SUPERVISED STATISTICAL INFERENCE FOR DATA OF VERSATILE DIMENSIONALITY WITH APPLICATION TO GWAS STUDIES

SHENG XU

PhD

The Hong Kong Polytechnic University

2020

The Hong Kong Polytechnic University

Department of Applied Mathematics

Supervised statistical inference for data of versatile dimensionality with application to GWAS studies

Sheng Xu

A thesis submitted in partial fulfilment of the requirements

for the degree of Doctor of Philosophy

May 2020

Certificate of Originality

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

_____(Signature)

Sheng Xu (Name of student)

Abstract

Genome-Wide Association Studies (GWAS) have been successful strategies of applying biological insights into diseases in epigenetics and epigenomics in the past two decades, by linking diseases or their traits with genomic variants, environmental confounders, and clinically relevant information. The companion data used to be of versatile dimensionality and of complex data structure, posing exciting challenges and opportunities for new statistical methodology and inference, coupled with new modeling and effective computing implementation. The thesis composes of three parts and aims to address several important regression problems of estimation, hypothesis testing, and classification arising from the prevailing GWAS data pool, to meet the increasing need of statistical analytic toolsets.

Part I focuses on regression with censored survival outcomes and is motivated by data of diffuse large B-cell lymphoma (DLBCL), which integrated a large number of gene expression variants and censored survival time of patients with low sample size. This calls for efficient algorithms for feature screening and delicate statistical inference for the selected subset of influenced variables after dimensionality reduction. In Chapter 2, we present the non-monotone proximal gradient (NPG) algorithm to speed up sure joint screening for ultrahigh-dimensional Cox proportional hazard model and prove its convergence with LASSO initiator. The accompanied R-package named coxnpgsjs is fast and efficient to select a designated number of influenced gene variants from the DLBCL data. In Chapter 3, we investigate the impact of such a

subset of genetic factors on the survival time through the single-index hazard (SIH) semiparametric regression model. The SIH model is robust but challenging in efficient statistical inference owing to the nested single index structure. We propose a censored version of multiple local linear regression to attain uniformly consistent estimator of the nonparametric component and the semiparametric efficient bound for the profile likelihood estimator of the parametric component. Two classes of estimations equations are derived as the practical alternative of the score equation from the perspective of double robustness. The proposed methods and results are applied to estimate the gene effects and to detect its significance on the aforementioned lymphoma.

Part II focuses on regression with sparse longitudinal responses and is motivated by large-scale longitudinal GWAS for Alzheimer's Disease in detecting Single Nucleotide Polymorphisms (SNPs) level genotype effects on the phenotype response. It is in urgent need of powerful test procedures to detect the significance at the GWAS P-value significant threshold to the wide community of associated researchers. To compare multiple treatments, Chapters 4 and 5 present practical strategies on bootstrap procedures and apply successfully on models with Gaussian and non-Gaussian phenotype response and gigantic SNP level genotypes. This unveils some interesting association discoveries of generic effect on the disease at the GWAS significance level for the well-known Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort.

Part III focuses on regression with binary outcomes and is motivated by labeling the Multiple Sclerosis disease precisely among a population where the projection scores are skewed. In Chapter 6, we define a general distance to incorporate existing optimal functional classifiers and interpret reasonably why our proposed quantile classifier is robust. The optimal property of near perfect is derived. The accompanied classification procedure is fast and accurate. A Shiny app is built for the convenience of clinical practitioners.

Key words: ADNI cohort; functional analysis of covariance; functional classification; GWAS p-value significant threshold; longitudinal GWAS; non-monotone proximal gradient descending algorithm; single index hazards model;

Acknowledgements

I would like to express my sincere gratitude to my chief supervisor Dr. Catherine Chunling Liu. It has been an honor to be her Ph.D. student. Catherine has taught me, both consciously and unconsciously, how to be a good researcher and a good person. The joy and enthusiasm she has for her research and life were contagious and motivational for me, during all the time in the Ph.D. pursuit. I appreciate all her contributions of time, ideas, and resources to make my Ph.D. experience productive and stimulating.

A special thank is owed to Prof. Yehua Li for his scientific advice and experience and knowledge in functional data analysis. I consider him as a second advisor. He has made a great effort supervising me and provided me many insightful discussions and suggestions in research. Without Yehua's help and supervision, I would never have accomplished as much as I have done.

I am also very grateful to Prof. Yiyuan She for providing me the opportunity to visit his research group at Florida State University. His vision and passion have urged me a lot. With his abundant experience and inspiring ideas in statistical optimization, Yiyuan spent plenty of time discussing with me, from which I learned how to think critically and gradually approach the target when facing a new topic.

My thanks are extended to Dr. Jicai Liu, who led me to the world of survival analysis in the first year. I thank Dr. Tao Zhang and Dr. Haiqiang Ma, who have been very helpful in providing ideas and advice many times in functional data

analysis. I also thank Dr. T.K. Pong and Dr. Lei Yang for their insightful advice and warm discussions in optimization. My special thanks go to my former fellow mate Dr. Jin Yang who has given me generous help all the time.

I would also like to thank my Ph.D. committee, Prof. Xiao Wang, Prof. Xinyuan Song, and Prof. Defeng Sun for their helpful advice and suggestions in general.

I gratefully acknowledge the funding sources that made my Ph.D. work possible. The research covered in this thesis is fully supported by the student scholarship of the University Grants Council (UGC) of Hong Kong and partially supported by the General Research Funding (GRF) 15327216 and 15301519, Research Grants Council (RGC), UGC, Hong Kong.

I am deeply thankful to my parents who raised me with unconditional love and supported me in all my pursuits constantly. And finally, to my loving, encouraging, and patient wife Joanna, who has been always non-judgmental of me and instrumental in instilling confidence. Jo has been a wise and true supporter and has unconditionally loved me during my good and bad times. I dedicate this thesis to my family.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

Genome-wide Association Studies (GWAS) have been successful strategies of integrating biological variants into diseases and their traits in epigenetics and epigenomics in the past two decades (Tam et al., 2019), generating data of various dimensionality covering from low- to ultrahigh- dimensional data, and intrinsically infinite curve or image data. The blend of versatile data types and complicated data structures in the GWAS provide many opportunities and poses new challenges for statisticians to develop effective algorithms, new models, and the accompanying methodology to meet the urgent need of statistical analytic toolsets. The thesis is motivated by three data sets arising from the prevailing GWAS studies and aims to address a series of regression analysis of estimation, testing, and classification.

Part I is motivated by microarray studies of censored survival time for diffuse large-B-cell lymphoma cancer (DLBCL), the most common type of lymphoma worldwide (Pasqualucci et al., 2011). The data set is quite typical in the GWAS, hundreds of thousands to millions of genetic variants like gene expressions across genomes of individuals in the human population together with a limited amount of patients, and censored patient survival outcomes. The primary scientific interest is to identify signature molecular features and to detect the influence of the important molecular

factors on time to death in patients with DLBCL. This triggers our research interest in two aspects.

On one hand, we present the non-monotone proximal gradient sure joint screening (NPGSJS) algorithm to speed up joint screening for the popular Cox proportional hazard model subject to ultrahigh-dimensionality. In the literature, there is rich literature for marginal sure screening methods for the data set of ultrahigh-dimensionality but the joint screening is sporadic until feature screening based on generalized linear models Xu and Chen (2014). Yang et al. (2016) firstly applied the spirit of joint feature screening into a survival model. However, the implementation of their algorithm for ultrahigh-dimensional Cox's model was slow and lacked the theory of convergence of the algorithm. It is known that the Cox proportional hazard model is the most popular survival model widely used in biomedical studies. Therefore computing convenient algorithms will benefit the community and theoretical proof of convergent algorithm is necessary for safe use.

On the other hand, notice that the feature aberration at survival times (FAST) screening procedure Gorst-Rasmussen and Scheike (2013) is an excellent and expedient procedure to implement feature screening. However, there lack of the delicate statistical inference for the single-index hazard regression model involved when all important features are selected. Lin et al. (2011) pointed out that, the popularity of Cox's model has mathematical expedience to deal with, but may suffer misspecification and incur bias. Thus more robust hazard regression will be remedial. It also brings the theoretical challenge in establishing efficient asymptotic estimation procedure. As evidence, Ding et al. (2013) showed that even strong consistency can hardly be guaranteed unless strong assumptions are given.

Part II is driven by the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort study where phenotypic responses were collected at multiple time points. The ADNI was launched in 2004 and has had a global impact by uniting multisite data to

investigate the progression of Alzheimer's Disease (AD) to seek for better prevention and treatment. The ADNI GWAS is longitudinal in the sense that the disease-related phenotype response is repeatedly measured on irregular and sparse time points, and thus correlation structure can not be avoided. Also, the Single Nucleotide Polymorphisms (SNPs) level genotypes are hundreds of thousand giants. The scientific interest is to detect associated genotype effects on longitudinal phenotype response at the GWAS significance threshold for multiple comparisons.

Longitudinal GWAS are quite recent and thus the methods for multiple comparisons of generic effects on longitudinal phenotypic responses are in short (Xu et al., 2014; He et al., 2015; Wang et al., 2017; Visscher et al., 2017). The computational feasibility is a great challenge in multiple comparisons of large-scale longitudinal GWAS. Traditional cross-sectional methods for phenotypic data observed at a single time point in most existing GWAS are not directly applicable to longitudinal GWAS. Modeling longitudinal outcomes as functional response is a stream to develop test procedures (Reimherr and Nicolae, 2014; Huang et al., 2017). Zhang (2013) summarized fANOVA methods for dense functional data and Gaussian-type responses. Tang et al. (2016) is the first one to raise the nonparametric likelihood ratio based-test for common generalized partial linear models with sparse functional outcomes that allow non-Gaussian. Zhu et al. (2020+) further improved the power of the generalized quasi-likelihood ratio (GQLR) test of Tang et al. (2016) by incorporating within-subject correlation.

Our contribution lies in addressing the computing feasibility for comparison of multiple treatments when there are giant bootstrap samples for large scale longitudinal GWAS. We provide practical strategies and detailed procedures for the two GQLR tests aforementioned to identify the SNP level generic effects on sparse longitudinal Gaussian and non-Gaussian responses, unveiling some interesting associations: We discover the 177 SNPs that are associated with the hippocampal volume

trajectory at GWAS significant level for the ADNI 1 GWAS, coincided with some known findings by independent studies. For the whole population of the ADNI1 cohort and MCI group, we detect out that the APOE allele significantly impacts the delayed score RAVLT of patients as the count response.

Part III is motivated by Multiple Sclerosis (MS), another progressive neurological disease. Diffusion tensor imaging (DTI) provides the tract profiles to discriminate MS cases from healthy controls. The scientific problem is to assign a group label to an observed random trajectory detected by the DTI technique. It looks as if a regression problem where responses are binary categories and predictors are random trajectories. Projection classifiers are one of the main research streams. Existing projection classifiers enjoy good theoretical properties and satisfactory misclassification rates in some cases. See Delaigle and Hall (2012), Dai et al. (2017), among others. These is no unified framework to incorporate existing optimal functional classifiers. In this part, we consider both computational efficiency and theoretical development to construct a unified framework of projection classifiers. The proposed robust and yet computational expedient classifiers will be a good tool, particularly for large scale community monitoring in biomedical studies.

1.2 Organization of the thesis

The remaining of the thesis is organized as follows.

Chapter 2 presented the non-monotone proximal gradient sure joint screening (NPGSJS) algorithm to speed up joint screening for the Cox proportional hazard model subject to ultrahigh-dimensionality. We establish the convergence of the algorithm with the LASSO initiator. An R package named coxnpgsjs is available.

Chapter 3 develops a censored version of multiple local linear regression to attain semiparametric efficient bound for the profile likelihood estimator of the paramet-

ric component, and uniformly consistent estimator of the nonparametric part in the single-index hazard (SIH) semiparametric regression model. Doubly robust estimation equations. Two classes of estimations equations from the perspective of double robustness are presented as practical alternatives.

Chapters 4 and 5 focus on adapting bootstrap procedures to make multiple comparisons to reach the GWAS significance level for longitudinal GWAS data.

Chapter 6 defines a general distance to incorporate existing optimal functional classifiers and propose the weighted-quantile classifier that is convenient to implement. The optimal property of perfect is derived. A Shiny app is built for the convenience of clinical practitioners.

Chapters 2, 3, 4, 5, 6 are based on manuscripts Chen et al. (2020+); Liu et al. (2020+); Li et al. (2020+); Zhu et al. (2020+), and Ma et al. (2020+), respectively. Chapter 2 is under revision, Chapters 3, 5, and 6 are under review, and Chapter 4 is to appear.

Part I

Chapter 2

An Efficient Algorithm for Joint Feature Screening in Ultrahigh-Dimensional Cox's Model

2.1 Introduction

At contemporary biology and gene epidemiology studies and multiple other fields with survival outcomes, there is a data phenomenon called ultrahigh-dimensionality, which is of data a large scale or a huge scale in exponentially increasing relative to the so-called large-p-small-n high-dimensional data. Such ultrahigh-dimensional data accompanied with survival outcomes are encountered in a wide range of applications, particularly in microarray gene expression studies. See the diffuse large B-cell lymphoma study of Rosenwald et al. (2002), the mantle cell lymphoma study of Rosenwald et al. (2003), the neuroblastoma study of Oberthuer et al. (2006), the cytogenetically normal acute myeloid leukaemia study of Metzeler et al. (2008), among others. Thus feature screening is inevitable before variable selection and applications of conventional statistical methods. In feature screening procedures, one first concern is the effectiveness of methodologies, and the other top concern is the computational feasibility in modern genetic studies.

Hazard regression is an important tool to model survival outcome data. Within the single-index hazard model, Gorst-Rasmussen and Scheike (2013) developed a so-called FAST feature screening method, which is computationally efficient. In practice, the semiparametric Cox model may be the most popular employed due to its interpretability and high recognition. Within the Cox model, there are quite a few feature screening methods such as SIS, P-SIS, and SJS (see Fan et al. (2010), Zhao and Li (2012), Yang et al. (2016), among others). In the aforementioned feature screening methods, Gorst-Rasmussen and Scheike (2013), Fan et al. (2010), and Zhao and Li (2012) belong to the class of marginal screening methods, which is performed in one-feature-at-a-time fashion by ranking some marginal utility, say the Pearson correlation coefficient. The merit is that it is computationally highly efficient in practice. Rather than such marginal sure independent screening (SIS) schemes, as an analog to Xu and Chen (2014), Yang et al. (2016) employed a sure joint screening (SJS) strategy in the sense that it jointly estimates the model coefficients. Comparatively, SJS idea can capture more information naturally, but at the cost of heavy computational burden. Considering the popularity of the Cox model, it is imperative to develop efficient algorithms that are computational expedient for feature screening under the ultrahigh-dimensional Cox model.

Recall that, Yang et al. (2016) presented a sparsity-restricted maximum partial likelihood estimator (SMPLE) under Cox's model with the idea analogue to the sparsity-restricted maximum likelihood estimator (SMLE) in generalized linear models of Xu and Chen (2014) because both models possess the likelihood structure. It is known that the likelihood-based procedure is usually computationally costly. Yang et al. (2016) applied directly the iterative hard-thresholding (IHT) algorithm adopted by Xu and Chen (2014) and developed by She (2009). However, their theory proof cannot ensure the local convergence of the partial likelihood sequence, and therefore cannot assure the sure screening of the IHT iteration. This drives our work

in this paper for the purpose of enhancement of the algorithm and establishment of its optimality theories.

In our algorithm, we replace the diagonal matrix of the Hessian matrix by the simple identity matrix in the objective function inside each iteration. Such modification saves the computational cost by avoiding the estimation of the second derivatives of the likelihood functions, but at the cost of scarifying the effectiveness and accuracy. As compensation, we borrow the strength of the solution to the locally Lipschitz optimization problems from Chen et al. (2016) and Yang (2017) and adapt the nonmonotone proximal gradient (NPG) method for the Cox model based on two facts in optimization theory: one is that the IHT algorithm can be viewed as a proximal gradient (PG) algorithm with monotone line search under a more general framework, and the other is that the NPG algorithm is able to significantly improve the efficiency and accuracy in contrast to the monotone PG algorithm.

The contribution of this chapter is two-fold. On one hand, the proposed algorithm is efficient in that it enjoys the advantage of joint feature screening as well as its implementation is computationally fast. This is demonstrated in our simulation study compared with the other existing methods. We develop R functions for the implementation of the proposed screening procedure. This provides obvious expedience for users in epigenetic studies. On the other hand, we establish the sure screening property of the iterative screening algorithm with LASSO initial estimator strictly. The selection of the initial value will impact the performance of an algorithm. The LASSO initial estimator is preferable in practice since LASSO is a standard convex relaxation formulation.

The remainder of this chapter is organized as follows. In Section 2.2, we set up the local optimality of the constrained likelihood sequence after a brief review of the ultrahigh-dimensional Cox model. In Section 2.3, we first derive the sure screening property of the LASSO-initiated IHT algorithm. And we also provide

an implementation procedure of the NPG algorithm. In Section 2.4, we carry out several numerical studies to demonstrate the improvement of the enhanced algorithm through comparing it with the existing methods. A real data example is illustrated in Section 2.5. All theoretical proofs are left to Section 2.6.

2.2 Local Optimality of the Constrained Partial Likelihood Sequence

Recall that, Cox's model, or the proportional hazard regression, is to characterize the dependence of survival time T on p-dimensional covariate $\mathbf{Z} = (Z_1, \cdots, Z_p)^T$ in the form of

$$\lambda(t|\mathbf{Z}) = \lambda_0(t)\exp(\mathbf{Z}^T\boldsymbol{\beta}), \tag{2.1}$$

where $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T$ is a p-vector of unknown regression coefficients, and $\lambda_0(t)$ is the unspecified baseline hazard function. The observed data is denoted by $(X_i, \delta_i, \mathbf{Z}_i)$, $i = 1, \cdots, n$, where $X_i = \min(T_i, C_i)$, C_i, and $\delta_i = I(X_i \leqslant C_i)$ are observed time, censoring time, and censoring indicator for the ith subject. We assume the general settings for survival outcomes. Briefly, 1) the parameter space $\mathcal{B}$ is a compact subset of $\mathbb{R}^p$ and contains the true parameter $\boldsymbol{\beta}^*$; 2) X_i and C_i are conditionally independent given the covariates $\mathbf{Z}_i$, i.e. noninformative censoring; 3) there are no ties in the observed failure time, otherwise the technique in Breslow (1974) can be adopted.

Let $N_i(t) = I(X_i \leqslant t, \delta_i = 1)$ and $Y_i(t) = I(X_i \geqslant t)$ be the counting and at risk processes, respectively. Also, we put $\bar{N}(t) = \sum_{i=1}^n N_i(t)$. Then the log partial likelihood function (Cox (1975) and Andersen and Gill (1982)), can be written as

$$l_n(\boldsymbol{\beta}) = \sum_{i=1}^n \int_0^\tau \left[\mathbf{Z}_i^T\boldsymbol{\beta} - \log\left\{nS^{(0)}(\boldsymbol{\beta}, t)\right\}\right] dN_i(t), \tag{2.2}$$

where τ is the maximum follow-up time and $S^{(l)}(\boldsymbol{\beta}, t) = n^{-1}\sum_{i=1}^n Y_i(t)\mathbf{Z}_i^{\otimes l}\exp(\mathbf{Z}_i^T\boldsymbol{\beta})$, $l = 0, 1, 2$ with $\otimes$ being the Kronecker product. Define $M_i(t) = N_i(t) - \int_0^t Y_i(s)\exp(\mathbf{Z}_i^T\boldsymbol{\beta}^*)$

$d\Lambda_0(s)$, where $\Lambda_0(\cdot)$ is the cumulative hazard function. Then $M_i(t)$ is an orthogonal local square integrable martingale with respect to filtration $\mathcal{F}_{ti} = \sigma\{N_i(s), \mathbf{Z}_i(s^+), Y_i(s^+), 0 \leqslant s \leqslant t\}$ for $i = 1, \cdots, n$. Let $\bar{\mathbf{Z}}(\boldsymbol{\beta}, t) = S^{(1)}(\boldsymbol{\beta}, t)/S^{(0)}(\boldsymbol{\beta}, t)$ and $V(\boldsymbol{\beta}, t) = S^{(2)}(\boldsymbol{\beta}, t)/S^{(0)}(\boldsymbol{\beta}, t) - \bar{\mathbf{Z}}(\boldsymbol{\beta}, t)^{\otimes 2}$. By differentiation and some simple algebraic manipulation, it is readily seen that the score function of (2) is $\dot{l}_n(\boldsymbol{\beta}) = \partial l_n(\boldsymbol{\beta})/\partial \boldsymbol{\beta} = \sum_{i=1}^n \int_0^\tau \{\mathbf{Z}_i - \bar{\mathbf{Z}}(\boldsymbol{\beta}, t)\} dN_i(t)$, and the Hessian matrix of $-l_n(\boldsymbol{\beta})$ is $-\ddot{l}_n(\boldsymbol{\beta}) = -\partial^2 l_n(\boldsymbol{\beta})/(\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T) = \sum_{i=1}^n \int_0^\tau V(\boldsymbol{\beta}, t) dN_i(t) = \int_0^\tau V(\boldsymbol{\beta}, t) d\bar{N}(t)$.

The above description is a brief review of the Cox model. Going back to the ultrahigh-dimensional setting, we assume that $\log p = O(n^m)$ for some $0 \leqslant m < 1/2$. Then $\boldsymbol{\beta}$ is a sparse parameter vector with $p \gg n$, the jth feature of which is referred to important if $\beta_j^* \neq 0$, otherwise unimportant. According to the sparsity principle, there are only a small number of non-zero β_j^*'s. Feature screening is to identify all the important features with moderate size so that more sophisticated regularized methods can be easily carried out. To this end, we introduce a few more notations. Let M denote an arbitrary subset of $\{1, \cdots, p\}$. Denote $\boldsymbol{\beta}_M = \{\beta_j, j \in M\}$ and M_0 to be the subset of indices of all important features, i.e., $M_0 = \{j : \beta_j^* \neq 0\}$. Throughout this paper, let $\|\cdot\|_0$ be the number of non-zero coordinates or cardinality where $\cdot$ represents a vector or a set, respectively. Then $\|M_0\|_0 = \|\boldsymbol{\beta}^*\|_0 := q$, $q < k$, where q is the number of the non-zero coordinates of the true model and k is a pre-given positive integer. The number q may vary with n. Let $\mathcal{B}(k) = \{\boldsymbol{\beta} \in \mathcal{B} | \|\boldsymbol{\beta}\|_0 \leqslant k\}$. Then the sparsity-restricted maximum partial likelihood estimator (SMPLE) of the SJS procedure by Yang et al. (2016) is defined as a sparse solution

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathcal{B}(k)}{\operatorname{argmin}} \{-l_n(\boldsymbol{\beta})\}, \tag{2.3}$$

which has equivalent representation in (2.3) in Yang et al. (2016). Yang et al. (2016) proved the sure screening of SMPLE in their Theorem 2 in subsection 2.2.

Besides the sure screening, Yang et al. (2016) also gave a result about the mono-

tonicity of the constrained partial likelihood sequence through a Taylor expansion to $l_n(\boldsymbol{\beta})$, but they have no evidence for the local optimality. We are able to obtain the local optimality because we apply a different technique ((2.7) in Section 2.6) so as to achieve the sufficient ascent of the constrained likelihood sequence. The result is reported in the theorem below.

Theorem 2.1. *Let* $\rho^{(t)} = \sup\{\lambda_{\max}\{\int_0^\tau V(\bar{\boldsymbol{\beta}}, t)d\bar{N}(t)\} : \bar{\boldsymbol{\beta}} = \alpha\boldsymbol{\beta}^{(t+1)} + (1-\alpha)\boldsymbol{\beta}^{(t)}, 0 \leqslant \alpha \leqslant 1\}$, *where* $\lambda_{\max}(A)$ *represents the maximum eigenvalue of a matrix* A, *and the iterated estimator of* $\boldsymbol{\beta}$ *is denoted by,*

$$\boldsymbol{\beta}^{(t+1)} = \underset{\boldsymbol{\beta}\in\mathcal{B}(k)}{\operatorname{argmin}}\{-((\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)})^T \dot{l}_n(\boldsymbol{\beta}^{(t)}) - \frac{u}{2}\|\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)}\|_2^2)\}, \tag{2.4}$$

for $u > 0$. *Then we have:*

(1) If $u \geqslant \rho^{(t)}$, *we have* $l_n(\boldsymbol{\beta}^{(t+1)}) \geqslant l_n(\boldsymbol{\beta}^{(t)})$.

(2) Furthermore, if $\rho = \sup_{t\geqslant 0}\rho^{(t)} < \infty$, $u \geqslant \rho$ *and* $\int_0^\tau V(\boldsymbol{\beta}_M, t)d\bar{N}(t)$ *is positive definite for any* $\boldsymbol{\beta}_M$ *with the cardinality of* M *being smaller than* $2k$, *then* $\{\boldsymbol{\beta}^{(t)}\}$ *converges to a local maximum of* $l_n(\boldsymbol{\beta})$ *subject to* $\|\boldsymbol{\beta}\|_0 \leqslant k$.

Remark 1. In the second result in afore Theorem 2.1, we give a value $2k$ as a sparse recovery guarantee so that we have the local convergence property of the algorithm. It further provides a theoretical support for the exercise of IHT in practice. The first result of the monotonicity of the likelihood sequence in afore Theorem 2.1 corresponds to Theorem 1 in subsection 2.1 of Yang et al. (2016).

Remark 2. Notice that the objective function in Yang et al. (2016) and ours can be unified in the form of $Q_n(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(t)}) = l_n(\boldsymbol{\beta}^{(t)}) + (\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)})^T\dot{l}_n(\boldsymbol{\beta}^{(t)}) - u/2(\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)})^T\boldsymbol{W}(\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)})$. The working matrix $\boldsymbol{W}$ of Yang et al. (2016) is $diag\{\ddot{l}(\boldsymbol{\beta}^{(t)})\}$, whereas ours is an identity matrix $\boldsymbol{I}$. The later reduces the heavy computation load caused by $\ddot{l}(\boldsymbol{\beta}^{(t)})$ to some extend. But the side effect is that it may give rise to the estimation inaccuracy. We address the problem in next section.

Remark 3. Our iterative estimation (2.4) corresponds to the constrained maximization problem (2.5) in Yang et al. (2016). In implementation, one first solves (2.4) without the L_0 norm constraint and denotes the solution by $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}^{(t)} + u^{-1}\dot{l}_n(\boldsymbol{\beta}^{(t)})$. Then, let

$$\boldsymbol{\beta}^{(t+1)} = \mathbf{H}(\tilde{\boldsymbol{\beta}}; k) \equiv \left\{H(\tilde{\beta}_1; \tilde{\beta}_{(k)}), \cdots, H(\tilde{\beta}_p; \tilde{\beta}_{(k)})\right\}^T, \tag{2.5}$$

where $H(\tilde{\beta}_j; \tilde{\beta}_{(k)}) = \tilde{\beta}_j I(|\tilde{\beta}_j| \geqslant \tilde{\beta}_{(k)})$ and $\tilde{\beta}_{(k)}$ is the kth largest component of $|\tilde{\boldsymbol{\beta}}|$. The iteration (2.5) is called the iterative hard-thresholding (IHT) algorithm adapted from the iterative thresholding algorithms in She (2009). Both Xu and Chen (2014) and Yang et al. (2016) used such algorithm to realize their sparse likelihood estimation procedure.

Remark 4. For the IHT algorithm, the selection of the step size, the reciprocal of u, is an inevitable step. Although Yang et al. (2016) proposed the ITH algorithm under the Cox model, they did not declare the method of selection u in practice. The monotone line search for u recommended in Xu and Chen (2014) can also be applied to the IHT iteration for the Cox model. In this way, the IHT algorithm with the monotone line search that was used both in Xu and Chen (2014) and in Yang et al. (2016) can be viewed as a proximal gradient (PG) algorithm with monotone linear search within a more general framework. It is known that the non-monotone proximal gradient (NPG), stacking up against PG, will be much more efficient in practice since it potentially reduces number of inner loops. This motivates us to employ the NPG to improve the IHT algorithm in the next section.

2.3 Sure Screening of Cox-LASSOi-IHT iteration and an NPG Algorithm

In previous section, we have derived the local optimality of the constrained partial likelihood sequence. Now it comes to the issues for computational feasibility of

the whole feature screening algorithm. Our aim here is twofold. One is about the selection of the initial value of the algorithm. We establish the sure screening for the IHT procedure initiated by LASSO for the Cox model (abbreviated as Cox-LASSOi-IHT hereinafter). The other is to present an NPG algorithm to settle down the efficiency and accuracy problems mentioned in the Remarks 2 and 4 of Theorem 2.1.

2.3.1 Sure Screening of Cox-LASSOi-IHT iteration

Yang et al. (2016) derived the sure screening property of the SMPLE and used the IHT algorithm. However, they left the unsolved question that whether the IHT algorithm necessarily leads to the SMPLE so that Theorem 2 in Yang et al. (2016) (sure screening property of SMPLE) is applicable. Unfortunately, there is no guarantee that $\hat{\boldsymbol{\beta}}$ is the outcome of a single run of the IHT due to the complexity of the minimization problem in (2.3).

It is known that an appropriate selection of the initial value is able to increase the chance of hitting the local maximizer and hence enhance the accuracy. In addition, a good initial setup can further save the computational cost in the sense that less number of the iterations are carried out. LASSO, as the standard convex relaxation formulation for sparse learning, is widely used in practice because it leads to sparse solution and is computational efficient (Zhang (2009)). Under certain conditions (Zhao and Yu (2006), Meinshausen and Bühlmann (2006)), LASSO is model selection consistency. Since LASSO considers the joint effects of the predictors, it can serve as an initial estimator in the IHT procedure.

The following Theorem 2.2 states the sure screening property of the Cox-LASSOi-IHT iteration.

Theorem 2.2. *Under Assumptions 2.1 to 2.7 in Section 2.6, let* $\log(p) = O(n^m)$, $\tau_1 + \tau_2 < \frac{1}{2} - m$, $0 \leqslant m < 1/2$, *and* $u > c_6 rn$ *with* $r = O(n^{\tau_3})$, *where* τ_3 *is defined in Assumption 2.7, and* c_6 *is a positive constant. Define* $\boldsymbol{\beta}^{(0)} = \text{argmin}_{\boldsymbol{\beta}}\{-l_n(\boldsymbol{\beta}) +$

$n\lambda\|\boldsymbol{\beta}\|_1\}$, where λ satisfies $\lambda n^{\frac{1}{2}-m} \to \infty$, $\lambda n^{\tau_1+\tau_2} \to 0$ and $\max_j \sigma_j^2 = O(\lambda n^{\frac{1}{2}})$, and let $M^{(t)} = \{j : \hat{\beta}_j^{(t)} \neq 0\}$, where σ_j^2 is defined in Assumption 2.3 and τ_1 and τ_2 are defined in Assumption 2.4. Then we have

$$\mathrm{pr}(M_0 \subset M^{(t)}) \to 1,$$

for any finite $t \geqslant 1$, as n goes to infinity.

Although the exact SMPLE $\hat{\boldsymbol{\beta}}$ could not be obtained, the above Theorem 2.2 guarantees the sure screening property of the Cox-LASSOi-IHT iteration. It is still hard to determine which estimator shall be employed in the entire solution path of the LASSO. We select λ that recruits the first k features in the solution path of LASSO as the chosen tuning parameter value. This selection works well for all the numerical examples in next section. In addition, from the proof of Theorem 2.2, we have $\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|_\infty = o_p(w)$ for any t, where $w^{-1} = O(n^{\tau_1})$ (see the proof of Theorem 2.2 in Section 2.6 for detail), implying that the SMPLE may have desirable estimation accuracy, as is validated in our numerical studies.

2.3.2 An NPG Algorithm

Theorem 1 in Yang et al. (2016) provides insight about the choice of the step size, but they did not declare their exact method of step size selection in the numerical study. In the following, we allow the step size in each iteration is different to other iterations. For the tth iteration, the step size is denoted by $1/u_t$. To provide a feasible way of choosing the step size $1/u_t$ in practice, we might apply the method of choosing u_t suggested in Xu and Chen (2014). That is, at each iteration, we will keep decreasing the step size by multiplying a constant less than 1 till the new partial likelihood value is larger than the last one. Under a more general framework, such method can be viewed as a proximal gradient method. Analogue to Xu and Chen

(2014), the algorithm in Yang et al. (2016) could be summarized as the proximal gradient (PG) method.

Algorithm 1 Proximal Gradient

Obtain an initial estimator $\boldsymbol{\beta}^{(0)}$. Choose $L_0 > 0$, $\tau > 1$, and set $t = 0$.

(1) Set $u_t = L_0$

(1a) Solve the subproblem

$$\boldsymbol{\gamma} = \underset{\boldsymbol{\beta} \in \mathcal{B}(k)}{\operatorname{argmin}} \{ -[l_n(\boldsymbol{\beta}^{(t)}) + (\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)})^T \dot{l}_n(\boldsymbol{\beta}^{(t)}) - \frac{u_t}{2} (\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)})^T \boldsymbol{W} (\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)})] \},$$

where $\boldsymbol{W} = diag\{-\ddot{l}(\boldsymbol{\beta}^{(t)})\}$.

(1b) If $l_n(\boldsymbol{\gamma}) > l_n(\boldsymbol{\beta}^{(t)})$, go to Step 2.

(1c) Set $u_t \leftarrow \tau u_t$ and go to step (1a).

(2) Set $\boldsymbol{\beta}^{(t+1)} \leftarrow \boldsymbol{\gamma}$, $t \leftarrow t + 1$, and go to step (1).

END

Specifically, Xu and Chen (2014) kept doubling u_t till the criterion 1(b) is satisfied, that is, they used $\tau = 2$. Because the computation cost of $\boldsymbol{W}$ is large and the line search method is monotone, the Algorithm 1 might be not efficient. To make the method more applicable in applications, we employ non-monotone proximal gradient (NPG) method (Algorithm 2), namely, the proximal gradient method with a non-monotone line search for solving the minimization problem (2.4) (See Wright et al. (2009) and Chen et al. (2016)).

In the step 1(a)'s of the both algorithms, the hard thresholding method of She (2009) is applied for solving the subproblems. Specifically, one may first compute the surrogate quantity without consider the constraint $\mathcal{B}(k)$. Then, the step 1(a) can be solved by keeping the largest k values of entries of the surrogate quantity. Notice that the working matrix $\boldsymbol{W}$ in Algorithm 1 is replaced by the identity matrix in Algorithm 2 so that the computational loads reduced in the later algorithm.

Algorithm 2 Non-monotone Proximal Gradient

Obtain an initial estimator $\boldsymbol{\beta}^{(0)}$. Choose $L_{max} > L_{min} > 0$, $\tau > 1$, $c > 0$ and and integer $M \geqslant 0$. Set $t = 0$.

(1) Choose $L_0^{(t)} \in [L_{min}, L_{max}]$ arbitrarily. Set $u_t = L_0^{(t)}$

(1a) Solve the subproblem

$$\boldsymbol{\gamma} = \underset{\boldsymbol{\beta} \in \mathcal{B}(k)}{\operatorname{argmin}} \{-Q_{n,t}(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(t)})\},$$

where $Q_{n,t}(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(t)}) = l_n(\boldsymbol{\beta}^{(t)}) + (\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)})^T \dot{l}_n(\boldsymbol{\beta}^{(t)}) - \frac{u_t}{2}\|\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)}\|_2^2$.

(1b) If

$$l_n(\boldsymbol{\gamma}) \geqslant \min_{(t-M)^+ \leqslant i \leqslant t} l_n(\boldsymbol{\beta}^{(i)}) + \frac{1}{c}\|\boldsymbol{\gamma} - \boldsymbol{\beta}^{(t)}\|_2^2,$$

go to Step 2.

(1c) Set $u_t \leftarrow \tau u_t$ and go to step (1a).

(2) Set $\boldsymbol{\beta}^{(t+1)} \leftarrow \boldsymbol{\gamma}$, $t \leftarrow t + 1$, and go to step (1).

END

NPG is an efficient strategy for accelerating PG method through finding a proper step size at each iteration. The main difference of these two algorithms lies in the line search step 1(b)'s. The monotone line search in PG method requires a larger number of loops to find an appropriate step size. While the non-monotone line search in NPG method relaxes the requirement for the choice of the step size. That is, instead of requiring the new log likelihood value greater than the previous one in PG, the NPG only requires the new log likelihood value greater than the minimum value of previous M log likelihood values, where M is a positive integer. Our numerical experience suggests that set M to be 4 works well for the proposed method. The parameter M is also commonly set to be 4 in the practice of using non-monotone proximal gradient methods, see Chen et al. (2016), Yang (2017), among others. Empirically, it has been shown that NPG can obtain better numerical performance in many applications. Our numerical experiments also demonstrate the advantage of the NPG method.

$$\min F(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(t)}) := -Q_n(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(t)}) + P(\boldsymbol{\beta}), \tag{2.6}$$

where $P(\boldsymbol{\beta})$ takes value 0 if $\boldsymbol{\beta} \in \mathcal{B}(k)$ and ∞ otherwise. Then $-Q_n(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(t)})$ and $P(\boldsymbol{\beta})$ satisfy the conditions in Assumption A.1 of Chen et al. (2016). Chen et al. (2016) provided the sparse optimization solutions to three scenarios of objective functions, nonconvex, locally Lipschitz, and non-Lipschitz. Our objective function belongs to the second. See Section 2.6 for assumption verification in details.

In practice we use the Barzilai-Borwein method (see Barzilai and Borwein (1988)) for the choice of $L_0^{(t)}$. That is,

$$L_0^{(t)} := \min\left\{\max\left\{\frac{[\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^{(t-1)}]^T[\dot{l}(\boldsymbol{\beta}^{(t-1)}) - \dot{l}(\boldsymbol{\beta}^{(t)})]}{\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^{(t-1)}\|_2^2}, L_{min}\right\}, L_{max}\right\},$$

for any $t \geqslant 1$. The Barzilai-Borwein step size approximates $\ddot{l}(\boldsymbol{\beta}^{(t)})$ through a single constant. Therefore, compared with Algorithm 1, where $\ddot{l}(\boldsymbol{\beta}^{(t)})$ is approximated by $u_t diag\{\ddot{l}(\boldsymbol{\beta}^{(t)})\}$, Algorithm 2 saves more computational cost. Numerically, there is often less than two loops in each non-monotone line search backtracking after the initial iteration when the Barzilai-Borwein method is applied.

We use the termination criterion which is commonly used in the general iterative shrinkage and thresholding algorithm. (See Gong et al. (2013) and Wright et al. (2009)) That is, we terminate the algorithm if

$$\frac{\|\dot{l}(\boldsymbol{\beta}^{(t)}) - \dot{l}(\boldsymbol{\beta}^{(t-1)})\|_2 + u_t\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^{(t-1)}\|_2}{\max\{1, \|\boldsymbol{\beta}^{(t)}\|_2\}} \leqslant 10^{-3}.$$

Another important problem in practice is choosing the hyperparameter k in the estimation procedure. Choosing $k = [n/(a_0)\log(n)]$, where a_0 is a positive constant, is widely used in practice, see Fan et al. (2010), Fan and Lv (2008), among others. In our proposed procedure, we adopt this rule. Based on our experience from the

numerical experience and real data analysis, a smaller k makes the proposed method slightly easier to succeed.

2.4 Simulation Studies

In this section, we examine and compare the empirical performance of the proposed NPGSJS method with several marginal independence screening alternatives including the SIS/ISIS of Fan and Song (2010), the FAST/IFAST of Gorst-Rasmussen and Scheike (2013), the SMPLE-based sure joint screening (abbreviated as SJS thereafter) of Yang et al. (2016), and the popular LASSO of Tibshirani (1997), through extensive Monte Carlo simulations. We do not intend to compare our NPGSJS screening with all the existing competitors, and believe that superiority of this approach over marginal screening and SJS methods could be fully exhibited through our comparisons to SIS/ISIS, FAST/IFAST, and SJS. To have a fair comparison with the SJS method, we use the LASSO as the initial estimator of the SJS, although there is no theoretical support that the algorithm of SJS with the LASSO initial has sure screening property.

We generate the survival time from model (2.1) with features, regression coefficients, and the baseline hazard function being $\lambda_0(t) = 1$ in Examples 2.1 to 2.3 and 2.5. In Example 2.4, the survival time follows the accelerated failure time model with standard normally distributed error, which means that the distribution of the survival time is misspecified by the Cox model. In all examples except Example 2.5, the censoring time follows the uniform distribution on $(0, c_0)$, where different c_0's are chosen to produce approximately 30% to 35% censoring rate. In Example 2.5, the censoring time is generated from the accelerated failure time model. Some other detailed elements of the simulation setup are given as follows:

Example 2.1, with mutually independent features, is the most straightforward

for variable screening, while Examples 2.2 and 2.3 allow for moderate and high correlation between features become somewhat difficult.

Example 2.1. *(The Cox model with mutually independent features)* $\mathbf{Z} = (Z_1, \cdots, Z_p)^T$ *follows the multivariate normal distribution with mean* $\mathbf{0}$ *and the covariance matrix* $\Sigma = (\sigma_{ij})_{p\times p}$ *with* $\sigma_{ii} = 1$ *and* $\sigma_{ij} = 0$ *for* $i \neq j$, $i, j = 1 \cdots, p$. $M_0 = \{1, 2, 3, 4, 5, 6\}$, $\boldsymbol{\beta}_{M_0} = (-1.6328, 1.3988, -1.6497, 1.6353, -1.4209, 1.7022)^T$, $\boldsymbol{\beta}_{M_0^c} = \mathbf{0}$, *and* $(n, p) = (120, 10000)$.

Example 2.2. *(The Cox model with moderately correlated features) The same as Example 2.1 except that* $\sigma_{ij} = 0.5$ *for* $i \neq j$, *and* $(n, p) = (120, 2000)$.

Example 2.3. *(The Cox model with highly correlated features) The same as Example 2.1 except that* $\sigma_{ij} = 0.8$ *for* $i \neq j$, *and* $(n, p) = (150, 2000)$.

We then use Example 2.4 to check the robustness of all the methods to model misspecification.

Example 2.4. *(The survival time follows the accelerated failure time model)* $\mathbf{Z} = (Z_1, \cdots, Z_p)^T$ *follows the multivariate normal distribution with mean* $\mathbf{0}$ *and the covariance matrix* $\Sigma = (\sigma_{ij})_{p\times p}$ *with* $\sigma_{ii} = 1$. $M_0 = \{1, 2, 3, 4\}$. *When* $i \neq j$, $\sigma_{ij} = 0.15$ *for* $i, j \in M_0$ *and* $\sigma_{ij} = 0.3$ *for* i *or* $j \in M_0^c$. $\boldsymbol{\beta}_{M_0} = (3, 3, 3, 3)^T$, $\boldsymbol{\beta}_{M_0^c} = \mathbf{0}$, *and* $(n, p) = (200, 1000)$. *In addition, the survival time* T *follows the accelerated failure time model* $\log(T) = Z^T\boldsymbol{\beta} + \varepsilon$ *with* ε *being* $N(0, 1)$ *random variable.*

Example 2.5. *(The censoring time follows the accelerated failure time model)* $\mathbf{Z} = (Z_1, \cdots, Z_p)^T$ *follows the multivariate normal distribution with mean* $\mathbf{0}$ *and the covariance matrix* $\Sigma = (\sigma_{ij})_{p\times p}$ *with* $\sigma_{ii} = 1$. $M_0 = \{1, 2, 3, 4\}$. *When* $i \neq j$, $\sigma_{ij} = 0.15$ *for* $i, j \in M_0$ *and* $\sigma_{ij} = 0.3$ *for* i *or* $j \in M_0^c$. $\boldsymbol{\beta}_{M_0} = (3, 3, 3, 3)^T$, *and* $(n, p) = (200, 1000)$. *Besides, the censoring time* C *follows the accelerated failure*

time model, $\log(C) = Z^T\boldsymbol{\alpha} + \varepsilon$, *where* $\boldsymbol{\alpha} = (1, 1, 0, 0, 0, 0, 1, 1, 0, \cdots, 0)^T$ *and* ε *follows* $U(3.5, 5)$.

In Example 2.5, the distribution of censoring time relies on the features in the way that the theoretical results in Gorst-Rasmussen and Scheike (2013) cannot be guaranteed.

For SIS/ISIS, we use the univariate maximum partial likelihood estimator as the marginal utility. Five loops are performed for ISIS, and in each loop features are selected by LASSO with tuning parameter chosen using the BIC criterion. We conduct FAST/IFAST directly by the function "ahazisis" in R package **ahaz**. In addition, for a fair comparison, we use IFAST in the same way that we perform for ISIS. As for LASSO, we firstly obtain its entire regularization path by the function "glmnet" in R package **glmnet**. Then the optimal tuning parameter is determined by the BIC criterion. With the help of the LASSO initial estimator, we implement the NPGSJS by the Algorithm 2 with $\tau = 2$, $M = 4$, $L_{max} = 10^8$, $L_{min} = 1$, and $c = 10^{-4}$. These settings are common for NPG, see Yang (2017), Chen et al. (2016), among others. Our simulation results also demonstrate that these values work well empirically. We developed R functions for the implementation of the proposed screening procedure.

As was suggested by Fan et al. (2010) and was widely used in the literature, we set the threshold to be $k = [n/(3\log(n))]$. In our limited experience, the NPGSJS screening is easier to succeed for smaller k.

All of our simulation results are based on $L = 1000$ independently simulated datasets. To summarize these results, we consider the following performance measures: RC, the retaining capacity of all important features, $L^{-1}\sum_{l=1}^{L} I(M_0 \subset \hat{M}_l)$; RC_j, the retaining capacity of the jth important feature, $L^{-1}\sum_{l=1}^{L} I(\hat{\beta}_j^{(l)} \neq 0)$ for $j = 1, \cdots, q$; PSR, the positive selection rate, $L^{-1}\sum_{l=1}^{L} \|M_0 \cap \hat{M}_l\|_0/q$; FDR, the

Table 2.1: Summary statistics for Examples 2.1 to 2.3.

Setup	Method	RC	PSR	FDR	AMS	TP	L1.err	L2.err	Time
Example 2.1	SIS	0.011	0.605	0.546	8	3.630	9.446	10.632	-
	ISIS	0.055	0.628	0.529	8	3.768	7.646	8.473	58.563
	FAST	0.016	0.628	0.529	8	3.769	-	-	-
	IFAST	0.092	0.668	0.499	8	4.009	8.702	11.427	0.768
	LASSO	0.402	0.838	0.319	7	5.033	8.583	12.224	0.189
	SJS	0.849	0.946	0.290	8	5.677	2.759	1.809	173.221
	NPGSJS	**0.926**	**0.969**	**0.274**	8	**5.812**	**2.515**	**1.388**	43.250
Example 2.2	SIS	0.013	0.523	0.608	8	3.135	9.859	12.241	-
	ISIS	0.309	0.749	0.438	7.736	4.493	5.403	5.403	15.317
	FAST	0.017	0.554	0.584	8	3.326	-	-	-
	IFAST	0.174	0.734	0.449	7.943	4.406	8.055	9.852	0.191
	LASSO	0.428	0.845	0.306	7.326	5.067	8.300	11.438	0.016
	SJS	0.675	0.918	0.323	8	5.506	3.407	2.447	23.643
	NPGSJS	**0.847**	**0.958**	**0.282**	8	**5.746**	**3.108**	**1.952**	1.555
Example 2.3	SIS	0.024	0.468	0.688	9	2.810	10.060	13.383	-
	ISIS	0.556	0.657	0.562	7.143	3.944	5.896	6.479	17.612
	FAST	0.016	0.496	0.669	9	2.976	-	-	-
	IFAST	0.365	0.735	0.510	8.214	4.413	7.037	7.361	0.199
	LASSO	0.357	0.832	**0.372**	7.968	4.992	8.013	10.588	0.018
	SJS	0.500	0.888	0.408	9	5.325	5.151	3.886	76.996
	NPGSJS	**0.754**	**0.941**	0.373	9	**5.643**	**4.131**	**2.836**	5.627

∗Above the dashed line, marginal screening methods; below the dashed line, joint screening methods.

false discovery rate, $L^{-1}\sum_{l=1}^{L}\|\hat{M}_l - M_0\|_0/\|\hat{M}_l\|_0$; AMS, the average model size, $L^{-1}\sum_{l=1}^{L}\sum_{j=1}^{p} I(\hat{\beta}_j^{(l)} \neq 0)$; TP, the average number of important feature selected, $L^{-1}\sum_{l=1}^{L}\sum_{j=1}^{q} I(\hat{\beta}_j^{(l)} \neq 0)$; L1err, $L^{-1}\sum_{l=1}^{L}\|\hat{\boldsymbol{\beta}}^{(l)} - \beta\|_1 = L^{-1}\sum_{l=1}^{L}\sum_{j=1}^{p}|\hat{\beta}_j^{(l)} - \beta_j|$; L2err, $L^{-1}\sum_{l=1}^{L}\|\hat{\boldsymbol{\beta}}^{(l)} - \beta\|_2 = L^{-1}\sum_{l=1}^{L}\sqrt{\sum_{j=1}^{p}(\hat{\beta}_j^{(l)} - \beta_j)^2}$; Time, the average time, in seconds, for each method. All the experiences are run in a server with an Intel Xeon E7-4890 v2 CPU (2.80 GHz).

In our simulation results, we do not present L1err and L2err for the method of FAST, because the function "ahazisis" in R packages **ahaz** does not compute the marginal regression coefficients, and the corresponding estimators from IFAST must be more accurate than those obtained from FAST.

Simulation results for Examples 2.1 to 2.3 are presented in Tables 2.1 to 2.3, from which we can see that the NPGSJS method outperforms the other methods uniformly in terms of the above-mentioned measures. Overall, the joint screening methods

Table 2.2: Retaining capability of Examples 2.1 to 2.3.

Setup	Method	RC_1	RC_2	RC_3	RC_4	RC_5	RC_6
Example 2.1	SIS	0.679	0.398	0.682	0.663	0.445	0.763
	ISIS	0.703	0.426	0.705	0.684	0.478	0.772
	FAST	0.696	0.444	0.696	0.684	0.48	0.769
	IFAST	0.750	0.494	0.747	0.728	0.506	0.784
	LASSO	0.898	0.710	0.878	0.890	0.745	0.912
	SJS	0.955	0.905	0.954	0.956	0.933	0.974
	NPGSJS	**0.980**	**0.951**	**0.975**	**0.972**	**0.955**	**0.979**
Example 2.2	SIS	0.560	0.415	0.570	0.583	0.394	0.613
	ISIS	0.808	0.614	0.807	0.819	0.614	0.831
	FAST	0.585	0.448	0.606	0.601	0.442	0.644
	IFAST	0.781	0.585	0.803	0.793	0.612	0.832
	LASSO	0.883	0.742	0.895	0.887	0.759	0.901
	SJS	0.934	0.863	0.944	0.945	0.858	0.962
	NPGSJS	**0.962**	**0.921**	**0.969**	**0.966**	**0.946**	**0.982**
Example 2.3	SIS	0.437	0.389	0.516	0.579	0.357	0.532
	ISIS	0.683	0.548	0.722	0.714	0.619	0.659
	FAST	0.492	0.405	0.532	0.611	0.389	0.548
	IFAST	0.746	0.571	0.818	0.833	0.683	0.762
	LASSO	0.841	0.714	0.921	0.929	0.730	0.857
	SJS	0.937	0.802	**0.944**	0.881	0.810	0.952
	NPGSJS	**0.952**	**0.905**	0.992	**0.984**	**0.889**	**0.921**

∗Above the dashed line, marginal screening methods;
below the dashed line, joint screening methods.

Table 2.3: Biases of coefficient estimators for Examples 2.1 to 2.3.

Setup	Method	β_1	β_2	β_3	β_4	β_5	β_6
Example 2.1	ISIS	1.062	-1.068	1.071	-1.074	1.053	-1.063
	IFAST	1.386	-1.256	1.405	-1.397	1.277	-1.434
	LASSO	1.461	-1.295	1.476	-1.467	1.314	-1.508
	SJS	0.185	-0.194	0.191	-0.184	0.184	-0.184
	NPGSJS	**0.071**	**-0.073**	**0.068**	**-0.064**	**0.079**	**-0.067**
Example 2.2	ISIS	0.630	-0.716	0.637	-0.625	0.716	-0.629
	IFAST	1.274	-1.160	1.275	-1.272	1.179	-1.303
	LASSO	1.410	-1.258	1.413	-1.411	1.275	-1.452
	SJS	0.343	-0.353	0.340	-0.341	0.376	-0.335
	NPGSJS	**0.325**	**-0.314**	**0.318**	**-0.322**	**0.296**	**-0.316**
Example 2.3	ISIS	0.616	-0.686	0.534	-0.584	0.622	-0.666
	IFAST	1.046	-0.985	0.961	-0.953	0.961	-1.035
	LASSO	1.353	-1.223	1.332	-1.337	1.234	-1.390
	SJS	0.564	-0.613	0.544	-0.622	0.587	-0.581
	NPGSJS	**0.131**	**-0.168**	**-0.003**	**-0.084**	**0.141**	**-0.147**

∗Above the dashed line, marginal screening methods;
below the dashed line, joint screening methods.

(LASSO, SJS, and NPGSJS) outperform the independent screening methods. The NPGSJS method has the highest overall RC and individual RCs, PSRs and TPs, the smallest L1errs and L2errs, and nearly the smallest FDRs except for Example 2.3. Comparing with the marginal screening methods and LASSO, we find that the improvements of the NPGSJS over the other methods are quite substantial. The second best SJS method performs just slightly worse than the NPGSJS, but the its computing time is much longer.

As for Examples 2.1 and 2.2, the RCs for the NPGSJS are larger than 0.92, which means that the NPGSJS can recognize all the important features more than 92 times in every 100 runs, while for other approaches, the values of RCs are always less than 0.92. The implication that NPGSJS can identify all important features can also be seen from the TPs. On the other hand, the lowest FDR implies that the proposed method selected least unimportant features. From the view of estimation accuracy, the NPGSJS still works well based on the values of L1err and L2err. In Example 2.3, results from all the other methods are not satisfactory, because of the presence of high correlations between features. However, the NPGSJS, with significant shorter computing time than that of SJS, is still working well.

From Table 2.2, we can see that the retaining capabilities of the NPGSJS for each important feature are close to 1. In addition, the retaining capabilities of the other methods for each important feature are not too small, while the overall retaining capabilities are remarkably smaller than each individual retaining capability. Our simulation results also suggest that moderate correlation between features could improve the performance of iterative version of marginal independence screening methods, which is also noted by Fan et al. (2010). The small values of L1err and L2err for the k-sparse MPLE screening method prompt us to see how accurate the estimated regression coefficients could be. Table 2.3 presents the biases of coefficient estimators of important features, from which we can see that the NPGSJS is unbiased

Table 2.4: Summary statistics for Examples 2.4 and 2.5.

Setup	Method	RC	PSR	FDR	AMS	TP	L1.err	L2.err	Time
Example 2.4	SIS	0.000	0.306	0.898	12	1.222	23.200	53.548	-
	ISIS	0.016	0.339	0.887	12	1.357	16.166	45.158	4.652
	FAST	0.000	0.304	0.899	12	1.214	-	-	-
	IFAST	0.103	0.423	0.859	11.992	1.691	12.767	37.307	0.208
	LASSO	0.055	0.417	0.834	9.691	1.667	**12.378**	**36.869**	0.010
	SJS	0.849	0.894	0.702	12	3.574	26.492	144.099	58.751
	NPGSJS	**0.992**	**0.992**	**0.669**	12	**3.968**	37.273	957.691	4.110
Example 2.5	SIS	0.008	0.351	0.883	12	1.405	19.870	37.467	-
	ISIS	0.087	0.389	0.870	12	1.556	13.120	29.676	5.160
	FAST	0.000	0.314	0.896	12	1.254	-	-	-
	IFAST	0.159	0.480	0.839	11.968	1.921	12.3334	33.510	0.219
	LASSO	0.103	0.470	0.812	9.492	1.881	12.042	34.921	0.010
	SJS	0.882	0.918	0.694	12	3.671	**4.960**	**5.127**	56.592
	NPGSJS	**0.992**	**0.992**	**0.669**	12	**3.968**	12.947	863.341	4.316

∗Above the dashed line, marginal screening methods; below the dashed line, joint screening methods.

numerically, while estimators from other methods might be biased. This phenomenon indicates that the NPGSJS can also be used for estimation.

Tables 2.4 displays the simulation results for Examples 2.4 and 2.5. These results suggest that the NPGSJS may not be sensitive to model misspecification and is reliable for complex censoring mechanism, at least for the scenarios under our consideration. Except the joint screening methods, all the other methods completely fail in Examples 2.4 and 2.5. For both examples, the proposed NPGSJS method obtains nearly 100% RCs and smallest FDRs. Although the performance of SJS is acceptable, the NPGSJS is always performing slightly better than SJS with a significantly shorter computing time. Although the accuracy of the proposed method is not satisfactory under these cases, the summary statistics like RC, PSR, and FDR should be the top propriety indeed.

As we mentioned in Section 2, the SJS method takes the diagonal elements of the Hessian matrix into account. As a result, theoretically, it should be more accurate than the proposed method, at least when the model is specified correctly. However, the simulation results for Examples 2.1 to 2.3 report that the proposed method is more accurate than SJS in the sense that it obtains smaller L1err and L2err. The

reason might be that the NPGSJS, benefits from the NPG algorithm, is able to obtain the numerically larger likelihood and thus estimates the coefficients more accurately.

Figure 2.1 demonstrates the advantage and improvement of the NPG method by comparing with the PG method. The left pane is the partial log likelihood value versus the iterations plot of NPG (red solid line) and PG (blue dashed line) for a single trail under data setting of Example 2.2. It typically illustrates how the function value changes during the iterations. The red solid line is not monotone but achieve a larger value faster than the blue dashed line. To show the average trend, we plot the mean of the function values of 100 trails at the right pane. It is easy to see that numerically, through the NPG method, we can obtain a larger function value in smaller number of iterations.

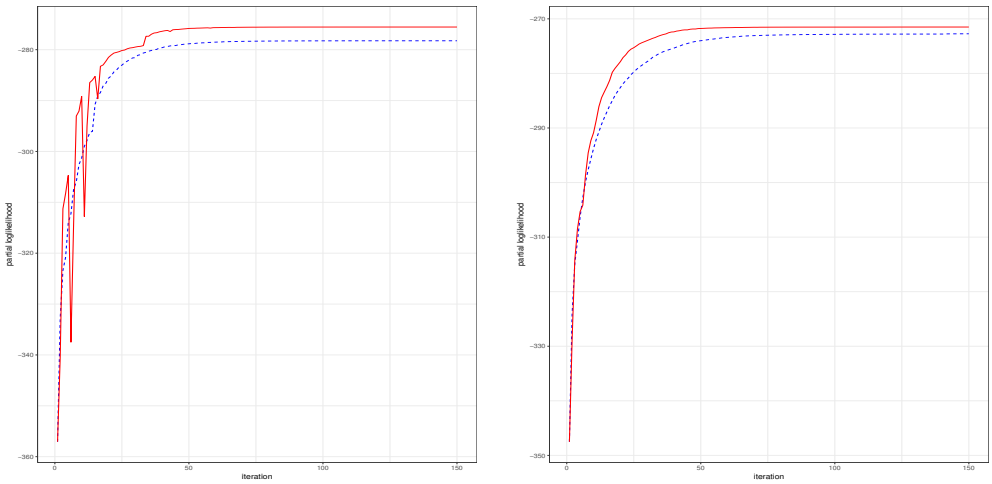

Figure 2.1: The partial log likelihood of NPG (red solid line) and PG (blue dashed line) under data setting 2. The left pane is a single data set and the right one is the average of 100 trails.

2.5 Analysis of Diffuse Large B-Cell Lymphoma

We apply our method for the diffuse large B-cell lymphoma data of Rosenwald et al. (2002). The diffuse large-B-cell lymphoma is the most common type of lymphoma

among adults. The survival time of patients with such kind of lymphoma after chemotherapy is affected by genes of the tumors. In this dataset of Rosenwald et al. (2002), 7399 gene expressions were obtained retrospectively from 240 patients with untreated diffuse large-B-cell lymphoma who had no previous history of lymphoma and receiving chemotherapy. The observed time after the chemotherapy ranges from 0 to 21.8 years. Of these 240 patients, 102 have censored survival times, causing 42.5% censoring rate.

Setting the threshold at $k = [240/(3\log(240))] = 14$, we employ various feature screening methods to select the influential genes on the survival time. Univariate feature selection method (Uni), as on of the most popular methods in medicine research (see van Wieringen et al. (2009)), is also included for comparison. The summary statistics reported below are computed through 5-fold cross-validation. The important genes are selected through different feature screening methods based on the training set. Then we fit the model through penalized partial likelihood with LASSO and SCAD penalties based on the selected variables for the testing set. The cross-validation procedure is repeated 100 times. The mean of log likelihood, AIC, and BIC of each selected model are reported in Table 2.5. The best results are in bolded.

Table 2.5: Log likelihood, AIC, and BIC of resulting models.

	Log Likelihood	AIC	BIC
NPGSJS-SCAD	**-85.89**	**174.48**	**177.83**
NPGSJS-LASSO	-86.75	177.06	178.93
SJS-SCAD	-86.94	175.87	178.60
SJS-LASSO	-87.59	178.06	179.48
ISIS-SCAD	-88.63	178.20	179.77
ISIS-LASSO	-88.75	179.63	180.19
IFAST-SCAD	-88.35	177.78	179.52
IFAST-LASSO	-88.65	179.41	180.08
Uni-SCAD	-88.42	177.87	179.53
Uni-LASSO	-88.68	179.44	180.11

It is easy to see from the Table 2.5 that the proposed NPGSJS procedure greatest log likelihood no matter SCAD or LASSO penalty is used. NPGSJS-SCAD also

obtained the smallest AIC and BIC. The SJS procedure performs the second best in the sense that the likelihood is greater than the other independent screening procedures . The reason is that considering the effects of the features jointly makes the selected model more effective. Benefit from the non-monotone proximal gradient algorithm, which is able to obtain better results numerically, the NPGSJS procedure outperforms the SJS procedure.

The selected gene IDs by different feature screening methods are report in Table 2.6. The Table 2.6 shows that the gene 1181 is selected by all the methods. Gene 1456 is selected by all the methods except SJS method, while genes 1826, 7069, and 7357 are failed to be selected by IFAST but are selected by the others. The reason might be that the single-index hazard model assumption by IFAST method is different from the Cox model assumption by the other methods. Thus the final selected features will be different numerically.

We are able to further investigate the selected features by the above methods, except the IFAST method, by making statistical inference based on the Cox model. Unlike the other methods which are based on the Cox model, the FAST(IFAST) method is based on a more flexible single-index hazards model. After the feature screening for such high- or ultrahigh- dimensional data setting by FAST, we still face the problem of making delicate statistical inference in low-dimensional single-index hazard models. To solve this problem, we will dive into details for the single-index hazards models in Chapter 3.

2.6 Assumptions and Proofs

Assumption 2.1. *There exist scalar, vector and matrix functions* $s^{(l)}(\beta, t)$ *defined on* $\mathcal{B}\times[0,\tau]$, $l = 0, 1, 2$, *that satisfy the following conditions: (i)* $\sup_{t\in[0,\tau],\boldsymbol{\beta}_1\in\mathcal{B}_1}\|S^{(l)}(\boldsymbol{\beta}_1, t)-s^{(l)}(\boldsymbol{\beta}_1, t)\|_2 \to 0$ *in probability as* $n \to \infty$ *for* $\mathcal{B}_1 \subset R^q$, $\mathcal{B}_1 \subset \mathcal{B}$; *(ii) The functions*

Table 2.6: Selected gene IDs by different feature screening methods

Uni	NPGSJS	SJS	IFAST	ISIS
1181	254	1072	197	1072
1456	454	1181	1188	1188
1662	1188	**1825**	**1456**	**1456**
1681	**1456**	3810	1681	**1825**
1825	1547	4547	1825	4131
3799	1681	5027	2107	5027
3810	**1825**	6507	2109	5043
4131	4966	6565	2240	5055
5055	5649	6701	2311	5301
5301	6607	6706	4131	6519
5614	6956	7018	4317	6706
5950	**7069**	**7069**	5054	6860
7069	7343	7307	5649	**7069**
7357	**7357**	**7357**	6519	**7357**

$s^{(l)}(\beta, t)$ *are bounded and* $s^{(0)}(\beta, t)$ *is bounded away from 0 on* $\mathcal{B} \times [0, \tau]$*; and the family of functions* $s^{(l)}(\cdot, t), 0 \leqslant t \leqslant \tau$*, is an equicontinuous family at* β^**.*

Assumption 2.2. *Let* $\bar{\mathbf{z}}(\boldsymbol{\beta}, t) = s^{(1)}(\boldsymbol{\beta}, t)/s^{(0)}(\boldsymbol{\beta}, t)$*. Denote* $d_n = \sup_{t\in[0,\tau]} \|\bar{\mathbf{Z}}(\boldsymbol{\beta}^*, t) - \bar{\mathbf{z}}(\boldsymbol{\beta}^*, t)\|_\infty$ *and* $e_n = \sup_{t\in[0,\tau]} \|S^{(0)}(\boldsymbol{\beta}^*, t) - s^{(0)}(\boldsymbol{\beta}^*, t)\|_\infty$*. The random sequences* d_n *and* e_n *are bounded almost surely.*

Assumption 2.3. *Define* $\varepsilon_{ij} = \int_0^\tau \{Z_{ij} - \bar{z}_j(\boldsymbol{\beta}^*, t)\} dM_i(t)$*, where* $\bar{z}_j(\boldsymbol{\beta}^*, t)$ *is the* j*th component of* $\bar{\mathbf{z}}(\boldsymbol{\beta}^*, t)$*. Suppose that the Cramér condition holds for* ε_{ij}*, i.e.,* $E|\varepsilon_{ij}|^l \leqslant 2^{-1} l! c_1^{l-2} \sigma_j^2$ *for all* j*, where* c_1 *is a positive constant,* $l \geqslant 2$*, and* $\sigma_j^2 = \mathrm{var}(\varepsilon_{ij}) < \infty$*.*

Assumption 2.4. *There exist positive constants* c_2, c_3, τ_1 *and* τ_2 *such that* $\min_{j\in M_0} |\beta_j^*| \geqslant c_2 n^{-\tau_1}$ *and* $q \leqslant k \leqslant c_3 n^{\tau_2}$*.*

Assumption 2.5. *When* n *is sufficiently large, for* $\boldsymbol{\beta}_M \in \{\boldsymbol{\beta}_M : \|\boldsymbol{\beta}_M - \boldsymbol{\beta}_M^*\|_2 \leqslant \delta\}$*,* $M \in \mathbf{M}_+^{2k}$*, it holds that* $\lambda_{\min}\{n^{-1} \int_0^\tau V(\boldsymbol{\beta}_M, t) d\bar{N}(t)\} \geqslant c_4$*, where* c_4 *and* δ *are positive constants depending on* k *but not* M*,* $\lambda_{\min}(A)$ *is the minimum eigenvalue of the matrix* A*, and* $V(\boldsymbol{\beta}_M, t)$ *is a version of* $V(\boldsymbol{\beta}, t)$ *based on the model* M*.*

Assumption 2.6. *There exists a positive constant* c_5 *such that, for sufficiently large* n,

$$\boldsymbol{\eta}^T \int_0^\tau V(\boldsymbol{\beta}, t) d\bar{N}(t) \boldsymbol{\eta} \geqslant c_5 n \|\boldsymbol{\eta}_{M_0}\|_2^2,$$

for any $\boldsymbol{\eta} \neq 0$, $\|\boldsymbol{\eta}_{M_0^c}\|_1 \leqslant 3\|\boldsymbol{\eta}_{M_0}\|_1$, *and* $\boldsymbol{\beta} \in \mathcal{B}$, *where* M_0^c *is the complement of* M_0 *in* $\{1, 2, \cdots, p\}$.

Assumption 2.7. $\sup_{(\boldsymbol{\beta},t)\in\mathcal{B}(\boldsymbol{\beta}^*,o(w))\times[0,\tau]} \|V(\boldsymbol{\beta}, t)\|_\infty = O_p(n^{\tau_3})$, *where* τ_3 *is a positive constant,* $w = \min_{j\in M_0} \|\beta_j^*\|$, *and* $\mathcal{B}(\boldsymbol{\beta}^*, o(w))$ *is a* p*-dimensional ball centered at* $\boldsymbol{\beta}^*$ *with radius* $o(w)$.

Assumptions 2.1 to 2.3 are mild, which are necessities to obtain a large deviation result, see Bradic et al. (2011) for more discussions of these assumptions. The first part of Assumption 2.4 states that the marginal signals are strong enough to be detected. Similar assumptions have been widely made in the ultrahigh-dimensional data analysis. In the second part, we allow the dimension of the important features to diverge to infinity at a polynomial speed. Assumption 2.5 is a very weak assumption since it usually holds that $n^{-1}\int_0^\tau V(\boldsymbol{\beta}_M, t)d\bar{N}(t)$ is positive definite when k is not too large by noting that the dimension of $\boldsymbol{\beta}_M$ is k. For example, if $k = 5$, $\boldsymbol{\beta}_M$ is a 5-dimensional vector parametric. Then $n^{-1}\int_0^\tau V(\boldsymbol{\beta}_M, t)d\bar{N}(t)$ is a 5×5 matrix, which is positive definite with high probability when the sample size is moderate.

Assumption 2.6 is needed for deriving an error bound for the LASSO estimator. There are many similar assumptions in the literature, such as Bickel et al. (2009) for the linear model, Xu and Chen (2014) for the generalized linear model, Huang et al. (2013) for the Cox model, and so on. As was discussed in Fleming and Harrington (2011), $V(\boldsymbol{\beta}, t)$ is an empirical covariance matrix of $\mathbf{Z}_i$ with the weights proportional to $Y_i(t)\exp\{\mathbf{Z}_i^T\boldsymbol{\beta}\}$. Hence, assumption 2.7 points out that the association of features should not be too strong. In particular, if all the features are mutually independent,

this assumption is easily met. Similar assumptions can be found in Bradic et al. (2011).

The proofs of the theorems can be obtained along the line of Xu and Chen (2014) by combining the large deviation result for martingales of Bradic et al. (2011) under our specific assumptions for the Cox model.

Proof of Theorem 2.1 Firstly, we prove the monotonicity. It is easy to see that

$$\begin{aligned}
& l_n(\boldsymbol{\beta}^{(t)}) \\
= \; & Q_n(\boldsymbol{\beta}^{(t)}|\boldsymbol{\beta}^{(t)}) \\
\leqslant \; & Q_n(\boldsymbol{\beta}^{(t+1)}|\boldsymbol{\beta}^{(t)}) \\
= \; & l_n(\boldsymbol{\beta}^{(t)}) + (\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)})^T \dot{l}_n(\boldsymbol{\beta}^{(t)}) - \frac{u}{2}\|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}\|_2^2 \\
= \; & l_n(\boldsymbol{\beta}^{(t+1)}) - \frac{u}{2}\|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}\|_2^2 + (\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)})^T \dot{l}_n(\boldsymbol{\beta}^{(t)}) \\
& + l_n(\boldsymbol{\beta}^{(t)}) - l_n(\boldsymbol{\beta}^{(t+1)}) \\
= \; & l_n(\boldsymbol{\beta}^{(t+1)}) - \frac{u}{2}\|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}\|_2^2 + \sum_{i=1}^{n} \delta_i \Big[\log(nS^{(0)}(\boldsymbol{\beta}^{(t+1)}, X_i)) \\
& - \log(nS^{(0)}(\boldsymbol{\beta}^{(t)}, X_i)) - (\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)})^{\mathrm{T}} \bar{\mathbf{Z}}(\boldsymbol{\beta}^{(t)}, X_i) \Big]
\end{aligned}$$

By the Taylor's expansion of $\sum_{i=1}^{n} \delta_i \log(nS^{(0)}(\boldsymbol{\beta}^{(t+1)}, X_i))$ at $\boldsymbol{\beta}^{(t)}$ and some algebraic manipulations, we have

$$\begin{aligned}
& l_n(\boldsymbol{\beta}^{(t)}) \\
\leqslant \; & l_n(\boldsymbol{\beta}^{(t+1)}) - \frac{u}{2}\|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}\|_2^2 + \\
& \frac{1}{2}(\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)})^T \sum_{i=1}^{n} \delta_i \Big(\frac{S^{(2)}(\boldsymbol{\beta}, X_i)}{S^{(0)}(\boldsymbol{\beta}, X_i)} - \big[\frac{S^{(1)}(\boldsymbol{\beta}, X_i)}{S^{(0)}(\boldsymbol{\beta}, X_i)}\big]^{\otimes 2} \Big) \Big|_{\boldsymbol{\beta}=\bar{\boldsymbol{\beta}}} (\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}) \\
= \; & l_n(\boldsymbol{\beta}^{(t+1)}) - \frac{u}{2}\|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}\|_2^2 + \frac{1}{2}(\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)})^T \int_0^{\tau} V(\bar{\boldsymbol{\beta}}, t) d\bar{N}(t) (\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)})
\end{aligned}$$

$$\leqslant \quad l_n(\boldsymbol{\beta}^{(t+1)}) - \frac{u}{2}\|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}\|_2^2 + \frac{1}{2}\lambda_{\max}\Big(\int_0^\tau V(\bar{\boldsymbol{\beta}}, t)d\bar{N}(t)\Big)\|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}\|_2^2. \quad (2.7)$$

So under the assumptions in Theorem 2.1, it finally arrives at

$$\begin{aligned} & l_n(\boldsymbol{\beta}^{(t)}) \\ \leqslant \quad & l_n(\boldsymbol{\beta}^{(t+1)}) - \frac{1}{2}(u - \rho^{(t)})\|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}\|_2^2 \\ \leqslant \quad & l_n(\boldsymbol{\beta}^{(t+1)}). \end{aligned}$$

Secondly, we will show that $\{\boldsymbol{\beta}^{(t)}\}$ converges to a local maximum of $l_n(\boldsymbol{\beta})$. It is noted that $l_n(\cdot)$ is bounded with $\boldsymbol{\beta}$ being confined in $\mathcal{B}$. From the proof of the first part, we can see

$$l_n(\boldsymbol{\beta}^{(t+1)}) - l_n(\boldsymbol{\beta}^{(t)}) \geqslant \frac{1}{2}(u - \rho)\|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}\|_2^2.$$

By the monotonicity and boundness of $l_n(\cdot)$, $\|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}\|_2 \to 0$ as t goes to ∞.

It is noted that k and p are constants in the proof for convergence of $\{\boldsymbol{\beta}^{(t)}\}$. For the finite sparse patterns of $\{\boldsymbol{\beta}^{(t)}\}$, we can always find a subsequence of $\{\boldsymbol{\beta}^{(t)}\}$, say $\{\boldsymbol{\beta}^{(t_m)}\}$, with a common sparse pattern M. Because $\int_0^\tau V(\boldsymbol{\beta}_M, t)d\bar{N}(t)$ is positive definite for any $\boldsymbol{\beta}_M$, $l_n(\cdot)$ is strictly concave for $\boldsymbol{\beta}_M$ with sparse pattern M. Then $\{\boldsymbol{\beta}^{(t_m)}\}$ is bounded and has at least one limiting point, denoted by $\boldsymbol{\beta}^\dagger = (\beta_1^\dagger, \beta_2^\dagger, \cdots, \beta_p^\dagger)^T$. Based on the facts that $\boldsymbol{\beta}^{(t_m+1)} = \text{argmax}_{\boldsymbol{\beta}\in\mathcal{B}(k)}\{Q_n(\boldsymbol{\beta}|\boldsymbol{\beta}^{(t_m)})\}$ and $\|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}\|_2 \to 0$, we have $\boldsymbol{\beta}^\dagger = \text{argmax}_{\boldsymbol{\beta}\in\mathcal{B}(k)}\{Q_n(\boldsymbol{\beta}|\boldsymbol{\beta}^\dagger)\}$ This indicates that $\boldsymbol{\beta}^\dagger$ maximizes $Q_n(\boldsymbol{\beta}|\boldsymbol{\beta}^\dagger)$ with respect to $\boldsymbol{\beta}$ and sparse pattern M.

If $\|\boldsymbol{\beta}^\dagger\|_0 < k$, i.e., $\boldsymbol{\beta}^\dagger$ has fewer than k non-zero entries, it is easy to see that $\dot{l}_n(\boldsymbol{\beta}^\dagger) = 0$. So $\boldsymbol{\beta}^\dagger$ is the unconstrained maximizer of $l_n(\boldsymbol{\beta})$ and satisfies $\|\boldsymbol{\beta}^\dagger\|_0 \leqslant k$.

If $\|\boldsymbol{\beta}^\dagger\|_0 = k$, we will prove that $\boldsymbol{\beta}^\dagger$ is the unique maximizer for the sparse pattern M. Note that

$$l_n(\boldsymbol{\beta}) = Q_n(\boldsymbol{\beta}|\boldsymbol{\beta}^\dagger) - R(\boldsymbol{\beta}|\boldsymbol{\beta}^\dagger),$$

where

$$R(\boldsymbol{\beta}|\boldsymbol{\beta}^\dagger) = -\frac{u}{2}\|\boldsymbol{\beta}-\boldsymbol{\beta}^\dagger\|_2^2 + (\boldsymbol{\beta}-\boldsymbol{\beta}^\dagger)^T\dot{l}_n(\boldsymbol{\beta}^\dagger) + l_n(\boldsymbol{\beta}^\dagger) - l_n(\boldsymbol{\beta}).$$

So it is easy to see that $\frac{\partial R(\boldsymbol{\beta}|\boldsymbol{\beta}^\dagger)}{\partial \boldsymbol{\beta}}|_{\boldsymbol{\beta}=\boldsymbol{\beta}^\dagger} = 0$. In addition, $\frac{\partial Q(\boldsymbol{\beta}|\boldsymbol{\beta}^\dagger)}{\partial \boldsymbol{\beta}_j}|_{\boldsymbol{\beta}=\boldsymbol{\beta}^\dagger} = 0$, because $\boldsymbol{\beta}^\dagger$ is the local maximizer for the sparse pattern M. Then we have $\frac{\partial l_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_j}|_{\boldsymbol{\beta}=\boldsymbol{\beta}^\dagger} = 0$, which implies $\boldsymbol{\beta}^\dagger$ is the unique maximizer of $l_n(\cdot)$ by the property of strict convexity.

Above all, any limiting point of $\{\boldsymbol{\beta}^{(t)}\}$ is a local maximizer satisfying $\|\boldsymbol{\beta}\|_0 \leqslant k$. By the finiteness of sparse patterns, there are at most finite many limiting points. Furthermore, similarly to Xu and Chen (2014), by the techniques of mathematical analysis, it can be shown that $\{\boldsymbol{\beta}^{(t)}\}$ has only one limiting point and thus converges.

Lemma 2.1. *Define* $\boldsymbol{\beta}^{(0)} = \text{argmax}_{\boldsymbol{\beta}}\{l_n(\boldsymbol{\beta}) - n\lambda\|\boldsymbol{\beta}\|_1\}$, *where* λ *satisfies* $\lambda n^{\frac{1}{2}-m} \to \infty$, $\lambda n^{\tau_1+\tau_2} \to 0$. *Under Assumptions 2.1 to 2.3 and 2.6, if* $\max_j \sigma_j^2 = O(\lambda n^{\frac{1}{2}})$, *we have*

$$\text{pr}(\|\boldsymbol{\beta}^{(0)} - \boldsymbol{\beta}^*\|_1 \leqslant 16c_5^{-1}\lambda q) \to 1,$$

where c_5 *is defined in Assumption 2.6.*

Proof. It is easy to see that

$$l_n(\boldsymbol{\beta}^{(0)}) - n\lambda\|\boldsymbol{\beta}^{(0)}\|_1 - (l_n(\boldsymbol{\beta}^*) - n\lambda\|\boldsymbol{\beta}^*\|_1) \geqslant 0,$$

or equivalently

$$l_n(\boldsymbol{\beta}^*) - l_n(\boldsymbol{\beta}^{(0)}) \leqslant n\lambda\|\boldsymbol{\beta}^*\|_1 - n\lambda\|\boldsymbol{\beta}^{(0)}\|_1.$$

Define $\boldsymbol{\delta} = (\boldsymbol{\beta}^{(0)} - \boldsymbol{\beta}^*) = (\delta_1, \cdots, \delta_p)^T$. By the Taylor's expansion of $l_n(\boldsymbol{\beta}^{(0)})$ at $\boldsymbol{\beta}^*$, we have

$$l_n(\boldsymbol{\beta}^{(0)}) - l_n(\boldsymbol{\beta}^*) = \boldsymbol{\delta}^T\dot{l}_n(\boldsymbol{\beta}^*) - \frac{1}{2}\boldsymbol{\delta}^T\int_0^\tau V(\boldsymbol{\beta}, t)d\bar{N}(t)\boldsymbol{\delta},$$

where $\boldsymbol{\beta}$ is between $\boldsymbol{\beta}_L$ and $\boldsymbol{\beta}^*$. So

$$\frac{1}{n}\boldsymbol{\delta}^T\int_0^\tau V(\boldsymbol{\beta}, t)d\bar{N}(t)\boldsymbol{\delta}$$

$$
\begin{aligned}
&= \frac{1}{n}\big[2\boldsymbol{\delta}^T \dot{l}_n(\boldsymbol{\beta}^*) + 2l_n(\boldsymbol{\beta}^*) - 2l_n(\boldsymbol{\beta}^{(0)})\big] \\
&\leqslant \frac{2}{n}\boldsymbol{\delta}^T \dot{l}_n(\boldsymbol{\beta}^*) + 2\lambda\|\boldsymbol{\beta}^*\|_1 - 2\lambda\|\boldsymbol{\beta}^{(0)}\|_1 \\
&\leqslant \frac{2}{n}|\boldsymbol{\delta}|^T|\dot{l}_n(\boldsymbol{\beta}^*)| + 2\lambda\|\boldsymbol{\beta}^*\|_1 - 2\lambda\|\boldsymbol{\beta}^{(0)}\|_1.
\end{aligned}
$$

Denote $\mathcal{A} = \{\max_{1\leqslant j\leqslant p}|\dot{l}_{nj}(\boldsymbol{\beta}^*)| \leqslant \frac{n\lambda}{2}\}$.

Because $\max_j \sigma_j^2 = O(\lambda n^{\frac{1}{2}})$, together with Assumptions 2.1 to 2.3, the result of the bound of tail probability in Theorem 3.1 in Bradic et al. (2011) can be applied here via substituting their ξ_j by $\dot{l}_{nj}(\boldsymbol{\beta}^*)$, that is, there exist positive constants c_7 and c_8 such that $\mathrm{pr}(|\dot{l}_{nj}(\boldsymbol{\beta}^*)| > \sqrt{n}u_n) \leqslant c_7\exp(-c_8u_n)$. Then we have

$$
\begin{aligned}
\mathrm{pr}(\mathcal{A}^c) &\leqslant \sum_{j=1}^{p}\mathrm{pr}(|\dot{l}_{nj}(\boldsymbol{\beta}^*)| > \frac{n\lambda}{2}) = \sum_{j=1}^{p}\mathrm{pr}(|\dot{l}_{nj}(\boldsymbol{\beta}^*)| > \sqrt{n}\frac{\sqrt{n}\lambda}{2}) \\
&\leqslant pc_7\exp(-c_8\frac{\sqrt{n}\lambda}{2}) \leqslant c_7\exp(c_{10}n^m - c_8\frac{\sqrt{n}\lambda}{2}) \to 0,
\end{aligned}
$$

where c_{10} is a positive constant. So we can conclude that $\mathrm{pr}(\mathcal{A}) \to 1$ and $\|\dot{l}_n(\boldsymbol{\beta}^*)\|_\infty = O_p(n\lambda)$. Under the event $\mathcal{A}$, it is easy to see that

$$
\frac{1}{n}\boldsymbol{\delta}^T\int_0^\tau V(\boldsymbol{\beta},t)d\bar{N}(t)\boldsymbol{\delta} \leqslant \lambda\|\boldsymbol{\delta}\|_1 + 2\lambda\|\boldsymbol{\beta}^*\|_1 - 2\lambda\|\boldsymbol{\beta}^{(0)}\|_1.
$$

So

$$
\begin{aligned}
&\frac{1}{n}\boldsymbol{\delta}^T\int_0^\tau V(\boldsymbol{\beta},t)d\bar{N}(t)\boldsymbol{\delta} + \lambda\|\boldsymbol{\delta}\|_1 \\
\leqslant\ & 2\lambda\|\boldsymbol{\delta}\|_1 + 2\lambda\|\boldsymbol{\beta}^*\|_1 - 2\lambda\|\boldsymbol{\beta}^{(0)}\|_1 \\
\leqslant\ & 2\lambda\sum_{j=1}^{p}\big(|\beta_j^{(0)} - \beta_j^*| + |\beta_j^*| - |\beta_j^{(0)}|\big) \\
=\ & 2\lambda\sum_{j\in M_0}\big(|\beta_j^{(0)} - \beta_j^*| + |\beta_j^*| - |\beta_j^{(0)}|\big)
\end{aligned}
$$

$$\begin{aligned} &\leqslant 4\lambda \sum_{j\in M_0} |\delta_j| \\ &\leqslant 4\lambda \|\boldsymbol{\delta}_{M_0}\|_1. \end{aligned}$$

It is easy to see that $\int_0^\tau V(\boldsymbol{\beta}, t)d\bar{N}(t)$ is semipositive definite. Thus $\|\boldsymbol{\delta}\|_1 \leqslant 4\|\boldsymbol{\delta}_{M_0}\|_1$, and furthermore $\|\boldsymbol{\delta}_{M_0^c}\|_1 \leqslant 3\|\boldsymbol{\delta}_{M_0}\|_1$. By the Cauchy-Schwarz inequality and Assumption 2.6,

$$\|\boldsymbol{\delta}_{M_0}\|_1^2 \leqslant q\|\boldsymbol{\delta}_{M_0}\|_2^2 \leqslant qc_5^{-1}n^{-1}\Big[\int_0^\tau V(\boldsymbol{\beta}, t)d\bar{N}(t)\Big] \leqslant 4c_5^{-1}\lambda q\|\boldsymbol{\delta}_{M_0}\|_1.$$

So $\|\boldsymbol{\delta}_{M_0}\|_1 \leqslant 4c_5^{-1}\lambda q$. Then finally we arrive at

$$\|\boldsymbol{\delta}\|_1 = \|\boldsymbol{\delta}_{M_0^c}\|_1 + \|\boldsymbol{\delta}_{M_0}\|_1 \leqslant 4\|\boldsymbol{\delta}_{M_0}\|_1 \leqslant 16c_5^{-1}\lambda q.$$

This finishes the proof.

Proof of Theorem 2.2 Recall that $w = \min_{j\in M_0}\|\beta_j^*\|$. We just need to show $\mathrm{pr}(\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|_\infty < \frac{w}{2}) \to 1$. It suffices to prove $\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|_\infty = o_p(w)$. As in Xu and Chen (2014), we use the method of mathematical induction to get this result.

When $t = 0$, by Lemma 1, we have

$$\mathrm{pr}(\|\boldsymbol{\beta}^{(0)} - \boldsymbol{\beta}^*\|_1 \leqslant 16c_5^{-1}\lambda q) \to 1.$$

Because $\lambda = o(n^{-(\tau_1+\tau_2)})$, $q = O(n^{\tau_2})$, $w^{-1} = O(n^{\tau_1})$, $\lambda q w^{-1} = o(n^{-(\tau_1+\tau_2)})O(n^{\tau_2})O(n^{\tau_1}) = o(1)$. Thus $\lambda q = o(w)$. So we have $\|\boldsymbol{\beta}^{(0)} - \boldsymbol{\beta}^*\|_1 = o_p(w)$. It is noted $\|\boldsymbol{\beta}^{(0)} - \boldsymbol{\beta}^*\|_\infty \leqslant \|\boldsymbol{\beta}^{(0)} - \boldsymbol{\beta}^*\|_1$. Then the desired result is obtained for $t = 0$.

Suppose that $\|\boldsymbol{\beta}^{(t-1)} - \boldsymbol{\beta}^*\|_\infty = o_p(w)$. In the following, we will show that $\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|_\infty = o_p(w)$ is also true. It is noted that $\boldsymbol{\beta}^{(t)} = \mathbf{H}(\tilde{\boldsymbol{\beta}}; k)$, where $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}^{(t-1)} + u^{-1}\dot{l}_n(\boldsymbol{\beta}^{(t-1)})$. If $\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_\infty = o_p(w)$ holds, it can be seen that elements of $\tilde{\boldsymbol{\beta}}_{M_0}$ are among the ones with top k largest absolute values in probability. Thus $\|\boldsymbol{\beta}^{(t-1)} - \boldsymbol{\beta}^*\|_\infty \leqslant \|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_\infty = o_p(w)$. So what remains is to prove $\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_\infty = o_p(w)$. Note

that $\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_\infty \leqslant \|\boldsymbol{\beta}^{(t-1)} - \boldsymbol{\beta}^*\|_\infty + \frac{1}{u}\|\dot{l}_n(\boldsymbol{\beta}^{(t-1)})\|_\infty$. By the Taylor's expansion of $\dot{l}_n(\boldsymbol{\beta}^{(t-1)})$ at β^*, we have

$$\begin{aligned}
& \|\dot{l}_n(\boldsymbol{\beta}^{(t-1)})\|_\infty \\
= \ & \|\dot{l}_n(\boldsymbol{\beta}^*) - \int_0^\tau V(\bar{\boldsymbol{\beta}}, t)d\bar{N}(t)(\boldsymbol{\beta}^{(t-1)} - \boldsymbol{\beta}^*)\|_\infty \\
\leqslant \ & \|\dot{l}_n(\boldsymbol{\beta}^*)\|_\infty + \| \int_0^\tau V(\bar{\boldsymbol{\beta}}, t)d\bar{N}(t)(\boldsymbol{\beta}^{(t-1)} - \boldsymbol{\beta}^*)\|_\infty \\
\leqslant \ & \|\dot{l}_n(\boldsymbol{\beta}^*)\|_\infty + n\|\frac{1}{n}\int_0^\tau V(\bar{\boldsymbol{\beta}}, t)d\bar{N}(t)\|_\infty\|(\boldsymbol{\beta}^{(t-1)} - \boldsymbol{\beta}^*)\|_\infty \\
\leqslant \ & \|\dot{l}_n(\boldsymbol{\beta}^*)\|_\infty + n\|(\boldsymbol{\beta}^{(t-1)} - \boldsymbol{\beta}^*)\|_\infty \sup_{(\boldsymbol{\beta},t)\in\mathcal{B}(\boldsymbol{\beta}^*,o(w))\times[0,\tau]}\|V(\boldsymbol{\beta}, t)\|_\infty,
\end{aligned}$$

where $\bar{\boldsymbol{\beta}}$ is between $\boldsymbol{\beta}^{(t-1)}$ and $\boldsymbol{\beta}^*$. So

$$\begin{aligned}
& \frac{1}{u}\|\dot{l}_n(\boldsymbol{\beta}^{(t-1)})\|_\infty \\
= \ & \frac{1}{u}O_p(n\lambda) + \frac{n}{u}o_p(w)O_p(n^{\tau_3}) \\
\leqslant \ & (c_6 rn)^{-1}n\lambda O_p(1) + (c_6 rn)^{-1}n^{1+\tau_3}o_p(w)O_p(1) \\
= \ & c_6^{-1}O(n^{-\tau_3})o(n^{-(\tau_1+\tau_2)})O_p(1) + c_6^{-1}O_p(n^{-\tau_3})n^{\tau_3}o_p(w) \\
= \ & o_p(w).
\end{aligned}$$

This ends up the proof.

Assumptions verification for optimization problem (2.6) We aim to verify that the optimization problem (2.6) satisfies the assumptions in A.1 of Chen et al. (2016) (page 1485-1486).

Let $\mathrm{Dom}(P) = \mathcal{B}(k)$. With a little abuse of definition, we redefine

$$-Q_n(\boldsymbol{\beta}|\boldsymbol{\beta}^{(t)}) = \begin{cases} -l_n(\boldsymbol{\beta}^{(t)}) - (\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)})^T\dot{l}_n(\boldsymbol{\beta}^{(t)}) + \frac{u}{2}\|\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)}\|_2^2 & \beta \in \mathrm{Dom}(P); \\ M^* & \beta \notin \mathrm{Dom}(P), \end{cases}$$

where M^* is a constant and larger than the maximum of $-l_n(\boldsymbol{\beta}^{(t)}) - (\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)})^T\dot{l}_n(\boldsymbol{\beta}^{(t)}) + \frac{u}{2}\|\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)}\|_2^2$ over $\mathrm{Dom}(P)$. Because $\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)}$ and $\|\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)}\|_2$ are continuous and

$\mathrm{Dom}(P)$ is a compact set, they are bounded over $\mathrm{Dom}(P)$. Thus M^* exists. This new definition will not affect the computation of our proposed algorithm.

(i) It is easy to be satisfied since $-Q_n(\boldsymbol{\beta}|\boldsymbol{\beta}^{(t)})$ is continuously differentiable and its derivative is bounded in a bounded closed set.

(ii) For any $\boldsymbol{\beta}^\dagger \in \mathbb{R}^p$, if $\boldsymbol{\beta}^\dagger \in \mathcal{B}(k)$, then $P(\boldsymbol{\beta}^\dagger) = 0$. Since $\liminf_{\boldsymbol{\beta}\to\boldsymbol{\beta}^\dagger} P(\boldsymbol{\beta}) \geqslant 0$, $P(\boldsymbol{\beta}^\dagger) \leqslant \liminf_{\boldsymbol{\beta}\to\boldsymbol{\beta}^\dagger} P(\boldsymbol{\beta})$. If $\boldsymbol{\beta}^\dagger \notin \mathcal{B}(k)$, then $P(\boldsymbol{\beta}^\dagger) = \infty$. The complement of $\mathrm{Dom}(P)$ is an open set, therefore, $\liminf_{\boldsymbol{\beta}\to\boldsymbol{\beta}^\dagger} P(\boldsymbol{\beta}) = \infty$. Then $P(\boldsymbol{\beta}^\dagger) \leqslant \liminf_{\boldsymbol{\beta}\to\boldsymbol{\beta}^\dagger} P(\boldsymbol{\beta})$. Hence P is a lower semicontinuous function in $\mathbb{R}^p$.

(iii) For $\boldsymbol{\beta}^\dagger \in \mathrm{Dom}(P)$, $\Omega(\boldsymbol{\beta}^\dagger) = \{\boldsymbol{\beta} \in \mathbb{R}^p : -Q_n(\boldsymbol{\beta}|\boldsymbol{\beta}^{(t)}) \leqslant Q_n(\boldsymbol{\beta}^\dagger|\boldsymbol{\beta}^{(t)})\} \subseteq \mathrm{Dom}(P)$. Then $-Q_n(\boldsymbol{\beta}|\boldsymbol{\beta}^{(t)})$ is bounded below in $\Omega(\boldsymbol{\beta}^\dagger)$.

(iv) Let $A = \sup_{\beta\in\Omega(\boldsymbol{\beta}^\dagger)} \|\nabla(-Q_n(\boldsymbol{\beta}|\boldsymbol{\beta}^{(t)}))\|$, $B = \sup_{\beta\in\Omega(\boldsymbol{\beta}^\dagger)} P(\beta)$ and $C = \inf_{\beta\in R^p} P(\beta)$. Because $\Omega(\boldsymbol{\beta}^\dagger) \subseteq \mathrm{Dom}(P)$ and $\nabla(-Q_n(\boldsymbol{\beta}|\boldsymbol{\beta}^{(t)}))$ is continuous, $A < \infty$. Furthermore, $B = \sup_{\beta\in\Omega(\boldsymbol{\beta}^\dagger)} P(\beta) \leqslant \sup_{\beta\in\mathrm{Dom}(P)} P(\beta) = 0 < \infty$. In addition, $C = 0 < \infty$.

2.7 coxnpgsjs: an R package

An R package called coxnpgsjs for sure joint feature screening for ultrahigh-dimensional Cox's model using a non-monotone proximal gradient algorithm is built. For even faster implementation, the core function is rewritten by C++. The package at https://github.com/iantsuising/coxnpgsjs. One can install it to the local library by R command "devtools::install_github('iantsuising/coxnpgsjs')".

Chapter 3

Estimation under Single-Index Hazard Model

3.1 Introduction

In contemporary GWAS studies, high- or ultrahigh- dimensional data with survival outcomes are very common. For example, the aforementioned DLBCL data set. To deal with such data, feature screening or variable selection techniques are often applied at the first stage in practice. Based on the Cox model assumption, a number of feature screening and variable selection methods are well studied. See Fan et al. (2010), Benner et al. (2010), among others. However, a specific functional form carries on a strong model assumption that maybe mostly for mathematical convenience. Such an assumption tends to be violated because many gene expressions may exhibit a more complicated effect than the log-linear effect. Based on a more flexible single-index hazards (SIH) model, Gorst-Rasmussen and Scheike (2013) proposed the feature aberration at survival times (FAST) statistic for feature screening. However, after such exercise for high or ultra high dimensional data setting, we still face the problem of making a delicate statistical inference in low-dimensional SIH models. This motivates us to establish effective and even efficient estimation procedures for the SIH model.

In this chapter, to study the right-censored survival outcomes with covariate variables, we consider the following semiparametric SIH model

$$\lambda\{t \mid \mathbf{Z}(t)\} = \lambda\{t, \beta^T \mathbf{Z}(t)\}, \tag{3.1}$$

where $\lambda(\cdot, \cdot)$ is an unspecified and positive bivariate hazards function, $\mathbf{Z}(t)$ is a p-dimensional vector of possibly time-dependent covariates, and β is an unknown regression coefficient vector.

The SIH model (3.1) is a preferable exploratory tool allowing nonparametric modeling of covariate effects in a parsimonious way via a single index. The index structure helps achieve dimension reduction. The unspecified functional form makes the model more robust in theory and more acceptable in applications when empirically the functional form is uncertain. Model (3.1) is flexible encompassing various existing intensity models. See Cox's proportional hazards model (Cox, 1972), additive hazards model (Lin and Ying, 1994), accelerated failure time model (Cox and Oakes, 1984), link-unknown proportional hazards model (Wang, 2004), accelerated hazards model (Chen and Wang, 2000), transformed hazards model (Zeng et al., 2005), and among others.

However, the index structure, the unknown form of bivariate function, and incomplete failure time observations together present great challenges for statistical inference in model (3.1). The index structure impacts the support of the infinite-dimensional "nuisance" nonparametric function $\lambda(\cdot, \cdot)$ in the SIH model (3.1). Consequently, it makes the partial derivative and the operation of convergence in probability not exchangeable in the hazard rate function $\lambda(\cdot, \cdot)$, as shown in our proof of Proposition 3.3. In other words, to estimate $\lambda(\cdot, \cdot)$ well means to "recover" the support of this bivariate function, to the extend possible. As evidence, Ding et al. (2013) validated that the estimated parametric component in model (3.1) can hardly keep consistency unless strong conditions are imposed when the nonparametric compo-

nent is estimated by the traditional univariate nonparametric likelihood estimation. To establish effective and even efficient estimation procedures for the vector of the index coefficient, we need to develop a new methodology to tackle the challenge in the inference procedure for model (3.1).

To deal with semiparametric models involving an index structure, there is vast literature. To handle the estimation of the parametric component, two-stage methods are quite popular such as generalized estimation equations and (local) profile likelihood. To tact estimation of the nonparametric component, there are ways like Hermite polynomial system, B-splines, local constant, backfitting iteration, weighted least squares. See Dong et al. (2015), Ma and Song (2015), Linton et al. (2003), Carroll et al. (1997), Ichimura (1993), and among others; However, notice that most work is under a mean regression structure, say within a generalized linear model framework or with explicit error terms. Also the data setting is assumed completely observed and no censoring. Kernel smoothing or local linear expansion of Fan and Gijbels (1996) is a common tool applied in the estimation of a univariate nonlinear function. See Xia (2006), Liang et al. (2010) and Liu et al. (2013); among others. In presence of random censorship, local linear expansion was also applied successfully to estimate a univariate nonlinear function withing a hazard regression model. See (Fan et al. (1997), Cai et al. (2008), Yin et al. (2008) and Lu and Zhang (2010)). Nevertheless, such smoothing techniques cannot be applied directly in model (3.1) since it lacks error terms and the hazard rate function is bivariate in essence.

Right in the form of model (3.1), there is relatively less work. Nielsen (1998) studied a nonparametric hazard estimation by local linear regression, but it could not extend to multivariate- and even bivariate- covariate situations due to the curse of dimensionality. The average derivatives method by Gørgens (2004, 2006) could not make the parametric component efficient, and so did the pseudo-integrated least squares estimation procedure by Chiang et al. (2017). Xia et al. (2002) extended

a dimension reduction technique Mave method to handle the hazard rate function. However, the resulting hMave estimator of β in model (3.1) can avoid estimation of variance and did not aim to achieve the information lower bound, thus not semiparametric efficient. Furthermore, they imposed strong restrictions on the state vector requiring that all the covariates are continuous. This is not realistic and the method may experience large variability when some covariates are discrete.

All in all, efficient estimation procedure in model (3.1) is essential but non-trivial (Newey (1990)). In our methodology, we carry out inference on index coefficients and other aspects relevant to the nonparametric component by first profiling the log-likelihood with respect to index coefficients. Rather than applying univariate local linear regression directly, to estimate the unknown functional form, we develop a new local linear kernel weighted least square estimator which naturally handles the random censorship and the bivariate local linear approximation jointly. This strategy proves to be effective because the estimator of the essentially bivariate hazard function is uniformly consistent. The consistency of $\lambda(\cdot, \cdot)$ ensures the consistency of the estimated index coefficients under quite mild conditions. Finally, the profiled likelihood estimator of the index coefficient vector achieves the information lower bound. In addition, the proof procedure inspires us to construct an efficient influential function. It leads to a class of estimation equations, the solution of which is efficient. Based on the doubly robust property, we also give another two sets of estimation equations of which the solutions are asymptotically equally efficient. The main concern is for computational feasibility in the sense that these two sets of estimations equations can escape away from either estimation of the nonlinear hazard rate function or estimation of its partial derivative.

The remainder of this chapter is organized as follows. In section 3.2, we introduce the semiparametric efficient inference procedure for β under the SIH model (3.1). That is, we derive the semiparametric efficient score and present that pro-

posed consistent estimators achieve the semiparametric efficient bound. In Section 3.3, we construct a class of efficient estimation equations and another two sets of estimation equations by using the components orthogonal to the nuisance tangent space and their doubly robust property. We introduce the adapted Newton-Raphson algorithm for computing β in Section 3.4. The estimation method and the asymptotic result for the nonparametric part of the model (3.1) are presented in Section 3.5. Simulation results and real data analysis are reported and discussed in Section 3.6 and 3.7, respectively. Section 3.8 concludes. Proofs of propositions and theorems are strictly derived in Section 3.9. All lemmas presented in Section 3.9 are strictly proved in Section 3.10.

3.2 Semiparametric Efficient Inference on β

Suppose a default occurs with random time T, and is censored by random time C impacted by credit risky state vector $\mathbf{Z}$. We then have a random sample of size n with the i-th observation $(X_i = \min(T_i, C_i), \delta_i = I(T_i \leqslant C_i), \mathbf{Z}_i(t))$ on a time domain $t \in [0, \tau]$. Here δ is called censoring indicator and $I(\cdot)$ is an indicator function. The censoring is noninformative in the sense that T_i and C_i are independent given state vector $\mathbf{Z}_i(t)$ and its observed history on $[0, \tau]$. Let $Y_i(t) = I(X_i \geqslant t)$ be an at-risk process. Based on model (3.1), the log-likelihood of observed data $\{X_i, \delta_i, \mathbf{Z}_i(t), t \in [0, \tau]\}_{i=1}^n$ can be written as, up to the constant n^{-1},

$$\ell_n(\beta) = \sum_{i=1}^{n} \Big[\delta_i \log \lambda\{X_i, \beta^T \mathbf{Z}_i(X_i)\} - \int_0^\tau \lambda\{s, \beta^T \mathbf{Z}_i(s)\} Y_i(s) ds \Big],$$

which is a function of the unknown index coefficient vector β and the unknown hazard function λ. Inference on β and other aspects relevant to the nonparametric component is usualy carried out by first profiling the log-likelihood $\ell_n(\beta)$ w.r.t. β. That is, once $\lambda(\cdot, \cdot)$ is known, the estimator of β can be obtained by maximizing the

log-likelihood function.

3.2.1 Randomly Censored Bivariate Local Linear Regression Estimation

Let $N(t) = I(X \leqslant t, \delta = 1)$ be the counting process and the filtration for the i-th observation be $\mathcal{F}_{t,i} = \sigma\{N_i(u), \mathbf{Z}_i(u), Y_i(u+), 0 \leqslant u \leqslant t\}$ with union σ-field $\mathcal{F}_t = \cup_{i=1}^n \mathcal{F}_{t,i}$. An insight on derivative of the classical Doob-Meyer decomposition $M_i(s) = N_i(s) - \int_0^s Y_i(t)\lambda_t\{\mathbf{Z}_i(t)\}dt$ within the counting process framework yields an equivalence of model (3.1)

$$E[Y(s)dN(s)|\mathcal{F}_{s-}] = Y(s)\lambda\{s, \beta^T\mathbf{Z}(s)\}ds, \quad s \in [0, \tau]. \tag{3.2}$$

The equation (3.2) is a "nominal" traditional nonparametric regression as if $Y(s)dN(s)$ is the response. This naturally motivates us to construct a kernel weighted least square estimator for the bivariate $\lambda(\cdot, \cdot)$ based on its local linear approximation

$$\begin{aligned} \lambda\{s, \beta^T\mathbf{Z}(s)\} &\approx \lambda(t,u) + \frac{\partial\lambda(t,u)}{\partial t}(s-t) + \frac{\partial\lambda(t,u)}{\partial u}\{\beta^T\mathbf{Z}(s) - u\} \\ &\triangleq \alpha_0 + \alpha_{10}(s-t) + \alpha_{01}\{\beta^T\mathbf{Z}(s) - u\}, \end{aligned} \tag{3.3}$$

for $(s, \beta^T\mathbf{Z}(s))^T$ in a neighborhood of $(t,u)^T$, where $\alpha_0 = \lambda(t,u)$, $\alpha_{10} = \partial\lambda(t,u)/\partial t \equiv \lambda_{10}(t,u)$ and $\alpha_{01} = \partial\lambda(t,u)/\partial u \equiv \lambda_{01}(t,u)$. The corresponding objective function is

$$\sum_{i=1}^n \Big[Y_i(s)\Delta N_i(s) - Y_i(s)[\alpha_0 + \alpha_{10}(s-t) + \alpha_{01}\{\beta^T\mathbf{Z}_i(s) - u\}]\Big]^2 K_{\mathbf{h}}\{s-t, \beta^T\mathbf{Z}_i(s) - u\}ds,$$

for any $s \in [0, \tau]$, where $\Delta N_i(s) = N_i(s+\Delta s) - N_i(s-)$ and $K_{\mathbf{h}}(\cdot,\cdot) = h_1^{-1}h_2^{-1}K(\cdot/h_1, \cdot/h_2)$ with $K(\cdot,\cdot)$ being a bivariate kernel function and bandwidth-vector $\mathbf{h} = (h_1, h_2)^T$. Integration over the time domain $[0, \tau]$ yields the final form of the criterion function

$$\sum_{i=1}^n \int_0^\tau \Big[\Delta N_i(s) - \alpha_0 - \alpha_{10}(s-t) - \alpha_{01}\{\beta^T\mathbf{Z}_i(s) - u\}\Big]^2 K_{\mathbf{h}}\{s-t, \beta^T\mathbf{Z}_i(s) - u\}Y_i(s)ds. \tag{3.4}$$

The minimizer of (3.4) with respect to $(\alpha_0, \alpha_{10}, \alpha_{01})$ is our anticipated local linear weighted least squares estimators of $\{(\lambda(\cdot,\cdot), \lambda_{10}(\cdot,\cdot), \lambda_{01}(\cdot,\cdot)\}$.

By some algebraic manipulation, the minimizer of (3.4) takes the explicit form

$$\widehat{\lambda}(t,u;\beta) = \frac{(S_{20}S_{02} - S_{11}^2)T_{00} + (S_{11}S_{01} - S_{10}S_{02})T_{10} + (S_{10}S_{11} - S_{20}S_{01})T_{01}}{2S_{01}S_{10}S_{11} - S_{02}S_{10}^2 - S_{01}^2S_{20} - S_{00}S_{11}^2 + S_{00}S_{20}S_{02}}; \quad (3.5)$$

$$\widehat{\lambda}_{10}(t,u;\beta) = \frac{(S_{01}S_{11} - S_{10}S_{02})T_{00} + (S_{00}S_{02} - S_{01}^2)T_{10} + (S_{10}S_{01} - S_{00}S_{11})T_{01}}{2S_{01}S_{10}S_{11} - S_{02}S_{10}^2 - S_{01}^2S_{20} - S_{00}S_{11}^2 + S_{00}S_{20}S_{02}}; \quad (3.6)$$

$$\widehat{\lambda}_{01}(t,u;\beta) = \frac{(S_{10}S_{11} - S_{01}S_{20})T_{00} + (S_{01}S_{10} - S_{00}S_{11})T_{10} + (S_{00}S_{20} - S_{10}^2)T_{01}}{2S_{01}S_{10}S_{11} - S_{02}S_{10}^2 - S_{01}^2S_{20} - S_{00}S_{11}^2 + S_{00}S_{20}S_{02}}, \quad (3.7)$$

where

$$S_{jk} = S_{jk}(t,u;\beta) = n^{-1}\sum_{i=1}^{n}\int_0^\tau K_{\mathbf{h}}\{s-t, \beta^T\mathbf{Z}_i(s) - u\}(s-t)^j\{\beta^T\mathbf{Z}_i(s) - u\}^k Y_i(s)ds;$$

$$T_{jk} = T_{jk}(t,u;\beta) = n^{-1}\sum_{i=1}^{n}\int_0^\tau K_{\mathbf{h}}\{s-t, \beta^T\mathbf{Z}_i(s) - u\}(s-t)^j\{\beta^T\mathbf{Z}_i(s) - u\}^k dN_i(s),$$

for $j,\ k = 0, 1, 2$. The forms of (3.5)-(3.7) under the random censorship are concise and corroborate the bivariate local linear regression estimators in Section 4.5, Härdle et al. (2004) for complete data. The explicit formulas are no doubt useful in practice. For the true parameter vector $\beta_0 = (\beta_1^0, \beta_2^0, \cdots, \beta_p^0)^T$, the local linear weighted least squares estimator of the unknown hazard rate function $\lambda(\cdot,\cdot)$ is uniformly consistent. Here comes the result.

Proposition 3.1. *Under Assumptions 3.1 to 3.4 in the section 3.9, then with* $h_1 \to 0, h_2 \to 0$ *and* $n \to \infty$ *such that* $nh_1h_2/\log n \to \infty$, *we have*

$$\sup_{t\in[0,\tau], \boldsymbol{z}\in\mathcal{Z}} \left|\hat{\lambda}(t, \beta_0^T\boldsymbol{z}; \beta_0) - \lambda(t, \beta_0^T\boldsymbol{z})\right| = O_p(\sqrt{\log n/(nh_1h_2)} + h_1^2 + h_2^2).$$

The consistency of $\hat{\lambda}(\cdot,\cdot)$ ensures the consistency of the the proposed profile likelihood estimators in the next adjacent subsection. The convergence rate of $\hat{\lambda}(\cdot,\cdot)$ at the true parameter β_0 has not been adequately discussed in the nonparametric hazards estimation literature since the proof is not trivial.

3.2.2 Profile Likelihood Estimation of Index Coefficient Vector

For the sake of model identifiability, throughout this chapter, we assume that β belongs to the parameter space $\Theta = \{\beta = (\beta_1, \beta_2, \cdots, \beta_p)^T : \|\beta\|_2 = 1, \beta_1 > 0\}$, where $\|\beta\|_2 = (\beta_1^2 + \beta_2^2 + \cdots + \beta_p^2)^{1/2}$. Note that the parameter space Θ is the boundary point of the p-dimensional unit sphere. Thus, the function $\lambda(t, \beta^T\mathbf{Z}(t))$ does not have derivative at point $\beta \in \Theta$. To this end, we delete the first component β_1 by $\beta_1 = (1 - \|\beta_{-1}\|^2)^{1/2}$, where $\beta_{-1} = (\beta_2, \cdots, \beta_p)^T \in \mathbb{R}^{p-1}$, and the parametric space of β is equivalent to

$$\Theta_{-1} = \{[(1 - \|\beta_{-1}\|^2)^{1/2}, \beta_2, \cdots, \beta_p]^T : \|\beta_{-1}\| < 1\}.$$

Then, the Jacobian matrix of β w.r.t β_{-1} is

$$\mathbf{J}(\beta_{-1}) = \frac{\partial \beta}{\partial \beta_{-1}^T} = \begin{pmatrix} -(1 - \|\beta_{-1}\|^2)^{-1/2}\beta_{-1}^T \\ \mathbf{I}_{p-1} \end{pmatrix},$$

where $\mathbf{I}_{p-1}$ is a unit matrix of size $(p-1) \times (p-1)$.

Applying the fact that $\int_0^\tau \Delta N_i(s) f(s) ds = \int_0^\tau f(s) dN_i(s)$ for any function $f(s)$, we plug $\widehat{\lambda}(t, u; \beta)$ in (3.5) to substitute unknown $\lambda(\cdot,\cdot)$ in the log-likelihood function, and obtain the profile log-likelihood

$$\hat{\ell}_n(\beta) = \sum_{i=1}^n \Big[\int_0^\tau \log \hat{\lambda}\{s, \beta^T\mathbf{Z}_i(s); \beta\} dN_i(s) - \int_0^\tau \hat{\lambda}\{s, \beta^T\mathbf{Z}_i(s); \beta\} Y_i(s) ds \Big]. \tag{3.8}$$

The profile likelihood estimator of β_{-1} is $\hat{\beta}_{-1} = \arg\max_{\beta_{-1} \in \Theta_{-1}} \hat{\ell}_n(\beta)$, leading to the estimator of β denoted by $\hat{\beta} = ((1 - \|\hat{\beta}_{-1}\|^2)^{1/2}, \hat{\beta}_{-1}^T)^T$.

The solver $\hat{\beta}_{-_1}$ is the solution to the following score equation

$$0 = \frac{\partial \hat{\ell}_n(\beta)}{\partial \beta_{-_1}} = \sum_{i=1}^{n} \Big[\int_0^\tau \frac{\partial \hat{\lambda}\{s, \beta^T \mathbf{Z}_i(s); \beta\}/\partial \beta_{-1}}{\hat{\lambda}\{s, \beta^T \mathbf{Z}_i(s); \beta\}} dN_i(s) - \int_0^\tau \frac{\partial \hat{\lambda}\{s, \beta^T \mathbf{Z}_i(s); \beta\}}{\partial \beta_{-1}} Y_i(s) ds \Big].$$

Let $\beta_0 = (\beta_1^0, \beta_2^0, \cdots, \beta_p^0)^T$ be the true vector parameter and $\beta_{-_1}^0 = (\beta_2^0, \cdots, \beta_p^0)^T$. As shown in Lemma 3.4 of the Section 3.9, we have

$$\begin{aligned} \frac{\partial \hat{\ell}_n(\beta_0)}{\partial \beta_{-_1}} &= \mathbf{J}^T(\beta_{-1}^0) \sum_{i=1}^{n} \int_0^\tau \Big\{ \mathbf{Z}_i(s) - \frac{E[Y(s)\mathbf{Z}(s)|\beta_0^T \mathbf{Z}_i(s)]}{E[Y(s)|\beta_0^T \mathbf{Z}_i(s)]} \Big\} \frac{\lambda_{01}\{s, \beta_0^T \mathbf{Z}_i(s)\}}{\lambda\{s, \beta_0^T \mathbf{Z}_i(s)\}} dM_i(s) \\ &\quad + o_p(\sqrt{n}). \end{aligned} \tag{3.9}$$

Therefore, the profile likelihood estimator $\hat{\beta}$ is asymptotically equivalent to the solution to the following estimated equation

$$0 = \mathbf{J}^T(\beta_{-_1}) \sum_{i=1}^{n} \int_0^\tau \Big\{ \mathbf{Z}_i(s) - \frac{\widehat{E}[Y(s)\mathbf{Z}(s)|\beta^T \mathbf{Z}_i(s)]}{\widehat{E}[Y(s)|\beta^T \mathbf{Z}_i(s)]} \Big\} \frac{\widehat{\lambda}_{01}\{s, \beta^T \mathbf{Z}_i(s); \beta\}}{\widehat{\lambda}\{s, \beta^T \mathbf{Z}_i(s); \beta\}} d\widehat{M}_i(s), \tag{3.10}$$

where $d\widehat{M}_i(s) = dN_i(s) - \widehat{\lambda}\{s, \beta^T \mathbf{Z}_i(s); \beta\} Y_i(s) ds$ and $\widehat{E}[\cdot|\beta^T \mathbf{Z}_i(s)]$ may be substituted for the kernel estimators mentioned later.

It is observed that the main term on the right-hand side of (3.9) is the estimator of the semiparametric efficient score of β_{-_1}. Recall that the semiparametric efficient score is defined as the projection of the score vector onto the orthogonal complement of the nuisance tangent space, spanned linearly by the nuisance score function, refer to pp.70, Bickel et al. (1997) or pp.47, Tsiatis (2006). This interesting finding is the stepping stone of our main results on the estimation of parameter vector within our model context.

Proposition 3.2. *Assume that $\lambda(t, u)$ is positive and differentiable w.r.t. u, where $u \in \mathcal{U}_\beta = \{u = \beta^T \mathbf{z} : z \in \mathcal{Z}\}$ and $\mathcal{Z}$ is a compact support set of $\boldsymbol{Z}(t)$. Then, the*

semiparametric efficient score for estimation of β_{-1} *in model* (3.1) *is*

$$S_{\text{eff}}\{X, \delta, \boldsymbol{Z}(X)\} = \boldsymbol{J}^T(\beta_{-1}^0) \int_0^\tau \Big\{ \boldsymbol{Z}(s) - \frac{E[Y(s)\boldsymbol{Z}(s)|\beta_0^T \boldsymbol{Z}(s)]}{E[Y(s)|\beta_0^T \boldsymbol{Z}(s)]} \Big\} \frac{\lambda_{01}\{s, \beta_0^T \boldsymbol{Z}(s)\}}{\lambda\{s, \beta_0^T \boldsymbol{Z}(s)\}} dM(s) \quad (3.11)$$

And $E[S_{\text{eff}}\{X, \delta, \boldsymbol{Z}(X)\}] = 0$. *The information matrix of* β_{-1} *is* $E[S_{\text{eff}}\{X, \delta, \boldsymbol{Z}(X)\}^{\otimes 2}] = \Sigma$, *where*

$$\Sigma = \boldsymbol{J}^T(\beta_{-1}^0) E\Big[\int_0^\tau \Big\{ \boldsymbol{Z}(s) - \frac{E[Y(s)\boldsymbol{Z}(s)|\beta_0^T \boldsymbol{Z}(s)]}{E[Y(s)|\beta_0^T \boldsymbol{Z}(s)]} \Big\}^{\otimes 2} \frac{\lambda_{01}^2\{s, \beta_0^T \boldsymbol{Z}(s)\}}{\lambda\{s, \beta_0^T \boldsymbol{Z}(s)\}} Y(s) ds \Big] \boldsymbol{J}(\beta_{-1}^0). \quad (3.12)$$

and $\alpha^{\otimes 2} = \alpha\alpha^T$ *for any column vector* α.

Note that the semiparametric efficient score function, $S_{\text{eff}}\{X, \delta, \mathbf{Z}(X)\}$ is model-dependent but does not depend on the observations.

Expressions (3.9) and (3.11) ensure that the profile likelihood estimator $\hat{\beta}_{-1}$ is semiparametric efficient, that is, $\hat{\beta}_{-1}$ has the smallest variance among a class of regular asymptotically linear estimators, shown in the following Theorem 3.1. Besides Proposition 1, another proposition also plays important role in attaining the efficiency in Theorem 3.1.

Proposition 3.3. *Under Assumption 3.1, 3.2, and 3.4 in the Section 3.9, then with* $h_1 \to 0, h_2 \to 0$ *and* $n \to \infty$ *such that* $nh_1h_2^3/\log n \to \infty$, *it holds*

$$\sup_{t\in[0,\tau], \boldsymbol{z}\in\mathcal{Z}} \Big\| \frac{\partial \hat{\lambda}(t, \beta_0^T \boldsymbol{z}; \beta_0)}{\partial \beta_{-1}} - \lambda_{01}(t, \beta_0^T \boldsymbol{z}) \boldsymbol{J}^T(\beta_{-1}^0) \Big\{ \boldsymbol{z} - \frac{E[Y(t)\boldsymbol{Z}(t)|U_0(t) = \beta_0^T \boldsymbol{z}]}{E[Y(t)|U_0(t) = \beta_0^T \boldsymbol{z}]} \Big\} \Big\|$$
$$= O_p(h_1^2 + h_2^2 + h_2^{-1}\sqrt{\log n/(nh_1h_2)}),$$

where $\frac{\partial \hat{\lambda}(t, \beta_0^T \boldsymbol{z}; \beta_0)}{\partial \beta_{-1}} = \frac{\partial \hat{\lambda}(t, \beta^T \boldsymbol{z}; \beta)}{\partial \beta_{-1}}\Big|_{\beta=\beta_0}$.

Proposition 3.3 suggests that $\partial \hat{\lambda}(t, \beta_0^T \mathbf{z}; \beta_0)/\partial \beta_{-1}$ does not converge in probability

to $\lambda_{01}(t, \beta_0^T \mathbf{z})\mathbf{J}^T(\beta_{-1}^0)$ owing to the fact that

$$\lim_{n\to\infty} \frac{\partial \hat{\lambda}(t, \beta_0^T \mathbf{z}; \beta_0)}{\partial \beta_{-1}} \neq \frac{\partial\{\lim_{n\to\infty} \hat{\lambda}(t, \beta^T \mathbf{z})\}}{\partial \beta_{-1}}\Big|_{\beta=\beta_0}.$$

The index coefficients actually impact the support of the nuisance nonparamtric function $\lambda(\cdot, \cdot)$. See Section 3.9 for the detail proof.

Now it is safe to show asymptotic efficient result for the profile estimator $\hat{\beta}_{-1}$ in the following theorem.

Theorem 3.1. *Under the regularity conditions in Section 3.9, if* $nh_1^8 \to 0$ *and* $nh_2^8 \to 0$, $nh_1h_2^3/\log n \to \infty$ *and* $nh_1^2h_2^2 \to \infty$, *then we have*

(i) $\hat{\beta}_{-1}$ *converges in probability to the true parameter* β_{-1}^0;

(ii) $\hat{\beta}_{-1}$ *is a semiparametrically efficient estimator. That is,*

$$\sqrt{n}(\hat{\beta}_{-1} - \beta_{-1}^0) \xrightarrow{\text{d}} N(0, \Sigma^{-1});$$

(iii) $\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{\text{d}} N(0, \boldsymbol{J}(\beta_{-1}^0)\Sigma^{-1}\boldsymbol{J}(\beta_{-1}^0)^T)$.

The strict theoretical derivation in Theorem 3.1 mainly relies on counting process theory. Here the common martingale method is no longer feasible for the proof due to the unpredictability in that the profile likelihood involves the estimation of nonparametric function $\lambda(\cdot, \cdot)$, which uses all observational information. The result of Theorem 4 in Mammen and Nielsen (2007) sheds light on our proof. See Section 3.9 for detail. Note that the profile estimator does not require undersmoothing of $\lambda(\cdot, \cdot)$ to obtain root-n consistency. This is primarily caused by the result in Proposition 3.3.

3.2.3 Significant Test of Index Coefficients

From Theorem 3.1, the asymptotic covariance of $\hat{\beta}$ achieves lower information bound. It can be estimated by $\mathbf{J}(\hat{\beta}_{-1})\widehat{\Sigma}^{-1}\mathbf{J}(\hat{\beta}_{-1})^T$, where $\widehat{\Sigma}$ has an empirical plug-in candidate estimator of Σ

$$\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \left\{\mathbf{Z}_i(s) - \frac{\widehat{E}[Y(s)\mathbf{Z}(s)|\hat{\beta}^T\mathbf{Z}_i(s)]}{\widehat{E}[Y(s)|\hat{\beta}^T\mathbf{Z}_i(s)]}\right\}^{\otimes 2} \frac{\widehat{\lambda}_{01}^2\{s,\hat{\beta}^T\mathbf{Z}_i(s);\hat{\beta}\}}{\widehat{\lambda}\{s,\hat{\beta}^T\mathbf{Z}_i(s);\hat{\beta}\}} Y_i(s)ds,$$

where $\widehat{E}[Y(s)|\hat{\beta}^T\mathbf{Z}(s) = u]$ and $\widehat{E}[Y(s)\mathbf{Z}(s)|\hat{\beta}^T\mathbf{Z}(s) = u]$ may be obtained by the kernel estimators,

$$\widehat{E}[Y(s)|\hat{\beta}^T\mathbf{Z}(s) = u] = \frac{\sum_{i=1}^n k_{h_3}\{\hat{\beta}^T\mathbf{Z}_i(s) - u\}I(X_i \geqslant s)}{\sum_{i=1}^n k_{h_3}\{\hat{\beta}^T\mathbf{Z}_i(s) - u\}};$$

$$\widehat{E}[Y(s)\mathbf{Z}(s)|\hat{\beta}^T\mathbf{Z}(s) = u] = \frac{\sum_{i=1}^n k_{h_3}\{\hat{\beta}^T\mathbf{Z}_i(s) - u\}\mathbf{Z}_i(s)I(X_i \geqslant s)}{\sum_{i=1}^n k_{h_3}\{\hat{\beta}^T\mathbf{Z}_i(s) - u\}},$$

with bandwidth h_3. Note that $\widehat{\Sigma}$ is shown to be consistent for Σ, refer to proof of Theorem 3.2. Then, looking $\lambda(\cdot,\cdot)$ as if a nuisance function, for the semiparametric testing problem

$$H_0 : \beta = \beta_0 \qquad \text{vs.} \qquad H_1 : \beta \neq \beta_0, \tag{3.13}$$

a generalized Wald test statistic W_n may be defined as

$$W_n = n(\hat{\beta} - \beta_0)^T\left\{\mathbf{J}(\hat{\beta}_{-1})\widehat{\Sigma}^{-1}\mathbf{J}(\hat{\beta}_{-1})^T\right\}^{-1}(\hat{\beta} - \beta_0).$$

This result may be applied to test whether a subset of coefficient variables is statistically significant in the semiparametric model. We give the asymptotic distribution in the following theorem.

Theorem 3.2. *Under conditions of Theorem 3.1, the asymptotic null distribution of W_n is $\chi^2(p)$, where p is the dimension of β.*

Another aspect in 3.1 is the selection of bandwidth. The conditions on bandwidth (h_1, h_2) indicates Theorem 3.1 is applicable for a reasonable range of bandwidths, where the bandwidths satisfy the optimal order $h_j = O(n^{-1/6}), j = 1, 2$. In real applications, we may use cross-validation method to select the optimal bandwidth. More discussions on the choice of bandwidth can be found in Section 3.5.

3.3 Efficient and Doubly Robust Estimation

The property of semiparametric efficient bound in Section 3.2 naturally leads to perspective of constructing efficient estimation equations and the application of doubly robust property in this section. In the proof of Proposition 3.2 in the previous section, we have obtained the space orthogonal to the nuisance tangent space Λ in 3.33 in Section 3.9, denoted by

$$\Lambda^{\perp} = \Big\{ \int_0^{\tau} \Big[\alpha\{s, \mathbf{Z}(s)\} - \frac{E[\alpha\{s, \mathbf{Z}(s)\}Y(s)|\beta^T\mathbf{Z}(s)]}{E[Y(s)|\beta^T\mathbf{Z}(s)]}\Big]\Big[dN(s) - \lambda\{s, \beta^T\mathbf{Z}(s)\}Y(s)ds\Big]\Big\},$$

for any arbitrary $p-1$ dimensional measurable function $\alpha(s, \mathbf{z})$ of $(s, \mathbf{z})$, $\mathbf{z} \in \mathcal{Z}$. This orthogonal component to the nuisance tangent space drives our study in this section. A class of efficient estimation equations are motivated to be presented which corroborate the semiparametric efficient score in Section 3.2. Meanwhile the insight into the doubly robust property that each element in the orthogonal component space possesses stimulates another two sets of estimation equations which enjoys computing convenience to some extent.

Notice that the expectation of each element in $\Lambda^{\perp}$ is equal to 0. This leads us to construct the influential equation

$$\frac{1}{n}\sum_{i=1}^{n} \int_0^{\tau} \Big[\alpha\{s, \mathbf{Z}_i(s)\} - \frac{E[\alpha\{s, \mathbf{Z}(s)\}Y(s)|\beta^T\mathbf{Z}_i(s)]}{E[Y(s)|\beta^T\mathbf{Z}_i(s)]}\Big]\Big[dN_i(s) - \lambda\{s, \beta^T\mathbf{Z}_i(s)\}Y_i(s)ds\Big] = 0,$$

refer to Chapter 5, Tsiatis (2006). Hence we are able to construct a class of efficient

estimating equations. Another observation is that each element in $\Lambda^{\perp}$ enjoys double robust property, that is, its expectation is equal to 0 if either the fraction part or $\lambda\{s, \beta_0^T \mathbf{Z}(s)\}$ is true, refer to Scharfstein et al. (1999), van der Laan and Robins (2003), Cao et al. (2009), among others. This motivates us to present consistent estimation equations for β without estimating $\lambda(\cdot, \cdot)$ or its partial derivative $\lambda_{01}(t, u)$.

The selection of $\alpha\{s, \mathbf{Z}(s)\}$ in $\Lambda^{\perp}$ will generate different estimation equations. Specifically we take it by

$$\alpha\{s, \mathbf{Z}(s)\} = \mathbf{J}^T(\beta_{-1})\gamma\{\mathbf{Z}(s)\}\omega\{s, \beta^T \mathbf{Z}(s)\},$$

where $\gamma(\cdot)$ is an arbitrary p-dimensional measurable function of $\mathbf{Z}(s)$ and $\omega\{s, \beta^T \mathbf{Z}(s)\}$ is any measurable weight function, yielding the following class of estimating equations of β_{-1}:

$$\mathbf{J}^T(\beta_{-1}) \sum_{i=1}^{n} \int_0^{\tau} \Big[\gamma\{\mathbf{Z}_i(s)\} - \frac{E[Y(s)\gamma\{\mathbf{Z}(s)\}|\beta^T \mathbf{Z}_i(s)]}{E[Y(s)|\beta^T \mathbf{Z}_i(s)]}\Big]\omega(s, \beta^T \mathbf{Z}_i(s))\Big[dN_i(s) - \lambda\{s, \beta^T \mathbf{Z}_i(s)\}Y_i(s)ds\Big] = 0. \tag{3.14}$$

3.3.1 Efficient Estimation

In this subsection we propose a class of efficient estimation equations which corroborate the rational of score equation (3.10) based on the equation (3.14). Substituting $\omega\{s, \beta^T \mathbf{Z}(s)\} = \lambda_{01}\{s, \beta^T \mathbf{Z}(s)\}/\lambda\{s, \beta^T \mathbf{Z}(s)\}$ and $\gamma\{\mathbf{Z}(s)\} = \mathbf{Z}(s)$ into the left side of (3.14), we have

$$\mathbf{J}^T(\beta_{-1}) \sum_{i=1}^{n} \int_0^{\tau} \Big\{\mathbf{Z}_i(s) - \frac{E[Y(s)\mathbf{Z}(s)|\beta^T \mathbf{Z}_i(s)]}{E[Y(s)|\beta^T \mathbf{Z}_i(s)]}\Big\}\frac{\lambda_{01}\{s, \beta^T \mathbf{Z}_i(s)\}}{\lambda\{s, \beta^T \mathbf{Z}_i(s)\}} dM_i(s),$$

which is exactly the linear approximation in the right-hand side of (3.9) through replacing β by the true parameter vector β_0. Substituting $\lambda\{s, \beta^T \mathbf{Z}(s)\}$, $\lambda_{01}\{s, \beta^T \mathbf{Z}(s)\}$

and $E[\cdot|\beta^T\mathbf{Z}(s)]$ with their estimators (3.5), (3.7) and kernel estimators aforementioned in previous section, we obtain a class of efficient estimation equations of β_{-1} below

$$0 = \mathbf{J}^T(\beta_{-1})\sum_{i=1}^{n}\int_0^\tau \Big\{\mathbf{Z}_i(s) - \frac{\widehat{E}[Y(s)\mathbf{Z}(s)|\beta^T\mathbf{Z}_i(s)]}{\widehat{E}[Y(s)|\beta^T\mathbf{Z}_i(s)]}\Big\}\frac{\widehat{\lambda}_{01}\{s,\beta^T\mathbf{Z}_i(s);\beta\}}{\widehat{\lambda}\{s,\beta^T\mathbf{Z}_i(s);\beta\}}\Big[dN_i(s) - \widehat{\lambda}\{s,\beta^T\mathbf{Z}_i(s);\beta\}Y_i(s)ds\Big], \tag{3.15}$$

which is identical with score equation (3.10). The efficiency of (3.15) can be demonstrated by some special cases. Take the popular proportional hazards model as the true model for instance. That is, substitute the ratio minuend in the first bracket with $\big[\sum_{i=1}^n Y_i(s)\exp\{\beta^T\mathbf{Z}_i(s)\}\mathbf{Z}_i(s)\big]/\big[\sum_{i=1}^n \exp\{\beta^T\mathbf{Z}_i(s)\}Y_i(s)\big]$. Up to a Jacobian coefficient matrix, (3.15) reduces to the partial likelihood score equation

$$\sum_{i=1}^{n}\int_0^\tau \Big[\mathbf{Z}_i(s) - \frac{\sum_{i=1}^n Y_i(s)\exp\{\beta^T\mathbf{Z}_i\}\mathbf{Z}_i}{\sum_{i=1}^n \exp\{\beta^T\mathbf{Z}_i\}Y_i(s)}\Big]dN_i(s) = 0.$$

Denote $\check{\beta}_{-1}$ to be the solution to (3.15). Let $\check{\beta} = ((1-\|\check{\beta}_{-1}\|^2)^{1/2}, \check{\beta}_{-1})^T$. As expected, the following theorem shows that under regular conditions, $\check{\beta}_{-1}$ and $\check{\beta}$ can reach the same asymptotic variance as $\hat{\beta}_{-1}$ and $\hat{\beta}$ do by the profile likelihood method in Section 3.2.

Theorem 3.3. *Under the mild regularity conditions given in Section 3.9, if* $nh_1^8 \to 0$ *and* $nh_2^8 \to 0$, $nh_1h_2^3/\log n \to \infty$ *and* $h_3 = O(n^{-1/5})$ *, then we have*

(i) $\sqrt{n}(\check{\beta}_{-1} - \beta^0_{-1}) \xrightarrow{\mathrm{d}} N(0, \Sigma^{-1})$;

(ii) $\sqrt{n}(\check{\beta} - \beta_0) \xrightarrow{\mathrm{d}} N\Big(0, \boldsymbol{J}(\beta^0_{-1})\Sigma^{-1}\boldsymbol{J}(\beta^0_{-1})^T\Big)$.

3.3.2 Doubly Robust Estimation

In this subsection we explore feasible estimation of β in two directories: estimation without estimating $\lambda(\cdot,\cdot)$ (NEst.lambda), or estimation without estimating the

partial derivative $\lambda_{01}(t,u)$ (NEst.Plambda).

Recall that in Section 3.2, we employ the multivariate local linear regression idea to estimate the nonparametric component, $\lambda(\cdot,\cdot)$ and its partial derivatives. However, estimation of the nonparametric component will affect substantially the properties of estimators of β. It therefore raises the question that whether consistency of estimator of the parameter for β can be achieved without estimating the partial derivative $\lambda_{01}(t,u)$ or even $\lambda(\cdot,\cdot)$ itself. The doubly robust property in equation (3.14) provides clues for presenting two sets of estimation equations.

From one perspective, we may avoid estimating $\lambda(\cdot,\cdot)$ but obtain the consistent estimators for β. Extremely, setting $\lambda\{s,\beta^T\mathbf{Z}(s)\}=0$, $\omega\{s,\beta^T\mathbf{Z}(s)\}=1$, and $\gamma\{\mathbf{Z}(s)\}=\mathbf{Z}(s)$, we obtain the estimating equation of $\beta_{_{-1}}$ as

$$\mathbf{J}^T(\beta_{_{-1}})\sum_{i=1}^{n}\int_0^{\tau}\Big\{\mathbf{Z}_i(s)-\frac{E[Y(s)\mathbf{Z}(s)|\beta^T\mathbf{Z}_i(s)]}{E[Y(s)|\beta^T\mathbf{Z}_i(s)]}\Big\}dN_i(s)=0. \tag{3.16}$$

Comparing to the profile likelihood in (3.8), the estimating equations (3.16) only require to estimate $E[\cdot|\beta^T\mathbf{Z}(s)]$ without estimating the bivariate function $\lambda(s,\beta^T\mathbf{Z}(s))$. Notice that the estimator of $E[\cdot|\beta^T\mathbf{Z}(s)]$ can be estimated by some classical univariate nonparametric regression methods, say, by kernel estimation, we have

$$\widehat{E}[Y(s)|\beta^T\mathbf{Z}(s)=u]=\frac{\sum_{i=1}^n k_{h_3}\{\beta^T\mathbf{Z}_i(s)-u\}I(X_i\geqslant s)}{\sum_{i=1}^n k_{h_3}\{\beta^T\mathbf{Z}_i(s)-u\}}$$

and

$$\widehat{E}[Y(s)\mathbf{Z}(s)|\beta^T\mathbf{Z}(s)=u]=\frac{\sum_{i=1}^n k_{h_3}\{\beta^T\mathbf{Z}_i(s)-u\}\mathbf{Z}_i(s)I(X_i\geqslant s)}{\sum_{i=1}^n k_{h_3}\{\beta^T\mathbf{Z}_i(s)-u\}},$$

separately with the bandwidth h_3. An estimator of $\beta_{_{-1}}$ can hence be obtained from the following estimation equations

$$\mathbf{J}^T(\beta_{_{-1}})\sum_{i=1}^{n}\int_0^{\tau}\Big\{\mathbf{Z}_i(s)-\frac{\widehat{E}[Y(s)\mathbf{Z}(s)|\beta^T\mathbf{Z}_i(s)]}{\widehat{E}[Y(s)|\beta^T\mathbf{Z}_i(s)]}\Big\}dN_i(s)=0. \tag{3.17}$$

Let $\hat{\beta}^*_{-1}$ be the solution to (3.17) and hence an estimator of β be $\hat{\beta}^* = ((1 - \|\hat{\beta}^*_{-1}\|^2)^{1/2}, \hat{\beta}^{*T}_{-1})^T$. We have the following asymptotic results for $\hat{\beta}^*_{-1}$ and $\hat{\beta}^*$.

Theorem 3.4. *Under the regularity conditions given in Section 3.9, if* $nh_3^4 \to 0$ *and* $nh_3/\log n \to \infty$, *then we have*

(i) $\sqrt{n}(\hat{\beta}^*_{-1} - \beta^0_{-1}) \xrightarrow{\mathrm{d}} N(0, A^{-1}BA^{-1})$;

(ii) $\sqrt{n}(\hat{\beta}^* - \beta_0) \xrightarrow{\mathrm{d}} N\Big(0, \boldsymbol{J}(\beta^0_{-1})A^{-1}BA^{-1}\boldsymbol{J}(\beta^0_{-1})^T\Big)$,

where

$$A = \boldsymbol{J}^T(\beta^0_{-1})E\Big[\int_0^\tau \Big\{\boldsymbol{Z}(s) - \frac{E[Y(s)\boldsymbol{Z}(s)|\beta_0^T\boldsymbol{Z}(s)]}{E[Y(s)|\beta_0^T\boldsymbol{Z}(s)]}\Big\}^{\otimes 2}\lambda_{01}\{s, \beta_0^T\boldsymbol{Z}(s)\}Y(s)ds\Big]\boldsymbol{J}(\beta^0_{-1}) \quad (3.18)$$

$$B = \boldsymbol{J}^T(\beta^0_{-1})E\Big[\int_0^\tau \Big\{\boldsymbol{Z}(s) - \frac{E[Y(s)\boldsymbol{Z}(s)|\beta_0^T\boldsymbol{Z}(s)]}{E[Y(s)|\beta_0^T\boldsymbol{Z}(s)]}\Big\}^{\otimes 2}\lambda\{s, \beta_0^T\boldsymbol{Z}(s)\}Y(s)ds\Big]\boldsymbol{J}(\beta^0_{-1}) \quad (3.19)$$

In order to obtain $\sqrt{n}$-consistency and asymptotic normality of $\hat{\beta}^*_{-1}$ and $\hat{\beta}^*$, Theorem 3.4 indicates that we need to use an undersmoothing bandwidth, i.e, $nh_3^4 \to 0$. This is a common requirement in semiparametric estimation problems, see Carroll et al. (1997). To this end, we may employ an ad hoc bandwidth $h_3 = \hat{h}_{3,\mathrm{opt}} \times n^{-2/15} = O_p(n^{-1/3})$, where $\hat{h}_{3,\mathrm{opt}} \propto n^{-1/5}$ is the optimal bandwidth for $E[Y(s)|\beta^T\mathbf{Z}(s)]$ and $E[Y(s)\mathbf{Z}(s)|\beta^T\mathbf{Z}(s)]$. In our simulation studies, such a method is insensitive to the choice of bandwidth h_3.

From Theorem 3.4, the asymptotic covariance matrix of $\hat{\beta}^*$ can be estimated by

$$\mathbf{J}(\hat{\beta}^*_{-1})\hat{A}^{*-1}\hat{B}^*\hat{A}^{*-1}\mathbf{J}(\hat{\beta}^*_{-1})^T,$$

where $\hat{A}^*$ and $\hat{B}^*$ are empirical plug-in estimators of A and B, taken by

$$\hat{A}^* = \mathbf{J}^T(\hat{\beta}^*_{-1})\frac{1}{n}\sum_{i=1}^n\int_0^\tau \Big\{\mathbf{Z}_i(s) - \frac{\widehat{E}[Y(s)\mathbf{Z}(s)|\hat{\beta}^{*T}\mathbf{Z}_i(s)]}{\widehat{E}[Y(s)|\hat{\beta}^{*T}\mathbf{Z}_i(s)]}\Big\}^{\otimes 2}\hat{\lambda}_{01}\{s, \hat{\beta}^{*T}\mathbf{Z}_i(s); \hat{\beta}^*\}$$

$$
\begin{aligned}
& Y_i(s)ds\mathbf{J}(\hat{\beta}^*_{-1}); \\
\hat{B}^* \quad = \quad & \mathbf{J}^T(\hat{\beta}^*_{-1})\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \Big\{\mathbf{Z}_i(s) - \frac{\widehat{E}[Y(s)\mathbf{Z}(s)|\hat{\beta}^{*T}\mathbf{Z}_i(s)]}{\widehat{E}[Y(s)|\hat{\beta}^{*T}\mathbf{Z}_i(s)]}\Big\}^{\otimes 2}\widehat{\lambda}\{s,\hat{\beta}^{*T}\mathbf{Z}_i(s);\hat{\beta}^*\} \\
& Y_i(s)ds\mathbf{J}(\hat{\beta}^*_{-1}).
\end{aligned}
$$

For the test problem (3.13), we may define another Wald test statistic W_n^*

$$
W_n^* = n(\hat{\beta}^* - \beta_0)^T\Big\{\mathbf{J}(\hat{\beta}^*_{-1})\hat{A}^{*-1}\hat{B}^*\hat{A}^{*-1}\mathbf{J}(\hat{\beta}^*_{-1})^T\Big\}^{-1}(\hat{\beta}^* - \beta_0).
$$

Similar to Theorem 3.2, under the regular conditions, the asymptotic null distribution of W_n^* is $\chi^2(p)$.

Remark 3.1. *Ignoring the Jacobian matrix coefficient, it is interesting to notice that the useness and rationale of* (3.17) *is echoed by the so called screening statistic FAST presented in Gorst-Rasmussen and Scheike (2013)*

$$
\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \Big\{\boldsymbol{Z}_i - \frac{\sum_{i=1}^n \boldsymbol{Z}_i Y_i(s)}{\sum_{i=1}^n Y_i(s)}\Big\}dN_i(s),
$$

of which the minuend in the bracket part can be looked as an approximation of the corresponding part in (3.16). *However, Gorst-Rasmussen and Scheike (2013) assumed* $E[\boldsymbol{Z}|\beta_0^T\boldsymbol{Z}] = c\beta_0^T\boldsymbol{Z}$ *for some constant* c *as a general condition for dimension deduction in feature screening, whereas it is not required for our methodology and theory derivation.*

From the second perspective, we may avoid estimating $\lambda_{01}(t,u)$ and obtain the consistent estimators through estimating $\lambda(\cdot,\cdot)$ only. Setting extremely $\gamma(\mathbf{Z}(s)) = \mathbf{Z}(s)$, $\omega\{s,\beta^T\mathbf{Z}(s)\} = 1$ and $E[Y(s)\gamma\{\mathbf{Z}(s)\}|\beta^T\mathbf{Z}(s)] = 0$ in equation (3.14), it yields

$$
\mathbf{J}^T(\beta_{-1})\sum_{i=1}^{n}\int_0^\tau \mathbf{Z}_i(s)\Big[dN_i(s) - Y_i(s)\lambda\{s,\beta^T\mathbf{Z}_i(s)\}ds\Big] = 0.
$$

Replacing $\lambda\{s, \beta^T\mathbf{Z}(s)\}$ with $\widehat{\lambda}\{s, \beta^T\mathbf{Z}(s); \beta\}$ defined in (3.5), another class of estimation equations of β_{-1} is given by

$$\mathbf{J}^T(\beta_{-1})\sum_{i=1}^{n}\int_0^\tau \mathbf{Z}_i(s)\Big[dN_i(s) - Y_i(s)\widehat{\lambda}\{s, \beta^T\mathbf{Z}_i(s); \beta\}ds\Big] = 0. \tag{3.20}$$

Let $\hat{\beta}_{-1}^{**}$ be the solution to (3.20) and an estimator of β be $\hat{\beta}^{**} = ((1-\|\hat{\beta}_{-1}^{**}\|^2)^{1/2}, \hat{\beta}_{-1}^{**T})^T$. We have the following asymptotic result for $\hat{\beta}_{-1}^{**}$.

Theorem 3.5. *Under the regularity conditions given in Section 3.9, if* $n^{1/2}(h_1^2 + h_2^2) \to 0$, $nh_1h_2^3/\log n \to \infty$ *and* $nh_1^2h_2^2 \to \infty$, *then*

(i) $\sqrt{n}(\hat{\beta}_{-1}^{**} - \beta_{-1}^0) \xrightarrow{\mathrm{d}} N(0, A^{-1}BA^{-1})$;

(ii) $\sqrt{n}(\hat{\beta}^{**} - \beta_0) \xrightarrow{\mathrm{d}} N\Big(0, \boldsymbol{J}(\beta_{-1}^0)A^{-1}BA^{-1}\boldsymbol{J}(\beta_{-1}^0)^T\Big)$,

where A *and* B *are defined in* (3.18) *and* (3.19).

From Theorem 3.5, the undersmoothing bandwidth is necessary to obtain a more accurate estimator of $\hat{\beta}^{**}$. Comparing Theorems 3.4 and 3.5, we notice that in theory, aforementioned two estimation methods can asymptotically reach the same efficiency in the sense that their asymptotic covariance matrices are identical. However, in practice, there are different application preference for these two estimation methods, which will be demonstrated in simulation studies.

3.4 Adapted Newton-Raphson Algorithm for β

In this section we introduce an adapted Newton-Raphson algorithm to obtain the efficient estimator $\hat{\beta}$ from equation (3.10) or (3.15). In literature for semiparametric estimation, some authors obtain estimator of the parameter vector β through minimizing the objective function $\hat{\ell}_n(\beta_{-1})$ in (3.2) subject to the constraint that the

norm of β_{-1} is less than 1. However, practical iteration is not robust due to involving $\hat{\beta}_1^{(m+1)} = \sqrt{1 - ||\hat{\beta}_{-1}^{(m+1)}||^2}$, where $\beta^{(m)}$ and $\beta_{-1}^{(m)}$ denote the m-th iterative estimators of β and β_{-1}, respectively. To solve the problem, we modify Newton-Raphson iterative algorithm by using a first order approximation between deviation of β and β_{-1}, refer to equation (3.23). Furthermore, implementation of the algorithm involves approximation of the first and second order derivatives of the profile likelihood $\hat{\ell}_n(\beta_{-1})$. Such derivatives are not easy to be computed since $\hat{\lambda}(t, \beta^T\mathbf{Z}(t); \beta)$ contained in function $\hat{\ell}_n(\beta_{-1})$ does not depend on β explicitly. Expression (3.9) from the result of Lemma 5 provides an asymptotically equivalence, and hence a feasible score equation of $\partial\hat{\ell}_n(\beta)/\partial\beta_{-1} = 0$. In addition, based on results of Lemma 3.5 in Section 3.9 and applying Theorem 2.4.5, Chapter 2 of Fleming and Harrington (2011), we have

$$\begin{aligned}\frac{\partial^2\hat{\ell}_n(\beta_{-1})}{\partial\beta_{-1}\partial\beta_{-1}^T} \approx & -\mathbf{J}^T(\beta_{-1})\sum_{i=1}^{n}\Big\{\int_0^\tau\Big\{\mathbf{Z}_i(s) - \frac{\widehat{E}[Y(s)\mathbf{Z}(s)|\beta^T\mathbf{Z}_i(s)]}{\widehat{E}[Y(s)|\beta^T\mathbf{Z}_i(s)]}\Big\} \\ & \times\frac{\widehat{\lambda}_{01}\{s, \beta^T\mathbf{Z}_i(s); \beta\}}{\widehat{\lambda}\{s, \beta^T\mathbf{Z}_i(s); \beta\}}d\widehat{M}_i(s)\Big\}^{\otimes 2}\mathbf{J}(\beta_{-1}), \end{aligned} \tag{3.21}$$

where $d\widehat{M}_i(s) = dN_i(s) - \widehat{\lambda}\{s, \beta^T\mathbf{Z}_i(s); \beta\}Y_i(s)ds$, and $\widehat{E}[Y(s)|\beta^T\mathbf{Z}(s)]$ and $\widehat{E}[Y(s)\mathbf{Z}(s)|\beta^T\mathbf{Z}(s)]$ are replaced by their kernel estimators.

Here comes our adapted Newton-Raphson algorithm.

Step 1. Let $\beta^{(1)} = (1, 1, \cdots, 1)^T/\sqrt{p}$ be an initial estimator of β. Set $m = 1$.

Step 2. Given $\beta^{(m)}$, calculate $\widehat{\lambda}\{t, \mathbf{Z}_i(t)^T\beta^{(m)}; \beta^{(m)}\}$ and $\widehat{\lambda}_{01}\{t, \mathbf{Z}_i(t)^T\beta^{(m)}; \beta^{(m)}\}$, $i = 1, \cdots, n$ by (3.5) and (3.7).

Step 3. Update $\beta^{(m)}$. In this step, we first update $\beta_{-1}^{(m)}$ by Newton-Raphson algorithm

$$\beta_{-1}^{(m+1)} = \beta_{-1}^{(m)} - \Big\{\frac{\partial^2\hat{\ell}_n(\beta_{-1}^{(m)})}{\partial\beta_{-1}\partial\beta_{-1}^T}\Big\}^{-1}\frac{\partial\hat{\ell}_n(\beta_{-1}^{(m)})}{\partial\beta_{-1}}, \tag{3.22}$$

where $\partial\hat{\ell}_n(\beta_{-1}^{(m)})/\partial\beta_{-1}$ and $\partial^2\hat{\ell}_n(\beta_{-1}^{(m)})/\partial\beta_{-1}\partial\beta_{-1}^T$ can be computed approximately through (3.10) and (3.21). Then, using Taylor expansion, we obtain

$$\beta^{(m+1)} - \beta^{(m)} = \mathbf{J}(\beta_{-1}^{(m)})(\beta_{-1}^{(m+1)} - \beta_{-1}^{(m)}) + O_p(\|\beta_{-1}^{(m+1)} - \beta_{-1}^{(m)}\|^2). \tag{3.23}$$

Combining (3.22) with (3.23), we obtain

$$\beta^{(m+1)} = \beta^{(m)} - \mathbf{J}(\beta_{-1}^{(m)})\Big\{\frac{\partial^2\hat{\ell}_n(\beta_{-1}^{(m)})}{\partial\beta_{-1}\partial\beta_{-1}^T}\Big\}^{-1}\frac{\partial\hat{\ell}_n(\beta_{-1}^{(m)})}{\partial\beta_{-1}}.$$

Finally, update $\beta^{(m)}$ with normalized $\beta^{(m+1)}$. That is, $\beta^{(m+1)} := \beta^{(m+1)}/\|\beta^{(m+1)}\|$. Set $m := m + 1$ and go to Step 2.

Step 4. Repeat Steps 2 and 3 until convergence.

3.5 Estimation for the Nonparametric Part

Given the profile estimator $\hat{\beta}$, we obtain the fitted estimator of $\lambda(t, u)$,

$$\hat{\lambda}(t,u) \equiv \hat{\lambda}(t,u;\hat{\beta}) = \frac{(\hat{S}_{20}\hat{S}_{02} - \hat{S}_{11}^2)\hat{T}_{00} + (\hat{S}_{11}\hat{S}_{01} - \hat{S}_{10}\hat{S}_{02})\hat{T}_{10} + (\hat{S}_{10}\hat{S}_{11} - \hat{S}_{20}\hat{S}_{01})\hat{T}_{01}}{2\hat{S}_{01}\hat{S}_{10}\hat{S}_{11} - \hat{S}_{02}\hat{S}_{10}^2 - \hat{S}_{01}^2\hat{S}_{20} - \hat{S}_{00}\hat{S}_{11}^2 + \hat{S}_{00}\hat{S}_{20}\hat{S}_{02}},$$

where

$$\hat{S}_{jk} = S_{jk}(t,u;\hat{\beta}) = n^{-1}\sum_{i=1}^{n}\int_0^\tau K_{\mathbf{h}}\{s - t, \hat{\beta}^T\mathbf{Z}_i(s) - u\}(s-t)^j\{\hat{\beta}^T\mathbf{Z}_i(s) - u\}^k Y_i(s)ds,$$

$$\hat{T}_{jk} = T_{jk}(t,u;\hat{\beta}) = n^{-1}\sum_{i=1}^{n}\int_0^\tau K_{\mathbf{h}}\{s - t, \hat{\beta}^T\mathbf{Z}_i(s) - u\}(s-t)^j\{\hat{\beta}^T\mathbf{Z}_i(s) - u\}^k dN_i(s),$$

for $j,\ k = 0, 1, 2, \cdots$.

In this section, we first study the asymptotic behaviors of the nonparametric estimator $\hat{\lambda}(t, u)$. For a kernel function $k(\cdot)$, define

$$\mu_j = \int u^j k(u)du, \quad \nu_j = \int u^j k^2(u)du, \quad j = 0, 1, 2.$$

Theorem 3.6. *Under the regularity assumptions given in Section 3.9, if $\hat{\beta}$ is $\sqrt{n}$-consistent,*

(i) *if $\sqrt{nh_1h_2}(h_1^2+h_2^2)$ is bounded and $nh_1h_2^3 \to \infty$ with $h_1 \to 0, h_2 \to 0$ and $n \to \infty$, then we have*

$$\sqrt{nh_1h_2}\Big\{\hat{\lambda}(t,\hat{\beta}^T\boldsymbol{z};\hat{\beta}) - \lambda(t,\beta_0^T\boldsymbol{z}) - b(t,\beta_0^T\boldsymbol{z})\Big\} \xrightarrow{\text{d}}$$

$$N\left(0, \frac{\nu_0^2\lambda(t,\beta_0^T\boldsymbol{z})}{E[Y(t)|\beta_0^T\mathbf{Z}(\mathbf{t}) = \beta_0^T\boldsymbol{z}]f_{\beta_0}(\beta_0^T\boldsymbol{z})}\right),$$

where the asymptotic bias is $b(t,\beta_0^T\boldsymbol{z}) = \frac{1}{2}h_1^2\mu_2\lambda_{20}(t,\beta_0^T\boldsymbol{z}) + \frac{1}{2}h_2^2\mu_2\lambda_{02}(t,\beta_0^T\boldsymbol{z})$.

(ii) *Moreover, as $nh_1h_2^3/\log n \to \infty$, we have*

$$\sup_{t\in[0,\tau],\boldsymbol{z}\in\mathcal{Z}}\left|\hat{\lambda}(t,\hat{\beta}^T\boldsymbol{z};\hat{\beta}) - \lambda(t,\beta_0^T\boldsymbol{z})\right| = O_p(\sqrt{\log n/(nh_1h_2)} + h_1^2 + h_2^2).$$

The uniform consistency as well as convergence rate of $\hat{\lambda}(t,u)$ has been established in afore Theorem 3.6. The asymptotic normality of $\hat{\lambda}(t,u)$ for our semiparametric model is the same as that of the local linear estimator in the nonparametric hazards model (Nielsen, 1998). It lies in, once we attain $\sqrt{n}$-consistency for $\hat{\beta}$, the local linear method may be carried out to fit the nonparametric part of equation (3.5) as if β is known. Compared with existing literature about nonparametric estimation of hazards function, say local constance method in Nielsen and Linton (1995), one challenge in proof that we need to overcome is the predictability issue raised in semiparametric models, where the integrand of martingale integrals is not predictable and thus the classical counting process theory of martingales is not directly applicable, refer to Mammen and Nielsen (2007).

Theorem 3.6 also indicates that the optimal bandwidth may be used. Note that the estimation procedure involves selection of the bandwidths $\mathbf{h} = (h_1,h_2)^T$ at two

different goals: one is to obtain the estimation of β and the other is to get the final fitted $\lambda(t,u)$. For the latter, the theoretic optimal bandwidth is obtained by minimizing the asymptotic mean squared error (AMSE) (Härdle et al., 2004)

$$\text{AMSE}(h_1,h_2) = \frac{1}{4}\mu_2^2\{h_1^2\lambda_{20}(t,\beta_0^T\mathbf{z}) + h_2^2\lambda_{02}(t,\beta_0^T\mathbf{z})\}^2 + \frac{1}{nh_1h_2}\frac{\nu_0^2\lambda(t,\beta_0^T\mathbf{z})}{E[Y(t)|\beta_0^T\mathbf{Z}(\mathbf{t})=\beta_0^T\mathbf{z}]f_{\beta_0}(\beta_0^T\mathbf{z})}.$$

Thus, if we assume that h_1 and h_2 have the same order, the optimal bandwidth is

$$\hat{h}_j^{\text{opt}} \propto n^{-1/6}, j = 1,2.$$

In practice, after we attain an estimator $\hat{\beta}$, bandwidths for estimating $\lambda(\cdot,\cdot)$ is chosen by minimizing the cross-validation score w.r.t. $\mathbf{h}$, as introduced by Nielsen and Linton (1995)

$$Q(\mathbf{h}) = n^{-1}\sum_{i=1}^{n}\Big[\int_0^\tau \hat{\lambda}_h^2\{s,\hat{\beta}^T\mathbf{Z}_i(s);\hat{\beta}\}Y_i(s)ds - 2\int_0^\tau \hat{\lambda}_h^{[-i]}\{s,\hat{\beta}^T\mathbf{Z}_i(s);\hat{\beta}\}dN_i(s)\Big], \tag{3.24}$$

where $\hat{\lambda}_h^{[-i]}(s,\hat{\beta}^T\mathbf{Z}_i(s);\hat{\beta})$ is the leave-one-out version of the estimator. By the general discrete approximation technique, we may calculate the minimizer of $Q(\mathbf{h})$ by the 2-dimensional grid search method, i.e. we define two-dimensional equally spaced grids of 100 bandwidths, with $h_1 \in [\tau/n, \tau/2]$ and $h_2 \in [\text{range}(\hat{\beta}^T\mathbf{Z}_i)/n, \text{range}(\hat{\beta}^T\mathbf{Z}_i)/2]$. In the situation that the practical result of the minimizer of $Q(\mathbf{h})$ is at the boundary, we adopt an indirect cross-validation and more robust bandwidth selection strategy, called do-validation, see Mammen et al. (2011) and Pérez et al. (2013).

3.6 Simulation Studies

In this section, we assess the finite sample performance of the proposed methods in Examples 1, 2 and 3. Example 3.1 aims to evaluate the accuracy and precision of

estimation β for the three proposed methods, profile likelihood estimation (PL), estimation without estimating $\lambda(\cdot,\cdot)$ (NEst.lambda), and estimation without estimating $\lambda_{01}(t,u)$ (NEst.Plambda). Example 3.2 compares performance of aforementioned profile likelihood method with hMAVE method. Example 3.3 demonstrates performance of the fitted hazards function. To save the computational cost, we use the method in Pérez et al. (2013) for discrete approximations of the hazard estimator and set the time grids approximately equal to quarter of the sample size. Each simulation is repeated 1000 times for various scenarios of censoring rate (C.rate) and sample size n.

Example 3.1. *We assume that* $\boldsymbol{Z} = (Z_1, Z_2, Z_3)^T$ *is generated from a multivariate normal distribution with mean* $\boldsymbol{0}$ *and variance-covariance matrix* $(\sigma_{ij})_{3\times 3}$*, where* $\sigma_{ij} = 0.5^{|i-j|}$ *for* $i, j = 1, 2, 3$*. The true parameter vector is* $\beta_0 = (1/2, -\sqrt{3}/2, 0)^T$*. The following three simulation settings are considered:*

(i) *Proportional hazards model (PHM):* $\lambda(t, \beta_0^T \boldsymbol{Z}) = 0.5t \exp(\beta_0^T \boldsymbol{Z})$*;*

(ii) *Additive hazards model (ADH):* $\lambda(t, \beta_0^T \boldsymbol{Z}) = 1 + \beta_0^T \boldsymbol{Z}$*;*

(iii) *Accelerated failure time model (AFT):* $\log(T) = \beta_0^T \boldsymbol{Z} + \varepsilon$ *with* $\varepsilon \sim N(0, 1)$*.*

The censoring time C *is generated from the uniform distribution* $[0,\ c_0]$*, where* c_0 *is modified to control an approximate censoring rate of 20% and 40%, respectively. The bandwidth of* $\boldsymbol{h} = (h_1, h_2)^T$ *is selected by the cross-validation method introduced in Section 3.5.*

Tables 3.1 and 3.2, for all proposed estimatorss, present the bias, sampling standard error of estimator of β_0 (SSE), sampling mean of the standard error estimator (SEE), and the empirical coverage probability (CP) of the 95% confidence interval. Table 3.1 evaluates performance of the semiparametric efficient estimator

by profile likelihood estimation procedure. According to the optimal bandwidth $\hat{h}_j^{\text{opt}} \propto n^{-1/6}, j = 1, 2$, in the previous section, and the fact that $200^{-1/6}$ and $400^{-1/6}$ are pretty close, we use a unified bandwidth for both sample sizes 200 and 400 to indicate that the accuracy and precision of estimators are not sensitive to the bandwidth. Generally speaking, it is evident from the results in table 3.1 that the profiled likelihood estimators have very small bias and yield reasonable SEEs and CPs. For all the three models, the SSEs and SEEs get closer to each other and the CPs are closer to the nominal level 95% as the sample size increases and censoring percentage decreases. Thus, the result of table 3.1 is confirmatory to the asymptotic normality of the parameter estimator established in Theorem 3.1. We also notice that, for Cox's model, SSEs and SEEs of a few estimators differ considerably, and thus CPs may be not that close to the nominal level 95% in the situation of small sample size or high censoring rate. We speculate that it is due to the fluctuation of the exponential tilt in the representation of the proportional hazard function.

Table 3.2 assesses performance of two estimators derived from the doubly robust property. The bandwidth selection is more stable than that for the profile likelihood estimator. This may owe to less complicated kernel smoothing in that we avoid estimation of the unknown $\lambda(\cdot, \cdot)$ or its partial derivative. It is noticeable that even though we reduce the estimation workload by not estimating either the hazards function or its partial derivative, the resulting estimators are still accurate and precise. Therefore, for expedient computation in practice, the two estimation equations built up based on the doubly robust property can be an alternative of the profile likelihood equation.

Example 3.2. *In this example, we compare the performance of our method with hMAVE method at sample size* $n = 200$*. We consider two scenarios for generating* $\boldsymbol{Z}$*: all the covariates are continuous and a mixture of discrete and continuous covariates:*

Table 3.1: Example 3.1: Profile likelihood estimator (semiparametric efficient estimation)

	True model		$n = 200$				$n = 400$			
C.rate	bandwidth	Par.	Bias	SSE	SEE	CP	Bias	SSE	SEE	CP
20%	(i) PHM	β_1	-0.0085	0.0829	0.0869	0.958	-0.0042	0.0542	0.0585	0.947
	(0.2019, 1.6755)	β_2	0.0086	0.0479	0.0510	0.941	0.0034	0.0315	0.0340	0.945
		β_3	-0.0029	0.1189	0.1251	0.937	-0.0012	0.0786	0.0853	0.950
	(ii) ADH	β_1	-0.0029	0.0477	0.0489	0.952	-0.0014	0.0289	0.0299	0.946
	(1.0021, 1.6931)	β_2	0.0027	0.0269	0.0285	0.957	0.0008	0.0165	0.0173	0.945
		β_3	-0.0044	0.0670	0.0716	0.945	-0.0015	0.0404	0.0435	0.952
	(iii) AFT	β_1	-0.0069	0.0755	0.0815	0.961	-0.0027	0.0506	0.0537	0.954
	(1.6609, 1.3383)	β_2	0.0071	0.0440	0.0480	0.963	0.0034	0.0297	0.0313	0.954
		β_3	-0.0055	0.1069	0.1180	0.954	-0.0027	0.0715	0.0781	0.957
40%	(i) PHM	β_1	-0.0104	0.0990	0.1006	0.949	-0.0044	0.0638	0.0664	0.960
	(0.1919, 1.4255)	β_2	0.0286	0.1614	0.0615	0.938	0.0056	0.0366	0.0389	0.963
		β_3	-0.0029	0.1521	0.1438	0.945	-0.0022	0.0933	0.0961	0.952
	(ii) ADH	β_1	-0.0074	0.0668	0.0684	0.950	-0.0032	0.0426	0.0424	0.947
	(1.9021, 1.7931)	β_2	0.0040	0.0376	0.0396	0.951	0.0015	0.0241	0.0245	0.944
		β_3	-0.0028	0.0917	0.0994	0.956	-0.0005	0.0582	0.0618	0.961
	(iii) AFT	β_1	-0.0066	0.0758	0.0849	0.967	-0.0029	0.0515	0.0546	0.955
	(3.0609, 1.3383)	β_2	0.0106	0.0735	0.0506	0.964	0.0039	0.0300	0.0319	0.957
		β_3	-0.0085	0.1167	0.1218	0.955	-0.0054	0.0772	0.0794	0.951

Case 1: all the covariates are continuous. *We take the PHM:*

$$T = \Lambda_0^{-1}\Big\{\varepsilon \exp\left(6\boldsymbol{Z}^T\beta_0 + 1\right)\Big\}, \quad \Lambda_0^{-1}(v) = \Phi\{5(v-2)\}, \tag{3.25}$$

where $\Phi(\cdot)$ *is the cumulative distribution function of* $N(0,1)$, $\varepsilon \sim \exp(1)$ *and* $\boldsymbol{Z} \sim N_7(0, I_7)$ *are independent. Let the true parameter* $\beta_0 = (1, 0, -1, 0, -1, 0, 1)^T/2$. *The censoring time* C *is generated from* $\Phi(2Z_2 + 2Z_3) + c_0$, *where* c_0 *is selected to control an approximate censoring rate of 0%, 20% or 40%. This design is similar to Example 5.1 in Xia et al. (2010). The results are summarized in Table 3.3.*

Case 2: some covariates are categorical. *We take the ADH model:*

$$\lambda(t, \beta_0^T\boldsymbol{Z}) = 1 + \beta_0^T\boldsymbol{Z}, \tag{3.26}$$

where $\beta_0 = (1, 1, \ -1, \ -1)^T/2$. *We generate* Z_3 *from nonlinear models:*

Table 3.2: Example 3.1: Estimation without estimating $\lambda(\cdot,\cdot)$ or without estimating $\lambda_{01}(t,u)$ (doubly robust property estimation)

	True model		$n=200$				$n=400$			
C.rate	Bandwidth	Par.	Bias	SSE	SEE	CP	Bias	SSE	SEE	CP
	Estimation without estimating $\lambda(\cdot,\cdot)$									
20%	(i) PHM	β_1	-0.0076	0.0716	0.0763	0.960	-0.0033	0.0488	0.0513	0.963
	(0.1571, 1.1764)	β_2	0.0063	0.0416	0.0446	0.951	0.0029	0.0283	0.0298	0.957
		β_3	-0.0021	0.1072	0.1105	0.954	-0.0002	0.0718	0.0745	0.956
	(ii) AHD	β_1	-0.0033	0.0613	0.0615	0.950	-0.0010	0.0419	0.0403	0.944
	(0.4864, 0.6383)	β_2	0.0058	0.0370	0.0362	0.952	0.0029	0.0242	0.0242	0.954
		β_3	-0.0052	0.0894	0.0901	0.950	-0.0027	0.0600	0.0600	0.946
	(iii) AFT	β_1	-0.0064	0.0707	0.0733	0.944	-0.0022	0.0486	0.0496	0.948
	(0.5609, 0.6383)	β_2	0.0063	0.0418	0.0432	0.943	0.0034	0.0287	0.0290	0.953
		β_3	-0.0088	0.1017	0.1064	0.951	-0.0042	0.0700	0.0721	0.946
40%	(i) PHM	β_1	-0.0080	0.0852	0.0911	0.962	-0.0015	0.0586	0.0605	0.958
	(0.1571, 1.1764)	β_2	0.0104	0.0499	0.0538	0.955	0.0060	0.0342	0.0356	0.958
		β_3	-0.0055	0.1267	0.1316	0.948	-0.0019	0.0858	0.0879	0.953
	(ii) AHD	β_1	-0.0061	0.0822	0.0802	0.947	-0.0033	0.0547	0.0532	0.945
	(0.4864, 0.6383)	β_2	0.0098	0.0482	0.0481	0.950	0.0039	0.0320	0.0310	0.942
		β_3	-0.0087	0.1176	0.1174	0.944	-0.0026	0.0772	0.0773	0.943
	(iii) AFT	β_1	-0.0059	0.0739	0.0746	0.949	0.0027	0.0506	0.0511	0.952
	(0.5609, 0.6383)	β_2	0.0076	0.0439	0.0442	0.943	0.0037	0.0298	0.0299	0.941
		β_3	-0.0081	0.1073	0.1080	0.945	-0.0048	0.0750	0.0738	0.943
	Estimation without estimating $\lambda_{01}(t,u)$									
20%	(i) PHM	β_1	-0.0118	0.0685	0.0886	0.982	-0.0061	0.0466	0.0567	0.989
	(1.2571, 1.6764)	β_2	0.0029	0.0394	0.0513	0.976	0.0010	0.0270	0.0329	0.984
		β_3	-0.0158	0.1009	0.1342	0.988	-0.0136	0.0683	0.0869	0.982
	(ii) AHD	β_1	-0.0041	0.0599	0.0708	0.982	-0.0020	0.0395	0.0442	0.969
	(2.1021, 1.7931)	β_2	0.0045	0.0344	0.0416	0.977	0.0019	0.0227	0.0257	0.970
		β_3	-0.0073	0.0843	0.1080	0.986	-0.0030	0.0572	0.0669	0.980
	(iii) AFT	β_1	-0.0036	0.0715	0.0820	0.964	0.0001	0.0474	0.0542	0.975
	(1.5609, 0.7383)	β_2	0.0080	0.0484	0.0483	0.965	0.0044	0.0278	0.0316	0.976
		β_3	-0.0053	0.0992	0.1247	0.974	0.0017	0.0674	0.0824	0.982
40%	(i) PHM	β_1	-0.0101	0.0793	0.1065	0.983	-0.0039	0.0544	0.0680	0.980
	(1.2571, 1.6764)	β_2	0.0074	0.0469	0.0626	0.981	0.0040	0.0318	0.0398	0.985
		β_3	-0.0160	0.1186	0.1612	0.985	-0.0118	0.0812	0.1035	0.980
	(ii) AHD	β_1	-0.0066	0.0783	0.0926	0.970	-0.0029	0.0524	0.0586	0.968
	(1.1864, 0.7703)	β_2	0.0082	0.0453	0.0547	0.976	0.0036	0.0304	0.0341	0.965
		β_3	-0.0090	0.1114	0.1414	0.979	-0.0033	0.0740	0.0892	0.977
	(iii) AFT	β_1	-0.0052	0.0723	0.0828	0.969	-0.0019	0.0487	0.0545	0.972
	(1.5609, 0.7383)	β_2	0.0075	0.0433	0.0487	0.962	0.0039	0.0286	0.0381	0.972
		β_3	-0.0036	0.1049	0.1255	0.971	0.0009	0.0731	0.0827	0.969

$Z_3 = |Z_1 + Z_2| + |Z_1|\varepsilon_1$, where ε_i's are independently generated from the standard normal population; Z_4 from a Bernoulli distribution with success probability $\exp(Z_1)/\{1+\exp(Z_1)\}$. *The censoring time C is generated from uniform distribution $U(0,\ c_0)$, where c_0 is selected to control an approximate censoring rate of 20% or 40%. The results are summarized in Table 3.4.*

Tables 3.3 and 3.4 show comparison of bias, standard error (SE) and mean square error (MSE) among our proposed estimators and the estimator by hMAVE method.

When all covariates are continuous, we report the comparison results in Table 3.3. All four estimators have sound bias and MSE. The two estimators based on doubly robust property are comparable with hMAVE. Nevertheless, the profile likelihood estimator outperforms hMAVE since it is semiparametric efficient, refer to the bolded magnitudes.

Table 3.4 comes to the case when there exist discrete covariates. hMAVE, as predicted, does not perform well because the continuity assumption is violated. In contrast to hMAVE, our three proposed estimators still perform well since our methods only need that at least one covariate is continuous. The semiparametric efficient estimator outplays all other methods in bias and MSE, refer to the bolded magnitudes.

Example 3.3. *In this example, we examine the performance of the fitted estimation of hazards function. Data are generated from the proportional hazards model in Example 3.1(i). In this setting, the true hazards function is*

$$\lambda(t,u) = 0.5t\exp(u).$$

We take the sample size $n = 400$ and the approximate censoring rate of 40% .

Figure 3.1 demonstrates the bias of $\hat{\lambda}(t,u)$ compared to $\lambda(t,u)$ at given grid points. The results show that the shape of curved surface is captured well.

Table 3.3: Case 1, Example 3.2: all the covariates are continuous.

Method	Censoring rate		β_1	β_2	β_3	β_4	β_5	β_6	β_7
PL	0%	Bias	**-0.0015**	**0.0013**	0.0039	0.0008	**0.0026**	**0.0006**	**-0.0017**
(0.2501, 1.7857)	($c_1 = 2$)	SE	0.0349	0.0387	0.0368	0.0410	0.0348	0.0393	0.0341
		MSE	**0.0012**	**0.0015**	**0.0014**	**0.0017**	**0.0012**	**0.0015**	0.0012
	20%	Bias	**-0.0021**	**-0.0017**	**0.0015**	**-0.0003**	0.0024	0.0018	-0.0032
(0.1111, 2.1889)	($c_1 = 0.5$)	SE	0.0341	0.0396	0.0330	0.0400	0.0321	0.0408	0.0329
		MSE	**0.0012**	**0.0016**	**0.0011**	**0.0016**	**0.0010**	**0.0017**	**0.0011**
	40%	Bias	**0.0029**	0.0233	0.0227	-0.0015	-0.0053	0.0002	0.0033
(0.1111, 2.1889)	($c_1 = 0.1$)	SE	0.0372	0.0407	0.0350	0.0404	0.0340	0.0420	0.0362
		MSE	**0.0014**	0.0022	0.0017	**0.0016**	**0.0012**	**0.0018**	**0.0013**
hMAVE	0%	Bias	-0.0029	-0.0032	0.0032	-0.0007	0.0032	0.0021	-0.0023
	($c_1 = 2$)	SE	0.0380	0.0433	0.0393	0.0422	0.0381	0.0445	0.0391
		MSE	0.0015	0.0019	0.0016	0.0018	0.0015	0.0020	0.0015
	20%	Bias	-0.0033	0.0039	0.0062	-0.0012	0.0015	0.0009	-0.0006
	($c_1 = 0.5$)	SE	0.0374	0.0450	0.0379	0.0432	0.0378	0.0445	0.0374
		MSE	0.0014	0.0020	0.0015	0.0019	0.0014	0.0020	0.0014
	40%	Bias	-0.0040	-0.0096	-0.0031	0.0010	0.0050	0.0017	-0.0063
	($c_1 = 0.1$)	SE	0.0381	0.0448	0.0393	0.0441	0.0400	0.0435	0.0396
		MSE	0.0015	0.0021	0.0016	0.0019	0.0016	0.0019	0.0016
NEst.lambda	0%	Bias	-0.0006	0.0008	0.0058	0.0030	0.0024	0.0018	-0.0041
(0.1111,3.3262)	($c_1 = 2$)	SE	0.0379	0.0427	0.0441	0.0467	0.0404	0.0459	0.0401
		MSE	0.0014	0.0018	0.0020	0.0022	0.0016	0.0021	0.0016
	20%	Bias	-0.0017	0.0002	0.0023	-0.0005	0.0028	-0.0011	-0.0057
(0.1111,1.9017)	($c_1 = 0.5$)	SE	0.0385	0.0461	0.0411	0.0457	0.0386	0.0442	0.0395
		MSE	0.0015	0.0021	0.0017	0.0021	0.0015	0.0020	0.0016
	40%	Bias	-0.0019	-0.0007	0.0066	-0.0009	0.0035	-0.0005	-0.0030
(0.1111,1.9017)	($c_1 = 0.1$)	SE	0.0395	0.0467	0.0508	0.0467	0.0489	0.0483	0.0410
		MSE	0.0016	0.0022	0.0026	0.0022	0.0024	0.0023	0.0017
NEst.Plambda	0%	Bias	-0.0049	0.0003	0.0033	-0.0032	0.0033	-0.0011	-0.0029
(0.1111, 1.9017)	($c_1 = 2$)	SE	0.0425	0.0491	0.0427	0.0485	0.0423	0.0489	0.0430
		MSE	0.0018	0.0024	0.0018	0.0024	0.0018	0.0024	0.0019
	20%	Bias	-0.0029	-0.0015	0.0064	0.0016	0.0010	-0.0019	-0.0050
(0.1111, 2.1889)	($c_1 = 0.5$)	SE	0.0433	0.0507	0.0442	0.0503	0.0434	0.0509	0.0423
		MSE	0.0019	0.0026	0.0020	0.0025	0.0019	0.0026	0.0018
	40%	Bias	-0.0009	0.0072	0.0105	-0.0026	0.0026	-0.0007	-0.0059
(0.1362, 1.4373)	($c_1 = 0.1$)	SE	0.0470	0.0560	0.0483	0.0550	0.0493	0.0602	0.0542
		MSE	0.0022	0.0032	0.0024	0.0030	0.0024	0.0036	0.0030

3.7 Analysis of Diffuse Large B-Cell Lymphoma

In this section, we apply the proposed method on the selected 14 features of the DLBCL data in Chapter 2 by the IFAST method. The results of the regression analysis by Model (3.1) is reported in Table 3.5. The genes with p-values less than 0.05 are in bold. These genes are very possibly associated with the survival time of lymphoma patients. The results verify that FAST is an effective feature screening method. Among the genes selected by FAST, there are nine out of fourteen are significant. Genes 1181 and 1456, which are also selected by most of Cox's model

Table 3.4: Case 2, Example 3.2: some covariates are categorical.

Method	C.rate		β_1	β_2	β_3	β_4
PL	20%	Bias	**-0.0031**	**0.0003**	**0.0015**	**0.0012**
(0.2401, 2.7749)	($c_0 = 14$)	SE	0.0285	0.0325	0.0318	0.0514
		MSE	**0.0008**	**0.0011**	**0.0010**	**0.0026**
	40%	Bias	**-0.0057**	**-0.0030**	0.0024	**0.0052**
(0.2401, 2.7749)	($c_0 = 3.5$)	SE	0.0480	0.0553	0.0552	0.0884
		MSE	**0.0023**	**0.0031**	**0.0030**	**0.0078**
hMAVE	20%	Bias	-0.0031	0.0042	-0.0025	0.0187
	($c_0 = 14$)	SE	0.0497	0.0553	0.0549	0.0789
		MSE	0.0025	0.0031	0.0030	0.0066
	40%	Bias	-0.0116	-0.0045	0.0014	0.0267
	($c_0 = 3.5$)	SE	0.0872	0.1069	0.0978	0.1214
		MSE	0.0077	0.0114	0.0096	0.0155
NEst.lambda	20%	Bias	-0.0010	0.0009	0.0014	0.0095
(0.1362, 1.1330)	($c_0 = 14$)	SE	0.0401	0.0460	0.0442	0.0726
		MSE	0.0016	0.0021	0.0020	0.0054
	40%	Bias	-0.0079	0.0002	0.0037	0.0124
(0.1362, 1.1330)	($c_0 = 3.5$)	SE	0.0615	0.0669	0.0634	0.1059
		MSE	0.0038	0.0045	0.0040	0.0114
NEst.Plambda	20%	Bias	-0.0030	0.0025	0.0005	0.0094
(0.1362, 1.1330)	($c_0 = 14$)	SE	0.0393	0.0452	0.0432	0.0698
		MSE	0.0016	0.0020	0.0019	0.0050
	40%	Bias	-0.0069	0.0021	0.0033	0.0126
(0.1362, 1.1330)	($c_0 = 3.5$)	SE	0.0560	0.0635	0.0609	0.0980
		MSE	0.0032	0.0040	0.0037	0.0098

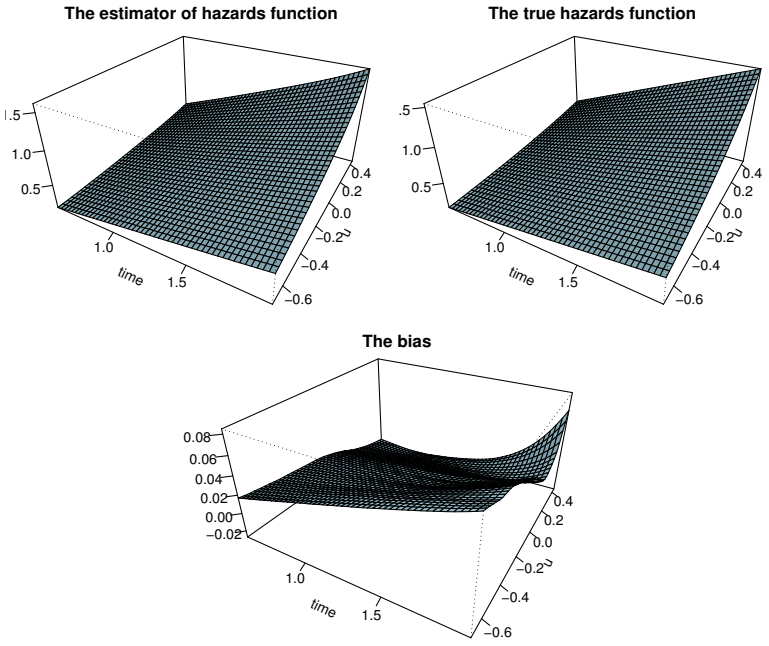

Figure 3.1: Example 3.1: Assume the true model is Cox's model. The bias measures the deviation between the proposed fitted hazard function and the true one.

based feature screening methods, are significant under the SIH model. Notice that genes 1681, 1825, 2311, 4317, 5649, 6519, which are only selected by FAST in Chapter 2, are significant, confirmed by the proposed method. These genes might be neglected by the conventional Cox's model but indeed deserve further investigation.

Table 3.5: Regression analysis for DLBCL data by Model (3.1)

Genes	coefficient	standard error	p-value
197	0.1516	0.0862	0.0784
1188	0.2981	0.1145	**0.0092**
1456	-0.2941	0.0992	**0.0030**
1681	-0.3435	0.1038	**0.0009**
1825	-0.3798	0.1001	**0.0001**
2107	0.1462	0.1543	0.3435
2109	-0.1215	0.1373	0.3760
2240	0.1558	0.1158	0.1785
2311	0.3197	0.1021	**0.0017**
4131	0.3162	0.0841	**0.0002**
4317	-0.3970	0.0949	**0.0000**
5054	0.0263	0.1093	0.8096
5649	0.2077	0.0975	**0.0332**
6519	0.2746	0.0812	**0.0007**

3.8 Discussion

In this chapter, we have systematically studied efficient estimation of the single-index hazards model with right-censored survival outcomes. Our proposed profile maximum likelihood estimator of the parameter component reaches semiparamatric efficient bound. Meanwhile observation on the doubly robustness of influence function class have motivated us to further develop a class of efficient estimation equations. Furthermore, two classes of estimation equations are presented as a practical substitute of the efficient score equations. Rigorous theory proof have been derived for the asymptotic results. univariate

Ding et al. (2013) and Chiang et al. (2017) estimated the cumulative rather than the hazard function directly, by nonparametric likelihood estimation and application of local constant estimation of Nielsen and Linton (1995) separately. Instead, we estimate $\lambda(\cdot,\cdot)$ directly by treating it as a bivariate function.

Local linear regression by Fan and Gijbels (1996) is a widely used method to estimate multivariate function for completely observed data. Nielsen (1998)'s non-parametric hazards estimation actually represents an equivalent or corrected kernel form to handle right censored data, refer to (3.28) in Section 3.9. In the presence of index structure, we further develop multivariate local linear approach that is suitable and even efficient to deal with right-censored failure time. The equivalent kernel nested in our method satisfies some basic properties and allows for the automatic adjustment near boundary regions, refer to the proof in the Section 3.9.

As supplementary, for equation (3.14), if we set $\omega(s, \beta^T\mathbf{Z}(s)) = 1$, we can obtain another estimating equation

$$\mathbf{J}^T(\beta_{-1})\sum_{i=1}^{n}\int_0^\tau \Big\{\mathbf{Z}_i(s) - \frac{\widehat{E}[Y(s)\mathbf{Z}(s)|\beta^T\mathbf{Z}_i(s)]}{\widehat{E}[Y(s)|\beta^T\mathbf{Z}_i(s)]}\Big\}\Big[dN_i(s) - Y_i(s)\widehat{\lambda}\{s,\beta^T\mathbf{Z}_i(s)\}ds\Big] = 0. \quad (3.27)$$

Similar to the proof of Theorems 3.4 and 3.5, we can show that the solution to (3.27) is asymptotically normally distributed with mean 0 and covariance matrix $\mathbf{J}(\beta^0_{-1})A^{-1}BA^{-1}\mathbf{J}(\beta^0_{-1})^T$ under certain regular conditions. However, the bandwidths h_1, h_2 and h_3 in (3.27), are different from the NEst.lambda and NEst.Plambda methods. The optimal bandwidths can still be used.

In practice, there are several types of high dimensional covariates. This motivates our future work to study multiple index hazards modeling that incorporates different types of effects of covariates.

3.9 Proofs of propositions and theorems

In order to study the asymptotic behavior of the proposed estimators, we give the following conditions:

Assumption 3.1. *Assume that the hazard function $\lambda(t,u)$ is three times continuously differentiable at $(t,u) \in [0,\tau] \times \mathcal{U}_\beta$, where $\mathcal{U}_\beta = \{\beta^T \mathbf{z} : z \in \mathcal{Z}\}$ and $\mathcal{Z}$ is a compact support set of $\mathbf{Z}(\mathbf{t})$.*

Assumption 3.2. *Assume that $\phi_\beta(t,u) = f_\beta(u)E[Y(t)|\beta^T \boldsymbol{Z}(t) = u]$ is positive and has bounded second derivatives on $[0,\tau]\times\mathcal{U}_\beta$ for β in some neighborhood of β_0, where $f_\beta(u)$ is the density function of $\beta^T \boldsymbol{Z}(t)$ at u.*

Assumption 3.3. *$f_\beta(u)E[Y(t)\boldsymbol{Z}(t)|\beta^T \boldsymbol{Z}(t) = u]$ has bounded second derivatives at $(t,u) \in [0,\tau] \times \mathcal{U}_\beta$.*

Assumption 3.4. *The kernel $k(\cdot)$ is symmetric density function, with a bounded derivative, and satisfies $\int_{-\infty}^{\infty} u^2 k(u)du < \infty$, $\int_{-\infty}^{\infty} u^{2j}k^2(u)du < \infty$, $j = 0,1,2$.*

Assumption 3.5. *Assume that, for any $\beta \neq \beta_0$, $\lambda(t,\beta^T \boldsymbol{Z}(t)) \neq \lambda(t,\beta_0^T \boldsymbol{Z}(t))$ holds, with probability one.*

Assumption 3.6. *The matrices Σ and A are finite and nonsingular, where Σ and A are defined in (3.12) and (3.18), respectively.*

All of these conditions are analogous to those in the traditional single index literature. Assumption 3.1 gives the smoothness condition of $\lambda(\cdot,\cdot)$. It is noteworthy that Assumption 3.2 is similar to Assumption (S) of Theorem 1 in Nielsen and Linton (1995). In fact, by Fubini's theorem, simple calculation yields

$$E\Big[\int_0^\tau K_{\mathbf{h}}(s-t,\beta^T\mathbf{Z}(s)-u)Y(s)ds\Big] \;=\; f_\beta(u)E[Y(t)|\beta^T\mathbf{Z}(t)=u] + o(1)$$

$$= \quad f^*(u,t)P(Y(t)=1)+o(1),$$

$h_j \to 0$, j=1, 2, where $f^*(u,t)$ is the density function of $F(u,t) = P(\beta^T\mathbf{Z}(t) \leqslant u|Y(t)=1)$. Notice that the main term $f^*(u,t)P(Y(t)=1)$ is the definition of $\varphi(x)$ in Nielsen and Linton (1995)'s Theorem 1. Assumption 3.3 is a mild smoothness condition similar to the condition $E[\mathbf{Z}|\beta^T\mathbf{Z}=u]$ in the classical single index model literature (Condition (C2) in Wang et al. (2010), Condition (a) in Cui et al. (2011) and Condition (C2) in Ma and Zhu (2012)). Assumption 3.4 is a commonly used assumption for second-order kernels. Assumption 3.5 ensures that the likelihood of β is identifiable.

For notational convenience, we write $U_0(t) = \beta_0^T\mathbf{Z}(t)$, $U(t) = \beta^T\mathbf{Z}(t)$, $U_{0i}(t) = \beta_0^T\mathbf{Z}_i(t)$ and $U_i(t) = \beta^T\mathbf{Z}_i(t)$. Let $c_n = \sqrt{\log n/(nh_1h_2)} + h_1^2 + h_2^2$. Define

$$\begin{aligned} V_{ni}(s,t,u;\beta) \quad = \quad & K_{\mathbf{h}}(s-t, U_i(s)-u)\Big\{[S_{20}(t,u)S_{02}(t,u) - S_{11}^2(t,u)] \\ & +[S_{11}(t,u)S_{01}(t,u) - S_{10}(t,u)S_{02}(t,u)](s-t) \\ & +[S_{10}(t,u)S_{11}(t,u) - S_{20}(t,u)S_{01}(t,u)](U_i(s)-u)\Big\} \end{aligned} \tag{3.28}$$

and

$$W_{ni}(s,t,u;\beta) = \frac{V_{ni}(s,t,u;\beta)}{\sum_{i=1}^n \int_0^\tau V_{ni}(s,t,u;\beta)Y_i(s)ds}.$$

Note that $W_{ni}(s,t,u;\beta)$ is similar to the bivariate equivalent kernel for the local linear regression Fan and Gijbels (1996). Then, it is easy to verify that $W_{ni}(s,t,u;\beta)$ satisfies the following basic properties:

(1) $\sum_{i=1}^n \int_0^\tau W_{ni}(s,t,u;\beta)Y_i(s)ds = 1$;

(2) $\sum_{i=1}^n \int_0^\tau W_{ni}(s,t,u;\beta)(s-t)Y_i(s)ds = 0$;

(3) $\sum_{i=1}^n \int_0^\tau W_{ni}(s,t,u;\beta)(U_i(s)-u)Y_i(s)ds = 0$;

(4) $\widehat{\lambda}(t,u;\beta)=\sum_{i=1}^{n}\int_0^\tau W_{ni}(s,t,u;\beta)dN_i(s)$.

To prove Proposition 3.1, we first introduce the following Lemmas 3.1-3.3.

Lemma 3.1. *Under Assumptions 3.2 and Assumption 3.4, we have, for* $j,k=0,1,2,3$

$$\sup_{(t,\boldsymbol{z},\beta)\in\mathcal{A}_n}|S_{jk}(t,\beta^T\boldsymbol{z};\beta)-E[S_{jk}(t,\beta^T\boldsymbol{z};\beta)]|=O_p(h_1^jh_2^k\sqrt{\log n/(nh_1h_2)}),$$

as $h_1\to 0, h_2\to 0$ *and* $(nh_1h_2)/\log n\to\infty$, *where* $\mathcal{A}_n=\{(t,\boldsymbol{z},\beta):(t,\boldsymbol{z},\beta)\in[0,\tau]\times\mathcal{Z}\times\mathcal{R}^p,\|\beta-\beta_0\|\leqslant cn^{-1/2}\}$ *for a given constant* $c>0$.

Lemma 3.2. *When Assumptions 3.2 and Assumption 3.4 hold, as* $h_1\to 0,h_2\to 0$ *and* $(nh_1h_2)/\log n\to\infty$, *for* $j,k=0,1,2,3$, *we have*

$$\begin{aligned}S_{jk}(t,u;\beta) &= \phi_\beta(t,u)h_1^jh_2^k\mu_j\mu_k+\frac{\partial\phi_\beta(t,u)}{\partial t}h_1^{j+1}h_2^k\mu_{j+1}\mu_k+\frac{\partial\phi_\beta(t,u)}{\partial u}h_1^jh_2^{k+1}\mu_j\mu_{k+1}\\ &\quad+O(h_1^jh_2^kc_n),\end{aligned}$$

uniformly for $t\in[0,\tau],u\in\mathcal{U}_\beta$ *and* $\|\beta-\beta_0\|\leqslant cn^{-1/2}$.

Lemma 3.3. *Under Assumptions 3.1, 3.2 and 3.4, as* $h_1\to 0,h_2\to 0$ *and* $(nh_1h_2)/\log n\to\infty$, *we have*

$$\begin{aligned}\hat{\lambda}(t,\beta^T\boldsymbol{z};\beta) &= \lambda_0(t,\beta_0^T\boldsymbol{z})+\lambda_{01}(t,\beta_0^T\boldsymbol{z})(\beta_{-1}-\beta_{-1}^0)^T\boldsymbol{J}^T(\beta_{-1}^0)\Big[\boldsymbol{z}-\frac{E[Y(t)\boldsymbol{Z}(t)|U(t)=u]}{E[Y(t)|U(t)=u]}\Big]\\ &\quad+\lambda_{20}(t,\beta_0^T\boldsymbol{z})h_1^2\mu_2+\lambda_{02}h_2^2\mu_2+\mathcal{E}_{n,1}^\beta(t,\boldsymbol{z})\\ &\quad+O(h_1^3+h_2^3+h_2\|\beta_0-\beta\|+\|\beta_0-\beta\|^2+c_n),\\ \hat{\lambda}_{10}(t,u;\beta)h_1 &= \lambda_{10}(t,\beta_0^T\boldsymbol{z})h_1+\mathcal{E}_{n,2}^\beta(t,\boldsymbol{z})+O(c_n),\\ \hat{\lambda}_{01}(t,u;\beta)h_2 &= \lambda_{01}(t,\beta_0^T\boldsymbol{z})h_2+\mathcal{E}_{n,3}^\beta(t,\boldsymbol{z})+O(c_n),\end{aligned}\tag{3.29}$$

uniformly for $t\in[0,\tau],u\in\mathcal{U}_\beta$ *and* $\|\beta-\beta_0\|\leqslant cn^{-1/2}$, *where*

$$\begin{pmatrix}\mathcal{E}_{n,1}^\beta(t,\boldsymbol{z})\\ \mathcal{E}_{n,2}^\beta(t,\boldsymbol{z})\\ \mathcal{E}_{n,3}^\beta(t,\boldsymbol{z})\end{pmatrix}=\frac{1}{n\phi(t,\beta^T\boldsymbol{z})}\sum_{i=1}^n\int_0^\tau K_{\boldsymbol{h}}(s-t,\beta^T(\boldsymbol{Z}_i(s)-\boldsymbol{z}))\begin{pmatrix}1\\ (s-t)/h_1\\ \beta^T(\boldsymbol{Z}_i(s)-\boldsymbol{z})/h_2\end{pmatrix}dM_i(s).$$

Proof of Proposition 3.1. Since $M_i(t)$ is a martingale process with the filtration $\mathcal{F}_{t,i}$ and $\mathcal{F}_t = \vee_{i=1}^n \mathcal{F}_{t,i}$, by a similar argument of Lemma 3.2, we have

$$\sup_{t\in[0,\tau],\ \mathbf{z}\in\mathcal{Z}} \left| \frac{\sum_{i=1}^n \int_0^\tau K_{\mathbf{h}}(s-t, \beta_0^T(\mathbf{Z}_i(s)-\mathbf{z}))dM_i(s)}{nE[Y(t)|U(t)=\beta_0^T\mathbf{z}]f_{\beta_0}(\beta_0^T\mathbf{z})} \right| = O_p(\sqrt{\log n/(nh_1h_2)}).$$

By Lemma 3.3, we obtain $\hat{\lambda}(t, \beta_0^T\mathbf{z}; \beta_0) = \lambda(t, \beta_0^T\mathbf{z}) + O_p(c_n)$, uniformly for $t \in [0,\tau],\ \mathbf{z} \in \mathcal{Z}$. □

Proof of Proposition 3.2. Let $p_{X,\delta,\mathbf{Z}(X)}(x,\delta,\mathbf{z})$ be the joint density function of $(X,\delta,\mathbf{Z}(X))$. Then, in model (1), the likelihood of one random observation $(x,\delta,\mathbf{z})$ is given by

$$\{\lambda(x,\beta^T\mathbf{z})\}^\delta \exp\Big\{-\int_0^x \lambda(s,\beta^T\mathbf{z})ds\Big\}\{p_{C|\mathbf{Z}(x)}(x|\mathbf{z})\}^{1-\delta}\Big\{\int_x^\infty p_{C|\mathbf{Z}(x)}(u|\mathbf{z})du\Big\}^\delta p_{\mathbf{Z}(x)}(\mathbf{z}), \tag{3.30}$$

where $\beta = \left((1-\|\beta_{-1}\|^2)^{1/2}, \beta_{-1}^T\right)^T$, $p_{\mathbf{Z}(x)}(\mathbf{z})$ denotes the marginal density of $\mathbf{Z}(x)$ and $p_{C|\mathbf{Z}(x)}(u|\mathbf{z})$ denotes the conditional density of censoring time C given $\mathbf{Z}(x)$. The density (3.30) is equivalent to

$$\begin{aligned} p_{X,\delta,\mathbf{Z}(X)}(x,\delta,\mathbf{z}) &= \lambda(x,\beta^T\mathbf{z})^\delta \exp\Big\{-\int_0^x \lambda(s,\beta^T\mathbf{z})ds\Big\}\lambda_{C|\mathbf{Z}(x)}(x|\mathbf{z})^{1-\delta} \\ &\quad \times \exp\Big\{-\Lambda_{C|\mathbf{Z}(x)}(x|\mathbf{z})\Big\}p_{\mathbf{Z}(x)}(\mathbf{z}), \end{aligned} \tag{3.31}$$

where

$$\lambda_{C|\mathbf{Z}(t)}(t|\mathbf{z}) = \lim_{h\to 0}\frac{P\{t\leqslant C<t+h|C\geqslant t, \mathbf{Z}(t)=\mathbf{z}\}}{h}, \quad \Lambda_{C|\mathbf{Z}(t)}(t|\mathbf{z}) = \int_0^t \lambda_{C|\mathbf{Z}(v)}(v|\mathbf{z})dv.$$

The log of the density in (3.31) is

$$\log p_{X,\delta,\mathbf{Z}(X)}(x,\delta,\mathbf{z}) = \delta\log\lambda(x,\beta^T\mathbf{z}) - \int_0^x \lambda(s,\beta^T\mathbf{z})ds + (1-\delta)\log\lambda_{C|\mathbf{Z}(x)}(x|\mathbf{z})$$

$$-\Lambda_{C|\mathbf{Z}(x)}(x|\mathbf{z}) + \log p_{\mathbf{Z}(x)}(\mathbf{z}). \tag{3.32}$$

Note that β_{-1} is the $p-1$ dimensional parameters of interest, and $\lambda(x,u)$, $\lambda_{C|\mathbf{Z}(x)}(x|\mathbf{z})$ and $p_{\mathbf{Z}(x)}(\mathbf{z})$ are the nuisance parameters. Denote the nuisance tangent space corresponding to $\lambda(x,u)$, $\lambda_{C|\mathbf{Z}(x)}(x|\mathbf{z})$ and $p_{\mathbf{Z}(x)}(\mathbf{z})$ to be Λ_{1s}, Λ_{2s} and Λ_{3s}, respectively. Obviously, the nuisance tangent space can be written as a direct sum of three orthogonal spaces

$$\Lambda = \Lambda_{1s} \oplus \Lambda_{2s} \oplus \Lambda_{3s}.$$

Let $dM_C(t,\mathbf{Z}(t)) = dI(X \leqslant t, \delta = 0) - I(X \geqslant t)\lambda_{0C|\mathbf{Z}(t)}(t|\mathbf{Z}(t))dt$ where $\lambda_{0C|\mathbf{Z}(t)}(t|\mathbf{Z}(t))$ denotes the true conditional hazard function of C given $\mathbf{Z}(t)$. By similar arguments in Chapter 5 of Tsiatis (2006), we can obtain

$$\Lambda_{2s} = \Big\{ \int_0^\tau \alpha(s,\mathbf{Z}(s))dM_C(s,\mathbf{Z}(s)), \textit{for all } \alpha(s,\mathbf{z}) \Big\},$$

$$\Lambda_{3s} = \Big\{ \alpha(\mathbf{Z}(X)) : E[\alpha(\mathbf{Z}(X))] = 0, \ \textit{for all } \alpha(\mathbf{z}) \Big\},$$

where $\alpha(s,\mathbf{z})$ and $\alpha(\mathbf{z})$ are $p-1$ dimensional measurable functions.

In order to derive the space Λ_{1s}, we consider the parametric submodel

$$\lambda(x,u,\gamma_1) = \lambda(x,u)\exp\{\gamma_1^T \alpha(x,u)\},$$

for any arbitrary $p-1$ dimensional measurable function $\alpha(x,u)$ of (x,u), for $u \in \mathcal{U}_\beta$. This parametric submodel is valid because it contains the true model (i.e., when $\gamma_1 = 0$) and $\lambda(x,u,\gamma_1)$ is positive. Subsisting this parametric submodel into (3.32), taking derivatives with respect to γ_1 and setting $\gamma_1 = 0$ and $\beta = \beta_0$, the score function is

$$\begin{aligned} S_{\gamma_1}(X,\delta,\mathbf{Z}(X)) &= \delta\alpha(X,\beta_0^T\mathbf{Z}(X)) - \int_0^X \lambda(s,\beta_0^T\mathbf{Z}(s))\alpha(s,\beta_0^T\mathbf{Z}(s))ds \\ &= \int_0^\tau \alpha(s,\beta_0^T\mathbf{Z}(s))dM(s,\beta_0^T\mathbf{Z}(s)), \end{aligned}$$

where $dM(s, \beta_0^T \mathbf{Z}(s)) = dM(s) = dN_i(s) - Y(s)\lambda(s, \beta_0^T \mathbf{Z}(s))ds$. We hence conjecture that

$$\Lambda_{1s} = \Big\{ \int_0^\tau \alpha(s, \beta_0^T \mathbf{Z}(s)) dM(s, \beta_0^T \mathbf{Z}(s)), \text{for all } \alpha(s,u), u \in \mathcal{U}_{\beta_0} \Big\},$$

where $\alpha(s,u)$ is $p-1$ dimensional measurable function. To justify this conjecture, we need to verify that, for any parametric submodel $\lambda(s,u,\gamma_1)$ (when $\gamma_1 = \gamma_{10}$, $\lambda(s,u,\gamma_{10}) = \lambda(s,u)$), the linear space spanned by its score vector with respect to γ_1 belongs to Λ_{1s}. Substituting the submodel $\lambda(s,u,\gamma_1)$ into (3.32) and setting $\gamma_1 = \gamma_{10}$ and $\beta = \beta_0$, the score function is

$$\begin{aligned} S_{\gamma_1}(X, \delta, \mathbf{Z}(X)) &= \frac{\partial}{\partial \gamma_1} \Big\{ \delta \log \lambda(X, \beta_0^T \mathbf{Z}(X), \gamma_{10}) - \int_0^X \lambda(s, \beta_0^T \mathbf{Z}(s), \gamma_{10}) ds \Big\} \\ &= \int_0^\tau \frac{\partial \log \lambda(s, \beta_0^T \mathbf{Z}(s), \gamma_{10})}{\partial \gamma_1} dM(s, \beta_0^T \mathbf{Z}(s)). \end{aligned}$$

Thus, for any $(p-1) \times r_1$ dimensional matrix B, where r_1 is the dimension of γ_1, we have $BS_{\gamma_1}(X, \delta, \mathbf{Z}(X)) \in \Lambda_{1s}$.

To find the orthogonal complement of the nuisance tangent space Λ, we can use the method in Theorem 5.5 of Tsiatis (2006) to construct the space

$$\Lambda_{1s}^* = \Big\{ \int_0^\tau \alpha(s, \mathbf{Z}(s)) dM(s, \mathbf{Z}(s)), \textit{for all } \alpha(s, \mathbf{z}) \Big\},$$

where $dM(s, \mathbf{Z}(s)) = dN_i(s) - Y(s)\lambda_{0T|\mathbf{Z}(s)}(s|\mathbf{Z}(s))ds$, and $\lambda_{0T|\mathbf{Z}(s)}(s|\mathbf{Z}(s))$ is the true conditional hazard function of T given $\mathbf{Z}(s)$ with no restrictions on the distribution of $(X, \delta, \mathbf{Z}(X))$. Theorem 5.5 in Tsiatis (2006) shows that $\Lambda^\perp \in \Lambda_{1s}^*$. Thus, elements of $\Lambda^\perp$ belong to Λ_{1s}^* and are orthogonal to Λ_{1s}. Therefore, to identify elements of $\Lambda^\perp$, we can take an arbitrary element of Λ_{1s}^*, denoted by

$$\int_0^\tau \alpha(s, \mathbf{Z}(s)) dM(s, \mathbf{Z}(s))$$

and find its residual after projecting it onto Λ_{1s}. Toward that end, we can derive $\alpha^*(s,u)$, $u \in \mathcal{U}_{\beta_0}$, so that

$$\int_0^\tau \alpha(s,\mathbf{Z}(s))dM(s,\mathbf{Z}(s)) - \int_0^\tau \alpha^*(s,\beta_0^T\mathbf{Z}(s))dM(s,\beta_0^T\mathbf{Z}(s))$$

is orthogonal to every element of Λ_{1s}. That is,

$$E[\int_0^\tau [\alpha(s,\mathbf{Z}(s)) - \alpha^*(s,\beta_0^T\mathbf{Z}(s))]^T dM(s,\mathbf{Z}(s)) \int_0^\tau \alpha(s,\beta_0^T\mathbf{Z}(s))dM(s,\beta_0^T\mathbf{Z}(s))] = 0$$

for any arbitrary $p-1$ dimensional measurable function $\eta(s,u)$ of (s,u), $u \in \mathcal{U}_{\beta_0}$. By the theory of martingale stochastic integrals, we have

$$\begin{aligned}
& E\Big[\int_0^\tau [\alpha(s,\mathbf{Z}(s)) - \alpha^*(s,\beta_0^T\mathbf{Z}(s))]^T dM(s,\mathbf{Z}(s)) \int_0^\tau \eta(s,\beta_0^T\mathbf{Z}(s))dM(s,\beta_0^T\mathbf{Z}(s))\Big] \\
= \; & E\Big[\int_0^\tau [\alpha(s,\mathbf{Z}(s)) - \alpha^*(s,\beta_0^T\mathbf{Z}(s))]^T \eta(s,\beta_0^T\mathbf{Z}(s))Y(s)\lambda(s,\beta_0^T\mathbf{Z}(s))ds\Big] \\
= \; & \int_0^\tau E\Big[\{\alpha(s,\mathbf{Z}(s)) - \alpha^*(s,\beta_0^T\mathbf{Z}(s))\}^T \eta(s,\beta_0^T\mathbf{Z}(s))Y(s)\lambda(s,\beta_0^T\mathbf{Z}(s))\Big]ds \\
= \; & \int_0^\tau E\Big[E\Big\{[\alpha(s,\mathbf{Z}(s)) - \alpha^*(s,\beta_0^T\mathbf{Z}(s))]^T Y(s)|\beta_0^T\mathbf{Z}(s)\Big\}\eta(s,\beta_0^T\mathbf{Z}(s))\lambda(s,\beta_0^T\mathbf{Z}(s))\Big]ds.
\end{aligned}$$

Since $\eta(s,u)$ is arbitrary, above equation implies that

$$E\Big[\{\alpha(s,\mathbf{Z}(s)) - \alpha^*(s,\beta_0^T\mathbf{Z}(s))\}^T Y(s)|\beta_0^T\mathbf{Z}(s)\Big] = 0. \tag{3.33}$$

Solving (3.33), we obtain

$$\alpha^*(s,\beta_0^T\mathbf{Z}(s)) = \frac{E[\alpha(s,\mathbf{Z}(s))Y(s)|\beta_0^T\mathbf{Z}(s)]}{E[Y(s)|\beta_0^T\mathbf{Z}(s)]}.$$

Therefore, the space orthogonal to the nuisance tangent space is given by

$$\Lambda^\perp = \Big\{\int_0^\tau \Big[\alpha(s,\mathbf{Z}(s)) - \frac{E[\alpha(s,\mathbf{Z}(s))Y(s)|\beta_0^T\mathbf{Z}(s)]}{E[Y(s)|\beta_0^T\mathbf{Z}(s)]}\Big]dM(s) \;\; for\ all\ \alpha(s,\mathbf{z})\Big\}, \tag{3.34}$$

for any arbitrary $p-1$ dimensional measurable function $\alpha(s,\mathbf{z})$ of $(s,\mathbf{z})$, $\mathbf{z}\in\mathcal{Z}$.

The efficient score β_{-1} is obtained by computing $S_{\beta_{-1}}(X,\delta,\mathbf{Z}(X))$ and projecting this onto Λ . It is straightforward to calculate

$$\begin{aligned} S_{\beta_{-1}}(X,\delta,\mathbf{Z}(X)) &= \frac{\partial}{\partial\beta_{-1}}\Big\{\delta\log\lambda(X,\beta_0^T\mathbf{Z}(X))-\int_0^\tau\lambda(s,\beta_0^T\mathbf{Z}(s))Y(s)ds\Big\} \\ &= \mathbf{J}^T(\beta_{-1}^0)\delta\frac{\lambda_{01}(X,\beta_0^T\mathbf{Z}(X))}{\lambda(X,\beta^T\mathbf{Z}(s))}\mathbf{Z}(s)-\int_0^\tau\lambda_{01}(s,\beta_0^T\mathbf{Z}(s))\mathbf{Z}(s)Y(s)ds \\ &= \mathbf{J}^T(\beta_{-1}^0)\int_0^\tau\mathbf{Z}(s)\frac{\lambda_{01}(s,\beta_0^T\mathbf{Z}(s))}{\lambda(s,\beta_0^T\mathbf{Z}(s))}dM(s,\beta_0^T\mathbf{Z}(s)). \end{aligned}$$

Note that $S_{\beta_{-1}}(X,\delta,\mathbf{Z}(X))$ is an element of Λ_{1s}^*, with

$$\alpha(s,\mathbf{Z}(s))=\mathbf{J}^T(\beta_{-1}^0)\mathbf{Z}(s)\frac{\lambda_{01}(s,\beta_0^T\mathbf{Z}(s))}{\lambda(s,\beta_0^T\mathbf{Z}(s))}.$$

Therefore, the efficient score, derived as the residual after projecting $S_{\beta_{-1}}(X,\delta,\mathbf{Z}(X))$ onto the nuisance tangent space, is given as

$$S_{\text{eff}}(X,\delta,\mathbf{Z}(X))=\int_0^\tau\Big[\mathbf{Z}(s)-\frac{E[\mathbf{Z}(s)Y(s)|\beta_0^T\mathbf{Z}(s)]}{E[Y(s)|\beta_0^T\mathbf{Z}(s)]}\Big]\frac{\lambda_{01}(s,\beta_0^T\mathbf{Z}(s))}{\lambda(s,\beta_0^T\mathbf{Z}(s))}dM(s,\beta_0^T\mathbf{Z}(s)).$$

Thus, the information matrix of β_{-1}, namely $E[S_{\text{eff}}(X,\delta,\mathbf{Z}(X))^{\otimes 2}]$ is

$$\mathbf{J}^T(\beta_{-1}^0)E\Big[\int_0^\tau\Big[\mathbf{Z}(s)-\frac{E[Y(s)\mathbf{Z}(s)|\beta_0^T\mathbf{Z}(s)]}{E[Y(s)|\beta_0^T\mathbf{Z}(s)]}\Big]^{\otimes 2}\frac{\lambda_{01}^2(s,\beta_0^T\mathbf{Z}(s))}{\lambda(s,\beta_0^T\mathbf{Z}(s))}Y(s)ds\Big]\mathbf{J}(\beta_{-1}^0).$$

□

Proof of Proposition 3.3. From the proof of Lemma 3.3, we have

$$\begin{aligned} &\hat\lambda(t,\beta^T\mathbf{z};\beta)-\hat\lambda(t,\beta_0^T\mathbf{z};\beta_0) \\ =\ &\lambda_{01}(t,\beta_0^T\mathbf{z})(\beta_{-1}-\beta_{-1}^0)^T\mathbf{J}^T(\beta_{-1}^0)\Big[\mathbf{z}-\frac{E[Y(t)\mathbf{Z}(t)|U(t)=u]}{E[Y(t)|U(t)=u]}\Big]+[\mathbf{R}_n(\beta)-\mathbf{R}_n(\beta_0)] \end{aligned}$$

$$+\{\mathcal{E}_{n,1}^{\beta}(t,\mathbf{z})-\mathcal{E}_{n,1}^{\beta_0}(t,\mathbf{z})\}+O(h_2\|\beta_0-\beta\|+\|\beta_0-\beta\|^2), \tag{3.35}$$

where

$$\begin{aligned}\mathbf{R}_n(\beta) &= \frac{1}{2}\lambda_{02}(t,\beta_0^T\mathbf{z})\sum_{i=1}^n\int_0^\tau W_{ni}(s,t,\beta^T\mathbf{z};\beta)(U_i(s)-\beta^T\mathbf{z})^2Y_i(s)ds\\ &\quad+\lambda_{20}(t,\beta_0^T\mathbf{z})\sum_{i=1}^n\int_0^\tau W_{ni}(s,t,\beta^T\mathbf{z};\beta)(s-t)(U_i(s)-\beta^T\mathbf{z})]Y_i(s)ds\\ &\quad+\frac{1}{2}\lambda_{20}(t,\beta_0^T\mathbf{z})\sum_{i=1}^n\int_0^\tau W_{ni}(s,t,\beta^T\mathbf{z};\beta)(s-t)^2Y_i(s)ds\\ &= \mathbf{R}_{n,1}(\beta)+\mathbf{R}_{n,2}(\beta)+\mathbf{R}_{n,3}(\beta).\end{aligned}$$

Note that the term $\mathbf{R}_n(\beta)$ comes from (3.63) in the proof of Lemma 3.3.

Firstly, we consider the difference

$$\begin{aligned}&\mathbf{R}_{n,1}(\beta)-\mathbf{R}_{n,1}(\beta_0)\\ =\ &\lambda_{02}(t,\beta_0^T\mathbf{z})\Big[\frac{1}{n\phi(t,\beta^T\mathbf{z})}\sum_{i=1}^n\int_0^\tau K_{\mathbf{h}}(s-t,\beta^T(\mathbf{Z}_i(s)-\mathbf{z}))(U_i(s)-\beta^T\mathbf{z})^2Y_i(s)ds\\ &-\frac{1}{n\phi(t,\beta_0^T\mathbf{z})}\sum_{i=1}^n\int_0^\tau K_{\mathbf{h}}(s-t,\beta_0^T(\mathbf{Z}_i(s)-\mathbf{z}))(U_{0i}(s)-\beta_0^T\mathbf{z})^2Y_i(s)ds\Big]\{1+o_p(1)\}.\end{aligned}$$

By Taylor expansion,

$$\begin{aligned}&K_{\mathbf{h}}(s-t,\beta^T(\mathbf{Z}_i(s)-\mathbf{z}))(U_i(s)-\beta^T\mathbf{z})^2-K_{\mathbf{h}}(s-t,\beta_0^T(\mathbf{Z}_i(s)-\mathbf{z}))(U_{0i}(s)-\beta_0^T\mathbf{z})^2\\ =\ &[h_2^{-1}K_{\mathbf{h},01}(s-t,\beta_0^T(\mathbf{Z}_i(s)-\mathbf{z}))(U_{0i}(s)-\beta_0^T\mathbf{z})^2\\ &+K_{\mathbf{h}}(s-t,\beta_0^T(\mathbf{Z}_i(s)-\mathbf{z}))(U_{0i}(s)-\beta_0^T\mathbf{z})](\beta_{-1}-\beta_{-1}^0)^T\mathbf{J}^T(\beta_{-1}^0)(\mathbf{Z}_i(s)-\mathbf{z})\\ &+O_p(\|\beta_{-1}-\beta_{-1}^0\|_2^2),\end{aligned}$$

where $K_{\mathbf{h},01}(t,u)=\partial K_{\mathbf{h}}(t,u)/\partial u$. This leads to

$$\mathbf{R}_{n,1}(\beta)-\mathbf{R}_{n,1}(\beta_0)=O_p(h_2^2\|\beta_{-1}-\beta_{-1}^0\|)+O_p(c_n\|\beta_{-1}-\beta_{-1}^0\|(1+h_2)). \tag{3.36}$$

By similar calculations, we can derive that

$$\mathbf{R}_{n,2}(\beta) - \mathbf{R}_{n,2}(\beta_0) = O_p(h_1^2\|\beta_{-1} - \beta_{-1}^0\|) + O_p(c_n\|\beta_{-1} - \beta_{-1}^0\|h_1^2 h_2^{-1}) \tag{3.37}$$

and

$$\mathbf{R}_{n,3}(\beta) - \mathbf{R}_{n,3}(\beta_0) = O_p(h_1^2\|\beta_{-1} - \beta_{-1}^0\|). \tag{3.38}$$

By Taylor expansion, we have

$$\begin{aligned}
&\mathcal{E}_{n,1}^{\beta}(t,\mathbf{z}) - \mathcal{E}_{n,1}^{\beta_0}(t,\mathbf{z}) \\
= \;& \frac{1}{n\phi(t,\beta^T\mathbf{z})}\sum_{i=1}^n\int_0^\tau K_{\mathbf{h}}(s-t,\beta^T(\mathbf{Z}_i(s)-\mathbf{z}))dM_i(s) \\
& -\frac{1}{n\phi(t,\beta_0^T\mathbf{z})}\sum_{i=1}^n\int_0^\tau K_{\mathbf{h}}(s-t,\beta_0^T(\mathbf{Z}_i(s)-\mathbf{z}))dM_i(s) \\
= \;& \frac{1}{n\phi(t,\beta_0^T\mathbf{z})}\sum_{i=1}^n\int_0^\tau [K_{\mathbf{h}}(s-t,\beta^T(\mathbf{Z}_i(s)-\mathbf{z})) - K_{\mathbf{h}}(s-t,\beta_0^T(\mathbf{Z}_i(s)-\mathbf{z}))]dM_i(s) \\
& -\frac{\phi(t,\beta^T\mathbf{z}) - \phi(t,\beta_0^T\mathbf{z})}{n\phi(t,\beta_0^T\mathbf{z})}\sum_{i=1}^n\int_0^\tau K_{\mathbf{h}}(s-t,\beta_0^T(\mathbf{Z}_i(s)-\mathbf{z}))dM_i(s) + O_p(\|\beta_{-1}-\beta_{-1}^0\|^2) \\
= \;& O_p(\sqrt{\log n/(nh_1h_2)}\|\beta_{-1}-\beta_{-1}^0\|h_2^{-1}).
\end{aligned} \tag{3.39}$$

Together with (3.35)-(3.39), the proof of Proposition 3.3 is completed by the definition of derivative. □

Lemma 3.4. *Under Assumption 3.1 to 3.4, as* $nh_1^8 \to 0$ *and* $nh_2^8 \to 0$, $nh_1h_2^3/\log n \to \infty$ *and* $nh_1^2h_2^2 \to \infty$, *we have*

$$\begin{aligned}
\frac{1}{\sqrt{n}}\frac{\partial\hat{\ell}_n(\beta_0)}{\partial\beta_{-1}} = \;& \frac{1}{\sqrt{n}}\boldsymbol{J}^T(\beta_{-1}^0)\sum_{i=1}^n\int_0^\tau\Big[\boldsymbol{Z}_i(s) - \frac{E[Y(s)\boldsymbol{Z}(s)|U_{0i}(s)]}{E[Y(s)|U_{0i}(s)]}\Big]\frac{\lambda_{01}(s,U_{0i}(s))}{\lambda(s,U_{0i}(s))}dM_i(s) \\
& +o_p(1).
\end{aligned}$$

Lemma 3.5. *Under Assumption 3.1 to 3.4, we have*

$$-\frac{1}{n}\frac{\partial^2\hat{\ell}_n(\beta_0)}{\partial\beta_{-1}\partial\beta_{-1}^T} = \boldsymbol{J}^T(\beta_{-1}^0)E\Big[\int_0^\tau\Big\{\boldsymbol{Z}(s)-\frac{E[Y(s)\boldsymbol{Z}(s)|\beta_0^T\boldsymbol{Z}(s)]}{E[Y(s)|\beta_0^T\boldsymbol{Z}(s)]}\Big\}^{\otimes 2} \times\frac{\lambda_{01}^2(s,\beta_0^T\boldsymbol{Z}(s))}{\lambda(s,\beta_0^T\boldsymbol{Z}(s))}Y(s)ds\Big]\boldsymbol{J}(\beta_{-1}^0)+o_p(1).$$

Proof of Theorem 3.1. (i) To prove the consistency of $\hat{\beta}_{-1}$, we only need to verify Exp. (5.8) in Theorem 5.7 of van der Vaart (2000). Specifically, we can show that, for any ε

$$\sup_{\beta_{-1}:\|\beta_{-1}-\beta_{-1}^0\|\geqslant\varepsilon}\ell(\beta_{-1})<\ell(\beta_{-1}^0), \tag{3.40}$$

$$\sup_{\beta_{-1}\in\Theta_{-1}}|\frac{1}{n}\hat{\ell}_n(\beta_{-1})-\ell(\beta_{-1})|=o_p(1), \tag{3.41}$$

where

$$\ell(\beta_{-1})=E\Big[\delta\log\lambda(X,\beta^T\mathbf{Z}(X))-\int_0^\tau\lambda(s,\beta^T\mathbf{Z}(s))Y(s)ds\Big].$$

Firstly, consider (3.40). Note that

$$\begin{aligned}
&\ell(\beta_{-1})-\ell(\beta_{-1}^0)\\
=\;&E\Big[\int_0^\tau\{\lambda(s,\beta_0^T\mathbf{Z}(s))\log\lambda(s,\beta^T\mathbf{Z}(s))-\lambda(s,\beta^T\mathbf{Z}(s))\}Y(s)ds\Big]\\
&-E\Big[\int_0^\tau\{\lambda(s,\beta_0^T\mathbf{Z}(s))\log\lambda(s,\beta_0^T\mathbf{Z}(s))-\lambda(s,\beta_0^T\mathbf{Z}(s))\}Y(s)ds\Big]\\
=\;&E\Big[\int_0^\tau\Big\{\lambda(s,\beta_0^T\mathbf{Z}(s))\log\frac{\lambda(s,\beta^T\mathbf{Z}(s))}{\lambda(s,\beta_0^T\mathbf{Z}(s))}-\{\lambda(s,\beta^T\mathbf{Z}(s))-\lambda(s,\beta_0^T\mathbf{Z}(s))\}\Big\}Y(s)ds\Big].
\end{aligned}$$

By the inequality $\log x<x-1$, for $x\neq 1$, we have

$$\lambda(s,\beta_0^T\mathbf{Z}(s))\log\frac{\lambda(s,\beta^T\mathbf{Z}(s))}{\lambda(s,\beta_0^T\mathbf{Z}(s))}-\{\lambda(s,\beta^T\mathbf{Z}(s))-\lambda(s,\beta_0^T\mathbf{Z}(s))\}<0,$$

if $\lambda(s, \beta^T \mathbf{Z}(s)) \neq \lambda(s, \beta_0^T \mathbf{Z}(s))$. Thus, under the condition (C7), we obtain that

$$\sup_{\beta_{-1}:\|\beta_{-1}-\beta_{-1}^0\|\geqslant\varepsilon} \ell(\beta_{-1}) < \ell(\beta_{-1}^0).$$

For Exp. (3.41), we can consider

$$\begin{aligned} \sup_{\beta_{-1}\in\Theta_{-1}} |\frac{1}{n}\hat{\ell}_n(\beta_{-1}) - \frac{1}{n}\ell(\beta_{-1})| &\leqslant \sup_{\beta_{-1}\in\Theta_{-1}} |\frac{1}{n}\hat{\ell}_n(\beta_{-1}) - \frac{1}{n}\ell_n(\beta_{-1})| \\ &\quad + \sup_{\beta_{-1}\in\Theta_{-1}} |\frac{1}{n}\ell_n(\beta_{-1}) - \ell(\beta_{-1})|. \end{aligned} \tag{3.42}$$

By Proposition 3.1, the first term on the right-hand of (3.42) is

$$\begin{aligned} \frac{1}{n}\hat{\ell}_n(\beta_{-1}) - \frac{1}{n}\ell_n(\beta_{-1}) &= \frac{1}{n}\sum_{i=1}^n \int_0^\tau \log \frac{\hat{\lambda}(s, \beta^T \mathbf{Z}_i(s); \beta)}{\lambda(s, \beta^T \mathbf{Z}(s))} dN_i(s) \\ &\quad -\frac{1}{n}\sum_{i=1}^n \int_0^\tau \Big\{\hat{\lambda}(s, \beta^T \mathbf{Z}_i(s); \beta)) - \lambda(s, \beta^T \mathbf{Z}(s))\Big\} Y_i(s) ds \\ &= o_p(1). \end{aligned}$$

By the uniform law of large numbers, we can obtain

$$\sup_{\beta_{-1}\in\Theta} |\frac{1}{n}\ell_n(\beta_{-1}) - \ell(\beta_{-1})| = o_p(1).$$

Therefore, by Theorem 5.7 of van der Vaart (2000), $\hat{\beta}_{-1}$ converges in probability to β_{-1}^0 .

(ii) By Taylor expansion

$$\frac{1}{n}\hat{\ell}_n(\hat{\beta}) = \frac{1}{n}\hat{\ell}_n(\beta_0) + (\hat{\beta}_{-1} - \beta_{-1}^0)^T \frac{1}{n}\frac{\partial \hat{\ell}_n(\beta_0)}{\partial \beta_{-1}} + \frac{1}{2}(\hat{\beta}_{-1} - \beta_{-1}^0)^T \frac{1}{n}\frac{\partial^2 \hat{\ell}_n(\beta^*)}{\partial \beta_{-1} \partial \beta_{-1}^T}(\hat{\beta}_{-1} - \beta_{-1}^0),$$

where β^* lies between $\hat{\beta}$ and β_0. This together with Lemma 3.4–3.5 and the consistency of $\hat{\beta}_{-1}$ yieds

$$\sqrt{n}(\hat{\beta}_{-1} - \beta_{-1}^0) = \Sigma^{-1}\frac{1}{\sqrt{n}}\frac{\partial \hat{\ell}_n(\beta_0)}{\partial \beta_{-1}} + o_p(1)$$

$$= \Sigma^{-1}\frac{1}{\sqrt{n}}\mathbf{J}^T(\beta_{-1}^0)\sum_{i=1}^{n}\int_0^\tau \Big\{\mathbf{Z}_i(s) - \frac{E[Y(s)\mathbf{Z}(s)|U_{0i}(s)]}{E[Y(s)|U_{0i}(s)]}\Big\}\frac{\lambda_{01}(s,U_{0i}(s))}{\lambda(s,U_{0i})}dM_i(s)$$

$$+o_p(1).$$

Then, by the martingale central limit theorem, we have

$$\sqrt{n}(\hat{\beta}_{-1} - \beta_{-1}^0) \xrightarrow{\mathrm{d}} N(0, \Sigma^{-1}).$$

(iii) By the delta-method, we can obtain the result of Theorem 3.1 (iii). □

Proof of Theorem 3.2. We first prove that $\widehat{\Sigma}$ is consistent to Σ. By similar arguments of Lemma 3.3, we obtain that

$$\sup_{\substack{s\in[0,\tau],\mathbf{z}\in\mathcal{Z},\\ \|\beta-\beta_0\|\leqslant cn^{-1/2}}} \|\widehat{E}[Y(s)\mathbf{Z}(s)|U(s)=\beta^T\mathbf{z}] - E[Y(s)\mathbf{Z}(s)|U(s)=\beta^T\mathbf{z})]\| = O_P(\sqrt{\log n/nh_3} + h_3^2),$$

$$\sup_{\substack{s\in[0,\tau],\mathbf{z}\in\mathcal{Z},\\ \|\beta-\beta_0\|\leqslant cn^{-1/2}}} |\widehat{E}[Y(s)|U(s)=\beta^T\mathbf{z}] - E[Y(s)|U(s)=\beta^T\mathbf{z}]| = O_P(\sqrt{\log n/nh_3} + h_3^2),$$

and

$$\sup_{\substack{s\in[0,\tau],\mathbf{z}\in\mathcal{Z},\\ \|\beta-\beta_0\|\leqslant cn^{-1/2}}} \Big\|\frac{\widehat{\lambda}_{01}^2(s,\beta^T\mathbf{z})}{\widehat{\lambda}(s,\beta^T\mathbf{z})} - \frac{\lambda_{01}^2(s,\beta_0^T\mathbf{z})}{\lambda(s,\beta_0^T\mathbf{z})}\Big\| = O_p(h_1^2+h_2^2) + O_p(c_n^2).$$

Thus,

$$\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \Big[\mathbf{Z}_i(s) - \frac{\widehat{E}[Y(s)\mathbf{Z}(s)|\hat{\beta}^T\mathbf{Z}_i(s)]}{\widehat{E}[Y(s)|\hat{\beta}^T\mathbf{Z}_i(s)]}\Big]^{\otimes 2}\frac{\widehat{\lambda}_{01}^2(s,\hat{\beta}^T\mathbf{Z}_i(s))}{\widehat{\lambda}(s,\hat{\beta}^T\mathbf{Z}_i(s))}Y_i(s)ds \longrightarrow \Sigma,$$

as $n\to\infty$. This together with Theorem 3.1 yields

$$W_n = n(\hat{\beta}-\beta_0)^T\Big[\mathbf{J}(\hat{\beta}_{-1})\widehat{\Sigma}^{-1}\mathbf{J}(\hat{\beta}_{-1})^T\Big]^{-1}(\hat{\beta}-\beta_0)$$

$$= \Big[\sqrt{n}[\mathbf{J}(\beta_{-1}^0)\Sigma^{-1}\mathbf{J}(\beta_{-1}^0)^T]^{-1/2}(\hat{\beta}-\beta_0)\Big]^T\Big[\sqrt{n}[\mathbf{J}(\beta_{-1}^0)\Sigma^{-1}\mathbf{J}(\beta_{-1}^0)^T]^{-1/2}(\hat{\beta}-\beta_0)\Big]$$

$$+O_P(\sqrt{\log n/nh_3}+h_3^2)+O_p(h_1^2+h_2^2)+O_p(c_n^2)\longrightarrow\chi^2(p),$$

under conditions of Theorem 3.1. □

Proof of Theorem 3.3. (i) By the definition of $\check{\beta}_{-1}$, we have

$$\begin{aligned}0 &= \mathbf{J}^T(\check{\beta}_{-1})\sum_{i=1}^n\int_0^\tau\Big[\mathbf{Z}_i(s)-\frac{\widehat{E}[Y(s)\mathbf{Z}(s)|\check{\beta}^T\mathbf{Z}_i(s)]}{\widehat{E}[Y(s)|\check{\beta}^T\mathbf{Z}_i(s)]}\Big]\frac{\widehat{\lambda}_{01}(s,\check{\beta}^T\mathbf{Z}_i(s);\check{\beta})}{\widehat{\lambda}(s,\check{\beta}^T\mathbf{Z}_i(s);\check{\beta})}\Big[dN_i(s)\\ &\quad -\widehat{\lambda}(s,\check{\beta}^T\mathbf{Z}_i(s);\check{\beta})Y_i(s)ds\Big],\end{aligned}\tag{3.43}$$

where $\check{\beta}=\Big((1-\|\check{\beta}_{-1}\|^2)^{1/2},\check{\beta}_{-1}\Big)^T$. Let

$$\begin{aligned}\mathbf{R}^*(\beta_{-1}) &= \sum_{i=1}^n\int_0^\tau\Big[\mathbf{Z}_i(s)-\frac{\widehat{E}[Y(s)\mathbf{Z}(s)|\beta^T\mathbf{Z}_i(s)]}{\widehat{E}[Y(s)|\beta^T\mathbf{Z}_i(s)]}\Big]\frac{\widehat{\lambda}_{01}(s,\beta^T\mathbf{Z}_i(s);\beta)}{\widehat{\lambda}(s,\beta^T\mathbf{Z}_i(s);\beta)}\Big[dN_i(s)\\ &\quad -\widehat{\lambda}(s,\beta^T\mathbf{Z}_i(s);\beta)Y_i(s)ds\Big].\end{aligned}$$

According to the following decomposition of the right-hand of above expression

$$\begin{aligned}&\sum_{i=1}^n\int_0^\tau\Big[\Big\{\mathbf{Z}_i(s)-\frac{E[Y(s)\mathbf{Z}(s)|U_i(s)]}{E[Y(s)|U_i(s)]}\Big\}-\Big\{\frac{\widehat{E}[Y(s)\mathbf{Z}(s)|U_i(s)]}{\widehat{E}[Y(s)|U_i(s)]}-\frac{E[Y(s)\mathbf{Z}(s)|U_i(s)]}{E[Y(s)|U_i(s)]}\Big\}\Big]\\ &\times\Big[\Big\{\frac{\widehat{\lambda}_{01}(s,\beta^T\mathbf{Z}_i(s);\beta)}{\widehat{\lambda}(s,\beta^T\mathbf{Z}_i(s);\beta)}-\frac{\lambda_{01}(s,U_i(s))}{\lambda(s,U_i(s))}\Big\}+\frac{\lambda_{01}(s,U_i(s))}{\lambda(s,U_i(s))}\Big]\\ &\times\Big[dM_i(s)+\{\lambda(s,U_{0i}(s))-\lambda(s,U_i(s))\}-\{\widehat{\lambda}(s,U_i(s);\beta)-\lambda(s,U_i(s))\}Y_i(s)ds\Big],\end{aligned}$$

some algebraic manipulation yields

$$\begin{aligned}&\mathbf{R}^*(\beta_{-1})\\ =\ &\sum_{i=1}^n\int_0^\tau\Big\{\mathbf{Z}_i(s)-\frac{E[Y(s)\mathbf{Z}(s)|U_i(s)]}{E[Y(s)|U_i(s)]}\Big\}\frac{\lambda_{01}(s,U_i(s))}{\lambda(s,U_i(s))}dM_i(s)\\ +\ &\sum_{i=1}^n\int_0^\tau\Big\{\mathbf{Z}_i(s)-\frac{E[Y(s)\mathbf{Z}(s)|U_i(s)]}{E[Y(s)|U_i(s)]}\Big\}\frac{\lambda_{01}(s,U_i(s))}{\lambda(s,U_i(s))}\Big[\lambda(s,U_{0i}(s))-\lambda(s,U_i(s))\Big]Y_i(s)ds\end{aligned}$$

$$
\begin{aligned}
&- \sum_{i=1}^{n}\int_0^\tau \Big\{\frac{\widehat{E}[Y(s)\mathbf{Z}(s)|U_i(s)]}{\widehat{E}[Y(s)|U_i(s)]} - \frac{E[Y(s)\mathbf{Z}(s)|U_i(s)]}{E[Y(s)|U_i(s)]}\Big\}\frac{\lambda_{01}(s,U_i(s))}{\lambda(s,U_i(s))}dM_i(s)\\
&- \sum_{i=1}^{n}\int_0^\tau \Big\{\frac{\widehat{E}[Y(s)\mathbf{Z}(s)|U_i(s)]}{\widehat{E}[Y(s)|U_i(s)]} - \frac{E[Y(s)\mathbf{Z}(s)|U_i(s)]}{E[Y(s)|U_i(s)]}\Big\}\frac{\lambda_{01}(s,U_i(s))}{\lambda(s,U_i(s))}\Big[\lambda(s,U_{0i}(s)) - \lambda(s,U_i(s))\Big]Y_i(s)ds\\
&+ \sum_{i=1}^{n}\int_0^\tau \Big\{\mathbf{Z}_i(s) - \frac{\widehat{E}[Y(s)\mathbf{Z}(s)|U_i(s)]}{\widehat{E}[Y(s)|U_i(s)]}\Big\}\Big[\frac{\widehat{\lambda}_{01}(s,U_i(s);\beta)}{\widehat{\lambda}(s,U_i(s);\beta)} - \frac{\lambda_{01}(s,U_i(s))}{\lambda(s,U_i(s))}\Big]dM_i(s)\\
&+ \sum_{i=1}^{n}\int_0^\tau \Big\{\mathbf{Z}_i(s) - \frac{\widehat{E}[Y(s)\mathbf{Z}(s)|U_i(s)]}{\widehat{E}[Y(s)|U_i(s)]}\Big\}\Big[\frac{\widehat{\lambda}_{01}(s,U_i(s);\beta)}{\widehat{\lambda}(s,U_i(s);\beta)} - \frac{\lambda_{01}(s,U_i(s))}{\lambda(s,U_i(s))}\Big]\Big[\lambda(s,U_{0i}(s)) - \lambda(s,U_i(s))\Big]Y_i(s)ds\\
&- \sum_{i=1}^{n}\int_0^\tau \Big\{\mathbf{Z}_i(s) - \frac{E[Y(s)\mathbf{Z}(s)|U_i(s)]}{E[Y(s)|U_i(s)]}\Big\}\frac{\lambda_{01}(s,U_i(s))}{\lambda(s,U_i(s))}\Big[\widehat{\lambda}(s,U_i(s);\beta) - \lambda(s,U_i(s))\Big]Y_i(s)ds\\
&+ \sum_{i=1}^{n}\int_0^\tau \Big\{\frac{\widehat{E}[Y(s)\mathbf{Z}(s)|U_i(s)]}{\widehat{E}[Y(s)|U_i(s)]} - \frac{E[Y(s)\mathbf{Z}(s)|U_i(s)]}{E[Y(s)|U_i(s)]}\Big\}\frac{\lambda_{01}(s,U_i(s))}{\lambda(s,U_i(s))}\Big[\widehat{\lambda}(s,U_i(s);\beta) - \lambda(s,U_i(s))\Big]Y_i(s)ds\\
&- \sum_{i=1}^{n}\int_0^\tau \Big\{\mathbf{Z}_i(s) - \frac{E[Y(s)\mathbf{Z}(s)|U_i(s)]}{E[Y(s)|U_i(s)]}\Big\}\Big[\frac{\widehat{\lambda}_{01}(s,U_i(s);\beta)}{\widehat{\lambda}(s,U_i(s);\beta)} - \frac{\lambda_{01}(s,U_i(s))}{\lambda(s,U_i(s))}\Big]\\
&\quad\times\Big[\widehat{\lambda}(s,U_i(s);\beta) - \lambda(s,U_i(s))\Big]Y_i(s)ds\\
&+ \sum_{i=1}^{n}\int_0^\tau \Big\{\frac{\widehat{E}[Y(s)\mathbf{Z}(s)|U_i(s)]}{\widehat{E}[Y(s)|U_i(s)]} - \frac{E[Y(s)\mathbf{Z}(s)|U_i(s)]}{E[Y(s)|U_i(s)]}\Big\}\Big[\frac{\widehat{\lambda}_{01}(s,U_i(s);\beta)}{\widehat{\lambda}(s,U_i(s);\beta)} - \frac{\lambda_{01}(s,U_i(s))}{\lambda(s,U_i(s))}\Big]\\
&\quad\times\Big[\widehat{\lambda}(s,U_i(s);\beta) - \lambda(s,U_i(s))\Big]Y_i(s)ds\\
&= \sum_{k=1}^{10}\mathbf{R}_k^*(\beta_{-1}).
\end{aligned}
\tag{3.44}
$$

Using a similar decomposition of (3.48), we have

$$
\begin{aligned}
\mathbf{R}_1^*(\check{\beta}_{-1}) &= \sum_{i=1}^{n}\int_0^\tau \Big\{\mathbf{Z}_i(s) - \frac{E[Y(s)\mathbf{Z}(s)|U_{0i}(s)]}{E[Y(s)|U_{0i}(s)]}\Big\}\frac{\lambda_{01}(s,U_{0i}(s))}{\lambda(s,U_{0i}(s))}dM_i(s)\\
&\quad + o_p(\sqrt{n}).
\end{aligned}
\tag{3.45}
$$

By Taylor expansion and the law of large numbers, we have

$$\begin{aligned}\mathbf{R}_2^*(\check{\beta}_{-1}) &= -\sum_{i=1}^n \int_0^\tau \Big\{\mathbf{Z}_i(s) - \frac{E[Y(s)\mathbf{Z}(s)|U_{0i}(s)]}{E[Y(s)|U_{0i}(s)]}\Big\}\mathbf{Z}_i^T(s)\frac{\lambda_{01}^2(s,U_{0i}(s))}{\lambda(s,U_{0i}(s))}Y_i(s)ds \\ &\quad \times \mathbf{J}(\beta_{-1}^0)(\check{\beta}_{-1}-\beta_{-1}^0) + o_p(\sqrt{n}) \\ &= -nE\Big[\int_0^\tau \Big[\mathbf{Z}(s) - \frac{E[Y(s)\mathbf{Z}(s)|\beta_0^T\mathbf{Z}(s)]}{E[Y(s)|\beta_0^T\mathbf{Z}(s)]}\Big]^{\otimes 2}\frac{\lambda_{01}^2(s,\beta_0^T\mathbf{Z}(s))}{\lambda(s,\beta_0^T\mathbf{Z}(s))}Y(s)ds\Big] \\ &\quad \times \mathbf{J}(\beta_{-1}^0)(\check{\beta}_{-1}-\beta_{-1}^0) + o_p(\sqrt{n}). \qquad (3.46)\end{aligned}$$

By similar arguments of Proposition 3.3, we obtain

$$\sup_{t\in[0,\tau],\mathbf{z}\in\mathcal{Z}}\Big|\hat{\lambda}_{01}(t,\beta_0^T\mathbf{z};\beta_0) - \lambda_{01}(t,\beta_0^T\mathbf{z})\Big| = O_p(h_1^2+h_2^2+\sqrt{\log n/(nh_1h_2^3)}).$$

Based on above facts and with similar techniques in the proofs of Theorems 3.4 and 3.5, it can be shown that

$$\mathbf{R}_k^*(\check{\beta}_{-1}) = o_p(\sqrt{n}), \quad k=3,\cdots,10. \qquad (3.47)$$

Thus, combining with (3.44), (3.45), (3.46) and (3.47), we have

$$\begin{aligned}\mathbf{R}^*(\check{\beta}_{-1}) &= \sum_{i=1}^n \int_0^\tau \Big\{\mathbf{Z}_i(s) - \frac{E[Y(s)\mathbf{Z}(s)|U_{0i}(s)]}{E[Y(s)|U_{0i}(s)]}\Big\}\frac{\lambda_{01}(s,U_{0i}(s))}{\lambda(s,U_{0i}(s))}dM_i(s) \\ &\quad -nE\Big[\int_0^\tau \Big[\mathbf{Z}(s) - \frac{E[Y(s)\mathbf{Z}(s)|\beta_0^T\mathbf{Z}(s)]}{E[Y(s)|\beta_0^T\mathbf{Z}(s)]}\Big]^{\otimes 2}\frac{\lambda_{01}^2(s,\beta_0^T\mathbf{Z}(s))}{\lambda(s,\beta_0^T\mathbf{Z}(s))}Y(s)ds\Big]\mathbf{J}(\beta_{-1}^0)(\check{\beta}_{-1}-\beta_{-1}^0) \\ &\quad +o_p(\sqrt{n}).\end{aligned}$$

Then, plugging the above equation into (3.43) and using the martingale central limit theorem, we can obtain the result of the theorem.

(ii) By the delta-method, we can obtain the result of 3.3 (ii). □

Proof of Theorem 3.4. (i) Note that the estimator $\hat{\beta}_{-1}^{*T}$ satisfies

$$0 = \mathbf{J}^T(\hat{\beta}_{-1}^*)\sum_{i=1}^n \int_0^\tau \Big\{\mathbf{Z}_i(s) - \frac{\widehat{E}[Y(s)\mathbf{Z}(s)|\hat{\beta}^{*T}\mathbf{Z}_i(s)]}{\widehat{E}[Y(s)|\hat{\beta}^{*T}\mathbf{Z}_i(s)]}\Big\}dN_i(s),$$

where $\hat{\beta}^* = \Big((1 - \|\hat{\beta}^*_{-1}\|^2)^{1/2}, \hat{\beta}^{*T}_{-1}\Big)^T$. To obtain the asymptotic properties of $\hat{\beta}^{*T}_{-1}$, we need to study the right-hand side of above expression. Separating this term, we have

$$
\begin{aligned}
0 &= \mathbf{J}^T(\hat{\beta}^*_{-1}) \sum_{i=1}^n \int_0^\tau \Big\{ \mathbf{Z}_i(s) - \frac{\widehat{E}[Y(s)\mathbf{Z}(s)|\hat{\beta}^{*T}\mathbf{Z}_i(s)]}{\widehat{E}[Y(s)|\hat{\beta}^{*T}\mathbf{Z}_i(s)]} \Big\} dN_i(s) \\
&= \mathbf{J}^T(\hat{\beta}^*_{-1}) \sum_{i=1}^n \int_0^\tau \Big\{ \Big(\mathbf{Z}_i(s) - \frac{E[Y(s)\mathbf{Z}(s)|U_{0i}(s)]}{E[Y(s)|U_{0i}(s)]} \Big) \\
&\quad - \Big(\frac{\widehat{E}[Y(s)\mathbf{Z}(s)|\hat{\beta}^{*T}\mathbf{Z}_i(s)]}{\widehat{E}[Y(s)|\hat{\beta}^{*T}\mathbf{Z}_i(s)]} - \frac{\widehat{E}[Y(s)\mathbf{Z}(s)|U_{0i}(s)]}{\widehat{E}[Y(s)|U_{0i}(s)]} \Big) \\
&\quad - \Big(\frac{\widehat{E}[Y(s)\mathbf{Z}(s)|U_{0i}(s)]}{\widehat{E}[Y(s)|U_{0i}(s)]} - \frac{E[Y(s)\mathbf{Z}(s)|U_{0i}(s)]}{E[Y(s)|U_{0i}(s)]} \Big) \Big\} \Big\{ dM_i(s) + Y_i(s)\lambda(s, U_{0i}(s))ds \Big\} \\
&= \mathbf{J}^T(\hat{\beta}^*_{-1}) \sum_{i=1}^n \int_0^\tau \Big\{ \mathbf{Z}_i(s) - \frac{E[Y(s)\mathbf{Z}(s)|U_{0i}(s)]}{E[Y(s)|U_{0i}(s)]} \Big\} dM_i(s) \\
&\quad - \mathbf{J}^T(\hat{\beta}^*_{-1}) \sum_{i=1}^n \int_0^\tau \Big\{ \frac{\widehat{E}[Y(s)\mathbf{Z}(s)|\hat{\beta}^{*T}\mathbf{Z}_i(s)]}{\widehat{E}[Y(s)|\hat{\beta}^{*T}\mathbf{Z}_i(s)]} - \frac{\widehat{E}[Y(s)\mathbf{Z}(s)|U_{0i}(s)]}{\widehat{E}[Y(s)|U_{0i}(s)]} \Big\} dN_i(s) \\
&\quad - \mathbf{J}^T(\hat{\beta}^*_{-1}) \sum_{i=1}^n \int_0^\tau \Big\{ \frac{\widehat{E}[Y(s)\mathbf{Z}(s)|U_{0i}(s)]}{\widehat{E}[Y(s)|U_{0i}(s)]} - \frac{E[Y(s)\mathbf{Z}(s)|U_{0i}(s)]}{E[Y(s)|U_{0i}(s)]} \Big\} dM_i(s) \\
&\quad - \mathbf{J}^T(\hat{\beta}^*_{-1}) \sum_{i=1}^n \int_0^\tau \Big\{ \frac{\widehat{E}[Y(s)\mathbf{Z}(s)|U_{0i}(s)]}{\widehat{E}[Y(s)|U_{0i}(s)]} - \mathbf{Z}_i(s) \Big\} Y_i(s)\lambda(s, U_{0i}(s))ds \\
&= \mathbf{L}_{n,1} - \mathbf{L}_{n,2} - \mathbf{L}_{n,3} - \mathbf{L}_{n,4}. \qquad (3.48)
\end{aligned}
$$

Next consider $\mathbf{L}_{n,j}, j = 2, 3, 4$. By Taylor expansion, we obtain

$$
\begin{aligned}
\mathbf{L}_{n,2} &= \mathbf{J}^T(\beta^0_{-1}) \sum_{i=1}^n \int_0^\tau \frac{\partial}{\partial \beta^T_{-1}} \Big\{ \frac{\widehat{E}[Y(s)\mathbf{Z}(s)|\beta_0^T\mathbf{Z}_i(s)]}{\widehat{E}[Y(s)|\beta_0^T\mathbf{Z}_i(s)]} \Big\} dN_i(s)(\hat{\beta}^*_{-1} - \beta^0_{-1}) + o_p(\sqrt{n}) \\
&= n\mathbf{J}^T(\beta^0_{-1}) E\Big[\int_0^\tau \frac{\partial}{\partial \beta^T_{-1}} \Big\{ \frac{E[Y(s)\mathbf{Z}(s)|\beta_0^T\mathbf{Z}(s)]}{E[Y(s)|\beta_0^T\mathbf{Z}(s)]} \Big\} dN(s) \Big] (\hat{\beta}^*_{-1} - \beta^0_{-1}) + o_p(\sqrt{n}) \\
&= n\mathbf{J}^T(\beta^0_{-1}) E\Big[\int_0^\tau \frac{\partial}{\partial \beta^T_{-1}} \Big\{ \frac{E[Y(s)\mathbf{Z}(s)|\beta_0^T\mathbf{Z}(s)]}{E[Y(s)|\beta_0^T\mathbf{Z}(s)]} \Big\} Y(s)\lambda(s, \beta_0^T\mathbf{Z}(s))ds \Big] (\hat{\beta}^*_{-1} - \beta^0_{-1})
\end{aligned}
$$

$$+o_p(\sqrt{n}).$$

Note that, using the product rule for derivatives, we have

$$\begin{aligned}
0 &= \frac{\partial}{\partial \beta_{-1}^T}\Big\{E\Big[\int_0^\tau \Big\{\mathbf{Z}(s) - \frac{E[Y(s)\mathbf{Z}(s)|\beta_0^T\mathbf{Z}(s)]}{E[Y(s)|\beta_0^T\mathbf{Z}(s)]}\Big\}Y(s)\lambda(s,\beta_0^T\mathbf{Z}(s))ds\Big]\Big\} \\
&= E\Big[\int_0^\tau \frac{\partial}{\partial \beta_{-1}^T}\Big\{\frac{E[Y(s)\mathbf{Z}(s)|\beta_0^T\mathbf{Z}(s)]}{E[Y(s)|\beta_0^T\mathbf{Z}(s)]}\Big\}Y(s)\lambda(s,\beta_0^T\mathbf{Z}(s))ds\Big] \\
&\quad + E\Big[\int_0^\tau \Big\{\mathbf{Z}(s) - \frac{E[Y(s)\mathbf{Z}(s)|\beta_0^T\mathbf{Z}(s)]}{E[Y(s)|\beta_0^T\mathbf{Z}(s)]}\Big\}Y(s)\frac{\partial}{\partial \beta_{-1}^T}\Big\{\lambda(s,\beta_0^T\mathbf{Z}(s))\Big\}ds\Big].
\end{aligned}$$

This leads to

$$\begin{aligned}
\mathbf{L}_{n,2} &= n\mathbf{J}^T(\beta_{-1}^0)E\Big[\int_0^\tau \Big\{\mathbf{Z}(s) - \frac{E[Y(s)\mathbf{Z}(s)|U_0(s)]}{E[Y(s)|U_0(s)]}\Big\}Y(s)\lambda_{01}(s,U_0(s))\mathbf{Z}^T(s)ds\Big] \\
&\quad \times \mathbf{J}(\beta_{-1}^0)(\hat{\beta}_{-1}^* - \beta_{-1}^0) + o_p(\sqrt{n}). \qquad (3.49)
\end{aligned}$$

By similar arguments of Lemma 3.3, it can be shown that

$$\sup_{s\in[0,\tau],\mathbf{z}\in\mathcal{Z}} \|\widehat{E}[Y(s)\mathbf{Z}(s)|U_0(s) = \beta_0^T\mathbf{z}] - E[Y(s)\mathbf{Z}(s)|U_0(s) = \beta_0^T\mathbf{z})]\| = O_P(\sqrt{\log n/nh_3} + h_3^2),$$

$$\sup_{s\in[0,\tau],\mathbf{z}\in\mathcal{Z}} |\widehat{E}[Y(s)|U_0(s) = \beta_0^T\mathbf{z}] - E[Y(s)|U_0(s) = \beta_0^T\mathbf{z}]| = O_P(\sqrt{\log n/nh_3} + h_3^2).$$

Note that $\frac{\widehat{E}[Y(s)\mathbf{Z}(s)|U_{0i}(s)]}{\widehat{E}[Y(s)|U_{0i}(s)]} - \frac{E[Y(s)\mathbf{Z}(s)|U_{0i}(s)]}{E[Y(s)|U_{0i}(s)]}$ is a predictable process with the filtration $\mathcal{F}_{t,i}$. Thus, by the martingale central limit theorem, we have

$$\mathbf{L}_{n,3} = O_P(\sqrt{\log n/nh_3} + h_3^2)O_p(\sqrt{n}). \qquad (3.50)$$

Let

$$w_{ni}(s,u;\beta) = \frac{(\beta^T\mathbf{Z}_i(s) - u)I(X_i \geqslant s)}{\sum_{j=1}^n k_{h_3}(\beta^T\mathbf{Z}_j(s) - u)I(X_j \geqslant s)}$$

and $\mathbf{L}_{n,4s}$ denote the s-th component of $\mathbf{L}_{n,4}$. Then,

$$\begin{aligned} E\Big[\frac{1}{\sqrt{n}}\mathbf{L}_{n,4s}\Big]^2 &= \frac{1}{n}E\Big[\sum_{i=1}^{n}\int_0^\tau\{\sum_{j=1}^{n}w_{nj}(s,\beta_0^T\mathbf{Z}_i(s);\beta_0)\mathbf{Z}_{js}(s)-\mathbf{Z}_{is}(s)\}Y_i(s)\lambda(s,U_{0i}(s))ds\Big]^2 \\ &\leqslant \sum_{i=1}^{n}E\Big[\int_0^\tau\{\sum_{j=1}^{n}w_{nj}(s,\beta_0^T\mathbf{Z}_i(s);\beta_0)\mathbf{Z}_{js}(s)-\mathbf{Z}_{is}(s)\}Y_i(s)\lambda(s,U_{0i}(s))ds\Big]^2 \\ &\leqslant \sum_{i=1}^{n}\int_0^\tau E\Big[\sum_{j=1}^{n}w_{nj}(s,\beta_0^T\mathbf{Z}_i(s);\beta_0)\mathbf{Z}_{js}(s)-\mathbf{Z}_{is}(s)\Big]^2ds \\ &\leqslant cnh_3^4. \end{aligned} \tag{3.51}$$

The last inequality holds from similar arguments of Lemma 1 in Zhu and Xue (2006).

Together with (3.48), (3.49), (3.50) and (3.51), we have

$$\begin{aligned} 0 &= \frac{1}{\sqrt{n}}\mathbf{J}^T(\beta_{-1}^0)\sum_{i=1}^{n}\int_0^\tau\Big\{\mathbf{Z}_i(s)-\frac{E[Y(s)\mathbf{Z}(s)|U_{0i}(s)]}{E[Y(s)|U_{0i}(s)]}\Big\}dM_i(s) \\ &\quad -\mathbf{J}^T(\beta_{-1}^0)E\Big[\int_0^\tau\Big\{\mathbf{Z}(s)-\frac{E[Y(s)\mathbf{Z}(s)|U_0(s)]}{E[Y(s)|U_0(s)]}\Big\}Y(s)\lambda_{01}(s,U_0(s))\mathbf{Z}^T(s)ds\Big] \\ &\quad \times\mathbf{J}(\beta_{-1}^0)\sqrt{n}(\hat{\beta}_{-1}^*-\beta_{-1}^0)+O_p(\sqrt{nh_3^4}). \end{aligned}$$

If $nh_3^4\to 0$ and $nh_3/\log n\to 0$, this leads to

$$\begin{aligned} \sqrt{n}A(\hat{\beta}_{-1}^*-\beta_{-1}^0) &= \frac{1}{\sqrt{n}}\mathbf{J}^T(\beta_{-1}^0)\sum_{i=1}^{n}\int_0^\tau\Big\{\mathbf{Z}_i(s)-\frac{E[Y(s)\mathbf{Z}(s)|U_{0i}(s)]}{E[Y(s)|U_{0i}(s)]}\Big\}dM_i(s) \\ \xrightarrow{\mathrm{d}} N\Big(0,\ \mathbf{J}^T(\beta_{-1}^0)E\Big[\int_0^\tau\Big\{\mathbf{Z}(s)-\frac{E[Y(s)\mathbf{Z}(s)|U_0(s)]}{E[Y(s)|U_0(s)]}\Big\}^{\otimes 2}\lambda(s,U_0(s))Y(s)ds\Big]\mathbf{J}(\beta_{-1}^0)\Big), \end{aligned}$$

by the martingale central limit theorem. Thus, the proof of Theorem 3.4 is completed if A is nonsingular.

(ii) By the delta-method, we can obtain the result of Theorem 3.4 (ii). □

Proof of Theorem 3.5. (i) By the definition of $\hat{\beta}^{**}_{-1}$, we have

$$0 = \mathbf{J}^T(\hat{\beta}^{**}_{-1})\sum_{i=1}^{n}\int_0^\tau \mathbf{Z}_i(s)\Big[dN_i(s) - Y_i(s)\widehat{\lambda}(s, \mathbf{Z}_i^T(s)\hat{\beta}^{**}; \hat{\beta}^{**})ds\Big],$$

where $\hat{\beta}^{**} = \Big((1 - \|\hat{\beta}^{**}_{-1}\|^2)^{1/2}, \hat{\beta}^{**T}_{-1}\Big)^T$. Note that

$$\begin{aligned}
0 &= \frac{1}{\sqrt{n}}\mathbf{J}^T(\hat{\beta}^{**}_{-1})\sum_{i=1}^{n}\int_0^\tau \mathbf{Z}_i(s)\Big[dN_i(s) - Y_i(s)\widehat{\lambda}(s, \mathbf{Z}_i^T(s)\hat{\beta}^{**}; \hat{\beta}^{**})ds\Big] \\
&= \frac{1}{\sqrt{n}}\mathbf{J}^T(\hat{\beta}^{**}_{-1})\sum_{i=1}^{n}\int_0^\tau \mathbf{Z}_i(s)\Big[dN_i(s) - Y_i(s)\lambda(s, \beta_0^T\mathbf{Z}_i(s))ds\Big] \\
&\quad -\frac{1}{\sqrt{n}}\mathbf{J}^T(\hat{\beta}^{**}_{-1})\sum_{i=1}^{n}\int_0^\tau \mathbf{Z}_i(s)\Big[\widehat{\lambda}(s, \mathbf{Z}_i^T(s)\hat{\beta}^{**}; \hat{\beta}^{**}) - \widehat{\lambda}(s, \beta_0^T\mathbf{Z}_i(s); \beta_0)\Big]Y_i(s)ds \\
&\quad -\frac{1}{\sqrt{n}}\mathbf{J}^T(\hat{\beta}^{**}_{-1})\sum_{i=1}^{n}\int_0^\tau \mathbf{Z}_i(s)\Big[\widehat{\lambda}(s, \beta_0^T\mathbf{Z}_i(s); \beta_0) - \lambda(s, \beta_0^T\mathbf{Z}_i(s))\Big]Y_i(s)ds \\
&= \mathbf{M}_{n,1} - \mathbf{M}_{n,2} - \mathbf{M}_{n,3}. \qquad (3.52)
\end{aligned}$$

Next consider $\mathbf{M}_{n,2}$ and $\mathbf{M}_{n,3}$. By Proposition 3.3, we have

$$\begin{aligned}
\mathbf{M}_{n,2} &= \frac{1}{\sqrt{n}}\mathbf{J}^T(\hat{\beta}^{**}_{-1})\sum_{i=1}^{n}\int_0^\tau \mathbf{Z}_i(s)Y_i(s)\frac{\partial\hat{\lambda}(s, \beta_0^T\mathbf{Z}_i(s); \beta_0)}{\partial\beta^T_{-1}}ds(\hat{\beta}^{**}_{-1} - \beta^0_{-1}) \\
&\quad +o_P(\sqrt{n}\|\hat{\beta}^{**}_{-1} - \beta^0_{-1}\|) \\
&= \frac{1}{n}\mathbf{J}^T(\beta^0_{-1})\sum_{i=1}^{n}\int_0^\tau \mathbf{Z}_i(s)Y_i(s)\lambda_{01}(s, U_{0i}(s))\Big[\mathbf{Z}_i(s) - \frac{E[Y(s)\mathbf{Z}(s)|U_{0i}(s)]}{E[Y(s)|U_{0i}(s)]}\Big]^T ds\mathbf{J}(\beta^0_{-1}) \\
&\quad \times\sqrt{n}(\hat{\beta}^{**}_{-1} - \beta^0_{-1}) + o_P(\sqrt{n}\|\hat{\beta}^{**}_{-1} - \beta^0_{-1}\|). \qquad (3.53)
\end{aligned}$$

By Lemma 3.3, we obtain

$$\mathbf{M}_{n,3} = \frac{1}{\sqrt{n}}\mathbf{J}^T(\beta^0_{-1})\sum_{i=1}^{n}\int_0^\tau \mathbf{Z}_i(s)Y_i(s)\frac{\sum_{j=1}^{n}\int_0^\tau K_h(t-s, U_{0j}(t) - U_{0i}(s))dM_j(t)}{nE[Y(s)|U_{0i}(s)]f_{\beta_0}(U_{0i}(s))}ds$$

$$
\begin{aligned}
&+O_p(n^{1/2}(h_1^2+h_2^2)) \\
= \;& \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\int_0^\tau \Big\{\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \frac{\mathbf{Z}_i(s)Y_i(s)K_h(t-s,U_{0j}(t)-U_{0i}(s))}{E[Y(s)|U_{0i}(s)]f_{\beta_0}(U_{0i}(s))}ds\Big\}dM_j(t) \\
&+O_p(n^{1/2}(h_1^2+h_2^2)).
\end{aligned}
$$

Note that the integrand

$$
\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \frac{\mathbf{Z}_i(s)Y_i(s)K_h(t-s,U_{0j}(t)-U_{0i}(s))}{E[Y(s)|U_{0i}(s)]f_{\beta_0}(U_{0i}(s))}ds
$$

is not predictable with the filtration $\mathcal{F}_{t,j}$. In addition, by Lemma 3.2,

$$
\max_{j=1,\cdots,n}\sup_{t\in[0,\tau]}\Big\|\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \frac{\mathbf{Z}_i(s)Y_i(s)K_h(t-s,U_{0j}(t)-U_{0i}(s))}{nE[Y(s)|U_{0i}(s)]f_{\beta_0}(U_{0i}(s))}ds-\frac{E[Y(t)\mathbf{Z}(t)|U_{0j}(t)]}{E[Y(t)|U_{0j}(t)]}\Big\| = O_P(c_n).
$$

Then, by similar arguments of Lemma 3.4, we can show that

$$
\mathbf{M}_{n,3} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^\tau \frac{E[Y(s)\mathbf{Z}(s)|U_{0i}(s)]}{E[Y(s)|U_{0i}(s)]}dM_i(s) + O_p(n^{1/2}(h_1^2+h_2^2)). \tag{3.54}
$$

Thus, combining with (3.52), (3.53) and (3.54), we obtain

$$
\begin{aligned}
\sqrt{n}A(\hat{\beta}^{**}_{-1}-\beta^0_{-1}) &= \frac{1}{\sqrt{n}}\mathbf{J}^T(\beta^0_{-1})\sum_{i=1}^{n}\int_0^\tau \Big\{\mathbf{Z}_i(s)-\frac{E[Y(s)\mathbf{Z}(s)|U_{0i}(s)]}{E[Y(s)|U_{0i}(s)]}\Big\}dM_i(s) \\
&\xrightarrow{\mathrm{d}} N\Big(0,\quad \mathbf{J}^T(\beta^0_{-1})E\Big[\int_0^\tau \Big\{\mathbf{Z}(s)-\frac{E[Y(s)\mathbf{Z}(s)|U_0(s)]}{E[Y(s)|U_0(s)]}\Big\}^{\otimes 2}\lambda(s,U_0(s))Y(s)ds\Big]\mathbf{J}(\beta^0_{-1})\Big),
\end{aligned}
$$

under the condition $n^{1/2}(h_1^2+h_2^2)\to 0$.

(ii) By the delta-method, we can obtain the result of Theorem 3.5 (ii). □

Proof of Theorem 3.6. (i) Note that

$$
\sqrt{nh_1h_2}\Big\{\hat{\lambda}(t,\hat{\beta}^T\mathbf{z};\hat{\beta})-\lambda(t,\beta_0^T\mathbf{z})\Big\} \;=\; \sqrt{nh_1h_2}\Big\{\hat{\lambda}(t,\hat{\beta}^T\mathbf{z};\hat{\beta})-\hat{\lambda}(t,\beta_0^T\mathbf{z};\beta_0)\Big\}
$$

$$+\sqrt{nh_1h_2}\Big\{\hat{\lambda}(t,\beta_0^T\mathbf{z};\beta_0)-\lambda(t,\beta_0^T\mathbf{z})\Big\}.$$

By Taylor expansion, we have

$$\hat{\lambda}(t,\hat{\beta}^T\mathbf{z};\hat{\beta})-\hat{\lambda}(t,\beta_0^T\mathbf{z};\beta_0)=(\hat{\beta}_{-1}-\beta_{-1}^0)^T\frac{\partial\hat{\lambda}(t,\beta_0^T\mathbf{z};\beta_0)}{\partial\beta_{-1}}\{1+o_P(1)\}.$$

By the similar arguments of Proposition 3.3, we obtain

$$\begin{aligned}\frac{\partial\hat{\lambda}(t,\beta_0^T\mathbf{z};\beta_0)}{\partial\beta_{-1}} &= \lambda_{01}(t,\beta_0^T\mathbf{z})\mathbf{J}^T(\beta_{-1}^0)\Big\{\mathbf{z}-\frac{E[Y(t)\mathbf{Z}(t)|U_0(t)=\beta_0^T\mathbf{z}]}{E[Y(t)|U_0(t)=\beta_0^T\mathbf{z}]}\Big\}\\ &\quad +O_p(h_1^2+h_2^2+h_2^{-1}\sqrt{1/(nh_1h_2)}),\end{aligned}$$

as $nh_1h_2^3\to\infty$. This, together with Theorem 3.1, yields

$$\sqrt{nh_1h_2}\Big\{\hat{\lambda}(t,\hat{\beta}^T\mathbf{z};\hat{\beta})-\hat{\lambda}(t,\beta_0^T\mathbf{z};\beta_0)\Big\}=O_P(\sqrt{nh_1h_2}/\sqrt{n}). \tag{3.55}$$

Using the same arguments of Lemma 3.3, we can show that

$$\begin{aligned}\hat{\lambda}(t,\beta_0^T\mathbf{z};\beta_0)-\lambda(t,\beta_0^T\mathbf{z}) &= \frac{1}{2}\sum_{i=1}^n\int_0^\tau W_{ni}(s,t,\beta_0^T\mathbf{z};\beta_0)\lambda_{02}(t,\tilde{U}_{1i}^*(s))(\beta_0^T[\mathbf{Z}_i(s)-\mathbf{z}])^2Y_i(s)ds\\ &\quad +\frac{1}{2}\sum_{i=1}^n\int_0^\tau W_{ni}(s,t,\beta_0^T\mathbf{z};\beta_0)\lambda_{20}(\tilde{t}^*,\tilde{U}_{2i}^*(s))(s-t)^2Y_i(s)ds\\ &\quad +\frac{\sum_{i=1}^n\int_0^\tau K_h(s-t,\beta_0^T[\mathbf{Z}_i(s)-\mathbf{z}])dM_i(s)}{nE[Y(t)|U_0(t)=\beta_0^T\mathbf{z}]f_{\beta_0}(\beta_0^T\mathbf{z})}+O_p(c_n^2)\\ &= \frac{1}{2}h_1^2\mu_2\lambda_{20}(t,\beta_0^T\mathbf{z})+\frac{1}{2}h_2^2\mu_2\lambda_{02}(t,\beta_0^T\mathbf{z})\\ &\quad +\frac{\sum_{i=1}^n\int_0^\tau K_h(s-t,\beta_0^T[\mathbf{Z}_i(s)-\mathbf{z}])dM_i(s)}{nE[Y(t)|U_0(t)=\beta_0^T\mathbf{z}]f_{\beta_0}(\beta_0^T\mathbf{z})}\\ &\quad +o_p(\sqrt{1/(nh_1h_2)}+h_1^2+h_2^2),\end{aligned}$$

where $\tilde{t}^*$ lies between s and t; $\tilde{U}_{1i}^*(s)$ and $\tilde{U}_{2i}^*(s)$ are values between $\beta_0^T\mathbf{Z}_i(s)$ and $\beta_0^T\mathbf{z}$. By the martingale central limit theorem,

$$\sqrt{nh_1h_2}\Big\{\hat{\lambda}(t,\beta_0^T\mathbf{z};\beta_0)-\lambda(t,\beta_0^T\mathbf{z})-b(t,\beta_0^T\mathbf{z})\Big\}$$

$$= \sqrt{\frac{h_1 h_2}{n}} \frac{\sum_{i=1}^n \int_0^\tau K_h(s-t, \beta_0^T[\mathbf{Z}_i(s) - \mathbf{z}]) dM_i(s)}{E[Y(t)|U_0(t) = \beta_0^T \mathbf{z}] f_{\beta_0}(\beta_0^T \mathbf{z})} \{1 + o_p(1)\}$$

$$\xrightarrow{\mathrm{d}} N\left(0, \frac{\nu_0^2 \lambda(t, \beta_0^T \mathbf{z})}{E[Y(t)|U_0(t) = \beta_0^T \mathbf{z}] f_{\beta_0}(\beta_0^T \mathbf{z})}\right), \tag{3.56}$$

where the asymptotic bias is

$$b(t, \beta_0^T \mathbf{z}) = \frac{1}{2} h_1^2 \mu_2 \lambda_{20}(t, \beta_0^T \mathbf{z}) + \frac{1}{2} h_2^2 \mu_2 \lambda_{02}(t, \beta_0^T \mathbf{z}).$$

Combining with (3.55) and (3.56),

$$\sqrt{n h_1 h_2}\Big\{\hat{\lambda}(t, \hat{\beta}^T \mathbf{z}; \hat{\beta}) - \lambda(t, \beta_0^T \mathbf{z}) - b(t, \beta_0^T \mathbf{z})\Big\} \xrightarrow{\mathrm{d}} N\left(0, \frac{\nu_0^2 \lambda(t, \beta_0^T \mathbf{z})}{E[Y(t)|U_0(t) = \beta_0^T \mathbf{z}] f_{\beta_0}(\beta_0^T \mathbf{z})}\right).$$

(ii) By Proposition 3.1, (3.55) holds uniformly for $t \in [0, \tau]$ and $\mathbf{z} \in \mathcal{Z}$, as $n h_1 h_2^3 / \log n \to \infty$. By Lemma 3.3, we have

$$\sup_{t \in [0,\tau], \mathbf{z} \in \mathcal{Z}} \left| \hat{\lambda}(t, \beta_0^T \mathbf{z}; \beta_0) - \lambda(t, \beta_0^T \mathbf{z}) - \frac{\sum_{i=1}^n \int_0^\tau K_h(s-t, \beta_0^T[\mathbf{Z}_i(s) - \mathbf{z}]) dM_i(s)}{n E[Y(t)|U_0(t) = \beta_0^T \mathbf{z}] f_{\beta_0}(\beta_0^T \mathbf{z})} \right|$$

$$= O_p(h_1^2 + h_2^2) + O_p(c_n^2),$$

as $h_1 \to 0, h_2 \to 0$ and $(n h_1 h_2)/\log n \to \infty$. Note that

$$\sup_{t \in [0,\tau],\ \mathbf{z} \in \mathcal{Z}} \left| \frac{\sum_{i=1}^n \int_0^\tau K_{\mathbf{h}}(s-t, \beta_0^T(\mathbf{Z}_i(s) - \mathbf{z})) dM_i(s)}{n E[Y(t)|U(t) = \beta_0^T \mathbf{z}] f_{\beta_0}(\beta_0^T \mathbf{z})} \right| = O_p(\sqrt{\log n/(n h_1 h_2)}).$$

Thus, we have

$$\sup_{t \in [0,\tau], \mathbf{z} \in \mathcal{Z}} \left| \hat{\lambda}(t, \beta_0^T \mathbf{z}; \beta_0) - \lambda(t, \beta_0^T \mathbf{z}) \right| = O_p(h_1^2 + h_2^2 + \sqrt{\log n/(n h_1 h_2)}) + O_p(c_n^2).$$

Therefore, if $n h_1 h_2^3 / \log n \to \infty$, we have

$$\sup_{t \in [0,\tau], \mathbf{z} \in \mathcal{Z}} \left| \hat{\lambda}(t, \hat{\beta}^T \mathbf{z}; \hat{\beta}) - \lambda(t, \beta_0^T \mathbf{z}) \right| \leqslant \sup_{t \in [0,\tau], \mathbf{z} \in \mathcal{Z}} \left| \hat{\lambda}(t, \hat{\beta}^T \mathbf{z}; \hat{\beta}) - \hat{\lambda}(t, \beta_0^T \mathbf{z}; \beta_0) \right|$$

$$+ \sup_{t\in[0,\tau],\mathbf{z}\in\mathcal{Z}} \left|\hat{\lambda}(t,\beta_0^T\mathbf{z};\beta_0) - \lambda(t,\beta_0^T\mathbf{z})\right|$$

$$= O_p(h_1^2 + h_2^2 + \sqrt{\log n/(nh_1h_2)}),$$

which proves Theorem 3.6(ii). □

3.10 Proofs of auxiliary lemmas

Proof of Lemma 3.1. To prove these results, we only need to verify (A.1) and (A.2) in Lemma A.1 of Wang et al. (2010). Let

$$\begin{aligned}
\xi_i(t,\beta^T\mathbf{z},\beta) &= \frac{\sqrt{nh_1h_2}}{h_1^j h_2^k\sqrt{\log n}}\Big\{\int_0^\tau K_{\mathbf{h}}(s-t,\beta^T(\mathbf{Z}_i(s)-\mathbf{z}))(s-t)^j\{\beta^T(\mathbf{Z}_i(s)-\mathbf{z})\}^k Y_i(s)ds \\
&\quad -E\Big[\int_0^\tau K_{\mathbf{h}}(s-t,\beta^T(\mathbf{Z}_i(s)-\mathbf{z}))(s-t)^j\{\beta^T(\mathbf{Z}_i(s)-\mathbf{z})\}^k Y_i(s)ds\Big]\Big\}, \\
f_{(t,\mathbf{z},\beta)}(V_i) &= \xi_i(t,\mathbf{z},\beta), \quad V_i = (X_i,\delta_i,\mathbf{Z}_i(X_i),t\in[0,\tau])^T, \quad i=1,2,\cdots,n.
\end{aligned}$$

By condition (C4) on the kernel function, we calculate that

$$\frac{1}{n}\sum_{i=1}^n |f_{(t,\mathbf{z},\beta)}(V_i) - f_{(t^*,\mathbf{z}^*,\beta^*)}(V_i)| \leqslant cn^a(\|\beta-\beta^*\| + \|\mathbf{z}-\mathbf{z}^*\| + |t-t^*|),$$

for some constant c and a. Thus, (A.1) in Wang et al. (2010) is satisfied.

We next verify that (A.2) in Wang et al. (2010) is satisfied. By Cauchy-Schwarz inequality, we obtain

$$\begin{aligned}
Var(\frac{1}{n}\sum_{i=1}^n \xi_i(t,\beta^T\mathbf{z},\beta)) &\leqslant h_1^{1-2j}h_2^{1-2k}(\log n)^{-1}E\Big[\int_0^\tau K_{\mathbf{h}}(s-t,\beta^T(\mathbf{Z}_i(s)-\mathbf{z})) \\
&\quad \times(s-t)^j[\beta^T(\mathbf{Z}_i(s)-\mathbf{z})]^k Y_i(s)ds\Big]^2 \\
&\leqslant h_1^{1-2j}h_2^{1-2k}(\log n)^{-1}E\Big[\int_0^\tau K_{\mathbf{h}}^2(s-t,\beta^T(\mathbf{Z}_i(s)-\mathbf{z})) \\
&\quad \times(s-t)^{2j}[\beta^T(\mathbf{Z}_i(s)-\mathbf{z})]^{2k} Y_i(s)ds\Big].
\end{aligned}$$

Furthermore, by Fubini's theorem, we have

$$\begin{aligned} Var(\frac{1}{n}\sum_{i=1}^{n}\xi_i(t,\beta^T\mathbf{z},\beta)) &\leqslant h_1^{1-2j}h_2^{1-2k}(\log n)^{-1}\int\int_{[0,\tau]\times\mathcal{U}_\beta} k_{h_1}^2(s-t)k_{h_2}^2(u-\beta^T\mathbf{z}) \\ &\quad\times(s-t)^{2j}(u-\beta^T\mathbf{z})^{2k}\phi_\beta(s,u)duds \\ &\leqslant (\log n)^{-1}\int\int_{[0,\tau]\times\mathcal{U}_\beta} k^2(r)k^2(w)r^{2j}w^{2k} \\ &\quad\times\phi_\beta(t+h_1r,\beta^T\mathbf{z}+h_2w)dwdr \\ &= O((\log n)^{-1}), \end{aligned} \tag{3.57}$$

holds uniformly for $(t,\mathbf{z},\beta)\in\mathcal{A}_n$, under conditions (C2) and (C4). Thus, given a $\varepsilon_n>0$, by Chevbychev's inequality and (3.57), we have

$$\begin{aligned} P\Big\{\Big|\frac{1}{n}\sum_{i=1}^{n}\xi_i(t,\beta^T\mathbf{z},\beta)\Big|>\frac{1}{2}\varepsilon_n\Big\} &\leqslant 4\varepsilon_n^{-2}Var(\frac{1}{n}\sum_{i=1}^{n}\xi_i(t,\beta^T\mathbf{z},\beta)) \\ &\leqslant c\varepsilon_n^{-2}(\log n)^{-1}<1/2, \end{aligned}$$

as $n\to\infty$. Hence, (A.2) in Wang et al. (2010) is satisfied. Again, using (3.57), we have

$$\frac{1}{n^2}\sum_{i=1}^{n}E[\xi_i^2(t,\beta^T\mathbf{z},\beta)]=Var(\frac{1}{n}\sum_{i=1}^{n}\xi_i(t,\beta^T\mathbf{z},\beta))=O_p((\log n)^{-1}).$$

Thus, given a sufficiently large $M>0$, by Markov's inequality, we have

$$P\Big\{\frac{1}{n^2}\sum_{i=1}^{n}\xi_i^2(t,\beta^T\mathbf{z},\beta)>M(\log n)^{-1}\Big\}\leqslant\frac{n^{-2}\sum_{i=1}^{n}E[\xi_i^2(t,\beta^T\mathbf{z},\beta)]}{M(\log n)^{-1}}\leqslant c\frac{1}{M}.$$

Hence,

$$\frac{1}{n^2}\sum_{i=1}^{n}\xi_i^2(t,\beta^T\mathbf{z},\beta)=O_p((\log n)^{-1}).$$

Then, from Lemma (A.1) in Wang et al. (2010), we have, for $j,k=0,1,2$

$$P\Big\{\sup_{(t,\mathbf{z},\beta)\in\mathcal{A}_n}\Big|\frac{1}{n}\sum_{i=1}^{n}\xi_i(t,\beta^T\mathbf{z},\beta)\Big|>\frac{1}{2}\varepsilon_n\Big\}\leqslant c_1n^{2pa}\varepsilon_n^{-2p}\exp\{-c_2\varepsilon_n^2\log n\}$$

for some constants c_1 and c_2. As $h_1 \to 0, h_2 \to 0$ and $n \to \infty$, by choosing a sufficiently large ε_n, it follows that the right-hand side of the above formula tends to zero. Therefore, we have

$$\sup_{(t,\mathbf{z},\beta)\in\mathcal{A}_n} \Big| S_{jk}(t, \beta^T\mathbf{z}; \beta) - E[S_{jk}(t, \beta^T\mathbf{z}; \beta)] \Big| = O_p(h_1^j h_2^k \sqrt{\log n/(nh_1h_2)}).$$

□

Proof of Lemma 3.2. By changing variables to $r = (s-t)/h_1$ and $v = (w-u)/h_2$, we have, for $j, k = 0, 1, 2$

$$\begin{aligned}
E[S_{jk}(t,u;\beta)] &= \int_0^\tau E\Big[E[K_{\mathbf{h}}(s-t, U(s)-u)(s-t)^j(U(s)-u)^k Y(s)|U(s)]\Big]ds \\
&= \int_0^\tau \int_{w\in\mathcal{U}_\beta} k_{h_1}(s-t)k_{h_2}(w-u)(s-t)^j(w-u)^k E[Y(s)|U(s)=w] \\
&\quad \times f_\beta(w)dwds \\
&= \int_{\frac{-t}{h_1}}^{\frac{\tau-t}{h_1}} \int_{\tilde{\mathcal{U}}_\beta} k(r)k(v)(h_1r)^j(h_2v)^k\phi_\beta(t+h_1r, u+h_2v)dvdr \\
&= \phi_\beta(t,u)h_1^j h_2^k \mu_j\mu_k + \frac{\partial\phi_\beta(t,u)}{\partial t}h_1^{j+1}h_2^k\mu_{j+1}\mu_k + \frac{\partial\phi_\beta(t,u)}{\partial u}h_1^j h_2^{k+1}\mu_j\mu_{k+1} \\
&\quad + O(h_1^j h_2^k(h_1^2+h_2^2)),
\end{aligned}$$

with $\tilde{\mathcal{U}}_\beta = \{v : v = (w-u)/h_2, w \in \mathcal{U}_\beta\}$. The last equality holds by a second-order Taylor expansion of $\phi_\beta(t+h_1r, u+h_2v)$. Combining this with Lemma 3.1, we complete the proof.

□

Proof of Lemma 3.3. Let

$$S_n^\beta(t,\mathbf{z}) = n^{-1}\sum_{i=1}^n \int_0^\tau K_{\mathbf{h}}(s-t, \beta^T(\mathbf{Z}_i(s)-\mathbf{z}))\begin{pmatrix} 1 \\ (s-t)/h_1 \\ \beta^T(\mathbf{Z}_i(s)-\mathbf{z})/h_2 \end{pmatrix}$$

$$\times \begin{pmatrix} 1 & (s-t)/h_1 & \beta^T(\mathbf{Z}_i(s)-\mathbf{z})/h_2 \end{pmatrix} Y_i(s)ds$$

and

$$\begin{pmatrix} T_{00}^{\beta}(t,\mathbf{z}) \\ T_{10}^{\beta}(t,\mathbf{z}) \\ T_{01}^{\beta}(t,\mathbf{z}) \end{pmatrix} = n^{-1}\sum_{i=1}^{n}\int_0^{\tau} K_{\mathbf{h}}(s-t,\beta^T(\mathbf{Z}_i(s)-\mathbf{z}))\begin{pmatrix} 1 \\ (s-t)/h_1 \\ \beta^T(\mathbf{Z}_i(s)-\mathbf{z})/h_2 \end{pmatrix} dN_i(s).$$

By Lemma 3.2, we have

$$S_n^{\beta}(t,\mathbf{z}) = \begin{pmatrix} \phi(t,\beta^T\mathbf{z}) & \frac{\partial\phi(t,\beta^T\mathbf{z})}{\partial t}h_1\mu_2 & \frac{\partial\phi(t,\beta^T\mathbf{z})}{\partial(\beta^T\mathbf{z})^T}h_2\mu_2 \\ \frac{\partial\phi(t,\beta^T\mathbf{z})}{\partial t}h_1\mu_2 & \phi(t,\beta^T\mathbf{z})\mu_2 & \mathbf{0} \\ \frac{\partial\phi(t,\beta^T\mathbf{z})}{\partial(\beta^T\mathbf{z})}h_2\mu_2 & \mathbf{0} & \phi(t,\beta^T\mathbf{z})\mu_2 \end{pmatrix} + O(c_n).$$

Using the matrix equality

$$\{\mathbf{A}+(\mathbf{A}_n-\mathbf{A})\}^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}(\mathbf{A}_n-\mathbf{A})[\mathbf{I}_p+\mathbf{A}^{-1}(\mathbf{A}_n-\mathbf{A})]^{-1}\mathbf{A}^{-1},$$

we have

$$\begin{aligned} S_n^{\beta}(t,\mathbf{z})^{-1} &= \phi(t,\beta^T\mathbf{z})^{-1}\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1/\mu_2 & 0 \\ 0 & 0 & 1/\mu_2 \end{pmatrix} \\ &\quad +\{\phi(t,\beta^T\mathbf{z})\mu_2\}^{-1}\begin{pmatrix} 0 & \frac{\partial\phi(t,\mathbf{z})}{\partial t}h_1 & \frac{\partial\phi(t,\beta^T\mathbf{z})}{\partial(\beta^T\mathbf{z})}h_2 \\ \frac{\partial\phi(t,\mathbf{z})}{\partial t}h_1 & 0 & 0 \\ \frac{\partial\phi(t,\beta^T\mathbf{z})}{\partial(\beta^T\mathbf{z})}h_2 & 0 & 0 \end{pmatrix} + O(c_n). \end{aligned} \tag{3.58}$$

Using $dN_i(s) = \lambda_0(s,\beta_0^T\mathbf{Z}_i)Y_i(s)ds + dM_i(s)$, we have

$$\begin{aligned} &\begin{pmatrix} T_{00}^{\beta}(t,\mathbf{z}) \\ T_{10}^{\beta}(t,\mathbf{z}) \\ T_{01}^{\beta}(t,\mathbf{z}) \end{pmatrix} \\ &= n^{-1}\sum_{i=1}^{n}\int_0^{\tau} K_{\mathbf{h}}(s-t,\beta^T(\mathbf{Z}_i(s)-\mathbf{z}))\begin{pmatrix} 1 \\ (s-t)/h_1 \\ \beta^T(\mathbf{Z}_i(s)-\mathbf{z})/h_2 \end{pmatrix} Y_i(s)\lambda_0(s,\beta_0^T\mathbf{Z}_i(s))ds \\ &\quad + n^{-1}\sum_{i=1}^{n}\int_0^{\tau} K_{\mathbf{h}}(s-t,\beta^T(\mathbf{Z}_i(s)-\mathbf{z}))\begin{pmatrix} 1 \\ (s-t)/h_1 \\ \beta^T(\mathbf{Z}_i(s)-\mathbf{z})/h_2 \end{pmatrix} dM_i(s). \end{aligned} \tag{3.59}$$

By the Taylor expansion of $\lambda_0(s, \beta_0^T \mathbf{Z}_i(s))$ at $(t, \beta_0^T \mathbf{z})$, we have

$$\begin{aligned}
& \lambda_0(s, \beta_0^T \mathbf{Z}_i(s)) \\
= \; & \lambda_0(t, \beta_0^T \mathbf{z}) + \lambda_{10}(t, \beta_0^T \mathbf{z})(s-t) + \lambda_{01}(t, \beta_0^T \mathbf{z}) \beta_0^T (\mathbf{Z}_i(s) - \mathbf{z}) \\
& + \frac{1}{2}(s-t, \beta_0^T (\mathbf{Z}_i(s) - \mathbf{z})) \mathbf{H}(t, \beta_0^T \mathbf{z})(s-t, \beta_0^T (\mathbf{Z}_i(s) - \mathbf{z}))^T \\
& + O(|s-t|^3 + |\beta_0^T (\mathbf{Z}_i(s) - \mathbf{z})|^3) \\
= \; & \lambda_0(t, \beta_0^T \mathbf{z}) + \lambda_{10}(t, \beta_0^T \mathbf{z})(s-t) + \lambda_{01}(t, \beta_0^T \mathbf{z}) \beta^T (\mathbf{Z}_i(s) - \mathbf{z}) \\
& \quad + \frac{1}{2}(s-t, \beta^T (\mathbf{Z}_i(s) - \mathbf{z})) \mathbf{H}(t, \beta_0^T \mathbf{z})(s-t, \beta^T (\mathbf{Z}_i(s) - \mathbf{z}))^T \\
& + \lambda_{01}(t, \beta_0^T \mathbf{z})(\beta_0 - \beta)^T (\mathbf{Z}_i(s) - \mathbf{z}) + R_n(s, \mathbf{Z}_i(s) - \mathbf{z}),
\end{aligned} \tag{3.60}$$

where $R_n(s, \mathbf{Z}_i(s) - \mathbf{z}) = O(|s-t|^3 + |\beta_0^T (\mathbf{Z}_i(s) - \mathbf{z}))|^3 + |\beta^T (\mathbf{Z}_i(s) - \mathbf{z}))||\mathbf{Z}_i(s) - \mathbf{z})||\beta_0 - \beta| + |\mathbf{Z}_i(s) - \mathbf{z})|^2 |\beta_0 - \beta|^2)$. It is easy to see that

$$\begin{aligned}
& \{n S_n^\beta(t, \mathbf{z})\}^{-1} \sum_{i=1}^n \int_0^\tau K_{\mathbf{h}}(s-t, \beta^T (\mathbf{Z}_i(s) - \mathbf{z})) \begin{pmatrix} 1 \\ (s-t)/h_1 \\ \beta^T (\mathbf{Z}_i(s) - \mathbf{z})/h_2 \end{pmatrix} \{\lambda_0(t, \beta_0^T \mathbf{z}) \\
& + \lambda_{10}(t, \beta_0^T \mathbf{z})(s-t) + \lambda_{01}(t, \beta_0^T \mathbf{z}) \beta^T (\mathbf{Z}_i(s) - \mathbf{z})\} Y_i(s) ds \\
= \; & \begin{pmatrix} \lambda_0(t, \beta_0^T \mathbf{z}) \\ \lambda_{10}(t, \beta_0^T \mathbf{z}) h_1 \\ \lambda_{01}(t, \beta_0^T \mathbf{z}) h_2 \end{pmatrix}.
\end{aligned} \tag{3.61}$$

By Lemma 3.2 and (3.58), we have

$$\begin{aligned}
& \{n S_n^\beta(t, \mathbf{z})\}^{-1} \sum_{i=1}^n \int_0^\tau K_{\mathbf{h}}(s-t, \beta^T (\mathbf{Z}_i(s) - \mathbf{z})) \begin{pmatrix} 1 \\ (s-t)/h_1 \\ \beta^T (\mathbf{Z}_i(s) - \mathbf{z})/h_2 \end{pmatrix} \\
& \quad \times \lambda_{01}(t, \beta_0^T \mathbf{z})(\beta_0 - \beta)^T (\mathbf{Z}_i(s) - \mathbf{z}) Y_i(s) ds \\
= \; & \begin{pmatrix} \lambda_{01}(t, \beta_0^T \mathbf{z})(\beta - \beta_0)^T [\mathbf{z} - \frac{E[Y(t)\mathbf{Z}(t)|U(t)=u]}{E[Y(t)|U(t)=u]}] \\ O(\|\beta_0 - \beta\| h_1) \\ O(\|\beta_0 - \beta\| h_2) \end{pmatrix} + O(\|\beta_0 - \beta\| c_n).
\end{aligned} \tag{3.62}$$

Again, by Lemma 3.2 and (3.58), we have

$$\begin{aligned}
&\{nS_n^{\beta}(t,\mathbf{z})\}^{-1}\sum_{i=1}^{n}\int_0^{\tau}K_{\mathbf{h}}(s-t,\beta^T(\mathbf{Z}_i(s)-\mathbf{z}))\begin{pmatrix}1\\(s-t)/h_1\\ \beta^T(\mathbf{Z}_i(s)-\mathbf{z})/h_2\end{pmatrix}\\
&\qquad (s-t,\beta^T(\mathbf{Z}_i(s)-\mathbf{z}))\mathbf{H}(t,\beta_0^T\mathbf{z})(s-t,\beta^T(\mathbf{Z}_i(s)-\mathbf{z}))^T Y_i(s)ds \qquad (3.63)\\
&=\begin{pmatrix}\lambda_{20}(t,\beta_0^T\mathbf{z})h_1^2\mu_2+\lambda_{02}(t,\beta_0^T\mathbf{z})h_2^2\mu_2\\ O(h_1^3)+O(h_1h_2^2)\\ O(h_2^3)+O(h_1^2h_2)\end{pmatrix}+O(\{h_1+h_2\}c_n)
\end{aligned}$$

and

$$\begin{aligned}
&\{nS_n^{\beta}(t,\mathbf{z})\}^{-1}\sum_{i=1}^{n}\int_0^{\tau}K_{\mathbf{h}}(s-t,\beta^T(\mathbf{Z}_i(s)-\mathbf{z}))\begin{pmatrix}1\\(s-t)/h_1\\ \beta^T(\mathbf{Z}_i(s)-\mathbf{z})/h_2\end{pmatrix}R_n(s,\mathbf{Z}_i(s)-\mathbf{z})Y_i(s)ds\\
&=O(h_1^3+h_2^3+h_2\|\beta_0-\beta\|+\|\beta_0-\beta\|^2). \qquad (3.64)
\end{aligned}$$

Consider the noise term in (3.59). By (3.58), we have

$$\begin{aligned}
&\{nS_n^{\beta}(t,\mathbf{z})\}^{-1}\sum_{i=1}^{n}\int_0^{\tau}K_{\mathbf{h}}(s-t,\beta^T(\mathbf{Z}_i(s)-\mathbf{z}))\begin{pmatrix}1\\(s-t)/h_1\\ \beta^T(\mathbf{Z}_i(s)-\mathbf{z})/h_2\end{pmatrix}dM_i(s)\\
&=\{n\phi_{\beta}(t,\beta^T\mathbf{z})\}^{-1}\sum_{i=1}^{n}\int_0^{\tau}K_{\mathbf{h}}(s-t,\beta^T(\mathbf{Z}_i(s)-\mathbf{z}))\begin{pmatrix}1\\(s-t)/h_1\\ \beta^T(\mathbf{Z}_i(s)-\mathbf{z})/h_2\end{pmatrix}dM_i(s) \qquad (3.65)\\
&\quad+O(c_n\sqrt{\log n/(nh_1h_2)}).
\end{aligned}$$

Combining (3.59)-(3.65), we complete Lemma 3.3.

□

Proof of Lemma 3.4. Without confusion, denote

$$\frac{\partial\hat{\lambda}(s,\beta_0^T\mathbf{Z}_i(s))}{\partial\beta_{-1}}=\frac{\partial\hat{\lambda}(s,\beta^T\mathbf{Z}_i(s);\beta)}{\partial\beta_{-1}}\Big|_{\beta=\beta_0}.$$

By the definition of $M_i(t)=N_i(t)-\int_0^t Y_i(s)\lambda(s,\beta_0^T\mathbf{Z}_i(s))ds$, $t\in[0,\tau]$, which is a martingale process with the filtration $\mathcal{F}_{t,i}$ and $\mathcal{F}_t=\vee_{i=1}^n\mathcal{F}_{t,i}$, we have

$$\frac{1}{\sqrt{n}}\frac{\partial\hat{\ell}_n(\beta_0)}{\partial\beta_{-1}}=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big[\int_0^{\tau}\frac{\partial\hat{\lambda}(s,\beta_0^T\mathbf{Z}_i(s))/\partial\beta_{-1}}{\hat{\lambda}(s,\beta_0^T\mathbf{Z}_i(s))}dN_i(s)-\int_0^{\tau}\frac{\partial\hat{\lambda}(s,\beta_0^T\mathbf{Z}_i(s))}{\partial\beta_{-1}}Y_i(s)ds\Big]$$

$$
\begin{aligned}
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^\tau \frac{\partial\hat{\lambda}(s,\beta_0^T\mathbf{Z}_i(s))/\partial\beta_{-1}}{\hat{\lambda}(s,\beta_0^T\mathbf{Z}_i(s))}dM_i(s) \\
&\quad -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^\tau \frac{\partial\hat{\lambda}(s,\beta_0^T\mathbf{Z}_i(s))/\partial\beta_{-1}}{\hat{\lambda}(s,\beta_0^T\mathbf{Z}_i(s))}\{\hat{\lambda}(s,\beta_0^T\mathbf{Z}_i(s))-\lambda(s,\beta_0^T\mathbf{Z}_i(s))\}Y_i(s)ds \\
&= \mathbf{R}_{n,1}-\mathbf{R}_{n,2}. \qquad (3.66)
\end{aligned}
$$

Note that $\partial\hat{\lambda}(s,\beta_0^T\mathbf{Z}_i(s))/\partial\beta_{-1}$ and $\hat{\lambda}(s,\beta_0^T\mathbf{Z}_i(s))$, $s\in[0,\tau]$ in $\mathbf{R}_{n,1}$ use all observational information and leads to unpredictability. Instead, we apply Theorem 4 in Mammen and Nielsen (2007) to address the problem. Toward this end, define

$$
\begin{aligned}
\bar{f}_i^{(n)}(s) &= \frac{1}{\sqrt{n}}\frac{\partial\hat{\lambda}(s,\beta_0^T\mathbf{Z}_i(s))/\partial\beta_{-1}}{\hat{\lambda}(s,\beta_0^T\mathbf{Z}_i(s))}, \\
\tilde{f}_i^{(n)}(s) &= \frac{1}{\sqrt{n}}\mathbf{J}^T(\beta_{-1}^0)\Big[\mathbf{Z}_i(s)-\frac{E[Y(s)\mathbf{Z}(s)|U_{0i}(s)]}{E[Y(s)|U_{0i}(s)]}\Big]\frac{\lambda_{01}(s,U_{0i})}{\lambda(s,U_{0i}(s))},
\end{aligned}
$$

and

$$
f_i^{*,(n)}(s)=\frac{1}{\sqrt{n}}\frac{\partial\hat{\lambda}^{[-i]}(s,\beta_0^T\mathbf{Z}_i(s))/\partial\beta_{-1}}{\hat{\lambda}^{[-i]}(s,\beta_0^T\mathbf{Z}_i(s))}, \quad f_{i,j}^{**,(n)}(s)=\frac{1}{\sqrt{n}}\frac{\partial\hat{\lambda}^{[-i,j]}(s,\beta_0^T\mathbf{Z}_i(s))/\partial\beta_{-1}}{\hat{\lambda}^{[-i,j]}(s,\beta_0^T\mathbf{Z}_i(s))},
$$

where $\hat{\lambda}^{[-i]}(\cdot,\cdot)$ and $\hat{\lambda}^{[-i,j]}(\cdot,\cdot)$ are the leave-one-out and leave-two-out versions of $\hat{\lambda}(\cdot,\cdot)$, respectively. Let $h_i^{(n)}(s)=f_i^{*,(n)}(s)-\tilde{f}_i^{(n)}(s)$ and $h_{i,j}^{(n)}(s)=f_{i,j}^{**,(n)}(s)-\tilde{f}_i^{(n)}(s)$. According to Theorem 4 in Mammen and Nielsen (2007), we need to calculate the orders of $\sum_{i=1}^n\int_0^\tau(\bar{f}_i^{(n)}(s)-f_i^{*,(n)}(s))dM_i(s)$, $\sum_{i=1}^n\rho_i^2$ and $\sum_{i=1}^n\varrho_i^2$, where

$$
\rho_i=\Big[E\int_0^\tau\{h_i^{(n)}(s)\}^2\lambda(s,\beta_0^T\mathbf{Z}_i(s))Y_ids\Big]^{1/2},
$$

$$
\varrho_i=\max_{1\leqslant j\leqslant n}\Big[E\int_0^\tau\{h_i^{(n)}(s)-h_{i,j}^{(n)}(s)\}^2\lambda(s,\beta_0^T\mathbf{Z}_i(s))Y_ids\Big]^{1/2}.
$$

Next, we consider these three approximate error. It follows from conditions (C1)

and (C4) that, $j,k=0,1,2$,

$$\sup_{\substack{s\in[0,\tau],u\in\mathcal{U}_\beta,\\ \|\beta-\beta_0\|\leqslant cn^{-1/2}}} |S_{jk}(t,u;\beta)-S_{jk}^{[-i]}(t,u;\beta)| = O_p((nh_1h_2)^{-1}),$$

$$\sup_{\substack{s\in[0,\tau],u\in\mathcal{U}_\beta,\\ \|\beta-\beta_0\|\leqslant cn^{-1/2}}} |T_{jk}(t,u;\beta)-T_{jk}^{[-i]}(t,u;\beta)| = O_p((nh_1h_2)^{-1}),$$

uniformly for $i=1,\cdots,n$. Thus, it is easy to show that

$$\sup_{\substack{s\in[0,\tau],u\in\mathcal{U}_\beta,\\ \|\beta-\beta_0\|\leqslant cn^{-1/2}}} \left|\hat{\lambda}(s,u)-\hat{\lambda}^{[-i]}(s,u)\right| = O_p((nh_1h_2)^{-1}).$$

This leads to

$$\sup_{s\in[0,\tau]} \left|\bar{f}_i^{(n)}(s)-f_i^{*,(n)}(s)\right| = O_p(n^{-1/2}(nh_1h_2)^{-1}). \tag{3.67}$$

Thought $\int^t(\bar{f}_i^{(n)}(s)-f_i^{*,(n)}(s))dM_i(s), t\in[0,\tau]$, may be not a martingale process with the filtration $\mathcal{F}_{t,i}$, it is a Lebesgue-Stieltjes integration. Then, by the law of large numbers, we obtain that

$$\begin{aligned}\sum_{i=1}^n\int_0^\tau(\bar{f}_i^{(n)}(s)-f_i^{*,(n)}(s))dM_i(s) &= nO_p(n^{-1/2}(nh_1h_2)^{-1})\\ &= O_p((n^{1/2}h_1h_2)^{-1}).\end{aligned} \tag{3.68}$$

Secondly, consider $\sum_{i=1}^n\rho_i^2$. Combining Proposition 3.3 and Exp. (3.67), we have

$$\begin{aligned}\{h_i^{(n)}(s)\}^2 &= [\{f_i^{*,(n)}(s)-\bar{f}_i^{(n)}(s)\}+\{\bar{f}_i^{(n)}(s)-\tilde{f}_i^{(n)}(s)\}]^2\\ &\leqslant 2\{f_i^{*,(n)}(s)-\bar{f}_i^{(n)}(s)\}^2+2\{\bar{f}_i^{(n)}(s)-\tilde{f}_i^{(n)}(s)\}^2\\ &\leqslant cn^{-1}[(nh_1h_2)^{-2}+(h_1^2+h_2^2+h_2^{-1}\sqrt{\log n/(nh_1h_2)})^2].\end{aligned}$$

Thus, we have

$$\sum_{i=1}^n\rho_i^2 = O_p((nh_1h_2)^{-2}+(h_1^2+h_2^2+h_2^{-1}\sqrt{\log n/(nh_1h_2)})^2). \tag{3.69}$$

Finally, consider $\sum_{i=1}^{n} \varrho_i^2$. By similar arguments of Exp. (3.67), we obtain that

$$\max_{1\leqslant j\leqslant n}\{h_i^{(n)}(s) - h_{i,j}^{(n)}(s)\}^2 = \max_{1\leqslant j\leqslant n}\{f_i^{*,(n)}(s) - f_{i,j}^{**,(n)}(s)\}^2 = O_p(n^{-1}(nh_1h_2)^{-2}).$$

Thus, we have

$$\sum_{i=1}^{n} \varrho_i^2 = O_p((nh_1h_2)^{-2}). \tag{3.70}$$

Combining (3.68), (3.69) and (3.70), Theorem 4 in Mammen and Nielsen (2007) implies that

$$\begin{aligned}
&\mathbf{R}_{n,1} - \frac{1}{\sqrt{n}}\mathbf{J}^T(\beta_{-1}^0)\sum_{i=1}^{n}\int_0^\tau \Big[\mathbf{Z}_i(s) - \frac{E[Y(s)\mathbf{Z}(s)|U_{0i}(s)]}{E[Y(s)|U_{0i}(s)]}\Big]\frac{\lambda_{01}(s,U_{0i}(s))}{\lambda(s,U_{0i}(s))}dM_i(s)\\
=\;& O_P\Big(\sum_{i=1}^{n}\int_0^\tau (\bar{f}_i^{(n)}(s) - f_i^{*,(n)}(s))dM_i(s) + \{\sum_{i=1}^{n}\rho_i^2\}^{1/2} + \{\sum_{i=1}^{n}\varrho_i^2\}^{1/2}\Big)\\
=\;& O_p((n^{1/2}h_1h_2)^{-1}) + O_p((nh_1h_2)^{-1} + h_1^2 + h_2^2 + h_2^{-1}\sqrt{\log n/(nh_1h_2)}) + O_p((nh_1h_2)^{-1}).
\end{aligned}$$

Thus, as $h_1 \to 0, h_2 \to 0$, $n \to \infty$, $nh_1h_2^3/\log n \to \infty$ and $nh_1^2h_2^2 \to \infty$, we have

$$\begin{aligned}
\mathbf{R}_{n,1} \;=\;& \frac{1}{\sqrt{n}}\mathbf{J}^T(\beta_{-1}^0)\sum_{i=1}^{n}\int_0^\tau \Big[\mathbf{Z}_i(s) - \frac{E[Y(s)\mathbf{Z}(s)|U_{0i}(s)]}{E[Y(s)|U_{0i}(s)]}\Big]\frac{\lambda_{01}(s,U_{0i}(s))}{\lambda(s,U_{0i}(s))}dM_i(s)\\
&+o_p(1).
\end{aligned} \tag{3.71}$$

For the term $\mathbf{R}_{n,2}$, by the proof of Proposition 3.3, we have

$$\begin{aligned}
\frac{\partial\hat{\lambda}(s,\beta_0^T\mathbf{Z}_i(s))/\partial\beta_{-1}}{\hat{\lambda}(s,\beta_0^T\mathbf{Z}_i(s))} \;=\;& \frac{\lambda_{01}(s,U_{0i}(s))}{\lambda(s,U_{0i}(s))}[\mathbf{Z}_i(s) - \frac{E[Y(s)\mathbf{Z}(s)|U_{0i}(s)]}{E[Y(s)|U_{0i}(s)]}]\\
&+O_p(h_1^2 + h_2^2 + h_2^{-1}\sqrt{\log n/(nh_1h_2)}),
\end{aligned}$$

uniformly for $s \in [0,\tau]$ and $i = 1,\cdots,n$. Note that the order $h_2^{-1}\sqrt{\log n/(nh_1h_2)}$ of above expression is based on (3.39), which is a martingale integral. By similar proofs

of $\mathbf{R}_{n,1}$ with respective to unpredictability, we obtain

$$\begin{aligned}\mathbf{R}_{n,2} &= O_p(c_n)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^\tau \frac{\lambda_{01}(s,U_{0i}(s))}{\lambda(s,U_{0i}(s))}\Big[\mathbf{Z}_i - \frac{E[Y(s)\mathbf{Z}(s)|U_{0i}(s)]}{E[Y(s)|U_{0i}(s)]}\Big]Y_i(s)ds \\ &\quad +\sqrt{n}O_p(c_n)O_p(h_1^2+h_2^2) + O_p(c_n)O_p(h_2^{-1}\sqrt{\log n/(nh_1h_2)}) \\ &= o_p(1),\end{aligned}$$

if $nh_1^8 \to 0$ and $nh_2^8 \to 0$. This together with (3.71) proves this lemma. □

Proof of Lemma 3.5. By the definition of $M_i(t) = N_i(t) - \int_0^t Y_i(s)\lambda(s,\beta_0^T\mathbf{Z}_i(s))ds$, $t \in [0,\tau]$, simple algebra gives that

$$\begin{aligned}\frac{1}{n}\frac{\partial^2\hat{\ell}_n(\beta_0)}{\partial\beta_{-1}\partial\beta_{-1}^T} &= \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau\Big[\frac{\partial^2\hat{\lambda}(s,\beta_0^T\mathbf{Z}_i(s))/\partial\beta_{-1}\partial\beta_{-1}^T}{\hat{\lambda}(s,\beta_0^T\mathbf{Z}_i(s))} - \Big(\frac{\partial\hat{\lambda}(s,\beta_0^T\mathbf{Z}_i(s))/\partial\beta_{-1}}{\hat{\lambda}(s,\beta_0^T\mathbf{Z}_i(s))}\Big)^{\otimes 2}\Big]dN_i(s) \\ &\quad -\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau\frac{\partial^2\hat{\lambda}(s,\beta_0^T\mathbf{Z}_i(s))}{\partial\beta_{-1}\partial\beta_{-1}^T}Y_i(s)ds \\ &= -\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau\Big(\frac{\partial\hat{\lambda}(s,\beta_0^T\mathbf{Z}_i(s))/\partial\beta_{-1}}{\hat{\lambda}(s,\beta_0^T\mathbf{Z}_i(s))}\Big)^{\otimes 2}dN_i(s) \\ &\quad +\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau\frac{\partial^2\hat{\lambda}(s,\beta_0^T\mathbf{Z}_i(s))}{\partial\beta_{-1}\partial\beta_{-1}^T}\Big[\frac{1}{\hat{\lambda}(s,\beta_0^T\mathbf{Z}_i(s))} - \frac{1}{\lambda(s,\beta_0^T\mathbf{Z}_i(s))}\Big]dM_i(s) \\ &= E_{n,1} + E_{n,2}.\end{aligned}$$

Clearly, Proposition 3.3 implies that

$$\sup_{s\in[0,\tau],Z(s)\in\mathcal{Z}}\Big\|\frac{\partial\hat{\lambda}(s,\beta_0^T\mathbf{Z}(s))/\partial\beta_{-1}}{\hat{\lambda}(s,\beta_0^T\mathbf{Z}(s))} - \mathbf{J}^T(\beta_{-1}^0)\Big[\mathbf{Z}(s) - \frac{E[Y(s)\mathbf{Z}(s)|\beta_0^T\mathbf{Z}(s)]}{E[Y(s)|\beta_0^T\mathbf{Z}(s)]}\Big]\frac{\lambda_{01}(s,\beta_0^T\mathbf{Z}(s))}{\lambda(s,\beta_0^T\mathbf{Z}(s))}\Big\| = o_p(1).$$

Thus, by the martingale central limit theorem and the law of large numbers, we have

$$E_{n,1} = -\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau\Big[\mathbf{J}^T(\beta_{-1}^0)\Big\{\mathbf{Z}(s) - \frac{E[Y(s)\mathbf{Z}(s)|\beta_0^T\mathbf{Z}(s)]}{E[Y(s)|\beta_0^T\mathbf{Z}(s)]}\Big\}\frac{\lambda_{01}(s,\beta_0^T\mathbf{Z}(s))}{\lambda(s,\beta_0^T\mathbf{Z}(s))}\Big]^{\otimes 2}dN_i(s) + o_p(1)$$

$$
\begin{aligned}
&= -\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \Big[\mathbf{J}^T(\beta_{-1}^0)\Big\{\mathbf{Z}(s) - \frac{E[Y(s)\mathbf{Z}(s)|\beta_0^T\mathbf{Z}(s)]}{E[Y(s)|\beta_0^T\mathbf{Z}(s)]}\Big\}\frac{\lambda_{01}(s,\beta_0^T\mathbf{Z}(s))}{\lambda(s,\beta_0^T\mathbf{Z}(s))}\Big]^{\otimes 2} dM_i(s) \\
&\quad -\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \Big[\mathbf{J}^T(\beta_{-1}^0)\Big\{\mathbf{Z}(s) - \frac{E[Y(s)\mathbf{Z}(s)|\beta_0^T\mathbf{Z}(s)]}{E[Y(s)|\beta_0^T\mathbf{Z}(s)]}\Big\}\frac{\lambda_{01}(s,\beta_0^T\mathbf{Z}(s))}{\lambda(s,\beta_0^T\mathbf{Z}(s))}\Big]^{\otimes 2} \\
&\quad \lambda(s,\beta_0^T\mathbf{Z}_i)Y_i(s)ds + o_p(1) \\
&= -\mathbf{J}^T(\beta_{-1}^0)E\Big[\int_0^\tau \Big\{\mathbf{Z}(s) - \frac{E[Y(s)\mathbf{Z}(s)|\beta_0^T\mathbf{Z}(s)]}{E[Y(s)|\beta_0^T\mathbf{Z}(s)]}\Big\}^{\otimes 2}\frac{\lambda_{01}^2(s,\beta_0^T\mathbf{Z}(s))}{\lambda(s,\beta_0^T\mathbf{Z}(s))}Y(s)ds\Big] \\
&\quad \mathbf{J}(\beta_{-1}^0) + o_p(1). \qquad (3.72)
\end{aligned}
$$

Under the regular conditions, it can be shown that $\frac{\hat{\lambda}(s,\beta_0^T\mathbf{Z}_i(s))}{\partial\beta_{-1}\partial\beta_{-1}^T} = O_p(1)$. This together with Proposition 3.3 yields

$$
\xi(s,\beta_0^T\mathbf{Z}_i(s)) = \frac{\partial^2\hat{\lambda}(s,\beta_0^T\mathbf{Z}_i(s))}{\partial\beta_{-1}\partial\beta_{-1}^T}\Big[\frac{1}{\hat{\lambda}(s,\beta_0^T\mathbf{Z}_i(s))} - \frac{1}{\lambda(s,\beta_0^T\mathbf{Z}_i(s))}\Big] = o_p(1).
$$

Therefore, even thought $\xi(s,\beta_0^T\mathbf{Z}_i(s))$ may be not a predictable process with the filtration $\mathcal{F}_{t,i}$, we have

$$
E_{n,2} = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \xi(s,\beta_0^T\mathbf{Z}_i(s))[dN_i(s) + Y_i(s)ds] = o_p(1), \qquad (3.73)
$$

by the law of large numbers. Combining (3.72) with (3.73), Lemma 3.5 is proved. □

Part II

for the Alzheimer's Disease Neuroimaging Initiative*

*Data used in preparation of this part were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

Chapter 4

Multiple Testing of Genetic Association with Longitudinal Phenotypes for Large-Scale ADNI GWAS

4.1 Introduction

The Alzheimer's Disease Neuroimaging Initiative (ADNI) is a longitudinal study uniting transdisciplinary research fields to investigate the progression of Alzheimer's Disease (AD) for better clinical prevention and treatment. One key aim of the ADNI is to provide researchers the combined genetics and clinical data to help investigate mechanisms of Alzheimer's disease. It has been extended once and again, and currently composes of ADNI 1 (launched in 2003 and lasted 5 years), ADNI GO (launched in 2009 and lasted 2 years), ADNI 2 (launched in 2011 and lasted 5 years), and ADNI 3 (launched in 2016 and will last 5 years), generating genotyping and sequencing data including ANDI 1 GWAS and ADNI GO/2 GWAS, and ADNI WGS, refer to the webpage, http://adni.loni.usc.edu/about/.

Phenotypes such as disease-progression or severity used to be longitudinally collected at multiple time points in the ADNI GWAS, and can not be analyzed through traditional association studies for cross-sectional phenotypic data observed at a single

time point in most existing GWAS. This raises the challenge to identify the association between Single Nucleotide Polymorphisms (SNPs) level genetic variants and repeatedly measured human phenotypes (He et al. (2015) and Visscher et al. (2017)). In this chapter, we aim to provide practical strategies and detailed procedures to test the association between a random trajectory of longitudinal phenotype outcomes and SNP-level genotypes from the perspective of functional data analysis. This is in the sense that we model the AD-related phenotype response observed on irregular time points as sparse functional data, and detect functional genotype effects while controlling the confounding effects of environmental covariates (Ramsay and Silverman (2005) and Yao et al. (2005)). The question turns to test if the mean phenotype trajectories differ across different genotypes.

Within the ADNI 1 GWAS, the real problem is to make multiple comparison

$$H_0 : \theta_1 = \theta_2 = \theta_3, \quad \text{vs} \quad H_1 : \text{not all } \theta_i\text{'s are equal}, \tag{4.1}$$

where θ_i's represent the functional genotype effects of three genetic groups according to genetic traits on the longitudinal phenotypes. In the ADNI cohort, a SNP is disease-related if $\theta_i's$ are different across the genotypes defined by a SNP. Such hypothesis testing problems widely exist in many non-GWAS settings, e.g. longitudinal AIDS clinical trial data Li (2011), and therefore is worth investigating in its own right.

There is a large volume of recent literature on methods and applications of functional linear models and functional analysis of variance (fANOVA) under various designs (Brumback and Rice, 1998; Morris and Carroll, 2006; Zhang and Chen, 2007; Zhou et al., 2010; Li et al., 2015; Xu et al., 2018). For the fANOVA test (4.1) and analog, existing work mainly considers dense functional data with Gaussian-type responses, where observations on each curve are made on a dense grid. A comprehensive account of fANOVA methods for dense functional data and Gaussian-type

responses is provided in the monograph of Zhang (2013). Reimherr and Nicolae (2014) and Huang et al. (2017) also applied similar test procedures in genetic studies. This brings out two concerns. One is that their test statistics were based on the integrated square error rather than the likelihood, and hence are not applicable to data with non-Gaussian type response; The other is that the available asymptotic theories were developed for dense functional data, which lead to χ^2 mixture limiting distributions for the test statistics, and are not applicable to sparse longitudinal data. The nonparametric test of Tang et al. (2016) is remedial for such longitudinal GWAS test by applying the working-independent estimation to build the generalized likelihood ratio (GQLR) test within a semiparametric partially generalized linear regression model (Lin and Carroll, 2001).

One will encounter two major challenges when applying the GQLR test to multiple hypotheses testing in large-scale longitudinal GWAS data. One is the computational infeasibility to run bootstrap for hundreds of thousands of SNPs. The other is that it requires a gigantic bootstrap sample to reach genome-wide significance levels 10^{-7} (Fadista et al., 2016; Huang et al., 2017). Notice that GQLR enjoys a property called the Wilks phenomenon (Fan et al., 2001), meaning that the null distribution of the test statistic does not depend on the unknown model parameters. We therefore suggest to select a small number of SNPs randomly and to fit a χ^2 distribution to the bootstrap sample using maximum likelihood estimation and use the fitted χ^2 distribution to determine the p-values for all SNPs. Meanwhile, within a quite general semiparametric generalized linear model, we present the complete F-test procedure based on the description of Zhang (2013) for the large-scale longitudinal test.

In the remainder of this chapter, we model the longitudinal phenotype data by a class of generalized functional concurrent linear models, where the responses are allowed to be either Gaussian or non-Gaussian. Then for a general hypothesis statement that incorporates the multiple treatment test in (4.1), we provide two test pro-

cedures and apply to the large-scale longitudinal ADNI data, where the responses are Alzheimer related phenotypes modeled as sparse functional data. Data used in the preparation of this and next chapters were obtained from the ADNI database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early AD. Simulation at the real data set is conducted to assess the performances of the two tests.

4.2 Functional Modeling of Longitudinal Phenotype Data and Estimation Procedure

4.2.1 Model and Hypotheses

Let $Y_i(t)$ be the phenotype of the ith subject observed at time $t \in \mathcal{T}$, $i = 1, \ldots, n$, where $\mathcal{T}$ is a closed time interval, $\boldsymbol{X}_i(t)$ is a p-dim subject-specific covariate vector representing environmental confounders, and $\mathbf{Z}_i$ is a time-variant q-dim genetic predictor. Suppose that $E\{Y_i(t)|\mathbf{Z}_i, \boldsymbol{X}_i(s), s \in \mathcal{T}\} = E\{Y_i(t)|\mathbf{Z}_i, \boldsymbol{X}_i(t)\} = \mu_i(t)$ (Pepe and Couper, 1997). We propose a generalized functional concurrent linear model

$$g\{\mu_i(t)\} = \boldsymbol{X}_i^{\mathrm{T}}(t)\boldsymbol{\beta} + \mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\theta}(t), \tag{4.2}$$

where $g(\cdot)$ is a known monotonic and differentiable link function, $\boldsymbol{\beta}$ is a p-vector of unknown coefficients representing environmental effects and $\boldsymbol{\theta}(t) \equiv (\theta_1, \ldots, \theta_q)^{\mathrm{T}} = (\theta_1(t), \ldots, \theta_q(t))^{\mathrm{T}}$ is a vector of unknown smooth functions representing the genotype effect. The parameter $\boldsymbol{\beta}$ is merely used to control the confounding effect of the environment covariates, and the primary interest is to make inference on the functional genotype effects $\boldsymbol{\theta}$. When $\mathbf{Z}$ is vector of group indicators, model (4.2) reduces to

the functional analysis of covariance (fANCOVA) model (Zhang, 2013; Tang et al., 2016) since the treatment effect for genotype k is represented by a nonparametric function $\theta_k(t)$. In semiparametric regression literature, model (4.2) is also referred to as a generalized partially linear varying coefficient model. We are interested in testing the following general hypotheses

$$H_0 : \boldsymbol{C}\boldsymbol{\theta}(t) = \boldsymbol{c}(t) \quad \text{vs.} \quad H_1 : \boldsymbol{C}\boldsymbol{\theta}(t) \neq \boldsymbol{c}(t), \tag{4.3}$$

where $\boldsymbol{C}$ is $r \times q$ matrix of linear contrasts, $\boldsymbol{c}(t)$ is an r-dim function, and $r = \text{rank}(\boldsymbol{C}) \leqslant q$. Hypotheses (4.3) reduce to

$$H_0 : \theta_1(t) = \cdots = \theta_q(t), \quad \text{vs} \quad H_1 : \text{not all } \theta_k\text{'s are equal}, \tag{4.4}$$

in Tang et al. (2016) when $\boldsymbol{C}$ is an identity matrix and $\boldsymbol{c}(t) = \boldsymbol{0}$.

In the ADNI, some of the most important Alzheimer-related phenotypes include the hippocampal volume, the decay of which is known to be related to memory loss (Schuff et al., 2009), and the Rey Auditory Verbal Learning Test (RAVLT) score; some environmental covariates include age, sex, education, marital status, etc.; and a genetic predictor can be the genotypes defined by a SNP, which is AA, AB or BB defined by the two alleles. To include the effects of one SNP in model (4.2), $\mathbf{Z}_i$ is a 3-dim vector of indicators for the three genotypes. The practical hypotheses (4.1) is a special case of (4.4) for $q = 3$ and a special case of (4.3) with $\boldsymbol{\theta}(t) = (\theta_1, \theta_2, \theta_3)^{\mathrm{T}}(t)$.

Although model (4.2) is defined in continuum, observations on $Y_i(t)$ and $\boldsymbol{X}_i(t)$ are, in practice, made on discrete and subject-specific time points. Let $\boldsymbol{T}_i = (T_{i1}, \cdots, T_{im_i})^{\mathrm{T}}$ be the random observation time points for subject i in genotype k, where m_i is the number of repeated measurements. Denote $\boldsymbol{Y}_i = (Y_{i1}, \cdots, Y_{im_i})^{\mathrm{T}}$, $\mu_i = (\mu_{i1}, \cdots, \mu_{im_i})^{\mathrm{T}}$, $\boldsymbol{X}_i = (\boldsymbol{X}_{i1}, \cdots, \boldsymbol{X}_{im_i})^{\mathrm{T}}$, where $Y_{ij} = Y_i(T_{ij})$, $\mu_{ij} = \mu_i(T_{ij})$ and $\boldsymbol{X}_{ij} = \boldsymbol{X}_i(T_{ij})$. We assume the conditional covariance of $Y_i(t)$ is a bivariate positive semidefinite function

$$\mathcal{R}(t_1, t_2) = \text{cov}\left\{Y_i(t_1), Y_i(t_2) | \boldsymbol{X}_i(s), s \in \mathcal{T}\right\}, \quad \text{for any } t_1, t_2 \in \mathcal{T}. \tag{4.5}$$

The covariance structure is assumed to be the same across subjects. Let $\boldsymbol{\Sigma}_i = \text{cov}(\boldsymbol{Y}_i \mid \boldsymbol{X}_i, \boldsymbol{T}_i) = \{\mathcal{R}(T_{ij}, T_{ij'})\}_{j,j'=1}^{m_i}$ be the subject-specific covariance matrices. Since the true covariance function $\mathcal{R}$ is unknown, the covariance model $\mathcal{V}(t_1, t_2)$ adopted in data analysis is commonly referred to as a "working" covariance, which is subject to misspecification. Historically, a working covariance model is usually assumed to be member of a parametric family, such as the Matérn family. Let $\boldsymbol{V}_i = \{\mathcal{V}(T_{ij}, T_{ij'})\}_{j,j'=1}^{m_i}$ be the "working" covariance matrix for subject (k, i), which is the interpolation of the continuous covariance function $\mathcal{V}$ on the subject-specific time points. The simplest working covariance is the working independence (WI), i.e. $\boldsymbol{V}_i = \boldsymbol{I}_{m_i}$. It is known that misspecified working covariance can still lead to consistent estimators, although such estimators are not semiparametric efficient (Wang et al., 2005).

We refer to the model under the null hypothesis in (4.3) as the reduced model and that under the alternative hypothesis as the full model. Denote $\widehat{\boldsymbol{\beta}}_R$ and $\widehat{\boldsymbol{\theta}}_R(t)$ as the estimators under the reduced model and $\widehat{\boldsymbol{\beta}}_F$ and $\widehat{\boldsymbol{\theta}}_F(t)$ as the estimators under the full model. Our estimation procedures under both models are based on profile-kernel estimating equations.

4.2.2 Estimation Under the Full Model

We first consider estimation under the full model. By Taylor's expansion, for any T_{ij} in a neighborhood h of t, $\boldsymbol{\theta}(T_{ij})$ can be approximated locally by a linear polynomial

$$\boldsymbol{\theta}(T_{ij}) \approx \boldsymbol{\theta}(t) + \boldsymbol{\theta}'(t)(T_{ij} - t) = \boldsymbol{\alpha}_0 + \boldsymbol{\alpha}_1(T_{ij} - t)/h.$$

Let $K(\cdot)$ be a symmetric probability density function and denote $K_h(t) = h^{-1}K(t/h)$ where h is the bandwidth. Put $\mathbb{T}_i = (\boldsymbol{T}_i - t)/h$, $\boldsymbol{U}_{ij}(t) = \{\mathbf{Z}_i^{\mathrm{T}}, \mathbf{Z}_i^{\mathrm{T}}(T_{ij} - t)/h\}^{\mathrm{T}}$, and $\boldsymbol{U}_i(t) = \{\boldsymbol{U}_{i1}(t), \ldots, \boldsymbol{U}_{im_i}(t)\}^{\mathrm{T}} = (\mathbf{1}\mathbf{Z}_i^{\mathrm{T}}, \mathbb{T}_i\mathbf{Z}_i^{\mathrm{T}})$. For a given $\boldsymbol{\beta}$, $\boldsymbol{\theta}(t)$ is estimated by

solving the following local linear estimating equation regarding $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_0^{\mathrm{T}}, \boldsymbol{\alpha}_1^{\mathrm{T}})^{\mathrm{T}}$,

$$\sum_{i=1}^{n} \boldsymbol{U}_i(t)^{\mathrm{T}} \Delta_i(t) \mathcal{W}_i^{-1} \boldsymbol{K}_h(\boldsymbol{T}_i - t)\{\boldsymbol{Y}_i - \mu_i(t)\} = 0, \tag{4.6}$$

where $\boldsymbol{K}_h(\boldsymbol{T}_i - t) = \mathrm{diag}\{K_h(T_{ij} - t)\}_{j=1}^{m_i}$, $\mu_i(t) = (\mu_{i1}, \ldots, \mu_{im_i})^{\mathrm{T}}(t)$, $\mu_{ij}(t) = g^{-1}\{\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{U}_{ij}^{\mathrm{T}}(t)\boldsymbol{\alpha}\}$, $\Delta_i(t) = \mathrm{diag}\{\mu_{ij}^{(1)}(t)\}_{j=1}^{m_i}$, $\mu_{k,ij}^{(1)}(t)$ is the first derivative of $\mu(\cdot) = g^{-1}(\cdot)$ evaluated at $\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{U}_{ij}^{\mathrm{T}}(t)\boldsymbol{\alpha}$, and $\mathcal{W}_i$ is a weight matrix to be specified below. The local linear estimator is given by $\widehat{\boldsymbol{\theta}}_F(t;\boldsymbol{\beta}) = \widehat{\boldsymbol{\alpha}}_0$, where $(\widehat{\boldsymbol{\alpha}}_0^{\mathrm{T}}, \widehat{\boldsymbol{\alpha}}_1^{\mathrm{T}})^{\mathrm{T}}$ is the solution of (4.6). Then $\widehat{\boldsymbol{\beta}}_F$ is obtained by solving

$$\mathbf{0} = \sum_{i=1}^{n} \left\{ \boldsymbol{X}_i^{\mathrm{T}} + \sum_{k=1}^{q} Z_{ik} \frac{\partial \widehat{\theta}_{k,F}(\boldsymbol{T}_i;\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right\} \Delta_i(\boldsymbol{T}_i) \mathcal{W}_i^{-1} [\boldsymbol{Y}_i - g^{-1}\{\boldsymbol{X}_i\boldsymbol{\beta} + \widehat{\boldsymbol{\theta}}_F^{\mathrm{T}}(\boldsymbol{T}_i;\boldsymbol{\beta})\mathbf{Z}_i\}]. \tag{4.7}$$

At convergence of the algorithm described above, the nonparametric estimator needs a final update as $\widehat{\boldsymbol{\theta}}_F(t) = \widehat{\boldsymbol{\theta}}_F(t, \widehat{\boldsymbol{\beta}}_F)$.

As shown by Lin and Carroll (2001), the most efficient estimators within the class defined by (4.6) are obtained by setting $\mathcal{W}_i = \mathrm{diag}\{\omega(\mu_{ij})\}_{j=1}^{m_i}$ where $\omega(\cdot)$ is a working variance function. Similar kernel estimators are widely used in longitudinal and functional data analysis, see Fan and Li (2004); Yao et al. (2005); Hall et al. (2006); Li and Hsing (2010). The working variance function $\omega(\cdot)$ can be replaced by a nonparametric variance estimator described in Section 4.4.

Under the special case of identical link, $g(x) = x$, the solution of (4.6) is

$$\widehat{\boldsymbol{\alpha}} = \left\{ \sum_{i=1}^{n} \boldsymbol{U}_i(t)^{\mathrm{T}} \mathcal{W}_i^{-1} \boldsymbol{U}_i(t) \right\}^{-1} \left\{ \sum_{i=1}^{n} \boldsymbol{U}_i(t)^{\mathrm{T}} \mathcal{W}_i^{-1} (\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta}) \right\},$$

then

$$\widehat{\boldsymbol{\theta}}(t,\boldsymbol{\beta}) = (\boldsymbol{I}, \mathbf{0}) \left\{ \sum_{i=1}^{n} \boldsymbol{U}_i(t)^{\mathrm{T}} \mathcal{W}_i^{-1} \boldsymbol{U}_i(t) \right\}^{-1} \left\{ \sum_{i=1}^{n} \boldsymbol{U}_i(t)^{\mathrm{T}} \mathcal{W}_i^{-1} (\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta}) \right\},$$

$$\frac{\partial \widehat{\boldsymbol{\theta}}}{\partial \boldsymbol{\beta}^{\mathrm{T}}}(t,\boldsymbol{\beta}) = -(\boldsymbol{I},\boldsymbol{0}) \left\{ \sum_{i=1}^{n} \boldsymbol{U}_i(t)^{\mathrm{T}} \mathcal{W}_i^{-1} \boldsymbol{U}_i(t) \right\}^{-1} \left\{ \sum_{i=1}^{n} \boldsymbol{U}_i(t)^{\mathrm{T}} \mathcal{W}_i^{-1} \boldsymbol{X}_i \right\}.$$

Define $\widetilde{\boldsymbol{X}}_i = (\widetilde{\boldsymbol{X}}_{i1}, \ldots, \widetilde{\boldsymbol{X}}_{im_i})^{\mathrm{T}}$ and $\widetilde{\boldsymbol{Y}}_i = (\widetilde{Y}_{i1}, \ldots, \widetilde{Y}_{im_i})^{\mathrm{T}}$, where

$$\begin{aligned}
\widetilde{\boldsymbol{X}}_{ij}^{\mathrm{T}} &= \boldsymbol{X}_{ij}^{\mathrm{T}} + \mathbf{Z}_i^{\mathrm{T}} \frac{\partial \widehat{\boldsymbol{\theta}}}{\partial \boldsymbol{\beta}^{\mathrm{T}}}(T_{ij},\boldsymbol{\beta}) \\
&= \boldsymbol{X}_{ij}^{\mathrm{T}} - (\mathbf{Z}_i^{\mathrm{T}},\boldsymbol{0}) \left\{ \sum_{i'=1}^{n} \boldsymbol{U}_{i'}(T_{ij})^{\mathrm{T}} \mathcal{W}_{i'}^{-1} \boldsymbol{U}_{i'}(T_{ij}) \right\}^{-1} \left\{ \sum_{i'=1}^{n} \boldsymbol{U}_{i'}^{\mathrm{T}}(T_{ij}) \mathcal{W}_{i'}^{-1} \boldsymbol{X}_{i'} \right\}, \\
\widetilde{Y}_{ij} &= Y_{ij} - (\mathbf{Z}_i^{\mathrm{T}},\boldsymbol{0}) \left\{ \sum_{i'=1}^{n} \boldsymbol{U}_{i'}(T_{ij})^{\mathrm{T}} \mathcal{W}_{i'}^{-1} \boldsymbol{U}_{i'}(T_{ij}) \right\}^{-1} \left\{ \sum_{i'=1}^{n} \boldsymbol{U}_{i'}^{\mathrm{T}}(T_{ij}) \mathcal{W}_{i'}^{-1} \boldsymbol{Y}_{i'} \right\}.
\end{aligned}$$

The solution of (4.7) is

$$\widehat{\boldsymbol{\beta}}_F = \left(\sum_{i=1}^{n} \widetilde{\boldsymbol{X}}_i^{\mathrm{T}} \mathcal{W}_i^{-1} \widetilde{\boldsymbol{X}}_i \right)^{-1} \left(\sum_{i=1}^{n} \widetilde{\boldsymbol{X}}_i^{\mathrm{T}} \mathcal{W}_i^{-1} \widetilde{\boldsymbol{Y}}_i \right).$$

4.2.3 Estimation Under the Reduced Model

We now consider estimation under the reduced model. We first partition the contrast matrix $\boldsymbol{C}$ in (4.3) into $\boldsymbol{C} = (\boldsymbol{C}_1, \boldsymbol{C}_2)$, where $\boldsymbol{C}_1$ is $r \times (q-r)$ and $\boldsymbol{C}_2$ is $r \times r$. Partition $\boldsymbol{\theta}$ accordingly into $(\boldsymbol{\theta}_1^{\mathrm{T}}, \boldsymbol{\theta}_2^{\mathrm{T}})^{\mathrm{T}}$, where $\dim(\boldsymbol{\theta}_2) = r$. Without loss of generality, suppose $\boldsymbol{C}_2$ is full rank, then under the null hypothesis $\boldsymbol{C}\boldsymbol{\theta}(t) = \boldsymbol{c}(t)$,

$$\boldsymbol{\theta}_1(t) = \boldsymbol{C}_1^{-1}\{\boldsymbol{c}(t) - \boldsymbol{C}_2\boldsymbol{\theta}_2(t)\}.$$

By a simple reparameterization, let $\boldsymbol{\vartheta}(t) = \boldsymbol{\theta}_2(t)$, then $\boldsymbol{\theta}(t) = \boldsymbol{c}^*(t) + \boldsymbol{D}\boldsymbol{\vartheta}(t)$, where

$$\boldsymbol{c}^*(t) = \begin{pmatrix} \boldsymbol{C}_1^{-1}\boldsymbol{c}(t) \\ \boldsymbol{0} \end{pmatrix}, \quad \boldsymbol{D} = \begin{pmatrix} -\boldsymbol{C}_1^{-1}\boldsymbol{C}_2 \\ \boldsymbol{I} \end{pmatrix}.$$

For a given $\boldsymbol{\beta}$, the profile local linear estimator for $\boldsymbol{\vartheta}(t)$ is given by $\widehat{\boldsymbol{\vartheta}}(t,\boldsymbol{\beta}) = \widehat{\boldsymbol{a}}_0$,

where $\widehat{\boldsymbol{a}} = (\widehat{\boldsymbol{a}}_0, \widehat{\boldsymbol{a}}_1)^{\mathrm{T}}$ is the solution of

$$\sum_{i=1}^{n} \mathscr{U}_i(t)^T \Delta_i(t) \mathcal{W}_i^{-1} \boldsymbol{K}_h(\boldsymbol{T}_i - t)\{\boldsymbol{Y}_i - \mu_i(t)\} = 0, \tag{4.8}$$

$\mathscr{U}_i(t) = \{\mathscr{U}_{i1}(t), \ldots, \mathscr{U}_{im_i}(t)\}$, $\mu_i(t) = (\mu_{i1}, \ldots, \mu_{im_i})^{\mathrm{T}}(t)$, $\mathscr{U}_{ij}(t) = \{\mathbf{Z}_i^{\mathrm{T}}\boldsymbol{D}, \mathbf{Z}_i^{\mathrm{T}}\boldsymbol{D}(T_{ij} - t)/h\}^{\mathrm{T}}$, $\mu_{ij}(t) = g^{-1}\{\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{Z}_i\boldsymbol{c}^*(t) + \mathscr{U}_{ij}^{\mathrm{T}}(t)\boldsymbol{a}\}$, $\Delta_i(t) = \mathrm{diag}\{\mu_{ij}^{(1)}(t)\}_{j=1}^{m_i}$, and $\mu_{k,ij}^{(1)}(t)$ is the first derivative of $\mu(\cdot) = g^{-1}(\cdot)$ evaluated at $\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{Z}_i\boldsymbol{c}^*(t) + \mathscr{U}_{ij}^{\mathrm{T}}(t)\boldsymbol{a}$.

Denote $\mathscr{Z}_i = \boldsymbol{D}^{\mathrm{T}}\mathbf{Z}_i = (\mathscr{Z}_{i1}, \ldots \mathscr{Z}_{i,q-r})^{\mathrm{T}}$ and $\boldsymbol{\vartheta}(t) = (\vartheta_1, \ldots, \vartheta_{q-r})^{\mathrm{T}}(t)$. The reduced model estimator $\widehat{\boldsymbol{\beta}}_R$ is obtained by solving

$$\begin{aligned}\mathbf{0} = \sum_{i=1}^{n} \left\{ \boldsymbol{X}_i^{\mathrm{T}} + \sum_{j=1}^{q-r} \mathscr{Z}_{ij} \frac{\partial \widehat{\vartheta}_j(\boldsymbol{T}_i; \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right\} \Delta_i(\boldsymbol{T}_i) \mathcal{W}_i^{-1} \\ \times [\boldsymbol{Y}_i - g^{-1}\{\boldsymbol{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\boldsymbol{c}^*(t) + \mathscr{Z}_i^{\mathrm{T}}\widehat{\boldsymbol{\vartheta}}(\boldsymbol{T}_i; \boldsymbol{\beta})\}]. \end{aligned} \tag{4.9}$$

At convergence, the nonparametric estimator is updated as $\widehat{\boldsymbol{\theta}}_R(t) = \boldsymbol{c}^*(t) + \boldsymbol{D}\widehat{\boldsymbol{\vartheta}}(t, \widehat{\boldsymbol{\beta}}_R)$.

4.3 Two Testing Procedures on Genotype Effects

We now direct our focus back to testing the hypotheses in (4.3) and we will discuss two test procedures the Generalized Quasi-Likelihood Ratio (GQLR) test and the functional F-test.

4.3.1 Generalized Quasi-Likelihood Ratio Test

For model (4.2) and longitudinal data set in the chapter, the quasi-likelihood function $\mathcal{Q}$ satisfies

$$\frac{\partial \mathcal{Q}(\mu, \boldsymbol{Y})}{\partial \mu} = \boldsymbol{V}(\mu)^{-1}(\boldsymbol{Y} - \mu), \tag{4.10}$$

where $\boldsymbol{Y}$ is an m-vector of response within a subject, $\mu = g^{-1}\{\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\theta}^{\mathrm{T}}(\boldsymbol{T})\mathbf{Z}\}$ is the conditional mean vector and $\boldsymbol{V}(\mu)$ is a working covariance matrix not necessarily the same as the true covariance $\boldsymbol{\Sigma}(\mu)$.

The quasi-likelihood of the data is

$$\ell(\boldsymbol{\theta}, \boldsymbol{\beta}) = \sum_{i=1}^{n} \mathcal{Q}[g^{-1}\{\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{\theta}^{\mathrm{T}}(\boldsymbol{T}_i)\mathbf{Z}_i\}, \boldsymbol{Y}_i]. \tag{4.11}$$

The GQLR test statistic under the model (4.2) is defined as the difference of the quasi-likelihoods under the full and reduced models

$$\lambda_n(H_0) = \ell(\widehat{\boldsymbol{\theta}}_F, \widehat{\boldsymbol{\beta}}_F) - \ell(\widehat{\boldsymbol{\theta}}_R, \widehat{\boldsymbol{\beta}}_R). \tag{4.12}$$

The GQLR test statistic $\lambda_n(H_0)$ in (4.12) meets the result of Wilks phenomenon if a working independence covariance model is used in both estimation and hypothesis testing and if the true covariance function of the functional response is used guaranteed by Theorem 1 in Tang et al. (2016). Therefore the distribution of $\lambda_n(H_0)$ does not depend on $\boldsymbol{\beta}_0$, $\theta_0(t)$ or the true correlation structure $\mathcal{R}(\boldsymbol{\tau})$, making it easy to assess the approximating distribution of $\lambda_n(H_0)$.

4.3.2 Functional F-Test

Zhang and Chen (2007) and Zhang (2013) (Chapter 6) proposed a functional F-test for hypothesis (4.3), however their test was developed fore Gaussian response under dense functional data setting and without covariates. We now extend their procedure into our setting with the link function $g(\cdot)$ restricted to be an identity link.

Define sum of squares

$$SSH_n(t) = \left\{\boldsymbol{C}\widehat{\boldsymbol{\theta}}(t) - \boldsymbol{c}(t)\right\}^{\mathrm{T}} \left\{\boldsymbol{C}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\boldsymbol{C}^{\mathrm{T}}\right\}^{-1} \left\{\boldsymbol{C}\widehat{\boldsymbol{\theta}}(t) - \boldsymbol{c}(t)\right\},$$

where $\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_n)^{\mathrm{T}}$. The F test statistic is defined as

$$F = \frac{\int SSH_n(t)dt/r}{\int \widehat{\mathcal{R}}(t,t)dt}, \tag{4.13}$$

where $\widehat{\mathcal{R}}(t,t)$ is an estimator of the variance function. For dense functional data, $\widehat{\mathcal{R}}(t,t)$ is expressed as mean square error in Zhang (2013). For sparse functional data, the covariance function need to be estimated using methods described in Section 4.4.

According to Zhang (2013) (page 217),

$$F \sim F_{r\widehat{\kappa},(n-p-q)\widehat{\kappa}} \quad \text{approximately}, \tag{4.14}$$

where $\widehat{\kappa} = \frac{\mathrm{tr}^2(\widehat{\mathcal{R}})}{\mathrm{tr}(\widehat{\mathcal{R}}^{\otimes 2})}$ and $\widehat{\mathcal{R}}$ is an estimator of the covariance function (4.5). When the covariance function yields a spectral decomposition $\widehat{\mathcal{R}}(s,t) = \sum_{k=1}^{\infty} \widehat{\omega}_k \widehat{\psi}_k(s)\widehat{\psi}_k(t)$, where $\{\widehat{\psi}_k(t), k = 1, 2, \ldots\}$ are orthonormal eigenfunctions, then

$$\mathrm{tr}(\widehat{\mathcal{R}}) = \sum_k \widehat{\omega}_k, \quad \mathrm{tr}(\widehat{\mathcal{R}}^{\otimes 2}) = \sum_k \widehat{\omega}_k^2.$$

The approximating distribution (4.14) were developed under dense functional data and was never previously tested on longitudinal or sparse functional data. This approximation also suggest that the F-test does not enjoy the Wilks phenomenon that its null distribution does depend on nuisance parameters such as the eigenvalues of the covariance function.

4.4 Nonparametric Covariance Estimation

In this section, we will focus on the nonparametric covariance models advocated by Yao et al. (2005); Li and Hsing (2010). For the rest of this section, we will focus on the case $g(\cdot)$ is an identical link, let $\epsilon_i(t) = Y_i(t) - \{\boldsymbol{X}_i(t)^{\mathrm{T}}\beta + \mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\theta}(t)\}$ be the error process.

We model $\mathcal{R}(t_1, t_2)$ as a bivariate nonparametric function, which is smooth except for the points on the diagonal line, $\{t_1 = t_2\}$, to allow possible nugget effects. To see this point, we assume that $\epsilon_i(t)$ can be decomposed into two independent components, $\epsilon_i(t) = \epsilon_{i0}(t) + \epsilon_{i1}(t)$, where $\epsilon_{i0}(\cdot)$ is a longitudinal process with smooth covariance function $\mathcal{R}_0(t_1, t_2)$, $\epsilon_{i1}(\cdot)$ is a white noise process usually caused by measurement errors. Let $\sigma_1^2(t) = \text{var}\{\epsilon_{i1}(t)\}$, then

$$\mathcal{R}(t_1, t_2) = \mathcal{R}_0(t_1, t_2) + \sigma_1^2(t_1) I(t_1 = t_2), \tag{4.15}$$

where $I(\cdot)$ is an indicator function. In equation (4.15), $\sigma_1^2(\cdot)$ is the nugget effect causing discontinuity in $\mathcal{R}(\cdot, \cdot)$. We assume that both $\mathcal{R}_0(\cdot, \cdot)$ and $\sigma_1^2(\cdot)$ are smooth functions. As a result, $\mathcal{R}(t_1, t_2)$ is a smooth surface except on the diagonal points where $t_1 = t_2$, and it is also smooth along the diagonal direction. For time series data, without additional assumptions, some confounding will occur if both the mean and covariance functions are modeled nonparametrically. However, this identifiability issue will not occur for longitudinal data, because of the independence structure between subjects.

Let $\widehat{\epsilon}_{ij} = Y_{ij} - \{\boldsymbol{X}_{ij}^{\mathrm{T}}\widehat{\boldsymbol{\beta}} + \mathbf{Z}_i^{\mathrm{T}}\widehat{\boldsymbol{\theta}}(T_{ij})\}$ be the residual of the full model in Section 4.2. The variance function is estimated applying a smoother to the squares of residuals. Let $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\theta}}(t)$ be the full model estimators described in Section 4.2, and define the residuals as $\widehat{\epsilon}_{ij} = Y_{ij} - \boldsymbol{X}_{ij}^{\mathrm{T}}\widehat{\beta} - \mathbf{Z}_i\widehat{\boldsymbol{\theta}}(T_{ij})$. Then $\sigma^2(t)$ can be estimated by a local linear estimator $\widehat{\sigma}^2(t) = \widehat{\alpha}_0$, where $(\widehat{\alpha}_0, \widehat{\alpha}_1)$ minimizes

$$\frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\{\widehat{\epsilon}_{ij}^2 - \alpha_0 - \alpha_1(T_{ij} - t)\}^2 K_h(T_{ij} - t), \tag{4.16}$$

and $K(\cdot)$ and h are the kernel function and bandwidth. Suppose $\mathcal{R}$ has a decomposition as in (4.15); we first estimate the smooth part $\mathcal{R}_0$ using a bivariate local linear

smoother. Let $\widehat{\mathcal{R}}_0(t_1,t_2)=\widehat{\alpha}_0$, where $(\widehat{\alpha}_0,\widehat{\alpha}_1,\widehat{\alpha}_2)$ minimizes

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{j'\neq j}\left\{\widehat{\epsilon}_{ij}\widehat{\epsilon}_{ij'}-\alpha_0-\alpha_1(T_{ij}-t_1)-a_2(T_{ij'}-t_2)\right\}^2 \times K_h(T_{ij}-t_1)K_h(T_{ij'}-t_2). \tag{4.17}$$

Define $N_R=\sum_{i=1}^{n}m_i(m_i-1)$,

$$\begin{aligned} S_{pq} &= \frac{1}{N_R}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{j'\neq j}\left(\frac{T_{ij}-t_1}{h}\right)^p\left(\frac{T_{ij'}-t_2}{h}\right)^q K_h(T_{ij}-t_1)K_h(T_{ij'}-t_2),\\ R_{pq} &= \frac{1}{N_R}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{j'\neq j}\widehat{\epsilon}_{ij}\widehat{\epsilon}_{ik}\left(\frac{T_{ij}-t_1}{h}\right)^p\left(\frac{T_{ij'}-t_2}{h}\right)^q \\ &\qquad \times K_h(T_{ij}-t_1)K_h(T_{ij'}-t_2). \end{aligned}$$

Then the following solution for (4.17) is given in Hall *et al.* (2006):

$$\widehat{\mathcal{R}}_0(s,t)=(\mathcal{A}_1R_{00}-\mathcal{A}_2R_{10}-\mathcal{A}_3R_{01})\mathcal{B}^{-1}, \tag{4.18}$$

where $\mathcal{A}_1=S_{20}S_{02}-S_{11}^2$, $\mathcal{A}_2=S_{10}S_{02}-S_{01}S_{11}$, $\mathcal{A}_3=S_{01}S_{20}-S_{10}S_{11}$, $\mathcal{B}=\mathcal{A}_1S_{00}-\mathcal{A}_2S_{10}-\mathcal{A}_3S_{01}$.

As described above, the diagonal values on $\mathcal{R}(\cdot,\cdot)$ require a special treatment. The variance function can be written as $\sigma^2(t)=\mathcal{R}_0(t,t)+\sigma_1^2(t)$, and be estimated by the local linear smoother in (4.16).

The covariance function is estimated by

$$\widehat{\mathcal{R}}(s,t)=\widehat{\mathcal{R}}_0(s,t)I(s\neq t)+\widehat{\sigma}^2(t)I(s=t). \tag{4.19}$$

Li and Hsing (2010) and Li (2011) proved that nonparametric covariance function estimator in (4.19) is uniformly consistent to the true covariance function

$$\sup_{s,t\in\mathcal{T}}|\widehat{\mathcal{R}}(s,t)-\mathcal{R}(s,t)|=O_p\left[h^2+\{\log n/(nh^2)\}^{1/2}\right].$$

The detailed convergence rate for the nonparametric covariance estimator can be found in Li (2011). However, as noted in previous literature (Hall *et al.*, 1994; Li *et al.*, 2007), the kernel covariance estimator in (4.19) is not guaranteed to be positive semi-definite, and therefore some adjustment is needed to enforce the condition. One possible adjustment is though a spectral decomposition of the covariance estimator.

A commonly used spectral decomposition of the covariance functions for longitudinal data is (Yao *et al.*, 2005; Hall *et al.*, 2006)

$$\mathcal{R}_0(s,t) = \sum_{k=1}^{\infty} \omega_k \psi_k(s)\psi_k(t),$$

where $\omega_1 \geqslant \omega_2 \geqslant \cdots \geqslant 0$ are the eigenvalues of the covariance function, and $\psi_k(t)$ are the corresponding eigenfunctions with $\int_{\mathcal{T}} \psi_k(t)\psi_{k'}(t)dt = I(k = k')$.

An adjustment procedure has been proposed and theoretically justified by Hall *et al.* (2008) to transform $\widehat{\mathcal{R}}_0$ into a valid covariance function. We take a spectral decomposition of $\widehat{\mathcal{R}}_0$ and truncate the negative components. Letting $\widehat{\omega}_k$ and $\widehat{\psi}_k(\cdot)$, $k = 1, 2, \ldots$, be the eigenvalues and eigenfunctions of $\widehat{\mathcal{R}}_0$, and $K_n = \max\{k; \widehat{\omega}_k > 0\}$, then the adjusted estimator for $\mathcal{R}$ is

$$\begin{aligned} \widetilde{\mathcal{R}}_0(s,t) &= \textstyle\sum_{k=1}^{K_n} \widehat{\omega}_k \widehat{\psi}_k(s)\widehat{\psi}_k(t), \\ \widetilde{\mathcal{R}}(s,t) &= \widetilde{\mathcal{R}}_0(s,t)I(s \neq t) + \widehat{\sigma}^2(t)I(s = t). \end{aligned} \tag{4.20}$$

4.5 Analysis of Longitudinal GWAS Data from ADNI

The Alzheimer's Disease Neuroimaging Initiative (ADNI) is an NIH-funded longitudinal observational study, the goal of which is to develop biomarkers to detect and track Alzheimer's Disease (AD). The original ADNI cohort included a total of 800 subjects, many of whom has repeated measurements on AD related biomarkers, such

as the hippocampal volume (HV) and Rey Auditory Verbal Learning Tests (RAVLT), over 10 years of followups. The HV is a Gaussian-type continuous variable, which can be modeled by Model (4.2) with an identity link; in contrast, the RAVLT data are non-Gaussian count data.

Genotype of 620,901 SNPs were measured for the ADNI subjects, and our goal is to identify the SNPs related to AD biomarkers such as HV and RAVLT. As mentioned before, the genotypes for each SNP are AA, AB, and BB, determined by the two alleles of the SNP. After excluding SNPs with large portions of missing values and unevenly distributions, our analysis focuses on 311,417 SNPs with at least 5% subjects in each of the three genotypes. Demographical variables, such as baseline age, sex, years of education, race, and marital status, are collected as covariates.

4.5.1 Analysis of the Hippocampal Volume Data

Among the biomarkers considered in ADNI, there has been some documented evidence that loss of hippocampal volume in human brain may be associated with memory loss and Alzheimer's Disease (Schuff et al., 2009). In the ADNI cohort, 629 subjects have repeatedly measured hippocampal volume using neuroimaging methods during the 10-year follow-up. The measurement times are irregular and random, and the number of repeated measures per subject ranges between 2 and 11 with a median of 4. The distribution of observation time is highly skewed and observations become increasingly sparse after year 6, we therefore take a log-transformation to time and let $t = \log(1+\text{actual visit time})$, which brings the time domain to $\mathcal{T} = [0, 2.4]$. Figure 4.1 shows twenty randomly selected hippocampal volume trajectories in log-transformed time.

Genotype of 620,901 SNPs were measured for the ADNI subjects. Our goal is to identify the SNPs related to hippocampal volume loss by testing hypothesis (4.3) for each SNP. the genotypes for each SNP are AA, AB, and BB, determined by two

Figure 4.1: Twenty randomly selected hippocampal volume trajectories from the ADNI cohort with log-transformed time.

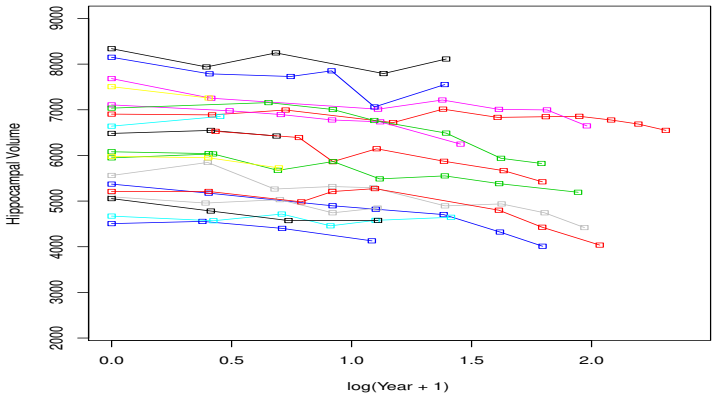

Table 4.1: Summary of the covariates in the ADNI data

Age					
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
55.10	70.90	74.80	74.81	79.70	90.90
Education					
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
6.00	14.00	16.00	15.63	18.00	20.00
Gender		Race		Marital	
Female	Male	White	Non-white	Married	Not married
256 (40.69%)	373 (59.30%)	588 (93.48%)	41 (6.52%)	490 (77.90%)	139 (22.10%)

alleles. After excluding SNPs with large portions of missing values and unevenly distributions, our analysis focuses on 311,417 SNPs with at least 5% subjects in each of the three genotypes among the 629 subjects. Demographical variables including age, gender, years of education, race, and marital status are considered as covariates in Model (4.2). Summary statistics of these covariates are provided in Table 4.1. We repeat the test for each SNP, taking into account of the multiple hypothesis testing issue by a Bonferroni procedure.

To perform SNP level tests for hundreds of thousands of times, we encounter two technical difficulties. First, it is computationally infeasible to run bootstrap for hundreds of thousands of SNPs. Second, the commonly used genome-wide significant levels are 10^{-5} or 10^{-7} (Fadista et al., 2016; Huang et al., 2017), hence it requires a

gigantic bootstrap sample for the bootstrap p-value to reach such accuracy.

Between the two tests discussed in Section 4.3, the GQLR test is more feasible to address these statistical challenges. First, the Wilks phenomenon in Section 4.3.1 implies that the null distribution the GQLR test is the same for all SNPs so that there is no need to repeat bootstrap for hundreds of thousands of SNPs. In contrast, no Wilks phenomenon is established for the F-test. Second, our simulation study also suggest that the GQLR test is more power than the F-test. We therefore perform the GQLR test to all SNPs, but only apply the bootstrap procedure on 20 randomly selected SNPs, with 1,000 bootstrap samples for each SNP. As the model exhibits heteroskedasticity, wild bootstrap is applied Mammen (1993). We follow the wild bootstrap procedure under the sparse functional data setting in Tang et al. (2016) for a single SNP. We then combine these bootstrap test statistics from the 20 SNPs and fit a χ^2 distribution to the combined sample using maximum likelihood estimation, and use the fitted χ^2 distribution to evaluate the p-values for all SNPs. Figure 4.2 shows the empirical distribution of the combined bootstrap sample for the working independent test statistic and its χ^2 approximation. The bandwidth is selected using leave-one-out cross validations on 20 randomly selected SNPs, the average of these selected bandwidths is adjusted by multiplying a factor $n^{-1/45}$ for undersmoothing and then fixed for all SNPs.

The GQLR test detects 3 SNPs associated with HV at 10^{-7} significance level, 52 SNPs at 10^{-5} significance level. The 50 most significant SNPs are reported in Table (4.5.1), where we report the name of the SNP, the corresponding gene and chromosome, and the position of the SNP on chromosome. The most significant SNP is in gene APOE, which has been identified by multiple independent studies to be related to HV and AD (Grupe et al., 2007; Ferencz et al., 2013). The second most significant SNP is located at gene ABLIM2, the association of which with AD was found by Gasparoni et al. (2018). For the third SNP 'rs2800235' on the list is

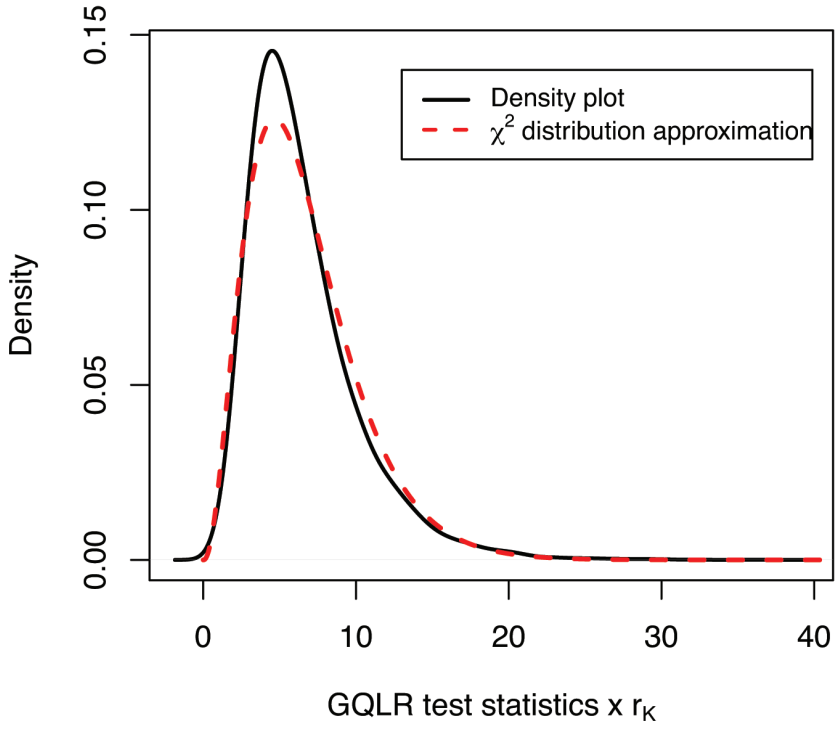

Figure 4.2: The empirical distributions (black sold line) and their χ_2 approximations (red dashed line).

not found in any existing literature, and thus merits further investigation. We also show in Figure 4.3 the functional genotype effects for the top three most significant SNPs. The solid curve in each plot is the overall mean function, while the dashed, dotted and dash-dot curves are the estimated genotype effects for AA, AB and BB, respectively.

4.5.2 Analysis of the RAVLT Data

The RAVLT is a neuropsychological assessment designed to evaluate verbal memory in patients and it can also be used to evaluate the nature and severity of memory dysfunction. During the test, the patient hears a list of 15 nouns (List A) and is asked

Table 4.2: Top 50 SNPs associated with HV, with the names of SNP and corresponding gene, the chromosome, and the position of the SNP on chromosome.

Order	SNP	Gene	Chr	Position	Order	SNP	Gene	Chr	Position
1	rs2075650	APOE	19	44892363	26	rs13420500	C2orf88	2	190070149
2	rs11936149	ABLIM2	4	8120770	27	rs7068990	LOC105378515	10	120329401
3	rs2800235	-	1	224861046	28	rs1673887	-	3	103882935
4	rs1673874	-	3	103868142	29	rs11247613	SLC9A1	1	27149295
5	rs2061345	-	3	103869583	30	rs2516104	-	6	117770467
6	rs17300532	LOC105379028	5	72084529	31	rs1474359	C2orf88	2	190068281
7	rs6044895	DSTN	20	17586934	32	rs4518082	-	3	139685328
8	rs1885082	RRBP1	20	17613340	33	rs4980200	-	10	123698107
9	rs2655997	TMEM63C	14	77204147	34	rs10936959	LINC02015	3	177873287
10	rs10495753	DTNB	2	25452827	35	rs1361417	-	6	102290539
11	rs10439990	ZBTB20	3	114396188	36	rs4920338	PAX7	1	18664300
12	rs4972625	-	2	173988067	37	rs433627	LOC105376126	9	87212499
13	rs10895739	-	11	97410246	38	rs2257468	ABR	17	1083503
14	rs7889761	FRMPD4	X	12520193	39	rs405509	APOE	19	44905580
15	rs4740801	-	9	4790166	40	rs3909086	-	20	6270993
16	rs10931440	C2orf88	2	190018635	41	rs12713521	AFTPH	2	64590123
17	rs10995440	-	10	63108822	42	rs29327	ANK2	4	113286511
18	rs1345516	-	2	64476153	43	rs1890202	MCF2L	13	112900737
19	rs228815	-	6	39139170	44	rs10518258	-	19	29185868
20	rs6136143	DSTN	20	17592113	45	rs4947936	-	7	50839056
21	rs9874829	-	3	139681798	46	rs7233189	EPB41L3	18	5480543
22	rs11633192	THSD4	15	71555974	47	rs6054058	-	20	6281541
23	rs733217	-	3	69671518	48	rs5030938	LOC101928994	10	69216161
24	rs2395891	LOC107985278	19	2032150	49	rs10992211	-	9	90129734
25	rs972795	FHIT	3	59774393	50	rs1062980	IREB2	15	78500186

to recall as many words from the list as possible. After five repetitions of free-recall, a second interference list (List B) is presented, and the participant is asked to recall as many words from List B as possible. The participant is asked to recall the words from List A immediately after the interference trial and after a 30 min delay. The delayed RAVLT score is the number of words that the participant correctly recall from List A after the delay interval, which has been used for identifying patients at

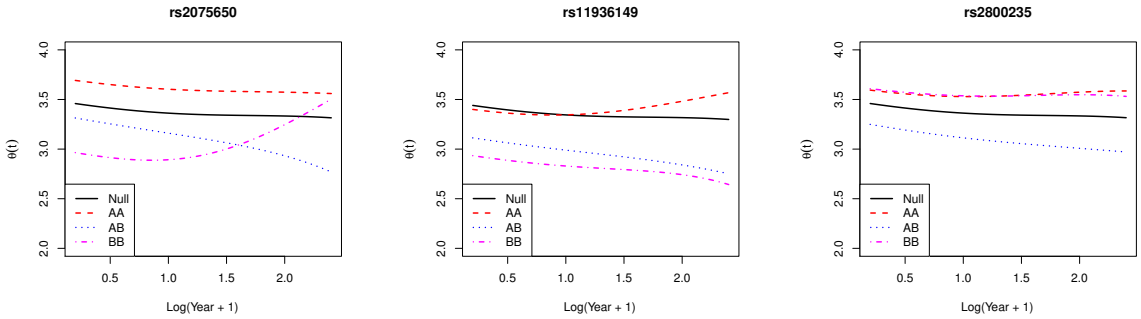

Figure 4.3: Estimated genotype effects for the top three SNPs related to HV in the ADNI data.

high risks of cognitive decline and subsequent dementia (Andersson et al., 2006).

In the ADNI cohort, 358 subjects with mild cognitive impairment (MCI) were administered the RAVLT test at months 0, 6, 12, 18, 24 and 36, but actual measurement times varied randomly around the scheduled dates. A histogram of delay RAVLT scores is provided in Figure 4.4. These scores are count data, skew to the right and obviously non-Gaussian. Our goal is the establish the association between RAVLT test score and the gene APOE.

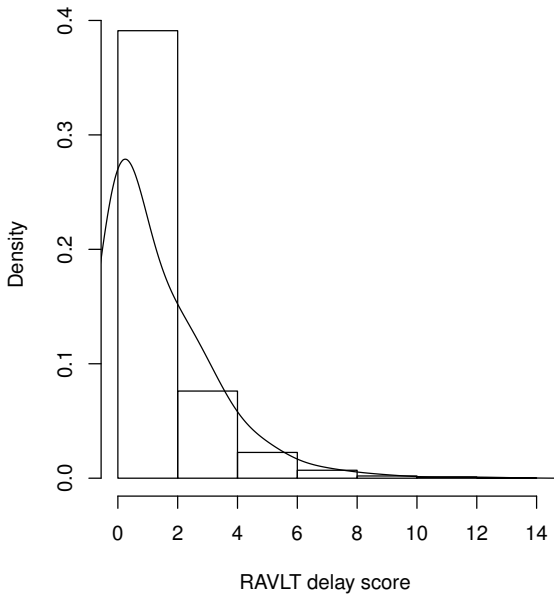

Figure 4.4: The histogram of all observed RAVLT delay scores.

We fit Model (4.2) with log link to RAVLT data, where $\boldsymbol{X}$ includes the baseline age and sex (0 for man and 1 for woman) and $\mathbf{Z}$ is a vector of indicators for APOE alleles numbers (0, 1 or 2). The estimated coefficients for age and sex are $\hat{\beta}_{\texttt{age}} = -0.0024$ and $\hat{\beta}_{\texttt{sex}} = -0.1992$ with standard errors 0.006382 and 0.09997, respectively, indicating a significant effect of sex. To test the significance of the effect of APOE on RAVLT scores, the proposed GQLR test is applied. The obtained p-value for the

null hypothesis: $\theta_1(t) = \theta_2(t) = \theta_3(t)$ is 0.039 by the wild bootstrap procedure with sample size 1000. This result illustrated that APOE is significant with the mean curve of RAVLT curve. The estimated functional effects of APOE are shown in Figure (4.5), where the dark solid curve is $\widehat{\vartheta}(t)$ under the null hypothesis representing the overall mean curve and the other three curves represent the group mean functions for APOE alleles 0, 1 and 2, respectively. It shows that RAVLT scores of MCI patients with APOE allele(s) decline dramatically over time, while scores of those without APOE alleles remain almost the same level.

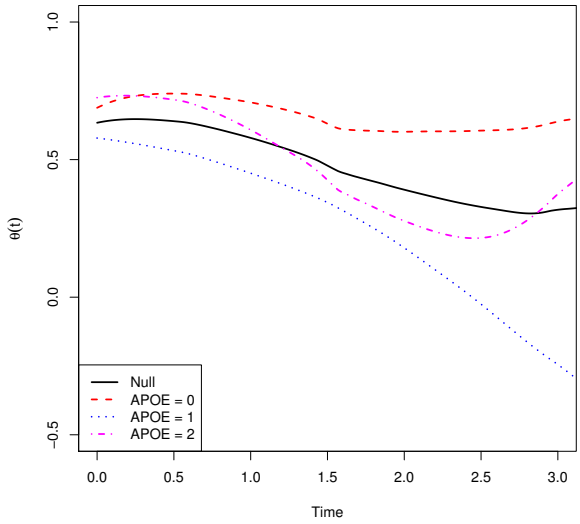

Figure 4.5: The estimated functional effects of APOE on RAVLT scores in the ADNI data.

4.6 Simulation Studies

4.6.1 Gaussian Case

For Gaussian-type sparse functional data, both GQLR test and functional F-test can be used to test hypothesis (4.3), however, their powers have not been previously compared. We now provide such a comparison through simulation studies.

We generate data from model (4.2) with an identity link, $p = 2$ environmental predictors and $q = 4$ genetic predictors. There are $m_i = 5$ repeated measures of Y on each subject, where the observation times are iid with $T_{ij} \sim Unif(0, 1)$. The first environmental predictor is time dependent with $X_{1,ij} = T_{ij} + U_{ij}$, where $U_{ij} \sim Unif(-1, 1)$; and the second environmental predictor $X_{2,i}$ is a binary, time-invariant covariate that equals 0 or 1 with probability 0.5. Suppose the subjects are classified into 4 groups according to genetic traits, and $\mathbf{Z}_i$ is a 4-dimensional vector of indicators for the groups. We simulate a total of $n = 200$ subjects with $n_k = 50$ subjects in each genetic group, $k = 1, \ldots, 4$. The goal is to test if there are any genetic effects, i.e.

$$H_0 : \theta_1(t) = \cdots = \theta_4(t) \equiv \theta_0(t). \tag{4.21}$$

We generate the errors ϵ_{ij} as discrete observations on a zero-mean Gaussian process $\epsilon_i(t)$ and consider two covariance settings: (i) ARMA(1,1) covariance with $\rho(t_1, t_2; \gamma, \varphi) = \gamma \exp(-|t_1 - t_2|/\varphi)$, where $\gamma = 0.75$, $\varphi = 1$, and variance function $\sigma^2(t) = 0.5$ (ii) a nonparametric covariance induced by the mixed model $\epsilon_{ij} = \xi_{0,ij} + \sum_{l=1}^{3} \xi_{li}\phi_l(T_{ij})$, where $\xi_{0,ij}, \xi_{li} \sim N(0, 0.3)$ are independent random effects and $\phi_1(t) = t^2 + 0.5$, $\phi_2(t) = \sin(3\pi t)$, $\phi_3(t) = \cos(3\pi t)$.

We set $\boldsymbol{\beta} = (1, 1)^{\mathrm{T}}$, and $\theta_1(t) = \theta_0(t) - 2\delta S(t)$, $\theta_2(t) = \theta_0(t) - \delta S(t)$, $\theta_3(t) = \theta_0(t) + \delta S(t)$ and $\theta_4(t) = \theta_0(t) + 2\delta S(t)$, where $\theta_0 = \sin(2\pi t)$ and $S(t) = \sin(6\pi t)$. We set $\delta = \{0, 0.05, 0.1, 0.15, 0.2, 0.25\}$, where $\delta = 0$ correspond to the null hypothesis (4.21) and the true model deviates further from H_0 as δ increases. For each value of δ, we simulate 200 datasets and apply both the GQLR test and the F-test to test the null hypothesis (4.21).

Note that the hypothesis (4.21) is an ANOVA hypothesis, under which case the

F test statistic can be written as

$$F = \frac{\int_{\mathcal{T}} \sum_{k=1}^{q} n_k \{\widehat{\theta}_{F,k}(t) - \widehat{\theta}_R(t)\}^2 dt/(q-1)}{\int_{\mathcal{T}} \widehat{\mathcal{R}}(t,t)dt}.$$

There are two ways to estimate null distribution of the F statistic: the asymptotic F distribution given in Section 4.3.2 (F**-asymp**) with the covariance function estimated using the nonparametric method described in Section 4.4 and the wild bootstrap method (F**-boot**). For the GQLR test statistic, we use a Gaussian quasi-likelihood $\mathcal{Q}(\mu, \boldsymbol{Y}) = -(\boldsymbol{Y}-\mu)^{\mathrm{T}}\boldsymbol{V}^{-1}(\boldsymbol{Y}-\mu)/2$, where $\boldsymbol{V}$ is a diagonal variance matrix using the estimated variance function (4.16) interpolated at subject-specific time points. The null distribution of the GQLR test is estimated by bootstrap.

The empirical powers of the three tests as functions of δ are shown in Figure 4.6, where the two panels correspond to the two covariance settings. The F test based on asymptotic theory does not hold the nominal size in our second covariance setting, which is understandable since the asymptotic distribution in Section 4.3.2 was developed under dense functional data. This result shows that the asymptotic F approximation for the F test may not be legitimate under our sparse functional data setting. Both the GQLR and F test hold the nominal size when the null hypothesis is estimated by bootstrap, however the GQLR test is more powerful under both simulation settings.

4.6.2 Non-Gaussian Response

To demonstrate the use of the methods described, we also simulate data from model (4.2) under a logarithm link. The covariates $\boldsymbol{X}$ and $\mathbf{Z}$ are simulated in the same way as in Section 4.6.1. Suppose there are $m_i = 4$ repeated measures on each subject with observation times uniformly distributed in $[0, 1]$. Conditional on $\boldsymbol{X}_{ij}$ and $\mathbf{Z}_i$, Y_{ij} are generated from Poisson distribution with mean $\mu_{ij} = \exp\{\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\theta}(T_{ij})\}$ and exchangeable correlation within the same subject.

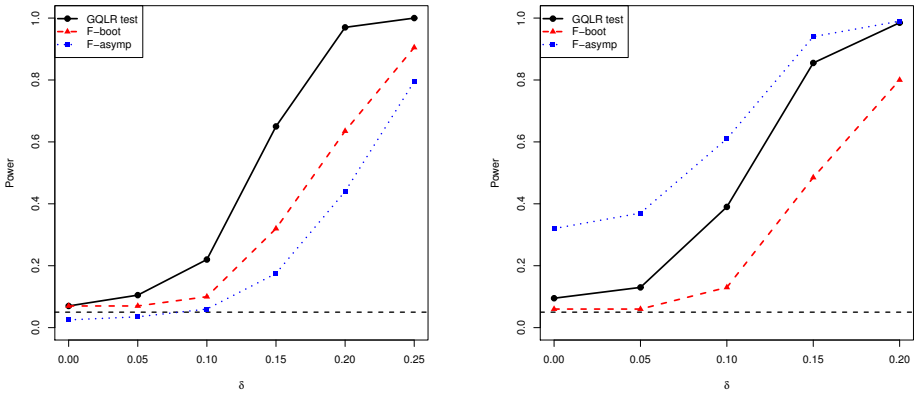

Figure 4.6: Empirical power of three tests. The horizontal dotted line is set at 0.05. The left panel is the result under covariance setting (i) where the true covariance is ARMA(1,1); the right panel is the result under covariance setting (ii) where the errors are generated from a mixed model with nonparametric factors.

The correlated multivariate Poisson random variables are simulated by the method of Yahav and Shmueli (2012), using auxiliary normal distributions. To be more specific, we generate standard normal distribution Y'_{ij} with exchangeable within-cluster correlation such that $corr(Y'_{ij}, Y'_{ij'}) = \rho_{jj'} = 0.3$ for $j \neq j'$, and generate $Y_{ij} = F_{ij}^{-1}\{\Phi(Y'_{ij})\}$ where $\Phi(\cdot)$ is the distribution function of standard normal and $F_{ij}(\cdot)$ is the distribution function of a poisson distribution with mean μ_{ij}.

We test the same ANOVA hypothesis in (4.21). Note that the F-test was not developed for non-Gaussian response, we therefore only consider the GQLR test using a Poisson quasi-likelihood $\mathcal{Q}(\mu_i, \boldsymbol{Y}_i) = \boldsymbol{Y}_i^{\mathrm{T}} \log(\mu_i) - \mu_i^{\mathrm{T}}\mathbf{1}$.

To demonstrate the Wilks phenomenon of the GQLR test, we consider the following three scenarios with different setting of $\boldsymbol{\beta}$ and θ_0:

$$\text{Scenario I}: \beta_1 = 1,\ \beta_2 = 1,\ \theta_0(t) = \sin(2\pi t);$$

$$\text{Scenario II}: \beta_1 = 0.5,\ \beta_2 = 1.5,\ \theta_0(t) = \sin(2\pi t);$$

$$\text{Scenario III}: \beta_1 = 1,\ \beta_2 = 1.5,\ \theta_0(t) = \cos(2\pi t).$$

We simulate 200 datasets for each scenario and apply the GQLR test to each simu-

lated data set. Figure 4.7 shows the estimated densities for $\lambda_n(H_0)$ under the three scenarios using kernel smoothing. We perform a k-sample Anderson-Darling test and find no significant difference among the three distributions (p-value: 0.53). These results show Wilks phenomenon holds for the GQLR test under non-Gaussian case that the distribution of $\lambda_n(H_0)$ does not depend on the true value of the parameters.

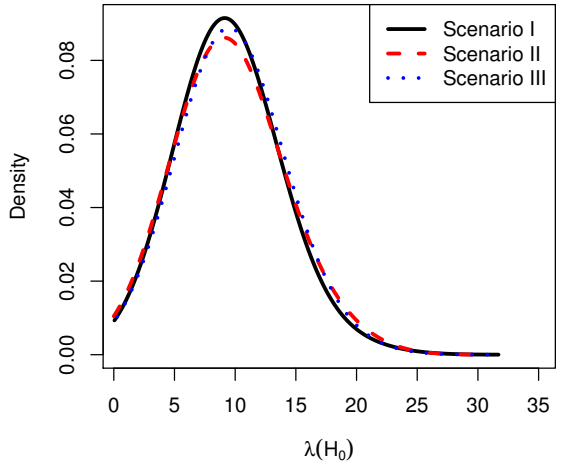

Figure 4.7: Simulation under non-Gaussian case: demonstration of the Wilks phenomenon. The three curves are the estimated distribution of $\lambda_n(H_0)$ under the three simulation scenarios.

Next, we study the power of the GQLR test. We focus on the simulation setting described in Scenario I and consider local alternatives with $\theta_1(t) = \theta_0(t) - 2\delta G(t)$, $\theta_2(t) = \theta_0(t) - \delta G(t)$, $\theta_3(t) = \theta_0(t) + \delta G(t)$, and $\theta_4(t) = \theta_0(t) + 2\delta G(t)$. We set $G(t) = \sin(4\pi t)$ and consider different δ values. The null hypothesis is true when $\delta = 0$ and as δ increases the model deviates further away from H_0. Specifically, we set the significance level at $\alpha = 0.05$ and use the wild bootstrap procedure to estimate the null distribution. Figure 4.8 shows the power of the GQLR test as a function of δ. As we can see, the test holds the nominal size and the power increases to 1 as δ increases.

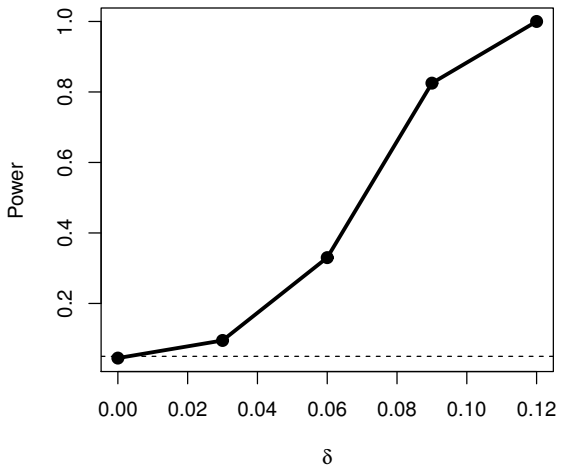

Figure 4.8: Simulation under non-Gaussian case: power of the GQLR test

4.7 Summary

In this chapter, we demonstrate the use of functional data modeling and inference methods to analyze longitudinal GWAS data, where aging disease related phenotypes are repeatedly measured over time. The method can be used to analyze both Gaussian-type (such as the HV data in ADNI) and non-Gaussian (such as the RAVLT scores) response, taking into account the parametric effects of environmental covariates and functional effects of genotypes. In testing the functional genotype effects, we compare the effectiveness of two widely used nonparametric tests and show advantages the GQLR test over the functional F-test when analyzing sparse functional data from longitudinal GWAS. First, the GQLR test can be used for both Gaussian and non-Gaussian responses, but the F-test was only developed for Gaussian response. Second, the GQLR test enjoys the Wilks property making it feasible for large scale multiple SNP-level hypotheses testing, but the F-test does not enjoy such property. Third, the GQLR test is shown to enjoy the minimax optimal power, while

the local power of the F-test is largely unknown. Our simulation studies suggest the GQLR test has higher power than the F-test, where there is inhomogeneity and correlation in the data.

Chapter 5

Improved Power for Multiple Testing of Genetic Association with Longitudinal Phenotypes for Large-Scale ADNI GWAS

5.1 Introduction

In the previous Chapter 4, for the hypotheses (4.1), GQLR test is more powerful than F-test but still limited detecting only three SNPs at the genome-wide significance level ($P < 10^{-7}$). This lies in the reason that the GQLR test for sparse longitudinal response in Chapter 4 was developed based on the working independence (WI) estimation of Lin and Carroll (2001), which ignored the correlation structure entirely. Noticed that Wang et al. (2005) modified WI of Lin and Carroll (2001) by incorporating within-subject correlation by imputing a correlation function into components of mean response vector. This motivates the seemingly unrelated functional analysis of covariance (SU-fANCOVA) test procedure in Zhu et al. (2020+). The bootstrap version of SU-fANCOVA can be a good solution for the hypothesis problem (4.1) for the longitudinal phenotypic ADNI 1 GWAS. The real data analysis on ADNI 1 GWAS again shows that the SU-fANCOVA is much powerful at GWAS significance level to detect out over 100 significant SNPs that may be related to AD.

The rest of this chapter is organized as follows. We first analyze the ADNI data for both Gaussian and non-Gaussian responses and compare the results with those in Chapter 4. Then we assess the performance of SU-fANCOVA test for non-Gaussian response by simulation studies.

5.2 Covariance Estimation for Non-Gaussian Response

In the previous chapter, one needs not to estimate the entire covariance function of the sparse functional response since the working independent covariance is used. However, in this chapter, the SU-fANCOVA test procedure employs working correlation covariance, and hence the entire covariance function should be estimated. When the longitudinal response is non-Gaussian, the nonparametric method described in Section 4.4 cannot be applied. We recommend the method by Lin (2007) for the covariance estimation for non-Gaussian response in this section.

For non-Gaussian longitudinal data, the variance-covariance structure usually depends on the conditional mean response. Therefore, to estimate the whole covariance function, we would prefer the semiparametric covariance models that respect the mean-covariance dependency relationship. For ADNI Rey Auditory Verbal Learning Test data, the longitudinal responses are count variables, and can be modeled by a Poisson Mixed Model (Lin, 2007). A Poisson functional analysis of variance model with a subject-specific random effect is $Y_{k,ij} \sim \text{Poisson}(\mu^u_{k,ij})$ and $\mu^u_{k,ij} = \exp\{\boldsymbol{X}^{\mathrm{T}}_{k,ij}\boldsymbol{\beta} + \theta_k(T_{k,ij}) + U_{k,i}\}$, where $U_{k,i} \sim$ independent $\text{Normal}(0, \sigma^2_u)$ are random effects. Integrating out the random effects, the marginal mean and within-subject variance-covariance structure are

$$\mu_{k,ij} = \exp\{\boldsymbol{X}^{\mathrm{T}}_{k,ij}\boldsymbol{\beta} + \theta_k(T_{k,ij}) + \sigma^2_u/2\},$$

$$\text{var}(Y_{k,ij}) = \mu_{k,ij} + \mu^2_{k,ij}\{\exp(\sigma^2_u) - 1\},$$

$$\text{cov}(Y_{k,ij_1}, Y_{k,ij_2}) = \mu_{k,ij_1}\mu_{k,ij_2}\{\exp(\sigma_u^2) - 1\}, \quad \text{for } j_1 \neq j_2. \tag{5.1}$$

Notice that the random effect only creates a shift $\sigma_u^2/2$ to the functional treatment effects $\theta_k(t)$ and hence will not affect the test results. Lin (2007) proposed to estimate the covariance parameter $\gamma = \sigma_u^2$ by maximizing a Gaussian quasi-likelihood

$$\widehat{\boldsymbol{\gamma}} = \text{argmax}_{\boldsymbol{\gamma}} - \tfrac{1}{2}\sum_{k=1}^{q}\sum_{i=1}^{n_k} \log|\boldsymbol{V}_{k,i}(\gamma)| + (\boldsymbol{Y}_{k,i} - \mu_{k,i})^{\mathrm{T}}\boldsymbol{V}_{k,i}^{-1}(\gamma)(\boldsymbol{Y}_{k,i} - \mu_{k,i}), \tag{5.2}$$

with $\mu_{k,i}$ substituted by working independent pilot estimators and $\boldsymbol{V}_{k,i}(\gamma)$ reconstructed from (5.1). For Gaussian type functional analysis of variance models, we assume covariance functions are equal across treatment groups. This assumption can be replaced by the assumption that the correlation parameter is equal across groups, and all of the theoretical results in Zhu et al. (2020+) still hold.

5.3 Analysis of Longitudinal GWAS Data from ADNI

We apply seemingly unrelated functional analysis of variance test to the same ADNI data in Chapter 4.

5.3.1 Analysis of the Hippocampal Volume Data

We first apply the working independent functional analysis of variance test to screen for the important SNPs. The bandwidth is selected using cross validations on 20 randomly selected SNPs, the average of these selected bandwidths is adjusted by multiplying a factor $n^{-1/45}$ for undersmoothing and then fixed for all SNPs. Following the bootstrap procedure by Zhu et al. (2020+), we perform wild bootstrap on 20 randomly selected SNPs, with 1,000 bootstrap samples for each SNP, and fit a χ^2 distribution to the combined bootstrap sample using maximum likelihood estimation. The left panel of Figure 5.1 shows the empirical distribution of the combined

bootstrap sample for the working independent test statistic and its χ^2 approximation. We then use the fitted χ^2 distribution to evaluate the p-values for all SNPs. At the 10^{-7} significance level, the working independent test detects 3 SNPs associated with hippocampal volume.

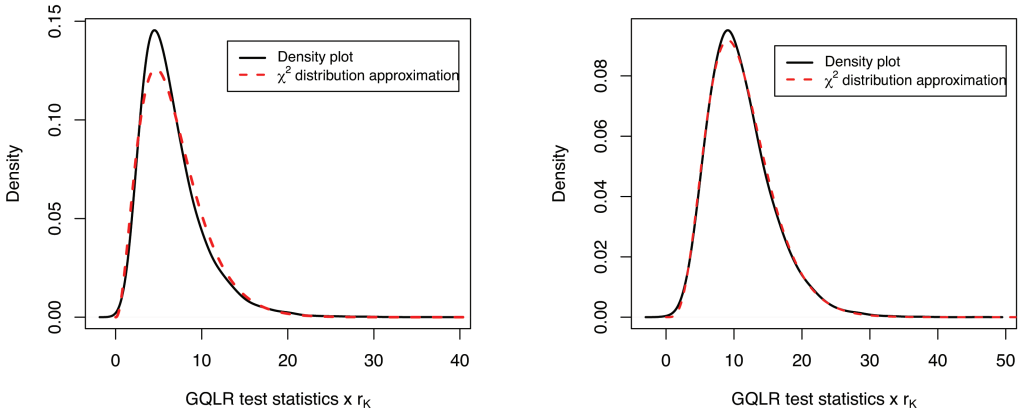

Figure 5.1: The empirical distributions (black sold line) and their χ_2 approximations (red dashed line) by the working independent method (the left panel) and nonparametric method (the right panel).

Next, we apply the seemingly unrelated functional analysis of variance test to the top 2000 SNPs screened by the working independent test. We adopt the same bandwidth for the mean estimation as the working independent procedure, estimate the covariance function separately for each SNP using the nonparametric procedure described in Section 4.4, where the bandwidth for covariance estimation is chosen by leave-one-out cross-validation in 20 randomly selected SNPs. To estimate the null distribution, we run wild bootstrap on 20 randomly selected SNPs; the empirical distributions of $r_K\lambda_n^*(H_0)$ from the combined bootstrap sample and its χ^2 approximation are shown in the right panel of Figure 5.1. The closeness of the two distributions corroborates of the results in Corollary 1 in Zhu et al. (2020+). At significance level 10^{-7}, the proposed test detects 177 SNPs that are associated with

Table 5.1: Top 50 SNPs associated with HV.

SNP	Chr	Position	Gene	SNP	Chr	Position	Gene
rs2075650	19	50087459	TOMM40	rs2722385	7	24393080	LOC107986777
rs3817959	1	14280602	KAZN	rs7922793	10	62496958	
rs1890202	13	112603051	MCF2L	rs11223157	11	131999715	OPCML
rs4646737	3	127328950	ALDH1L1	rs10995440	10	64538587	
rs1938590	11	58668028	FAM111A	rs815845	9	83405910	TLE1
rs17033413	2	45393229		rs2865297	2	57111147	
rs4649222	1	231558060	MAP3K21	rs3888289	11	71017315	
rs4356778	3	28646361	LINC00693	rs447479	21	14258290	ANKRD20A11P
rs1439930	2	224088364		rs12045968	1	33463285	
rs11936149	4	8173396	ABLIM2	rs1429310	2	57126215	
rs1147917	10	42398979	ZNF33B	rs11851025	14	50873678	LINC00640
rs2054365	6	107184195	RTN4IP1	rs2473113	10	42504435	LINC01518
rs11589265	1	27468449	WDTC1	rs1032669	3	1158679	CNTN6
rs2705594	2	217564387	LOC101928278	rs7594454	2	237202674	
rs7939969	11	58620351	LOC105369315	rs17625895	16	25682603	HS3ST4
rs994883	12	17243041	VWF	rs2605877	8	74309320	
rs10069076	5	93815065	KIAA0825	rs982003	10	18747302	CACNB2
rs7530701	1	27448495	WDTC1	rs9295895	6	30546205	
rs10510816	3	59630570	LOC105377110	rs916775	7	24399278	LOC107986777
rs10869183	9	74325645	TMC1	rs10928003	1	14290773	KAZN
rs4849996	2	3880801		rs2940556	5	8402784	LINC02226
rs2940554	5	8395988	LINC02226	rs6491729	13	102492286	
rs965921	4	150276699		rs9407390	9	7805985	
rs2257468	17	933492	ABR	rs4787760	16	25678156	HS3ST4
rs9534812	13	47235539		rs5906966	X	44178162	

hippocampal volume. These SNPs deserve further investigation using independent studies. We summarize the top 50 SNPs detected by the proposed test in Table 5.1. The SNPs are ranked by their significance level. We provide the names of the SNPs, the chromosomes they are on, and the gene names for SNPs located in known genes.

The most significant SNP is rs2075650 located in gene APOE and some other top genes include MCF2L, OPCML, TLE1, FAM111A, and ALDH1L1, all of which have been identified by multiple independent studies to be related to hippocampal volume and Alzheimer's Disease. References of these genes are listed in the Supplementary Material. On the other hand, the proposed method also finds some new genes, such as LOCI107986777 and KAZN, which we could not find in the existing literature and merit further investigation. Figure 5.2 shows the estimated functional genotype effects for the top three SNPs, rs2075650, rs2722385, and rs3817959, located in genes APOE, LOCI107986777, and KAZN, respectively. In each panel of Figure 5.2, the solid curve is the overall mean function, while the dashed, dotted and dash-dot curves

are the estimated mean functions for different genotypes.

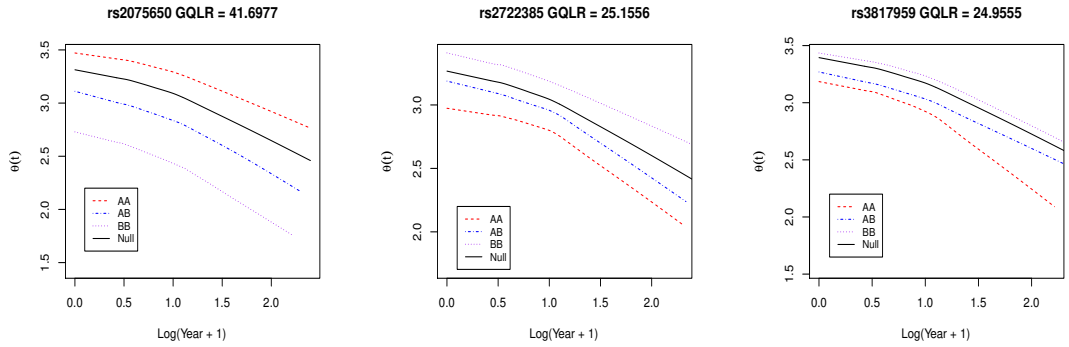

Figure 5.2: Estimated genotype effects for the top three SNPs in the ADNI data.

5.3.2 Analysis of the Rey Auditory Verbal Learning Test Data

Different from that in Section 4.5.2 we here apply the SU-fANCOVA test to the whole subjects in ADNI cohort rather than the MCI patients.

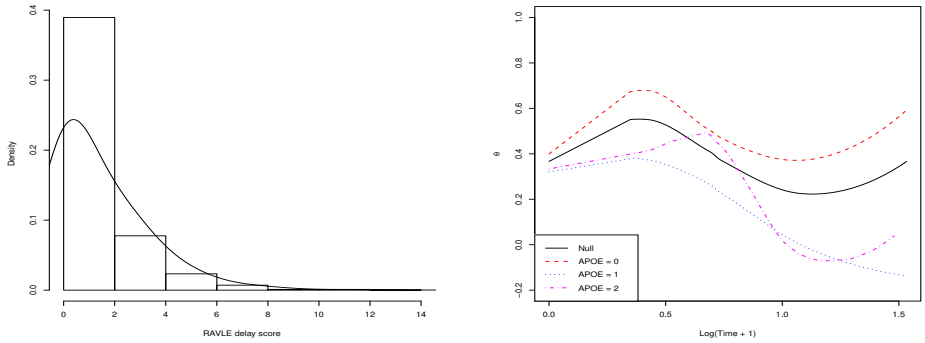

Figure 5.3: ADNI Rey Auditory Verbal Learning Test data: The left panel is the histogram of all observed scores; the right panel contains estimated functional effects of APOE on the scores.

In the ADNI cohort, 721 subjects were administered the Rey Auditory Verbal Learning Test at months 0, 6, 12, 18, 24, and 36, but actual measurement times varied randomly around the scheduled dates. A histogram of the test scores is provided in

the left panel of Figure 5.3. These scores are count data, skew to the right and obviously non-Gaussian. We fit the following model

$$\log\{\mu_{k,i}(t)\} = \boldsymbol{X}_{k,i}^{\mathrm{T}}(t)\boldsymbol{\beta} + \theta_k(t), \tag{5.3}$$

to the data using a logarithm link, where the covariates include baseline age and sex (0 for man and 1 for woman) and the treatment groups are defined by the alleles of APOE. The within-subject correlation is modeled by the Poisson Mixed Model described in Section 5.2, with the correlation parameter estimated by the quasi maximum likelihood method. The estimated coefficients for age and sex are $\widehat{\beta}_{\text{age}} = -0.0399$ and $\widehat{\beta}_{\text{sex}} = -0.0522$ with standard errors 0.025 and 0.071, respectively. The estimated functional effects of APOE are shown in the right panel of Figure 5.3, where the dark solid curve is $\widehat{\theta}(t)$ under the null hypothesis representing the overall mean curve and the other three curves represent the group mean functions for APOE allele numbers 0, 1 and 2, respectively. By wild bootstrap with sample size 1000, the p-value for hypothesis (4.1) is 0.005, which suggests a significant relationship between APOE and the test scores.

5.4 Simulation Studies

To demonstrate the proposed methods under the non-Gaussian response, we also simulate data from the model (5.3). We simulate data for $q = 4$ treatment groups with $n_k = 50$ subjects in each group, and $m_i = 4$ repeated measures on each subject. The responses$Y_{k,ij}$ are generated from Poisson distribution with mean $\mu_{k,ij} = \exp\{\boldsymbol{X}_{k,ij}^{\mathrm{T}}\boldsymbol{\beta} + \theta_k(T_{k,ij})\}$ and an exchangeable correlation structure, where $T_{k,ij} \sim Unif(0,1)$, $X_{1,k,ij} = T_{k,ij} + U_{k,ij}$ is a time varying covariate with $U_{k,ij} \sim Unif(-1,1)$, and $X_{2,k,i}$ is a binary, time-invariant covariate that equals 0 or 1 with probability 0.5. The correlated multivariate Poisson random variables are simulated by the method of Yahav and Shmueli (2012), using auxiliary normal distributions. To

be more specific, we generate standard normal distribution $Z_{k,ij}$ with exchangeable within-subject correlation such that $\text{corr}(Z_{k,ij}, Z_{k,ij'}) = 0.3$ for $j \neq j'$, and generate $Y_{k,ij} = F_{k,ij}^{-1}\{\Phi(Z_{k,ij})\}$ where $\Phi(\cdot)$ is the distribution function of standard normal and $F_{k,ij}(\cdot)$ is the Poisson distribution with mean μ_{ij}.

We set $\beta_1 = \beta_2 = 1$ and $\theta_k(t)$ to be $\theta_0(t) \pm \delta S(t)$ and $\theta_0(t) \pm 2\delta S(t)$, where $\theta_0(t) = 1.5 + \sin(2\pi t)$, $S(t) = \sin(4\pi t)$ and $\delta = \{0, 0.01, 0.02, 0.03, 0.04, 0.05\}$. Since the F-test was not developed for non-Gaussian response, we consider two versions of the proposed GQLR test to test the null hypothesis in (4.1) both based on a Gaussian quasi-likelihood as in (5.2), one under working independence and the other under the compound symmetry correlation with the correlation parameter estimated using the QMLE method. We set the significance level at $\alpha = 0.05$ and use the wild bootstrap procedure with sample size 1000 to estimate the null distribution. Figure 5.4 shows the power of the GQLR test as a function of δ. As we can see, both tests hold the nominal size and the test take into account the correlation is far more powerful than the WI test.

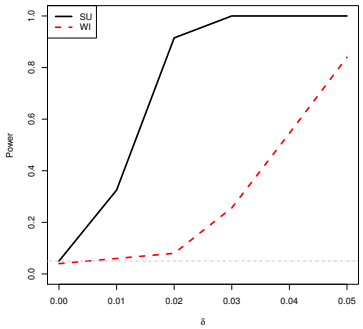

Figure 5.4: Simulation under non-Gaussian case: power of the GQLR test

Acknowledgments for Part II

Data collection and sharing for this part was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Part III

Chapter 6

Weighted Multiple-Quantile Classifiers for Functional Data

6.1 Introduction

Classification for functional data is a challenging and yet appealing research field that has a rich literature and steady study interest over the past two decades because functional data are increasingly encountered in medical studies (Delaigle and Hall (2012), Dai et al. (2017), Berrendero et al. (2018), among others). In this chapter, we are interested in constructing a unified framework by minimizing the expected loss to incorporate existing projection classifiers for functional data. It sheds light on the proposed quantile based nonparametric functional classifier.

There are two main streams of dimension reduction in classification of functional data, by transforming the functional data into multivariate vectors and applying discriminant analysis (Park and Simpson (2019)), or by projecting random trajectories into random variables (Kraus and Stefanucci (2019). The projection scores are an excellent approximation to the original random process, and thus play an important role in difference methods of classification (Delaigle and Hall (2012) and Dai et al. (2017)). Thus it is natural to measure the deviation between projection scores instead of random trajectories.

To construct a suitable distance measure is a popular spirit for clustering and classification for various data sets. The L_2-norm distance is a usual choice between random curves or its derivatives but not applicable to projection scores (Alonso et al. (2012)). We are motivated by a probability operator to define a generalized distance based on probabilistic components of projection scores. It is intuitive to interpret the underlying mechanisms of existing projection classifiers. For instance, the strong assumptions of at least one distinct for mean and variance in Bayes classifier (Dai et al. (2017)) are necessary; the centroid classifier (Delaigle and Hall (2012) is invalid if one does not assume identical variances of the projection scores. We also involve weights into the proposed method which can be obtained accurately hitting upon the important projection scores for classification. Theoretically, it helps settle down the consistency property of the presented classifier.

Particularly we take the quantile rather than moments of projection scores into consideration to develop a weighted multiple-quantile classifier (weMulQ). The quantile technique is a guarantee of robustness against outliers or skewness. For heavy-tailed or non-exponential family distributions, the weMulQ classifier outperforms the existing methods, particularly for mixture distributions. This weakens the assumptions for Bayes classifier based on density ratio of distributions of projection scores. The weMulQ classifier is accurate even when the means of scores are approximately equal in that multiple quantile locations make the classifier take account of almost the full vision of the distribution shape of a projection score. Simultaneously it simplifies the implementation procedure and gains computing expedience. The weMulQ is compatible with other classifier competitors for Gaussian-process scenarios.

In literature, other studies of classification of functional data include classification of sparsely or irregular sampled functional data (Park and Simpson (2019)) and incompletely observed functional curve or fragments (Delaigle and Hall (2013) and Kraus and Stefanucci (2019)), among others. The weMulQ classifier works for dense

functional samples.

6.2 Generalized Distance Minimizing the Risk

We focus on binary classification. Let $X = X(t)$ be a square integrable random function defined on a compact interval $\mathcal{I}$ and Y be a Bernoulli random variable indicating a group label. The observed random trajectories of X come from a mixture of sub-populations Π_0 and Π_1. That is, $X \overset{d}{=} X^{[k]}$ if X is from population k, where $k = 0, 1$ and $\overset{d}{=}$ means equivalence in distribution. Projection scores of an arbitrary $X_0(\cdot) \in \mathcal{L}^2(\mathcal{I})$ and the stochastic process $X^{[k]}(\cdot)$ are

$$x_j^0 = \int_{\mathcal{I}} X_0(s)\psi_j(s)ds, \quad \xi_j^{[k]} = \int_{\mathcal{I}} X^{[k]}(s)\psi_j(s)ds, \quad j = 1, 2, \ldots, \quad k = 0, 1,$$

respectively, where $\psi_j(s)$'s are the orthonomal basises of $\mathcal{L}^2(\mathcal{I})$.

Given a new infinite dimensional random trajectory $X_0 \in \mathcal{L}^2(\mathcal{I})$, one needs to identify whether X_0 comes from Π_0 or Π_1. This can be discriminated by measuring the distance of X_0 away from the specific functional group. Instead, we inspect the distance of their pertinent projection scores following the optimal decision rule of minimizing the risk or expected loss. Let $L(\theta, a)$ be the loss function for the true state θ and an action a, and $\mathbb{E}\{L(\theta, d(\mathbf{X}))\}$ be the expected loss with a decision rule d based on random functional data $\mathbf{X}$ (chapter 2, Young et al. (2005)). Denote the expectation operator of a random variable Z by

$$\mathcal{E}(Z) := \arg\min_{q \in \mathbb{R}} \mathbb{E}\{L(Z, q)\}.$$

This comes up with a generalized distance between X_0 and population Π_k as

$$D(X_0, \Pi_k) := \sum_{j=1}^{\infty} w_j L\{x_j^0, \mathcal{E}(\xi_j^{[k]})\}, \tag{6.1}$$

with the weights w_j's measuring the effect of deviation between projection scores of X_0 and $X^{[k]}$. Let $I(\cdot)$ be the indicator function and $q_j^{[k]}(\tau) = \inf\{u : F_{\xi_j^{[k]}}(u) \geqslant \tau\}$ for $\tau \in \mathcal{S} = [a_0, 1 - a_0]$ with $0 < a_0 < 0.5$. Taking the asymmetric absolute function

$$\rho_\tau(s) = \tau s - sI(s < 0) = |s|[\tau I(s > 0) + (1 - \tau)I(s \leqslant 0)]$$

as the loss function, we have the optimal solution

$$q_j^{[k]}(\tau) \equiv \mathcal{E}(\xi_j^{[k]}) = \arg\min_{q \in \mathbb{R}} \mathbb{E}\{\rho_\tau(\xi_j^{[k]} - q)\}, \tag{6.2}$$

when the τ-quantile of the projection score $\xi_j^{[k]}$ acts as the minimizer. This motivates the following weighted quantile classification criterion based on the first J projection scores $\{x_j\}_{j=1}^J$ of $X = x$,

$$Q_J(x, \tau, \mathbf{w}) = \sum_{j=1}^{J} w_j[L_{1j}(x, \tau) - L_{0j}(x, \tau)], \tag{6.3}$$

where $\mathbf{w} = (w_1, \ldots, w_J)$, and $L_{kj}(x, \tau) := L[x_j, \mathcal{E}(\xi_j^{[k]})] = \rho_\tau\{x_j - q_j^{[k]}(\tau)\}$ can be viewed as an $\mathcal{L}_1$-distance between x_j and the τ-quantile $q_j^{[k]}(\tau)$, $k = 0, 1$. The weight w_j is imposed on the jth component of the quantile-based classifiers, reflecting the relative importance of the jth projection score for classification.

Consequently, for a fixed quantile level $\tau \in \mathcal{S}$, given two sets of observations from populations Π_0 and Π_1 along with a new observation $x \in \mathcal{L}^2(\mathcal{I})$, x is assigned to Π_0 if $Q_J(x, \tau, \mathbf{w}) > 0$, and to Π_1 otherwise. Let $\pi_k = P(Y = k)$ be the prior probability that an observation belongs to population Π_k, $k = 0, 1$. In order to choose the optimal values of J, τ and the weight $\mathbf{w}$, we consider the probability of correct classification based on (6.3),

$$\Psi(J, \tau, \mathbf{w}) = \pi_0 P[Q_J(X, \tau, \mathbf{w}) > 0 | Y = 0] + \pi_1 P[Q_J(X, \tau, \mathbf{w}) \leqslant 0 | Y = 1]. \tag{6.4}$$

Equation (6.4) represents the theoretical rate of correct classification based on the first J true projection scores and the true quantiles. The theoretical optimal values of J, τ and the weight $\mathbf{w}$ can be obtained by maximizing the probability of correct classification (6.4).

Notice that, the generalized distance framework (6.1) provides a unified framework to incorporate existing projection functional classifiers by assigning distinct loss functions. For example, assuming that projection scores $\xi_j^{[k]}$'s are independent and identically normally distributed with mean $\mu_j^{[k]}$ and variance $\lambda_j^{[k]}$, the Bayesian classifier by Dai et al. (2017) can be obtained by taking the square error loss function $L[x_j, \mathcal{E}(\xi_j^{[k]})] = (x_j - \mu_j^{[k]})^2$ and the weight $w_j = (2\lambda_j^{[k]})^{-1}$, $j = 1, \ldots, J$. Likewise, taking the unity weight, the centroid classifier by Delaigle and Hall (2012) can be derived by taking square error loss function $L(\theta, a) = (\theta - a)^2$ since the corresponding expectation operator $E(Z) = \arg\min_{z_0 \in \mathbb{R}} \int_{-\infty}^{\infty} (z - z_0)^2 dF_Z(z)$ induces the distance measurement $D^c(x, X^{(k)}) = \sum_{j=1}^{\infty} w_j \{x_j - E(\xi_j)\}^2$. The classifiers for high-dimensional data can also be interpreted by taking the absolute loss $L(\theta, a) = |\theta - a|$ and the median operator $\text{median}(Z) = \arg\min_{z_0 \in \mathbb{R}} \int_{-\infty}^{\infty} |z - z_0| dF_Z(z)$, leading to the companion distance $D^m(x, X^{(k)}) = \sum_{j=1}^{\infty} w_j |x_j - \text{median}(\xi_j)|$. See Hall et al. (2009) and its extension Henning and Viroli (2016). Such a perspective based on (6.1) gives an intuitive interpretation of why the quantile-based classifier is notable more informative because it is based on 'location' (quantile) of distribution rather than some kind of i-th order moments.

6.3 Implementation of Functional Multiple-Quantile Classification

In this section, we first estimate the projection scores of functional predictors, and then obtain the estimate of the proposed weighted multiple-quantile classifier. We provide an computation procedure for implementation afterwards.

Denote $n = n_0 + n_1$, where n_k is the number of functional predictors from population k, $k = 0, 1$. In practice, the entire predictor trajectories $\{X_i\}_{i=1}^n$ are not observable. Instead, the observations of the ith subject are measured at m_i time points $\{T_{il}\}_{l=1}^{m_i}$ and contaminated with measurement errors, i.e., $\tilde{X}_{il} = X_i(T_{il}) + e_{il}$, where e_{il}'s are identically and independently distributed and independent of predictor trajectories X_i's. For the observed $\{(T_{il}, \tilde{X}_{il}, Y_i)\}_{l=1}^{m_i}$, we may get the smoothed estimate of X_i by the local linear smoothing technique (Fan and Gijbels (1996)),

$$(\hat{a}_{0i}, \hat{a}_{1i}) = \underset{(a_{0i}, a_{1i})}{\arg\min} \sum_{l=1}^{m_i} \left\{ \tilde{X}_{il} - a_{0i} - a_{1i}(T_{il} - t) \right\}^2 K\left(\frac{T_{il} - t}{h_i} \right), \ i = 1, \ldots, n, \quad (6.5)$$

leading to the local linear estimator $\hat{X}_i(t) = \hat{a}_{0i}$, where $K(\cdot)$ is a symmetric kernel function and h_i is bandwidth. The smoothed trajectories can be then regarded as a fully observed random curves.

Denote the covariance function of population k by $G^{[k]}(s,t) = \mathrm{Cov}[X^{[k]}(s), X^{[k]}(t)]$, and assume $G^{[k]}(s,t)$ is continuous. It follows from Mercer's theorem that $G^{[k]}(s,t) = \sum_{j=1}^{\infty} \lambda_j^{[k]} \psi_j^{[k]}(s) \psi_j^{[k]}(t)$, where the orthonormal eigendecomposition yields eigenfunctions $\psi_j^{[k]}(\cdot)$'s and eigenvalues $\lambda_1^{[k]} \geqslant \lambda_2^{[k]} \geqslant \cdots \geqslant 0$ satisfying $\sum_{j=1}^{\infty} \lambda_j^{[k]} < \infty$ for $k = 0, 1$. Applying the weighted local linear smoothing approach of Li and Hsing (2010), we have the local linear estimators $\hat{\mu}^{[k]}(t) = \hat{b}_{0k}$ and $\hat{G}^{[k]}(s,t) = \hat{c}_{0k} - \hat{\mu}^{[k]}(s)\hat{\mu}^{[k]}(t)$ of

mean and covariance functions of population k, $k = 0, 1$, based on

$$(\hat{b}_{0k}, \hat{b}_{1k}) = \underset{(b_{0k}, b_{1k})}{\arg\min} \frac{1}{n_k} \sum_{i:\, Y_i = k} \frac{1}{m_i} \sum_{l=1}^{m_i} \left\{ \tilde{X}_{il} - b_{0k} - b_{1k}(T_{il} - t) \right\}^2 K\left(\frac{T_{il} - t}{h_\mu} \right), \tag{6.6}$$

and

$$\begin{aligned}(\hat{c}_{0k}, \hat{c}_{1k}, \hat{c}_{2k}) = \underset{(c_{0k}, c_{1k}, c_{2k})}{\arg\min} \quad & \frac{1}{n_k} \sum_{i:\, Y_i = k} \frac{1}{m_i(m_i - 1)} \sum_{l_1 \neq l_2} \left\{ \hat{X}_{il_1} \hat{X}_{il_2} - c_{0k} - c_{1k}(T_{il_1} - s) \right. \\ & \left. - c_{2k}(T_{il_2} - t) \right\}^2 K\left(\frac{T_{il_1} - s}{h_G} \right) K\left(\frac{T_{il_2} - t}{h_G} \right), \end{aligned} \tag{6.7}$$

where $\hat{X}_{il} = \hat{X}_i(T_{il})$, h_μ and h_G are bandwidths.

Under the common eigenfunction assumption, the estimate of joint covariance operator $G = \pi_0 G^{[0]} + \pi_1 G^{[1]}$ can then be denoted by $\hat{G}(s,t) = \hat{\pi}_0 \hat{G}^{[0]}(s,t) + \hat{\pi}_1 \hat{G}^{[1]}(s,t)$, with $\hat{\pi}_k = n_k/n$. Let $(\hat{\lambda}_j, \hat{\psi}_j)$ be the jth eigenvalue-eigenfunction pair of $\hat{G}$. The projection scores for the ith functional predictor $X_i^{[k]}$ can be estimated by $\hat{\xi}_{ij}^{[k]} = \int_{\mathcal{I}} \hat{X}_i^{[k]}(s) \hat{\psi}_j(s) ds$, $i = 1, \ldots, n_k$, $j = 1, 2, \ldots$.

6.3.1 The Weighted-Multiple Quantile Classifier

Denote the empirical τ-quantile of $\hat{\xi}_j^{[k]}$ by $\hat{q}_{jn}^{[k]}(\tau)$, and the empirical representation of $L_{kj}(x, \tau)$ by $\hat{L}_{kjn}(x, \tau) = |\hat{x}_j - \hat{q}_{jn}^{[k]}(\tau)| [\tau I\{\hat{x}_j > \hat{q}_{jn}^{[k]}(\tau)\} + (1 - \tau) I\{\hat{x}_j \leqslant \hat{q}_{jn}^{[k]}(\tau)\}]$, $k = 0, 1$, where $\hat{x}_j = \int_{\mathcal{I}} x(s) \hat{\psi}_j(s) ds$ is the projection score of the new functional observation $x(s)$ along the direction $\hat{\psi}_j$. For a fixed quantile level $\tau \in \mathcal{S}$, the estimated criterion function for classification is given by

$$\hat{Q}_J(x, \tau, \mathbf{w}) = \sum_{j=1}^{J} w_j [\hat{L}_{1jn}(\hat{x}_j, \tau) - \hat{L}_{0jn}(\hat{x}_j, \tau)]. \tag{6.8}$$

In order to determine the empirical optimal values of J, τ and $\mathbf{w}$ for classification, we use the observed rate of correct classification

$$\hat{\Psi}_n(J,\tau,\mathbf{w}) = n^{-1}\Bigg[\sum_{i:\ Y_i=0} I\Big\{\hat{Q}_J(X_i,\tau,\mathbf{w})>0\Big\}+\sum_{i:\ Y_i=1} I\Big\{\hat{Q}_J(X_i,\tau,\mathbf{w})\leqslant 0\Big\}\Bigg], \quad (6.9)$$

and select

$$(\hat{J},\hat{\tau}_n,\hat{\mathbf{w}}) = \underset{\tau\in\mathcal{S},J\in\mathbb{Z},w_j\in[0,1],\sum_{j=1}^{J}w_j=1}{\arg\max}\ \hat{\Psi}_n(J,\tau,\mathbf{w}),$$

as the estimated optimum of J, τ and $\mathbf{w}$. Therefore, the empirically optimal quantile classifier for functional data can be defined by assigning x to Π_0 if

$$\hat{Q}_{\hat{J}}(x,\hat{\tau}_n,\hat{\mathbf{w}}) = \sum_{j=1}^{\hat{J}} w_j[\hat{L}_{1jn}(\hat{x}_j,\hat{\tau}_n)-\hat{L}_{0jn}(\hat{x}_j,\hat{\tau}_n)]>0. \quad (6.10)$$

Note that in (6.8), we fix a single quantile level only for all projection scores. It is known that quantiles on various locations can reflect the whole vision of distribution information. Thus, naturally, we have the following accumulated quantile-based criterion function using multiple quantile levels

$$\hat{Q}_J^M(x,\tau^M,\mathbf{w}^M) = \sum_{j=1}^{J}\left\{\sum_{m=1}^{M_0} w_{jm}\left[\hat{L}_{1jn}(\hat{x}_j,\tau_m)-\hat{L}_{0jn}(\hat{x}_j,\tau_m)\right]\right\}, \quad (6.11)$$

where we use M_0 quantile levels $\tau^M=(\tau_1,\ldots,\tau_{M_0})$ for every component of quantile-based classifiers, and $\mathbf{w}^M=(w_{11},\ldots,w_{1M_0},\ldots,w_{J1},\ldots,w_{JM_0})$. As for the choice of the quantile levels $(\tau_1,\ldots,\tau_{M_0})$, the number J of projection scores, and the weight $\mathbf{w}^M$, we can select the optimal values $(\hat{J},\hat{\tau}^M,\hat{\mathbf{w}}^M)$ by maximizing the observed rate of correct classification $\hat{\Psi}_n(J,\tau^M,\mathbf{w}^M)$, which is obtained by replacing $\hat{Q}_J(x,\tau,\mathbf{w})$ with the criterion function (6.11) in (6.9). Thus, the corresponding empirical optimal multiple-quantile classifier is defined by assigning x to Π_0 if $\hat{Q}_{\hat{J}}^M(x,\hat{\tau}^M,\hat{\mathbf{w}}^M)>0$, and to Π_1 otherwise.

6.3.2 Implementation Procedure

In this subsection, we provide the detailed steps for implementation of the weighted multiple-quantile classifiers based on (6.11).

Step 0: Smooth the discretely observed data $\{(T_{il}, \tilde{X}_i(T_{il}))\}_{l=1}^{m_i}$ by (6.5), for $i = 1, \ldots, n$. The smoothed random trajectories are denoted as $\{\hat{X}_i\}_{i=1}^n$.

Step 1: Estimate the mean and covariance functions of the two populations $\Pi_k, k = 0, 1$ by (6.6) and (6.7), and obtain the pooled covariance function by $\hat{G}(s,t) = \hat{\pi}_0 \hat{G}^{[0]}(s,t) + \hat{\pi}_1 \hat{G}^{[1]}(s,t)$.

Step 2: Estimate the eigenfunctions $\{\psi_j\}_{j=1}^{\infty}$ by solving the eigen-equation

$$\int_{\mathcal{I}} \hat{G}(s,t)\hat{\psi}_j(s)ds = \hat{\lambda}_j \hat{\psi}_j(t),\ j = 1, 2, \ldots,$$

where $\hat{\psi}_j$'s are subject to $\int_{\mathcal{I}} \hat{\psi}_j^2(s)ds = 1$, $\int_{\mathcal{I}} \hat{\psi}_{j_1}(s)\hat{\psi}_{j_2}(s)ds = 0$, $j_1 \neq j_2$, and the projection scores $\hat{\xi}_{ij}^{[k]}$ of random curve X_i can be obtained by $\hat{\xi}_{ij}^{[k]} = \int_{\mathcal{I}} \hat{X}_i^{[k]}(s)\hat{\psi}_j(s)ds$, $i = 1, \ldots, n$, $j = 1, 2, \ldots$.

Step 3: Given the number J of projection scores, the number M_0 of quantile levels, the weight $\mathbf{w}^M$, and the quantile levels $\tau^M = (\tau_1, \ldots, \tau_{M_0})$, obtain the empirical τ_j-quantile $\hat{q}_{jn}^{[k]}(\tau_j)$ of $\hat{\xi}_j^{[k]}$, $j = 1, \ldots, M_0$, and the estimated multiple quantile classification criterion function (6.11).

Step 4: Select the optimal number J of projection scores by controlling the fraction of variances explained. Let $\tau_1, \ldots, \tau_{M_0}$ be equally spaced on the subset $\mathcal{S}$, where $\tau_1 = a_0$, $\tau_{M_0} = 1 - a_0$. Set $a_0 = \frac{1}{2n}$. One can select M_0 by grid search. Let m_0 be a proper integer. Select M_0 and the optimal weight $\mathbf{w}^M$ by maximizing the following observed rate of correct classification

$$(\hat{M}_0, \hat{\mathbf{w}}^M) = \underset{\{M_0 \in \{1,\ldots,m_0\}, \mathbf{w}^M \in [0,1]^{JM_0}\}}{\arg\max} \hat{\Psi}_n(J, \tau^M, \mathbf{w}^M), \tag{6.12}$$

where $\mathbf{w}^M$ is subject to $\sum_{j=1}^{JM_0} w_j = 1$.

Step 5: For a new functional trajectory $X = x$, which is independent of the training data, we classify it to Π_0 if and only if $\hat{Q}^M_{\hat{J}}(x, \hat{\tau}^M, \hat{\mathbf{w}}^M) > 0$.

In Steps 0 and 1, the bandwidth used in each smoothing step is chosen by generalized cross-validation. In Step 4, we look for the optimal values of $\mathbf{w}^M$ for τ^M in $\mathcal{S}$, a closed interval not containing zero. In practice, we should choose a_0 as small as possible while ensuring that the estimated τ-quantiles are still of some use. In contrast, we should choose M_0 as large as possible in order to characterize more distribution information while ensuring that the proposed procedure is still efficient. Given M_0, one can use the constrained optimization algorithm (e.g., we use fmincon() from MATLAB's optimization toolbox) for maximizing (6.12) in Step 4, thus to obtain the optimal value $\hat{\mathbf{w}}^M$. Empirically, when $M_0 = 5$, i.e. 5 equally spaced quantile levels, the proposal classifier generally performs well. The reason might be that, like five number summary of Boxplot, even 5 equally spaced quantile levels are able to characterize the whole distribution of the principle scores approximately. Both simulation studies and real data analysis show such that implementation procedure obtains good classification results.

Remark 6.1. *The single quantile classifier in (6.8) is analogous to the quantile-based classifiers for high-dimensional situation in Henning and Viroli (2016) but not a trivial extension from cross-sectional variables to infinite dimensional observations. For example, projecting of infinite dimensional data into a random variable adds the extra task of estimation of the projection scores for functional data. Also, Henning and Viroli (2016) only validated for finite p-dim vector, leaving consistency unsolved when p goes to infinity for high-dimensional classification. Nevertheless under functional data setting, the unknown truncated number J (corresponding to p) of*

projection scores needs to be estimated, and meanwhile theoretically it shall tend to infinity, thus posing a serious challenge in proof. Furthermore, our proposed single quantile classifier (6.8) serves as a stepping stone for functional classification. Involvement of weights together with multiple quantile locations make it more powerful in classification.

Remark 6.2. *One may also try the adaptive weighted quantile criterion based on the fact that we use the first J projection scores to construct the weighted quantile-based classifiers and their distributions may be different. Thus different quantile levels correspond to different projection scores. Let $\tau^A = (\tau_1, \ldots, \tau_J)$. Then we have*

$$\hat{Q}_J^A(x, \tau^A, \boldsymbol{w}) = \sum_{j=1}^{J} w_j \left[\hat{L}_{1jn}(\hat{x}_j, \tau_j) - \hat{L}_{0jn}(\hat{x}_j, \tau_j) \right], \tag{6.13}$$

where $\{\tau_j\}_{j=1}^J$ may be different from each other. The optimal values $(\hat{J}, \hat{\tau}^A, \hat{\boldsymbol{w}})$ for classification can be also computed by maximizing the observed rate of correct classification $\hat{\Psi}_n(J, \tau^A, \boldsymbol{w})$ obtained by replacing $\hat{Q}_J(x, \tau, \boldsymbol{w})$ with the criterion function (6.13) in (6.9). Thus the corresponding empirical optimal adaptive quantile-based classifier is defined by assigning x to Π_0 if $\hat{Q}_j^A(x, \hat{\tau}^A, \hat{\boldsymbol{w}}) > 0$, and to Π_1 otherwise. However it is computationally expensive. This indicates that treatments for high-dimensional and functional data are quite different.

6.4 Asymptotic Properties

In this section, we consider the case where J tends to infinity as $n \to \infty$, and establish some asymptotic properties for the weighted quantile-based classifier (6.10).

Denote the covariance function of population k by $G^{[k]}(s,t) = \mathrm{Cov}[X^{[k]}(s), X^{[k]}(t)]$. Assume that $G^{[k]}(s,t)$ is continuous. Mercer's theorem tells $G^{[k]}(s,t) = \sum_{j=1}^{\infty} \lambda_j^{[k]} \psi_j^{[k]}(s) \psi_j^{[k]}(t)$,

where the orthonormal eigendecomposition yields eigenfunctions $\psi_j^{[k]}(\cdot)$'s and eigenvalues $\lambda_1^{[k]} \geqslant \lambda_2^{[k]} \geqslant \cdots \geqslant 0$ satisfying $\sum_{j=1}^{\infty} \lambda_j^{[k]} < \infty$ for $k = 0, 1$. Assume that the two populations Π_0 and Π_1 share the same set of eigenfunctions, not necessarily with the same order, following the thought of projecting the data from both groups onto the same basis in Hall et al. (2001). We reorder the eigenfunctions such that $\psi_j^{[k]} \equiv \psi_j$ holds, but $\lambda_j^{[k]}$'s are not necessarily in descending order, for $k = 0, 1$.

Let $\mathcal{S} = [a_0, 1 - a_0]$ for arbitrarily small $0 < a_0 < 0.5$, and $\mathcal{W}_J = \{\mathbf{w} | \sum_{j=1}^{J} w_j = 1, 0 \leqslant w_j \leqslant 1, j = 1, \ldots, J\}$. To proceed, we need the following conditions:

Assumption 6.1. *For all $j = 1, \ldots, J$, $k = 0, 1$, $q_j^{[k]}(\tau)$ is a continuous function of $\tau \in \mathcal{S}$.*

Assumption 6.2. *For all $\tau \in \mathcal{S}$, $\sup_{J \in \mathbb{Z}} \sup_{\boldsymbol{w} \in \mathcal{W}_J} P\big\{Q_J(X, \tau, \boldsymbol{w}) = 0\big\} = 0$.*

Assumption 6.3. *The quantile functions $q_j^{[k]}(\tau)$ have bounded derivative, and satisfy*

$$\sup_{j \geqslant 1} \sup_{\tau \in \mathcal{S}} \left| (q_j^{[k]})'(\tau) \right| < \infty, \ \ k = 0, 1.$$

Assumption 6.4. *The weights $\{w_j\}_{j=1}^J$ satisfy $\sum_{j=1}^J w_j = 1$, $w_j \geqslant 0$, $j = 1, \ldots, J$, for any $J \in \mathbb{Z}$.*

Assumption 6.5. *The covariance operators $G_k(s, t)$ under the populations Π_0 and Π_1 have common eigenfunctions.*

Assumption 6.6. $J\sqrt{\frac{\log n}{n}} \to 0$, *as* $J \to \infty$, $n \to \infty$.

Assumption 6.1 is the same as Assumption 1 in Henning and Viroli (2016). Assumption 6.2 is similar to Assumption 2 in Henning and Viroli (2016), enforced uniformly for $\tau \in \mathcal{S}$, $J \in \mathbb{Z}, \mathbf{w} \in \mathcal{W}_J$. Assumption 6.3 ensures the derivations of

$q_j^{[k]}(\tau)$ are bounded uniformly. Assumption 6.4 concerns a standard weight condition, and Assumption 6.5 is identical to Condition 1 in Dai et al. (2017). Assumption 6.6 controls the convergence rate of the number J of projection scores as $n \to \infty$.

For given $J \in \mathbb{Z}$ and $\mathbf{w} \in \mathcal{W}_J$, define $\hat{\tau}_n(J, \mathbf{w}) \equiv \arg\max_{\tau \in \mathcal{S}} \hat{\Psi}_n(J, \tau, \mathbf{w})$, and $\tau_0(J, \mathbf{w}) \equiv \arg\max_{\tau \in \mathcal{S}} \Psi(J, \tau, \mathbf{w})$. We drop the argument $(J, \mathbf{w})$ for $\hat{\tau}_n(J, \mathbf{w})$ and $\tau_0(J, \mathbf{w})$ when no confusion arises. Then we have the following asymptotic result.

Theorem 6.1. *Under Assumptions 6.1-6.6 and 6.13-6.16 in the Section 6.8, for any $\epsilon > 0$, there exists a sequence $J = J(n, \epsilon) \to \infty$ such that*

$$\inf_{\boldsymbol{w} \in \mathcal{W}_J} P\Big\{ |\Psi(J, \hat{\tau}_n, \boldsymbol{w}) - \Psi(J, \tau_0, \boldsymbol{w})| \leqslant \epsilon \Big\} \to 1, \text{ as } n \to \infty.$$

Remark 6.3. *This theorem means that, the empirical optimal $\hat{\tau}_n$ in the weighted quantile-based classifiers achieves the true correct classification probability for the true optimal τ_0, as $n \to \infty$ and $J \to \infty$.*

Let $\zeta = (\zeta_1, \zeta_2, \ldots)$ denote an infinite sequence of random variables, where each ζ_j has τ-quantiles $q_j(\tau)$ for all $\tau \in \mathcal{S}$ and median zero. Assume that there is at most value u with $F_{\zeta_j}(u) = \tau$ for all $\tau \in \mathcal{S}$, $j = 1, 2, \ldots$. For infinite sequences of constants $(\upsilon_{01}, \upsilon_{02}, \ldots)$ and $(\upsilon_{11}, \upsilon_{12}, \ldots)$, assume that for each $J \in \mathbb{Z}$, the J-dimensional vector $(\xi_1^{[0]}, \ldots, \xi_J^{[0]})$ is identically distributed as $(\upsilon_{01} + \zeta_1, \ldots, \upsilon_{0J} + \zeta_J)$, and the J-dimensional vector $(\xi_1^{[1]}, \ldots, \xi_J^{[1]})$ is identically distributed as $(\upsilon_{11} + \zeta_1, \ldots, \upsilon_{1J} + \zeta_J)$, respectively. Thus, the τ-quantile of $\xi_j^{[0]}$ is $q_j^{[0]}(\tau) = \upsilon_{0j} + q_j(\tau)$, and the τ-quantile of $\xi_j^{[1]}$ is $q_j^{[1]}(\tau) = \upsilon_{1j} + q_j(\tau)$. We also assume that $(X_i, Y_i)_{i=1}^n$ are independent and identically distributed. The following assumption are needed in Theorem 6.2:

Assumption 6.7. *The differences $|q_j^{[1]}(\tau) - q_j^{[0]}(\tau)|$ are uniformly bounded.*

Assumption 6.8. *For sufficiently small $c > 0$, the proportion of values of $j \in [1, J]$ for which $|q_j^{[1]}(\tau) - q_j^{[0]}(\tau)| > c$ for all $\tau \in \mathcal{S}$ is bounded away from zero as $J \to \infty$.*

Assumptions 6.7 and 6.8 are closely related to Assumptions (7) and (8) in Henning and Viroli (2016). Assumption 6.7 imposes the condition that the differences of the quantiles of projection scores between the two groups are uniformly bounded, and Assumption 6.8 requires that a non-negligible proportion of the componentwise differences of the quantiles be bounded away from zero.

The next result states that the proposed weighted quantile-based classifiers achieve near perfect classification under certain conditions.

Theorem 6.2. *Under Assumptions 6.7- 6.8, Assumptions 6.9- 6.12 and 6.13-6.16 given in Section 6.8, with probability converging to one as $n \to \infty$ and $J \to \infty$, the weighted quantile-based classifier (6.10) makes the correct decision,*

$$\sup_{\tau \in \mathcal{S}} \sup_{\boldsymbol{w} \in \mathcal{W}_J} \left[P_{\Pi_0}\{\hat{Q}_J(X, \tau, \boldsymbol{w}) \leqslant 0\} + P_{\Pi_1}\{\hat{Q}_J(X, \tau, \boldsymbol{w}) > 0\} \right] \to 0.$$

This theorem extends the previous results on near perfect classification for high-dimensional data, such as those in Hall et al. (2009) and Henning and Viroli (2016), to the proposed classifiers for functional data. From Theorem 6.2, it is easy to see that near perfect classification occurs if there are infinitely many projection scores relevant to classifying the groups apart. Conditions $B1$-$B2$ are different from that in Dai et al. (2017), in which they assumed that there are sufficient differences in the mean or variance functions under the two groups for achieving near perfect classification. However, we only require that there are sufficient differences between two groups in the quantile function of projection scores in the directions of tail eigenfunctions. Thus, Assumptions 6.7-6.8 are weaker than that in Dai et al. (2017). In addition, since we cannot observe the entire predictor trajectories $\{X_i\}_{i=1}^n$, but rather obtain the irregular/regular repeated measurements of the predictors, the

smoothing errors caused by smoothing the discrete observations for every subject and its influence carried over to the principle components scores estimates should be taken into account. Furthermore, the eigenfunctions are estimated from the observed data, and the estimated errors resulted from the estimates of eigenfunctions must be also considered in the proofs of Theorems 6.1 and 6.2. Thus, in order to obtain theoretical results under presmoothing, we need Assumptions 6.13-6.16 in Section 6.8, which are identical to those in Kong et al. (2016b) and Dai et al. (2017).

6.5 Simulation Studies

To assess the performance of the proposed weMulQ classifier, we consider two scenarios with six examples of Monte Carlo simulation experiments in total, and compare them with some of the existing methods, including

(a) Bayes classifier (Bayes) by Dai et al. (2017);

(b) centroid classifier (Cent) by Delaigle and Hall (2012);

(c) functional logistic regression (Logistic) by Araki et al. (2009).

The random curves of classification objects are discretized at 51 equally spaced time-points over $\mathcal{I} = [0, 1]$, and are disturbed with independent small normal measurement errors with mean zero and standard deviation 0.1. In each case, a training sample curves with moderate sample sizes of $n = 50, 100$ based on 200 replications are generated for training the classifiers, and the same number of curves for evaluating the predictive performance. Each curve is equally likely drawn from either preset population.

We compare the performances of the classifiers in terms of their misclassification rate (MCR). We also draw the Box plots so as to compare their standard error trends.

All simulations are coded in MATLAB 2017b 9.3.0.713579, and executed on a Unix laptop with an Inter Core i7-6700HQ processor and 16 Gbites memory.

6.5.1 Scenario I

In this subsection, the random samples from two populations are generated in the form of $X^{[k]}(s) = \mu^{[k]}(s) + e^{[k]}(s)$, $s \in \mathcal{I}$, $k = 0, 1$, where $\mu^{[k]}(s) = E[X^{[k]}(s)]$, and $e_i^{[k]}(s)$ are Gaussian processes or generalized Gaussian processes (Student-t processes, see Shah et al. (2014)). Using generalized Gaussian processes (Student-t processes) allows to simulate potentially heavy-tailed error terms. Such heavy-tailed error terms are more general and have important usage in stock return modeling and financial time series.

In the first example, we consider two populations with the same mean but different covariances. One error process in Example 1 is Gaussian and the other is heavy-tailed. In the next two examples, we consider two Student-t processes with different means but the same covariance.

Example 6.1. (Gaussian process versus Student-t process (GP vs tP)). *In this example, we have the following two functional populations:*

$$Group\,0: \quad X_i^{[0]}(s) = e_i^{[0]}(s) \quad and \quad Group\,1: \quad X_i^{[1]}(s) = e_i^{[1]}(s),$$

where $e_i^{[0]}$ is a Gaussian process with mean zero and covariance function

$$\sigma(t, s) = 0.25 \exp(-|t - s|^2),$$

and $e_i^{[1]}$ is a Student-t process with mean zero, shape parameter $\sigma(t, s)$, and degree of freedom 3.

Example 6.2. (More complicated mean curves). *In this example, we have the following two functional populations:*

$$Group\,0: \quad X_i^{[0]}(s) = U_i h_1(s) + (1 - U_i) h_2(s) + e_i^{[0]}(s),$$

$$Group\,1: \quad X_i^{[1]}(s) = V_i h_1(s) + (1 - V_i)h_3(s) + e_i^{[1]}(s),$$

where U_i *and* V_i *are uniform random variables on the interval* $[0, 1]$,

$$\begin{aligned} h_1(s) &= \max(6 - |s - 10|, 0)/20, \\ h_2(s) &= h_1\{(s - 4)/20\}, \\ h_3(s) &= h_1\{(s + 4)/20\}, \end{aligned}$$

and $e_i^{[k]}, k = 0, 1$, *are Student-t processes with mean zero, degree of freedom* 3, *and shape parameter* $\tilde{\sigma}(t, s) = 1$, *for* $t = s$, *and* $\tilde{\sigma}(t, s) = 0$, *otherwise. This example is similar to that in Alonso et al. (2012), whereas* $\{e_i^{[k]}(s), s \in \mathcal{I}\}$ *in Alonso et al. (2012) are Gaussian white noise instead of heavy-tailed t distribution.*

Example 6.3. (Mean curves with different pulses). *In this example, we have the following two functional populations:*

$$\begin{aligned} Group\,0: \quad & X_i^{[0]}(s) = 4s + \frac{1}{100} f(s) + e_i^{[0]}(s), \\ Group\,1: \quad & X_i^{[1]}(s) = 4s + e_i^{[1]}(s), \end{aligned}$$

where $f(s)$ *is the probability density function of* $N(0, 0.001^2)$, *and* $\{e_i^{[k]}, k = 0, 1\}$ *are the same as the Student-t process in Example 6.1.*

Table 6.1 shows the means of misclassification rates with empirical standard errors in brackets for the above three examples. The best results in all the tables are in bold. In Example 6.1, two groups have no difference between the mean curves but differ in the setting of the error processes. In this case, the proposed classifier is optimal in the sense that the MCRs and the standard errors are the smallest. The Bayes classifier is comparable with ours. However, the performance of centroid and logistics classifiers perform rather worse. This may be attributed to that the centroid

and Logistic methods only work in the cases where the differences are exclusively in the mean.

In Example 6.2, the mean curves are different but the error processes are the same for the two populations. The weMulQ classifier still has superior performance, although the centroid and Logistic classifiers obtain comparable results. The Bayes method performs relatively poorly and the corresponding MCRs are approximately triple as much as those of the proposed method.

Example 6.3 is similar to the previous one but the difference in the mean curves only comes from a pulse. This case is more difficult to classify than the case in Example 2, and thus the magnitude of misclassification rates are larger than those in Example 6.2. The proposed classifier still outperforms the others. Particularly, when $n = 50$, the weMulQ classifier obtains MCR 21.48%, which is 88% of MCR of the second best classifier, i.e., the Cent classifier. The results of the centroid and logistic classifiers are comparable with those of the proposed method while the misclassification rates of the Bayes classifier are approximately 50% worse than the proposed method.

To demonstrate the robustness of the proposed classifier, the boxplots of the misclassification for the above examples are shown in Figure 6.1. It is easy to see that the quantile classifier has small medians and standard errors of the misclassification rates. Especially, when the sample size is large, the quantile classifier works even better.

Overall, the proposed quantile classifier works well in the sense that it has the smallest misclassification rates, medians, and standard errors. The quantile and Bayes methods work well when the mean curves are the same; and all the methods except the Bayes classifier perform well when the mean curves are different.

Table 6.1: The means and standard errors (in brackets) of the MRCs for Examples 6.1-6.3 Scenario I

n	Quantile	Bayes	Cent	Logistic
	Example 6.1 (Mean curves same, GP vs tP)			
50	**0.1169** (**0.0480**)	0.1238 (0.0520)	0.4280 (0.0716)	0.4611 (0.0776)
100	**0.1080** (**0.0330**)	0.1163 (0.0405)	0.4356 (0.0493)	0.4687 (0.0591)
	Example 6.2 (More complicated mean curves)			
50	**0.0545** (**0.0347**)	0.1481 (0.1643)	0.0668 (0.0366)	0.0724 (0.0427)
100	**0.0532** (**0.0238**)	0.1618 (0.1846)	0.0597 (0.0258)	0.0584 (0.0289)
	Example 6.3 (Mean curves with different pulses)			
50	**0.2148** (**0.0632**)	0.3507 (0.1044)	0.2438 (0.0873)	0.2528 (0.0845)
100	**0.1960** (**0.0440**)	0.2988 (0.0955)	0.1980 (0.0559)	0.2201 (0.0566)

6.5.2 Scenario II

In this subsection, all the random curves are generated in a more ideal way, that is, the curves come from linear combinations of the presetted eigenbases. The random samples from the two populations are generated in the form of $X^{[k]}(s) = \mu^{[k]}(s) + \sum_{j=1}^{50} \xi_j^{[k]} \phi_j(s)$, $s \in \mathcal{I}$, $k = 0, 1$, where $\mu^{[k]}(s)$ is the mean function of $X^{[k]}(s)$, $\xi_j^{[k]}$'s are projections scores with mean zero and variance $\lambda_j^{[k]}$ but from different distributions, and $\phi_j(\cdot)$ is the jth function in the Fourier basis, that is, $\phi_1(s) = 1$, $\phi_2(s) = \sqrt{2}\cos(2\pi s)$, $\phi_3(s) = \sqrt{2}\sin(2\pi s)$, and so on. For the mean functions $\mu^{[k]}(s)$, we set $\mu^{[0]}(s) = 0$ and set $\mu^{[1]}(s) = 0$ or t for the same or different mean settings, respectively. The variance of $\xi_j^{[k]}$ under the population Π_0 are $\lambda_j^{[0]} = \exp(-j/3)$, and those under the population Π_1 are $\lambda_j^{[1]} = \exp(-j/3)$ or $\lambda_j^{[1]} = \exp(-j/2)$ for the same or different variance settings, respectively.

Example 6.4. (Mixture Gaussian). *The samples are generated using* $X_i^{[k]}(s) = \mu^{[k]}(s) + \sum_{j=1}^{50} \xi_{ij}^{[k]} \phi_j(s)$, *where* $\mu^{[0]}(s) = 0$ *and* $\mu^{[1]}(s) = 0$ *or* t *for the same or different mean settings, respectively, and* $\xi_{ij}^{[k]}$ *are independent normal random variables with mean zero and variance* $\lambda_j^{[k]}$.

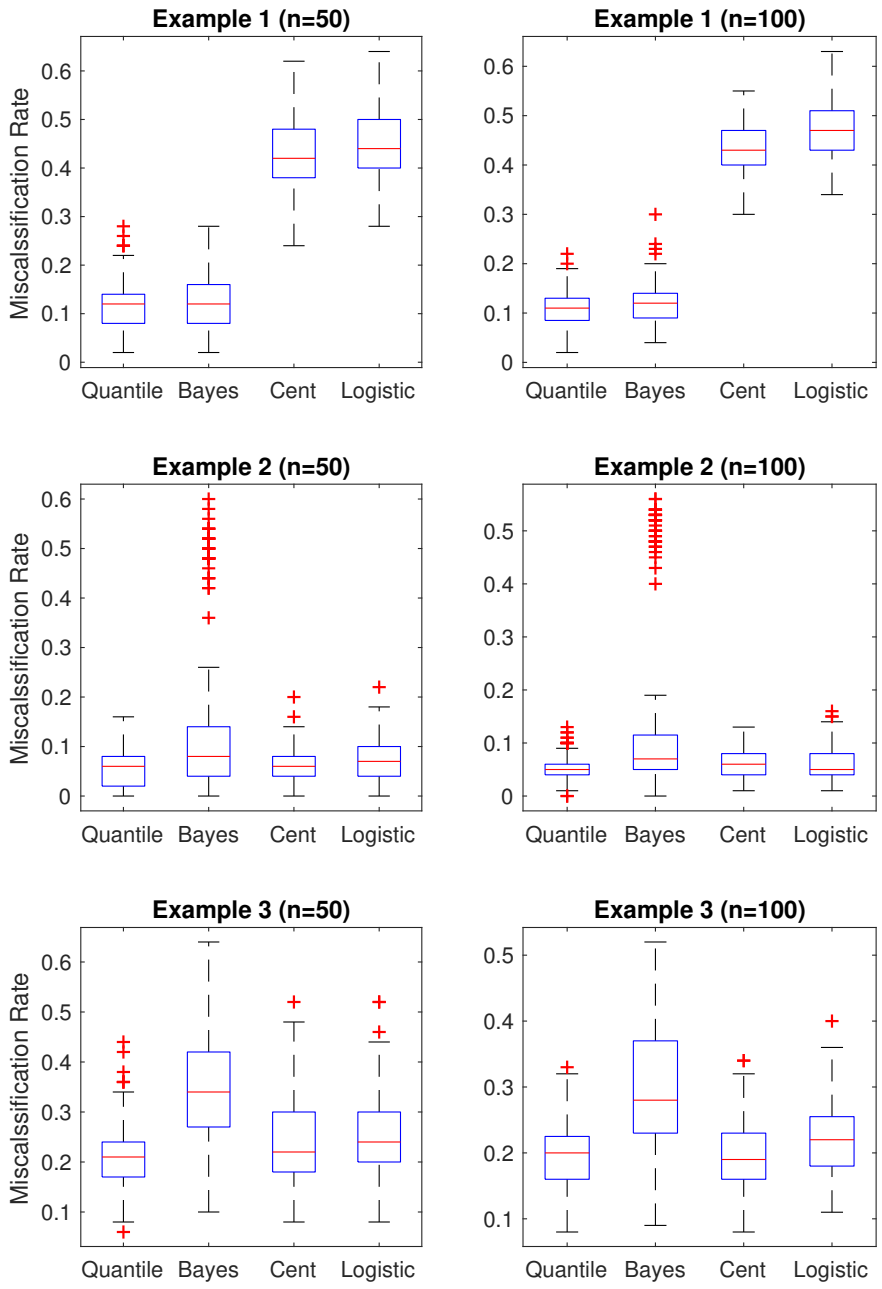

Figure 6.1: The boxplots of misclassification rates for Examples 6.1-6.3 in Scenario I.

Table 6.2: The means and standard errors (in brackets) of the MCRs for Examples 6.4-6.6 in Scenario II ($n = 50$)

μ	λ	Quantile	Bayes	Cent	Logistic
			Example 6.4 (Mixture Gaussian)		
same	diff	0.2676 (0.0739)	**0.2176** (0.0680)	0.4916 (0.0692)	0.5003 (**0.0671**)
diff	same	**0.3881** (0.0734)	0.4246 (0.0786)	0.4007 (0.0769)	0.4141 (**0.0727**)
diff	diff	0.2116 (0.0677)	**0.1772** (**0.0660**)	0.3750 (0.0760)	0.3845 (0.0801)
			Example 6.5 (Norm-t(3))		
same	same	**0.4427** (**0.0671**)	0.4548 (0.0724)	0.4993 (0.0672)	0.5108 (0.0697)
same	diff	0.3753 (0.0719)	**0.3476** (0.0947)	0.4958 (0.0683)	0.5161 (**0.0672**)
diff	same	**0.3376** (**0.0715**)	0.3819 (0.0834)	0.3738 (0.0735)	0.3853 (0.0798)
diff	diff	**0.2640** (0.0743)	0.2726 (0.0777)	0.3572 (**0.0717**)	0.3725 (0.0729)
			Example 6.6 (Norm-Cauchy)		
–	–	**0.0669** (**0.0397**)	0.1438 (0.0663)	0.4029 (0.0643)	0.4257 (0.0819)

Example 6.5. (Gaussian and Student-t (Norm-t(3))). *The samples are generated in the way similar to that in Example 4 except that* $\xi_{ij}^{[k]}, k = 0, 1$ *are independent and identically distributed random variables following normal distribution and* t*-distribution with degree of freedom* 3*, respectively.*

Example 6.6. (Gaussian and Cauchy (Norm-Cauchy)). *The samples are generated using* $X_i^{[k]}(s) = \sum_{j=1}^{50} \xi_{ij}^{[k]} \phi_j(s)$*, where* $\xi_{ij}^{[0]}$ *and* $\xi_{ij}^{[1]}$ *are independent normal random variables and Cauchy random variables, respectively, with the same scale parameter* $\exp(-j/6)$.

Example 6.4 was designed by Dai et al. (2017). While the last two examples are designed to demonstrate the advantages of the proposed method in the cases where the distributions of the scores in different populations are different.

Tables 6.2 and 6.3 display the misclassification rates results with empirical standard errors in brackets for the above three examples in Scenario II for sample size 50 and 100, respectively.

In Example 4, the result coincides with our anticipation that the quantile classifier might not perform the best while the Bayes method has the best performance. When

Table 6.3: The means and standard errors (in brackets) of the MCRs for Examples 6.4-6.6 in Scenario II ($n = 100$)

μ	λ	Quantile	Bayes	Cent	Logistic
			Example 6.4 (Mixture Gaussian)		
same	diff	0.2094 (0.0694)	**0.1479** (**0.0391**)	0.4905 (0.0501)	0.5063 (0.0514)
diff	same	0.3747 (0.0533)	0.3924 (0.0664)	0.3695 (**0.0538**)	**0.3659** (0.0543)
diff	diff	0.1687 (0.0582)	**0.1206** (**0.0366**)	0.3358 (0.0555)	0.3361 (0.0548)
			Example 6.5 (Norm-t(3))		
same	same	**0.4184** (0.0510)	0.4404 (0.0652)	0.5061 (**0.0474**)	0.5101 (0.0509)
same	diff	0.3354 (0.0616)	**0.2827** (**0.0478**)	0.5061 (0.0517)	0.5035 (0.0482)
diff	same	**0.3137** (0.0476)	0.3642 (0.0620)	0.3515 (**0.0472**)	0.3503 (0.0514)
diff	diff	0.2306 (0.0601)	**0.2201** (0.0499)	0.3247 (**0.0495**)	0.3288 (0.0545)
			Example 6.6 (Norm-Cauchy)		
–	–	**0.0499** (**0.0234**)	0.1540 (0.0543)	0.3918 (0.0520)	0.4267 (0.0661)

$n = 50$, the MCR of weMulQ classifier is 1.23 and 1.19 times of the that of the Bayes classifier under the same-diff and diff-diff cases, respectively. Under the diff-same case, the proposed classifier has the best performance. Although the quantile method is not the best under the Gaussian cases, our method still performs better than the centroid and Logistic methods. The boxplots of the MCRs in Figures 6.2 and 6.3 indicate that the Bayes classifier outperforms the others when the variances of scores in two populations are different, while the quantile classifier performs slightly worse but significantly better than the other two classifiers. In the case that the means are different but the variances of scores are the same for the two populations, the four methods are comparable but the proposed one is better for small samples.

For other settings in Dai et al. (2017), where the distributions of the scores are the same for the two populations, the proposed method is still comparable with the Bayes method. Therefore, we do not report these cases here but emphasize on the cases where the distributions of scores are different.

Example 6.5 is designed for generating the curves whose distributions of the scores are different. Tables 6.2 and 6.3 show that the proposed method outperforms the others generally. In the cases when the means are different, the quantile methods

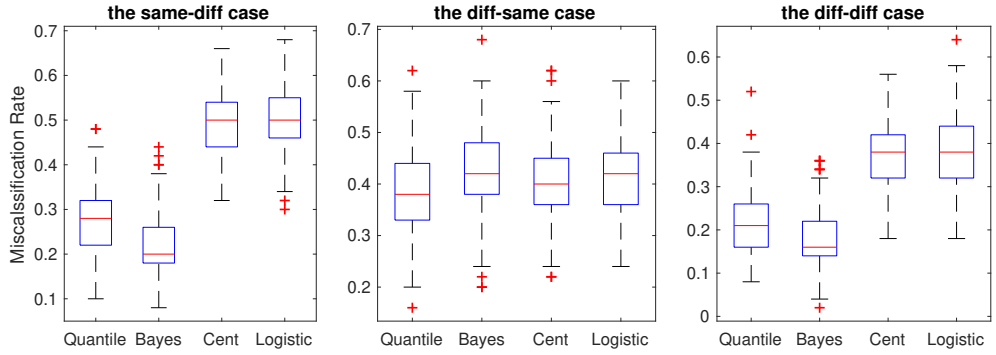

Figure 6.2: The boxplots of misclassification rates for Example 6.4, the mixture Gaussian case ($n = 50$).

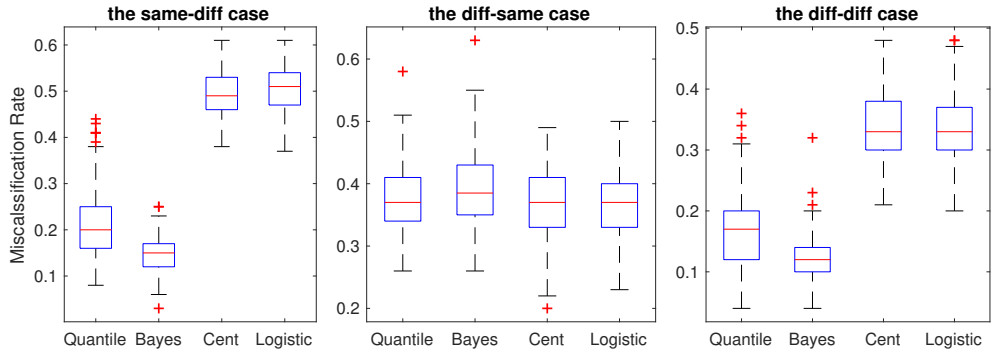

Figure 6.3: The boxplots of misclassification rates for Example 6.4, the mixture Gaussian case ($n = 100$).

outperforms the others. In the case where the means are the same but the variances of scores are different, the Bayes method obtains better results. The MCR of proposed classifier is comparable with that of the Bayes one, for example, when the sample size is 50, the ratio between the MCRs of the weMulQ and the Bayes classifier is 1.08. In the most difficult case, where both the means and variances of the scores are the same, the proposed classifier is able to obtain smaller misclassification rates. When the means and variances of scores are the same, the classification methods based on the first two moments fail but the quantile classifier is still able to work well since the quantile of the distributions of the scores are still different, and thus can be used to classify the curves. Also, we observe from the boxplots of the MCRs in Figures

6.4 and 6.5 that the proposed classifier and the Bayes classifier are comparable when the variances of the scores are different while the proposed classifier outperforms the others when the variances of the scores are the same.

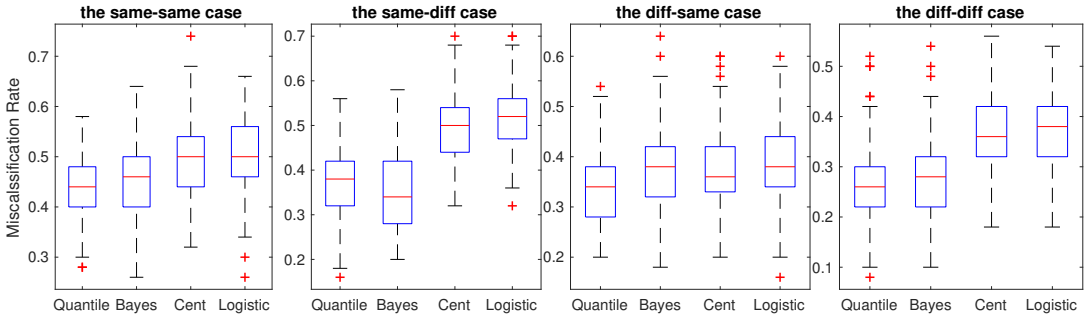

Figure 6.4: The boxplots of misclassification rates for Example 6.5, the norm-t case ($n = 50$).

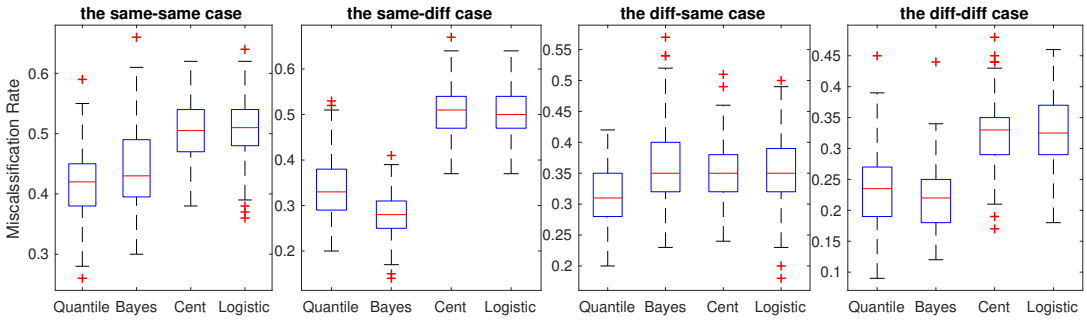

Figure 6.5: The boxplots of misclassification rates for Example 6.5, the norm-t case ($n = 100$).

Example 6.6 is similar to Example 6.5 but the t-distribution is replaced by the Cauchy distribution. Therefore, the mean and variance of the scores in one population are not controllable. From Tables 6.2 and 6.3, it can be seen that the proposed method has much smaller MCRs. The boxplots of the MRCs for Example 6.6 for sample size 50 and 100 are shown in Figures 6.6 and 6.7, respectively. The boxplots indicate that the quantile classifier is the most accurate and robust one among others.

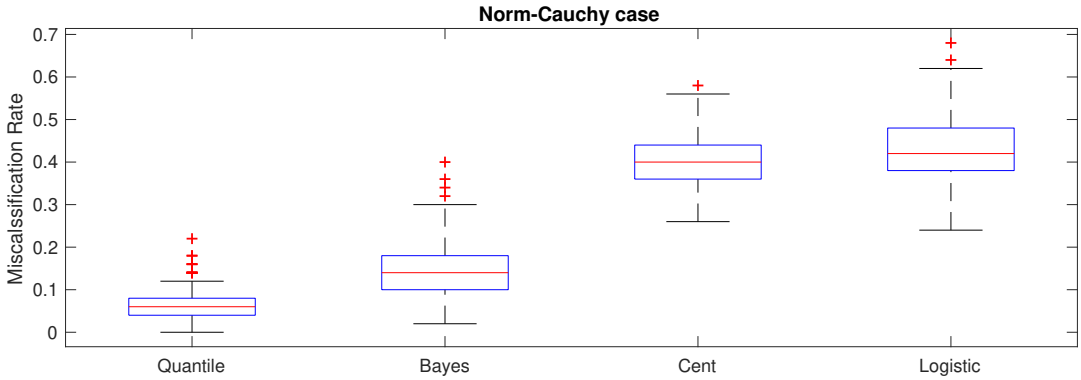

Figure 6.6: The boxplots of misclassification rates for Example 6.6, the norm-cauchy case ($n = 50$).

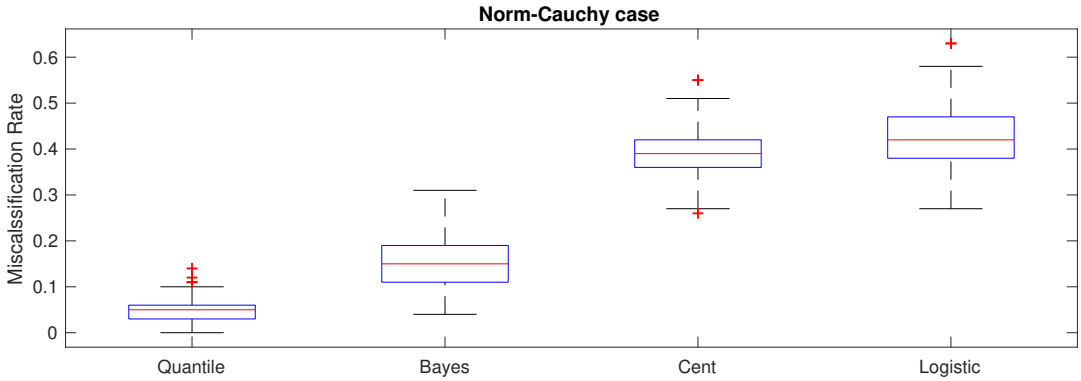

Figure 6.7: The boxplots of misclassification rates for Example 6.6, the norm-cauchy case ($n = 100$).

Overall, the proposed method has good performance in this subsection, especially when the sample size is small. The quantile method works the best when the means are different but the variances of scores are the same or both the means and variances of scores are the same. This conclusion coincides with that in Scenario I.

From the simulation results, it can be seen that the MCRs of the quantile method decrease as the sample size increases. However, we observe that in some cases (Examples 6.2 and 6.6), the performance of the Bayes method is not robust since the corresponding MCRs increase as the sample size increases. In addition, the computing time of the proposed method is much smaller than that of the others. Figure 6.8

shows the average CPU time over 200 simulations for the four methods in Example 6.6. In Figure 6.8, the line corresponding to the quantile method lies lower than the others, which indicates that the proposed weighted multiple quantile classifier is computationally efficient. The small-scaled grid search for M_0 in (6.12), combined with the constrained optimization algorithm, makes the implementation fast. The method does not require a complicated procedure for parameter tuning but is able to obtain a small misclassification rate. The hyperparameter tuning procedures of the classifiers by Dai et al. (2017) require a large number of loops which cost much more computation resources. Thus the proposed classifier is faster than the others. The R package named QuiCFun for the proposal WeMulQ is developed and is available upon request.

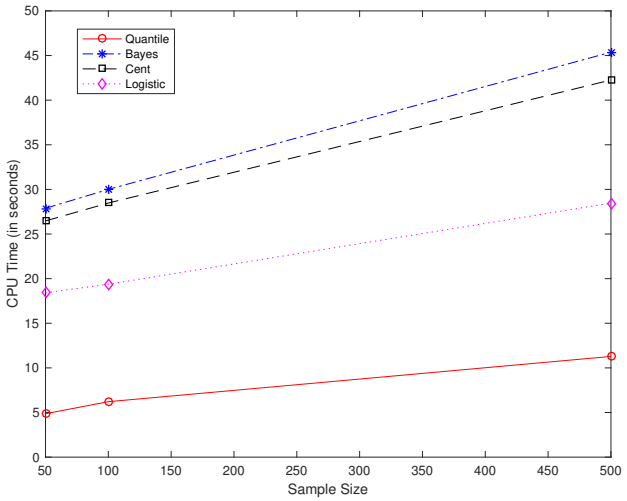

Figure 6.8: The CPU time (in seconds) of the classifiers.

6.6 Analysis of Diffusion Tensor Imaging data

Multiple sclerosis (MS) is the most prevalent chronic neurological disease of the central nervous system that disrupts the flow of information within the brain, and between the brain and body. The US annual 2012 MS extrapolated population was

403,630 according to National Multiple Sclerosis Society (2017). Typical symptoms range from numbness and tingling to blindness and paralysis owing to multifocal demyelinating lesions in the while matter as well as gray matter lesions. Most people with MS are diagnosed between the ages of 20 and 50. Once an individual is detected for MS, clinically early treatment may delay the onset of future attack. Diffusion tensor imagining (DTI) is a quantitative technique that has been widely applied to measure and grade the clinical manifestation and evolution of MS at different stages of the disease (Rovaris et al., 2005; Filippi et al., 2016). It is crucial to classify the degenerative progression of MS using DTI data, for the purpose of effectively treating and managing MS patients(Chen et al. (2017), Vafajoo et al. (2018), Dilokthornsakul et al. (2016), Zwibel and Smrtka (2011), Miller (2004)).

There are several well identified white matter tracts such as right/left corticospinal tract (rCST, lCST), corpus callosum (CCA), and right/left optic radiations tract (rOPR, lOPR) (see Pomann et al. (2016)). Along the afore white matter tracts, there are several modalities that DTI provides: fractional anisotropy (FA), parallel diffusivity, and perpendicular diffusivity (Goldsmith et al. (2012), among others). Goldsmith et al. (2011) and McLean et al. (2014) analyzed tract profiles to discriminate multiple sclerosis cases from healthy controls through functional generalized linear models and functional generalized additive models. Goldsmith et al. (2012) studied the relationship between the white matter tracts in MS patients and cognitive impairment over time through longitudinal penalized functional regression. Pomann et al. (2016) tested the distributions of white matter tract profiles between MS and control groups. Gertheiss et al. (2013) selected important tracts that are associated with the disease status. Other studies related to the DTI data in MS can be found in Morris (2015), Ivanescu et al. (2015), Kong et al. (2016a), Scheipl et al. (2015), among others. However, to the best of our knowledge, there is no methodology developed from the view of classification of functional data.

In this section, we inspect the white matter tract CCA and the modality FA profile to classify MS patients. The data set consists of 100 subjects with MS and 42 healthy controls. For each subject, FA profile at 93 locations along CCA were collected. The data set is available in the R package **refund**.

Figure 6.9 shows the random curves and the mean curves for both groups. Figure 6.10 shows the difference between the covariance functions of the two groups. From Figures 6.9 and 6.10, we see that there is a shift between the two mean curves but the covariance functions are approximately the same since the difference of the covariance surfaces fluctuates around zero. In this case, the method by Dai et al. (2017) may not perform well from our experience in Scenario I of the simulation study.

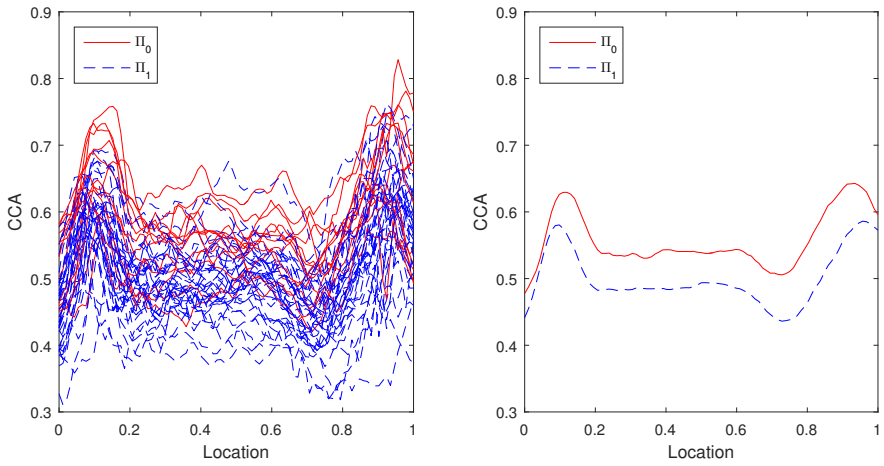

Figure 6.9: White matter measurement trajectories, left panel; mean curves for both groups, right panel.

Since the proposed classifier as well as most of the existing classifiers for functional data are developed based on the projection scores, we check whether the projection scores are different between the two groups. The means and variances of the functional principle component (FPC) scores, which explain 97.01% of the variability, for both groups are plotted in Figure 6.11. The figure indicates that the means and variances of the FPC scores are approximately the same. This might raise the diffi-

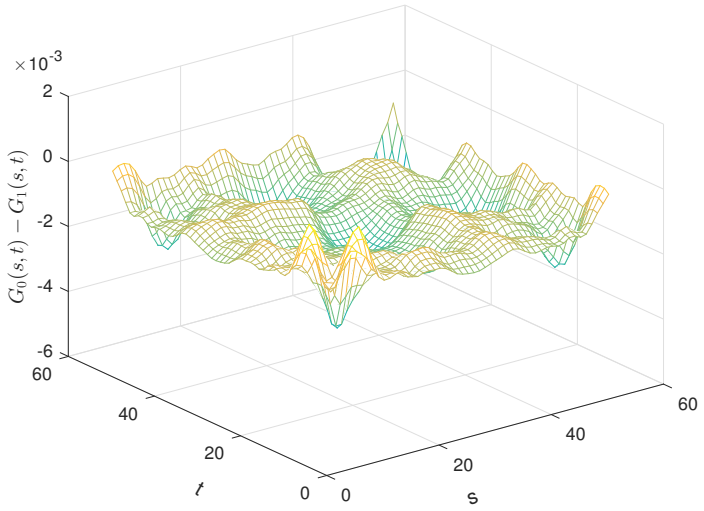

Figure 6.10: The difference between two covariance functions.

culty of classifying the curves based on the FPC scores. Nevertheless, the same mean and variance do not imply the same distribution. Therefore, we report the empirical probability density functions of the projection scores in Figure 6.12. It is obvious that some of the scores have different distributions. Based on our experience from Scenario II in our simulation study, the proposed classifier is able to outperform the existing classifiers in the cases where the projection scores have the same mean and variance but different distributions.

Figure 6.13 presents the misclassification rates of the four methods. The markers corresponding to the best results are solid. We can see that the quantile classifier is the most precious one in the real data case. The weighted indicator (WI) method (Alonso et al. (2012)), a classification method with a semi-distance based on the functional curves directly, is good at rate of false positive. However, more importantly, one should concern about the false negative rate because this retards the treatment to prevent deterioration, incurring loss of labor force and more medical insurance and heath resources. For example, if the rate of misspecification is 3%

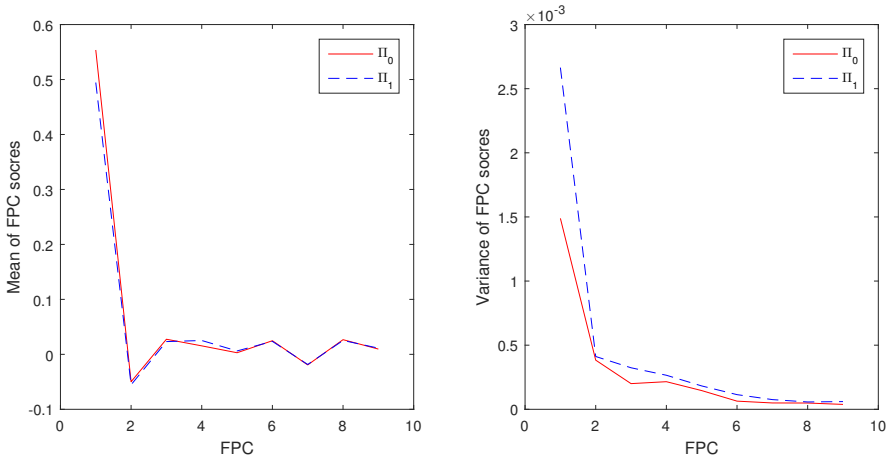

Figure 6.11: Mean and variance for the first nine FPC scores (Π_1 blue dashed line, Π_0 red solid line).

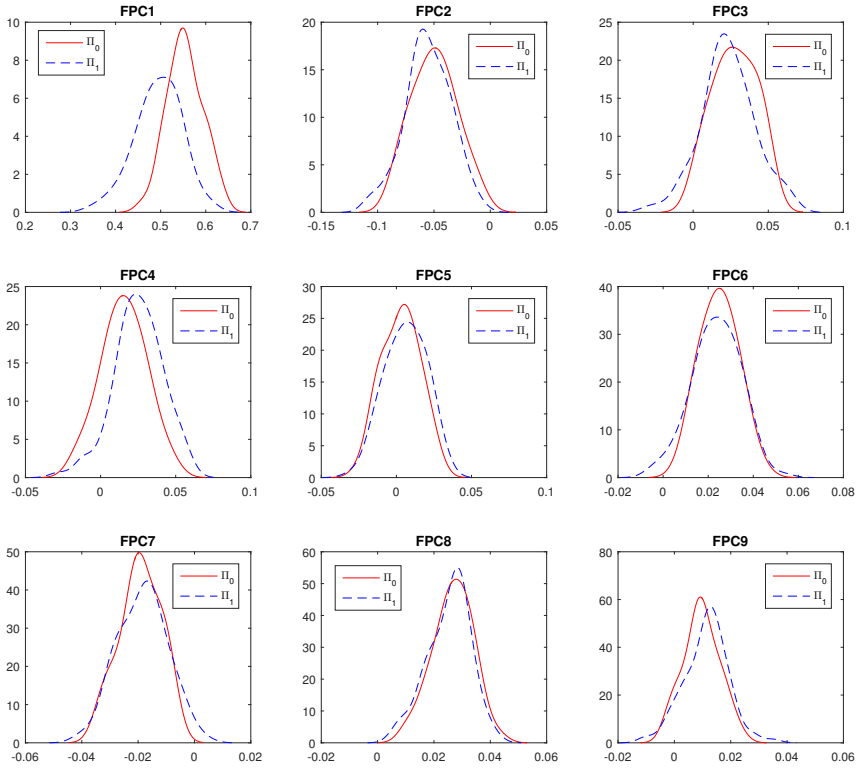

Figure 6.12: pdfs for the first nine FPC scores.

difference, then every 100 subjects, one may diagnose three MS patients as healthy individuals. This kind of misclassification, macroscopically, significantly increases the economic burden of society and the loss of labor force, and hence it is much serious than the false positive error. The other three existing methods were developed based on projection scores. These methods have preferably low false negative rates. However, their false positive rate are relatively high. The proposed classier, which is developed based on the projection scores and the generalized distance, gives the lowest false negative rate and the lowest false positive rate among all projection score-based methods. Overall, the total misclassification rate of our classifier is the lowest. Consequently, the application of the proposed classifier in MS screening will benefit the patients and the medical management of the government because the health risk of the patients and the economic burden of the society can be reduced while the resources of hospitalizations and emergency case can be saved.

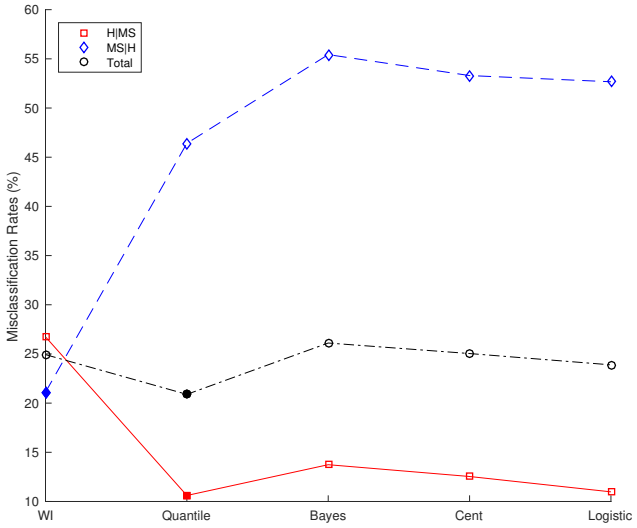

Figure 6.13: Misclassification rates of the existing methods: H|MS indicates the event that an MS patient is misdiagnosed as healthy one, and verse visa for MS|H .

6.7 Discussion

Although the proposed weighted quantile-based classifiers are developed for problems of binary functional supervised classification, the extension of (6.8) to multiple-class functional classification is straightforward. Specifically, we first assume that the covariance operators $G_k(s,t)$ under $\{\Pi_k\}_{k=1}^K$, $K > 2$, have common eigenfunctions, we then project all observations onto this shared set of eigenfunctions, and estimate the weighted quantile distances $\sum_{j=1}^J w_j L_{jk}(x,\tau)$, $k = 1,\ldots,K$. At last, by definition, the weighted quantile-based classifier rule for allocating an new observation $X = x$ to one of K populations $\Pi_1,\ldots,\Pi_K$ is to allocate $X = x$ to the population $\Pi_{k^\star}$, which gives the lowest weighted loss measurement $\sum_{j=1}^J w_j L_{jk}(x,\tau)$, $k \in \{1,\ldots,K\}$. Similarly, (6.11) and (6.13) can be extended to multiple-class functional classification problems. Although such extension is theoretically straightforward, the numerical results would be unstable and not satisfactory in some cases according to our simulation experience. Another solution is to minimize the softmax function instead of the misclassification rate. Softmax loss is widely used for multi-class classification in deep learning community (He et al., 2016). Softmax loss is a multi-class classification version of the binary cross-entropy loss, which is the likelihood function of the logistic regression model under a traditional statistical view. Designing the projection classifier departing from the softmax loss directly might obtain a good numerical result. But it also poses challenges in theoretical development. Such extension deserves further investigation.

6.8 Proofs

In this section, we list some mild assumptions, which are used in Theorems 6.1 and 6.2, and provide some Lemmas and the detailed proofs of the asymptotic results.

6.8.1 Assumptions

Again, let $\mathcal{S} = [a_0, 1 - a_0]$ for arbitrarily small $0 < a_0 < 0.5$. Let $\zeta = (\zeta_1, \zeta_2, \ldots)$ denote an infinite sequence of random variables, where each ζ_j has τ-quantiles $q_j(\tau)$ for all $\tau \in \mathcal{S}$ and median zero. Assume that there is at most value u with $F_{\zeta_j}(u) = \tau$ for all $\tau \in \mathcal{S}$, $j = 1, 2, \ldots$ For infinite sequences of constants $(\upsilon_{01}, \upsilon_{02}, \ldots)$ and $(\upsilon_{11}, \upsilon_{12}, \ldots)$, assume that for each $J \in \mathbb{Z}$, the J-dimensional vector $(\xi_1^{[0]}, \ldots, \xi_J^{[0]})$ is identically distributed as $(\upsilon_{01} + \zeta_1, \ldots, \upsilon_{0J} + \zeta_J)$, and the J-dimensional vector $(\xi_1^{[1]}, \ldots, \xi_J^{[1]})$ is identically distributed as $(\upsilon_{11} + \zeta_1, \ldots, \upsilon_{1J} + \zeta_J)$, respectively. Thus, the τ-quantiles of $\xi_j^{[0]}$ is $q_j^{[0]}(\tau) = \upsilon_{0j} + q_j(\tau)$, and the τ-quantiles of $\xi_j^{[1]}$ is $q_j^{[1]}(\tau) = \upsilon_{1j} + q_j(\tau)$. We also assume that $(X_i, Y_i)_{i=1}^n$ are independent and identically distributed. The following assumptions are needed in Theorem 6.2.

Assumption 6.9. $\lim_{\lambda \to \infty} \sup_{j \geqslant 1} E\{|\zeta_j| I(|\zeta_j| > \lambda)\} = 0$.

Assumption 6.10. *Let* $L_j\{\zeta, \tau, q_j(\tau)\} = \{\tau + (1 - 2\tau) I[\zeta_j \leqslant q_j(\tau)]\} |\zeta_j - q_j(\tau)|$. *For each* $c > 0$,

$$\inf_{j \geqslant 1} \inf_{|u| \geqslant c} \inf_{\tau \in \mathcal{S}} \left(E[L_j\{\zeta, \tau, q_j(\tau) + u\}] - E[L_j\{\zeta, \tau, q_j(\tau)\}] \right) > 0.$$

Assumption 6.11.

$$\inf_{j \geqslant 1} \inf_{\tau \in \mathcal{S}} \left(\min[\tau - P\{\zeta_j \leqslant q_j(\tau) - c\}, 1 - \tau - P\{\zeta_j \geqslant q_j(\tau) + c\}] \right) > 0.$$

Assumption 6.12.

$$\lim_{N \to \infty} \sup_{j_1, j_2 : |j_1 - j_2| \geqslant N} \sup_{B_1, B_2 \in \mathcal{B}} |P(\zeta_{j_1} \in B_1, \zeta_{j_2} \in B_2) - P(\zeta_{j_1} \in B_1) P(\zeta_{j_2} \in B_2)| = 0.$$

Assumptions 6.9-6.12 is similar to Assumptions 3-6 in Henning and Viroli (2016). Assumption 6.9 requires that the first moments of the variables ζ_j be uniformly

bounded in a strong sense, for example, Assumption 6.9 holds if the ζ_j's are identical distributed with finite mean. Assumptions 6.10-6.11 concern uniform continuity and well-definedness of the quantiles. If variables ζ_j's are identically distributed, then these conditions hold under the basic assumption of uniquely defined τ-qunatile. Assumption 6.12 is a strong α-mixing condition, which implies that variables with different index numbers will be approximately independent.

Since we can not observe the entire predictor trajectories $\{X_i\}_{i=1}^n$, but rather obtain the irregular/regular repeated measurements of the predictors, contaminated with additional measurement errors, we must implement the smoothing step with local polynomial fitting to obtain smooth estimates of the predictor trajectories $\{X_i\}_{i=1}^n$. In order to obtain the asymptotic results of Theorem 6.1 and 6.2 under presmoothing, we also need the following additional mild assumptions.

Assumption 6.13. *For $k = 0, 1$, $X^{[k]}(s)$ is twice continuously differentiable on $\mathcal{I}$ with probability approaching one, such that $\int_{\mathcal{J}_X} \mathrm{E}\{d^2X^{[k]}(s)/dt^2\}\, ds < \infty$.*

Assumption 6.14. *For $i = 1, \ldots, n$, the measurement times $\{T_{il}, l = 1, \ldots, m_i\}$ can be generated by $T_{il} = G_i^{-1}\{(l-1)/(m_i-1)\}$, where $G_i(t) = \int_{-\infty}^t g_i(s)ds$, and the density function $g_i(\cdot)$ is uniformly smooth over i, satisfying $\int_{\mathcal{I}} g_i(s)ds = 1$, $0 < C_1 < \inf_i\{\inf_{s\in\mathcal{I}} g_i(s)\} < \sup_i\{\sup_{s\in\mathcal{I}} g_i(s)\} < C_2 < \infty$.*

Assumption 6.15. *There exist a common sequence of bandwidths h such that $0 < \inf_{i=1,\ldots,n} h_i/h < \sup_{i=1,\ldots,n} h_i/h < \infty$, where h_i is the bandwidth for smoothing X_i. The kernel function $\kappa(\cdot)$ is smooth and compactly supported on $\mathcal{I}$.*

Assumption 6.16. *Let $\mathcal{I} = [a_0, b_0]$, $T_{i0} = a_0$, $T_{im_i} = b_0$, $\Delta_i = \sup\{T_{il+1} - T_{il}, l = 1, \ldots, m_i\}$, $m_0 = \inf_{i=1,\ldots,n} m_i$, we have $\sup_i \Delta_i = O(m_0^{-1})$, $h \sim m_0^{-1/5}$ and $m_0 n^{-5/4} \to \infty$, as $n \to \infty$.*

Assumption 6.13 is standard for local linear smoothers. Assumptions 6.14 and

6.16 concern how the functional predictors are sampled and smoothed. Assumption 6.15 is aimed to guarantee that the smooth estimates $\hat{X}_i$ serve as well as the true functional predictors X_i. Assumptions 6.13-6.16 are also used in Kong et al. (2016b) and Dai et al. (2017). From Assumption 6.16, we know that the repeated observations are sufficiently dense for each subject.

6.8.2 Lemmas

To prove Theorem 6.1, we need the following Lemma.

Lemma 6.1. *For any given $\boldsymbol{w} \in \mathcal{W}_J$, and any $\epsilon > 0$, there exists a sequence $J \to \infty$ such that*

$$P\left(\sup_{\tau \in \mathcal{S}} |\hat{\Psi}_n(J, \tau, \boldsymbol{w}) - \Psi(J, \tau, \boldsymbol{w})| \leqslant \epsilon \right) \to 1, \quad n \to \infty, J \to \infty.$$

Proof of Lemma 6.1. Denote

$$\begin{aligned} \Psi_n(J, \tau, \mathbf{w}) \quad = \quad & \frac{1}{n} \Bigg\{ \sum_{i:Y_i=0} I\left[\sum_{j=1}^{J} w_j \{L_{1jn}(\xi_{ij}, \tau) - L_{0jn}(\xi_{ij}, \tau)\} > 0 \right] \\ & + \sum_{i:Y_i=1} I\left[\sum_{j=1}^{J} w_j \{L_{1jn}(\xi_{ij}, \tau) - L_{0jn}(\xi_{ij}, \tau)\} \leqslant 0 \right] \Bigg\}, \end{aligned} \tag{6.14}$$

where $L_{kjn}(\xi_{ij}, \tau) \equiv L_j(\xi_{ij}, \tau, q_{jn}^{[k]}(\tau)) = [\tau + (1 - 2\tau) I\{\xi_{ij} \leqslant q_{jn}^{[k]}(\tau)\}] |\xi_{ij} - q_{jn}^{[k]}(\tau)|$, and $q_{jn}^{[k]}(\tau)$ is the empirical τ-quantile of $\xi_j^{[k]}, k = 0, 1,\ j = 1, 2, \ldots$ In words, we use the true projection scores ξ_{ij} and the true empirical quantile $q_{jn}^{[k]}(\tau)$ to define the observed rate of correct classification (6.14). From the definition of $\hat{\Psi}_n(J, \tau, \mathbf{w})$ and $\Psi(J, \tau, \mathbf{w})$, we obtain

$$\begin{aligned} & \sup_{\tau \in \mathcal{S}} |\hat{\Psi}_n(J, \tau, \mathbf{w}) - \Psi(J, \tau, \mathbf{w})| \\ \leqslant & \sup_{\tau \in \mathcal{S}} |\hat{\Psi}_n(J, \tau, \mathbf{w}) - \Psi_n(J, \tau, \mathbf{w})| + \sup_{\tau \in \mathcal{S}} |\Psi_n(J, \tau, \mathbf{w}) - \Psi(J, \tau, \mathbf{w})| \qquad (6.15) \\ \equiv & U_{1n} + U_{2n}. \end{aligned}$$

We first consider the term U_{2n}. Suppose that $U_{2n} \xrightarrow{p} 0$ does not hold as $n \to \infty$ and $J \to \infty$. It means that there exist $\epsilon > 0$, $\delta > 0$, $J \in \mathbb{Z}$, a sequence $\mathcal{M}$ of $\{1, 2, \ldots\}$, and $\{\tau_m^*\}_{m \in \mathcal{M}}$ such that $P\{|\Psi_m(J, \tau, \mathbf{w}) - \Psi(J, \tau, \mathbf{w})| > \epsilon\} \geqslant \delta$, for any $m \in \mathcal{M}$. Since $\{\tau_m^*\}_{m \in \mathcal{M}}$ is bounded and at least a subsequence has a limit, there exists $\tau^* = \lim_{m \to \infty} \tau_m^*$. Note that for any $m \in \mathcal{M}$ and $J \in \mathbb{Z}$,

$$\begin{aligned} & |\Psi_m(J, \tau_m^*, \mathbf{w}) - \Psi(J, \tau_m^*, \mathbf{w})| \\ \leqslant\ & |\Psi_m(J, \tau_m^*, \mathbf{w}) - \Psi_m(J, \tau^*, \mathbf{w})| + |\Psi_m(J, \tau^*, \mathbf{w}) - \Psi(J, \tau^*, \mathbf{w})| \\ & + |\Psi(J, \tau^*, \mathbf{w}) - \Psi(J, \tau_m^*, \mathbf{w})| \\ \equiv\ & U_{21m} + U_{22m} + U_{23m}. \end{aligned} \tag{6.16}$$

Since $\Psi(\cdot, \tau, \cdot)$ is continuous with respect to τ, for any $J \in \mathbb{Z}$ and $\mathbf{w} \in \mathcal{W}_J$, the term U_{23m} converges to zero as $m \to \infty$. For the term U_{22m}, we define a true version of $\Psi_n(J, \tau, \mathbf{w})$, using the true quantiles instead of the empirical ones, that is,

$$\begin{aligned} \tilde{\Psi}_n(J, \tau, \mathbf{w}) = \frac{1}{n} \Bigg\{ & \sum_{i: Y_i = 0} I\Bigg[\sum_{j=1}^{J} w_j (L_{1j}(\xi_{ij}, \tau) - L_{0j}(\xi_{ij}, \tau)) > 0 \Bigg] \\ & + \sum_{i: Y_i = 1} I\Bigg[\sum_{j=1}^{J} w_j (L_{1j}(\xi_{ij}, \tau) - L_{0j}(\xi_{ij}, \tau)) > 0 \Bigg] \Bigg\}, \end{aligned}$$

where $L_{kj}(\xi_{ij}, \tau) = [\tau + (1 - 2\tau) I(\xi_{ij} \leqslant q_j^{[k]}(\tau))] |\xi_{ij} - q_j^{[k]}(\tau)|$. Thus, we have

$$\begin{aligned} U_{22m} &= |\Psi_m(J, \tau^*, \mathbf{w}) - \Psi(J, \tau^*, \mathbf{w})| \\ &\leqslant |\Psi_m(J, \tau^*, \mathbf{w}) - \tilde{\Psi}_m(J, \tau^*, \mathbf{w})| + |\tilde{\Psi}_m(J, \tau^*, \mathbf{w}) - \Psi(J, \tau^*, \mathbf{w})| \\ &\equiv U_{221m} + U_{222m}. \end{aligned}$$

Because of the strong law of large numbers, $\lim_{m \to \infty} |\tilde{\Psi}_m(J, \tau^*, \mathbf{w}) - \Psi(J, \tau^*, \mathbf{w})| \xrightarrow{a.s.} 0$, for any given $J \in \mathbb{Z}$, $\mathbf{w} \in \mathcal{W}_J$. Since $L_{kj}(\xi_{ij}, \tau) \equiv L_{kj}[\xi_{ij}, \tau, q_j^{[k]}(\tau)]$ is continuous on

$q_j^{[k]}(\tau)$, and $\lim_{m\to\infty} q_{jm}^{[k]}(\tau) \overset{a.s.}{=} q_j^{[k]}(\tau)$, $j = 1, 2, \ldots$ Thus, for any given $J \in \mathbb{Z}$, $\mathbf{w} \in \mathcal{W}_J$, $U_{221m} \xrightarrow{a.s.} 0$, and then $U_{22m} \xrightarrow{a.s.} 0$ as $m \to \infty$.

We now consider the term $U_{21m} = |\Psi_m(J, \tau^*, \mathbf{w}) - \Psi_m(J, \tau^*, \mathbf{w})|$. Note that

$$\begin{aligned}\Psi_m(J, \tau, \mathbf{w}) = \frac{1}{m}\Bigg(&\sum_{i:Y_i=0} I\left[\sum_{j=1}^{J} w_j\{L_{1jm}(\xi_{ij}, \tau) - L_{0jm}(\xi_{ij}, \tau)\} > 0\right] \\ &+ \sum_{i:Y_i=1} I\left[\sum_{j=1}^{J} w_j\{L_{1jm}(\xi_{ij}, \tau) - L_{0jm}(\xi_{ij}, \tau)\} \leqslant 0\right]\Bigg),\end{aligned} \tag{6.17}$$

where $L_{kjm}(\xi_{ij}, \tau) = [\tau + (1 - 2\tau)I(\xi_{ij} \leqslant q_{jm}^{[k]}(\tau))]|\xi_{ij} - q_{jm}^{[k]}(\tau)|$. Note that

$$\begin{aligned}&|q_{jm}^{[k]}(\tau_m^*) - q_{jm}^{[k]}(\tau^*)| \\ \leqslant &|q_{jm}^{[k]}(\tau^*) - q_j^{[k]}(\tau^*)| + |q_{jm}^{[k]}(\tau_m^*) - q_j^{[k]}(\tau_m^*)| + |q_j^{[k]}(\tau_m^*) - q_j^{[k]}(\tau^*)|.\end{aligned} \tag{6.18}$$

It follows from Theorem 3 in Mason (1982) that $\lim_{m\to\infty} \sup_{\tau\in\mathcal{S}} |q_{jm}^{[k]}(\tau) - q_j^{[k]}(\tau)| \xrightarrow{a.s} 0$, for $j \in \mathbb{Z}$. Hence, the first two terms on the right-hand side of (6.18) converge to zero almost surely as $m \to \infty$. For the last term of (6.18), since $\tau_m^\star \to \tau^\star$ as $m \to \infty$, applying conditions (A1) and (A3) gives $\sup_{j\geqslant 1} \left|q_j^{[k]}(\tau_m^*) - q_j^{[k]}(\tau^*)\right| \xrightarrow{a.s} 0$ as $m \to \infty$. Then it follows from (6.18) that

$$|q_{jm}^{[k]}(\tau_m^*) - q_{jm}^{[k]}(\tau^*)| \xrightarrow{a.s} 0, \text{ as } m \to \infty, j = 1, 2, \ldots. \tag{6.19}$$

Let

$$Q_J(X, \tau, \mathbf{w}) = \sum_{j=1}^{J} w_j[L_{1j}(\xi_j, \tau) - L_{0j}(\xi_j, \tau)],$$

$$\tilde{Q}_J(X, \tau, \mathbf{w}) = \sum_{j=1}^{J} w_j[L_{1jn}(\xi_j, \tau) - L_{0jn}(\xi_j, \tau)].$$

For any fixed $\epsilon > 0$, define

$$X_\epsilon(J, \mathbf{w}) = \left\{ X \middle| |Q_J(X, \tau, \mathbf{w})| > \epsilon \right\} \cap \left\{ X \middle| \sum_{j=1}^{J} w_j |\xi_j| \leqslant \epsilon^{-1} \right\}.$$

Then we have

$$\begin{aligned}
&|\Psi_m(J, \tau_m^*, \mathbf{w}) - \Psi_m(J, \tau^*, \mathbf{w})| \\
=&\frac{1}{m} \Bigg\{ \sum_{\substack{i:Y_i=0 \\ X_i \in X_\epsilon(J,\mathbf{w})}} \left(I[\tilde{Q}_J(X_i, \tau_m^*, \mathbf{w}) > 0] - I[\tilde{Q}_J(X_i, \tau^*, \mathbf{w}) > 0] \right) \\
&+ \sum_{\substack{i:Y_i=1 \\ X_i \in X_\epsilon(J,\mathbf{w})}} \left(I[\tilde{Q}_J(X_i, \tau_m^*, \mathbf{w}) \leqslant 0] - I[\tilde{Q}_J(X_i, \tau^*, \mathbf{w}) \leqslant 0] \right) \\
&+ \sum_{\substack{i:Y_i=0 \\ X_i \notin X_\epsilon(J,\mathbf{w})}} \left(I[\tilde{Q}_J(X_i, \tau_m^*, \mathbf{w}) > 0] - I[\tilde{Q}_J(X_i, \tau^*, \mathbf{w}) > 0] \right) \\
&+ \sum_{\substack{i:Y_i=1 \\ X_i \notin X_\epsilon(J,\mathbf{w})}} \left(I[\tilde{Q}_J(X_i, \tau_m^*, \mathbf{w}) \leqslant 0] - I[\tilde{Q}_J(X_i, \tau^*, \mathbf{w}) \leqslant 0] \right) \Bigg\} \\
\equiv& m^{-1}(V_{1m} + V_{2m} + V_{3m} + V_{4m}).
\end{aligned}$$

For large m and arbitrarily $\delta > 0$, we have

$$m^{-1}(V_{3m} + V_{4m}) \leqslant 1 - P\{X_\epsilon(J, \mathbf{w})\} + \delta, \quad a.s.$$

As for the term $m^{-1}(V_{1m} + V_{2m})$, we note that for $X \in X_\epsilon(J, \mathbf{w})$,

$$\begin{aligned}
&|\tilde{Q}_J(X, \tau_m^*, \mathbf{w}) - \tilde{Q}_J(X, \tau^*, \mathbf{w})| \\
\leqslant\ & 2 \sum_{j=1}^{J} w_j |\xi_j| |\tau_m^* - \tau^*| + 8 \sum_{j=1}^{J} w_j |q_{jm}^{[k]}(\tau_m^*) - q_{jm}^{[k]}(\tau^*)|. \qquad (6.20)
\end{aligned}$$

Since $|\tau_m^* - \tau^*| \to 0$ as $m \to \infty$, and $\sum_{j=1}^{J} w_j |\xi_j| \leqslant \epsilon^{-1}$ for $X \in X_\epsilon(J, \mathbf{w})$, the first term on the right hand side of (6.20) can be arbitrarily small, for large enough m

and $X \in X_\epsilon(J, \mathbf{w})$. Applying the well-known results in Csörgo and Révész (1981), we have

$$\limsup_{n\to\infty} (\log\log n)^{-1/2} n^{1/2} \sup_{\tau\in T} \left|q_{jn}^{[k]}(\tau) - q_j^{[k]}(\tau)\right| \stackrel{a.s}{=} 2^{-1/2}, \ k = 0, 1, \tag{6.21}$$

for $j = 1, 2, \ldots$ For the second term on right hand side of (6.20), by applying (6.18), (6.21) and Condition (A3), we have

$$\begin{aligned} & \sum_{j=1}^{J} w_j \left|q_{jm}^{[k]}(\tau_m^*) - q_{jm}^{[k]}(\tau^*)\right| \\ \leqslant & \sum_{j=1}^{J} w_j \left[|q_{jm}^{[k]}(\tau^*) - q_j^{[k]}(\tau^*)| + |q_{jm}^{[k]}(\tau_m^*) - q_j^{[k]}(\tau_m^*)| + |q_j^{[k]}(\tau_m^*) - q_j^{[k]}(\tau^*)|\right] \\ = & \left(\sum_{j=1}^{J} w_j\right) O_{a.s.}\left(\sqrt{\frac{\log\log n}{n}}\right) + \left(\sum_{j=1}^{J} w_j\right) |\tau_m^* - \tau^*| \xrightarrow{a.s.} 0, \end{aligned}$$

as $m \to \infty$, for any $J \in \mathbb{Z}$. Then, $|\tilde{Q}_J(X_i, \tau_m^*, \mathbf{w}) - \tilde{Q}_J(X_i, \tau^*, \mathbf{w})| \xrightarrow{a.s.} 0$ for large enough m, $X_i \in X_\epsilon(J, \mathbf{w})$, and all $J \in \mathbb{Z}$, $i = 1, \ldots, n$. Thus, for $X_i \in X_\epsilon(J, \mathbf{w})$, $\tilde{Q}_J(X_i, \tau_m^*, \mathbf{w})$ and $\tilde{Q}_J(X_i, \tau^*, \mathbf{w})$ are identical in sign, and the corresponding indicator functions therefore are the same, almost surely. It follows from Condition (A.2) that $P\{X_\epsilon(J, \mathbf{w})\} \to 1$, as $\epsilon \to 0$, for any $J \in \mathbb{Z}$ and $\mathbf{w} \in \mathcal{W}_J$, which implies $U_{21m} \xrightarrow{a.s.} 0$, for large m and any $J \in \mathbb{Z}$. Thus, we have $U_{2n} \xrightarrow{a.s.} 0$, as $m \to \infty$.

We next consider the term U_{1n}. Note that

$$\begin{aligned} \hat{\Psi}_n(J, \tau, \mathbf{w}) = \frac{1}{n}\Bigg\{ & \sum_{i:Y_i=0} I\left[\sum_{j=1}^{J} w_j \{\hat{L}_{1jn}(\hat{\xi}_{ij}, \tau) - \hat{L}_{0jn}(\hat{\xi}_{ij}, \tau)\} > 0\right] \\ & + \sum_{i:Y_i=1} I\left[\sum_{j=1}^{J} w_j \{\hat{L}_{1jn}(\hat{\xi}_{ij}, \tau) - \hat{L}_{0jn}(\hat{\xi}_{ij}, \tau)\} \leqslant 0\right]\Bigg\}, \end{aligned}$$

where $\hat{L}_{kjn}(\hat{\xi}_{ij}, \tau) = [\tau + (1 - 2\tau) I(\hat{\xi}_{ij} \leqslant \hat{q}_{jn}^{[k]}(\tau))]|\hat{\xi}_{ij} - \hat{q}_{jn}^{[k]}(\tau)|$. Since $\tilde{Q}_J(X_i, \tau, \mathbf{w}) = \sum_{j=1}^{J} w_j [L_{1jn}(\xi_{ij}, \tau) - L_{0jn}(\xi_{ij}, \tau)]$, and $\hat{Q}_J(X_i, \tau, \mathbf{w}) = \sum_{j=1}^{J} w_j [\hat{L}_{1jn}(\hat{\xi}_{ij}, \tau) - \hat{L}_{0jn}(\hat{\xi}_{ij}, \tau)]$,

we have

$$\begin{aligned}
&\hat{\Psi}_n(J,\tau,\mathbf{w}) - \Psi_n(J,\tau,\mathbf{w}) \\
=&\frac{1}{n}\sum_{i:Y_i=0}\big\{I[\hat{Q}_J(X_i,\tau,\mathbf{w})>0] - I[\tilde{Q}_J(X_i,\tau,\mathbf{w})>0]\big\} \\
&+\frac{1}{n}\sum_{i:Y_i=0}\big\{I[\hat{Q}_J(X_i,\tau,\mathbf{w})\leqslant 0] - I[\tilde{Q}_J(X_i,\tau,\mathbf{w})\leqslant 0]\big\}.
\end{aligned}$$

Note that

$$\begin{aligned}
&\hat{Q}_J(X_i,\tau,\mathbf{w}) - \tilde{Q}_J(X_i,\tau,\mathbf{w}) \\
=&\sum_{j=1}^{J} w_j[\hat{L}_{1jn}(\hat{\xi}_{ij},\tau) - \hat{L}_{0jn}(\hat{\xi}_{ij},\tau)] - \sum_{j=1}^{J} w_j[L_{1jn}(\xi_{ij},\tau) - L_{0jn}(\xi_{ij},\tau)] \\
=&\sum_{j=1}^{J} w_j\Big\{[\tau+(1-2\tau)I(\hat{\xi}_{ij}\leqslant \hat{q}_{jn}^{[1]}(\tau))]|\hat{\xi}_{ij}-\hat{q}_{jn}^{[1]}(\tau)| \\
&-[\tau+(1-2\tau)I(\hat{\xi}_{ij}\leqslant \hat{q}_{jn}^{[0]}(\tau))]|\hat{\xi}_{ij}-\hat{q}_{jn}^{[0]}(\tau)|\Big\} \\
&-\sum_{j=1}^{J} w_j\Big\{[\tau+(1-2\tau)I(\xi_{ij}\leqslant q_{jn}^{[1]}(\tau))]|\xi_{ij}-q_{jn}^{[1]}(\tau)| \\
&-[\tau+(1-2\tau)I(\xi_{ij}\leqslant q_{jn}^{[0]}(\tau))]|\xi_{ij}-q_{jn}^{[0]}(\tau)|\Big\}.
\end{aligned}$$

Thus, we focus on

$$\begin{aligned}
\hat{\xi}_{ij} - \hat{q}_{jn}^{[k]}(\tau) &= (\hat{\xi}_{ij}-\xi_{ij}) + [\xi_{ij} - q_j^{[k]}(\tau)] - [\hat{q}_{jn}^{[k]}(\tau) - q_{jn}^{[k]}(\tau)] - [q_{jn}^{[k]}(\tau) - q_j^{[k]}(\tau)] \\
&\equiv D_1 + D_2 - D_3 - D_4,
\end{aligned}$$

and $\xi_{ij} - q_{jn}^{[k]}(\tau) = \{\xi_{ij} - q_j^{[k]}(\tau)\} - \{q_{jn}^{[k]}(\tau) - q_j^{[k]}(\tau)\} \equiv D_2 - D_4$, for $k = 0, 1$. For

any $X \in \mathcal{L}^2(\mathcal{I})$, we have

$$\begin{aligned}\hat{Q}_J(X,\tau,\mathbf{w}) - \tilde{Q}_J(X,\tau,\mathbf{w}) &= \sum_{j=1}^{J} w_j \Big\{ [\tau + (1-2\tau) I(\hat{x}_j \leqslant \hat{q}_{jn}^{[1]}(\tau))] |\hat{x}_j - \hat{q}_{jn}^{[1]}(\tau)| \\ &\quad - [\tau + (1-2\tau) I(x_j \leqslant q_{jn}^{[1]}(\tau))] |x_j - q_{jn}^{[1]}(\tau)| \Big\} \\ &\quad - \sum_{j=1}^{J} w_j \Big\{ [\tau + (1-2\tau) I(\hat{x}_j \leqslant \hat{q}_{jn}^{[0]}(\tau))] |\hat{x}_j - \hat{q}_{jn}^{[0]}(\tau)| \\ &\quad - [\tau + (1-2\tau) I(x_j \leqslant q_{jn}^{[0]}(\tau))] |x_j - q_{jn}^{[0]}(\tau)| \Big\} \\ &= \sum_{j=1}^{J} w_j [R_{1j}(x,\tau) - R_{0j}(x,\tau)],\end{aligned}$$

where

$$\begin{aligned}R_{kj}(x,\tau) &= [\tau + (1-2\tau) I(\hat{x}_j \leqslant \hat{q}_{jn}^{[k]}(\tau))] |\hat{x}_j - \hat{q}_{jn}^{[k]}(\tau)| \\ &\quad - [\tau + (1-2\tau) I(x_j \leqslant q_{jn}^{[k]}(\tau))] |x_j - q_{jn}^{[k]}(\tau)|,\end{aligned}$$

for $k = 0, 1$. Denote $\triangle(\tau) \equiv -(\hat{x}_j - x_j) + \hat{q}_{jn}^{[k]}(\tau) - q_{jn}^{[k]}(\tau)$. Then we have

$$\begin{aligned}|R_{kj}(x,\tau)| =& |\{\tau + (1-2\tau) I[x_j \leqslant q_{jn}^{[k]}(\tau) + \triangle(\tau)]\} |x_j - q_{jn}^{[k]}(\tau) - \triangle(\tau)| \\ &- [\tau + (1-2\tau) I(x_j \leqslant q_{jn}^{[k]}(\tau))] |x_j - q_{jn}^{[k]}(\tau)|.\end{aligned}$$

Let $\rho_\tau(u) = u[\tau - I(u < 0)]$. We then have

$$|R_{kj}(x,\tau)| = |\rho_\tau[x_j - q_{jn}^{[k]}(\tau) - \triangle(\tau)] - \rho_\tau[x_j - q_j^{[k]}(\tau)]|.$$

Applying the following identity

$$\rho_\tau(u - \nu) - \rho_\tau(u) = -\nu\varphi_\tau(u) + \int_0^\nu [I(u \leqslant s) - I(u \leqslant 0)] ds,$$

where $\varphi_\tau(u) = \tau - I(u < 0)$, $\tau \in (0,1)$, we obtain

$$R_{kj}(x,\tau) = -\triangle(\tau)\varphi_\tau[x_j - q_{jn}^{[k]}(\tau)] + \int_0^{\triangle(\tau)} \big\{ [I(x_j - q_{jn}^{[k]}(\tau) \leqslant s)] - [I(x_j - q_{jn}^{[k]}(\tau) \leqslant 0)] \big\} ds.$$

Since $\varphi_\tau(\cdot)$ and the indicator function are bounded, we have

$$|R_{kj}(x,\tau)| \leqslant |\triangle(\tau)| + 2|\triangle(\tau)| = 3|\triangle(\tau)|.$$

Therefore, we only need to study the property of $\triangle(\tau)$. Since

$$\begin{aligned}
|\triangle(\tau)| &\leqslant |\hat{x}_j - x_j| + |\hat{q}_{jn}^{[k]}(\tau) - q_{jn}^{[k]}(\tau)| \\
&= \left| \int_{\mathcal{I}} \hat{x}(s)\hat{\psi}_j(s)ds - \int_{\mathcal{I}} x(s)\psi_j(s)ds \right| + |\hat{q}_{jn}^{[k]}(\tau) - q_{jn}^{[k]}(\tau)|,
\end{aligned}$$

$$\begin{aligned}
\hat{\xi}_{ij} - \xi_{ij} &= \int_{\mathcal{I}} \hat{X}_i(s)\hat{\psi}_j(s)ds - \int_{\mathcal{I}} X_i(s)\psi_j(s)ds \\
&= \int_{\mathcal{I}} \hat{X}_i(s)[\hat{\psi}_j(s) - \psi_j(s)]ds + \int_{\mathcal{I}} [\hat{X}_i(s) - X_i(s)]\psi_j(s)ds,
\end{aligned}$$

we have

$$\begin{aligned}
|\hat{\xi}_{ij} - \xi_{ij}| &\leqslant \left| \int_{\mathcal{I}} X_i(s)[\hat{\psi}_j(s) - \psi_j(s)]ds \right| + \left| \int_{\mathcal{I}} (\hat{X}_i(s) - X_i(s))[\hat{\psi}_j(s) - \psi_j(s)]ds \right| \\
&\quad + \left| \int_{\mathcal{I}} (\hat{X}_i(s) - X_i(s))[\psi_j(s)]ds \right| \\
&\leqslant \|X_i(s)\|\|\hat{\psi}_j - \psi_j\| + \|\hat{X}_i(s) - X_i\|\|\hat{\psi}_j - \psi_j\| + \|\hat{X}_i(s) - X_i(s)\|.
\end{aligned}$$

For fixed $\epsilon > 0$, set c such that $P(\|X\| > c) = P\{X \notin \mathcal{F}_0(c)\} \leqslant \frac{\epsilon}{2}$, where $\mathcal{F}_0(c) = \left\{X \middle| \|X\| \leqslant c\right\}$ for $c > 0$ with $\|\cdot\|$ being the L^2 norm. According to the Lemma 1 of Kong et al. (2016b), we know that

$$E[\|\hat{X}_i - X_i\|^2] = o(n^{-1}), \quad E\left\{ \int [\hat{X}_i(s) - X_i(s)]^4 ds \right\} = o(n^{-2}). \tag{6.22}$$

Thus, $\|\hat{X}_i - X_i\| = o_p(n^{-1/2})$, for $i = 1, \ldots, n$. By applying Corollary 3.7 of Li and Hsing (2010), we have $\|\hat{\psi}_j - \psi_j\| \overset{a.s.}{=} O(\sqrt{\frac{\log n}{n}})$.

To prove $\sup_{\tau\in\mathcal{S}}|\hat{\Psi}_n(J,\tau,\mathbf{w})-\Psi_n(J,\tau,\mathbf{w})|\xrightarrow{p}0$ as $n\to\infty$ and $J\to\infty$, for any $\mathbf{w}\in\mathcal{W}_J$, it is suffice to prove that

$$\sup_{\mathbf{w}\in\mathcal{W}_J}\sup_{\tau\in\mathcal{S}}\frac{1}{n}\sum_{i:Y_i=0}\left|I[\hat{Q}_J(X_i,\tau,\mathbf{w})>0]-I[\tilde{Q}_J(X_i,\tau,\mathbf{w})>0]\right|\xrightarrow{p}0,$$

and

$$\sup_{\mathbf{w}\in\mathcal{W}_J}\sup_{\tau\in\mathcal{S}}\frac{1}{n}\sum_{i:Y_i=1}\left|I[\hat{Q}_J(X_i,\tau,\mathbf{w})>0]-I[\tilde{Q}_J(X_i,\tau,\mathbf{w})>0]\right|\xrightarrow{p}0.$$

Thus, we only need to prove

$$\begin{aligned}&\sup_{\mathbf{w}\in\mathcal{W}_J}\frac{1}{n}\sum_{i:Y_i=0}\sup_{\tau\in\mathcal{S}}\left|I[\hat{Q}_J(X_i,\tau,\mathbf{w})>0]-I[\tilde{Q}_J(X_i,\tau,\mathbf{w})>0]\right|\xrightarrow{p}0,\\&\sup_{\mathbf{w}\in\mathcal{W}_J}\frac{1}{n}\sum_{i:Y_i=1}\sup_{\tau\in\mathcal{S}}\left|I[\hat{Q}_J(X_i,\tau,\mathbf{w})>0]-I[\tilde{Q}_J(X_i,\tau,\mathbf{w})>0]\right|\xrightarrow{p}0.\end{aligned}\tag{6.23}$$

According to the aforementioned results, we have

$$\hat{Q}_J(X_i,\tau,\mathbf{w})-\tilde{Q}_J(X_i,\tau,\mathbf{w})=\sum_{j=1}^{J}w_j[R_{1j}(X_i,\tau)-R_{0j}(X_i,\tau)],$$

$$|R_{kj}(X_i,\tau))|\leqslant 3|\triangle_i(\tau)|,$$

$$|\triangle_i(\tau)|\leqslant\|X_i\|\|\hat{\psi}_j-\psi_j\|+\|\hat{X}_i-X_i\|\|\hat{\psi}_j-\psi_j\|+\|\hat{X}_i-X_i\|+|\hat{q}_{jn}^{[k]}(\tau)-q_{jn}^{[k]}(\tau)|,$$

where

$$\begin{aligned}R_{kj}(x,\tau)&=[\tau+(1-2\tau)I(\hat{x}_j\leqslant\hat{q}_{jn}^{[k]}(\tau))]|\hat{x}_j-\hat{q}_{jn}^{[k]}(\tau)|\\&\quad-[\tau+(1-2\tau)I(x_j\leqslant q_{jn}^{[k]}(\tau))]|x_j-q_{jn}^{[k]}(\tau)|\\&=[\tau+(1-2\tau)I(x_j\leqslant q_{jn}^{[k]}(\tau)-(\hat{x}_j-x_j))+\hat{q}_{jn}^{[k]}(\tau)-q_{jn}^{[k]}(\tau)]\\&\quad|x_j-q_{jn}^{[k]}(\tau)+(\hat{x}_j-x_j)-(\hat{q}_{jn}^{[k]}(\tau)-q_{jn}^{[k]}(\tau))|\\&\quad-[\tau+(1-2\tau)I(x_j\leqslant q_{jn}^{[k]}(\tau))]|x_j-q_{jn}^{[k]}(\tau)|.\end{aligned}$$

Since $\triangle(X,\tau) \equiv \triangle(\tau) = -(\hat{x}_j - x_j) + [\hat{q}_{jn}^{[k]}(\tau) - q_{jn}^{[k]}(\tau)]$, $|R_{kj}(X,\tau)| \leqslant 3|\triangle(\tau)|$, and

$$|\triangle(\tau)| \leqslant |\hat{x}_j - x_j| + |\hat{q}_{jn}^{[k]}(\tau) - q_{jn}^{[k]}(\tau)|,$$

we have

$$\begin{aligned} R_{kj}(X_i,\tau) &\leqslant 3|\triangle_i(\tau)| \equiv 3|\triangle(\tau)| \leqslant 3\big\{|\hat{\xi}_{ij} - \xi_{ij}| + |\hat{q}_{jn}^{[k]}(\tau) - q_{jn}^{[k]}(\tau)|\big\} \\ &= 3\big\{\|X_i\|\|\hat{\psi}_j - \psi_j\| + \|\hat{X}_i - X_i\|\|\hat{\psi}_j - \psi_j\| + \|\hat{X}_i - X_i\| + |\hat{q}_{jn}^{[k]}(\tau) - q_{jn}^{[k]}(\tau)|\big\}. \end{aligned}$$

For (6.23), we have

$$\begin{aligned} &\frac{1}{n}\sum_{i:Y_i=0}\sup_{\tau\in\mathcal{S}}\Big|I[\hat{Q}_J(X_i,\tau,\mathbf{w}) > 0] - I[\tilde{Q}_J(X_i,\tau,\mathbf{w}) > 0]\Big| \\ =&\frac{1}{n}\sum_{i:Y_i=0}\sup_{\tau\in\mathcal{S}}\Big|I[\tilde{Q}_J(X_i,\tau,\mathbf{w}) > -\{\hat{Q}_J(X_i,\tau,\mathbf{w}) - \tilde{Q}_J(X_i,\tau,\mathbf{w})\}] - I[\tilde{Q}_J(X_i,\tau,\mathbf{w}) > 0]\Big|. \end{aligned}$$

Thus, in order to prove the equation (6.23), it is sufficient to show that

$$\sup_{\mathbf{w}\in\mathcal{W}_J}\sup_{\tau\in\mathcal{S}}|\hat{Q}_J(X_i,\tau,\mathbf{w}) - \tilde{Q}_J(X_i,\tau,\mathbf{w})| \xrightarrow{p} 0, \tag{6.24}$$

as $n \to \infty$ and $J \to \infty$. For any $\mathbf{w} \in \mathcal{W}_J$, we have

$$\begin{aligned} &\sup_{\tau\in\mathcal{S}}|\hat{Q}_J(X_i,\tau,\mathbf{w}) - \tilde{Q}_J(X_i,\tau,\mathbf{w})| \\ =& \sum_{j=1}^{J} w_j \sup_{\tau\in\mathcal{S}}[|R_{1j}(X_i,\tau)| + |R_{0j}(X_i,\tau)|] \\ =& \sum_{j=1}^{J} w_j \sup_{\tau\in\mathcal{S}}|R_{1j}(X_i,\tau)| + \sum_{j=1}^{J} w_j \sup_{\tau\in\mathcal{S}}|R_{0j}(X_i,\tau)| \qquad (6.25) \\ =& \sum_{k=0,1}\Big\{\sum_{j=1}^{J} w_j \sup_{\tau\in\mathcal{S}}|R_{kj}(X_i,\tau)|\Big\} \\ \leqslant& C\sum_{k=0,1}\Big\{(\sum_{j=1}^{J} w_j\|\hat{\psi}_j - \psi_j\|)\|X_i\| + \|\hat{X}_i - X_i\|\sum_{j=1}^{J} w_j\|\hat{\psi}_j - \psi_j\| \end{aligned}$$

$$+\sum_{j=1}^{J} w_j(\|\hat{X}_i - X_i\|) + \sum_{j=1}^{J} \sup_{\tau\in\mathcal{S}} |\hat{q}_{jn}^{[k]}(\tau) - q_{jn}^{[k]}(\tau)|\Bigg\}$$

$$= C\|X_i\|\left(\sum_{j=1}^{J} w_j\|\hat{\psi}_j - \psi_j\|\right) + C\|\hat{X}_i - X_i\|\left(\sum_{j=1}^{J} w_j\|\hat{\psi}_j - \psi_j\|\right)$$

$$+C\|\hat{X}_i - X_i\|\left(\sum_{j=1}^{J} w_j\right) + C\sum_{j=1}^{J} \sup_{\tau\in\mathcal{S}} |\hat{q}_{jn}^{[k]}(\tau) - q_{jn}^{[k]}(\tau)|,$$

For the first three terms in (6.25), it follows from (6.22) and Theorem 3.6 in Li and Hsing (2010) that the first three terms in (6.25) are $o_p(1)$ as $J \to \infty$ and $n \to \infty$. Thus, if we want to derive (6.24), we need to prove that $\sum_{j=1}^{J} \sup_{\tau\in\mathcal{S}} |\hat{q}_{jn}^{[k]}(\tau) - q_{jn}^{[k]}(\tau)| \xrightarrow{p} 0$ as $J \to \infty$. Note that $\hat{q}_{jn}^{[k]}(\tau) = \inf\{x|\hat{F}_{nj}(x) \geqslant \tau\}$ and $q_{jn}^{[k]}(\tau) = \inf\{x|F_{nj}(x) \geqslant \tau\}$, where $\hat{F}_{nj}(x) = n^{-1}\sum_{i=1}^{n} I(\hat{\xi}_{ij} \leqslant x)$, $F_{nj}(x) = n^{-1}\sum_{i=1}^{n} I(\xi_{ij} \leqslant x)$, and $F_{\xi_j^{[k]}}(u) := F_j^{[k]}(u)$. So, we have

$$\sup_{x\in\mathbb{R}} |\hat{F}_{nj}(x) - F_{nj}(x)| = \sup_{x\in\mathbb{R}} \left| n^{-1}\sum_{i=1}^{n} [I(\hat{\xi}_{ij} \leqslant x) - I(\xi_{ij} \leqslant x)] \right|$$

$$= \sup_{x\in\mathbb{R}} n^{-1}\sum_{i=1}^{n} |I[x - (\hat{\xi}_{ij} - \xi_{ij}) \leqslant \xi_{ij} \leqslant x]|$$

$$= n^{-1}\sum_{i=1}^{n} \sup_{x\in\mathbb{R}} |I[x - (\hat{\xi}_{ij} - \xi_{ij}) \leqslant \xi_{ij} \leqslant x]|.$$

Denote $\pi_{ij} = \hat{\xi}_{ij} - \xi_{ij}$, $|\pi_{ij}| = |\hat{\xi}_{ij} - \xi_{ij}| \leqslant \|X_i\|\|\hat{\psi}_j - \psi_j\| + \|\hat{X}_i - X_i\|\|\hat{\psi}_j - \psi_j\| + \|\hat{X}_i - X_i\|$. Then we have

$$\sup_{x\in\mathbb{R}} |\hat{F}_{nj}(x) - F_{nj}(x)| = \sup_{x\in\mathbb{R}} \left| n^{-1}\sum_{i=1}^{n} I[x - \pi_{ij} \leqslant \xi_{ij} \leqslant x] \right|.$$

Denote $\mathcal{F}_i(c) = \{\|X_i\| \leqslant c, \sqrt{n}\|\hat{X}_i - X_i\| \leqslant c\}$ for $c > 0$. On one hand, letting $\mathcal{F}(c) = \mathcal{F}_1(c) \cap \mathcal{F}_2(c) \cap \ldots \cap \mathcal{F}_n(c)$, we have $\|\pi_{ij}\| \leqslant c\|\hat{\psi}_j - \psi_j\| + c/\sqrt{n}$ on $\mathcal{F}(c)$. On

the other hand, by applying Theorem 3.6 in Li and Hsing (2010), we get $\|\hat{\psi}_j - \psi_j\| \overset{a.s.}{=} O\left(\sqrt{\frac{\log n}{n}}\right)$. Thus, we have $|\pi_{ij}| \overset{a.s.}{=} O\left(\sqrt{\frac{\log n}{n}}\right)$ on $\mathcal{F}(c)$.

Similar to the proof of Glivenko's theorem, we have $\sup_{x\in\mathbb{R}} |\hat{F}_{nj}(x) - F_{nj}(x)| \overset{a.s.}{=} O\left(\sqrt{\frac{\log n}{n}}\right)$, $n^{-1}\sum_{i=1}^{n} I(x - \pi_{ij} \leqslant \xi_{ij} \leqslant x) \overset{a.s.}{\to} cf_j(x)\sqrt{\frac{\log n}{n}}$ on $\mathcal{F}(c)$ as $n \to \infty$, where $f_j(u)$ denotes the density function of the projection score ξ_j. Assume that $\sup_{j\geqslant 1}\sup_{u\in\mathbb{R}} |f_j(u)| < \infty$. Then one can obtain

$$\sup_{x\in\mathbb{R}} |\hat{F}_{nj}(x) - F_{nj}(x)| \overset{a.s.}{=} O\left(\sqrt{\frac{\log n}{n}}\right). \tag{6.26}$$

Thus, we have $\sum_{j=1}^{J}\sup_{\tau\in T} |\hat{q}_{jn}^{[k]}(\tau) - q_{jn}^{[k]}(\tau)| = O(J\sqrt{\frac{\log n}{n}})$ as $n \to \infty$ and $J \to \infty$. From the fact X is a square integrable random function and (6.22), it holds that $P\{\mathcal{F}(c)\} \to 1$ as $c \to \infty$. We then obtain $I_{1n} \xrightarrow{p} 0$, and

$$\sup_{\tau\in\mathcal{S}} |\hat{\Psi}_n(J, \tau, \mathbf{w}) - \Psi(J, \tau, \mathbf{w})| \xrightarrow{p} 0, \ as \ J \to \infty, \ n \to \infty.$$

Thus, we have $\Psi(J, \hat{\tau}_n, \mathbf{w}) - \Psi(J, \tau_0, \mathbf{w}) \xrightarrow{p} 0$, as $n \to \infty$ and $J \to \infty$. We complete the proof of Lemma 1.

6.8.3 Proofs of Main Results

Proof of Theorem 6.1. Define

$$L_j(x, \tau, q) \equiv [\tau + (1 - 2\tau)I(x_j \leqslant q)]|x_j - q|,$$

and abbreviate $L_j\{x, \tau, q_j^{[k]}(\tau)\}$ as $L_{kj}(x, \tau)$. It follows from Lemma 3 in Henning and Viroli (2016) that

$$|L_j(x, \tau_1, q_1) - L_j(x, \tau_2, q_2)| \leqslant |x_j||\tau_1 - \tau_2| + 4|q_1 - q_2|, \tag{6.27}$$

where $q_1 \equiv q_1(\tau_1)$ and $q_2 \equiv q_2(\tau_2)$ for $j = 1, \ldots, J$ and $x \in \mathcal{L}^2(\mathcal{I})$. From (6.27), we know that $L_{kj}(x, \tau)$ is a continuous function with respect to τ, implying the continuity of $\Psi(J, \tau, \mathbf{w})$ with respect to τ. Then the convergence of the integrals of indicator functions within Ψ can be guaranteed by the dominated convergence theorem for $\hat{\tau}_n \to \tau$, $\tau \in \mathcal{S}$. For any $J \in \mathbb{Z}$ and $\mathbf{w} \in W_J$,

$$
\begin{aligned}
&|\Psi(J, \hat{\tau}_n, \mathbf{w}) - \Psi(J, \tau_0, \mathbf{w})| \\
\leqslant &|\Psi(J, \hat{\tau}_n, \mathbf{w}) - \hat{\Psi}_n(J, \hat{\tau}_n, \mathbf{w})| + |\hat{\Psi}_n(J, \tau_0, \mathbf{w}) - \Psi(J, \tau_0, \mathbf{w})| + |\hat{\Psi}_n(J, \hat{\tau}_n, \mathbf{w}) - \hat{\Psi}_n(J, \tau_0, \mathbf{w})| \\
\equiv &I_{1n} + I_{2n} + I_{3n}.
\end{aligned}
$$

By applying Lemma 6.1, for any $\mathbf{w} \in W_J$, we have $I_{1n} \xrightarrow{p} 0$ and $I_{2n} \xrightarrow{p} 0$ as $n \to \infty$, $J \to \infty$. For the term I_{3n}, by definition, we have $\hat{\Psi}_n(J, \hat{\tau}_n, \mathbf{w}) \geqslant \hat{\Psi}_n(J, \tau_0, \mathbf{w})$, $\Psi(J, \tau_0, \mathbf{w}) \geqslant \Psi(J, \hat{\tau}_n, \mathbf{w})$. Thus, it follows that

$$
\begin{aligned}
&\hat{\Psi}_n(J, \hat{\tau}_n, \mathbf{w}) - \hat{\Psi}_n(J, \tau_0, \mathbf{w}) \\
= &\hat{\Psi}_n(J, \hat{\tau}_n, \mathbf{w}) - \Psi(J, \tau_0, \mathbf{w}) + \Psi(J, \tau_0, \mathbf{w}) - \hat{\Psi}_n(J, \tau_0, \mathbf{w}) \\
\equiv &I_{31n} + I_{32n} \leqslant |I_{31n}| + |I_{32n}|.
\end{aligned}
$$

By applying Lemma 6.1 again, for any $\mathbf{w} \in W_J$, it is easy to show that $|I_{32n}| \xrightarrow{p} 0$ as $n \to \infty$ and $J \to \infty$. As for the term I_{31n}, note that

$$
\hat{\Psi}_n(J, \tau_0, \mathbf{w}) - \Psi(J, \tau_0, \mathbf{w}) \leqslant \hat{\Psi}_n(J, \hat{\tau}_n, \mathbf{w}) - \Psi(J, \tau_0, \mathbf{w}) < \hat{\Psi}_n(J, \hat{\tau}_n, \mathbf{w}) - \Psi(J, \hat{\tau}_n, \mathbf{w}).
$$

From Lemma 6.1 and the Sandwich theorem, we have $I_{31n} \xrightarrow{p} 0$ as $n \to \infty$ and $J \to \infty$. Thus, for any $\epsilon > 0$, we can conclude that

$$
\inf_{\mathbf{w} \in \mathcal{W}_J} P\left\{ |\Psi(J, \hat{\tau}_n, \mathbf{w}) - \Psi(J, \tau_0, \mathbf{w})| \leqslant \epsilon \right\} \to 1, \text{ as } n \to \infty, \ J \to \infty.
$$

This completes the proof of Theorem 1.

Proof of Theorem 6.2. To prove the asymptotic result of Theorem 2, it is sufficient to show that

$$\inf_{\tau\in\mathcal{S}}\inf_{\mathbf{w}\in\mathcal{W}_J}\left[P_{\Pi_0}\{\hat{Q}_J(X,\tau,\mathbf{w})>0\}\to 1,\ as\ n\to\infty, J\to\infty,\right.$$

$$\left.\inf_{\tau\in\mathcal{S}}\inf_{\mathbf{w}\in\mathcal{W}_J}\left[P_{\Pi_1}\{\hat{Q}_J(X,\tau,\mathbf{w})\leqslant 0\}\to 1,\ as\ n\to\infty, J\to\infty.\right.\right. \tag{6.28}$$

Note that $\hat{Q}_J(X,\tau,\mathbf{w}) = \tilde{Q}_J(X,\tau,\mathbf{w}) + \{\hat{Q}_J(X,\tau,\mathbf{w}) - \tilde{Q}_J(X,\tau,\mathbf{w})\}$. By applying Assumptions 6.7-6.12 and using arguments similar to those in Henning and Viroli (2016) and Hall et al. (2009), one can show that

$$\inf_{\tau\in\mathcal{S}}\inf_{\mathbf{w}\in\mathcal{W}_J}\left[P_{\Pi_0}\{\tilde{Q}_J(X,\tau,\mathbf{w})>0\}\to 1,\ as\ n\to\infty, J\to\infty,\right.$$

and

$$\inf_{\tau\in\mathcal{S}}\inf_{\mathbf{w}\in\mathcal{W}_J}\left[P_{\Pi_1}\{\tilde{Q}_J(X,\tau,\mathbf{w})\leqslant 0\}\to 1,\ as\ n\to\infty, J\to\infty.\right.$$

Thus, to prove (6.28), we only need to prove that, for any $\epsilon>0$,

$$\sup_{\tau\in\mathcal{S}}\sup_{\mathbf{w}\in\mathcal{W}_J}|\hat{Q}_J(X,\tau,\mathbf{w}) - \tilde{Q}_J(X,\tau,\mathbf{w})| \xrightarrow{p} 0,\ as\ n\to\infty, J\to\infty. \tag{6.29}$$

Then it follows from (6.25) that (6.29) holds. Hence, the proof is complete.

6.9 An R Package and Shiny App for Quantiles-Based Classifier for Functional Data

An R package quicfun (QUantIles-based Classifier for FUNctional data) is available at https://github.com/iantsuising/quickfun. For easy implementation and tuning, we also develop a Shiny App, a web-based interface for practitioners or non-R users. The App is available at https://ianxu.shinyapps.io/quicfun/. The source of the web interface is available at https://github.com/iantsuising/quickfun-app.

Chapter 7

Future Work

In Part I, we study the single-index hazard model for survival outcomes. In modern biomedical and GAWS studies, for example, breast cancer studies, the outcomes could be recurrent. The covariates could be divided into several groups based on prior information, for example, a group of demographical factors, a group of genetic variates, a group of clinical assessments, among others. The complexity of both response and covariates motivates us to extend the single-index hazard model to a multiple-index hazard model by developing a sufficient dimension reduction paradigm for counting process data. The proposed single-index hazard model could be regarded as a special case of the more general multiple-index hazard model where the number of the occurrence of the event is less or equal one and the number of the group of the covariate is one.

In Part II, the applied methods are limited to sparse functional data where the number of repeated measure m on each curve is bounded by a fixed constant. In model biomedical studies, for example, functional magnetic resonance imaging studies, the data could be treated as dense and regular observed functional data. For dense functional data, there has been some literature on nonparametric tests including Zhang and Chen (2007) and Wang et al. (2018), which are also based on the working independence principle. Taking into account the within-subject covari-

ance can potentially improve the power of these tests as well. However, extending our methodology to dense functional data may encounter some technical difficulties, since the SU-based method requires inverting the within-subject covariance matrix, which becomes a high dimensional random matrix if m goes to infinity. One possible solution is to reduce the rank of the covariance matrix by functional principal component analysis. This is an important problem that calls for future research.

In Part III, we consider the classification problem for the subject with only one functional biomarker. In real-world applications, the number of functional predictors could be larger than one. For example, one may obtain several DTI functional predictors from different brain regions of each subject. In this case, combining the functional biomarkers can naturally increase the diagnosis accuracy. One of our future work is to extend the proposed quantile-based functional classifier to a quantile-based combination of functional predictors to improve diagnosis accuracy.

Bibliography

Alonso, A. M., Casado, D. and Romo, J. (2012) Supervised classification for functional data: A weighted distance approach. *Computational Statistics and Data Analysis*, **56**, 2334–2346.

Andersen, P. K. and Gill, R. D. (1982) Cox's regression model for counting processes: a large sample study. *The Annals of Statistics*, 1100–1120.

Andersson, C., Lindau, M., Almkvist, O., Engfeldt, P., Johansson, S.-E. and Jönhagen, M. E. (2006) Identifying patients at high and low risk of cognitive decline using rey auditory verbal learning test among middle-aged memory clinic outpatients. *Dementia and Geriatric Cognitive Disorders*, **21**, 251–259.

Araki, Y., Konishi, S., Kawano, S. and Matsui, H. (2009) Functional logistic discrimination via regularized basis expansions. *Communications in Statistics – Theory and Methods*, **38**, 2944–2957.

Barzilai, J. and Borwein, J. M. (1988) Two-point step size gradient methods. *IMA Journal of Numerical Analysis*, **8**, 141–148.

Benner, A., Zucknick, M., Hielscher, T., Ittrich, C. and Mansmann, U. (2010) High-dimensional cox models: the choice of penalty as part of the model building process. *Biometrical Journal*, **52**, 50–69.

Berrendero, J. R., Cuevas, A. and Torrecilla, J. L. (2018) On the use of reproducing kernel hilbert spaces in functional classification. *Journal of the American Statistical Association*, **113**, 1210–1218.

Bickel, P., Klaassen, C., Ritov, Y. and Wellner, J. (1997) *Testing the Equality of Covariance Operators*. Baltimore: Johns Hopkins University Press.

Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009) Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, **37**, 1705–1732.

Bradic, J., Fan, J. and Jiang, J. (2011) Regularization for cox's proportional hazards model with np-dimensionality. *The Annals of Statistics*, **39**, 3092–3120.

Breslow, N. (1974) Covariance analysis of censored survival data. *Biometrics*, 89–99.

Brumback, B. and Rice, J. A. (1998) Smoothing spline models for the analysis of nested and crossed samples of curves (with discussion). *Journal of the American Statistical Association*, **93**, 961–994.

Cai, J., Fan, J., Jiang, J. and Zhou, H. (2008) Partially linear hazard regression with varying coefficients for multivariate survival data. *Journal of Royal Statistical Society, Series B*, **70**, 141–158.

Cao, W., Tsiatis, A. and Davidian, M. (2009) Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika*, **96(3)**, 723–734.

Carroll, R., Fan, J., Gijbels, I. and Wand, M. (1997) Generalized partially linear single-index models. *Journal of the American Statistical Association*, **92**, 477–489.

Chen, A. Y., Chonghasawat, A. O. and Leadholm, K. L. (2017) Multiple sclerosis: Frequency, cost, and economic burden in the United States. *Journal of Clinical Neuroscience*, **45**, 180–186.

Chen, X., Lu, Z. and Pong, T. K. (2016) Penalty methods for a class of non-lipschitz optimization problems. *SIAM Journal on Optimization*, **26**, 1465–1492.

Chen, X., Xu, S. and Liu, C. (2020+) An efficient algorithm for joint feature screening in ultrahigh-dimensional Cox's model. *Technical Report*.

Chen, Y. and Wang, M. (2000) Analysis of accelerated hazards models. *Journal of the American Statistical Association*, **95**, 349–372.

Chiang, C.-T., Wang, S.-H. and Huang, M.-Y. (2017) Versatile estimation in censored single-index hazards regression. *Annals of the Institute of Statistical Mathematics*, 1–29.

Cox, D. (1972) Regression models and life tables (with discussion). *Journal of Royal Statistical Society, Series B*, **34**, 187–220.

Cox, D. and Oakes, D. (1984) *Analysis of Survival Data*. London: Chapman and Hall.

Cox, D. R. (1975) Partial likelihood. *Biometrika*, **62**, 269–276.

Csörgo, M. and Révész, P. (1981) Strong approximation in probability and statistics. *Academic Press*, New York.

Cui, X., Härdle, W. and Zhu, L. (2011) The efm approach for single-index models. *The Annals of Statistics*, **39**, 1658–1688.

Dai, X., Müller, H. and Yao, F. (2017) Optimal bayes classifiers for functional data and density ratios. *Biometrika*, **104**, 545–560.

Delaigle, A. and Hall, P. (2012) Achieving near perfect classification for functional data. *The Journal of the Royal Statistical Society, Series B*, **74**, 267–286.

Delaigle, A. and Hall, P. (2013) Classification using censored functional data. *Journal of the American Statistical Association*, **108**, 1269–1283.

Dilokthornsakul, P., Valuck, R. J., Nair, K. V., Corboy, J. R., Allen, R. R. and Campbell, J. D. (2016) Multiple sclerosis prevalence in the united states commercially insured population. *Neurology*, **86**, 1014–1021.

Ding, K., Kosorok, M. and Zeng, D. (2013) On the local and stratified likelihood approaches in single-index hazards model. *Communications in Mathematics and Statistics*, **1**, 115–132.

Dong, C., Gao, J. and Peng, B. (2015) Semiparametric single-index panel data models with cross-sectional dependence. *Journal of Econometrics*, **188**, 301–312.

Fadista, J., Manning, A. K., Florez, J. C. and Groop, L. (2016) The (in) famous GWAS P-value threshold revisited and updated for low-frequency variants. *European Journal of Human Genetics*, **24**, 1202.

Fan, J., Feng, Y. and Wu, Y. (2010) High-dimensional variable selection for cox's proportional hazards model. In *Borrowing Strength: Theory Powering Applications–A Festschrift for Lawrence D. Brown*, 70–86. Institute of Mathematical Statistics.

Fan, J. and Gijbels, I. (1996) *Local Polynomial Modelling and Its Applications*. Chapman & Hall/CRC.

Fan, J., Gijbels, I., King, M. et al. (1997) Local likelihood and local partial likelihood in hazard regression. *The Annals of Statistics*, **25**, 1661–1690.

Fan, J. and Li, R. (2004) New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. *Journal of the American Statistical Association*, **99**, 710–723.

Fan, J. and Lv, J. (2008) Sure independence screening for ultrahigh dimensional feature space. *The Journal of the Royal Statistical Society, Series B*, **70**, 849–911.

Fan, J. and Song, R. (2010) Sure independence screening in generalized linear models with np-dimensionality. *The Annals of Statistics*, **38**, 3567–3604.

Fan, J., Zhang, C. and Zhang, J. (2001) Generalized likelihood ratio statistics and Wilks phenomenon. *The Annals of Statistics*, **29**, 153–193.

Ferencz, B., Laukka, E. J., Lövdén, M., Kalpouzos, G., Keller, L., Graff, C., Wahlund, L.-O., Fratiglioni, L. and Bäckman, L. (2013) The influence of APOE and TOMM40 polymorphisms on hippocampal volume and episodic memory in old age. *Frontiers in Human Neuroscience*, **7**, 198.

Filippi, M., Pagani, E., Preziosa, P. and Rocca, M. A. (2016) The role of DTI in multiple sclerosis and other demyelinating conditions. In *Diffusion Tensor Imaging*, 331–341. Springer.

Fleming, T. R. and Harrington, D. P. (2011) *Counting processes and survival analysis*, vol. 169. John Wiley & Sons.

Gasparoni, G., Bultmann, S., Lutsik, P., Kraus, T. F., Sordon, S., Vlcek, J., Dietinger, V., Steinmaurer, M., Haider, M., Mulholland, C. B. et al. (2018) Dna methylation analysis on purified neurons and glia dissects age and alzheimer's disease-specific changes in the human cortex. *Epigenetics & Chromatin*, **11**, 41.

Gertheiss, J., Maity, A. and Staicu, A.-M. (2013) Variable selection in generalized functional linear models. *Stat*, **2**, 86–101.

Goldsmith, J., Crainiceanu, C. M., Caffo, B. and Reich, D. (2012) Longitudinal penalized functional regression for cognitive outcomes on neuronal tract measurements. *The Journal of the Royal Statistical Society, Series C*, **61**, 453–469.

Goldsmith, J., Crainiceanu, C. M., Caffo, B. S. and Reich, D. S. (2011) Penalized functional regression analysis of white-matter tract profiles in multiple sclerosis. *NeuroImage*, **57**, 431–439.

Gong, P., Zhang, C., Lu, Z., Huang, J. and Ye, J. (2013) A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems. In *International Conference on Machine Learning*, 37–45.

Gørgens, T. (2004) Average derivatives for hazard functions. *Econometric Theory*, **20**, 437–463.

Gørgens, T. (2006) Semiparametric estimation of single-index hazard functions without proportional hazards. *The Econometrics Journal*, **9**, 1–22.

Gorst-Rasmussen, A. and Scheike, T. (2013) Independent screening for single-index hazard rate models with ultrahigh dimensional features. *The Journal of the Royal Statistical Society, Series B*, **75**, 217–245.

Grupe, A., Abraham, R., Li, Y., Rowland, C., Hollingworth, P., Morgan, A., Jehu, L., Segurado, R., Stone, D. and Schadt, E. (2007) Evidence for novel susceptibility genes for late-onset alzheimer's disease from a genome-wide association study of putative functional variants. *Human Molecular Genetics*, **16**, 865–873.

Hall, P., Müller, H. G. and Wang, J. L. (2006) Properties of principal component methods for functional and longitudinal data analysis. *The Annals of Statistics*, **34**, 1493–1517.

Hall, P., Poskitt, D. S. and Presnell, B. (2001) A functional data analytic approach to signal discrimination. *Technometrics*, **43**, 1–9.

Hall, P., Titteringon, D. M. and Xue, J. H. (2009) Median-based classifiers for high-dimensional data. *Journal of the American Statistical Association*, **104**, 1597–1608.

Härdle, W., Müller, M., Sperlich, S. and Werwatz, A. (2004) *Nonparametric and semiparametric models*. Springer Verlag, Heidelberg.

He, K., Zhang, X., Ren, S. and Sun, J. (2016) Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

He, Z., Zhang, M., Lee, S., Smith, J. A., Guo, X., Palmas, W., Kardia, S. L., Roux, A. V. D. and Mukherjee, B. (2015) Set-based tests for genetic association in longitudinal studies. *Biometrics*, **71**, 606–615.

Henning, C. and Viroli, C. (2016) Quantile-based classifiers. *Biometrika*, **103**, 435–446.

Huang, C., Thompson, P., Wang, Y., Yu, Y., Zhang, J., Kong, D., Colen, R. R., Knickmeyer, R. C., Zhu, H. and Initiative, T. A. D. N. (2017) FGWAS: Functional genome wide association analysis. *Neuroimage*, **159**, 107–121.

Huang, J., Sun, T., Ying, Z., Yu, Y. and Zhang, C.-H. (2013) Oracle inequalities for the lasso in the cox model. *The Annals of Statistics*, **41**, 1142.

Ichimura, H. (1993) Semiparametric least squares (sls) and weighted sls estimation of single-index models. *Journal of Economic*, **58**, 71–120.

Ivanescu, A. E., Staicu, A.-M., Scheipl, F. and Greven, S. (2015) Penalized function-on-function regression. *Computational Statistics*, **30**, 539–568.

Kong, D., Staicu, A.-M. and Maity, A. (2016a) Classical testing in functional linear models. *Journal of Nonparametric Statistics*, **28**, 813–838.

Kong, D., Xue, K., Yao, F. and Zhang, H. H. (2016b) Partially functional linear regression in high dimensions. *Biometrika*, **103**, 147–159.

Kraus, D. and Stefanucci, M. (2019) Classification of functional fragments by regularized linear classifiers with domain selection. *Biometrika*, **106**, 161–180.

van der Laan, M. J. and Robins, J. M. (2003) *Unified methods for censored longitudinal data and causality*. Springer Science & Business Media.

Li, H., Keadle, S. K., J., S., Assaad, H., Huang, J. Z. and Carroll, R. J. (2015) Methods to assess an exercise intervention trial based on 3-level functional data. *Biostatistics*, **16**, 754–771.

Li, Y. (2011) Efficient semiparametric regression for longitudinal data with nonparametric covariance estimation. *Biometrika*, **98**, 355–370.

Li, Y. and Hsing, T. (2010) Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data. *The Annals of Statistics*, **38**, 3321–3351.

Li, Y., Xu, S. and Liu, C. (2020+) Functional data modeling and hypothesis testing for longitudinal alzheimer genome-wide association studies. In *Springer Book on New Frontiers of Biostatistics and Bioinformatics Research*, To appear. Springer.

Liang, H., Liu, X., Li, R. and Tsai, C.-L. (2010) Estimation and testing for partially linear single-index models. *The Annals of Statistics*, **38**, 3811–3836.

Lin, D. and Ying, Z. (1994) Semiparametric analysis of the additive risk model. *Biometrika*, **81**, 61–71.

Lin, X. (2007) Estimation using penalized quasilikelihood and quasi-pseudo-likelihood in poisson mixed models. *Lifetime Data Analysis*, **13**, 533–544.

Lin, X., Cai, T., Wu, M. C., Zhou, Q., Liu, G., Christiani, D. C. and Lin, X. (2011) Kernel machine snp-set analysis for censored survival outcomes in genome-wide association studies. *Genetic epidemiology*, **35**, 620–631.

Lin, X. and Carroll, R. J. (2001) Semiparametric regression for clustered data using generalized estimating equations. *Journal of the American Statistical Association*, **96**, 1045–1056.

Linton, O. B., Nielsen, J. P. and van de Geer, S. (2003) Estimating multiplicative and additive hazard functions by kernel methods. *The Annals of Statistics*, **31**, 464–492.

Liu, J., Xu, S., Liu, C. and K.C., Y. (2020+) Estimation under single-index hazard model. *Technical Report*.

Liu, J., Zhang, R., Zhao, W. and Lv, Y. (2013) A robust and efficient estimation method for single index models. *Journal of Multivariate Analysis*, **122**, 226–238.

Lu, W. and Zhang, H. H. (2010) On estimation of partially linear transformation models. *Journal of the American Statistical Association*, **105**, 683–691.

Ma, H., Xu, S., Liu, C. and C., Y. K. (2020+) Weighted multiple-quantile classifiers for functional data. *Technical Report*.

Ma, S. and Song, P. X.-K. (2015) Varying index coefficient models. *Journal of the American Statistical Association*, **110**, 341–356.

Ma, Y. and Zhu, L. (2012) A semiparametric approach to dimension reduction. *Journal of the American Statistical Association*, **107**, 168–179.

Mammen, E. (1993) Bootstrap and wild bootstrap for high dimensional linear models. *The Annals of Statistics*, **21**, 255–285.

Mammen, E., Martínez-Miranda, M., Nielsen, J. and Sperlich, S. (2011) Do-validation for kernel density estimation. *Journal of the American Statistical Association*, **106**, 651–660.

Mammen, E. and Nielsen, J. (2007) A general approach to the predictability issue in survival analysis with applications. *Biometrika*, **94**, 873–892.

Mason, D. M. (1982) Some characterizations of almost sure bounds for weighted multidimensional empirical distributions and a glivenkocantelli theorem for sample quantiles. *Probability Theory and Related Fields*, **59**, 505–513.

McLean, M. W., Hooker, G., Staicu, A.-M., Scheipl, F. and Ruppert, D. (2014) Functional generalized additive models. *The Annals of Statistics*, **23**, 249–269.

Meinshausen, N. and Bühlmann, P. (2006) High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 1436–1462.

Metzeler, K. H., Hummel, M., Bloomfield, C. D., Spiekermann, K., Braess, J., Sauerland, M.-C., Heinecke, A., Radmacher, M., Marcucci, G., Whitman, S. P. et al. (2008) An 86-probe-set gene-expression signature predicts survival in cytogenetically normal acute myeloid leukemia. *Blood*, **112**, 4193–4201.

Miller, J. R. (2004) The importance of early diagnosis of multiple sclerosis. *Journal of Managed Care & Specialty Pharmacy*, **10**, 4–11.

Morris, J. S. (2015) Functional regression. *The Annual Review of Statistics and Its Application*, **2**, 321–359.

Morris, J. S. and Carroll, R. J. (2006) Wavelet-based functional mixed models. *Journal of the Royal Statistical Society, Series B*, **68**, 179–199.

National Multiple Sclerosis Society (2017) Preliminary results of ms prevalence study estimate nearly 1 million living with ms in the U.S. URL www.nationalmssociety.org/About-the-Society/News/Preliminary-Results-of-MS-Prevalence-Study.

Newey, W. K. (1990) Semiparametric efficiency bounds. *Journal of applied econometrics*, **5**, 99–135.

Nielsen, J. (1998) Marker dependent kernel hazard estimation from local linear estimation. *Scandinavian Actuarial Journal*, **2**, 113–124.

Nielsen, J. and Linton, O. (1995) Kernel estimation in a nonparametric marker dependent hazard model. *The Annals of Statistics*, **23**, 1735–1748.

Oberthuer, A., Berthold, F., Warnat, P., Hero, B., Kahlert, Y., Spitz, R., Ernestus, K., Konig, R., Haas, S., Eils, R. et al. (2006) Customized oligonucleotide microarray gene expression–based classification of neuroblastoma patients outperforms current clinical risk stratification. *Journal of Clinical Oncology*, **24**, 5070–5078.

Park, Y. and Simpson, D. G. (2019) Robust probabilistic classification applicable to irregularly sampled functional data. *Computational statistics & data analysis*, **131**, 37–49.

Pasqualucci, L., Trifonov, V., Fabbri, G., Ma, J., Rossi, D., Chiarenza, A., Wells, V. A., Grunn, A., Messina, M., Elliot, O. et al. (2011) Analysis of the coding genome of diffuse large b-cell lymphoma. *Nature genetics*, **43**, 830.

Pepe, M. S. and Couper, D. (1997) Modeling partly conditional means with longitudinal data. *Journal of the American Statistical Association*, **92**, 991–998.

Pérez, M. L. G., Janys, L., Miranda, M. D. M. and Nielsen, J. P. (2013) Bandwidth selection in marker dependent kernel hazard estimation. *Computational Statistics & Data Analysis*, **68**, 155–169.

Pomann, G.-M., Staicu, A.-M. and Ghosh, S. (2016) A two-sample distribution-free test for functional data with application to a diffusion tensor imaging study of multiple sclerosis. *The Journal of the Royal Statistical Society, Series C*, **65**, 395–414.

Ramsay, J. O. and Silverman, B. W. (2005) *Functional data analysis*. New York: Springer, 2nd edn.

Reimherr, M. and Nicolae, D. (2014) A functional data analysis approach for genetic association studies. *Annals of Applied Statistics*, **8**, 406–429.

Rosenwald, A., Wright, G., Chan, W. C., Connors, J. M., Campo, E., Fisher, R. I., Gascoyne, R. D., Muller-Hermelink, H. K., Smeland, E. B., Giltnane, J. M. et al. (2002) The use of molecular profiling to predict survival after chemotherapy for diffuse large-b-cell lymphoma. *The New England Journal of Medicine*, **346**, 1937–1947.

Rosenwald, A., Wright, G., Wiestner, A., Chan, W. C., Connors, J. M., Campo, E., Gascoyne, R. D., Grogan, T. M., Muller-Hermelink, H. K., Smeland, E. B. et al. (2003) The proliferation gene expression signature is a quantitative integrator of

oncogenic events that predicts survival in mantle cell lymphoma. *Cancer Cell*, **3**, 185–197.

Rovaris, M., Gass, A., Bammer, R., Hickman, S., Ciccarelli, O., Miller, D. and Filippi, M. (2005) Diffusion MRI in multiple sclerosis. *Neurology*, **65**, 1526–1532.

Scharfstein, D. O., Rotnitzky, A. and Robins, J. M. (1999) Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, **94**, 1096–1120.

Scheipl, F., Staicu, A.-M. and Greven, S. (2015) Functional additive mixed models. *The Annals of Statistics*, **24**, 477–501.

Schuff, N., Woerner, N., Boreta, L., Kornfield, T., Shaw, L., Trojanowski, J., Thompson, P., Jack Jr, C., Weiner, M. and Initiative, A. D. N. (2009) MRI of hippocampal volume loss in early Alzheimer's disease in relation to ApoE genotype and biomarkers. *Brain*, **132**, 1067–1077.

Shah, A., Wilson, A. and Ghahramani, Z. (2014) Student-t processes as alternatives to gaussian processes. *Artificial intelligence and statistics*, 877–885.

She, Y. (2009) Thresholding-based iterative selection procedures for model selection and shrinkage. *The Annals of Statistics*, **3**, 384–415.

Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G. and Meyre, D. (2019) Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics*, **20**, 467–484.

Tang, J., Li, Y. and Guan, Y. (2016) Generalized quasi-likelihood ratio tests for semiparametric analysis of covariance models in longitudinal data. *Journal of the American Statistical Association*, **111**, 736–747.

Tibshirani, R. (1997) The lasso method for variable selection in the cox model. *Statistics in Medicine*, **16**, 385–395.

Tsiatis, A. (2006) *Semiparametric Theory and Missing Data*. New York: Springer.

van der Vaart, A. W. (2000) *Asymptotic Statistics*. New York: Cambridge University Press.

Vafajoo, A., Rostami, A., Parsa, S. F., Salarian, R., Rabiee, N., Rabiee, G., Rabiee, M., Tahriri, M., Vashaee, D., Tayebi, L. et al. (2018) Early diagnosis of disease using microbead array technology: A review. *Analytica Chimica Acta*, https://doi.org/10.1016/j.aca.2018.05.011.

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A. and Yang, J. (2017) 10 years of GWAS discovery: biology, function, and translation. *The American Journal of Human Genetics*, **101**, 5–22.

Wang, H., Zhong, P.-S., Cui, Y. and Li, Y. (2018) Unified empirical likelihood ratio tests for functional concurrent linear models and the phase transition from sparse to dense functional data. *Journal of the Royal Statistical Society, Series B*, **80**, 343–364.

Wang, J., Xue, L., Zhu, L. and Chong, Y. (2010) Estimation for a partial-linear single-index model. *The Annals of Statistics*, **38**, 246–274.

Wang, N., Carroll, R. J. and Lin, X. (2005) Efficient semiparametric marginal estimation for longitudinal/clustered data. *Journal of the American Statistical Association*, **100**, 147–157.

Wang, W. (2004) Proportional hazards regression models with unknown link function and time-dependent covariates. *Statistica Sinica*, 885–905.

Wang, Z., Xu, K., Zhang, X., Wu, X. and Wang, Z. (2017) Longitudinal snp-set association analysis of quantitative phenotypes. *Genetic epidemiology*, **41**, 81–93.

van Wieringen, W. N., Kun, D., Hampel, R. and Boulesteix, A.-L. (2009) Survival prediction using gene expression data: a review and comparison. *Computational Statistics and Data Analysis*, **53**, 1590–1603.

Wright, S. J., Nowak, R. D. and Figueiredo, M. A. (2009) Sparse reconstruction by separable approximation. *IEEE Transactions on Signal Processing*, **57**, 2479–2493.

Xia, Y. (2006) Asymptotic distributions for two estimators of the single-index model. *Econometric Theory*, **22**, 1112–1137.

Xia, Y., Tong, H., Li, W. and Zhu, L.-X. (2002) An adaptive estimation of dimension reduction space. *Journal of Royal Statistical Society, Series B*, **64**, 363–410.

Xia, Y., Zhang, X. and Xu, J. (2010) Dimension reduction and semiparametric estimation of survival models. *Journal of the American Statistical Association*, **105**, 278–290.

Xu, C. and Chen, J. (2014) The sparse mle for ultrahigh-dimensional feature screening. *Journal of the American Statistical Association*, **109**, 1257–1269.

Xu, Y., Li, Y. and Nettleton, D. (2018) Nested hierarchical functional data modeling and inference for the analysis of functional plant phenotypes. *Journal of the American Statistical Association*, **113**, 593–606.

Xu, Z., Shen, X., Pan, W., Initiative, A. D. N. et al. (2014) Longitudinal analysis is more powerful than cross-sectional analysis in detecting genetic association with neuroimaging phenotypes. *PloS one*, **9**.

Yahav, I. and Shmueli, G. (2012) On generating multivariate poisson data in management science applications. *Applied Stochastic Models in Business and Industry*, **28**, 91–102.

Yang, G., Yu, Y., Li, R. and Buu, A. (2016) Feature screening in ultrahigh dimensional cox's model. *Statistica Sinica*, **26**, 881.

Yang, L. (2017) Proximal gradient method with extrapolation and line search for a class of nonconvex and nonsmooth problems. *arXiv preprint arXiv:1711.06831*.

Yao, F., Müller, H.-G. and Wang, J.-L. (2005) Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, **100**, 577–590.

Yin, G., Li, H. and Zeng, D. (2008) Partially linear additive hazards regression with varying coefficients. *Journal of the American Statistical Association*, **103**, 1200–1213.

Young, G. A., Young, G. A., Severini, T. A., Smith, R., Smith, R. L. et al. (2005) *Essentials of statistical inference*, vol. 16. Cambridge University Press.

Zeng, D., Yin, G. and Ibrahim, J. (2005) Inference for a class of transformed hazards model. *Journal of the American Statistical Association*, **100**, 1000–1008.

Zhang, J. T. (2013) *Analysis of Variance for Functional Data*. CRC Press, New York.

Zhang, J. T. and Chen, J. W. (2007) Statistical inferences for functional data. *The Annals of Statistics*, **35**, 1052–1079.

Zhang, T. (2009) Multi-stage convex relaxation for learning with sparse regularization. In *Advances in Neural Information Processing Systems*, 1929–1936.

Zhao, P. and Yu, B. (2006) On model selection consistency of lasso. *Journal of Machine Learning Research*, **7**, 2541–2563.

Zhao, S. D. and Li, Y. (2012) Principled sure independence screening for cox models with ultra-high-dimensional covariates. *Journal of Multivariate Analysis*, **105**, 397–411.

Zhou, L., Huang, J., Martinez, J. G., Maity, A., Baladandayuthapani, V. and Carroll, R. J. (2010) Reduced rank mixed effects models for spatially correlated hierarchical functional data. *Journal of the American Statistical Association*, **105**, 390–400.

Zhu, L. and Xue, L. (2006) Empirical likelihood confidence regions in a partially linear single-index model. *Journal of Royal Statistical Society, Series B*, **68**, 549–570.

Zhu, W., Xu, S., Li, Y. and Liu, C. (2020+) Minimax powerful functional tests for longitudinal genome-wide association studies. *Technical Report*, Submitted.

Zwibel, H. L. and Smrtka, J. (2011) Improving quality of life in multiple sclerosis: an unmet need. *The American Journal of Managed Care*, **17**, 139–145.