Mediation analysis is regularly used to explore the relationships among a main covariate, a mediating variable, and a response. It investigates the effect of the mediator on the relationship between the covariate and the response. Mediation occurs when the covariate affects the mediator, which in turn affects the response. The relationship between the covariate and the response that is affected by the mediator is called the indirect effect and is the effect in which we are specifically interested. The indirect effect studies how the response changes due to the mediator when the covariate remains the same. We consider performing mediation analysis with a scalar response variable, a scalar and continuous mediating variable, and a functional covariate, meaning the covariate is a continuous curve rather than a single scalar measure. We incorporate functional regression methods with mediation analysis to account for a functional covariate under several different settings.

First, we consider the case where the response variable is continuous. We model the relationship between the mediator and the functional covariate using a functional linear model (FLM) and use functional principal component regression (FPCR) to model the FLM. We also model the effects of the functional covariate and the mediator on the response variable using a FLM and FPCR. We develop an estimation procedure for the indirect effect, along with a method to create simultaneous confidence intervals. We also develop three testing procedures for testing the nullity of the indirect effect: a Wald test, a bootstrap test, and a test based on simultaneous confidence intervals. We extend our method to account for a generalized response using a generalized functional linear model. Finally, we apply the methodology to a genetics application regarding multiple myeloma to investigate the indirect effect of gene expression on the relationship between functional copy number variation data and $\beta_2$ microglobulin, a known biomarker for multiple myeloma.

Next, we consider the case where the response variable is the time until an event occurs and survival analysis methods are needed. We continue to model the relationship between the mediator and the
functional covariate using a FLM and FPCR since the mediator is continuous. However, now we model
the effects of the functional covariate and the mediator on the time-to-event response using survival
techniques. We implement the accelerated failure time model and the Cox proportional hazards model
and use FPCR to account for the functional covariate. Again, we develop an estimation procedure for
the indirect effect and simultaneous confidence intervals. We also investigate the same three testing
procedures as in the continuous response case.

Lastly, we consider the case where we allow the functional covariate to be modeled nonlinearly. We
consider both a continuous response variable and a generalized response variable. We model the effect
of the covariate on the mediator and the effects of the covariate and the mediator on the response using
a functional generalized additive model as it does not make any assumptions regarding linearity. We
use a bivariate spline basis system in order to model the functional coefficient surface. We develop an
estimation procedure for the indirect effect. Finally, we apply our nonlinear methodology using functional
generalized additive models to the genetic copy number application regarding multiple myeloma.
Functional Mediation Analysis with an Application to Multiple Myeloma and Copy Number Variation Data

by
Sarah Louise Hale

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
2016

APPROVED BY:

_________________________  ___________________________
Jung-Ying Tzeng             Spencer Muse

_________________________  ___________________________
Eric Stone                  Arnab Maity
Chair of Advisory Committee
DEDICATION

To my family and friends for their endless support over the last six years.

In loving memory of Thomas Pautsch.
BIOGRAPHY

Sarah Louise Hale was born on April 18, 1988, to parents Brian and Laura Hale. She grew up outside of Milwaukee, Wisconsin, in the city of Muskego. In 2006, she graduated from Muskego High School and left the state of Wisconsin to continue her education at Eastern Kentucky University. While at college, Sarah was a member of the EKU Honors Program and majored in statistics and economics, with a minor in mathematics. After graduating summa cum laude from Eastern Kentucky University in 2010, she enrolled in the statistics PhD program at North Carolina State University. En route to her PhD, Sarah earned a Masters of Statistics in 2012. While working on her dissertation, she was a fellow on the National Institutes of Health grant Biostatistics Training in the Omics Era, directed by Dr. Spencer Muse. Under the direction of Dr. Arnab Maity, Sarah is set to earn her PhD in the summer of 2016. Upon graduation, Sarah will start working for the Duke Translational Medicine Institute Biostatistics Core within the Department of Biostatistics and Bioinformatics at Duke University. There she will support an interdisciplinary team of clinical investigators conducting research at Duke by providing expertise in study design, the implementation of statistical methodology, and the interpretation of results.
ACKNOWLEDGEMENTS

I would first like to thank my advisor, Dr. Arnab Maity. I am grateful for your guidance and support. This dissertation would not have been possible without your patience. I would also like to thank Dr. Jung-Ying Tzeng, Dr. Spencer Muse, and Dr. Eric Stone for taking the time to serve on my dissertation committee. Your valuable feedback has undoubtedly strengthened this dissertation.

There are numerous NC State faculty members I’d like to thank. Specifically, I’d like to thank Dr. John Monahan, Dr. Leonard Stefanski, Dr. Dennis Boos, and Dr. Eric Stone for being influential professors in my development at State and believing in me. I’d also like to thank Dr. Howard Bondell for his substantial help and support while serving as a graduate program director. In addition, I’d like to thank Dr. Spencer Muse for his extensive mentorship. Your support and direction during my time on the biostatistics training grant, as well as guidance while developing the ST 307 course, has been monumental. I’d also like to thank the incredible staff within the NCSU Department of Statistics. A special thank you to Alison McCoy for being such a sweet, caring person and always being there when I need help or advice.

I am also extremely thankful for the education and mentorship I received during my undergraduate studies at Eastern Kentucky University. Thank you to the wonderful statistics professors Dr. Patti Costello, Dr. Lisa Kay, and Dr. Michelle Smith. You three helped foster my passion for statistics and encouraged me to pursue a graduate degree. In addition, I’d like to greatly thank Dr. Fred Ruppel. You revived my love of learning with your joy and dedication to your students. My economics courses were some of my favorite courses to date.

I am very fortunate to have had one of the best mentors back in high school as well. A huge thank you to Kenneth Bosch for taking a vested interest in my education and life in general. You encouraged and inspired me to pursue greatness and always believed in me. Your mentorship, and ultimately friendship, helped guide me throughout the years.

My time in graduate school would not have been complete without the help and support of my fellow classmates and friends. I especially want to thank Becky Hager, Colleen McKendry, Michele Balik-Meisner, Adam Suarez, Brian Gaines, Kristin Linn, Sarah Mikol, Stephanie Moody, Jen Deis, Josh
Katz, Matt Phelan, and Andrew Wilcox. Your support and encouragement inside the classroom as well as your invaluable friendship outside helped me survive graduate school. I will always cherish our amazing grad school memories together.

I would also like to thank my friends outside of graduate school. Thank you to my closest friends from EKU, Emily Adams and Megan and Kyle Wright, and my oldest friends back home in Wisconsin, Amy Murphy, Emily Fumo, Renee Aschenbrenner, and Heather Keck. Thank you for your support even though we don’t get to see each other often anymore. No matter how much time has passed, we always pick up right where we left off. I also want to thank all of my friends in Raleigh from Meetup. I want to give a special thanks to Simon Evangelista, Kevin Garofalo, Allie Jacobs, Alex Cozort, Nakita Pasour, and numerous others for supporting me and helping me unwind from graduate school. I am grateful for your friendship and for such memorable experiences with you all, including whitewater rafting, cabin trips, and beach trips. I also participate on a women’s volleyball team and want to thank the beautiful and talented ladies of Tenacious D-fense for supporting me and accepting me into their family. I am thankful that a shared love of volleyball turned into such wonderful friendship.

Most importantly, I’d like to thank my parents, Brian and Laura, and my brother, Scott, for their constant love and support. You have always supported me and encouraged me to pursue my dreams. You pushed me to challenge myself and instilled in me the value of hard work. The successes you have achieved in life motivate me to push harder and excel further. I am grateful to have you all in my life.
TABLE OF CONTENTS

LIST OF TABLES ................................................................. viii

LIST OF FIGURES .............................................................. ix

Chapter 1  Introduction to Mediation Analysis ........................................... 1
  1.1 Literature Review ........................................................... 1
    1.1.1 Overview ............................................................ 1
    1.1.2 Standard Mediation Versus Causal Mediation Using a Counterfactual Model 4
    1.1.3 Testing the Indirect Effect ........................................ 7

Chapter 2  Introduction to Functional Regression Methods .......................... 9
  2.1 Literature Review .......................................................... 9
    2.1.1 Functional Linear Models ........................................... 9
    2.1.2 Generalized Functional Linear Models ............................ 14
    2.1.3 Nonlinear Functional Generalized Additive Models ............... 14
  2.2 Combining Mediation Analysis and Functional Regression Methods ........... 17

Chapter 3  Standard Functional Mediation Analysis Using Principal Components for Gaussian and Generalized Responses .................. 18
  3.1 Model and Framework ..................................................... 18
    3.1.1 Product of Coefficients Method for the Indirect Effect Function 18
    3.1.2 Testing the Indirect Effect Function ............................... 20
    3.1.3 Modeling a Generalized Response Using a Generalized Functional Linear Model 22
  3.2 Simulation Study .......................................................... 23
    3.2.1 Design and Data Generation ....................................... 23
    3.2.2 Estimation Results .................................................. 24
    3.2.3 Testing Results ..................................................... 26
  3.3 Application to Multiple Myeloma and Copy Number Variation Data ......... 32
  3.4 Summary ................................................................. 35

Chapter 4  Causal Functional Mediation Analysis Using Principal Components for Time-to-Event Data ..................................... 37
  4.1 Model and Framework for Time-to-Event Data Using a Counterfactual Approach 37
  4.2 Details on Time-to-Event Methods ....................................... 41
  4.3 Estimation of and Testing the Indirect Effect Function .................... 42
    4.3.1 Functional Principal Component Regression ....................... 42
    4.3.2 Estimation of the Indirect Effect Function ....................... 43
    4.3.3 Testing the Indirect Effect Function ............................... 44
  4.4 Simulation Study ........................................................ 45
    4.4.1 Purpose and Predictions ......................................... 45
    4.4.2 Design and Data Generation ....................................... 46
    4.4.3 Estimation Results ................................................ 47
### LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 3.1</td>
<td>ISB, IMSE, and IC for $\delta(\cdot)$ for IEF (1) and IEF (2). Coverages are given in terms of percent. The maximum standard error for the estimates in this table is 0.003.</td>
<td>25</td>
</tr>
<tr>
<td>Table 3.2</td>
<td>Type 1 error and power results for the Wald test and simultaneous interval test for IEF (1) for a significance level of $\alpha = 0.05$. The top table refers to the case when the response variable is Gaussian and $n = 100$. The bottom table refers to the case when the response variable is binary and $n = 250$.</td>
<td>28</td>
</tr>
<tr>
<td>Table 3.3</td>
<td>Summary information for the 11 genes that have a significant IEF using the Wald test, including gene name, the chromosome on which the gene is located, the p-value from the Wald test, the number of FPCs needed to account for 99% of the variability, the number of CNV probes associated with the gene, gene function, and known associated diseases. Information is from the NCBI (<a href="http://www.ncbi.nlm.nih.gov/">http://www.ncbi.nlm.nih.gov/</a>).</td>
<td>34</td>
</tr>
<tr>
<td>Table 4.1</td>
<td>ISB, IMSE, and IC for $\delta(\cdot)$ for IEF (1) and IEF (2) for all three survival models. Coverages are given in terms of percent. The mean standard error for the estimates in this table is 0.003.</td>
<td>49</td>
</tr>
<tr>
<td>Table 4.2</td>
<td>Type 1 error and power results for the Wald test and simultaneous interval test for IEF (1) under the exponential data generation scheme when $n = 250$ for a significance level of $\alpha = 0.05$.</td>
<td>51</td>
</tr>
<tr>
<td>Table 5.1</td>
<td>Bias, MSE, and 95% coverage for the IE when the response variable is Gaussian. Coverages are given in terms of percent. The mean standard error for the estimates in this table is 0.00003.</td>
<td>64</td>
</tr>
<tr>
<td>Table 5.2</td>
<td>Bias, MSE, and 95% coverage for the IE when the response variable is binary. Coverages are given in terms of percent. The maximum standard error for the estimates in this table is 0.00002.</td>
<td>65</td>
</tr>
<tr>
<td>Table 5.3</td>
<td>The 11 genes with a standardized IE estimate above the cut-off value of 4.389 when $\kappa = 2$ from a FGAM fit.</td>
<td>69</td>
</tr>
<tr>
<td>Table 5.4</td>
<td>The six genes with a standardized IE estimate above the cut-off value of 4.389 when $\kappa = 2$ from a FLM fit.</td>
<td>70</td>
</tr>
<tr>
<td>Table 6.1</td>
<td>Type 1 error results when $\beta_M = 0$ for the Wald test, simultaneous interval test, and bootstrap test for a standard mediation model using the POC method when the response variable is Gaussian and $n = 100$ for significance level $\alpha = 0.05$.</td>
<td>75</td>
</tr>
</tbody>
</table>
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1.1</td>
<td>An illustration of a simple mediation model.</td>
<td>2</td>
</tr>
<tr>
<td>Figure 3.1</td>
<td>Power curves for the results in Table 3.2 for IEF (1) when the response variable is Gaussian and ( n = 100 ).</td>
<td>29</td>
</tr>
<tr>
<td>Figure 3.2</td>
<td>Power curves for the results in Table 3.2 for IEF (1) when the response variable is binary and ( n = 250 ).</td>
<td>29</td>
</tr>
<tr>
<td>Figure 3.3</td>
<td>Pointwise (dashed) and simultaneous (dotted) confidence intervals for one simulation of IEF (1) when ( n = 250 ) and the response variable is Gaussian. In the top graph, ( \beta_M = 2 ) and ( \gamma_K = (0, 0, 0, 0)^T ) so the IEF is ( \delta(s) = 0 ). In the bottom graph, ( \beta_M = 2 ) and ( \gamma_K = (0, 0, 2, 0, 0)^T ) so the IEF is ( \delta(s) = 0.4[\sqrt{2}\cos(2\pi s)] ).</td>
<td>30</td>
</tr>
<tr>
<td>Figure 3.4</td>
<td>Pointwise (dashed) and simultaneous (dotted) confidence intervals for one simulation of IEF (1) when ( n = 250 ) and the response variable is binary. In the top graph, ( \beta_M = 2 ) and ( \gamma_K = (0, 0, 0, 0)^T ) so the IEF is ( \delta(s) = 0 ). In the bottom graph, ( \beta_M = 2 ) and ( \gamma_K = (0, 0, 2, 0, 0)^T ) so the IEF is ( \delta(s) = 0.4[\sqrt{2}\cos(2\pi s)] ).</td>
<td>31</td>
</tr>
<tr>
<td>Figure 3.5</td>
<td>Graphs of the IEF estimate (solid) and its pointwise confidence interval (dashed) for the top six significant genes as ranked by p-value from the Wald test. There is also a line at zero (dotted) for reference. Three genes require two FPCs while the other three genes only use one FPC.</td>
<td>36</td>
</tr>
<tr>
<td>Figure 4.1</td>
<td>Power curves for the results in Table 4.2 for IEF (1) for an exponential AFT model under the exponential data generation scheme when ( n = 250 ).</td>
<td>52</td>
</tr>
<tr>
<td>Figure 4.2</td>
<td>Pointwise (dashed) and simultaneous (dotted) confidence intervals for one simulation of IEF (1), ( \delta(s) = 0.4[\sqrt{2}\cos(2\pi s)] ), when ( n = 500 ) for each of the three models under the exponential data generation scheme. The top graph fits an exponential AFT model, the middle graph fits a lognormal AFT model, and the bottom graph fits a Cox model.</td>
<td>53</td>
</tr>
<tr>
<td>Figure 5.1</td>
<td>The empirical mean and 95% confidence interval of the IE for both a FGAM fit and a FLM fit for the two nonlinear functions we investigate for a Gaussian response when ( n = 500 ) for varying levels of ( \kappa ).</td>
<td>66</td>
</tr>
<tr>
<td>Figure 5.2</td>
<td>The empirical mean and 95% confidence interval of the IE for both a FGAM fit and a FLM fit for the two linear functions we investigate for a Gaussian response when ( n = 500 ) for varying levels of ( \kappa ).</td>
<td>67</td>
</tr>
<tr>
<td>Figure 5.3</td>
<td>Standardized estimates of the IE for the 11 genes given in Table 5.3 from a FGAM fit for varying levels of ( \kappa ).</td>
<td>70</td>
</tr>
<tr>
<td>Figure 5.4</td>
<td>Standardized estimates of the IE for the six genes given in Table 5.4 from a FLM fit for varying levels of ( \kappa ).</td>
<td>71</td>
</tr>
</tbody>
</table>
1.1 Literature Review

1.1.1 Overview

Mediation analysis is typically used to examine the various pathways among primary and intermediate variables and help identify factors that affect the relationship between a main covariate and a response. The notion of mediation is simple - that a third, intermediate variable transmits the effect of one variable to another. However, the investigation and analysis of the relationships among these variables is complex. Mediation analysis is mainly used in one of two ways (MacKinnon, 2008). The first way uses mediation to address how and why two things are related by considering mediating variables that help explain the relationship. Researchers seek to explain the processes underlying the observed relation between an independent variable and a dependent variable. The second approach to mediation attempts to manipulate mediating variables that are causally related to the dependent variable in an effort to change the dependent
variable. The first approach is more common in basic explanatory research to analyze an observed relationship, whereas the second method is an experimental approach used in applied research.

We consider a simple mediation model with one mediator $M$, one covariate $X$, and one response $Y$ (see Figure 1.1). The theories and methods of mediation analysis under this three variable framework assume the model involves only these three variables and no other outside factors affect the model (MacKinnon, 2008). Mediation analysis investigates the intermediate effect of the mediator on the relationship between the covariate and the response. Mediation occurs when the covariate affects the mediator, which in turn affects the response. The effect of $X$ on $Y$ (pathway 1 in Figure 1.1) is called the direct effect, whereas the relationship between $X$ and $Y$ that is mediated by $M$ (pathway 4) is called the indirect effect (IE) and is the effect in which we are specifically interested. The IE describes how the relationship between the covariate and the response is affected by the mediator. Specifically, it studies how the response changes due to a change in the mediating variable when the covariate remains the same.

An important distinction of mediation from other statistical models is the direction of the relationships among the variables. Consider three variables $X$, $Y$, and $Z$. There are a wide variety of ways and directions in which the three variables can be related, e.g. $X$ to $Z$, $Z$ to $X$, $Y$, etc. (MacKinnon, 2008). In mediation, the third variable $Z$ is intermediate in the chain relating $X$ and $Y$, meaning $X$ affects $Z$, which then affects $Y$. Therefore, the third variable is called a mediating variable and is often denoted by $M$. Note this is different than a confounding variable. A confounder is a third variable that changes
the relationship between $X$ and $Y$ because it affects both $X$ and $Y$.

Mediation analysis with all scalar variables is well-studied and implemented frequently, especially in the fields of psychology, agriculture, and medicine, among others. One of the earliest examples of mediation studies how an organism mediates the relation of a stimulus to a response, known as the stimulus-organism-response (S-O-R) model (MacKinnon, 2008). Mediation mechanisms are what determines how an organism responds to a stimulus. For example, a stimulus may trigger a memory identifying the stimulus as a threat, leading to an avoidance response, or a stimulus may trigger an attraction that leads to an approach response (MacKinnon, 2008).

Mediation analysis is often used in clinical psychology for the development and evaluation of treatment programs. For example, Huey Jr. et al. (2000) study the effect of their treatment program ($X$) on delinquent behavior ($Y$) and try to identify possible mediators that affect the effectiveness of the program, such as association with deviant peers ($M$). They want to determine if their treatment program is effective in deterring future delinquent behavior, i.e. there is a direct effect between their treatment program and delinquent behavior. However, they find that affiliation with delinquent peers mediated the effectiveness of their treatment program. This shows that there is an indirect effect of association with deviant peers on the effectiveness of the treatment program.

In agricultural studies, mediation analysis is also commonly used. Variables such as amount of fertilizer and insecticide ($X$) are hypothesized to be related to both germination ($M$) and crop yield ($Y$) (Smith, 1957). Researchers study the direct effect of fertilizer and insecticide on crop yield and the indirect effect of germination on this relationship. In addition, Cochran and Cox (1957) describe an experiment in which different types of fumigation ($X$) are used to reduce eelworms ($M$), which then increases oat yield ($Y$). Here, they are primarily concerned with the indirect effect of eelworms.

There are also several examples of mediation models in medicine and epidemiology. For example, Susser (1973) describes the theory that maternal diet ($X$) affects maternal weight ($M$) that in turn affects birth weight ($Y$). Mediation analysis is also used to study how mediators influence the effect of a risk factor for a disease ($X$) on disease occurrence ($Y$). Ikram and VanderWeele (2015) discuss how gastric acid reflux ($X$) is a risk factor for esophagus cancer ($Y$). Reflux esophagitis ($M$), which is the inflammation
and necrosis of esophageal epithelial cells, is caused by gastric acid reflux and can affect the disease pathway. They also detail how high radiation exposure ($X$) can lead to disease of the reproductive glands ($M$), which can lead to infertility ($Y$).

We extend scalar mediation analysis to functional mediation analysis to account for functional data. We investigate performing functional mediation analysis with a scalar mediating variable $M$, a scalar response variable $Y$, and a functional covariate $X(\cdot)$, meaning the covariate is a continuous curve rather than a single scalar measure. Since functional mediation analysis involves the use of functional regression methods, we review necessary functional regression methods in Chapter 2. First, however, in this rest of this chapter, we review the main concepts of mediation analysis.

### 1.1.2 Standard Mediation Versus Causal Mediation Using a Counterfactual Model

There are two main schools of thought in mediation analysis: a standard approach and a causal approach using a counterfactual model (Pearl, 2012). The beginnings of standard mediation analysis stem from the concepts developed by Baron and Kenny (1986) and Judd and Kenny (1981). They use a series of tests to establish conditions for and test the presence of mediation. Referencing Figure 1.1, the four regression equations coinciding with the four pathways when all of the variables are continuous are

\[
Y = \alpha_1 + \beta_1 X + \epsilon_1;
\]
\[
M = \alpha_2 + \beta_2 X + \epsilon_2;
\]
\[
Y = \alpha_3 + \beta_3 M + \epsilon_3;
\]
\[
Y = \alpha_4 + \beta_4 X + \beta_M M + \epsilon_4.
\]

Baron and Kenny (1986) first test $\beta_1$ and $\beta_2$ to determine if significant relationships exist among the variables. If both are significant, then $\beta_M$ is tested in order to establish if mediation exists. If it is significant and mediation is found to exist, then $\beta_4$ is tested to determine whether the mediation is full or partial. If $\beta_4$ is not significant, full mediation is supported, whereas if it is significant, then only partial mediation is said to exist. This method of a series of stepwise tests only tests whether mediation exists
and does not quantify the IE in which we are specifically interested. In addition, while many scholars
generally accept the notion that $\beta_2$ and $\beta_M$ should not equal zero in order to establish mediation, they
criticize the thought that $\beta_1$ needs to be significant (MacKinnon, 2008). The relationship between $X$ and
$Y$ does not need to be statistically significant for $M$ to be a mediator. The reason for this is that the effect
of $X$ on $Y$ may be zero or close to zero when the direct effect and indirect effect, as described later in this
section, have opposite signs. This phenomenon is sometimes called inconsistent mediation (VanderWeele,
2015). This stepwise method to mediation is also problematic in practice as the likelihood of establishing
mediation from testing four coefficients for significance can sometimes be very small, even if mediation
is present.

Therefore, there are two other main methods in standard mediation analysis, and they are named after
the manner in which the IE is quantified: the product of coefficients (POC) method and the difference
of coefficients (DOC) method. The POC method is based on path analysis and calculates the IE as
the product of the effect of $X$ on $M$ and the effect of $M$ on $Y$ when $X$ is in the model. Mathematically
speaking,

$$IE = \beta_2 \beta_M.$$  

The DOC method quantifies the IE as the difference between the effect of $X$ on $Y$ and the effect of $X$ on
$Y$ when $M$ is included in the model; effectively, what is the difference of including the mediator? The
DOC method calculates the IE as

$$IE = \beta_1 - \beta_4.$$  

The DOC method is used more commonly in epidemiology, whereas the POC method is primarily used in
the social sciences. In the case of a continuous response, the two methods, POC and DOC, are equivalent;
see details in MacKinnon et al. (1995). However, for a generalized response, these two methods may
not be equivalent. MacKinnon et al. (2007) note that in the logistic setting with a binary response, the
two methods of the IE are not equivalent. The POC method is a better choice as it is more stable than
the DOC method; it produces an asymptotically unbiased estimate whereas the DOC method requires
additional scaling in order to reduce bias (MacKinnon et al., 2007).
Unfortunately, the standard approaches to mediation only suffice under certain conditions. They don’t handle non-continuous data well; they don’t allow for interactions or nonlinearities in the data to be modeled; and they don’t allow for a causal interpretation of the model. Therefore, a causal approach using a counterfactual model is necessary in these cases. The counterfactual approach to mediation considers potential outcomes in order to decompose the total effect of \( X \) on \( Y \) into a direct effect and an indirect effect and to allow for a causal interpretation of the model. Since it is more structured than the standard approach, it can also account for a covariate-mediator interaction and nonlinearities in the data. The counterfactual approach requires the assumption of several fairly strong conditions regarding no unmeasured confounding among the relationships of the covariate, mediator, and response (VanderWeele, 2011). These assumptions are imperative in order to have a causal interpretation of the model. In a counterfactual model, we denote \( Y_x \) and \( M_x \), respectively, to be the counterfactual values of the response and mediator if \( X \) had been set, possibly contrary to fact, to \( x \). Let \( Y_{xm} \) be the counterfactual value of the response that would have been observed if \( X \) and \( M \) had been set, possibly contrary to fact, to \( x \) and \( m \), respectively. We also consider nested counterfactual responses such as \( Y_{xM_x} \), which is the response if the covariate had been set to \( x \) and the mediator had been set to the level it would have taken had \( X \) been \( x \). On the other hand, \( Y_{xM_x^*} \) is the response if \( X \) had been set to \( x \) and the mediator had been set to the level it would have taken had \( X \) been \( x^* \). We assume composition, meaning \( Y_x = Y_{xM_x} \). A counterfactual approach to mediation analysis is often preferred when the covariate \( X \) takes a discrete value, such as a treatment or level of exposure. When the covariate is a treatment variable, researchers can only impose one treatment on a subject, but in order to determine the treatments causal effect, they must consider the potential outcome given a different treatment. Therefore, it is also known as a potential outcomes model.

The total causal effect of the covariate on the response can be decomposed into the indirect effect (IE) and the direct effect (DE) as shown below:

\[
E(Y_x) - E(Y_{x^*}) = E(Y_{xM_x}) - E(Y_{x^*M_{x^*}})
= E(Y_{xM_x}) - E(Y_{xM_{x^*}}) + E(Y_{xM_{x^*}}) - E(Y_{x^*M_{x^*}})
= E(Y_{xM_x} - Y_{xM_{x^*}}) + E(Y_{xM_{x^*}} - Y_{x^*M_{x^*}}),
\]
where \( E(Y_{xM_i} - Y_{xM_i^*}) \) is the IE and \( E(Y_{xM_i^*} - Y_{x'M_i^*}) \) is the DE (VanderWeele and Vansteelandt, 2010). Here, the IE is the expected change in \( Y \) when the covariate is held constant at \( X = x \) and the mediator is changed from \( M_i \) to the value it would have taken had \( X \) been set to \( x^* \), \( M_i^* \). On the other hand, the DE is the expected change in \( Y \) when \( X \) is changed from \( X = x \) to \( X = x^* \) and the mediator remains constant. We focus only on the IE of the mediator:

\[
IE = E(Y_{xM_i} - Y_{xM_i^*}).
\] (1.1)

The standard POC method and the counterfactual method will coincide in some settings. For instance, for linear models, they will coincide when there is no covariate-mediator interaction. Assuming the four mediation equations given previously in this section when all of the variables are continuous, the counterfactual definition of the IE is

\[
IE = E(Y_{xM_i} - Y_{xM_i^*})
\]

\[
= E(\alpha_4 + \beta_4x + \beta_M M_i + \varepsilon_4) - E(\alpha_4 + \beta_4x + \beta_M M_{i^*} + \varepsilon_4)
\]

\[
= \beta_M E(M_i - M_{i^*})
\]

\[
= \beta_M (\alpha_2 + \beta_2 x - \alpha_2 - \beta_2 x^*)
\]

\[
= \beta_M \beta_2 (x - x^*).
\]

This is equivalent to the POC method with the addition of \((x - x^*)\), which are known quantities. The counterfactual approach to mediation is greatly preferable, however, when the data are not continuous, such as binary responses or survival times, and it is necessary when including an interaction term or fitting a nonlinear model. The counterfactual approach also adds the benefit of having a causal interpretation.

### 1.1.3 Testing the Indirect Effect

There are many tests in mediation analysis. MacKinnon et al. (2002) compare 14 testing procedures in the standard mediation framework in an extensive simulation study. The 14 tests are partitioned into three
categories: (1) those based on a series of stepwise tests, (2) those based on the difference of coefficients (DOC) method, and (3) those based on the product of coefficients (POC) method. The stepwise tests are inferior because they only establish conditions for mediation and do not provide an estimate or test of the IE. In addition, they often have low Type 1 error and power due to the use of multiple hypothesis tests in order to test all of the steps. The methods based on the DOC method often underestimate the true standard error or have low Type 1 error (MacKinnon et al., 2002).

The most commonly used POC test is a $Z$-test based on the large-sample variance formula derived by Sobel (1982) for the first-order variance approximation of the product of two coefficients based on first derivatives using the multivariate delta method. MacKinnon (2008) describes the formula for calculating Sobel’s approximation of the variance of the quantity $ab$ as $\sigma_{ab}^2 = b^2 \sigma_a^2 + a^2 \sigma_b^2$, where $\sigma_a^2$ and $\sigma_b^2$ are the variances of $a$ and $b$, respectively. In simulations in MacKinnon et al. (2002), this test has low Type 1 error but adequate power. They note that while the two coefficients separately are asymptotically normal, the product of two normal quantities is not guaranteed to be normal; unfortunately, the POC $Z$-test assumes the POC has a normal distribution when it may not, leading to possibly low Type 1 error and power. Therefore, they suggest variants to the typical product of coefficients tests. They propose a test based on the empirical distribution of the test statistic and a test based on the distribution of the product of two normal variables. These tests have better properties when both coefficients equal zero or both do not equal zero, but they have unacceptably high Type 1 error when one coefficient equals zero and the other does not. Lastly, Bollen and Stine (1990) and Shrout and Bolger (2002) recommend using a bootstrap procedure to test the IE as it does not require any distributional assumptions. Shrout and Bolger (2002) use the bootstrap percentiles of the distribution of the POC to create asymmetric confidence intervals and test whether zero is included in the interval. They find this bootstrap method gives more accurate results compared to the conventional tests that assume normality when the bootstrap distribution is distinctly skewed. Under a counterfactual model, the most commonly used test is also a large sample $Z$-test based on Sobel’s first-order variance approximation (VanderWeele, 2011). Other methods include bootstrap procedures, such as those in Shrout and Bolger (2002) and Tchetgen Tchetgen (2011).
CHAPTER 2

INTRODUCTION TO FUNCTIONAL REGRESSION METHODS

2.1 Literature Review

2.1.1 Functional Linear Models

Functional covariates arise when multiple measurements are taken on a subject; these observations can be viewed as discrete realizations of some unknown smooth, underlying continuous function. Specifically, for subject \( i = 1, \ldots, n \), the covariate is a real function \( X_i(\cdot) \) defined on a closed interval \( S \). The curves \( X_i(s) \) for all \( s \in S \) are assumed to be square-integrable over \( S \), i.e. \( X_i(s) \in L^2[S] \). Further, the covariate function \( X_i(\cdot) \) is often measured with error and observed on a grid of points, meaning one often observes \( \{W_i1, \ldots, W_iJ_i\} \) corresponding to possibly irregular and sparse grid points \( \{s_{i1}, \ldots, s_{iJ_i}\} \) in \( S \), where \( J_i \) corresponds to the number of points observed for subject \( i \). Therefore, \( W_i(s_{ij}) = X_i(s_{ij}) + e_i(s_{ij}) \), where \( X_i(\cdot) \) is the true underlying process that generates the realizations \( W_i(\cdot) \) and \( e_i(\cdot) \) is random mean zero noise with variance \( \sigma_X^2 \), where \( \sigma_X^2 \) determines the amount of covariate measurement error.
Often, one wants to quantify the relationship between a functional covariate and a scalar response. The most common model for scalar on function regression is a functional linear model (FLM) (Ramsay and Dalzell, 1991). A FLM quantifies the effect of a functional covariate by the inner product between the covariate and an unknown functional parameter. A standard FLM for the case of one functional covariate and a scalar, continuous response is

$$Y_i = \alpha_0 + \int_{\mathcal{S}} X_i(s) \alpha_1(s) ds + \epsilon_i, \quad i = 1, \ldots, n,$$

(2.1)

where \( \epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2) \), \( \alpha_0 \) is an unknown intercept, and \( \alpha_1(\cdot) \) is an unknown coefficient function that weights the functional covariate across the domain \( \mathcal{S} \) to emphasize portions of the curve functionally contributing to the model to predict \( Y_i \).

In order to produce an interpretable and uniquely identifiable estimator for \( \alpha_1(\cdot) \), one must use some method of regularization (Ramsay and Silverman, 2005). There are two common methods of regularization in FLMs, leading to two main classes of estimators. The first method performs regularization though dimension reduction by projecting the functional data onto a finite sequence of functions of \( L^2[\mathcal{S}] \). The most popular method in this class is functional principal component regression (FPCR). The second main type of estimators involves expanding \( \alpha_1(\cdot) \) in some basis of functions of \( L^2[\mathcal{S}] \) and minimizing a penalized least squares criterion.

**Functional Principal Component Regression**

FPCR uses functional principal component analysis (FPCA) to capture the main modes of variation in the data and to reduce the dimensionality of the function. Assume that \( X_i(\cdot) \) is a real valued smooth process defined over \( \mathcal{S} \) with true mean function \( \mu(\cdot) \). FPCA is a stepwise procedure based on finding functional probes of maximum variance. The corresponding linear probes are called functional principal components (FPCs). The first step involves finding the weight function \( \phi_1(\cdot) \) such that \( \xi_{i1} = \int_{\mathcal{S}} \phi_1(s) \{X_i(s) - \mu(s)\} ds \) has the largest variance subject to \( \|\phi_1\|^2 = 1 \). The second step finds the weight function \( \phi_2(\cdot) \) such that \( \xi_{i2} = \int_{\mathcal{S}} \phi_2(s) \{X_i(s) - \mu(s)\} ds \) has the largest variance subject to \( \|\phi_2\|^2 = 1 \) and \( \phi_2 \) is orthogonal to \( \phi_1 \).
This process is repeated until a sufficient number of weight functions has been chosen. Each successive weight function must be orthogonal to all previous weight functions and will capture successively smaller modes of variation in the curves.

Finding the FPCs of a curve is the same as finding the eigenfunctions of its covariance function (Ramsay and Silverman, 2005). Denote the continuous covariance function of \( X_i(\cdot) \) by \( V(s,u) = \text{Cov}\{X_i(s),X_i(u)\} \). By Mercer’s theorem, there exists an orthonormal set of continuous eigenfunctions in \( L^2[\mathcal{S}] \), \( \{\phi_k(\cdot),k \geq 1\} \), and a non-increasing set of eigenvalues, \( \lambda_1 \geq \lambda_2 \geq \ldots \geq 0 \), such that \( V(s,u) = \sum_{k=1}^{\infty} \lambda_k \phi_k(s) \phi_k(u) \), for all \( s,u \in \mathcal{S} \) (Ramsay and Silverman, 2005). Using the basis functions determined by the eigenfunctions of \( V(s,u) \) and the Karhunen-Loève theorem, the random process \( X_i(\cdot) \) can be represented by \( X_i(s) = \mu(s) + \sum_{k=1}^{\infty} \xi_{ik} \phi_k(s) \), where \( \phi_k(s) \) is the \( k^{th} \) eigenfunction of \( V(s,u) \) and \( \xi_{ik} = \int_{\mathcal{S}} X_i(s) \phi_k(s) ds \) is the random FPC score with \( E(\xi_{ik}) = 0 \) and \( \text{Var}(\xi_{ik}) = \lambda_k \). In practice, the infinite sum is truncated such that \( X_i(s) \approx \mu(s) + \sum_{k=1}^{K_n} \xi_{ik} \phi_k(s) \), where \( K_n \) is the truncation point that determines the closeness of the approximation. There are several methods to choose the truncation point \( K_n \). We choose \( K_n \) by selecting the minimum number of eigenfunctions that account for a pre-specified percentage of variation explained (PVE) by the truncated expansion. Specifically, \( \text{PVE} = \frac{\sum_{k=1}^{K_n} \lambda_k}{\sum_{k=1}^{\infty} \lambda_k} \), where \( \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{K_n} \) are the ordered eigenvalues that correspond to the eigenfunctions.

Recall that often we do not observe \( X_i(\cdot) \); instead, we observe \( W_i(s_{ij}) = X_i(s_{ij}) + e_i(s_{ij}) \), where \( e_i(\cdot) \) is a white noise process with variance \( \sigma_X^2 \), meaning the observations are measured with error on a possibly sparse grid. In order to account for such measurement error and sparsity in the observed functional data, we must use smoothed FPCA. There are several smoothing approaches in the FPCA literature (Staniswalis and Lee, 1998; Yao et al., 2005; Di et al., 2009; Goldsmith et al., 2013; Xiao et al., 2016). We adapt the one developed by Xiao et al. (2016) referred to as fast covariance estimation (FACE), which is a fast implementation of the sandwich smoother (Xiao et al., 2013) for covariance matrix smoothing. The sandwich smoother proposed by Xiao et al. (2013) is a penalized spline method for bivariate smoothing that uses a tensor product structure to allow for fast computations. FPCA using FACE is implemented using the ‘fpca.face’ function from the ‘refund’ package in R (Crainiceanu et al., 2016).

Once we have performed FPCA, we can use the results to perform FPCR. The eigenfunctions
from FPCA can be used to expand the coefficient function in the FLM in equation (2.1) as
\[ \alpha_1(s) \approx \sum_{k=1}^{K_n} \alpha_{1k} \phi_k(s), \]
where \( \{ \alpha_{1k}, k = 1, \ldots, K_n \} \) are unknown scalar basis coefficients. A key assumption in
the above step is that the unknown regression coefficient \( \alpha_1(\cdot) \) resides in the same function space as the
covariate and can be represented using the same basis functions. The model in equation (2.1) can be
written as the approximate model
\[ Y_i = \alpha_0 + \sum_{k=1}^{K_n} \hat{\xi}_{ik} \alpha_{1k} + \epsilon_i, \quad i = 1, \ldots, n. \]

This is a simple linear regression model with \( \hat{\xi}_{ik} = (\hat{\xi}_{i1}, \ldots, \hat{\xi}_{iK_n})^T \) as random covariates and \( \alpha_{K_n} = (\alpha_{11}, \ldots, \alpha_{1K_n})^T \) as unknown regression coefficients. The unknown parameters can be estimated using
standard least squares estimation. The regression function can be reconstructed by
\[ \hat{\alpha}_1(s) = \sum_{k=1}^{K_n} \hat{\alpha}_{1k} \hat{\phi}_k(s), \]
where \( \{ \hat{\phi}_k(\cdot), k = 1, \ldots, K_n \} \) are the first \( K_n \) estimated eigenfunctions from FPCA.

### Penalized Least Squares

The second class of estimators uses restricted basis functions and allows for the expansion of \( X_i(\cdot) \)
and \( \alpha_1(\cdot) \) to be with different basis systems, \( \{ \psi_k(\cdot), k \geq 1 \} \) and \( \{ \theta_l(\cdot), l \geq 1 \} \), respectively. Therefore,
\[ X_i(s) = \sum_{k=1}^{K_n} c_{ik} \psi_k(s) \quad \text{and} \quad \alpha_1(s) = \sum_{l=1}^{L_n} \alpha_{1l} \theta_l(s), \]
where \( c_{ik} \) are covariate basis coefficients and \( \alpha_{1l} \) are
unknown coefficients (Ramsay and Silverman, 2005). An issue with this approach is as the number of
basis functions used to estimate \( \alpha_1(\cdot) \), \( L_n \), increases, the roughness of the estimated function also
increases. Therefore, in order to control the roughness of the estimate and obtain an interpretable estimate,
several methods adapt a roughness penalty approach, which requires the minimization of a penalized
least squares criterion.

Let the roughness of the function \( \alpha_1(\cdot) \) be measured by
\[ R\{ \alpha_1(s) \} = \int_{\cal x} (L \alpha_1(s))^2 ds, \]
where \( L \) is a linear differential operator (Ramsay and Silverman, 2005). For example, one might use the second
derivative to look at departure from linearity. The operator \( L \) is chosen according to the specific type
of problem. In order to account for this penalty, it is included in the sum of squares that needs to be
minimized with respect to \( \{\alpha_0, \alpha_1(\cdot)\} \):

\[
\sum_{i=1}^{n} \left\{ Y_i - \alpha_0 - \int_{\mathcal{Y}} X_i(s) \alpha_1(s) ds \right\}^2 + \lambda \int_{\mathcal{Y}} \{L\alpha_1(s)\}^2 ds,
\]

where \( \lambda \) is a penalty parameter (Ramsay and Silverman, 2005). If the penalty parameter \( \lambda \) is small, the estimate of \( \alpha_1(\cdot) \) can be rough as not much emphasis is put on the penalty term. However, a large penalty parameter results in a smoother, flatter estimate as more weight is placed on the penalty term.

Note that the basis systems used to expand \( X_i(\cdot) \) and \( \alpha_1(\cdot) \) may not be orthogonal as in FPCR. Therefore, \( \int_{\mathcal{Y}} X_i(s) \alpha_1(s) ds \) can be written as

\[
\sum_{k=1}^{K_n} \sum_{l=1}^{L_n} c_{ik} \left[ \int_{\mathcal{Y}} \psi_k(s) \theta_l(s) ds \right] \alpha_{il} = c_i^T J \alpha,
\]

where \( c_i = (c_{i1}, \ldots, c_{iK_n})^T \), \( \alpha = (\alpha_{11}, \ldots, \alpha_{1L_n})^T \) and \( J \) is a \( K_n \times L_n \) matrix with \( (k,l) \)-th element \( J_{kl} = \int_{\mathcal{Y}} \psi_k(s) \theta_l(s) ds \) (Ramsay and Silverman, 2005). The penalty term \( L\alpha_1(s) \) can be written as \( \sum_{l=1}^{L_n} \alpha_{il} \{L\theta_l(s)\} \), meaning \( \int_{\mathcal{Y}} \{L\alpha_1(s)\}^2 ds \) can be written as

\[
\sum_{l=1}^{L_n} \sum_{v=1}^{L_n} \alpha_{il} \left[ \int_{\mathcal{Y}} \{L\theta_l(s)\} \{L\theta_v(s)\} ds \right] \alpha_{iv} = \alpha^T R \alpha,
\]

where \( R \) is a \( L_n \times L_n \) matrix with \( (l,v) \)-th element \( R_{lv} = \int_{\mathcal{Y}} \{L\theta_l(s)\} \{L\theta_v(s)\} ds \). Therefore, the penalized sum of squares can be written as (Ramsay and Silverman, 2005)

\[
\sum_{i=1}^{n} \{Y_i - \alpha_0 - c_i^T J \alpha \}^2 + \lambda \alpha^T R \alpha.
\] (2.2)

This equation is a least squares problem with a penalty term such that given a fixed \( \lambda \), \( \alpha_0 \) and \( \alpha \) can be computed. The choice of the penalty term \( \lambda \) can be determined in many different ways. A popular method is to use generalized cross validation (GCV), which requires the minimization, with respect to \( \lambda \), of a criterion involving the sum of the squared differences between the true responses and the predicted responses (Cardot et al., 2003).
2.1.2 Generalized Functional Linear Models

A natural extension of a FLM is a generalized functional linear model (GFLM) where $Y_i$ can be a non-continuous response variable that has discrete positive values. GFLMs allow for discrete responses with binomial, multinomial or Poisson distributions, leading to functional versions of logistic, binomial or Poisson regression. A GFLM is written as

$$Y_i = g \left( \alpha_0 + \int \mathcal{S} X_i(s) \alpha_1(s) ds \right) + \varepsilon_i, \quad i = 1, \ldots, n,$$

where $g(\cdot)$ is a link function and is a monotone and twice continuously differentiable function and $\varepsilon_i$ are independent and identically distributed random variables such that $E(\varepsilon_i|X_i(s), s \in \mathcal{S}) = 0$. Some common examples of link functions are the identity link $g(\mu) = \mu$, which is used for Gaussian data and results in the FLR model in equation (2.1), the logit link $g(\mu) = \log\{\mu/(1-\mu)\}$, which is used for binary data and results in a logistic model, and the log link $g(\mu) = \log(\mu)$, which is used for count data and results in a Poisson model. Both classes of estimators described in the previous section can be implemented in GFLMs as well.

2.1.3 Nonlinear Functional Generalized Additive Models

A limitation of the popular FLM is that the assumed linear relationship between the functional covariate and response may be too restrictive. Therefore, functional regression models that model nonlinear functional relationships are desirable in some settings. Functional generalized additive models (FGAMs) allow the effect of the covariate to be modeled nonlinearly (McLean et al., 2014). Compared to a FLM, a FGAM provides greater flexibility as it does not make the strong assumption of linearity between the functional covariate and the functional coefficient. Just as a FLM is the natural extension of a linear model to functional data, a FGAM is the natural extension of a generalized additive model (GAM) to functional data.

The present literature on using nonparametric, additive models for scalar on function regression models is minimal. Müller and Yao (2012) consider a GAM that uses a finite number of principal
components as linear functionals of the predictor curves as covariates. James and Silverman (2005) study a similar approach using projection pursuit to search for linear functionals. While not an additive model, Ferraty and Vieu (2006) propose nonparametric kernel based approaches for scalar on function regression models based on conditional expectation, median, and mode. Compared to earlier nonparametric, additive functional models, FGAMs regress on the functional covariate directly (McLean et al., 2014). Early, standard additive models focus on smoothing splines and backfitting (see Hastie and Tibshirani (1990)). More recently, penalized regression splines have been used, most often using penalized iteratively reweighted least squares (P-IRLS) with smoothing parameters chosen by generalized cross validation (GCV) for estimation purposes (see Marx and Eilers (1998) and Wood (2006)). This is the approach McLean et al. adopt to estimate the FGAM. Overall, additive models allow for increased flexibility and potentially lower estimation bias than linear models while also having less variance in estimation and being less susceptible to the curse of dimensionality than models that make no additivity assumptions (McLean et al., 2014).

Consider a simple FGAM with a scalar, continuous response and a functional covariate (McLean et al., 2014)

\[ Y_i = \alpha_0 + \int F\{X_i(s), s\} ds + \epsilon_i, \quad i = 1, \ldots, n, \]

where \( \epsilon_i \sim \text{iid} \ N(0, \sigma^2_\epsilon) \), \( \alpha_0 \) is an unknown intercept, and \( F(\cdot, \cdot) \) is an unknown smooth bivariate function. Notice this method incorporates the functional covariate directly. A special case occurs when \( F(x, s) = X_i(s) \beta(s) \) and the resulting model is a FLM. As described in McLean et al. (2014), the derivation of a FGAM is as follows. Consider an additive model of the form \( E\{Y_i|X_i(s_1), \ldots, X_i(s_J)\} = \alpha_0 + \sum_{j=1}^J f_j\{X_i(s_j)\} \), where the \( f_j \)'s are unspecified smooth functions. The model can be written as \( E\{Y_i|X_i(s_1), \ldots, X_i(s_J)\} = \alpha_0 + J^{-1} \sum_{j=1}^J F\{X_i(s_j), s_j\} \) using a Reimann sum approximation. Let \( J \to \infty \), and the resulting model is the FGAM above.

To overcome the curse of dimensionality, both the \( x \) and \( s \) components of \( F(\cdot, \cdot) \) are smoothed. A bivariate spline model is used for \( F(\cdot, \cdot) \) such that

\[ Y_i = \alpha_0 + \int F\{X_i(s), s\} ds + \epsilon_i, \quad i = 1, \ldots, n, \]
\[ F(x, s) = \sum_{j=1}^{K_x} \sum_{k=1}^{K_s} \theta_{jk} B^X_j(x) B^S_k(s), \]

where \( \{B^X_j(x) : j = 1, \ldots, K_x\} \) and \( \{B^S_k(x) : k = 1, \ldots, K_s\} \) are spline bases on \( \mathcal{S} \). McLean et al.’s FGAM procedure can incorporate several types of bases and penalties to estimate the regression function \( F(\cdot, \cdot) \). They use tensor product P-splines with a second order difference penalty. Using this expansion, we can write the FGAM as the following linear model

\[ E(Y_i | X_i) = \alpha_0 + \int_{\mathcal{S}} F\{X_i(s), s\} ds = \alpha_0 + \sum_{j=1}^{K_x} \sum_{k=1}^{K_s} \theta_{jk} Z_{jk}(i), \]

where \( Z_{jk}(i) = \int_{\mathcal{S}} B^X_j(X_i(s)) B^S_k(s) ds \). Each \( Z_{jk}(i) \) can be approximated by Simpson’s rule. For specific details on the penalized iteratively reweighted least squares (P-IRLS) estimation procedure of FGAMs, see McLean et al. (2014). For identifiability, we require the condition \( \sum_{i=1}^{n} \int_{\mathcal{S}} F\{X_i(s), s\} ds = 0 \) (McLean et al., 2014). We note that even with this constraint, the model is not fully identifiable. To see this, consider any function \( g(\cdot) \) such that \( \int g(s) ds = 0 \), and define \( F^*\{X(\cdot), \cdot\} = F\{X(\cdot), \cdot\} + g(\cdot) \). Then we have \( \int F^*\{X(s), s\} ds = \int F\{X(s), s\} ds \). Thus, \( F(\cdot, \cdot) \) is identifiable only up to functions intergrating to zero.

Depending on the number of splines used for each axis, there could be a particular tensor product of splines that has no observed data on its support, causing \( Z_{jk} = 0 \) and leading to a rank deficient design matrix. To remedy this, McLean et al. (2014) use a pointwise quantile transformation of the functional covariate using the empirical cumulative distribution function (cdf) to ensure each tensor product has observed data on its support. The transformation certifies that for any \( s \), the transformed points will lie uniformly between \([0, 1]\). Other transformations, including a center and scale transformation, may also be used. Parameter estimates can still be found when the design matrix has a column of zeros due to the penalization used when fitting a FGAM. However, McLean et al. (2014) devise that transforming the covariate will improve the numerical and statistical stability of the estimates. It is also important to note
that the FGAM is invariant to transformations of the covariate, unlike the FLM (McLean et al., 2014).

2.2 Combining Mediation Analysis and Functional Regression Methods

In the following three chapters, we integrate the theories of mediation analysis and functional regression methods to investigate the IE of a scalar mediator on the relationship between a functional covariate and a scalar response. We first consider a Gaussian response (or generalized response) in a FLM (or GFLM) using FPCR in Chapter 3. We use the standard approach to mediation analysis and define the indirect effect function (IEF) using the POC method. We propose three testing procedures of the IEF and compare their results. In Chapter 4, we use a counterfactual approach to mediation to analyze the IEF when the response is time-to-event data. We examine the accelerated failure time (AFT) model and Cox proportional hazards model in the framework of FPCR in order to estimate and test the presence of an IE. Lastly, in Chapter 5, we consider a nonlinear model in a counterfactual framework where we allow the functional covariate to be modeled nonlinearly. We use a FGAM and bivariate spline basis system in order to model the covariate in a nonlinear fashion.
CHAPTER 3

STANDARD FUNCTIONAL MEDIATION ANALYSIS USING PRINCIPAL COMPONENTS FOR GAUSSIAN AND GENERALIZED RESPONSES

3.1 Model and Framework

3.1.1 Product of Coefficients Method for the Indirect Effect Function

The classical mediation model is pictured in Figure 1.1; this mediation model can account for a functional covariate. For \( i = 1, \ldots, n \), the data consist of a real valued scalar response \( Y_i \), a scalar mediating variable \( M_i \), and a real function \( X_i(\cdot) \). For simplicity, we initially only consider a Gaussian response. We apply the standard mediation approach and use the POC method to define the IE. The POC method requires two
In equation (3.1), $M_i$ is regressed on $X_i(s)$; here, $\gamma_0$ is an unknown intercept, $\gamma_1(s)$ is the unknown coefficient function of $X_i(s)$ on the mediator $M_i$, and $\eta_i \overset{iid}{\sim} N(0, \sigma^2_\eta)$. In equation (3.2), $Y_i$ is regressed on both the scalar mediator and the functional covariate. Similarly, $\beta_0$ is an unknown intercept, $\beta_1(s)$ is the unknown coefficient function quantifying the effect of the functional covariate on the response when the mediating variable is included in the model, $\beta_M$ is the scalar partial relationship between the mediating variable and the response, and $\zeta_i \overset{iid}{\sim} N(0, \sigma^2_\zeta)$.

We use FPCR to model the coefficient functions in equation (3.1) and equation (3.2). Recall from Section 2.1.1, FPCR uses FPCA to highlight the most important modes of variation in the data and to reduce the dimensionality of the function. The eigenfunctions from FPCA can be used to expand the coefficient function in equation (3.1) as $\gamma_1(s) \approx \sum_{k=1}^{K_n} \hat{\gamma}_{1k} \phi_k(s)$, where $\{\gamma_{1k}, k = 1, \ldots, K_n\}$ are unknown scalar basis coefficients. The model in equation (3.1) then can be approximated by

$$M_i = \gamma_0 + \sum_{k=1}^{K_n} \hat{\gamma}_{1k} \zeta_{ik} \gamma_{1k} + \eta_i, \quad i = 1, \ldots, n,$$

where $\hat{\zeta}_{ik} = \int_{\mathcal{X}} X_i(s) \hat{\phi}_k(s) ds$ are the estimated FPC scores. The regression function can be reconstructed by $\hat{\gamma}_1(s) = \sum_{k=1}^{K_n} \hat{\gamma}_{1k} \hat{\phi}_k(s)$. Similarly, the coefficient function in equation (3.2) can be represented by $\beta_1(s) \approx \sum_{k=1}^{K_n} \beta_{1k} \phi_k(s)$. The model in equation (3.2) can be approximated by

$$Y_i = \beta_0 + \sum_{k=1}^{K_n} \hat{\zeta}_{ik} \beta_{1k} + \beta_M M_i + \zeta_i, \quad i = 1, \ldots, n.$$

Then $\beta_1(s)$ can be reconstructed using $\hat{\beta}_1(s) = \sum_{k=1}^{K_n} \hat{\beta}_{1k} \hat{\phi}_k(s)$.

We extend the POC definition of the IE to the functional case and define the indirect effect function
(IEF) as \( \delta(s) = \beta_M \gamma_1(s) \). Its estimate is

\[
\hat{\delta}(s) = \sum_{k=1}^{K} \hat{\beta}_M \hat{\gamma}_k \hat{\phi}_k(s) = \sum_{k=1}^{K} \hat{\delta}_k \hat{\phi}_k(s),
\]

where \( \hat{\delta}_k = \hat{\beta}_M \hat{\gamma}_k \). In order to obtain a variance estimate of the IEF, we extend Sobel’s first-order variance approximation of the product of two coefficients to the IE vector \( (\delta_1, \ldots, \delta_{1K_n})^T = \delta_{K_n} = \beta_M \gamma_{K_n} \), where \( \gamma_{K_n} = (\gamma_1, \ldots, \gamma_{K_n})^T \). We define the approximate variance as \( \Sigma_\delta = \beta_M^2 \Sigma_Y + \sigma_{\beta_M}^2 \gamma_{K_n} \gamma_{K_n}^T \), where \( \Sigma_Y \) is the covariance matrix of \( \gamma_{K_n} \) and \( \sigma_{\beta_M}^2 \) is the variance of \( \beta_M \). Using this definition of the variance of \( \delta_{K_n} \) and the estimated eigenfunctions from FPCA, \( \Phi(\cdot) \), we estimate the variance of the IEF as \( \hat{\Sigma}_\delta(s) = \hat{\Phi}(s) \hat{\Sigma}_\delta \hat{\Phi}^T(s) \). We find pointwise confidence intervals of the IEF using this estimated variance.

We also calculate simultaneous confidence intervals for the IEF using a method proposed by Crainiceanu et al. (2012). Given the estimate of the IEF, \( \hat{\delta}(s) = \sum_{k=1}^{K} \hat{\beta}_M \hat{\gamma}_k \hat{\phi}_k(s) \), and its variance estimate based on Sobel’s first-order approximation, \( \hat{\Sigma}_\delta(s) = \hat{\Phi}(s) \hat{\Sigma}_\delta \hat{\Phi}^T(s) \), we simulate \( d_u(s) \) from a multivariate normal distribution, \( \text{MVN}\{\hat{\delta}(s), \hat{\Sigma}_\delta(s)\} \). We calculate \( x_u = \max_s \{|d_u(s) - \hat{\delta}(s)|/\hat{\sigma}(s)\} \), where \( \hat{\sigma}(s) \) is the \( s \)-th diagonal element of \( \hat{\Sigma}_\delta(s) \). We repeat this process \( u = 1, \ldots, U \) times and obtain \( q_{1-\alpha} \), the \( (1 - \alpha) \times 100\% \) empirical quantile of the sample \( \{x_u : u = 1, \ldots, U\} \). We calculate the joint confidence intervals using \( \hat{\delta}(s) \pm q_{1-\alpha} \hat{\sigma}(s) \). Crainiceanu et al. (2012) note that with enough samples, the sampling variability in \( \hat{\delta}(s) \) and \( \hat{\Sigma}_\delta(s) \) can be ignored.

### 3.1.2 Testing the Indirect Effect Function

We wish to test if the mediator has an effect on the relationship between the functional covariate and the response by testing if the IEF differs from zero. The hypothesis in which we are interested is \( H_0 : \delta(\cdot) = 0 \), which is equivalent to testing \( H_0 : \delta_{K_n} = 0 \). First, it is important to note that since \( \delta_{K_n} = \beta_M \gamma_{K_n} \), there are three ways in which \( \delta_{K_n} \) can equal zero: (1) \( \beta_M = 0 \) with \( \gamma_{K_n} \neq 0 \), (2) \( \gamma_{K_n} = 0 \) with \( \beta_M \neq 0 \), and (3) \( \beta_M = \gamma_{K_n} = 0 \). The complexity of this hypothesis can lead to flawed statistical properties of its test, hence why testing in mediation analysis is still problematic. However, for mediation to exist, \( \beta_M \) should not equal zero; otherwise the mediator has no effect on \( Y \) when \( X \) is in the model. This is also noted by Baron.
and Kenny (1986) and generally accepted by many scholars. Therefore, we focus on case (2).

Another issue is the unknown distribution of $\delta_{K_n}$. Since $\hat{\beta}_M$ and $\hat{\gamma}_{K_n}$ are estimated coefficients from a typical linear model involving ordinary least squares, they are asymptotically normal by the law of large numbers and central limit theorem:

$$\sqrt{n}(\hat{\beta}_M - \beta_M) \overset{d}{\rightarrow} N(0, \sigma_{\hat{\beta}_M}^2);$$

$$\sqrt{n}(\hat{\gamma}_{K_n} - \gamma_{K_n}) \overset{d}{\rightarrow} N(0, \Sigma).$$

However, the product of two normal random variables, $\delta_{K_n} = \hat{\beta}_M \hat{\gamma}_{K_n}$, is not guaranteed to be normal under all conditions; in fact, it is often asymmetric with high kurtosis (MacKinnon et al., 2002). Shrout and Bolger (2002) further say that the product of normal variables with positive means will tend to have a positive skew while the product of normal variables with means of opposite signs often has a negative skew. However, Aroian (1947) shows that the product of two normal quantities is asymptotically normal if either component (here, $\beta_M$ or $\gamma_{K_n}$) is large. Often, researchers testing for mediation expect this to be the case; however, it may not be. This can cause issues with any test that assumes the product is normal if the skew or kurtosis is large.

We examine three testing procedures of the IEF. First, we examine a Wald test based on Sobel’s variance approximation:

$$W_n = \delta_{K_n} \hat{\Sigma}_{\delta}^{-1} \hat{\delta}_{K_n} \sim \chi^2_d, \quad d = \text{trace}(\hat{\Sigma}_{\delta}) = K_n,$$

where $\hat{\Sigma}_{\delta}$ is the variance of $\delta_{K_n}$ based on the first-order Sobel approximation of the variance of the product of two coefficients. Therefore, the test statistic $W_n$ is compared to a $\chi^2$-distribution with $d$ degrees of freedom in this one-sided alternative. This Wald test assumes the product of coefficients tends to be normally distributed as the sample size gets large.

The second test we consider is a bootstrap method based on the procedure described by Shrout and Bolger (2002) that uses the bootstrap percentiles of the distribution of the product of coefficients to
create asymmetric confidence intervals to test the IE. A benefit of the bootstrap method is it does not require any distributional assumptions. After performing FPCA on the observed data, the data are in the form $[Y_i, (\hat{\xi}_{i1}, \ldots, \hat{\xi}_{iK_n}), M_i], i = 1, \ldots, n$. Using this dataset, we create a bootstrap sample of size $n$ by randomly sampling observations with replacement. We estimate the IE vector for each bootstrap sample, $\hat{\delta}_{K_n}^{(b)} = \hat{\beta}_M^{(b)} \gamma_{K_n}^{(b)}$. We repeat this process $b = 1, \ldots, B$ times. Note in classical mediation analysis the IE is scalar, while here it is a vector of length $K_n$. Therefore, we must correct for $K_n$ multiple tests. We find the $[\alpha/(2K_n)] \times 100\%$ and $[1 - \alpha/(2K_n)] \times 100\%$ percentiles of the distribution for each of the $K_n$ estimates and determine if zero is contained in all of the $K_n$ intervals. If zero is outside at least one of the $K_n$ intervals, we reject the null hypothesis. We note that performing the bootstrap procedure on the estimated FPC scores may lead to biased results as we do not account for the variation of the FPCA decomposition. However, bootstrapping outside the FPCA loop can cause issues with the decomposition since observations may be repeated several times.

Lastly, we use simultaneous confidence intervals for the IEF to test the hypothesis based on whether zero is included in the interval. Using the method described earlier in Section 3.1.1, the joint confidence intervals are defined by $\hat{\delta}(s) \pm q_{1 - \alpha} \hat{\sigma}(s)$. If zero is outside the joint interval at one or more of the $J$ time points, we reject the null hypothesis. This method also assumes the distribution of the product of coefficients tends to be normally distributed for large samples.

### 3.1.3 Modeling a Generalized Response Using a Generalized Functional Linear Model

We can extend the above concepts to a generalized response using a GFLM, detailed in Section 2.1.2. Specifically, we consider a binary response and a logit link, leading to a functional logistic model. Following the POC method, the two equations needed are given below:

logit $\Pr(Y_i = 1|X_i(s), M_i) = \beta_0 + \int_{\mathcal{S}} X_i(s) \beta_1(s) ds + \beta_M M_i$.

$M_i = \gamma_0 + \int_{\mathcal{S}} X_i(s) \gamma_1(s) ds + \eta_i.$

22
The POC definition of the IEF is the same as before, \( \delta(s) = \beta_M \gamma_i(s) \). We use FPCR to model the above equations and obtain an estimate. We also examine the same three testing procedures as we do in the Gaussian response case; they are described in Section 3.1.2.

3.2 Simulation Study

3.2.1 Design and Data Generation

The purpose of this simulation study is to investigate the estimation and testing of the IEF in standard functional mediation analysis in two cases: with a Gaussian response and with a binary response. First, we consider the Gaussian response case. Consider \( S \) to be the interval \([0, 1]\) and the functions are observed fully at \( J \) equally spaced points on the interval, where \( J \) increases as sample size \( n \) increases. We generate the functional covariate with four trigonometric basis functions as

\[
X_i(s) = a_{1i} \sqrt{2} \sin(2\pi s) + a_{2i} \sqrt{2} \cos(2\pi s) + a_{3i} \sqrt{2} \sin(4\pi s) + a_{4i} \sqrt{2} \cos(4\pi s),
\]

where \( s = \{s_1, s_2, \ldots, s_J\} \) and \( a_{1i}, a_{2i}, a_{3i}, \) and \( a_{4i} \) are normally distributed, independent random variables with mean zero and respective variances 16, 8, 4, and 2. The covariate is often observed with error, i.e. \( W_i(s) \) is observed where

\[
W_i(s) = X_i(s) + e_i(s) \quad \text{and} \quad e_i \sim N(0, \sigma^2_X).
\]

The value of \( \sigma^2_X \) determines the amount of covariate measurement error. We set \( \sigma^2_X = 1 \), which corresponds with moderate measurement error. We investigate two IEFs:

1. \( \delta(s) = 0.4 \sqrt{2} \cos(2\pi s) \) and
2. \( \delta(s) = 0.4 \sqrt{2} \sin(2\pi s) + \sqrt{2} \cos(2\pi s) \).

IEF (2) is more complex and has more curvature than IEF (1).

We set the sample size \( n \) and the number of grid points \( J \) at two different combinations. As \( n \) increases, we increase \( J \) at a similar rate, leading to the following two combinations of \( n \) and \( J \): \( n = 100 \) with \( J = 81 \) and \( n = 250 \) with \( J = 201 \).

In our simulation study, the intercepts \( \gamma_0 \) and \( \alpha_0 \) are set to zero, and the error variances \( \sigma^2_\eta \) and \( \sigma^2_\zeta \) are set to one. The scalar mediating variable is generated from equation (3.1) using

\[
M_i = A_i g + \eta_i, \quad \text{where} \quad \eta_i \sim N(0, \sigma^2_\eta), \quad A_i = (a_{1i}, a_{2i}, a_{3i}, a_{4i}) \quad \text{and} \quad g^T \text{ equals one of the following vectors corresponding to the two IEFs above:}
\]

1. \( g^T = (0, 0.2, 0, 0) \) or
2. \( g^T = (0.2, 0.2, 0, 0) \).

These choices of \( g \) are used to regulate the amount of correlation between the covariate and the mediator. Under this model, the correlation between them is approximately 50% for IEF (1) and 70% for IEF (2).
The scalar response variable is generated from equation (3.2) using $Y_i = A_i b + \beta M_i + \zeta_i$, where $\zeta_i \overset{iid}{\sim} N(0, \sigma^2_\zeta)$, $A_i$ is the same as above, $\beta M = 2$, and $b^T$ is one of the following vectors corresponding to the two IEFs: (1) $b^T = (0, 3, 0, 0)$ or (2) $b^T = (3, 3, 0, 0)$. The choice of $b$ is arbitrary and does not affect the calculation of the IEF in which we are interested. The first IEF is calculated by multiplying $\beta M = 2$ by $\gamma_1(s) = g^T \Phi(s) = 0.2\sqrt{2}\cos(2\pi s)$, leading to $\delta(s) = 0.4\sqrt{2}\cos(2\pi s)$. The other IEF is found similarly. The number of simulations for each combination of IEF and sample size is $N = 1,000$. Also, the truncation number in FPCA, $K_n$, is defined as the minimum number of FPCs needed to account for 95% of the variability of the data.

In the case of a binary response variable, the set-up is very similar but with a few small changes. In order to keep $P(Y_i = 1)$ away from its boundaries of zero and one, we reduce the variance of the FPC scores. We generate $X_i(\cdot)$ such that $a_{i1}, a_{i2}, a_{i3}$, and $a_{i4}$ are normally distributed, independent random variables with mean zero and respective variances 1.0, 0.5, 0.25, and 0.125. In addition, we increase the sample size and the corresponding number of observed time points of the functional covariate. We use the following two combinations of $n$ and $J$: $n = 250$ with $J = 201$ and $n = 500$ with $J = 401$. The binary scalar response variable is generated such that $P(Y_i = 1) = e^{A_i b + \beta M_i} / (1 + e^{A_i b + \beta M_i})$, where $A_i$, $b$, and $\beta M$ are previously defined. Under the binary response model, the correlation between the functional covariate and the mediator is approximately 15% for IEF (1) and 25% for IEF (2).

### 3.2.2 Estimation Results

In order to analyze the results of the estimation of the IEF, we examine integrated squared bias (ISB), integrated mean squared error (IMSE), and integrated coverage (IC) for $\delta(\cdot)$. The estimates of these three criteria are below:

\[
\hat{\text{ISB}}(\delta) = \frac{1}{J} \sum_{j=1}^{J} \left\{ \frac{1}{N} \sum_{k=1}^{N} \left( \delta^{(k)}(s_j) - \delta(s_j) \right) \right\}^2 ;
\]

\[
\hat{\text{IMSE}}(\delta) = \frac{1}{J} \sum_{j=1}^{J} \frac{1}{N} \sum_{k=1}^{N} \left( \delta^{(k)}(s_j) - \delta(s_j) \right)^2 ;
\]

\[
\hat{\text{IC}}(\delta) = \frac{1}{J} \sum_{j=1}^{J} \frac{1}{N} \sum_{k=1}^{N} \left[ \delta(s_j) \in \{ \delta^{(k)}(s_j) \pm Z_{\alpha/2} \sqrt{\hat{\Sigma}_{\delta(s_j)}} \} \right] .
\]
Table 3.1: ISB, IMSE, and IC for $\delta(\cdot)$ for IEF (1) and IEF (2). Coverages are given in terms of percent. The maximum standard error for the estimates in this table is 0.003.

<table>
<thead>
<tr>
<th></th>
<th>$n$</th>
<th>ISB</th>
<th>IMSE</th>
<th>95% IC</th>
<th>90% IC</th>
<th>85% IC</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gaussian Response</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IEF (1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>0.000</td>
<td>0.041</td>
<td>94.2</td>
<td>89.1</td>
<td>84.2</td>
<td></td>
</tr>
<tr>
<td>250</td>
<td>0.000</td>
<td>0.016</td>
<td>95.1</td>
<td>90.3</td>
<td>85.3</td>
<td></td>
</tr>
<tr>
<td>IEF (2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>0.000</td>
<td>0.042</td>
<td>94.3</td>
<td>89.1</td>
<td>84.1</td>
<td></td>
</tr>
<tr>
<td>250</td>
<td>0.000</td>
<td>0.016</td>
<td>95.2</td>
<td>90.2</td>
<td>85.2</td>
<td></td>
</tr>
<tr>
<td><strong>Binary Response</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IEF (1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>250</td>
<td>0.001</td>
<td>0.306</td>
<td>95.8</td>
<td>90.9</td>
<td>85.6</td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>0.000</td>
<td>0.137</td>
<td>95.3</td>
<td>90.0</td>
<td>84.7</td>
<td></td>
</tr>
<tr>
<td>IEF (2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>250</td>
<td>0.001</td>
<td>0.317</td>
<td>95.9</td>
<td>90.9</td>
<td>85.7</td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>0.000</td>
<td>0.145</td>
<td>95.1</td>
<td>89.8</td>
<td>84.7</td>
<td></td>
</tr>
</tbody>
</table>

In the above equations, $\hat{\delta}^{(k)}(\cdot)$ is the estimate of $\delta(\cdot)$ for simulation $k$, $J$ is the number of time points at which the curves are evaluated, $N$ is the number of simulations, $I$ is an indicator variable equaling one if $\delta(\cdot)$ is in the specified interval and zero if it is not, $Z_{\alpha/2}$ is the normal Z-score for a significance level of $\alpha$, and $\hat{\Sigma}_{\delta(\cdot)}$ is the estimate of the variance of the IEF based on Sobel’s first-order approximation. We calculate 95%, 90%, and 85% confidence intervals using normal Z-scores of 1.96, 1.645, and 1.44, respectively, which assume the distribution of $\delta_{K_n}$ is normal. We examine the normality assumption and determine the distribution of $\hat{\delta}_{K_n}$ is close enough to normal in our simulation to permit using Z-scores. This may not always be true, however.

Table 3.1 describes the estimation results of the simulation study for both a Gaussian and a binary response. It shows the ISB, IMSE, and IC for $\delta(\cdot)$ for the two IEFs detailed previously at each level of sample size. For each sample size and IEF, the estimate is unbiased and has good coverage. As sample size increases, IMSE decreases as expected. When the response is binary, IMSE is much higher in all cases compared to the Gaussian response results. Overall, the estimation results are good for the two IEFs we investigate regardless of sample size or type of response variable.
### 3.2.3 Testing Results

In order to compare the results of the three procedures described earlier for testing the IEF, we investigate Type 1 error and power using a significance level of \( \alpha = 0.05 \). As we note previously for the hypothesis \( H_0 : \delta_{K_n} = 0 \), there are three ways in which \( \delta_{K_n} = \beta_M \gamma_{K_n} \) can equal zero: (1) \( \beta_M = 0 \) with \( \gamma_{K_n} \neq 0 \), (2) \( \gamma_{K_n} = 0 \) with \( \beta_M \neq 0 \), and (3) \( \beta_M = \gamma_{K_n} = 0 \). However, since \( \beta_M \) should not equal zero for mediation to exist, we focus on case (2). Therefore, we set \( \beta_M \) equal to a constant and vary \( \gamma_{K_n} \) to examine Type 1 error and power. For the Gaussian response case, we set \( \beta_M = 1 \) and 2, and for each case of \( \beta_M \), we vary \( \gamma_{K_n} = 0, 0.05, 0.1, 0.15, \) and 0.2. For the binary response case, we still set \( \beta_M = 1 \) and 2, but we vary \( \gamma_{K_n} = 0, 0.1, 0.2, 0.3, 0.4, \) and 0.5. We also look at larger levels of \( \beta_M \) up to \( \beta_M = 50 \) and find similar results as those described below. For the bootstrap test, we perform the bootstrap procedure \( B = 200 \) times. Similarly, for the simultaneous interval test, we perform the process \( U = 200 \) times.

First, we discuss the Type 1 error results for the bootstrap based test. For all of the cases we consider, Type 1 error is highly inflated, ranging from 0.110 to 0.119. This is over double the nominal amount of \( \alpha = 0.05 \). Therefore, we do not suggest using the bootstrap test as it does not control Type 1 error. This may be a result of bootstrapping on the estimated FPC scores and losing a source of variation. Therefore, we focus on the Wald test and simultaneous interval test. Table 3.2 displays the Type 1 error and power results corresponding to the Wald test and simultaneous interval test for IEF (1). The top table refers to the case when the response is Gaussian and \( n = 100 \). The bottom table refers to the case when the response is binary and \( n = 250 \). Figure 3.1 and Figure 3.2 depict the power curves for Table 3.2 for a Gaussian response and a binary response, respectively. These graphs help illustrate the differences between the Wald test and simultaneous interval test, as well as the changes within a test as \( \beta_M \) changes. The results for other combinations of sample size and IEF are similar and therefore, are not shown for convenience.

We consider the Gaussian response case first. Both the Wald test using Sobel’s first-order approximation of the variance of the product of two coefficients and the simultaneous interval test have Type 1 error close to the nominal amount of \( \alpha = 0.05 \). Type 1 error remains around \( \alpha = 0.05 \) even when we increase \( \beta_M \) to higher levels. The Wald test and the simultaneous interval test also have good power,
although the power of the Wald test is higher. Keeping $\beta_M$ constant, the power increases as $\gamma_{K_n}$ increases. Figure 3.3 depicts pointwise and simultaneous confidence intervals for one simulation of IEF (1) for a Gaussian response when $n = 250$ for the cases when the IEF equals $\delta(s) = 0$ and when it equals $\delta(s) = 0.4[\sqrt{2}\cos(2\pi s)]$. In both cases, the simultaneous interval is wider than the pointwise interval as expected. The results are similar when the response is binary. However, Type 1 error and power are much lower in the binary case, especially when the sample size is small. Figure 3.4 depicts pointwise and simultaneous confidence intervals for one simulation of IEF (1) when the response is binary and $n = 250$. Again, the simultaneous interval is wider than the pointwise one. Note the intervals in the binary case are wider than in the Gaussian case due to higher variance estimates.

For our null hypothesis $H_0 : \delta_{K_n} = 0$, while we focus on the case when $\beta_M \neq 0$, we acknowledge there are two other cases when $\beta_M = 0$. For the Wald test and simultaneous interval test when $\beta_M = 0$, Type 1 error is extremely conservative. Type 1 error is also conservative for the bootstrap test, although it is slightly higher. Therefore, we cannot believe any test to be proper when $\beta_M = 0$. In previous studies with all scalar variables, such as the one in MacKinnon et al. (2002), many tests of the IE have cases where Type 1 error is very conservative. Using our simulation estimates, we investigate the distribution of each component of the IE vector $\delta_{K_n}$ in several cases. We note that when $\beta_M \neq 0$, the distributions are relatively normal, even when $\gamma_{K_n} = 0$, which is an assumption of the Wald test and simultaneous interval test. However, when $\beta_M = 0$, the distributions are highly concentrated around zero with little to no variation. Similarly, we explore the bootstrap distributions of $\hat{\delta}_{K_n}$ in several situations. Again, when $\beta_M = 0$, the bootstrap distributions are highly concentrated around zero, although they have more variation than the empirical distributions of $\delta_{K_n}$. When $\beta_M \neq 0$, the bootstrap distributions are relatively normal, just like their empirical counterparts. These results may help explain the conservative Type 1 error when $\beta_M = 0$ and why the assumption of normality for the Wald test and simultaneous interval test isn’t an issue when $\beta_M \neq 0$.

Overall, when $\beta_M \neq 0$, the Wald test and the simultaneous interval test have better Type 1 error rates than the bootstrap test. In addition, the Wald test has higher power than the simultaneous interval test. Therefore, we recommend using the Wald test.
Table 3.2: Type 1 error and power results for the Wald test and simultaneous interval test for IEF (1) for a significance level of $\alpha = 0.05$. The top table refers to the case when the response variable is Gaussian and $n = 100$. The bottom table refers to the case when the response variable is binary and $n = 250$.

<table>
<thead>
<tr>
<th>$\beta_M$</th>
<th>Gaussian response</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma_K_n$</td>
<td>Wald Test</td>
<td>Interval Test</td>
<td></td>
</tr>
<tr>
<td>0.00</td>
<td>0.039</td>
<td>0.042</td>
<td></td>
</tr>
<tr>
<td>0.05</td>
<td>0.139</td>
<td>0.107</td>
<td></td>
</tr>
<tr>
<td>0.10</td>
<td>0.500</td>
<td>0.302</td>
<td></td>
</tr>
<tr>
<td>0.15</td>
<td>0.903</td>
<td>0.600</td>
<td></td>
</tr>
<tr>
<td>0.20</td>
<td>0.994</td>
<td>0.882</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\beta_M$</th>
<th>Binary response</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma_K_n$</td>
<td>Wald Test</td>
<td>Interval Test</td>
<td></td>
</tr>
<tr>
<td>0.00</td>
<td>0.007</td>
<td>0.014</td>
<td></td>
</tr>
<tr>
<td>0.1</td>
<td>0.020</td>
<td>0.029</td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>0.116</td>
<td>0.080</td>
<td></td>
</tr>
<tr>
<td>0.3</td>
<td>0.374</td>
<td>0.179</td>
<td></td>
</tr>
<tr>
<td>0.4</td>
<td>0.710</td>
<td>0.349</td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td>0.892</td>
<td>0.593</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\beta_M$</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma_K_n$</td>
<td>Wald Test</td>
<td>Interval Test</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.00</td>
<td>0.019</td>
<td>0.031</td>
<td></td>
</tr>
<tr>
<td>0.1</td>
<td>0.066</td>
<td>0.060</td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>0.247</td>
<td>0.148</td>
<td></td>
</tr>
<tr>
<td>0.3</td>
<td>0.607</td>
<td>0.312</td>
<td></td>
</tr>
<tr>
<td>0.4</td>
<td>0.905</td>
<td>0.547</td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td>0.994</td>
<td>0.786</td>
<td></td>
</tr>
</tbody>
</table>
Figure 3.1: Power curves for the results in Table 3.2 for IEF (1) when the response variable is Gaussian and \( n = 100 \).

Figure 3.2: Power curves for the results in Table 3.2 for IEF (1) when the response variable is binary and \( n = 250 \).
Figure 3.3: Pointwise (dashed) and simultaneous (dotted) confidence intervals for one simulation of IEF (1) when \( n = 250 \) and the response variable is Gaussian. In the top graph, \( \beta_M = 2 \) and \( \gamma_K = (0,0,0,0)^T \) so the IEF is \( \delta(s) = 0 \). In the bottom graph, \( \beta_M = 2 \) and \( \gamma_K = (0,0.2,0,0)^T \) so the IEF is \( \delta(s) = 0.4[\sqrt{2}\cos(2\pi s)] \).
Figure 3.4: Pointwise (dashed) and simultaneous (dotted) confidence intervals for one simulation of IEF (1) when \( n = 250 \) and the response variable is binary. In the top graph, \( \beta_M = 2 \) and \( \gamma_K \equiv (0,0,0,0)^T \) so the IEF is \( \delta(s) = 0 \). In the bottom graph, \( \beta_M = 2 \) and \( \gamma_K \equiv (0,0.2,0,0)^T \) so the IEF is \( \delta(s) = 0.4[\sqrt{2}\cos(2\pi s)] \).
3.3 Application to Multiple Myeloma and Copy Number Variation Data

In order to demonstrate the methods presented in this chapter, we apply them to a dataset from the Multiple Myeloma Research Consortium. Genetic abnormalities in the number of copies of DNA, known as copy number variations (CNVs), are associated with the development and progression of cancer. Multiple myeloma (MM) is a blood cancer that causes cancerous plasma cells to accumulate in the bone marrow. The data consist of over 17,000 gene expression and copy number profiles for 235 individuals with MM. The CNV profile for each patient consists of approximately 244,000 serially correlated measurements along the genome, leading to it being modeled as a functional covariate. The data also include measurements of $\beta_2$ microglobulin ($\beta_2M$), a protein that is widely recognized as a clinical biomarker of MM and is related to abnormalities such as CNVs. We apply functional mediation analysis to investigate the IE of gene expression (the mediator) on the relationship between functional CNV data and the response $\beta_2M$. We also adjust for the effect of two demographic covariates, age and gender, which does not change the estimation or testing procedure. We note that this application is for demonstrative purposes only.

Because we model CNV as a functional covariate, any gene that has fewer than 35 associated CNV probes is dropped from the analysis, leaving 627 genes. In addition, only 172 patients have a non-missing $\beta_2M$ measurement. Thus, for each of the 172 subjects, we have 627 gene expressions and each gene expression has a corresponding CNV profile with at least 35 measurements. We denote $\beta_2M$ as $Y_i$, the response, and gene expression as $M_i$, the mediator. We assume that $X_i(\cdot)$ is the random process that produces the observed CNV profiles that we model as the functional covariate.

We apply our proposed FPCA estimation procedure and the corresponding Wald test to all 627 genes to determine which genes have a significant IEF. Using the Wald test, 11 genes have a significant IEF at an unadjusted significance level of $\alpha = .05$. We do not correct for multiple testing because doing so results in zero significant genes due to the very large number of tests we perform. We do not recommend ignoring multiple testing in practice. Table 3.3 displays relevant summary information for the 11 genes that have a significant IEF, including gene name, the chromosome on which the gene is located, the
p-value from the Wald test, the number of FPCs needed to account for 99% of the variability, the number of CNV probes associated with the gene, gene function, and known associated diseases. The majority of the information in the table comes from the National Center for Biotechnology Information (NCBI: http://www.ncbi.nlm.nih.gov/) and the references given within. Referencing Table 3.3, out of the 11 significant genes, eight are known to be associated with various forms of cancer and other diseases, mainly brain diseases.

In addition, five genes require two FPCs in order to account for 99% of the variability in the CNV profiles associated with those genes while the other six genes only need one FPC. The genes that only require one FPC suggest the functional aspect of the model is not needed in those cases. Averaging over all of the measurements in the CNV profile retains the necessary data as all probes contribute the same amount of information to the mediated effect. Even though the functional nature of our model is not needed for these six genes, we do not lose any information by modeling them functionally. On the other hand, we gain detail about the IEF for the five genes that require two FPCs by modeling them functionally. For instance, for the genes WDR70 and BRE, we learn that the probes between locations 10 and 30 contribute more to the mediated effect than the other probes. Furthermore, for gene PFDN6, the probes near the beginning of the gene affect the IE more than the rest of the probes. This can be seen in Figure 3.5, which illustrates the estimated IEF and its pointwise confidence interval for the top six significant genes as ranked by p-value from the Wald test. The graphs also include a dotted line at zero to better illustrate where the IEF does not equal zero. In the figure of the top six genes, three genes require two FPCs and the other three only use one FPC. The graphs of the other five significant genes look similar to the ones presented here.
Table 3.3: Summary information for the 11 genes that have a significant IEF using the Wald test, including gene name, the chromosome on which the gene is located, the p-value from the Wald test, the number of FPCs needed to account for 99% of the variability, the number of CNV probes associated with the gene, gene function, and known associated diseases. Information is from the NCBI (http://www.ncbi.nlm.nih.gov/).

<table>
<thead>
<tr>
<th>Gene</th>
<th>Gene Name</th>
<th>Chromosome</th>
<th>P-value</th>
<th>Number of FPCs</th>
<th>Number of CNV Probes</th>
</tr>
</thead>
<tbody>
<tr>
<td>DAB1</td>
<td>disabled homolog 1</td>
<td>1</td>
<td>0.00086</td>
<td>1</td>
<td>159</td>
</tr>
<tr>
<td>PFDN6</td>
<td>prefoldin subunit 6</td>
<td>6</td>
<td>0.00625</td>
<td>2</td>
<td>46</td>
</tr>
<tr>
<td>RNF115</td>
<td>ring finger protein 115</td>
<td>1</td>
<td>0.00717</td>
<td>1</td>
<td>37</td>
</tr>
<tr>
<td>GMDS</td>
<td>GDP-mannose 4, 6-dehydratase</td>
<td>6</td>
<td>0.01223</td>
<td>1</td>
<td>68</td>
</tr>
<tr>
<td>WDR70</td>
<td>WD repeat domain 70</td>
<td>5</td>
<td>0.01230</td>
<td>2</td>
<td>40</td>
</tr>
<tr>
<td>BRE</td>
<td>brain and reproductive organ-expressed</td>
<td>2</td>
<td>0.01398</td>
<td>2</td>
<td>49</td>
</tr>
<tr>
<td>XPR1</td>
<td>xenotropic and polytropic retrovirus receptor 1</td>
<td>1</td>
<td>0.01691</td>
<td>1</td>
<td>38</td>
</tr>
<tr>
<td>POLR3C</td>
<td>polymerase (RNA) III (DNA directed) polypeptide C</td>
<td>1</td>
<td>0.01773</td>
<td>1</td>
<td>40</td>
</tr>
<tr>
<td>PDSS2</td>
<td>prenyl (decaprenyl) diphosphate synthase, subunit 2</td>
<td>6</td>
<td>0.01887</td>
<td>2</td>
<td>42</td>
</tr>
<tr>
<td>PIAS3</td>
<td>protein inhibitor of activated STAT, 3</td>
<td>1</td>
<td>0.02447</td>
<td>1</td>
<td>43</td>
</tr>
<tr>
<td>FARS2</td>
<td>phenylalanyl-tRNA synthetase 2, mitochondrial</td>
<td>6</td>
<td>0.02480</td>
<td>2</td>
<td>55</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gene</th>
<th>Gene Function</th>
<th>Associated Diseases</th>
</tr>
</thead>
<tbody>
<tr>
<td>DAB1</td>
<td>regulates Reelin signaling (key in brain development)</td>
<td>brain malformations; mental retardation; autism; Alzheimer’s</td>
</tr>
<tr>
<td>PFDN6</td>
<td>pertains to unfolded protein and chaperone binding</td>
<td>unknown</td>
</tr>
<tr>
<td>RNF115</td>
<td>controls ubiquitin liagase activity</td>
<td>breast cancer</td>
</tr>
<tr>
<td>GMDS</td>
<td>produces enzymes for metabolism</td>
<td>colorectal cancer; leukocyte adhesion deficiency (LAD)</td>
</tr>
<tr>
<td>WDR70</td>
<td>protein-coding gene</td>
<td>unknown</td>
</tr>
<tr>
<td>BRE</td>
<td>promotes tumor growth</td>
<td>liver cancer</td>
</tr>
<tr>
<td>XPR1</td>
<td>receptor for murine leukemia viruses</td>
<td>prostate cancer</td>
</tr>
<tr>
<td>POLR3C</td>
<td>controls DNA directed RNA polymerase activity</td>
<td>unknown</td>
</tr>
<tr>
<td>PDSS2</td>
<td>tumor suppressor that relates to coenzyme Q</td>
<td>lung cancer; coenzyme Q10 deficiency</td>
</tr>
<tr>
<td>PIAS3</td>
<td>pertinent to transcription</td>
<td>lung cancer; brain tumors; thrombocytopenia with absent radius (TAR syndrome)</td>
</tr>
<tr>
<td>FARS2</td>
<td>charges tRNA localized to the mitochondrion</td>
<td>Alpers’ syndrome (degeneration of grey matter, causing dementia and seizures)</td>
</tr>
</tbody>
</table>
3.4 Summary

In this chapter, we integrate the theories of mediation analysis and functional regression methods and develop a FPCR based method to estimate and test the IEF under a standard functional mediation framework with Gaussian and generalized responses. Through numerical simulations, we show our method properly estimates the IEF while the results of the three proposed testing procedures vary. We suggest using the Wald test based on Sobel’s variance approximation when $\beta_M \neq 0$ as it has adequate Type 1 error and good power. However, when $\beta_M = 0$, the Wald test and all other proposed tests have extremely conservative Type 1 error. In addition, we apply our methodology to the MM data to investigate the IE of gene expression on the relationship between functional CNV data and $\beta_2 M$.

The broad purpose of many genetic studies is to be able to identify genes that have a significant effect on diseases in order to gain knowledge about the development and progression of those diseases. This study tries to show that the mediated effect of gene expression may be important when modeling the relationship between CNV data and a response relating to a disease. In addition, we propose that modeling CNV functionally can help pinpoint which locations within a gene are deserving of further investigation. The results presented here are for demonstrative purposes and do not provide much evidence to support our hypotheses. However, these concepts are still noteworthy for future studies.
Figure 3.5: Graphs of the IEF estimate (solid) and its pointwise confidence interval (dashed) for the top six significant genes as ranked by p-value from the Wald test. There is also a line at zero (dotted) for reference. Three genes require two FPCs while the other three genes only use one FPC.
CHAPTER 4

CAUSAL FUNCTIONAL MEDIATION
ANALYSIS USING PRINCIPAL
COMPONENTS FOR TIME-TO-EVENT
DATA

4.1 Model and Framework for Time-to-Event Data Using a Counterfactual Approach

For a functional mediation model with a right-censored time-to-event outcome, the data consist of many components. For \( i = 1, \ldots, n \), let \( T_i \) be the survival (or failure) time and \( C_i \) be the censoring time for the \( i^{th} \) subject. We observe the bivariate vector \((Y_i, \Delta_i)\), where \( Y_i = \min(T_i, C_i) \) and \( \Delta_i = 1 \) if \( Y_i = T_i \) and \( \Delta_i = 0 \) otherwise. The censoring times \( C_i \) are assumed to be independent of the failure times \( T_i \). We also observe
a scalar and continuous mediating variable $M_i$ and a functional covariate $X_i(\cdot)$ defined on a closed interval $\mathcal{S}$.

In this chapter, we use a counterfactual approach to mediation analysis in order to have a causal interpretation, which is highly beneficial, especially for time-to-event data. The outcome is frequently about a disease or death, and often, the length of time it takes for a disease to occur and low incidence rates of the disease make it difficult to conduct research on predictors of the disease. Therefore, studies often instead use a surrogate for the disease rather than the disease itself (MacKinnon, 2008). For example, in the study of colon cancer, the number of precancerous cells is often investigated rather than colon cancer itself. Other examples include congestive heart failure, cholesterol levels, or blood pressure as surrogates for death by cardiovascular disease and bone mineral density as a surrogate for osteoporosis. These surrogate endpoints are generally easier to study than the ultimate disease outcome since they are affected earlier than the disease outcome and the incidence of cases is larger than those for the disease. Using a surrogate endpoint requires an assumption of a causal relationship between the surrogate and the disease. The surrogate is a mediator between a predictor and the disease, so studies are designed to change the surrogate under the assumption that will change the ultimate outcome of the disease. This is one reason why causal interpretation and a counterfactual model is advantageous.

Recall the counterfactual framework presented in Section 1.1.2. The total causal effect of the covariate on the response can be decomposed into the indirect effect (IE) and the direct effect (DE):

$$E(Y_x) - E(Y_x^*) = E(Y_{xM_x} - Y_{xM_x^*}) + E(Y_{xM_x^*} - Y_{x^*M_x^*}),$$

where $E(Y_{xM_x} - Y_{xM_x^*})$ is the IE and $E(Y_{xM_x^*} - Y_{x^*M_x^*})$ is the DE (VanderWeele and Vansteelandt, 2010). Here, the IE is the expected change in $Y$ when the covariate is held constant at $X = x$ and the mediator is changed from $M_x$ to the value it would have taken had $X$ been set to $x^*$, $M_x^*$. On the other hand, the DE is the expected change in $Y$ when $X$ is changed from $X = x$ to $X = x^*$ and the mediator remains constant. We focus only on the IE of the mediator: $\text{IE} = E(Y_{xM_x} - Y_{xM_x^*})$.

In order to calculate the IE in our functional mediation model with time-to-event data, two models
are required; the first models the continuous mediating variable as a functional linear model (FLM) of the covariate $X(\cdot)$. Following Ramsay and Silverman (2005), the standard FLM in this case is

$$M_i = \gamma_0 + \int_S X_i(s) \gamma_1(s) ds + \eta_i, \quad (4.1)$$

where $\eta_i \sim N(0, \sigma^2)$, $\gamma_0$ is an unknown intercept, and $\gamma_1(\cdot)$ is an unknown coefficient function that highlights portions of the covariate functionally contributing to the model to predict the mediating variable. The second model regresses the survival time response data on both the functional covariate and the scalar mediating variable using either an AFT or Cox model. For an AFT model, the logarithm of the survival time is modeled linearly as

$$\log(T_i) = \theta_0 + \int_S X_i(s) \theta_1(s) ds + \theta_M M_i + \nu \omega_i, \quad (4.2)$$

where $\theta_0$ is an unknown intercept, $\theta_1(\cdot)$ is an unknown coefficient function relating the functional covariate and response when the mediator is in the model, $\theta_M$ is a scalar partial coefficient between the mediator and response, $\nu$ is a scale parameter, and $\omega_i$ are independently and identically distributed error variables assumed to have some known density function $f(\omega)$. For a Cox model, the hazard function is modeled as

$$\lambda(T_i | X_i(s), M_i) = \lambda_0(T_i) \exp \left[ \int_S X_i(s) \kappa_1(s) ds + \kappa_M M_i \right], \quad (4.3)$$

where $\lambda_0(\cdot)$ is an unknown baseline hazard function when all of the covariates are 0, $\kappa_1(\cdot)$ is an unknown coefficient function expressing the effect of the functional covariate on the response when the mediator is included in the model, and $\kappa_M$ is a scalar partial coefficient relating the mediator and response. Note that in equation (4.2) and equation (4.3), we assume there is no interaction between the effects of the covariate function and the mediator on the response. Unlike a standard approach, a counterfactual approach can incorporate the interaction between a mediator and covariate, but we choose to assume the simpler model.

Under a counterfactual approach to mediation analysis, the IE is defined in equation (1.1) as the expected change in the response due to a change in the mediating variable when the covariate remains
the same. Considering $Y = \log(T)$ and using equation (4.1) and equation (4.2), the derivation of the IE under an AFT model is as follows:

$$
IE = E(Y_{xM} - Y_{xM^*})
$$

$$
= E(\theta_0 + \int_{\mathcal{S}} X(s)\theta_1(s)ds + \theta_M M_x + \nu \omega) - E(\theta_0 + \int_{\mathcal{S}} X(s)\theta_1(s)ds + \theta_M M_{x^*} + \nu \omega)
$$

$$
= \theta_M E(M_x - M_{x^*})
$$

$$
= \theta_M (\gamma_0 + \int_{\mathcal{S}} X(s)\gamma_1(s)ds - \gamma_0 - \int_{\mathcal{S}} X^*(s)\gamma_1(s)ds)
$$

$$
= \int_{\mathcal{S}} \theta_M \gamma_1(s) \{X(s) - X^*(s)\} ds,
$$

The derivation of the IE under a Cox model is similar but uses the logarithm of the hazard function, which involves the ratio of the conditional density to the survival function, which involves the cumulative baseline hazard function. The derivation under a Cox model only holds when the cumulative baseline hazard function is small, i.e. the outcome is rare (VanderWeele, 2011). Otherwise, the IE does not have a simple analytic expression; this causes issues with the estimation of the IE under a Cox model. In addition, it is easy to see this counterfactual framework can implement the interaction between the covariate and the mediator; it just includes additional terms, leading to a more complicated formula for the IE. Readers should see VanderWeele (2011) for additional details.

We generalize the derivation of the IE above for either an AFT model or a Cox model as

$$
IE = \int_{\mathcal{S}} \psi_M \gamma_1(s) \{X(s) - X^*(s)\} ds,
$$

where $\psi_M = \theta_M$ for an AFT model and $\psi_M = \kappa_M$ for a Cox model. We want to investigate this IE over grid points $s \in \mathcal{S}$, so we define the indirect effect function (IEF) as

$$
IEF = \delta(s) = \psi_M \gamma_1(s).
$$

(4.4)

Therefore, the IE is equal to the integral of the IEF multiplied by $\{X(s) - X^*(s)\}$. Since the IE can
be found for any given $X(\cdot)$ and $X^*(\cdot)$, we focus on the unknown quantities $\psi_M$ and $\gamma_1(\cdot)$. We briefly compare the counterfactual derivation of the IE to that of the standard POC approach for our time-to-event model. Assuming no interaction, the only difference between the counterfactual formula and the standard POC method is the addition of $\{X(s) - X^*(s)\}$. If one includes a mediator-covariate interaction in the model, the counterfactual derivation of the IE is more complicated, as shown in VanderWeele (2011), whereas the standard POC approach is not valid. The counterfactual approach also allows for a causal interpretation whereas the standard approach does not.

4.2 Details on Time-to-Event Methods

Specifically, we consider three survival models. First, we consider two different AFT models, shown in equation (4.2), which are parametric log-linear models and assume a known error distribution $f(\omega)$. This class of log-linear models assumes the covariates act additively on the logarithm of the survival time $T$ and therefore, multiplicatively on the actual survival time (Moeschberger and Klein, 2003). We consider the exponential AFT model, which assumes $T$ is exponentially distributed, meaning the scale parameter $\nu = 1$ and assumes $f(\omega)$ is the extreme value distribution. The exponential AFT model assumes a constant hazard rate over time. While we don’t consider the Weibull AFT model, we note it is similar except the scale parameter $\nu$ is estimated rather than set equal to one and allows for a non-constant hazard rate. We also consider the lognormal AFT model; the survival times $T_i$ are assumed to come from a lognormal distribution with scale parameter $\nu$, meaning the error variables $\omega_i$ come from a standard normal distribution. AFT models use maximum likelihood to estimate the coefficients $\theta_1(\cdot)$ and $\theta_M$ in equation (4.2). Third we consider the Cox proportional hazards model, which models the hazard function in a multiplicative manner. As previously mentioned, the Cox model is semi-parametric and does not require any distributional assumptions. Partial maximum likelihood is used to estimate the coefficients of a Cox model (Kalbfleisch and Prentice, 2011).
Moeschberger and Klein (2003) describe the relationship between AFT and Cox models. They show the hazard function of an individual with survival time $T_i$ under the AFT model in equation (4.2) is

$$\lambda(T_i|X_i(s),M_i) = \lambda_0(T_i \exp \left(- \int_{s} X_i(s) \theta_1(s) ds - \theta_M M_i \right) \exp \left(- \int_{s} X_i(s) \theta_1(s) ds - \theta_M M_i \right),$$

which is very similar to the hazard function under a Cox model, seen in equation (4.3). In fact, both the exponential and Weibull AFT models are specific cases of the Cox proportional hazards model. Rather than the baseline hazard function being unknown as with the Cox model, it is of a specific known form under an exponential or Weibull model. The lognormal AFT model, however, does not satisfy proportional hazards. All other AFT models, including log-logistic, generalized gamma, and generalized $F$, also do not satisfy the proportional hazards property of the Cox model.

### 4.3 Estimation of and Testing the Indirect Effect Function

#### 4.3.1 Functional Principal Component Regression

In order to estimate the coefficient functions in the mediating equations in Section 4.1 (equation (4.1), equation (4.2), and equation (4.3)), we use FPCR, detailed in Section 2.1.1. The eigenfunctions from FPCA can be used to expand the coefficient function in equation (4.1) as

$$\gamma_1(s) \approx \sum_{k=1}^{K_n} \gamma_{1k} \phi_k(s),$$

where $\{\gamma_{1k}, k = 1, \ldots, K_n\}$ are unknown scalar basis coefficients. Therefore, the model in equation (4.1) can be approximated by

$$M_i = \gamma_0 + \sum_{k=1}^{K_n} \hat{\xi}_{ik} \gamma_{1k} + \eta_i, \quad i = 1, \ldots, n,$$

where $\hat{\xi}_{ik} = \int_{s} X_i(s) \hat{\phi}_k(s) ds$ are the estimated FPC scores. The coefficient function can be reconstructed using $\hat{\gamma}_1(s) = \sum_{k=1}^{K_n} \hat{\gamma}_{1k} \hat{\phi}_k(s)$. Similarly, the coefficient function in equation (4.2) can be represented by

$$\theta_1(s) \approx \sum_{k=1}^{K_n} \theta_{1k} \phi_k(s).$$

The model in equation (4.2) can be approximated by

$$\log(T_i) = \theta_0 + \sum_{k=1}^{K_n} \hat{\xi}_{ik} \theta_{1k} + \theta_M M_i + \nu \omega_i.$$
The above equation is an AFT model with \( \hat{\xi}_{iK_n} \) and \( M_i \) as covariates and \( \theta_{K_n} = (\theta_{11}, \ldots, \theta_{1K_n})^T \) and \( \theta_M \) as unknown coefficients. Reconstructing the coefficient function gives \( \hat{\theta}_1(s) = \sum_{k=1}^{K_n} \hat{\theta}_{1k} \hat{\phi}_k(s) \). Equivalently, the coefficient function in equation (4.3) can be represented by \( \kappa_1(s) \approx \sum_{k=1}^{K_n} \kappa_{1k} \phi_k(s) \). The model in equation (4.3) can be written as

\[
\lambda(T_i|X_i(s), M_i) = \lambda_0(T_i) \exp \left[ \sum_{k=1}^{K_n} \hat{\xi}_{ik} \kappa_{1k} + \kappa_M M_i \right].
\]

This is a Cox model with \( \hat{\xi}_{iK_n} \) and \( M_i \) as covariates and \( \kappa_{K_n} = (\kappa_{11}, \ldots, \kappa_{1K_n})^T \) and \( \kappa_M \) as unknown coefficients, the estimates of which can be used to estimate \( \kappa_1(\cdot) \) in the same manner as above.

### 4.3.2 Estimation of the Indirect Effect Function

Earlier in equation (4.4), we define the IEF as \( \delta(s) = \psi_M \gamma_1(s) \), where \( \psi_M = \theta_M \) for an AFT model and \( \psi_M = \kappa_M \) for a Cox model. As we mention previously, under a Cox model, this derivation only holds when the cumulative baseline hazard function is small, i.e. the outcome is rare. Using the FPCR procedure above for the estimation of \( \psi_M \) and \( \gamma_1(\cdot) \), the IEF is estimated by

\[
\hat{\delta}(s) = \sum_{k=1}^{K_n} \hat{\psi}_M \hat{\gamma}_1 \hat{\phi}_k(s) = \sum_{k=1}^{K_n} \hat{\delta}_{1k} \hat{\phi}_k(s),
\]

where \( \{ \hat{\delta}_{1k} = \hat{\psi}_M \hat{\gamma}_1 \hat{\phi}_k, \ k = 1, \ldots, K_n \} \). Denote the true values \( (\delta_{11}, \ldots, \delta_{1K_n})^T \) by \( \delta_{K_n} = \psi_M \gamma_{K_n} \), which we refer to as the IE vector.

Sobel (1982) provides a platform to obtain a variance estimate of the IE vector \( \delta_{K_n} \). He derives a first-order approximation for the variance of the product of two scalar coefficients based on first derivatives using the multivariate delta method. Extending his method for one scalar coefficient and one vector of coefficients, Sobel’s approximate variance of \( \delta_{K_n} \) is defined as \( \Sigma_\delta = \psi_M^2 \Sigma_\gamma + \sigma_{\psi_M}^2 \gamma_{K_n} \gamma_{K_n}^T \), where \( \Sigma_\gamma \) is the variance-covariance matrix of \( \gamma_{K_n} \) and \( \sigma_{\psi_M}^2 \) is the variance of \( \psi_M \). Using the estimate of this variance and the estimated eigenfunctions from FPCA, \( \hat{\Phi}(\cdot) \), we obtain an estimate of the variance of \( \delta(\cdot) \): \( \hat{\Sigma}_\delta(s) = \hat{\Phi}(s) \hat{\Sigma}_\delta \hat{\Phi}^T(s) \). We find pointwise confidence intervals of the IEF using this estimated
variance.

We also create simultaneous confidence intervals for the IEF, which are calculated using the method illustrated in Crainiceanu et al. (2012). This method requires the estimation of the IEF, \( \hat{\delta}(s) = \sum_{k=1}^{K_n} \hat{\psi}_M \hat{\gamma}_n \hat{\phi}_k(s) \), and its variance estimate based on Sobel’s first-order variance approximation and the estimated eigenfunctions from FPCA, \( \hat{\Sigma}_{\delta}(s) = \hat{\Phi}(s) \hat{\Sigma}_\delta \hat{\Phi}^T(s) \). We simulate \( d_u(s) \) from a multivariate normal distribution, \( \text{MVN}\{\hat{\delta}(s), \hat{\Sigma}_{\delta}(s)\} \) and calculate \( x_u = \max_s \{|d_u(s) - \hat{\delta}(s)| \hat{\sigma}(s)\} \), where \( \hat{\sigma}^2(s) \) is the \( s \)-th diagonal element of \( \hat{\Sigma}_{\delta}(s) \). Repeating this process \( u = 1, \ldots, U \) times and obtaining \( q_{1-\alpha} \), the \((1 - \alpha) \times 100\% \) empirical quantile of the sample \( \{x_u: u = 1, \ldots, U\} \), we calculate the joint confidence intervals using \( \hat{\delta}(s) \pm q_{1-\alpha} \hat{\sigma}(s) \). Crainiceanu et al. (2012) note that with enough samples, the sampling variability in \( \hat{\delta}(s) \) and \( \hat{\Sigma}_{\delta}(s) \) can be ignored.

**4.3.3 Testing the Indirect Effect Function**

We wish to test if the mediating variable has an effect on the relationship between the functional covariate and the time-to-event outcome by testing if the IEF differs from zero. The null hypothesis is \( H_0: \delta(\cdot) = 0 \), which is equivalent to testing \( H_0: \delta_K = 0 \). Since \( \delta_K = \psi_M \gamma_K \), there are three ways in which \( \delta_K \) can equal zero: (1) \( \psi_M = 0 \) with \( \gamma_K \neq 0 \), (2) \( \gamma_K = 0 \) with \( \psi_M \neq 0 \), and (3) \( \psi_M = \gamma_K = 0 \). However, for mediation to exist, \( \psi_M \) should not equal 0; otherwise the mediator does not have any effect on \( Y \) when \( X \) is in the model. Therefore, we focus on case (2). We consider the same three testing procedures described in Section 3.1.2.

First, we study a Wald test:

\[
W_n = \hat{\delta}_K^T \hat{\Sigma}_\delta^{-1} \hat{\delta}_K \sim \chi^2_d, \quad d = \text{trace}(\hat{\Sigma}_\delta) = K_n,
\]

where \( \hat{\Sigma}_\delta \) is the variance of \( \hat{\delta}_K \) based on Sobel’s approximation. Therefore, the test statistic \( W_n \) is compared to a \( \chi^2 \)-distribution with \( d \) degrees of freedom in this one-sided alternative. This Wald test assumes the IE vector \( \delta_K \) tends to be normally distributed as the sample size gets large.

Secondly, we consider a bootstrap test based on the procedure in Shrout and Bolger (2002) that does
not require any distributional assumptions. We bootstrap the dataset \([Y_i, \Delta_i, (\hat{\xi}_{i1}, \ldots, \hat{\xi}_{iK_n})], M_i], i = 1, \ldots, n\) and calculate the IE vector for each bootstrap sample \(\hat{\delta}_{k_n}^{(b)} = \psi_M^{(b)} \hat{\gamma}_{k_n}^{(b)}, b = 1, \ldots, B\). For each of the \(K_n\) estimates of the IE vector, we use the bootstrap percentiles to calculate \(K_n\) asymmetric confidence intervals. Since we perform \(K_n\) tests, we correct for multiple testing using a Bonferroni correction. We note that performing the bootstrap procedure on the estimated FPC scores may lead to biased results as we do not account for the variation of the FPCA decomposition. However, bootstrapping outside the FPCA loop can cause issues with the decomposition since observations may be repeated several times.

Finally, we examine a test based on simultaneous confidence intervals for the IEF; we previously describe the method used to calculate the simultaneous intervals in Section 4.3.2. We calculate the joint confidence intervals using \(\hat{\delta}(s) \pm q_{1-\alpha} \hat{\sigma}(s)\). If zero is outside the joint interval at one or more of the \(J\) time points, we reject the null hypothesis. Also, this test is based on Sobel’s first-order variance approximation and assumes the distribution of \(\delta_{k_n}\) is normal.

### 4.4 Simulation Study

#### 4.4.1 Purpose and Predictions

The purpose of this simulation study is to examine the estimation of the IEF and the three proposed testing procedures in the case of causal functional mediation analysis with a time-to-event outcome. We consider three survival models: an exponential AFT model, a lognormal AFT model, and a Cox model. We predict good estimation for the correctly-specified AFT models and plan to inspect the effect of misspecifying the error distribution and modeling the incorrect AFT model. We predict poor estimation for the Cox model because as we discuss previously, the derivation does not always hold, and therefore, it should not be used for estimation purposes as it may result in a biased estimate, especially when the outcome is not rare.

VanderWeele (2011) explains that while a Cox model often results in poor estimation, it still leads to a valid test of the IE, even with a common outcome. He shows that using a test based on the POC method still provides a proper test, even when the outcome is common. Therefore, we predict the Wald test based
on Sobel’s variance approximation performs well for both the AFT model and the Cox model. Also, it works well in standard functional mediation analysis with Gaussian and generalized responses as shown in Section 3.2.3. In Section 3.2.3, the simultaneous interval test also performs well, while the bootstrap test results in much too high Type 1 error. We study whether the results of these three tests change when the response is time-to-event data and an AFT model and a Cox model are implemented.

### 4.4.2 Design and Data Generation

The set-up of this simulation study is very similar to the one in Section 3.2. For all three models, we consider $\mathcal{S}$ to be the interval $[0, 1]$ and the functional covariate is observed fully at $J$ equally spaced points on the interval, where $J$ increases as sample size increases. The functional covariate is generated with four trigonometric basis functions as $X_i(s) = a_{i1} \sqrt{2} \sin(2\pi s) + a_{i2} \sqrt{2} \cos(2\pi s) + a_{i3} \sqrt{2} \sin(4\pi s) + a_{i4} \sqrt{2} \cos(4\pi s)$, where $s = \{s_1, s_2, \ldots, s_J\}$ and $a_{i1}, a_{i2}, a_{i3},$ and $a_{i4}$ are normally distributed, independent random variables with mean zero and respective variances 1.0, 0.5, 0.25, and 0.125. We set the covariate error variance $\sigma^2_X = 1$. We investigate two IEFs: (1) $\delta(s) = 0.4 \sqrt{2} \cos(2\pi s)$ and (2) $\delta(s) = 0.4 \sqrt{2} \sin(2\pi s) + \sqrt{2} \cos(2\pi s)$. IEF (2) is more complex and has more curvature than IEF (1).

We set the sample size $n$ and the number of grid points $J$ at two different combinations. As $n$ increases, we increase $J$ at a similar rate, leading to the following two combinations of $n$ and $J$: $n = 250$ with $J = 201$ and $n = 500$ with $J = 401$. In our simulation study, the intercepts $\gamma_0$ and $\theta_0$ are set to zero and the error variance $\sigma^2_\eta$ is set to one. The scalar, continuous mediating variable is generated using $M_i = A_i g + \eta_i$, where $\eta_i \overset{iid}{\sim} N(0, \sigma^2_\eta)$, $A_i = (a_{i1}, a_{i2}, a_{i3}, a_{i4})$, and $g^T$ equals one of the following vectors corresponding to the two IEFs above: (1) $g^T = (0, 0.2, 0, 0)$ or (2) $g^T = (0.2, 0.2, 0, 0)$. These choices of $g$ are used to regulate the amount of correlation between the covariate and the mediator. Under this model, the correlation between the two is approximately 15% for IEF (1) and 25% for IEF (2).

We generate the time-to-event response variable in two different ways and model all three of the survival methods for each way, leading to a total of six combinations. First, we generate the time-to-event data under an exponential distribution (Bender et al., 2005). The survival times $T_i$ are generated.
from an exponential distribution with mean parameter $e^{-A_i \cdot b - \psi M_i}$, where $A_i$ is the same as above, $\psi_M = 2$, and $b^T$ is one of the following vectors corresponding to the two IEFs: (1) $b^T = (0, 3, 0, 0)$ or (2) $b^T = (3, 3, 0, 0)$. The censoring times $C_i$ are generated from an exponential distribution with mean parameter .05, which results in approximately 25%-30% censoring, which is a moderate amount. Using this data, we fit an exponential AFT model, a lognormal AFT model, and a Cox model. The exponential AFT model correctly specifies the parametric model for the data. On the other hand, the Cox model requires no distributional assumptions. We also study misspecification of the model by generating from an exponential distribution and fitting a lognormal model.

Secondly, we generate the time-to-event data under a lognormal model. The survival times $T_i$ are generated using $Y_i = \exp\{A_i b + \psi M_M + \omega_i\}$, where $A_i$, $b$, and $\psi_M$ are the same as above and $\omega_i$ are standard normal random variables. The censoring times are generated in the same manner as above, leading to approximately 25%-30% censoring. Again, all three models are fit: a correctly specified lognormal AFT model, a semi-parametric Cox model, and a misspecified exponential AFT model. IEF (1) is calculated by multiplying $\psi_M = 2$ by $\gamma_1(s) = g^T \Phi(s) = 0.2 \sqrt{2} \cos(2\pi s)$, leading to $\delta(s) = 0.4 \sqrt{2} \cos(2\pi s)$. IEF (2) is found similarly. The number of simulations for each combination of IEF and sample size is $N = 1,000$. Also, the truncation number in FPCA, $K_n$, is defined as the minimum number of FPCs needed to account for 95% of the variability of the data.

### 4.4.3 Estimation Results

We examine the estimation of the IEF by looking at integrated squared bias (ISB), integrated mean squared error (IMSE), and integrated coverage (IC) for $\delta(\cdot)$. We calculate 95%, 90%, and 85% confidence intervals for the IEF using Sobel’s variance approximation, $\hat{\Sigma}(s) = \hat{\Phi}(s) \hat{\Sigma} \hat{\Phi}^T(s)$, which assumes normality of $\delta_{K_n}$. Table 4.1 presents the estimation results of the simulation study for both IEFs for all three models at each level of sample size. The top table refers to the case when the data is exponentially generated, and the bottom table refers to the case when the data is lognormally generated. Estimation is good for both AFT models, even when the error distribution is misspecified. However, it is biased for the Cox model, leading to low coverage. The Cox model should not be implemented when an estimate of the IEF
is desired; this is consistent with VanderWeele’s finding in the scalar case.

### 4.4.4 Testing Results

We compare the three testing procedures described earlier in Section 4.3.3 by analyzing Type 1 error and power using a significance level of $\alpha = .05$. We focus on the case when $\psi_M \neq 0$ as it should not equal zero for mediation to exist. Therefore, we set $\psi_M$ equal to a constant and vary $\gamma_K$ to examine Type 1 error and power. We set $\psi_M = 1$ and 2, and for each case of $\psi_M$, we vary $\gamma_K = 0, 0.1, 0.2, 0.3, \text{ and } 0.4$. We also look at larger levels of $\psi_M$ and find similar results as those described below. In our simulation, we perform the bootstrap procedure $B = 200$ times and the simultaneous interval process $U = 200$ times.

First, we discuss the Type 1 error results for the bootstrap based test. For all of the cases we consider, Type 1 error is highly inflated, ranging from 0.091 to 0.096. This is almost double the nominal amount of $\alpha = 0.05$. Therefore, we do not suggest using the bootstrap test as it does not control Type 1 error. This may be a result of bootstrapping on the estimated FPC scores and losing a source of variation. Therefore, we focus on the Wald test and the simultaneous interval test. Table 4.2 displays the simulation results for Type 1 error and power for the Wald test and the simultaneous interval test for IEF (1) under the exponential data generation scheme with $n = 250$. The top table gives the results when an exponential AFT model is fit, the middle table shows the results when a lognormal AFT model is fit, and the bottom table displays the results when a Cox model is fit. Figure 4.1 depicts the power curves associated with Table 4.2 when an exponential AFT model is fit. Since the other cases show very similar results, we do not present those power curves here. These power curves help illustrate the differences in power between the Wald test and the simultaneous interval test. The results for IEF (2), for $n = 500$, and under the lognormal data generation scheme are similar; therefore, they are not shown for convenience. All three survival models yield similar results for each test, and the three proposed tests provide results analogous to those in Section 3.2.3, which studies standard functional mediation models with Gaussian and generalized responses. We note that misspecifying the parametric AFT model does not affect the estimation or testing results in our simulation study.

The Wald test using Sobel’s first-order approximation of the variance results in proper Type 1 error
Table 4.1: ISB, IMSE, and IC for $\delta(\cdot)$ for IEF (1) and IEF (2) for all three survival models. Coverages are given in terms of percent. The mean standard error for the estimates in this table is 0.003.

### Exponentially Distributed Data

<table>
<thead>
<tr>
<th></th>
<th>$n$</th>
<th>ISB</th>
<th>IMSE</th>
<th>95% IC</th>
<th>90% IC</th>
<th>85% IC</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Exponential AFT Model</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IEF (1)</td>
<td>250</td>
<td>0.001</td>
<td>0.268</td>
<td>95.2</td>
<td>90.1</td>
<td>85.0</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.000</td>
<td>0.130</td>
<td>94.6</td>
<td>89.5</td>
<td>84.4</td>
</tr>
<tr>
<td>IEF (2)</td>
<td>250</td>
<td>0.001</td>
<td>0.268</td>
<td>95.1</td>
<td>90.0</td>
<td>85.0</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.000</td>
<td>0.131</td>
<td>94.5</td>
<td>89.5</td>
<td>84.3</td>
</tr>
<tr>
<td><strong>Lognormal AFT Model</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IEF (1)</td>
<td>250</td>
<td>0.001</td>
<td>0.272</td>
<td>95.2</td>
<td>90.1</td>
<td>85.0</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.000</td>
<td>0.132</td>
<td>94.6</td>
<td>89.5</td>
<td>84.3</td>
</tr>
<tr>
<td>IEF (2)</td>
<td>250</td>
<td>0.001</td>
<td>0.272</td>
<td>95.1</td>
<td>90.0</td>
<td>84.9</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.000</td>
<td>0.132</td>
<td>94.5</td>
<td>89.5</td>
<td>84.3</td>
</tr>
<tr>
<td><strong>Cox Model</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IEF (1)</td>
<td>250</td>
<td>0.621</td>
<td>0.877</td>
<td>65.9</td>
<td>55.6</td>
<td>48.3</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.638</td>
<td>0.764</td>
<td>43.9</td>
<td>35.6</td>
<td>30.5</td>
</tr>
<tr>
<td>IEF (2)</td>
<td>250</td>
<td>1.213</td>
<td>1.452</td>
<td>47.4</td>
<td>39.3</td>
<td>34.0</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>1.239</td>
<td>1.362</td>
<td>31.3</td>
<td>25.5</td>
<td>21.9</td>
</tr>
</tbody>
</table>

### Lognormally Distributed Data

<table>
<thead>
<tr>
<th></th>
<th>$n$</th>
<th>ISB</th>
<th>IMSE</th>
<th>95% IC</th>
<th>90% IC</th>
<th>85% IC</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Exponential AFT Model</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IEF (1)</td>
<td>250</td>
<td>0.000</td>
<td>0.268</td>
<td>94.8</td>
<td>89.9</td>
<td>84.8</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.000</td>
<td>0.121</td>
<td>95.4</td>
<td>90.9</td>
<td>86.9</td>
</tr>
<tr>
<td>IEF (2)</td>
<td>250</td>
<td>0.001</td>
<td>0.270</td>
<td>94.8</td>
<td>89.8</td>
<td>84.9</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.000</td>
<td>0.122</td>
<td>95.4</td>
<td>90.9</td>
<td>86.9</td>
</tr>
<tr>
<td><strong>Lognormal AFT Model</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IEF (1)</td>
<td>250</td>
<td>0.000</td>
<td>0.268</td>
<td>94.8</td>
<td>89.9</td>
<td>84.8</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.000</td>
<td>0.121</td>
<td>95.5</td>
<td>90.9</td>
<td>86.0</td>
</tr>
<tr>
<td>IEF (2)</td>
<td>250</td>
<td>0.000</td>
<td>0.268</td>
<td>94.8</td>
<td>89.8</td>
<td>84.8</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.000</td>
<td>0.121</td>
<td>95.4</td>
<td>90.9</td>
<td>85.9</td>
</tr>
<tr>
<td><strong>Cox Model</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IEF (1)</td>
<td>250</td>
<td>0.683</td>
<td>0.817</td>
<td>44.8</td>
<td>36.4</td>
<td>31.2</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>1.328</td>
<td>1.610</td>
<td>48.8</td>
<td>40.3</td>
<td>35.0</td>
</tr>
<tr>
<td>IEF (2)</td>
<td>250</td>
<td>1.339</td>
<td>1.471</td>
<td>32.2</td>
<td>26.2</td>
<td>22.5</td>
</tr>
</tbody>
</table>
when $\psi_M \neq 0$, regardless of IEF, sample size, model fit, or data generation scheme. In addition, power increases as $\gamma_K_n$ increases. Power is also higher when $n = 500$ and for IE (2). Power does not increase, however, when $\psi_M$ increases, except for a slight increase in power when a Cox model is fit. The results of the simultaneous interval test are very similar to those of the Wald test. However, power is slightly lower for this test than for the Wald test. Figure 4.2 illustrates pointwise and simultaneous confidence intervals for one simulation of IEF (1), $\delta(s) = 0.4[\sqrt{2}\cos(2\pi s)]$, when $n = 500$ for each of the three models fit under the exponential data generation scheme. In all three cases, the simultaneous interval is wider than the pointwise interval as expected. Similar graphs are noted under the lognormal data generation scheme.

For our testing procedures, while we focus on the case when $\psi_M \neq 0$, we note there are other cases when $\psi_M = 0$. The Type 1 error results when $\psi_M = 0$ are extremely grim. For the Wald test and simultaneous interval test, Type 1 error ranges from immensely conservative for a Cox fit to woefully large in some situations for an AFT fit. The bootstrap test remains conservative for all fits. Therefore, we do not suggest any of our three tests when $\psi_M = 0$. In addition, as we note previously, a possible issue with the Wald test and simultaneous interval test is the assumption that $\delta_K_n = \psi_M \gamma_K_n$ is normally distributed. While this may be an issue when $\psi_M = 0$, it does not seem to affect the results when $\psi_M \neq 0$ in our simulation.

Overall, for the case when $\psi_M \neq 0$, the exponential AFT model, lognormal AFT model, and Cox model perform similarly for the three testing procedures we investigate, even when the model is misspecified. The Wald test and the simultaneous interval test have better Type 1 error rates than the bootstrap test. In addition, the Wald test has higher power than the simultaneous interval test. Therefore, we recommend using the Wald test. It properly tests the IEF under either an AFT or Cox model as expected. While estimation is poor for a Cox model, it results in a proper test, even with a common outcome; this is also consistent with the findings in VanderWeele (2011).
Table 4.2: Type 1 error and power results for the Wald test and simultaneous interval test for IEF (1) under the exponential data generation scheme when $n = 250$ for a significance level of $\alpha = 0.05$.

<table>
<thead>
<tr>
<th></th>
<th>$\psi_M = 1$</th>
<th>Wald Test</th>
<th>Interval Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma_K_n = 0.0$</td>
<td></td>
<td>0.057</td>
<td>0.060</td>
</tr>
<tr>
<td>$\gamma_K_n = 0.1$</td>
<td></td>
<td>0.126</td>
<td>0.091</td>
</tr>
<tr>
<td>$\gamma_K_n = 0.2$</td>
<td></td>
<td>0.372</td>
<td>0.219</td>
</tr>
<tr>
<td>$\gamma_K_n = 0.3$</td>
<td></td>
<td>0.739</td>
<td>0.422</td>
</tr>
<tr>
<td>$\gamma_K_n = 0.4$</td>
<td></td>
<td>0.960</td>
<td>0.683</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$\psi_M = 2$</th>
<th>Wald Test</th>
<th>Interval Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma_K_n = 0.0$</td>
<td></td>
<td>0.057</td>
<td>0.060</td>
</tr>
<tr>
<td>$\gamma_K_n = 0.1$</td>
<td></td>
<td>0.126</td>
<td>0.091</td>
</tr>
<tr>
<td>$\gamma_K_n = 0.2$</td>
<td></td>
<td>0.372</td>
<td>0.219</td>
</tr>
<tr>
<td>$\gamma_K_n = 0.3$</td>
<td></td>
<td>0.739</td>
<td>0.423</td>
</tr>
<tr>
<td>$\gamma_K_n = 0.4$</td>
<td></td>
<td>0.960</td>
<td>0.683</td>
</tr>
</tbody>
</table>

Exponential AFT Model

Lognormal AFT Model

Cox Model

<table>
<thead>
<tr>
<th></th>
<th>$\psi_M = 1$</th>
<th>Wald Test</th>
<th>Interval Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma_K_n = 0.0$</td>
<td></td>
<td>0.037</td>
<td>0.045</td>
</tr>
<tr>
<td>$\gamma_K_n = 0.1$</td>
<td></td>
<td>0.094</td>
<td>0.072</td>
</tr>
<tr>
<td>$\gamma_K_n = 0.2$</td>
<td></td>
<td>0.327</td>
<td>0.180</td>
</tr>
<tr>
<td>$\gamma_K_n = 0.3$</td>
<td></td>
<td>0.695</td>
<td>0.373</td>
</tr>
<tr>
<td>$\gamma_K_n = 0.4$</td>
<td></td>
<td>0.937</td>
<td>0.636</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$\psi_M = 2$</th>
<th>Wald Test</th>
<th>Interval Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma_K_n = 0.0$</td>
<td></td>
<td>0.048</td>
<td>0.055</td>
</tr>
<tr>
<td>$\gamma_K_n = 0.1$</td>
<td></td>
<td>0.109</td>
<td>0.076</td>
</tr>
<tr>
<td>$\gamma_K_n = 0.2$</td>
<td></td>
<td>0.344</td>
<td>0.202</td>
</tr>
<tr>
<td>$\gamma_K_n = 0.3$</td>
<td></td>
<td>0.718</td>
<td>0.397</td>
</tr>
<tr>
<td>$\gamma_K_n = 0.4$</td>
<td></td>
<td>0.948</td>
<td>0.657</td>
</tr>
</tbody>
</table>
4.5 Summary

This chapter combines the methodologies of FPCR, survival analysis, and causal mediation analysis. We establish a FPCR based method to estimate and test the IEF under a counterfactual functional mediation model with a survival response. In order to model the survival data, we consider the parametric AFT model and the semi-parametric Cox model under two data generation schemes, exponential and lognormal. We show that while the AFT model properly estimates the IEF, even under misspecification, the Cox model can result in a highly biased estimate of the IEF because its derivation only holds when the cumulative baseline hazard function is small, i.e. the outcome is rare. Therefore, we do not suggest using a Cox model when an estimate of the IEF is desired. We also examine three testing procedures of the IEF. Through numerical simulations, we conclude the Wald test based on Sobel’s variance approximation renders an adequate test for both the AFT and Cox models when $\psi_M \neq 0$, even if the parametric AFT model is misspecified. However, when $\psi_M = 0$, none of the testing procedures have proper Type 1 error, and we do not suggest using any of them.
Figure 4.2: Pointwise (dashed) and simultaneous (dotted) confidence intervals for one simulation of IEF (1), \( \delta(s) = 0.4[\sqrt{2}\cos(2\pi s)] \), when \( n = 500 \) for each of the three models under the exponential data generation scheme. The top graph fits an exponential AFT model, the middle graph fits a lognormal AFT model, and the bottom graph fits a Cox model.
CHAPTER 5

CAUSAL NONLINEAR FUNCTIONAL MEDIATION ANALYSIS USING FUNCTIONAL GENERALIZED ADDITIVE MODELS

5.1 Model and Framework

In this chapter, we extend functional mediation analysis to include nonlinear functional models where we allow the functional covariate to be modeled nonlinearly. The standard approach to mediation is not applicable in the case of nonlinearities, so we must use the counterfactual approach. We use functional generalized additive models (FGAMs), presented in Section 2.1.3, to allow the effect of the covariate to be modeled nonlinearly (McLean et al., 2014). We consider the following mediation model; for $i = 1, \ldots, n$, the data consist of a scalar response $Y_i$, a scalar and continuous mediator $M_i$, and a functional covariate
\(X_i(\cdot)\) defined on a closed interval \(\mathcal{S}\). Again, the covariate function \(X_i(\cdot)\) is often observed sparsely and measured with error, meaning one observes \(\{W_{i1}, \ldots, W_{iJ_i}\}\) corresponding to possibly irregular time points \(\{s_{ij}: s_{ij} \in \mathcal{S}\}\), where \(J_i\) corresponds to the number of points observed for subject \(i\). Here, \(W_i(s_{ij}) = X_i(s_{ij}) + e_i(s_{ij})\), where \(e_i\) are random noise variables with mean zero and variance \(\sigma^2_X\). Thus, for subject \(i\), data are typically of the form \([Y_i, M_i, \{W_i(s_{ij}): s_{ij} \in \mathcal{S}\}, i = 1, \ldots, n, j = 1, \ldots, J_i]\). For simplicity, we assume the covariate function is observed fully on a dense grid of points. In addition, we initially consider a continuous response; however, a FGAM can easily account for a generalized response using a link function. We also consider a binary response in our simulation study using a logit link.

An important aspect of mediation analysis is the indirect effect (IE) of the mediator on the relationship between the covariate and the response. Recall the counterfactual model presented in Section 1.1.2. The total causal effect of the covariate on the response can be decomposed into the indirect effect (IE) and the direct effect (DE):

\[
E(Y_x) - E(Y^*_x) = E(Y_{xM_x} - Y_{x'M_x}) + E(Y_{x'M_x} - Y_{x'M^*_x}),
\]

where \(E(Y_{xM_x} - Y_{x'M_x})\) is the IE and \(E(Y_{x'M_x} - Y_{x'M^*_x})\) is the DE (VanderWeele and Vansteelandt, 2010). Here, the IE is the expected change in \(Y\) when the covariate is held constant at \(X = x\) and the mediator is changed from \(M_x\) to the value it would have taken had \(X\) been set to \(x^*\), \(M_{x^*}\). On the other hand, the DE is the expected change in \(Y\) when \(X\) is changed from \(X = x\) to \(X = x^*\) and the mediator remains constant. We focus only on the IE of the mediator: 

\[
IE = E(Y_{xM_x} - Y_{x'M_x}).
\]

In order to calculate the IE and allow for a nonlinear effect of the functional covariate, we model the data using two functional generalized additive models (FGAMs):

\[
M_i = \alpha_1 + \int_{\mathcal{S}} F_1\{X_i(s), s\}ds + \varepsilon_{1i}; \tag{5.1}
\]

\[
Y_i = \alpha_2 + M_i\beta_M + \int_{\mathcal{S}} F_2\{X_i(s), s\}ds + \varepsilon_{2i}. \tag{5.2}
\]

The model in equation (5.1) regresses the mediating variable on the functional covariate; \(\alpha_1\) is an intercept,
\(F_1(\cdot, \cdot)\) is an unspecified smooth function, and \(\varepsilon_{i1} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_{\varepsilon_1}^2)\). The model in equation (5.2) regresses \(Y_i\) on both the scalar mediator and the functional covariate. Here, \(\alpha_2\) is an intercept, \(\beta_M\) is a scalar partial relationship between the mediator and the response, \(F_2(\cdot, \cdot)\) is an unknown smooth regression function, and \(\varepsilon_{i2} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_{\varepsilon_2}^2)\). When \(Y_i\) is a generalized response, equation (5.2) requires the addition of an appropriate link function.

As mentioned previously in Section 2.1.3, the FGAM procedure can incorporate several types of bases and penalties to estimate the regression function \(F(\cdot, \cdot)\). We use tensor product cubic P-splines (Eilers and Marx, 1996) with a second order difference penalty on the coefficients. In addition, we also previously note it is preferable to transform the covariate in order to ensure each tensor product has observed data on its support. We use a pointwise transformation to standardize the covariate by centering and scaling it. An advantage of this transformation approach over the quantile transformation used by McLean et al. (2014) is that future observations falling outside of the range of the original data at a particular \(s\) are not all assigned the value zero or one. We define the pointwise center and scale transformation of \(X(\cdot)\) by

\[
X^{C/S}(s) = \{X(s) - \mu_X(s)\} / \sigma_X(s),
\]

where \(\mu_X(\cdot)\) and \(\sigma_X(\cdot)\) are the mean and standard deviation of \(X(\cdot)\). We use functional principal component analysis (FPCA) to estimate the mean and variance and then center and scale the FPC approximation of the functional covariate. We choose the number of FPCs to use by the proportion of variance explained. For simplicity and ease of notation, we drop the \(C/S\) superscript for the remainder of the chapter with the knowledge that the transformed covariate is used.

Under a counterfactual approach to mediation analysis, the IE is defined in equation (1.1) as the expected change in response due to a change in the mediating variable when the covariate remains the same. Using equation (5.1) and equation (5.2), we derive the IE for causal nonlinear functional mediation.
analysis using FGAMs as

\[
\text{IE} = E(Y_{sM_x} - Y_{sM_x^*}) \\
= \alpha_2 + \beta_M E(M_x) + \int \mathcal{F}_2 \{X(s), s\} ds - \alpha_2 - \beta_M E(M_x^*) - \int \mathcal{F}_2 \{X(s), s\} ds \\
= \beta_M E(M_x - M_x^*) \\
= \beta_M \left( \alpha_1 + \int \mathcal{F}_1 \{X(s), s\} ds - \alpha_1 - \int \mathcal{F}_1 \{X^*(s), s\} ds \right) \\
= \beta_M \int \mathcal{F}_1 \{X(s), s\} - \mathcal{F}_1 \{X^*(s), s\} \} ds.
\]

In the following section, we estimate this quantity.

### 5.1.1 Estimation of the Indirect Effect

Above we define the IE = \( \beta_M \int \mathcal{F}_1 \{X(s), s\} - \mathcal{F}_1 \{X^*(s), s\} \} ds \). In order to estimate the IE, we use the FGAM procedure presented in Section 2.1.3. Since \( \int \mathcal{F}_1 \{X(s), s\} ds = \sum_{j=1}^{K_x} \sum_{k=1}^{K_x} \theta_{jk} Z_{jk} \), where \( Z_{jk} = \int \mathcal{F}_1 \{X(s)\} \mathcal{B}_j(s) \mathcal{B}_k(s) ds \), the IE is estimated by

\[
\hat{\text{IE}} = \hat{\beta}_M \left( \sum_{j=1}^{K_x} \sum_{k=1}^{K_x} \hat{\theta}_{jk} (Z_{jk} - Z^*_{jk}) \right),
\]

where \( Z^*_{jk} = \int \mathcal{F}_1 \{X^*(s)\} \mathcal{B}_j(s) \mathcal{B}_k(s) ds \). For convenience, we can write this in matrix form, \( \hat{\text{IE}} = \hat{\beta}_M (Z - Z^*)^T \hat{\theta} \).

Sobel (1982) provides a basis to obtain a variance estimate of the IE. He derives a first-order approximation for the variance of the product of two scalar coefficients based on first derivatives using the multivariate delta method. Extending his method for one scalar coefficient and one vector of coefficients,
Sobel’s approximate variance of \( \hat{IE} \) is defined as

\[
\text{Var}(\hat{IE}) = \hat{\beta}_M^2 \text{Var}\left\{(Z - Z^*)^T \hat{\theta}\right\} + \{(Z - Z^*)^T \hat{\theta}\}^2 \text{Var}(\hat{\beta}_M)
\]

\[
= \hat{\beta}_M^2 (Z - Z^*)^T \text{Var}(\hat{\theta})(Z - Z^*) + \{(Z - Z^*)^T \hat{\theta}\}^2 \text{Var}(\hat{\beta}_M)
\]

\[
= \hat{\beta}_M^2 (Z - Z^*)^T \Sigma_\theta (Z - Z^*) + \{(Z - Z^*)^T \hat{\theta}\}^2 \sigma_{\hat{\beta}_M}^2,
\]

where \( \Sigma_\theta \) is the variance-covariance matrix of \( \hat{\theta} \) and \( \sigma_{\hat{\beta}_M}^2 \) is the variance of \( \hat{\beta}_M \). McLean et al. (2014) use the Bayesian approach by Wahba (1983) to estimate these variances and we follow suit; see McLean et al. (2014) for details.

### 5.2 Comparison to Functional Linear Models

We specifically use FGAMs in order to model the covariate non-linearly. In our simulation study, though, we wish to compare the results from a FGAM to those from a functional linear model (FLM) using splines. We use the penalized least squares method detailed in Section 2.1.1. Using two FLMs, the equations we need to find the IE are

\[
M_i = \alpha_3 + \int X_i(s) \beta_1(s) ds + \epsilon_{3i}; \quad (5.3)
\]

\[
Y_i = \alpha_4 + \beta_{M_2} M_i + \int X_i(s) \beta_2(s) ds + \epsilon_{4i}, \quad (5.4)
\]

where \( \alpha_3 \) and \( \alpha_4 \) are unknown intercepts, \( \beta_1(\cdot) \) and \( \beta_2(\cdot) \) are unknown coefficient functions, \( \beta_{M_2} \) is a scalar partial coefficient, \( \epsilon_{3i} \sim \text{iid } N(0, \sigma_{\epsilon_3}^2) \), and \( \epsilon_{4i} \sim \text{iid } N(0, \sigma_{\epsilon_4}^2) \). Again, when the response is generalized, a link function is needed to model equation (5.4), resulting in a generalized functional linear model (GFLM). Using the counterfactual definition of the IE in equation (1.1), the IE under the FLM framework
in equation (5.3) and equation (5.4) is

\[
IE = E(Y_{sM_1} - Y_{sM_2})
\]

\[
= \beta_{M_2} \int_0^\tau \beta_1(s) \{X(s) - X^*(s)\} \, ds.
\]

Following Ramsay and Silverman (2005), for equation (5.3), and similarly for equation (5.4), assume that given two basis systems \( \{\theta_l(\cdot), l \geq 1\} \) and \( \{\psi_k(\cdot), k \geq 1\} \), the functions \( \beta_1(\cdot) \) and \( X_i(\cdot) \) can be written as

\[
\beta_1(s) = \sum_{l=1}^L \beta_{1l} \theta_l(s), \quad X_i(s) = \sum_{k=1}^K c_{ik} \psi_k(s),
\]

where \( \beta_{1l} \) are unknown coefficients and \( c_{ik} \) are covariate basis coefficients. Thus, we write

\[
\int X_i(s) \beta_1(s) \, ds = \sum_{k=1}^K \sum_{l=1}^L c_{ik} \int \psi_k(s) \theta_l(s) \, ds \beta_{1l} = c_i^T J \beta,
\]

where \( c_i = (c_{i1}, \ldots, c_{iK})^T \), \( \beta = (\beta_{11}, \ldots, \beta_{1L})^T \), and \( J \) is a \( K \times L \) matrix with the \((k,l)\)-th element \( J_{kl} = \int \psi_k(s) \theta_l(s) \, ds \). We then write the FLM in equation (5.3) in the form

\[
M_i = \alpha_3 + Z_i \beta + \epsilon_{3i},
\]

where \( Z_i = c_i^T J \). We include a penalty term for smoothness and use penalized least squares to estimate the model (see equation (2.2)). A similar procedure is used to model equation (5.4). We use 40 cubic P-splines with a second order difference penalty in our simulation study to model \( \beta_1(\cdot) \) and \( \beta_2(\cdot) \). Using the above procedure, the IE is estimated by

\[
\hat{IE} = \hat{\beta}_{M_2} (c - c^*)^T J \hat{\beta}.
\]

The variance of \( \hat{IE} \) is

\[
\text{Var}(\hat{IE}) = \hat{\beta}_{M_2}^2 (c - c^*)^T J \Sigma_{\beta} J^T (c - c^*) + (c - c^*)^T J \hat{\beta})^2 \sigma_{\beta_{M_2}}^2,
\]

59
where $\Sigma_\beta$ is the variance-covariance matrix of $\hat{\beta}$ and $\sigma^2_{\beta M}$ is the variance of $\hat{\beta}_M^2$.

5.3 Simulation Study

5.3.1 Design and Data Generation

In this simulation study, we investigate the estimation of the IE using FGAMs for several different choices of bivariate functions, $F(\cdot, \cdot)$. We consider $\mathcal{I}$ to be the interval $[0, 1]$ and the functional covariate is observed fully at $J$ equally spaced points on the interval, where $J$ increases as sample size increases. We consider both a Gaussian response and a binary response using a logit link. The functional covariate is generated as $X_i(s) = a_{i1} + a_{i2}\sqrt{2}\sin(2\pi s) + a_{i3}\sqrt{2}\cos(2\pi s)$, where $s = \{s_1, s_2, \ldots, s_J\}$ and $a_{i1}$, $a_{i2}$, and $a_{i3}$ are normally distributed, independent random variables with mean zero and respective variances 9, 4, and 1. We set the covariate error variance $\sigma^2_X = 1$. We set the sample size $n$ and the number of grid points $J$ at two combinations: $n = 300$ with $J = 101$ and $n = 500$ with $J = 201$. The intercepts $\alpha_1$ and $\alpha_2$ are set to zero, and the error variances $\sigma^2_{\epsilon_1}$ and $\sigma^2_{\epsilon_2}$ are set to one.

We set $F_1(\cdot, \cdot) = F_2(\cdot, \cdot) = F(\cdot, \cdot)$ and consider four different functions:

- Nonlinear (1): $F(x, s) = 10 \exp\{-(x/5)^2 - ((s-.5)/.3)^2\}$;
- Nonlinear (2): $F(x, s) = 1 + x + s + 2x^2s$;
- Linear (1): $F(x, s) = 1 + x + s$;
- Linear (2): $F(x, s) = xs$.

The first two are nonlinear in nature, while the last two are linear functions. The core part of Nonlinear (1) comes from a function McLean et al. (2014) consider in their simulation study; they also study the function Linear (2). Nonlinear (2) and Linear (1) come from the simulation study in Kim et al. (2016). Finally, we set $\beta_M = 2$. The number of simulations for each combination of $F(\cdot, \cdot)$ and sample size is $N = 1,000$. The number of basis functions used for $F(\cdot, \cdot)$ is set to $K_x = 7$ and $K_s = 7$. As we state previously, we use a tensor product basis with cubic P-splines with a second order difference penalty on...
the coefficients. The proportion of variance explained used in FPCA to transform the covariate is 95%. We extensively use the ‘fgam’ and ‘predict.gam’ functions, along with others, from the ‘refund’ and ‘mgcv’ packages in R (Crainiceanu et al., 2016; Wood, 2016).

Since the IE in our model consists of the integral of the difference of $F_1\{X(s),s\}$ and $F_1\{X^*(s),s\}$, we choose to model $X^*(\cdot)$ as a baseline and set it equal to the population mean of the covariate, in this case $X^*(s) = X_{\text{base}}(s) = 0$. We then look at the IE of a new covariate that is moved $\kappa$ standard deviations in the direction of the first FPC, where $\kappa = 1, 2, \text{ or } 3$. Therefore, $X_\kappa(s) = X_{\text{base}}(s) + \kappa\{\sqrt{\lambda_1}\phi_1(s)\}$. In our simulation study, $\lambda_1 = 9$ and $\phi_1(s) = 1$. We transform $X_{\text{base}}(s), X_1(s), X_2(s)$, and $X_3(s)$ using the center and scale method with the population mean and variance. Using the FGAM fit from equation (5.1), we estimate $\hat{M}_{X_{\text{base}}}, \hat{M}_{X_1}, \hat{M}_{X_2}$, and $\hat{M}_{X_3}$. Subtracting $\hat{M}_{X_\kappa} - \hat{M}_{X_{\text{base}}}$ gives $\int_S [\hat{F}_1\{X_\kappa(s),s\} - \hat{F}_1\{X_{\text{base}}(s),s\}]ds$, which we then multiply by $\hat{\beta}_M$ to estimate the IE.

As we discuss in Section 5.2, we compare the FGAM results to those using a FLM. We use 40 cubic P-splines with a second order difference penalty to model $\beta_1(\cdot)$ and $\beta_2(\cdot)$ in equation (5.3) and equation (5.4). In order to estimate the IE, we use the FLM fit from equation (5.3) to estimate $\hat{M}_{X_{\text{base}}}, \hat{M}_{X_1}, \hat{M}_{X_2}$, and $\hat{M}_{X_3}$ and follow a similar procedure as detailed above. All of the other settings remain the same as in the FGAM fit.

In addition, we also consider our nonlinear functional mediation model in the case of a binary response. In order to keep $P(Y_i = 1)$ away from its boundaries of zero and one, we alter the coefficient functions we investigate. Due to the odd nature of Nonlinear (1) and issues with $P(Y_i = 1)$, we do not include it in our simulation study for the case of a binary response. We modify Nonlinear (2) by dividing it by five and Linear (1) by dividing it by 10 in order to reduce $P(Y_i = 1)$ to reasonable values. Linear (2) remains the same. We denote the modified $F(\cdot, \cdot)$’s for the binary response case as

Nonlinear (2*) : $F(x, s) = (1 + x + s + 2x^2s)/5$;

Linear (1*) : $F(x, s) = (1 + x + s)/10$;

Linear (2) : $F(x, s) = xs$. 

61
The rest of the set-up for the binary response case is the same as the Gaussian response case detailed above.

### 5.3.2 Estimation Results

We examine the estimation of the IE by looking at the bias, mean squared error (MSE), and 95% coverage. Table 5.1 presents the estimation results for the IE for all four $F(·, ·)$’s when the response variable is Gaussian. The top table gives the results when a FGAM is fit, and the bottom table shows the results when a FLM is fit. When a FGAM is fit, as $\kappa$ increases, the bias and MSE also increase, whereas when the sample size increases, bias and MSE decrease. There are a few cases where coverage is slightly low when $n = 300$, but coverage increases when $n = 500$. When a FLM is fit, as expected, the two linear cases continue to perform well. In fact, the results show a FGAM performs just as well as a FLM when the truth is linear, although a FLM has lower MSEs than a FGAM, specifically when $\kappa = 3$, but overall nothing is lost by modeling a linear function nonlinearly. On the other hand, both of the nonlinear functions perform extremely poorly. Bias and MSE are very large and coverage is almost nonexistent. A FLM cannot handle the nonlinear nature of the functions.

Table 5.2 displays the results for the three $F(·, ·)$’s we investigate when the response variable is binary. The top table gives the results when a logistic FGAM is fit, and the bottom table shows the results when a logistic FLM is fit. The results paint a similar picture as in the Gaussian response case. When a FGAM is fit, bias and MSE increase as $\kappa$ increases but decrease as sample size increases. Also, for the few cases when coverage is slightly low when $n = 300$, it increases, albeit still a little low, when $n = 500$. Again, a FGAM and FLM perform similarly in the two linear cases, with a FLM having slightly lower MSE when $\kappa = 3$. However, the nonlinear case performs poorly when a FLM is fit. Clearly, a FGAM is desirable when the function is of a nonlinear nature and performs just as well as a FLM when the function is in fact linear.

Figure 5.1 and Figure 5.2 display the estimation curves of the IE for the nonlinear and linear cases, respectively, of $F(·, ·)$ for a Gaussian response for both a FGAM fit and a FLM fit when $n = 500$. They show the empirical mean of the $N$ IE estimates and the truth, along with a 95% empirical confidence
interval, for each level of $\kappa$. Clearly, as $\kappa$ increases, the variance, and therefore the length of the confidence interval, also increases. Also, the FLM fits for the two nonlinear functions are distinctly poor.

### 5.4 Application to Multiple Myeloma and Copy Number Variation Data

We apply causal nonlinear functional mediation analysis using FGAMs to the multiple myeloma (MM) data introduced in Section 3.3. The data consist of over 17,000 gene expression and CNV profiles for 235 individuals with MM. We wish to investigate the IE of gene expression on the relationship between functional CNV data and $\beta_2$ microglobulin ($\beta_2 M$), a known biomarker for MM, where we allow the covariate CNV data to be modeled non-linearly using FGAMs. We adjust for the effect of two demographic covariates, age and gender, which does not change the derivation of the IE or the estimation procedure. Because we model CNV as a functional covariate, any gene that has fewer than 20 associated CNV probes is dropped from the analysis, leaving 1,463 genes. In addition, only 172 patients have a non-missing $\beta_2 M$ measurement. Thus, for each of the 172 patients, we have 1,463 gene expressions and each gene expression has a corresponding CNV profile with at least 20 measurements. We denote $\beta_2 M$ as $Y_i$, the response, and gene expression as $M_i$, the mediator. We assume that $X_i(\cdot)$ is the random process that produces the observed CNV profiles that we model as the functional covariate. We note that this application is for demonstrative purposes only.

We apply our FGAM estimation procedure to all 1,463 genes to investigate the IE of each gene. We first transform the CNV data using the FPCA center and scale method we explain earlier, setting the proportion of variance explained equal to 95%, and perform mediation analysis using FGAMs. We use tensor product cubic P-splines with a second order difference penalty and set $K_x = K_s = 7$. We use the FPCA results to estimate the covariate mean, eigenvalues, and eigenfunctions in order to create $X_{\text{base}}(s), X_1(s), X_2(s)$, and $X_3(s)$. We transform them and predict $\hat{M}_{X_{\text{base}}}, \hat{M}_{X_1}, \hat{M}_{X_2}$, and $\hat{M}_{X_3}$ using the original FGAM fit. From there, we can estimate the IE, and its variance, for $\kappa = 1, 2, 3$, corresponding to a new CNV profile that is moved 1, 2, or 3 standard deviations in the direction of the first eigenfunction compared to the overall mean. We note that we set the gender equal to male and age equal to 60 years old,
Table 5.1: Bias, MSE, and 95% coverage for the IE when the response variable is Gaussian. Coverages are given in terms of percent. The mean standard error for the estimates in this table is 0.00003.

<table>
<thead>
<tr>
<th>FGAM</th>
<th>$n = 300$</th>
<th>$n = 500$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\kappa$</td>
<td>Bias</td>
</tr>
<tr>
<td></td>
<td>$\kappa = 1$</td>
<td>0.021</td>
</tr>
<tr>
<td></td>
<td>$\kappa = 2$</td>
<td>0.144</td>
</tr>
<tr>
<td></td>
<td>$\kappa = 3$</td>
<td>0.401</td>
</tr>
<tr>
<td></td>
<td>$\kappa = 1$</td>
<td>0.068</td>
</tr>
<tr>
<td></td>
<td>$\kappa = 2$</td>
<td>0.114</td>
</tr>
<tr>
<td></td>
<td>$\kappa = 3$</td>
<td>-0.247</td>
</tr>
<tr>
<td></td>
<td>$\kappa = 1$</td>
<td>-0.008</td>
</tr>
<tr>
<td></td>
<td>$\kappa = 2$</td>
<td>-0.011</td>
</tr>
<tr>
<td></td>
<td>$\kappa = 3$</td>
<td>-0.010</td>
</tr>
<tr>
<td></td>
<td>$\kappa = 1$</td>
<td>-0.004</td>
</tr>
<tr>
<td></td>
<td>$\kappa = 2$</td>
<td>-0.003</td>
</tr>
<tr>
<td></td>
<td>$\kappa = 3$</td>
<td>-0.001</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>FLM</th>
<th>$n = 300$</th>
<th>$n = 500$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\kappa$</td>
<td>Bias</td>
</tr>
<tr>
<td></td>
<td>$\kappa = 1$</td>
<td>0.269</td>
</tr>
<tr>
<td></td>
<td>$\kappa = 2$</td>
<td>1.036</td>
</tr>
<tr>
<td></td>
<td>$\kappa = 3$</td>
<td>2.183</td>
</tr>
<tr>
<td></td>
<td>$\kappa = 1$</td>
<td>-0.891</td>
</tr>
<tr>
<td></td>
<td>$\kappa = 2$</td>
<td>-4.422</td>
</tr>
<tr>
<td></td>
<td>$\kappa = 3$</td>
<td>-10.591</td>
</tr>
<tr>
<td></td>
<td>$\kappa = 1$</td>
<td>0.008</td>
</tr>
<tr>
<td></td>
<td>$\kappa = 2$</td>
<td>0.016</td>
</tr>
<tr>
<td></td>
<td>$\kappa = 3$</td>
<td>0.024</td>
</tr>
<tr>
<td></td>
<td>$\kappa = 1$</td>
<td>0.003</td>
</tr>
<tr>
<td></td>
<td>$\kappa = 2$</td>
<td>0.005</td>
</tr>
<tr>
<td></td>
<td>$\kappa = 3$</td>
<td>0.008</td>
</tr>
</tbody>
</table>
Table 5.2: Bias, MSE, and 95% coverage for the IE when the response variable is binary. Coverages are given in terms of percent. The maximum standard error for the estimates in this table is 0.00002.

<table>
<thead>
<tr>
<th></th>
<th>FGAM</th>
<th></th>
<th></th>
<th>FLM</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$n = 300$</td>
<td>$n = 500$</td>
<td>$n = 300$</td>
<td>$n = 500$</td>
<td>$n = 500$</td>
</tr>
<tr>
<td>$\kappa$</td>
<td>Bias</td>
<td>MSE</td>
<td>95% Cov.</td>
<td>Bias</td>
<td>MSE</td>
<td>95% Cov.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nonlinear (2*) $\kappa = 1$</td>
<td>0.027</td>
<td>0.041</td>
<td>96.8</td>
<td>0.038</td>
<td>0.027</td>
<td>97.0</td>
</tr>
<tr>
<td>$\kappa = 2$</td>
<td>-0.014</td>
<td>0.252</td>
<td>89.9</td>
<td>0.021</td>
<td>0.136</td>
<td>93.5</td>
</tr>
<tr>
<td>$\kappa = 3$</td>
<td>-0.279</td>
<td>1.031</td>
<td>86.1</td>
<td>-0.206</td>
<td>0.561</td>
<td>90.6</td>
</tr>
<tr>
<td>Linear (1*) $\kappa = 1$</td>
<td>0.000</td>
<td>0.019</td>
<td>96.6</td>
<td>0.001</td>
<td>0.011</td>
<td>95.2</td>
</tr>
<tr>
<td>$\kappa = 2$</td>
<td>-0.004</td>
<td>0.096</td>
<td>95.9</td>
<td>-0.001</td>
<td>0.052</td>
<td>94.7</td>
</tr>
<tr>
<td>$\kappa = 3$</td>
<td>-0.012</td>
<td>0.302</td>
<td>95.7</td>
<td>-0.005</td>
<td>0.156</td>
<td>94.6</td>
</tr>
<tr>
<td>Linear (2) $\kappa = 1$</td>
<td>0.022</td>
<td>0.032</td>
<td>95.1</td>
<td>0.006</td>
<td>0.016</td>
<td>95.1</td>
</tr>
<tr>
<td>$\kappa = 2$</td>
<td>0.041</td>
<td>0.150</td>
<td>89.1</td>
<td>0.010</td>
<td>0.073</td>
<td>90.5</td>
</tr>
<tr>
<td>$\kappa = 3$</td>
<td>0.054</td>
<td>0.427</td>
<td>87.8</td>
<td>0.012</td>
<td>0.203</td>
<td>90.1</td>
</tr>
</tbody>
</table>
Figure 5.1: The empirical mean and 95% confidence interval of the IE for both a FGAM fit and a FLM fit for the two nonlinear functions we investigate for a Gaussian response when $n = 500$ for varying levels of $\kappa$. 

66
Figure 5.2: The empirical mean and 95% confidence interval of the IE for both a FGAM fit and a FLM fit for the two linear functions we investigate for a Gaussian response when \( n = 500 \) for varying levels of \( \kappa \).
the mean of the covariate age in our dataset. However, again, the covariates age and gender do not affect the calculation of the IE and therefore, are irrelevant. If the model included the interaction of any of these covariates with $X_i(\cdot)$ or $M_i$, then the derivation of the IE would be affected and the specific covariate settings would matter.

In order to determine which genes have a relevant IE, we use an ad hoc testing procedure as there are not yet any formal tests in this case. We standardize the IEs by taking their absolute value, dividing them by their standard error, and ranking them in decreasing order to see which have the largest effect. In total, we have 4,389 IEs (1,463 genes × three levels of $\kappa$). Using an experiment-wise significant level of $\alpha = .05$, a two-sided test, and Bonferroni’s conservative correction for multiple comparisons, we use a cutoff value of $Z_{\alpha^{*}/2} = 4.389$, where $\alpha^{*} = .05/4,389$. When $\kappa = 1$, none of the 1,463 genes have a significant IE. The gene with the largest standardized IE is NDUFAF2 with 3.509. When $\kappa = 2$, 11 genes have a standardized IE larger than the cut-off value. These genes are presented in Table 5.3. Lastly, when $\kappa = 3$, the same 11 genes in Table 5.3 plus 17 more genes, totaling 28 genes, have a significant IE.

Clearly, we question the validity of this provisionary test, but we wanted some way to choose a few top genes to possibly investigate further. We focus on the 11 significant genes when $\kappa = 2$. Figure 5.3 shows the standardized estimates of the IE for each level of $\kappa$ for these 11 genes, with a cutoff line at 4.389 to help gauge how the IE changes as $\kappa$ changes. While for each gene the standardized IE increases as $\kappa$ increases, it appears to start to stabilize, although this observation is only based on three levels of $\kappa$.

We also apply the FLM with splines mediation framework to the CNV data in order to compare the above FGAM results with their linear counterparts. We model the coefficient functions using 40 cubic P-splines with a second order difference penalty. After standardizing and ranking the IE estimates, when $\kappa = 1$, none of the 1,463 genes have a significant IE. Just as in the FGAM case, the gene with the largest standardized IE is NDUFAF2 with 3.300. When $\kappa = 2$, six genes have a standardized IE larger than the cut-off value of 4.389. These genes are presented in Table 5.4. Five out of six of the genes in this case are also included in the 11 significant genes from the FGAM fit when $\kappa = 2$. Lastly, when $\kappa = 3$, the same six genes in Table 5.4 plus 11 more genes, totaling 17 genes, have a significant IE. Figure 5.4 shows the standardized estimates of the IE for each level of $\kappa$ for the six significant genes when $\kappa = 2$ for the
Table 5.3: The 11 genes with a standardized IE estimate above the cut-off value of 4.389 when $\kappa = 2$ from a FGAM fit.

| Gene     | $|\hat{IE}|/\hat{\text{se}}(\hat{IE})$ |
|-----------|----------------------------------|
| NDUFAF2   | 5.818                            |
| BCKDHB    | 5.151                            |
| FTO       | 5.033                            |
| RNF115    | 4.898                            |
| DYM       | 4.822                            |
| ELP4      | 4.799                            |
| MSH3      | 4.602                            |
| RPS6KC1   | 4.523                            |
| OTUD7B    | 4.470                            |
| POLR3C    | 4.402                            |
| MNAT1     | 4.390                            |

FLM fit, along with a cutoff line at 4.389. Again, while for each gene the standardized IE increases as $\kappa$ increases, the increase appears to lessen.

We briefly compare the CNV application results from a FGAM and FLM with splines to those from a FLM using FPCR from Section 3.3. As shown in Section 3.3, fitting a FLM using FPCR to the MM data and calculating the IEF using the standard product of coefficients (POC) method results in 11 significant genes using the Wald test with no multiple comparison correction. Only two of those genes are also significant in the FGAM fit when $\kappa = 2$, RNF115 and POLR3C. An additional five genes from the results in Section 3.3 are significant in the FGAM fit when $\kappa = 3$. None are in the top 17 genes from the FLM with splines fit when $\kappa = 3$. It is important to note, however, that the methods from Section 3.3 are not based on a counterfactual framework, but rather they are based on the standard POC method. It could be of interest to further investigate the difference in results between fitting a counterfactual model and a standard POC model.

5.5 Summary

In this chapter, we explore causal nonlinear functional mediation analysis using a counterfactual approach. We use a FGAM to allow the functional covariate to be modeled nonlinearly. We use a bivariate spline...
Figure 5.3: Standardized estimates of the IE for the 11 genes given in Table 5.3 from a FGAM fit for varying levels of $\kappa$.

Table 5.4: The six genes with a standardized IE estimate above the cut-off value of 4.389 when $\kappa = 2$ from a FLM fit.

| Gene       | $|\text{IE}|/\hat{s}_\text{e}(\text{IE})$ |
|------------|---------------------------------|
| NDUFAF2    | 4.941                           |
| FTO        | 4.780                           |
| ELP4       | 4.687                           |
| FAF1       | 4.666                           |
| MSH3       | 4.460                           |
| BCKDHB     | 4.398                           |
Figure 5.4: Standardized estimates of the IE for the six genes given in Table 5.4 from a FLM fit for varying levels of $\kappa$.

model in order to estimate the functional coefficient surface. We also compare the estimation results from a FGAM to those from a FLM with splines through numerical simulations. As expected, a FGAM significantly outperforms a FLM when the function is in fact nonlinear. When the function is linear, the two methods, FGAM and FLM, perform similarly. We also apply both the nonlinear causal methodology using FGAMs and the linear causal methodology using FLMs with splines to the same genetics application regarding MM as in the standard FLM using FPCR case. While we are able to use the methodology to estimate the IE, the results presented here are only for demonstrative purposes.
6.1 Summary

In this dissertation, we integrate the theories of mediation analysis and functional regression methods to investigate the IE of a scalar mediator on the relationship between a functional covariate and a scalar response. In Chapter 3, we consider Gaussian and generalized responses using generalized functional linear models. We use the standard approach to mediation analysis and define the IEF using the POC method. We use FPCR to model the mediation equations and estimate the IEF. We also propose three testing procedures: a Wald test, a bootstrap test, and a simultaneous confidence interval test. Through numerical simulations, we show our method properly estimates the IEF while the results of the three testing procedures vary. We suggest using the Wald test based on Sobel’s variance approximation when $\beta_M \neq 0$ as it has proper Type 1 error and good power. The simultaneous interval test also has proper Type 1 error but less power than the Wald test. The bootstrap test, though, has much too high Type 1 error. However, the Wald test and interval test both assume the POC is approximately normally distributed when it is not guaranteed to be. In addition, when $\beta_M = 0$, all of the tests have extremely conservative Type 1 error.
error. We further discuss these issues in testing below in Section 6.2. Lastly, we apply our methodology to a dataset regarding patients with multiple myeloma (MM). We investigate the IE of gene expression on the relationship between functional CNV data and $\beta_2$ microglobulin, a known biomarker for MM.

In Chapter 4, we use a counterfactual approach to mediation analysis to analyze the IEF when the response is time-to-event. The counterfactual approach to mediation is largely preferred over the standard approach as it allows for a mediator-covariate interaction, a nonlinear model, and a causal interpretation. Here, we explore an exponential AFT model, a lognormal AFT model, and a Cox proportional hazards model in the context of FPCR in order to estimate and test the presence of an IE. We show that while an AFT model properly estimates the IEF, even under misspecification, a Cox model reports a highly biased estimate and should not be used for estimation purposes. This is because the derivation of the IE under a Cox model only holds when the cumulative baseline hazard function is small, i.e. the outcome is rare. Therefore, we do not suggest using a Cox model when an estimate of the IEF is desired. We also examine three testing procedures of the IEF. Through numerical simulations, we conclude the Wald test based on Sobel’s variance approximation renders an adequate test for both the AFT and Cox models when $\psi_M \neq 0$, even if the parametric AFT model is misspecified. The simultaneous interval test also provides a proper test but has less power than the Wald test. The bootstrap test, however, results in much too high Type 1 error and should not be used. Again, the Wald test and simultaneous interval test assume the IE tends to be normally distributed. In addition, when $\psi_M = 0$, none of the testing procedures have proper Type 1 error, and we do not suggest using any of them.

Lastly, in Chapter 5, we consider a nonlinear mediation model in a counterfactual framework where we allow the functional covariate to be modeled nonlinearly. We use a FGAM and bivariate spline basis system in order to estimate the functional coefficient surface. We also compare the estimation results from a nonlinear FGAM to those from a FLM with splines. Unsurprisingly, a FGAM significantly outperforms a FLM when the function is in fact nonlinear. When the function is linear, the two methods perform similarly. We also apply both our nonlinear methodology using FGAMs and the linear methodology using FLMs with splines to the genetic CNV application regarding MM.
6.2 Future Work

This research is an introduction to functional mediation analysis; there is still much work to be done in the field. As discussed above, testing in mediation analysis is still problematic. First, the main variance estimator used is based on Sobel’s variance approximation for the product of two coefficients, which assumes the POC tends to be normally distributed when the sample size is large. However, the product of two normal quantities is not guaranteed to be normal. Both the Wald test and simultaneous confidence interval test utilize this variance estimator and assume the POC is normally distributed. While this does not seem to cause any issues in our simulation study when \( \beta_M \neq 0 \) (or \( \psi_M \neq 0 \) in the survival response case), it still is not optimal. Further research needs to be done to develop a proper variance estimator and test that does not assume normality of the POC. We propose trying to modify the bootstrap test so that it results in nominal Type 1 error.

Another issue in testing in mediation analysis occurs when \( \beta_M = 0 \) (or \( \psi_M = 0 \)), resulting in poor Type 1 error. We focus on the standard functional mediation case with a Gaussian response. In our simulation study in Section 3.2, Type 1 error is conservative for all of the tests we consider when \( \beta_M = 0 \). Table 6.1 displays the Type 1 error results when \( \beta_M = 0 \) for the Wald test, simultaneous interval test, and bootstrap test for a standard mediation model using the POC method when the response variable is Gaussian and \( n = 100 \). Clearly, the Wald test and interval test have awfully conservative Type 1 error. The bootstrap test also has very conservative Type 1 error, although it is slightly higher. Therefore, we cannot believe any test to be proper when \( \beta_M = 0 \). As discussed in Section 3.2.3, we investigate the distribution of each component of the IE vector \( \delta_{Kn} \) in several cases. We note that when \( \beta_M \neq 0 \), the distributions are relatively normal, even when \( \gamma_{Kn} = 0 \). However, when \( \beta_M = 0 \), the distributions are highly concentrated around zero with little to no variation. Similarly, we explore the bootstrap distributions of \( \hat{\delta}_{Kn} \) in several situations. Again, when \( \beta_M = 0 \), the bootstrap distributions are highly concentrated around zero, and when \( \beta_M \neq 0 \), the bootstrap distributions are relatively normal. These results may help explain the conservative Type 1 error when \( \beta_M = 0 \). Baron and Kenny (1986) note that for mediation to exist, \( \beta_M \) should not equal zero; otherwise the mediator has no effect on the response when the covariate is in the model. While this
Table 6.1: Type 1 error results when $\beta_M = 0$ for the Wald test, simultaneous interval test, and bootstrap test for a standard mediation model using the POC method when the response variable is Gaussian and $n = 100$ for significance level $\alpha = 0.05$.

<table>
<thead>
<tr>
<th>$\beta_M = 0$</th>
<th>Wald Test</th>
<th>Interval Test</th>
<th>Bootstrap Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma_{K_0} = 0.00$</td>
<td>0.000</td>
<td>0.000</td>
<td>0.001</td>
</tr>
<tr>
<td>$\gamma_{K_0} = 0.05$</td>
<td>0.000</td>
<td>0.000</td>
<td>0.004</td>
</tr>
<tr>
<td>$\gamma_{K_0} = 0.10$</td>
<td>0.000</td>
<td>0.000</td>
<td>0.019</td>
</tr>
<tr>
<td>$\gamma_{K_0} = 0.15$</td>
<td>0.001</td>
<td>0.000</td>
<td>0.026</td>
</tr>
<tr>
<td>$\gamma_{K_0} = 0.20$</td>
<td>0.001</td>
<td>0.001</td>
<td>0.030</td>
</tr>
</tbody>
</table>

is generally accepted by most scholars, it is still greatly preferable to establish a proper test of the IE that incorporates the case when $\beta_M = 0$. This issue in testing still needs to be heavily researched.

In addition, we only consider the case when there is assumed to be no mediator-covariate interaction. The addition of an interaction term between the mediator and covariate is well-studied using a counterfactual model in scalar mediation analysis and should be investigated in functional mediation analysis. Consider the following functional mediation equations where we include the mediator-covariate interaction for the case of a Gaussian response:

$$
M_i = \gamma_0 + \int_{\mathcal{X}} X_i(s) \gamma_1(s) ds + \eta_i;
$$

$$
Y_i = \beta_0 + \int_{\mathcal{X}} X_i(s) \beta_1(s) ds + M_i \beta_M + M_i \int_{\mathcal{X}} X_i(s) \beta_2(s) ds + \zeta_i.
$$

Following the counterfactual definition of the IE, we have

$$
\text{IE} = E(Y_{iM_i} - Y_{iM_i^*})
$$

$$
= E(M_i)\beta_M + E(M_i) \int_{\mathcal{X}} X(s) \beta_2(s) ds - E(M_i^*)\beta_M - E(M_i^*) \int_{\mathcal{X}} X(s) \beta_2(s) ds
$$

$$
= \{\beta_M + \int_{\mathcal{X}} X(s) \beta_2(s) ds\} E(M_i - M_i^*)
$$

$$
= \{\beta_M + \int_{\mathcal{X}} \beta_2(s) X(s) ds\} \int_{\mathcal{X}} \gamma_1(s) \{X(s) - X^*(s)\} ds
$$

$$
= \int_{\mathcal{X}} \{\beta_M \gamma_1(s) + \beta_2(s) \gamma_1(s) X(s)\} \{X(s) - X^*(s)\} ds.
$$
As expected, this IE is comprised of the IE when the interaction is not included, \( \beta_M \int_{\mathcal{R}} \gamma_1(s) \{ X(s) - X^*(s) \} ds \), plus an additional term, \( \int_{\mathcal{R}} X(s)\beta_2(s)\gamma_1(s) \{ X(s) - X^*(s) \} ds \), leading to a much more complicated formula. The IE with a generalized or survival response is similar. There is much research to be done under a functional mediation model with a mediator-covariate interaction.

Finally, here are a few last points of interest for future research in functional mediation analysis. In this dissertation, we only consider the covariate to be functional. Accounting for a functional mediator and a functional response is also of interest, along with more complicated pathways that include multiple mediators. In Chapter 3 and Chapter 4, we choose to use FPCR to estimate the IEF, but other estimation schemes, such as spline based methods, may be used. It could also be of interest to study the difference in results under a standard mediation framework versus a counterfactual model. Lastly, we desire to find more suitable applications for the methods described in this literature to better showcase their capabilities and strengthen their validity.
REFERENCES


