SEMIPARAMETRIC REGRESSION ANALYSIS OF SURVIVAL AND LONGITUDINAL DATA

by

Mengdie Yuan
A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Statistical Science

Committee:

	Dr. Guoqing Diao, Dissertation Director
	Dr. Jacqueline M. Hughes-Oliver, Committee Member
	Dr. Anand N. Vidyashankar, Committee Member
	Dr. Ali A. Weinstein, Committee Member
	Dr. William F. Rosenberger, Department Chair
	Dr. Kenneth S. Ball, Dean, Volgenau School of Engineering
Date: April 28th, 2014	Spring Semester 2014 George Mason University Fairfax, VA

Semiparametric Regression Analysis of Survival and Longitudinal Data

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

By

Mengdie Yuan
Master of Science
George Mason University, 2011
Bachelor of Science
University of Science and Technology of China, 2009

Director: Dr. Guoqing Diao, Professor
Department of Statistics

Spring Semester 2014
George Mason University
Fairfax, VA

Dedication

I dedicate this dissertation to my parents.

Acknowledgments

It has been almost five years since I started graduate school at George Mason University. In this memorable journey of my life, fortunately I have met the many great colleagues and friends. Here I give my deepest thanks to Dr. Guoqing Diao, my doctoral advisor, who has been the guiding force behind me during my stay here. The enlightening lectures he gave and the inspiring discussions we had got me engrossed to this field and strongly influenced what I set to do after completing my doctoral degree. I am grateful to him for the thoughtful advice and indispensable help that he gave throughout my research, dissertation and job searching.

I would also like to thank some other people who have played deepseated roles in my academic life here at George Mason University. I owe my sincerest gratitude to Dr. Rosenberger for taking time to discuss with me on my study plans and offering me his kindly help on seizing the many opportunities. I would like to thank Dr. Hughes-Oliver, Dr. Vidyashankar and Dr. Weinstein for serving on my thesis committee and giving valuable feedback on my dissertation. I would like to thank the fellow graduate students at GMU, especially Brad Patterson and Yang Wang, for the help they gave and the delightful discussions we had over these years. The last but not the least, I give my sincere thanks to the administrative staff in the department of statistics, whose hard work and kindly help made my stay here very enjoyable.

I feel indebted to my parents and my brother, who are thousand miles away but always close to my heart. I also would like to thank my best friend Sin-Shuen, a passionate cook and rising computer scientist, for her unlimited support and encouragement over these years.

Table of Contents

List of Tables

List of Figures

Abstract

SEMIPARAMETRIC REGRESSION ANALYSIS OF SURVIVAL AND LONGITUDINAL DATA

Mengdie Yuan, PhD

George Mason University, 2014

Dissertation Director: Dr. Guoqing Diao

The proportional odds model is used in survival analysis when the odds ratios over time are constant. The assumption, however, is often violated in many applications. We propose a novel semiparametric general odds ratio model for the analysis of right-censored survival data. The proposed model incorporates the short-term and long-term covariate effects on the failure time data, and includes the proportional odds model as a special case. We derive efficient likelihood-based inference procedures and establish the large sample properties of the proposed nonparametric maximum likelihood estimators. Extensive simulation studies demonstrate the proposed methods perform well in practical settings. An application to a breast cancer study is provided.

Secondly, we study the generalized linear models with an unknown link function for both data with independent observations and longitudinal data. We propose sieve maximum likelihood estimation procedures for both fixed effects models and random effects models by using B-splines. We establish the consistency and asymptotic normality of the proposed sieve maximum likelihood estimators. Extensive simulation studies along with an application to an epileptic study are provided to evaluate the finite-sample performance of the proposed methods.

Chapter 1: Introduction

Semiparametric regression models are useful alternatives to standard parametric models in that they are generally robust to violation of certain model assumptions. These models typically involve finite dimensional regression coefficients and one or more infinite dimensional functional parameters. In this dissertation, we investigate several semiparametric regression models for survival data and longitudinal data and develop statistically efficient likelihood-based estimation and inference procedures. We derive the asymptotic properties of the proposed estimators through advanced empirical process theories. The performance of the proposed methodologies is evaluated through extensive simulation studies and applications to real examples.

1.1 Proportional Odds Model

Survival data arise in many scientific areas, such as health studies, biomedical studies, and insurance. Examples of survival data include time to death, time to develop a disease, and time to default of an insurance company, etc. Survival time are often subject to censoring because of loss to follow-up, limited duration of study, or failure due to competing causes. One main objective of the analysis of survival data is to evaluate the effects of covariates of interest on the survival time. The proportional odds (PO) model (Bennett, 1983a,b; Pettitt, 1984) is a commonly-used regression model for analyzing survival data. The PO model specifies that the survival function of the failure time T given the covariates $\mathbf{X}$ takes the form

$$S(t|\mathbf{X}) = \frac{e^{-\boldsymbol{\beta}^T\mathbf{X}}}{H(t) + e^{-\boldsymbol{\beta}^T\mathbf{X}}},$$

where $H(t)$ is a strictly increasing function with $H(0) = 0$, and $\boldsymbol{\beta}$ is a vector of regression parameters. Unlike the Cox proportional hazards model (Cox, 1972), the hazard ratio between two sets of covariates is time-varying under the PO model. Instead, the odds ratio of survival is assumed to be constant over time under the PO model. Consequently, the hazard ratio converges to one as time goes to infinity. The converging relative risk property makes the PO model favorable to the proportional hazard model, another commonly used survival model, in practice.

The performance of the PO model relies on the assumption of constant odds ratios over time. This assumption, however, is violated in many applications. The violation of the proportional odds assumption can lead to biased results and may compromise the power of identifying covariates with significant time-varying effects. In Chapter 2, we propose a novel semiparametric odds rate model which allows for time-varying covariate effects and one can explicitly model the short-term and long-term odds ratios.

1.2 Generalized Linear Models and Generalized Linear Mixed Models

The classical linear models assume the response variable is continuous and and is normally distributed. These assumptions are often not true in practice. For example, we often encounter binary, ordinal, and count data, none of which follow normal distributions. Linear models are not appropriate for these types of outcomes.

Generalized linear models (GLMs) are an extension of linear models. The responses are allowed to follow a distribution in the exponential family without being restricted to normal as in linear models. A link function is introduced when regressing the expected value of the response variable on a linear combination of a set of covariates. It is therefore commonly used when analyzing discrete and non-normal responses.

The general form of the distributions in the exponential family is given by

$$f_Y(y) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} - c(y,\phi)\right\},$$

where $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$ are known functions and ϕ is a scaler. We call θ the canonical parameter. The mean and variance of Y can be obtained as

$$E(Y) = b'(\theta), Var(Y) = b''(\theta)a(\phi),$$

where $b'(\cdot)$ and $b''(\cdot)$ denote the first and second derivatives of $b(\cdot)$, respectively. The exponential family includes common distributions such as Normal, Binomial, Poisson, Exponential, Gamma, and inverse Gaussian. To relate the outcome variable Y with covariates $\mathbf{X}$, we denote $E(Y|\mathbf{X}) = \mu$ and define the linear predictor $\eta = \mathbf{X}^T\boldsymbol{\beta}$. If we have the model as $g(\mu) = \eta$, we call $g(\cdot)$ the link function and $h(\cdot) = g^{-1}(\cdot)$ the inverse link function. A particular link function $b'^{-1}(\cdot)$ is known as the canonical link function, where $b'^{-1}(\cdot)$ is the inverse function of the first derivative of $b(\cdot)$. It is intuitive and common to use the canonical link.

Generalized linear models have also been widely applied to longitudinal data with repeated measurements. For longitudinal data, it is necessary to account for the correlations among repeated measurements. There are mainly two methods to account for heterogeneity across time, namely, the marginal approach and the conditional approach. In the fixed effects models, the marginal distribution of the outcome at each time point is specified, and a variance-covariance structure is set for the repeated measurements. That is, the mean and variance-covariance are constructed separately. There are no subject-specific coefficients or random effects in the mean structure. All subjects share the same fixed effects. Thus the inference is made on population averages. The coefficients can be estimated by generalized estimating equations (GEE). Thall and Vail (1990), Pourahmadi (1999) and Pourahmadi (2000) described some possible choices of the variance-covariance structures.

Mixed models incorporate subject-specific random effects in the mean structure in addition to the fixed effects. This approach allows the regression coefficients to vary across subjects according to the distribution for the random effects. The variance-covariance matrix for the repeated measurements from a subject therefore is contributed by two parts: the subject-specific random effects and the random errors. The response vectors of different subjects conditional on the subject-specific random effects are assumed to be independent. With non-linear link functions, there is no model-based interpretation for the population mean. The expectation of the linear predictor does not necessarily explain the expected mean through the non-linear link function as the case with the identity link. Usually we refer to this kind of random effects models as the generalized linear mixed models (GLMMs). One issue for random effects models is that the model misspecification may cause bias results (Heagerty and Kurland, 2001).

Zeger et al. (1988), Heagerty and Zeger (2000) and Heagerty and Kurland (2001) investigated the relationship between the random effects models and the fixed effects models. They showed that we could associate the random effects model with a fixed effects model by imposing an adjustment function in the fixed part of the linear predictors and use the same link function for both models. However Chiou and Müller (2005) pointed out that the link functions for the random effects model and the corresponding fixed effects model are not in general coincident due to the impact of the random effects. In Chapters 3 and 4, we develop sieve maximum likelihood estimation procedures for GLMs and GLMMs with an unknown link function using B-spline smoothing.

1.2.1 Fixed Effects Models

Suppose we have n subjects with k_i repeated measurements for the ith subject. Let Y_{ij} denote the jth outcome of the ith subject with the corresponding p-dimensional vector of covariates $\mathbf{X}_{ij}$, where $i = 1, \cdots, n$ and $j = 1, \cdots, k_i$. The distribution of Y_{ij} belongs to the exponential family

$$f_Y(y_{ij}) = \exp\left\{\frac{y_{ij}\theta_{ij} - b(\theta_{ij})}{a(\phi)} - c(y_{ij}, \phi)\right\}.$$

For simplicity, suppose that $a(\phi) = \phi$ which is a scalar parameter. Let $E(Y_{ij}) = \mu_{ij}$ and $Var(Y_{ij}) = \phi v_{ij}(\mu_{ij})$, where $\{v_{ij}(\cdot) : i = 1, \cdots, n; j = 1, \cdots, k_i\}$ be known functions. Generalized linear models employ a one-to-one continuously differentiable function $g(\cdot)$ to adjust the mean structure as a linear model. For fixed effects models, we have

$$g(\mu_{ij}) = \eta_{ij} = \mathbf{X}_{ij}^T\boldsymbol{\beta}, i = 1, \cdots, n, j = 1, \cdots, k_i,$$

or

$$\mu_{ij} = h(\eta_{ij}) = h(\mathbf{X}_{ij}^T\boldsymbol{\beta}), i = 1, \cdots, n, j = 1, \cdots, k_i, \tag{1.1}$$

where $g(\cdot)$ is the link function and $h(\cdot) = g^{-1}(\cdot)$ is the inverse link function.

Unlike linear models for normal data, the likelihood cannot be fully determined by only specifying the mean and the covariance. In this regard, the inference based on a conventional likelihood function is not suggested. Alternatively, quasi-likelihood is preferred in the absence of a full distribution assumption for the data.

Let $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \cdots, Y_{ik_i})^T$ and the corresponding mean vector $\boldsymbol{\mu}_i = (\mu_{i1}, \mu_{i2}, \cdots, \mu_{ik_i})$. We define

$$\mathbf{D}_i = \left(\frac{\partial\boldsymbol{\mu}_i}{\partial\boldsymbol{\beta}}\right)$$

which is a $k_i \times p$ partial derivative matrix. The $k_i \times k_i$ covariance matrix of $\mathbf{Y}_i$ is denoted by $\boldsymbol{\Sigma}_i$. Suppose that $\boldsymbol{\Sigma}_i(\phi, \boldsymbol{\mu}_i) = \phi\mathbf{V}_i(\boldsymbol{\mu}_i)$, where $\mathbf{V}_i(\boldsymbol{\mu}_i)$ is a known matrix function of $\boldsymbol{\mu}_i$. The quasi-likelihood $Q_i(\boldsymbol{\mu}_i|\mathbf{Y}_i)$ for the ith subject satisfies

$$\frac{\partial Q_i(\boldsymbol{\mu}_i|\mathbf{Y}_i)}{\partial\boldsymbol{\mu}_i} = \mathbf{V}_i^{-1}(\boldsymbol{\mu}_i)(\mathbf{Y}_i - \boldsymbol{\mu}_i).$$

Since we assume independence between subjects, the quasi-likelihood can be written as

$$Q(\boldsymbol{\mu}|\mathbf{Y}) = \sum_{i=1}^{n} Q_i(\boldsymbol{\mu}_i|\mathbf{Y}_i).$$

The corresponding quasi-score equations are given by

$$\sum_{i=1}^{n} \mathbf{D}_i^T \mathbf{V}_i^{-1}(\boldsymbol{\mu}_i)(\mathbf{Y}_i - \boldsymbol{\mu}_i) = \mathbf{0}.$$

The equations are non-linear functions of $\boldsymbol{\beta}$ which therefore may not be solved analytically.

An extension of the quasi-likelihood-based method is the generalized estimating equations (GEE) approach. In GEE, the covariance matrix is specified as

$$\boldsymbol{\Sigma}_i(\phi, \boldsymbol{\mu}_i, \boldsymbol{\alpha}) = \phi \mathbf{V}_i(\boldsymbol{\mu}_i, \boldsymbol{\alpha}) = \phi \mathbf{A}_i^{\frac{1}{2}}(\boldsymbol{\mu}_i)\mathbf{R}_i(\boldsymbol{\alpha})\mathbf{A}_i^{\frac{1}{2}}(\boldsymbol{\mu}_i),$$

where $\mathbf{A}_i(\boldsymbol{\mu}_i) = \mathbf{Diag}\left(v_{i1}(\mu_{i1}), v_{i2}(\mu_{i2}), \cdots, v_{ik_i}(\mu_{ik_i})\right)$, and $\mathbf{R}_i(\boldsymbol{\alpha})$ is a working correlation matrix indexed by some unknown parameters $\boldsymbol{\alpha}$. Usually, we assume $\mathbf{R}_i(\boldsymbol{\alpha}) = \mathbf{R}(\boldsymbol{\alpha})$ among subjects for balanced data. The unknown parameters include $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. The generalized estimating equations are given by

$$\sum_{i=1}^{n} \mathbf{D}_i^T \mathbf{V}_i^{-1}(\boldsymbol{\mu}_i, \boldsymbol{\alpha})(\mathbf{Y}_i - \boldsymbol{\mu}_i) = \mathbf{0}.$$

Typically, the estimators are obtained by iterative algorithms. First, given current estimates of $\widehat{\phi}^{(k)}$ and $\widehat{\boldsymbol{\alpha}}^{(k)}$, we obtain $\widehat{\boldsymbol{\beta}}^{(k)}$ to be the solution to the generalized estimating equations. Initially, we may use an independent structure for $\mathbf{V}_i$, where $\mathbf{R}(\boldsymbol{\alpha}) = \mathbf{I}$. Then given $\widehat{\boldsymbol{\beta}}^{(k)}$, we update $\widehat{\phi}^{(k+1)}$ and $\widehat{\boldsymbol{\alpha}}^{(k+1)}$ on the basis of current standardized residuals

$$r_{ij}^{(k)} = \frac{y_{ij} - \widehat{\mu}_{ij}^{(k)}}{\sqrt{v_{ij}(\widehat{\mu}_{ij}^{(k)})}}.$$

Generally,

$$\widehat{\phi}^{(k+1)} = \frac{1}{n-p} \sum_{i=1}^{n} \sum_{j=1}^{k_i} \left(r_{ij}^{(k)} \right)^2 .$$

The formula for $\widehat{\boldsymbol{\alpha}}^{(k+1)}$ depends on the assumed structure of $\mathbf{R}(\boldsymbol{\alpha})$. We iterate between the above two steps until convergence. The resulting estimator $\widehat{\boldsymbol{\beta}}$ can be shown to be asymptotically unbiased and normally distributed.

Let $\mathbf{Var}(\widehat{\boldsymbol{\beta}})$ be the covariance matrix of $\widehat{\boldsymbol{\beta}}$. Denote $\widehat{\boldsymbol{\Sigma}}_i = \boldsymbol{\Sigma}_i(\widehat{\phi}, \widehat{\boldsymbol{\mu}}_i, \widehat{\boldsymbol{\alpha}})$ and

$$\mathbf{S}_0 = \sum_{i=1}^{n} \mathbf{D}_i^T \widehat{\boldsymbol{\Sigma}}_i^{-1} \mathbf{D}_i.$$

Then

$$\widehat{\boldsymbol{\Sigma}} = \mathbf{S}_0^{-1}$$

is called a "model-based" or naive estimator of $\mathbf{Var}(\widehat{\boldsymbol{\beta}})$. However the naive covariance estimator $\widehat{\boldsymbol{\Sigma}}$ is not a consistent estimator of $\mathbf{Var}(\widehat{\boldsymbol{\beta}})$ if $\mathbf{R}(\boldsymbol{\alpha})$ is incorrectly specified.

A "sandwich", robust or empirical estimator is defined as

$$\widehat{\boldsymbol{\Sigma}}_s = \mathbf{S}_1^{-1} \mathbf{S}_0 \mathbf{S}_1^{-1},$$

where

$$\mathbf{S}_1 = \sum_{i=1}^{n} \mathbf{D}_i^T \widehat{\boldsymbol{\Sigma}}_i^{-1} (\mathbf{y}_i - \widehat{\boldsymbol{\mu}}_i)(\mathbf{y}_i - \widehat{\boldsymbol{\mu}}_i)^T \widehat{\boldsymbol{\Sigma}}_i^{-1} \mathbf{D}_i.$$

This empirical estimator is preferred because it provides a consistent estimator of $\mathbf{Var}(\widehat{\boldsymbol{\beta}})$ even if the correlation structure is misspecified.

1.2.2 Random Effects Models

The GLMMs specify

$$g(E(Y_{ij}|\mathbf{b}_i)) = \mathbf{X}_{ij}^T\boldsymbol{\beta} + \mathbf{Z}_{ij}^T\mathbf{b}_i, i = 1, \cdots, n, j = 1, \cdots, k_i$$

or

$$E(Y_{ij}|\mathbf{b}_i) = h\left(\mathbf{X}_{ij}^T\boldsymbol{\beta} + \mathbf{Z}_{ij}^T\mathbf{b}_i\right), i = 1, \cdots, n, j = 1, \cdots, k_i, \tag{1.2}$$

where $g(\cdot)$ is the link function and $h(\cdot)$ is the inverse link function. In the model, $\mathbf{Z}_{ij}$ is a set of covariates, and $\mathbf{b}_i$'s are q-dimensional subject-specific random vectors which are independently and identically distributed with a joint distribution f_b. Typically, $\mathbf{b}_i$'s are assumed to be normal distributed with mean $\mathbf{0}$ and variance-covariance matrix $\mathbf{M}$ although other distributions can be considered too. We also assume $Y_{i1}, \cdots, Y_{ik_i}$ are mutually independent given $\mathbf{b}_i$.

As mentioned before, the coefficients in the random effects model do not necessarily have the same interpretation as in the fixed effects model. It follows from

$$E(g(E(Y_{ij}|\mathbf{b}_i))) = \mathbf{X}_{ij}^T\boldsymbol{\beta} \neq g(E(Y_{ij}))$$

except for the case of linear mixed models. Actually, we cannot simplify the marginal mean and variance of the outcomes without knowing about $g(\cdot)$ and the distribution of $\mathbf{b}_i$. Notice that using the same link function in a fixed effects model and in a GLMM does not lead to the same model for μ_{ij} in non-normal cases.

The inference is typically likelihood-based under the GLMMs. The marginal likelihood is given by

$$\prod_{i=1}^{n} \int f(\mathbf{Y}_i|\mathbf{b}_i) f_b(\mathbf{b}_i) d\mathbf{b}_i.$$

In general, there is no closed form for the integral. Numerical or Monte-Carlo approaches can be used to approximate the marginal likelihood. We can use numerical methods to obtain the maximum likelihood estimators by maximizing the marginal likelihood. Given the MLEs ($\widehat{\boldsymbol{\beta}}$,$\widehat{\mathbf{M}}$), we can predict the random effects as $\widehat{\mathbf{b}}_i = E(\mathbf{b}_i|\mathbf{Y}_i, \widehat{\boldsymbol{\beta}}, \widehat{\mathbf{M}})$.

1.3 B-spline Smoothing

B-spline smoothing in statistics is a technique to estimate a nonparametric real-valued function using a linear combination of basis splines (B-splines) of certain degree. A B-spline is a piecewise polynomial function, where the pieces connected at specified knots.

Suppose there are K interior knots $\xi_1 < \xi_2 < \cdots < \xi_K$ in an interval bounded by $\xi_0 < \xi_1 < \xi_2 < \cdots < \xi_K < \xi_{K+1}$. We augment the K interior knots $\boldsymbol{\xi} = (\xi_1, \xi_2, \cdots, \xi_K)$ to a $K + 2M$ knot sequence $\boldsymbol{\tau}$ defined below:

- $\tau_1 \le \tau_2 \le \cdots \tau_M \le \xi_0$;
- $\tau_{j+M} = \xi_j$, $j = 1, \cdots, K$;
- $\xi_{K+1} \le \tau_{K+M+1} \le \tau_{K+M+2} \le \cdots \tau_{K+2M}$.

Notice that the knots must be in an ascending order. Typically, the additional knots are made all the same and equal to ξ_0 and ξ_{K+M}, respectively.

Suppose $B_{i,m}(x)$ is the ith B-spline basis function of order m, where $m \le M$. The basis functions can be derived recursively by the following formula,

$$B_{i,1}(x) = \begin{cases} 1 & \text{if } \tau_i \le x \le \tau_{i+1} \\ 0 & \text{otherwise} \end{cases}, \ i = 1, \cdots, K + 2M - 1$$

and

$$B_{i,m}(x) = \frac{x - \tau_i}{t_{i+m-1} - \tau_i} B_{i,m-1}(x) + \frac{\tau_{i+m} - x}{\tau_{i+m} - \tau_{i+1}} B_{i+1,m-1}(x),$$

for $i = 1, \cdots, K + 2M - m$. Particularly, with $M = 4$, we call $B_{i,4}$, $i = 1, \cdots, K + 4$ the cubic B-spline basis functions. Therefore, the B-spline of degree m consists of $m + 1$ m-degree polynomial pieces jointing at m internal knots. Using the B-spline to fit the curve $y = g(x)$, we can write

$$g(x) = \sum_{i=1}^{K+M} \gamma_i B_{i,M}(x),$$

where γ_i's are unknown coefficients.

1.4 Kronecker Product

Let $\mathbf{A} = (a_{ij})$ be a $n \times p$ matrix and $\mathbf{B} = (b_{ij})$ be a $m \times q$ matrix. Then the Kronecker product is defined by

$$\mathbf{A} \otimes \mathbf{B} = \begin{pmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \cdots & a_{1p}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \cdots & a_{2p}\mathbf{B} \\ \vdots & \vdots & \cdots & \vdots \\ a_{n1}\mathbf{B} & a_{n2}\mathbf{B} & \cdots & a_{np}\mathbf{B} \end{pmatrix}_{mn \times pq},$$

and

$$\mathbf{B} \otimes \mathbf{A} = \begin{pmatrix} b_{11}\mathbf{A} & b_{12}\mathbf{A} & \cdots & b_{1p}\mathbf{A} \\ b_{21}\mathbf{A} & b_{22}\mathbf{A} & \cdots & b_{2p}\mathbf{A} \\ \vdots & \vdots & \cdots & \vdots \\ b_{m1}\mathbf{A} & b_{m2}\mathbf{A} & \cdots & b_{mp}\mathbf{A} \end{pmatrix}_{mn \times pq}.$$

1.5 Sieve Maximum Likelihood Estimators

Let Θ be the parameter space, and $L_n(\theta)$, $\theta \in \Theta$ be the likelihood. The maximum likelihood estimator $\hat{\theta}_n$ may be inconsistent if the size of the underlying parameter space is too large. Therefore we often maximize the likelihood over a space Θ_n that is an approximation of Θ, where the approximation error decreases to zero as the sample size increases. Then Θ_n is called the sieve space, and the corresponding maximizer is called the sieve maximum likelihood estimators.

1.6 Gaussian Quadrature Approximation

1.6.1 One-dimensional Gauss-Hermite Quadrature

Gaussian quadrature is commonly applied to numerically approximate the integral against the probability measures. The integral can be presented as

$$\int_{-\infty}^{\infty} f(x)p(x)dx,$$

where $p(x)$ is a density function. When $p(x)$ has the kernel e^{-x^2}, we can use the Gauss-Hermite quadrature. Let N_q denote the number of quadrature points, ω_i and x_i are the quadrature weights and the abscissas, $i = 1, \cdots, N_q$. Then

$$\int_{-\infty}^{\infty} f(x)p(x)dx \approx \sum_{i=1}^{N_q} \omega_i f(x_i).$$

1.6.2 High-dimensional Gauss-Hermite Quadrature

Let $\mathbf{b}_{r\times 1}$ follow a multivariate normal distribution with density $p(\mathbf{b})$. To get the r-dimensional quadrature grid, we first obtain the empirical Bayes estimates $\hat{\mathbf{b}}$ of $\mathbf{b}$ by minimizing

$$-\log(f(x|\mathbf{b})p(\mathbf{b})) = g(x;\mathbf{b}).$$

Let $\boldsymbol{\omega} = (\omega_1, \omega_2, \cdots, \omega_{N_q})$ and $\mathbf{x} = (x_1, x_2, \cdots, x_{N_q})$ which are defined above. The for a point on the r-dimensional quadrature grid $\mathbf{x}_k^* = (x_{k_1}, x_{k_2}, \cdots, x_{k_r})$, the centered and scaled abscissas are given by

$$\mathbf{a}_j = \hat{\mathbf{b}} + \sqrt{2} g''(x;\hat{\mathbf{b}})^{-\frac{1}{2}} \mathbf{x}^*,$$

where

$$g''(x;\mathbf{b}) = \frac{\partial^2 g(x;\mathbf{b})}{\partial \mathbf{b} \partial \mathbf{b}^T}\Big|_{\hat{\mathbf{b}}}.$$

Then the integral can be approximated as

$$\int f(x|\mathbf{b})p(\mathbf{b})d\mathbf{b} \approx (\sqrt{2})^r |g''(x;\hat{\mathbf{b}})|^{-\frac{1}{2}} \sum_{k_1=1}^{N_q} \cdots \sum_{k_r=1}^{N_q} \left[f(x|\mathbf{a}_j)p(\mathbf{a}_j) \prod_{l=1}^{r} \omega_{k_l} e^{x_{k_l}^2} \right].$$

1.7 Outline

In Chapter 2, we propose a novel semiparametric general odds ratio model for the analysis of right-censored survival data by incorporating the short-term and long-term covariate effects. The baseline odds function is unspecified and we derive the nonparametric maximum likelihood estimators and establish their large sample properties. In Chapter 3, we study the generalized linear models with an unknown link function for independent data. We propose to use B-splines to approximate the unknown link function and derive the sieve maximum likelihood estimators (MLEs). In Chapter 4, we extend the proposed methods in Chapter 3 to longitudinal data by incorporating random effects in the model. In both Chapters 3 and

Chapter 4 , we derive the convergence rate of the sieve MLEs and establish the asymptotic normality and asymptotic efficiency of the estimators of the finite dimensional parameters. We conclude this dissertation with discussions and future work in Chapter 5.

Chapter 2: Semiparametric Odds Rate Model for Modeling Short-term and Long-term Effects with Application to a Breast Cancer Genetic Study

2.1 Introduction

Much attention has been paid to the development of inference procedures under the PO model and its generalizations. Murphy et al. (1997) provided statistical inference for the maximum likelihood estimation with right-censored data under the PO model and demonstrated that the estimators of the regression parameters provided by Bennett (1983a) are efficient. The sieve maximum likelihood estimation and the estimation based on weighted empirical odds functions were developed by Shen (1998) and Yang and Prentice (1999), respectively. Computational issues in the PO models are discussed in Hunter and Lange (2002). The PO model has also been extensively studied for current status data and interval censored data, see Rossini and Tsiatis (1996), Huang and Rossini (1997), Rabinowitz et al. (2000) and Sun et al. (2007), among others. For correlated failure time data, the PO model with random effects has been investigated in literature (Lam et al., 2002; Zeng et al., 2005).

All the aforementioned methods are developed under the assumption of constant odds ratio over time. However, in some applications, this assumption is violated. Violation of this assumption may lead to biased and misleading results. A motivating example is from a copy number alteration (CNA) study with breast cancer patients (Zhang et al., 2009). The objective of the study was to determine whether the CNAs influence the time to develop distant metastasis among breast cancer patients. We expect the log odds curves of survival to be parallel for two groups if the proportional odds assumption is true. Figure 2.1 presents the Kaplan-Meier curves for two groups partitioned by the median of copy number variation data at two SNP loci. It appears that the proportional odds assumption is reasonable at SNP #2124; however, the two curves cross at SNP #39687 indicating a violation of the proportional odds assumption. The PO model failed to capture the phenomenon of the crossing survival at SNP #39687.

To test the proportional odds assumption, Dauxois and Kirmani (2003) developed a formal analytical test for two-sample problems. Resampling approaches using observed score processes for the regression coefficients have been developed by Scheike (2006) to check the goodness-of-fit of the PO model. Liu et al. (2009) provided graphical diagnostics to check model misspecification for the PO model. To alleviate the proportional odds assumption, Scheike (2006) proposed a flexible transformation model which allows for time-varying coefficients for some covariates in the model. They developed estimation and inference procedures based on martingale estimating equations. Additionally, they proposed to use a robust sandwich-type covariance matrix for the estimates of the regression coefficients. More recently, Chen et al. (2012) proposed an extended PO model by incorporating "external" time-varying covariates and developed appropriate inference procedures for the regression parameters. Note that both Liu et al. (2009) and Chen et al. (2012) do not incorporate time-varying coefficients.

In this chapter, we propose an alternative approach for relating covariates of interest to the odds of survival for right-censored data. Specifically, we propose a novel semiparametric odds rate model to incorporate the short-term and long-term covariate effects. This model includes the PO model as a special case, leading to a new diagnostic tool for checking the proportional odds assumption. We establish efficient likelihood-based estimation and inference procedures and derive the asymptotic properties of the proposed estimators. Results from numerical studies demonstrate that the proposed methods perform well in practical settings.

The remaining of this chapter is organized as follows. In Section 2.2, we propose the new model and derive the nonparametric maximum likelihood estimation. The model assumptions and the large sample properties are presented in Section 2.3. In Section 2.4, we conduct extensive simulation studies to investigate the finite sample properties of the proposed estimators. An application to the motivating CNA study is provided in Section 2.5 and we conclude with a brief discussion in Section 2.6.

2.2 Methods

Suppose there are n independent subjects in a random sample. For the ith subject, let T_i be the failure time, C_i be the censoring time, and $\mathbf{X}_i$ be a $p \times 1$ vector of time-independent covariates. The observed data contain $\{(Y_i = \min(T_i, C_i), \Delta_i = I(T_i \leq C_i), \mathbf{X}_i), i = 1, ..., n\}$, where $I(\cdot)$ is the indicator function.

Let $F(t|\mathbf{X})$ and $S(t|\mathbf{X})$ be the conditional distribution function and survival function of the failure time T given the covariates $\mathbf{X}$, respectively. We propose the following generalized odds model

$$\frac{F(t|\mathbf{X})}{S(t|\mathbf{X})} = H(t)\left[e^{\boldsymbol{\beta}^T\mathbf{X}}F(t) + e^{\boldsymbol{\gamma}^T\mathbf{X}}S(t)\right], \tag{2.1}$$

where $H(t)$ is a non-decreasing function with $H(0) = 0$. In model (2.1), $H(t)$ is the baseline odds function of failure, $S(t)$ is the baseline survival function which is the survival function for a subject with $\mathbf{X} = \mathbf{0}$, and $F(t) = 1 - S(t)$ is the baseline cumulative distribution function.

Under the proposed model (2.1), the regression coefficients $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ have attractive and intuitive interpretations. For two sets of covariates $\mathbf{X}$ and $\mathbf{X}^*$, we can show that

$$\lim_{t\to 0}\frac{F(t|\mathbf{X})/S(t|\mathbf{X})}{F(t|\mathbf{X}^*)/S(t|\mathbf{X}^*)} = e^{\boldsymbol{\gamma}^T(\mathbf{X}-\mathbf{X}^*)}, \ \lim_{t\to\infty}\frac{F(t|\mathbf{X})/S(t|\mathbf{X})}{F(t|\mathbf{X}^*)/S(t|\mathbf{X}^*)} = e^{\boldsymbol{\beta}^T(\mathbf{X}-\mathbf{X}^*)}.$$

Therefore, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ can be interpreted as the long-term and short-term log-odds ratios, respectively. When $\boldsymbol{\beta} = \boldsymbol{\gamma}$, model (2.1) reduces to the PO model. Hence, this model provides a useful diagnostic tool for testing the PO model. To be more specific, we can check the PO model by testing the null hypothesis $H_0 : \boldsymbol{\beta} = \boldsymbol{\gamma}$. The test can be performed by using Wald, score or likelihood ratio test statistics.

We now derive the likelihood function for the unknown parameters $(\boldsymbol{\beta}, \boldsymbol{\gamma}, H)$. It can be shown that

$$S(t) = \frac{1}{H(t)+1}, \ F(t) = \frac{H(t)}{H(t)+1}.$$

It follows that

$$S(t|\mathbf{X}) = \frac{H(t)+1}{e^{\boldsymbol{\beta}^T\mathbf{X}}H^2(t) + (e^{\boldsymbol{\gamma}^T\mathbf{X}}+1)H(t)+1}.$$

Consequently, the hazard function is given by

$$\lambda(t|\mathbf{X}) = \frac{H'(t)\left\{e^{\boldsymbol{\beta}^T\mathbf{X}}H^2(t) + 2e^{\boldsymbol{\beta}^T\mathbf{X}}H(t) + e^{\boldsymbol{\gamma}^T\mathbf{X}}\right\}}{\{H(t)+1\}\left\{e^{\boldsymbol{\beta}^T\mathbf{X}}H^2(t) + (e^{\boldsymbol{\gamma}^T\mathbf{X}}+1)H(t)+1\right\}},$$

where $H'(t)$ is the first derivative of $H(t)$. Under the conditional independent censoring assumption, the likelihood function for $(\boldsymbol{\beta}, \boldsymbol{\gamma}, H)$ has the form

$$\begin{aligned} L_n(\boldsymbol{\beta}, \boldsymbol{\gamma}, H) &= \prod_{i=1}^{n} \lambda^{\Delta_i}(Y_i|\mathbf{X}_i)S(Y_i|\mathbf{X}_i) \\ &= \prod_{i=1}^{n}\left[\frac{H'(Y_i)\left\{e^{\boldsymbol{\beta}^T\mathbf{X}_i}H^2(Y_i) + 2e^{\boldsymbol{\beta}^T\mathbf{X}_i}H(Y_i) + e^{\boldsymbol{\gamma}^T\mathbf{X}_i}\right\}}{\{H(Y_i)+1\}\left\{e^{\boldsymbol{\beta}^T\mathbf{X}_i}H^2(Y_i) + (e^{\boldsymbol{\gamma}^T\mathbf{X}_i}+1)H(Y_i)+1\right\}}\right]^{\Delta_i} \qquad (2.2) \\ &\quad \times \frac{H(Y_i)+1}{e^{\boldsymbol{\beta}^T\mathbf{X}_i}H^2(Y_i) + (e^{\boldsymbol{\gamma}^T\mathbf{X}_i}+1)H(Y_i)+1}. \end{aligned}$$

We would like to make inference based on the likelihood function. Note that the maximum of the likelihood function does not exist since we can always choose $H'(Y_i) = 0$ for some Y_i with $\Delta_i = 1$. Therefore, following the techniques of nonparametric maximum likelihood estimation in the Cox proportional hazards model (Murphy, 1994, 1995), the proportional odds model (Murphy et al., 1997; Zeng et al., 2005), and the semiparametric transformation model (Zeng et al., 2008), we allow H to be a right continuous function. Therefore, we allow the estimator of H to be a step function with positive jumps only at the observed failure time points. Let $H\{Y_i\}$ denote the jump size of $H(t)$ at Y_i. We replace $H'(Y_i)$ with $H\{Y_i\}$ in the above equation (2.2) to obtain the nonparametric likelihood function, still denoted by $L_n(\boldsymbol{\beta}, \boldsymbol{\gamma}, H)$ for simplicity,

$$L_n(\boldsymbol{\beta}, \boldsymbol{\gamma}, H) = \prod_{i=1}^{n} \left[\frac{H\{Y_i\} \left\{ e^{\boldsymbol{\beta}^T \mathbf{X}_i} H^2(Y_i) + 2e^{\boldsymbol{\beta}^T \mathbf{X}_i} H(Y_i) + e^{\boldsymbol{\gamma}^T \mathbf{X}_i} \right\}}{\{H(Y_i) + 1\} \left\{ e^{\boldsymbol{\beta}^T \mathbf{X}_i} H^2(Y_i) + (e^{\boldsymbol{\gamma}^T \mathbf{X}_i} + 1) H(Y_i) + 1 \right\}} \right]^{\Delta_i}$$

$$\times \frac{H(Y_i) + 1}{e^{\boldsymbol{\beta}^T \mathbf{X}_i} H^2(Y_i) + (e^{\boldsymbol{\gamma}^T \mathbf{X}_i} + 1) H(Y_i) + 1}.$$

Assume the data have m unique observed failure time points $Y_{(1)} < Y_{(2)} < \cdots < Y_{(m)}$. Easily we have $H(t) = \sum_{k: Y_{(k)} \le t} H\{Y_{(k)}\}$. Then

$$S(t) = \frac{1}{\sum_{k: Y_{(k)} \le t} H\{Y_{(k)}\} + 1}, \;\; F(t) = \frac{\sum_{k: Y_{(k)} \le t} H\{Y_{(k)}\}}{\sum_{k: Y_{(k)} \le t} H\{Y_{(k)}\} + 1}.$$

In this case, the odds function $H(t|\mathbf{X}) = F(t|\mathbf{X})/S(t|\mathbf{X})$ is given by

$$\sum_{k: Y_{(k)} \le t} \left\{ e^{\boldsymbol{\beta}^T \mathbf{X}} F(Y_{(k)}) + e^{\boldsymbol{\gamma}^T \mathbf{X}} S(Y_{(k)}) \right\} H\{Y_{(k)}\}.$$

Hence, to estimate the unknown parameters, we maximize the nonparametric likelihood function with respect to the regression parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, and jump sizes $H\{Y_{(1)}\}$, $H\{Y_{(2)}\}, \cdots, H\{Y_{(m)}\}$. The resulting nonparametric maximum likelihood estimators (NPMLEs) are denoted by $(\hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{\gamma}}_n, \hat{H}_n)$.

The maximization procedure can be performed by using the quasi-Newton optimization algorithm in Press et al. (1992). Based on Fisher information theory, the covariance matrix for ($\hat{\boldsymbol{\beta}}_n$, $\hat{\gamma}_n$, $\hat{H}_n$) can be obtained by inverting the observed information matrix, which is the negative second derivatives of the nonparametric likelihood function with respect to the regression parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, and the jump sizes $H\{Y_{(1)}\}$, $H\{Y_{(2)}\}$, $\cdots$, $H\{Y_{(m)}\}$. The justification follows the arguments in Theorem 2 of Parner (1998).

2.3 Asymptotic Properties

Before we derive the asymptotic properties of the NPMLEs, we impose the following regularity conditions:

C1. There exists some positive constant δ_0 such that $Pr(C_i \geq \tau | \mathbf{X}_i) \geq \delta_0$ almost everywhere, where τ is the end of the study. In addition, $Pr(T_i > \tau | \mathbf{X}_i) > 0$.

C2. Conditional on the vector of covariates $\mathbf{X}_i$, the censoring time C_i is independent of T_i for $i = 1, \cdots, n$.

C3. $\mathbf{X}_i$ are bounded for all $i = 1, \cdots, n$. Furthermore, $\mathbf{a}^T \mathbf{X}_i = 0$ almost everywhere if and only if $\mathbf{a} = \mathbf{0}$.

C4. The true values $\boldsymbol{\beta}_0$ and $\boldsymbol{\gamma}_0$ for parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ belong to the interior of a known compact set $\mathcal{B}_0$ in $\mathcal{R}^{2p}$.

C5. The true baseline odds function H_0 belongs to

$$\mathcal{H}_0 = \{H : \ H \text{ is strictly increasing in } [0, \tau] \text{ and is continuously differentiable}$$
$$\text{with } H(0) = 0 \text{ and } H(\tau) < \infty\}.$$

Remark 2.1. Condition C3 implies that $\mathbf{X}_i$ is linearly independent, which ensures the identifiability of the model. Condition C1 implies that any subject who survives at the end of the study is considered censored at time τ, and every subject has a positive probability to relapse after the study. Conditions C4)and C5 guarantee the existence of the NPMLEs.

We first establish the consistency of the NPMLEs.

Theorem 2.1. Under conditions C1 - C5, the NPMLEs $(\hat{\boldsymbol{\beta}}_n, \hat{\gamma}_n, \hat{H}_n)$ exist. Furthermore $||\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0|| \to 0$, $||\hat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_0|| \to 0$ and $\sup_{t\in[0,\tau]} |\hat{H}_n(t) - H_0(t)| \to 0$ almost surely as $n \to \infty$, where $|| \cdot ||$ is the Euclidean norm.

The existence of $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ can be proved from the compactness of the parameter space and the boundedness of covariates. Suppose for some Y_i with $\Delta_i = 1$, we have $\hat{H}_n\{Y_i\} = \infty$. In that case, the likelihood function must be 0, which implies $\hat{H}_n\{Y_i\}$ must be finite. Therefore, the NPMLEs exist. The key part of the proof of consistency is to establish the boundedness of $\hat{H}_n(\tau)$. Then we can construct a limit, $(\boldsymbol{\beta}^*, \boldsymbol{\gamma}^*, H^*)$, for the sequences $(\hat{\boldsymbol{\beta}}_n, \hat{\gamma}_n, \hat{H}_n)$ and prove that the limit is actually $(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0, H_0)$ by using the Kullback-Leibler information.

Proof. Suppose two sets of parameters $(\boldsymbol{\beta}, \gamma, H)$ and $(\tilde{\boldsymbol{\beta}}, \tilde{\gamma}, \tilde{H})$ give the same likelihood function for any observed data. That is,

$$\begin{aligned}
&\left[\frac{H'(Y)\left\{e^{\boldsymbol{\beta}^T\mathbf{X}}H^2(Y) + 2e^{\boldsymbol{\beta}^T\mathbf{X}}H(Y) + e^{\boldsymbol{\gamma}^T\mathbf{X}}\right\}}{\{H(Y)+1\}\left\{e^{\boldsymbol{\beta}^T\mathbf{X}}H^2(Y) + (e^{\boldsymbol{\gamma}^T\mathbf{X}}+1)H(Y)+1\right\}}\right]^{\Delta} \\
&\times \frac{H(Y)+1}{e^{\boldsymbol{\beta}^T\mathbf{X}}H^2(Y) + (e^{\boldsymbol{\gamma}^T\mathbf{X}}+1)H(Y)+1} \\
=&\left[\frac{\tilde{H}'(Y)\left\{e^{\tilde{\boldsymbol{\beta}}^T\mathbf{X}}\tilde{H}^2(Y) + 2e^{\tilde{\boldsymbol{\beta}}^T\mathbf{X}}\tilde{H}(Y) + e^{\tilde{\boldsymbol{\gamma}}^T\mathbf{X}}\right\}}{\{\tilde{H}(Y)+1\}\left\{e^{\tilde{\boldsymbol{\beta}}^T\mathbf{X}}\tilde{H}^2(Y) + (e^{\tilde{\boldsymbol{\gamma}}^T\mathbf{X}}+1)\tilde{H}(Y)+1\right\}}\right]^{\Delta} \\
&\times \frac{\tilde{H}(Y)+1}{e^{\tilde{\boldsymbol{\beta}}^T\mathbf{X}}\tilde{H}^2(Y) + (e^{\tilde{\boldsymbol{\gamma}}^T\mathbf{X}}+1)\tilde{H}(Y)+1}.
\end{aligned} \tag{2.3}$$

Since we assume (2.3) holds for any Y and $\mathbf{X}$, setting $\Delta = 1$, $Y = 0$ and $\mathbf{X} = \mathbf{0}$, we have

$H'(0) = \tilde{H}'(0)$. And for any $\mathbf{X} = \mathbf{0}$ with $\Delta = 1$, $Y = 0$, we have $e^{\boldsymbol{\gamma}^T \mathbf{X}} = e^{\tilde{\boldsymbol{\gamma}}^T \mathbf{X}}$, i.e.

$$(\boldsymbol{\gamma} - \tilde{\boldsymbol{\gamma}})^T \mathbf{X} = 0.$$

By (C3), we obtain $\boldsymbol{\gamma} = \tilde{\boldsymbol{\gamma}}$. With $\Delta = 0$ and $\mathbf{X} = 0$, for any $Y = y$, we have $H(y) = \tilde{H}(y)$. It follows that $e^{\boldsymbol{\beta}^T \mathbf{X}} = e^{\tilde{\boldsymbol{\beta}}^T \mathbf{X}}$, which implies $\boldsymbol{\beta} = \tilde{\boldsymbol{\beta}}$ with (C3). Therefore, the model is identifiable.

By constructing a new step function $\tilde{H}_n(t) = \hat{H}_n(t)/\hat{H}_n(\tau)$, we show that its corresponding likelihood will be larger than the likelihood of the NPMLEs if $\hat{H}_n(\tau)$ goes to infinity. In fact, let $l_n(\boldsymbol{\beta}, \boldsymbol{\gamma}, H)$ denote the log-likelihood with sample size n. We have

$$\begin{aligned}
&\frac{1}{n}\left[l_n(\hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{\gamma}}_n, \hat{H}_n) - l_n(\hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{\gamma}}_n, \tilde{H}_n)\right] = \frac{1}{n}\left[l_n(\hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{\gamma}}_n, \hat{H}_n(\tau)\tilde{H}_n) - l_n(\hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{\gamma}}_n, \tilde{H}_n)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\left\{\Delta_i \log\left[\frac{\hat{H}_n(\tau)\left\{e^{\hat{\boldsymbol{\beta}}_n^T \mathbf{X}_i}\hat{H}_n^2(\tau)\tilde{H}_n^2(Y_i) + 2e^{\hat{\boldsymbol{\beta}}_n^T \mathbf{X}_i}\hat{H}_n(\tau)\tilde{H}_n(Y_i) + e^{\hat{\boldsymbol{\gamma}}_n^T \mathbf{X}_i}\right\}}{\{\hat{H}_n(\tau)\tilde{H}_n(Y_i)+1\}\left\{e^{\hat{\boldsymbol{\beta}}_n^T \mathbf{X}_i}\hat{H}_n^2(\tau)\tilde{H}_n^2(Y_i) + (e^{\hat{\boldsymbol{\gamma}}_n^T \mathbf{X}_i}+1)\hat{H}_n(\tau)\tilde{H}_n(Y_i)+1\right\}}\right.\right. \\
&\qquad \left.\frac{\{\tilde{H}_n(Y_i)+1\}\left\{e^{\hat{\boldsymbol{\beta}}_n^T \mathbf{X}_i}\tilde{H}_n^2(Y_i) + (e^{\hat{\boldsymbol{\gamma}}_n^T \mathbf{X}_i}+1)\tilde{H}_n(Y_i)+1\right\}}{\left\{e^{\hat{\boldsymbol{\beta}}_n^T \mathbf{X}_i}\tilde{H}_n^2(Y_i) + 2e^{\hat{\boldsymbol{\beta}}_n^T \mathbf{X}_i}\tilde{H}_n(Y_i) + e^{\hat{\boldsymbol{\gamma}}_n^T \mathbf{X}_i}\right\}}\right] \\
&\qquad + \log\left[\frac{\hat{H}_n(\tau)\tilde{H}_n(Y_i)+1}{e^{\hat{\boldsymbol{\beta}}_n^T \mathbf{X}_i}\hat{H}_n^2(\tau)\tilde{H}_n^2(Y_i) + (e^{\hat{\boldsymbol{\gamma}}_n^T \mathbf{X}_i}+1)\hat{H}_n(\tau)\tilde{H}_n(Y_i)+1}\right. \\
&\qquad \left.\left.\frac{e^{\hat{\boldsymbol{\beta}}_n^T \mathbf{X}_i}\tilde{H}_n^2(Y_i) + (e^{\hat{\boldsymbol{\gamma}}_n^T \mathbf{X}_i}+1)\tilde{H}_n(Y_i)+1}{\tilde{H}_n(Y_i)+1}\right]\right\}.
\end{aligned}$$

Given the compactness of the parameter space and the boundedness of covariates and $\tilde{H}_n$, it is easy to show when $\hat{H}_n(\tau)$ goes to infinity, the first log term goes to a bounded limit, and the second log term goes to $-\infty$. That is,

$$0 \le \frac{1}{n}\left[l_n(\hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{\gamma}}_n, \hat{H}_n) - l_n(\hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{\gamma}}_n, \tilde{H}_n)\right] \to -\infty, \text{ as } n \to \infty.$$

By the contradiction, $\hat{H}_n(\tau)$ is bounded almost everywhere. In addition, $\tilde{H}_n(t)$ is bounded by 1. This holds for arbitrary n.

Hence, it follows from compactness theorems that there exist convergent subsequences of ($\hat{\boldsymbol{\beta}}_n$, $\hat{\boldsymbol{\gamma}}_n$, $\hat{H}_n$) with limits ($\boldsymbol{\beta}^*$, $\boldsymbol{\gamma}^*$, H^*).

We next show that $\boldsymbol{\beta}^* = \boldsymbol{\beta}_0$, $\boldsymbol{\gamma}^* = \boldsymbol{\gamma}_0$, and $H^* = H_0$. We begin with constructing a new sequence of step functions $\bar{H}_n(t)$ with jumps at those Y_i's for which $\Delta_i = 1$. We define the jump size as

$$\frac{\Delta_i}{\bar{H}_n\{Y_i\}} = \sum_{k=1}^{n} I(Y_k \ge Y_i)\left[\frac{\Delta_k A_k(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0, H_0) + \left\{e^{\boldsymbol{\beta}_0^T \mathbf{X}_k} H_0^2(Y_k) + 2e^{\boldsymbol{\beta}_0^T \mathbf{X}_k} H_0(Y_k) + e^{\boldsymbol{\gamma}_0^T \mathbf{X}_k}\right\}^2}{B_k(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0, H_0)}\right],$$

where

$$\begin{aligned} A_k(\boldsymbol{\beta}, \boldsymbol{\gamma}, H) =& e^{2\boldsymbol{\beta}^T \mathbf{X}_k} H^4(Y_k) + 4e^{2\boldsymbol{\beta}^T \mathbf{X}_k} H^3(Y_k) + (2e^{2\boldsymbol{\beta}^T \mathbf{X}_k} + 4e^{\boldsymbol{\beta}^T \mathbf{X}_k + \boldsymbol{\gamma}^T \mathbf{X}_k}) H^2(Y_k) \\ &+ (2e^{\boldsymbol{\beta}^T \mathbf{X}_k + \boldsymbol{\gamma}^T \mathbf{X}_k} - 2e^{\boldsymbol{\beta}^T \mathbf{X}_k} + 2e^{2\boldsymbol{\gamma}^T \mathbf{X}_k} + 2e^{\boldsymbol{\gamma}^T \mathbf{X}_k}) H(Y_k) \\ &- 2e^{\boldsymbol{\beta}^T \mathbf{X}_k} + e^{2\boldsymbol{\gamma}^T \mathbf{X}_k} + 2e^{\boldsymbol{\gamma}^T \mathbf{X}_k} \end{aligned}$$

and

$$\begin{aligned} B_k(\boldsymbol{\beta}, \boldsymbol{\gamma}, H) =& (H(Y_k) + 1)\left\{e^{\boldsymbol{\beta}^T \mathbf{X}_k} H^2(Y_k) + (e^{\boldsymbol{\gamma}^T \mathbf{X}_k} + 1) H(Y_k) + 1\right\} \\ &\times \left\{e^{\boldsymbol{\beta}^T \mathbf{X}_k} H^2(Y_k) + 2e^{\boldsymbol{\beta}^T \mathbf{X}_k} H(Y_k) + e^{\boldsymbol{\gamma}^T \mathbf{X}_k}\right\}. \end{aligned}$$

The above equation is derived similarly from the score functions of $l_n(\boldsymbol{\beta}, \boldsymbol{\gamma}, H)$ with respect to $H\{Y_i\}$ as follows:

$$\frac{\Delta_i}{\hat{H}_n\{Y_i\}} = \sum_{k=1}^{n} I(Y_k \geq Y_i) \left[\frac{\Delta_k A_k(\hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{\gamma}}_n, \hat{H}_n) + \left\{ e^{\hat{\boldsymbol{\beta}}_n^T \mathbf{x}_k} \hat{H}_n^2(Y_k) + 2e^{\hat{\boldsymbol{\beta}}_n^T \mathbf{x}_k} \hat{H}_n(Y_k) + e^{\hat{\boldsymbol{\gamma}}_n^T \mathbf{x}_k} \right\}^2}{B_k(\hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{\gamma}}_n, \hat{H}_n)} \right],$$

where $\bar{H}_n(t) = \sum_{i=1}^n \bar{H}_n\{Y_i\} I(Y_i \leq t)$ and $\hat{H}_n(t) = \sum_{i=1}^n \hat{H}_n\{Y_i\} I(Y_i \leq t)$. Let

$$M_k(\boldsymbol{\beta}, \boldsymbol{\gamma}, H) = \frac{\Delta_k A_k(\boldsymbol{\beta}, \boldsymbol{\gamma}, H) + \left\{ e^{\boldsymbol{\beta}^T \mathbf{x}_k} H^2(Y_k) + 2e^{\boldsymbol{\beta}^T \mathbf{x}_k} H(Y_k) + e^{\boldsymbol{\gamma}^T \mathbf{x}_k} \right\}^2}{B_k(\boldsymbol{\beta}, \boldsymbol{\gamma}, H)}.$$

We have

$$\hat{H}_n(t) = \int_0^t \frac{\sum_{k=1}^n I(Y_k \geq Y_i) M_k(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0, H_0)}{\sum_{k=1}^n I(Y_k \geq Y_i) M_k(\hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{\gamma}}_n, \hat{H}_n)} d\bar{H}_n(t). \tag{2.4}$$

Therefore $\hat{H}_n(t)$ is absolutely continuous with respect to $\bar{H}_n(t)$. If we could verify that $\bar{H}_n(t)$ converges uniformly to $H_0(t)$ in $[0, \tau]$ with probability 1, then in (2.4), letting $n \to \infty$, we obtain

$$H_n^*(t) = \int_0^t \frac{\sum_{k=1}^n I(Y_k \geq Y_i) M_k(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0, H_0)}{\sum_{k=1}^n I(Y_k \geq Y_i) M_k(\boldsymbol{\beta}^*, \boldsymbol{\gamma}^*, H^*)} dH_0(t).$$

Therefore $H_n^*(t)$ is differentiable with respect to $H_0(t)$. It follows that $d\hat{H}_n(t)/d\bar{H}_n(t)$ converges uniformly to $dH^*(t)/dH_0(t)$ in $[0, \tau]$. Here we introduce a lemma before the following proof.

Lemma 2.1. The class

$$\mathcal{F} = \{I(Y_1 \geq y) M_1(\boldsymbol{\beta}, \boldsymbol{\gamma}, H) : y \in [0, \tau], (\boldsymbol{\beta}, \boldsymbol{\gamma}) \in \mathcal{B}, H \in \mathcal{H}\}$$

is a bounded p-Donsker class, therefore a Glivenko-Cantelli class, where $\mathcal{B}$ is a compact set and

$$\mathcal{H} = \{h : h \text{ is increasing in } [0,\tau] \text{ with } h(0) = 0, h(\tau) < B_0$$

$$\text{where } B_0 \text{ is a positive constant such that } \hat{H}_n(\tau) \leq B_0 \text{ with probability one }\}.$$

This lemma can be proved by definition. Recall that

$$\mathcal{F} = \{I(Y_1 \geq y)M_1(\boldsymbol{\beta}, \boldsymbol{\gamma}, H) : y \in [0,\tau], (\boldsymbol{\beta}, \boldsymbol{\gamma}) \in \mathcal{B}, H \in \mathcal{H}\}.$$

First we prove that

$$\mathcal{F}^* = \{M_1(\boldsymbol{\beta}, \boldsymbol{\gamma}, H) : y \in [0,\tau], (\boldsymbol{\beta}, \boldsymbol{\gamma}) \in \mathcal{B}, H \in \mathcal{H}\}$$

is a bounded p-Donsker class. And $\{I(Y_1 \leq y), y \in [0,\tau]\}$ is uniformly bounded and monotone on the real line, therefore is p-Donsker. By the preservation of Donsker property under product, we can prove Lemma 2.1.

It remains to prove that $\mathcal{F}^*$ is a bounded p-Donsker class. Despite the complex expression for $M_1(\boldsymbol{\beta}, \boldsymbol{\gamma}, H)$, it is easy to know that it is bounded and continuously differentiable with respect to $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ for any $(\boldsymbol{\beta}, \boldsymbol{\gamma}) \in \mathcal{B}$ and the norm of the derivative are also bounded. It is also dominated by H. Therefore, by the mean-value theorem, $|M_1(\boldsymbol{\beta}, \boldsymbol{\gamma}, H) - M_1(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\gamma}}, \tilde{H})|$ is dominated by $||\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}|| + ||\boldsymbol{\gamma} - \tilde{\boldsymbol{\gamma}}|| + \sup_{t \in (0,\tau)} |H(t) - \tilde{H}(t)|$. Then $\mathcal{F}^*$ is a p-Donsker class (van der Vaart and Wellner, 1996). It follows that $\mathcal{F}$ is a p-Donsker class. This completes the proof of Lemma 2.1.

By the Glivenko-Cantelli theorem given by van der Vaart and Wellner (1996),

$$\bar{H}_n(t) \to E\left[\frac{I(Y_1 \leq t)\Delta_1}{\mu(Y_1)}\right] \text{ uniformly, where } \mu(y) = E\left[I(Y_1 \geq y)M_1(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0, H_0)\right].$$

Let $S_c(\cdot|\mathbf{X})$ be the survival function for the censoring time. We can show that

$$\mu(y) = E\left[\frac{H'(y)\left\{e^{\boldsymbol{\beta}_0^T\mathbf{X}}H_0^2(y) + 2e^{\boldsymbol{\beta}_0^T\mathbf{X}}H_0(y) + e^{\boldsymbol{\gamma}_0^T\mathbf{X}}\right\}S_c(y|\mathbf{X})}{\{H_0(y)+1\}\left\{e^{\boldsymbol{\beta}_0^T\mathbf{X}}H_0^2(y) + (e^{\boldsymbol{\gamma}_0^T\mathbf{X}}+1)H_0(y)+1\right\}}\right].$$

Then we have

$$\begin{aligned}
E\left[\frac{I(Y_1 \le t)\Delta_1}{\mu(Y_1)}\right] =& E\left[\int_0^t \frac{H'(y)\left\{e^{\boldsymbol{\beta}_0^T\mathbf{X}}H_0^2(y) + 2e^{\boldsymbol{\beta}_0^T\mathbf{X}}H_0(y) + e^{\boldsymbol{\gamma}_0^T\mathbf{X}}\right\}S_c(y|\mathbf{X})}{\mu(y)\{H_0(y)+1\}\left\{e^{\boldsymbol{\beta}_0^T\mathbf{X}}H_0^2(y) + (e^{\boldsymbol{\gamma}_0^T\mathbf{X}}+1)H_0(y)+1\right\}}dH_0(y)\right] \\
=& \int_0^t dH_0(y) = H_0(t).
\end{aligned}$$

Thus, $\bar{H}_n(t)$ converges uniformly to $H_0(t)$ in $[0,\tau]$ with probability 1. By the argument above, $d\hat{H}_n(t)/d\bar{H}_n(t)$ converges uniformly to $dH^*(t)/dH_0(t)$ in $[0,\tau]$.

On the other hand,

$$\begin{aligned}
& \frac{1}{n}\left[l_n(\hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{\gamma}}_n, \hat{H}_n) - l_n(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0, \bar{H}_n)\right] \\
= \quad & \frac{1}{n}\sum_{i=1}^n \left\{\Delta_i \log\left[\frac{\hat{H}_n\{Y_i\}\left\{e^{\hat{\boldsymbol{\beta}}_n^T\mathbf{X}_i}\hat{H}_n^2(Y_i) + 2e^{\hat{\boldsymbol{\beta}}_n^T\mathbf{X}_i}\hat{H}_n(Y_i) + e^{\hat{\boldsymbol{\gamma}}_n^T\mathbf{X}_i}\right\}}{\{\hat{H}_n(Y_i)+1\}\left\{e^{\hat{\boldsymbol{\beta}}_n^T\mathbf{X}_i}\hat{H}_n^2(Y_i) + (e^{\hat{\boldsymbol{\gamma}}_n^T\mathbf{X}_i}+1)\hat{H}_n(Y_i)+1\right\}}\right.\right. \\
& \left.\frac{\{\bar{H}_n(Y_i)+1\}\left\{e^{\boldsymbol{\beta}_0^T\mathbf{X}_i}\bar{H}_n^2(Y_i) + (e^{\boldsymbol{\gamma}_0^T\mathbf{X}_i}+1)\bar{H}_n(Y_i)+1\right\}}{\bar{H}_n\{Y_i\}\left\{e^{\boldsymbol{\beta}_0^T\mathbf{X}_i}\bar{H}_n^2(Y_i) + 2e^{\boldsymbol{\beta}_0^T\mathbf{X}_i}\bar{H}_n(Y_i) + e^{\boldsymbol{\gamma}_0^T\mathbf{X}_i}\right\}}\right] \\
& + \log\left[\frac{\hat{H}_n(Y_i)+1}{e^{\hat{\boldsymbol{\beta}}^T\mathbf{X}_i}\hat{H}_n^2(Y_i) + (e^{\hat{\boldsymbol{\gamma}}^T\mathbf{X}_i}+1)\hat{H}_n(Y_i)+1}\right.
\end{aligned}$$

$$\left.\left.\frac{e^{\boldsymbol{\beta}_0^T\mathbf{X}_i}\bar{H}_n^2(Y_i)+(e^{\boldsymbol{\gamma}_0^T\mathbf{X}_i}+1)\bar{H}_n(Y_i)+1}{\bar{H}_n(Y_i)+1}\right]\right\}$$

$$\geq \quad 0.$$

Letting $n\to\infty$, given $\mathbf{X}$, we have

$$E\left[\frac{1}{n}\sum_{i=1}^{n}\lim_{n\to\infty}\left\{l_{n,i}(\hat{\boldsymbol{\beta}}_n,\hat{\boldsymbol{\gamma}}_n,\hat{H}_n)-l_{n,i}(\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0,\bar{H}_n)\right\}\right]$$

$$= E\left\{\log\left[\left\{\frac{H^{*\prime}(Y_1)\left\{e^{\boldsymbol{\beta}^{*T}\mathbf{X}}H^{*2}(Y_1)+2e^{\boldsymbol{\beta}^{*T}\mathbf{X}}H^{*}(Y_1)+e^{\boldsymbol{\gamma}^{*T}\mathbf{X}}\right\}}{(H^*(Y_1)+1)P_1(\boldsymbol{\beta}^*,\boldsymbol{\gamma}^*,H^*)}\right\}^{\Delta_1}\left\{\frac{H^*(Y_1)+1}{P_1(\boldsymbol{\beta}^*,\boldsymbol{\gamma}^*,H^*)}\right\}\right.\right.$$

$$\left.\left.\Big/\left\{\frac{H_0'(Y_1)\left\{e^{\boldsymbol{\beta}_0^T\mathbf{X}}H_0^2(Y_1)+2e^{\boldsymbol{\beta}_0^T\mathbf{X}}H_0(Y_1)+e^{\boldsymbol{\gamma}_0^T\mathbf{X}}\right\}}{(H_0(Y_1)+1)P_1(\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0,H_0)}\right\}^{\Delta_1}\left\{\frac{H_0(Y_1)+1}{P_1(\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0,H_0)}\right\}\right]\right\}$$

$$\geq 0,$$

where $l_{n,i}(\boldsymbol{\beta},\boldsymbol{\gamma},H)$ is the log-likelihood contributed from the ith subject and

$$P_1(\boldsymbol{\beta},\boldsymbol{\gamma},H)=e^{\boldsymbol{\beta}^T\mathbf{X}}H^2(Y_1)+(e^{\boldsymbol{\gamma}^T\mathbf{X}}+1)H(Y_1)+1.$$

This expectation is the negative Kullback-Leibler information, which is non-positive. As a result, it equals zero. Consequently, we have

$$\left[\frac{H^{*\prime}(Y_1)\left\{e^{\boldsymbol{\beta}^{*T}\mathbf{X}}H^{*2}(Y_1)+2e^{\boldsymbol{\beta}^{*T}\mathbf{X}}H^{*}(Y_1)+e^{\boldsymbol{\gamma}^{*T}\mathbf{X}}\right\}}{(H^*(Y_1)+1)P_1(\boldsymbol{\beta}^*,\boldsymbol{\gamma}^*,H^*)}\right]^{\Delta_1}\left\{\frac{H^*(Y_1)+1}{P_1(\boldsymbol{\beta}^*,\boldsymbol{\gamma}^*,H^*)}\right\}$$

$$=\left[\frac{H_0'(Y_1)\left\{e^{\boldsymbol{\beta}_0^T\mathbf{X}}H_0^2(Y_1)+2e^{\boldsymbol{\beta}_0^T\mathbf{X}}H_0(Y_1)+e^{\boldsymbol{\gamma}_0^T\mathbf{X}}\right\}}{(H_0(Y_1)+1)P_1(\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0,H_0)}\right]^{\Delta_1}\left\{\frac{H_0(Y_1)+1}{P_1(\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0,H_0)}\right\}.$$

Using the same argument in the proof of identifiability, we have $\boldsymbol{\beta}^* = \boldsymbol{\beta}_0$, $\boldsymbol{\gamma}^* = \boldsymbol{\gamma}_0$, and $H^* = H_0$.

Therefore, $||\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0|| \to 0$, $||\hat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_0|| \to 0$ and $|\hat{H}_n(t) - H_0(t)| \to 0$ pointwise, almost surely as $n \to \infty$. Since H_0 is a continuous function, we have further uniform convergence $\sup_{t\in[0,\tau]} |\hat{H}_n(t) - H_0(t)| \to 0$ almost surely as $n \to \infty$.

□

We next establish the asymptotic normality of the NPMLEs. Let $l^\infty[0,\tau]$ be a space consisting of all uniformly bounded functions on $[0,\tau]$ and the norm on it is defined as the supremum absolute value on $[0,\tau]$.

Theorem 2.2. Under conditions C1-C5, $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0, \hat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_0, \hat{H}_n - H_0)^T$ converges weakly to a Gaussian process with mean zero in the metric space $l^\infty(\mathcal{A})$ where

$$\mathcal{A} = \{(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, g) : \boldsymbol{\theta}_1 \in \mathcal{R}^p, \boldsymbol{\theta}_2 \in \mathcal{R}^p, g \text{ is a function on } [0,\tau],$$

$$\text{and } ||\boldsymbol{\theta}_1|| \leq 1, ||\boldsymbol{\theta}_2|| \leq 1, |g|_V \leq 1,$$

$$\text{where } |g|_V \text{ is the total variation of } g \text{ on } [0,\tau]\}.$$

Proof. First, we introduce some notation in the context of empirical process. Let $\mathbb{P}_n$ and $\mathbb{P}$ be the empirical measure and the population distribution of n i.i.d. observations $\mathbf{O}_1, \mathbf{O}_2, \cdots, \mathbf{O}_n$. Let $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - \mathbb{P})$ denotes the empirical process. Then for any measurable function h,

$$\mathbb{P}_n h = \frac{1}{n}\sum_{i=1}^n h(\mathbf{O}_i), \mathbb{P}h = E_{\mathbb{P}}[h],$$

and

$$\mathbb{G}_n h = \sqrt{n}(\mathbb{P}_n h - \mathbb{P}h).$$

By the consistency theorem, $||\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0|| + ||\hat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_0|| + \sup_{t\in[0,\tau]} |\hat{H}_n(t) - H_0(t)| < \epsilon$ with probability 1 when sample size n is large enough. Therefore we consider a neighborhood of $(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0, H_0)$,

$$\mathcal{U} = \{(\boldsymbol{\beta}, \boldsymbol{\gamma}, H) : ||\boldsymbol{\beta} - \boldsymbol{\beta}_0|| + ||\boldsymbol{\gamma} - \boldsymbol{\gamma}_0|| + \sup_{t \in [0,\tau]} |H(t) - H_0(t)| < \epsilon\}$$

for a small constant ϵ, and define a sequence of functions K_n mapping $\mathcal{U}$ to $l^\infty(\mathcal{A})$ as

$$K_n(\boldsymbol{\beta}, \boldsymbol{\gamma}, H)[\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, g] = \frac{1}{n}\frac{d}{d\epsilon} l_n \left(\boldsymbol{\beta} + \epsilon\boldsymbol{\theta}_1, \boldsymbol{\gamma} + \epsilon\boldsymbol{\theta}_2, H(t) + \epsilon \int_0^t g dH(s)\right)\bigg|_{\epsilon=0}.$$

Particularly, consider a single observation $(Y, \Delta, \mathbf{X})$, we have

$$\begin{aligned}
& K(\boldsymbol{\beta}, \boldsymbol{\gamma}, H)[\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, g] \\
&= \Delta \left[g(Y) + \frac{e^{\boldsymbol{\beta}^T \mathbf{X}} \left\{ \boldsymbol{\theta}_1^T \mathbf{X} H^2(Y) + 2\boldsymbol{\theta}_1^T \mathbf{X} H(Y) + 2(H(Y)+1) E_H g(Y) \right\}}{e^{\boldsymbol{\beta}^T \mathbf{X}} H(Y) \left(H(Y)+2\right) + e^{\boldsymbol{\gamma}^T \mathbf{X}}} \right. \\
&\qquad \left. + \frac{e^{\boldsymbol{\gamma}^T \mathbf{X}} \boldsymbol{\theta}_2^T \mathbf{X}}{e^{\boldsymbol{\beta}^T \mathbf{X}} H(Y) \left(H(Y)+2\right) + e^{\boldsymbol{\gamma}^T \mathbf{X}}} \right] \\
&- (\Delta+1) \left[\frac{e^{\boldsymbol{\beta}^T \mathbf{X}} \left(\boldsymbol{\theta}_1^T \mathbf{X} H^2(Y) + 2H(Y) E_H g(Y) \right)}{e^{\boldsymbol{\beta}^T \mathbf{X}} H^2(Y) + (e^{\boldsymbol{\gamma}^T \mathbf{X}} + 1) H(Y) + 1} \right. \\
&\qquad \left. + \frac{e^{\boldsymbol{\gamma}^T \mathbf{X}} \left(\boldsymbol{\theta}_2^T \mathbf{X} H(Y) + E_H g(Y) \right) + E_H g(Y)}{e^{\boldsymbol{\beta}^T \mathbf{X}} H^2(Y) + (e^{\boldsymbol{\gamma}^T \mathbf{X}} + 1) H(Y) + 1} \right] \\
&- (\Delta - 1) \frac{E_H g(Y)}{H(Y)+1},
\end{aligned} \tag{2.5}$$

where $E_H g(Y) = \int_0^Y g(s) dH(s)$.

Essentially, the empirical process $\mathbb{G}_n K$ also maps from $\mathcal{U}$ to $l^\infty(\mathcal{A})$, and

$$\mathbb{P}_n K(\hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{\gamma}}_n, \hat{H}_n)[\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, g] = 0, \ \mathbb{P} K(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0, H_0)[\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, g] = 0.$$

This is implied from the definition of K_n and the consistency of NPMLEs.

To prove the asymptotic normality, we verify three Properties $a) - c)$ below. Then it follows from Theorem 3.3.1 of van der Vaart and Wellner (1996) that $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0, \hat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_0, \hat{H}_n - H_0)^T$ converges in distribution to a Gaussian process in $l^\infty(\mathcal{A})$.

a) $\mathbb{G}_n K(\hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{\gamma}}_n, \hat{H}_n)[\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, g] - \mathbb{G}_n K(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0, H_0)[\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, g] = o_{\mathbb{P}}[1 + \sqrt{n}(||\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0|| + ||\hat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_0|| + \sup_{t \in [0,\tau]} |\hat{H}_n(t) - H_0(t)|)]$.

b) $\mathbb{G}_n K(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0, H_0)[\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, g]$ converges to a tight random element $\boldsymbol{\xi}$.

c) $\mathbb{P} K(\boldsymbol{\beta}, \boldsymbol{\gamma}, H)[\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, g]$ is Frechet-differentiable at $(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0, H_0)$ with a continuously invertible derivative $(\dot{\mathbb{P}K})(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0, H_0)$.

Note that we assume that $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ belong to the interior of a known compact set $\mathcal{B}_0$ in $\mathcal{R}^{2p}$, $H(\tau) < \infty$, and $\mathbf{X}$ are bounded. Therefore it is easy to see from the explicit expression (2.5) where $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, g) \in \mathcal{A}$ that $K(\boldsymbol{\beta}, \boldsymbol{\gamma}, H)$ is continuously differentiable with respect to $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ and the norm of the derivative is bounded. Furthermore, $K(\boldsymbol{\beta}, \boldsymbol{\gamma}, H_1) - K(\boldsymbol{\beta}, \boldsymbol{\gamma}, H_2)$ is dominated by

$$\left\{ |H_1(Y) - H_2(Y)| + \int_0^Y d|H_1(s) - H_2(s)| \right\}.$$

Therefore, we have

$$\sup_{(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, g) \in \mathcal{A}} \mathbb{P} \left(K(\boldsymbol{\beta}, \boldsymbol{\gamma}, H)[\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, g] - K(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0, H_0)[\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, g] \right)^2 \to 0,$$

as $||\boldsymbol{\beta} - \boldsymbol{\beta}_0|| + ||\boldsymbol{\gamma} - \boldsymbol{\gamma}_0|| + \sup_{t \in [0,\tau]} |H(t) - H_0(t)| \to 0$. Using the similar arguments in the proof of Lemma 2.1, we prove that the class

$$\{K(\boldsymbol{\beta},\boldsymbol{\gamma},H)[\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,g]-K(\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0,H_0)[\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,g]:(\boldsymbol{\beta},\boldsymbol{\gamma},H)\in\mathcal{U},(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,g)\in\mathcal{A}\}$$

is p-Donsker. Therefore, by applying Lemma 3.3.5 of van der Vaart and Wellner (1996), Property a) holds.

Similarly, we can prove that the class $\{K(\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0,H_0)[\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,g]:(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,g)\in\mathcal{A}\}$ is p-Donsker. Then Property b) holds. In fact, $\boldsymbol{\xi}$ is a Gaussian process indexed by $(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,g)\in\mathcal{A}$. The covariance between $\boldsymbol{\xi}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,g)$ and $\boldsymbol{\xi}(\boldsymbol{\theta}_1^*,\boldsymbol{\theta}_2^*,g^*)$ is denoted by

$$\mathbb{P}\left(K(\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0,H_0)[\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,g]\times K(\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0,H_0)[\boldsymbol{\theta}_1^*,\boldsymbol{\theta}_2^*,g^*]\right).$$

The smoothness of $\mathbb{P}(\boldsymbol{\beta},\boldsymbol{\gamma},H)$ implies the Frechet differentiability in Property c). According to the proof of Theorem 2 in Zeng and Lin (2007), to show the continuously invertibility of $(\dot{\mathbb{P}K})$ can be implemented by showing that $(\dot{\mathbb{P}K})$ is one-to-one; that is,

$$K(\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0,H_0)[\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,g]=0$$

implies $\boldsymbol{\theta}_1=0,\boldsymbol{\theta}_2=0,g=0$. In fact, to prove the continuously invertibility, it suffices to prove the Fisher information along the path $\boldsymbol{\beta}_0+\epsilon\boldsymbol{\theta}_1,\boldsymbol{\gamma}_0+\epsilon\boldsymbol{\theta}_2,H_0+\epsilon E_{H_0}g$ is nonsingular. This can be proved by using the idea of proving the identifiability.

Let $\Delta=1,Y=0$, we have $\boldsymbol{\theta}_2^T\mathbf{X}+g(0)=0$. By Condition (C3), we have $\boldsymbol{\theta}_2=0$ and $g(0)=0$. Let $\Delta=0,\mathbf{X}=\mathbf{0}$, we obtain $E_Hg=0$. While when $\Delta=1,\mathbf{X}=\mathbf{0}$, $g-\frac{2E_Hg}{H+1}=0$. Thus, $g=0$. Plugging in $\boldsymbol{\theta}_2=\mathbf{0},g=0$, we have

$$\begin{aligned}&K(\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0,H_0)[\boldsymbol{\theta}_1,\mathbf{0},0]\\=&\Delta\frac{e^{\boldsymbol{\beta}_0^T\mathbf{X}}(\boldsymbol{\theta}_1^T\mathbf{X}H^2+2\boldsymbol{\theta}_1^T\mathbf{X}H)}{e^{\boldsymbol{\beta}_0^T\mathbf{X}}H(H+2)+e^{\boldsymbol{\gamma}_0^T\mathbf{X}}}-(\Delta+1)\frac{e^{\boldsymbol{\beta}_0^T\mathbf{X}}\boldsymbol{\theta}_1^T\mathbf{X}}{e^{\boldsymbol{\beta}_0^T\mathbf{X}}H^2+(e^{\boldsymbol{\gamma}_0^T\mathbf{X}}+1)H+1}.\end{aligned}$$

By letting $\Delta = 0$, we immediately have $\boldsymbol{\theta}_1 = \mathbf{0}$ by Condition (C3). Therefore, Property c) is proved.

It follows from Theorem 3.3.1 of van der Vaart and Wellner (1996) that $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0, \hat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_0, \hat{H}_n - H_0)^T$ converges in distribution to a tight Gaussian random element $(\mathbb{P}\dot{K})^{-1}\boldsymbol{\xi}$ in $l^\infty(\mathcal{A})$.

$\square$

Moreover, we can show that $(\hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{\gamma}}_n)$ is an asymptotic linear estimator for $(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0)$ and the corresponding influence functions are in the spanned space via score functions, which implies $(\hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{\gamma}}_n)$ is semiparametrically efficient (Bickel et al., 1993).

2.4 Simulation Studies

We conduct simulation studies to examine the finite-sample performance of our proposed methods. We generate the failure time from the following model

$$F(t|X) = \frac{e^{\beta X}t^2 + e^{\gamma X}t}{e^{\beta X}t^2 + e^{\gamma X}t + t + 1},$$

where X is set to be uniform$(-0.5, 0.5)$. Under the above model, $H(t) = t$. Results are obtained under four scenarios of regression parameters (β, γ): (0.5, 0.5), (0, 0.5), (-0.5, 0.5), and (-0.5, 0). The long-term and short-term effects are equal under the first scenario where the model reduces to the PO model; there is no long-term effect under scenario 2; the long-term and short-term effects are in opposite directions under scenario 3; and there is no short-term effect under scenario 4. We generated the censoring time from an exponential distribution with mean 6. The censoring time is truncated at 4. The censoring rates were about 30% under all scenarios. For each setting, we consider the sample sizes of 100 and 200 and generate 10000 replicates. The variance-covariance matrix of the NPMLEs is approximated by the inverse of the observed information matrix.

Tables 2.1 and 2.2 present the summary statistics of the NPMLEs of the unknown regression parameters and the baseline odds functions at the time points 0.3, 0.8, 2, and 4, the first three of which correspond to the approximate first quartile, median and third quartile of the failure time, and the last one is the end of the study. In the table, $Bias$ is the sampling average of the biases of the estimates; SEE denotes the average of the standard error estimates; SE denotes the empirical standard deviations of the parameter estimates; CP is the coverage probability of 95% confidence interval estimates; and PO denotes the average of the parameter estimates under the PO model.

From Tables 2.1 and 2.2, we observe that the biases of the NPMLEs are small under all simulation settings. The standard deviation estimate exhibit the true variation. The coverage probabilities of the 95% confidence interval estimates are close to the nominal level. As expected, as sample size increases, the coverage probabilities improve. The PO model yield biased results when the proportional odds assumption is violated. Particularly, when the long-term and short-term effects are in opposite directions, the PO model have little power to detect covariate effects.

The next set of simulation studies are concerned with the performance of our proposed procedures for testing the covariate effects and the proportional odds assumption. We use the same simulation settings as the previous simulation study except that the censoring time is from a uniform distribution in $(0, 20)$, yielding approximate censoring rates of 20% under all scenarios. We consider five hypothesis testings here: (H1) $H_0 : \beta = 0$; (H2) $H_0 : \gamma = 0$; (H3) $H_0 : \beta = \gamma = 0$; (H4) $H_0 : \beta = \gamma$; (H5) $H_0 : \beta_{PO} = 0$. The first four tests are conducted based on our proposed model. The last one is conducted based on the PO model. We use Wald test statistics for testing (H1), (H2), (H4), and (H5) and likelihood ratio test statistics for testing (H3). The data are generated under the same model as above with $n = 400$.

Table 2.3 presents the sizes/powers at the nominal significance level of 0.05 based on 10000 replicates. Our method provides accurate control of type I error rates. We also observe reasonable powers of the proposed tests for testing (H3) compared to the test using the PO model when the proportional odds assumption is true; the powers for testing the covariate effect under both scenarios $\beta = \gamma = -1$ and $\beta = \gamma = 1$ are approximately 84.2% and 90.7% using the proposed model and the PO model, respectively. However, when the proportional odds assumption is violated, the proposed tests tend to be more powerful than the test based on the PO model, especially when the long-term and short-term effects are in opposite directions. For example, when $(\beta, \gamma) = (-1, 1)$, the powers for testing (H1), (H2), and (H3) are 0.4553, 0.5213, and 0.5804, respectively, compared to the power of 0.0569 from the PO model. The proposed test for the proportional odds assumption also yielded reasonable powers; for example, we obtained powers of 0.6755 and 0.6713 for $(\beta, \gamma) = (-1, 1)$ and $(\beta, \gamma) = (1, -1)$, respectively.

In the last set of simulation studies, we compare the performance of our method with the method by Scheike (2006). We consider the simulation settings similar to those in Scheike (2006). Specifically, we generate data from the following model

$$\frac{F(t|X, Z)}{S(t|X, Z)} = \{A_0(t) + XA_1(t)\} \exp(Z\beta),$$

where Z is a standard normal random variable, X is a log-normal random variable truncated at 30 with mean $0.3Z$ and standard deviation 0.4, $A_0(t) = 0.1t$, and $A_1(t) = 0.05t^2$. We vary the value of β from -0.2 to 0.2. The failure times are censored at 12 yielding approximate censoring rates of 30%. For each setting, we consider the sample sizes of 200 and 400. We fit the proposed model with the constraint of equal short-term and long-term effects of Z and compare its performance with that of Scheike (2006).

Table 2.4 summarizes the results for estimating the effect of Z. The NPMLE of the effect of Z under the proposed model appears to have little bias suggesting that the proposed method is robust to moderate model mis-specifications. Compared to the method of Scheike (2006), the NPMLE has slightly larger biases but the empirical standard deviations are smaller. Additionally, the NPMLE appears to be more efficient in terms of the mean squared error since the method of Scheike (2006) involves the estimation of two unknown functions $A_0(t)$ and $A_1(t)$ instead of one unknown function $H(t)$ in the proposed model.

2.5 Example

In this Section, we apply the proposed methods to the motivating CNA study (Zhang et al., 2009) described in the Introduction Section. In this study, the copy number data for 313 lymph node-negative breast cancer patients who were selected from the tumor bank at the Erasmus Medical Center are available. Survival analysis on time to develop distant metastasis (DM) was conducted at 56859 SNP loci in the breast cancer genome. The follow up time ranged from 20 to 169 months with a median of 99 months. Among the 313 breast cancer patients, 114 patients developed distant metastasis yielding a censoring rate of 63.6%.

There are two types of estrogen receptor (ER) for tumors, ER positive or ER negative. Among 313 patients, 199 tumors were ER positive; and 161 patients experienced menopause. The ages (years) of patients at the time of surgery were recorded ranging from 26 to 83 with a median of 54. As suggested by Zhang et al. (2009), we conduct the analysis for ER positive group and ER negative group separately. For each group, we include surgery age and menopause status as confounders and evaluate the long-term and short-term effects of the CNV data at each of the 56859 SNP loci. As a common practice, we standardize all the covariates in the numerical implementation.

Using the proposed method, the proportional odds assumption holds at about 99.5% of SNP loci for age and about 66.4% for menopause status. Table 2.5 presents the results for the ER positive group. The first part of Table 2.5 presents the top ten SNPs with the smallest P-values for testing the proportional odds assumption. It evidently shows that the PO model is inappropriate when the long-term and short-term effects are in opposite directions. At these loci, the proposed methods are able to detect significant short-term and/or long-term genetic effects missed by the PO model. The second part of Table 2.5

presents the top ten SNPs with the most significant copy number effect under the PO model. The results demonstrate that our proposed model can also detect significant long-term and/or short-term effects at those SNP loci. In contrast to the first part, we present the first 10 SNPs with the largest P-values for testing the proportional odds assumption in the third part of Table 2.5. The proposed methods yield short-term and long-term effects similar to the estimates obtained from the PO model, suggesting that the proposed methods work well when the proportional odds assumption is true.

As described in the Introduction Section, the proportional odds assumption appears to be true at SNP #2124, but is violated at SNP #39687. In Table 2.5, at SNP #2124, the estimates for the long-term and short-term effects were both -0.203 with standard error estimates of 0.285 and 0.242, respectively. The estimate for the effect under the PO model was -0.193, which is close to the estimates under the proposed model. At SNP #39687, we obtained significant long-term and short-term effects in opposite directions with estimates 1.482 and -1.130. However, the PO model yielded insignificant results with regression parameter estimate of 0.006 (P-value = 0.967). These results suggest that in this case the proposed model works as well as the PO model when the proportional odds assumption is true and outperforms the PO model when the assumption is violated.

We also calculated the adjusted odds ratio comparing patients with copy number value of 1 against 0 while fixing the two confounders at their sample means. Figure 4.1 displays the estimated curves of the log odds ratios with their pointwise 95% confidence bands at SNP #2124 and SNP #39687. It appears that the log odds ratio curve at SNP #2124 is close to a horizontal line whereas the log odds ratio curve at SNP #39687 is non-decreasing and crosses 0 around 60 months suggesting violation of the proportional odds assumption.

2.6 Discussion

We have proposed a novel odds rate model incorporating short-term and long-term covariate effects and have established efficient likelihood-based nonparametric estimation and inference procedures. The proposed model allows one to explicitly model the long-term and short-term covariate effects on the odds of survival without the use of smoothing. The simulation studies demonstrate that our model can detect the short-term and long-term effects when the proportional odds assumption is violated, while the PO model is inappropriate in this circumstance.

One limitation of the proposed method is that we assume inexplicitly that the odds ratio function is monotonic over time. Consider two univariate covariates $X > X^*$. It can be shown that

$$\begin{aligned} \frac{F(t|X)/S(t|X)}{F(t|X^*)/S(t|X^*)} &= e^{\beta(X-X^*)}F(t) + e^{\gamma(X-X^*)}S(t) \\ &= e^{\beta(X-X^*)}F(t) + e^{\gamma(X-X^*)}\{1 - F(t)\}. \end{aligned} \tag{2.6}$$

When $\beta < \gamma$, the odds ratio in (2.6) is decreasing, whereas when $\beta > \gamma$, the odds ratio is increasing. On the other hand, the transformation model proposed by Scheike (2006) postulates that

$$\frac{F(t|\mathbf{X},\mathbf{Z})}{S(t|\mathbf{X},\mathbf{Z})} = \mathbf{X}^T\mathbf{A}(t)\exp(\mathbf{Z}^T\boldsymbol{\beta}),$$

where $\mathbf{Z}$ is another set of covariates which satisfy the proportional odds assumption, $\mathbf{A}(\cdot)$ is a vector of functions, and $H(t|\mathbf{X}) \equiv \mathbf{X}^T\mathbf{A}(t)$ is assumed to be increasing for all $\mathbf{X}$. This model is more general than the proposed model in the sense that it does not impose any assumption on the odds ratio functions $\mathbf{A}(t)$. However, it is not clear whether the estimators of $\mathbf{A}(t)$ are constrained to be monotonic (see Equation (7) in Scheike (2006)). Alternatively, one may consider another model described by Scheike (2006)

$$\frac{F(t|\mathbf{X},\mathbf{Z})}{S(t|\mathbf{X},\mathbf{Z})} = H(t)\exp\{\mathbf{Z}^T\boldsymbol{\beta} + \mathbf{X}^T\boldsymbol{\alpha}(t)\}.$$

This model is not identifiable as $t \to 0$. Furthermore, it is not clear how one can ensure that the estimator of $S(t|\mathbf{X}, \mathbf{Z})$ is non-increasing.

Table 2.1: Summary statistics of the proposed NPMLEs

	Parameter	Bias	SEE	SE	CP	PO
		Scenario 1: $(\beta, \gamma) = (0.5, 0.5)$				
n=100	β	0.024	1.277	1.404	0.955	0.498
	γ	-0.024	1.068	1.132	0.941	
	$H(0.3)$	-0.013	0.071	0.073	0.908	
	$H(0.8)$	-0.031	0.172	0.173	0.921	
	$H(2.0)$	-0.054	0.486	0.501	0.913	
	$H(4.0)$	-0.028	1.273	1.353	0.898	
n=200	β	0.012	0.849	0.868	0.959	0.499
	γ	-0.012	0.739	0.741	0.952	
	$H(0.3)$	-0.006	0.051	0.051	0.937	
	$H(0.8)$	-0.014	0.120	0.119	0.938	
	$H(2.0)$	-0.020	0.335	0.337	0.934	
	$H(4.0)$	-0.001	0.871	0.893	0.929	
		Scenario 2: $(\beta, \gamma) = (0, 0.5)$				
n=100	β	-0.010	1.275	1.411	0.956	0.268
	γ	-0.015	1.064	1.130	0.941	
	$H(0.3)$	-0.013	0.072	0.073	0.911	
	$H(0.8)$	-0.031	0.172	0.174	0.919	
	$H(2.0)$	-0.053	0.487	0.504	0.917	
	$H(4.0)$	-0.022	1.278	1.378	0.901	
n=200	β	0.000	0.849	0.869	0.959	0.270
	γ	-0.009	0.737	0.739	0.951	
	$H(0.3)$	-0.005	0.051	0.051	0.936	
	$H(0.8)$	-0.013	0.120	0.120	0.938	
	$H(2.0)$	-0.020	0.336	0.338	0.934	
	$H(4.0)$	0.000	0.873	0.902	0.928	

Continued in next table

Table 2.2: Summary statistics of the proposed NPMLEs

	Parameter	Bias	SEE	SE	CP	PO
		Scenario 3: $(\beta, \gamma) = (-0.5, 0.5)$				
n=100	β	-0.049	1.285	1.419	0.954	0.038
	γ	-0.005	1.065	1.125	0.940	
	$H(0.3)$	-0.013	0.072	0.073	0.909	
	$H(0.8)$	-0.030	0.173	0.176	0.919	
	$H(2.0)$	-0.053	0.491	0.509	0.916	
	$H(4.0)$	-0.038	1.277	1.357	0.898	
n=200	β	-0.016	0.856	0.878	0.958	0.041
	γ	-0.006	0.736	0.739	0.949	
	$H(0.3)$	-0.006	0.051	0.051	0.935	
	$H(0.8)$	-0.013	0.121	0.121	0.938	
	$H(2.0)$	-0.019	0.340	0.344	0.934	
	$H(4.0)$	-0.004	0.877	0.905	0.932	
		Scenario 4: $(\beta, \gamma) = (-0.5, 0)$				
n=100	β	-0.016	1.277	1.416	0.955	-0.234
	γ	-0.012	1.066	1.131	0.939	
	$H(0.3)$	-0.013	0.071	0.072	0.908	
	$H(0.8)$	-0.032	0.172	0.174	0.918	
	$H(2.0)$	-0.056	0.486	0.504	0.914	
	$H(4.0)$	-0.039	1.268	1.345	0.896	
n=200	β	-0.004	0.849	0.871	0.957	-0.231
	γ	-0.007	0.738	0.741	0.948	
	$H(0.3)$	-0.006	0.051	0.051	0.934	
	$H(0.8)$	-0.014	0.120	0.119	0.938	
	$H(2.0)$	-0.021	0.336	0.339	0.936	
	$H(4.0)$	-0.004	0.871	0.897	0.930	

Bias is the bias of the parameter estimate; SEE is the average of the standard error estimates; SE is the empirical standard deviation of the parameter estimates; CP is the coverage probability of the 95% confidence interval estimates; and PO is the estimated covariate effect under the PO model.

Table 2.3: Empirical sizes/powers at significance level of 0.05

β	γ	$H_0 : \beta = 0$	$\gamma = 0$	$\beta = \gamma = 0$	$\beta = \gamma$	$\beta_{PO} = 0$
-1	-1	0.4789	0.5126	0.8457	0.0494	0.9065
-1	0	0.4812	0.0537	0.4212	0.2249	0.3548
-1	1	0.4553	0.5213	0.5804	0.6755	0.0569
0	-1	0.0511	0.5223	0.4534	0.2361	0.3932
0	0	0.0502	0.0521	0.0533	0.0488	0.0518
0	1	0.0488	0.5169	0.4532	0.2358	0.3952
1	-1	0.4497	0.5259	0.5787	0.6713	0.0568
1	0	0.4727	0.0536	0.4181	0.2261	0.3586
1	1	0.4767	0.5022	0.8422	0.0492	0.9065

Table 2.4: Comparison between the proposed method and the method of Scheike (2006)

		New			Scheike(2006)			
	True	Bias	SE	MSE	Bias	SE	MSE	RE
n=200	-0.2	0.009	0.122	0.015	0.004	0.133	0.018	1.170
	-0.1	0.005	0.122	0.015	0.002	0.131	0.017	1.169
	0.0	0.001	0.121	0.015	0.001	0.131	0.017	1.172
	0.1	-0.003	0.122	0.015	0.000	0.132	0.017	1.172
	0.2	-0.007	0.123	0.015	-0.002	0.133	0.018	1.170
n=400	-0.2	0.012	0.086	0.007	0.003	0.094	0.009	1.196
	-0.1	0.007	0.085	0.007	0.003	0.094	0.009	1.211
	0.0	0.002	0.085	0.007	0.002	0.094	0.009	1.217
	0.1	-0.003	0.085	0.007	0.001	0.094	0.009	1.216
	0.2	-0.009	0.085	0.007	0.000	0.094	0.009	1.203

True is the true value of the parameter; Bias is the bias of the parameter estimate; SE is the empirical standard deviation of the parameter estimates; MSE is the mean squared error; and RE is the MSE relative efficiency of the proposed estimator compared to the estimator of Scheike (2006).

Table 2.5: Results for the CNA study

SNP loci with the smallest P-values for testing the PO assumption

SNP	$\hat{\beta}_3$	$se(\hat{\beta}_3)$	$\hat{\gamma}_3$	$se(\hat{\gamma}_3)$	$\beta_3 = 0$	$\gamma_3 = 0$	$\beta_3 = \gamma_3 = 0$	$\beta_3 = \gamma_3$	$\hat{\beta}_{3PO}$	$\beta_{3PO} = 0$
39687	1.482	0.249	-1.130	0.212	2.55E-09	1.01E-07	2.93E-10	1.82E-12	0.006	0.967
39734	1.481	0.249	-1.130	0.212	2.55E-09	1.01E-07	2.94E-10	1.82E-12	0.006	0.967
36091	1.463	0.249	-1.139	0.212	4.52E-09	8.01E-08	3.02E-10	2.48E-12	0.001	0.996
11222	1.404	0.246	-1.118	0.209	1.14E-08	8.88E-08	9.13E-10	4.75E-12	-0.022	0.885
27435	-2.410	0.341	0.708	0.195	1.68E-12	2.84E-04	4.21E-07	5.12E-12	-0.172	0.222
7879	1.411	0.251	-1.122	0.209	1.82E-08	7.88E-08	9.19E-10	6.54E-12	-0.028	0.851
7934	1.411	0.251	-1.122	0.209	1.82E-08	7.89E-08	9.19E-10	6.54E-12	-0.028	0.851
52493	1.356	0.235	-1.151	0.221	8.15E-09	2.00E-07	1.10E-09	7.42E-12	0.029	0.844
55032	1.355	0.235	-1.152	0.222	8.43E-09	2.00E-07	1.12E-09	7.63E-12	0.029	0.844
54335	1.355	0.235	-1.152	0.222	8.43E-09	2.00E-07	1.12E-09	7.63E-12	0.029	0.844

SNP loci with the most significant effects in the PO model

SNP	$\hat{\beta}_3$	$se(\hat{\beta}_3)$	$\hat{\gamma}_3$	$se(\hat{\gamma}_3)$	$\beta_3 = 0$	$\gamma_3 = 0$	$\beta_3 = \gamma_3 = 0$	$\beta_3 = \gamma_3$	$\hat{\beta}_{3PO}$	$\beta_{3PO} = 0$
38136	0.856	0.201	0.426	0.188	2.07E-05	2.38E-02	1.39E-06	0.114	0.623	5.04E-06
9940	-0.974	0.531	0.902	0.168	6.63E-02	8.37E-08	8.48E-06	0.002	0.620	5.64E-06
9732	-0.978	0.530	0.902	0.168	6.49E-02	8.14E-08	8.35E-06	0.002	0.620	5.64E-06
9889	-0.975	0.531	0.902	0.168	6.60E-02	8.34E-08	8.47E-06	0.002	0.620	5.65E-06
24812	-0.373	0.364	-0.789	0.218	3.05E-01	3.04E-04	5.69E-06	0.399	-0.649	6.79E-06
8495	0.984	0.287	0.418	0.235	6.05E-04	7.51E-02	2.28E-06	0.194	0.611	6.93E-06
25940	0.968	0.319	0.439	0.245	2.43E-03	7.35E-02	3.01E-06	0.275	0.614	7.04E-06
31353	0.969	0.319	0.438	0.246	2.41E-03	7.48E-02	3.06E-06	0.274	0.613	7.16E-06
31412	0.969	0.319	0.438	0.246	2.42E-03	7.48E-02	3.06E-06	0.274	0.613	7.17E-06
31253	0.969	0.319	0.438	0.246	2.42E-03	7.48E-02	3.07E-06	0.274	0.613	7.18E-06

SNP loci with the largest P-values for testing the PO assumption

SNP	$\hat{\beta}_3$	$se(\hat{\beta}_3)$	$\hat{\gamma}_3$	$se(\hat{\gamma}_3)$	$\beta_3 = 0$	$\gamma_3 = 0$	$\beta_3 = \gamma_3 = 0$	$\beta_3 = \gamma_3$	$\hat{\beta}_{3PO}$	$\beta_{3PO} = 0$
19252	0.108	0.354	0.108	0.243	0.760	0.656	0.437	1.000	0.113	0.420
33094	0.057	0.429	0.057	0.206	0.894	0.780	0.676	1.000	0.053	0.699
19200	0.108	0.354	0.108	0.243	0.760	0.656	0.437	1.000	0.113	0.420
33046	0.057	0.429	0.057	0.205	0.894	0.780	0.676	1.000	0.053	0.699
38039	-0.322	0.445	-0.322	0.230	0.470	0.160	0.027	1.000	-0.326	0.025
38133	-0.322	0.446	-0.322	0.230	0.471	0.161	0.027	1.000	-0.326	0.026
18446	0.219	0.251	0.219	0.210	0.382	0.297	0.095	1.000	0.219	0.096
8897	-0.128	0.304	-0.128	0.231	0.674	0.578	0.373	1.000	-0.127	0.378
2124	-0.203	0.285	-0.203	0.242	0.476	0.401	0.172	1.000	-0.193	0.196
26366	0.224	0.279	0.224	0.191	0.421	0.241	0.075	1.000	0.217	0.083

$\hat{\beta}_3$ and $\hat{\gamma}_3$ are the estimates of the long-term genetic effect and short-term genetic effect under our proposed model. $se(\hat{\beta}_3)$ and $se(\hat{\gamma}_3)$ are their estimated standard errors. $\hat{\beta}_{3PO}$ is the estimate of the genetic effect under the PO model. Columns 6, 7, 8, 9 and 11 are the P-values for testing the corresponding null hypotheses.

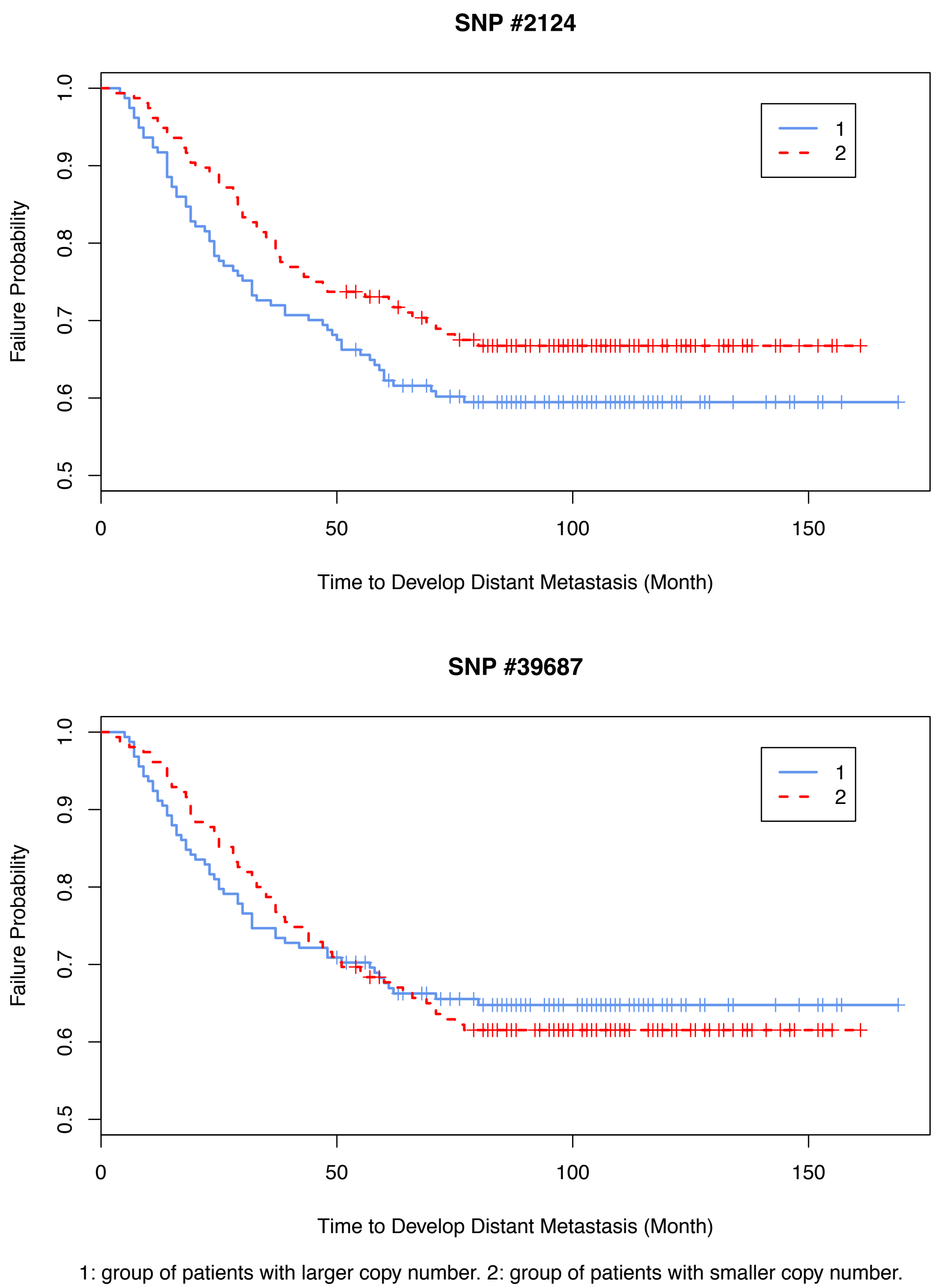

Figure 2.1: The Kaplan-Meier curves of time to develop distant metastasis for the CNA study at SNP #2124 and SNP #39687.

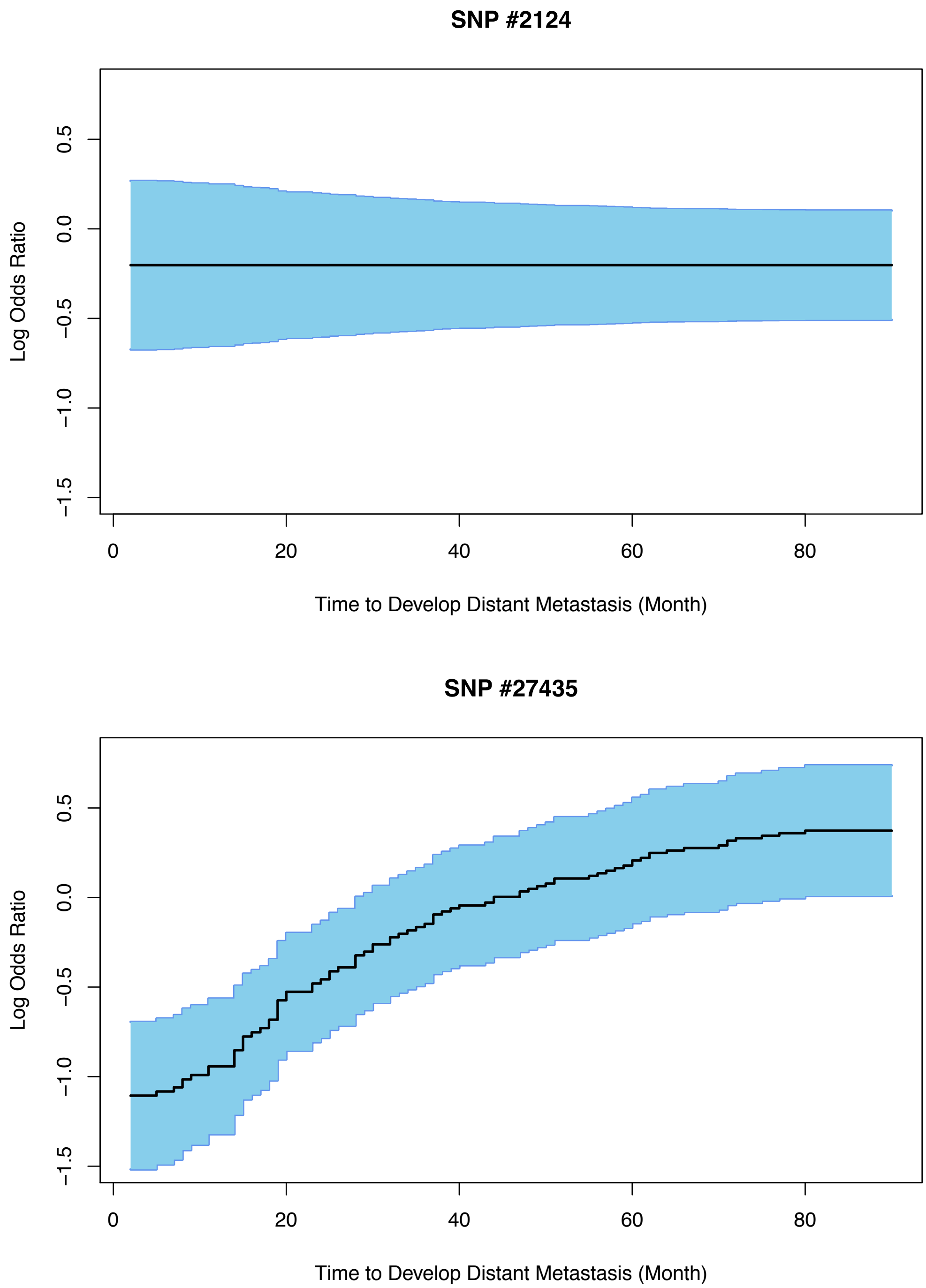

Figure 2.2: Adjusted log odds ratios (black curves) and the 95% pointwise confidence bands (blue) at SNP #2124 and SNP #39687.

Chapter 3: Sieve Maximum Likelihood Estimation Using B-spline Smoothing in Generalized Linear Models with an Unknown Link Function

3.1 Introduction

Canonical link functions are commonly used in practice for the generalized linear models. However, more concerns have risen about the impact of misspecification of the link function on the inference procedure, say the consistency and robustness of the regression parameter estimators. Pregibon (1981) provided data analytic procedures as a tool to assess the adequacy of the assumed link function for GLMs. Aranda-Ordaz (1981) proposed families of parametric link functions for binary outcomes by considering departures from the logistic model. Scallan et al. (1984) extended these ideas to other GLMs. However parametric models are not adequate in many applications (Chiou and Müller, 2005). A Bayesian approach assuming a strictly increasing link function was introduced by Mallick and Gelfand (1994).

Beyond the Bayesian approach, smoothing techniques were another widely-used procedure to estimate the unknown link in the GLMs without a full distribution assumption. Weisberg and Welsh (1994) introduced a Kernel-based method to estimate the unknown link function in the GLMs. Other smoothing techniques without full model specification have also been considered in Fan et al. (1995), Chiou and Müller (1998) and Muggeo and Ferrara (2007), among others.

For the GLMs, alternative inferential methods using the Quasi-likelihood have been well studied in literature. Chiou and Müller (1998) pointed out that specifying the variance function as an unknown function of the mean would not lose efficiency based on the quasi-likelihood. They used the local polynomial smoothing method fitting by locally weighted least squares, and suggested well-working data-adaptive bandwidth choices. Muggeo and Ferrara (2007) presented an estimating algorithm by using the P-spline smoothing.

In this chapter, we consider the GLMs with an unknown link function. We propose to estimate the unknown parameters by B-spline smoothing. Specifically, we maximize the log-likelihood function over a sieve space induced by the B-splines. We establish the consistency of the resulting sieve MLEs. In addition, we show that the sieve MLEs of the regression coefficients are asymptotically normally distributed, and the variance-covariance attains the semiparametric efficiency bound. Extensive simulation studies demonstrate that the proposed estimators perform well.

3.2 Methods

Suppose there are n subjects in the study. Let Y_i denote the outcome of the ith subject and $\mathbf{X}_i = (X_{i1}, \cdots, X_{ip})^T$ denote the corresponding p-dimensional vector of covariates, $i = 1, \cdots, n$. We assume that the outcome variable Y has a probability distribution belonging to the exponential family. Specifically, the density of Y_i in the general form of the exponential family is given by

$$f(y) = \exp\left\{\frac{y\phi_i - b(\phi_i)}{a(\varphi)} - c(y, \varphi)\right\},$$

Let $\mu_i = E(Y_i|\mathbf{X}_i)$. Consider the following generalized linear model

$$g(\mu_i) = h(\mathbf{X}_i^T\boldsymbol{\beta}), i = 1, \cdots, n, \tag{3.1}$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown regression parameters and $h^{-1} \circ g(\cdot)$ is the link function. In this model, $g(\cdot)$ is a known transformation function and $h(\cdot) : \mathcal{R} \to \mathcal{R}$ is an unknown function which needs to be estimated. The transformation function $g(\cdot)$ can be taken as the canonical link function, that is, $g(\mu_i) = \phi_i$. For example, $g(x) = x$ for normal data, $g(x) = \log(x)$ for count data, and $g(x) = \text{logit}(x)$ for binary data. To ensure the identifiability of the model, we impose the constraint on $\boldsymbol{\beta}$ such that $||\boldsymbol{\beta}|| = 1$ and $\beta_p > 0$, where $||\cdot||$ is the Euclidean norm.

The log-likelihood for $(\boldsymbol{\beta}, h)$ is

$$
\begin{aligned}
l_n(\boldsymbol{\beta}, h) &= \sum_{i=1}^{n} \log f(y_i|\phi_i, \mathbf{x}_i) \\
&= \sum_{i=1}^{n} \left[\frac{1}{a(\varphi)} \left\{ y_i h(\mathbf{x}_i^T \boldsymbol{\beta}) - b\left(h(\mathbf{x}_i^T \boldsymbol{\beta})\right) \right\} \right] + C,
\end{aligned}
$$

where C is a constant as a function of $(\boldsymbol{\beta}, h)$ for the given $\mathbf{x} = (\mathbf{x}_1^T, \mathbf{x}_2^T, \cdots, \mathbf{x}_n^T)^T$ and $\mathbf{y} = (y_1, y_2, \cdots, y_n)^T$. We propose to approximate $h(\cdot)$ with B-splines.

Suppose $\mathbf{X}_i^T \boldsymbol{\beta}$ are bounded by (a, b). For $\nu \in (0, 0.5)$, consider K_n interior knots $K_1, K_2, \cdots, K_{K_n}$ with $K_n = O(n^{\nu})$ and $\max_{0 \le j \le K_n} |K_{j+1} - K_j| = O(n^{-\nu})$, and $a = K_0 \le K_1 \le K_2 \le \cdots \le K_{K_n} \le K_{K_n+1} = b$. Let $\mathcal{S}_n(\mathbf{K}, K_n, M)$ denote the space of polynomial splines of order M defined in Schumaker (1981), where $\mathbf{K} = (K_0, K_1, \cdots, K_{K_n+1})$. Then there exists a local basis $\{B_j : 1 \le j \le K_n + M\}$ such that for any $s \in \mathcal{S}_n(\mathbf{K}, K_n, M)$, we have

$$
s(t) = \sum_{j=1}^{K_n+M} \gamma_j B_{M,j}(t) = \boldsymbol{\gamma}^T \mathbf{B}_M(t),
$$

where $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \cdots, \gamma_{K_n+M})^T$ are the smoothing coefficients and $B_M(t) = \big\{B_{M,1}(t), B_{M,2}(t), \cdots, B_{M,K_n+M}(t)\big\}^T$ are the M-degree B-spline basis functions. Under some smoothness assumptions, the unknown univariate function $h(\cdot)$ can be well approximated by an M-degree

B-spline in $\mathcal{S}_n(\mathbf{K}, K_n, M)$. Replacing $h(t)$ by $\boldsymbol{\gamma}^T \mathbf{B}_M(t)$, we obtain

$$l_n(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_{i=1}^{n} \left[\frac{1}{a(\varphi)} \left\{ y_i \boldsymbol{\gamma}^T \mathbf{B}_M(\mathbf{x}_i^T \boldsymbol{\beta}) - b\left(\boldsymbol{\gamma}^T \mathbf{B}_M(\mathbf{x}_i^T \boldsymbol{\beta}) \right) \right\} \right] + C.$$

To maximize the above sieve log-likelihood, we solve the following score equations:

$$\frac{\partial l_n(\boldsymbol{\beta}, \boldsymbol{\gamma})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \frac{1}{a(\varphi)} \left\{ y_i - b'(\phi_i) \right\} \sum_{j=1}^{K_n+M} \gamma_j B'_{M,j}(\mathbf{x}_i^T \boldsymbol{\beta}) \mathbf{x}_i = \mathbf{0},$$

$$\frac{\partial l_n(\boldsymbol{\beta}, \boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}} = \sum_{i=1}^{n} \frac{1}{a(\varphi)} \left\{ y_i - b'(\phi_i) \right\} \mathbf{B}_M(\mathbf{x}_i^T \boldsymbol{\beta}) = \mathbf{0},$$

where $b'(\cdot)$ and $B'_{M,j}(\cdot)$ are the derivatives of $b(\cdot)$ and $B_{M,j}(\cdot)$, respectively. We use an iterative algorithm to estimate $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. We refer to the resulting estimator $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}, \hat{h})$ as the sieve MLEs, where $\hat{h}(t) = \hat{\boldsymbol{\gamma}}^T \mathbf{B}_M(t)$. Note that the estimation of $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ does not depend on $a(\varphi)$. However, the estimation of the variance of $\hat{\boldsymbol{\beta}}_n$ depends on $a(\varphi)$ because the efficient score function of $\boldsymbol{\beta}_0$ involves $a(\varphi)$. Without loss of generality, we assume φ is known. Also note that in the likelihood, the basis are functions of $\mathbf{X}^T \boldsymbol{\beta}$. We recommend to choose the knots based on the equally spaced sample quantiles of $\mathbf{X}^T \boldsymbol{\beta}$. As a result, the knots and basis change as $\boldsymbol{\beta}$ changes. We describe the detailed algorithm in Section 3.4.

To make statistical inference of $\boldsymbol{\beta}_0$, the distribution of $\hat{\boldsymbol{\beta}}_n$ can be approximated by a multivariate normal distribution when the sample size is large enough. The variance-covariance matrix of $\hat{\boldsymbol{\beta}}_n$ can be estimated by the inverse of the observed information matrix under the efficient score function of $\boldsymbol{\beta}_0$. We show that the asymptotic variance-covariance matrix of $\hat{\boldsymbol{\beta}}_n$ attains the semiparametric efficiency bound. Ding and Nan (2011) suggested to use the observed information matrix by taking into account the nuisance parameters $\boldsymbol{\gamma}$. In our simulation, we used the inverse matrix of the estimated variance-covariance matrix of the score functions also by taking into account the nuisance parameters $\boldsymbol{\gamma}$. Extensive simulations suggest that this variance-covariance estimator works well. When φ is unknown, we denote $\boldsymbol{\alpha} = (\boldsymbol{\beta}, \varphi)$ and similar techniques can be used to obtain the sieve MLEs of $\boldsymbol{\alpha}$, denoted by $\hat{\alpha}_n$, and to establish the asymptotic properties of $(\hat{\boldsymbol{\alpha}}_n, \hat{h}_n)$.

3.3 Asymptotic Properties

Let $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0, h_0)$ denote the true parameters. Before we establish the asymptotic properties of the sieve MLEs, we impose the following regularity conditions.

C1. The true values $\boldsymbol{\beta}_0$ for parameters $\boldsymbol{\beta}$ belong to the interior of a compact set

$$\mathcal{B}_0 = \left\{ \boldsymbol{\beta}^* : \boldsymbol{\beta}^* \in \mathcal{R}^p, \beta_p^* > 0, \text{ and } ||\boldsymbol{\beta}^*|| = 1 \right\}.$$

C2. The domain of $\mathbf{X}$ is a bounded subset of $\mathcal{R}^p$. Furthermore, $E(\mathbf{X}\mathbf{X}^T)$ is nonsingular.

C3. The true function $h_0(\cdot)$ belongs to $\mathcal{H}^q$, where the functional space $\mathcal{H}^q$ is the collection of bounded functions h on a bounded interval $[a, b]$ with bounded jth derivative $h^{(j)}$, $j = 1, \cdots, k$, such that $h^{(k)}$ satisfies the Lipschitz continuity condition with exponent m $(0 < m \leq 1)$:

$$|h^{(k)}(s) - h^{(k)}(t)| \leq L|s - t|^m, \text{ for } s, t \in [a, b],$$

where $L < \infty$ is a positive constant, and $q = k + m \geq 3$.

C4. For some $\rho \in (0, 1)$, $u^T Var(\mathbf{X}|\mathbf{X}^T\boldsymbol{\beta}_0)u \geq \rho u^T E(\mathbf{X}\mathbf{X}^T|\mathbf{X}^T\boldsymbol{\beta}_0)u$ almost surly for all $u \in \mathcal{R}^p$.

C5. $E\left[\left\{h_0'(\mathbf{X}^T\boldsymbol{\beta}_0)\right\}^2 \mathbf{X}\mathbf{X}^T\right]$ is nonsingular.

Remark 3.1. Condition C1 is a common regularity assumption in literature. Conditions C1 and C2 ensure the identifiability of the model. The restriction $q \geq 3$ in the condition C3 is needed to provide a desirable control of the B-spline approximation error rates of h_0 as well as the first and second derivatives of h_0 for the proof of the asymptotic normality of $\hat{\boldsymbol{\beta}}_n$. Conditions C4 and C5 are technical assumptions and can be justified in many applications. In addition, Condition C5 ensures that the semiparametric efficiency information matrix is invertible.

Let $\mathcal{H}_n^q = \mathcal{H}^q \cap \mathcal{S}_n$. Clearly, we have $\mathcal{H}_n^q \subseteq \mathcal{H}_{n+1}^q \subseteq \cdots \subseteq \mathcal{H}^q$ for all $n \geq 1$. Denote $\Theta_n^q = \mathcal{B} \times \mathcal{H}_n^q$ as the sieve space. The sieve MLEs are the maximizer of the log-likelihood function over the sieve space Θ_n^q. Define the norm $||\cdot||_2$ over the space $\mathcal{H}^q$ as

$$||h||_2 = \left[E\{h^2(\mathbf{X}^T\boldsymbol{\beta}_0)\}\right]^{\frac{1}{2}}.$$

Let $||\cdot||_\infty$ denote the supremum norm. We define the distance over the space $\Theta^q = \mathcal{B} \times \mathcal{H}^p$ as follows

$$d(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \left(||\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2||^2 + ||h_1 - h_2||_2^2\right)^{\frac{1}{2}}, \text{ for } \boldsymbol{\theta}_1 = (\boldsymbol{\beta}_1, h_1), \boldsymbol{\theta}_2 = (\boldsymbol{\beta}_2, h_2) \in \Theta^q.$$

Next, we introduce some notation in the context of empirical process. Let $\mathbb{P}_n$ and $\mathbb{P}$ be the empirical measure and the population distribution of n i.i.d. observations $\mathbf{O}_1, \mathbf{O}_2, \cdots, \mathbf{O}_n$, where $\mathbf{O}_i = (Y_i, \mathbf{X}_i), i = 1, \cdots, n$. Let $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - \mathbb{P})$ denote the empirical process. Then for any measurable function $h(\mathbf{O})$,

$$\mathbb{P}_n h(\mathbf{O}) = \frac{1}{n}\sum_{i=1}^{n} h(\mathbf{O}_i), \mathbb{P}h(\mathbf{O}) = E_{\mathbb{P}}[h(\mathbf{O})],$$

and

$$\mathbb{G}_n h(\mathbf{O}) = \sqrt{n}(\mathbb{P}_n h(\mathbf{O}) - \mathbb{P}h(\mathbf{O})),$$

where $\mathbf{O} = (Y, \mathbf{X})$. Let $K_n = O(n^\nu)$, and $\frac{1}{2(1+q)} < \nu < \frac{1}{2q}$, where q is the smoothness parameter in Condition C3. The following theorem establishes the convergence rate of the sieve MLEs $\hat{\boldsymbol{\theta}}_n = (\hat{\boldsymbol{\beta}}_n, \hat{h}_n)$.

Theorem 3.1. Under the regularity conditions C1-C6,

$$d(\hat{\boldsymbol{\theta}}_n, \theta_0) = O_p(n^{-\min(q\nu, \frac{1-\nu}{2})}).$$

Remark 3.2. The proof is provided in the following by verifying the assumptions in Theorem 1 in Shen and Wong (1994). According to Ding and Nan (2011), the sieve space $\mathcal{H}_n^q$ does not have to be restricted to $\mathcal{S}_n$ as long as the estimator $\hat{\boldsymbol{\theta}}_n$ satisfies the assumptions of Theorem 1 in Shen and Wong (1994). Notice when $\nu = \frac{1}{1+2q}$, we have the optimal convergence rate of $n^{-\frac{q}{1+2q}}$ in the nonparametric regression setting. In general, the convergence rate in this theorem is lower than $n^{-\frac{1}{2}}$, yet as stated in the next theorem, the proposed estimators of the regression parameters still possess asymptotic normality and are semiparametrically efficient.

Proof. We prove Theorem 3.1 by verifying the assumptions in Theorem 1 in Shen and Wong (1994). For simplicity of the notation, we assume $a(\varphi) = \varphi$ is known. It is straightforward to generalize the results to the case when φ needs to be estimated.

Let $l(\boldsymbol{\beta}, h; \mathbf{O}) = \frac{1}{\varphi}\left\{Y h(\mathbf{X}^T\boldsymbol{\beta}) - b\left(h(\mathbf{X}^T\boldsymbol{\beta})\right)\right\}$, and

$$\begin{aligned}
\dot{l}_{\boldsymbol{\beta}}(\boldsymbol{\beta}, h; \mathbf{O}) &= \frac{\partial l(\boldsymbol{\beta}, h; \mathbf{O})}{\partial \boldsymbol{\beta}} \\
&= \frac{1}{\varphi}\left\{Y - b'\left(h(\mathbf{X}^T\boldsymbol{\beta})\right)\right\} h'(\mathbf{X}^T\boldsymbol{\beta})\mathbf{X}, \\
\dot{l}_h(\boldsymbol{\beta}, h; \mathbf{O})[h_1] &= \frac{\partial l(\boldsymbol{\beta}, h + \epsilon h_1; \mathbf{O})}{\partial \epsilon}\Big|_{\epsilon=0} \\
&= \frac{1}{\varphi}\left\{Y - b'\left(h(\mathbf{X}^T\boldsymbol{\beta})\right)\right\} h_1(\mathbf{X}^T\boldsymbol{\beta}),
\end{aligned}$$

$$
\begin{aligned}
\ddot{l}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\beta}, h; \mathbf{O}) &= \frac{\partial^2 l(\boldsymbol{\beta}, h; \mathbf{O})}{\partial \boldsymbol{\beta}\boldsymbol{\beta}^T} \\
&= \frac{1}{\varphi}\Big\{Y h''(\mathbf{X}^T\boldsymbol{\beta}) - b'\left(h(\mathbf{X}^T\boldsymbol{\beta})\right) h''(\mathbf{X}^T\boldsymbol{\beta}) \\
&\qquad -b''\left(h(\mathbf{X}^T\boldsymbol{\beta})\right)\left(h'(\mathbf{X}^T\boldsymbol{\beta})\right)^2\Big\}\mathbf{X}\mathbf{X}^T, \\
\ddot{l}_{hh}(\boldsymbol{\beta}, h; \mathbf{O})[h_1, h_2] &= \frac{\partial}{\partial \varepsilon}\dot{l}_h(\boldsymbol{\beta}, h + \varepsilon h_2; \mathbf{O})[h_1] \\
&= \frac{1}{\varphi}\left\{-b''\left(h(\mathbf{X}^T\boldsymbol{\beta})\right) h_1(\mathbf{X}^T\boldsymbol{\beta}) h_2(\mathbf{X}^T\boldsymbol{\beta})\right\}, \\
\ddot{l}_{\boldsymbol{\beta}h}(\boldsymbol{\beta}, h; \mathbf{O})[h_1] &= \frac{\partial}{\partial \varepsilon}\dot{l}_{\boldsymbol{\beta}}(\boldsymbol{\beta}, h + \varepsilon h_1; \mathbf{O}) \\
&= \frac{1}{\varphi}\Big\{Y h_1'(\mathbf{X}^T\boldsymbol{\beta}) - b'\left(h(\mathbf{X}^T\boldsymbol{\beta})\right) h_1'(\mathbf{X}^T\boldsymbol{\beta}) \\
&\qquad -b''\left(h(\mathbf{X}^T\boldsymbol{\beta})\right) h_1(\mathbf{X}^T\boldsymbol{\beta}) h'(\mathbf{X}^T\boldsymbol{\beta})\Big\}\mathbf{X}, \\
\ddot{l}_{h\boldsymbol{\beta}}(\boldsymbol{\beta}, h; \mathbf{O})[h_1] &= \ddot{l}^T_{\boldsymbol{\beta}h}(\boldsymbol{\beta}, h; \mathbf{O})[h_1], \\
\ddot{l}_{hh}(\boldsymbol{\beta}, h; \mathbf{O})[h_1, \mathbf{h}^*] &= \left(\ddot{l}_{hh}(\boldsymbol{\beta}, h; \mathbf{O})[h_1, h_1^*], \ddot{l}_{hh}(\boldsymbol{\beta}, h; \mathbf{O})[h_1, h_2^*], \cdots, \ddot{l}_{hh}(\boldsymbol{\beta}, h; \mathbf{O})[h_1, h_p^*]\right), \\
\ddot{l}_{hh}(\boldsymbol{\beta}, h; \mathbf{O})[\mathbf{h}^*, h_1] &= \ddot{l}^T_{hh}(\boldsymbol{\beta}, h; \mathbf{O})[h_1, \mathbf{h}^*], \\
\ddot{l}_{\boldsymbol{\beta}h}(\boldsymbol{\beta}, h; \mathbf{O})[\mathbf{h}^*] &= \left(\ddot{l}_{\boldsymbol{\beta}h}(\boldsymbol{\beta}, h; \mathbf{O})[h_1^*], \ddot{l}_{\boldsymbol{\beta}h}(\boldsymbol{\beta}, h; \mathbf{O})[h_2^*], \cdots, \ddot{l}_{\boldsymbol{\beta}h}(\boldsymbol{\beta}, h; \mathbf{O})[h_p^*]\right), \\
\ddot{l}_{h\boldsymbol{\beta}}(\boldsymbol{\beta}, h; \mathbf{O})[\mathbf{h}^*] &= \ddot{l}^T_{\boldsymbol{\beta}h}(\boldsymbol{\beta}, h; \mathbf{O})[\mathbf{h}^*],
\end{aligned}
$$

where $h_1, h_2 \in \mathbf{H} = \{g : g = \frac{\partial h_\epsilon}{\partial \epsilon}\big|_{\epsilon=0}, h_\epsilon \in \mathcal{H}^p\}$, and $\mathbf{h}^* = (h_1^*, h_2^*, \cdots, h_p^*) \in \mathbf{H}^p$. Under conditions C1-C3, all the above derivatives are continuous and bounded. Moreover, the first derivative of $\mathbb{P}l(\boldsymbol{\beta}, h; \mathbf{O})$ at $(\boldsymbol{\beta}_0, h_0)$ is $\mathbf{0}$. Then by applying the Taylor series expansion, we obtain

$$
\mathbb{P}l(\boldsymbol{\beta}, h; \mathbf{O}) - \mathbb{P}l(\boldsymbol{\beta}_0, h_0; \mathbf{O})
$$

$$= \frac{1}{2}\mathbb{P}\Big\{(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T\ddot{l}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\beta}_0,h_0;\mathbf{O})(\boldsymbol{\beta}-\boldsymbol{\beta}_0)+2(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T\ddot{l}_{\boldsymbol{\beta}h}(\boldsymbol{\beta}_0,h_0,\mathbf{O})[h-h_0]$$

$$+\ddot{l}_{hh}(\boldsymbol{\beta}_0,h_0;\mathbf{O})[h-h_0,h-h_0]\Big\}+o(d^2(\boldsymbol{\theta},\boldsymbol{\theta}_0))$$

$$= A+o(d^2(\boldsymbol{\theta},\boldsymbol{\theta}_0)).$$

It is easy to show that

$$E(Y|\mathbf{X}) = b'\left(h(\mathbf{X}^T\boldsymbol{\beta})\right)$$

and

$$Var(Y|\mathbf{X}) = b''\left(h(\mathbf{X}^T\boldsymbol{\beta})\right)\varphi > 0.$$

Then there exists a positive number c_1 such that $-b''\left(h_0(\mathbf{X}^T\boldsymbol{\beta}_0)\right) \leq -c_1$. It follows that

$$\begin{aligned}
A &= \frac{1}{2\varphi}\mathbb{P}\Big[-b''\left(h_0(\mathbf{X}^T\boldsymbol{\beta}_0)\right)\left(h_0'(\mathbf{X}^T\boldsymbol{\beta}_0)\right)^2(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T\mathbf{X}\mathbf{X}^T(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\\
&\quad -2b''\left(h_0(\mathbf{X}^T\boldsymbol{\beta}_0)\right)\left\{h(\mathbf{X}^T\boldsymbol{\beta}_0)-h_0(\mathbf{X}^T\boldsymbol{\beta}_0)\right\}h_0'(\mathbf{X}^T\boldsymbol{\beta}_0)(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T\mathbf{X}\\
&\quad -b''\left(h_0(\mathbf{X}^T\boldsymbol{\beta}_0)\right)\left\{h(\mathbf{X}^T\boldsymbol{\beta}_0)-h_0(\mathbf{X}^T\boldsymbol{\beta}_0)\right\}^2\Big]\\
&\leq -\frac{c_1}{2\varphi}\mathbb{P}\Big[(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T\mathbf{X}\mathbf{X}^T(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\left(h_0'(\mathbf{X}^T\boldsymbol{\beta}_0)\right)^2\\
&\quad +2\left\{h(\mathbf{X}^T\boldsymbol{\beta}_0)-h_0(\mathbf{X}^T\boldsymbol{\beta}_0)\right\}h_0'(\mathbf{X}^T\boldsymbol{\beta}_0)(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T\mathbf{X}\\
&\quad +\left\{h(\mathbf{X}^T\boldsymbol{\beta}_0)-h_0(\mathbf{X}^T\boldsymbol{\beta}_0)\right\}^2\Big]\\
&= -\frac{c_1}{2\varphi}\mathbb{P}\left\{h_0'(\mathbf{X}^T\boldsymbol{\beta}_0)\mathbf{X}^T(\boldsymbol{\beta}-\boldsymbol{\beta}_0)+\left(h(\mathbf{X}^T\boldsymbol{\beta}_0)-h_0(\mathbf{X}^T\boldsymbol{\beta}_0)\right)\right\}^2\\
&\leq -\frac{c_1}{2\varphi}\Big[\mathbb{P}\left\{h_0'(\mathbf{X}^T\boldsymbol{\beta}_0)\mathbf{X}^T(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\right\}^2+\mathbb{P}\left\{h(\mathbf{X}^T\boldsymbol{\beta}_0)-h_0(\mathbf{X}^T\boldsymbol{\beta}_0)\right\}^2\\
&\quad -2\Big|\mathbb{P}\left\{h_0'(\mathbf{X}^T\boldsymbol{\beta}_0)\mathbf{X}^T(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\left(h(\mathbf{X}^T\boldsymbol{\beta}_0)-h_0(\mathbf{X}^T\boldsymbol{\beta}_0)\right)\right\}\Big|\Big].
\end{aligned}$$

Let $B_1 = h_0'(\mathbf{X}^T\boldsymbol{\beta}_0)\mathbf{X}^T(\boldsymbol{\beta}-\boldsymbol{\beta}_0)$ and $B_2 = h(\mathbf{X}^T\boldsymbol{\beta}_0)-h_0(\mathbf{X}^T\boldsymbol{\beta}_0)$. Under condition C4 and by Cauchy-Schwartz inequality, we have

$$\left|\mathbb{P}(B_1 B_2)\right| \leq \sqrt{(1-\rho)\mathbb{P}(B_1^2)\mathbb{P}(B_2^2)} \text{ for some } \rho \in (0,1).$$

Then

$$A \leq -\frac{c_1}{2\varphi}\left[\left(1-\sqrt{1-\rho}\right)\left\{\mathbb{P}(B_1^2)+\mathbb{P}(B_2^2)\right\}\right].$$

In addition, condition (C5) ensures that the smallest eigenvalue of $\mathbb{P}\left[\left\{h_0'(\mathbf{X}^T\boldsymbol{\beta}_0)\right\}^2 \mathbf{X}\mathbf{X}^T\right]$ is positive, and therefore is greater than a positive number c_2. It follows that

$$\mathbb{P}(B_1^2) = (\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T \mathbb{P}\left[\left\{h_0'(\mathbf{X}^T\boldsymbol{\beta}_0)\right\}^2 \mathbf{X}\mathbf{X}^T\right](\boldsymbol{\beta}-\boldsymbol{\beta}_0) \geq c_2 ||\boldsymbol{\beta}-\boldsymbol{\beta}_0||^2.$$

By definition, $\mathbb{P}(B_2^2) = ||h-h_0||_2^2$. Thus, there exists a positive constant C_1 such that

$$\mathbb{P}(l(\boldsymbol{\beta}_0, h_0; \mathbf{O}) - l(\boldsymbol{\beta}, h; \mathbf{O})) \geq C_1 d^2(\boldsymbol{\theta}, \boldsymbol{\theta}_0).$$

It implies that

$$\inf_{\{d(\boldsymbol{\theta},\boldsymbol{\theta}_0)\geq\varepsilon, \boldsymbol{\theta}\in\Theta_n^p\}} \mathbb{P}(l(\boldsymbol{\theta}_0; \mathbf{O}) - l(\boldsymbol{\theta}; \mathbf{O})) \geq C_1\varepsilon^2.$$

Hence the condition C1 of Shen and Wong (1994) holds with $\alpha = 1$ in their notation.

Next we verify the condition C2 in Shen and Wong (1994). We have

$$\begin{aligned}
&\varphi\left\{l(\boldsymbol{\beta}, h; \mathbf{O}) - l(\boldsymbol{\beta}_0, h_0; \mathbf{O})\right\}^2 \\
&= \left[Y\left\{h(\mathbf{X}^T\boldsymbol{\beta}) - h_0(\mathbf{X}^T\boldsymbol{\beta}_0)\right\} - \left\{b\left(h(\mathbf{X}^T\boldsymbol{\beta})\right) - b\left(h_0(\mathbf{X}^T\boldsymbol{\beta}_0)\right)\right\}\right]^2 \\
&\leq 2\left[Y\left\{h(\mathbf{X}^T\boldsymbol{\beta}) - h_0(\mathbf{X}^T\boldsymbol{\beta}_0)\right\}\right]^2 + 2\left[\left\{b\left(h(\mathbf{X}^T\boldsymbol{\beta})\right) - b\left(h_0(\mathbf{X}^T\boldsymbol{\beta}_0)\right)\right\}\right]^2 \\
&\leq 2\left[Y\left\{h(\mathbf{X}^T\boldsymbol{\beta}) - h(\mathbf{X}^T\boldsymbol{\beta}_0)\right\}\right]^2 + 2\left[Y\left\{h(\mathbf{X}^T\boldsymbol{\beta}_0) - h_0(\mathbf{X}^T\boldsymbol{\beta}_0)\right\}\right]^2
\end{aligned}$$

$$+2\left[\left\{b\left(h(\mathbf{X}^T\boldsymbol{\beta})\right)-b\left(h_0(\mathbf{X}^T\boldsymbol{\beta}_0)\right)\right\}\right]^2.$$

By the Mean Value Theorem, there exists a point c_3 between $\mathbf{X}^T\boldsymbol{\beta}$ and $\mathbf{X}^T\boldsymbol{\beta}_0$ such that

$$h(\mathbf{X}^T\boldsymbol{\beta})-h(\mathbf{X}^T\boldsymbol{\beta}_0)=h'(c_3)\mathbf{X}^T(\boldsymbol{\beta}-\boldsymbol{\beta}_0).$$

Similarly, there exists a point c_4 between $h(\mathbf{X}^T\boldsymbol{\beta})$ and $h_0(\mathbf{X}^T\boldsymbol{\beta}_0)$ such that

$$b\left(h(\mathbf{X}^T\boldsymbol{\beta})\right)-b\left(h_0(\mathbf{X}^T\boldsymbol{\beta}_0)\right)=b'(c_4)\left(h(\mathbf{X}^T\boldsymbol{\beta})-h_0(\mathbf{X}^T\boldsymbol{\beta}_0)\right).$$

By the boundness of h and the continuity of b' and b'', we can show that $|E(Y|\mathbf{X})|$ and $Var(Y|\mathbf{X})$ are bounded by a positive number c_5, $b'(c_4)$ is bounded by a positive number c_6, and $h'(c_3)$ is bounded by a positive number c_7. Without loss of generality, let $\varphi=1$. Then

$$\begin{aligned}
&\mathbb{P}\left\{l(\boldsymbol{\beta},h;\mathbf{O})-l(\boldsymbol{\beta}_0,h_0;\mathbf{O})\right\}^2\\
&\le c_5^2\mathbb{P}\left[h(\mathbf{X}^T\boldsymbol{\beta})-h(\mathbf{X}^T\boldsymbol{\beta}_0)\right]^2+c_5^2\mathbb{P}\left[h(\mathbf{X}^T\boldsymbol{\beta}_0)-h_0(\mathbf{X}^T\boldsymbol{\beta}_0)\right]^2\\
&\quad+c_6^2\mathbb{P}\left[h(\mathbf{X}^T\boldsymbol{\beta})-h(\mathbf{X}^T\boldsymbol{\beta}_0)\right]^2+c_6^2\mathbb{P}\left[h(\mathbf{X}^T\boldsymbol{\beta}_0)-h_0(\mathbf{X}^T\boldsymbol{\beta}_0)\right]^2\\
&=(c_5^2+c_6^2)\mathbb{P}\left[h(\mathbf{X}^T\boldsymbol{\beta})-h(\mathbf{X}^T\boldsymbol{\beta}_0)\right]^2+(c_5^2+c_6^2)\mathbb{P}\left[h(\mathbf{X}^T\boldsymbol{\beta}_0)-h_0(\mathbf{X}^T\boldsymbol{\beta}_0)\right]^2\\
&\le(c_5^2+c_6^2)c_7^2(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T\mathbb{P}(\mathbf{X}\mathbf{X}^T)(\boldsymbol{\beta}-\boldsymbol{\beta}_0)+(c_5^2+c_6^2)||h-h_0||_2^2.
\end{aligned}$$

Since the largest eigenvalue of $\mathbb{P}(\mathbf{X}\mathbf{X}^T)$ is bounded, there exists a positive number C_2 such that

$$\mathbb{P}\left\{l(\boldsymbol{\beta},h;\mathbf{O})-l(\boldsymbol{\beta}_0,h_0;\mathbf{O})\right\}^2\le C_2(||\boldsymbol{\beta}-\boldsymbol{\beta}_0||^2+||h-h_0||_2^2).$$

Therefore we verity the second condition in Shen and Wong (1994) with $\beta=1$ in their notation.

Finally, we introduce a lemma to verify the condition C3 in Shen and Wong (1994). Let

$q_n = K_n + M$ for simplicity. According to the Corollary 6.21 of Schumaker (1981), there exists a sequence $h_{0,n} \in \mathcal{H}_n^q$ such that

$$||h_{0,n} - h_0||_\infty = O(q_n^{-q}) = O(n^{-q\nu}).$$

Let $\boldsymbol{\theta}_{0,n} = (\boldsymbol{\beta}_0, h_{0,n})$, and $\mathcal{F}_n = \{l(\boldsymbol{\theta}; \mathbf{O}) - l(\boldsymbol{\theta}_{0,n}; \mathbf{O}) : \boldsymbol{\theta} \in \Theta_n^p\}$.

Lemma 3.1. Under the regularity conditions C1-C3, the ε-bracketing number associated with $|| \cdot ||_\infty$ norm for $\mathcal{F}_n$, $N_{[]}(\varepsilon, \mathcal{F}_n, || \cdot ||_\infty)$, is bounded by $c_0(1/\varepsilon)^{cq_n+p}$ for some constant $c_0 > 0$ and $c > 0$.

We continue to verify the condition C3 in Shen and Wong (1994) before providing the proof of Lemma 3.1.

Actually, the covering number is bounded by the bracketing number, therefore the L_∞-metric entropy of the space $\mathcal{F}_n$ follows

$$H(\varepsilon, \mathcal{F}_n, || \cdot ||_\infty) = \log N_{[]}(\varepsilon, \mathcal{F}_n, || \cdot ||_\infty) \leq c_7(cq_n + p) \log(1/\varepsilon) \leq c_8 n^\nu \log(1/\varepsilon),$$

where c_7 and c_8 are positive constants. Hence the condition C3 in Shen and Wong (1994) is verified with $2r_0 = \nu$ and $r = 0^+$ in their notation.

Meanwhile, by definition, $\hat{\boldsymbol{\theta}}_n$ satisfies (1.1) in Shen and Wong (1994) with $\eta_n = 0$ in their notation.

Thus all the assumptions in Theorem 1 in Shen and Wong (1994) are satisfied. The constant τ in their theorem is $\frac{1-\nu}{2} - \frac{\log\log n}{2\log n}$ with $\frac{\log\log n}{2\log n} \to 0$ as $n \to \infty$. Moreover, the Kullback-Leilber pseudodistance of $\boldsymbol{\theta}_{0,n}$ and $\boldsymbol{\theta}_0$ is given by

$$\begin{aligned} K(\boldsymbol{\theta}_{0,n}, \boldsymbol{\theta}_0) \quad &= \mathbb{P}\left\{l(\boldsymbol{\theta}_0; \mathbf{O}) - l(\boldsymbol{\theta}_{0,n}; \mathbf{O})\right\} \\ &= -\mathbb{P}\left\{\ddot{l}_{hh}[h_{0,n} - h_0, h_{0,n} - h_0]\right\} + o(||h_{0,n} - h_0||_2^2) \end{aligned}$$

$$
\begin{aligned}
&= b''\left(h_0(\mathbf{X}^T\boldsymbol{\beta}_0)\right)\left\{h_{0,n}(\mathbf{X}^T\boldsymbol{\beta}_0) - h_0(\mathbf{X}^T\boldsymbol{\beta}_0)\right\}^2 + o(||h_{0,n} - h_0||_2^2) \\
&\leq c_9||h_{0,n} - h_0||_2^2 + o(||h_{0,n} - h_0||_2^2) \\
&\leq c_9||h_{0,n} - h_0||_\infty^2 + o(||h_{0,n} - h_0||_\infty^2) \\
&= O(n^{-2q\nu}).
\end{aligned}
$$

Hence

$$K^{\frac{1}{2}}(\boldsymbol{\theta}_{0,n}, \boldsymbol{\theta}_0) = O(n^{-q\nu}).$$

Then by Theorem 1 of Shen and Wong (1994), we obtain the convergence rate for $\hat{\boldsymbol{\theta}}_n$ as

$$d(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0) = O_p\left(\max(n^{-\frac{1-\nu}{2}}, n^{-q\nu}, n^{-q\nu})\right) = O_p(n^{-\min(q\nu, \frac{1-\nu}{2})}).$$

This completes the proof of Theorem 3.1. It remains to prove Lemma 3.1.

Let $\lceil x \rceil$ be the ceiling of a number x. Then for any ε, and $n_{b1} = \lceil (1/\varepsilon)^{a_1 q_n} \rceil$, there exists a set of brackets $\{[h_{iL}, h_{iU}] : i = 1, \cdots, n_{b1}\}$, such that for any $h \in \mathcal{H}_n^q$, we have $h_{iL}(t) < h(t) < h_{iU}$ for all t for some $1 \leq i \leq n_{b1}$, and $||h_{iU} - h_{iL}||_\infty \leq \varepsilon$. The compact set $\mathcal{B}$ can be covered by $n_{b2} = \lceil a_2(1/\varepsilon)^p \rceil$ balls with radius ε. That is, for any $\boldsymbol{\beta} \in \mathcal{B}$, there exists an $1 \leq s \leq n_{b2}$ such that $||\boldsymbol{\beta} - \boldsymbol{\beta}_s|| \leq \varepsilon$. It follows $|\mathbf{X}^T(\boldsymbol{\beta} - \boldsymbol{\beta}_s)| \leq a_3\varepsilon$. Assume $h_{iL}(\mathbf{X}^T\boldsymbol{\beta}_s + c_{iL}\varepsilon)$ and $h_{iU}(\mathbf{X}^T\boldsymbol{\beta}_s + c_{iU}\varepsilon)$ are the minumum and maximum values of h_{iL} and h_{iU} within the interval $[\mathbf{X}^T\boldsymbol{\beta}_s - a_3\varepsilon, \mathbf{X}^T\boldsymbol{\beta}_s + a_3\varepsilon]$, where the two constants $|c_{iL}|, |c_{iU}| \leq a_3$. So we can construct a set of brackets

$$\{[m_{isL}(\mathbf{O}), m_{isU}(\mathbf{O})] : i = 1, \cdots, n_{b1}, s = 1, \cdots, n_{b2}\},$$

and for any $m(\boldsymbol{\theta}; \mathbf{O}) \in \mathcal{F}_n$, there exists a pair (i, s) such that for any $\mathbf{O}$, $m(\boldsymbol{\theta}; \mathbf{O}) \in$

$[m_{isL}(\mathbf{O}), m_{isU}(\mathbf{O})]$, where

$$m_{isL}(\mathbf{O}) = Y h_{iL}(\mathbf{X}^T\boldsymbol{\beta}_s + c_{iL}\varepsilon) - b\left(h_{iL}(\mathbf{X}^T\boldsymbol{\beta}_s + c_{iL}\varepsilon)\right) - l(\boldsymbol{\theta}_{0,n}; \mathbf{O}),$$

and

$$m_{isU}(\mathbf{O}) = Y h_{iU}(\mathbf{X}^T\boldsymbol{\beta}_s + c_{iU}\varepsilon) - b\left(h_{iU}(\mathbf{X}^T\boldsymbol{\beta}_s + c_{iU}\varepsilon)\right) - l(\boldsymbol{\theta}_{0,n}; \mathbf{O}),$$

Still, we suppose $\varphi = 1$ without loss of generality. Then using the Mean Value Theorem and the similar technique in the proof of the condition C2, we can show

$$\begin{aligned}
|m_{isU}(\mathbf{O}) - m_{isL}(\mathbf{O})| =& |Y\left\{h_{iU}(\mathbf{X}^T\boldsymbol{\beta}_s + c_{iU}\varepsilon) - h_{iL}(\mathbf{X}^T\boldsymbol{\beta}_s + c_{iL}\varepsilon)\right\} \\
& - \left\{b\left(h_{iU}(\mathbf{X}^T\boldsymbol{\beta}_s + c_{iU}\varepsilon)\right) - b\left(h_{iL}(\mathbf{X}^T\boldsymbol{\beta}_s + c_{iL}\varepsilon)\right)\right\}| \\
\leq& |Y\left\{h_{iU}(\mathbf{X}^T\boldsymbol{\beta}_s + c_{iU}\varepsilon) - h_{iL}(\mathbf{X}^T\boldsymbol{\beta}_s + c_{iL}\varepsilon)\right\}| \\
& + |b\left(h_{iU}(\mathbf{X}^T\boldsymbol{\beta}_s + c_{iU}\varepsilon)\right) - b\left(h_{iL}(\mathbf{X}^T\boldsymbol{\beta}_s + c_{iL}\varepsilon)\right)| \\
\leq& a_4|h_{iU}(\mathbf{X}^T\boldsymbol{\beta}_s + c_{iU}\varepsilon) - h_{iL}(\mathbf{X}^T\boldsymbol{\beta}_s + c_{iL}\varepsilon)| \\
\leq& a_5|c_{iU} - c_{iL}|\varepsilon \\
\leq& a_6\varepsilon.
\end{aligned}$$

It follows $||m_{isU} - m_{isL}||_\infty \leq a_6\varepsilon$. Then the ε-bracketing number associated with the supreme norm for the class $\mathcal{F}_n$ satisfies

$$N_{[]}(\varepsilon, \mathcal{F}_n, ||\cdot||_\infty) \leq (1/\varepsilon)^{a_1 q_n} a_2(1(1/\varepsilon)^p = a_2(1/\varepsilon)^{a_1 q_n + p}.$$

In the proof, $a_i, i = 1, \cdots, 6$ are constant positive numbers.

$\square$

Theorem 3.2. We define $\mathbf{a}^{\otimes 2} = \mathbf{a}\mathbf{a}^T$ for a vector $\mathbf{a}$. Under the regularity conditions C1-C6,

$$n^{\frac{1}{2}}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \to N(\mathbf{0}, I^{-1}(\boldsymbol{\beta}_0))$$

in distribution, where $I(\boldsymbol{\beta}_0) = \mathbb{P}\left\{\tilde{l}_{\boldsymbol{\beta}}(\boldsymbol{\theta}_0; \mathbf{O})^{\otimes 2}\right\}$, and

$$\tilde{l}_{\boldsymbol{\beta}}(\boldsymbol{\theta}_0; \mathbf{O}) = \left\{Y - b'\left(h_0(\mathbf{X}^T\boldsymbol{\beta}_0)\right)\right\}\left\{h_0'(\mathbf{X}^T\boldsymbol{\beta}_0)\left(\mathbf{X} - \mathbb{P}(\mathbf{X}|\mathbf{X}^T\boldsymbol{\beta}_0)\right)\right\}$$

is the efficient score function of $\boldsymbol{\beta}_0$.

Remark 3.3. To prove Theorem 3.2, we check the similar assumptions of Theorem 6.1 in Wellner and Zhang (2007). Specifically, we need to verify the consistency and rate of convergence, positive information, stochastic equicontinuity of the estimators and the smoothness of the model. The proof heavily employs the empirical process theory in van der Vaart and Wellner (1996).

Proof. Following the proof of Theorem 6.1 in Wellner and Zhang (2007), we verify the following assumptions to prove Theorem 3.2.

(i) $d(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0) = O_p(n^{-\delta})$ for some $\delta > 0$.

(ii) $\mathbb{P}\dot{l}_{\boldsymbol{\beta}}(\boldsymbol{\theta}_0; \mathbf{O}) = 0$ and $\mathbb{P}\dot{l}_h(\boldsymbol{\theta}_0; \mathbf{O})[h] = 0$ for all $h \in \mathbf{H}$.

(iii) There exists an $\mathbf{h}^* = (h_1^*, h_2^*, \cdots, h_p^*)^T$, where $h_j^* \in \mathbf{H}$ for $j = 1, \cdots, p$, such that $\mathbb{P}\ddot{l}_{\boldsymbol{\beta}h}(\boldsymbol{\theta}_0; \mathbf{O})[h] - \mathbb{P}\ddot{l}_{hh}(\boldsymbol{\theta}_0; \mathbf{O})[\mathbf{h}^*, h] = \mathbf{0}$, for all $h \in \mathbf{H}$. Furthermore, the matrix $\mathbb{P}\left\{\ddot{l}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\theta}_0; \mathbf{O}) - \ddot{l}_{h\boldsymbol{\beta}}(\boldsymbol{\theta}_0; \mathbf{O})[\mathbf{h}^*]\right\}$ is nonsingular.

(iv) $\mathbb{P}_n\dot{l}_{\boldsymbol{\beta}}(\hat{\boldsymbol{\theta}}_n, \mathbf{O}) = o_p(n^{-\frac{1}{2}})$ and $\mathbb{P}_n\dot{l}_h(\hat{\boldsymbol{\theta}}_n; \mathbf{O})[\mathbf{h}^*] = o_p(n^{-\frac{1}{2}})$.

(v) For any $c > 0$,

$$\sup_{\{d(\boldsymbol{\theta},\boldsymbol{\theta}_0)\leq cn^{-\delta},\boldsymbol{\theta}\in\Theta_n^p\}} \left|\mathbb{G}_n \dot{l}_{\boldsymbol{\beta}}(\boldsymbol{\theta};\mathbf{O}) - \mathbb{G}_n \dot{l}_{\boldsymbol{\beta}}(\boldsymbol{\theta}_0;\mathbf{O})\right| = o_p(1),$$

and

$$\sup_{\{d(\boldsymbol{\theta},\boldsymbol{\theta}_0)\leq cn^{-\delta},\boldsymbol{\theta}\in\Theta_n^p\}} \left|\mathbb{G}_n \dot{l}_g(\boldsymbol{\theta};\mathbf{O})[\mathbf{h}^*] - \mathbb{G}_n \dot{l}_g(\boldsymbol{\theta}_0;\mathbf{O})[\mathbf{h}^*]\right| = o_p(1).$$

(vi) For some $\zeta > 1$ and $\delta\zeta > 1/2$, consider a neighborhood of $\boldsymbol{\theta}_0$:

$$\left\{\boldsymbol{\theta}_c : |\boldsymbol{\beta}_c - \boldsymbol{\beta}_0| + ||h_c - h_0||_2 \leq cn^{-\delta}\right\}.$$

Then

$$\left|\mathbb{P}\dot{l}_{\boldsymbol{\beta}}(\boldsymbol{\theta}_c;\mathbf{O}) - \mathbb{P}\dot{l}_{\boldsymbol{\beta}}(\boldsymbol{\theta}_0;\mathbf{O}) - \mathbb{P}\ddot{l}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\theta}_0;\mathbf{O})(\boldsymbol{\beta}-\boldsymbol{\beta}_0) - \mathbb{P}\ddot{l}_{\boldsymbol{\beta}h}(\boldsymbol{\theta}_0;\mathbf{O})[h_c - h_0]\right|$$
$$= O\left((|\boldsymbol{\beta}_c - \boldsymbol{\beta}_0| + ||h_c - h_0||_2)^\zeta\right),$$

and

$$\left|\mathbb{P}\dot{l}_h(\boldsymbol{\theta};\mathbf{O})[\mathbf{h}^*] - \mathbb{P}\dot{l}_h(\boldsymbol{\theta}_0;\mathbf{O})[\mathbf{h}^*] - \mathbb{P}\ddot{l}_{h\boldsymbol{\beta}}(\boldsymbol{\theta}_0;\mathbf{O})(\boldsymbol{\beta}-\boldsymbol{\beta}_0)[\mathbf{h}^*](\boldsymbol{\beta}-\boldsymbol{\beta}_0)\right.$$
$$\left. - \mathbb{P}\ddot{l}_{hh}(\boldsymbol{\theta}_0;\mathbf{O})[\mathbf{h}^*, h - h_0]\right| = O\left((|\boldsymbol{\beta}_c - \boldsymbol{\beta}_0| + ||h_c - h_0||_2)^\zeta\right).$$

The assumption (i) is a direct result of Theorem 3.1 with $\delta = \min(q\nu, \frac{1-\nu}{2})$. The assumption (ii) is also obvious for the score functions. We next find $\mathbf{h}^*$ such that the assumption (iii) holds. Using the same calculation in the proof of Theorem 3.1, we have

$$
\begin{aligned}
&\mathbb{P}\ddot{l}_{\boldsymbol{\beta}h}(\boldsymbol{\theta}_0;\mathbf{O})[h]-\mathbb{P}\ddot{l}_{hh}(\boldsymbol{\theta}_0;\mathbf{O})[\mathbf{h}^*,h]\\
&=-\mathbb{P}\left\{b''\left(h_0(\mathbf{X}^T\boldsymbol{\beta}_0)\right)h(\mathbf{X}^T\boldsymbol{\beta}_0)h_0'(\mathbf{X}^T\boldsymbol{\beta}_0)\mathbf{X}\right\}\\
&\quad+\mathbb{P}\left\{b''\left(h_0(\mathbf{X}^T\boldsymbol{\beta}_0)\right)h(\mathbf{X}^T\boldsymbol{\beta}_0)\mathbf{h}^*(\mathbf{X}^T\boldsymbol{\beta}_0)\right\}\\
&=\mathbb{P}\left\{b''\left(h_0(\mathbf{X}^T\boldsymbol{\beta}_0)\right)h(\mathbf{X}^T\boldsymbol{\beta}_0)\left\{h^*(\mathbf{X}^T\boldsymbol{\beta}_0)-h_0'(\mathbf{X}^T\boldsymbol{\beta}_0)\mathbf{X}\right\}\right\}.
\end{aligned}
\tag{3.2}
$$

To let (3.2)=0 holds for all $h\in\mathbf{H}$, setting $t_0=\mathbf{X}^T\boldsymbol{\beta}_0$ and using the Double Expectation Theorem we obtain

$$
\mathbf{h}^*(t_0)=h_0'(t_0)\mathbb{P}(\mathbf{X}|t_0).
$$

By the fact of zero-mean for a score function, we have

$$
\begin{aligned}
0&=\frac{\partial}{\partial\varepsilon}\mathbb{P}\dot{l}_{\boldsymbol{\beta}}(\boldsymbol{\beta}_0,h_0+\varepsilon h;\mathbf{O})\Big|_{\varepsilon=0}\\
&=\mathbb{P}\ddot{l}_{\boldsymbol{\beta}h}(\boldsymbol{\beta}_0,h_0;\mathbf{O})[h]+\mathbb{P}\left\{\dot{l}_{\boldsymbol{\beta}}(\boldsymbol{\beta}_0,h_0;\mathbf{O})\dot{l}_h(\boldsymbol{\beta}_0,h_0;\mathbf{O})[h]\right\}.
\end{aligned}
$$

Thus, we have

$$
\mathbb{P}\ddot{l}_{\boldsymbol{\beta}h}(\boldsymbol{\beta}_0,h_0;\mathbf{O})[\mathbf{h}^*]=-\mathbb{P}\left\{\dot{l}_{\boldsymbol{\beta}}(\boldsymbol{\beta}_0,h_0;\mathbf{O})\dot{l}_h^T(\boldsymbol{\beta}_0,h_0;\mathbf{O})[\mathbf{h}^*]\right\}.
$$

Similarly, we can also obtain

$$
\begin{aligned}
\mathbb{P}\ddot{l}_{h\boldsymbol{\beta}}(\boldsymbol{\beta}_0,h_0;\mathbf{O})[\mathbf{h}^*]&=-\mathbb{P}\left\{\dot{l}_g(\boldsymbol{\beta}_0,h_0;\mathbf{O})[\mathbf{h}^*]\dot{l}_{\boldsymbol{\beta}}^T(\boldsymbol{\beta}_0,h_0;\mathbf{O})\right\},\\
\mathbb{P}\ddot{l}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\beta}_0,h_0;\mathbf{O})&=-\mathbb{P}\left\{\dot{l}_{\boldsymbol{\beta}}(\boldsymbol{\beta}_0,h_0;\mathbf{O})\dot{l}_{\boldsymbol{\beta}}^T(\boldsymbol{\beta}_0,h_0;\mathbf{O})\right\},\\
\mathbb{P}\ddot{l}_{hh}(\boldsymbol{\beta}_0,h_0;\mathbf{O})[\mathbf{h}^*,\mathbf{h}^*]&=-\mathbb{P}\left\{\dot{l}_h(\boldsymbol{\beta}_0,h_0;\mathbf{O})[\mathbf{h}^*]\dot{l}_h^T(\boldsymbol{\beta}_0,h_0;\mathbf{O})[\mathbf{h}^*]\right\}.
\end{aligned}
$$

Then

$$
\begin{aligned}
& -\mathbb{P}\ddot{l}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\theta}_0;\mathbf{O}) + \mathbb{P}\ddot{l}_{h\boldsymbol{\beta}}(\boldsymbol{\theta}_0;\mathbf{O})[\mathbf{h}^*] \\
& = -\mathbb{P}\ddot{l}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\theta}_0;\mathbf{O}) + \mathbb{P}\ddot{l}_{h\boldsymbol{\beta}}(\boldsymbol{\theta}_0;\mathbf{O})[\mathbf{h}^*] \\
& \quad + \mathbb{P}\ddot{l}_{\boldsymbol{\beta}h}(\boldsymbol{\theta}_0;\mathbf{O})[h] - \mathbb{P}\ddot{l}_{hh}(\boldsymbol{\theta}_0;\mathbf{O})[\mathbf{h}^*, h] \\
& = \mathbb{P}\left\{\dot{l}_{\boldsymbol{\beta}}(\boldsymbol{\beta}_0, h_0;\mathbf{O})\dot{l}_{\boldsymbol{\beta}}^T(\boldsymbol{\beta}_0, h_0;\mathbf{O}) - \dot{l}_{\boldsymbol{\beta}}(\boldsymbol{\beta}_0, h_0;\mathbf{O})\dot{l}_h^T(\boldsymbol{\beta}_0, h_0;\mathbf{O})[\mathbf{h}^*]\right. \\
& \quad \left. -\dot{l}_g(\boldsymbol{\beta}_0, h_0;\mathbf{O})[\mathbf{h}^*]\dot{l}_{\boldsymbol{\beta}}^T(\boldsymbol{\beta}_0, h_0;\mathbf{O}) + \dot{l}_h(\boldsymbol{\beta}_0, h_0;\mathbf{O})[\mathbf{h}^*]\dot{l}_h^T(\boldsymbol{\beta}_0, h_0;\mathbf{O})[\mathbf{h}^*]\right\} \\
& = \mathbb{P}\left\{\dot{l}_{\boldsymbol{\beta}}(\boldsymbol{\beta}_0, h_0;\mathbf{O}) - \dot{l}_h(\boldsymbol{\beta}_0, h_0;\mathbf{O})[\mathbf{h}^*]\right\}^{\otimes 2}.
\end{aligned}
\tag{3.3}
$$

Note that $\tilde{l}_{\boldsymbol{\beta}}(\boldsymbol{\theta}_0;\mathbf{O}) = \dot{l}_{\boldsymbol{\beta}}(\boldsymbol{\beta}_0, h_0;\mathbf{O}) - \dot{l}_h(\boldsymbol{\beta}_0, h_0;\mathbf{O})[\mathbf{h}^*]$ is the efficient score function of $\boldsymbol{\beta}_0$. The last term in (3.3) is the information matrix for $\boldsymbol{\beta}_0$ under semiparametric efficiency and it is nonsingular.

Next we prove the assumption (iv). The first part holds since

$$\mathbb{P}_n\dot{l}_{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}_n, \hat{h}_n;\mathbf{O}) = 0.$$

Meanwhile,

$$\mathbb{P}_n\dot{l}_h(\hat{\boldsymbol{\beta}}_n, \hat{h}_n;\mathbf{O})[h_j^*] = \mathbb{P}_n\left\{Yh_j^*(\mathbf{X}^T\boldsymbol{\beta}) - b'\left(\hat{h}_n(\mathbf{X}^T\boldsymbol{\beta})\right)h_j^*(\mathbf{X}^T\boldsymbol{\beta})\right\}.$$

According to the Corollary 6.21 of Schumaker (1981), we can also show that there exists a sequence $h_{j,n}^* \in \mathcal{H}_n^2$ such that $||h_j^* - h_{j,n}^*||_\infty = O(n^{-2\nu})$. Also,

$$\mathbb{P}_n\left\{Yh_{j,n}^*(\mathbf{X}^T\boldsymbol{\beta}) - b'\left(\hat{h}_n(\mathbf{X}^T\boldsymbol{\beta})\right)h_{j,n}^*(\mathbf{X}^T\boldsymbol{\beta})\right\} = 0.$$

So it suffices to show that for $j = 1, \cdots, p$,

$$\mathbb{P}_n \dot{l}_h(\hat{\boldsymbol{\beta}}_n, \hat{g}_n; \mathbf{O})[h_j^* - h_{j,n}^*] = o_p(n^{-\frac{1}{2}}).$$

Notice

$$\mathbb{P}_n \dot{l}_h(\hat{\boldsymbol{\beta}}_n, \hat{g}_n; \mathbf{O})[h_j^* - h_{j,n}^*]$$

$$= (\mathbb{P}_n - \mathbb{P})\dot{l}_h(\hat{\boldsymbol{\beta}}_n, \hat{g}_n; \mathbf{O})[h_j^* - h_{j,n}^*] + \mathbb{P}\left\{\dot{l}_h(\hat{\boldsymbol{\beta}}_n, \hat{g}_n; \mathbf{O})[h_j^* - h_{j,n}^*] - \dot{l}_h(\boldsymbol{\beta}_0, g_0; \mathbf{O})[h_j^* - h_{j,n}^*]\right\}$$

$$= A_{1n} + A_{2n}.$$

Before further proof, we introduce a lemma.

Lemma 3.2. For $j = 1, \cdots, p$, define

$$\mathcal{F}_n^j(\xi) = \left\{\dot{l}_h(\boldsymbol{\theta}; \mathbf{O})[h_j^* - h] : \boldsymbol{\theta} \in \Theta_n^p, h \in \mathcal{H}_n^2 \text{ and } d(\boldsymbol{\theta}, \boldsymbol{\theta}_0) \le \xi, ||h - h_j^*||_\infty \le \xi\right\}.$$

Under the regularity conditions, $N_{[]}(\varepsilon, \mathcal{F}_n^j(\xi), ||\cdot||_\infty) \le c_0(\xi/\varepsilon)^{cq_n+p}$ for some positive constant c_0 and c.

The proof of Lemma 3.2 is similar to the proof of Lemma 3.1 by constructing a set of baskets for $\mathcal{F}_n^j(\xi)$ and proving their radii are bounded by ε times a positive number.

It follows from the lemma that the bracketing integral

$$J_{[]}(\xi, \mathcal{F}_n^j(\xi), L_2(\mathbb{P})) = \int_0^\xi \sqrt{1 + \log N_{[]}(\varepsilon, \mathcal{F}_n^j(\xi), L_2(\mathbb{P}))}d\varepsilon$$

$$\le \int_0^\xi \sqrt{1 + \log N_{[]}(\varepsilon, \mathcal{F}_n^j(\xi), ||\cdot||_\infty)}d\varepsilon \le b_1 q_n^{\frac{1}{2}} \xi.$$

Let $\xi = \xi_n = O\left(n^{-\min\left(2\nu, \frac{1-\nu}{2}\right)}\right)$. Then since $q \geq 3$,

$$d(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0) = O_p\left(n^{-\min\left(q\nu, \frac{1-\nu}{2}\right)}\right) \leq O_p\left(n^{-\min\left(2\nu, \frac{1-\nu}{2}\right)}\right) = \xi_n,$$

and

$$||h_j^* - h_{j,n}^*||_\infty = O(n^{-2\nu}) \leq O_p\left(n^{-\min\left(2\nu, \frac{1-\nu}{2}\right)}\right) = \xi_n.$$

Therefore $\dot{l}_h(\hat{\boldsymbol{\beta}}_n, \hat{h}_n; \mathbf{O})[h_j^* - h_{j,n}^*] \in \mathcal{F}_n^j(\xi)$. Similar to the proof technique in Theorem 3.1, it is easy to show

$$\mathbb{P}\left\{\dot{l}_h(\boldsymbol{\theta}; \mathbf{O})[h_j^* - h_{j,n}^*]\right\}^2 \leq b_2 ||h_j^* - h_{j,n}^*||_\infty^2 \leq b_2 \xi_n.$$

In addition, by direct calculation, we can show that $||\dot{l}_h(\boldsymbol{\theta}; \mathbf{O})[h_j^* - h_{j,n}^*]||_\infty$ is bounded by some positive constant M. Then by the maximal inequality in Lemma 3.4.2 in van der Vaart and Wellner (1996), we have

$$E_{\mathbb{P}} ||\mathbb{G}_n||_{\mathcal{F}_n^j(\xi_n)} \leq b_3 J_{[]}(\xi, \mathcal{F}_n^j(\xi), L_2(\mathbb{P})) \left(1 + \frac{J_{[]}(\xi, \mathcal{F}_n^j(\xi), L_2(\mathbb{P}))}{\xi_n^2 \sqrt{n}} M\right)$$

$$\leq b_4 \left(q_n^{\frac{1}{2}} \xi_n + q_n n^{-\frac{1}{2}}\right) = O\left(n^{-\min\left(\frac{3\nu}{2}, \frac{1}{2} - \nu\right)}\right) + O\left(n^{\nu - \frac{1}{2}}\right) = o(1).$$

Thus

$$A_{1n} = n^{-\frac{1}{2}} \mathbb{G}_n \dot{l}_h(\hat{\boldsymbol{\theta}}_n; \mathbf{O})[h_j^* - h_{j,n}^*] = o_p(n^{-\frac{1}{2}}).$$

The Taylor series expansion for $\dot{l}_h(\hat{\boldsymbol{\theta}}_n; \mathbf{O})[h_j^* - h_{j,n}^*]$ at $\boldsymbol{\theta}_0$ is given by

$$\dot{l}_h(\hat{\boldsymbol{\theta}}_n; \mathbf{O})[h_j^* - h_{j,n}^*] - \dot{l}_h(\hat{\boldsymbol{\theta}}_0; \mathbf{O})[h_j^* - h_{j,n}^*]$$

$$= (\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)^T \ddot{l}_{\boldsymbol{\beta} h}(\boldsymbol{\theta}_{cn}; \mathbf{O})[h_j^* - h_{j,n}^*] + \ddot{l}_{hh}(\boldsymbol{\theta}_{cn}; \mathbf{O})[h_j^* - h_{j,n}^*, \hat{h}_n - h_0],$$

where $\boldsymbol{\theta}_{cn}$ is the intermediate value between $\boldsymbol{\theta}_0$ and $\hat{\boldsymbol{\theta}}_n$. Using similar techniques in the proof of Theorem 3.1, we obtain

$$\left|\dot{l}_h(\hat{\boldsymbol{\theta}}_n; \mathbf{O})[h_j^* - h_{j,n}^*]\right| \leq b_5 \left(||h_j^* - h_{j,n}^*||_\infty + ||h_j^{*\prime} - h_{j,n}^{*\prime}||_\infty\right).$$

According to the Corollary 6.21 of Schumaker (1981), $||h_j^{*\prime} - h_{j,n}^{*\prime}||_\infty = O(n^{-\nu})$. Thus,

$$\begin{aligned}
\left|(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)^T \dot{l}_h(\hat{\boldsymbol{\theta}}_n; \mathbf{O})[h_j^* - h_{j,n}^*]\right| &= |\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0| \left\{O(n^{-\nu}) + O(n^{-2\nu})\right\} \\
&= O_p\left(n^{-\min\left(q\nu, \frac{1-\nu}{2}\right)}\right) \cdot O(n^{-\nu}) \\
&= O_p\left(n^{-\min\left((q+1)\nu, \frac{1+3\nu}{2}\right)}\right).
\end{aligned}$$

In addition,

$$\begin{aligned}
\left|\ddot{l}_{hh}(\boldsymbol{\theta}_{cn}; \mathbf{O})[h_j^* - h_{j,n}^*, \hat{h}_n - h_0]\right| &\leq b_6 ||h_j^* - h_{j,n}^*||_\infty \cdot ||\hat{h}_n - h_0||_2 \\
&= O(n^{-2\nu}) \cdot O_p\left(n^{-\min\left(q\nu, \frac{1-\nu}{2}\right)}\right) \\
&= O_p\left(n^{-\min\left((q+2)\nu, \frac{1+3\nu}{2}\right)}\right).
\end{aligned}$$

Since $\frac{1}{2(1+q)} < \nu < \frac{1}{1+2q}$, it follows that

$$A_{2n} = o_p(n^{-\frac{1}{2}}).$$

Therefore, we have shown that the assumption (iv) holds.

To verify the next assumption, we need to use the following lemma.

Lemma 3.3. For $j = 1, \cdots, p$, define

$$\mathcal{F}_{nj,\boldsymbol{\beta}}(\xi) = \left\{\dot{l}_{\beta_j}(\boldsymbol{\theta}; \mathbf{O}) - \dot{l}_{\beta_j}(\boldsymbol{\theta}_0; \mathbf{O}) : \boldsymbol{\theta} \in \Theta_n^p, h' \in \mathcal{H}_n^{p-1} \text{ and } d(\boldsymbol{\theta}, \boldsymbol{\theta}_0) \leq \xi, ||h' - h_0'||_\infty \leq \xi\right\}.$$

and

$$\mathcal{F}_{nj,h}(\xi) = \left\{ \dot{l}_h(\boldsymbol{\theta}; \mathbf{O})[h_j^*] - \dot{l}_h(\boldsymbol{\theta}_0; \mathbf{O})[h_j^*] : \boldsymbol{\theta} \in \Theta_n^p \text{ and } d(\boldsymbol{\theta}, \boldsymbol{\theta}_0) \leq \xi \right\}.$$

Under the regularity conditions,

$$N_{[]}(\varepsilon, \mathcal{F}_{nj,\boldsymbol{\beta}}(\xi), ||\cdot||_\infty) \leq c_1(\xi/\varepsilon)^{c_2 q_n + p},$$

and

$$N_{[]}(\varepsilon, \mathcal{F}_{nj,h}^j(\xi), ||\cdot||_\infty) \leq c_3(\xi/\varepsilon)^{c_4 q_n + p},$$

for some positive constant c_1, c_2, c_3 and c_4.

The proof of Lemma 3.3 is similar to the proof of Lemma 3.1 by constructing two sets of baskets for $\mathcal{F}_{nj,h}^j(\xi)$ and $\mathcal{F}_{nj,h}^j(\xi)$ respectively, and proving the radii of the baskets are bounded by ε times a positive number.

From Lemma 3.3, we have

$$J_{[]}(\xi, \mathcal{F}_{nj,\boldsymbol{\beta}}(\xi), L_2(\mathbb{P})) \leq b_7 q_n^{1/2} \xi$$

and

$$J_{[]}(\xi, \mathcal{F}_{nj,h}^j(\xi), L_2(\mathbb{P})) \leq b_8 q_n^{1/2} \xi.$$

By the Cauchy-Schwartz inequality,

$$\begin{aligned}
&\left\{ \dot{l}_{\beta_j}(\boldsymbol{\theta}; \mathbf{O}) - \dot{l}_{\beta_j}(\boldsymbol{\theta}_0; \mathbf{O}) \right\}^2 \\
&= \left[\left\{ Y - b'\left(h(\mathbf{X}^T\boldsymbol{\beta}) \right) \right\} h'(\mathbf{X}^T\boldsymbol{\beta}) X_j - \left\{ Y - b'\left(h_0(\mathbf{X}^T\boldsymbol{\beta}_0) \right) \right\} h_0'(\mathbf{X}^T\boldsymbol{\beta}_0) X_j \right]^2 \\
&\leq \left[Y X_j (h - h_0)(\mathbf{X}^T\boldsymbol{\beta}) \right]^2 + \left[Y X_j \left\{ h_0(\mathbf{X}^T\boldsymbol{\beta}) - h_0(\mathbf{X}^T\boldsymbol{\beta}_0) \right\} \right]^2 \\
&\quad + X_j^2 \left[b'\left(h(\mathbf{X}^T\boldsymbol{\beta}) \right) h'(\mathbf{X}^T\boldsymbol{\beta}) - b'\left(h_0(\mathbf{X}^T\boldsymbol{\beta}) \right) h'(\mathbf{X}^T\boldsymbol{\beta}) \right]^2
\end{aligned}$$

$$
\begin{aligned}
&+X_j^2\left[b'\left(h_0(\mathbf{X}^T\boldsymbol{\beta})\right)h'(\mathbf{X}^T\boldsymbol{\beta})-b'\left(h_0(\mathbf{X}^T\boldsymbol{\beta}_0)\right)h'(\mathbf{X}^T\boldsymbol{\beta})\right]^2\\
&+X_j^2\left[b'\left(h_0(\mathbf{X}^T\boldsymbol{\beta}_0)\right)h'(\mathbf{X}^T\boldsymbol{\beta})-b'\left(h_0(\mathbf{X}^T\boldsymbol{\beta}_0)\right)h_0'(\mathbf{X}^T\boldsymbol{\beta})\right]^2 \qquad (3.4)\\
&+X_j^2\left[b'\left(h_0(\mathbf{X}^T\boldsymbol{\beta}_0)\right)h_0'(\mathbf{X}^T\boldsymbol{\beta})-b'\left(h_0(\mathbf{X}^T\boldsymbol{\beta}_0)\right)h_0'(\mathbf{X}^T\boldsymbol{\beta}_0)\right]^2\\
&=B_1+B_2+B_3+B_4+B_5+B_6.
\end{aligned}
$$

We again apply the technique used in the proof of Theorem 3.1, and obtain

$$
\begin{aligned}
\mathbb{P}B_1 &\le ||h-h_0||_2^2 \le \xi^2,\\
\mathbb{P}B_2 &\le |\boldsymbol{\beta}-\boldsymbol{\beta}_0|^2 \le \xi^2,\\
\mathbb{P}B_3 &\le ||h-h_0||_2^2 \le \xi^2,\\
\mathbb{P}B_4 &\le |\boldsymbol{\beta}-\boldsymbol{\beta}_0|^2 \le \xi^2,\\
\mathbb{P}B_5 &\le ||h'-h_0'||_2^2 \le \xi^2,\\
\mathbb{P}B_6 &\le |\boldsymbol{\beta}-\boldsymbol{\beta}_0|^2 \le \xi^2,
\end{aligned}
$$

where we omit the positive constant on the right-hand side. Hence $\mathbb{P}\left\{\dot{l}_{\beta_j}(\boldsymbol{\theta};\mathbf{O})-\dot{l}_{\beta_j}(\boldsymbol{\theta}_0;\mathbf{O})\right\}^2$ $\le \xi^2$. Using the similar argument, it is easy to show $\mathbb{P}\left\{\dot{l}_h(\boldsymbol{\theta};\mathbf{O})[h_j^*]-\dot{l}_h(\boldsymbol{\theta}_0;\mathbf{O})[h_j^*]\right\}^2 \le \xi^2$. Also, $||\dot{l}_{\beta_j}(\boldsymbol{\theta};\mathbf{O})-\dot{l}_{\beta_j}(\boldsymbol{\theta}_0;\mathbf{O})||_\infty$ and $||\dot{l}_h(\boldsymbol{\theta};\mathbf{O})[h_j^*]-\dot{l}_h(\boldsymbol{\theta}_0;\mathbf{O})[h_j^*]||_\infty$ are bounded. Pick $\xi=\xi_n=O\left(n^{-\min\left(q\nu,\frac{1-\nu}{2}\right)}\right)$, then by the maximal inequality in Lemma 3.4.2 of van der Vaart and Wellner (1996),

$$
E_{\mathbb{P}}||\mathbb{G}_n||_{\mathcal{F}_{nj,\boldsymbol{\beta}}(\xi_n)} \le b_9\left(q_n^{\frac{1}{2}}\xi_n+q_n n^{-\frac{1}{2}}\right)=O\left(n^{-\min\left((q-\frac{1}{2})\nu,\frac{1}{2}-\nu\right)}\right)+O\left(n^{\nu-\frac{1}{2}}\right)=o(1).
$$

Similarly,

$$E_{\mathbb{P}}||\mathbb{G}_n||_{\mathcal{F}^j_{nj,h}(\xi_n)} = o(1).$$

Thus taking $\delta = \min\left(q\nu, \frac{1-\nu}{2}\right)$, by the Markov's inequality,

$$\sup_{\left\{d(\boldsymbol{\theta},\boldsymbol{\theta}_0)\le cn^{-\delta},\boldsymbol{\theta}\in\Theta_n^p\right\}} \left|\mathbb{G}_n \dot{l}_{\boldsymbol{\beta}}(\boldsymbol{\theta};\mathbf{O}) - \mathbb{G}_n \dot{l}_{\boldsymbol{\beta}}(\boldsymbol{\theta}_0;\mathbf{O})\right| = o_p(1),$$

and

$$\sup_{\left\{d(\boldsymbol{\theta},\boldsymbol{\theta}_0)\le cn^{-\delta},\boldsymbol{\theta}\in\Theta_n^p\right\}} \left|\mathbb{G}_n \dot{l}_g(\boldsymbol{\theta};\mathbf{O})[\mathbf{h}^*] - \mathbb{G}_n \dot{l}_g(\boldsymbol{\theta}_0;\mathbf{O})[\mathbf{h}^*]\right| = o_p(1),$$

which completes the verification of assumption (v). For the last assumption, we consider a neighborhood of $\boldsymbol{\theta}_0$: $\left\{\boldsymbol{\theta} : |\boldsymbol{\beta}_c - \boldsymbol{\beta}_0| + ||h_c - h_0||_2 \le cn^{-\delta}\right\}$ with $\delta = \min\left(q\nu, \frac{1-\nu}{2}\right)$. By the Taylor series expansion, we have

$$\dot{l}_{\boldsymbol{\beta}}(\boldsymbol{\beta}, h;\mathbf{O}) - \dot{l}_{\boldsymbol{\beta}}(\boldsymbol{\beta}_0, h_0;\mathbf{O}) = \ddot{l}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\beta}_c, h_c;\mathbf{O})(\boldsymbol{\beta}-\boldsymbol{\beta}_0) + \ddot{l}_{\boldsymbol{\beta}h}(\boldsymbol{\beta}_c, h_c;\mathbf{O})[h-h_0],$$

where $(\boldsymbol{\beta}_c, h_c)$ are an intermediate value between $(\boldsymbol{\beta}_0, h_0)$ and $(\boldsymbol{\beta}, h)$. Then

$$\begin{aligned}
&\left|\mathbb{P}\dot{l}_{\boldsymbol{\beta}}(\boldsymbol{\theta};\mathbf{O}) - \mathbb{P}\dot{l}_{\boldsymbol{\beta}}(\boldsymbol{\theta}_0;\mathbf{O}) - \mathbb{P}\ddot{l}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\theta}_0;\mathbf{O})(\boldsymbol{\beta}-\boldsymbol{\beta}_0) - \mathbb{P}\ddot{l}_{\boldsymbol{\beta}h}(\boldsymbol{\theta}_0;\mathbf{O})[h-h_0]\right| \\
&= \left|\mathbb{P}\left\{\ddot{l}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\beta}_c, h_c;\mathbf{O}) - \ddot{l}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\beta}_0, h_0;\mathbf{O})\right\}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\right. \\
&\qquad \left. + \mathbb{P}\left\{\ddot{l}_{\boldsymbol{\beta}h}(\boldsymbol{\beta}_c, h_c;\mathbf{O}) - \ddot{l}_{\boldsymbol{\beta}h}(\boldsymbol{\beta}_0, h_0;\mathbf{O})[h-h_0]\right\}\right| \\
&\le \left|\mathbb{P}\left[\left\{b''\left(h_0(\mathbf{X}^T\boldsymbol{\beta}_0)\right)h_0'^2(\mathbf{X}^T\boldsymbol{\beta}_0) - b''\left(h_c(\mathbf{X}^T\boldsymbol{\beta}_c)\right)h_c'^2(\mathbf{X}^T\boldsymbol{\beta}_c)\right\}\mathbf{X}\mathbf{X}^T(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\right]\right| \\
&\qquad + \left|\mathbb{P}\left[\left\{b''\left(h_0(\mathbf{X}^T\boldsymbol{\beta}_0)\right)h_0'(\mathbf{X}^T\boldsymbol{\beta}_0)(h-h_0)(\mathbf{X}^T\boldsymbol{\beta}_0)\right.\right.\right. \\
&\qquad\quad \left.\left.\left. -b''\left(h_c(\mathbf{X}^T\boldsymbol{\beta}_c)\right)h_c'(\mathbf{X}^T\boldsymbol{\beta}_c)(h-h_0)(\mathbf{X}^T\boldsymbol{\beta}_c)\right\}\mathbf{X}\right]\right|.
\end{aligned} \tag{3.5}$$

Using the same technique in (3.4) by adding and subtracting some terms, we continue to work on (3.5) as follows

$$
\begin{aligned}
(3.5) &\leq \left(|\boldsymbol{\beta}_c - \boldsymbol{\beta}_0| + ||h'_c - h'_0||_2 + ||h_c - h_0||_2\right) \cdot |\boldsymbol{\beta} - \boldsymbol{\beta}_0| + |\boldsymbol{\beta}_c - \boldsymbol{\beta}_0| + ||h'_c - h'_0||_2 + ||h_c - h_0||_2 \\
&= \left\{O(n^{-\delta}) + O(n^{-\min\left((q-1)\nu, \frac{1-\nu}{2}\right)})\right\} \cdot O(n^{-\delta}) + O(n^{-\delta}) + O(n^{-\min\left((q-1)\nu, \frac{1-\nu}{2}\right)}) \\
&= O(n^{-\min\left((q-1)\nu, \frac{1-\nu}{2}\right)}) \cdot O(n^{-\delta}) + O(n^{-\min\left((q-1)\nu, \frac{1-\nu}{2}\right)}) \\
&= O(n^{-\min\left((2q-1)\nu, \frac{1}{2} + (q - \frac{2}{5}\nu), 1-\nu\right)}) \\
&= O(n^{-\zeta\delta}) = o(n^{-\frac{1}{2}}),
\end{aligned}
$$

where $\zeta = \min\left((2q-1)\nu, \frac{1}{2} + (q - \frac{2}{5}\nu), 1 - \nu\right)/\delta > 1$ and $\zeta\delta > \frac{1}{2}$. Similarly, the second equation in the assumption (vi) can be proved. Therefore, we have verified all the assumptions. Notice in the proof of assumption (iii), the A matrix in Wellner and Zhang (2007) is given by $\mathbb{P}\tilde{l}_{\boldsymbol{\beta}}(\boldsymbol{\theta}_0; \mathbf{O})^{\otimes 2} = I^{-1}(\boldsymbol{\beta}_0)$. Therefore, we complete the proof of Theorem 3.2.

□

A consistent estimator for the asymptotic variance-covariance matrix is provided in the following theorem.

Theorem 3.3. Under the regularity conditions C1-C6,

$$
\mathbb{P}_n\left\{\tilde{l}_{\boldsymbol{\beta}}(\hat{\boldsymbol{\theta}}_n, \mathbf{O})^{\otimes 2}\right\} \to \mathbb{P}\left\{\tilde{l}_{\boldsymbol{\beta}}(\boldsymbol{\theta}_0, \mathbf{O})^{\otimes 2}\right\}
$$

in probability, where

$$
\tilde{l}_{\boldsymbol{\beta}}(\hat{\boldsymbol{\theta}}_n, \mathbf{O}) = \left\{Y - b'\left(\hat{h}_n(\mathbf{X}^T\hat{\boldsymbol{\beta}}_n)\right)\right\}\left\{\hat{h}'_n(\mathbf{X}^T\hat{\boldsymbol{\beta}}_n)\left(\mathbf{X} - \mathbb{P}(\mathbf{X}|\mathbf{X}^T\hat{\boldsymbol{\beta}}_n)\right)\right\}.
$$

Proof. Define

$$I_{jk}(\boldsymbol{\theta}_0) = \mathbb{P}\left[\left\{\dot{l}_{\beta_j}(\boldsymbol{\theta}_0;\mathbf{O}) - \dot{l}_h(\boldsymbol{\theta}_0;\mathbf{O})[h_j^*]\right\} \times \left\{\dot{l}_{\beta_k}(\boldsymbol{\theta}_0;\mathbf{O}) - \dot{l}_h(\boldsymbol{\theta}_0;\mathbf{O})[h_k^*]\right\}\right] = \mathbb{P}A_{jk}(\boldsymbol{\theta}_0;\mathbf{O})$$

and

$$I_{jkn}(\hat{\boldsymbol{\theta}}_n) = \mathbb{P}_n\left[\left\{\dot{l}_{\beta_j}(\hat{\boldsymbol{\theta}}_n;\mathbf{O}) - \dot{l}_h(\hat{\boldsymbol{\theta}}_n;\mathbf{O})[h_j^*]\right\} \times \left\{\dot{l}_{\beta_k}(\hat{\boldsymbol{\theta}}_n;\mathbf{O}) - \dot{l}_h(\hat{\boldsymbol{\theta}}_n;\mathbf{O})[h_k^*]\right\}\right] = \mathbb{P}_n A_{jkn}(\hat{\boldsymbol{\theta}}_n;\mathbf{O}).$$

Then

$$\begin{aligned} I_{jkn}(\hat{\boldsymbol{\theta}}_n) - I_{jk}(\boldsymbol{\theta}_0) &= \mathbb{G}_n A_{jkn}(\hat{\boldsymbol{\theta}}_n;\mathbf{O}) + \mathbb{P}\left\{A_{jkn}(\hat{\boldsymbol{\theta}}_n;\mathbf{O}) - A_{jk}(\boldsymbol{\theta}_0;\mathbf{O})\right\} \\ &= I_{1n} + I_{2n}. \end{aligned}$$

The following work is similar to the technique details of the verification of the assumption (iv) in the proof of Theorem 3.2. We construct functional spaces

$$\begin{aligned} \mathcal{F}_{nj,\boldsymbol{\beta},h}(\xi) = \Big\{&\dot{l}_{\beta_j}(\boldsymbol{\theta};\mathbf{O}) - \dot{l}_h(\boldsymbol{\theta};\mathbf{O})[h_j] : \boldsymbol{\theta} \in \Theta_n^p, h_j' \in \mathcal{H}_n^2 \\ &\text{and } d(\boldsymbol{\theta},\boldsymbol{\theta}_0) \le \xi, ||h' - h_0'||_2 \le \xi, ||h_j - h_j^*||_\infty \le \xi\Big\} \end{aligned}$$

and

$$\begin{aligned} \mathcal{F}_{njk,\boldsymbol{\beta},h}(\xi) = \Big\{ &\left(\dot{l}_{\beta_j}(\boldsymbol{\theta};\mathbf{O}) - \dot{l}_h(\boldsymbol{\theta};\mathbf{O})[h_j]\right)\left(\dot{l}_{\beta_k}(\boldsymbol{\theta};\mathbf{O}) - \dot{l}_h(\boldsymbol{\theta};\mathbf{O})[h_k]\right) : \boldsymbol{\theta} \in \Theta_n^p, h_j', h_k' \in \mathcal{H}_n^2 \\ &\text{and } d(\boldsymbol{\theta},\boldsymbol{\theta}_0) \le \xi, ||h' - h_0'||_2 \le \xi, ||h_j - h_j^*||_\infty \le \xi, ||h_k - h_k^*||_\infty \le \xi\Big\}. \end{aligned}$$

Moreover,

$$N_{[]}(\varepsilon, \mathcal{F}_{nj,\boldsymbol{\beta},h}(\xi), ||\cdot||_\infty) \le (\xi/\varepsilon)^{cq_n+p},$$

and

$$N_{[]}(\varepsilon, \mathcal{F}_{njk,\boldsymbol{\beta},h}(\xi), ||\cdot||_\infty) \leq (\xi/\varepsilon)^{cq_n+p},$$

where we omit the positive number on the right-hand side and c is a positive constant. This result can be easily obtained by following the proof of Lemma 3.1 by constructing a set ε-baskets. Using the similar argument in the verification of the assumption (iv) in the proof of Theorem 3.2, it is straightforward to show $I_{1n} = o_p(1)$ and $I_{2n} = o_p(1)$.

In the above proof, $b_i, i = 1, \cdots, 9$ are constant positive numbers.

□

3.4 Simulation Studies

We conduct simulation studies to examine the finite-sample performance of the proposed method. In the first set of simulation studies, we consider the normal response. We include three covariates X_1, X_2 and X_3 in model (3.1). Specifically, X_1 is generated from a normal distribution with mean 1 and variance 0.5^2, X_2 is generated from the uniform distribution $U(-4.25, 2.25)$, and X_3 is generated from a bernoulli distribution with success probability 0.5. We set the true values $\boldsymbol{\beta}_0 = \frac{1}{\sqrt{2}}(0.6, 0.8, 1)$. Then the outcome Y is generated from a normal distribution with mean $h(\mathbf{X}^T\boldsymbol{\beta}_0)$ and variance 0.2. We consider two different scenarios for $h(\cdot)$: (i) $h(t) = t$; and (ii) $h(t) = cos(t)$. We adopt three data-adaptive interior knots that were placed at the equally spaced sample quantiles of $\mathbf{X}_i^T\boldsymbol{\beta}, i = 1, \cdots, n$ for a given $\boldsymbol{\beta}$, and employ the cubic ($M = 4$) B-spline to approximate the unknown function. Therefore, we obtain the sieve MLEs iteratively by using the following procedures.

(1) Choose the initial values $\boldsymbol{\beta}$.

(2) Given $\boldsymbol{\beta}$, calculate the knots and basis functions.

(3) Given the knots and basis functions, obtain $\hat{\boldsymbol{\gamma}}$ by solving $\partial l_n(\boldsymbol{\beta}, \boldsymbol{\gamma})/\partial\boldsymbol{\gamma} = \mathbf{0}$.

(4) Given the knots, basis functions and $\hat{\boldsymbol{\gamma}}$, obtain $\hat{\boldsymbol{\beta}}$ by solving $\partial l_n(\boldsymbol{\beta}, \boldsymbol{\gamma})/\partial\boldsymbol{\beta} = \mathbf{0}$.

(5) Set the initial values for next iteration as $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$.

(6) Repeat steps (2)-(5) until the maximum of the component-wise absolute value difference of $\hat{\boldsymbol{\beta}}$ between two consecutive iterations is less than 10^{-3}.

The quasi-Newton algorithm is used in steps (3) and (4) to update the estimates of $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$. We estimate the variance-covariance matrix for the regression parameters $\boldsymbol{\beta}$ using the inverse of the consistent variance-covariance matrix estimator for the score functions. For comparison, we also analyze the data using the standard GLM based on the normal distribution with an intercept and the identity link function. For fair comparison, we also standardize the estimates of the regression parameters (excluding the intercept) of the GLM such that the Euclidean norm is 1. For each case, we consider the sample sizes $n = 300$, 500 and 1000 with 10000 replicates.

Tables 3.1 and 3.2 present the summary statistics of the sieve MLEs of the unknown regression parameters and the MLEs of the regression parameters under the GLM. In these tables, Bias is the sampling average of the biases of the estimates; SE denotes the empirical standard deviations of the parameter estimates; SEE denotes the average of the standard error estimates; CP is the coverage probability of 95% confidence interval estimates; and MSE is the mean squared error of the parameter estimates. The last column RE is the MSE relative efficiency of the proposed estimator compared to the estimator of the GLM.

From Tables 3.1 and 3.2, we observe that the biases of the proposed sieve MLEs are small under all simulation settings. The standard error estimates exhibit the true variation well. The coverage probabilities of the 95% confidence interval estimates are close to the nominal level. The relative efficiency with $h(t) = t$ in Table 3.1 are all less than 1. It implies that model (3.1) loses some efficiency when the true model is the GLM. The loss of efficiency is caused by estimating the unknown link function in model (3.1). We also notice that when the sample size increases, the relative efficiency improves. The sieve MLEs appear to be more efficient with $h(t) = cos(t)$ in Table 3.2 in terms of the mean squared error. Moreover, the estimates of the GLM have larger biases in Table 3.2 compared to those in Table 3.1.

In the next set of the simulation studies, we consider the count data. We still include three covariates X_1, X_2 and X_3 in model (3.1), and use the same distributions to generate the covariates. We set the true values $\boldsymbol{\beta}_0 = \frac{1}{\sqrt{3}}(1, 1, 1)$. Then the outcome Y is generated from a Poisson distribution with mean $e^{h(\mathbf{X}^T\boldsymbol{\beta}_0)}$. In the simulations, we still consider two different scenarios for $h(\cdot)$: (i) $h(t) = t$; and (ii) $h(t) = cos(t)$. We also adopt three data-adaptive interior knots that were placed at the equally spaced sample quantiles of $\mathbf{X}_i^T\boldsymbol{\beta}, i = 1, \cdots, n$ for a given $\boldsymbol{\beta}$, and the cubic ($M = 4$) B-spline to approximate the unknown function. We used model (3.1) to analyze the generated data and compared the results with that of the GLM based on the Poisson distribution with an intercept and the log link function. For each case, we considered the sample sizes $n = 300$, 500 and 1000 with 10000 replicates.

Tables 3.3 and 3.4 present the summary statistics of the sieve MLEs of the unknown regression parameters under model (3.1) and the MLEs of the regression parameters under the GLM. We observe similar results to the case of normal outcome data.

3.5 Discussion

We have proposed the sieve MLEs of the unknown parameters for the GLMs with an unknown link function using B-spline smoothing, and have established the asymptotic properties of our proposed estimators. The sieve MLEs of the regression coefficients and the unknown link function are consistent. Furthermore, the sieve MLEs of the regression parameters are asymptotically normally distributed and their covariance attains the semiparametric efficiency bound. The simulation studies demonstrate that the proposed sieve MLEs outperform the MLEs based on the ordinary GLMs with mis-specified link functions while the loss of efficiency is limited under correctly specified link functions.

We propose the sieve estimation approach in the framework of the GLMs. However, the proposed likelihood-based methods are not restricted to the GLMs and can be easily extended to other regression models in which the distribution of the outcome may not belong to the exponential family.

In Chapter 4, we extend the GLMs with an unknown link function to the longitudinal data by incorporating random effects.

Table 3.1: Comparison between the Sieve MLEs and the MLEs from the GLM with $h(t) = t$ for normal data

n		True	New					GLM			RE
			Bias 10^{-3}	SEE 10^{-2}	SE 10^{-2}	CP %	MSE 10^{-3}	Bias 10^{-3}	SE 10^{-2}	MSE 10^{-3}	
300	β_1	0.424	-0.62	4.76	4.81	94.3	2.32	-1.52	4.55	2.08	0.90
	β_2	0.566	-0.65	3.17	3.19	94.5	1.02	-0.82	3.08	0.95	0.93
	β_3	0.707	-2.28	3.35	3.39	94.5	1.15	-1.32	3.25	1.06	0.92
500	β_1	0.424	0.09	3.65	3.66	94.7	1.34	-1.16	3.55	1.26	0.94
	β_2	0.566	-0.39	2.43	2.40	95.1	0.58	-0.40	2.35	0.55	0.96
	β_3	0.707	-1.56	2.56	2.56	95.0	0.66	-0.70	2.50	0.62	0.95
1000	β_1	0.424	0.63	2.56	2.56	94.5	0.66	-0.56	2.51	0.63	0.96
	β_2	0.566	-0.16	1.70	1.70	95.1	0.29	-0.18	1.67	0.28	0.97
	β_3	0.707	-1.15	1.80	1.80	94.9	0.32	-0.38	1.76	0.31	0.96

True is the true value for the parameter; Bias is the bias of the parameter estimate; SEE is the average of the standard error estimates; SE is the empirical standard deviation of the parameter estimates; MSE is the mean squared error; CP is the coverage probability of the 95% confidence interval estimates; and RE is the relative efficiency of the proposed sieve MLEs with the estimators under the GLM with identity link function.

Table 3.2: Comparison between the Sieve MLEs and the MLEs from the GLM with $h(t) = \cos(t)$ for normal data

n		True	New					GLM			RE
			Bias 10^{-3}	SEE 10^{-2}	SE 10^{-2}	CP %	MSE 10^{-3}	Bias 10^{-3}	SE 10^{-2}	MSE 10^{-3}	
300	β_1	0.424	-1.33	5.86	6.08	93.3	3.70	-18.75	6.34	4.37	1.18
	β_2	0.566	0.56	3.74	3.85	94.1	1.48	10.41	4.22	1.89	1.27
	β_3	0.707	-4.55	4.01	4.15	93.8	1.75	-2.88	4.40	1.95	1.11
500	β_1	0.424	0.20	4.52	4.61	94.0	2.13	-17.76	4.94	2.75	1.29
	β_2	0.566	0.18	2.87	2.88	94.7	0.83	10.35	3.21	1.14	1.37
	β_3	0.707	-3.05	3.07	3.12	94.5	0.98	-1.18	3.38	1.15	1.16
1000	β_1	0.424	0.70	3.18	3.20	94.7	1.02	-17.37	3.48	1.51	1.48
	β_2	0.566	0.12	2.01	2.01	94.9	0.41	10.46	2.29	0.63	1.56
	β_3	0.707	-1.86	2.16	2.17	94.8	0.47	0.14	2.39	0.57	1.20

True is the true value for the parameter; Bias is the bias of the parameter estimate; SEE is the average of the standard error estimates; SE is the empirical standard deviation of the parameter estimates; MSE is the mean squared error; CP is the coverage probability of the 95% confidence interval estimates; and RE is the relative efficiency of the proposed sieve MLEs with the estimators under the GLM with identity link function.

Table 3.3: Comparison between the Sieve MLEs and the MLEs from the GLM with $h(t) = t$ for poisson data

n		True	New					GLM			RE
			Bias 10^{-3}	SEE 10^{-2}	SE 10^{-2}	CP %	MSE 10^{-3}	Bias 10^{-3}	SE 10^{-2}	MSE 10^{-3}	
300	β_1	0.577	-2.37	4.64	4.61	94.7	2.14	-2.03	4.40	1.94	0.91
	β_2	0.577	-0.61	3.66	3.67	94.3	1.35	-0.85	3.51	1.23	0.92
	β_3	0.577	-1.80	4.63	4.51	95.2	2.03	-1.55	4.41	1.95	0.96
500	β_1	0.577	-1.88	3.52	3.52	94.9	1.24	-1.45	3.40	1.16	0.93
	β_2	0.577	0.17	2.80	2.80	94.6	0.78	-0.46	2.71	0.73	0.93
	β_3	0.577	-1.09	3.53	3.46	95.5	1.20	-0.74	3.43	1.17	0.98
1000	β_1	0.577	-0.94	2.45	2.47	94.9	0.61	-0.59	2.38	0.57	0.93
	β_2	0.577	0.06	1.96	1.95	94.9	0.38	-0.54	1.90	0.36	0.95
	β_3	0.577	-0.48	2.47	2.43	95.0	0.59	-0.19	2.42	0.59	1.00

True is the true value for the parameter; Bias is the bias of the parameter estimate; SEE is the average of the standard error estimates; SE is the empirical standard deviation of the parameter estimates; MSE is the mean squared error; CP is the coverage probability of the 95% confidence interval estimates; and RE is the relative efficiency of the proposed sieve MLEs with the estimators under the GLM with log link function.

Table 3.4: Comparison between the Sieve MLEs and the MLEs from the GLM with $h(t) = \cos(t)$ for poisson data

			New					GLM			
n		True	Bias 10^{-3}	SEE 10^{-2}	SE 10^{-2}	CP %	MSE 10^{-3}	Bias 10^{-3}	SE 10^{-2}	MSE 10^{-3}	RE
300	β_1	0.577	-9.24	9.25	10.99	88.6	12.17	-42.01	11.16	14.21	1.17
	β_2	0.577	-2.40	6.84	7.78	90.1	6.06	28.50	8.38	7.83	1.29
	β_3	0.577	-14.00	9.10	10.58	89.6	11.38	-16.16	10.94	12.22	1.07
500	β_1	0.577	-5.27	7.39	8.13	91.0	6.63	-38.92	8.59	8.89	1.34
	β_2	0.577	-1.38	5.41	5.80	92.8	3.37	31.42	6.52	5.24	1.56
	β_3	0.577	-7.29	7.22	7.78	92.3	6.11	-10.99	8.43	7.23	1.18
1000	β_1	0.577	-2.76	5.32	5.54	93.6	3.07	-36.25	6.10	5.04	1.64
	β_2	0.577	-0.41	3.87	3.99	94.0	1.59	33.40	4.64	3.27	2.06
	β_3	0.577	-3.31	5.17	5.31	94.0	2.83	-7.49	5.98	3.63	1.28

True is the true value for the parameter; Bias is the bias of the parameter estimate; SEE is the average of the standard error estimates; SE is the empirical standard deviation of the parameter estimates; MSE is the mean squared error; CP is the coverage probability of the 95% confidence interval estimates; and RE is the relative efficiency of the proposed sieve MLEs with the estimators under the GLM with log link function.

Chapter 4: Sieve Maximum Likelihood Estimation Using B-spline Smoothing in Generalized Linear Mixed Models with an Unknown Link Function for the Longitudinal Data

4.1 Introduction

For longitudinal data with repeated measurements, the misspecification problem of the link function also attracts attentions in literature. Chiou and Müller (2005) explored the methodology dealing with GLMs with an unknown link function and unknown variance-covariance functions. They used the local polynomial fitting to estimate the link function and variance-covariance functions. Instead of using GEE, they proposed "estimated estimating equations" (EEE) to obtain regression parameter estimators. Large sample properties were derived for the coefficient estimators. Bai et al. (2009) also developed estimating equation-based procedures to estimate the unknown link function under the GLMs using penalized P-spline smoothing. Xu and Zhu (2012) proposed profile-type estimating functions for the coefficients by applying the Kernel smoothing method to estimate the unknown link function. The penalized spline method shows no boundary effects, while Kernel smoothing tends to have boundary effects. Various work has been done for other models with unknown link functions as well. Jiang and Wang (2011) employed Kernel smoothing to investigate the functional single index models for longitudinal data. Kernel smoothing method for the linear mixed effect models (partially linear single-index models) with unknown link functions are also studied by Pang and Xue (2012). Essentially no work on the GLMMs with an unknown link function is available in literature.

In this chapter, we consider the GLMMs with an unknown function. We propose a sieve maximum likelihood estimation procedure to estimate the unknown parameters including the regression coefficients, variance-covariance matrix of the random effects, and the unknown link function by using the B-spline smoothing. We derive the convergence rate of the sieve MLEs and establish the asymptotic normality and semiparametric efficiency of the sieve MLEs of the regression coefficients and the variance-covariance matrix of the random effects. Extensive simulation studies and an application to an epileptic study (Thall and Vail, 1990) are provided.

4.2 Methods

Suppose there are n subjects in the study with k_i observations for the ith subject. Let Y_{ij} denote the jth outcome of the ith subject, $\mathbf{X}_{ij} = (X_{ij1}, \cdots, X_{ijp_x})^T$ denote the corresponding p_x-dimensional vector of covariates and $\mathbf{Z}_{ij} = (Z_{ij1}, \cdots, Z_{ijp_z})^T$ denote the corresponding p_z-dimensional vector of covariates for the jth observation of the ith subject $j = 1, \cdots, k_i$, $i = 1, \cdots, n$, and Y_{ij} and $\mathbf{X}_{ij}$ be sampled from Y and $\mathbf{X}$. We assume that the outcome variable Y has a probability distribution belonging to the exponential family. Specifically, the density of Y_{ij} in the general form of the exponential family is given by

$$f(y_{ij}) = \exp\left\{\frac{y_{ij}\phi_{ij} - b(\phi_{ij})}{a(\varphi)} - c(y_{ij}, \varphi)\right\}, j = 1, \cdots, k_i, i = 1, \cdots, n.$$

Consider the following generalized linear mixed model

$$g\left(E(Y_{ij}|\mathbf{X}_{ij}, \mathbf{Z}_{ij}, \mathbf{b}_i)\right) = h(\mathbf{X}_{ij}^T\boldsymbol{\beta} + \mathbf{Z}_{ij}^T\mathbf{b}_i), j = 1, \cdots, k_i, i = 1, \cdots, n, \tag{4.1}$$

where $\boldsymbol{\beta}$ is a $p_x \times 1$ vector of unknown regression parameters, $g(\cdot)$ is a known transformation function and $h(\cdot) : \mathcal{R} \to \mathcal{R}$ is an unknown function which needs to be estimated. Moreover, $\mathbf{b}_i$'s are p_z-dimensional subject-specific random vectors which are independently and identically distributed with a joint distribution ψ and $\mathbf{b}_i$ can include a random intercept. The transformation function $g(\cdot)$ can be taken as the canonical link function. For example,

$g(x) = x$ for normal data, $g(x) = \log(x)$ for count data, and $g(x) = \text{logit}(x)$ for binary data. Notice that the actual link function is $h^{-1} \circ g(\cdot)$. To ensure the identifiability of the model, we do not include intercept in the term $\mathbf{X}_{ij}^T \boldsymbol{\beta}$, and impose the constrint on $\boldsymbol{\beta}$ such that $||\boldsymbol{\beta}|| = 1$ and $\beta_{p_x} > 0$, where $|| \cdot ||$ is the Euclidean norm. In this chapter, we assume $\mathbf{b}_i$ follows a multivariate normal distribution with mean $\mathbf{0}$ and variance-covariance matrix $\boldsymbol{\Sigma}$. We also assume $Y_{i1}, \cdots, Y_{ik_i}$ are mutually independent given $\mathbf{b}_i$. Let $\mu_{ij} = E(Y_{ij}|\mathbf{X}_{ij}, \mathbf{Z}_{ij}, \mathbf{b}_i)$. Then for the canonical link function, we have $g(\mu_{ij}) = \phi_{ij}$.

The conditional likelihood given $\mathbf{b}_i$ for the ith subject is given by

$$L_{ci}(\boldsymbol{\beta}, \boldsymbol{\gamma}|\mathbf{x}_i, \mathbf{z}_i, \mathbf{y}_i, \mathbf{b}_i) = C \prod_{j=1}^{k_i} \exp\left[\frac{1}{a(\varphi)}\left\{y_{ij} h(\mathbf{x}_{ij}^T\boldsymbol{\beta} + \mathbf{z}_{ij}^T\mathbf{b}_i) - b\left(h(\mathbf{x}_{ij}^T\boldsymbol{\beta} + \mathbf{z}_{ij}^T\mathbf{b}_i)\right)\right\}\right],$$

where $\mathbf{x}_i = (\mathbf{x}_{i1}^T, \mathbf{x}_{i2}^T, \cdots, \mathbf{x}_{ik_i}^T)^T$, $\mathbf{z}_i = (\mathbf{z}_{i1}^T, \mathbf{z}_{i2}^T, \cdots, \mathbf{z}_{ik_i}^T)^T$, $\mathbf{y}_i = (y_{i1}, y_{i2}, \cdots, y_{ik_i})^T$ and given the data, C is a constant . It follows that the log-likelihood given the observed data takes the form

$$l_n(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\Sigma}|\mathbf{x}, \mathbf{z}, \mathbf{y}) = \sum_{i=1}^{n} \log \int L_{ci}(\boldsymbol{\beta}, \boldsymbol{\gamma}|\mathbf{x}_i, \mathbf{z}_i, \mathbf{y}_i, \mathbf{b}_i)\psi(\mathbf{b}_i)d\mathbf{b}_i,$$

where $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n)^T$, $\mathbf{z}_i = (\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_n)^T$, $\mathbf{y} = (y_1, y_2, \cdots, y_n)^T$, and $\psi(\cdot)$ is the multivariate normal density function with mean $\mathbf{0}$ and variance-covariance matrix $\boldsymbol{\Sigma}$.

Consider a bounded interval (a, b). We also consider K_n interior knots $K_1, K_2, \cdots, K_{K_n}$ with $K_n = O(n^{\nu})$ and $\max_{0 \le j \le K_n} |K_{j+1} - K_j| = O(n^{-\nu})$, for $\nu \in (0, 0.5)$, and $a = K_0 \le K_1 \le K_2 \le \cdots \le K_{K_n} \le K_{K_n+1} = b$. Let $\mathcal{S}_n(\mathbf{K}, K_n, M)$ denote the space of polynomial splines of order M defined in Schumaker (1981), where $\mathbf{K} = (K_0, K_1, \cdots, K_{K_n+1})$. Then there exists a local basis $\{B_j : 1 \le j \le K_n + M\}$ such that for any $s \in \mathcal{S}_n(\mathbf{K}, K_n, M)$, we have

$$s(t) = \sum_{j=1}^{K_n+M} \gamma_j B_{M,j}(t) = \boldsymbol{\gamma}^T \mathbf{B}_M(t),$$

where $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \cdots, \gamma_{K_n+M})^T$ are the smoothing coefficients and $B_M(t) = \big\{B_{M,1}(t), B_{M,2}(t), \cdots, B_{M,K+M}(t)\big)^T$ are the M-degree B-spline basis functions. Under some smoothness assumptions, the unknown univariate function $h(\cdot)$ can be well approximated by an M-degree B-spline in $\mathcal{S}_n(\mathbf{K}, K_n, M)$, that is,

$$h(t) \approx \sum_{j=1}^{K_n+M} \gamma_j B_{M,j}(t) = \boldsymbol{\gamma}^T \mathbf{B}_M(t).$$

Replacing $h(t)$ by $\boldsymbol{\gamma}^T \mathbf{B}_M(t)$ in the conditional likelihood function, we obtain

$$\begin{aligned} & L_{ci}(\boldsymbol{\beta}, \boldsymbol{\gamma} | \mathbf{x}_i, \mathbf{z}_i, \mathbf{y}_i, \mathbf{b}_i) \\ = \; & C \prod_{j=1}^{k_i} \exp\left[\frac{1}{a(\varphi)}\left\{y_{ij}\boldsymbol{\gamma}^T \mathbf{B}_M(\mathbf{x}_{ij}^T\boldsymbol{\beta} + \mathbf{z}_{ij}^T\mathbf{b}_i) - b\left(\boldsymbol{\gamma}^T \mathbf{B}_M(\mathbf{x}_{ij}^T\boldsymbol{\beta} + +\mathbf{z}_{ij}^T\mathbf{b}_i)\right)\right\}\right]. \end{aligned}$$

In order to estimate the unknown parameters, we need to maximize the observed-data log-likelihood. Notice that there is no closed form for the observed-data log-likelihood in general except for normal responses. We suggest to use the Gauss-Hermite quadrature to numerically approximate the log-likelihood function. Let $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \cdots, \sigma_r)$ be a $r \times 1$ vector of parameters that contain all the unknown parameters involved in the variance-covariance matrix $\boldsymbol{\Sigma}$. For simplicity, let $l_n(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\sigma}) = l_n(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\Sigma} | \mathbf{x}, \mathbf{z}, \mathbf{y})$. The score equations are then given by

$$\frac{\partial l_n(\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\sigma})}{\partial \boldsymbol{\beta}} = \mathbf{0},$$

$$\frac{\partial l_n(\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\sigma})}{\partial \boldsymbol{\gamma}} = \mathbf{0},$$

$$\frac{\partial l_n(\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\sigma})}{\partial \boldsymbol{\sigma}} = \mathbf{0}.$$

We use an iterative procedure to estimate the unknown parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\Sigma}, h)$. The resulting sieve MLEs are denoted by $\hat{\boldsymbol{\theta}}_n = (\hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{\Sigma}}_n, \hat{h}_n)$, where $\hat{\boldsymbol{\Sigma}}$ is determined by $\hat{\boldsymbol{\sigma}}_n$ and $\hat{h}_n(t) = \hat{\boldsymbol{\gamma}}_n \mathbf{B}_M(t)$. Notice that the likelihood involves the value of the basis functions at $\mathbf{x}_{ij}^T\boldsymbol{\beta} + \mathbf{z}_{ij}^T\mathbf{b}_i$, $j = 1, \cdots, k_i, i = 1, \cdots, n$. We recommend to choose the knots based on the equally spaced sample quantiles of $\mathbf{x}_{ij}^T\boldsymbol{\beta} + \mathbf{z}_{ij}^T\mathbf{b}_k^*$, $j = 1, \cdots, k_i, i = 1, \cdots, n, k = 1, \cdots, N_q^{p_z}$, N_q is the number of quadrature points and $\mathbf{b}_k^*$ are the kth abscissas of the Gaussian-Hermite quadrature. As a result, the knots and basis change as $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ change. We describe the detailed algorithm in Section 4.4.

To make statistical inferences of $\boldsymbol{\beta}_0$ and $\boldsymbol{\Sigma}_0$, the distribution of $(\hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{\sigma}}_n)$ can be approximated by a multivariate normal distribution when the sample size is large enough. The variance-covariance matrix of $(\hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{\sigma}}_n)$ can be estimated by the inverse of the observed information matrix under the efficient score function of $(\boldsymbol{\beta}_0, \boldsymbol{\sigma}_0)$. We show that the asymptotic variance-covariance matrix of $(\hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{\sigma}}_n)$ attains the semiparametric efficiency bound. Ding and Nan (2011) suggested to use the observed information matrix by taking into account the parameter $\boldsymbol{\gamma}$. In our simulation, we used the inverse matrix of the estimated variance-covariance matrix of the score functions by taking into account all the unknown parameter $(\boldsymbol{\beta}, \boldsymbol{\sigma}, \boldsymbol{\gamma})$. Extensive simulations suggest that the variance-covariance estimator works well.

4.3 Asymptotic Properties

Let $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0, h_0)$ denote the true parameters. Before we establish the asymptotic properties of the sieve MLEs, we impose the following regularity conditions.

C1. The true values $(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0)$ for parameters $(\boldsymbol{\beta}, \boldsymbol{\Sigma})$ belong to the interior of a compact set

$$\mathcal{B}_0 = \{ (\boldsymbol{\beta}, \boldsymbol{\Sigma}) : \boldsymbol{\beta} \in \mathcal{R}^p, \beta_{p_x} > 0, ||\boldsymbol{\beta}|| = 1, \text{ and}$$

$$\boldsymbol{\Sigma} \text{ is positive definite and its eigenvalues are bounded away from 0 and } \infty \} .$$

C2. The domain of $\mathbf{X}$ is a bounded subset of $\mathcal{R}^{p_x}$. Therefore, $\mathbf{Z}$ is also bounded. Furthermore, $E(\mathbf{X}_{j_1}\mathbf{X}_{j_2}^T)$ are nonsingular for $j_1, j_2 = 1, \cdots, k$.

C3. The number of observations k for each subject is random. Moreover, there exists a positive integer k_0 such that $1 \leq k \leq k_0$ and $Pr(k \geq 2) > 0$.

C4. The true function $h_0(\cdot)$ belongs to $\mathcal{H}^q$, where the functional space $\mathcal{H}^q$ is the collection of all bounded functions h on a bounded interval $[a, b]$ with bounded jth derivative $h^{(j)}$, $j = 1, \cdots, k$, such that $h^{(k)}$ satisfies the Lipschitz continuity condition with exponent m $(0 < m \leq 1)$:

$$|h^{(k)}(s) - h^{(k)}(t)| \leq L|s - t|^m, \text{ for } s, t \in [a, b],$$

where $L < \infty$ is a positive constant, and $q = k + m \geq 3$.

C5. For some $\eta \in (0, 1)$, $u^T Var(\mathbf{X}|\mathbf{X}^T\boldsymbol{\beta}_0 + \mathbf{Z}^T\mathbf{b})u \geq \eta u^T E(\mathbf{X}\mathbf{X}^T|\mathbf{X}^T\boldsymbol{\beta}_0 + \mathbf{Z}^T\mathbf{b})u$ and $u^T Var(\mathbf{Z}|\mathbf{X}^T\boldsymbol{\beta}_0 + \mathbf{Z}^T\mathbf{b})u \geq \eta u^T E(\mathbf{Z}\mathbf{Z}^T|\mathbf{X}^T\boldsymbol{\beta}_0 + \mathbf{Z}^T\mathbf{b})u$ almost surly for all $u \in \mathcal{R}^p$.

C6. $E\left[\left\{h_0'(\mathbf{X}^T\boldsymbol{\beta}_0 + \mathbf{Z}^T\mathbf{b})\right\}^2 \mathbf{X}\mathbf{X}^T\right]$ and $E\left[\left\{h_0'(\mathbf{X}^T\boldsymbol{\beta}_0 + \mathbf{Z}^T\mathbf{b})\right\}^2 \mathbf{Z}\mathbf{Z}^T\right]$ are nonsingular.

Remark 4.1. Condition C1 is a common regularity assumption in literature. Conditions C1 and C2 ensure the identifiability of the model. The restriction $q \geq 3$ in the condition C4 is needed to provide desirable control for the B-spline approximation error rates of h_0 as well as the first and second derivatives of h_0 for the proof of the normality. Conditions C4 and C5 are technical assumptions and can be justified in many applications. Condition C3 implies that the number of observations for each subject is bounded and some of them have at least two observations. Note that although the domain of $\mathbf{X}^T\boldsymbol{\beta}+\mathbf{Z}^T\mathbf{b}$ is $(-\infty, \infty)$, we focus on the estimation and inference of h on a bounded interval (a, b). In addition, Condition C6 ensures that the semiparametric efficiency information matrix is invertible.

Let $\mathcal{H}_n^q = \mathcal{H}^q \cap \mathcal{S}_n$. Clearly, we have $\mathcal{H}_n^q \subseteq \mathcal{H}_{n+1}^q \subseteq \cdots \subseteq \mathcal{H}^q$ for all $n \geq 1$. Denote $\Theta_n^q = \mathcal{B} \times \mathcal{H}_n^q$ as the sieve space. The sieve MLEs are the maximizer of the log-likelihood function over the sieve space Θ_n^q. Define the norm $||\cdot||_2$ over the space $\mathcal{H}^q$ as

$$||h||_2 = \left[E\{h^2(\mathbf{X}^T\boldsymbol{\beta}_0)\}\right]^{\frac{1}{2}}.$$

Let $||\cdot||_\infty$ denote the supremum norm. We define the distance over the space $\Theta^q = \mathcal{B} \times \mathcal{H}^p$ as follows

$$d(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \left(||\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2||^2 + ||\boldsymbol{\sigma}_1 - \boldsymbol{\sigma}_2||^2 + ||h_1 - h_2||_2^2\right)^{\frac{1}{2}},$$

for $\boldsymbol{\theta}_1 = (\boldsymbol{\beta}_1, \boldsymbol{\sigma}_1, h_1), \boldsymbol{\theta}_2 = (\boldsymbol{\beta}_2, \boldsymbol{\sigma}_2, h_2) \in \Theta^q$.

Next, we introduce some notation in the context of empirical process. Let $\mathbb{P}_n$ and $\mathbb{P}$ be the empirical measure and the population distribution of n i.i.d. observations $\mathbf{O}_1, \mathbf{O}_2, \cdots, \mathbf{O}_n$ for $\mathbf{O}$=($\mathbf{Y}$, $\mathbf{X}$, $\mathbf{Z}$), $i = 1, \cdots, n$. Let $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - \mathbb{P})$ denote the empirical process. Then for any measurable function $h(\mathbf{O})$,

$$\mathbb{P}_n h(\mathbf{O}) = \frac{1}{n}\sum_{i=1}^n h(\mathbf{O}_i), \mathbb{P}h(\mathbf{O}) = E_{\mathbb{P}}[h(\mathbf{O})],$$

and

$$\mathbb{G}_n h(\mathbf{O}) = \sqrt{n}(\mathbb{P}_n h(\mathbf{O}) - \mathbb{P} h(\mathbf{O})).$$

Let $K_n = O(n^{\nu})$, and $\frac{1}{2(1+q)} < \nu < \frac{1}{2q}$, where q is the smoothness parameter in C4. Suppose $\boldsymbol{\sigma}$ contains the least number of parameters that can determine $\boldsymbol{\Sigma}$, and $\boldsymbol{\alpha}$=($\boldsymbol{\beta}$, $\boldsymbol{\sigma}$). The following theorem establishes the convergence rate of the sieve MLEs $\hat{\boldsymbol{\theta}}_n = (\hat{\boldsymbol{\alpha}}_n, \hat{h}_n)$ to the true parameter $\boldsymbol{\theta}_0 = (\boldsymbol{\alpha}_0, h_0)$.

Theorem 4.1. Under the regularity conditions C1-C5,

$$d(\hat{\boldsymbol{\theta}}_n, \theta_0) = O_p(n^{-\min(q\nu, \frac{1-\nu}{2})}).$$

Remark 4.2. The proof is similar to the proof of Theorem 3.1 in Chapter 3. The main difference is that we have an integral in the log-likelihood. In fact, the integral of a bounded function against a probability measure is still bounded. Due to this fact, the properties of the log-likelihood and its derivatives remain the same. In addition, the condition C1 indicates that the eigenvalues of $\boldsymbol{\Sigma}_0$ are also bounded. Then by verifying the conditions in Shen and Wong (1994), we obtain the convergence rate of $\hat{\boldsymbol{\theta}}_n$..

Proof. Suppose k is the number of observations for a single subject, which is random. For a single subject, let $\mathbf{Y} = (Y_1, Y_2, \cdots, Y_k)$ be the outcome variable. Correspondingly, the covariate vector for the jth observation is denoted by $\mathbf{X}_j$ and $\mathbf{Z}_j$. Then $\mathbf{X} = (\mathbf{X}_1^T, \mathbf{X}_2^T, \cdots, \mathbf{X}_k^T)$, and $\mathbf{Z} = (\mathbf{Z}_1^T, \mathbf{Z}_2^T, \cdots, \mathbf{Z}_k^T)$. Suppose $\mathbf{O} = (\mathbf{Y}, \mathbf{X}, \mathbf{Z}, k)$, and $\mathbf{O}_i, i = 1, \cdots, n$ are n i.i.d. observations. For simplicity, let $T_j(\mathbf{b}) = \mathbf{X}_j^T \boldsymbol{\beta} + \mathbf{Z}_j^T \mathbf{b}$, $T_{0j}(\mathbf{b}) = \mathbf{X}_j^T \boldsymbol{\beta}_0 + \mathbf{Z}_j^T \mathbf{b}$. Then the likelihood for a single subject given $\mathbf{O}$ takes the form

$$l(\boldsymbol{\beta}, \boldsymbol{\Sigma}, h; \mathbf{O}) = -\frac{1}{2}\log|\boldsymbol{\Sigma}| + \log\int \exp\left\{-\frac{1}{2}\mathbf{b}^T\boldsymbol{\Sigma}^{-1}\mathbf{b} + \sum_{j=1}^{k}\left\{Y_j h(T_j(\mathbf{b})) - b\left(h(T_j(\mathbf{b}))\right\}\right\} d\mathbf{b}.$$

Before further proof, we recall that for independent data ($k = 1$), the likelihood and its derivatives for a single observation are given by

$$
\begin{aligned}
l(\boldsymbol{\beta}, h; \mathbf{O}) &= \left\{ Y h(\mathbf{X}^T\boldsymbol{\beta}) - b\left(h(\mathbf{X}^T\boldsymbol{\beta})\right)\right\}, \\
\dot{l}_{\boldsymbol{\beta}}(\boldsymbol{\beta}, h; \mathbf{O}) &= \frac{\partial l(\boldsymbol{\beta}, h; \mathbf{O})}{\partial \boldsymbol{\beta}} \\
&= \left\{ Y - b'\left(h(\mathbf{X}^T\boldsymbol{\beta})\right)\right\} h'(\mathbf{X}^T\boldsymbol{\beta})\mathbf{X}, \\
\dot{l}_{h}(\boldsymbol{\beta}, h; \mathbf{O})[h_1] &= \frac{\partial l(\boldsymbol{\beta}, h + \epsilon h_1; \mathbf{O})}{\partial \epsilon}\Big|_{\epsilon=0} \\
&= \left\{ Y - b'\left(h(\mathbf{X}^T\boldsymbol{\beta})\right)\right\} h_1(\mathbf{X}^T\boldsymbol{\beta}), \\
\ddot{l}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\beta}, h; \mathbf{O}) &= \frac{\partial^2 l(\boldsymbol{\beta}, h; \mathbf{O})}{\partial \boldsymbol{\beta}\boldsymbol{\beta}^T} \\
&= \Big\{ Y h''(\mathbf{X}^T\boldsymbol{\beta}) - b'\left(h(\mathbf{X}^T\boldsymbol{\beta})\right) h''(\mathbf{X}^T\boldsymbol{\beta}) \\
&\quad - b''\left(h(\mathbf{X}^T\boldsymbol{\beta})\right)\left(h'(\mathbf{X}^T\boldsymbol{\beta})\right)^2 \Big\}\mathbf{X}\mathbf{X}^T, \\
\ddot{l}_{\boldsymbol{\beta}h}(\boldsymbol{\beta}, h; \mathbf{O})[h_1] &= \frac{\partial^2 l(\boldsymbol{\beta}, h; \mathbf{O})}{\partial \boldsymbol{\beta}\partial h} \\
&= \Big\{ Y h_1'(\mathbf{X}^T\boldsymbol{\beta}) - b'\left(h(\mathbf{X}^T\boldsymbol{\beta})\right) h_1'(\mathbf{X}^T\boldsymbol{\beta}) \\
&\quad - b''\left(h(\mathbf{X}^T\boldsymbol{\beta})\right) h_1(\mathbf{X}^T\boldsymbol{\beta}) h'(\mathbf{X}^T\boldsymbol{\beta}) \Big\}\mathbf{X}, \\
\ddot{l}_{hh}(\boldsymbol{\beta}, h; \mathbf{O})[h_1, h_2] &= \frac{\partial^2 l(\boldsymbol{\beta}, h; \mathbf{O})}{\partial h \partial h} \\
&= -b''\left(h(\mathbf{X}^T\boldsymbol{\beta})\right) h_1(\mathbf{X}^T\boldsymbol{\beta}) h_2(\mathbf{X}^T\boldsymbol{\beta}).
\end{aligned}
$$

Then for the jth observation, given $\mathbf{b}$, we define

$$
\begin{aligned}
l^j(\boldsymbol{\beta}, h; \mathbf{O}, \mathbf{b}) &= Y_j h(T_j(\mathbf{b})) - b\left(h(T_j(\mathbf{b}))\right), \\
\dot{l}^j_{\boldsymbol{\beta}}(\boldsymbol{\beta}, h; \mathbf{O}, \mathbf{b}) &= \left\{ Y_j - b'\left(h(T_j(\mathbf{b}))\right)\right\} h'(T_j(\mathbf{b}))\mathbf{X}_j,
\end{aligned}
$$

$$
\begin{aligned}
\dot{l}^j_h(\boldsymbol{\beta}, h; \mathbf{O}, \mathbf{b})[h_1] &= \left\{Y_j - b'\left(h(T_j(\mathbf{b}))\right)\right\} h_1(T_j(\mathbf{b})), \\
\ddot{l}^j_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\beta}, h; \mathbf{O}, \mathbf{b}) &= \Big\{Y_j h''(T_j(\mathbf{b})) - b'\left(h(T_j(\mathbf{b}))\right) h''(T_j(\mathbf{b})) \\
&\quad -b''\left(h(T_j(\mathbf{b}))\right)\left(h'(T_j(\mathbf{b}))\right)^2 \Big\} \mathbf{X}_j \mathbf{X}_j^T, \\
\ddot{l}^j_{\boldsymbol{\beta}h}(\boldsymbol{\beta}, h; \mathbf{O}, \mathbf{b})[h_1] &= \Big\{Y_j h_1'(T_j(\mathbf{b})) - b'\left(h(T_j(\mathbf{b}))\right) h_1'(T_j(\mathbf{b})) \\
&\quad -b''\left(h(T_j(\mathbf{b}))\right) h_1(T_j(\mathbf{b})) h'(T_j(\mathbf{b}))\Big\} \mathbf{X}_j, \\
\ddot{l}^j_{hh}(\boldsymbol{\beta}, h; \mathbf{O}, \mathbf{b})[h_1, h_2] &= -b''\left(h(T_j(\mathbf{b}))\right) h_1(T_j(\mathbf{b})) h_2(T_j(\mathbf{b})).
\end{aligned}
$$

Let

$$
\begin{aligned}
A(\mathbf{b}) &= -\frac{1}{2}\mathbf{b}^T \boldsymbol{\Sigma}^{-1}\mathbf{b} + \sum_{j=1}^k \left\{Y_j h(T_j(\mathbf{b})) - b\left(h(T_j(\mathbf{b})\right)\right\} \\
&= -\frac{1}{2}\mathbf{b}^T \boldsymbol{\Sigma}^{-1}\mathbf{b} + \sum_{j=1}^k \left\{l^j(\boldsymbol{\beta}, h; \mathbf{O}, \mathbf{b}))\right\},
\end{aligned}
$$

and

$$
B = \int e^{A(\mathbf{b})} d\mathbf{b}.
$$

Then

$$
\dot{l}_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\Sigma}, h; \mathbf{O}) = \frac{\partial l(\boldsymbol{\beta}, \boldsymbol{\Sigma}, h; \mathbf{O})}{\partial \boldsymbol{\beta}}
$$

$$
= \frac{1}{B}\int e^{A(\mathbf{b})}\left[\sum_{j=1}^k \dot{l}^j_{\boldsymbol{\beta}}(\boldsymbol{\beta}, h; \mathbf{O}, \mathbf{b})\right] d\mathbf{b},
$$

$$
\dot{l}_{\boldsymbol{\Sigma}}(\boldsymbol{\beta}, \boldsymbol{\Sigma}, h; \mathbf{O}) = \frac{\partial l(\boldsymbol{\beta}, \boldsymbol{\Sigma}, h; \mathbf{O})}{\partial \boldsymbol{\Sigma}}
$$

$$= -\frac{1}{2}\mathbf{\Sigma}^{-1} + \frac{1}{2B}\int e^{A(\mathbf{b})}(\mathbf{\Sigma}^{-1}\mathbf{b}\mathbf{b}^T\mathbf{\Sigma}^{-1})d\mathbf{b},$$

$$\dot{l}_h(\boldsymbol{\beta}, \mathbf{\Sigma}, h; \mathbf{O})[h_1] = \frac{\partial l(\boldsymbol{\beta}, \mathbf{\Sigma}, h + \epsilon h_1; \mathbf{O})}{\partial \epsilon}\Big|_{\epsilon=0}$$

$$= \frac{1}{B}\int e^{A(\mathbf{b})}\left[\sum_{j=1}^{k}\dot{l}_h^j(\boldsymbol{\beta}, h; \mathbf{O}, \mathbf{b})[h_1]\right]d\mathbf{b},$$

$$\ddot{l}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\beta}, \mathbf{\Sigma}, h; \mathbf{O}) = \frac{\partial^2 l(\boldsymbol{\beta}, \mathbf{\Sigma}, h; \mathbf{O})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}$$

$$= \frac{1}{B}\int e^{A(\mathbf{b})}\left[\sum_{j=1}^{k}\ddot{l}_{\boldsymbol{\beta}\boldsymbol{\beta}}^j(\boldsymbol{\beta}, h; \mathbf{O}, \mathbf{b}) + \left\{\sum_{j=1}^{k}\dot{l}_{\boldsymbol{\beta}}^j(\boldsymbol{\beta}, h; \mathbf{O}, \mathbf{b})\right\}\left\{\sum_{j=1}^{k}\dot{l}_{\boldsymbol{\beta}}^{j\,T}(\boldsymbol{\beta}, h; \mathbf{O}, \mathbf{b})\right\}\right]d\mathbf{b},$$

$$-\frac{1}{B^2}\int e^{A(\mathbf{b})}\left\{\sum_{j=1}^{k}\dot{l}_{\boldsymbol{\beta}}^j(\boldsymbol{\beta}, h; \mathbf{O}, \mathbf{b})\right\}d\mathbf{b}\int e^{A(\mathbf{b})}\left[\sum_{j=1}^{k}\dot{l}_{\boldsymbol{\beta}}^{j\,T}(\boldsymbol{\beta}, h; \mathbf{O}, \mathbf{b})\right]d\mathbf{b},$$

$$\ddot{l}_{\boldsymbol{\beta}\mathbf{\Sigma}}(\boldsymbol{\beta}, \mathbf{\Sigma}, h; \mathbf{O}) = \frac{\partial^2 l(\boldsymbol{\beta}, \mathbf{\Sigma}, h; \mathbf{O})}{\partial\boldsymbol{\beta}\otimes\partial\mathbf{\Sigma}}$$

$$= \frac{1}{2B}\int e^{A(\mathbf{b})}\left[\sum_{j=1}^{k}\dot{l}_{\boldsymbol{\beta}}^j(\boldsymbol{\beta}, h; \mathbf{O}, \mathbf{b})\otimes\left\{\mathbf{\Sigma}^{-1}\mathbf{b}\mathbf{b}^T\mathbf{\Sigma}^{-1}\right\}\right]d\mathbf{b}$$

$$-\frac{1}{2B^2}\int e^{A(\mathbf{b})}\mathbf{1}\otimes\left\{\mathbf{\Sigma}^{-1}\mathbf{b}\mathbf{b}^T\mathbf{\Sigma}^{-1}\right\}d\mathbf{b},$$

$$\ddot{l}_{\mathbf{\Sigma}\boldsymbol{\beta}}(\boldsymbol{\beta}, \mathbf{\Sigma}, h; \mathbf{O}) = \frac{\partial^2 l(\boldsymbol{\beta}, \mathbf{\Sigma}, h; \mathbf{O})}{\partial\mathbf{\Sigma}\otimes\partial\boldsymbol{\beta}}$$

$$= \frac{1}{2B}\int e^{A(\mathbf{b})}\left[\left\{\mathbf{\Sigma}^{-1}\mathbf{b}\mathbf{b}^T\mathbf{\Sigma}^{-1}\right\}\otimes\sum_{j=1}^{k}\dot{l}_{\boldsymbol{\beta}}^j(\boldsymbol{\beta}, h; \mathbf{O}, \mathbf{b})\right]d\mathbf{b}$$

$$-\frac{1}{2B^2}\int e^{A(\mathbf{b})}\left\{\mathbf{\Sigma}^{-1}\mathbf{b}\mathbf{b}^T\mathbf{\Sigma}^{-1}\right\}\otimes\mathbf{1}d\mathbf{b},$$

$$\ddot{l}_{\mathbf{\Sigma}\mathbf{\Sigma}}(\boldsymbol{\beta}, \mathbf{\Sigma}, h; \mathbf{O}) = \frac{\partial^2 l(\boldsymbol{\beta}, \mathbf{\Sigma}, h; \mathbf{O})}{\partial\mathbf{\Sigma}\otimes\partial\mathbf{\Sigma}}$$

$$= -\frac{1}{2}\mathbf{I}_{p_z}\otimes\mathbf{D}$$

$$+\frac{1}{2B}\int e^{A(\mathbf{b})}\Big[2\boldsymbol{\Sigma}^{-1}\mathbf{b}\mathbf{b}^T\otimes\mathbf{D}+\frac{1}{2}\left\{\boldsymbol{\Sigma}^{-1}\mathbf{b}\mathbf{b}^T\boldsymbol{\Sigma}^{-1}\right\}\otimes\left\{\boldsymbol{\Sigma}^{-1}\mathbf{b}\mathbf{b}^T\boldsymbol{\Sigma}^{-1}\right\}\Big]d\mathbf{b}$$

$$-\frac{1}{4B^2}\int e^{A(\mathbf{b})}(\boldsymbol{\Sigma}^{-1}\mathbf{b}\mathbf{b}^T\boldsymbol{\Sigma}^{-1})d\mathbf{b}\otimes\int e^{A(\mathbf{b})}(\boldsymbol{\Sigma}^{-1}\mathbf{b}\mathbf{b}^T\boldsymbol{\Sigma}^{-1})d\mathbf{b},$$

$$\ddot{l}_{\boldsymbol{\beta}h}(\boldsymbol{\beta},\boldsymbol{\Sigma},h;\mathbf{O})[h_1]=\frac{\partial}{\partial\varepsilon}\dot{l}_{\boldsymbol{\beta}}(\boldsymbol{\beta},\boldsymbol{\Sigma},h+\varepsilon h_1;\mathbf{O})$$

$$=\frac{1}{B}\int e^{A(\mathbf{b})}\left[\sum_{j=1}^{k}\ddot{l}^j_{\boldsymbol{\beta}h}(\boldsymbol{\beta},h;\mathbf{O},\mathbf{b})[h_1]+\left\{\sum_{j=1}^{k}\dot{l}^j_{\boldsymbol{\beta}}(\boldsymbol{\beta},h;\mathbf{O},\mathbf{b})\right\}\left\{\sum_{j=1}^{k}\dot{l}^j_{h}(\boldsymbol{\beta},h;\mathbf{O},\mathbf{b})[h_1]\right\}\right]d\mathbf{b},$$

$$-\frac{1}{B^2}\int e^{A(\mathbf{b})}\left\{\sum_{j=1}^{k}\dot{l}^j_{\boldsymbol{\beta}}(\boldsymbol{\beta},h;\mathbf{O},\mathbf{b})\right\}d\mathbf{b}\int e^{A(\mathbf{b})}\left[\sum_{j=1}^{k}\dot{l}^j_{h}(\boldsymbol{\beta},h;\mathbf{O},\mathbf{b})[h_1]\right]d\mathbf{b},$$

$$\ddot{l}_{\boldsymbol{\Sigma}h}(\boldsymbol{\beta},\boldsymbol{\Sigma},h;\mathbf{O})[h_1]=\frac{\partial}{\partial\varepsilon}\dot{l}_{\boldsymbol{\Sigma}}(\boldsymbol{\beta},\boldsymbol{\Sigma},h+\varepsilon h_1;\mathbf{O})$$

$$=\frac{1}{2B}\int e^{A(\mathbf{b})}\boldsymbol{\Sigma}^{-1}\mathbf{b}\mathbf{b}^T\boldsymbol{\Sigma}^{-1}\left[\sum_{j=1}^{k}\dot{l}^j_{h}(\boldsymbol{\beta},h;\mathbf{O},\mathbf{b})[h_1]d\mathbf{b}\right.$$

$$\left.-\frac{1}{2B^2}\int e^{A(\mathbf{b})}\boldsymbol{\Sigma}^{-1}\mathbf{b}\mathbf{b}^T\boldsymbol{\Sigma}^{-1}d\mathbf{b}\int e^{A(\mathbf{b})}\left[\sum_{j=1}^{k}\dot{l}^j_{h}(\boldsymbol{\beta},h;\mathbf{O},\mathbf{b})[h_1]\right]d\mathbf{b},\right.$$

$$\ddot{l}_{hh}(\boldsymbol{\beta},\boldsymbol{\Sigma},h;\mathbf{O})[h_1,h_2]=\frac{\partial}{\partial\varepsilon}\dot{l}_{h}(\boldsymbol{\beta},\boldsymbol{\Sigma},h+\varepsilon h_2;\mathbf{O})[h_1]$$

$$=\frac{1}{B}\int e^{A(\mathbf{b})}\left[\sum_{j=1}^{k}\ddot{l}^j_{hh}(\boldsymbol{\beta},h;\mathbf{O},\mathbf{b})[h_1,h_2]\right.$$

$$+\left\{\sum_{j=1}^{k}\dot{l}^j_{h}(\boldsymbol{\beta},h;\mathbf{O},\mathbf{b})[h_1]\right\}\left\{\sum_{j=1}^{k}\dot{l}^j_{h}(\boldsymbol{\beta},h;\mathbf{O},\mathbf{b})[h_2]\right\}$$

$$-\frac{1}{B^2}\int e^{A(\mathbf{b})}\left\{\sum_{j=1}^{k}\dot{l}^j_{h}(\boldsymbol{\beta},h;\mathbf{O},\mathbf{b})[h_1]\right\}d\mathbf{b}\int e^{A(\mathbf{b})}\left\{\sum_{j=1}^{k}\dot{l}^j_{h}(\boldsymbol{\beta},h;\mathbf{O},\mathbf{b})[h_2]\right\}d\mathbf{b},$$

where $h_1,h_2\in\mathbf{H}=\{g:g=\frac{\partial h_\epsilon}{\partial\epsilon}\big|_{\epsilon=0},h_\epsilon\in\mathcal{H}^p\}$, $\otimes$ is the Kronecker product, $\mathbf{I}_{p_z}$ is the

p_z-dimensional identity matrix,

$$\mathbf{D} = \begin{pmatrix} \mathbf{D}_{11} & \mathbf{D}_{12} & \cdots & \mathbf{D}_{1p_z} \\ \mathbf{D}_{21} & \mathbf{D}_{22} & \cdots & \mathbf{D}_{2p_z} \\ \vdots & \vdots & \cdots & \vdots \\ \mathbf{D}_{p_z1} & \mathbf{D}_{p_z2} & \cdots & \mathbf{D}_{p_zp_z} \end{pmatrix}$$

with

$$\mathbf{D}_{ij} = \begin{pmatrix} a_{1i}a_{j1} & a_{1i}a_{j2} & \cdots & a_{1i}a_{jp_z} \\ a_{2i}a_{j1} & a_{2i}a_{j2} & \cdots & a_{2i}a_{jp_z} \\ \vdots & \vdots & \cdots & \vdots \\ a_{p_zi}a_{p_z1} & a_{p_zi}a_{p_z2} & \cdots & a_{p_zi}a_{p_zp_z}, \end{pmatrix}, \ i,j = 1, \cdots, p_z,$$

and a_{ij} is the (i,j)th element of $\mathbf{\Sigma}^{-1}$. It should be mentioned that instead of the ordinary Kronecker product, we use $\mathbf{R} \otimes \mathbf{D}$ denote

$$\begin{pmatrix} \mathbf{RD}_{11} & \mathbf{RD}_{12} & \cdots & \mathbf{RD}_{1p_z} \\ \mathbf{RD}_{21} & \mathbf{RD}_{22} & \cdots & \mathbf{RD}_{2p_z} \\ \vdots & \vdots & \cdots & \vdots \\ \mathbf{RD}_{p_z1} & \mathbf{RD}_{p_z2} & \cdots & \mathbf{RD}_{p_zp_z} \end{pmatrix},$$

for a $p_z \times p_z$ matrix $\mathbf{R}$.

Let $\mathbf{h}^* = (h_1^*, h_2^*, \cdots, h_p^*) \in \mathbf{H}^p$, then we have

$$\ddot{l}_{h\boldsymbol{\beta}}(\boldsymbol{\beta}, \mathbf{\Sigma}, h; \mathbf{O})[h_1] = \ddot{l}^{\,T}_{\boldsymbol{\beta}h}(\boldsymbol{\beta}, \mathbf{\Sigma}, h; \mathbf{O})[h_1],$$

$$\ddot{l}_{\boldsymbol{\beta}h}(\boldsymbol{\beta}, \mathbf{\Sigma}, h; \mathbf{O})[\mathbf{h}^*]$$
$$= \left(\ddot{l}_{\boldsymbol{\beta}h}(\boldsymbol{\beta}, \mathbf{\Sigma}, h; \mathbf{O})[h_1^*], \ddot{l}_{\boldsymbol{\beta}h}(\boldsymbol{\beta}, \mathbf{\Sigma}, h; \mathbf{O})[h_2^*], \cdots, \ddot{l}_{\boldsymbol{\beta}h}(\boldsymbol{\beta}, \mathbf{\Sigma}, h; \mathbf{O})[h_p^*]\right),$$

$$\ddot{l}_{h\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\Sigma}, h; \mathbf{O})[\mathbf{h}^*] = \ddot{l}^{\,T}_{\boldsymbol{\beta}h}(\boldsymbol{\beta}, \boldsymbol{\Sigma}, h; \mathbf{O})[\mathbf{h}^*],$$

$$\ddot{l}_{\boldsymbol{\Sigma}h}(\boldsymbol{\beta}, \boldsymbol{\Sigma}, h; \mathbf{O})[\mathbf{h}^*]$$

$$= \left(\ddot{l}_{\boldsymbol{\Sigma}h}(\boldsymbol{\beta}, \boldsymbol{\Sigma}, h; \mathbf{O})[h_1^*], \ddot{l}_{\boldsymbol{\Sigma}h}(\boldsymbol{\beta}, \boldsymbol{\Sigma}, h; \mathbf{O})[h_2^*], \cdots, \ddot{l}_{\boldsymbol{\Sigma}h}(\boldsymbol{\beta}, \boldsymbol{\Sigma}, h; \mathbf{O})[h_p^*]\right),$$

$$\ddot{l}_{h\boldsymbol{\Sigma}}(\boldsymbol{\beta}, \boldsymbol{\Sigma}, h; \mathbf{O})[\mathbf{h}^*] = \ddot{l}^{\,T}_{\boldsymbol{\Sigma}h}(\boldsymbol{\beta}, \boldsymbol{\Sigma}, h; \mathbf{O})[\mathbf{h}^*],$$

$$\ddot{l}_{hh}(\boldsymbol{\beta}, \boldsymbol{\Sigma}, h; \mathbf{O})[h_1, \mathbf{h}^*]$$

$$= \left(\ddot{l}_{hh}(\boldsymbol{\beta}, \boldsymbol{\Sigma}, h; \mathbf{O})[h_1, h_1^*], \ddot{l}_{hh}(\boldsymbol{\beta}, \boldsymbol{\Sigma}, h; \mathbf{O})[h_1, h_2^*], \cdots, \ddot{l}_{hh}(\boldsymbol{\beta}, \boldsymbol{\Sigma}, h; \mathbf{O})[h_1, h_p^*]\right),$$

$$\ddot{l}_{hh}(\boldsymbol{\beta}, \boldsymbol{\Sigma}, h; \mathbf{O})[\mathbf{h}^*, h_1] = \ddot{l}^{\,T}_{hh}(\boldsymbol{\beta}, \boldsymbol{\Sigma}, h; \mathbf{O})[h_1, \mathbf{h}^*].$$

For simplicity, $\lesssim$ denotes that the left-hand side is bounded above by the right-hand side times a constant; and $\gtrsim$ denotes that the left-hand side is bounded below by the right-hand side times a constant.

In the methodology, $h(\cdot)$ is defined on a bounded interval. To make the integral meaningful, we extended $h(\cdot)$ to the whole real line with $h(x) = 0$ if x is outside the bounded interval. Then under conditions C1-C3, the likelihood and all of its derivatives above are continuous and bounded. In fact, from the proof in Chapter 3, the likelihood for each single subject is bounded. Therefore, the integral against a probability measure is still bounded. Consequently, B is bounded by two positive numbers. We will use this property for bounded functions throughout the proof.

It is straightforward to get all the derivatives with respect to $\boldsymbol{\theta} = (\boldsymbol{\alpha}, h)$ from the above formulae. And this operation will not change the results. Under this parameterization, by applying the Taylor expansion, we obtain

$$\mathbb{P}\left\{l(\boldsymbol{\theta}; \mathbf{O}) - l(\boldsymbol{\theta}_0; \mathbf{O})\right\}$$

$$= \frac{1}{2}\mathbb{P}\Big\{(\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)^T \ddot{l}_{\boldsymbol{\alpha}\boldsymbol{\alpha}}(\boldsymbol{\theta}_0; \mathbf{O})(\boldsymbol{\alpha} - \boldsymbol{\alpha}_0) + \ddot{l}_{hh}(\boldsymbol{\beta}_0, h_0; \mathbf{O})[h - h_0, h - h_0]$$

$$+2(\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)^T \ddot{l}_{\boldsymbol{\alpha}h}(\boldsymbol{\theta}_0, \mathbf{O})[h - h_0]\Big\} + o(d^2(\boldsymbol{\theta}, \boldsymbol{\theta}_0)).$$

Since B is bounded by two positive numbers, we can omit the terms $\frac{1}{B}$, $\frac{1}{B^2}$ in the proof. By the condition C1, the eigenvalues of $\mathbf{\Sigma}_0$ are bounded from 0 and ∞. Therefore, the eigenvalues of $\mathbf{\Sigma}_0^{-1}$ are also bounded from 0 and ∞. Also, using the fact that the integral of a bounded function over a probability space is still bounded and the proof techniques in Chapter 3, we can verify the three conditions in Shen and Wong (1994):

1) $\inf_{\{d(\boldsymbol{\theta},\boldsymbol{\theta}_0)\geq\varepsilon,\boldsymbol{\theta}\in\Theta_n^p\}} \mathbb{P}(l(\boldsymbol{\theta}_0;\mathbf{O}) - l(\boldsymbol{\theta};\mathbf{O})) \gtrsim \varepsilon$.

2) $\mathbb{P}\left\{l(\boldsymbol{\beta},h;\mathbf{O}) - l(\boldsymbol{\beta}_0,h_0;\mathbf{O})\right\}^2 \leq C_2(||\boldsymbol{\beta}-\boldsymbol{\beta}_0||^2 + ||h-h_0||_2^2)$.

3) Let $\boldsymbol{\theta}_{0,n} = (\boldsymbol{\beta}_0, \mathbf{\Sigma}_0, h_{0,n})$, and $\mathcal{F}_n = \{l(\boldsymbol{\theta};\mathbf{O}) - l(\boldsymbol{\theta}_{0,n};\mathbf{O}) : \boldsymbol{\theta} \in \Theta_n^p\}$, where

$$||h_{0,n} - h_0||_\infty = O(q_n^{-q}) = O(n^{-q\nu}).$$

Then the L_∞-metric entropy of the space $\mathcal{F}_n$ satisfies

$$H(\varepsilon, \mathcal{F}_n, ||\cdot||_\infty) = \log N_{[]}(\varepsilon, \mathcal{F}_n, ||\cdot||_\infty) \leq c_7(cq_n + p)\log(1/\varepsilon) \leq c_8 n^\nu \log(1/\varepsilon).$$

It follows from Theorem 1 in Shen and Wong (1994) that

$$d(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0) = O_p\left(\max(n^{-\frac{1-\nu}{2}}, n^{-q\nu}, n^{-q\nu})\right) = O_p(n^{-\min(q\nu,\frac{1-\nu}{2})}).$$

□

The following theorem establishes the asymptotic normality of the sieve MLEs of $\boldsymbol{\alpha}$.

Theorem 4.2. Under the regularity conditions C1-C5,

$$n^{\frac{1}{2}}(\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0) \rightarrow N(\mathbf{0}, \mathbf{I}^{-1}(\boldsymbol{\alpha}_0))$$

in distribution, where $I(\boldsymbol{\alpha}_0) = \mathcal{P}\left\{\tilde{l}_{\boldsymbol{\alpha}}(\boldsymbol{\theta}_0; \mathbf{O})^{\otimes 2}\right\}$, and $\tilde{l}_{\boldsymbol{\alpha}}(\boldsymbol{\theta}_0; \mathbf{O})$ is the efficient score function of $\boldsymbol{\alpha}_0$.

Remark 4.3. To prove Theorem 4.2, we check the assumptions of Theorem 6.1 in Wellner and Zhang (2007). Specifically, we need to verify the consistency and rate of convergence, positive information, stochastic equicontinuity of the estimators and the smoothness of the model. The key idea is that we need to construct ε-baskets in which the target functions can be bounded by ε times a positive number. Following the proof of Theorem 6.1 in Wellner and Zhang (2007), we need to verify the following assumptions.

(i) $d(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0) = O_p(n^{-\delta})$ for some $\delta > 0$.

(ii) $\mathbb{P}\dot{l}_{\boldsymbol{\alpha}}(\boldsymbol{\theta}_0; \mathbf{O}) = 0$ and $\mathbb{P}\dot{l}_h(\boldsymbol{\theta}_0; \mathbf{O})[h] = 0$ for all $h \in \mathbf{H}$.

(iii) There exists an $\mathbf{h}^* = (h_1^*, h_2^*, \cdots, h_(p_x + p_z^2)^*)^T$, where $h_j^* \in \mathbf{H}$ for $j = 1, \cdots, p_x + p_z^2$, such that $\mathbb{P}\ddot{l}_{\boldsymbol{\alpha}h}(\boldsymbol{\theta}_0; \mathbf{O})[h] - \mathbb{P}\ddot{l}_{hh}(\boldsymbol{\theta}_0; \mathbf{O})[\mathbf{h}^*, h] = \mathbf{0}$, for all $h \in \mathbf{H}$. Furthermore, the matrix $\mathbb{P}\left\{\ddot{l}_{\boldsymbol{\alpha}\boldsymbol{\alpha}}(\boldsymbol{\theta}_0; \mathbf{O}) - \ddot{l}_{h\boldsymbol{\alpha}}(\boldsymbol{\theta}_0; \mathbf{O})[\mathbf{h}^*]\right\}$ is nonsingular.

(iv) $\mathbb{P}_n\dot{l}_{\boldsymbol{\alpha}}(\hat{\boldsymbol{\theta}}_n, \mathbf{O}) = o_p(n^{-\frac{1}{2}})$ and $\mathbb{P}_n\dot{l}_h(\hat{\boldsymbol{\theta}}_n; \mathbf{O})[\mathbf{h}^*] = o_p(n^{-\frac{1}{2}})$.

(v) For any $c > 0$,

$$\sup_{\left\{d(\boldsymbol{\theta}, \boldsymbol{\theta}_0) \leq cn^{-\delta}, \boldsymbol{\theta} \in \Theta_n^p\right\}} \left|\mathbb{G}_n\dot{l}_{\boldsymbol{\alpha}}(\boldsymbol{\theta}; \mathbf{O}) - \mathbb{G}_n\dot{l}_{\boldsymbol{\alpha}}(\boldsymbol{\theta}_0; \mathbf{O})\right| = o_p(1),$$

and

$$\sup_{\left\{d(\boldsymbol{\theta}, \boldsymbol{\theta}_0) \leq cn^{-\delta}, \boldsymbol{\theta} \in \Theta_n^p\right\}} \left|\mathbb{G}_n\dot{l}_g(\boldsymbol{\theta}; \mathbf{O})[\mathbf{h}^*] - \mathbb{G}_n\dot{l}_g(\boldsymbol{\theta}_0; \mathbf{O})[\mathbf{h}^*]\right| = o_p(1).$$

(vi) For some $\zeta > 1$ and $\delta\zeta > 1/2$, consider a neighborhood of $\boldsymbol{\theta}_0$:

$$\left\{\boldsymbol{\theta}_c : |\boldsymbol{\alpha}_c - \boldsymbol{\alpha}_0| + ||h_c - h_0||_2 \leq cn^{-\delta}\right\}.$$

Then

$$\left|\mathbb{P}\dot{l}_{\boldsymbol{\alpha}}(\boldsymbol{\theta};\mathbf{O}) - \mathbb{P}\dot{l}_{\boldsymbol{\alpha}}(\boldsymbol{\theta}_0;\mathbf{O}) - \mathbb{P}\ddot{l}_{\boldsymbol{\alpha}\boldsymbol{\alpha}}(\boldsymbol{\theta}_0;\mathbf{O})(\boldsymbol{\alpha}_c - \boldsymbol{\alpha}_0) - \mathbb{P}\ddot{l}_{\alpha h}(\boldsymbol{\theta}_0;\mathbf{O})[h_c - h_0]\right|$$

$$= O\left((|\boldsymbol{\alpha}_c - \boldsymbol{\alpha}_0| + ||h_c - h_0||_2)^{\zeta}\right),$$

and

$$\left|\mathbb{P}\dot{l}_h(\boldsymbol{\theta};\mathbf{O})[\mathbf{h}^*] - \mathbb{P}\dot{l}_h(\boldsymbol{\theta}_0;\mathbf{O})[\mathbf{h}^*] - \mathbb{P}\ddot{l}_{h\boldsymbol{\alpha}}(\boldsymbol{\theta}_0;\mathbf{O})(\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)[\mathbf{h}^*](\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)\right.$$

$$\left.- \mathbb{P}\ddot{l}_{hh}(\boldsymbol{\theta}_0;\mathbf{O})[\mathbf{h}^*, h - h_0]\right| = O\left((|\boldsymbol{\alpha}_c - \boldsymbol{\alpha}_0| + ||h_c - h_0||_2)^{\zeta}\right).$$

Notice that the term $\tilde{l}_{\boldsymbol{\alpha}}(\boldsymbol{\theta}_0;\mathbf{O})$ in Theorem 4.2 is given by $\dot{l}_{\boldsymbol{\alpha}}(\boldsymbol{\theta}_0;\mathbf{O}) - \dot{l}_h(\boldsymbol{\theta}_0;\mathbf{O})[\mathbf{h}^*]$. The proof mimics the proof of 3.2 in Chapter 3.

A consistent estimator for the asymptotic variance-covariance matrix is provided in the following theorem.

Theorem 4.3. Under the regularity conditions C1-C5,

$$\mathbb{P}_n\left\{\tilde{l}_{\boldsymbol{\alpha}}(\hat{\boldsymbol{\theta}}_n;\mathbf{O})^{\otimes 2}\right\} \to \mathbb{P}\left\{\tilde{l}_{\boldsymbol{\alpha}}(\boldsymbol{\theta}_0;\mathbf{O})^{\otimes 2}\right\}$$

in probability, where $\tilde{l}_{\boldsymbol{\alpha}}(\hat{\boldsymbol{\theta}}_n;\mathbf{O})$ is the estimator of $\tilde{l}_{\boldsymbol{\alpha}}(\boldsymbol{\theta}_0;\mathbf{O})$.

The proof is similar to the proof of Theorem 3.3 in Chapter 3.

4.4 Simulation Studies

We conduct simulation studies to examine the finite-sample performance of our proposed method. In the study, we consider the count data. We include three covariates (p_x=3) X_1, X_2 and X_3 in the model. Specifically, X_1 is generated from a normal distribution with mean 1 and variance 0.5^2, X_2 is generated from the uniform distribution $U(-1, 2.25)$, and

X_3 is generated from a bernoulli distribution with success probability 0.5. We set the true values $\boldsymbol{\beta}_0 = \frac{1}{\sqrt{3}}(1,1,1)$ and $g(t) = \log(t)$. We include a random intercept in the model which follows $N(0, 0.6^2)$. The outcome Y is generated from a Poisson distribution with mean $\exp\left\{h(\mathbf{X}^T\boldsymbol{\beta}_0 + b)\right\}$. In the simulations, we consider two different scenarios for $h(\cdot)$: (i) $h(t) = t$; (ii) $h(t) = \cos(t)$. For scenario (i), we divide X_1, X_2 and X_3 by 2. The log-likelihood is approximated by the Gauss-Hermite quadrature with $N_q = 10$ quadrature points. Suppose $a_k, k = 1, \cdots, N_q$ are the abscissas. Let σ denote the standard deviation of the random intercept b. Then we adopt three data-adaptive interior knots that were placed at the equally spaced quantiles of $\mathbf{X}_{ij}^T\boldsymbol{\beta} + \sqrt{2}\sigma a_k$, $j = 1, \cdots, k_i, i = 1, \cdots, n, k = 1, \cdots, N_q$ given $\boldsymbol{\beta}$ and σ, and employ the cubic ($M = 4$) B-spline to approximate the unknown function. We obtain the sieve MLEs iteratively by using the following iterative procedures.

(1) Choose the initial values $\boldsymbol{\beta}$ and σ.

(2) Given $\boldsymbol{\beta}$, σ and the abscissas $a_k, k = 1, \cdots, N_q$ of the Gauss-Hermite quadrature, calculate the knots and basis functions.

(3) Given the knots and basis functions, obtain $\hat{\boldsymbol{\gamma}}$ by solving $\partial l_n(\boldsymbol{\beta}, \sigma, \boldsymbol{\gamma})/\partial\boldsymbol{\gamma} = \mathbf{0}$.

(4) Given the knots, basis functions and $\hat{\boldsymbol{\gamma}}$, obtain $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}$ by solving $\partial l_n(\boldsymbol{\beta}, \boldsymbol{\sigma}, \boldsymbol{\gamma})/\partial\boldsymbol{\beta} = \mathbf{0}$ and $\partial l_n(\boldsymbol{\beta}, \sigma, \boldsymbol{\gamma})/\partial\sigma = 0$.

(5) Set the initial values for next iteration as $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$, $\sigma = \hat{\sigma}$.

(6) Repeat steps (2)-(5) until the maximum of the component-wise absolute value difference of $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}$ between two consecutive iterations is less than 10^{-3}.

The maximization in steps (3) and (4) is carried out by the quasi-Newton algorithm. We estimate the variance-covariance matrix for the regression parameter $\boldsymbol{\beta}$ using the inverse of the consistent variance-covariance matrix estimator for the score functions. We compare the sieve MLEs based on model (4.1) with the sieve MLEs based on model (3.1) without random effects and the standard Poisson GLMM with a random intercept and a canonical log link function. For fair comparison, we standardize the estimates of the regression parameters (excluding the fixed intercept) of the GLMM such that the Euclidean norm is 1. For each case, we considered the sample sizes $n = 120$, 200 with 5 observations for each subject. The results are based on 1000 replicates.

Tables 4.1 and 4.3 present the summary statistics of the sieve MLEs of the unknown regression parameters $\boldsymbol{\beta}$ and σ using our method with $h(t) = t$ and $h(t) = \cos(t)$ respectively. In these tables, Bias is the sampling average of the biases of the estimates; SE denotes the empirical standard deviations of the parameter estimates; SEE denotes the average of the standard error estimates; CP is the coverage probability of 95% confidence interval estimates; and MSE is the mean squared error of the parameter estimates. We observe that the biases of the proposed sieve MLEs are small under all simulation settings. The standard error estimates exhibit the true variation well. The coverage probabilities of the 95% confidence interval estimates are close to the nominal level.

Tables 4.2 and 4.4 present the results from New_0 (model (3.1) without random effects) and the Poisson GLMM with a random intercept and a log link function. Notice that the results for σ under the New_0 are not applicable. The column RE is the MSE relative efficiency of our proposed estimator compared to the estimators based on other models. In Table 4.2, the relative efficiencies for all parameters compared to the New_0 are greater than 1. These results are expected because the model (3.1) ignores the within-subject correlations. As expected, the proposed sieve MLEs are less efficient than the MLEs from the GLMM with the correctly specified link function under most cases, with the efficiency loss of less than 10%. The loss of efficiency is caused by estimating the unknown link function in our method. We also notice that when the sample size increases, the relative efficiency compared to the GLMM improves. The sieve MLEs appear to be more efficient with $h(t) = \cos(t)$ in Table 4.4 compared to both the estimators based on model (3.1) and the standard GLMM. Additionally, the estimates under the GLMM in Table 4.4 have larger biases compared to those in Table 4.3 under our proposed estimators.

4.5 Example

In this section, we apply the proposed method to an epileptic data study (Thall and Vail, 1990). The objective of this study is to explore the treatment/drug effects on patients' seizures. Fifty-nine epileptics were enrolled in this study. The numbers of seizures for each patient suffering from epileptic episodes were recorded at the baseline, then followed every two weeks for an eight-week period. Thus, each patient had 4 observations. Patients in this study were randomized to the test group to receive the drug Progabide (Trt= 1) or the control group to receive a placebo (Trt= 0).

Besides the treatment, two additional covariates are also of interest, including Age in years and Base in the logarithm of the baseline counts divided by 4. Variable Age was log-transformed.

In our analysis, we include Age, Trt, Base and the interaction Trt*Base between Trt and Base as covariates, and evaluate the covariate effects using the proposed method with a random intercept and without random effects. We also apply a) the GLMM based on the Poisson distribution with a random intercept and a log link function and b) the GLM based on the Poisson distribution with a log link function to the data. The parameter estimates are summarized in Table 4.5 for the four different approaches. In the table, New denotes our proposed method with a random intercept, while New_0 refers to model (4.1) without random effects. The results using our methods do not include the intercept because of the constraint on the regression parameters for the model identifiability. Models GLM and GLMM were fitted by rountines glm() and glmer() in R, respectively.

The estimate of the standard error σ of the random intercept and the standard error estimate for the estimator are 0.448 and 0.108 using our method, and are 0.501 and 0.058 under the GLMM. Both models detected significant within-subject correlations. By comparing the Wald test statistic for testing $H_0 : \sigma = 0$ against a half-half mixture of a point mass at 0 and a χ^2_1 distribution, we obtain the p-values of 1.67E$-$05 and $<$ 1.0E$-$6 for model (4.1) and the GLMM, respectively. As expected, the tests based on the standard

GLM and model (3.1) tend to be liberal in the presence of within-subject correlations. At the significance level of 0.05, the proposed method detected significant Base effect whereas the GLMM detected both significant Base effect and significant treatment effect. The directions of the estimates are the same from these two models with a random intercept. Notice that the interpretation of the parameter estimates that are obtained by our method is intrinsically tied to the link function.

Figure 4.1 presents the estimated curve $\hat{h}_n(t)$ in the region of $\mathbf{X}_{ij}^T\hat{\boldsymbol{\beta}}_n+\hat{b}_{in}$, $i = 1, \cdots, n, j = 1, \cdots, k_i$, where $\hat{b}_{in} = E(b_i|\mathbf{Y}_i, \mathbf{X}_i, \hat{\boldsymbol{\theta}}_n)$ is the predictor of the random intercept for the ith subject given the observed data and the sieve MLEs $\hat{\boldsymbol{\theta}}_n$. It appears that $\hat{h}_n(t)$ is close to a straight line within the interval $[-2, 2]$ suggesting a link function close to the canonical log link function in the Poisson GLMM. To check the fit of the proposed method to the epileptic data, in Figure 4.2 we also plot the model-fitted mean curve $\exp\{\hat{h}_n(t)\}$ and compare it with the observed data as well as the empirically estimated mean curve. The model-fitted mean curve and the empirically estimated mean curve agree very well indicating a good fit of the proposed method.

4.6 Discussion

We have proposed the sieve MLEs to estimate the regression parameters for the GLMMs with an unknown link function using B-spline smoothing, and have established the asymptotic properties of our proposed estimators. The estimators of the regression coefficients and the variance-covariance matrix of the random effects achieve the semiparametric efficiency bound. The simulation studies demonstrate that model (4.1) outperforms the ordinary GLMMs when the link function is misspecified and there is little loss of efficiency with a correctly specified link function. We have applied our method to the epileptic data and an ad hoc model checking procedure by comparing the model-fitted mean curve with the empirical estimate mean curve suggests the proposed method fits the data well. show that the proposed method can notably handle the longitudinal data and yield satisfying results.

It is often of interest to interpret the model by increasing one unit in one covariate. Usually the interpretation can be straightforward for the canonical link. Discussion on the interpretation of a nonparametric link function can be found in Climov et al. (2002) and Chiou and Müller (2005).

Table 4.1: Summary Statistics of the proposed sieve MLEs with $h(t) = t$ for poisson data

n	Parameter	True	Bias	SE	SEE	MSE	CP
120	β_1	0.577	-0.0161	0.146	0.158	0.0217	0.938
	β_2	0.577	-0.0091	0.110	0.119	0.0121	0.953
	β_3	0.577	-0.0226	0.145	0.158	0.0215	0.940
	σ	0.600	-0.0201	0.139	0.140	0.0197	0.932
200	β_1	0.577	-0.0042	0.113	0.116	0.0128	0.938
	β_2	0.577	-0.0137	0.089	0.089	0.0080	0.930
	β_3	0.577	-0.0113	0.113	0.117	0.0129	0.941
	σ	0.600	-0.0223	0.096	0.102	0.0097	0.920

True is the true value for the parameter; Bias is the bias of the parameter estimate; SE is the empirical standard deviation of the parameter estimates; SEE is the average of the standard error estimates; MSE is the mean squared error; and CP is the coverage probability of the 95% confidence interval estimates.

Table 4.2: Comparison of the proposed sieve MLEs with the sieve MLEs assuming independence (New_0) and the MLEs from the GLMM with $h(t) = t$ for poisson data

		New_0				GLMM			
n	Parameter	Bias	SE	MSE	RE	Bias	SE	MSE	RE
120	β_1	-0.0449	0.194	0.0396	1.83	-0.0164	0.142	0.0203	0.94
	β_2	-0.0080	0.146	0.0214	1.77	-0.0063	0.108	0.0117	0.97
	β_3	-0.0317	0.189	0.0367	1.71	-0.0222	0.139	0.0198	0.92
	σ					-0.0050	0.125	0.0157	0.80
200	β_1	-0.0225	0.139	0.0198	1.54	-0.0046	0.111	0.0123	0.96
	β_2	-0.0070	0.112	0.0127	1.58	-0.0098	0.090	0.0082	1.01
	β_3	-0.0163	0.142	0.0205	1.59	-0.0141	0.111	0.0125	0.97
	σ					-0.0022	0.100	0.0099	1.03

Bias is the bias of the parameter estimate; SE is the empirical standard deviation of the parameter estimates; MSE is the mean squared error; and RE is the relative efficiency of the proposed sieve MLEs with the estimators under the indicating model.

Table 4.3: Summary statistics of the proposed sieve MLEs with $h(t) = \cos(t)$ for poisson data

n	Parameter	True	Bias	SE	SEE	MSE	CP
120	β_1	0.577	-0.0128	0.091	0.093	0.0084	0.952
	β_2	0.577	0.0024	0.067	0.068	0.0045	0.948
	β_3	0.577	-0.0073	0.086	0.090	0.0075	0.946
	σ	0.600	-0.0095	0.116	0.106	0.0135	0.942
200	β_1	0.577	-0.0074	0.068	0.071	0.0046	0.961
	β_2	0.577	0.0031	0.050	0.052	0.0025	0.953
	β_3	0.577	-0.0056	0.066	0.069	0.0043	0.958
	σ	0.600	-0.0026	0.081	0.079	0.0065	0.942

True is the true value for the parameter; Bias is the bias of the parameter estimate; SE is the empirical standard deviation of the parameter estimates; SEE is the average of the standard error estimates; MSE is the mean squared error; and CP is the coverage probability of the 95% confidence interval estimates.

Table 4.4: Comparison of the proposed sieve MLEs with the sieve MLEs assuming independence (New_0) and the MLEs from the GLMM with $h(t) = \cos(t)$ for poisson data

		New_0				GLMM			
n	Parameter	Bias	SE	MSE	RE	Bias	SE	MSE	RE
120	β_1	-0.0129	0.104	0.0110	1.31	-0.0433	0.106	0.0131	1.56
	β_2	0.0001	0.077	0.0060	1.33	0.0282	0.079	0.0070	1.55
	β_3	-0.0110	0.102	0.0106	1.40	-0.0116	0.103	0.0116	1.43
	σ					-0.0125	0.123	0.0154	1.14
200	β_1	-0.0073	0.079	0.0062	1.34	-0.0436	0.085	0.0092	1.97
	β_2	-0.0003	0.058	0.0033	1.31	0.0316	0.064	0.0051	2.02
	β_3	-0.0055	0.075	0.0056	1.29	-0.0063	0.082	0.0068	1.58
	σ					-0.0260	0.097	0.0100	1.54

Bias is the bias of the parameter estimate; SE is the empirical standard deviation of the parameter estimates; MSE is the mean squared error; and RE is the relative efficiency of the proposed sieve MLEs with the estimators under the indicating model.

Table 4.5: Results for the analysis of the epileptic data

	Est	SE	P-value	Est	SE	P-value
	New			New_0		
Age	0.478	0.291	1.00E-01	0.622	0.032	<1.00E-06**
Trt	-0.297	0.168	7.70E-02	-0.245	0.034	<1.00E-06**
Base	0.766	0.184	3.05E-05**	0.688	0.039	<1.00E-06**
Trt*Base	0.310	0.253	2.20E-01	0.282	0.045	<1.00E-06**
σ	0.448	0.108	1.67E-05**			
	GLMM			GLM		
Intercept	1.795	0.104	<1.00E-06**	1.860	0.041	<1.00E-06**
Age	0.481	0.346	1.64E-01	0.888	0.117	<1.00E-06**
Trt	-0.334	0.147	2.33E-02**	-0.346	0.061	<1.00E-06**
Base	0.883	0.131	<1.00E-06**	0.949	0.044	<1.00E-06**
Trt*Base	0.339	0.202	9.40E-02	0.562	0.064	<1.00E-06**
σ	0.501	0.058	<1.00E-06**			

** indicates significant effect at the 5% significance level.

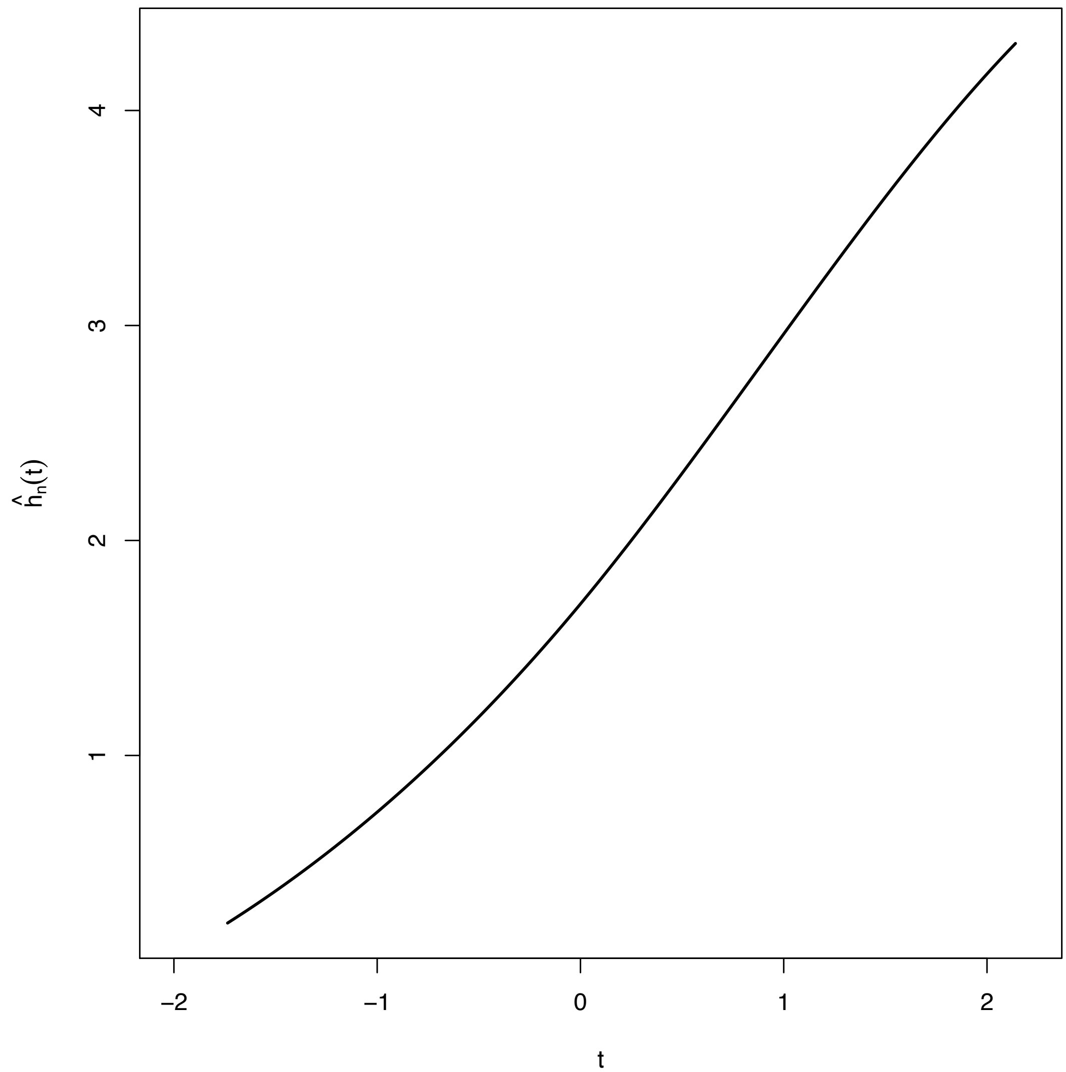

Figure 4.1: Estimated curve $\hat{h}_n(t)$ in the region of $\mathbf{X}_{ij}^T\hat{\boldsymbol{\beta}}_n + \hat{b}_{in}$, $i = 1, \cdots, n, j = 1, \cdots, k_i$, where $\hat{b}_{in} = E(b_i|\mathbf{Y}_i, \mathbf{X}_i, \hat{\boldsymbol{\theta}}_n)$ is the predictor of the random intercept for the ith subject given the observed data and the sieve MLEs $\hat{\boldsymbol{\theta}}_n$.

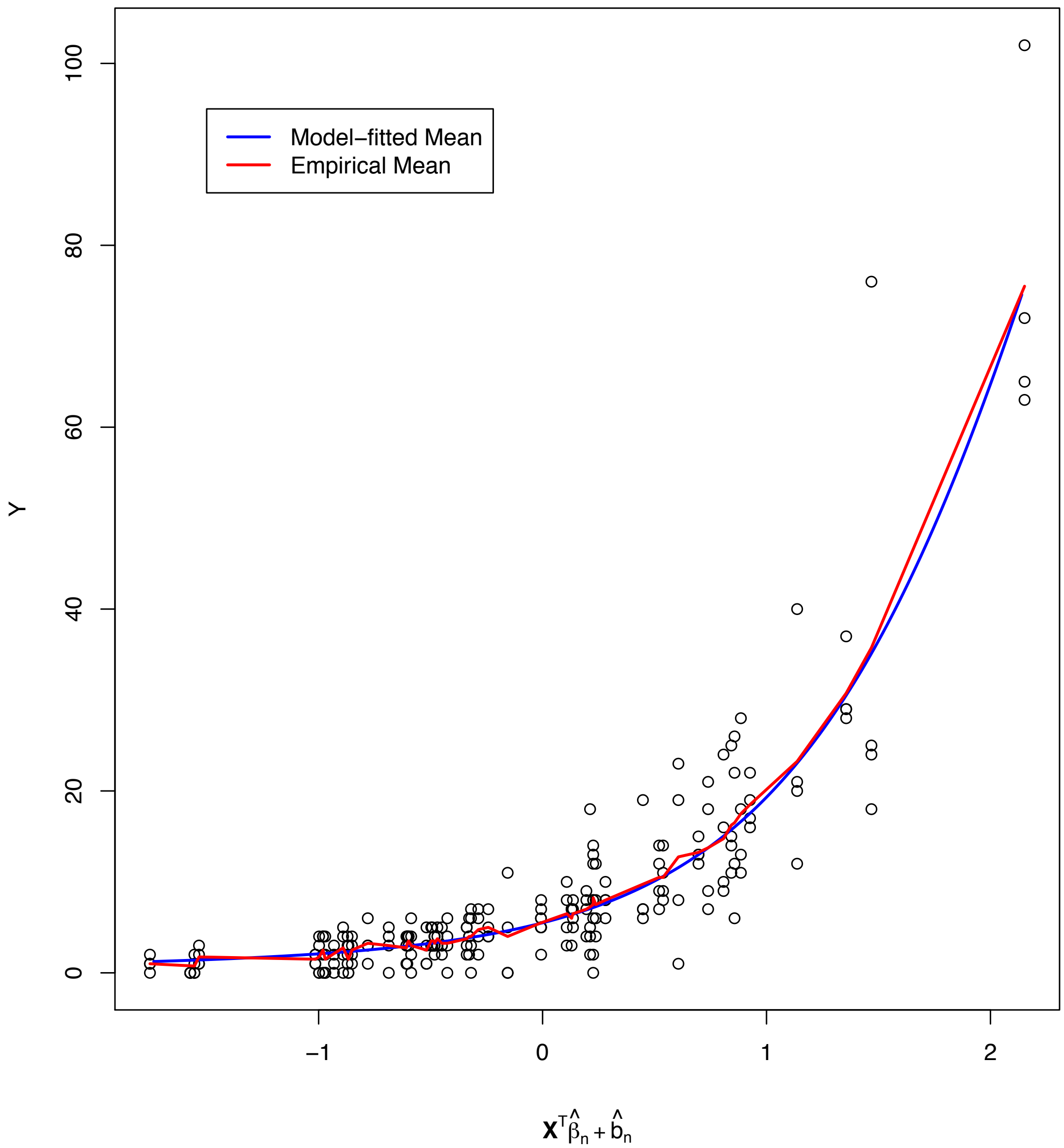

Figure 4.2: Model fitted mean curve $\exp\{\hat{h}_n(t)\}$ (blue solid curve) and the empirical estimated mean curve (red solid curve). The circles correspond to the observed data Y_{ij} against $\mathbf{X}_{ij}^T\hat{\boldsymbol{\beta}}_n + \hat{b}_{in}$, $i = 1, \cdots, n, j = 1, \cdots, k_i$.

Chapter 5: Conclusions

5.1 Discussion

We have proposed a generalized odds rate model incorporating short-term and long-term covariate effects for the survival data and derived the NPMLEs of the unknown parameters for the model estimation. We have also proposed the sieve MLEs for the GLMs/GLMMs with an unknown link function using B-splines smoothing method for both data with independent observations and longitudinal data. We derive the asymptotic properties of the proposed estimators. Extensive simulation studies demonstrate that the proposed methods perform well in practical settings. Specifically, the odds rate model is substantially more powerful than the PO model when the PO assumption is violated especially when the short-term and long-term effects are in opposite directions. Furthermore, the proposed sieve MLEs have smaller biases and standard errors than the MLEs based on the GLMs and GLMMs when the link function is mis-specified. The efficiency loss of the proposed sieve MLEs is limited compared to the MLEs under the GLMs and GLMMs with correctly specified link functions. Applications show that the proposed methods can notably handle the real data and yield satisfying results. We have also implemented the proposed methods in computer programs in C language.

5.2 Future work

We assume in Chapter 2 that the failure time and the censoring time are conditionally independent given the covariates. Violation of this assumption may lead to biased results. For example, in the analysis of time to develop distant metastasis in the CNA study, censoring is likely to be informative because some local recurrences might censor the observations while being driven by the same tumor. To account for such dependent censoring, one may assume certain dependence structures between the failure time and the censoring time (e.g., see Tanaka and Rao (2005) and Li et al. (2007)). Alternatively, we can adopt the technique proposed by Othus et al. (2009) through the use of the conditional crude hazard function of the censoring time given the covariates. Future research along this direction is warranted.

In Chapter 2, we assume independent observations. In several applications, we may encounter clustered or family data. One interesting future work is to extend our proposed method by incorporating random effects to account for the correlations within a cluster/family. Suppose $\mathbf{X}_{ij}$ is the set of covariates for the jth subject in the ith family. We can incorporate random effects in the following model

$$\frac{F(t|\mathbf{X}_{ij}, \omega_i)}{S(t|\mathbf{X}_{ij}, \omega_i)} = H(t)\omega_i \left[e^{\boldsymbol{\beta}^T \mathbf{X}_{ij}} F(t) + e^{\boldsymbol{\gamma}^T \mathbf{X}_{ij}} S(t) \right],$$

where ω_i denotes the shared frailty for the ith family. We are currently investigating the theoretical and computational aspects of this model.

Yang and Prentice (2005) proposed a two-sample model that accommodates the short-term and long-term covariate effects on the hazard function. More recently, Diao et al. (2013) extended the above two-sample model to a regression setting accommodating possibly external time-dependent covariates. While the models of Yang and Prentice (2005) and Diao et al. (2013) are extensions of the Cox proportional hazards model, the proposed model renders direct interpretations of short-term and long-term covariate effects on the survival function and includes the proportional odds model as a special case. It is well known that both the proportional hazards model and the proportional odds model belong to a general class of semiparamtric transformation models (Zeng et al., 2008). One may be interested in investigating the short-term and long-term covariate effects under this class of transformation models. Future research is warranted.

The flat lines after about 80 months in Figure 2.1 suggest that a proportion of patients in the CNA study may never develop distant metastasis, i.e., there exists a cured sub-population. To account for the presence of a cured sub-population, we can combine the proposed model and the standard mixture cure model. Specifically, given the covariates $\mathbf{Z}_i$ and $\mathbf{X}_i$, the mixture cure model postulates that the population survival function takes the form

$$S_{pop}(t|\mathbf{Z}_i, \mathbf{X}_i) = 1 - \theta(\mathbf{Z}_i) + \theta(\mathbf{Z}_i)S(t|\mathbf{X}_i),$$

where

$$\theta(\mathbf{Z}_i) = \frac{\exp(\beta^T \mathbf{Z}_i)}{1 + \exp(\beta^T \mathbf{Z}_i)}$$

and $S(t|\mathbf{X}_i)$ is the conditional survival function in the un-cured sub-population. Note that $1 - \theta(\mathbf{Z}_i)$ is the cure fraction. We assume the proposed model for $S(t|\mathbf{X}_i)$. The likelihood function is then given by

$$\prod_{i=1}^{n} \left[\left\{ \theta(\mathbf{Z}_i) f(Y_i|\mathbf{X}_i) \right\}^{\Delta_i} (1 - \theta(\mathbf{Z}_i) + \theta(\mathbf{Z}_i) S(Y_i|\mathbf{X}_i))^{1-\Delta_i} \right]^{I(Y_i < \infty)} \left\{ 1 - \theta(\mathbf{Z}_i) \right\}^{I(Y_i = \infty)},$$

where $f(t|\mathbf{X}_i)$ is the conditional density function of the failure time in the un-cured sub-population. We can perform nonparametric likelihood estimation and inference using techniques similar to those in Zeng et al. (2006) and Diao and Yin (2012).

In the CNV study in Chapter 2, we performed hypothesis testing at each SNP locus to search for the significant location at which the CNV impacts the time to develop distant metastasis, leading to the so-called "large p small n" problem. Appropriate procedures are needed to control for the family-wise error rate introduced by the multiple comparisons. Many statistical methods have been proposed in literature to account for multiple testing. Particularly, Lin (2005) developed a Monte Carlo approach to approximate the joint distribution of the test statistics along the genome. It would be interesting to adopt this approach in our model setting to adjust for multiple testing. Moreover, resampling methods can be used to approximate the joint distribution of the test statistics, which provides another solution for the "large p small n" problem (Diao and Vidyashankar, 2013).

An interesting area is focused on the selection of important variables that impact the failure time. Recently, Lu and Zhang (2007) proposed variable selection procedures using adaptive LASSO for the PO model. It would be desirable to extend the adaptive LASSO procedure to select variables with significant long-term and/or short-term effects.

It is known that using splines method to approximate the unknown function could cause undersmoothing problem. It would be interesting to develop an estimation procedure based on the penalized likelihood and compare the results. Penalized spline smoothing method has been well discussed by Eilers and Marx (1996). They proposed to base the penalty on finite differences of the coefficients of adjacent B-spline basis functions. Suppose $\gamma_i, i = 1, \cdots, K_n + M$ are the coefficients of the B-spline basis functions. Then the penalty takes the form

$$\frac{\lambda}{2} \sum_{j=k+1}^{K_n+M} \left(\Delta^k \gamma_j\right)^2 = \frac{\lambda}{2} \boldsymbol{\gamma}^T \mathbf{D} \boldsymbol{\gamma},$$

where $\Delta^2 \gamma_j = \Delta\Delta\gamma_j = \gamma_j - 2\gamma_{j-1} + \gamma_{j-2}$.

One important issue in the use of penalized likelihood is how to choose an appropriate value for the tuning parameter λ. There are a number of criteria that can be used to choose the tuning parameter λ. The generalized cross-validation (GCV) score (Golub et al., 1979) and the Akaike's information criterion (AIC) are two mainly used criteria. Other criteria are also investigated for selecting the smoothing parameter, see Hurvich et al. (1998), Wood (2000) and Ueki and Fueda (2010).

The AIC is defined as the corrected log-likelihood of a fitted model for the effective dimension of the parameter vector. Therefore we need to determine the effective dimension of the parameter vector under the model setting. For B-spline methods, Eilers and Marx (1996) used the trace of the influence matrix as an approximation of the effective degree of the parameters. Imoto and Konishi (2003) also discussed the approximation based on the empirical influence function of the estimators and the score function. The cross-validation score is defined based on the mean square error. Unlike the ordinary cross-validation score, the GCV score weights the sample differently adjusting the mean square error for periodicity and non-equidistant samples. Eilers and Marx (1996) discussed the formula of the GCV score for spline smoothing. Bai et al. (2009), Ruppert and Carroll (2000) and Yu and Ruppert (2002) adopted the GCV score to select the tuning parameter for spline fitting with truncated power function basis. Sakamoto and Shirahata (1999) also studied the smoothing parameter selection using cross-validation.

The results in our data analysis in Chapter 4 do not provide strong evidence that the log link assumption in the Poisson GLMM is violated. It would be interesting to further investigate the link function. We can obtain the estimates of the link function using our method and then compare it to the canonical link function. It would be interesting to develop a formal procedure to test $H_0 : h'(t) = c, \forall t \in [a, b]$ for some constant c. Future research is warranted toward this direction.

Bibliography

Aranda-Ordaz, F. J. (1981): “On two families of transformations to additivity for binary response data,” *Biometrika*, 68, 357–363.

Bai, Y., W. K. Fung, and Z. Y. Zhu (2009): “Penalized quadratic inference functions for single-index models with longitudinal data,” *Journal of Multivariate Analysis*, 100, 152–161.

Bennett, S. (1983a): “Analysis of survival data by the proportional odds model,” *Statistics in Medicine*, 2, 273–277.

Bennett, S. (1983b): “Log-logistic regression-models for survival data,” *Applied Statistics - Journal of the Royal Statistical Society Series C*, 32, 165–171.

Bickel, P. J., C. A. J. Klaassen, Y. Ritov, and J. A. Wellner (1993): *Efficient and Adaptive Estimation for Semiparametric Models*, Baltimore: Johns Hopkins University Press.

Chen, Y. Q., N. Hu, S. C. Cheng, P. Musoke, and L. P. Zhao (2012): “Estimating regression parameters in an extended proportional ddds model,” *Journal of the American Statistical Association*, 107, 318–330.

Chiou, J. M. and H. G. Müller (1998): “Quasi-likelihood regression with unknown link and variance functions,” *Journal of the American Statistical Association*, 93, 1376–1387.

Chiou, J. M. and H. G. Müller (2005): “Estimated estimating equations: semiparametric inference for clustered and longitudinal data,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67, 531–553.

Climov, D., J. Hart, and L. Simar (2002): "Automatic smoothing and estimation in single index poisson regression," *Journal of Nonparametric Statistics*, 14, 307–323.

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

Dauxois, J. Y. and S. Kirmani (2003): "Testing the proportional odds model under random censoring," *Biometrika*, 90, 913–922.

Diao, G. and A. N. Vidyashankar (2013): "Assessing genome-wide statistical significance for large p small n problems," *Genetics*, 194, 781–783.

Diao, G. and G. Yin (2012): "A general transformation class of semiparametric cure rate frailty models," *Annals of the Institute of Statistical Mathematics*, 64, 959–989.

Diao, G., D. Zeng, and S. Yang (2013): "Efficient semiparametric estimation of short-term and long-term hazard ratios with right-censored data," *Biometrics*, in press.

Ding, Y. and B. Nan (2011): "A sieve m-theorem for bundled parameters in semiparametric models, with application to the efficient estimation in a linear model for censored data," *Annals of statistics*, 39, 2795.

Eilers, P. H. and B. D. Marx (1996): "Flexible smoothing with b-splines and penalties," *Statistical science*, 89–102.

Fan, J. Q., N. E. Heckman, and M. P. Wand (1995): "Local polynomial kernel regression for generalized linear models and quasi-likelihood functions," *Journal of the American Statistical Association*, 90, 141–150.

Golub, G. H., M. Heath, and G. Wahba (1979): "Generalized cross-validation as a method for choosing a good ridge parameter," *Technometrics*, 21, 215–223.

Heagerty, P. J. and B. F. Kurland (2001): "Misspecified maximum likelihood estimation and generalized linear mixed models," *Biometrika*, 88, 973–985.

Heagerty, P. J. and S. L. Zeger (2000): "Marginalized multilevel models and likelihood inference," *Statistical Science*, 15, 1–26.

Huang, J. and A. Rossini (1997): "Sieve estimation for the proportional-odds failure-time regression model with interval censoring," *Journal of the American Statistical Association*, 92, 960–967.

Hunter, D. R. and K. Lange (2002): "Computing estimates in the proportional odds model," *Annals of the Institute of Statistical Mathematics*, 54, 155–168.

Hurvich, C., J. Simonoff, and C. Tsai (1998): "Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60, 271–293.

Imoto, S. and S. Konishi (2003): "Selection of smoothing parameters in B-spline nonparametric regression models using information criteria," *Annals of the Institute of Statistical Mathematics*, 55, 671–687.

Jiang, C. R. and J. L. Wang (2011): "Functional single index models for longitudinal data," *The Annals of Statistics*, 39, 362–388.

Lam, K. F., Y. W. Lee, and T. L. Leung (2002): "Modeling multivariate survival data by a semiparametric random effects proportional odds model," *Biometrics*, 58, 316–323.

Li, Y., R. C. Tiwari, and S. Guha (2007): "Mixture cure survival models with dependent censoring," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69, 285–306.

Lin, D. Y. (2005): "An efficient monte carlo approach to assessing statistical significance in genomic studies," *Bioinformatics*, 21, 781–787.

Liu, I., B. Mukherjee, T. Suesse, D. Sparrow, and S. K. Park (2009): "Graphical diagnostics to check model misspecification for the proportional odds regression model," *Statistics in Medicine*, 28, 412–429.

Lu, W. B. and H. H. Zhang (2007): "Variable selection for proportional odds model." *Statistics in Medicine*, 26, 3771–3781.

Mallick, B. K. and A. E. Gelfand (1994): "Generalized linear models with unknown link functions," *Biometrika*, 81, 237–245.

Muggeo, V. M. R. and G. Ferrara (2007): "Fitting generalized linear models with unspecified link function: A p-spline approach," *Computational Statistics and Data Analysis*, 52, 2529–2537.

Murphy, S. A. (1994): "Consistency in a proportional hazards model incorporating a random effect," *The Annals of Statistics*, 22, 712–731.

Murphy, S. A. (1995): "Asymptotic theory for the frailty model," *The Annals of Statistics*, 23, 182–198.

Murphy, S. A., A. J. Rossini, and A. W. van der Vaart (1997): "Maximum likelihood estimation in the proportional odds model," *Journal of the American Statistical Association*, 92, 968–976.

Othus, M., Y. Li, and R. C. Tiwari (2009): "A class of semiparametric mixture cure survival models with dependent censoring," *Journal of the American Statistical Association*, 104, 1241–1250.

Pang, Z. and L. Xue (2012): "Estimation for the single-index models with random effects," *Computational Statistics & Data Analysis*, 56, 1837–1853.

Parner, E. (1998): "Asymptotic theory for the correlated gamma-frailty model," *The Annals of Statistics*, 26, 183–214.

Pettitt, A. N. (1984): "Proportional odds models for survival data and estimates using ranks," *Applied Statistics - Journal of the Royal Statistical Society Series C*, 33, 169–175.

Pourahmadi, M. (1999): "Joint mean-covariance models with applications to longitudinal data: unconstrained parameterisation," *Biometrika*, 86, 677–690.

Pourahmadi, M. (2000): "Maximum likelihood estimation of generalized linear models for multivariate normal covariance matrix," *Biometrika*, 87, 425–435.

Pregibon, D. (1981): "Logistic regression diagnostics," *The Annals of Statistics*, 9, 705–724.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery (1992): *Numerical recipes in C: the art of scientific computing*, New York: Cambridge University Press, 2 edition.

Rabinowitz, D., R. A. Betensky, and A. A. Tsiatis (2000): "Using conditional logistic regression to fit proportional odds models to interval censored data," *Biometrics*, 56, 511–518.

Rossini, A. J. and A. A. Tsiatis (1996): "A semiparametric proportional odds regression model for the analysis of current status data," *Journal of the American Statistical Association*, 91, 713–721.

Ruppert, D. and R. J. Carroll (2000): "Theory & methods: Spatially-adaptive penalties for spline fitting," *Australian & New Zealand Journal of Statistics*, 42, 205–223.

Sakamoto, W. and S. Shirahata (1999): "Likelihood-based cross-validation score for selecting the smoothing parameter in maximum penalized likelihood estimation," *Communications in Statistics-Theory and Methods*, 28, 1671–1698.

Scallan, A., R. Gilchrist, and M. Green (1984): "Fitting parametric link functions in generalised linear models," *Computational Statistics and Data Analysis*, 2, 37–49.

Scheike, T. H. (2006): "A flexible semiparametric transformation model for survival data," *Lifetime Data Analysis*, 12, 461–480.

Schumaker, L. L. (1981): *Spline functions: basic theory*, Wiley New York.

Shen, X. (1998): "Proportional odds regression and sieve maximum likelihood estimation," *Biometrika*, 85, 165–177.

Shen, X. and W. Wong (1994): "Convergence rate of sieve estimates," *The Annals of Statistics*, 22, 580–615.

Sun, J. G., L. Q. Sun, and C. Zhu (2007): "Testing the proportional odds model for interval-censored data," *Lifetime Data Analysis*, 13, 37–50.

Tanaka, Y. and P. V. Rao (2005): "A proportional hazards model for informatively censored survival times," *Journal of statistical planning and inference*, 129, 253–262.

Thall, P. F. and S. C. Vail (1990): "Some covariance models for longitudinal count data with overdispersion," *Biometrics*, 46, 657–671.

Ueki, M. and K. Fueda (2010): "Optimal tuning parameter estimation in maximum penalized likelihood method," *Annals of the Institute of Statistical Mathematics*, 62, 413–438.

van der Vaart, A. W. and J. A. Wellner (1996): *Week convergence and empirical processes: with applications to statistics*, New York: Springer-Verlag.

Weisberg, S. and A. H. Welsh (1994): "Adapting for the missing link," *The Annals of Statistics*, 22, 1674–1700.

Wellner, J. A. and Y. Zhang (2007): "Two likelihood-based semiparametric estimation methods for panel count data with covariates," *The Annals of Statistics*, 35, 2106–2142.

Wood, S. (2000): "Modelling and smoothing parameter estimation with multiple quadratic penalties," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62, 413–428.

Xu, P. and L. Zhu (2012): "Estimation for a marginal generalized single-index longitudinal model," *Journal of Multivariate Analysis*, 105, 285–299.

Yang, S. and R. Prentice (2005): "Semiparametric analysis of short-term and long-term hazard ratios with two-sample survival data," *Biometrika*, 92, 1–17.

Yang, S. and R. L. Prentice (1999): "Semiparametric inference in the proportional odds regression model," *Journal of the American Statistical Association*, 94, 125–136.

Yu, Y. and D. Ruppert (2002): "Penalized spline estimation for partially linear single-index models," *Journal of the American Statistical Association*, 97, 1042–1054.

Zeger, S. L., K. Y. Liang, and P. S. Albert (1988): "Models for longitudinal data: a generalized estimating equation approach," *Biometrics*, 44, 1049–1060.

Zeng, D. and D. Y. Lin (2007): "Semiparametric transformation models with random effects for recurrent events," *Journal of the American Statistical Association*, 102, 167–180.

Zeng, D., D. Y. Lin, and X. Lin (2008): "Semiparametric transformation models with random effects for clustered failure time data," *Statistica Sinica*, 18, 355–377.

Zeng, D., D. Y. Lin, and G. Yin (2005): "Maximum likelihood estimation for the proportional odds model with random effects," *Journal of the American Statistical Association*, 100, 470–483.

Zeng, D., G. Yin, and J. G. Ibrahim (2006): "Semiparametric transformation models for survival data with a cure fraction," *Journal of the American Statistical Association*, 101, 670–684.

Zhang, Y., J. W. M. Martens, J. X. Yu, J. Jiang, A. M. Sieuwerts, M. Smid, J. G. M. Klijn, Y. Wang, and J. Foekens (2009): "Copy number alterations that predict metastatic capability of human breast cancer," *Cancer Research*, 69, 3795–3801.

BIOGRAPHY

Mengdie earned her B.S. degree in Statistics from the University of Science and Technology of China (USTC) in 2009. She joined the doctoral program in statistics at George Mason University (GMU) afterwards and received her Master's degree in Statistical Science from GMU in 2011. In Summer 2012 and Summer 2013, she worked as a student fellow intern at the Food and Drug Administration (FDA). She focused her research interest on survival analysis, longitudinal data analysis, semiparametric models, statistical genetics and biosimilarity studies. Her publications and preprints include the following:

1. Yuan, M. and Diao, G. (2011). Joint Association Analysis of Bivariate Quantitative and Qualitative Traits. *BMC Proceedings*, 5 (Suppl 9): S74.
2. Tsong, Y., Yuan, M., Dong, X., Wu, Y. and Shen, M. (2012). Comparing the Response Rates for Superiority, Non-inferiority and Equivalence Testing with Multiple-to-one Matched Binary Data. *Journal of Biopharmaceutical Statistics*, 23:1, 98-109.
3. Yuan, M. and Diao, G. (2014a). Semiparametric Odds Rate Model for Modeling Short-term and Long-term Effects with Application to a Breast Cancer Genetic Study. *International Journal of Biostatistics*, in press. (An earlier version of this paper won the 2013 ICSA/ISBS Student Paper Award.)
4. Yuan, M. and Diao, G. (2014b). Association Analysis of Censored Failure Time Data with a Cured Fraction in Family Studies. To be submitted.
5. Yuan, M. and Diao, G. (2014c). Sieve Maximum Likelihood Estimation for Semiparametric GLMMs with Bundled Fixed Effects and Random Effects. To be submitted.

She is the recipient of the following awards, fellowships and scholarships:

1. Distinguished Academic Achievement, Ph.D., Statistical Science (GMU) Spring 2014
2. Outstanding Graduate Student (WSS) Spring 2014
3. Dissertation Completion Grant (GMU) Spring 2014
4. ICSA/ISBS Student Paper Award (2013 ICSA/ISBS) 2013
5. ORISE Fellow Certificate (FDA) Summer 2012, Summer 2013
6. Merck Tuition Fellowship (GMU) Fall 2010, Spring 2011
7. IT&E Academic Fellowship (GMU) Fall 2009, Fall 2010, Fall 2011
8. Merck Graduate Fellowship (GMU) Fall 2009
9. Outstanding Student Scholarship (USTC) 2006–2008