the results for $n = 100$ and $n = 200$, respectively.

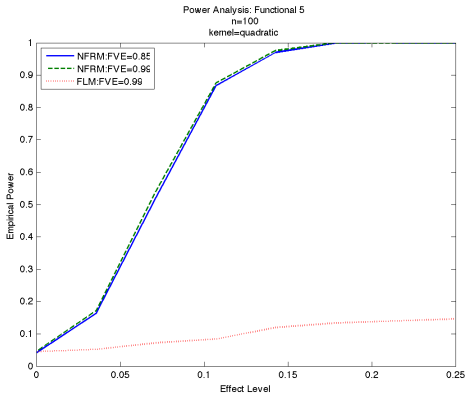


(a) Sample Size = 100

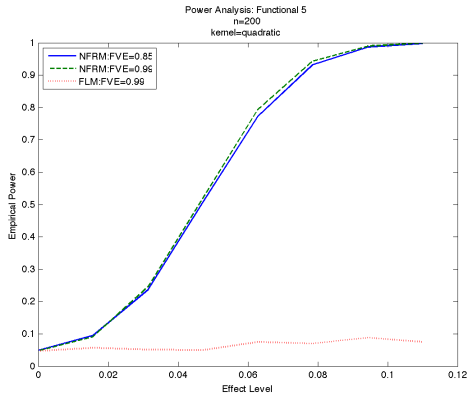


(b) Sample Size = 200

Figure B.2: Simulation results of the rejection probability as a function of b for functional 4. The left and right panel shows the results for $n = 100$ and $n = 200$, respectively.

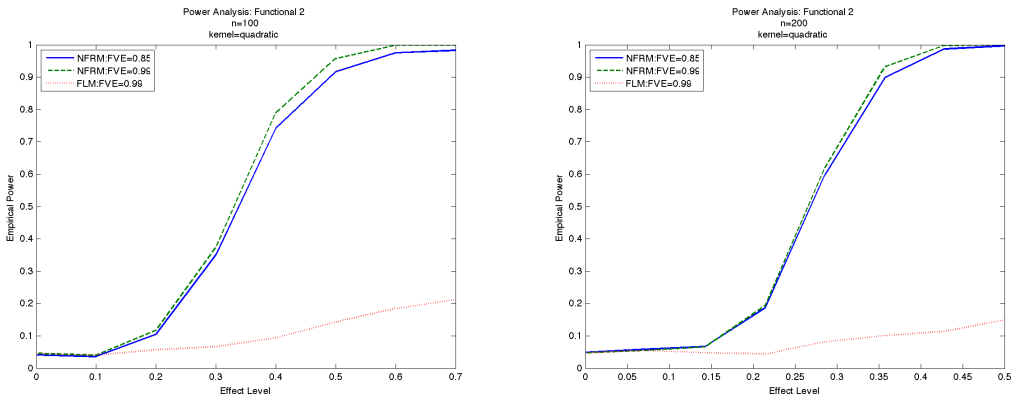


(a) Sample Size = 100



(b) Sample Size = 200

Figure B.3: Simulation results of the rejection probability as a function of b for functional 5. The left and right panel shows the results for $n = 100$ and $n = 200$, respectively.



(a) Sample Size = 100

(b) Sample Size = 200

Figure B.4: Simulation results of the rejection probability as a function of b for the quadratic functional. The left and right panel shows the results for $n = 100$ and $n = 200$, respectively.

B.2 Data Analysis Results

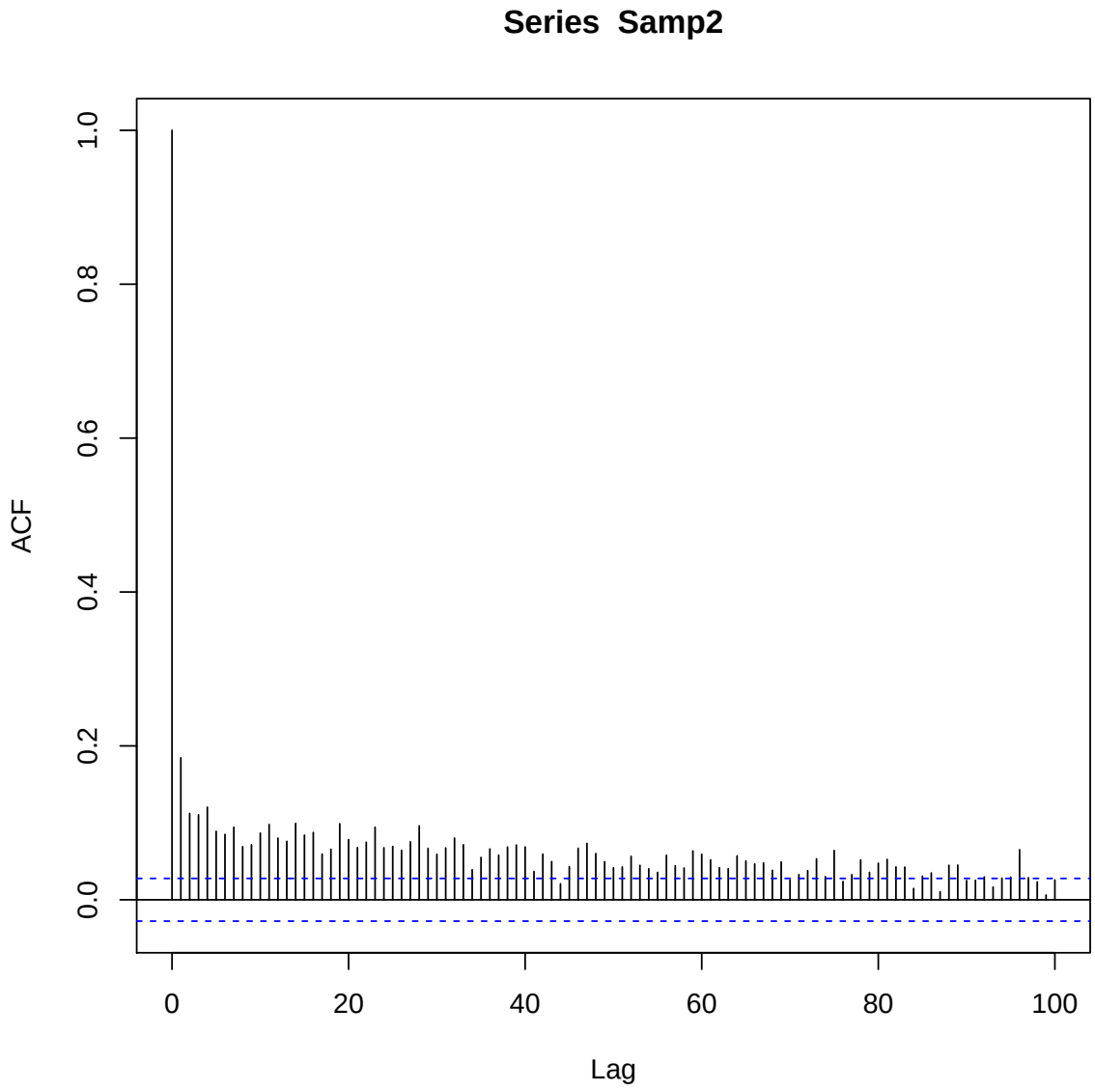


Figure B.5: Serial correlation in the copy number profile of the chromosome 1 p-arm for a second random sample.

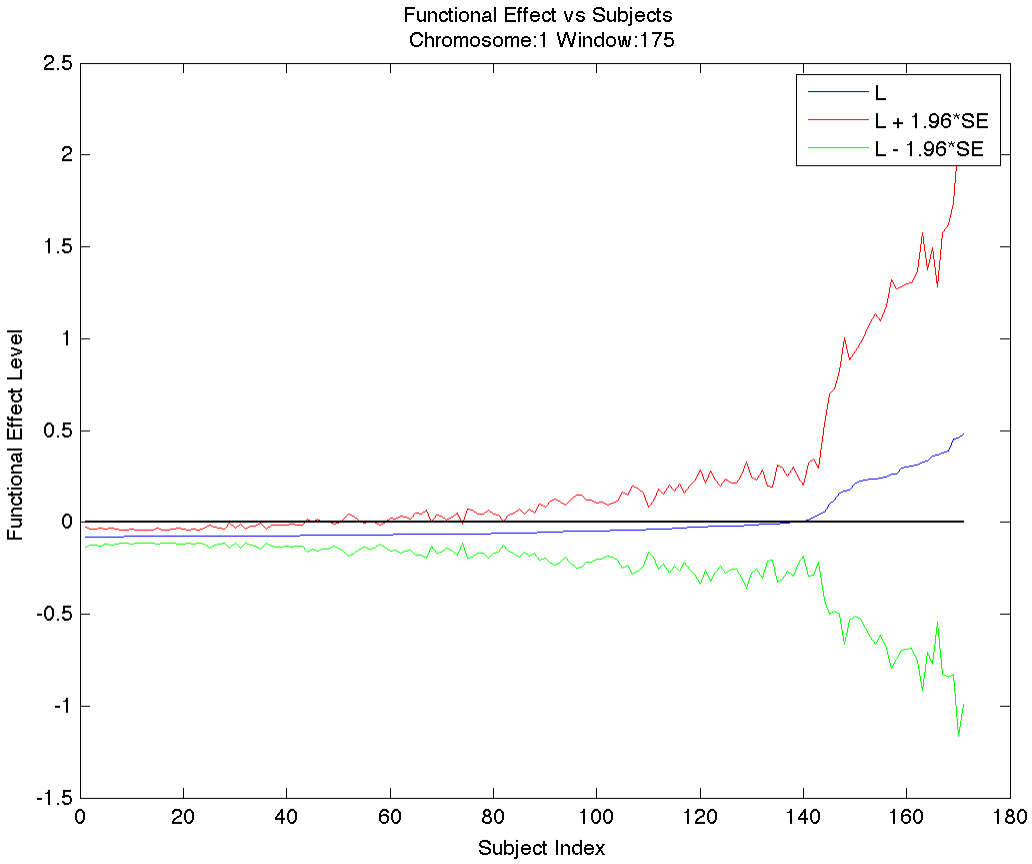


Figure B.6: Ordered estimated copy number profile effect (along with pointwise 95% confidence bands) by subject for window 175 of chromosome 1 p-arm.

Table B.5: Significant genomic locations related to B2M identified using NFRM and FLM. The reference for the genomic locations is Human genome build hg18.

Chromosome	Start	End	Predicted By
1	97823290	100099870	NFRM
3	89509246	90391757	NFRM
4	78039006	83941746	NFRM
4	88325567	94816389	NFRM
14	60326236	61895272	NFRM
1	19233342	20305757	FLM
1	14237369	157747420	FLM
1	15849160	162845350	FLM
1	16451612	194825560	FLM
1	19694553	199626870	FLM
1	20046867	201327430	FLM
1	20822932	214439910	FLM
1	21815588	219308770	FLM
1	22128829	226997220	FLM
1	22905672	229977540	FLM
1	24445878	245785230	FLM
2	16593000	166859220	FLM
2	20165777	202446160	FLM
2	20296970	203920420	FLM
2	23147497	232830660	FLM
4	76740441	77543461	FLM
4	78039006	83941746	FLM
4	88325567	94816389	FLM
4	10110103	102680240	FLM
4	10438888	106300820	FLM
4	10896645	110102640	FLM
4	19102782	191173880	FLM
14	44424842	45975216	FLM
14	60326236	61250997	FLM
21	26311881	27514879	FLM

Appendix C

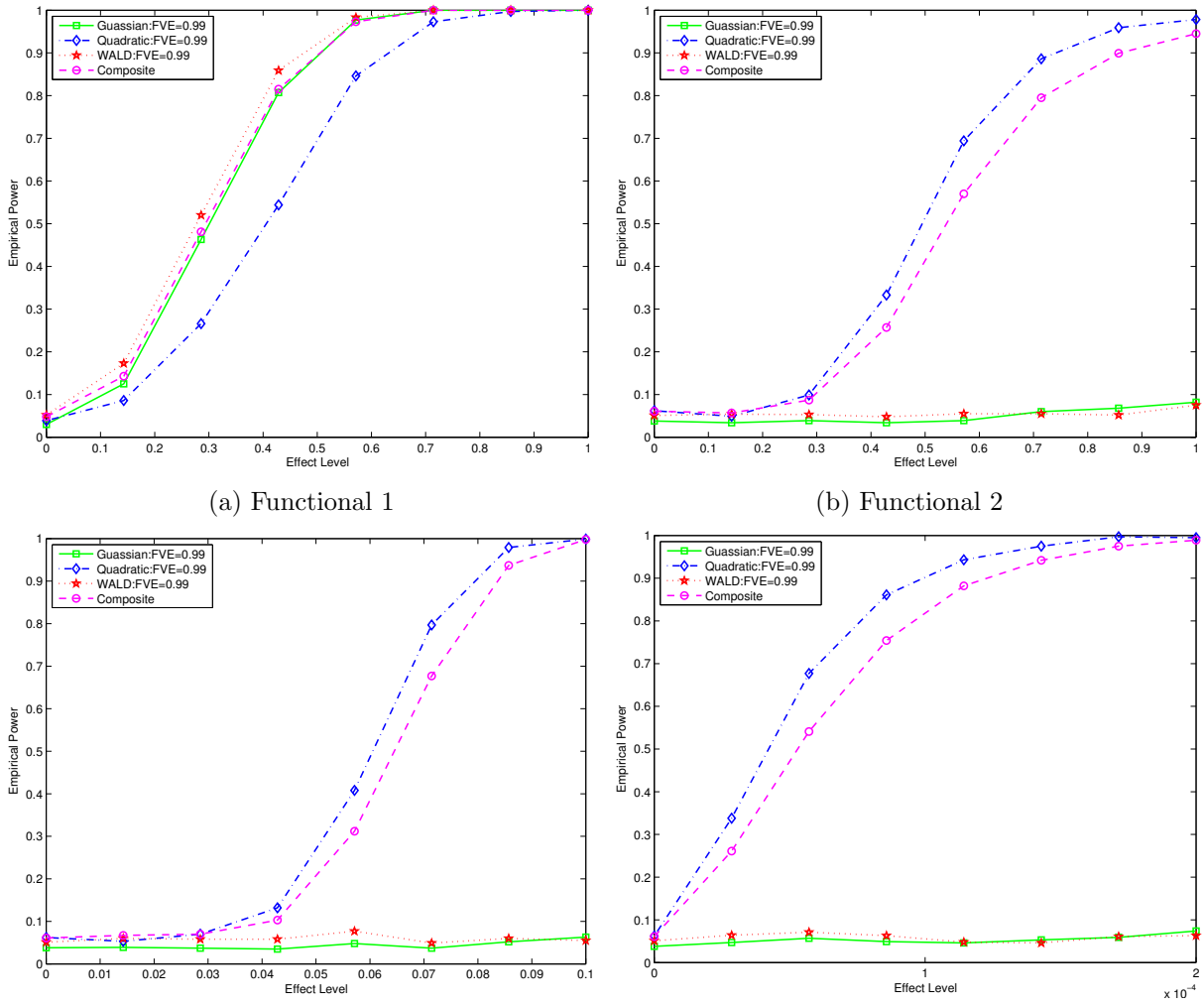
GNFRM: Additional Results

C.1 Simulation Results

Table C.1: Simulation results for type I error based on 1,000 generated datasets and $n = 200$. Standard errors for each estimate < 0.001.

Type I Error Rate	GNFRM_G	GNFRM_Q	GNFRM_CK	GFLRM
0.01	0.011	0.009	0.013	0.014
0.05	0.030	0.039	0.048	0.052
0.10	0.060	0.090	0.089	0.091

Definitions: GNFRM_G, GNFRM with Gaussian kernel; GNFRM_Q, GNFRM with quadratic kernel; GNFRM_CK, GNFRM with composite kernel



(a) Functional 1

(b) Functional 2

(c) Functional 3

(d) Functional 4

Figure C.1: This figure displays the Type I Error and power for each functional at FVE = 0.99 and $n = 200$. Here the composite kernel is included in the analysis. The solid line represents GNFRM with the Gaussian kernel; the dotted-dashed line represents GNFRM with the quadratic kernel; the dashed line represents GNFRM with a composite kernel; the dotted line represents the WALD test for the FLM.

C.2 Data Analysis Results

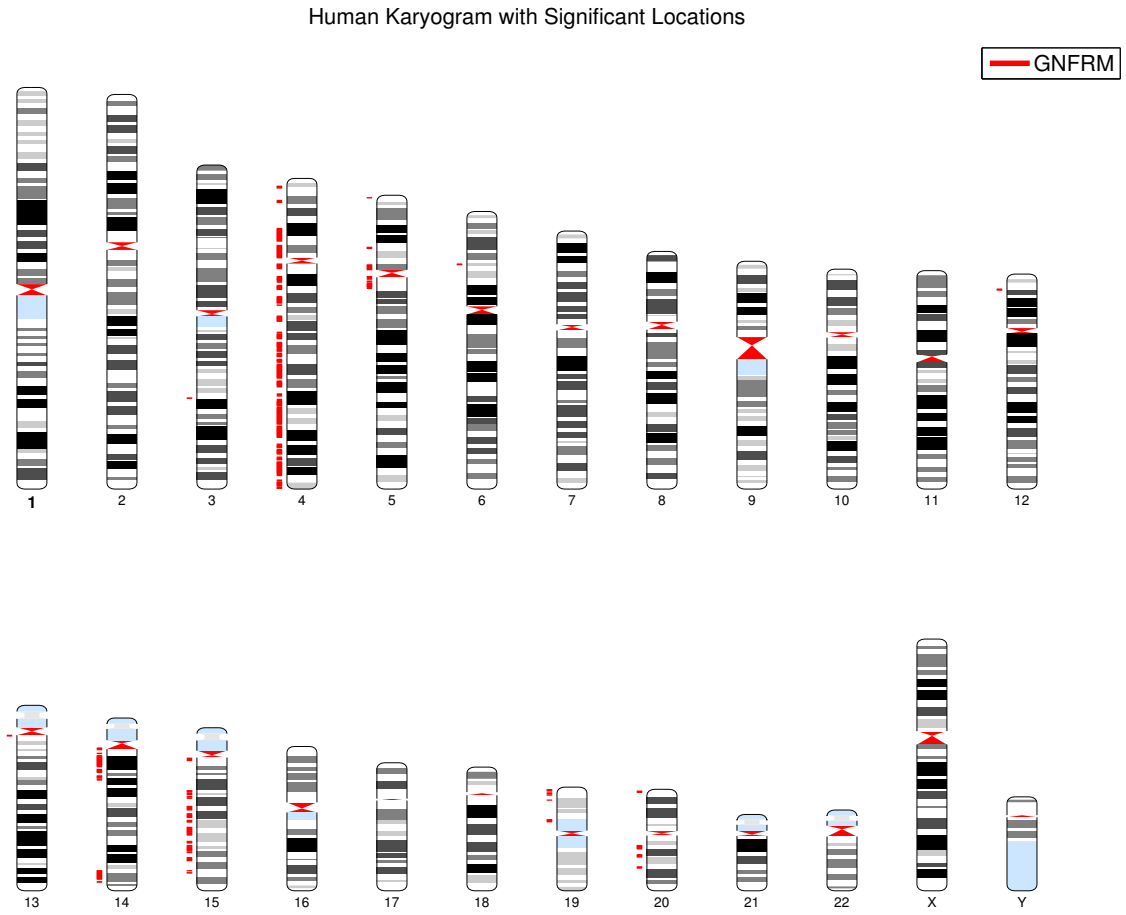


Figure C.2: Test results for the MM application. The figure is a karyogram that depicts the test results for GNFRM across the genome using a Bonferroni correction for multiple tests. Red regions to the left of each chromosome were identified by GNFRM.

Appendix D

FNCPH: Additional Results

D.1 Simulation Results

Table D.1: Empirical size of the tests for main effects at $n = 300$ at multiple type I error rates when censoring is moderate (25%) and high (50%). Testing was performed using (1) FNCPH with a linear kernel (Q_{lin}), (2) FNCPH with a quadratic kernel, (Q_{quad}), and (3) naive linear approach (LRT_L). The results are based on $50,000$ generated datasets.

Censoring	Type I Error Rate	Q_{lin}	Q_{quad}	LRT_L
25%	0.05	0.0565	0.0449	0.0746
	0.01	0.0109	0.0069	0.0178
	0.005	0.0054	0.0029	0.0096
	0.001	0.0011	0.0004	0.0021
50%	0.05	0.0536	0.0496	0.0687
	0.01	0.0102	0.0087	0.0149
	0.005	0.0051	0.0041	0.0081
	0.001	0.0010	0.00077	0.0018

Table D.2: Empirical size of the tests for interaction effects at $n = 300$ at multiple type I error rates when censoring is moderate (25%) and high (50%). Testing was performed using (1) FNCPH with a linear kernel (Q_{lin}), (2) FNCPH with a quadratic kernel, (Q_{quad}), and (3) naive linear approach (LRT_L). The results are based on $50,000$ generated datasets.

Censoring	Type I Error Rate	Q_{lin}	Q_{quad}	LRT_L
25%	0.05	0.0408	0.0360	0.1206
	0.01	0.0098	0.0073	0.0362
	0.005	0.0054	0.0035	0.0214
	0.001	0.0010	0.0007	0.0066
50%	0.05	0.0318	0.0257	0.1111
	0.01	0.0069	0.0043	0.0323
	0.005	0.0032	0.0018	0.0194
	0.001	0.0007	0.0004	0.0053

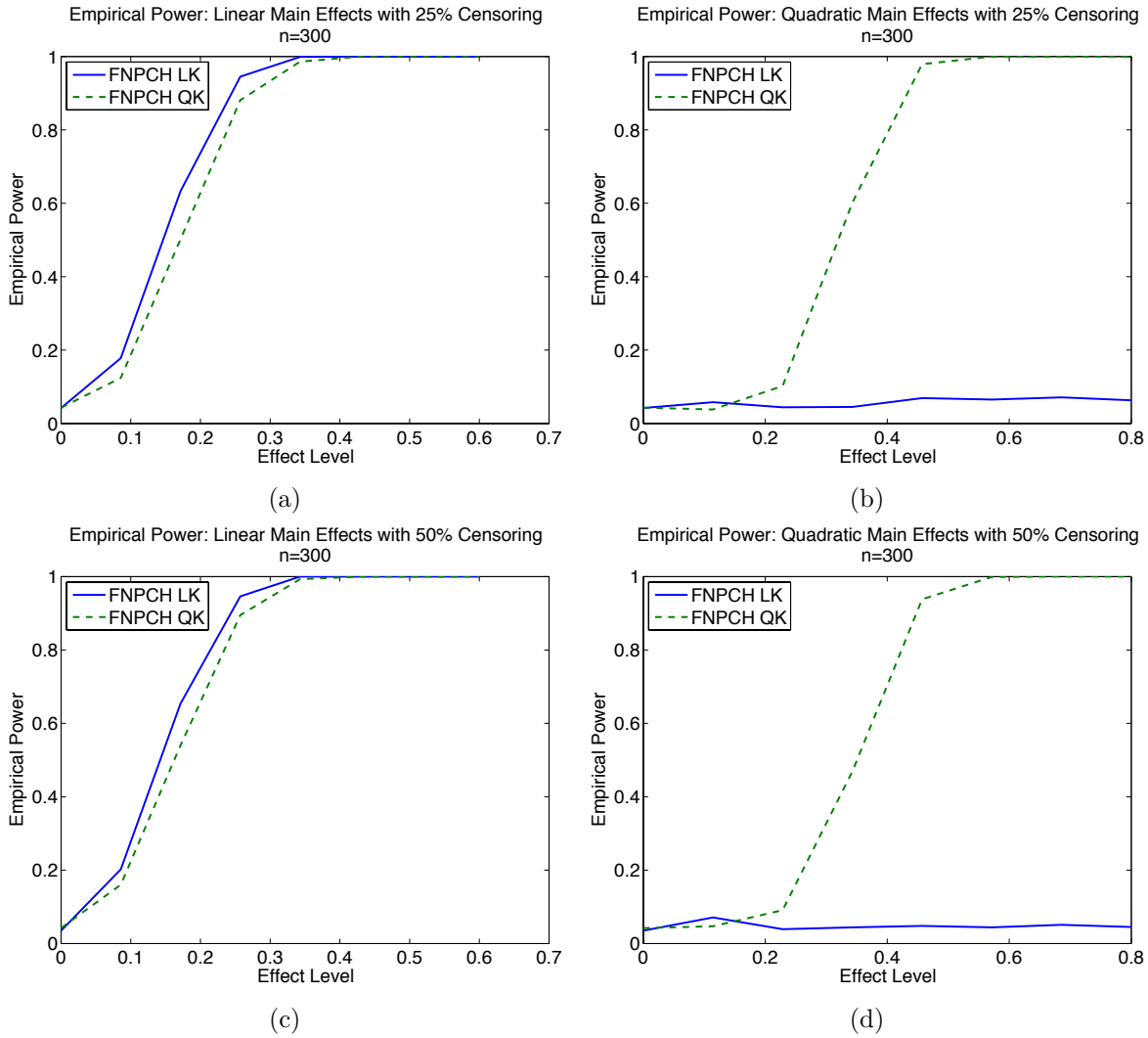


Figure D.1: Empirical power for main effects model at $n = 300$. The dashed line corresponds to FNCPH model with a quadratic kernel. The solid line corresponds to a FNCPH with a linear kernel.

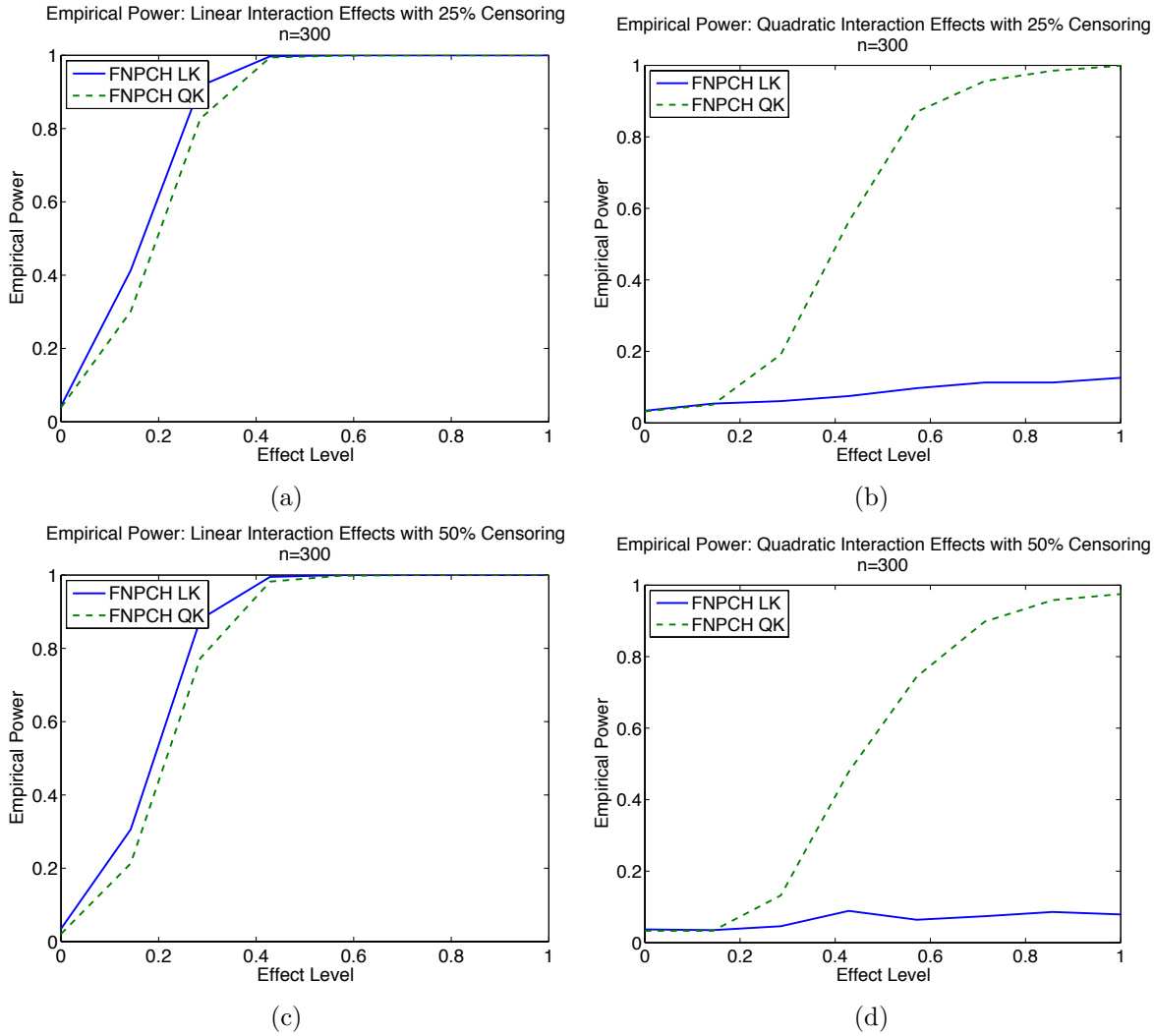


Figure D.2: Empirical power for interaction effects model at $n = 200$. The dashed line corresponds to FNCPH model with a quadratic kernel. The solid line corresponds to a FNCPH with a linear kernel.

D.2 Data Analysis Results

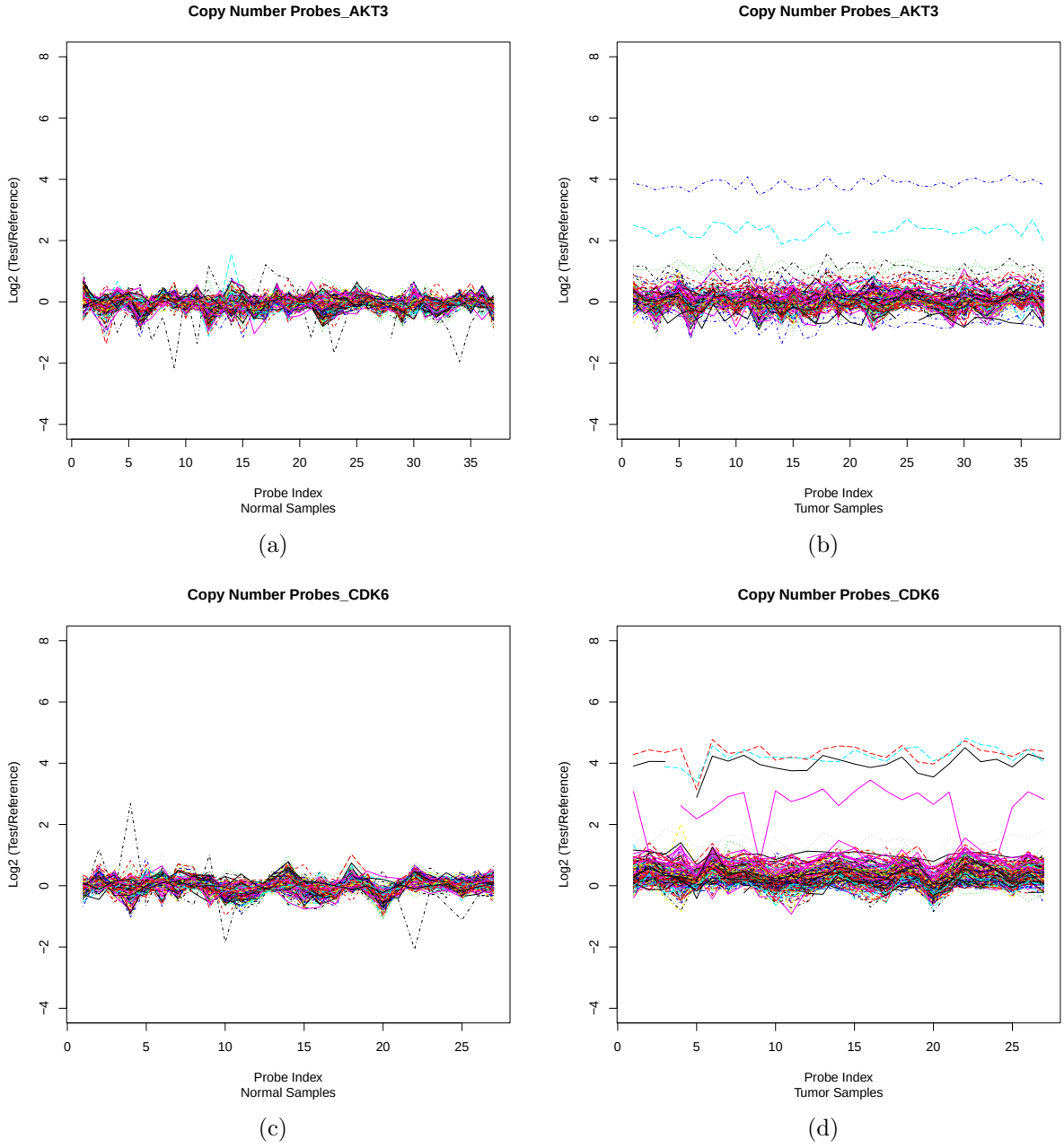


Figure D.3: Copy number intensities over two genes in the GBM pathway. The left panels display the copy number intensities measured in each patient's normal tissue sample. The right panels display the copy number intensities measured in each patient's diseased tissue sample.

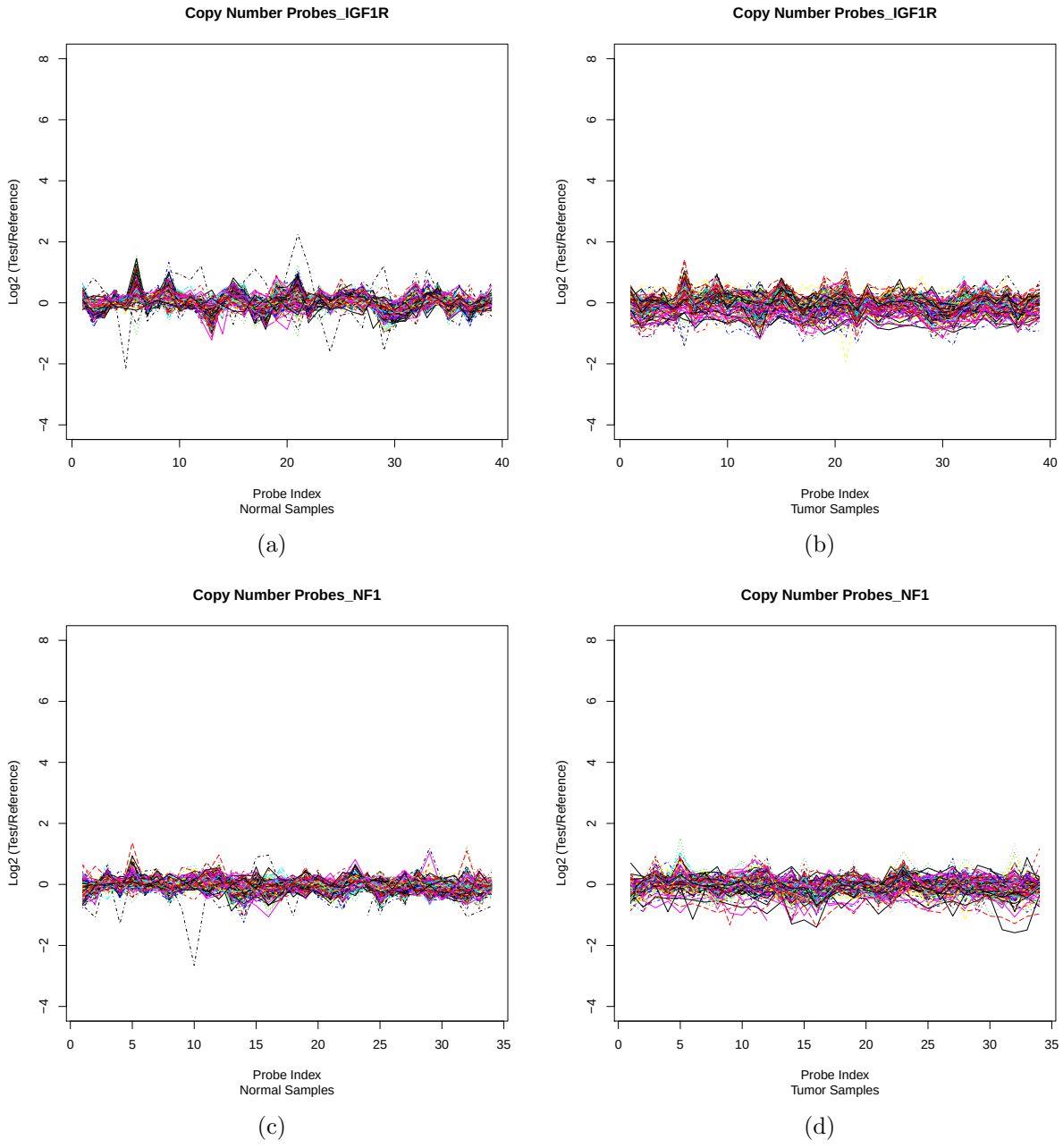
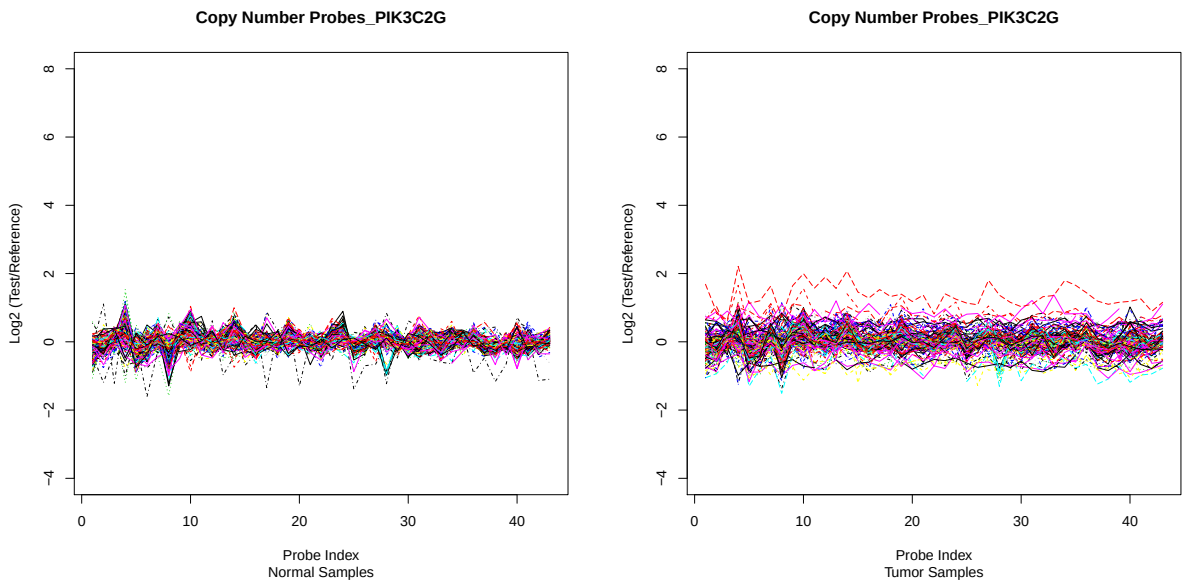
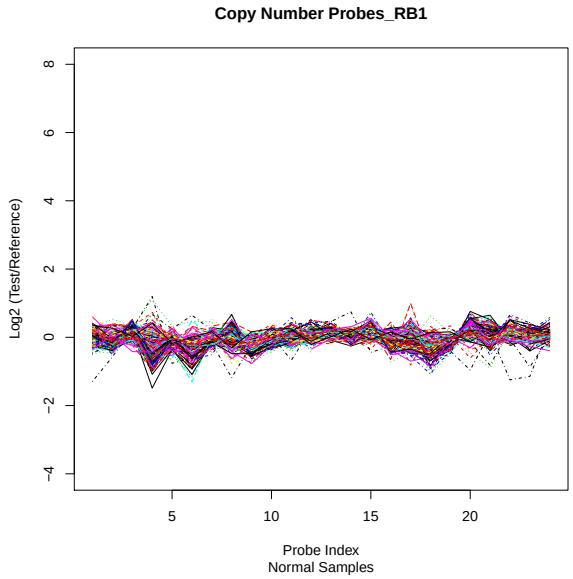


Figure D.4: Copy number intensities over two genes in the GBM pathway. The left panels display the copy number intensities measured in each patient's normal tissue sample. The right panels display the copy number intensities measured in each patient's diseased tissue sample.

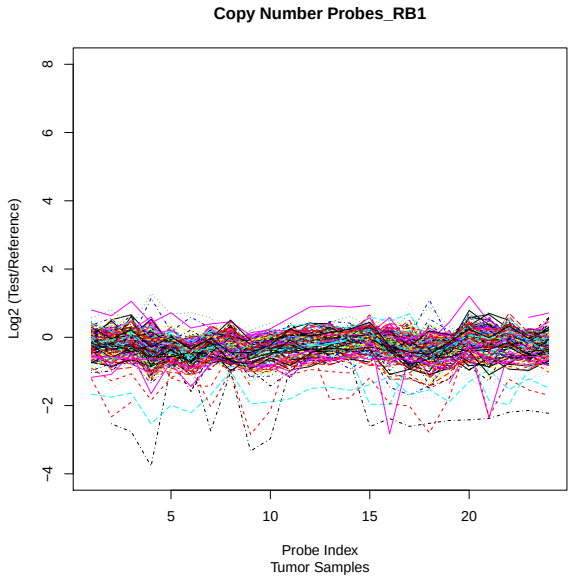


(a)

(b)



(c)



(d)

Figure D.5: Copy number intensities over two genes in the GBM pathway. The left panels display the copy number intensities measured in each patient's normal tissue sample. The right panels display the copy number intensities measured in each patient's diseased tissue sample.