Contributions to Nonparametric Quantile Analysis and Quantile-Based Mediation Analysis, With Applications to Lifecourse Analysis in Human Biology

by

Sanjana Gupta

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in the University of Michigan
2022

Doctoral Committee:

Professor Kerby Shedden, Chair
Assistant Professor Snigdha Panigrahi
Assistant Professor Jeffrey Regier
Professor Beverly I Strassmann

Sanjana Gupta
gsan@umich.edu
ORCID iD: 0000-0003-2916-0350

© Sanjana Gupta 2022

DEDICATION

To my grandmother, dadima.

ACKNOWLEDGMENTS

I would like to thank my advisor, Kerby Shedden, for his endless patience, guidance and support without which I can not imagine having completed my thesis. Kerby's work ethic, vast and deep knowledge of different fields of Statistics, along with his positive attitude never cease to amaze me and will always inspire me to grow both professionally and personally.

I would also like to thank my family for always encouraging me to pursue my academic persuits. Thank you to my parents for their unconditional love, and for standing by me through thick and thin. Thank you to my biggest support, best friend, confidant, and my personal cheerleader - my sister; I can not imagine making it through anything in life without you. My brother who has been a constant source of moral support and funny anecdotes from his own PhD journey, thank you for always being there. Friends who are like family - thank you to Sukhman Kaur and Anviti Mohan for living through these five years with me and supporting me across continents.

I am very fortunate to have made some amazing friends in Ann Arbor. Thank you to Olivia Hackworth for being my pillar in Ann Arbor, and for sending me care packages when I needed them the most. Thank you to Caleb Ki, Byoungwook Jang, Dan Kessler, and everyone else in my cohort for embarking on this journey with me; I hope we continue to make new memories together.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE

LIST OF TABLES

TABLE

LIST OF APPENDICES

ABSTRACT

This thesis develops and assesses new ways to study the conditional quantiles of a population using a sample of data that represents the population. All methods presented here build on a recently-proposed non-parametric approach to quantile regression that is analogous to local linear regression in the least-squares setting. A major challenge is that the raw local quantile estimates are cumbersome to interpret and gain insight from directly. Aiming to overcome this challenge, there are four main contributions herein. First, we demonstrate how a low-rank additive regression analysis can produce insight into a collection of local nonparametric quantile estimates. The low rank structure regularizes the noisy quantile estimates and facilitates interpretation of the findings. Second, we show how a multivariate dimension reduction approach provides a different type of insight into a collection of estimated conditional quantile functions. The third contribution of the thesis leverages the combination of nonparametric quantile estimation and low-rank regression in the context of mediation analysis. We show that this produces a novel quantile-based approach to mediation analysis that expresses direct and indirect effects in a concise and interpretable way. The final methodological contribution of the thesis is a framework for moment-based estimation of conditional covariance functions for stochastic processes. Throughout the thesis, we motivate our work through analyses looking at the proximal and distal factors predicting human blood pressure.

CHAPTER 1

Introduction

This thesis contributes novel and practical statistical methodology that can be used to assess associations between multiple explanatory variables and the quantile structure of a single quantitative response variable. We focus on both surface-level "regression effects" as well as on attaining more mechanistic insights by considering hypothesized mediating relationships involving the set of observed covariates. Quantile analysis is a central theme, as we argue that focusing only on conditional means and conditional variances, as in linear and generalized linear modeling, fails to capture the full set of relationships between explanatory variables and outcomes. While this potential failing has been noted before, challenges remain in terms of practical methodology for exploiting quantile structure, especially when moving beyond parametric linear models, and when considering mechanistic hypotheses involving mediation.

1.1 Motivating Scientific Application

The methodological work in this thesis was motivated by a long line of research in human biology considering the relationship between anthropometry (body size measures such as BMI and height) and blood pressure. This is a setting where it is especially natural to consider the full distribution of a response as opposed to considering only its conditional mean, as we are especially interested in the upper quantiles of the conditional blood pressure distribution since this part of othe distribution corresponds to the greatest risk for adverse health outcomes. As in much previous work, we are interested in anthropometry as a primary risk factor (predictor) for elevated blood pressure. We argue that quantile analysis, and especially nonparametric quantile analysis may prove to be fruitful, as it is unclear whether the effects of variables such as BMI and height are equal at different quantiles, or are linear throughout the their range.

The Dogon Longitudinal Study

Throughout this thesis, we will use data from the Dogon Longitudinal Study, led by Professor

Beverly Strassmann [91] and based at the University of Michigan. This is a longitudinal cohort study following (initially) around 1,800 children from the Dogon ethnic group in Mali, who were born between the years 1992 and 1998. These children, referred to as the "F1 cohort", are now adults, and many of the F1 females have given birth to children of their own, referred to as the "F2 cohort". The study also captures limited, mostly retrospective data from the parents of the F1 cohort, referred to as the "F0 cohort". The dataset contains roughly annual observations for the F1 cohort children beginning shortly after birth. However most individuals do not have highly regular observations, which complicates the analysis. A vast set of measurements were collected at each F1 assessment. Here we focus in particular on anthropometry measures including height, weight, body mass index (BMI), weight-for-age Z-score (WAZ), height-for-age Z-score (HAZ), and BMI-for-age Z-score (BAZ). Our key outcome variable is systolic blood pressure (SBP), which is measured from age 11. For further details about the study and data please refer to [91].

The Dogon are a relatively isolated ethnic group living in a rural region of Mali. The Dogon traditionally support themselves by subsistence farming, although in recent years substantial numbers of younger Dogon have moved to cities including to the capital city Bamako, where they may engage in other forms of work. Undernutrition is common in the Dogon population, but exposure to undernutrition varies between individuals. A primary reason for interest in this population is that important research questions for the field of human biology center on the consequences of undernutrition in early life for adult metabolic and cardiovascular health . This is the primary substantive question addressed by the methods developed in this thesis.

We use systolic blood pressure (SBP) as an indicator of adult cardiovascular health, and our main goal is to explain variation in this measure. Although the F1 cohort is currently too young to exhibit much cardiovascular morbidity or disease, elevated systolic blood pressure (SBP) in young adults (in their late teens and twenties) may track to cardiovascular problems later in life. Cross-sectional relationships between anthropometry and SBP are well-established – taller and heavier individuals tend to have greater current SBP, although the nature of this relationship may differ based on genetic and environmental factors. More complex is the role of early life exposures including undernutrition, and its observable surrogates captured through the anthropometry measure that are collected as part of this study.

A basic hypothesis is that childhood growth relates to adult SBP via two antagonistic routes. First, larger children "track" into being larger adults, and larger adults tend to have greater SBP. In this way, well-nourished children are at greater risk for elevated adult SBP. However, well-nourished children also have the capacity to develop healthy tissue and organ systems, which may protect them from having high SBP as adults. One possible mechanism at work here is that undernourished children may be at risk for having under-developed vascular and renal tissues, putting them at risk for elevated SBP regardless of body size.

These antagonistic relationships have been extensively considered in previous work with other populations, leading to some controversies about proper analysis and interpretation of the data. A so-called "reversal paradox" has been debated, in which early-life body size x is positively associated with adult SBP y in marginal terms, but this becomes an inverse association when conditioning on adult body size z. The paradox is resolved by recognizing that it may result from a specific set of causal relationships among x, z, and y. This issue is discussed in more detail in chapter 3 using methods of mediation analysis. In this chapter, we always condition on adult anthropometry variables (z) and interpret the association of childhood body size with adult SBP as a "direct effect" that controls for body size tracking. We note that these issues seem mainly to have been considered through the conditional mean, making our analysis involving conditional quantiles a novel contribution to this debate.

The National Health and Nutrition Examination Survey

We also use data from the 2015-2016 wave of the National Health and Nutrition Examination Survey (NHANES) [14], [2] to benchmark our methods. NHANES is a high-quality study run by the United States Centers for Disease Control (CDC), conducted in waves over many decades. NHANES cross-sectionally surveys a probability sample of the US population, with the overall goal of assessing the health and nutritional status of people within the United States of all ages. NHANES is not longitudinal, so cannot be used to address the role of early-life exposures in determining adult outcomes. However, NHANES has a larger sample size compared to the Dogon longitudinal study ($\approx 8,000$ versus $\approx 1,400$) and presents numerous opportunities to explore cross-sectional relationships. We use the NHANES study to assess and benchmark the findings of the frameworks developed here in a context where the ground truth is very well explored.

1.2 Literature review of quantile regression

Quantile regression is an extremely vast topic. It is outside the scope of this study to provide an extensive literature review of this topic. For such reviews, some references include [53], [52]. Here we provide a brief review of techniques relevant to the methods explored in this thesis.

Linear Quantile Regression

Linear Quantile regression is an effective analytic technique since it can drive inferences about individuals in the tails of the population, and can focus on properties of the response beyond its

central tendency. The linear quantile function can be written as follows

$$Y_i = \boldsymbol{X}_i'\boldsymbol{\beta}_{pi} + \epsilon_{pi}, \;\; \epsilon_{pi} \sim H_{pi}, \;\; i = 1, \cdots, N, \; 0 < p < 1.$$

Here H_{pi} is some arbitrary distribution such that $H_{pi}(0) = p$, ie the distribution function is zero equals p. This parametric model describes the conditional qunatile function of the outcome Y_i conditional on covariate(s) $\boldsymbol{X}_i$ at given quantile $p \in [0, 1]$ as

$$Q_{Y_i}(p|\boldsymbol{X}_i) = F^{-1}_{Y_i}(p|\boldsymbol{X}_i) = \boldsymbol{X}_i'\boldsymbol{\beta}_{pi} + \epsilon_{pi}, \;\; \text{with } Q_{\epsilon_{pi}} \sim F_{pi}.$$

Additive Quantile Regression

A semi-parametric extension of linear quantile regression to include non-linear terms led to additive quantile regression (AQR) [95]. A brief review of this can be found in [53]. In the classical additive models, each covariate contributes through its own functions additively. The model is obtained by minimizing

$$\arg\min_{\boldsymbol{\beta}_p} \sum_{i=1}^{N} \rho_p(y_i - Q_{Y_i|\boldsymbol{X}_i,\boldsymbol{Z}_i}(p|\boldsymbol{x}_i, \boldsymbol{z}_i)),$$

and has the general form

$$Q_{Y_i|\boldsymbol{X}_i,\boldsymbol{Z}_i}(p|\boldsymbol{x}_i) = \sum_{j=1}^{J} f_{pj}(\boldsymbol{x_i}),$$

where $\boldsymbol{x_i} \in \mathbb{R}^J$. In these methods, the non-linear terms are generally estimated non-parametrically. This is feasible because the additive structure of the model restricts the non-parametric components to be low-dimensional, which helps overcome some aspects of curse of dimensionality. This was extended by [51], who expanded the fitted functions into the sum of univariate and bivariate functions. The model proposed in [51] was of the form

$$Q_{Y_i|\boldsymbol{X}_i,\boldsymbol{Z}_i}(p|\boldsymbol{x}_i, \boldsymbol{z}_i) = \boldsymbol{x}_i^T\boldsymbol{\beta}_p + \sum_{j=1}^{J} f_{pj}(\boldsymbol{z}_i). \tag{1.1}$$

The additive model above consists of a linear term in $\boldsymbol{X}$ and a sum of non-linear terms formed of arbitrary functions $f_{pj}, \;\; j = 1, 2, \cdots, J$ of $\boldsymbol{Z}_i$. The vector $\boldsymbol{Z} = (Z_1, \cdots, Z_r)$ may consist of additional terms; f_{pj} could be any functions, such as a non-linear smooth function of some Z_i ie $f_{pj}(\boldsymbol{Z}_i) = f_{pj}(Z_i)$, or a bivariate function in $\boldsymbol{Z}$, ie $f_{pj}(\boldsymbol{Z}_i) = f_{pj}(Z_{i_{r_1}}, Z_{i_{r_2}})$.

Generally $f_{\cdot}'s$ are estimated via splines since they are non parametric. Quantile B-splines are not

used as they requires determining the optimal number of b-splines in the basis. Instead, smoothing splines are used along with a roughness penalty to prevent overfitting. The updated equation which is minimize to obtain the estimates is

$$\arg\min_{\boldsymbol{\beta}_p} \sum_{i=1}^{N} \rho_p\Big(Y_i - \boldsymbol{X}_i^T\boldsymbol{\beta}_p - \sum_{i=1}^{J}\big(f_{pi}(\boldsymbol{Z})\big)\Big) + \sum_{j=1}^{J}\lambda_j V(f'_{pj}), \tag{1.2}$$

where $V(f'_{pi})$ represents the total variation of f'_{pi}, such that $V(f'_{pi}) = \int |f''_{ip}(\boldsymbol{Z})| \cdot d\boldsymbol{Z}$. The smoothing penalties λ_j can be tuned using quantile regression adapted information criterion such as AIC and Schwarz-type (SIC) [51], [55], or via bootstrap [27]. Alternate ways to estimate $\{f_{pk}\}_{k=1}^{q}$ include backfitting [25], marginal integration [20], local polynomial splines [43].

Additive Quantile Regression for time-varying data

AQRs were originally extended to Longitudinal data by including fixed individual-specific intercept and slopes in [27]. They were further extended by including multiple correlated random effects to account for the individual-level correlation with a general covariance matrix, such that the model allows for automatic data-driven selection of the smoothing penalty [32]. The general form of the model is given below

$$Q_{y_{ik}|\boldsymbol{u}_i,\boldsymbol{x}_i,\boldsymbol{z}_i}(p) = \beta_{p,0} + \sum_{j=1}^{q} f_p^j(x_{ijk}) + z'_{ik}u_{p,i},$$

where $\boldsymbol{X}_i = \begin{bmatrix} —x'_{ij1}— \\ \vdots \\ —x'_{ijn_i}— \end{bmatrix} \in \mathbb{R}^{n_i\times q}$, $\boldsymbol{Z}_i = \begin{bmatrix} —z'_{ij1}— \\ \vdots \\ —z'_{ijn_i}— \end{bmatrix} \in \mathbb{R}^{n_i\times p}$ for the i^{th} subject. $f_p^j(\cdot)$ is a centered, p-specific, twice differentiable smooth function of the j^{th} component of $\boldsymbol{x}$, and $u_{p,i}$ is a $p \times 1$ vector of the random effect coefficients for z_{ik} and its distribution may contain p-specific parameters. In their paper, they consider a spline model for the non-linear functions in each component of $\boldsymbol{x}$. They further assume that $\boldsymbol{u}_{pi}$ and $\boldsymbol{v}_p$ follow zero-centered multivariate Gaussian distributions which are independent of each other. For further details please refer to [32]. An application of this method can be found in [100]. More on quantile mixed-effects models can be found in [30], [33], [56].

AQRs have also been developed for time series data, such as Generalized Autoregressive Conditional Heteroscedasticity (GARCH)-type models [96]. For time series data, another dimensional reduction technique with functional-coefficients is the smooth coefficients model [13]. This model assumes stationarity and has an additive form as follows

$$Q_p(\boldsymbol{X}_t, \boldsymbol{U}_t) = \sum_{j=1}^{J} a_{jp}(\boldsymbol{U}_t) Xj(t) = \boldsymbol{X}_t' \boldsymbol{a}_p(\boldsymbol{U}_t),$$

where $\boldsymbol{U}_t$, called the smoothing variable, can contain some of the covariates $\boldsymbol{X}_t = (X_1(t), \cdots, X_J(t))$, or time, or other exogenous or lagged variables. Most importantly, $\boldsymbol{a}_p(\cdot) = (a_{0p}(\cdot), \cdots, a_{Jp}(\cdot))'$, where some of the smooth coefficient functions $\{a_{jp}(\cdot)\}$ can be quantile-specific (dependent on p). The smoothing functions are estimated non-parametrically using a local polynomial fit. This setting covers many quantile regression models, such as autoregressive QAR, a specific class of autoregressive conditional heteroscedasticity (ARCH) models and the ordinary non-parametric quantile regression model if $J = 0$.

Non-parametric Quantile Regression

While linear quantile regression is easy to interpret and not intensive computationally, it is not flexible due to the strict linearity assumption. Non-parametric quantile regression has become a popular alternate to avoid restrictive parametric assumptions. [54] defined the class of Non-parametric Quantile Regression models as close to the linear quantile model as possible. They presented the following form

$$Q_{Y_i}(p|\boldsymbol{X}_i) = f(\boldsymbol{X}_i, \beta_i(p)),$$

where the p^{th} non-parametric quantile regression estimate is obtained by

$$\hat{\boldsymbol{\beta}}_i(p) = \arg\min_{\beta} \sum_{i=1}^{N} \rho_p(y_i - f(\boldsymbol{x}_i, \beta(p))).$$

Several techniques have been proposed for non-parametric quantile regression modelings, including smoothing techniques. Two broad categories of smoothing techniques are methods using splines and kernels. Details on different strategies employed to obtain estimates in non-parametric quantile regression can be found in many studies, including [19], [54].

Other models for non-parametric quantile regression include nearest-neighbour estimates [7], [26], quantile random forests [69], [69] and estimates using neural nets [78].

1.3 Mediation analysis

Mediation analysis is a framework for exploring mechanistic hypotheses about the relationships among measured variables. It aims to expand beyond what can be learned from conventional

regression analysis. In the standard form of mediation analysis, we have three types of variables: the *exposure*, the *mediators*, and the *outcome*. We hypothesize that the exposure causally affects the mediators and the mediators causally affect the outcome, giving rise to an *indirect* effect from the exposure to the outcome. In addition, a *direct effect* may exist in which the exposure directly affects the outcome. The primary goal of mediation analysis is to estimate and quantify these direct and indirect effects, and to assess the uncertainty in these estimates.

In early frameworks for mediation analysis, the exposure, mediator, and outcome were all scalar-valued. More recently, methods have been developed in which any of the three types of variables can be multivariate. In this thesis, we consider multivariate mediators but the exposure and outcome are always scalars for us.

Guarantees about the causality of mediation structure generally require the exposure to be randomized, in which case it may be referred to as a *treatment*. Our motivating data application involves the life history of human beings which is impossible to randomize, so our mediation analysis takes place in an observational setting. Therefore the causality of any findings cannot be guaranteed from the data and statistical analysis alone.

In a conventional regression analysis with this type of data, we would regress the outcome on the mediators and exposure using some form of multiple regression. This type of analysis does not consider mechanistic relationships between the exposure and the mediators, as both are treated as covariates predicting the outcome. In regression-based approaches to mediation analysis, two separate regression models are fit, one regressing the mediators on the exposure (which could involve separate models for each mediator), and another regressing the outcome on the mediators and exposure. Mediation analysis integrates the point estimates and uncertainty information about these models to produce an overall picture of the mediation structure. We note that there also exist frameworks to mediation analysis that use a single integrated model such as a structural equations model. The approaches we develop here build on the regression approaches so we focus on that line of work here.

The earliest form of regression-based mediation analysis was based on linear models of the form $y = a + b \cdot x + c \cdot m + \epsilon$ and $m = d + e \cdot x + \eta$, where ϵ and η are random variables capturing the unexplained variation. The *product coefficient* estimate of the indirect effect is simply $\hat{e} \cdot \hat{c}$. In a linear model, the product $e \cdot c$ is the expected change in y resulting from a one-unit change in x if the direct effect captured by the slope parameter b is "blocked". The slope parameter b itself represents the expected increase in y resulting from a one-unit change in x if the indirect effect is blocked. In this setting, we can also define a total effect $b + c \cdot e$ capturing the expected change in y resulting from a one-unit change in x through either direct or indirect routes.

The mediation structure captured by linear models is somewhat limited by the simple form of these models. It is possible to attain more insight by introducing nonlinear relationships, for

example with spline basis functions, or by including interactions in the regressions. Moreover any approach based on least-squares targets conditional means, leaving unexplored any mediation structure existing in the tails of the distribution that differs from the mediation at the conditional mean. This has motivated a small number of researchers to consider mediation analysis involving quantile regression. As noted above, this is especially pertinent for analyses involving variables such as blood pressure, since the health consequences of high and low blood pressure are quite distinct.

Much of the existing work on quantile-based mediation analysis has utilized linear quantile regression, in which the conditional quantiles of the response are modeled as linear functions of the explanatory variables. Here we leverage a recently-proposed non-parametric approach to quantile regression. By utilizing this approach we aim to capture mediation effects that may be non-linear and/or heterogeneous at different probability points.

An important conceptual challenge is the interpretation of quantile-based mediation analysis. The conditional mean can be interpreted as a prediction of one observed value. Thus, when working with conditional means we can speak of the indirect and direct effects as corresponding to expected changes in the observed response that can be attributed to changes in the exposure variable. Such an interpretation also can be made for the conditional median. However outer conditional quantiles are not naturally viewed as predictions of the data, but rather as population parameters describing collections of individuals (resulting from hypothetical perturbations to the overall population of interest). For example, we may find that the indirect effect of a one-unit change of the exposure x on the 90^{th} percentile of the outcome y through a mediator m is 0.5 units, while the analogous effect at the median is 0.25 units. This means that among all people exposed to the one-unit change in x, and blocking any direct effect of x on y, the indirect effect of this exposure is to increase the 90^{th} percentile by two times the increase in the median. This is a statement about the heterogeneity of responses to the exposure across individuals in the population, rather than being a statement about impacts on individual people's outcomes.

1.4 Baseline analyses

To provide a reference point for the results of the novel analytic frameworks discussed in this thesis, in this section we present a limited analysis of the Dogon anthropometry and blood pressure data using conventional analytic methods. We use generalized estimating equations (GEE) for this purpose although mixed effects regression or other established methods for repeated measures data could also have been used. The results presented here are for a linear mean structure model in which systolic blood pressure (SBP) is the dependent variable. The dataset includes repeated SBP measures per person, taken from ages 11 to 27. The SBP measurements are irregular in

time, with subjects having between 1 and 12 longitudinal SBP measurements, with a median of 9 measurements. An exchangeable working correlation model is used in the GEE analysis to accommodate residual correlation among the SBP values. We note that by using GEE, estimation and inference for the mean structure parameters are robust to misspecification of the working correlation model. Since this is a linear mean structure, this analysis can also be viewed as a form of generalized least squares (GLS), and the mean structure parameters capture changes in the conditional mean.

The mean structure includes both cross-sectional and longitudinal effects. The cross sectional effects correspond to age, various anthropometry measures, and other relevant variables measured at the same time that the SBP is measured. For example, we control for pregnancy status in females, and we control for the number of prior occasions on which a subject's blood pressure was measured to control for "white coat hypertension". Capturing the effects of early life exposures is more difficult as these are measured at different ages and differing numbers of times in different subjects. To accommodate this, we use data imputation based on a "process regression" model discussed in detail in chapter 4. Details of this model are given in that chapter, but to briefly summarize here, we estimate the conditional mean of a given childhood anthropometry measure at ages $1, 2, \ldots, 10$, given all observed data for that individual and conditioned on the structure of the fitted process regression model. We then used principal components analysis (PCA) to reduce these 10-dimensional trajectories to univariate scores. The dominant such score was included in the GEE regression analysis as a covariate. We do not claim that only the dominant factor is informative, but for simplicity only that term is considered here. We note that this is a single imputation analysis and a full multiple imputation analysis would give a fuller picture of the uncertainty.

Due to profound sex effects, separate models were fit for female and for male subjects. We considered a total of 24 models – 6 childhood anthropometry measures $\times$ two levels of covariate adjustment $\times$ two sexes. The most relevant findings that provide context for the remainder of this thesis are the direction, size, and significance levels of cross-sectional and longitudinal associations between SBP and anthropometry, therefore that is what we focus on here. These results are summarized graphically in figure 1.1, with all model coefficients shown as standardized effects, i.e. the estimated change in SBP corresponding to a 1-standard deviation change in an explanatory variable. Note that the horizontal axis in this plot has units of mm Hg (the standard unit for SBP) per standard deviation of body size.

Across the 24 models, we see that main effects for childhood body size are always negative, although in a few cases, mainly for height variables, they are not statistically significantly different from zero. This reflects the conditional inverse association that has been noted before and is the subject of some debate, as discussed above. In contrast, and unsurprisingly, measures of adult body size (height and BMI) obtained at the same time as the SBP measure always have positive

point estimates that are statistically different from zero. In terms of effect sizes, for females BMI shows a much stronger association with SBP than with height, but for males these two distinct aspects of body size have similar associations with SBP. Also, the current body size measures have coefficients with magnitude that are 3-5 times greater than those for childhood body size.

As discussed in detail in chapter 3 on mediation analysis, debate has taken place regarding the proper interpretation of models such as these. Childhood body size is not independent of adult body size, and while such independence is not a requirement for regression analysis, it complicates reasoning about these models. An intervention that aims to raise childhood body size would potentially also raise adult body size, resulting in an "indirect efect" that acts together with the "direct effect" reflected in the estimated regression coefficient for childhood body size. The opposing effects of these two plausibly mechanistic pathways may combine to do either harm or good depending on their respective magnitudes. Here we interpret the coefficients for childhood body size in these analyses as a type of "direct effect". Strictly speaking, they may be viewed as corresponding to an intervention that raises childhood body size (implying better nutrition during childhood), while blocking the concomitant expected increase in adult body size. While not impossible, such an intervention may be difficult to achieve in practice.

Figure 1.1: Regression coefficients for adult height and weight, and for various measures of childhood body size, with adult Systolic Blood Pressure as the response variable. Each row corresponds to a different fitted model. Coefficients for females are shown in orange above each axis line, and coefficients for males are shown in purple below each axis line. The horizontal grey bars depict 95% confidence intervals. The specific body size variable included in each model is given to the left of each axis, and to the right of each axis an "A" appears if the model is adjusted for village of birth, the number of previous SBP measurements, and pregnancy status for females.

CHAPTER 2

Additive Low Rank and Dimension Reduction Approaches to Joint Nonparametric Quantile Regression

2.1 Introduction

In this chapter we develop new tools leveraging recent advances in nonparametric quantile regression to understand complex conditional relationships between the full distribution of an outcome variable $Y \in \mathbb{R}$ and multiple explanatory variables $\boldsymbol{X} \in \mathbb{R}^J$. Recent work in nonparametric quantile regression permits estimation of the p^{th} conditional quantile of an outcome Y given explanatory variables X without specification of strong models. In this work we make use of a local method for conditional quantile estimation that we refer to as the "Quantile Nearest Neighbor" (QNN) method [99]. QNN is analogous to local polynomial least squares methods [36, 90], which have been extensively explored. The ideas developed here should also be applicable when used in concert with other nonparametric quantile regression approaches such as recently proposed tree-based approaches [69].

Our goal is not prediction or forecasting, but rather attaining scientific insight, specifically in the motivating human biology application discussed below. The main challenge we face is that raw local estimates of conditional quantiles are both too voluminous and too imprecise to permit direct interpretation. The main contribution of this chapter is the development of two practical frameworks for post-processing and imposing structure on these raw quantile estimates to achieve scientific insights.

The work presented in this chapter is motivated by questions arising in human biology about the relationship between anthropometry (body size) and blood pressure. Background on this topic was provided in section 1.1 of the Introductory chapter. To briefly review, while it is well-known that cross-sectionally, individuals who are taller and/or have greater height and/or adiposity tend to have greater blood pressure (BP), these relationships are modified in complex ways by other fac-

tors, including by sex, age, and life history. Moreover, when considering longitudinal trajectories of anthropometry variables as predictors of subsequent blood pressure (e.g. looking at anthropometry over the developmental span from birth to adulthood), there may be complex inter-relationships among anthropometry variables taken at different ages, and among different types of anthropometry variables (e.g. between height and adiposity). Most importantly for our goals here, factors associated with either exceptionally high or exceptionally low blood pressure (i.e. outer BP quantiles) may operate differently than factors associated with the mean or with inner BP quantiles such as the median. This motivates our focus on conditional quantiles in this and subsequent chapters of the thesis.

Let $Q_{Y|\boldsymbol{X}}(p|\boldsymbol{x})$ denote the p^{th} quantile of a response Y given $J \geq 1$ explanatory variables $\boldsymbol{X} \in \mathbb{R}^J$. Absent a strong model, this function can exhibit very different behavior at different values of p for fixed $\boldsymbol{x}$, and at different values of $\boldsymbol{x}$ for fixed p. For example, the quantiles may depend strongly on a specific variable x_j at the median $p = 0.5$, but exhibit much weaker relationships with x_j at outer quantiles such as $p = 0.1, 0.9$ (or vice-versa). Similarly, the shapes of $Q_{Y|\boldsymbol{X}}(p|\boldsymbol{x})$ viewed as functions of p for fixed x may be simple translates of each other as in a "location family", or may be heterogeneous, having qualitatively different shapes at different values of $\boldsymbol{x}$. Local estimates of $Q_{Y|\boldsymbol{X}}(p|\boldsymbol{x})$ have the potential to capture these and other forms of heterogeneity. However like all local regression procedures, a bias/variance tradeoff is present. To achieve low bias in a local estimate of $Q_{Y|\boldsymbol{X}}(p|\boldsymbol{x})$, it will generally be the case that the individual point estimates will have substantial variance.

Any multiple regression procedure faces trade-offs with regard to the handling of multiple explanatory variables x. Fully nonparametric analysis allowing arbitrary non-additive and non-linear covariate effects suffers from the "curse of dimensionality". There is a rich body of methodology that allows weak constraints to be introduced, enabling such modeling to proceed. Major advances of this type include single index models [47], [39], [81], [24], dimensionally reduced models [64], [18], [28], and additive models [41], [85], [12].

Quantile regression introduces the additional complexity that we have a model relating x to y for each probability point p. Our goal here is to understand the joint behavior of the quantiles over multiple probability points. One previously-explored issue is the so-called "quantile crossing problem", in which estimated quantiles violate the requirement that $\hat{Q}_{Y|\boldsymbol{X}}(p'|\boldsymbol{x}) \geq \hat{Q}_{Y|\boldsymbol{X}}(p|\boldsymbol{x})$ when $p' \geq p$. While important in some cases, this is not our main concern here. Regardless of whether minor quantile crossings occur, it remains challenging to interpret the joint variation of multiple conditional quantiles with respect to multiple explanatory variables. Therefore, our goal here is to provide practical methods that permit accessible insights to be drawn about a collection of conditional quantiles with respect to variation in multiple explanatory variables.

Specifically, we develop the following two frameworks for joint understanding of a collection

of conditional quantiles:

- **Additive Factor Quantile Regression** (AFQR) – here we model the covariate effects additively, leading to a dramatic reduction in one form of complexity, permitting us to shift our focus to another form of complexity. Specifically, by disallowing non-additive relationships among the covariates, we are able to recover a wide range of easily-interpretable nonlinear relationships between each covariate and the full range of conditional quantiles of the response.

- **Dimension Reduction Quantile Regression** (DRQR) – here we treat the set of conditional quantiles over multiple probability points at a covariate value $\boldsymbol{x}$ as a multivariate response, and employ existing methods for nonparametric multivariate regression that leverage dimension reduction. These methods result in low-dimensional "multi-index" models that can capture non-additive relationships between the covariates and the quantiles, at the cost of a more complex representation (compared to AFQR) that is somewhat challenging to interpret. We explore and illustrate some effective visualization techniques to ease the process of interpretation.

Chapter organization : in this chapter we present two methods that are very different from each other (AFQR, DRQR) but share the same aim, which is to jointly study the conditional quantiles of an outcome in an interpretable manner. This chapter is best viewed as consisting of two subchapters, one per technique. We start by describing the common motivation, aim and notation for both the techniques in the remainder of this section. We then present AFQR in section 2.3 and DRQR in section 2.4. Per method, we provide the outline of the technique and detail the estimation procedure. We also describe the tools built to aid interpretation of the results, and explain the inference procedure. We perform simulation studies to assess the performance of our methods, and illustrate each method on a longitudinal dataset (from the Dogon longitudinal study) and a cross sectional dataset (using the NHANES data). We end the chapter with common concluding remarks, and future directions in section 2.5.

2.1.1 Motivation and basic setup

We are interested in modelling and visualising the associations between explanatory variables and quantiles of an outcome variable, jointly at a given set of probability points. We adopt a quantile based approach, since this allows us to study full conditional distributions of the outcome given the predictors. This allows us to discover relationships in which, for instance, an exposure predicts the median and the outer quantiles of an outcome differently. Furthermore, we wish to accommodate

nonlinearity, as when a fixed increase in an exposure corresponds to different changes in a quantile of the outcome based on the initial value of the exposure.

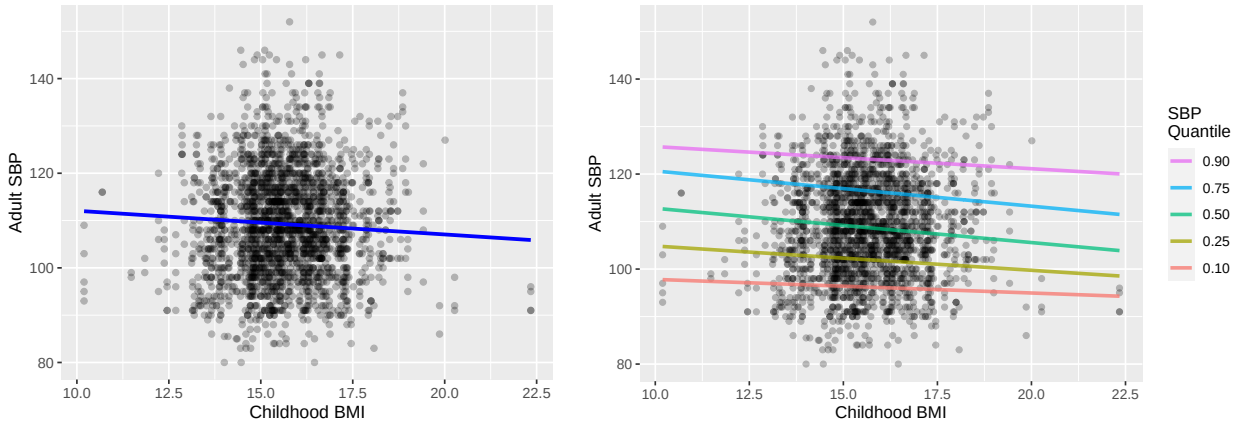
Why study quantiles?

Let us consider an example to illustrate the importance of modelling full distributions via quantiles instead of means. To keep it easy, we present an example in which we focus on a single covariate. We use this as a running example through-out the chapter. In the Dogon data, we consider adult SBP among men, and model its relationship with childhood BMI while controlling for adult body size via adult BMI and adult height (HT).

In fig (2.1) we can see that childhood BMI for males around age 3, impacts different quantiles of adult SBP around age 21 differently. Subfigure (2.1a) contains the fit of a linear regression model with kernel weights corresponding to age. In particular, we consider a smaller bandwidth of 0.5 years for childhood age and a larger bandwidth of 2 years for adult age. The second subfigure (2.1b) presents a quantile regression fit at different quantiles of the outcome, with the same weights as above. Both the models control for adult body size via the formula

$$\texttt{Adult SBP} \sim \texttt{childhood BMI + adult HT + adult BMI + childhood Age + adult Age}.$$

The predicted values in these images are calculated at median adult HT and BMI, across childhood BMI. As we can see, a least squares fit (2.1a) does not give us the full story. For instance, it does not tell us (1) about the change in the shape of the SBP distribution across childhood BMI, or (2) whether higher SBP (0.75) and lower SBP (0.25) quantiles vary similarly as functions of childhood BMI. From fig (2.1b), while all the quantiles of adult SBP are decreasing functions of childhood BMI, the median has a much steeper slope than the outer quantiles (both lower and upper), which means that there is more signal at median adult SBP, as compared to its outer quantiles. We can see that the mean trend (obtained from OLS) is not able to capture this difference in shape of the curve across different quantiles of the outcome, and does not give us all the information.



(a) Linear Regression: OLS

(b) Linear Quantile Regression

Figure 2.1: Adult SBP between ages 21 ± 2.0 years as a function of childhood BMI around ages 3 ± 0.5 years, for males only.

Notation

Here we establish notation. We consider data that consists of an outcome Y and a set of J explanatory variables $\boldsymbol{X} = \begin{bmatrix} X_1 & X_2 & \cdots & X_J \end{bmatrix}^T \in \mathbb{R}^J$, where X_j represents the j^{th} covariate. If this is longitudinal data, all variables would be functions of time with repeated measures per individual. We omit time points in this subsection for ease of notation and generalizability. The data are of the form $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^n = \left\{ \left(y_i, \{x_{ij}\}_{j=1}^J\right) \right\}_{i=1}^n$, with total number of observations n. We will denote the observed data by $\boldsymbol{X}^o$ short for $\boldsymbol{X}$ observed.

Throughout this document, all vectors will be bold and all matrices will be capital and bold. Random variables are capital, and vector-values random variables will be capital and bold. It will be easy to distinguish between vector-values random variables and matrices per context. Finally, observed values of random variables will be denoted by small letters. A subtle distinction worth noting is that $\boldsymbol{x}_i$ represents the i^{th} observation in the dataset, whereas x_j represents an observed value of the j^{th} independent variable X_j, and $\boldsymbol{x}$ represents a generic observed value of $\boldsymbol{X}$.

Our approach

We are interested in exploring the full distribution of the outcome and explaining its variation at all quantiles in terms of explanatory factors. As we briefly discussed in the Introductory chapter, non-parametric methods have been developed to flexibly estimate conditional quantiles of the outcome. However, many of these methods are opaque, and thus provide results which may be difficult to interpret. As explained earlier, QNN is one such non-parametric qunatile regression technique that we use through-out this chapter to obtain non-parametric estimates.

In particular, per observation $\boldsymbol{x}_i$, we obtain conditional quantile estimates of the outcome over a grid of probability points, say $\boldsymbol{p} = [p_1, \cdots, p_m]$, via QNN. This gives us a matrix of estimated conditional quantiles $\hat{\boldsymbol{Q}} \in \mathbb{R}^{n \times m}$. We discuss methods to understand and visualize the association between this matrix $\hat{\boldsymbol{Q}}$ and the observed data $\boldsymbol{X}^o \in \mathbb{R}^{n \times J}$. So our methods treat the conditional quantile estimates obtained from QNN (which correspond to elements of $\hat{\boldsymbol{Q}}$) as the dependent variable, and the dataset "used" by our proposed methods is of the form $\{(\hat{\boldsymbol{Q}}_{i\cdot}, \boldsymbol{x}_i)\}_{i=1}^n$ where the i^{th} row of $\hat{\boldsymbol{Q}}$ corresponds to the conditional quantiles estimated at $\boldsymbol{x}_i$, i.e. $\hat{\boldsymbol{Q}}_{i\cdot} = \left(\hat{Q}_{Y|\boldsymbol{X}}(p_1|\boldsymbol{x}_i), \cdots, \hat{Q}_{Y|\boldsymbol{X}}(p_m|\boldsymbol{x}_i)\right)$ instead of the observed data $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$. Thus, obtaining the QNN estimates can be viewed as a pre-processing step to the proposed techniques (AFQR/ DRQR) or the proposed techniques can be viewed as post-processing steps to the outcome of QNN.

2.2 Properties of local non-parametric quantile estimates

Our main contribution is a data analysis framework that consists of a pipeline in which first the data are first "transformed" to predictions of their conditional quantiles, and then these predicted conditional quantiles are modeled and/or analyzed in various ways. Such pipelines may be difficult to analyze and justify from a theoretical perspective, since bias, variance, and other forms of statistical uncertainty propagate through the pipeline in a complex manner. Consider for example a two-stage procedure in which the observed data $\boldsymbol{y}$ are first transformed into fitted values $\tilde{\boldsymbol{y}}$ (e.g. quantile estimates) using a strong parametric model. The $\tilde{y}$ values would likely exhibit strong global correlations – e.g. an observation $\tilde{y}_i$ may have strong dependence with every other observation $\tilde{y}_j$, even if x_i and x_j are highly dissimilar. Furthermore, a strong model is almost certainly an incorrect model, and therefore each $\tilde{y}$ value will potentially be strongly biased.

We argue that in the present setting, the challenge of adopting a two-stage approach to data analysis is made less severe through the use of local non-parametric procedures in the first stage of the pipeline. In a local approach, there should only be weak dependence between the $\tilde{y}_i$, beyond any dependence that was already present in the data (if not an independent sample). The pipeline analysis method itself induces additional dependence only in small neighborhoods defined through the covariates $\boldsymbol{x}$. Furthermore, we will demonstrate that the $\tilde{y}$ are minimally biased in relation to their target values, and thus behave more like "data" than like "estimates". These properties follow from the local, nonparametric nature of the quantile regression analysis employed here, and are supported by simulation studies reported below. We employ a non-parametric quantile estimation approach previously developed and analyzed in [99], which we refer to as the "QNN" (Quantile Nearest Neighbor) procedure. The inventors of the QNN procedure provide asymptotic results demonstrating that the quantile estimates concentrate around their target values. We augment this

theoretical work with empirical studies presented below.

The goal is to estimate the p^{th} conditional quantile of a random variable Y given explanatory variables $\boldsymbol{x}$, and we denote this target of estimation as $Q_{Y|\boldsymbol{X}}(p|\boldsymbol{x})$. For each observation y_i ($i = 1, \ldots, n$), a free parameter θ_i is introduced, and the check function $\rho_p(x) = (p - \mathcal{I}_{x<0})x$ is used to assess the fit of the θ_i to the y_i – that is, $\sum_i \rho_p(y_i - \theta_i)$ should be small. However, minimizing this loss function yields degenerate quantile estimates $\hat{\theta}_i \equiv y_i$ which both overfit the data, and that fail to unbiasedly estimate the target quantile (unless $p = 0.5$ where these degenerate estimates are median unbiased).

Regularization is used to address this issue. Specifically, for each i a neighborhood N_i is constructed so that for a chosen dissimilarity function $d(\cdot)$, $d(x_i, x_j)$ is small when $j \in N_i$. The overall loss function becomes

$$\sum_i \rho_p(y_i - \theta_i) + \lambda \sum_i \sum_{j \in N_i} |\theta_i - \theta_j|.$$

The regularization involves two tuning parameters – the neighborhood size $|N_i|$ and the penalty weight λ. The inventors of QNN claimed that a neighborhood size of 5 and penalty weight of $\lambda = 0.1$ work well in a broad range of settings. For computation, the inventors of QNN favor a proximal gradient type of algorithm for speed, but for moderate-sized data sets as we have here, traditional linear programming methods are easy to implement and sufficiently fast, so that is what we employ here. Our implementation of the QNN procedure is available at `github.com/kshedden/QuantileNN.jl`.

In our pipeline, the goal of QNN is not to reduce variance – that goal is achieved through the second stage of the pipeline (an additive factor-type model or multivariate dimension reduction analysis). In fact, in a perfect world, QNN would transform the observed data y_i into quantile estimates $\tilde{y}_{pi}$ that are unbiased but that have approximately the same variance as the data – that is, we would be willing to accept $\text{var}(\tilde{y}_{ip})$ being comparable or even somewhat greater than $\text{var}(y_i)$, especially for extreme p, as long as $E[\tilde{y}_{ip}] \approx Q_{Y|X}(p|X = x_i)$. This would justify viewing the $\tilde{y}_{ip}$ as a "transformation" of the data, analogous to the use of a logarithm transformation in basic regression.

Our main concerns are bias and dependence, and we focus on bias here. Bias is the systematic part of estimation error. We first consider by analogy local polynomial regression – QNN is most analogous to local constant regression since the θ_i are shrunk together within each neighborhood N_i without regard to local linear trends or curvature. Bias in this type of procedure is largely determined by the trend and curvature of the target function f as measured by f' and f''. When f is locally constant, there should be minimal bias as long as the bandwidth is small. Thus, one contributor to the bias would be any component of the estimation error that varies systematically

with $Q'_{Y|\boldsymbol{X}}(p|\boldsymbol{x})$ and $Q''_{Y|\boldsymbol{X}}(p|\boldsymbol{x})$, where derivatives are taken with respect to x.

Local polynomial regression targets the conditional expectation while QNN targets conditional quantiles. In quantile regression, an additional source of bias arises as the targeted quantile becomes more extreme (i.e. further from the conditional median $p = 0.5$). In our studies this form of bias appears to dominate bias induced by the local trend and curvature in $Q_{Y|\boldsymbol{X}}(p|\boldsymbol{x})$ (unless $p \approx 0.5$). We hypothesized that a component of the estimation error can be explained as bias toward the conditional median – that is, upper quantiles ($p > 0.5$) are biased downward and lower quantiles ($p < 0.5$) are biased upward. Specifically, we considered bias as the component of prediction error that can be predicted from the following function:

$$g(p, x) \equiv Q_{Y|\boldsymbol{X}}(p|\boldsymbol{x}) - Q_{Y|\boldsymbol{X}}(0.5|\boldsymbol{x}). \tag{2.1}$$

We assessed this hypothesis empirically using simulation studies, working with populations where the population values of all conditional quantiles are known exactly.

2.2.1 Simulation Study

We conducted a simulation study to assess the bias in the QNN procedure. The bias in $\hat{Q}_Y(p; x)$ as an estimate of $Q_{Y|\boldsymbol{X}}(p|\boldsymbol{x})$ will be denoted $\text{bias}(p, x) \equiv E\hat{Q}_{Y|\boldsymbol{X}}(p|\boldsymbol{x}) - Q_{Y|\boldsymbol{X}}(p|\boldsymbol{x})$. We simulated data from a heteroscedastic Gaussian population as follows. The explanatory variables x are simulated as standard independent multivariate Gaussian values of dimension d. The conditional mean of the simulated population is either x_1 or x_1^2 (to consider the effect of curvature in the population conditional quantile functions), and the conditional variance is $\sigma^2(1 + x_2^2)$. We population design as well as the population quantile function are presented for both the cases below, where $\Phi^{-1}(p)$ is the quantile function of the standard normal distribution.

Linear:

Data Generation	$\boldsymbol{X} = \mathcal{N}_d(\mathbf{0}_d,\ \mathbb{I}(d))$
	$Y = X_1 + \epsilon$
	$\epsilon \sim \mathcal{N}(0,\ \sigma^2(1 + X_2^2))$.
Quantile function	$Q_{Y|\boldsymbol{X}}(p; \boldsymbol{x}) = \sigma(1 + x_2^2)^{1/2} \cdot \Phi^{-1}(p) + x_1$.

Non Linear:

Data Generation	$\boldsymbol{X} = \mathcal{N}_d(\mathbf{0}_d,\ \mathbb{I}(d))$
	$Y = X_1^2 + \epsilon,$
	$\epsilon \sim \mathcal{N}(0,\ \sigma^2(1 + X_2^2)).$
Quantile function	$Q_{Y\|\boldsymbol{X}}(p; \boldsymbol{x}) = \sigma(1 + x_2^2)^{1/2} \cdot \Phi^{-1}(p) + x_1^2.$

We generate $m = 100$ datasets from the population described above with sample size $n = 1500$, considering dimensions $d = 2$ and $d = 5$. Note that we only sample $\boldsymbol{X}$ once per dimension d, and generate m sets of Y from the same $\boldsymbol{X}$, for each setting $\sigma = 1$ and $\sigma = 2$. We then use QNN to estimate the conditional quantiles at $p = 0.5, 0.75$, and 0.9 for each simulated dataset. We do not consider quantiles below 0.5 as the quantiles behave similarly based on their distance from 0.5, i.e. $p = 0.2$ behaves similar to $p = 0.8$.

2.2.1.1 Estimate evaluation

The estimation error $\hat{Q}_{Y|\boldsymbol{X}}(p|\boldsymbol{x_i}) - Q_{Y|\boldsymbol{X}}(p|\boldsymbol{x_i})$ reflects both the systematic effects (bias) and purely random effects (variance). To disentangle these two sources of error, we use the estimation errors as the dependent variable in a least squares regression, with the following mean structure

$$\hat{Q}_{Y|\boldsymbol{X}}(p|\boldsymbol{x_i}) - Q_{Y|\boldsymbol{X}}(p|\boldsymbol{x_i}) \ \sim\ g_p(\boldsymbol{x_i}) + g_p(\boldsymbol{x_i})^2 + g_p(\boldsymbol{x_i})^3. \tag{2.2}$$

This mean structure model is fit using ordinary least squares to a dataset consisting of simulated values, pooling over the three values for p and over the m replicates (so the overall sample size for the least squares fit was $3mn$ since we are considering three values for p). We adopt this regression approach for two reasons. First, it requires fewer simulation runs than a direct approach that estimates the bias as the sample mean of $\hat{Q} - Q$ over Monte Carlo replicates. Second, and more important, it helps us understand the structure of the biases, by identifying the factors that predict the systematic component of the estimation errors. We do note that we cannot claim that our empirical model 2.2 completely captures the bias, since systematic errors could result from factors unrelated to g_p.

Let $\hat{\beta}_0, \ldots, \hat{\beta}_3$ denote the four parameters of model 2.2, and let $\widehat{\text{relbias}}(p; \boldsymbol{x}) = \hat{\beta}_1 + \hat{\beta}_2 g_p(\boldsymbol{x}) + \hat{\beta}_3 g_p(\boldsymbol{x})^2$ denote our estimate of the relative bias for probability point p at covariate value $\boldsymbol{x}$, where $\text{relbias}(p; \boldsymbol{x}) = \text{bias}(p; \boldsymbol{x})/g_p(\boldsymbol{x})$. We treat $\beta_0 = 0$ in this definition and consider the value of this intercept separately. For a fixed value of p, we can now summarize the bias as $\text{relbias}(p) = \text{median}\{|\widehat{\text{relbias}}(p; \boldsymbol{x_i})|; i = 1, \ldots, n\}$, which we refer to as the *median absolute relative bias*. Below we focus on the relative bias at $p = 0.75$.

2.2.1.2 Simulation study findings

The fitted intercepts in model 2.2 capture the bias for estimating the conditional median. In our results, these intercepts (table 2.1) were very small relative to the range of true conditional medians over $\boldsymbol{x}$. All fitted intercepts were smaller in magnitude than 0.03, while the true conditional medians had standard deviations (over the sample of $\boldsymbol{x}$ values) ranging from 0.98 to 1.46. This suggests that the estimated conditional medians are nearly unbiased. Arguably, any bias in the estimated conditional medians should result from oversmoothing that fails to capture the local curvature of the true quantile function, and hence would only be present in the model where the true conditional median is quadratic in $\boldsymbol{x}$ (when the true conditional median is linear in $\boldsymbol{x}$ there is no local curvature). However in our simulation study, when the true quantile function was quadratic versus linear, the intercepts of the regression model (equation 2.2) were very similar in magnitude, suggesting that curvature and oversmoothing have minimal contributions to estimation bias of the conditional quantiles.

Our simulation results support our hypothesis that the bias primarily takes the form of attenuation toward the conditional median - that is, when $g_p(x) > 0$ the bias is negative and with $g_p(x) < 0$ the bias is positive. An example is presented in figure 2.2. Figure 2.2 displays the model fit at $p = 0.75$ for the non-linear case ($d = 5$, $\sigma = 2$). Here $g(0.75, \boldsymbol{x}) = \sigma(1 + x_2^2)^{1/2} \cdot \Phi^{-1}(0.75) > 0$ for all values of $\boldsymbol{x}$. The fitted curve corresponds to the estimate of the bias; each point on the curve corresponds to the estimated bias at a point $\boldsymbol{x}$, such that the x-axis value at that point corresponds to $g(0.75, \boldsymbol{x})$. As we can see, here $g(0.75, \boldsymbol{x}) > 0$ everywhere and the estimated bias is negative throughout.

To summarize the overall extent of bias, we report the median absolute bias (as defined above) for $p = 0.75$. The results are presented in table 2.1. The median relative bias is found to increase with σ, and with d. However in the situations that we consider (across all p), the relative bias tends to be less than 10% when $d = 2$ and less than 20% when $d = 5$.

Form	d	σ	Intercept	R^2	relbias(0.75)
Linear	2	1	0.002	0.04	0.06
Linear	2	2	0.005	0.06	0.06
Quadratic	2	1	−0.003	0.03	0.06
Quadratic	2	2	−0.004	0.03	0.06
Linear	5	1	−0.017	0.06	0.12
Linear	5	2	−0.027	0.14	0.13
Quadratic	5	1	−0.011	0.06	0.21
Quadratic	5	2	−0.028	0.12	0.17

Table 2.1: Simulation results for $p = 0.75$.

Finally, we consider the magnitude of the bias in relation to the estimation variance, again noting that we are only detecting the component of the bias that is predictable from g_p as defined above. We find that the estimation bias is a small fraction of the estimation error. This can be seen in figure 2.2 for a specific data setting. The scaterplot in the figure 2.2 corresponds to the errors (which are the dependent variable of the OLS model). This plot clearly indicates that the variance of the estimates (captues by the vertical spread of the errors (points), per $g(p, \boldsymbol{x})$) is much larger than the bias of the estimates, ofcourse when the bias is modelled in terms of $g(p, \boldsymbol{x})$. Formally, we do this using the R^2 values from the linear regression defined above. The R^2 values tend to be less than 10% for $d = 2$ and less than 15% for $d = 5$. These small R^2 values suggest that our estimates are very unstable, and the estimation variance comprises atleast 85% of the estimation error (bias comprises the remainder).

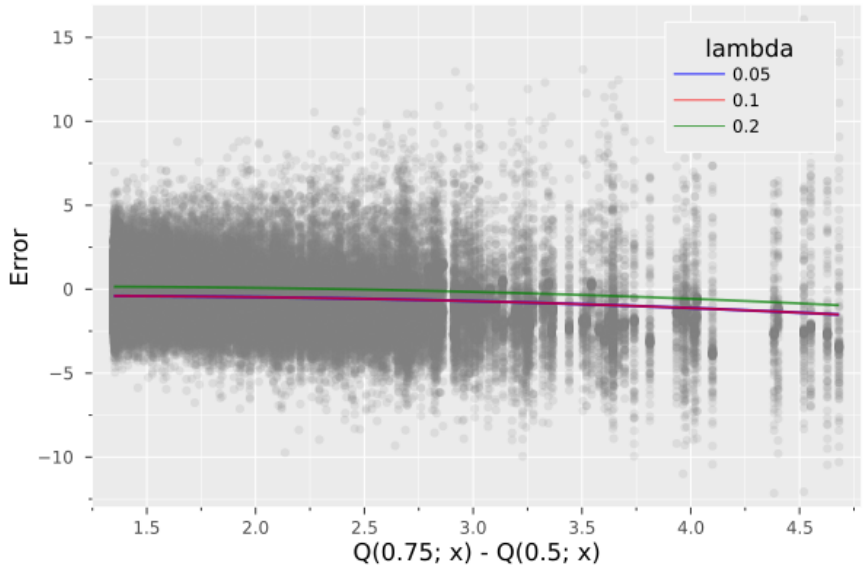


Figure 2.2: OLS fit of the estimation errors in the QNN estimates for the non-linear case with 5 features ($d = 5$) and variance parameter $\sigma = 2$.

2.3 Functional Low Rank Additive Regression

In this section we develop an analytic approach called *Additive Factor Quantile Regression* (AFQR) to represent the collection of all conditional quantile functions in a coherent way using a rank-1 additive structure. We aim to demonstrate here that low rank additive regression is a natural choice for post-processing and making sense of the QNN estimates, which constitute the input to the AFQR approach.

An additive model for quantiles has the form $Q_{Y|X}(p; X = x) = \sum_j h_j(x_j, p)$, representing

the "effect" of each explanatory variable x_j in terms of a function $h_j : \mathbb{R}^2 \to \mathbb{R}$, mapping (x, p) to the effect of x on the p^{th} quantile of the response. Placing a rank-one constraint on each h_j further simplifies the representation so that $h_j(x_j, p) = f(x_j) \cdot g(p)$. Now the role of a covariate x_j is completely represented through the function f_j that maps the variable's value to its score, and the function g_j that captures how changes in the score result in changes to each conditional quantile of the response. The score functions f_j and loading functions g_j are all smooth functions from $\mathbb{R} \to \mathbb{R}$.

We argue that the AFQR approach brings advantages in terms of both interpretability and statistical power. With regard to interpretability, as discussed above AFQR provides an explicit closed-form model for jointly representing all conditional quantiles of the outcome in terms of the explanatory variables. In this way, it is able to reduce the voluminous output of QNN in terms of two functions per covariate (the score function and the loading function) that are relatively easy to interpret. With regard to the statistical power, an important property of AFQR is that the low rank structure reduces the variance of estimates. This is particularly important since local non-parametric techniques (which provide the input data to AFQR) exhibit high variability. This is in fact the case with QNN; our simulation studies in section 2.2.1 produced estimates with variance greater than the squared bias. Such low-bias/high-variance estimates are ideal for use as inputs to AFQR, which can reduce sampling variance but is unlikely to reduce any bias that is present in its inputs. The resulting estimates of these functions benefit from pooling information across all outcome quantiles across the entire range of each covariate.

2.3.1 Introduction to AFQR

AFQR is an additive low rank model that is structured as a sum of products between smooth functions of a single covariate and smooth functions of the outcome quantile (the latter indexed by $0 < p < 1$). This structure allows the model to borrow information across quantiles of the outcome, unlike most other quantile regression techniques which do not link the conditional quantile functions for different probability points $p \in (0, 1)$, except perhaps imposing the very weak and necessary constraint that the estimated quantile functions do not cross. This "no crossing" constraint is a minimal form of regularization, and we argue that when working with modest amounts of data, a stronger regularization is needed in order to gain more stable estimates, especially of the outer quantiles.

The population version of the AFQR model has the form

$$Q_{Y|\boldsymbol{X}}(p|\boldsymbol{x}) = \mu(p) + \sum_{j=1}^{J} f_j(x_j) \cdot g_j(p), \quad p \in (0, 1),\ \boldsymbol{x} \in \mathbb{R}^J, \tag{2.3}$$

where $\{f_j, g_j\}_{j=1}^J$ are smooth functions such that $f_j(0.5) = 0$ for each j. The additive structure mitigates the curse of dimensionality, promotes model interpretability, and makes it feasible to visualize the association between the outcome and different covariates through the smooth functions f, g. We refer to $\mu(p)$ as the *central axis*, as it is the conditional quantile function of the outcome evaluated at the median of each component of $\boldsymbol{X}$, i.e.

$$\mu(p) = Q_{Y|\boldsymbol{X}}\big(p|X_1 = \text{median}(X_1), \cdots, X_J = \text{median}(X_J)\big). \tag{2.4}$$

The constraints $f_j(0.5) = 0$ are imposed to make the central axis identifiable. The reason for fixing these functions at $X_j = 0.5$ is that in practice we quantile transform each covariate to be uniform on $(0, 1)$, so this amounts to constraining the score at the median value of each covariate to be zero. With these constraints in mind, a natural interpretation of the model parameters $\{f_j, g_j\}_{j=1}^J$ is in terms of deviations from the central axis. We discuss the interpretation of the model in more detail in section 2.3.5.

Connections to multivariate regression

There are connections between our approach and the large body of work on low-rank methods for multivariate regression. In that setting, a vector of outcomes $\boldsymbol{Y} \in \mathbb{R}^q$ is regressed on a vector of predictors $\boldsymbol{X} \in \mathbb{R}^q$. Focusing on linear relationships, a common model is $E[Y|\boldsymbol{X}] = \boldsymbol{B}\boldsymbol{X}$, where $\boldsymbol{B}$ is a $q \times J$ matrix. Since the goal is to estimate conditional means, least squares methods are appropriate, but to regularize the estimation it is common to impose rank constraints on the coefficient matrix $\boldsymbol{B}$. Moving to the functional setting, $Y = Y(t)$ and $X = X(t)$ may be viewed as functions rather than as vectors, and the conditional mean can be modeled in linear form as $E[Y(t)|\boldsymbol{X}] = \int B(s,t)X(s)ds$. In our setting, the conditional quantile function plays the role of Y, and is indeed a function of a continuous index (p). However we discretize the domain of p and work with a gridded vector of conditional quantiles, establishing a connection between our work and the setting of multivariate regression with vector responses.

Reduced rank approaches to multivariate regression are discussed in [3, 4, 74, 94], and have been studied more recently by many [101, 71, 73, 15]. A brief review of recent work is included in [16]. Our proposed method (AFQR) can be expressed as a low rank regression model, but we partition the parameters into blocks and impose a rank-one constraint on each block, instead of constraining the rank of the entire parameter matrix $\boldsymbol{B}$ as is typically done.

The AFQR model is fit by minimizing an objective function of the form given below in equation 2.5. This is developed in detail in section 2.3.3 (see in parrticular equations 2.10 and 2.12), but we omit the full derivation here and adopt simpler notation to draw a parallel between AFQR and low rank regression. Our approach minimizes the objective function

$$\|\boldsymbol{Y} - \sum_{j=1}^{J} \boldsymbol{X_j}\boldsymbol{\beta_j}\boldsymbol{v_j'}\|^2 + \text{ penalty}, \tag{2.5}$$

where $\boldsymbol{Y}$ is an $n \times m$ matrix whose columns correspond to distinct quantiles, i.e. $\boldsymbol{Y}$ is a column-centered version of $\hat{\boldsymbol{Q}}$, where the columns are centered around estimates of the marginal median value of each variable (see section 2.3.3 for details). The matrix $\boldsymbol{X_j}$ is an $n \times d_j$ matrix of basis functions for the j^{th} explanatory variable, allowing us to represent the smooth functions f_j (see section 2.3.3.2). The vector $\boldsymbol{\beta_j}$ is a d_j dimensional vector of regression coefficients and $\boldsymbol{v}_j$ is a J dimensional vector of loadings.

To establish a connection with low-rank regression, note that it is equivalent to view this problem as minimizing

$$\|\boldsymbol{Y} - \sum_{j=1}^{J} \boldsymbol{X_j}\boldsymbol{F_j}\|^2 + \text{ penalty}$$

where $\boldsymbol{F_j}$ is a rank-1 matrix. If the $\boldsymbol{X_j}$ matrices are concatenated horizontally to produce $\boldsymbol{X} = [\boldsymbol{X_1}, \boldsymbol{X_2}, \cdots, \boldsymbol{X_J}]$ and the $\boldsymbol{F_j}$ matrices are concatenated vertically to obtain $\boldsymbol{F} = [\boldsymbol{F_1'}, \boldsymbol{F_2'}, \ldots, \boldsymbol{F_J'}]'$ then the objective function can be written as $\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{F}\|_F^2$. This is similar to a standard low-rank regression, however we constrain each $\boldsymbol{F_j}$ to have rank exactly 1 whereas a standard low-rank regression would control the overall rank of $\boldsymbol{F}$. We could constrain the rank of F to be J, but this does not guarantee that each F_j will have rank one.

2.3.2 Model Structure

Let us start by motivating why the additive low rank structure is sensible to adopt in a quantile-regression setting. By analogy with factor analysis, this model can be understood in terms of *scores* and *loadings*. The score functions f_j capture the role of each covariate and the loading functions g_j capture the manner in which these scores impact the response quantiles. Three properties of the additive low-rank structure are discussed below.

- **Additive** Each covariate X_j has a distinct relationship with the conditional quantiles of the outcome variable Y, and these effects are additive over the covariates. In other words, there are no interactions among the covariates.

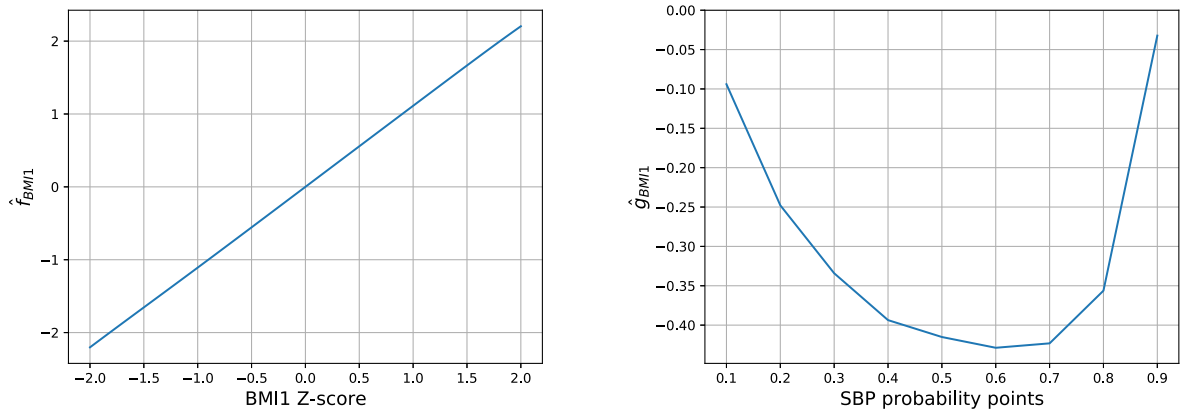
- **Non-Linear** Since the score functions f_j are smooth but not necessarily linear, changes in a covariate do not in general lead to linear changes in the quantiles. Further, non-linearity of the loading functions g_j implies that changes in a covariate may have heterogeneous effects at different quantiles of the outcome variable, implying heteroscedasticity.

- **Low rank** Each covariate is represented by a single term in the additive representation (2.3), thereby simplifying and regularizing its contribution to the model. Equation 2.3 implies that as we vary X_j, the quantile function varies exclusively in the direction g_j, with the the amount of variation given by f_j.

We note that the setting $g_j \propto 1$ is a special case that corresponds to a *location family* in which all quantiles move together with the median (and with the mean). Suppose that $Y = \alpha + \beta X + \epsilon$ for jointly distributed random variables X, ϵ. In this case, $Q_{Y|X}(p|X = x) = \alpha + \beta X + Q_\epsilon(p)$, where Q_ϵ is the marginal quantile function of ϵ. This population follows an AFQR with $J = 1$, a central axis $\mu(p) = \alpha + \beta \cdot \text{med}(X) + Q_\epsilon(p)$, a single score function $f_1(x) = \beta \cdot (x - \text{med}(X))$, and corresponding loading function $g_j \equiv 1$.

Example with anthropometry and blood pressure As an example foreshadowing our data analysis below, recall the discussion from section 2.1.1. In this context, the low rank representation takes the form

$$Q_{SBP_2}(p|\boldsymbol{x}) = \mu(p) + f_{BMI_1}(x_1) \cdot g_{BMI_1}(p) + f_{HT_2}(x_2) \cdot g_{HT_2}(p) + f_{BMI_2}(x_3) \cdot g_{BMI_2}(p).$$



(a) f_1 for childhood BMI

(b) g_1 for childhood BMI

Figure 2.3: Adult SBP between ages $[18, 22]$ years as a function of childhood BMI around ages $[2.5, 3.5]$ years for males.

The functional estimates corresponding to childhood BMI $(\hat{f}_{BMI1}, \hat{g}_{BMI1})$ are presented in figure 2.3. In this, f_1 is essentially a straight line from -2 to 2, and g_1 is U-shaped. This means

that, at a specific $\boldsymbol{x}$, the inner quantiles of Y (adult SBP in the figure) will always deviate from the median curve $\{\mu(p) : p \in [0,1]\}$ more than the outer quantiles (both lower and higher), regardless of $\boldsymbol{x}$. Recall that the "median curve" at each point $p \in [0,1]$ refers to the quantile of Y (adult SBP) for median(X_j) (i.e. a person with median childhood BMI, and median adult BMI and HT). The curve in fig 2.3b displays the direction of the deviation from the median curve for a person with median adult body size; this only depends on childhood BMI due to the additive nature of the AFQR.

So if we look at the distribution of adult SBP for individuals with low childhood BMI ($x^\star = -1.5$ z-score of childhood BMI) or high childhood BMI ($x^\star = 1.5$ z-score of childhood BMI), both distributions $Q_{Y|\boldsymbol{X}}\big(p|X_1 = x^\star, \boldsymbol{X_{-1}} = \text{median}(\boldsymbol{X_{-1}})\big)$ will vary from the the central axis $\mu(p)$ in a similar way – they will vary most near the median and will vary less as we move to the tails. The extent of deviation at a specific X_j is determined by the curvature of the curve $f_j(x_j) \cdot g_j(p)$ $p \in [0,1]$. Since $\{g_j\}_j$ are not necessarily linear, the curvature can be inflated or deflated, and thus lead to conditional quantile curves of different shapes (as functions of outcome probability points for a fixed $\boldsymbol{X}$).

Focusing on the magnitude of change in the quantile function at any point p, if x_j changes from x_{j1} to x_{j2}, then the conditional quantile function will change by an amount proportional to $f_j(x_{j1}) - f_j(x_{j2})$, i.e.

$$Q_{Y|\boldsymbol{X}}(p|\boldsymbol{x}_2) = Q_{Y|\boldsymbol{X}}(p|\boldsymbol{x}_1) + \Big(f_j(x_{j2}) - f_j(x_{j1})\Big) \cdot g_j(p).$$

2.3.3 Model Estimation

In this section, we provide a complete development of the estimation procedure. Our goal is to fit the low rank representation given in (2.3) to the data.

2.3.3.1 Outline of the estimation pipeline

Let $Q^c_{Y|\boldsymbol{X}} \equiv Q_{Y|\boldsymbol{X}}(p;\cdots) - \mu(p)$ denote conditional quantiles that are centered around the central axis $\mu(p)$. The representation 2.3 now becomes

$$Q^c_{Y|\boldsymbol{X}}(p; x_1, \cdots, x_J) = \sum_{j=1}^{J} f_j(x_j) g_j(p). \tag{2.6}$$

Our estimation proceeds as follows

1. Estimate $\mu(p)$ as $\hat{\mu}_j = \hat{\mu}(p_j)$ over a grid of m probability points $\boldsymbol{p} = (p_1, \ldots, p_j)$.
2. Estimate $Q_{Y|\boldsymbol{X}}(p|\boldsymbol{x_i})$ for each observed $\boldsymbol{x_i}$ for each p in $\boldsymbol{p}$, yielding a $n \times m$ matrix $\hat{\boldsymbol{Q}}$.

3. Construct the $n \times m$ array of centered quantiles $\hat{\boldsymbol{Q}}^c_{ij} \equiv \hat{\boldsymbol{Q}} - \mathbf{1}_n \hat{\mu}'$.

4. Fit a smooth low-rank model to $\hat{\boldsymbol{Q}}^c$.

The population version of our model (as in equation 2.6) evaluated at the observed covariate values and the grid of m probability points takes the form

$$\begin{aligned} Q^c_{Y|\boldsymbol{X}}(p; \boldsymbol{x}_i) &= \sum_{j=1}^{J} f_j(x_{ij}) g_j(p), \\ \implies \boldsymbol{Q}^c &= \sum_{j=1}^{J} \begin{bmatrix} f_j(x_{1j}) \\ \vdots \\ f_j(x_{nj}) \end{bmatrix} \begin{bmatrix} g_j(p_1) & \cdots & g_j(p_m) \end{bmatrix} \\ &= \sum_{j=1}^{J} \boldsymbol{u_j} \cdot \boldsymbol{v'_j}. \end{aligned} \tag{2.7}$$

We have now parameterized the infinite dimensional model (2.6), which was parameterized in terms of smooth functions $\{f_j, g_j\}_{j=1}^J$, in terms of finite-dimensional parameters $\mathbf{u}_j \in \mathbb{R}^n$ and $\mathbf{v}_j \in \mathbb{R}^m$, for $j = 1, \ldots, J$. While in principle we may be able to fit this model directly to the data, the number of parameters grows linearly with the sample size (specifically, it is $J \cdot m \cdot n$). Further, we would like u_j to vary smoothly with x_j. Therefore, the dimensionality of the model will be further reduced and regularized as discussed in the subsequent two sections.

2.3.3.2 Basis representation

We use basis functions to nonparametrically represent the parameter spaces for smooth unknown functions. Let $\tilde{\boldsymbol{X}}_j \in \mathbb{R}^{n \times d_j}$ denote a basis matrix of dimension d_j for the j^{th} additive term in (2.6). Each column of $\tilde{\boldsymbol{X}}_j$ is a smooth function evaluated at the observed covariate values $x_{1j}, \ldots, x_{nj}$. For example, a quadratic polynomial basis of order 2 would give us a matrix in which the first column would consist of a constant, the second column would consist of the observed values $\boldsymbol{x_{\cdot j}} \in \mathbb{R}^n$, and the third column would be the observed values squared. Spline bases are another possibility. More generally, let the k^{th} column of $\tilde{\boldsymbol{X}}_j$ be formed by evaluating the basis function h_k on $\boldsymbol{x_{\cdot j}}$. Then the fitted function $\hat{f}_j$ corresponding to the parameter f_j can be written

$$\hat{f}_j(\cdot) = \sum_{k=1}^{d_j} \hat{\beta}_{jk} h_k(\cdot) \implies \hat{\boldsymbol{u}}_j = \begin{bmatrix} \hat{f}_j(x_{1j}) \\ \vdots \\ \hat{f}_j(x_{nj}) \end{bmatrix} = \begin{bmatrix} \sum_{k=1}^{d_j} h_k(x_{1j}) \hat{\beta}_{jk} \\ \vdots \\ \sum_{k=1}^{d_j} h_k(x_{nj}) \hat{\beta}_{jk} \end{bmatrix} = \tilde{\boldsymbol{X}}_j \hat{\boldsymbol{\beta}}_j. \tag{2.8}$$

Due to its lower dimensionality, we do not use basis functions for the loading vectors v_j (although it would be straightforward to do so). Equivalently, we are using $I_{m\times m}$ as a collection of non-smooth basis vectors for the loading functions. The additive low rank model now becomes

$$\boldsymbol{Q^c} \approx \sum_{j=1}^{J} \underbrace{\tilde{\boldsymbol{X}}_j \boldsymbol{\beta}_j}_{\boldsymbol{u}_j} \boldsymbol{v}_j', \tag{2.9}$$

where $\boldsymbol{\beta}_j \in \mathbb{R}^{d_j}$ and $\boldsymbol{v}_j' \in \mathbb{R}^m$ are the parameters. $\boldsymbol{v}_j$ is an m-dimensional vector whose k^{th} entry is $g_j(p_k)$ and can be thought of as the loading corresponding to the j^{th} additive factor. Each observation i has a score $f_j(x_{ij})$ for each component j, which are stacked in a vector to form $\boldsymbol{u}_j \in \mathbb{R}^n$. The loss function corresponding to the model in equation 2.9 is

$$\|\hat{\boldsymbol{Q}}^{\boldsymbol{c}} - \sum_{j=1}^{J} \tilde{\boldsymbol{X}}_j \boldsymbol{\beta}_j \boldsymbol{v}_j'\|_F^2, \tag{2.10}$$

where the β_j and v_j are free variables.

2.3.3.3 Regularization

In order to obtain smooth estimates of the functions f_j and g_j, we utilize regularization analogous to that used in various functional data analysis (FDA) tools and in Generalized Additive Models (GAMs) [95]. The fact that the basis functions used to construct the matrices $\tilde{\boldsymbol{X}}_j$ are smooth functions of the underlying covariate whose scores they represent implies that the fitted basis functions will be somewhat smooth. However an arbitrary linear combination of smooth functions may not be sufficiently smooth, so in addition we penalize the curvature of each fitted score function and loading function. Specifically, we impose an additive smoothing penalty equal to the integrated squared second derivative of each functional parameter that we wish to be smooth. The penalty for a twice differentiable function $\hat{f}$ is $\int \hat{f}''^2(x)dx$.

The penalty expressed above as an integral can be numerically approximated using the midpoint rule, which amounts to constructing an equispaced grid, say of 100 points on the range of each variable, and constructing the basis matrix $\tilde{\boldsymbol{X}}_j^p$ corresponding to this grid (the same basis functions used to construct $\tilde{\boldsymbol{X}}_j$ are used here). Now let $D^2 \in \mathcal{R}^{n-2\times n}$ denote the second difference operator (row j of D^2 contains the stencil $(1, -2, 1)$ beginning in column j). The quadratic form

$$\beta \to \beta' \tilde{\boldsymbol{X}}_j^{pT} F^T F \tilde{\boldsymbol{X}}_j^p \boldsymbol{\beta}_j = \|D^2 \tilde{\boldsymbol{X}}_j^p \boldsymbol{\beta}_j\|^2 \tag{2.11}$$

approximates the integrated squared second derivative of the function represented (on a grid) by $\tilde{\boldsymbol{X}}_j^p \boldsymbol{\beta}$.

An additional complexity in specifying the regularization penalties is that the scores $\boldsymbol{u}_j = \tilde{\boldsymbol{X}}_j\boldsymbol{\beta}_j$ and loadings $\boldsymbol{v}_j$ only impact the fitted values through the outer product $\boldsymbol{u}_j\boldsymbol{v}_j'$. Any representation of the form $c \cdot \boldsymbol{u}_j$, $c^{-1} \cdot \boldsymbol{v}_j$ is equivalent as long as $c \neq 0$. Using $\|D^2\tilde{\boldsymbol{X}}_j^p\boldsymbol{\beta}_j\|^2$ as a penalty is problematic because the penalty is sensitive to the specific representation that is used (i.e. it depends on c). To resolve this, we use a scaled version of the penalty that is invariant to the value of c. Specifically, the smoothing penalty for $\boldsymbol{\beta}_j$ is $\|\boldsymbol{v}_j\|^2 \cdot \|D^2\tilde{\boldsymbol{X}}_j^p\boldsymbol{\beta}_j\|^2$ and the smoothing penalty for v_j is $\|\tilde{\boldsymbol{X}}_j^p\boldsymbol{\beta}_j\|^2 \cdot \|D^2\boldsymbol{v}_j\|^2$.

2.3.3.4 Loss function and optimization

The overall objective function that we wish to minimize over $\{\boldsymbol{\beta}_j, \boldsymbol{v}_j\}_{j=1}^J$ is

$$\frac{1}{nm}\|\hat{\boldsymbol{Q}}^{\boldsymbol{c}} - \sum_{j=1}^{J}\tilde{\boldsymbol{X}}_j\boldsymbol{\beta}_j\boldsymbol{v}'\|_F^2 + \sum_{j=1}^{J} c_{u_j}\|\boldsymbol{v}_j\|^2 \cdot \|D^2\tilde{\boldsymbol{X}}_j^p\boldsymbol{\beta}_j\|^2 + \sum_{j=1}^{J} c_{v_j}\|\tilde{\boldsymbol{X}}_j^p\boldsymbol{\beta}_j\|^2 \cdot \|D^2\boldsymbol{v}_j\|^2, \quad (2.12)$$

where $c_{u_j}, c_{v_j} \in \mathbb{R}$ are tuning parameters that determine the smoothness of the estimated curves. We note that the loss function (2.12) is invariant to the ambiguity in the rank-one representation, i.e. replacing $\boldsymbol{\beta}_j$ with $c \cdot \boldsymbol{\beta}_j$ and $\boldsymbol{v}_j$ with $c^{-1} \cdot \boldsymbol{v}_j$ does not change the value of the objective function (2.12).

The loss function (2.12) is a quartic polynomial of the parameters $\{\boldsymbol{\beta}_j, \boldsymbol{v}_j\}$ and hence smooth, but is not convex. However if we fix all the $\boldsymbol{\beta}_j$ parameters and view the loss in terms of the $\boldsymbol{v}_j$ parameters alone, the restricted loss becomes a convex quadratic function of $\{\boldsymbol{v}_j\}_{j=1}^J$. Analogously, if we fix all the $\boldsymbol{v}_j$ parameters and view the loss in terms of the $\boldsymbol{\beta}_j$ parameters alone, it becomes a convex quadratic function of $\{\boldsymbol{\beta}_j\}_{j=1}^J$. This motivates the use of alternating least squares (ALS) to find a local minimum. This algorithm has a long history, including in the calculation of rank-reduced approximations to arrays [92].

Since the loss function is not globally convex, there may be multiple local minima and we do not aim here to find a global minimum. However it is important in practice to have good starting values for any local optimizer. We accomplish this using a two-stage procedure in which the first stage involves solving a least-squares problem and the second stage is a matrix factorization. To accomplish the first stage, the following objective function can be easily minimized over the $\boldsymbol{B}_j$ using least squares techniques, yielding estimates $\hat{\boldsymbol{\beta}}_j$:

$$\|\hat{\boldsymbol{Q}}^{\boldsymbol{c}} - \sum_{j=1}^{J}\tilde{\boldsymbol{X}}_j\boldsymbol{\beta}_j\|_F^2. \quad (2.13)$$

We then use the dominant singular values to approximate each $\hat{\boldsymbol{\beta}}_j \approx \hat{\boldsymbol{\beta}}_j^0\hat{\boldsymbol{v}}_j^{0'}$ in rank-1 form, for

vectors $\hat{\boldsymbol{\beta}}_j^0$ and $\hat{\boldsymbol{v}}_j^0$ which become the starting values for the ALS procedure discussed above.

2.3.4 Inference

It is important to have a means for assessing the uncertainty in the estimated AFQR parameters. This is made complicated in our setting for at least three reasons. First, the input data used to fit the AFQR model is not the observed data, but rather results from "preprocessing" the observed data into estimates of conditional quantiles. If $\boldsymbol{Q}$ denotes the true quantiles and $\hat{\boldsymbol{Q}}$ denotes the stage-1 estimated conditional quantiles obtained from the QNN procedure, then the "errors" $\hat{\boldsymbol{Q}} - \boldsymbol{Q}$ play an important role in determining the uncertainty in the AFQR parameter estimates. These errors are not independent or identically distributed. Specifically, the elements of $\hat{\boldsymbol{Q}}$ are correlated through the QNN procedure, although these correlations should be local. Also, it is likely that the magnitudes of the errors for quantiles at different probability points p differ.

A second reason that inference is difficult in this setting is that we are working with longitudinal data. This will be discussed further below, but as a consequence, observations are correlated if they are made on the same subject. This is an additional source of non-independence in the elements of $\hat{\boldsymbol{Q}}$ that operates in tandem with the dependence induced by the QNN algorithm.

A third reason that inference is difficult in this setting is that the AFQR model in its simplest form is not identified. To mitigate this, as discussed above, we use a loss function that is invariant to choosing among different members of an equivalence class of parameter settings that yield identical fitted values. Minimizing this loss function will reach one (arbitrary) member of such an equivalence class. As a result of this ambiguity, it is generally more effective to interpret the products $\tilde{\boldsymbol{X}}_{\boldsymbol{j}}\boldsymbol{\beta}_{\boldsymbol{j}}\boldsymbol{v}_{\boldsymbol{j}}'$ rather than attempting to interpret the scores parameters $\boldsymbol{\beta}_j$ and loading parameters $\boldsymbol{v}_j$ in isolation. When it is desirable to interpret either the scores or the loadings, we impose the constraint $\|\boldsymbol{v}_j\| = 1$ to make the parameters identifiable. However doing this makes $\boldsymbol{v}_j$ into a parameter whose domain is a manifold (the unit sphere), not a Euclidean space. Inference procedures should respect the domain of a parameter and it is more complicated to consider notions such as confidence sets and standard errors on a non-Euclidean domain.

We address these challenges using what is essentially a modified parametric bootstrap. The conventional parametric bootstrap is used in a setting where we are fitting a generative model f_θ to the observed data. Since this is a generative model, we can simulate datasets from a fitted model $F_{\hat{\theta}}$. This would allow us to generate datasets $Y^{(j)} \sim F_{\hat{\theta}^{(j)}}$, where $\hat{\theta}^{(j)}$ is an estimate of θ. As long as we have a means to obtain the $\hat{\theta}^{(j)}$, thereby yielding the $Y^{(j)}$, we can then fit the AFQR model to each $Y^{(j)}$, yielding parameter estimates $\hat{\boldsymbol{\beta}}^{(j)}, \hat{\boldsymbol{v}}^{(j)}$, and standard errors of any function of the parameters can be calculated empirically.

One popular way to obtain $\hat{\theta}^{(j)}$ is to set $\hat{\theta}^{(j)} \equiv \hat{\theta} + \eta^{(j)}$, where $\hat{\theta}$ is an estimate based on the

observed data, and $\eta^{(j)} \sim N(0, \Psi)$ where Ψ is the sampling variance/covariance matrix of $\hat{\theta}$. An alternative approach is to use the nonparametric bootstrap (sampling from the observations with replacement) to obtain $\hat{\theta}^{(j)}$.

The standard parametric bootstrap is not applicable in our setting for two reasons. One is that the AFQR model is not a generative model, as it describes the conditional quantiles of Y but does not capture the dependence among the observations which should be respected when using the parametric bootstrap. In addition, we do not have a means to obtain the $\hat{\theta}^{(j)}$, as neither of the two approaches presented above are applicable for us. Our procedure does not provide a means to obtain Ψ (in fact, this is the very challenge being discussed here). Moreover, having longitudinal data, the non-parametric bootstrap is not directly applicable, although we note that approaches for non-parametric bootstrapping of longitudinal data do exist [89], [58], [49].

We resolve these challenges by using a surrogate model to "emulate" the data, and then proceed with the remaining parametric bootstrapping steps outlined above. The surrogate model is a Gaussian process regression model fit using parametric maximum likelihood techniques. This model is discussed in chapter 4.

2.3.5 Interpretation

Most non-parametric quantile regression techniques are difficult to interpret because their output is voluminous and unstable, and they often operate as black-box techniques designed for prediction and not interpretation. We have repeatedly mentioned and motivates our techniques as producing results that facilitate interpretation and visualization. In this section we explain how to interpret AFQR, and what hypotheses it can be used to examine.

Since the conditional quantile function is broadly a function of two variables, p and $\boldsymbol{x}$, it can lead to two partial functions. Both these functions provide very different information about the quantile structure of the data, and we discuss them in detail to understand how they can be used to extract the information captured by our model.

2.3.5.1 Quantile curves

Quantiles curves are obtained directly from equation (2.3), by holding X_j fixed $\forall j \in \{1, \cdots, J\}$. These are functions of probability points at which we can compute outcome quantiles, at fixed values of the covariate. In particular, the quantile curve tells us how all the quantiles of the outcome will vary across the range of probability points from $[0, 1]$, conditioned on a fixed value of $\boldsymbol{x}$. We always look at quantile curves in terms of one variable at a time; per covariate j, we fix all the other covariates at their median values. In such a case, the quantile curve is a function with one paarmeter, which is the fixed value of the the covariate being varied (x_j). We define two forms of

quantile curves. The first form includes the central axis and can be defined as follows

$$\begin{aligned} QC(x_j) \ &: [0,1] \to \mathbb{R} \\ QC(p|x_j) &= Q_{Y|\boldsymbol{X}}\big(p|X_j = x_j, \boldsymbol{X_{-j}} = \text{median}(\boldsymbol{X_{-j}})\big) \\ &= g_j(p) \cdot f_j(x_j) + \mu(p). \end{aligned}$$

The second form of quantile curves focuses on the residual quantiles by omitting the central axis, defined as follows

$$\begin{aligned} QC^{\text{resid}}(x_j) \ &: [0,1] \to \mathbb{R} \\ QC^{\text{resid}}(p|x_j) &= Q_{Y|\boldsymbol{X}}\big(p|X_j = x_j, \boldsymbol{X_{-j}} = \text{median}(\boldsymbol{X_{-j}})\big) - \mu(p) \\ &= g_j(p) \cdot f_j(x_j). \end{aligned}$$

The contrast of two quantile curves is a function of $p \in [0,1]$, which takes the following form. For fixed $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X}, \forall p \in [0,1]$

$$QC(p|\boldsymbol{x}_1) - QC(p|\boldsymbol{x}_2) = Q_{Y|\boldsymbol{X}}(p|\boldsymbol{x}_1) - Q_{Y|\boldsymbol{X}}(p|\boldsymbol{x}_2) = \sum_{j=1}^{J} g_j(p) \cdot \big(f_j(\boldsymbol{x}_1) - f_j(\boldsymbol{x}_2)\big).$$

We can see that the contrast is proportional to $g(p)$, and thus lies in a one-dimensional subspace. Recall that we discussed that g is a constant function for a location family in section 2.3.2; this results in parallel quantile curves . Similarly for $J > 1$, if g_j is constant $\forall\ 1 \leq j \leq J$, then the quantile curves at different $\boldsymbol{x}_\cdot$ will be parallel to each other, i.e. they will have the same shape (not necessarily linear) with intercepts determined by the value of the conditioning $\boldsymbol{x}_\cdot$. It is important to note that for non-constant $\{g_j(p)\}_{j=1}^{J}$, the curves can have different shapes; the structure of the model does not impose parallel quantile curves.

2.3.5.2 Effect curves

Effect curves view the conditional quantile estimates as functions of X_j at fixed quantiles of the outcome. For instance, modelling the median outcome as a function of the independent variables is an example of an effect curve at $p = 0$. This is most commonly of interest in classical quantile regression analysis, and so is familiar to most of us. In fact most quantile regression techniques model effect curves individually for each $p \in [0,1]$, as in linear quantile regression. Similar to quantile curves, we always look at one covariate at a time and fix all the others at their median

values. The effect curve upon varying covariate j is defined as a function of x_j with a parameter that lies in $[0, 1]$ as follows

$$\begin{aligned} EC(p) \ &: Range(X_j) \to \mathbb{R} \\ EC(x_j|p) &= Q_{Y|\boldsymbol{X}}\big(p|X_j = x_j, \boldsymbol{X_{-j}} = \text{median}(\boldsymbol{X_{-j}})\big) \\ &= g_j(p) \cdot f_j(x_j) + \mu(p). \end{aligned}$$

Note that we do not need another form of effect curves as the contribution of the central axis $\mu(p)$ is a constant for fixed p, and so does not impact the interpretation or visualization of effect curves much. Contrasting different effect curves at different values of p against the same range of $\boldsymbol{x}$, gives the form

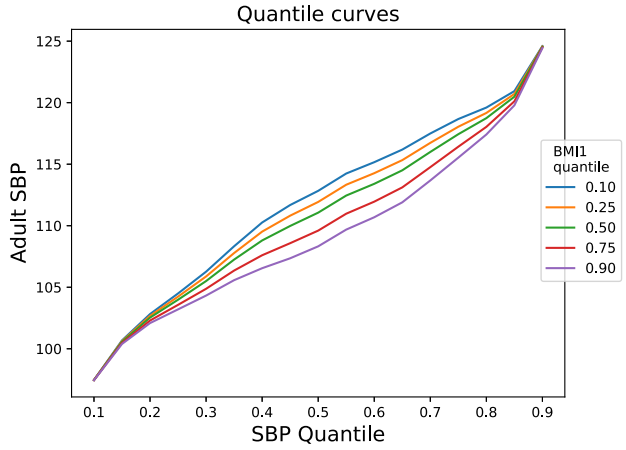
$$Q_y(p_1|\boldsymbol{x}) - Q_y(p_2|\boldsymbol{x}) = \Big(\mu(p_1) - \mu(p_2)\Big) \ + \ \sum_{j=1}^{J} f_j(x_j) \cdot \big(g_j(p_1) - g_j(p_2)\big). \tag{2.14}$$

It is important to note that the difference in equation (2.14) is a function of $\boldsymbol{x}_{\cdot}$. It lies in a two-dimensional subspace which is spanned by $\mathbf{1}$ (the constant function in x_j), and $f_j(x_j)$ (estimated by the data).

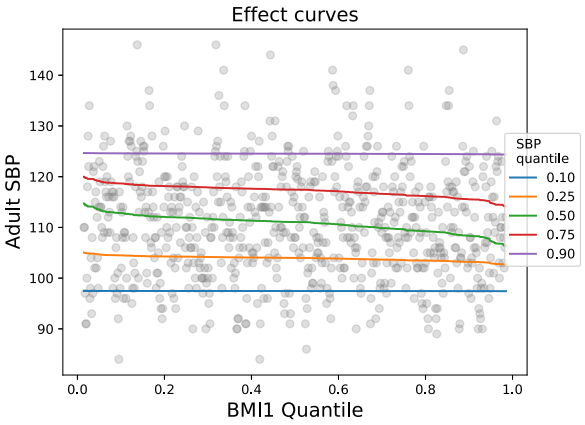
2.3.5.3 Example with anthropometry and blood pressure

We continue to discuss the example introduced in section 2.1.1. An example of effects curves obtained from linear quantile regression was presented in fig (2.4b). The corresponding figure for effect curves obtained from QNN is presented in fig 2.4. Note that effect curves are functions of individual X_j's but here we present the plots in terms of quantiles of X_j so that the median and extreme values of the independent variable (corresponding to outer quantiles) are easily referenced.

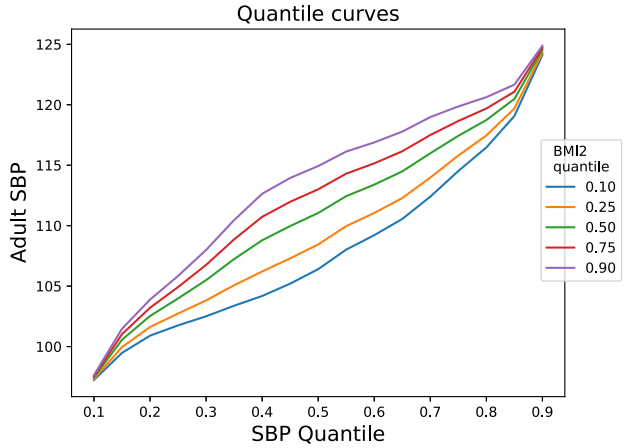
Let us start by commenting on the shape of the two sets of curves presented above. We will focus our attention on the subplots 2.4c, 2.4d, which correspond to the impact of X_2 on Y. Recall that in the example being discussed here, the outcome Y is adult SBP and X_2 is adult HT. The other covariates (X_1, X_3) are childhood BMI and adult HT respectively. Beginning with the quantile curves in figure 2.4c, the green curve represents the median function $\mu(p)$. Each of the other curves are quantile curves at different values of X_2 (adult BMI), which deviate from the median, but are similar in shape to the median function. We can see that the extreme outer quantiles are quite similar for all X_2, however quantiles between $0.20 - 0.80$ deviate from one another. They all seem to have a sigmoidal shape, but with different curvatures. Change in curvature depends on the value of $f_2(X_2)$ which is multiplied by the direction of deviation given by $g_2(p)$ $p \in [0, 1]$. Larger



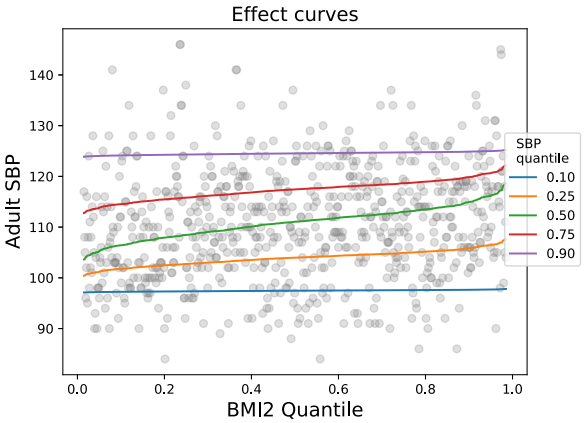
(a) Quantile curves: $\text{SBP}_2 \sim \text{BMI}_1$



(b) Effect curves: $\text{SBP}_2 \sim \text{BMI}_1$



(c) Quantile curves: $\text{SBP}_2 \sim \text{BMI}_2$

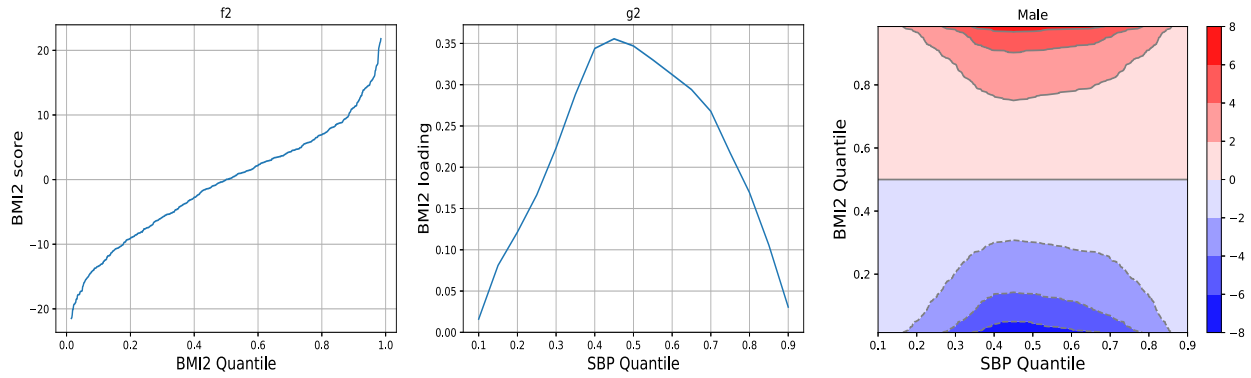


(d) Effect curves: $\text{SBP}_2 \sim \text{BMI}_2$

Figure 2.4: Adult SBP between ages $[18, 24]$ years as a function of (row 1) childhood BMI around ages $[1.5, 4.5]$ years, and (row 2) adult BMI. All cases are for males only.

$f_2(X_2)$ leads to greater deviations of the quantile curve from the central axis.

To understand the link between the quantile/ effect curves and $\hat{f}_2, \hat{g}_2$ estimated from the low rank structure, we present $\hat{\boldsymbol{u}}_{\boldsymbol{2}}, \hat{\boldsymbol{v}}_{\boldsymbol{2}}$, which are discretized vectors of $\hat{f}_2, \hat{g}_2$ in image 2.5 below. The heatmap in figure 2.5 represents $\hat{f}_2(Y) * \hat{g}_2(X_2)$, where Y (adult SBP) varies along the x-axis and the independent variable X_2 (adult BMI) varies along the y-axis. Each point on this heatmap corresponds to a certain quantile of the outcome (say p_y) and independent variable (say p_x), and gives the estimated deviation of the p_y^{th} quantile of SBP conditioned on the p_x^{th} quantile of adult BMI from the y^{th} quantile of SBP at median adult BMI. Recall that all other covariates (X_1 (childhood BMI), X_3 (adult HT) and both childhood and adult ages) are held at their medians in the results presented in fig 2.5.



(a) Estimated $\boldsymbol{u_2}$ (b) Estimated $\boldsymbol{v_2}$ (c) Estimated $\boldsymbol{uv'}$

Figure 2.5: Low rank decomposition of adult SBP among males, in terms of childhood BMI and adult body size (BMI, HT). Impact of adult BMI on adult SBP is displayed here. Childhood ages $\in [2.5, 3.5]$ years, and adult ages $\in [19, 23]$ years.

All the information in the quantile curves and effect curves is actually present in the $\boldsymbol{u}$, $\boldsymbol{v}$ estimates and can be inferred from figure 2.5. We explain this link in more detail below.

Quantile Curves

Recall that adult HT (X_2) impacts the outcome quantiles in the form $\mu(p) + f_2(X_2) \cdot g_2(p)$ $p = p_y \in [0, 1]$, where $\mu(p)$ is the quantile function of adult SBP (Y) conditioned on the median value of all the covariates, including median adult HT. So if we look at the distribution of adult SBP for those having low adult HT ($p_x = 0.1$ quantile of adult HT) or high adult HT ($p_x = 0.9$ quantile of adult HT), all distributions will vary from the median $\mu(p)$ in a similar way - they will vary most towards the middle and will vary less as we move to the tails. This is determined entirely by the shape of $\boldsymbol{v_2}$. As seen in figure 2.5 $\hat{\boldsymbol{v}}_{\boldsymbol{2}}$ is close to 0.1 at the outer quantiles and around 0.4 at

the median, implying that the quantile curves at different X_2 will have smaller deviations from the central axis at the tail probabilities, roughly $0.1/0.4 = 1/4^{th}$ of the deviations at the inner quantiles ($p_y \in [0.4, 0.6]$).

The extent of deviation at a specific $X_2 = x_2$ is determined by the curvature of $f_2(x_2) \cdot g_2(p),\ p \in [0, 1]$. In general, as $\{g_j\}_j$ are not necessarily linear, the curvature can be inflated or deflated, and thus lead to conditional quantile curves of different shapes. The curvature is determined by $\boldsymbol{u_2}$. By design, if X_2 takes its median value, the quantile curve is the same as the central axis; this is always satisfied and is the reason why f_2, like all f functions, crosses the x-axis at the median of X_2. Lastly, unlike effect curves which are described in more detail below, quantile curves do not need to be stacked one on top of the other. Going back to our example in figure 2.5, $\hat{\boldsymbol{u}}_{\boldsymbol{2}}$ is an increasing function which crosses 0 at 0.5, and since $\hat{\boldsymbol{v}}_2$ is positive throughout, the deviations of quantile curves from the central axis at $X_2 < Q_{X_2}(0.5)$ will be negative and for $X_2 > Q_{X_2}(0.5)$ will be positive. This means that for quantiles $p_x^1 < 0.5$, the quantile curves corresponding to $X_2 = Q_{X_2}(p_x^1)$ will lie below the central axis μ, and vice-versa. We see exactly this in subfigure 2.4c.

Effect curves

An effect curve is evaluated at a fixed p and is a function of X_j of the form $\mu(p) + f_j(X_j) \cdot g_j(p) = c_1 + f_j(X_j) \cdot c_2$ for constants $c_1, c_2 \in \mathbb{R}$. The shape of the effect curve is purely determined by $f_j(X_j)$, the intercept is determined by $\mu(p)$ and the curvature by $g_j(p)$. In subfigure 2.5 we can see that $\hat{\boldsymbol{v}}_{\boldsymbol{2}}$ is sigmoidal, which is why we'd expect the effect curves to be sigmoidal as well. Next, let us explain the varying curvature and intercepts of the curve.

The intercept of the effect curves is exactly equal to $\mu(p) \in \mathbb{R}$. We expect $\mu(p)$ to be increasing in p, which is indeed the case here and is why the effect curves are stacked one on top of the other in increasing order of p. Lastly, the curvature is determined by $g_2(p)$. We can see in subfigure 2.5 that $\hat{\boldsymbol{v}}_{\boldsymbol{2}}$ (discretized $\hat{g}_2$) has an inverted-U shape. This means that the curvature will be the largest for the inner quantiles, and will become smaller as we move towards the outer quantiles (a flat line has curvature 0), since $\hat{\boldsymbol{v}}_{\boldsymbol{2}}$ goes to zero towards both ends of its domain. Since $\hat{\boldsymbol{v}}_{\boldsymbol{2}}$ is positive throughout, the slope of all effect curves (each at a distinct quantile of the outcome) will be positive, however the effect curves will become more constant at the outer quantiles. This is exactly what we see in 2.4d.

We would like to clarify that the regularization in AFQR does not prevent the quantile crossing problem. Our hope is that the regularization brings the estimates close to the truth, and so we would expect the quantiles to not cross. Also, we do not impose a monotonicity condition on the quantile curves, i.e. the quantile curves at fixed $\boldsymbol{x}$ are not forced to be non-decreasing functions. Again, we hope our estimates will achieve this due to the quality of our model. Both these conditions are

met in our applications of this method on the Dogon and NHANES datasets, as well in simulation studies.

2.3.6 Simulation study

In this section we discuss simulation studies conducted to assess the performance of AFQR. We provide the integrated mean squared error for the low rank estimates. We consider homogeneous as well as non-homogeneous data, and evaluate the estimates for both. We start by describing the simulation setup, followed by the results.

2.3.6.1 Data generation

We consider two data types, a location family which gives homogeneous data, along with a scale family which yields non-homogeneous data. Recall that the conditional quantile functions of the outcome (per p) are parallel to each-other in the homogeneous case. The two data generating models are given below. The distributions used to generate the data in this study have been picked such that the conditional quantile function of the outcome has a closed form.

Case 1: Gaussian Case

$$\begin{aligned}
&X_1 \sim \mathcal{U}(1,2) \\
&X_2 = X_1 + \epsilon_{X_2}, && \epsilon_{X_2} \sim \mathcal{U}(-1,1) \\
&X_3 = X_1 + \epsilon_{X_3}, && \epsilon_{X_3} \sim \mathcal{U}(-1,1) \\
&Y = a_1 X_1 + a_2 X_2 + a_3 X_3 + \epsilon_Y, && \epsilon_Y \sim \mathcal{N}(0, 0.5)
\end{aligned}$$

In this case, the conditional quantile function has a closed form given by

$$Q_{Y|(X_1,X_2,X_3)}(p|x_1,x_2,x_3) = a_1 x_1 + a_2 x_2 + a_3 x_3 + 0.5 \cdot Q_p(\mathcal{N}(0,1)).$$

Case 2: Exponential Case

$$\begin{aligned}
&X_1 \sim \mathcal{U}(1,2) \\
&X_2 = X_1 + \epsilon_{X_2}, && \epsilon_{X_2} \sim \mathcal{U}(-1,1) \\
&X_3 = X_1 + \epsilon_{X_3}, && \epsilon_{X_3} \sim \mathcal{U}(-1,1) \\
&Y \sim exp(\lambda), \quad 1/\lambda = X_1 + X_2 + X_3
\end{aligned}$$

We selected the exponential distribution since it is among the few distributions that have a closed form conditional quantile function

$$Q_{Y|(X_1,X_2,X_3)}(p|x_1,x_2,x_3) = -log(1-p)(x_1+x_2+x_3),$$

that is not additive in the probability point (p), and the covariate values (x_1, x_2, x_3). This is unlike the conditional quantile function in the Gaussian case.

2.3.6.2 Simulation Design

We generate data with $J = 3$ covariates and an outcome variable. The number of observations per setting varies along $n^s \in \boldsymbol{n} = [2000, 4000]$. Per sample size, we use the setup described above to randomly sample n^s iid copies of (X_1, X_2, X_3, Y). In this section, let $\boldsymbol{X} = (X_1, X_2, X_3)$, and denote the median z-scored simulated data by $\boldsymbol{X}^z = (X_1^z, X_2^z, X_3^z)$.

In step 1 of AFQR, we pick m to be an odd number so that the median along any grid of length m is an entry in the grid, and not the average of two entries. In other words, $m/2 \in \mathbb{N}$. In particular, we set $m = 9$ and define $\boldsymbol{p}$ as $[0.1, 0.2, 0.3, \cdots, 0.9]$. We use the same smoothing penalty for all $\boldsymbol{u}_j$ and all $\boldsymbol{v}_j$, i.e. $c_{u_j} = c_1, c_{v_j} = c_2, \ \forall j$, where c_1, c_2 are picked via grid search over a sparse grid of length 6. The results discussed here correspond to $c_1 = 1000, c_2 = 1000$.

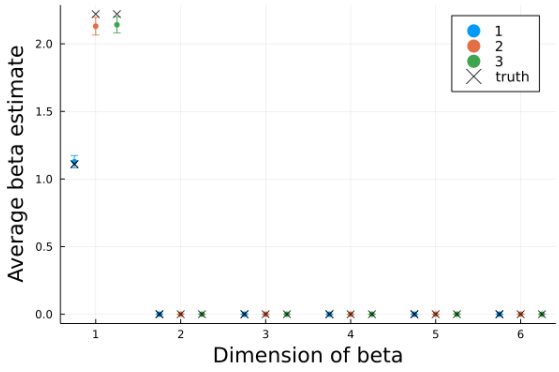
2.3.6.3 Simulation Results

In the Gaussian case, we explore the performance of AFQR when the underlying structure is homogeneous across quantiles. Good performance in this case demonstrates our method's ability to obtain consistent results even at the outer quantiles, which is often difficult due to limited data in the tails. Our method provides more stable results (reduced variance) for outer quantiles by partially pooling information from other quantiles, however our simulations show that this borrowing of information does not induce bias on our estimates.

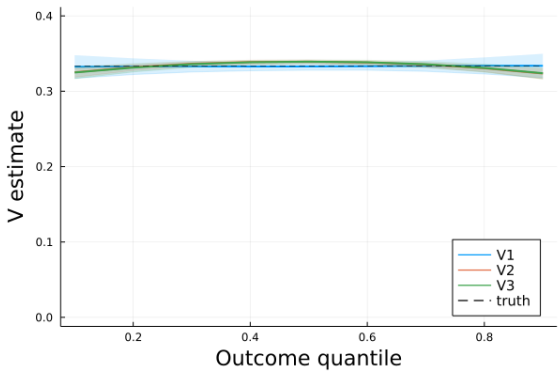
In the Exponential case, good performance demonstrates the method's ability to model heterogeneous effects characterized by non-trivial $\{g_j\}_j$ functions. We compare the parameter estimates $(\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}, \ \hat{\boldsymbol{v}}_{\boldsymbol{j}})$ with the underlying value of the parameters, which is easily derived.

We are able to estimate both $\boldsymbol{\beta}_{\boldsymbol{j}}$ and $\boldsymbol{v}_{\boldsymbol{j}}$ well in terms of bias and variance. The point-wise average estimate along with standard deviation bands for both the Gaussian and Exponential cases are presented in figure 2.6 below. The black dashed line or cross represents the population parameter, and is labeled as"truth" in all the plots. These results correspond to a sample size of $n^2 = 2000$. The results for $n^s = 4000$ were better and scaled by roughly $1/\sqrt{2}$ as expected. We present results for $n^s = 2000$ since the sample size in one of our applications is of this order and we want to illustrate that AFQR does not need a large sample size to work well. The relative MSEs for the first entry of β_j and $\boldsymbol{v}$ are summarized in table 2.2, which are all below 15% for the exponential case and below 5% for the Gaussian case. Specifically $\boldsymbol{v}$ is estimated over a grid of

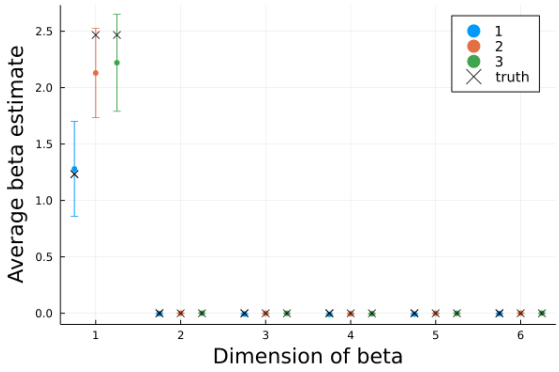
$m = 9$ probability points, so we present the average pointwise absolute relative bias and average pointwise relative mse.



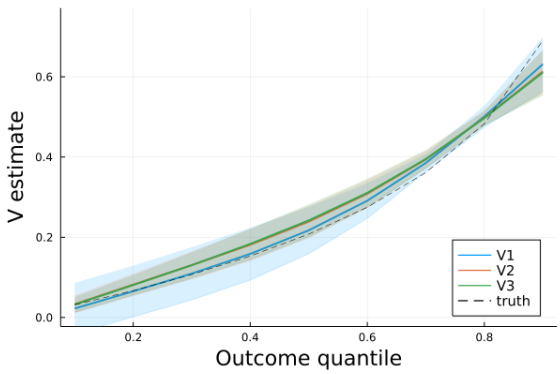
(a) Gaussian: $\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}$



(b) Gaussian: $\hat{\boldsymbol{v}}_{\boldsymbol{j}}$



(c) Exponential: $\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}$



(d) Exponential: $\hat{\boldsymbol{v}}_{\boldsymbol{j}}$

Figure 2.6: Simulation results for flr_reg. The figures present average estimates over 100 replicates, with estimate standard deviation bands.

		β_1		$\boldsymbol{v}$	
Distribution	Dimension	rel bias	rel mse	Avg |rel bias|	Avg rel mse
exp	1	0.037	0.144	0.074	0.033
exp	2	-0.136	0.109	0.132	0.011
exp	3	-0.099	0.098	0.131	0.011
normal	1	0.016	0.002	0.002	0.001
normal	2	-0.040	0.005	0.015	0.001
normal	3	-0.036	0.004	0.015	0.001

Table 2.2: Bias and variance of AFQR parameter estimates from the simulation study.

2.3.7 Illustration using Dogon Longitudinal Study data

In this section, we apply AFQR to the Dogon data. We fit the model separately for men ($n \approx 2400$) and women ($n \approx 2250$), across different feature settings. Recall that we are interested in exploring how the quantiles of adult SBP vary with childhood body size and adult body size. Thus, our primary features of interest are childhood and adult body size measures. We begin by describing the variables used in the models in more detail. Note that we always include childhood and adult ages in all fitted models, but omit mentioning them in model specification throughout this section.

2.3.7.1 Independent variables

Adult body size measures

We use adult height (HT) and adult BMI to control for adult body size as this accounts for the two different drivers of SBP. Taller people generally tend to have higher blood pressure. This is also true for heavier people. The advantage of using BMI over weight (WT) is that BMI by definition controls for the expected increase/decrease in weight due to an individuals height. Thus, we expect BMI to be less correlated with HT than WT.

Childhood body size measures

For childhood body size, we consider six different childhood body size measures namely childhood height, weight, BMI, height adjusted z-score (HAZ), weight adjusted z-score (WAZ) and BMI adjusted z-score (BAZ). In our dataset, all features (including the six childhood body size variables) are observed at different ages for different individuals. Because there is no common age at which all individuals have a measurement, we define childhood body size in the two ways presented below, which allow us to model age functionally.

Fixed age via kernel smoothing The first way is by localizing to a childhood age and accommodating arbitrary non-temporal sampling via kernel smoothing. We pick a childhood age, say $\mathrm{age}_1 = 5$ and consider all observations weighted by the difference between the age at which the observation is recorded and age_1. We have seen results from this model sprinkled throughout the document as the running example, such as in figures 2.3, 2.4, 2.5.

Growth trajectory The second way aims at accounting for each child's growth trajectory instead of their body size at a specific age (age_1). As mentioned in the introduction, there are two broad steps involved in doing so. First, we use multiple imputation to impute the childhood body size per individual at a grid of ages 1 through 10. We use an MLE-based Gaussian Process model to impute the values. The details of the model can be found in Chapter 4. Second, we conduct PCA

on the imputed childhood body size variable at the 10 ages, and report the dominant score. This score is used as a summary variable capturing the growth trajectory of a child, and is referred to as "pcscore". For instance, if the growth trajectory is computed in terms of childhood body size measure BMI, then the score will be referred to as *BMI pcscore*.

2.3.7.2 Working with longitudinal data

We adopt a marginal approach to obtain conditional quantile estimates via QNN on longitudinally observed data. This entails three steps: (i) localize to two ages, e.g., $\text{age}_1 = 3$ and $\text{age}_2 = 20$ to represent childhood and adult states with a caliper for each age, (ii) construct a long-format dataset with all within-subject records obtained by combining data within the calipers, and (iii) use kernel weights to focus the analysis on the target ages. For instance, in figure 2.5, the ages are $2, 21$ and the calipers are $0.5, 2$ respectively. Table 2.3 displays the number of individuals retained in the reformatted dataset(s) for which we provide results in this section. Notice that localizing to even a small childhood age such as 1 does not lead to a drastic loss in the number of individuals; we retain a bit less than 50% of the individuals in the filtered datasets for $\text{age}_1 = 1$ and around 50% for $\text{age}_1 = 3$.

Childhood body size	age1	age2	sex	No. of unique ID	No. of observations per ID average	sd	maximum
WT	1	20	Female	339	4.811	2.916	15
WT	1	20	Male	393	4.463	2.965	15
BMI	1	20	Female	338	4.790	2.913	15
BMI	1	20	Male	392	4.436	2.942	15
WT	3	20	Female	390	5.818	3.510	18
WT	3	20	Male	426	5.603	3.408	18
BMI	3	20	Female	390	5.769	3.482	18
BMI	3	20	Male	426	5.601	3.411	18
			Female	833	13.858	3.206	19
Raw data			Male	866	13.708	2.851	20

Table 2.3: Number of individuals (No of unique IDs), and the average, standard deviation and maximum value of the number of observations per ID in the Dogon data with and without localizing to specific ages. The last two rows of the table correspond to the raw Dogon data after removing missing values for the childhood body size variable, HT, BMI and SBP (at all ages in the data).

2.3.7.3 Incorporating binary variables

In our application, we consider whether females are pregnant at the time of their SBP measurement since pregnancy is known to have an impact on SBP especially in the later months. Pregnancy is a binary variable which takes a value of 1 for pregnant and 0 otherwise. For a binary variable in AFQR, say X_b, there will be no score ($\boldsymbol{u}_{X_b}$) since the variable only takes two values. There is a vector of loadings $\boldsymbol{v}_{X_b}$, which is equal to the additive component of the binary variable in the low rank model, as $(1-0)\cdot \boldsymbol{v}_{X_b} = \boldsymbol{v}_{X_b}$. Thus, in the additive model structure as in equation 2.3, $f_{X_b} \equiv 1$. This means that the p^{th} quantile of Y will be $g_{X_b}(p)$ units greater for pregnant women ($X_b = 1$) as compared to non-pregnant women ($X_b = 0$), if for instance X_b is the indicator for pregnancy among women.

2.3.7.4 Method parameters

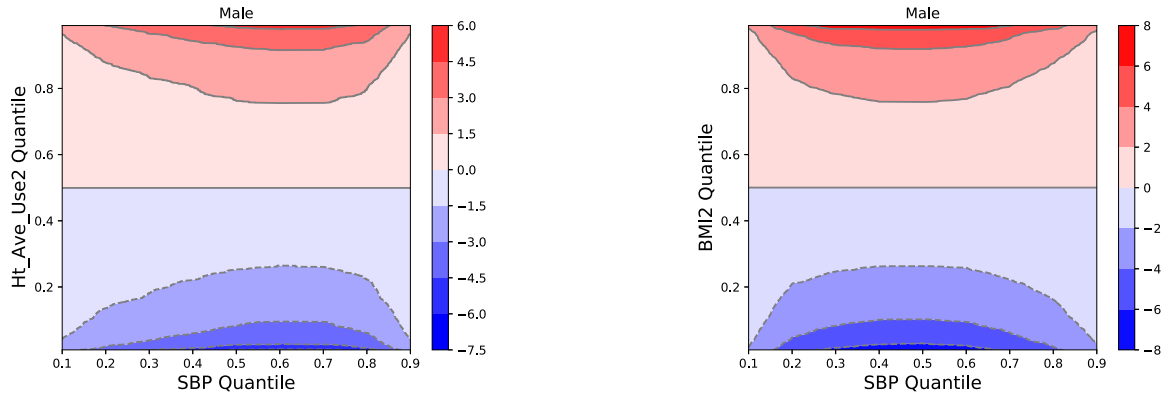
We obtain the conditional quantile estimates for adult SBP via QNN using the recommended parameters - lasso penalty of $\lambda = 0.1$ and $k = 5$ neighbors. In AFQR, we consider a radial basis transformation of dimension five. In particular, the first column is a linear transformation. The other five radial transformations are based on the observed values of the variables. In particular, for the j^{th} variable, we consider an equi-spaced grid of five points through the range the $\boldsymbol{x_{\cdot j}}$ and use these points as the centers of the radial transformations. We use the sample standard deviation as the scale of the transformation. For instance, the k^{th} radial transform for the j^{th} variable is given below, where $2s$ is the sample standard deviation of $\boldsymbol{x}_{\cdot j}$.

$$h_k : \boldsymbol{x_{\cdot j}} \rightarrow \exp\left(-\frac{(\boldsymbol{x_{\cdot j}} - c_k)^2}{2s^2}\right) - \exp\left(-\frac{c_j^2}{2s^2}\right),$$
$$c_k = \min(\boldsymbol{x_{\cdot j}}) + \frac{(k-1)}{4}(\max(\boldsymbol{x_{\cdot j}}) - \min(\boldsymbol{x_{\cdot j}})), \quad k = 1, \cdots, 5.$$

In addition to this, We vary the smoothing bandwidth in AFQR along a grid of length three, $(c_{u_j}, c_{v_j}) \in [(1, 500), (100, 1000), (200, 1000)]$ where the smoothing bandwidth per component is the same. Here we present the results for a smoothing bandwidth with $c_{u_j} = 100$ and $c_{v_j} = 1000$, for all $j = 1, 2, 3$.

2.3.7.5 Results

In this section we exclusively present point estimates. We provide results for six varying childhood body size measures (cbs) and for fixed adult body size measures (HT and BMI) across the two



(a) SBP $\sim$ Adult HT

(b) SBP $\sim$ Adult BMI

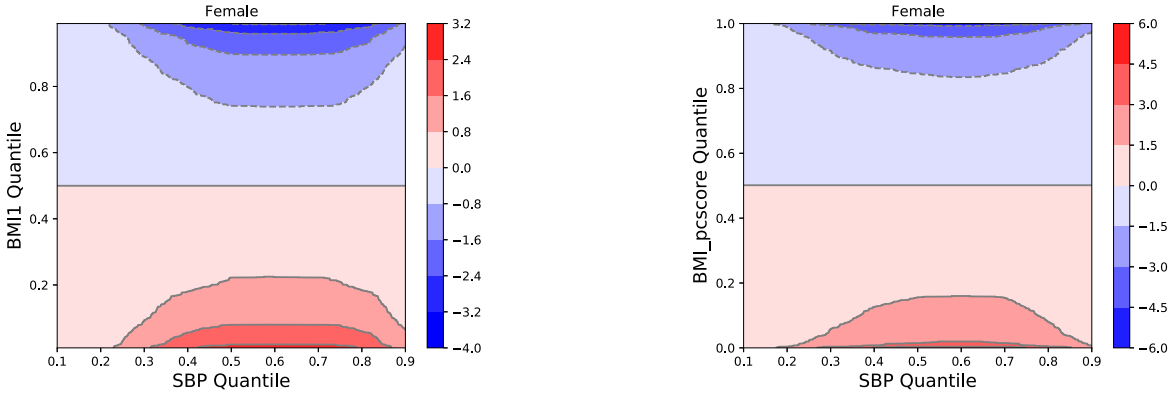
Figure 2.7: Impact of adult height (left) and adult BMI (right) on adult SBP for males, at ages 3 and 21.

genders. Thus, the general form of all models we will consider will be

$$\text{SBP} \sim \text{cbs} + \text{HT}_2 + \text{BMI}_2 + \text{age}_1 + \text{age}_2,$$

where $A_1,\ A_2$ denote variable A measured at a childhood age (age_1), adult age (age_2) respectively. We can divide the independent variables into two categories, adult body size and childhood body size. The impact of adult body size onto adult SBP is expected to be positive (as they have a cross-sectional relationship). This means that we expect bigger (taller and heavier) people to have higher SBP (see figure 2.7b). We see this across all the six childhood body measures considered in this data application.

The impact of childhood body size on adult SBP is less straightforward, and is more interesting. We find that childhood body size has an inverse relationship with adult SBP. For instance for childhood BMI, children with below-median BMI either at a specific age or at their BMI pc-scores tend to have higher adult SBP and children with higher BMI tend to have lower adult SBP. This can be clearly seen in the heatmaps presented in figure 2.8; the heatmaps display the fitted low rank structure corresponding to childhood BMI. Every point on the heatmap displays the deviation of the adult SBP quantile (at a probability point corresponding to the x-axis) for the subpopulation of people whose childhood BMI was at some quantile (at a probability point on the y-axis) from the same quantile of adult SBP among people who had median childhood BMI. In both cases, all the other covariates are held at their median values. Thus, each point on the heatmap (figure 2.8) is of the form (p_y, p_x) and corresponds to $Q^c_{Y|\boldsymbol{X}}\big(p_y|X_1 = Q_{X_1}(p_x), \boldsymbol{X}_{-1} = median(\boldsymbol{X}_{-1})\big)$, which is the deviation of the conditional quantile from the central axis at p_y. Here



(a) Childhood age $\in [1.5, 4.5]$

(b) Childhood PC score

Figure 2.8: Effect of childhood BMI on adult SBP (a) when localizing to childhood age 3 and adult age 20, and (b) by considering the BMI pcscores as a summary of childhood growth trajectory and all adults above the age of 12. Results are for females.

$\boldsymbol{X} = (BMI_1, HT_2, BMI_2, age_1, age_2)$.

Figure 2.8a displays the effect of childhood BMI on SBP when localizing to specific childhood and adult ages. Plot 2.8b displays this effect when we do not localize to either childhood or adult age, and instead use BMI pcscore to model childhood body size, and simply consider all individuals above the age of 12. We can see that the direct effect is larger in magnitude if we do not restrict to a specific adult age or childhood age. We would like to point out that the effect size is in between the two presented here for the other case (with localized childhood age and all adult ages above twelve).

Next, we note that the effect of childhood BMI (modelled by localizing to a specific childhood age or via the BMI pcscore) is a lot more prominent in the inner and upper quantiles of adult SBP than the lower quantiles. While this trend can already be identified in the heatmaps above, it is more easily seen via quantile curves. In figure 2.9, we provide the quantile curves for the impact of childhood BMI localized to ages 1, 3, and 5 on the adult SBP among females. It appears that the lower quantiles of adult SBP are not affected as much by childhood BMI, but the upper quantiles deviate more. For instance, at childhood age 3 (ref figures 2.9c, 2.9d), SBP quantiles at any probability point beyond 0.2 deviate from each other as childhood BMI varies but they do not vary as much below the 20th percentile of adult SBP. The same trend is observed at childhood age 5 with deviations after the 30th percentile of SBP (ref figures 2.9e, 2.9f).

This shows the adverse effect of being undernourished in childhood, which is relatively common in Mali. Undernutrition may lead to poor development of organs, which may affect adult health in the form of SBP (lead to elevated SBP). The results in figure 2.9 show that upper SBP

quantiles are substantially lower for bigger children than for smaller children. Bigger children having a lower upper quantile of SBP shows that bigger children who grow up to be "unhealthier" in terms of SBP (have high SBP), still have lower SBP than malnourished children who may grow up to be small adults. This is against the cross-sectional effect of body size and SBP which is well established, expected, and generally much stronger than the longitudinal-effect of childhood body size on SBP. While the lower end of adult SBP is usually not of interest, we observe that bigger children also have a lower low-quantile of adult SBP than smaller children, which is desirable.

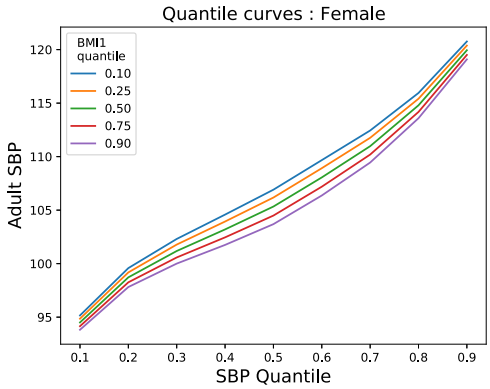
In other words, for the subpopulation of malnourished children versus healthy children, not only is the median SBP of the former subpopulation larger than the latter, but the upper end of SBP quantiles for the undernourished subpopulation is also larger. Since high SBP is risky and the main area of interest in most SBP-related studies, these results suggest that being bigger as a child is only helpful and being undernourished can be dangerous in the long run. This provides motivation for the mediation analysis approach in the next chapter of this thesis; we will come back to this and explain it in further detail later.

For females at childhood age 1, we see the same pattern of lack of dispersion in the lower quantiles of adult SBP. However, in this case, the lower and upper quantiles of SBP both have less dispersion and most of the action is in the inner quantiles. For males, the impact of childhood BMI is smaller but it too leads to more dispersion in the mid-upper quantiles of SBP with virtually no impact of BMI on the lower quantiles.

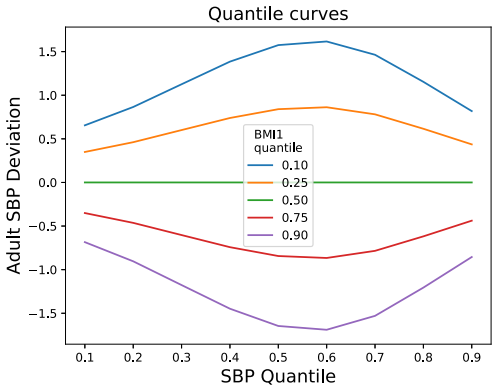
So far we have discussed childhood BMI as the childhood body size measure, but childhood weight has a greater impact on adult SBP than childhood BMI amongst females. The impact of childhood WT is also a lot more dispersed at the mid-upper quantiles of adult SBP, while the lower quantiles of adult SBP seem to not be affected by childhood WT (ref figure 2.10). Also, cross sectional HT has a bigger impact on adult SBP, especially at the lower quantiles of adult HT meaning shorter females of median BMI have lower adult SBP upon controlling for childhood WT as opposed to controlling for childhood BMI. This can be seen in figure 2.11.

The impact of childhood WT on adult SBP among males is much weaker, however it captures a different trend. In this case, it appears that the distribution of adult SBP for lighter children is not as dispersed as the SBP distribution for heavier children. In particular, for lighter boys the lower quantiles of adult SBP are high and the upper quantiles are low (less dispersion) whereas heavier boys have lower low-SBP quantiles and higher upper SBP quantiles (meaning heavier tails and more dispersion). Further, we see the opposite trend in terms of dispersion among the adult SBP quantiles due to varying childhood weight, as compared to that seen among females in figure 2.10. We present quantile curves at ages 1, 3, and 5 to see the difference in patterns among males in figure 2.12.

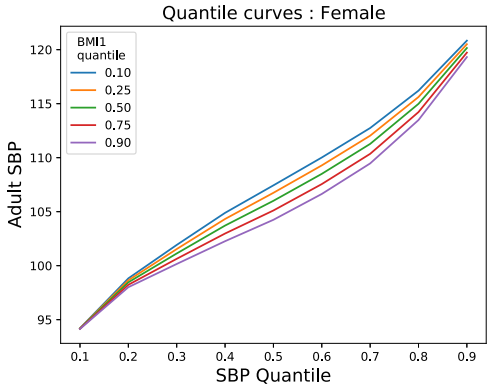
We can see that weight centered at age 1 has the opposite effect, wherein the lower quantiles



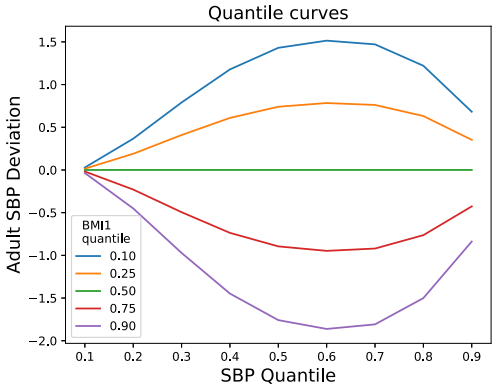
(a) SBP : age 1



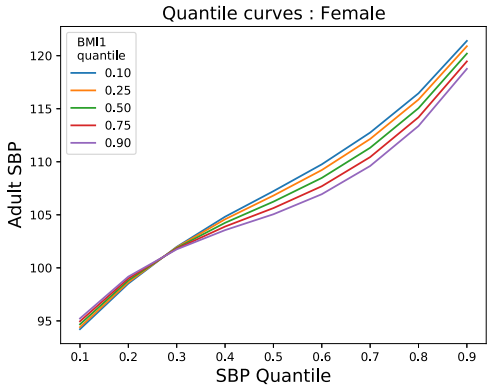
(b) SBP deviations : age 1



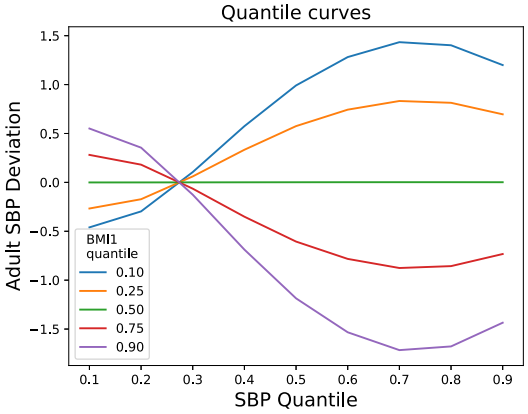
(c) SBP : age 3



(d) SBP deviations : 3

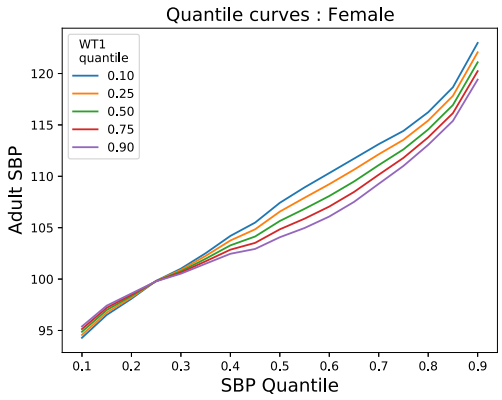


(e) SBP : age 5

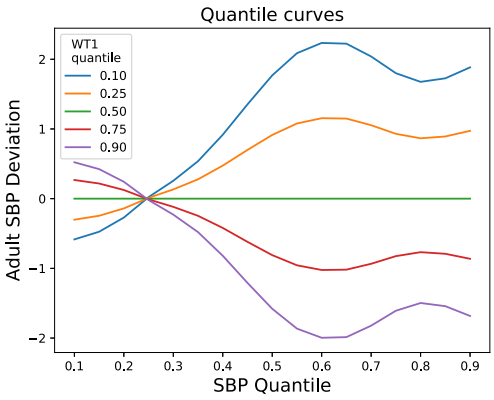


(f) SBP deviations : age 5

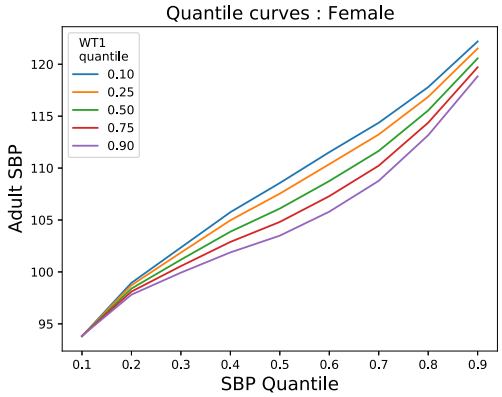
Figure 2.9: Quantile curves capturing the effect of childhood BMI measured at age groups centered at different childhood ages. Each row corresponds to a childhood age of 1, 3, or 5. The first column (left) displays the SBP (with the central axis added to the deviations): $\mu(p)+f_{BMI_1}(bmi_0)g_{BMI_1}(p)$ and the second columns (right) displayed deviations from the central axis : $f_{BMI_1}(bmi_0)g_{BMI_1}(p)$ at 5 fixed BMI values (bmi_0).



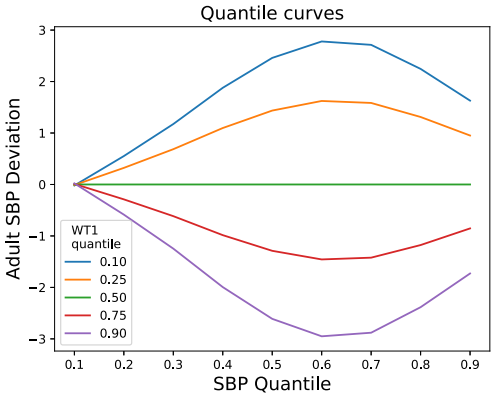
(a) SBP : age 1



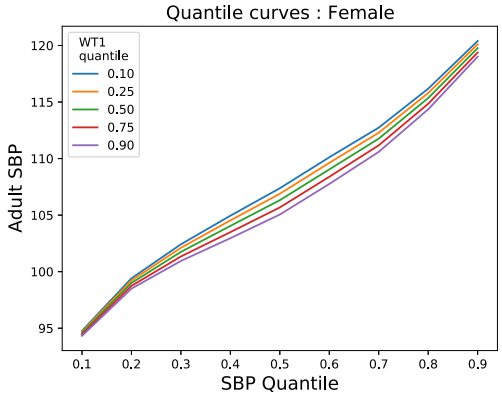
(b) SBP deviations : age 1



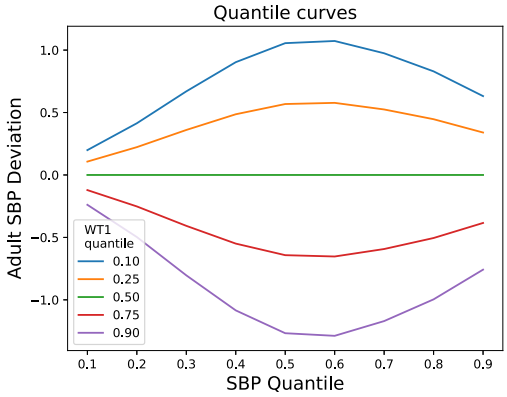
(c) SBP : age 3



(d) SBP deviations : 3

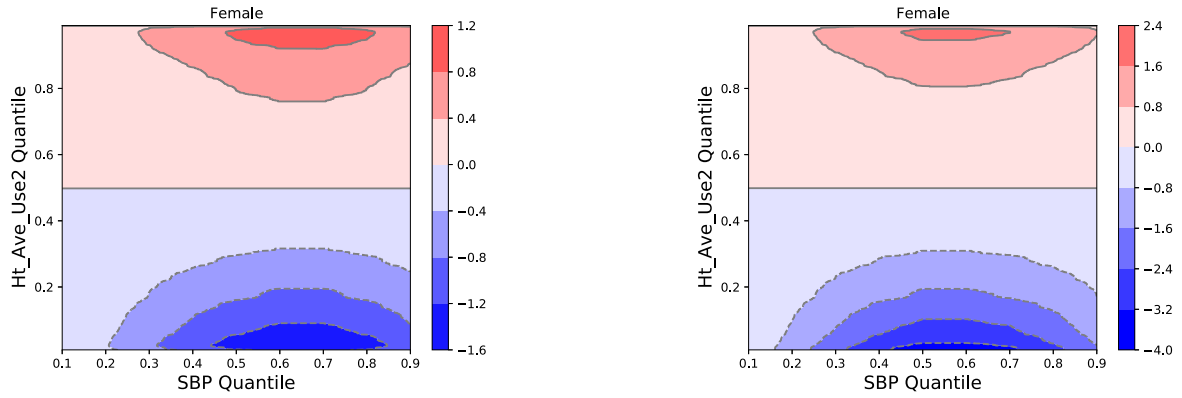


(e) SBP : age 5



(f) SBP deviations : age 5

Figure 2.10: Quantile curves capturing the effect of childhood WT measured at age groups centered at different childhood ages. Each row corresponds to a childhood age 1, 3, or 5. The first column (left) displays the SBP (with the central axis added to the deviations): $\mu(p) + f_{WT_1}(wt_0)g_{WT_1}(p)$ and the second columns (right) displayed deviations from the central axis : $f_{WT_1}(wt_0)g_{WT_1}(p)$ at 5 fixed WT (wt_0) values.



(a) Childhood BMI

(b) Childhood WT

Figure 2.11: Impact of adult height on adult SBP for females when controlling for childhood BMI verses childhood weight at age 3.

of adult SBP deviate more while the upper quantiles converge for varying childhood WT among males. At age 3, childhood WT no longer has an inverse relationship with adult SBP throughout the SBP quantiles. Instead, variation is present in the childhood WT distribution; adult SBP is less dispersed for lighter males and more dispersed for heavier males. This can be seen more easily through effect curves, which are presented in figure 2.12f. Here, we can see that SBP quantiles are less dispersed towards the lower quantiles of childhood WT (lighter males) and more dispersed at the upper quantiles (heavier males).

Results for childhood body size height and HAZ were very weak especially for the across-time association, which is why we do not include those results here.

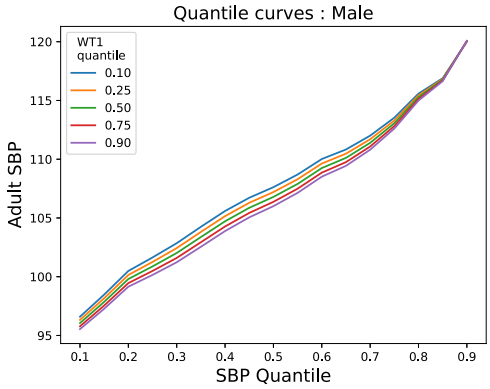
2.3.7.6 Illustration of inference results

Here we present some limited results for applying the inference procedure discussed above to the Dogon data. Figure 2.13 contain plots of the functional components in AFQR for the model

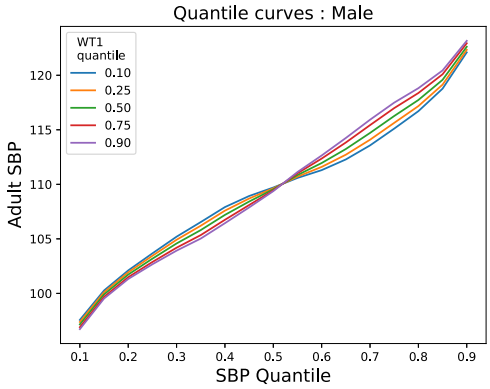
$$\text{SBP} \sim \text{WT}_1 + \text{HT}_2 + \text{BMI}_2 + \text{age}_1 + \text{age}_2$$

at a childhood age $\text{age}_1 = 1$ and adult age $\text{age}_2 = 20$ for females. The point estimates of the effect curves corresponding to this setting were previously seen in figures 2.10a, 2.10b.

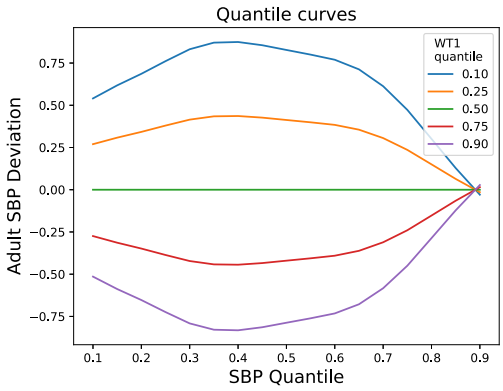
Figure 2.13 contains "simultaneous" bands for the low rank vectors $\boldsymbol{u}_j, \boldsymbol{v}_j$ for $j = 1, 2, 3$. The simultaneous band for a vector $\boldsymbol{w}$ is constructed using the point-wise bootstrap estimates of $\boldsymbol{w}$. In particular, let $\widehat{\mu(\boldsymbol{u})}$ and $\widehat{\sigma(\boldsymbol{u})}$ denote the point-wise bootstrap average value and standard deviation of the $\boldsymbol{u}$ estimates. The simultaneous band is formed by finding the smallest value of $\gamma > 0$ such



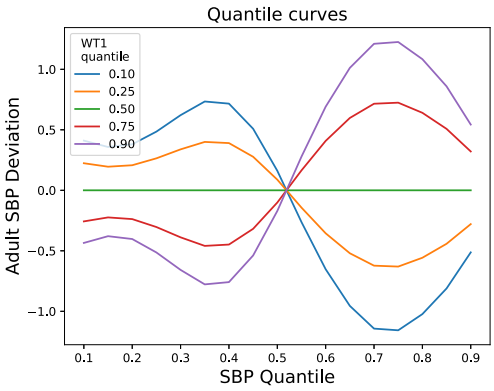
(a) Quantile curve : age 1



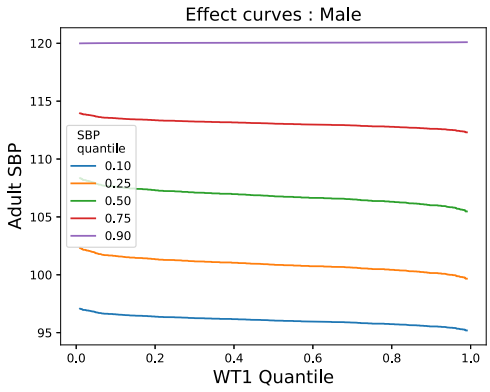
(b) Quantile curve : age 3



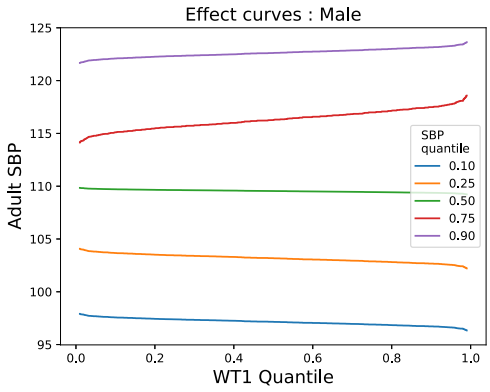
(c) Quantile curve deviations : age 1



(d) Quantile curve deviations : age 3



(e) Effect curve : age 1



(f) Effect curve : age 3

Figure 2.12: Results for childhood WT on adult SBP among males. The columns correspond to childhood ages of 1 and 3 respectively. The first row presents the quantile curves obtained by fixing WT values in the additive components of childhood WT in AFQR : $\mu(p) + f_{WT_1}(wt_0) \cdot g_{WT_1}(p)$. The second row displays the deviations of quantiles curves at the same WT values from the central axis $\mu(p)$. The last row displays the effect curves which are functions of childhood WT at fixed SBP probability points : $\mu(p_0) + g_{WT_1}(p_0) \cdot f_{WT_1}$.

that with probability p, $\widehat{\mu(\boldsymbol{u})} \pm \gamma \cdot \widehat{\sigma(\boldsymbol{u})}$ contains all the bootstrap estimates. The heatmaps in the third column of figure 2.13 display the outer-product of the corresponding functional components (in the first two columns of the same row). We use the point-wise bootstrap standard deviations to construct (point-wise) 95% confidence intervals around each point estimate; the blank spots on the heatmap correspond to the points for which the confidence interval contained 0.

The results in figure 2.13 mean (for row 1) that childhood WT impacts the upper quantiles of adult SBP but not the lower quantiles, (for row 2) all quantiles of adult HT impact adult SBP "significantly", and (for row 3) lower quantiles of adult BMI have a "significant" impact on all quantiles of adult SBP (above the 20^{th} quantile), and the uppermost quantile of adult BMI (quantiles above 95%) seem to impact on the mid to upper quantiles of adult SBP.

2.3.8 Illustration using NHANES data

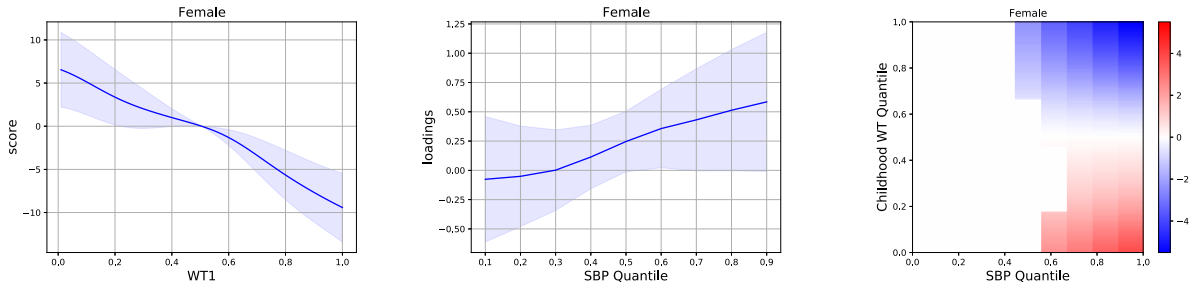
In this section, we apply AFQR to the NHANES dataset. We explore the relationship between SBP and body size measurements such as a persons height and weight. Note that unlike the Dogon study, this is a cross sectional dataset. We start by describing the models we fit, in terms of the independent variables and the model parameters.

The NHANES data contains information of people of all ages. For confidentiality reasons, all age above 80 are hard coded as 80. In this analysis, we only consider people above the age of 18, since it's not common to measure and analyze SBP for children. Further, we divide people of ages $18 - 80$ into three categories, young people below the age of 40, middle-aged people between ages 40 and 60, and old people above the age of 60.

We use the same model parameters as we did for the Dogon Longitudinal study 2.3.7.4. We present results for a smoothing bandwidth of 1000 for $f.$ and 5000 for $g.$.

2.3.8.1 Results

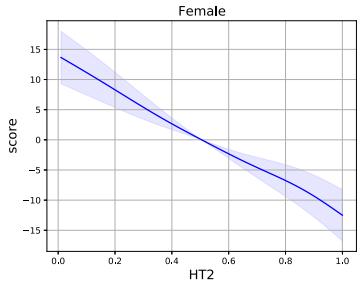
Results for young people are as expected, i.e. all quantiles of SBP increase with age, BMI and HT. The results for young males are presented in figure 2.14c. The effect of height is the weakest and BMI is undoubtedly the biggest driver of SBP in this sample. Most of the variation is seen across different quantiles of the independent variables (for instance as BMI deviates from the median BMI), and not much variation is present among the quantiles of SBP. Thus, we can look at the estimates as a function of the exposure z-scored values at fixed quantiles of the outcome via effect curves. We find that these are fairly parallel curves indicating that all outcome quantiles vary along the range of the exposure in a similar manner. The heterogeneity among the exposure z-scores can be seen via quantile curves, which are stacked one on top of the other indicating that all quantiles of the outcome tend to be higher or lower together, based on the value of the exposure.



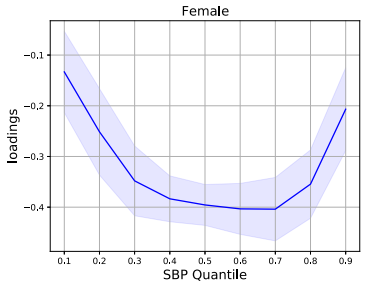
(a) $\boldsymbol{u_1}$

(b) $\boldsymbol{v_1}$

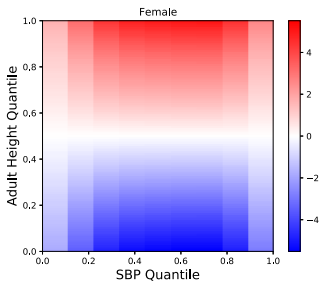
(c) $\boldsymbol{u_1 v_1'}$



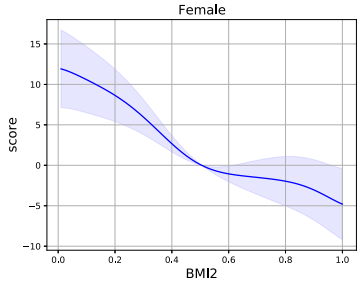
(d) $\boldsymbol{u_2}$



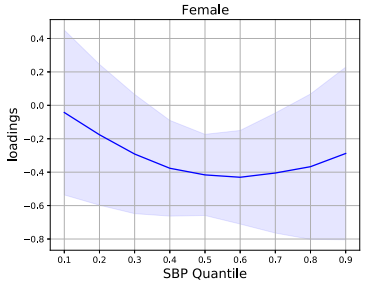
(e) $\boldsymbol{v_2}$



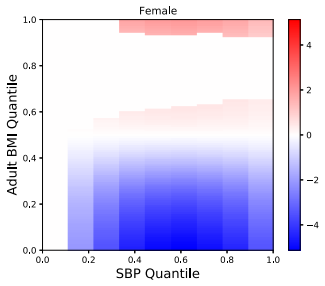
(f) $\boldsymbol{u_2 v_2'}$



(g) $\boldsymbol{u_3}$

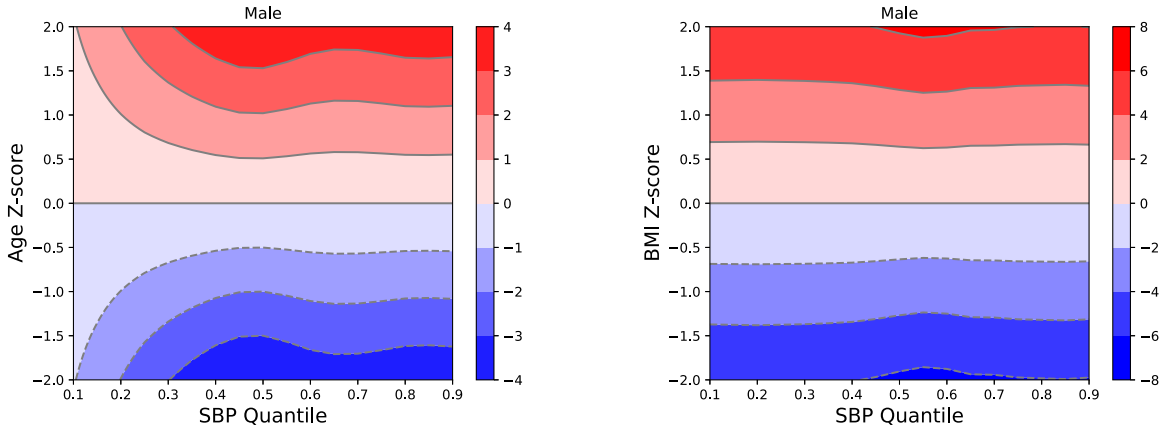


(h) $\boldsymbol{v_3}$



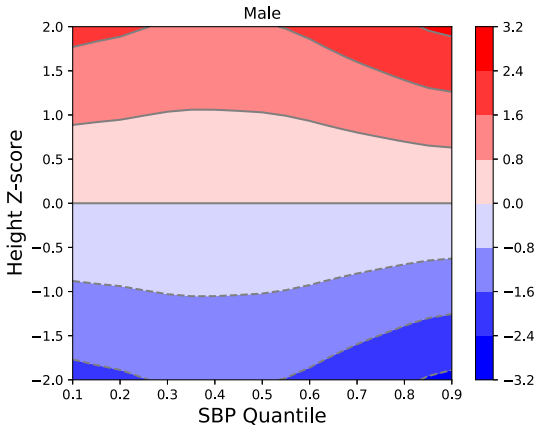
(i) $\boldsymbol{u_3 v_3'}$

Figure 2.13: Emulation-based bootstrap confidence bands for the AFQR functional components $(\boldsymbol{u}, \boldsymbol{v})$ corresponding to childhood WT, adult HT and adult BMI on adult SBP among females.



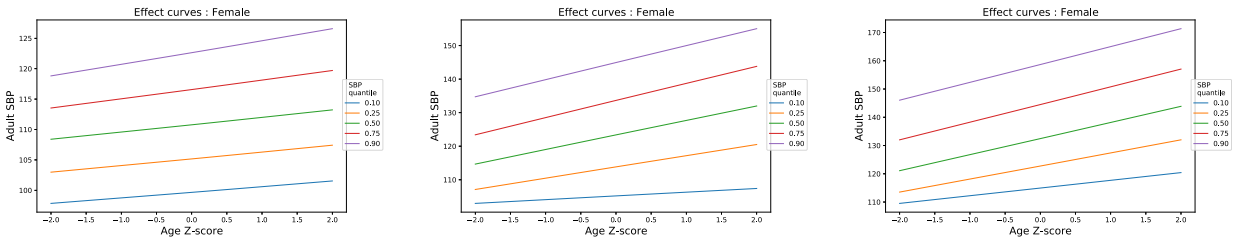
(a) SBP $\sim$ Age

(b) SBP $\sim$ BMI



(c) SBP $\sim$ HT

Figure 2.14: AFQR estimates for young males in the NHANES dataset. Each subplot corresponds to an additive component $\mu(p) + f_j(x)g_j(p)$ in the low rank representation ($X = \text{Age}, \text{BMI}, \text{HT}$ clockwise from 2.14a). Per component j, the other 2 variables are held at their medians. Per subfigure, f_j varies along the y-axis and g_j along the x-axis.



(a) Age < 40 (b) Age $\in [40, 60]$ (c) Age > 60

Figure 2.15: Effect curves measuring the impact of adult age on adult SBP via AFQR. The other variables are held at their median values. Results for females.

Variation with age

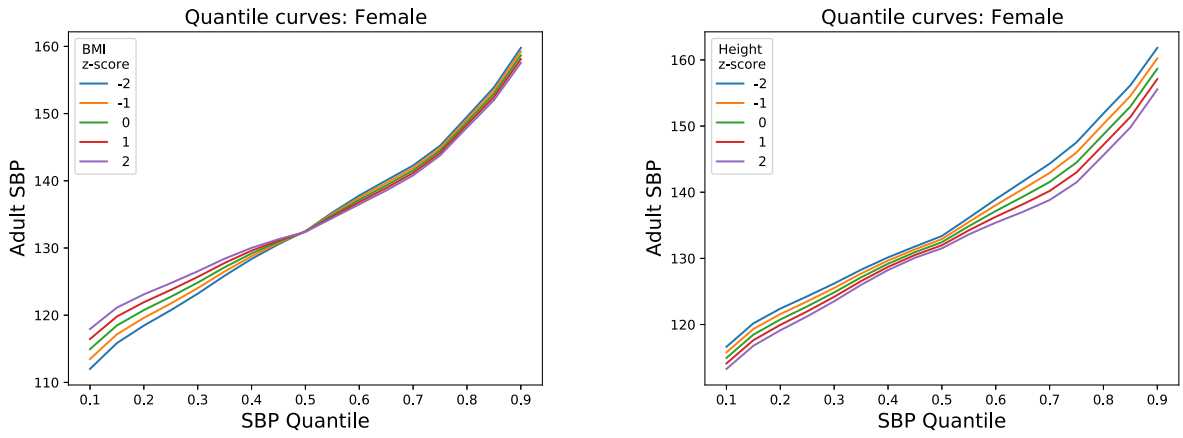
For all age groups, older people have higher blood pressure. In addition, the outer quantiles of SBP are impacted more by age than lower quantiles. This trend is missing among younger women; besides them, it means that the upper quantiles of SBP increase sharply with age, but the lower quantiles do not increase as steeply and that people with the lowest SBP among all ages tend to have similar SBPs. This can be seen through effect curves. presented in figure 2.15.

Variation with height

This gets more interesting. A stark difference between middle-aged, and younger people is that the cross-sectional effect of height on SBP gets flipped between the two groups. That is, among people of median BMI and median age above 40, being taller leads is favorable in terms of SBP and shorter people tend to have higher SBP and median-height or taller people tend to have lower SBP. This trend is seen in both groups of people above the age of 40 (ages 40-60 and above 60), but the opposite effect is seen among younger people (ref figure 2.14).

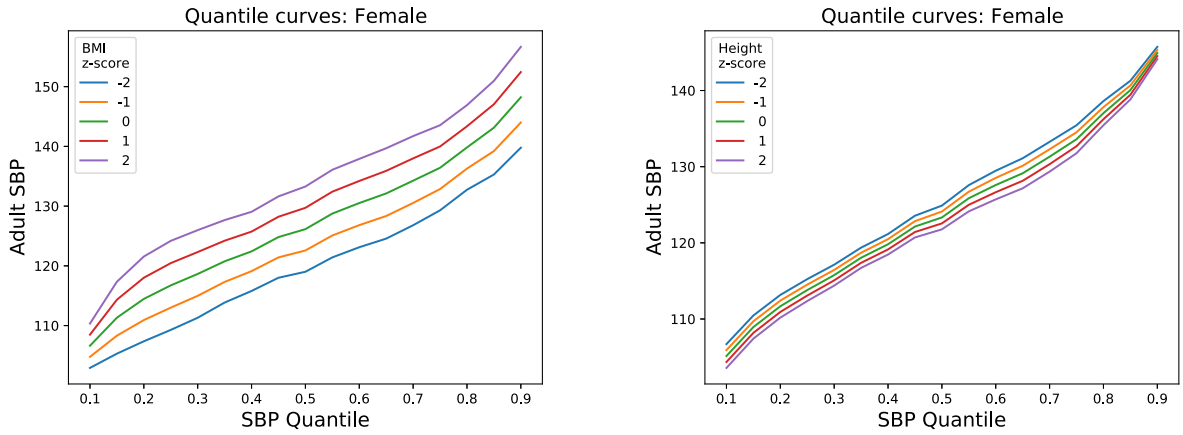
Another distinction between middle-aged and old people, and young people is that while the effects seen among the older people are still mostly heterogeneous along the exposure distribution, they are also heterogeneous along the outcome quantiles, unlike the corresponding effects for young people. We present some results for older women (ages 40 and above) in figure 2.16 to further explore this heteroscedasticity among the outcome quantiles. Figure 2.16 contains the quantiles curves as a function of SBP quantiles points, at fixed values of BMI and HT.

Figure 2.16 displays opposite trends in heteroscedasticity along the outcome quantiles between middle aged people and old age people. In particular, among old people (let's focus on women, as in figure 2.16b), the inner quantiles of adult SBP are more diverse as BMI varies whereas the upper quantiles are similar regardless of the persons age or body size. The opposite trend is seen in height (ref figure 2.16a), where lower-to-mid quantiles of SBP are similar across different heights, but the upper quantiles fan out as the height z-score varies. This seems to imply that the upper quantiles of SBP are more impacted by height, whereas the lower quantiles are more impacted by



(a) Age above 60 (BMI)

(b) Age above 60 (HT)



(c) Age 40-60 (BMI)

(d) Age 40-60 (HT)

Figure 2.16: Quantile curves for old women (row 1) and middle aged women (row 2), at varying adult BMIs (column 1) and varying adult HT (column 2).

BMI. Looking at the variation in SBP at the lower end for BMI and upper end for HT, we observe that lower quantiles of adult SBP are lower for smaller women (BMI zscore = -2 has the lowest SBP towards the left of figure 2.16b), whereas the upper quantiles of adult SBP are lowest for taller females (HT zscore = +2 has the lowest SBP towards the right of figure 2.16a). In other words, the variation in SBP quantiles at different HT and BMI, seems flipped in terms of which SBP quantiles (upper vs lower) are more similar or dispersed.

2.4 Dimension Reduction Regression

In this section we develop a second and independent framework for understanding the conditional quantiles of a response variable in terms of a set of explanatory variables. This framework is based on dimension reduction regression (DRR) and is completely distinct and complementary to the AFRQ approach with the same goals that was discussed above.

Dimension reduction regression has been extensively studied; a review of this topic can be found in [64, 18]. Dimension reduction regression is based on the premise that the relationship between explanatory variables $\boldsymbol{x}$ and a response y can be simplified by reducing $\boldsymbol{x}$ to a few *variates*, which are linear functions of $\boldsymbol{x}$. If the *sufficient dimension* of the regression is k, then there exist basis or "direction" vectors $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k$ such that $y \perp\!\!\!\perp \boldsymbol{x} \mid \boldsymbol{\theta}_1'\boldsymbol{x}, \ldots, \boldsymbol{\theta}_k'\boldsymbol{x}$. In one extreme case where $k = 1$, we have a *single index model* wherein y only depends on $\boldsymbol{x}$ through a single *linear predictor* $\boldsymbol{\theta}_1'\boldsymbol{x}$, and in this case we write $\boldsymbol{\theta} = \theta_1$ for simplicity. In the opposite extreme, the sufficient dimension of the regression is equal to the dimension of $\boldsymbol{x}$, which leads us to a fully unconstrained nonparametric analysis with resulting challenges due to the curse of dimensionality.

Linear regression and generalized linear models (GLMs) are well-known types of single index models that require explicit specification of a *link function* g such that $g(E[y|\boldsymbol{x}]) = \boldsymbol{\theta}'\boldsymbol{x}$. In a GLM or linear model, g is specified and the estimates of the coefficient vector $\boldsymbol{\theta}$ depend on this specification. In contrast to these parametric approaches to low-index regression, the methods that we employ here aim to estimate the directions $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k$ without pre-specifying or explicitly estimating the link function. While this strategy generally results in some loss of efficiency, it reduces the risk that a specification error in the choice of link function will lead to biased or inconsistent estimates of the basis vectors $\boldsymbol{\theta}_k$. We also note that absent a specific link function, we can only identify the *effective dimension reduction* (EDR) subspace $\mathrm{span}(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k)$, not the individual basis vectors.

We can use the QNN approach to estimate each conditional quantile in a probability grid $\boldsymbol{p} \in \mathbb{R}^m$, for all n observations in our sample of data. This yields a $n \times m$ matrix $\hat{\boldsymbol{Q}}$ containing the estimates $\hat{\boldsymbol{Q}}_p(y_i; \boldsymbol{x}_i)$. Since each row of $\hat{\boldsymbol{Q}}$ and of $\boldsymbol{X}^o$ corresponds to an observation, and both $\hat{\boldsymbol{Q}}$ and $\boldsymbol{X}^o$ have multiple columns, it is natural to consider some form of multivariate regression. This

provides the rationale behind the two approaches developed below. First, in the next section we consider the classical approach of canonical correlation analysis (CCA). Then we consider a more modern approach based on *sliced inverse regression* and the notion of "most predictable variates". This latter approach may better capture non-linear relationships compared to CCA. For ease of notation, in this section we always mean $\boldsymbol{X} = \boldsymbol{X}^o$.

2.4.1 Canonical Correlation Analysis

CCA [29] seeks to find vectors $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_r$ and $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_r$ such that (i) the vectors $\boldsymbol{X}\boldsymbol{\beta}_j$ are pairwise uncorrelated, (ii) the vectors $\hat{\boldsymbol{Q}}\boldsymbol{\eta}_j$ are pairwise uncorrelated, and (iii) subject to (i)-(ii), the correlation between each $\boldsymbol{X}\boldsymbol{\beta}_j$ and $\hat{\boldsymbol{Q}}\boldsymbol{\eta}_j$ is maximized. That is, for each $j = 1, \ldots, d$, CCA optimizes

$$\arg\max_{\boldsymbol{\beta}_j, \boldsymbol{\eta}_j} \text{Corr}(\boldsymbol{X}\boldsymbol{\beta}_j, \hat{\boldsymbol{Q}}\boldsymbol{\eta}_j), \tag{2.15}$$

subject to

$$\text{Corr}(\boldsymbol{X}\boldsymbol{\beta_i}, \boldsymbol{X}\boldsymbol{\beta_j}) = 0, \;\; \text{Corr}(\boldsymbol{X}\eta_i, \boldsymbol{X}\eta_j) = 0, \;\; \forall i < j.$$

To produce estimates from data, all instances of Corr above are replaced with the corresponding sample correlation coefficient $\widehat{\text{Corr}}$.

The solution to this constrained optimization problem is efficiently obtained using the singular value decomposition (SVD). Let $\boldsymbol{S_{XQ}}$ denote the cross-covariance between $\boldsymbol{X}$ and $\hat{\boldsymbol{Q}}$, and let $\boldsymbol{S_{XX}}$ and $\boldsymbol{S_{QQ}}$ denote the covariance matrices of $\boldsymbol{X}$ and $\hat{\boldsymbol{Q}}$ respectively. The Cholesky decomposition of the covariance matrices can be written $\boldsymbol{S_{XX}} = \boldsymbol{R'_X R_X}$ and $\boldsymbol{S_{QQ}} = \boldsymbol{R'_Q R_Q}$. The Cholesky decomposition of the decorrelated cross-covariance matrix is $\boldsymbol{R_X^{-T} S_{XQ} R_Q^{-1}} = \boldsymbol{USV'}$, where $\boldsymbol{U}$, $\boldsymbol{V}$ are orthogonal matrices and $\boldsymbol{S}$ is a diagonal matrix with non-increasing values along the main diagonal. The CCA solution is given by the loading vectors $\boldsymbol{\beta}_j = \boldsymbol{R_X^{-1} U}_j$ and $\boldsymbol{\eta}_j = \boldsymbol{R_Q^{-1} V}_j$, where $\boldsymbol{U}_j$ and $\boldsymbol{V}_j$ are the j^{th} columns of $\boldsymbol{U}$ and $\boldsymbol{V}$. The corresponding *canonical scores* given by $\boldsymbol{X\beta}$ and $\hat{\boldsymbol{Q}}\boldsymbol{\eta}$.

As described above, CCA can be intepreted as solving a sequence of constrained optimizations. The dominant CCA loadings are two vectors $\boldsymbol{\beta}_1 \in \mathbb{R}^J, \boldsymbol{\eta}_1 \in \mathbb{R}^m$ that maximize the correlation in equation 2.15. These are called the first set of *canonical coefficients* or *canonical loadings*. The second canonical coefficients are obtained by maximizing the same correlation subject to the constraint that $\boldsymbol{X}_c\boldsymbol{\beta}_1$ and $\boldsymbol{X_c}\boldsymbol{\beta_2}$ are uncorrelated and $\hat{\boldsymbol{Q}}_{\boldsymbol{c}}\boldsymbol{\eta_1}$ and $\hat{\boldsymbol{Q}}_c\boldsymbol{\eta}_2$ are uncorrelated. These properties hold up to a maximum of $\min(m, J)$ sets of loadings.

CCA is defined in terms of correlation coefficients which are location and scale-invariant. Therefore, without loss of generality the columns of $\boldsymbol{X}$ and $\boldsymbol{Q}$ can be taken to be centered to have mean zero, and the magnitude of the loading vectors $\boldsymbol{\beta}_j$ and $\boldsymbol{\eta}_j$ are scaled so that $\|\boldsymbol{\beta}_j\| = 1$ and $\|\boldsymbol{\eta}_j\| = 1$ for all j. When required for emphasis, we write $\hat{\boldsymbol{Q}}_c$ and $\boldsymbol{X}_c$ to emphasize that we are working with column-centered matrices. As a result of this centering, CCA can be seen as a way to understand the deviation of each conditional quantile function from an overall average quantile function. These "$\boldsymbol{Q}$-side" deviations are explained in terms of the centered "X-side" residuals.

2.4.1.1 Combining CCA with PCA

One issue that can arise when applying CCA in moderate or higher dimensional settings is that the variance for a CCA-optimal variate, say $\mathrm{var}(\hat{\boldsymbol{Q}}\boldsymbol{\eta}_j)$, may be very small compared to $\mathrm{var}(\hat{\boldsymbol{Q}}\tilde{\boldsymbol{\eta}})$ for some other unit vector $\tilde{\boldsymbol{\eta}}$. That is, CCA may find relationships for which the correlations between the $\boldsymbol{X}$ and $\hat{\boldsymbol{Q}}$ variates are strong, but where either variate is a very small part of the overall variation (for $\boldsymbol{X}$ or $\hat{\boldsymbol{Q}}$ respectively). This is a form of overfitting that may lead to spurious results. To address this, we can pre-process the data with Principal Components Analysis (PCA) as detailed in the next section.

The columns of $\hat{\boldsymbol{Q}}$ contain estimated conditional quantiles of the outcome, and as a result they may vary in a somewhat constrained manner. Each row of $\hat{\boldsymbol{Q}}$ contains a sequence of quantiles and hence must be non-decreasing, however we are working with a column-centered version of $\boldsymbol{Q}$ and the resulting residuals need not be non-decreasing. Nevertheless, it is possible that a lower-dimensional subspace of $\mathbb{R}^m$ contains most of the relevant variation in the conditional quantile functions. In this case, the irrelevant part of the space could contribute to spurious overfitting in the CCA as discussed above. To avoid such a circumstance, we conduct CCA on $\boldsymbol{X}$ relative to the scores for a subset of dominant principal components of $\hat{\boldsymbol{Q}}$, instead of on $\hat{\boldsymbol{Q}}$ itself.

PCA is a dimension reduction technique, which provides an alternate basis for the space spanned by the columns of $\hat{\boldsymbol{Q}}_c$, such that the basis vectors are pairwise orthogonal to each other and capture the directions of maximum variation within the data. Specifically, we have $\tilde{\boldsymbol{Q}} = \hat{\boldsymbol{Q}}_c\boldsymbol{P}$, where $\boldsymbol{P}$ is the $m \times k$ matrix of PCA loadings that maps the matrix $\hat{\boldsymbol{Q}}_c$ of column-centered conditional quantile estimates to the principal component scores $\tilde{\boldsymbol{Q}} \in \mathbb{R}^{n\times k}$. Here k is a chosen dimension for the reduced space of PC scores. The PCA loadings $\boldsymbol{P}$ can be obtained via the spectral decomposition of the sample covariance matrix

$$\hat{\boldsymbol{Q}}_c^\top \hat{\boldsymbol{Q}}_c = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}'.$$

If $\tilde{\boldsymbol{\eta}}_k$ is a CCA coefficient vector for $\tilde{\boldsymbol{Q}}$ in this analysis, then $\boldsymbol{\eta}_k \equiv \boldsymbol{P}\tilde{\boldsymbol{\eta}}_k/\|\boldsymbol{P}\tilde{\boldsymbol{\eta}}_k\|$ represents the same loading vector expressed with respect to the original $\hat{\boldsymbol{Q}}$. Further, since $\boldsymbol{\eta}_k$ maps $\hat{\boldsymbol{Q}}$ into the

span of its dominant k left singular vectors, the variance of $\hat{\boldsymbol{Q}}\boldsymbol{\eta}_k$ is no less than $\boldsymbol{\Lambda}_{kk}/n$, where $\boldsymbol{\Lambda}_{kk}$ is the k^{th} largest singular value of $\hat{\boldsymbol{Q}}_c^\top \hat{\boldsymbol{Q}}_c$. Thus the "PCA-CCA" pipeline can be seen as a way to obtain direction vectors $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ such that the corresponding variates $\boldsymbol{X}_c\boldsymbol{\theta}$ and $\boldsymbol{Q}_c\boldsymbol{\eta}$ are highly correlated, while still reflecting a substantial fraction of the total variance.

2.4.2 Dimension reduction and most-predictable variates

Classical methods for dimension reduction including PCA and CCA are most effective when the link function g is approximately linear. To better capture nonlinear structure, various nonparametric dimension reduction algorithms have been developed. While most of this literature has focused on the familiar setting of a scalar response that is predicted by one or more explanatory variables, a few approaches have been devised for the setting where both the response and explanatory variables are multivariate. Here we consider the "most-predictable variates" approach built on Sliced Inverse Regression (SIR), denoted here as MP-SIR [60]. SIR [59] was the first dimension reduction method to be proposed. It focuses on the "inverse regression function" $E[\boldsymbol{x}|y]$, and specifically on its marginal covariance $\text{var}E[\boldsymbol{x}|y]$. The goal is to find a matrix $\boldsymbol{B} \in \mathbb{R}^{p\times d}$ where $\boldsymbol{x} \in \mathbb{R}^p$, so that $\boldsymbol{B}'\boldsymbol{x}$ captures all of the information in $\boldsymbol{x}$ that is relevant for predicting y. Formally, we seek $\boldsymbol{B}$ so that $y \perp\!\!\!\perp \boldsymbol{x}|\boldsymbol{B}'\boldsymbol{x}$. Under certain conditions, the generalized eigenvalues of $\widehat{\text{cov}}E[\boldsymbol{x}|y]$ with respect to $\widehat{\text{cov}}(\boldsymbol{x})$ to estimate such a basis. In SIR, $\widehat{\text{cov}}[\boldsymbol{x}|y]$ is estimated using a simple slicing estimator, and $\widehat{\text{cov}}(\boldsymbol{x})$ is estimated using the standard moment estimator.

The MP-SIR approach generalizes SIR to accommodate multivariate responses. It is an alternating approach that iteratively applies SIR to y as predicted by $\boldsymbol{B}'_{\boldsymbol{x}}$, and then to $\boldsymbol{x}$ as predicted by $\boldsymbol{B}'_y y$. Here, $\boldsymbol{B}_x \in \mathbb{R}^{J\times r}$ and $\boldsymbol{B}_{\boldsymbol{y}} \in \mathbb{R}^{p_y\times r}$ are current values of the basis matrices. The slicing approach of SIR is extended to allow slicing on multiple columns. Under certain conditions, $\boldsymbol{x} \perp\!\!\!\perp y|(\boldsymbol{B}'_{\boldsymbol{x}}\boldsymbol{x}, \boldsymbol{B}'_y y)$, meaning that all of the dependence between $\boldsymbol{x}$ and y is captured through the linearly reduced variates $\boldsymbol{B}'_{\boldsymbol{x}}\boldsymbol{x}$, $\boldsymbol{B}'_y y$. Unlike in CCA, the relationships among the reduced $\boldsymbol{x}$ and reduced y variates may be substantially nonlinear.

2.4.3 Interpretation

2.4.3.1 Rotation

The loading vectors $\boldsymbol{\eta_k}$ for $\hat{\boldsymbol{Q}}$ correspond to factors in the space of conditional quantile functions that are predictable from $\boldsymbol{X}$. Note here that by "quantile function" we mean the deviation between a specific conditional quantile function and the average of all conditional quantile functions. To further aid interpretation, we note that if $\boldsymbol{\eta_k} \propto 1$, all quantiles change to the same extent as $\boldsymbol{X}$ varies, as in a location family. We have found that data of interest often approximately exhibit such

location behavior, although not always in the dominant component. Therefore, we have found it useful to rotate the first few loading vectors to identify an approximate location relationship, and then interpret this component in contrast to the remaining factors which capture non-location effects (e.g. effects on quantiles at different probability points to different extents).

Rotations are an important part of classical factor analysis [1, 75]. In this spirit, we develop a procedure to rotate the CCA or MP-SIR solutions so that the first factor loadings for $\hat{\boldsymbol{Q}}$ become approximately constant, reflecting a location relationship as discussed above. This involves constructing a transformation matrix $\boldsymbol{F}$ and replacing $\boldsymbol{\eta}$ with $\boldsymbol{\eta F}$, and replacing $\boldsymbol{\beta}$ with $\boldsymbol{\beta F}$. In doing so, we wish to preserve the property that the scores $\boldsymbol{X\beta_j}$ remain pairwise uncorrelated. It is not possible in general when rotating to also preserve pairwise uncorrelatedness of the scores $\hat{\boldsymbol{Q}}\boldsymbol{\eta_j}$, but this is arguably less important for interpretability.

$$\boldsymbol{X\beta} \longleftrightarrow \hat{\boldsymbol{Q}}\boldsymbol{\eta} \qquad \underrightarrow{rotation} \qquad \boldsymbol{X\beta}F \longleftrightarrow \hat{\boldsymbol{Q}}\boldsymbol{\eta F}, \quad \boldsymbol{F'F} = \boldsymbol{I}.$$

Let $\boldsymbol{\beta}$ denote the matrix of all X-loadings $\boldsymbol{\beta_j}$ and $\boldsymbol{\eta}$ denote the matrix of Q-loadings $\boldsymbol{\eta_j}$, i.e. $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \cdots, \boldsymbol{\beta}_r')' \in \mathbb{R}^{J\times r}$ and $\boldsymbol{\eta} = (\boldsymbol{\eta}_1', \cdots, \boldsymbol{\eta}_r')' \in \mathbb{R}^{m\times r}$. Here r is the number of canonical coefficients in the model, which is a tuning parameter.

Let us now discuss how to construct $\boldsymbol{F} = [f_{jk}] \in \mathbb{R}^{r\times r}$. The rotation algorithm begins by using least squares to regress a column of $1's$ onto the Q-loading vectors $\boldsymbol{\eta}_j$ which form the columns on matrix $\boldsymbol{\eta}$, i.e. $\mathbf{1}_m \sim \boldsymbol{\eta}_1 \implies \mathbf{1}_m = \sum_{j=1}^{r} f_{j1}\boldsymbol{\eta}_j$. Let $\hat{\boldsymbol{\eta}}_\mathbf{1}$ denote the fitted values, and let $\boldsymbol{f}_1 = (f_{11}, \cdots, f_{1r})$ denote the coefficient vector such that $\hat{\boldsymbol{\eta}}_\mathbf{1} = \boldsymbol{\eta f}_1$. The vector $\boldsymbol{f}_1$ becomes the first column of the transformation matrix $\boldsymbol{F}$. Subsequent columns of $\boldsymbol{F}$ are obtained using the Gram-Schmidt procedure to preserve pairwise uncorrelatedness among the $\boldsymbol{X}$-side scores.

2.4.3.2 Support points

Interpreting the results of CCA or MP-SIR is a somewhat complex task in that we must understand how (i) variation in the $\boldsymbol{X}$-scores relates to variation in the observed $\boldsymbol{X}$ variables, (ii) variation in the $\boldsymbol{X}$ scores relates to variation in the $\hat{\boldsymbol{Q}}$ scores, and (iii) variation in the $\hat{\boldsymbol{Q}}$ scores relates to variation in the observed $\hat{\boldsymbol{Q}}$ variables. In CCA this process is somewhat simplified by two facts. First, the relationships in (ii) are treated in CCA as linear. Also, in standard CCA there are no "cross-relationships" in (ii), i.e. the $\boldsymbol{X}$-scores for factor j are uncorrelated with the $\hat{\boldsymbol{Q}}$-scores for factor j' if $j \neq j'$. However after rotating as discussed above the second property will no longer hold. In MP-SIR neither of these properties holds.

To aid in the interpretation of the relationships uncovered by CCA or MP-SIR, we make use of "support points" as recently developed [67]. Support points are a set of points that are distributed over the important regions of a joint distribution. Using a distance-minimization approach, a set of

support points of a chosen size can be constructed that optimally cover the support of a distribution. For our purposes, we construct a limited set of support points in the joint space of $\boldsymbol{X}$-scores, consisting of say five points. We then identify a neighborhood of observed data points in the space of $\boldsymbol{X}$-scores that fall close to each support point, and average their corresponding quantiles in $\hat{\boldsymbol{Q}}$. This produces a nonparametric estimate of the expected quantile function corresponding to a given point in the $\boldsymbol{X}$-score space. By considering 5-10 support points, the relationship between explanatory variables $\boldsymbol{X}$ and quantiles $\hat{\boldsymbol{Q}}$ can be elucidated.

To provide sharper insight, we select a "focus variable", say adult BMI, take each support point $\boldsymbol{P}$ and then consider the points $\boldsymbol{P}_+$ and $\boldsymbol{P}_-$ that correspond to one standard deviation changes in BMI relative to $\boldsymbol{P}$. Specifically, suppose that $\boldsymbol{B}_{\boldsymbol{x}}$ is the MP-SIR or PCA-CCA basis matrix for the explanatory variables $\boldsymbol{X}$, and standardized BMI occupies column j^* of $\boldsymbol{X}$. Then $\boldsymbol{P}_+ = \boldsymbol{P} + \boldsymbol{B}'_{\boldsymbol{x}}\boldsymbol{e}_{j^*}$, where $\boldsymbol{e}_{j^*} \in \{0,1\}^J$ is the indicator vector of column j^*. We then identify the n^* nearest neighbors of $\boldsymbol{P}_+$ and $\boldsymbol{P}_-$ in the set of observed $\{\boldsymbol{B}'_{\boldsymbol{x}}\boldsymbol{x}_i\}$ values, and let I_+ and I_i denote the indices of these nearest neighbors. Finally, we average the rows of $\hat{\boldsymbol{Q}}$ corresponding to $\boldsymbol{I}_+$ and $\boldsymbol{I}_-$. This is a nonparameteric estimate of the difference of connditional quantile functions $\boldsymbol{Q}_y(p; x = \boldsymbol{P}_+) - \boldsymbol{Q}_y(p; x = \boldsymbol{P}_-)$.

2.4.4 Inference for dimension reduction analysis of conditional quantiles

A natural goal is to assess how the conditional quantiles of Y given $\boldsymbol{X}$ vary with $\boldsymbol{X}$. Since the estimated quantiles in $\hat{\boldsymbol{Q}}$ were explicitly constructed using $\boldsymbol{X}$, there is an inbuilt relationship between $\hat{\boldsymbol{Q}}$ and $\boldsymbol{X}$ which may lead to spurious relationships. The question of interest is what form of relationship exists in the population – that is, were we to have the true quantiles $\boldsymbol{Q_0}$, what would be the relationship between $\boldsymbol{X}$ and $\boldsymbol{Q_0}$? In this section we discuss a simple method for addressing this question when using CCA in the dimension reduction analysis. An extension of this approach could be used with MP-SIR, but this requires us to quantify the relationship between the $\boldsymbol{X}$-side and $\boldsymbol{Q}$-side variates in a way that captures non-linear as well as linear relationships. For simplicity, we consider only the linear CCA setting here.

As noted above, $\hat{\boldsymbol{Q}}$ is a matrix of estimated quantiles that are constructed partially using $\boldsymbol{X}$. Our goal is to assess how the underlying true quantiles $\boldsymbol{Q_0}$ vary with $\boldsymbol{X}$. To address this question in a way that overcomes the inbuilt dependence between $\hat{\boldsymbol{Q}}$ and $\boldsymbol{X}$, we use a randomization approach. We randomize the rows of $\boldsymbol{X}$ holding Y fixed, re-estimate the quantiles in $\hat{\boldsymbol{Q}}$, then finally conduct the CCA with rotation. We then consider the correlations between $\hat{\boldsymbol{Q}}\boldsymbol{\eta_j}$ and $\boldsymbol{X}\boldsymbol{\beta_j}$ in the observed and randomized data. To the extent that the correlations in observed data are greater than the correlations in randomized data, the apparent relationship between $\boldsymbol{X}$ and $\hat{\boldsymbol{Q}}$ is unlikely to be spurious.

Specifically, we randomly permute the rows of $\boldsymbol{X}$ n_{rep} times and ultimately obtain a set of n_{rep} correlations between $\hat{\boldsymbol{Q}}\boldsymbol{\eta_j}$ and $\boldsymbol{X}\boldsymbol{\beta_j}$. We use these to conduct a bootstrap one-sided hypothesis test with a null hypothesis that the correlation between $\hat{\boldsymbol{Q}}\boldsymbol{\eta_j}$ and $\boldsymbol{X}\boldsymbol{\beta_j}$ is zero, and the alternate hypothesis that it's greater than zero. We reject the null hypothesis if the correlation estimate obtained from the data is greater than the empirical 95^{th} quantile of correlations.

2.4.5 Illustration using NHANES data

In this section we analyze the NHANES data using the procedures described above, with systolic blood pressure (SBP) as the dependent variable and age, BMI, and height as explanatory variables. The NHANES data are cross-sectional and the covariates are all measured at the same time that SBP is measured. We conduct the analysis twice, once including all subjects of age 18 and over (2,757 females and 2,590 males), and the second time including subjects with age between 18 and 40 (1,064 females and 1,001 males), to better align with the demographics of the Dogon Longitudinal Study participants.

We first consider the evidence that there is any linear relationship at all between the explanatory variables $\boldsymbol{X}$ and the conditional SBP quantiles $\boldsymbol{Q_0}$. As discussed above, we approach this using randomization. Table 2.4 contains the estimated loading parameters for the $\boldsymbol{X}$-side variables and the results of this randomization analysis. For each sex, we fit CCA with $d = 1$, 2, and 3 components and report the results for each component, indexed by c in the table. For inference, we focus on the columns labeled R and R_p. The values labeled R are the canonical correlations between each $\boldsymbol{X}$-side and corresponding $\boldsymbol{Q}$-side component. As noted above, these are biased upward and to assess the extent of bias we randomized the SBP values in Y 100 times and repeated the entire procedure. The 95^{th} percentiles of the randomized canonical correlation coefficients is shown in the column labeled R_p. When R exceeds R_p, we interpret this as providing substantial evidence that the component is not spurious.

The results shown in table 2.4 show that in the younger cohort for both females and males, only one pair of variates are linearly related. However in the cohort with no age restriction, correlations for at least three pairs of correlated variates are found in females, while the males have evidence for only two pairs of correlated variates. This indicates that the relationship between explanatory variables and SBP quantiles is more complex if we mix people across a wide age range.

Figure 2.17 shows the estimated loading patterns for the NHANES data. The first row of the figure shows the loading patterns for the 18-40 age cohort, which are estimated to have only one CCA component. These loading patterns essentially reflect a location relationship, in which some people, based on their covariate values, have either higher or lower SBP quantiles across all probability points. The second row of the figure shows the loading patterns for the 18-80 age cohort,

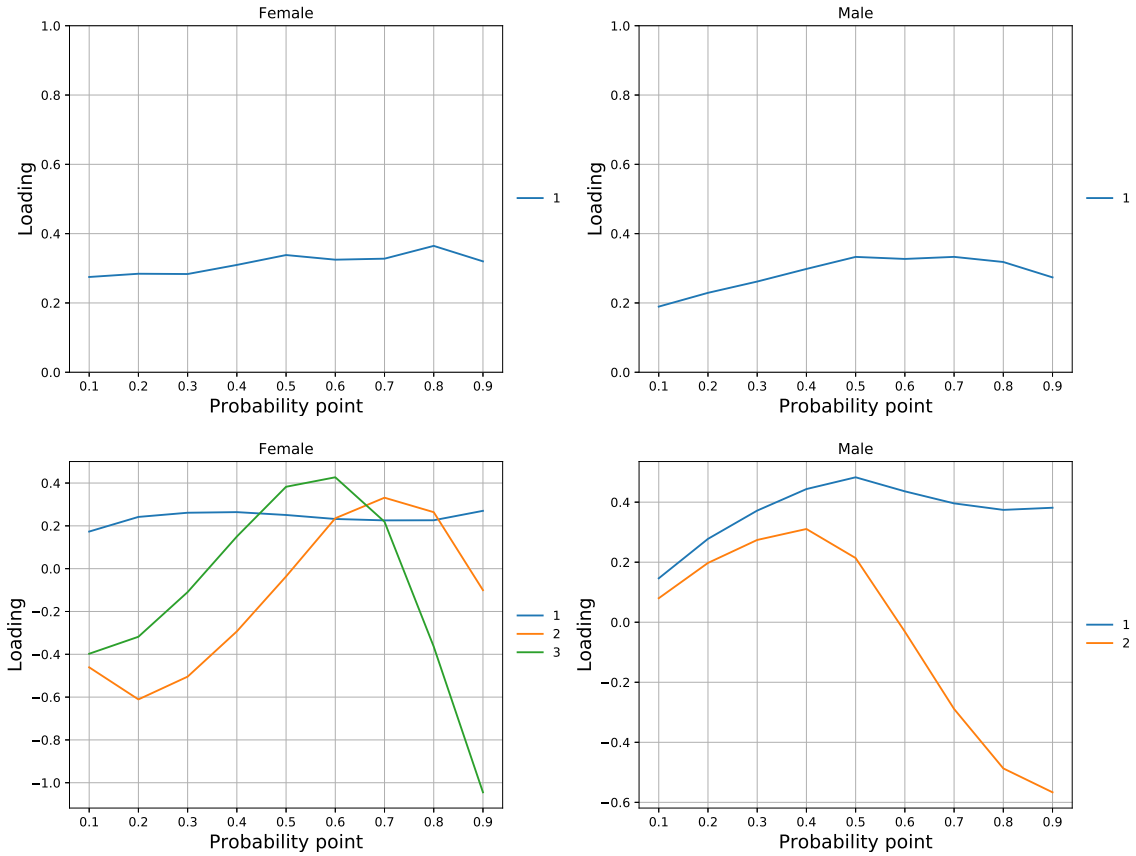


Figure 2.17: Loading patterns for the 18-40 year old (top row) and 18 and over (second row) subsets of NHANES.

			18-40		18-80	
Sex	d	c	R	R_p	R	R_p
Female	1	1	0.87	0.35	0.92	0.24
Female	2	1	0.87	0.37	0.92	0.24
Female	2	2	0.10	0.19	0.27	0.14
Female	3	1	0.89	0.45	0.93	0.30
Female	3	2	0.13	0.27	0.31	0.19
Female	3	3	0.08	0.13	0.09	0.07
Male	1	1	0.83	0.35	0.84	0.22
Male	2	1	0.83	0.39	0.84	0.27
Male	2	2	0.04	0.23	0.24	0.14
Male	3	1	0.89	0.41	0.85	0.31
Male	3	2	0.18	0.24	0.25	0.18
Male	3	3	0.03	0.11	0.02	0.09

Table 2.4: Randomization analysis of linear relationships between explanatory variables and quantiles in the NHANES data. Models were fit with d components, indexed by c; R denotes the observed correlation between corresponding variates and R_p denotes the 95^{th} percentile of correlations under randomization.

where the relationship between blood pressure and explanatory variables is more complex. The first loading function was rotated to be approximately constant, as discussed above. The remaining components (two for females and one for males) capture relationships in which some people have more dispersed conditional quantiles than others.

2.4.6 Illustration using Dogon Longitudinal Study data

As in the previously-described AFQR analyses using the Dogon study cohort, our overall goal here is to better understand the relationship between systolic blood pressure (SBP) and explanatory variables related to body size. For each observed SBP value, we used the QNN algorithm to estimate 9 conditional quantiles of SBP (spanning from probability points $p = 0.1$ to $p = 0.9$). These conditional quantiles comprise the $\hat{\boldsymbol{Q}}$ matrix as defined above. The analysis was applied in a series of configurations. Each analysis includes four explanatory ($\boldsymbol{X}$) variables – age, current height, current BMI, and one of six measures of childhood body size. The "current" height and BMI are the height and BMI as measured at the time of SBP measurement. The childhood body size was one of six variables – height, weight, BMI, HAZ, WAZ, and BAZ. The latter three variables are age-specific Z-scores based on World Health Organization growth charts.

The analysis was conducted separately for females ($n = 4,910$ observations) and males ($n = 5,178$ observations), in a "marginal" mode, meaning that moment matrices required by PCA,

CCA, and MP-SIR were constructed as simple aggregations over all observations. We note that while repeated measures per-subject induce correlation among the values observed on one subject, these marginal moments remain valid summaries when interpreted with respect to the marginal distribution of observations, pooled over subjects.

We first reduced the 9 estimated conditional SBP quantiles to 4 dimensions using PCA, and we then obtained the 3-dimensional MP-SIR fit. The results of the analysis are best appreciated visually. We start with childhood weight (PC scores for the trajectory from age 1-10) as the childhood body size variable, use age as the focus variable, and consider the results for female subjects. The three panels in the first column of figure 2.18 show scatterplots of the X-side variates. Biplotting is used to show how the reduced variates relate to the original four variables. Text labels A, B, C, D show the locations of four support points. Note that the coordinates of these support points are computed in $\mathbb{R}^3$ and are then projected into each of the displayed two-dimensional views. Corresponding to each support point, we construct two opposing points labeled $\oplus$ and $\ominus$ – these are constructed by displacing each support point by $+1$ or -1 standard deviation in the direction of the focus variable (here age).

The distribution of observed points in the first panel of column 1 of figure 2.18 reveals an inhomogeneity in the joint distribution of body sizes and age – younger subjects "fork" into two nearly disconnected subsets based largely on BMI. No such inhomogeneity is present for observations made on older subjects. This is likely explained by the fact that the timing of puberty onset varies by subject and generally takes place in the first few years after SBP measurements commence at around age 11 (i.e. in the younger observations within this dataset). We identified four support points, and each support point gives rise to one displacement in the positive age direction and one displacement in the negative age directions.

Panel 1 of figure 2.19 shows the estimated conditional quantile functions for the four pairs of displaced support points, plotted as deviations (residuals) of a specific conditional quantile function relative to the mean conditional quantile function. The extremes in panel 1 are the D- trajectory, which has high SBP quantiles, and the A+ trajectory, having low SBP quantiles. Referring to panels 1-3 we see that the D- trajectory corresponds to younger females with higher BMI and the A+ trajectory corresponds to older females with lower BMI.

To better understand how SBP quantiles vary when contrasting across ages, panel 2 of figure 2.19 shows the difference between the estimated conditional quantile function for each displaced P_+ support point relative to its displaced P_- support point. These differences are always negative, showing that when holding BMI, height, and childhood weight fixed, an older female subject always has lower SBP quantiles than a younger female subject. These differences are most pronounced for the central quantiles. A key feature of interest is the heterogeneity in these differences. The A support point has the least difference and the C support point has the greatest difference.

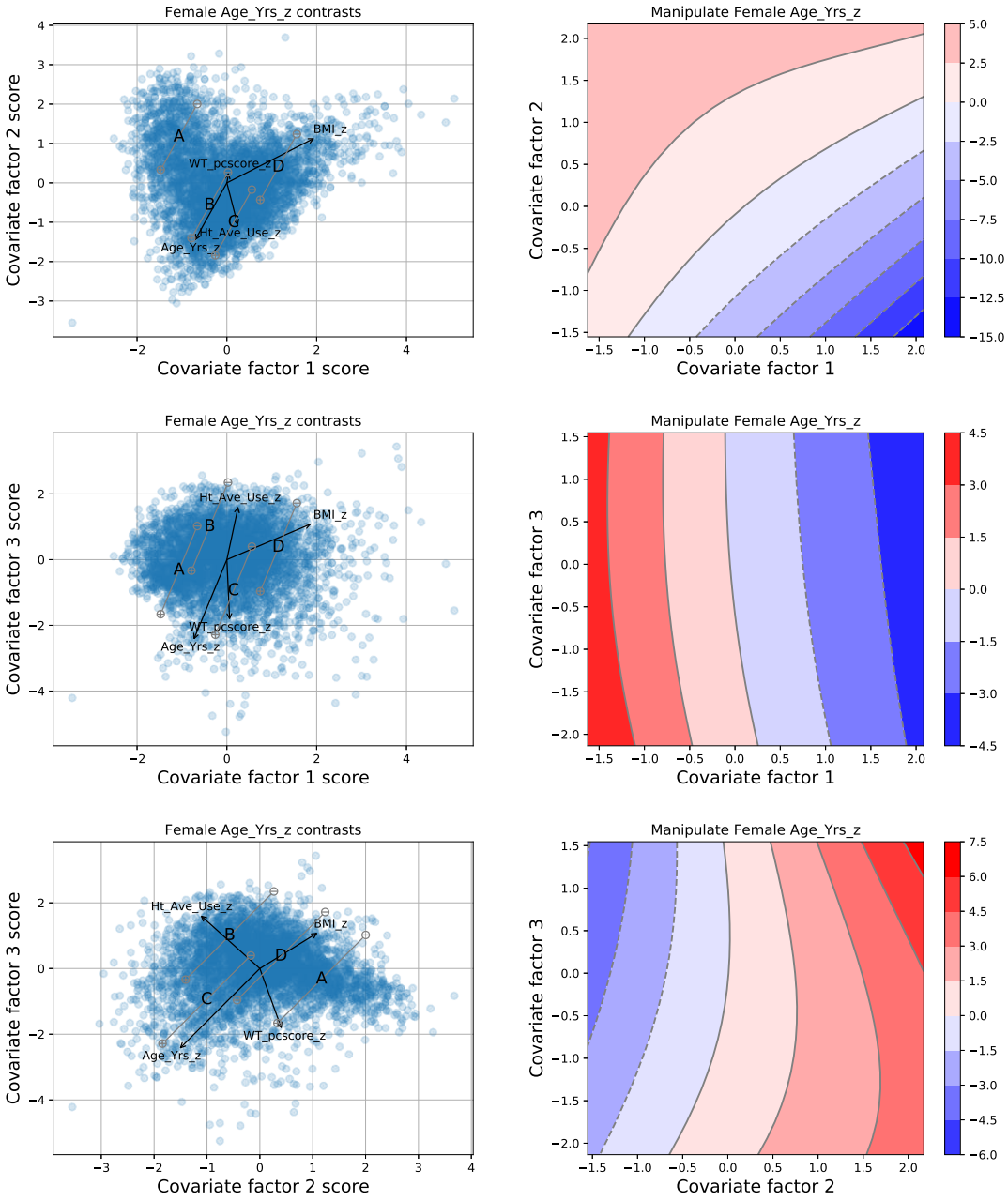


Figure 2.18: Left column: $\boldsymbol{X}$-side biplots with support points; Right column: heatmaps of inhomogeneity scores.

This "effect heterogeneity" suggests that age does not operate in an equivalent manner at all levels of the anthropometry variables.

To further probe the effect heterogeneity for age revealed in figure 2.19, we constructed 50 support points and contrasted their conditional quantile functions when translated by one unit in the increasing or decreasing age direction. The mean of these 50 differences is shown as panel 3 in figure 2.18. This mean is entirely negative and shows strongest contrasts at the central quantiles. We then used PCA to find a dominant linear summary of the variation in these differences. We call the corresponding PC scores "inhomogeneity scores" since they reflect the inhomogeneity in the "age effects" as the other covariates are varied. Panel 4 of figure 2.19 shows the loadings from which the inhomogeneity scores were derived. The pattern is entirely positive, meaning that the inhomogeneity generally takes a form in which differences at all quantiles are either greater, or lesser, based on the values of the covariates (excluding age). However the differences at the upper quantiles are as much as four times greater than the differences at the lower quantiles, suggesting greater inhomogeneity in the upper quantiles.

Next we sought to more systematically understand how the inhomogeneity scores vary with respect to the covariates. Using the inhomogeneity scores calculated at the 50 support points, we used local linear least squares regression to estimate the mean inhomogeneity score at each point in the covariate space. These estimated inhomogeneity scores are plotted in the right column of figure 2.18. Recall that a greater inhomogeneity score corresponds to a greater different in SBP quantiles when contrasting an older to a younger subject. Using the original 4 support points labeled A-D as landmarks, we see that there is a pattern in which positions close to support point A tend to have positive inhomogeneity scores. Since the mean quantile difference (panel 3 of figure 2.19 is negative, this means that positions close to support point A have the smallest differences in SBP quantiles as age changes. In contrast, positions close to support point C tend to have negative inomogeneity scores, and adding these negative scores to the mean results in a greater difference between conditional quantiles as age changes. Considering that the maximum magnitude of inhomogeneity score loadings (panel 4 of figure 2.19) exceeds 0.4, and the imhomogeneity scores themselves (right column of figure 2.18) often exceed 6 in magnitude, we see that the age contrasts often exceed 3 mm Hg, which is not a negligible effect size for SBP.

We applied the analysis framework described above using each of the six available childhood body size variables, separately for females and males, and yielding results for each of the four $\boldsymbol{X}$ covariates as the "focus variable". We next give a brief overview of these results, which are too voluminous to present in their entirety. Figure 2.20 shows the estimated differences in SBP quantile functions corresponding to ± 1 SD differences in BMI and in height, separately for females and males. The BMI and height variables are included in all analyses, along with one of the six childhood anthropometry variables, but here we only show the results when using weight as the

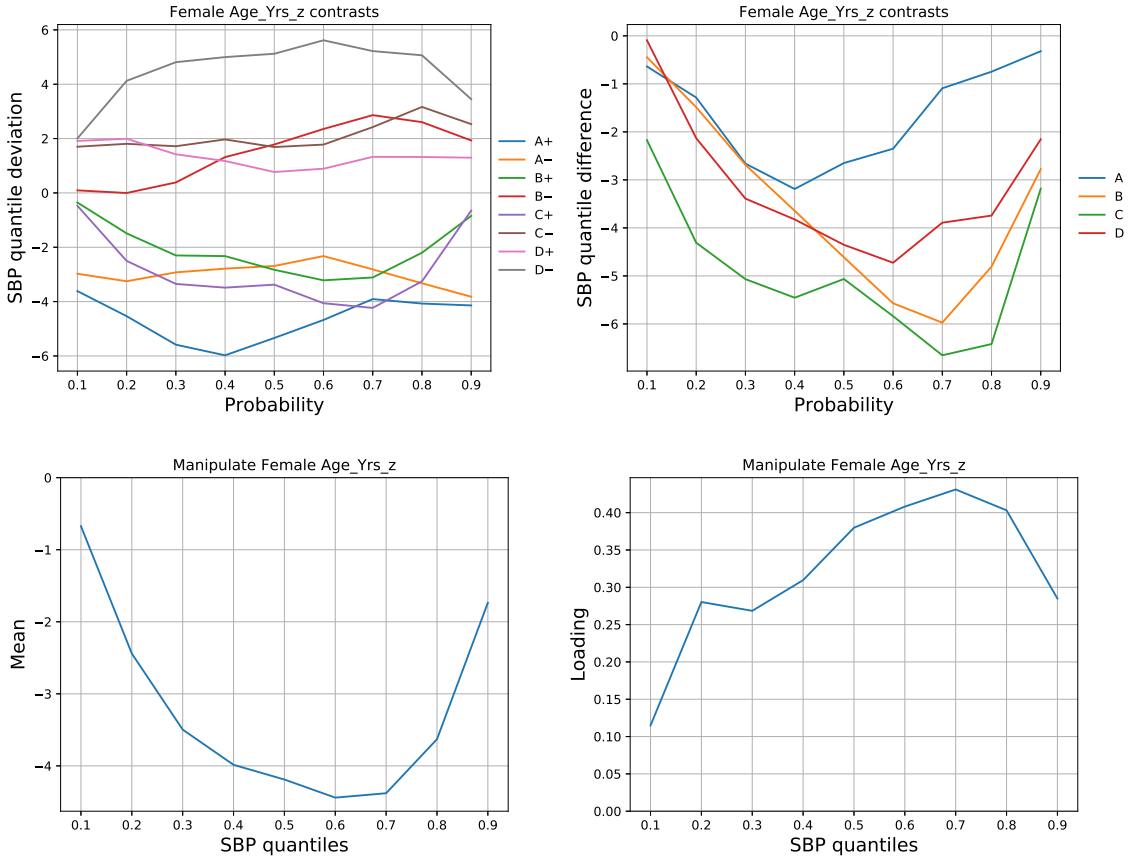


Figure 2.19: Panel 1: estimated conditional quantile functions corresponding to the support points; Panel 2: difference of opposing pairs of estimated conditional quantile functions; Panel 3: mean conditional quantile function over 50 support points; Panel 4: dominant PC loading over 50 support points.

childhood body size. Note that these are the "current" BMI and height measured at the same time that SBP was measured. Most contrasts are positive, showing that greater BMI or height associates with greater SBP, when holding the other variables fixed. However substantial heterogeneity is suggested in these plots. For example, for males the A support point shows little responsiveness to BMI changes and the D support point shows little responsiveness to height changes. Also, for females the BMI responses are much stronger at the central quantiles, while in males the BMI responses are approximately uniform over all quantiles.

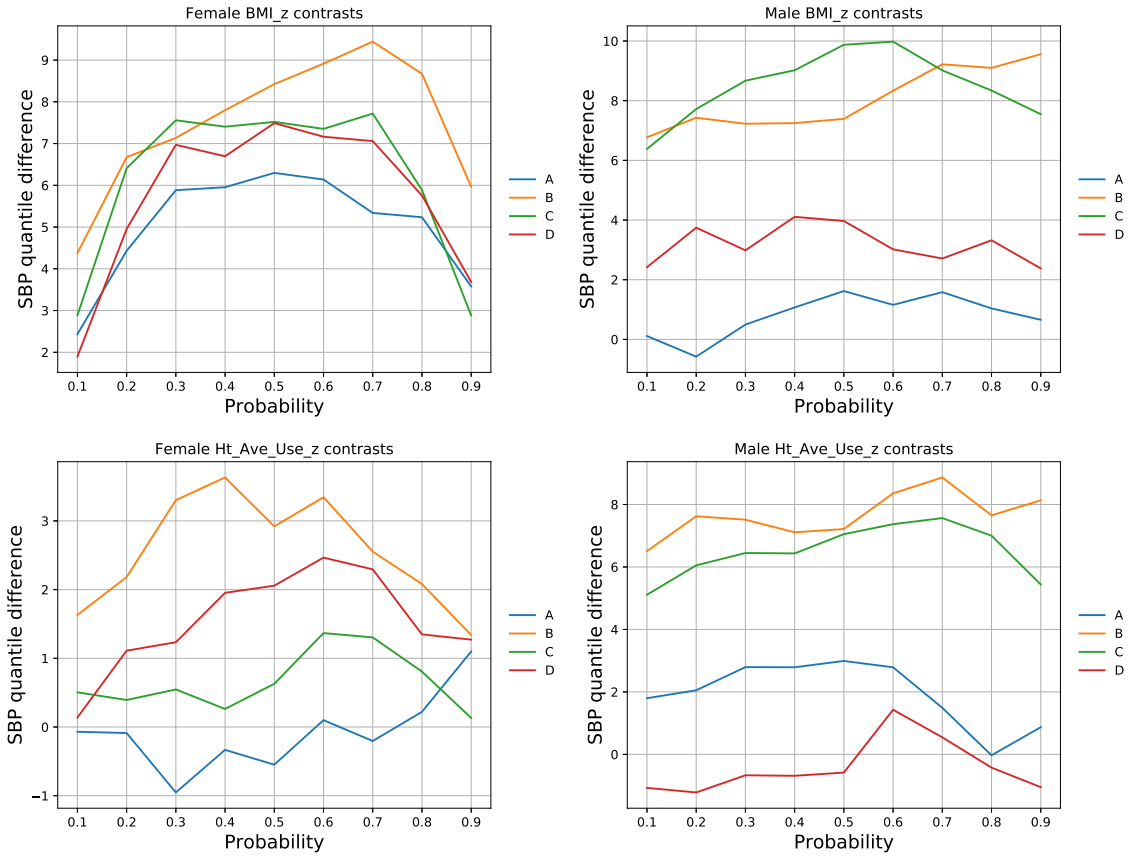


Figure 2.20: Estimated differences between quantile functions at four support points, in response to ± 1 SD changes in BMI or height. Results are shown separately for females and for males.

The greatest point of interest in these analyses is the relationship between childhood traits and adult SBP. The Dogon Longitudinal Study is uniquely suitable for assessing this relationship, since many other studies are cross sectional or have short follow-up times. Figure 2.21 shows selected results for these effects, using the methodology developed here. This figure shows results for the three variables related to adiposity (weight, BMI, BAZ). At most support points, the SBP quantiles are lower when childhood body size is greater. However there is substantial heterogeneity in these differences, and for some support points there is minimal relationship, or even an inverted

relationship. For example, support point A for males shows an inverted relationship in which larger childhood body size is associated with elevated adult blood pressure.

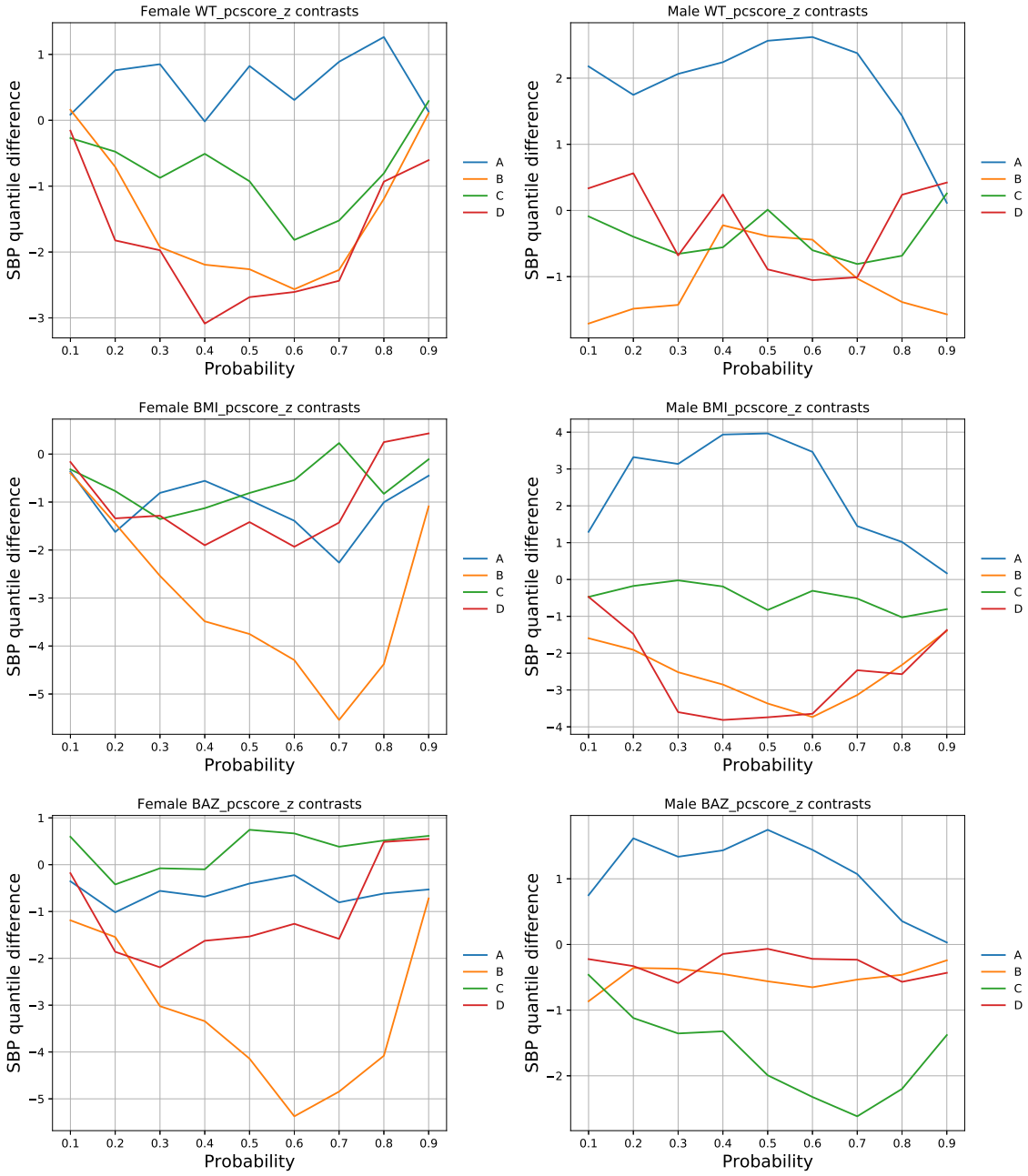


Figure 2.21: Estimated differences between quantile functions at four support points, in response to ± 1 SD changes in childhood body size variables. Results are shown separately for females and for males.

2.5 Conclusion

In this chapter we presented two methods to jointly model the conditional quantiles of an outcome variable across the distribution of multiple explanatory variables. Both the methods are designed to aid the interpretation of quantile-based analyses, unlike most existing quantile regression techniques that are aimed at prediction. Among methods aimed at interpretation, our methods are unique in that they can be applied as a post-processing step to any non-parametric (or parametric) quantile regression technique, and are thus flexible and able to accommodate non-parametric methods to produce interpretable results. The first step for both our methods includes estimating a collection of conditional quantiles that are low-bias, with potentially high variance. Our methods then find the joint association between this collection of quantiles and the covariates, suing dimension reduction techniques that ultimately lead to low-variance estimates. Thus, the final estimates are low-bias due to the non-parametric quantile regression in step 1, and low variance due to the dimension reduction in step 2.

In general, modelling all conditional quantiles (in terms of the quantiles of the covariates) completely non-parametrically is very challenging. Our techniques prioritize different kinds of complexities in order to achieve this goal flexibly (though not entirely non-parametrically). The key distinction between the two methods is that AFQR models the additive rank-1 impact of each covariate (with no restrictions on the shape of the functional rank 1 components), but the additive structure does not accommodate interactions. DRQR on the other hand can model interactions and accommodate the joint behaviour of any number of features, but it is a bit tedious to interpret. We develop tools for visualizing the results of both techniques, and demonstrate our methods on two common datasets (one cross sectional and one longitudinal). We also perform simulation studies for both techniques which display good performance of our methods.

As part of future work, the rank-1 model in AFQR can be extended to rank 2 or in general a higher rank (say r) per covariate. Such a model would take the following form

$$Q_{Y|\boldsymbol{X}}(p|\boldsymbol{x}) = \sum_{j=1}^{J} \sum_{k=1}^{r} f_{jr}(x_j) g_{jr}(p).$$

Additionally, our data-emulation based approach to inference for AFQR is cumbersome and there remain many challenges to be solved. Currently both our methods are two-stage procedures, but we could try to fit the AFQR model as a one-step process by minimizing the check function directly, eliminating the first step of estimating the conditional quantiles non-parametrically. Such

an approach would involve minimizing the following expression

$$\int \sum_{i=1}^{n} \tau_p\big(y_i - \sum_{j=1}^{J} h_j(x_{ij} \cdot g_j(p)\big) dp.$$

Lastly, the smoothing bandwidth optimization in AFQR can be formalized and potentially automated.

CHAPTER 3

Quantile-Based Mediation Analysis Using Factor-Structured Regression

3.1 Introduction

Mediation analysis is a framework for assessing structured hypotheses about associations and conditional associations (see [6], [65] for representative early papers). The essential idea is that an outcome can be influenced by explanatory variables at the same time that the explanatory variables are influencing each other. The relationships among the explanatory variables induce dependence among them, and through this dependence mediation analysis aims to assess hypotheses about causal relationships in the population. Specifically, suppose that two explanatory variables x_1 and x_2 are not only correlated, but it is thought that changes in x_1 induce changes in x_2, and not the other way around. In this case, we can define *direct* and *indirect* effects of x_1 on y, with the indirect effects operating "through" the mediator x_2 and the direct effects operating "independently" of the mediator.

Scientific hypotheses about associations in longitudinal data are often expressed in terms of mediation analysis. However with a few exceptions, mediation analysis has mostly focused on the conditional mean value of the outcome, which may fail to capture effects that occur in the tails of the outcome's conditional distribution [66, 82]. Moreover, many conventional forms of mediation analysis are linear in the exposure and mediators, meaning that the model treats a one-unit change in an exposure or mediator as having the same consequences regardless of where within the overall distribution of the variable this one-unit change occurs [6].

To address these challenges, in this chapter we introduce a novel framework for exploring mediation effects based on conditional quantiles of the relevant variables (the exposure, mediator, and outcome). This approach to mediation analysis reveals the extent of direct and indirect effects at every quantile of the outcome, through every quantile of the mediator(s). Perhaps the best way to interpret this is through the lens of heterogeneity. Changes to the exposure variable do not impact

the response variable deterministically, but rather alter the conditional distributions of the mediators and the response. The changes in the conditional distributions can be different at different quantiles – for example there may be strong direct effects at the median while the indirect effects are stronger at the outer quantiles. To the best of our knowledge, no existing analysis approaches allow a researcher to conduct mediation analysis across the quantiles of all exposure, mediator, and outcome variables, accommodating continuous exposures and multiple mediators.

The work presented here builds on the QNN and AFQR methods discussed in detail in chapter 2 (sections 2.2, 2.3). A key insight is that the AFQR framework yields relatively simple expressions for the direct, indirect, and total effects that can be seen as resulting from a "remapping" of the AFQR score functions to reflect the way that an "intervention" changes the distribution of the exposure, and the manner in which these changes propagate to change the marginal distributions of the mediators.

Throughout this chapter we motivate our methods using data from the Dogon Longitudinal study. Specifically, we consider the important question of how adult SBP is associated with growth and development from childhood into adulthood. This analysis contributes to the ongoing debate about the relationship between body size trajectories and adult SBP [17, 31, 46, 57]. On the one hand, there is a well-established cross-sectional relationship between SBP and current body size – larger (heavier and taller) people tend to have greater blood pressure [23, 5, 35, 93]. On the other hand, undernourished children experience stresses that may directly increase adult SBP, for example through impaired development of vascular tissues [68, 72, 86]. Yet a small child will tend to be a smaller adult (due to "tracking"). Thus, it is plausible that there are two antagonistic mechanisms connecting childhood undernutrition and adult SBP – larger children become larger adults and as a result tend to have greater SBP, but larger children are protected from developmental stresses which may result in lower adult SBP.

This chapter is organized as follows: we start by explaining the motivating application in detail, as this clarifies the goals of our mediation approach. We then introduce the quantile mediation model in section 3.2 and the direct, indirect and total effects. We establish notation in section 3.2.1, and navigate through the steps used to define a mediation model, in which (i) we establish the conditional quantile models of the mediators and the outcome using AFQR in section 3.2.2, then (ii) we introduce our hypothetical intervention in section, and redefine our conditional models in terms of the perturbed densities in section 3.2.4. Finally (iii) we define the mediation effects in section 3.2.5. We end this section by commenting on our marginal approach to model Longitudinal data. Next we access the performance of the entire pipeline consisting of AFQR models used in the quantile mediation analysis approach via a simulation study in section 3.3. We illustrate our method by applying it to the Dogon data in section 3.4. We end with concluding remarks and future directions in section 3.5.

3.1.1 Motivating Data Application

The impact of childhood growth on adult health has multiple facets. On the one hand, body size changes (growth) occur continuously throughout early life, often with strong "tracking", where tracking refers to temporal autocorrelation where small (big) children tend to become small (big) adults [79]. This is a longitudinal association, and at the same time there is a cross-sectional association between adult body size and adult health, i.e. at a fixed age, people who are taller and/or have greater body mass tend to have higher SBP. Thus, an individual's growth trajectory influences their health outcomes as an adult via two routes, longitudinally and cross-sectionally. The two associations at play here are summarized below

- *Tracking:* longitudinally, in an adverse environment under-nourished children grow up to be smaller adults, i.e. small children (short/ low BMI) become small adults (short/ low BMI).
- *Simultaneous Association:* cross sectionally, smaller people (short/ low BMI) have lower SBP.

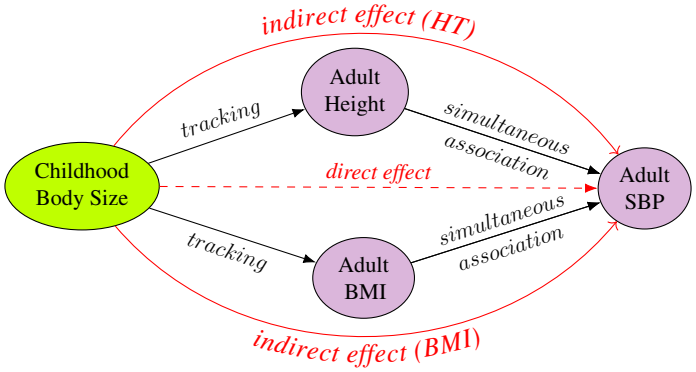


Figure 3.1: Direct and indirect effects in the motivating study.

Based on the tracking mechanism alone, we might expect undernourished children to have lower adult SBP. This is an indirect effect of childhood under-nutrition which results from the combined effects of tracking and simultaneous association. However, childhood under-nutrition may also have direct effects on adult SBP. There is strong evidence that childhood under-nutrition impacts mental and physical development, which in turn can have long-lasting health impacts including increasing risk for hypertension, lower fat oxidation, and reduced capacity for manual work, among other impairments [68, 86, 8, 42, 72]. Moreover, undernutrition is one of the leading causes of childhood mortality [10, 86, 40]. Focusing on SBP as an adult health outcome, the "direct effect" of childhood under-nutrition, which lacking direct measures of diet and nutrition can be captured by the proxy variable of childhood body size, is anticipated to be an increase in

SBP (hypertension). This direct effect may be partially or fully masked by the aforementioned "tracking effect" in which smaller children tend to have lower adult SBP. One of the main insights provided by a mediation analysis is to clarify whether these two antagonistic effects are in fact present, and to what extent one dominates the other.

Looking more closely at the tracking effect, it is important to note that tracking effects via different anthropometric attributes may have differing impacts on adult SBP. In this thesis we focus on indirect effects through height and Body Mass Index (BMI). It is plausible that BMI and height have independent contributions to adult SBP, and this is supported by the analysis using GEE presented in chapter 1. Moreover, by design BMI is largely independent of height. While out of scope for our analysis, we note that it may be argued that gains in SBP from increased BMI have a different overall cost than gains in SBP from increased height. Especially in under-resourced contexts, greater height is associated with benefits unrelated to SBP such as higher social status, greater educational attainment, and lower mortality risk during childbirth. These beneficial effects of greater height may possibly offset the adverse effects of greater SBP. On the other hand, except at the extreme low end of the range, effects of increased BMI are mostly or entirely adverse.

The Dogon Longitudinal Study, like most real-world longitudinal studies, presents several further complications that we wish to be able to address in a rigorous and principled way. First, we wish to view our population as developing continuously in time, even though measurements for each subject are taken at a finite number of time points. In the Dogon Longitudinal Study, individuals are assessed at roughly 1-year intervals, although there is substantial variation in the actual timing. We therefore adopt a kernel method to focus the analysis on two pre-specified time points, one in early childhood and one in adulthood. The subjects need not be observed at these exact time points since continuity and pooling are used to borrow information from nearby time points where observations were made. Second, we wish to quantify the uncertainty in our estimates in a way that accounts for the presence of unbalanced repeated measures per subject. We take a marginal approach, treating the observation rather than the person to be the unit of analysis for estimation purposes, and then use data emulation and bootstrapping to quantify uncertainty in a way that accounts for data correlations. Our mediation approach is used to decompose the impact of childhood body size at a pre-specified childhood age (exposure) on adult blood pressure at a pre-specified adult age (outcome), into its direct longitudinal effect and the indirect cross-sectional effect through adult body size measures, such as height and BMI measured at the same adult age (mediators).

3.1.2 Literature review of quantile approaches to mediation analysis

Early efforts to consider quantiles in mediation analysis involved modelling the conditional quantiles of only the outcome (our approach also considers quantiles of the exposure and mediators). This was proposed in [48], [22], and applied in [88], [44], [50]. The general strategy is to define parametric conditional models for the outcome given the mediator(s) and exposure $Y|M, X$ and the mediator(s) $M|X$, and then estimate the model parameters at a specific probability point by minimizing the check-function $\rho_\tau(x) = (\tau - \mathcal{I}_{x<0})x$. The only departure from least squares analysis is that the check function is minimized instead of minimizing the L2 loss. Alternatively, some researchers have used propensity score weighting rather than regression adjustment [87]. The advantages and disadvantages of these two approaches in a quantile estimation setting parallel what occurs in the setting of mean estimation.

Quantiles have also been considered in the mediator instead of, or alongside those of the outcome. In [34] they model the expected value of the outcome, and decompose the average indirect effect based on the quantiles of the mediator. They refer to the indirect effect of the exposure on the outcome at a specific quantile of the mediator as "u-specific" indirect effects. They show that the average indirect effect can be conditioned on the quantiles of the mediator and can be written as the average of the u-specific indirect effects. They also define u-specific total and direct effects, which are the conditional expected total and direct effect respectively, conditioned on the quantile of the mediator. The average total effect and average direct effect are similarly expressed as averages consisting of the u-specific effects.

Quantiles of both the mediator and the outcome are considered in [9]. They specify a linear model with random intercepts for the mediator, and a linear random intercept model with an exposure-mediator interaction for the outcome. Their method is also applicable to cross-sectional data, by excluding the random intercept from the model. Since the mean structures of their models are linear, the direct and indirect effects have closed forms in the absence of the interaction. They define direct effects as follows: for fixed quantiles of the outcome and mediator at p_Y, p_M, the direct effect is defined as the contrast between the p_Y^{th} quantile of the potential outcomes upon varying exposure $X = x, x^\star$ while holding the mediator at it is p_M^{th} quantile under $x^\star$. The indirect effect is defined as the contrast in the p_Y^{th} quantile of the potential outcomes when the exposure is fixed, say $X = x$, but the mediator changes from its p_M^{th} quantile conditioned on $X = x^\star$ to the p_M^{th} quantile conditioned on $X = x$.

For a binary exposure, [21] propose a way to find the direct effect if the distribution of the mediator (under the two categories of the exposure) can be made identical, referred to as residual disparity. They do this by proposing hypothetical interventions applied to the mediator. They use a Bayesian Linear Dependent Dirichlet Process model to estimate the density of the mediator, and explain that only the density in one of the exposure categories is needed - the category who's

mediation distribution is to be matched (the more advantageous category). They show that residual disparity is different from the natural expected total effect. While this is very different from our analysis, it is worth noting since it is one of the few analyses that consider the entire distribution of the mediator instead of focusing on its average value.

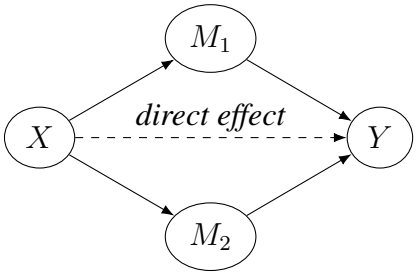
3.2 Mediation analysis using additive conditional quantiles

In this section we introduce our approach to mediation analysis using the additive functional quantile models (AFQR) developed in chapter 2 (section 2.3. The key insight is that the low-rank and additive structure of these models gives rise to very concise notions of direct and indirect effect that can be interpreted as "remapping" the AFQR factor scores in response to a hypothetical intervention.

3.2.1 Notation

The data consists of the outcome Y and a set of covariates measured for n subjects. Let Y denote the outcome, X denote the exposure, M_k denote the k^{th} mediator, and $\boldsymbol{Z}$ denote the set of other covariates. The data consists of n observations and is of the form $\{(Y_i, X_i, \{M_k\}_i), \boldsymbol{Z}_i\}_{i=1}^n$. For simplicity, we do not write $\boldsymbol{Z}$ anywhere but it is understood that all regression relationships adjust for $\boldsymbol{Z}$ where appropriate. We focus on a mediation model with two mediators; a graph summarizing the relationships among key variables is presented in figure 3.2.

Let $\tilde{M}_1$, $\tilde{M}_2$, and $\tilde{X}$ denote quantile-transformed versions of M_1, M_2, and X, so that $P(\tilde{M}_1 \leq t) = P(\tilde{M}_2 \leq t) = P(\tilde{X} \leq t) = t$. Also denote $\boldsymbol{W} = \{M_1, M_2, X\}$ and $\tilde{\boldsymbol{W}} = \{\tilde{M}_1, \tilde{M}_2, \tilde{X}\}$.



In our application
X : childhood body size
M_1 : adult HT
M_2 : adult BMI
Y : adult SBP

Figure 3.2: Directed acyclic graph representing the relationships between the variables in the mediation model. The nodes represent variables, and the arrows represent causal effects.

3.2.2 Conditional quantile models with additive low-rank structure

In this section we describe the models that we use to represent the conditional quantiles of the mediators given the exposure, $\tilde{M}_j|\tilde{X}$, and the conditional quantiles of the outcome given the mediators and exposure, $Y|(\tilde{X}, \tilde{M}_1, \tilde{M}_2) = Y|\tilde{\boldsymbol{W}}$. These in turn are used to carry out a quantile-based mediation analysis as discussed in subsequent sections.

As noted earlier, most methods for quantile regression estimate each conditional quantile function corresponding to a probability point $p \in (0, 1)$ separately. In this type of approach, there is no explicit consideration of the relationships between quantiles at different probability points, and in some methods there can even be crossings of the estimated quantile functions that cannot reflect the behavior of the population quantile functions. Our two-stage approach imposes a global structure on the collection of all quantile functions indexed by $p \in (0, 1)$. This approach partially pools information for different probability points, providing more stable estimation especially for the outer quantiles.

Our model for the conditional quantiles has an additive/multiplicative structure that focuses on two-way interactions between a function of one covariate and a function of the outcome variable's probability points. Specifically, we represent the conditional quantiles of the outcome in the form given below

$$\begin{aligned} &Q_{Y|(\tilde{M}_1,\tilde{M}_2,\tilde{X})}(p \,|\, \tilde{M}_1 = \tilde{m}_1, \tilde{M}_2 = \tilde{m}_2, \tilde{X} = \tilde{x}) = \\ &\quad \mu(p) + f_1(\tilde{m}_1)g_1(p) + f_2(\tilde{m}_2)g_2(p) + f_3(\tilde{x})g_3(p), \end{aligned} \tag{3.1}$$

where f_j and g_j are smooth functions and $p \in (0, 1)$ are probability points. To ensure that the model is identified, we impose the constraint $f_j(0.5) = 0,\ j = 1, 2, 3$. Here $\tilde{M}_1$, $\tilde{M}_2$, and $\tilde{X}$ marginally follow uniform distributions on the unit interval $(0, 1)$, as they have been quantile transformed. Thus the constraint is that the "score" function f at the unconditional median of its argument is zero. In this model, the (functional) parameters of interest are $\{f_j\}_{j=1}^3$ and $\{g_j\}_{j=1}^3$.

In practice we will need to work on a finite grid of values, so we produce a tensor of order 4 containing gridded values of $Q_{Y|(\tilde{M}_1,\tilde{M}_2,\tilde{X})}$ covering the support of $\tilde{\mathbf{W}}$, which is $(0, 1)^3$. Each component of $\tilde{\mathbf{W}}$ (i.e. $\tilde{X}$, $\tilde{M}_1$, $\tilde{M}_2$) is represented by a grid of probability points $1/(m+1), 2/(m+1), \ldots, m/(m+1)$ for an integer m, taken to be odd so that $p = 0.5$ is in the grid. The values of $Q_{Y|(\tilde{M}_1,\tilde{M}_2,\tilde{X})}$ are represented through points in this $m \times m \times m \times m$ tensor, with the first three axes of the tensor corresponding to probability points of $\tilde{X}$, $\tilde{M}_1$, and $\tilde{M}_2$, and the final axis corresponding to probability points of Y.

The conditional quantiles of the mediators given the exposure $\tilde{X}$ are also represented in a functional rank-1 form

$$Q_{\tilde{M}_1|\tilde{X}}(p \mid \tilde{X} = \tilde{x}) = \mu_1(p) + f_4(\tilde{x})g_4(p)$$
$$Q_{\tilde{M}_2|\tilde{X}}(p \mid \tilde{X} = \tilde{x}) = \mu_2(p) + f_5(\tilde{x})g_5(p), \tag{3.2}$$

and an order 2 $m \times m$ tensor (a matrix) can be used to represent each of these functions on their natural domains on a finite grid of points.

3.2.3 Hypothetical intervention

Mediation analysis can be viewed as a way to understand how an observed, untreated population would respond to a hypothetical intervention. We consider a hypothetical deterministic intervention that changes the distribution of the exposure. Let

$$h_{0,\theta}(x) \equiv \Phi(\Phi^{-1}(x) + \theta), \tag{3.3}$$

where θ is a parameter that determines the size of the intervention. Recall that our quantile transformed exposure $\tilde{X}$ follows a uniform distribution on the interval $[0, 1]$. The random variable

$$\tilde{X}' \equiv h_{0,\theta}(\tilde{X}) \tag{3.4}$$

can be seen as the result of mapping uniform random values to normal scores, shifting the normal scores by θ, and then converting back to the original scale. Thus, θ can be interpreted as an additive effect size on a standard normal scale. Further, it is easy to see that the CDF of $\tilde{X}'$ is $h^{-1}_{0,\theta}(\cdot)$ and the quantile function of $\tilde{X}'$ is $h_{0,\theta}(\cdot)$.

The parameter θ in equation 3.3 controls the mapping between the quantiles of $\tilde{X}$ and the quantiles of $\tilde{X}'$. If $\theta > 0$, $\tilde{X}'$ is stochastically greater than $\tilde{X}$, and if $\theta < 0$, $\tilde{X}$ is stochastically greater than $\tilde{X}'$. If X and $\tilde{X}$ are anthropometric variables representing childhood body size in the observed (unperturbed) population, then transforming $\tilde{X}$ using $\theta > 0$ produces a new population with greater body sizes at every quantile (analogous to a more "advantaged" population). This could result from an intervention that provides nutritional support to the population. Conversely, transforming $\tilde{X}$ using $\theta < 0$ produces a new population with smaller body sizes at every quantile (analogous to a less "advantaged" population). To understand the mediated relationship of the exposure (childhood nutrition) on the outcome (adult SBP) through the mediators (adult body size measures), we contrast two perturbed populations at values of θ and $-\theta$, that are symmetrically arranged with respect to the observed, unperturbed population.

3.2.4 Quantile remapping induced by an intervention

The generative models given above for $\tilde{M}_j|\tilde{X}$ and $Y|\tilde{X}, \tilde{M}_1, \tilde{M}_2$ (equations 3.1 and 3.2) are assumed to hold in the observed population as well as following an intervention – the intervention only changes the values (and hence the distribution) of $\tilde{X}$. We also note that in our observed population, the exposure $\tilde{X}$ was transformed (from X) to become uniformly distributed on the unit interval. This is simply a data transformation that is best viewed as a pre-processing step to our mediation framework (to understand the essence of our approach, it may be best to focus on $\tilde{X}$ as if it were the observed value). In contrast, when we transform $\tilde{X}$ to $\tilde{X}'$, this action changes the actual value of a variable, which in turn changes the corresponding values of the mediators and the outcome. Please note this subtle distinction between a transformation and a change in the variable – the intervention is represented as a function which looks like a transformation (in equation 3.3), but it is in fact not a transformation of the exposure or a change of units, it leads to a change in the value of the exposure expressed in the same units as the unperturbed exposure.

Let $\tilde{X}' \equiv h(\tilde{X})$ denote the value of a random variable after applying a given transforming function $h : \mathbb{R} \to \mathbb{R}$. It is important to note that we interpret $\tilde{X}$ and $\tilde{X}'$ as operating on the same scale, i.e. the intervention does not change the scale or units of the measurements, but rather redistributes the population along the same scale. Since $\tilde{X}$ is uniform on $(0, 1)$, the quantile function of $\tilde{X}'$ is h. Thus, for a given probability point $q \in (0, 1)$, the value of $h(q)$ conveys (on the common scale of $\tilde{X}$ and $\tilde{X}'$) the value of the q^{th} quantile of the exposure following the intervention. If we have a model $f_p(x) = Q_{Y|\tilde{X}}(p|\tilde{X} = \tilde{x})$ for the quantiles of $Y|\tilde{X} = \tilde{x}$, then $f_p(h(q))$ expresses the quantiles of Y (indexed by p) in terms of the quantiles of $\tilde{X}'$ (indexed by q). This fact is used repeatedly below.

Based on this logic, and denoting the hypothetical intervention by $h_0 = h_{0,\theta}$ for ease of notation, the conditional quantiles of $\tilde{M}_1$ and $\tilde{M}_2'$ given $\tilde{X}'$ can be expressed as follows:

$$
\begin{aligned}
Q_{\tilde{M}_1|\tilde{X}'}(p \,|\, \tilde{X}' = Q_{\tilde{X}'}(q)) &= \mu_1(p) + f_4(h_0(q))g_4(p), \\
Q_{\tilde{M}_2|\tilde{X}'}(p \,|\, \tilde{X}' = Q_{\tilde{X}'}(q)) &= \mu_2(p) + f_5(h_0(q))g_5(p).
\end{aligned} \tag{3.5}
$$

Note that in the expression above, q represents a probability point for $\tilde{X}'$ and p represents a probability point for $\tilde{M}_1$ or $\tilde{M}_2$.

Our model for the observed population expresses the joint distribution of $(\tilde{X}, \tilde{M}_1, \tilde{M}_2)$ in terms of the marginal distribution of $\tilde{X}$ and the conditional distribution $\tilde{M}_j|\tilde{X}, j = 1, 2$ referring to 3.5??? . As noted above, in our framework the intervention does not impact the conditional distribution $\tilde{M}_j|\tilde{X}$, but does change the marginal distribution of $\tilde{X}$, and therefore also changes the marginal distribution of each $\tilde{M}_j$. Let $\tilde{M}_1'$ and $\tilde{M}_2'$ denote random values of $\tilde{M}_j$ that are generated in this way

(i.e. following transformation of $\tilde{X}$). Again we emphasize that the scales of $\tilde{M}_j$ and $\tilde{M}'_j$ have the same meaning – the intervention leads to a redistribution of the population along these common scales. Let h_1 and h_2 denote the marginal quantile functions of $\tilde{M}'_1$ and $\tilde{M}'_2$ (constructing h_1 and h_2 is a minor technical challenge that we discuss below in section 3.2.6). The conditional quantile function of Y with respect to quantiles of the perturbed $\tilde{X}'$, $\tilde{M}'_1$, and $\tilde{M}'_2$ is

$$Q_{Y|(\tilde{M}'_1,\tilde{M}'_2,\tilde{X}')}(p \mid \tilde{M}'_1 = Q_{\tilde{M}'_1}(r_1), \tilde{M}'_2 = Q_{\tilde{M}'_2}(r_2), \tilde{X}' = Q_{\tilde{X}'}(q)) = \\ \mu(p) + f_1(h_1(r_1))g_1(p) + f_2(h_2(r_2))g_2(p) + f_3(h_0(q))g_3(p). \tag{3.6}$$

In the expression above, q is a probability point for the post-intervention exposure ($\tilde{X}'$), r_1 and r_2 are marginal probability points for the post-intervention mediators ($\tilde{M}_1$ and $\tilde{M}_2$), and p is a conditional probability point for the outcome. It is important to index the quantiles of the mediators $\tilde{M}'_j$ by marginal rather than conditional quantiles so that they have a common scale across all values of the exposure.

3.2.5 Mediation effects

Mediation effects are of three types: total, direct, and indirect. The direct effect assesses how changes to the exposure impact the outcome, while blocking any impact of the exposure on the mediators (i.e. in assessing the direct effect the mediators are held fixed even though in reality they may change in response to changes in the exposure). The indirect effect does the opposite; it assesses how changes to the exposure impact the mediators, which thereby impact the outcome, while blocking any direct effect of the changes to the exposure on the outcome. The total effect is essentially the sum of the direct and indirect effects. Our proposed framework for mediation analysis is compatible with these three notions, but achieves them in a novel way leveraging the notion of quantile remapping discussed above.

We define the direct, indirect and total effects as the contrast between two populations that have stochastically greater and stochastically smaller distributions of the exposure, such that the observed population lies between these two hypothetical populations. All effects are computed at fixed quantiles of the outcome at probability p, mediators at probabilities r_1, r_2 respectively, and the exposure at probability q. With $h_{0,\theta}(\tilde{x})$ as defined in equation 3.3, consider two populations corresponding to $h_{0,+\alpha}(\tilde{x})$ and $h_{0,-\alpha}(\tilde{x})$, for some $\alpha > 0$. In our application, we use $\alpha = 0.5$. Recall that the quantile function of the perturbed exposure $\tilde{X}'_\theta = h_{0,\theta}(\tilde{X})$ is exactly $h_{0,\theta}$. Let $h_{1,\theta}$ and $h_{2,\theta}$ denote the marginal quantile functions of $\tilde{M}'_{1,\theta}$ and $\tilde{M}'_{2,\theta}$ when $\tilde{X}'_\theta = h_{0,\theta}(\tilde{X})$.

The total effect is defined as the contrast between the p^{th} conditional quantile of the outcome at the

marginal q, r_1, and r_2 quantiles of the exposure and mediators, respectively, contrasting between the settings of the two hypothetical interventions. Using the additive rank-1 representation of the conditional quantile functions, we obtain the following expression for the total effect:

$$\begin{aligned} &Q_{Y|\tilde{\boldsymbol{W}}'_{+\alpha}}(p \,|\, \tilde{M}'_{1,+\alpha} = Q_{\tilde{M}'_{1,+\alpha}}(r_1), \tilde{M}'_{2,+\alpha} = Q_{\tilde{M}'_{2,+\alpha}}(r_2), \tilde{X}'_{+\alpha} = Q_{\tilde{X}'_{+\alpha}}(q)) - \\ &\quad Q_{Y|\tilde{\boldsymbol{W}}'_{-\alpha}}(p \,|\, \tilde{M}'_{1,-\alpha} = Q_{\tilde{M}'_{1,-\alpha}}(r_1), \tilde{M}'_{2,-\alpha} = Q_{\tilde{M}'_{2,-\alpha}}(r_2), \tilde{X}'_{-\alpha} = Q_{\tilde{X}'_{-\alpha}}(q)) \end{aligned}$$

$$\begin{aligned} &= f_1(h_{1,+\alpha}(r_1))g_1(p) + f_2(h_{2,+\alpha}(r_2))g_2(p) + f_3(h_{0,+\alpha}(q))g_3(p) - \\ &\quad f_1(h_{1,-\alpha}(r_1))g_1(p) - f_2(h_{2,-\alpha}(r_2))g_2(p) - f_3(h_{0,-\alpha}(q))g_3(p). \end{aligned}$$

The direct effect in our framework is obtained by blocking the indirect effects, and consists of one additive term in the expression for the total effect given above:

$$(f_3(h_{0,+\alpha}(q)) - f_3(h_{0,-\alpha}(q)))g_3(p). \tag{3.7}$$

The indirect effect in our framework is as follows:

$$(f_1(h_{1,+\alpha}(r_1)) - f_1(h_{1,-\alpha}(r_1)))g_1(p) + (f_2(h_{2,+\alpha}(r_2)) - f_2(h_{2,-\alpha}(r_2)))g_2(p). \tag{3.8}$$

To better understand the motivation for this construction, consider the setting where all quantiles correspond to medians, i.e. $q = r_1 = r_2 = 1/2$. We first discuss the direct effects. A median person under the $+\alpha$ intervention will have a value of the exposure equal to $h_{0,+\alpha}(1/2)$. This in turn gives them a "direct score" in our rank-1 model (corresponding to direct effects from $\tilde{X}$ to Y) that is equal to $f_3(h_{0,+\alpha}(q))$. This direct score impacts the conditional quantiles of Y in proportion to the loadings $g_3(p)$, which do not change under the intervention. We contrast this effect to what would happen under the $-\alpha$ intervention, leading to a difference as given in (3.7).

The indirect effects are slightly more subtle and we focus on indirect effects through the first mediator. In the population of $\tilde{M}_1$ values induced by the $+\alpha$ intervention, a median person will have a mediator value of $h_{1,+\alpha}(1/2)$, and an indirect score of $f_1(h_{1,+\alpha}(1/2))$. This impacts the outcome in proportion to the loading $g_1(p)$ (which is not altered by the intervention). Under the $-\alpha$ intervention, the median mediator value is $h_{1,-\alpha}(1/2)$, which has an indirect score of $f_1(h_{1,-\alpha}(1/2))$. The first term of (3.8) thus reflects the contrast between these two interventions when propagated through the first mediator.

3.2.6 Marginalizing a conditional quantile function

It remains to show how to obtain the marginal quantile functions $h_{j,\theta}$ of $\tilde{M}'_{j,\theta}$ ($j = 1, 2$) required above when constructing our proposed indirect effects. Recall that when $\tilde{X}$ (which is uniform on $(0, 1)$) is transformed by an invertible function $g : (0, 1) \to \mathbb{R}$, the quantiles of $\tilde{X}' = g(\tilde{X})$ are simply given by $g(q)$ for probability points q. Thus we easily obtain expressions for the marginal quantiles of $\tilde{X}'$ when $\tilde{X}$ is transformed by either $h_{-1,\theta}, h_{+1,\theta}$, corresponding to two interventions. It is more challenging to obtain expressions for the marginal quantiles of $\tilde{M}'_j$, although all relevant information is encoded in the marginal quantiles of $\tilde{X}'$ and the conditional quantiles of $\tilde{M}_j|\tilde{X}$. A method for accomplishing this is provided next.

A key fact is that the quantile function is the inverse of the cumulative distribution function (CDF), with the latter being easier to marginalize. For jointly distributed random variables X, Z, the conditional CDF of X given Z is $h(t; z) \equiv P(X \leq t|Z = z)$. Using the "smoothing theorem", the unconditional CDF of X is simply $E_Z P(X \leq t|Z)$, where E_Z denotes expectation with respect to the marginal distribution of Z. Access to the marginal quantiles of Z allows us to evaluate this expectation, and then we invert the resulting function to obtain the marginal quantile function of X. In practice, our representations of the relevant quantile functions are restricted to finite grids, so we use the midpoint method from elementary numerical analysis to approximate all integrals.

3.2.7 Working with longitudinal data

While some researchers conduct mediation analyses using cross-sectional data, it is generally considered to be more rigorous to use longitudinal data for this purpose. The preference for longitudinal data is motivated by the observation that when the exposure temporally precedes the mediators, and similarly if the mediators temporally precede the outcome, then it is more plausible to argue that the direct and indirect effects are causal – the temporal ordering of measurements precludes overt reverse causation. We note however that unmeasured confounders could still induce spurious associations even in longitudinal data.

Beyond the context of mediation analysis, there are two main bodies of methodology for longitudinal data [70]. One set of approaches known as "marginal modeling" aims to model only the conditional mean structure, and utilizes only observed variables. Generalized Estimating Equations (GEE) and marginal regression using "robust" inference are popular ways to proceed in this vein [38]. Importantly, these approaches do not deny that within-subject correlations exist, but account for the correlations indirectly by adjusting the standard errors, and sometimes by reweighting or decorrelating the observations. In contrast, latent variable and multilevel approaches, such as mixed effects regression, introduce random effects that allow a model to explicitly capture the dependence in the data along with its mean structure. In our context, the local non-parametric

quantile estimates provided by the QNN procedure are marginal since we estimate the quantiles by pooling observations over all subjects.

The basis of a marginal approach to mediation analysis is that even when working with longitudinal data, it still makes sense to consider marginal relationships among variables, e.g. as expressed through a conditional quantile function $Q_{Y|\boldsymbol{X}}(p|\boldsymbol{x})$. It is important to note that the proper interpretation of this object in a marginal setting is that it provides the marginal p^{th} quantile of Y among all subjects with $\boldsymbol{X} = \boldsymbol{x}$. In this sense it is a "population averaged" rather than a "subject-specific" quantity.

Statistical inference remains a challenge, since while we can justify ignoring the clustering of data within subjects when constructing point estimates, this cannot be justified when carrying out statistical inference. Our approach to inference is based on data emulation and is essentially the parametric bootstrap. However we use a surrogate model for emulation, since emulated data should aim to exhibit the dependence structure of the population as well as its mean structure, and our quantile-based analysis deliberately excludes consideration of correlations or other forms of within-subject dependence. In this thesis we only provide a limited exploration of this approach to inference, focusing primarily on methods for obtaining point estimates and their interpretation.

A second challenge to working with longitudinal data is that some longitudinal covariates may not be observed at a common set of time points for all subjects. We address this challenge using kernel weighting as follows. Suppose that Y_{it}, X_{it} are a response and predictor observed at time t for subject i. Our goal is to estimate $Q_{Y|X}(p|X_{t^*} = x^*)$, where t^* is a given time value and x^* is a given covariate value. A regularized estimation approach based on the check function ρ_p would minimize

$$L(\theta; t^*, x^*) \equiv \sum_{i=1}^{n} \sum_{j=1}^{n_i} w_{ij} \rho_p(\theta - Y_{it_{ij}}) \tag{3.9}$$

over θ, where p is a probability point, and the w_{ij} are weights of the form

$$w_{ij} = \omega_{\lambda_1}(x_{it_j} - x^*) \cdot \omega_{\lambda_2}(t_{ij} - t^*), \tag{3.10}$$

with $\omega_\lambda(x) = \exp(-x^2/\lambda)$, and λ_1, λ_2 are bandwidth parameters. In principle this approach can be extended to accommodate multiple covariates however the weights become very skewed leading to high variance estimators unless the sample size is very large.

3.3 Simulation

In this section we access the performance of our entire pipeline that uses QNN to obtain a tensor of conditional quantile estimates, which is decomposed using AFQR to obtain the functional components that are further used to obtain the mediation effects in the mediation analysis. We consider three data settings, given below.

Case 1: Homoscedastic Data

$$
\begin{aligned}
&X_j \sim \mathcal{N}(0,1), \\
&M_1 = X + \epsilon_{M_1}, && \epsilon_{M_1} \sim \mathcal{N}(0,1) \\
&M_2 = X + \epsilon_{M_2}, && \epsilon_{M_1} \sim \mathcal{N}(0,1) \\
&Y = a_1 X + a_2 M_1 + a_3 M_2 + \epsilon_Y, && \epsilon_Y \sim \mathcal{N}(0, 0.5)
\end{aligned}
$$

In this case, the conditional quantile function has a closed form given by

$$Q_{Y|(M_1,M_2,X)}(p|m_1, m_2, x) = a_1 x + a_2 m_1 + a_3 m_2 + 0.5 \cdot Q\mathcal{N}(0,1)(p).$$

Case 2: Heteroscedastic Data (Gaussian)

$$
\begin{aligned}
&X \sim \mathcal{N}(0,1), \\
&M_1 = X + \epsilon_{M_1}, && \epsilon_{M_1} \sim \mathcal{N}(0,1) \\
&M_2 = X + \epsilon_{M_2}, && \epsilon_{M_1} \sim \mathcal{N}(0,1) \\
&Y = a_1 X + a_2 M_2 + \sigma_s(M_1)\epsilon_Y && \epsilon_Y \sim \mathcal{N}(0, 0.5) \\
&\sigma_s(m_1) = \sqrt{(1+m_1)^2}
\end{aligned}
$$

In this case, the conditional quantile function has a closed form given by

$$Q_{Y|(M_1,M_2,X)}(p|m_1, m_2, x) = a_1 x + a_2 m_2 + 0.5\sigma_s(m_1) \cdot Q_{\mathcal{N}(0,1)}(p).$$

Case 3: Heteroscedastic Data (Exponential)

$$
\begin{aligned}
&X \sim \mathcal{U}(1,2), \\
&M_1 = X + \epsilon_{M_1}, && \epsilon_{M_1} \sim \mathcal{U}(-1,1) \\
&M_2 = X + \epsilon_{M_2}, && \epsilon_{M_1} \sim \mathcal{U}(-1,1) \\
&Y \sim exp(\mu), && \mu = X + M_1 + M_2
\end{aligned}
$$

In this case too, the conditional quantile function has a closed form given by

$$Q_{Y|(M_1,M_2,X)}(p|m_1, m_2, x) = -log(1-p)(x + m_1 + m_2).$$

In case 1 above, the direct and indirect effects are constant (derived later in the section). This is in part due to the linearity of the low rank component g. In chapter 2 we discuss in detail that if the conditional outcome belongs to a location family, then $\boldsymbol{v}$ (discretized g functions) in the low rank decomposition are constant. This results in the conditional quantile curves being parallel to each other at different values of the independent variable. Homoscedastic data (location family) leads to simple (constant) mediation effects. This is why we consider non-homogeneous data in cases 2 and 3, which leads to non-trivial $\boldsymbol{v}$ functions, leading to non-constant mediation effects.

3.3.1 Simulation parameters

We evaluate and assess the mediation effects for each of the three cases above. Per case, we generate iid samples of size 4000, and aggregate the results over 50 Monte Carlo replicates. First, we perform QNN to obtain a tensor of conditional quantile estimates. QNN contains 3 tuning parameters, the lasso smoothing penalty, the neighborhood size and the bandwidth for obtaining smooth estimates; we use the recommended values for the first two (refer to chapter 2 section 2.2 for more details), and use a bandwidth of **1** for the last parameter. The results are not too sensitive to this bandwidth for values between $[0.5, 1.0]$. Next, we decompose the estimated tensor of quantiles into a low-rank functional additive model, using AFQR. This model contains a set of functional components for each independent variable, i.e. for independent variable i, we obtain $(\boldsymbol{u}_i, \boldsymbol{v}_i)$. AFQR contains one tuning parameter which controls the degree of regularization ($c_{u,j}, c_{v,j}$. We use $c_{u,\cdot} = 0, c_{v,\cdot} = 1000$ for cases 1 and 2, and $c_{u,\cdot} = 1000, c_{v,\cdot} = 1000$ for case 3. Lastly, we estimate the mediation effects which are defined as contrasts of two perturbations expressed in terms of the discretized functional components of AFQR $(\boldsymbol{u}, \boldsymbol{v})$. The only parameter here is the degree of perturbation (α in section 3.2.5), and we use $\alpha = 0.5$. This is discussed in more detail towards the end of this section.

The true forms of the direct effects can be derived for all the 3 cases, but doing so for the indirect effects is a lot trickier. This is because the latter involves finding the marginal quantile function of the perturbed mediators which is tricky as it does not necessarily have a closed form. We do however know the underlying forms of the $\{\boldsymbol{u}_j, \boldsymbol{v}_j; j = 1, 2, 3\}$ functions for all the cases, which can be used to compute the empirical true values of the indirect effects. In the simulation results presented below, we compare the estimates with the empirical true values of the indirect effects computed by using the theoretical values of $\{\boldsymbol{u}_j, \boldsymbol{v}_j; j = 1, 2, 3\}$ in the code used to obtain the mediation effects.

3.3.2 Simulation study results

All mediation effects are bivariate functions of probability points. Thus, the direct and indirect effects are functional matrices; we jointly model these matrices to find a rank-one estimate per matrix. In reality, we evaluate the mediation effects on a discrete probability grid $\boldsymbol{p} \in \mathbb{R}^m$ to obtain discretized versions of the direct and indirect effects which are $m \times m$ matrices. Therefore, estimating the mediation effects can be viewed as a matrix estimation problem. At the same time, since each element of this matrix has its own meaning and properties, the estimate of each point of the matrix is important (and not just the overall matrix estimate).

We consider two broad criterion to evaluate the performance of our method. First, we assess the fit at a matrix level and consider two criterion. We report the relative Frobenius norm of the error as defined in 3.11, and the correlation of the estimated curves with the functional matrix (vectorized matrices). The Frobenius norm of the error gives us an idea about how "far" the estimated functional matrix curve (vectorized matrix) is from the true functional matrix, and the correlation gives us a sense for the similarity in the shape of the two curves.

Per replicate i, the Frobenius norm of the error, referred to as the Frobenius RMSE, of an effect $E \in \mathbb{R}^{m \times m}$ is given as follows, where E is the direct or indirect effect.

$$\text{RMSE}_i = \frac{\|\text{vec}(\hat{E}) - \text{vec}(E)\|_2}{\|\text{vec}(E)\|_2} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{m}\left(\hat{E}_{ij} - E_{ij}\right)^2}{\sum_{i=1}^{m}\sum_{j=1}^{m} E_{ij}^2}. \tag{3.11}$$

Next, we evaluate the fit at a pointwise level. Note that the underlying values of the effects are mostly very small, often close to 0, making a point-wise analysis harder to interpret than the Frobenius RMSE which integrates over all points (regardless of how small the underlying value is), providing a one-number summary of the fit.

The simulation results assessed at the matrix level are presented in table 3.1. As we can see, the relative integrated MSE is less than 10% for all the mediation effects (DE, IDE1, IDE2) in case 1, and below or around 20% for cases 2 and 3. The correlation of the curves is above 90% for case 3, and above 55% for the other cases. This is not a matter of concern because we know that our estimates are close to the truth (implied by low integrated RMSE), but they are a bit less smooth in that they are more bouncy. This can be controlled by the regularization penalty in AFQR if one wishes.

Next, we present the point-wise RMSE and point-wise relative bias (rbias) of the mediation effects. Specifically for any mediation effect $\text{ME}(p_y, p_w)$ (where $W = X$ for direct effect and $W = M_1(M_2)$ for indirect effect 1(2)), the estimation error $r(p_y, p_w) := \widehat{\text{ME}}(p_y, p_w) - \text{ME}(p_y, p_w)$ has a systematic component (bias) and a random component (variance). But the systematic component

Mediation effect	RMSE		Correlation	
	Mean	SD	Mean	SD
Simulation Case 1				
DE	0.085	0.017	0.842	0.125
IDE1	0.073	0.020	0.584	0.350
IDE2	0.070	0.021	0.615	0.262
Simulation Case 2				
DE	0.103	0.024	0.585	0.090
IDE1	0.217	0.049	0.981	0.008
IDE2	0.094	0.023	0.552	0.260
Simulation Case 3				
DE	0.184	0.092	0.983	0.017
IDE1	0.181	0.065	0.977	0.021
IDE2	0.178	0.060	0.967	0.029

Table 3.1: Results of simulation study for the mediation analysis pipeline (at a matrix level).

of the error can be modelled, and we consider a least squares model with the estimation errors as the dependent variable, as follows

$$\widehat{\mathrm{ME}}(p_y, p_w) - \mathrm{ME}(p_y, p_w) \sim \Big((p_y - 0.5) - (p_w - 0.5)\Big)^2 + \mathrm{ME}(p_y, p_w) + \\ \mathrm{ME}(p_y, p_w)^2 + h_z\big(\mathrm{ME}(p_y, p_w)\big) + h_z\big(\mathrm{ME}(p_y, p_w)\big)^2$$

where $h_z(X) = X - \text{median}(X)$. This model helps us understand the factors that drive the bias. Naturally, it is not a complete model for the bias, since the systematic errors could depend on other factors, however the adjusted R^2 values of the regression models for different mediation effects (MEs) are quite high suggesting that this model can be used for predicting the bias. The model is fit using ordinary least squares to a dataset that consist of m^2 estimates of the ME per Monte Carlo replicate stacked together, along with the corresponding probability points (p_y, p_w) at which the ME in each row is evaluated. The fitted value at each point (p_y, p_w) is an estimate of $E[\widehat{\mathrm{ME}}(p_y, p_w) - \mathrm{ME}(p_y, p_w)]$, which is the bias at that point. Diving the fitted value by the true underlying value $\mathrm{ME}(p_y, p_w)$ yields an estimate of the relative bias.

We wish to report a one-number summary of the relative bias and relative mse for the mediation effects. However, we get a relative measure per observation, giving us a total of m^2 estimates. So we can report the average or median value of these estimates, and we go with the former. While computing the average, positive and negative biases cancel out which may mislead us into thinking that the relative bias is lower than it actually is. To avoid this, we compute the average of the

squared bias, and report it is square root so that it is on the same scale as the relative bias. To compute the mse, we denote the fitted value of the error at (p_y, p_w) for the i^{th} replicate by

$$\hat{r}^{(i)}(p_y, p_w) = E\big[\widehat{\mathrm{ME}}^{(i)}(p_y, p_w) - \mathrm{ME}(p_y, p_w)\big]. \tag{3.12}$$

We can compute the sample variance of $\{\hat{r}^{(i)}(p_y, p_w) : p. \in \boldsymbol{p}, i = 1, \cdots, n\}$, and obtain the bias square from below, to compute the MSE and relative MSE. The formulae for all these measures are given below

$$\begin{aligned}
\text{Avg bias} &= \underset{i,p_y,p_w}{\text{mean}} \left(\hat{r}^{(i)}(p_y, p_w) \right) \\
\text{Avg rbias} &= \underset{i,p_y,p_w}{\text{mean}} \left(\frac{\hat{r}^{(i)}(p_y, p_w)}{\mathrm{ME}(p_y, p_w)} \right) \\
\text{Avg rbias}^2 &= \sqrt{\underset{i,p_y,p_w}{\text{mean}} \left(\frac{\hat{r}^{(i)}(p_y, p_w)}{\mathrm{ME}(p_y, p_w)} \right)^2} \\
\text{Av rel MSE} &= \underset{i,p_y,p_w}{\text{mean}} \left(\frac{\hat{r}^{(i)}(p_y, p_w)^2 + var_{p_y,p_w,i}\big(r^{(i)}(p_y, p_w)\big)}{|\mathrm{ME}(p_y, p_w)|} \right)
\end{aligned}$$

Another advantage of modeling the bias aside from obtaining a one-number summary of the relative bias and relative MSE is that it requires fewer simulation runs than a direct approach that estimates the bias as the sample mean of $\widehat{\mathrm{ME}} - \mathrm{ME}$ over Monte Carlo replicates, since each Monte Carlo replicate contributes m^2 estimates instead of a single one.

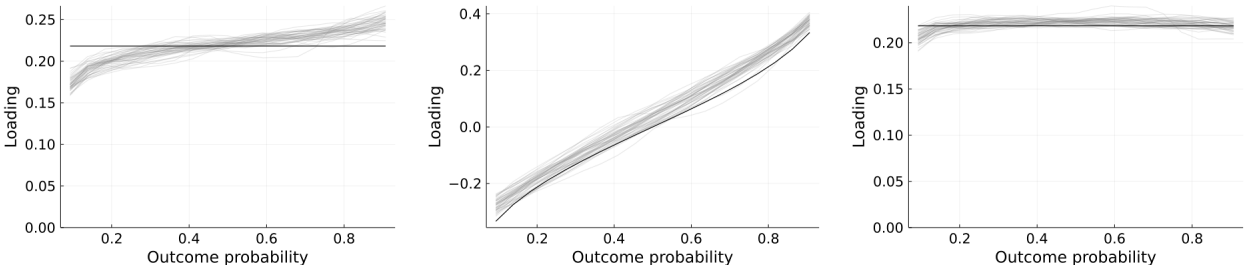
Mediation Effect	Sim. Case	Avg R^2	Avg rbias	Avg rel bias^2	MSE
DE	1	0.04	0.084	0.085	0.009
IDE1	1	0.07	−0.052	0.056	0.005
IDE2	1	0.07	−0.057	0.060	0.005
DE	2	0.09	−0.076	0.088	0.032
IDE1	2	0.44	−0.058	0.149	0.148
IDE2	2	0.13	−0.070	0.084	0.015
DE	3	0.06	0.142	0.162	0.092
IDE1	3	0.70	−0.116	0.156	0.045
IDE2	3	0.72	−0.080	0.144	0.047

Table 3.2: Model-based Monte-Carlo estimates of sampling bias and MSE for estimates in the mediation analysis pipeline.

The average relative bias for the homoscedastic Gaussian case (case 1) is below 10% and for the other two cases is roughly below 17%. The average relative MSE appears to be below 15% for all cases. The point-wise heatmaps of all the estimates (direct effects and indirect effects) for the 3 cases are included in appendix A.

3.3.3 Remarks

Statistical bias : Statistical estimates of regression effects based on local fitting tend to be biased toward zero ("attenuation bias") due to oversmoothing. This means that the relative bias will generally be negative. The effect of covariate j in our model is represented by the matrix $u_j v_j'$, and we note that the relative bias for the estimates presented above is not always negative, although the magnitude of any positive bias is less than 10% in most cases. Estimates of the score vector ($\boldsymbol{u}$) are biased towards zero, consistent with attenuation bias. However the loading vector $\boldsymbol{v}$ is a norm-1 direction vector, so it has a non-Euclidean domain which complicates interpretation of its bias. This is displayed in figure 3.3 which contains the estimates of $\boldsymbol{v}$ for case 2. Recall that $\boldsymbol{v}$ is a function of probability points evaluated at a probability grid $\boldsymbol{p}$; regions of $\boldsymbol{p}$ for which $\hat{\boldsymbol{v}}$ is biased away from 0 may lead to mediation effects also biased away from 0. Since AFQR performs well and leads to good estimates for $\boldsymbol{u}, \boldsymbol{v}$, the bias away from 0 (or positive relative bias) in $\boldsymbol{v}$ is quite small, leading to small positive relative bias in the mediation effects.



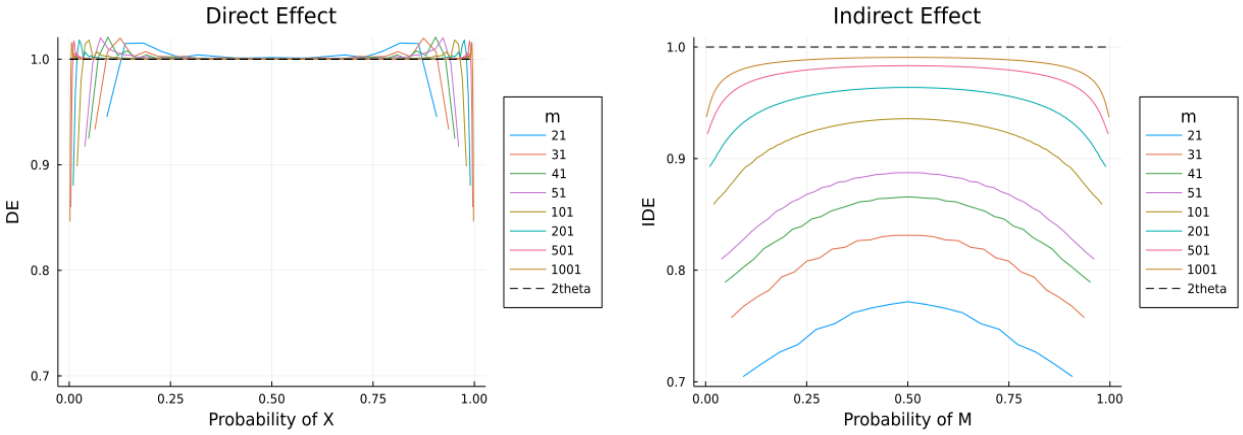
(a) $\hat{V}_1$ (b) $\hat{V}_2$ (c) $\hat{V}_3$

Figure 3.3: AFQR estimate V for simulation case 2.

Discretization bias : One source of bias in our analysis seems to arise from discretization effects. All quantile functions studied here are represented with vectors that are their evaluations on a grid of probability points. In calculating the indirect effect, we use a marginalization procedure that involves inverting the quantile function to a cumulative distribution function, marginalizing the conditional CDF over a covariate, then inverting back to a quantile function. It appears that coarse discretization induces bias in this marginalization.

If we evaluate the direct and indirect effects using the true forms of $\boldsymbol{u}, \boldsymbol{v}$ on a probability grid $\boldsymbol{p}$, then due to the discretization we do not get the exact same mediation effects as their underlying theoretical values. Recall that the direct and indirect effects at fixed probability points of the (exposure/mediator, outcome) (p, p_y), are simply the outer products of V with the contrast of the $\boldsymbol{u}$ functions evaluated at two distinct probability points $p_1 < p < p_2$. Obtaining the probability points p_1, p_2 is relatively straightforward for the direct effect, but more complicated for the indirect effect. This is because the perturbed probability points are obtained from the marginal quantile function of the perturbed mediator which is tricky to calculate. As an example, let us consider one specific simulation case.

For the homoscedastic simulation case 1, the mathematical error for the direct and indirect effects is presented in figure 3.4. In this case, the theoretical value of the direct and indirect effects is equal to 1, which is 2θ where $\theta = 0.5$ is the parameter of perturbation. These underlying values of the mediation effects can be derived analytically for this simple case. Since we perturb the exposure by $\pm\theta$, and the mediators are linear in the exposure, they too get perturbed by $\pm\theta$, and as a result the outcome gets perturbed by $\pm 3\theta$. The mediation effects are defined as the contrast of two perturbations corresponding to $+\theta = 0.5$ and $-\theta = -0.5$. Therefore, the direct effect, which is the contrast of the two perturbed outcomes due to the exposure (blocking the impact of the mediators) will be $\theta - (-\theta) = 2\theta$. Similarly the indirect effects are the contrasts in the perturbed outcomes due to each mediator, which will also be $\alpha - (-\alpha) = 2\theta$. Since we use $\theta = 0.5$, the true mediation effects in case 1 are equal to one (for all DE, IDE1, IDE2).



(a) Mathematical bias (DE)

(b) Mathematical bias (IDE)

Figure 3.4: Mathematical error in the direct and indirect effect estimates for simulation case 1.

We can combat this by using a finer grid of probability points along which the estimates of $\boldsymbol{u}, \boldsymbol{v}$ are computed. This is denoted by m in the plots above. A probability grid of length above $m \geq 200$

points seems to have less than 10% mathematical bias, however using such a fine probability grid is not computationally feasible. This is because our low rank decomposition takes place on a cartesian product of these probability grids. Recall that we construct a tensor of quantile estimates, where each dimension of the tensor is indexed by $\boldsymbol{p}$ and is thus of length m. The dimension of the low rank optimization problem is m^4 in the case of 2 mediators, which grows exponentially with m. In particular, as we go from $m = 21$ to $m = 41$, the optimization problems grows from $21^4 = 194,481$ parameters to $41^4 = 2,825,761$ parameters, which is $\sim 2^4 = 16$ times bigger. This is why in the simulation study above we have presented results for $m = 21$ and evaluated their performance based on the oracle.

3.4 Illustration using Dogon Longitudinal Study data

In this section we present the results of the entire mediation analysis pipeline on data from the Dogon Longitudinal study. As previously described in section 3.1.1, our goal is to disentangle the effects of childhood and adult body size measures onto adult systolic blood pressure (SBP). We apply the entire mediation analysis pipeline starting from QNN to AFQR to the mediation analysis, in order to get the direct effect of childhood body size on adult SBP, and the indirect effects of adult body size (adult height and adult BMI) on adult SBP. Please note that in our application, the adult body sizes are measured at the same time (age) as adult SBP. The unit of time in this longitudinal data set is the age at which individuals are measured.

To get "childhood" and "adult" measures, we focus on two ages, one in childhood ($\text{age}_1 < 10$) and one in adulthood ($\text{age}_2 > 12$). While individuals aged $12 - 16$ years are generally not considered adults, we refer to them as adults here since we have SBP measurements after the age of 12. Note that every individual is measured at a different set of time points (ages), which are noted as floats (and not rounded to the nearest integer). Thus, upon fixing a childhood age, almost no one in the dataset will be measured at exactly that age. As mentioned earlier, we use kernel methods to incorporate measurements taken at nearby ages to the fixed target ages. In particular, we consider a childhood caliper of 1.5 years, and an adult caliper of 2 years around the target ages. We only consider observations within these neighborhoods (for instance consider observation at time t if $t \in [\text{age}_1 - 1.5, \text{age}_1 + 1.5]$, or $t \in [\text{age}_2 - 2, \text{age}_2 + 2]$), and use all within individuals pairs of observations (each pair contains one childhood and one adult observation). So if an individual only has one (or multiple) childhood observations without any adult observations in the desired neighborhoods, they will not contribute to the mediation estimates.

We conduct the entire analysis pipeline for multiple settings. We consider 3 different measures for childhood body size, namely Body Mass Index (BMI), Height adjusted z-score (HAZ), and weight adjusted z-score (WAZ). These variables form the exposure (X). The mediators are the

adult body size measures, which are always fixed to adult HT (M_1), and adult BMI (M_2). We consider all childhood body size measures across 3 different childhood ages – 1, 3, and 5 years. The analysis is conducted independently for females and males.

Estimation pipeline

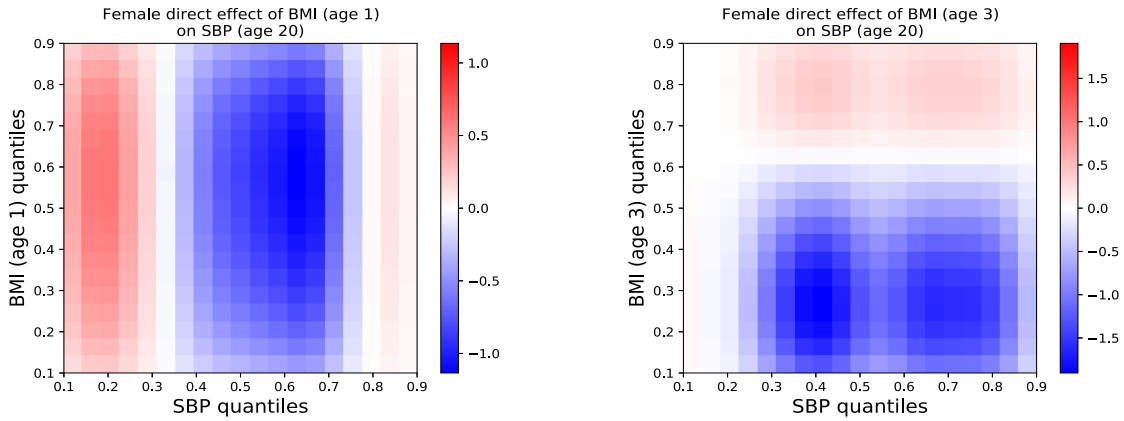
Per setting, we start by estimating a tensor of conditional quantiles using QNN. The tensor is indexed by probability points along a grid $\boldsymbol{p} = \{p_i\}_{i=1}^{m}$ of length $m = 21$. We consider an equispaced grid spanning points from $p = 1/m$ to $p = 1 - 1/m$. The tensor is then of dimension $m \times m \times m \times m$, such that its axes corresponds to probabilities of the exposure, each of the adult body sizes (HT and BMI resp), and the outcome respectively. We use default values of the tuning parameters in the QNN algorithm to obtain the m^4 estimates that form this tensor. In particular, cell (i_1, i_2, i_3, i_4) contains the estimate of the $p_{i_4}^{th}$ quantile of the outcome, conditioned on the $p_{i_1}^{th}$ quantile of observed X, $p_{i_2}^{th}$ quantile of observed M_1, and $p_{i_3}^{th}$ quantile of observed M_2, i.e., $\hat{Q}_{(Y|M_1,M_2,X)}(Q_X(p_{i_1}), Q_{M_1}(p_{i_2}), Q_{M_2}(p_{i_3}))$. The marginal quantiles of the independent variables are their order statistics.

We then fit the AFQR model to this tensor of quantiles, and compute the mediation effects using the AFQR functional component estimates. We present the direct and indirect effects for select cases below.

3.4.1 Results

The direct effect is a bivariate function of the probability points for the exposure quantiles (childhood body size, e.g. as measured by BMI), which are denoted by p_x, and the probability points for adult SBP quantiles, which are denoted by p_y. The direct effect at p_x, p_y reflects how the p_y^{th} conditional quantile of adult SBP at age age_2 for someone who is at the marginal p_x^{th} quantile of childhood BMI at age age_1 differs between the two intervention populations, when blocking any indirect effects through the mediators.

Direct effects Figure 3.5a displays the direct effect of childhood BMI at age 1 on adult SBP at age 20, among females. We can see that the direct effect is negative for the mid-upper quantiles of adult SBP $(0.4 \leq p_y \geq 0.75)$ whereas it is almost 0 elsewhere. This means that the lower quantiles of adult SBP are practically not affected by these hypothetical interventions, and people who are at the lower spectrum of adult SBP potentially have other factors besides childhood body size that determine their adult SBP. On the other hand, the mid-upper quantiles of adult SBP which are of more interest in any case, are lowered by such a hypothetical contrast. Regardless of childhood BMI at age 1, if the females in the observed population had higher childhood BMI rather than



(a) Direct Effect at $\text{age}_1 = 1$

(b) Direct Effect at $\text{age}_1 = 3$

Figure 3.5: Direct effects for childhood body size variable BMI measured at ages 1 and 3 on adult SBP at age 20 among females.

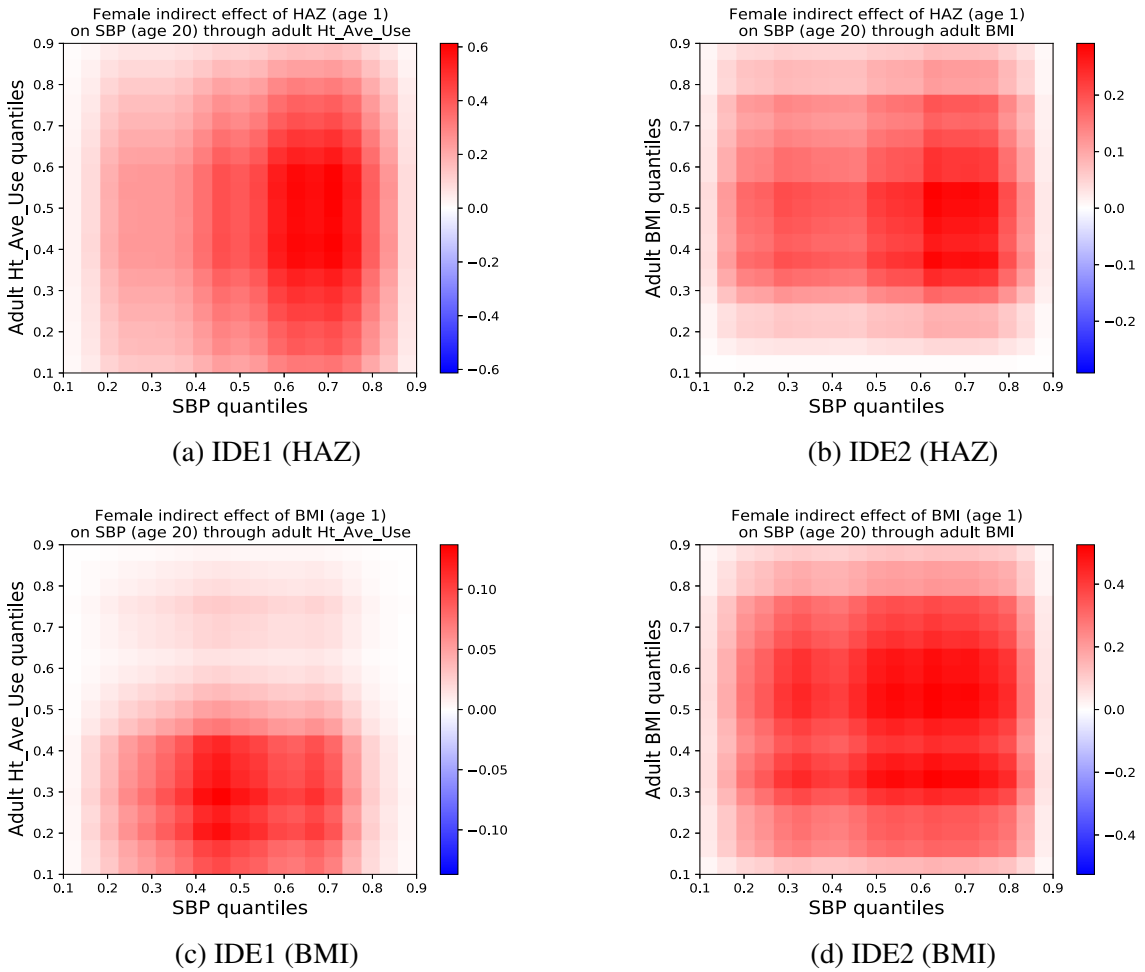
having lower childhood BMI, the mid-upper quantiles of adult SBP for women would have been between 1 and 1.2 mm Hg lower than in the observed population.

The direct effect of childhood BMI at age 3 among females is even stronger, and is much more concentrated toward the lower quantiles of childhood BMI, as seen in figure 3.5b. The impact of making smaller children bigger vs even smaller on the mid-upper quantiles of adult SBP ($p_y \geq 0.3$), blocking any indirect impact on SBP through adult body size, will be a decrease in SBP by an amount between 1 and 1.75 mm Hg.

The mediation effects for females at childhood age 1, using childhood body size measures HAZ and WAZ have similar trends to those using childhood BMI (figure 3.5). The direct effects are negative in the upper quantiles of adult SBP throughout the range of childhood body size, and are between -1 and -1.2 around $0.6 - 0.7$ quantiles of adult SBP. The indirect effects through both adult body size variables are positive, and very small (between 0 and 0.4).

Indirect effects The indirect effect is a bivariate function of the probability points of the mediator quantiles (e.g. adult HT or BMI), denoted p_m, and the probability points for quantiles of adult SBP outcome, denoted p_y. The indirect effect at p_m, p_y captures the intervention effects that propagate through a mediator. Specifically, it tells us something about individuals who end up at the p_m^{th} quantile of the mediator distributions, under the two interventions. For these people, the p_y^{th} quantiles of their SBP distributions will differ by the amount given by the indirect effect.

The indirect effects for all three childhood body size measures as exposure are always either positive or essentially zero. This is expected in our analysis as we know that cross-sectionally bigger (smaller) people have higher (lower) SBP, referred to as simultaneous correlation. A hypo-



(a) IDE1 (HAZ)

(b) IDE2 (HAZ)

(c) IDE1 (BMI)

(d) IDE2 (BMI)

Figure 3.6: Indirect effects through adult HT (IDE1, left) and adult BMI (IDE2, right) when the exposure is childhood HAZ (top row) and childhood BMI (bottom row), both measured at 1 year, on adult SBP at 20 years for females.

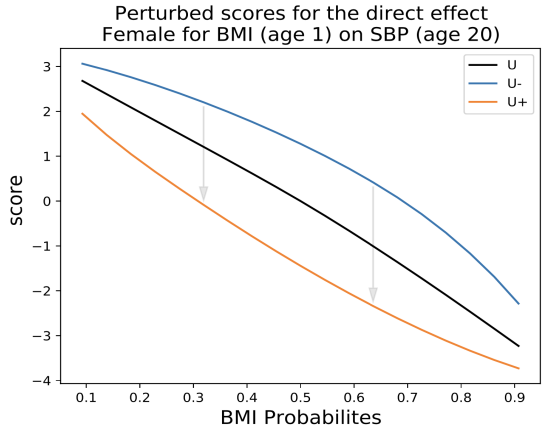
thetical intervention that increases the childhood body size of the entire population will also lead to an increase in the adult body size due to tracking. The impact of this increase in the adult body size on adult SBP is captured by the indirect effect. We consider two adult body size measures as mediators, namely HT and BMI. The indirect effect due to each of them is referred to as the indirect effect through mediator (HT/BMI).

Figure 3.6 presents the indirect effects of HT and BMI on adult SBP when the childhood measure is HAZ and BMI for females at age one. The first row corresponds to childhood HAZ and the second to BMI. In the first case, the indirect effect through HT is more than double the indirect effect through BMI almost everywhere, as HAZ tracks stronger with adult HT than adult BMI. The reverse is seen for childhood BMI, which naturally tracks stronger with adult BMI than adult HT. In fact this can be confirmed by fitting marginal AFQR models for the quantiles of adult BMI

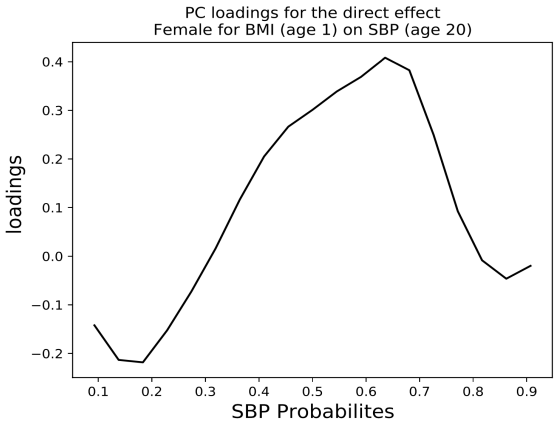
and HT, each regressed on childhood BMI and HAZ. Upon doing so, we find that all the trends are positive, but the effects are much larger for BMI to BMI and HAZ to HT than the other way around.

Note that the change in the distribution of adult body size due to the hypothetical intervention will likely not be proportional to the change in the exposure distribution. Moreover, the two mediators (HT, BMI) will track to varying degrees for different childhood body size measures, leading to unequal and differently distributed indirect effects. For instance, childhood BMI tracks more strongly to adult BMI than adult HT, and vice-versa for childhood HAZ.

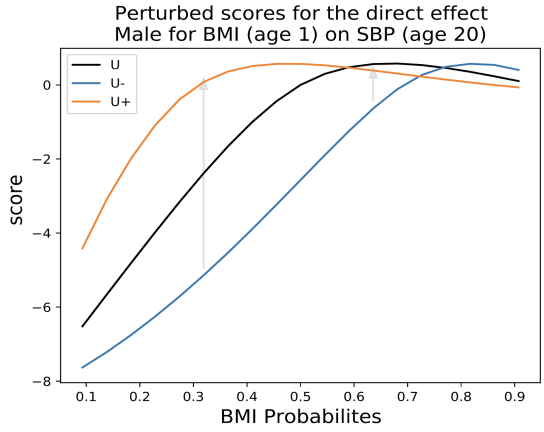
Understanding the effects and the perturbed scores contrast Next, we explore the workings/ details of the statistical method behind the direct and indirect effects, with the help of an example. We note that the direct effect of BMI at age one on adult SBP at age 20 for males is postive, unlike the corresponding result for females (presented in figure 3.5). The anthropometric implication of this is that among all children of age one, females will benefit from a food intervention that increases their BMI more than males. In fact, this is also the case for childhood HAZ since the direct effect of HAZ at age one was weaker amongst males than females.



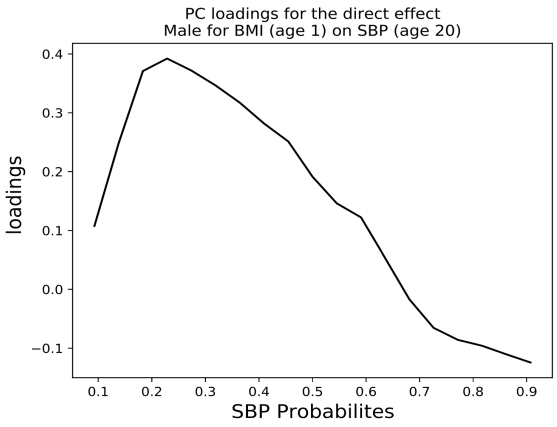
(a) Females : Perturbed scores



(b) Females : loading



(c) Males : Perturbed scores



(d) Males : loading

Figure 3.7: Perturbed score vectors $(\boldsymbol{u}^+, \boldsymbol{u}^-)$ and common loadings $(\boldsymbol{v})$ for childhood BMI at age 1 on adult SBP at age 20, among females (first row) and males (second row). The score vector for the observed population is $\boldsymbol{u}$. The direct effect is defined as $(\boldsymbol{u}^+ - \boldsymbol{u}^-) \otimes \boldsymbol{v}$.

Focusing on the statistical side of the results, let us explore why we are getting opposite effects for different genders, while all other variables/parameters in the model are the same. Figure 3.7 presents the plots for the perturbed scores and the common loadings they are multiplied with for both genders. Recall that our effects are defined as the contrast between $\boldsymbol{u}^+$ and $\boldsymbol{u}^-$ at every quantile of the exposure (p_x). The impact of such a contrast on the p_y^{th} quantile of the outcome is obtained by multiplying $\boldsymbol{u}^+ - \boldsymbol{u}^-$ at probability point p_x by the value of the loading $(\boldsymbol{v})$ at p_y. In terms of the definition of direct effect given in equation 3.7, $\boldsymbol{u}^+, \boldsymbol{u}^-$ reference $\boldsymbol{u}_3^+, \boldsymbol{u}_3^-$, which are the discretized versions of $f_3(h_{0,+\alpha}(q)), f_3(h_{0,-\alpha}(q))$ on $\boldsymbol{p}$ respectively, while $\boldsymbol{v}$ references $\boldsymbol{v}_3$ which is the discretized version of g_3 on $\boldsymbol{p}$. As an example, the direct effect at median childhood BMI onto the 0.7 quantile of adult SBP is $\big(1.7 - (-1.5)\big) \times 0.38 = 1.2$ mmHg for females (ref figures 3.7a, 3.7b) and $\big(0.5 - (-2.4)\big) \times -0.02 \approx 0$ mmHg for males (ref figures 3.7c, 3.7d).

In general, if $\boldsymbol{u}^+(p_x) > \boldsymbol{u}^-(p_x)$ then $\boldsymbol{v}(p_y) > 0 \implies DE(p_x, p_y) > 0$ and $\boldsymbol{v}(p_y) < 0 \implies DE(p_x, p_y) < 0$, and vice verse. This is summarized in figure 3.8.

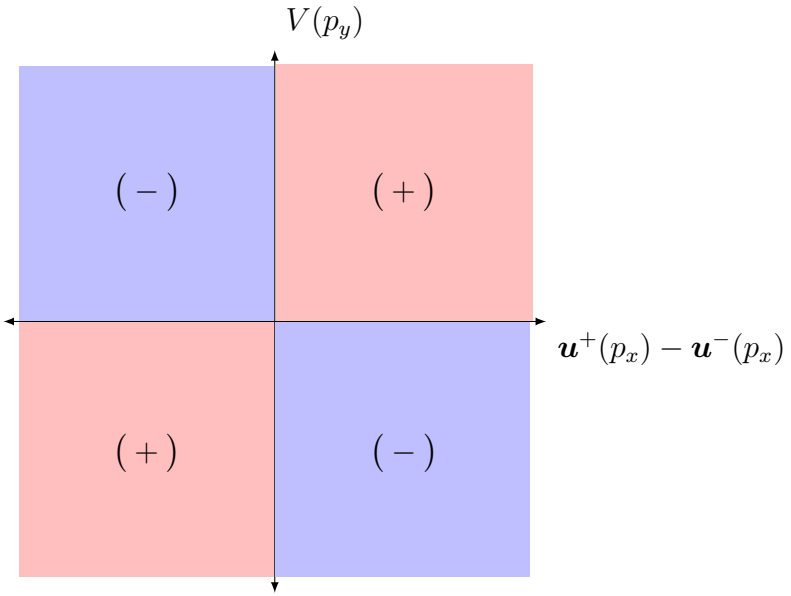


Figure 3.8: Sign of $DE(p_x, p_y)$ based on the signs of the score contrast ($sign(\boldsymbol{u}^+(p_x) - \boldsymbol{u}^-(p_x))$) and the loading ($V(p_y)$) that are multiplied to obtain the direct effect (DE) at the point (p_x, p_y).

Lastly, we comment on the trends in $\boldsymbol{u}^+, \boldsymbol{u}^-, \boldsymbol{u}$. One might expect $\boldsymbol{u}$ to be between $\boldsymbol{u}^+$ and $\boldsymbol{u}^-$ as the hypothetical interventions are designed in a way that $h_{0,+}$ corresponds to a more advantaged population in which all the exposure quantiles are higher, and $h_{0,-}$ corresponds to a population with lower exposure quantiles than the observed population (at every probability point). This design leads to two populations that have stochastically greater and stochastically smaller distributions of the exposure than the observed population, and $\boldsymbol{u}^-, \boldsymbol{u}, \boldsymbol{u}^+$ are the scores corresponding to the less advantaged $(h_{0,-})$, observed, and more advantaged $(h_{0,+})$ populations respectively. However, we can see that $\boldsymbol{u}^+ > \boldsymbol{u} > \boldsymbol{u}^-$ need not necessarily be the case as seen in figures 3.7c, 3.9a and 3.9c. Let us understand why this happens. There are different mechanisms at play for the direct and indirect effects, and thus different reasons for the trends in $\boldsymbol{u}^+, \boldsymbol{u}^-, \boldsymbol{u}$ for the exposure and mediator variables.

For the exposure (and direct effect), we expect $\boldsymbol{u}^+ \geq \boldsymbol{u} \geq \boldsymbol{u}^-$ if $\boldsymbol{u}$ is an increasing function. We explain $\boldsymbol{u}^+ \geq \boldsymbol{u}$, and $\boldsymbol{u} \geq \boldsymbol{u}^-$ follows by the same logic. Recall that by definition,

$$\boldsymbol{u}(p_x) = f_x(p_x) = f_x(\tilde{x})$$

for some smooth function f_x. Further, perturbed x corresponding to $+\theta$

$$\tilde{x'}_{+\theta} = \phi(\phi^{-1}(\tilde{x}) + \theta),$$

and then $\boldsymbol{u}^+$ is

$$\boldsymbol{u}^+ = \boldsymbol{u}(\tilde{x}_{+\theta} = f_x(\phi(\phi^{-1}(\tilde{x}) + \theta)) = \phi(\phi^{-1}(\tilde{p_x}) + \theta).$$

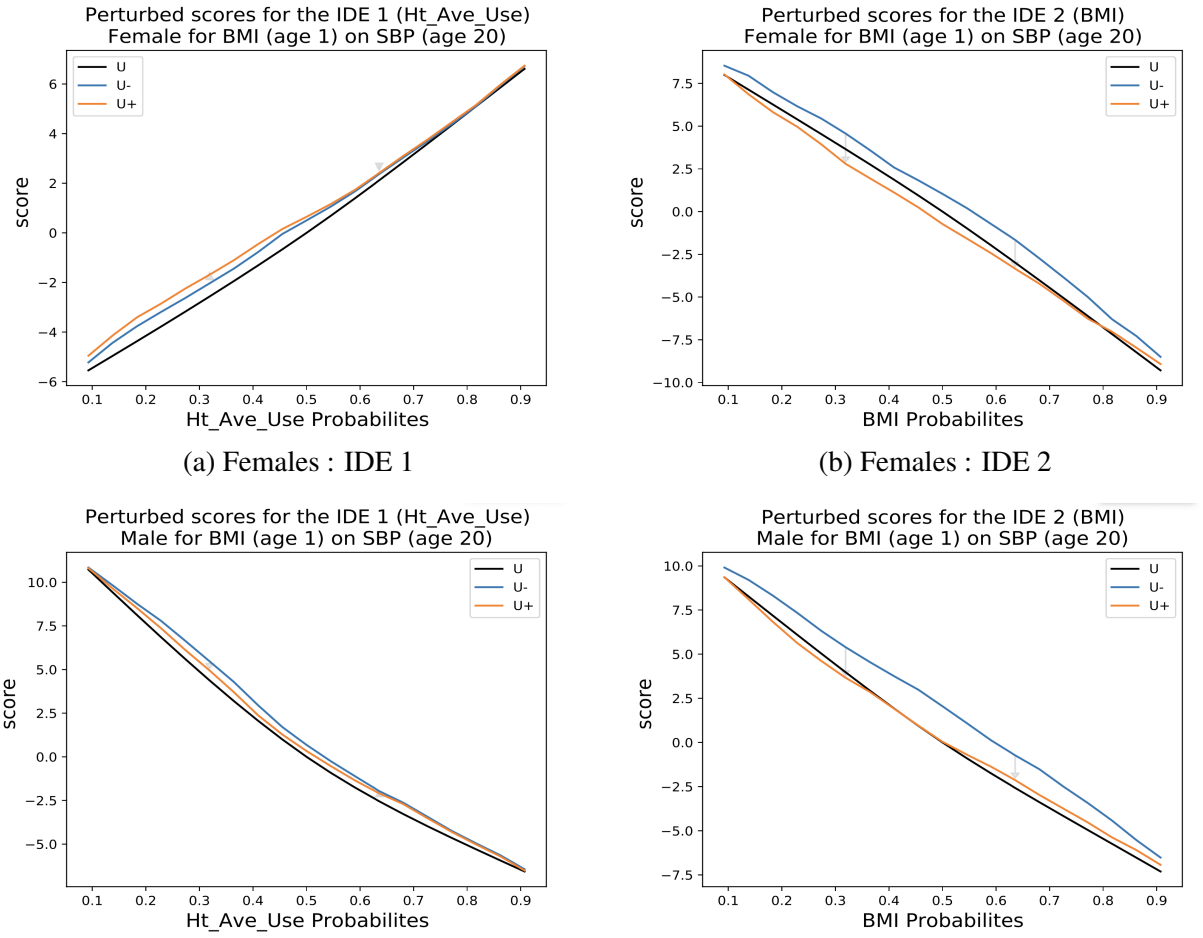
Note that $h_\theta(x) = \phi(\phi^{-1}(\cdot) + \theta)$ is an increasing function. Thus, $\boldsymbol{u}^+ > \boldsymbol{u}$ if and only if f_x is an increasing function. Similarly $\boldsymbol{u}^- \leq \boldsymbol{u}$ if and only if f_x is a decreasing function. We can see in figure 3.7a that whenever $\boldsymbol{u}$ is monotone, we observe this trend (this was true for all direct effect results, across all childhood body size measure, but we only provide limited examples here).

The same logic does not follow for the score functions $(\boldsymbol{u}^+, \boldsymbol{u}^-, \boldsymbol{u})$ for the mediators (which define the indirect effect). This is due to the extra setup involving the estimation of the marginal distribution of the perturbed mediation ($\tilde{M'}$). As a result, we no longer only require $\boldsymbol{u}$ to be a monotone function but also require $\tilde{M'}$ to be a monotone function of $\tilde{x}'(= p_x)$ in the same direction as $\boldsymbol{u}$. If both $\boldsymbol{u}^+$ and $\tilde{M'}$ are increasing then $\boldsymbol{u}^+ \geq \boldsymbol{u} \geq \boldsymbol{u}^-$, and vice versa. Looking at some examples, we can see that the perturbed scores for the indirect effect through adult HT (IDE1) in figures 3.9a and 3.9c do not sandwich $\boldsymbol{u}$ (the score for the observed data), even thought $\boldsymbol{u}$ is increasing. This is because M_1 is not an increasing function of x (childhood BMI). This can be seen from a marginal AFQR model with outcome adult HT, regressed on childhood BMI controlling for childhood and adult ages giving a model of the form

$$\mathrm{HT_2(p)} = \mathrm{u_1(BMI_1)v_1(p)} + \mathrm{u_2(age_1)v_2(p)} + \mathrm{u_3(age_2)v_3(p)}.$$

The AFQR models for both males and females confirmed that $\boldsymbol{u}_1$ is not an increasing function as HT does not track strongly to BMI.

Looking at the indirect effects through BMI (ref figures 3.9b, 3.9d), the perturbed $\boldsymbol{u}$'s sandwich $\boldsymbol{u}$ nicely for females, as adult BMI tracks strong with childhood BMI (ref figure 3.9b), i.e. adult BMI is an increasing function of childhood BMI. Since $\boldsymbol{u}$ is increasing, we get that $\boldsymbol{u}^+ \geq \boldsymbol{u} \geq \boldsymbol{u}^-$ as argued above.



(a) Females : IDE 1

(b) Females : IDE 2

(c) Males : IDE 1

(d) Males : IDE 2

Figure 3.9: Perturbed score vectors $(\boldsymbol{u}^+, \boldsymbol{u}^-)$ for the indirect effects through adult HT (IDE1) and adult BMI (IDE2) for exposure childhood BMI at age 1 on adult SBP at age 20, among females (first row) and males (second row). The score vector for the observed population is $\boldsymbol{u}$.

For males, we see a similar trend in the lower quantiles of audlt SBP but not in the upper quantiles (ref figure 3.9d). A possible explanation for this is that lower quantiles of adult BMI track childhood BMI more strongly but not the upper quantiles. This would mean that children who grow up to be above-median sized adults in terms of BMI have factors besides childhood body size influencing their adult BMI. A potential influencing factor could be migration to Bamoko. People who migrate to Bamoko tend to experience weight gain due to a change in lifestyle. To explore the extent of tracking, we fit a marginal AFQR model of the form

$$\text{adult BMI}(\text{BMI}_2) \sim \text{childhood BMI}(\text{BMI}_1) + \text{age}_1 + \text{age}_2.$$

The impact of BMI_1 on BMI_2 is presented in figure 3.10, and we can see that the tracking is

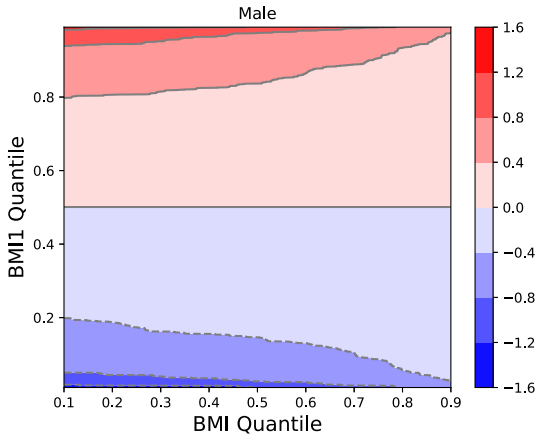


Figure 3.10: Marginal AFQR model for adult BMI regressed on childhood BMI and ages, to assess how BMI tracks from age 1 to age 20 along males.

stronger for all lower quantiles of adult SBP, but is weaker among the upper quantiles. While the upper quantiles of adult SBP still track positively, the indirect effects involve multiple non-linear transformations, due to which weak association between M_1' and X_1 may not propagate through to M_1' and X_1' leading to crossing of the perturbed $\boldsymbol{u}$ curves over observed $\boldsymbol{u}$ as seen in figure 3.9d.

3.5 Conclusion and Future Directions

In this chapter we proposed a novel framework for studying, defining and interpetting mediation effects through-out the distributions of the features involved. We do so by using a novel additive low rank model (AFQR) for the joint estimation of all conditional quantiles in the data generating models for the outcome given the mediators and exposure, and the mediators given the exposure; AFQR is a structured non-parametric method presented in chapter 2. We showed that the additive nature of AFQR yields simple expressions for the mediation effects, while the non-parametric nature of the functional components allows for flexible mediation estimates.

Our method for mediation analysis can accommodate continuous and time-varying variables (exposure, mediators and outcome) as well as multiple mediators. This was illustrated through a data application.

As part of future work, alternate hypothetical interventions can be explored. We propose one form for such an intervention (in equation 3.3), but other definitions could be constructed, especially based on the aim of the analysis. A general guideline for alternate interventions (using our intervention definition) is given below in table 3.3.

In the table above, $\tilde{x} \in [0, 1]$ denotes the quantile transformed observed exposure, and θ is the

Intervention	Impact of intervention
$\Phi(\Phi^{-1}(\tilde{x}) + \theta)$	uniformly boosts everyone by θ
$\Phi(\Phi^{-1}(\tilde{x}) + \lvert\frac{1}{2} - \tilde{x}\rvert\theta)$	boosts outer quantiles more and inner quantiles less
$\Phi(\Phi^{-1}(\tilde{x}) + (1 - \tilde{x})\theta)$	boosts lower quantiles more and upper quantiles less
$\Phi(\Phi^{-1}(\tilde{x}) + \tilde{x}\theta)$	boosts upper quantiles more and lower quantiles less

Table 3.3: Guideline for alternate interventions based on the desired impact of the hypothetical intervention. All interventions lead to perturbations in the same direction as the sign of θ.

parameter of the perturbation function. Also, currently the mediation effects are defined in terms of the rank one behaviour of the covariates, since AFQR provides an additive rank-1 representation. AFQR can potentially be extended to higher ranks, giving us mediation effects as a sum of differences over the rank of the model.

Immediate future work involves performing the inference procedure for the mediation effects, and reducing the discretization bias in the mediation effects. We believe the discretization bias may be reduced substantially by estimating the conditional quantile models in the mediation analysis via the regression-based AFQR, instead of evaluating them on a tensor of dimension 4. Thus, incorporating the regression-based AFQR in the mediation analysis estimation procedure is the immediate next task.

CHAPTER 4

Process Regression Data Emulator

4.1 Introduction

The majority of this thesis focuses on quantile regression, but in this work we also made use of another flexible regression approach based on Gaussian processes. The purpose of introducing this approach is to support statistical inference for the methods developed in other chapters of the thesis. Specifically, we take a data emulation approach to inference, mimicking what is done in the parametric bootstrap. Consider any of the techniques from chapters 2 or 3, where the goal is to estimate a quantity, say a direct or indirect effect in a mediation analysis, based on some mathematical processing of nonparametric quantile estimates. Statistical inference for these analyses is challenging for at least two reasons. First, inference for multi-step estimators is often difficult due to challenges in propagating uncertainty through the steps. In some cases analytic approaches to variance propagation are possible, but it is unclear how to do that in our setting. Second, as discussed in earlier chapters we are taking a marginal perspective to analyzing longitudinal data. We model the data at the level of observations, but the fact that there are dependencies among the observations cannot be ignored when assessing uncertainty.

The population of interest here is one in which a stochastic outcome $Y_i(t)$ varies continuously in time t for each observed unit i (t is a subject's age in our setting). In addition we have covariates $X_i(t)$ which determine the distribution of the $Y_i(t)$. The data for unit i consists of $Y_i(t)$ evaluated at finitely many time points $t_{i1}, \ldots, t_{in_i}$, and different subjects are observed at different time points t_{ij}, which are positive real numbers. Since $Y_i(t)$ is taken to be continuous in t, the dependence between $Y_i(s)$ and $Y_i(t)$ is strongly influenced by the magnitude of $s - t$, with the dependence becoming very strong as $s-t$ tends to zero. The main idea is that we would like the covariates X to determine the conditional mean of Y, the conditional variance of Y, and the degree of smoothness in the random functions $Y(t)$.

The model considered below has a Gaussian form with distinct specifications for the conditional mean and conditional variance/covariance structures. The mean structure is simply

$E[Y(t)|X = x] = b_\mu(t,x)'\beta$, where $b_\mu(t,x)$ consists of basis functions in time. The variance/covariance structure is harder to specify in a way that guarantees a valid (positive-definite) covariance model. This is resolved using methods introduced by [77] as discussed in detail below. We refer to the procedure here as "Gaussian Process Regression" (GPR) but note that it is somewhat different from the GPR used in machine learning, since we parameterize the family of covariance matrices and use likelihood-based approaches for fitting the models.

4.2 Structure of the model

The population model is expressed explicitly in terms of the finite-dimensional observations. It was shown in [77] that the finite-dimensional models discussed below are proper marginalizations of a common parent model for $Y(\cdot) \mid X(\cdot)$ for time t in $\mathcal{R}^+$. The model depends on four quantities that can be obtained from the covariates x_t and unknown parameters β_μ, β_λ, β_s, and β_u:

$$\begin{aligned}
&\text{Mean} && \mu(\boldsymbol{x}_t) = \beta_\mu' b_\mu(\boldsymbol{x}_t, t) \\
&\text{Scale} && \lambda(\boldsymbol{x}_t) = \exp(\beta_\lambda' b_\lambda(\boldsymbol{x}_t, t)) \\
&\text{Smoothing} && s(\boldsymbol{x}_t) = \exp(\beta_s' b_s(\boldsymbol{x}_t, t)) \\
&\text{Unexplained} && u(\boldsymbol{x}_t) = \exp(\beta_u' b_u(\boldsymbol{x}_t, t)).
\end{aligned} \tag{4.1}$$

Since this is a Gaussian model we can specify the mean and variance separately. The mean is simply $\mu(\cdot)$ as defined above. The covariance matrix of the finite dimensional observed vector $(Y(t_{k_1}), \ldots, Y(t_{k_d}))$ is given by

$$\Psi(k_1,k_2) = \lambda(\boldsymbol{z}_{t_{k_1}})\lambda(\boldsymbol{z}_{t_{k_2}}) \exp\left(\frac{-2(t_{k_1}-t_{k_2})^2}{s(\boldsymbol{z}_{t_{k_1}}) + s(\boldsymbol{z}_{t_{k_2}})}\right) \cdot \frac{s(\boldsymbol{z}_{t_{k_1}})^{-\frac{1}{4}} s(\boldsymbol{z}_{t_{k_2}})^{-\frac{1}{4}}}{\left[\frac{s(\boldsymbol{z}_{t_{k_1}}) + s(\boldsymbol{z}_{t_{k_2}})}{2}\right]^{1/2}} + \mathbf{1}_{k_1=k_2} u(\boldsymbol{z}_{t_{k_1}}). \tag{4.2}$$

If $u(\cdot) \equiv 0$ then the trajectories $Y(t)$ are Gaussian processes and in this case are also infinitely differentiable due to the specification of 4.2. For some variables like height it is reasonable to adopt such a model since the true value of height varies smoothly in time and there is little measurement error for this variable. For variables such as blood pressure, however, it is arguably essential to use a non-continuous model. While there may be a latent "true" blood pressure that varies smoothly in time, the observed blood pressure can deviate from this value due to occasion-specific factors. For that reason we have introduced the "white noise" or "unexplained" term determined by $u(\cdot)$. This is intended to capture, for example, independent occasion-specific measurement error or transient

perturbations of the true blood pressure.

4.3 Parameter estimation

We use a penalized maximum likelihood estimator, with a finite dimensional approximation to the squared second derivative as the penalty. Each of the four functional parameters $\mu(\cdot)$, $\lambda(\cdot)$, $s(\cdot)$, $u(\cdot)$ is subject to the penalty. At the population level, the penalty for, e.g. $\mu(\cdot)$ is $\int \mu''(t)^2 dt$. In practice, we define a grid $t_1^*, \ldots, t_q^*$ and penalize the finite dimensional vectors $\mu(t^*) \equiv (\mu(t_1^*), \ldots, \mu(t_d^*))^T$. The penalty has the structure of a quadratic form $\mu(t^*)^T G \mu(t^*)$, where G is a $d \times d$ positive semi-definite matrix. Specifically, if F is a $d-2 \times d$ matrix with i^{th} row equal to the discretized second difference operator with values $(1, -2, 1)$ at positions i, $i+1$, and $i+2$, and zeros elsewhere, then $G = F'F$. In practice we make use of a hybrid penalty $G_\eta = \eta G + (1-\eta) I_{d \times d}$ which combines the functional smoothing penalty and a ridge-type penalty.

The model parameters β_μ, β_λ, β_s, and β_u are estimated using penalized maximum likelihood estimation. Gradients of the log-likelihood and penalty terms can be analytically derived, permitting conjugate gradient-type algorithms to be used for likelihood optimization.

4.4 Data emulation and inference

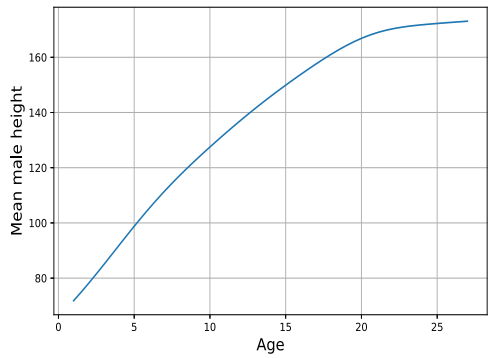
As noted above, the purpose of introducing this model is to support statistical inference for the procedures developed elsewhere in this thesis. The approach is analogous to parametric bootstrapping. Here we focus on the setting of mediation analysis (and the corresponding AFQR models) where we have three time-varying quantities: the exposure $X(t)$, two mediators $M_1(t)$, $M_2(t)$, and the outcome $Y(t)$. Concretely, in our application $X(t)$ and $M_1(t)$ are both BMI, $M_2(t)$ is height, and $Y(t)$ is systolic blood pressure. Thus there are three variables to consider, and the time variable t is age. Note that $\text{BMI} = 10000 \times \text{WT}/\text{HT}^2$, and so we always only have three independent variables in our application, regardless of which feature we consider for measuring childhood body size. The goal is to emulate m independent copies of $[(X(t), M_2(t), Y(t)); t = t_{i_1}, \ldots, t_{i_{d_i}}]_{i=1}^n$, evaluated at the same time points as the observed data.

To emulate the data, we use a factorized form of the joint distribution of X, M, and Y: $P(X, M, Y) = P(X) \cdot P(M|X) \cdot P(Y|X, M)$. Each of the three factors in the factored joint distribution is estimated using the GPR technique discussed above using all observed data. We then simulate $X_i^{(j)} \sim \hat{P}(X)$, $M_{2i}^{(j)} \sim \hat{P}(M_2|X)$, and $Y_i^{(j)} \sim \hat{P}(Y|X, M)$, where i indexes people and j indexes emulated data sets. Finally, for each emulated data set we estimate the AFQR functional parameters $\{f_j, g_j\}_{j=1}^J$, then obtain point-wise standard deviations for these estimates that can be interpreted as standard errors of the effects of interest.

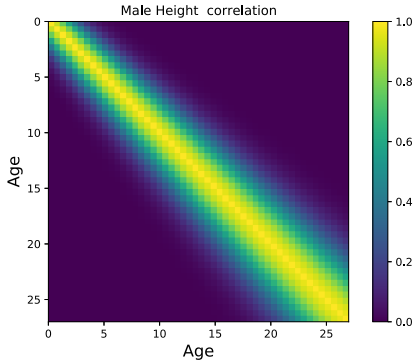
4.4.1 Illustration using the Dogon Longitudinal Study

Here we present the estimated parameters used to emulate datasets mimicking the Dogon Logitudinal Study, for the variables height (HT), weight (WT), and systolic blood pressure (SBP). The simulation proceeds by factoring the joint distributions, first sampling HT, then sampling WT given HT, then sampling SBP given HT and WT. All conditional models also take account of age. The four model parameter vectors β_μ, β_λ, β_s, and β_u inform the mean and variance functions in the model as given in equation 4.1.

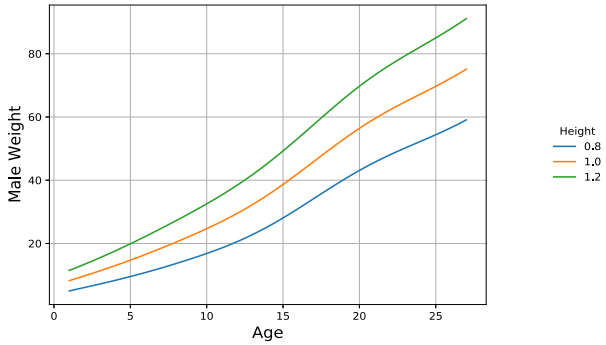
Figure 4.1 presents the fitted mean functions for the male subpopulation, and the temporal correlation matrices for the unexplained variation in these variables. The parameters correspond to a spline basis of age, with interactions used to capture conditional relationships among the variables. In particular, β_u consists of a basis of dimensions seven, β_λ, β_s a basis of dimension two and the unexplained variance parameter β_u is modeled using a basis of dimension one. The full specification of the parameter basis and the fitted regression coefficients for this example are provided in the appendix B, along with plots of simulated data.



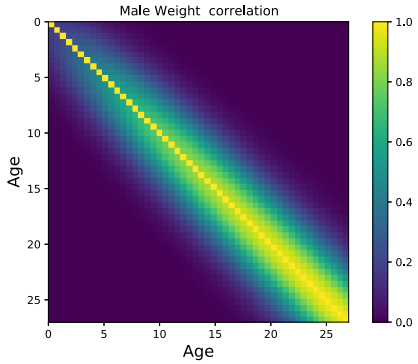
(a) HT|Age : Estimated Mean



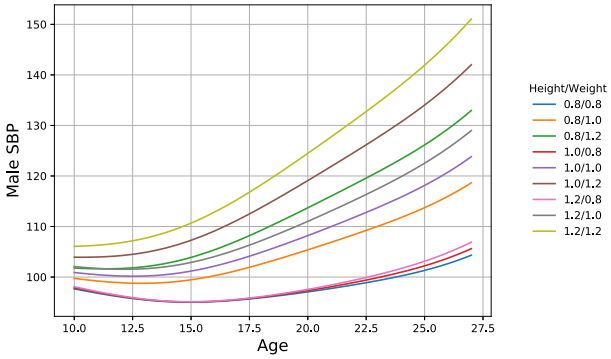
(b) HT|Age : Estimated Covariance



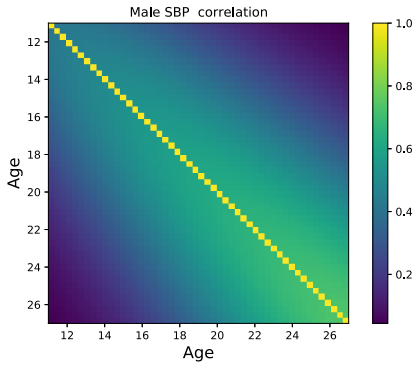
(c) WT|(HT, Age) : Estimated Mean



(d) WT|(HT, Age) : Estimated Covariance



(e) SBP : Mean



(f) SBP : Estimated Covariance

Figure 4.1: Fitted conditional mean functions for height (HT), weight (WT), and systolic blood pressure (SBP) for males in the Dogon Longitudinal Study, estimated using the GPR procedure. The conditional mean functions are modelled using a spline basis of age.

CHAPTER 5

Conditional Covariance Estimation for Multivariate Functional Data

5.1 Introduction

The methods developed here aim to understand the properties of a vector-valued random function of time denoted $\mathbf{X}$. For a time point t, $\mathbf{X}(t) \in \mathcal{R}^p$ is a vector containing the quantitative states of several attributes at time t. The primary focus is on conditional correlations involving specific attributes and time points in $\mathbf{X}$, as discussed in detail below. The data have the form $\{\mathbf{Y}_{ij}\}$, where $\mathbf{Y}_{ij} = \mathbf{X}_i(T_{ij}) + \epsilon_{ij}$, where $i = 1, 2, .., n$ indexes subjects, T_{ij} is the j^{th} observation time point for the i^{th} individual, $\mathbf{X}_i$ is the realization of $\mathbf{X}$ for the i^{th} individual, and the ϵ_{ij} are mean zero homoscedastic errors.

This setting has connections to important branches of statistics including longitudinal data analysis, functional data analysis and stochastic process modelling. Our work focuses on semi-parametric and non-parametric estimation of conditional correlations involving $\mathbf{Y}$, which has not received much attention in the literature. These covariance estimators are worth investigating as they provide a means for assessing correlations of high relevance for science and policy while avoiding strong modeling assumptions.

This report is organized as follows: in Section 2, we review local polynomial regression for mean estimation in functional data and provide the details for applying this method to marginal covariance estimation. Section 3 explains the plug-in inference procedure for the marginal covariance estimates. Two methods to model conditional covariances and correlations are introduced in Section 4, along with the inference procedures for both these methods. In Section 5, we evaluate the performance of one of these techniques on simulated data, and finally conclude with Section 6 which includes the summary, limitations and future work.

5.1.1 Motivating Application

Our motivation stems from fundamental questions in human biology relating to the interplay between growth during childhood and blood pressure in adulthood; specifically, $\mathbf{X}(t) \in \mathbb{R}^2$ represents body size (height or weight) and blood pressure of individual i measured at age t. It is well-established that blood pressure and body size are correlated within subjects when measured at a single time point, for instance, people who are overweight tend to have substantially higher blood pressure than people of healthy weight, and to a lesser extent, people who are tall tend to have higher blood pressure compared to people of average height. However, more interesting is the role of life history as a possible driver of adult blood pressure, which is just beginning to be explored. In this setting, we can ask whether growth and body size in childhood are correlated with adult blood pressure when controlling for adult body size. This can be formalized as the conditional correlation

$$\text{Cor}(\text{HT}(s), \text{SBP}(t) | \text{HT}(t)), \tag{5.1}$$

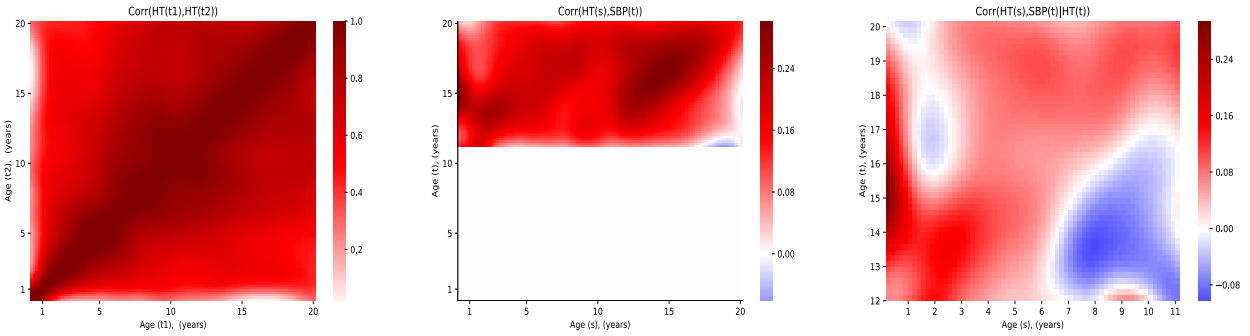
where height (HT) and systolic blood pressure (SBP) are measured at time points $s < t \in \mathbb{R}$. This correlation helps us single out the possible direct impact of undernourishment during early childhood on adult blood pressure. Undernourishment remains common in the developing world, and we will be illustrating our methods using data from a longitudinal study of around 1800 children born in a rural part of Mali, a resource-poor country in Africa.

We would like to draw attention to the importance of conditioning here. It is well-established that weight and height "track" within subjects. Due to this autocorrelation, taller children tend to become taller adults and heavier children tend to become heavier adults. This can be seen in the second moment structure of the data, in that we anticipate $\text{Cor}(\text{HT}(s), \text{HT}(t)) \gg 0$ for ages $s < t \in \mathbb{R}$ (years). As mentioned above, $\text{Cor}(\text{HT}(t), \text{SBP}(t)) \gg 0$ is an anticipated relationship that is commonly referred to as *simultaneous correlation*. By this logic, it may follow that $\text{Cor}(\text{HT}(s), \text{SBP}(t)) > 0$ for $s < t$ years, i.e. being smaller (physically) in childhood is good if the goal is to moderate adult blood pressure. Our goal is to isolate the "direct effect" of childhood growth and nutrition, net of any impacts of tracking. This would capture physiological phenomena such as children with poor childhood nutrition having underdeveloped organs (especially kidney and heart) that may place them at risk for elevated adult blood pressure.

In the literature in this area, there is discussion of a "reversal paradox", in which the relationship between childhood height and adult blood pressure flips from a positive to a negative association when conditioning on adult height. The methods proposed and developed here can be used to understand both the direct and total effects of childhood undernutrition in relation to adult blood pressure, and can more fully explain why the reversal "paradox" arises.

5.1.2 Flexible covariance estimation

The work in progress presented here aims to develop new methodology for estimation and inference of conditional correlation functions when the available data are observed at finite and irregular time points for a collection of independent subjects. The data are conceptualized as a finite collection of observations made on a collection of random continuous functions of time (with each subject being one realization from the distribution of random functions). Since we are interested in multiple attributes (e.g. height and blood pressure), these functions are multivariate at each time point, in addition to being repeated observations over time. Finally, it may happen that at some time points, some but not all of the attributes are observed. This is referred to as "missing data" below.



(a) Cor$(HT(t_1), HT(t_2))$ (b) Cor$(HT(s), SBP(t))$ (c) Cor$(HT(s), SBP(t)|HT(t))$

Figure 5.1: Data exploration: plots L-R are to motivate tracking, simultaneous correlation and finally reversal in the conditional correlation of interest.

The covariance function $\mathrm{Cov}(s,t) \in \mathcal{R}^{p\times p}$ of a random vector-valued function with p attributes is a complex object and it is desirable in practice to be able to adapt to many different possible population structures. Moreover, it is desirable to also accommodate various forms of data collection. In particular, $\mathbf{X}(t)$ will never be observed in entirety. Rather, we will observe $\mathbf{X}(t)$ at a finite set of time points which may vary between subjects, both in terms of the number and locations of the time points. Furthermore, what we observe is a noisy version of $\mathbf{X}(t)$.

Much existing work on covariance function estimation exploits structural constraints. Our goal is to estimate the covariance function without any such constraints. Here we briefly review some of the more common structural constraints and comment on why they would not be likely to hold in our application of interest.

- *Homoscedasticity:* A covariance function is homoscedastic if $\mathrm{diag}(\mathrm{Cov}(t,t))$ does not depend on t. This condition states that the marginal variance of each attribute is the same at

each time point. Growth data do not normally satisfy this condition, and instead often exhibit a mean/variance relationship in which the variance is greater when the mean is greater.

- *Stationarity:* A covariance function is stationary if $\mathrm{Cov}(s,t) \equiv \mathrm{Cov}(s-t,0)$. This means that the covariance between $\mathbf{X}(s)$ and $\mathbf{X}(t)$ depends only on the time displacement $t-s$. This is very unlikely to hold in our setting, since growth of very young children, say of ages 2-4, is quite unlike growth of older children, say of ages 8-11.

- *Symmetry:* A covariance function is symmetric if

$$\mathrm{Cov}(\mathbf{X}(s),\mathbf{X}(t))_{jk} = \mathrm{Cov}(\mathbf{X}(t),\mathbf{X}(s))_{jk}.$$

  In our setting, this would imply, for example, that the correlation between the height of younger person and the blood pressure of an older person would be equal to the correlation between the blood pressure of a younger person and the height of an older person.

- *Separability:* A covariance function in our setting is separable if $\mathrm{Cov}(s,t) = c(s,t)\cdot A$, where A is a fixed $p \times p$ matrix defining the covariances among the p attributes, and $c(s,t) : \mathcal{R}^2 \to \mathcal{R}$ is a continuous function defining the role of time. This is unlikely to hold in our setting, since it implies that the conditional correlation of interest is always zero (see section 5.1.6 for details).

With this in mind, we develop covariance and conditional covariance estimators that can capture a variety of smooth covariance functions, using irregularly sampled data with missingness.

5.1.3 Introduction to Methodology

Our approach can be seen as a form of the method of moments in that it matches empirical moments of the observed data to the unknown second moment structure of the process of interest. However due to the irregularity of the observation times, there is no evident low-dimensional summary statistic that can be used to construct a small number of moment equations to solve. Within-individual products of observed values, for example, Y_{ij}^2 and $Y_{ij}Y_{ik}$, are informative about the covariance function evaluated at specific points, here, (T_{ij},T_{ij}) and (T_{ij},T_{ik}) respectively. We use local polynomial regression to estimate the marginal variances and covariances as continuous functions of time. Local polynomial regression is widely used for estimating mean functions, but is much less commonly used for estimating covariances and correlations. Here we use local polynomial regression to integrate the information in a large collection of products and squares of data to obtain a coherent estimate of the underlying covariance function.

Local regression for covariance estimation

Let us start by building some intuition for variance and covariance estimation using local polynomial smoothing in univariate data. Covariance estimation in data with multiple attributes, arbitrary missingness and subject-specific time points, as we are interested in modelling, is harder to visualize and comes with additional complexity.

We start with the simplest case of estimating the variance function for univariate Gaussian data with mean zero. The data were simulated from a population with heteroscedastic variance given by $\mathrm{Var}[Y(t)] = (1+t)^2$. Figure 5.2 shows a scatterplot of the squared data $Y(t)^2$ against time t. Local linear regression was used to estimate $E[Y(t)^2] = \mathrm{Var}[Y(t)]$, using the `loess` function in `R`. Note that the values of $Y(t)^2$ are skewed as they follow scaled χ^2 distributions. The figure shows that the local linear estimate of the variance function lies quite close to the true variance function.

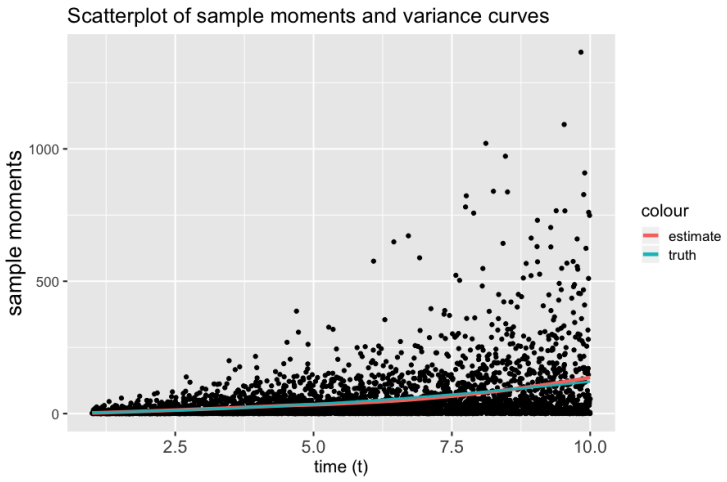
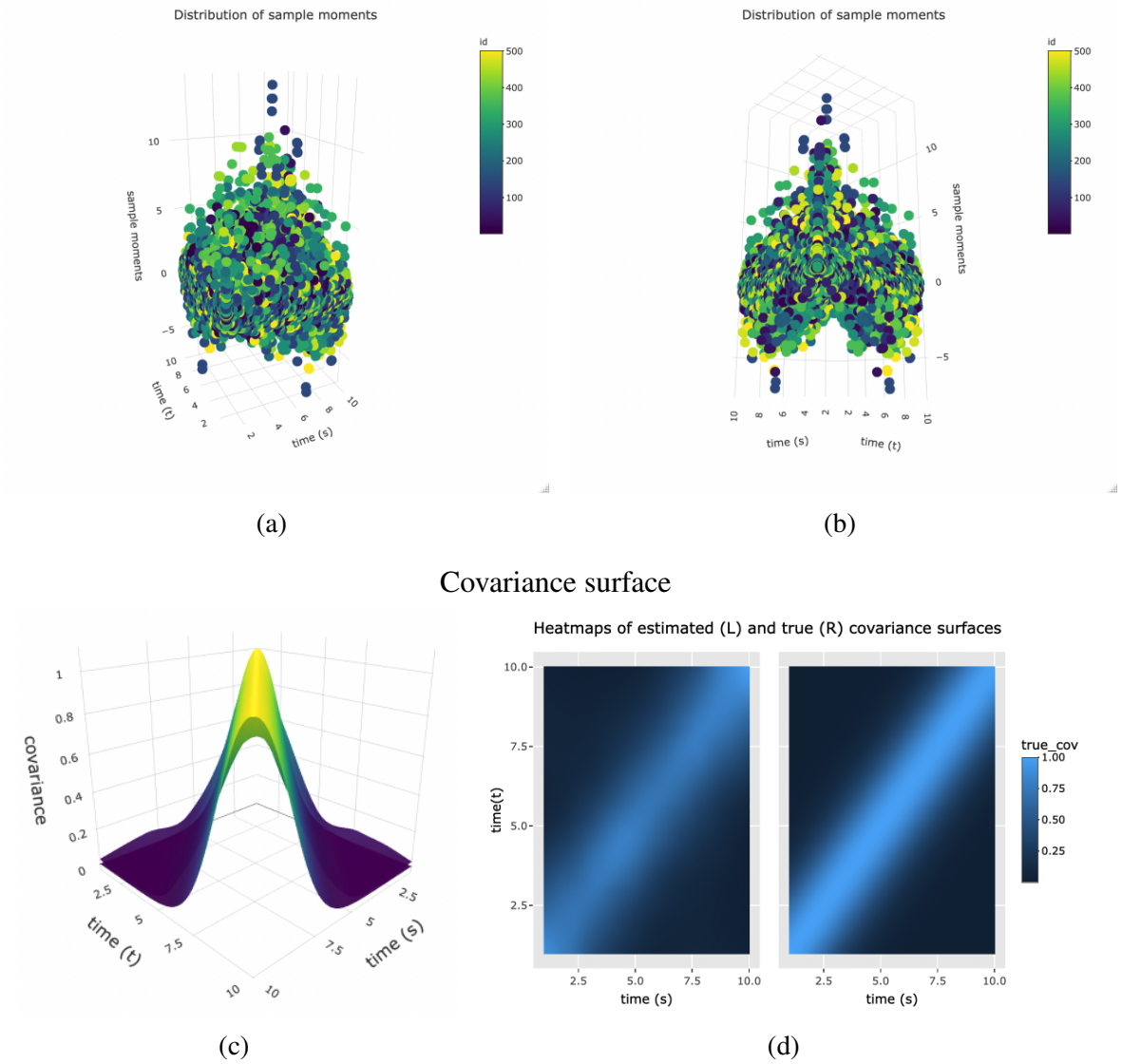


Figure 5.2: Variance estimation using local linear smoothing

Going one step further, figure 5.3 illustrates covariance estimation in stationary univariate data. The simulated dataset consists of 500 iid copies of a mean 0 stationary function with covariance $\mathrm{Cov}(Y(s), Y(t)) = \exp(s-t)^2/5$, each being observed at 10 distinct time points normalized to lie in $[1, 10]$. Estimating the covariance surface is equivalent to estimating the second moments of the underlying function. Thus, local linear regression now considers all 500×10^2 pair-wise products and smooths over these points (in 3D space) to find the mean surface, which estimates the covariance function. This becomes much harder to visualize directly, as can be seen in figures 5.3a and 5.3b. Subfigures 5.3c and 5.3d show that the method is able to recover the underlying covariance surface.



(a) (b)

(c) (d)

Figure 5.3: Local linear covariance estimation in univariate stationary data

When estimating covariance functions of multivariate longitudinal data such as ours, the 'product data' now consists of all within-individual between time and between attribute pairwise products of the original data. Thus, the sample size for the local regression is much larger than the sample size of measured data points. Combined with the well-known computational challenges of large-scale local regression, this can lead to a very demanding computation. We also note that the within-individual products of observed data values are unbiased for the underlying variances and covariances, but have high variance, and may also exhibit skewed distributions and complex patterns of interdependence. These factors do not produce bias in the local regression estimates, but they do make it difficult to achieve high statistical efficiency, and to conduct rigorous statistical inference for the resulting estimates.

There are some theoretical results on convergence rates and the asymptotic properties of marginal covariance estimates in similar settings, e.g. [102]. However, we are not aware of any work focusing specifically on the computational aspects of this approach. Also, we are not aware of prior work examining the estimation of conditional covariances and correlations. The limited existing work is summarized in section 5.1.4 below.

5.1.4 Literature review

This section includes some relevant work on covariance/ correlation estimation using moment-based kernel regression in functional data. This is by no means an exhaustive literature review. As far as we are aware, there is no prior work on estimating the conditional covariance function in longitudinal data.

[103] estimate the Pearson correlation for paired longitudinal data (that is, bivatiate data wherein both variables are observed at the same subject-specific time points) and provide asymptotic results for the same. They model the Pearson correlation between the two random functions at the same time point by estimating the marginal variances and cross-covariances at the same time point using local linear regression. This means that their kernel weights are univariate, corresponding to one-dimensional smoothing. We are interested in modelling cross-time covariances and correlations, which involves two-dimensional smoothing.

A non-parametric estimator of the covariance function in multivariate longitudinal data using linear smoothing is established in [62]. They model stationary covariances and impose a symmetry constraint. Under this setting, they provide asymptotic results for both separable and non-separable cases, however, their simulation and data application explore a seperable covariance structure.

Marginal covariance estimation for longitudinal data is often studied as part of functional principal component analysis (FPCA). FPCA is a common tool used to analyze functional data which involves modelling the mean and covariance functions of the random process, before performing PCA on the data. Many have used local polynomial regression for covariance estimation in FPCA, including [98, 37, 61]. Both [98] and [61] estimate the covariance surface of non-gridded unidimensional functional data using local linear regression. These papers provide theoretical results on the covariance estimates and consequently the functional principal components. While we are interested in modelling the marginal covariance function, our ultimate aim, as mentioned in section 5.1.1 is to extend this to model the conditional covariance and correlation function (across attribute across time), and to perform inference on these estimates.

Theoretical findings on covariance estimation using local regression have been established in the literature. Asymptotic results and convergence rates for both mean and covariance estimation in noisy functional data were studied by [61, 97, 102], among others. [97] provide asymptotic

normality results for one and two-dimensional smoothers for noisy functional data. They also apply these to the Nadaraya-Watson and local linear estimators to give expressions for their asymptotic means and variances (respectively). [102] improve upon their results by increasing the convergence rates for the asymptotic estimates. They derive the theoretical expressions for asymptotic bias and variance of covariance estimates in uni-dimensional data common to sparse, dense and ultra-dense data. Under the sparse setting, which is what we're interested in, both mean and variance estimation attain the classical non-parametric rates for univariate and bivariate functions respectively.

The convergence rates for both mean and covariance estimation improve as the degree of smoothness increases, as explained in [61]. However, we are not aware of convergence rates for multidimensional data.

5.1.5 Alternative approaches

In this section we discuss some of the existing techniques for estimating covariance functions. This is a brief survey of such techniques and is by no means an exhaustive review.

Parametric covariance models

Parametric covariance estimation can proceed by specifying a model $\Sigma_\theta(s,t) : \mathcal{R}^2 \to \mathcal{R}^{p\times p}$ for the covariance between observations taken on the same individual at times s and t, in terms of a finite dimensional parameter $\theta \in \Theta \subset \mathcal{R}^q$. Relatively few such forms exist since it is difficult to parametrize a family of functions that are positive semidefinite (PSD). Familiar covariance models of this type include the first order autoregressive model $\Sigma_\theta(s,t) = \theta^{|s-t|}$, and the exchangeable covariance model $\Sigma_\theta(s,t) = \theta$, both for $0 \leq \theta \leq 1$. A more flexible and realistic parametric model is the Matérn covariance. A non-stationary covariance function for data indexed by multiple variables was provided by [76]. All of these parametric covariance models are finite dimensional and are unable to capture any covariance structure that cannot be expressed in the specified parametric form.

If the data are Gaussian and have zero mean, then maximum likelihood estimation can be used to estimate θ based on data observed from one or multiple independent replicates of the process, each observed at a finite set of time points. Non-Gaussian likelihood-based analysis is sometimes possible but presents additional difficulties.

Kronecker products and sums

We are interested in covariance models for data with multiple attributes, indexed by time. This opens up the possibility of using Kronecker products and Kronecker sums to create large families of covariance functions from simple building blocks. The Kronecker product has the form

$\mathrm{Cov}(\mathbf{X}(s), \mathbf{X}(t)) = c(s,t) \cdot F$, where $c(\cdot,\cdot) : \mathcal{R}^2 \to \mathcal{R}$ is a covariance function and F is a $p \times p$ covariance matrix. Kronecker products are said to be *separable* with respect to the covariance for time and the covariance among attributes. The Kronecker product model turns out to be too strong for our purposes, since it implies that the conditional covariances of interest are always zero under Gaussianity.

Kronecker sums generalize Kronecker products and are potentially more useful to us. In the Kronecker sum model, $\mathrm{Cov}(\mathbf{X}(s), \mathbf{X}(t)) = \sum_{k=1}^{K} c_k(s,t) \cdot F_k$, where $c_k(\cdot,\cdot) : \mathcal{R}^2 \to \mathcal{R}$ are covariance functions and $F_k \in \mathcal{R}^{p \times p}$ are covariance matrices. In general if $K > 1$ the conditional covariances of a Kronecker sum are not necessarily equal to zero.

Multilevel modeling

A different type of parametric covariance model can be obtained using random effects. This is a large topic, and we will only discuss it here to the extent that it is related to the approaches developed below. Suppose that the data for the i^{th} replicate are generated as $Z_i\gamma_i + \epsilon_i$, where Z_i is a fixed $m \times q$ matrix, $\gamma_i \sim [0, \Psi]$ is a random q-vector, and $\epsilon_i \sim [0_q, \sigma^2 I_{q\times q}]$ is a random m-vector. Then the covariance matrix of our data is $Z_i \Psi Z_i' + \sigma^2 I_{m\times m}$. This is a finite dimensional parametric model with parameter $\theta = (\Psi, \sigma^2)$. A common choice for Z_i would be $Z_i(j,k) = t_{ij}^{k-1}$, where t_{ij} is the time at which observation j was made for subject i. Thus, the value for the trait of interest is a random polynomial of the form $\sum_{j=1}^{q} \gamma_j t^{j-1} + \epsilon$. This longitudinal multilevel model does not require the data to be observed at a common set of time points for all subjects.

Multilevel modeling is usually fit using maximum likelihood (or restricted maximum likelihood), which is efficient if the model is correctly-specified but may perform poorly otherwise. Also, the likelihood function for a multilevel model can be strongly non-convex, nearly flat, and multi-modal. An earlier approach to estimating multilevel models called MINQUE (minimum quadratic unbiased estimation) [84] is more similar in spirit to the approach that we develop below. MINQUE is based on the fact that the product of any two observations for the same subject has an expected value that is a linear combination of the model parameters. Equating all observed products to their expectations yields an overdetermined system of equations that can be solved using least squares.

Efficient moment-based approaches

Generalized Estimating Equations (GEE) is an approach for estimating mean parameters based on efficient estimating equations that involve moments of the data. GEE can estimate the mean structure in the presence of arbitrary covariance, but does not estimate the covariance structure itself efficiently. An extension to GEE known as GEE2 [63] estimates both the mean and covariance structures efficiently using efficient estimating equations. Another approach of this

type is the approach for joint estimation of mean and covariance parameters based on the quadratic exponential quasi-likelihood, due to [83]. These approaches are notable since they indicate how moments can be used to efficiently estimate covariance structures. However it is not clear how to generalize any of these approaches to the problem of estimating covariance functions, which is our goal here.

Non-parametric techniques

Kernel regression is often applied to model the covariance function in longitudinal data. Section 5.1.4 talks about previous work relating to this technique.

An alternate set of techniques model matrix decompositions instead of the covariance matrix itself. For instance, instead of estimating the covariance matrix directly, [45] propose estimating its modified cholesky decomposition for univariate data. This diagonalizes the covariance matrix with a lower triangular matrix with unit diagonal entries. The elements of both these matrices (the diagonal and lower triangular matrices) can be estimated as they are associated with a varying-coefficients regression model [80]. The estimated covariance matrix thus obtained is guaranteed to be PSD. [11] generalize this to incorporate non-regular time points on multidimensional data. They extend the varying-coefficients linear model corresponding to the Cholesky decomposition and use smoothing splines to estimate the parameters in the regression by normal maximum likelihood. The extended linear model involves regressing each observation per individual on all previous observations (for that individual) with universal (non-subject specific) coefficients given by the autoregressive function evaluated at the two time-points. We do not believe this applies to our setting, since the number of times an individual is sampled prior to a specific time point, say t_0, should not relate to their measurements at t_0. That is, $\mathbf{Y}_{ij} = \begin{bmatrix}\text{HT}_i(T_{ij}) & \text{SBP}_i(T_{ij})\end{bmatrix}$ should not depend on j and should only depend on the value of T_{ij}.

5.1.6 Separability and conditional covariance

We show here that separable covariance functions have diagonal conditional covariances, and thus are unsuitable for our goal which is to flexibly model the conditional covariance and conditional correlation functions. As per our motivating application (5.1), we want to model the conditional correlation

$$\text{Cor}(X_1(s), X_2(t)|X_1(t)) = \frac{\text{Cov}(X_1(s), X_2(t)|X_1(t))}{[\text{Var}(X_1(s)|X_1(t)) \cdot \text{Var}(X_2(t)|X_1(t))]^{1/2}}. \tag{5.2}$$

where $\mathbf{X}_k(t)$ is the random function for attribute k evaluated at time t. A subtle clarification in notation is the distinction between $\mathbf{X}_i$, the realization of $\mathbf{X}$ for the i^{th} individual, and X_i, which refers

to the i^{th} component of the random function $\mathbf{X}$. Thus, $\mathbf{X}(T_{ij}) = \Big[X_1(T_{ij}) \quad X_2(T_{ij}) \cdots X_p(T_{ij})\Big]^T$.

For a mean-zero p-dimensional process $\mathbf{X}(t) = \{[X_1(t) \cdots X_p(t)]^T : t \in \mathbb{R}^p\}$, a separable covariance function leads to the conditional correlation of interest (5.2) always being zero, at any set of time points $s, t \in \mathbb{R}$. We show this result by first deriving an expression for the correlation in (5.2) when the covariance function can be expressed as a Kronecker sum. Since a separable covariance is representable as a Kronecker product, and a Kronecker product is a special case of a Kronecker sum, the result will follow directly.

Suppose we observe a 2-dimensional random function at n points to obtain a random vector

$$Z \equiv \big(Z_1(t_1), Z_2(t_1), \cdots, Z_1(t_n), Z_2(t_n)\big) = (z_{11}, z_{12}, \cdots, z_{n1}, z_{n2}) \in \mathcal{R}^{2n}.$$

We express the covariance matrix of Z as the Kronecker sum

$$\sum_{k=1}^{K} A_k \otimes B_k,$$

where B_k is a 2×2 exchangeable matrix with correlation $\theta^{(k)}$and A_k is a positive semidefinite (PSD) matrix with parameter $\eta^{(k)}$, i.e.

$$B_k = \begin{bmatrix} 1 & \theta^{(k)} \\ \theta^{(k)} & 1 \end{bmatrix}; \quad A_k(l_1, l_2) = \eta^{(k)}(l_1, l_2).$$

For instance, if A_k is stationary, then $A_k(l_1, l_2) = \eta^{(k)}_{|l_1 - l_2|}$. Then, for Gaussian data,

$$\begin{aligned} \text{Cov}(Z_1(s), Z_2(t)|Z_1(t)) &= \text{Cov}\Big(Z_1(s), Z_2(t)\Big) - \text{Cov}\Big(Z_1(s), Z_1(t)\Big) \cdot \text{Cov}\Big(Z_2(t), Z_1(t)\Big)/\text{Var}\Big(Z_1(t)\Big) \\ &= \sum_k \theta^{(k)}\eta^{(k)}(s,t) - \sum_k \theta^{(k)} \cdot \sum_k \eta^{(k)}(s,t)/K \end{aligned}$$

Define $u = [\eta^{(k)}(s,t); k = 1, 2, \cdots, K]^T$, $v = [\theta^{(k)}; k = 1, 2, \cdots, K]^T$ and $u_\cdot = \mathbf{1}^T u, v_\cdot = \mathbf{1}^T v$. Then, the relevant relative conditional variances are

$$\begin{aligned} \text{Cov}(Z_1(s), Z_2(t)|Z_1(t)) &= \sum_k \theta^{(k)}\eta^{(k)}(s,t) - \sum_k \theta^{(k)} \sum_k \eta^{(k)}(s,t)/K = u'v - u_\cdot v_\cdot/K, \\ \text{Var}(Z_2(t)|Z_1(t)) &= K - \big(\sum_k \theta^{(k)}\big)^2/K = K - u_\cdot^2/K, \\ \text{Var}(Z_1(s)|Z_1(t)) &= K - \big(\sum_k \eta^{(k)}(s,t)\big)^2/K = K - v_\cdot^2/K. \end{aligned}$$

Hence, with $-1 \leq u_k, v_k \leq 1 \ \forall k$, the conditional correlation of interest as in (5.2),

$$\text{Cor}(Z_1(s), Z_2(t)|Z_1(t)) = \frac{u'v - u_\cdot v_\cdot / K}{[(K - u_\cdot^2/K) \cdot (K - v_\cdot^2/K)]^{1/2}}. \tag{5.3}$$

As we can see, a separable covariance function implies $K = 1$ which would make the conditional correlation zero (since $u_\cdot = u, v_\cdot = v$ and $u'v = uv$). For non-zero conditional correlation, we can optimize the expression in (5.3) over all feasible u and v, with the only constraints being $-1 \leq u_k \leq 1$ and $-1 \leq v_k \leq 1 \ \forall k$. More details in Section 5.5.

5.2 Local Polynomial Regression for Covariance Estimation

In this section we introduce local polynomial regression applied to functional data. We start by reviewing mean estimation in functional data and then introduce the more challenging problem of covariance estimation.

In local polynomial regression, weighted least squares (WLS) is used to fit a polynomial mean structure model locally at each point in the domain. This allows familiar polynomial regression methods to be used to estimate non-linear regression functions. We start by presenting the model for mean estimation in functional data. Let $\{\mathbf{X}(t) : t \in \mathbb{R}\}$ be a p-variate random function with mean function $\boldsymbol{\mu}(t) = \mathbb{E}(\mathbf{X}(t)) \in \mathbb{R}^p$ and covariance function $\Sigma(s,t) = \text{Cov}(\mathbf{X}(s), \mathbf{X}(t)) \in \mathbb{R}^{p \times p}$ for time points $s, t \in \mathbb{R}$. The p components of $\mathbf{X}(t)$ correspond to the p measured attributes of interest. We can then decompose $\mathbf{X}(t)$ as follows:

$$\mathbf{X}(t) = \boldsymbol{\mu}(t) + \mathbf{U}(t)$$

where $\boldsymbol{\mu}(t)$ is a deterministic function and $\mathbf{U}(t)$ is a centered random function, i.e. $\mathbb{E}(\mathbf{U}(t)) = 0 \ \forall t \in \mathbb{R}$, which also has covariance function $\text{Cov}(\mathbf{U}(s), \mathbf{U}(t)) = \Sigma(s,t)$.

If our dataset has n individuals, then each subject has a realization of the underlying random function $\mathbf{X}(t)$, i.e. for individual $i = 1, 2, \cdots, n$,

$$\mathbf{X}_i(t) = \boldsymbol{\mu}(t) + \mathbf{U}_i(t). \tag{5.4}$$

However, since the entire random function $\mathbf{X}(t)$ is not directly observable, for each individual i, we observe $\mathbf{X}_i(t)$ at a finite number of subject-specific time points. Further, these observations are contaminated with some additive measurement errors ϵ. In particular, if there are n iid individuals, with the i^{th} individual having n_i number of observations at time points $\{T_{i1}, T_{i2}, \cdots, T_{in_i}\}$, we observe $\{\mathbf{Y}_{ij}\}_{j=1}^{n_i}$, which is a noisy version of $\mathbf{X}_i(t)$ at these n_i distinct points. Here, we can

decompose $\mathbf{Y}_{ij}$ as follows:

$$\mathbf{Y}_{ij} = \mathbf{X}_i(T_{ij}) + \boldsymbol{\epsilon}_{ij} = \boldsymbol{\mu}(T_{ij}) + \mathbf{U}_i(T_{ij}) + \boldsymbol{\epsilon}_{ij}, \tag{5.5}$$

where $\boldsymbol{\epsilon}_i = [\boldsymbol{\epsilon}_{ij}]_{j=1}^{n_i}$, with $\boldsymbol{\epsilon}_{ij}$ are iid copies of measurement error $\boldsymbol{\epsilon}$, so that $\boldsymbol{\epsilon}_i \sim [\mathbf{0}, \sigma_\epsilon^2 \mathbb{I}]$.

5.2.1 Mean estimation

Local polynomial regression is used to estimate the mean structure $\boldsymbol{\mu}(t)$, motivated by Taylor's expansion

$$\boldsymbol{\mu}(t) \approx \boldsymbol{\mu}(t_0) + \boldsymbol{\mu}'(t_0)(t - t_0) + \cdots + \frac{\boldsymbol{\mu}^{(m)}(t_0)}{m!}(t - t_0)^p.$$

Thus we can estimate $\boldsymbol{\mu}(t)$ by estimating the terms in the truncated Taylor expansion above. This can be done using weighted least squares by solving the following,

$$(\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{\beta}}_{-0}) = \underset{\boldsymbol{\beta}_0, \cdots, \boldsymbol{\beta}_m \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} \sum_{j=1}^{n_i} ||\mathbf{Y}_{ij} - \boldsymbol{\beta}_0 - \sum_{l=1}^{m} \boldsymbol{\beta}_l (T_{ij} - t)^l||_2^2 K_h(T_{ij} - t), \tag{5.6}$$

where $K_h = K(\frac{\cdot}{h})/h$, h controls the size of the neighbourhood around t and $K(\cdot)$ is a kernel function that determines the weights of the data points in WLS based on their distance from the point of interest. In this case, K is a function of the difference between the times of the observations (T_{ij}) and the time point of interest (t). Then for $l = 1, 2, \cdots, m$, $\hat{\boldsymbol{\mu}}^{(l)}(t_0) \equiv l!\hat{\boldsymbol{\beta}}_l$. Using this, we can obtain the estimate for $\hat{\boldsymbol{\mu}}(t) = \hat{\boldsymbol{\beta}}_0$.

The WLS problem above is more easily understood in matrix notation. A design matrix for subject i can be expressed as

$$\mathbb{X}_i = \begin{bmatrix} \mathbb{I}_p & (T_{i1} - t)\mathbb{I}_p & (T_{i1} - t)^2\mathbb{I}_p & \cdots & (T_{i1} - t)^m\mathbb{I}_p \\ \vdots & \vdots & \vdots & & \vdots \\ \mathbb{I}_p & (T_{in_i} - t)\mathbb{I}_p & (T_{in_i} - t)^2\mathbb{I}_p & \cdots & (T_{in_i} - t)^m\mathbb{I}_p \end{bmatrix}_{pn_i \times p(m+1).} \tag{5.7}$$

The weight matrix $\mathbb{W}_i$ for subject i is the $n_i \times n_i$ diagonal matrix with entries $\mathbb{W}_i|_{j,j} = K_h(T_{ij} - t)$. $\mathbf{Y}_i$ is a vector of length n_i which contains the observations for individual i. Since the subjects are independent, the minimization problem (5.6) becomes

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{n} (\mathbf{Y}_i - \mathbb{X}_i(t)\boldsymbol{\beta})^T \mathbb{W}_i(t)(\mathbf{Y}_i - \mathbb{X}_i(t)\boldsymbol{\beta}). \tag{5.8}$$

This can also be expressed as

$$\underset{\boldsymbol{\beta}}{\text{argmin}}(\mathbf{Y} - \mathbb{X}(t)\boldsymbol{\beta})^T \mathbb{W}(t)(\mathbf{Y} - \mathbb{X}(t)\boldsymbol{\beta}), \tag{5.9}$$

where $\mathbb{X}(t)$ is the $(p\sum_i n_i) \times p(m+1)$ dimensional design matrix obtained as a vertical stacking of the design matrices per individual $\mathbb{X}_i(t)$ as follows

$$\mathbb{X}(t) = \begin{bmatrix} \mathbb{X}_1^T(t) & \mathbb{X}_2^T(t) & \cdots & \mathbb{X}_i^T(t) & \cdots \mathbb{X}_n^T(t) \end{bmatrix}^T, \tag{5.10}$$

The weight matrix $\mathbb{W}(t)$ is the block diagonal concatenation of individual weight matrices $\mathbb{W}_i(t)$. Lastly, $\mathbf{Y}$ is simply a length $\sum_{i=1}^n n_i$ vector containing all the observations for all individuals. The solution to (5.9) is given by

$$\hat{\boldsymbol{\beta}} = \Big(\sum_{i=1}^n \mathbb{X}_i^T \mathbb{W} \mathbb{X}_i \Big)^{-1} \sum_{i=1}^n \mathbb{X}_i^T \mathbb{W}_i \mathbf{Y}_i = (\mathbb{X}^T \mathbb{W} \mathbb{X})^{-1} \mathbb{X}^T \mathbb{W} \mathbf{Y} = \mathbb{A}\mathbf{Y}. \tag{5.11}$$

Note that we can decompose $\hat{\boldsymbol{\beta}} = [\hat{\boldsymbol{\beta}}_0^T, \hat{\boldsymbol{\beta}}_1^T, \cdots, \hat{\boldsymbol{\beta}}_m^T]^T \in \mathbb{R}^{p(m+1)}$. Finally, we can obtain $\hat{\boldsymbol{\mu}}(t)$ by taking the first p elements of $\hat{\boldsymbol{\beta}}$ which exactly corresponds to $\hat{\boldsymbol{\beta}}_0$.

5.2.2 Marginal covariance estimation

From (5.4) we note that all individuals have the same mean structure and individual-specific deviations are captured by $\mathbf{U}_i(t)$. Thus, in order to estimate the covariance function $\Sigma(s,t)$, also denoted by $\text{Cov}(s,t)$, we need to isolate $\mathbf{U}_i(t)$ for each individual. We can do just this since we now know how to estimate $\boldsymbol{\mu}$! In this subsection we describe how to obtain covariance estimates using the residuals which can be obtained using our estimated $\hat{\mu}$.

Before proceeding further we clarify what we mean by 'residual' here and establish some notation. If we knew the true underlying mean function common to all individuals $\boldsymbol{\mu}(t)$, we could get ideal residuals per individual $(i = 1, 2, \cdots, n)$ per observation $(j = 1, 2, \cdots, n_i)$,

$$\mathbf{r}_{ij} = \mathbf{Y}_{ij} - \boldsymbol{\mu}(T_{ij}).$$

These residuals can be used to estimate the covariance function of interest since $\mathbb{E}(\mathbf{r}_{ij}\mathbf{r}_{ik}) = \text{Cov}(T_{ij}, T_{ik})$. In practice, we instead use the estimated $\hat{\boldsymbol{\mu}}(t)$ to obtain observed residuals

$$\hat{\mathbf{r}}_{ij} = \mathbf{Y}_{ij} - \hat{\boldsymbol{\mu}}(T_{ij}).$$

and rely on the approximation $\mathbb{E}(\hat{\mathbf{r}}_{ij}\hat{\mathbf{r}}_{ik}) \approx \text{Cov}(T_{ij}, T_{ik})$.

In the remainder of this document we take the population mean to be zero, $\boldsymbol{\mu} = 0$. Further, in this subsection we we omit the measurement error term, i.e. $\boldsymbol{\epsilon} \equiv 0$ which means that the

underlying random function $\mathbf{X}_i(t)$ is directly observed for every individual i at a finite number of subject-specific time points. In this setting, we describe joint variance and covariance estimation as below. The next subsection provides particulars on covariance estimation in the more realistic case of non-zero measurement error.

In order to estimate the covariance function, we perform local polynomial regression on products of the observed residuals within individuals, which arise from elements of the outer product matrices $\hat{\mathbf{r}}_{ij}\hat{\mathbf{r}}_{ik}^T$ for all $j, k = 1, 2, \cdots, n_i$. Since we are taking the population mean to be 0, we solve the following optimization problem to obtain an estimate of $\text{Cov}(s, t)$

$$(\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{\beta}}_{-0}) = \underset{\boldsymbol{\beta}_0, \boldsymbol{\beta}_{-0}}{\text{argmin}} \sum_{i=1}^{n} \sum_{j,k=1}^{n_i} ||\mathbf{Y}_{ij}\mathbf{Y}_{ik}^T - \boldsymbol{\beta}_0 - \boldsymbol{\beta}_1(T_{ij} - s) - \boldsymbol{\beta}_2(T_{ik} - t) - \boldsymbol{\beta}_3(T_{ij} - s)^2 - \boldsymbol{\beta}_4(T_{ik} - t)^2 - \boldsymbol{\beta}_5(T_{ij} - s)(T_{ik} - t)||_F^2 K_h(T_{ij} - s)K_h(T_{ik} - t). \tag{5.12}$$

The optimization problem above corresponds to quadratic local smoothing, whereas the expression without the red terms corresponds to linear smoothing. In this report we only model local linear and quadratic smoothers, however this can be trivially extended to any polynomial of degree m by simply adding the corresponding smoothing terms. We can re-express (5.12) in conventional WLS notation, as in (5.8), by defining

$$\underset{\sim}{\mathbf{Y}}_i = [\text{vec}(\mathbf{Y}_{ij}\mathbf{Y}_{ik}^T)]_{j,k=1,2,\cdots,n_i}. \tag{5.13}$$

We denote $\mathbf{Y}$ by $\underset{\sim}{\mathbf{Y}}$ and correspondingly $\mathbf{Y}_i$ by $\underset{\sim}{\mathbf{Y}}_i$ to highlight the involvement of products. $\underset{\sim}{\mathbf{Y}}_i$ here is a vector of length $p^2 \times \sum_i n_i^2$ containing the vectorized sample covariance matrices between all possible sets of time points observed for the i^{th} individual. The design matrix per individual is now

$$\mathbb{X}_i = \begin{bmatrix} \mathbb{I}_q & (T_{i1} - s)\mathbb{I}_q & (T_{i1} - t)\mathbb{I}_q & (T_{i1} - s)^2\mathbb{I}_q & (T_{i1} - t)^2\mathbb{I}_q & (T_{i1} - s)(T_{i1} - t)\mathbb{I}_q \\ \mathbb{I}_q & (T_{i1} - s)\mathbb{I}_q & (T_{i2} - t)\mathbb{I}_q & (T_{i1} - s)^2\mathbb{I}_q & (T_{i2} - t)^2\mathbb{I}_q & (T_{i1} - s)(T_{i2} - t)\mathbb{I}_q \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbb{I}_q & (T_{ij} - s)\mathbb{I}_q & (T_{ik} - t)\mathbb{I}_q & (T_{ij} - s)^2\mathbb{I}_q & (T_{ik} - t)^2\mathbb{I}_q & (T_{ij} - s)(T_{ik} - t)\mathbb{I}_q \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbb{I}_q & (T_{in_i} - s)\mathbb{I}_q & (T_{in_i} - t)\mathbb{I}_q & (T_{in_i} - s)^2\mathbb{I}_q & (T_{in_i} - t)^2\mathbb{I}_q & (T_{in_i} - s)(T_{in_i} - t)\mathbb{I}_q \end{bmatrix}_{j,k=1,2,\cdots,n_i}$$

which is a matrix with $q \times n_i^2$ rows where $q = p^2$. Lastly, the weight matrix per individual is a

diagonal matrix of dimension $n_i^2 q \times n_i^2 q$ of the form

$$\mathbb{W}_i = \textit{diag}\big(\{K_h(T_{ij} - s)K_h(T_{ik} - t)\}_{j,k=1,2,\cdots,n_i}\big) \otimes \mathbb{I}_q.$$

The parameters in (5.9) for covariance estimation include $\mathbf{Y}$, now denoted by $\underset{\sim}{\mathbf{Y}}$, which is a length $p^2 \times \sum_i n_i^2$ vector formed by vertically stacking the n vectors $\underset{\sim}{\mathbf{Y}}_i$ as

$$\underset{\sim}{\mathbf{Y}} = \begin{bmatrix} \underset{\sim}{\mathbf{Y}_1}^T & \underset{\sim}{\mathbf{Y}_2}^T & \cdots & \underset{\sim}{\mathbf{Y}_n}^T \end{bmatrix}^T. \tag{5.14}$$

The design matrix $\mathbb{X}$ is formed by stacking $\mathbb{X}_i$ as in (5.10) and the weight matrix $\mathbb{W}$ is a diagonal weight matrix which can be thought of as containing n diagonal blocks, where the i^{th} block $\mathbb{W}_i$ corresponds to the i^{th} individual.

The covariance estimate $\hat{\Sigma}(s,t)$ is then obtained by taking the first q elements of $\hat{\boldsymbol{\beta}}$, as in (5.11), and reshaping them to form a $p \times p$ matrix.

5.2.3 Variance estimation

In this subsection, we describe covariance estimation in the presence of additive measurement noise ($\boldsymbol{\epsilon} \neq 0$). Start by noting that in the previous section, we estimated both variances and covariances using the exact same algorithm. It might, however, make more sense to treat variances and covariance estimation separately (as previously modelled by [98]). This is because, from (5.5) we know that for any individual i with observations j, k,

$$\begin{aligned} \text{Cov}(\mathbf{Y}_{ij}, \mathbf{Y}_{ik}) = \text{Cov}(\mathbf{X}_i(T_{ij}), \mathbf{X}_i(T_{ik})) &= \text{Cov}(\mathbf{U}_i(T_{ij}), \mathbf{U}_i(T_{ik})) + \text{Cov}(\epsilon_{ij}, \epsilon_{ik}) \\ &= \Sigma(T_{ij}, T_{ik}) \;\; + \;\; \delta_{jk}\sigma_\epsilon^2 \mathbb{I}_q, \end{aligned}$$

where $\delta_{jk} = 1$ if $j = k$ and 0 otherwise. Hence, for $s, t \in \mathbb{R}$, if some individual i is observed at distinct time points s and t with $T_{ij} = s$ and $T_{ik} = t$, then

$$\begin{aligned} &\text{Cov}(s,t) = \text{Cov}(\mathbf{X}(s), \mathbf{X}(t)) = \text{Cov}(\mathbf{Y}_{ij}, \mathbf{Y}_{ik}) = \Sigma(s,t) \qquad \text{and} \\ &\text{Cov}(s,s) = \text{Var}(\mathbf{Y}_{ij}) = \Sigma(s,s) + \sigma_\epsilon^2 \mathbb{I}_q = \text{Var}(\mathbf{X}(s)) + \sigma_\epsilon^2 \mathbb{I}_q. \end{aligned}$$

If we estimate variances and covariances separately, the norm in (5.12) changes and consequently the definitions of parameters $(\mathbb{X}, \underset{\sim}{\mathbf{Y}}, \mathbb{W})$ in (5.9) change. In particular, for $s = t$, the Frobenius norm is replaced by the diagonal-only Frobenius norm in (5.12) which leads to the

following

$$\underset{\sim}{\mathbf{Y}}_i = [\text{vec}(\mathbf{Y}_{ij}\mathbf{Y}_{ij}^T)]_{j=1,2,\cdots,n_i}, \tag{5.15}$$

$$\mathbb{X}_i = \begin{bmatrix} \vdots & \vdots & \vdots \\ \mathbb{I}_q & (T_{ij}-s)\mathbb{I}_q & (T_{ij}-s)^2\mathbb{I}_q \\ \vdots & \vdots & \vdots \end{bmatrix}_{j=1,2,\cdots,n_i}$$

and

$$\mathbb{W}_i = \mathit{diag}\big(\{K_h(T_{ij}-s)K_h(T_{ij}-t)\}_{j=1,2,\cdots,n_i}\big) \otimes \mathbb{I}_q.$$

For $s \neq t$, we consider the hollow Frobenius norm in (5.12). $\underset{\sim}{\mathbf{Y}}_i$, $\mathbb{X}_i$ and $\mathbb{W}$ are now defined as in the case of the joint estimation except only sets of distinct time-points are considered, i.e. the term n_i^2 anywhere in the expression for the dimensions of either $\underset{\sim}{\mathbf{Y}}_i$ or $\mathbb{X}_i$ is replaced by $n_i(n_i-1)$. For the remainder of this report, we will only refer to and focus on the separate estimation case. The implementation of the method includes only this case as well.

5.3 Inference

In this section we propose a plug-in based method to estimate the standard error (SE) of the marginal covariance estimate obtained from local polynomial regression, as described in Section 2. Since the marginal estimates ($\hat{\Sigma}_{s,t}$) are linear functions of the collection of product data $\underset{\sim}{\mathbf{Y}}$, the standard errors for these can be estimated using simple moment calculations, leading to familiar "sandwich expressions". That is, recall from (5.11) that $\hat{\boldsymbol{\beta}} = \mathbb{A}\underset{\sim}{\mathbf{Y}}$ and $\hat{\Sigma}_{st} = \hat{\boldsymbol{\beta}}|_{1:q}$ reshaped into a $p \times p$ matrix. Therefore,

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \mathbb{A}\text{Cov}(\underset{\sim}{\mathbf{Y}})\mathbb{A}^T = \sum_{i=1}^{n} \mathbb{A}_i \text{Cov}(\underset{\sim}{\mathbf{Y}}_i)\mathbb{A}_i^T \tag{5.16}$$

The sandwich term $\text{Cov}(\underset{\sim}{\mathbf{Y}})$ contains fourth moments, which can be calculated purely non-parametrically, or via plug-in, by expressing the fourth moments in terms of second moments. Here we use a plug-in procedure.

The plug-in procedure for obtaining standard errors that is developed here requires that the fourth and second moments are related just as they are in a Gaussian distribution. Previous work in this area has either estimated the higher-order moments directly from the data, which requires

a large sample size to obtain stable results, or linked the higher-order and lower-order moments, as we do here. Some work has used asymptotic approximations to isolate the dominant terms, somewhat simplifying the implementation. Notably however, most published work in this area uses the bootstrap for inference, possibly due to the difficulty of implementing any of the analytically-derived inference procedures. As far as we are aware, there are no empirical results assessing the performance of non-asymptotic analytic approaches to inference such as those that we are implementing. Thus, our work will provide a useful assessment of the tradeoffs between analytic approaches and alternatives such as the bootstrap. It has been proven that the covariance estimates obtained as described in section 2 are asymptotically Gaussian ([102]), justifying the use of these standard errors for forming confidence intervals and conducting hypothesis tests.

As discussed in section 5.1.4, there are theoretical results on the consistency, asymptotic distributions and the rates of convergence of the covariance estimates obtained using local linear regression. However, we are not aware of computational methods for performing inference on such estimates, besides bootstrap techniques to obtain standard errors. Our inference technique models these (standard errors) mathematically, which allows us to understand the intricacies of the SE, for eg. to understand which aspect of estimation contributes most to the standard error, and how this can be modelled better/ improved via tuning.

Plug-in inference procedure

We start by describing the setup and the assumption. We assume that $\mathbf{X}(t)$ is a p-dimensional Gaussian process with mean 0 and covariance function $\text{Cov}(\mathbf{X}(s), \mathbf{X}(t)) = \text{Cov}(s,t) = \Sigma(s,t)$, and that $\mathbf{Y}_{ij}$ denotes the observed value of $\mathbf{X}_i(T_{ij})$ corresponding to observation $j = 1, 2, \cdots, n_i$ for individual $i = 1, 2, \cdots, n$ which is contaminated with measurement errors. That is, the setup remains as in (5.5) with the added assumption that $\mathbf{U}(t) \sim \mathcal{N}_p(\mathbf{0}_p, \Sigma(t,t))$ for any $t \in \mathbb{R}^+$. Denote $\Sigma(s,t)$ by Σ_{st} and recall that $\mathbf{Y}_{ij} = \begin{pmatrix} Y_{ij1} & Y_{ij2} & \cdots & Y_{ijp} \end{pmatrix}^T$. Again, we know $\hat{\Sigma}_{st} = \hat{\boldsymbol{\beta}}|_{1:q}$ reshaped into a $p \times p$ matrix, i.e. $\text{vec}(\hat{\Sigma}_{st}) = \hat{\boldsymbol{\beta}}|_{1:q}$ where $\hat{\boldsymbol{\beta}} = \underset{\sim}{\mathbb{A}}\mathbf{Y}$. (recall $p^2 = q$)

To get the standard error of an arbitrary estimate $\hat{\Sigma}_{st}$, we perform the following steps:

1. Find the variance of $\hat{\boldsymbol{\beta}}$, denoted by $\text{Var}(\hat{\boldsymbol{\beta}})$.
2. Extract the $1 : q \times 1 : q$ submatrix of $\text{Var}(\hat{\boldsymbol{\beta}})$ corresponding to the first q elements of $\hat{\boldsymbol{\beta}}$ (recall that these first q elements exactly constitute the vectorized form of $\hat{\Sigma}_{st}$, i.e. $\hat{\boldsymbol{\beta}}|_{1:q} = \text{vec}(\hat{\Sigma}_{st})$).
3. Take the diagonal elements of this matrix to obtain the variances of each of these elements, $\text{Diag}(\text{Var}(\hat{\boldsymbol{\beta}})|_{1:q \times 1:q})$.

4. Reshape this vector of length $q(=p^2)$ into a $p \times p$ matrix to obtain $\text{Var}(\hat{\Sigma}_{st})$ which is a $p \times p$ matrix that contains the estimated variance of each element of $\hat{\Sigma}_{st}$ and take the square root.

This seems fairly straightforward, however the difficulty arises in estimating $\text{Var}(\hat{\boldsymbol{\beta}})$. This is because estimating $\text{Var}(\hat{\boldsymbol{\beta}})$ involves finding the variance of products of the original variables which is equivalent to computing the fourth-order moments of the original variables. This will become clear in a moment. From (5.16) we know that

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \mathbb{A}\text{Cov}(\underset{\sim}{\mathbf{Y}})\mathbb{A}^T = \sum_{i=1}^{n} \mathbb{A}_i \text{Cov}(\underset{\sim}{\mathbf{Y}}_i)\mathbb{A}_i^T, \tag{5.17}$$

where each element of the vector $\underset{\sim}{\mathbf{Y}}$ is a product of observed values of the form $Y_{ijp_1}Y_{ikp_2}$ where $p_m = 1, 2, \cdots, p$ and $j, k = 1, 2, \cdots, n_i$ for some subject $i = 1, 2, \cdots, n$ (recall from (5.15)). Then $\mathbb{Y} = \text{Cov}(\underset{\sim}{\mathbf{Y}})$ is a $len(\underset{\sim}{\mathbf{Y}}) \times len(\underset{\sim}{\mathbf{Y}})$ matrix whose entires are of the form

$$\text{Cov}(Y_{ijp_1}Y_{ikp_2}, Y_{i'j'p_3}Y_{i'k'p_4}), \tag{5.18}$$

for possibly distinct subjects i, i' and attributes $p_m = 1, 2, \cdots, p$. Note that $\mathbb{Y}$ is block diagonal due to the assumption of different individuals being independent, thus we need only consider the case when $i = i'$ in (5.18).

In order to estimate terms of the form given by (5.18), we can decompose these into simpler forms using the definition of covariance as follows

$$\begin{aligned}
&\text{Cov}(Y_{ijp_1}Y_{ikp_2}, Y_{i'j'p_3}Y_{i'k'p_4}) \\
&\quad = \mathbb{E}(Y_{ijp_1}Y_{ikp_2}Y_{i'j'p_3}Y_{i'k'p_4}) - \mathbb{E}(Y_{ijp_1}Y_{ikp_2})\mathbb{E}(Y_{i'j'p_3}Y_{i'k'p_4}) \\
&\quad = \mathbb{E}(Y_{ijp_1}Y_{ikp_2}Y_{i'j'p_3}Y_{i'k'p_4}) - \text{Cov}(Y_{ijp_1}, Y_{ikp_2})\text{Cov}(Y_{i'j'p_3}, Y_{i'k'p_4}) \\
&\qquad \text{since the data are centered (so } \mathbb{E}(Y_{p_m}(t_{ij})) = 0 \ \forall i, j, m) \\
&\quad = \mathbb{E}(Y_{ijp_1}Y_{ikp_2}Y_{i'j'p_3}Y_{i'k'p_4}) - [\text{Cov}(t_{ij}, t_{ik})]_{p_1,p_2}\left[\text{Cov}(t_{i'j'}, t_{i'k'})\right]_{p_3,p_4}.
\end{aligned} \tag{5.19}$$

To estimate $\mathbb{Y}$ we must estimate all terms of the form (5.18) which corresponds to the LHS of (5.19). Estimates for the latter part of the RHS of (5.19) can be obtained directly from local polynomial regression. Estimating the fourth-order moment term is the main challenge which is solved by further decomposing this term using higher-moment formulae and Gaussianity. Since we have assumed that $\mathbf{Y}(t)$ is Gaussian for any $t \in \mathbb{R}$, we know that $(Y_{ijp_1}, Y_{ikp_2}, Y_{i'j'p_3}, Y_{i'k'p_4})^T$

will be jointly Gaussian. For any multivariate Gaussian vector

$$(X_1, X_2, X_3, X_4)^T \sim \mathcal{N}(\mathbf{0}_4, \Sigma)$$

with $\text{Cov}(X_i, X_j) = \sigma_{ij}$, the following hold from Isserlis' theorem

$$\mathbb{E}(X_i^4) = 3\sigma_{ii}^2$$
$$\mathbb{E}(X_i^3 X_j) = 3\sigma_{ii}\sigma_{ij}$$
$$\mathbb{E}(X_i^2 X_j^2) = \sigma_{ii}\sigma_{jj} + 2\sigma_{ij}^2$$
$$\mathbb{E}(X_i^2 X_j X_k) = \sigma_{ii}\sigma_{jk} + 2\sigma_{ij}\sigma_{ik}$$
$$\mathbb{E}(X_i X_j X_k X_n) = \sigma_{ij}\sigma_{kn} + \sigma_{ik}\sigma_{jn} + \sigma_{in}\sigma_{jk}$$

Thus, using the above decomposition of fourth-moments in terms of second-order moments for multivariate Gaussian variables with mean $\mathbf{0}$, we can decompose the fourth-order term in the RHS of (5.19) in terms of second-order moments. Since all the second-order terms can be estimated directly by our non-parametric method, we can estimate $\mathbb{Y}$ directly from our lower-order covariance estimates. We can then plug-in $\hat{\mathbb{Y}}$ into (5.16) to obtain $\widehat{\text{Var}(\hat{\boldsymbol{\beta}})}$. Once we have this, steps 2-4 of finding the $\text{SE}(\hat{\Sigma}_{st})$ are straight-forward.

5.3.1 Computation

5.3.1.1 Computational complexity

Marginal covariance estimates:
While the scale of the problem is not large, i.e. the regression is still very optimized, the formation of design matrices requires book-keeping which is the main challenge. Tracking how sample and population moments are linked leads to a combinatorial explosion. Additionally, implementing the technique in a way that can be generalized to other data settings, including multiple attributes, multiple-indices, different data designs etc, means abstracting the moment matching process, which makes it harder to simplify this/ reduce it to a simpler form. To find a way around this, we plan on finding estimates element-wise (instead of estimating the entire $p \times p$ covariance matrix) when we re-implement this technique. To put this is context, obtaining a covariance estimate for a dataset with 1500 individuals, each having 10 entries takes about 7 seconds.

Consequent challenges in estimating the SE:
SE estimates involve computing $\text{Cov}(\underset{\sim}{\mathbf{Y}}_i)$ per individual i, which corresponds to finding $p^2 n_i^2(n_i^2 + 1)/2$ fourth order moments. These in turn require $\text{Cov}(s,t)\ \forall s,t \in \{T_{i1}, \cdots, T_{in_i}\}$ (about n_i^2 esti-

mates). Therefore, a single SE estimate requires obtaining $p^2 \sum_{i=1}^{n} n_i^4$ estimates which implicitly depend on a total of $p^2 \sum_{i=1}^{n} n_i^2$ marginal estimates themselves. This is on the order of n_i^4 which increases very fast as n_i increases. For instance, for a dataset with 1500 individuals, each having 10 entries, obtaining a single SE estimates involves finding $3,030,000$ fourth order moments, which depend on a total of $150,000$ marginal estimates in different ways.

Note that the above corresponds to a single set of values for (s,t) and the entire procedure needs to be repeated for every different (s,t) pair, i.e. for every $\widehat{\text{SE}}(\hat{\Sigma}_{st})$. Thus, to make the computation of standard errors more feasible, we adopt two techniques in additional to parallelism. First, we calculate selective marginal estimates. We simply don't compute those marginal estimates whose associated kernel weights are negligible, i.e. to obtain $\widehat{\text{SE}}(\hat{\Sigma}_{st})$, skip computing $\text{Cov}(s_0,t_0)$ if $K_h(t-t_0)K_h(s-s_0)$ is negligible.

Next, we use a plug-in approach along with an approximation technique similar to memoization. In particular, we plug-in an already computed estimate, say $\widehat{\text{Cov}}(s_0,t_0)$, for $\widehat{\text{Cov}}(s,t)$ in the formation of $\text{Cov}(\underset{\sim}{\mathbf{Y}}_i)$, provided $(s-s_0),(t-t_0) < \delta_h$ simultaneously for some prefixed $\delta_h > 0$. We are able to do this because the covariance surface being estimated here is a smooth function on a compact set.

This significantly reduces the computational burden and speeds up the process. The idea is that since $\text{Cov}(\underset{\sim}{\mathbf{Y}}_i)$ is sandwiched in equation 5.17, a majority of the terms probably cancel out with each other and thus an approximate estimate works quite as well. Preliminary tests show that the SE estimates obtained by computing $1000-2000$ distinct marginal estimates are essentially the same as those obtained by computing marginal estimates at $10,000$ points and then interpolating.

5.3.1.2 Treating missing data

Regardless of the missingness pattern, no information from any of the observations is thrown away. For instance, if one of the attributes is not observed at a time point, say the $j^{th} > 1$ time point for an individual i, then the other attributes measured at t_{ij} will still be considered in forming the sample second moments and thus, will contribute to the model-fitting regression.

For instance, if p_2 is the only missing attribute, i.e. $\mathbf{Y}_{ij} = [Y_{ij1}\ ?\ Y_{ij3} \cdots Y_{ijp}]$, then the parameters in the regression for covariance estimation (5.12) will incorporate the missingness as follows. For ease of explanation, we describe the parameters assuming only one missing entry for individual i. Starting with $\underset{\sim}{\mathbf{Y}}_i$, all products corresponding to Y_{ij2} will be omitted. This means that first entry omitted in $\underset{\sim}{\mathbf{Y}}_i$ as compared to its form in (5.13) will be the $p^2(j-1)+2^{th}$ entry. In general, missing entries among the first $p^2 n_i$ entries of $\underset{\sim}{\mathbf{Y}}_i$ as in (5.13), corresponding to products with $\mathbf{Y}_{i1}$ will be at positions $p^2(j-1)+2+(m-1)p$ where $m = 1,2,\cdots,p$. For each missing entry in $\underset{\sim}{\mathbf{Y}}_i$, the corresponding row in $\mathbb{X}_i$ and $\mathbb{W}_i$ is omitted as well. For instance, corresponding to the first

missing entry in $\underset{\sim}{\mathbf{Y}}_i$ which is at the $p^2(j-1)+2^{th}$ position, the 2^{nd} row in the j^{th} row-block of $\mathbb{X}_i$ and $\mathbb{W}_i$ will be omitted. This is precisely the $p^2(j-1)+2^{th}$ row.

5.4 Conditional covariance

The general form of the conditional correlation function for a random function $\mathbf{X}$ with p attributes, (recall $\mathbf{X}(t) = [X_{p_m}(t)]^T_{m=1,\cdots,p} \in \mathbb{R}^p$ is given by

$$\text{Cor}(X_{p_1}(t_1), X_{p_2}(t_2)|X_{p_3}(t_3)) = \frac{\text{Cov}(X_{p_1}(t_1), X_{p_2}(t_2)|X_{p_3}(t_3))}{[\text{Var}(X_{p_1}(t_1)|X_{p_3}(t_3)) \cdot \text{Var}(X_{p_2}(t_2)|X_{p_3}(t_3))]^{1/2}}. \tag{5.20}$$

Motivated by our application, we focus on a specific form of this conditional correlation. This is precisely (5.2), which is reproduced below

$$\text{Cor}(X_1(s), X_2(t)|X_1(t)) = \frac{\text{Cov}(X_1(s), X_2(t)|X_1(t))}{[\text{Var}(X_1(s)|X_1(t)) \cdot \text{Var}(X_2(t)|X_1(t))]^{1/2}}. \tag{5.2}$$

To estimate (5.2) and its standard error, we propose two techniques, both of which use kernel regression. The first technique directly models the conditional covariance, whereas the second technique estimates the marginal covariances, and uses the conditional Gaussian identity to obtain an estimate of the conditional covariance. Both approaches use the variance estimation techniques from Section 5.2.3. We present these methods in detail below, and briefly comment on their ability to estimate the broader conditional correlation (5.20). As for the inference, we extend the technique from Section 5.3 to find the SE of the conditional estimates using linearization.

5.4.1 Technique 1: indirect kernel estimation

This technique models the conditional covariances required in estimating (5.2) for a p-dimensional Gaussian process $\mathbf{X}(t)$. Specifically, using the decomposition of conditional covariances for Gaussian data, we have

$$\text{Cov}\left(\begin{bmatrix} X_1(s) \\ X_2(t) \end{bmatrix} |X_1(t) \right) = \begin{bmatrix} \Sigma^{11}_{ss} - \frac{(\Sigma^{11}_{st})^2}{\Sigma^{11}_{tt}} & \Sigma^{12}_{st} - \frac{\Sigma^{11}_{st}\Sigma^{12}_{tt}}{\Sigma^{11}_{tt}} \\ \Sigma^{12}_{st} - \frac{\Sigma^{11}_{st}\Sigma^{12}_{tt}}{\Sigma^{11}_{tt}} & \Sigma^{22}_{tt} - \frac{(\Sigma^{12}_{tt})^2}{\Sigma^{11}_{tt}} \end{bmatrix}, \tag{5.21}$$

where

$$\Sigma^{p_1p_2}_{t_1t_2} = \text{Cov}(X_{p_1}(t_1), X_{p_2}(t_2)) = \text{Cov}(\mathbf{X}(t_1), \mathbf{X}(t_2))|_{p_1p_2}.$$

We can easily obtain an estimate for $\text{Cov}([X_1(s)\ X_2(t)]^T|X_1(t))$, denoted by $\hat{\Sigma}_{s_1t_2|t_1}$ by plugging-in marginal estimates for terms in the RHS of (5.21), obtained using the technique described in Section 5.2.3. Note that using this notation, our correlation of interest, as in (5.2), can be denoted by $\text{Cor}^{12}_{s_1t_2|t_1}$.

5.4.2 Technique 2: direct kernel estimation

Note that each of the covariances in (5.2) is of the form $\text{Cov}(Z_1(t_1), Z_2(t_2)|Z_3(t_3))$ for some combination of time-points $t_1, t_2 \in \{s, t\}$ and random functions $Z_1, Z_2 \in \{X_1, X_2\}$ with $Z_3 \equiv X_1$ and $t_3 = t$. With this in mind, we wish to use kernel regression to directly estimate the specific conditional covariances

$$\text{Cov}(Z_1(t_1), Z_2(t_2)|Z_3(t_3) = c), \tag{5.22}$$

to estimate directly

$$\text{Corr}(Z_1(t1), Z_2(t_2)|Z_3(t_3) = c)) = \frac{\text{Cov}(Z_1(t_1), Z_2(t_2)|Z_3(t_3) = c)}{\text{Var}(Z_1(t_1)|Z_3(t_3) = c)\text{Var}(Z_2(t_2)|Z_3(t_3) = c)}. \tag{5.23}$$

For a Gaussian process, this conditional correlation does not depend on the value of c. However for non-Gaussian data, the conditional covariances, and hence the conditional correlation can vary with c. To attain an overall summary, we average (5.23) over the marginal distribution of $Z_3(t_3)$ to obtain estimates for the averaged conditional correlation of interest:

$$\mathbb{E}_{Z_3(t_3)}[\text{Corr}(Z_1(t_1), Z_2(t_2)|Z_3(t_3))]. \tag{5.24}$$

Below we provide a local regression approach that provides an estimate $\widehat{\text{Cov}}(Z_1(t_1), Z_2(t_2)|Z_3(t_3) = c)$ of (5.22). This can then be averaged with weights to obtain an estimate of (5.24), via

$$\delta^{-1} \sum_c \widehat{\text{Corr}}(Z_1(t_1), Z_2(t_2)|Z_3(t_3) = c) \cdot \hat{\pi}(c), \tag{5.25}$$

where $\hat{\pi}$ is an estimate of the marginal density of $Z_3(t_3)$, obtained, for example, using kernel density estimation, and c runs over a grid with mesh δ.

The approach for estimating (5.22) using kernel regression is similar to that in section 5.2.2, except that we now need a kernel with three terms. Let z_{ijk} represent the observed value of $Z_k(T_{ij})$ for the i^{th} individual. Then an estimate β_0 of the conditional covariance $\text{Cov}(Z_1(t_1), Z_2(t_2)|Z_3(t_3) = c)$ is obtained by solving the following minimization problem

$$\underset{\beta_0,\beta_{-0}}{\text{argmin}} \sum_{i=1}^{n}\sum_{k_1=1}^{n_{i1}}\sum_{k_2=1}^{n_{i2}}\sum_{k_3=1}^{n_{i3}} ||z_{i1k_1}z_{i2k_2}-\beta_0-\beta_1(T_{i1k_1}-t_1)-\beta_2(T_{i2k_2}-t_2)-\beta_3(T_{i3k_3}-t_3)-\beta_4(z_{i3j_3}-c)||_F^2 W_{ik_1k_2k_3} \tag{5.26}$$

where the weight

$$W_{ik_1k_2k_3} = K_h(T_{ik_1}-t_1)K_h(T_{ik_2}-t_2)K_h(T_{ik_3}-t_3)K_h(Z_3(T_{ik_3})-c). \tag{5.27}$$

The important thing to note here is that, all conditional covariances in (5.2) have at most two distinct time points. Thus, in estimating $\mathbb{E}(Z_1(t_1), Z_2(t_2)|Z_3(t_3)=c)$ above, there are at most two distinct time points, say t_1' and t_2' and $t_1, t_2, t_3 \in \{t_1', t_2'\}$. Thus, the kernel weight (5.27) reduces to the 3-way product

$$W_{ik_sk_t} = K_h(T_{ik_{t_1'}}-t_1')K_h(T_{ik_{t_2'}}-t_2')K_h(Z_3(T_{ik_{t_3}})-c), \tag{5.28}$$

where again, $t_3 \in \{t_1', t_2'\}$.

Specifically, for estimating the cross covariance of interest $\text{Cov}(X_1(s), X_2(t)|X_1(t))$ using our data, (5.26) becomes

$$\underset{\beta_0,\beta_{-0}}{\text{argmin}} \sum_{i=1}^{n}\sum_{k_1=1}^{n_i}\sum_{k_2=1}^{n_i}\sum_{k_3=1}^{n_i} ||Y_{i1k_1}Y_{i2k_2}-\beta_0-\beta_1(T_{ik_1}-s)-\beta_2(T_{ik_2}-t)-\beta_3(T_{ik_3}-t)-\beta_4(Y_{i1k_3}-c)||_F^2 W_{ik_1k_2}, \tag{5.29}$$

where the weight has a similar form to (5.28), which is,

$$W_{ik_1k_2,k_3} = K_h(T_{i1k_1}-s)K_h(T_{i2k_2}-t)K_h(T_{i3k_3}-t)K_h(X_1(T_{i1k_3})-c).$$

The other conditional covariances in (5.2) can be computed in a similar manner. Note that both (5.26,5.29) correspond to local linear smoothing. Additional product terms can be added, just as in (5.12), for different degrees of smoothing.

Instead of using polynomials in the local linear regression which are increasing functions of the distance between time points, we can form polynomials of functions which are decrease with

increase difference in time. One such example is as follows:

$$finv(T-t) = \begin{cases} 2 * sign(T-t) & \text{if } T-t < \frac{1}{thr} \\ \frac{1}{(T-t)} * \frac{2}{thr} & \text{if } T-t > \frac{1}{thr} \end{cases}$$

for some pre-determined threshhold *thr*, for instance, $thr = 100$. This definition ensures that the polynomial arguments lie between ± 2.

Comparison of the two approaches

An advantage of technique 2 over technique 1 is that it allows us to directly model the conditional covariances, and thus the conditional correlation of interest, without any Gaussianity assumption. However, the drawback is that the smoothing now involves a 3-way kernel. This is not only harder to tune but also results in a less efficient estimator as compared to technique 1. We thus need a larger dataset to obtain the an estimate with the same efficiency (SE) as that of the estimate obtain assuming Gaussinaity under technique 1 from a smaller dataset.

Technique 1 can be easily generalized to model (5.20) by decomposing the relevant conditional covariance matrix in terms of marginal covariances. On the other hand, while technique 2 can theoretically be used to model the underlying covariances in (5.20), for distinct $t_1 \neq t_2 \neq t_3$, the weights would be products of 4 kernel functions, as in (5.27), which is not desirable. Due to the curse of dimensionality, we would need a sample size exponential in base 4 to obtain the desired estimates. We are in the process of implementing this technique. It would be interesting to compare the two techniques via simulations, both in terms of accuracy and relative efficiency, in the future.

5.4.3 Inference

Here, we extend the inference procedure from section 5.3 to the conditional estimates obtained using technique 1, and we provide the details for both conditional covariance and conditional correlation estimates. Note that this does not apply to estimates obtained from the direct kernel technique.

5.4.3.1 Conditional covariance

In order to calculate $SE(\hat{\Sigma}_{s_1 t_2 | t_1})$, we use the Delta method to get a linear approximation of the variance of each term of $\hat{\Sigma}_{s_1 t_2 | t_1}$ (given by (5.21)). For each of the four terms, the Delta method requires finding the joint covariance matrix of the elements constituting each term. These elements, ofcourse, are the estimated covariances $\hat{\Sigma}^{p_1 p_2}_{t_1 t_2}$, for some time points $t_1, t_2 \in \mathbb{R}$, whose joint covari-

ance matrix is obtained by element-wise plug-ins using estimates of the marginal standard errors as described in the previous section. Here, we present the details involved in finding the variance of $\widehat{\text{Cov}}(X_1(s)X_2(t)|X_1(t))$, i.e. $\hat{\Sigma}^{12}_{s_1t_2|t_1} = \Sigma^{12}_{st} - \frac{\Sigma^{11}_{st}\Sigma^{12}_{tt}}{\Sigma^{11}_{tt}}$ and the procedure for the other terms is similar.

Details of the multivariate Delta method

To find the variance of $\hat{\Sigma}^{12}_{s_1t_2|t_1} = \hat{\Sigma}^{12}_{st} - \frac{\hat{\Sigma}^{11}_{st}\hat{\Sigma}^{12}_{tt}}{\hat{\Sigma}^{11}_{tt}}$, define

$$f(\hat{\Sigma}^{12}_{st}, \hat{\Sigma}^{11}_{st}, \hat{\Sigma}^{12}_{tt}, \hat{\Sigma}^{11}_{tt}) = \hat{\Sigma}^{12}_{st} - \frac{\hat{\Sigma}^{11}_{st}\hat{\Sigma}^{12}_{tt}}{\hat{\Sigma}^{11}_{tt}}.$$

Then

$$\nabla f = \begin{pmatrix} 1 & -\frac{\hat{\Sigma}^{12}_{tt}}{\hat{\Sigma}^{11}_{tt}} & -\frac{\hat{\Sigma}^{11}_{st}}{\hat{\Sigma}^{11}_{tt}} & \frac{\hat{\Sigma}^{11}_{st}\hat{\Sigma}^{12}_{tt}}{(\hat{\Sigma}^{11}_{tt})^2} \end{pmatrix}.$$

Then by the Delta method,

$$\text{Var}(f) = \nabla f \text{Cov} \begin{pmatrix} \hat{\Sigma}^{12}_{st} \\ \hat{\Sigma}^{11}_{st} \\ \hat{\Sigma}^{12}_{tt} \\ \hat{\Sigma}^{11}_{tt} \end{pmatrix} \nabla f^T.$$

We now describe how to estimate the covariance matrix above. Note that each element of this matrix is of the form

$$\text{Cov}(\hat{\Sigma}^{p_1p_2}_{t_1t_2}, \hat{\Sigma}^{p_1'p_2'}_{t_1't_2'}) = \text{Cov}(\text{vec}(\hat{\Sigma}_{t_1t_2}), \text{vec}(\hat{\Sigma}_{t_1't_2'}))|_{p(p_1-1)+p_2, p(p_1'-1)+p_2'} \tag{5.30}$$

for some time points $t_i, t_i' \in \mathbb{R}$ and some attributes $p_i, p_i' = 1, 2, \cdots, p$ and here $p = 2$. Recall from section 2.1 that

$$\begin{aligned} &\text{vec}(\widehat{\text{Cov}}(t_1, t_2)) = \text{vec}(\hat{\Sigma}_{t_1t_2}) = \hat{\boldsymbol{\beta}}|_{1:q} \text{ where } \hat{\boldsymbol{\beta}} = \mathbb{A}\underset{\sim}{\mathbf{Y}} \\ &\text{and let} \qquad\qquad \text{vec}(\hat{\Sigma}_{t_1't_2'}) = \hat{\boldsymbol{\beta}}'|_{1:q} \text{ where } \hat{\boldsymbol{\beta}}' = \mathbb{A}'\underset{\sim}{\mathbf{Y}}, \end{aligned}$$

which gives

$$\text{Cov}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}}') = \text{Cov}(\mathbb{A}\underset{\sim}{\mathbf{Y}}, \mathbb{A}'\underset{\sim}{\mathbf{Y}}) = \mathbb{A}\text{Cov}(\underset{\sim}{\mathbf{Y}})\mathbb{A}'^T = \mathbb{A}\mathbb{Y}\mathbb{A}'^T.$$

We can then estimate (5.30) by extracting the $(p(p_1-1)+p_2, p(p_1'-1)+p_2')^{th}$ term from the following estimate

$$\widehat{\text{Cov}}(\text{vec}(\hat{\Sigma}_{t_1t_2}), \text{vec}(\hat{\Sigma}_{t_1't_2'})) = \widehat{\text{Cov}}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}}')|_{1:q,1:q} = \mathbb{A}\hat{\mathbb{Y}}\mathbb{A}'^T,$$

where $\hat{\mathbb{Y}}$ can be obtained as described in Section 5.3. In this way, we can get each term in the required covariance matrix for the Delta method and plug-in the values to obtain the required variance estimate.

Inference for estimates obtained using technique 2 is part of future work.

5.4.3.2 Conditional correlation

For technique 1, the conditional correlation of interest is of the form

$$\begin{aligned}\text{Cor}(X_1(s), X_2(t)|X_1(t)) &= \frac{\text{Cov}(X_1(s), X_2(t)|X_1(t))}{[\text{Var}(X_1(s)|X_1(t)) \cdot \text{Var}(X_2(t)|X_1(t))]^{1/2}} \\ &= \frac{\Sigma^{12}_{st} - \frac{\Sigma^{11}_{st}\Sigma^{12}_{tt}}{\Sigma^{11}_{tt}}}{\sqrt{\left(\Sigma^{11}_{ss} - \frac{(\Sigma^{11}_{st})^2}{\Sigma^{11}_{tt}}\right)\left(\Sigma^{22}_{tt} - \frac{(\Sigma^{12}_{tt})^2}{\Sigma^{11}_{tt}}\right)}}\end{aligned}$$

We follow the exact same strategy as the one outlined above for obtaining the SE of conditional covariance estimates, i.e. we use the delta method.

Details of the multivariate Delta method

To find the variance of the conditional correlation, define

$$f(\Sigma^{12}_{st}, \Sigma^{11}_{st}, \Sigma^{11}_{ss}, \Sigma^{11}_{tt}, \Sigma^{12}_{tt}, \Sigma^{22}_{tt}) = \frac{\Sigma^{12}_{st} - \frac{\Sigma^{11}_{st}\Sigma^{12}_{tt}}{\Sigma^{11}_{tt}}}{\sqrt{\left(\Sigma^{11}_{ss} - \frac{(\Sigma^{11}_{st})^2}{\Sigma^{11}_{tt}}\right)\left(\Sigma^{22}_{tt} - \frac{(\Sigma^{12}_{tt})^2}{\Sigma^{11}_{tt}}\right)}} = \frac{f_n}{f_d}.$$

Then,

$$\nabla f = \left(\frac{\partial f}{\partial \Sigma^{12}_{st}}, \quad \frac{\partial f}{\partial \Sigma^{11}_{st}}, \quad \frac{\partial f}{\partial \Sigma^{11}_{ss}}, \quad \frac{\partial f}{\partial \Sigma^{11}_{tt}}, \quad \frac{\partial f}{\partial \Sigma^{12}_{tt}}, \quad \frac{\partial f}{\partial \Sigma^{rr}_{tt}}\right)'$$

$$= \left(\frac{\Sigma^{11}_{tt}}{f_d}, \frac{-\Sigma^{12}_{tt} * f_d + \Sigma^{11}_{st} * f}{f_d^2}, -\frac{1}{2}\frac{\Sigma^{11}_{tt} * f}{\Sigma^{11}_{tt}\Sigma^{11}_{ss} - (\Sigma^{11}_{st})^2}, \frac{\Sigma^{12}_{st}}{f_d} - \frac{f}{2}\frac{2\Sigma^{11}_{tt}\Sigma^{11}_{ss}\Sigma^{22}_{tt} - \Sigma^{22}_{tt}(\Sigma^{11}_{st})^2 - \Sigma^{11}_{ss}(\Sigma^{12}_{tt})^2}{f_d^2}\right.$$

$$\left.\frac{-\Sigma^{11}_{st}}{f_d} + \frac{\left(\Sigma^{11}_{ss}\Sigma^{11}_{tt} - (\Sigma^{11}_{st})^2\right)\Sigma^{12}_{tt} f}{f_d^2}, -\frac{1}{2}\frac{\Sigma^{11}_{tt} f}{\Sigma^{11}_{tt}\Sigma^{22}_{tt} - (\Sigma^{12}_{tt})^2}\right)'.$$

Then by the Delta method,

$$\text{Var}(f) = \nabla f \text{Cov} \begin{pmatrix} \Sigma_{st}^{12} \\ \Sigma_{st}^{11} \\ \Sigma_{ss}^{11} \\ \Sigma_{tt}^{11} \\ \Sigma_{tt}^{12} \\ \Sigma_{tt}^{22} \end{pmatrix} \nabla f^T.$$

We now describe how to estimate the covariance matrix above. Note that each element of this matrix is of the form

$$\text{Cov}(\hat{\Sigma}_{t_1t_2}^{a_1a_2}, \hat{\Sigma}_{t_3t_4}^{a_3a_4}) \tag{5.31}$$

Recall that we obtain each estimate directly as follows

$$\widehat{\text{Cov}}(X_{a_1}(t_1), X_{a_2}(t_2)) = \hat{\Sigma}_{t_1t_2}^{a_1a_2} = \hat{\boldsymbol{\beta}}|_1 \text{ where } \hat{\boldsymbol{\beta}} = \Sigma_{i=1}^n \mathbb{A}_i(t_1, t_2) \underset{\sim}{\mathbf{Y}}_i(a_1, a_2) = \mathbb{A} \underset{\sim}{\mathbf{Y}}$$

$$\text{and let} \qquad \hat{\Sigma}_{t_3t_4}^{a_3a_4} = \hat{\boldsymbol{\beta}}'|_1 \text{ where } \hat{\boldsymbol{\beta}}' = \Sigma_{i=1}^n \mathbb{A}_i(t_3, t_4) \underset{\sim}{\mathbf{Y}}_i(a_3, a_4) = \mathbb{A}' \underset{\sim}{\mathbf{Y}}',$$

which gives

$$\text{Cov}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}}') = \text{Cov}(\mathbb{A} \underset{\sim}{\mathbf{Y}}, \mathbb{A}' \underset{\sim}{\mathbf{Y}}') = \mathbb{A} \text{Cov}(\underset{\sim}{\mathbf{Y}}, \underset{\sim}{\mathbf{Y}}') \mathbb{A}'^T = \Sigma_{i=1}^n \mathbb{A}_i \text{Cov}(\underset{\sim}{\mathbf{Y}}_i, \underset{\sim}{\mathbf{Y}}_i') \mathbb{A}_i'^T.$$

In obtaining $\text{Cov}(\underset{\sim}{\mathbf{Y}}_i, \underset{\sim}{\mathbf{Y}}_i')$, it is important to note that $\mathbf{Y}_i = \mathbf{Y}_i(a_1, a_2)$ and $\mathbf{Y}_i' = \mathbf{Y}_i'(a_3, a_4)$ may be of different lengths. Regardless, the cross-covariance matrix will contain higher order moments as it consists of terms of the form

$$\begin{aligned} &\text{Cov}(Y_{ia_1k_1} Y_{ia_2k_2}, Y_{ia_3k_3} Y_{ia_4k_4}) \\ &\quad = \mathbb{E}(Y_{ia_1k_1} Y_{ia_2k_2} Y_{ia_3k_3} Y_{ia_4k_4}) - [\Sigma_{t_{ik_1}t_{ik_2}}^{a_1a_2}][\Sigma_{t_{ik_3}t_{ik_4}}^{a_3a_4}] \\ &\quad = [\Sigma_{t_{ik_1}t_{ik_3}}^{a_1a_3}][\Sigma_{t_{ik_2}t_{ik_4}}^{a_2a_4}] + [\Sigma_{t_{ik_1}t_{ik_4}}^{a_1a_4}][\Sigma_{t_{ik_2}t_{ik_3}}^{a_2a_3}] \end{aligned}$$

5.5 Simulation study

In this section, we evaluate the statistical performance of the marginal and the conditional covariance estimates obtained using technique 1 (section 5.4.2). We assess the performance for non-separable stationary data, both with and without missingness. For all the estimates, we provide

plots of the root mean squared error (RMSE) and the standard deviation, as the number of subjects increases. We are in the process of implementing technique 2 (section 5.4.1) for covariance estimation and therefore have no simulation results for it yet.

5.5.1 Data Generation

We start by describing the data-generating model used for this simulation study. We aim to mimic the main features of the Dogon Longitudinal Study. For simplicity, we only consider gridded data, that is, we take all individuals to be observed at the same time points. Under this gridded setting, we generate non-separable stationary Gaussian data as described below.

Optimization of non-separable covariance in simulation

Recall from Section 5.1.6 that a separable covariance function results in a trivial conditional correlation of interest (5.2), i.e. $\text{Cor}^{12}_{s_1t_2|t_1} = 0$. Since we'd like to evaluate our models ability to estimate non-zero conditional covariances, the underlying convariance function must be non-separable. To construct a non-separable covariance matrix for our simulation, we take the simplest case of a Kronecker sum with $K = 2$ (notation as in Section 5.1.6). We consider a stationary process and take A_k to be an autoregressive matrix with correlation sequence $\eta^{(k)}$, i.e. $A_k(l_1, l_2) = \eta^{(k)}_{|l_1-l_2|}$. In practice, since simulating data corresponding to $u_k = \eta^{(k)} < 0$ is tricky, we instead optimize the problem within the following framework,

$$0 \leq \eta^{(1)}_{|t-s|} \leq 1, \eta^{(2)} = 0 \text{ and } \theta^{(2)} = -\theta^{(1)} \implies u = \begin{bmatrix} \eta_{|t-s|} & 0 \end{bmatrix}, v = \begin{bmatrix} \theta & -\theta \end{bmatrix}.$$

Recall that we derived a closed form for the conditional correlation of interest, $\text{Cor}^{12}_{s_1t_2|t_1}$ in (5.3). Under the given framework, the expression for this conditional correlation reduces to

$$\text{Cor}^{12}_{s_1t_2|t_1} = \frac{\theta\eta_{|t-s|}}{\sqrt{4 - \eta^2_{|t-s|}}},$$

which attains a maximum of $1/\sqrt{(3)} = 0.58$. To see this, note that $|\text{Cor}^{12}_{s_1t_2|t_1}|$ is an increasing function of both $\eta_{|t-s|}$ and θ. Therefore, $\text{Cor}^{12}_{s_1t_2|t_1}$ has a maximum magnitude of $\pm 1/\sqrt{(3)} = 0.58$ corresponding to $\max \eta_{|t-s|} = 1$ and $\max |\theta| = 1 \implies \theta = \pm 1$. Since we only expect the conditional correlation in our application to be at most in the range of 0.1 to 0.3, we are satisfied with doing simulations and checking the performance of our method on conditional correlations ranging between 0 and 0.52.

Particulars of data generation

Starting with the time grid, we normalize the time points to lie between 0 and 1, and consider the time grid $0, 0.1, \ldots, 0.9, 1$ for all individuals. Since we are interested in two-dimensional Gaussian data, each individual has 11 2-dimensional observations which are generated as an instance of a 22-dimensional multivariate Gaussian variable with mean 0 and a non-separable covariance matrix Σ, taken as a 2-Kronecker sum, as described above. We conduct simulations for the conditional correlation of interest (5.2), i.e. $\text{Cor}^{12}_{s_1 t_2|s_1}$ being $0, 0.1, 0.2, 0.3, 0.4$ and 0.51, for fixed $s = 0.2$ and $t = 0.6$. Here we present results only for $\text{Cor}^{12}_{0.2_1 0.6_2|0.6_1} = 0.51$. Results for the other correlations are similar.

The true underlying marginal and conditional correlation matrices for the simulated data are

$$\Sigma_{0.2,0.6} = \begin{bmatrix} 0.95 & 0.90 \\ 0.90 & 0.95 \end{bmatrix} \text{ and } \text{Cor}_{0.2_1,0.6_2|0.6_1} = \begin{bmatrix} 1 & 0.51 \\ 0.51 & 1 \end{bmatrix}. \tag{5.32}$$

Note that here $\Sigma_{0.2,0.6}$ is symmetric, but this need not be the case and we make no such assumptions anywhere.

5.5.2 Simulation design

We fix two time points, $s = 0.2, t = 0.6$ and estimate the marginal covariances $\hat{\Sigma}_{0.2,0.6} = \widehat{\text{Cov}}(0.2, 0.6)$, $\hat{\Sigma}_{0.2,0.2}$ and $\hat{\Sigma}_{0.6,0.6}$ along with the conditional correlation matrix $\widehat{\text{Cor}}_{0.2_1,0.6_2|0.6_1}$ for non-separable stationary data, as described above, both with and without missingness. The estimates are averaged over 500 Monte Carlo replicates per data setting. We present plots for the root MSE, the bias and the standard deviation of the estimates as the number of subjects increases from 200 to 1000 along $[200, 400, 800, 1000]$, only under the missing data setting.

Model parameters

We only present results for local linear regression, using a fixed bandwidth of $h = 0.05$, which corresponds to 1 year in the application of interest where the data span 20 years. We use the squared exponential (Gaussian) kernel to assign weights in the local regression.

Induced missingness pattern

The pattern of missingness mimics that in the Dogon Longitudinal Study. We assume that all entries for the second attribute of the Gaussian process are missing at time points less than 0.5. That is, $Y_2(t)$ is missing for $t < 0.5$. Again, this is similar to our motivating dataset wherein blood pressure (SBP) is not available for subjects until age 11.

5.5.3 Simulation results

Marginal estimates

For marginal covariances, it is important to note that there are essentially four covariances we're trying to estimate, (a) between-time within attribute (of the form $\text{Cov}(Y_1(s), Y_1(t))$), (b) within time between attribute (of the form $\text{Cov}(Y_1(s), Y_2(s))$), (c) between time between attribute (of the form $\text{Cov}(Y_1(s), Y_2(t))$) and finally (d) the variance (of the form $\text{Cov}(Y_1(s), Y_1(s))$).

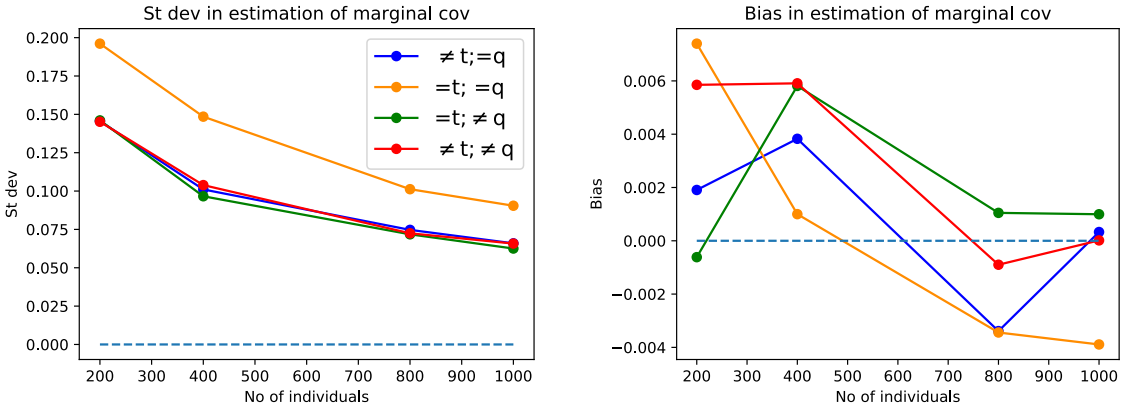
In particular, the 4 marginal estimates in figure 5.4 corresponding to $s = 0.2$ and $t = 0.6$ are:

(a) Between time within attribute [$\neq t; = q$]: this is taken as the average between estimates of $\text{Cov}(Y_1(0.2), Y_1(0.6))$ and $\text{Cov}(Y_2(0.2), Y_2(0.6))$.

(b) Within time between attribute [$= t; \neq q$]: this is taken as the average between estimates of $\text{Cov}(Y_1(0.2), Y_2(0.2))$ and $\text{Cov}(Y_1(0.6), Y_2(0.6))$.

(c) Between time between attribute[$\neq t; \neq q$]: this is taken as the average between estimates of $\text{Cov}(Y_1(0.2), Y_2(0.6))$ and $\text{Cov}(Y_1(0.6), Y_2(0.2))$.

(d) Variance (within type within attribute) [$= t; = q$]: this is taken as the average between estimates of $\text{Cov}(Y_1(0.2), Y_1(0.2))$ and $\text{Cov}(Y_2(0.6), Y_2(0.6))$.



(a) (b)

Figure 5.4: Standard deviation and bias in the 4 types of marginal covariance estimates corresponding to s=0.2, t=0.6.

We notice here that the marginal estimates simultaneously achieve quite low standard deviation ($\sim 10\%$) and bias ($\sim 0.4\%$). Note that since the true marginal values to be estimated are $0.95, 2, 0$ and 0.9, the bias is almost negligible and the standard deviation, and hence the Root MSE, are roughly between $5 - 10\%$ for 1000 subjects. Thus, the method seems to perform well for marginal

covariance estimates.

Conditional estimates

For conditional covariances, there are 3 different types of covariances we're trying to estimate as well; these correspond to the 3 different elements of the conditional covariance matrix (ref to RHS of (5.21)). Since our primary interest is in the off-diagonal term of the correlation matrix, we discuss the estimates for that in some detail here.

We start with the estimate for the conditional correlation of interest, illustrated in figure 5.5. Fig 5.5a compares the Root MSE of the conditional correlation estimates obtained using our technique (labelled ker_LM) and using the standard method of moments. Since the conditional correlation involves both time points that are observed ($s, t = 0.2, 0.6 \in [0, 0.1, 0.2, \cdots, 1]$ which is the grid of 11 time points used to simulate the data), the empirical correlation can be calculated by simply considering the data at these two time-points. This is labelled as *MOM* in the plot. The RMSE of both the kernel and MOM estimates are presented for the true underlying conditional correlation, $\text{Cor}^{12}_{0.2_1 0.6_2|0.6_1}$ being -0.1 and 0.51. Fig 5.5b provides the standard deviation of the same set of estimates. There plots indicate that the estimates obtained using our method perform atleast as well as the naive method of moments estimates.

Figures 5.5c and 5.5d show the distribution of the estimates corresponding to an underlying conditional correlation of -0.1 and 0.5 respectively. We clearly see that the estimates converge to the truth as the number of subjects increases.

Note that the conditional correlation of interest here is a non-linear function of the elements of the conditional covariance matrix, which means it is fairly complex to estimate. We take one step back and look at the estimates for the conditional covariance, which by itself is a non-linear function of 4 marginal covariances. Fig 5.6 presents the element-wise RMSE for the three elements of the conditional covariance matrix, $\hat{\Sigma}_{0.2_1 0.6_2|0.6_1}$. Since the underlying covariance matrix corresponding to a cross conditional correlation of 0.52 is

$$\Sigma_{0.2_1 0.6_2|0.6_1} = \begin{bmatrix} 1.59 & 0.90 \\ 0.90 & 2.00 \end{bmatrix},$$

we note that the relative RMSEs for the diagonal terms, columns 2 and 3 of figure 5.6, reduce by half upon taking the square root, which is what is used in the conditional correlation estimation (they reduce from $4.7, 5.3\%$ to $2.2 - 2.5\%$). This illuminates that it is okay for the relative RMSE of certain terms of the conditional covariance matrix to be higher than those of the conditional correlation of interest, even though the latter is a slightly complicated (non-linear) function of the former.

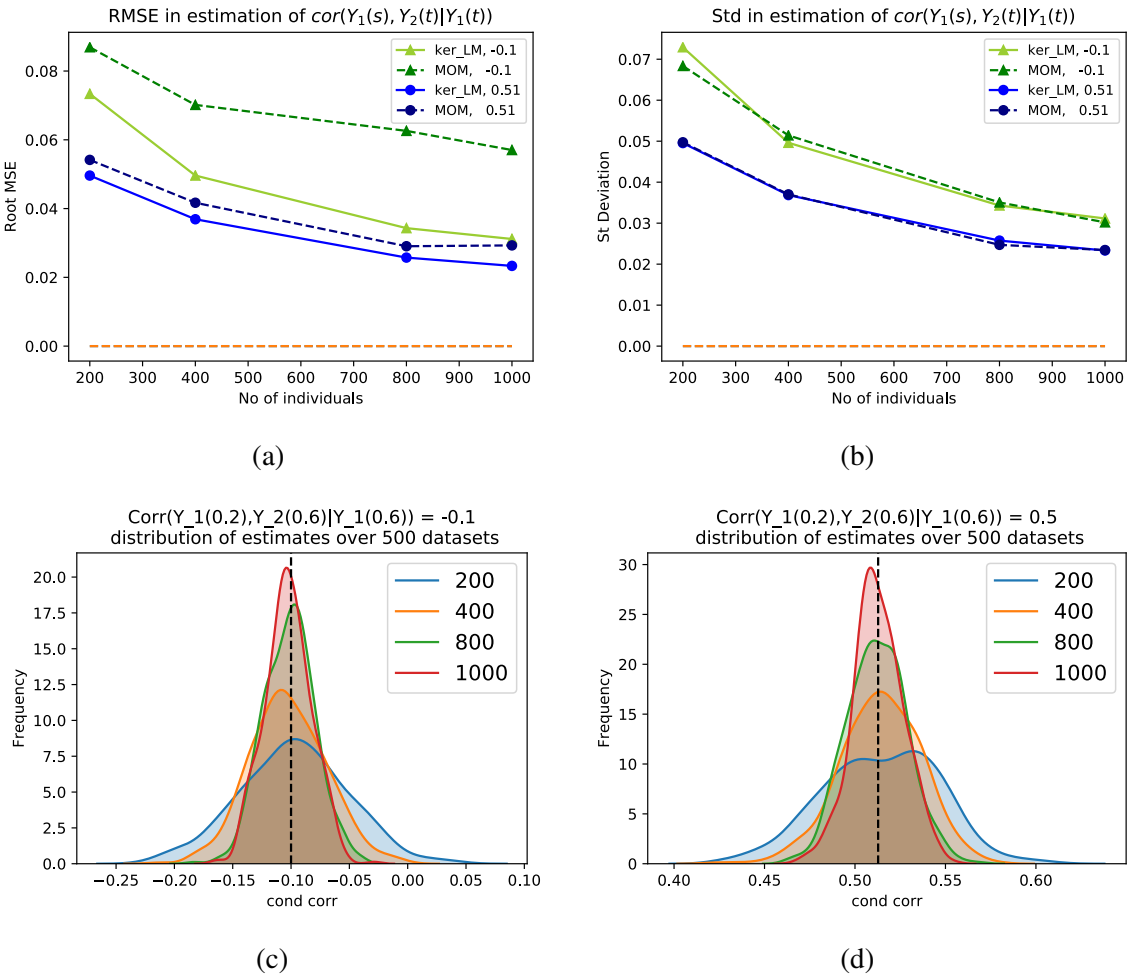


Figure 5.5: Conditional correlation of interest, $\widehat{\text{Cor}}^{12}_{0.2_1 0.6_2|0.6_1}$.

Overall, we conclude that the method performs well on the scale of the motivating dataset. That is, it works well as long as the number of observations per individual comply with the bound $\max_i n_i^4 < 10,000$. For datasets of larger scale, while the covariance estimates can be obtained, the standard errors become infeasible to calculate. This is because the problem blows up (recall the discussion on computational complexity from section 5.3.1.1).

5.6 Summary and future directions

In this report we implement a non-parametric technique to estimate the covariance structure of longitudinal data. There are no restrictions imposed on the data or the covariance function to be estimated which make this technique very broad and enables it to model any kind of covariance structure. In particular, we carry out local polynomial regression on products of residuals within

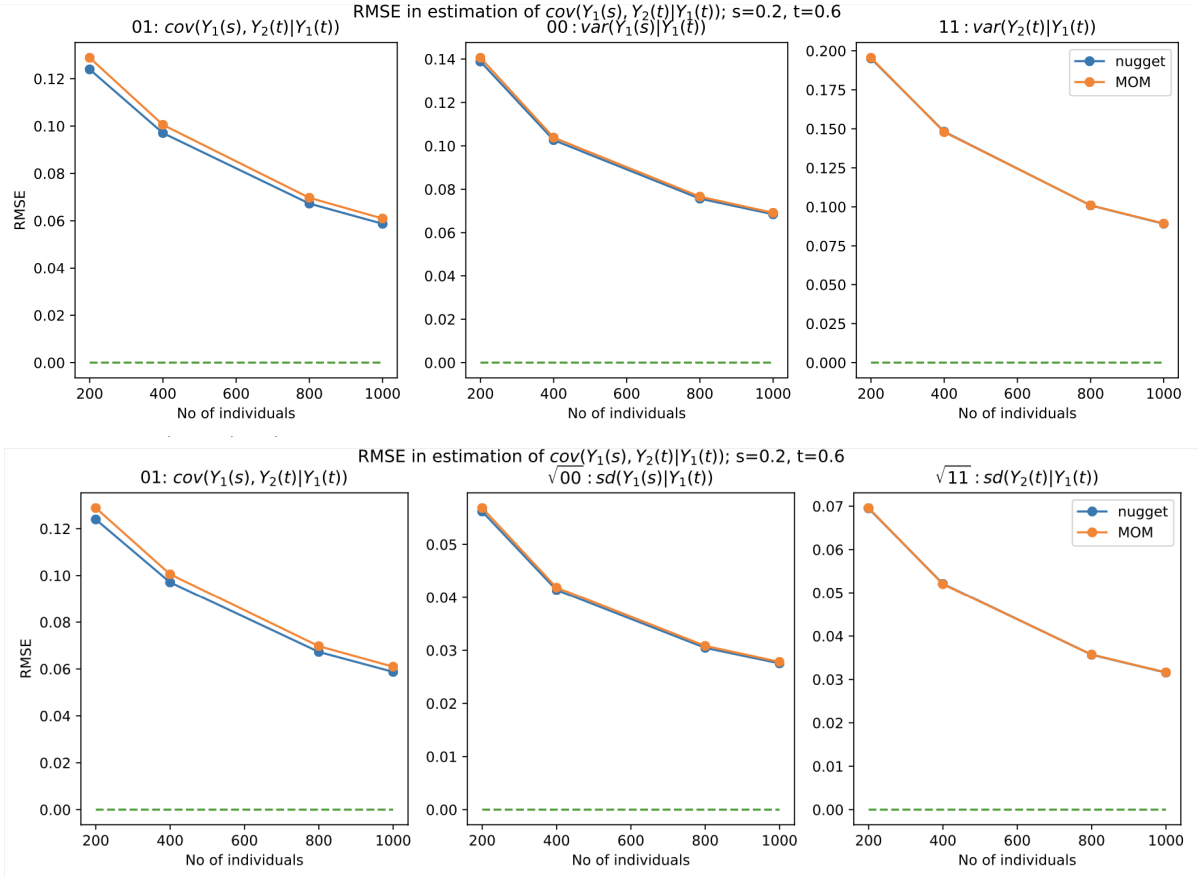


Figure 5.6: Term-wise RMSE for the estimates of the conditional covariance matrix of interest $\hat{\Sigma}_{0.2_1 0.6_2|0.6_1}$ in row 1. Row 2 contains the RMSE of the estimates of the square of the diagonal terms in the conditional covariance matrix.

each individual to model the covariance structure. Under the Gaussianity assumption, we extend this technique to model the conditional covariance matrix in longitudinal data and propose a plug in inference procedure for both the marginal and conditional estimates. We provide simulation studies to assess the performance of the estimates and finally apply it to a real world dataset to obtain covariance and correlation estimates on the same.

One of the major drawbacks of kernel regression is that it does not give a functional form for the estimate. That is, for each new estimate (at a different time-point) the entire method needs to be refitted at the end of which an estimate is produced. This makes the technique extremely computationally heavy. Thus in the future, we need to re-implement this technique to make it run faster! Also, due to a fixed (global) bandwidth, local polynomial regression can be a bit too local in certain cases, rendering it unable to model a global trend, if one exists. Future work includes exploring bandwidth selection for covariance estimation.

A common difficulty in estimating covariances is the PSD constraint. We impose no such

constraint in the local regression here. This is because we are ultimately interested in modelling the conditional covariance and as long as we can get meaningful estimates for these, we do not mind it if some of the intermediate marginal covariance estimates are non-PSD. Our belief is that, as long as the marginal estimates are consistent and reasonably accurate, they will eventually be PSD as the sample size grows. This is definitely a limitation especially in using this technique for marginal covariance estimation, however the flexibility of the method which gives meaningful estimates is still very attractive and powerful, due to lack of other such broad techniques. Improving upon this and including a PSD constraint is part of future work.

In the inference procedure, we realise that the Normal formulae for relating the fourth order moments to the second order moments may be too stringent, and would like to consider alternatives to this, such as parametrizing this relationship using some broader distribution, perhaps the spherical distribution, and carrying out sensitivity analysis.

Motivated by our data application, we would like to expand the analysis to model the covariance between a linear functional of the variables (height in this case, which captures a persons growth trajectory) instead of computing the correlation at specific, distinct time points. That is, modelling

$$\int_0^{10} \text{Cov}(HT(s), SBP(t)|HT(t))w(s)ds = \text{Cov}(\int_0^{10} HT(s)w(s)ds, SBP(t)|HT(t)).$$

This will be possible, along with exploring bandwidth selection once we have a more efficient implementation of this technique.

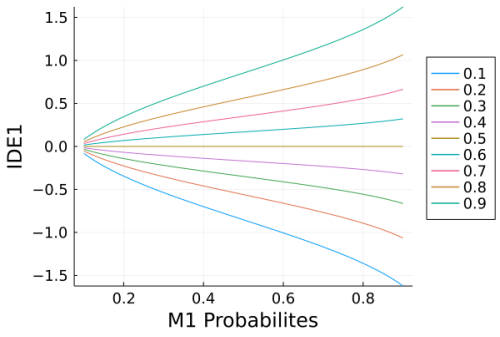
As mentioned in section 2, we would like to model uncertainty due to mean estimation into our covariance estimator. Future work also includes incorporating cluster analysis into this technique to model the heirarchical structure of the underlying data.

APPENDIX A

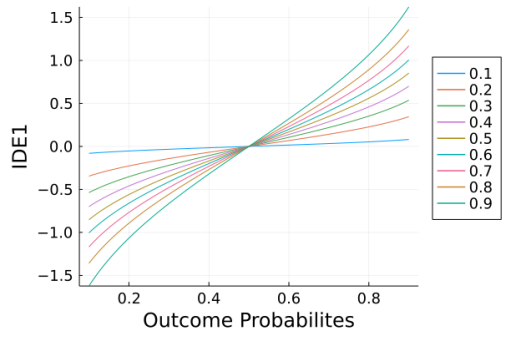
Mediation Simulation

We provide the underlying form of the direct and indirect effects for simulation cases 2 and 3 in the simulation study presented in section 3.3.

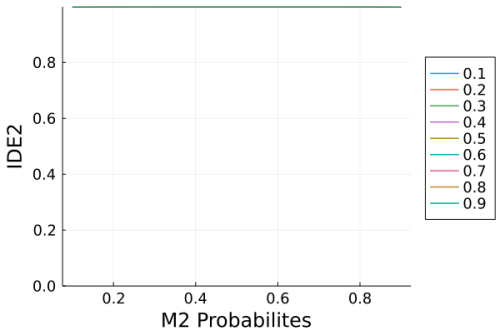
For case 2 (heteroscedastic Gaussian), the direct effect is constant at 1, as derived in section 3.3. Both the indirect effects are presented in figure A.1. The case 3 (exponential), indirect effects 1 and 2 are identical, which is why indirect effect 2 is omitted in figure A.2.



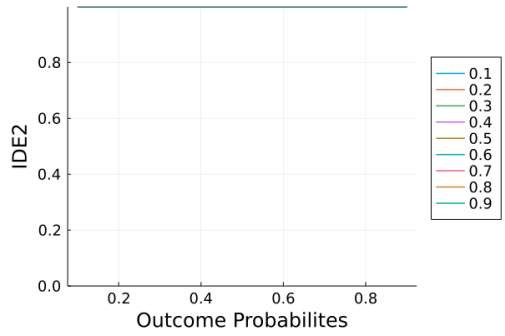
(a) Indirect effect 1



(b) Indirect effect 1

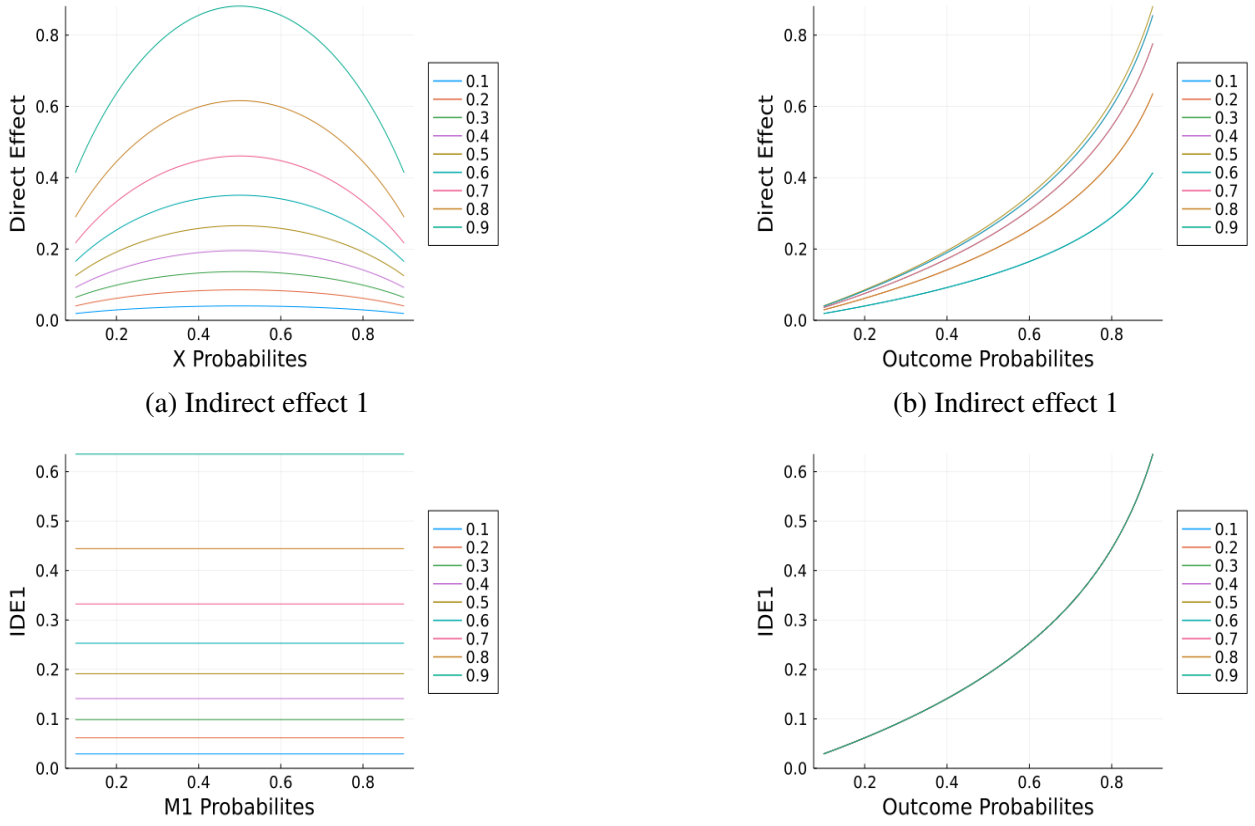


(c) Indirect effect 1



(d) Indirect effect 1

Figure A.1: Underlying values of the indirect effects for case 2.



(a) Indirect effect 1

(b) Indirect effect 1

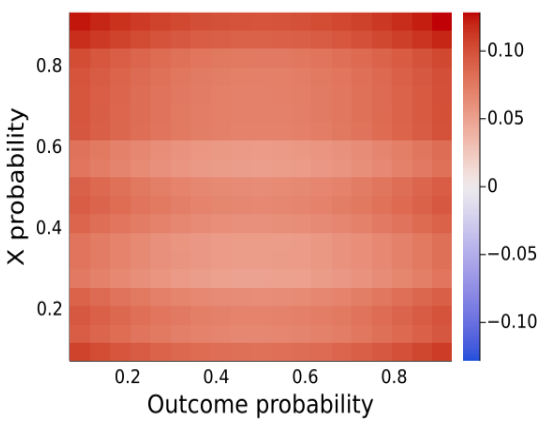
(c) Indirect effect 1

(d) Indirect effect 1

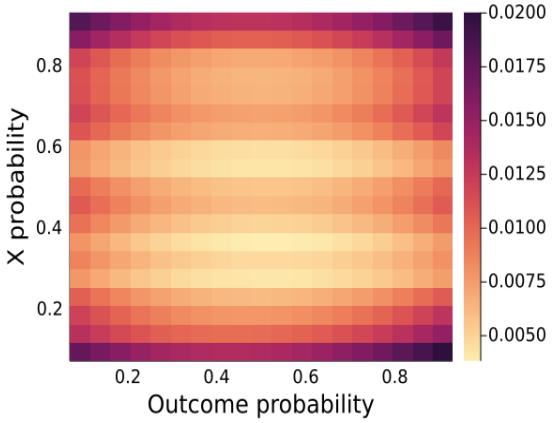
Figure A.2: Underlying values of the direct and indirect effects for case 3.

A.0.1 Pointwise Results

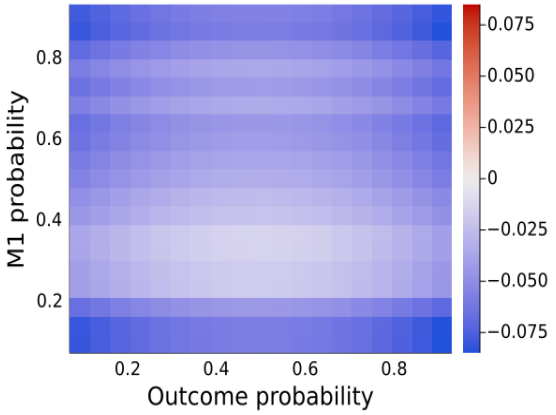
The pointwise estimates for the mediation effects, across all three settings take the form of $m]timesm$ heatmaps which are included below. We do not compute the realative measures (rbias, rrmse) for points which have a true underlying value of less than 0.1, and these points are blank in the heatmaps.



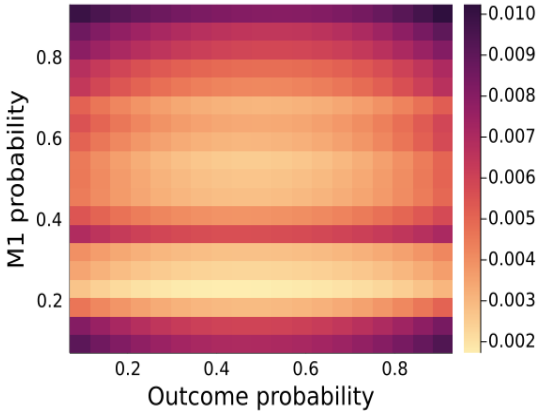
(a) DE : Relative bias



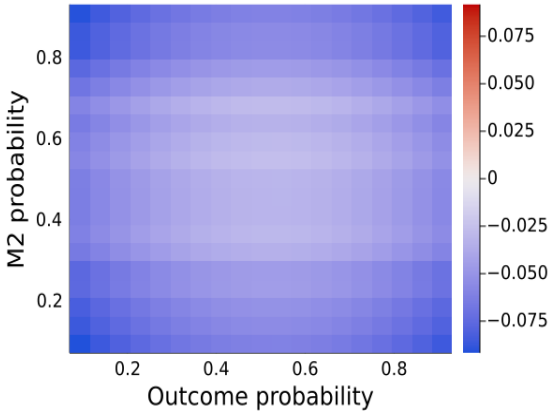
(b) DE : Relative MSE



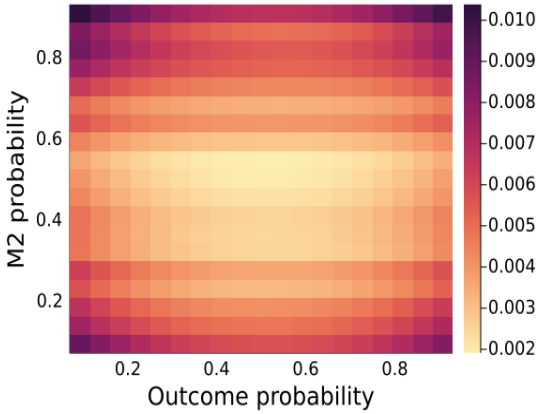
(c) IDE1 : Relative bias



(d) IDE1 : Relative MSE

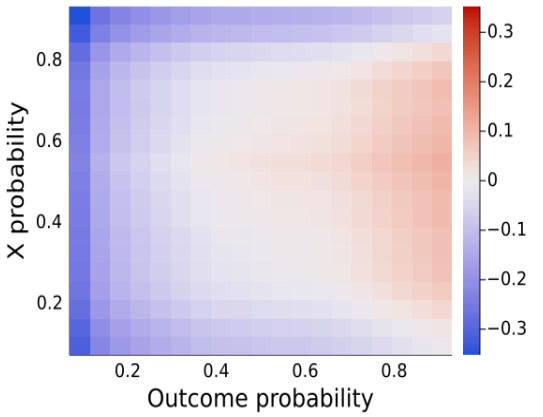


(e) IDE2 : Relative bias

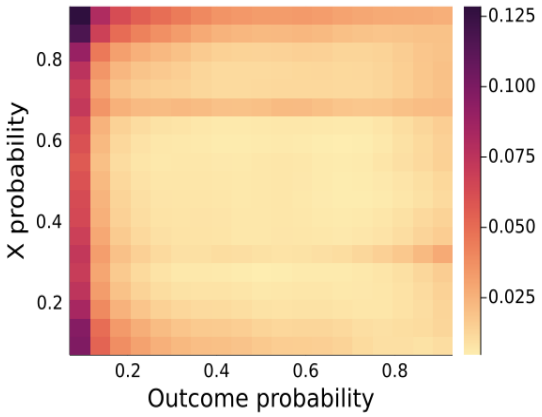


(f) IDE2 : Relative RMSE

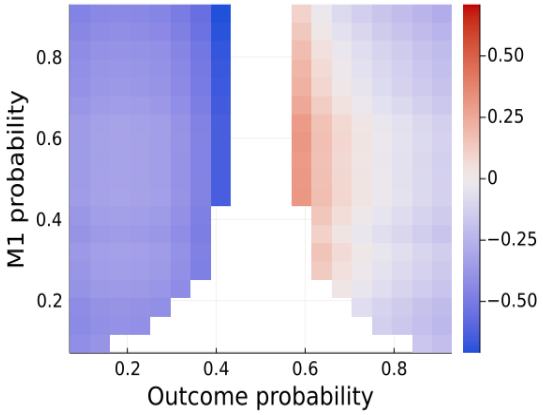
Figure A.3: Pointwise evaluation of the simulation results for the entire mediation pipeline corresponding to simulation case 1.



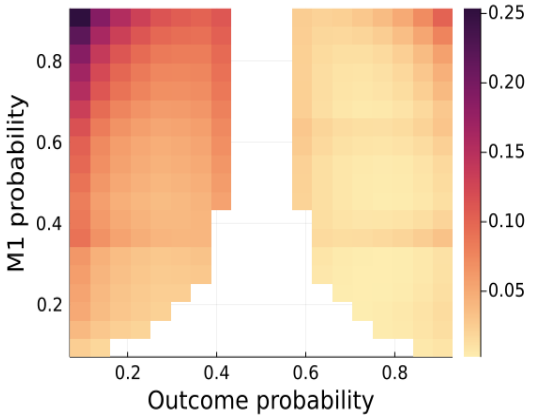
(a) DE : Relative bias



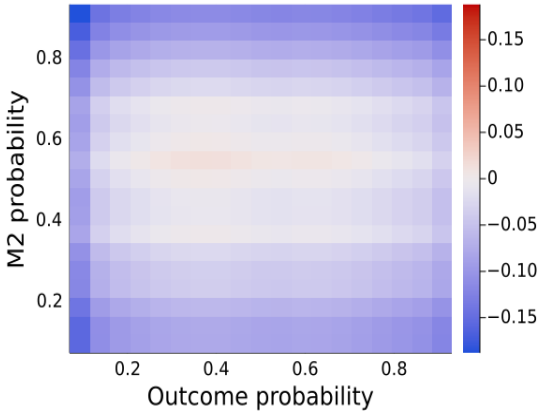
(b) DE : Relative MSE



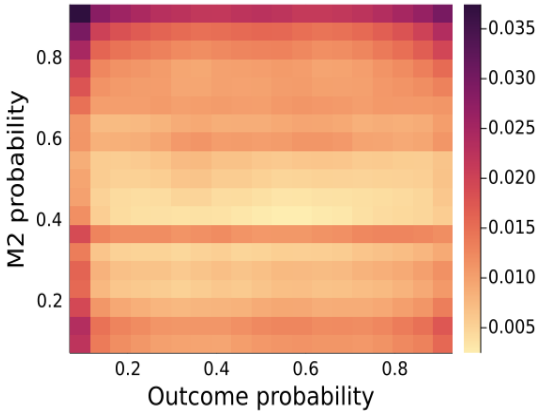
(c) IDE1 : Relative bias



(d) IDE1 : Relative MSE

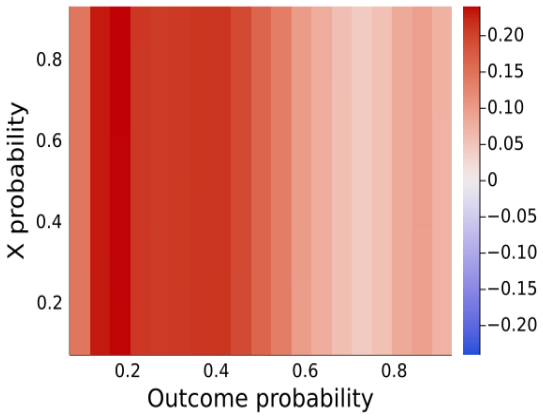


(e) IDE2 : Relative bias

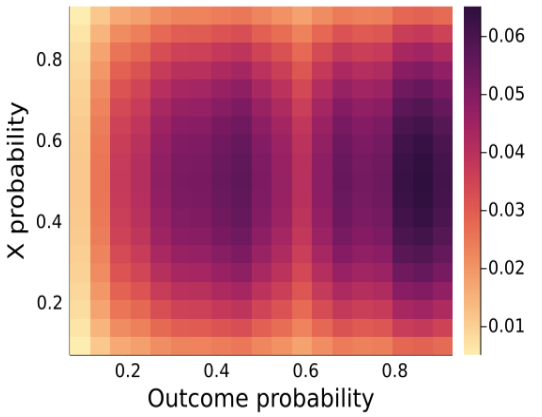


(f) IDE2 : Relative MSE

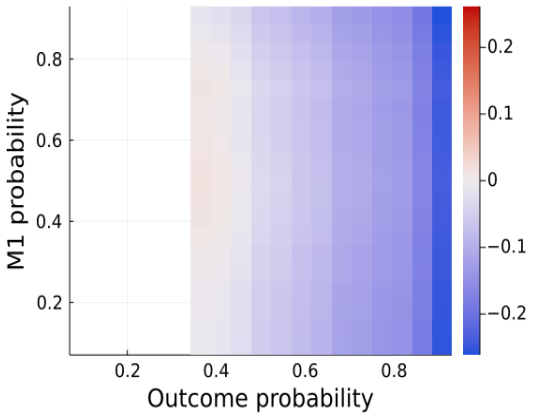
Figure A.4: Pointwise evaluation of the simulation results for the entire mediation pipeline corresponding to simulation case 2.



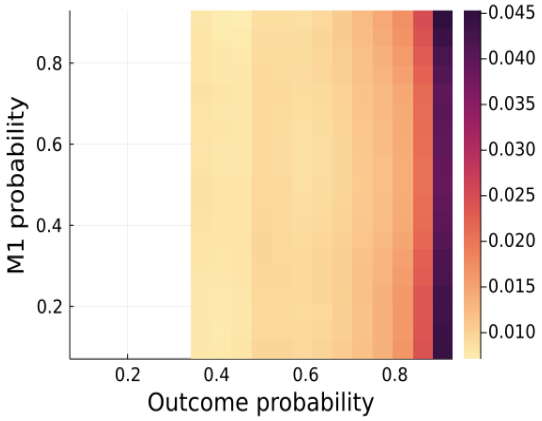
(a) DE : Relative bias



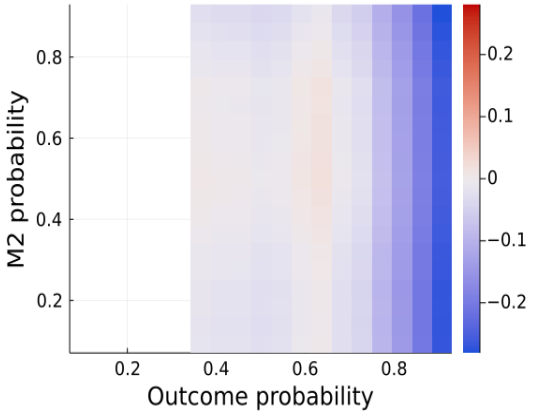
(b) DE : Relative MSE



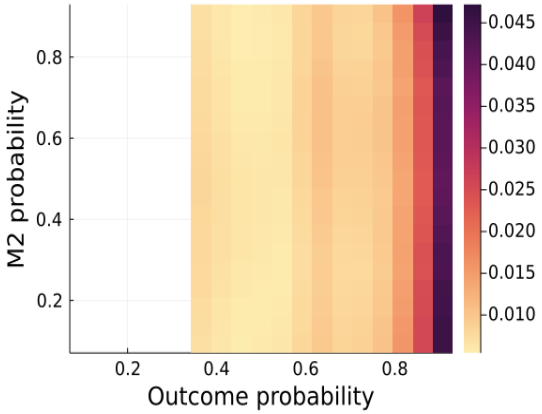
(c) IDE1 : Relative bias



(d) IDE1 : Relative MSE



(e) IDE2 : Relative bias



(f) IDE2 : RelMSE

Figure A.5: Pointwise evaluation of the simulation results for the entire mediation pipeline corresponding to simulation case 3.

In case 1, both the relative bias and RRMSE are less than 15% for all the 3 mediation effects. In case 2, the relative bias and RRMSE for the direct effect are mostly below 20%. They get as high

as 35% in the extreme lower quantile of the the outcome ($p_y = 0.1$) and extreme outer quantiles of the exposure ($p_x = 0.1, 0.9$), but are below 20% everywhere else, including outer quantiles (which are no as extreme) around $p_x, p_y \in [0.2, 0.8]$. For indirect effect 1, we must note that IDE1 consists of smooth functions that cross 0 at $p_y = 0.5$. Due to this, the values of IDE1 around the inner quantiles of the outcome ($p_y \sim 0.5$) are very small. The relative bias is mostly below 20% and the RRMSE is below 30% for the most part of IDE1; both values are high (rbias and RRMSE) around 65% for $p_y = 0.42$. It is important to note that the underlying value of IDE1 in this region is very small and is about 0.15, and so an error of 70% means the average estimate at this point would be around 0.05. We observe similar but less extreme patterns in case 3. IDE2 estimates have below 15% rbias and RRMSE (IDE2 is never equal to 0 in this setup).

In case 3, we see that that the relative bias for the direct effect is always less than 20%. The RRMSE gets as big as 42%, but this is when the underlying value of the direct effect is between 0.1 and 0.2. Note that relative RMSE of around 40% means that on average, the estimated value of 0.15 might be 0.21. The relative errors seem big as the underlying values are so small, but the estimates are infact close to the truth. This is also captured by the Frobenius RRMSE of the estimates presented in table 3.1. The indirect effect estimates are of similar quality. For both IDE1 and IDE2, the relative bias is mostly below 20%; it is between 30-40% at the extreme outer quantiles of both the outcome and the mediator. Due to less data in this region, these estimates are already very difficult to compute.

A.1 Simulation study for conditional quantile models

In this section we conduct a simulation study to assess the performance of the analysis pipeline that uses QNN and FLR to understand the conditional quantile structure of a population. We use this to model the conditional quantiles of $Y|M_1, M_2, X$, and $M_j|X$ in the mediation analysis (ref sections $3.2.2$, $3.2.4$) and the functional components of these models are further used to define the mediation effects. Since we have presented simulation studies to evaluate the performance of QNN in the previous chapter, we do not access the fit of the QNN estimates and instead focus our attention on the results of the FLR, which are the final results from the joint pipeline.

Please note that the simulation study presented in chapter 3 section 3.3 relies on the results of the conditional quantile models, and so it implicitly checks the quantile of the conditional quantile estimates. So the simulation presented here tests 2/3rd of the mediation pipeline. Also note that we already conducted a simulation study for the AFQR model in chapter 2 section $2.3.6$, which assesses the same pipeline. The distinction between that and this is that here we apply AFQR to a tensor of quantiles evaluated on a much sparser grid for the covariates (X, M_1, M_2).

A.1.0.1 Data Generation

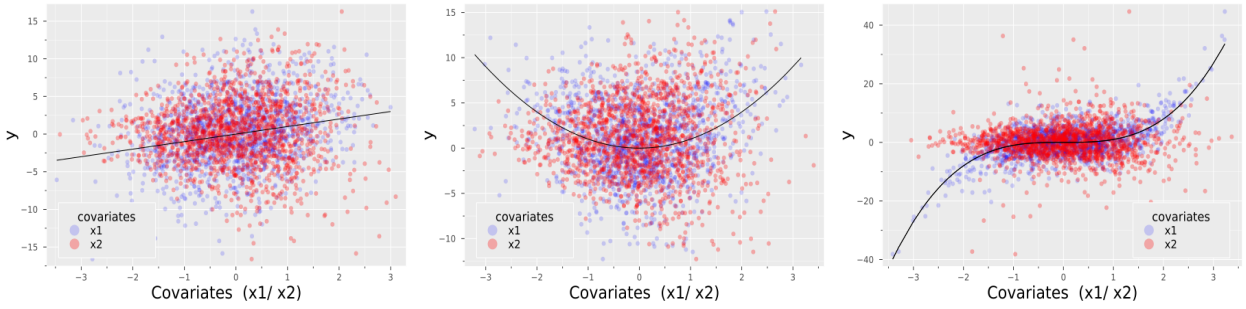
We consider a Gaussian outcome with covariate-dependent mean and variance functions; this has location and scale behaviour. Thus, the generated data will have a homogeneous component determined by the mean structure, and a non-homogeneous component (meaning a non-trivial g function) corresponding to the variable that determines the variance. We will explain this in more detail later. We pick a Gaussian distribution for the outcome since its quantile function has a closed form, and so the components of the low rank decomposition can be derived mathematically.

If the model has p covariates,

$$\begin{aligned} X_j &\sim \mathcal{N}(0,1) & j &= 1, \cdots, p \\ Y &= \mu_s(X_1) + \sigma_s(X_2)\epsilon_Y & \epsilon_Y &\sim \mathcal{N}(0,1) \\ \mu_s(x_1) &= x_1^q & q &\in \{1,2,3\} \\ \sigma_s(x_2) &= \sqrt{(4+x_2)^2} \end{aligned}$$

In this case, the conditional quantile function has a closed form given by

$$Q_{Y|\boldsymbol{X}}(p|X_1 = x_1, X_2 = x_2, \cdots, X_p = x_p) = \mu(x) + \sigma(x_2) \cdot Q_{\mathcal{N}(0,1)}(p).$$



(a) $\mu_s(x_1) = x_1$ (b) $\mu_s(x_1) = x_1^2$ (c) $\mu_s(x_1) = x_1^3$

Figure A.6: Simulated data corresponding to $q = 1, 2, 3$

A.1.0.2 Simulation Design

We generate data with $q = 2$ and $q = 4$ covariates. The number of observations per setting is 1500. For each number of covariates, we vary the mean function μ_s which is a monomial of X_1, by varying its degree along $[1, 2, 3]$. Per setting, we use the setup described above to randomly sample 10 iid copies of $(X_1, \cdots, X_q, Y)$. As in previous sections, let $\boldsymbol{X} = (X_1, \cdots, X_q)$.

We evaluate the conditional quantiles along a probability grid of length $m = 11$. This is done so that that the median value corresponds to an entry on the probability grid, and does not have to be evaluated as the mean of two estimates. The probability grid, $(p) = [0.1, \cdots, 0.9]$. We use the same smoothing penalty for all $\boldsymbol{u}_j$ and all $\boldsymbol{v}_j$, ie $c_{u_j} = 1, c_{v_j} = 1, \ \forall j$.

A.1.0.3 True parameters

In this section we give the form of μ, h_j, g_j in the low rank decomposition as per the data generating model. Recall that the quantile function takes the form

$$Q_{Y|\boldsymbol{X}}(p|X_1 = x_1, \cdots, X_q = x_q) \quad = \quad \mu_s(x) + \sigma_s(x_2) \cdot \Phi^{-1}(p).$$

The central axis is defined as the conditional quantile function evaluated at the median value of all the covariates. Since $X_j \sim \mathcal{N}(0,1), \ \text{median}(X_j) \quad = \quad 0$. And so

$$\mu(p) = Q_{Y|\boldsymbol{X}}\big(p|\text{median}(\boldsymbol{X})\big) \quad = \quad \mu_s(0) + \sigma_s(0) \cdot \Phi^{-1}(p) \quad = \quad 4\Phi^{-1}(p).$$

So the centered conditional quantile function is of the form

$$Q^c_{Y|\boldsymbol{X}}(p|x_1, \cdots, x_q) = \mu_s(x_1) + \big(\sigma_s(x_2) - 4\big)\Phi^{-1}(p).$$

The low rank functional components are then

$$\begin{aligned} &f_1(x_1) = \mu_s(x_1) && g_1(p) \equiv 1 && \\ &f_2(x_2) = \sigma_s(x_2) - 4 && g_2(p) = \Phi^{-1}(p) && \\ &f_j \equiv 0 && g_j \equiv 0 && \forall j > 2. \end{aligned}$$

Since $V_\cdot$'s are constraint to be norm 1, the low rank discretized components $U_\cdot, V_\cdot \in \mathbb{R}^m$ take the following form

$$\begin{aligned} &U_1 = \mu_s\big(Q_{X_1}(\boldsymbol{p})\big) \cdot \sqrt{m} && V_1 = \mathbf{1}_m/\sqrt{m}, \\ &U_2 = \sigma_s\big(Q_{X_2}(\boldsymbol{p}) - 4\big) \cdot \|\Phi^{-1}(\boldsymbol{p})\|_2 && V_2 = \frac{\Phi^{-1}(\boldsymbol{p})}{\|\Phi^{-1}(\boldsymbol{p})\|_2}, \\ &\text{where } Q_{X_\cdot}(\boldsymbol{p}) = \Phi^{-1}(\boldsymbol{p}). \end{aligned}$$

A.1.0.4 Simulation Results

As mentioned earlier, we are able to get both location and scale effects here. The first dimension in the additive low rank model is homogeneous across quantiles. This case tests our methods ability

to obtain good quality estimates at the outer quantiles, and explores if our method has any inherent bias at the outer quantiles. This is possible as there is limited data in the outer tails, and so the model structure could impose bias. The second dimension of the additive low rank form has a non-trivial g_2 function. This tests our methods ability to detect non-homogeneous effects.

We present the results for the simulation study in table A.1. For each Monte Carlo replicate, we get an estimate of the central axis μ, and the low rank components f_j, g_j where $j = 1, \cdots, q$ and $q \in \{2, 4\}$. For the central axis, we consider the relative integrated root mean squared error (RRMSE) over the grid of probabilites $\boldsymbol{p}$, and present the mean and the standard deviation of the same over the MC replicates. If $\hat{\mu}_j(\boldsymbol{p}) \in \mathbb{R}^m$ denotes the m-vector estimate of the central axis for the i^{th} replicate, then we define RRMSE of the central axis as

$$\text{RRMSE}_i = \frac{\sqrt{mean\big[\big(\hat{\mu}_i(\boldsymbol{p}) - \mu(\boldsymbol{p})\big)^2\big]}}{\|\mu(\boldsymbol{p})\|_2} = \frac{\sqrt{\frac{1}{m}\sum_{k=1}^{m}\big(\hat{\mu}_i(p_k) - \mu(p_k)\big)^2}}{\sum_{k=1}^{m}\mu(p_k)^2}.$$

For the score and loading vector estimates over discrete probability grid $\boldsymbol{p}$, $\boldsymbol{u}_j = \hat{h}_j, \boldsymbol{v}_j = \hat{g}_j$ we consider the RRMSE of the outer product - $f_j(\boldsymbol{p}) \otimes g_j(\boldsymbol{p})'$ which is an estimate for the slice of the tensor of quantiles along the j^{th} axis and the last axis, while all the other axes are held at their $\rfloor m/2 = 5^{th}$ entries (corresponding to median of the variables along those dimensions). This is the estimated surface for the conditional quantile function $Q_{Y|\boldsymbol{X}}\big(p|X_j, \boldsymbol{X}_{-\boldsymbol{j}} = median(\boldsymbol{x}_{-\boldsymbol{j}})\big)$. We evaluate the outer product of the components h_j, g_j instead of individually evaluating them since they are sign invariant.

For the outer product per covariate, we compute the RRMSE of the vectorized matrix per iteration, similar to the definition for the central axis above. We present the average value and the standard deviation of the estimated RRMSE measures. Note that we can do this for a small number of Monte Carlo replicates, since each RRMSE measure in itself depends on either m (for μ), or m^2 (for $f_j \otimes g_j$) estimates. In addition to this, we access the similarity in the shape of the vectorized matrix curves by reporting the correlation between the vectorized estimated outer products and the underlying value. The results are summarized in the table A.1.

		μ		$f_1 \otimes g_1$				$f_2 \otimes g_2$			
		RRMSE		Correlation		RRMSE		Correlation		RRMSE	
q	p	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
1	2	0.086	0.031	0.983	0.007	0.009	0.006	0.928	0.059	0.0027	0.0030
1	4	0.064	0.026	0.952	0.014	0.005	0.004	0.887	0.031	0.0007	0.0011
2	2	0.083	0.034	0.808	0.285	0.015	0.020	0.972	0.016	0.0006	0.0006
2	4	0.064	0.026	0.714	0.098	0.043	0.007	0.815	0.094	0.0021	0.0027
3	2	0.097	0.042	0.924	0.003	0.004	0.004	0.916	0.105	0.0051	0.0113
3	4	0.056	0.015	0.896	0.005	0.001	0.001	0.888	0.049	0.0012	0.0012

Table A.1: Results of simulation study for tensor functional low rank decomposition

The average RRMSE is below 10% for the central axis, below 5% for the first dimension which determines the mean structure, and below 0.5% for the second dimension which determines the variance structure. The correlation is above 70% for the first dimension and above 80% for the second dimension. Overall for the mean and the first dimension, the average RRMSE is greater than its SD. This shows that the variance of the FLR estimates is very low, which is exactly what we expect.

APPENDIX B

Gaussian Process Emulator Fit

We fit the Gaussian Process Emulator to the Dogon data for varying settings. Specifically, we explore the fit for Males with SBP as the outcome, and height (HT) and weight (WT) as the independent variables. Note that the other 4 measures of childhood body size considered by us throughout this document are BMI, HAZ, WAZ, and BAZ, and these are all functions of HT and WT. Thus we can simulate data for any combination of features among males using the model discussed here.

Recall that the emulator has four parameters, each of which may consists of a basis of potentially time-varying features. In our models, we use a basis of dimension nine for the mean parameter; the basis consists of six functions of age – a constant and linear term, along with a radial basis of dimension five. The scale and smoothing parameters consists of a basis of dimension two, containing only the constant and linear terms in age, and the white noise parameter only consists of a linear term in its basis. The fitted coefficients are printed in tables B.1, B.2. Note that the number of mean parameters increase as the models condition on more independent variables. For instance, the model $\mathrm{P}(\mathrm{WT}|(\mathrm{HT}, \mathrm{Age}))$ has nine instead of seven mean basis functions (refer to table B.2). This is because we add two interactions to the basis for the mean parameter with every additional covariate that is conditioned on in the model; the observed value of the variables being conditioned on are interacted with the constant and linear terms of age in the original basis. So in this case, the original model is $\mathrm{P}(\mathrm{HT}|\mathrm{Age})$ with a mean basis of dimension seven, and the next model $\mathrm{P}(\mathrm{WT}|(\mathrm{HT}, \mathrm{Age}))$ add two terms to the mean basis, observed HT and HT$\times$ Age. The model corresponding to $\mathrm{P}(\mathrm{SBP}|(\mathrm{WT}, \mathrm{HT}, \mathrm{Age}))$ will further add WT and WT$\times$Age to the mean basis making its dimension eleven (refer to table).

Ten simulated copies for afew randomly sampled individuals are presented in figure B.1. As we can see, our estimates are close to the population mean, and seem to be of good quality.

Parameter	Coef	Std. Error	z	Pr($>\|z\|$)	Lower 95%	Upper 95%
mean_1	111.08	1.27	87.52	$< 10^{-99}$	111.08	111.08
mean_2	101.48	1.94	52.23	$< 10^{-99}$	101.48	101.48
mean_3	0.65	1.24	0.53	0.59	0.65	0.65
mean_4	20.35	0.94	21.55	$< 10^{-99}$	20.35	20.35
mean_5	25.78	0.89	28.94	$< 10^{-99}$	25.78	25.78
mean_6	28.83	0.92	31.32	$< 10^{-99}$	28.83	28.83
mean_7	14.35	1.01	14.22	$< 10^{-45}$	14.35	14.35
scale_1	1.45	0.11	13.56	$< 10^{-41}$	1.45	1.45
scale_2	0.11	0.04	3.29	0.001	0.11	0.11
smooth_1	-2.83	0.29	-9.71	$< 10^{-21}$	-2.83	-2.83
smooth_2	1.53	0.09	16.82	$< 10^{-62}$	1.53	1.53
unexplained_1	0.02	0.02	1.38	0.167	0.021	0.021

Table B.1: Summary table for the fitted parameters in the model corresponding to $\mathrm{HT}|\mathrm{Age}$

Parameter	Coef	Std. Error	z	Pr($>\|z\|$)	Lower 95%	Upper 95%
mean_1	-0.05	1.054	0.55	0.58	-0.05	-0.052
mean_2	11.50	1.564	-5.32	$< 10^{-6}$	11.50	11.50
mean_3	-0.07	0.701	-17.18	$< 10^{-65}$	-0.07	-0.07
mean_4	-6.71	0.625	-23.23	$< 10^{-99}$	-6.71	-6.71
mean_5	-10.96	0.601	-27.76	$< 10^{-99}$	-10.96	-10.96
mean_6	-7.39	0.589	-14.80	$< 10^{-48}$	-7.39	-7.39
mean_7	-7.71	0.605	-12.49	$< 10^{-35}$	-7.71	-7.71
mean_8	0.35	0.006	64.94	$< 10^{-99}$	0.35	0.35
mean_9	0.27	0.013	17.84	$< 10^{-70}$	0.27	0.27
scale_1	-5.95	0.012	-22.97	$< 10^{-99}$	-5.95	-5.96
scale_2	2.22	0.037	31.54	$< 10^{-99}$	2.22	2.22
smooth_1	1.20	Inf	0.00	$< 10^{-99}$	1.20	1.20
smooth_2	0.34	Inf	0.00	$< 10^{-99}$	0.34	0.34
unexplained_1	0.19	0.011	21.68	$< 10^{-99}$	0.19	0.19

Table B.2: Summary table for the fitted parameters in the model corresponding to $\mathrm{WT}|(\mathrm{HT}, \mathrm{Age})$

Parameter	Coef	Std. Error	z	Pr($>\|z\|$)	Lower 95%	Upper 95%
mean_1	104.16	0.593	175.73	$< 10^{-99}$	102.993	105.317
mean_2	0.07	0.595	-0.13	0.900	-1.092	1.242
mean_3	-0.41	0.291	-1.42	0.157	-0.984	0.158
mean_4	-0.57	0.270	-2.12	0.034	-1.101	-0.043
mean_5	-0.77	0.280	-2.74	0.006	-1.319	-0.219
mean_6	-0.61	0.262	-2.34	0.012	-1.127	-0.099
mean_7	-0.44	0.240	-1.84	0.065	-0.914	0.028
mean_8	-0.23	0.011	-21.15	$< 10^{-98}$	-0.255	-0.212
mean_9	-0.43	0.049	-8.79	$< 10^{-17}$	-0.523	-0.332
mean_10	0.85	0.035	24.22	$< 10^{-99}$	0.783	0.921
mean_11	0.79	0.142	5.57	$< 10^{-7}$	0.513	1.070
scale_1	-3.45	0.544	-6.35	$< 10^{-9}$	-4.521	-2.387
scale_2	1.67	0.163	10.26	$< 10^{-23}$	1.355	1.995
smooth_1	0.81	Inf	0.00	1.000	-Inf	Inf
smooth_2	7.24	0.980	7.39	$< 10^{-12}$	5.321	9.162
unexplained_1	1.95	0.013	149.67	$< 10^{-99}$	1.925	1.976

Table B.3: Summary table for the fitted parameters in the model corresponding to $\mathrm{SBP}|(\mathrm{WT}, \mathrm{HT}, \mathrm{Age})$

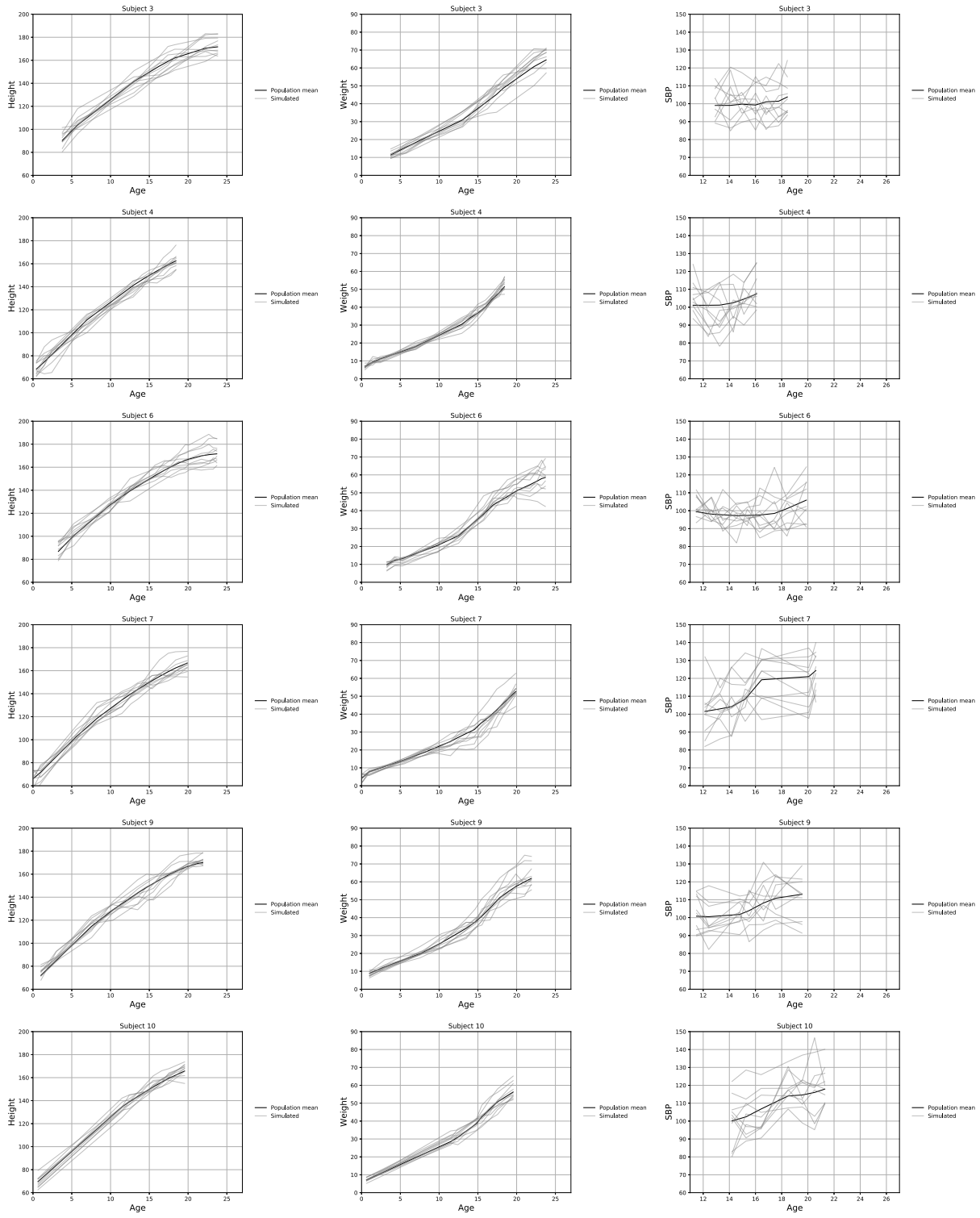


Figure B.1: Simulated datasets for afew individuals

APPENDIX C

Moments Standard Error

Some primary simulation results on marginal SE are presented below. We obtain SE estimates for the 4 kinds of marginal covariances as mentioned in section 5.5.3, again corresponding to $s = 0.2$ and $t = 0.6$, using 200 MC replicates. The relative SE is around 0.5%.

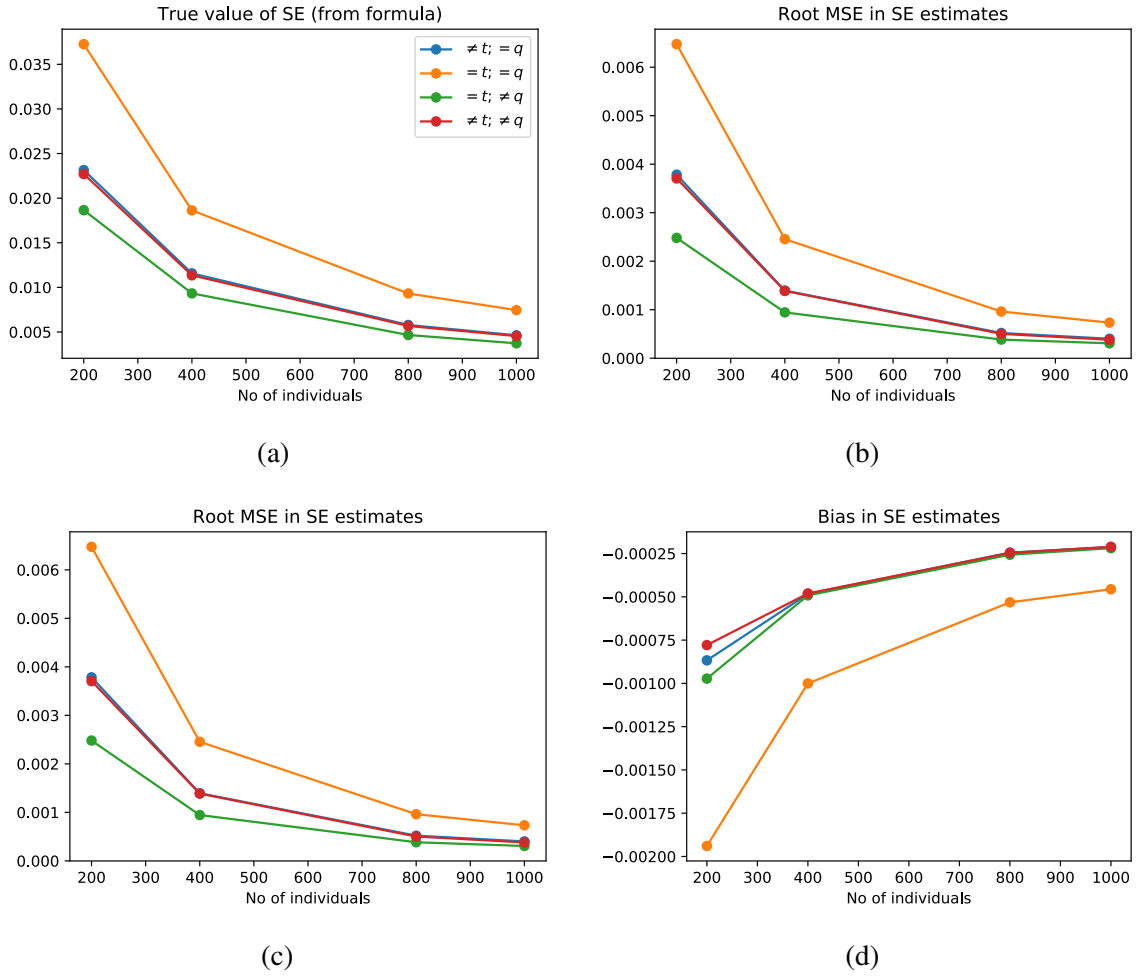


Figure C.1: SE estimates of the four kinds of marginal covariance estimates.

BIBLIOGRAPHY

[1] Hervé Abdi. Factor rotations in factor analyses. *Encyclopedia for Research Methods for the Social Sciences. Sage: Thousand Oaks, CA*, pages 792–795, 2003.

[2] Namanjeet Ahluwalia, Johanna Dwyer, Ana Terry, Alanna Moshfegh, and Clifford Johnson. Update on nhanes dietary data: focus on collection, release, analytical considerations, and uses to inform public policy. *Advances in nutrition*, 7(1):121–134, 2016.

[3] Theodore Wilbur Anderson. Estimating linear restrictions on regression coefficients for multivariate normal distributions. *The Annals of Mathematical Statistics*, pages 327–351, 1951.

[4] Theodore Wilbur Anderson. Asymptotic distribution of the reduced rank regression estimator under general conditions. *The Annals of Statistics*, 27(4):1141–1154, 1999.

[5] Hajer Aounallah-Skhiri, Jalila El Ati, Pierre Traissac, Habiba Ben Romdhane, Sabrina Eymard-Duvernay, Francis Delpeuch, Noureddine Achour, and Bernard Maire. Blood pressure and associated factors in a north african adolescent population. a national cross-sectional study in tunisia. *BMC public health*, 12(1):1–10, 2012.

[6] Reuben M Baron and David A Kenny. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology*, 51(6):1173, 1986.

[7] Pallab K Bhattacharya and Ashis K Gangopadhyay. Kernel and nearest-neighbor estimation of a conditional quantile. *The Annals of Statistics*, pages 1400–1415, 1990.

[8] Zulfiqar A Bhutta. Early nutrition and adult outcomes: pieces of the puzzle. *Lancet (London, England)*, 382(9891):486–487, 2013.

[9] M-A Bind, TJ VanderWeele, JD Schwartz, and BA Coull. Quantile causal mediation analysis allowing longitudinal data. *Statistics in medicine*, 36(26):4182–4195, 2017.

[10] Robert E Black, Cesar G Victora, Susan P Walker, Zulfiqar A Bhutta, Parul Christian, Mercedes De Onis, Majid Ezzati, Sally Grantham-McGregor, Joanne Katz, Reynaldo Martorell, et al. Maternal and child undernutrition and overweight in low-income and middle-income countries. *The lancet*, 382(9890):427–451, 2013.

[11] Tayler A Blake and Yoonkyung Lee. Nonparametric covariance estimation with shrinkage toward stationary models. 2019.

[12] Andreas Buja, Trevor Hastie, and Robert Tibshirani. Linear smoothers and additive models. *The Annals of Statistics*, pages 453–510, 1989.

[13] Zongwu Cai and Xiaoping Xu. Nonparametric quantile estimations for dynamic smooth coefficient models. *Journal of the American Statistical Association*, 103(484):1595–1608, 2008.

[14] CDC. Nhanes, 2022.

[15] Kun Chen, Kung-Sik Chan, and Nils Chr Stenseth. Reduced rank stochastic regression with a sparse singular value decomposition. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(2):203–221, 2012.

[16] Kun Chen, Hongbo Dong, and Kung-Sik Chan. Reduced rank regression via adaptive nuclear norm penalization. *Biometrika*, 100(4):901–920, 2013.

[17] Arnaud Chiolero, Gilles Paradis, and Jay S Kaufman. Assessing the possible direct effect of birth weight on childhood blood pressure: a sensitivity analysis. *American journal of epidemiology*, 179(1):4–11, 2014.

[18] Pádraig Cunningham. Dimension reduction. In *Machine learning techniques for multimedia*, pages 91–112. Springer, 2008.

[19] Cristina Davino, Marilena Furno, and Domenico Vistocco. *Quantile regression: theory and applications*, volume 988. John Wiley & Sons, 2013.

[20] Jan G De Gooijer and Dawit Zerom. On additive conditional quantiles with high-dimensional covariates. *Journal of the American Statistical Association*, 98(461):135–146, 2003.

[21] Katrina L Devick, Linda Valeri, Jarvis Chen, Alejandro Jara, Marie-Abèle Bind, and Brent A Coull. The role of body mass index at diagnosis of colorectal cancer on black–white disparities in survival: a density regression mediation approach. *Biostatistics*, 2020.

[22] Francesca Dominici, Scott L Zeger, Giovanni Parmigiani, Joanne Katz, and Parul Christian. Estimating percentile-specific treatment effects in counterfactual models: a case-study of micronutrient supplementation, birth weight and infant mortality. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 55(2):261–280, 2006.

[23] Suman Dua, Monika Bhuker, Pankhuri Sharma, Meenal Dhall, and Satwanti Kapoor. Body mass index relates to blood pressure among adults. *North American journal of medical sciences*, 6(2):89, 2014.

[24] Naihua Duan and Ker-Chau Li. Slicing regression: a link-free regression method. *The Annals of Statistics*, pages 505–530, 1991.

[25] Jianqing Fan and Irene Gijbels. *Local polynomial modelling and its applications*. Routledge, 2018.

[26] Yanqin Fan and Ruixuan Liu. Symmetrized multivariate k-nn estimators. *Econometric Reviews*, 34(6-10):828–848, 2015.

[27] Nora Fenske, Ludwig Fahrmeir, Torsten Hothorn, Peter Rzehak, and Michael Höhle. Boosting structured additive quantile regression for longitudinal childhood obesity data. *The international journal of biostatistics*, 9(1):1–18, 2013.

[28] Imola K Fodor. A survey of dimension reduction techniques. Technical report, Lawrence Livermore National Lab., CA (US), 2002.

[29] Wing Kam Fung, Xuming He, Li Liu, and Peide Shi. Dimension reduction based on canonical correlation. *Statistica Sinica*, pages 1093–1113, 2002.

[30] Christian Galarza Morales. *Quantile regression for mixed-effects models*. PhD thesis, Universidade Estadual de Campinas, 03 2015.

[31] Michael Gamborg, Per Kragh Andersen, Jennifer L Baker, Esben Budtz-Jørgensen, Torben Jørgensen, Gorm Jensen, and Thorkild IA Sørensen. Life course path analysis of birth weight, childhood growth, and adult systolic blood pressure. *American journal of epidemiology*, 169(10):1167–1178, 2009.

[32] Marco Geraci. Additive quantile regression for clustered data with an application to children's physical activity. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 68(4):1071–1089, 2019.

[33] Marco Geraci and Matteo Bottai. Quantile regression for longitudinal data using the asymmetric laplace distribution. *Biostatistics*, 8(1):140–154, 2007.

[34] Marco Geraci and Alessandra Mattei. A novel quantile-based decomposition of the indirect effect in mediation analysis with an application to infant mortality in the us population. *arXiv preprint arXiv:1710.00720*, 2017.

[35] Louise GH Goh, Satvinder S Dhaliwal, Timothy A Welborn, Andy H Lee, and Phillip R Della. Anthropometric measurements of general and central obesity and the prediction of cardiovascular disease risk in women: a cross-sectional study. *BMJ open*, 4(2):e004138, 2014.

[36] Ling Guo, Akil Narayan, and Tao Zhou. Constructing least-squares polynomial approximations. *SIAM Review*, 62(2):483–508, 2020.

[37] Peter Hall, Hans-Georg Müller, and Jane-Ling Wang. Properties of principal component methods for functional and longitudinal data analysis. *The annals of statistics*, pages 1493–1517, 2006.

[38] James W Hardin and Joseph M Hilbe. *Generalized estimating equations*. chapman and hall/CRC, 2002.

[39] Wolfgang Hardle, Peter Hall, and Hidehiko Ichimura. Optimal smoothing in single-index models. *The annals of Statistics*, 21(1):157–178, 1993.

[40] Kenneth Harttgen and Mark Misselhorn. A multilevel approach to explain child mortality and undernutrition in south asia and sub-saharan africa. Technical report, IAI Discussion Papers, 2006.

[41] Trevor J Hastie and Robert J Tibshirani. *Generalized additive models*. Routledge, 2017.

[42] John Hoddinott, John Maluccio, Jere R Behrman, Reynaldo Martorell, Paul Melgar, Agnes R Quisumbing, Manuel Ramirez-Zea, Aryeh D Stein, Kathryn M Yount, et al. The consequences of early childhood growth failure over the life course. *Washington, DC: International Food Policy Research Institute Discussion Paper*, 1073:5, 2011.

[43] Joel L Horowitz and Sokbae Lee. Nonparametric estimation of an additive quantile regression model. *Journal of the American Statistical Association*, 100(472):1238–1249, 2005.

[44] Tzu-Kuang Hsu. Exploring the impact of openness on inflation in taiwan: Mediation analysis with a quantile approach. *WSEAS TRANSACTIONS ON BUSINESS AND ECONOMICS*, 2020.

[45] Jianhua Z Huang, Naiping Liu, Mohsen Pourahmadi, and Linxu Liu. Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*, 93(1):85–98, 2006.

[46] Andraea Van Hulst, Tracie A Barnett, Gilles Paradis, Marie-Hélène Roy-Gagnon, Lilianne Gomez-Lopez, and Mélanie Henderson. Birth weight, postnatal weight gain, and childhood adiposity in relation to lipid profile and blood pressure during early adolescence. *Journal of the American Heart Association*, 6(8):e006302, 2017.

[47] Hidehiko Ichimura. *Estimation of single index models*. PhD thesis, Massachusetts Institute of Technology, 1987.

[48] Kosuke Imai, Luke Keele, and Dustin Tingley. A general approach to causal mediation analysis. *Psychological methods*, 15(4):309, 2010.

[49] Hyunsu Ju. Moving block bootstrap for analyzing longitudinal data. *Communications in Statistics-Theory and Methods*, 44(6):1130–1142, 2015.

[50] Ben Kelcey, Heather C Hill, and Mark J Chin. Teacher mathematical knowledge, instructional quality, and student outcomes: a multilevel quantile mediation analysis. *School Effectiveness and School Improvement*, 30(4):398–431, 2019.

[51] Roger Koenker. Additive models for quantile regression: Model selection and confidence bandaids. *Brazilian Journal of Probability and Statistics*, 25(3):239–262, 2011.

[52] Roger Koenker. Quantile regression: 40 years on. *Annual Review of Economics*, 9:155–176, 2017.

[53] Roger Koenker, Victor Chernozhukov, Xuming He, and Limin Peng. *Handbook of quantile regression*. CRC press, 2017.

[54] Roger Koenker and Kevin F Hallock. Quantile regression. *Journal of economic perspectives*, 15(4):143–156, 2001.

[55] Roger Koenker, Pin Ng, and Stephen Portnoy. Quantile smoothing splines. *Biometrika*, 81(4):673–680, 1994.

[56] Victor H Lachos, Ming-Hui Chen, Carlos A Abanto-Valle, and Caio LN Azevedo. Quantile regression for censored mixed-effects models with applications to hiv studies. *Statistics and its interface*, 8(2):203, 2015.

[57] Debbie A Lawlor, David A Leon, and Finn Rasmussen. Growth trajectory matters: interpreting the associations among birth weight, concurrent body size, and systolic blood pressure in a cohort study of 378,707 swedish men. *American journal of epidemiology*, 165(12):1405–1412, 2007.

[58] Rien van der Leeden, Erik Meijer, and Frank MTA Busing. Resampling multilevel models. In *Handbook of multilevel analysis*, pages 401–433. Springer, 2008.

[59] Ker-Chau Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991.

[60] Ker-Chau Li, Yve Aragon, Kerby Shedden, and Christine Thomas Agnan. Dimension reduction for multivariate response data. *Journal of the American Statistical Association*, 98(461):99–109, 2003.

[61] Yehua Li, Tailen Hsing, et al. Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data. *The Annals of Statistics*, 38(6):3321–3351, 2010.

[62] Yehua Li, Naisyin Wang, Meeyoung Hong, Nancy D Turner, Joanne R Lupton, and Raymond J Carroll. Nonparametric estimation of correlation functions in longitudinal and spatial data, with application to colon carcinogenesis experiments. *The Annals of Statistics*, pages 1608–1643, 2007.

[63] Kung-Yee Liang, Scott L Zeger, and Bahjat Qaqish. Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 54(1):3–24, 1992.

[64] Yanyuan Ma and Liping Zhu. A review on dimension reduction. *International Statistical Review*, 81(1):134–150, 2013.

[65] David P MacKinnon. *Introduction to statistical mediation analysis*. Routledge, 2012.

[66] David P MacKinnon, Amanda J Fairchild, and Matthew S Fritz. Mediation analysis. *Annu. Rev. Psychol.*, 58:593–614, 2007.

[67] Simon Mak and V Roshan Joseph. Support points. *The Annals of Statistics*, 46(6A):2562–2592, 2018.

[68] Vinicius JB Martins, Telma MM Toledo Florêncio, Luciane P Grillo, Maria Do Carmo P Franco, Paula A Martins, Ana Paula G Clemente, Carla DL Santos, Maria de Fatima A Vieira, and Ana Lydia Sawaya. Long-lasting effects of undernutrition. *International journal of environmental research and public health*, 8(6):1817–1846, 2011.

[69] Nicolai Meinshausen and Greg Ridgeway. Quantile regression forests. *Journal of Machine Learning Research*, 7(6), 2006.

[70] Stefanie Muff, Leonhard Held, and Lukas F Keller. Marginal or conditional regression models for correlated non-normal data? *Methods in Ecology and Evolution*, 7(12):1514–1524, 2016.

[71] Ashin Mukherjee and Ji Zhu. Reduced rank ridge regression and its kernel extensions. *Statistical analysis and data mining: the ASA data science journal*, 4(6):612–622, 2011.

[72] Iván Díaz Muñoz and Mark Van Der Laan. Population intervention causal effects based on stochastic interventions. *Biometrics*, 68(2):541–549, 2012.

[73] Sahand Negahban and Martin J Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, 39(2):1069–1097, 2011.

[74] Chung Wo Ong, X-A Zhao, YC Tsang, CL Choy, and PW Chan. Effects of substrate temperature on the structure and properties of reactive pulsed laser deposited cnx films. *Thin Solid Films*, 280(1-2):1–4, 1996.

[75] Jason W Osborne. What is rotating in exploratory factor analysis? *Practical Assessment, Research, and Evaluation*, 20(1):2, 2015.

[76] Christopher J Paciorek and Mark J Schervish. Nonstationary covariance functions for gaussian process regression. In *Advances in neural information processing systems*, pages 273–280, 2004.

[77] Christopher J Paciorek and Mark J Schervish. Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics: The official journal of the International Environmetrics Society*, 17(5):483–506, 2006.

[78] Oscar Hernan Madrid Padilla, Wesley Tansey, and Yanzhen Chen. Quantile regression with deep relu networks: Estimators and minimax rates. *arXiv preprint arXiv:2010.08236*, 2020.

[79] Kirsi H Pietiläinen, Jaakko Kaprio, Maija Räsänen, Torsten Winter, Aila Rissanen, and Richard J Rose. Tracking of body size from birth to late adolescence: contributions of birth length, birth weight, duration of gestation, parents' body size, and twinship. *American journal of epidemiology*, 154(1):21–29, 2001.

[80] Mohsen Pourahmadi. Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika*, 86(3):677–690, 1999.

[81] James L Powell, James H Stock, and Thomas M Stoker. Semiparametric estimation of index coefficients. *Econometrica: Journal of the Econometric Society*, pages 1403–1430, 1989.

[82] Kristopher J Preacher. Advances in mediation analysis: A survey and synthesis of new developments. *Annual review of psychology*, 66:825–852, 2015.

[83] Ross L Prentice and Lue Ping Zhao. Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics*, pages 825–839, 1991.

[84] C Radhakrishna Rao. Estimation of variance and covariance components—minque theory. *Journal of multivariate analysis*, 1(3):257–275, 1971.

[85] Pradeep Ravikumar, John Lafferty, Han Liu, and Larry Wasserman. Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5):1009–1030, 2009.

[86] Ana L Sawaya, Ricardo Sesso, Telma M de Menezes Toledo Florêncio, Maria TB Fernandes, and Paula A Martins. Association between chronic undernutrition and hypertension. *Maternal & child nutrition*, 1(3):155–163, 2005.

[87] Bernhard Schmidpeter. Involuntary unemployment and the labor market returns to interim jobs. *Available at SSRN 3046577*, 2017.

[88] Ernest Shen, Chih-Ping Chou, Mary Ann Pentz, and Kiros Berhane. Quantile mediation models: A comparison of methods for assessing mediation across the outcome distribution. *Multivariate Behavioral Research*, 49(5):471–485, 2014.

[89] Kesar Singh. On the asymptotic accuracy of efron's bootstrap. *The Annals of Statistics*, pages 1187–1195, 1981.

[90] Kirstine Smith. On the standard deviations of adjusted and interpolated values of an observed polynomial function and its constants and the guidance they give towards a proper choice of the distribution of observations. *Biometrika*, 12(1/2):1–85, 1918.

[91] Beverly I Strassmann. Cooperation and competition in a cliff-dwelling people. *Proceedings of the National Academy of Sciences*, 108(Supplement 2):10894–10901, 2011.

[92] Madeleine Udell, Corinne Horn, Reza Zadeh, Stephen Boyd, et al. Generalized low rank models. *Foundations and Trends® in Machine Learning*, 9(1):1–118, 2016.

[93] Elaine M Urbina, Samuel S Gidding, Weihang Bao, Arthur S Pickoff, Kaliope Berdusis, and Gerald S Berenson. Effect of body size, ponderosity, and blood pressure on left ventricular growth in children and young adults in the bogalusa heart study. *Circulation*, 91(9):2400–2406, 1995.

[94] Raja Velu and Gregory C Reinsel. *Multivariate reduced-rank regression: theory and applications*, volume 136. Springer Science & Business Media, 2013.

[95] Simon N Wood. *Generalized additive models: an introduction with R*. chapman and hall/CRC, 2006.

[96] Zhijie Xiao and Roger Koenker. Conditional quantile estimation for generalized autoregressive conditional heteroscedasticity models. *Journal of the American Statistical Association*, 104(488):1696–1712, 2009.

[97] Fang Yao. Asymptotic distributions of nonparametric regression estimators for longitudinal or functional data. *Journal of Multivariate Analysis*, 98(1):40–56, 2007.

[98] Fang Yao, Hans-Georg Müller, and Jane-Ling Wang. Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100(470):577–590, 2005.

[99] Steven Siwei Ye and Oscar Hernan Madrid Padilla. Non-parametric quantile regression via the k-nn fused lasso, 2020.

[100] Ashenafi A Yirga, Sileshi F Melesse, Henry G Mwambi, and Dawit G Ayele. Additive quantile mixed effects modelling with application to longitudinal cd4 count data. *Scientific reports*, 11(1):1–12, 2021.

[101] Ming Yuan, Ali Ekici, Zhaosong Lu, and Renato Monteiro. Dimension reduction and coefficient estimation in multivariate linear regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(3):329–346, 2007.

[102] Xiaoke Zhang, Jane-Ling Wang, et al. From sparse to dense functional data and beyond. *The Annals of Statistics*, 44(5):2281–2321, 2016.

[103] Yang Zhou, Shu-Chin Lin, and Jane-Ling Wang. Local and global temporal correlations for longitudinal data. *Journal of Multivariate Analysis*, 167:1–14, 2018.