University of California
Santa Barbara

Application of the single index methodology to the local Fréchet regression in the context of Object oriented data analysis (OODA)

A dissertation submitted in partial satisfaction
of the requirements for the degree

Doctor of Philosophy
in
Statistics and Applied Probability

by

Aritra Ghosal

Committee in charge:

Dr. Wendy Meiring, Committee Co-Chair
Dr. Alexander Petersen, Committee Co-Chair
Dr. Yuedong Wang
Dr. Andrew Carter

September 2023

The Dissertation of Aritra Ghosal is approved.

Dr. Yuedong Wang

Dr. Andrew Carter

Dr. Alexander Petersen, Committee Co-Chair

Dr. Wendy Meiring, Committee Co-Chair

September 2023

Application of the single index methodology to the local Fréchet regression in the context of Object oriented data analysis (OODA)

by

Aritra Ghosal

By the grace of Almighty God, I would like to dedicate this dissertation to my parents Manika Ghosal, and Amartya Ghosal.

Acknowledgements

I would like to express deepest appreciation to my esteemed advisors and co-chairs of my committee Dr. Wendy Meiring, and Dr. Alexander Petersen for their countless invaluable feedback sessions, patience, and guidance. I also could not have undertaken this journey without the rest of my defense committee, also my esteemed professors Dr. Andrew Carter, and Dr. Yuedong Wang, who generously provided me knowledge, expertise and suggestions. Additionally, this endeavor would not have been possible without the generous support from the Department of Statistics and Applied Probability, University of California Santa Barbara for giving me the various opportunities to teach courses, attend lectures, conferences, for encouraging further research throughout my stay as an aspiring PhD student.

I am also sincerely grateful to my collaborator and friend Marcos Matabuena, for his endless encouragement to put in my best efforts, through his regular feedback sessions. Thanks should also go to the librarians, fellow graduate students of the department who impacted and inspired me.

I would be remiss in not mentioning my family, especially my parents. Their belief in me has kept my spirits up and motivation high during this process.

Curriculum Vitæ

Aritra Ghosal

Education

2023	Ph.D. in Statistics and Applied Probability (Expected), University of California, Santa Barbara.
2014	M.Stat. in Statistics, Indian Statistical Institute, Kolkata.
2011	B.Sc. (Hons.) in Statistics, Presidency College, Calcutta University.

Publications

1. Ghosal, Aritra, Meiring, Wendy and Petersen, Alexander. (2023). *Fréchet single index models for object response regression. Electronic Journal of Statistics* 17(1), 1074 – 1112.
2. Ghosal, Aritra, Matabuena, Marcos, Meiring, Wendy and Petersen, Alexander. *Predicting distributional profiles of physical activity in the NHANES database using a partially linear single-index Fréchet regression model, arXiv preprint: 2302.07692.*

Abstract

Application of the single index methodology to the local Fréchet regression in the context of Object oriented data analysis (OODA)

by

Aritra Ghosal

In the context of Object oriented data analysis (OODA), the local Fréchet regression was formulated in analogy to the local linear regression to model the conditional Fréchet mean of the response on a single covariate in $\mathbb{R}$. To accommodate $p \geq 2$ covariates we introduced the Fréchet single index (FSI) model in analogy to the single index model already existing for responses in $\mathbb{R}$. We discussed the model performance on simulated spherical data and on observed mortality distributions belonging to the L^2-Wasserstein space. We also discussed the consistency of the coefficient vector estimate by combining the Fréchet regression and the M-estimation methods. We discovered the potential of our model to analyze the biomedical data obtained from the wearable accelerometer devices, available for the US population from the NHANES website for the period 2011-14. The physical activity profiles, transformed into quantile distributions, were considered Object responses in the partially linear Fréchet single index (PL-FSI) model which allowed an additive linear part with the single index part. The semi-parametric character of the model allows us to introduce non-linear effects for such covariates as Age, BMI, while the inclusion of a linear part retains the advantage of interpretability for other categorical variables such as diet score, ethnicity, sex and their interaction.

Contents

Chapter 1

Introduction

In recent times, sophisticated technology has enabled us to collect massive amounts of data in various fields of study. Very often, such data may belong to non-euclidean spaces whose geometric or structural properties may or may not be explicitly defined, only a measure of distance between two elements in the space is known. Such intrinsic properties (or their absence) ought to be considered when attempting to analyze such data or to draw a more comprehensive statistical inference from it. A challenge for the modern statisticians is to develop better methodologies to handle such data. To understand the complexity of the data through measures of central tendency, and dispersion, we first consider the metric that define the space. The earliest attempt in this investigation has to be credited to Maurice Fréchet, whose work in [1] introduced the concepts of Fréchet mean and Fréchet variance for the elements of a metric space. In recent studies, such responses in a metric space have been termed as objects [2] and such methods of analysis have been collectively termed as Object Oriented Data Analysis (OODA) as discussed by [3, 4]. As the methods of analysis vary with metric spaces, such concepts have been discussed when the response is a covariance matrix, residing in the space of symmetric positive definite matrices [5], probability distribution [6, 7], and networks [8].

The study of modeling random objects on the covariates in $\mathbb{R}^p$ has been growing steadily in the recent times, this is reflected in the studies of circular/spherical data [9, 10, 11], smooth manifolds [12, 13, 14, 15, 5, 16, 17], and more recently in general metric spaces [18, 19]. The regression models discussed in the cases of Smooth Riemannian Manifolds can be fully parametric, semi-parametric or non-parametric. Recently [19] has developed the global Fréchet model and the local Fréchet regression as generalizations of the multiple linear regression and local linear regression for objects for the general metric spaces. The object we are modelling has similar mathematical principles as the classical Fréchet mean.

As local and global Fréchet regression models are inspired from their respective counterparts local linear regression and the multiple linear regression when the responses are in $\mathbb{R}$, we looked for inspiration from similar scalar response models to balance the strengths and weaknesses of these two methods. The model we proposed here for responses in the general metric space has been discussed extensively for scalar responses as the single index models. Specifically we followed the approach studied in [20] estimating the model parameters. Consider the random pair $(\boldsymbol{X}, Y) \in \mathbb{R}^p \times \mathbb{R}$. The regression function for the scalar response model asserts that:

$$m(\boldsymbol{x}) := E(Y|\boldsymbol{X} = \boldsymbol{x}) = g(\boldsymbol{\theta}_0^T \boldsymbol{x}) \tag{1.1}$$

where $\boldsymbol{\theta}_0 \in \mathbb{R}^p$ is an unknown parameter and g is smooth and unknown function residing in an infinite-dimensional univariate function space. In multiple linear regression the g is assumed to be linear while in the fully non-parametric regression it lends interpretibility to the covariates as the parameter $\boldsymbol{\theta}_0$ is estimable, and it adds flexibility by allowing the effect of $\boldsymbol{\theta}^T \boldsymbol{X}$ to be non-linear. The model discussed in [20] was able to estimate the parameter $\boldsymbol{\theta}_0$ and show the consistency of the estimate with its parametric

rate of convergence, even proving asymptotic normality under certain conditions.

We proposed the Fréchet single index (FSI) regression model generalizing the standard single index model by modelling the Fréchet means of the random objects, conditional on the covariates in $\mathbb{R}^p$ [21]. We simulated spherical response data to illustrate the sampling variability of both the estimators of the index parameter and the overall regression function respectively. As an application of this method we discussed the modelling of the yearly distribution of age-at-death for various countries as elements of the L^2-Wasserstein space, endowed with the Wasserstein metric (1.3). As for covariates we chose some economic indicators which made sense to utilize in mortality analysis through various literature.

Having formulated the new model, we looked to the frontier of medical science for a possible application of our model. With the vast research conducted in the field of precision medicine [22, 23, 24, 25], its applicability is likely to grow exponentially in the near future. The patient information is recorded from wearable devices (accelerometers) and such data on physical activity profiles are represented as probability distributions [26, 27, 28, 29]. This format of data representation contains far more information about the patient's health condition than simple traditional biomarkers, thus improving the reliability of the inference significantly [27, 28, 29, 30] and introducing more personalized approach to the treatment. With this logic, distributional representations of physical activity can be considered a direct extension of the classical summary metrics [31, 32].

One of the key objectives of our exercise was to understand the factors characterizing the physical activity patterns for the American population while modelling the physical activity representation as objects in the L^2-Wasserstein space. However, the covariates are known to affect the physical activity in different ways, some linearly, while others may exhibit various non-linear association, e.g. age and some other anthropomorphic measures affect the physical energy expenditure non-linearly [33, 34]. We also wanted to

include categorical variables like sex and ethnicity which not only explain the variation in physical activity levels significantly in the American population, they tend to interact as well [35, 36, 37]. Hence we improved the Fréchet single index model by introducing a linear component adding further interpretability and flexibility to the existing model. For classical regression models with univariate response data, the partially linear single index model has been a topic with considerable popularity among researchers in the fields of statistical and econometrics in the last twenty years [38]. Several other works in this direction discussed recent extensions of the model to functional data [38, 39, 40, 41]. However, the author was not aware of any existing extension to response data in metric spaces. To analyze such data we proposed the partially linear Fréchet single index (PLFSI) regression model [42].

The NHANES (National Health And Nutrition Examination Survey) is a major program undertaken by National Center for Health Statistics (NCHS). NCHS is in turn a part of the CDC (Center for Disease Control and Prevention). The goal of these surveys are designed to evaluate the health and nutritional status of the adults and children in the United States. We apply the complex survey sampling design employed by the NHANES to obtain more reliable population based estimate of the model [43]. We were not aware of any work that incorporated the complex survey design into the analysis for the partially linear single index model with responses in metric spaces.

The findings from these studies are important from the perspective of public health since it elucidates the variables that impact the physical activity among the American population in all levels of the accelerometer intensities. Moreover, these new findings can be useful to refine and plan specific health interventions and policies that reduce the gap in physical inactivity in different US sub-populations.

The structure of the thesis is as follows. Chapter 2 lays the foundation to discuss the Fréchet regression, defining Fréchet mean and Fréchet variance, thereby defining

marginal and conditional Fréchet mean. Sections 2.2 and 2.3 formally define the global Fréchet model and the local Fréchet regression respectively and present these techniques as generalizations of the multiple linear regression and the non-parametric local linear model to estimate the conditional Fréchet mean of the Object data. Chapter 3 introduces the Fréchet single index model for the object response regression. Section 3.1 formally defines the Fréchet single index model for response in metric spaces by utilizing the properties of the local Fréchet model. Section 3.2 discusses the estimation method of the object-valued regression function and the single index coefficient vector. Section 3.3 discusses the simulation of the response data on the surface of 3-dimensional sphere and the sampling variability of the relevant estimators of the Fréchet single index model. Section 3.4 discusses an example of application of the FSI model where the distribution of yearly age-at-death is modelled on various economic indicators. Chapter 4 discusses the NHANES data that will be analyzed by the PL-FSI regression model, it formulates the quantile distribution of the physical activity data for each participant as the response for our model, discusses the covariates used in the model, it formulates the quantile distribution of the physical activity representation on an equidistant grid over $t \in [0, 1]$, while being members of the L^2-Wasserstein space, thereby formally defines the PL-FSI regression model for every t. Section 4.3 discusses the estimation procedure of the single index coefficient vector by considering a B-spline basis expansion of the unknown single index function and the parameters of regression. Section 4.4 discusses some findings and interesting aspects of the inference obtained from our model. Section 4.5 discusses the criteria for the selection of covariates in the PL-FSI model. Here we explore how the men and women among different ethnicities differ with respect to physical activity, if there is possible interaction between sex and ethnicities. It also explores the association of physical activity with the covariates in the non-linear component of the model, e.g. age and BMI.

The compilation of all the codes for the simulated spherical data example and the real data application related to the FSI model can be found on Github (`https://github.com/aghosal89/Frechet_SingleIndex`). The compilation of all the codes for the application of the PL-FSI model can be found in the repository on Github (`https://github.com/aghosal89/FSI_NHANES_Application`).

1.1 Some examples of metrics

We will discuss the analysis of object data in the general metric space (Ω, d). However, in the following subsections, we define some specific metric spaces and their characteristic metrics to aid us in discussion for the rest of this paper:

1.1.1 Geodesic distance in spherical data

Consider the data belonging to the space $\Omega = S^{p-1} = \{\boldsymbol{x} \in \mathbb{R}^p : ||\boldsymbol{x}||_E = 1\}$, where $||\cdot||_E$ is the euclidean norm, and $p \geq 2$. Hence, S^{p-1} represents the surface of the unit sphere with the center as the origin. Then the shortest distance on the spherical surface between two arbitrary points $\boldsymbol{x}_1, \boldsymbol{x}_2 \in S^2$ is defined as,

$$\mathrm{d}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \arccos(\boldsymbol{x}_1^T \boldsymbol{x}_2) \tag{1.2}$$

Also known as the geodesic distance between $\boldsymbol{x}_1, \boldsymbol{x}_2$.

1.1.2 Distributional data in L^2- Wasserstein space

Let Ω be the set of probability distribution functions with finite second moment. The metric between two distributions is called the Wasserstein metric. Let $\omega \in \Omega$ be a distribution on $\mathbb{R}$, then $\int_{\mathbb{R}} x^2 \, d\omega(x) < \infty$. For two distributions $\omega_1, \omega_2 \in \Omega$, the squared

Wasserstein distance [44] between them is

$$\mathrm{d}^2(\omega_1, \omega_2) = \int_0^1 \left(\omega_1^{-1}(t) - \omega_2^{-1}(t)\right)^2 dt, \tag{1.3}$$

where $\omega_1^{-1}, \omega_2^{-1}$ are the quantile functions corresponding to ω_1, ω_2 respectively, also known as the survival functions. The above form of the metric makes obvious the point that the L^2-Wasserstein space is isometric to a subset of the Hilbert space $L^2[0,1]$. Thus, it is a flat Hadamard space, though it is convex and not linear.

1.2 Permissions and Attributions

The contents of chapters 2, 3, 4, and 5 are the result of a collaboration with my co-advisors Dr. Alexander Petersen and Dr. Wendy Meiring, and has previously appeared as [21]. The contents of chapters 6 and 7 are the result of my work with my advisors mentioned above along with Marcos Matabuena (email: mmatabuena@hsph.harvard.edu) which earlier appeared as [42].

Chapter 2

Background on Fréchet Regression

2.1 Setup of the problem

Let (Ω, d) be a bounded metric space. The response $Y \in \Omega$ is to be modeled conditionally on a p-dimensional covariate $\boldsymbol{X} \in \mathbb{R}^p$. Assume $(\boldsymbol{X}, Y) \sim F$, with F being a joint distribution on $\mathbb{R}^p \times \Omega$ such that $\boldsymbol{\Sigma} = \mathrm{Var}(\boldsymbol{X})$ exists with $\boldsymbol{\Sigma}$ positive definite and $\boldsymbol{\mu} = E(\boldsymbol{X})$. When Ω is a Euclidean space such as $\mathbb{R}^p$ or $L^2[0,1]$ as would be the typical case for multivariate or functional data, one can utilize the usual notions of expectation arising from Lebesgue integration to quantify the mean and variance of Y. For arbitrary metric spaces Ω, the concepts of mean and variance of a random variable are replaced by the Fréchet mean and the Fréchet variance [1], respectively, defined as

$$\omega_\oplus = \underset{\omega \in \Omega}{\operatorname{argmin}} E\left(\mathrm{d}^2(Y, \omega)\right), \quad V_\oplus = E\left(\mathrm{d}^2\left(Y, \omega_\oplus\right)\right). \tag{2.1}$$

Existence and uniqueness of the Fréchet mean is not guaranteed for general metric spaces. However, in special cases such as certain Riemannian manifolds [45, 46] or spaces with negative curvature [47, 48], Fréchet means exist and are unique. For the moment,

we assume at least that a minimizer exists, with the consequence that $\omega_\oplus$ and $V_\oplus$ are not vacuous, and the latter is unique. Extending these concepts to regression, define the Fréchet regression function Y given $\boldsymbol{X} = \boldsymbol{x} \in \mathbb{R}^p$ as

$$m_\oplus(\boldsymbol{x}) = \underset{\omega \in \Omega}{\operatorname{argmin}} M_\oplus(\omega, \boldsymbol{x}), \quad M_\oplus(\cdot, \boldsymbol{x}) = E(\mathrm{d}^2(Y, \cdot)|\boldsymbol{X} = \boldsymbol{x}). \tag{2.2}$$

Two different approaches were proposed by [19] to estimate the conditional Fréchet mean $m_\oplus(\boldsymbol{x})$. First, a global model was proposed in which $m_\oplus(\boldsymbol{x})$ can be written as the minimizer of an alternative objective function motivated by multiple linear regression in the case $\Omega = \mathbb{R}$. The result is that $m_\oplus(\boldsymbol{x})$ can be viewed as a weighted Fréchet mean, where the weights depend on the joint distribution F and the input $\boldsymbol{x}$. As a direct generalization of linear regression, global Fréchet regression similarly can be overly restrictive for random object responses. Thus, in a second approach, [19] also demonstrated how to generalize local linear regression to estimate $m_\oplus(\boldsymbol{x})$ under less restrictive assumptions on the function $m_\oplus$. Both these approaches, termed local and global Fréchet regression, will now be described.

2.2 Global Fréchet regression model for Object data response

First we introduce the concept of global Fréchet regression in the context of OODA. An essential task in statistics is to find some regression relationship between the response Y and the covariate $\boldsymbol{X}$. For our case, when $Y \in \Omega$, modeling and estimating the Fréchet regression function $m_\oplus(\boldsymbol{x})$ from equation (2.2) is often of interest. The motivation stems from considering a scalar predictor $X \in \mathbb{R}$ and response $Y \in \Omega = \mathbb{R}$, so that the target $m_\oplus(x) =: m(x)$ in (2.2) is just the usual conditional expectation, as used in local polyno-

mial estimation. However, the global (parametric) modeling may not be straightforward, especially when Ω lacks a useful algebraic structure, such as an inner product. For instance, in classical linear regression analysis with $\Omega = \mathbb{R}$, the distribution of $(Y|\boldsymbol{X} = \boldsymbol{x})$ is normally distributed with a mean of $m(\boldsymbol{x}) = \alpha + \boldsymbol{\beta}^T\boldsymbol{x}$ and variance σ_Y^2, where α and $\boldsymbol{\beta}$ represent the regression coefficients. Similarly, when Ω possesses a linear-algebraic structure, one can specify a class of regression functions that quantifies the association between the expected outcome and covariates in an additive or multiplicative manner. However, the lack of an algebraic structure in general metric spaces may prevent us from characterizing $m_\oplus(\boldsymbol{x})$ with respect to the covariate $\boldsymbol{x}$ in the same way classical regression analysis determines the conditional expected value of the response with changing covariates.

Therefore, to tackle this challenge, [19] formulated the global Fréchet regression model such that for the covariates in the model, we can exploit the structure of the space $\mathbb{R}^p$ instead of Ω. Specifically, they formulated the Fréchet regression function as:

$$m_\oplus(\boldsymbol{x}) = \underset{\omega\in\Omega}{\operatorname{argmin}}\, M_\oplus(\cdot, \boldsymbol{x}), \quad M_\oplus(\omega, \boldsymbol{x}) = \mathrm{E}(s(\boldsymbol{X}, \boldsymbol{x})\, \mathrm{d}_E^2(Y, \omega)), \tag{2.3}$$

where $s(\cdot,\cdot) : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ is an arbitrary weight function such that $s(\boldsymbol{X}, \boldsymbol{x})$ denotes the influence of $\boldsymbol{X}$ on $\boldsymbol{x}$. The choice of the function s is considered following [19], in section 2.2.1.

2.2.1 Generalizing Linear Regression

[19] formulated the global Fréchet regression model in analogy to the multiple linear regression model for response in $\mathbb{R}$ so that we may implement and interpret the model and perform overall inference and testing on it with considerable ease. Global Fréchet model is fitted under the assumption that there is no bias. Hence, unlike any local model

fitting, global Fréchet regression does not involve choosing a tuning parameter. Here we follow the notations of [19]. Consider the standard setup of linear regression when $\Omega = \mathbb{R}$, the regression function, considering $m = m_{\oplus}$ in 2.2 becomes:

$$m(\boldsymbol{x}) := E(Y \mid \boldsymbol{X} = \boldsymbol{x}) = \beta_0^* + (\boldsymbol{\beta}_1^*)^T (\boldsymbol{x} - \boldsymbol{\mu}) \tag{2.4}$$

where, β_0^* and $\boldsymbol{\beta}_1^*$ are the solutions to the following equations:

$$(\beta_0^*, \boldsymbol{\beta}_1^*) = \underset{\beta_0 \in \mathbb{R}, \boldsymbol{\beta}_1 \in \mathbb{R}^p}{\operatorname{argmin}} \int \left[\int y \, dF_{Y|\boldsymbol{X}}(\boldsymbol{x}, y) - \left(\beta_0 + \boldsymbol{\beta}_1^T (\boldsymbol{x} - \boldsymbol{\mu})\right) \right]^2 dF_{Y|\boldsymbol{X}}(\boldsymbol{x}) \tag{2.5}$$

Setting $\boldsymbol{\mu} = E(\boldsymbol{X}), \boldsymbol{\Sigma} = \operatorname{Var}(\boldsymbol{X})$ and $\sigma_{Y\boldsymbol{X}} = E[Y(\boldsymbol{X} - \boldsymbol{\mu})]$, we get the solution to equation 2.5 are: $\boldsymbol{\beta}_1^* = \boldsymbol{\Sigma}^{-1}\sigma_{Y\boldsymbol{X}}$ and $\beta_0^* = E(Y)$. Plugging in the solutions into equation 2.4, we get,

$$m(\boldsymbol{x}) = E(Y) + \sigma_{Y\boldsymbol{X}}^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) = \int y s(\boldsymbol{z}, \boldsymbol{x}) \, dF(\boldsymbol{z}, y) \tag{2.6}$$

where the weight function is:

$$s(\boldsymbol{z}, \boldsymbol{x}) = 1 + (\boldsymbol{z} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \tag{2.7}$$

Notice that, $\int s(\boldsymbol{z}, \boldsymbol{x}) \, dF(\boldsymbol{z}, y) = 1$. Hence we characterized the regression in 2.4 such that the parameters are estimated via a weighted least square method where the weights are characterized by the covariates and the squared distance depends on the objects in the metric space.

2.2.2 Estimation

Let $i = 1, 2, \ldots, n$ be the index for observations, where n is the total number of observations under study. Consider the sample $(\boldsymbol{X}_1, Y_1), (\boldsymbol{X}_2, Y_2), \ldots, (\boldsymbol{X}_n, Y_n)$ be inde-

pendent and identically distributed according to F, and an arbitrary $\boldsymbol{x} \in \mathbb{R}^p$. for $i = 1, 2$, We estimate $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ by their corresponding empirical versions: $\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i$ and $\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})^T$ respectively. Hence, the empirical weights take the following form:

$$s_{in}(\boldsymbol{x}) := 1 + \left(\boldsymbol{X}_i - \bar{\boldsymbol{X}}\right)^T \hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{x} - \bar{\boldsymbol{X}}) \tag{2.8}$$

such that $\sum_{i=1}^{n} s_{in}(\boldsymbol{x}) = 1$. Hence, we get the regression function estimator:

$$\hat{m}_{\oplus}(\boldsymbol{x}) = \underset{\omega \in \Omega}{\operatorname{argmin}}\, M_n(\omega, \boldsymbol{x}) \tag{2.9}$$

where, $m_{\oplus}(\boldsymbol{x})$ for $\boldsymbol{x} \in \mathbb{R}^p$, and $M_n(\,\cdot\,, \boldsymbol{x}) = n^{-1}\sum_{i=1}^{n} s_{in}(\boldsymbol{x})\,\mathrm{d}^2\,(Y_i, \omega)$.

2.3 Local Fréchet Regression

Local regression is often preferred over the global regression fitting due to its inherent flexibility but requires the choice of a smoothing parameter that balances the bias and variance of the regression estimate. Under the similar setting described earlier, let K be a probability density kernel, and h be a bandwidth, and $K_h(\cdot) = h^{-1}K(\cdot/h)$. For our exposition we will consider covariates in the space $\mathbb{R}$ and follow the notation structure closely to those in [19].

Similar to the global Fréchet regression, local Fréchet regression is derived from the non-parametric local linear regression when $\Omega = \mathbb{R}$ and extended to the case when the responses are in a general metric space. The target of estimation again being (2.2), for the specific case $\Omega = \mathbb{R}$, we write $m = m_{\oplus}$.

According to [49], the local linear estimate of $m(x)$ is $\hat{l}(x) = \hat{\beta}_0$, where,

$$\left(\hat{\beta}_0, \hat{\beta}_1\right) = \underset{\beta_0,\beta_1}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} K_h\left(X_i - x\right)\left(Y_i - \beta_0 - \beta_1\left(X_i - x\right)\right)^2$$

by the properties of M-estimation ([50]), $\hat{\beta}_0, \hat{\beta}_1$ can be viewed as M-estimators of

$$(\beta_0^*, \beta_1^*) = \underset{\beta_0,\beta_1}{\operatorname{argmin}} \int K_h(z-x) \left[\int y \; dF_{Y|X}(z,y) - (\beta_0 + \beta_1(z-x))\right]^2 \; dF_X(z) \quad (2.10)$$

Define, for $j = 0, 1, 2, \ldots;$

$$\mu_j = E\left[K_h(X-x)(X-x)^j\right], r_j = E\left[K_h(X-x)(X-x)^j Y\right], \text{ and } \sigma_0^2 = \mu_0\mu_2 - \mu_1^2$$

the solutions to 2.10 are:

$$\beta_0^* = \sigma_0^{-2}\left(\mu_2 r_0 - \mu_1 r_1\right), \quad \beta_1^* = \sigma_0^{-2}\left(\mu_0 r_1 - \mu_1 r_0\right)$$

Hence, we get an intermediate target, which may be considered to be a smoothed version of the true regression $m(x)$.

$$\begin{aligned} \tilde{l}(x) = \beta_0^* &= \frac{\mu_2 r_0 - \mu_1 r_1}{\sigma_0^2} = \frac{1}{\sigma_0^2} \int y K_h(z-x)\left[\mu_2 - \mu_1(z-x)\right] \, dF(z,y) \\ &= E[s_h(X,x)Y] \end{aligned} \quad (2.11)$$

where the weight function

$$s_h(z,x) = \frac{1}{\sigma_0^2}\left\{K_h(z-x)\left[\mu_2 - \mu_1(z-x)\right]\right\}.$$

Since, $\int s_h(z,x)\,dF(z,y) \equiv 1$, the $\tilde{l}(x)$ in 2.11 corresponds to a localized Fréchet mean.

$$\tilde{l}(x) = \underset{y\in\mathbb{R}}{\operatorname{argmin}}\, E\left[s_h(X,x)(Y-y)^2\right] \tag{2.12}$$

Hence the well-known local linear estimator of $m(x)$ in order to motivate the local Fréchet technique is:

$$\hat{l}(x) = \frac{1}{n}\sum_{i=1}^{n} \hat{s}_h(X,x)Y_i = \underset{y\in\mathbb{R}}{\operatorname{argmin}} \sum_{i=1}^{n} \hat{s}_h(X_i,x)(Y_i-y)^2. \tag{2.13}$$

The empirical weight function $\hat{s}_h$, as derived from the local linear least squares criterion, is;

$$\hat{s}_h(z,x) = \hat{\varrho}^{-2} K_h(z-x)\left[\hat{\mu}_2 - \hat{\mu}_1(z-x)\right], \tag{2.14}$$

where

$$\hat{\mu}_j = n^{-1}\sum_{i=1}^{n} K_h(X_i-x)(X_i-x)^j, \quad \hat{\varrho}^2 = \hat{\mu}_0\hat{\mu}_2 - \hat{\mu}_1^2,$$

and thus satisfies $n^{-1}\sum_{i=1}^{n} \hat{s}_h(X_i,x) = 1$. Hence, $\hat{l}(x)$ is a weighted average of the observed responses.

Now consider the case when $Y \in \Omega$ is a general metric space. The local Fréchet regression estimator of $m_{\oplus}(x)$ in (2.2) for a general metric space Ω is obtained by replacing the squared difference $(Y-y)^2$ in (2.13) by its appropriate counterpart in metric spaces, the squared distance. Hence the definition of the local Fréchet regression:

$$\tilde{l}_{\oplus}(x) = \underset{\omega\in\Omega}{\operatorname{argmin}}\, \tilde{L}_n(\omega),\ \tilde{L}_n(\omega) = E\left[s_h(X,x)\ \mathrm{d}^2(Y,\omega)\right].$$

2.3.1 Estimation

Consider the sample $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ be independent and identically distributed according to F, and an arbitrary $x \in \mathbb{R}$. The local Fréchet estimator is

$$\hat{l}_{\oplus}(x) = \underset{\omega \in \Omega}{\operatorname{argmin}} \sum_{i=1}^{n} \hat{s}_h(X_i, x)\ \mathrm{d}^2\left(Y_i, \omega\right) \tag{2.15}$$

where the weights are again given by (2.14). The criterion minimized in the right-hand side of (2.15) is, for each x and ω, a local linear estimator of the conditional expected value represented by $M_{\oplus}(\omega, x)$ in (2.2). Thus, the local Fréchet regression approach is equivalent to pointwise estimation of $M_{\oplus}$ by local linear regression, followed by its minimization over Ω.

Chapter 3

The Fréchet single index models for object response regression

The extension of the local Fréchet estimator for a covariate $\boldsymbol{x} \in \mathbb{R}^p$ for $p > 1$ is mathematically straightforward. However, its performance deteriorates quickly as the dimension p increases due to the curse of dimensionality. Thus the global Fréchet model may be preferable for a moderate p, despite its bias due to increased stability in the estimation procedure. Hence we attempted to utilize the strengths and mitigate the weaknesses of these two Fréchet approaches in the same way the semiparametric method does so for parametric and non-parametric estimators in the classical models. More specifically, we proposed the FSI model which assumes that the Fréchet regression function depends on the covariate $\boldsymbol{x} \in \mathbb{R}^p$, and the index parameter $\boldsymbol{\theta}_0 \in \mathbb{R}^p$ only through the index $\boldsymbol{\theta}_0^T \boldsymbol{x} \in \mathbb{R}$.

3.1 Model Definition

The estimation of the parameter $\boldsymbol{\theta}_0 \in \mathbb{R}^p$ was the primary target of our new model. It lends interpretability by specifying the contribution of each predictor in the model.

For identifiability [51] we define the parameter space as follows:

$$\Theta_p = \{\boldsymbol{\theta} \in \mathbb{R}^p : \text{the first non-zero element of } \boldsymbol{\theta} \text{ is positive, and } ||\boldsymbol{\theta}||_E = 1\}$$

Hence, $\boldsymbol{\theta}$ belongs to the surface of the unit sphere in $\mathbb{R}^p$. Hence, by this convention, $\Theta_1 = \{1\}$ for which the necessary theoretical foundation had been laid out as local Fréchet regression in [19]. Here our focus will be on analyzing $p \geq 2$.

A comprehensive discussion of a large class of single index models and their applications can be found in [20] when the response data is in $\mathbb{R}$, where the parameter $\boldsymbol{\theta}_0$ is estimated using the Semiparametric Least Square (SLS) method. The procedure that will be described for estimating the coefficient in the proposed model is inspired by this intuitive technique, and leverages local Fréchet regression and standard distance-based least squares.

Now we formally define the new model. Let $F_{\boldsymbol{X}}$ denote the marginal distribution of $\boldsymbol{X}$, with support $\mathcal{X} \subset \mathbb{R}^p$. For any $\boldsymbol{\theta} \in \Theta_p$, define the Fréchet regression function conditional on the projected variable $\boldsymbol{\theta}^T\boldsymbol{X}$ as

$$g_{\oplus}(u, \boldsymbol{\theta}) = \underset{\omega \in \Omega}{\operatorname{argmin}} \Lambda_{\oplus}(\omega, u, \boldsymbol{\theta}), \quad \Lambda_{\oplus}(\cdot, u, \boldsymbol{\theta}) = E(\mathrm{d}^2(Y, \cdot)|\boldsymbol{\theta}^T\boldsymbol{X} = u) \tag{3.1}$$

where $u \in \mathcal{U}_{\boldsymbol{\theta}} =: \{\boldsymbol{\theta}^T\boldsymbol{x} : \boldsymbol{x} \in \mathcal{X}\}$ and a minimizer is assumed to exist. Thus, the FSI model for $m_{\oplus}(\boldsymbol{x})$ in (2.2) is

$$m_{\oplus}(\boldsymbol{x}) = g_{\oplus}(\boldsymbol{\theta}_0^T\boldsymbol{x}, \boldsymbol{\theta}_0) \tag{3.2}$$

Given existence of the minimizers in (2.1), the identifiability of the index parameter $\boldsymbol{\theta}_0$ is equivalent to the statement

$$P(g_{\oplus}(\boldsymbol{\theta}^T\boldsymbol{X}, \boldsymbol{\theta}) \neq g_{\oplus}(\boldsymbol{\theta}_0^T\boldsymbol{X}, \boldsymbol{\theta_0})) > 0,$$

from which it can be deduced that

$$W(\boldsymbol{\theta}) = E\left(\mathrm{d}^2(Y, g_{\oplus}(\boldsymbol{\theta}^T\boldsymbol{X}, \boldsymbol{\theta}))\right), \tag{3.3}$$

the natural generalization of the least-squares criterion for metric spaces, is uniquely minimized at $\boldsymbol{\theta}_0$. Thus, the above criterion will be used to construct an M-estimator for $\boldsymbol{\theta}_0$. In a recent preprint, [52] independently investigated model (3.2), though using a slightly different strategy to estimate $W(\boldsymbol{\theta})$ than that employed in this paper.

3.2 Estimation

Suppose a random sample $(\boldsymbol{X}_i, Y_i)$, $i = 1, \ldots, n$, distributed according to F is available. As the true parameter $\boldsymbol{\theta}_0$ is unknown, we proceed to estimate the target in (3.2) in two steps. First, $g_{\oplus}(\boldsymbol{\theta}^T\boldsymbol{x}, \boldsymbol{\theta})$ is estimated for fixed $\boldsymbol{\theta}$ using local Fréchet regression, followed by optimization over $\boldsymbol{\theta}$. Let $h > 0$ be a given bandwidth and K a univariate probability density kernel, as before. The estimates in this section depend on h, although we suppress this dependence for simplicity in several formulae.

For a fixed $\boldsymbol{\theta} \in \Theta_p$, repurposing (2.14) and (2.15) for use with the predictors $\boldsymbol{\theta}^T\boldsymbol{X}_i$, we obtain the estimator

$$\hat{g}_{\oplus}(\boldsymbol{\theta}^T\boldsymbol{x}, \boldsymbol{\theta}) = \underset{\omega \in \Omega}{\operatorname{argmin}}\, \hat{\Lambda}_{\oplus}(\omega, \boldsymbol{\theta}^T\boldsymbol{x}, \boldsymbol{\theta}),\ \hat{\Lambda}_{\oplus}(\omega, \boldsymbol{\theta}^T\boldsymbol{x}, \boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \hat{r}_h(\boldsymbol{X}_i, \boldsymbol{x}, \boldsymbol{\theta})\, \mathrm{d}^2\left(Y_i, \omega\right). \tag{3.4}$$

Here, the weight function $\hat{r}_h : \mathbb{R}^p \times \mathbb{R}^p \times \Theta_p \to \mathbb{R}$ is

$$\hat{r}_h(\boldsymbol{z}, \boldsymbol{x}, \boldsymbol{\theta}) = \hat{\sigma}_{\boldsymbol{\theta}}^{-2}(\boldsymbol{x}) K_h(\boldsymbol{\theta}^T(\boldsymbol{z} - \boldsymbol{x}))\left[\hat{\mu}_{2,\boldsymbol{\theta}}(\boldsymbol{x}) - \hat{\mu}_{1,\boldsymbol{\theta}}(\boldsymbol{x})(\boldsymbol{\theta}^T(\boldsymbol{z} - \boldsymbol{x}))\right], \tag{3.5}$$

where, for $j = 0, 1, 2$,

$$\hat{\mu}_{j,\boldsymbol{\theta}}(\boldsymbol{x}) = n^{-1} \sum_{i=1}^{n} K_h(\boldsymbol{\theta}^T(\boldsymbol{X}_i - \boldsymbol{x}))(\boldsymbol{\theta}^T(\boldsymbol{X}_i - \boldsymbol{x}))^j \tag{3.6}$$

and $\hat{\sigma}^2_{\boldsymbol{\theta}}(\boldsymbol{x}) = \hat{\mu}_{0,\boldsymbol{\theta}}(\boldsymbol{x})\hat{\mu}_{2,\boldsymbol{\theta}}(\boldsymbol{x}) - \hat{\mu}_{1,\boldsymbol{\theta}}(\boldsymbol{x})^2$.

Utilizing this result, we construct a criterion for estimating $\boldsymbol{\theta}_0$ by defining an empirical version of (3.3). Replacing the expectation with the empirical distribution, and replacing $g_{\oplus}(\boldsymbol{\theta}^T\boldsymbol{X}_i, \boldsymbol{\theta})$ with the fitted value $\hat{Y}_i(\boldsymbol{\theta}, h) = \hat{g}_{\oplus}(\boldsymbol{\theta}^T\boldsymbol{X}_i, \boldsymbol{\theta})$ yields

$$W_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{d}^2(Y_i, \hat{Y}_i(\boldsymbol{\theta}, h)). \tag{3.7}$$

The coefficient vector $\boldsymbol{\theta}_0$ is then estimated by

$$\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(h) = \underset{\boldsymbol{\theta} \in \Theta_p}{\operatorname{argmin}}\, W_n(\boldsymbol{\theta}). \tag{3.8}$$

As is typically the case in this type of semi-parametric estimation approach, the bandwidth h cannot decay too quickly if one is to obtain a consistent estimator of $\boldsymbol{\theta}_0$. Indeed, Theorem 1 stated in [21] restricts the decay of h in a way that depends on the dimension p as well as the sample size n. Nevertheless, in constructing the final estimator $\hat{m}_{\oplus}(\boldsymbol{x})$ of the regression function $m_{\oplus}(\boldsymbol{x})$, a different smoothing bandwidth may be used, potentially improving the overall rate of convergence. Specifically, denote by $\tilde{g}(\boldsymbol{\theta}^T\boldsymbol{x}, \boldsymbol{\theta})$ the estimator in (3.4) for any θ and x using a bandwidth $\tilde{h} > 0$. Then the final regression estimator is

$$\hat{m}_{\oplus}(\boldsymbol{x}) = \tilde{g}_{\oplus}(\hat{\boldsymbol{\theta}}^T\boldsymbol{x}, \hat{\boldsymbol{\theta}}). \tag{3.9}$$

3.3 Simulation Study on Spherical Data

We implement our methodology when the responses lie on a Riemannian manifold object space. Let $\Omega = S^2$, the surface of the unit sphere in $\mathbb{R}^3$, with origin being the center. For any two points, the geodesic distance between them is given by the equation 1.2. We refer to a simulation setting as a unique combination of the sample size n, covariate dimension p, and noise level $\sigma^2 > 0$ that will be defined below.

3.3.1 Data Generation

For a given setting (n, p, σ^2), independent and identically distributed data pairs $(\boldsymbol{X}_i, Y_i) \in \mathbb{R}^p \times S^2$, $i = 1, \ldots, n$ were generated according to the following steps.

1. Independently generate predictor components X_{ij}, $j = 1, \ldots, p$, as $X_{ij} = W_{ij}/\sqrt{p}$, where $W_{ij} \overset{\text{iid}}{\sim} \mathcal{U}(-1, 1)$.

2. With $\boldsymbol{\theta}_0 = (\theta_{01}, \theta_{02}, ..., \theta_{0p})^T$ being the true parameter, compute the latent predictor $U_i = \boldsymbol{\theta}_0^T \boldsymbol{X}_i$.

3. Compute the conditional Fréchet mean at $\boldsymbol{X}_i$, depending only on U_i, as

$$m_{\oplus}(\boldsymbol{X}_i) = \left(\sqrt{\left(1 - \frac{U_i^2}{p}\right)} \cos\left(\frac{\pi U_i}{\sqrt{p}}\right), \sqrt{\left(1 - \frac{U_i^2}{p}\right)} \sin\left(\frac{\pi U_i}{\sqrt{p}}\right), \frac{U_i}{\sqrt{p}} \right).$$

4. Generate a noise vector $\boldsymbol{\mathcal{Z}}_i$ as follows. First, let $(\boldsymbol{V}_{i1}, \boldsymbol{V}_{i2})$ be an orthonormal basis for the tangent space $\text{span}\{m_{\oplus}(\boldsymbol{X}_i)\}^{\perp}$. Next, for a given noise level σ^2, generate $\boldsymbol{C}_i = (c_{i1}, c_{i2})^T \overset{iid}{\sim} N_2(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_2)$. Finally, set $\boldsymbol{\mathcal{Z}}_i = c_{i1}\boldsymbol{V}_{i1} + c_{i2}\boldsymbol{V}_{i2}$.

5. Generate the spherical response variable as

$$Y_i = \cos\left(\|\boldsymbol{\mathcal{Z}}_i\|_E\right) m_{\oplus}(\boldsymbol{X}_i) + \sin\left(\|\boldsymbol{\mathcal{Z}}_i\|_E\right) \frac{\boldsymbol{\mathcal{Z}}_i}{\|\boldsymbol{\mathcal{Z}}_i\|_E}.$$

Steps 4 and 5 produce a point Y_i on the sphere with conditional Fréchet mean equal to $m_{\oplus}(\boldsymbol{X}_i)$. To give an idea of what the responses look like relative to the conditional Fréchet mean function for a given noise level, Figure 3.1 shows example data sets and corresponding estimates for $p = 5$ under two noise scenarios ($\sigma^2 = 0.4$ and $\sigma^2 = 0.8$) and three sample sizes ($n = 50, 100, 200$).

3.3.2 Computational Details

For each simulated data set, estimation was performed using a grid for the bandwidth h. For given values $\boldsymbol{\theta}$ and h, the local Fréchet estimate $\hat{g}_{\oplus}(u, \boldsymbol{\theta})$ in (3.4) was obtained for values $u = \boldsymbol{\theta}^T \boldsymbol{X}_i$, $i = 1, \ldots, n$, using a non-convex optimization trust region algorithm as implemented in ManOpt toolbox for Matlab [53, 19]. As the algorithm requires an initial estimate, we computed the leave-one-out Nadaraya-Watson estimate

$$\tilde{Y}_{(i)}^{(NW)}(h, \boldsymbol{\theta}) = \frac{\sum_{l \neq i} Y_l K\left(\left[\boldsymbol{X}_i^T \boldsymbol{\theta} - \boldsymbol{X}_l^T \boldsymbol{\theta}\right]/h\right)}{\sum_{l \neq i} K\left(\left[\boldsymbol{X}_i^T \boldsymbol{\theta} - \boldsymbol{X}_l^T \boldsymbol{\theta}\right]/h\right)}$$

for each observed predictor values $\boldsymbol{X}_i$. Then, the initial estimate that is entered into the algorithm is obtained by projecting onto the sphere, i.e.

$$\hat{Y}_{(i)}^{(0)}(h, \boldsymbol{\theta}) = \frac{\tilde{Y}_{(i)}^{(NW)}(h, \boldsymbol{\theta})}{\|\tilde{Y}_{(i)}^{(NW)}(h, \boldsymbol{\theta})\|_E}.$$

Computation of the estimate $\hat{\boldsymbol{\theta}}(h)$ by optimizing the criterion W_n is the more challenging task, particularly for larger values of p, since there is no explicit form for the gradient or Hessian. Numeric evaluation of the gradient can also be quite expensive when n is large due to the need to repeatedly perform local Fréchet regression for each data point. In addition, any optimization procedure is sensitive to the starting value for $\boldsymbol{\theta}$, particularly for larger p, further increasing the computational burden since multiple

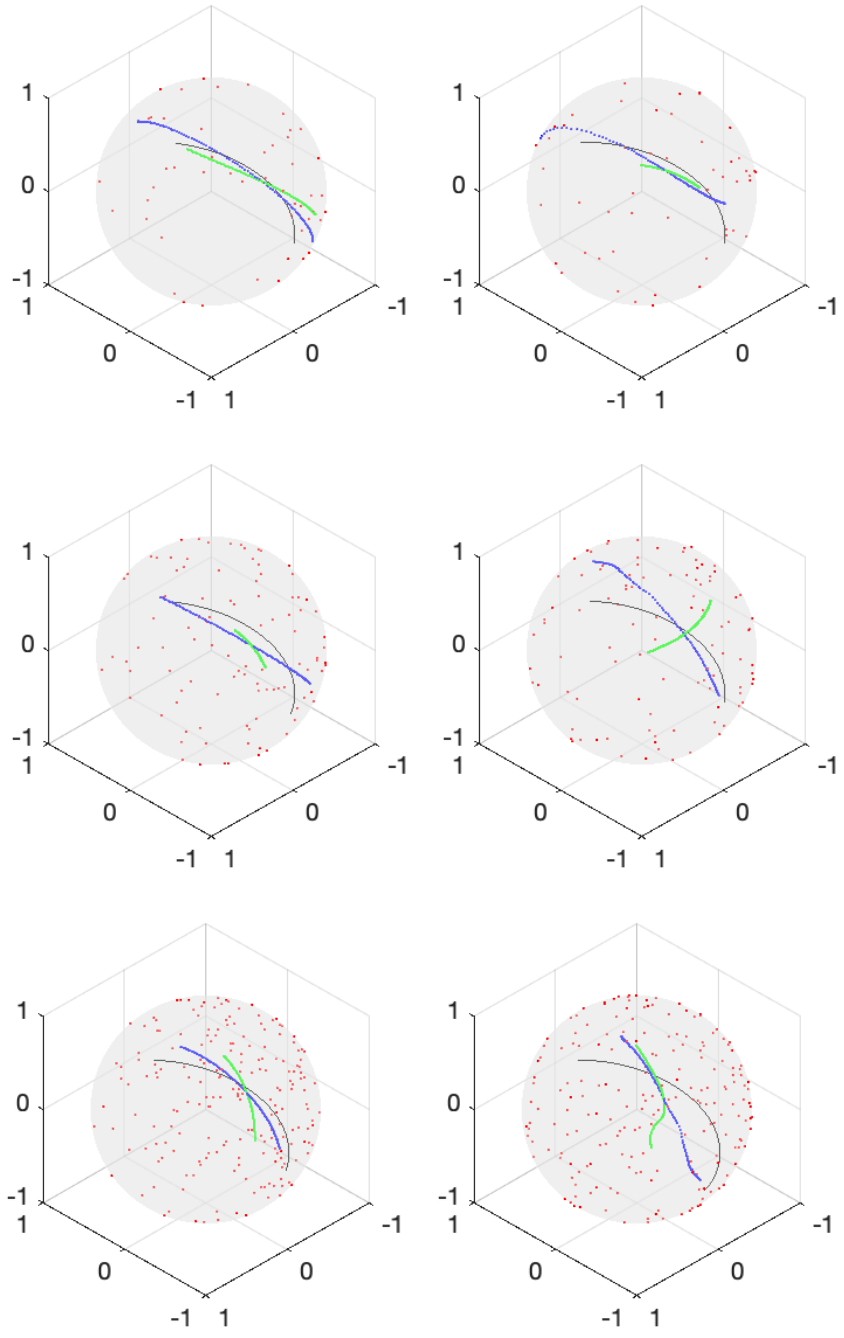

Figure 3.1: Examples of simulated data sets for covariate dimension $p = 5$, corresponding to sample sizes $n = 50$ (top row), $n = 100$ (middle row), and $n = 200$ (bottom row), and noise levels $\sigma^2 = 0.4$ (left column) and $\sigma^2 = 0.8$ (right column). The red dots represent values of Y_i in the sample, while the regression function values $m_{\oplus}(\boldsymbol{x})$ are shown by the black curve for $\boldsymbol{x} \in [0,1]^p$. The blue dots represent the FSI fitted responses for n observations using (3.9). The green dots are the fitted responses obtained by computing (3.4) for a value of $\boldsymbol{\theta}$ far from the true value $\boldsymbol{\theta}_0$.

starting values must be used. Therefore, we took the following approach.

First, a collection $\{\boldsymbol{\theta}_k : k = 1, \ldots, K_p\}$, of starting values was randomly generated for each setting (n, p, σ^2), with the same starting values being used for all data sets under that setting. The number of starting points was taken to be $K_2 = 10$, $K_5 = 50$, and $K_{10} = 100$, so that these increase with the dimension p. We then reduce this initial pool of starting values by optimizing a proxy to W_n given by

$$W_n^*(\boldsymbol{\theta}) = n^{-1} \sum_{i=1}^{n} d^2(Y_i, Y_i^*(h, \boldsymbol{\theta})), \tag{3.10}$$

where

$$Y_i^*(h, \boldsymbol{\theta}) = \frac{\sum_{j=1}^n r_h(\boldsymbol{X}_j, \boldsymbol{X}_i, \boldsymbol{\theta}) Y_j}{\|\sum_{j=1}^n r_h(\boldsymbol{X}_j, \boldsymbol{X}_i, \boldsymbol{\theta}) Y_j\|_E}$$

is the projection onto the sphere of the local linear estimate of the Euclidean regression function $E(Y|\boldsymbol{\theta}^T\boldsymbol{X} = u)$ at $u = \boldsymbol{\theta}^T\boldsymbol{X}_i$. The advantage of using this proxy is that an analytic gradient and Hessian for W_n^* are available, so that optimization of W_n^* is relatively fast. Using each of the K_p starting values, we obtain as many initial estimates $\tilde{\boldsymbol{\theta}}_k(h)$, $k = 1, \ldots, K_p$. This optimization was executed using the `fmincon` function in Matlab with the `trust-region-reflective` option for the optimizer. In this optimization, $\boldsymbol{\theta}$ was represented by its polar coordinates to handle the constraints in a simple way.

In the final optimization step, $\tilde{K}_p$ of the initial estimates $\tilde{\boldsymbol{\theta}}_k$ are retained as starting values based on having the lowest values of the proxy criterion W_n^*, with $\tilde{K}_2 = 2$, $\tilde{K}_5 = 3$, and $\tilde{K}_{10} = 5$. For each starting value, W_n is directly optimized using `fmincon` with the `SQP` option for the optimizer that does not require a gradient input, again using the polar representation of $\boldsymbol{\theta}$. The value of $\boldsymbol{\theta}$ that, at convergence, attains the lowest value of W_n is taken to be the estimate $\hat{\boldsymbol{\theta}}(h)$ for that bandwidth. Lastly, fitted values are computed using (3.9) by setting $\tilde{h} = h$, $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(h)$, and $\hat{Y}_i(h) = \hat{m}_{\oplus}(\boldsymbol{X}_i)$.

As a competitor to the FSI model, we also implemented a multivariate local Fréchet estimator. The estimator is defined as in (2.15), with the only difference being that the weights $\hat{s}_h(\boldsymbol{X}_i, \boldsymbol{x})$ are computed from multivariate local linear regression, since $\boldsymbol{X}_i \in \mathbb{R}^p$, using a product Gaussian kernel with the same bandwidth for each predictor. The optimization for this estimator was performed using the ManOpt trust region algorithm described above.

3.3.3 Performance Evaluation

Data were generated under 18 unique parameter settings using samples sizes $n = 50, 100, 200$, for noise levels $\sigma^2 = 0.4, 0.8$, and for dimensions $p = 2, 5, 10$, with 200 simulation runs per setting. Let $s = 1, \ldots, 200$ be the index for simulations within a given setting, and $(\boldsymbol{X}_i^s, Y_i^s)$ denoted the simulated data. Then, from each simulated data set and bandwidth we obtain an estimate $\hat{\boldsymbol{\theta}}^s(h) \in \mathbb{R}^p$ and fitted values $\hat{Y}_i^s(h)$, $i = 1, \ldots, n$ from the FSI model, as well as fitted values $\check{Y}_i^s(h)$ from the multivariate local Fréchet (mLF) estimator. The following performance metrics were computed for each simulated data set across the entire range of bandwidths.

1. As the parameter space Θ_p is a subset of the $(p-1)$-dimensional unit sphere, a natural measurement of empirical squared error for the s-th simulated data set is

$$\mathrm{SE}(\hat{\boldsymbol{\theta}}^s(h)) = \left[\arccos\left(\left|\boldsymbol{\theta}_0^T \hat{\boldsymbol{\theta}}^s(h)\right|\right)\right]^2, \tag{3.11}$$

   where we have introduced the absolute value to account for the fact that $\boldsymbol{\theta}_0$ and $-\boldsymbol{\theta}_0$ are indistinguishable from the data.

2. To evaluate the estimation error in regression for the FSI model, the mean square

estimation error (MSEE) for the s-th simulated data set was quantified by

$$\text{MSEE}^{(s)}_{\oplus,\text{FSI}}(h) = \frac{1}{n}\sum_{i=1}^{n}\left[\arccos\left(m_{\oplus}(\boldsymbol{X}_i^s)^T\hat{Y}_i^s(h)\right)\right]^2 \tag{3.12}$$

3. To evaluate the estimation error in regression for the multivariate local Fréchet estimator, the mean square estimation error (MSEE) for the s-th simulated data set was quantified by

$$\text{MSEE}^{(s)}_{\oplus,\text{mLF}}(h) = \frac{1}{n}\sum_{i=1}^{n}\left[\arccos\left(m_{\oplus}(\boldsymbol{X}_i^s)^T\check{Y}_i^s(h)\right)\right]^2 \tag{3.13}$$

Tables 3.1 and 3.2 show empirical performance metrics for the various simulation settings considered. In these tables, the average and standard deviation of each metric across simulations is reported. For each metric, the reported values are for the bandwidth value in the chosen grid that minimizes the corresponding average across simulations. We observe that the average squared estimation errors and their standard deviations for the FSI estimator of the coefficient $\boldsymbol{\theta}_0$, and both FSI and mLF estimators of the regression function $m_{\oplus}(x)$, all behave in the expected fashion. Namely, they decay toward zero with increasing sample size and are larger for higher values of p and for the higher noise level. However, the FSI regression estimation errors are overall smaller than those of the multivariate local Fréchet regression estimator when both are evaluated using their optimal bandwidth, with differences becoming more pronounced for larger covariate dimensions p.

Next, we more closely examine the empirical sampling distribution of $\hat{\boldsymbol{\theta}}(h)$ across different values of n for $p = 2$, since these can be easily visualized via histograms of the (scalar) polar coordinate representations $\hat{\eta}(h)$. Specifically, Figure 3.2 shows the empirical distribution of $\hat{\eta}^{(s)}(h)$ for different values of n and σ^2, where h is the same

Table 3.1: Simulation results for settings with low noise, $\sigma^2 = 0.4$. Here p and n are covariate dimension and sample size, respectively. The third column is the average of the values $\mathrm{SE}(\hat{\boldsymbol{\theta}}^s(h))$ from (3.11) across simulations, with standard deviation in parentheses. Columns 4 and 5 give the averages of $\mathrm{MSEE}^{(s)}_{\oplus,\mathrm{FSI}}(h)$ and $\mathrm{MSEE}^{(s)}_{\oplus,\mathrm{mLF}}(h)$ from (3.12) and (3.13), respectively, across simulations, with standard deviation given in parentheses. For each of the metrics in columns 3–5, results are shown for the bandwidth that minimizes the reported average of that metric and are rounded to 3 significant digits.

p	n	Avg. MSE	Avg. $\mathrm{MSEE}_{\oplus,\mathrm{FSI}}$	Avg. $\mathrm{MSEE}_{\oplus,\mathrm{mLF}}$
2	50	0.032 (0.047)	0.063 (0.039)	0.078 (0.042)
	100	0.014 (0.020)	0.030 (0.017)	0.040 (0.020)
	200	0.006 (0.008)	0.016 (0.008)	0.021 (0.010)
5	50	0.326 (0.283)	0.100 (0.051)	0.143 (0.054)
	100	0.168 (0.132)	0.050 (0.026)	0.074 (0.029)
	200	0.071 (0.056)	0.023 (0.012)	0.036 (0.015)
10	50	0.938 (0.519)	0.166 (0.064)	0.251 (0.081)
	100	0.544 (0.386)	0.082 (0.038)	0.128 (0.038)
	200	0.285 (0.152)	0.039 (0.016)	0.065 (0.018)

minimizing bandwidth used to compute the average of the $\mathrm{SE}(\hat{\boldsymbol{\theta}}^s(h))$ values for $p = 2$ in Tables 3.1 and 3.2 for $\sigma^2 = 0.4$ and $\sigma^2 = 0.8$, respectively. For reference, the true polar coordinate parameter η_0 is superimposed as the red vertical line. In all cases, as n increases the empirical sampling distribution becomes more concentrated near η_0.

Finally, to more fully examine the estimation performance of the overall regression function $m_\oplus(\boldsymbol{x})$ more closely, Figure 3.3 juxtaposes the boxplots of $\mathrm{MSEE}^{(s)}_{\oplus,\mathrm{FSI}}(h)$ from (3.12) for each simulation setting on the log scale, where h is the minimizing bandwidth used for this metric in Tables 3.1 and 3.2. The variation increases with p, but under each p it decreases with n. These reflect the numerical summaries given in Tables 3.1 and 3.2.

3.4 Regression of Mortality Distributions

3.4.1 Fréchet Regression with the Distributions as response

To demonstrate the application of our method, we consider human mortality data at the country level. The goal is to model the dependence of age-at-death distributions for a given year based on country-specific covariates. For this illustration, the year 2013 was selected, and human mortality data were sourced for 39 countries from the Human Mortality Database (HMD, [54] www.mortality.org) for this year. The HMD provides data for 41 countries; Hong Kong and Taiwan were omitted due to lack of availability of records for all covariates used in this illustrative example. The data for each country are structured as life-tables; for integer-valued age j, $0 \leq j \leq 110$, the life table provides the size of the population m_j which is at least j years old, normalized so that the total

Table 3.2: Simulation results for the settings with high noise, $\sigma^2 = 0.8$. Descriptions of column names and contents correspond to those given in Table 3.1.

p	n	Avg. MSE	Avg. $\text{MSEE}_{\oplus,\text{FSI}}$	Avg. $\text{MSEE}_{\oplus,\text{mLF}}$
2	50	0.154 (0.284)	0.231 (0.160)	0.285 (0.177)
	100	0.090 (0.244)	0.130 (0.107)	0.163 (0.100)
	200	0.025 (0.037)	0.063 (0.038)	0.084 (0.048)
5	50	1.038 (0.624)	0.376 (0.180)	0.558 (0.238)
	100	0.680 (0.555)	0.208 (0.118)	0.298 (0.142)
	200	0.367 (0.350)	0.100 (0.056)	0.143 (0.057)
10	50	1.481 (0.528)	0.496 (0.190)	0.927 (0.302)
	100	1.297 (0.578)	0.298 (0.104)	0.535 (0.171)
	200	0.869 (0.477)	0.160 (0.069)	0.276 (0.083)

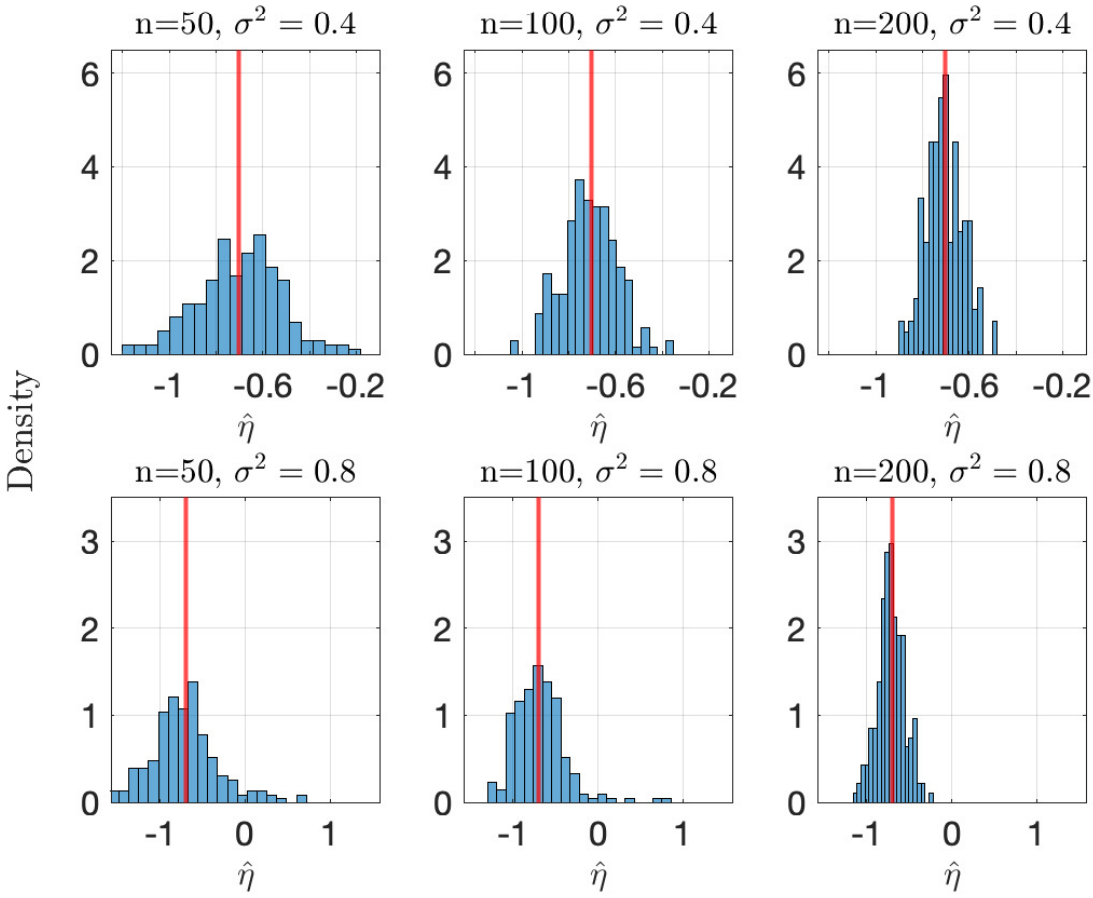

Figure 3.2: For $p = 2$ and sample sizes $n = 50$ (left panels), $n = 100$ (middle panels), $n = 200$ (right panels) the simulated empirical distributions of $\hat{\eta}(h)$, the polar coordinate of $\hat{\boldsymbol{\theta}}^s(h)$, are represented by histograms, with h chosen to minimize the average of $\mathrm{SE}(\hat{\boldsymbol{\theta}}^s(h))$ across simulations. In the top and bottom rows we have low noise ($\sigma^2 = 0.4$) and high noise ($\sigma^2 = 0.8$) scenarios respectively. The vertical red line represents the polar coordinate of $\boldsymbol{\theta}_0$, η_0 on the floor of the plot.

population is $m_0 = 100,000$. By computing differences, one can compute histograms of age-at-death that are specific to each country and year. In order to focus on adult mortality, we consider the histogram over the age range $[20, 110]$.

The impacts of many socioeconomic, environmental, and other variables on health outcomes have been extensively researched. For this illustration, we chose five covariates that, intuitively, have strong potential to influence mortality patterns of a nation. These

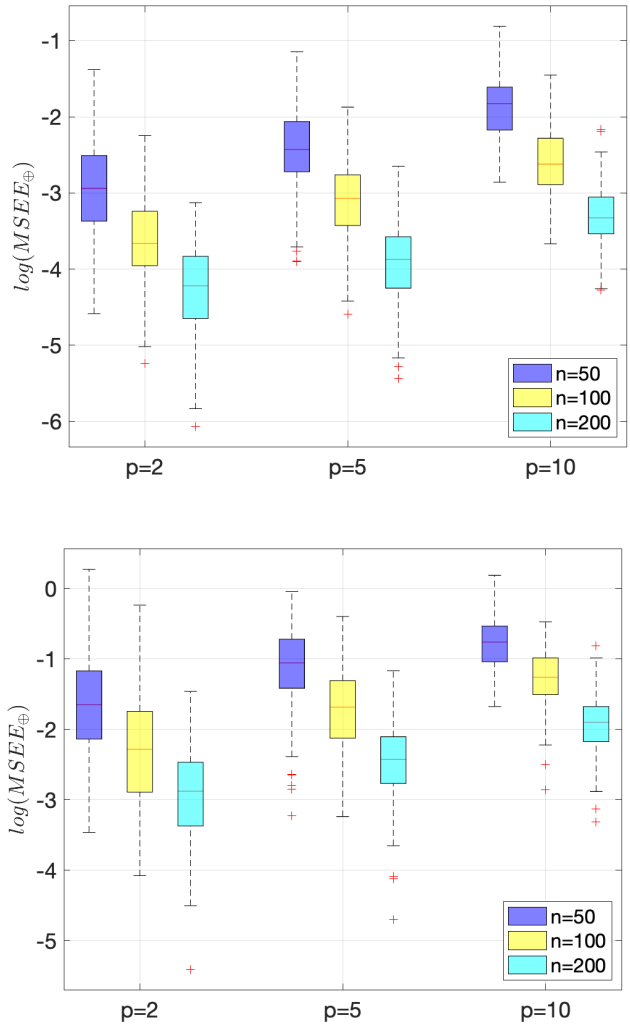

Figure 3.3: For each covariate dimension $p = 2, 5, 10$; boxplots of $\log(\text{MSEE}^{(s)}_{\oplus,\text{FSI}}(h))$ from (3.12) are given over all simulations for the optimizing bandwidths used in Tables 3.1 and 3.2, in each panel from left to right for sample sizes $n = 50, 100, 200$ as indicated by blue, yellow, and cyan in the plot, respectively. The top and bottom panels correspond to low and high noise scenarios, respectively, with different vertical axis ranges.

include year-on-year percentage change in GDP (GDPC [55]), carbon dioxide emissions in metric tons per capita (CO2E [56]), current health care expenditure as a percentage of GDP (HCE [57]), the human development index (HDI [58]), and infant mortality per 1000 live births (IM [59]) [60, 61, 62, 63, 64, 65, 66, 67, 68, 69]. Hence, $\boldsymbol{X}_i \in \mathbb{R}^5$ constitutes the covariate vector for the i-th country, $i = 1, \ldots, 39$.

To apply the proposed FSI model, Let Y_i represent the empirical quantile distribution of the age-at-death data for the i-th country, and $\boldsymbol{X}_i$ the vector of covariates for the i-the country for the year 2013. As random object responses Y_i are assumed to belong to Ω, the L^2-Wasserstein space, whose metric between two points is given by 1.3. The density histograms constructed from the lifetables were smoothed and then used to produce a quantile function for each country. This smoothing step was performed using the `CreateDensity` function in the R package `frechet` in order to obtain a smooth density, with the default cross-validated bandwidth choice, followed by conversion to a quantile function using the function `dens2quantile` in the package `fdadensity` [70, 71]. These constructed distributions will be referred to as observed distributions, and are visualized in Figure 3.4.

While one may, to some extent, employ linear methods to analyze such data, practical and theoretical problems emerge even in this simple case. From a practical standpoint, certain critical outputs, such as fitted values, that should be distribution-valued may not be so when linear methods are applied. These may be easily remedied using an ad hoc correction, but this is a clear disadvantage compared to the object treatment provided by Fréchet methods that will always respect such constraints. Beyond estimation, use of the non-linear geometry has distinct advantages when it comes to inference, particularly in the formulation of error models and uncertainty assessment, even in the setting of univariate distributions [72, 73, 7]. In addition, although univariate distributions are employed in this illustrative example, the model is equally applicable to multivariate

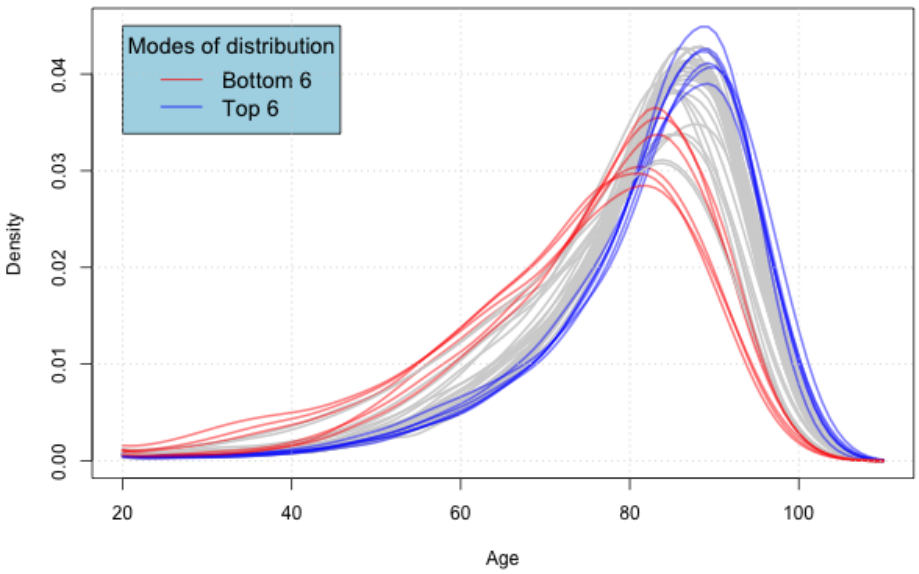

Figure 3.4: The estimated densities for each country for year 2013 over the age interval [20,110]; the countries with top 6 and bottom 6 mode ages are highlighted in blue and red colors respectively. The red colored densities include Russian Federation, Belarus, Ukraine, Hungary, Slovakia, Latvia. The densities in blue include Australia, Canada, Spain, France, Japan, Switzerland.

distributions [74], in which case the Wasserstein space is no longer flat.

Letting $(\boldsymbol{X}, Y)$ denote a generic covariate-distribution pair, the target is the Fréchet regression function $m_{\oplus}$ as defined in (2.2), for which we will assess seven competing models for object data. Specifically, $m_{\oplus}$ was estimated using global and local Fréchet regression techniques, the latter for each individual predictor, yielding six competitors to the proposed FSI model in (3.2).

3.4.2 Computational details

The computations for global and local Fréchet estimates, the letter for any fixed bandwidth, were carried out using the existing functionalities of the `frechet` package [70]. For the FSI model, for any specified $\boldsymbol{\theta}$ and bandwidth h, this package was also used

to compute $\hat{g}_{\oplus}$ in (3.4). To estimate $\boldsymbol{\theta}_0$ via (3.8), the `optim` command was used with option `"L-BFGS-B"` [75] with a lattice of $3^4 = 81$ starting points of polar coordinates $\eta \in [-\pi/2, \pi/2]^4$. The predictors were each centered and scaled to have sample mean zero and unit sample variance prior to fitting all models. For simplicity we use the same acronyms for the standardized covariates as previously given for the unstandardized ones, with the $\boldsymbol{X}_i$ values in each model being on the standardized scale.

As a first step, for each of the local Fréchet regression fits and the FSI model fit, a single bandwidth was selected by leave-one-out cross validation on the entire data set; no bandwidth is needed for global Fréchet regression. With m denoting a model index corresponding to the FSI model or one of the local Fréchet fits, let $\hat{Y}_i^{(m,-i)}(h)$ denote the fitted value for the i-th country produced by the estimate of model m using all countries except county i and with bandwidth h. Then the chosen bandwidth is

$$h_m^* = \underset{h \in \mathfrak{H}_m}{\operatorname{argmin}} \sum_{i=1}^{n} d_W^2(Y_i, \hat{Y}_i^{(m,-i)}(h)), \tag{3.14}$$

where $\mathfrak{H}_m$ is a grid of potential bandwidth choices for the given model. For the local Fréchet fits of each individual predictor, this step was executed using built-in functionalities of the `frechet` package. For the FSI bandwidth, the model was fit for each bandwidth in a pre-defined grid as described above, then h_{FSI}^* was computed as in (3.14).

3.4.3 Model Comparisons

To assess model performance, two metrics were computed. The first metric, termed the Fréchet R^2, quantifies the quality of model fit by in-sample performance. Specifically, for a given model m, let $\hat{Y}_i^{(m)}$ denote the fitted value that it produces for the i-th country.

Furthermore, let

$$\hat{\omega}_{\oplus} = \underset{\omega \in \Omega}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \mathrm{d}_W^2(Y_i, \omega)$$

denote the sample Fréchet mean. Indeed, this is simple to compute due to the nature of d_W in (1.3), as it is known that $\hat{\omega}_{\oplus}$ is the distribution with quantile function $n^{-1} \sum_{i=1}^{n} Y_i$. The Fréchet R^2 for model m is

$$R^2_{\oplus,m} = 1 - \frac{\sum_{i=1}^{n} \mathrm{d}_W^2(Y_i, \hat{Y}_i^{(m)})}{\sum_{i=1}^{n} \mathrm{d}_W^2(Y_i, \hat{\omega}_{\oplus})}, \tag{3.15}$$

which measures the proportion of Wasserstein-Fréchet variability in the data that is explained by the model.

The second performance metric is based on out-of-sample performance, in which the data were randomly split into a testing set of size 10 and training set of size 29, with 30 distinct random splits being executed. With $k = 1, \ldots, 30$ representing the index of each unique split of the data, denote by $Y_{[k,j]}$, $j = 1, \ldots, 10$, the age-at-death distribution for the j-th country in the k-th testing set, and by $\hat{Y}^{(m)}_{[k,j]}$ the predicted distribution for the same country using the fit of model m produced by the k-th training set. The error for the k-th split and model m is then quantified by

$$\mathrm{MSPE}_k^{(m)} = \frac{1}{10} \sum_{j=1}^{10} \mathrm{d}_W^2 \left(Y_{[k,j]}, \hat{Y}^{(m)}_{[k,j]} \right). \tag{3.16}$$

For local Fréchet and FSI model fits, the bandwidth used for each training set was fixed to be the value h_m^* in (3.14).

Table 3.3 gives the computed metrics for all models. The top three models in terms of Fréchet R^2 are the proposed FSI model, the local Fréchet fit using the HDI covariate, and the global Fréchet model. Figure 3.5 plots the fitted distributions (as densities) for these three models, along with the observed densities. The plot provides a visual

Table 3.3: Performance metrics for comparing seven Fréchet regression fits in three classes of models: (GF) global Fréchet, (LF) local Fréchet, (FSI) Fréchet single index. The predictor used for each local Fréchet fit is indicated for each subcolumn below LF: (HDI) human development index; (HCE) current health care expenditure as a percentage of GDP; (GDPC) GDP year-over-year percentage change in GDP; (IM) infant mortality; (CO2E) carbon dioxide emissions in metric tonnes per capita. The $R^2_\oplus$ row gives the Fréchet R^2 values defined in (3.15). The MSPE row gives the average out-of-sample mean-square prediction error (MSPE), defined in (3.16), across the 30 data splits. The SD (MSPE) row gives the standard deviation of the out of sample prediction errors across the 30 data splits.

Evaluation Measures	GF	LF					FSI
		HDI	HCE	GDPC	IM	CO2E	
$R^2_\oplus$	0.697	0.688	0.521	0.132	0.433	0.162	0.827
MSPE	6.23	6.87	6.93	13.51	12.08	13.74	4.35
SD(MSPE)	2.45	5.45	3.44	4.82	10.03	5.40	2.11

reinforcement of the Fréchet R^2 findings as these three models all produce distribution fits that approximate the observed distributions reasonably well.

Using out-of-sample performance, the FSI model emerges as the best model with the lowest average MSPE of 4.35. The left panel of Figure 3.6 shows boxplots of the 30 different $\text{MSPE}_k^{(m)}$ values for each model across splits, reinforcing the metrics in Table 3.3. In addition to having the smallest median MSPE value, the dispersion across folds for the FSI is among the lowest, second only to the global Fréchet model. The global Fréchet model suffers from model-induced bias, while the local Fréchet estimates using HDI lack relevant information from other variables and suffer from poor prediction in certain data splits. As designed, the FSI model balances the strengths of these two models. However, these results do not examine the relative performance of these models for each individual

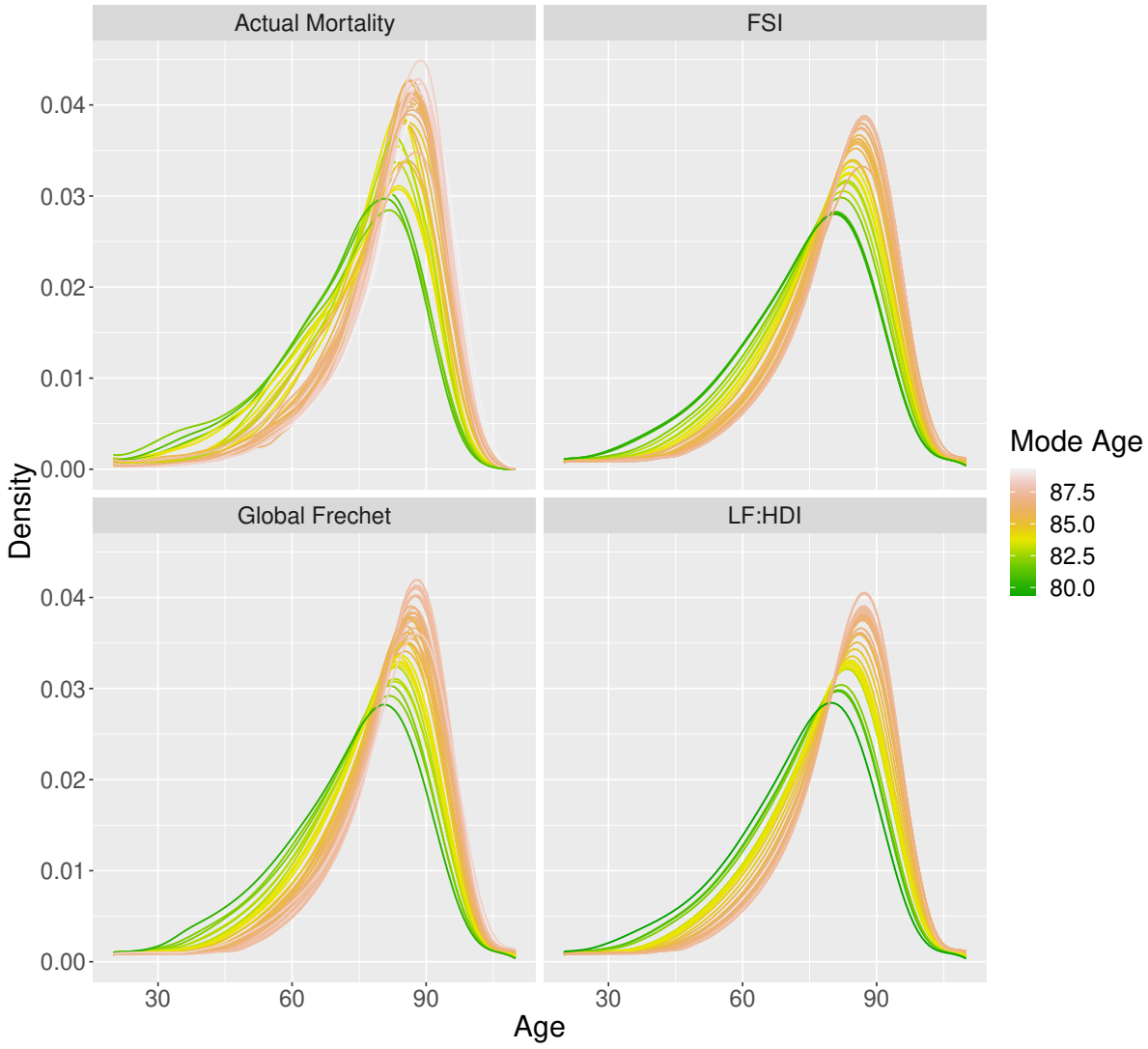

Figure 3.5: Observed smooth densities (top left) along with their fits produced by the proposed FSI model (top right), global Fréchet model (bottom left), and local Fréchet regression with HDI as predictor (bottom right). Densities are colored by the mode of the age-at-death distribution.

split of the data. The right panel of Figure 3.6 shows the boxplots of the logarithm of the ratio of MSPEs for each of three competing models (global Fréchet and local Fréchet estimates using HDI and HCE, respectively) to the MSPEs of FSI across splits. This comparison shows FSI as the best in overall out-of-sample prediction, as its prediction error is smaller than that of the other top-performing models for the majority of the 30

training/test data splits.

Next, we interpret the coefficient estimate for the FSI model. Rounded to three digits after the decimal, this was

$$\hat{\boldsymbol{\theta}} = (0.667,\ 0.741,\ -0.067,\ 0.005,\ 0.046)^T.$$

with the order of standardized covariates being human development index (HDI), health-care expenditure as percentage of GDP (HCE), year-on-year percentage change in GDP (GDPC), infant mortality per 1000 live births (IM), carbon dioxide emissions metric tons per capita (CO2E). The estimated coefficients for HDI and HCE have the highest magnitudes of 0.667 and 0.741 respectively, indicating their heavy influence relative to the other three predictors on the index $\hat{U}_i = \hat{\boldsymbol{\theta}}^T \boldsymbol{X}_i$ that drives the FSI fit, when all variables are in the model. As the FSI fit can be viewed as a local Fréchet estimate based on the univariate predictor $\hat{U}_i$, the superiority of the FSI model to the local Fréchet fit using either the HDI or HCE as predictor indicates that the combined predictive power of HDI and HCE, as quantified by the projection direction $\hat{\boldsymbol{\theta}}$, is stronger than either individual predictor when using local Fréchet regression. On the other hand, the global Fréchet model also combines the influence of all predictors, but does so less efficiently due to bias in the underlying model.

Since HDI and HCE appear to have relatively higher importance as predictors of mortality distributions for the local Fréchet regression as well as for the FSI model in terms of both in-sample and out-of-sample performance, it was interesting to explore how a small change in standardized value of HDI or HCE would affect the mortality distribution prediction of FSI model, while keeping all other covariates fixed at their median values. Figure 3.7 shows the age-at-death distributions predicted by the fitted FSI model. As expected, higher HDI or HCE are associated with increased longevity.

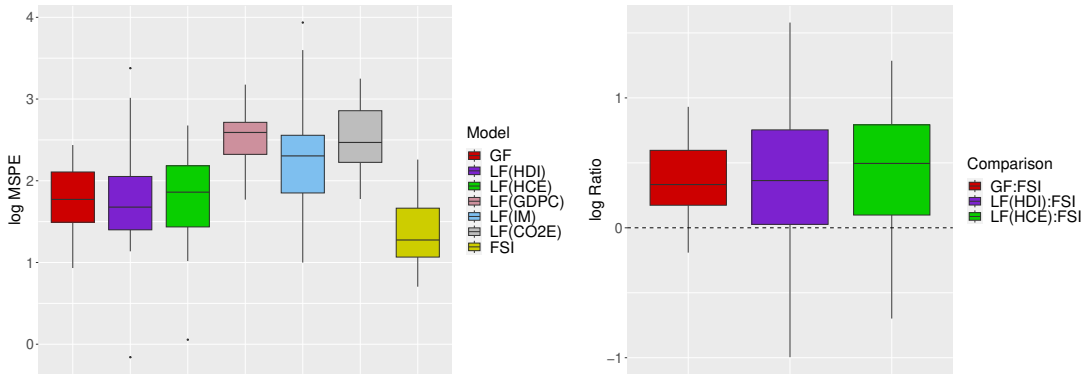

Figure 3.6: Left panel: boxplots of $\text{MSPE}_k^{(m)}$ values from (3.16) across splits for the following estimates from left to right: global Fréchet (GF); local Fréchet for each of the predictors human development index (HDI), healthcare expenditure as percentage of GDP (HCE), year-over-year percentage change in GDP (GDPC), infant mortality per 1000 live births (IM), and CO_2 emissions in metric tonnes per capita (CO2E); and Fréchet single index (FSI). Right panel: boxplots of log of ratio of the MSPEs from global Fréchet estimates (dark red, left), local Fréchet estimates using HDI (dark purple, midde), and HCE (green, right) relative to those of FSI are shown. The MSPE values of each competitor are higher than the FSI values for more than 75% of the folds, shown by the first quartile of the log-ratios being above the dotted horizontal line.

In particular, the plots suggests that the mode of mortality distributions increases for higher values of HDI or HCE, keeping other covariates fixed.

3.5 Discussion

The Fréchet single index model developed in this paper offers an alternative to global and local Fréchet regression for random object response data with vector-valued predictors in the spirit of semiparametric regression. While global Fréchet regression comfortably accommodates multiple predictors, it can be unduly rigid for many complex data settings. Indeed, even in the special case $\Omega = \mathbb{R}$, in which global Fréchet is multiple linear regression, such a model is often inadequate, so that its inadequacy in more complex

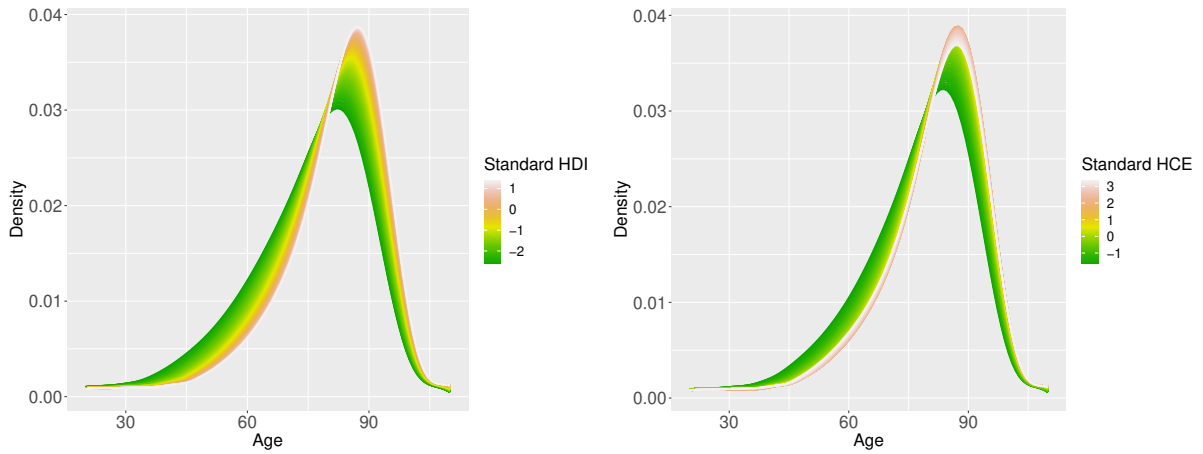

Figure 3.7: Age-at-death densities fitted by the FSI model for varying values of HDI (human development index, left) and HCE (health care expenditure, right), with other variables at their sample median. Colors indicate regularly spaced standardized values of the covariate.

metric spaces Ω is more likely than not. Local Fréchet regression, on the other hand, is unattractive when multiple predictors are present on both theoretical and practical grounds, despite its flexibility. Indeed, the data illustration involving mortality profiles demonstrates that the FSI model outperforms both global Fréchet regression and the best single-predictor model fitted using local Fréchet regression. Future extensions of the FSI model to handle more complex predictors, such as high-dimensional, functional, or object-valued data, will be valuable assets.

The technical issue surrounding existence and uniqueness of Fréchet means, whether marginal or conditional, has been circumvented in this work by assumption, although specific concrete examples of spaces satisfying the relevant assumption (M) have been provided due to the work of others on this challenging topic. Nevertheless, as pointed out by reviewers, a particular limitation of the FSI model is its requirement that the conditional Fréchet means $m_{\oplus}(\boldsymbol{x})$ in (2.2) not only exist for each $\boldsymbol{x}$, but that those conditional on $\boldsymbol{\theta}^T\boldsymbol{x}$, namely $g_{\oplus}(\boldsymbol{\theta}^T\boldsymbol{x}, \boldsymbol{\theta})$ in (3.1), exist and are unique for every $\boldsymbol{\theta}$. Examples can be quickly constructed in which the FSI model holds while $g_{\oplus}(\boldsymbol{\theta}^T\boldsymbol{x}, \boldsymbol{\theta})$ are only unique for

$\boldsymbol{\theta}$ equal to or in a neighborhood of $\boldsymbol{\theta}_0$. It seems plausible that one should still be able to estimate $\boldsymbol{\theta}_0$ in this setting, yet the methods proposed in this paper are inadequate. It is likely that criterion functions less restrictive than (3.3) may provide a path, and we leave this for future work.

While we have used a generalized version of semiparametric least squares for the estimation of the coefficient vector and local Fréchet regression to estimate $g_{\oplus}$ in (3.1), other options are of course available. For example, projection pursuit [76, 77], average derivatives [78], the conditional minimum average variance estimation (MAVE) technique [79], and sliced inverse regression [80], among others, have been validated practically and theoretically for scalar responses. Such approaches could conceivably work for object responses as alternatives to the method presented here for estimating the coefficient in the FSI model. More broadly, alternative smoothing methods could be developed for the estimation of the link function $g_{\oplus}$, although local Fréchet regression and the Nadaraya-Watson estimator [81] seem to be the only available options to date for a general metric space. Other semiparametric approaches for scalar data, such as multiple index models, may well prove to be adaptable to this scenario, although their extensions are less obvious.

Chapter 4

The partially linear Fréchet single index Regression model

4.1 Quantile distributional physical activity representations

We adopt a novel representation of the resulting data that extends previous compositional metrics to a functional setting [28], aimed at overcoming their dependency on certain physical activity intensity thresholds. This approach also overcomes some previously known limitations of more traditional approaches. Let $i \in \{1, 2, \ldots, n\}$ be the index for participants, where n is the total number of participants in the study. For the i-th participant, let M_i indicate the number of days (including partial days) for which accelerometer records are available and n_i be the number of observations recorded in the form of pairs (m_{ij}, A_{ij}), $j = 1, \ldots, n_i$. Here, the m_{ij} are a sequence of time points in the interval $[0, M_i]$ in which the accelerometer records activity information and A_{ij} is the measurement of the accelerometer at time m_{ij}.

In this paper, each individual's accelerometer measurements $\{A_{ij}\}_{j=1}^{n_i}$, $i = 1, \ldots, n$, are studied without regard for their ordering. We consider the empirical quantile function, $Y_i(t) = \hat{Q}_i(t)$, for $t \in [0, 1]$, as the response in the regression model. Here, $\hat{Q}_i(p) = \inf\{t \in \mathbb{R} : \hat{F}_i(p) \geq p\}$ is the generalized inverse of the empirical cumulative distribution function $\hat{F}_i(a) = \frac{1}{n_i} \sum_{j=1}^{n_i} 1\{A_{ij} \leq a\}$, $a \in \mathbb{R}$ for the physical activity values for the i-th individual. In order to illustrate clearly the difficulty of analyzing raw physical activity data in which different participants are monitored during different periods and in different experimental conditions, Figure 4.1 shows the plot of observed A_{ij} against m_{ij} for an arbitrary participant in our study. In Figure 4.2 the left panel shows the empirical quantile of the physical activity distribution of the participant whose raw measurements are shown in Figure 4.1, the right panel shows the empirical quantile functions of all participants after transforming the raw time series physical activity data into distributional quantile physical activity representation. Quantile physical activity representation overcomes the problem of summary physical activity when the raw time series have different lengths. In addition, the new representation uses all accelerometer intensities (over a continuum) to construct the new physical activity functional profile, unlike traditional representations of physical activity that summarize the information in a compositional vector.

4.1.1 Details of covariates

Four thousand six hundred sixteen individuals were chosen for our analyses, with physical activity monitoring available for at least 10 hours per day for four days. The covariates used in the model include socio-demographic, physical activity, dietary, and clinical variables such as age, Body Mass Index (BMI), Healthy Eating Index (HEI), along with the categorical variables Ethnicity and Sex. Age at the time of the analysis was 20

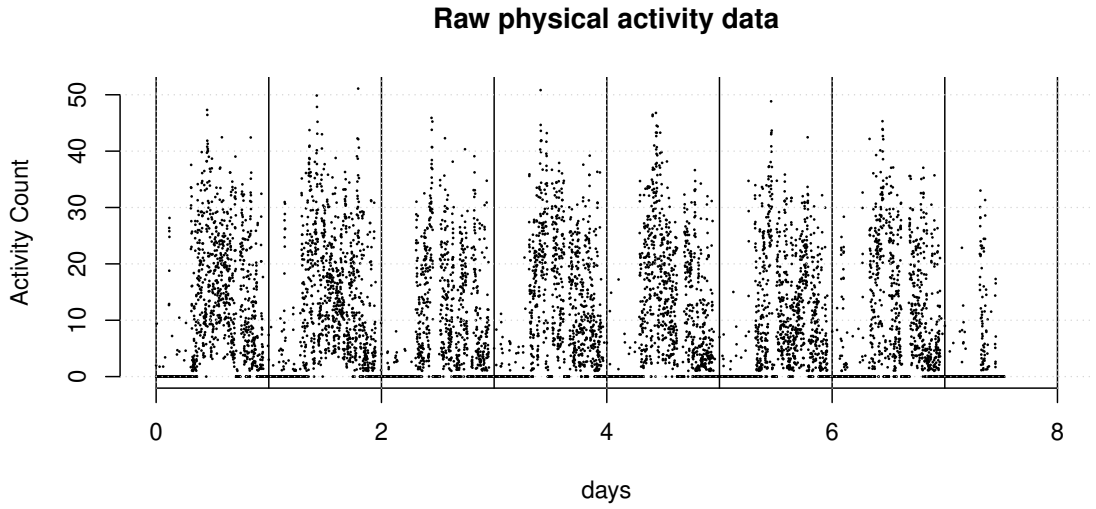

Figure 4.1: The plot of physical activity time series A_{ij} of one representative participant (one chosen i) in the NHANES 2011-2014 study monitored during 8 days are plotted over the observed time intervals m_{ij}, when the physical activity measurements are counted as described in the section 4.1.

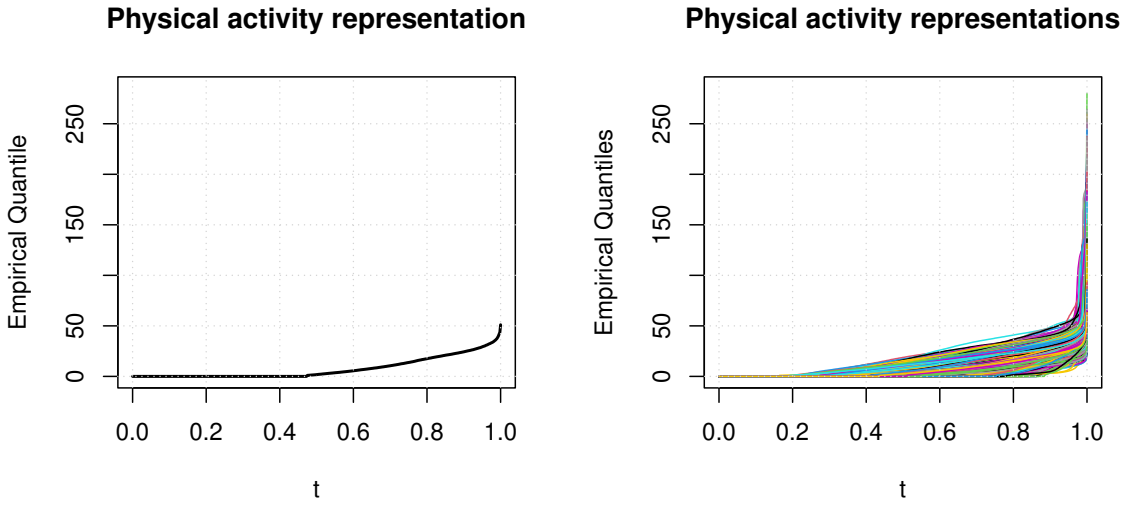

Figure 4.2: (Left panel) The empirical quantile representation $\hat{Q}_i$ (described in 4.1), of the activity profile of the participant (chosen i) in figure 4.1. (Right panel) The estimated empirical quantiles of the physical activity profiles are computed for all the 4616 participants in the study and plotted here. This helps to visualize the quantile representation of the participant i in comparison with the rest of the participants in our study.

to 80 years. The BMI (Kg/m^2) was restricted to $18.5 - 40$ to study individuals ranging from healthy to highly overweight/obese. The variable (HEI) was utilized, indicating a global score about the diet quality. The ethnicity variable reported the racial origin of the participants divided into the following seven categories: Mexican American, Other Hispanic, Non-Hispanic White, Non-Hispanic Black, Non-Hispanic Asian, and Other Races, including Multi-Racial. To understand the distribution of the covariates between the levels of the Sex variable, we constructed Table 4.1.

Table 4.1: Summaries of the predictor variables Age, BMI (Body Mass Index), HEI (Healthy Eating Index) and Ethnicity used in the regression analysis, separated by the Sex. In the first column we distinguish the levels of the categorical variable ethnicity from the numerical covariates Age, BMI, HEI which are designated in the second column. In the third and fourth columns, the first four rows present the means and, in brackets, the standard deviations of the continuous variables (Age, BMI, HEI) for Men and Women respectively. The percentage breakdown of the sub-populations of Men and Women into their respective ethnicities is also provided. The description of the covariates are found in Section 4.1.1.

	Covariates	Men	Women
Numeric Variables	Age	47.45 (16.45)	48.082 (16.50)
	Body Mass Index	28.72 (5.73)	29.18 (7.41)
	Healthy Eating Index	53.013 (14.13)	56.63 (14.75)
Ethnicities	Mexican American	8.55 %	6.42 %
	Other Hispanic	5.42 %	5.28 %
	Non-Hispanic White	70.65 %	72.3 %
	Non-Hispanic Black	8.82 %	10.29 %
	Non-Hispanic Asian	3.77 %	3.37 %
	Other Races Including Multi-racial	2.79 %	2.35 %

This paper aims to create a parsimonious and straightforward regression model to interpret the several central aspects of energetic expenditure captured by the Age and BMI variables that are expected to behave in a non-linear way with the response. At the same time, we are interested in assessing the diet's effect on physical exercise. We have observed that sex and ethnicity differences in the U.S. population tend to interact concerning physical activity; e.g., women tend to be physically less active than men for the White, Black, Asian, Other races including multi-racial ethnicities, however, for his-

panic population, men and women show similar levels of physical activity. Hence, we considered an interaction between sex and ethnicity to obtain reliable population-based conclusions about the relationships between these predictors and physical activity. The sampling design of NHANES provides essential advantages to obtaining reliable population measurements that we cannot guarantee due to selection bias with observational cohorts such as the UK-Biobank. In order to properly exploit this advantage, however, we must incorporate survey design in the estimation procedure, as described in Section 4.3 below.

4.2 The partially linear Fréchet single-index Regression model

Let Y_i be the quantile function of daily activity levels corresponding to the i-th participant. In what follows, we will build the regression by directly modeling the pointwise mean function of $Y_i(t)$, $t \in [0, 1]$ on the covariates. Using the quantile function to characterize the physical activity distribution can be explained as follows. First, a density representation that ignores inactivity time is inappropriate because the distributions represented by the Y_i are a mixture of a mass at 0 and a continuous distribution for positive values. Moreover, the quantile function is practically less restrictive than, for example, the cumulative distribution function, which must take values between 0 and 1. Finally, and perhaps most importantly, the quantile function is known to be intimately connected to the well-established Wasserstein geometry on the space of distributions [82, 83, 84]. As a consequence, under L^2-Wasserstein metric 1.3 [1], the Fréchet mean [1] measure of

[1]We don't need them to be absolutely continuous. It is true that, if both are discrete, there may be uniqueness issues for the optimal transport map, but the optimal cost is unique and the quantile distance is still equivalent.

a random measure is characterized by the pointwise mean of the corresponding random quantile process. Hence, by proposing a regression model for the random quantile function Y_i, we are implicitly constructing a model for the conditional (Wasserstein-)Fréchet mean of the underlying random physical activity distribution measure [7].

Hence, we formally define the partially linear Fréchet single index model. For convenience of description, the model shall be denoted as PL-FSI for the remainder of the paper. Let $\boldsymbol{X_i} \in \mathbb{R}^p$ denote the p-dimensional covariate vector that will appear in the single index part of the model, while $\boldsymbol{Z_i} \in \mathbb{R}^q$ is the covariate vector considered for the linear part. The PL-FSI model is

$$E(Y_i(t)|\boldsymbol{X}_i, \boldsymbol{Z}_i) = \alpha(t) + \boldsymbol{\beta}(t)^T \boldsymbol{Z}_i + g(\boldsymbol{\theta}_0^T \boldsymbol{X}_i, t), \quad t \in [0,1], \tag{4.1}$$

where the vector $\boldsymbol{\theta_0} \in \mathbb{R}^p$, intercept function α, coefficient function $\boldsymbol{\beta}$ and link function g are the unknown parameters.

4.3 Model Estimation

For estimating the parameter $\boldsymbol{\theta}_0$, consider the parameter space Θ_p defined in 3.1. To facilitate estimation of the smooth bivariate function g, we will use the expansion

$$g(u,t) \approx \sum_{k=1}^{K+\varrho} \gamma_k(t)\phi_k(u), \tag{4.2}$$

where $\{\phi_k\}_{k=1}^{K+\varrho}$ are the basis functions of the B-spline of order ϱ on degrees of freedom K, and $\gamma_k(t)$ are the coefficients of the basis as a function of t. Hence, the knot sequence is of length $K + \varrho - 2$. For your clarification, a more detailed explanation of the B-spline is provided in the Appendix A. With this discussion, the approximation to (4.1) that

will motivate our estimator is

$$E(Y_i(t)|\boldsymbol{X}_i, \boldsymbol{Z}_i) \approx \alpha(t) + \boldsymbol{\beta}(t)^T \boldsymbol{Z}_i + \boldsymbol{\gamma}(t)^T \boldsymbol{U}_i(\boldsymbol{\theta}_0), \quad t \in [0,1], \tag{4.3}$$

where $\boldsymbol{\gamma}(t) = (\gamma_1(t), \ldots, \gamma_{K+\varrho}(t))^T$ and, for any $\boldsymbol{\theta} \in \Theta_p$, $\boldsymbol{U}_i(\boldsymbol{\theta}) = (\phi_1(\boldsymbol{\theta}^T \boldsymbol{X}_i), \ldots, \phi_{K+\varrho}(\boldsymbol{\theta}^T \boldsymbol{X}_i))^T$.

The linear form of (4.3) suggests a semi-parametric least-squares approach for estimation. However, one must remember that the individuals we analyze from the NHANES database do not represent a simple random sample of the US population. Instead, they are the result of a structured sample of a complex survey design from a finite population of individuals. Therefore, in order to perform inference correctly and obtain reliable results according to the specific sample design of the NHANES data set [43], it is necessary to adapt the usual estimation approach.

Assume that a sample $\mathcal{D} = \{(Y_i, \boldsymbol{X}_i, \boldsymbol{Z}_i) : i \in S\}$ is available, where Y_i is a response variable, and $\boldsymbol{X}_i, \boldsymbol{Z}_i$ are vectors of covariates taking values in a finite-dimensional space. The index set S represents a sample of n units from a finite population. To account for this sampling, each individual $i \in S$ will be associated with a positive weight w_i. In our analyses, these weights were taken to be the inverse of the probability $\pi_i > 0$ of being selected into the sample, i.e. $w_i = 1/\pi_i$ [85, 86]. These weights are used to construct an estimator of Horvitz-Thompson type [87, 88], by constructing a weighted least squares criterion.

The full procedure can be broken down into two steps. In the first step, for any unit-norm vector $\boldsymbol{\theta} \in \Theta_p$ and any $t \in [0,1]$, we can readily compute

$$\left(\hat{\alpha}_{\boldsymbol{\theta}}(t), \hat{\boldsymbol{\beta}}_{\boldsymbol{\theta}}(t), \hat{\boldsymbol{\gamma}}_{\boldsymbol{\theta}}(t)\right) = \underset{a \in \mathbb{R}, \boldsymbol{b} \in \mathbb{R}^q, \boldsymbol{c} \in \mathbb{R}^{K+s}}{\operatorname{argmin}} \sum_{i=1}^{n} w_i \left[Y_i(t) - a - \boldsymbol{b}^T z_i - \boldsymbol{c}^T \boldsymbol{U}_i(\boldsymbol{\theta})\right]^2. \tag{4.4}$$

These estimates lead to initial fitted quantile functions

$$Y_i^*(\boldsymbol{\theta}, t) = \hat{\alpha}_{\boldsymbol{\theta}}(t) + \hat{\boldsymbol{\beta}}_{\boldsymbol{\theta}}^T(t)\boldsymbol{Z}_i + \hat{\boldsymbol{\gamma}}_{\boldsymbol{\theta}}^T(t)\boldsymbol{U}_i(\boldsymbol{\theta}), \quad t \in [0,1]. \tag{4.5}$$

However, as a function of t, it may happen that $Y_i^*(\boldsymbol{\theta}, t)$ is not monotonically increasing, and hence is not a valid quantile function. The typical solution for this is to project, in the $L^2[0,1]$ sense, this fitted value onto the nearest monotonic function [19, 7], yielding valid fitted quantile functions $\hat{Y}_i(\boldsymbol{\theta}, t)$. Once these initial quantities are formed for any $\boldsymbol{\theta}$ and t, one can proceed to the estimation of $\boldsymbol{\theta}_0$ as justified in [21], one can use a generalized version of the residual sums of squares to obtain the estimate. In the current context, we propose the survey-weighted criterion

$$W_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} w_i \int_0^1 \left\{ Y_i(t) - \hat{Y}_i(\boldsymbol{\theta}, t) \right\}^2 dt \tag{4.6}$$

that constitutes a weighted average of the squared L^2 norms of the quantile residuals (or, equivalently, of the squared Wasserstein distances between observed and fitted physical activity distributions). Then the estimated parameter is

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \Theta_p}{\operatorname{argmin}} W_n(\boldsymbol{\theta}). \tag{4.7}$$

From this estimate of the index parameter, given any covariate pair $(\boldsymbol{z}, \boldsymbol{x})$, we can estimate the conditional Wasserstein-Fréchet mean quantile function as follows. First, the basis functions are evaluated at the relevant input by computing $\hat{\boldsymbol{u}} = (\phi_1(\hat{\boldsymbol{\theta}}^T\boldsymbol{x}), \ldots, \phi_{K+\varrho}(\hat{\boldsymbol{\theta}}^T\boldsymbol{x}))^T$. Then, as in (4.5), we construct the preliminary estimate

$$Y^*(t; \boldsymbol{z}, \boldsymbol{x}) = \hat{\alpha}_{\hat{\boldsymbol{\theta}}}(t) + \hat{\boldsymbol{\beta}}_{\hat{\boldsymbol{\theta}}}^T(t)\boldsymbol{z} + \hat{\boldsymbol{\gamma}}_{\hat{\boldsymbol{\theta}}}^T(t)\hat{\boldsymbol{u}}. \tag{4.8}$$

Finally, the estimated quantile function $\hat{Y}(t; \boldsymbol{z}, \boldsymbol{x})$ is obtained by projecting, in the L^2 sense, $Y^*(t; \boldsymbol{z}, \boldsymbol{x})$ onto the space of quantile functions, meaning the nearest monotonically increasing function. In particular, for any set of observed covariates $(\boldsymbol{Z_i}, \boldsymbol{X_i})$, we obtain fitted values $\hat{Y}_i(t) = \hat{Y}(t; \boldsymbol{Z_i}, \boldsymbol{X_i})$.

4.4 Computational Details

We now provide details regarding our implementation of our estimator for the NHANES data base. In the models implemented below in chapter 4.5, the non-linear covariate $\boldsymbol{X_i}$ for the i-th individual consists of their BMI and age, so the dimension for this component of the model is $q = 2$, i.e., $\boldsymbol{\theta}_0 \in \Theta_2$. For the spline basis in (4.2), computations were internally performed using the `dbs` function in the package `splines2` [89]. Knot placement was determined internally by the default option of the `dbs` function, and varied with the value of $\boldsymbol{\theta}$. Specifically, for any K, equally spaced values r_k, $k = 1, \ldots, K$, were computed, where $r_0 = 0 < r_1 < \cdots < r_K < r_{K+1} = 1$; the k-th interior knot was then taken as the r_k-th empirical quantile of the values $\{\boldsymbol{\theta}^T \boldsymbol{X}_i;\, i = 1, \ldots, n\}$. In our experiments, we set $\varrho = 4$ and $K = 5$, so the number of spline regression parameters is $K + \varrho = 9$. The covariates in the linear component $\boldsymbol{Z}_i$ consist of HEI (continuous) and indicator variables for sex and ethnicity, as well as the interaction between these. The total number of covariates in this component is thus $p = 12$. For the the i-th participant, let Z_{i1} be the continuous variable denoting the HEI, let $\text{Sex}_i, \text{Ethnicity}_i$ be the categorical

variables for Sex and Ethnicity of the i-th participant respectively, that is,

$$\text{Sex}_i = \begin{cases} 1, & \text{Male,} \\ 2, & \text{Female,} \end{cases} \qquad \text{Ethnicity}_i = \begin{cases} 1, & \text{Mexican American,} \\ 2, & \text{Other Hispanic,} \\ 3, & \text{Non-Hispanic White,} \\ 4, & \text{Non-Hispanic Black,} \\ 5, & \text{Non-Hispanic Asian,} \\ 6, & \text{Other Races including Multi-Racial.} \end{cases}$$

Then define the variables $Z_{i2} = \mathbf{1}_{\{\text{Sex}_i=\text{Female}\}}$, $Z_{ir} = \mathbf{1}_{\{\text{Ethnicity}_i=r-1\}}$ for $r = 3, 4, 5, 6, 7$; corresponding to the ethnicities Mexican American, Other Hispanic, Non-Hispanic White, Non-Hispanic Black, Non-Hispanic Asian, Other Races including Multi-Racial respectively. Then, define $\boldsymbol{Z}_i = (Z_{i1}, Z_{i2}, Z_{i3}, \ldots, Z_{i7}, Z_{i2} * Z_{i3}, \ldots, Z_{i2} * Z_{i7})$ and finally the covariate set $(\boldsymbol{X}_i, \boldsymbol{Z}_i)$.

(4.3) can be rewritten as follows:

$$E(Y_i(t)|\boldsymbol{X}_i, \boldsymbol{Z}_i) \approx \alpha(t) + \beta_1(t)Z_{i1} + \beta_2(t)Z_{i2} + \sum_{r=2}^{6} \beta_{3r}(t)Z_{i,(r+1)} + \sum_{r=2}^{6} \beta_{4r}(t)Z_{i,(r+1)} * Z_{i2} + \boldsymbol{\gamma}(t)^T \boldsymbol{U}_i(\boldsymbol{\theta}_0), \quad t \in [0,1]. \quad (4.9)$$

The estimates of parameters in (4.4) can be efficiently computed as a weighted least squares problem for any fixed $\boldsymbol{\theta}$ and $t \in [0,1]$, but in practice this can only be done for a finite ordered grid of values $t \in T_m = \{t_1, \ldots, t_m\} \subset [0,1]$. Let $\boldsymbol{D}$ be the design matrix of order $n \times 22$ and the parameter $\boldsymbol{\varphi}(t) = (\alpha(t), \beta_1(t), \beta_2(t), \beta_{32}(t), \ldots, \beta_{36}(t), \beta_{42}(t), \ldots, \beta_{46}(t),$

$\gamma_1(t), \ldots, \gamma_9(t))^T$. Hence, the equation (4.4) can be reframed as follows:

$$\hat{\boldsymbol{\varphi}}(t) = \underset{\boldsymbol{\varphi}(t)\in\mathbb{R}^{22}}{\operatorname{argmin}} \sum_{i=1}^{n} w_i \left(Y_i(t) - \boldsymbol{D}_i^T \boldsymbol{\varphi}(t)\right)^2 \tag{4.10}$$

where $\boldsymbol{D}_i^T$ is the i-th row of the design matrix $\boldsymbol{D}$. These initial survey-weighted least squares computations were done using R package `survey` [86, 43, 90], that allows us to introduce splines into the regression model while simultaneously incorporating the weights w_i that are necessitated by the complex sampling designs of NHANES.

For any given $\boldsymbol{\theta}$ and grid point t, computation of (4.5) is straightforward. To execute the projection step, observe that monotonicity can only be achieved in the discrete sense in dependence on the chosen grid T_m. We refer to [7] for a simple description of this projection algorithm, which can be done using any basic quadratic program solver. Consequently, for a given $\boldsymbol{\theta}$, (4.6) is approximated by numerical integration. Finally, to perform the optimization in (4.7), we use the function `optim` in R with the L-BFGS-B algorithm by repeatedly performing the above steps to evaluate $W_n(\boldsymbol{\theta})$ for different values of $\boldsymbol{\theta}$ across iterations. To deal with the possibility of local minima, four different starting values (taken to be equally spaced in their angular representation) in Θ_2 were used for this optimization step, yielding (potentially) four local minimizers. The final estimator was taken as the one among these yielding the smallest value of W_n. The algorithm 1 is provided in the Appendix B to demonstrate the flow of the computation.

Due to the characteristic of the index parameter $\boldsymbol{\theta}$ (i.e. $||\boldsymbol{\theta}||_E = 1$ and first non-zero element being positive), for convenience of estimation, we considered the transformation $\boldsymbol{\theta} \to (1, \eta)$, for $\eta \in [\pi/2, \pi/2]$ to leverage the advantages of the polar coordinate.

The following are some examples of intercepts in different cases are as follows for $t \in [0, 1]$:

- $\alpha(t)$ for Male, ethnicity: Mexican American.

- $\alpha(t) + \beta_2(t)$ for Female, ethnicity: Mexican American.
- $\alpha(t) + \beta_{32}(t)$ for Male, ethnicity: Other Hispanic.
- $\alpha(t) + \beta_2(t) + \beta_{32}(t) + \beta_{42}(t)$ for Female, ethnicity: Other Hispanic.
- $\alpha(t) + \beta_{33}(t)$ for Male, ethnicity: Non-Hispanic White.
- $\alpha(t) + \beta_2(t) + \beta_{33}(t) + \beta_{43}(t)$ for Female, ethnicity: Non-Hispanic White.
- $\alpha(t) + \beta_{34}(t)$ for Male, ethnicity: Non-Hispanic Black.
- $\alpha(t) + \beta_2(t) + \beta_{34}(t) + \beta_{44}(t)$ for Female, ethnicity: Non-Hispanic Black.
- $\alpha(t) + \beta_{35}(t)$ for Male, ethnicity: Non-Hispanic Asian.
- $\alpha(t) + \beta_2(t) + \beta_{35}(t) + \beta_{45}(t)$ for Female, ethnicity: Non-Hispanic Asian.
- $\alpha(t) + \beta_{36}(t)$ for Male, ethnicity: Other races, including Multi-racial.
- $\alpha(t) + \beta_2(t) + \beta_{36}(t) + \beta_{46}(t)$ for Female, ethnicity: Other races, including Multi-racial.

However, it was interesting to study the differences in the intercepts among these combinations of sex and ethnicities considered above. Hence, we computed the pairwise differences of these intercepts and their pointwise confidence intervals as function of t.

Let us introduce key results to derive the asymptotic point-wise confidence intervals for the previous linear combinations of the variables that constitute the the different intercepts. Consider the intercept part for the Female, ethnicity: Other Races, including Multi-Racial: $\alpha(t)+\beta_2(t)+\beta_{36}(t)+\beta_{46}(t)$. We obtain it by the linear combination $\boldsymbol{c}^T\boldsymbol{\varphi}(t)$ where $\boldsymbol{c} = (1,0,1,0,0,0,0,1,0,0,0,0,1,0,0,0,0,0,0,0,0,0)$. Hence,

$$\boldsymbol{c}^T\hat{\boldsymbol{\varphi}}(t) \sim \mathrm{N}(\boldsymbol{c}^T\boldsymbol{\varphi}(t), \boldsymbol{c}^T\Phi(t)\boldsymbol{c}) \tag{4.11}$$

where $\Phi(t)$ is the variance-covariance matrix of the estimate $\hat{\boldsymbol{\varphi}}(t)$. It was computed using the internal functions of the `svyglm` package.

To understand the interaction of sex and ethnicity in physical activity levels, we computed the model estimates of the intercept for the different cases and have taken their respective differences in 4.5.2. The findings we got were crucial to the Figures 4.3, 4.4, 4.5.

4.5 Experimental Results on the PL-FSI model

In order to examine the advantages of the newly proposed PL-FSI model, we compare it with the global Fréchet (GF) model of [19] which we slightly modify by introducing the specific survey weights in the estimation criterion. The covariates used were the same in each model. In fact, the global Fréchet model can be considered as a special case of the PL-FSI model in which all covariates are included in the linear component. To begin, we evaluate the capacity of the models to explain differences in physical activity distributions across individuals using the survey-weighted Fréchet R^2, denoted as $R^2_{\oplus}$ in contrast to (3.15), given by,

$$R^2_{\oplus} = 1 - \frac{\sum_{i=1}^{n} w_i \int_0^1 (Y_i(t) - \hat{Y}_i(t))^2 \mathrm{d}t}{\sum_{i=1}^{n} w_i \int_0^1 (Y_i(t) - \overline{Y}(t))^2 \mathrm{d}t} \tag{4.12}$$

where w_i is the survey weight corresponding to i-th observation and

$$\overline{Y}(t) = \left(\sum_{i=1}^{n} w_i \right)^{-1} \sum_{i=1}^{n} w_i Y_i(t)$$

is the weighted sample Wasserstein-Fréchet mean of the observed physical activity distributions. To compare models with different numbers of predictors, we define the adjusted

Fréchet R^2 as

$$\bar{R}^2_{\oplus} = R^2_{\oplus} - \left(1 - R^2_{\oplus}\right) \frac{q}{n - q - 1} \tag{4.13}$$

where n is the number of observations and q is the number of unknown parameters in the model [19].

4.5.1 Regression results

We first fitted a global Fréchet model with linear terms for all covariates, and then compared its performance with that of the PL-FSI regression model that included non-linear terms for BMI and Age. To facilitate comparison, we centered and standardized all numerical covariates before analysis.

We evaluated the goodness-of-fit of each model using $\overline{R}^2_{\oplus}$ (4.13). The PL-FSI model had a higher value of 0.146, which is 24% higher relative to the 0.118 value obtained by the global Fréchet model. This suggests that although the predictive capacity of both models is moderate, the additional parameters introduced by the single index and spline representation improved the variance explained.

The PL-FSI model was used to estimate the index parameter, resulting in $\hat{\boldsymbol{\theta}} = (0.2661, 0.9639)$ for the variables BMI and Age respectively. This indicates that Age had a greater influence than BMI on the physical activity levels of each individual. To examine the interaction of sex and ethnicity in physical activity levels, we computed the model estimates of the intercept for different cases.

4.5.2 Interpretation of the categorical covariates

To tackle significant epidemiological and public health inquiries, we leveraged the linear as well as the semi-parametric nature of our model to examine variations in physical

activity patterns across diverse populations in the United States. Our focus was on investigating how physical activity varies across various population groups, encompassing all ranges of physical activity intensities measured by the physical activity quantile function. Hence the inclusion of the covariates sex, ethnicity and their interactions into the linear part of the PL-FSI model. Specifically, we aimed to address the following queries:

1. Are there differences in levels of physical activity between men and women, and how do these differences vary among different ethnicities?

2. Do women of different ethnicities exhibit variations in physical activity levels, and if so, how do these differences differ within ethnicities?

3. Do men of different ethnicities demonstrate discrepancies in physical activity levels, and if so, how do these differences vary within ethnicities?

To answer these questions, we estimated the regression parameters $\hat{\alpha}_{\hat{\boldsymbol{\theta}}}(t)$ and $\hat{\boldsymbol{\beta}}_{\hat{\boldsymbol{\theta}}}(t)$ (4.8) and their 95% confidence interval for each $t \in [0, 0.98]$. This restriction in t was necessary due to boundary effects of B-spline basis functions in estimating the quantiles in the far right tail.

If the order of splines of the non-linear component is fixed and assuming no bias, the asymptotic distribution of the estimator $\hat{\boldsymbol{\beta}}_{\boldsymbol{\theta}}(t)$ has been shown to be Gaussian in the settings similar to [91, 7]. We can then use the standard outputs of the `survey` package, which adjusts for the survey weights, to construct pointwise confidence bands for the functional coefficients in (4.8). corresponding to the covariates in the linear part. Notice that the standard re-sampling strategies like a naive bootstrap do not work here with the two-step-sampling design of the NHANES because the observational units are not exchangeable. We emphasize that these confidence intervals merely guide our qualitative assessment of uncertainty. However, their asymptotic precision has not been theoretically

guaranteed due to the different sources of variability, which include the effects of spline parameters and the estimation of $\boldsymbol{\theta}_0$, for which this procedure does not provide concrete inference.

To answer the question (1) above, Figure 4.3 shows the estimated model intercepts for male participants subtracted from the estimated intercepts for female participants within each ethnic group. The computations were performed using the various functionalities of the `survey` package [86, 43, 90]. The difference as a function of t for each ethnicity were plotted along with their 95% pointwise confidence intervals also as functions of t. For the lower quantiles of physical activity, the plots are not informative due to the responses having positive mass at 0. For sufficiently high $t \in [0.30, 0.98]$ pointwise results indicate that men are more physically active than women across the ethnicities White, Black, Asian, Other Races including Multi-Racial such that the poitwise confidence intervals exclude 0 for in moderate to high physical activity ranges. However, in the Mexican American and Other Hispanic categories, men and women show similar levels of physical activity.

In response to the question (2) above, in figure (4.4), the estimated model intercepts for females were computed as a function of t for each ethnicity and their pairwise differences were computed along with their 95% pointwise confidence intervals. Due to similar reasoning in figure (4.3), the plot is informative for only $t \in [0.30, 0.98]$, where Mexican American and Other Hispanic women show higher levels of physical activity compared to other races, for each t on the chosen grid. Women of White, Black, Asian, and Other Races, including Multi-Racial individuals, exhibit similar levels of physical activity. Mexican American and Other Hispanic women show similar levels of physical activity but higher than the rest of the ethnicities.

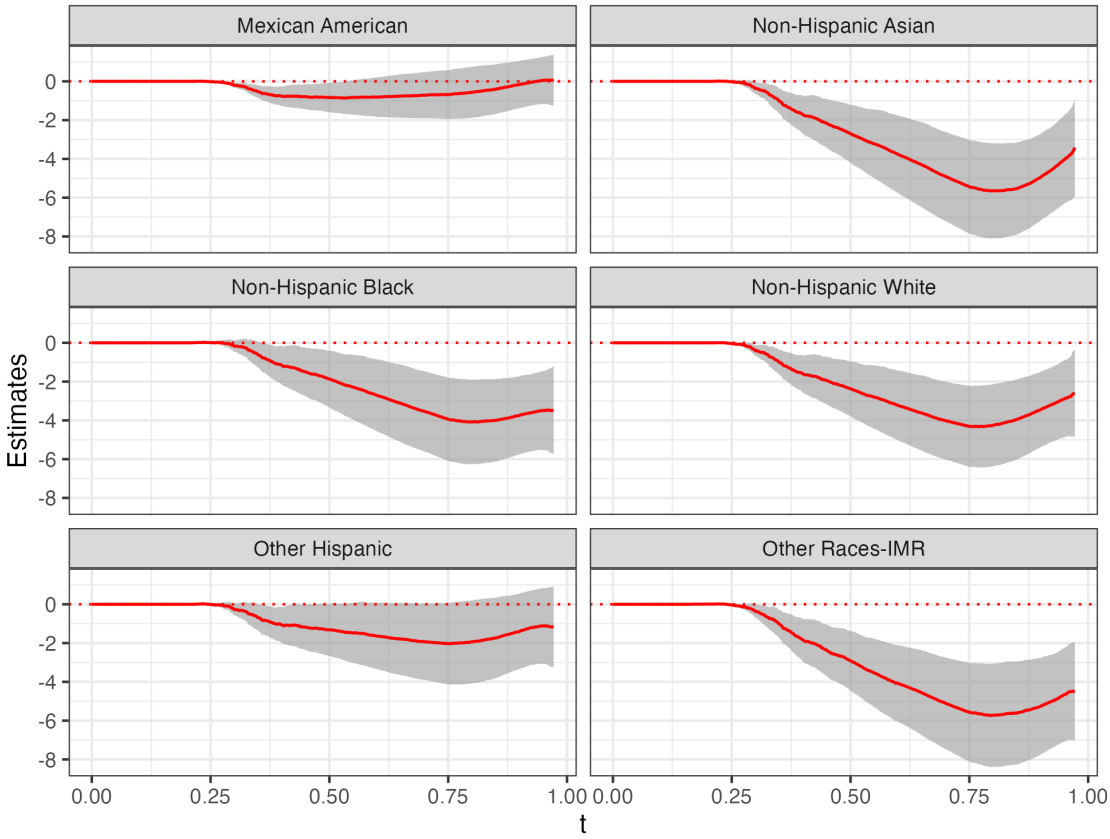

Figure 4.3: The intercepts for the PL-FSI regression model (4.3), are considered for male and female participants of different ethnic backgrounds. The intercepts for the males are subtracted from the intercepts of the females for each ethnicity, considering the numeric variables HEI, Age and BMI are fixed. The respective parameter combinations are computed along with their 95% Confidence Intervals and plotted as solid red lines and grey shade respectively. The dotted red line at 0 is for reference. The differences are considered for the ethnicities: Mexican American, Other Hispanic, Non-Hispanic White, Non-Hispanic Black, Non-Hispanic Asian, Other races including Multi-Racial.

In response to the question (3) above, in figure (4.5) shows the corresponding results of Figure 4.4 but for the male participants. The results indicate that for moderate to high physical activity levels, i.e. $t \in [0.30, 0.98]$ male participants of Mexican American and Other Hispanic ethnicities are more physically active than the males of of Black, White, Asian, and Other Races, including Multi-Racial ethnicities, since the pointwise confidence intervals exclude 0 in many activity levels. Mexican American and Other Hispanic men

exhibit similar physical activity levels and they remain more active individuals than the rest of the male participants of other ethnicities. There is some evidence that show men of White and Black ethnicities show higher levels of physical activity compared to Asian men in higher ranges of t. Men of Other Races, including Multi-Racial individuals, exhibit slightly lower levels of physical activity compared to men of Black and White ethnicities, but are similar to Asian men.

From a public health perspective, this suggests that specific interventions for promoting physical activity must be tailored differently based on accelerometer intensities of various sexes and ethnicities. Hence, different strategies may be required to address the diverse physical activity patterns exhibited by individuals, distinguished by sex and ethnicities.

In Figure 4.6, we conducted an examination of the influence of the HEI variable, which serves as a measure of diet quality, on physical activity. Our analysis revealed notable distinctions primarily in large quantile probabilities that are associated with high-intensity exercise areas. For instance, individuals engaging in high-intensity cardio and/or resistance training were found to more frequently adhere to a strictly healthy diet. These findings highlight the connection between diet quality and physical activity, especially in contexts where intense exercise regimens are prevalent.

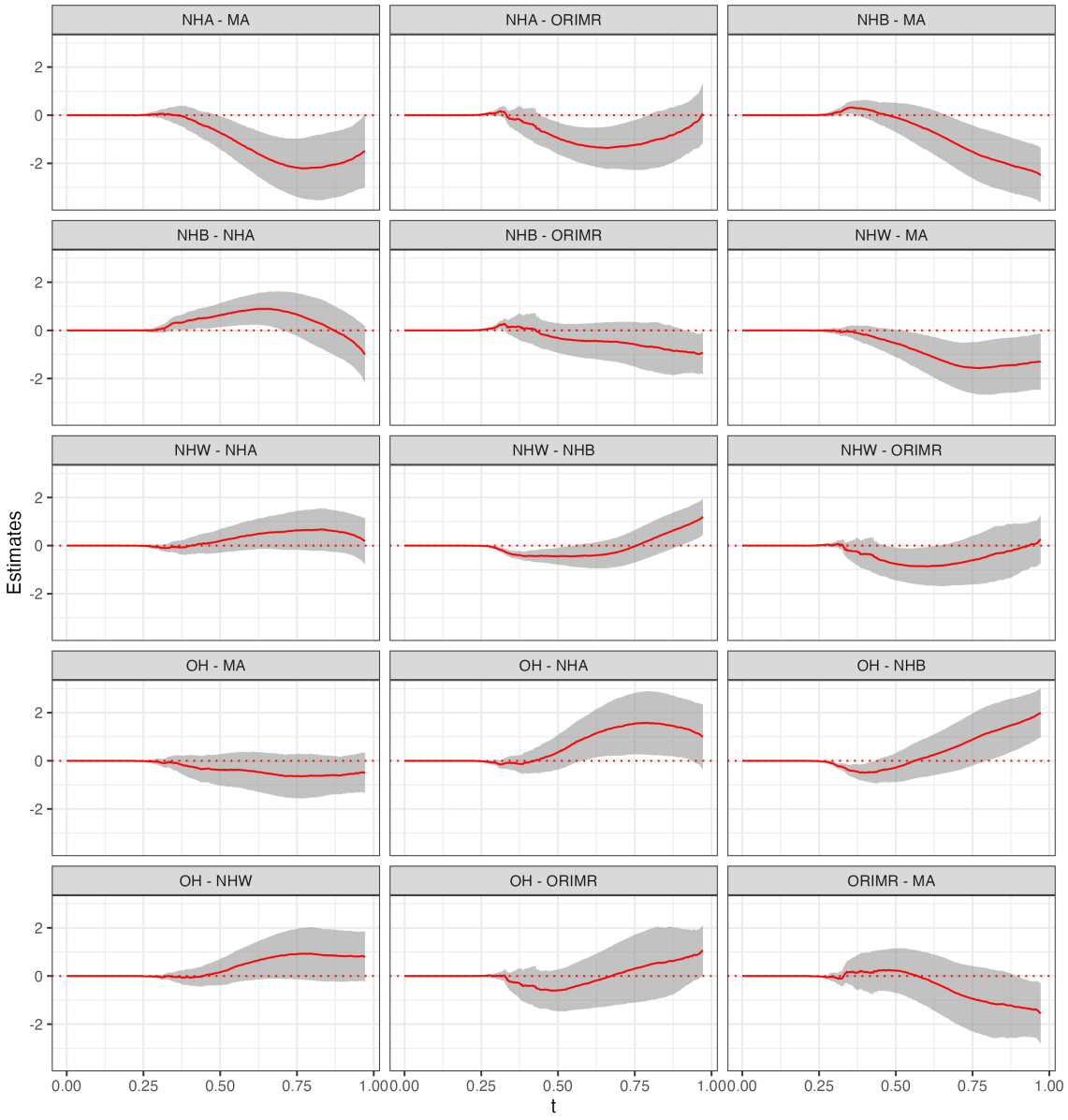

Figure 4.4: The intercepts for the PL-FSI regression model (4.3) are considered for females of different ethnic backgrounds. The pairwise differences of such intercepts are computed along with their 95% Confidence Intervals and plotted here as solid red lines and grey shade respectively, considering the numeric variables HEI, Age and BMI are fixed. The dotted red line at 0 is for reference. The title for each panel indicates the order of the differences of the intercepts. The abbreviations for the ethnicities are, OH: Other Hispanic, MA: Mexican American, NHW: Non-Hispanic White, NHB: Non-Hispanic Black, NHA: Non-Hispanic Asian, and ORIMR: Other Races Including Multi-Racial. The the title, e.g. 'OH - MA' indicates that the estimated intercepts for females of Mexican American was subtracted from the estimated intercepts for females of Other Hispanic ethnicity.

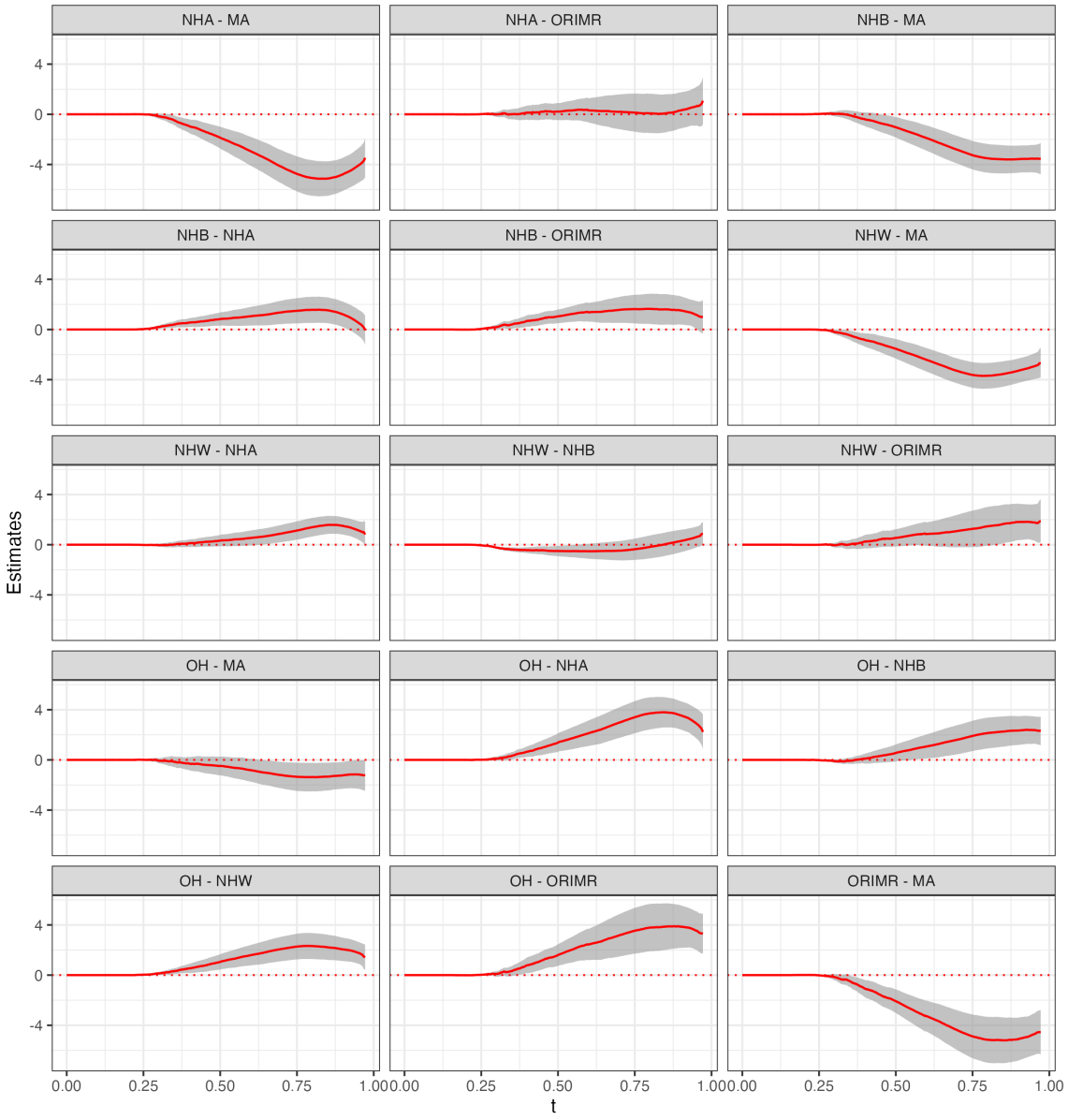

Figure 4.5: The intercepts for the PL-FSI regression model (4.3) are considered for males of different ethnic backgrounds. The pairwise differences of such intercepts are computed along with their 95% Confidence Intervals and plotted here as solid red lines and grey shade respectively, considering the numeric variables HEI, Age and BMI are fixed. The dotted red line at 0 is for reference. The abbreviations for ethnicities as well as the order of the differences are same as in the Figure 4.4.

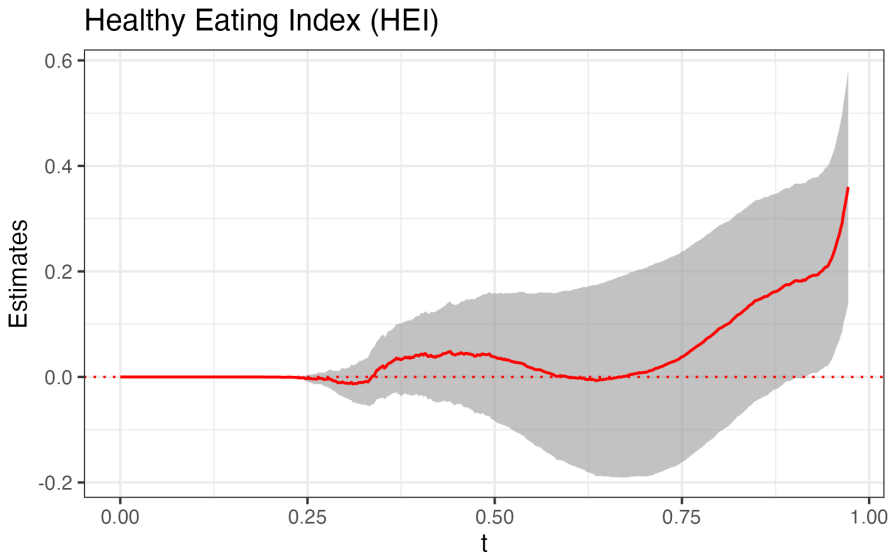

Figure 4.6: The estimated functional coefficients for the covariate HEI in the model (4.3) is computed and plotted here as the solid red line and its 95% pointwise confidence band is given by the grey shade. The dotted red line at 0 is for reference.

4.5.3 Association with the non-linear covariates

In this section, we focus on exploring the modeling advantages of the new semi-parametric models to gain more detailed insights into non-linear relationships, while retaining the information from the linear components of the models. We examined the behavior of fitted quantiles for unstandardized values of the covariates (Age and BMI) within the respective ranges of $[20, 80]$ and $[18.5, 40]$. To achieve this, we used the formula in (4.8), where the covariates were utilized after being scaled and centered for any $t \in [0, 0.98]$. This analysis allows us to uncover the dynamics of the relationship of Age and BMI with the fitted quantiles and provides valuable insights into the non-linear aspects of the model fits. As discussed earlier, for given values of the covariates $\boldsymbol{x}$, $\boldsymbol{z}$, $Y^*(t, \boldsymbol{z}, \boldsymbol{x})$ may not be a valid quantile function in t. Hence, we project to obtain the nearest quantile function $\hat{Y}(t, \boldsymbol{z}, \boldsymbol{x})$. But the latter is not amenable to further computation exercises. Therefore, for the next few exercises we will consider the response Y^*, prior to

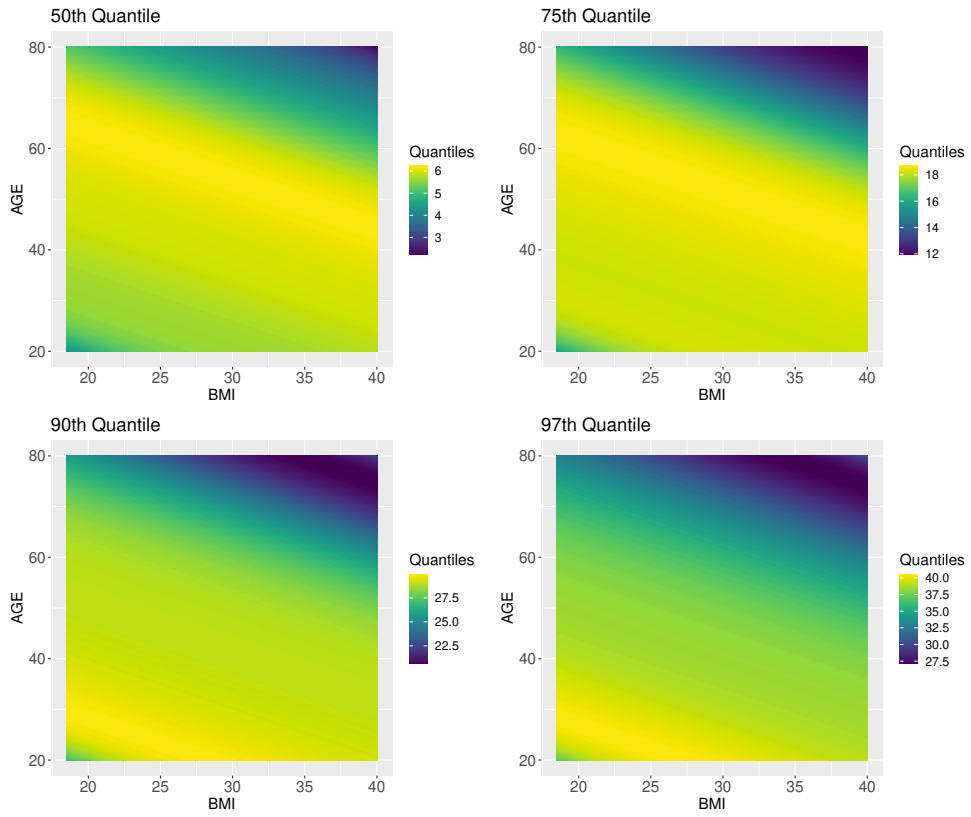

Figure 4.7: Heatmap plot of $\hat{Y}(\hat{\boldsymbol{\theta}}, t)$ across different quantiles, $t = 0.50$ (top left), 0.75 (top right), 0.90 (second row left), 0.97 (second row right) respectively. A 2-dimensional grid was considered for the covariates BMI (in range [18.5,40]) and age (in range [20,80]) for the single index component of the PL-FSI regression model. The categorical covariates in the linear component were fixed at their baseline levels (i.e. sex male, ethnicity Mexican American) while the numerical covariate HEI was fixed at median level.

the transformation.

To create the four panels of Figure 4.7, we considered equidistant grids of length 500 each, over the standardized ranges for BMI (for horizontal axis) and Age (for vertical axis). Let $\tilde{\boldsymbol{x}}$ be the 2-dimensional vectors on the grid. The fitted quantile values $Y^*(t, \boldsymbol{z}, \tilde{\boldsymbol{x}})$ across the aforementioned region of standardized values, with $\boldsymbol{z}$ being the covariates in the linear part fixed to represent the reference groups and HEI. The four panels are for $t = 0.50, 0.75, 0.90, 0.97$; lower values of t were omitted because the nonlinear component had a negligible effect. These panels indicate that the non-linear relationship

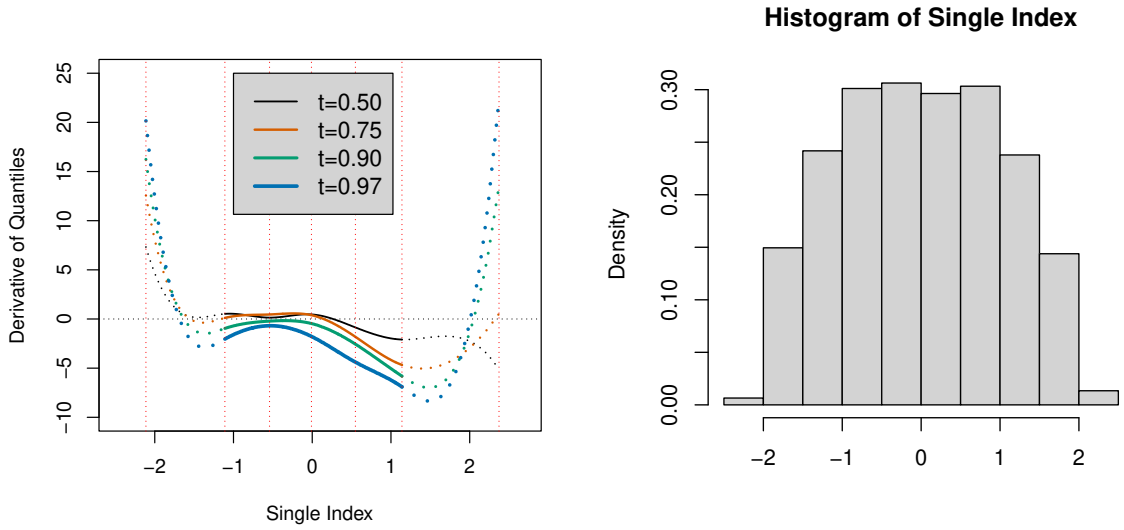

Figure 4.8: (Left) Plot of $(\partial\hat{g}/\partial u)(u,t)$, the derivative of the spline-based estimate of the nonlinear component defined in (4.14), for u across the empirical range of observed index values $\hat{\boldsymbol{\theta}}^T\boldsymbol{X}_i$, $i=1,\ldots,n$, and for quantile levels $t=0.50$ (black), 0.75 (red), 0.90(green),and 0.97 (blue). Knot locations are shown as vertical dotted red lines, and the derivative curves are depicted as solid (respectively, dotted) within (resp., outside of) the interior knot range. (Right) Histogram of observed index values $\hat{\boldsymbol{\theta}}^T\boldsymbol{X}_i$, $i=1,\ldots,n$.

between age and BMI is more pronounced for larger values of t.

At the lowest level of BMI (< 20), the people in age range $55-70$ perform the highest median to 75^{th} (i.e. $t = 0.50, 0.75$) quantile level of physical activity, but for the same lowest BMI range, people in the age range $25-35$ perform the highest physical activity in the quantiles $t = 0.90, 0.97$. In each of the panels (or, quantiles) the age range for highest physical activity linearly decreases with increase in BMI. For the highest BMI in our study (≈ 40), the highest physical activity in the quantiles $t = 0.50$, 0.75 are shown by the age range $40-55$. However, for the highest physical activity in the quantiles $t = 0.90$, 0.97 are shown by the BMI range $25-30$ in the age range $20-25$.

As an additional visualization of the nonlinear effect of these covariates in the model,

we directly analyze the derivative of the nonlinear fit, namely

$$\frac{\partial \hat{g}}{\partial u}(u,t) = \sum_{k=1}^{K+\varrho} \hat{\gamma}_{\hat{\boldsymbol{\theta}},k}(t) \left[\frac{d}{du}\phi_k(u)\right]. \tag{4.14}$$

To compute the derivative estimates, we utilized the R package `splines2`'s `dbs` function, which computes derivatives of B-spline functions. The left panel of Figure 4.8 displays the behavior of (4.14) across the relevant range of values $\hat{\boldsymbol{\theta}}^T \boldsymbol{X}_i$ (whose distribution is depicted via a histogram in the right panel of the figure) that were used to generate the estimates, for $t = 0.5, 0.75, 0.9, 0.97$. To facilitate reliable interpretation, we focus on the behavior of the curves within the interior knots, indicated by the solid portion of each curve in the figure. The various derivative curves suggest that, given the covariates in the linear term, there is little to no association between the BMI and age index and the physical activity quantile response until the single index value becomes positive, at which point the association becomes negative. Since both elements of $\hat{\boldsymbol{\theta}}$ are positive, these findings imply that the model reflects a negative association between physical activity quantiles and BMI/Age when at least one of these is large. Furthermore, the strength of this negative association increases as the quantile level t becomes larger, as evidenced by the increasingly negative derivative estimates in the left panel as t increases. For instance, the derivative for $t = 0.5$ (the median physical activity quantile) is only slightly negative for values of u near zero, whereas it steadily decreases as one examines the curves for $t = 0.75, 0.9, 0.97$.

4.6 Discussion

The core contribution of this chapter is to propose a new PL-FSI regression model to analyze responses of distributional functional nature. This new methods have been

implemented to analyze the physical activity data from the NHANES database 2011-2014, for participants in Age group $20 - 80$ and BMI range $18.5 - 40$. We incorporated the NHANES survey weights within the new PL-FSI algorithm according to the sampling mechanism introduced by the Horvitz-Thompson type estimator [85] to construct a weighted least square criterion to estimate the model parameters.

Our new findings in through this literature are summarized below:

1. We examine the discrepancies in the physical activity levels between men and women of different ethnicities in the American population. We also attempted to understand the impact of the continuous variables HEI, BMI, and Age, in all ranges of human physical activity intensities thanks to the new quantile distributional representations of physical activity. For example, we show that diet is important only in the high-intensity levels of physical activity range; a better diet, according to the HEI score, is related to more exercise. We also show that the Mexican American and Other Hispanic groups are the most active individuals in the American population for both men and women. We discover a non-linear interaction between Age and BMI in the energetic expenditure, specifically in the higher quantiles of physical activity profiles.

2. We show the modeling advantages of the new PL-FSI algorithm over the classical global Fréchet regression model in terms of adjusted Fréchet R-squared as well as in terms of interpretability with the new tools introduced, e.g., the gradient of a conditional mean function.

From a methodological point of view, we propose the first PL-FSI regression model in the context of Object data analysis to bridge the gap between the global Fréchet regression [19] and the Fréchet single index model [21], while preserving the interpretability of the predictors and parameter estimates. To the best of our knowledge, this is also the

first regression model to incorporate survey data in the context of Object data analysis and the first work that computes the gradient of the Fréchet regression function in order to interpret each predictor in the model.

The most popular approach to analyzing accelerometer data is through finite dimensional compositional metrics. Here we used their functional extension [28] to capture more information about physical activity from an individual by adopting the mathematical framework of the L^2-Wasserstein space endowed with the Wasserstein metric. To overcome the problem of the positive probability at zero physical activity level, we used compositional data. In addition, the range of values measured by the accelerometer varies widely among individuals and groups, which can present difficulties when trying to apply the standard distributional data analysis methods in our setting [28]. For example, functional compositional transformations can be an alternative strategy to creating a regression model about physical activity in a linear space [92, 93, 94]. However, the distributional physical activity representation arises from a mixed-stochastic process (see Figures 4.1, 4.2 for more details) that prevents the use of the linear functional data methods based on considering a basis of functions due to the discontinuity of the quantile function in the transition of the inactivity to activity in the physical exercise.

The analysis of complex statistical objects in biomedical science provides an excellent opportunity to create new clinical biomarkers that enrich the information more than the existing variables that monitor the health and evolution of diseases. For example, distributional representations are a significant advancement in digital medicine [95] as a digital biomarker [96, 29]. However, the generality of techniques introduced provides users the opportunity to use the methods developed here with other complex statistical objects such as connectivity graphs, shape, and directional objects that can introduce new clinical findings in a broad list of clinical situations, for example in the brain and phylogenetic tree analysis [97, 98, 99, 100, 5].

Furthermore, with the increasing analysis of large cohorts with richer designs, such as complex survey design, the methods provided here will gain more popularity among practitioners. The use of complex statistical objects will undoubtedly be a daily statistical practice in biomedical applications.

Appendix A

B-Spline and its application in the partially linear Fréchet regression model

This section discusses in somewhat selective detail the fundamental concepts of the Basis spline and how it was applied in the estimation and inference process of the PL-FSI model. More specifically, in 4.1, the single index portion $g(\boldsymbol{\theta}_0^T \boldsymbol{X}_i, t)$, $t \in [0,1]$ is estimated in 4.2 for an arbitrary $\boldsymbol{\theta}$ and an arbitrary covariate point $\boldsymbol{x}$, with the single index $u = \boldsymbol{\theta}^T \boldsymbol{x}$. The reason to choose B-splines for the estimation over other competitor non-parametric methods e.g. kernel regression or wavelet regression is simplicity; B-splines are easier and faster to compute, allowing the estimation of all the relevant parameters simultaneously. Many softwares and packages perform spline computation that provide user-friendly interface. The kernel regression has a multistage estimation process where a smoothing bandwidth has to be chosen first, the rest of the model estimated subsequently. The wavelets regression allow for simultaneous estimation of its parameters but are easy to pick up on local patterns in the data and it is usually good for noisy or volatile data.

The relevant software for wavelet computation are not well-developed either.

A.0.1 Definition of B-splines

In the rest of the discussion we try to lay out the definitions and theoretical treatment of the B-splines that is relevant to the PL-FSI model. The treatment closely follows the description of [101]. Let $\xi_0 < \xi_1 \leq \xi_2 \leq \ldots \leq \xi_{M_1} < \xi_{M_1+1}$ be a sequence of numbers such that ξ_0, ξ_{M_1+1} are called the boundary knots which typically define the domain over which we want to define the spline regression. And $\xi_1 \leq \xi_2 \leq \ldots \leq \xi_{M_1}$ are called the interior knots corresponding to the spline regression. Then, define the augmented knot sequence ϑ such that

- $\vartheta_1 \leq \vartheta_2 \leq \cdots \leq \vartheta_{M_2} \leq \xi_0$;
- $\vartheta_{j+M_2} = \xi_j$, for $j = 1, 2, ..., M_1$
- $\xi_{M_1+1} \leq \vartheta_{M_1+M_2+1} \leq \vartheta_{M_1+M_2+2} \leq \cdots \leq \vartheta_{M_1+2M_2}$

The actual values of these additional knots beyond the boundary knots are arbitrary and it is customary to make them all the same and equal to ξ_0 and ξ_{M_1+1} respectively.

Let $B_{i,\varrho}(x)$ the ith B-spline basis function of order ϱ for the knot-sequence ϑ, $\varrho \leq M_2$. They are defined recursively as follows:

$$B_{i,1}(x) = \begin{cases} 1 & \text{if } \vartheta_i \leq x < \vartheta_{i+1} \\ 0 & \text{otherwise} \end{cases} \tag{A.1}$$

for $i = 1, \ldots, M_1 + 2M_2 - 1$. These are also known as the Haar basis functions.

Hence, the B-spline basis functions of order ϱ are defined as;

$$B_{i,\varrho}(x) = \frac{x - \vartheta_i}{\vartheta_{i+\varrho-1} - \vartheta_i} B_{i,\varrho-1}(x) + \frac{\tau_{i+\varrho} - x}{\vartheta_{i+\varrho} - \vartheta_{i+1}} B_{i+1,\varrho-1}(x) \tag{A.2}$$

for $i = 1, 2, ..., M_1 + 2M_2 - \varrho$.

Thus with $M_2 = 5$, $B_{i,4}$, $i = 1, 2, ..., M_1 + 6$ are the $M_1 + 6$ cubic B-spline basis functions for the knot sequence $\{\xi_0, \xi_1, ..., \xi_{M_1}, \xi_{M_1+1}\}$. In fact notice that with these knots only the subset $B_{i,\varrho}$, $i = M_1 - \varrho + 1, \ldots, M_1 + M_2$ are required for the B-spline basis of order $\varrho < M_1$. To generate B-spline basis functions of any order this recursion can be used and they do span the space of cubic splines for the knot sequence.

To avoid division by zeros in (A.2) due to duplicated knots, we adopt the convension that $B_{i,1} = 0$ if $\vartheta_i = \vartheta_{i+1}$, then by induction $B_{i,\varrho} = 0$ if $\vartheta_i = \vartheta_{i+1} = \cdots = \vartheta_{i+\varrho}$.

As mentioned earlier, for the PL-FSI model (4.2), $\varrho = 4$ and $M_1 = 5$ and $M_2 = 4$ the B-splines bases $B_{i,4}$ for $i = 1, 2, ..., 9$ are respectively denoted by $\phi_1, \phi_2, ..., \phi_9$. Hence, for every order of quantile, t, the coefficients of the B-spline basis are estimated by the weighted generalized linear regression.

Appendix B

Algorithms of the partially linear Fréchet single index regression model.

Here we present the algorithm of the computation flow for the PL-FSI model between the equations (4.1) to (4.8), given in sections 4.2 and 4.3. It is meant to provide the reader more details of every step of computation of the locally estimated single index parameter $\hat{\boldsymbol{\theta}} \in \Theta_p$ from a starting value $\boldsymbol{\theta}_{st} \in \Theta_p$. To get a global estimate we consider a grid of starting values spanning Θ_p, each yielding an estimate $\hat{\boldsymbol{\theta}}(\boldsymbol{\theta}_{st})$. We choose the estimate which minimizes the M-estimation criterion $W_n(\hat{\boldsymbol{\theta}}(\boldsymbol{\theta}_{st}))$ in (4.6). After obtaining the globally optimized $\hat{\boldsymbol{\theta}}$, we get the fitted response $Y^*(t; \boldsymbol{Z}_i, \boldsymbol{X}_i)$ in (4.8) and project to the nearest quantile function $\hat{Y}_i = \hat{Y}_i(t, \boldsymbol{Z}_i, \boldsymbol{X}_i)$ into the L^2-Wasserstein space.

Algorithm 1: Estimation of $\hat{\boldsymbol{\theta}}(\boldsymbol{\theta}_{st})$ from a starting value $\boldsymbol{\theta}_{st} \in \Theta_p$.

Inputs:

- $\varrho \leftarrow 4$ (order of B-spline basis), $K \leftarrow 5$ (number of internal knots), n be the number of participants.
- $tol \leftarrow 10^{-12}$, $miter \leftarrow 1000$;
- The equidistant grid $\{0 = t_0, t_1, t_2, ..., t_{500} = 1\}$.
- Let $Y_i(t_j)$ be the t_j-th quantile of the physical activity representation of the i-th participant, $j = 1, 2, ..., 500$, $i = 1, 2, ..., n$.
- Let $\boldsymbol{X}_i \in \mathbb{R}^p$ and $\boldsymbol{Z}_i \in \mathbb{R}^q$ for $i = 1, 2, ..., n$ be the n observations in the non-linear and the linear part of the PL-FSI model.
- NHANES survey weights $\boldsymbol{w} = \{w_1, w_2, ..., w_n\}$ for the participants in PL-FSI model.

Computation:

1. Set $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}_{st}$.
2. Set $iter = 1$.
3. Consider $u_i = \boldsymbol{\theta}^T \boldsymbol{X}_i$, for $i = 1, 2, ..., n$.

   For the smooth and unknown function g, compute the expansion $g(u_i, t_j) \approx \sum_{k=1}^{K+\varrho} \gamma_k(t_j)\phi_k(u_i) = \boldsymbol{\gamma}(t_j)^T \boldsymbol{U}_i(\boldsymbol{\theta})$ in (4.2).
   Consider the parameters $\alpha(t_j)$, $\boldsymbol{\beta}(t_j)$ in the regression $E(Y_i(t_j)|\boldsymbol{X}_i, \boldsymbol{Z}_i) \approx \alpha(t_j) + \boldsymbol{\beta}(t_j)^T \boldsymbol{Z}_i + \boldsymbol{\gamma}(t_j)^T \boldsymbol{U}_i(\boldsymbol{\theta})$, $j = 1, 2, ..., 500$ in (4.3)
4. Estimate the parameters $\hat{\alpha}_{\boldsymbol{\theta}_{st}}(t_j), \hat{\boldsymbol{\beta}}_{\boldsymbol{\theta}}(t_j), \hat{\boldsymbol{\gamma}}_{\boldsymbol{\theta}}(t_j)$ in (4.4) and with that compute $Y_i^*(\boldsymbol{\theta}, t_j) = \hat{\alpha}_{\boldsymbol{\theta}}(t_j) + \hat{\boldsymbol{\beta}}_{\boldsymbol{\theta}}^T(t_j)\boldsymbol{Z}_i + \hat{\boldsymbol{\gamma}}_{\boldsymbol{\theta}}^T(t_j)\boldsymbol{U}_i(\boldsymbol{\theta})$ in (4.5) for $j = 1, 2, ..., 500$.
5. $\{Y_i^*(\boldsymbol{\theta}, t_j) : j = 1, 2, ..., 500\}$ has to be a valid quantile function for every $i = 1, 2, ..., n$. Otherwise, project $\{Y_i^*(\boldsymbol{\theta}, t_j) : j = 1, 2, ..., 500\} \rightarrow \{\hat{Y}_i(\boldsymbol{\theta}, t_j) : j = 1, 2, ..., 500\}$ in $L^2[0, 1]$ sense to the nearby monotonic function.
6. Compute the $W_n(\boldsymbol{\theta})$ in (4.6). Set $\mathcal{W}_{iter} \leftarrow \mathrm{W}_n(\boldsymbol{\theta})$.
7. Obtain $\hat{\boldsymbol{\theta}}_{iter}$ from (4.7) using the *L-BFGS-B* algorithm.

while $iter < miter$ ***and*** $\mathcal{W}_{iter} > tol$ **do**
  $\boldsymbol{\theta} \leftarrow \hat{\boldsymbol{\theta}}_{iter}$;
  $iter \leftarrow iter + 1$;
  repeat steps *3* - *7*.
end

Outputs: $\hat{\boldsymbol{\theta}}(\boldsymbol{\theta}_{st}) = \hat{\boldsymbol{\theta}}_{iter}$, $iter$, $\mathcal{W}_{iter}$.

Bibliography

[1] M. Fréchet, *Les éléments aléatoires de nature quelconque dans un espace distancié*, in *Annales de l'institut Henri Poincaré*, vol. 10(4), pp. 215–310, 1948.

[2] H.-G. Müller, *Peter Hall, functional data analysis and random objects*, *The Annals of Statistics* **44** (2016), no. 5 1867–1887.

[3] J. S. Marron and A. M. Alonso, *Overview of object oriented data analysis*, *Biometrical Journal* **56** (2014), no. 5 732–753.

[4] V. Patrangenaru and L. Ellingson, *Nonparametric Statistics on Manifolds and Their Applications to Object Data Analysis*. CRC Press, 2016.

[5] Y. Yuan, H. Zhu, W. Lin, and J. Marron, *Local polynomial regression for symmetric positive definite matrices*, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **74** (2012), no. 4 697–719.

[6] Y. Chen, Z. Lin, and H.-G. Müller, *Wasserstein regression*, *Journal of the American Statistical Association* **118** (2023), no. 542 869–882, [https://doi.org/10.1080/01621459.2021.1956937].

[7] A. Petersen, X. Liu, and A. A. Divani, *Wasserstein F-tests and confidence bands for the Fréchet regression of density response curves*, *The Annals of Statistics* **49** (2021), no. 1 590–611.

[8] P. Dubey and H.-G. Müller, *Fréchet analysis of variance for random objects*, *Biometrika* **106** (2019), no. 4 803–821, [https://academic.oup.com/biomet/article-pdf/106/4/803/30646779/asz052.pdf].

[9] N. I. Fisher, *Statistical Analysis of Circular Data*. Cambridge University Press, 1995.

[10] N. Fisher, T. Lewis, and B. J. Embleton, *Statistical Analysis of Spherical Data*. Cambridge University Press, Cambridge, 1987.

[11] T. Chang, *Spherical regression with errors in variables*, *The Annals of Statistics* **17** (1989), no. 1 293–306.

[12] B. Pelletier, *Non-parametric regression estimation on closed Riemannian manifolds*, *Journal of Nonparametric Statistics* **18** (2006), no. 1 57–67.

[13] X. Shi, M. Styner, J. Lieberman, J. G. Ibrahim, W. Lin, and H. Zhu, *Intrinsic regression models for manifold-valued data*, in *International Conference on Medical Image Computing and Computer-Assisted Intervention – MICCAI 2009, 12th International Conference, London, UK, September 20-24, 2009, Proceedings, Part II; Lecture Notes in Computer Science* (G.-Z. Yang, D. Hawkes, D. Rueckert, A. Noble, and C. Taylor, eds.), vol. 5762, pp. 192–199, Springer, 2009.

[14] M. Niethammer, Y. Huang, and F.-X. Vialard, *Geodesic regression for image time-series*, in *International Conference on Medical Image Computing and Computer-Assisted Intervention – MICCAI 2011, 14th International Conference, Toronto, Canada, September 18-22, 2011, Proceedings, Part II; Lecture Notes in Computer Science* (G. Fichtinger, A. Martel, and T. Peters, eds.), vol. 6892, pp. 655–662. Springer, 2011.

[15] J. Hinkle, P. Muralidharan, P. T. Fletcher, and S. Joshi, *Polynomial Regression on Riemannian Manifolds*, in *Computer Vision – ECCV 2012. Lecture Notes in Computer Science* (A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, eds.), vol. 7574, pp. 1–14, Springer, 2012.

[16] P. T. Fletcher, *Geodesic Regression and the Theory of Least Squares on Riemannian Manifolds*, *International Journal of Computer Vision* **105** (2013) 171–185.

[17] E. Cornea, H. Zhu, P. Kim, and J. G. Ibrahim, *Regression models on Riemannian Symmetric Spaces*, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79** (2017), no. 2 463–482.

[18] J. J. Faraway, *Regression for non-Euclidean data using distance matrices*, *Journal of Applied Statistics* **41** (2014), no. 11 2342–2357.

[19] A. Petersen and H.-G. Müller, *Fréchet regression for random objects with Euclidean predictors*, *The Annals of Statistics* **47** (2019), no. 2 691–719.

[20] H. Ichimura, *Semiparametric least squares (SLS) and weighted SLS estimation of single-index models*, *Journal of Econometrics* **58** (1993), no. 1-2 71–120.

[21] A. Ghosal, W. Meiring, and A. Petersen, *Fréchet single index models for object response regression*, *Electronic Journal of Statistics* **17** (2023), no. 1 1074 – 1112.

[22] M. R. Kosorok and E. B. Laber, *Precision medicine*, *Annual Review of Statistics and Its Application* **6** (2019) 263–286.

[23] E. J. Topol, *A decade of digital medicine innovation*, *Science Translational Medicine* **11** (2019), no. 498 [https://stm.sciencemag.org/content/11/498/eaaw7610.full.pdf].

[24] X. Li, J. Dunn, D. Salins, G. Zhou, W. Zhou, S. M. Schüssler-Fiorenza Rose, D. Perelman, E. Colbert, R. Runge, S. Rego, R. Sonecha, S. Datta, T. McLaughlin, and M. P. Snyder, *Digital Health: Tracking physiomes and Activity Using Wearable Biosensors Reveals Useful Health-Related Information*, *PLOS Biology* **15** (2017), no. 1 e2001402.

[25] A. Javaid, F. Zghyer, C. Kim, E. M. Spaulding, N. Isakadze, J. Ding, D. Kargillis, Y. Gao, F. Rahman, D. E. Brown, *et. al.*, *Medicine 2032: The future of cardiovascular disease prevention with machine learning and digital health technology*, *American Journal of Preventive Cardiology* **12** (2022) 100379.

[26] R. Ghosal, V. R. Varma, D. Volfson, J. Urbanek, J. M. Hausdorff, A. Watts, and V. Zipunnikov, *Scalar on time-by-distribution regression and its application for modelling associations between daily-living physical activity and cognitive functions in alzheimer's disease*, *Scientific Reports* **12** (2022), no. 1 1–16.

[27] R. Ghosal, V. R. Varma, D. Volfson, I. Hillel, J. Urbanek, J. M. Hausdorff, A. Watts, and V. Zipunnikov, *Distributional data analysis via quantile functions and its application to modeling digital biomarkers of gait in Alzheimer's Disease*, *Biostatistics* **24** (2021), no. 3 539–561.

[28] M. Matabuena and A. Petersen, *Distributional data analysis of accelerometer data from the NHANES database using nonparametric survey regression models*, *Journal of the Royal Statistical Society Series C: Applied Statistics* **72** (2023), no. 2 294–313, [https://academic.oup.com/jrsssc/article-pdf/72/2/294/50296582/qlad007.pdf].

[29] M. Matabuena, A. Petersen, J. C. Vidal, and F. Gude, *Glucodensities: a new representation of glucose profiles using distributional data analysis*, *Statistical Methods in Medical Research* **30** (2021), no. 6 1445–1464.

[30] M. Matabuena, P. Félix, Z. A. A. Hammouri, J. Mota, and B. del Pozo Cruz, *Physical activity phenotypes and mortality in older adults: a novel distributional data analysis of accelerometry in the NHANES*, *Aging Clinical and Experimental Research* **34** (2022), no. 12 3107–3114.

[31] L. Biagi, A. Bertachi, M. Giménez, I. Conget, J. Bondia, J. A. Martín-Fernández, and J. Vehí, *Individual categorisation of glucose profiles using compositional data analysis*, *Statistical Methods in Medical Research* **28** (2019), no. 12 3550–3567. PMID: 30380996.

[32] T. Battelino, T. Danne, R. M. Bergenstal, S. A. Amiel, R. Beck, T. Biester, E. Bosi, B. A. Buckingham, W. T. Cefalu, K. L. Close, *et. al.*, *Clinical Targets for Continuous Glucose Monitoring Data Interpretation: Recommendations From the International Consensus on Time in Range*, *Diabetes Care* **42** (2019), no. 8 1593–1603.

[33] J. A. Schrack, E. M. Simonsick, P. H. Chaves, and L. Ferrucci, *The Role of Energetic Cost in the Age-Related Slowing of Gait Speed*, *Journal of the American Geriatrics Society* **60** (2012), no. 10 1811–1816.

[34] J. N. Mehta, A. V. Gupta, N. G. Raval, N. Raval, and N. Hasnani, *Physiological cost index of different body mass index and age of an individual*, *National Journal of Physiology, Pharmacy and Pharmacology* **7** (2017), no. 12 1313–1317.

[35] A. Kriska, *Ethnic and Cultural Issues in Assessing Physical Activity*, *Research Quarterly for Exercise and Sport* **71** (2000), no. sup2 47–53.

[36] S. Dogra, B. A. Meisner, and C. I. Ardern, *Variation in mode of physical activity by ethnicity and time since immigration: a cross-sectional analysis*, *International Journal of Behavioral Nutrition and Physical Activity* **75** (2010), no. 7 1–11.

[37] C. J. Caspersen, M. A. Pereira, and K. M. Curran, *Changes in physical activity patterns in the United States, by sex and cross-sectional age*, *Medicine & Science in Sports & Exercise* **32** (2000), no. 9 1601–1609.

[38] R. K. W. Wong, Y. Li, and Z. Zhu, *Partially Linear Functional Additive Models for Multivariate Functional Data*, *Journal of the American Statistical Association* **114** (2019), no. 525 406–418.

[39] J.-L. Wang, J.-M. Chiou, and H.-G. Müller, *Functional Data Analysis*, *Annual Review of Statistics and Its Application* **3** (2016), no. 1 257–295.

[40] W. Xiao, Y. Wang, and H. Liu, *Generalized partially functional linear model*, *Scientific reports* **11** (2021), no. 1 1–14.

[41] H. Zhu, R. Zhang, Y. Liu, and H. Ding, *Robust estimation for a general functional single index model via quantile regression*, *Journal of the Korean Statistical Society* **51** (2022), no. 4 1041–1070.

[42] A. Ghosal, M. Matabuena, W. Meiring, and A. Petersen, *Predicting distributional profiles of physical activity in the NHANES database using a Partially Linear Single-Index Fréchet Regression model*, *arXiv preprint arXiv:2302.07692* (2023).

[43] T. Lumley, *Complex Surveys: A Guide to Analysis Using R.* John Wiley and Sons, 2010.

[44] C. Villani, *Topics in Optimal Transportation*. American Mathematical Society, Graduate Studies in Mathematics, Volume 58, 2003.

[45] B. Afsari, *Riemannian L^p center of mass: Existence, uniqueness, and convexity*, *Proceedings of the American Mathematical Society* **139** (2011), no. 2 655–673.

[46] X. Pennec, *Barycentric subspace analysis on manifolds*, *The Annals of Statistics* **46** (2018), no. 6A 2711–2746.

[47] R. Bhattacharya and V. Patrangenaru, *Large sample theory of intrinsic and extrinsic sample means on manifolds*, *The Annals of Statistics* **31** (2003), no. 1 1–29.

[48] R. Bhattacharya and V. Patrangenaru, *Large sample theory of intrinsic and extrinsic sample means on manifolds: II*, *The Annals of Statistics* **33** (2005), no. 3 1225–1259.

[49] J. Fan and I. Gijbels, *Local Polynomial Modelling and its Applications: Monographs on Statistics and Applied Probability*, vol. 66. Chapman and Hall, 1996.

[50] A. W. Van Der Vaart and J. A. Wellner, *Weak convergence*, in *Weak convergence and empirical processes*, pp. 16–28. Springer, 1996.

[51] W. Lin and K. Kulasekera, *Identifiability of single-index models and additive-index models*, *Biometrika* **94** (2007), no. 2 496–501.

[52] S. Bhattacharjee and H.-G. Müller, *Single index Fréchet regression*, *arXiv preprint arXiv:2108.05437* (2021).

[53] N. Boumal, B. Mishra, P.-A. Absil, and R. Sepulchre, *Manopt, a Matlab toolbox for optimization on manifolds*, *Journal of Machine Learning Research* **15** (2014), no. 42 1455–1459.

[54] Human Mortality Database, "Max Planck Institute for Demographic Research (Germany), University of California, Berkeley (USA), and French Institute for Demographic Studies (France).." Available at `www.mortality.org` (data downloaded on August 18, 2020).

[55] The World Bank, "GDP year-on-year percentage change." Available at `https://data.worldbank.org/indicator/NY.GDP.MKTP.KD.ZG` (accessed September 12, 2022).

[56] The World Bank, "CO_2 Emissions in metric tonnes per capita." Available at `https://data.worldbank.org/indicator/EN.ATM.CO2E.PC` (accessed September 12, 2022).

[57] The World Bank, "Current healthcare expenditure (% of GDP)." Available at `https://data.worldbank.org/indicator/SH.XPD.CHEX.GD.ZS` (accessed September 12, 2022).

[58] United Nations Development Programme, "Human Development Index for 2013." Available at `https://hdr.undp.org/data-center/documentation-and-downloads` (accessed September 12, 2022).

[59] United Nations Inter-agency Group for Child Mortality Estimation, "UN IGME estimate for 2013.5." Available at `https://childmortality.org/data` (accessed September 16, 2022).

[60] S. Hassanipour-Azgomi, A. Mohammadian-Hafshejani, M. Ghoncheh, F. Towhidi, S. Jamehshorani, and H. Salehiniya, *Incidence and mortality of prostate cancer and their relationship with the Human Development Index worldwide*, *Prostate International* **4** (2016), no. 3 118–124.

[61] M. Ghoncheh, M. Mirzaei, and H. Salehiniya, *Incidence and mortality of breast cancer and their relationship with the Human Development Index (HDI) in the world in 2012*, *Asian Pacific Journal of Cancer Prevention* **16** (2016), no. 18 8439–8443.

[62] E. Rasoulinezhad, F. Taghizadeh-Hesary, and F. Taghizadeh-Hesary, *How is mortality affected by fossil fuel consumption, CO_2 emissions and economic factors in CIS region?*, *Energies* **13** (2020), no. 9 2255.

[63] J. A. T. Granados, *Recessions and mortality in Spain, 1980–1997*, *European Journal of Population/Revue Européenne de Démographie* **21** (2005), no. 4 393–422.

[64] J. Eyer, *Prosperity as a cause of death*, *International Journal of Health Services* **7** (1977), no. 1 125–150.

[65] R. Higgs, *Cycles and trends of mortality in 18 large American cities, 1871–1900*, *Explorations in Economic History* **16** (1979), no. 4 381–408.

[66] J. D. Graham, B.-H. Chang, and J. S. Evans, *Poorer is riskier*, *Risk Analysis* **12** (1992), no. 3 333–337.

[67] P. A. Owusu, S. A. Sarkodie, and P. A. Pedersen, *Relationship between mortality and health care expenditure: Sustainable assessment of health care system*, *PLOS ONE* **16** (2021), no. 2 e0247413.

[68] G. Lippi, C. Mattiuzzi, and G. Cervellin, *No correlation between health care expenditure and mortality in the European Union*, *European Journal of Internal Medicine* **32** (2016) e13–e14.

[69] M. B. Rothberg, J. Cohen, P. Lindenauer, J. Maselli, and A. Auerbach, *Little evidence of correlation between growth in health care spending and reduced mortality*, *Health Affairs* **29** (2010), no. 8 1523–1531.

[70] Y. Chen, A. Gajardo, J. Fan, Q. Zhong, P. Dubey, K. Han, S. Bhattacharjee, and H.-G. Müller, *frechet: Statistical Analysis for Random Objects and Non-Euclidean Data*, 2020. R package version 0.2.0, available at `https://CRAN.R-project.org/package=frechet`.

[71] A. Petersen, P. Z. Hadjipantelis, and H.-G. Müller, *fdadensity: Functional Data Analysis for Density Functions by Transformation to a Hilbert Space*, 2019. R package version 0.1.2.

[72] V. M. Panaretos and Y. Zemel, *Amplitude and phase variation of point processes*, *The Annals of Statistics* **44** (2016), no. 2 771–812.

[73] A. Petersen and H.-G. Müller, *Wasserstein covariance for multiple random densities*, *Biometrika* **106** (2019), no. 2 339–351.

[74] Y. Zemel and V. M. Panaretos, *Fréchet means and Procrustes analysis in Wasserstein space*, *Bernoulli* **25** (2019), no. 2 932–976.

[75] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu, *A limited memory algorithm for bound constrained optimization*, *SIAM Journal on Scientific Computing* **16** (1995), no. 5 1190–1208.

[76] J. H. Friedman and W. Stuetzle, *Projection pursuit regression*, *Journal of the American Statistical Association* **76** (1981) 817–823, [https://www.tandfonline.com/doi/pdf/10.1080/01621459.1981.10477729].

[77] P. Hall, *On projection pursuit regression*, *The Annals of Statistics* **17** (1989), no. 2 573–588.

[78] W. Härdle and T. M. Stoker, *Investigating smooth multiple regression by the method of average derivatives*, *Journal of the American Statistical Association* **84** (1989), no. 408 986–995.

[79] Y. Xia, *Asymptotic distributions for two estimators of the single-index model*, *Econometric Theory* **22** (2006), no. 6 1112–1137.

[80] K.-C. Li, *Sliced inverse regression for dimension reduction*, *Journal of the American Statistical Association* **86** (1991), no. 414 316–327.

[81] M. Hein, *Robust nonparametric regression with metric-space valued output*, in *Advances in Neural Information Processing Systems* (Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, eds.), vol. 22, Curran Associates, Inc., 2009.

[82] C. Villani, *Optimal transport: old and new*, vol. 338. Springer, 2009.

[83] G. Peyré and M. Cuturi, *Computational Optimal Transport: With Applications to Data Science*, *Foundations and Trends® in Machine Learning* **11** (2019), no. 5-6 355–607.

[84] V. M. Panaretos and Y. Zemel, *Statistical aspects of Wasserstein distances*, *Annual Review of Statistics and its Application* **6** (2019), no. 1 405–431.

[85] L. Kish, *Survey sampling*. No. 04; HN29, K5. 1965.

[86] T. Lumley, *Analysis of Complex Survey Samples*, *Journal of Statistical Software* **9** (2004), no. 8 1–19.

[87] D. G. Horvitz and D. J. Thompson, *A Generalization of Sampling Without Replacement from a Finite Universe*, *Journal of the American Statistical Association* **47** (1952), no. 260 663–685.

[88] S. Rabe-Hesketh and A. Skrondal, *Multilevel Modelling of Complex Survey Data*, *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **169** (2006), no. 4 805–827.

[89] W. Wang and J. Yan, *Shape-Restricted Regression Splines with R Package splines2*, *Journal of Data Science* **19** (2021), no. 3 498–517.

[90] T. Lumley, *survey: analysis of complex survey samples*, 2020. R package version 4.2-1.

[91] E. Cui, A. Leroux, E. Smirnova, and C. M. Crainiceanu, *Fast Univariate Inference for Longitudinal Functional Models*, *Journal of Computational and Graphical Statistics* **31** (2022), no. 1 219–230.

[92] K. G. Van den Boogaart, J. J. Egozcue, and V. Pawlowsky-Glahn, *Bayes hilbert spaces*, *Australian & New Zealand Journal of Statistics* **56** (2014), no. 2 171–194.

[93] A. Petersen, H.-G. Müller, *et. al.*, *Functional data analysis for density functions by transformation to a hilbert space*, *The Annals of Statistics* **44** (2016), no. 1 183–218.

[94] K. Hron, A. Menafoglio, M. Templ, K. Hruuzova, and P. Filzmoser, *Simplicial principal component analysis for density functions in bayes spaces*, *Computational Statistics & Data Analysis* **94** (2016) 330–350.

[95] A. Javaid, F. Zghyer, C. Kim, E. M. Spaulding, N. Isakadze, J. Ding, D. Kargillis, Y. Gao, F. Rahman, D. E. Brown, S. Saria, S. S. Martin, C. M. Kramer, R. S. Blumenthal, and F. A. Marvel, *Medicine 2032: The future of cardiovascular disease prevention with machine learning and digital health technology*, *American Journal of Preventive Cardiology* **12** (2022) 100379.

[96] J. Zhang, K. R. Merikangas, H. Li, and H. Shou, *Two-sample tests for multivariate repeated measurements of histogram objects with applications to wearable device data*, *The Annals of Applied Statistics* **16** (2022), no. 4 2396–2416.

[97] J. D. A. Relión, D. Kessler, E. Levina, and S. F. Taylor, *Network classification with applications to brain connectomics*, *The Annals of Applied Statistics* **13** (2019), no. 3 1648–1667.

[98] Y. Zhou and H.-G. Müller, *Network regression with graph Laplacians*, *The Journal of Machine Learning Research* **23** (2022), no. 1 14383–14423.

[99] T. M. Nye, X. Tang, G. Weyenberg, and R. Yoshida, *Principal component analysis and the locus of the fréchet mean in the space of phylogenetic trees*, *Biometrika* **104** (2017), no. 4 901–922.

[100] P. Dubey and H.-G. Müller, *Modeling time-varying random objects and dynamic networks*, *Journal of the American Statistical Association* **117** (2022), no. 540 2252–2267, [https://doi.org/10.1080/01621459.2021.1917416].

[101] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, New York, New York, 2nd ed., 2009.