12-2016

Functional regression models in the frame work of reproducing kernel Hilbert space

Simeng Qu
Purdue University

Graduate School Form
30 Updated 12/26/2015

PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By Simeng Qu

Entitled

FUNCTIONAL REGRESSION MODELS IN THE FRAME WORK OF REPRODUCING KERNEL HILBERT SPACE

For the degree of Doctor of Philosophy

Is approved by the final examining committee:

Xiao Wang
Chair

Mary Ellen Bock

Chuanhai Liu

Lingsong Zhang

To the best of my knowledge and as understood by the student in the Thesis/Dissertation Agreement, Publication Delay, and Certification Disclaimer (Graduate School Form 32), this thesis/dissertation adheres to the provisions of Purdue University's "Policy of Integrity in Research" and the use of copyright material.

Approved by Major Professor(s): Xiao Wang

Approved by: Hao Zhang 11/29/2016

Head of the Departmental Graduate Program Date

FUNCTIONAL REGRESSION MODELS IN THE FRAME WORK OF REPRODUCING KERNEL HILBERT SPACE

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Simeng Qu

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

December 2016

Purdue University

West Lafayette, Indiana

For my family.

ACKNOWLEDGMENTS

First of all, I would like to thank my advisor Dr. Xiao Wang, for his unending support and guidance over the years, for all the opportunities he has provided, and for constantly motivating and pushing me to achieve more. I would also like to thank the other members of my thesis committee: Dr. Mary Ellen Bock, Dr. Chuanhai Liu and Dr. Lingsong Zhang, for your advice and encouragement. I have learned a great deal from all of you and would not be where I am today without all your help.

I would also like to thank Dr. Jane-Ling Wang. Though somewhat short, it was a fruitful collaboration, and I truly enjoyed working with you and your group.

To all my dear colleagues, I thank you for constantly helping and supporting me, at work and in life. Yixuan Qiu, your technical and computing knowledge is boundless, and I am glad you were happy to share it with me. Chen Chen, Yaowu Liu, Bing Yu, Qi Wang and Rongrong Zhang, I will miss working with you all on homework and projects, sitting with you all in lectures, and most importantly, having fun with you all. Libo Wang, Kelly Ann Dixon, and Longjie Cheng, you were great officemates, and I appreciate all your advice and suggestions. And to my members of my group, Yixi Xu, Yao Chen, and Shuang He, it was truly a pleasure working with you all.

And last but not least, I would like to acknowledge my parents for their unconditional love.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

SYMBOLS

$\circ$	Element-wise product
$\otimes$	Kronecker product
$\|\| \cdot \|\|_0$	$\mathcal{L}_0$ norm or the number of non-zero elements in a vector
$\|\| \cdot \|\|_2$	$\mathcal{L}_2$ norm or Euclidean norm
$\|\| \cdot \|\|_K$	Norm associated with reproducing kernel Hilbert space and reproducing kernel K
F_m^{-1}	Inverse cumulative distribution function of the χ-square distribution with degree of freedom m
$\mathcal{H}(K)$	Reproducing kernel Hilbert space with reproducing kernel K

ABBREVIATIONS

ACE	Alternating conditional expectation
ADMM	Alternating Direction Method of Multipliers
ECG	Electrocardiogram
EEG	Electroencephalogram
FDA	Functional Data Analysis
fMRI	Functional Magnetic Resonance Imaging
GCV	Generalized Cross Validation
GSCAD	Grouped Smoothly Clipped Absolute Deviation
MSE	Mean Squared Error
OMP	Orthogonal Matching Pursuit
RKHS	Reproducting kernel Hilbert space

ABSTRACT

Qu, Simeng PhD, Purdue University, December 2016. Functional Regression Models in the Frame Work of Reproducing Kernel Hilbert Space . Major Professor: Xiao Wang.

The aim of this thesis is to systematically investigate some functional regression models for accurately quantifying the effect of functional predictors. In particular, three functional models are studied: functional linear regression model, functional Cox model, and function-on-scalar model. Both theoretical properties and numerical algorithms are studied in depth. The new models find broad applications in many areas.

For the functional linear regression model, the focus is on testing the nullity of the slope function, and a generalized likelihood ratio test based on easily implementable data-driven estimate is proposed. The quality of the test is measured by the minimal distance between the null and the alternative space that still allows a possible test. The lower bound of the minimax decay rate of this distance is derived, and test with a distance that decays faster than the lower bound would be impossible. It is shown that the minimax optimal rate is jointly determined by the reproducing kernel and the covariance kernel and our test attains this optimal rate. Later, the test is applied to the effect of the trajectories of oxides of nitrogen (NOx) on the level of ozone (O3).

In the functional Cox model, the aim is to study the Cox model with right-censored data in the presence of both functional and scalar covariates. Asymptotic properties of the maximum partial likelihood estimator is established and it is shown that the estimator achieves the minimax optimal rate of convergence under a weighted L2-risk. Implementation of the estimation approach and the selection of the smoothing parameter are discussed in detail. The finite sample performance is illustrated by simulated examples and a real application.

The function-on-scalar model concentrates on developing the simultaneous model selection and estimation technique. A novel regularization method called the Grouped Smoothly Clipped Absolute Deviation (GSCAD) is proposed. The initial problem can be transferred into a dictionary learning problem, where the GSCAD can be directly applied to simultaneously learn a sparse dictionary and select the appropriate dictionary size. Efficient algorithm is designed based on the alternative direction method of multipliers (ADMM) which decomposes the joint non-convex problem with the non-convex penalty into two convex optimization problems. Several examples are presented for image denoising and image inpainting, which are competitive with the state of the art methods.

1. INTRODUCTION

1.1 Functional Data

Functional data refer to data in form of functions such as curves, surfaces or more general objects, where a sample element is considered to be a function. The concept of functional data can be broad. Traditional functional data can be described as observations of trajectories at discrete points along time line (or more general continuum), where the trajectories are generated from underlying smooth stochastic process. They typically consist of a random sample of independent real-valued functions, $X_1(t), ..., X_n(t)$, on a compact interval $I = [0, T]$ on the real line. These real-valued functions can be viewed as the realizations of a one-dimensional stochastic process $X(t)$. This type of functional data is also referred to as the first general functional data [1]. Typical examples include children's growth curves, daily temperature and precipitation records. Figure 1.1 shows the mean monthly temperature curves for four selected Canadian weather stations. Functional data is also very common in various medical and biomedical fields. These data can take fairly simple forms, such as 2-dimensional electrocardiogram (ECG) and electroencephalogram (EEG) traces, or be highly complex, like functional magnetic resonance imaging data (fMRI). Such functional data are also referred to as the next generation functional data, that are part of complex data objects, and possibly are multivariate, correlated, or involve images or shapes.

As modern technology produces increasingly larger volumes of functional data with higher quality, demand for more powerful and sophisticated statistical methods is growing rapidly.

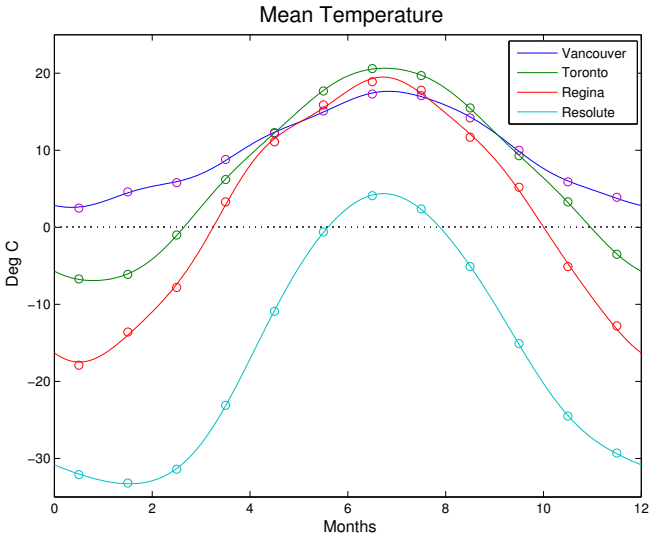


Figure 1.1. Mean monthly temperatures for four selected Canadian weather stations.

1.2 Reproducing Kernel Hilbert Space

Reproducing kernel Hilbert space (RKHS) has been an important tool to studying functional data. In this section, we will introduce some basic concepts of RKHS and list a few properties that we will use in later sections. More details about RKHS can be found in [2] and [3].

1.2.1 Definition

RKHS is a Hilbert space of functions in which point evaluation is a continuous linear functional. That is, if two functions f and g in the RKHS are close in norm, i.e., $||f-g||$ is small, then f and g are also pointwise close, i.e., $|f(x)-f(g)|$ is small for all x. The reverse may not be true. Definition of RKHS is described as follows and we will give out two examples of RKHS later in section 1.2.3.

Definition 1.1 *A reproducing kernel Hilbert space is a Hilbert space $\mathcal{H}$ of functions on domain $\mathcal{X}$, such that for each $x \in \mathcal{X}$, the evaluation function $L_x:\ L_x f = f(x)$, is a bounded linear functional. The boundedness means that there exists an $M = M_x$, such that*

$$|L_x f| = |f(x)| \leq M||f||, \text{ for all } f \in \mathcal{H}, \tag{1.1}$$

where $||\cdot||$ is the norm in the Hilbert space.

The condition of L_x being bounded is equivalent to that of L_x being continuous in $\mathcal{H}$, and some references also define RKHS by L_x being continuous. By the Riesz representation theorem of Hilbert space, for every $x \in \mathcal{X}$, there exists an element $K_x \in \mathcal{H}$ with the property that

$$L_x f = \langle K_x, f\rangle = f(x), \quad \forall f \in \mathcal{H}.$$

K_x is called the representer of evaluation at x. Here $\langle\cdot,\cdot\rangle$ denote the inner product of $\mathcal{H}$. The symmetric bivariate function $K(x,y) = \langle K_x, K_y\rangle$ is called the reproducing kernel of the space $\mathcal{H}$ as it has the reproducing property that,

$$\langle K(x,\cdot), f(\cdot)\rangle = f(x) \quad \forall x \in \mathcal{X}, \text{ and } \forall f \in \mathcal{H}.$$

In fact $K(x, y)$ is a non-negative definite function, and there is a one-to-one correspondence between reproducing kernel Hilbert spaces and non-negative definite functions.

Theorem 1.2.1 *For every RKHS $\mathcal{H}$ of functions on $\mathcal{X}$, there corresponds a unique non-negative definite reproducing kernel $K(x, y)$; conversely, given a non-negative definite function K on $\mathcal{X} \times \mathcal{X}$, we construct a unique RKHS $\mathcal{H}$, that has $K(x, y)$ as its reproducing kernel.*

1.2.2 Useful Properties

Suppose reproducing kernel $K(x, y)$ is continuous and satisfying

$$\int_{\mathcal{X}} \int_{\mathcal{X}} K^2(x, y) dx dy < \infty. \tag{1.2}$$

By Mercer theorem [4], there exist eigenvalues $\lambda_1 \geq \lambda_2 \geq ... \geq 0$, and an orthonormal sequence of continuous eigenfunctions $\phi_1, \phi_2, ...$ in the $\mathcal{L}_2$ space whose elements are functions defined on $\mathcal{X}$, such that

$$\int_{\mathcal{X}} K(x, y)\phi_v(y) dy = \lambda_v \phi_v(x), \quad v = 1, 2, ...,$$

$$K(x, y) = \sum_{v=1}^{\infty} \lambda_v \phi_v(x)\phi_v(y),$$

$$\int_{\mathcal{X}} \int_{\mathcal{X}} K^2(x, y) dx dy = \sum_{v=1}^{\infty} \lambda_v^2 < \infty.$$

Then it is easy to verify the following proposition.

Proposition 1.2.1 *Suppose (1.2) holds. Let $f_v = \int_{\mathcal{X}} f(x)\phi_v(x)dx$, then $f \in \mathcal{H}(K)$ if and only if*

$$\sum_{v=1}^{\infty} \frac{f_v^2}{\lambda_v} < \infty,$$

and

$$||f||_K^2 = \sum_{v=1}^{\infty} \frac{f_v^2}{\lambda_v}.$$

Here $|| \cdot ||_K$ denote the norm defined by RKHS $\mathcal{H}(K)$.

Proposition 1.2.1 shows that, if we begin with K satisfying 1.2, we can construct an RKHS of functions as

$$\{f|\ f(\cdot) = \sum_{v=1}^{\infty} f_v \phi_v(\cdot) \ \text{ and } \ \sum_{v=1}^{\infty} \frac{f_v^2}{\lambda_v} < \infty\}.$$

The following theorem shows that RKHS can be decomposed into tensor sums.

Proposition 1.2.2 *If the reproducing kernel K of a RKHS $\mathcal{H}$ on domain $\mathcal{X}$ can be decomposed into $K = K_0 + K_1$, where K_0 and K_1 are both non-negative definite, $K_0(x,\cdot),\ K_1(x,\cdot) \in \mathcal{H}$, for every $x \in \mathcal{X}$, and $\langle K_0(x,\cdot), K_1(x,\cdot)\rangle = 0$, for every $x, y \in \mathcal{X}$, then the spaces $\mathcal{H}_0$ and $\mathcal{H}_1$ corresponding respectively to K_0 and K_1 form a tensor sum decomposition of $\mathcal{H}$. Conversely, if K_0 and K_1 are both non-negative definite and $\mathcal{H}_0 \cap \mathcal{H}_1 = \{0\}$, then $\mathcal{H} = \mathcal{H}_0 + \mathcal{H}_1$ has a reproducing kernel $K = K_0 + K_1$.*

Proposition 1.2.2 will make constructing estimators for coefficient functions in functional regression models a lot easier, as will be presented in later sections. In general, we assume the coefficient functions (denoted as β) reside in a RHKS $\mathcal{H}(K)$, and $\mathcal{H}(K)$ can be decomposed according to a penalty function J, which is applied to control the smoothness of β. More specifically, $\mathcal{H}(K) = \mathcal{H}_0(K_0) + \mathcal{H}_1(K_1)$, where $\mathcal{H}_0$ is the null space of J,

$$\mathcal{H}_0 = \{\beta \in \mathcal{H}(K) : J(\beta) = 0\},$$

and $\mathcal{H}_1$ is its orthogonal complement in $\mathcal{H}$. Then coefficient function β can be represented by a finite set of basis consisting basis of K_0 and inner products of K_1 and observed predictor functions. In this case, the infinite target space $\mathcal{H}$ has been reduced to a subspace spanned by a finite set of basis.

1.2.3 Examples of RKHS

Before introducing any examples, I would like to point out that the familiar Hilbert space $\mathcal{L}_2[0,1]$ of square integrable functions on $[0,1]$ is not a RKHS as it does not satisfy condition (1.1). In fact, elements in $\mathcal{L}_2[0,1]$ are not even defined point-wise.

A finite-dimensional Hilbert space, on the other hand, is always a reproducing kernel Hilbert space since all linear functionals are continuous.

Consider the continuous function space $C^{(m)}[0,1]$ defined as

$$C^{(m)}[0,1] = \{f : [0,1] \to \mathbb{R} | f, f', ..., f^{(m-1)} \text{ are absolutely continuous and } f^{(m)} \in \mathcal{L}_2[0,1]\}.$$

I am going to introduce two inner products, equipped with either of which, space $C^{(m)}[0,1]$ becomes a RKHS.

A nature way to construct a RKHS on space $C^{(m)}[0,1]$ is based on taylor expansion. For $f \in C^{(m)}[0,1]$, Taylor expansion gives

$$f(x) = \sum_{v=0}^{m-1} \frac{x^v}{v!} f^{(v)}(0) + \int_0^1 \frac{(x-u)_+^{m-1}}{(m-1)!} f^{(m)}(u) du,$$

where $(\cdot)_+ = \max(0, \cdot)$.

Example 1.2.1 *If we define inner product of* $C^{(m)}[0,1]$ *as*

$$\langle f, g \rangle = \sum_{v=0}^{m-1} f^{(v)}(0) g^{(v)}(0) + \int_0^1 f^{(m)}(x) g^{(m)}(x) dx, \quad f, g \in C^{(m)}[0,1],$$

then $C^{(m)}[0,1]$ *becomes an RKHS with kernel*

$$K(x,y) = \sum_{v=0}^{m-1} \frac{x^v}{v!} \frac{y^v}{v!} + \int_0^1 \frac{(x-u)_+^{m-1}}{(m-1)!} \frac{(y-u)_+^{m-1}}{(m-1)!} du.$$

To check this, using the fact that $K_x^{(v)}(0) = x^{(v)}/v!$, $v = 1, ..., m-1$, $K_x^{(m)}(y) = (x-y)_+^{m-1}/(m-1)!$, therefore

$$\begin{aligned} \langle K(x,y), f \rangle &= \sum_{v=0}^{m-1} \frac{x^v}{v!} f^{(v)}(0) + \int_0^1 \frac{(x-u)_+^{m-1}}{(m-1)!} f^{(m)}(u) du \\ &= f(x). \end{aligned}$$

Write K as $K = K_0 + K_1$, with

$$K_0(x,y) = \sum_{v=0}^{m-1} \frac{x^v}{v!} \frac{y^v}{v!},$$

and

$$K_1(x,y) = \int_0^1 \frac{(x-u)_+^{m-1}}{(m-1)!}\frac{(y-u)_+^{m-1}}{(m-1)!}du.$$

K_0 corresponds to a polynomial spaces $\mathcal{H}_0$,

$$\mathcal{H}_0 = \{f : f^{(m)} = 0\},$$

with inner product

$$\langle f, g\rangle_0 = \sum_{v=0}^{m-1} f^{(v)}(0)g^{(v)}(0),$$

and K_1 associates with its orthogonal complement $\mathcal{H}_1$

$$\mathcal{H}_1 = \{f : f^{(v)}(0) = 0,\ v = 1, ..., m-1,\ \int_0^1 (f^{(m)})^2 dx < \infty\},$$

with inner product

$$\langle f, g\rangle_1 = \int_0^1 f^{(m)}g^{(m)}dx.$$

Since K_0 and K_1 are both non-negative definite and $\mathcal{H}_0 \cap \mathcal{H}_1 = \{0\}$, by Proposition 1.2.2, we can decompose $\mathcal{H}(K) = \mathcal{H}_0(K_0) + \mathcal{H}_1(K_1)$.

When $m = 1$, $K_0(x,y) = 1$ and

$$K_1(x,y) = \int_0^1 I_{[u<x]}I_{[u<y]}du = x \wedge y,$$

where $x \wedge y = \min(x,y)$. When $m = 2$, $K_0(x,y) = 1 + xy$ and

$$\begin{aligned} K_1(x,y) &= \int_0^1 (x-u)_+(y-u)_+du \\ &= (x\wedge y)^2(3(x\vee y) - (x\wedge y))/6, \end{aligned}$$

where $x \vee y = \max(x,y)$.

We can also construct another RKHS on $C^{(m)}[0,1]$ by assigning it a different inner product.

Example 1.2.2 *If we define inner product of $C^{(m)}[0,1]$ as*

$$\langle f, g\rangle = \sum_{v=0}^{m-1}(\int_0^1 f^{(v)}dx)(\int_0^1 g^{(v)}dx) + \int_0^1 f^{(m)}(x)g^{(m)}(x)dx, \quad \forall f, g \in C^{(m)}[0,1], \tag{1.3}$$

then $C^{(m)}[0,1]$ is an RKHS.

We now decompose $\mathcal{H}$ as $\mathcal{H} = \mathcal{H}_0 + \mathcal{H}_1$ and obtain its reproducing kernel in form of $K = K_0 + K_1$. Define $\mathcal{H}_0 = \{f : f^{(m)} = 0\}$ with inner product

$$\langle f, g\rangle_0 = \sum_{v=0}^{m-1} (\int_0^1 f^{(v)} dx)(\int_0^1 g^{(v)} dx), \tag{1.4}$$

and let $\mathcal{H}_1$ be

$$\mathcal{H}_1 = \{f : \int_0^1 f^{(v)} dx = 0,\ v = 1, ..., m-1,\ f^{(m)} \in \mathcal{L}_2[0,1]\}, \tag{1.5}$$

with inner product

$$\langle f, g\rangle_1 = \int_0^1 f^{(m)} g^{(m)} dx.$$

Denote

$$k_r(x) = -(\sum_{\mu=-\infty}^{-1} + \sum_{\mu=1}^{\infty}) \frac{\exp(2\pi \mathbf{i}\mu x)}{(2\pi \mathbf{i}\mu)^r}, r = 1, 2, ...,$$

where $\mathbf{i} = \sqrt{-1}$. The k_r functions are actually scaled Bernoulli polynomials, $k_r(x) = B_r(x)/r!$. They are well defined, real-valued and periodic with period 1. Moreover, $k_v, v = 0, ..., m-1$, form an orthonormal basis of $\mathcal{H}_0$ and the reproducing kernel of $\mathcal{H}_0$ under norm (1.4) can be represented as

$$K_0(x, y) = \sum_{v=0}^{m-1} k_v(x) k_v(y).$$

For $\mathcal{H}_1$ in (1.5), its reproducing kernel is given by

$$K_1(x, y) = k_m(x) k_m(y) + (-1)^{m-1} k_{2m}(x - y),$$

and finally, the reproducing kernel of $\mathcal{H} = C^{(m)}[0, 1]$ with norm (1.3) can be obtained as $K = K_0 + K_1$.

Here are a few examples of function $k_r(x)$,

$$\begin{aligned}
k_0(x) &= 1 \\
k_1(x) &= x - 0.5, \quad x \in (0, 1) \\
k_2(x) &= \frac{1}{2}\big(k_1^2(x) - \frac{1}{12}\big) \\
k_4(x) &= \frac{1}{24}\big(k_1^4(x) - \frac{k_1^2(x)}{2} + \frac{7}{240}\big).
\end{aligned}$$

When $m = 1$, $K_0(x, y) = 1$ and

$$K_1(x, y) = k_1(x)k_1(y) + k_2(x - y).$$

When $m = 2$, $K_0(x, y) = 1 + k_1(x)k_1(y)$ and

$$K_1(x, y) = k_2(x)k_2(y) - k_4(x - y).$$

1.3 Overview of Later Chapters

Three functional regression models are covered in this thesis.

Chapter 2 introduces Functional Linear Regression Model, which is a core technique in functional data analysis(FDA). My focus is on testing the nullity of the slope function. In Section 2.2, a smoothing spline estimate for the slope function is introduced, and a generalized likelihood ratio test based on this smoothing spline estimate is proposed. The quality of the test is measured by the minimal distance between the null and the alternative space that still allows a possible test. In Section 2.3, a lower bound of the minimax decay rate of this distance is derived. Test with a distance that decays faster than the lower bound would be impossible. We will also show that the minimax optimal rate is jointly determined by the reproducing kernel and the covariance kernel and our test attains this optimal rate. Section2.4 demonstrates the finite sample performance of the test under different simulated setups. Then the test is applied to study the effect of the trajectories of oxides of nitrogen (NOx) on the level of ozone (O3) in an California air quality example. All the proofs are displayed in Section 2.6.

In Chapter 3, the Functional Cox Model is studied and our work has been published in [5]. Functional covariates are common in many medical, biodemographic, and neuroimaging studies, while Cox proportional hazard model has been widely used in survival analysis. The Functional Cox Model incorporates functional covariates in to Cox model, and models the right-censored survival response with both functional and scalar covariates. Section 3.2 summarizes the asymptotic properties of

the maximum partial likelihood estimator that we established. It is shown that the estimator achieves the minimax optimal rate of convergence under a weighted $\mathcal{L}_2$-risk. Implementation of the estimation approach is discussed in Section 3.3, including a generalized cross-validation (GCV) method to select the smoothing parameter and a method of calculating the information bound of θ based on the alternating conditional expectations (ACE) algorithm. Section 3.4 contains numerical studies, including simulations and a data application. All the proofs are relegated to Section 3.5.

The model being considered in Chapter 4 is the Function-on-scalar Model. In this Chapter, we concentrated on developing the simultaneous model selection and estimation technique. It starts with the Function-on-scalar Model with both model selection and knots selection problems. This motives me to develop a novel regularization method called the Grouped Smoothly Clipped Absolute Deviation (GSCAD), which tackles both model selection and knots selection problems simultaneously. Function-on-scalar Model, and GSCAD are introduced in Section 4.1 and Section 4.2. It turns out the initial problem can be transferred into a dictionary learning problem, where the GSCAD can be directly applied to simultaneously learn a sparse dictionary and select the appropriate dictionary size. Formulation of the dictionary learning problem under matrix factorization framework is introduced in Section 4.3. Efficient algorithm is designed based on the alternative direction method of multipliers (ADMM) which decomposes the joint non-convex problem with the non-convex penalty into two convex optimization problems. Synthetic Experiments are presented in Section 4.4, follows by image denoising application in Section 4.5 and image inpainting application in Section 4.6.

2. OPTIMAL GLOBAL TEST FOR FUNCTIONAL LINEAR REGRESSION

2.1 Introduction

2.1.1 Functional Linear Regression Model

Functional linear regression model, which relates functional predictors to a scalar response, is one of the most useful tools in FDA. The model is stated as follows,

$$Y = \alpha_0 + \int_0^1 \beta_0(t)X(t)dt + \epsilon, \tag{2.1}$$

where Y is a scalar response, $X : [0,1] \to \mathbb{R}$ is a square integrable random functional predictor, $\alpha_0 \in \mathbb{R}$ is the intercept, $\beta_0 : [0,1] \to \mathbb{R}$ is the slope function, and ϵ is the random error with mean zero and variance σ^2. Since our main focus is on the coefficient function $\beta(t)$, we assume both X and Y are centered, i.e., $E(Y) = 0$ and $E(X(t)) = 0$ for all t, and therefore by taking expectation over both sides of (2.1), we have $\alpha_0 = 0$. Let $(X_i, Y_i), i = 1, \ldots, n$ be independent and identically distributed observations sampled from the model. Then model (2.1) can be rewritten as

$$Y_i = \int_0^1 \beta_0(t)X_i(t)dt + \epsilon_i, \qquad i = 1, \ldots, n. \tag{2.2}$$

2.1.2 Motivation

Although the asymptotic properties of estimators of β_0 are widely discussed in the literature, there is little research on testing whether β_0 resides in a given finite dimensional linear subspace, or more specifically, $\beta_0 \equiv 0$.

Take the study of California air quality data as an example. In this study, we focus on the effect of the trajectories of oxides of nitrogen (NO_x) on the levels of ozone (O_3).

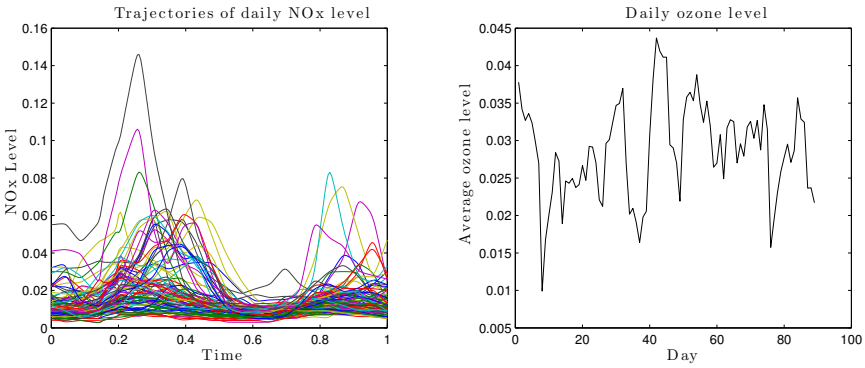


Figure 2.1. Left: the daily trajectories of NO_x levels. Right: average O_3 level each day.

Levels of ground-level concentrations of NO_x in the city of Sacramento is observed hourly every day from June 1 to August 31 in 2005, and records of Sacramento's daily average ground-level concentrations of O_3 during the same time period are obtained. Figure 2.1 displays the daily trajectories of NO_x levels as well as the daily average O_3 levels. We are interested in whether the level of NO_x trajectory has any effect on the O_3 level and, if it does, how long this effect lasts.

If we take daily NO_x trajectory as predictor $X(t)$ and average O_3 level as Y, then an absent effect will be indicated by a zero slope function in model (2.2). The estimated slope functions are shown in Figure 2.2. We see that when response Y is taken as the O_3 level of the same day as NO_x level, the estimated slope function has a large magnitude and a clear curve, which indicates that the true slope function in this model is very unlikely to be a zero function. On the other hand, when response is taken as the O_3 level five days later after the recorded NO_x trajectory, the estimated slope function stays close to zero. The slight curvature of this estimated slope function maybe due to randomness of the data, with the true β_0 residing in a zero null space. However to draw a statistical conclusion under a certain significant level on whether there is still some effect on the O_3 level from the NO_x level five days ago, we need a well-designed testing procedure.

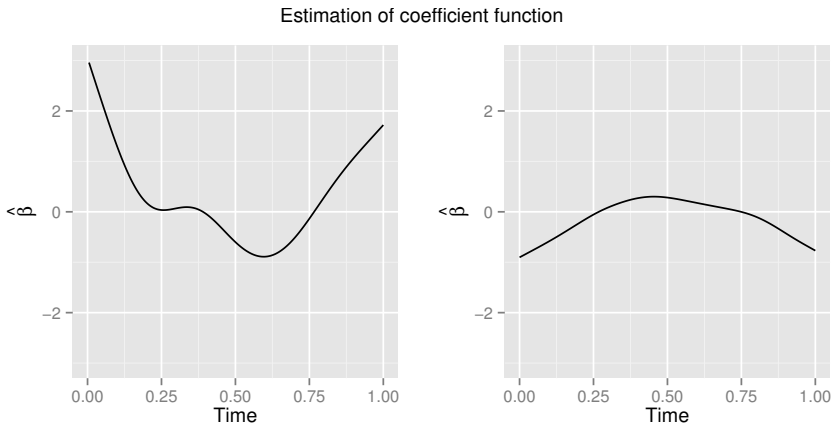


Figure 2.2. The estimated slope functions. Left panel: response Y is taken as the average O_3 level of the same day as NO_x level. Right panel: response Y is taken as the average O_3 level 5 days later after the recorded NO_x trajectory.

2.1.3 Related works

[6] proposed a test statistic based on the first k functional components of X, and derived a limiting distribution under the null and the corresponding power. It is well-known that selection of k is a difficult problem. Some computational methods have been studied to resolve this issue without theoretical guarantee on the power ([7,8]). For more recent work, [9] used the functional principle component approach to test the nullity of the slope function, and established that their procedures are minimax adaptive to the unknown regularity of the slope. In particular, they assumed that $\beta_0 \in \mathcal{E}_a(L)$ where

$$\mathcal{E}_a(L) = \Big\{\beta \in L_2[0,1] : \sum_{k=1}^{\infty} a_k^{-2}\big\langle \beta, \varphi_k \big\rangle^2 \leq L^2\Big\},$$

with $\langle \beta, \varphi_k \rangle = \int_0^1 \beta(t)\varphi_k(t)dt$, and φ_k's are eigenfunctions of the covariance Γ. The smoothness of β_0 is characterized by the decay rate of a_k. $\mathcal{E}_a(L)$ is essentially a reproducing kernel Hilbert space (RKHS), denoted by $\mathcal{H}(K)$, with a specific reproducing kernel $K(t,s) = \sum_{k=1}^{\infty} a_k^2 \varphi_k(t)\varphi_k(s)$. When their underline assumption that, kernel K and Γ are well aligned, is not satisfied, their methods may not perform

well. [10] developed a method simultaneously testing the slope vectors in a sequence of functional principal components regression models, and showed that under certain conditions, his method is uniformly powerful over a class of smooth alternatives. However, the principal-component-based methods are successful upon the assumption that the slope function $\beta(t)$ can be well represented by the leading functional principal components of X. [11] showed that, for the benchmark Canadian weather data, the estimated Fourier coefficients of the slope function with respect to the eigenfunctions of the sample covariance function do not decay at all, which is a typical example for the case that the slope function is not well represented by the leading principal components.

For nonparametric regression, the nonparametric testing has been studied by a series of papers of [12–15]. Other related papers include [16], [17], [18] and [19]. For a more detailed review, see [20].

2.1.4 Problem statement

We study adaptive and minimax optimal testing procedures on detecting the nullity of the slope function in functional linear model within the framework of reproducing kernel Hilbert space. Let $\Gamma(s,t)$ denote the covariance function of X. Γ can also be taken as a nonnegative definite operator with $\Gamma f = \int_0^1 \Gamma(\cdot,\ t)f(t)dt$ for $f \in L_2$. We wish to test the null hypothesis $H_0 : \beta \equiv 0$ against the composite nonparametric alternative that β_0 is separated away from zero in terms of a L_2-norm induced by the operator Γ, i.e. $\|\beta_0\|_\Gamma \geq \varrho_n$, where $\|\beta\|_\Gamma^2 = \langle \Gamma\beta, \beta\rangle$ with $\langle \beta, \gamma\rangle = \int_0^1 \beta(t)\gamma(t)dt$. Then assuming that the unknown slope function β_0 possesses some smoothness properties such that it belongs to a reproducing kernel Hilbert space $\mathcal{H}(K)$ with a reproducing kernel K, therefore, we arrive at the following alternative:

$$H_1 : \mathcal{F}_{K,\Gamma}(\rho_n) = \Big\{\beta \in \mathcal{H}(K) : \|\beta\|_\Gamma \geq \rho_n\Big\}.$$

It should be emphasized that in the present paper we do not consider the usual L_2 norm in the alternative when specifying β_0 being separated away from zero. On one

hand, if there is no additional condition linking the smoothness of β_0 to the random curve X, $\|\hat{\beta} - \beta_0\|_2^2$ may not even be consistent by some standard approaches ([21]). On the other hand, the $||\cdot||_\Gamma$ norm is a more natural option in the sense that $\|\hat{\beta} - \beta_0\|_\Gamma^2$ represents prediction error.

The radius ρ_n characterizes the sensitivity of the test. We investigate the optimal decay rate of the radius ρ_n, under which the test with prescribed probabilities of errors is still possible. The minimax rate is established in a general setting with no constraint on the relationship between the reproducing kernel K and the covariance function Γ of the random predictor X. We show that the optimal ρ_n is jointly determined by both kernels K and Γ. In particular, the alignment of K and Γ can significantly affect the optimal rate of ρ_n. Similar phenomena occurs when studying prediction in the functional linear model ([11, 21, 22]). In particular, the optimal rate for prediction is associated with the decay rate of the eigenvalues of operator $K^{1/2}\Gamma K^{1/2}$.

We also propose a testing procedure that is shown to be asymptotically optimal by obtaining the previously described minimax optimal rate of ρ_n. We first develop a new smoothing spline estimator of the slope function β, and then construct a generalized likelihood ratio test statistic based on the estimated slope function $\hat{\beta}$. It is worth mentioning that this testing procedure can be easily generalized to the case when functional predictor is observed with a measurement error. In this case, on top of the proposed testing procedure, we only need to add a step to estimate the true predictor functions, which could be done by the commonly used regularized method. The optimal properties of our test are expected to be maintained.

2.2 Generalized Likelihood Ratio Test

2.2.1 Notation and definitions

We focus on the Sobolev space W_2^m of order m as the parameter space, defined by

$$W_2^m = \left\{ \beta : [0,1] \to \mathbb{R} \middle| \beta, \beta', \ldots, \beta^{(m-1)} \text{ are absolutely continuous and } \beta^{(m)} \in L_2[0,1] \right\}.$$

W_2^m is a reproducing kernel Hilbert space $\mathcal{H}(K)$ with the reproducing kernel ([2])

$$K(t,s) = \sum_{k=0}^{m-1} \frac{s^k t^k}{(k!)^2} + R(t,s),$$

where

$$R(t,s) = \int_0^1 \frac{(s-u)_+^{m-1}(t-u)_+^{m-1}}{\{(m-1)!\}^2} du.$$

Let T_0 and T_1 be operators on $L_2[0,1]$ such that

$$T_0 X(t) = \int_0^t X(s)ds \qquad \text{and} \qquad T_1 X(t) = \int_t^1 X(s)ds.$$

It follows Fubini's theorem that $\langle f, T_0 g\rangle = \langle T_1 f, g\rangle$, and thus T_0 is the adjoint operator to T_1. Further, define that $T_0^k X(t) = T_0 T_0^{k-1} X(t)$ and $T_1^k X(t) = T_1 T_1^{k-1} X(t)$ for $k \geq 2$. Therefore, T_0^k is the adjoint operator to T_1^k, and

$$T_0^k X(t) = \int_0^1 \frac{(t-s)_+^{k-1}}{(k-1)!} X(s)ds, \qquad T_1^k X(t) = \int_0^1 \frac{(s-t)_+^{k-1}}{(k-1)!} X(s)ds.$$

In particular,

$$R = T_0^m T_1^m.$$

Observe that R differs from K only by a polynomial of degree less than or equal to m. Therefore, their eigenvalues have the same decay rate.

The following notations will be used in estimating slope function and then constructing test statistic. Denote $\mathbf{X}(t) = (X_1(t), \ldots, X_n(t))^T$ and sample covariance function $\hat{\Gamma}(t,s) = n^{-1}\mathbf{X}(t)^T\mathbf{X}(s)$. Let $\widetilde{X}(1) \in \mathbb{R}^{m\times n}$ be an m by n matrix with the $(i,j)'s$ element $(\widetilde{X}(1))_{i,j} = T_0^i X_j(1)$ and $\hat{H} = n^{-1}\widetilde{X}(1)\widetilde{X}(1)^T$. Define a matrix $\hat{B} = \frac{1}{n}\widetilde{X}(1)^T\hat{H}^{-1}\widetilde{X}(1)$, then $\hat{B}$ is an $n \times n$ idempotent matrix with $\hat{B}^2 = \hat{B}$. Finally, define an operator $\hat{Q}$ as $\hat{Q}(t,s) = n^{-1}\hat{U}(t)^T\hat{U}(s)$, where $\hat{U}(t)$ is a random function vector such that

$$\hat{U}(t) = (I_n - \hat{B})T_0^m \mathbf{X}(t).$$

It is easy to see that

$$\hat{Q} = n^{-1}T_0^m \mathbf{X}^T(I_n - \hat{B})T_0^m \mathbf{X} = T_0^m\big(\hat{\Gamma} - \hat{\Gamma}_0\big)T_1^m,$$

where

$$\hat{\Gamma}_0(t,s) = \frac{1}{n}\mathbf{X}(t)^T B\mathbf{X}(s),$$

is a degenerated operator with at most m eigenvalues. Hence, the eigenvalues of $\hat{Q}$, $T_0^m\hat{\Gamma}T_1^m$ and further $T\Gamma T^*$ have the same decay rate.

2.2.2 The smoothing spline estimator

In this section, we study the smoothing spline estimate which will be used to construct the generalized likelihood ratio test in the next session. Let $\hat{\beta}$ be the smoothing spline estimate such that $\hat{\beta} \in W_2^m$ minimizes

$$\frac{1}{n}\sum_{i=1}^{n}\Big\{Y_i - \int_0^1 \beta(t)X_i(t)dt\Big\}^2 + \lambda\int_0^1 \Big\{\beta^{(m)}(s)\Big\}^2 ds, \tag{2.3}$$

where $\lambda > 0$ is the smoothing parameter. Next theorem provides the characterization of $\hat{\beta}$.

Theorem 2.2.1 *Denote* $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ *and operator* $\hat{Q}^+ = (\lambda I + \hat{Q})^{-1}$.

(a). The mth derivative of $\hat{\beta}$ *is*

$$\hat{\beta}^{(m)} = (-1)^m\frac{1}{n}\hat{Q}^+\hat{U}^T\mathbf{Y}.$$

(b). Let $\widehat{\Upsilon}(1) = \Big[\hat{\beta}(1), -\hat{\beta}'(1), \ldots, (-1)^{m-1}\hat{\beta}^{(m-1)}(1)\Big]^T$. *We have*

$$\widehat{\Upsilon}(1) = \frac{1}{n}\hat{H}^{-1}\widetilde{X}(1)\Big\{I_n - \frac{1}{n}\int_0^1 T_0^m\mathbf{X}(s)\ \hat{Q}^+\hat{U}(s)^T ds\Big\}\mathbf{Y}.$$

Theorem 2.2.1 provides a brand new approach to compute $\hat{\beta}$ explicitly over the infinitely dimensional function space $\mathcal{H}(K)$. This observation is important to both numerical implementation and asymptotic analysis. The explicit formula for $\hat{\beta}$ is

$$\hat{\beta}(t) = \widehat{\Upsilon}(1)^T\zeta(t) + (-1)^m\int_0^1 \hat{\beta}^{(m)}(s)\frac{(s-t)_+^{m-1}}{(m-1)!}ds = \Pi_t\mathbf{Y} \tag{2.4}$$

where $\zeta(t) = \Big[1, (1-t), \frac{(1-t)^2}{2!}, \ldots, \frac{(1-t)^{m-1}}{(m-1)!}\Big]^T$, and

$$\Pi_t = \frac{1}{n}\zeta(t)^T\hat{H}^{-1}\widetilde{X}(1)\Big\{I_n - \frac{1}{n}\int_0^1 T_0^m\mathbf{X}(s)\ \hat{Q}^+\hat{U}(s)^T ds\Big\} + \frac{1}{n}T_1^m\hat{Q}^+\hat{U}(t)^T.$$

Therefore, $\hat{\beta}$ is a linear function of the response $\mathbf{Y}$ with Π_t as the hat matrix.

2.2.3 Generalized likelihood ratio test

Assuming that ϵ_i follows normal distribution, the conditional log-likelihood function for (2.2) becomes

$$\ell_n(\beta, \sigma) = -n \log(\sqrt{2\pi}\ \sigma) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \Big(Y_i - \int \beta X_i \Big)^2.$$

Define the residual sum of squares under the null and alternative hypothesis as follows:

$$\mathrm{RSS}_0 = \sum_{i=1}^{n} Y_i^2, \quad \mathrm{RSS}_1 = \sum_{i=1}^{n} (Y_i - \int \hat{\beta} X_i)^2.$$

Then the logarithm of the conditional maximum likelihood ratio test statistic is given by

$$\tau_{n,\lambda} = \ell_n(\hat{\beta}, \hat{\sigma}_1) - \ell_n(0, \hat{\sigma}_0) = \frac{n}{2} \log \frac{\mathrm{RSS}_0}{\mathrm{RSS}_1}, \tag{2.5}$$

where $\hat{\sigma}_1^2 = \mathrm{RSS}_1/n$ and $\hat{\sigma}_0^2 = \mathrm{RSS}_0/n$. Define an $n \times n$ matrix $A_n = A_n(\boldsymbol{X})$ as

$$A_n = \frac{1}{n} \int_0^1 \hat{U}(t) \hat{Q}^{+} \hat{U}(t)^T dt - \frac{1}{2n} \int_0^1 \int_0^1 \hat{Q}^{+} \hat{U}(t) \hat{Q}(t,s) \hat{Q}^{+} \hat{U}(s)^T dt ds + \frac{1}{2} \hat{B}.$$

Next theorem shows the properties of the test statistic $\tau_{n,\lambda}$.

Theorem 2.2.2 . *If $tr(A_n) = o_p(n)$, we have the following results,*

(a). Under $H_0 : \beta \equiv 0$, the likelihood ratio test statistic $\tau_{n,\lambda}$ is of the form

$$\tau_{n,\lambda} = z^T A_n z + o_p\big(1\big),$$

where $z = \epsilon/\sigma$. Furthermore, if ϵ_i, $i = 1, ..., n$ are independent and identically distributed following $\mathcal{N}(0, \sigma^2)$, then $\tau_{n,\lambda}$ has an asymptotic normal distribution with mean $\mu_n = tr(A_n)$ and variance $\sigma_n^2 = 2\, tr(A_n^2)$.

(b). Under $H_1' : \mathcal{F}'_{K,\Gamma}(\rho_n) = \Big\{ \beta \in \mathcal{H}(K) : \|\beta\|_\Gamma = \rho_n \Big\}$, if $\rho_n^2 = o(n^{-1/2})$ and $\lambda = o(n^{-1/2})$, then

$$\tau_{n,\lambda} = z^T A z + \frac{n}{2\sigma^2} \|\beta_0\|_{\hat{\Gamma}}^2 + O_p\Big(n\lambda + n^{1/2}\lambda^{1/2} + n^{1/2} \|\beta_0\|_{\hat{\Gamma}} \Big).$$

The condition that $tr(A_n) = o_p(n)$ in Theorem 2.2.2 can be satisfied in many cases. In fact, $tr(A_n)$ can be computed explicitly. Consider the spectral decomposition of operator $\hat{Q}$, $\hat{Q}(t,s) = \sum_{j=1}^{\infty} \hat{\kappa}_j \hat{\phi}_j(t)\hat{\phi}_j(s)$, where $(\hat{\kappa}_j, \hat{\phi}_j)$ are (eigenvalue, eigenfunction) pairs, ordered such that $\hat{\kappa}_1 \geq \hat{\kappa}_2 \geq \cdots \geq 0$. We may write $\hat{U}_{X_i}(t) = \sum_{k=1}^{\infty} \hat{\xi}_{ik}\hat{\phi}_k(t)$. Since $\hat{Q}(t,s) = n^{-1}\sum_{i=1}^{n} \hat{U}_{X_i}(t)\hat{U}_{X_i}(s)$, we have $n^{-1}\sum_{i=1}^{n} \hat{\xi}_{ik}^2 = \hat{\kappa}_k$ and $n^{-1}\sum_{i=1}^{n} \hat{\xi}_{ik}\hat{\xi}_{ij} = 0$ for $k \neq j$. It is not hard to obtain that

$$tr(A_n) = \sum_{k=1}^{\infty} \frac{\hat{\kappa}_k(\lambda + \frac{1}{2}\hat{\kappa}_k)}{(\lambda + \hat{\kappa}_k)^2} + \frac{m}{2}.$$

Proposition 2.2.1 *If $\lambda^{-1} = O(n)$, then $tr(A)$ is of the same order of $\sum_{k=1}^{\infty} \frac{s_k}{\lambda + s_k}$.*

Proposition 2.2.1 shows that $tr(A_n)$ is determined by the order of λ and the decay rate of s_k, the sorted eigenvalues of linear operator $T\Gamma T^*$. More specifically, if s_k has a polynomial decay rate as $s_k \asymp k^{-2r}$, for some $r > 1/2$, then $tr(A_n) = O_p(\lambda^{-1/2r})$, while if s_k has an exponential decay rate as $s_k \asymp e^{-2rk}$ for some $r > 0$, then $tr(A_n) = O(\log \lambda^{-1})$. In both cases, $tr(A_n) = o_p(n)$ will be satisfied once we choose a proper λ. The optimal order of λ will be shown later in Theorem 2.3.2, followed by a data-driven procedure of choosing λ.

Based on Theorem 2.2.2, we have an α level testing procedure that, we reject H_0 when $\frac{\tau_{n,\lambda} - \mu_n}{\sigma_n} > z_\alpha$ where z_α is the upper α quantile of the standard normal distribution. In the next section, we will show that the power function of this test is asymptotically one at the minmax optimal rate.

2.3 Optimal Test

2.3.1 Minimax lower bound

Let ϕ_n be a measurable function of the observations taking values at two points $\{0, 1\}$. We accept H_0 if $\phi_n = 0$, and reject H_0 if $\phi_n = 1$. The probability of type I error, denoted by $\alpha_0(\phi_n)$, is

$$\alpha_0(\phi_n) = \mathbb{P}_0(\phi_n = 1),$$

where $\mathbb{P}_0$ is the probability measure on the space of observations corresponding to H_0. The probability of type II error, denoted by $\alpha_1(\phi_n)$, is

$$\alpha_1(\phi_n, \rho_n) = \sup_{\beta \in \mathcal{F}_{K,\Gamma}(\rho_n)} \mathbb{P}_\beta(\phi_n = 0),$$

where $\mathbb{P}_\beta$ is the probability measure corresponding to a particular slope function β. Let

$$\gamma_n(\phi_n, \rho_n) = \alpha_0(\phi_n) + \alpha_1(\phi_n, \rho_n),$$

which measures the error of the test ϕ_n by summarizing probability of the type I and type II errors. Fix a number $0 < \gamma < 1$. A sequence $\rho_n \to 0$ as $n \to \infty$ is called the minimax rate of testing if:

(i) For any sequence ρ_n' such that $\rho_n'/\rho_n \to 0$, we have $\liminf_{n\to\infty} \inf_{\phi_n} \gamma_n(\phi_n, \rho_n') \geq \gamma$;

(ii) There exists a test ϕ_n^* such that $\limsup_{n\to\infty} \gamma_n(\phi_n^*, \rho_n) \leq \gamma$.

For the given reproducing kernel K, let T and T^* be two operators acting on $L_2[0,1]$ such that $K = TT^*$, where T^* is the adjoint operator to T with $\langle f, Tg\rangle = \langle T^* f, g\rangle$. Consider the linear operator $T\Gamma T^*$. It follows from the spectral theorem that

$$T\Gamma T^*(t,s) = \sum_{k=1}^{\infty} s_k \varphi_k(t)\varphi_k(s),$$

where $s_1 \geq s_2 \geq \cdots > 0$ are the eigenvalues of the operator $T\Gamma T^*$ and φ_k's are the corresponding eigenfunctions. For any two sequences $a_k, b_k > 0$, $a_k \asymp b_k$ means that a_k/b_k is bounded away from zero and infinity as $k \to \infty$.

Theorem 2.3.1 *Assume ϵ_i, $i = 1, ..., n$ are independent and identically distributed following $\mathcal{N}(0, \sigma^2)$. Let $\{s_k : k \geq 1\}$ be the sorted eigenvalues of the linear operator $T\Gamma T^*$.*

(a). When $s_k \asymp k^{-2r}$ for some constant $r > 1/2$, let

$$\rho_n = n^{-2r/(1+4r)}. \tag{2.6}$$

If ρ_n' is such that $\rho_n'/\rho_n \to 0$ as $n \to \infty$, then

$$\liminf_{n\to\infty} \inf_{\phi_n} \gamma_n(\phi_n, \rho_n') \geq 1.$$

(b). When $s_k \asymp e^{-2rk}$ for some constant $r > 0$, let

$$\rho_n = \Big(\frac{\log n}{2rn^2}\Big)^{1/4}. \tag{2.7}$$

If ρ_n' is such that $\rho_n'/\rho_n \to 0$ as $n \to \infty$, then

$$\liminf_{n\to\infty} \inf_{\phi_n} \gamma_n(\phi_n, \rho_n') \geq 1.$$

The cholesky decomposition of the operator $K = TT^*$ is not unique, and T is not necessarily a symmetric operator. If we would like T to be a symmetric operator, we may choose $T = T^* = K^{1/2}$. It is shown in the next proposition that the decay rate of the eigenvalues of the operator $T\Gamma T^*$ and $K^{1/2}\Gamma K^{1/2}$ have the same asymptotic order.

Proposition 2.3.1 *Let $K = TT^*$, where T^* is adjoint to T. The eigenvalues of the two operators $T\Gamma T^*$ and $K^{1/2}\Gamma K^{1/2}$ have the same decay rate.*

The minimax lower bound for the excess prediction risk has been established by [11]. Suppose the k^{th} eigenvalues of the linear operator $K^{1/2}\Gamma K^{1/2}$ is of order k^{-2r} for some constant $0 < r < \infty$, then

$$\lim_{a\to 0} \lim_{n\to\infty} \inf_{\hat{\beta}} \sup_{\beta_0 \in H(K)} \mathbb{P}\Big(\big\| \hat{\beta} - \beta_0 \big\|_\Gamma \geq a n^{-\frac{r}{2r+1}} \Big) = 1.$$

It turns out that the optimal separating rate ρ_n for testing differs from the optimal rate for the problem of prediction. Similar situation arises in the setting of nonparametric regression.

Consider a special case that the reproducing kernel K is perfectly aligned with Γ, i.e., $K(s,t) = \sum_{k=1}^\infty a_k^2 \psi_k(t)\psi_k(s)$ and $\Gamma(t,s) = \sum_{k=1}^\infty \eta_k \psi_k(t)\psi_k(s)$. In this case, it is easy to see that $K^{1/2}\Gamma K^{1/2}(t,s) = \sum_{k=1}^\infty \eta_k a_k^2 \psi_k(t)\psi_k(s)$, which indicates that $s_k = \eta_k a_k^2$. This special case has been studied in [9].

2.3.2 Optimal adaptive test

Now back to the generalized likelihood ratio test. Recall that the test statistic $\tau_{n,\lambda}$ has an asymptotic normal distribution with mean $\mu_n = tr(A_n)$ and variance $\sigma_n^2 = 2\,tr(A_n^2)$. Concerning the distribution of the random function X, we shall assume that

(A1). X has a finite fourth moment, i.e., $\int_0^1 E(X^4) < \infty$ and

$$E\Big(\big\langle X, \psi_k\big\rangle^4\Big) \le C\Big(E\big\langle X, \psi_k\big\rangle^2\Big)^2 \quad \text{for } k \ge 1,$$

where $C > 0$ is a constant and ψ_k's are eigenfunctions of Γ.

Theorem 2.3.2 *Assume (A1) holds and ϵ_i, $i = 1, ..., n$ are independent and identically distributed following $\mathcal{N}(0, \sigma^2)$. Let $\{s_k : k \ge 1\}$ be the sorted eigenvalues of the linear operator $T\Gamma T^*$.*

(a). When $s_k \asymp k^{-2r}$ for some constant $r > 1/2$. Choose

$$\lambda = cn^{-4r/(4r+1)},$$

for some $c > 0$. Then μ_n and σ_n^2 are of order $O_p(n^{2/(4r+1)})$, and for any sequence $c_n \to \infty$, the power function of the generalized likelihood ratio test is asymptotically one:

$$\inf_{\beta \in \mathcal{F}_{K,\Gamma}(c_n \cdot \rho_n): \|\beta\|_\Gamma \ge c_n n^{-2r/(4r+1)}} \mathbb{P}_\beta\Big(\frac{\tau_{n,\lambda} - \mu_n}{\sigma_n} > z_\alpha\Big) \longrightarrow 1,$$

where z_α is the upper α quantile of the standard normal distribution and ρ_n is given in (2.6).

(b). Assume $s_k \asymp \exp(-2rk)$ for some constant $r > 0$. Choose λ such that

$$\log \lambda^{-1} = O(\log n), \qquad \lambda^{-1} n^{-1} = O(1), \qquad \text{and} \quad \lambda = o(n^{-1/2}).$$

Then μ_n and σ_n^2 are of order $O_p\{\log n/(2r)\}$, and for any sequence $\tilde{c}_n \to \infty$,

$$\inf_{\beta \in \mathcal{F}_{K,\Gamma}: \|\beta\|_\Gamma \ge \tilde{c}_n \{\log n/(2rn^2)\}^{1/4}} \mathbb{P}_\beta\Big(\frac{\tau_{n,\lambda} - \mu_n}{\sigma_n} \ge z_\alpha\Big) \longrightarrow 1.$$

The optimal smoothing parameters for prediction and testing are different. When $\kappa_k \asymp k^{-2r}$, if we choose $\lambda = \tilde{\lambda}$ to be of order $n^{-2r/(2r+1)}$, which is the optimal order for prediction, the rate of the testing will be slower than the optimal rate given in Theorem 2.3.1. Specifically, there exists a $\beta \in \mathcal{F}_{K,\Gamma}$ satisfying $\|\beta\|_\Gamma = n^{-(r+d)/(2r+1)}$ with $d > 1/8$ such that the power function of the test at the point β is bounded by α, namely

$$\limsup_{n\to\infty} \mathbb{P}_\beta\Big(\tau_{n,\tilde{\lambda}} > \mu_n + z_\alpha \sigma_n\Big) \le \alpha.$$

As we see in part (b), when s_k is exponentially decayed, the choice of λ is more flexible. For example, any n^d for $-1 \le d < -\frac{1}{2}$, could guarantee an optimal test.

Considering λ^* such that

$$\lambda^* = \arg\min_{\lambda\ge 0} \Big(\lambda + \frac{1}{n}\sum_{k=1}^{\infty} \frac{\kappa_k}{\sqrt{\lambda} + \kappa_k}\Big),$$

where κ_k's are eigenvalues of $Q = T_0^m \Gamma T_1^m$. λ^* is well-defined, since

$$\sum_{k=1}^{\infty} \kappa_k = \int_0^1 Q(t,t)dt = E\big\langle T_0^m X, T_1^m X\big\rangle \le C_1 \int_0^1 E(X^2) < \infty.$$

It is not hard to see that $\lambda^* \asymp n^{-4r/(4r+1)}$ if $\kappa_k \asymp k^{-2r}$, while $\lambda^* \asymp n^{-1}$ if $\kappa_k \asymp e^{-2rk}$. Therefore an estimated λ^* can be used as our choice of the smoothing parameter. It is natural to use $\tilde{Q} = T_0^m \hat{\Gamma} T_1^m$ as an estimate of Q. The following Theorem gives an adaptive estimation of λ.

Theorem 2.3.3 *Assume (A1) holds. Denote by $\tilde{\kappa}_1 \ge \tilde{\kappa}_2 \ge \cdots \ge 0$ the eigenvalues of $\tilde{Q}$. Choosing $\tilde{\lambda}$ as*

$$\tilde{\lambda} = \arg\min_{\lambda\ge 0} \Big(\lambda + \frac{1}{n}\sum_{k=1}^{\infty} \frac{\tilde{\kappa}_k}{\sqrt{\lambda} + \tilde{\kappa}_k}\Big). \tag{2.8}$$

When $s_k \asymp k^{-2r}$ for some constant $r > 1/2$, there exist constants $0 < c_1 < c_2 < \infty$ such that

$$\lim_{n\to\infty} \mathbb{P}\Big(c_1 < \frac{\tilde{\lambda}}{\lambda_o} < c_2\Big) = 1$$

where $\lambda_o = cn^{-4r/(4r+1)}$ for some $c > 0$.

Theorem 2.3.3 verifies that $\tilde{\lambda}$ chosen by (2.8) is of the proper order. Simulations also show that as long as $X(s)$ and Y are at a proper scale, say ranging at the level of $[-10, 10]$, we can directly use the $\tilde{\lambda}$ without worrying about multiplying a constant. However we need to be more careful when X and Y are numerically at a different scale. As for the case when κ_k is exponentially decayed, the proper λ has a much larger range. We can still use (2.8) to get a proper λ.

2.4 Numerical Studies

2.4.1 Simulation

Consider the case that slope function $\beta(t)$ is in the Soblev space W_2^2. The penalty function in (2.3) becomes $\lambda \int_0^1 \beta''(s)^2 ds$. Following a similar setup as that in Yuan and Cai (2010), we generate the covariate function $X(\mathrm{t})$ by:

$$X(t) = \sum_{k=1}^{50} \zeta_k Z_k \phi_k(t).$$

where Z_k's are independently sampled from $Unif[-\sqrt{3}, \sqrt{3}]$ and ϕ_k's are Fourier basis with $\phi_1 = 1$ and $\phi_{k+1}(t) = \sqrt{2}\cos(k\pi t)$ for $k \geq 1$. We have two settings for ζ_k. For setup 1, let

$$\zeta_k = (-1)^{k+1} k^{-v/2} / ||\zeta||$$

where $\zeta = (\zeta_1, ..., \zeta_{50})^T$ and $||\cdot||$ indicates $\mathcal{L}_2$ norm. The normalizing term $||\zeta||^{-1}$ is added to rule out the potential effect from the magnitude of $X(s)$. For setup 2, ζ is chosen as

$$\zeta_k = \begin{cases} 1 & k = 1 \\ 0.2(-1)^{k+1}(1 - 0.0001k) & 2 \leq k \leq 4 \\ 0.2(-1)^{k+1}[5(k/5)]^{-v/5} - 0.0001(k \, mod \, 5) & k \geq 5 \end{cases} .$$

Table 2.1.
Size of the test under setup 1.

	n=50	n=100	n=200
ν=1.1	0.087	0.089	0.067
ν=1.5	0.085	0.078	0.072
ν=2	0.076	0.085	0.079
ν=4	0.075	0.070	0.079

The eigenvalues of the covariance function of $X(t)$ are ζ_k^2's, the decay rate of which is determined by ν. In both cases, let $\nu = 1.1, 1.5, 2, 4$. With the same basis, the true slope function β_0 is generated as:

$$\beta_0 = B \cdot \sum_{i=1}^{50} (-1)^{k+1} k^{-2} \phi_k$$

where B is a constant to control the norm of β_0. For both setups, a set of B ranging from 0 to 1 is examined. Response Y is generated through the functional regression model with $\varepsilon \sim N(0, 1)$. Sample size $n = 50, 100, 200$ are adopted to appreciate the effect of sample size.

For each simulated dataset, smoothing parameter λ is chosen based on (2.8), $\hat{\beta}(t)$ is estimated by (2.4), and the testing statistic $\tau_{n,\lambda}$ is calculated as shown in (2.5). According to Theorem 2.2.2, we reject H_0 if $\frac{\tau_{n,\lambda} - \mu_n}{\sigma_n} > z_\alpha$, with $\alpha = 0.05$. To estimate the size and power of our testing procedure, each setting is repeated 1000 times to get the percentage of rejecting H_0.

For setup 1, Table 2.1 shows the size of the test under different decay rate ν and sample size n. As we see, the size of test is slightly larger than what we expect under $\alpha = 0.05$. The reason is that with a finite sample size, $\tau_{n,\lambda}$ tends to be slightly larger than a random variable that follows exactly normal distribution. Recall Theorem 2.2.2, we conclude that under H_0, $\tau_{n,\lambda} = z^T A_n z + o_p(1)$, where the quadratic form $z^T A_n z$ is asymptotic normal. The small positive term $o_p(1)$, that we drop, plays its

Table 2.2.
Size of the test under setup 1 using the correction rule.

	n=50	n=100	n=200
ν=1.1	0.066	0.058	0.059
ν=1.5	0.048	0.055	0.041
ν=2	0.051	0.055	0.045
ν=4	0.067	0.053	0.041

role, when we treat $\tau_{n,\lambda}$ as the quadratic form $z^T A_n z$. To make a correction, we can use a two-sided test instead, which is to reject H_0 if $|\frac{\tau_{n,\lambda}-\mu_n}{\sigma_n}| > z_{\alpha/2}$. Under this correction rule, the size of the test stays closer around 0.05 as shown in Table 2.2.

Under alternative hypothesis $H_1 : \beta_0 \in F_{K,\Gamma}(\rho_n)$, the power function of test under different decay rate ν and sample size n are shown in Figure 2.3. It is very clear that as B increases, $||\beta_0||_\Gamma$ increases, and therefore the power of the test increases to 1. Also as expected, under the same setting, when sample size n goes up, the power should increase, which manifests a steeper slope of the power function in the figure. What is more interesting in the figure, is how the power is affected by the decay rate of the eigenvalues of $T_0^m \Gamma T_1^m$, which in our setting is determined by ν. As shown in the figure, power function with $\nu = 4$ always lies on top while that with $\nu = 1.1$ always stays the lowest, which perfectly matches Theorem 2.3.1 that the larger the ν, the faster the decay rate, and therefore the more powerful the test.

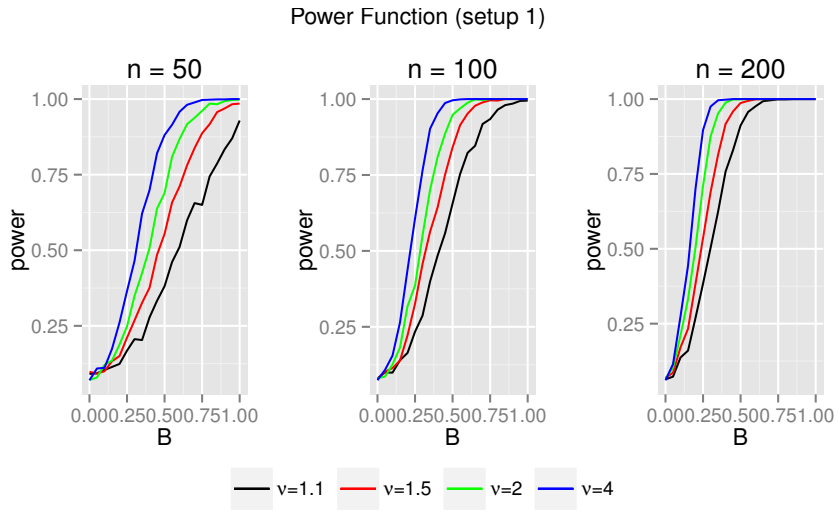


Figure 2.3. power function of the test under setup 1 for n=50, 100, 200

For setup 2, the size of the test and its correction version are shown in Table 2.3 and Table 2.4. Plots of power functions for different sample size n and decay rate ν are shown in Figure 2.4. Similarly as the previous results of setup 1, the power of the test goes up when sample size n and $||\beta_0||_\Gamma$ increase. However the effect of the decay

rate ν can be hardly seen this time. The reason is that when choosing ζ we did not normalize it as we did in setup1. Therefore even though a larger ν could lead to a more powerful test, the magnitude of $X(s)$ is significantly decreased due to the faster decay rate, and this counter balanced the effect of ν.

Table 2.3.
Size of the test under setup 2.

	n=50	n=100	n=200
ν=1.1	0.094	0.074	0.079
ν=1.5	0.088	0.067	0.073
ν=2	0.090	0.066	0.070
ν=4	0.091	0.066	0.065

Table 2.4.
Size of the test under setup 2 using the correction rule.

	n=50	n=100	n=200
ν=1.1	0.065	0.052	0.054
ν=1.5	0.063	0.046	0.057
ν=2	0.059	0.051	0.051
ν=4	0.065	0.044	0.050

2.4.2 California air quality data

Back to the California air quality example, as mentioned in the introduction, we are interest in testing the effect of trajectories of oxides of nitrogen (NO_x) on the level of ground-level concentrations of ozone (O_3). Data we are using is from the database of California Air Quality Data. NO_x levels and O_3 levels of city Sacramento

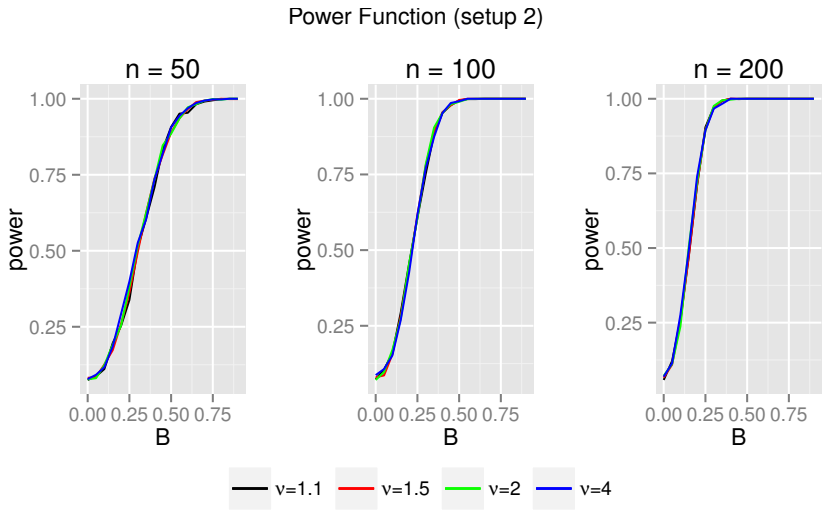


Figure 2.4. Power function of the test under setup 2 for n=50, 100, 200.

are recorded from June 1 to August 31 in 2015. There are 91 days on the record, and 3 days are removed due to severe missing data. For the rest 89 days, levels of NO_x are observed at each hour except for 4am and average O_3 level can also be obtained through the recorded data. The left panel of Figure 2.1 displays the daily trajectories of NO_x levels, and the right panel shows the average O_3 level each day during the same time period. When applying the proposed testing procedure, every record is rescaled by multiplying 100 due to the small magnitude.

Let $X_i(s)$, $i = 1, ..., 89$ denote the daily trajectories of NO_x levels after pre-smoothing and centering, and rescale s so that $s \in [0, 1]$. In the introduction, two types of response variables are considered, the average O_3 level of the same day as the NO_x level, and the average O_3 level five days later after the recorded NO_x trajectory. More generally we can examine the relation between the O_3 level of a certain day and the NO_x level d days before that day. If we take Y_i, $i = 1, ..., 89$ as the corresponding O_3 level of the day when X_i is recorded. Then the regression function is written as

$$Y_{i+d} = \int_0^1 X_i(s)\beta(s)ds + \epsilon_i,$$

for a fixed d.

Table 2.5.
P-value

d	0	1	2	3	4	5	6
p-value	3.07e-5	6.78e-9	2.30e-5	3.13e-4	0.0031	0.36	0.70

We go through the proposed testing procedure for $d = 0, 1, ..., 5$ and all the p-value are listed in Table 2.5. We can see that for d up to 4, the test returns a significant result at level $\alpha = 0.05$, which indicates that daily NO_x level is significantly related to the O_3 level up to four days later. Noting that Bonferroni correction for multiple comparison is applied here when identifying significance. It is also interesting to see that the smallest p-value occurs at $d = 1$. A possible way to interpret it is that instead of the current NO_x level, the average O_3 level depends more on the NO_x level the day before. That is to say there is a delayed effect of NO_x level on O_3 level.

2.5 Discussion

We have so far focused on the case with continuously observed functional predictors. If we have densely observed functional predictors, our framework can be applied similarly. An interesting extension of the current work would be to study the case when having sparsely observed functional predictors with/without measurement error. The ideas of [23] can be applied. A common strategy is to first have a pre-smoothing step and then apply our methodology. How the number of sparse observations affects the power of the test is beyond the scope of this paper and will be explored in future works.

A continuation of this paper is to study the optimal testing for the generalized functional linear model with a scalar response and a functional predictor ([24]). Given the functional predictor, the response is assumed to follow some distribution from the exponential family. The main difficulty is that the characterization conditions of the

slope estimator becomes complex and nontrivial. This problem hinders further studies in the asymptotic properties. We conjecture that the generalized likelihood ratio test will achieve the optimal rate of testing and the optimal rate still depends on the decay rate of $K^{1/2}\Gamma K^{1/2}$. This issue will be addressed in detail in the future.

2.6 Proofs of Theorems

2.6.1 Proof of Theorem 2.2.1

We prove this theorem using the calculus of variation. Denote

$$L(\beta) = \frac{1}{n}\sum_{i=1}^{n}\Big\{Y_i - \int_0^1 X_i(s)\beta(s)ds\Big\}^2 + \lambda\int_0^1\Big\{\beta^{(m)}(s)\Big\}^2 ds.$$

For any $\beta, \beta_1 \in W_2^m$ and $\delta \in \mathbb{R}$,

$$L(\beta + \delta\beta_1) - L(\beta) = 2\delta L_1(\beta, \beta_1) + O(\delta^2), \tag{2.9}$$

where

$$\begin{aligned} L_1(\beta, \beta_1) &= -\frac{1}{n}\sum_{i=1}^{n}\big\{Y_i - \int_0^1 X_i(s)\beta(s)ds\big\}\big\{\int_0^1 X_i(s)\beta_1(s)ds\big\} \\ &\quad + \lambda\int_0^1 \beta^{(m)}(s)\beta_1^{(m)}(s)ds. \end{aligned} \tag{2.10}$$

By Lemma 1, if $L_1(\beta, \beta_1) = 0$ for all $\beta_1 \in W_2^m$, letting $\mathcal{I}_1 = \{t \in [0,1] : L_2(\beta) \neq 0\}$ and $\beta_1^{(m)}(t) = -I_{\mathcal{I}_1}(t)$ gives

$$L_1(\beta, \beta_1) = \int_{\mathcal{I}_1} L_2(\beta)dt \neq 0,$$

unless $\mathcal{I}_1$ is of measure zero. This shows $L_2(\beta) = 0$ a.e.. This complete the proof of the first part of the theorem.

If $\hat{\beta}$ is the optimal solution, we have

$$\hat{\beta}^{(m)} = \frac{(-1)^m}{n}\hat{Q}^{+}\hat{U}^T\mathbf{Y}.$$

It follows from (2.19) that

$$\hat{H}\widehat{\Upsilon}(1) + \frac{(-1)^m}{n}\widetilde{X}(1)\int_0^1 T_0^m\mathbf{X}(s)\hat{\beta}^{(m)}(s)ds = \frac{1}{n}\widetilde{X}(1)\mathbf{Y}.$$

Therefore, the second part of the theorem follows from these two facts.

■

2.6.2 Proof of Theorem 2.2.2

For part (a), under H_0 with $\beta_0 \equiv 0$, we have

$$\frac{1}{n}\mathrm{RSS}_0 = \frac{1}{n}\epsilon^T\epsilon,$$
$$\frac{1}{n}\mathrm{RSS}_1 = \frac{1}{n}\epsilon^T\epsilon + \|\hat{\beta}-\beta_0\|^2_{\hat{\Gamma}} - \frac{2}{n}\epsilon^T\int(\hat{\beta}-\beta_0)\mathbf{X}.$$

It follows from Lemma 2 that,

$$\begin{aligned}
&\frac{1}{n}\mathrm{RSS}_1 - \frac{1}{n}\mathrm{RSS}_0 \\
&\quad = \|\hat{\beta}-\beta_0\|^2_{\hat{\Gamma}} - \frac{2}{n}\epsilon^T\int(\hat{\beta}-\beta_0)\mathbf{X} \\
&\quad = \frac{1}{n^2}\epsilon^T\Big\{\int_0^1\int_0^1 \hat{Q}^+\hat{U}(t)\hat{Q}(t,s)\hat{Q}^+\hat{U}(s)^T dtds - 2\int_0^1 \hat{U}(t)\hat{Q}^+\hat{U}(t)^T dt\Big\}\epsilon \\
&\qquad - \frac{1}{n^2}\epsilon^T\widetilde{X}(1)^T\hat{H}^{-1}\widetilde{X}(1)\epsilon \\
&\quad = -\frac{2}{n}\epsilon^T A_n\epsilon = o_p(n^{-1/2}),
\end{aligned}$$

provided that $tr(A_n^2) = o(n)$. Hence, with the fact that under H_0, $\sigma^2 = \mathrm{RSS}_0/n + O_p(n^{-1/2})$, the likelihood ratio test statistic $\tau_{n,\lambda}$ becomes

$$\begin{aligned}
\tau_{n,\lambda} &= -\frac{n}{2}\log\frac{\mathrm{RSS}_1/n}{\mathrm{RSS}_0/n} = -\frac{n}{2\sigma^2}\Big(\frac{1}{n}\mathrm{RSS}_1 - \frac{1}{n}\mathrm{RSS}_0\Big)\big(1+o_p(n^{-\frac{1}{2}})\big) \\
&= z^T A_n z + o_p(1),
\end{aligned}$$

where $z = \epsilon/\sigma$.

To show that $\tau_{n,\lambda}$ has an asymptotic normal distribution with mean $\mu_n = tr(A_n)$ and variance $\sigma_n^2 = 2tr(A_n^2)$, we need to show that

$$Trace(A_n^4)/\sigma_n^4 \to 0.$$

Let

$$A_I = \frac{1}{n}\int_0^1 \hat{U}(t)\hat{Q}^+\hat{U}(t)^T dt - \frac{1}{2n}\int_0^1\int_0^1 \hat{Q}^+\hat{U}(t)\hat{Q}(t,s)\hat{Q}^+\hat{U}(s)^T dtds,$$

and

$$A_{II} = \frac{1}{2}B.$$

So $tr(A) = tr(A_I) + tr(A_{II})$. Noting that $tr(A_{II}) = m/2$, $tr(A)$ is of the same order as $tr(A_I)$. Recall that $\hat{Q}(t,s) = \sum_{j=1}^{\infty} \hat{\kappa}_j \hat{\phi}_j(t)\hat{\phi}_j(s)$. and $\hat{U}_{X_i}(t) = \sum_{k=1}^{\infty} \hat{\xi}_{ik}\hat{\phi}_k(t)$ with $n^{-1}\sum_{i=1}^{n} \hat{\xi}_{ik}^2 = \hat{\kappa}_k$ and $n^{-1}\sum_{i=1}^{n} \hat{\xi}_{ik}\hat{\xi}_{ij} = 0$ for $k \neq j$. Therefore

$$(A_I)_{ij} = \frac{1}{n}\sum_{k=1}^{\infty} \frac{(2\lambda + \hat{\kappa}_k)\hat{\xi}_{ik}\hat{\xi}_{jk}}{2(\lambda + \hat{\kappa}_k)^2}.$$

Further

$$tr(A_I) = \sum_{k=1}^{\infty} \frac{\hat{\kappa}_k(2\lambda + \hat{\kappa}_k)}{2(\lambda + \hat{\kappa}_k)^2} \asymp \sum_{k=1}^{\infty} \frac{\hat{\kappa}_k}{\lambda + \hat{\kappa}_k}.$$

Similarly, we can show that

$$(A_I^2)_{ij} = \frac{1}{n}\sum_{k=1}^{\infty} \frac{(2\lambda + \hat{\kappa}_k)^2\hat{\kappa}_k\hat{\xi}_{ik}\hat{\xi}_{jk}}{4(\lambda + \hat{\kappa}_k)^4}, \qquad (A_I^4)_{ij} = \frac{1}{n}\sum_{k=1}^{\infty} \frac{(2\lambda + \hat{\kappa}_k)^4(\hat{\kappa}_k)^3\hat{\xi}_{ik}\hat{\xi}_{jk}}{16(\lambda + \hat{\kappa}_k)^8},$$

and

$$tr(A_I^2) \asymp \sum_{k=1}^{\infty} \frac{\hat{\kappa}_k^2}{(\lambda + \hat{\kappa}_k)^2}, \qquad tr(A_I^4) \asymp \sum_{k=1}^{\infty} \frac{\hat{\kappa}_k^4}{(\lambda + \hat{\kappa}_k)^4}.$$

Since $\frac{\hat{\kappa}_k^4}{(\lambda+\hat{\kappa}_k)^4} \leq \frac{\hat{\kappa}_k^2}{(\lambda+\hat{\kappa}_k)^2}$, therefore $tr(A_n^4) = O(\sigma_n^2)$, and further $tr(A_n^4)/\sigma_n^4 \to 0$.

For part (b), Under H_1',

$$\begin{aligned}
\frac{1}{n}\mathrm{RSS}_0 &= \frac{1}{n}\epsilon^T\epsilon + \|\beta_0\|_{\hat{\Gamma}}^2 + \frac{2}{n}\epsilon^T \int \beta_0 \mathbf{X}, \\
\frac{1}{n}\mathrm{RSS}_1 &= \frac{1}{n}\epsilon^T\epsilon + \|\hat{\beta} - \beta_0\|_{\hat{\Gamma}}^2 - \frac{2}{n}\epsilon^T \int (\hat{\beta} - \beta_0)\mathbf{X} \\
&= \sigma^2 - \frac{2}{n}\epsilon^T A\epsilon \\
&\quad + \lambda^2 \int_0^1 \int_0^1 \hat{Q}(t,s)\hat{Q}^+\beta_0^{(m)}(t)\hat{Q}^+\beta_0^{(m)}(s)dtds \\
&\quad + (-1)^m \frac{2\lambda}{n}\epsilon^T \int_0^1 \hat{U}(t)\hat{Q}^+\beta_0^{(m)}(t)dt - \frac{2}{n}\epsilon^T \int \beta_0 \mathbf{X} \\
&\quad + (-1)^{m+1} \frac{2\lambda}{n}\epsilon^T \int_0^1 \int_0^1 \hat{Q}(t,s)\hat{Q}^+\beta_0^{(m)}(t)\hat{Q}^+\hat{U}(s)dtds.
\end{aligned}$$

For $\frac{1}{n}\mathrm{RSS}_0$,

$$Var(\frac{2}{n}\epsilon^T \int \beta_0 \mathbf{X}) = \frac{4}{n^2} Var(\sum_{i=1}^{n} \epsilon_i \int \beta_0 X_i) = \frac{4\sigma^2}{n^2}\sum_{i=1}^{n} \{\int \beta_0(s)X_i(s)ds\}^2 = O(\frac{1}{n}||\beta_0||_{\hat{\Gamma}}^2).$$

For $\frac{1}{n}\text{RSS}_1$, write $\beta_0^{(m)}(t) = \sum_{j=1}^{\infty} \hat{\eta}_j \hat{\phi}_j(t)$. Since $\beta_0^{(m)} \in L_2$, we have $\sum_{j=1}^{\infty} \hat{\eta}_j^2 < \infty$. In the above expansion of $\frac{1}{n}\text{RSS}_1$,

$$\lambda^2 \int_0^1 \int_0^1 \hat{Q}(t,s)\hat{Q}^+\beta_0^{(m)}(t)\hat{Q}^+\beta_0^{(m)}(s)dtds$$
$$= \lambda^2 \sum_{j=1}^{\infty} \frac{\hat{\kappa}_j \hat{\eta}_j^2}{(\lambda + \hat{\kappa}_j)^2} \leq \lambda^2 \sum_{j=1}^{\infty} \hat{\eta}_j^2 \sup_{x \geq 0} \frac{x}{(\lambda + x)^2} \leq \frac{\lambda}{4} \sum_{j=1}^{\infty} \hat{\eta}_j^2 = O(\lambda).$$

Further,

$$(-1)^m \frac{2\lambda}{n} \epsilon^T \int_0^1 \hat{U}(t)\hat{Q}^+\beta_0^{(m)}(t)dt = (-1)^m \frac{2\lambda}{n} \sum_{i=1}^{n} \epsilon_i \sum_{k=1}^{\infty} \frac{\hat{\xi}_{ik}\hat{\eta}_k}{\lambda + \hat{\kappa}_k},$$

and its variance is

$$\frac{4\lambda^2\sigma^2}{n} \sum_{k=1}^{\infty} \frac{\hat{\kappa}_k \eta_k^2}{(\lambda + \hat{\kappa}_k)^2} \leq \frac{4\lambda^2\sigma^2}{n} \sup_{x \geq 0} \frac{x}{(\lambda + x)^2} \sum_{j=1}^{\infty} \hat{\eta}_j^2 \leq \frac{\lambda\sigma^2}{n} \sum_{j=1}^{\infty} \hat{\eta}_j^2 = O(\lambda/n).$$

The last term becomes

$$(-1)^{m+1} \frac{2\lambda}{n} \epsilon^T \int_0^1 \int_0^1 \hat{Q}(t,s)\hat{Q}^+\beta_0^{(m)}(t)\hat{Q}^+\hat{U}(s)dtds = (-1)^{m+1} \frac{2\lambda}{n} \sum_{i=1}^{n} \epsilon_i \sum_{k=1}^{\infty} \frac{\hat{\kappa}_k \hat{\eta}_k \hat{\xi}_{ik}}{(\lambda + \hat{\kappa}_k)^2}.$$

Since $\sum_{k=1}^{\infty} \frac{\hat{\kappa}_k \hat{\eta}_k \hat{\xi}_{ik}}{(\lambda + \hat{\kappa}_k)^2} \leq \sum_{k=1}^{\infty} \frac{\hat{\eta}_k \hat{\xi}_{ik}}{\lambda + \hat{\kappa}_k}$, the variance of last term is controlled by $O(\lambda/n)$. So altogether,

$$\frac{1}{n}\text{RSS}_1 - \frac{1}{n}\text{RSS}_0 = -\frac{2}{n}\epsilon^T A \epsilon - \|\beta_0\|_{\hat{\Gamma}}^2 + O\Big(\lambda\Big) + O_p\Big(n^{-1/2}\lambda^{1/2}\Big) + O_p\Big(n^{-1/2}\|\beta_0\|_{\hat{\Gamma}}\Big).$$

Since $\rho_n^2 = o(n^{-1/2})$ and $\lambda = o(n^{-1/2})$, therefor $\frac{1}{n}\text{RSS}_1 - \frac{1}{n}\text{RSS}_0 = o_p(n^{-1/2})$ and

$$\tau_{n,\lambda} = z^T A z + \frac{n}{2\sigma^2}\|\beta_0\|_{\hat{\Gamma}}^2 + O\Big(n\lambda\Big) + O_p\Big(n^{1/2}\lambda^{1/2}\Big) + O_p\Big(n^{1/2}\|\beta_0\|_{\hat{\Gamma}}\Big).$$

■

2.6.3 Proof of Theorem 2.3.1

The proof follows [15]. First show part (a). Let

$$\rho_n = n^{-2r/(1+4r)},$$

and suppose that $\rho_n'/\rho_n \to 0$. We show that, for any test ϕ_n,

$$\liminf_{n\to\infty} \gamma_n(\phi_n, \rho_n') \geq 1.$$

The idea of deriving the lower bound is standard. Let π_n be a probability measure on $\mathcal{F}_{K,\Gamma}(\rho_n')$. Then the lower bound is based on the inequality

$$\sup_{f\in\mathcal{F}_{K,\Gamma}(\rho_n')} \mathbb{P}_f(\phi_n = 0) \geq \mathbb{P}_{f,\pi_n}(\phi_n = 0),$$

where $\mathbb{P}_{f,\pi_n} = \int \mathbb{P}_f d\pi_n$. Write

$$\gamma_{n,\pi_n} = \mathbb{P}_0(\phi_n = 1) + \mathbb{P}_{f,\pi_n}(\phi_n = 0).$$

Denote by ℓ_{n,π_n} the likelihood ratio,

$$\ell_{n,\pi_n} = \frac{d\ \mathbb{P}_{f,\pi_n}}{d\ \mathbb{P}_0} = \int \frac{d\ \mathbb{P}_f}{d\ \mathbb{P}_0} d\pi_n.$$

For any $f \in \mathcal{F}_{K,\Gamma}(\rho_n)$, direct calculation yields that

$$\log \frac{d\ \mathbb{P}_f}{d\ \mathbb{P}_0} = \frac{1}{\sigma^2}\sum_{i=1}^n Y_i \int X_i f - \frac{n}{2\sigma^2}\|f\|_{\hat{\Gamma}}^2,$$

where $\hat{\Gamma}$ is the empirical covariance function such as

$$\hat{\Gamma}(t,s) = \frac{1}{n}\sum_{i=1}^n X_i(t)X_i(s).$$

It is convenient to use the following inequalities [14]:

$$\gamma_{n,\pi_n}(\phi_n, \rho_n') = 1 - \frac{1}{2}\mathrm{var}(\mathbb{P}_0, \mathbb{P}_{f,\pi_n}) \geq 1 - \frac{1}{2}\delta_{n,\pi_n},$$

where $\mathrm{var}(\mathbb{P}_0, \mathbb{P}_{f,\pi_n})$ stands for L_1 distance between two measures, and

$$\delta_{n,\pi_n}^2 = \mathbb{E}_0(\ell_{n,\pi_n} - 1)^2.$$

In the following, we select a probability measure π_n for which γ_{n,π_n} can be effectively estimated. Recall that $K = TT^*$, where T^* is the adjoint operator to T such

that $\langle f, Tg\rangle = \langle T^*f, g\rangle$. Define the linear operator $T\hat{\Gamma}T^*$ and let $\hat{s}_1 \geq \hat{s}_2 \geq \cdots \geq 0$ be the eigenvalues of $T\hat{\Gamma}T^*$ and the $\hat{\varphi}_k$ be the corresponding eigenfunctions. Consider

$$f_\xi = u\sum_{k=1}^{M} \xi_k g_k, \tag{2.11}$$

where $\xi = (\xi_1, \ldots, \xi_M)$ and $\xi_k = \pm 1$ with probability 1/2, and $g_k = \hat{s}_k^{-1/2} T^*\hat{\varphi}_k$. In (2.11), we choose $M = 2n^{2/(4r+1)}$ and $u = n^{-1/(4r+1)}\rho_n'$. Note that

$$\Big\langle g_k, g_j \Big\rangle_{\hat{\Gamma}} = (\hat{s}_k\hat{s}_j)^{-1/2}\Big\langle T^*\hat{\varphi}_k, T^*\hat{\varphi}_j \Big\rangle_{\hat{\Gamma}} = (\hat{s}_k\hat{s}_j)^{-1/2}\Big\langle T\hat{\Gamma}T^*\hat{\varphi}_k, \hat{\varphi}_j \Big\rangle_{L_2} = \delta_{jk},$$

where $\delta_{jk} = 1$ for $j = k$, and 0 for $j \neq k$. Further,

$$\Big\langle g_k, g_j \Big\rangle_{\mathcal{H}(K)} = (\hat{s}_k\hat{s}_j)^{-1/2}\Big\langle T^*\hat{\varphi}_k, T^*\hat{\varphi}_j \Big\rangle_{\mathcal{H}(K)} = (\hat{s}_k\hat{s}_j)^{-1/2}\Big\langle \hat{\varphi}_k, \hat{\varphi}_j \Big\rangle_{L_2} = (\hat{s}_k\hat{s}_j)^{-1/2}\delta_{jk}.$$

It is easy to check that

$$\|f_\xi\|^2_{\mathcal{H}(K)} = u^2\sum_{k=1}^{M} \hat{s}_k^{-1} \leq u^2 M\hat{s}_M^{-1} = 2(\rho_n')^2 s_M^{-1}(1 + o_p(1)),$$

which is bounded since s_M has the same order with $\rho_n^2 = n^{-4r/(4r+1)}$ and $\rho_n'/\rho_n = o(1)$. For any $\varphi \in L_2$, $T^*\varphi \in \mathcal{H}(K)$ ([25]). Therefore, $f_\xi \in \mathcal{H}(K)$. On the other hand,

$$\|f_\xi\|^2_{\hat{\Gamma}} = Mu^2 = 2(\rho_n')^2.$$

So, $\|f_\xi\|^2_{\Gamma} = 2(\rho_n')^2(1 + o(1)) \geq (\rho_n')^2$ and it shows that $f_\xi \in \mathcal{F}_{K,\Gamma}(\rho_n')$.

For this case, the likelihood ratio is

$$\begin{aligned}\ell_{n,\pi_n} = \mathbb{E}_\xi \frac{d\,\mathbb{P}_{f_\xi}}{d\,\mathbb{P}_0} &= \exp\Big(-\frac{nMu^2}{2\sigma^2}\Big)\mathbb{E}_\xi \exp\Big(\frac{u}{\sigma^2}\sum_{k=1}^{M}\sum_{i=1}^{n} Y_i x_{ik}\xi_k\Big) \\ &= \exp\Big(-\frac{nMu^2}{2\sigma^2}\Big)\prod_{k=1}^{M} cosh\Big(\frac{u}{\sigma^2}\sum_{k=1}^{M} Y_i x_{ik}\Big).\end{aligned}$$

where x_{ik} is denoted as $x_{ik} = \int X_i g_k$. Note that $\sum_{i=1}^n x_{ik}^2 = n\|g_k\|_{\hat{\Gamma}}^2 = n$. Given $X_1, \ldots, X_n$, we have

$$\begin{aligned}
\mathbb{E}_0\Big(\ell_{n,\pi_n}\Big|X_1,\ldots,X_n\Big) &= \mathbb{E}_0\mathbb{E}_\xi \frac{d\,\mathbb{P}_{f_\xi}}{d\,\mathbb{P}_0} = \mathbb{E}_\xi\mathbb{E}_0 \frac{d\,\mathbb{P}_{f_\xi}}{d\,\mathbb{P}_0} \\
&= \exp\Big(-\frac{nMu^2}{2\sigma^2}\Big)\mathbb{E}_\xi \prod_{i=1}^n \mathbb{E}_0 \exp(\frac{u}{\sigma^2}\sum_{k=1}^M \xi_k x_{ik}\cdot Y_i) \\
&= \exp\Big(-\frac{nMu^2}{2\sigma^2}\Big)\mathbb{E}_\xi \prod_{i=1}^n \exp\{\frac{1}{2}\sigma^2(\frac{u}{\sigma^2}\sum_{k=1}^M \xi_k x_{ik})^2\} \\
&= \exp(-\frac{nMu^2}{2\sigma^2})\exp(\frac{nMu^2}{2\sigma^2}) = 1.
\end{aligned}$$

Noting that

$$\begin{aligned}
\ell_{n,\pi_n}^2 &= \exp\Big(-\frac{nMu^2}{\sigma^2}\Big)\prod_{k=1}^M cosh(\frac{u}{\sigma^2}\sum_{k=1}^M Y_i x_{ik})^2 \\
&= \exp\Big(-\frac{nMu^2}{\sigma^2}\Big)\prod_{k=1}^M (\frac{1}{4}e^{\frac{2u}{\sigma^2}\sum_{k=1}^M Y_i x_{ik}} + \frac{1}{4}e^{-\frac{2u}{\sigma^2}\sum_{k=1}^M Y_i x_{ik}} + \frac{1}{2}) \\
&= \exp\Big(-\frac{nMu^2}{\sigma^2}\Big)\mathbb{E}_\varsigma \exp(\frac{2u}{\sigma^2}\sum_{k=1}^M\sum_{i=1}^n Y_i x_{ik}\varsigma_k),
\end{aligned}$$

where random variable ς takes value $-1, 0, 1$ with probability $1/4$, $1/2$, $1/4$. Therefore we can calculate $\mathbb{E}_0\Big(\ell_{n,\pi_n}^2\Big|X_1,\ldots,X_n\Big)$ as

$$\begin{aligned}
\mathbb{E}_0\Big(\ell_{n,\pi_n}^2\Big|X_1,\ldots,X_n\Big) &= \mathbb{E}_0 \exp\Big(-\frac{nMu^2}{\sigma^2}\Big)\mathbb{E}_\varsigma \exp(\frac{2u}{\sigma^2}\sum_{k=1}^M\sum_{i=1}^n Y_i x_{ik}\varsigma_k) \\
&= \exp\Big(-\frac{nMu^2}{\sigma^2}\Big)\mathbb{E}_\varsigma \prod_{i=1}^n \mathbb{E}_0 \exp(\frac{2u}{\sigma^2}\sum_{k=1}^M x_{ik}\varsigma_k\cdot Y_i) \\
&= \exp\Big(-\frac{nMu^2}{\sigma^2}\Big)\mathbb{E}_\varsigma \exp(\frac{2u^2 n}{\sigma^2}\sum_{k=1}^M \varsigma_k^2) \\
&= \exp\Big(-\frac{nMu^2}{\sigma^2}\Big)\prod_{k=1}^M \{\frac{1}{2}\exp(\frac{2u^2 n}{\sigma^2}) + \frac{1}{2}\} \\
&= \{\cosh(\frac{u^2 n}{\sigma^2})\}^M,
\end{aligned}$$

and

$$\mathbb{E}_0(\ell_{n,\pi_n}-1)^2=\mathbb{E}_0\Big(\ell_{n,\pi_n}^2\Big|X_1,\ldots,X_n\Big)-2\mathbb{E}_0\Big(\ell_{n,\pi_n}\Big|X_1,\ldots,X_n\Big)+1$$
$$=\{\cosh(\frac{u^2n}{\sigma^2})\}^M-1.$$

Using the inequality $\log\cosh x\le Bx^2$ for a certain B,

$$\{\cosh(\frac{u^2n}{\sigma^2})\}^M-1\le\exp\Big(\frac{BMu^4n^2}{\sigma^4}\Big)-1.$$

Hence

$$\mathbb{E}_0(\ell_{n,\pi_n}-1)^2\le\exp\Big(\frac{Bn^2Mu^4}{\sigma^4}\Big)-1.$$

Our choices of M and u guarantees that $n^2Mu^4\to 0$, so $\liminf_{n\to}\gamma_n(\phi_n,\rho_n')=1$. This completes the proof of part (a).

Next, we prove part (b). The proof is similar. In particular, in (2.11) we choose $M=\log n/(2r)$ and $u=2\rho_n'\sqrt{2r/\log n}$, where $\rho_n'/\rho_n\to 0$ with $\rho_n=n^{-1/2}(\log n/(2r))^{1/4}$. It is easy to see that $n^2Mu^4\to 0$, so that $\liminf_{n\to}\gamma_n(\phi_n,\rho_n')=1$. This completes the proof of part (b).

■

2.6.4 Proof of Theorem 2.3.2

Recall that $H_1':\mathcal{F}'_{K,\Gamma}(\rho_n)=\big\{\beta\in\mathcal{H}(K):\ \|\beta\|_\Gamma=\rho_n\big\}$, we only need to show that

$$\lim_{c_n\to\infty}\inf_{\beta_0\in\mathcal{F}'_{K,\Gamma}(c_n\rho_n)}\mathbb{P}_{\beta_0}\Big(\frac{\tau_{n,\lambda}-\mu_n}{\sigma_n}>z_\alpha\Big)=1.$$

The power function under H_1' can be written as

$$\mathbb{P}_{\beta_0}\Big(\frac{\tau_{n,\lambda}-\mu_n}{\sigma_n}\ge z_\alpha\Big)$$
$$=\mathbb{P}_{\beta_0}\Big\{\frac{z^TAz-\mu_n}{\sigma_n}+\frac{\frac{n}{2\sigma^2}\|\beta_0\|_{\hat\Gamma}^2+O\Big(n\lambda\Big)+O_p\Big(n^{1/2}\lambda^{1/2}\Big)+O_p\Big(n^{1/2}\|\beta_0\|_{\hat\Gamma}\Big)}{\sigma_n}\ge z_\alpha\Big\}.$$

Recall that $\sigma_n^2=tr(A^2)=O(tr(A))$ as shown in the proof as Theorem 2, and by Lemma 3, we have

$$\mu_n=O_p(\sum_{k=1}^\infty\frac{s_k}{\lambda+s_k})\qquad and\qquad \sigma_n^2=O_p(\sum_{k=1}^\infty\frac{s_k}{\lambda+s_k}).$$

Therefore μ_n and σ_n^2 are of order $O_p(\lambda^{-1/2r})$ when $s_k \asymp k^{-2r}$, or of order $O_p\{(2r)^{-1}\log\lambda^{-1}\}$ when $s_k \asymp e^{-2rk}$. Recall that when $\kappa_k \asymp k^{-2r}$, the optimal λ is of order $n^{-4r/(4r+1)}$; when $\kappa_k \asymp exp^{-2rk}$, $\log\lambda^{-1} = O(\log n)$. So, when $\kappa_k \asymp k^{-2r}$,

$$\lim_{c_n\to\infty}\inf_{\beta\in\mathcal{F}_{K,\Gamma}:\|\beta\|_\Gamma\geq c_n n^{-2r/(4r+1)}}\mathbb{P}_\beta\Big(\frac{\tau_{n,\lambda}-\mu_n}{\sigma_n}\geq z_\alpha\Big)=1,$$

and when $\kappa_k \asymp e^{-2rk}$,

$$\lim_{c_n\to\infty}\inf_{\beta\in\mathcal{F}_{K,\Gamma}:\|\beta\|_\Gamma\geq c_n\{\log n/(2rn^2)\}^{1/4}}\mathbb{P}_\beta\Big(\frac{\tau_{n,\lambda}-\mu_n}{\sigma_n}\geq z_\alpha\Big)=1.$$

This finishes the proof of the theorem. ■

2.6.5 Proof of Theorem 2.3.3

First noting that s_k and κ_k have the same decay rate, so we can replace s_k in condition $s_k \asymp k^{-2r}$ by κ_k.

Given a symmetric bivariate function M, let $|||M||| = (\int\int M^2)^{1/2}$. Define $\delta_k = \min_{1\leq j\leq k}(\kappa_j - \kappa_{j+1})$ which is of order k^{-2r-1}. $\tilde{\Delta} = |||\tilde{Q}-Q|||$, $\tilde{\Delta}_j = \|\int(\tilde{Q}-Q)\phi_j\|$, and

$$\tilde{\Delta}_{jj} = \int_0^1\int_0^1(\tilde{Q}(t,s)-Q(t,s))\phi_j(t)\phi_j(s)dtds.$$

It follows from Equation (5.7) of [26] that

$$\Big|\tilde{\kappa}_j-\kappa_j-\tilde{\Delta}_{jj}\Big|\leq\delta_j^{-1}\tilde{\Delta}(\tilde{\Delta}+\tilde{\Delta}_j),$$

and we also have $\mathbb{E}\tilde{\Delta}_{jj}^2 \leq C_1 n^{-1}\kappa_j^2$ and $\mathbb{E}(\tilde{\Delta}^2+\tilde{\Delta}_j^2)\leq C_2 n^{-1}$ where C_1 and C_2 do not depend on j. Observe that

$$\sum_{j=1}^{\varrho}|\tilde{\kappa}_j-\kappa_j|\leq\sum_{j=1}^{\varrho}|\tilde{\Delta}_j|+\tilde{\Delta}\sum_{j=1}^{\varrho}\delta_j^{-1}(\tilde{\Delta}+\tilde{\Delta}_j).$$

Further, $\sum_{j=1}^{\varrho}|\tilde{\Delta}_{jj}|$ is of order $O_p(n^{-1/2})$ since

$$\mathbb{E}\sum_{j=1}^{\varrho}|\tilde{\Delta}_{jj}|\leq\sum_{j=1}^{\varrho}\sqrt{\mathbb{E}\tilde{\Delta}_{jj}^2}\leq C_1 n^{-1/2}\sum_{j=1}^{\varrho}\kappa_j=O(n^{-1/2}),$$

and $\tilde{\Delta}\sum_{j=1}^{\varrho}\delta_j^{-1}(\tilde{\Delta}+\tilde{\Delta}_j)$ is or order $O_p(n^{-1}\varrho^{2r+2})$ since

$$\mathbb{E}\sum_{j=1}^{\varrho}|\delta_j^{-1}(\tilde{\Delta}+\tilde{\Delta}_j)| \leq \sum_{j=1}^{\varrho}\delta_j^{-1}\sqrt{2\mathbb{E}(\tilde{\Delta}^2+\tilde{\Delta}_j^2)} \leq \sqrt{2C_2n^{-1}}\sum_{j=1}^{\varrho}\delta_j^{-1} = O(n^{-1/2}\varrho^{2r+2}).$$

Hence,

$$\sum_{j=1}^{\varrho}\Big|\tilde{\kappa}_j-\kappa_j\Big| = O_p\Big(n^{-1/2}+n^{-1}\varrho^{2r+2}\Big).$$

On the other hand, since $E(\tilde{Q}-Q)^2 = O(n^{-1})$ uniformly on $[0,1]^2$,

$$\begin{aligned}\Big|\sum_{j=\varrho+1}^{\infty}(\tilde{\kappa}_j-\kappa_j)\Big| &= \Big|\int\int(\tilde{Q}-Q)(s,t)dsdt-\sum_{j=1}^{\varrho}(\tilde{\kappa}_j-\kappa_j)\Big| \\ &\leq [\int\int(\tilde{Q}-Q)^2]^{1/2}+\Big|\sum_{j=1}^{\varrho}\Big(\tilde{\kappa}_j-\kappa_j\Big)\Big| \\ &= O_p(n^{-1/2}+n^{-1}\varrho^{2r+2}).\end{aligned}$$

If we choose $\rho \asymp n^{1/(4r+1)}$, we have

$$\Big|\sum_{j=1}^{\varrho}(\tilde{\kappa}_j-\kappa_j)\Big| = O_p(n^{(-2r+1)/(4r+1)}), \qquad \Big|\sum_{j=\varrho+1}^{\infty}(\tilde{\kappa}_j-\kappa_j)\Big| = O_p(n^{(-2r+1)/(4r+1)}).$$

Define the event $\mathcal{E}_\varrho$ by

$$\mathcal{E}_\varrho = \mathcal{E}_\varrho(n) = \{\frac{1}{2}\kappa_\varrho \geq \tilde{\Delta}\}.$$

Since $\sup_{k\geq 1}|\tilde{\kappa}_k-\kappa_k| \leq \tilde{\Delta}$ [27], if $\mathcal{E}_\varrho$ holds, we have $\tilde{\kappa}_k \geq \frac{1}{2}\kappa_k$ for $1 \leq k \leq \varrho$. Here, we choose $\varrho \asymp n^{1/(4r+1)}$, which implies that $n^{1/2}\kappa_\varrho \to \infty$ as $n \to \infty$. Since $\tilde{\Delta} = O_p(n^{-1/2})$, we have $\mathbb{P}(\mathcal{E}_\varrho) \to 1$. Therefore, since the result we wish to prove only relates to probabilities of differences (not to moments of differences), it suffices to work with bounds that are established under the assumption that $\mathcal{E}_\varrho$ holds. The optimal choice $\tilde{\lambda}$ is the root of

$$\frac{1}{n}\sum_{k=1}^{\infty}\frac{\tilde{\kappa}_k}{(\sqrt{\lambda}+\tilde{\kappa}_k)^2} = 2\sqrt{\lambda}.$$

In the following, we derive the asymptotic order of $\sum_{k=1}^{\infty} \frac{\tilde{\kappa}_k}{(\sqrt{\lambda}+\tilde{\kappa}_k)^2}$. Note that

$$\begin{aligned}\sum_{k=1}^{\infty} \frac{\tilde{\kappa}_k}{(\sqrt{\lambda}+\tilde{\kappa}_k)^2} &= \sum_{k=1}^{\varrho} \frac{\tilde{\kappa}_k}{(\sqrt{\lambda}+\tilde{\kappa}_k)^2} + \sum_{k=\varrho+1}^{\infty} \frac{\tilde{\kappa}_k}{(\sqrt{\lambda}+\tilde{\kappa}_k)^2} \\ &\leq \sum_{k=1}^{\varrho} \tilde{\kappa}_k^{-1} + \lambda^{-1} \sum_{k=\varrho+1}^{\infty} \tilde{\kappa}_k \\ &\leq 2\sum_{k=1}^{\varrho} \kappa_k + \lambda^{-1} \sum_{k=\varrho+1}^{\infty} \kappa_k + \lambda^{-1} \Big| \sum_{j=\varrho+1}^{\infty} (\tilde{\kappa}_j - \kappa_j) \Big| \\ &= O_p\Big(n^{(2r+1)/(4r+1)} + \lambda^{-1} n^{(-2r+1)/(4r+1)} \Big). \end{aligned} \tag{2.12}$$

We also need the lower bound for $\sum_{k=1}^{\infty} \frac{\tilde{\kappa}_k}{(\sqrt{\lambda}+\tilde{\kappa}_k)^2}$. This follows from

$$\begin{aligned}\sum_{k=1}^{\infty} \frac{\tilde{\kappa}_k}{(\sqrt{\lambda}+\tilde{\kappa}_k)^2} &\geq \sum_{k=\varrho+1}^{\infty} \frac{\tilde{\kappa}_k}{(\sqrt{\lambda}+\tilde{\kappa}_k)^2} \\ &\geq \frac{1}{(\sqrt{\lambda}+\tilde{\kappa}_\varrho)^2} \sum_{k=\varrho+1}^{\infty} \tilde{\kappa}_k \\ &\geq \frac{1}{2(\lambda + O_p(n^{-4r/(4r+1)}))} O_p(n^{(-2r+1)/(4r+1)}). \end{aligned} \tag{2.13}$$

Combining (2.12) and (2.13), we obtain that $\tilde{\lambda}$ is of order $O_p(n^{4r/(4r+1)})$.

■

2.6.6 Proof of Proposition 2.2.1

In Theorem 2, we have shown that

$$tr(A) = O(\sum_{k=1}^{\infty} \frac{\hat{\kappa}_k}{\lambda + \hat{\kappa}_k}).$$

Define that $\tilde{Q} = T_0^m \hat{\Gamma} T_1^m$. Noting that the eigenvalues of $\hat{Q} = T_0^m \big(\hat{\Gamma} - \hat{\Gamma}_0\big) T_1^m$ and $\tilde{Q}$ have the same decay rate. If we write $\tilde{Q}(t,s) = \sum_{j=1}^{\infty} \tilde{\kappa}_j \tilde{\phi}_j(t) \tilde{\phi}_j(s)$, then $tr(A)$ is of the same order as $\sum_{k=1}^{\infty} \frac{\tilde{\kappa}_k}{\lambda+\tilde{\kappa}_k}$. On the other hand, recall that linear operator $Q = T_0^m \Gamma T_1^m$. Following spectral theorem, we have $Q(t,s) = \sum_{j=1}^{\infty} \kappa_j \phi_j(t) \phi_j(s)$. $\{\kappa_k\}$ and $\{s_k\}$ have the same decay rate. So we only need to show that $\sum_{k=1}^{\infty} \frac{\tilde{\kappa}_k}{\lambda+\tilde{\kappa}_k} = O_p(\sum_{k=1}^{\infty} \frac{\kappa_k}{\lambda+\kappa_k})$.

Let $Q^+ = (Q + \lambda I)^{-1}$ and $\tilde{Q}^+ = (\tilde{Q} + \lambda I)^{-1}$. It is easy to see that $Q^+(t,s) = \sum_{j=1}^{\infty} \frac{1}{\lambda+\kappa_j}\phi_j(t)\phi_j(s)$ and $\tilde{Q}^+(t,s) = \sum_{j=1}^{\infty} \frac{1}{\lambda+\tilde{\kappa}_j}\tilde{\phi}_j(t)\tilde{\phi}_j(s)$. Then

$$\begin{aligned}
\sum_{k=1}^{\infty} \frac{\tilde{\kappa}_k}{\lambda + \tilde{\kappa}_k} &= \int\int \tilde{Q}(s,t)\tilde{Q}^+(s,t)dsdt \\
&= \int\int Q(s,t)Q^+(s,t)dsdt \\
&\quad + \int\int Q^+(s,t)(\tilde{Q} - Q)(s,t)dsdt + \int\int (\tilde{Q}^+ - Q^+)(s,t)Q(s,t)dsdt \\
&\quad + \int\int (\tilde{Q} - Q)(s,t)(\tilde{Q}^+ - Q^+)(s,t)dsdt.
\end{aligned}$$

We are going to show that all four terms above in the last equation are either of the same order of or of $\sum_{k=1}^{\infty} \frac{\kappa_k}{\lambda+\kappa_k}$ or smaller than that.

For the first term, it is easy to see that

$$\int\int Q(s,t)Q^+(s,t)dsdt = \sum_{k=1}^{\infty} \frac{\kappa_k}{\lambda + \kappa_k}.$$

For the second term, let $\Delta(s,t) = (\tilde{Q} - Q)(s,t)$ and $\hat{\Delta}_{jk} = |\int\int \Delta(s,t)\phi_j(s)\phi_k(t)dsdt|$. It follows Section 5.3 of [26] that

$$\hat{\Delta}_{jj} = |\int\int \Delta(s,t)\phi_j(s)\phi_j(t)| = O_p(n^{-1/2}\kappa_j).$$

And similarly we can show that $\hat{\Delta}_{jk} = O_p(n^{-1/2}\kappa_j^{1/2}\kappa_k^{1/2})$ for any $j \neq k$, which will be used later in calculating the order of the fourth term. The second term becomes

$$\begin{aligned}
&\int\int Q^+(s,t)(\tilde{Q} - Q)(s,t)dsdt \\
&= \sum_{k=1}^{\infty} \frac{1}{\lambda + \kappa_k} \int\int \Delta(s,t)\phi_j(s)\phi_j(t)dsdt \\
&\leq O_p(n^{-1/2} \sum_{k=1}^{\infty} \frac{\kappa_k}{\lambda + \kappa_k}).
\end{aligned}$$

For the third term, we refer to (6.7) of [28] that $||(I + Q^+\Delta)^{-1}|| = O_p(1)$. Here $||\cdot||$ as a norm of a functional from $L_2[0,1]$ to itself, is defined as

$$||\chi|| = \sup_{\phi \in L_2[0,1], ||\phi||=1} ||\chi(\phi)||.$$

Noting that $\tilde{Q}^+ - Q^+ = -(I + Q^+\Delta)^{-1}Q^+\Delta Q^+$, then

$$\begin{aligned}
&\int\int(\tilde{Q}^+ - Q^+)(s,t)Q(s,t)dsdt \\
&= -\sum_{k=1}^{\infty}\kappa_k \int\int (I + Q^+\Delta)^{-1}Q^+\Delta Q^+(s,t)\phi_k(s)\phi_k(t)dsdt \\
&= -\sum_{k=1}^{\infty}\frac{\kappa_k}{\lambda+\kappa_k} \int\int (I + Q^+\Delta)^{-1}Q^+\Delta(s,t)\phi_k(s)\phi_k(t)dsdt \\
&\leq \sum_{k=1}^{\infty}\frac{\kappa_k}{\lambda+\kappa_k} ||(I + Q^+\Delta)^{-1}Q^+\Delta(s,t)\phi_k(s)||\,||\phi_k(t)|| \\
&= O_p(\sum_{k=1}^{\infty}\frac{\kappa_k}{\lambda+\kappa_k}).
\end{aligned}$$

The last equation follows from the fact that

$$\begin{aligned}
&||(I + Q^+\Delta)^{-1}Q^+\Delta(s,t)\phi_k(s)|| \\
&\quad = ||\phi_k(s) - (I + Q^+\Delta)^{-1}\phi_k(s)|| \\
&\quad \leq ||\phi_k(s)|| + ||(I + Q^+\Delta)^{-1}||\,||\phi_k(s)|| \\
&\quad = 1 + ||(I + Q^+\Delta)^{-1}||.
\end{aligned}$$

For the last term, by Cauchy-Schwarz inequality

$$\begin{aligned}
&\int\int(\tilde{Q} - Q)(s,t)(\tilde{Q}^+ - Q^+)(s,t)dsdt \\
&\qquad \leq \Big\{\int\int(\tilde{Q} - Q)^2(s,t)dsdt \cdot \int\int(\tilde{Q}^+ - Q^+)^2(s,t)dsdt\Big\}^{1/2} \\
&\qquad \leq n^{-1/2}\Big\{\int\int(\tilde{Q}^+ - Q^+)^2(s,t)dsdt\Big\}^{1/2} \\
&\qquad = n^{-1/2}\Big\{\sum_{k=1}^{\infty}|| - (I + Q^+\Delta)^{-1}Q^+\Delta Q^+\phi_k||^2\Big\}^{1/2} \\
&\qquad \leq n^{-1/2}||(I + Q^+\Delta)^{-1}||^{-1/2}\Big\{\sum_{k=1}^{\infty}||Q^+\Delta Q^+\phi_k||^2\Big\}^{1/2}.
\end{aligned}$$

Recall that $\hat{\Delta}_{jk} = O_p(n^{-1/2}\kappa_j^{1/2}\kappa_k^{1/2})$ for any $j \neq k$. Then,

$$\begin{aligned}
||Q^+\Delta Q^+\phi_k||^2 &= \int \Big\{ \int\int Q^+\Delta(s,u) \sum_{j=1}^{\infty} \frac{1}{\lambda+\kappa_j}\phi_j(u)\phi_j(t)\phi_k(t)dtdu \Big\}^2 ds \\
&= \frac{1}{(\lambda+\kappa_k)^2} \int \Big\{ \int Q^+\Delta(s,u)\phi_k(u)du \Big\}^2 ds \\
&= \frac{1}{(\lambda+\kappa_k)^2} \int \Big\{ \int\int \sum_{j=1}^{\infty} \frac{1}{\lambda+\kappa_j}\phi_j(s)\phi_j(v)\Delta(v,u)\phi_k(u)dudv \Big\}^2 ds \\
&= \frac{1}{(\lambda+\kappa_k)^2} \sum_{j=1}^{\infty} \frac{\hat{\Delta}_{jk}^2}{(\lambda+\kappa_j)^2} \\
&= O_p\Big(\frac{\kappa_k}{(\lambda+\kappa_k)^2} n^{-1} \sum_{j=1}^{\infty} \frac{\kappa_j}{(\lambda+\kappa_j)^2}\Big) \\
&= O_p\Big(n^{-1}\lambda^{-2}\Big(\sum_{j=1}^{\infty} \frac{\kappa_j}{\lambda+\kappa_j}\Big)\frac{\kappa_k}{\lambda+\kappa_k}\Big).
\end{aligned}$$

Therefore

$$\int\int (\tilde{Q}-Q)(s,t)(\tilde{Q}^+-Q^+)(s,t)dsdt = O_p(n^{-1}\lambda^{-1}\sum_{j=1}^{\infty}\frac{\kappa_j}{\lambda+\kappa_j}).$$

All together we show that $\sum_{k=1}^{\infty}\frac{\tilde{\kappa}_k}{\lambda+\tilde{\kappa}_k} = O_p(\sum_{k=1}^{\infty}\frac{\kappa_k}{\lambda+\kappa_k})$ provided that $\lambda^{-1} = O(n)$.

■

2.6.7 Proof of Proposition 2.3.1

Let $\mathcal{D}_k = \text{span}\{f_1,\ldots,f_k : K^{1/2}f_j = T^*\varphi_j, j = 1,\ldots,k\}$. It follows from the minimax principal that

$$\begin{aligned}
\tilde{s}_k &\leq \sup_{f\in\mathcal{D}_k^{\perp}} \frac{\Big\langle K^{1/2}\Gamma K^{1/2}f,\ f\Big\rangle}{\langle f,f\rangle} = \sup_{f\in\mathcal{D}_k^{\perp}} \frac{\Big\langle \Gamma K^{1/2}f,\ K^{1/2}f\Big\rangle}{\langle f,f\rangle} \\
&= \sup_{g\in\tilde{\mathcal{D}}_k^{\perp}} \frac{\Big\langle \Gamma T^*g,\ T^*g\Big\rangle}{\langle g,g\rangle}\frac{\langle g,g\rangle}{\langle T^*g,T^*g\rangle} \leq cs_k,
\end{aligned}$$

where $\tilde{s}_k$ is the kth eigenvalue of $K^{1/2}\Gamma K^{1/2}$, $\tilde{\mathcal{D}}_k = \text{span}\{\varphi_1,\ldots,\varphi_k\}$ and the constant $c > 0$ does not depend on k. Using a similar argument, we may show that $s_k \leq c\tilde{s}_k$. Therefore, the eigenvalues of $T\Gamma T^*$ and $K^{1/2}\Gamma K^{1/2}$ have the same decay rate. ■

2.6.8 Proof of Lemmas

Lemma 2.1 *The following statements are true:*

(a). The $\beta \in W_2^m$ minimizes $L(\beta)$, if and only if, $L_1(\beta, \beta_1) = 0$ for all $\beta_1 \in W_2^m$.

(b). If $\beta \in W_2^m$ minimizes $L(\beta)$, then for all $\beta_1 \in W_2^m$,

$$L_1(\beta, \beta_1) = \int_0^1 L_2(\beta)(t)\ \beta_1^{(m)}(t)dt, \tag{2.14}$$

where

$$L_2(\beta) = (\lambda I + \hat{Q})\beta^{(m)} - \frac{(-1)^m}{n}\hat{U}^T\mathbf{Y}. \tag{2.15}$$

Proof First show part (a). If $\hat{\beta} \in W_2^m$ minimizes $L(\beta)$, then $L(\hat{\beta} + \delta\beta_1) - L(\hat{\beta}) \geq 0$ for all $\beta_1 \in W_2^m$ and any $\delta \in \mathbb{R}$. Then $L_1(\hat{\beta}, \beta_1) = 0$ follows since δ can be either negative or positive. On the other hand, if $L_1(\hat{\beta}, \beta_1) = 0$, we have $L(\hat{\beta}+\delta\beta_1)-L(\hat{\beta}) \geq 0$ by (2.9). Thus, $\hat{\beta}$ minimizes $L(\beta)$. Therefore, part (a) follows.

Let $\beta_1(t) = t^{(k-1)}$, $k = 1, \ldots, m$ in (2.10). If $\hat{\beta}$ minimizes $L(\beta)$, then

$$\frac{1}{n}\sum_{i=1}^n \big\{Y_i - \int_0^1 X_i(s)\hat{\beta}(s)ds\big\}\big\{\int_0^1 X_i(s)s^{(k-1)}ds\big\} = 0. \tag{2.16}$$

Let $X_i^{(-k)}(t) = T_0^k X_i(t) = \int_0^1 \frac{(t-s)_+^{(k-1)}}{(k-1)!}X_i(s)ds$. When $k = 1$, $\int_0^1 X_i(s)s^{(k-1)}ds = X_i^{(-1)}(1)$ and further (2.16) becomes

$$\frac{1}{n}\sum_{i=1}^n X_i^{(-1)}(1)\big\{Y_i - \int_0^1 X_i(s)\hat{\beta}(s)ds\big\} = 0.$$

When $k = 2$, we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n\big\{Y_i - \int_0^1 X_i(s)\hat{\beta}(s)ds\big\}\big\{\int_0^1 X_i(s)sds\big\} \\
&\quad = -\frac{1}{n}\sum_{i=1}^n \big\{Y_i - \int_0^1 X_i(s)\hat{\beta}(s)ds\big\}\big\{\int_0^1 X_i(s)(1-s)ds\big\} \\
&\quad = -\frac{1}{n}\sum_{i=1}^n \big\{Y_i - \int_0^1 X_i(s)\hat{\beta}(s)ds\big\}\big\{X_i^{(-2)}(1)\big\}.
\end{aligned}$$

Hence,

$$\frac{1}{n}\sum_{i=1}^n X_i^{(-2)}(1)\big\{Y_i - \int_0^1 X_i(s)\hat{\beta}(s)ds\big\} = 0.$$

Following the same procedure, it can be shown that

$$\frac{1}{n}\sum_{i=1}^{n} X_i^{(-k)}(1)\big\{Y_i - \int_0^1 X_i(s)\hat{\beta}(s)ds\big\} = 0,\ k = 1, \ldots, m. \tag{2.17}$$

Considering that

$$\beta_1(s) = \sum_{k=0}^{m-1}(-1)^k \frac{\beta_1^{(k)}(1)}{k!}(1-s)^k + (-1)^m \int_0^1 \frac{\beta_1^{(m)}(t)}{(m-1)!}(t-s)_+^{m-1}dt.$$

Therefore

$$\begin{aligned}
&\int_0^1 X_i(s)\beta_1(s)ds \\
&\quad = \sum_{k=0}^{m-1}(-1)^k \beta_1^{(k)}(1)\int_0^1 X_i(s)\frac{(1-s)^k}{k!}ds \\
&\qquad + (-1)^m \int_0^1\int_0^1 X_i(s)\frac{\beta_1^{(m)}(t)}{(m-1)!}(t-s)_+^{m-1}dtds \\
&\quad = \sum_{k=1}^{m}(-1)^{k-1}\beta_1^{(k-1)}(1)\ X_i^{(-k)}(1) + (-1)^m \int_0^1 \beta_1^{(m)}(t)X_i^{(-m)}(t)dt.
\end{aligned} \tag{2.18}$$

If (2.17) holds, direct calculation yields

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\big\{Y_i - \int_0^1 X_i(s)\hat{\beta}(s)ds\big\}\big\{\int_0^1 X_i(s)\beta_1(s)ds\big\} \\
&\qquad = \frac{(-1)^m}{n}\sum_{i=1}^{n}\big\{Y_i - \int_0^1 X_i(s)\hat{\beta}(s)ds\big\}\big\{\int_0^1 X_i^{(-m)}(t)\beta_1^{(m)}(t)dt\big\}.
\end{aligned}$$

Recall the definition of $L_2(\beta)$, we have

$$\begin{aligned}
L_2(\hat{\beta}) &= \lambda\ \hat{\beta}^{(m)}(t) + \frac{(-1)^m}{n}\sum_{i=1}^{n} X_i^{(-m)}(t)\big\{\int_0^1 X_i(s)\hat{\beta}(s)ds - Y_i\big\} \\
&= \lambda\ \hat{\beta}^{(m)}(t) + \frac{(-1)^m}{n}T_0^{(m)}\mathbf{X}(t)^T\Big\{\int_0^1 \mathbf{X}(s)\hat{\beta}(s)ds - \mathbf{Y}\Big\}.
\end{aligned}$$

Similar to (2.18),

$$\int_0^1 X_i(s)\hat{\beta}(s)ds = \widehat{\Upsilon}(1)^T\widetilde{X}_i(1) + (-1)^m\int_0^1 X_i^{(-m)}(s)\hat{\beta}^{(m)}(s)ds, \quad j = 1, \ldots, m.$$

which gives

$$\int_0^1 \mathbf{X}(s)\hat{\beta}(s)ds = \widetilde{X}(1)^T\widehat{\Upsilon}(1) + (-1)^m \int_0^1 T_0^m \mathbf{X}(s)\hat{\beta}^{(m)}(s)ds.$$

This, combining (2.17), gives

$$\hat{H}\widehat{\Upsilon}(1) + \frac{(-1)^m}{n}\widetilde{X}(1)\int_0^1 T_0^m \mathbf{X}(s)\hat{\beta}^{(m)}(s)ds = \frac{1}{n}\widetilde{X}(1)\mathbf{Y}. \tag{2.19}$$

So for $\beta \in W_2^m$ minimizes $L(\beta)$,

$$L_2(\beta) = \lambda\ \beta^{(m)} + \hat{Q}\beta^{(m)} - \frac{(-1)^m}{n}\hat{U}^T\mathbf{Y}.$$

So, part (b) follows. ■

Lemma 2.2 *Let $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$. The following statements hold:*

(a)

$$\int_0^1 \mathbf{X}(t)\Big\{\hat{\beta}(t) - \beta_0(t)\Big\}dt \tag{2.20}$$
$$= (-1)^{m+1}\lambda\int_0^1 \hat{U}(t)\hat{Q}^+\beta_0^{(m)}(t)dt + \frac{1}{n}\Big\{\int_0^1 \hat{U}(t)\hat{Q}^+\hat{U}(t)^T dt + \widetilde{X}(1)^T\hat{H}^{-1}\widetilde{X}(1)\Big\}\epsilon;$$

(b)

$$\Big\|\hat{\beta} - \beta_0\Big\|_{\hat{\Gamma}}^2 = \lambda^2\int_0^1\int_0^1 \hat{Q}(t,s)\hat{Q}^+\beta_0^{(m)}(t)\hat{Q}^+\beta_0^{(m)}(s)dtds \tag{2.21}$$
$$+ \frac{1}{n^2}\epsilon^T\Big\{\int_0^1\int_0^1 \hat{Q}^+\hat{U}(s)\hat{Q}(t,s)\hat{Q}^+\hat{U}(t)^T dsdt + \widetilde{X}(1)^T\hat{H}^{-1}\widetilde{X}(1)\Big\}\epsilon$$
$$+ (-1)^{m+1}\frac{2\lambda}{n}\epsilon^T\int_0^1\int_0^1 \hat{Q}(t,s)\hat{Q}^+\beta_0^{(m)}(t)\hat{Q}^+\hat{U}(s)dtds.$$

Proof Denote

$$\Upsilon_0(1) = \Big[\beta_0(1), -\beta_0'(1), \ldots, (-1)^{m-1}\beta_0^{(m-1)}(1)\Big]^T.$$

Direct calculation yields

$$\frac{1}{n}\widetilde{X}(1)\mathbf{Y} = \hat{H}\Upsilon_0(1) + (-1)^m\frac{1}{n}\widetilde{X}(1)\int_0^1 T_0^m\mathbf{X}(s)\beta_0^{(m)}(s)ds + \frac{1}{n}\widetilde{X}(1)\epsilon.$$

Combining this with (2.19) gives

$$\widehat{\Upsilon}(1)-\Upsilon_0(1)=(-1)^{m+1}\frac{1}{n}\hat{H}^{-1}\widetilde{X}(1)\int_0^1 T_0^m X(s)\Big\{\hat{\beta}^{(m)}(s)-\beta_0^{(m)}(s)\Big\}ds+\frac{1}{n}\hat{H}^{-1}\widetilde{X}(1)\epsilon.$$

Therefore,

$$\begin{aligned}
&\int_0^1 \mathbf{X}(s)\Big\{\hat{\beta}(s)-\beta_0(s)\Big\}ds\\
&\quad=\widetilde{X}(1)^T\Big\{\widehat{\Upsilon}(1)-\Upsilon_0(1)\Big\}+(-1)^m\int_0^1 T_0^m\mathbf{X}(s)\Big\{\hat{\beta}^{(m)}(s)-\beta_0^{(m)}(s)\Big\}ds\\
&\quad=(-1)^m\int_0^1\hat{U}(s)\Big\{\hat{\beta}^{(m)}(s)-\beta_0^{(m)}(s)\Big\}ds+\frac{1}{n}\widetilde{X}(1)^T\hat{H}^{-1}\widetilde{X}(1)\epsilon. \qquad (2.22)
\end{aligned}$$

Recall that $\hat{Q}^+=(\lambda I+\hat{Q})^{-1}$. It follows from Theorem 1 that

$$\begin{aligned}
\hat{\beta}^{(m)}-\beta_0^{(m)}&=(-1)^m n^{-1}\mathbf{Y}^T\hat{Q}^+\hat{U}-\beta_0^{(m)}\\
&=\frac{(-1)^m}{n}\hat{Q}^+\hat{U}^T\{\int_0^1\mathbf{X}(s)\beta_0(s)ds\}-\beta_0^{(m)}+(-1)^m\frac{1}{n}\epsilon^T\hat{Q}^+\hat{U}\\
&=\frac{(-1)^m}{n}\hat{Q}^+\hat{U}^T\widetilde{X}(1)^T\Upsilon_0(1)+\hat{Q}^+\hat{Q}\beta_0^{(m)}-\beta_0^{(m)}+(-1)^m\frac{1}{n}\epsilon^T\hat{Q}^+\hat{U}\\
&=-\lambda\hat{Q}^+\beta_0^{(m)}+(-1)^m\frac{1}{n}\epsilon^T\hat{Q}^+\hat{U}.
\end{aligned}$$

The last equation follows from the fact that $\widetilde{X}(1)\hat{U}(s)=0$. Then, this, combing with (2.22), leads to part (a). Furthermore, part (b) follows that

$$\Big\|\hat{\beta}-\beta_0\Big\|_{\hat{\Gamma}}^2=\frac{1}{n}\Big[\int_0^1\mathbf{X}(t)^T\Big\{\hat{\beta}(t)-\beta_0(t)\Big\}dt\Big]\ \Big[\int_0^1\mathbf{X}(s)\Big\{\hat{\beta}(s)-\beta_0(s)\Big\}ds\Big].$$

This completes the proof of the lemma. ■

3. OPTIMAL ESTIMATION FOR THE FUNCTIONAL COX MODEL

3.1 Introduction

3.1.1 Background

The proportional hazard model, known as the Cox model, was introduced by [29], where the hazard function of the survival time T for a subject with covariate $Z(t) \in \mathbb{R}^p$ is represented by

$$h(t|Z) = h_0(t)e^{\theta_0' Z(t)}, \tag{3.1}$$

where h_0 is an unspecified baseline hazard function and $\theta_0 \in \mathbb{R}^p$ is an unknown parameter. Some or all of the p components in Z may be time-independent, meaning that they are constant over time t, or may depend on t.

Many people have studied parametric, nonparametric, or semiparametric modeling of the covariate effects using the Cox model (e.g. [30–34] and references therein) and Cox ([29]) proposed to use partial likelihood to estimate θ in (3.1). The advantage of using partial likelihood is that it estimates θ without knowing or involving the functional form of h_0. The asymptotic equivalence of the partial likelihood estimator and the maximum likelihood estimator has been established by several authors ([35–39]).

3.1.2 Functional Cox Model

The aim of my work is to develop a different type of model, the functional Cox model, by incorporating functional predictors along with scalar predictors. [40] first proposed such a model when studying the survival of diffuse large-B-cell lymphoma

(DLBCL) patients, which is thought to be influenced by genetic differences. The functional predictor, denoted by $X(\cdot) : \mathcal{S} \to \mathbb{R}$ on a compact domain $\mathcal{S}$, is a smooth stochastic process related to the high-dimensional microarray gene expression of DLBCL patients. The entire trajectory of X has an effect on the hazard function, which makes it different from the Cox model (3.1) with time-varying covariates, where only the current value of X at time t affects the hazard function at time t.

Specifically, the functional Cox model with a vector covariate Z and functional covariate $X(t)$ represents the hazard function by

$$h(t\,|X) = h_0(t) \exp\Big\{\theta_0' Z + \int_{\mathcal{S}} X(s)\beta_0(s)ds\Big\}, \tag{3.2}$$

where β_0 is an unknown coefficient function. Without loss of generality, we take $\mathcal{S}$ to be $[0, 1]$.

Under the right censorship model and letting T^u and T^c be, respectively, the failure time and censoring time, we observe i.i.d. copies of $(T,\, \Delta,\, X(s), s \in \mathcal{S})$, $(T_1,\, \Delta_1,\, X_1), \ldots, (T_n,\, \Delta_n,\, X_n)$, where $T = \min\{T^u,\, T^c\}$ is the observed time event and $\Delta = I\{T^u \le T^c\}$ is the censoring indicator.

3.1.3 Problem statement

Our goal is to estimate $\alpha_0 = (\theta_0, \beta_0(\cdot))$ to reveal how the functional covariates $X(\cdot)$ and other scalar covariates Z relate to survival.

Let $\hat{\alpha} = (\hat{\theta}, \hat{\beta}(\cdot))$ be an estimate from the data. It is critical to define the risk function to measure the accuracy of the estimate. Let $W = (Z, X)$ and

$$\eta_\alpha(W) = \theta' Z + \int_0^1 \beta(s)X(s)ds.$$

Define an L_2-distance such that

$$d^2(\hat{\alpha}, \alpha_0) = \mathbb{E}\Big\{\Delta\Big(\eta_{\hat{\alpha}}(W) - \eta_{\alpha_0}(W)\Big)^2\Big\}. \tag{3.3}$$

Based on this L_2-distance, we show that the accuracy of $\hat{\theta}$ is measured by the usual L_2-norm $\|\hat{\theta}-\theta\|_2$ and the accuracy of $\hat{\beta}$ is measured by a weighted L_2-norm $||\hat{\beta}-\beta_0||_{C_\Delta}$, where

$$C_\Delta(s,t) = \mathrm{Cov}\Big(\Delta X(s),\ \Delta X(t)\Big), \quad \text{and} \quad \|\beta\|^2_{C_\Delta} = \int\int \beta(s)C_\Delta(s,t)\beta(t)dsdt.$$

It worth noting that we do not consider the convergence of $\hat{\beta}$ with respect to the usual L_2-norm in the present paper. In general, $\|\hat{\beta}-\beta_0\|_2^2 = \int_0^1(\hat{\beta}(t)-\beta_0(t))^2dt$ may not converge to zero in probability, and to obtain the convergence of $\|\hat{\beta}-\beta_0\|_2^2$ one needs additional smoothness conditions linking β to the functional predictor X; see [41] for a discussion of this phenomenon for functional linear models. On the other hand, in the presence of censoring, the Kullback-Leibler distance between two probability measures $\mathbb{P}_{h_0,\hat{\alpha}}$ and $\mathbb{P}_{h_0,\alpha_0}$ is equivalent to the L_2 distance d in (3.3). When failure times T^u are fully observed, i.e. $\Delta = 1$ is true regardless of $X(s)$, the $\|\cdot\|_{C_\Delta}$ norm becomes $\|\cdot\|_C$, where $C(t,s) = \mathrm{Cov}(X(t),X(s))$ is the covariance function of X. This norm $\|\cdot\|_C$ has been widely used for functional linear models (e.g. [42]).

Recently, [43] studied a similar functional Cox model to establish some asymptotic properties but without investigating the optimality property. Moreover, their estimate of the parametric component converges at a rate which is slower than root-n. Thus, it is desirable to develop new theory to systematically investigate properties of the estimates and establish their optimal asymptotic properties. In addition, instead of assuming that both β_0 and X can be represented by the same set of basis functions, we adopt a more general reproducing kernel Hilbert space framework to estimate the coefficient function.

In this chapter, we will discuss the convergence of the estimator $\hat{\alpha} = (\hat{\theta},\hat{\beta})$ under the framework of the reproducing kernel Hilbert space and the Cox model. The true coefficient function β_0 is assumed to reside in a reproducing kernel Hilbert space $\mathcal{H}(K)$ with the reproducing kernel K, which is a subspace of the collection of square integrable functions on $[0,1]$. There are two main challenges for our asymptotic analysis, the nonlinear structure of the Cox model, and the fact that the reproducing kernel K and the covariance kernel C_Δ may not share a common ordered set of

eigenfunctions, so β_0 can not be represented effectively by the leading eigenfunctions of C_Δ. We obtain the estimator by maximizing a penalized partial likelihood and establish $\sqrt{n}$-consistency, asymptotic normality, and semi-parametric efficiency of the estimator $\hat{\theta}$ of the finite-dimensional regression parameter.

A second optimality result is on the estimator of the coefficient function, which achieves the minimax optimal rate of convergence under the weighted L_2-risk. The optimal rate of convergence is established in the following two steps. First, the convergence rate of the penalized partial likelihood estimator is calculated. Second, in the presence of the nuisance parameter h_0, the minimax lower bound on the risk is derived, which matches the convergence rate of the partial likelihood estimator. Therefore the estimator is rate-optimal. Furthermore, an efficient algorithm is developed to estimate the coefficient function. Implementation of the estimation approach, selection of the smoothing parameter, as well as calculation of the information bound $I(\theta)$ are all discussed in detail.

3.2 Main Results

We estimate $\alpha_0 = (\theta_0, \beta_0) \in \mathbb{R}^p \times \mathcal{H}(K)$ by maximizing the penalized log partial likelihood,

$$\hat{\alpha}_\lambda = \arg\min_{\alpha \in \mathbb{R}^p \times \mathcal{H}(K)} l_n(\alpha) + \lambda\ J(\beta), \tag{3.4}$$

where the negative log partial likelihood is given by

$$l_n(\alpha) = -\frac{1}{n}\sum_{i=1}^{n} \Delta_i \Big\{ \eta_\alpha(W_i) - \log \sum_{T_j \geq T_i} \exp(\eta_\alpha(W_j)) \Big\}, \tag{3.5}$$

J is a penalty function controlling the smoothness of β, and λ is a smoothing parameter that balances the fidelity to the model and the plausibility of β. The choice of the penalty function $J(\cdot)$ is a squared semi-norm associated with $\mathcal{H}$ and its norm. In general, $\mathcal{H}(K)$ can be decomposed with respect to the penalty J as $\mathcal{H} = \mathcal{N}_J + \mathcal{H}_1$, where $\mathcal{N}_J$ is the null space defined as

$$\mathcal{N}_J = \{\beta \in \mathcal{H}(K) :\ J(\beta) = 0\},$$

and $\mathcal{H}_1$ is its orthogonal complement in $\mathcal{H}$. Correspondingly, the kernel K can be decomposed as $K = K_0 + K_1$, where K_0 and K_1 are kernels for the subspace $\mathcal{N}_J$ and $\mathcal{H}_1$ respectively. For example, for the Sobolev space,

$$\mathcal{W}_{2,m} = \Big\{ f : [0,1] \to R |\ f, f', \dots f^{(m-1)} \text{are absolutely continuous, } f^{(m)} \in L_2 \Big\},$$

endowed with the norm

$$||f||_{\mathcal{W}_{2,m}} = \sum_{v=0}^{m-1} f^{(v)}(0) + \int_0^1 (f^{(m)}(s))^2 ds, \tag{3.6}$$

where the penalty $J(\cdot)$ in this case can be assigned as $J(f) = \int_0^1 (f^{(m)}(s))^2 ds$.

We first present some main assumptions:

(A1) Assume $\mathbb{E}(\Delta Z) = 0$ and $\mathbb{E}(\Delta X(s)) = 0, s \in [0,1]$.

(A2) The failure time T^u and the censoring time T^c are conditionally independent given W.

(A3) The observed event time $T_i,\ 1 \le i \le n$ is in a finite interval, say $[0, \tau]$, and there exists a small positive constant ε such that: (i) $\mathbb{P}(\Delta = 1|W) > \varepsilon$, and (ii) $\mathbb{P}(T^c > \tau|W) > \varepsilon$ almost surely with respect to the probability measure of W.

(A4) The covariate Z takes values in a bounded subset of $\mathbb{R}^p$, and the L_2-norm $||X||_2$ of X is bounded almost surely.

(A5) Let $0 < c_1 < c_2 < \infty$ be two constants. The baseline joint density $f(t, \Delta = 1)$ of $(T, \Delta = 1)$ satisfies $c_1 < f(t, \Delta = 1) < c_2$ for all $t \in [0, \tau]$.

Condition (A1) requires Z and X to be suitably centered. Since the partial likelihood function (3.5) does not change when centering Z_i as $Z_i - \sum \Delta_i Z_i / \sum \Delta_i$ or X_i as $X_i - \sum \Delta_i X_i / \sum \Delta_i$, centering does not impose any real restrictions. In addition, centering by $\mathbb{E}(\Delta Z)$ and $\mathbb{E}(\Delta X)$, instead of centering by $\mathbb{E}(Z)$ and $\mathbb{E}(X)$, simplifies the asymptotic analysis. Conditions (A2) and (A3) are common assumptions for analyzing right-censored data, where (A2) guarantees the censoring mechanism

to be non-informative while (A3) avoids the unboundedness of the partial likelihood at the end point of the support of the observed event time. This is a reasonable assumption since the experiment can only last for a certain amount of time in practice. Assumption (A3)(i) further ensures the probability of being uncensored to be positive regardless of the covariate and (A3)(ii) controls the censoring rate so that it will not be too heavy. Assumption (A4) places a boundedness restriction on the covariates. This assumption can be relaxed to the sub-Gaussianity of $||X||_2$, which implies that with a large probability, $||X||_2$ is bounded. Condition (A5) and condition (A1) together guarantee the identifiability of the model. Moreover the joint density $f(T, Z, X, \Delta = 1)$ is bounded away from zero and infinity under assumptions (A3)-(A5), which is used to calculate the information bound and convergence rate later in Theorem 3.2.1 and Theorem 3.2.2.

Let $r(W) = \exp(\eta_\alpha(W))$, then the counting process martingale associated with model (1) is:

$$M(t) = M(t|W) = \Delta I\{T \leq t\} - \int_0^t I\{T \geq u\} r(W) dH_0(u),$$

where $H_0(t) = \int_0^t h_0(u)du$ is the baseline cumulative hazard function. For two sequences $a_k : k \geq 1$ and $b_k : k \geq 1$ of positive real numbers, we write $a_k \asymp b_k$ if there are positive constants c and C independent of k such that $c \leq a_k/b_k \leq C$ for all $k \geq 1$.

Theorem 3.2.1 *Under (A1)-(A5), the efficient score for the estimation of θ is*

$$l^*_\theta(T, \Delta, W) = \int_0^\tau (Z - a^*(t) - \eta_{g^*}(X)) dM(t)$$

where $(a^*, g^*) \in L_2 \times \mathcal{H}(K)$ is a solution that minimizes

$$\mathbb{E}\Big\{\Delta \|Z - a(T) - \eta_g(X)\|^2\Big\}.$$

Here a^* can be expressed as $a^*(t) = \mathbb{E}[Z - \eta_{g^*}(X)|\, T = t, \Delta = 1]$. The information bound for the estimation of θ is

$$I(\theta) = \mathbb{E}[l^*_\theta(T, \Delta, W)]^{\otimes 2} = \mathbb{E}\{\Delta[Z - a^*(T) - \eta_{g^*}(X)]^{\otimes 2}\},$$

where $y^{\otimes 2} = yy'$ *for column vector* $y \in \mathbb{R}^d$.

Recall that K and C_Δ are two real, symmetric, and nonnegative definite functions. Define a new kernel $K^{1/2}C_\Delta K^{1/2} : [0,1]^2 \to \mathbb{R}$, which is a real, symmetric, square integrable, and nonnegative definite function. Let $L_{K^{1/2}C_\Delta K^{1/2}}$ be the corresponding linear operator $L_2 \to L_2$. Then Mercers theorem [4] implies that there exists a set of orthonomal eigenfunctions $\{\phi_k : k \geq 1\}$ and a sequence of eigenvalues $s_1 \geq s_2 \geq \ldots > 0$ such that

$$K^{1/2}C_\Delta K^{1/2}(s,t) = \sum_{k=1}^{\infty} s_k \phi_k(s)\phi_k(t), \qquad L_{K^{1/2}C_\Delta K^{1/2}}(\phi_k) = s_k.$$

Theorem 3.2.2 *Assume (A1)-(A5) hold.*

(i) (consistency) $d(\hat{\alpha}, \alpha_0) \xrightarrow{p} 0$, *provided that* $\lambda \to 0$ *as* $n \to \infty$.

(ii) (convergence rate) If the eigenvalues $\{s_k : k \geq 1\}$ *of* $K^{1/2}C_\Delta K^{1/2}$ *satisfy* $s_k \asymp k^{-2r}$ *for some constant* $0 < r < \infty$, *then for* $\lambda = O(n^{-\frac{2r}{2r+1}})$ *we have*

$$d(\hat{\alpha}, \alpha_0) = O_p(n^{-\frac{r}{2r+1}}).$$

(iii) If $I(\theta)$ *is nonsingular, then* $\|\hat{\theta} - \theta_0\|_2 = O_p(n^{-\frac{r}{2r+1}})$ *and*

$$\lim_{A\to\infty} \lim_{n\to\infty} \sup_{\beta_0 \in \mathcal{H}(K)} \mathbb{P}_{h_0\beta_0}\left\{ \|\hat{\beta}_\lambda - \beta_0\|_{C_\Delta} \geq An^{-\frac{r}{2r+1}} \right\} = 0.$$

Theorem 3.2.2 indicates that the convergence rate is determined by the decay rate of the eigenvalues of $K^{1/2}C_\Delta K^{1/2}$, which is jointly determined by the eigenvalues of both reproducing kernel K and the conditional covariance function C_Δ as well as by the alignment between K and C_Δ. When K and C_Δ are perfectly aligned, meaning that K and C_Δ have the same ordered eigenfunctions, the decay rate of $\{s_k : k \geq 1\}$ equals to the summation of the decay rates of the eigenvalues of K and C_Δ. [42] established a similar result for functional linear models, for which the optimal prediction risk depends on the decay rate of the eigenvalues of $K^{1/2}CK^{1/2}$, where C is the covariance function of X.

The next theorem establishes the asymptotic normality of $\hat{\theta}$ with root-n consistency.

Theorem 3.2.3 *Suppose (A1)-(A5) hold, and that the Fisher information $I(\theta_0)$ is nonsingular. Let $\hat{\alpha} = (\hat{\theta}, \hat{\beta})$ be the estimator given by (3.4) with $\lambda = O(n^{-\frac{2r}{2r+1}})$. Then*

$$\sqrt{n}(\hat{\theta} - \theta_0) = n^{-1/2} I^{-1}(\theta_0) \sum_{i=1}^{n} l^*_{\theta_0}(T_i, \Delta_i, W_i) + o_p(1) \xrightarrow{d} \mathcal{N}(0, \Sigma),$$

where $\Sigma = I^{-1}(\theta_0)$.

For the nonparametric coefficient function β, it is of interest to see whether the convergence rate of $\hat{\beta}$ in Theorem 3.2.2 is optimal. In the following, we derive a minimax lower bound for the risk.

Theorem 3.2.4 *Assume that the baseline hazard function $h_0 \in \mathcal{F} = \{h : H(t) = \int_0^t h(s)ds < \infty,$ for any $0 < t < \infty\}$. Suppose that the eigenvalues $\{s_k : k \geq 1\}$ of $K^{1/2} C_\Delta K^{1/2}$ satisfy $s_k \asymp k^{-2r}$ for some constant $0 < r < \infty$. Then,*

$$\lim_{a \to 0} \lim_{n \to \infty} \inf_{\hat{\alpha}} \sup_{\alpha_0 \in \mathbb{R}^p \times \mathcal{H}(K)} \sup_{h_0 \in \mathcal{F}} \mathbb{P}_{\alpha_0, h_0} \left\{ \left\| \hat{\beta} - \beta_0 \right\|_{C_\Delta} \geq a n^{-\frac{r}{2r+1}} \right\} = 1,$$

where the infimum is taken over all possible predictors $\hat{\alpha}$ based on the observed data.

Theorem 3.2.4 shows that the minimax lower bound of the convergence rate for estimating β_0 is $n^{-r/(2r+1)}$, which is determined by r and the decay rate of the eigenvalues of $K^{1/2} C_\Delta K^{1/2}$. We have shown that this rate is achieved by the penalized partial likelihood predictor and therefore this estimator is rate-optimal.

3.3 Computation of the Estimator

3.3.1 Penalized partial likelihood

In this section, we present an algorithm to compute the penalized partial likelihood estimator. Let $\{\xi_1, \ldots \xi_m\}$ be a set of orthonormal basis of the null space with $m = \dim(\mathcal{N}_J)$. The next theorem provides a closed form representation of $\hat{\beta}$ from the penalized partial likelihood method.

Theorem 3.3.1 *The penalized partial likelihood estimator of the coefficient function is given by*

$$\hat{\beta}_\lambda(t) = \sum_{k=1}^{m} d_k \xi_k(t) + \sum_{i=1}^{n} c_i \int_0^1 X_i(s) K_1(s,t) ds, \tag{3.7}$$

where d_k $(k = 1, \ldots m)$ *and* c_i $(i = 1, \ldots n)$ *are constant coefficients.*

Theorem 3.3.1 is a direct application of the generalized version of the well-known representer lemma for smoothing splines (see [44] and [45]). We omit the proof here. In fact, the algorithm can be made more efficient without using all n bases $\int_0^1 X_i(s)K_1(s,t)ds$, $i = 1, \ldots, n$ in (3.7). [3] showed that, under some conditions, a more efficient estimator, denoted by β_λ^*, sharing the same convergence rate with $\hat{\beta}_\lambda$, can be calculated in the data-adaptive finite-dimensional space

$$\mathcal{H}^* = \mathcal{N}_J \oplus \big\{K_1(\tilde{X}_j, \cdot),\, j = 1, \ldots, q\big\},$$

where $\{\tilde{X}_j\}$ is a random subset of $\{X_i : \Delta_i = 1\}$ and

$$K_1(\tilde{X}_j, \cdot) = \int_0^1 \tilde{X}_j(s) K_1(s, \cdot) ds.$$

Here, $q = q_n \asymp n^{2/(ps+1)+\epsilon}$ for some $s > 1$ and $p \in [1, 2]$, and for any $\epsilon > 0$. Therefore, β_λ^* is given by

$$\beta_\lambda^*(t) = \sum_{k=1}^{m} d_k \xi_k(t) + \sum_{j=1}^{q} c_j K_1(\tilde{X}_j, t).$$

The computational efficiency is more prominent when n is large, as the number of coefficients is significantly reduced from $n + m$ to $q + m$.

For the Sobolev space $\mathcal{W}_{2,m}$, the penalty function $J(\cdot)$ is $J(f) = \int_0^1 (f^{(m)}(s))^2 ds$, and (3.4) becomes

$$\begin{aligned}(\hat{\theta}, \hat{\beta}_\lambda) \;=\; & \underset{\theta \in \mathbb{R}^p, \beta \in \mathcal{W}_{2,m}}{\arg\min} \; -\frac{1}{n} \sum_{i=1}^{n} \Delta_i \big\{\eta_\alpha(W_i) - \log \sum_{T_j > T_i} \exp(\eta_\alpha(W_j)\big\} \\ & + \lambda \int_0^1 (\beta^{(m)}(s))^2 ds. \end{aligned} \tag{3.8}$$

Let $\xi_\nu = t^{\nu-1}/(\nu-1)!$, $\nu = 1, \ldots m$, be the orthonormal basis of the null space

$$\mathcal{N}_J = \Big\{\beta \in \mathcal{W}_{2,m},\, \int_0^1 (\beta^{(m)}(s))^2 ds = 0\Big\}.$$

Write $G_m(t,u) = (t-u)_+^{m-1}/(m-1)!$, then the kernels are in forms of

$$K_0(s,t) = \sum_{\nu=1}^{m} \xi_\nu(s)\xi_\nu(t), \quad \text{and} \quad K_1(s,t) = \int_0^1 G_m(s,u)G_m(t,u)du.$$

Hence, the estimator is given by

$$\hat{\beta}_\lambda(t) = \sum_{\nu=1}^{m} d_v\xi_\nu(t) + \sum_{i=1}^{n} c_i \int_0^1 X_i(s)K_1(s,t)ds. \tag{3.9}$$

We may obtain the constants c_i and d_j as well as the estimator $\hat{\theta}$ by maximizing the objective function (3.8) after plugging $\hat{\beta}_\lambda(t)$ back into the objective function.

3.3.2 Choosing the smoothing parameter

The choice of the smoothing parameter λ is always a critical but difficult question. In this section, we borrow ideas from [3] and provide a simple GCV method to choose λ. The key idea is to draw an analogy between the partial likelihood estimation and weighted density estimation, which then allows us to define a criterion analogous to the Kullback-Leibler distance to select the best performing smoothing parameter. Below we provide more details.

Let $i_1, \dots i_N$ be the index for the uncensored data, i.e $\Delta_{i_k} = 1$, for $k = 1, \dots N$ and $N = \sum_1^n \Delta_i$. Define weights $w_{i_k}(\cdot)$ as $w_{i_k}(t) = I\{t \geq T_{i_k}\}$ and

$$f_{\alpha|i_k}(t,w) = \frac{w_{i_k}(t)e^{\eta_\alpha(w)}}{\sum_{k=1}^{N} w_{i_k}(t)e^{\eta_\alpha(w)}}.$$

Following the suggestion in Section 8.5 of [3], we extend the Kullback-Leibler distance for density functions to the partial likelihood as follows,

$$\begin{aligned} KL(\hat{\alpha}_\lambda, \alpha) &= \frac{1}{N}\sum_{k=1}^{N} \mathbb{E}_{f_{\alpha_0|i_k}}\Big\{ \log \frac{f_{\alpha_0|i_k}(T_{i_k}, W_{i_k})}{f_{\hat{\alpha}|i_k}(T_{i_k}, W_{i_k})} \Big\} \\ &= \frac{1}{N}\sum_{k=1}^{N} \mathbb{E}_{f_{\alpha_0|i_k}}\Big\{ \log \frac{e^{\eta_{\alpha_0}(W_{i_k})}}{\sum_{j=1}^{n} w_{i_k}(T_j)e^{\eta_{\alpha_0}(W_j)}} - \log \frac{e^{\eta_{\hat{\alpha}_\lambda}(W_{i_k})}}{\sum_{j=1}^{n} w_{i_k}(T_j)e^{\eta_{\hat{\alpha}_\lambda}(W_j)}} \Big\}. \end{aligned}$$

Dropping off terms not involving $\hat{\alpha}_\lambda$, we have a relative KL distance

$$RKL(\hat{\alpha}_\lambda, \alpha) = -\frac{1}{N}\sum_{k=1}^{N} \mathbb{E}_{f_{\alpha_{0|i_k}}} \eta_{\hat{\alpha}_\lambda}(W) + \frac{1}{N}\sum_{k=1}^{N} \log \sum_{j=1}^{n} w_{i_k}(T_j) e^{\eta_{\hat{\alpha}_\lambda}(W_j)}.$$

The second term is ready to be computed once we have an estimate $\hat{\alpha}_\lambda$, but the first term involves α_0 and needs to be estimated. We approximate the RKL by

$$\widehat{RKL}(\hat{\alpha}_\lambda, \alpha_0) = -\frac{1}{n}\sum_{i=1}^{n} \eta^{[i]}_{\hat{\alpha}_\lambda}(W_i) + \frac{1}{N}\sum_{i=1}^{n} \Delta_i \log \sum_{T_j \geq T_i} \exp\{\eta_{\hat{\alpha}_\lambda}(W_j)\}.$$

Based on this $\widehat{RKL}(\hat{\alpha}_\lambda, \alpha_0)$, a function GCV($\lambda$) can be derived analytically when replacing the penalized partial likelihood function by its quadratic approximation,

$$\begin{aligned} GCV(\lambda) &= -\frac{1}{n}\sum_{i=1}^{n} \eta_{\hat{\alpha}_\lambda}(W_i) + \frac{1}{n(n-1)} \mathrm{tr}[(SH^{-1}S)(\mathrm{diag}\Delta - \Delta\mathbf{1}'/n)] \\ &\quad + \frac{1}{N}\sum_{i=1}^{n} \Delta_i \log \sum_{T_j \geq T_i} \exp\{\eta_{\hat{\alpha}_\lambda}(W_j)\}. \end{aligned}$$

Details of deriving GCV(λ) are given in Section 3.5.5.

3.3.3 Calculating the information bound $I(\theta)$

To calculate the information bound $I(\theta)$, we apply the ACE method [46], the estimator of which is shown to converge to (a^*, g^*). For simplicity, we take Z as a one-dimensional scalar. When Z is a vector, we just need to apply the following procedure to all dimensions of Z separately.

Theorem 3.2.1 shows that

$$I(\theta) = \mathbb{E}\{\Delta[Z - a^*(t) - \eta_{g^*}(X)]^{\otimes 2}\}$$

with $(a^*, g^*) \in L_2 \times \mathcal{H}(K)$ being the unique solution that minimizes

$$\mathbb{E}\Big\{\Delta||Z - a(T) - \eta_g(X)||^2\Big\}.$$

Furthermore, the proof of Theorem 3.2.1 reveals that this is equivalent to the following: (a^*, g^*) is the unique solution to the equations:

$$\mathbb{E}(Z - a^* - \eta_{g^*}|T, \Delta = 1) = 0, \quad a.s.\, P_T^{(u)},$$

$$\mathbb{E}(Z - a^* - \eta_{g^*}|X, \Delta = 1) = 0, \quad a.s.\, P_X^{(u)},$$

where $P_T^{(u)}$ and $P_X^{(u)}$ represent, respectively, the measure space of $(T, \Delta = 1)$ and $(X, \Delta = 1)$.

The idea of ACE is to update a and g alternatively until the objective function $e(a, g) = \mathbb{E}\Delta||Z - a(T) - \eta_g(X)||^2$ stops to decrease. In our case, the procedure is as follows:

(i) Initialize a and g,

(ii) Update a by

$$a(T) = \mathbb{E}(Z - \eta_g|T, \Delta = 1)$$

(iii) Update g such that

$$\eta_g(X) = \mathbb{E}(Z - a|X, \Delta = 1)$$

(iv) Calculate $e(a, g) = \mathbb{E}\Delta||Z - a(T) - \eta_g(X)||^2$ and repeat (ii) and (iii) until $e(a, g)$ fails to decrease.

In practice, we replace $\mathbb{E}\Delta||Z - a(T) - \eta_g(X)||^2$ by the sample mean

$$e(a, g) = \frac{1}{n}\sum_{i=1}^{n} \Delta_i||Z_i - a(T_i) - \eta_g(X_i)||^2.$$

As for a and g, we need to employ some smoothing techniques. For a given $g \in \mathcal{H}(K)$ we calculate

$$\tilde{a}_i = \sum_{T_j = T_i} \Delta_j[Z_j - \eta_g(X_j)] / \sum_{T_j = T_i} \Delta_j,$$

and update $a(t)$ as the local polynomial regression estimator for the data $(T_1, \tilde{a}_1), ..., (T_n, \tilde{a}_n)$. For a given $a \in L_2$ we calculate

$$y_i = Z_i - a(T_i), \text{ for all } \Delta_i = 1,$$

and update g by fitting a functional linear regression

$$y = \int g(s)X(s)ds + \epsilon,$$

based on the data (y_i, X_i) with $\Delta_i = 1$. More details can be find in [45]. When $(a^*,\ g^*)$ is obtained, $I(\theta)$ is estimated by

$$\widehat{I(\theta)} = \frac{1}{n}\sum_{i=1}^{n} \Delta_i [Z_i - a^*(T_i) - \eta_{g^*}(X_i)]^{\otimes 2}.$$

3.4 Numerical Studies

In this session, we first carry out simulations under different settings to study the finite sample performance of the proposed method and to demonstrate practical implications of the theoretical results. In the second part, we apply the proposed method to data that were collected to study the effect of early reproduction history to the longevity of female Mexican fruit flies.

3.4.1 Simulations

We adopt a similar design as that in [45]. The functional covariate X is generated by a set of cosine basis functions, $\phi_1 = 1$ and $\phi_{k+1}(s) = \sqrt{2}\cos(k\pi s)$ for $k \geq 1$, such that

$$X(s) = \sum_{k=1}^{50} \zeta_k U_k \phi_k(s),$$

where the U_k are independently sampled from the uniform distribution on $[-3, 3]$ and $\zeta_k = (-1)^{k+1} k^{-v/2}$ with $v = 1,\ 1.5,\ 2,\ 2.5$. In this case, the covariance function of X is $C(s,t) = \sum_{k=1}^{50} 3k^{-v}\phi_k(s)\phi_k(t)$. The coefficient function β_0 is

$$\beta_0 = \sum_{i=1}^{50} (-1)^k k^{-3/2} \phi_k,$$

which is from a Sobolov space $\mathcal{W}_{2,2}$. The reproducing kernel takes the form:

$$K(s,t) = 1 + st + \int_0^1 (s-u)_+ (t-u)_+ du,$$

and $K_1 = \int_0^1 (s-u)_+(t-u)_+ du$. The null space becomes $\mathcal{N}_J = \text{span}\{1, s\}$. The penalty function as mentioned before is $J(f) = \int (f'')^2$. The vector covariate Z is set

to be univariate with distribution $\mathcal{N}(0,1)$ and corresponding slope $\theta = 1$. The failure time T^u is generated based on the hazard function

$$h(t) = h_0(t) \exp\Big\{\theta' Z + \int_0^1 X(s)\beta_0(s)ds\Big\},$$

where $h_0(t)$ is chosen as a constant or a linear function t. Given X, T^u follows an exponential distribution when h_0 is a constant, and follows a Weibull distribution when $h_0(t) = t$. The censoring time T^c is generated independently, following an exponential distribution with parameter γ which controls the censoring rate. When $h_0(t)$ is constant, $\gamma = 19$ and 3.4 lead to censoring rates around 10% and 30% respectively. Similar censoring rates result from $\gamma = 15$ and 3.9 for the case when $h_0(t) = t$. (T, Δ) is then generated by $T = \min\{T^u, T^c\}$ and $\Delta = I\{T^u \leq T^c\}$.

The criterion to evaluate the performance of the estimators $\hat{\beta}$ is the mean squared error, defined as

$$MSE(\hat{\beta}) = \Big\{\frac{1}{\sum_{i=1}^n \Delta_i} \sum_{i=1}^n \Delta_i \Big(\eta_{\hat{\beta}}(X_i) - \eta_{\beta_0}(X_i)\Big)\Big\}^{1/2},$$

which is an empirical version of $||\hat{\beta} - \beta_0||_{C_\Delta}$. To study the trend as the sample size increases, we vary the sample size n according to $n = 50, 100, 150, 200$ for each value $v = 1$, 1.5, 2, 2.5. For each combination of censoring rate, h_0, v and n, the simulation is repeated 1000 times, and the average mean squared error was obtained for each scenario.

Note that for a fixed γ, $\mathbb{E}(\Delta|X)$ is roughly a constant for different values of v. Therefore $C_\Delta(s,t)$ is approximately proportional to $C(s,t) = \sum_{k=1}^{50} k^{-v}\phi_k(s)\phi_k(t)$. In this case, v controls the decay rate of the eigenvalues of C_Δ and $K^{1/2}C_\Delta K^{1/2}$. It follows from Theorem 3.2.2 that a faster decay rate of the eigenvalues leads to a faster convergence rate. Figure 3.1 displays the average MSE based on 1000 simulations. The simulation results are in agreement with Theorem 3.2.2; it is very clear that when v increases from 1 to 2.5 with the remaining parameters fixed, the average MSEs decrease steadily. The average MSEs also decrease with the sample sizes. Besides, for both the exponential and Weibull distribution, the average MSEs are lower for each

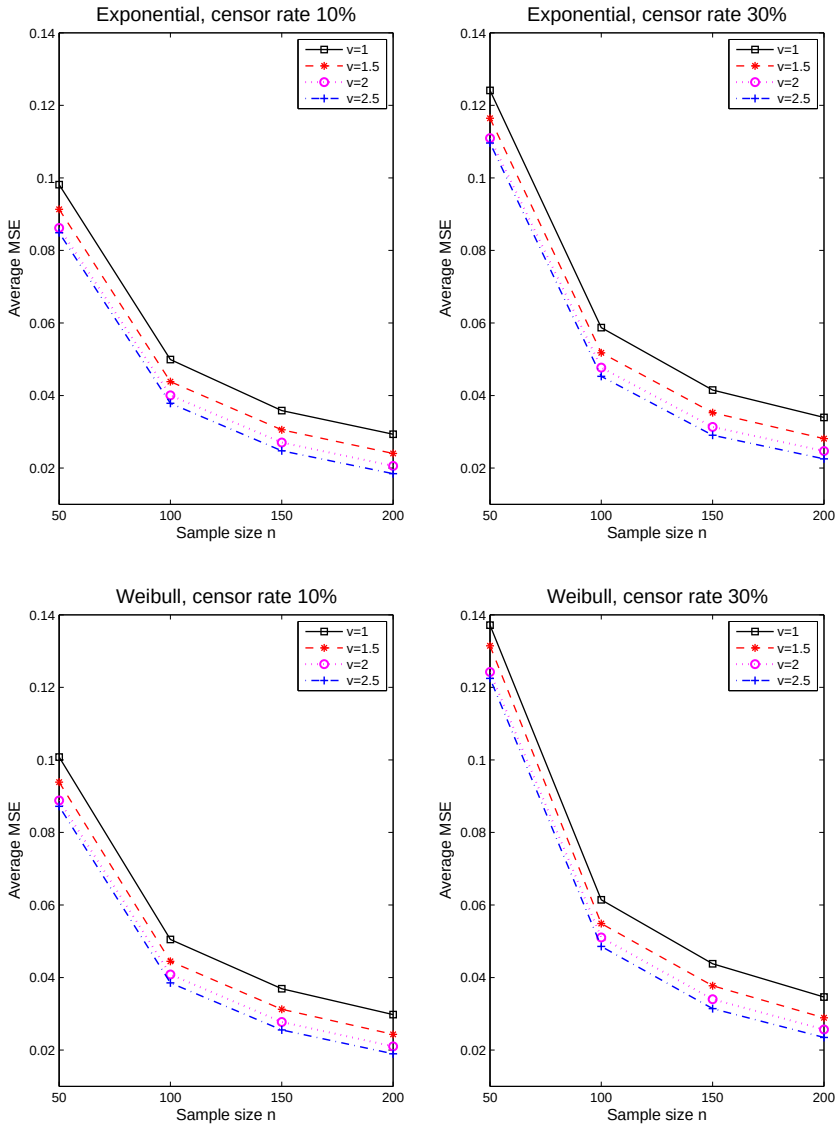


Figure 3.1. The average MSE based on 1000 simulations. The top panel is for the constant baseline hazard function and the bottom panel is for the linear baseline hazard function. For each panel, from left to right, the censoring rate is controlled to be around 10% and 30%. The sample sizes are $n = 50, 100, 150, 200$ and the decay rate parameters are $v = 1, 1.5, 2, 2.5$.

setting at the 10% censoring rate comparing to the values for the 30% censoring rate. This is consistent with the expectation that the lower the censoring rate is, the more accurate the estimate will be.

Averages and standard deviations of the estimated $\hat{\theta}$, for each setting of v and n over 1000 repetition for the case of $h_0 = c$ and 30% censoring rate, are given in Table 3.1. For each case of v, as n increases, the average of $\hat{\theta}$ gets closer to the true value and the standard deviation decreases. Noting that the results do not vary much across different values of v, as v is specially designed to examine the estimation of β and has little effect on the estimation of θ.

For each simulated dataset, we also calculated the information bound $I(\theta)$ based on the ACE method proposed in Section 3.3. The inverse of this information bound, as suggested by Theorem 3.2.3, can be used to estimate the asymptotic variance of $\hat{\theta}$. We further used these asymptotic variance estimates to construct a 95% confidence interval for θ. Table 3.2 shows the observed percentage the constructed 95% confidence interval covered the true value 1 for the various settings. As expected, the covering rates increase towards 95% as n gets larger. Results for other choices of h_0 and censoring rates were about the same and are omitted.

3.4.2 Mexican Fruit Fly Data

We now apply the proposed method to the Mexican fruit fly data in [47]. There were 1152 female flies in that paper coming from four cohorts, for illustration purpose we are using the data from cohort 1 and cohort 2, which consist of the lifetime and daily reproduction (in terms of number of eggs laid daily) of 576 female flies.

We are interested in whether and how early reproduction will affect the lifetime of female Mexican fruit flies. For this reason, we exclude 28 infertile flies from cohort 1 and 20 infertile flies from cohort 2. The period for early reproduction is chosen to be from day 6 to day 30 based on the average reproduction curve (Figure 3.2), which shows that no flies laid any eggs before day 6 and the peak of reproduction was day

Table 3.1.
Average and standard deviation of $\hat{\theta}$. ($h_0 = c$, 30% censoring rate)

n	$v = 1$	$v = 1.5$	$v = 2$	$v = 2.5$
50	1.061	1.064	1.064	1.065
	(0.264)	(0.265)	(0.264)	(0.265)
100	1.027	1.030	1.031	1.031
	(0.164)	(0.164)	(0.164)	(0.163)
150	1.013	1.016	1.017	1.018
	(0.133)	(0.132)	(0.131)	(0.131)
200	1.011	1.013	1.015	1.016
	(0.111)	(0.111)	(0.110)	(0.110)

Table 3.2.
Covering rate of the 95% confidence intervals for θ. ($h_0 = c$, 30% censoring rate)

n	$v = 1$	$v = 1.5$	$v = 2$	$v = 2.5$
50	91.5%	91.9%	92.0%	91.5%
100	93.3%	92.4%	92.4%	93.0%
150	93.5%	93.1%	93.9%	93.4%
200	93.6%	93.7%	93.9%	93.8%

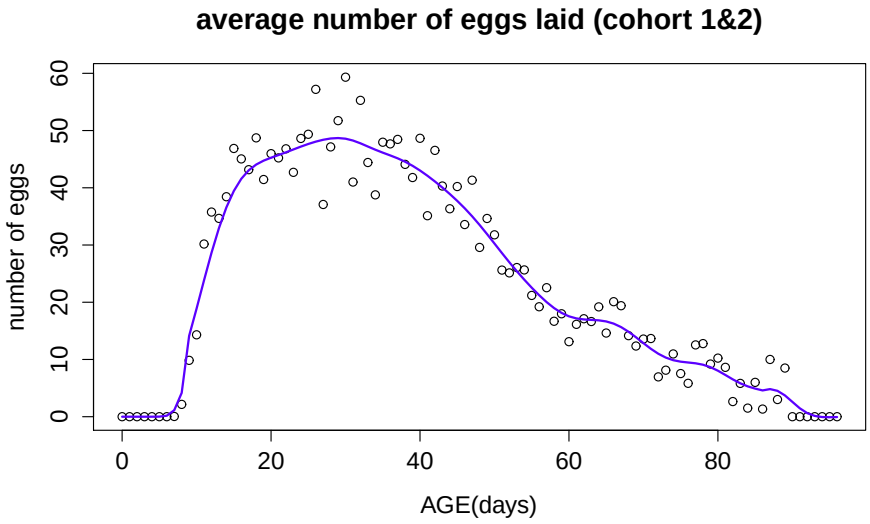


Figure 3.2. Average number of eggs laid daily for both cohorts

30. Once the period of early reproduction was determined to be $[6, 30]$, we further excluded flies that died before day 30 to guarantee a fully observed trajectory for all flies and this leaves us with a total of 479 flies for further exploration of the functional Cox model. The mean and median lifetime of the remaining 224 flies in cohort 1 is 56.41 and 58 days respectively; the mean and the median lifetime of the remaining 255 flies in cohort 2 is 55.78 and 55 days respectively.

The trajectories of early reproduction for these 479 flies are of interest to researchers but they are very noisy, so for visualization we display the smoothed egg-laying curves for the first 100 flies (Figure 3.3). The data of these 100 flies were individually smoothed with a local linear smoother, but the subsequent data analysis for all 479 flies was based on the original data without smoothing.

Using the original egg-laying curves from day 6 to day 30 as the longitudinal covariates and the cohort indicator as a time-independent covariate, the functional Cox model resulted in an estimate $\hat{\theta} = 0.0562$ with 95% confidence interval $[-0.1235, 0.2359]$. Since zero is included in the interval, we conclude that the cohort effect is not significant. Figure 3.4 shows the estimated coefficient function $\hat{\beta}$ for the longitudinal

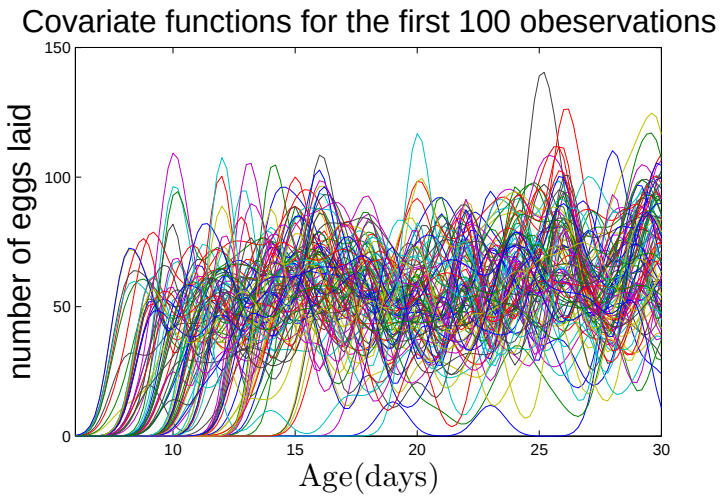


Figure 3.3. Pre-smoothed individual curves for the first 100 observations.

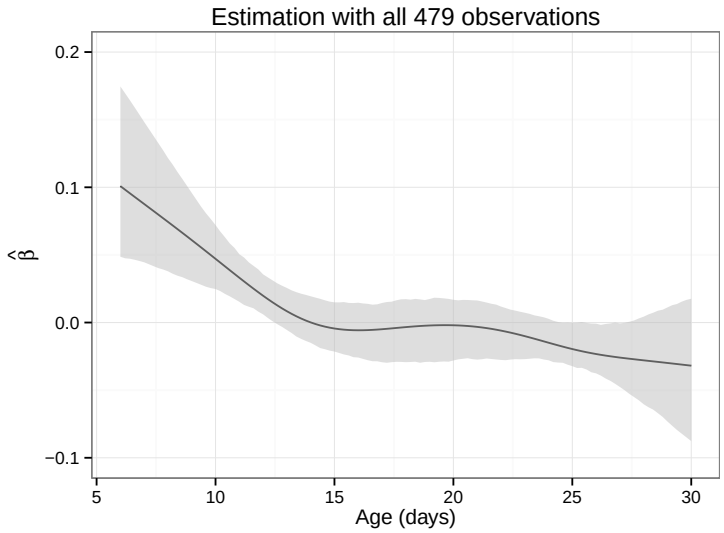


Figure 3.4. Estimated coefficient function $\hat{\beta}(s)$ using all 479 observations and 95% pointwise c.i. for $\beta(s)$.

covariate. The shaded area is the 95% pointwise bootstrap confidence interval. Under the functional Cox model, a positive $\hat{\beta}(s)$ yields a larger hazard function and a decreased probability of survival and vice versa for a negative $\hat{\beta}(s)$.

Checking the plot of $\hat{\beta}(s)$, we can see that $\hat{\beta}(s)$ starts with a large positive value, but decreases fast to near zero on day 13 and stays around zero till day 22, then declines again mildly towards day 30. The pattern of $\hat{\beta}(s)$ indicates that higher early reproduction before day 13 results in a much higher mortality rate suggesting the high cost of early reproduction, whereas a higher reproduction that occurs after day 22 tends to lead to a relatively lower mortality rate, suggesting that reproduction past day 22 might be sign of physical fitness. However, the latter effect is less significant than the early reproduction effect as indicated by the bootstrap confidence interval. Reproduction between day 13 and day 22 does not have a major effect on the mortality rate. In other words, flies that lay a lot of eggs in their early age (before day 13) and relatively fewer eggs after day 22 tend to die earlier, while those with the opposite pattern tend to have a longer life span.

Table 3.3.
Values of fixed cut-off point and parameters for generating random cut-off point, followed by the actual censored percentage for both cohorts and the whole data.

	fixed cut-off point		random cut-off point	
	$T^c = 71$ (10%)	$T^c = 62$ (30%)	$T^c \sim \exp(450)$ (10%)	$T^c \sim \exp(150)$ (30%)
Cohort 1	0.138	0.339	0.0.071	0.353
Cohort 2	0.067	0.259	0.110	0.251
Total	0.100	0.296	0.092	0.300

The Mexfly data contains no censoring, so it is easy to check how the proposed method works in the presence of censored data. We artificially randomly censor the data by 10% and then again by 30% using an exponential censoring distribution with parameter $\gamma = 450$ and 150, respectively. The estimated coefficient $\hat{\theta}$ and corresponding 95% confidence intervals are given in Table 3.4. Regardless of the censoring conditions, all the confidence intervals contain zero and therefore indicate a non-significant cohort effect. This is consistent with the previous result for non-censored data. The estimated coefficient functions $\hat{\beta}$ and the corresponding pointwise bootstrap confidence intervals are displayed in Figure 3.5. Despite the slightly different results for different censoring proportions and choice of tuning parameters, all the $\hat{\beta}$ have a similar pattern. This indicates that the proposed method is quite stable with respect to right censorship, as long as the censoring rate is below 30%.

3.5 Technical Proofs

We first introduce some notations by denoting $d(\beta_1, \beta_2) = ||\beta_1 - \beta_2||_{C_\Delta}$, for any $\beta_1, \beta_2 \in \mathcal{H}(K)$; $Y(t) = 1_{\{T \geq t\}}$; $Y_j(t) = 1_{\{T_j \geq t\}}$, $1 \leq j \leq n$; and $\eta_\beta(X_i) = \int_0^1 \beta(s) X_i(s) ds$.

Table 3.4.
The estimated $\hat{\theta}$ and 95% confidence interval for θ under different censoring conditions.

	10% censoring	30% censoring
fixed cut-off point	0.0929 [-0.0914, 0.2772]	0.0757 [-0.1268 0.2870]
random cut-off point	0.0104 [-0.1705, 0.1913]	0.1863 [-0.0177,0.3903]

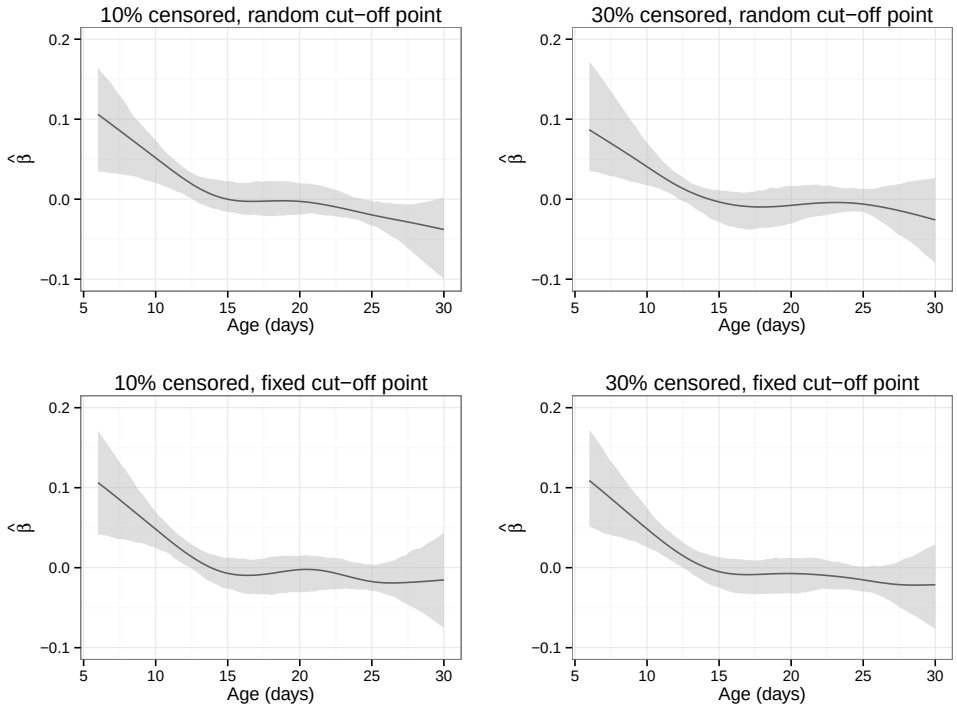


Figure 3.5. Estimation for $\beta(s)$ with censored data and 95% pointwise c.i.

Recall that $W = (Z, X)$ represents the covariates, $\alpha = (\theta, \beta)$ represents the corresponding regression coefficient with θ the coefficient for Z and β the coefficient function for $X(\cdot)$, and the true coefficient is denoted as $\alpha_0 = (\theta_0, \beta_0)$. The index $\eta_\alpha(W) = \theta' Z + \int_0^1 \beta(s)X(s)ds$ summarizes the information carried by the covariate W. To measure the distance between two coefficients α_1 and α_2 we use

$$d(\alpha_1, \alpha_2)^2 = \mathbb{E}(\Delta[\eta_{\alpha_1}(W) - \eta_{\alpha_2}(W)]^2).$$

Furthermore, we denote

$$S_{0n}(t, \alpha) = \frac{1}{n}\sum_{j=1}^{n} Y_j(t)e^{\eta_\alpha(W_j)}, \qquad S_0(t, \alpha) = \mathbb{E}\{Y(t)e^{\eta_\alpha(W)}\},$$

and for $\widetilde{\alpha} \in L_2 \times \mathcal{H}(K)$,

$$S_{1n}(t, \alpha)[\widetilde{\alpha}] = \frac{1}{n}\sum_{j=1}^{n} Y_j(t)e^{\eta_\alpha(W_j)}\eta_{\widetilde{\alpha}}(W_j)), \qquad S_1(t, \alpha)[\widetilde{\alpha}] = \mathbb{E}[Y(t)e^{\eta_\alpha(W)}\eta_{\widetilde{\alpha}}(W)].$$

Define

$$m_n(t, W, \alpha) = [\eta_\alpha(W) - \log S_{0n}(t, \alpha)]1_{\{0 \le t \le \tau\}},$$

and

$$m_0(t, W, \alpha) = [\eta_\alpha(W) - \log S_0(t, \alpha)]1_{\{0 \le t \le \tau\}}.$$

Let P_n and P be the empirical and probability measure of (T_i, Δ_i, W_i) and (T, Δ, W), respectively, and $P_{\Delta n}$ and P_Δ be the subprobability measure with $\Delta_i = 1$ and $\Delta = 1$ accordingly. The logarithm of the partial likelihood is $M_n(\alpha) = P_{\Delta n} m_n(\cdot, \alpha)$. Let $M_0(\alpha) = P_\Delta m_0(\cdot, \alpha)$. Note that P_Δ is restricted to $T \in [0, \tau]$ due to the $1\{0 \le t \le \tau\}$ term.

A useful identity due to Lemma 2 in [31] is

$$\frac{S_1(t, \alpha)[\widetilde{\alpha}]}{S_0(t, \alpha)} = \mathbb{E}[\eta_{\widetilde{\alpha}}(W) | \, T = t, \Delta = 1]. \tag{3.10}$$

3.5.1 Proof of Theorem 3.2.1

The log-likelihood for a single sample $(t, \Delta, Z, X(\cdot))$ is

$$l(h_0, \theta, \beta) = \Delta[\log h_0(t) + Z'\theta + \int_0^1 X(s)\beta(s)ds] - H_0(t)\exp[Z'\theta + \int_0^1 X(s)\beta(s)ds],$$

where $H_0(t) = \int_0^t h_0(u)du$ is the baseline cumulative hazard function. Consider a parametric and smooth sub-model $\{h_{(\mu_1)} : \mu_1 \in \mathbb{R}\}$ satisfying $h_{(0)} = h_0$ and

$$\frac{\partial \log h_{(\mu_1)}}{\partial \mu_1}(t)\Big|_{\mu_1=0} = a(t).$$

Let $\eta_{(\mu_2)}(X) = \eta_\beta(X) + \eta_{\mu_2 g}(X)$, for $g \in \mathcal{H}(K)$. Therefore $\eta_{(0)} = \eta_\beta(X)$ and

$$\frac{\partial \eta(\mu_2)}{\partial \mu_2}(X)\Big|_{\mu_2=0} = \eta_g(X).$$

Recall that $r(W) = \exp(\eta_\alpha(W))$, and $M(t)$ is the counting process martingale associated with model (1),

$$M(t) = M(t|W) = \Delta I\{T \le t\} - \int_0^t I\{T \ge u\} r(W) dH_0(u).$$

The score operators for the cumulative hazard H_0, coefficient function β, and the score vector for θ are the partial derivatives of the likelihood $l(h_{(\mu_1)}, \theta, \eta_{(\mu_2)})$ with respect to μ_1, μ_2 and θ evaluated at $\mu_1 = \mu_2 = 0$,

$$i_H a := \Delta a(T) - r(W) \int_0^\infty Y(t) a(t) dH_0(t) = \int_0^\infty a(t) dM(t),$$

$$i_\beta g := \eta_g(X)[\Delta - r(W) H_0(T)] = \int_0^\infty \eta_g(X)\, dM(t),$$

$$i_\theta := Z[\Delta - r(W) H_0(T)] = \int_0^\infty Z\, dM(t).$$

Define $L(P_T^{(u)}) := \{a \in \mathcal{L}_2 : \mathbb{E}[\Delta a^2(T)] < \infty\}$ and $L(P_X^{(u)}) := \{g \in \mathcal{H}(K) : \mathbb{E}[\Delta \eta_g(X)] = 0; \mathbb{E}[\Delta \eta_g^2(X) < \infty]\}$. Let

$$A_H = \{i_H a : a \in L(P_T^{(u)})\},$$

and

$$G = \{i_\beta g : g \in L(P_X^{(u)})\}.$$

To calculate the information bound for θ, we need to find the (least favorable) direction (a^*, g^*) such that $i_\theta - i_H a^* - i_\beta g^*$ is orthogonal to the sum space $\boldsymbol{A} = A_H + G$. That is, (a^*, g^*) must satisfy

$$\mathbb{E}[(i_\theta - i_H a^* - i_\beta g^*) i_H a] = 0, \quad a \in L(P_T^{(u)}),$$

$$\mathbb{E}[(i_\theta - i_H a^* - i_\beta g^*) i_\beta g] = 0, \quad g \in L(P_X^{(u)}).$$

Following the proof of Theorem 3.1 in [34], we can show that (a^*, g^*) satisfies

$$\mathbb{E}[\Delta(Z - a^* - \eta_{g^*})a] = 0, \quad a \in L(P_T^{(u)}), \tag{3.11}$$

$$\mathbb{E}[\Delta(Z - a^* - \eta_{g^*})\eta_g] = 0, \quad g \in L(P_X^{(u)}). \tag{3.12}$$

Therefore, (a^*, g^*) is the solution to the following equations:

$$\mathbb{E}(Z - a^* - \eta_{g^*} | T, \Delta = 1) = 0, \quad a.s.\, P_T^{(u)},$$

$$\mathbb{E}(Z - a^* - \eta_{g^*} | X, \Delta = 1) = 0, \quad a.s.\, P_X^{(u)}.$$

So, $(a^*, g^*) \in L(P_T^{(u)}) \times L(P_X^{(u)})$ minimizes

$$\mathbb{E}\Big\{\Delta \big\| Z - a(T) - \eta_g(X) \big\|^2\Big\}. \tag{3.13}$$

It follows from Conditions A3 and A4 that the space $L(P_T^{(u)}) \times L(P_X^{(u)})$ is closed, so that the minimizer of (3.13) is well-defined. Further, the solution can be obtained by the population version of the ACE algorithm of [46]. ■

3.5.2 Proof of Theorem 3.2.2

For some large number M, such that $||\theta_0||_\infty < M$ and $||\beta_0||_K < M$, define $\mathbb{R}_M = \{\theta \in \mathbb{R}^p, ||\theta||_\infty < M\}$ and $\mathcal{H}^M = \{\beta \in \mathcal{H}(K), ||\beta||_K < M\}$. Let $\alpha^M = (\theta^M, \beta^M)$ be the penalized partial likelihood estimator with minimum taken over $L^M \times \mathcal{H}^M$, i.e.

$$\alpha^M = \arg \min_{\alpha \in \mathbb{R}_M \times \mathcal{H}^M} -n^{-1} \sum_{i=1}^{n} \Delta_i \big\{ \eta_\alpha(W_i) - \log \sum_{T_j > T_i} \exp\{\eta_\alpha(W_j)\} \big\} + \lambda \cdot J(\beta). \tag{3.14}$$

We first prove that

$$\sup_{\alpha \in \mathbb{R}_M \times \mathcal{H}^M} |M_n(\alpha) - M_0(\alpha)| \xrightarrow{P} 0. \tag{3.15}$$

Observe that

$$
\begin{aligned}
&|M_n(\alpha) - M_0(\alpha)| \\
&\leq |P_{\Delta n} m_n(\cdot, \alpha) - P_{\Delta n} m_0(\cdot, \alpha)| + |P_{\Delta n} m_0(\cdot, \alpha) - P_{\Delta} m_0(\cdot, \alpha)| \\
&\leq P_{\Delta n} |\log S_{0n}(T, \alpha) - \log S_0(T, \alpha)| 1_{\{0 \leq T \leq \tau\}} + |(P_n - P)\Delta m_0(\cdot, \alpha)| \\
&\lesssim \sup_{0 \leq t \leq \tau} |S_{0n}(t, \alpha) - S_0(t, \alpha)| + |(P_n - P)\Delta m_0(\cdot, \alpha)| \\
&= \sup_{0 \leq t \leq \tau} |(P_n - P) Y(t) e^{\eta_\alpha(W)}| + |(P_n - P)\Delta m_0(\cdot, \alpha)|.
\end{aligned}
$$

Lemma 3.3 shows that $\mathcal{F}_1 = \{\Delta m_0(t, W, \alpha) : \alpha \in \mathbb{R}_M \times \mathcal{H}^M\}$ and $\mathcal{F}_2 = \{Y(t) e^{\eta_\alpha(W)} : \alpha \in \mathbb{R}_M \times \mathcal{H}^M,\ 0 \leq t \leq \tau\}$ are P-Glivenko-Cantelli, which means that both terms on the righthand side above converge to zero in probability uniformly with respect to $\alpha \in \mathbb{R}_M \times \mathcal{H}^M$. Therefore (3.15) holds.

The definition of α^M in (3.14) indicates that

$$
-M_n(\alpha^M) + \lambda J(\beta^M) \leq -M_n(\alpha_0) + \lambda J(\beta_0).
$$

Rearranging the inequality with $M_n(\alpha^M)$ on one side and the fact that $\lambda \to 0$ as $n \to \infty$ lead to

$$
M_n(\alpha^M) \geq M_n(\alpha_0) - o_p(1). \tag{3.16}
$$

On the other hand, lemma 3.2 implies that $\sup_{d(\alpha, \alpha_0) \geq \epsilon} M_0(\alpha) < M_0(\alpha_0)$. Combining this with (3.15) and (3.16) and by the consistency result in [48, Theorem 5.7 on Page 45], we can show that α^M is consistent, i.e. $d(\alpha^M, \alpha_0) \xrightarrow{P} 0$.

Part (i) now follows from

$$
d(\hat{\alpha}, \alpha_0) \leq d(\hat{\alpha}, \alpha^M) + d(\alpha^M, \alpha_0),
$$

and $P(\hat{\alpha} = \alpha^M) = P(||\hat{\beta}||_K < M, ||\hat{\theta}||_\infty < M) \to 1$, as $M \to \infty$, i.e. $d(\hat{\alpha}, \alpha^M) \to 0\, a.s.$.

For part (ii), we follow the proof of Theorem 3.4.1 in [49]. We first show that

$$
E^* \sup_{\delta/2 \leq d(\alpha, \alpha_0)) \leq \delta} \sqrt{n} |(M_n - M_0)(\alpha - \alpha_0)| \lesssim \phi_n(\delta), \tag{3.17}
$$

where $\phi_n(\delta) = \delta^{\frac{2r-1}{2r}}$. Direct calculation yields that

$$
\begin{aligned}
&(M_n - M_0)(\alpha - \alpha_0) \\
&= P_{\Delta n} m_n(\cdot, \alpha) - P_{\Delta n} m_n(\cdot, \alpha_0) - P_\Delta m_0(\cdot, \alpha) + P_\Delta m_0(\cdot, \alpha_0) \\
&= (P_{\Delta n} - P_\Delta)(m_0(\cdot, \alpha) - m_0(\cdot, \alpha_0)) \\
&\quad + P_{\Delta n}(m_n(\cdot, \alpha) - m_n(\cdot, \alpha_0) - m_0(\cdot, \alpha) + m_0(\cdot, \alpha_0)) \\
&= (P_{\Delta n} - P_\Delta)(m_0(\cdot, \alpha) - m_0(\cdot, \alpha_0)) \\
&\qquad + P_{\Delta n}(\log \frac{S_0(T, \alpha)}{S_0(T, \alpha_0)} - \log \frac{S_{0n}(T, \alpha)}{S_{0n}(T, \alpha_0)}) \\
&= I + II.
\end{aligned}
$$

For the first term, $I = (P_{\Delta n} - P_\Delta)(m_0(\cdot, \alpha) - m_0(\cdot, \alpha_0))$. By Lemma 3.4, we have

$$\sup_{\delta/2 \le d(\alpha, \alpha_0)) \le \delta} |I| = O(\delta^{\frac{2r-1}{2r}} n^{-1/2}).$$

For the second term II, we have

$$
\begin{aligned}
&\sup_{\delta/2 \le d(\alpha, \alpha_0)) \le \delta} |II| \\
\le & \sup_{\substack{\delta/2 \le d(\alpha, \alpha_0)) \le \delta \\ t \in [0, \tau]}} \Big| \log \frac{S_0(t, \alpha)}{S_0(t, \alpha_0)} - \log \frac{S_{0n}(t, \alpha)}{S_{0n}(t, \alpha_0)} \Big| \\
\le & \sup_{\substack{\delta/2 \le d(\alpha, \alpha_0)) \le \delta \\ t \in [0, \tau]}} c \Big| \frac{S_{0n}(t, \alpha)}{S_{0n}(t, \alpha_0)} - \frac{S_0(t, \alpha)}{S_0(t, \alpha_0)} \Big| \\
= & \sup_{\substack{\delta/2 \le d(\alpha, \alpha_0)) \le \delta \\ t \in [0, \tau]}} c \Big| \frac{S_{0n}(t, \alpha) S_0(t, \alpha_0) - S_{0n}(t, \alpha_0) S_0(t, \alpha)}{S_0(t, \alpha_0) S_{0n}(t, \alpha_0)} \Big|.
\end{aligned}
$$

For $t \in [0, \tau]$, the denominator $S_0(t, \alpha_0)S_{0n}(t, \alpha_0)$ is bounded away from zero with probability tending to one. The numerator satisifes

$$\begin{aligned}
&S_{0n}(t, \alpha)S_0(t, \alpha_0) - S_{0n}(t, \alpha_0)S_0(t, \alpha) \\
&= S_0(t, \alpha_0)[S_{0n}(t, \alpha) - S_{0n}(t, \alpha_0) - S_0(t, \alpha) + S_0(t, \alpha_0)] \\
&\quad -[S_{0n}(t, \alpha_0) - S_0(t, \alpha_0)][S_0(t, \alpha) - S_0(t, \alpha_0)].
\end{aligned}$$

For the first term on the right side, we have $S_0(t, \alpha_0) = O(1)$ and

$$\begin{aligned}
&[S_{0n}(t, \alpha) - S_{0n}(t, \alpha_0) - S_0(t, \alpha) + S_0(t, \alpha_0)] \\
&\quad = (P_n - P)\big\{Y(t)\big[\exp(\eta_\alpha(W)) - \exp(\eta_{\alpha_0}(W))\big]\big\}.
\end{aligned}$$

Define the above $(P_n - P)\big\{Y(t)\big[\exp(\eta_\alpha(W)) - \exp(\eta_{\alpha_0}(W))\big]\big\} \stackrel{def}{=} III$.

Lemma 3.4 implies that

$$\sup_{\delta/2 \le d(\alpha, \alpha_0)) \le \delta} |III| = O(\delta^{\frac{2r-1}{2r}} n^{-1/2}).$$

For the second term, the Central Limit Theorem implies $S_{0n}(t, \alpha_0) - S_0(t, \alpha_0) = O_p(n^{-1/2})$, and

$$\begin{aligned}
|S_0(t, \alpha) - S_0(t, \alpha_0)| &\le E\big\{Y(t)\big|\exp(\eta_\alpha(W)) - \exp(\eta_{\alpha_0}(W))\big|\big\} \\
&\lesssim \Big(E[\eta_\alpha(W) - \eta_{\alpha_0}(W)]^2\Big)^{1/2} \\
&\lesssim d(\alpha, \alpha_0).
\end{aligned}$$

Therefore

$$\sup_{\delta/2 \le d(\alpha, \alpha_0)) \le \delta} |II| \le O(\delta^{\frac{2r-1}{2r}} n^{-1/2}) + O(\delta n^{-1/2}) = O(\delta^{\frac{2r-1}{2r}} n^{-1/2}).$$

Combining I and II yields

$$E^* \sup_{\delta/2 \le d(\alpha, \alpha_0)) \le \delta} \sqrt{n}|(M_n - M_0)(\alpha - \alpha_0)| \lesssim O(\delta^{\frac{2r-1}{2r}}).$$

Furthermore, Lemma 3.2 implies

$$\sup_{\delta/2 \le d(\alpha, \alpha_0)) \le \delta} P_\Delta m_0(\cdot, \alpha) - P_\Delta m_0(\cdot, \alpha_0) \lesssim -\delta^2.$$

Let $r_n = n^{\frac{r}{2r+1}}$. It is easy to check that r_n satisfies $r_n^2 \phi_n(\frac{1}{r_n}) \leq \sqrt{n}$, and

$$M_n(\hat{\alpha}_\lambda) \geq M_n(\alpha_0) + \lambda[J(\hat{\beta}_\lambda) - J(\beta_0)] \geq M_n(\alpha_0) - O_p(r_n^{-2})$$

with $\lambda = O(r_n^{-2}) = O(n^{-\frac{2r}{2r+1}})$.

So far we have verified all the conditions in Theorem 3.4.1 of [49] and thus conclude that

$$d(\hat{\alpha}, \alpha_0) = O_p(r_n^{-1}) = O_p(n^{-\frac{r}{2r+1}}).$$

For part (iii), recall the projections a^* and g^* defined in Theorem 3.2.1, then

$$\begin{aligned}
d(\hat{\alpha}, \alpha_0)^2 &= \mathbb{E}\Delta[\eta_{\hat{\alpha}}(W) - \eta_{\alpha_0}(W)]^2 \\
&= \mathbb{E}\Delta[Z^{'}(\hat{\theta} - \theta_0) + (\eta_{\hat{\beta}}(X) - \eta_{\beta_0}(X))]^2 \\
&= \mathbb{E}\Delta[(Z - a^*(T) - \eta_{g^*}(X))^{'}(\hat{\theta} - \theta_0) + (a^*(T) + \eta_{g^*}(X))(\hat{\theta} - \theta_0) \\
&\qquad + (\eta_{\hat{\beta}}(X) - \eta_{\beta_0}(X))]^2 \\
&= \mathbb{E}\Delta[(Z - a^*(T) - \eta_{g^*}(X))^{'}(\hat{\theta} - \theta_0)]^2 \\
&\qquad + \mathbb{E}\Delta[(a^*(T) + \eta_{g^*}(X))(\hat{\theta} - \theta_0) + (\eta_{\hat{\beta}}(X) - \eta_{\beta_0}(X))]^2. \qquad (3.18)
\end{aligned}$$

Since $I(\theta)$ is non-singular, it follows that$||\hat{\theta} - \theta_0||^2 = O_p(n^{-\frac{2r}{2r+1}})$. This in turn implies

$$d(\hat{\beta}, \beta_0)^2 = O_p(n^{-\frac{2r}{2r+1}}).$$

■

3.5.3 Proof of Theorem 3.2.3

Let $u = (t, Z, X(\cdot))$. For $g \in \mathcal{H}(K)$, define

$$s_n(u, \alpha)[g] = \eta_g(X) - \frac{S_{1n}(t, \alpha)[g]}{S_{0n}(t, \alpha)}, \qquad s(u, \alpha)[g] = \eta_g(X) - \frac{S_1(t, \alpha)[g]}{S_0(t, \alpha)},$$

and for $Z \in \mathbb{R}^d$ and the identify map $I(Z) = Z$, define

$$s_n(u, \alpha)[Z] = Z - \frac{S_{1n}(t, \alpha)[I]}{S_{0n}(t, \alpha)}, \qquad s(u, \alpha)[Z] = \eta_g(X) - \frac{S_1(t, \alpha)[I]}{S_0(t, \alpha)},$$

where $S_{1n}(t, \alpha)[I] = \frac{1}{n}\sum_{j=1}^n Y_j(t) e^{\eta_\alpha(W_j)} Z_j$ and $S_1(t, \alpha)[I] = \mathbb{E}Y(t) e^{\eta_\alpha(W)} Z$.

By analogy to the score function, we call the derivatives of the partial likelihood with respect to the parameters the partial score functions. The partial score function based on the partial likelihood for θ is

$$i_{n\theta}(\alpha) = P_{\Delta n} s_n(\cdot, \alpha)[Z].$$

The partial score function based on the partial likelihood for β in a direction $g \in \mathcal{H}(K)$ is

$$i_{n\beta}(\alpha)[g] = P_{\Delta n} s_n(\cdot, \alpha)[g].$$

Recall that $(\hat{\theta}, \hat{\beta})$ is defined to maximize the penalized partial likelihood, i.e.

$$-P_{\Delta n} m_n(\cdot, \hat{\theta}, \hat{\beta}) + \lambda J(\hat{\beta}) \leq -P_{\Delta n} m_n(\cdot, \theta, \beta) + \lambda J(\beta),$$

for all $\theta \in \mathbb{R}^p$ and $\beta \in \mathcal{H}(K)$. Since the penalty term is unrelated to θ, the partial score function should satisfy

$$i_{n\theta}(\hat{\alpha}) = P_{\Delta n} s_n(\cdot, \hat{\alpha})[Z] = 0.$$

On the other hand, the partial score function for β satisfies

$$i_{n\beta}(\hat{\alpha})[g] = P_{\Delta n} s_n(\cdot, \alpha)[g] = O(\lambda) = o_p(n^{-\frac{1}{2}}), \quad \text{for all } g \in \mathcal{H}(K).$$

Combining this with Lemma 3.5 and Lemma 3.6, we have

$$n^{1/2} P_\Delta \{s(\cdot, g_0)[Z - h^*]\}^{\otimes 2}(\hat{\theta} - \theta_0) = -n^{1/2} P_{\Delta n} s_n(\cdot, \alpha_0)[Z - g^*] + o_p(1).$$

Let

$$M_i(t) = \Delta_i I\{T_i \leq t\} - \int_0^t Y_i(u) \exp(\eta_{\alpha_0}(W_i)) dH_0(u), \quad 1 \leq i \leq n.$$

We can write

$$n^{1/2} P_{\Delta n} s_n(\cdot, \alpha_0)[Z - g^*] = n^{-1/2} \sum_{i=1}^n \int_0^\tau [Z_i - \eta_{h^*}(X_i) - \frac{S_{1n}(t, \alpha_0)[Z - g^*]}{S_{0n}(t, \alpha_0)}] dM_i(t).$$

Thus

$$\begin{aligned} & n^{1/2} P_{\Delta n} s_n(\cdot, \alpha_0)[Z - g^*] - n^{-1/2} \sum_{i=1}^n \int_0^\tau [Z_i - \eta_{h^*}(X_i) - \frac{S_1(t, \alpha_0)[Z - g^*]}{S_0(t, \alpha_0)}] dM_i(t) \\ & = n^{-1/2} \sum_{i=1}^n \int_0^\tau [\frac{S_1(t, \alpha_0)[Z - g^*]}{S_0(t, \alpha_0)} - \frac{S_{1n}(t, \alpha_0)[Z - g^*]}{S_{0n}(t, \alpha_0)}] dM_i(t). \end{aligned}$$

Because

$$n^{-1}\sum_{i=1}^{n}\int_0^\tau [\frac{S_1(t,\alpha_0)[Z-g^*]}{S_0(t,\alpha_0)} - \frac{S_{1n}(t,\alpha_0)[Z-g^*]}{S_{0n}(t,\alpha_0)}]Y_i(t)\exp[\eta_{\alpha_0}(W_i)]dH_i(t) \xrightarrow{P} 0,$$

by Lenglart's inequality, as stated in Theorem 3.4.1 and Corollary 3.4.1 of [50], we have

$$\begin{aligned} & n^{1/2}P_{\Delta n}s_n(\cdot,\alpha_0)[Z-g^*] \\ & \quad = n^{-1/2}\sum_{i=1}^{n}\int_0^\tau [Z_i - \eta_{h^*}(X_i) - \frac{S_1(t,\alpha_0)[Z-g^*]}{S_0(t,\alpha_0)}]dM_i(t) + o_p(1). \end{aligned}$$

Recall that

$$\frac{S_1(t,\alpha_0)[Z-g^*]}{S_0(t,\alpha_0)} = E[Z-\eta_{g^*}(W)]|T=t,\Delta=1] = a^*(t).$$

By the definition of the efficient score function l_θ^*, we have

$$n^{1/2}P_{\Delta n}s_n(\cdot,\alpha_0)[Z-g^*] = n^{-1/2}\sum_{i=1}^{n} l_\theta^*(T_i,\Delta_i,W_i) + o_p(1) \to \mathcal{N}(\overset{d}{0}, I(\theta_0)) \ .$$

■

3.5.4 Proof of Theorem 3.2.4

To get the minmax lower bound, it suffices to show that, when the true baseline hazard function h_0 and the true θ_0 are fixed and known, for a subset $\mathcal{H}^*$ of $\mathcal{H}(K)$,

$$\lim_{a\to 0}\lim_{n\to\infty}\inf_{\hat\beta}\sup_{\beta_0\in\mathcal{H}^*}\mathbb{P}_{h_0,\theta_0,\beta_0}\{d(\hat\beta,\beta_0)\geq an^{-\frac{r}{2r+1}}\} = 1. \tag{3.19}$$

If we can find a subset $\{\beta^{(0)},\ldots,\beta^{(N)}\}\subset\mathcal{H}^*$ with N increasing with n, such that for some positive constant c and all $0\leq i<j\leq N$,

$$d^2(\beta^{(i)},\beta^{(j)}) \geq c\gamma^{\frac{2r}{2r+1}}n^{-\frac{2r}{2r+1}}, \tag{3.20}$$

and

$$\frac{1}{N}\sum_{j=1}^{N} KL(P_j,P_0) \quad \leq \quad \gamma\log N, \tag{3.21}$$

then we can conclude, according to [51] (Theorem 2.5 on page 99), that

$$\inf_{\hat{\beta}} \sup_{\beta \in \mathcal{H}^*} \mathbb{P}(d^2(\beta^{(i)}, \beta^{(j)}) \geq c\gamma^{\frac{2r}{2r+1}} n^{-\frac{2r}{2r+1}}) \geq \frac{\sqrt{N}}{1+\sqrt{N}}(1 - 2\gamma - \sqrt{\frac{2\gamma}{\log N}}),$$

which yields

$$\lim_{a \to 0} \lim_{n \to \infty} \inf_{\hat{\beta}} \sup_{\beta_0 \in \mathcal{H}^*} \mathbb{P}(d(\beta^{(i)}, \beta^{(j)}) \geq a n^{-\frac{r}{2r+1}}) \geq 1.$$

Hence Theorem 3.2.4 will be proved.

Next, we are going to construct the set $\mathcal{H}^*$ and the subset $\{\beta^{(0)}, \ldots, \beta^{(N)}\} \subset \mathcal{H}^*$, and then show that both (3.20) and (3.21) are satisfied.

Consider the function space

$$\mathcal{H}^* = \{\beta = \sum_{k=M+1}^{2M} b_k M^{-1/2} L_{K^{1/2}} \varphi_k : (b_{M+1}, \ldots b_{2M}) \in \{0,1\}^M\}, \tag{3.22}$$

where $\{\varphi_k : k \geq 1\}$ are the orthonomal eigenfunctions of $T(s,t) = K^{1/2} C_\Delta K^{1/2}(s,t)$ and M is some large number to be decided later.

For any $\beta \in \mathcal{H}^*$, observe that

$$\begin{aligned} ||\beta||_K^2 &= || \sum_{k=M+1}^{2M} b_k M^{-1/2} L_{K^{1/2}} \varphi_k ||_K^2 \\ &= \sum_{k=M+1}^{2M} b_k^2 M^{-1} ||L_{K^{1/2}} \varphi_k||_K^2 \\ &\leq \sum_{k=M+1}^{2M} M^{-1} ||L_{K^{1/2}} \varphi_k||_K^2 \\ &= 1, \end{aligned}$$

which follows from the fact that

$$< L_{K^{1/2}} \varphi_k, L_{K^{1/2}} \varphi_l >_K = < L_K \varphi_k, \varphi_l >_K = < \varphi_k, \varphi_l >_{L_2} = \delta_{kl}.$$

Therefore $\mathcal{H}^* \subset \mathcal{H}(K) = \{\beta : ||\beta||_k < \infty\}$.

The Varshamov-Gilbert bound shows that for any $M \geq 8$, there exists a set $\mathcal{B} = \{b^{(0)}, b^{(1)}, \ldots, b^{(N)}\} \subset \{0,1\}^M$such that

1. $b^{(0)} = (0, \ldots, 0)'$;

2. $H(b,b') > M/8$ for any $b \neq b' \in \mathcal{B}$, where $H(\cdot,\cdot) = \frac{1}{4}\sum_{i=1}^{M}(b_i - b'_i)^2$ is the Hamming distance;

3. $N \geq 2^{M/8}$.

The subset $\{\beta^{(0)},\ldots,\beta^{(N)}\} \subset \mathcal{H}^*$ is chosen as $\beta^{(i)} = \sum_{k=M+1}^{2M} b_{k-M}^{(i)} M^{-1/2} L_{K^{1/2}\varphi_k}$, $i = 0,\ldots N$.

For any $0 \leq i < j \leq N$, observe that

$$\begin{aligned} d^2(\beta^{(i)},\beta^{(j)}) &= \mathbb{E}\Delta\big(\eta_{\beta^{(i)}}(X) - \eta_{\beta^{(j)}}(X)\big)^2 \\ &= ||L_{C_\Delta^{1/2}} \sum_{k=M+1}^{2M} (b_{k-M}^{(i)} - b_{k-M}^{(j)}) M^{-1/2} L_{K^{1/2}\varphi_k}||_{L_2}^2 \\ &= \sum_{k=M+1}^{2M} (b_{k-M}^{(i)} - b_{k-M}^{(j)})^2 M^{-1} ||L_{C_\Delta^{1/2}} L_{K^{1/2}\varphi_k}||_{L_2}^2 \\ &= \sum_{k=M+1}^{2M} (b_{k-M}^{(i)} - b_{k-M}^{(j)})^2 M^{-1} s_k. \end{aligned}$$

On one hand, we have

$$\begin{aligned} d^2(\beta^{(i)},\beta^{(j)}) &= \sum_{k=M+1}^{2M} (b_{k-M}^{(i)} - b_{k-M}^{(j)})^2 M^{-1} s_k \\ &\geq s_{2M} M^{-1} \sum_{k=1}^{M} (b_k^{(i)} - b_k^{(j)})^2 \\ &= 4 s_{2M} M^{-1} H(b^{(i)}, b^{(j)}) \\ &\geq s_{2M}/2. \end{aligned}$$

On the other hand, we have

$$\begin{aligned} d^2(\beta^{(i)},\beta^{(j)}) &= \sum_{k=M+1}^{2M} (b_{k-M}^{(i)} - b_{k-M}^{(j)})^2 M^{-1} s_k \\ &\leq s_M M^{-1} \sum_{k=1}^{M} (b_k^{(i)} - b_k^{(j)})^2 \\ &\leq s_M. \end{aligned}$$

So altogether,

$$s_{2M}/2 \leq d^2(\beta^{(i)},\beta^{(j)}) \leq s_M. \tag{3.23}$$

Let $P_j,\ j = 1, \dots N$, be the likelihood function with data $\{(T_i, \Delta_i, W_i(s)), i = 1, \dots n\}$ and $\beta^{(j)}$, i.e

$$P_j = \prod_{i=1}^{n} \left[f_{T^u|W}(T_i) S_{T^c|W}(T_i)\right]^{\Delta_i} \cdot \left[f_{T^c|W}(T_i) S_{T^u|W}(T_i)\right]^{1-\Delta_i}.$$

Let $c_{T^c} = \prod_{i=1}^{n} \left[S_{T^c|W}(T_i)\right]^{\Delta_i} \left[S_{T^u|W}(T_i)\right]^{1-\Delta_i}$, which does not depend on $\beta^{(j)}$, then

$$P_j = c_{T^c} \prod_{i=1}^{n} \left[h_0(T_i) \exp(\theta_0' Z_i + \eta_{\beta^{(j)}}(X_i))\right]^{\Delta_i} \cdot \exp\{-H_0(T_i) \cdot e^{\theta_0' Z_i + \eta_{\beta^{(j)}}(X_i)}\}.$$

We calculte the Kullback-Leibler distance between P_j and P_0 as

$$\begin{aligned}
KL(P_j, P_0) &= \mathbb{E}_{P_j} \log \frac{P_j}{P_0} \\
&= \mathbb{E}_{P_j} \Big\{ \Delta_i \sum_{i=1}^{n} \{\eta_{\beta^{(j)} - \beta^{(0)}}(X_i)\} + \sum_{i=1}^{n} H_0(T_i)\, e^{\theta_0' Z_i} [\exp(\eta_{\beta^{(0)}}(X_i)) \\
&\qquad - \exp(\eta_{\beta^{(j)}}(X_i))] \Big\} \\
&= n \mathbb{E}_{P_j} \Delta [\eta_{\beta^{(j)} - \beta^{(0)}}(X)] + n \mathbb{E}_{P_j} H_0(T)\, e^{\theta_0' Z} [\exp(\eta_{\beta^{(0)}}(X)) \\
&\qquad - \exp(\eta_{\beta^{(j)}}(X))] \\
&= n \mathbb{E}_{P_j}^{W} \mathbb{E}_{P_j}^{T,\Delta} \{H_0(T)\,|W\}\, e^{\theta_0' Z} [\exp(\eta_{\beta^{(0)}}(X)) - \exp(\eta_{\beta^{(j)}}(X))] \Big],
\end{aligned}$$

where

$$\begin{aligned}
\mathbb{E}_{p_j}^{T,\Delta}(H_0(T)\,|W) &= \mathbb{E}^{T^c} \big\{ \mathbb{E}_{p_j}^{T,\Delta}(H_0(T)\,|T^c, W) \big| W \big\} \\
&= \mathbb{E}^{T^c} \big\{ \int_0^{T^c} H_0(t) f_{T^u|W}(t) dt + H_0(T^c) \mathbb{P}(T^u > T^c | T^c, W) \big| W \big) \big\},
\end{aligned}$$

$$\begin{aligned}
&\int_0^{T^c} H_0(t) f_{T^u|W}(t) dt \\
&= \int_0^{T^c} H_0(t) \cdot h_0(t) \exp[\theta_0' Z + \eta_{\beta^{(j)}}(X)] \exp\{-H_0(t) e^{\theta_0' Z + \eta_{\beta^{(j)}}(X)}\} dt \\
&= e^{-\theta_0' Z - \eta_{\beta^{(j)}}(X)} \int_0^{T^c} e^{\theta_0' Z + \eta_{\beta^{(j)}}(X)} H_0(T) \exp\{-H_0(T) \cdot e^{\theta_0' Z + \eta_{\beta^{(j)}}(X)}\} \\
&\qquad de^{\theta_0' Z + \eta_{\beta^{(j)}}(X)} H_0(T) \\
&= \exp(-\theta_0' Z + \eta_{\beta^{(j)}}(X)) \int_0^{a} u e^{-u} du \ \Big|_{a = e^{\theta_0' Z + \eta_{\beta^{(j)}}(X)} H_0(T^c)} \\
&= \exp(-\theta_0' Z - \eta_{\beta^{(j)}}(X)) [1 - e^{-a} - a\, e^{-a}] \Big|_{a = e^{\theta_0' Z + \eta_{\beta^{(j)}}(X)} H_0(T^c)},
\end{aligned}$$

and

$$\begin{aligned}\mathbb{P}(T^u > T^c|T^c, W) &= S_{T^u|W}(T^c) \\ &= \exp\{-H_0(T^c)e^{\theta_0' Z+\eta_{\beta^{(j)}}(X)}\}.\end{aligned}$$

Therefore

$$\begin{aligned}&\mathbb{E}_{p_j}^{T,\Delta}(H_0(T)\,|T^c, W) \\ &= e^{-\theta_0' Z-\eta_{\beta^{(j)}}(X)}[1-\exp\{-H_0(T^c)e^{\theta_0' Z+\eta_{\beta^{(j)}}(X)}\}] - H_0(T^c)\cdot \\ &\qquad \exp\{-H_0(T^c)e^{\theta_0' Z+\eta_{\beta^{(j)}}(X)}\} + H_0(T^c)\exp\{-H_0(T^c)e^{\theta_0' Z+\eta_{\beta^{(j)}}(X)}\} \\ &= e^{-\theta_0' Z-\eta_{\beta^{(j)}}(X)}[1-\exp\{-H_0(T^c)e^{\theta_0' Z+\eta_{\beta^{(j)}}(X)}\}] \\ &= e^{-\theta_0' Z-\eta_{\beta^{(j)}}(X)}[F_{T^u|W}(T^c)] \\ &= e^{-\theta_0' Z-\eta_{\beta^{(j)}}(X)}\mathbb{P}(T^u \le T^c|T^c, W),\end{aligned}$$

and further

$$\begin{aligned}\mathbb{E}_{p_j}^{T,\Delta}(H_0(T)\,|W) &= \mathbb{E}^{T^c}\left\{\mathbb{E}_{p_j}^{T,\Delta}(H_0(T)\,|T^c, W)\middle|W\right\} \\ &= e^{-\theta_0' Z-\eta_{\beta^{(j)}}(X)}\mathbb{P}(T^u \le T^c|W) \\ &= e^{-\theta_0' Z-\eta_{\beta^{(j)}}(X)}\mathbb{E}[\Delta|W].\end{aligned}$$

Then the KL distance becomes

$$\begin{aligned}KL(P_j, P_0) &= n\mathbb{E}_{P_j}^{W}\mathbb{E}[\Delta|W]e^{-\theta_0' Z-\eta_{\beta^{(j)}}(X)}\, e^{\theta_0' Z}[\exp(\eta_{\beta^{(0)}}(X)) - \exp(\eta_{\beta^{(j)}}(X))]\Big] \\ &= n\mathbb{E}_{P_j}^{W,\Delta}\Delta[\exp(\eta_{\beta^{(0)}}(X) - \eta_{\beta^{(j)}}(X)) - 1\Big] \\ &= n\mathbb{E}_{P_j}^{W,\Delta}[\frac{1}{2}\Delta(\eta_{\beta^{(0)}}(X) - \eta_{\beta^{(j)}}(X))^2 + o(\Delta(\eta_{\beta^{(0)}}(X) - \eta_{\beta^{(j)}}(X))^2)] \\ &\le n\mathbb{E}_{P_j}^{X}[\frac{1}{2}(\eta_{\beta^{(0)}}(X) - \eta_{\beta^{(j)}}(X))^2 + o((\eta_{\beta^{(0)}}(X) - \eta_{\beta^{(j)}}(X))^2)] \\ &\lesssim nd^2(\beta^{(j)}, \beta^{(0)}) \\ &\lesssim ns_M.\end{aligned}$$

Therefore for some positive constant c_1,

$$KL(P_j, P_0) \le c_1 nM^{-2r}.$$

By taking M to be the smallest integer greater than $c_2\gamma^{-\frac{1}{2r+1}}n^{\frac{1}{2r+1}}$ with $c_2 = (c_1 \cdot 8\log 2)^{1/(1+2r)}$, we verified (3.21) that

$$\frac{1}{N}\sum_{j=1}^{N} KL(P_j, P_0) \;\leq\; \gamma \log N.$$

Meanwhile, since $d^2(\beta^{(i)}, \beta^{(j)}) \geq s_{2M}/2$ and $s_{2M} \asymp (2M)^{-2r}$, condition (3.20) is verified by plugging in M. ■

3.5.5 Derivation of $GCV(\lambda)$

Recall that given the observations $\{(T_i, \Delta_i, W_i)\}_{i=1}^n$, $\hat{\beta}_\lambda$ can be written in the form of

$$\hat{\beta}_\lambda(t) = \sum_{k=1}^{m} d_k\xi_k(t) + \sum_{i=1}^{n} c_i \int_0^1 X_i(s)K_1(s,t)ds.$$

For simplicity, let $\xi_{k+j}(t) = \int_0^1 X_j(s)K_1(s,t)ds$, $j = 1, \ldots n$, then write $\beta(t) = \sum_{k=1}^{m+n} c_k^{(\beta)}\xi_k(t)$. In this way,

$$\eta_\alpha(W_i) = \sum_{k=1}^{p} \theta_k Z_{ik} + \sum_{k=1}^{m+n} c_k^{(\beta)} \int X_i(t)\xi_k(t)dt.$$

Let $S^{(\beta)}$ be an $n\times(m+n)$ matrix with the $(i,j)th$ entry defined as $S_{ij}^{(\beta)} = \int X_i(s)\xi_j(s)ds$, and $\mathbf{Z} = (Z_1, \cdots, Z_n)_{n\times p}$. Denote $S = (\mathbf{Z}, S^{(\beta)})$, a $n \times (p+m+n)$ matrix, and $(\eta_\alpha(W_1), \ldots \eta_\alpha(W_n))^T = S\cdot c$ with $c = (c_1, \ldots, c_{p+n+m})^T = (\theta_1, \ldots \theta_p, c_1^{(\beta)}, \ldots, c_{m+n}^{(\beta)})^T$.

Since $\xi_1, \ldots \xi_m$ are the bases of the null space with the semi-norm $J(\cdot)$, we can write J as $J(\beta) = c^TQc$, with Q a $(p+m+n)\times(p+m+n)$ diagonal block matrix whose non-zero entries only occur in the $n \times n$ submatrix $(Q_{i,j})_{i,j=p+m+1}^{p+m+n}$.

Let $\Delta = (\Delta_1, \ldots, \Delta_n)^T$ and $Y_j(t) = I\{t \geq T_j\}$. Under the above expressions, we can write the penalized partial likelihood as a function of the coefficient c:

$$A_\lambda(c) = -\Delta' S\cdot c/n + \frac{1}{n}\sum_{i=1}^{n} \Delta_i \log\{\sum_{j=1}^{n} Y_j(T_i)e^{S_{j\cdot}c}\} + \lambda c^TQc,$$

where $S_{j\cdot}$ is the j^{th} row of S.

For any $\alpha \in \mathbb{R}^p \times \mathcal{H}(K)$, functions $f, g \in \mathcal{H}(K)$, and $z, z^* \in \mathbb{R}^{n\times 1}$ define

$$\mu_\alpha(f|t) = \frac{\sum_{j=1}^n Y_j(t)e^{\eta_\alpha(W_j)}\eta_f(X_j)}{\sum_{j=1}^n Y_j(t)e^{\eta_\alpha(W_j)}}, \qquad \mu_\alpha(z|t) = \frac{\sum_{j=1}^n Y_j(t)e^{\eta_\alpha(W_j)}z_j}{\sum_{j=1}^n Y_j(t)e^{\eta_\alpha(W_j)}},$$

and

$$\mu_\alpha(f,\, g|t) = \frac{\sum_{j=1}^n Y_j(t)e^{\eta_\alpha(W_j)}\eta_f(X_j)\cdot\eta_g(X_j)}{\sum_{j=1}^n Y_j(t)e^{\eta_\alpha(W_j)}},$$

$$\mu_\alpha(z,\, z^*|t) = \frac{\sum_{j=1}^n Y_j(t)e^{\eta_\alpha(W_j)}z_j\, z_j^*}{\sum_{j=1}^n Y_j(t)e^{\eta_\alpha(W_j)}},$$

$$\mu_\alpha(f,\, z|t) = \frac{\sum_{j=1}^n Y_j(t)e^{\eta_\alpha(W_j)}\eta_f(X_j)\, z_j}{\sum_{j=1}^n Y_j(t)e^{\eta_\alpha(W_j)}}.$$

Define $\mu_\alpha(g) = \frac{1}{n}\sum_{i=1}^n \mu_\alpha(g|T_i)$, $V_\alpha(f,g|t) = \mu_\alpha(f,\, g|t) - \mu_\alpha(f|t)\mu_\alpha(g|t)$, and $V_\alpha(f,g) = \frac{1}{n}\sum_{i=1}^n V_\alpha(f,g|T_i)$, and define by analogy $\mu_\alpha(z)$, $V_\alpha(z,z^*|t)$, $V_\alpha(f,z|t)$, $V_\alpha(z,z^*)$, and $V_\alpha(f,z)$. Now take the derivative of $A_\lambda(c)$ at $\tilde{\alpha} = S\cdot\tilde{c}$ with respect to c, we have

$$\frac{\partial A_\lambda(c)}{\partial c}\Big|_{\tilde{\alpha}} = -S^T\Delta/n + \mu_{\tilde{\alpha}}(\varsigma) + 2\lambda Q\tilde{c},$$

and

$$\frac{\partial^2 A_\lambda(c)}{\partial c^2}\Big|_{\tilde{\alpha}} = V_{\tilde{\alpha}}(\varsigma,\varsigma^T) + 2\lambda Q,$$

where $\varsigma = (Z_{\cdot 1},\ldots,Z_{\cdot p},\xi_1(s),\ldots,\xi_{m+n}(s))^T$. To obtain the minimum of $A_\lambda(c)$, we apply the Newton-Raphson algorithm to $\partial A_\lambda(c)/\partial c$. That is,

$$[V_{\tilde{\alpha}}(\varsigma,\varsigma^T) + 2\lambda Q](c-\tilde{c}) = S^T\Delta/n - \mu_{\tilde{\alpha}}(\varsigma) - 2\lambda Q\tilde{c}.$$

To simplify the notations, let $H = [V_{\tilde{\alpha}}(\varsigma,\varsigma^T) + 2\lambda Q]$ and $h = -\mu_{\tilde{\alpha}}(\varsigma) + V_{\tilde{\alpha}}(\varsigma,\varsigma^T)\tilde{c}$, so $\hat{c} \approx H^{-1}(S'\Delta/n + h)$ and

$$\hat{c}^{[i]} \approx H^{-1}(\frac{S^T\Delta - \Delta_i S_{i\cdot}^T}{n-1} + h) = \hat{c} - \Delta_i\cdot\frac{H^{-1}S_{i\cdot}^T}{n-1} + \frac{H^{-1}S'\Delta}{n(n-1)}.$$

Then the first term of $\widehat{RKL}$ becomes

$$\sum_{i=1}^n \eta_{\tilde{\alpha}_\lambda}^{[i]}(W_i) = \sum_{i=1}^n \eta_{\tilde{\alpha}_\lambda}(W_i) - \sum_{i=1}^n[\Delta_i\cdot\frac{S_{i\cdot}H^{-1}S_{i\cdot}^T}{n-1} + \frac{S_{i\cdot}H^{-1}S'\Delta}{n(n-1)}].$$

Simplifying this leads to

$$\sum_{i=1}^n \eta_{\tilde{\alpha}_\lambda}^{[i]}(W_i) = \sum_{i=1}^n \eta_{\tilde{\alpha}_\lambda}(W_i) - \frac{1}{(n-1)}tr[(SH^{-1}S)(diag\Delta - \Delta\mathbf{1}'/n)],$$

where $diag\Delta$ is an $n \times n$ diagonal matrix with diagonal entries $\Delta_1, \ldots, \Delta_n$. Plugging this back to $\widehat{RKL}$, then $GCV(\lambda)$ is obtained.

If the efficient estimator β^*_λ is used instead, the derivation and therefore the main result remain the same by adjusting the definition of ξ and $S^{(\beta)}$ accordingly.

3.5.6 Proofs of Lemmas

Lemma 3.1 *Following the former notations, for* $0 \leq s \leq 1$, *let*

$$g(t,s) = \frac{S_1(t, \alpha_0 + s\widetilde{\alpha})[\alpha^*]}{S_0(t, \alpha_0 + s\widetilde{\alpha})}.$$

Denote $R_s(t) = Y(t)\exp(\eta_{\alpha_0} + s\eta_{\widetilde{\alpha}})/S_0(t, \alpha_0 + s\widetilde{\alpha})$. *We have*

$$\begin{aligned}\frac{\partial}{\partial s}g(t,s) &= \mathbb{E}[R_s(t)\eta_{\widetilde{\alpha}}\eta_{\alpha^*}] - \mathbb{E}[R_s(t)\eta_{\widetilde{\alpha}}]\mathbb{E}[R_s(t)\eta_{\alpha^*}] \\ &= \mathbb{E}\big\{R_s(t)\big(\eta_{\widetilde{\alpha}} - \mathbb{E}[R_s(t)\eta_{\widetilde{\alpha}}]\big)\big(\eta_{\alpha^*} - \mathbb{E}[R_s(t)\eta_{\alpha^*}]\big),\end{aligned}$$

and

$$\begin{aligned}\frac{\partial^2}{\partial s^2}g(t,s) = \mathbb{E}[R_s(t)\eta_{\alpha^*}\eta^2_{\widetilde{\alpha}}] - 2\mathbb{E}[R_s(t)\eta_{\widetilde{\alpha}}]\mathbb{E}[R_s(t)\eta_{\alpha^*}\eta_{\widetilde{\alpha}}] \\ - \mathbb{E}[R_s(t)\eta_{\alpha^*}]\mathbb{E}[R_s(t)\eta^2_{\widetilde{\alpha}}] + 2\mathbb{E}[R_s(t)\eta_{\alpha^*}]\mathbb{E}[R_s(t)\eta_{\widetilde{\alpha}}]^2.\end{aligned}$$

Proof The lemma follows by direct calculation. ■

Lemma 3.2 *Let* α_0 *be the true coefficients. Under assumption A(1)-A(4), we have*

$$P_\Delta m_0(\cdot, \alpha) - P_\Delta m_0(\cdot, \alpha_0) \asymp -d^2(\alpha, \alpha_0).$$

Proof Observe that

$$\begin{aligned}&P_\Delta m_0(\cdot, \alpha) - P_\Delta m_0(\cdot, \alpha_0) \\ &= P_\Delta(m_0(\cdot, \alpha) - m_0(\cdot, \alpha_0)) \\ &= P_\Delta\{\eta_{\alpha-\alpha_0}(W) - \log S_0(\cdot, \alpha) + \log S_0(\cdot, \alpha_0)\}1_{\{0 \leq T \leq \tau\}} \\ &= -P_\Delta\{\log S_0(\cdot, \alpha) - \log S_0(\cdot, \alpha_0)\}1_{\{0 \leq T \leq \tau\}}.\end{aligned}$$

Let $\widetilde{\alpha} = (\theta - \theta_0, \beta - \beta_0)$ and $G(t,s) = \log(S_0(t, \alpha_0 + s\widetilde{\alpha}))$, then

$$P_\Delta m_0(\cdot, \alpha) - P_\Delta m_0(\cdot, \alpha_0) = -P_\Delta(G(\cdot, 1) - G(\cdot, 0))1_{\{0 \le T \le \tau\}}.$$

For fixed t, take the derivative of $G(t,s)$ with respect to s, we have

$$\frac{\partial}{\partial s}G(t,s) = \frac{S_1(t, \alpha_0 + s\widetilde{\alpha})[\widetilde{\alpha}]}{S_0(t, \alpha_0 + s\widetilde{\alpha})} \stackrel{def}{=} g(t,s).$$

Noting that $P_\Delta \frac{\partial}{\partial s}G(\cdot, 0) = P_\Delta g(\cdot, 0) = 0$, then lemma 3.1 implies,

$$\frac{\partial^2}{\partial s^2}G(t,s) = \frac{\partial}{\partial s}g(t,s) = \mathbb{E}[R_s(t)\eta_{\widetilde{\alpha}}^2] - \big(\mathbb{E}[R_s(t)\eta_{\widetilde{\alpha}}]\big)^2 = \mathbb{E}R_s(t)\big(\eta_{\widetilde{\alpha}} - \mathbb{E}[R_s(t)\eta_{\widetilde{\alpha}}]\big)^2,$$

where $R_s(t) = \frac{Y(t)e^{\eta_{\alpha_0} + s\eta_{\widetilde{\alpha}}}}{S_0(t, \eta_{\alpha_0} + s\eta_{\widetilde{\alpha}})}$. Therefore for some $\gamma \in [0,1]$,

$$\begin{aligned} G(t,1) - G(t,0) &= G_s'(t,0) + \frac{1}{2}G_s''(t,\gamma) \\ &= g(t,0) + \frac{1}{2}\mathbb{E}R_\gamma(t)\big(\eta_{\widetilde{\alpha}} - \mathbb{E}[R_\gamma(t)\eta_{\widetilde{\alpha}}]\big)^2 \\ &= g(t,0) + \frac{1}{2}\mathbb{E}^W\mathbb{E}\big(R_\gamma(t)|W\big)\big(\eta_{\widetilde{\alpha}} - \mathbb{E}[R_\gamma(t)\eta_{\widetilde{\alpha}}]\big)^2. \end{aligned}$$

By the definition of $R_s(t)$,

$$\mathbb{E}\big(R_\gamma(t)|W\big) = P(T \ge t|W)\exp(\eta_{\alpha_0 + \gamma\widetilde{\alpha}}(W))/S_0(t, \eta_{\alpha_0 + \gamma\widetilde{\alpha}}).$$

By the assumptions and for $t \in [0, \tau]$, there exists constants $c_1 > c_2 > 0$ not depending on t, such that

$$c_2 \le \mathbb{E}[R_\gamma(t)|W] \le c_1.$$

On one hand,

$$\begin{aligned} &G(t,1) - G(t,0) \\ &\ge g(t,0) + \frac{1}{2}c_2\mathbb{E}\big(\eta_{\widetilde{\alpha}} - \mathbb{E}[R_\gamma(t)\eta_{\widetilde{\alpha}}]\big)^2 \ge g(t,0) + \frac{1}{2}c_2\mathbb{E}\Delta\big(\eta_{\widetilde{\alpha}} - \mathbb{E}[R_\gamma(t)\eta_{\widetilde{\alpha}}]\big)^2 \\ &= g(t,0) + \frac{1}{2}c_2\mathbb{E}\Delta\eta_{\widetilde{\alpha}}^2 - 2\mathbb{E}\Delta\eta_{\widetilde{\alpha}}\mathbb{E}[R_\gamma(t)\eta_h] + \mathbb{E}[R_\gamma(t)\eta_{\widetilde{\alpha}}]^2 \\ &\ge g(t,0) + \frac{1}{2}c_2 d^2(\alpha, \alpha_0), \end{aligned}$$

which follows from the fact that $E\Delta\eta_{\widetilde{\alpha}} = 0$. So

$$\begin{aligned} P_\Delta m_0(\cdot,\alpha) - P_\Delta m_0(\cdot,\alpha_0) &= -P_\Delta\{G(\cdot,1) - G(\cdot,0)\}1_{\{0\le T\le\tau\}} \\ &\le -P_\Delta d^2(\alpha,\alpha_0)1_{\{0\le T\le\tau\}} \\ &\lesssim -d^2(\alpha,\alpha_0). \end{aligned} \tag{3.24}$$

On the other hand,

$$\begin{aligned} G(t,1) - G(t,0) &\le g(t,0) + \frac{1}{2}c_1\mathbb{E}\big(\eta_{\widetilde{\alpha}}^2 - E[R_\gamma(t)\eta_{\widetilde{\alpha}}]\big)^2 \\ &\le g(t,0) + c_1\{E\eta_{\widetilde{\alpha}}^2 + (E[R_\gamma(t)\eta_{\widetilde{\alpha}}])^2\}. \end{aligned}$$

Since $(E[R_\gamma(t)\eta_{\widetilde{\alpha}}])^2 = (E^W E[R_\gamma(t)|W]^2 \cdot \eta_{\widetilde{\alpha}}^2) \le c_1^2\epsilon^{-1}E\Delta\eta_{\widetilde{\alpha}}^2$, we arrive at

$$\begin{aligned} P_\Delta m_0(\cdot,\alpha) - P_\Delta m_0(\cdot,\alpha_0) &= -P_\Delta\{G(\cdot,1) - G(\cdot,0)\}1_{\{0\le T\le\tau\}} \\ &\gtrsim -P_\Delta d^2(\alpha,\alpha_0)1_{\{0\le T\le\tau\}} \\ &\gtrsim -d^2(\alpha,\alpha_0). \end{aligned} \tag{3.25}$$

Combining (3.24) and (3.25) we have

$$P_\Delta m_0(\cdot,\alpha) - P_\Delta m_0(\cdot,\alpha_0) \asymp -d^2(\alpha,\alpha_0).$$

■

Lemma 3.3 *$\mathcal{F}_1 = \{\Delta m_0(t,W,\alpha) : \alpha \in \mathbb{R}_M \times \mathcal{H}^M\}$ and $\mathcal{F}_2 = \{Y(t)e^{\eta_\alpha(W)} : \alpha \in \mathbb{R}_M \times \mathcal{H}^M,\ 0 \le t \le \tau\}$ are P-Glivenko-Cantelli.*

Proof Given that $\eta_\alpha(W) = \theta' Z + \eta_\beta(X)$ is bounded almost surely, it is easy to see that $\Delta m_0(t,W,\alpha) = \Delta[\eta_\alpha(W) - \log S_0(t,\alpha)]1_{\{0\le t\le\tau\}}$ and $Y(t)e^{\eta_\alpha(W)}$ are bounded. So following Theorem 19.13 in [48], it is sufficient to show that $\mathcal{N}(\epsilon,\mathcal{F}_i,L_1(P)) < \infty$ for $i = 1,2$.

For any $f = \Delta m_0(t, W, \alpha)$, and $f_1 = \Delta m_0(t, W, \alpha_1)$ in $\mathcal{F}_1$,

$$\begin{aligned}
||f - f_1||_{L_1(P)} &= P|f - f_1| = P|\Delta m_0(\cdot, \alpha) - \Delta m_0(\cdot, \alpha_1)| \\
&= P|\Delta\big[\eta_\alpha(W) - \eta_{\alpha_1}(W) - \log \frac{S_0(\cdot, \alpha)}{S_0(\cdot, \alpha_1)}\big] 1_{\{0 \le T \le \tau\}}| \\
&\le P|\eta_\alpha(W) - \eta_{\alpha_1}(W)| + P|[\log S_0(\cdot, \alpha) - \log S_0(\cdot, \alpha_1)] 1_{\{0 \le T \le \tau\}}| \\
&\lesssim P|\eta_\alpha(W) - \eta_{\alpha_1}(W)| + \sup_{0 \le t \le \tau} |S_0(t, \alpha) - S_0(t, \alpha_1)| \\
&\lesssim P|\eta_\alpha(W) - \eta_{\alpha_1}(W)| + \sup_{0 \le t \le \tau} |E(Y(t)e^{\eta_\alpha(W)} - Y(t)e^{\eta_{\alpha_1}(W)})| \\
&\lesssim P|\eta_\alpha(W) - \eta_{\alpha_1}(W)|.
\end{aligned}$$

Therefore $\mathcal{N}(\epsilon, \mathcal{F}_1, L_1(P)) \asymp \mathcal{N}(\epsilon, \{\eta_\alpha(W) : \alpha \in \mathbb{R}_M \times \mathcal{H}^M\}, L_1(P))$.

Similarly for $f = Y(t)e^{\eta_\alpha(W)}$, and $f_1 = Y(t)e^{\eta_{\alpha_1}(W)}$: in $\mathcal{F}_2$,

$$\begin{aligned}
||f - f_1||_{L_1(P)} &= P|f - f_1| \\
&\le P|e^{\eta_\alpha(W)} - e^{\eta_{\alpha_1}(W)}| \\
&\lesssim P|\eta_\alpha(W) - \eta_{\alpha_1}(W)|,
\end{aligned}$$

and $\mathcal{N}(\epsilon, \mathcal{F}_2, L_1(P)) \asymp \mathcal{N}(\epsilon, \{\eta_\alpha(W) : \alpha \in \mathbb{R}_M \times \mathcal{H}^M\}, L_1(P))$.

So it suffices to show that $\mathcal{N}(\epsilon, \{\eta_\alpha(W) : \alpha \in \mathbb{R}_M \times \mathcal{H}^M\}, L_1(P)) < \infty$, which is obvious since $\eta_\alpha(W)$ is bounded almost surely for $\alpha \in \mathbb{R}_M \times \mathcal{H}^M$. ■

Lemma 3.4 *Let I and III be defined as*

$$\begin{aligned}
I &= (P_{\Delta n} - P_\Delta)(m_0(\cdot, \alpha) - m_0(\cdot, \alpha_0)), \\
III &= (P_n - P)\big\{Y(t)\big[\exp(\eta_\alpha(W)) - \exp(\eta_{\alpha_0}(W))\big]\big\},
\end{aligned}$$

and $\mathcal{B}_\delta = \{\alpha \in \mathbb{R}^p \times \mathcal{H}(K) : \delta/2 \le d(\alpha, \alpha_0) \le \delta\}$, then

$$\begin{aligned}
\sup_{\alpha \in \mathcal{B}_\delta} I &= O(\delta^{\frac{2r-1}{2r}} n^{-1/2}), \\
\sup_{\alpha \in \mathcal{B}_\delta} III &= O(\delta^{\frac{2r-1}{2r}} n^{-1/2}), \quad \textit{for for } t \in [0, \tau].
\end{aligned}$$

Proof Consider

$$\begin{aligned}
\mathcal{M}_{\delta 1} &= \{\Delta[m_0(t, W, \alpha) - m_0(t, W, \alpha_0)] 1_{\{0 \le t \le \tau\}},\ \alpha \in \mathcal{B}_\delta\}, \\
\mathcal{M}_{\delta 2} &= \{Y(t)\big[\exp(\eta_\alpha(W)) - \exp(\eta_\alpha(W))\big],\ \alpha \in \mathcal{B}_\delta,\ t \in [0, \tau]\},
\end{aligned}$$

with $L_2(P)$ norm, i.e for any $f \in \mathcal{M}_{\delta 1}$, $||f||_{P,2} = (\int f^2 dP)^{1/2} = \big(E^{t,W} f^2(t, W, \alpha)\big)^{1/2}$, and for any $f \in \mathcal{M}_{\delta 2}$, $||f||_{P,2} = (\int f^2 dP)^{1/2} = \big(E^{T,W} f^2(T, W, t, \alpha)\big)^{1/2}$. Then it suffices to show that

$$\big|\big|||\mathbb{G}_n||_{\mathcal{M}_{\delta 1}}\big|\big|_{P,2} = O(\delta^{\frac{2r-1}{2r}}),$$

$$\big|\big|||\mathbb{G}_n||_{\mathcal{M}_{\delta 2}}\big|\big|_{P,2} = O(\delta^{\frac{2r-1}{2r}}),$$

where $\mathbb{G}_n = \sqrt{n}(P_n - P)$ and $||\mathbb{G}_n||_{\mathcal{M}_{\delta i}} = \sup_{f \in \mathcal{M}_{\delta i}} |\mathbb{G}_n f|$, $i = 1, 2$.

We first show that

$$\log \mathcal{N}(\epsilon, \mathcal{M}_{\delta 1}, ||\cdot||_{p,2}) \leq O((p + \epsilon^{-1/r}) \log(\frac{\delta}{\epsilon})),$$

and

$$\log \mathcal{N}(\epsilon, \mathcal{M}_{\delta 2}(t), ||\cdot||_{p,2}) \leq O((p + \epsilon^{-1/r}) \log(\frac{\delta}{\epsilon})), \quad \text{for all } t \in [0, \tau].$$

Suppose there exist functions $f_1, \ldots, f_m \in \mathcal{M}_{\delta 1}$, such that

$$\min_{1 \leq i \leq m} ||f - f_i||_{p,2} < \epsilon, \quad \text{for all } f \in \mathcal{M}_{\delta 1}.$$

This is equivalent to the existence of $\alpha_1, \ldots \alpha_m \in \mathcal{B}_\delta$, s.t

$$\min_{1 \leq i \leq m} \big|\big|\Delta[m_0(\cdot, \alpha) - m_0(\cdot, \alpha_i)] 1_{\{0 \leq T \leq \tau\}}\big|\big|_{p,2} < \epsilon, \quad \text{for all } \alpha \in \mathcal{B}_\delta.$$

Observe that

$$\begin{aligned}
& \{\Delta[m_0(t, W, \alpha) - m_0(t, W, \alpha_i)] 1_{\{0 \leq t \leq \tau\}}\}^2 \\
= \;& \Delta\big[\eta_\alpha(W) - \eta_{\alpha_i}(W) - \log \frac{S_0(t, \alpha)}{S_0(t, \alpha_i)}\big]^2 1_{\{0 \leq t \leq \tau\}} \\
\leq \;& 2\Delta\big\{[\eta_\alpha(W) - \eta_{\alpha_i}(W)]^2 + [\log \frac{S_0(t, \alpha)}{S_0(t, \alpha_i)}]^2\big\} 1_{\{0 \leq t \leq \tau\}} \\
\leq \;& 2\Delta\big\{[\eta_\alpha(W) - \eta_{\alpha_i}(W)]^2 + c[S_0(t, \alpha) - S_0(t, \alpha_i)]^2\big\} 1_{\{0 \leq t \leq \tau\}} \\
= \;& 2\Delta\big\{[\eta_\alpha(W) - \eta_{\alpha_i}(W)]^2 + c[\mathbb{E}Y(t)\{\exp(\eta_\alpha(W)) - \exp(\eta_{\alpha_i}(\alpha))\}]^2\big\} 1_{\{0 \leq t \leq \tau\}} \\
\leq \;& 2\Delta\big\{[\eta_\alpha(W) - \eta_{\alpha_i}(W)]^2 + c\mathbb{E}Y^2(t)\mathbb{E}\big[\exp(\eta_\alpha(W)) - \exp(\eta_{\alpha i}(W))\big]^2\big\} 1_{\{0 \leq t \leq \tau\}} \\
\leq \;& 2\Delta\big\{[\eta_\alpha(W) - \eta_{\alpha_i}(W)]^2 + c_1\mathbb{E}Y(t)\mathbb{E}\big[\eta_\alpha(W) - \eta_{\alpha_i}(W)\big]^2\big\} 1_{\{0 \leq t \leq \tau\}}.
\end{aligned}$$

Then

$$\begin{aligned}
&\left|\left|\Delta[m_0(\cdot,\alpha) - m_0(\cdot,\alpha_i)]1_{\{0\le T\le\tau\}}\right|\right|_{p,2}^2 \\
&= P\{\Delta[m_0(\cdot,\alpha) - m_0(\cdot,\alpha_i)]1_{\{0\le T\le\tau\}}\}^2 \\
&\lesssim d^2(\alpha,\alpha_i).
\end{aligned}$$

Therefore, the covering number for $\mathcal{M}_{\delta 1}$ is of the same order as that for $\mathcal{B}_\delta$. To be more specific,

$$N(\epsilon, \mathcal{M}_{\delta 1}, ||\cdot||_{p,2}) \le N(\epsilon/C, \mathcal{B}_\delta, d). \tag{3.26}$$

In addition, we know that

$$d^2(\alpha,\alpha_i) \le 2\mathbb{E}\Delta[(\theta-\theta_i)'Z]^2 + 2d^2(\beta,\beta_i),$$

and it follows that $N(\epsilon/C, \mathcal{B}_\delta, d) \le N(\epsilon/2C, \mathcal{B}_\delta^\theta, d_\theta)\cdot N(\epsilon/2C, \mathcal{B}_\delta^\beta, d_\beta)$, where $d_\theta^2(\theta_1,\theta_2) = \mathbb{E}\Delta[(\theta_1-\theta_2)'Z]^2$ and $d_\beta(\beta_1,\beta_2) = d(\beta_1,\beta_2)$. Here $\mathcal{B}_\delta^\theta$ and $\mathcal{B}_\delta^\beta$ are defined as

$$\mathcal{B}_\delta^\theta = \{\theta\in\mathbb{R}^p, d_\theta(\theta,\theta_0)) \le \delta\}, \qquad \mathcal{B}_\delta^\beta = \{\beta\in\mathcal{H}(K), d_\beta(\beta,\beta_0)\le\delta\},$$

with $\mathcal{B}_\delta^\theta \times \mathcal{B}_\delta^\beta \supset \mathcal{B}_\delta$.

It is easy to see that $N(\epsilon/2C, \mathcal{B}_\delta^\theta, d_\theta) = O((\frac{\delta}{\epsilon})^p)$. For $N(\epsilon/2C, \mathcal{B}_\delta^\beta, d_\beta)$, noticing that $\mathcal{H}(K) = L_{K^{1/2}}(L_2) = \{\sum_k b_k L_{K^{1/2}}\phi_k : (b_k)\in l_2\}$, then for any $\beta = \sum_{k\ge1} b_k L_{K^{1/2}}\phi_k \in \mathcal{H}(K)$, we have

$$\begin{aligned}
d^2(\beta,\beta_0) &= \mathbb{E}\Delta\eta^2_{\beta-\beta_0}(X) \\
&= <\beta-\beta_0, L_{C_\Delta}\beta-\beta_0>_{L_2} \\
&= <\sum_{k\ge1}(b_k-b_k^0)L_{K^{1/2}}\phi_k, \sum_{k\ge1}(b_k-b_k^0)L_{C_\Delta K^{1/2}}\phi_k>_{L_2} \\
&= <\sum_{k\ge1}(b_k-b_k^0)\phi_k, \sum_{k\ge1}(b_k-b_k^0)L_{K^{1/2}C_\Delta K^{1/2}}\phi_k>_{L_2} \\
&= <\sum_{k\ge1}(b_k-b_k^0)\phi_k, \sum_{k\ge1}(b_k-b_k^0)s_k\phi_k>_{L_2} \\
&= \sum_{k\ge1} s_k(b_k-b_k^0)^2.
\end{aligned}$$

If we further let $\gamma_k = \sqrt{s_k} b_k$, then $d(\beta, \beta_0) = \sum_{k\geq 1}(\gamma_k - \gamma_k^0)^2$ and $\mathcal{B}_\delta^\beta = \{\beta \in \mathcal{H}(K) : d(\beta, \beta_0)) \leq \delta\}$ can be rewritten as

$$\mathcal{B}_\delta = \{\beta = \sum_{k\geq 1} s_k^{-1/2}\gamma_k L_{K^{1/2}}\phi_k : (s_k^{-1/2}\gamma_k) \in l_2, \sum_{k\geq 1}(\gamma_k - \gamma_k^0)^2 \leq \delta^2\}.$$

Let $M = (\frac{\epsilon}{4C})^{-1/r}$, and

$$\mathcal{B}_\delta^{\beta *} = \{\beta = \sum_{k=1}^{M} s_k^{-1/2}\gamma_k L_{K^{1/2}}\phi_k : (s_k^{-1/2}\gamma_k)_{k=1}^M \in l_2, \sum_{k=1}^{M}(\gamma_k - \gamma_k^0)^2 \leq \delta^2\}.$$

For any $\beta = \sum_{k\geq 1} s_k^{-1/2}\gamma_k L_{K^{1/2}}\phi_k \in \mathcal{B}_\delta$, let $\beta^* = \sum_{k=1}^{M} s_k^{-1/2}\gamma_k L_{K^{1/2}}\phi_k \in \mathcal{B}_\delta^*$. It's easy to see that

$$d^2(\beta, \beta^*) \;=\; \sum_{k>M}\gamma_k^2 = \sum_{k>M} s_k b_k^2 \leq s_M \sum_{k>M} b_k^2 \asymp M^{-2r} = (\frac{\epsilon}{4C})^2,$$

where $\sum_{k>M} b_k^2$ is some small number when M is large, since $(b_k) \in l_2$. So if we can find a set $\{\beta_i^*\}_{i=1}^m \subset \mathcal{B}_\delta^*$ satisfying

$$\min_{1\leq k\leq m} d(\beta^*, \beta_i^*) \leq \epsilon/4C \;\text{ for all } \beta^* \in \mathcal{B}_\delta^*,$$

then it also guarantees that

$$\min_{1\leq k\leq m} d(\beta, \beta_i^*) \leq \min_{1\leq k\leq m}[d(\beta, \beta^*) + d(\beta^*, \beta_i^*)] \lesssim \epsilon/2C \;\text{ for all } \beta \in \mathcal{B}_\delta,$$

i.e.

$$N(\epsilon/2C, \mathcal{B}_\delta^\beta, d_\beta) \lesssim N(\epsilon/4C, \mathcal{B}_\delta^*, d). \tag{3.27}$$

We know that $N(\epsilon/4C, \mathcal{B}_\delta^*, d) \leq (\frac{4\delta+\epsilon/4C}{\epsilon/4C})^M$ is the covering number for a ball in $\mathbb{R}^M$. Therefore combining with (3.26), we have

$$\begin{aligned}
\log \mathcal{N}(\epsilon, \mathcal{M}_{\delta 1}, ||\cdot||_{p,2}) &\leq \log N(\epsilon/C, \mathcal{B}_\delta, d) \\
&\leq \log N(\epsilon/2C, \mathcal{B}_\delta^\theta, d_\theta) + \log N(\epsilon/2C, \mathcal{B}_\delta^\beta, d_\beta) \\
&\leq (\frac{\epsilon}{4C})^{-1/r}\log(\frac{4\delta+\epsilon/4C}{\epsilon/4C}) + \log O((\frac{\delta}{\epsilon})^p) \\
&= O((p + \epsilon^{-1/r})\log(\frac{\delta}{\epsilon})).
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\big|\big|Y(t)\big[\exp(\eta_{\alpha_1}(W)) - \exp(\eta_{\alpha_2}(W))\big]\big|\big|_{p,2}^2 \\
&= P^{TW}\{Y(t)\big[\exp(\eta_{\alpha_1}(W)) - \exp(\eta_{\alpha_2}(W))\big]\}^2 \\
&\leq Cd^2(\alpha_1, \alpha_2), \quad \text{for all } t \in [0, \tau].
\end{aligned}$$

Following the same procedure, we have

$$\log \mathcal{N}(\epsilon, \mathcal{M}_{\delta 2}, ||\cdot||_{p,2}) \leq O((p + \epsilon^{-1/r}) \log(\frac{\delta}{\epsilon})).$$

Now we are able to calculate $J(1, \mathcal{M}_{\delta 1})$,

$$\begin{aligned}
J(1, \mathcal{M}_{\delta 1}) &= \int_0^1 \sqrt{1 + \log \mathcal{N}(\epsilon, \mathcal{M}_{\delta 1}, ||\cdot||_{p,2})} d\epsilon \\
&= \int_0^1 \sqrt{1 + (p + \epsilon^{-1/r}) \log(\frac{\delta}{\epsilon})} d\epsilon \\
&\asymp \int_0^1 \sqrt{\epsilon^{-1/r} \log(\frac{\delta}{\epsilon})} d\epsilon, \\
\text{and for } u = \sqrt{\log(\frac{\delta}{\epsilon})}, &\asymp \int_{\sqrt{\log \delta}}^{\infty} (\frac{\delta}{e^{u^2}})^{-\frac{1}{2r}} u^2 \cdot 2\delta e^{-u^2} du \\
&= O(\delta^{\frac{2r-1}{2r}}) \int_{\sqrt{\log \delta}}^{\infty} (e^{-u^2})^{(1-\frac{1}{2r})} u^2 \cdot du \\
&= O(\delta^{\frac{2r-1}{2r}}), \qquad \text{for } r > \frac{1}{2}.
\end{aligned}$$

The last inequality follows from the fact that the integral above can be seen as the second order moment of a standard normal times some constant, hence it is a constant not depending on δ. Since functions in $\mathcal{M}_{\delta 1}$ are bounded and $J(1, \mathcal{M}_{\delta 1}) = O(\delta^{\frac{2r-1}{2r}})$, Theorem 2.14.1 in [49] implies

$$\big|\big|\,||\mathbb{G}_n||_{\mathcal{M}_{\delta 1}}\big|\big|_{P,2} \lesssim J(1, \mathcal{M}_{\delta 1}) = O(\delta^{\frac{2r-1}{2r}}).$$

Similarly we have

$$\big|\big|\,||\mathbb{G}_n||_{\mathcal{M}_{\delta 2}(t)}\big|\big|_{P,2} = O(\delta^{\frac{2r-1}{2r}}), \quad \text{for all } t \in [0, \tau].$$

■

Lemma 3.5

$$P_{\Delta n}\{s_n(\cdot,\hat{\alpha})[Z]-s_n(\cdot,\alpha_0)[Z]\}-P_{\Delta}\{s(\cdot,\hat{\alpha})[Z]-s(\cdot,\alpha_0)[Z]\}=o_p(n^{-1/2}), \quad (3.28)$$

$$P_{\Delta n}\{s_n(\cdot,\hat{\alpha})[g^*]-s_n(\cdot,\alpha_0)[g^*]\}-P_{\Delta}\{s(\cdot,\hat{\alpha})[g^*]-s(\cdot,\alpha_0)[g^*]\}=o_p(n^{-1/2}). \quad (3.29)$$

Proof We only prove (3.29) as the proof of (3.28) is similar. The right-hand side of (3.29) can be bounded by the sum of the following two terms

$$I_{1n}=\big|(P_{\Delta n}-P_{\Delta})\{s(\cdot,\hat{\alpha})[g^*]-s(\cdot,\alpha_0)[g^*]\}\big|,$$

and

$$I_{2n}=\big|P_{\Delta n}\{s_n(\cdot,\hat{\alpha})[g^*]-s_n(\cdot,\alpha_0)[g^*]-s(\cdot,\hat{\alpha})[g^*]+s(\cdot,\alpha_0)[g^*]\}\big|.$$

We are going to show that $I_{1n}=o_p(n^{-\frac{1}{2}})$ and $I_{2n}=o_p(n^{-\frac{1}{2}})$.

For the first term, since $S_0(\cdot,\hat{\alpha})$, $S_0(\cdot,\alpha_0)$ and $S_1(t,\alpha_0)[g^*]$ are bounded almost surely, we have

$$\begin{aligned}
I_{1n}&=\Big|(P_n-P)\{\Delta[\frac{S_1(\cdot,\hat{\alpha})[g^*]}{S_0(\cdot,\hat{\alpha})}-\frac{S_1(\cdot,\alpha_0)[g^*]}{S_0(\cdot,\alpha_0)}]\}\Big|\\
&=\Big|(P_n-P)\{\Delta[S_0(\cdot,\hat{\alpha})]^{-1}\big[S_1(\cdot,\hat{\alpha})[g^*]-S_1(\cdot,\alpha_0)[g^*]\big]\\
&\qquad+\Delta[S_0(\cdot,\hat{\alpha})S_0(\cdot,\alpha_0)]^{-1}S_1(\cdot,\alpha_0)[g^*][S_0(\cdot,\hat{\alpha})-S_0(\cdot,\alpha_0)]\}\Big|\\
&\lesssim\Big|(P_n-P)\{\Delta\big[S_1(\cdot,\hat{\alpha})[g^*]-S_1(\cdot,\alpha_0)[g^*]\big]\}\Big|\\
&\qquad+\Big|(P_n-P)\{\Delta[S_0(\cdot,\hat{\alpha})-S_0(\cdot,\alpha_0)]\}\Big|.
\end{aligned}$$

Considering $\mathcal{M}_{\delta 3}=\big\{\Delta\big[S_1(t,\alpha)[g^*]-S_1(t,\alpha_0)[g^*]\big],\ \alpha\in\mathcal{B}_{\delta}\big\}$, for any $f_1,f_2\in\mathcal{M}_{\delta 1}$, we have

$$\begin{aligned}
||f_1-f_2||_{p,2}&=\mathbb{E}\Delta^2\{S_1(\cdot,\alpha_1)[g^*]-S_1(\cdot,\alpha_2)[g^*]\}^2\\
&=\mathbb{E}^{\Delta,t,X}\Delta\{\mathbb{E}Y(t)(e^{\eta_{\alpha_1}(W)}-e^{\eta_{\alpha_2}(W)})\eta_{g^*}(X)\}^2\\
&\lesssim d^2(\alpha_1,\alpha_2).
\end{aligned}$$

Following the same proof as Lemma 3.4, we can show that

$$\big|(P_n-P)\{\Delta\big[S_1(\cdot,\hat{\alpha})[g^*]-S_1(\cdot,\alpha_0)[g^*]\big]\}\big|=O(d^{\frac{2r-1}{2r}}(\hat{\alpha},\alpha_0)n^{-\frac{1}{2}})=o_p(n^{-\frac{1}{2}}),$$

given that $d(\hat{\alpha}, \alpha_0) = O_p(n^{-\frac{2r}{2r+1}})$. Similarly,

$$\left|(P_n - P)\{\Delta[S_0(\cdot, \hat{\alpha}) - S_0(\cdot, \alpha_0)]\}\right| = o_p(n^{-\frac{1}{2}}),$$

and altogether we have shown that $I_{1n} = o_p(n^{-\frac{1}{2}})$.

For the second term, the quantity inside the empirical measure $P_{\Delta n}$ is

$$II_{2n}(t) := \frac{S_{1n}(t, \hat{\alpha})[g^*]}{S_{0n}(t, \hat{\alpha})} - \frac{S_{1n}(t, \alpha_0)[g^*]}{S_{0n}(t, \alpha_0)} - \frac{S_1(t, \hat{\alpha})[g^*]}{S_0(t, \hat{\alpha})} + \frac{S_1(t, \alpha_0)[g^*]}{S_0(t, \alpha_0)}.$$

It follows from the same proof as in Lemma A.7 of [34] that

$$\sup_{0 \le t \le 1} |II_{2n}(t)| = o_p(n^{-\frac{1}{2}}).$$

■

Lemma 3.6

$$\begin{aligned}
&P_\Delta\{s(\cdot, \hat{\alpha})[Z - g^*] - s(\cdot, \alpha_0)[Z - g^*]\} \\
&= P_\Delta\{s(\cdot, \alpha_0)[Z - g^*]\}^{\otimes 2}(\hat{\theta} - \theta_0) + O(||\hat{\theta} - \theta_0||^2 + ||\hat{\beta} - \beta)||^2_{C_\Delta}) \\
&= P_\Delta\{s(\cdot, \alpha_0)[Z - g^*]\}^{\otimes 2}(\hat{\theta} - \theta_0) + o_p(n^{-1/2}).
\end{aligned}$$

Proof By lemma 3.1, direct calculation implies

$$\begin{aligned}
&P_\Delta\{s(\cdot, \hat{\alpha})[Z - g^*] - s(\cdot, g_0)[Z - h^*]\} \\
=&P_\Delta\{s(\cdot, \alpha_0)[Z - g^*]s(\cdot, g_0)[\hat{\alpha} - \alpha_0]\} + O(d^2(\hat{\alpha}, \alpha_0)) \\
=&P_\Delta\{s(\cdot, \alpha_0)[Z - g^*]s(\cdot, g_0)[Z]\}(\hat{\theta} - \theta_0) \\
&+ P_\Delta\{s(\cdot, \alpha_0)[Z - g^*]s(\cdot, g_0)[\eta_{\hat{\beta}} - \eta_{\beta_0}]\} \\
&+ O(d^2(\hat{\alpha}, \alpha_0)),
\end{aligned}$$

while by (3.11) , (3.12) and (3.10), we have

$$\begin{aligned}
&P_\Delta\{s(\cdot, \alpha_0)[Z - g^*]s(\cdot, g_0)[\eta_{\hat{\beta}} - \eta_{\beta_0}]\} \\
&= P_\Delta[Z - \eta_{g^*}(X) - \frac{S_1(t, \alpha_0)[Z - g^*]}{S_0(t, \alpha_0)}][\eta_{\hat{\beta}} - \eta_{\beta_0} - \frac{S_1(t, \alpha_0)[\hat{\beta} - \beta_0]}{S_0(t, \alpha_0)}] \\
&= P_\Delta\{Z - \eta_{g^*}(X) - \mathbb{E}[Z - \eta_{g^*}(X)|T, \Delta = 1]\}\{\eta_{\hat{\beta}-\beta_0}(X) - \mathbb{E}[\eta_{\hat{\beta}-\beta_0}(X)|T, \Delta = 1]\} \\
&= P_\Delta[Z - \eta_{g^*}(X) - a^*(T)][\eta_{\hat{\beta}-\beta_0}(X) - a(T)] \\
&= 0,
\end{aligned}$$

and

$$P_\Delta\{s(\cdot,\alpha_0)[Z-g^*]s(\cdot,g_0)[Z]\} = P_\Delta\{s(\cdot,\alpha_0)[Z-g^*]\}^{\otimes 2}.$$

The lemma now follows from from Theorem 3.2.2 which asserts that $d^2(\hat{\alpha},\alpha_0) = o_p(n^{-1/2})$. ■

4. SIMULTANEOUS MODEL SELECTION AND ESTIMATION WITH GSCAD

4.1 Simultaneously Model and Knots Selection in Function-on-scalar Regression

4.1.1 Function-on-scalar regression model

Functional imaging data are common in various medical and biomedical fields, where massive imaging data can be observed over both time and space. Such imaging techniques include functional magnetic resonance imaging (fMRI), electroencephalography (EEG), diffusion tensor imaging (DTI) among many other imaging techniques. Along with the imaging data, scalar predictors such as age, gender, or even gene expression information are recorded to explore their potential effects on the functional response. Therefore, regression models with functional responses and scalar predictors are routinely encountered in practice. A nature model to address such problem is the function-on-scalar regression, stated as

$$Y(t) = \sum_{j=1}^{p} X_j \beta_j(t) + \epsilon(t) \quad t \in \mathcal{T} \tag{4.1}$$

where $Y(t)$ is the functional response on domain $\mathcal{T}$, $X_1, ..., X_p$ are a large number of scalar predictors. $(\mathbf{x}_i, \mathbf{y}_i),\ i = 1, ..., n$ are n observations, with $\mathbf{x}_i = (x_{i1}, ..., x_{ip})^T \in \mathbb{R}^p$ being a vector of scalar predictors and $\mathbf{y}_i = (y_i(t_1), ..., y_i(t_M))$, being real-valued realization of function $Y(t)$ at points $t_m \in \mathcal{T},\ m = 1, ..., M$. $\epsilon(t)$ is the error function with $\epsilon_i(t_m) \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$,$i = 1, ..., n$ and $m = 1, ..., M$. We can also take into account the within-function covariance by adding a certain structure to the covariance of $\epsilon(t)$ [52].

4.1.2 Model selection in Function-on-scalar regression

Like many regression models with high dimensional predictors, function-on-scalar regression faces the model selection challenge of how to identify the important predictors among a potentially large collection. Standard solution to deal with model selection problem is to add a penalty function to the objective function. In case when $\beta(t)$ is assumed to be in an reproducing kernel Hilbert space $\mathcal{H}(K)$, we consider the penalty corresponding to the norm of $\mathcal{H}(K)$, i.e. $||\beta(t)||_K$.

The objective function is stated as

$$\min_{\beta\in\mathcal{H}(K)} \frac{1}{2}\sum_{i=1}^{n}\sum_{m=1}^{M}(y_i(t_m) - \sum_{j=1}^{p} x_{ij}\beta_j(t_m))^2 + \lambda\sum_{j=1}^{p}||\beta_j(t)||_K. \tag{4.2}$$

Based on the properties of RKHS, we have the following representative theory.

Theorem 4.1.1 *The solution of* 4.2 *is in form of*

$$\hat{\beta}_j(t) = \sum_{m=1}^{M} b_{jm}K(t_m, t), \quad j = 1, ..., p,$$

where $\mathbf{b}_j = (b_{j1}, ..., b_{jM})^T \in \mathbb{R}^M$.

Theorem 4.1.1 allows us to reformat the problem. Let $\mathbf{y} = (\mathbf{y}_1^T, ..., \mathbf{y}_n^T)^T \in \mathbb{R}^{(mn)}$ be the vectorized observation of the functional response, $\mathbf{b} = (\mathbf{b}_1^T, ..., \mathbf{b}_p^T)^T \in \mathbb{R}^{(mp)}$ be the vectorized coefficient for $\beta(t)$, and $\tilde{K} = \{K(t_i, t_j)\}_{i,j=1,...,M}$ be a M by M matrix realization of kernel K at points $t_1, ..., t_M \in \mathcal{T}$. Then (4.2) can be rewrite as

$$\min_{\mathbf{b}} \frac{1}{2}||\mathbf{y} - (X \otimes \tilde{K})\mathbf{b}||^2 + \lambda\sum_{j=1}^{p}||\mathbf{b}_j||_{\tilde{K}},$$

where $\otimes$ is the Kronecker product and $||\mathbf{b}_j||_{\tilde{K}} = (b_j^T\tilde{K}b_j)^{1/2}$. Since $K(\cdot,\cdot)$ is the reproducing kernel, matrix $\tilde{K}$ is symmetric and positive definite. Write the spectral decomposition of $\tilde{K}$ as $\tilde{K} = \sum_{l=1}^{M}\tilde{\rho}_l\tilde{\phi}_l^T\tilde{\phi}_l$, where $(\tilde{\rho}_l, \tilde{\phi}_l), l = 1, ..., M$ are pairs of eigenvalue and eigenvector. Define $\tilde{K}^{1/2}$ as

$$\tilde{K}^{1/2} = \sum_{l=1}^{M}\tilde{\rho}_l^{1/2}\tilde{\phi}_l^T\tilde{\phi}_l.$$

Therefore $\tilde{K}^{1/2}$ is also symmetric and positive definite, and satisfies $\tilde{K} = \tilde{K}^{1/2}\tilde{K}^{1/2}$. Make an transformation of $\mathbf{b}_j$ as

$$\alpha_j = \tilde{K}^{1/2}\mathbf{b}_j,$$

and let $\alpha = (\alpha_1^T, ..., \alpha_p^T)^T$. Denoting $\mathbf{D} = X \otimes \tilde{K}^{1/2}$, the objective function 4.2 can be further simplified as

$$\min \frac{1}{2}||\mathbf{y} - \mathbf{D}\alpha||^2 + \lambda \sum_{j=1}^{p} ||\alpha_j||_2,$$

where $||\cdot||_2$ is the $\mathcal{L}_2$ norm.

4.1.3 Knots selection in Function-on-scalar regressions

Unlike traditional models, functional-on-scalar regress, and more generally, functional response models, face an additional challenge of knots selection. The urge of knots selection comes from many aspects. For example, some functional responses are spacial inhomogeneous with different smoothness level over it domain, like the Doppler Curve shown in Figure 4.1 (left). In medical field, signals like EEG or ECG, typically exhibit periodic sharp spikes between waves, see Figure 4.1 (right). Knots selection technique can characterize such inhomogeneity by selecting more knots in areas with dramatic changes, say around the spike in the ECG plot, while keep less knots for smoother areas like the right side of the Doppler curve.

Besides, knots selection can lead to better interpretations of models. In the case of function-on-scalar regression model in (4.1), a proper knots selection in coefficient function $\beta(t)$ can help us understand how each predictor affects the functional response, and which specific region of the response, a predictor has the most effect on.

One way to proceed knots selection is to represent the functional data by a set of base functions with small supports, such as B-splines, and only select a small number of corresponding coefficients to be non-zero. In the setting of function-on-scalar model

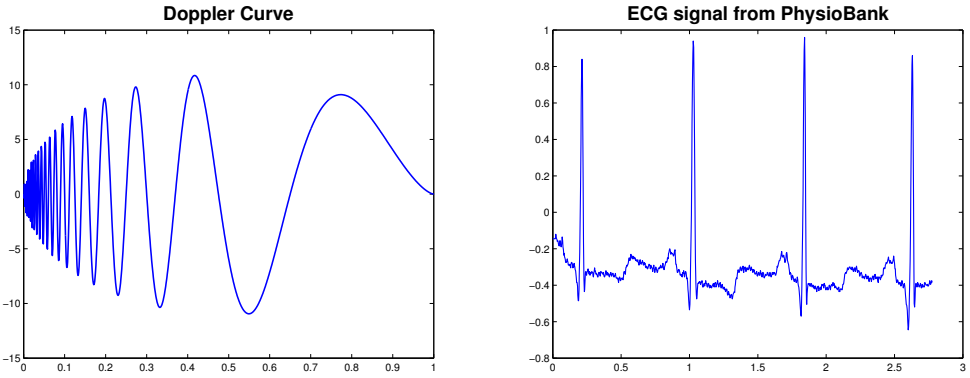


Figure 4.1. Example of spacial inhomogeneous.

(4.1), suppose $\{\phi_1(t), ..., \phi_K(t)\}$ is a set of such basis. The coefficient functions $\beta_j(t)$, $j = 1, ..., p$ can be expanded as

$$\beta_j(t) = \sum_{k=1}^{K} b_{jk}\phi_k(t).$$

Hence, selecting knots for β_j is equivalent to obtaining a sparse estimation of $\mathbf{b}_j = (b_{j1}, ..., b_{jk})$.

Taking into account the need of model selection in function-on-scalar, we will need a penalty function $p(\cdot)$ that could, (1) bring down some of the entire vectors $\mathbf{b}_j$ to zero to produce a zero coefficient function $\beta_j(t) = 0$ and thus select the proper predictor X_js; (2) bring down only part of the elements b_{jk}s for $\beta_j(t)$ corresponding to the important predictor X_j, to do knots selection and furthermore, to show which region on $\mathcal{T}$, predictor X_j has an effect on. Under such situation, Grouped Smoothly Clipped Absolute Deviation(GSCAD) is developed to meet the need of simultaneously selecting model and knots. In fact, GSCAD goes beyond function-on-scalar model and can be applied to the more general problem setting of dictionary learning.

4.2 GSCAD Penalty

4.2.1 Review of the Smoothly Clipped Absolute Deviation (SCAD) penalty.

SCAD penalty is first proposed by [53] in the context of high dimensional linear regression. SCAD has some desired properties: (i) Unbiasedness: the resulting estimator is nearly unbiased when the true unknown parameter is large; (ii) Sparsity: The resulting estimator is able to sets small estimated coefficients to zero to reduce model complexity; (iii) Continuity: The resulting estimator is continuous in data to avoid instability in model prediction. Defined as

$$\psi_\lambda(d) = \begin{cases} \lambda|d|, & if\ |d| \le \lambda \\ -\frac{|d|^2 - 2c\lambda|d| + \lambda^2}{2(c-1)}, & if\ \lambda < |d| \le c\lambda\ , \\ \frac{(c+1)\lambda^2}{2}, & if\ |d| > c\lambda \end{cases} \tag{4.3}$$

for some $\lambda > 0$ and $c > 2$, the SCAD contains three segments. When d is small (less than λ), it acts exactly like the Lasso penalty; when d is big (greater than 3λ), it becomes a constant so that no extra penalty is applied to truly significant parameters; these two segments are connected by a quadratic function which results in a continuous differentiable SCAD penalty function $\psi_\lambda(\cdot)$.

4.2.2 GSCAD penalty

Even though the SCAD penalty possesses many good properties, it only treats parameters individually and does not address any group effect among parameters. With respect to the structure of the dictionary, we propose a new penalty, GSCAD, where G stands for group. Let θ be a vector in $\mathbb{R}^m$. The GSCAD penalty is defined as

$$\Psi_\lambda(\theta) = \log\{1 + \sum_{k=1}^{m} \psi_\lambda(\theta_k)\},$$

where ψ_λ is the SCAD penalty defined in (4.3). It inherits all three merits of SCAD, unbiasedness, sparsity and continuity, and at the same time takes into account both individual parameters and group effect among parameters. Individually, the GSCAD penalty tends to set small estimated θ_k to zero. Group-wise, if all elements in θ are small, the penalty will penalize the entire vector θ to zero. In addition, if some of the θ_k is significantly large, the penalty will have more tolerance of smaller elements appearing in θ.

To better understand GSCAD, let us consider a penalized least squares problem with an orthogonal design

$$\min_\theta \frac{1}{2}\|z - \theta\|_2^2 + p_\lambda(|\theta|),$$

where z and θ are vectors in $\mathbb{R}^m$. For GSCAD, SCAD and LASSO, the penalty $p_\lambda(|\theta|)$ is, respectively,

$$p_\lambda(|\theta|) = \log\{1 + \sum_{k=1}^{m} \psi_\lambda(\theta_k)\}, \quad p_\lambda(|\theta|) = \sum_{k=1}^{m} \psi_\lambda(\theta_k), \quad p_\lambda(|\theta|) = \sum_{k=1}^{m} |\theta_k|.$$

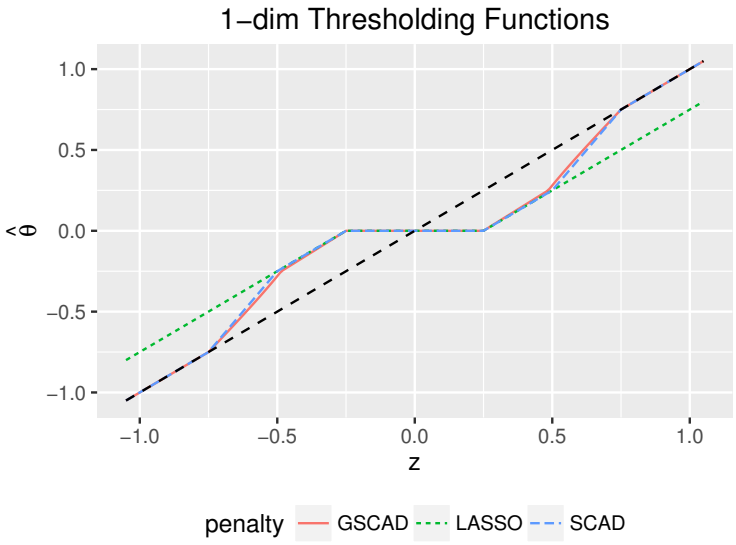


Figure 4.2. 1-dim threshold function.

Estimators of θ when $m = 1$ are shown in Figure 4.2, where GSCAD performs very similar to SCAD. All three penalties shows sparsity properties since they all set $\hat{\theta}$ to zero when $|z| \leq \lambda$. While the soft-thresholding from LASSO has the inherent bias issue, SCAD and GSCAD give $\hat{\theta} = z$ when $|z| \geq c\lambda$ and and avoid bias. In a two-dimensional case when $m = 2$ and $z = (z_1, z_2)$, we investigate partitions of the space according to the number of non-zero element in the resulting estimator $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2)$, see Figure 4.3. While SCAD and Lasso treat each coordinate individually, GSCAD takes into account the whole group. It is less likely to set the estimator of one coordinate to zero as the estimator of another coordinate gets away from zero.

Convexity. Even though GSCAD is built upon the non-convex penalty function SCAD, our development uncovers a surprising fact that the optimization problem of GSCAD under orthogonal design is a convex problem. This will greatly facilitates the implementation of GSCAD.

Theorem 4.2.1 *Define* $\hat{\theta} = (\hat{\theta}_1, ..., \hat{\theta}_m)$ *as the minima of optimization problem*

$$\min_{\theta \in \mathbb{R}^m} \frac{\varrho}{2} \sum_{k=1}^{m} (z_k - \theta_k)^2 + \log\{1 + \sum_{k=1}^{m} \psi_\lambda(\theta_k)\}, \quad \textit{with constant } \varrho > 0. \tag{4.4}$$

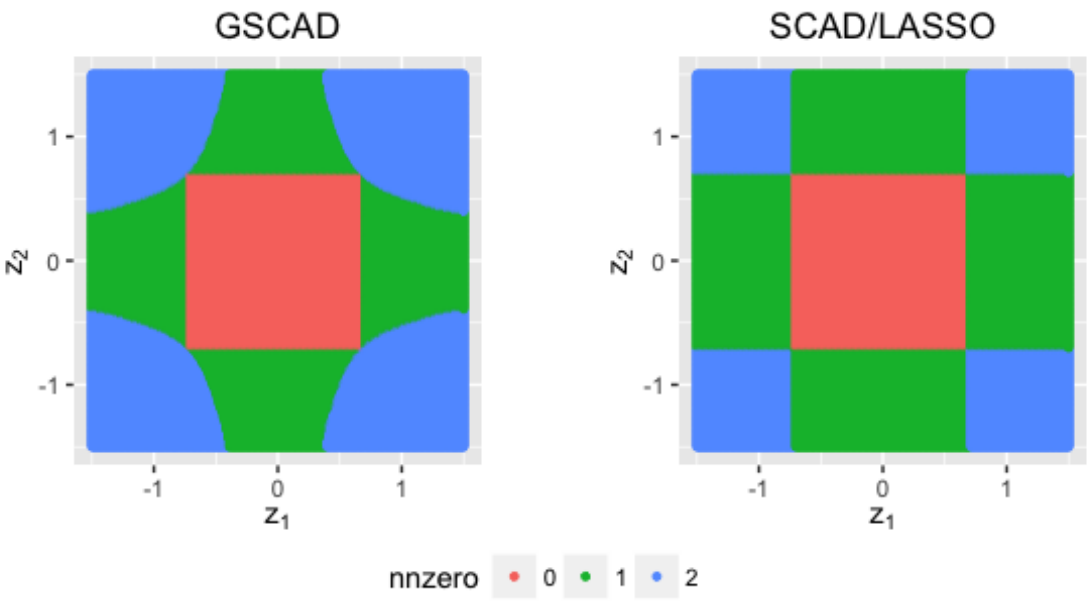


Figure 4.3. Partitions of the 2-dim space $(z_1, z_2) \in \mathbb{R}^2$ according to the number of nonzero elements in $\hat{\theta}$.

Then,

(1) $sign(\hat{\theta}_k) = sign(z_k)$, and $|\hat{\theta}_k| \leq |z_k|$. Denote $\tilde{K} = \{1 \leq k \leq K : \ z_k \neq 0\}$, and let Θ_k be the open interval between z_k and 0. Then problem (4.4) is equivalent to

$$\min_{\theta_k \in \Theta_k \cup \{0\}, k \in \tilde{K}} \frac{\varrho}{2} \sum_{k \in \tilde{K}} (z_k - \theta_k)^2 + \log\{1 + \sum_{k \in \tilde{K}} \psi_\lambda(\theta_k)\} \tag{4.5}$$

(2) Let $c_0 = card(\tilde{K})$, be the number of non-zero element in z. If

$$\lambda^2 \leq \varrho c_0^{-1} \quad and \quad (c-1)\{\varrho(1+\lambda^2)^2 - c_0\lambda^2\} \geq 1 + \lambda^2, \tag{4.6}$$

then optimization problem (4.5) is convex, and $\hat{\theta}$ is continuous in data z.

Remarks on Theorem 4.2.1. (i) Adding a constant ϱ in (4.4) makes the problem more general such that the convexity result can be directly applied to the algorithms in Section 4.3.3, where ϱ plays a role of penalty parameter in the Augmented Lagrangian method. (ii) Condition (4.6) can be satisfied easily under a wide range of circumstances. For instance, in the previous two-dimensional example with $\varrho = 1$, $c_0 = 2$, and $c = 3$, Condition (4.6) will be satisfied as long as $\lambda \leq 2^{-1/2}$.

4.3 Dictionary Learning with GSCAD

4.3.1 Introduction to Dictionary Learning

Sparse coding, which represents signals as sparse linear combinations of basis in a dictionary, has been successfully applied to many signal processing tasks, such as image restoration [54, 55], image classification [56, 57], to name a few. The dictionary is crucial to the success of sparse representation. Most of the compressive sensing literatures take off-the-shelf bases such as wavelets as the dictionary [58, 59]. In contrast, dictionary learning assumes that a signal can be sparsely represented by a learned and usually over-completed dictionary. The pre-specified dictionary might be universal but will not be effective enough for specific tasks such as face recognition

[60, 61]. Instead, using the learned dictionary has recently led to state-of-the-art results in many practical applications, such as image denoising [54, 62–64], image inpainting [65–67], and image compression [68].

Determining a proper size for the to-be-learned dictionary is crucial for both precision and efficiency of the process. However, there is not much existing work discussing the selection of the dictionary size while most algorithms fix the number of atoms in the dictionary. In general, a two-stage procedure may be used to infer the dictionary size, namely first learning a dictionary with a fixed size and then defining a new objective function penalizing the model complexity [69]. The Bayesian technique can be also employed by putting a prior on the dictionary size [70].

Our work is to introduce the novel regularization method GSCAD to Dictionary Learning, and propose an algorithm that could learn a sparse dictionary and select the appropriate dictionary size simultaneously. The algorithm is based on the alternative direction method of multipliers (ADMM) [71]. There are several merits of our approach. First, it imposes sparsity-enforcing constraints on the learned atoms, which improves interpretability of the results and achieves variable selection in the input space. Second, this is a one-stage procedure to learn a sparse dictionary and the dictionary size jointly. Third, the convexity property of GSCAD allow us to decompose the joint non-convex problem with the non-convex penalty into two convex optimization problems, both of which can be solved easily and efficiently. Besides, compared with other state-of-the-art dictionary learning methods, GSCAD has better or competitive performance in image denoising and inpainting.

4.3.2 Matrix Factorization Framework

Dictionary learning problems are commonly specified under the framework of matrix factorization. Consider a vectorized clean signal $\mathbf{x} \in \mathbb{R}^m$ and a dictionary $\mathbf{D} = (\mathbf{d}_1, ..., \mathbf{d}_p) \in \mathbb{R}^{m \times p}$, with its p columns referred to as atoms. Sparse representa-

tion theory assumes that signal $\mathbf{x}$ can be well approximated by a linear combination of a few atoms in $\mathbf{D}$, i.e.

$$\mathbf{x} \approx \mathbf{D}\alpha,$$

where the number of non-zero elements in α is far less than the number of atoms m. In most of the cases, the clean signal $\mathbf{x}$ won't be available, and instead, we will only be able to observe a noisy signal $\mathbf{y} = \mathbf{x} + \epsilon$, where ϵ represents noise with mean zero and variance σ^2. Suppose we have n signals $\mathbf{Y} = (\mathbf{y}_1, ..., \mathbf{y}_n) \in \mathbb{R}^{m \times n}$, and we want to retrieve the corresponding clean signals $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_n)$. This can be summarized as a matrix factorization model

$$\mathbf{Y} = \mathbf{D}\mathbf{A} + \epsilon,$$

where $\mathbf{A} = (\alpha_1, ..., \alpha_m)$. To make the problem identifiable, we require the dictionary $\mathbf{D}$ belongs to a convex set $\mathcal{D}$

$$\mathcal{D} = \{\mathbf{D} \in \mathbb{R}^{m \times p} \; s.t. \; \forall j = 1, ..., p, \; ||\mathbf{d}_j||_\infty \leq 1\}.$$

Dictionary learning aims to obtain estimations of dictionary $\hat{\mathbf{D}}$ and sparse coding $\hat{\mathbf{A}}$, and then reconstruct the clean signal as $\hat{\mathbf{x}} = \hat{\mathbf{D}}\hat{\mathbf{A}}$. This is usually done by minimizing the total squared error:

$$\min ||\mathbf{Y} - \mathbf{D}\mathbf{A}||_F^2, \quad \text{subject to additional sparsity constrains on } \alpha,$$

where $||\cdot||_F$ is the Frobenius norm. Constrains such as $||\alpha||_0 \leq L$ (l_0-penalty) and $||\alpha||_1 \leq \lambda$ (Lasso penalty) for some positive constants L and λ are widely adopted by dictionary learning literature. Experiments have shown that Lasso penalty provides better results when used for learning the dictionary, while l_0 norm should always be used for the final reconstruction step [72].

4.3.3 Simultaneous Sparse Dictionary Learning and Pruning

Compared with sparse coding, regularization on dictionary size is less studied. Most of the existing methods, such as K-SVD and Online Learning, estimate the

dictionary directly with a fixed dictionary size. They usually require the size of the dictionary to be specified before learning, and this will end up with a solution of over completed dictionary with $p > m$, which may not be very helpful if we want to better understand the mechanism. In addition, learning a sparse dictionary can lower the model complexity and improve interpretability of the results. All these issues can be addressed with the help of GSCAD penalty, that could reveal the real size of the dictionary and at the same time obtain an estimated sparse dictionary. More specifically, denote dictionary as $\mathbf{D}$ with p atoms $\mathbf{d}_i = (d_{i1}, \ldots, d_{im})^T \in \mathbb{R}^m, 1 \leq i \leq p$. The GSCAD penalty on dictionary $\mathbf{D}$ is defined by

$$\Psi_\lambda(\mathbf{D}) = \sum_{j=1}^{p} \log\{1 + \sum_{k=1}^{m} \psi_{\lambda_1}(d_{jk})\}$$

where ψ_λ is the SCAD penalty defined in (4.3). The objective function for our problem is formulated as

$$\min_{\mathbf{D}\in\mathcal{D},\alpha_i\in\mathbb{R}^p} \frac{1}{2}\sum_{i=1}^{n} ||\mathbf{y}_i - \mathbf{D}\alpha_i||_2^2 + \Psi_{\lambda_1}(\mathbf{D}) + \lambda_2 \sum_{j=1}^{p} ||\alpha_j||_1. \tag{4.7}$$

Firstly, the GSCAD penalty tends to set small estimated d_{ij} to zero, and reduces the complexity of the estimated dictionary. If all elements in $\mathbf{d}_i$ are small, GSCAD will lead to $\mathbf{d}_i = 0$. Therefore, when starting with a relatively large p, GSCAD will be able to prune the dictionary by penalizing useless atoms to zero. In this way, the true size of the dictionary can be approximated by the number of non-zero columns in the resulting dictionary. In addition, if GSCAD detects some significant d_{ij}s in $\mathbf{d}_i$, it will exert less penalty on the whole $\mathbf{d}_i$ to avoid mistakenly truncating any real signals.

To solve the optimization problem (4.7), we follow the classic iterative two steps approach. Given the dictionary $\mathbf{D}$, we update $\mathbf{A} = (\alpha_1, ..., \alpha_n)$ by solving the Lasso problem,

$$\min_{\alpha_i\in\mathbb{R}^p} \frac{1}{2}||\mathbf{y}_i - \mathbf{D}\alpha_i||_2^2 + \lambda_2||\alpha_i||_1$$

for all signals $1 \leq i \leq n$. Given $\mathbf{A}$, the optimization problem (4.7) becomes

$$\arg\min_{\mathbf{D}\in\mathcal{C}} \frac{1}{2}\sum_{i=1}^{n} ||\mathbf{y}_i - \mathbf{D}\alpha_i||_2^2 + \Psi_{\lambda_1}(\mathbf{D}), \tag{4.8}$$

which is addressed by the ADMM algorithm. Once $\mathbf{D}$ is updated, we remove all zero columns of $\mathbf{D}$ and reset p to the number of current atoms. Algorithm 1 demonstrates this whole procedure. It should be noted that (4.8) is a non-convex problem. Recently, the global convergence of ADMM in non-convex optimization is discussed in [73], which shows that several ADMM algorithms including SCAD are guaranteed to converge.

Problem (4.8) is equivalent to

$$\min \frac{1}{2}\sum_{i=1}^{n} ||\mathbf{y}_i - \mathbf{D_1}\alpha_i||_2^2 + \Psi_{\lambda_1}(\mathbf{D_2})$$
$$s.t. \quad \mathbf{D_1} = \mathbf{D_2}.$$

We form the augmented Lagrangian as

$$L_\varrho(\mathbf{D_1}, \mathbf{D_2}, \xi) = \frac{1}{2}\sum_{i=1}^{n} ||\mathbf{y}_i - \mathbf{D_1}\alpha_i||_2^2 + \frac{\varrho}{2}||\mathbf{D_1} - \mathbf{D_2}||_F^2 + \varrho||\xi \circ (\mathbf{D_1} - \mathbf{D_2})||_F + \Psi_{\lambda_1}(\mathbf{D_2}).$$

where $\circ$ is the element-wise multiplication operator of two matrices, and $\xi \in \mathbb{R}^{d\times p}$. The ADMM algorithm consists three steps in each iteration

$$\mathbf{D_1}^{(t+1)} = \arg\min_{\mathbf{D_1}} L_\varrho(\mathbf{D_1}, \mathbf{D_2}^{(t)}, \xi^{(t)}) \tag{4.9}$$
$$\mathbf{D_2}^{(t+1)} = \arg\min_{\mathbf{D_2}} L_\varrho(\mathbf{D_1}^{(t+1)}, \mathbf{D_2}, \xi^{(t)}) \tag{4.10}$$
$$\xi^{(t+1)} = \xi^{(k)} + (\mathbf{D_1}^{(k+1)} - \mathbf{D_2}^{(k+1)}).$$

Problem (4.9) bears an explicit solution

$$\mathbf{D_1}^{(t+1)} \leftarrow \{\mathbf{y}\mathbf{A}^T + \varrho(\mathbf{D_2}^{(t)} - \xi^{(t)})\}(\mathbf{A}\mathbf{A}^T + \varrho I_p)^{-1}. \tag{4.11}$$

$\mathbf{D}_2$ in (4.10) can be solved by column-wise optimization such as

$$\mathbf{d_2}_j^{(t+1)} = \arg\min_{\mathbf{d_2}_j} \frac{\varrho}{2}||\mathbf{d_2}_j - (\mathbf{d_1}_j^{(t+1)} + \xi_j^{(t)}||_2^2 + \log\{1 + \Psi_{\lambda_1}(\mathbf{d_2}_j)\},$$

for $1 \le j \le p$. In theorem 4.2.1, we have shown that this is a convex problem under Condition (4.6), and can be solved easily by exiting convex optimization algorithms. The ADMM algorithm for updating dictionaries is summarized in Algorithm 2.

Algorithm 1: Dictionary Learning with GSCAD

Input : Training samples $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_n]$, parameter λ_1,λ_2,c,m,p_0

initialize $\mathbf{D}^{(0)} \in \mathbb{R}^{m \times p_0}$;

while *not converge* **do**

Sparse Coding Stage: for $i = 1, ..., n$, update α_i by solving Lasso problem

$$\min_{\alpha_i \in \mathbb{R}^p} \frac{1}{2}||\mathbf{y}_i - \mathbf{D}\alpha_i||_2^2 + \lambda_2||\alpha_i||_1; \tag{4.12}$$

Dictionary Update Stage: update $\mathbf{D}$ using Algorithm 2;

Number of atoms: $p \leftarrow$ # columns of $\mathbf{D}$

end

Output: $\mathbf{D}$, p

Algorithm 2: Update dictionary using ADMM

Input : Training samples $\mathbf{Y}$, current $\mathbf{A} = (\alpha_1, ..., \alpha_n)$, parameter λ_1,c,ϱ

Initialize $\mathbf{D_2}^{(0)} = \xi = \mathbf{0} \in \mathbb{R}^{n \times p}$, set $t = 0$

while *not converge* **do**

$\mathbf{D_1}^{(t+1)} \leftarrow \{\mathbf{y}\mathbf{A}^T + \varrho(\mathbf{D_2}^{(t)} - \xi^{(t)})\}(\mathbf{A}\mathbf{A}^T + \varrho I_r)^{-1}$

Normalize each column of $\mathbf{D_1}$ as $\mathbf{d_1}_j \leftarrow \frac{1}{\max(||\mathbf{d_1}_j||_\infty, 1)} \mathbf{d_1}_j$;

Update $\mathbf{D_2}$: for $1 \leq j \leq p$,

$$\mathbf{d_2}_j^{(t+1)} = \arg\min_{\mathbf{d_2}_j} \frac{\varrho}{2}||\mathbf{d_2}_j - (\mathbf{d_1}_j^{(t+1)} + \xi_j^{(t)})||_2^2 + \log\{1 + \Psi_{\lambda_1}(\mathbf{d_2}_j)\}; \tag{4.13}$$

$\xi^{(t+1)} \leftarrow \xi^{(k)} + (\mathbf{D_1}^{(k+1)} - \mathbf{D_2}^{(k+1)})$;

$t = t + 1$;

end

Remove the zero columns of $\mathbf{D_2}$;

Output: $\mathbf{D_2}$

We define the convergence of the algorithm by the differences of $\mathbf{D}$ and the differences of $\mathbf{A}$ between two consecutive iterations. If they are both below a certain threshold, the algorithm stops. However, in implementation, we add an extra rule on the maximum number of iterations, since GSCAD may get stuck to a region where $\mathbf{D}$ keeps alternating from two local minima and never converge due to a bad initiation. Fortunately, the performance of local minima is mostly decent in terms of denoising. During the dictionary updating stage after we obtain a new dictionary from ADMM, if any two atoms are highly correlated, correlation greater than 0.95 for example, we only keep one of them. Some experiments have shown that this does not have much effect on the results, but will speed up convergence of the algorithm.

4.4 Synthetic Experiments

We design a simple example to check the performance of GSCAD from two aspects: (i) whether GSCAD could recover the true size of the dictionary, and (ii) its denoising performance compared with other methods.

Data is generate from dictionary $\mathbf{D}_0 \in \mathbb{R}^{10\times 100}$, which contains 10 atoms. Each atom is a vectorized 10×10 patch shown in Figure 4.4. Then 1500 signals $\{\mathbf{y}_i\}_{i=1}^{1500}$ in $\mathbb{R}^{100}$ are generated, each created by a linear combination of three different generating dictionary atoms picked randomly, with identically independently distributed coefficients following $Unif(0, 1/3)$. Gaussian noises $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ are added, with signal-to-noise ratio (SNR) controlled by the Gaussian variance σ^2. Four levels of noise $\sigma \in \{5, 10, 20, 50\}$ are adopted for pixel values in the range [0,255].

In order to examine GSCAD's ability to prune dictionaries to the right size, dictionaries are initialized with varying number of atoms p_0, namely, 10 (true size), 15, 20 and 50. Each setting is repeated 1000 times, and each time a dictionary $\hat{\mathbf{D}} \in \mathbb{R}^{m\times\hat{p}}$ and its proper size $\hat{p}(\leq p_0)$ are the learned. Table 4.1 summarizes the size of the learned dictionary. It can be seen that when noise level is small to mod-

$p_0 \backslash \sigma$	5	10	20	50
10	9.98(0.14)	9.98(0.14)	9.99(0.10)	10(0)
15	10.45(0.59)	10.7(0.66)	11.3(0.80)	13.92(0.85)
20	10.71(0.77)	11.1(0.77)	11.74(0.85)	15.92(1.19)
50	11.29(0.10)	11.55(1.31	11.99(1.39)	19.77(2.21)

Table 4.1.

Average number of atoms in the resulting dictionary. Numbers in the parenthesis are corresponding standard deviations.

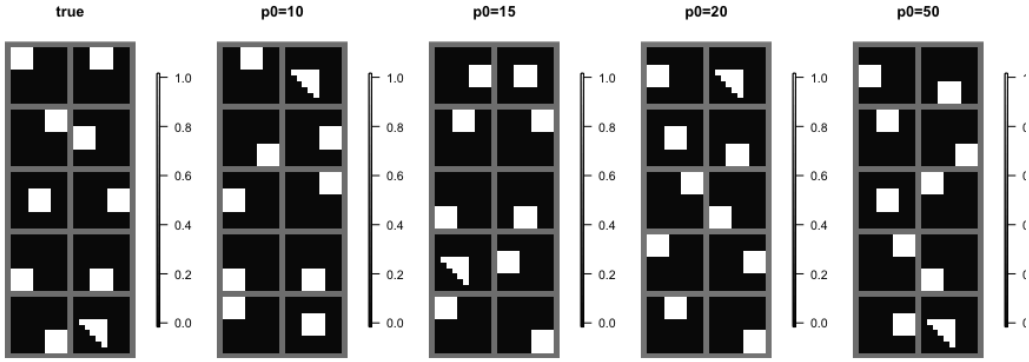


Figure 4.4. From left to right, (1) the generating dictionary $\mathbf{D}_0$ (2)-(5) learned dictionaries using clean data under initialization size $p_0 = 10, 15, 20, 50$. Each atom corresponds to a 10×10 patch with white region representing 1 and black region representing 0.

erate, GSCAD algorithm is able to recover the true size of the dictionary, and its performance is stable across initial dictionaries with different sizes. The result also indicates that as the noise level gets larger, a larger dictionary is needed to process denoising task. Examples of the learned dictionaries with clean data ($\sigma = 0$) under different initial size p_0 are also shown in Figure 4.4.

For comparison, we also run the K-SVD algorithm using the Matlab Toolbox associated its original paper [64], and Online Learning algorithm [74] using the SPAMS

package. Since neither K-SVD nor Online Learning would prune the dictionary, the learned dictionary $\hat{\mathbf{D}}$ will be the same size as its initial value, i.e. $\hat{p} = p_0$.

Once a dictionary $\hat{\mathbf{D}}$ is learned, we obtain the sparse coding α in two ways,

$$\min_{\alpha_i \in \mathbb{R}^{\hat{p}}} ||\alpha_i||_0 \quad s.t. \; ||\mathbf{y}_i - \hat{\mathbf{D}}\alpha_i||_2^2 \leq \epsilon, \tag{4.14}$$

and

$$\min_{\alpha_i \in \mathbb{R}^{\hat{p}}} ||\mathbf{y}_i - \hat{\mathbf{D}}\alpha_i||_2^2 \quad s.t. \; ||\alpha_i||_0 = L, \tag{4.15}$$

using the Orthogonal Matching Pursuit(OMP) algorithm. ϵ in (4.14) is set heuristically by $\epsilon = \sigma^2 F_m^{-1}(\tau)$ [75], where F_m^{-1} is the inverse cumulative distribution function of the χ-square distribution with $m = 100$ and $\tau = 0.9$. L is set to 3 in (4.15). Then denoised signals are reconstructed as $\hat{\mathbf{x}}_i = \hat{\mathbf{D}}\hat{\alpha}_i$, and PSNR is calculated as

$$\text{PSNR} = 10 \log_{10}(\frac{255^2}{\text{MSE}}),$$

where MSE denotes the mean squared-error for images whose intensities are between 0 and 255.

Average PSNR over 100 repeats are shown in Figure 4.5(noting that the scale of axis is shifted downwards figures in the last column). Generally, GSCAD performs better than the other two methods across varying initial size p_0 and SNR levels controlled by σ(sigma). When σ is small, advantage of GSCAD is very clear; when sigma reaches 50, all three method gives similar results with GSCAD performs slightly better. Inspecting the result agains initial size p0, we find that the performance of GSCAD is very stable across p0. Online Learning performs reasonable stable when sigma is small, but when sigma goes as big as 50, a bad p0, say p0=50, hurts more compared with GSCAD. In contrary, the performance of K-SVD depends largely on the initial size of the dictionary; when sigma is small, it benefits more from a over-sized initialization, but when sigma is large, an over-sized initialization does more harm to it comparing with GSCAD. Finally, comparing the first row with the second row of Figure 4.5, we can see that PSNRs obtained from (4.14) is smaller than that from (4.15) for GSCAD, which goes along with our intuition that extra information

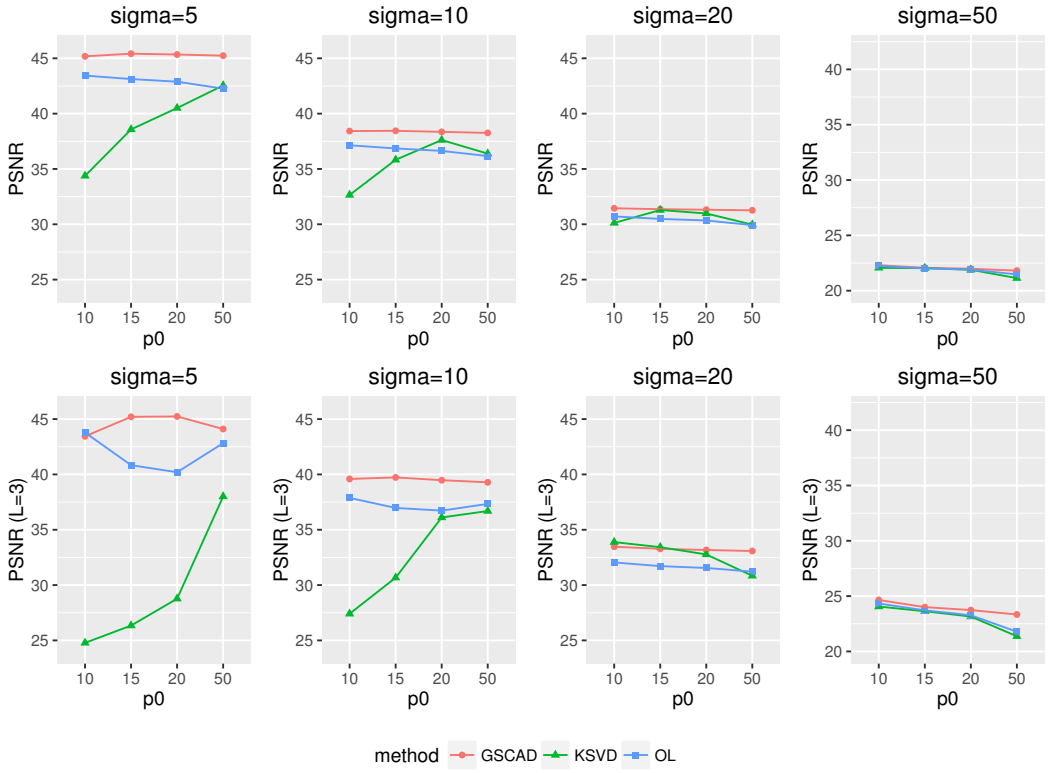


Figure 4.5. Synthetic results. First row, sparse coding is obtained by 4.14. Second row, sparse coding is obtained by 4.15 with $L = 3$.

of $L = 3$ for (4.15) should lead to better results. However, results of the other two methods seem not to follow this intuition. A possible explanation is that to benefit from this extra information of $L = 3$, the learned dictionary needs to be close enough to the truth, and this might not always be the case, especially for K-SVD.

4.5 Image Denoising with GSCAD

To denoise image using GSCAD, we follow the denoising scheme proposed by [54].

1. Split the corrupted image into $\sqrt{m} \times \sqrt{m}$ overlapped patches, which will be treated independently. Let $\mathbf{y}_i \in \mathbb{R}^m$, $i = 1, ..., n$, denote the vectorized small patches.

2. Center $\mathbf{y}_i$ as
$$\mathbf{y}_i^c = \mathbf{y}_i - \mu_i \mathbf{1}_m \quad \text{with } \mu_i = \frac{1}{n}\mathbf{1}_m^T \mathbf{y}_i.$$

3. Train dictionary on the centered $\mathbf{y}_i^c, i = 1, ..., n$, using the proposed Algorithm 1. In the sparse coding step, (4.12) is replaced by its equivalent formula
$$\min_{\alpha_i \in \mathbb{R}^p} \lambda_2 ||\alpha_i||_1, \quad s.t. \ ||\mathbf{y}_i - \mathbf{D}\alpha_i||_2^2 \leq \epsilon, \tag{4.16}$$
with $\epsilon = \sigma^2 F_m^{-1}(\tau)$. Let $\hat{\mathbf{D}}$ denote the learned dictionary.

4. Estimate the final sparse coding $\hat{\alpha}_i$ by (4.14).

5. Add back the mean component to obtain the clean estimate $\hat{\mathbf{x}}_i$:
$$\hat{\mathbf{x}}_i = \hat{\mathbf{D}}\hat{\alpha}_i + \mu_i \mathbf{1}_m.$$

6. Reconstruct the image using the clean estimate $\hat{\mathbf{x}}_i$. Since patches overlap, each pixel belongs to m different patches and admits m estimates. The pixel is thus estimated by the average of its m estimates.

More details about the scheme can be found in [72].

Now, we are ready to compare the denoising performance of GSCAD with K-SVD and Online learning. We follow the same set-up as [72]. Twelve benchmarks images are used in the image denoising, see Figure 4.6. Each image is corrupted with a set of Gaussian noise with its standard deviation σ in $\{5, 10, 15, 20, 25, 50, 100\}$. Patch size m is set to be $\{6^2, 8^2, 10^2, 12^2, 14^2, 16^2\}$ separately. Dictionary size is initialized at $p_0 = 256$ for all three methods. For every noise level, the parameter m is selected such that it maximizes the average PSNR obtained on the last 5 images of the dataset. Then the mean PSNR over all 12 images are reported in Table 4.2. For K-SVD and Online Learning, results are borrowed from [72]. For GSCAD, redundant DCT of size $p = 256$ is used as initialization. For the penalty function, parameter c is set to 3.7 as [53] suggested and λ_1 is picked from $\{0.1, 0.05, 0.01, 0.001\}$. In most cases, a λ_1 of 0.05 would give descent results. The reported PSNR for GSCAD are the averages

σ	5	10	15	20	25	50	100
GSCAD	37.53	33.79	31.75	30.36	29.26	25.81	22.37
Online	37.60	33.90	31.90	30.51	29.43	26.20	22.72
K-SVD	37.42	33.62	31.58	30.18	29.10	25.61	22.10

Table 4.2.
Denoising performance in PSNR

taken over the highest PSNR of each image. Results for all three methods are very close to each other in general, with Online learning performs slightly better, then follows GSCAD.

Figure 4.7 shows how the patch size m affects the denoising result under different noise levels. We can see that when $\sigma = 5$, slitting image into smaller sized patches, like $m = 8\times 8$, works better, and as noise level σ increases, this advantage of smaller m diminishes. We also notice that the fingerprint image reacts differently to the change of patch sizes. When σ is larger than 25, there is a clear pattern of PSNR increasing with m. Besides, the pattern for the flinstones image also deviates slightly from the majority for σ between 15 and 25. This is not surprising as the structure of both images are quite different from all the other nature images. In general, $m = 8 \times 8$ is a decent choice for denoising under all noise levels, and for higher noise level ($\sigma \geq 20$), a patch of size 16×16 can also be considered.

Under patch size $m = 64$, and penalty parameter $\lambda_1 = 0.05$, we examine the dictionary pruning effect of GSCAD. The size of the learned dictionary under different noise levels are plotted in Figure 4.8. It is shown clearly that as noise level increases, a larger-sized dictionary is expected. On the other hand, when the noise level is small, GSCAD gives competitive denoising results with the learned dictionaries only half the sizes of those used by the other two methods.

In the end, we are going to show some denoising examples. Image lena and house are corrupted using Gaussian noise with $\sigma = 25$, see Figure 4.9. We denoise both

(a) house

(b) peppers

(c) Cameraman

(d) lena

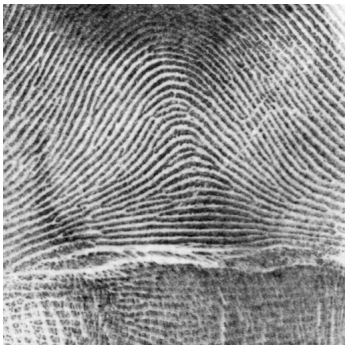
(e) barbara

(f) boat

(g) hill

(h) couple

(i) man



(j) fingerprint

(k) bridge

(l) flinstones

Figure 4.6. Benchmark images for image denoising.

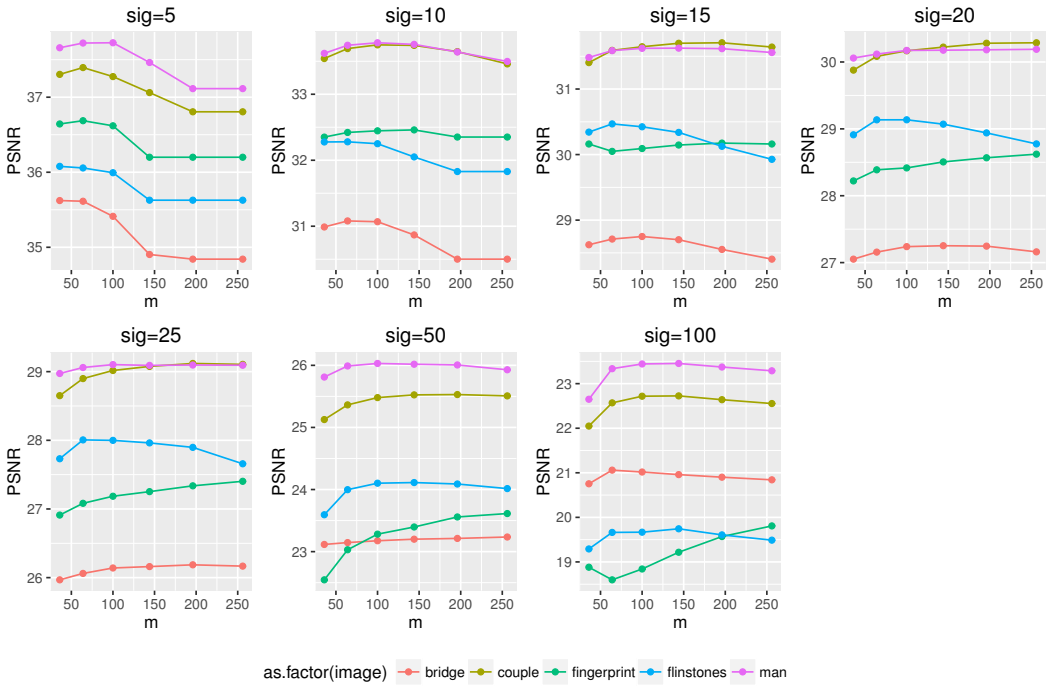


Figure 4.7. Denoising result agains different m

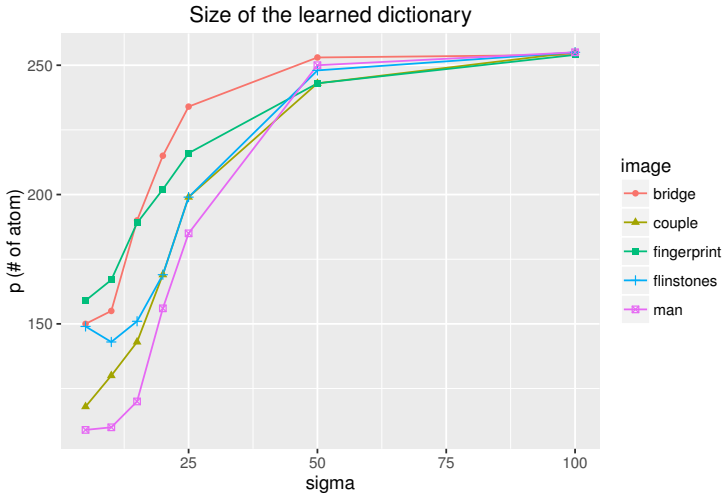


Figure 4.8. Size of the learned dictionary for GSCAD under m=64, $\lambda = 0.05$

lena

house

Figure 4.9. Corrupted Image using Gaussian Noise with $\sigma = 25$

clean

GSCAD(31.28)

Online (31.57)

K-SVD (31.32)

Figure 4.10. Denise lena with patch size $m = 64$, noise level $\sigma = 25$. Numbers in the parenthesis are the resulting PSNR.

clean

GSCAD(31.59)

Online(31.77)

K-SVD(31.41)

Figure 4.11. Denise lena with patch size $m = 256$, noise level $\sigma = 25$. Numbers in the parenthesis are the resulting PSNR.

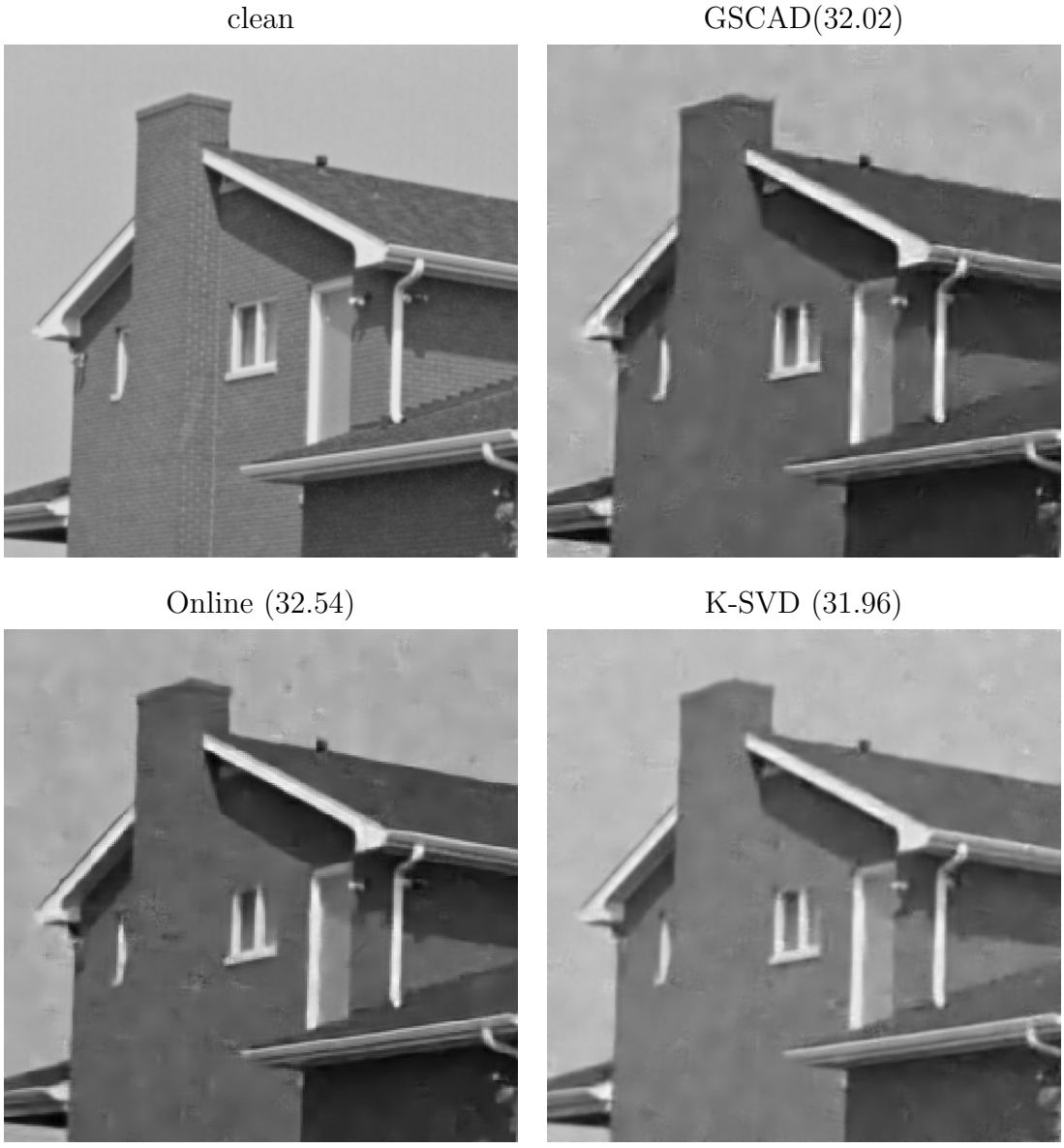


Figure 4.12. Denise house with patch size $m = 64$, noise level $\sigma = 25$. Numbers in the parenthesis are the resulting PSNR.

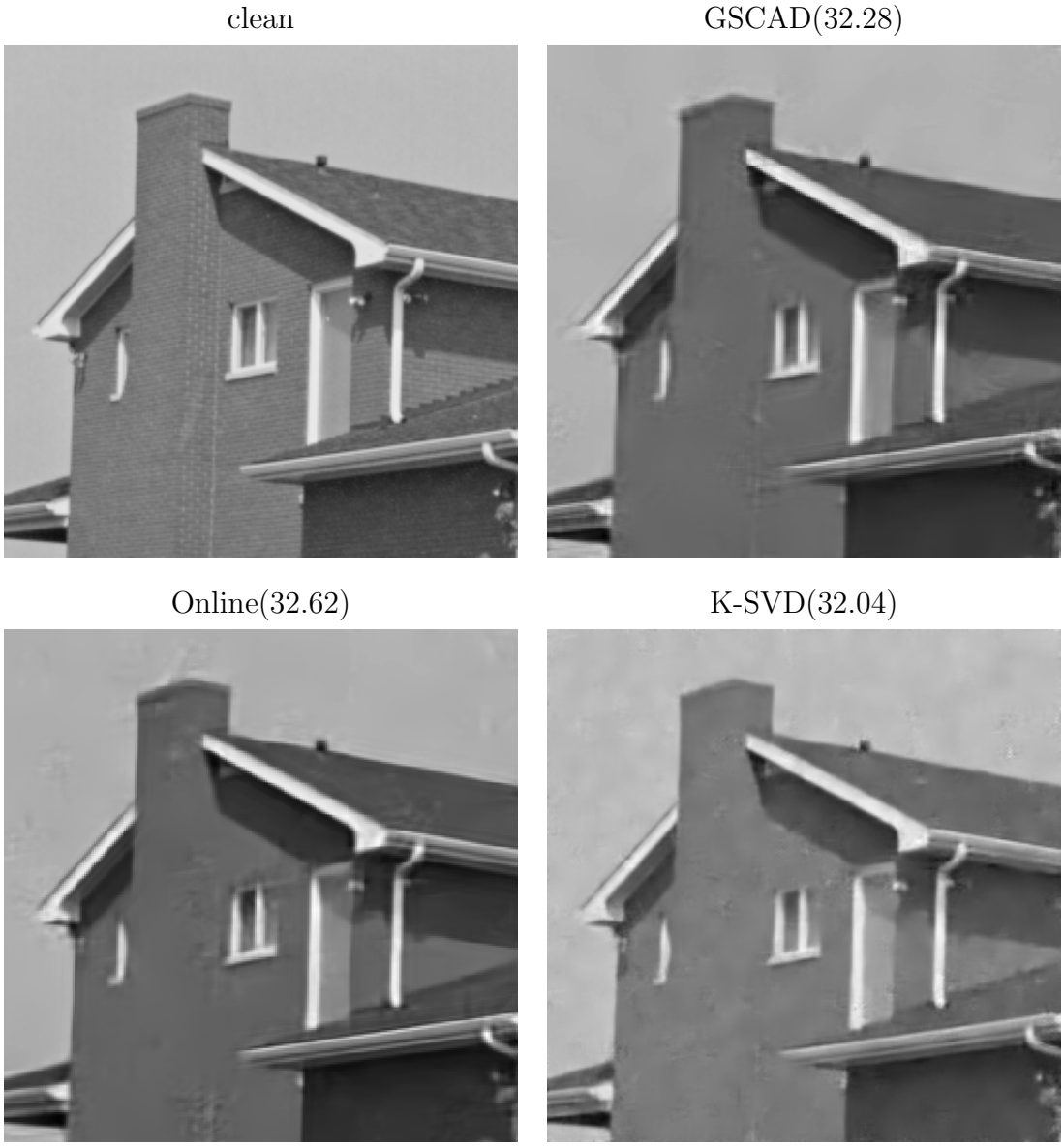


Figure 4.13. Denise house with patch size $m = 256$, noise level $\sigma = 25$. Numbers in the parenthesis are the resulting PSNR.

images using patches of size $m = 8 \times 8$ and $m = 16 \times 16$ respectively. Denoised images obtained using GSCAD, Online Learning and K-SVD are shown in Figure 4.10, Figure 4.11, Figure 4.12 and Figure 4.13.

4.6 Image Inpainting with GSCAD

Image inpainting refers to the task of filling in the missing pixels in a image. When the missing pixels form small holes that are smaller than the patch sizes, the GSCAD algorithm 1 can be easily extended to deal with such unobserved information like many other dictionary learning methods. Define a binary mask $\mathbf{M} \in \mathbb{R}^{m \times m}$ such that

$$\mathbf{M}_{ij} = \begin{cases} 1 & \text{if the } j^{th} \text{ pixel of } \mathbf{y}_i \text{ is observed} \\ 0 & \text{otherwise.} \end{cases}$$

Then the original dictionary learning formulation can be modified as

$$\min_{\mathbf{D} \in \mathcal{D}, \alpha_i \in \mathbb{R}^p} \frac{1}{2} ||\mathbf{M} \circ (\mathbf{y} - \mathbf{D}\alpha)||_F^2 + \Psi_{\lambda_1}(\mathbf{D}) + \lambda_2 \sum_{j=1}^{p} ||\alpha_j||_1. \tag{4.17}$$

Following the previous two steps approach, given the dictionary $\mathbf{D}$, we update $\mathbf{A} = (\alpha_1, ..., \alpha)$ by solving the masked Lasso problem,

$$\min_{\alpha_i \in \mathbb{R}^p} \frac{1}{2} ||\mathbf{M}_{\cdot i} \circ (\mathbf{y}_i - \mathbf{D}\alpha_i)||_2^2 + \lambda_2 ||\alpha_i||_1$$

for all signals $1 \leq i \leq n$. And given $\mathbf{A}$, the optimization problem (4.17) becomes

$$\arg\min_{\mathbf{D} \in \mathcal{C}} ||\mathbf{M} \circ (\mathbf{y} - \mathbf{D}\alpha)||_F^2 + \Psi_{\lambda_1}(\mathbf{D}), \tag{4.18}$$

which can still be addressed by the ADMM algorithm with a slightly modification for updating $\mathbf{D}_1$. Now that mask $\mathbf{M}$ is involved in our ADMM, $\mathbf{D}_1$ needs to be updated one row at a time. Let $\mathbf{M}_{j\cdot}$ denote the j^{th} row of mask $\mathbf{M}$, and the rows of other matrix defined in the same fashion. For sample $\mathbf{y}_i$'s, let $\mathbf{y} = (\mathbf{y}_1, ..., \mathbf{y}_n) \in \mathbb{R}^{m \times n}$, so $\mathbf{y}_{j\cdot}$ indicates the j^{th} row of matrix $\mathbf{y}$. For $1 \leq j \leq m$, the j^{th} row of $\mathbf{D}_1$ is updated as

$$\mathbf{D_1}_{j\cdot}^{(t+1)} = \{\mathbf{y}_{j\cdot}\text{diag}(\mathbf{M}_{j\cdot})\mathbf{A}^T + \varrho(\mathbf{D_2}_{j\cdot}^{(t)} - \xi_{j\cdot}^{(t)})\}\{\mathbf{A}\text{diag}(\mathbf{M}_{j\cdot})\mathbf{A}^T + \varrho I_p\}^{-1}.$$

Algorithm 3: Dictionary Learning with GSCAD (Inpainting)

Input : Training samples $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_n]$, mask $\mathbf{M}$, parameter λ_1,λ_2,c,m,p_0

initialize $\mathbf{D}^{(0)} \in \mathbb{R}^{m \times p_0}$

while *not converge* **do**

Sparse Coding Stage: for $i = 1, ..., n$, update α_i by solving the masked Lasso problem

$$\min_{\alpha_i \in \mathbb{R}^p} \frac{1}{2} ||\mathbf{M}_{\cdot i} \circ (\mathbf{y}_i - \mathbf{D}\alpha_i)||_2^2 + \lambda_2 ||\alpha_i||_1; \tag{4.19}$$

Dictionary Update Stage: update $\mathbf{D}$ using the Algorithm 4;

Number of atoms: $p \leftarrow$ # columns of $\mathbf{D}$

end

Output: $\mathbf{D}$, p

Algorithm 4: Update dictionary using ADMM (Inpainting)

Input : Training samples $\mathbf{Y}$, mask $\mathbf{M}$, current $\mathbf{A} = (\alpha_1, ..., \alpha_n)$, parameter λ_1,c,ϱ

Initialize ${\mathbf{D_2}}^{(0)} = \xi = \mathbf{0} \in \mathbb{R}^{n \times p}$, set $t = 0$

while *not converge* **do**

  Update $\mathbf{D}_1$ row by row

$$\mathbf{D_1}_{j\cdot}^{(t+1)} = \{\mathbf{y}_{j\cdot}\text{diag}(\mathbf{M}_{j\cdot})\mathbf{A}^T + \varrho(\mathbf{D_2}_{j\cdot}^{(t)} - \xi_{j\cdot}^{(t)})\}\{\mathbf{A}\text{diag}(\mathbf{M}_{j\cdot})\mathbf{A}^T + \varrho I_p\}^{-1}.$$

  Normalize each column of $\mathbf{D_1}$ as $\mathbf{d_1}_j \leftarrow \frac{1}{\max(||\mathbf{d_1}_j||_\infty, 1)}\,\mathbf{d_1}_j$;

  Update $\mathbf{D_2}$: for $1 \le j \le p$,

$$\mathbf{d_2}_j^{(t+1)} = \arg\min_{\mathbf{d_2}_j} \frac{\varrho}{2}||\mathbf{d_2}_j - (\mathbf{d_1}_j^{(t+1)} + \xi_j^{(t)})||_2^2 + \log\{1 + \Psi_{\lambda_1}(\mathbf{d_2}_j)\}; \quad (4.20)$$

  $\xi^{(t+1)} \leftarrow \xi^{(k)} + ({\mathbf{D_1}}^{(k+1)} - {\mathbf{D_2}}^{(k+1)})$;

  $t = t + 1$;

end

Remove the zero columns of $\mathbf{D_2}$;

Output: $\mathbf{D_2}$

The whole inpainting algorithm is summarized in Algorithm 3 and Algorithm 4.

When inpainting a corrupted image, we can follow a similar scheme as image denoising with a few modification.

1. Split the corrupted image into $\sqrt{m} \times \sqrt{m}$ overlapped patches, which will be treated independently. Let $\mathbf{y}_i \in \mathbb{R}^m$, $i = 1, ..., n$, denote the vectorized small patches.

2. Center $\mathbf{y}_i$ with respect to the missing pixels

$$\mathbf{y}_i^c = \mathbf{y}_i - \mu_i \mathbf{1}_m \quad \text{with } \mu_i = \mathbf{M}_i^T \mathbf{y}_i / \mathbf{M}_i^T \mathbf{1}_m,$$

   where $\mathbf{M}_i$ is the i^{th} column of mask $\mathbf{M}$.

3. Train dictionary on the centered $\mathbf{y}_i^c, i = 1, ..., n$, using the proposed Algorithm 3. In the sparse coding step, (4.12) is replaced by its equivalent formula

$$\min_{\alpha_i \in \mathbb{R}^p} \lambda_2 ||\alpha_i||_1, \quad s.t. \ ||\mathbf{M}_i \circ (\mathbf{y}_i - \mathbf{D}\alpha_i)||_2^2 \leq \epsilon, \tag{4.21}$$

   with ϵ chosen heuristically as $F_m^{-1}(0.9)$. Let $\hat{\mathbf{D}}$ denote the learned dictionary.

4. Estimate the final sparse coding $\hat{\alpha}_i$ by

$$\min_{\alpha_i \in \mathbb{R}^p} \lambda_2 ||\alpha_i||_1, \quad s.t. \ ||\mathbf{M}_i \circ (\mathbf{y}_i - \mathbf{D}\alpha_i)||_2^2 \leq \epsilon.$$

5. Add back the mean component to obtain the clean estimate $\hat{\mathbf{x}}_i$:

$$\hat{\mathbf{x}}_i = \mathbf{M}_i \circ \mathbf{y}_i + (1 - \mathbf{M}_i) \circ (\hat{\mathbf{D}}\hat{\alpha}_i + \mu_i \mathbf{1}_m).$$

6. Reconstruct the image using the clean estimate $\hat{\mathbf{x}}_i$. Since patches overlap, each pixel belongs to m different patches and admits m estimates. The pixel is thus estimated by the average of its m estimates.

Like other inpainting algorithms, when the missing wholes follow a regular pattern, the proposed algorithm may face a possible problem of absorbing this pattern in

Figure 4.14. Image Inpainting. Left: lena with 50% of the data removed. Right: Inpainting result from global learned dictionary using GSCAD.

its dictionary. The common strategy to fix this problem is to first learn a global dictionary $\mathbf{D}_g$ using clean image from a standard image bank. Then take $\mathbf{D}_g$ as an initial dictionary to learn an adaptive dictionary $\mathbf{D}_a$ using patches extracted from the corrupted image. When it comes to the step of recovering the missing pixels, the joint dictionary $\mathbf{D}_g \cup \mathbf{D}_a$ is used.

As we need to update $\mathbf{D}_1$ one row at a time, the inpainting algorithm is slower than the denoising one. However, experiments have shown that using the global learned dictionary $\mathbf{D}_g$ directly to inpaint the corrupted image still gives decent results. Figure 4.14 and Figure 4.15 show some inpainting examples using global learned dictionary. The global dictionary is learned from 240000 natural image patches of size $m = 8 \times 8$ extracted from the Kodak PhotoCD images. Algorithm 1 and the denoising scheme in Section 4.5 are employed with parameters set to $c = 3.7$, $\lambda = 0.05$ and $\epsilon = F_m^{-1}(0.9)$. In Figure 4.14, 50% of the original pixels in image lena are removed randomly, and in Figure 4.15, text with two fonts are added to the original image. The resulting PSNR are 33.84 and 35.21 respectively.

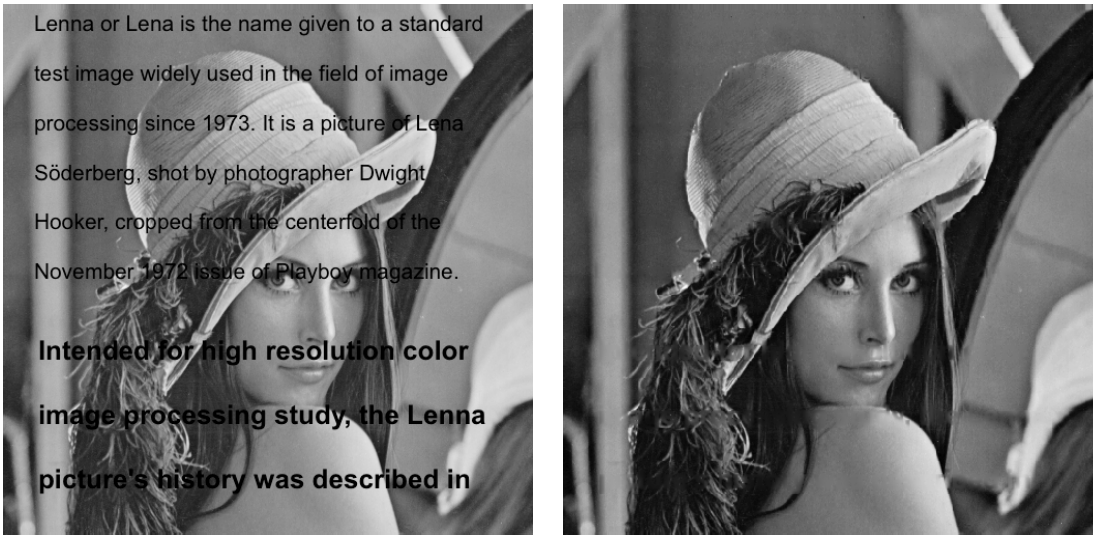


Figure 4.15. Text removal result from global learned dictionary using GSCAD

4.7 The GSCAD Package

R package GSCAD is developed to run image denosing and inpainting task with GSCAD. Major functions include *gsacd.DL*, an implementation of Algorithm 1, and *gsacd.DLmask*, an implementation of Algorithm 3. Schemes of image denoising and inpainting mentioned in Section 4.5 and Section 4.6 can be carried out by function *denoiseImage* and *inpaintImage*. In addition, some basic evaluation functions are also provided, such as function *PSNR* to calculated the PSNR for the processed image and function *plotDic* to visualize a dictionary.

4.8 Discussion

The GSCAD method has been presented to learn a sparse dictionary and select the dictionary size simultaneously. The experimental analysis has demonstrated very encouraging results relative to the state-of-the-art methods. This new framework may also be applied to the general subspace clustering problem for imaging clustering, which assumes that similar points are described as points lying in the same subspace. The proposed formulation can learn the clustering and the number of clusters at the same time. This framework may also be applied to the architecture design of deep learning. The new GSCAD penalty can learn a sparse connection between units of two layers in the deep neural network to improve efficiency.

4.9 Proofs of Theorems

4.9.1 Proof of Theorem 4.1.1.

Let $\hat{\beta}_j(s), j = 1, ..., p$ be the solution of 4.2. Since $\beta_j(s) \in \mathcal{H}(K)$, we can write

$$\hat{\beta}_j(s) = \sum_{l=1}^{\infty} \hat{b}_{jl}\phi_l(s),$$

and

$$||\hat{\beta}_j||_K^2 = \sum_{l=1}^{\infty} \hat{b}_{jl}^2/\rho_l$$

Denote $\hat{\beta}_j(s_m) = z_m,\ m = 1, ..., M$. Then $\hat{b}_{jl}$'s are the solution to

$$\min \sum_{l=1}^{\infty} b_{jr}^2(l)/\rho_l \qquad s.t.\ \sum_{l=1}^{\infty} b_{jl}\phi_l(s_m) = z_m, \text{ for all } m = 1, ..., M.$$

Applying the largrange method, we have

$$L(b_j, \zeta) = \sum_{l=1}^{\infty} b_{jl}^2/\rho_l + \sum_{m=1}^{M} \zeta_m \{\sum_{l=1}^{\infty} b_{jl}\phi_l(s_m) - z_m\}.$$

Taking derivative

$$\frac{\partial L}{\partial b_{jl}} = 2b_{jl}/\rho_l + \sum_{m=1}^{M} \zeta_m \phi_l(s_m) = 0.$$

Therefore

$$\hat{b}_{jl} = -\rho_l \sum_{m=1}^{M} \zeta_m \phi_l(s_m),$$

and

$$\begin{aligned}
\hat{\beta}_j(s) &= \sum_{l=1}^{\infty} -\{\rho_l \sum_{m=1}^{M} \zeta_m \phi_l(s_m)\}\phi_l(s) \\
&= -\sum_{m=1}^{M} \zeta_m \sum_{l=1}^{\infty} \rho_l \phi_l(s_m)\phi_l(s) \\
&= -\sum_{m=1}^{M} \zeta_m K(s_m, s).
\end{aligned}$$

4.9.2 Proof of Theorem 4.2.1.

1. When $z_k = 0$, we have $(z_k - 0)^2 \leq (z_k - \theta_k)^2$, and further

$$\log\{1 + \psi_\lambda(0) + \sum_{l \neq k} \psi_\lambda(\theta_l)\} \leq \log\{1 + \psi_\lambda(\theta_k) + \sum_{l \neq k} \psi_\lambda(\theta_l)\},$$

for any $\theta_k \in \mathbb{R}$. When $z_k \neq 0$, we have

$$\{z_k - sign(z_k)|\theta_k|\}^2 \leq [z_k - \{-sign(z_k)|\theta_k|\}]^2,$$

and further

$$\log\{1+\psi_\lambda(sign(z_k)|\theta_k|)+\sum_{l\neq k}\psi_\lambda(\theta_l)\}=\log\{1+\psi_\lambda(-sign(z_k)|\theta_k|)+\sum_{l\neq k}\psi_\lambda(\theta_l)\}.$$

Therefore to minimize 4.4, $\hat{\theta}_k$ has to satisfy that $sign(\hat{\theta}_k)=sign(z_k)$. If we denote $\tilde{K}=\{1\leq k\leq K:\ z_k\neq 0\}$ and Θ_k as the open interval between z_k and 0, i.e.

$$\Theta_k=\begin{cases}(0,z_k), & if\ z_k>0\\ (z_k,0), & if\ z_k<0\end{cases},$$

then optimization problem (2) is equivalent to

$$\min_{\theta_k\in\Theta_k\cup\{0\},k\in\tilde{K}}\frac{\varrho}{2}\sum_{k\in\tilde{K}}(z_k-\theta_k)^2+\log\{1+\sum_{k\in\tilde{K}}\psi_\lambda(\theta_k)\}.$$

2. To simplify the notation, we rewrite $z=(z_{i_1},...,z_{i_{c_0}})\in\mathbb{R}^{c_0}$ and $\theta=(\theta_{i_1},...,\theta_{i_{c_0}})\in\mathbb{R}^{c_0}$ as with $\tilde{K}=\{i_1,i_2,...,i_{c_0}\}$ and $c_0=card(\tilde{K})$. Define $L:\mathbb{R}^{c_0}\to\mathbb{R}$ as

$$L(\theta)=\frac{\varrho}{2}||z_k-\theta_k||^2+\log\{1+\sum_{k=1}^{c_0}\psi_\lambda(\theta_k)\}.$$

We expend Θ_k to the whole half plane as

$$\tilde{\Theta}_k=\begin{cases}(0,\infty), & if\ z_k>0\\ (-\infty,0), & if\ z_k<0\end{cases}.$$

If we can show that L is convex in $\tilde{\Theta}_1\times...\times\tilde{\Theta}_{c_0}$, this will imply that L is convex over $\prod_{k=1}^{c_0}\Theta_k\cup\{0\}$, as$L$ is continuous all over $\mathbb{R}^{c_0}$.

To show that the optimization problem within $\times^o=\tilde{\Theta}_1\times...\times\tilde{\Theta}_{c_0}$ is convex, we are going to verify the inequality

$$L((1-t)x+ty)\leq(1-t)L(x)+tL(y),\quad t\in[0,1],$$

for any $x,y\in\Theta^o$. This is trivial for $x=y$, and for $x\neq y$, we consider the following cases.

Case 1: $x,y\in\times_1^o=\{x\in\times^o:|x_i|\notin\{\lambda,c\lambda\}$ for any $1\leq i\leq c_0\}$. Therefore only a finite number of points in set $\{tx+(1-t)y:t\in[0,1]\}$ such that L does not have

a second-order derivative. Let $v = x - y$. Define $\varphi(t) = L(x + tv), t \in [0, 1]$. If we can show that $\varphi'(t)$ is continuous on $[0, 1]$, and $\varphi''(t) \geq 0$ except at a finite number of points, therefore $\varphi'(t)$ is non-decreasing. Furthermore $\varphi(t)$ is convex on $[0, 1]$. By definition, for any $t \in [0, 1]$,

$$L((1-t)x + ty) = L(x + tv) = \varphi(t) \leq t\varphi(1) + (1-t)\varphi(0) = tL(y) + (1-t)L(x).$$

Therefore f is convex.

Now we are going to show that $\varphi'(t)$ is continuous and $\varphi''(t) \geq 0$ except at a finite number of points, where $\varphi''(t)$ does not exist. Taking derivative of L, we get

$$\begin{aligned}
L'_{x_i} &= sign(x_i)\{\varrho|x_i| + \frac{\dot{\psi}_\lambda(x_i)}{1 + \sum_k \psi_\lambda(x_k)}\} - \varrho z_k, \\
L''_{x_i x_i} &= \varrho + \frac{\ddot{\psi}_\lambda(x_i)}{1 + \sum_k \psi_\lambda(x_k)} - \frac{\dot{\psi}^2_\lambda(x_i)}{\{1 + \sum_k \psi_\lambda(x_k)\}^2}, \quad |x_i| \notin \{\lambda, c\lambda\}, \\
L''_{x_i x_j} &= -\frac{\dot{\psi}_\lambda(x_i) \cdot \dot{\psi}_\lambda(x_j)}{\{1 + \sum_k \psi_\lambda(x_k)\}^2}, \quad |x_i|, |x_j| \notin \{\lambda, c\lambda\}
\end{aligned}$$

where

$$\dot{\psi}_\lambda(x_i) = \begin{cases} \lambda \cdot sign(x_i), & if\ |x_i| \leq \lambda \\ \frac{c\lambda - |x_i|}{(c-1)} \cdot sign(x_i), & if\ \lambda < |x_i| \leq c\lambda \\ 0, & if\ |x_i| > c\lambda \end{cases} \quad and \quad \ddot{\psi}_\lambda(x_i) = \begin{cases} -\frac{1}{(c-1)}, & if\ \lambda < |x_i| \leq c\lambda \\ 0, & o.w. \end{cases}.$$

Since L'_{x_i} is continuous for all $1 \leq i \leq c_0$ and $x \in \times^o$,

$$\varphi'(t) = \sum_i \frac{\partial L}{\partial x_i}(x + tv) \cdot v_i$$

is continuous. Except a finite number of $t \in [0,1]$, such that $L''_{x_i x_j}$ does not exist at $x + tv$, we have

$$\begin{aligned}
\varphi''(t) &= \sum_{i,j} \frac{\partial^2 L}{\partial x_i \partial x_j}(x+tv) v_i v_j \\
&= \sum_{i=1}^{c_0} \{\varrho + \frac{\ddot{\psi}_\lambda(x_i)}{1+\sum_k \psi_\lambda(x_k)}\} v_i^2 - \{1 + \sum_k \psi_\lambda(x_k)\}^{-2} \{\sum_{i=1}^{c_0} \dot{\psi}_\lambda(x_i) v_i\}^2 \\
&\geq \sum_{i=1}^{c_0} \{\varrho + \frac{\ddot{\psi}_\lambda(x_i)}{1+\sum_k \psi_\lambda(x_k)}\} v_i^2 - \{1 + \sum_k \psi_\lambda(x_k)\}^{-2} c_0 \sum_{i=1}^{c_0} \dot{\psi}_\lambda^2(x_i) v_i^2 \\
&= \sum_{i=1}^{c_0} \{\varrho + \frac{\ddot{\psi}_\lambda(x_i)}{1+\sum_k \psi_\lambda(x_k)} - \frac{c_0 \dot{\psi}_\lambda^2(x_i)}{\{1+\sum_k \psi_\lambda(x_k)\}^2}\} v_i^2.
\end{aligned}$$

Let

$$f_i(x_i) = \varrho + \frac{\ddot{\psi}_\lambda(x_i)}{1+\sum_l \psi_\lambda(b_l)} - \frac{c_0\, \dot{\psi}_\lambda^2(x_i)}{\{1+\sum_l \psi_\lambda(b_l)\}^2}, \quad 1 \leq i \leq c_0.$$

To show that $\varphi''(t) \geq 0$, we only need to show that $f_i(x_i) \geq 0$. Since $f_i(x_i) = f_i(-x_i)$, without loss of generality, we are only going to show that $f_i(x_i) \geq 0$, for $x_i > 0$.

Take derivative of f_i,

$$f_i'(x_i) = -\frac{\dddot{\psi}_\lambda(x_i)\dot{\psi}_\lambda(x_i)}{1+\sum_l \psi_\lambda(x_l)} - \frac{2c_0\, \dot{\psi}_\lambda^2(x_i)\ddot{\psi}_\lambda(x_i)}{\{1+\sum_l \psi_\lambda(x_l)\}^2} + \frac{2c_0\, \dot{\psi}_\lambda^3(x_i)}{\{1+\sum_l \psi_\lambda(x_l)\}^3}, \quad x_i \notin \{\lambda, c\lambda\}.$$

Since $\ddot{\psi}_\lambda(x_i) \leq 0$ and $\dot{\psi}_\lambda(x_i) \geq 0$, we have $f_i'(x_i) \geq 0$ for all $x_i \in \tilde{\Theta}_k \backslash \{\lambda, c\lambda\}$. Observe that $f_i(x_i)$ is continuous on $(0, \infty)$. For $x_i \in (0, \lambda)$,

$$f_i(x_i) \geq \lim_{x_i \to 0^+} f_i(x_i) = \varrho - \frac{c_0 \lambda^2}{\{1+\sum_{l \in \tilde{K}, l \neq k} p_\lambda(x_l)\}^2} \geq \varrho - c_0 \lambda^2 \geq 0.$$

For $x_i \in (\lambda, c\lambda)$

$$\begin{aligned}
f_i(x_i) &\geq \lim_{x_i \to \lambda^+} f_i(x_i) \\
&= \varrho - \frac{1}{(c-1)\{1+\lambda^2+\sum_{l \neq k} \psi_\lambda(x_l)\}} - \frac{c_0 \lambda^2}{\{1+\lambda^2+\sum_{l \neq k} \psi_\lambda(x_l)\}^2} \\
&\geq \varrho - \frac{1}{(c-1)(1+\lambda^2)} - \frac{c_0 \lambda^2}{(1+\lambda^2)^2} \\
&= \frac{\varrho(c-1)(1+\lambda^2)^2 - (1+\lambda^2) - c_0(c-1)\lambda^2}{(c-1)(1+\lambda^2)^2} \\
&\geq 0.
\end{aligned}$$

For $x_i \in (c\lambda, \infty)$,

$$f_i(x_i) \geq \lim_{x_i \to c\lambda^+} f_i(x_i) = \varrho > 0.$$

Therefore $f_i(x_i) \geq 0$, for $x_i > 0$. Furthermore, we show that $\varphi''(t) \geq 0$ except a finite number of $t \in [0,1]$ and finish the proof of case 1.

Case 2: $x \in \times_0^o$ or $y \in \times_0^o$, where $\times_0^o = \times^o \backslash \times_1^o = \{x \in \times^o : |x_i| \in \{\lambda, c\lambda\} \text{ for some } 1 \leq i \leq c_0\}$. Without loss of generality, we assume that the last $c_0 - k,\ 1 \leq k \leq n$ elements of x and y are the same, and the rest are not, i.e. $x_i \neq y_i$ for $1 \leq i \leq k$ and $x_i = y_i$ for $k+1 \leq i \leq c_0$. Let $x^* = (x_1, ..., x_k)$, $y^* = (y_1, ..., y_k)$ and $v^* = y^* - x^*$. Therefore only a finite number of $t \in [0,1]$ such that point $(1-t)x^* + ty^*$ belongs to $\mathcal{D}^k = \{x \in \tilde{\Theta}_{i_1} \times ... \times \tilde{\Theta}_{i_k} : |x_i| \in \{\lambda, c\lambda\} \text{ for some } 1 \leq i \leq k\}$.

Let $w = (w_1, ..., w_k)$, and define $g : \tilde{\Theta}_{i_1} \times ... \times \tilde{\Theta}_{i_k} \to \mathbb{R}$, as

$$g(w) = L((w, x_{k+1}, ..., x_{c_0})).$$

Define $\varphi^*(t) = g(x^* + tv^*), t \in [0,1]$. Then similar to Case 1, we can show that

$$\frac{d\varphi^*}{dt} = \sum_i \frac{\partial g}{\partial x_i^*}(x^* + tv^*) \cdot v_i^* = \sum_{i=1}^k \frac{\partial L}{\partial x_i}\big((x^* + tv^*, x_{k+1}, ..., x_n)\big) \cdot v_i^*$$

is continuous, and

$$\begin{aligned}
\frac{d^2\varphi^*}{dt^2} &= \sum_{i,j} \frac{\partial^2 g}{\partial x_i^* \partial x_j^*}(x^* + tv^*) v_i^* v_j^* \\
&= \sum_{i,j=1}^k \frac{\partial^2 L}{\partial x_i \partial x_j}\big((x^* + tv^*, x_{k+1}, ..., x_n)\big) v_i^* v_j^* \\
&\geq 0
\end{aligned}$$

except a finite number of $t \in [0,1]$. Therefore $d\varphi^*/dt$ is non-decreasing, and further $\varphi^*(t)$ is convex on $[0,1]$. By definition, for any $t \in [0,1]$,

$$\begin{aligned}
L((1-t)x + ty) = L(x + tv) &= g(x^* + tv^*) \\
&= \varphi^*(t) \leq t\varphi^*(1) + (1-t)\varphi^*(0) = tL(y) + (1-t)L(x)
\end{aligned}$$

REFERENCES

REFERENCES

[1] Jane-Ling Wang, Jeng-Min Chiou, and Hans-Georg Mueller. Review of functional data analysis. *arXiv preprint arXiv:1507.05135*, 2015.

[2] Grace Wahba. *Spline models for observational data*, volume 59. Siam, 1990.

[3] Chong Gu. *Smoothing Spline ANOVA Models*, volume 297 of *Springer Series in Statistics*. Springer, New York, NY, 2013.

[4] F Riesz and B. Sz-Nagy. Functional analysis. *Ungar, New York*, 1990.

[5] Simeng Qu, Jane-Ling Wang, and Xiao Wang. Optimal estimation for the functional cox model. *Annals of Statistics*, 44(4):1708–1738, 08 2016.

[6] Hervé Cardot, Frédéric Ferraty, André Mas, and Pascal Sarda. Testing hypotheses in the functional linear model. *Scandinavian Journal of Statistics*, 30(1):241–255, 2003.

[7] Hervé Cardot, Aldo Goia, and Pascal Sarda. Testing for no effect in functional linear regression models, some computational approaches. *Communications in Statistics-Simulation and Computation*, 33(1):179–199, 2004.

[8] Wenceslao González-Manteiga and Adela Martínez-Calvo. Bootstrap in functional linear regression. *Journal of Statistical Planning and Inference*, 141(1):453–461, 2011.

[9] Nadine Hilgert, André Mas, and Nicolas Verzelen. Minimax adaptive tests for the functional linear model. *The Annals of Statistics*, 41(2):838–869, 2013.

[10] Jing Lei. Adaptive global testing for functional linear models. *Journal of the American Statistical Association*, 109(506):624–634, 2014.

[11] T Tony Cai and Ming Yuan. Minimax and adaptive prediction for functional linear regression. *Journal of the American Statistical Association*, 107(499):1201–1216, 2012.

[12] Yuri Izmailovich Ingster. On the minimax nonparametric detection of signals in white gaussian noise. *Problemy Peredachi Informatsii*, 18(2):61–73, 1982.

[13] Yu I Ingster. An asymptotic minimax test of nonparametric hypotheses about spectral density. *Theory of Probability & Its Applications*, 29(4):846–847, 1985.

[14] Yu I Ingster. Minimax testing of nonparametric hypotheses on a distribution density in the l_p metrics. *Theory of Probability & Its Applications*, 31(2):333–337, 1987.

[15] Yuri I Ingster. Asymptotically minimax hypothesis testing for nonparametric alternatives. i, ii, iii. *Mathematical Methods of Statistics*, 2(2):85–114, 1993.

[16] Vladimir G Spokoiny. Adaptive hypothesis testing using wavelets. *The Annals of Statistics*, 24(6):2477–2498, 1996.

[17] Jianqing Fan. Test of significance based on wavelet thresholding and neyman's truncation. *Journal of the American Statistical Association*, 91(434):674–688, 1996.

[18] Oleg V Lepski and Vladimir G Spokoiny. Minimax nonparametric hypothesis testing: the case of an inhomogeneous alternative. *Bernoulli*, 5(2):333–358, 1999.

[19] Jianqing Fan, Chunming Zhang, and Jian Zhang. Generalized likelihood ratio statistics and wilks phenomenon. *Annals of Statistics*, pages 153–193, 2001.

[20] Jeffrey D Hart. *Nonparametric smoothing and lack-of-fit tests*. Springer New York, 1997.

[21] Christophe Crambes, Alois Kneip, and Pascal Sarda. Smoothing splines estimators for functional linear regression. *The Annals of Statistics*, pages 35–72, 2009.

[22] T Tony Cai and Ming Yuan. Optimal estimation of the mean function based on discretely sampled functional data: Phase transition. *The annals of statistics*, 39(5):2330–2355, 2011.

[23] Fang Yao, Hans-Georg Müller, and Jane-Ling Wang. Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100(470):577–590, 2005.

[24] Pang Du and Xiao Wang. Penalized likelihood functional regression. *Statistica Sinca*, 24(10.5705):1017–1041, 2014.

[25] F. Cucker and S. Smale. On the mathematical foundations of learning. *American Mathematical Society*, 39(1):1–49, 2002.

[26] Peter Hall and Joel L Horowitz. Methodology and convergence rates for functional linear regression. *The Annals of Statistics*, 35(1):70–91, 2007.

[27] Rajendra Bhatia, Chandler Davis, and Alan McIntosh. Perturbation of spectral subspaces and solution of linear operator equations. *Linear Algebra and its Applications*, 52:45–67, 1983.

[28] Peter Hall and Joel L Horowitz. Nonparametric methods for inference in the presence of instrumental variables. *The Annals of Statistics*, 33(6):2904–2929, 2005.

[29] D. R. Cox. Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B*, 34(2):187–220, 1972.

[30] P. Sasieni. Information bounds for the conditional hazard ratio in a nested family of regression models. *Journal of the Royal Statistical Society. Series B*, 54(2):617–635, 1992.

[31] P. Sasieni. Non-Orthogonal Projections and Their Application to Calculating the Information in a Partly Linear Cox Model. *Scandinavian Journal of Statistics*, 19(3):215–233, 1992.

[32] Trevor J Hastie and R. Tibshirani. Generalized additive models. *Statistical Science*, 1:297–318, 1986.

[33] Trevor J Hastie and R. Tibshirani. Exploring the nature of covariate effects in proportional hazards model. *Biometrics*, 46:1005–1016, 1990.

[34] Jian Huang. Efficient Estimation of the Partly Linear Additive Cox Model. *The Annals of Statistics*, 27(5):1536–1563, 1999.

[35] D. R. Cox. Partial Likelihood. *Biometrika*, 62(2):269–276, 1975.

[36] Anastasios A. Tsiatis. A Large Sample Study of Cox's Regression Model. *The Annals of Statistics*, 9(1):93–108, 1981.

[37] P. K. Andersen and R. D. Gill. Cox's Regression Model for Counting Processes: A Large Sample Study. *The Annals of Statistics*, 10(4):1100–1120, 1982.

[38] S Johansen. An Extension of Cox's Regression Model. *International Statistical Review*, 51(2):165–174, 1983.

[39] Martin Jacobsen. Maximum Likelihood Estimation in the Multiplicative Intensity Model: A Survey. *International Statistical Review*, 52(2):193–207, 1984.

[40] Kun Chen, Kehui Chen, Hans-Georg Müller, and Jane-Ling Wang. Stringing High-Dimensional Data for Functional Analysis. *Journal of the American Statistical Association*, 106(493):275–284, 2011.

[41] Christophe Crambes, Alois Kneip, and Pascal Sarda. Smoothing splines estimators for functional linear regression. *The Annals of Statistics*, 37(1):35–72, 2009.

[42] TT Cai and Ming Yuan. Minimax and adaptive prediction for functional linear regression. *Journal of the American Statistical Association*, 107:1–35, 2012.

[43] D. Kong, J. Ibrahim, E. Lee, and H. Zhu. FLCRM: Functional Linear Cox Regression Models. *Submitted*, 2014.

[44] Grace Wahba. *Spline models for observational data*, volume 59. Siam, 1990.

[45] Ming Yuan and T. Tony Cai. A reproducing kernel Hilbert space approach to functional linear regression. *The Annals of Statistics*, 38(6):3412–3444, 2010.

[46] L. Breiman and J. H. Friedman. Estimating Optimal Transformations for Multiple Regression and Correlation. *Journal of the American Statistical Association*, 80(391):580–598, 1985.

[47] James R Carey, Pablo Liedo, Hans-Georg Müller, Jane-Ling Wang, Damla Senturk, and Lawrence Harshman. Biodemography of a long-lived tephritid: reproduction and longevity in a large cohort of female Mexican fruit flies, Anastrepha ludens. *Experimental gerontology*, 40(10):793–800, 2005.

[48] AW van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 2000.

[49] AW van der Vaart and JA Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, New York, NY, 1996.

[50] T.R. Fleming and D.P. Harrington. *Counting Processes and Survival Analysis.* Wiley, New York., 1991.

[51] AB Tsybakov and V Zaiats. *Introduction to nonparametric estimation.* Springer, New York, NY, 2009.

[52] Jeffrey S Morris. Functional regression. *arXiv preprint arXiv:1406.4068*, 2014.

[53] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

[54] Michael Elad and Michal Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *Image Processing, IEEE Transactions on*, 15(12):3736–3745, 2006.

[55] Jianchao Yang, John Wright, Thomas Huang, and Yi Ma. Image super-resolution as sparse representation of raw image patches. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[56] John Wright, Allen Y Yang, Arvind Ganesh, Shankar S Sastry, and Yi Ma. Robust face recognition via sparse representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(2):210–227, 2009.

[57] Jianchao Yang, Kai Yu, Yihong Gong, and Tingwen Huang. Linear spatial pyramid matching using sparse coding for image classification. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1794–1801, 2009.

[58] Emmanuel J Candes, Justin K Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on pure and applied mathematics*, 59(8):1207–1223, 2006.

[59] David Donoho. Compressed sensing. *Information Theory, IEEE Transactions on*, 52(4):1289–1306, 2006.

[60] Qiang Zhang and Baoxin Li. Discriminative k-svd for dictionary learning in face recognition. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2691–2698. IEEE, 2010.

[61] Shu Kong and Donghui Wang. A dictionary learning approach for classification: separating the particularity and the commonality. In *Computer Vision–ECCV 2012*, pages 186–199. Springer, 2012.

[62] Julien Mairal, Francis Bach, and Jean Ponce. Task-driven dictionary learning. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(4):791–804, 2012.

[63] Mingyuan Zhou, Haojun Chen, John Paisley, Lu Ren, Lingbo Li, Zhengming Xing, David Dunson, Guillermo Sapiro, and Lawrence Carin. Nonparametric bayesian dictionary learning for analysis of noisy and incomplete images. *Image Processing, IEEE Transactions on*, 21(1):130–144, 2012.

[64] Michal Aharon, Michael Elad, and Alfred Bruckstein. K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *Signal Processing, IEEE Transactions on*, 54(11):4311–4322, 2006.

[65] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online dictionary learning for sparse coding. In *Proceedings of the 26th annual international conference on machine learning*, pages 689–696. ACM, 2009.

[66] Julien Mairal, Jean Ponce, Guillermo Sapiro, Andrew Zisserman, and Francis R Bach. Supervised dictionary learning. In *Advances in neural information processing systems*, pages 1033–1040, 2009.

[67] MarcAurelio Ranzato, Christopher Poultney, Sumit Chopra, Yann L Cun, et al. Efficient learning of sparse representations with an energy-based model. In *Advances in neural information processing systems*, pages 1137–1144, 2006.

[68] Ori Bryt and Michael Elad. Compression of facial images using the k-svd algorithm. *Journal of Visual Communication and Image Representation*, 19(4):270–282, 2008.

[69] Elisabeth Gassiat and Ramon Van Handel. Consistent order estimation and minimal penalties. *Information Theory, IEEE Transactions on*, 59(2):1115–1128, 2013.

[70] Mingyuan Zhou, Haojun Chen, Lu Ren, Guillermo Sapiro, Lawrence Carin, and John W Paisley. Non-parametric bayesian dictionary learning for sparse image representations. In *Advances in neural information processing systems*, pages 2295–2303, 2009.

[71] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

[72] Julie Mairal and Francis Bach. Sparse Modeling for Image and Vision Processing. *Foundations and Trends in Computer Graphics and Vision*, 8(2):85–283, 2014.

[73] Yu Wang, Wotao Yin, and Jinshan Zeng. Global convergence of admm in nonconvex nonsmooth optimization. *arXiv preprint arXiv:1511.06324*, 2015.

[74] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online learning for matrix factorization and sparse coding. *The Journal of Machine Learning Research*, 11:19–60, 2010.

[75] Julien Mairal, Francis Bach, Jean Ponce, Guillermo Sapiro, and Andrew Zisserman. Non-local sparse models for image restoration. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 2272–2279. IEEE, 2009.

VITA

VITA

CONTACT

Department of Statistics, Purdue University
150 N. University St., West Lafayette, IN 47907

Cell: (765) 637-6580
Email: qu20@purdue.edu

EDUCATION

Purdue University, West Lafayette, IN, USA

Ph.D. in Statistics (GPA: 4.0/4.0) Aug 2012-Dec 2016

Advisor: Xiao Wang

Dissertation Topic: Some Functional Regression Models in the Frame Work of Reproducing Kernel Hilbert Space

Research Interest: functional data analysis, non-parametric statistics, big data, dictionary learning, splines, high-dimensional data analysis, survival analysis

M.S. in Mathematical Statistics (GPA: 4.0/4.0) Aug 2012-Dec 2014

University of Science and Technology of China, Hefei, Anhui, China

B.S. in Applied Mathematics (GPA: 3.92/4.3) Aug 2008-Jun 2012

PUBLICATIONS

Qu, S., Wang, J.L., Wang, X.(2016). Optimal Estimation for the Functional Cox Model. *Annals of Statistics, 44*(4), 1708-1738.

Qu, S., Wang, X.(2016). Simultaneous Sparse Dictionary Learning and Pruning. *To be submitted.*

Qu, S., Wang, X.(2016). Optimal Global Test for Functional Linear Regression Models. *To be submitted.*

PROFESSIONAL EXPERIENCE

Research Assistant Aug 2013-Present

Department of Statistics, Purdue University, West Lafayette, IN

- Conducting innovative research on functional data analysis related topics, including Functional Cox Model, Functional Linear Model and Function-on-scalar models
- Proposing and validating methods for estimation and hypothesis testing
- Implementing fast algorithms for the proposed methods to solve practical problems with real-world data.

Research Assistant Jan 2015-May 2016

Regenstrief Center for Healthcare Engineering, Purdue University, West Lafayette, IN

- Providing statistical and analytical support for Bundle Pricing project and Population Health Management project
- Making practical suggestions on healthcare efficiency using clinical data and medical claim data

Predictive Modelling Summer Intern Jun 2015-Aug 2015

CNA Financial Corporation, Chicago, IL

- Modelling experience rate for aging services policies using blended models
- Implementing the model using SAS Enterprise Guide (data preparation) and R (modelling)
- Writing documentation for the model

Teaching Assistant Aug 2014-Dec 2014

Department of Statistics, Purdue University, West Lafayette, IN

- Class STAT524: Applied Multivariate Statistical Analysis
- Grading homework reports and holding office hours

Consultant Aug 2012-Dec 2014

Statistics in Community (StatCom), Purdue University, West Lafayette, IN

- Providing statistical consulting services to Child And Parent Services (CAPS)
- Modelling and analyzing data to evaluate CAPS's service

CONFERENCE TALKS

Joint Statistical Meetings (JSM), Chicago, IL Aug 2014

Contributed Session: Nonparametric Methods for High-Dimensional Data

Optimal Estimation and Variable Selection for Multivariate Varying Coefficient Model with Functional Response

ICSA/Graybill Applied Statistics Symposium and Graybill Conference, Fort Collins, Co Jun 2015

Invited Session: New Frontier of Functional Data Analysis

Topics: Optimal estimation for the functional Cox model

Joint Statistical Meetings (JSM), Boston, MA Aug 2014

Topics: Optimal Global Test for Functional Linear Regression Models

Finalists in ASA non-parametric statistics paper competition

Joint Statistical Meetings (JSM), Montreal, Quebec, Canada Aug 2013

Contributed Session: Functional Analysis and Mixed Models

Topics: Optimal estimation for the functional Cox model

ICSA/ISBS Joint Statistics Conference, Bethesda, MD Jun 2013

Invited Session: Functional Data Analysis: Applications and New Advances

Topics: Optimal estimation for the functional Cox model

HONORS AND AWARDS

Bilsland Dissertation Fellowship Purdue University, 2016

William J. Studden Publication Award Purdue University, 2016

JSM Student Paper Award Finalist American Statistical Association, 2014

Ross Fellowship Purdue University, 2012

Bao Gang Outstanding Student Award Hefei, China, 2011

Zhang Zong Zhi Technology Student Award	Hefei, China, 2010
National Inspiration Student Award	Hefei, China, 2009