Novel Methods for Estimation and Inference in Varying Coefficient Models

by

Yuan Yang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2020

Doctoral Committee:

Associate Professor Jian Kang, Co-Chair
Professor Yi Li, Co-Chair
Associate Professor Chad Brummett
Professor Timothy Johnson
Professor Ji Zhu

Yuan Yang

yuanyang@umich.edu

ORCID iD: 0000-0002-8207-5529

DEDICATION

To my sisters, my parents, and my husband

ACKNOWLEDGEMENTS

I want to express my deepest gratitude to Dr. Yi Li and Dr. Jian Kang, for their guidance throughout my Ph.D. life. Without their support and encouragement, I would not come so far and decide to go further in my future career. Whenever I had questions and concerns, they were always there to help me. I have learned a lot from working with them: independent thinking, collaborating with others, being enthusiastic, etc. It is an honor for me to be their student.

I thank Dr. Chad Brummett, Dr. Timothy Johnson, and Dr. Ji Zhu for their kindness of being in my committee. Their professional comments have helped complete this dissertation.

I want to thank Dr. Kevin He and Dr. Jack D. Kalbfleisch. They have been helping me in my research assistant work and gave me strong support in my job search.

I would also like to thank the faculties and staff, fellow students, and friends for their help during the last six years in Michigan.

In the end, I have to thank my sisters, Fei and Mei, my parents, and my husband, Tianyu, for their love and support throughout my life.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

LIST OF APPENDICES

Appendix

ABSTRACT

Function type parameters relax many model assumptions because of the flexibility and the size of the parameter space. However, the curse of dimensionality has been the biggest challenge in the nonparametric regression area. An advantageous approach to dimension reduction is using basis expansion to approximate infinite parameter space. An even more challenging problem is estimating functions with unique structures, such as functions with zero-effect regions. The main part of this dissertation is working on varying coefficients with zero-effect regions. We propose a novel model that can detect zero-effect regions and estimate the non-zero effects simultaneously. We provide theoretical support for the inference of our proposed estimators. Simulation studies and real data analyses demonstrate the advantage of our models. This dissertation also introduces a new model that considers the additive effects from a novel aspect: estimating the dynamic effect changes. Simulations and real data applications provide comparisons between our model and the existing model.

CHAPTER I

Introduction

Varying coefficient models have been used to explore dynamic effect patterns in many scientific areas, such as in biomedicine, finance, and epidemiology. When the effect is changing from positive to negative (or vice versa), detection of transition regions (zero regions) is of practical importance for proper intervention. As the most existing models ignore the existence of zero regions, the first chapter proposes a new soft-thresholded varying coefficient model, where the varying coefficient functions are piecewise smooth with zero regions. Our new modeling approach enables us to perform variable selection and detect the zero regions of selected variables simultaneously, obtain point estimates of the varying coefficients with zero regions and construct a new type of sparse confidence intervals that accommodate zero regions. We prove the asymptotic properties of the estimator, and our simulation study reveals that the confidence intervals achieve the desired coverage probability. We apply the proposed method to analyze a large scale preoperative opioid use study and obtain some interesting results on opioid use.

The second chapter is considering the generalized additive models (GAMs). GAMs have been widely used for modeling nonlinear effects of predictors on a variety of outcomes. However, the explanation of covariates' effects in GAMs is intriguing and statistical inference on effects is challenging. Extending GAMs, we propose a new class

of models that can directly characterize the dynamic effect change of each predictor. Our model, which incorporates derivatives of nonlinear effects as functional parameters of interest, is termed the generalized additive dynamic effect change model. We develop an efficient statistical procedure for inferring functional parameters embedded in the reproducing Hilbert kernel space. As opposed to GAMs, our derivative-based model renders a straightforward interpretation of model parameters. We establish the large sample properties for the proposed method and show its superior performance compared to GAM in various simulation scenarios. We apply our method to construct an individualized risk prediction model for opioid use, which provides a better understanding of dynamic effect changes of potential risk factors.

In the third chapter, we consider the high-dimensional Cox models with time-varying effects that contain zero regions. As opposed to the commonly used regularization methods, we apply the idea of the soft-thresholding operator from our first chapter in the space of smooth functions. This leads to a more interpretable model with a straightforward inference procedure. We develop an efficient algorithm for inference in the target functional space. We show that the proposed method enjoys good theoretical properties. The method is further illustrated and evaluated via extensive simulation studies and a data analysis of Boston Lung Cancer Study.

CHAPTER II

A New Soft-Thresholded Varying Coefficient Model to Predict Opioid Use with Risk Factors that Have Zero-effect Regions

2.1 Introduction

According to the World Drug Report [84], opioid use for pain treatment has risen sharply, but without much improvement in reducing the severity of chronic pain. The rapid rise in opioid use was strongly associated with the incidence of emergency department visits and deaths [12]. Patients with preoperative opioid use have worse surgical outcomes, greater postoperative pain, more pronounced morbidity, higher rates of use of health care services [14, 63, 102], and are less likely to stop opioid-based therapy after surgery [22, 35]. To avoid unnecessary opioid use and prevent possible opioid addiction, effective strategies for opioid prescription management are needed for both patients and physicians. For obese patients, effective prescription management is especially important because of complex co-comorbidities and high prevalence of obstructive sleep apnea [70]. It is critical to understand whether and how the association between preoperative opioid use and pain is modified by the level of body mass index (BMI) [70].

This work was motivated by a preoperative opioid use study, which assessed the

association of preoperative opioid use and the characteristics of patients in a broadly representative surgical cohort [43]. We did a preliminary analysis that considers a varying coefficient model using B-spline approximation. The preliminary analysis shows that the dose-relationship between opioid use and pain level is changing from negative to positive when BMI is increasing from 15.5 to 20.0, and the non-significant and significant regions are not well separated. A practical explanation is that there may exist zero-effect regions in terms of BMI for pain on opioid use. The zero-effect regions of BMI may hint at possible opioid addiction among those with BMI less than 20.0. However, most existing methods ignore the existence of zero-effect regions. There is a need to develop a varying coefficient model that enables us to estimate zero-effect regions and quantify the associated uncertainty.

Varying coefficient models [39] are commonly used to characterize the dynamic changes of regression effects. Framing the model in the context of opioid use, we denote by Y the total amount of preoperative opioids, and by $X_1, \ldots, X_p$ the p covariates, consisting of demographic information and clinical symptoms, such as preoperative pain. The following model detects how the covariate effects on opioid use are modified by BMI (denoted by W):

$$Y = \sum_{j=1}^{p} X_j \beta_j(W) + \epsilon, \tag{2.1}$$

where $\beta_j(W)$ is the varying coefficient function representing the effect of X_j, and ϵ is a random variable with mean zero and variance σ^2. We set X_1 to be 1, corresponding to the intercept function. The challenge lies in how to detect zero-effect regions and draw inference on varying coefficient functions simultaneously, and is aggravated when p is large.

Local log-likelihood approaches have been proposed to estimate $\beta_j(W)$. For example, Hoover et al. (1998) [45] used the smoothing spline and local polynomial methods;

Fan and Zhang (1999) [33] and Fan et al. (2000) [32] proposed a two-step procedure to allow more flexibility of coefficient functions; Wu et al. (2000) [94] and Chiang et al. (2001) [17] proposed component-based kernel and smoothing spline estimators for varying coefficient models with repeated measurements. In high-dimensional settings, variable selection and screening with varying coefficient models were studied [52, 55, 57]. The local polynomial estimators may not provide adequate smoothing for all the coefficients simultaneously and the computational burden of smoothing splines can be heavy. Other alternative methods were proposed. They included global estimation and variable selection for varying coefficient models based on basis approximations [49, 50] and penalized spline-based models [15, 16, 28, 31, 41, 47, 74, 88, 89, 97]. However, none of these methods is able to detect zero-effect regions.

Model (2.1) differs from functional linear models [64, 65, 100] and scalar-on-image regression models [51], of which both coefficients and covariates are functional. The roles and interpretations of functional coefficients deviate from those in model (2.1), as the latter is designed to characterize the varying effects of scalar covariates. Moreover, the methods of drawing statistical inference on zero-effect regions for functional linear models [51, 100] are not applicable to model (2.1), which addresses the particular question on opioid use.

We propose a soft-thresholded varying coefficient model, where coefficients in model (2.1) are constructed by applying soft-thresholding operators to smooth functions. The soft-thresholded varying coefficients are continuous, piecewise smooth, and with zero-effect regions. The smooth functions before soft-thresholding can be approximated using B-splines [27, 73, 76], or some other basis functions, such as smoothing splines or reproducing kernel Hilbert space splines [8, 86]. The soft-thresholded function, originally introduced to construct estimators for the wavelet coefficients [24, 25], has been widely used for effect shrinkage: Chiang et al. (2001) [17] proposed an adaptive, data-driven threshold for image denoising in a Bayesian framework with

the generalized Gaussian distribution prior based on wavelet soft-thresholding; Tibshirani (1996) [81] also pointed out that the lasso estimator is a soft-thresholded estimator when the covariate matrix has an orthonormal design. As all of these estimators were designed for parameters with finite dimensions, their usage for functional coefficients, including varying coefficients, remains elusive.

Our approach is new in the following aspects. First, our method involves a novel application of a soft-thresholding operator in a functional space, which enables us to uncover zero-effect regions of varying coefficients. The soft-thresholded estimates are continuous, piecewise smooth and with zero-effect regions, and possess an easy interpretation for a range of applications. Second, our new modeling framework enables us to estimate varying coefficients and draw the statistical inference. We particularly develop a new type of confidence interval, termed sparse confidence intervals, which can be degenerated to a singleton with a non-zero probability. Finally, we have established theoretical properties, which inform valid statistical inference for high-dimensional varying coefficient models.

2.2 Method

2.2.1 Varying coefficient models with zero-effect regions

In model (2.1), we write $\boldsymbol{\beta}(w) = \{\beta_1(w), \ldots, \beta_p(w)\}^T$ as a vector of varying coefficients, where p may grow with the sample size. Without loss of generality, we assume $W \in \mathbb{D} = [0, 1]$. To detect zero-effect regions of $\boldsymbol{\beta}(w)$, we assume that $\beta_j, j = 1, \ldots, p$, is continuous everywhere, with zero-effect regions consisting of at least one interval, and is smooth over regions where its effect is non-zero. Specifically, let $R_0(\beta) = \{w : \beta(w) = 0, w \in \mathbb{D}\}$, $R_-(\beta) = \{w : \beta(w) < 0, w \in \mathbb{D}\}$, $R_+(\beta) = \{w : \beta(w) > 0, w \in \mathbb{D}\}$, and $\overline{R}$ be the closure of any set $R \subseteq \mathbb{D}$. The functional space $\mathbb{H}$ containing β_j is defined as follows.

Definition 2.2.1 $\mathbb{H}$ *contains* $\beta(w)$ *with: (continuity)* $\lim_{w \to w_0} \beta(w) = \beta(w_0)$*, for any* $w_0 \in \mathbb{D}$*; (zero-effect regions)* $\overline{R}_0(\beta)$ *contains at least one interval with a non-zero Lebesgue measure; (piecewise smoothness)* $\overline{R}_+(\beta) \bigcup \overline{R}_-(\beta)$ *can be partitioned as a union of disjoint intervals, each with a non-zero Lebesgue measure. The dth derivative of* $\beta(w)$ *exists and satisfies the Lipschitz condition on each interval:*

$$|\beta^{(d)}(s) - \beta^{(d)}(w)| \leq C|s - w|^t,$$

where d *is a non-negative integer, and* $t \in (0, 1]$ *such that* $m \equiv d + t > 0{\cdot}5$*.*

The smoothness requirement for β in our definition is weaker than that in Kang et al. (2018) [51]. The full-zero coefficients are those with $R_0 = \mathbb{D}$, and partial-zero coefficients are those with $R_0 \subsetneq \mathbb{D}$. Definition 2.2.1 implies a "buffer zone" when an effect switches signs, reflecting gradual degradation in real life. We assume that the true parameter is β_0 and $\beta_{0j} \in \mathbb{H}$ for all j. Let $p_0 = \sum_{j=1}^{p} \mathcal{I}\{\beta_{0j}(w) \equiv 0\}$ be the number of full-zero coefficients, and $\tilde{p} = p - p_0$ be the number of partial-zero and non-zero coefficients. Here, $\mathcal{I}\{\cdot\}$ is the indicator function. Without loss of generality, we assume the first $\tilde{p}$ coefficients are either partial-zero or non-zero. Certain sparsity conditions will be imposed on $\tilde{p}$ later.

2.2.2 Soft-thresholding operator

Representing zero-effect regions for varying coefficients, we propose a soft-thresholding operator ζ:

$$\zeta_{\{\theta,\alpha\}}(w) = \{\theta(w) - \alpha\}\, \mathcal{I}\{\theta(w) > \alpha\} + \{\theta(w) + \alpha\}\, \mathcal{I}\{\theta(w) < -\alpha\},$$

where $\alpha > 0$ is the thresholding parameter and $\theta(w)$ is a real-valued function. Our proposal resembles Donoho and Johnstone (1994) [25], which was designed for denoising wavelet coefficients. However, our proposal is a functional operator which

transforms a function to a function (Figure 2.1).

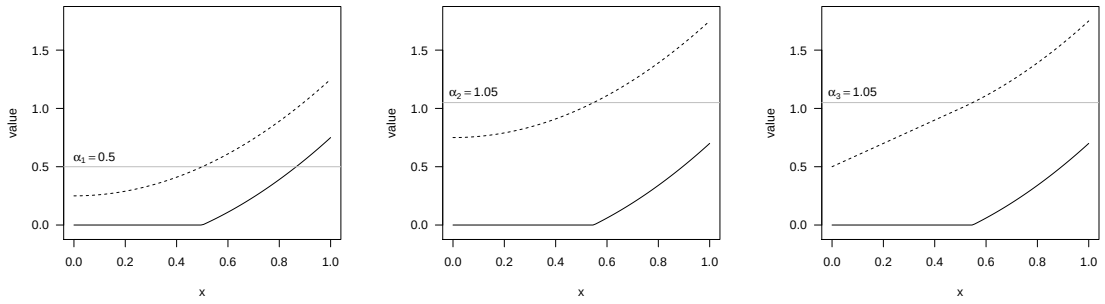


Figure 2.1: Demonstration of the soft-thresholding operator, which maps different smooth functions (in dash) with different thresholding values (α_1,α_2,α_3) to the same curve with a zero-effect region (in solid).

Let $\mathbb{F}_0$ be a class of functions θ defined on $\mathbb{D}$, with the dth derivative $\theta^{(d)}$ satisfying the Lipschitz condition in Definition 2.2.1. According to Lemma 1 in the Supplementary Material, we have that for any function $\beta(w) \in \mathbb{H}$ and any $\alpha > 0$, there exists at least one $\theta(w) \in \mathbb{F}_0$ such that $\beta(w) = \zeta_{\{\theta,\alpha\}}(w)$.

As illustrated by Figure 2.1, the soft-thresholding operator maps different smooth $\theta(w)$'s with different thresholding parameter α's to the same $\beta(w)$. Even for a fixed α, $\theta(w)$ may not be uniquely defined and, hence, is not estimable without further constraints. Our strategy is to consider a sieve space that approximates $\mathbb{F}_0$ and shows that, within the sieve space, a penalized loss function can uniquely determine a $\theta(w)$, which after soft thresholding will approximate the desired $\beta(w)$. In theory, we may set α to be any positive number, but our numerical experience suggests that choosing an appropriate α, which is comparable to the scale of $\beta(w)$, lead to more stable and efficient estimates. Thus, in a regression setting, we specify covariate-specific α's.

2.2.3 Spline approximation and differentiable approximation

We specify a B-spline function sieve space, denoted by $\mathbb{F}$, to approximate $\mathbb{F}_0$. Let $K = O(n^\nu)$ be an integer with $0 < \nu < 0{\cdot}5$. Following Schumaker (2007) [71], we

let $B_k(w)(1 \leq k \leq q)$ with $q = K + d$ be the B-spline basis functions of degree $d + 1$ associated with the knots $0 = w_0 < w_1 < \cdots < w_{K-1} < w_K = 1$, satisfying $\max_{1 \leq k \leq K}(w_k - w_{k-1}) = O(n^{-\nu})$.

Definition 2.2.2 *Let* $\boldsymbol{B}(w) = \{B_1(w), \ldots, B_q(w)\}^T$ *be a functional vector of the B-spline bases. We define*

$$\mathbb{F} = \left\{ \sum_{k=1}^{q} \gamma_k B_k(w), w \in \mathbb{D}, \gamma_k \in \mathbb{R}, k = 1, \ldots, q \right\}.$$

Let $\boldsymbol{X} = (X_1, \ldots, X_p)^T$. With the observed data $\{(Y_i, W_i, \boldsymbol{X}_i)\}_{i=1}^n$ being independent samples of $\{(Y, W, \boldsymbol{X})\}$, we specify

$$Y_i = \sum_{j=1}^{p} X_{ij} \zeta_{\left\{\sum_{k=1}^{q} \gamma_{jk} B_k, \alpha_j\right\}}(W_i) + \epsilon_i. \tag{2.2}$$

Compared to model (2.1), model (2.2) should be viewed as a "working" model, wherein γ_{jk} may not be unique or estimable. But with a penalized loss function specified below, we show that γ_{jk} can be uniquely estimated. Thus, the soft thresholded estimate based on a working sieve model can approximate the true β_0.

We define a penalized least-squares loss function:

$$l(\boldsymbol{\gamma}; \boldsymbol{X}, Y, W) = \left[Y - \sum_{j=1}^{p} X_j \zeta_{\{\boldsymbol{B}^T \boldsymbol{\gamma}_j, \alpha_j\}}(W) \right]^2 + \rho \sum_{j=1}^{p} \left\{ \boldsymbol{B}(W)^T \boldsymbol{\gamma}_j \right\}^2,$$

where $\rho > 0$ and $\gamma = (\boldsymbol{\gamma}_1^T, \ldots, \boldsymbol{\gamma}_p^T)^T$ are the coefficients of bases. The penalty term aims to select zero-effect regions and identify the unique inner functions in $\mathbb{F}$. Although we use the same q for all coefficient functions, different q can be chosen for different covariates.

Let f be a non-random function, and $\xi_1, \ldots, \xi_n$ be i.i.d. copies of random vector ξ. We denote by $\mathrm{E}f(\xi)$ the theoretical mean of $f(\xi)$ and by $\mathrm{E}_n f(\xi) = n^{-1} \sum_{i=1}^{n} f(\xi_i)$

the empirical mean of $f(\xi)$. Define

$$\tilde{\boldsymbol{\gamma}} = \arg\min_{\boldsymbol{\gamma}} \mathrm{E} l(\boldsymbol{\gamma}; \boldsymbol{X}, Y, W)$$

as the true sieve parameters to estimate. Let $\tilde{\theta}_j = \boldsymbol{B}^T \tilde{\boldsymbol{\gamma}}_j$ and $\tilde{\boldsymbol{\beta}}(w) = (\tilde{\beta}_1, \ldots, \tilde{\beta}_p)^T$ with $\tilde{\beta}_j = \zeta_{\{\tilde{\theta}_j, \alpha_j\}}(w)$.

For given α and q, we define the thresholded sieve space

$$\mathbb{S}_{q,\alpha} = \left\{ \beta(w) = \zeta_{\{\theta,\alpha\}}(w) : \theta(w) = \sum_{k=1}^{q} \gamma_k B_k(w), w \in \mathbb{D}, \gamma_k \in \mathbb{R}, k = 1, \ldots, q \right\}.$$

By Lemma 2 in the Supplementary Material, if $\beta_{0j} \in \mathbb{S}_{q,\alpha_j}$ for $j = 1, \ldots, \tilde{p}$ with q and α_j the same as in the penalized likelihood, $||\tilde{\boldsymbol{\beta}} - \beta_0||_\infty = O((\tilde{p}\rho)^{1/2})$; if $\beta_{0j} \notin \mathbb{S}_{q,\alpha_j}$ for $j = 1, \ldots, \tilde{p}$, we have $||\tilde{\boldsymbol{\beta}} - \beta_0||_\infty = O((\tilde{p}\rho + \tilde{p}q^{-2m})^{1/2})$, where m is the smoothness parameter as in Definition 2.2.1.

As ζ is not differentiable everywhere, we consider a smooth approximation of it.

Definition 2.2.3 *A smooth approximation of $\zeta_{(\theta,\alpha)}$, denoted by $H_\eta(\theta, \alpha)$ ($\eta > 0$), is continuous and twice differentiable with respect to θ everywhere and $\sup_{w \in \mathbb{D}} |H_\eta(\theta, \alpha) - \zeta_{(\theta,\alpha)}| = \nabla(\eta)$, where $\nabla(\eta) \geq 0$ and $\lim_{\eta \to 0+} \nabla(\eta) = 0$.*

For example, a smooth approximation of $\zeta_{(\theta,\alpha)}$ is defined as

$$\begin{aligned} H_\eta\{\theta(w), \alpha\} =& \frac{1}{2}\Bigg(\left[1 + \frac{2}{\pi}\arctan\{\theta_-(w)/\eta\}\right] \theta_-(w) + \\ & \left[1 - \frac{2}{\pi}\arctan\{\theta_+(w)/\eta\}\right] \theta_+(w) \Bigg), \end{aligned}$$

where $\alpha > 0$, $\eta > 0$ and $\theta_\pm(w) = \theta(w) \pm \alpha$. The approximation error between $H_\eta\{\theta(w), \alpha\}$ and $\zeta_{(\theta,\alpha)}$ is bounded by $\eta + O(\eta^3)$ and H is continuous and differentiable. The proof can be found in the Supplementary Material.

For simplicity, we drop α and η and write $\boldsymbol{h}(w, \gamma) = \{h_1(w, \gamma_1), \ldots, h_p(w, \gamma_p)\}^T$

with $h_j(w, \boldsymbol{\gamma}_j) = H_\eta\{\boldsymbol{B}(w)^T\boldsymbol{\gamma}_j, \alpha_j\}$. Then, we define a smoothed loss function:

$$l^s(\boldsymbol{\gamma}; \boldsymbol{X}, Y, W) = \{Y - \boldsymbol{X}^T\boldsymbol{h}(W, \gamma)\}^2 + \rho\sum_{j=1}^{p}\{\boldsymbol{B}(W)^T\boldsymbol{\gamma}_j\}^2. \tag{2.3}$$

2.2.4 Estimation

We minimize the empirical mean of (2.3) to obtain an estimate of $\tilde{\boldsymbol{\gamma}}$:

$$\hat{\boldsymbol{\gamma}} = \arg\min_{\gamma} \mathrm{E}_n l^s(\boldsymbol{\gamma}; \boldsymbol{X}, Y, W).$$

Then the estimate for $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)^T$, where $\hat{\beta}_j = \zeta_{(\boldsymbol{B}^T\hat{\boldsymbol{\gamma}}_j, \alpha_j)}(w)$.

Computation of $\hat{\boldsymbol{\gamma}}$ can be implemented by gradient-based methods and a coordinate descent algorithm. With appropriate initial values, global optimizers can be reached. Specifically, for each $j = 1, \ldots, p$, we set the initial $\boldsymbol{\gamma}_j^{(0)}$ to be the sample correlation between Y and $X_j\boldsymbol{B}(W)$, i.e. $\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})X_{ij}\boldsymbol{B}(W_i)$, where $\bar{Y} = \sum_{i=1}^{n} Y_i/n$. We choose the pre-specified parameters as follows. As a value of α comparable to the scale of true coefficients works well, we set α_j to be the absolute value of the corresponding coefficient estimate from a parametric model. The choices of η and ρ can be specified in accordance with Condition (**Ch2.C6**) in Section 2.3.1. The knots of B-spline are equally spaced over $\mathbb{D}$. The number of basis functions, q, can be determined through R-fold cross-validation. That is, partition the full data D into R equal-sized groups, denoted by D_r, for $r = 1 \ldots, R$, and let $\hat{\boldsymbol{\beta}}_{-r}^{(q)}(W)$ be the estimate obtained with q bases using all the data except for D_r. We obtain the optimal q by minimizing the cross-validation error

$$\mathrm{CV}(q) = \sum_{r=1}^{R}\sum_{i\in D_r}\left\{Y_i - \boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}_{-r}^{(q)}(W_i)\right\}^2. \tag{2.4}$$

2.3 Inference

2.3.1 Asymptotic properties

Let $l_n = \mathrm{E}_n l^s(\boldsymbol{\gamma}; \boldsymbol{X}, Y, W)$, and denote by $l_n'(\boldsymbol{\gamma})$ and $l_n''(\boldsymbol{\gamma})$ the first and second derivatives of l_n with respect to $\boldsymbol{\gamma}$ respectively. It follows that $l_n'(\boldsymbol{\gamma}) = 2\mathrm{E}_n\{-(Y - \boldsymbol{X}^T\boldsymbol{h})\boldsymbol{U} \otimes \boldsymbol{B}(W) + \rho\boldsymbol{\theta} \otimes \boldsymbol{B}(W)\}$ and $l_n''(\boldsymbol{\gamma}) = 2\mathrm{E}_n\big[\big\{\boldsymbol{U}\boldsymbol{U}^T + \rho I_p - (Y - \boldsymbol{X}^T\boldsymbol{h})\Lambda\big\} \otimes (\boldsymbol{B}\boldsymbol{B}^T)\big]$, where $\boldsymbol{U} = \boldsymbol{U}(\boldsymbol{\gamma}; \boldsymbol{X}, W) = \{X_1 h_1'(\boldsymbol{\gamma}; W), \ldots, X_p h_p'(\boldsymbol{\gamma}; W)\}^T$, and $\Lambda = \mathrm{diag}(X_1 h_1'', \ldots, X_p h_p'')$ is a diagonal matrix. Let $\boldsymbol{V}_n = \boldsymbol{V}_n(\boldsymbol{\gamma})$ be the n by pq matrix $\{\boldsymbol{v}_1(\boldsymbol{\gamma}), \ldots, \boldsymbol{v}_n(\boldsymbol{\gamma})\}^T$ with $\boldsymbol{v}_i(\boldsymbol{\gamma}) = \boldsymbol{U}(\boldsymbol{\gamma}; \boldsymbol{X}_i, W_i) \otimes \boldsymbol{B}(W_i)$. Let $\hat{\theta}_j = \boldsymbol{B}(w)^T\hat{\boldsymbol{\gamma}}_j$ and $\tilde{\theta}_j = \boldsymbol{B}(w)^T\tilde{\boldsymbol{\gamma}}_j$. Technical conditions (**Ch2.C1**)–(**Ch2.C7**) are listed in the Supplementary Material.

Theorem 2.3.1 (Convergence Rate) *Under Conditions (**Ch2.C1**), (**Ch2.C4**), (**Ch2.C6**) and (**Ch2.C7**), given α_j $(j = 1, \ldots, p)$, if $\beta_{0j}(w) \in \mathbb{S}_{q,\alpha_j}$ for $j = 1, \ldots, \tilde{p}$ with q and α_j be the same as in $l(\boldsymbol{\gamma}; \boldsymbol{X}, Y, W)$, then*

$$||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0||_2 = O_p\left((\tilde{p}q/n)^{1/2}\right);$$

if $\beta_{0j}(w) \notin \mathbb{S}_{q,\alpha_j}$ for $j = 1, \ldots, \tilde{p}$,

$$||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0||_2 = O_p\left((\tilde{p}q/n)^{1/2} + \tilde{p}^{1/2}q^{-m}\right).$$

By Condition (**Ch2.C6**) and $m > 0{\cdot}5$, Theorem 2.3.1 implies convergence of $\hat{\beta}$. If $\tilde{p}$ is $O(1)$, Theorem 2.3.1 coincides with theories in nonparametric regression. If the true curves are in the thresholded sieve space, then there is no approximation error; and if q is $O(1)$, Theorem 2.3.1 suggests root-n consistency.

Let $\sigma_{nj}^2(w) = \sigma^2/n^2\{\boldsymbol{e}_j \otimes \boldsymbol{B}(w)\}^T\{l_n''(\tilde{\boldsymbol{\gamma}})\}^{-1}\left\{\boldsymbol{V}_n^T(\tilde{\boldsymbol{\gamma}})\boldsymbol{V}_n(\tilde{\boldsymbol{\gamma}})\right\}\{l_n''(\tilde{\boldsymbol{\gamma}})\}^{-1}\{\boldsymbol{e}_j \otimes \boldsymbol{B}(w)\}$, where $\boldsymbol{e}_j$ is p-dimensional vector with j-th entry being one and others being zero. We have the following theorem.

Theorem 2.3.2 *Under Conditions (**Ch2.C1**)–(**Ch2.C7**), then for any $w \in \mathbb{D}$, the limiting distribution of $\hat{\beta}_j(w) = \zeta_{\{\hat{\theta}_j, \alpha_j\}}(w)$ ($j = 1, \ldots, p$) satisfies*

$$\lim_{n\to\infty} \left| \Pr(\hat{\beta}_j(w) \le x) - G_{nj}(w, x) \right| \;=\; 0,$$

where $G_{nj}(w,x) = \Phi\left\{\frac{x+\alpha_j-\tilde{\theta}_j(w)}{\sigma_{nj}(w)}\right\}\mathcal{I}(x \ge 0) + \Phi\left\{\frac{x-\alpha_j-\tilde{\theta}_j(w)}{\sigma_{nj}(w)}\right\}\mathcal{I}(x < 0)$, and $\Phi(\cdot)$ is the cumulative distribution function for $N(0,1)$.

The limiting distribution in Theorem 2.3.2 reveals that the probability of $\hat{\beta}_j(w) = 0$ is greater than 0, which enables us to detect zero-effect regions even with finite sample size.

2.3.2 Sparse confidence intervals

We need to gauge the uncertainty of the point estimates and draw valid statistical inference on the selection and zero-effect region detection in high-dimensional varying coefficient models. Classical confidence intervals are non-applicable as the limiting distribution of the estimates involves zero point-mass. This motivates us to develop a new type of confidence interval for the varying coefficients with zero-effect regions.

Definition 2.3.1 (Sparse confidence interval) *For any $w \in \mathbb{D}$, let $u_n(w)$ and $v_n(w)$ be the lower and upper bound estimates of $\beta(w)$, and let $\xi \in (0,1)$.*

i) when $\beta(w) \neq 0$, $[u_n(w), v_n(w)]$ is a $(1-\xi)$ level sparse confidence interval if, for any $w \in \mathbb{D}$, $\lim_{n\to\infty} \Pr\{u_n(w) \le \beta(w) \le v_n(w)\} = 1-\xi$;

ii) when $\beta(w) = 0$, $[u_n(w), v_n(w)]$ is a $(1-\xi)$ level sparse confidence interval if there exists an integer $N > 0$, such that $\Pr\{u_n(w) = 0 \text{ or } v_n(w) = 0\} > 0$ for any $n > N$, and $\lim_{n\to\infty} \Pr\{u_n(w) \le \beta(w) \le v_n(w)\} \ge 1-\xi$.

When $\beta(w) = 0$, a sparse confidence interval allows the upper bound or the lower bound or both to be zero with a non-zero probability; see Figure 2.2. This unique

property distinguishes the sparse confidence interval from its classical counterpart and provides a useful means to draw inference on estimated zero-effect regions, which also differs from the post-selection inference [53, 79, 82].

The derivation of sparse confidence intervals utilizes Lemma 5 and can be found in the Supplementary Material. Under Conditions (**Ch2.C1**)-(**Ch2.C7**) and given α_j, for any $w \in \mathbb{D}$ we construct a pointwise $(1-\xi)$ level asymptotic sparse confidence interval for $\beta_{0j}(w)$, denoted by $[u_{nj}(w), v_{nj}(w)]$. Let $z_{\xi/2}$ and Φ be the $(1-\xi/2)$ quantile and the cumulative distribution function of $N(0,1)$, respectively, and $\hat{\sigma}_{nj}$ be σ_{nj} with $\tilde{\gamma}$ replaced by $\hat{\boldsymbol{\gamma}}$. Let $P_+ = \Pr\{\hat{\beta}_j(w) > 0\}$ and $P_- = \Pr\{\hat{\beta}_j(w) < 0\}$, which can be estimated by $\hat{P}_+ = 1 - \Phi\{(\alpha_j - \hat{\theta}_j)/\hat{\sigma}_{nj}\}$ and $\hat{P}_- = \Phi\{-(\alpha_j + \hat{\theta}_j)/\hat{\sigma}_{nj}\}$ using Theorem 2.3.2. We construct $[u_{nj}(w), v_{nj}(w)]$ as follows:

- if $\hat{P}_+ + \hat{P}_- \le \xi$, $u_{nj}(w) = v_{nj}(w) = 0$;
- else if $\hat{P}_+ < \xi/2$ and $\hat{P}_- < 1 - \xi/2$, $[u_{nj}(w), v_{nj}(w)] = \big[\hat{\beta}_j(w) - \hat{\sigma}_{nj}\hat{B}, 0\big]$ with $\hat{B} = \Phi^{-1}\big\{1 - \xi + \Phi(-\hat{\sigma}_{nj}^{-1}\alpha_j + \hat{\sigma}_{nj}^{-1}\hat{\theta}_j)\big\}$ and $\hat{\sigma}_{nj}(w)$ as defined in Lemma 5;
- else if $\hat{P}_- < \xi/2$ and $\hat{P}_+ < 1 - \xi/2$, $[u_{nj}(w), v_{nj}(w)] = \big[0, \hat{\beta}_j(w) + \hat{\sigma}_{nj}\hat{A}\big]$ with $\hat{A} = -\Phi^{-1}\big\{\xi - 1 + \Phi(\hat{\sigma}_{nj}^{-1}\alpha_j + \hat{\sigma}_{nj}^{-1}\hat{\theta}_j)\big\}$;
- else $[u_{nj}(w), v_{nj}(w)] = \big[\hat{\beta}_j(w) - \hat{\sigma}_{nj}z_{\xi/2}, \hat{\beta}_j(w) + \hat{\sigma}_{nj}z_{\xi/2}\big]$.

Theorem 2.3.3 *Under Conditions (**Ch2.C1**)-(**Ch2.C7**), $[u_{nj}(w), v_{nj}(w)]$ is a $(1-\xi)$ level sparse confidence interval of $\beta_{0j}(w)$ for $j = 1, \ldots, p$ and any $w \in \mathbb{D}$.*

2.4 Simulation Studies

2.4.1 Low dimensional covariates

With $p = 3$, we focus on the accuracy in estimation and inference. We compare with two competing methods: the regular B-spline method [27] and the local polynomial method [33]. We simulate data from (2.1), where W_i are generated from a

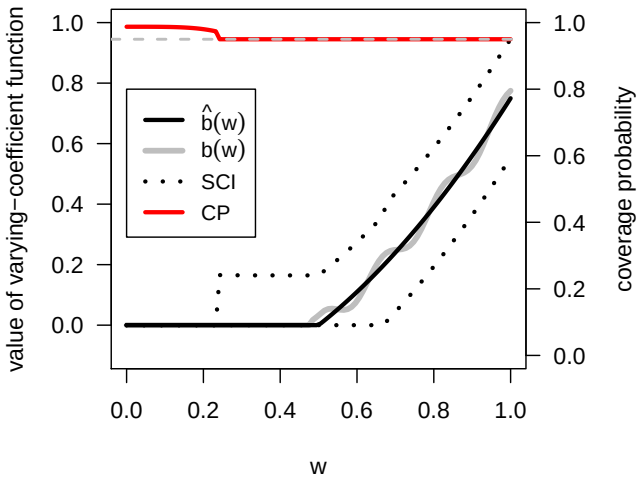


Figure 2.2: An illustrated example of sparse confidence intervals (SCI), with the true varying coefficient $\beta(w)$, estimation $\hat{\beta}(w)$ and coverage probability (CP). For $w \in [0, 0.23]$, $\Pr\{\hat{\beta}(w) = 0\} > 0.95$, then the 95% sparse confidence interval degenerates to $[0, 0]$, but with a coverage probability (CP) between 0.95 and 1.0 at each $w \in [0, 0.23]$. The coverage probability is 0.95 at every $w > 0.23$.

uniform distribution on $[0, 3]$, the covariates are generated from a multivariate normal distribution with mean zero and $\text{cov}(X_{ij}, X_{ij^*}) = 2\mathcal{I}(j = j^*) + 0{\cdot}5\mathcal{I}(j \neq j^*)$, and ϵ_i are generated from a standard normal distribution such that the noise to effect ratio is 0·1. The coefficient functions are $\beta_1(w) = (-w^2 + 3)\mathcal{I}(w \leq \sqrt{3})$, $\beta_2(w) = 2\log(w + 0{\cdot}01)\mathcal{I}(w \geq 1)$, and $\beta_3(w) = \{-6/(w + 1) + 2\}\,\mathcal{I}(w \leq 2)$.

We choose $n = 200, 500$ and 1,000 and replicate 200 times for each setting. We set $\eta = 0{\cdot}001$, $\rho = 1/n^2$, α_j to be half of the absolute value of the least-squares estimate. The number of knots, q, is selected through cross-validation. For evaluation criteria, we use the integrated squared errors and the averaged integrated squared errors, defined as $\text{ISE}(\beta_j) = n_g^{-1} \sum_{g=1}^{n_g} \{\hat{\beta}_j(w_g) - \beta_j(w_g)\}^2$ and $\text{AISE} = p^{-1} \sum_{j=1}^{p} \text{ISE}(\beta_j)$, respectively, where w_g $(g = 1, \ldots, n_g)$ are the grid points on $\mathbb{D}$. Table 2.1 summarizes the results, showing that the soft-thresholded varying coefficient model has smaller integrated squared errors and averaged integrated squared errors than the other two methods. Figure 2.3 compares the true coefficients and the median estimates ob-

Table 2.1: Simulation results for three models with $p = 3$

	n	ISE(β_1)	ISE(β_2)	ISE(β_3)	AISE
STV		21 (16)	21 (16)	22 (17)	21 (12)
B-spline	200	30 (19)	28 (18)	24 (16)	28 (13)
local polynomial		31 (15)	23 (13)	29 (16)	28 (10)
STV		7 (5)	7 (6)	8 (5)	8 (4)
B-spline	500	13 (6)	11 (7)	10 (6)	11 (5)
local polynomial		15 (6)	11 (5)	15 (8)	14 (4)
STV		4 (2)	4 (3)	4 (3)	4 (2)
B-spline	1000	8 (2)	6 (3)	4 (3)	6 (2)
local polynomial		9 (3)	7 (3)	9 (4)	8 (2)

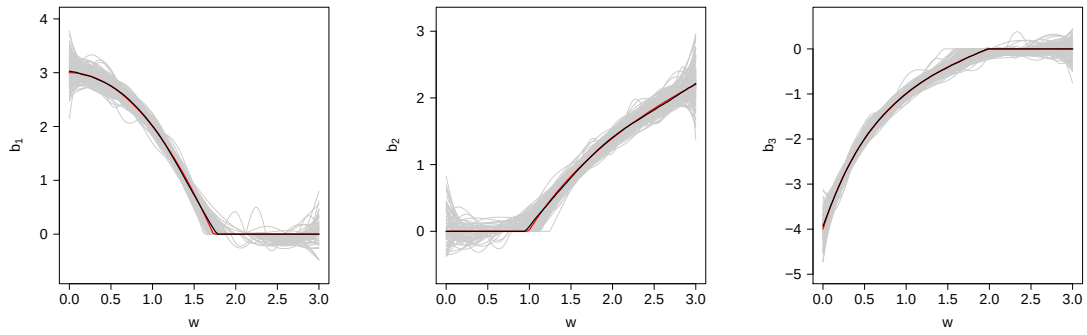
ISE: the integrated squared errors; AISE: the averaged integrated squared errors. Values are means and standard deviations from 200 replications and multiplied by 10^3.

tained by the competing methods. Only the median estimates obtained by the soft-thresholded varying coefficient model overlap with the truth, indicating the usefulness of our proposed method, particularly when estimating the zero-effect regions. Let $|A|$ denote the cardinality of set A. To compare zero-effect region detection, we define two quantities, estimation-based true positive ratio and estimation-based true negative ratio:

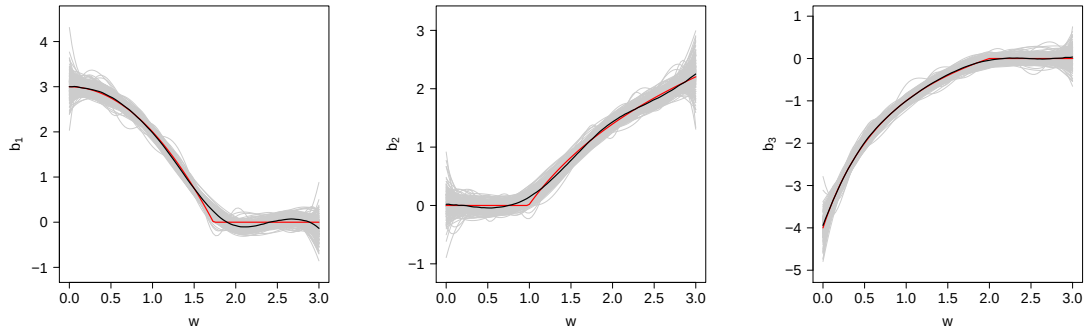
$$\text{ETPR}(\beta) = \frac{|\{w : \hat{\beta}(w) \neq 0 \text{ and } \beta(w) \neq 0\}|}{|\{w : \beta(w) \neq 0\}|},$$
$$\text{ETNR}(\beta) = \frac{|\{w : \hat{\beta}(w) = 0 \text{ and } \beta(w) = 0\}|}{|\{w : \beta(w) = 0\}|}.$$

Since the B-spline and local polynomial methods do not yield exactly zero estimates, the above definitions are not applicable. Instead, we introduce inference-based true positive ratio and true negative ratio:

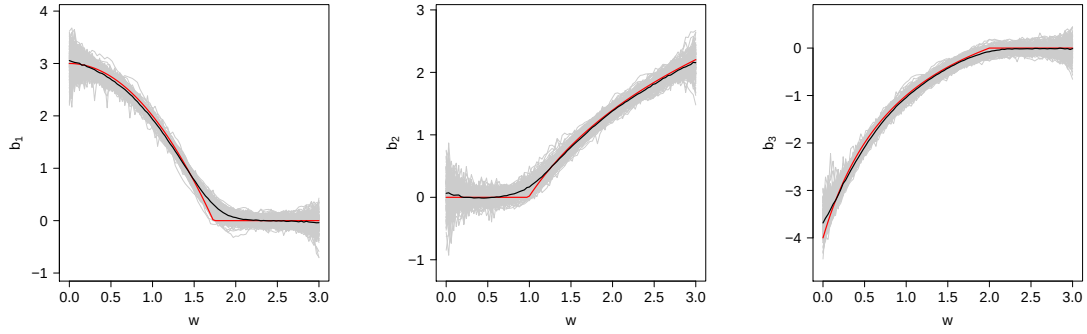
$$\text{ITPR}(\beta) = \frac{|\{w : 0 \notin \text{CI}\{\hat{\beta}(w)\} \text{ and } \beta(w) \neq 0\}|}{|\{w : \beta(w) \neq 0\}|},$$
$$\text{ITNR}(\beta) = \frac{|\{w : 0 \in \text{CI}\{\hat{\beta}(w)\} \text{ and } \beta(w) = 0\}|}{|\{w : \beta(w) = 0\}|},$$



(a) STV



(b) B-spline method



(c) local polynomial method

Figure 2.3: Comparisons of three methods. The red solid line is the true β curve, the gray solid lines are the estimated β curves, and the black solid line is the median of the estimated curves.

where $\mathrm{CI}\{\hat{\beta}(w)\}$ is the 95% confidence interval of $\beta(w)$.

We choose 100 grid points on $[0, 3]$ and count the number of W in each set as its cardinality. The Benjamini-Hochberg procedure [7] is adopted to control the false discovery rate in the calculation of the inference-based true positive ratio and the inference-based true negative ratio. Table 2.2 shows that the soft-thresholded varying coefficient model has higher values of the inference-based true negative ratio than the B-spline varying coefficient model and the local polynomial varying coefficient model. The inference-based true negative ratio of our method is improving as n becomes larger. Although the inference-based true positive and negative ratios are more reliable with controlled false discovery rates, their computational burden increases when the sample size becomes larger. Therefore, the estimation-based true positive and negative ratios are favorable for large datasets as their calculation merely depends on the estimations. In particular, we compare the non-zero-effect region selection accuracy between our estimation-based method and our inference-based method in Table 2.3. The estimation-based true positive ratio is slightly higher than the inference-based true positive ratio, but both of them closely approach to 1 as n increases. The estimation-based method is computationally much faster than the inference-based method.

Figure 2.4 shows the coverage probability of b_1 at each grid point for all three methods based on their 95% confidence intervals when $n = 500$. Among the three methods, the soft-thresholded varying coefficient model makes more accurate inference on zero-effect regions and non-zero-effect regions, as the coverage probabilities are closer to 95% on average compared to the others. At the transitions between zero and non-zero-effect regions, all the three methods draw less accurate inference, but our method still outperforms the competing methods. Specifically, the B-spline varying coefficient model and the local polynomial varying coefficient model have considerably small coverage probabilities around 50% to 60%, while our method can

Table 2.2: Comparisons of true positive ratios and true negative ratios among three methods for non-zero-effect region detection

n	Method	ITPR(β_1)	ITPR(β_2)	ITPR(β_3)	ITNR(β_1)	ITNR(β_2)	ITNR(β_3)
200	STV	936 (44)	919 (54)	816 (83)	987 (44)	967 (104)	976 (76)
	B-spline	977 (30)	930 (49)	833 (71)	928 (105)	952 (118)	969 (100)
	local polynomial	992 (23)	974 (38)	891 (78)	854 (141)	870 (161)	930 (127)
500	STV	962 (26)	949 (37)	883 (62)	990 (37)	980 (75)	985 (53)
	B-spline	993 (17)	970 (35)	897 (57)	876 (124)	954 (95)	967 (103)
	local polynomial	996 (12)	984 (24)	933 (54)	858 (112)	863 (123)	926 (133)
1000	STV	974 (18)	963 (25)	911 (48)	992 (24)	985 (45)	981 (69)
	B-spline	997 (9)	991 (15)	929 (43)	772 (152)	907 (129)	961 (90)
	local polynomial	996 (11)	989 (19)	951 (45)	857 (122)	836 (139)	921 (102)

ITPR: the inference-based true positive ratio; ITNR: the inference-based true negative ratio. Values are generated from 200 replications and multiplied by 10^3.

Table 2.3: Comparisons of true positive ratios and true negative ratios between the estimation-based method and the inference-based method using the soft-thresholded varying coefficient model for non-zero-effect region detection

		200	500	1000	2000	5000	10000
b_1	ETPR	997 (7)	998 (5)	997 (7)	997 (7)	999 (4)	1000 (2)
	ITPR	977 (14)	980 (12)	977 (14)	977 (14)	985 (10)	989 (9)
	ETNR	853 (125)	880 (104)	853 (125)	853 (125)	892 (100)	915 (84)
	ITNR	992 (30)	996 (18)	992 (30)	992 (30)	992 (27)	992 (28)
b_2	ETPR	989 (15)	989 (16)	989 (15)	989 (15)	992 (11)	993 (10)
	ITPR	962 (20)	963 (23)	962 (20)	962 (21)	972 (14)	975 (11)
	ETNR	900 (149)	872 (157)	900 (149)	900 (149)	955 (91)	981 (57)
	ITNR	991 (41)	990 (29)	991 (41)	991 (41)	994 (29)	999 (12)
b_3	ETPR	981 (30)	978 (33)	981 (30)	981 (30)	989 (20)	991 (16)
	ITPR	933 (42)	920 (40)	933 (42)	933 (42)	958 (31)	970 (24)
	ETNR	713 (282)	694 (267)	713 (282)	713 (282)	777 (266)	829 (265)
	ITNR	984 (51)	980 (64)	984 (51)	984 (51)	980 (55)	980 (60)

ETPR: the estimation-based true positive ratio; ITPR: the inference-based true positive ratio; ETNR: the estimation-based true negative ratio; ETPR: the inference-based true negative ratio. Values are multiplied by 10^3.

still achieve a coverage probability of at least 80%.

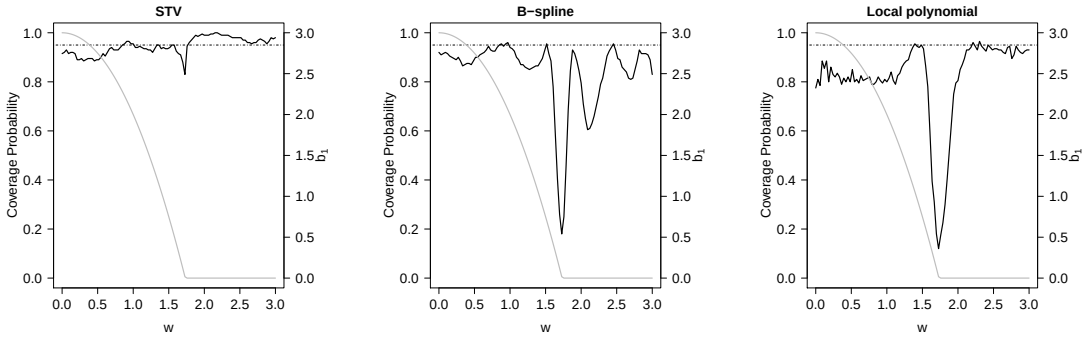


Figure 2.4: Empirical coverage probabilities (black curves) of the soft-thresholding varying coefficient model (STV), the regular B-spline varying coefficient model (B-spline) and the local polynomial varying coefficient model (local polynomial) in low dimensional covariates simulations. The grey curves are the true values of varying coefficients. The horizontal lines indicate the target coverage probability of 0·95.

2.4.2 High dimensional covariates

We focus on the variable selection and the prediction accuracy, and compare the soft-thresholded varying coefficient model to the penalized spline procedures with the group smoothly-clipped-absolute-deviation (SCAD) penalty and the group lasso penalty presented in Wei et al. (2011) [89]. We simulate data from (2.1), where W_i are generated from a uniform distribution on $[0, 3]$, the covariates are generated from a multivariate normal distribution with mean zero and covariance $\text{cov}(X_{ij}, X_{ij^*}) = \mathcal{I}(j = j^*)$ (independent) or $0{\cdot}5^{|j-j^*|}$ (autoregressive) or $\mathcal{I}(j = j^*) + 0{\cdot}5\mathcal{I}(j \neq j^*)$ (compound symmetry), and the random errors ϵ_i are generated from a standard normal distribution such that the noise to effect ratio is 0·1. The coefficient functions are $\beta_1(w) = -\beta_4(w) = 1{\cdot}2(-w^2 + 3)\mathcal{I}(w \leq \sqrt{3})$, $\beta_2(w) = -\beta_5(w) = 0{\cdot}8(-w^2 + 2)\mathcal{I}(w \geq \sqrt{2})$, and $\beta_3(w) = -\beta_6(w) = 2{\cdot}5\sin(w)$ and $\beta_j(w) = 0$ for $j = 7, \ldots, p$. We consider various (n, p): $(200, 250)$, $(500, 750)$ and $(1000, 1500)$. For each setting, a testing dataset with the same n is also generated. A total of 100 repetitions are made.

We use the R package *grpreg* [10] to implement the group SCAD penalized B-spline model, and the group lasso penalized B-spline model. The penalty tuning parameters are chosen through 10-fold cross-validation with a default option in the *grpreg* package. The number of knots q for B-spline is selected to be 12 and is fixed across all the methods for computational convenience. As the results are not sensitive to the choice of α_j in the soft-thresholded varying coefficient model, we set $\alpha_j = 2$ for all coefficients. Table 2.4 summarizes selection and estimation accuracy, including the total integrated squared errors between $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}_0$, which is defined as $\text{TISE} = \sum_{j=1}^{p} \text{ISE}(\beta_j)$, the predictive mean squared errors between y and $\hat{y}$ on the testing data, the number of false positives and false negatives, and the percentages of correct-fitting, over-fitting and under-fitting. Following Xue and Qu (2012) [97], we label a model correct-fitting if the selected set equals the true signal set, over-fitting if the selected set includes but is not equal to the true signal set, and under-fitting otherwise.

For $n = 200$, the soft-thresholded varying coefficient model has smaller total integrated squared errors and predictive mean squared errors; the percentages of correct-fitting of the soft-thresholded varying coefficient model are higher than those of the group lasso penalized model but lower than those of the group SCAD penalized model; the standard deviations of the number of false positives for the group SCAD penalized B-spline varying coefficient model and the group lasso penalized B-spline varying coefficient model are higher than those of the soft-thresholded varying coefficient model, indicating the soft-thresholded varying coefficient model is more stable than the other two methods for feature selection.

For $n = 500$, the soft-thresholded varying coefficient model outperforms the group SCAD penalized B-spline model and the group lasso penalized B-spline model with higher percentages of correct-fitting and fewer false positives; when comparing the total integrated squared errors and the predictive mean squared errors, the soft-

thresholded varying coefficient model is always better than the group lasso penalized model, and outperforms the group SCAD penalized model for the independent case, and has similar results as the group SCAD penalized method for the autoregressive and compound symmetry cases.

For $n = 1000$, the soft-thresholded varying coefficient model has 100% correct-fitting for all the three covariance matrices, indicating that with a large n, our method performs well even with complex covariance matrices.

We compare the computing time of the competing methods using the R package *grpreg* [10] on a laptop with a CPU of 2·7 GHz and a memory of 8 GB. Consider, for example, the case of independent covariates. When $n = 200$, the soft-thresholded varying coefficient model, the group SCAD penalized model and the group lasso penalized model respectively take 4·56, 12·48, and 5·36 seconds on average; when $n = 500$, they take 14·98, 25·03, and 23·24 seconds on average, respectively.

2.5 Analysis of the Preoperative Opioid Use Data

The motivating data on opioid use of preoperative patients were collected from 2010 to 2016, as part of the Michigan Genomics Initiative and Analgesic Outcome Study [43]. The raw data include 34,186 patients. After removing subjects with missing values in pre-surgical pain and other covariates of interest, the final analyzable data contain 13,787 patients, along with the records of preoperative opioid use and other characteristics before surgery; see Table 2.5 for the summary characteristics of the patients included in data analysis. Hilliard et al. (2018) [43] identified nine significant risk factors for preoperative opioid use, including, for example, pain severity, Fibromyalgia (FM) survey score (on a scale of 0 to 30 measuring centralized pain) and American Society of Anesthesiology score (ASA; on a scale of 0 to 4 measuring health conditions). Body mass index (BMI), which may reflect an individual's socioeconomic status [78] as well as overall fitness [2], is a major effect modifier for

Table 2.4: Simulation results under the high dimensional settings

cov(X)	Method	C (%)	O (%)	U (%)	FP	FN	TISE	PMSE
				$n = 200,\ p = 250$				
	STV	7	93	0	2·6 (1·7)	0·0 (0·0)	1·6 (0·6)	3·6 (0·6)
Ind	grscad	58	42	0	1·8 (4·8)	0·0 (0·0)	3·5 (1·7)	6·4 (3·5)
	grlasso	0	100	0	33·1 (11·7)	0·0 (0·0)	7·6 (1·6)	7·8 (1·4)
	STV	1	99	0	5·9 (2·6)	0·0 (0·0)	3·6 (1·1)	3·7 (0·7)
AR(1)	grscad	15	82	3	7·5 (7·6)	0·0 (0·2)	8·0 (5·3)	7·2 (4·6)
	grlasso	0	100	0	36·5 (13·1)	0·0 (0·0)	11·9 (1·8)	7·4 (1·2)
	STV	4	93	3	4·0 (2·8)	0·0 (0·2)	2·7 (1·6)	2·4 (0·8)
CS	grscad	69	27	4	2·2 (5·3)	0·1 (0·3)	4·8 (3·6)	4·6 (4·2)
	grlasso	0	100	0	40·1 (11·1)	0·0 (0·0)	9·2 (1·7)	4·8 (1·2)
				$n = 500,\ p = 750$				
	STV	99	1	0	0·0 (0·1)	0·0 (0·0)	0·3 (0·1)	2·7 (0·3)
Ind	grscad	66	34	0	6·0 (13·9)	0·0 (0·0)	0·7 (0·5)	2·9 (0·3)
	grlasso	0	100	0	48·2 (20·3)	0·0 (0·0)	2·9 (0·8)	3·8 (0·4)
	STV	88	12	0	0·1 (0·4)	0·0 (0·0)	0·9 (0·3)	2·4 (0·3)
AR(1)	grscad	62	38	0	5·5 (11·0)	0·0 (0·0)	0·7 (0·3)	2·2 (0·2)
	grlasso	0	100	0	67·9 (21·8)	0·0 (0·0)	5·2 (1·0)	3·5 (0·4)
	STV	59	40	1	1·2 (2·1)	0·0 (0·1)	0·8 (0·6)	1·6 (0·3)
CS	grscad	57	43	0	6·3 (10·9)	0·0 (0·0)	0·6 (0·3)	1·4 (0·2)
	grlasso	0	100	0	56·0 (18·3)	0·0 (0·0)	3·2 (0·8)	2·0 (0·3)
				$n = 1000,\ p = 1500$				
	STV	100	0	0	0·0 (0·0)	0·0 (0·0)	0·2 (0·0)	2·5 (0·2)
Ind	grscad	61	39	0	7·3 (16·8)	0·0 (0·0)	0·3 (0·3)	2·6 (0·2)
	grlasso	0	100	0	55·7 (24·7)	0·0 (0·0)	1·6 (0·6)	2·9 (0·2)
	STV	100	0	0	0·0 (0·0)	0·0 (0·0)	0·6 (0·1)	2·2 (0·2)
AR(1)	grscad	58	42	0	5·0 (11·3)	0·0 (0·0)	0·3 (0·2)	2·0 (0·1)
	grlasso	0	100	0	89·3 (30·5)	0·0 (0·0)	2·9 (0·7)	2·5 (0·2)
	STV	100	0	0	0·0 (0·0)	0·0 (0·0)	0·4 (0·1)	1·4 (0·1)
CS	grscad	59	41	0	4·8 (13·1)	0·0 (0·0)	0·3 (0·3)	1·3 (0·1)
	grlasso	0	100	0	64·3 (25·4)	0·0 (0·0)	1·7 (0·6)	1·5 (0·1)

STV: the soft-thresholded varying coefficient model; grscad: B-spline varying coefficient model with group SCAD penalty; grlasso: B-spline varying coefficient model with group lasso penalty; C: the percentage of correct-fitting; U: the percentage of under-fitting; O: the percentage of over-fitting; FP: the number of false positives; FN: the number of false negatives; TISE: the total integrated squared errors between $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}_0$; PMSE: the predictive mean squared errors between y and $\hat{y}$ on testing data; Ind, AR(1), and CS represent independent, autoregressive and compound symmetry correlation of covariates, respectively. Results are from 100 replications.

these risk factors. In our preliminary analysis, we fitted a varying coefficient model on the preoperative data with BMI being the index variable. In the model, the coefficient functions were expanded by a set of cubic B-spline basis functions, which were commonly used because of its nice approximation property [26]. Our preliminary analysis did show that the effects of these factors varied by BMI; see Figure 2.5, and Figures A.1-A.5. We suspect that zero-effect regions might exist around transition points and where the estimates were near 0. To properly characterize the possible BMI-dependent effects and identify the corresponding zero-effect regions, we apply the proposed soft-thresholded varying coefficient model.

Table 2.5: Patient Characteristics by Preoperative Opioid Use

Characteristics	No Preoperative Opioid Use ($n = 10,804$)	Preoperative Opioid Use ($n = 2,983$)
Age	52.74 (16.59)	53.14 (15.29)
BMI	29.58 (6.52)	30.48 (7.00)
Worst pain	2.02 (3.14)	4.60 (3.88)
Average pain	1.41 (2.30)	3.40 (3.05)
Fibromyalgia survey score	4.50 (4.02)	8.23 (5.16)
Life satisfaction	7.30 (2.55)	5.94 (2.67)
Male	4,996(46.2%)	1,329 (44.6%)
Depression	1,902 (17.6%)	1,153 (38.7%)
Race		
White	9,752 (90.3%)	2,658 (89.1%)
Black	457(4.2%)	191 (6.4%)
Asian	169 (1.6%)	17 (0.6%)
Other	426 (3.9%)	117 (3.9%)
Anxiety	3,746 (34.7%)	1,523 (51.1%)
Charlson Comorbidity Index		
$= 0$	8,074 (74.7%)	2,176 (72.9%)
$(0, 3)$	801 (7.4%)	365 (12.2%)
≥ 3	1,929 (17.9%)	442 (14.8%)
Alcohol	4,906 (45.0%)	1,218 (40.8%)
Apnea	2,461 (23.0%)	857 (28.7%)
Illicit drug use	369 (3.4%)	227 (7.6%)
Tobacco use	4,074 (37.7%)	1,613 (54.1%)
ASA score		
< 3	7,225 (66.9%)	1,535 (51.5%)
≥ 3	3,579 (33.1%)	1,448 (48.5%)

Continuous variables are presented in mean (standard deviation), and categorical variables are presented in count (percentage).

We consider the daily dose level of preoperative opioid use, measured in morphine milligram equivalents (MME), as the outcome Y in (2.1). The covariates $\boldsymbol{X}$ include categorical variables: sex, race, depression status, anxiety status, alcohol use, apnea status, illicit drug use, tobacco use, and ASA score; and continuous variables: age, worst pain score, Charlson comorbidity index (a weighted combination of comorbidity conditions), Fibromyalgia survey scores, average overall body pain score, and life satisfaction score (higher values meaning more satisfied with life). BMI, ranging from 15·0 to 55·0, is used as W. We set the initial values required by STV to be the estimates from a linear regression model and set the thresholding parameter α_j to be the half of the absolute value of the corresponding coefficient estimate from this model. The knots of B-spline are equally spaced over the range of BMI, while the number of basis functions, $q = 8$, is determined by minimizing the 10-fold cross-validation error in equation (2.4) over a candidate set $\{6, 8, 12, 16, 20, 24\}$. We set the penalty parameter ρ to be $1/n^2$. We also apply the local polynomial method and the regular B-spline method for comparisons. When implementing the regular B-spline method, we use the same spline bases as in STV. For the local polynomial method, we choose the bandwidth parameter using the same cross validation method and candidate set as in Fan and Zhang (2000) [32].

The STV method selects sex (female as the reference), race (white as the reference), worst pain score, Fibromyalgia survey score, depression, Charlson comorbidity index, alcohol use, apnea, illicit drug use, tobacco use, and ASA score (ASA < 3 as the reference group) into the final model. For the competing methods, the local polynomial method cannot detect zero-effect regions, and B-spline method has larger variation in the boundary. In contrast, STV has zero-effect region detection and stable boundary estimation. For example, the effect of worst pain score estimated by STV is statistically significant for the entire BMI range, while the B-spline method detects the significant effect of worst pain score only when BMI is between 18 and 40.

For FM survey score, STV is able to detect zero-effect region as BMI < 20 where the local polynomial could have false positive. For tobacco use, the estimated effect by the B-spline method switches signs without a transition region, which may not be biologically reasonable. In contrast, the effect estimated by STV contains a zero-effect region at the tail of the BMI range, which means the effect of tobacco use gradually disappears as BMI increases to extremity [5]. The details can be seen in Figure 2.5, and the results for the other factors are relegated to the Supplementary Material.

For ease of presentation, Table 2.6 summarizes the effects of risk factors based on the BMI categories and identifies several patterns related to opioid use. We use STV to identify the cut-offs for BMI and define four BMI categories, which largely coincide with the underweight, normal/overweight and obese categories as defined by WHO, except that our category includes a super obese group with BMI$>$ 49.5 (see Table 2.6 for details). When BMI is less than 18.0, ASA, alcohol, anxiety, and race (non-white versus white) have significantly positive effects on opioid use, indicating that underweight patients with severe systemic diseases, drinking history, and/or anxiety may tend to take more opioids than those without. Among patients with BMI between 30.0 and 49.5, FM, Tobacco use, illicit drug use, ASA, and race are all significantly associated with opioid use. When BMI is greater than 49.5, ASA, illicit drug use, and alcohol use are significantly associated with opioid use, suggesting that the supper obese patients with severe systemic disease, illicit drug use history, and/or drinking history likely take more opioids than others. Both pain and depression are significantly associated with increased opioid use for all patients, regardless of BMI levels. Some of our findings are consistent with the conclusions from the existing literature, for example, previous studies [6] have reported that the ASA category is significantly related to opioid use; and alcohol use may significantly increase the odds of opioid use only for underweight and obese patients, but has minimum or no effects among the normal weight or overweight patients.

Table 2.6: Effects of risk factors based on BMI categories.

	BMI			
	(<18)	(18.0-30.0)	(30.0- 49.5)	(> 49.5)
Worst Pain	+∗	+∗	+∗	+∗
FM	0	+∗	+∗	+∗
Tobacco use	+∗	+∗	+∗	0
ASA > 3	+∗	+∗	+∗	+
Illicit drug use	0	+∗	+∗	+∗
Apnea	0	+	0	0
Alcohol	+∗	0	0	+∗
Anxiety	+∗	0	0	+∗
Depression	+∗	+∗	+∗	+∗
Sex (male)	+∗	0	0	0
Age	0	0	0	0
Race (black)	+∗	+∗	0	+∗
Race (Asian)	+∗	0	0	0
Race (other)	+∗	0	+∗	0
Average overall pain	0	0	0	0
Life satisfaction	0	0	0	0
Comorbidity (>3)	+∗	0	0	+∗
Comorbidity (1-3)	+∗	+	+∗	0

0: no effects +: positive ∗: significant

To further confirm our findings, we conduct subgroup analyses by fitting linear regression within each BMI category; see Table A.1 in the Supplementary Material. Alcohol use is significantly associated with preoperative opioid use among the supper obese patients ($\hat{\beta} = 0.205, p = 0.046$), consistent with the STV results, while the other two methods fail to capture the association; FM score is not significant for the underweight population ($\hat{\beta} = 0.024, p = 0.079$), confirmed by STV and the B-spline method, while the local polynomial method gives a false positive estimation; both sub-group regression analysis and STV conclude that the illicit drug use is significantly associated with preoperative drug use among the overweight population ($\hat{\beta} = 0.084, p = 0.010$), while the other two methods do not detect this association.

In summary, leveraging a large-scale dataset, we have examined the conjectures proposed from the previous literature [20, 37, 44, 58, 77] and, in particular, elucidated the effect changes over BMI on opioid use. The obtained results can potentially inform pain management, aid in physicians' prescription, and eventually relieve the persistent use of opioids.

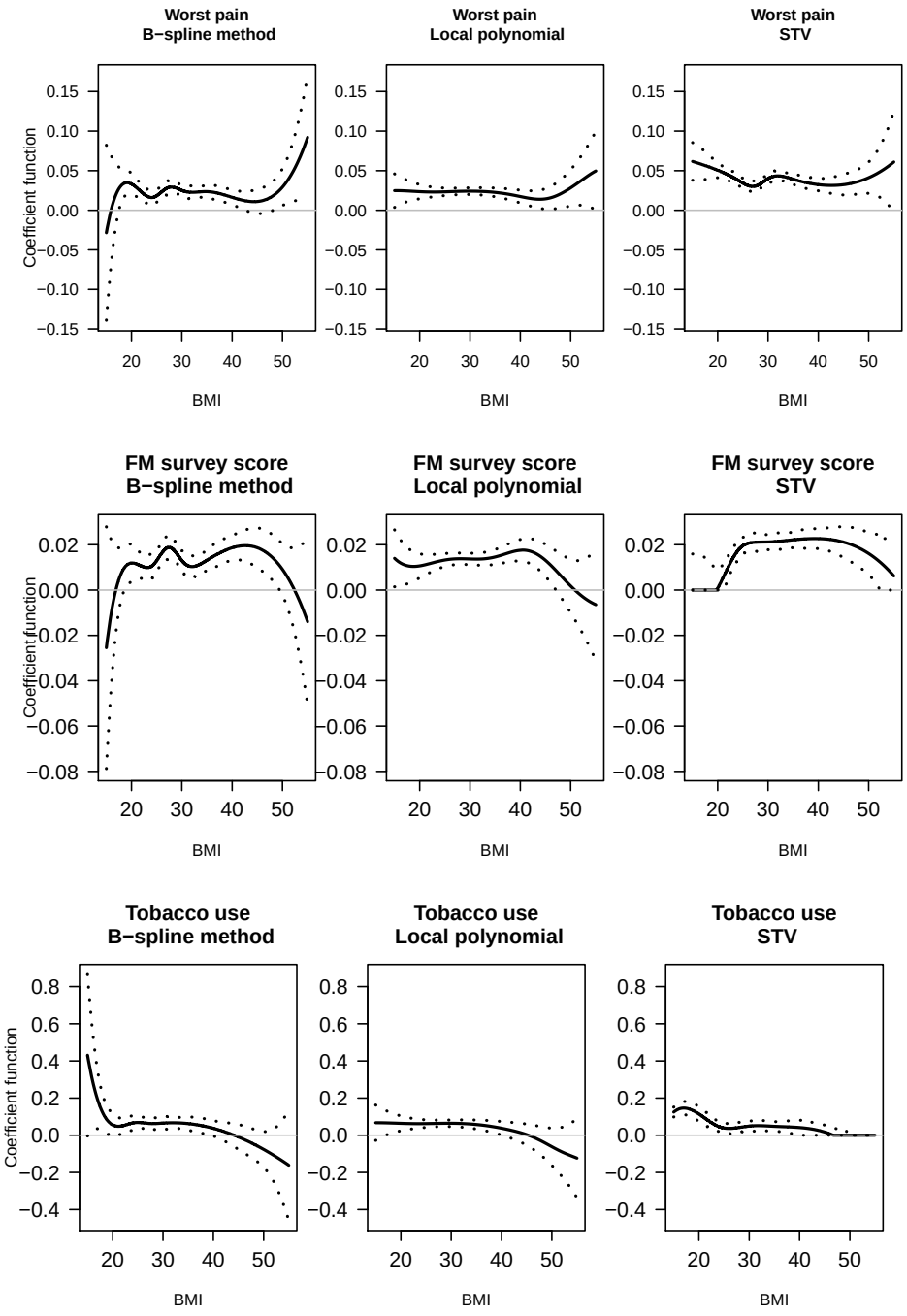


Figure 2.5: Estimation results (I) for the preoperative opioid use data using the B-spline method, the local polynomial method and the STV method: the black solid lines are the estimated coefficient function curves for each variable; the dotted lines are the pointwise (sparse) confidence intervals.

2.6 Discussion

To address the challenge of modeling varying coefficients with zero-effect regions, we proposed a new soft-thresholded varying coefficient model, where the varying coefficients are piecewise smooth with zero-effect regions. We have designed an efficient estimation method and a novel sparse confidence interval, which extends classical confidence intervals by accommodating the exact zero estimates. Our flexible framework enables us to perform variable selection and detect the zero-effect regions of selected variables simultaneously, and to obtain point estimates of the varying coefficients with zero-effect regions and construct the associated sparse confidence intervals. The future work lies in extending the model to accommodate more general settings and more types of data, such as discrete data, censored data, and functional data.

CHAPTER III

Generalized Dynamic Effect Change Model: An Interpretable Extension of GAM

3.1 Introduction

The varying coefficient model (VCM) [39] is commonly used to characterize the dynamic changes of regression effects. It is more powerful and flexible than classical linear regression models, as it considers the coefficient as a function instead of a constant value. The varying coefficient functions either depend on one index variable or multiple index variables [34, 48, 49, 50, 95]. Let Y be the outcome, $\boldsymbol{X} = (X_1, \ldots, X_p)^T$ the covariates, and $\boldsymbol{Z} = (Z_1 \ldots, Z_p)^T$ some other covariates that may or may not be the same as $\boldsymbol{X}$. With data $(Y, \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Z} = \boldsymbol{z})$, those models often have the form

$$g(m(\boldsymbol{x}, \boldsymbol{z})) = x_1 f_1(z_1) + \ldots + x_p f_p(z_p), \tag{3.1}$$

where g is some link function, and $m(\boldsymbol{x}, \boldsymbol{z}) = \mathrm{E}(Y|\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Z} = \boldsymbol{z})$. Some other works consider partially varying coefficient models, where part of the coefficients are constant and the remainders are varying functions [1, 30]. Xue and Qu (2012) [97] studied the marginal integration method for a slightly generalized version of model (3.1), where each $f_j(z_j)$ is replaced by a multivariate function $f_j(\boldsymbol{z}) = f_{j1}(z_1) + \ldots + f_{j1}(z_k)$ with an additive structure. Lee et al. (2012) [54] studied a fully extended version

of model (3.1), where each $f_j(Z_j)$ is replaced by a multivariate function $f_j(\boldsymbol{z}) = \sum_{k\in I_j} f_{jk}(z_k)$ with each index set I_j known and excluding j.

The aforementioned models are interaction models in the sense that each coefficient function depends on some other covariates. A more straightforward model is the generalized additive model (GAM) [38, 40, 96]. The GAM has the form

$$g(m(\boldsymbol{x})) = \alpha + f_1(x_1) + \ldots + f_p(x_p), \tag{3.2}$$

where g is some link function, and $m(\boldsymbol{x}) = \mathrm{E}(Y|\boldsymbol{X} = \boldsymbol{x})$. Constraints $\mathrm{E}\{f_j(x_j)\} = 0$ are commonly assumed for model identification. GAMs are widely used in practice (see, e.g., [29, 67]) with their popularity resting in part on the availability of statistically well founded smoothing parameter estimation methods that are numerically efficient and robust [91, 93] and perform the important task of estimating how smooth the component functions of a model should be. However, GAM often focuses on the prediction and the explanation of the model is unclear due to the constraints. For example, assume the true function is $g(m(\boldsymbol{x})) = \alpha_0 + f_1(x_1) + f_2(x_2)$, where $\alpha_0 = 1$, $f_1(x_1) = 1 + \sin(x_1)$, and $f_2(x_2) = 2 + \cos(x_2)$ with both x_1 and x_2 taking values on $[0, 2\pi]$. The GAM is trying to estimate $\tilde{\alpha}_0 = 4$, $\tilde{f}_1 = \sin(x_1)$ and $\tilde{f}_2 = \cos(x_2)$. The values of those parameters don't represent the true contributions of each covariate.

In our motivating data, a survey study about the use of preoperative opioids, we are interested in tangling the relationship between the opioid use and risk factors and then propose the optimal pain management plan. It is reasonable to assume some relationships are nonlinear. Although the above models have their advantages, limitations also exist: VCM assumes a linear relationship between the predictor and response, and the varying coefficient functions depend on other covariates; GAM has a good prediction but lacks a good explanation of functional components since the values of GAM estimations can not represent the actual relationships. Therefore,

we propose a new model termed as the generalized dynamic effect change model (GDECM), which directly estimates the derivative function of the nonlinear effect, $\partial f_j(x_j)/\partial x_j$. It can be viewed as an extension of GAM. The advantages of our model are constraint-free and having a good model explanation. Our estimations can provide a new way to understand the associations between the opioid use and the risk factors. For example, if we assume the association between the opioid use and BMI is nonlinear and the estimated $\partial f(\text{BMI})/\partial \text{BMI}$ is positive when BMI is 35, then for patients with BMI= 35, increase of BMI will increase the preoperative opioid use.

Conventional methods for estimating functions include B-splines, smoothing splines, and P-splines [87]. Here in this study, we apply the reproducing kernel Hilbert space (RKHS) basis to model the functions in our model. RKHS bases are widely used in the functional data analysis [99]. Ravikumar et al. (2009) [66] considered sparse additive model under a Hilbert space. In this paper, we examine the properties of RKHS in our model and show the estimation performances by solid theorems and simulation studies.

The remainder of the paper proceeds as follows. In Section 3.2, we propose a new generalized dynamic effect change model and describe the estimation method. Section 3.3 provides theoretical results. In Section 3.4, we conduct simulations to demonstrate the advantages of our method by comparing it with GAM. Section 3.5 includes an analysis of the Preoperative Opioid Use data and Section 3.6 concludes the paper with brief discussions. Theoretical proofs and technical derivations are in the Appendix for Chapter III.

3.2 Method

3.2.1 Generalized Dynamic Effect Change Model

Let i $(i = 1, \ldots, n)$ be the index of subjects, and $X_{ij} \in \mathbb{R}(j = 1, \ldots, p)$ the risk factor j for subject i. The response Y_i is assumed to follow a exponential family distribution with density

$$f_Y(Y_i; \psi, \phi) = \exp\left\{\frac{Y_i\psi - b(\psi)}{a(\phi)} + c(Y_i, \phi)\right\},$$

where ψ is the natural parameter and ϕ is the scale parameter. We propose the following generalized dynamic effect change model (GDECM):

$$g(\mu_i) = \alpha_0 + \sum_{j=1}^{p} \int_0^{X_{ij}} \{\alpha_j + \beta_j(x)\}\, \mathrm{d}x, \tag{3.3}$$

where $\mu_i = \mathrm{E}(Y_i|\boldsymbol{X}_i) = \dot{b}(\psi)$, $g(\cdot)$ is a monotonic differentiable link function, $\alpha_0, \ldots, \alpha_p$ are scalar parameters, and $\beta_1, \ldots, \beta_p$ are function parameters. Comparing model (3.3) with GAM, then

$$\alpha_j + \beta_j(x) = \partial f_j(x)/\partial x$$

is the effect change, and the parameter β_j is the dynamic part.

In practice, covariate X_j can be categorical or continuous. If X_j is categorical, the representation of the varying coefficients is easy to be formed as combinations of indicator functions. Then the estimation of the effects becomes the estimation of coefficients associated with indicator functions. The model is more complicated when X_j is continuous. Therefore, we will focus on the case when all X_j are continuous. We shall assume all X_j are bounded on $\mathbb{R}$. For the simplicity of notation, we assume $X_j \in \mathcal{T}$ for all j, and $\mathcal{T}$ is a compact subset of $\mathbb{R}$. Without loss of generality, let $\mathcal{T} = [0, 1]$. Let $\boldsymbol{\theta}(\boldsymbol{x}) = \{\alpha_0, \alpha_1, \ldots, \alpha_p, \beta_1(x_1), \ldots, \beta_p(x_p)\}^T$ be the parameter of

interest and $\boldsymbol{\theta}_0$ the true parameter. Estimation of $\boldsymbol{\theta}_0$ won't involve the scale parameter ϕ, so for simplicity ϕ will be assumed known and equal to 1. Let

$$h(\boldsymbol{X};\boldsymbol{\theta}) = g(\mu_i) = \alpha_0 + \sum_{j=1}^{p} \int_0^{X_{ij}} \{\alpha_j + \beta_j(x)\}\, \mathrm{d}x. \tag{3.4}$$

3.2.2 Estimation

We assume that β_j $(j = 1, \ldots, p)$ belong to a reproducing kernel Hilbert space (RKHS) $\mathbb{H}$, a subspace of the collection of square integrable functions on $\mathcal{T}$. More specifically, $\mathbb{H}$ is the Gaussian RKHS. Let $K_j(x,t) = \exp(-\sigma_j^2||x-t||_2^2)$ denote the Gaussian kernel, where $x, t \in \mathcal{T}$ and $\sigma_j > 0$ is a free parameter whose inverse $1/\sigma_j$ is called the width of K_j. Let $\{u_{jk}\}_{k=1}^{\infty}$ be the orthonormal eigenfunction set of K_j with respect to the integral operator and $\{\lambda_{jk}\}_{k=1}^{\infty}$ the corresponding eigenvalue set. Without loss of generality, we assume $\lambda_{j1} \geq \lambda_{j2} \geq \ldots \geq 0$. Following Mercer's theorem, the set $\{u_{jk}\}_{k=1}^{\infty}$ forms an orthonormal basis for $\mathbb{H}$. The Gaussian RKHS is defined as

$$\mathbb{H}_j = \left\{ \beta_j : \beta_j(x_j) = \sum_{k=1}^{\infty} \gamma_{jk} u_{jk}(x_j), ||\beta_j||_{K_j}^2 = \sum_{k=1}^{\infty} \frac{\gamma_{jk}^2}{\lambda_{jk}} < \infty \right\},$$

where $||\cdot||_{K_j}$ is the norm induced by K_j. To simplify the notation, let $B_{jk} = \sqrt{\lambda_{jk}} u_{jk}$, then

$$\mathbb{H}_j = \left\{ \beta_j : \beta_j(x_j) = \sum_{k=1}^{\infty} \gamma_{jk} B_k(x_j), ||\beta_j||_{K_j}^2 = ||\boldsymbol{\gamma}_j||_2^2 = \sum_{k=1}^{\infty} \gamma_{jk}^2 < \infty \right\}.$$

For the simplicity of notation, we shall assume that all $\mathbb{H}_j$ are the same. The extension to different $\mathbb{H}_j$ is straightforward and not of interest in this study. Without further clarification, we will drop j and denote $\mathbb{H}_j$ as $\mathbb{H}$. Therefore, the parameter space for our model is $\Theta = \mathbb{R}^{p+1} \times \mathbb{H}^p$.

Let q be an integer number that increases with n, $\boldsymbol{B}(x) = (B_1(x), \ldots, B_q(x))^T$ the

basis vector, and $\boldsymbol{\gamma}_j = (\gamma_1, \ldots, \gamma_q)^T$ the basis coefficients such that $\tilde{\beta}_j(x) = \boldsymbol{B}^T(x)\boldsymbol{\gamma}_j$. Consider a truncated RKHS,

$$\mathbb{H}_n = \left\{ \beta = \sum_{k=1}^{q} \gamma_k b_k : ||\beta||_K^2 = ||\boldsymbol{\gamma}||_2^2 = \sum_{k=1}^{q} \gamma_k^2 < \infty \right\}.$$

A common choice of q is $O(n^{1/5})$, which results in a truncation bias $||\beta_j - \tilde{\beta}_j||^2 = O(n^{-4/5})$ for the Sobolev space of order 2.

The sieve space for our model can be expressed as $\Theta_n = \mathbb{R}^{p+1} \times \mathbb{H}_n^p$. Let

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \Theta_n} \frac{1}{n} \sum_{i=1}^{n} l(Y_i, \boldsymbol{X}_i; \boldsymbol{\theta}), \tag{3.5}$$

where $l(Y_i, \boldsymbol{X}_i; \boldsymbol{\theta}) = Y_i h(\boldsymbol{X}_i; \boldsymbol{\theta}) - b\{h(\boldsymbol{X}_i; \boldsymbol{\theta})\}$, and $\dot{b}^{-1} = g$. We denote $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T$, and $\boldsymbol{\gamma} = (\alpha_0, \alpha_1, \ldots, \alpha_p, \boldsymbol{\gamma}_1^T, \ldots, \boldsymbol{\gamma}_p^T)^T$. Solving (3.5) over the sieve space Θ_n is equivalent to iteratively solving the problem at k iteration

$$\min_{\boldsymbol{\gamma}} \frac{1}{n} ||\boldsymbol{W}^{[k]^{1/2}} (\boldsymbol{U}\boldsymbol{\gamma} - \boldsymbol{z}^{[k]})||_2^2, \tag{3.6}$$

where $\boldsymbol{U} = (U(\boldsymbol{X}_1), \ldots, U(\boldsymbol{X}_n))^T$ with $U(\boldsymbol{X}_i) = (1, \boldsymbol{X}_i^T, \int_0^{X_{i1}} \boldsymbol{B}_1^T(x)\mathrm{d}x, \ldots, \int_0^{X_{ip}} \boldsymbol{B}_p^T(x)\mathrm{d}x)^T$, $\boldsymbol{z}^{[k]} = \boldsymbol{U}\boldsymbol{\gamma}^{[k]} + \boldsymbol{\Gamma}^{[k]}(\boldsymbol{Y} - \boldsymbol{\mu}^{[k]})$, $\boldsymbol{\Gamma}^{[k]}$ is a diagonal matrix with $\boldsymbol{\Gamma}_{ii}^{[k]} = g'(\mu_i^{[k]})$, $\boldsymbol{W}^{[k]}$ is a diagonal matrix with $\boldsymbol{W}_{ii}^{[k]} = \left\{ (\boldsymbol{\Gamma}_{ii}^{[k]})^2 V(\mu_i^{[k]}) \right\}^{-1}$ and $V(\mu_i)$ is the variance of Y_i.

3.3 Inference

We begin with some notation. Let

$$\boldsymbol{\theta}_0 = (\alpha_0, \ldots, \alpha_p, \beta_{01}, \ldots, \beta_{0p}) = \arg\min_{\boldsymbol{\theta} \in \Theta} \mathrm{E} l(Y, \boldsymbol{X}; \boldsymbol{\theta}),$$

$$\tilde{\boldsymbol{\theta}} = (\alpha_0, \ldots, \alpha_p, \tilde{\beta}_1, \ldots, \tilde{\beta}_p) = \arg\min_{\boldsymbol{\theta} \in \Theta_n} \mathrm{E} l(Y, \boldsymbol{X}; \boldsymbol{\theta}),$$

$$\text{and} \quad \hat{\boldsymbol{\theta}} = (\hat{\alpha}_0, \ldots, \hat{\alpha}_p, \hat{\beta}_1, \ldots, \hat{\beta}_p) = \arg\min_{\boldsymbol{\theta} \in \Theta_n} \frac{1}{n} \sum_{i=1}^{n} l(Y_i, \boldsymbol{X}_i; \boldsymbol{\theta}),$$

where $l(Y, \boldsymbol{X}; \boldsymbol{\theta}) = Yh(\boldsymbol{X}; \boldsymbol{\theta}) - b\{h(\boldsymbol{\theta})\}$. Let $\Lambda_{min}(M)$ and $\Lambda_{max}(M)$ be the smallest and maximum eigenvalues of matrix M respectively.

Conditions:

Ch3.C1 The covariates $\boldsymbol{X}$ take values in a bounded subset of $\mathbb{R}^p$.

Ch3.C2 The eigenvalues of $\mathrm{E}(\boldsymbol{X}\boldsymbol{X}^T)$ are bounded away from zero and infinity; that is, there are positive constants C_1 and C_2 such that $C_1 \leq \Lambda_{min}(\mathrm{E}(\boldsymbol{X}\boldsymbol{X}^T)) \leq \ldots \leq \Lambda_{max}(\mathrm{E}(\boldsymbol{X}\boldsymbol{X}^T)) \leq C_2$. Consequently, the eigenvalues of $\mathrm{E}(W_{ii}U_iU_i^T)$ are bounded away from zero and infinity.

Ch3.C3 The eigenvalues $\lambda_k^2 \lesssim k^{-2\eta}$ with $\eta > 1$ for $(k = 1, \ldots, \infty)$.

Ch3.C4 The number of bases $q = O(n^a)$ with $(2\eta - 1)/(8\eta) \leq a < 1$.

Condition **Ch3.C**1 requires that the support of each X_j is bounded, which is commonly assumed for asymptotic analysis in nonparametric regression [9]. Note that Condition **Ch3.C**1 implies $\mathrm{E}_{\boldsymbol{X}} h(\boldsymbol{X}; \boldsymbol{\theta}) < \infty$. Condition **Ch3.C**2 is regularity condition which was used in [33, 49]. Ravikumar (2009) [66] also requires Condition **Ch3.C**3 for the convergence of the estimator. Condition **Ch3.C**4 is required to control the size of parameter space.

Let $\boldsymbol{e}_{l,m}$ be the l-dimensional vector with the m-th element taken to be one and zero elsewhere, $\mathbf{0}_{p+1}$ the $p+1$ dimension vector of 0s, $\boldsymbol{c}_j(x) = (\mathbf{0}_{p+1}, \boldsymbol{e}_{p,j} \otimes \boldsymbol{B}(x))^T$, and

$$\boldsymbol{C}(\boldsymbol{x}) = (\boldsymbol{e}_{pq+p+1,0}, \boldsymbol{e}_{pq+p+1,1}, \ldots, \boldsymbol{e}_{pq+p+1,p}, \boldsymbol{c}_1(x_1), \ldots, \boldsymbol{c}_p(x_p))^T.$$

Then, the estimator of $\boldsymbol{\theta} = (\alpha_0, \ldots, \alpha_p, \beta_1(x_1), \ldots, \beta_p(x_p))^T$ is $\hat{\boldsymbol{\theta}} = \boldsymbol{C}\hat{\boldsymbol{\gamma}}$.

Theorem 3.3.1 *Suppose Conditions* ***Ch3.C1-Ch3.C4*** *hold, then*

$$||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0||_\infty = O_p(n^{-(2\eta-1)/(4\eta)} + (q/n)^{1/2} + 1/q^2).$$

Theorem 3.3.1 suggests that the larger the decay rate, the faster the convergence. The optimal number of bases is $q = O(n^{-1/5})$, which results in an approximation error $O(n^{-2/5})$. The result is consistent with Ravikumar (2009) [66].

Recall that $\boldsymbol{U} = (U_1, \ldots, U_n)^T$ with $U_i = (1, \boldsymbol{X}_i^T, \int_0^{X_{i1}} \boldsymbol{B}_1^T(x)\mathrm{d}x, \ldots, \int_0^{X_{ip}} \boldsymbol{B}_p^T(x)\mathrm{d}x)^T$. Let $\boldsymbol{\Gamma}$ be the diagonal matrix with $\boldsymbol{\Gamma}_{ii} = g^{'}(\hat{\mu}_i)$, $\boldsymbol{W}$ the diagonal matrix with $\boldsymbol{W}_{ii} = \left\{(\boldsymbol{\Gamma}_{ii})^2 \hat{V}(\hat{\mu}_i)\right\}^{-1}$, and $\hat{V}(\hat{\mu}_i)$ the variance estimator of Y_i.

Theorem 3.3.2 *Under Conditions* ***Ch3.C1-Ch3.C4****, for any* $\boldsymbol{x} = (x_1, \ldots, x_p)^T \in \mathcal{T}^p$*, we have*

$$\boldsymbol{\Sigma}^{-1/2}(\hat{\boldsymbol{\theta}}(\boldsymbol{x}) - \boldsymbol{\theta}_0(\boldsymbol{x})) \to_d N(0, \boldsymbol{I}), \quad \text{as } n \to \infty,$$

where $\boldsymbol{\Sigma} = \boldsymbol{C}^T(\boldsymbol{U}^T\boldsymbol{W}\boldsymbol{U})^{-1}\boldsymbol{C}$.

Proofs of Theorem 3.3.1 and 3.3.2 can be found in the Appendix of Chapter III. With Theorem 3.3.2, the derivation of the pointwise confidence interval is straightforward. Here we omit the derivation, but provide the results below. Let $z_{\xi/2}$ be the $1 - \xi/2$ quantile of the standard normal distribution:

(i) For any $j = 0, 1, \ldots, p$, the $1-\xi$ level confidence interval for α_j is $\hat{\alpha}_j \pm z_{\xi/2}\hat{\sigma}_j(\alpha_j)$, where $\hat{\sigma}_j(\alpha_j) = \sqrt{\boldsymbol{e}_{pq+p+1,j}^T(\boldsymbol{U}^T\boldsymbol{W}\boldsymbol{U})^{-1}\boldsymbol{e}_{pq+p+1,j}}$.

(ii) For any $j = 1, \ldots, p$ and any $x_j \in \mathcal{T}$, the $1 - \xi$ level confidence interval for $\beta_j(x_j)$ is $\hat{\beta}_j(x_j) \pm z_{\xi/2}\hat{\sigma}_j(\beta_j, x_j)$, where $\hat{\sigma}_j(\beta_j, x_j) = \sqrt{\boldsymbol{c}_j(x_j)^T(\boldsymbol{U}^T\boldsymbol{W}\boldsymbol{U})^{-1}\boldsymbol{c}_j(x_j)}$.

3.4 Simulations

In this section, we will show how GDECM performs under different settings and compare the results with GAM. We will use common βs throughout the simulations. The detailed functions are $\beta_1(x) = 1 + 12\sin(24x)$, $\beta_2(x) = 1 - 12\cos(24x)$, $\beta_3(x) = 1 + 12\sin(24x)$, and $\beta_4(x) = 1 - 12\cos(24x)$. We consider three types of covariance matrices for covariates: independent, compound symmetry and autoregressive. In independent case, X_1 to X_4 independently follow the uniform distribution $U(0, 1)$. In compound symmetry and autoregressive cases, we first let $\boldsymbol{X} \sim N(0, \Sigma)$ with diagonal entries being 1 and off-diagonal entries Σ_{ij} being ρ and $\rho^{|i-j|}$, respectively. Then we transform $\boldsymbol{X}$ to correlated uniform variables using copula transformation. In our simulations, ρ is set to be 0.3. The comparison metrics are the mean squared errors: $\text{MSE}(f_j) = \sum_{g=1}^{G}(f_j(X_{jg}) - \hat{f}(X_{jg}))^2/G$ and $\text{MSE}(\beta_j) = \sum_{g=1}^{G}(\beta_j(X_{jg}) - \hat{\beta}_j(X_{jg}))/G$ on a set of grid points $\{X_{j1}, \ldots, X_{jG}\}$ on $\mathcal{T}$ for $(j = 1, \ldots, p)$. Metric $\text{MSE}(f_j)$ stands for prediction accuracy and $\text{MSE}(\beta_j)$ is a measure of estimation accuracy.

3.4.1 Gaussian outcomes

We first simulate outcomes Y_i following Gaussian distribution using model

$$Y_i = \int_0^{X_{1i}} \beta_1(x)\mathrm{d}x + \int_0^{X_{2i}} \beta_2(x)\mathrm{d}x + \int_0^{X_{3i}} \beta_4(x)\mathrm{d}x + \int_0^{X_{4i}} \beta_4(x)\mathrm{d}x + e_i,$$

where $e_i \sim N(0, \sigma^2)$ and σ is chosen such that the noise to signal ratio is 0.2. GAM estimation is obtained using R package *mgcv* with default optimizing options. The results are summarized in Table 3.1 and 3.2. In both tables, GDECM has smaller mean squared errors than GAM. Therefore, GDECM has better better estimation and prediction accuracy compared to GAM when the model outcome following normal distributions.

Figure 3.1 plots the estimation results from GDECM and GAM, under the setting

Table 3.1: Comparisons of mean squared errors of β from GAM and GDECM for Gaussian Outcome.

Covariance	n	Model	β_1	β_2	β_3	β_4
	200	GDECM	4.55 (1.8)	6.62 (2.3)	4.6 (1.8)	6.60 (2.6)
		GAM	10.48 (7.0)	13.13 (5.2)	10.54 (7.5)	12.87 (3.8)
Ind	500	GDECM	2.85 (0.6)	4.74 (1.0)	2.84 (0.5)	4.81 (1.0)
		GAM	8.22 (1.7)	10.48 (2.1)	8.32 (1.8)	10.61 (2.0)
	1000	GDECM	2.35 (0.3)	4.31 (0.7)	2.36 (0.3)	4.35 (0.8)
		GAM	7.65 (1.3)	9.89 (1.4)	7.84 (1.3)	9.89 (1.4)
	200	GDECM	5.70 (2.3)	8.02 (3.5)	5.76 (2.5)	7.79 (3.2)
		GAM	12.65 (11.6)	14.66 (6.8)	13.04 (12.1)	14.95 (7.9)
CS	500	GDECM	3.16 (0.8)	5.25 (1.4)	3.26 (0.9)	5.26 (1.5)
		GAM	8.55 (2.0)	10.82 (2.3)	8.56 (2.0)	10.91 (2.2)
	1000	GDECM	2.54 (0.4)	4.46 (0.8)	2.52 (0.4)	4.49 (0.8)
		GAM	7.8 (1.4)	10.05 (1.4)	7.8 (1.4)	10.07 (1.5)
	200	GDECM	5.32 (2.1)	7.21 (2.5)	5.16 (2.1)	7.19 (2.7)
		GAM	11.97 (9.9)	14.13 (6.1)	11.28 (7.5)	13.65 (5.4)
AR(1)	500	GDECM	3.05 (0.7)	5.08 (1.3)	3.04 (0.7)	5.05 (1.2)
		GAM	8.35 (1.9)	10.74 (2.2)	8.37 (1.9)	10.67 (2.1)
	1000	GDECM	2.49 (0.4)	4.47 (0.8)	2.47 (0.3)	4.42 (0.8)
		GAM	7.83 (1.4)	9.99 (1.5)	7.83 (1.4)	9.9 (1.4)

Ind: independent covariance; CS: compound symmetry covariance; AR(1): autoregressive covariance.

with Gaussian outcome, independent covariate covariance, and sample size of 500. It shows that GDECM estimations are more stable and accurate than GAM estimations.

3.4.2 Poisson outcomes

We then simulate outcomes Y_i following Poisson(λ_i), where

$$\log(\lambda_i) = \int_0^{X_{1i}} \beta_1(x)\mathrm{d}x + \int_0^{X_{2i}} \beta_2(x)\mathrm{d}x + \int_0^{X_{3i}} \beta_4(x)\mathrm{d}x + \int_0^{X_{4i}} \beta_4(x)\mathrm{d}x.$$

The noise-to-signal ratio is approximately 0.2. GAM estimation is obtained through R package *mgcv* with default optimizing options. Table 3.3 and 3.4 summarize the

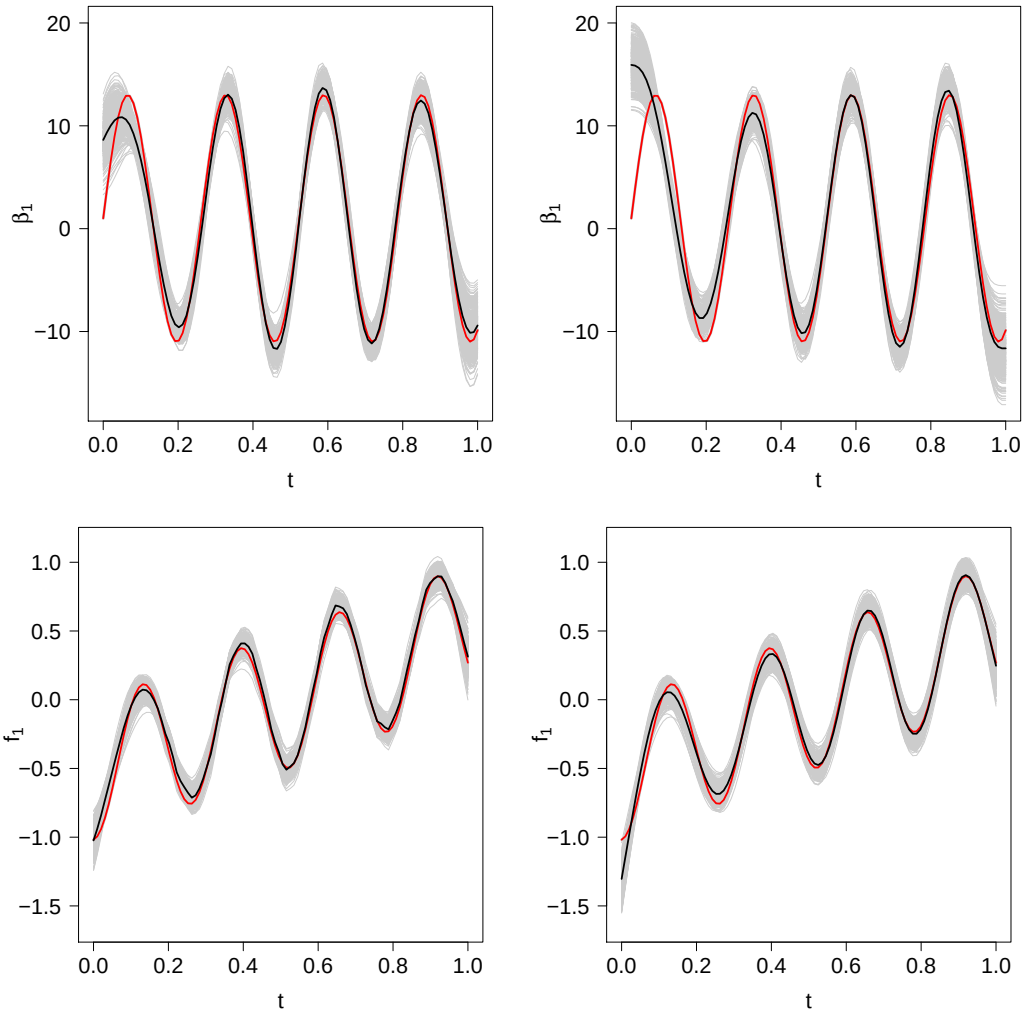


Figure 3.1: Comparisons of simulation estimations for one effect between GDECM (left) and GAM (right) with Gaussian outcome and independent covariate covariance ($n = 500$): gray lines are 200 estimations, black lines are their mean, and red lines are the truth.

Table 3.2: Comparisons of mean squared errors of f from GAM and GDECM for Gaussian Outcome.

Covariance	n	Model	f_1	f_2	f_3	f_4
	200	GDECM	8.42 (3.2)	11.69 (3.6)	8.50 (3.3)	11.77 (3.9)
		GAM	11.79 (11.9)	12.83 (8.8)	12.16 (12.7)	12.28 (4.6)
Ind	500	GDECM	4.24 (1.1)	7.36 (1.5)	4.23 (1.1)	7.41 (1.6)
		GAM	5.65 (1.5)	6.47 (1.8)	5.76 (1.5)	6.56 (1.7)
	1000	GDECM	3.07 (0.5)	6.19 (1.0)	3.05 (0.5)	6.24 (1.0)
		GAM	4.23 (0.9)	4.97 (0.9)	4.27 (0.9)	4.95 (0.9)
	200	GDECM	11.36 (4.7)	14.92 (5.8)	11.61 (4.8)	14.61 (5.3)
		GAM	17.04 (20.3)	16.43 (11.8)	17.73 (21.1)	16.78 (14.7)
CS	500	GDECM	5.21 (1.6)	8.60 (2.1)	5.34 (1.7)	8.47 (2.2)
		GAM	6.60 (1.9)	7.57 (2.1)	6.69 (1.9)	7.39 (2.1)
	1000	GDECM	3.53 (0.7)	6.75 (1.2)	3.50 (0.7)	6.76 (1.3)
		GAM	4.64 (1.0)	5.38 (1.1)	4.65 (1.0)	5.45 (1.2)
	200	GDECM	10.32 (4.1)	13.44 (4.3)	10.06 (3.9)	13.27 (4.6)
		GAM	15.01 (17.3)	15.01 (10.6)	13.72 (13)	14.30 (9.0)
AR(1)	500	GDECM	4.78 (1.3)	8.07 (1.9)	4.88 (1.4)	8.04 (1.8)
		GAM	6.18 (1.7)	7.15 (2.1)	6.38 (1.8)	7.11 (1.8)
	1000	GDECM	3.38 (0.7)	6.63 (1.2)	3.33 (0.7)	6.54 (1.1)
		GAM	4.57 (1.0)	5.30 (1.0)	4.49 (1.0)	5.14 (1.0)

Ind: independent covariance; CS: compound symmetry covariance; AR(1): autoregressive covariance. The mgcv R package is used to estimate GAM model. Values are multiplied by 10^3.

mean squared errors from 200 simulations. Both Table 3.3 and 3.4 show that the mean squared errors in GDCEM are smaller than that in GAM, showing that GDECM has better estimation accuracy and prediction accuracy than GAM under Poisson setting.

Figure 3.2 plots the estimation results from GDECM and GAM with Poisson outcome, independent covariate covariance, and sample size of 500. From the figure, GDECM estimations has smaller empirical standard errors and smaller biases than GAM estimations. Therefore, we can conclude that GDECM is more stable and more accurate than GAM.

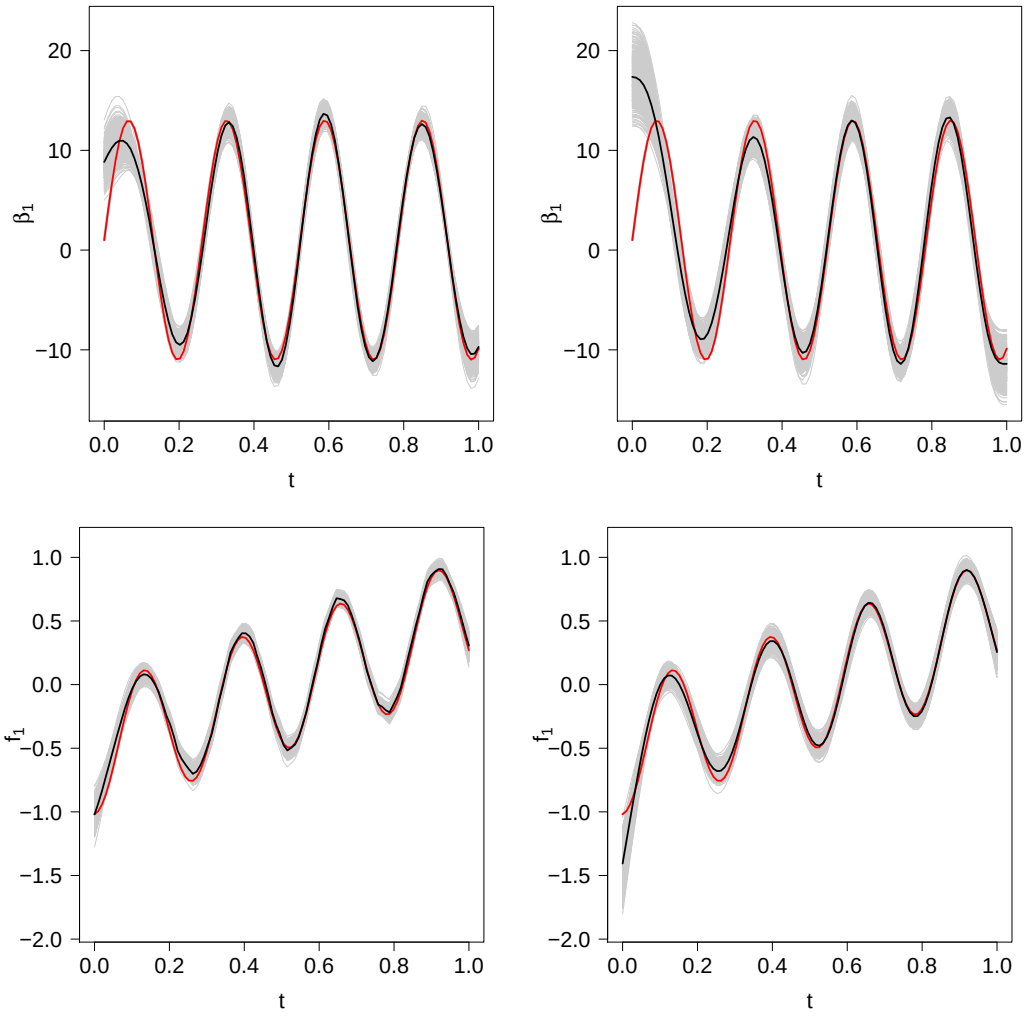


Figure 3.2: Comparisons of simulation estimations for one effect between GDECM (left) and GAM (right) with Poisson outcome and independent covariate covariance ($n = 500$): gray lines are 200 estimations, black lines are their mean, and red lines are the truth.

Table 3.3: Comparisons of mean squared errors of β from GAM and GDECM for Poisson Outcome

Covariance	n	Model	β_1	β_2	β_3	β_4
	200	GDECM	3.28 (0.9)	4.95 (1.2)	3.27 (1.0)	4.96 (1.1)
		GAM	10.46 (3.6)	11.77 (3.1)	10.31 (3.6)	11.61 (3.0)
Ind	500	GDECM	2.41 (0.3)	3.92 (0.5)	2.39 (0.3)	3.93 (0.5)
		GAM	9.34 (2.3)	9.77 (1.7)	9.24 (2.2)	9.68 (1.6)
	1000	GDECM	2.17 (0.2)	3.68 (0.3)	2.17 (0.2)	3.65 (0.3)
		GAM	9.00 (1.5)	9.21 (1.0)	8.90 (1.5)	9.09 (1.0)
	200	GDECM	3.38 (1.0)	4.93 (1.1)	3.43 (1.1)	4.90 (1.3)
		GAM	11.48 (4.4)	11.79 (3.5)	11.63 (4.4)	11.45 (2.8)
CS	500	GDECM	2.47 (0.4)	3.94 (0.5)	2.43 (0.3)	3.87 (0.4)
		GAM	10.31 (2.7)	9.68 (1.8)	10.40 (2.8)	9.64 (1.7)
	1000	GDECM	2.25 (0.2)	3.64 (0.2)	2.23 (0.2)	3.63 (0.2)
		GAM	10.03 (2.0)	8.96 (1.0)	10.09 (2.0)	8.98 (1.0)
	200	GDECM	3.26 (1.0)	4.95 (1.2)	3.35 (0.9)	4.86 (1.1)
		GAM	10.71 (3.8)	11.81 (3.6)	11.04 (4.1)	11.69 (3.3)
AR(1)	500	GDECM	2.47 (0.4)	3.90 (0.4)	2.44 (0.3)	3.86 (0.4)
		GAM	9.78 (2.5)	9.67 (1.6)	10.01 (2.6)	9.62 (1.6)
	1000	GDECM	2.22 (0.2)	3.64 (0.2)	2.20 (0.2)	3.63 (0.2)
		GAM	9.44 (1.8)	8.98 (1.1)	9.83 (1.8)	9.10 (1.0)

Ind: independent covariance; CS: compound symmetry covariance; AR(1): autoregressive covariance.

3.4.3 Coverage probability

Conventionally, the inference for GAM is often conducted by the Bayesian approach [92]. In this section, we compare our confidence interval with the classical Bayesian approach. Figure 3.3 shows that the Bayesian approach always has larger coverage probabilities compared to level 0.95, indicating an overestimate of the variance. In contrast, our method gives close to 0.95 coverage probabilities, showing the validity of our inference approach.

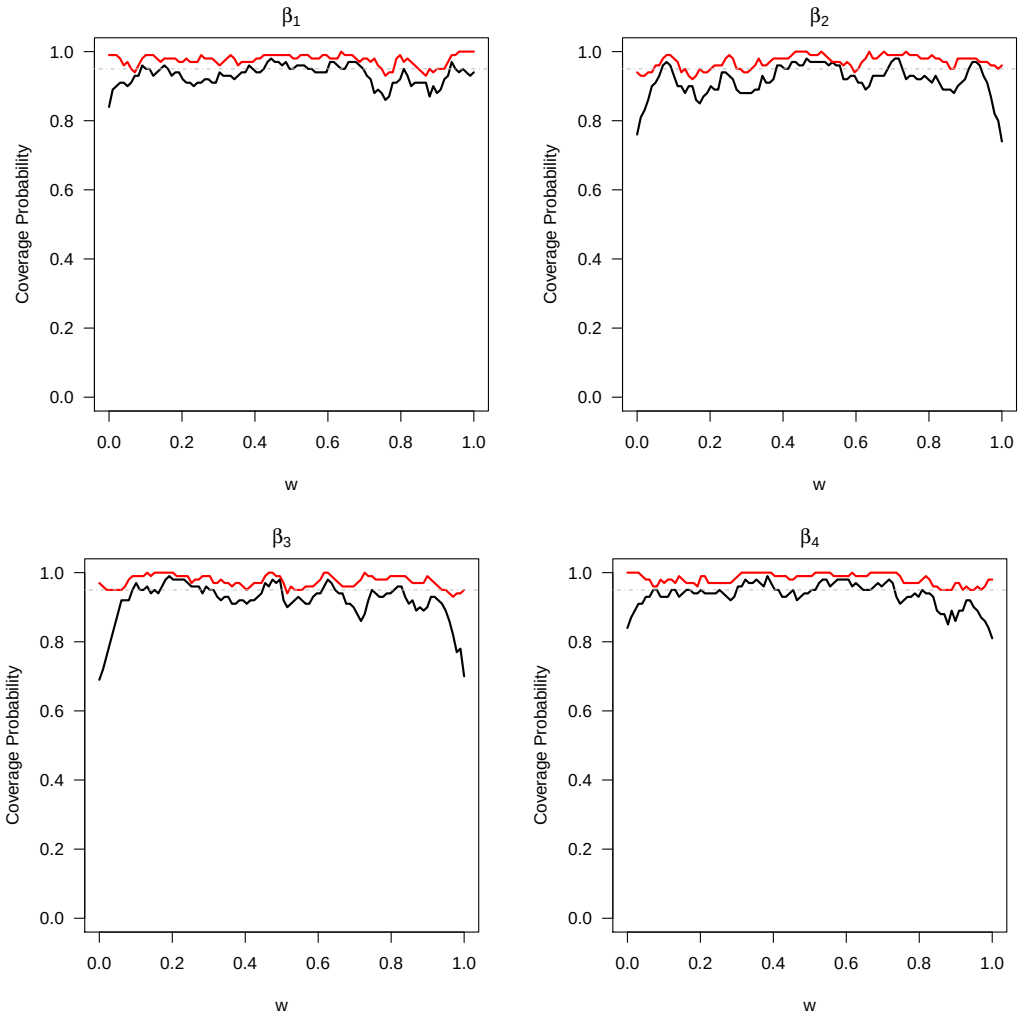


Figure 3.3: Comparisons of coverage probability between the Bayesian approach (in red) and GDECM (in black).

Table 3.4: Comparisons of mean squared errors of f from GAM and GDECM for Poisson Outcome

Covariance	n	Model	f_1	f_2	f_3	f_4
	200	GDECM	5.39 (1.8)	8.37 (2.5)	5.42 (1.9)	8.17 (2.4)
		GAM	9.52 (4.6)	9.72 (3.7)	9.55 (4.8)	9.55 (3.8)
Ind	500	GDECM	3.26 (0.7)	5.93 (1.0)	3.22 (0.6)	5.99 (1.1)
		GAM	5.76 (2.0)	5.80 (1.5)	5.63 (2.2)	5.64 (1.5)
	1000	GDECM	2.69 (0.4)	5.36 (0.7)	2.70 (0.4)	5.36 (0.6)
		GAM	4.65 (1.3)	4.58 (0.8)	4.62 (1.2)	4.52 (0.8)
	200	GDECM	5.69 (1.9)	8.31 (2.9)	6.00 (2.5)	8.67 (3.4)
		GAM	10.70 (5.3)	10.33 (4.4)	10.84 (5.9)	10.10 (4.0)
CS	500	GDECM	3.45 (0.8)	6.03 (1.2)	3.48 (0.9)	5.95 (1.2)
		GAM	6.58 (2.7)	6.10 (1.7)	6.70 (2.8)	6.08 (1.7)
	1000	GDECM	2.88 (0.5)	5.43 (0.8)	2.89 (0.5)	5.42 (0.8)
		GAM	5.50 (1.9)	4.85 (0.9)	5.49 (1.8)	4.89 (0.9)
	200	GDECM	5.40 (2.0)	8.55 (2.9)	5.85 (2.5)	8.11 (2.5)
		GAM	9.77 (4.9)	10.25 (4.4)	10.06 (4.8)	9.75 (3.9)
AR(1)	500	GDECM	3.36 (0.8)	6.09 (1.4)	3.52 (0.8)	5.76 (1.0)
		GAM	6.09 (2.3)	5.94 (1.6)	6.36 (2.7)	5.77 (1.7)
	1000	GDECM	2.71 (0.4)	5.56 (0.8)	2.91 (0.5)	5.21 (0.7)
		GAM	5.04 (1.6)	4.74 (0.9)	5.28 (1.6)	4.69 (0.8)

Ind: independent covariance; CS: compound symmetry covariance; AR(1): autoregressive covariance. Values are multiplied by 10^3.

3.5 Analysis of Preoperative Opioid data

We apply the generalized linear model (GLM) and GDECM to analyze the preoperative opioid use data collected from 2010 to 2016, as part of the Michigan Genomics Initiative and Analgesic Outcome Study [43]. The data include 13, 787 patients, along with the records of the preoperative opioid use and other characteristics before surgery. Table 2.5 in Chapter II summarizes the descriptive statistics of the data. Hilliard et al. (2018) [43] identified nine significant risk factors for preoperative opioid use, including pain severity, Fibromyalgia survey score, American Society of Anesthesiology score, etc. However, they did not consider the case that part of the

effects has nonlinear forms. Therefore, we consider GDECM with partially nonlinear effects.

We consider the daily dose level of preoperative opioid use, measured in morphine milligram equivalents (MME), as the Poisson outcome Y in (3.3). The covariates $\boldsymbol{X}$ include categorical variables: male, race, depression status, anxiety status, alcohol use, apnea status, illicit drug use, tobacco use, and American Society of Anesthesiologists category; continuous variables: age, BMI, worst pain score, Charlson Comorbidity Index, Fibromyalgia survey scores, average overall body pain score, and life satisfaction score. We only consider the effects of age and BMI nonparametric and others constants.

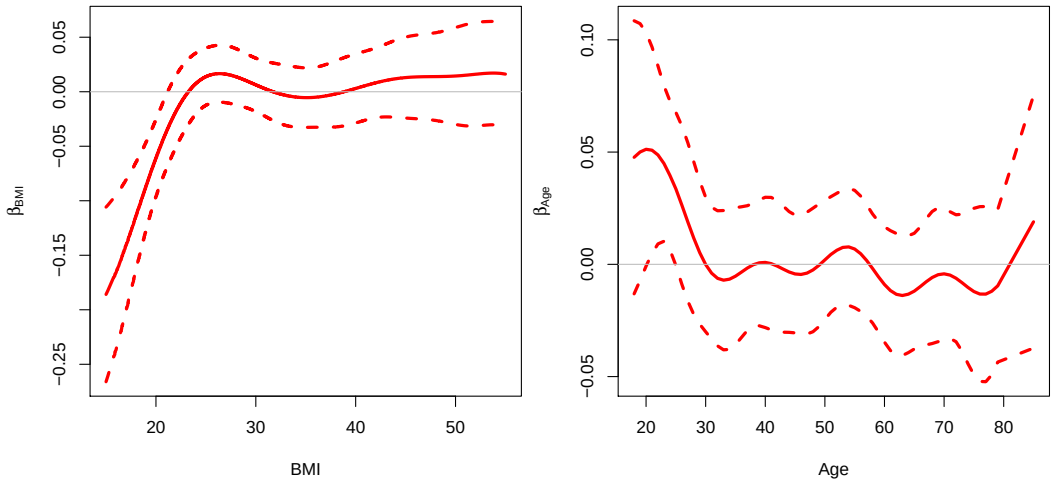


Figure 3.4: Effect coefficients of BMI and age from GDECM for the preoperative opioids data. Solid lines are estimation, and dashed lines are confidence intervals.

Figure 3.4 shows the dynamic effect changes of BMI and age from GDECM. The effect of BMI is significantly negative when BMI < 22, indicating that an increase in patients' BMI results in a decrease in the preoperative opioid use when BMI is smaller than 22. The effect change of age is not significant over the most age region. Figure 3.5 includes all other non-varying coefficients from GDECM. After adjusting for BMI and age in GDECM, variables with significantly positive coefficients include ASA, smoke, illicit drug use, depression, FM score, pain, and male. Adjusting for

Table 3.5: Comparisons of other regression estimates between GLM and GDECM for the preoperative opioids data.

	GDECM		GLM	
	β	sd(β)	β	sd(β)
Male	0.08	0.04	0.08	0.03
Black	0.01	0.07	0.02	0.07
Asian	-0.50	0.24	-0.52	0.23
Other race	-0.03	0.09	-0.03	0.08
Pain	0.13	0.01	0.13	0.01
FM	0.06	0.00	0.06	0.00
Satisfaction	-0.04	0.01	-0.04	0.01
Depression	0.14	0.05	0.14	0.04
Anxiety	-0.02	0.04	-0.01	0.04
Comorbidities $>$ 3	0.04	0.05	0.03	0.05
Alcohol	-0.08	0.04	-0.06	0.03
Apnea	-0.02	0.04	-0.02	0.04
Drug	0.22	0.07	0.22	0.06
Smoke	0.25	0.04	0.26	0.03
ASA	0.33	0.04	0.33	0.04

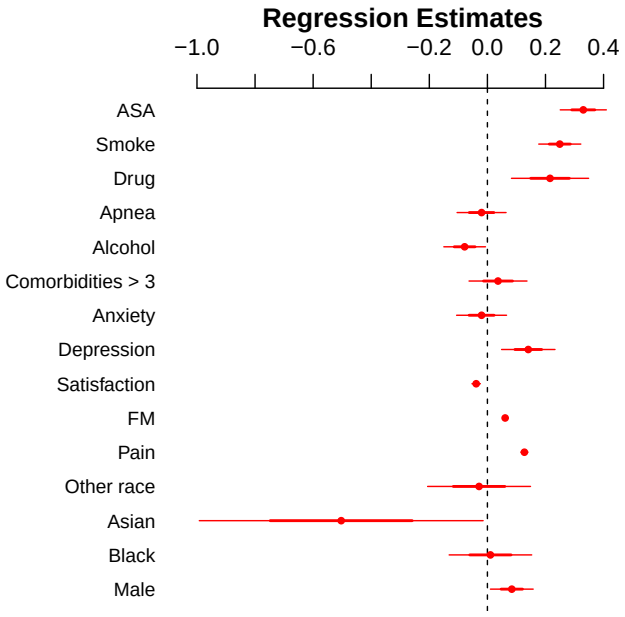


Figure 3.5: Other regression estimates from GDECM for the preoperative opioids data.

other variables, an increase in those variables lead to an increase in preoperative opioid use. The coefficients for alcohol and life satisfaction are significantly negative.

With all other variables constant, patients satisfied with life would take fewer opioids than those not satisfied with life. Table 3.5 summarizes the actual values of those coefficients and also provides the estimates from GLM as a comparison. Estimates from the two models are consistent.

3.6 Discussion

In this study, we propose a new model called the generalized dynamic effect change model (GDECM), which is an extension of GAM. Compared to GAM, our model is easy to implement and interpret. Our simulation studies also showed that GDECM has better estimation and prediction accuracy than GAM. As for application, by using GDECM, we were able to find an extra significant variable in predicting the preoperative opioid use, showing a promising usage in real data analyses. An R package is under construction to further expand its application.

For our current settings, we assume that the reproducing kernel is known since the selection of the kernel is not of interest in this study. However, the choice of kernels in regression splines is an exciting topic, and future work in this direction is considered.

CHAPTER IV

Soft-Thresholding Operator for Modeling Sparse Time-Varying Effects in Survival Analysis

4.1 Introduction

In analyses of survival, the Cox proportional hazards model by Cox (1992) [21] is a popular and useful tool. The key assumption of the Cox model is that the effect of a given covariate is constant over time. However, that assumption does not always hold. In fact, non-proportionality is very common in survival analyses and has been widely studied [23, 46, 59, 60, 61, 90]. The time-dependent coefficients Cox model is a typical method to adjust for the non-proportionality [39].

In this study, we are particularly interested in time-varying effects with sparsity, which means the covariate effects can be zero on specific time intervals and can be time-varying on others. Many studies mentioned this special scenario. Anderson and Gill (1982) [3] noticed the effects of some covariates disappeared in the later follow-up in a vulvar cancer study. Gore et al. (1982) [36] found that the influence of signs recorded at diagnosis waned with time in the Western General breast cancer study. Tian et al. (2005) [80] noted sparsity in the edema effect during the early stage and also showed the effect from log(prothrombin time) on survival diminished over time in the Mayo Clinic primary biliary cirrhosis dataset.

The challenge here is how to detect no-effects period and estimate effects in other periods simultaneously. Existing methods to handle the time-dependent Cox model cannot achieve the goal here. Spline based models with penalizations [42, 56, 98, 101] focused on variable selection for different covariates. They may detect one covariate as time-varying or time-constant. However, they are not able to detect the no-effects period within each covariate. Kernel weighted likelihood approach [11, 80] do not assume sparsity design for covariate effects. Ideally, this approach can handle the detection of no-effects region and estimation simultaneously. However, this method suffered a huge computational burden when the sample size is large, which makes it not pleasant to use. Therefore, we aim to develop a new statistical method that can efficiently model sparse time-varying effects in survival analysis.

Our method uses the idea of soft-thresholding to represent the time-varying effects in the Cox model. The concept of soft thresholding was introduced in [24, 25]; they applied this estimator to the coefficients of a wavelet transform of a function measured with noise. The soft-thresholding function is widely used for effect shrinkage: Chang et al. (2000) [13] proposed an adaptive, data-driven threshold for image denoising in a Bayesian framework with the generalized Gaussian distribution prior based on wavelet soft-thresholding; Tibshirani (1996) [81] also pointed out that the Lasso estimator is a soft-thresholding estimator when the covariate matrix has an orthonormal design. Kang et al. (2018) [51] first uses the soft-thresholding operator for modeling sparse, continuous, and piecewise smooth functions in image data analysis. However, their method is not developed for survival. Extending to survival needs more effort.

In this study, the soft-thresholding function is used to model the time-varying effects of covariates. The B splines approximate the kernel smoothing part. Estimation is obtained by maximizing the partial likelihood. The asymptotic properties of our proposed estimator will be provided. A novel inference approach will be introduced to quantify the uncertainty of the proposed estimator.

This paper is organized as follows. In Section 4.2, we give the model definitions and derivation of our algorithm. Section 4.3 provides the theoretical support for our method. We present simulation results in Section 4.4 to demonstrate the advantage of our methods. In Section 4.5, we analyze the Boston Lung Cancer Data using our proposed model. Section 4.6 concludes this study.

4.2 Methods

4.2.1 Model

Let T_i^u and T_i^c represent the survival and censoring times, respectively, for the ith patient. Observation times are denoted by $T_i = T_i^u \wedge T_i^c$, where $a \wedge b = \min\{a, b\}$. The observed death indicators are denoted by $\Delta_i = \mathcal{I}(T_i^u \leq T_i^c)$, where $\mathcal{I}(A)$ is an indicator function taking the value 1 when condition A holds and 0 otherwise. Let $\boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{ip})^T$ be a p-dimensional covariate vector for sample i. Let $\boldsymbol{\beta}(t) = (\beta_1(t), \ldots, \beta_p(t))^T$ be a p-dimensional vector of potentially time-varying coefficients. The observed data consist of n independent vectors, $(T_i, \Delta_i, \boldsymbol{Z}_i)$, where $T_i \in [0, \tau]$.

Let $\lambda(t|\boldsymbol{Z}_i)$ be the hazard function given $\boldsymbol{Z}_i$ and the time-varying effects survival model is specified as

$$\lambda(t|\boldsymbol{Z}_i) = \lambda_0(t) \exp(\boldsymbol{Z}_i^T \boldsymbol{\beta}(t)),$$

where $\lambda_0(t)$ is the baseline hazard.

The log partial likelihood with time-varying coefficients is

$$\mathrm{PL}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \Delta_i \left\{ \sum_{j=1}^{p} Z_{ij} \beta_j(T_i) - \log \left[\sum_{l \in R_i} \exp \left\{ \sum_{j=1}^{p} Z_{lj} \beta_j(T_i) \right\} \right] \right\}, \tag{4.1}$$

where $R_i = \{l : T_l > T_i\}$ is the risk set for sample i.

Following Definition 2.2.1 in Chapter II, we assume that $\beta_j, j = 1, \ldots, p$, is continuous everywhere, with zero-effect regions (R_0) consisting of at least one interval,

and is smooth over regions (positive R_+ and negative R_-) where its effect is non-zero. On the non-zero regions, the dth derivative of $\beta(t)$ exists and satisfies the Lipschitz condition on each interval:

$$|\beta^{(d)}(s) - \beta^{(d)}(t)| \leq C|s-t|^w, \tag{4.2}$$

where d is a non-negative integer, and $w \in (0,1]$ such that $m \equiv d + w > 0{\cdot}5$. Let $\mathbb{H}$ be the set of all such $\beta(t)$ and $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0p})^T$ the true coefficient vector to be estimated, where $\beta_{0j} \in \mathbb{H}$.

Following Chapter II, we use the soft-thresholding operator ζ to represent the varying coefficient:

$$\zeta\{\theta(t), \alpha\} = \{\theta(t) - \alpha\}\,\mathcal{I}\{\theta(t) > \alpha\} + \{\theta(t) + \alpha\}\,\mathcal{I}\{\theta(t) < -\alpha\},$$

where $\alpha > 0$ is the thresholding parameter and $\theta(t)$ is a real-valued function.

According to Lemma 1 in Appendix A, we have that for any function $\beta(t) \in \mathbb{H}$ and any $\alpha > 0$, there exists at least one $\theta(t) \in \mathbb{F}_0$ such that $\beta(t) = \zeta\{\theta, \alpha\}(t)$, where $\mathbb{F}_0$ is the class of functions θ defined on $[0, \tau]$, with the dth derivative $\theta^{(d)}$ satisfying the Lipschitz condition (4.2).

Combining all above results, we introduce a new penalized likelihood for estimation

$$\mathrm{PL}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \Delta_i \left\{ \sum_{j=1}^{p} Z_{ij} \zeta\{\theta_j(T_i), \alpha_j\} - \log \left[\sum_{l \in R_i} \exp \left\{ \sum_{j=1}^{p} Z_{lj} \zeta\{\theta_j(T_l), \alpha_j\} \right\} \right] \right\} - \rho ||\boldsymbol{\theta}||_2^2, \tag{4.3}$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T$, and $\rho > 0$ is the predetermined penalization coefficient.

With the soft-thresholding representation, we can convert the problem from estimating non-smooth functions β to estimating smooth functions. Many approaches can reduce the dimensions in estimating smooth functions. In this study, we will

utilize the B spline basis to model the smooth functions.

Let $\mathbb{F}$ be the B-spline function sieve space. With the same notation as in the Chapter II, we let $K = O(n^{\nu})$ be an integer with $0 < \nu < 0{\cdot}5$, and let $B_k(t)(1 \leq k \leq q,$ and $q = K + d)$ be the B-spline basis functions of degree $d+1$ associated with the knots $0 = t_0 < t_1 < \cdots < t_{K-1} < t_K = 1$, satisfying $\max_{1 \leq k \leq K}(t_k - t_{k-1}) = O(n^{-\nu})$. Let $\boldsymbol{B}(t) = \{B_1(t), \ldots, B_q(t)\}^T$ be a functional vector of the B-spline bases. Then, we have

$$\mathbb{F} = \left\{ \theta : \theta = \sum_{k=1}^{q} \gamma_k B_k(t), t \in [0, \tau], \gamma_k \in \mathbb{R}, k = 1, \ldots, q \right\}.$$

For given α and q, we define the thresholded sieve space

$$\mathbb{S}_{q,\alpha} = \left\{ \beta(t) = \zeta\{\theta(t), \alpha\} : \theta(t) = \sum_{k=1}^{q} \gamma_k B_k(t), t \in [0, \tau], \gamma_k \in \mathbb{R}, k = 1, \ldots, q \right\}.$$

Let $\theta_j = \boldsymbol{B}(t)^T \boldsymbol{\gamma}_j$, then the penalized log partial likelihood becomes

$$\begin{aligned} \mathrm{PL}(\boldsymbol{\gamma}) = & \sum_{i=1}^{n} \Delta_i \left\{ \sum_{j=1}^{p} Z_{ij} \zeta\{\boldsymbol{B}(T_i)^T \boldsymbol{\gamma}_j, \alpha_j\} - \log \left[\sum_{l \in R_i} \exp \left\{ \sum_{j=1}^{p} Z_{lj} \zeta\{\boldsymbol{B}(T_i)^T \boldsymbol{\gamma}_j, \alpha_j\} \right\} \right] \right\} \\ & - \rho \sum_{j=1}^{p} \sum_{i=1}^{n} \{\boldsymbol{B}(T_i)^T \boldsymbol{\gamma}_j\}^2. \end{aligned} \tag{4.4}$$

4.2.2 Estimation

In order to estimate the coefficients, we consider the same smooth approximation of the thresholding operator, h, as in Chapter II. The smooth approximation of $\zeta(\theta, \alpha)$ is defined as

$$\begin{aligned} h_\eta\{\theta(t), \alpha\} = & \frac{1}{2} \Bigg(\left[1 + \frac{2}{\pi} \arctan\{\theta_-(t)/\eta\} \right] \theta_-(t) + \\ & \left[1 - \frac{2}{\pi} \arctan\{\theta_+(t)/\eta\} \right] \theta_+(t) \Bigg), \end{aligned}$$

where $\alpha > 0$, $\eta > 0$ and $\theta_{\pm}(t) = \theta(t) \pm \alpha$. It has been verified in Appendix for Chapter II that h is continuous and differentiable, and the approximation error between $h_\eta\{\theta(t), \alpha\}$ and $\zeta(\theta, \alpha)$ is bounded by $\eta + O(\eta^3)$. We drop η hereafter for simplicity of notation. Then, we obtain the smoothed log partial likelihood function:

$$\begin{aligned}\mathrm{PL}(\boldsymbol{\gamma}) = &\sum_{i=1}^{n} \Delta_i \left\{ \sum_{j=1}^{p} Z_{ij} h\{\boldsymbol{B}(T_i)^T \boldsymbol{\gamma}_j, \alpha_j\} - \log \left[\sum_{l \in R_i} \exp \left\{ \sum_{j=1}^{p} Z_{lj} h\{\boldsymbol{B}(T_i)^T \boldsymbol{\gamma}_j, \alpha_j\} \right\} \right] \right\} \\ &- \rho \sum_{j=1}^{p} \sum_{i=1}^{n} \{\boldsymbol{B}(T_i)^T \boldsymbol{\gamma}_j\}^2. \end{aligned} \tag{4.5}$$

Let $\tilde{\boldsymbol{\gamma}} = \arg\max_{\boldsymbol{\gamma}} \mathrm{E}_{T,\boldsymbol{Z},\Delta} \mathrm{PL}(\boldsymbol{\gamma})$. An estimate of $\tilde{\boldsymbol{\gamma}}$ is obtained by maximizing the likelihood (4.5) to:

$$\hat{\boldsymbol{\gamma}} = \arg\max_{\boldsymbol{\gamma}} \ \mathrm{PL}(\boldsymbol{\gamma}).$$

Then the estimate of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)^T$, where $\hat{\beta}_j(t) = \zeta(\boldsymbol{B}(t)^T \hat{\boldsymbol{\gamma}}_j, \alpha_j)$.

Computation of $\hat{\boldsymbol{\gamma}}$ can be implemented by gradient-based methods and a coordinate descent algorithm. With appropriate initial values, global optimizers can be reached. Specifically, for each $j = 1, \ldots, p$, we obtain the non-varying coefficients $(a_1, \ldots, a_p)^T$ from the Cox model, then we set the initial $\boldsymbol{\gamma}_j^{(0)}$ to be a vector of a_j with length q. We choose the pre-specified parameters as follows. As a value of α comparable to the scale of true coefficients works well, we set α_j to be $|a_j|$. The choices of η and ρ can be specified in accordance with Condition **Ch4.C**6. The knots of B-spline are equally spaced over $[0, \tau]$. The number of basis functions, q, can be determined through R-fold cross-validation. That is, partition the full data D into R equal-sized groups, denoted by D_r, for $r = 1 \ldots, R$, and let $\hat{\boldsymbol{\beta}}_{-r}^{(q)}(t)$ be the estimate obtained with q bases using all the data except for D_r. We obtain the optimal q by minimizing the cross-validation error, which is a mean of the negative objective function over all D_r

with $\hat{\boldsymbol{\beta}}_{-r}^{(q)}(t)$.

4.3 Inference

We begin this section with some notation and key conditions. Let

$$
\begin{aligned}
g(\boldsymbol{\beta}, \boldsymbol{Z}, t) &= \sum_{j=1}^{p} Z_j \beta_j(t), \\
g_n(\boldsymbol{\gamma}, \boldsymbol{Z}, t) &= \sum_{j=1}^{p} Z_j h_j(\boldsymbol{B}(t)\boldsymbol{\gamma}_j), \\
S_{0n}(\tilde{\boldsymbol{\gamma}}, t) &= \frac{1}{n}\sum_{i=1}^{n} Y_i(t) \exp(g_n(\tilde{\boldsymbol{\gamma}}, \boldsymbol{Z}_i, t)), \\
S_0(t) &= \mathrm{E}Y(t) \exp(g(\boldsymbol{\beta}, \boldsymbol{Z}, t)), \\
S_{1n}(\tilde{\boldsymbol{\gamma}}, t) &= \frac{1}{n}\sum_{i=1}^{n} Y_i(t) \exp(g_n(\tilde{\boldsymbol{\gamma}}, \boldsymbol{Z}_i, t))\boldsymbol{Z}_i \otimes \boldsymbol{B}_i, \\
S_1(t) &= \mathrm{E}Y(t) \exp(g(\boldsymbol{\beta}, \boldsymbol{Z}, t))\boldsymbol{Z} \otimes \boldsymbol{B}, \\
S_{2n}(\tilde{\boldsymbol{\gamma}}, t) &= \frac{1}{n}\sum_{i=1}^{n} Y_i(t) \exp(g_n(\tilde{\boldsymbol{\gamma}}, \boldsymbol{Z}_i, t))(\boldsymbol{Z}_i\boldsymbol{Z}_i^T) \otimes (\boldsymbol{B}_i\boldsymbol{B}_i^T), \\
\text{and} \quad S_2(t) &= \mathrm{E}Y(t) \exp(g(\boldsymbol{\beta}, \boldsymbol{Z}, t))(\boldsymbol{Z}\boldsymbol{Z}^T) \otimes (\boldsymbol{B}\boldsymbol{B}^T).
\end{aligned}
$$

Conditions:

Ch4.C1 The failure time T^u and the censoring time T^c are conditionally independent given the covariate $\boldsymbol{Z}$.

Ch4.C2 Only the observations for which the event time T_i, $1 \le i \le n$ is in a finite interval, say $[0, \tau]$, are used in the partial likelihood. At this point τ, the baseline cumulative hazard function $\lambda_0(\tau) \equiv \int_0^\tau \lambda_0(s)ds < \infty$.

Ch4.C3 The covariates $\boldsymbol{Z}$ takes value in a bounded subset of $\mathbb{R}^p$ and $\Pr(Z_j = 0) < 1$. Also, $\sum_{j=1}^{p} |Z_j| = O_p(1)$.

Ch4.C4 There exists a small positive constant ϵ such that $\Pr(\Delta = 1|\boldsymbol{Z}) > \epsilon$ and $\Pr(T^c > \tau|\boldsymbol{Z}) > \epsilon$ almost surely with respect to the probability measure of $\boldsymbol{Z}$.

Ch4.C5 Let $0 < c_1 < c_2 < \infty$ be two constants. The joint density $f(t, \boldsymbol{z}, \Delta = 1)$ of $(T, \boldsymbol{Z}, \Delta = 1)$ satisfies $c_1 \leq f(t, \boldsymbol{z}, \Delta = 1) < c_2$ for all $(t, \boldsymbol{z}) \in [0, \tau] \times \mathbb{R}^p$.

Ch4.C6 $\eta = o(q^{-m})$, $\rho = O(n^a)$ with $a \leq -1$, and $q = o(n)$.

Ch4.C7 There exists a neighborhood Θ of $\tilde{\boldsymbol{\gamma}}$ and scalar, vector and matrix functions s_0, s_1 and s_2 defined on $\boldsymbol{\gamma} \times [0, \tau]$ such that for $j = 0, 1, 2$

$$\sup_{0 \leq t \leq \tau, \boldsymbol{\gamma} \in \Theta} ||S_j(\boldsymbol{\gamma}, t) - s_{(j)}(\boldsymbol{\gamma}, t)|| \to_p 0$$

Ch4.C8 Let Θ, s_0, s_1 and s_2 be as in Condition **Ch4.C**7 and define $e = s_1/s_0$ and $v = s_2/s_0 - e^{\otimes 2}$. For all $\boldsymbol{\gamma} \in \Theta$, $t \in [0, \tau]$:

$$s_1(\boldsymbol{\gamma}, t) = \frac{\partial}{\partial \boldsymbol{\gamma}} s_0(\boldsymbol{\gamma}, t), \quad s_2(\boldsymbol{\gamma}, t) = \frac{\partial^2}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}^T} s_0(\boldsymbol{\gamma}, t),$$

$s_0(\cdot, t)$, $s_1(\cdot, t)$, $s_2(\cdot, t)$ are continuous functions of $\boldsymbol{\gamma} \in \Theta$, uniformly in $t \in [0, \tau]$, s_0, s_1, and s_2 are bounded on $\Theta \times [0, \tau]$, and the matrix

$$\Sigma(\tilde{\boldsymbol{\gamma}}, \tau) = \int_0^\tau v(\tilde{\boldsymbol{\gamma}}, t) s_0(\tilde{\boldsymbol{\gamma}}, t) \tilde{\boldsymbol{\gamma}}(t) dt$$

is positive definite.

Ch4.C9 There exists $\delta > 0$ such that

$$n^{-1/2} \sup_{i,t} ||\boldsymbol{Z}_i||_\infty |Y_i(t) \mathcal{I}\{\boldsymbol{Z}_i^T \boldsymbol{\beta} > -\delta ||\boldsymbol{Z}_i||_\infty\}| \to_p 0.$$

Condition **Ch4.C**1 is a common assumption in analyzing right-censored data for the

censoring mechanism to be non-informative. The finite interval Condition **Ch4.C**2 is assumed in many studies [4]. Condition **Ch4.C**3 is often assumed in nonparametric regression and makes sense in practical situations since we do not observe infinite covariates. Condition **Ch4.C**4 controls the censoring rate to be not too large [69]. Condition **Ch4.C**5 is needed for model identifiability and used in Huang (1999) [46]. Condition **Ch4.C**6 controls the estimation bias and ensures the convergence. Conditions **Ch4.C**7, **Ch4.C**8, and **Ch4.C**9 are regularity conditions. Similar conditions were present in Anderson and Gill (1982) [4].

4.3.1 Asymptotic theory

Theorem 4.3.1 *Suppose Conditions* ***Ch4.C1-Ch4.C6*** *hold, if* $\beta_{0j}(t) \in \mathbb{S}_{q,\alpha_j}$ *for* $j = 1, \ldots, p$ *with* q *and* α_j *be the same as in* $\mathrm{PL}(\boldsymbol{\theta})$*, then*

$$||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0||_2 = O_p\left((q/n)^{1/2}\right);$$

if $\beta_{0j}(t) \notin \mathbb{S}_{q,\alpha_j}$ *for* $j = 1, \ldots, p$*,*

$$||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0||_2 = O_p\left(r_n^{1/2}\right),$$

where $r_n = q/n + q^{-2m}$*.*

Theorem 4.3.1 implies convergence of $\hat{\boldsymbol{\beta}}$ by Condition **Ch4.C**6 and $m > 0{\cdot}5$. If the true curves are in the thresholded sieve space, then there is no approximation error; and if q is $O(1)$, Theorem 2.3.1 suggests root-n consistency.

Let $\boldsymbol{e}_j$ be a vector of length p with jth entry as 1 and others 0. For any $t \in [0, \tau]$, let $\boldsymbol{a}(t) = \boldsymbol{e}_j \otimes \boldsymbol{B}(t)$, then $\hat{\theta}_j(t) = \boldsymbol{a}(t)^T \hat{\boldsymbol{\gamma}}$.

Theorem 4.3.2 *Under Conditions* ***Ch4.C1-Ch4.C9*** *, we have for any* $t \in [0, \tau]$

and $j = 1, \ldots, p$,

$$\frac{\hat{\theta}_j(t) - \theta_j(t)}{\sigma_{nj}(t)} \to_d N(0,1), \quad \text{as } n \to \infty,$$

where $\sigma_{nj}^2(t) = n\boldsymbol{a}(t)^T \left[\{\mathrm{PL}\}''(\tilde{\boldsymbol{\gamma}})\right]^{-1} \Sigma(\tilde{\boldsymbol{\gamma}}, 1) \left[\{\mathrm{PL}\}''(\tilde{\boldsymbol{\gamma}})\right]^{-1} \boldsymbol{a}(t)$.

With Theorem 4.3.2, we can then obtain the asymptotic distribution of $\hat{\beta}_j(t)$ based on $\hat{\beta}_j(t) = \zeta\{\hat{\theta}_j(t), \alpha_j\}$.

Theorem 4.3.3 *Under Conditions* ***Ch4.C1***–***Ch4.C9****, for any* $t \in [0, \tau]$*, the limiting distribution of* $\hat{\beta}_j(t)$ ($j = 1, \ldots, p$) *satisfies*

$$\lim_{n\to\infty} \left| \Pr(\hat{\beta}_j(t) \le x) - G_{nj}(x) \right| \; = \; 0,$$

where $G_{nj}(x) = \left[\Phi\left\{\frac{x+\alpha_j-\tilde{\theta}_j(t)}{\sigma_{nj}}\right\}\mathcal{I}(x \ge 0) + \Phi\left\{\frac{x-\alpha_j-\tilde{\theta}_j(t)}{\sigma_{nj}}\right\}\mathcal{I}(x < 0)\right]$ *and* $\Phi(\cdot)$ *is the cumulative distribution function for* $N(0,1)$.

The limiting distribution in Theorem 4.3.3 guarantees the zero-effect detection ability of our proposed estimator, since the probability of $\hat{\beta}_j(t) = 0$ is greater than 0 even with finite sample size.

4.3.2 Sparse confidence intervals

Following Chapter II, we introduce the sparse confidence intervals to gauge the uncertainty of the point estimates and make valid statistical inferences on the selection and the zero-effect region detection. Here we will only introduce the construction of the sparse confidence intervals. For detailed derivation, see Section 2.3 in Chapter II.

Given α_j, for any $t \in [0, \tau]$ we construct a pointwise $(1 - \xi)$ level asymptotic sparse confidence interval for $\beta_j(t)$, denoted by $[u_{nj}(t), v_{nj}(t)]$. Let $z_{\xi/2}$ and Φ be the $(1 - \xi/2)$ quantile and the cumulative distribution function of $N(0,1)$, respectively. Let $P_+ = \Pr\{\hat{\beta}_j(t) > 0\}$ and $P_- = \Pr\{\hat{\beta}_j(t) < 0\}$, which can be estimated by

$\hat{P}_+ = 1 - \Phi\{(\alpha_j - \hat{\theta}_j)/\hat{\sigma}_{nj}\}$ and $\hat{P}_- = \Phi\{(-\alpha_j - \hat{\theta}_j)/\hat{\sigma}_{nj}\}$ using Theorem 4.3.3. We construct $[u_{nj}(t), v_{nj}(t)]$ as follows:

- if $\hat{P}_+ + \hat{P}_- \leq \xi$, $u_{nj}(t) = v_{nj}(t) = 0$;
- else if $\hat{P}_+ < \xi/2$ and $\hat{P}_- < 1 - \xi/2$, $[u_{nj}(t), v_{nj}(t)] = \big[\hat{\beta}_j(t) - \hat{\sigma}_{nj}\hat{B}, 0\big]$ with $\hat{B} = \Phi^{-1}\big\{1 - \xi + \Phi(-\hat{\sigma}_{nj}^{-1}\alpha_j + \hat{\sigma}_{nj}^{-1}\hat{\theta}_j)\big\}$ and $\hat{\sigma}_{nj}(t)$ as defined in Theorem 4.3.2;
- else if $\hat{P}_- < \xi/2$ and $\hat{P}_+ < 1 - \xi/2$, $[u_{nj}(t), v_{nj}(t)] = \big[0, \hat{\beta}_j(t) + \hat{\sigma}_{nj}\hat{A}\big]$ with $\hat{A} = -\Phi^{-1}\big\{\xi - 1 + \Phi(\hat{\sigma}_{nj}^{-1}\alpha_j + \hat{\sigma}_{nj}^{-1}\hat{\theta}_j)\big\}$;
- else $[u_{nj}(t), v_{nj}(t)] = \big[\hat{\beta}_j(t) - \hat{\sigma}_{nj}z_{\xi/2}, \hat{\beta}_j(t) + \hat{\sigma}_{nj}z_{\xi/2}\big]$.

Theorem 4.3.4 *Under Conditions* ***Ch4.C1-Ch4.C9****,* $[u_{nj}(t), v_{nj}(t)]$ *is a* $(1-\xi)$ *level sparse confidence interval of* $\beta_j(t)$ *for* $j = 1, \ldots, p$ *and any* $t \in [0, \tau]$.

The proof of Theorem 4.3.4 is very similar to the proof of Theorem 2.3.3. Therefore, we omit the proof here.

4.4 Simulations

In this section, we will compare our method with the regular time-varying Cox model. We design some special varying coefficient functions which contain zero-effect regions. The detailed formulas are provided below:

$$\begin{aligned} \beta_1(t) &= (-t^2 + 3)\mathcal{I}(t \leq \sqrt{3}), \\ \beta_2(t) &= 2\log(t + 0.01)\mathcal{I}(t \geq 1), \\ \text{and} \qquad \beta_3(t) &= (\frac{-6}{t+1} + 2)\mathcal{I}(t \leq 2). \end{aligned} \tag{4.6}$$

We first simulate $\boldsymbol{Z_i} = (Z_{i1}, \ldots, Z_{ip)} \sim \mathrm{N}(\mathbf{0}, \Sigma)$, where Σ can be independent with $\mathrm{cov}(Z_{ij}, Z_{ij^*}) = \mathcal{I}(j = j^*)$, autoregressive with $\mathrm{cov}(Z_{ij}, Z_{ij^*}) = 0{\cdot}5^{|j-j^*|}$, or

compound symmetry with $\text{cov}(Z_{ij}, Z_{ij^*}) = \mathcal{I}(j = j^*) + 0{\cdot}5\mathcal{I}(j \neq j^*)$. We then simulate $U_i \sim U(0,1)$, and solve T_i^u using $U_i = 1 - \exp\left\{-\int_0^{T_i^u} \lambda_0(u)\exp(\sum_j^p Z_j\beta_j(u))\text{d}u\right\}$, where $\lambda_0(u)$ is set to be some constant in $(0,1)$. The censoring times C_i are generated from $U(0,10)$, then $T_i^c = \min(C_i, 3)$. Then Event indicator $\Delta_i = \mathcal{I}(T_i < C_i)$ and observation time $T_i = \min\{T_i^u, T_i^c\}$.

We choose sample size $n = 500,\ 2,000$ and $5,000$. Each setting is replicated 200 times. We set $\eta = 0{\cdot}001$, $\rho = 1/n^2$, α_j to be half of the absolute value of the least-squares estimate. The number of knots, q, is selected through cross-validation. For evaluation criteria, we use the integrated squared errors (ISE) and the averaged integrated squared errors (AISE), defined as $\text{ISE}(\beta_j) = n_g^{-1}\sum_{g=1}^{n_g}\{\hat{\beta}_j(t_g) - \beta_j(t_g)\}^2$ and $\text{AISE} = p^{-1}\sum_{j=1}^{p}\text{ISE}(\beta_j)$, respectively, where t_g $(g = 1, \ldots, n_g)$ are the grid points on $(0,3)$. Table 4.1 summarizes the results. The results show that the soft-thresholded time-varying Cox model has smaller integrated squared errors and averaged integrated squared errors than the regular time-varying Cox model, indicating the soft-thresholded time-varying Cox model has better estimation accuracy than the regular time-varying Cox model.

Figure 4.1 plots the estimation curves and their median for the soft-thresholded time-varying Cox model and the regular time-varying Cox model. From the figure, the medium estimation curves from the soft-thresholded time-varying Cox model cover the truth, while the regular time-varying Cox model fail to estimate the zero effect. It shows that the soft-thresholded time-varying Cox model has the zero-effect detection ability.

Figure 4.2 compares the estimation of coverage probability from the soft-thresholded time-varying Cox model and the regular time-varying Cox model. The figure shows that the soft-thresholded time-varying Cox model has reasonable coverage probability in both zero-effect region and non-zero-effect region. In the region around the transition point, the soft thresholded time-varying Cox model has a higher coverage

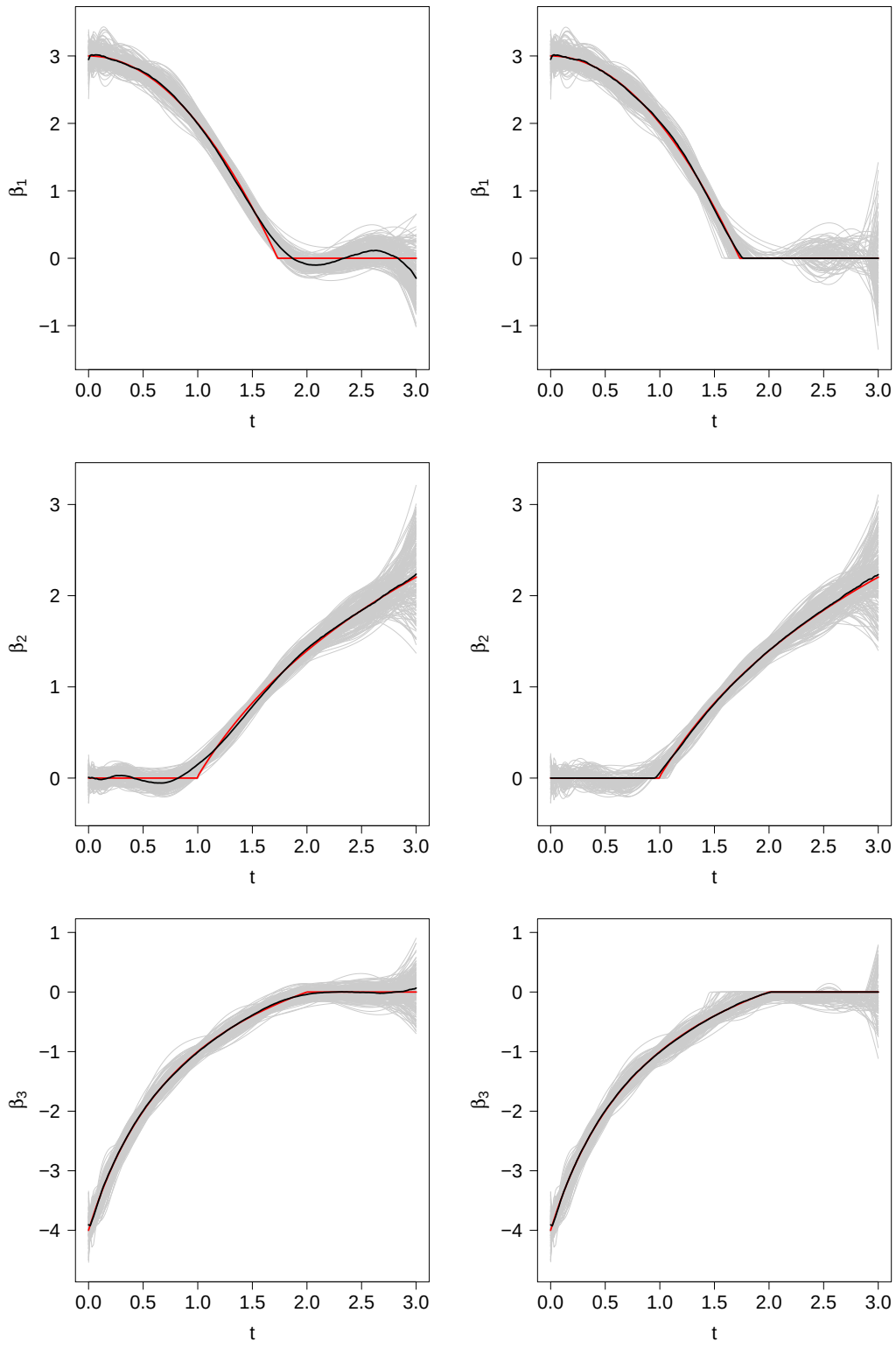


Figure 4.1: Comparison of estimation result from the soft-thresholded time-varying Cox model (right panel) and the regular time-varying Cox model (left panel). The gray curves are estimation curves from 200 simulations, the black curves are the medium estimation curves, and the red curves are the simulation truth. The data sample size is N=5,000 and the average event rate is 0.88.

Table 4.1: Comparisons of estimation accuracy for the soft-thresholded time-varying Cox model and the regular time-varying Cox model.

Covariance	n	Model	ISE(β_1)	ISE(β_2)	ISE(β_3)	AISE
Ind	500	STTV	62.6 (77.1)	53.1 (43.5)	58.7 (59.5)	58.1 (39.7)
		RegTV	75.5 (94.6)	56.6 (44.3)	61.9 (60.6)	65.4 (46.2)
	2000	STTV	12.4 (9.7)	12.0 (8.5)	13.1 (10.4)	12.5 (5.7)
		RegTV	13.9 (8.2)	11.8 (8.6)	12.4 (8.8)	12.7 (5.1)
	5000	STTV	4.2 (3.2)	4.1 (2.8)	4.0 (2.7)	4.1 (1.7)
		RegTV	5.6 (3.0)	4.2 (2.8)	4.5 (2.7)	4.7 (1.6)
AR(1)	500	STTV	16.2 (16.0)	18.2 (47.1)	15.2 (11.1)	16.5 (18.7)
		RegTV	16.3 (14.1)	20.9 (50.3)	13.4 (8.2)	16.9 (19.4)
	2000	STTV	3.6 (2.2)	2.6 (2.0)	3.9 (2.3)	3.3 (1.5)
		RegTV	3.7 (2.2)	2.8 (2.5)	3.1 (1.6)	3.2 (1.4)
	5000	STTV	1.3 (1.0)	1.1 (0.9)	1.2 (0.8)	1.2 (0.6)
		RegTV	1.9 (0.9)	1.3 (0.9)	1.3 (0.8)	1.5 (0.6)
CS	500	STTV	18.9 (24.6)	19.1 (30.2)	16.5 (14.6)	18.2 (16.2)
		RegTV	19.1 (15.5)	20.4 (30.3)	17.0 (12.2)	18.8 (13.2)
	2000	STTV	3.6 (2.6)	2.7 (2.5)	3.8 (2.7)	3.4 (1.8)
		RegTV	4.0 (2.3)	2.8 (2.4)	3.2 (1.6)	3.4 (1.4)
	5000	STTV	1.2 (0.8)	1.1 (0.9)	1.0 (0.6)	1.1 (0.5)
		RegTV	1.8 (0.7)	1.1 (0.9)	1.2 (0.7)	1.4 (0.5)

STTV: the soft-thresholded time-varying Cox model; RegTV: the regular time-varying Cox model; ISE: the integrated squared errors; AISE: the averaged integrated squared errors. Values are multiplied by 100.

probability estimation than the regular time-varying Cox model. It further confirms that the soft-thresholded time-varying Cox model has better estimation and inference than the regular time-varying Cox model.

Let $|A|$ denote the cardinality of set A. We use same comparison metrics defined

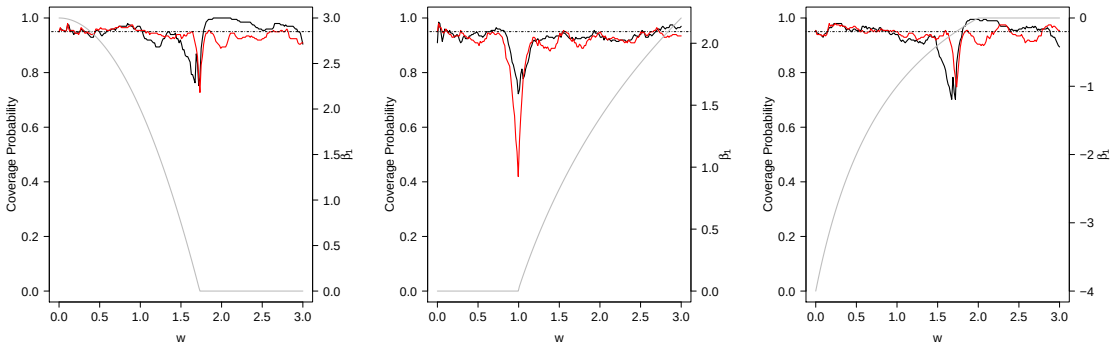


Figure 4.2: Comparisons of coverage probability (cp) from the regular time-varying Cox model (RegTV) and the soft-thresholded time-varying Cox model (STTV). The data sample size is $N = 5,000$ and the average event rate is 0.88.

in Chapter II to compare zero-effect region detection:

$$\text{Estimation-based true positive ratio: } \text{ETPR}(\beta) = \frac{|\{t : \hat{\beta}(t) \neq 0 \text{ and } \beta(t) \neq 0\}|}{|\{t : \beta(t) \neq 0\}|},$$

$$\text{Estimation-based true negative ratio: } \text{ETNR}(\beta) = \frac{|\{t : \hat{\beta}(t) = 0 \text{ and } \beta(t) = 0\}|}{|\{t : \beta(t) = 0\}|},$$

$$\text{Inference-based true positive ratio: } \text{ITPR}(\beta) = \frac{|\{t : 0 \notin \text{CI}\{\hat{\beta}(t)\} \text{ and } \beta(t) \neq 0\}|}{|\{t : \beta(t) \neq 0\}|},$$

and

$$\text{Inference-based true negative ratio: } \text{ITNR}(\beta) = \frac{|\{t : 0 \in \text{CI}\{\hat{\beta}(t)\} \text{ and } \beta(t) = 0\}|}{|\{t : \beta(t) = 0\}|},$$

where $\text{CI}\{\hat{\beta}(t)\}$ is the 95% confidence interval of $\beta(t)$.

We choose 100 grid points on $[0, 3]$ and count the number of t_g in each set as its cardinality. Table 4.2 shows that the soft-thresholded time-varying Cox model has higher values of the inference-based true negative ratio than the regular time-varying Cox model. Although the inference-based true positive and negative ratios are more reliable with controlled false discovery rates, their computational burden increases when the sample size increases. Therefore, the estimation-based true positive and negative ratios are favorable for large datasets as their calculation merely depends on the estimations. The estimation-based true negative ratio in our method has stable

Table 4.2: Comparisons of true positive ratios and true negative ratios for zero-effect region detection

			STTV				RegTV	
	n	β	ETPR	ETNR	ITPR	ITNR	ITPR	ITNR
Ind	500	β_1	0.96 (0.08)	0.44 (0.25)	0.81 (0.10)	0.94 (0.11)	0.81 (0.09)	0.95 (0.11)
		β_2	0.96 (0.05)	0.22 (0.13)	0.57 (0.17)	0.94 (0.08)	0.57 (0.18)	0.94 (0.12)
		β_3	0.95 (0.12)	0.37 (0.28)	0.55 (0.12)	0.94 (0.12)	0.55 (0.12)	0.94 (0.12)
	2000	β_1	0.95 (0.06)	0.61 (0.23)	0.89 (0.07)	0.95 (0.10)	0.91 (0.06)	0.94 (0.10)
		β_2	0.97 (0.04)	0.34 (0.16)	0.85 (0.08)	0.94 (0.08)	0.86 (0.08)	0.95 (0.08)
		β_3	0.96 (0.12)	0.50 (0.24)	0.70 (0.10)	0.94 (0.11)	0.72 (0.10)	0.95 (0.13)
	5000	β_1	0.98 (0.03)	0.64 (0.27)	0.95 (0.04)	0.94 (0.10)	0.96 (0.04)	0.93 (0.11)
		β_2	0.98 (0.03)	0.46 (0.18)	0.92 (0.05)	0.94 (0.09)	0.93 (0.04)	0.94 (0.10)
		β_3	0.97 (0.09)	0.50 (0.31)	0.81 (0.09)	0.96 (0.10)	0.80 (0.08)	0.96 (0.10)
AR(1)	500	β_1	0.96 (0.05)	0.60 (0.22)	0.90 (0.07)	0.95 (0.10)	0.93 (0.06)	0.93 (0.12)
		β_2	0.98 (0.04)	0.32 (0.18)	0.85 (0.08)	0.92 (0.13)	0.86 (0.08)	0.93 (0.12)
		β_3	0.97 (0.14)	0.51 (0.27)	0.69 (0.14)	0.95 (0.13)	0.73 (0.12)	0.95 (0.12)
	2000	β_1	0.97 (0.04)	0.71 (0.19)	0.94 (0.04)	0.95 (0.08)	0.99 (0.02)	0.92 (0.11)
		β_2	0.99 (0.02)	0.49 (0.19)	0.95 (0.04)	0.94 (0.10)	0.97 (0.03)	0.93 (0.11)
		β_3	0.96 (0.09)	0.62 (0.25)	0.77 (0.10)	0.92 (0.13)	0.86 (0.07)	0.94 (0.13)
	5000	β_1	1.00 (0.01)	0.79 (0.17)	0.98 (0.02)	0.96 (0.08)	1.00 (0.00)	0.85 (0.11)
		β_2	1.00 (0.01)	0.56 (0.17)	0.98 (0.02)	0.94 (0.09)	1.00 (0.01)	0.87 (0.11)
		β_3	0.97 (0.05)	0.63 (0.30)	0.90 (0.05)	0.97 (0.09)	0.91 (0.05)	0.96 (0.10)
CS	500	β_1	0.96 (0.06)	0.58 (0.23)	0.90 (0.07)	0.96 (0.10)	0.92 (0.07)	0.94 (0.12)
		β_2	0.98 (0.03)	0.32 (0.19)	0.85 (0.07)	0.93 (0.12)	0.86 (0.07)	0.94 (0.13)
		β_3	0.98 (0.13)	0.51 (0.29)	0.70 (0.13)	0.96 (0.11)	0.71 (0.12)	0.95 (0.11)
	2000	β_1	0.97 (0.04)	0.68 (0.21)	0.94 (0.04)	0.96 (0.08)	0.98 (0.02)	0.92 (0.12)
		β_2	0.99 (0.02)	0.48 (0.18)	0.96 (0.04)	0.94 (0.09)	0.97 (0.03)	0.94 (0.09)
		β_3	0.97 (0.11)	0.65 (0.25)	0.78 (0.11)	0.95 (0.09)	0.86 (0.07)	0.94 (0.13)
	5000	β_1	0.99 (0.01)	0.73 (0.18)	0.98 (0.02)	0.96 (0.08)	1.00 (0.01)	0.87 (0.11)
		β_2	1.00 (0.01)	0.55 (0.16)	0.98 (0.02)	0.92 (0.10)	1.00 (0.01)	0.89 (0.10)
		β_3	0.96 (0.06)	0.66 (0.30)	0.89 (0.06)	0.97 (0.08)	0.90 (0.05)	0.96 (0.11)

STTV: the soft-thresholded time-varying Cox model; RegTV: the regular time-varying Cox model.

positive values indicating the zero-effect region detection probability ability of our approach. We also compare the non-zero-effect region selection accuracy between our estimation-based method and our inference-based method in Table 4.2. The estimation-based true positive ratio is slightly higher than the inference-based true positive ratio, but both of them closely approach to 1 as n increases. The estimation-based method is computationally much faster than the inference-based method.

4.5 Real data application

In this section, we will apply our method to a subset of the Boston Lung Cancer Survivor Cohort (BLCSC) [18]. The data consists of $n = 599$ individuals, among which 148 were alive, and 451 were dead before the end of the study. The endpoint of the study is death, and the survival outcome is the survival time from the diagnosis of lung cancer to death or the end of the study. Patients in the alive group were younger than that of those in the dead group (average age in years: 55.4 vs. 61.2). In both groups, most patients were Caucasian (89.9% and 95.8%). Early-stage lung cancer is defined as lung cancer with stage lower than II, including 1A, 1B, IIA, and IIB. Late-stage lung cancer is those with stage higher than III, including IIIA, IIIb, and IV. In the alive group, 64.2% of the patients had early-stage lung cancer, higher than the rate of early-stage patients in the dead group (62.3%). The rate of patients who had surgery in the alive group was 83.8%, higher than that in the dead group (63.0%). More information can be found in Table 4.3.

We include the variable age, race, education, sex, smoking status, cancer stage, and treatments received (surgery, chemotherapy, and radiotherapy) into the time-varying Cox model. The estimation results from the Cox model, the regular time-varying Cox model (RegTV) and the soft-thresholded time-varying Cox model (STTV) are shown in Figure 4.3, 4.4 and 4.5.

Compared with the regular time-varying Cox model, the soft-thresholded time-varying Cox model is more consistent with the constant effect Cox model. For some non-significant coefficients in the constant effect Cox model, STTV estimates those to be all zero over the time, such as for chemotherapy, radiotherapy, and education above high school. Holding all other factors constant, receiving surgery has a protective effect for lung cancer patients. There is no evidence that chemotherapy and radiotherapy are protective factors in increasing lung cancer patients' survival time. Compared to female patients, the expected hazard is significantly higher in male pa-

Table 4.3: Summary statistics table for the Boston Lung Cancer Data

Variable	Alive ($n = 148$)	Dead ($n = 451$)
Time (days)	2729.9 (1793.1)	1414.9 (1488.3)
Age	55.4 (10.1)	61.2 (10.8)
Pack years	34.4 (29.7)	51.6 (38.5)
Race		
White (ref)	133 (89.9%)	432 (95.8%)
Others	15 (10.1%)	19 (4.2%)
Education		
Under high school (ref)	10 (6.8%)	72 (16%)
High school graduate	30 (20.3%)	113 (25.1%)
Above high school	108 (73.0%)	266 (59.0%)
Sex		
Female (ref)	113 (76.4%)	256 (56.8%)
Male	35 (23.6%)	195 (43.2%)
Smoking status		
Ever or never (ref)	96 (64.9%)	281 (62.3%)
Current	52 (35.1%)	170 (37.7%)
Cancer stage		
Early (ref)	95 (64.2%)	190 (42.1%)
Late	53 (35.8%)	261 (57.9%)
Surgery	124 (83.8%)	284 (63.0%)
Chemotherapy	48 (32.4%)	206 (45.7%)
Radiotherapy	35 (23.6%)	184 (40.8%)

Continuous variables are presented in mean (standard deviation), and categorical variables are presented in count (percentage). Due to rounding, some summations of percentages for one variable are not one. Reference groups in the model are marked by (ref).

tients, adjusting for all other factors. Older patients have a significantly higher hazard compared to younger patients when other factors are the same. Non-white patients have a lower hazard than white patients adjusting for others. Smoking and the late cancer stage are predictive factors for lung cancer death. There are no significant associations between education levels and the lung cancer patient's survival adjusting for others. In conclusion, STTV is consistent with the Cox model and also accurately capture the time-varying effects of each factor.

To verify that our model, STTV, has better performance than RegTV, we calculate the C statistics for both models. The C statistics for the survival model is definded as

$$C_{stat} = \frac{\sum_{i<j} \triangle_i \mathcal{I}\{D_i > D_j, T_i < T_j\}}{\sum_{i<j} \triangle_i},$$

where $D_i = \sum_{j=1}^{p} Z_{ij}\beta_j(T_i)$ for $i = 1, \ldots, n$. The C statistics is 0.61 for RegTV, and 0.65 for STV. Since higher C statistics indicates a better model fitting, our model outperforms RegTV for this dataset.

4.6 Discussion

To address the challenge of modeling time-varying coefficients with zero-effect regions in survival analysis, we proposed a new soft-thresholded time-varying coefficient model, where the varying coefficients are piecewise smooth with zero-effect regions. We have designed an efficient estimation method and a novel sparse confidence interval, which extends classical confidence intervals by accommodating the exact zero estimates. Our flexible framework enables us to perform variable selection and detect the zero-effect regions of selected variables simultaneously, and to obtain point estimates of the varying coefficients with zero-effect regions and construct the associated sparse confidence intervals.

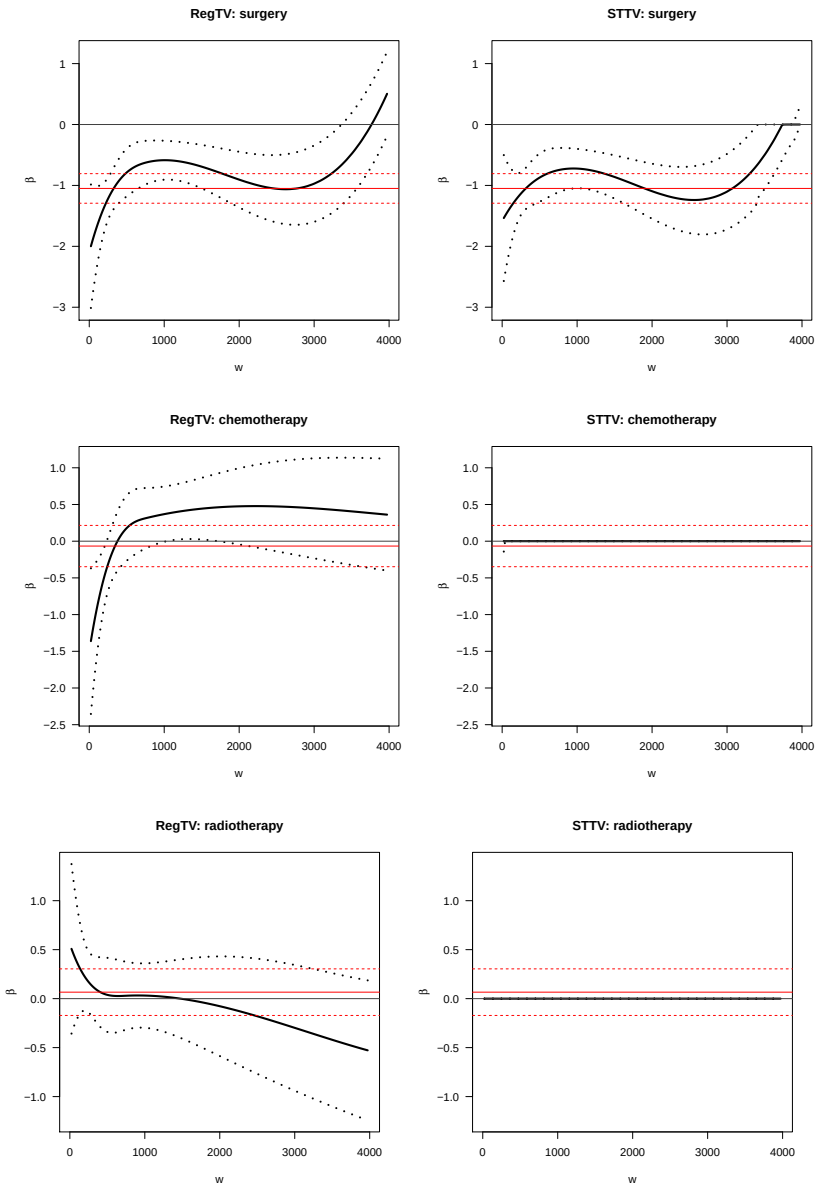


Figure 4.3: Estimation results (part I) for the BLCSC data using the regular time-varying Cox model (RegTV) and the soft-thresholded time-varying Cox model (STTV): the solid lines are the estimated coefficient function curves; the dotted lines are the pointwise (sparse) confidence intervals; black lines are from varying coefficient modesl; red lines are from the constant effect Cox model.

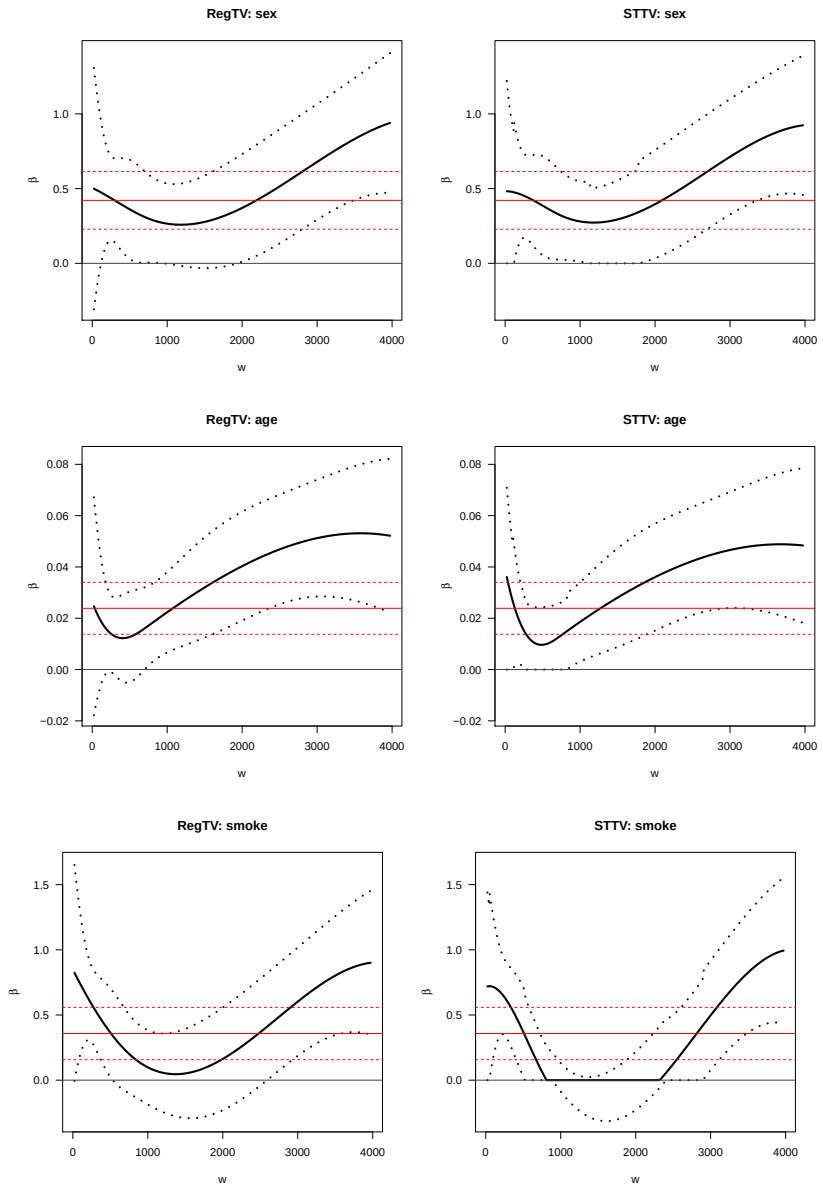


Figure 4.4: Estimation results (part II) for the BLCSC data using the regular time-varying Cox model (RegTV) and the soft-thresholded time-varying Cox model (STTV): the solid lines are the estimated coefficient function curves; the dotted lines are the pointwise (sparse) confidence intervals; black lines are from varying coefficient modesl; red lines are from the constant effect Cox model.

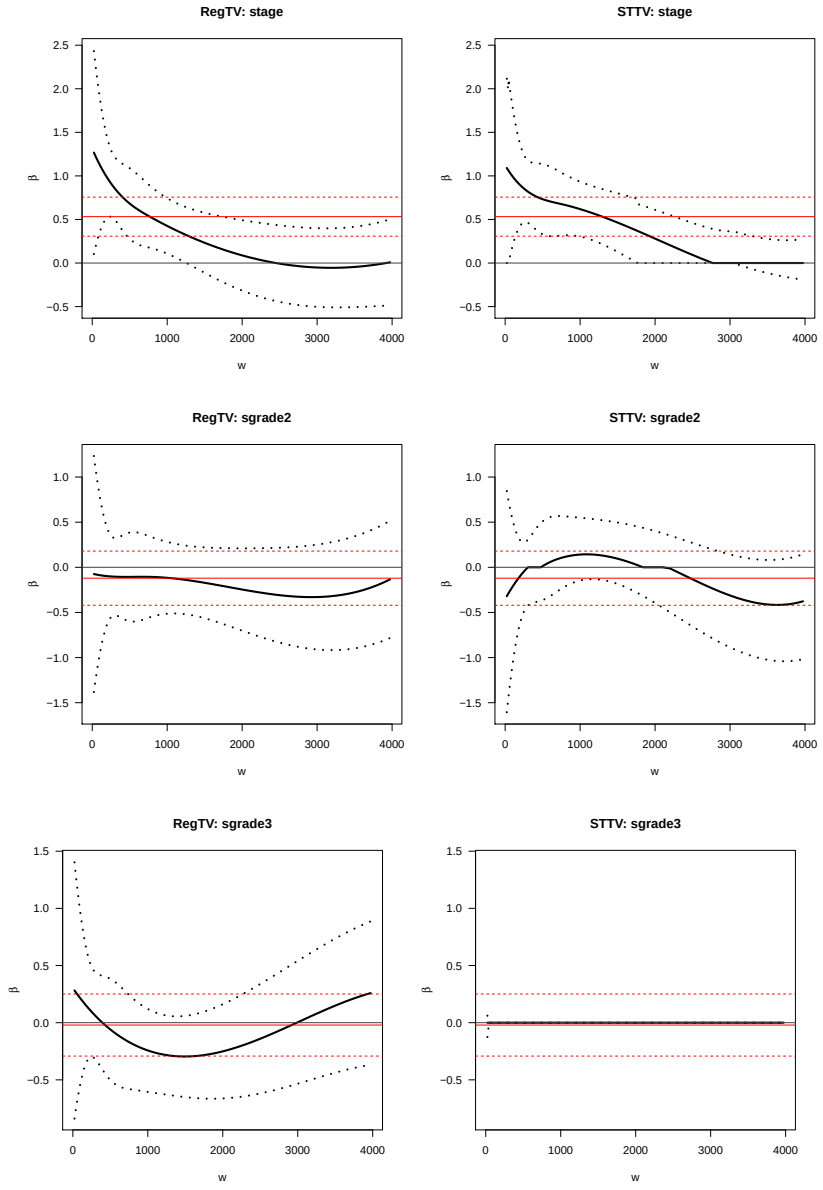


Figure 4.5: Estimation results (part III) for the BLCSC data using the regular time-varying Cox model (RegTV) and the soft-thresholded time-varying Cox model (STTV): the solid lines are the estimated coefficient function curves; the dotted lines are the pointwise (sparse) confidence intervals; black lines are from varying coefficient modesl; red lines are from the constant effect Cox model.

APPENDICES

APPENDIX A

Appendices for Chapter II

A.1 Appendix for A New Soft-Thresholded Varying Coefficient Model to Predict Opioid Use with Risk Factors that Have Zero-effect Regions

We first introduce some common notation that will be used throughout the Appendix. Let a_{1n} and a_{2n} be two sequences of real numbers indexed by positive integers and a_{2n} is positive for all n. For a real number a_1, say a_{1n} tends to a limit a_1 in symbols: $a_{1n} \to a_1$ as $n \to \infty$. We say $a_{1n} = O(a_{2n})$ if there exist an $M > 0$ and a finite $N > 0$ such that $|a_{1n}/a_{2n}| < M$ when $n > N$. We say $a_{1n} = o(a_{2n})$ if $|a_{1n}/a_{2n}| \to 0$ as $n \to \infty$. For a sequence of random variables Z_n, we say $Z_n = O_p(a_{1n})$ if for any $\delta > 0$, there exist a finite $M > 0$ and a finite $N > 0$ such that $\Pr(|Z_n/a_{1n}| > M) < \delta$ when $n > N$; and $Z_n = o_p(a_{1n})$ if for any $\delta > 0$, $\Pr(|Z_n/a_{1n}| > \delta) \to 0$ as $n \to \infty$. The convergence of Z_n in distribution to a random variable Z is denoted by $Z_n \to_d Z$, which implies that $\lim F_n(z) = F(z)$ as $n \to \infty$ for every z at which F is continuous, where F_n and F are the cumulative distribution functions of random variables Z_n and Z, respectively. Let $\mathrm{E}_n f(\cdot) = n^{-1} \sum_{i=1}^{n} f(\cdot)$ be the empirical mean of f, and $\mathrm{E} f$ the

theoretical mean of f. Let $\otimes$ denote the Kronecker product. Let f' and f'' denote the first and second derivatives of f function, respectively. Let $N(\mu, \sigma^2)$ denote the normal distribution with mean μ and variance σ^2. Let $\mathcal{I}(\mathcal{A})$ be an event indicator function, where $\mathcal{I}(\mathcal{A}) = 1$ if event $\mathcal{A}$ is true and $\mathcal{I}(\mathcal{A}) = 0$ otherwise. Let I_d be a $d \times d$ identity matrix. For a real valued function θ on $\mathbb{D}$, $||\theta||_\infty = \sup_{w \in \mathbb{D}} |\theta(w)|$ denotes its supreme norm and $||\theta||_2 = \{\int_{w \in \mathbb{D}} |\theta(w)|^2\}^{1/2}$ denotes its $\mathcal{L}_2$ norm. For a vector valued function $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T$, let $||\boldsymbol{\theta}||_2 = \left\{ \sum_j ||\theta_j||_2^2 \right\}^{1/2}$ and $||\boldsymbol{\theta}||_\infty = \max_{1 \le j \le p} ||\theta_j||_\infty$.

A.1.1 TECHNICAL DERIVATIONS

A.1.2 Properties of $H_\eta(\theta, \alpha)$

For any $\eta > 0$, $\alpha > 0$, and a real function θ, we have

$$
\begin{aligned}
&\left| \zeta_{(\theta,\alpha)} - H_\eta(\theta, \alpha) \right| \\
=&\Bigg| (\theta - \alpha)\mathcal{I}(\theta > \alpha) + (\theta + \alpha)\mathcal{I}(\theta < -\alpha) - \frac{1}{2}\left\{ 1 + \frac{2}{\pi}\arctan\left(\frac{\theta - \alpha}{\eta} \right) \right\}(\theta - \alpha) - \\
&\frac{1}{2}\left\{ 1 - \frac{2}{\pi}\arctan\left(\frac{\theta + \alpha}{\eta} \right) \right\}(\theta + \alpha) \Bigg| \\
=&\Bigg| (\theta - \alpha)\left[\mathcal{I}(\theta > \alpha) - \frac{1}{2}\left\{ 1 + \frac{2}{\pi}\arctan\left(\frac{\theta - \alpha}{\eta} \right) \right\} \right] + \\
&(\theta + \alpha)\left[\mathcal{I}(\theta < -\alpha) - \frac{1}{2}\left\{ 1 + \frac{2}{\pi}\arctan\left(\frac{\theta + \alpha}{\eta} \right) \right\} \right] \Bigg| \\
\le&\left| (\theta - \alpha)\left[\mathcal{I}(\theta > \alpha) - \frac{1}{2}\left\{ 1 + \operatorname{sign}(\theta - \alpha) + \frac{\eta}{\theta - \alpha} + O(\eta^3) \right\} \right] \right| + \\
&\left| (\theta + \alpha)\left[\mathcal{I}(\theta < -\alpha) - \frac{1}{2}\left\{ 1 + \operatorname{sign}(\theta + \alpha) + \frac{\eta}{\theta + \alpha} + O(\eta^3) \right\} \right] \right| \\
=&\eta + O(\eta^3).
\end{aligned}
$$

Therefore, the bias due to approximation is bounded by $\eta + O(\eta^3)$.

When α and η are fixed, the first derivative of h function in terms of θ is

$$H'_\eta(\theta,\alpha) = \frac{1}{\pi}\cdot\frac{(\theta-\alpha)/\eta}{1+(\theta-\alpha)^2/\eta^2} + \frac{1}{2}\left\{1+\frac{2}{\pi}\arctan\left(\frac{\theta-\alpha}{\eta}\right)\right\} - \frac{1}{\pi}\cdot\frac{(\theta-\alpha)/\eta}{1+(\theta-\alpha)^2/\eta^2} + \frac{1}{2}\left\{1-\frac{2}{\pi}\arctan\left(\frac{\theta+\alpha}{\eta}\right)\right\},$$

and the second derivative is

$$H''_\eta(\theta,\alpha) = \frac{2}{\pi}\cdot\frac{(\eta-\theta+\alpha)/\eta^2}{1+(\theta-\alpha)^2/\eta^2} - \frac{2}{\pi}\cdot\frac{(\eta-\theta-\alpha)/\eta^2}{1+(\theta+\alpha)^2/\eta^2}.$$

To facilitate the ensuing proofs, we also provide the approximation of H' here. For $-\alpha<\theta<\alpha$, by the Taylor expansion of H' around $\eta=0$, we have

$$H'_\eta(\theta,\alpha) = \frac{1}{\pi}\left\{\frac{2(\theta-\alpha)^2-8}{(\theta-\alpha)^5} - \frac{2(\theta+\alpha)^2-8}{(\theta+\alpha)^5}\right\}\eta^3 + o(\eta^3).$$

A.1.3 TECHNICAL PROOFS

Some conditions are assumed for the proofs.

Conditions:

(**Ch2.C1**) The covariates $\boldsymbol{X}$ take values in a bounded subset of $\mathbb{R}^p$. That is, there exist finite real numbers C_1 and C_2 such that $\Pr(C_1 < X_j < C_2, \text{for all } j = 1,\ldots,p) = 1$.

(**Ch2.C2**) The eigenvalues $\lambda_1 \le \ldots \le \lambda_p$ of $\mathrm{E}(\boldsymbol{X}\boldsymbol{X}^T)$ are bounded away from zero and infinity; that is, there are positive constants M_1 and M_2 such that $M_1 \le \lambda_1 \le \ldots \le \lambda_p \le M_2$. Consequently, the eigenvalues of $\mathrm{E}(\boldsymbol{V}_n\boldsymbol{V}_n^T)$ are bounded away from zero and infinity.

(**Ch2.C3**) The error satisfies $\lim_{\lambda\to\infty}\mathrm{E}\{\epsilon^2\mathcal{I}(|\epsilon|>\lambda)\}=0$ and $\mathrm{E}\{\exp(t\epsilon)\}\le\exp(\sigma^2t^2/2)$ for any t in $\mathbb{R}$.

(**Ch2.C4**) $l''_n(\boldsymbol{\gamma})$ is bounded and has a bounded inverse around $\tilde{\boldsymbol{\gamma}}$; $\mathrm{E}(\tilde{\boldsymbol{U}}\boldsymbol{X}^T)$ is invertible, where $\tilde{\boldsymbol{U}} = \boldsymbol{U}(\tilde{\boldsymbol{\gamma}};\boldsymbol{X},W)$.

(**Ch2.C5**) The distribution of W is absolutely continuous and its density is bounded away from zero and infinity on $\mathbb{D}$. Moreover, W is independent of $\boldsymbol{X}$.

(**Ch2.C6**) $\tilde{p} = o(\min\{n/q, q^{2m}\})$, $\rho = o(\tilde{p}^{1/2}q^{-m})$ and $\nabla(\eta) = o(q^{-m})$.

(**Ch2.C7**) The true varying coefficients $\beta_{0j}\,(j = 1, \ldots, p)$ are bounded.

Conditions (**Ch2.C1**)–(**Ch2.C4**) are mild regularity conditions used in the existing literature [33, 49]. Condition (**Ch2.C5**) guarantees that observations are randomly scattered [50]. Condition (**Ch2.C6**) is a technical assumption that controls convergence rate, estimation bias, and model sparsity. Condition (**Ch2.C7**) is reasonable for a wide range of applications.

Let $M_n(\boldsymbol{\theta}) = -\mathrm{E}_n l^s(\boldsymbol{\theta})$ and $M_0(\boldsymbol{\theta}) = -\mathrm{E} l^s(\boldsymbol{\theta})$ be the empirical and theoretical mean of $l^s(\boldsymbol{\theta})$. Let $|v|$ denote the Euclidean norm of a real valued vector v. For a real valued function θ on $\mathbb{D}$, $||\theta||_\infty = \sup_{w\in\mathbb{D}} |\theta(w)|$ denotes its supreme norm and $||\theta||_2 = \{\int_{w\in\mathbb{D}} |\theta(w)|^2\}^{1/2}$ denotes its $\mathcal{L}_2$ norm. For a vector valued function $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T$, let $||\boldsymbol{\theta}||_2 = \left\{\sum_j ||\theta_j||_2^2\right\}^{1/2}$ and $||\boldsymbol{\theta}||_\infty = \max_{1\le j\le p} ||\theta_j||_\infty$. Let $N_{[]}(\delta, \mathbb{S}, \mathcal{L}_p)$ be the δ-bracketing number for $\mathbb{S}$ under norm $\mathcal{L}_p$ and $\mathrm{E}^*(g)$ denote the outer expectation of process g. For two sequences a_n and b_n, we say $a_n \simeq b_n$ if $a_n/b_n = O(1)$. The convergence of Z_n in distribution to a random variable Z is denoted by $Z_n \to_d Z$, which implies that $\lim F_n(z) = F(z)$ as $n \to \infty$ for every z at which F is continuous, where F_n and F are the cumulative distribution functions of random variables Z_n and Z, respectively. The convergence of Z_n in probability to a random variable Z is denoted by $Z_n \to_p Z$, which implies that $\lim \Pr(|Z_n - Z| > \epsilon) = 0$ as $n \to \infty$ for all $\epsilon > 0$. A sequence of random vectors or matrices converge to a random vector or matrix if and only if each component of random vectors or matrices converges in probability to each component of the vector or matrix.

Lemma 1 *For any function $\beta(w) \in \mathbb{H}$ and any $\alpha > 0$, there exists at least one $\theta(w) \in \mathbb{F}_0$ such that $\beta(w) = \zeta_{\{\theta,\alpha\}}(w)$.*

Proof of Lemma 1:

We show that Lemma 1 is valid when $\beta(w)$ has only one zero region (w_0, w_1), where $w_0, w_1 \in (0,1)$. The proof can be easily extended to more general settings. Without loss of generality, we further assume $\beta(w) < 0$ on $[0, w_0)$ and $\beta(w) > 0$ on $(w_1, 1]$. The definition of $\beta(w)$ implies that $\beta^{(j)}$ exists on $[0, w_0]$ and $[w_1, 1]$, and that there exists a constant $M > 0$ such that $|\beta^{(j)}(w_k)| < M$ for $j = 1, \ldots, d$ and $k = 0, 1$.

In the following, we construct a θ satisfying: (i) $\theta(w) = b(w) - \alpha$ on $[0, w_0]$ and $\theta(w) = b(w) + \alpha$ on $[w_1, 1]$; (ii) for $j = 1, \ldots, d$, $\theta^{(j)}(w_0) = \beta^{(j)}(w_0)$, and $\theta^{(j)}(w_1) = \beta^{(j)}(w_1)$; (iii) $|\theta(w)| < \alpha$ on (w_0, w_1); and (iv) $|\theta^{(d)}(s) - \theta^{(d)}(w)| \leq C|s - w|^t$ for s, w in $[0, 1]$ and some constant C, where $0 < t \leq 1$.

Let $f(w) = e^{-1/w}\mathcal{I}(w > 0)$. It follows that $f(w) \in [0, 1)$ and $f^{(d)}(0) = 0$ for any $d \geq 1$. Define $f_0(w, a_0) = f(-w + a_0)/\{f(-w + a_0) + f(-w_0 + w)\}$ and $f_1(w, a_1) = f(w - a_1)/\{f(w - a_1) + f(w_1 - w)\}$, where $a_0 \in (w_0, (w_0 + w_1)/2)$ and $a_1 \in ((w_0 + w_1)/2, w_1)$. As $f(w)$ is infinitely differentiable over the real line, so is $f_k(w)$ for $k = 0, 1$. It is easy to verify that $f_k(w, a_1)$ satisfies that $f_k(w_k, a_k) = 1$, $f_k(a_k, a_k) = 0$, $f_k^{(j)}(w, a_k) = 0$ when $w = a_k$ or w_k, and $0 \leq f_k(w, a_k) \leq 1$ for $k = 0, 1$ and $j \geq 1$.

Let $\theta_0^*(w) = -\alpha + \sum_{j=1}^{d} \frac{\beta^{(j)}(w_0)}{j!}(w - w_0)^j$ and $\theta_1^*(w) = \alpha + \sum_{j=1}^{d} \frac{\beta^{(j)}(w_1)}{j!}(w - w_1)^j$. We define

$$\theta(w) = \begin{cases} b(w) - \alpha, & w \in [0, w_0] \\ \theta_0^*(w) * f_0(w, a_0), & w \in (w_0, a_0] \\ 0, & w \in (a_0, a_1) \\ \theta_1^*(w) * f_1(w, a_1), & w \in [a_1, w_1) \\ b(w) + \alpha, & w \in [w_1, 1] \end{cases},$$

and show that there exist a_0 and a_1 which ensure the above $\theta(w)$ satisfies conditions (i)-(iv).

It is obvious that $\theta(w)$ satisfies (i) and $\theta(w)$ is continuous. Since $f_k(w_k, a_k) = 1$

and $f_k^{(j)}(w_k, a_k) = 0$ for $j \geq 1$, we have that $\theta^{(j)}(w_k) = \theta_k^{*(j)}(w_k) = \beta^{(j)}(w_k)$ for $j = 1, \ldots, d$, where $k = 0, 1$. Therefore, condition (ii) is satisfied.

Since $\theta_0^*(w)$ and $f_0(w, a_0)$ are infinitely differentiable over (w_0, a_0), so is $\theta(w)$ over (w_0, a_0). Similarly, $\theta(w)$ is also infinitely differentiable over (a_1, w_1). Because $f_k^{(j)}(a_k, a_k) = 0$ for $j \geq 0$ and $k = 0, 1$, we have that $\theta^{(j)}(a_k) = 0$ for $j \geq 0$ and $k = 0, 1$. Therefore, $\theta(w)$ is infinitely differentiable over (w_0, w_1), which implies $\theta(w)$ also satisfies condition (iv) over (w_0, w_1).

Apparently, condition (iv) is satisfied when w and s are in the same region (zero or non-zero region) by taking $t = 1$. We only verify that condition (iv) is valid when $w \in [0, w_0)$ and $s \in [w_0, w_1]$. The other situations can be verified similarly.

To proceed, we notice

$$\begin{aligned}
|\theta^{(d)}(w) - \theta^{(d)}(s)| &= |\theta^{(d)}(w) - \theta^{(d)}(w_0) + \theta^{(d)}(w_0) - \theta^{(d)}(s)| \\
&\leq |\theta^{(d)}(w) - \theta^{(d)}(w_0)| + |\theta^{(d)}(w_0) - \theta^{(d)}(s)| \\
&\leq C_1|w - w_0| + C_2|w_0 - s| \\
&\leq \max\{C_1, C_2\}|w - s|.
\end{aligned}$$

Hence, condition (iv) is valid for $t = 1$.

To prove condition (iii), we just need to find a_0 and a_1 such that $\theta'(w) \geq 0$ over $[w_0, w_1]$. By the construction of $\theta(w)$, we have $\theta'(w) = 0$ over $[a_0, a_1]$. When

$w \in (a_1, w_1)$, we let $r_1 = w_1 - a_1$ and show

$$\begin{aligned}
|\theta_1^{*'}(w)| &= \left| \sum_{j=1}^{d} \frac{b^{(j)}(w_1)}{(j-1)!}(w - w_1)^{j-1} \right| \\
&\leq \sum_{j=1}^{d} \left| \frac{b^{(j)}(w_1)}{(j-1)!}(w - w_1)^{j-1} \right| \leq M \sum_{j=1}^{d} r_1^{j-1} \leq \frac{M}{1 - r_1}, \\
\theta_1^*(w) &\geq \alpha - \left| \sum_{j=1}^{d} \frac{\beta^{(j)}(w_1)}{j!}(w - w_1)^{j} \right| \\
&\geq \alpha - \sum_{j=1}^{d} \left| \frac{\beta^{(j)}(w_1)}{j!}(w - w_1)^{j} \right| \geq \alpha - \frac{M r_1}{1 - r_1}, \\
\text{and } f_1^{'}(w, a_1) &= \frac{e^{-1/(w-a_1)-1/(w_1-w)} \left\{ 1/(w-a_1)^2 + 1/(w_1-w)^2 \right\}}{\left\{ e^{-1/(w-a_1)} + e^{-1/(w_1-w)} \right\}^2} \\
&\geq \frac{1/(w-a_1)^2 + 1/(w_1-w)^2}{2^2} \\
&\geq \frac{1}{2r_1^2}.
\end{aligned}$$

Then

$$\begin{aligned}
\theta^{'}(w) &= \theta_1^{*'}(w) f_1(w, a_1) + \theta_1^*(w) f_1^{'}(w, a_1) \\
&\geq \theta_1^*(w) f_1^{'}(w, a_1) - |\theta_1^{*'}(w) f_1(w, a_1)| \\
&\geq \left(\alpha - \frac{M r_1}{1 - r_1} \right) \frac{1}{2r_1^2} - \frac{M}{1 - r_1}.
\end{aligned}$$

Let

$$g(r) = \left(\alpha - \frac{Mr}{1 - r} \right) \frac{1}{2r^2} - \frac{M}{1 - r},$$

then when $0 < r < 1$,

$$g^{'}(r) = -\frac{\alpha}{r^3} - \frac{M}{2r^2(1-r)^2} - \frac{M}{(1-r)^2} < 0.$$

Therefore, $g(r)$ is strictly decreasing on $(0, 1)$. As $\lim_{r \downarrow 0} g(r) = \infty$ and $\lim_{r \uparrow 1} g(r) = -\infty$, there exists a unique $r^* \in (0, 1)$ such that $g(r^*) = 0$. Therefore, $g(r) > 0$ over

$(0, r^*)$. Let $r_1 = \min\{r^*, (w_1 - w_0)/2\}$, and we have $\theta'(w) > 0$ over $(w_1 - r_1, w_1)$. Thus, we find an $a_1 = w_1 - r_1$ such that $|\theta(w)| \leq \alpha$ over (a_1, w_1). Similarly, we can find an a_0 such that $|\theta(w)| \leq \alpha$ over (w_0, a_0). Therefore, condition (iii) is satisfied.

Combining all the results, we have found a $\theta \in \mathbb{F}_0$ such that $\zeta_{(\theta,\alpha)}(w) = \beta(w)$, which completes the proof. ■

Lemma 2 *Under Conditions (**Ch2.C1**), (**Ch2.C5**), and (**Ch2.C7**), if $\beta_j \in \mathbb{S}_{q,\alpha_j}$ for $j = 1, \ldots, \tilde{p}$ with q and α_j the same as in the penalized likelihood, then $||\tilde{\boldsymbol{\beta}} - \beta_0||_\infty = O((\tilde{p}\rho)^{1/2})$; if $\beta_j \notin \mathbb{S}_{q,\alpha_j}$ for $j = 1, \ldots, \tilde{p}$, we have $||\tilde{\boldsymbol{\beta}} - \beta_0||_\infty = O((\tilde{p}\rho + \tilde{p}q^{-2m})^{1/2})$, where m is the smoothness parameter as in Definition 2.2.1.*

Proof of Lemma 2:

Let $l_0(\beta; \boldsymbol{X}, Y, W) = \left[Y - \sum_{j=1}^p X_j b_j(W)\right]^2$. By model assumption, we have $\mathrm{E}_{Y|\boldsymbol{X},W} Y = \sum_{j=1}^p X_j b_{0j}(W)$, then the true parameter $\beta_0 = (\beta_{01}, \ldots, \beta_{0p})^T = \arg\min_{\beta \in \mathbb{H}^p} \mathrm{E} l_0(\beta; \boldsymbol{X}, Y$

By definition, we have $l(\theta; \boldsymbol{X}, Y, W) = \left[Y - \sum_{j=1}^p X_j \zeta_{\{\theta_j,\alpha_j\}}(W)\right]^2 + \rho \sum_{j=1}^p \{\theta_j(W)\}^2$ and $\tilde{\theta} = (\boldsymbol{B}^T\tilde{\boldsymbol{\gamma}}_1, \ldots, \boldsymbol{B}^T\tilde{\boldsymbol{\gamma}}_p)^T = \arg\min_{\theta \in \mathbb{F}^p} \mathrm{E} l(\theta; \boldsymbol{X}, Y, W)$. Since $\beta_{0j} = 0$ for $j > \tilde{p}$, we can infer that $\tilde{\theta}_j = 0$ for $j > \tilde{p}$, and thus $\tilde{\beta}_j = 0$ for $j > \tilde{p}$.

Then by calculation,

$$
\begin{aligned}
&\mathrm{E} l_0(\beta_0; \boldsymbol{X}, Y, W) - \mathrm{E} l(\tilde{\theta}; \boldsymbol{X}, Y, W) \\
=&\mathrm{E}\left[Y - \sum_{j=1}^p X_j b_{0j}(W)\right]^2 - \mathrm{E}\left[Y - \sum_{j=1}^p X_j \tilde{\beta}_j(W)\right]^2 - \rho \mathrm{E} \sum_{j=1}^p \left\{\tilde{\theta}_j(W)\right\}^2 \\
=&\mathrm{E}\left[\sum_{j=1}^p X_j \left\{\tilde{\beta}_j(W) - b_{0j}(W)\right\}\right]\left[2Y - \sum_{j=1}^p X_j b_{0j}(W) - \sum_{j=1}^p X_j \tilde{\beta}_j(W)\right] - \rho \mathrm{E} \sum_{j=1}^p \left\{\tilde{\theta}_j(W)\right\}^2 \\
=&-\mathrm{E}\left[\sum_{j=1}^p X_j \left\{\tilde{\beta}_j(W) - b_{0j}(W)\right\}\right]^2 - \rho \mathrm{E} \sum_{j=1}^p \left\{\tilde{\theta}_j(W)\right\}^2.
\end{aligned}
\tag{A.1}
$$

According to Lemma 1, for $j = 1, \ldots, \tilde{p}$, there exists $\theta_j \in \mathbb{F}_0$ such that $\zeta_{\{\theta_j,\alpha_j\}} =$

β_{0j}. If $\theta_j \notin \mathbb{F}$, then we can find $\theta_j^* \in \mathbb{F}$ such that $||\theta_j - \theta_j^*||_2 = O(q^{-m})$. When $j > \tilde{p}$, let $\theta_j^* = 0$, then we have $\zeta_{\{\theta_j^*,\alpha_j\}} = 0 = \beta_{0j}$. Let $\beta^*(w) = (\beta_1^*, \ldots, \beta_p^*)^T$, where $\beta_j^* = \zeta_{\left\{\theta_j^*,\alpha_j\right\}}(w)$. Then by Condition (**Ch2.C1**) and (**Ch2.C5**), we have

$$
\begin{aligned}
&\mathrm{E}l_0(\beta^*; \boldsymbol{X}, Y, W) - \mathrm{E}l_0(\beta_0; \boldsymbol{X}, Y, W) = \mathrm{E}\left[\sum_{j=1}^{p} X_j \left\{\beta_j^*(W) - b_{0j}(W)\right\}\right]^2 \\
=&\mathrm{E}\left[\sum_{j,k} X_j X_k \left\{\beta_j^*(W) - b_{0j}(W)\right\}\left\{\beta_k^*(W) - b_k(W)\right\}\right] \\
=&\mathrm{E}\left[\left\{\beta_1^*(W) - b_{01}(W), \ldots, \beta_p^*(W) - b_{0p}(W)\right\} \mathrm{E}(\boldsymbol{X}\boldsymbol{X}^T) \left\{\beta_1^*(W) - b_{01}(W), \ldots, \beta_p^*(W) - b_{0p}(W)\right\}^T\right] \\
\leq&\lambda_p \mathrm{E}\sum_{j=1}^{p}(\beta_j^*(W) - b_{0j}(W))^2 = \lambda_p \sum_{j=1}^{\tilde{p}} ||\beta_j^*(W) - b_{0j}(W)||_2^2 \\
=&O(\tilde{p}q^{-2m}).
\end{aligned}
\tag{A.2}
$$

If for $j = 1, \ldots, \tilde{p}$, $\theta_j \in \mathbb{F}$, let $\theta_j^* = \theta_j$, then we have $\beta^* = \beta$ and $\mathrm{E}l_0(\beta^*; \boldsymbol{X}, Y, W) - \mathrm{E}l_0(\beta_0; \boldsymbol{X}, Y, W) = 0$. Here, we assume all β_{0j} $(j = 1, \ldots, \tilde{p})$ have the same smoothness, either $\beta_j \in \mathbb{S}_{q,\alpha_j}$ for $j = 1, \ldots, \tilde{p}$, or $\beta_j \notin \mathbb{S}_{q,\alpha_j}$ for $j = 1, \ldots, \tilde{p}$.

By definition of $\tilde{\theta}$, we have $\mathrm{E}l(\tilde{\theta}) \leq \mathrm{E}l(\theta^*) = \mathrm{E}l_0(\beta^*) + \rho\mathrm{E}\sum_{j=1}^{p}\left\{\theta_j^*(W)\right\}^2$. Therefore, $\mathrm{E}l(\tilde{\theta}) - \mathrm{E}l_0(\beta^*) \leq \rho\mathrm{E}\sum_{j=1}^{p}\left\{\theta_j^*(W)\right\}^2$. If $\theta_j \notin \mathbb{F}$ for all $j \leq \tilde{p}$, based on equation (A.1), (A.2) and Condition (**Ch2.C7**), we have

$$
\begin{aligned}
\mathrm{E}\left[\sum_{j=1}^{p} X_j \left\{\tilde{\beta}_j(W) - b_{0j}(W)\right\}\right]^2 &= \mathrm{E}l(\tilde{\theta}) - \mathrm{E}l_0(\beta_0) - \rho\mathrm{E}\sum_{j=1}^{p}\left\{\tilde{\theta}_j(W)\right\}^2 \\
&\leq \mathrm{E}l(\tilde{\theta}) - \mathrm{E}l_0(\beta^*) + \mathrm{E}l_0(\beta^*) - \mathrm{E}l_0(\beta) - \rho\mathrm{E}\sum_{j=1}^{p}\left\{\tilde{\theta}_j(W)\right\}^2 \\
&\leq \rho\mathrm{E}\sum_{j=1}^{p}\left\{\theta_j^*(W)\right\}^2 - \rho\mathrm{E}\sum_{j=1}^{p}\left\{\tilde{\theta}_j(W)\right\}^2 + \mathrm{E}l_0(\beta^*) - \mathrm{E}l_0(\beta) \\
&= O(\tilde{p}\rho + \tilde{p}q^{-2m}).
\end{aligned}
$$

If $\theta_j \in \mathbb{F}$ for all j, then $\mathrm{E}\left[\sum_{j=1}^{p} X_j \left\{\tilde{\beta}_j(W) - b_{0j}(W)\right\}\right]^2 = O(\tilde{p}\rho)$.

By Condition (**Ch2.C1**) and (**Ch2.C5**), we also have

$$\mathrm{E}\left[\sum_{j=1}^{p} X_j \left\{\tilde{\beta}_j(W) - b_{0j}(W)\right\}\right]^2 = \mathrm{E}\left[\sum_{j,k} X_j X_k \left\{\tilde{\beta}_j(W) - b_{0j}(W)\right\}\left\{\tilde{\beta}_k(W) - b_{0k}(W)\right\}\right]$$

$$= \mathrm{E}\left[\left\{\tilde{\beta}_1(W) - b_{01}(W), \ldots, \tilde{\beta}_p(W) - b_{0p}(W)\right\} \mathrm{E}(\boldsymbol{X}\boldsymbol{X}^T) \left\{\tilde{\beta}_1(W) - b_{01}(W), \ldots, \tilde{\beta}_p(W) - b_{0p}(W)\right\}^T\right]$$

$$\geq \lambda_1 \mathrm{E} \sum_{j=1}^{p} (\tilde{\beta}_j(W) - b_{0j}(W))^2 = \lambda_1 \sum_{j=1}^{p} ||\tilde{\beta}_j(W) - b_{0j}(W)||_2^2.$$

Therefore, $\max_{1\leq j\leq p} ||\tilde{\beta}_j - b_{0j}||_2^2 = O(\mathrm{E}\left[\sum_{j=1}^{p} X_j \left\{\tilde{\beta}_j(W) - b_{0j}(W)\right\}\right]^2)$. In addition, $||\tilde{\boldsymbol{\beta}} - \beta_0||_\infty = \max_{1\leq j\leq p} ||\tilde{\beta}_j - b_{0j}||_\infty \leq \max_{1\leq j\leq p} ||\tilde{\beta}_j - b_{0j}||_2$. Combining all above results, we conclude: if $\beta_j \notin \mathbb{S}_{q,\alpha_j}$ for $j = 1, \ldots, \tilde{p}$, we have $||\tilde{\boldsymbol{\beta}} - \beta_0||_\infty = O((\tilde{p}\rho + \tilde{p}q^{-2m})^{1/2})$; if $\beta_j \in \mathbb{S}_{q,\alpha_j}$ for $j = 1, \ldots, \tilde{p}$, $||\tilde{\boldsymbol{\beta}} - \beta_0||_\infty = O((\tilde{p}\rho)^{1/2})$. ■

We introduce two important lemmas in order to prove our main theorems. Lemma 3 is a variation of the Lyapunov central limit theorem and will be used in the proof of Lemma 5, and Lemma 4 is used in the proof of Theorem 2.3.1.

Lemma 3 *Suppose ϵ_i are independent with mean 0 and variance 1, and ϵ_i satisfy Condition (**Ch2.C3**). If $\max_i a_i^2/(\sum_i a_i^2) \to 0$, then*

$$\frac{\sum_i a_i \epsilon_i}{\sqrt{(\sum_i a_i^2)}} \to_d N(0,1).$$

Lemma 4 (Consistency) *Under Conditions (**Ch2.C1**), (**Ch2.C2**), (**Ch2.C4**), (**Ch2.C6**) and (**Ch2.C7**),*

$$||\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}||_2^2 = o_p(\tilde{p}q^{-1}),$$

where $\tilde{\boldsymbol{\theta}} = \boldsymbol{B}\tilde{\boldsymbol{\gamma}}$.

Proof of Lemma 4:

Let $\boldsymbol{\theta}^* = (\theta_1^*, \ldots, \theta_p^*)^T$. We choose $\theta_j^* \in \mathbb{F}$ such that $||\boldsymbol{\theta}_j^*||_2^2 = O(q^{-1})$ for $j = 1, \ldots, p$. Let $T_n(a) = M_n(\tilde{\boldsymbol{\theta}} + a\boldsymbol{\theta}^*)$. The derivative of T_n with respect to a is

$$T_n'(a) = -2\mathrm{E}_n\left[\left\{Y - \sum_{j=1}^{p} X_j h_j(\tilde{\theta}_j + a\theta_j^*)\right\}\sum_{j=1}^{p} X_j h_j'(\tilde{\theta}_j + a\theta_j^*)\theta_j^* - \rho\sum_{j=1}^{p}(\tilde{\theta}_j + a\theta_j^*)\theta_j^*\right]. \tag{A.3}$$

When a is sufficiently small, T_n is convex. Thus, T_n' is non-decreasing. Therefore, we only need to show that for any small $a_0 > 0$, $-T_n'(a_0) < 0$ and $-T_n'(-a_0) > 0$. Then, $||\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}||_2 \leq a_0||\boldsymbol{\theta}^*||_2$. Since $\tilde{\boldsymbol{\gamma}} = \arg\min_{\boldsymbol{\gamma}} \mathrm{E}l(\boldsymbol{\gamma}; \boldsymbol{X}, Y, W)$, then $\tilde{\theta}_j = \boldsymbol{B}\tilde{\boldsymbol{\gamma}}_j \equiv 0$ for $l > \tilde{p}$. By Condition (**Ch2.C6**), $\alpha_j > ||\theta_j^*||_2$ for $l > \tilde{p}$. Thus, $h_j(\tilde{\theta}_j + a\theta_j^*) \equiv 0$ for $l > \tilde{p}$. Then, we have $\sum_{j=1}^{p} X_j h_j(\tilde{\theta}_j + a\theta_j^*) = \sum_{j=1}^{\tilde{p}} X_j h_j(\tilde{\theta}_j + a\theta_j^*)$.

From (A.3), we have

$$\begin{aligned}
-\frac{1}{2}T_n'(a_0) =& \mathrm{E}_n\left[\left\{Y - \sum_{j=1}^{p} X_j h_j(\tilde{\theta}_j + a_0\theta_j^*)\right\}\cdot\left\{\sum_{j=1}^{p} X_j h_j'(\tilde{\theta}_j + a_0\theta_j^*)\theta_j^*\right\} - \rho\sum_{j=1}^{p}(\tilde{\theta}_j + a_0\theta_j^*)\theta_j^*\right] \\
=& \mathrm{E}_n\left\{Y - \sum_{j=1}^{p} X_j\tilde{\beta}_j\right\}\cdot\left\{\sum_{j=1}^{p} X_j h_j'(\tilde{\theta}_j + a_0\theta_j^*)\theta_j^*\right\} + \\
& \mathrm{E}_n\left\{\sum_{j=1}^{p} X_j\tilde{\beta}_j - \sum_{j=1}^{p} X_j h_j(\tilde{\theta}_j)\right\}\cdot\left\{\sum_{j=1}^{p} X_j h_j'(\tilde{\theta}_j + a_0\theta_j^*)\theta_j^*\right\} + \\
& \mathrm{E}_n\left\{\sum_{j=1}^{p} X_j h_j(\tilde{\theta}_j) - \sum_{j=1}^{p} X_j h_j(\tilde{\theta}_j + a_0\theta_j^*)\right\}\cdot\left\{\sum_{j=1}^{p} X_j h_j'(\tilde{\theta}_j + a_0\theta_j^*)\theta_j^*\right\} - \\
& \rho\mathrm{E}_n\sum_{j=1}^{p}(\tilde{\theta}_j + a_0\theta_j^*)\theta_j^* \\
=& A_1 + A_2 + A_3 + A_4,
\end{aligned}$$

where $\tilde{\beta}_j = \zeta_{(\tilde{\theta}_j, \alpha_j)}$.

By the definition of h_j, we have that $|h_j'(\tilde{\theta}_j + a_0\theta_j^*)| \leq 1$ for $j = 1, \ldots, \tilde{p}$ and $|h_j'(\tilde{\theta}_j + a_0\theta_j^*)| \equiv 0$ for $j = \tilde{p} + 1, \ldots, p$. Let $h_n = \sum_{j=1}^{p} X_j h_j'(\tilde{\theta}_j + a_0\theta_j^*)\theta_j^*$. Then

$\mathrm{E}(h_n^2) = O(\sum_{j=1}^{\tilde{p}} ||\boldsymbol{\theta}_j^*||_2^2) = O(\tilde{p}q^{-1})$ by Condition (**Ch2.C2**). Since $Y - \sum_{j=1}^{p} X_j\tilde{\beta}_j = \epsilon$, by Chebyshev's inequality, we have

$$\Pr(|A_1| > 1/\sqrt{n}) \leq \frac{\mathrm{E}(\mathrm{E}_n h_n \epsilon)^2}{1/n} \leq \frac{\mathrm{E}\{(\mathrm{E}_n h_n)^2(\mathrm{E}_n \epsilon)^2\}}{1/n} = \frac{O(\mathrm{E}h_n^2)\mathrm{E}\epsilon^2/n}{1/n} = O(\tilde{p}q^{-1})\sigma^2.$$

Therefore, $|A_1| = o_p(n^{-1/2}) = o_p(\tilde{p}q^{-1})$.

By the definition of $\tilde{\boldsymbol{\gamma}}$, it satisfies the score equation

$$0 = \mathrm{E}l^{s'} = -2\mathrm{E}\left\{(Y - \boldsymbol{X}^T\tilde{\boldsymbol{h}}) \cdot \tilde{\boldsymbol{U}} \otimes \boldsymbol{B}(W) - \rho\tilde{\boldsymbol{\theta}} \otimes \boldsymbol{B}(W)\right\}, \tag{A.4}$$

where $\tilde{\boldsymbol{h}}$, $\tilde{\boldsymbol{U}}$, $\tilde{\boldsymbol{\theta}}$ are $\boldsymbol{h}$, $\boldsymbol{U}$, $\boldsymbol{\theta}$ with $\boldsymbol{\gamma}$ replaced by $\tilde{\boldsymbol{\gamma}}$ respectively. Since $\boldsymbol{B}(W) \neq 0$ for any $W \in \mathbb{D}$, equation (A.4) becomes $\mathrm{E}\left\{(Y - \boldsymbol{X}^T\tilde{\boldsymbol{h}}) \cdot \tilde{\boldsymbol{U}} - \rho\tilde{\boldsymbol{\theta}}\right\} = 0$. We then have $\mathrm{E}\big[\tilde{\boldsymbol{U}}\boldsymbol{X}^T\{\tilde{\boldsymbol{\beta}} - \boldsymbol{h}(\tilde{\boldsymbol{\gamma}})\} - \rho\tilde{\boldsymbol{\theta}}\big] = 0$, because $Y - \boldsymbol{X}^T\tilde{\boldsymbol{\beta}} = \epsilon$. Note that $\mathrm{E}(\tilde{\boldsymbol{U}}\boldsymbol{X}^T)$ is invertible according to Condition (**Ch2.C4**), then we have $(\tilde{\boldsymbol{\beta}} - \tilde{\boldsymbol{h}}) = \rho\{\mathrm{E}(\tilde{\boldsymbol{U}}\boldsymbol{X}^T)\}^{-1}\tilde{\boldsymbol{\theta}}$. By the Cauchy-Schwarz inequality and Condition (**Ch2.C1**), (**Ch2.C2**) and (**Ch2.C6**),

$$\begin{aligned}|A_2|^2 &\leq \left(\frac{1}{n}\sum_{i=1}^{n} h_n^2\right)\left(\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{j=1}^{p} X_j\left\{\tilde{\beta}_j - h_j(\tilde{\theta}_j)\right\}\right]^2\right) = O_p(q^{-1})O_p\left(\mathrm{E}\left\{\boldsymbol{X}^T(\tilde{\boldsymbol{\beta}} - \tilde{\boldsymbol{h}})\right\}^2\right) \\ &= O_p(q^{-1})O_p\left(\mathrm{E}\left[\rho\boldsymbol{X}^T\left\{\mathrm{E}(\tilde{\boldsymbol{U}}\boldsymbol{X}^T)\right\}^{-1}\tilde{\boldsymbol{\theta}}\right]^2\right) = O_p(\rho^2\tilde{p}q^{-1}).\end{aligned}$$

Hence, $A_2 = o_p(\tilde{p}q^{-1})$.

Moreover, we have $A_3 = O\Big(-\mathrm{E}_n\big\{\sum_{j=1}^{\tilde{p}} X_j\theta_j^*\big\}^2\Big) = -a_0O_p(\tilde{p}q^{-1})$ and $A_4 = -O_p(\rho\tilde{p} + \rho a_0 pq^{-1}) = o_p(\tilde{p}q^{-1})$ by Condition (**Ch2.C6**).

Therefore, we have

$$-\frac{1}{2}T_n'(a_0) = o_p(\tilde{p}q^{-1}) + o_p(\tilde{p}q^{-1}) - a_0O_p(\tilde{p}q^{-1}) + o_p(\tilde{p}q^{-1}) = -a_0O_p(\tilde{p}q^{-1}) < 0,$$

if $a_0 > 0$ and $H_n'(a_0) > 0$, if $a_0 < 0$. Thus, $||\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}||_2^2 = o_p(\tilde{p}q^{-1})$. The proof is

completed. ■

Proof of Theorem 2.3.1:

By the definitions of M_n and M_0, we have

$$
\begin{aligned}
&(\mathrm{M}_n - \mathrm{M}_0)(\boldsymbol{\theta}) \\
=&(\mathrm{E}_n - \mathrm{E})\Big[- \Big\{Y - \sum_{j=1}^{p} X_j h_j(\theta_j)\Big\}^2 - \rho \sum_{j=1}^{p} \theta_j^2 \Big] \\
=&(\mathrm{E}_n - \mathrm{E})\Big[- \Big(Y - \sum_{j=1}^{p} X_j \tilde{\beta}_j\Big)^2 - \Big\{ \sum_{j=1}^{p} X_j \tilde{\beta}_j - \sum_{j=1}^{p} X_j h_j(\theta_j)\Big\}^2 - \\
&2(Y - \sum_{j=1}^{p} X_j \tilde{\beta}_j)\Big\{ \sum_{j=1}^{p} X_j \tilde{\beta}_j - \sum_{j=1}^{p} X_j h_j(\theta_j)\Big\} - \rho \sum_{j=1}^{p} \theta_j^2 \Big] \\
=&(\mathrm{E}_n - \mathrm{E})\Big[- \epsilon^2 - \Big\{ \sum_{j=1}^{p} X_j \tilde{\beta}_j - \sum_{j=1}^{p} X_j h_j(\theta_j)\Big\}^2 - 2\Big\{ \sum_{j=1}^{p} X_j \tilde{\beta}_j - \sum_{j=1}^{p} X_j h_j(\theta_j)\Big\}\epsilon - \rho \sum_{j=1}^{p} \theta_j^2 \Big].
\end{aligned}
$$

Therefore, we have

$$
\begin{aligned}
&(\mathrm{M}_n - \mathrm{M}_0)(\boldsymbol{\theta}) - (\mathrm{M}_n - \mathrm{M}_0)(\tilde{\boldsymbol{\theta}}) \\
=&2\mathrm{E}_n\Big[\Big\{ \sum_{j=1}^{p} X_j h_j(\theta_j) - \sum_{j=1}^{p} X_j h_j(\tilde{\theta}_j)\Big\}\epsilon\Big] - (\mathrm{E}_n - \mathrm{E})\Big\{ \Big[\sum_{j=1}^{p} X_j \big\{h_j(\theta_j) - h_j(\tilde{\theta}_j)\big\}\Big]^2 \Big\} + \\
&2(\mathrm{E}_n - \mathrm{E})\Big[\sum_{j=1}^{p} X_j \Big\{h_j(\theta_j) - h_j(\tilde{\theta}_j)\Big\}\Big]\Big[\sum_{j=1}^{p} X_j \Big\{\tilde{\beta}_j - h_j(\tilde{\theta}_j)\Big\}\Big] - \\
&\rho(\mathrm{E}_n - \mathrm{E})\left\{ \sum_{j=1}^{p} (\theta_j - \tilde{\theta}_j)(\theta_j + \tilde{\theta}_j) \right\} \\
=&B_1 + B_2 + B_3 + B_4
\end{aligned}
$$

For $j = 1, \ldots, p$, let

$$G_j = \left\{\theta_j : ||\theta_j - \tilde{\theta}_j||_2 \leq \delta, 0 < \delta < 1, \theta_j \in \mathbb{F}\right\},$$
$$H_j = \left\{h_j(\theta_j) : ||\theta_j - \tilde{\theta}_j||_2 \leq \delta, 0 < \delta < 1, \theta_j \in \mathbb{F}\right\},$$
$$S_j = \left\{X_j h_j(\theta_j) : ||\theta_j - \tilde{\theta}_j||_2 \leq \delta, 0 < \delta < 1, \theta_j \in \mathbb{F}\right\},$$

and

$$S = \left\{\sum_{j=1}^{p} X_j h_j(\theta_j) : ||\theta_j - \tilde{\theta}_j||_2 \leq \delta, 0 < \delta < 1, \theta_j \in \mathbb{F}, j = 1, \ldots p\right\}, \tag{A.5}$$

where $\mathbb{F} = \left\{\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T : \theta_j \in \mathbb{F}, j = 1, \ldots, p\right\}$.

Since $|h_j(\theta_j) - h_j(\tilde{\theta}_j)| \leq |\theta_j - \tilde{\theta}_j|$, we have $N_{[]}\{\delta_1, H_j, \mathcal{L}_2(\mathbb{D})\} \simeq N_{[]}\{\delta_1, G_j, \mathcal{L}_2(\mathbb{D})\}$. By Condition (**Ch2.C1**), we further have $N_{[]}\{(C_2 - C_1)\delta_1, S_j, \mathcal{L}_2(\mathbb{D})\} \simeq N_{[]}\{\delta_1, G_j, \mathcal{L}_2(\mathbb{D})\}$.

By Condition (**Ch2.C6**), we have $\alpha_j > \delta$ for $j = 1, \ldots, p$. Then by the definition of $\tilde{\theta}_j$, we have

$$S = \left\{\sum_{j=1}^{\tilde{p}} X_j h_j(\theta_j) : ||\theta_j - \tilde{\theta}_j||_2 \leq \delta, 0 < \delta < 1, \theta_j \in \mathbb{F}, j = 1, \ldots \tilde{p}\right\}.$$

According to the construction of S, we have that

$$N_{[]}\big(\tilde{p}(C_2 - C_1)\delta_1, S, \mathcal{L}_2(\mathbb{D})\big) \simeq \left\{N_{[]}((C_2 - C_1)\delta_1, S_j, \mathcal{L}_2(\mathbb{D}))\right\}^{\tilde{p}} \simeq \left\{N_{[]}(\delta_1, G_j, \mathcal{L}_2(\mathbb{D}))\right\}^{\tilde{p}},$$

since the bracket numbers are the same over j for S_j as well as G_j.

From the calculation by Shen and Wong (1994) [72], $\log N_{[]}\{\delta_1, G_j, \mathcal{L}_2(\mathbb{D})\} = c_1 q \log(\delta/\delta_1)$, we have $\log N_{[]}\left\{\tilde{p}(C_2 - C_1)\delta_1, S, \mathcal{L}_2(\mathbb{D})\right\} \simeq c_1 \tilde{p} q \log(\delta/\delta_1)$.

By Condition (**Ch2.C3**), the stochastic process $\Big\{\sqrt{n}\mathrm{E}_n\big[\{\sum_{j=1}^{\tilde{p}} X_j h_j(\theta_j) - \sum_{j=1}^{\tilde{p}} X_j h_j(\tilde{\theta}_j)\}\epsilon\big], \theta_j \in \mathbb{F}, j = 1, \ldots, \tilde{p}\Big\}$ is sub-Gaussian for the $\mathcal{L}_2(\mathbb{D})$-semimetric on S. According to Corol-

lary 2.2.8 of Van Der Vaart and Wellner (1996) [85], we have

$$\mathrm{E}^*\left\{\sup_{||\boldsymbol{\theta}-\tilde{\boldsymbol{\theta}}||_2<\delta,\boldsymbol{\theta}\in\mathbb{F}_n^{Ln}}\sqrt{n}|B_1|\right\}\simeq\int_0^{\delta}\sqrt{\log N_{[]}\left\{\tilde{p}\delta_1,S,\mathcal{L}_2(\mathbb{D})\right\}}\,\mathrm{d}(\tilde{p}\delta_1)\simeq(\tilde{p}q)^{1/2}\delta.$$

With the similar calculation of the bracketing number and Lemma 3.4.2 of Van Der Vaart and Wellner (1996) [85], we have

$$\mathrm{E}^*\left\{\sup_{||\boldsymbol{\theta}-\tilde{\boldsymbol{\theta}}||_2<\delta,\boldsymbol{\theta}\in\mathbb{F}_n^{Ln}}\sqrt{n}|B_2|\right\}\simeq(\tilde{p}q)^{1/2}\delta.$$

Since $\boldsymbol{X}^T(\tilde{\boldsymbol{\beta}}-\tilde{\boldsymbol{h}})=\rho\boldsymbol{X}^T\{\mathrm{E}(\tilde{\boldsymbol{U}}\boldsymbol{X}^T)\}^{-1}\tilde{\boldsymbol{\theta}}=O_p(\rho||\tilde{\boldsymbol{\theta}}||)$ is bounded, we can also have

$$\mathrm{E}^*\left\{\sup_{||\boldsymbol{\theta}-\tilde{\boldsymbol{\theta}}||_2<\delta,\boldsymbol{\theta}\in\mathbb{F}_n^{Ln}}\sqrt{n}|B_3|\right\}\simeq\mathrm{E}^*\left\{\sup_{||\boldsymbol{\theta}-\tilde{\boldsymbol{\theta}}||_2<\delta,\boldsymbol{\theta}\in\mathbb{F}_n^{Ln}}\sqrt{n}|B_1|\right\}\simeq(\tilde{p}q)^{1/2}\delta.$$

By Condition (**Ch2.C7**), $|\theta_j+\tilde{\theta}_j|$ is bounded, then

$$\mathrm{E}^*\left\{\sup_{||\boldsymbol{\theta}-\tilde{\boldsymbol{\theta}}||_2<\delta,\boldsymbol{\theta}\in\mathbb{F}_n^{Ln}}\sqrt{n}|B_4|\right\}\simeq\mathrm{E}^*\left\{\sup_{||\boldsymbol{\theta}-\tilde{\boldsymbol{\theta}}||_2<\delta,\boldsymbol{\theta}\in\mathbb{F}_n^{Ln}}\sqrt{n}|B_1|\right\}\simeq(\tilde{p}q)^{1/2}\delta.$$

According to Theorem 3.4.1 of Van Der Vaart and Wellner (1996) [85], the key function $\phi(\delta)$ takes the form of $\phi_n(\delta)=(\tilde{p}q)^{1/2}\delta$. Therefore, $||\hat{\boldsymbol{\theta}}-\tilde{\boldsymbol{\theta}}||_2=O_p\left((\tilde{p}q/n)^{1/2}\right)$.

By Lemma 2 and Condition (**Ch2.C6**), If $\beta_j\notin\mathbb{S}_{q,\alpha_j}$ for $j=1,\ldots,\tilde{p}$, then

$$\begin{aligned}
||\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0||_2&=||\zeta_{(\hat{\boldsymbol{\theta}},\alpha)}-\boldsymbol{\beta}_0||_2\\
&\leq||\zeta_{(\hat{\boldsymbol{\theta}},\alpha)}-h(\hat{\boldsymbol{\theta}})||_2+||h(\hat{\boldsymbol{\theta}})-h(\tilde{\boldsymbol{\theta}})||_2+||h(\tilde{\boldsymbol{\theta}})-\tilde{\boldsymbol{\beta}}||_2+||\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_0||_2\\
&=O_p(\tilde{p}^{1/2}\nabla(\eta))+O_p(||\hat{\boldsymbol{\theta}}-\tilde{\boldsymbol{\theta}}||_2)+O(\tilde{p}^{1/2}\nabla(\eta))+O((\tilde{p}\rho+\tilde{p}q^{-2m})^{1/2})\\
&=O_p\left(\tilde{p}^{1/2}\nabla(\eta)+(\tilde{p}q/n)^{1/2}+\tilde{p}^{1/2}\nabla(\eta)+(\tilde{p}\rho+\tilde{p}q^{-2m})^{1/2}\right)\\
&=O_p\left((\tilde{p}q/n)^{1/2}+\tilde{p}^{1/2}q^{-m}\right);
\end{aligned}$$

if $\beta_j \in \mathbb{S}_{q,\alpha_j}$ for $j = 1, \ldots, \tilde{p}$, $||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0||_2 = O_p\left(\tilde{p}^{1/2}\nabla(\eta) + (\tilde{p}q/n)^{1/2} + (\tilde{p}\rho)^{1/2}\right) = O_p\left(\tilde{p}^{1/2}r(\eta) + (\tilde{p}q/n)^{1/2}\right)$. The proof is completed. ■

Lemma 5 (Normality) *Under Conditions (**Ch2.C1**)–(**Ch2.C7**), for $j = 1, \ldots, p$, and any $w \in \mathbb{D}$,*

$$\{\sigma_{nj}^2(w)\}^{-1/2}\left\{\hat{\theta}_j(w) - \tilde{\theta}_j(w)\right\} \to_d N(0,1),$$

where $\sigma_{nj}^2(w) = \sigma^2[n^2\{\boldsymbol{e}_j \otimes \boldsymbol{B}(w)\}^T\{l_n''(\tilde{\boldsymbol{\gamma}})\}^{-1}\left\{\boldsymbol{V}_n^T(\tilde{\boldsymbol{\gamma}})\boldsymbol{V}_n(\tilde{\boldsymbol{\gamma}})\right\}\{l_n''(\tilde{\boldsymbol{\gamma}})\}^{-1}\{\boldsymbol{e}_j \otimes \boldsymbol{B}(w)\}]^{-1}$.

Proof of Lemma 5:

By the Mean Value Theorem, there exists a $\boldsymbol{\gamma}^*$ between $\tilde{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\gamma}}$, such that

$$0 = l_n'(\hat{\boldsymbol{\gamma}}) = l_n'(\tilde{\boldsymbol{\gamma}}) + l_n''(\boldsymbol{\gamma}^*)(\hat{\boldsymbol{\gamma}} - \tilde{\boldsymbol{\gamma}}). \tag{A.6}$$

According to the previous calculation,

$$\begin{aligned} l_n'(\boldsymbol{\gamma}) &= -2\mathrm{E}_n\left\{(Y - \boldsymbol{X}^T\boldsymbol{h}) \cdot \boldsymbol{U} \otimes \boldsymbol{B}(W) - \rho\boldsymbol{\theta} \otimes \boldsymbol{B}(W)\right\} \\ &= -2\mathrm{E}_n\left\{\boldsymbol{U} \otimes \boldsymbol{B}(W)\epsilon + \boldsymbol{U} \otimes \boldsymbol{B}(W) \cdot \boldsymbol{X}^T(\boldsymbol{\beta} - \boldsymbol{h}) - \rho\boldsymbol{\theta} \otimes \boldsymbol{B}(W)\right\} \\ &= -2\mathrm{E}_n\left\{\boldsymbol{v}\epsilon + \boldsymbol{v} \cdot \boldsymbol{X}^T(\boldsymbol{\beta} - \boldsymbol{h}) - \rho\boldsymbol{\theta} \otimes \boldsymbol{B}(W)\right\}. \end{aligned} \tag{A.7}$$

Since $l_n''(\boldsymbol{\gamma}^*)$ is invertible, then we have $\hat{\boldsymbol{\gamma}} - \tilde{\boldsymbol{\gamma}} = -\{l_n''(\boldsymbol{\gamma}^*)\}^{-1}l_n'(\tilde{\boldsymbol{\gamma}})$. To prove the theorem, it suffices to show that for any $\boldsymbol{c}_n \in \mathbb{R}_{q*p}$ whose components are not all zero and $\boldsymbol{c}_n^T\boldsymbol{c}_n = O_p(q)$, $\boldsymbol{c}_n^T(\hat{\gamma} - \tilde{\gamma})/\mathrm{SD}\left\{\boldsymbol{c}_n^T(\hat{\gamma} - \tilde{\gamma})\right\} \to_d N(0,1)$, where

$$\mathrm{SD}\left\{\boldsymbol{c}_n^T(\hat{\gamma} - \tilde{\gamma})\right\} = \sqrt{(1/n^2)\boldsymbol{c}_n^T\left\{l_n''(\tilde{\boldsymbol{\gamma}})\right\}^{-1}\left\{\boldsymbol{V}_n^T(\tilde{\boldsymbol{\gamma}})\boldsymbol{V}_n(\tilde{\boldsymbol{\gamma}})\right\}\left\{l_n''(\tilde{\boldsymbol{\gamma}})\right\}^{-1}\boldsymbol{c}_n\sigma^2}.$$

By some algebra, we have

$$\begin{aligned}\boldsymbol{c}_n^T(\hat{\boldsymbol{\gamma}}-\tilde{\boldsymbol{\gamma}}) =& -\boldsymbol{c}_n^T\big\{l_n''(\boldsymbol{\gamma}^*)\big\}^{-1}l_n'(\tilde{\boldsymbol{\gamma}})\\ =&\sum_{i=1}^n a_i\epsilon_i^* + \boldsymbol{c}_n^T\big\{l_n''(\boldsymbol{\gamma}^*)\big\}^{-1}\mathrm{E}_n\big\{\boldsymbol{v}(\tilde{\boldsymbol{\gamma}})\cdot\boldsymbol{X}^T(\tilde{\boldsymbol{\beta}}-\tilde{\boldsymbol{h}})-\rho\tilde{\boldsymbol{\theta}}\otimes\boldsymbol{B}(W)\big\}\\ =&A_1+A_2,\end{aligned}$$

where $a_i = \boldsymbol{c}_n^T\big\{l_n''(\boldsymbol{\gamma}^*)\big\}^{-1}\boldsymbol{v}_i(\tilde{\boldsymbol{\gamma}})\sigma/n$ and ϵ_i^* are independent with mean zero and variance one conditioning on $\{\theta_i, W_i, i=1,\ldots,n\}$.

Since $\mathrm{E}_n\big\{\tilde{\boldsymbol{U}}\otimes\boldsymbol{B}(W)\boldsymbol{X}^T(\tilde{\boldsymbol{\beta}}-\tilde{\boldsymbol{h}})-\rho\tilde{\boldsymbol{\theta}}\otimes\boldsymbol{B}(W)\big\} = \rho\mathrm{E}_n\Big\{\Big[\tilde{\boldsymbol{U}}\boldsymbol{X}^T\big\{\mathrm{E}(\tilde{\boldsymbol{U}}\boldsymbol{X}^T)\big\}^{-1}\tilde{\boldsymbol{\theta}}-\tilde{\boldsymbol{\theta}}\Big]\otimes\boldsymbol{B}\Big\}$, we have $A_2 = o_p(\rho q^{1/2})$. Moreover,

$$\begin{aligned}\sum_{i=1}^n a_i^2 =& \frac{\sigma^2}{n^2}\sum_{i=1}^n \boldsymbol{c}_n^T\big\{l_n''(\boldsymbol{\gamma}^*)\big\}^{-1}\boldsymbol{v}_i(\tilde{\boldsymbol{\gamma}})\boldsymbol{v}_i^T(\tilde{\boldsymbol{\gamma}})\big\{l_n''(\boldsymbol{\gamma}^*)\big\}^{-1}\boldsymbol{c}_n\\ =&\frac{\sigma^2}{n}\boldsymbol{c}_n^T\big\{l_n''(\boldsymbol{\gamma}^*)\big\}^{-1}\frac{1}{n}\sum_{i=1}^n\boldsymbol{v}_i(\tilde{\boldsymbol{\gamma}})\boldsymbol{v}_i^T(\tilde{\boldsymbol{\gamma}})\cdot\big\{l_n''(\boldsymbol{\gamma}^*)\big\}^{-1}\boldsymbol{c}_n\\ =&O_p(\boldsymbol{c}_n^T\boldsymbol{c}_n/n) = O_p(q/n),\end{aligned}$$

thus we have $A_2/\sqrt{(\sum a_i^2)} = o_p(n\rho/q) = o_p(1)$ by Condition (**Ch2.C6**).

By Slutsky's Theorem, we then only need to prove $A_1/\sqrt{(\sum a_i^2)}$ follows a Normal distribution. By Condition (**Ch2.C3**) and Lemma 3, we only need to verify that $\max_i a_i^2/\sum_{i=1}^n a_i^2 \to_p 0$. With some calculations, we have

$$\begin{aligned}\max_{1\le i\le n} a_i^2 =& \frac{\sigma^2}{n^2}\max_{1\le i\le n}\Big[\boldsymbol{c}_n^T\left\{-l_n''(\boldsymbol{\gamma}^*)\right\}^{-1}\left\{\boldsymbol{V}_n^T(\tilde{\boldsymbol{\gamma}})\boldsymbol{V}_n(\tilde{\boldsymbol{\gamma}})\right\}^{1/2}\left\{\boldsymbol{V}_n^T(\tilde{\boldsymbol{\gamma}})\boldsymbol{V}_n(\tilde{\boldsymbol{\gamma}})\right\}^{-1/2}\boldsymbol{v}_i(\tilde{\boldsymbol{\gamma}})\Big]^2\\ \le&\frac{\sigma^2}{n^2}\boldsymbol{c}_n^T\left\{l_n''(\boldsymbol{\gamma}^*)\right\}^{-1}\cdot\boldsymbol{V}_n^T(\tilde{\boldsymbol{\gamma}})\boldsymbol{V}_n(\tilde{\boldsymbol{\gamma}})\cdot\left\{l_n''(\boldsymbol{\gamma}^*)\right\}^{-1}\boldsymbol{c}_n\cdot\\ &\max_{1\le i\le n}\boldsymbol{v}_i^T(\tilde{\boldsymbol{\gamma}})\left\{\boldsymbol{V}_n^T(\tilde{\boldsymbol{\gamma}})\boldsymbol{V}_n(\tilde{\boldsymbol{\gamma}})\right\}^{-1}\boldsymbol{v}_i(\tilde{\boldsymbol{\gamma}}).\end{aligned}$$

According to Condition (**Ch2.C2**), we have

$$\frac{\max_i a_i^2}{\sum_{i=1}^n a_i^2} = \max_{1\le i\le n} \boldsymbol{v}_i^T (\boldsymbol{V}_n^T \boldsymbol{V}_n)^{-1} \boldsymbol{v}_i \to_p 0,$$

as $n \to \infty$.

Because $\hat{\boldsymbol{\gamma}} \to_p \tilde{\boldsymbol{\gamma}}$, we have $\boldsymbol{\gamma}^* \to_p \tilde{\boldsymbol{\gamma}}$. Since for any $w \in \mathbb{D}$, $\hat{\theta}_j(w) = (\boldsymbol{e}_j \otimes \boldsymbol{B}(w))^T \hat{\boldsymbol{\gamma}}$, then let $\boldsymbol{c}_n = \boldsymbol{e}_j \otimes \boldsymbol{B}(w)$, we have

$$\{\sigma_{nj}^2(w)\}^{-1/2} \left\{\hat{\theta}_j(w) - \tilde{\theta}_j(w)\right\} \to_d N(0,1),$$

where $\sigma_{nj}^2(w) = \sigma^2/n^2 \left\{\boldsymbol{e}_j \otimes \boldsymbol{B}(w)\right\}^T \{l_n''(\tilde{\boldsymbol{\gamma}})\}^{-1} \left\{\boldsymbol{V}_n^T(\tilde{\boldsymbol{\gamma}}) \boldsymbol{V}_n(\tilde{\boldsymbol{\gamma}})\right\} \{l_n''(\tilde{\boldsymbol{\gamma}})\}^{-1} \{\boldsymbol{e}_j \otimes \boldsymbol{B}(w)\}$. The proof is completed. ■

Proof of Theorem 2.3.2:

It is straightforward to show that if $(Z-\mu)/\sigma \sim N(0,1)$, then

$$\Pr\left\{\zeta_{(Z,\alpha)} < x\right\} = \Phi\left(\frac{x+\alpha-\mu}{\sigma}\right) \mathcal{I}(x \ge 0) + \Phi\left(\frac{x-\alpha-\mu}{\sigma}\right) \mathcal{I}(x < 0).$$

Under regularity conditions and by Lemma 5, for $1 \le j \le p$ and any $w \in \mathbb{D}$, we have $\lim_{n\to\infty} \Pr\left(\sigma_{nj}^{-1}\hat{\theta}_j(w) - \sigma_{nj}^{-1}\tilde{\theta}_j(w) < x\right) = \Phi(x)$. Note that $\sigma_{nj}^{-1}\zeta_{\left\{\hat{\theta}_j,\alpha_j\right\}}(w) =$

$\zeta_{\left\{\sigma_{nj}^{-1}\hat{\theta}_j,\sigma_{nj}^{-1}\alpha_j\right\}}(w)$, then we have

$$\begin{aligned}
&\lim_{n\to\infty}\left|\Pr\left[\zeta_{\left\{\hat{\theta}_j,\alpha_j\right\}}(w)\le x\right]-\Phi\left(\frac{x+\alpha_j-\tilde{\theta}_j(w)}{\sigma_{nj}}\right)\mathcal{I}(x\ge 0)-\right.\\
&\qquad\left.\Phi\left(\frac{x-\alpha_j-\tilde{\theta}_j(w)}{\sigma_{nj}}\right)\mathcal{I}(x<0)\right|\\
=&\lim_{n\to\infty}\left|\Pr\left[\zeta_{\left\{\sigma_{nj}^{-1}\hat{\theta}_j,\sigma_{nj}^{-1}\alpha_j\right\}}(w)\le \sigma_{nj}^{-1}x\right]-\Phi\left\{\frac{x+\alpha_j-\tilde{\theta}_j(w)}{\sigma_{nj}}\right\}\mathcal{I}(x\ge 0)-\right.\\
&\qquad\left.\Phi\left\{\frac{x-\alpha_j-\tilde{\theta}_j(w)}{\sigma_{nj}}\right\}\mathcal{I}(x<0)\right|\\
=&\,0.
\end{aligned}$$

■

Proof of Theorem 2.3.3:

Let $u_{nj}^*=\hat{\theta}_j-\hat{\sigma}_{nj}z_{\xi/2}$ and $v_{nj}^*=\hat{\theta}_j+\hat{\sigma}_{nj}z_{\xi/2}$.

(a). When $P_+>\xi/2$ and $P_->\xi/2$, or $P_-<\xi/2$ and $P_+>1-\xi/2$, or $P_+<\xi/2$ and $P_->1-\xi/2$, then $\zeta_{(u_{nj}^*,\alpha_j)}\neq 0$ and $\zeta_{(v_{nj}^*,\alpha_j)}\neq 0$. Therefore $\hat{\theta}_j-\hat{\sigma}_{nj}z_{\xi/2}\le\tilde{\theta}_j\le\hat{\theta}_j+\hat{\sigma}_{nj}z_{\xi/2}$ is equivalent to $\zeta(\hat{\theta}_j,\alpha_j)-\hat{\sigma}_{nj}z_{\xi/2}\le\zeta_{(\tilde{\theta}_j,\alpha_j)}\le\zeta(\hat{\theta}_j,\alpha_j)+\hat{\sigma}_{nj}z_{\xi/2}$. Therefore,

$$\begin{aligned}
&\lim_{n\to\infty}\Pr\left\{\zeta_{(\hat{\theta}_j,\alpha_j)}-\hat{\sigma}_{nj}z_{\xi/2}\le\zeta_{(\tilde{\theta}_j,\alpha_j)}\le\zeta_{(\hat{\theta}_j,\alpha_j)}+\hat{\sigma}_{nj}z_{\xi/2}\right\}\\
=&\lim_{n\to\infty}\Pr\left(\hat{\theta}_j-\hat{\sigma}_{nj}z_{\xi/2}\le\tilde{\theta}_j\le\hat{\theta}_j+\hat{\sigma}_{nj}z_{\xi/2}\right)\\
=&\lim_{n\to\infty}\Pr\left(\tilde{\theta}_j-\hat{\sigma}_{nj}z_{\xi/2}\le\hat{\sigma}_{nj}^{-1}\hat{\theta}_j\le\tilde{\theta}_j+\hat{\sigma}_{nj}z_{\xi/2}\right)\\
=&1-\xi.
\end{aligned}$$

That is, $\left[\zeta_{(\hat{\theta}_j,\alpha_j)}-\hat{\sigma}_{nj}z_{\xi/2},\zeta_{(\hat{\theta}_j,\alpha_j)}+\hat{\sigma}_{nj}z_{\xi/2}\right]$ is the $1-\xi$ confidence interval for $\zeta_{(\tilde{\theta}_j,\alpha_j)}$.

(b). When $P_+<\xi/2$ and $\xi-P_+<P_-<1-\xi/2$, then $\zeta_{(u_{nj}^*,\alpha_j)}\neq 0$ and $\zeta_{(v_{nj}^*,\alpha_j)}=0$. Let $A=\hat{\sigma}_{nj}^{-1}\alpha_j+\delta_0-\hat{\sigma}_{nj}^{-1}\hat{\theta}_j$ and B satisfy $\Pr(z<-A)+\Pr(z>B)=\xi$, where

$z \sim N(0,1)$ and $\delta_0 > 0$ is small enough such that $\hat{\sigma}_{nj}^{-1}\hat{\theta} - B < -\hat{\sigma}_{nj}^{-1}\alpha_j$. Then, $\lim_{n\to\infty} \Pr\left\{-A \le \hat{\sigma}_{nj}^{-1}(\hat{\theta} - \tilde{\theta}_j) \le B\right\} = 1-\xi$, i.e. $\lim_{n\to\infty} \Pr(\hat{\theta}_j - \hat{\sigma}_{nj}B \le \tilde{\theta}_j \le \hat{\theta}_j + \hat{\sigma}_{nj}A) = 1-\xi$. By the definitions of A and B, we have $\zeta_{(\hat{\theta}_j+\hat{\sigma}_{nj}A,\alpha_j)} > 0$ and $\zeta_{(\hat{\theta}_j-\hat{\sigma}_{nj}B)} < 0$. Therefore, similar to part (a), we have,

$$\begin{aligned}
&\lim_{n\to\infty} \Pr\left\{\zeta_{(\hat{\theta},\alpha_j)} - \hat{\sigma}_{nj}B \le \zeta_{(\tilde{\theta}_j,\alpha_j)} \le \hat{\sigma}_{nj}\delta_0\right\} \\
=&\lim_{n\to\infty} \Pr\left(\hat{\theta}_j - \hat{\sigma}_{nj}B \le \tilde{\theta}_j \le \hat{\theta}_j + \hat{\sigma}_{nj}A\right) \\
=&1-\xi.
\end{aligned}$$

Then, $\left[\zeta_{(\hat{\theta},\alpha_j)} - \hat{\sigma}_{nj}B, \hat{\sigma}_{nj}\delta_0\right]$ is the $1-\xi$ confidence interval for $\zeta_{(\tilde{\theta}_j,\alpha_j)}$, where $B = \Phi^{-1}\left\{1-\xi+\Phi(-\hat{\sigma}_{nj}^{-1}\alpha_j + \hat{\sigma}_{nj}^{-1}\hat{\theta}_j + \delta_0)\right\}$.

(c). When $P_- < \xi/2$ and $\xi - P_- < P_+ < 1-\xi/2$, then $\zeta_{(u_{nj}^*,\alpha_j)} = 0$ and $\zeta_{(v_{nj}^*,\alpha_j)} \neq 0$. Let $B = \hat{\sigma}_{nj}^{-1}\alpha_j + \delta_0 + \hat{\sigma}_{nj}^{-1}\hat{\theta}_j$ and A satisfy $\Pr(z < -A) + \Pr(z > B) = \xi$, where $z \sim N(0,1)$ and $\delta_0(\delta_0 > 0)$ is small enough such that $\hat{\sigma}_{nj}^{-1}\hat{\theta} + A > \hat{\sigma}_{nj}^{-1}\alpha_j$. Similar to part (b), we have

$$\begin{aligned}
&\lim_{n\to\infty} \Pr\left\{-\hat{\sigma}_{nj}\delta_0 \le \zeta_{(\tilde{\theta}_j,\alpha_j)} \le \zeta_{(\hat{\theta},\alpha_j)} + \hat{\sigma}_{nj}A\right\} \\
=&\lim_{n\to\infty} \Pr\left(\hat{\sigma}_{nj}^{-1}\tilde{\theta} - A \le \hat{\sigma}_{nj}^{-1}\hat{\theta}_j \le \hat{\sigma}_{nj}^{-1}\tilde{\theta} + B\right) \\
=&1-\xi.
\end{aligned}$$

Then, $\left[-\hat{\sigma}_{nj}\delta_0, \zeta_{(\hat{\theta},\alpha_j)} + \hat{\sigma}_{nj}A\right]$ is the $1-\xi$ confidence interval for $\zeta_{(\tilde{\theta}_j),\alpha_j}$, where $A = -\Phi^{-1}\left\{\xi-1+\Phi(\hat{\sigma}_{nj}^{-1}\alpha_j + \hat{\sigma}_{nj}^{-1}\hat{\theta}_j + \delta_0)\right\}$.

(d). When $P_+ + P_- < \xi$, then $\zeta_{(u_{nj}^*,\alpha_j)} = 0$ and $\zeta_{(v_{nj}^*,\alpha_j)} = 0$. Therefore, $\hat{\sigma}_{nj}^{-1}\hat{\theta}_j - z_{\xi/2} \le \hat{\sigma}_{nj}^{-1}\tilde{\theta}_j \le \hat{\sigma}_{nj}^{-1}\hat{\theta}_j + z_{\xi/2}$ implies that $0 = \zeta_{(\hat{\sigma}_{nj}^{-1}\hat{\theta}_j - z_{\xi/2},\hat{\sigma}_{nj}^{-1}\alpha_j)} \le \zeta_{(\hat{\sigma}_{nj}^{-1}\tilde{\theta}_j,\hat{\sigma}_{nj}^{-1}\alpha_j)} \le \zeta_{(\hat{\sigma}_{nj}^{-1}\hat{\theta}_j + z_{\xi/2},\hat{\sigma}_{nj}^{-1}\alpha_j)} = 0$. Therefore, $\Pr\left\{\zeta_{(\tilde{\theta}_j,\alpha_j)} = 0\right\} \ge \lim_{n\to\infty} \Pr(\hat{\sigma}_{nj}^{-1}\tilde{\theta}_j - z_{\xi/2} \le \hat{\sigma}_{nj}^{-1}\hat{\theta}_j \le \hat{\sigma}_{nj}^{-1}\tilde{\theta}_j + z_{\xi/2}) = \lim_{n\to\infty} \Pr(\hat{\sigma}_{nj}^{-1}\tilde{\theta}_j - z_{\xi/2} \le \hat{\sigma}_{nj}^{-1}\hat{\theta}_j \le \hat{\sigma}_{nj}^{-1}\tilde{\theta}_j + z_{\xi/2}) = 1-\xi$. Then $[0,0]$ is a confidence interval for $\zeta_{(\tilde{\theta}_j,\alpha_j)}$ with at least $1-\xi$ coverage probability.

As δ_0 in (b) and (c) can be arbitrarily small, the results remain valid when δ_0 goes to 0. Let $\delta_0 \to 0$, then the confidence interval for $\zeta_{(\tilde{\theta}_j,\alpha_j)}$ with at least $1-\xi$ coverage probability is

$$
\begin{aligned}
&[u_{nj}(w), v_{nj}(w)] \\
&= \begin{cases}
\big[\hat{\beta}_j(w) - \hat{\sigma}_{nj} z_{\xi/2}, \hat{\beta}_j(w) + \hat{\sigma}_{nj} z_{\xi/2}\big], & P_+ > \xi/2 \text{ and } P_- > \xi/2, \\
& \text{or } P_- < \xi/2 \text{ and } P_+ > 1-\xi/2, \\
& \text{or } P_+ < \xi/2 \text{ and } P_- > 1-\xi/2 \\
\big[\hat{\beta}_j(w) - \hat{\sigma}_{nj}\hat{B}, 0\big], & P_+ < \xi/2 \text{ and } \xi - P_+ < P_- < 1-\xi/2 \\
\big[0, \hat{\beta}_j(w) + \hat{\sigma}_{nj}\hat{A}\big], & P_- < \xi/2 \text{ and } \xi - P_- < P_+ < 1-\xi/2 \\
[0,0], & P_+ + P_- < \xi
\end{cases}
\end{aligned}
\tag{A.8}
$$

where $\hat{A} = -\Phi^{-1}\Big\{\xi - 1 + \Phi(\hat{\sigma}_{nj}^{-1}\alpha_j + \hat{\sigma}_{nj}^{-1}\hat{\theta}_j)\Big\}$ and $\hat{B} = \Phi^{-1}\Big\{1 - \xi + \Phi(-\hat{\sigma}_{nj}^{-1}\alpha_j + \hat{\sigma}_{nj}^{-1}\hat{\theta}_j)\Big\}$.

Since the bias $\beta_j - \zeta_{(\tilde{\theta}_j,\alpha_j)}$ is asymptotically negligible relative to the variance of $\hat{\theta}_j$, and $\hat{P}_+ \to P_+$ and $\hat{P}_- \to P_-$ as $n \to \infty$, the asymptotic $1-\xi$ confidence interval (A.8) for $\zeta_{(\tilde{\theta}_j,\alpha_j)}$ is also an asymptotic $1-\xi$ confidence interval for β_j with P_+ and P_- replaced by $\hat{P}_+$ and $\hat{P}_-$.

When $\beta_j(w) \neq 0$, the boundary points will not be zero as we defined in (a) and the limiting coverage probability is $1-\epsilon$. When $\beta_j(w) = 0$, since $\hat{\beta}_j(w) \to \beta_j(w)$ as $n \to \infty$. Therefore, there exists $N > 0$ such that when $n > N$, $P_+ < \epsilon/2$ or $P_- < \epsilon/2$ and $P_+ + P_- < 1 - \epsilon/2$ by their definition. Then $u_{nj}(w) = 0$ and (or) $v_{nj}(w) = 0$.

We have

$$\begin{aligned}
\Pr(u_{nj} = 0 \text{ or } v_{nj} = 0) &= \Pr\Big\{\zeta_{(u^*_{nj},\alpha_j)} = \zeta_{(v^*_{nj},\alpha_j)} = 0\Big\} \\
&= \Pr\Big\{|\hat{\theta}_j - \hat{\sigma}_{nj} z_{\xi/2}| \le \alpha_j \text{ or } |\hat{\theta}_j + \hat{\sigma}_{nj} z_{\xi/2}| \le \alpha_j\Big\} \\
&= \Pr\big\{ -\alpha_j + \hat{\sigma}_{nj} z_{\xi/2} \le \hat{\theta}_j \le \alpha_j + \hat{\sigma}_{nj} z_{\xi/2} \text{ or} \\
&\qquad -\alpha_j - \hat{\sigma}_{nj} z_{\xi/2} \le \hat{\theta}_j \le \alpha_j - \hat{\sigma}_{nj} z_{\xi/2}\big\} \\
&\ge \Pr\big\{ -\alpha_j + \hat{\sigma}_{nj} z_{\xi/2} \le \hat{\theta}_j \le \alpha_j + \hat{\sigma}_{nj} z_{\xi/2}\big\} \\
&> 0.
\end{aligned}$$

Therefore, $[u_{nj}, v_{nj}]$ is a sparse confidence interval for β_j. ■

A.1.4 Additional results for preoperative opioid study

In this section, we describe additional results for real data analysis. As opposed to the literature on opioid use [19, 20, 37, 44, 58, 83], our results shed light on how the effects of opioid risk factors are modified by the level of BMI [62]. For example, both worst pain score shows significant impacts on the dose of preoperatively used opioids across the whole range of BMI (Figure 2.5 in main text). The coefficient function of ASA score is significantly positive except for patients with extremely large BMIs; when BMI is 21 and other covariates remain unchanged, the daily dose level of preoperative opioid use will significantly increase 0.13 units comparing patients having severe systemic disease or worse with those having mild systemic disease or healthy patients, indicating worse health condition increases the preoperative opioid use (Figure A.1 below). Depression [44] is positively associated with opioid use regardless of the level of BMI, but the effect is only significant when BMI is smaller than 42; when BMI is 25 and adjusting for other covariates, patients with depression take 0.1 more units of opioids before the surgery (Figure A.1). Illicit drug use [58] has positive effects only for patients with BMI greater than 19·6 (Figure A.1 below).

Both anxiety and alcohol use [37] have positive effects only when BMI is large or small; apnea status [20] is positively associated with preoperative opioid use only for patients whose BMI is within 24·0 to 27·0 (Figure A.2 below).

The results from sub-group analyses are summarized in Table A.1 below.

Table A.1: Estimation results from sub-group analysis

BMI group	< 18		$[18, 30)$		$[30, 49.5)$		≥ 49.5	
	β	p	β	p	β	p	β	p
(Intercept)	-0.293	0.258	0.065	0.014	0.029	0.417	0.064	0.794
Sex	0.305	0.021	0.024	0.016	0.017	0.182	-0.138	0.197
Age	0.010	0.009	-0.001	0.000	-0.001	0.041	0.001	0.711
Race (black)	-0.191	0.636	0.017	0.509	0.014	0.577	-0.431	0.015
Race (Asian)	0.376	0.379	-0.028	0.426	-0.098	0.244	-0.652	0.199
Race (other)	0.388	0.136	-0.012	0.615	-0.017	0.606	-0.013	0.945
Worst pain	0.008	0.757	0.024	0.000	0.020	0.000	0.047	0.034
FM	**0.024**	**0.079**	0.014	0.000	0.014	0.000	-0.002	0.825
Average overall pain	0.070	0.022	0.035	0.000	0.038	0.000	0.042	0.067
Life satisfaction	-0.058	0.018	-0.008	0.000	-0.004	0.146	-0.011	0.571
Depression	-0.106	0.428	0.054	0.000	0.066	0.000	-0.084	0.520
Anxiety	0.172	0.188	-0.016	0.181	-0.007	0.629	0.039	0.728
Comorbidity (>3)	0.075	0.704	0.032	0.017	0.021	0.218	0.143	0.267
Comorbidity (0-3)	-0.051	0.784	-0.014	0.491	0.013	0.533	-0.026	0.834
Alcohol	-0.056	0.632	-0.009	0.357	-0.020	0.099	**0.205**	**0.046**
Apnea	-0.113	0.734	-0.007	0.617	-0.004	0.767	0.065	0.496
Drug	0.056	0.792	**0.075**	**0.002**	**0.084**	**0.010**	0.304	0.206
Tobacco use	0.132	0.312	0.066	0.000	0.053	0.000	-0.097	0.315
ASA > 3	-0.054	0.677	0.101	0.000	0.075	0.000	0.092	0.398

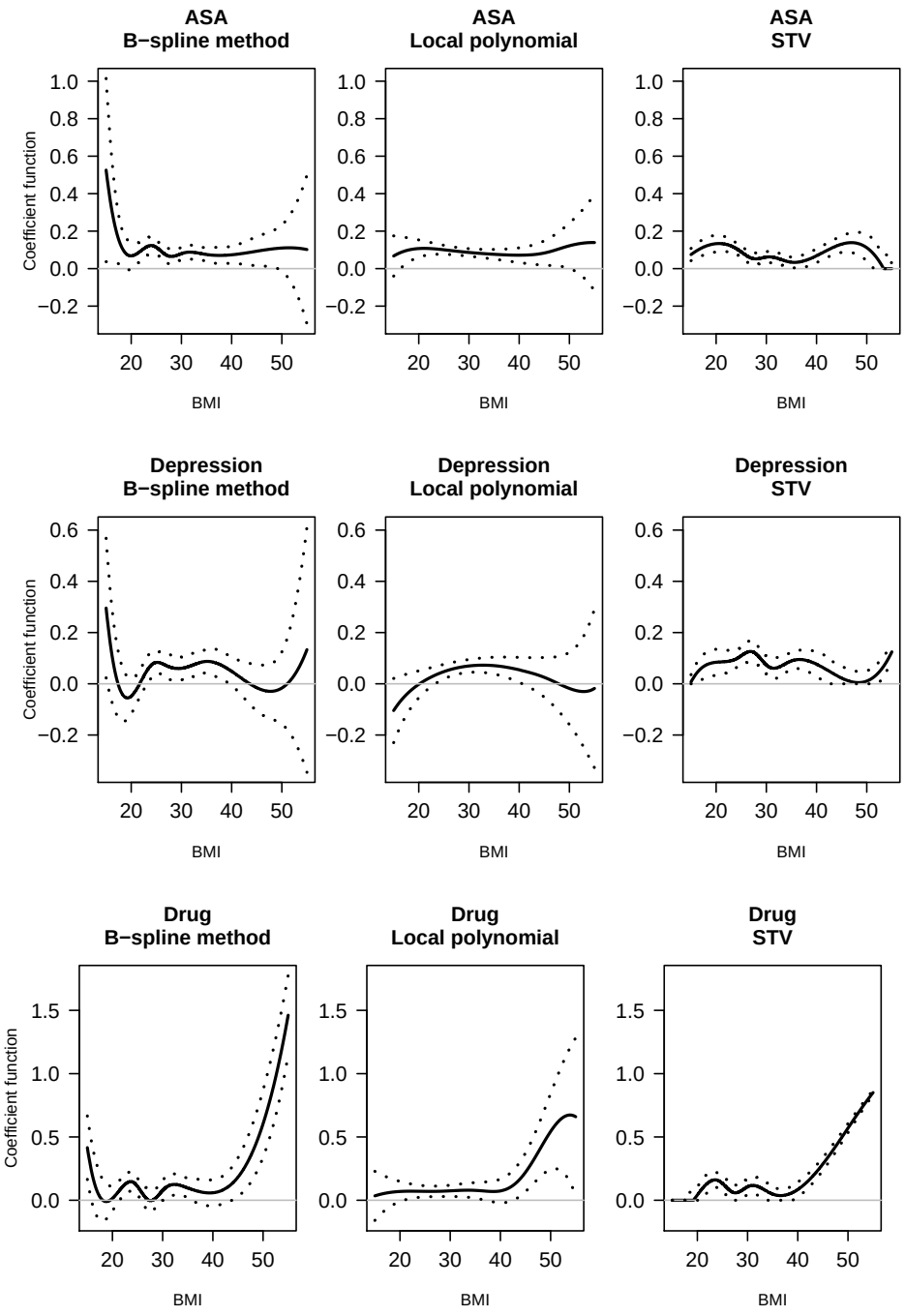


Figure A.1: Estimation results (II) for the preoperative opioid use data using the B-spline method, the local polynomial method and the STV method: the black solid lines are the estimated coefficient function curves for each variable; the dotted lines are the pointwise (sparse) confidence intervals.

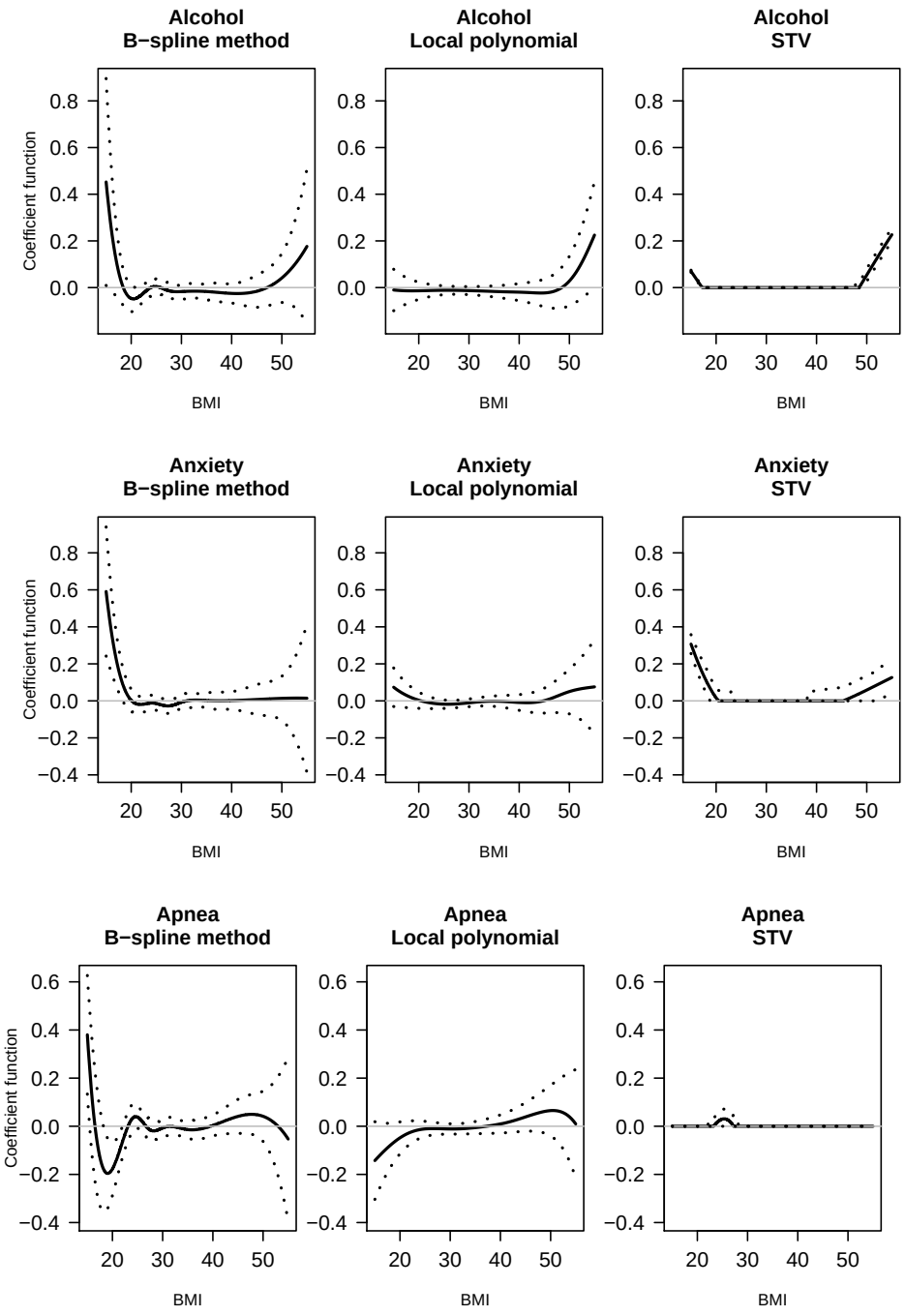


Figure A.2: Estimation results (III) for the preoperative opioid use data using the B-spline method, the local polynomial method and the STV method: the black solid lines are the estimated coefficient function curves for each variable; the dotted lines are the pointwise (sparse) confidence intervals.

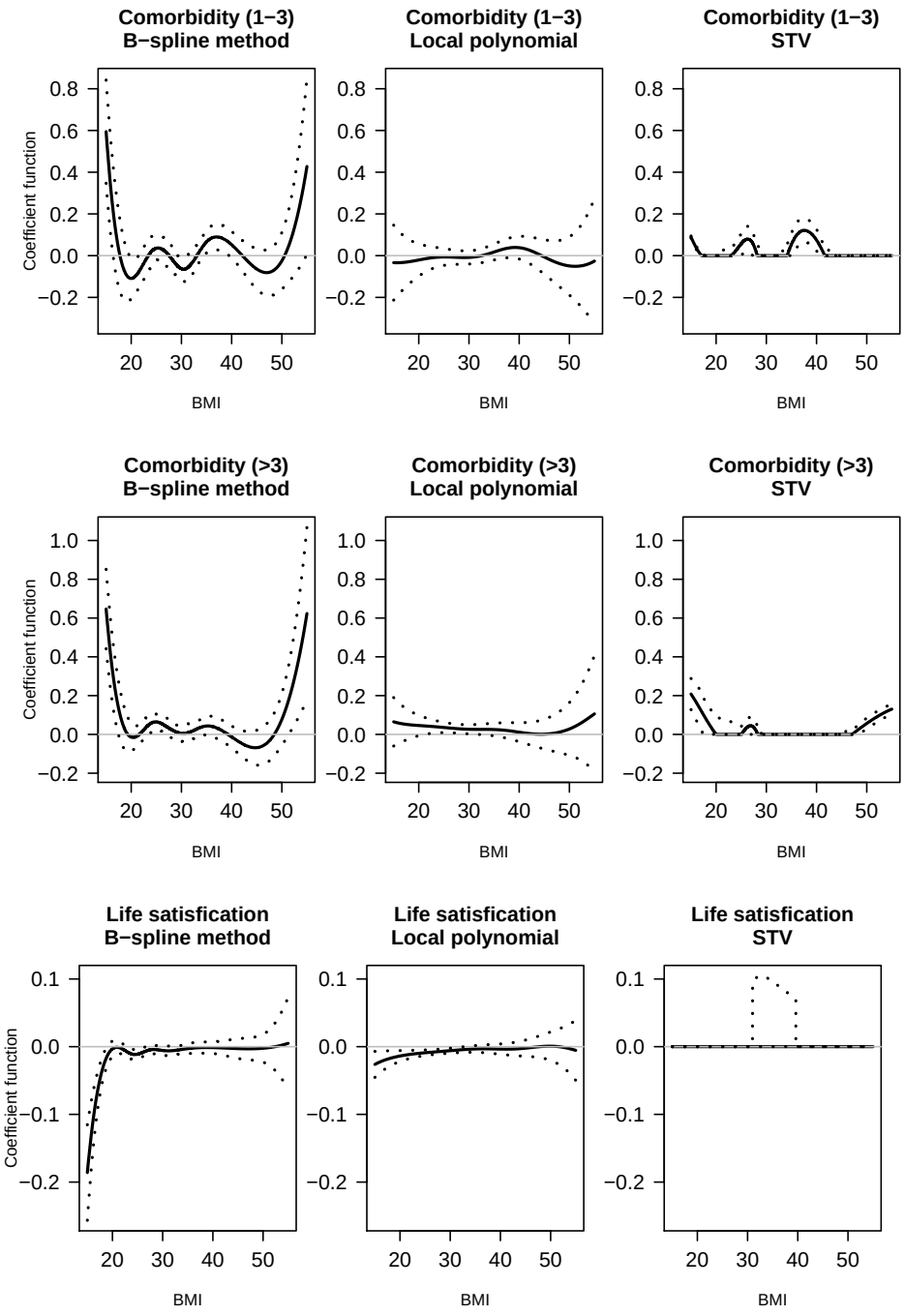


Figure A.3: Estimation results (IV) for the preoperative opioid use data using the B-spline method, the local polynomial method and the STV method: the black solid lines are the estimated coefficient function curves for each variable; the dotted lines are the pointwise (sparse) confidence intervals.

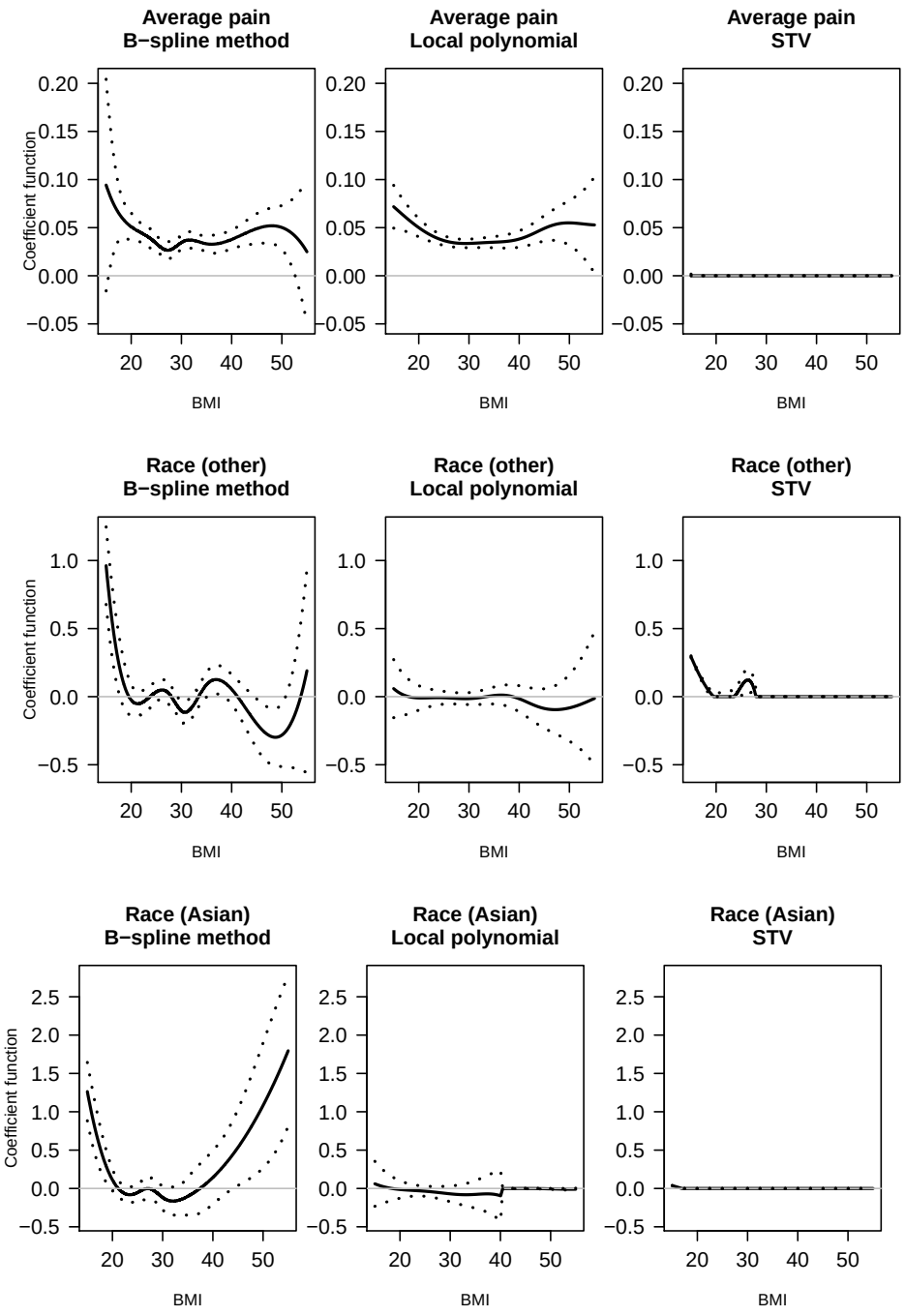


Figure A.4: Estimation results (V) for the preoperative opioid use data using the B-spline method, the local polynomial method and the STV method: the black solid lines are the estimated coefficient function curves for each variable; the dotted lines are the pointwise (sparse) confidence intervals.

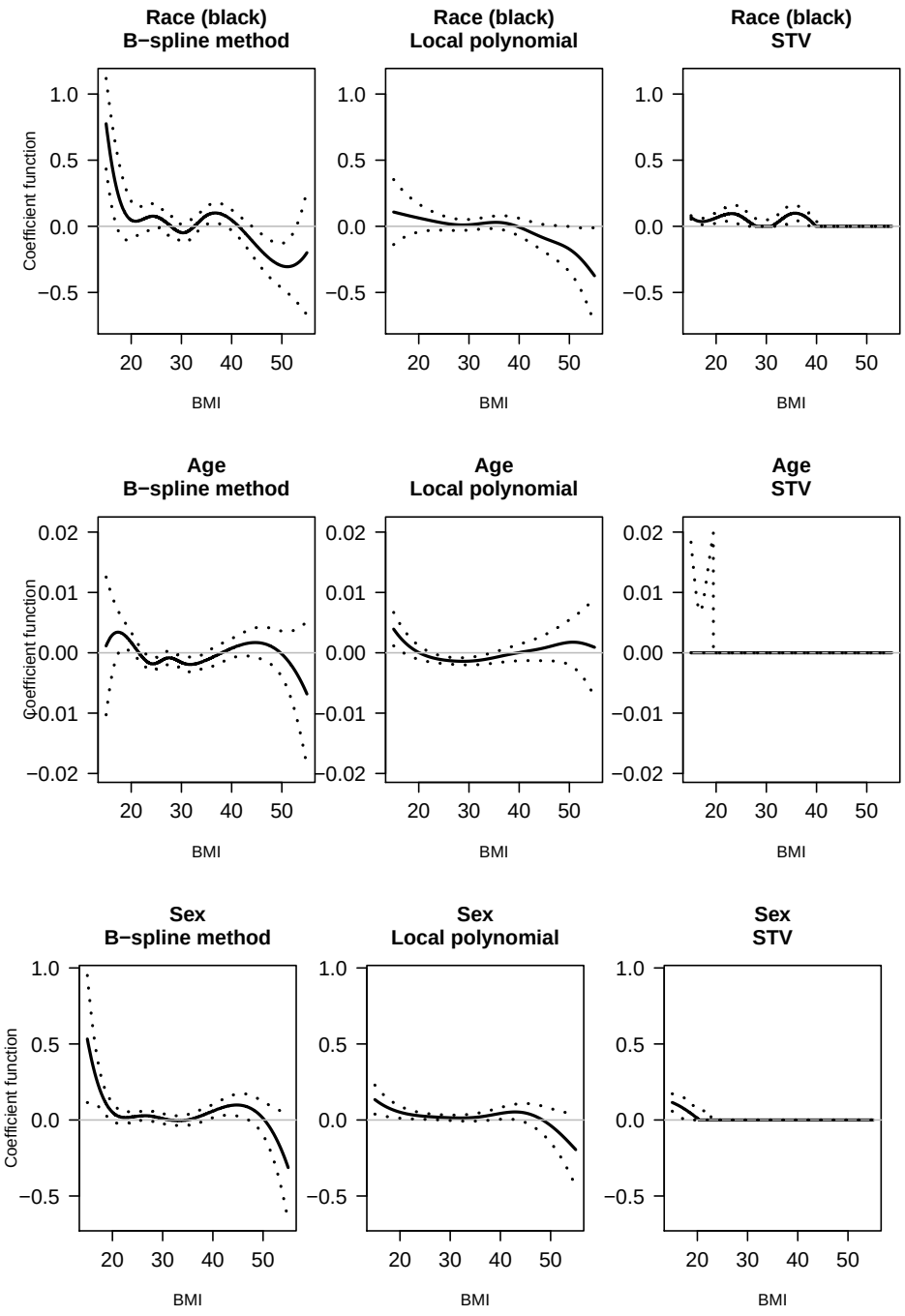


Figure A.5: Estimation results (VI) for the preoperative opioid use data using the B-spline method, the local polynomial method and the STV method: the black solid lines are the estimated coefficient function curves for each variable; the dotted lines are the pointwise (sparse) confidence intervals.

APPENDIX B

Appendices for Chapter III

B.1 Appendix for Generalized Dynamic Effect Change Model: An Interpretable Extension of GAM

We begin with some notation. Let $L_r(P)$ denote the collection of functions $g : \mathcal{X} \mapsto \mathbb{R}$ such that $||g||_{r,P} = \left[\int_{\mathcal{X}} |g(x)|^r dP(x)\right]^{1/r} < \infty$. Let

$$S = \{b\{h(x;\boldsymbol{\theta})\} - b\{h(x;\boldsymbol{\theta}_0)\} : ||\boldsymbol{\theta} - \boldsymbol{\theta}_0|| \leq \delta_n, \boldsymbol{\theta} \in \Theta\}$$

with $\delta_n > 0$. An ϵ-bracket in $L_r(P)$ is a pair of functions $l, u \in L_r p$ with $P\{l(X) \leq f(X) \leq u(X)\}$ and with $||l - r||_{r,P} \leq \epsilon$. The bracketing number $N_{[]}(\epsilon, \mathbb{F}, L_r(P))$ is the minimum number of ϵ-brackets in $L_r(P)$ needed to ensure that every $f \in \mathbb{F}$ lies in at least one bracket. In the following, $d_i > 0$ are constants. Let $f(\xi)$ be a function of random variable ξ, $\mathrm{E} f(\xi) = \int f(\xi) \mathrm{d}\Pr(\xi)$, and $\mathrm{E}_n f(\xi) = \sum_{i=1}^n f(\xi_i)/n$.

Lemma 6 *Under Conditions **Ch3.C3** and **Ch3.C4**, the entropy of $\mathbb{H}_n$ is bounded by $(1/\epsilon)^{2/(2\eta-1)} + q \log(1/\epsilon)$.*

Proof of Lemma 6: Let $S_k = \{cb_k : ||cb_k|| \leq B\}$, where $B < \infty$ is a constant. The bracketing number

$$N_{[]}(\epsilon, S_k, ||\cdot||_\infty) \lesssim \frac{B}{\lambda_k \epsilon} = \frac{k^\eta B}{\epsilon}.$$

Therefore, $N_{[]}(\epsilon, \mathbb{H}_n, ||\cdot||_\infty) \lesssim (1/\epsilon)^q \prod_{k=1}^q Bk^\eta$ by Condition **Ch3.C**4. The entropy $\log N_{[]}(\epsilon, \mathbb{H}_n, ||\cdot||_\infty) \lesssim q \log(1/\epsilon)$. By Lemma 2 in [68], we have $\log N_{[]}(\epsilon, \mathbb{H}_n, ||\cdot||_\infty) \lesssim (1/\epsilon)^{2/(2\eta-1)}$ by Condition **Ch3.C**3. Combining above results, we have that $\log N_{[]}(\epsilon, \mathbb{H}_n, ||\cdot||_\infty) \lesssim \min\{q \log(1/\epsilon), (1/\epsilon)^{2/(2\eta-1)}\} \lesssim q \log(1/\epsilon) + (1/\epsilon)^{2/(2\eta-1)}$.

■

Proof of Theorem 3.3.1:

We let $G_n(\theta) = \mathrm{E}_n l(\theta; \boldsymbol{X}, Y) - \mathrm{E} l(\theta; \boldsymbol{X}, Y) = (\mathrm{E}_n - \mathrm{E})\left[Yh(\theta; \boldsymbol{X}) - b\{h(\theta; \boldsymbol{X})\}\right]$, then

$$\begin{aligned} G_n(\theta) - G_n(\theta_n) &= (\mathrm{E}_n - \mathrm{E})\left\{l(\theta; \boldsymbol{X}, Y) - l(\theta_n; \boldsymbol{X}, Y)\right\} \\ &= (\mathrm{E}_n - \mathrm{E})\left\{Yh(\theta; \boldsymbol{X}) - Yh(\theta_n; \boldsymbol{X})\right\} - (\mathrm{E}_n - \mathrm{E})\left\{b\{h(\theta; \boldsymbol{X})\} - b\{h(\theta_n; \boldsymbol{X})\}\right\} \\ &= A + B, \end{aligned}$$

where θ_n is the projection of θ in $\boldsymbol{\Theta}_n$.

Let $\mathcal{M} = \left\{\int_0^x \beta(t)\mathrm{d}t : \beta \in \mathbb{H}_n\right\}$. We claim that $N_{[]}(\epsilon, \mathcal{M}, ||\cdot||_2) \leq N_{[]}(\epsilon, \mathbb{H}_n, ||\cdot||_2)$. For $\beta_1 \in \mathbb{H}_n$, we assume that $l_1 \leq \beta_1 \leq r_1$ and $||l_1 - r_1|| \leq \epsilon$, then l_1, r_1 is an ϵ-bracket for β_1. By integration, we have $\int_0^x l_1(t)\mathrm{d}t \leq \int_0^x \beta_1(t)\mathrm{d}t \leq \int_0^x r_1(t)\mathrm{d}t$. Since

$$\begin{aligned} \left|\int_0^x l_1(t)\mathrm{d}t - \int_0^x r_1(t)\mathrm{d}t\right| &\leq \int_0^x |l_1(t) - r_1(t)|\mathrm{d}t \\ &= \int_0^1 |l_1(t) - r_1(t)| I(t \leq x)\mathrm{d}t \\ &\leq \sqrt{\int_0^1 |l_1(t) - r_1(t)|^2 \mathrm{d}t} \sqrt{\int_0^1 [I(t \leq x)]\mathrm{d}t} \\ &= \epsilon \cdot x, \end{aligned}$$

we have $|| \int_0^x l_1(t)\mathrm{d}t - \int_0^x r_1(t)\mathrm{d}t||_2 = \sqrt{\int_0^1 | \int_0^x l_1(t)\mathrm{d}t - \int_0^x r_1(t)\mathrm{d}t|^2 \mathrm{d}x} \leq \sqrt{\int_0^1 \epsilon^2 x^2 \mathrm{d}x} = \epsilon/\sqrt{3}$. Therefore, $N_{[]}(\epsilon, \mathcal{M}, ||\cdot||_2) \leq N_{[]}(\epsilon, \mathbb{H}_n, ||\cdot||_2)$. By Lemma 6, $\log N_{[]}(\epsilon, \mathcal{M}, ||\cdot||_\infty) \lesssim (1/\epsilon)^{2/(2\eta-1)} + q\log(1/\epsilon)$.

Let $\mathcal{M}_1 = \left\{h(\theta; \boldsymbol{X}) = \alpha_0 + \sum_{j=1}^p \int_0^{X_{ij}} \{\alpha_j + \beta_j(x)\}\,\mathrm{d}x : \beta_j \in \mathbb{H}_n\right\}$. Since $p = O(1)$, we have $\log N_{[]}(\epsilon, \mathcal{M}_1, ||\cdot||_\infty) \lesssim (1/\epsilon)^{2/(2\eta-1)} + q\log(1/\epsilon)$. The bracketing integral of $\mathcal{M}_1$ is

$$J_{[]}(\delta, \mathcal{M}_1, ||\cdot||_\infty) = \int_0^\delta \sqrt{1 + \log N_{[]}(\epsilon, \mathcal{M}_1, ||\cdot||_\infty)} \lesssim \delta^{(2\eta-2)/(2\eta-1)} + q^{1/2}\delta.$$

By Lemma 3.4.2 in Van Der Vaart and Wellner (1996) [85], we have

$$E^* \sup_{||\theta-\theta_n||<\delta^2, \theta\in\Theta_n} |A| \lesssim J_{[]}(\delta, \mathcal{M}_1, ||\cdot||_\infty) \lesssim \delta^{(2\eta-2)/(2\eta-1)} + q^{1/2}\delta.$$

Since function b is monotone, we also have

$$E^* \sup_{||\theta-\theta_n||<\delta^2, \theta\in\Theta_n} |B| \lesssim J_{[]}(\delta, \mathcal{M}_1, ||\cdot||_\infty) \lesssim \delta^{(2\eta-2)/(2\eta-1)} + q^{1/2}\delta.$$

According to Theorem 3.4.1 of Van Der Vaart and Wellner (1996) [85], we can choose $\phi_n(\delta) = \delta^{(2\eta-2)/(2\eta-1)}$, and then $||\hat{\theta}_n - \theta_n||_\infty = n^{-(2\eta-1)/(4\eta)} + (q/n)^{1/2}$.

Since $||\beta_j - \beta_{nj}||^2 = O(1/q^4)$ [66], we have $||\theta_0 - \theta_n||^2 = O(1/q^4)$ with p fixed. Therefore, $||\hat{\theta} - \theta_0||_\infty \leq ||\hat{\theta}_n - \theta_n||_\infty + ||\theta_n - \theta_0||_\infty = O(n^{-(2\eta-1)/(4\eta)} + (q/n)^{1/2} + 1/q^2)$.

■

Proof of Theorem 3.3.2:

The iterative least squares approach suggests working on the random vector $\boldsymbol{z} = \boldsymbol{U}\boldsymbol{\gamma} + \boldsymbol{\Gamma}(\boldsymbol{Y} - \boldsymbol{\mu})$. Then

$$\hat{\boldsymbol{\gamma}} = \frac{1}{n}(\frac{1}{n}\boldsymbol{U}^T\boldsymbol{W}\boldsymbol{U})^{-1}\boldsymbol{U}^T\boldsymbol{W}\boldsymbol{z}.$$

To prove the theorem, we first prove that for any non-zero $(pq+p+1)$ dimensional constant vector $\boldsymbol{c}$, we have

$$\boldsymbol{c}^T(\hat{\boldsymbol{\gamma}}-\tilde{\boldsymbol{\gamma}})/\mathrm{SD}(\boldsymbol{c}^T\hat{\boldsymbol{\gamma}}) \to_d N(0,1)$$

as $n\to\infty$, where $\mathrm{SD}(\boldsymbol{c}^T\hat{\boldsymbol{\gamma}}) = \{\mathrm{Var}(\boldsymbol{c}^T\boldsymbol{\gamma})\}^{1/2}$.

By calculations, we have

$$\begin{aligned}\boldsymbol{c}^T(\hat{\boldsymbol{\gamma}}-\tilde{\boldsymbol{\gamma}}) =& \frac{1}{n}\boldsymbol{c}^T(\frac{1}{n}\boldsymbol{U}^T\boldsymbol{W}\boldsymbol{U})^{-1}\boldsymbol{U}^T\boldsymbol{W}\boldsymbol{z} - \boldsymbol{c}^T\tilde{\boldsymbol{\gamma}}\\ =& \boldsymbol{c}^T(\boldsymbol{U}^T\boldsymbol{W}\boldsymbol{U})^{-1}\boldsymbol{U}^T\boldsymbol{W}\Gamma(\boldsymbol{Y}-\boldsymbol{\mu}) + \boldsymbol{c}^T\left\{(\frac{1}{n}\boldsymbol{U}^T\boldsymbol{W}\boldsymbol{U})^{-1}\frac{1}{n}\boldsymbol{U}^T\boldsymbol{W}\boldsymbol{U} - \boldsymbol{I}\right\}\tilde{\boldsymbol{\gamma}}\\ =& A_1 + A_2,\end{aligned}$$

where $\boldsymbol{I}$ is the identity matrix of dimension $pq+p+1$.

We rewrite $A_1 = \frac{1}{n}\sum_i^n \boldsymbol{c}^T(\frac{1}{n}\boldsymbol{U}^T\boldsymbol{W}\boldsymbol{U})^{-1}U_iW_{ii}\Gamma_{ii}(Y_i-\mu_i) = \sum_i^n a_i\xi_i$ with $a_i = \frac{1}{n}\boldsymbol{c}^T(\frac{1}{n}\boldsymbol{U}^T\boldsymbol{W}\boldsymbol{U})^{-1}W^{1/2}U_i$ and $\xi_i = (Y_i-\mu_i)/\sqrt{V(\mu_i)}$. Since ξ_i are independent with mean zero and variance one, we only need to verify the Lindeberg condition

$$\frac{\max_i a_i^2}{\sum_i a_i^2} \to_p 0, \qquad \text{as } n\to\infty,$$

to prove $A_1/\sqrt{\sum_i a_i^2}$ is asymptotically $N(0,1)$.

By calculations, we have

$$\begin{aligned}\sum_i^n a_i^2 &= \frac{1}{n^2}\sum_i^n\left\{\boldsymbol{c}^T(\frac{1}{n}\boldsymbol{U}^T\boldsymbol{W}\boldsymbol{U})^{-1}\boldsymbol{W}^{1/2}U_i\right\}^2\\ &= \frac{1}{n}\boldsymbol{c}^T(\frac{1}{n}\boldsymbol{U}^T\boldsymbol{W}\boldsymbol{U})^{-1}\frac{1}{n}\sum_i^n W_{ii}U_iU_i^T(\frac{1}{n}\boldsymbol{U}^T\boldsymbol{W}\boldsymbol{U})^{-1}\boldsymbol{c}\\ &= \frac{1}{n}\boldsymbol{c}^T(\frac{1}{n}\boldsymbol{U}^T\boldsymbol{W}\boldsymbol{U})^{-1}(\frac{1}{n}\boldsymbol{U}^T\boldsymbol{W}\boldsymbol{U})(\frac{1}{n}\boldsymbol{U}^T\boldsymbol{W}\boldsymbol{U})^{-1}\boldsymbol{c}\\ &= O_p(\frac{\boldsymbol{c}^T(\frac{1}{n}\boldsymbol{U}^T\boldsymbol{W}\boldsymbol{U})^{-1}\boldsymbol{c}}{n}),\end{aligned}$$

and

$$\max_i a_i^2 = \frac{1}{n^2}\max_i \left\{ \boldsymbol{c}^T(\frac{1}{n}\boldsymbol{U}^T\boldsymbol{W}\boldsymbol{U})^{-1}W^{1/2}U_i \right\}^2$$
$$\leq \frac{1}{n^2}\boldsymbol{c}^T(\frac{1}{n}\boldsymbol{U}^T\boldsymbol{W}\boldsymbol{U})^{-1}\boldsymbol{c}\cdot\max_i U_i^T(\frac{1}{n}\boldsymbol{U}^T\boldsymbol{W}\boldsymbol{U})^{-1}U_i.$$

Since the eigenvalues of $\boldsymbol{U}^T\boldsymbol{W}\boldsymbol{U}/n$ are bounded above by Condition **Ch3.C**2, we have

$$\frac{\max_i a_i^2}{\sum_i^n a_i^2} \leq o(\frac{1}{n\max_i U_i^T(\frac{1}{n}\boldsymbol{U}^T\boldsymbol{W}\boldsymbol{U})^{-1}U_i}) \to 0,$$

as $n \to \infty$.

Therefore, $A_1/\sqrt{\sum_i a_i^2}$ is asymptotically $N(0,1)$.

$$|A_2| = \left| \boldsymbol{c}^T \left\{ (\frac{1}{n}\boldsymbol{U}^T\boldsymbol{W}\boldsymbol{U})^{-1}\frac{1}{n}\boldsymbol{U}^T\boldsymbol{W}\boldsymbol{U} - \boldsymbol{I} \right\}\boldsymbol{\gamma} \right|$$
$$\leq \sqrt{\boldsymbol{c}^T(\frac{1}{n}\boldsymbol{U}^T\boldsymbol{W}\boldsymbol{U})^{-1}\boldsymbol{c}}\sqrt{\boldsymbol{\gamma}^T(\frac{1}{n}\boldsymbol{U}^T\boldsymbol{W}\boldsymbol{U})^{-1}\boldsymbol{\gamma}}$$

Therefore, $|A_2|/\sqrt{\sum_i a_i^2} = 1/n^{1/2} = o(1)$ by Condition **Ch3.C**2.

By Slusky's Theorem, we have

$$\boldsymbol{c}^T(\hat{\boldsymbol{\gamma}} - \tilde{\boldsymbol{\gamma}})/\mathrm{SD}(\boldsymbol{c}^T\hat{\boldsymbol{\gamma}}) \to_d N(0,1)$$

as $n \to \infty$, where

$$\mathrm{SD}(\boldsymbol{c}^T\hat{\boldsymbol{\gamma}}) = \sqrt{\sum_i a_i^2} = \left\{ \frac{1}{n}\boldsymbol{c}^T(\frac{1}{n}\boldsymbol{U}^T\boldsymbol{W}\boldsymbol{U})^{-1}\boldsymbol{c} \right\}^{1/2}.$$

Since $\boldsymbol{e}_{l,m}$ is the l-dimensional vector with the m-th element taken to be one and zero elsewhere, $\mathbf{0}_{p+1}$ is the $p+1$ dimension vector of 0s, $\boldsymbol{c}_j(x) = (\mathbf{0}_{p+1}, \boldsymbol{e}_{p,j}\otimes\boldsymbol{B}(x))^T$, and

$$\boldsymbol{C} = (\boldsymbol{e}_{pq+p+1,0}, \boldsymbol{e}_{pq+p+1,1}, \ldots, \boldsymbol{e}_{pq+p+1,p}, \boldsymbol{c}_1(x_1), \ldots, \boldsymbol{c}_p(x_p))^T,$$

the estimator of $\boldsymbol{\theta} = (\alpha_0, \ldots, \alpha_p, \beta_1(x_1), \ldots, \beta_p(x_p))^T$ is $\hat{\boldsymbol{\theta}} = \boldsymbol{C}\hat{\boldsymbol{\gamma}}$. Then by Cramer-Wold device Theorem, we have

$$\boldsymbol{\Sigma}^{-1/2}(\hat{\boldsymbol{\theta}}(\boldsymbol{x}) - \boldsymbol{\theta}_0(\boldsymbol{x})) \to_d N(0, \boldsymbol{I}),$$

where $\boldsymbol{\Sigma} = \boldsymbol{C}^T(\boldsymbol{U}^T\boldsymbol{W}\boldsymbol{U})^{-1}\boldsymbol{C}$. The proof is completed. ■

APPENDIX C

Appendices for Chapter IV

C.1 Appendix for Chapter Soft-Thresholding Operator for Modeling Sparse Time-Varying Effects in Survival Analysis

We introduce some notation that are needed in the proofs. Let

$$m_n(t,z,g) = [g(z) - \log S_{0n}(t,g)]\, I[0 \le t \le \tau],$$

$$m_0(t,z,g) = [g(z) - \log S_0(t,g)]\, I[0 \le t \le \tau],$$

$$M_n(g) = P_{\Delta_n} m_n(\cdot, g),$$

$$M_0(g) = P_{\Delta} m_0(\cdot, g),$$

$$\mathbb{G} = \{g(z) : g(z,t) = \sum_{j=1}^{p} z_j \beta_j(t), \beta_j \in \mathbb{H}, 1 \le j \le p, ||g - g_n|| \le \delta\},$$

$$\text{and } E_{n,\delta} = \{m_0(\cdot, g) - m_0(\cdot, g_n), g \in \mathbb{G}, ||g, g_n|| \le \delta\}.$$

Hereafter, $c_i\,(i = 1, \ldots, 4)$ are some constants.

Proof of Theorem 4.3.1:

For every $f \in \mathbb{F}_0$, by Corollary 6.21 of Schumaker (2007) [71] , there exists an $f_n \in \mathbb{F}$, $||f_n - f||_\infty = O(q^{-m})$. For any $\delta_1 > 0$ and $\delta_2 > 0$, $\exists\, \eta$ constructing h, such that $||h(f_n) - \zeta(f_n)|| < \delta_1$ and $|h(f) - \zeta(f)|| < \delta_2$. Then we have,

$$\begin{aligned} ||h(f_n) - h(f)|| &< ||h(f_n) - \zeta(f_n)|| + ||\zeta(f_n) - \zeta(f)|| + ||\zeta(f) - h(f)|| \\ &= A_1 + A_2 + A_3. \end{aligned}$$

Let $\delta_1 = O(q^{-m})$ and $\delta_2 = O(q^{-m})$, then $A_1 < \delta_1 = O(q^{-m})$ and $A_2 < \delta_2 = O(q^{-m})$. We have $A_3 \leq ||f_n - f|| = O(q^{-m})$ because the Lipschitz continuous property in Lemma 1 of Kang et al. (2018) [51]. Therefore, $||h(f_n) - h(f)||_\infty = O(q^{-m})$. For simplicity of notation, let h_{nj} denote $h(f_{nj})$ and h_{0j} denote $h(f_{0j})$.

Let $g_n = \sum_{j=1}^p Z_j h_{nj}$. Then, given $\boldsymbol{Z}$, $|g_n - g_0| = |\sum_{j=1}^p Z_j(h_{nj} - h_{0j})| \leq \sum_{j=1}^p |Z_j||(h_{nj} - h_{0j})| = O_p(q^{-m})$. Thus, we have $||g_n - g_0|| = O_p(q^{-m})$.

By Lemma 5.1 of Huang (1999) [46], $||\hat{g}_n - g_n||_2^2 = o_p(1)$. We then only need to prove

$$\mathrm{E} \sup_{\delta/2 < ||g - g_n|| \leq \delta} |M_n(g) - M_n(g_n) - (M_0(g) - M_0(g_n))| = O_p(n^{-\frac{1}{2}}\delta(q^{\frac{1}{2}} + log^{\frac{1}{2}}(1/\delta))). \tag{C.1}$$

By derivation, we have

$$\begin{aligned} & M_n(g) - M_n(g_n) - \{M_0(g) - M_0(g_n)\} \\ =& P_{\Delta_n} m_n(\cdot, g) - P_{\Delta_n} m_n(\cdot, g_n) - P_\Delta m_0(\cdot, g) + P_\Delta m_0(\cdot, g_n) \\ =& P_{\Delta_n} m_n(\cdot, g) - P_\Delta m_0(\cdot, g) - P_{\Delta_n} m_n(\cdot, g_n) + P_\Delta m_0(\cdot, g_n) \\ =& P_{\Delta_n}\{\log S_{0n}(\cdot, g) - \log S_0(\cdot, g)\} - P_{\Delta_n}\{\log S_0(\cdot, g_n) - \log S_0(\cdot, g_n)\} \\ =& J_{1n} + J_{2n}. \end{aligned}$$

Since by Lemma 1 in Chapter II, for any $\beta \in \mathbb{H}_n$ and any $\alpha > 0$, we can find at least one $f \in \mathbb{F}_n$ such that $\beta = \zeta(f, \alpha)$, then $\log N_{[]}(\epsilon, \mathbb{H}_n, \delta) \leq \log N_{[]}(\epsilon, \mathbb{F}_n, \delta) \lesssim$

$c_1 q \log(\delta/\epsilon)$ by calculation in Shen and Wong (1994) [72]. Therefore, we can also obtain $\log N_{[]}(\epsilon, \mathbb{H}_n, \delta) \lesssim c_2 q \log(\delta/\epsilon)$ according to its construction. Because both exp and log are monotone functions, we have $\log N_{[]}(\epsilon, E_{n,\delta}, \delta) \lesssim c_2 q \log(\delta/\epsilon) + c_3 q \log(\delta/\epsilon) \lesssim c_4 q \log(\delta/\epsilon)$, where $c_4 = \max(c_2, c_3)$.

Therefore, $J_{[]}(\delta, \epsilon_{n,\delta}, \rho) = \int_0^\delta \sqrt{1 + \log N_{[]}(\epsilon, E_{n,\delta}, \rho)} \mathrm{d}\epsilon \lesssim \delta q^{\frac{1}{2}}$. By Lemma 3.4.2 of Van Der Vaart and Wellner (1996) [85], we have

$$\mathrm{E}||J_{1n}|| \lesssim n^{-\frac{1}{2}} q^{\frac{1}{2}} \delta (1 + \frac{q^{\frac{1}{2}}\delta}{\delta^2 \sqrt{n}} c_5) = O(n^{-\frac{1}{2}} q^{\frac{1}{2}} \delta). \tag{C.2}$$

On the other hand, we have

$$\begin{aligned}
\sup_{||g-g_n|| \le \delta} |J_{2n}| &\le 2 \sup_{0 \le t \le \tau, ||g-g_n|| \le \delta} \left| log \frac{S_{0n}(\cdot, g)}{S_{0n}(\cdot, g_n)} - log \frac{S_0(\cdot, g)}{S_0(\cdot, g_n)} \right| \\
&\lesssim \sup_{0 \le t \le \tau, ||g-g_n|| \le \delta} \left| \frac{S_{0n}(\cdot, g)}{S_{0n}(\cdot, g_n)} - \frac{S_0(\cdot, g)}{S_0(\cdot, g_n)} \right| \\
&\lesssim \sup_{0 \le t \le \tau, ||g-g_n|| \le \delta} \left| \frac{S_{0n}(\cdot, g) S_0(\cdot, g_n) - S_{0n}(\cdot, g_n) S_0(\cdot, g)}{S_{0n}(\cdot, g_n) S_0(\cdot, g_n)} \right|.
\end{aligned}$$

Since the denominator is bounded away from 0 with probability approaching to 1, we only need to consider the numerator. By calculation, we have

$$\begin{aligned}
& S_{0n}(\cdot, g) S_0(\cdot, g_n) - S_{0n}(\cdot, g_n) S_0(\cdot, g) \\
=& S_0(t, g_n)\{S_{0n}(t, g) - S_{0n}(t, g_n) - S_0(t, g) + S_0(t, g_n)\} - \\
& \{S_{0n}(t, g_n) - S_0(t, g_n)\}\{S_0(t, g) - S_{0n}(t, g)\} \\
=& I_{1n} - I_{2n}.
\end{aligned}$$

Since $I_{1n} = S_0(t, g_n) Y(t) [\exp(g(z)) - \exp(g_n(z))]$, we consider the class of function $Y(t) \exp(g(z))$. Since exp is monotone and the entropy of the class of indicator function $Y(t) = I[0 \le t \le \tau]$ is $\delta \log^{\frac{1}{2}}(1/\delta)$, we have that the entropy of the class of function $Y(t) exp(g(z))$ is $\delta(q^{\frac{1}{2}} + log^{\frac{1}{2}}(1/\delta))$. By Lemma 3.4.2 of Van Der Vaart and

Wellner (1996) [85], $I_{1n} \lesssim n^{-\frac{1}{2}}\delta(q^{\frac{1}{2}} + \log^{\frac{1}{2}}(1/\delta))$.

By Taylor's expansion and Jensen's inequality, we have

$$\begin{aligned} |S_0(t,g) - S_0(t,g_n)| &\leq E(Y(t)[\exp(g) - \exp(g_n)]) \\ &\leq \mathrm{E}(\exp(g_n)|g - g_n|) \\ &\lesssim (\mathrm{E}(g - g_n)^2)^{\frac{1}{2}} = O_p(\delta). \end{aligned}$$

Since $S_n(t,g_n) - S_0(t,g_n) = O_p(n^{-\frac{1}{2}}q^{\frac{1}{2}})$, we obtain $I_{2n} = O_p(n^{-\frac{1}{2}}q^{\frac{1}{2}}\delta)$.

Therefore, $\sup_{||g-g_n||\leq\delta}|J_{2n}| \lesssim n^{-\frac{1}{2}}\delta(q^{\frac{1}{2}} + \log^{\frac{1}{2}}(1/\delta))$. Thus, we have $M_n(g) - M_n(g_n) - \{M_0(g) - M_0(g_n)\} = O_p(n^{-\frac{1}{2}}\delta(q^{\frac{1}{2}} + \log^{\frac{1}{2}}(1/\delta)))$.

According to Theorem 3.4.1 of Van Der Vaart and Wellner (1996) [85], the key function $\phi(\delta)$ takes the form $\phi_n(\delta) = \delta(q^{\frac{1}{2}} + log^{\frac{1}{2}}(1/\delta))$. Therefore, $||(\hat{g}_n - g_n)||_2 = O_p((q/n)^{\frac{1}{2}})$.

Therefore, we have

$$\begin{aligned} ||\hat{g}_n - g_0||_2^2 &\leq ||\hat{g}_n - g_n||_2^2 + ||g_n - g_0||_2^2 \\ &\leq O_p(q/n) + O_P(q^{-2m}) \\ &\leq O_p(r_n), \end{aligned} \tag{C.3}$$

where $r_n = q/n + q^{-2m}$.

Then by Lemma 1 of Stone (1985) [75], we have

$$E(Z_j\hat{h}_j(t) - Z_jh_j(t))^2 = O_p(r_n), \quad 1 \leq j \leq p. \tag{C.4}$$

By Condition **Ch4.C**3, there exists $\delta, \epsilon > 0$, $\Pr(|Z_j| > \delta) > \epsilon$. Then

$$\begin{aligned} E(Z_j\hat{h}_j(t) - Z_jh_j(t))^2 &> \Pr(|Z_j| > \delta)\delta^2(\hat{h}_j(t) - h_j(t))^2 \\ &> \epsilon\delta^2(\hat{h}_j(t) - h_j(t))^2. \end{aligned} \tag{C.5}$$

Therefore for any t, we have $(\hat{\beta}_j(t)-\beta_j(t))^2 = O_p(r_n)$, i.e. $|\hat{\beta}_j(t)-\beta_j(t)| = O_p(r_n^{1/2})$. Then we have $||\hat{\beta}_j - \beta_j||_\infty = O_p(r_n^{1/2})$ for $j = 1, \ldots, p$. ■

Proof of Theorem 4.3.2:

We show Theorem 4.3.2 is true when $\tau = 1$. The extension to any $\tau < \infty$ is straightforward so it is omitted here.

Following the counting process notation in Anderson and Gill (1982) [4], we let

$$C(\boldsymbol{\gamma}, t) = \sum_{i=1}^{n} \int_0^t \sum_{j=1}^{p} Z_{ij} h_j(\boldsymbol{\gamma}_j, s) \mathrm{d}N_i(s) - \int_0^t \log \left\{ \sum_{i=1}^{n} Y_i(s) \exp\{ \sum_{j=1}^{p} Z_{ij}(s) h_j(\boldsymbol{\gamma}_j, s) \} \right\} \mathrm{d}\bar{N}(s),$$

then we have,

$$\mathrm{PL}(\boldsymbol{\gamma}) = C(\boldsymbol{\gamma}, 1) - \rho ||\boldsymbol{\theta}||_2^2.$$

Then for any $\boldsymbol{\gamma}$,

$$\mathrm{PL}^{'}(\boldsymbol{\gamma}) = C^{'}(\boldsymbol{\gamma}, 1) - \rho \sum_{i=1}^{n} \boldsymbol{\theta} \otimes \boldsymbol{B}(T_i).$$

By Taylor's expansion, we have that

$$\{\mathrm{PL}\}^{'}(\hat{\boldsymbol{\gamma}}) - \mathrm{PL}^{'}(\tilde{\boldsymbol{\gamma}}) = \{\mathrm{PL}\}^{''}(\boldsymbol{\gamma}^*)(\boldsymbol{\gamma} - \tilde{\boldsymbol{\gamma}}),$$

where $\boldsymbol{\gamma}^*$ is on the line segment between $\hat{\boldsymbol{\gamma}}$ and $\tilde{\boldsymbol{\gamma}}$. Since $\{\mathrm{PL}\}^{'}(\hat{\boldsymbol{\gamma}}) = 0$, we have

$$\begin{aligned}
\boldsymbol{\gamma} - \tilde{\boldsymbol{\gamma}} &= - \left[\{\mathrm{PL}\}^{''}(\boldsymbol{\gamma}^*) \right]^{-1} \mathrm{PL}^{'}(\tilde{\boldsymbol{\gamma}}) \\
&= - \left[\{\mathrm{PL}\}^{''}(\boldsymbol{\gamma}^*) \right]^{-1} \left\{ C^{'}(\tilde{\boldsymbol{\gamma}}, 1) - \rho \sum_{i=1}^{n} \boldsymbol{\theta}_0 \otimes \boldsymbol{B}(T_i) \right\} \\
&= - \left[\{\mathrm{PL}\}^{''}(\boldsymbol{\gamma}^*) \right]^{-1} C^{'}(\tilde{\boldsymbol{\gamma}}, 1) + \rho \left[\{\mathrm{PL}\}^{''}(\boldsymbol{\gamma}^*) \right]^{-1} \sum_{i=1}^{n} \boldsymbol{\theta} \otimes \boldsymbol{B}(T_i).
\end{aligned}$$

The goal is to prove that for any non-zero $\boldsymbol{a}$,

$$\frac{\boldsymbol{a}^T(\hat{\boldsymbol{\gamma}} - \tilde{\boldsymbol{\gamma}})}{\hat{\sigma}(\boldsymbol{a})} \to_d N(0,1),$$

where $\hat{\sigma}(\boldsymbol{a}) = n\boldsymbol{a}^T \left[\{\mathrm{PL}\}''(\tilde{\boldsymbol{\gamma}})\right]^{-1} \Sigma(\tilde{\boldsymbol{\gamma}}, 1) \left[\{\mathrm{PL}\}''(\tilde{\boldsymbol{\gamma}})\right]^{-1} \boldsymbol{a}$.

We claim that

$$\frac{\boldsymbol{a}^T \left[-\{\mathrm{PL}\}''(\boldsymbol{\gamma}^*)\right]^{-1} C'(\tilde{\boldsymbol{\gamma}}, 1)}{\hat{\sigma}(\boldsymbol{a})} \to_d N(0,1) \tag{C.6}$$

and

$$\rho \boldsymbol{a}^T \left[\{\mathrm{PL}\}''(\boldsymbol{\gamma}^*)\right]^{-1} \sum_{i=1}^{n} \boldsymbol{\theta} \otimes \boldsymbol{B}(T_i)/\hat{\sigma}(\boldsymbol{a}) \to_p 0. \tag{C.7}$$

To show (C.6), we will utilize the martingale theories in Anderson and Gill (1982) [4] to prove that $\boldsymbol{a}^T \left[-\{\mathrm{PL}\}''(\boldsymbol{\gamma}^*)\right]^{-1} C'(\tilde{\boldsymbol{\gamma}}, t)/\hat{\sigma}(\boldsymbol{a})$ is converging to a Gaussian process. By calculation, we have

$$C'(\tilde{\boldsymbol{\gamma}}, t) = \sum_{i=1}^{n} \int_0^t \{A_i(\tilde{\boldsymbol{\gamma}}, s) - E(\tilde{\boldsymbol{\gamma}}, s)\}\, dM_i(s),$$

where $A_i(\tilde{\boldsymbol{\gamma}}, s) = \boldsymbol{U}_i \otimes \boldsymbol{B}_i$ and $E(\tilde{\boldsymbol{\gamma}}, s) = S_1(\tilde{\boldsymbol{\gamma}}, s)/S_0(\tilde{\boldsymbol{\gamma}}, s)$.

Then we have

$$\frac{\boldsymbol{a}^T \left[-\{\mathrm{PL}\}''(\boldsymbol{\gamma}^*)\right]^{-1}}{\hat{\sigma}(\boldsymbol{a})} C'(\tilde{\boldsymbol{\gamma}}, t) = \sum_{i=1}^{n} \int_0^t \frac{\boldsymbol{a}^T \left[-\{\mathrm{PL}\}''(\boldsymbol{\gamma}^*)\right]^{-1}}{\hat{\sigma}(\boldsymbol{a})} \{A_i(\tilde{\boldsymbol{\gamma}}, s) - E(\tilde{\boldsymbol{\gamma}}, s)\}\, dM_i(s).$$

Let

$$H_i(s) = \frac{\boldsymbol{a}^T \left[-\{\mathrm{PL}\}''(\boldsymbol{\gamma}^*)\right]^{-1}}{\hat{\sigma}(\boldsymbol{a})} \{A_i(\tilde{\boldsymbol{\gamma}}, s) - E(\tilde{\boldsymbol{\gamma}}, s)\},$$

we then can show claim C.6 is true by applying Theorem I.2 in Anderson and Gill (1982) [4]. Condition (I.3) of Theorem I.2 is valid because by Condition **Ch4.C**2,

Ch4.C7 and **Ch4.C**8, we have

$$\begin{aligned}\int_0^t \sum_{i=1}^n H_i^2(s)\lambda_i(s)\mathrm{d}s =& \boldsymbol{a}^T \left[-\{\mathrm{PL}\}''(\boldsymbol{\gamma}^*)\right]^{-1} \cdot \\ & \int_0^t \sum_{i=1}^n \{A_i(\tilde{\boldsymbol{\gamma}},s) - E(\tilde{\boldsymbol{\gamma}},s)\}\{A_i(\tilde{\boldsymbol{\gamma}},s) - E(\tilde{\boldsymbol{\gamma}},s)\}^T \lambda_i(s)\mathrm{d}s \cdot \\ & \left[-\{\mathrm{PL}\}''(\boldsymbol{\gamma}^*)\right]^{-1} \boldsymbol{a}/\hat{\sigma}^2(\boldsymbol{a}) \\ \to_p & \, r(t),\end{aligned}$$

where $r(t)$ is some positive function of t and $r(1) = 1$.

By similar arguments in Anderson and Gill (1982), condition (I.4) of Theorem I.2 is true by Condition **Ch4.C**2, **Ch4.C**7, and **Ch4.C**9. Then claim (C.6) is valid.

Claim (C.7) is valid because

$$\rho \left| \boldsymbol{a}^T \left[\{\mathrm{PL}\}''(\boldsymbol{\gamma}^*)\right]^{-1} \sum_{i=1}^n \boldsymbol{\theta} \otimes \boldsymbol{B}(T_i)/\hat{\sigma}(\boldsymbol{a}) \right| \leq O_p(n\rho) \to_p 0. \tag{C.8}$$

by Condition **Ch4.C**6

Therefore, for any non-zero $\boldsymbol{a}$,

$$\frac{\boldsymbol{a}^T(\hat{\boldsymbol{\gamma}} - \tilde{\boldsymbol{\gamma}})}{\hat{\sigma}(\boldsymbol{a})} \to_d N(0,1),$$

where $\hat{\sigma}(\boldsymbol{a}) = n\boldsymbol{a}^T \left[\{\mathrm{PL}\}''(\tilde{\boldsymbol{\gamma}})\right]^{-1} \Sigma(\tilde{\boldsymbol{\gamma}},1) \left[\{\mathrm{PL}\}''(\tilde{\boldsymbol{\gamma}})\right]^{-1} \boldsymbol{a}$.

Since for any $t \in [0,\tau]$, $\hat{\theta}_j(t) = (\boldsymbol{e}_j \otimes \boldsymbol{B}(t))^T \hat{\boldsymbol{\gamma}}$, then let $\boldsymbol{a} = \boldsymbol{e}_j \otimes \boldsymbol{B}(t)$, we have for any $t \in [0,\tau]$,

$$\frac{\hat{\theta}_j(t) - \theta_j(t)}{\sigma_{nj}(t)} \to_d N(0,1),$$

where $\sigma_{nj}^2(t) = n\{\boldsymbol{e}_j \otimes \boldsymbol{B}(t)\}^T \left[-\{\mathrm{PL}\}''(\boldsymbol{\gamma}^*)\right]^{-1} \Sigma(\tilde{\boldsymbol{\gamma}},1) \left[-\{\mathrm{PL}\}''(\boldsymbol{\gamma}^*)\right]^{-1} \{\boldsymbol{e}_j \otimes \boldsymbol{B}(t)\}$. The proof is completed. ■

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Ahmad, I., Leelahanon, S., and Li, Q. I. (2005), "Efficient Estimation of a Semiparametric Partially Linear Varying Coefficient Model," *Annals of Statistics*, 33, 258–283.

[2] Aires, L., Silva, P., Santos, R., Santos, P., Ribeiro, J., Mota, J., et al. (2008), "Association of Physical Fitness and Body Mass Index in Youth," *Minerva Pediatrica*, 60, 397–406.

[3] Anderson, J. A. and Senthilselvan, A. (1982), "A Two-Step Regression Model for Hazard Functions," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 31, 44–51.

[4] — (1982), "A Two-Step Regression Model for Hazard Functions," *Applied Statistics*, 31, 44–51.

[5] Barry, D. and Petry, N. M. (2009), "Associations Between Body Mass Index and Substance Use Disorders Differ by Gender: Results From the National Epidemiologic Survey on Alcohol and Related Conditions," *Addictive Behaviors*, 34, 51–60.

[6] Bartels, K., Fernandez-Bustamante, A., McWilliams, S. K., Hopfer, C. J., and Mikulich-Gilbertson, S. K. (2018), "Long-Term Opioid Use After Inpatient Surgery–a Retrospective Cohort Study," *Drug and Alcohol Dependence*, 187, 61–65.

[7] Benjamini, Y. and Hochberg, Y. (1995), "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 57, 289–300.

[8] Berlinet, A. and Thomas-Agnan, C. (2011), *Reproducing Kernel Hilbert Spaces in Probability and Statistics*, Berlin: Springer Science & Business Media.

[9] Bickel, P., Zhang, P., and Zhang, P. (1992), "Variable selection in nonparametric regression with categorical covariates," *Journal of the American Statistical Association*, 87, 90–97.

[10] Breheny, P. and Zeng, Y. (2019), *grpreg: Regularization Paths for Regression Models with Grouped Covariates*, r package version 3.2-1.

[11] Cai, Z. and Sun, Y. (2003), "Local Linear Estimation for Time-Dependent Coefficients in Cox's Regression Models," *Scandinavian Journal of Statistics*, 30, 93–111.

[12] CDCP (2007), "Unintentional Poisoning Deaths–United States, 1999-2004," *Morbidity and Mortality Weekly Report*, 56, 93–96.

[13] Chang, S. G., Member, S., Yu, B., Member, S., and Vetterli, M. (2000), "Adaptive Wavelet Thresholding for Image Denoising and Compression," 9, 1532–1546.

[14] Chapman, C. R., Davis, J., Donaldson, G. W., Naylor, J., and Winchester, D. (2011), "Postoperative Pain Trajectories in Chronic Pain Patients Undergoing Surgery: The Effects of Chronic Opioid Pharmacotherapy on Acute Pain," *The Journal of Pain*, 12, 1240–1246.

[15] Cheng, M.-Y., Honda, T., Li, J., Peng, H., et al. (2014), "Nonparametric Independence Screening and Structure Identification for Ultra-High Dimensional Longitudinal Data," *The Annals of Statistics*, 42, 1819–1849.

[16] Cheng, M.-Y., Honda, T., and Zhang, J.-T. (2016), "Forward Variable Selection for Sparse Ultra-High Dimensional Varying Coefficient Models," *Journal of the American Statistical Association*, 111, 1209–1221.

[17] Chiang, C.-T., Rice, J. a., and Wu, C. O. (2001), "Smoothing Spline Estimation for Varying Coefficient Models With Repeatedly Measured Dependent Variables," *Journal of the American Statistical Association*, 96, 605–619.

[18] Christiani, D. C. (2017), "The Boston lung cancer survival cohort," Tech. rep.

[19] Clarke, H., Soneji, N., Ko, D. T., Yun, L., and Wijeysundera, D. N. (2014), "Rates and Risk Factors for Prolonged Opioid Use After Major Surgery: Population Based Cohort Study," *Bmj*, 348, g1251.

[20] Correa, D., Farney, R. J., Chung, F., Prasad, A., Lam, D., and Wong, J. (2015), "Chronic Opioid Use and Central Sleep Apnea: A Review of the Prevalence, Mechanisms, and Preoperative Considerations," *Anesthesia & Analgesia*, 120, 1273–1285.

[21] Cox, D. R. (1992), "Regression Models and Life-Tables," 34, 527–541.

[22] Cron, D. C., Englesbe, M. J., Bolton, C. J., Joseph, M. T., Carrier, K. L., Moser, S. E., Waljee, J. F., Hilliard, P. E., Kheterpal, S., and Brummett, C. M. (2017), "Preoperative Opioid Use Is Independently Associated With Increased Costs and Worse Outcomes After Major Abdominal Surgery," *Annals of Surgery*, 265, 695–701.

[23] D. Schoenfeld (1982), "Partial Residuals for The Proportionnal Hazards Regression Model," *Biometrika*, 69, 239–241.

[24] Donoho, D. L. (1995), "De-Noising by Soft-Thresholding," *IEEE Transactions on Information Theory*, 41, 613–627.

[25] Donoho, D. L. and Johnstone, J. M. (1994), "Ideal Spatial Adaptation by Wavelet Shrinkage," *Biometrika*, 81, 425–455.

[26] Eilers, P. H. and Marx, B. D. (2010), "Splines, Knots, and Penalties," *Wiley Interdisciplinary Reviews: Computational Statistics*, 2, 637–653.

[27] Eilers, P. H. C. and Marx, B. D. (1996), "Flexible Smoothing With B-Splines and Penalties," *Statistical Science*, 11, 89–102.

[28] Eubank, R. L., Huang, C., Maldonado, Y. M., Wang, N., Wang, S., and Buchanan, R. J. (2004), "Smoothing Spline Estimation in Varying-Coefficient Models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66, 653–667.

[29] Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. (2013), *Regression: Models, Methods and Applications*, Springer Science & Business Media.

[30] Fan, J. and Huang, T. (2005), "Profile Likelihood Inferences on Semiparametric Varying-coefficient Partially Linear Models," *Bernoulli*, 11, 1031–1057.

[31] Fan, J., Ma, Y., and Dai, W. (2014), "Nonparametric Independence Screening in Sparse Ultra-High-Dimensional Varying Coefficient Models," *Journal of the American Statistical Association*, 109, 1270–1284.

[32] Fan, J. and Zhang, J.-T. (2000), "Two-Step Estimation of Functional Linear Models With Applications to Longitudinal Data," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62, 303–322.

[33] Fan, J. and Zhang, W. (1999), "Statistical Estimation in Varying Coefficient Models," *The Annals of Statistics*, 27, 1491–1518.

[34] — (2008), "Statistical Methods with Varying Coefficient Models." *Statistics and its interface*, 1, 179–195.

[35] Goesling, J., Moser, S. E., Zaidi, B., Hassett, A. L., Hilliard, P., Hallstrom, B., Clauw, D. J., and Brummett, C. M. (2016), "Trends and Predictors of Opioid Use Following Total Knee and Total Hip Arthroplasty," *PAIN*, 157, 1259–1265.

[36] Gore, S. M., Pocock, S. J., and Kerr, G. R. (1984), "Regression Models and Non-Proportional Hazards in the Analysis of Breast Cancer Survival," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 33, 176–195.

[37] Grant, B. F., Stinson, F. S., Dawson, D. A., Chou, S. P., Dufour, M. C., Compton, W., Pickering, R. P., and Kaplan, K. (2004), "Prevalence and Co-occurrence of Substance use Disorders and Independentmood and Anxiety disorders: Results from the National Epidemiologic Survey on Alcohol and Relatedconditions," *Archives of General Psychiatry*, 61, 807–816.

[38] Hastie, T. and Tibshirani, R. (1987), "Generalized Additive Models: Some Applications," *Journal of the American Statistical Association*, 82, 371–386.

[39] — (1993), "Varying-Coefficient Models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 55, 757–796.

[40] Hastie, T. J. (2017), "Generalized Additive Models," in *Statistical Models in S*, pp. 249–307.

[41] He, K., Lian, H., Ma, S., and Huang, J. Z. (2018), "Dimensionality Reduction and Variable Selection in Multivariate Varying-Coefficient Models With a Large Number of Covariates," *Journal of the American Statistical Association*, 113, 746–754.

[42] He, K., Yang, Y., yan, L., Zhu, J., and Li, Y. (2017), "Modeling Time-varying Effects with Large-scale Survival Data : An Efficient Quasi-Newton Approach," *Journal of Computational and Graphical Statistics*, 26.

[43] Hilliard, P. E., Waljee, J., Moser, S., Metz, L., Mathis, M., Goesling, J., Cron, D., Clauw, D. J., Englesbe, M., Abecasis, G., and Brummett, C. M. (2018), "Prevalence of Preoperative Opioid Use and Characteristics Associated With Opioid Use Among Patients Presenting for Surgery," *JAMA Surgery*, 153, 929–937.

[44] Hooten, W. M., Shi, Y., Gazelka, H. M., and Warner, D. O. (2011), "The Effects of Depression and Smoking on Pain Severity and Opioid Use in Patients with Chronic Pain," *PAIN®*, 152, 223–229.

[45] Hoover, D. R., Rice, J. A., Wu, C. O., and Yang, L.-P. (1998), "Nonparametric Smoothing Estimates of Time-Varying Coefficient Models With Longitudinal Data," *Biometrika*, 85, 809–822.

[46] Huang, J. (1999), "Efficient estimation of the partly linear additive Cox model," *Annals of Statistics*, 27, 1536–1563.

[47] Huang, J., Breheny, P., and Ma, S. (2012), "A Selective Review of Group Selection in High-Dimensional Models," *Statistical Science*, 27.

[48] Huang, J. Z. (2003), "Local Asymptotics for Polynomial Spline Regression," *Annals of Statistics*, 31, 1600–1635.

[49] Huang, J. Z., Wu, C. O., and Zhou, L. (2002), "Varying-Coefficient Models and Basis Function Approximations for the Analysis of Repeated Measurements," *Biometrika*, 89, 111–128.

[50] — (2004), "Polynomial Spline Estimation and Inference for Varying Coefficient Models With Longitudinal Data," *Statistica Sinica*, 14, 763–788.

[51] Kang, J., Reich, B. J., and Staicu, A.-M. (2018), "Scalar-on-Image Regression via the Soft-Thresholded Gaussian Process," *Biometrika*, 105, 165–184.

[52] Lee, E. R., Mammen, E., et al. (2016), "Local Linear Smoothing for Sparse High Dimensional Varying Coefficient Models," *Electronic Journal of Statistics*, 10, 855–894.

[53] Lee, J. D., Sun, D. L., Sun, Y., Taylor, J. E., et al. (2016), "Exact Post-Selection Inference, With Application to the Lasso," *The Annals of Statistics*, 44, 907–927.

[54] Lee, Y. K., Mammen, E., and Park, B. U. (2012), "Flexible Generalized Varying Coefficient Regression Models," *Annals of Statistics*, 40, 1906–1933.

[55] Li, D., Ke, Y., and Zhang, W. (2015), "Model Selection and Structure Specification in Ultra-High Dimensional Generalised Semi-Varying Coefficient Models," *The Annals of Statistics*, 43, 2676–2705.

[56] Lian, H., Lai, P., and Liang, H. (2013), "Partially linear structure selection in cox models with varying coefficients," *Biometrics*, 69, 348–357.

[57] Liu, J., Li, R., and Wu, R. (2014), "Feature Selection for Varying Coefficient Models With Ultrahigh-Dimensional Covariates," *Journal of the American Statistical Association*, 109, 266–274.

[58] Manchikanti, L., Damron, K., McManus, C., and Barnhill, R. (2004), "Patterns of Illicit Drug Use and Opioid Abuse in Patients with Chronic Pain at Initial Evaluation: A Prospective, Observational Study." *Pain Physician*, 7, 431–437.

[59] Martinussen, T., Scheike, T. H., and Skovgaard, I. M. (2002), "Efficient estimation of fixed and time-varying covariate effects in multiplicative intensity models," *Scandinavian Journal of Statistics*, 29, 57–74.

[60] Marzec, L. (1997), "On fitting Cox's regression model with time-dependent coefficients," *Biometrika*, 84, 901–908.

[61] Murphy, S. (1993), "Testing for a time dependent coefficient in Cox's regression model," *Scandinavian Journal of Statistics*, 20, 35–50.

[62] Nafiu, O. O., Shanks, A., Abdo, S., Taylor, E., and Tremper, T. T. (2013), "Association of High Body Mass Index in Children With Early Post-Tonsillectomy Pain," *International Journal of Pediatric Otorhinolaryngology*, 77, 256–261.

[63] Pivec, R., Issa, K., Naziri, Q., Kapadia, B. H., Bonutti, P. M., and Mont, M. A. (2014), "Opioid Use Prior to Total Hip Arthroplasty Leads to Worse Clinical Outcomes," *International Orthopaedics*, 38, 1159–1165.

[64] Ramsay, J. O. and Dalzell, C. J. (1991), "Some Tools for Functional Data Analysis," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 53, 539–561.

[65] Ramsay, J. O. and Silverman, B. W. (2007), *Applied Functional Data Analysis: Methods and Case Studies*, New York: Springer.

[66] Ravikumar, P., Lafferty, J., Liu, H., and Wasserman, L. (2009), "Sparse additive models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71, 1009–1030.

[67] Ruppert, D., Wand, M. P., and Carroll, R. J. (2003), *Semiparametric Regression*, no. 12, Cambridge university press.

[68] Sancetta, A. (2016), "Inference for Additive Models in the Presence of Possibly Infinite Dimensional Nuisance Parameters," *arXiv preprint arXiv:1611.02199*.

[69] Sasieni, P. (1992), "Non-orthogonal projections and their application to calculating the information in a partly linear Cox model," *Scandinavian Journal of Statistics*, 215–233.

[70] Schug, S. A. and Raymann, A. (2011), "Postoperative Pain Management of the Obese Patient," *Best Practice and Research Clinical Anaesthesiology*, 25, 73–81.

[71] Schumaker, L. (2007), *Spline Functions: Basic Theory*, Cambridge: Cambridge University Press.

[72] Shen, X. and Wong, W. H. (1994), "Convergence Rate of Sieve Estimates," *The Annals of Statistics*, 22, 580–615.

[73] Silverman, B. W. (1985), "Some Aspects of the Spline Smoothing Approach to Non-Parametric Regression Curve Fitting," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 47, 1–52.

[74] Song, R., Yi, F., and Zou, H. (2014), "On Varying-Coefficient Independence Screening for High-Dimensional Varying-Coefficient Models," *Statistica Sinica*, 24, 1735.

[75] Stone, C. J. (1985), "Additive Regression and Other Nonparametric Models," *The Annals of Statistics*, 3, 689–705.

[76] — (1986), "The Dimensionality Reduction Principle for Generalized Additive Models," *The Annals of Statistics*, 14, 590–606.

[77] Sun, E. C., Darnall, B. D., Baker, L. C., and Mackey, S. (2016), "Incidence of and Risk Factors for Chronic Opioid Use Among Opioid-Naive Patients in the Postoperative Period," *JAMA Internal Medicine*, 176, 1286–1293.

[78] Sundquist, J. and Johansson, S.-E. (1998), "The Influence of Socioeconomic Status, Ethnicity and Lifestyle on Body Mass Index in a Longitudinal Study," *International Journal of Epidemiology*, 27, 57–63.

[79] Taylor, J. and Tibshirani, R. (2018), "Post-Selection Inference For-Penalized Likelihood Models," *Canadian Journal of Statistics*, 46, 41–61.

[80] Tian, L., Zucker, D., and Wei, L. (2005), "On the Cox Model With Time-Varying Regression Coefficients," *Journal of the American Statistical Association*, 100, 172–183.

[81] Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 58, 267–288.

[82] Tibshirani, R. J., Taylor, J., Lockhart, R., and Tibshirani, R. (2016), "Exact Post-Selection Inference for Sequential Regression Procedures," *Journal of the American Statistical Association*, 111, 600–620.

[83] Tu, C. Y., Park, J., and Wang, H. (2018), "Estimation of Functional Sparsity in Nonparametric Varying Coefficient Models for Longitudinal Data Analysis," *Statistica Sinica*.

[84] UNDOC (2014), *World Drug Report 2014*.

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

[86] Wahba, G. (1990), *Spline Models for Observational Data*, vol. 59, Philadelphia: SIAM.

[87] — (1990), *Spline Models for Observational Data*, vol. 59, Siam.

[88] Wang, L., Liu, X., Liang, H., and Carroll, R. J. (2011), "Estimation and Variable Selection for Generalized Additive Partial Linear Models," *Annals of Statistics*, 39, 1827–1851.

[89] Wei, F., Huang, J., and Li, H. (2011), "Variable Selection and Estimation in High-Dimensional Varying-Coefficient Models," *Statistica Sinica*, 21, 1515–1540.

[90] Winnett, A. and Sasieni, P. (2003), "Iterated residuals and time-varying covariate effects in Cox regression," *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 65, 473–488.

[91] Wood, S. N. (2000), "Modelling and Smoothing Parameter Estimation With Multiple Quadratic Penalties," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62, 413–428.

[92] — (2006), "On confidence intervals for generalized additive models based on penalized regression splines," *Australian & New Zealand Journal of Statistics*, 48, 445–464.

[93] — (2011), "Fast Stable Restricted Maximum Likelihood and Marginal Likelihood Estimation of Semiparametric Generalized Linear Models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73, 3–36.

[94] Wu, C. O., Yu, K. F., and Chiang, C.-T. (2000), "A Two-Step Smoothing Method for Varying-Coefficient Models With Repeated Measurements," *Annals of the Institute of Statistical Mathematics*, 52, 519–543.

[95] Wu, H. and Liang, H. (2004), "Backfitting Random Varying-Coefficient Models with Time-Dependent Smoothing Covariates," *Scandinavian Journal of Statistics*, 31, 3–19.

[96] Xue, L. and Liang, H. (2010), "Polynomial Spline Estimation for a Generalized Additive Coefficient Model," *Scandinavian Journal of Statistics*, 37, 26–46.

[97] Xue, L. and Qu, A. (2012), "Variable Selection in High-Dimensional Varying-Coefficient Models With Global Optimality," *Journal of Machine Learning Research*, 13, 1973–1998.

[98] Yan, J. and Huang, J. (2012), "Model Selection for Cox Models with Time-Varying Coefficients," *Biometrics*, 68, 419–428.

[99] Yuan, M., Cai, T. T., et al. (2010), "A Reproducing Kernel Hilbert Space Approach to Functional Linear Regression," *The Annals of Statistics*, 38, 3412–3444.

[100] Zhou, J., Wang, N.-Y., and Wang, N. (2013), "Functional Linear Model with Zero-Value Coefficient Function at Sub-Regions," *Statistica Sinica*, 23, 25—-50.

[101] Zucker, D. M. and Karr, A. F. (1990), "Nonparametric Survival Analysis with Time-Dependent Covariate Effects: A Penalized Partial Likelihood Approach," *The Annals of Statistics*, 18, 329–353.

[102] Zywiel, M. G., Stroh, D. A., Lee, S. Y., Bonutti, P. M., and Mont, M. A. (2011), "Chronic Opioid Use Prior to Total Knee Arthroplasty," *The Journal of Bone and Joint Surgery*, 93, 1988–1993.