November 2023

NONPARAMETRIC DERIVATIVE ESTIMATION USING PENALIZED SPLINES: THEORY AND APPLICATION

Bright Antwi Boasiako
University of Massachusetts Amherst

NONPARAMETRIC DERIVATIVE ESTIMATION USING PENALIZED SPLINES: THEORY AND APPLICATION

A Dissertation Presented

by

BRIGHT ANTWI BOASIAKO

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

September 2023

Department of Mathematics and Statistics

NONPARAMETRIC DERIVATIVE ESTIMATION USING PENALIZED SPLINES: THEORY AND APPLICATION

A Dissertation Presented

by

BRIGHT ANTWI BOASIAKO

Approved as to style and content by:

John W. Staudenmayer, Chair

Anna Liu, Member

Ted Westling, Member

Chi Hyun Lee, Member

Tom Braden, Graduate Program Director
Department of Mathematics and Statistics

DEDICATION

Dedicated to my beloved family and friends.

ACKNOWLEDGMENTS

First, I would like to express my deepest appreciation to my thesis advisor, Prof. John Staudenmayer whose guidance, patience, and profound knowledge have been invaluable throughout this journey. His unwavering confidence in me has inspired and motivated me, and I am truly grateful for his mentorship.

Again, I would like to thank the members of my dissertation committee for their insightful comments and suggestions, and for their time and dedication, which significantly contributed to improving the quality of my research. My heartfelt thanks also go to my department for providing a stimulating and supportive environment for my studies.

I want to express my deepest gratitude to my parents and my grandmother. Even though they are no longer with us, their love, values, and belief in the power of education continue to guide me. I carry their memory with me as I embark on this new chapter in my academic journey.

My uncle Naval Capt. I. Y. Kwantwi-Mensah deserves a special mention for his unwavering support, both financially and emotionally. His faith in my potential and his encouragement have been a constant source of strength, for which I am sincerely grateful.

I also wish to express my gratitude to my friends and all those who were a part of this journey. Their constant companionship, understanding, and shared wisdom have provided much-needed support throughout my academic journey.

ABSTRACT

NONPARAMETRIC DERIVATIVE ESTIMATION USING PENALIZED SPLINES: THEORY AND APPLICATION

SEPTEMBER 2023

BRIGHT ANTWI BOASIAKO

B.Sc., KWAME NKRUMAH UNIVERSITY OF SCIENCE AND TECHNOLOGY

M.Sc., UNIVERSITY OF MASSACHUSETTS AMHERST

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor John W. Staudenmayer

This dissertation is in the field of Nonparametric Derivative Estimation using Penalized Splines. It is conducted in two parts. In the first part, we study the L_2 convergence rates of estimating derivatives of mean regression functions using penalized splines. In 1982, Stone provided the optimal rates of convergence for estimating derivatives of mean regression functions using nonparametric methods. Using these rates, Zhou et. al. in their 2000 paper showed that the MSE of derivative estimators based on regression splines approach zero at the optimal rate of convergence. Also, in 2019, Xiao showed that, under some general conditions, penalized spline estimators of mean regression functions achieve optimal L_2 rates of convergence. We extend this result to derivative estimators. In particular, we show that under similar conditions, penalized spline estimators of derivatives of mean regression functions achieve optimal L_2 rates of convergence.

In the second part of the thesis, we estimate the amount of association between physical activity and all-cause mortality in US adults using penalized splines. We introduce a novel nonparametric isotemporal substitution model to investigate the dose-response relationship between daily time allocations across physical activity and sedentary behaviors, and all-cause mortality. Our method reveals that the association between such daily time allocations and mortality depends on one's level of physical activity. We apply our method to data from the 2003-2006 wave of the US National Health and Nutrition Examination Survey (NHANES) with mortality follow-up through December 31st, 2019, a nationally representative survey. Among US adults with less than 6 hours of daily activity, replacing 1 hour of physical activity with sedentary behaviors is associated with up to 68% increase in mortality risks after adjusting for sleep time and baseline demographic and health covariates. In addition, for those with sedentary time above 50% of non-sleep time, replacing 1% of moderate-to-vigorous activity (MVPA) time with sedentary time is associated with up to 18% increase in mortality risks. Therefore, to better understand mortality risk associations, US adults may consider their full daily activity time allocations before replacing one activity type with another.

TABLE OF CONTENTS

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

CHAPTER 1

INTRODUCTION

This dissertation aims to contribute to the field of Nonparametric Derivative Estimation, specifically focusing on Penalized Splines. The research is split into two parts; the first focuses on the theoretical properties of penalized spline derivative estimators while the second part uses a novel penalized spline-based survival analysis to study the association between physical activity and all-cause mortality in a nationally representative survey among US adults.

1.1 Optimal L_2 convergence rates of penalized spline derivative estimators.

In a standard regression problem with independent and identically distributed (IID) data, represented as $\{x_i, y_i\}_{i=1}^n$, we use derivatives to estimate the associations between dependent and independent variables. To put it in context, consider the standard linear model represented by the equation $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ (ϵ_i's are IID with mean 0 and variance σ^2). Here, β_1, which is the first derivative of the mean regression function $\mathbb{E}(y_i|x_i) = \beta_0 + \beta_1 x_i$, serves as a measure of the association between y (the dependent variable) and x (the independent variable).

Moving beyond linear models, nonparametric methods allow us to handle mean regression functions in a more flexible way. Instead of adhering to the linearity assumption in the standard linear model, these methods model the mean regression function as $\mathbb{E}(y_i|x_i) = f(x_i)$, where f is an unknown function. This effectively relaxes

the linearity constraint, offering robust modeling of complex relationships with minimal assumptions on f such as smoothness. However, this does not mean we abandon the use of derivatives. Quite the contrary, we still use them to estimate associations in these models. The primary difference lies in the fact that, unlike the constant derivative in a linear model (β_1), the derivatives in these nonparametric models can change depending on the value of the independent variable x. For instance, we can represent the first derivative as $f'(x)$, signifying that the derivative of the mean regression function is a function of x.

In the context of nonparametric models, there are two primary method categories employed to estimate derivatives of mean regression functions. The first category involves a two-step process: initially, the mean regression function is estimated. Then, the obtained estimator is differentiated to estimate the derivatives ([49, 60, 30]). The second category of methods takes a more direct approach. Rather than first estimating the mean regression function, these methods aim to estimate the derivatives directly and thus avoid the need for the initial step of estimating the mean regression function ([13, 11]). In the first part of this thesis, we focus on spline-based nonparametric methods for derivative estimation, more specifically, we focus on penalized spline derivative estimators. These methods fall under the first category of nonparametric derivative estimators.

A spline function is made up of piece-wise polynomials that are joined together at various locations (referred to as knots) with some continuity conditions at the knot locations. With the flexibility that polynomials provide and the continuity conditions at the knots, splines provide a rich class of smooth functions that are robust to model misspecification to model non-linear mean regression functions in a wide range of applications.

The number of polynomial pieces and the placement of knots in a spline function form the basis for different spline-based data smoothing methods. In Regression

Splines, a predetermined number of knots is selected (often through cross-validation) and the least squares method is used to determine the coefficients of the spline function that best fits the given data. On the other hand, Smoothing Splines avoid the knot placement issue by placing a knot at every unique value of x. Penalization is then used to balance the bias-variance trade-off. Lastly, Penalized Splines straddle the line between Regression Splines and Smoothing Splines by using a large but non-exhaustive number of knots and employing penalization to balance the bias-variance trade-off.

1.1.1 Asymptotic properties of spline-based derivative estimators

There are a number of asymptotic results in the literature for spline-based estimators of mean regression functions and their derivatives. For instance, [60] in 2000 showed that the mean squared error (MSE) of a regression spline derivative estimator approaches zero at the optimal rate of convergence. Here, optimality is based on the work by Stone in 1982 which provides global optimal rates of convergence for nonparametric derivative estimators of mean regression functions. Also, [48] studied the asymptotics of smoothing spline derivative estimators and found that the optimal smoothing parameter depends on the derivative being estimated.

In this first work of the thesis, we contribute to the asymptotic understanding of penalized spline derivative estimators by studying their L_2 convergence rates. Building on the work of [58, 8], we show that the penalized spline derivative estimator of the r^{th} derivative of the mean regression function, $\mathbb{E}(y_i|x_i) = f(x_i)$, achieves the optimal L_2 rate of convergence [45]. We remark that the L_2 asymptotics of the penalized spline derivative estimator are similar to those of regression splines when the number of knots increases slowly with the sample size, n, and are similar to smoothing spline results when the number of knots increases faster with n. This remark is consistent

with asymptotic results of penalized spline estimators of the mean regression function itself [8].

Even though we show that the penalized spline derivative estimator converges at the optimal L_2 rate, this result does not cover the performance of the derivative estimator in finite samples. We performed an extensive simulation study to understand how the estimator performs in finite samples. By comparing the estimator with an estimator that uses knowledge of the true derivative to choose its smoothing parameter (referred to as the "oracle" estimator), we observe that there may be quite substantial differences between the MSE of the penalized spline derivative estimator and the oracle estimator, especially for higher derivatives.

1.2 Nonparametric estimation of the association between objectively measured physical activity and mortality.

The significance of physical activity in promoting longevity and health is widely recognized. Physical activity is often categorized into three main types based on the intensity level. These are sedentary behaviors, light physical activity, and moderate-to-vigorous physical activity (MVPA). Numerous epidemiological studies have established a clear association between physical activity and reduced risk of chronic illnesses such as stroke, diabetes, and heart disease ([27, 50]). Furthermore, it has been established that regular exercise can influence overall mortality risks [22]. However, a crucial problem remains: the exact dose-response relationship between physical activity and mortality remains unknown ([34]).

Understanding this dose-response relationship is further complicated by two fundamental challenges. Firstly, the finite nature of time – with only 24 hours in a day – means that the total time a person spends engaging in various types of physical activity sums up to a constant. This constraint leads to perfect multicollinearity, posing a challenge to traditional statistical techniques. Data exhibiting such charac-

teristics are referred to as "compositions" in the literature ([2]). To address this issue, isotemporal substitution methods ([28, 52]) are typically used. In such approaches, one variable from the compositional covariates is omitted from the regression. As a result, increasing any of the remaining compositional covariates by 1 unit, while keeping the other variables constant, implies a decrement of 1 unit in the excluded variable.

The other challenge to understanding the dose-response relationship between physical activity and mortality is data quality. Relying on self-reported measures of physical activity intensity levels is problematic, as individuals have been found to overestimate the time they spend engaging in physical activity ([41, 46, 22]).

A more objective and reliable estimation of one's physical activity can be achieved through the use of accelerometer measurements, when available. Accelerometers offer a precise way of capturing physical activity intensity levels. For instance, a study by Troiano et al. ([46]), using accelerometry data, revealed a stark discrepancy between self-reported and objectively measured physical activity. They found that less than 10% of US adults engage in the recommended 30 minutes per day of MVPA for most days of the week (at least five days). This figure stands in contrast to self-reported data, where more than 30% claimed to meet these recommendations. Such discrepancies underscore the importance of reliable and objective physical activity measurements in studying the relationship between physical activity and mortality.

In 2016, Fishman et al. ([22]) used objectively measured physical activity data from the 2003-2006 wave of the US National Health and Nutrition Examination Survey (NHANES) with mortality follow-up through December 31st, 2011 [32]. One of the goals of that research was to study the isotemporal substitution effects of physical activity time allocations on all-cause mortality among US adults. The authors used a Cox model [9] where they represented the effect of physical activity time in the log-hazard function as a linear function. While the study by [22] and others like it,

deepen our understanding of the association between replacing one physical activity time with another and mortality, there are two main drawbacks to their approach that we want to highlight and address in this work.

The first drawback pertains to the assumption of uniformity in the estimated associations. Specifically, the existing models assume that the impact of adjustments in physical activity time allocations remains constant, irrespective of an individual's current activity allocations. For instance, the linear model predicts identical effects when a person increases their daily sedentary time from 1 hour to 2 hours, as compared to increasing it from 7 hours to 8 hours. In practical applications, however, it may not be reasonable to assume such a uniform effect. The second drawback lies in the underestimation of substitution effects. Specifically, the linear model is liable to predict smaller effects from substituting one type of physical activity for another on mortality, which may not accurately represent the real associations.

In the second part of this dissertation, we present a novel nonparametric approach for quantifying the relationships between physical activity and mortality. This method can be applied to any regression involving a time-to-event response with compositional covariates. In a Cox model, our approach estimates the effects of such compositional covariates in the log-hazard function using a multivariate penalized tensor product spline. The incorporation of this flexible functional form allows our model to detect varying effects across different regions of the hazard function. Furthermore, the substitution effects within our model are determined by the partial derivatives of the fitted tensor product spline surface, providing a more nuanced understanding of the relationships within the data.

We apply our method to the NHANES data used in Fishman et al. ([22]) but with mortality follow-up through December 31st, 2019 [31]. To facilitate comparison, we conduct a parallel analysis that, similar to [22], represents the physical activity covariates as linear terms in the log-hazard function. Our analysis results show that,

among U.S. adults, the impact of substituting one type of physical activity for another on mortality can differ greatly depending on the individual's current level of physical activity. Such an insight is easily missed by the linear model. In all the analyses conducted in this work, adjustments were made for baseline demographic and health-related covariates. These included age, education level, body mass index (BMI), smoking status, presence or absence of diseases such as diabetes, heart disease, stroke, and cancer, as well as whether an individual has a mobility limitation.

The rest of the dissertation is structured as follows: In Chapter 2, we present the asymptotic study and results of the L_2 convergence rates of the Penalized Spline derivative estimator. In this chapter, we definite the estimator more clearly and state the assumptions we make for our main result. The result appears in Theorem 1. Next, we provide a formal proof of the theorem with needed technical lemmas appearing in the appendix. We end the chapter with an extensive simulation study to assess the L_2 convergence rates of the estimator for the first two derivatives and how the estimator performs in finite samples. Chapter 3 contains our nonparametric method to estimate the substitution effects of compositional variables in a Cox model and the associated partial derivatives used to measure such substitution effects. The chapter also contains the analyses that apply our nonparametric method to the NHANES data. We conclude the dissertation with a discussion of our results and the directions for future research in Chapter 4.

CHAPTER 2

OPTIMAL L_2 CONVERGENCE OF PENALIZED SPLINE DERIVATIVE ESTIMATORS

2.1 Introduction

We consider the situation where data $\{x_i, y_i\}_{i=1}^n$, sampled from the model:

$$y_i = f(x_i) + \varepsilon_i, \quad \forall i = 1, 2, \ldots, n \tag{2.1}$$

with $f \in \mathcal{C}^p(\mathcal{K})$, the space of functions with p continuous derivatives over $\mathcal{K} = [0, 1]$, the x_i's, for $x_i \in \mathcal{K}$, are either random or deterministic, and ε_i's are independent and identically distributed random error terms with $E[\varepsilon_i] = 0$ and $Var[\varepsilon_i] = \sigma^2$. There are many cases when it is of interest to estimate some derivative of the mean regression function f, with minimal assumptions on the functional form of f. For example, in human growth studies, the first derivative of the function relating height and age indicates the speed of growth ([29, 35]). Additionally, [7] apply derivative estimation in the development of a visual mechanism for studying curve structures, and [33] compare regression curves using those structures. In economics, derivatives are used to calculate the marginal propensity to consume, which measures the effect of changes in disposable income on personal consumption ([21]). In addition, average derivatives of mean regression functions are used to empirically validate the so-called "law of demand" ([25]). It is sufficient for a random matrix composed of average derivatives to be positive definite for the law of demand to hold ([24]). In nonparametric regression itself, estimates of derivatives of the true function, f, are used in plug-in bandwidth

selection techniques such as in local polynomial regression ([36]) and to construct confidence bands for nonparametric estimators ([18]).

Previous work has taken three major approaches to estimating derivatives of functions nonparametrically: local polynomial regression, empirical derivatives, and spline-based methods. In local polynomial regression, a derivative of $f(x)$ can be estimated using a coefficient of the fitted local polynomial at x ([19], page 22). Empirical methods generally transform the data and smooth difference-based estimates of derivatives. Earlier works include [30], which used Kernel-based approaches to estimate the derivatives of the mean regression function while utilizing difference quotients to identify the best kernel bandwidth via cross-validation. More recently, [14] used symmetric difference quotients to estimate derivatives of mean regression functions and showed that their approach improved the asymptotic order of the variance. Spline-based methods use the fact that splines are piecewise polynomials. As a result, differentiating the basis function with respect to the covariate gives a basis function for the derivative, and an estimate of that function then can be obtained using a subset of the estimated coefficients ([12], page 115, [16, 17]).

While it is straightforward to compute nonparametric derivative estimates, the challenge is that those estimates also require some sort of regularization to balance estimation bias and overfitting, and methods to choose that regularization are usually designed for estimating the function itself, not derivatives ([38, 16]). Several authors have designed methodologies to address that problem (e.g. [6, 43]), but it has also been suggested that methods that choose the amount of smoothing for the derivatives as if the function were of interest often work well in practice ([38], page 154, [10]). We call such methods *naive*. In this work, we add evidence to that debate by exploring the asymptotic behavior of naive nonparametric derivative estimators, focusing on penalized splines.

The past few decades have seen some progress in related areas. In local polynomial regression, results from [37] can be used to show that when a bandwidth is chosen to minimize the integrated mean squared error (IMSE) of a p^{th} degree polynomial estimate of f, and that bandwidth is used to estimate $f^{(r)}$ ($p \geq r$, i.e. a naive estimator), then the derivative estimate's IMSE converges at an optimal rate if r is even. Otherwise, the naive bandwidth over- or under-smooths. In particular, a naive estimator of the first derivative using cubic local polynomials under-smooths. A derivation of this is given in the appendix.

Also, [14] showed that the asymptotic order of the bias of empirical estimators does not depend on the order of the derivative being estimated, and they employ a method by [13] to deal with correlations that result from creating the empirical dataset for the derivative. [11] generalized those results by considering linear combinations of observations to better fit both interior and boundary points. They also demonstrated that their method achieves optimal rates of convergence ([45]).

The asymptotic properties of two types of spline-based derivative estimators have been considered too. [48] studied smoothing splines and found that the optimal smoothing parameter depends on the order of the derivative being estimated. In contrast, [60] considered the asymptotics of regression spline-based derivative estimators where the number of knots increases with the sample size. They showed that the MSE goes to zero at the optimal rate ([45]), and the required rate of increase in the number of knots does not depend on the order of the derivative.

Somewhat surprisingly though, comparable results about penalized spline estimators of derivatives do not seem to exist. Building on work on the asymptotics of penalized spline estimators of functions ([58]), we derive the apparently new result that naive methods to estimate derivatives with penalized splines achieve optimal global L_2 rates of convergence ([45]).

The rest of the chapter is structured as follows: in Section 2.2, we give some background on splines, penalized splines, and the naive derivative estimator. In Section 2.3, we present our main result and remark that depending on the rate at which the number of knots increases with the sample size, n, the L_2 convergence of the naive estimator is similar to that of regression splines or smoothing splines. We also provide the proof of theorem 1 in this section. In Section 2.4, we present a simulation study of the L_2 rate of convergence of the naive estimator, and we conclude with a discussion in Section 2.5. Proofs of the technical lemmas for the theorem are given in the appendix. We also provide a note on the asymptotic rates of convergence on local polynomial derivative estimators in the appendix.

2.2 Penalized Splines & the Naive Derivative Estimator

2.2.1 Splines

Splines provide a flexible mechanism to estimate derivatives of the mean regression function f, and in the case of estimating the function itself, they have been shown to do so at the best possible rates of convergence ([58, 57, 45]). A spline is a piece-wise polynomial with continuity conditions at the points where the pieces join together (called knots).

More specifically, for $q \geq 2$, we let

$$\boldsymbol{\mathcal{S}}(q, \underline{\mathbf{t}}) = \left\{ s \in \mathcal{C}^{q-2}(\mathcal{K}) : s \text{ is a } q\text{-order polynomial on each } [t_i, t_{i+1}] \right\}$$

be a space of $q-$order splines over $\mathcal{K} = [0, 1]$ with knot locations $\underline{\mathbf{t}} = (t_0, t_1, \ldots, t_{K+1})$ where $t_0 = 0$, $t_{K+1} = 1$ and $t_i < t_j \ \forall_{i<j}$. For $q = 1$, $\boldsymbol{\mathcal{S}}(q, \underline{\mathbf{t}})$ consists of step functions with jumps at the knots.

This space has a number of equivalent bases and one notable for having stable numerical properties is the *B-Spline* basis ([12, 38, 40]).

de Boor ([12]) defines the q^{th} order B-spline basis function $B_{j,q}(x)$ over the knot locations $\underline{\mathbf{t}}$ through a recurrence relation. [17] show that when the distance between the knots is constant, $B_{j,q}(x)$ reduces to

$$B_{j,q}(x) = \frac{(-1)^q \Delta^q (x - t_j)_+^{q-1}}{(q-1)! h^{q-1}}$$

where Δ is the backward difference operator ($\Delta t_j = t_j - t_{j-1}$), and h is the common distance between the knots. Observe that $B_{j,q}(x)$, in this case, is a rescaled q-order difference of truncated polynomials. To get a complete set of B-Spline basis, we need $2q$ extra knots with q knots on each side of $[0, 1]$. This is referred to as the expanded basis ([17]).

Without losing generality, we will assume a B-Spline basis for $\boldsymbol{\mathcal{S}}(q, \underline{\mathbf{t}})$ for the rest of this work. We refer the reader to [12, 40, 16] for an introduction to B-Splines, and [17] for how the B-Spline basis compares to the Truncated Polynomial Functions (TPF) on metrics including fit quality, numerical stability, and multidimensional smoothing.

2.2.2 Penalized Splines & the Naive Estimator

Penalized splines are often viewed as a compromise between regression and smoothing splines because they combine penalization and low-rank bases to achieve computational efficiency. They vary slightly based on the basis functions used and the object of penalization. For example, P-Splines ([16]) use B-Spline basis functions and penalize differences of the coefficients to a specific order. In this section, we will focus on P-Splines. Our later results hold for the general penalized spline estimator defined by [58].

A P-Spline estimator of f in (2.1) based on an iid sample of size n finds a *spline* function $g(x) = B(x)\underline{\boldsymbol{\alpha}}$, that minimizes:

$$Q(\underline{\boldsymbol{\alpha}}, \lambda_n) = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - B(x_i)\underline{\boldsymbol{\alpha}}\right)^2 + \lambda_n \underline{\boldsymbol{\alpha}}^T \boldsymbol{P_m} \underline{\boldsymbol{\alpha}} \tag{2.2}$$

where, $\underline{\boldsymbol{\alpha}} = (\alpha_1, \alpha_2, \ldots, \alpha_{K+q})$ is a vector of coefficients, and $B(x_i) = [B_{1,q}(x_i), B_{2,q}(x_i), \ldots, B_{K+q,q}(x_i)] \in \mathbb{R}^{K+q}$ is a vector of basis functions at x_i, for $i = 1, 2, \ldots, n$. The penalty matrix $\boldsymbol{P_m} = \boldsymbol{D_m^T D_m} \in \mathbb{R}^{(K+q)\times(K+q)}$ where $\boldsymbol{D_m}\underline{\boldsymbol{\alpha}}$ is a vector of m^{th} order differences of $\underline{\boldsymbol{\alpha}}$. Finally, $\lambda_n \geq 0$ is the smoothing parameter and needs to be chosen. Three prevalent methods for choosing λ_n are Generalized Cross Validation (GCV), Maximum Likelihood (ML), and Restricted (or residual) Maximum Likelihood (REML); we refer the reader to [53] chapter 4 and [38] chapters 4 and 5 for details.

Minimizing (2.2) with respect to $\underline{\boldsymbol{\alpha}}$ gives $\underline{\hat{\boldsymbol{\alpha}}} = \left(\frac{\boldsymbol{B}^T\boldsymbol{B}}{n} + \lambda_n \boldsymbol{P_m}\right)^{-1} \frac{\boldsymbol{B}^T y}{n}$ which results in $\hat{f}(x) = B(x)\underline{\hat{\boldsymbol{\alpha}}}$. Here, $\boldsymbol{B} = [B(x_1), B(x_2), \ldots, B(x_n)]^T \in \mathbb{R}^{n\times K+q}$. From $\hat{f}$, we can derive the naive estimator of the r^{th} derivative of f as follows:

$$\begin{aligned} \hat{f}^{(r)}(x) &= \frac{d^{(r)}}{dx} B(x)\underline{\hat{\boldsymbol{\alpha}}} \\ &= \frac{d^{(r)}}{dx}\left(\sum_{j=1}^{K+q} \hat{\alpha}_j B_{j,q}(x)\right) \\ &= \sum_{j=1}^{K+q-r} \hat{\alpha}_j^{(r)} B_{j,q-r}(x) \end{aligned} \tag{2.3}$$

where $\hat{\alpha}_j^{(r)} = (q-r)\frac{\left(\hat{\alpha}_{j+1}^{(r-1)} - \hat{\alpha}_j^{(r-1)}\right)}{t_j - t_{j-q+r}}$, with $\hat{\alpha}_j^{(0)} = \hat{\alpha}_j$ for $1 \leq j \leq K+q-r$, and $r = 1, 2, \ldots, q-2$. ([12, 60]).

[58] showed that under some conditions on the distribution of the knots and λ_n, $\hat{f}$ achieves the optimal L_2 rate of convergence ([45]) to the true f but they do not discuss derivative estimators. We extend this result to show that under same conditions that do not depend of the order of the derivative, the naive derivative estimator $\hat{f}^{(r)}$ of $f^{(r)}$ achieves optimal L_2 rates of convergence.

2.3 Main Results

In this section, we provide our main result in Theorem 1 and remark on how this result relates to regression and smoothing splines. Note that the findings in this section apply to the general penalized spline estimator as defined by [58]. This general estimator is based on the realization that the various types of penalized splines differ mainly by their penalty matrices. However, the eigenvalues of the penalty matrices decay at similar rates, making their unified asymptotic study tractable. We refer the interested reader to a derivation of the decay rates of various penalty matrices in [58].

2.3.1 Notation

We start by defining the following notations relating to norms and limits. For a real matrix A, $||A||_\infty = \max_i \sum_j |a_{ij}|$ is the largest row absolute sum. $||A||_2$ is the operator norm of A induced by the vector norm $||.||_2$. $||A||_F = \sqrt{tr(A^T A)}$ is the Frobenius norm. For a real vector, $||\underline{\boldsymbol{a}}|| = \max_i |a_i|$. For a real-valued function $g(x)$ defined on $\mathcal{K} \subset \mathbb{R}$, $||g|| = \sup_{x \in \mathcal{K}} |g(x)|$ and $||g||_{L2} = \left(\int_{x \in \mathcal{K}} (g(x))^2 \, dx \right)^{1/2}$ is the L_2-norm of g. For two real sequences $\{a_n\}_{n \geq 1}$ and $\{b_n\}_{n \geq 1}$, $a_n \sim b_n$ means $\lim_{n \to \infty} \frac{a_n}{b_n} = 1$.

2.3.2 Assumptions

Next, we state assumptions on the knot placement and penalty matrix. We note that these assumptions are the same as those made in [58] for the asymptotic analysis of estimates of functions rather than derivatives.

1. $K = o(n)$.

2. $\max_{1 \leq i \leq K} |h_{i+1} - h_i| = o(K^{-1})$, where $h_i = t_i - t_{i-1}$.

3. $\frac{h}{\min\limits_{1\leq i\leq K} h_i} \leq M$, where $h = \max\limits_{1\leq i\leq K} h_i$ and $M > 0$ is some predetermined constant.

4. For a deterministic design,

$$\sup_{x\in[0,1]} |Q_n(x) - Q(x)| = o(K^{-1})$$

where $Q_n(x)$ is the empirical CDF of x and $Q(x)$ is a distribution with continuously differentiable positive density $q(x)$.

5. The penalty matrix $\boldsymbol{P_m}$ is a banded symmetric positive semi-definite square matrix with a finite bandwidth and $||\boldsymbol{P_m}||_2 = O(h^{1-2m})$. This assumption is similar to Assumption 3 of [58] where it is stated in terms of the eigenvalues of $\boldsymbol{P_m}$. This assumption is verifiable for P-Splines, O-Splines, and T-Splines. See Propositions 4.1 and 4.2 of [58]. Also, we assume $\underline{\boldsymbol{\beta}}^T \boldsymbol{P_m} \underline{\boldsymbol{\beta}} = O(1)$ where $\underline{\boldsymbol{\beta}}$ is the coefficient vector for approximating f in 2.1 with the best approximating spline function $\boldsymbol{s_f}$ in $\boldsymbol{\mathcal{S}}(q, \underline{\mathbf{t}})$ (see Lemma 4).

6. $\lambda_n = o(1)$.

Assumptions (2) and (3) are necessary conditions on the placements of the knots and also imply that $h \sim K^{-1}$. This ensures that $M^{-1} < Kh < M$ and is necessary for numerical computations ([57]).

Theorem 1. *Let the mean regression function in* (2.1) *be such that* $f \in \mathcal{C}^p(\mathcal{K})$. *Under Assumptions* (1) - (6) *above, and for* $m \leq \min(p, q)$*:*

$$\begin{aligned}\mathbb{E}\left(||\hat{f}^{(r)} - f^{(r)}||_{L_2}^2\right) &= O\left(\frac{K_e}{n}\right) + O\left(K^{-2(q-r)}\right) + o(K^{-2(p-r)}) \\ &\quad + O\{\min(\lambda_n^2 K^{2m+2r}, \lambda_n K^{2r})\}\end{aligned}$$

where $K_e = \min\left\{K^{2r+1}, K^{2r}\lambda_n^{-1/2m}\right\}$ *and* $r = 1, 2, \ldots, q-2$.

2.3.3 Remarks

Remark 1. The asymptotics of penalized splines are either similar to those of regression splines or smoothing splines depending on how fast the number of knots increases as the sample size increases ([8, 58]). This creates two scenarios: the small number of knots scenario with asymptotics similar to regression splines and the large number of knots scenario with asymptotics similar to smoothing splines. We explore the rates of convergence of the naive estimator under each of these scenarios in Remarks 1a and 1b below.

Remark 1a (Small number of knots scenario): Suppose the mean regression function is q-times continuously differentiable, where q is the order of the spline used to estimate f. Thus, $f \in \mathcal{C}^q(\mathcal{K})$. Also suppose $\lambda_n K^{2m} = O(1)$, then

$$\begin{aligned}\mathbb{E}\left(||\hat{f}^{(r)} - f^{(r)}||_{L_2}^2\right) &= O\left(\frac{K_e}{n}\right) + O\left(K^{-2(q-r)}\right) + o(K^{-2(p-r)}) \\ &\quad + O\{\min(\lambda_n^2 K^{2m+2r}, \lambda_n K^{2r})\} \\ &= O\left(\frac{K^{2r+1}}{n}\right) + O\left(K^{-2(q-r)}\right) + O(\lambda_n^2 K^{2m+2r}).\end{aligned}$$

Choosing K such that $K \sim n^{\frac{1}{2q+1}}$ and $\lambda_n = O(n^{-(q+m)/(2q+1)})$, the estimator $\hat{f}^{(r)}$ of $f^{(r)}$ converges at the optimal L_2 rate of $n^{-\frac{(q-r)}{2q+1}}$. In the above, we have used the fact that $p = q$ and that $\min\{\lambda_n^2 K^{2m+2r}, \lambda_n K^{2r}\} = \lambda_n^2 K^{2m+2r}$, $K_e = K^{2r+1}$ for $\lambda_n K^{2m} = O(1)$. We note that the λ_n's rate of decrease does not depend on (r), the order of the derivative.

Remark 1b (Large number of knots scenario): Suppose $f \in \mathcal{C}^m(\mathcal{K})$, and there exists a sufficiently large constant, C, independent of K such that for $K \geq C^{1/2m}\lambda_n^{-1/2m} = C^{1/2m} n^{\frac{1}{2m+1}}$, with $m \leq q$, we have

$$\begin{aligned}\mathbb{E}\left(||\hat{f}^{(r)} - f^{(r)}||_{L_2}^2\right) &= O\left(\frac{K_e}{n}\right) + O\left(K^{-2(q-r)}\right) + o(K^{-2(p-r)}) \\ &\quad + O\{\min(\lambda_n^2 K^{2m+2r}, \lambda_n K^{2r})\} \\ &= O\left(\frac{K^{2r}\lambda_n^{-1/2m}}{n}\right) + O\left(K^{-2(q-r)}\right) + o\left(K^{-2(m-r)}\right) \\ &\quad + O(\lambda_n K^{2r}).\end{aligned}$$

Choosing λ_n such that $\lambda_n \sim n^{-2m/(2m+1)}$, the estimator $\hat{f}^{(r)}$ of $f^{(r)}$ converges at the optimal L_2 rate of $n^{-\frac{(m-r)}{2m+1}}$. Again, we note that the λ_n's rate of decrease does not depend on (r).

Remark 2. While the naive estimator of the derivative achieves an optimal rate of convergence, that does not mean that the naive approach is optimal in a finite sample. We compare the performance of the naive estimator to an "oracle estimator" that minimizes mean integrated squared error in Section 4.1.4.

Remark 3. The theorem is derived under conditions on the growth in the number of knots, the spacings between them, and the smoothing parameter (λ_n). Specific rates of growth for K and for λ_n in Remarks 1a and 1b led to optimal rates of convergence. That said, it is not clear whether standard ways of choosing smoothing parameters would lead to optimal rates of converged. This too is explored in Section 4.

2.3.4 Proof of Theorem

The proof proceeds in two steps. We first derive the L_2 rate of convergence for the bias of the naive estimator and then we derive that of the variance. The approach of the proof closely follows the proof for the L_2 rate of convergence of the mean regression function itself found in [58] with a bit more clarity. We start by defining some terms to simplify the notation.

Let $G_{n,q} = \boldsymbol{B}^T\boldsymbol{B}/n$ and $H_n = G_{n,q} + \lambda_n \boldsymbol{P_m}$. To ease exposition, we follow [60] and write $\hat{f}^{(r)}(x)$ as

$$\hat{f}^{(r)}(x) = B_{q-r}(x)D^{(r)}(G_{n,q} + \lambda_n \boldsymbol{P_m})^{-1}\boldsymbol{B}^T\boldsymbol{y}/n$$

where $B_{q-r}(x) \in \mathbb{R}^{K+q-r}$ is a vector of B-Spline basis functions of order $q-r$ and $D^{(r)}$ is defined as $D^{(r)} = M_r^T \times M_{r-1}^T \times \cdots \times M_1^T$ with

$$M_l = (q-1)\begin{bmatrix} \frac{-1}{t_1 - t_{1-q+l}} & 0 & 0 & \dots & 0 \\ \frac{1}{t_1 - t_{1-q+l}} & \frac{-1}{t_2 - t_{2-q+l}} & 0 & \dots & 0 \\ 0 & \frac{1}{t_2 - t_{2-q+l}} & \frac{-1}{t_3 - t_{3-q+l}} & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \frac{1}{t_{K+q-l} - t_K} \end{bmatrix}$$

for $1 \leq l \leq r \leq q-2$.

Let $B_q^{(r)}(x) = B_{q-r}(x)D^{(r)}$, implying

$$\hat{f}^{(r)}(x) = B_q^{(r)}(x)\left(G_{n,q} + \lambda_n \boldsymbol{P_m}\right)^{-1}\boldsymbol{B}^T\boldsymbol{y}/n$$

We use the identity $(A+B)^{-1} = A^{-1} - A^{-1}B(A+B)^{-1}$ to expand the inverse term in the estimator. This later allows us to split the bias term into the part due to approximating $f^{(r)}(x)$ with a spline (approximation bias) and the other part due to penalization (shrinkage bias).

$$\begin{aligned} (G_{n,q} + \lambda_n \boldsymbol{P_m})^{-1} &= G_{n,q}^{-1} - G_{n,q}^{-1}(\lambda_n \boldsymbol{P_m})H_n^{-1} \\ &= G_{n,q}^{-1} - H_n^{-1}(\lambda_n \boldsymbol{P_m})G_{n,q}^{-1} \end{aligned}$$

where the last equality is by symmetry.

Substituting into $\hat{f}^{(r)}(x)$, we have:

$$\begin{aligned} \implies \hat{f}^{(r)}(x) &= B_q^{(r)}(x)\left(G_{n,q}^{-1} - H_n^{-1}(\lambda_n \boldsymbol{P_m}) G_{n,q}^{-1}\right) \boldsymbol{B}^T \boldsymbol{y}/n \\ &= B_q^{(r)}(x) G_{n,q}^{-1} \boldsymbol{B}^T \boldsymbol{y}/n - B_q^{(r)}(x) H_n^{-1}(\lambda_n \boldsymbol{P_m}) G_{n,q}^{-1} \boldsymbol{B}^T \boldsymbol{y}/n \end{aligned}$$

We now focus on the bias of the naive estimator, $E\left[\hat{f}^{(r)}(x)\right] - f^{(r)}(x)$.

From Lemma 1, $\exists s_f \in \boldsymbol{\mathcal{S}}(q, \underline{\mathbf{t}})$, the space of spline functions of order q defined on knots $\underline{\mathbf{t}}$ such that $||f^{(r)} - s_f^{(r)}|| = O(h^{q-r}) + o(h^{p-r})$. The bias of $\hat{f}^{(r)}(x)$ can be written as:

$$E\left[\hat{f}^{(r)}(x)\right] - f^{(r)}(x) = \left[E\left(\hat{f}^{(r)}(x)\right) - s_f^{(r)}(x)\right] + \left[s_f^{(r)}(x) - f^{(r)}(x)\right] \tag{2.4}$$

Equation (2.4) above allows us to separately evaluate the approximation bias and shrinkage bias for estimating $f^{(r)}(x)$. Notice that Lemma 1 provides information on the rate of convergence on the second term in (2.4), we will next focus expressing the first term in a form that isolates the effect of penalization on the bias. Substituting the previously derived expression for $\hat{f}^{(r)}(x)$ into the first term, we have

$$\begin{aligned} E\hat{f}^{(r)}(x) - s_f^{(r)}(x) &= B_q^{(r)}(x) G_{n,q}^{-1} \boldsymbol{B}^T \boldsymbol{f}/n - s_f^{(r)}(x) \\ &\quad - B_q^{(r)}(x) H_n^{-1}(\lambda_n \boldsymbol{P_m}) G_{n,q}^{-1} \boldsymbol{B}^T \boldsymbol{f}/n \\ &= B_q^{(r)}(x) \underline{\boldsymbol{\gamma}} - s_f^{(r)}(x) - B_q^{(r)}(x) H_n^{-1}(\lambda_n \boldsymbol{P_m}) \underline{\boldsymbol{\gamma}} \end{aligned}$$

where $\underline{\boldsymbol{\gamma}} = G_{n,q}^{-1} \boldsymbol{B}^T \boldsymbol{f}/n$ and $\boldsymbol{f} = E[\boldsymbol{y}]$.

But $s_f^{(r)}(x) = B_q^{(r)}(x) G_{n,q}^{-1} \boldsymbol{B}^T \boldsymbol{s_f}/n$, where $\boldsymbol{s_f} = \{s_f(x_1), s_f(x_2), \ldots, s_f(x_n)\}$

$$
\begin{aligned}
\implies E\hat{f}^{(r)}(x) - s_f^{(r)}(x) &= B_q^{(r)}(x)\underline{\boldsymbol{\gamma}} - B_q^{(r)}(x)H_n^{-1}(\lambda_n \boldsymbol{P_m})\underline{\boldsymbol{\gamma}} \\
&\quad -B_q^{(r)}(x)G_{n,q}^{-1}\boldsymbol{B}^T\boldsymbol{s_f}/n \\
&= B_q^{(r)}(x)G_{n,q}^{-1}\boldsymbol{B}^T\boldsymbol{f}/n - B_q^{(r)}(x)G_{n,q}^{-1}\boldsymbol{B}^T\boldsymbol{s_f}/n \\
&\quad -B_q^{(r)}(x)H_n^{-1}(\lambda_n \boldsymbol{P_m})\underline{\boldsymbol{\gamma}} \\
&= B_q^{(r)}(x)G_{n,q}^{-1}\boldsymbol{B}^T(\boldsymbol{f} - \boldsymbol{s_f})/n - B_q^{(r)}(x)H_n^{-1}(\lambda_n \boldsymbol{P_m})\underline{\boldsymbol{\gamma}} \\
&= B_q^{(r)}(x)G_{n,q}^{-1}\underline{\boldsymbol{\alpha}} - B_q^{(r)}(x)H_n^{-1}(\lambda_n \boldsymbol{P_m})\underline{\boldsymbol{\gamma}} \qquad (2.5)
\end{aligned}
$$

where $\underline{\boldsymbol{\alpha}} = \boldsymbol{B}^T(\boldsymbol{f} - \boldsymbol{s_f})/n$

Let $Q(x)$ be a distribution of x on $[0,1]$ with positive continuous density $q(x)$. Then substituting (2.5) into (2.4) and using the triangle inequality, we can evaluate the squared bias of $\hat{f}^{(r)}(x)$ as:

$$
\begin{aligned}
\frac{1}{3}\int_0^1 \left(E\left[\hat{f}^{(r)}(x)\right] - f^{(r)}(x)\right)^2 q(x)dx &\leq \int_0^1 \left(s_f^{(r)}(x) - f^{(r)}(x)\right)^2 q(x)dx \\
&\quad +\underline{\boldsymbol{\alpha}}^T G_{n,q}^{-1} G_q^{(r)} G_{n,q}^{-1}\underline{\boldsymbol{\alpha}} \qquad (2.6) \\
&\quad +\underline{\boldsymbol{\gamma}}^T(\lambda_n \boldsymbol{P_m})H_n^{-1}G_q^{(r)}H_n^{-1}(\lambda_n \boldsymbol{P_m})\underline{\boldsymbol{\gamma}}
\end{aligned}
$$

where $G_q^{(r)} = \int_0^1 B_q^{(r)T}(x)B_q^{(r)}(x)q(x)dx$. The first and second terms in (2.6) represent the part of the bias due to using spline functions to estimate $f^{(r)}(x)$, and the last term represents the part of the bias due to penalization.

Observe that, by Lemma 1,

$$
\begin{aligned}
\int_0^1 \left(s_f^{(r)}(x) - f^{(r)}(x)\right)^2 q(x)dx &\leq q_{\max}\int_0^1 \left(s_f^{(r)}(x) - f^{(r)}(x)\right)^2 dx \\
&= O\left(h^{2(q-r)}\right) + o\left(h^{2(p-r)}\right)
\end{aligned}
$$

where, $q_{\max} = \max_{0\leq x\leq 1} q(x) < \infty$.

For the second term in (2.6), we use the result $||G_q^{(r)}||_\infty = O(h^{-2r})$, from Lemma 2. We also use $||G_{n,q}^{-1}||_\infty = O(h^{-1})$ from Lemma 3 and Lemma 6.10 of [1] that $||\underline{\boldsymbol{\alpha}}||_{\max} = o(h^{p+1})$.

Let $G_q^{(r)\frac{1}{2}}$ be a square and symmetric matrix such that $G_q^{(r)} = G_q^{(r)\frac{1}{2}} G_q^{(r)\frac{1}{2}}$.

We write

$$\begin{aligned}
\underline{\boldsymbol{\alpha}}^T G_{n,q}^{-1} G_q^{(r)} G_{n,q}^{-1} \underline{\boldsymbol{\alpha}} &= \left(G_q^{(r)\frac{1}{2}} G_{n,q}^{-1} \underline{\boldsymbol{\alpha}} \right)^T \left(G_q^{(r)\frac{1}{2}} G_{n,q}^{-1} \underline{\boldsymbol{\alpha}} \right) \\
&= ||G_q^{(r)\frac{1}{2}} G_{n,q}^{-1} \underline{\boldsymbol{\alpha}}||_2^2 \\
&= ||G_q^{(r)\frac{1}{2}} G_{n,q}^{-1}||_2^2 ||\underline{\boldsymbol{\alpha}}||_2^2
\end{aligned}$$

Using the fact that for a real matrix A, $||A||_2^2 = \rho(A^T A) \leq ||A^T A||_\infty$, here, $\rho(A^T A)$ is the largest eigen value of $A^T A$, we write:

$$\begin{aligned}
||G_q^{(r)\frac{1}{2}} G_{n,q}^{-1}||_2^2 &\leq ||G_{n,q}^{-1} G_q^{(r)\frac{1}{2}} G_q^{(r)\frac{1}{2}} G_{n,q}^{-1}||_\infty \\
&= ||G_{n,q}^{-1} G_q^{(r)} G_{n,q}^{-1}||_\infty \\
&\leq ||G_{n,q}^{-1}||_\infty ||G_q^{(r)}||_\infty ||G_{n,q}^{-1}||_\infty \\
&= O(h^{-1}) O(h^{-2r}) O(h^{-1})
\end{aligned}$$

Also, from $||\underline{\boldsymbol{\alpha}}||_{\max} = o(h^{p+1})$, we have:

$$\begin{aligned}
\underline{\boldsymbol{\alpha}}^T G_{n,q}^{-1} G_q^{(r)} G_{n,q}^{-1} \underline{\boldsymbol{\alpha}} &= O(h^{-1}) O(h^{-2r}) O(h^{-1}) o(h^{2(p+1)}) \\
&= o(h^{2p-2r})
\end{aligned}$$

Next, we focus on the part of the bias due to penalization as given by the third term in (2.6). First, note that from [12] and Lemma 5.2 of [60], $D^{(r)}$ in $B_q^{(r)}(x) = B_{q-r}(x) D^{(r)}$, is such that

$$||D^{(r)}||_\infty = O(h^{-r})$$

This can be easily seen by inspecting the elements of $D^{(r)}$.

$$\therefore B_q^{(r)T}(x) B_q^{(r)}(x) = D^{(r)T} B_{q-r}^T(x) B_{q-r}(x) D^{(r)} = O(h^{-2r}) B_{q-r}^T(x) B_{q-r}(x)$$

Thus, we can write

$$\begin{aligned} G_q^{(r)} &= \int_0^1 B_q^{(r)T}(x) B_q^{(r)}(x) q(x) dx \\ &= O(h^{-2r}) \int_0^1 B_{q-r}^T(x) B_{q-r}(x) q(x) dx \\ &= O(h^{-2r}) G_{q-r} \end{aligned}$$

where $G_{q-r} = \int_0^1 B_{q-r}^T(x) B_{q-r}(x) q(x) dx$.

Also, by the WLLN, $G_{n,q-r} = G_{q-r} + o(1)$.

Therefore:

$$\begin{aligned} \underline{\boldsymbol{\gamma}}^T(\lambda_n \boldsymbol{P_m}) H_n^{-1} G_q^{(r)} H_n^{-1}(\lambda_n \boldsymbol{P_m})\underline{\boldsymbol{\gamma}} &= O(h^{-2r}) \underline{\boldsymbol{\gamma}}^T(\lambda_n \boldsymbol{P_m}) H_n^{-1} \\ &\quad \times G_{q-r} H_n^{-1}(\lambda_n \boldsymbol{P_m})\underline{\boldsymbol{\gamma}} \\ &= O(h^{-2r}) \underline{\boldsymbol{\gamma}}^T(\lambda_n \boldsymbol{P_m}) H_n^{-1} \\ &\quad \times G_{n,q-r} H_n^{-1}(\lambda_n \boldsymbol{P_m})\underline{\boldsymbol{\gamma}} \end{aligned}$$

where $G_{n,q-r} = B_{q-r}^T B_{q-r}/n$, the version of $G_{n,q}$ based on B-splines of order $q-r$. Note that the decay of the eigenvalues of $G_{n,q}$ does not depend on q (see Lemma 3). Therefore, we will use $G_{n,q}$ instead of $G_{n,q-r}$ in the derivations that follow for asymptotic order. This simplifies the calculations since H_n^{-1} depends on $G_{n,q}$.

From

$$\begin{aligned} H_n^{-1} &= [G_{n,q} + (\lambda_n \boldsymbol{P_m})]^{-1} \\ &= \left[G_{n,q}^{\frac{1}{2}} \left(G_{n,q}^{\frac{1}{2}} + \lambda_n G_{n,q}^{-\frac{1}{2}} \boldsymbol{P_m}\right)\right]^{-1} \\ &= \left(G_{n,q}^{\frac{1}{2}} + \lambda_n G_{n,q}^{-\frac{1}{2}} \boldsymbol{P_m}\right)^{-1} G_{n,q}^{-\frac{1}{2}} \end{aligned}$$

we can write

$$
\begin{aligned}
(\lambda_n \boldsymbol{P_m}) H_n^{-1} G_{n,q} H_n^{-1} (\lambda_n \boldsymbol{P_m}) &= (\lambda_n \boldsymbol{P_m}) \left(G_{n,q}^{\frac{1}{2}} + \lambda_n G_{n,q}^{-\frac{1}{2}} \boldsymbol{P_m} \right)^{-1} G_{n,q}^{-\frac{1}{2}} G_{n,q} \\
&\quad \times \left(G_{n,q}^{\frac{1}{2}} + \lambda_n G_{n,q}^{-\frac{1}{2}} \boldsymbol{P_m} \right)^{-1} G_{n,q}^{-\frac{1}{2}} (\lambda_n \boldsymbol{P_m})
\end{aligned}
$$

Let $\tilde{P} = G_{n,q}^{-\frac{1}{2}} (\lambda_n \boldsymbol{P_m}) G_{n,q}^{-\frac{1}{2}} \implies \tilde{P} G_{n,q}^{\frac{1}{2}} = G_{n,q}^{-\frac{1}{2}} (\lambda_n \boldsymbol{P_m})$

Substituting into (2.7), we have

$$
\begin{aligned}
(\lambda_n \boldsymbol{P_m}) H_n^{-1} G_{n,q} H_n^{-1} (\lambda_n \boldsymbol{P_m}) &= (\lambda_n \boldsymbol{P_m}) \left(G_{n,q}^{\frac{1}{2}} + \tilde{P} G_{n,q}^{\frac{1}{2}} \right)^{-1} G_{n,q}^{\frac{1}{2}} \\
&\quad \times \left(G_{n,q}^{\frac{1}{2}} + \tilde{P} G_{n,q}^{\frac{1}{2}} \right)^{-1} \tilde{P} G_{n,q}^{\frac{1}{2}} \\
&= (\lambda_n \boldsymbol{P_m}) G_{n,q}^{-\frac{1}{2}} \left(I + \tilde{P} \right)^{-1} G_{n,q}^{\frac{1}{2}} G_{n,q}^{-\frac{1}{2}} \\
&\quad \times \left(I + \tilde{P} \right)^{-1} \tilde{P} G_{n,q}^{\frac{1}{2}} \\
&= G_{n,q}^{\frac{1}{2}} \tilde{P} (I + \tilde{P})^{-2} \tilde{P} G_{n,q}^{\frac{1}{2}}
\end{aligned}
$$

where in the second equality, we have used the fact that $G_{n,q}^{\frac{1}{2}} + \tilde{P} G_{n,q}^{\frac{1}{2}} = (I + \tilde{P}) G_{n,q}^{\frac{1}{2}}$ and that $(\lambda_n \boldsymbol{P_m}) G_{n,q}^{-\frac{1}{2}} = G_{n,q}^{\frac{1}{2}} \tilde{P}$ in the last equality.

Using the above, we can then write:

$$
\underline{\boldsymbol{\gamma}}^T (\lambda_n \boldsymbol{P_m}) H_n^{-1} G_{n,q} H_n^{-1} (\lambda_n \boldsymbol{P_m}) \underline{\boldsymbol{\gamma}} = \underline{\boldsymbol{\gamma}}^T G_{n,q}^{\frac{1}{2}} \tilde{P} \left(I + \tilde{P} \right)^{-2} \tilde{P} G_{n,q}^{\frac{1}{2}} \underline{\boldsymbol{\gamma}}
$$

It follows from the symmetry of $\tilde{P}$ that $||\tilde{P}||_2 \tilde{P} - \tilde{P} \left(I + \tilde{P} \right)^{-2} \tilde{P}$ and $\tilde{P} - \tilde{P}(I + \tilde{P})^{-2} \tilde{P}$ are positive semidefinite.

First, for $||\tilde{P}||_2 \tilde{P} - \tilde{P} \left(I + \tilde{P} \right)^{-2} \tilde{P}$ positive semidefinite, we have

$$
\begin{aligned}
\underline{\boldsymbol{\gamma}}^T (\lambda_n \boldsymbol{P_m}) H_n^{-1} G_q^{(r)} H_n^{-1} (\lambda_n \boldsymbol{P_m}) \underline{\boldsymbol{\gamma}} &= O(h^{-2r}) ||\tilde{P}||_2 \underline{\boldsymbol{\gamma}}^T G_{n,q}^{\frac{1}{2}} \tilde{P} G_{n,q}^{\frac{1}{2}} \underline{\boldsymbol{\gamma}} \\
&= O(h^{-2r}) ||\tilde{P}||_2 \underline{\boldsymbol{\gamma}}^T (\lambda_n \boldsymbol{P_m}) \underline{\boldsymbol{\gamma}} \\
&= O(h^{-2r}) ||G_{n,q}^{-\frac{1}{2}} (\lambda_n \boldsymbol{P_m}) G_{n,q}^{-\frac{1}{2}}||_2 \underline{\boldsymbol{\gamma}}^T (\lambda_n \boldsymbol{P_m}) \underline{\boldsymbol{\gamma}} \\
&= O(h^{-2r}) ||G_{n,q}^{-1}||_2 ||(\lambda_n \boldsymbol{P_m})||_2 \underline{\boldsymbol{\gamma}}^T (\lambda_n \boldsymbol{P_m}) \underline{\boldsymbol{\gamma}}
\end{aligned}
$$

where we have used $G_{n,q}^{\frac{1}{2}}\tilde{P}G_{n,q}^{\frac{1}{2}} = (\lambda_n \boldsymbol{P_m})$ in the second equality and substituted $\tilde{P}$ in the third.

By Assumption 5, $||\boldsymbol{P_m}||_2 = O(h^{1-2m})$ and from Lemma 4, $\underline{\boldsymbol{\gamma}}^T \boldsymbol{P_m} \underline{\boldsymbol{\gamma}} = O(1)$. Therefore:

$$\begin{aligned} \underline{\boldsymbol{\gamma}}^T(\lambda_n \boldsymbol{P_m})H_n^{-1}G_q^{(r)}H_n^{-1}(\lambda_n \boldsymbol{P_m})\underline{\boldsymbol{\gamma}} &= O(h^{-2r})O(h^{-1})O(\lambda_n h^{1-2m})O(\lambda_n) \\ &= O(\lambda_n^2 h^{-2m-2r}). \end{aligned}$$

Also, $\tilde{P} - \tilde{P}(I+\tilde{P})^{-2}\tilde{P}$ positive semidefinite, we have

$$\begin{aligned} \underline{\boldsymbol{\gamma}}^T(\lambda_n \boldsymbol{P_m})H_n^{-1}G_q^{(r)}H_n^{-1}(\lambda_n \boldsymbol{P_m})\underline{\boldsymbol{\gamma}} &= O(h^{-2r})\underline{\boldsymbol{\gamma}}^T G_{n,q}^{\frac{1}{2}}\tilde{P}(I+\tilde{P})^{-2}\tilde{P}G_{n,q}^{\frac{1}{2}}\underline{\boldsymbol{\gamma}} \\ &= O(h^{-2r})\underline{\boldsymbol{\gamma}}^T G_{n,q}^{\frac{1}{2}}\tilde{P}G_{n,q}^{\frac{1}{2}}\underline{\boldsymbol{\gamma}} \\ &= O(h^{-2r})\underline{\boldsymbol{\gamma}}^T(\lambda_n \boldsymbol{P_m})\underline{\boldsymbol{\gamma}} \\ &= O(\lambda_n h^{-2r}) \end{aligned}$$

$$\therefore \underline{\boldsymbol{\gamma}}^T(\lambda_n \boldsymbol{P_m})H_n^{-1}G_q^{(r)}H_n^{-1}(\lambda_n \boldsymbol{P_m})\underline{\boldsymbol{\gamma}} = O\left\{\min\left(\lambda_n^2 h^{-2m-2r}, \lambda_n h^{-2r}\right)\right\}$$

This concludes the proof for bias in (2.6).

Next, we look at the variance part:

$$\begin{aligned} Var(\hat{f}^{(r)}(x)) &= Var\left(B_q^{(r)}(x)H_n^{-1}\boldsymbol{B}^T\boldsymbol{y}/n\right) \\ &= B_q^{(r)}(x)H_n^{-1}\boldsymbol{B}^T Var(\boldsymbol{y}/n)\boldsymbol{B}H_n^{-1}B_q^{(r)T}(x) \\ &= \frac{\sigma^2}{n}tr\left\{B_q^{(r)}(x)H_n^{-1}(\boldsymbol{B}^T\boldsymbol{B}/n)H_n^{-1}B_q^{(r)T}(x)\right\} \\ &= \frac{\sigma^2}{n}tr\left\{B_q^{(r)}(x)H_n^{-1}G_{n,q}H_n^{-1}B_q^{(r)T}(x)\right\} \\ &= \frac{\sigma^2}{n}tr\left\{H_n^{-1}G_{n,q}H_n^{-1}B_q^{(r)T}(x)B_q^{(r)}(x)\right\} \end{aligned}$$

Note that we have used the rotation property of the trace in the last equality.

Therefore,

$$\begin{aligned}\int_0^1 Var(\hat{f}^{(r)}(x))q(x)dx &= \frac{\sigma^2}{n} tr\left\{H_n^{-1}G_{n,q}H_n^{-1}G_q^{(r)}\right\} \\ &= O(h^{-2r})\frac{\sigma^2}{n} tr\left\{H_n^{-1}G_{n,q}H_n^{-1}G_{n,q}\right\}\end{aligned}$$

where in the last equality, we have used the fact that $G_q^{(r)} = O(h^{-2r})G_{q-r}$ and that the decay rates of $G_{n,q}$ do not depend on q.

From

$$\begin{aligned}H_n^{-1} &= (G_{n,q} + (\lambda_n \boldsymbol{P_m}))^{-1} \\ &= \left[G_{n,q}\left(I + G_{n,q}^{-1}(\lambda_n \boldsymbol{P_m})\right)\right]^{-1} \\ &= \left[I + G_{n,q}^{-1}(\lambda_n \boldsymbol{P_m})\right]^{-1} G_{n,q}^{-1}\end{aligned}$$

$$\implies H_n^{-1}G_{n,q} = \left[I + G_{n,q}^{-1}(\lambda_n \boldsymbol{P_m})\right]^{-1}.$$

Note that $G_{n,q}^{-1}(\lambda_n \boldsymbol{P_m}) = G_{n,q}^{-\frac{1}{2}}G_{n,q}^{-\frac{1}{2}}(\lambda_n \boldsymbol{P_m})$ and by the rotation property of the trace,

$$\begin{aligned}tr\left[G_{n,q}^{-1}(\lambda_n \boldsymbol{P_m})\right] &= tr\left[G_{n,q}^{-\frac{1}{2}}G_{n,q}^{-\frac{1}{2}}(\lambda_n \boldsymbol{P_m})\right] \\ &= tr\left[G_{n,q}^{-\frac{1}{2}}(\lambda_n \boldsymbol{P_m})G_{n,q}^{-\frac{1}{2}}\right] \\ &= tr\left[\tilde{P}\right]\end{aligned}$$

$$
\begin{aligned}
\therefore \int_0^1 Var(\hat{f}^{(r)}(x))q(x)dx &= O(h^{-2r})\frac{\sigma^2}{n} tr\left[(I+\tilde{P})^{-2}\right] \\
&= O(h^{-2r})\frac{\sigma^2}{n}\left|\left|(I+\tilde{P})^{-2}\right|\right|_F^2 \\
&= O(h^{-2r})\frac{\sigma^2}{n} O\left\{\frac{1}{\max(h,\lambda_n^{1/2m})}\right\} \\
&= O(h^{-2r})\frac{\sigma^2}{n} O\left\{\min(h^{-1},\lambda_n^{-1/2m})\right\} \\
&= O(K^{2r})\frac{\sigma^2}{n} O\left\{\min(K,\lambda_n^{-1/2m})\right\} \\
&= O\left(\frac{K_e}{n}\right)
\end{aligned}
$$

Where in the above, we have used $\left|\left|(I+\tilde{P})^{-2}\right|\right|_F^2 = O\left\{\frac{1}{\max(h,\lambda^{1/2m})}\right\}$ from Lemma 5.2 of [58], $K \sim h^{-1}$, and $K_e = \min\left\{K^{2r+1}, K^{2r}\lambda_n^{-1/2m}\right\}$

This completes the proof of the theorem.

2.4 Simulations

2.4.1 Overview

In this section, we present a simulation to assess the naive estimator's rate of convergence and its finite-sample performance. The simulation is divided into three parts. The first part examines the L_2 rates of convergence of the naive estimator when GCV and REML are used to choose the smoothing parameter. The second part of this section focuses on the finite sample performance of the naive estimator. We compared it to an "oracle" method that uses knowledge of the true function (or derivatives) to choose the optimal smoothing parameter. That "oracle" method is not a practical estimator, but it provides an upper bound benchmark for P-spline performance. Finally, the third part of this section compares the naive method to other derivative estimation methods in the literature.

Except where noted, we use the same mean regression function f as [14]. We simulated data $\{x_i, y_i\}_{i=1}^n$ from the model:

$$Y_i = f(x_i) + \varepsilon_i, \ \forall \ 1 \leq i \leq n$$

where x_i's are a grid over $\mathcal{K} = [0, 1]$, ε_i's are iid with $\varepsilon_i \sim N(0, \sigma^2 = 0.1^2)$ and

$$f(x) = 32e^{-8(1-2x)^2}(1 - 2x) \tag{2.7}$$

Figure 2.1 shows the mean regression function in (2.7) and its first two derivatives. We use a range of sample sizes as shown in the results.

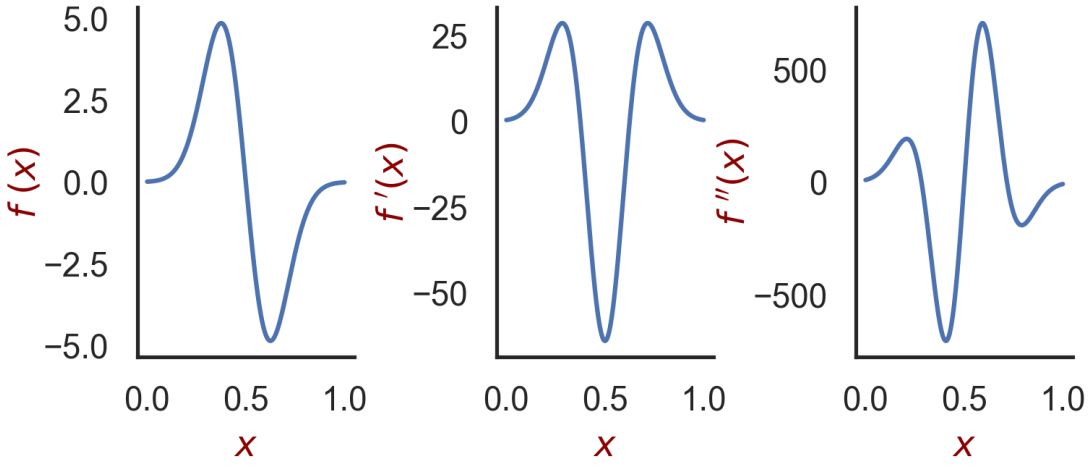

Figure 2.1: Mean regression function with its first two derivatives.

As discussed in [58], [8], and our Remark (1), the asymptotics of the penalized spline estimator are similar to those of Regression Splines (small K scenario) or Smoothing Splines (large K scenario), depending on the rate at which the number of knots, K, increases with the sample size, n. In our simulation, we considered these two scenarios: when K increases slowly with n, and when K increases at a faster rate with n. For the slow K scenario we use $K \sim n^{\frac{1}{2p+1}}$, and K in the fast scenario is chosen such that $K \geq C^{1/p}\lambda_n^{-1/2p}$ for some large constant, C.

We investigated the L_2 rate of convergence for the first two derivatives of the mean regression function in (2.7) using a P-Spline with 2^{nd} ($m = 2$) order penalty ([16]). Note that with $m = 2$, the equivalent kernel methodology ([42], Lemma 9.13 of [59]) implies that the assumed differentiability of f is $p = 2m = 4$.

[45] provided optimal rates of convergence for non-parametric regression estimators. The optimal rate of convergence for a non-parametric estimator of the r^{th} derivative of $g : \mathbb{R}^d \to \mathbb{R}$ where $g \in \mathcal{C}^p$ is given by $n^{-\frac{p-r}{2p+d}}$, in our simulations, we have the optimal L_2 rate of convergence for estimating the r^{th} derivative of f as:

$$n^{-\frac{p-r}{2p+d}} = n^{-\frac{4-r}{2\times 4+1}} = n^{-\frac{1}{9}(4-r)}$$

2.4.1.1 L_2 Convergence of the Naive Estimator

Figure 2.2 illustrates the L_2 rate of the naive estimator when the smoothing parameter λ_n is chosen by the GCV approach. The naive estimator achieves the optimal L_2 rates of convergence for the mean regression function and its first two derivatives when GCV is used to choose the smoothing parameter, but it is slightly slower for REML. This deviation from the optimal rate using REML appears to worsen for higher derivatives. Also, we observed that the fast K scenario was overall slightly slower than the slow K scenario for REML. These results agree with known results in the literature for smoothing splines when estimating the mean regression function. For instance, [10] showed that GCV achieves the optimal rate of convergence when used to choose the smoothing parameter in smoothing splines. However, [47] found that Maximum Likelihood (ML) based methods may be slower than GCV for sufficiently smooth functions.

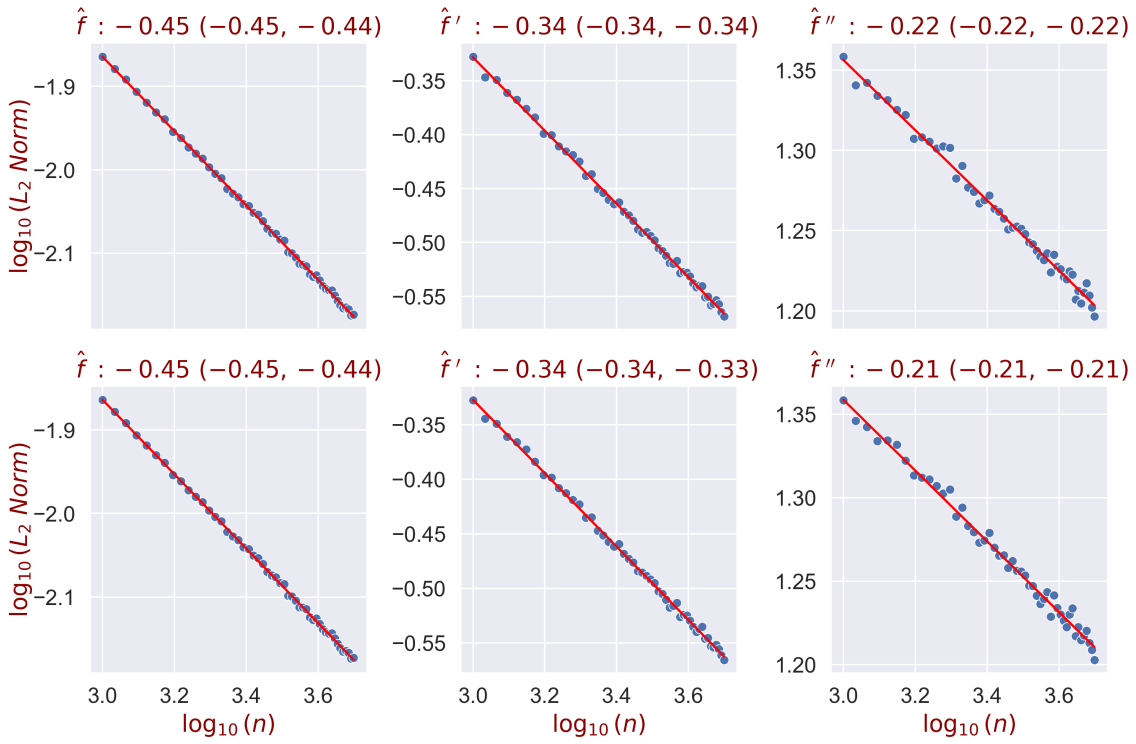

Figure 2.2: L_2 convergence rates for f and its first two derivatives under two scenarios for increasing K with n. The top three figures show results for the slowly increasing K scenario while the bottom three show results for the fast increasing K scenario. The smoothing parameter λ_n is chosen by the GCV method.

Table 2.1 below summarizes the rates of convergence of the naive estimator for estimating derivatives of the mean regression function in (2.7) under the various scenarios of the number of knots K as n increases.

λ_n **Method**	**Target**	**Optimal** L_2 **Rate**	**Slow** K	**Fast** K
GCV	f	-0.44	$-0.45(-0.45, -0.44)$	$-0.45(-0.45, -0.44)$
	f'	-0.33	$-0.34(-0.34, -0.34)$	$-0.34(-0.34, -0.33)$
	f''	-0.22	$-0.22(-0.22, -0.22)$	$-0.21(-0.21, -0.21)$
REML	f	-0.44	$-0.44(-0.44, -0.43)$	$-0.43(-0.44, -0.43)$
	f'	-0.33	$-0.32(-0.32, -0.31)$	$-0.31(-0.32, -0.31)$
	f''	-0.22	$-0.19(-0.19, -0.18)$	$-0.18(-0.18, -0.17)$

Table 2.1: Summary of L_2 rates of convergence for estimating the mean regression function in (2.7) and its first two derivatives.

2.4.2 Finite sample performance of naive estimator.

In this section we compare the naive estimator to an "oracle" method that uses knowledge of the true form of the target (mean regression function or its derivatives) to choose the optimal amount of smoothing, which we did with a grid search. While this "oracle" is not an estimator, it shows the minimum loss when estimating the function in question with a penalized spline. GCV was used to choose the appropriate smoothing parameter for the various spline-based estimators in what follows.

In Figure 2.3 below, we show that the naive estimator corresponds to the median MSE in the Monte Carlo experiment. To summarize, we see that the naive estimator appears to accurately estimate both the true mean regression function (f) and its first derivative (f'). However, we observe some lack of fit around the boundaries of the second derivative, (f'').

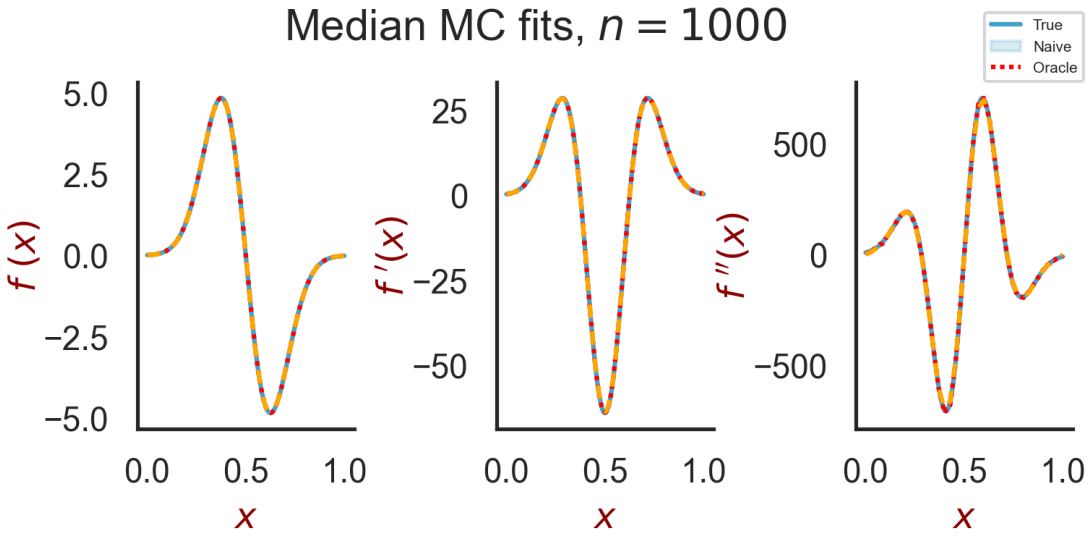

Figure 2.3: Median Monte-Carlo fits of the mean regression function in (2.7) with its first two derivatives using the naive and oracle estimators.

Next, Figure 2.4 compares the naive and oracle methods for the mean regression function in (2.7) and its first two derivatives across the two increasing K scenarios. Overall, in comparison to the oracle method, the naive estimator's finite sample

performance degrades with increasing derivatives, with an average error difference (logarithmic scale) of about 0.5% for the mean regression function, 17% for its first derivative, and 29% for its second derivative. While the naive penalized spline derivative estimator is shown to converge at the optimal L_2 rate of convergence (Theorem 1), it may also have higher mean squared error in finite samples, especially for higher derivatives. This is largely driven by a higher variance of the naive estimator, compared with the oracle method. We note that the results summarized in Figure 2.4 are similar for the two increasing K scenarios.

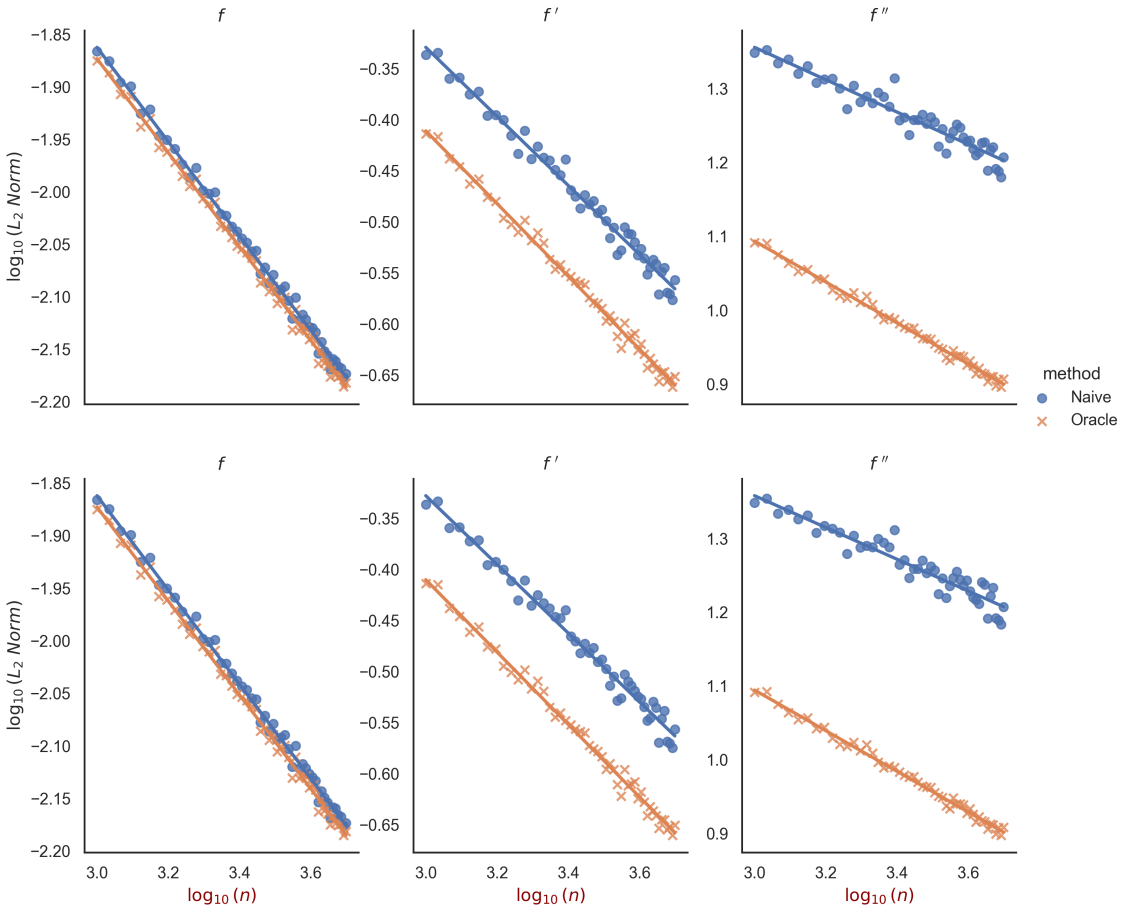

Figure 2.4: L_2 convergence rates for f and its first two derivatives with two scenarios for increasing K with n and how they compare with their corresponding oracle estimators. Figures in the top row show results for slowly increasing K scenario while figures in the bottom row show results for the fast increasing K scenario. The smoothing parameter λ_n is chosen by the GCV method.

2.4.3 Comparison with other methods

In this section, we compare the finite sample MSE of the naive estimator to other derivative estimation methods in the literature. We considered the adaptive penalty penalized spline estimator by [43]. We also used the linear combination method of [11], but it consistently had higher MSE values and results are not shown.

We evaluated the methods using three mean regression functions from the literature ([14, 11]). As proxies for low, medium, and high noise scenarios, we considered noise levels that were 10 percent, 30 percent, and 60 percent of the range of each function. This was to understand how the methods compare at different levels of noise. The following are the three functions considered:

$$f_1(x) = \sin^2(2\pi x) + \log(4/3 + x) \quad \text{for} \quad x \in [-1, 1],$$

$$f_2(x) = 32e^{-8(1-2x)^2}(1 - 2x) \quad \text{for} \quad x \in [0, 1],$$

and the doppler function

$$f_3(x) = \sqrt{x(1-x)}\sin\left(\frac{2.1\pi}{x + 0.05}\right) \quad \text{for} \quad x \in [0.25, 1].$$

Figure 2.5 below shows the results for estimating the first (panel a) and second (panel b) derivatives of the three mean regression functions across the three noise levels. These results indicate that the adaptive penalty (S) methods and the naive (N) method often perform similarly, depending on the form of the function, the noise level, and the order of the derivative. We also note that the adaptive penalty method sometimes performs better than the oracle method (O). This is possible since the oracle method only finds the best P-splines estimate based on a single smoothing penalty.

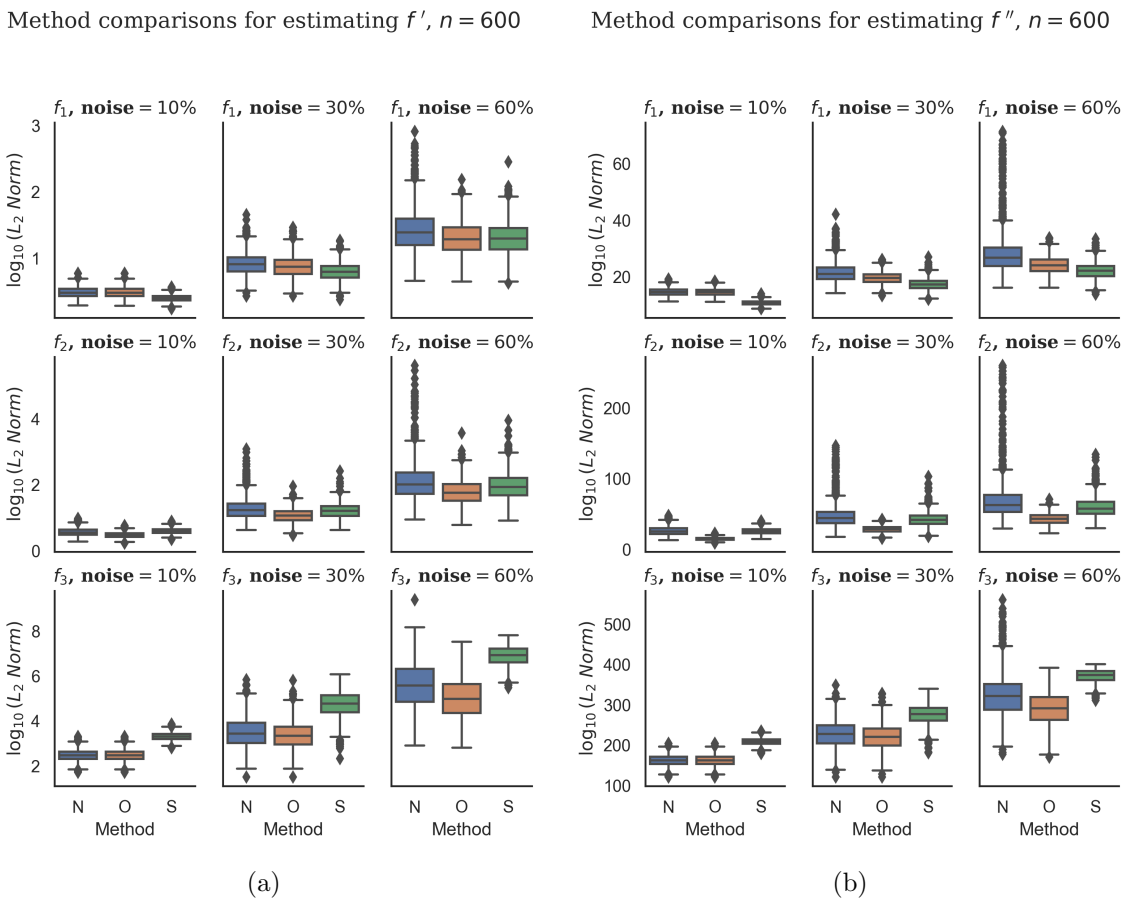

Figure 2.5: Comparing derivative estimation methods in reference to the oracle estimator across different functions and noise levels. Panel (a) shows results for estimating first derivatives of the mean regression functions f_1, f_2 and f_3 while Panel (b) shows results for estimating second derivatives. The first row is for f_1, second row is for f_2 and the last row is for f_3.

2.5 Discussion

We have shown that the naive penalized spline estimator of the r^{th} derivative of the mean regression function achieves the optimal L_2 rate of convergence ([45]) under standard assumptions on knot placement and the penalty matrix. This builds on the work by [58] which derived the L_2 rate of convergence for estimating the mean regression function. As stated in Remark 1 and noted by others ([8, 58]), the rate at which the number of knots, K, increases with n gives rise to two scenarios: the

fast K scenario, which is similar to smoothing spline asymptotics, and the slow K scenario, which is similar to regression spline asymptotics.

Using simulations, we investigated how two prevalent methods for choosing the smoothing parameter (GCV and REML) affect the L_2 convergence of the naive estimator. We found that, for both slow and fast K scenarios, the naive estimator achieves the optimal L_2 rate of convergence when GCV is used. For REML, the estimator did not quite achieve the optimal rate.

To access the finite sample performance of the naive estimator, we compared the MSEs of the estimator with an "oracle" method that uses information about the true function to be estimated to choose the P-spline's smoothing parameter. We found that, in finite samples, the naive estimator may have noticeably larger mean squared errors, especially for higher derivatives, but the estimates can still be quite visually similar. We found that the adaptive penalty penalized spline estimator by [43] performed similarly to the naive estimator.

CHAPTER 3

NONPARAMETRIC APPROACH TO ESTIMATING THE ASSOCIATION BETWEEN OBJECTIVELY MEASURED PHYSICAL ACTIVITY AND MORTALITY

3.1 Introduction

Physical activity is widely recognized as an important contributor to overall health and longevity. Several studies have shown that regular physical activity can lower the risk of chronic diseases such as heart disease, stroke, and diabetes, as well as reduce the risk of all-cause mortality ([27, 50, 5]), but the precise dose-response relationship between physical activity and sedentary behaviors remains largely unknown ([34]).

Understanding the dose-response relationship is complicated by the fact that physical activity is also part of an inherently compositional (e.g. [2]) structure. For instance, the total time spent in various levels of physical activity, sedentary behavior, and sleep on any given day is 24 hours, and that results in perfect multicollinearity among those covariates. Adapting ideas from nutritional epidemiology (e.g. [52]), [28] developed a method called isotemporal substitution regression. In that method, one of the components in the 24-hour composition is dropped from the regression, and the remaining coefficients estimate the expected effect of increasing that covariate by one unit while decreasing the one dropped from the model by one unit, i.e. a substitution effect.

Using data from the NHANES 2003-2006 survey and the corresponding mortality follow-up through December 31st, 2011, [22] used an isotemporal substitution Cox proportional hazard model ([9]) to estimate the effect of substituting one level of physical activity for another while adjusting for baseline covariates. To objectively

measure physical activity in the NHANES survey, participants were provided with physical activity monitoring devices and were instructed to wear them throughout the day, excluding periods of sleep or when engaging in activities where the monitor would get wet. The NHANES 2003-2006 survey data captured the time spent engaging in physical activity, further classified into light and moderate-to-vigorous activity (MVPA), as well as sedentary behavior on a given day. Any remaining hours were considered sleep or non-wear time.

In this chapter, we extend the analysis in [22] in two ways: our primary contribution is to develop a novel nonparametric version of the isotemporal substitution model. More generally, this is a novel nonparametric model that is appropriate when the covariates are compositions. In addition to proportional hazards, our model only assumes a smooth nonparametric form for the dependence of the hazard function on the composition of physical activity.

Compared with the linear Cox model, our model is more robust to misspecification and can capture different substitution effects at different levels of activity. For instance, our model can capture different mortality associations of going from three to four hours of sedentary time versus going from nine to ten hours. We also extend the analysis in [22] by using mortality follow-up through December 31st, 2019 ([31]) which gives us almost an extra decade of mortality follow-up data which may result in increased statistical power and improved parameter estimates.

To facilitate our analysis, we represented the three-dimensional physical activity data in a trilinear coordinate system that was transformed into a Cartesian coordinate system to enable visualization of the three-dimensional surface in two dimensions. We then fitted a smooth tensor product spline to the sedentary and sleep & non-wear time variables, with activity time serving as the baseline in the isotemporal substitution regression. With this approach, the effects of substituting one activity level for another are given by partial derivatives of the fitted surface and by applying

the chain rule we derived the effects on the original scale of the data. In addition, baseline covariates such as age, gender, and the presence of chronic diseases were adjusted for in the Cox model as linear terms.

3.2 Methods

3.2.1 Statistical Model

Our methods address time-to-event data where at least some of the covariates describe the distribution of a three-dimensional composition for each observation. The data are independent across observations and consist of $\{t_i, \delta_i, w_i, \underline{\boldsymbol{z}}_i, x_{i1}, x_{i2}, x_{i3}\}_{i=1}^{n}$ where t_i is the observed event time or follow-up time for censored individuals, δ_i is a binary indicator where $\delta_i = 1$ means the participant experienced the event and, w_i is a survey weight, $\underline{\boldsymbol{z}}_i$ is a $p-$dimensional covariate vector (full rank over the observations), and x_{i1}, x_{i2}, x_{i3} are compositional covariates. In our motivating example, $x_{i1} + x_{i2} + x_{i3} = 24$ for all i, and $x_{ij} \in [0, 24]$ for all i and j. We note that our model readily generalizes to m-dimensional compositions, but estimation becomes more difficult since it involves fitting an $(m-1)$ dimensional nonparametric function.

We use the following hazard function for the Cox model: $h(t; x_{i1}, x_{i2}, x_{i3}, \underline{\boldsymbol{z}}_i) = h_0(t) \exp\left\{ f(x_{i1}, x_{i2}, x_{i3}) + \underline{\boldsymbol{z}}_i^\top \underline{\boldsymbol{\alpha}} \right\}$, where $h_0(t)$ is the unknown baseline hazard function, and $f(.,.,.)$ is an unknown smooth function defined for values of x_{i1}, x_{i2}, x_{i3} that satisfy the two constraints on x_{ij} above. Without loss of generality, we use the fact that the x_{ij}s sum to a constant number for each i and let f be a function of only the first two arguments. That is, $f(x_{i1}, x_{i2}, x_{i3}) = g(x_{i1}, x_{i2})$ for some smooth function g, and the hazard function model is:

$$h(t; x_{i1}, x_{i2}, x_{i3}, \underline{\boldsymbol{z}}_i) = h_0(t) \exp\left\{ g(x_{i1}, x_{i2}) + \underline{\boldsymbol{z}}_i^\top \underline{\boldsymbol{\alpha}} \right\}. \tag{3.1}$$

In order to facilitate model fitting, we represent the compositional covariates with a trilinear coordinate system which we describe in the next section. We also note that linear compositional Cox models such as the ones used in [22] are special cases of our model.

3.2.2 Trilinear Coordinate System

A trilinear coordinate system ([2]) represents a given point $P = (x_1, x_2, x_3)$ as the intersection of lines perpendicular to the relative sides of each coordinate in an equilateral triangle. An example is given below in Figure 3.1. In order to estimate g, the effects of the compositional covariates of the log hazard ratio in our Cox model, we embed the trilinear coordinate system into a two-dimensional Cartesian coordinate system. With this approach, $g(x_1, x_2) = m\{x_1^*(x_1, x_2), x_2^*(x_1, x_2)\}$ using the transformed coordinates $\{x_1^*(x_1, x_2), x_2^*(x_1, x_2)\}$. Using trigonometry, it can be shown that for a given trilinear point $P = (x_1, x_2, x_3)$, the corresponding Cartesian coordinates are $x_1^*(x_1, x_2) = \frac{x_1+2x_2}{\sqrt{3}}$ and $x_2^*(x_1, x_2) = x_1$. This is illustrated in Figure 3.1 where the trilinear point $(12, 8, 4)$ is transformed into the Cartesian point $\left(\frac{12+2(8)}{\sqrt{3}}, 12\right)$.

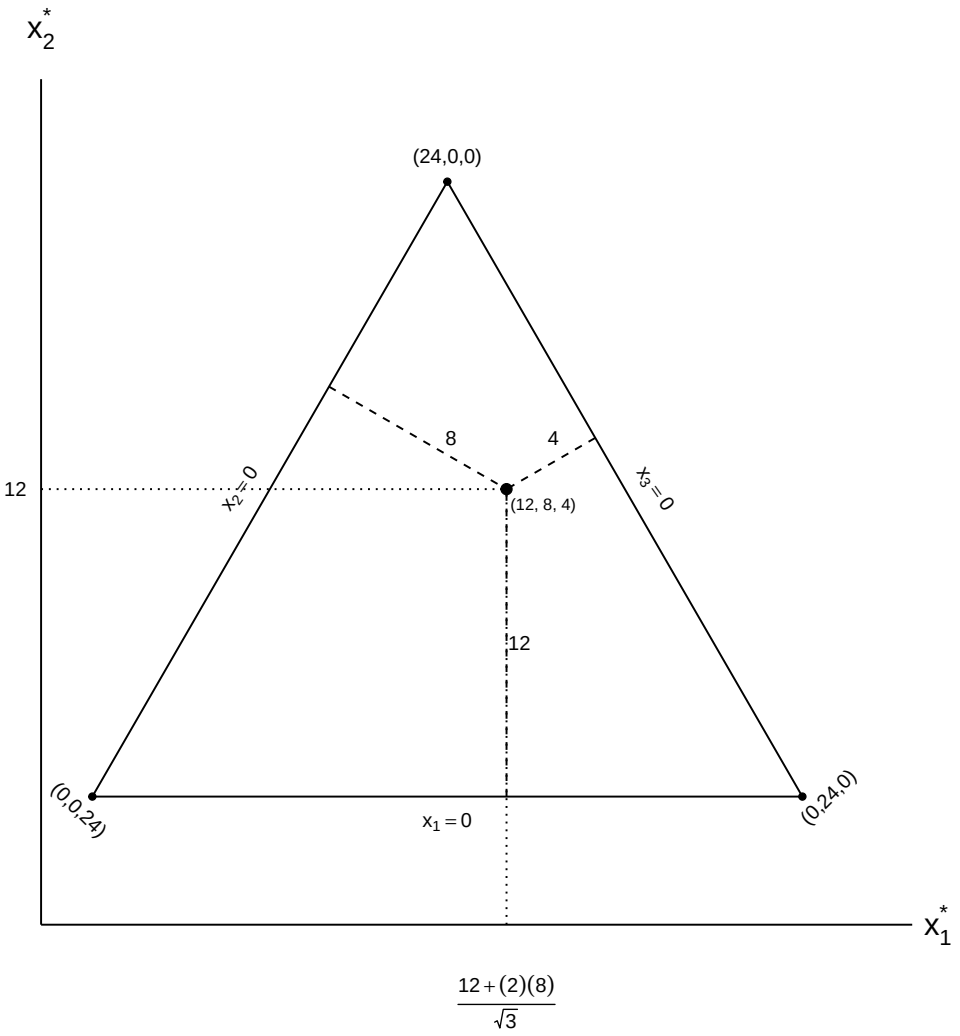

Figure 3.1: Example of a three-dimensional point $P = (12, 8, 4)$ in a trilinear coordinate system embedded in a two-dimensional Cartesian coordinate system with "x-axis" X_1^* and "y-axis" X_2^*.

From the mapping above, our original Cox model in equation (3.2.1) becomes the equivalent working model below:

$$h(t; x_1^*, x_2^*, \underline{\boldsymbol{z}}_i) = h_0(t) \exp\left\{g(x_1, x_2) + \underline{\boldsymbol{z}}_i^\top \underline{\boldsymbol{\alpha}}\right\} \tag{3.2}$$

$$= h_0(t) \exp\left[m\left\{(x_1^*(x_1, x_2), x_2^*(x_1, x_2)\right\} + \underline{\boldsymbol{z}}_i^\top \underline{\boldsymbol{\alpha}}\right]. \tag{3.3}$$

Directional derivatives of $g(x_1, x_2)$ have interesting interpretations related to the effects associated with substitutions among the compositional covariates. We discuss this interpretation more precisely in the next subsection. From now on we use x_1^* and x_2^* and suppress their dependence on (x_1, x_2) to simplify notation.

3.2.3 Partial Derivatives as Substitution Effects

In the usual linear Cox model, a coefficient is interpreted as the partial derivative of the log-hazard function with respect to its associated covariate when all other covariates are held constant. As a result, partial deriatives of $g(x_1, x_2)$ are also partial derivatives of the log-hazard function,

$$\frac{\partial}{\partial x_k} \log\left\{h(t; x_1, x_2, x_3, \underline{\boldsymbol{z}})\right\} = \frac{\partial}{\partial x_k} g(x_1, x_2).$$

Furthermore, since the compositional covariates sum to twenty-four for each observation, an increase in one component must be offset by decreasing in the sum of the others by the same amount, and directional derivatives can be used to estimate substitution effects. Table 3.1 summarizes the directional derivatives that estimate substitutions among the three components of the composition. We note that opposite directions are given by the negative values of the directional derivatives.

Table 3.1: Isotemporal Substitution Effects given by Partial Derivatives

Increase	Decrease	Constant	Partial Derivative (g)	Partial Derivative (m)
x_1	x_2	x_3	$\frac{\partial g}{\partial x_1} - \frac{\partial g}{\partial x_2}$	$-\frac{1}{\sqrt{3}}\frac{\partial m}{\partial x_1^*} + \frac{\partial m}{\partial x_2^*}$
x_3	x_1	x_2	$-\frac{\partial g}{\partial x_1}$	$-\frac{1}{\sqrt{3}}\frac{\partial m}{\partial x_1^*} - \frac{\partial m}{\partial x_2^*}$
x_2	x_3	x_1	$\frac{\partial g}{\partial x_2}$	$\frac{2}{\sqrt{3}}\frac{\partial m}{\partial x_1^*}$

Since our estimated model in equation (3.3) is based on $m(x_1^*, x_2^*)$ instead of $g(x_1, x_2)$, we use the multivariate chain rule to obtain the partial derivatives of $g(x_1, x_2)$:

$$\frac{\partial g}{\partial x_k} = \frac{\partial m}{\partial x_1^*}\frac{\partial x_1^*}{\partial x_k} + \frac{\partial m}{\partial x_2^*}\frac{\partial x_2^*}{\partial x_k} \tag{3.4}$$

for $k = 1, 2$. The last column of Table 3.1 gives the substitution effects in terms of $m\{x_1^*(x_1, x_2), x_2^*(x_1, x_2)\}$. The next section describes how we model and estimate $m(x_1^*, x_2^*)$ and its partial derivatives.

3.2.4 Bivariate smoothing using P-Splines

In statistical modeling, several methods are available to represent a bivariate function $m(.,.)$. Common approaches include kernel smoothing ([49]), as well as spline-based techniques, such as thin-plate splines and tensor product splines which are based on univariate B-splines ([54, 12]). Within the domain of Cox models, [23] demonstrated the application of a tensor product spline composed of two univariate cubic B-splines to model continuous-by-continuous interaction effects. In this work, we have chosen to utilize this particular approach to estimate the effects of physical activity time on mortality in a nationally representative survey.

In spline-based methods, two primary factors generally influence the quality of the fitted surface. The first is the number and location of the knots—the points where the individual polynomial pieces that make up the spline are joined together. The second factor is how much, if any, additional penalization should be imposed on the smoothness of the fitted function, and what kind of penalty to adopt. In our application, we adopt the approach by [16, 17] (known as penalized B-splines or P-Splines) which have desirable computational advantages. Also, the derivative of a B-splines is a linear combination of lower order B-splines (see Section 3.2.6), which is useful in our application. P-splines use a reasonably large number of knots placed at equally-spaced intervals over the range of each univariate variable in the smooth.

In addition, the method penalizes differences in the coefficients of the spline function to ensure goodness-of-fit. We represent $m(x_1^*, x_2^*)$ as a bivariate B-spline in the Cox model in equation (3.3).

Let $k_1 = 1, 2, \ldots, K_1$ and $k_2 = 1, 2, \ldots, K_2$ be equally-spaced knots locations over x_1^* and x_2^* respectively. Also, let $X_{1k_1}^{q_1}(x_{i1}^*)$ be a B-spline basis function of order q_1 at location k_1 defined at x_{i1}^* along the x_1^* axis and let $X_{2k_2}^{q_2}(x_{i2}^*)$ be defined similarly along the x_2^* axis. Then, for a given point (x_{i1}^*, x_{i2}^*) the bivariate tensor product spline representation of $m(x_{i1}^*, x_{i2}^*)$ is given by:

$$m(x_{i1}^*, x_{i2}^*) = \sum_{k_1=1}^{K_1} \sum_{k_2=1}^{K_2} X_{1k_1}^{q_1}(x_{i1}^*) X_{2k_2}^{q_2}(x_{i2}^*) \theta_{k_1 k_2} \tag{3.5}$$

where $\theta_{k_1 k_2}$ is a real unknown coefficient.

Collecting the B-spline basis functions into vectors for each i, we can rewrite equation (3.5) in matrix notation as

$$m(x_1^*, x_2^*) = (\boldsymbol{X_1}^{q_1} \odot \boldsymbol{X_2}^{q_2})\, \underline{\boldsymbol{\theta}} \tag{3.6}$$

where $\boldsymbol{X_1}^{q_1} = \left(\underline{\boldsymbol{x}}_{11}^{q_1}, \underline{\boldsymbol{x}}_{12}^{q_1}, \ldots, \underline{\boldsymbol{x}}_{1K_1}^{q_1}\right)$, with dimension $n \times K_1$ and $\underline{\boldsymbol{x}}_{1j}^{q_1}$ (for $j = 1, 2, \ldots, K_1$) is a vector of the B-spline basis functions evaluated for all observations ($i = 1, 2, \ldots, n$) at the knot location j. For instance, $\underline{\boldsymbol{x}}_{1j}^{q_1} = \left(X_{1j}^{q_1}(x_{11}^*), X_{1j}^{q_1}(x_{12}^*), \ldots, X_{1j}^{q_1}(x_{1n}^*)\right)^\top$. $\boldsymbol{X_2}^{q_2}$ is defined similarly, $\underline{\boldsymbol{\theta}}$ is a $K_1 K_2$ dimensional vector of real unknown coefficients, and the operator $\odot$ is used as the row-Kronecker product operator (e.g. [54] Appendix B). The row-Kronecker product of two real matrices $A_{n\times m}$ and $B_{n\times p}$ is the $n \times mp$ matrix C with each row i formed by multiplying each element of row i of A by the i^{th} row of B. Further, we collect the vectors $\underline{\boldsymbol{z}}_i$ in equation (3.3) as rows of a matrix $\boldsymbol{Z}_{n\times p}$ resulting in the following form of equation (3.3) in matrix notation.

$$h(t; \boldsymbol{X_1}^{q_1}, \boldsymbol{X_2}^{q_2}, \boldsymbol{Z}) = h_0(t) \exp\left\{(\boldsymbol{X_1}^{q_1} \odot \boldsymbol{X_2}^{q_2})\underline{\boldsymbol{\theta}} + \boldsymbol{Z}\underline{\boldsymbol{\alpha}}\right\} \tag{3.7}$$

As shown in equation (3.7), the exponentiated terms are linear in $\underline{\boldsymbol{\theta}}$ and $\underline{\boldsymbol{\alpha}}$, and to mitigate against overfitting, we impose a penalty on $\underline{\boldsymbol{\theta}}$, the coefficient vector of the tensor product spline basis, during estimation. There are many kinds of penalties to impose on a spline-based method such as a tensor product spline. We choose one that penalizes each univariate smooth marginally based on differences of the corresponding coefficients ([16, 54]). Construction of such difference-based P-spline penalties is straightforward. Suppose Δ is the first-order backward difference operator such that $\Delta\gamma_k = \gamma_k - \gamma_{k-1}$ and let $\underline{\boldsymbol{\theta}}_k$ be the coefficients vector of the k^{th} univariate B-spline basis matrix of order q, $\boldsymbol{X}_{\boldsymbol{k}}^q$, and define $\boldsymbol{D}_m\underline{\boldsymbol{\theta}}_k$ as a vector of m^{th} order differences of the coefficient vector $\underline{\boldsymbol{\theta}}_k$. Here, the rows of the matrix $\boldsymbol{D}_m$ are simply the m^{th} order differencing operations. Then, the m^{th} order smoothness penalty on the univariate B-spline basis $\boldsymbol{X}_{\boldsymbol{k}}^q$ is given by the quadratic form $\lambda\underline{\boldsymbol{\theta}}_k^\top \boldsymbol{D}_m^\top \boldsymbol{D}_m \underline{\boldsymbol{\theta}}_k$. Here, $\lambda \geq 0$ is the smoothing parameter that determines the level of penalization. A thorough discussion on how such difference-based univariate penalties are constructed can be found in [16].

For our application, we form a penalty matrix for the full tensor product spline coefficients vector $\underline{\boldsymbol{\theta}}$ in equation (3.7) by simply combining the univariate penalty matrices defined above ([54]). To simplify exposition, we denote $\boldsymbol{D}_m^\top \boldsymbol{D}_m$ as $\boldsymbol{P}_m$. Let $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$ be two smoothing parameters along the two univariate axes, we use the smoothing penalty matrix, $\boldsymbol{P}_\lambda = \lambda_1 (\boldsymbol{P}_m \otimes I_{K_1}) + \lambda_2 (I_{K_2} \otimes \boldsymbol{P}_m)$. Here, I_{K_j} $(j = 1, 2)$ is the identity matrix of dimension K_j and $\otimes$ is the Kronecker product. Once a representation is chosen for the smooth function $m(x_1^*, x_2^*)$ in our Cox model in equation (3.3), and a smoothness penalty is determined, estimating our Cox model proceeds by utilizing well-established procedures for fitting a Cox model. We discuss this next with a few practical nuances for complex multi-stage survey data like that provided by NHANES.

3.2.5 Estimating the Cox Model

In order to further simplify the notation for fitting our Cox model, we denote $\boldsymbol{X} = [(\boldsymbol{X_1}^{q_1} \odot \boldsymbol{X_2}^{q_2}) : \boldsymbol{Z}]$ and define $\underline{\boldsymbol{\beta}} = \begin{pmatrix} \underline{\boldsymbol{\theta}} \\ \underline{\boldsymbol{\alpha}} \end{pmatrix}$, a new coefficient vector that stacks $\underline{\boldsymbol{\theta}}$ and $\underline{\boldsymbol{\alpha}}$ together appropriately. With this new notation, the hazard function in equation (3.7) is equivalently $h(t; \boldsymbol{X}) = h_0(t) \exp\left(\boldsymbol{X}\underline{\boldsymbol{\beta}}\right)$ with an updated block-diagonal penalty matrix $\boldsymbol{S}_\lambda = \begin{pmatrix} \boldsymbol{P}_\lambda & \\ & \boldsymbol{O}_p \end{pmatrix}$. Here, $\boldsymbol{O}_p$ is a $p \times p$ zero matrix.

We estimate the parameters in equation (3.7) by optimizing the penalized partial log-likelihood function of the Cox model. For complex survey designs like those in large-scale studies such as NHANES, [4] provided a weighted estimator for the parameters in the Cox model. The weighted partial log-likelihood of our model is given by:

$$l(\underline{\boldsymbol{\beta}}|\boldsymbol{X}, \underline{\boldsymbol{w}}) = \sum_{i=1}^{n} \delta_i w_i \left\{ \underline{\boldsymbol{x}}_i^\top \underline{\boldsymbol{\beta}} - \log \sum_{j=1}^{n} w_j Y_j(t_i) \exp(\underline{\boldsymbol{x}}_i^\top \underline{\boldsymbol{\beta}}) \right\} \tag{3.8}$$

where $\underline{\boldsymbol{w}}$ is a vector of survey weights, δ_i is the event or censoring indicator which is 1 if participant i died or zero otherwise, and $\underline{\boldsymbol{x}}_i$ is the i^{th} row of $\boldsymbol{X}$. $Y_j(t_i) = 1$ if $t_j \geq t_i$ and zero otherwise. $Y_j(t_i)$ is used to determine the risk set at the event time for the i^{th} individual. The solution to the penalized partial log-likelihood optimization in equation (3.9) below gives the estimated coefficients needed to fit a surface of mortality risk over physical activity time. Given smoothing parameters λ_1 and λ_2, the optimal coefficient vector $\widehat{\underline{\boldsymbol{\beta}}}$ is given by:

$$\widehat{\underline{\boldsymbol{\beta}}} = \arg\max_{\underline{\boldsymbol{\beta}}} \; l(\underline{\boldsymbol{\beta}}|\boldsymbol{X}, \underline{\boldsymbol{w}}) - \frac{1}{2}\underline{\boldsymbol{\beta}}^\top \boldsymbol{S}_\lambda \underline{\boldsymbol{\beta}}. \tag{3.9}$$

Estimating the smoothing parameters λ_1 and λ_2 is usually the most challenging part of the estimation process in this class of semiparametric models ([54]). However, there are some well-studied and efficient methods available to estimate the smoothing

parameters. Examples include a nested iteration method that attempts to optimize both the smoothing parameters and the coefficient vector $\underline{\boldsymbol{\beta}}$ directly ([56]). Another approach is a generalization of the Fellner-Schall method ([20, 39]) by [55] to accommodate tensor product smooths where the penalty components are not easily separated. The highly capable R package *mgcv* ([54]) is equipped to handle such optimization routines through its *gam* and *bam* (typically used for large datasets) procedures. Unfortunately though, at the time of writing, the default routines in *mgcv* for survival analysis using Cox models do not support weighted estimation of the model parameters, which is required for our application in Section 3.

To address this challenge, we use a method from [51] to reformulate our Cox model as a Poisson generalized linear model. That allows us to use routines in *mgcv* to estimate a penalized model and incorporate observation weights. The reformulation is done by creating a new response variable r_{ij} for each non-censoring event time t_i where $r_{ij} = 1$ for all observations, j, with event time t_i and $r_{ij} = 0$ otherwise. Using this new response variable, [51] showed that the Poisson model $r_{ij} \sim \text{Poisson}(\mu_{ij})$, where $\log(\mu_{ij}) = \eta_i + \boldsymbol{x}_j(t_i)\underline{\boldsymbol{\beta}}$ yields the same estimates and inference for $\underline{\boldsymbol{\beta}}$ as in the original Cox model. Here, η_i's are intercept terms for the distinct event times and $\boldsymbol{x}_j(t_i)$ is the row of the covariate matrix $\boldsymbol{X}$ for observation j for the data created for event time t_i. This approach may not be computationally efficient for large datasets with more distinct event times as the creation of such artificial Poisson data may result in much larger datasets ([54]). For our application with $3,035$ records though, this resulted in $456,337$ records. The model was easily fit using the *bam* procedure in the *mgcv* package.

3.2.6 Partial derivatives of $\widehat{m}(x_1^*, x_2^*)$

Next, we describe the computation of the substitution effects in Table 3.1. For univariate B-spline smoothing with equal distance between the knots, [16] gave deriva-

tives of a $q-$order B-spline function using $q-1$ order B-spline functions and differences of the estimated coefficients. This derivation was based on a formula earlier developed by [12]. These derivative formulas can be adapted to find partial derivatives for our tensor product representation of $m(x_1^*, x_2^*)$. Let $\boldsymbol{D}_1$ be a $(K_1 - 1) \times K_1$ differencing matrix and $\boldsymbol{D_2}$ be a $(K_2 - 1) \times K_2$ differencing matrix. Also let I_{K_1} and I_{K_2} be identity matrices of dimensions K_1 and K_2 respectively.

From equation (3.6), we can write the estimated function $\widehat{m}(x_1^*, x_2^*) = (\boldsymbol{X_1^{q_1}} \odot \boldsymbol{X_2^{q_2}})\,\underline{\widehat{\boldsymbol{\theta}}}$. We give the partial derivatives of $\widehat{m}$ as below:

$$\frac{\partial \widehat{m}(x_1^*, x_2^*)}{\partial x_1^*} = \frac{1}{h_1}\left(\boldsymbol{X_1^{q_1-1}} \odot \boldsymbol{X_2^{q_2}}\right)(\boldsymbol{D_1} \otimes \boldsymbol{I_{K_2}})\,\underline{\widehat{\boldsymbol{\theta}}} \tag{3.10}$$

and

$$\frac{\partial \widehat{m}(x_1^*, x_2^*)}{\partial x_2^*} = \frac{1}{h_2}\left(\boldsymbol{X_1^{q_1}} \odot \boldsymbol{X_2^{q_2-1}}\right)(\boldsymbol{I_{K_1}} \otimes \boldsymbol{D_2})\,\underline{\widehat{\boldsymbol{\theta}}} \tag{3.11}$$

where h_1 is the constant difference between the knots on the x_1^*-axis and h_2 is similarly defined on the x_2^*-axis. $\boldsymbol{X_1^{q_1-1}}$ is a matrix of $q_1 - 1$ order B-spline basis functions defined over $K_1 - 1$ total knots and $\boldsymbol{X_2^{q_2-1}}$ is similarly defined over $K_2 - 1$ total knots.

We note that the partial derivatives in equations (3.10) and (3.11) are linear combinations of the estimated coefficient vector $\underline{\widehat{\boldsymbol{\theta}}}$. Therefore, it is straightforward to obtain standard errors of the estimated substitution effects using the standard errors of $\underline{\widehat{\boldsymbol{\theta}}}$. Standard errors of $\underline{\widehat{\boldsymbol{\theta}}}$ follow from the sandwich weighted variance estimator in [4] given as $\mathcal{I}^{-1}U^TWU\mathcal{I}^{-1}$. Here, $\mathcal{I}$ is the observed information matrix, U is the unweighted score residual matrix and W is a diagonal matrix with the squared weights as elements.

3.3 Analysis of NHANES Data

The US National Health and Nutrition Examination Survey (NHANES) is a series of cross-sectional surveys of the US population which use a multi-stage sampling de-

sign to ensure that the data are representative of the non-incarcerated US population. NHANES measures many health-related factors, and the 2003-2006 waves assessed physical activity / sedentary behavior using Actigraph AM-7164 accelerometers. Participants were instructed to wear the device around their waist for seven days, only removing it when sleeping or bathing. Data from those devices can be used to estimate aspects of participants' physical activity / sedentary behaviors. Combined with baseline characteristics and links to the National Death Index to record deaths up to December 31st, 2011 ([32]), [22] used those data and linear Cox models to investigate the association between physical activity / sedentary behaviors and mortality hazard in people who were 50-79 years old at the time of the NHANES assessment. The paper found that replacing sedentary time with either light activity or moderate to vigorous physical activity was associated with lower mortality hazards.

One drawback of [22]'s approach is that it makes the parametric assumption that the physical activity / sedentary time substitution associations do not depend on the initial levels of those behaviors. For instance, they assume that the association with increasing sedentary time from 11 to 12 hours and decreasing activity from 2 to 1 is the same as increasing sedentary time from 1 to 2 hours and decreasing activity from 12 to 11 hours. Our nonparametric approach does not make that assumption. (See the top panels of Figures 3.2 and 3.6.)

As described in [22] we removed participants who did not wear the device long enough and whose data were invalid. Descriptive statistics are in Table 3.2. Our final sample consisted of $3,035$ participants, including 21 who were censored at the time of accidental death. We also updated the dataset to include deaths up to December 31st, 2019. A summary of the leading causes of death is in Table 3.3.

Table 3.2: Summary statistics of analysis data (NHANES 2003-2006)

n (sample size)	3,035
Demographic Measures:	
Age (yrs)	61.9(8.4)
Female (%)	53%
White (%)	79%
Black (%)	10%
Hispanic (%)	7%
Education Level:	
Less than High School	18%
More than High School	55%
Health Conditions	
BMI (kg/m^2)	29.0(6.2)
Current Smoker(%)	18.5%
CHD(%)	7.2%
CHF(%)	4.8%
Stroke(%)	4.5%
Cancer(%)	15.4%
Mobility Problem(%)	26.8%
Diabetes(%)	14.4%
Daily Physical Activity Measures:	
Sendentary Time (hrs)	8.6(2.2)
Light Time (hrs)	5.4(1.7)
MVPA Time (hrs)	0.26(0.3)

CHD is Coronary Heart Disease, CHF is Congestive Heart Failure, BMI is Body Mass Index. Mobility Problem is defined as having difficulty walking for a quarter mile. Values are survey-weighted.

Table 3.3: Causes of death in the 2019 Mortality Follow-up data

Heart Disease	276
Malignant Neoplasms	266
Chronic Lower Respiratory Diseases	68
Cerebrovascular Diseases	52
Alzheimer's Disease	45
Diabetes	40
Nephritis	26
Influenza and Pneumonia	19
All other causes	235

We fit two models to these data. The first considered substitutions among sleep & non-wear time, sedentary time, and activity, and the second considers substitutions among sedentary time, light activity, and moderate to vigorous activity with sleep & non-wear time held constant. The methods by which time in those categories are estimated are described in [22].

The first Cox model used the following hazard function:

$$h(t|x_{i1}, x_{i2}, x_{i3}, \underline{\boldsymbol{z}}_i) = h_0(t) \exp\left\{g(x_{i1}, x_{i2}) + \underline{\boldsymbol{z}}_i^\top \underline{\boldsymbol{\alpha}}\right\} \text{ for } i = 1, 2, \ldots, 3035, \quad (3.12)$$

where: x_{i1} is the average number of hours participant i was either asleep or not wearing the device per day, x_{i2} is the average number of hours participant i spent in sedentary behaviors per day, x_{i3} is the average number of hours participant i spent physically active per day. The linear covariates, $\underline{\boldsymbol{z}}_i$, are the same ones used in [22] and include age, sex, race/ethnicity, education level, BMI, smoking, and the presence of chronic diseases and mobility limitations. We modeled the smooth surface with a fourth-degree bivariate tensor product P-splines with $K_1 = K_2 = 20$. The weighted coefficients in the Cox model were estimated using equations (3.8) and (3.9) through the equivalent Poisson model ([51]) using the *bam* procedure in the *mgcv* package in R. To provide a comparison, we also fit a linear version of the model:

$$h(t|x_{i1}, x_{i2}, x_{i3}, \underline{\boldsymbol{z}}_i) = h_0(t) \exp\left(x_{i1}\nu_1 + x_{i2}\nu_2 + \underline{\boldsymbol{z}}_i^\top \underline{\boldsymbol{\alpha}}\right) \text{ for } i = 1, 2, \ldots, 3035. \quad (3.13)$$

where ν_1 and ν_2 are unknown coefficients and the other terms are as defined in equation (3.12).

The estimated $g(x_1, x_2)$ and the estimated linear surface are in Figure 3.2 (top and bottom panels respectively). The lack of parallel contour lines in the top panel illustrates the benefit of the nonparametric model. That model allows a rapid increase in the association with mortality hazard when activity levels are low, but smaller increases for higher levels of activity. The linear model assumes the log hazard ratio is a plane in any two of x_1, x_2, and x_3. We also note that the range of the log-hazard ratios is smaller in the linear model compared to the spline approach. We posit that this is because the linear approach in effect over-smooths the log-hazard ratio rather than letting the data determine the shape of the function.

Directional derivatives (Section 3.2.6) for both models are in Figures 3.3-3.5. The arrows in the pictures indicate the direction of the substitution that associates with an increased log-hazard ratio. The length is proportional to the size of the association and black indicates a pointwise p-value less than 0.05. The top panel of Figure 3.3 illustrates that increasing sedentary time and decreasing activity associates with increased mortality hazard, but the effect is only seen when activity is approximately less than 6 hours per day. The bottom panel of Figure 3.3 shows that the standard linear model both misses this subtlety since it only estimates one constant derivative and it also estimates a much smaller effect.

The maximum partial derivative of the nonparametric model is 0.52, which is equivalent to a hazard ratio of 1.68. This corresponds to approximately 68% increase in mortality risk when an hour of activity is replaced with an hour of sedentary behavior while holding sleep & non-wear time and baseline covariates constant. The linear model, on the other hand, uniformly estimates the partial derivative as 0.2

which corresponds to a hazard ratio of 1.22, thus, only a 22% increase in mortality risk.

Figure 3.4 shows the associations with increasing sleep & non-wear time and decreasing activity. Again, the top panel shows that the associations are strongest for low levels of activity, but in this case, the associations begin at about 8 hours of activity per day. Similar to the effects in Figure 3.3, the maximum partial derivative is 0.49. This means that substituting an hour of sleep & non-wear time for activity is associated with up to approximately 63% increase in mortality risks after adjusting for sedentary time and baseline covariates. The drawbacks of the linear model are similar to those described above and in this case, it only estimates the increase in mortality risks at 21%.

Finally, Figure 3.5 (top panel) shows the associations with increasing sedentary time and decreasing sleep & non-wear time while holding activity fixed. The linear model (bottom panel of Figure 3.5) shows no significant effect here, and the nonparametric approach only shows positive effects when sedentary time is above approximately 12 hours and activity is around 6 hours. The maximum partial derivative, in this case, is 0.16. This means, being sedentary instead of sleeping (& non-wear) for one hour is associated with mortality risks of up to approximately 17% after adjusting for activity time and baseline covariates.

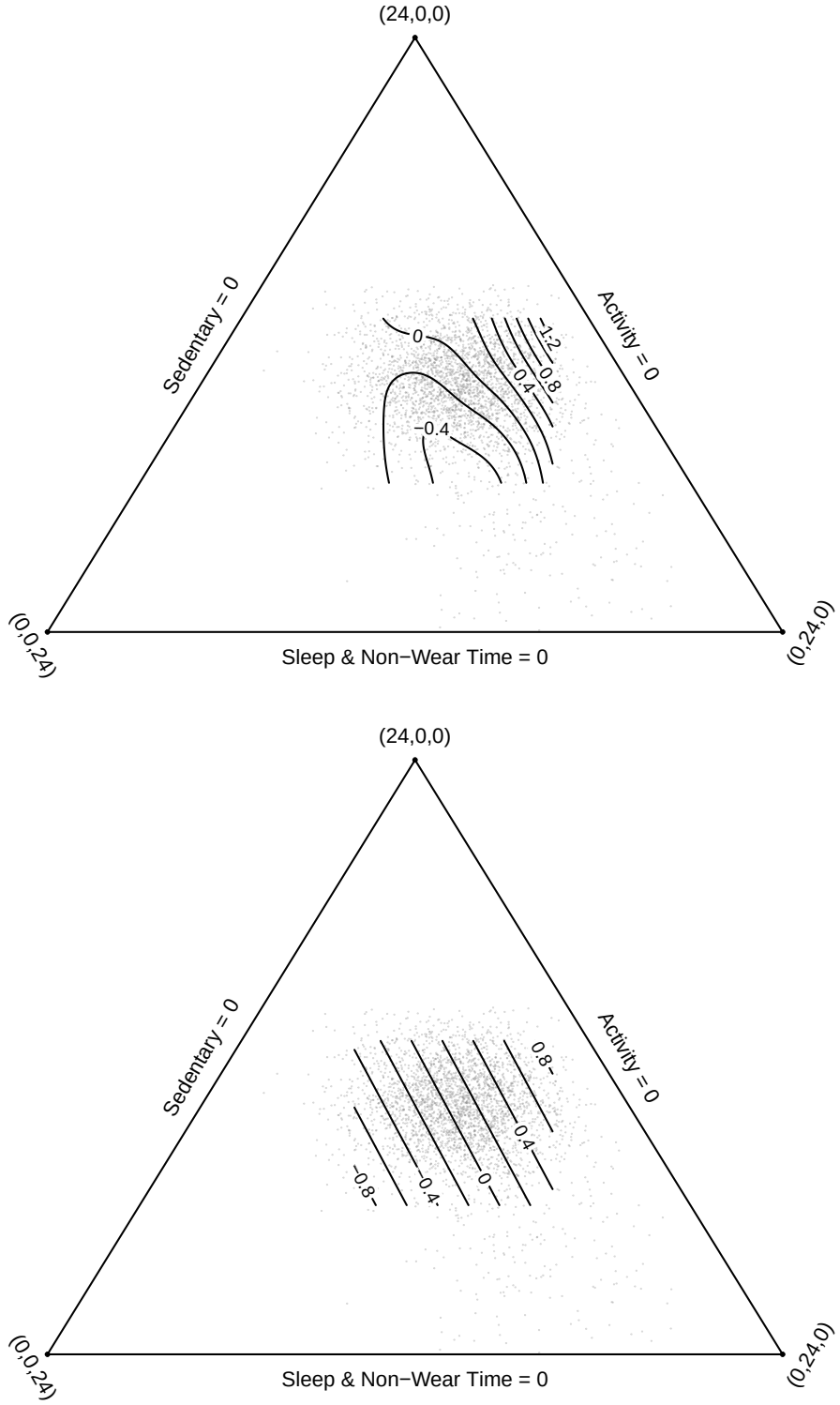

Figure 3.2: Top panel: Estimated log-hazard tensor product surface of the effects of sedentary, sleep & non-wear and activity time on all-cause mortality. Bottom panel: Estimated log-hazard surface using the linear model.

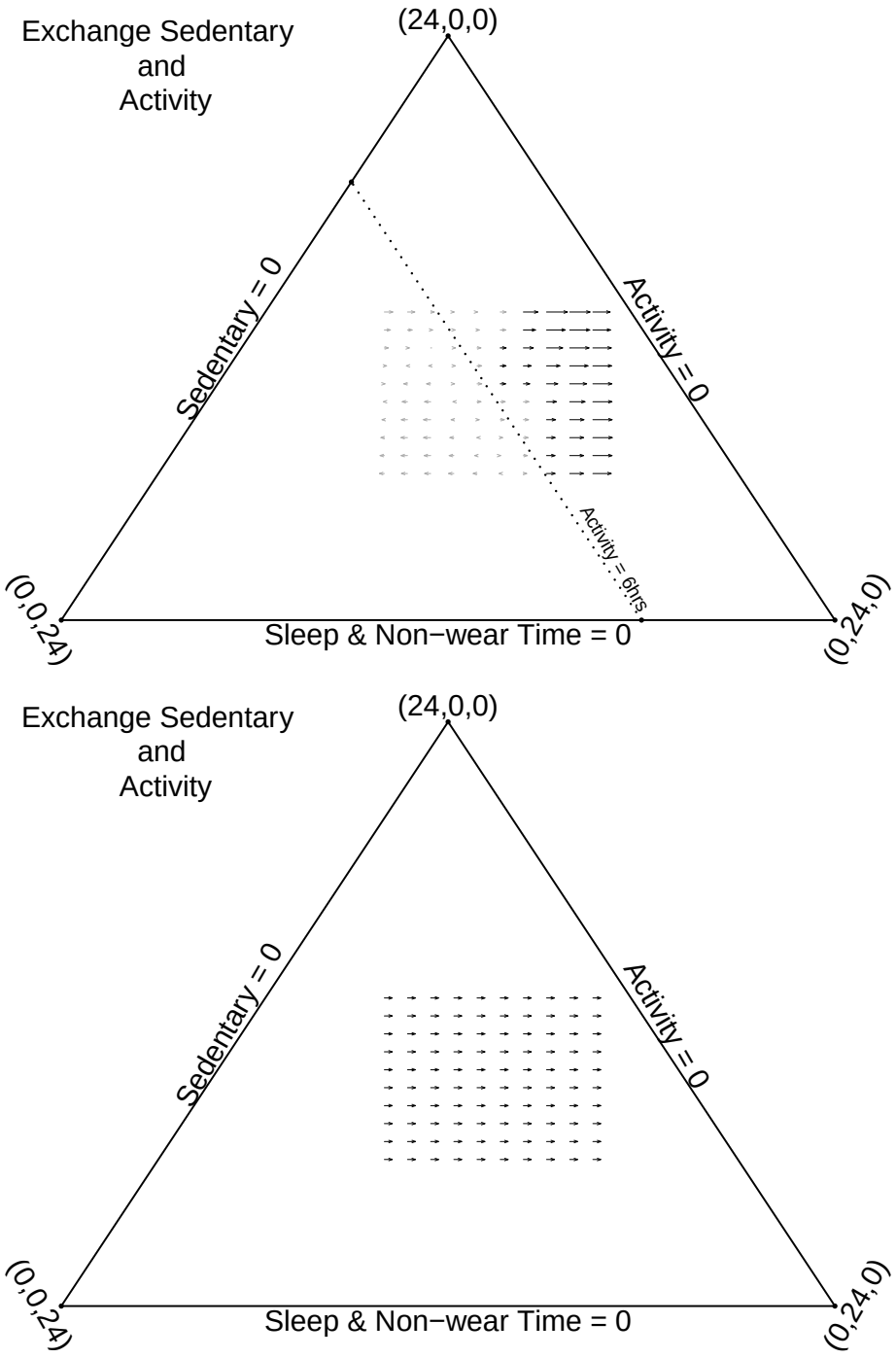

Figure 3.3: Top panel: Estimated effects of exchanging sedentary behavior and activity while holding sleep & non-wear and baseline covariates constant. The maximum partial derivative is 0.52. Bottom panel: Same substitution effects estimated by the linear model. The partial derivative is 0.20.

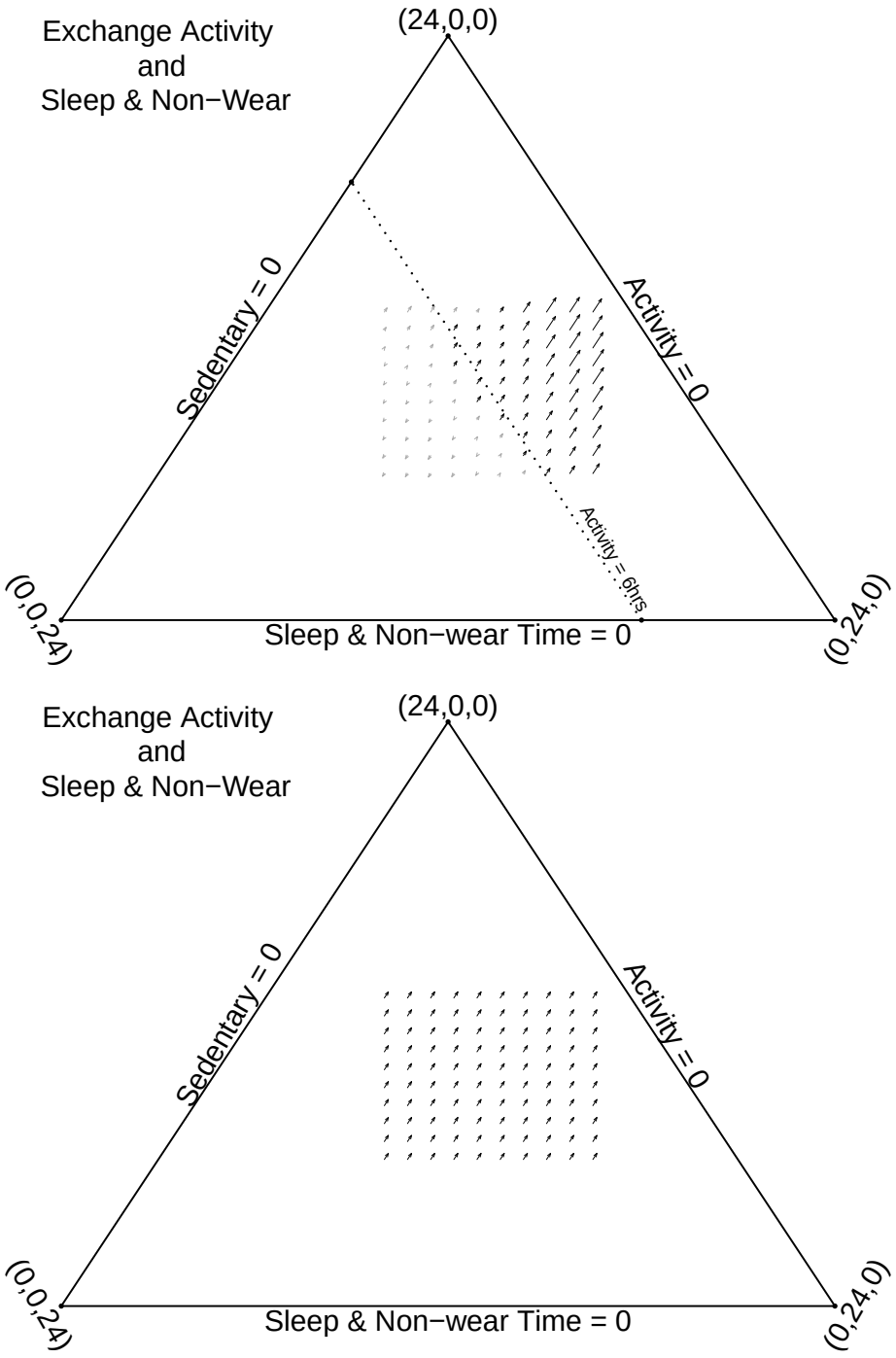

Figure 3.4: Top panel: Estimated effects of exchanging sleep & non-wear and activity while holding sedentary behavior and baseline covariates constant. The maximum partial derivative is 0.49. Bottom panel: The same substitution effects as estimated by the linear model. The partial derivative is 0.19.

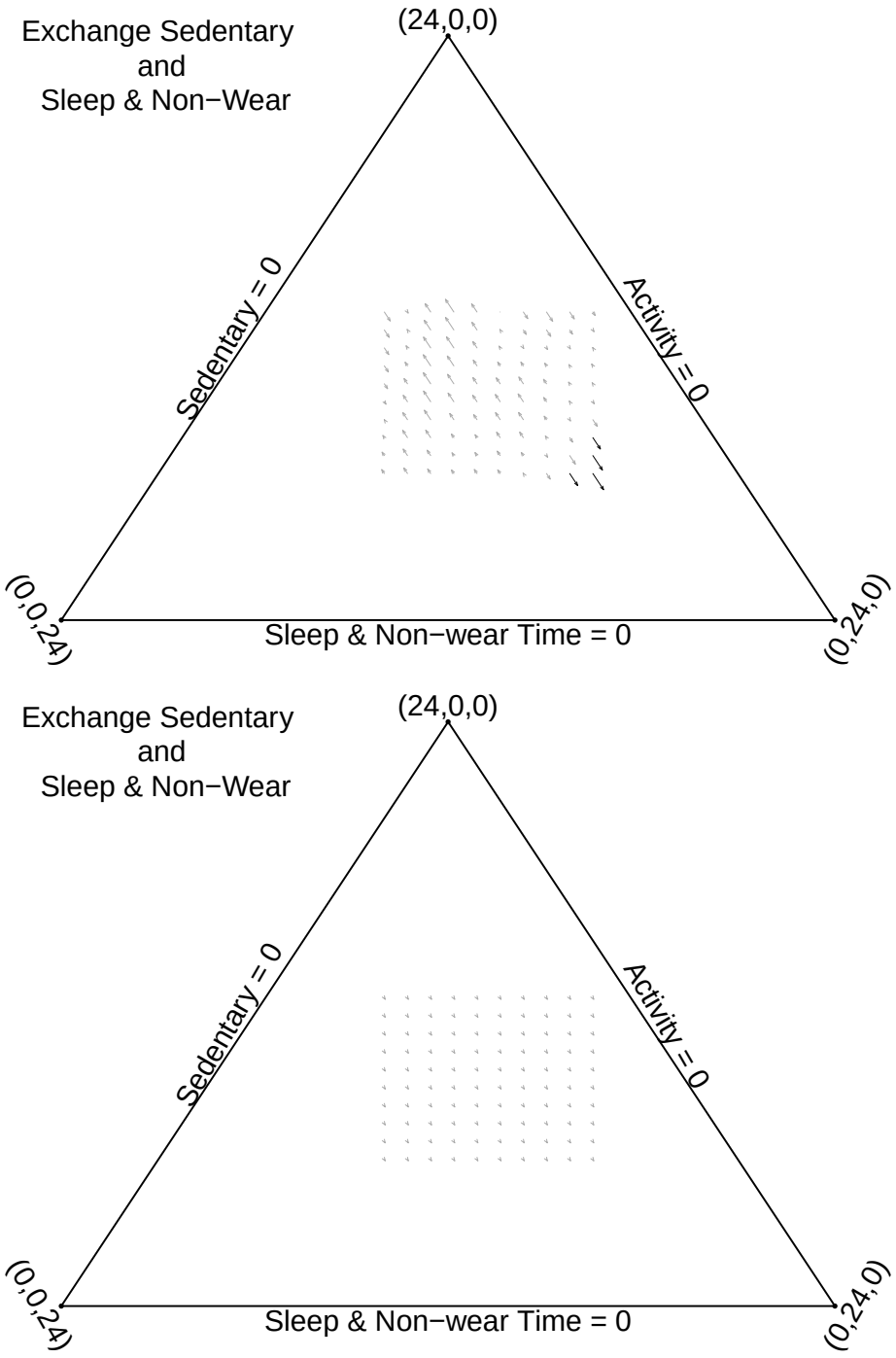

Figure 3.5: Top panel: Estimated effects of exchanging sedentary behavior and sleep & non-wear while holding activity and baseline covariates constant. The maximum partial derivative is 0.16. Bottom panel: The same substitution effects as estimated by the linear model. The partial derivative is 0.01.

In a second analysis, we subdivided total wear time into sedentary, light and moderate-to-vigorous physical activity (MVPA), and then considered the percentage of each as the composition; sleep & non-wear time was included as a linear term. Let x_{i1} be sedentary hours, x_{i2} be light hours, x_{i3} be MVPA hours, and x_{i4} be sleep & non-wear hours for person i. Further, let $r_{i1} = \frac{x_{i1}}{x_{i1}+x_{i2}+x_{i3}}$ be the proportion of wear time spent engaging in sedentary behaviors, and $r_{i2} = \frac{x_{i2}}{x_{i1}+x_{i2}+x_{i3}}$ be the proportion of wear time spent engaging in light activity. With that notation, our model for the hazard is

$$h(t|x_{i1}, x_{i2}, x_{i3}, x_{i4}, \underline{\boldsymbol{z}}_i) = h_0(t) \exp\left(g(r_{i1}, r_{i2}) + x_{i4}\gamma + \underline{\boldsymbol{z}}_i^\top \underline{\boldsymbol{\alpha}}\right) \text{ for } i = 1, 2, \ldots, 3035. \tag{3.14}$$

By consolidating the linear terms in the model in equation (3.14), the model is essentially the same as in equation (3.12), and we can estimate and make inferences with previously described methods. As done in the previous analysis, we also fit a linear model for comparison.

Our analysis reveals that a higher mortality risk is associated with decreasing time spent on MVPA, similar to [22]. This is illustrated in Figure 3.6 (top panel), which showcases the estimated surface of the log hazard ratio of the associations between proportions of device-wear time spent in sedentary, light activity, and MVPA, and mortality. The contours in the figure increase from -0.8 to 0.8 as the duration decreases along the MVPA axis, specifically moving towards the line denoted by $MVPA = 0$ from the point $(0, 0, 1)$. Again, the non-parallel nature of these contours implies that the associations can vary, even when located along the same axis. This is an advantage of our spline model as it allows for a nuanced interpretation of data, as compared to a standard model which linearly represents proportions of wear time allocations. In the linear model, as demonstrated in the bottom panel of Figure 3.6, associations remain constant along the same axis.

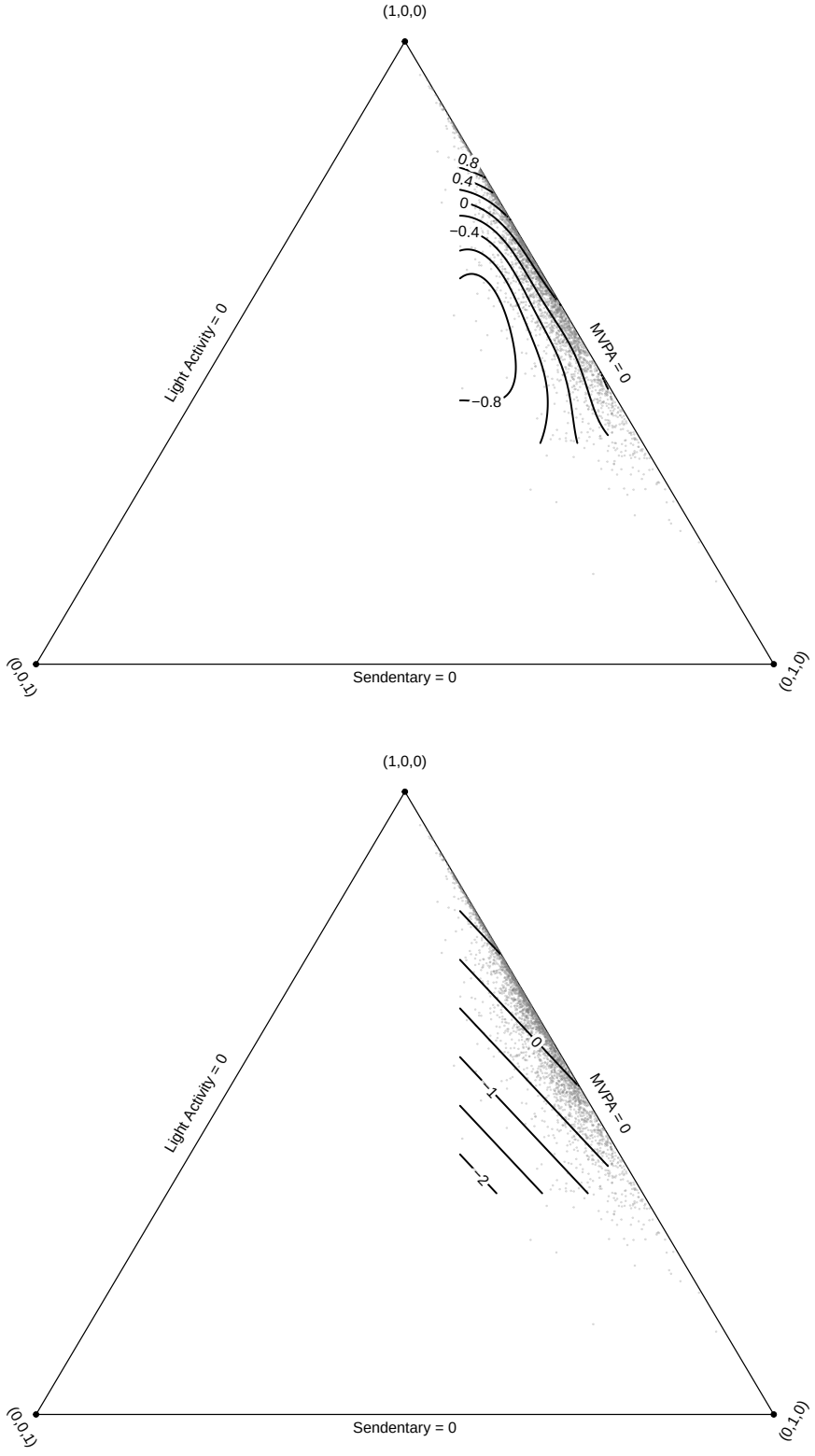

Figure 3.6: Top panel: Estimated log-hazard tensor product surface of the association between the proportion of hours of wear time spent in sedentary, light activity and MVPA on all-cause mortality. Bottom panel: Estimated log-hazard surface using the linear model.

We present the directional derivatives in Figures 3.7-3.9. These represent the substitution effects of the proportional time allocations across sedentary, light activity, and MVPA on mortality. Similar to our earlier analysis, the arrows in the figures depict the magnitude of the partial derivatives, while the directions signify the order of substitutions associated with an increase in mortality risks. Also, as before, black arrows indicate a pointwise p-value less than 0.05.

Figure 3.7 (top panel) shows that substituting light activity time for MVPA is associated with high mortality risks. However, these associations are significant only when the time spent on MVPA is below 12.5% of total device-wear time. The maximum partial derivative observed is 10.46, suggesting that substituting 1% of light activity time for 1% of MVPA time could be associated with an approximately 11% increase in mortality risks, assuming the same level of sedentary time, sleep (& non-wear) time, and baseline covariates among US adults. Equivalently, for a US adult with about 10 hours of sleep (& non-wear) time, substituting 10 minutes of MVPA for 10 minutes of light activity—while keeping sedentary time constant—is roughly associated with a decrease in mortality risks by up to 12%, after factoring in baseline covariates. The directional derivatives of exchanging light activity time and MVPA using the standard linear model are presented in the bottom panel of Figure 3.7. As with previous results, the linear model estimates a lower partial derivative of 8.08, irrespective of the level of light activity time or MVPA.

Similarly, the top panel of Figure 3.8 illustrates that substituting sedentary time for MVPA, while maintaining the same level of light activity time, is associated with higher mortality risks. Statistically significant associations are observed when MVPA time is below 10% of total wear time, and sedentary time is above 50%. The maximum partial derivative here is 11.12, which implies that increasing sedentary time by 1% while decreasing MVPA by 1%, given the same level of light activity, sleep (& non-wear) time, and baseline covariates, is associated with up to 18% increase in mortality

risks among US adults. The standard linear model produces a partial derivative of 10.44 and is presented in the bottom panel of Figure 3.8.

Finally, we show in Figure 3.9 (top panel) that exchanging sedentary behaviors for light activity, while holding MVPA time constant, is also associated with higher mortality risks after adjusting for sleep (& non-wear) time and baseline covariates. These associations are significant when sedentary behaviors make up over 20% of wear time. Moreover, the maximum partial derivative for this substitution is 9.45. This means that substituting 1% of sedentary time for light activity is associated with up to 10% increase in mortality risks. The linear model, in this case, is presented in the bottom panel of Figure 3.9 below. It only measures a partial derivative of 2.36 which potentially significantly underestimates the association.

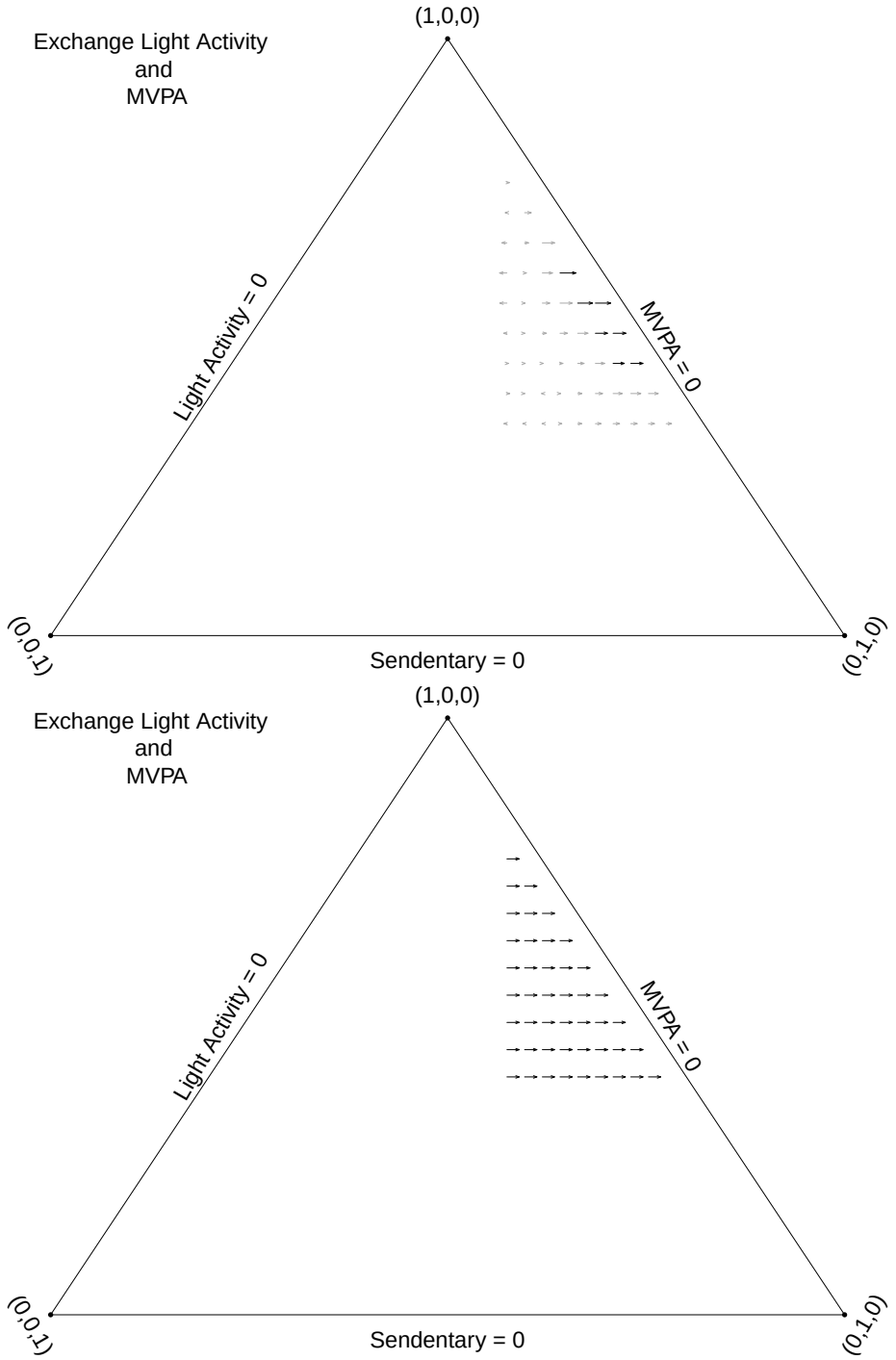

Figure 3.7: Top panel: Estimated effects of exchanging proportions of light activity and MVPA while holding sedentary time, sleep & non-wear time and baseline covariates constant. The maximum partial derivative is 10.46. Bottom panel: Same substitution effects estimated by the linear model. The partial derivative is 8.08.

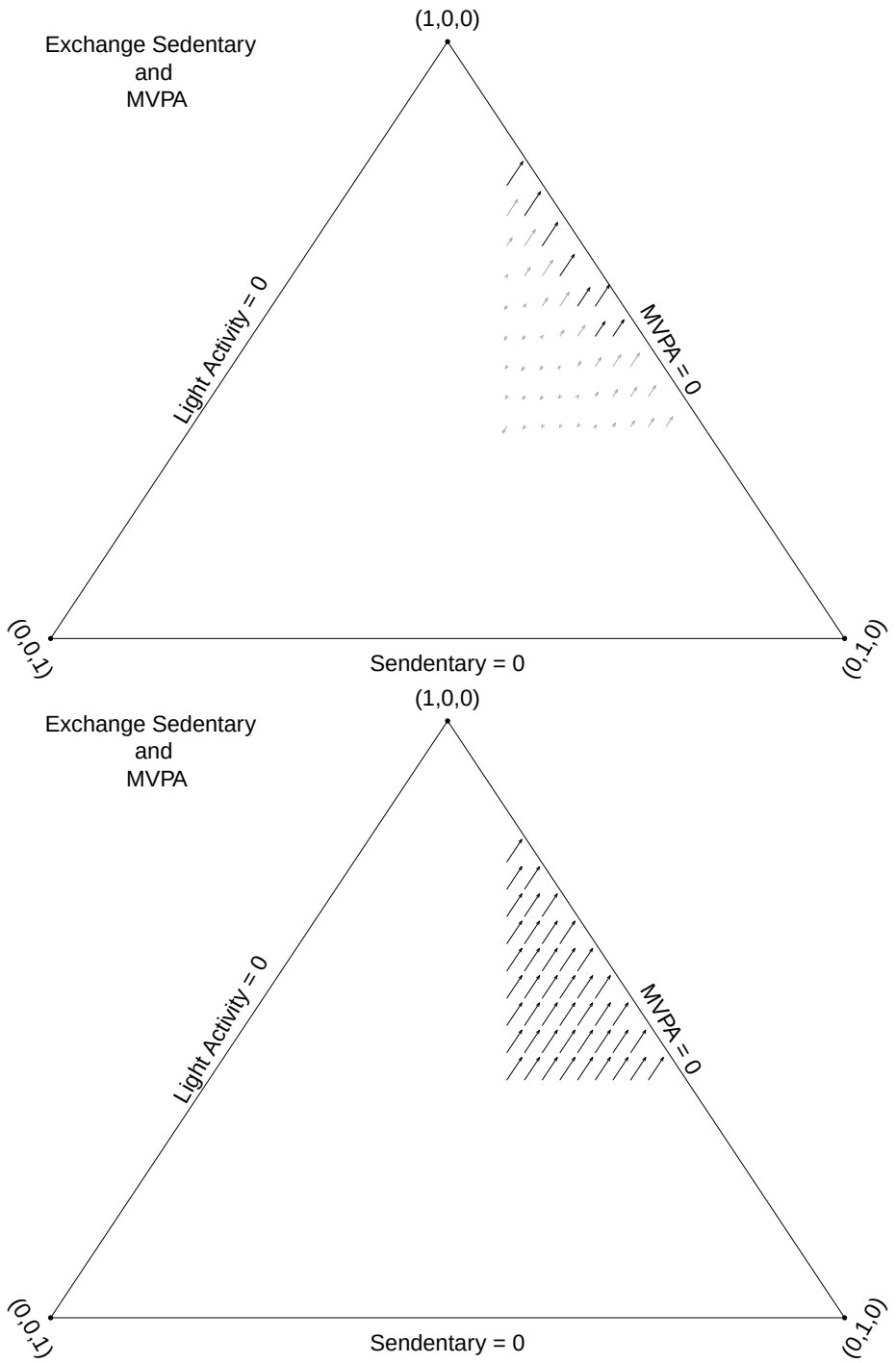

Figure 3.8: Top panel: Estimated effects of exchanging proportions of sedentary time and MVPA while holding light activity, sleep & non-wear time and baseline covariates constant. The maximum partial derivative is 11.12. Bottom panel: Same substitution effects estimated by the linear model. The partial derivative is 10.44.

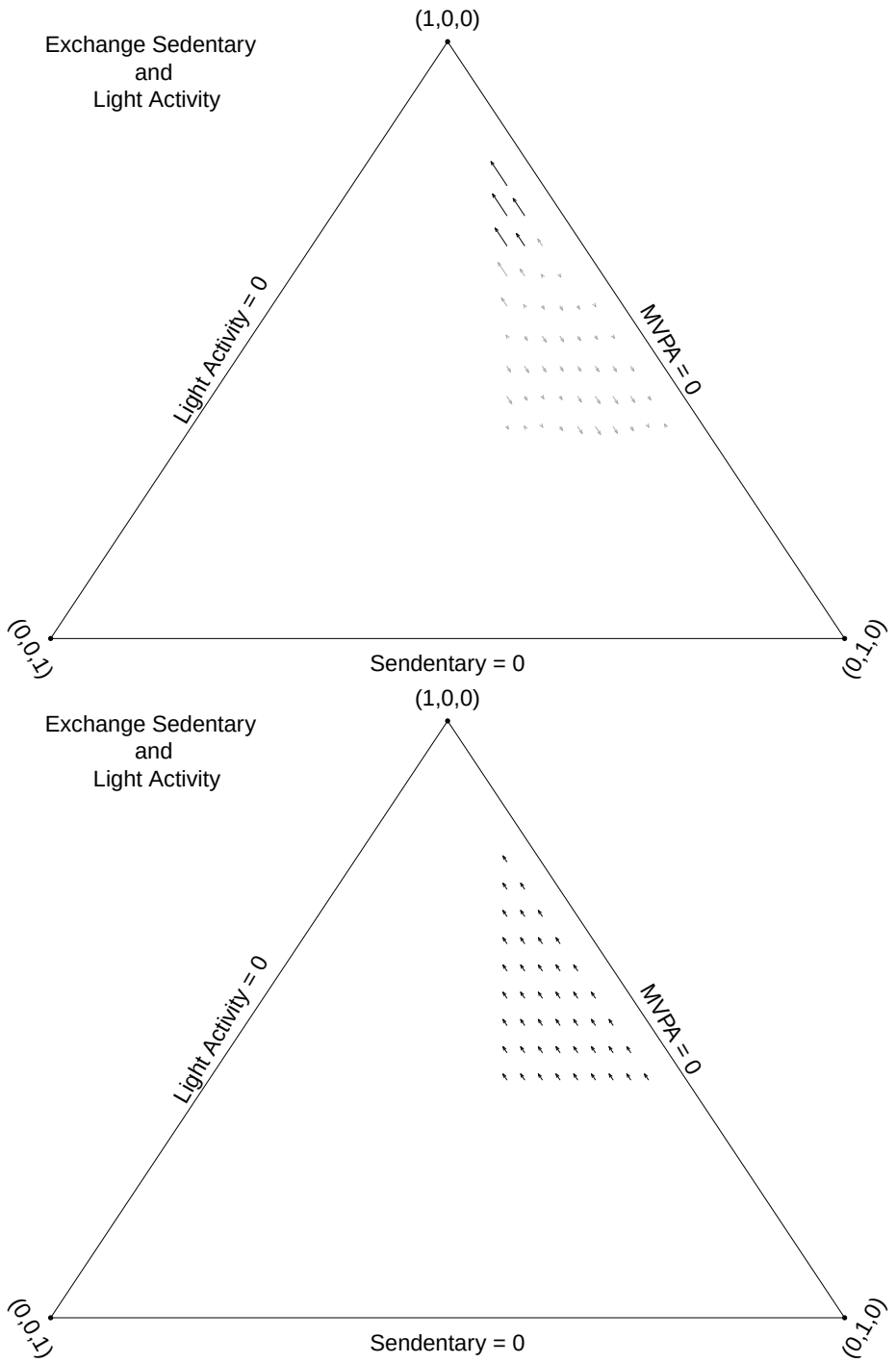

Figure 3.9: Top panel: Estimated effects of exchanging proportions of sedentary time and light activity time while holding MVPA, sleep & non-wear time and baseline covariates constant. The maximum partial derivative is 9.45. Bottom panel: Same substitution effects as estimated by the linear model. The partial derivative is 2.36.

3.4 Discussion

This work developed a novel nonparametric method to estimate the effects of a compositional covariate on a time-to-event outcome. We applied our method to data from the 2003-2006 wave of the National Health and Nutrition Examination Survey (NHANES) with mortality follow-up through December 31st, 2019. These methods estimated the associations between substitutions among aspects of physical activity and sedentary behavior and mortality risk. Our methods improved over previous analysis by showing that the association with a substitution can depend on the starting level.

Our results showed that substituting sedentary time (or sleep & non-wear time) for physical activity results in increased mortality risks among US adults. This finding is consistent with previous research that has shown that sedentary behavior is a risk factor for all-cause mortality [22, 44, 26]. In particular, our method estimated that, among US adults, replacing one hour of sedentary time for physical activity is associated with up to approximately 68% increase in mortality hazards after adjusting for baseline covariates. However, this effect is notable for those engaging in less than four hours of activity per day. Also, substituting 1% of sedentary time for light activity or moderate-to-vigorous activity (MVPA) is associated with approximately 10% and 18% increase in mortality risks respectively.

In interpreting the findings of the analysis on physical activity and mortality, it is important to consider the limitations regarding causality and potential unmeasured confounding. While this work may show an association between physical activity and mortality risks, establishing a causal relationship requires rigorous experimental design. Observational studies like this one are susceptible to unmeasured confounding, meaning there may be unknown factors influencing both physical activity levels and mortality risks. To strengthen the study's conclusions, future research should explore advanced statistical methods and sensitivity analyses to address potential

confounders. Caution should be exercised when inferring causality, and understanding the impact of unmeasured variables is crucial in making informed decisions for public health interventions.

Our approach has an inherent computational limitation due to the so-called curse of dimensionality; a composition with m levels would require one to fit an $m-1$ dimensional nonparametric model. If each dimension used $K = 20$ knots, approximately $20^{(m-1)}$ would need to be estimated. That was feasible for $m = 3$ and a sample size of over 3000, but we would not have been able to fit a fully nonparametric model with $m = 4$. Future work will explore the use of additive models that ignore the highest level of interactions.

CHAPTER 4

CONCLUSION

4.1 Summary

Nonparametric derivative estimation methods are robust to model misspecification as they are able to capture different effects along the same axis of a covariate in a regression model. We have shown in the first part of this dissertation that the penalized spline estimator of derivatives of mean regression functions achieves the optimal L_2 rates of convergence under some usual conditions on the placements of the knots in the spline and the penalty matrix.

Also, in the second part, we have introduced a novel nonparametric method based on a multivariate penalized tensor product spline to model time-to-event data when compositional covariates (i.e. covariates that add up to a constant) are present. We use partial derivatives of the estimated tensor product surface to measure the effects of substituting one compositional covariate for another on the event of interest (for example mortality).

Further, we apply our method to data from the US National Health and Nutrition Examination Survey (NHANES), a representative national survey, to estimate the effects of exchanging sedentary behaviors and physical activity on all-cause mortality. Our analysis shows that while replacing physical activity with sedentary behaviors is generally associated with increased mortality risks, the extent of the association depends vastly on one's current level of physical activity, among US adults.

4.2 Limitations and Future Work

Beyond L_2 asymptotics, one possible direction for the work on the penalized spline derivative estimator is to understand if a different level of penalization would yield a better finite sample performance of the derivative estimator. This is particularly interesting since the estimator is optimized for estimating the mean regression function itself, not its derivatives. One approach is the adaptive penalty methodology by Simpkin in 2013. However, we observed in our simulation study that, this approach may have higher MSE for some functional forms. Even though this is only simulation evidence, it suggests the question may still be open to future research.

In our nonparametric method for modeling time-to-event data with compositional covariates, one limitation is the computational problems that arise for compositions exceeding three variables. As stated in the discussion section of Chapter 3, one approach is to ignore higher-level interactions and use additive models of lower-dimensional smooth functions. We hope to explore this and other directions, including a Bayesian model in the future.

APPENDIX A

PROOF OF TECHNICAL LEMMAS FOR THEOREM 1

A.0.1 Technical Lemmas

Lemma 1. *Let $f \in \mathcal{C}^p$, then there exists $s_f \in \boldsymbol{\mathcal{S}}(q, \underline{\mathbf{t}})$ such that*

$$||f^{(r)} - s_f^{(r)}|| = O(h^{q-r}) + o(h^{p-r})$$

for all $r = 0, 1, \ldots, q-2$ and $p \leq q$.

Here, $b(x) = -\frac{f^{(q)}(x)h_i^q}{q!} B_q\left(\frac{x-t_i}{h_i}\right)$ for $t_i \leq x < t_{i+1}$ where $B_q(.)$ is the q^{th} Bernoulli polynomial defined as $B_0(x) = 1$, and $B_k(x) = \int_0^x kB_{k-1}(x)dx + B_k$

and B_k is chosen such that $\int_0^1 B_k(x)dx = 0$.

B_k is known as the k^{th} Bernoulli number ([3]). This Lemma also appears in [3] and [58] adopts the general result in [3] to prove the case where $p < q$.

Proof of Lemma 1

We provide a proof for the case where $p = q$ and refer to Remark 3.1 of [58] for the case where $p < q$. [58] showed that when $p < q$, $||f^{(r)} - s_f^{(r)}|| = o(h^{p-r})$.
For $p = q$, first note that under Assumption 3, [3] showed that

$$\inf_{s(x) \in \boldsymbol{\mathcal{S}}(q, \underline{\mathbf{t}})} ||f^{(r)}(x) - s^{(r)}(x) + b^{*(r)}(x)||_{L_\infty} = o(h^{q-r})$$

This means, there exists an $s_f(x) \in \boldsymbol{\mathcal{S}}(q, \underline{\mathbf{t}})$ such that

$$||f^{(r)}(x) - s_f^{(r)}(x) + b^{*(r)}(x)|| = o(h^{q-r})$$

where $b^*(x) = -\frac{f^{(q)}(t_i)h_i^q}{q!} B_q\left(\frac{x-t_i}{h_i}\right)$, for $t_i \leq x < t_{i+1}$ and $b^{*(r)}$ is the r^{th} derivative of b^*. Note that $f^{(q)}(x)$ in $b(x)$ is replaced with $f^{(q)}(t_i)$ in $b^*(x)$.

With $p = q$, we have that $f \in \mathcal{C}^q[0,1]$. Therefore, from Taylor's theorem, $f^{(q)}(x) = f^{(q)}(t_i) + o(1)$.

$$\implies b(x) = b^*(x) + o(h^q)$$

The derivative of the Bernoulli polynomial of order k is given by $\boldsymbol{B}'_k(x) = \boldsymbol{B}_{k-1}(x)$ ([3]), it therefore follows that

$$b^{(r)}(x) = b^{*(r)}(x) + o(h^{q-r})$$

for $r = 0, 1, 2, \ldots, q-2$. But $||b^*|| = O(h^q)$ by definition, giving $||b^{(r)}|| = O(h^{q-r})$.

Combining this with the case where $p < q$, we have that $||f^{(r)} - s_f^{(r)}|| = O(h^{q-r}) + o(h^{p-r})$ for all $p \leq q$.

Lemma 2. *Given* $G_q^{(r)} = \int_0^1 B^{(r)}(x) B_q^{(r)}(x) q(x) dx$,

$$||G_q^{(r)}||_\infty = O(h^{-2r})$$

Proof of Lemma 2

Note that $B_q^{(r)}(x) = B_{q-r}(x) D^{(r)}$

$$\begin{aligned} \therefore G_q^{(r)} &= \int_0^1 B_{q-r}(x) D^{(r)} D^{T(r)} B_{q-r}^T(x) q(x) dx \\ &= O(h^{-2r}) \int_0^1 B_{q-r}(x) B_{q-r}^T(x) q(x) dx \\ &= O(h^{-2r}) \times q_{\max} \\ &= O(h^{-2r}) \end{aligned}$$

Where $q_{\max} = \max_{x \in [0,1]} q(x) < \infty$. Also, note that B-spline bases are bounded by 1 $\forall x \in [0,1]$.

Lemma 3. *Let* $G_{n,q} = \boldsymbol{B}^T\boldsymbol{B}/n$ *where* $\boldsymbol{B} = [B(x_1), B(x_2), \ldots, B(x_n)]^T \in \mathbb{R}^{n\times K}$ *is a matrix of basis functions with each* $B(x) \in \mathbb{R}^K$ *being a vector of basis functions of order* q *at* x.

Then

$$||G_{n,q}^{-1}||_\infty = O(h^{-1})$$

Proof of Lemma 3

This Lemma is adapted from Lemma 6.3 of [57] and the key idea is to show that the elements of $G_{n,q}^{-1}$ decay exponentially and of order h^{-1}. We provide the proof here for convenience.

Let $\lambda_{\max}$ and $\lambda_{\min}$ be the maximum and minimum eigenvalues of $G_{n,q}$ respectively. Since $G_{n,q}$ is a band matrix, Theorem 2.2 of [15] is used. First, we need to satisfy the conditions of the theorem.

Note that

$$\begin{aligned} ||\lambda_{\max}^{-1}G_{n,q}||_2 &= \lambda_{\max}^{-1}||G_{n,q}||_2 \\ &= \lambda_{\max}^{-1}\max_{\sum_{i=1}^K z_i^2=1}||G_{n,q}\underline{\boldsymbol{z}}||_2 \\ &\leq 1 \end{aligned}$$

Where the max term in the second equality gives some eigenvalue that is at most $\lambda_{\max}^{-1}$.

Also,

$$\begin{aligned} ||\lambda_{\max}G_{n,q}^{-1}||_2 &= \frac{\lambda_{\max}}{\lambda_{\min}}\left|\left|\lambda_{\min}G_{n,q}^{-1}\right|\right|_2 \\ &\leq \frac{\lambda_{\max}}{\lambda_{\min}} \end{aligned}$$

Lemma 6.2 of [57] provides bounds on the eigenvalues of $G_{n,q}$. In particular, for large n, there exist constants c_1 and c_2 such that

$$c_1 h/2 \leq \lambda_{\min} \leq \lambda_{\max} \leq 2c_2 h$$

Therefore by Theorem 2.2 of [15], there exists constants $c > 0$ and $\gamma \in (0, 1)$ which depend only on c_1, c_2 and q such that:

$$|\lambda_{\max} g_{ij}| \leq c\gamma^{|i-j|} \tag{A.1}$$

where g_{ij} is the (i, j)th element of $G_{n,q}^{-1}$.

From equation (A.1),

$$|g_{ij}| \leq c\lambda_{\max}^{-1}\gamma^{|i-j|} \leq 2(c/c_1)h^{-1}\gamma^{|i-j|}$$

which means that $G_{n,q}^{-1} = O(h^{-1})$. This completes the proof of Lemma 3.

Lemma 4. *Suppose* $\underline{\gamma} = G_{n,q}^{-1}\boldsymbol{B}^T\boldsymbol{f}/n$ *and* $\boldsymbol{P_m}$ *is the penalty matrix for the penalized spline estimator in* (2.3)*,*

then

$$\underline{\gamma}^T \boldsymbol{P_m} \underline{\gamma} = O(1)$$

Proof of Lemma 4

Again, this Lemma is adapted from Lemma 8.4 of [58] which puts a bound on the penalty matrix of the penalized spline estimator. The proof follows closely from the proof in [58] with a bit more clarity.

Observe that we can write

$$\begin{aligned}\underline{\boldsymbol{\gamma}} = G_{n,q}^{-1}\boldsymbol{B}^T\boldsymbol{f}/n &= G_{n,q}^{-1}\boldsymbol{B}^T(\boldsymbol{f}-\boldsymbol{s_f})/n + G_{n,q}^{-1}\boldsymbol{B}^T\boldsymbol{s_f}/n \\ &= G_{n,q}^{-1}\boldsymbol{B}^T(\boldsymbol{f}-\boldsymbol{s_f})/n + \underline{\boldsymbol{\beta}} \\ &= G_{n,q}^{-1}\underline{\boldsymbol{\alpha}} + \underline{\boldsymbol{\beta}}\end{aligned}$$

where $\underline{\boldsymbol{\beta}} = G_{n,q}^{-1}\boldsymbol{B}^T\boldsymbol{s_f}/n$ and $\underline{\boldsymbol{\alpha}} = \boldsymbol{B}^T(\boldsymbol{f}-\boldsymbol{s_f})/n$.

Since $\boldsymbol{P_m}$ is positive semi-definite, we can use the Cauchy-Schwarz inequality defined for an inner product $\langle x, y\rangle_{\boldsymbol{P_m}} = x^T\boldsymbol{P_m}y$ and write:

$$\left(\underline{\boldsymbol{\gamma}}\boldsymbol{P_m}\underline{\boldsymbol{\gamma}}\right)^{\frac{1}{2}} \leq \left(\underline{\boldsymbol{\alpha}}^T G_{n,q}^{-1}\boldsymbol{P_m}G_{n,q}^{-1}\underline{\boldsymbol{\alpha}}\right)^{\frac{1}{2}} + \left(\underline{\boldsymbol{\beta}}^T\boldsymbol{P_m}\underline{\boldsymbol{\beta}}\right)^{\frac{1}{2}} \tag{A.2}$$

By Assumption, $\underline{\boldsymbol{\beta}}^T\boldsymbol{P_m}\underline{\boldsymbol{\beta}} = O(1)$, therefore showing that the first term in A.2 is O(1) completes the proof.

In the following, we use the following matrix relations. Let $A \in \mathbb{R}^{m\times n}$, then

$$\frac{1}{\sqrt{n}}||A||_\infty \leq ||A||_2 \leq \sqrt{m}||A||_\infty \tag{A.3}$$

Also, let $P_m^{\frac{1}{2}}$ be a square symmetric matrix such that $\boldsymbol{P_m} = P_m^{\frac{1}{2}}P_m^{\frac{1}{2}}$.

Observe that

$$
\begin{aligned}
\underline{\boldsymbol{\alpha}}^T G_{n,q}^{-1} \boldsymbol{P_m} G_{n,q}^{-1} \underline{\boldsymbol{\alpha}} &= \left(P_m^{\frac{1}{2}} G_{n,q}^{-1} \underline{\boldsymbol{\alpha}} \right)^T \left(P_m^{\frac{1}{2}} G_{n,q}^{-1} \underline{\boldsymbol{\alpha}} \right) && \text{(A.4)} \\
&= ||P_m^{\frac{1}{2}} G_{n,q}^{-1} \underline{\boldsymbol{\alpha}}||_2^2 && \text{(A.5)} \\
&\leq ||\underline{\boldsymbol{\alpha}}||_2^2 ||P_m^{\frac{1}{2}} G_{n,q}^{-1}||_2^2 && \text{(A.6)} \\
&\leq ||\underline{\boldsymbol{\alpha}}||_2^2 ||P_m^{\frac{1}{2}}||_2^2 ||G_{n,q}^{-1}||_2^2 && \text{(A.7)} \\
&\leq ||\underline{\boldsymbol{\alpha}}||_2^2 ||P_m^{\frac{1}{2}}||_2^2 K ||G_{n,q}^{-1}||_\infty^2 && \text{(A.8)} \\
&= o(h^{2p+2}) O(h^{1-2m}) O(h^{-1}) O(h^{-2}) && \text{(A.9)} \\
&= o(h^{2p-2m}) && \text{(A.10)} \\
&= O(1) && \text{(A.11)}
\end{aligned}
$$

for $p \geq m$. The first and second inequalities are by Cauchy Schwartz inequality, and we have used the matrix identity in (A.3) in the third inequality. Also, We have used the result by [1] for $||\underline{\boldsymbol{\alpha}}||_2^2$ and the assumption that $||\boldsymbol{P_m}||_2 = O(h^{1-2m})$. Finally, we have used Lemma 3 in the third inequality for $||G_{n,q}^{-1}||_\infty$ as well.

APPENDIX B

RATES OF CONVERGENCE FOR LOCAL POLYNOMIAL DERIVATIVE ESTIMATORS

When estimating the r^{th} derivative of the mean regression function with a local polynomial of degree p, several authors ([19, 37]) recommend using odd $p-r$. In this section, we lay out an argument that the naive bandwidth under- or over- smooths when p and $p-r$ have different parities and that only even derivatives can be optimally estimated by the naive estimator. Table B.1 below shows the four (4) potential parity combinations for p and $p-r$. We show next that the naive estimator achieves the optimal rate of convergence when used to estimate $p-r$ only for cases I and IV (where p and $p-r$ have same parity, equivalently, when r is even). Let $\hat{m}_r(x)$ be a p^{th}-degree

		$\boldsymbol{p-r}$	
		odd	even
$\boldsymbol{p}$	odd	I	II
	even	III	IV

Table B.1: Parity combinations of p and $p-r$ when estimating the r^{th} derivative of a mean regression function with p^{th} degree local polynomial regression.

local polynomial estimate of the r^{th} ($r \leq p$) derivative of the mean regression function, $m(x)$ at a point x such that $m^{(p+1)}(\cdot)$ is continuous in a neighborhood of x. Let also h be the bandwidth of $\hat{m}_r(x)$ such that $h = o(n)$ and $nh \to \infty$, then we know from [37] that

$$\text{IMSE}\left(\hat{m}_r(x)\right) = o\left(h^{2(p+1-r)}\right) + O\left(\frac{1}{nh^{2r+1}}\right)$$

for odd $p-r$ and

$$\text{IMSE}\left(\hat{m}_r(x)\right) = o\left(h^{2(p+2-r)}\right) + O\left(\frac{1}{nh^{2r+1}}\right)$$

for even $p-r$.

Note that the naive estimator uses the optimal bandwidth when estimating $m(\cdot)$ itself, thus, when $r=0$. In the above, IMSE is the integrated mean squared error. First, we will derive the rates of convergence for the optimal bandwidth for the naive estimator ($r=0$) for both the odd p and even p cases. We will then compare how these naive rates of convergence compare with the optimal bandwidths for estimating $p-v$ for both parity scenarios.

For odd p (thus, $r=0$ and $p-r$ is odd),

$$\text{IMSE}\left(\hat{m}_0(x)\right) = o\left(h^{2(p+1)}\right) + O\left(\frac{1}{nh}\right)$$

To get the rate of convergence of the optimal bandwidth, we derive the h that minimizes the IMSE (ignoring constants).

From:

$$\begin{aligned} 2(p+1)h^{2p+1} - n^{-1}h^{-2} &= 0 \\ 2(p+1)h^{2p+1} &= \frac{1}{nh^2} \\ h^{2p+3} &= \frac{n^{-1}}{2(p+1)} \end{aligned}$$

$\therefore h^o_{naive} = O\left(n^{-\frac{1}{2p+3}}\right)$. Here, we use h^o_{naive} for the optimal bandwidth for the naive estimator when p is odd.

For even p (thus, $r = 0$ and $p - r$ is even),

$$\text{IMSE}\left(\hat{m}_0(x)\right) = o\left(h^{2(p+2)}\right) + O\left(\frac{1}{nh}\right)$$

From:

$$\begin{aligned} 2(p+2)h^{2p+3} - n^{-1}h^{-2} &= 0 \\ 2(p+2)h^{2p+3} &= \frac{1}{nh^2} \\ h^{2p+5} &= \frac{n^{-1}}{2(p+2)} \end{aligned}$$

$\therefore h^e_{naive} = O\left(n^{-\frac{1}{2p+5}}\right)$. h^e_{naive} is the optimal bandwidth for the naive estimator when p is even.

We now analyse the achieved rates of convergence for estimating the r^{th} derivative of the mean regression function, m and how those rates compare with the naive estimator. We consider the four (4) cases in Table B.1 above.

Case I: p **odd and** $p - r$ **odd (thus,** r **is even).**

$$\text{IMSE}\left(\hat{m}_r(x)\right) = o\left(h^{2(p+1-r)}\right) + O\left(\frac{1}{nh^{2r+1}}\right)$$

From

$$\begin{aligned} 2(p+1-r)h^{2p-2r+1} - (2r+1)n^{-1}h^{-2r-2} &= 0 \\ 2(p+1-r)h^{2p-2r+1} &= \frac{2r+1}{nh^{2r+2}} \\ h^{2p+3} &= \frac{2r+1}{2(p+1-r)n} \end{aligned}$$

$\therefore h_{opt} = O\left(n^{-\frac{1}{2p+3}}\right)$, this is the same rate achieved by h^o_{naive}. Therefore, the naive bandwidth achieves the same rate as the optimal bandwidth for estimating $p - r$ in this case.

Case II: p **odd and** $p-r$ **even (thus,** r **is odd).**

$$\text{IMSE}\left(\hat{m}_r(x)\right) = o\left(h^{2(p+2-r)}\right) + O\left(\frac{1}{nh^{2r+1}}\right)$$

By similar approach as in Case I, we get $h_{opt} = O\left(n^{-\frac{1}{2p+5}}\right)$, this rate is different from that achieved by the naive estimator h^o_{naive} for odd p. The consequence of using the naive bandwidth in this case is that, it shrinks faster than the optimal rate which may result in over-smoothing.

Case III: p **even and** $p-r$ **odd (thus,** r **is odd).**

$$\text{IMSE}\left(\hat{m}_r(x)\right) = o\left(h^{2(p+1-r)}\right) + O\left(\frac{1}{nh^{2r+1}}\right)$$

Again, similar to Cases I and II above, $h_{opt} = O\left(n^{-\frac{1}{2p+3}}\right)$, this rate is different from that achieved by the naive estimator h^e_{naive} for even p which is $O\left(n^{-\frac{1}{2p+5}}\right)$. Unlike in case II, the consequence of using the naive bandwidth in this case is that, it shrinks at a slower rate than the optimal rate which may result in over-smoothing.

Case IV: p **even and** $p-r$ **even (thus,** r **is even).**

$$\text{IMSE}\left(\hat{m}_r(x)\right) = o\left(h^{2(p+2-r)}\right) + O\left(\frac{1}{nh^{2r+1}}\right)$$

From

$$\begin{aligned} 2(p+2-r)h^{2p-2r+3} - (2r+1)n^{-1}h^{-2r-2} &= 0 \\ 2(p+2-r)h^{2p-2r+3} &= \frac{2r+1}{nh^{2r+2}} \\ h^{2p+5} &= \frac{2r+1}{2(p+2-r)n} \end{aligned}$$

$\therefore h_{opt} = O\left(n^{-\frac{1}{2p+5}}\right)$, this is the same rate achieved by h^e_{naive} for even p. Therefore, the naive bandwidth achieves the same rate as the optimal bandwidth for estimating

$p - r$ in this case. Thus, the naive estimator can only optimally estimate even-order derivatives for Local Polynomial Regression.

BIBLIOGRAPHY

[1] Agarwal, Girdhar G., and Studden, W. J. Asymptotic integrated mean square error using least squares and bias minimizing splines. *The Annals of Statistics 8*, 6 (1980), 1307–1325.

[2] Aitchison, John. *The Statistical Analysis of Compositional Data*. Chapman & Hall, 1986.

[3] Barrow, D.L., and Smith, P.W. Efficient L^2 approximation by splines. *Numerische Mathematik 33* (1979), 101–114.

[4] Binder, David A. Fitting cox's proportional hazards models from survey data. *Biometrika 79*, 1 (1992), 139–147.

[5] Booth, Frank W, Roberts, Christian K, and Laye, Matthew J. Lack of exercise is a major cause of chronic diseases. *Comprehensive Physiology 2*, 2 (2012), 1143–1211.

[6] Charnigo, Richard, Hall, Benjamin, and Srinivasan, Cidambi. A generalized cp criterion for derivative estimation. *Technometrics 53*, 3 (2011), 238–253.

[7] Chaudhuri, Probal, and Marron, J. S. Sizer for exploration of structures in curves. *Journal of the American Statistical Association 94*, 447 (1999), 807–823.

[8] Claeskens, Gerda, Krivobokova, Tatyana, and Opsomer, Jean D. Asymptotic properties of penalized spline estimators. *Biometrika 96*, 3 (09 2009), 529–544.

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

[10] Craven, Peter, and Wahba, Grace. Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik 31*, 4 (Dec. 1978), 377–403.

[11] Dai, Wenlin, Tong, Tiejun, and Genton, Marc G. Optimal estimation of derivatives in nonparametric regression. *Journal of Machine Learning Research 17*, 164 (2016), 1–25.

[12] de Boor, C. *A Practical Guide to Splines*. Applied Mathematical Sciences. Springer New York, 1978.

[13] De Brabanter, Kris, Brabanter, Jos De, Suykens, Johan A.K., and Moor, Bart De. Kernel regression in the presence of correlated errors. *Journal of Machine Learning Research 12*, 55 (2011), 1955–1976.

[14] De Brabanter, Kris, De Brabanter, Jos, De Moor, Bart, and Gijbels, Irène. Derivative estimation with local polynomial fitting. *J. Mach. Learn. Res. 14*, 1 (Jan. 2013), 281–301.

[15] Demko, Stephen. Inverses of band matrices and local convergence of spline projections. *SIAM Journal on Numerical Analysis 14*, 4 (1977), 616–619.

[16] Eilers, Paul H. C., and Marx, Brian D. Flexible smoothing with b -splines and penalties. *Statist. Sci. 11*, 2 (05 1996), 89–121.

[17] Eilers, Paul H. C., and Marx, Brian D. Splines, knots, and penalties. *WIREs Computational Statistics 2*, 6 (2010), 637–653.

[18] Eubank, R. L., and Speckman, P. L. Confidence bands in nonparametric regression. *Journal of the American Statistical Association 88*, 424 (1993), 1287–1301.

[19] Fan, J., and Gijbels, I. *Local Polynomial Modelling and Its Applications: Monographs on Statistics and Applied Probability 66*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1996.

[20] Fellner, William H. Robust estimation of variance components. *Technometrics 28*, 1 (1986), 51–60.

[21] Fisher, Jonathan D., Johnson, David S., Smeeding, Timothy M., and Thompson, Jeffrey P. Estimating the marginal propensity to consume using the distributions of income, consumption, and wealth. *Journal of Macroeconomics 65* (2020), 103218.

[22] Fishman, Ezra I, Steeves, Jeremy A, Zipunnikov, Vadim, Koster, Annemarie, Berrigan, David, Harris, Tamara A, and Murphy, Rachel. Association between objectively measured physical activity and mortality in nhanes. *Medicine and Science in Sports and Exercise 48*, 7 (2016), 1303–1311.

[23] Gray, Robert J. Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *Journal of the American Statistical Association 87*, 420 (1992), 942–951.

[24] Hildenbrand, Werner. Facts and ideas in microeconomic theory. *European Economic Review 33*, 2 (1989), 251–276.

[25] Härdle, W., Hart, J., Marron, J. S., and Tsybakov, A B. Bandwidth choice for average derivative estimation. *Journal of the American Statistical Association 87*, 417 (1992), 218–226.

[26] Katzmarzyk, PT, Church, TS, Craig, CL, and Bouchard, C. Sitting time and mortality from all causes, cardiovascular disease, and cancer. *Medicine and Science in Sports and Exercise 41*, 5 (2009), 998–1005.

[27] Lee, I-Min, Shiroma, Eric J, Lobelo, Felipe, Puska, Pekka, Blair, Steven N, Katzmarzyk, Peter T, and Lancet Physical Activity Series Working Group. Effect of physical inactivity on major non-communicable diseases worldwide: an analysis of burden of disease and life expectancy. *The Lancet 380*, 9838 (2012), 219–229.

[28] Mekary, Rania A, Willett, Walter C, Hu, Frank B, and Ding, Eric L. Isotemporal substitution paradigm for physical activity epidemiology and weight change. *American Journal of Epidemiology 170*, 4 (2009), 519–527.

[29] Müller, Hans-Georg. *Nonparametric Regression Analysis of Longitudinal Data.* Lecture Notes in Statistics: 46. Springer New York, 1988.

[30] Müller, Hans-Georg, Stadtmüller, U., and Schmitt, Thomas. Bandwidth choice and confidence intervals for derivatives of noisy data. *Biometrika 74*, 4 (1987), 743–749.

[31] National Center for Health Statistics Division of Analysis and Epidemiology. NHANES Public-Use Linked Mortality Files (2019), 2019. Available from https://www.cdc.gov/nchs/data-linkage/mortality-public.htm.

[32] National Health and Nutrition Examination Survey. NHANES Public-Use Linked Mortality Files (2011), 2011. Available from https://www.cdc.gov/nchs/data_access/data_linkage_activities.htm.

[33] Park, Cheolwoo, and Kang, Kee-Hoon. Sizer analysis for the comparison of regression curves. *Computational Statistics & Data Analysis 52* (04 2008), 3954–3970.

[34] Physical Activity Guidelines Advisory Committee. 2018 Physical Activity Guidelines Advisory Committee Scientific Report. Tech. rep., U.S. Department of Health and Human Services, Washington, DC, 2018.

[35] Ramsay, James O., and Silverman, Bernard W. *Applied Functional Data Analysis: Methods and Case Studies.* 2002.

[36] Ruppert, D., Sheather, S. J., and Wand, M. P. An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association 90*, 432 (1995), 1257–1270.

[37] Ruppert, D., and Wand, M. P. Multivariate Locally Weighted Least Squares Regression. *The Annals of Statistics 22*, 3 (1994), 1346 – 1370.

[38] Ruppert, David, Wand, M. P., and Carroll, R. J. *Semiparametric Regression.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2003.

[39] Schall, Robert. Estimation in generalized linear models with random effects. *Biometrika 78*, 4 (1991), 719–727.

[40] Schumaker, Larry. *Spline Functions: Basic Theory*, 3 ed. Cambridge Mathematical Library. Cambridge University Press, 2007.

[41] Schuna Jr, John M, Johnson, William D, and Tudor-Locke. Adult self-reported and objectively monitored physical activity and sedentary behavior: Nhanes 2005–2006. *International Journal of Behavioral Nutrition and Physical Activity 10*, 1 (2013), 1–10.

[42] Silverman, B. W. Spline smoothing: The equivalent variable kernel method. *The Annals of Statistics 12*, 3 (1984), 898–916.

[43] Simpkin, Andrew, and Newell, John. An additive penalty p-spline approach to derivative estimation. *Comput. Stat. Data Anal. 68* (Dec. 2013), 30–43.

[44] Stamatakis, Emmanuel, Rogers, Kris, Ding, Ding, Berrigan, David, Chau, Josephine, Mark, Hamer, and Bauman, A. All-cause mortality effects of replacing sedentary time with physical activity and sleeping using an isotemporal substitution model: a prospective study of 201,129 mid-aged and older adults. *International Journal of Behavioral Nutrition and Physical Activity 12*, 1 (2015), 1.

[45] Stone, Charles J. Optimal global rates of convergence for nonparametric regression. *Ann. Statist. 10*, 4 (12 1982), 1040–1053.

[46] Troiano, R. P., Berrigan, D., Dodd, K. W., Masse, L. C., Tilley, T., and McDowell, M. Physical activity in the united states measured by accelerometer. *Medicine & Science in Sports & Exercise 40*, 1 (2008), 181–8.

[47] Wahba, Grace. A Comparison of GCV and GML for Choosing the Smoothing Parameter in the Generalized Spline Smoothing Problem. *The Annals of Statistics 13*, 4 (1985), 1378 – 1402.

[48] Wahba, Grace, and Wang, Yonghua. When is the optimal regularization parameter insensitive to the choice of the loss function? *Communications in Statistics - Theory and Methods 19*, 5 (1990), 1685–1700.

[49] Wand, Matt P, and Jones, M Chris. *Kernel smoothing*. Chapman and Hall/CRC, 1994.

[50] Warburton, Darren E.R., Nicol, Cally W., and Bredin, Shannon S.D. Health benefits of physical activity: the evidence. *Canadian Medical Association Journal 174*, 6 (2006), 801–809.

[51] Whitehead, John. Fitting cox's regression model to survival data using glim. *Journal of the Royal Statistical Society. Series C (Applied Statistics) 29*, 3 (1980), 268–275.

[52] Willett, W., and Stampfer, M.J. Total energy intake: implications for epidemiologic analyses. *American Journal of Epidemiology 124*, 1 (July 1986), 17–27.

[53] Wood, Simon N. *Generalizd Additive Models, second edition*. Texts in Statistical Science. CRC Press, 2017.

[54] Wood, Simon N. *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC, 2017.

[55] Wood, Simon N., and Fasiolo, Matteo. A generalized fellner-schall method for smoothing parameter optimization with application to tweedie location, scale and shape models. *Biometrics 73*, 4 (2017), 1071–1081.

[56] Wood, Simon N., Pya, Natalya, and Säfken, Benjamin. Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association 111*, 516 (2016), 1548–1563.

[57] X. Shen, D. A. Wolfe, and Zhou, S. Local asymptotics for regression splines and confidence regions. *The Annals of Statistics 26*, 5 (1998), 1760–1782.

[58] Xiao, Luo. Asymptotic theory of penalized splines. *Electron. J. Statist. 13*, 1 (2019), 747–794.

[59] Xiao, Luo, Li, Yingxing, Apanasovich, Tatiyana V., and Ruppert, David. Local Asymptotics of P-splines. *arXiv e-prints* (Jan. 2012), arXiv:1201.0708.

[60] Zhou, Shanggang, and Wolfe, Douglas. On derivative estimation in spline regression. *Statistica Sinica 10* (01 2000), 93–108.