The Pennsylvania State University

The Graduate School

STATISTICAL MODELS FOR SCALAR RESPONSE WITH LONGITUDINAL COVARIATES

A Dissertation in

Statistics

by

Hanyu Yang

Submitted in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

December 2014

The dissertation of Hanyu Yang was reviewed and approved* by the following:

Runze Li
Distinguished Professor of Statistics
Dissertation Advisor and Chair of Committee

Naomi S. Altman
Professor of Statistics

Zhibiao Zhao
Associate Professor of Statistics

Stephanie T. Lanza
Research Associate Professor of Health and Human Development

Aleksandra B. Slavkovic
Associate Professor of Statistics
Associate Head for Graduate Studies of the Department of Statistics

*Signatures are on file in the Graduate School.

Abstract

Motivated by several health behavior studies, this dissertation is concerned with modeling distal outcome with longitudinal covariates. Specifically, the covariates for each subject are repeatedly observed at a sequence of time points, while the response is measured at one single time point, typically, at the end of the study. Such data are different from those in standard longitudinal models, where both response and covariates are repeatedly collected for each of many subjects.

We first develop models for scalar response with discrete longitudinal covariates, when the covariates can be characterized by a distribution from an exponential family. Recognizing that the observed longitudinal covariates are often subject to measurement errors while the true subject-specific profiles are unobservable, we propose a two-stage calibration regression procedure to estimate the effect function using the natural cubic smoothing spline technique. The performance of the proposed estimation procedure is compared under various circumstances via a set of simulation studies. The consistency and asymptotic normality of the estimated population profile function in the longitudinal covariate model are established, while allowing the number of observation time points to diverge as the sample size increases. A slight extension of the model is further proposed to accommodate an extra clustered random effect, then adopted in a study of alcoholic couples.

In many drug and alcohol studies, the longitudinal covariates are counted values with excess of zeros. Thus, special techniques for handling the zero inflation are considered. Specifically, we propose estimating the longitudinal covariate processes through a hurdle model with zero-truncated Poisson distribution. Simulation experiments are conducted to evaluate the performance and feasibility of this particular estimation procedure. The method is then applied to data from an alcohol study for further illustration.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to express the deepest appreciation to my advisor, Professor Runze Li, for his continuous support and guidance throughout the course of my PhD dissertation research. His intelligence and experience, as well as enthusiasm, patience and generosity, bring me to the right track, and help me all the time of my research and writing of this dissertation. His inputs and advices on both of my research and career are more than words can tell.

I want to sincerely thank Professor Anne Buu for her mentorship on my research projects. Her knowledge and experience enlighten my work, and her encouragement and motivation help to build up my confidence.

I would show my gratitude to Professors Naomi S. Altman, Stephanie T. Lanza and Zhibiao Zhao, for their serving on my dissertation committee, as well as their suggestions and insightful comments on my work.

I also want to extend the thanks to my friends and colleagues, as well as faculty and staff at the Department of Statistics, for making my time at State College a treasurable experience in my life.

Finally but most importantly, I would like to gratefully thank my parents, for their love and support all the time.

This dissertation research is supported by National Institutes of Health (NIH) grants P50-DA010075, P50-DA036107 and R01-CA168676. The content is solely the responsibility of the author and does not necessarily represent the official views of the NIH.

Chapter 1

Introduction

1.1 Background

Longitudinal data analysis has been widely adopted in a variety of research fields. Typically, a longitudinal dataset consists of repeated observations of a response and a vector of covariates for each of many subjects. Longitudinal studies usually have advantages in increased power and robustness to model selection. However, the within-subject correlations, as well as the often highly unbalanced data collected at irregular and possibly subject-specific time points, bring special challenges to the analysis.

In order to account for the within-subject correlations, longitudinal data are usually analyzed through one of the three extensions of generalized linear models in the parametric framework (Diggle et al., 2002): a marginal model separately assumes the regression of response on covariates, and the covariance structures; a transition model treats past outcomes as additional explanatory variables; and a random-effects model addresses the within-subject correlations by introducing a random effects term to the systematic component that yields a generalized linear mixed model.

The interest in longitudinal studies often lies in the pattern of the dependence of response on covariates. Thus nonparametric techniques could be employed to build more flexible and accurate models. For a longitudinal model with one single covariate, Lin and Carroll (2000) extend the generalized estimating equations (Liang and Zeger, 1986) by incorporating a kernel weight function, and propose a local polynomial kernel estimator for the regression function. The working independence estimator, which is obtained when ignoring the correlation structures within subjects, is found to be most efficient. Realizing that the result of Lin and Carroll (2000) is a consequence that not all correlated elements in each subject but asymptotically one datum point could contribute to the estimator, Wang (2003) proposes a marginal kernel method to develop an alternative estimator that properly uses correlations among observations from same subject. Linton et al. (2003) improve the working independence estimator of Lin and Carroll (2000) in another way by constructing a two-stage kernel estimator with a transformation leading to independence. Chen and Jin (2005) exploit generalized inverses of correlation matrices, and develop a local polynomial smoothing method to improve the accuracy of the working independence estimator. Furthermore, Chen et al. (2008) take advantages of methods in Wang (2003) and Chen and Jin (2005), and propose a local linear smoothing estimator that attains linear minimax efficiency when within-subject correlations are correctly specified. By Cholesky decomposition, Yao and Li (2013) propose a different method to simultaneously estimate the correlation structure and regression function, which is more efficient compared with the working independence estimator.

In order to accommodate models with multiple covariates, an additive model is introduced by assuming an additive combination of regression functions. For a genralized additive model (GAM), Hastie and Tibshirani (1986) develop a local scoring

method, a generalization of Fisher scoring algorithm, to iteratively estimate each of the regression functions. A quasi-scoring algorithm is then proposed by Berhane and Tibshirani (1998) to deal with GAMs for longitudinal data problems. Lin and Zhang (1999) study generalized additive mixed models (GAMMs), and propose to estimate regression functions by maximizing the double penalized quasi-likelihood, while smoothing parameters and variance component are obtained simultaneously from the marginal quasi-likelihood.

Another nonparametric approach for analyzing longitudinal data is through varying-coefficient models (Hastie and Tibshirani, 1993), which typically assumes that the effect for each of the covariates varies over time. Hoover et al. (1998) propose a smoothing spline estimator and a local weighted polynomial estimator for varying-coefficient functions. Huang et al. (2002) then consider the number of basis functions used to approximate each coefficient function as a smoothing parameter, and develop a less computationally demanding estimation procedure. Fan and Zhang (2000), alternatively, propose a two-step estimation by first obtaining raw estimates at each individual time point and then smoothing them over time. In a further step, Qu and Li (2006) develop an algorithm that incorporates within-subject correlations, and achieve a more efficient estimator.

Several semiparametric models are then developed to provide more flexible techniques for analyzing longitudinal data. Zhang et al. (1998) consider a semiparametric stochastic mixed model, which simultaneously handles overall regression coefficients and function, and subject-specific random effects and stochastic processes. Fan and Li (2004) study partial linear models for longitudinal data by proposing two estimations along with a variable selection procedure. Furthermore, a varying-coefficient partial linear model is investigated by Fan et al. (2007).

Different from standard longitudinal data problems, another type of questions

arises to handle situations when the covariate is repeatedly measured at multiple time points but the response is collected at one single time point. It is of interest to evaluate the effect of the longitudinal covariates on the future response. In practice, the effect might be a complex function of time; meanwhile the observed longitudinal covariates are often subject to measurement errors and the true profiles are unobservable. James (2002) develops a functional generalized linear model, and estimates the effect function using the EM algorithm by treating latent profiles as missing data. Zhang et al. (2007) propose a two-stage functional mixed model, of which the first stage estimates subject-specific profiles through a nonparametric measurement error model, and the second stage plugs estimated profiles in a functional linear model then estimates the effect function. Bhadra et al. (2012) propose another estimation in a hierarchical Bayesian framework, and adopt a Markov chain Monte Carlo algorithm on the joint likelihood to account for uncertain progression within two stages of the model.

1.2 Contribution

Our work is motivated by several health behavior studies, in which the covariate is observed multiple times and the response is collected at one single future time point. The Michigan Longitudinal Study (MLS) is an ongoing multi-wave prospective study of people at high risk for substance use disorders. In particular, for each of the children in the study, the number of drinking days in the past month of the assessment was recorded annually during ages 13 - 20, and a binary value of alcohol dependence diagnosis was obtained in adulthood. The research aims to delineate the time-varying effect of adolescent alcohol use patterns on the alcohol dependence diagnosis in adulthood, adjusting for other factors such as gender and family history of alcoholism. The problem has discrete values for both longitudinal

covariates and responses. Another longitudinal study recruited alcoholic married couples from the University of Michigan Addiction Treatment Services or local community, where either spouse met DSM-IV diagnosis of past year alcohol use disorder. In each of the next 14 days after the recruitment, participants reported their daily moods and alcohol involvement through an interactive voice response (IVR) system, and a binary covariate, the urge to drink, was then summarized. A continuous scale measurement of depression was taken 6 months after the IVR assessment as the response. The research is to characterize the change in self-reported urge to drink during IVR assessment, as well as its time-varying effect on the depression diagnosis 6 months later. The data contain discrete longitudinal covariates and continuous responses. These problems motivate us to develop new models to handle such data.

We first study models with longitudinal covariate processes that have discrete values. The method proposed in James (2002) has a generalized linear model framework for the response variable, but only handles continuous longitudinal covariates and estimates the latent profiles without smoothness. The model in Zhang et al. (2007) is designed for continuous longitudinal covariates and continuous scalar response, while the one in Bhadra et al. (2012) has similar model settings and considers continuous longitudinal covariates and binary scalar response. We thus extend the two-stage estimation procedure of Zhang et al. (2007) to accommodate situations where either or both of the longitudinal covariates and responses are discrete, allowing the observation time points of longitudinal covariates to be unbalanced among subjects. Particularly, longitudinal profiles are fitted by a GAMM, and quasi-likelihood and the Laplace approximation as those used in Lin and Zhang (1999) are introduced to facilitate computation. Simulation studies under a variety of circumstances are conducted to assess the performance of the

proposed estimation. The consistency and asymptotic normality of the estimated profile population function in the longitudinal covariate model are established, allowing the number of observation time points to increase as the sample size increases. A slight extension of this model is then proposed and applied to a study of alcoholic couples, in which the data were obtained through an IVR system.

Furthermore, in many psychology and sociology studies, longitudinal covariates are usually counted-values with excess of zeros. For example, in the MLS, a significant abundance of zeros for the number of drinking days in one month is observed. Thus, special methods are required to handle such problems. With some mild assumptions on the zero and positive components, we propose to estimate longitudinal covariate processes through a hurdle model. The model is then applied to the MLS alcohol abuse data, and leads to results consistent with empirical findings.

1.3 Organization

This dissertation is organized as follows. Chapter 2 provides literature review on related topics. We begin with some commonly used parametric models for analyzing longitudinal data. Several popular smoothing techniques are then introduced, followed by a series of nonparametric and semiparametric longitudinal models. This chapter also reviews some recently developed models that deal with longitudinal covariates and scalar response.

Models of scalar response with discrete longitudinal covariates are studied in Chapter 3. With the natural cubic smoothing spline method, we propose a two-stage estimation by calibration regression, as well as bias correction procedures. Specifically, the first stage of the proposed method considers the measurement error issues, and estimates latent profile functions by fitting a GAMM; the second

stage plugs in the estimates and achieves effect function for either continuous or discrete responses. A series of simulation studies are presented, and estimation performances under different circumstances are compared. A slight extension of the model that accommodates clustered random effects is then proposed, and adopted for an alcoholic couples study.

For the general model settings considered in Chapter 3, we explore the asymptotic properties of the estimate of the population profile function in the longitudinal covariate model. The problem formulations, conclusions and proofs are discussed in Chapter 4.

In Chapter 5, we focus on a particular model setting, in which the longitudinal covariates are zero-inflated count values. Technical and computational details of handling such models are discussed. We then provide simulation experiments to evaluate the proposed estimation. An application to an alcohol study of the MLS is included to demonstrate the proposed method.

However, under the regression calibration framework, there remain difficulties in characterizing the properties of estimates of regression coefficient and effect function in the response model. We leave these discussions as future research topics in Chapter 6, along with further improvements for the estimation procedures, and possible extensions of the current model settings.

Chapter 2

Literature Review

Longitudinal studies are widely adopted in a variety of research fields. Typically, a longitudinal dataset consists of repeated observations of a response and a vector of covariates for each of many subjects. The interest often lies in either the pattern of change in response over time, or the dependence of response on covariates. Longitudinal studies have advantages in increased power and robustness to model selection. However, the within-subject correlations, as well as the often highly unbalanced data collected at irregular and possibly subject-specific time points, bring special challenges to the analysis. A variety of models have been studied for longitudinal data under different circumstances.

There are also some statistical settings, which are referred to as models with longitudinal covariates, concerning situations when the covariate is repeatedly measured at many time points but the response is collected at one single time point for each subject. Under such settings, the predictor for a subject is in fact a latent profile function, whose value is observed at several distinct time points but possibly subject to measurement error, and the time-varying effect of the profile function on the response is often of interest. Some statistical procedures have been developed to handle these models.

This chapter is organized as follows. In Section 2.1, we review some of the most commonly used parametric models, with emphasis on their applications in longitudinal data analysis. In Section 2.2, we introduce several popular nonparametric smoothing techniques. Various nonparametric and semiparametric models for longitudinal data are reviewed in Sections 2.3 and 2.4, respectively. Section 2.5 reviews the particular models with longitudinal covariates. In additional, a few models dealing with zero-inflated count data are reviewed in Section 2.6.

2.1 Parametric Models

Regression is one of the most important areas in statistics, and is widely used to explore the associations between responses and covariates. Suppose that Y is a response and $\mathbf{x}$ is a vector of covariates, a regression model assumes that Y depends on $\mathbf{x}$ through

$$g(\mathrm{E}(Y)) = m(\mathbf{x}),$$

for a link function $g(\cdot)$, and refers to $m(\cdot)$ as the regression function.

A parametric regression model assumes that $m(\cdot)$ is known except for finitely many parameters, thus inference on $m(\cdot)$ is equivalent to inference on these parameters. In Sections 2.1.1 and 2.1.2, we review a family of linear models, where $m(\cdot)$ depends on the parameters linearly. However, the regression function may depend on the parameters in a nonlinear fashion as well.

2.1.1 Linear Models and Linear Mixed Models

Suppose that data points $(Y_i, \mathbf{x}_i)$, $i = 1, \ldots, n$, are observed, where Y_i is a response and $\mathbf{x}_i$ is a p-dimensional vector of covariates for i-th subject. A linear model (LM)

often assumes that

$$Y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i,$$

where $\boldsymbol{\beta}$ is a p-dimensional vector of fixed effects, and ϵ_i's are independent with $\mathrm{E}(\epsilon_i) = 0$ and $\mathrm{var}(\epsilon_i) = \sigma^2$. The least squares estimate (LSE) of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y},$$

and σ^2 is estimated by using mean squared error (MSE) as

$$\hat{\sigma}^2_{\mathrm{MSE}} = \frac{1}{n-p}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}),$$

where $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ and $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$. If an additional normality assumption is imposed on ϵ_i's, one may obtain the maximum likelihood estimate (MLE) of $\boldsymbol{\beta}$, which coincides with the LSE, and

$$\hat{\sigma}^2_{\mathrm{ML}} = \frac{1}{n}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}).$$

A linear mixed model (LMM) is regarded as a variant of an LM, where some effects are treated not as fixed but as realizations of random variables. Such random effects are usually denoted by a q-dimensional vector $\mathbf{u}$ following some distribution of $\mathrm{N}(\mathbf{0}, \mathbf{D})$. Conditional on the unobservable but realized value of $\mathbf{u}$, an LMM then assumes that

$$Y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{z}_i^T \mathbf{u} + \epsilon_i,$$

where $\mathbf{z}_i$ is another q-dimensional vector of covariates for i-th subject, and ϵ_i's are conditionally independent. By introducing random effects into the model, one can conveniently deal with variances and covariances attributable to factors acknowledged to be affecting the responses, in particular, $\mathrm{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$, and $\mathbf{V} =$

$\text{Var}(\mathbf{Y}) = \mathbf{ZDZ}^T + \sigma^2\mathbf{I}$ with $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)^T$. A normality assumption is usually imposed, leading to $\mathbf{Y} \sim \text{N}(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$. Thus, for a given $\mathbf{V}$, an MLE is derived as

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{Y}.$$

Typically, when refer to longitudinal settings, we denote Y_{ij} as the response and $\mathbf{X}_{ij} = (X_{ij1}, \ldots, X_{ijp})^T$ as the vector of covariates for j-th observation of i-th subject at time t_{ij}, and stack them as $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^T$, $\mathbf{Y} = (\mathbf{Y}_1^T, \ldots, \mathbf{Y}_n^T)^T$, $\mathbf{X}_i = (\mathbf{X}_{i1}, \ldots, \mathbf{X}_{in_i})^T$, $\mathbf{X} = (\mathbf{X}_1^T, \ldots, \mathbf{X}_n^T)^T$, $\mathbf{t}_i = (t_{i1}, \ldots, t_{in_i})^T$, and $\mathbf{t}^0 = (t_1^0, \ldots, t_r^0)^T$ representing an r-dimensional vector of ordered distinct values of all time points t_{ij}'s. Additionally, corresponding notation of $\mathbf{Z}_{ij} = (Z_{ij1}, \ldots, Z_{ijq})^T$ will be used if random effects or time-invariant effects are included in the model.

The LMs and LMMs can be adopted to explore longitudinal data problems. Considering the within-subject correlations in longitudinal data, the model is usually assumed as

$$Y_{ij} = \mathbf{X}_{ij}^T\boldsymbol{\beta} + \mathbf{Z}_{ij}^T\mathbf{u}_i + \epsilon_{ij},$$

where $\mathbf{u}_i$'s used to account for the correlations within subjects are independently distributed as $\text{N}(\mathbf{0}, \mathbf{D})$. The inferences are relatively straightforward. One may refer to Chapters 4 - 6 of Diggle et al. (2002) for the details.

2.1.2 Generalized Linear Models and Generalized Linear Mixed Models

Since the response Y may not be normally distributed, and some function of mean rather than the mean itself might be taken as a linear combination of parameters, it is thus necessary to extend LMs and LMMs to generalized linear models (GLMs) and generalized linear mixed models (GLMMs).

Building a GLM typically involves three components: response distribution, link function, and linear predictor. It is assumed that the density function of Y_i belongs to an exponential family as

$$f_Y(Y_i) = \exp(\frac{Y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(Y_i, \phi)),$$

where θ_i is a canonical parameter, ϕ is a dispersion parameter, and $a_i(\cdot)$, $b(\cdot)$ and $c(\cdot, \cdot)$ are known functions. The parameter in the distribution is then related to the covariates by modeling a transformation of $\mu_i = \mathrm{E}(Y_i)$, which is a function of θ_i, as a linear model of $\mathbf{x}_i$, that is,

$$g(\mu_i) = \eta_i = \mathbf{x}_i^T\boldsymbol{\beta},$$

where $g(\cdot)$ is a known link function, and $\eta_i = \mathbf{x}_i^T\boldsymbol{\beta}$ is referred as linear predictor or systematic component.

The log-likelihood can be derived from the GLM, which is then maximized to yield an MLE of $\boldsymbol{\beta}$. Since the score function does not have a closed solution, the MLE $\hat{\boldsymbol{\beta}}$ is estimated using an iterative reweighted least squares (IRLS) algorithm. In practice, ϕ is treated as a nuisance parameter and estimated by its moment estimate.

However, in many circumstances, it is unlikely that the exact distribution of the response is known. Another concern is that, for sake of robustness, one may prefer to impose some moment conditions, rather than a distribution, on the response. Without having specific distribution assumptions, quasi-likelihood (Wedderburn, 1974) is used to provide inferential methods. Given the covariate $\mathbf{x}_i$, the response Y_i has mean μ_i and variance $a_i(\phi)v(\mu_i)$, where $v(\cdot)$ is a specified variance function, and the function $q_i = \frac{Y_i - \mu_i}{a_i(\phi)v(\mu_i)}$ behaves similarly to the score function of log-

likelihood with respect to Bartlett first and second identities (Bartlett, 1953a,b). The log-quasi-likelihood is then defined as

$$Q_i = \int_{Y_i}^{\mu_i} \frac{Y_i - u}{a_i(\phi)v(u)} du.$$

Thus, inferences can be obtained based on the corresponding quasi-likelihood similarly to the likelihood. One may refer to McCullagh and Nelder (1989) for the complete details of inference procedures in GLMs.

An extension of GLM for longitudinal data problems is the marginal model, in which the regression and within-subject correlations are modeled separately. Specifically, the marginal expection of Y_{ij}, $\mu_{ij} = \mathrm{E}(Y_{ij})$, is related to $\mathbf{X}_{ij}$ by

$$g(\mu_{ij}) = \eta_{ij} = \mathbf{X}_{ij}^T \boldsymbol{\beta},$$

the marginal variance is a function of μ_{ij} as $\mathrm{var}(Y_{ij}) = \phi v(\mu_{ij})$, and the correlation within a subject is assumed to be $\mathrm{corr}(Y_{ij}, Y_{ij'}) = \rho(\mu_{ij}, \mu_{ij'}; \boldsymbol{\alpha})$ for a known function $\rho(\cdot, \cdot)$. The generalized estimating equation (GEE) method (Liang and Zeger, 1986), a multivariate analogue of quasi-likelihood, is developed to estimate $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ by simultaneously solving

$$\mathbf{S}_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{n} \{(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}})^T (\mathrm{Var}(\mathbf{Y}_i))^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i)\} = \mathbf{0},$$
$$\mathbf{S}_{\boldsymbol{\alpha}}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{n} \{(\frac{\partial \boldsymbol{\eta}_i}{\partial \boldsymbol{\alpha}})^T \mathbf{H}_i^{-1} (\mathbf{W}_i - \boldsymbol{\eta}_i)\} = \mathbf{0},$$

where $\mathbf{H}_i$ is a weight matrix, $\mathbf{W}_i$ is a vector of all products of residual pairs, and $\boldsymbol{\eta}_i = \mathrm{E}(\mathbf{W}_i)$. The estimator $\hat{\boldsymbol{\beta}}$ from the GEE method is found to be nearly efficient relative to its MLE counterpart, and consistent even if the correlation structure is incorrectly specified (Diggle et al., 2002).

Random effects can be incorporated into GLMs, which leads to GLMMs. Conditional on the realized value of random effects vector $\mathbf{u}$, a GLMM is specified as

$$f_{Y|\mathbf{U}}(Y_i|\mathbf{u}) = \exp(\frac{Y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(Y_i, \phi)),$$

and the conditional mean $\mu_i = \mathrm{E}(Y_i|\mathbf{u})$ is modeled through

$$g(\mu_i) = \eta_i = \mathbf{x}_i^T\boldsymbol{\beta} + \mathbf{z}_i^T\mathbf{u}.$$

Due to the often intractable calculations of high-dimensional integrals, the use of likelihood is very limited in GLMMs. As an alternative, Breslow and Clayton (1993) propose a penalized quasi-likelihood (PQL) criterion, which applies the Laplace method to approximate the integrated quasi-likelihood of a GLMM, and estimates regression coefficients by the Fisher scoring algorithm. The estimators for variance components are then obtained by maximizing the restricted maximum likelihood (REML; Harville, 1977). In addition, Lin and Breslow (1996) propose bias correction procedures for the estimates of regression coefficients as well as variance components obtained from the PQL method.

Rather than the marginal model, the random-effects model (Laird and Ware, 1982) is another approach to address the within-subject correlations in longitudinal data. The underlying assumption is that there exists natural heterogeneity across subjects in their regression coefficients, which can be represented by some distribution, and the correlation within a subject is assumed to arise from this heterogeneity (Diggle et al., 2002). This model introduces a random effects term $\mathbf{u}_i$ to the systematic component as

$$g(\mathrm{E}(Y_{ij}|\mathbf{u}_i)) = \mathbf{X}_{ij}^T\boldsymbol{\beta} + \mathbf{Z}_{ij}^T\mathbf{u}_i,$$

and assumes that the responses Y_{ij}'s in i-th subject are mutually independent conditional on $\mathbf{u}_i$. Furthermore, the random effects $\mathbf{u}_i$'s are often assumed to follow a distribution of $\mathrm{N}(\mathbf{0}, \mathbf{D}(\boldsymbol{\theta}))$ with a variance component $\boldsymbol{\theta}$.

2.2 Nonparametric Smoothing Techniques

A nonparametric regression model, on the other hand, allows great flexibility and only assumes that the regression function $m(\cdot)$ belongs to some infinite dimensional collection of functions, for example, $m(\cdot)$ is twice-differentiable. The form of $m(\cdot)$ thus is not specified but will be determined by the observed data analytically. Compared with parametric models, nonparametric regression models have several advantages, such as flexibility that may reduce possible modeling bias, and estimability by means of not introducing too many parameters in estimation. In practice, nonparametric regression models can also suggest a suitable parametric model, and provide diagnostic tools for parametric models.

A variety of methods have been developed to make inference on nonparametric regression models. We will describe some of them in this section. In the remaining of this section, without loss of generality, models with one single covariate are considered. For a more straightforward illustration in smoothing techniques, we focus on regression models with the identity link function, which assume $\mathrm{E}(Y|x) = m(x)$ and $\mathrm{var}(Y|x) = \sigma^2(x)$.

2.2.1 Kernel Smoothing and Local Polynomial Regression

We start the review of kernel smoothing method with kernel density estimators (KDEs). Suppose that a sample $\{X_1, \ldots, X_n\}$ is drawn from a density function $f(\cdot)$, and the estimation of $f(\cdot)$ is of interest. Without assuming $f(\cdot)$ belongs to

some distribution family, a KDE for $f(\cdot)$ is defined as

$$\hat{f}_h(x_0) = \frac{1}{n}\sum_{i=1}^{n} K_h(X_i - x_0),$$

for any x_0 within the range, where $K_h(\cdot) = \frac{1}{h}K(\frac{\cdot}{h})$ is a rescaling function, $K(\cdot)$ is a kernel function satisfying $\int K(u)du = 1$, and $h > 0$ is a bandwidth parameter. From the definition, each $K_h(X_i - x_0)$ smoothly redistributes the point mass at X_i, and the KDE is then the average over these density functions. Under some mild conditions, when $n \to \infty$, $h \to 0$ and $nh \to \infty$, the bias and variance of $\hat{f}_h(\cdot)$ are

$$\mathrm{E}(\hat{f}_h(x_0)) - f(x_0) = \frac{\mu_2 f''(x_0)}{2}h^2 + o(h^2),$$

and

$$\mathrm{var}(\hat{f}_h(x_0)) = \frac{\nu_0 f(x_0)}{nh} + o(\frac{1}{nh}),$$

respectively, where $\mu_k = \int u^k K(u)du$ and $\nu_k = \int u^k K^2(u)du$ (Wand and Jones, 1995).

The kernel function usually takes form of a Gaussian density function on $\mathbb{R}$, or a symmetric Beta family function on $[-1, 1]$, although KDE is not very sensitive to the choice of kernel function (Marron and Nolan, 1988). However, the choice of the bandwidth parameter is crucial, and an appropriate h would balance the bias and the variance of the estimator. The ideal bandwidth is assessed by minimizing the asymptotic mean integrated squared error. Silverman (1986) provides rule-of-thumb bandwidths for the Gaussian and symmetric Beta family kernels.

Now suppose that data (Y_i, X_i), $i = 1, \ldots, n$, are independently sampled from a regression model with $\mathrm{E}(Y|x) = m(x)$ and $\mathrm{var}(Y|x) = \sigma^2(x)$. Traditional kernel approaches usually assume that a more remote datum point contains less information about $m(x_0)$, and view a kernel regression estimator as a weighted average,

that is,

$$\hat{m}_h(x_0) = \frac{\sum_{i=1}^n (w_i(x_0) Y_i)}{\sum_{i=1}^n w_i(x_0)},$$

where $w_i(\cdot)$'s are weight functions. In particular, the Nadaraya-Watson (NW) estimator (Nadaraya, 1964; Watson, 1964) takes the weight as

$$w_i^{\mathrm{NW}}(x_0) = K_h(X_i - x_0),$$

and is most widely used; and, alternatively, the Gasser-Muller (GM) estimator (Gasser and Muller, 1979) uses

$$w_i^{\mathrm{GM}}(x_0) = \int_{s_{i-1}}^{s_i} K_h(u - x_0) du,$$

where $s_i = \frac{1}{2}(X_{(i)} + X_{(i+1)})$ with $X_{(0)} = -\infty$ and $X_{(n+1)} = \infty$. The asymptotic biases and variances of both estimators can be found in Chapter 2 of Fan and Gijbels (1996); in particular, the NW estimator yields a larger bias, while the asymptotic variance of the GM estimator is much greater.

Another local smoothing approach is local polynomial regression. Suppose that $m(\cdot)$ has up to d-th derivative, then by the Taylor expansion, one may have an approximation

$$m(\cdot) \approx \sum_{k=0}^{d} \{\beta_k (\cdot - x_0)^k\},$$

where $\beta_k = \frac{m^{(k)}(x_0)}{k!}$ for $k = 0, \ldots, d$. With notation of $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_d)^T$, $\mathbf{x}_i = (1, (X_i - x_0), \ldots, (X_i - x_0)^d)^T$ for $i = 1, \ldots, n$, and $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$, $\mathbf{W} = \mathrm{diag}\{K_h(X_i - x_0)\}$, local polynomial estimation yields

$$\hat{\boldsymbol{\beta}} = \mathrm{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^{n} \{(Y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 K_h(X_i - x_0)\} = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{Y}.$$

Thus, the local estimate of $m^{(k)}(x_0)$ is $\hat{m}^{(k)}(x_0) = k!\hat{\beta}_k$, for $k = 0, \ldots, d$; $\hat{m}(x_0) = \hat{\beta}_0$ is often of primary interest. Under some regularity conditions, when $h \to 0$ and $nh \to \infty$, the asymptotic bias and variance of $\hat{m}^{(k)}(x_0)$ are

$$\mathrm{E}(\hat{m}^{(k)}(x_0)) - m^{(k)}(x_0) = \mathbf{e}_{k+1}^T \mathbf{S}^{-1} \mathbf{c}_d \frac{k!}{(d+1)!} m^{(d+1)}(x_0) h^{d+1-k} + o_P(h^{d+1-k}),$$

and

$$\mathrm{var}(\hat{m}^{(k)}(x_0)) = \mathbf{e}_{k+1}^T \mathbf{S}^{-1} \mathbf{S}^* \mathbf{S}^{-1} \mathbf{e}_{k+1} \frac{(k!)^2 \sigma^2(x_0)}{f(x_0) n h^{1+2k}} + o_P(\frac{1}{nh^{1+2k}}),$$

respectively, where $\mathbf{e}_j$ is a $(d+1)$-dimensional vector with 1 at j-th element and 0 elsewhere, $\mathbf{c}_d = (\mu_{d+1}, \ldots, \mu_{2d+1})^T$, and both $\mathbf{S}$ and $\mathbf{S}^*$ are $(d+1) \times (d+1)$ matrices with $\mathbf{S}_{(i,j)} = \mu_{i+j}$ and $\mathbf{S}^*_{(i,j)} = \nu_{i+j}$, respectively (Fan and Gijbels, 1996).

In fact, local polynomial regression and kernel smoothing are closely related. One may notice that, the local constant estimator coincides with the NW estimator, and a local linear estimator overcomes the disadvantages of the NW and GM kernel estimators.

2.2.2 Spline Smoothing

Spline smoothing methods approximate the regression function by imposing a polynomial spline structure on $m(\cdot)$. Typically, a d-degree piecewise polynomial function divides the range into contiguous intervals by κ knots $X_{(0)} < x_1 < \ldots < x_\kappa < X_{(n)}$, then represents $m(\cdot)$ by a polynomial of degree d within each interval, and enforces continuity conditions of up to $(d-1)$-th derivative at each knot. Thus, d-degree spline functions form a functional space of dimension $K = \kappa + d + 1$, for example, the commonly-used cubic spline functions correspond to a $(\kappa + 4)$-dimensional functional space. Some exceptions include the natural cubic spline (NCS; Green and Silverman, 1994), where the additional natural boundary con-

ditions are imposed, thus $m(\cdot)$ is linear beyond two boundary knots and is κ-dimensional.

A piecewise polynomial function $m(\cdot)$ is usually expressed through a set of bases $\{B_1(\cdot), \ldots, B_K(\cdot)\}$, that is,

$$m(\cdot) \approx \sum_{k=1}^{K} (\beta_k B_k(\cdot)),$$

the estimator $\hat{m}(x_0)$ is thus obtained by estimating β_k's. In practice, the power basis, which takes $\{1, x, \ldots, x^d, (x - x_1)_+^d, \ldots, (x - x_\kappa)_+^d\}$, and the B-spline basis (de Boor, 1978), which does not have an expressive form but allows more stable and efficient computation, are two popular choices.

In practice, one needs to choose the number and locations of knots; smoothing spline, regression spline, and penalized spline are three popular techniques used to deal with this issue. The smoothing spline chooses $\{X_{(2)}, \ldots, X_{(n-1)}\}$ as the $n-2$ knots, and estimates $m(\cdot)$ by minimizing

$$\sum_{i=1}^{n} (Y_i - m(X_i))^2 + \lambda \int (m^{(r)}(u))^2 du,$$

or equivalently, by finding

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}} \{||\mathbf{Y} - \mathbf{B}\boldsymbol{\beta}||^2 + \lambda \boldsymbol{\beta}^T \mathbf{Q} \boldsymbol{\beta}\} = (\mathbf{B}^T\mathbf{B} + \lambda \mathbf{Q})^{-1} \mathbf{B}^T \mathbf{Y},$$

where $\mathbf{B}$ is an $n \times K$ matrix with $\mathbf{B}_{(i,j)} = B_j(X_i)$, $\mathbf{Q}$ is a $K \times K$ matrix with $\mathbf{Q}_{(i,j)} = \int_{X_{(1)}}^{X_{(n)}} B_i^{(r)}(u) B_j^{(r)}(u) du$, and $\lambda \geq 0$ is a smoothing parameter that can be selected by cross-validation (CV; Stone, 1974) or generalized cross-validation (GCV; Goluba et al., 1979; Wahba, 1980). In particular, for a cubic spline smoothing problem, the penalty is often imposed on the second derivative $m''(\cdot)$.

The regression spline initially chooses a large number of knots, and builds the regression function through the power basis, which leads to

$$Y_i \approx \sum_{l=0}^{d} (\beta_l X_i^l) + \sum_{k=1}^{\kappa} \{\beta_{d+k} (X_i - x_k)_+^d\} + \epsilon_i.$$

The method then treats the smoothing problem as a linear regression model, and applies variable selection procedures to select significant terms, which turns out to select some of the knots.

The penalized spline, a linear smoother technique, first has a few but evenly distributed knots $\{x_1, \ldots, x_\kappa\}$, and builds the regression function through power basis as

$$m(x) \approx \sum_{l=0}^{d} (\beta_l x^l) + \sum_{k=1}^{\kappa} \{\beta_{d+k} (x - x_k)_+^d\},$$

then obtains $\hat{m}(\cdot)$, or $\hat{\beta}_k$'s, by minimizing

$$\sum_{i=1}^{n} (Y_i - m(X_i))^2 + \lambda \sum_{k=1}^{\kappa} |\beta_{d+k}|^2,$$

where $\lambda \geq 0$ is a smoothing parameter usually determined by GCV.

2.3 Nonparametric Longitudinal Data Models

Nonparametric techniques have been exploited to build more flexible models for longitudinal data. In this section, we first review several methods for a traditional nonparametric model with one single covariate, and introduce a series of additive models that account for multiple covariates, as well as a couple of varying-coefficient models.

2.3.1 Traditional Nonparametric Regression Models

A variety of nonparametric regression models have been proposed for longitudinal data. Typically, for the one single covariate case, a nonparametric model assumes

$$Y_{ij} = m(X_{ij}) + \epsilon_{ij}, \tag{2.1}$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, n_i$, where $m(\cdot)$ is a nonparametric smooth regression function, and ϵ_{ij} is a random error; or under generalized linear model settings, assumes the marginal mean $\mu_{ij} = \mathrm{E}(Y_{ij})$ through

$$g(\mu_{ij}) = m(X_{ij}), \tag{2.2}$$

where $g(\cdot)$ is a known link function.

Lin and Carroll (2000) extend the GEE method by incorporating a kernel weight function, and propose a local polynomial kernel estimator for $m(\cdot)$. The nonparametric regression function at any given x_0 is approximated using a local polynomial that satisfies

$$m(\cdot) \approx \mathbf{G}_d^T(\cdot - x_0)\boldsymbol{\beta},$$

where $\mathbf{G}_d(t) = (1, t, \ldots, t^d)^T$ and $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_d)^T$, then the estimate is given by $\hat{m}(x_0) = \hat{\beta}_0$.

With variance assumed as $\mathrm{var}(Y_{ij}) = \phi_j w_{ij}^{-1} v(\mu_{ij})$ for dispersion ϕ_j and weight w_{ij}, as well as $\mathbf{G}_{id} = (\mathbf{G}_d(X_{i1} - x_0), \ldots, \mathbf{G}_d(X_{in_i} - x_0))^T$, $\mathbf{K}_{ih} = \mathrm{diag}\{K_h(X_{ij} - x_0)\}$, $\boldsymbol{\mu}_i = ((g^{-1})(m(X_{i1})), \ldots, (g^{-1})(m(X_{in_i})))^T$, $\boldsymbol{\Delta}_i = \mathrm{diag}\{(g^{-1})'(m(X_{ij}))\}$, $\mathbf{S}_i = \mathrm{diag}\{\phi_j w_{ij}^{-1} v(\mu_{ij})\}$, $\mathbf{V}_i = \mathbf{S}_i^{\frac{1}{2}} \mathbf{R}_i \mathbf{S}_i^{\frac{1}{2}}$ for working correlation $\mathbf{R}_i$, two different ker-

nel GEEs for $m(x_0)$ are proposed as

$$\sum_{i=1}^{n}\{\mathbf{G}_{id}^T\boldsymbol{\Delta}_i\mathbf{V}_i^{-1}\mathbf{K}_{ih}(\mathbf{Y}_i-\boldsymbol{\mu}_i)\}=\mathbf{0},$$

and

$$\sum_{i=1}^{n}\{\mathbf{G}_{id}^T\boldsymbol{\Delta}_i\mathbf{K}_{ih}^{\frac{1}{2}}\mathbf{V}_i^{-1}\mathbf{K}_{ih}^{\frac{1}{2}}(\mathbf{Y}_i-\boldsymbol{\mu}_i)\}=\mathbf{0}.$$

A Fisher scoring algorithm can be constructed to update $\hat{\boldsymbol{\beta}}$, while the bandwidth parameter h can be selected by an extension of empirical bias bandwidth selection (Ruppert, 1997). In particular, several versions of the working correlation matrix $\mathbf{R}_i$ might be taken into consideration. By studying the asymptotic performance, Lin and Carroll (2000) claim that the most efficient estimator for $m(\cdot)$, the weighted pooled estimator, is obtained when ignoring the correlation structure within subjects and assuming the working independence of $\mathbf{R}_i=\mathbf{I}$, while correctly specifying the correlation matrix results in an asymptotically less efficient estimator.

Realizing that the result of Lin and Carroll (2000) is a natural consequence of the standard estimating equations, when not all correlated elements in each subject but asymptotically only one datum point could contribute to the estimator $\check{m}(x_0)$, Wang (2003) proposes a marginal kernel method, which properly uses the correlations among observations from the same subject, to develop an alternative estimator. This method first uses the working independence estimator $\check{m}(x_0)$ from Lin and Carroll (2000) as an initial estimator, then constructs a local estimator $\hat{m}(x_0)$ with all Y_{ij}'s in $\mathbf{Y}_i$ if some Y_{ij_0} is used in $\check{m}(x_0)$, while the contributions of Y_{ij}'s but Y_{ij_0} are through their residuals from $\check{m}(x_0)$. Specifically, under the local linear setting, $\hat{\boldsymbol{\beta}}=(\hat{\beta}_0,\hat{\beta}_1)^T$ is solved from kernel-weighted estimating equations

of

$$\sum_{i=1}^{n}\sum_{j=1}^{n_i}[K_h(X_{ij}-x_0)\{(g^{-1})'(\beta_0+\beta_1(X_{ij}-x_0))\mathbf{G}_{i,*j}^T\}\mathbf{V}_i^{-1} \times\{\mathbf{Y}_i-(g^{-1})_{*j}(\mathbf{X}_i,\boldsymbol{\beta},\check{m}(\mathbf{X}_i))\}]=\mathbf{0},$$

where $\mathbf{G}_{i,*j}$ and $(g^{-1})_{*j}(\mathbf{X}_i,\boldsymbol{\beta},\check{m}(\mathbf{X}_i))$, as defined in Wang (2003), specify the local contribution of $\mathbf{Y}_i$. With the derived asymptotic properties, Wang (2003) proves that, the estimator using the true correlation is the optimal one and has a uniformly smaller variance than its working independence counterpart.

Under the balanced design setting of model (2.1), Linton et al. (2003) improve the working independence estimator in another way by constructing a two-stage kernel estimator. By assuming that

$$\mathbf{Y}_i=m(\mathbf{X}_i)+\boldsymbol{\epsilon}_i=m(\mathbf{X}_i)+\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\varepsilon}_i,$$

where $\mathrm{Cov}(\boldsymbol{\varepsilon}_i)=\mathbf{I}$ and $\boldsymbol{\Sigma}$ is a known $n_0\times n_0$ matrix, the proposed method first fits the working independence estimator $\check{m}(\cdot)$ from Lin and Carroll (2000), with bandwidth h_I and weights of $\boldsymbol{\Sigma}_{jj}$'s. A linear transformation

$$\mathcal{Z}_i(f)=\mathbf{Y}_i+\boldsymbol{\Lambda}^{-1}(\boldsymbol{\Sigma}^{-\frac{1}{2}}-\boldsymbol{\Lambda})(\mathbf{Y}_i-f(\mathbf{X}_i)),$$

where $\boldsymbol{\Lambda}=\mathrm{diag}(\boldsymbol{\Sigma}^{-\frac{1}{2}})$, is then applied to $\check{m}(\cdot)$, and yields $\mathcal{Z}_i(\check{m})=\check{m}(\mathbf{X}_i)+\boldsymbol{\Lambda}^{-1}\boldsymbol{\varepsilon}_i$. In the second stage, a working independence estimator $\hat{m}(\cdot)$ is achieved by fitting a local linear regression on $\mathcal{Z}_i(\check{m})$ with bandwidth h and weights ζ_j's, whose optimal choice is suggested as $\zeta_j=\boldsymbol{\Lambda}_{(j,j)}{}^2$. Linton et al. (2003) prove that, the proposed estimator $\hat{m}(\cdot)$ always has a smaller asymptotic variance for any arbitrary covariance $\boldsymbol{\Sigma}$, thus uniformly improves the working independence kernel estimator of Lin and Carroll (2000) in terms of MSE.

Chen and Jin (2005) develop another local polynomial smoothing method to improve the estimation, by exploiting generalized inverses of correlation matrices. Instead of $m(\cdot)$, $\mu(\cdot) = (g^{-1} \circ m)(\cdot)$ is studied for its clearer interpretation, thus the model can be formulated as

$$Y_{ij} = \mu(X_{ij}) + \sigma(X_{ij})\varepsilon_{ij},$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, n_0$, where $\sigma(\cdot) = \{(v \circ \mu)(\cdot)\}^{\frac{1}{2}}$ for given variance function $v(\cdot)$, and the mean-zero errors ε_{ij}'s satisfy $\text{var}(\varepsilon_{ij}) = \phi_j$ for dispersion parameter ϕ_j. With weight matrices $\mathbf{W}_i$'s defined as

$$\mathbf{W}_i = \mathbf{K}_{ih}^{\frac{1}{2}} \mathbf{\Phi}^{-\frac{1}{2}} (\mathbf{I}_i \mathbf{R}_i \mathbf{I}_i)^{-} \mathbf{\Phi}^{-\frac{1}{2}} \mathbf{K}_{ih}^{\frac{1}{2}},$$

where $\mathbf{\Phi}_i = \text{diag}\{\phi_1, \ldots, \phi_{n_0}\}$ and $\mathbf{I}_i = \text{diag}\{\text{I}(|X_{i1} - x_0| \leq h), \ldots, \text{I}(|X_{in_0} - x_0| \leq h)\}$, the weighted least squares estimate of $\boldsymbol{\beta}$ is obtained as

$$\hat{\boldsymbol{\beta}} = \{\sum_{i=1}^{n} (\mathbf{G}_{id}^T \mathbf{W}_i \mathbf{G}_{id})\}^{-1} \sum_{i=1}^{n} (\mathbf{G}_{id}^T \mathbf{W}_i \mathbf{Y}_i),$$

which leads to $\hat{\mu}^{(k)}(x_0) = k!\hat{\beta}_k$. Particularly, the expression of the estimator has a closed form which brings computational simplicity, and the framework of generalized settings is less essential in estimating thus provides a more general estimation procedure. The asymptotic bias and variance of $\hat{\mu}^{(k)}(x_0)$ are discussed in Chen and Jin (2005), and the estimation is optimized when the working correlation matrix $\mathbf{R}_i$ is the true correlation. It has been shown that, utilizing the generalized inverse $(\mathbf{I}_i \mathbf{R}_i \mathbf{I}_i)^{-}$, instead of $\mathbf{R}_i^{-1}$ as that in Lin and Carroll (2000), improves the accuracy of $\hat{\mu}^{(k)}(x_0)$.

Chen et al. (2008) take advantage of both estimators proposed in Wang (2003)

and Chen and Jin (2005), and propose a local linear smoothing estimator $\hat{\mu}(\cdot)$, which attains linear minimax efficiency when the within-subject correlations are correctly specified. With the local weight matrix

$$\mathbf{W}_i = \mathbf{K}_{ih}\{(\mathbf{I} - \bar{\mathbf{1}}_i)\mathbf{V}_i(\mathbf{I} - \bar{\mathbf{1}}_i)\}^-,$$

for $\mathbf{V}_i$ being the working covariance matrix and $\bar{\mathbf{1}}_i$ as defined in Chen et al. (2008), the weighted least squares estimator for $(\mu(x_0), \mu^{'}(x_0))^T$, or equivalently $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$, is obtained as

$$\hat{\boldsymbol{\beta}} = \{\sum_{i=1}^{n}(\mathbf{G}_{i1}^T\mathbf{W}_i\mathbf{G}_{i1})\}^{-1}\sum_{i=1}^{n}(\mathbf{G}_{i1}^T\mathbf{W}_i\mathbf{Y}_i).$$

The asymptotic variance and bias of $\hat{\mu}(x_0)$ are derived, with a conclusion that using true covariance for $\mathbf{V}_i$ leads to better estimation performance. Chen et al. (2008) also establish the linear minimax efficiency of the proposed local linear smoother under true covariance, claiming that such estimator cannot be further improved by using linear procedures.

Yao and Li (2013) propose a different method to simultaneously estimate the correlation structure and the regression function by the Cholesky decomposition and profile least squares techniques. Based on the fact that there exists a lower triangle matrix $\boldsymbol{\Phi}$ with $\boldsymbol{\Phi}_{(j,j)} = 1$, such that $\mathrm{Cov}(\mathbf{e}_i) = \boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Phi}^T = \mathbf{D} = \mathrm{diag}\{d_1^2, \ldots, d_{n_0}^2\}$ for $\mathbf{e}_i = \boldsymbol{\Phi}\boldsymbol{\epsilon}_i$ and $\mathrm{Cov}(\boldsymbol{\epsilon}_i) = \boldsymbol{\Sigma}$, model (2.1) under the balanced design setting of $n_i = n_0$ can be approximated as

$$\begin{aligned} Y_{i1} &= m(X_{i1}) + e_{i1}, \\ Y_{ij} &\approx m(X_{ij}) + \sum_{l=1}^{j-1}(\phi_{jl}\hat{\epsilon}_{il}) + e_{ij}, \text{ for } j = 2, \ldots, n_0, \end{aligned}$$

where $\phi_{jl} = -\mathbf{\Phi}_{(j,l)}$, and $\hat{\epsilon}_{il} = Y_{ij} - \check{m}(X_{ij})$ is estimated from the initial working independence estimator $\check{m}(\cdot)$ of Lin and Carroll (2000), or in matrix form as

$$\mathbf{Y}_{-} \approx m(\mathbf{X}_{-}) + \hat{\mathbf{F}}_{-}\boldsymbol{\phi} + \mathbf{e}_{-},$$

where $\mathbf{Y}_{-}$, $\mathbf{X}_{-}$ and $\mathbf{e}_{-}$ are stacked over observations of $j = 2, \ldots, n_0$, $i = 1, \ldots, n$, and $\hat{\mathbf{F}}_{-}$ and $\boldsymbol{\phi}$ are correspondingly defined.

A profile least squares estimator for $\boldsymbol{\phi}$ is proposed as

$$\hat{\boldsymbol{\phi}} = \{\hat{\mathbf{F}}_{-}^T(\mathbf{I} - \mathbf{S}_{-})^T\hat{\mathbf{D}}_{-}^{-1}(\mathbf{I} - \mathbf{S}_{-})\hat{\mathbf{F}}_{-}\}^{-1}\hat{\mathbf{F}}_{-}^T(\mathbf{I} - \mathbf{S}_{-})^T\hat{\mathbf{D}}_{-}^{-1}(\mathbf{I} - \mathbf{S}_{-})\mathbf{Y}_{-},$$

where $\hat{\mathbf{D}}_{-} = \text{diag}\{\hat{d}_2, \ldots, \hat{d}_{n_0}, \ldots, \hat{d}_2, \ldots, \hat{d}_{n_0}\}$ with some consistent estimates of $\hat{d}_j$'s, $\mathbf{S}_{-}$ is a smoothing matrix depending on $\mathbf{X}_{-}$, and the bandwidth h is selected using the plug-in bandwidth selector of Ruppert et al. (1995) on a difference-based estimate of $\boldsymbol{\phi}$. The local linear smoother $\hat{m}(x_0) = \hat{\beta}_0$ is then achieved from

$$\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1)^T = \text{argmin}_{\boldsymbol{\beta}}\{(\mathbf{Y} - \hat{\mathbf{F}}\hat{\boldsymbol{\phi}} - \mathbf{A}\boldsymbol{\beta})^T\mathbf{W}(\mathbf{Y} - \hat{\mathbf{F}}\hat{\boldsymbol{\phi}} - \mathbf{A}\boldsymbol{\beta})\}.$$

Here, $\mathbf{Y}$, $\hat{\mathbf{F}}$, $\mathbf{A}$ and $\mathbf{W}$ are based on all observations including $j = 1$; in particular, $\mathbf{A}$ is a local linear regression design matrix, $\mathbf{W}$ is a kernel weight matrix, $\mathbf{Y}$ and $\hat{\mathbf{F}}$ are similarly defined as $\mathbf{Y}_{-}$ and $\hat{\mathbf{F}}_{-}$ respectively. The asymptotic normality for $\hat{\boldsymbol{\phi}}$ and $\hat{m}(\cdot)$ is established, while $\hat{m}(\cdot)$ is more efficient compared with the working independence estimator of Lin and Carroll (2000). In addition, Yao and Li (2013) provide the estimation of $m(\cdot)$ under unbalanced settings, as well as the asymptotic bias and variance of the proposed $\hat{m}(\cdot)$.

2.3.2 Additive Models

In more general settings, models with multiple covariates, say $\mathbf{x} = (x_1, \ldots, x_p)^T$, will be of interest. For a standard linear regression model, it is usually assumed that $Y = \beta_0 + \sum_{k=1}^{p}(\beta_k x_k) + \epsilon$, or simply $Y = \sum_{k=1}^{p}(\beta_k x_k) + \epsilon$ if x_1 is set to be 1. An additive model extends the linear model by replacing the linear combination $\sum_{k=1}^{p}(\beta_k x_k)$ with an additive combination $\sum_{k=1}^{p} m_k(x_k)$, where $m_k(\cdot)$'s are centered smooth functions. Similarly, a generalized additive model (GAM; Hastie and Tibshirani, 1986) extends a GLM by incorporating smoothing techniques, and assumes the mean $\mu = \mathrm{E}(Y)$ as

$$g(\mu) = \eta = \beta_0 + \sum_{k=1}^{p} m_k(x_k).$$

The local scoring method, a generalization of Fisher scoring algorithm that replaces IRLS by a weighted additive modeling algorithm via backfitting on partial residuals, is illustrated in Hastie and Tibshirani (1986) for estimating $m_k(\cdot)$'s iteratively.

When dealing with longitudinal data, the expection $\mu_{ij} = \mathrm{E}(Y_{ij})$ in a marginal model is instead assumed as

$$g(\mu_{ij}) = \eta_{ij} = \beta_0 + \sum_{k=1}^{p} m_k(X_{ijk}).$$

Berhane and Tibshirani (1998) propose a local quasi-scoring algorithm, which takes versions of GEE and local scoring, to estimate $m_k(\cdot)$'s by iterating between a weighted additive model and moment estimation of the nuisance parameters.

A generalized additive mixed model (GAMM; Lin and Zhang, 1999) is then proposed by incorporating random effects into a GAM, or as an additive extension of a GLMM. It uses additive nonparametric functions to model covariate effects, while

accounting for overdispersion and within-subject correlations by random effects.

Specifically, Lin and Zhang (1999) consider a GAMM that assumes $\mathrm{E}(Y_i|\mathbf{u}) = \mu_i$ and $\mathrm{var}(Y_i|\mathbf{u}) = \frac{\phi}{w_i} v(\mu_i)$, and models μ_i through

$$g(\mu_i) = \eta_i = \beta_0 + \sum_{k=1}^{p} m_k(x_{ik}) + \mathbf{z}_i^T \mathbf{u}.$$

Here, ϕ is a scale parameter, $v(\cdot)$ is a specified variance function, w_i's are prior weights, $m_k(\cdot)$'s are centered twice-differentiable smooth functions, and $\mathbf{u}$ is distributed as $\mathrm{N}(\mathbf{0}, \mathbf{D}(\boldsymbol{\theta}))$ with a variance component vector of $\boldsymbol{\theta}$. The integrated log-quasi-likelihood is derived from the model, and the natural cubic smoothing spline technique is applied to estimate nonparametric functions $m_k(\cdot)$'s, or equivalently their NCS coefficient vectors $\mathbf{m}_k$'s, through the penalized log-quasi-likelihood, which is then approximated by the Laplace method to avoid numerical integration. Consequently, estimates $\hat{\beta}_0$, $\hat{m}_k(\cdot)$'s and $\hat{\mathbf{u}}$ are obtained by maximizing the double penalized quasi-likelihood

$$-\frac{1}{2\phi} \sum_{i=1}^{n} \tilde{d}_i(Y_i; \mu_i) - \frac{1}{2} \mathbf{u}^T \mathbf{D}^{-1} \mathbf{u} - \frac{1}{2} \sum_{k=1}^{p} (\lambda_k \mathbf{m}_k^T \mathbf{K}_k \mathbf{m}_k)$$

through the Fisher scoring algorithm, where $\tilde{d}_i(Y_i; \mu_i)$'s are conditional deviance functions, $\mathbf{K}_k$'s are smoothing matrices, and λ_k's are non-negative smoothing parameters controlling the goodness-of-fit and smoothness of $\hat{m}_k(\cdot)$'s. It is suggested that the estimators along with their covariance matrices can be obtained through an equivalent GLMM representation of

$$\mathcal{Y} = \mathcal{X}\boldsymbol{\beta} + \mathcal{B}\mathbf{a} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon},$$

where $\mathcal{Y}$ is a modified GLM working vector, $\mathcal{X}$ and $\mathcal{B}$ are design matrices for fixed

term $\boldsymbol{\beta}$ and random term $\mathbf{a}$ that are obtained through mixed effects representations of $\mathbf{m}_k$'s, and $\boldsymbol{\epsilon}$ is an error vector. Meanwhile, the smoothing parameters λ_k's and the variance component $\boldsymbol{\theta}$ can be jointly estimated by marginal quasi-likelihood, an extension of REML.

2.3.3 Varying-Coefficient Models

Another nonparametric approach to analyzing the longitudinal data is through varying-coefficient models (Hastie and Tibshirani, 1993). Typically, a varying-coefficient model for longitudinal data assumes

$$Y_{ij} = \mathbf{X}_{ij}^T \boldsymbol{\beta}(t_{ij}) + \epsilon_i(t_{ij}), \tag{2.3}$$

for mean-zero processes $\epsilon_i(\cdot)$'s; or under the generalized linear model framework, models the mean $\mu_{ij} = \mathrm{E}(Y_{ij})$ as

$$g(\mu_{ij}) = \mathbf{X}_{ij}^T \boldsymbol{\beta}(t_{ij}), \tag{2.4}$$

for a known link function $g(\cdot)$. Here, $\boldsymbol{\beta}(\cdot) = (\beta_1(\cdot), \ldots, \beta_p(\cdot))^T$ is a p-dimensional vector of coefficient functions.

Hoover et al. (1998) propose two estimators, a smoothing spline estimator and a local weighted polynomial estimator, for $\boldsymbol{\beta}(\cdot)$ in model (2.3). The smoothing spline method represents each $\beta_k(\cdot)$ in terms of a set of spline basis functions $\mathbf{B}(\cdot) = (B_1(\cdot), \ldots, B_V(\cdot))^T$, that is,

$$\beta_k(\cdot) \approx \sum_{v=1}^{V} (\gamma_{kv} B_v(\cdot)) = \mathbf{B}^T(\cdot)\boldsymbol{\gamma}_k,$$

where $\boldsymbol{\gamma}_k = (\gamma_{k1}, \ldots, \gamma_{kV})^T$, then obtains $\hat{\beta}_k^{\mathrm{SS}}(\cdot) = \mathbf{B}^T(\cdot)\hat{\boldsymbol{\gamma}}_k$ by minimizing

$$\sum_{i=1}^{n}[\{\mathbf{Y}_i - \sum_{k=1}^{p}(\mathcal{X}_{ik}\mathcal{B}_i\boldsymbol{\gamma}_k)\}^T\{\mathbf{Y}_i - \sum_{k=1}^{p}(\mathcal{X}_{ik}\mathcal{B}_i\boldsymbol{\gamma}_k)\}] + \sum_{k=1}^{p}(\lambda_k\boldsymbol{\gamma}_k^T\mathbf{Q}\boldsymbol{\gamma}_k),$$

with respect to $\boldsymbol{\gamma}_k$'s, where $\mathcal{X}_{ik} = \mathrm{diag}\{X_{ijk}\}$, $\mathcal{B}_i$ is a design matrix for $\boldsymbol{\gamma}_k$ corresponding to the spline basis $\mathbf{B}(\cdot)$, $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_k)^T$ is a vector of smoothing parameters, and $\mathbf{Q}$ is a smoothing matrix. On the other hand, the local weighted polynomial estimates $\hat{\beta}_k^{\mathrm{LWP}}(t_0)$'s, or equivalently $\hat{b}_{0k}$'s, are obtained from

$$\mathrm{argmin}_{(\mathbf{b}_1,\ldots,\mathbf{b}_p)}(\sum_{i=1}^{n}[\{\mathbf{Y}_i - \sum_{k=1}^{p}(\mathcal{X}_{il}\mathbf{G}_{id}\mathbf{b}_k)\}^T\mathbf{W}_i\{\mathbf{Y}_i - \sum_{k=1}^{p}(\mathcal{X}_{il}\mathbf{G}_{id}\mathbf{b}_k)\}]),$$

where $\mathbf{b}_k = (b_{0k}, \ldots, b_{dk})^T$, and $\mathbf{G}_{id}$ is its local polynomial design matrix at t_0, $\mathbf{W}_i$ is a kernel weight matrix. Typically, as a local constant fit, the kernel estimator yields

$$\hat{\boldsymbol{\beta}}^{\mathrm{K}}(t_0) = \{\sum_{i=1}^{n}(\mathbf{X}_i^T\mathbf{K}_i\mathbf{X}_i)\}^{-1}\sum_{i=1}^{n}(\mathbf{X}_i^T\mathbf{K}_i\mathbf{Y}_i),$$

where $\mathbf{K}_i = \mathrm{diag}\{K_h(t_{ij} - t_0)\}$. Both smoothing parameters of $\boldsymbol{\lambda}$ in $\hat{\boldsymbol{\beta}}^{\mathrm{SS}}$ and h in $\hat{\boldsymbol{\beta}}^{\mathrm{LWP}}$ are selected by CV. Hoover et al. (1998) further discuss the asymptotic properties of the proposed kernel estimator $\hat{\boldsymbol{\beta}}^{\mathrm{K}}(\cdot)$, then find that the asymptotic bias is influenced by the smoothness of $\mathbf{X}(\cdot)$ and $\boldsymbol{\beta}(\cdot)$, and the density of time points, while the asymptotic variance is affected by n_i's and within-subject covariances.

However, both estimators in Hoover et al. (1998) have drawbacks: the local weighted estimator $\hat{\boldsymbol{\beta}}^{\mathrm{LWP}}(\cdot)$ may lack smoothness since only one smoothing parameter h is introduced, while the computation of the smoothing spline estimator $\hat{\boldsymbol{\beta}}^{\mathrm{SS}}(\cdot)$ is often intensive. Huang et al. (2002) thus propose a less computationally expensive one-step procedure for estimating $\boldsymbol{\beta}(\cdot)$. Each $\beta_k(\cdot)$ is approximated by

a different set of basis functions, that is,

$$\beta_k(\cdot) \approx \sum_{v=1}^{V_k} (\gamma_{kv} B_{kv}(\cdot)) = \mathbf{B}_k^T(\cdot)\boldsymbol{\gamma}_k, \tag{2.5}$$

where $\mathbf{B}_k(\cdot) = (B_{k1}(\cdot), \ldots, B_{kV_k}(\cdot))^T$ and $\boldsymbol{\gamma}_k = (\gamma_{k1}, \ldots, \gamma_{kV_k})^T$, and V_k's are regarded as smoothing parameters and selected by CV. For a set of non-negative weights w_i's satisfying $\sum_{i=1}^{n}(n_i w_i) = 1$, the proposed method achieves the least squares basis estimator $\hat{\boldsymbol{\beta}}(\cdot)$ by minimizing

$$\sum_{i=1}^{n}\sum_{j=1}^{n_i}[w_i\{Y_{ij} - \sum_{k=1}^{p}\sum_{v=1}^{V_k}(X_{ijk}\gamma_{kv}B_{kv}(t_{ij}))\}^2],$$

with respect to $\boldsymbol{\gamma} = (\gamma_{11}, \ldots, \gamma_{1V_1}, \ldots, \gamma_{p1}, \ldots, \gamma_{pV_p})^T$, which leads to

$$\hat{\boldsymbol{\gamma}} = \{\sum_{i=1}^{n}(\mathbf{U}_i^T\mathbf{W}_i\mathbf{U}_i)\}^{-1}\sum_{i=1}^{n}(\mathbf{U}_i^T\mathbf{W}_i\mathbf{Y}_i),$$

where $\mathbf{W}_i = w_i\mathbf{I}_{n_i\times n_i}$, and $\mathbf{U}_i$ is i-th block of the design matrix for $\boldsymbol{\gamma}$ and calculated from X_{ijk}'s and $B_{kv}(t_{ij})$'s. The conditional bias and variance of $\hat{\boldsymbol{\beta}}(\cdot)$ are then derived from the expression of $\hat{\boldsymbol{\gamma}}$. In addition, the consistency and the convergence rates of the estimator with general choices of bases are established.

Fan and Zhang (2000), alternatively, propose a two-step procedure for estimating the coefficient functions. The first step obtains a raw estimate of $\boldsymbol{\beta}(\cdot)$ at each time point t_l^0 of $\mathbf{t}^0$, from $\tilde{\mathbf{X}}_l$ and $\tilde{\mathbf{Y}}_l$ that stack over the collection of $\{(Y_{ij}, \mathbf{X}_{ij})\}$ with $t_{ij} = t_l^0$, that is,

$$\check{\boldsymbol{\beta}}(t_l^0) = (\tilde{\mathbf{X}}_l^T\tilde{\mathbf{X}}_l)^{-1}\tilde{\mathbf{X}}_l^T\tilde{\mathbf{Y}}_l.$$

Several special treatments for t_l^0 with rank deficiency of $\tilde{\mathbf{X}}_l$ are then discussed in detail. Considering that the raw estimate has several drawbacks such as under-

smoothness and inefficiency, a second step is proposed to refine $\check{\boldsymbol{\beta}}(t_l^0)$ by smoothing it over time, which gives a linear estimate of q-th derivative $\beta_k^{(q)}(t_0)$ as

$$\hat{\beta}_k^{(q)}(t_0) = \sum_{l=1}^{r} (w_{qk}(t_l^0, t_0)\check{\beta}_k(t_l^0)).$$

In particular, a local d-degree polynomial fitting constructs the weight vector $\mathbf{w}_q(t_l^0, t_0) = (w_{q1}(t_l^0, t_0), \ldots, w_{qp}(t_l^0, t_0))^T$ as

$$\mathbf{w}_q(t_l^0, t_0) = q!\mathbf{e}_{(q+1)(d+1)}^T(\mathbf{C}^T\mathbf{W}\mathbf{C})^{-1}\mathbf{C}_l\mathbf{W}_l,$$

where $\mathbf{W}$ is the kernel weight matrix at t_0, $\mathbf{C}$ is the local polynomial design matrix calculated from $\mathbf{t}^0$, and $\mathbf{e}_{ij}$ is the error vector. Fan and Zhang (2000) then derive the asymptotic bias of the local polynomial estimator $\hat{\boldsymbol{\beta}}^{(q)}(t_0)$, as well as its asymptotic variance under various situations of the number of time points in each collection.

Qu and Li (2006) consider model (2.4), and assume that each $\beta_k(\cdot)$ can be approximated by a combination of basis functions similarly to equation (2.5), then propose to estimate $\boldsymbol{\gamma}$ by minimizing a penalized quadratic inference function of

$$\bar{\mathbf{g}}_n^T\bar{\mathbf{C}}_n^{-1}\bar{\mathbf{g}}_n + \lambda\boldsymbol{\gamma}^T\mathbf{D}\boldsymbol{\gamma},$$

where $\bar{\mathbf{g}}_n$ is a vector of elements of quasi-likelihood estimating equation, $\bar{\mathbf{C}}_n$ is a consistent estimator for its covariance matrix, $\mathbf{D}$ is a diagonal matrix penalizing coefficient $\boldsymbol{\gamma}$, and $\lambda \geq 0$ is a smoothing parameter selected by GCV. In fact, the within-subject correlations are incorporated into this estimation through $\mathbf{g}_i$'s, which are used to construct $\bar{\mathbf{g}}_n$ and $\bar{\mathbf{C}}_n$. The strong consistency of the estimator $\hat{\boldsymbol{\gamma}}$, and its $\sqrt{n}$-consistency and asymptotical normality, have been proved under some regularity conditions. Additionally, the estimator is more efficient than the GEE

estimator if the working correlation is misspecified. Qu and Li (2006) also provide a nonparametric goodness-of-fit test and a model selection procedure.

2.4 Semiparametric Longitudinal Data Models

A semiparametric model contains both parametric and nonparametric components, and provides a more advanced technique to study longitudinal data.

Zhang et al. (1998) study a semiparametric stochastic mixed model for Gaussian responses, that is,

$$Y_{ij} = \mathbf{X}_{ij}^T\boldsymbol{\beta} + m(t_{ij}) + \mathbf{Z}_{ij}^T\mathbf{u}_i + S_i(t_{ij}) + \epsilon_{ij},$$

where $\boldsymbol{\beta}$ is a p-dimensional vector of coefficients, $m(\cdot)$ is a twice-differentiable smooth function, $\mathbf{u}_i$'s are independent q-dimensional vectors of random effects, $S_i(\cdot)$'s are independent stochastic processes modeling within-subject serial correlations, and ϵ_{ij}'s are random errors. In addition, $\mathbf{u}_i$ is distributed as $\mathrm{N}(\mathbf{0}, \mathbf{D}(\boldsymbol{\phi}))$, $S_i(\cdot)$ is a mean-zero Gaussian process with covariance function $\mathrm{cov}(S_i(t), S_i(t')) = \gamma(t, t'; \boldsymbol{\xi}, \alpha)$, and ϵ_{ij} is independent and identically distributed as $\mathrm{N}(0, \sigma^2)$. By stacking over j and i, the model becomes

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{N}\mathbf{m} + \mathbf{Z}\mathbf{u} + \mathbf{S} + \boldsymbol{\epsilon},$$

where $\mathbf{N}$ is the incidence matrix, $\boldsymbol{\beta}$, $\mathbf{m}$, $\mathbf{u}$, $\mathbf{S}$ are used to model mean components, and $\boldsymbol{\theta} = (\boldsymbol{\phi}^T, \boldsymbol{\xi}^T, \alpha, \sigma^2)^T$ consists of variance components.

Zhang et al. (1998) apply the smoothing spline technique with NCS to smooth $m(\cdot)$, or equivalently $\mathbf{m}$, and derive the maximum penalized likelihood estimators

for regression coefficient $\boldsymbol{\beta}$ and nonparametric function $\mathbf{m}$ as

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{W}_x\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}_x\mathbf{Y},$$

and

$$\hat{\mathbf{m}} = (\mathbf{N}^T\mathbf{W}_n\mathbf{N} + \lambda\mathbf{K})^{-1}\mathbf{N}^T\mathbf{W}_n\mathbf{Y},$$

where $\lambda \geq 0$ is a smoothing parameter, $\mathbf{K}$ is the nonnegative definite smoothing matrix defined as equation (2.3) of Green and Silverman (1994), and $\mathbf{W}_x$, $\mathbf{W}_n$ are weight matrices for $\boldsymbol{\beta}$ and $\mathbf{m}$ respectively. By calculating their conditional expectations given $\mathbf{Y}_i$'s, estimators for the subject-specific random effects $\mathbf{u}_i$'s and the stochastic processes $\mathbf{S}_i(\mathbf{t})$'s, where $\mathbf{t}$ is an arbitrary vector of time points, are

$$\hat{\mathbf{u}}_i = \mathbf{D}\mathbf{Z}_i^T\mathbf{W}_i(\mathbf{Y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}} - \mathbf{N}_i\hat{\mathbf{m}}),$$

and

$$\hat{\mathbf{S}}_i(\mathbf{t}) = \boldsymbol{\Gamma}_i(\mathbf{t}, \mathbf{t}_i)\mathbf{W}_i(\mathbf{Y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}} - \mathbf{N}_i\hat{\mathbf{m}}),$$

respectively. In addition, both frequentist and Bayesian versions of covariances of $\hat{\boldsymbol{\beta}}$, $\hat{\mathbf{m}}$, $\hat{\mathbf{u}}_i - \mathbf{u}_i$, and $\hat{\mathbf{S}}_i(\mathbf{t}) - \mathbf{S}_i(\mathbf{t})$ are provided. Furthermore, the smoothing parameter $\tau = \lambda^{-1}$ is treated as an extra variance component, then τ and $\boldsymbol{\theta}$ are simultaneously estimated using REML with the Fisher scoring algorithm, while the covariance of $\hat{\boldsymbol{\theta}}$ is obtained from the corresponding block of the inverse of Fisher information matrix.

Fan and Li (2004) consider another semiparametric model

$$Y_{ij} = m(t_{ij}) + \mathbf{X}_{ij}^T\boldsymbol{\beta} + \epsilon(t_{ij}),$$

where $\epsilon(\cdot)$ is a mean-zero stochastic process, and propose two estimation methods along with a variable selection procedure. The difference-based estimation method sorts all observations by the order of t_{ij}'s as

$$Y_{i^*} = m(t_{i^*}) + \mathbf{X}_{i^*}^T \boldsymbol{\beta} + \epsilon(t_{i^*}),$$

for $i^* = 1, \ldots, \sum_{i=1}^{n} n_i$, and achieves $\hat{\boldsymbol{\beta}}_{\mathrm{DBE}}$ by fitting a marginal model of their sequential differences, where a linear approximation is imposed on $m(\cdot)$. On the other hand, considering that the linear estimate of $m(\cdot)$ is linear in $\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}$, the profile least squares estimator is obtained as

$$\hat{\boldsymbol{\beta}}_{\mathrm{PLSE}} = \{\mathbf{X}^T(\mathbf{I}-\mathbf{S})^T\mathbf{W}(\mathbf{I}-\mathbf{S})\mathbf{X}\}^{-1}\mathbf{X}^T(\mathbf{I}-\mathbf{S})^T\mathbf{W}(\mathbf{I}-\mathbf{S})\mathbf{Y},$$

where $\mathbf{S}$ is a smoothing matrix for $\hat{m}(\cdot)$, and $\mathbf{W}$ is a diagonal matrix of weight functions. Such $\hat{\boldsymbol{\beta}}_{\mathrm{PLSE}}$ is proved to be $\sqrt{n}$-consistent under some mild conditions. The nonparametric function $m(\cdot)$ is then estimated by smoothing the partial residuals of $Y_{ij} - \mathbf{X}_{ij}^T\hat{\boldsymbol{\beta}}$ using a local linear approximation. In practice, Fan and Li (2004) suggest to first have an initial $\hat{\boldsymbol{\beta}}_{\mathrm{DBE}}$, next select an appropriate bandwidth h, then obtain $\hat{m}(\cdot)$ and $\hat{\boldsymbol{\beta}}_{\mathrm{PLSE}}$ as estimators.

Fan and Li (2004) also provide a model selection procedure with a penalized least squares approach by eliminating $m(\cdot)$ as a nuisance parameter, that is, to minimize

$$l(\boldsymbol{\beta}) + n\sum_{k=1}^{p} p_{\lambda_k}(|\beta_k|),$$

where $l(\boldsymbol{\beta})$ is a weighted squares function of $\boldsymbol{\beta}$, λ_k's are tuning parameters, and $p_\cdot(\cdot)$'s are penalty functions. Specifically, each penalty function $p_{\lambda_k}(\cdot)$ is locally approximated by quadratic functions, while the tuning parameter $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_k)^T$ is set as $\boldsymbol{\lambda} = \lambda \mathrm{SE}(\hat{\boldsymbol{\beta}})$ and then estimated by GCV. Fan and Li (2004) prove that,

such penalized weighted least squares estimator $\hat{\boldsymbol{\beta}}$ is $\sqrt{n}$-consistent, and has oracle properties (Fan and Li, 2001).

Fan et al. (2007) further extend a semiparametric model with varying coefficients, and propose a semiparametric varying-coefficient partially linear model as

$$Y_{ij} = \mathbf{X}_{ij}^T\boldsymbol{\beta} + \mathbf{Z}_{ij}^T\boldsymbol{\alpha}(t_{ij}) + \epsilon(t_{ij}),$$

where $\boldsymbol{\beta}$ is a p-dimensional parameter vector, $\boldsymbol{\alpha}(\cdot) = (\alpha_1(\cdot), \ldots, \alpha_q(\cdot))^T$ is a q-dimensional vector of coefficient functions, and the error term has a semiparametric structure with $\mathrm{E}(\epsilon(\cdot)) = 0$, $\mathrm{var}(\epsilon(\cdot)) = \sigma^2(\cdot)$ and $\mathrm{corr}(\epsilon(t), \epsilon(t^{'})) = \rho(t, t^{'}; \boldsymbol{\theta})$.

Fan et al. (2007) propose to estimate $\boldsymbol{\alpha}(\cdot)$, $\boldsymbol{\beta}$, $\sigma^2(\cdot)$, and $\boldsymbol{\theta}$ by iterating between $(\hat{\boldsymbol{\alpha}}(\cdot), \hat{\boldsymbol{\beta}})$ and $(\hat{\sigma}^2(\cdot), \hat{\boldsymbol{\theta}})$. An NW kernel estimator $\hat{\sigma}^2(\cdot)$ is obtained from estimated residuals of $\hat{r}_{ij} = Y_{ij} - \mathbf{X}_{ij}^T\hat{\boldsymbol{\beta}} - \mathbf{Z}_{ij}^T\hat{\boldsymbol{\alpha}}(t_{ij})$, with a bandwidth parameter h_{σ^2}. Two methods are proposed to estimate $\boldsymbol{\theta}$: the quasi-likelihood estimator $\hat{\boldsymbol{\theta}}_{\mathrm{QL}}$ maximizes the quasi-likelihood with known $\hat{\boldsymbol{\alpha}}(\cdot)$, $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2(\cdot)$, but has limitations when the correlation structure is misspecified; the minimum generalized variance estimator $\hat{\boldsymbol{\theta}}_{\mathrm{MGV}}$ minimizes the determinant of the estimated covariance matrix of $\hat{\boldsymbol{\beta}}$, so greatly improves the efficiency for estimating $\boldsymbol{\beta}$. The smooth functions $\alpha_k(\cdot)$'s are estimated with the local linear regression technique, that is, by minimizing the sum of weighted squares

$$\sum_{i=1}^{n}\sum_{j=1}^{n_i}\{(Y_{ij} - \mathbf{X}_{ij}^T\hat{\boldsymbol{\beta}} - \sum_{k=1}^{q}[\{a_{k0} + a_{k1}(t_{ij} - t_0)\}Z_{ijk}])^2 K_h(t_{ij} - t_0)\},$$

with respect to (a_{k0}, a_{k1})'s, where h is a bandwidth parameter, then obtaining $\hat{\alpha}_k(t_0) = \hat{a}_{k0}$. Moreover, with such local linear approximation of $\boldsymbol{\alpha}(\cdot)$, the profile

weighted least squares estimator for $\boldsymbol{\beta}$ is achieved as

$$\hat{\boldsymbol{\beta}} = \{\mathbf{X}^T(\mathbf{I}-\mathbf{S})^T\mathbf{W}(\mathbf{I}-\mathbf{S})\mathbf{X}\}^{-1}\mathbf{X}^T(\mathbf{I}-\mathbf{S})^T\mathbf{W}(\mathbf{I}-\mathbf{S})\mathbf{Y},$$

where $\mathbf{W}$ is a weighted working covariance matrix, and $\mathbf{S}$ is a smoothing matrix. It is also proved that, under some regularity conditions, $\hat{\boldsymbol{\beta}}$ is $\sqrt{n}$-consistent, $\hat{\boldsymbol{\alpha}}(\cdot)$ is $\sqrt{nh}$-consistent, and $\hat{\sigma}^2(\cdot)$ is $\sqrt{nh_{\sigma^2}}$-consistent, while $\hat{\boldsymbol{\beta}}$ is asymptotically unbiased but $\hat{\boldsymbol{\alpha}}(\cdot)$ and $\hat{\sigma}^2(\cdot)$ are not.

2.5 Models with Longitudinal Covariates

A more interesting problem arises when the predictor is in fact a function $x_i(\cdot)$ instead of a vector $\mathbf{X}_i$, and $x_i(\cdot)$ is observed at n_i time points as $(W_{i1},\ldots,W_{in_i})$, while the response Y_i is assessed at one single time. The problem thus becomes a model with longitudinal covariates. Typically, in such problems, a longitudinal dataset $\{(Y_i,\mathbf{Z}_i,\mathbf{W}_i)\}$ is collected from n subjects, where Y_i is a scalar response, $\mathbf{Z}_i$ is a p-dimensional vector of time-invariant covariates, and $\mathbf{W}_i=(W_{i1},\ldots,W_{in_i})^T$ is a vector of observed longitudinal covariates of $x_i(\cdot)$ measured at time points t_{ij}'s.

James (2002) develops a functional generalized linear model that directly models the relationship between the response Y_i from some exponential family distribution and the functional predictor $x_i(\cdot)$. In particular, the mean $\mu_i=\mathrm{E}(Y_i)$ is modeled by

$$g(\mu_i)=\alpha+\int x_i(t)\gamma(t)dt,$$

where α is an intercept, and $\gamma(\cdot)$ is a smooth effect function. The observed W_{ij}'s are assumed to relate to the latent profile $x_i(\cdot)$ via

$$W_{ij}=x_i(t_{ij})+\varepsilon(t_{ij}),$$

where $\varepsilon(\cdot)$ is a mean-zero stationary Gaussian process regarded as the deviation of observation from true profile due to measurement error or other factors. James (2002) parameterizes $x_i(\cdot)$ with NCS as

$$x_i(\cdot) = \mathbf{x}_i^T \mathbf{c}(\cdot),$$

and

$$\mathbf{x}_i \sim \mathrm{N}(\boldsymbol{\eta}, \boldsymbol{\Gamma}),$$

where $\mathbf{c}(\cdot)$ is an NCS basis, $\mathbf{x}_i$ is the spline coefficients vector for $x_i(\cdot)$, and $\boldsymbol{\eta}$ and $\boldsymbol{\Gamma}$ are defined as mean and variance of $\mathbf{x}_i$'s. Thus, the model becomes

$$g(\mu_i) = \alpha + \mathbf{x}_i^T \boldsymbol{\beta},$$

where $\boldsymbol{\beta} = \int \gamma(t)\mathbf{c}(t)dt$, and

$$\mathbf{W}_i = \mathbf{C}_i \mathbf{x}_i + \boldsymbol{\varepsilon}_i,$$

$$\boldsymbol{\varepsilon}_i \sim \mathrm{N}(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I}),$$

where $\mathbf{C}_i = (\mathbf{c}(t_{i1}), \ldots, \mathbf{c}(t_{in_i}))^T$.

The unobserved $\mathbf{x}_i$'s are treated as missing data, then $\mathbf{W}_i$'s and Y_i's are conditionally independent, and the likelihood can be factored into three distinct parts for $(\alpha, \boldsymbol{\beta}, \phi)$, σ_ε^2 and $(\boldsymbol{\eta}, \boldsymbol{\Gamma})$, respectively. The EM algorithm (Dempster et al., 1977) is then applied, where the M-step maximizes three parts of likelihood separately to estimate parameters, and the E-step calculates the expected value and variance of $\mathbf{x}_i$'s in each iteration.

Zhang et al. (2007) focus on Gaussian response Y_i, and assume that Y_i is related to an unobserved T-periodic latent profile $x_i(\cdot)$ through a partial functional linear

model

$$Y_i = \alpha + \mathbf{Z}_i^T \boldsymbol{\delta} + \int_0^T x_i(t)\gamma(t)dt + \epsilon_i,$$

where α is an intercept, $\boldsymbol{\delta}$ is a p-dimensional vector of regression coefficients, $\gamma(\cdot)$ is a T-periodic smooth effect function, and ϵ_i's are independent and identically distributed errors with $\mathrm{N}(0, \sigma_\epsilon^2)$. The observed longitudinal covariates W_{ij}'s, on the other hand, are related to $x_i(\cdot)$ by an additive measurement error model as

$$W_{ij} = x_i(t_{ij}) + \varepsilon_{ij}, \tag{2.6}$$

where ε_{ij}'s are independent and identically distributed measurement errors with $\mathrm{N}(0, \sigma_\varepsilon^2)$.

Zhang et al. (2007) propose a two-stage nonparametric regression calibration (TSNRC) method to estimate coefficients α, $\boldsymbol{\delta}$, $\gamma(\cdot)$, and variance components σ_ϵ^2, σ_ε^2. In particular, a smoothing spline technique with a periodic cubic spline is applied to smooth $\hat{x}_i(\cdot)$'s and $\hat{\gamma}(\cdot)$. In the first stage, the TSNRC method decomposes the subject-specific profile as $x_i(\cdot) = x_0(\cdot) + g_i(\cdot)$, where $x_0(\cdot)$ is the periodic population profile and $g_i(\cdot)$ is a random subject-specific deviation from the population profile, then estimates $x_0(\cdot)$ and $g_i(\cdot)$ by maximizing the penalized likelihood function following Zhang et al. (1998). With $\hat{x}_i(\cdot) = \hat{x}_0(\cdot) + \hat{g}_i(\cdot)$, the TSNRC method next jointly estimates $\gamma(\cdot)$ along with α and $\boldsymbol{\delta}$ by maximizing the penalized pseudo-likelihood function in the same fashion as that in the first stage. This method also provides a score test for testing $\gamma(\cdot)$ as a constant versus a smooth function.

Bhadra et al. (2012) propose a Bayesian semiparametric approach for continuous W_{ij}'s and binary Y_i's. In a hierarchical Bayesian framework, the joint likelihood formulation would properly account for uncertainties associated with both stages

of the estimation process. In particular, the observed longitudinal covariates are considered through an additive measurement error model as equation (2.6), the binary response is then assumed to be of the form

$$\text{logit}(\mu_i) = \alpha + a_i^d \delta + \int_{-c_1}^{-c_2} x_i(t + a_i^d)\gamma(t)dt,$$

where a_i^d is a separate covariate, and $\mu_i = \text{P}(Y_i = 1)$.

For the nonparametric functions, Bhadra et al. (2012) represent $x_0(\cdot)$, $g_i(\cdot)$'s and $\gamma(\cdot)$ using regression splines, where $x_0(\cdot)$ and $g_i(\cdot)$'s follow the same decomposition of $x_i(\cdot) = x_0(\cdot) + g_i(\cdot)$ as that in Zhang et al. (2007). In addition to $\boldsymbol{\Omega}$, the set of unknown parameters corresponding to the covariate and response models, Bhadra et al. (2012) regard the numbers (k_1, k_2) and locations ($\boldsymbol{\tau}$, $\boldsymbol{\xi}$) of knots used in splines as extra parameters to be estimated. With another set of hyperparameters $\boldsymbol{\Omega}_h$, Bhadra et al. (2012) assume prior distributions on $\boldsymbol{\Omega}$, then specify the joint posterior distribution $p(\boldsymbol{\Omega}, k_1, k_2, \boldsymbol{\tau}, \boldsymbol{\xi}, \boldsymbol{\Omega}_h | \mathbf{Y}, \mathbf{W})$ for the parameters as well as the numbers and locations of the knots. A reversible jump Markov chain Monte Carlo algorithm (Green, 1995) is then adopted to simultaneously sample the parameters and knots in an integrated manner from their respective full conditionals.

2.6 Models for Zero-Inflated Count Data

Typically, count-valued response data are modeled using discrete distributions, such as Poisson, binomial, and negative binomial distributions. However, in psychology and sociology studies, it is common to encounter data where the proportion of observations with zero is often much larger than what standard distributions could allow. In such situations, the data are considered as zero inflated relative to ordinary count distributions.

A two-part model is usually adopted for zero-inflated data, in order to address both the abundance of zeros and the distribution of non-zero counts. While various two-part models have been developed, hurdle model (Mullahy, 1986) and zero-inflated model (Lambert, 1992) are most commonly used.

The zero-inflated model assumes that the data are generated from a mixture of a degenerate zero distribution and an ordinary count distribution F, then the probability function of response Y becomes

$$\begin{aligned} \mathrm{P}(Y = 0) &= (1 - \pi) + \pi f(0), \\ \mathrm{P}(Y = y) &= \pi f(y), \text{ for } y > 0, \end{aligned}$$

where $f(\cdot)$ is the probability function of the count distribution F.

The hurdle model, on the other hand, is a mixture of a degenerate zero distribution and a zero-truncated count distribution F^+, and gives the probability function as

$$\begin{aligned} \mathrm{P}(Y = 0) &= 1 - \pi, \\ \mathrm{P}(Y = y) &= \pi \frac{f(y)}{1 - f(0)}, \text{ for } y > 0, \end{aligned}$$

where $f(\cdot)$ is the probability function of the corresponding untruncated distribution F.

One advantage of the hurdle model over the zero-inflated model is that the former one can handle not only zero-inflated data but also zero-deflation, where the proportion of zeros is less than that a standard count distribution would predict. However, this advantage is negligible because zero-deflated data are extremely rare in practice (Buu et al., 2012a). When only accounting for zero-inflated data, the two models are mathematically equivalent, for one is a reparameterization of the other. The hurdle model is more appropriate when all subjects are at risk of an event and the realization of the event represents a hurdle, whereas the zero-inflated

model is more intuitive when the population consists of a group at risk and another group out of risk (Rose et al., 2006).

As for the estimation issue, if there are no random effects, two components have to be fitted simultaneously in the zero-inflated model, but can be estimated separately for the hurdle model, which brings computational simplicity for the latter one. Nevertheless, when random effects are included in both components and are correlated, the two components must be fitted simultaneously for both models.

Chapter 3

Modeling Scalar Response with Discrete Longitudinal Covariates: Method and Application

Many considerations are needed to handle situations when the covariate is repeatedly measured at multiple time points but the response is collected at one single time point for each subject. In the setting of such model with longitudinal covariates, standard longitudinal methods designed for longitudinal responses are not applicable. In practice, the developmental pattern from these measures of covariates is often highly predictive of the distal outcome, but the effect might be through a complex function of time. Additionally, the observed longitudinal covariates are often subject to measurement errors, while the true profiles are unobservable.

James (2002) develops a functional generalized linear model, Zhang et al. (2007) study a two-stage functional mixed model, and Bhadra et al. (2012) propose a Bayesian semiparametric approach; however, all of the methods focus on longitudinal covariates of continuous values. A methodology issue comes when the

longitudinal covariates are discrete values, thus the model needs to be fitted under a generalized linear model framework. From the view of regression calibration, we propose a two-stage estimation procedure to accommodate situations that either or both of longitudinal covariates and responses are discrete, while allowing observation time points of longitudinal covariates to be unbalanced among subjects.

This chapter is organized as follows. In Section 3.1, we introduce our model and its corresponding notation, as well as the natural cubic smoothing spline technique that will be used to approximate nonparametric functions in the estimation. Section 3.2 describes a two-stage estimation procedure in detail. A series of simulation studies to compare estimation performance under different circumstances are provided in Section 3.3. Section 3.4 slightly extends the model to accommodate an extra clustered random effect, and illustrates it with an empirical study on alcoholic couples data.

3.1 Preliminaries

3.1.1 Model and Notation

Suppose that a longitudinal dataset $\{(Y_i, \mathbf{Z}_i, \mathbf{W}_i)\}$ is collected from n subjects. For i-th subject, we have a response variable Y_i, a p-dimensional vector $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{ip})^T$ of time-invariant covariates, and an observed longitudinal covariate process $\mathbf{W}_i = (W_{i1}, \ldots, W_{in_i})^T$ of count-valued W_{ij}'s, which, in addition, are usually assumed as measurements of a latent profile function $x_i(\cdot)$ at time points $\{t_{ij}\} \in [T_1, T_2]$.

A typical example of such datasets is from a study on youth at high risk for alcohol abuse, which is a part of the Michigan Longitudinal Study. Table 3.1 shows records of a few subjects from the study, where Y_i is a binary response of the DSM-

Table 3.1: Example Data from a Youth Alcohol Abuse Study

i	Y_i	$\mathbf{Z}_i$		$\mathbf{W}_i$							
		1	2	13	14	15	16	17	18	19	20
4	0	1	1			2				4	
11	0	0	0	0	0	1	0	1	0		
14	0	0	1	0	0	1	1	4		6	4
15	0	0	1	0	0	0	0	10	1	6	
18	0	1	0	0	0	0	0	0	0		1
38	1	1	1		0	0	0	4	2		10
70	0	0	0		0		0	0			

IV alcohol dependence diagnosis (American Psychiatric Association, 1994) during early adulthood, $\mathbf{Z}_i$ consists of two binary time-invariant covariates, gender (Z_{i1}) and parental lifetime alcohol use disorder diagnosis (Z_{i2}), and $\mathbf{W}_i$ is an observed longitudinal covariate process of the number of drinking days in one typical month during ages of 13 - 20.

The response variable Y_i is assumed to relate to the latent profile $x_i(\cdot)$ and the time-invariant covariate $\mathbf{Z}_i$ through a generalized functional linear model (Muller and Stadtmuller, 2005), that is,

$$Y_i \sim f_Y(\cdot; \eta_i), \tag{3.1}$$

and

$$g(\eta_i) = \mathbf{Z}_i^T \boldsymbol{\delta} + \int_{T_1}^{T_2} x_i(t)\gamma(t)dt, \tag{3.2}$$

where $f_Y(\cdot)$ is a distribution belonging to an exponential family, $g(\cdot)$ is a known link function, $\boldsymbol{\delta}$ is a p-dimensional vector of regression coefficients, and $\gamma(\cdot)$ is a smooth effect function for the latent profile $x_i(\cdot)$. In particular, if Y_i is Gaussian with the identity link function, model (3.1,3.2) is equivalent to a partial functional

linear model (Ramsay and Silverman, 2005)

$$Y_i = \mathbf{Z}_i^T \boldsymbol{\delta} + \int_{T_1}^{T_2} x_i(t)\gamma(t)dt + \epsilon_i, \tag{3.3}$$

where ϵ_i's are assumed to be independent and identically distributed as $\mathrm{N}(0, \sigma_\epsilon^2)$.

On the other hand, the observed longitudinal covariate process $\mathbf{W}_i$ is related to the latent profile $x_i(\cdot)$ by

$$W_{ij} \sim f_W(\cdot; \mu_{ij}), \tag{3.4}$$

and

$$h(\mu_{ij}) = x_i(t_{ij}), \tag{3.5}$$

where $f_W(\cdot)$ is another distribution of exponential family, $h(\cdot)$ is a known link function and not necessarily same as $g(\cdot)$. Our model is focused on discrete-valued W_{ij}'s; however, for continuous W_{ij}'s with the identity link function, the model is equivalent to an additive measurement error model of

$$W_{ij} = x_i(t_{ij}) + \varepsilon_{ij}, \tag{3.6}$$

where ε_{ij}'s are measurement errors assumed to be independent and identically distributed as $\mathrm{N}(0, \sigma_\varepsilon^2)$.

The proposed approach is to estimate the regression coefficient $\boldsymbol{\delta}$ and the effect function $\gamma(\cdot)$, as well as the variance component σ_ϵ^2 if applicable, from the data $\{(Y_i, \mathbf{Z}_i, \mathbf{W}_i)\}$.

3.1.2 Natural Cubic Spline and Smoothing Spline Technique

In the model, both the subject-specific latent profiles $x_i(\cdot)$'s and the effect function $\gamma(\cdot)$ are nonparametric and infinite dimensional, thus we propose to approximate them by a spline method.

The spline smoothing method allows us to express each function by a linear combination of a set of basis functions, that is, $x_i(\cdot) = \dot{\mathbf{x}}_i^T \mathbf{B}^x(\cdot)$ and $\gamma(\cdot) = \dot{\boldsymbol{\gamma}}^T \mathbf{B}^\gamma(\cdot)$, where $\mathbf{B}^x(\cdot) = (B_1^x(\cdot), \ldots, B_{K_x}^x(\cdot))^T$ and $\mathbf{B}^\gamma(\cdot) = (B_1^\gamma(\cdot), \ldots, B_{K_\gamma}^\gamma(\cdot))^T$ are not necessarily same. Thus, the functional predictor term in equation (3.2) can be equivalently written as

$$\int_{T_1}^{T_2} x_i(t)\gamma(t)dt = \dot{\mathbf{x}}_i^T \mathbf{B} \dot{\boldsymbol{\gamma}},$$

where $\mathbf{B}$ is a $K_x \times K_\gamma$ matrix with $\mathbf{B}_{(i,j)} = \int_{T_1}^{T_2} B_i^x(t) B_j^\gamma(t)dt$.

In particular, we propose to adopt the natural cubic spline (NCS; Green and Silverman, 1994) in our estimation procedure. Specifically, all NCS functions in $[T_1, T_2]$ with knots of $\mathbf{t}^0$, where $\mathbf{t}^0 = (t_1^0, \ldots, t_r^0)^T$ is an r-dimensional vector of ordered distinct values of all time points t_{ij}'s, span a function space $\mathcal{S}_{\text{NCS}}(\mathbf{t}^0)$. For any function $s(\cdot) \in \mathcal{S}_{\text{NCS}}(\mathbf{t}^0)$, there exists a set of r piecewise cubic polynomial basis functions $\mathbf{c}(\cdot) = (c_1(\cdot), \ldots, c_r(\cdot))^T$ such that $s(\cdot) = \mathbf{s}^T \mathbf{c}(\cdot)$, where $\mathbf{s} = s(\mathbf{t}^0)$ is an r-dimensional vector formed by evaluating $s(\cdot)$ at $\mathbf{t}^0$. Thus, $\mathbf{s} \to s(\cdot)$ is a 1-1 mapping from $\mathbb{R}^r$ to $\mathcal{S}_{\text{NCS}}(\mathbf{t}^0)$. Particularly, each $c_l(\cdot)$ itself is an NCS function satisfying $c_l(\mathbf{t}^0) = \mathbf{e}_l$, where $\mathbf{e}_l$ is an r-dimensional vector with 1 in l-th element and 0 elsewhere. Hence, by assuming that $x_i(\cdot)$'s and $\gamma(\cdot)$ belong to the function space $\mathcal{S}_{\text{NCS}}(\mathbf{t}^0)$, we have $x_i(\cdot) = \mathbf{x}_i^T \mathbf{c}(\cdot)$ and $\gamma(\cdot) = \boldsymbol{\gamma}^T \mathbf{c}(\cdot)$, then equation (3.2) becomes

$$g(\eta_i) = \mathbf{Z}_i^T \boldsymbol{\delta} + \mathbf{x}_i^T \mathbf{C} \boldsymbol{\gamma}, \tag{3.7}$$

where $\mathbf{C} = \int_{T_1}^{T_2} \mathbf{c}(t)\mathbf{c}^T(t)dt$ is an $r \times r$ matrix, whose element can be calculated straightforwardly.

In order to smooth the function $s(\cdot)$, the smoothing spline technique introduces a roughness penalty term of

$$J_s = \frac{1}{2}\lambda \int_{T_1}^{T_2} (s''(t))^2 dt,$$

where $\lambda \geq 0$ is a smoothing parameter controling the goodness-of-fit of the model and the smoothness of $s(\cdot)$. In general, the linear combination expression of $s(\cdot) = \dot{\mathbf{s}}^T \mathbf{B}^s(\cdot)$ leads to

$$\int_{T_1}^{T_2} (s''(t))^2 dt = \dot{\mathbf{s}}^T \mathbf{K}^s \dot{\mathbf{s}},$$

where $\mathbf{K}^s$ is a $K_s \times K_s$ matrix with $\mathbf{K}^s_{(i,j)} = \int_{T_1}^{T_2} (B_i^s)''(t)(B_j^s)''(t)dt$. For the particular condition of $s(\cdot) \in \mathcal{S}_{\mathrm{NCS}}(\mathbf{t}^0)$, equation (2.3) in Green and Silverman (1994) specifies an $r \times r$ symmetric matrix $\mathbf{K}$ that only depends on $\mathbf{t}^0$ and has rank of $r - 2$.

An equivalent expression of the roughness penalty term can be derived from a view of linear mixed model. Explicitly, following Green (1987) and Lin and Zhang (1999), we denote $\mathbf{T} = (\mathbf{1}_r, \mathbf{t}^0)$ as the non-trivial null space of $\mathbf{K}$, and decompose $\mathbf{K}$ as $\mathbf{K} = \mathbf{L}\mathbf{L}^T$, where $\mathbf{L}$ is an $r \times (r-2)$ full-rank matrix satisfying $\mathbf{L}^T\mathbf{T} = \mathbf{O}$. With $\mathbf{B} = \mathbf{L}(\mathbf{L}^T\mathbf{L})^{-1}$, the NCS vector $\mathbf{s}$ can be represented as

$$\mathbf{s} = \mathbf{T}\boldsymbol{\alpha}_s + \mathbf{B}\mathbf{a}_s, \tag{3.8}$$

where $\boldsymbol{\alpha}_s$ is a 2-dimensional vector regarded as a fixed effect, and $\mathbf{a}_s$ is an $(r-2)$-dimensional random vector following $\mathrm{N}(\mathbf{0}, \lambda^{-1}\mathbf{I})$. Hence, a nonparametric or semiparametric model with $s(\cdot)$ can be considered as a modified mixed effects model with the decomposition of $\mathbf{s}$ as equation (3.8), and the roughness penalty is

indeed the corresponding term in the log-likelihood of the modified mixed effects model, that is,

$$J_s = \frac{1}{2}\lambda \mathbf{s}^T \mathbf{K}\mathbf{s} = \frac{1}{2}\lambda \mathbf{a}_s^T \mathbf{a}_s.$$

3.2 Two-Stage Estimation Procedure

In this section, we propose a two-stage procedure to estimate the regression coefficient $\boldsymbol{\delta}$ and the effect function $\gamma(\cdot)$: first, subject-specific $\hat{x}_i(\cdot)$'s, or equivalently their NCS vectors $\hat{\mathbf{x}}_i$'s, are obtained under model (3.4,3.5); next, we estimate $\boldsymbol{\delta}$ and $\gamma(\cdot)$ by fitting a calibration model

$$g(\eta_i) = \mathbf{Z}_i^T \boldsymbol{\delta} + \hat{\mathbf{x}}_i^T \mathbf{C}\boldsymbol{\gamma}, \tag{3.9}$$

or

$$g(\boldsymbol{\eta}) = \mathbf{Z}\boldsymbol{\delta} + \hat{\mathbf{X}}\boldsymbol{\gamma}, \tag{3.10}$$

where $g(\boldsymbol{\eta}) = (g(\eta_1), \ldots, g(\eta_n))^T$, $\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_n)^T$ and $\hat{\mathbf{X}} = (\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_n)^T\mathbf{C}$ are stacked over i.

3.2.1 Stage-I

In practice, it is usually the case that all subjects follow a general trend over time, while each individual has a deviation from the population profile that can be considered as random. From this point of view, the individual profile $x_i(\cdot)$ is decomposed as

$$x_i(\cdot) = x_0(\cdot) + d_i(\cdot), \tag{3.11}$$

where $x_0(\cdot)$ is the population profile, and $d_i(\cdot)$ is the random deviation of an individual profile from the population profile. We assume that $x_0(\cdot)$ and $d_i(\cdot)$'s are

all NCS functions, and $d_i(\cdot)$'s are in addition independent mean-zero Gaussian processes. By this decomposition, equation (3.5) can be written as

$$h(\mu_{ij}) = x_0(t_{ij}) + d_i(t_{ij}),$$

or equivalently, by stacking over j,

$$h(\boldsymbol{\mu}_i) = \mathbf{N}_i\mathbf{x}_0 + \mathbf{N}_i\mathbf{d}_i,$$

where $h(\boldsymbol{\mu}_i) = (h(\mu_{i1}), \ldots, h(\mu_{in_i}))^T$, $\mathbf{x}_0 = x_0(\mathbf{t}^0)$, $\mathbf{d}_i = d_i(\mathbf{t}^0)$, and $\mathbf{N}_i$ is an $n_i \times r$ incidence matrix mapping $(t_{i1}, \ldots, t_{in_i})^T$ to $\mathbf{t}^0$ such that (j, l)-th element is 1 if $t_{ij} = t_l^0$ and 0 otherwise.

In order to account for the smoothness of the random Gaussian processes $d_i(\cdot)$'s, we take a transformation of $\mathbf{d}_i = \mathbf{B}_*\mathbf{b}_i$. Here, $\mathbf{B}_* = (\mathbf{T}, \mathbf{B})$ is an $r \times r$ matrix with $\mathbf{T}$, $\mathbf{B}$ defined in Section 3.1.2, and $\mathbf{b}_i$'s are r-dimensional random vectors independently distributed as

$$\mathbf{b}_i \sim \mathrm{N}(\mathbf{0}, \begin{pmatrix} \sigma_1^2 & & \\ & \sigma_2^2 & \\ & & \tau_d\mathbf{I}_{(r-2)\times(r-2)} \end{pmatrix}).$$

The covariance matrix is then denoted as an $r \times r$ matrix of $\mathbf{D}(\boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\sigma_1^2, \sigma_2^2, \tau_d)^T$ is a vector of variance components. With such transformation, the smoothness of $d_i(\cdot)$ is characterized by a parameter τ_d along with the matrix $\mathbf{B}$.

From the constructions of $\mathbf{T}$, $\mathbf{L}$ and $\mathbf{B}$ in Section 3.1.2, one can show that, $\mathbf{B}_*$ is a full-rank matrix, then $\mathbf{b}_i \to \mathbf{d}_i$ is a 1-1 transformation, which implies that we may focus on estimation of $\mathbf{b}_i$'s instead of $\mathbf{d}_i$'s, that is,

$$h(\boldsymbol{\mu}_i) = \mathbf{N}_i\mathbf{x}_0 + \mathbf{Q}_i\mathbf{b}_i,$$

where $\mathbf{Q}_i = \mathbf{N}_i\mathbf{B}_*$. Furthermore, by denoting $h(\boldsymbol{\mu}) = (h(\boldsymbol{\mu}_1)^T, \ldots, h(\boldsymbol{\mu}_n)^T)^T$, $\mathbf{N} = (\mathbf{N}_1^T, \ldots, \mathbf{N}_n^T)^T$ and $\mathbf{Q} = \text{diag}\{\mathbf{Q}_1, \ldots, \mathbf{Q}_n\}$, we have a generalized additive mixed model

$$h(\boldsymbol{\mu}) = \mathbf{N}\mathbf{x}_0 + \mathbf{Q}\mathbf{b}, \tag{3.12}$$

where $\mathbf{b} = (\mathbf{b}_1^T, \ldots, \mathbf{b}_n^T)^T \sim \text{N}(\mathbf{0}, \mathcal{D}(\boldsymbol{\theta}))$ with $\mathcal{D} = \text{diag}\{\mathbf{D}, \ldots, \mathbf{D}\}$.

We will provide details of the estimation for the profile functions and variance components in Sections 3.2.1.1 and 3.2.1.2, respectively. However, one may refer to Section 3.2.1.3 for an overview of the Stage-I estimation procedure.

3.2.1.1 Estimation of Profile Functions

In order to estimate $\mathbf{x}_0$ and $\mathbf{b}$ in equation (3.12), the logarithm of the quasi-likelihood (Wedderburn, 1974) is derived from model (3.4,3.12) as

$$ql_I = \log(\int \exp(-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\tilde{d}_{ij} - \frac{1}{2}\mathbf{b}^T\mathcal{D}^{-1}\mathbf{b})d\mathbf{b}),$$

and $\tilde{d}_{ij}$ is the conditional deviance such that

$$\tilde{d}_{ij} = -2\int_{W_{ij}}^{\mu_{ij}} \frac{w_{ij}(W_{ij} - s)}{\phi_W v_W(s)}ds, \tag{3.13}$$

where ϕ_W is a specified dispersion parameter, $v_W(\cdot)$ is the variance function determined by the distribution $f_W(\cdot)$, and w_{ij}'s are known prior weights.

A smoothing spline technique is applied in estimating $x_0(\cdot)$ by including a roughness penalty term $\frac{1}{2}\lambda_x\int_{T_1}^{T_2}(x_0''(t))^2dt$ to the log-quasi-likelihood ql_I, where λ_x is a smoothing parameter for $x_0(\cdot)$. In addition, the Laplace approximation is adopted to avoid complicated numerical integrations. Following Lin and Zhang (1999), the estimates $\hat{\mathbf{x}}_0$ and $\hat{\mathbf{b}}$ maximize the double penalized quasi-likelihood

(DPQL) of

$$-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\tilde{d}_{ij}-\frac{1}{2}\mathbf{b}^T\mathcal{D}^{-1}\mathbf{b}-\frac{1}{2}\lambda_x\mathbf{x}_0^T\mathbf{K}\mathbf{x}_0, \tag{3.14}$$

for given λ_x and $\boldsymbol{\theta}$. Explicitly, $\hat{\mathbf{x}}_0$ and $\hat{\mathbf{b}}$ solve the estimating equations of

$$\mathbf{N}^T\mathcal{W}\boldsymbol{\Delta}(\mathbf{W}-\boldsymbol{\mu})-\lambda_x\mathbf{K}\mathbf{x}_0=\mathbf{0},$$
$$\mathbf{Q}^T\mathcal{W}\boldsymbol{\Delta}(\mathbf{W}-\boldsymbol{\mu})-\mathcal{D}^{-1}\mathbf{b}=\mathbf{0},$$

where $\mathbf{W}=(\mathbf{W}_1^T,\ldots,\mathbf{W}_n^T)^T$, $\boldsymbol{\Delta}=\mathrm{diag}\{\boldsymbol{\Delta}_1,\ldots,\boldsymbol{\Delta}_n\}$ with $\boldsymbol{\Delta}_i=\mathrm{diag}\{h^{'}(\mu_{ij})\}$, and $\mathcal{W}=\mathrm{diag}\{\mathcal{W}_1,\ldots,\mathcal{W}_n\}$ with $\mathcal{W}_i=\mathrm{diag}\{\frac{w_{ij}}{\phi_W v_W(\mu_{ij})(h^{'}(\mu_{ij}))^2}\}$. Considering the mutual independence of $\mathbf{b}_i$'s, the estimating equation for $\mathbf{b}$ can be separated as

$$\mathbf{Q}_i^T\mathcal{W}_i\boldsymbol{\Delta}_i(\mathbf{W}_i-\boldsymbol{\mu}_i)-\mathbf{D}^{-1}\mathbf{b}_i=\mathbf{0},$$

for $i=1,\ldots,n$, which will facilitate the computation.

As discussed in Section 3.1.2, the nonparametric function $x_0(\cdot)$ can equivalently take the form of

$$\mathbf{x}_0=\mathbf{T}\boldsymbol{\alpha}_x+\mathbf{B}\mathbf{a}_x, \tag{3.15}$$

with $\mathbf{a}_x\sim\mathrm{N}(\mathbf{0},\lambda_x^{-1}\mathbf{I})$. This representation then suggests a generalized linear mixed model (GLMM) of

$$h(\boldsymbol{\mu})=\mathbf{N}\mathbf{T}\boldsymbol{\alpha}_x+\mathbf{N}\mathbf{B}\mathbf{a}_x+\mathbf{Q}\mathbf{b}. \tag{3.16}$$

Similarly, with the Laplace approximation, the log-quasi-likelihood can be derived from this model as

$$-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\tilde{d}_{ij}-\frac{1}{2}\mathbf{b}^T\mathcal{D}^{-1}\mathbf{b}-\frac{1}{2}\lambda_x\mathbf{a}_x^T\mathbf{a}_x,$$

which coincides to the DPQL of equation (3.14).

For a further step, the solutions to the estimating equations, $\hat{\boldsymbol{\alpha}}_x$, $\hat{\mathbf{a}}_x$ and $\hat{\mathbf{b}}$, correspond to the estimates under a linear mixed model (LMM) of

$$\tilde{\mathbf{W}} = \mathbf{NT}\boldsymbol{\alpha}_x + \mathbf{NB}\mathbf{a}_x + \mathbf{Q}\mathbf{b} + \mathbf{e}_\varepsilon, \tag{3.17}$$

where $\tilde{\mathbf{W}} = \mathbf{N}\mathbf{x}_0 + \mathbf{Q}\mathbf{b} + \boldsymbol{\Delta}(\mathbf{W} - \boldsymbol{\mu})$ is a working vector, and $\mathbf{e}_\varepsilon \sim \mathrm{N}(\mathbf{0}, \mathcal{W}^{-1})$. This representation allows us to compute the covariance as

$$\mathrm{Cov}((\hat{\boldsymbol{\alpha}}_x^T, \hat{\mathbf{a}}_x^T)^T) = \mathbf{H}^{-1}\mathbf{H}_0\mathbf{H}^{-1},$$

where $\mathbf{R} = \mathbf{Q}\mathcal{D}\mathbf{Q}^T + \mathcal{W}^{-1}$, $\mathbf{H}_0 = (\mathbf{NB}_*)^T\mathbf{R}^{-1}(\mathbf{NB}_*)$ and

$$\mathbf{H} = \begin{pmatrix} (\mathbf{NT})^T\mathbf{R}^{-1}(\mathbf{NT}) & (\mathbf{NT})^T\mathbf{R}^{-1}(\mathbf{NB}) \\ (\mathbf{NB})^T\mathbf{R}^{-1}(\mathbf{NT}) & (\mathbf{NB})^T\mathbf{R}^{-1}(\mathbf{NB}) + \lambda_x\mathbf{I} \end{pmatrix}.$$

Thus, the approximate covariance of $\hat{\mathbf{x}}_0$ is derived as

$$\mathrm{Cov}(\hat{\mathbf{x}}_0) = \mathbf{B}_*\mathrm{Cov}((\hat{\boldsymbol{\alpha}}_x^T, \hat{\mathbf{a}}_x^T)^T)\mathbf{B}_*^T. \tag{3.18}$$

3.2.1.2 Estimation of Variance Components

In estimating $x_0(\cdot)$ and $d_i(\cdot)$'s, it is assumed that smoothing parameter λ_x and variance component $\boldsymbol{\theta}$ are given, whilst both are usually unknown but need to be estimated as well. In particular, we estimate $\boldsymbol{\theta}$ by maximizing the corresponding marginal quasi-likelihood (Lin and Zhang, 1999), and select λ_x by generalized cross validation (GCV).

Following Lin and Zhang (1999), if $x_0(\cdot)$ is represented as equation (3.15), with $\boldsymbol{\alpha}_x$ having a uniform prior distribution and $\mathbf{a}_x \sim \mathrm{N}(\mathbf{0}, \lambda_x^{-1}\mathbf{I})$, the marginal quasi-

likelihood is

$$ql_{\mathrm{MI}} = -\frac{1}{2}\log(|\mathcal{D}|) + \frac{r-2}{2}\log(\lambda_x) + \log(\int\int\int \exp(-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\tilde{d}_{ij} - \frac{1}{2}\mathbf{b}^T\mathcal{D}^{-1}\mathbf{b} - \frac{1}{2}\lambda_x\mathbf{a}_x^T\mathbf{a}_x)d\mathbf{b}d\mathbf{a}_x d\boldsymbol{\alpha}_x),$$

which, using the Laplace method, is approximated as

$$\begin{aligned} ql_{\mathrm{MI}} \approx & -\frac{1}{2}\log|\mathcal{V}| - \frac{1}{2}\log|(\mathbf{NT})^T\mathcal{V}^{-1}(\mathbf{NT})| \\ & -\frac{1}{2}\{\tilde{\mathbf{W}} - (\mathbf{NT})\hat{\boldsymbol{\alpha}}_x\}^T\mathcal{V}^{-1}\{\tilde{\mathbf{W}} - (\mathbf{NT})\hat{\boldsymbol{\alpha}}_x\}, \end{aligned} \tag{3.19}$$

where $\mathcal{V} = \lambda_x^{-1}(\mathbf{NB})(\mathbf{NB})^T + \mathbf{R}$. The estimate of $\boldsymbol{\theta}$ is thus obtained by maximizing ql_{MI}, that is, for each element $\boldsymbol{\theta}_l$ of $\boldsymbol{\theta} = (\sigma_1^2, \sigma_2^2, \tau_d)^T$, $\hat{\boldsymbol{\theta}}_l$ is the solution to

$$-\frac{1}{2}\mathrm{tr}(\mathbf{P}\frac{\partial\mathbf{R}}{\partial\boldsymbol{\theta}_l}) + \frac{1}{2}(\tilde{\mathbf{W}} - \mathbf{N}\hat{\mathbf{x}}_0)^T\mathbf{R}^{-1}\frac{\partial\mathbf{R}}{\partial\boldsymbol{\theta}_l}\mathbf{R}^{-1}(\tilde{\mathbf{W}} - \mathbf{N}\hat{\mathbf{x}}_0) = 0,$$

where $\mathbf{P} = \mathbf{R}^{-1} - \mathbf{R}^{-1}(\mathbf{NB}_*)\mathbf{H}^{-1}(\mathbf{NB}_*)^T\mathbf{R}^{-1}$. In addition, the Fisher information matrix $\mathcal{I}$ is derived as

$$\mathcal{I}_{(l,k)} = \frac{1}{2}\mathrm{tr}(\mathbf{P}\frac{\partial\mathbf{R}}{\partial\boldsymbol{\theta}_l}\mathbf{P}\frac{\partial\mathbf{R}}{\partial\boldsymbol{\theta}_k}),$$

which is then used to construct the approximate covariance matrix of $\hat{\boldsymbol{\theta}}$.

In fact, the use of the Laplace approximation in the quasi-likelihood ql_{MI} will bring some bias in the estimation, particularly when the data are sparse. Following Lin and Breslow (1996) and Lin and Zhang (1999), a correction procedure is taken on $\hat{\boldsymbol{\theta}}$. Specifically, with $h(\boldsymbol{\mu}^*) = \mathbf{Nx}_0$, we obtain $\mathbf{A}_0 = \mathrm{diag}\{\mathbf{A}_{01}, \ldots, \mathbf{A}_{0n}\}$ where $\mathbf{A}_{0i} = \mathrm{diag}\{\frac{w_{ij}}{\phi_W}v_W(\mu_{ij}^*)\}$; and diagonal matrices $\mathbf{A}_1$ and $\mathbf{A}_2$ are similarly defined with diagonal elements being $\frac{w_{ij}}{\phi_W}v_W'(\mu_{ij}^*)v_W(\mu_{ij}^*)$ and $\frac{w_{ij}}{\phi_W}\{v_W''(\mu_{ij}^*)(v_W(\mu_{ij}^*))^2 + (v_W'(\mu_{ij}^*))^2 v_W(\mu_{ij}^*)\}$, respectively. Further, by denoting $\mathbf{H}^{(2)}$ such that $\mathbf{H}^{(2)}_{(i,j)} = \mathbf{H}_{(i,j)}{}^2$ for any matrix $\mathbf{H}$, and $\mathbf{J} = \mathrm{diag}\{1, 1, \mathbf{1}_{r-2}\}$, the estimate of $\boldsymbol{\theta}$ can be cor-

rected by

$$\hat{\boldsymbol{\theta}}_C = (\mathbf{C}_1 + \mathbf{C}_2 - \mathbf{C}_3)^{-1}\mathbf{C}_1\hat{\boldsymbol{\theta}},$$

where matrices $\mathbf{C}_1$, $\mathbf{C}_2$ and $\mathbf{C}_3$ are evaluated at $\hat{\mathbf{x}}_0$, with

$$\mathbf{C}_1 = \frac{1}{2}\mathbf{J}^T(\mathbf{Q}^T\mathbf{A}_0\mathbf{Q})^{(2)}\mathbf{J},$$

$$\mathbf{C}_2 = \frac{1}{4}\mathbf{J}^T{\mathbf{Q}^{(2)}}^T\mathbf{A}_2\mathbf{Q}^{(2)}\mathbf{J},$$

and

$$\mathbf{C}_3 = \frac{1}{4}\mathbf{J}^T{\mathbf{Q}^{(2)}}^T\mathbf{A}_1\mathbf{N}(\mathbf{N}^T\mathbf{A}_0\mathbf{N})^{-1}\mathbf{N}^T\mathbf{A}_1\mathbf{Q}^{(2)}\mathbf{J}.$$

Lin and Zhang (1999) estimate the smoothing parameter λ_x through a generalization of maximum likelihood (GML; Wahba, 1985); however, due to numerical difficulties, we will select it by GCV. We assume independence in defining the GCV statistic, even though the data are in fact correlated within subjects. Specifically, for $\hat{\boldsymbol{\mu}}$ of equation (3.12) evaluated at $\hat{\mathbf{x}}_0$ and $\hat{\mathbf{b}}$, with

$$\mathbf{A}^x = \frac{\partial^2}{\partial\mathbf{x}_0\partial\mathbf{x}_0^T}(\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\tilde{d}_{ij}) = \mathbf{N}^T\mathcal{A}^x\mathbf{N},$$

where

$$\mathcal{A}_i^x = \mathrm{diag}\{\frac{w_{ij}\{v_W(\hat{\mu}_{ij})h^{'}(\hat{\mu}_{ij}) + (W_{ij} - \hat{\mu}_{ij})(v_W^{'}(\hat{\mu}_{ij})h^{'}(\hat{\mu}_{ij}) + v_W(\hat{\mu}_{ij})h^{''}(\hat{\mu}_{ij}))\}}{\phi_W v_W^2(\hat{\mu}_{ij})(h^{'}(\hat{\mu}_{ij}))^3}\},$$

and $\mathcal{A}^x = \mathrm{diag}\{\mathcal{A}_1^x, \ldots, \mathcal{A}_n^x\}$, we have the effective degrees of freedom

$$edf_x = \mathrm{tr}((\mathbf{A}^x + \lambda_x\mathbf{K})^{-1}\mathbf{A}^x),$$

then calculate the GCV statistic as

$$GCV(\lambda_x) = \frac{\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\tilde{d}_{ij}}{\dot{n}(1-\frac{edf_x}{\dot{n}})^2}, \tag{3.20}$$

for $\dot{n} = \sum_{i=1}^{n} n_i$ being the total number of observations. Hence, with the GCV statistic defined as above, the smoothing parameter λ_x can be estimated by a grid search.

3.2.1.3 Summary of Stage-I

In the first stage, we obtain the subject-specific $\hat{x}_i(\cdot)$'s from the observed longitudinal covariate processes $\mathbf{W}_i$'s, by assuming they are NCS functions.

Specifically, $\mathbf{x}_0$ and $\mathbf{b}_i$'s are estimated by maximizing the DPQL of (3.14), or from the equivalent mixed model representation of equation (3.16) or (3.17); and an approximate covariance matrix for $\hat{\mathbf{x}}_0$ is derived as equation (3.18). For each $\hat{x}_i(\cdot)$, the NCS vector is computed as $\hat{\mathbf{x}}_i = \hat{\mathbf{x}}_0 + \mathbf{B}_*\hat{\mathbf{b}}_i$. The estimate of variance component $\boldsymbol{\theta} = (\sigma_1^2, \sigma_2^2, \tau_d)^T$ is obtained by maximizing the marginal log-quasi-likelihood of (3.19), and a bias-correction procedure is often followed. The smoothing parameter λ_x is selected through GCV, which is defined in equation (3.20).

In addition, it is worth noting that, when model (3.6) instead of model (3.4,3.5) is assumed, one may refer to Zhang et al. (2007) for the details of a similar estimation procedure.

3.2.2 Stage-II

With subject-specific profile functions $\hat{x}_i(\cdot)$'s estimated as $\hat{\mathbf{x}}_i = \hat{\mathbf{x}}_0 + \mathbf{B}_*\hat{\mathbf{b}}_i$ from the first stage, the logarithm of the pseudo-quasi-likelihood can be derived from

model (3.1,3.10) as

$$ql_{II} = -\frac{1}{2}\sum_{i=1}^{n}\tilde{d}_i,$$

where

$$\tilde{d}_i = -2\int_{Y_i}^{\eta_i}\frac{w_i(Y_i - s)}{\phi_Y v_Y(s)}ds$$

is defined with prior weights w_i's, dispersion parameter ϕ_Y, and variance function $v_Y(\cdot)$ specified from the distribution $f_Y(\cdot)$.

The estimation procedure is similar to that for the first stage in Section 3.2.1. A roughness penalty term, $\frac{1}{2}\lambda_\gamma\int_{T_1}^{T_2}(\gamma''(t))^2dt$, is introduced to smooth $\gamma(\cdot)$, where λ_γ is a smoothing parameter that will be selected by GCV. For a given λ_γ, we may estimate the parametric coefficient $\boldsymbol{\delta}$ and the nonparametric function $\gamma(\cdot)$ by maximizing the penalized pseudo-quasi-likelihood of

$$-\frac{1}{2}\sum_{i=1}^{n}\tilde{d}_i - \frac{1}{2}\lambda_\gamma\boldsymbol{\gamma}^T\mathbf{K}\boldsymbol{\gamma}. \tag{3.21}$$

Specifically, $\hat{\boldsymbol{\delta}}$ and $\hat{\boldsymbol{\gamma}}$ are the solutions to estimating equations of

$$\begin{aligned}\mathbf{Z}^T\mathcal{U}\boldsymbol{\Xi}(\mathbf{Y} - \boldsymbol{\eta}) &= \mathbf{0},\\ \hat{\mathbf{X}}^T\mathcal{U}\boldsymbol{\Xi}(\mathbf{Y} - \boldsymbol{\eta}) - \lambda_\gamma\mathbf{K}\boldsymbol{\gamma} &= \mathbf{0},\end{aligned}$$

where $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$, $\boldsymbol{\Xi} = \mathrm{diag}\{g'(\eta_i)\}$, and $\mathcal{U} = \mathrm{diag}\{\frac{w_i}{\phi_Y v_Y(\eta_i)(g'(\eta_i))^2}\}$.

Typically, through a mixed effects transformation of

$$\boldsymbol{\gamma} = \mathbf{T}\boldsymbol{\alpha}_\gamma + \mathbf{B}\mathbf{a}_\gamma,$$

where $\mathbf{a}_\gamma \sim \mathrm{N}(\mathbf{0}, \lambda_\gamma^{-1}\mathbf{I})$, the model has an equivalent GLMM expression as

$$g(\boldsymbol{\eta}) = \mathbf{Z}\boldsymbol{\delta} + \hat{\mathbf{X}}\mathbf{T}\boldsymbol{\alpha}_\gamma + \hat{\mathbf{X}}\mathbf{B}\mathbf{a}_\gamma.$$

Furthermore, the estimation problem corresponds to an LMM with the form of

$$\tilde{\mathbf{Y}} = \mathbf{Z}\boldsymbol{\delta} + \hat{\mathbf{X}}\mathbf{T}\boldsymbol{\alpha}_\gamma + \hat{\mathbf{X}}\mathbf{B}\mathbf{a}_\gamma + \mathbf{e}_\epsilon,$$

where $\tilde{\mathbf{Y}} = \mathbf{Z}\boldsymbol{\delta}+\hat{\mathbf{X}}\boldsymbol{\gamma}+\boldsymbol{\Xi}(\mathbf{Y}-\boldsymbol{\eta})$ is a working vector, and $\mathbf{e}_\epsilon \sim \mathrm{N}(\mathbf{0},\mathcal{U}^{-1})$. Following the LMM representation, we obtain an approximate covariance as

$$\mathrm{Cov}((\hat{\boldsymbol{\delta}}^T, \hat{\boldsymbol{\alpha}}_\gamma^T, \hat{\mathbf{a}}_\gamma^T)^T) = \mathbf{G}^{-1}\mathbf{G}_0\mathbf{G}^{-1}, \tag{3.22}$$

where $\mathbf{G}_0 = (\mathbf{Z}, \hat{\mathbf{X}}\mathbf{B}_*)^T\mathcal{U}(\mathbf{Z}, \hat{\mathbf{X}}\mathbf{B}_*)$, and

$$\mathbf{G} = \begin{pmatrix} \mathbf{Z}^T\mathcal{U}\mathbf{Z} & \mathbf{Z}^T\mathcal{U}(\hat{\mathbf{X}}\mathbf{T}) & \mathbf{Z}^T\mathcal{U}(\hat{\mathbf{X}}\mathbf{B}) \\ (\hat{\mathbf{X}}\mathbf{T})^T\mathcal{U}\mathbf{Z} & (\hat{\mathbf{X}}\mathbf{T})^T\mathcal{U}(\hat{\mathbf{X}}\mathbf{T}) & (\hat{\mathbf{X}}\mathbf{T})^T\mathcal{U}(\hat{\mathbf{X}}\mathbf{B}) \\ (\hat{\mathbf{X}}\mathbf{B})^T\mathcal{U}\mathbf{Z} & (\hat{\mathbf{X}}\mathbf{B})^T\mathcal{U}(\hat{\mathbf{X}}\mathbf{T}) & (\hat{\mathbf{X}}\mathbf{B})^T\mathcal{U}(\hat{\mathbf{X}}\mathbf{B}) + \lambda_\gamma\mathbf{I} \end{pmatrix}.$$

Thus, $\mathrm{Cov}(\hat{\boldsymbol{\delta}})$ and $\mathrm{Cov}((\hat{\boldsymbol{\alpha}}_\gamma^T, \hat{\mathbf{a}}_\gamma^T)^T)$ are the corresponding blocks of $\mathbf{G}^{-1}\mathbf{G}_0\mathbf{G}^{-1}$, and

$$\mathrm{Cov}(\hat{\boldsymbol{\gamma}}) = \mathbf{B}_*\mathrm{Cov}((\hat{\boldsymbol{\alpha}}_\gamma^T, \hat{\mathbf{a}}_\gamma^T)^T)\mathbf{B}_*^T. \tag{3.23}$$

We may select the smoothing parameter λ_γ by GCV. To be specific, with $\hat{\boldsymbol{\eta}}$ of equation (3.10) evaluated at $\hat{\boldsymbol{\delta}}$ and $\hat{\boldsymbol{\gamma}}$, we let

$$\mathbf{A}^\gamma = \frac{\partial^2}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^T}(\frac{1}{2}\sum_{i=1}^n \tilde{d}_i) = \hat{\mathbf{X}}^T\mathcal{A}^\gamma\hat{\mathbf{X}},$$

where

$$\mathcal{A}^\gamma = \mathrm{diag}\{\frac{w_i\{v_Y(\hat{\eta}_i)g'(\hat{\eta}_i) + (Y_i - \hat{\eta}_i)(v_Y'(\hat{\eta}_i)g'(\hat{\eta}_i) + v_Y(\hat{\eta}_i)g''(\hat{\eta}_i))\}}{\phi_Y v_Y^2(\hat{\eta}_i)(g'(\hat{\eta}_i))^3}\},$$

then have the effective degrees of freedom

$$edf_\gamma = p + \mathrm{tr}((\mathbf{A}^\gamma + \lambda_\gamma \mathbf{K})^{-1}\mathbf{A}^\gamma),$$

and the GCV statistic

$$GCV(\lambda_\gamma) = \frac{\frac{1}{2}\sum_{i=1}^{n} \tilde{d}_i}{n(1 - \frac{edf_\gamma}{n})^2}. \tag{3.24}$$

Thus, the optimal choice of the smoothing parameter λ_γ is selected by minimizing $GCV(\lambda_\gamma)$ on a grid of points.

Rather than model (3.1,3.2), the problem is often to deal with Gaussian distributed responses Y_i's. From equation (3.3), the corresponding calibration model becomes

$$\mathbf{Y} = \mathbf{Z}\boldsymbol{\delta} + \hat{\mathbf{X}}\boldsymbol{\gamma} + \boldsymbol{\epsilon}^*. \tag{3.25}$$

By treating ϵ_i^*'s as independent with $\boldsymbol{\epsilon}^* \sim \mathrm{N}(\mathbf{0}, \sigma^2_{\epsilon^*}\mathbf{I})$, a pseudo-log-likelihood is derived as

$$l_{II} = -\frac{n}{2}\log(2\pi\sigma^2_{\epsilon^*}) - \frac{1}{2\sigma^2_{\epsilon^*}}(\mathbf{Y} - \mathbf{Z}\boldsymbol{\delta} - \hat{\mathbf{X}}\boldsymbol{\gamma})^T(\mathbf{Y} - \mathbf{Z}\boldsymbol{\delta} - \hat{\mathbf{X}}\boldsymbol{\gamma}).$$

A roughness penalty term for smoothing $\gamma(\cdot)$ is introduced, then $\hat{\boldsymbol{\delta}}$ and $\hat{\boldsymbol{\gamma}}$ are computed by maximizing the penalized pseudo-log-likelihood of

$$-\frac{1}{2\sigma^2_{\epsilon^*}}(\mathbf{Y} - \mathbf{Z}\boldsymbol{\delta} - \hat{\mathbf{X}}\boldsymbol{\gamma})^T(\mathbf{Y} - \mathbf{Z}\boldsymbol{\delta} - \hat{\mathbf{X}}\boldsymbol{\gamma}) - \frac{1}{2}\lambda_\gamma \boldsymbol{\gamma}^T\mathbf{K}\boldsymbol{\gamma},$$

which yields maximum penalized likelihood estimators (Zhang et al., 1998) of

$$\hat{\boldsymbol{\delta}} = (\mathbf{Z}^T\mathbf{W}_z\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{W}_z\mathbf{Y}, \tag{3.26}$$

and

$$\hat{\boldsymbol{\gamma}} = (\hat{\mathbf{X}}^T \mathbf{W}_x \hat{\mathbf{X}} + \lambda_\gamma \mathbf{K})^{-1} \hat{\mathbf{X}}^T \mathbf{W}_x \mathbf{Y}, \tag{3.27}$$

where

$$\mathbf{W}_z = \frac{1}{\sigma^2_{\epsilon^*}} \{\mathbf{I} - \hat{\mathbf{X}}(\hat{\mathbf{X}}^T \hat{\mathbf{X}} + \sigma^2_{\epsilon^*} \lambda_\gamma \mathbf{K})^{-1} \hat{\mathbf{X}}^T\},$$

and

$$\mathbf{W}_x = \frac{1}{\sigma^2_{\epsilon^*}} \{\mathbf{I} - \mathbf{Z}(\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T\}.$$

Moreover, the covariance matrices are derived as

$$\mathrm{Cov}(\hat{\boldsymbol{\delta}}) = \sigma^2_{\epsilon^*} (\mathbf{Z}^T \mathbf{W}_z \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{W}_z \mathbf{W}_z \mathbf{Z} (\mathbf{Z}^T \mathbf{W}_z \mathbf{Z})^{-1}, \tag{3.28}$$

and

$$\mathrm{Cov}(\hat{\boldsymbol{\gamma}}) = \sigma^2_{\epsilon^*} (\hat{\mathbf{X}}^T \mathbf{W}_x \hat{\mathbf{X}} + \lambda_\gamma \mathbf{K})^{-1} \hat{\mathbf{X}}^T \mathbf{W}_x \mathbf{W}_x \hat{\mathbf{X}} (\hat{\mathbf{X}}^T \mathbf{W}_x \hat{\mathbf{X}} + \lambda_\gamma \mathbf{K})^{-1}. \tag{3.29}$$

The variance component, $\sigma^2_{\epsilon^*}$, is estimated from the resticted maximum likelihood (REML), and provides an approximation to σ^2_ϵ. Explicitly, $\hat{\sigma}^2_{\epsilon^*}$ solves the estimating equation

$$-\frac{1}{2}\mathrm{tr}(\mathbf{P}) + \frac{1}{2\sigma^4_{\epsilon^*}} (\mathbf{Y} - \mathbf{Z}\hat{\boldsymbol{\delta}} - \hat{\mathbf{X}}\hat{\boldsymbol{\gamma}})^T (\mathbf{Y} - \mathbf{Z}\hat{\boldsymbol{\delta}} - \hat{\mathbf{X}}\hat{\boldsymbol{\gamma}}) = 0, \tag{3.30}$$

where

$$\mathbf{P} = \frac{1}{\sigma^2_{\epsilon^*}} \{\mathbf{I} - (\mathbf{Z}, \hat{\mathbf{X}}) \begin{pmatrix} \mathbf{Z}^T \mathbf{Z} & \mathbf{Z}^T \hat{\mathbf{X}} \\ \hat{\mathbf{X}}^T \mathbf{Z} & \hat{\mathbf{X}}^T \hat{\mathbf{X}} + \sigma^2_{\epsilon^*} \lambda_\gamma \mathbf{K} \end{pmatrix}^{-1} (\mathbf{Z}, \hat{\mathbf{X}})^T\},$$

and $\mathrm{var}(\hat{\sigma}^2_{\epsilon^*})$ corresponds to the inverse of Fisher information $\mathcal{I} = \frac{1}{2}\mathrm{tr}(\mathbf{P}^2)$.

Similarly, the smoothing parameter λ_γ is selected by GCV, instead of a GML approach suggested in Zhang et al. (1998). With $\mathbf{W}_x$ evaluated at $\hat{\sigma}^2_{\epsilon^*}$, we have

the effective degrees of freedom

$$edf_\gamma = p + \mathrm{tr}((\hat{\mathbf{X}}^T\mathbf{W}_x\hat{\mathbf{X}} + \lambda_\gamma\mathbf{K})^{-1}\hat{\mathbf{X}}^T\mathbf{W}_x\hat{\mathbf{X}}),$$

and the GCV statistic

$$GCV(\lambda_\gamma) = \frac{\frac{1}{2\hat{\sigma}^2_{\epsilon^*}}(\mathbf{Y} - \mathbf{Z}\hat{\boldsymbol{\delta}} - \hat{\mathbf{X}}\hat{\boldsymbol{\gamma}})^T(\mathbf{Y} - \mathbf{Z}\hat{\boldsymbol{\delta}} - \hat{\mathbf{X}}\hat{\boldsymbol{\gamma}})}{n(1 - \frac{edf_\gamma}{n})^2}. \tag{3.31}$$

Hence, λ_γ is selected by searching the minimum of $GCV(\lambda_\gamma)$ over a grid of points.

In summary, the second stage estimation allows the responses to be either discrete or continuous. If Y_i's are discrete so that model (3.1,3.2) is assumed, $\boldsymbol{\delta}$ and $\boldsymbol{\gamma}$ can be estimated by maximizing the penalized pseudo-quasi-likelihood of (3.21), and their approximate covariance matrices are given in equations (3.22) and (3.23); the smoothing parameter λ_γ is selected by GCV as defined in equation (3.24). If Y_i's are Gaussian and model (3.3) is instead assumed, $\hat{\boldsymbol{\delta}}$ and $\hat{\boldsymbol{\gamma}}$ are derived in equations (3.26) and (3.27), with covariance matrices as equations (3.28) and (3.29); through REML, $\hat{\sigma}^2_{\epsilon^*}$ solves the estimating equation of (3.30), and provides an approximation to σ^2_ϵ; the smoothing parameter λ_γ is obtained through $GCV(\lambda_\gamma)$ as equation (3.31).

3.3 Simulation Studies

In this section, we assess the finite sample performance of the proposed two-stage estimation for models with discrete longitudinal covariates, via simulations under different data generation settings. Specifically, we first conduct simulation under a base setting, then manipulate each of several factors and compare the performance separately.

In the base setting, both longitudinal covariates and responses are generated from Poisson distributions. We consider the sample size of $n = 500$, and set $\mathbf{t}^0$ as a vector of $r = 16$ equally spaced knots over the interval $[-1.5, 1.5]$. The longitudinal covariates are generated from

$$W_{ij} \sim \text{Poisson}(\mu_{ij}),$$

and

$$\log(\mu_{ij}) = x_0(t_{ij}) + d_i(t_{ij});$$

in particular, W_{ij}'s are observed at all time points of $\mathbf{t}^0$ for all subjects (i.e., $n_i = n_0 = r$ for every i). Here, we set $x_0(t) = 0.5 + \sin(\frac{2\pi}{3}t)$, and have $d_i(\cdot)$ by $\mathbf{d}_i = \mathbf{B}_*\mathbf{b}_i$ with $\mathbf{b}_i$'s being a random sample from $\text{N}(\mathbf{0}, \text{diag}\{\sigma_1^2, \sigma_2^2, \tau_d\mathbf{I}_{(r-2)\times(r-2)}\})$, where $\sigma_1^2 = 0.5^2$, $\sigma_2^2 = 0.4^2$ and $\tau_d = 0.6^2$. The response data Y_i's are then generated from

$$Y_i \sim \text{Poisson}(\eta_i),$$

and

$$\log(\eta_i) = \int_{-1.5}^{1.5} x_i(t)\gamma(t)dt,$$

where $\gamma(t) = \sin(\frac{2\pi}{3}t)$.

We generate $N = 100$ datasets and apply the proposed two-stage method in estimation. For each parameter, mean and standard error of N estimates, and mean squared error (MSE) and its empirical standard error are calculated; for nonparametric functions $x_0(\cdot)$ and $\gamma(\cdot)$, mean integrated squared error (MISE), instead of MSE, is used to summarize the estimation results. We also present the estimated curves for $x_0(\cdot)$ and $\gamma(\cdot)$, which are derived from evaluating the fitted functions of the N replications at a set of grid points and connecting means at all

grid points.

Table 3.2: Estimation Results of Base Simulation Setting

	mean (se)	MSE/MISE (se)
$\sigma_1^2 = 0.5^2$	0.2375 (0.0174)	0.0005 (0.0005)
$\sigma_2^2 = 0.4^2$	0.1483 (0.0121)	0.0003 (0.0003)
$\tau_d = 0.6^2$	0.3436 (0.0500)	0.0027 (0.0041)
$x_0(\cdot)$		0.0111 (0.0074)
$\gamma(\cdot)$		0.0482 (0.0388)

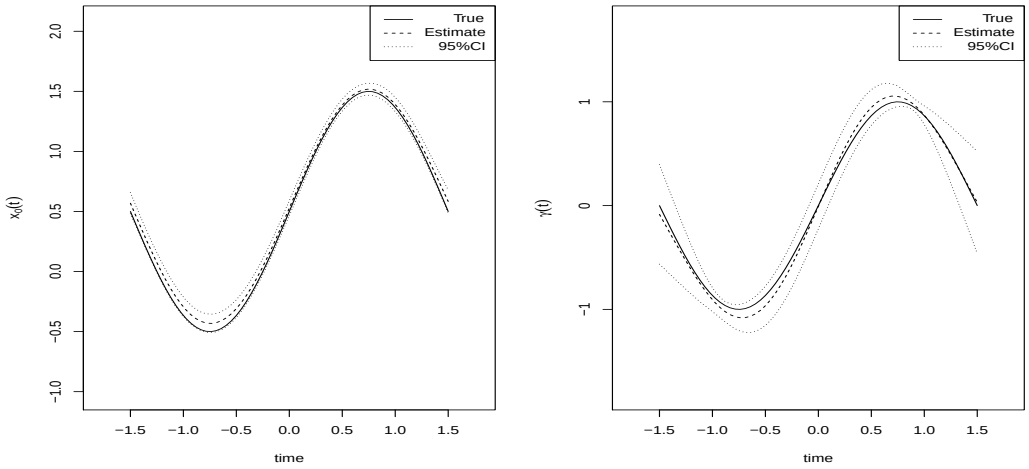

Figure 3.1: Estimated Functions of Base Simulation Setting: left, population profile function $x_0(\cdot)$; right, effect function $\gamma(\cdot)$

The estimation results under the base setting are summarized in Table 3.2 and Figure 3.1. In general, Table 3.2 shows that each of the variance components is estimated closely to its true value; Figure 3.1 indicates that our proposed procedure estimates $x_0(\cdot)$ and $\gamma(\cdot)$ with relatively small bias, and the true functions are covered by the 95% pointwise empirical confidence intervals. It is observed that this method tends to slightly underestimate variance components, which, however, are not crucial in estimating functions $x_0(\cdot)$ and $\gamma(\cdot)$.

In order to assess the estimation performance under a variety of circumstances,

we manipulate several factors. For each of the factors, we compare the results to those from the base setting, by holding the other factors invariant.

Table 3.3: Estimation Results of Simulation Setting (i)

	$n = 250$	$n = 500$ [Base]	$n = 750$
		mean (se)	
$\sigma_1^2 = 0.5^2$	0.2411 (0.0260)	0.2375 (0.0174)	0.2349 (0.0135)
$\sigma_2^2 = 0.4^2$	0.1483 (0.0178)	0.1483 (0.0121)	0.1491 (0.0100)
$\tau_d = 0.6^2$	0.3456 (0.0688)	0.3436 (0.0500)	0.3390 (0.0402)
		MSE/MISE (se)	
σ_1^2	0.0007 (0.0010)	0.0005 (0.0005)	0.0004 (0.0005)
σ_2^2	0.0004 (0.0006)	0.0003 (0.0003)	0.0002 (0.0003)
τ_d	0.0049 (0.0072)	0.0027 (0.0041)	0.0020 (0.0031)
$x_0(\cdot)$	0.0173 (0.0119)	0.0111 (0.0074)	0.0093 (0.0058)
$\gamma(\cdot)$	0.0931 (0.1477)	0.0482 (0.0388)	0.0362 (0.0267)

(i) We first manipulate the sample size with $n = 250$ and $n = 750$, and have results summarized in Table 3.3. Our simulation results indicate that, when the sample size increases, the performance of the estimation will improve in terms of MSE or MISE.

Table 3.4: Estimation Results of Simulation Setting (ii)

	$r = 8$	$r = 16$ [Base]	$r = 32$
		mean (se)	
$\sigma_1^2 = 0.5^2$	0.2333 (0.0178)	0.2375 (0.0174)	0.2381 (0.0145)
$\sigma_2^2 = 0.4^2$	0.1468 (0.0150)	0.1483 (0.0121)	0.1544 (0.0116)
$\tau_d = 0.6^2$	0.3261 (0.0640)	0.3436 (0.0500)	0.3541 (0.0349)
		MSE/MISE (se)	
σ_1^2	0.0006 (0.0006)	0.0005 (0.0005)	0.0003 (0.0005)
σ_2^2	0.0004 (0.0004)	0.0003 (0.0003)	0.0002 (0.0002)
τ_d	0.0052 (0.0058)	0.0027 (0.0041)	0.0012 (0.0017)
$x_0(\cdot)$	0.0226 (0.0112)	0.0111 (0.0074)	0.0059 (0.0045)
$\gamma(\cdot)$	0.0821 (0.1101)	0.0482 (0.0388)	0.0345 (0.0321)

(ii) The number of knots over $[-1.5, 1.5]$ is then manipulated by letting $r = 8$ and $r = 32$. From Table 3.4, we find out that the proposed estimation gets improved with more time knots introduced into the model.

Table 3.5: Estimation Results of Simulation Setting (iii)

	$k = 0.5$	$k = 0.75$	$k = 1$ [Base]
		mean (se)	
$\sigma_1^2 = 0.5^2$	0.2367 (0.0182)	0.2352 (0.0176)	0.2375 (0.0174)
$\sigma_2^2 = 0.4^2$	0.1477 (0.0132)	0.1494 (0.0115)	0.1483 (0.0121)
$\tau_d = 0.6^2$	0.3376 (0.0523)	0.3443 (0.0483)	0.3436 (0.0500)
		MSE/MISE (se)	
σ_1^2	0.0005 (0.0006)	0.0005 (0.0007)	0.0005 (0.0005)
σ_2^2	0.0003 (0.0004)	0.0002 (0.0003)	0.0003 (0.0003)
τ_d	0.0032 (0.0042)	0.0026 (0.0030)	0.0027 (0.0041)
$x_0(\cdot)$	0.0130 (0.0077)	0.0120 (0.0073)	0.0111 (0.0074)
$\gamma(\cdot)$	0.0638 (0.0616)	0.0588 (0.0546)	0.0482 (0.0388)

(iii) In order to assess the influence of missing schedule, we allow longitudinal covariates to be unbalanced among subjects. Specifically, each n_i is an integer randomly chosen from $\{\lceil kr \rceil, \ldots, r\}$ for $k = 0.75$ or $k = 0.5$. Once n_i is chosen for i-th subject, $\{t_{i1}, \ldots, t_{in_i}\}$ is a set of n_i out of r distinct points randomly drawn from $\mathbf{t}^0$. As demonstrated in Table 3.5, estimations of $x_0(\cdot)$ and σ_1^2, σ_2^2, τ_d do not vary much when some longitudinal covariates are not scheduled to be measured; however, estimation of $\gamma(\cdot)$ gets improved with less unobserved longitudinal covariates. A possible explanation is that, longitudinal covariates are in fact correlated, and existing ones would provide "enough" information for estimating population profile function and variance components in the first stage; whilst, since more $x_i(t_{ij})$'s are estimated from the data, the performance of $\hat{\gamma}(\cdot)$ will be improved.

(iv) We also compare estimation performance for different choices of the effect function $\gamma(\cdot)$. Particularly, we instead use $\gamma(t) = \sin(\frac{2\pi}{3}t) - 0.8t$, $\gamma(t) = t^2 - 1$, and $\gamma(t) = t$ in simulations. The estimations obtained in Stage-I are same under all settings, thus we only provide those of $\gamma(\cdot)$ in Figure 3.2. As illustrated, $\gamma(\cdot)$'s are estimated well with small bias, while the true functions are covered by the 95% pointwise empirical confidence intervals. This suggests that our proposed method can be widely adopted in a variety of effect functions.

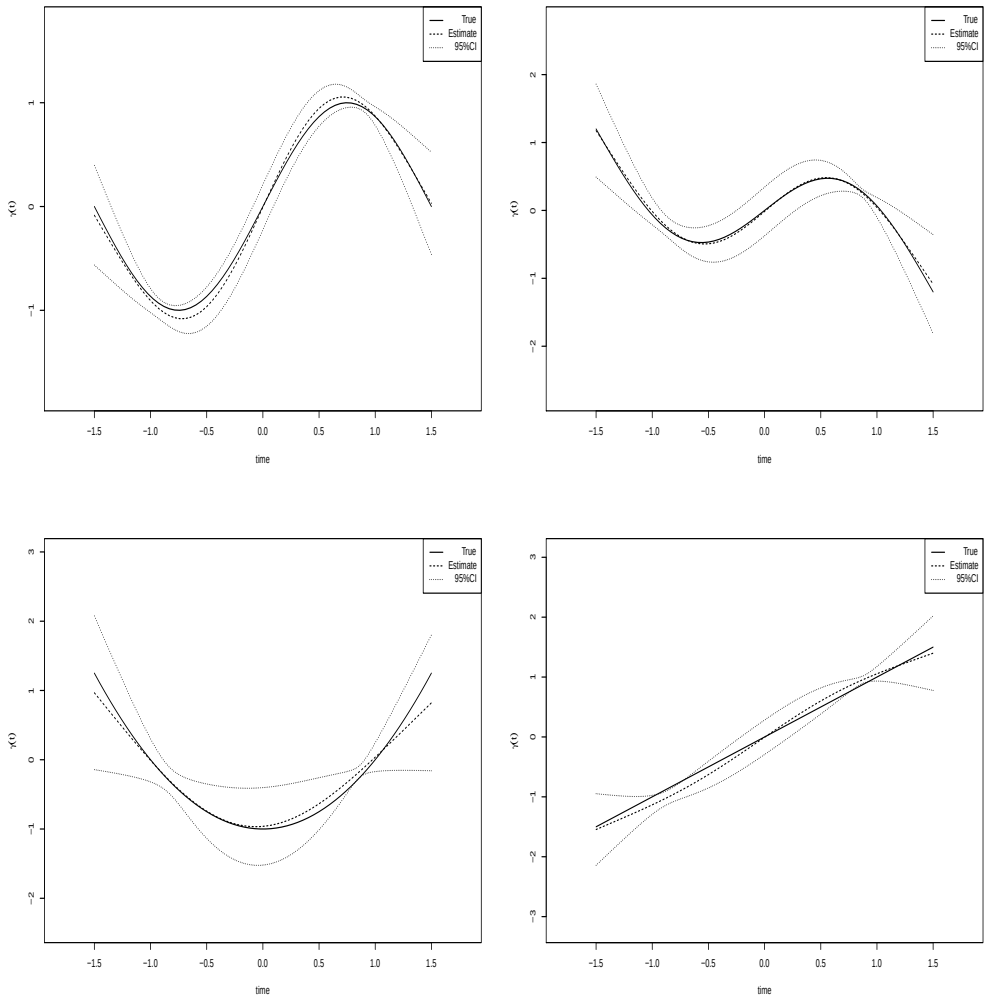

Figure 3.2: Estimated Functions of Simulation Setting (iv): upper-left, $\gamma(t) = \sin(\frac{2\pi}{3}t)$ [Base]; upper-right, $\gamma(t) = \sin(\frac{2\pi}{3}t) - 0.8t$; lower-left, $\gamma(t) = t^2 - 1$; lower-right, $\gamma(t) = t$

(v) The impact of introducing time-invariant effects is also assessed, when responses Y_i's are instead from

$$Y_i \sim \text{Poisson}(\eta_i),$$

Table 3.6: Estimation Results of Simulation Setting (v)

	[Base]	[v-1]	[v-2]	[v-3]
		mean (se)		
$\delta_1 = 0.5$	NA	0.4992 (0.0461)	NA	0.5067 (0.0507)
$\delta_2 = 0.5$	NA	NA	0.5003 (0.0277)	0.4992 (0.0219)
		MISE (se)		
$\gamma(\cdot)$	0.0482 (0.0388)	0.0455 (0.0376)	0.0505 (0.0446)	0.0530 (0.0391)

and

$$\log(\eta_i) = \mathbf{Z}_i^T \boldsymbol{\delta} + \int_{-1.5}^{1.5} x_i(t)\gamma(t)dt.$$

Specifically, we generate $\mathbf{Z}_i$'s in three ways: [v-1] Z_i is 1-dimensional and randomly drawn from $\{0, 1\}$ with $\mathrm{P}(Z_i = 1) = \frac{2}{3}$ independently for each subject, and $\delta_1 = 0.5$; [v-2] Z_i is 1-dimensional and independently distributed as $\mathrm{N}(0, 1^2)$, and $\delta_2 = 0.5$; [v-3] $\mathbf{Z}_i = (Z_{i1}, Z_{i2})^T$ is 2-dimensional with Z_{i1}'s from [v-1] and Z_{i2}'s from [v-2], while Z_{i1}'s and Z_{i2}'s are mutually independent, and $\boldsymbol{\delta} = (0.5, 0.5)^T$. Since estimation of $x_0(\cdot)$ in Stage-I is not affected by the introduction of $\mathbf{Z}_i$'s, we only summarize results regarding $\boldsymbol{\delta}$ and $\gamma(\cdot)$ as Table 3.6. Based on the simulation results, we find that the proposed procedure can estimate the additional coefficient $\boldsymbol{\delta}$ for the time-invariant covariate $\mathbf{Z}_i$, without losing efficiency in estimating the effect function $\gamma(\cdot)$.

(vi) Our proposed models allow either or both of longitudinal covariates and responses to be generated from other distributions rather than Poisson. Particularly, one setting simulates responses from normal distribution with

$$Y_i = \int_{-1.5}^{1.5} x_i(t)\gamma(t)dt + \epsilon_i,$$

where ϵ_i's are independently distributed as $\mathrm{N}(0, 0.5^2)$; another setting instead gen-

Table 3.7: Estimation Results of Simulation Setting (vi)

	Poisson-Poisson [Base]	Poisson-normal	Bernoulli-Poisson
		mean (se)	
$\sigma_1^2 = 0.5^2$	0.2375 (0.0174)	0.2375 (0.0174)	0.2375 (0.0269)
$\sigma_2^2 = 0.4^2$	0.1483 (0.0121)	0.1483 (0.0121)	0.1609 (0.0280)
$\tau_d = 0.6^2$	0.3436 (0.0500)	0.3436 (0.0500)	0.6743 (0.1509)
		MSE/MISE (se)	
σ_1^2	0.0005 (0.0005)	0.0005 (0.0005)	0.0009 (0.0010)
σ_2^2	0.0003 (0.0003)	0.0003 (0.0003)	0.0008 (0.0011)
τ_d	0.0027 (0.0041)	0.0027 (0.0041)	0.1213 (0.1221)
$x_0(\cdot)$	0.0111 (0.0074)	0.0111 (0.0074)	0.0227 (0.0151)
$\gamma(\cdot)$	0.0482 (0.0388)	0.0806 (0.0503)	0.2472 (0.2185)

erates longitudinal covariates by Bernoulli distrbution, that is,

$$W_{ij} \sim \text{Bernoulli}(\mu_{ij}),$$

and

$$\text{logit}(\mu_{ij}) = x_0(t_{ij}) + d_i(t_{ij}).$$

The comparisons are summarized in Table 3.7 and Figure 3.3. Generally, Figure 3.3 shows that, under each circumstance, estimates $\hat{x}_0(\cdot)$ and $\hat{\gamma}(\cdot)$ are close to their true functions, which turn out to be covered by the corresponding 95% pointwise empirical confidence intervals, thus our proposed method can be adopted with a wide range of distribution families. From the results in Table 3.7, and the comparisons between Figures 3.2 and 3.7, we notice that $\gamma(\cdot)$ is estimated better if responses are generated from Poisson distributions than from normal distributions. On the other hand, if longitudinal covariates are generated according to Bernoulli distributions, estimation performance of $x_0(\cdot)$ and its associated variance components are worse compared with those from Poisson distribution, which leads to a relatively inferior estimation of $\gamma(\cdot)$.

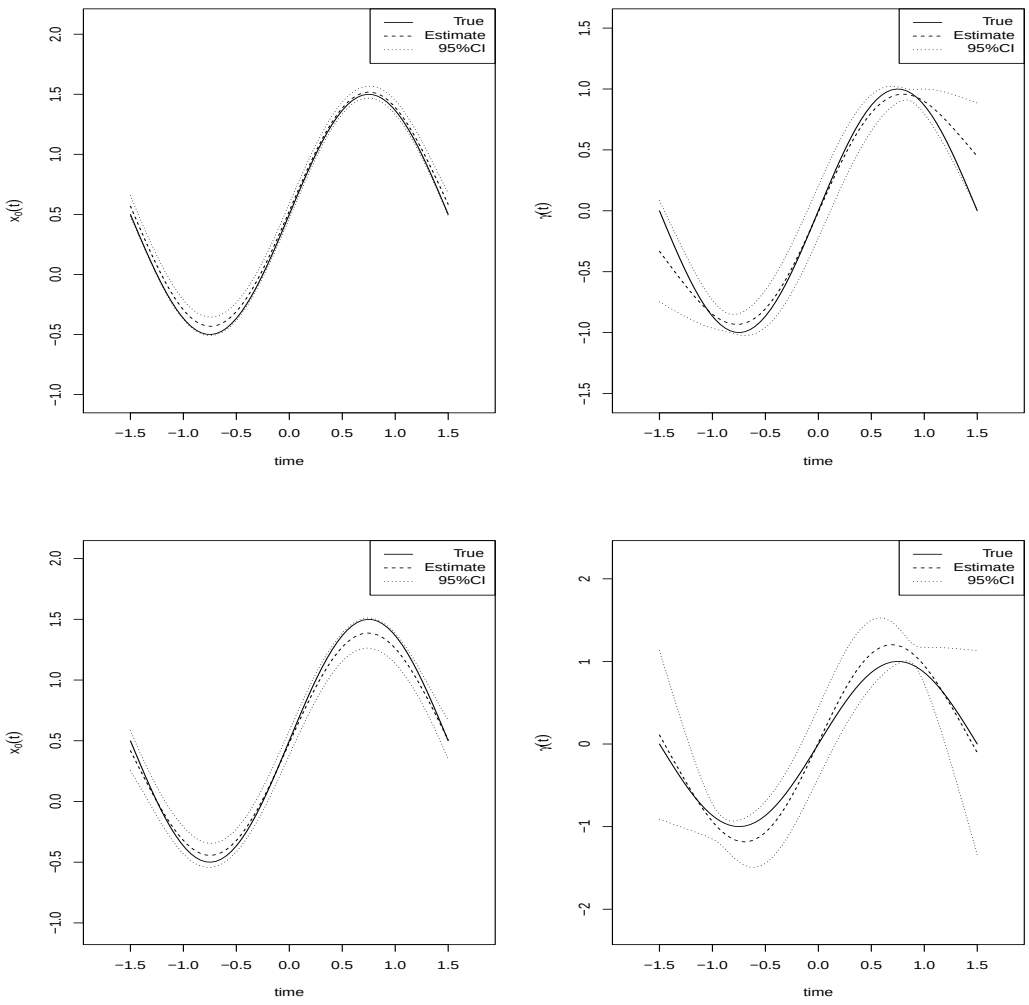

Figure 3.3: Estimated Functions of Simulation Setting (vi): upper, Poisson-normal; lower, Bernoulli-Poisson

3.4 Application

In this section, we extend the models proposed in Section 3.1.1 to accommodate an extra clustered random effect, then adopt the extended model to health risk behavior data of alcoholic couples collected through an interactive voice response (IVR) system.

3.4.1 Extended Model

In order to consider models for clustered subjects, we slightly modify the notation in Section 3.1.1 by denoting $i = 1, \ldots, n$ for cluster, $k = 1, \ldots, n_i$ for single subject nested in i-th cluster, and $\mathbf{W}_{ik} = (W_{ik1}, \ldots, W_{ikn_{ik}})^T$ for longitudinal covariate process of (ik)-th subject observed at $\{t_{ik1}, \ldots, t_{ikn_{ik}}\}$. With this notation, the longitudinal covariate model of (3.4,3.5) or (3.6) becomes

$$W_{ikj} \sim f_W(\cdot; \mu_{ikj}),$$

$$h(\mu_{ikj}) = x_{ik}(t_{ikj}),$$

or

$$W_{ikj} = x_{ik}(t_{ikj}) + \varepsilon_{ikj};$$

accordingly, the response model of (3.1,3.2) or (3.3) has the form

$$Y_{ik} \sim f_Y(\cdot; \eta_{ik}),$$

$$g(\eta_{ik}) = \mathbf{Z}_{ik}^T \boldsymbol{\delta} + \int_{T_1}^{T_2} x_{ik}(t)\gamma(t)dt + a_i,$$

or

$$Y_{ik} = \mathbf{Z}_{ik}^T \boldsymbol{\delta} + \int_{T_1}^{T_2} x_{ik}(t)\gamma(t)dt + a_i + \epsilon_{ik},$$

where a_i is a clustered random effect following $a_i \sim \mathrm{N}(0, \sigma_a^2)$.

With $\hat{x}_{ik}(\cdot)$'s obtained similarly to Section 3.2.1, $\boldsymbol{\delta}$ and $\gamma(\cdot)$ can be estimated by fitting a calibration model

$$g(\eta_{ik}) = \mathbf{Z}_{ik}^T \boldsymbol{\delta} + \int_{T_1}^{T_2} \hat{x}_{ik}(t)\gamma(t)dt + a_i, \tag{3.32}$$

or

$$Y_{ik} = \mathbf{Z}_{ik}^T \boldsymbol{\delta} + \int_{T_1}^{T_2} \hat{x}_{ik}(t)\gamma(t)dt + a_i + \epsilon_{ik}^*, \tag{3.33}$$

where ϵ_{ik}^*'s are treated as independent with $\mathrm{N}(0, \sigma_{\epsilon^*}^2)$.

We have the estimation for $\boldsymbol{\delta}$ and $\gamma(\cdot)$ following the logic in Section 3.2.2. Specially, for model (3.32), $\hat{\boldsymbol{\delta}}$, $\hat{\boldsymbol{\gamma}}$ and $\hat{a}_i$'s maximize

$$-\frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{n_k} d_{ik} - \frac{1}{2}\sigma_a^2 \sum_{i=1}^{n} a_i^2 - \frac{1}{2}\lambda_\gamma \boldsymbol{\gamma}^T \mathbf{K} \boldsymbol{\gamma},$$

for given λ_γ and σ_a^2; meanwhile, $\hat{\sigma}_a^2$ is estimated from the marginal quasi-likelihood, and λ_γ is selected based on a grid search of minimum GCV. For model (3.33), the estimates $\hat{\boldsymbol{\delta}}$ and $\hat{\boldsymbol{\gamma}}$ have the same forms as equations (3.26) and (3.27), where $\mathbf{N}_a = \mathrm{diag}\{\mathbf{1}_{n_1}, \ldots, \mathbf{1}_{n_n}\}$, and

$$\mathbf{W}_z = \mathbf{W} - \mathbf{W}\hat{\mathbf{X}}(\hat{\mathbf{X}}^T\mathbf{W}\hat{\mathbf{X}} + \lambda_\gamma \mathbf{K})^{-1}\hat{\mathbf{X}}^T\mathbf{W},$$

$$\mathbf{W}_x = \mathbf{W} - \mathbf{W}\mathbf{Z}(\mathbf{Z}^T\mathbf{W}\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{W},$$

$$\mathbf{W} = (\sigma_{\epsilon^*}^2\mathbf{I} + \sigma_a^2\mathbf{N}_a\mathbf{N}_a^T)^{-1};$$

their covariance matrices are

$$\mathrm{Cov}(\hat{\boldsymbol{\delta}}) = (\mathbf{Z}^T\mathbf{W}_z\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{W}_z\mathbf{W}^{-1}\mathbf{W}_z\mathbf{Z}(\mathbf{Z}^T\mathbf{W}_z\mathbf{Z})^{-1},$$

and

$$\mathrm{Cov}(\hat{\boldsymbol{\gamma}}) = (\hat{\mathbf{X}}^T\mathbf{W}_x\hat{\mathbf{X}} + \lambda_\gamma\mathbf{K})^{-1}\hat{\mathbf{X}}^T\mathbf{W}_x\mathbf{W}^{-1}\mathbf{W}_x\hat{\mathbf{X}}(\hat{\mathbf{X}}^T\mathbf{W}_x\hat{\mathbf{X}} + \lambda_\gamma\mathbf{K})^{-1};$$

the variance components σ_a^2 and $\sigma_{\epsilon^*}^2$ are estimated through REML, and the smooth-

ing parameter λ_γ is selected by GCV.

3.4.2 Application to a Study on Alcoholic Couples

In this section, we conduct statistical analysis on the data from a study using IVR technology to collect daily data from alcoholic couples for 14 consecutive days (Cranford et al., 2010), in order to demonstrate the application of the proposed extended model. The analysis also examines two important research questions, measurement reactivity and predictive validity, for daily process data. In particular, the results suggest that, the effect of measurement reactivity may only be evident in the first week of the IVR assessment, and fade away afterwards; in addition, the level of urge to drink before measurement reactivity may be more predictive in terms of the effect on the level of depression 6 months later.

Daily patterns of health risk behaviors, such as substance use, can be used to evaluate the risk of developing health problems and examine the dynamics of intervention effects over time (Mundt et al., 1995; Gwaltney et al., 2011). Such data are usually collected using retrospective methods, due to the high cost and heavy participant burden associated with collecting prospective data. However, prospective daily data collection using IVR has advantages of cutting costs of staff time, as well as minimizing recall bias and tendency to underreport socially undesirable behaviors (Bardone et al., 2000).

In this study, a total of 54 alcoholic married couples were recruited from the University of Michigan Addiction Treatment Services or local community, where either spouse met DSM-IV diagnosis (American Psychiatric Association, 1994) of past year alcohol use disorder. At baseline, participants completed questionnaires about their moods, marital interactions and drinking behaviors in the past month, as well as received an IVR training session. In each of the following 14 days, all

participants were instructed to call a toll-free telephone number during a designated time window, and had 15 minutes of privacy to report daily moods, marital interactions and alcohol involvement. The answers were automatically entered into database. A binary value, the urge to drink, was summarized from the answers in each day as the longitudinal covariate. A continuous scale of depression, the Beck Depression Inventory (BDI; Beck et al., 1996), was measured 6 months after the IVR assessment.

In the study on alcoholic couples, we apply the proposed method in Section 3.4.1 to characterize the change in self-reported urge to drink during the 14 days of IVR assessment, and delineate its time-varying effect on the BDI. In particular, we model the binary longitudinal covariate W_{ikj}, urge to drink in a certain day, as

$$W_{ikj} \sim \text{Bernoulli}(\mu_{ikj}), \tag{3.34}$$

and

$$\text{logit}(\mu_{ikj}) = x_{ik}(t_{ikj}) = x_0(t_{ikj}) + d_{ik}(t_{ikj}). \tag{3.35}$$

The continuous outcome Y_{ik}, BDI, is related to the recruitment setting Z_{ik} (1 for treatment and 0 for community) and $x_{ik}(\cdot)$ through

$$Y_{ik} = Z_{ik}\delta + \int_{T_1}^{T_2} x_{ik}(t)\gamma(t)dt + a_i + \epsilon_{ik}, \tag{3.36}$$

where $a_i \sim \text{N}(0, \sigma_a^2)$ characterizes the random family effect.

Figure 3.4 presents the population covariate trajectory $x_0(\cdot)$ and the time-varying effect $\gamma(\cdot)$. The left panel shows that, although participants' tendency to feel an urge to drink reduces in the first week of the IVR assessment, it rebounds during the second week. This implies a short-term effect of measurement reactivity. The right panel indicates that, the initial level of urge to drink (i.e., before the

measurement reactivity takes effect) is more predictive of the depression outcome measured in 6 months after the IVR assessment.

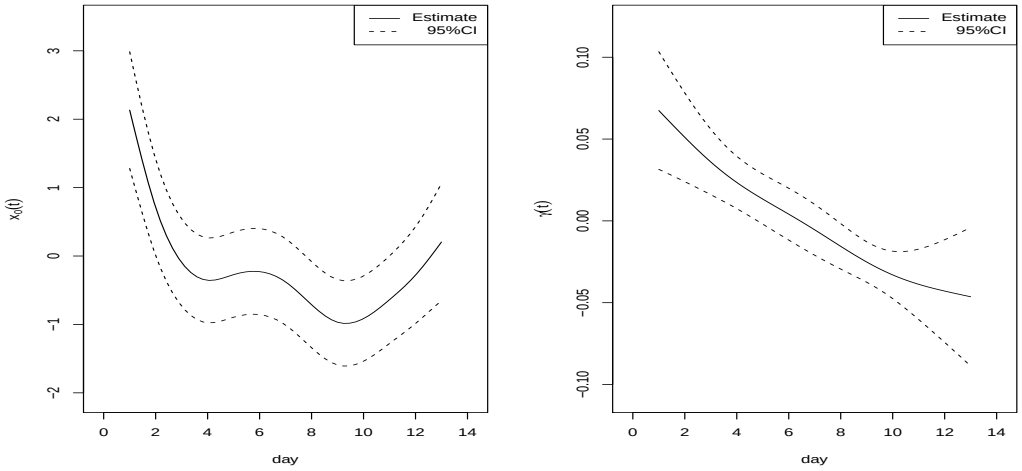

Figure 3.4: Estimated Functions for the Alcoholic Couples Study Dataset: left, the overall developmental trajectory of urge to drink, $x_0(\cdot)$; right, the time-varying effect of urge to drink on depression, $\gamma(\cdot)$

Furthermore, our analysis suggests that the participants recruited from the University of Michigan Addiction Treatment Services have a higher level of depression than those from the community sample ($\hat{\delta} = 0.4105$ with 95% confidence interval of $[0.2657, 0.5552]$). The variance components in the model are estimated as $\hat{\sigma}_1^2 = 2.60^2$, $\hat{\sigma}_2^2 = 1.02^2$, $\hat{\tau}_d = 3.64^2$, $\hat{\sigma}_a^2 = 0.05^2$, $\hat{\sigma}_{\epsilon^*}^2 = 0.48^2$, and the smoothing parameters are selected as $\lambda_x = 0.4$, $\lambda_\gamma = 20$.

3.4.3 Simulation Study of the Extended Model

This simulation experiment is designed to evaluate the performance of the particular model proposed in Section 3.4.1 under different situations, based on the features of the real data used in Section 3.4.2.

Particularly, three factors are manipulated: (1) sample size, $n^{(1)} = 100$, $n^{(2)} =$

200 and $n^{(3)} = 400$; (2) number of time points, $r^{(1)} = 14$, $r^{(2)} = 21$ and $r^{(3)} = 28$; and (3) proportion of zeros in the longitudinal covariates: 50%, 70% and 90%, which can be achieved by adjusting the population profile function $x_0(\cdot)$. The rest of the design is in general based on the data features in Section 3.4.2.

We set $\mathbf{t}^0$ to be r equally spaced time points over interval $[-1.5, 1.5]$, where r is 14, 21 or 28. It is worth noting that, observation time points t_{ikj}'s are not necessarily balanced among all individuals. Particularly, for ik-th subject, the longitudinal covariate process $\mathbf{W}_{ik}$ is observed at $\{t_{ikj}\}$ for $j = 1, \ldots, n_{ik}$, where n_{ik} is an integer randomly chosen from $\{\lceil \frac{4}{5}r \rceil, \ldots, r\}$, and t_{ikj}'s are n_{ik} distinct time points drawn from $\mathbf{t}^0$.

The longitudinal covariates W_{ikj}'s are generated by model (3.34,3.35). Here, we consider three choices of $x_0(\cdot)$, $x_0^{(1)}(t) = 0.8t^4 - t^2 - 0.5t - 0.1$, $x_0^{(2)}(t) = 0.8t^4 - t^2 - 0.5t - 1.3$, and $x_0^{(3)}(t) = 0.8t^4 - t^2 - 0.5t - 3$, which correspond to the proportions of zeros in W_{ikj}'s of about 50%, 70% and 90%, respectively. The random process $d_{ik}(\cdot)$ is determined by $\mathbf{d}_{ik} = \mathbf{B}_* \mathbf{b}_{ik}$ with $\mathbf{b}_{ik}$'s being a random sample from $\mathrm{N}(\mathbf{0}, \mathrm{diag}\{1^2, 0.6^2, 1.5^2 \mathbf{I}_{(r-2)\times(r-2)}\})$. The response data Y_{ik}'s are from model (3.36), where $T_1 = -1.5$, $T_2 = 1.5$, $\delta = 0.4$, $\gamma(t) = -0.6\arctan(0.8t)$, and $a_i \sim \mathrm{N}(0, 0.2^2)$, $e_{ik} \sim \mathrm{N}(0, 0.6^2)$. In addition, in order to generate the time-invariant covariate, Z_{ik}'s are randomly drawn from $\{0, 1\}$ with $\mathrm{P}(Z_{ik} = 1) = 0.5$ for each individual.

In summary, three factors are manipulated, leading to 27 combinations in total. Under each of the situations, $N = 100$ datasets are generated, then the proposed two-stage estimation method is applied. For each parameter, the MSE and its empirical standard error are calculated from the N replications. In terms of the nonparametric functions $x_0(\cdot)$ and $\gamma(\cdot)$, the MISE and its empirical standard error are used to summarize the results.

Table 3.8: Estimation Results of Simulation for the Extended Model: varying n ($r^{(2)} = 21$, $x_0^{(1)}(t)$)

	$n^{(1)} = 100$	$n^{(2)} = 200$	$n^{(3)} = 400$
	MSE/MISE (se)		
δ	0.0171 (0.0224)	0.0115 (0.0147)	0.0047 (0.0067)
σ_1^2	0.0385 (0.0486)	0.0404 (0.0363)	0.0366 (0.0280)
σ_2^2	0.0089 (0.0108)	0.0046 (0.0059)	0.0032 (0.0038)
τ_d	0.7086 (0.9980)	0.4322 (0.6667)	0.2875 (0.3938)
σ_a^2	0.0038 (0.0061)	0.0024 (0.0044)	0.0011 (0.0012)
σ_ϵ^2	0.0196 (0.0225)	0.0185 (0.0167)	0.0243 (0.0140)
$x_0(\cdot)$	0.2454 (0.1768)	0.1445 (0.0689)	0.0965 (0.0437)
$\gamma(\cdot)$	0.0290 (0.0406)	0.0148 (0.0116)	0.0108 (0.0085)

Table 3.9: Estimation Results of Simulation for the Extended Model: varying r ($n^{(2)} = 200$, $x_0^{(1)}(t)$)

	$r^{(1)} = 14$	$r^{(2)} = 21$	$r^{(3)} = 28$
	MSE/MISE (se)		
δ	0.0118 (0.0146)	0.0115 (0.0147)	0.0089 (0.0118)
σ_1^2	0.0515 (0.0543)	0.0404 (0.0363)	0.0311 (0.0325)
σ_2^2	0.0063 (0.0086)	0.0046 (0.0059)	0.0042 (0.0054)
τ_d	0.5026 (0.7004)	0.4322 (0.6667)	0.3069 (0.5239)
σ_a^2	0.0021 (0.0037)	0.0024 (0.0044)	0.0021 (0.0034)
σ_ϵ^2	0.0272 (0.0249)	0.0185 (0.0167)	0.0167 (0.0158)
$x_0(\cdot)$	0.1502 (0.0885)	0.1445 (0.0689)	0.1328 (0.0703)
$\gamma(\cdot)$	0.0195 (0.0204)	0.0148 (0.0116)	0.0123 (0.0099)

The results of the simulation are provided in Tables 3.8, 3.9 and 3.10. Table 3.8 shows the results for the three different sample sizes, holding the other two factors constant at $r^{(2)} = 21$ and $x_0^{(1)}(t) = 0.8t^4 - t^2 - 0.5t - 0.1$. The smaller MSE with a larger sample size indicates that the performance of the proposed method improves as the sample size increases.

Similarly, Table 3.9 demonstrates that the performance of the proposed method is better when the longitudinal covariates are collected more frequently. Table 3.10 shows that, when the binary longitudinal covariates are more concentrated in one value (i.e., contain more 0's in our settings), the proposed method performs less

Table 3.10: Estimation Results of Simulation for the Extended Model: varying $x_0(t)$ ($n^{(2)} = 200$, $r^{(2)} = 21$)

	$x_0^{(1)}(t)$ (50% of zeros)	$x_0^{(2)}(t)$ (70% of zeros)	$x_0^{(3)}(t)$ (90% of zeros)
	MSE/MISE (se)		
δ	0.0115 (0.0147)	0.0118 (0.0143)	0.0142 (0.0155)
σ_1^2	0.0404 (0.0363)	0.0459 (0.0435)	0.0723 (0.0721)
σ_2^2	0.0046 (0.0059)	0.0060 (0.0074)	0.0082 (0.0101)
τ_d	0.4322 (0.6667)	0.3673 (0.5391)	0.8843 (1.6865)
σ_a^2	0.0024 (0.0044)	0.0024 (0.0036)	0.0029 (0.0047)
σ_ϵ^2	0.0185 (0.0167)	0.0231 (0.0202)	0.0309 (0.0241)
$x_0(\cdot)$	0.1445 (0.0689)	0.2169 (0.1629)	0.3147 (0.2074)
$\gamma(\cdot)$	0.0148 (0.0116)	0.0175 (0.0218)	0.0361 (0.0618)

well, compared with a balanced occasion.

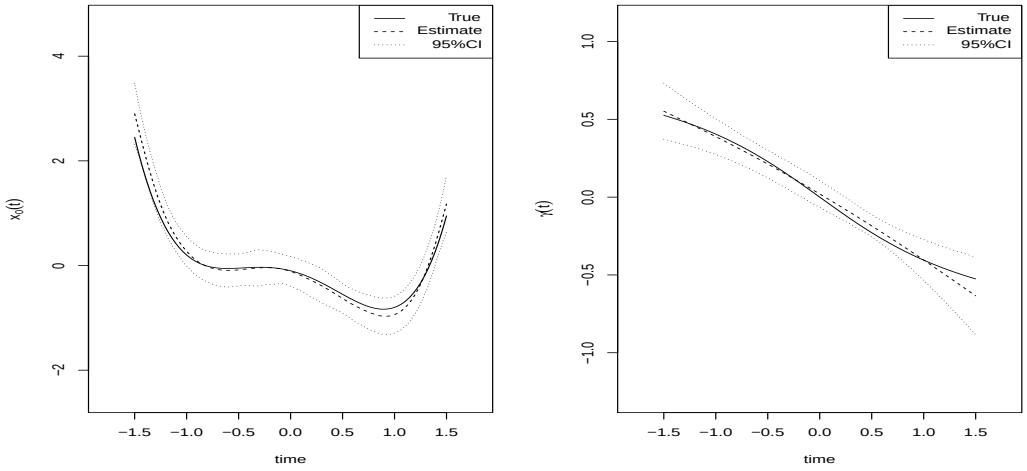

Figure 3.5: Estimated Functions of Simulation for the Extended Model: left, population profile function $x_0(\cdot)$; right, effect function $\gamma(\cdot)$

We also evaluate the performance of the proposed method by comparing the estimated curves of $x_0(\cdot)$ and $\gamma(\cdot)$ with the true curves for each setting. Figure 3.5 demonstrates the estimated curves derived from evaluating the fitted functions of the N replications at a set of grid points and connecting means at all grid points, for the setting of $n^{(2)} = 200$, $r^{(2)} = 21$ and $x_0^{(1)}(t) = 0.8t^4 - t^2 - 0.5t - 0.1$. It

indicates that our proposed method estimates $x_0(\cdot)$ and $\gamma(\cdot)$ well with relatively small bias, and the true functions are covered by the corresponding 95% pointwise empirical confidence intervals.

Chapter 4

Modeling Scalar Response with Discrete Longitudinal Covariates: Theory

In this chapter, we explore some asymptotic properties for the general model proposed in Chapter 3. Particularly, we establish the consistency and asymptotic normality of the estimate of the population profile function in the longitudinal covariate model, by first providing general results from generalized linear mixed models (GLMMs).

However, to best of our knowledge, there remain difficulties in characterizing the properties of the estimates of the regression coefficient and effect function from the calibration model. We leave them as future studies in Chapter 6, along with some discussions on the potential approaches.

We formulate the problem in two steps, and organize this chapter as follows. In Section 4.1, we study the asymptotic properties when the observation time points are scheduled in advance and fixed. Section 4.2 further allows the number of total

observation time points to diverge as sample size increases. The corresponding proofs of certain theorems, lemmas and corollaries are provided at the end of each of Sections 4.1 and 4.2.

4.1 Asymptotic Properties for Fixed Observation Time Points

In this section, we study the asymptotic properties of $\hat{x}_0(\cdot)$, the estimate of the population profile function as in model (3.11), when the possible observation time points are scheduled in advance and so will not change as new subjects enter the study. In particular, we first provide some regularity conditions on the likelihood and derive the asymptotic properties for a GLMM with finite parameters. When a GLMM is fitted via the penalized quasi-likelihood (PQL) as in Breslow and Clayton (1993), some additional conditions are imposed in order to obtain the asymptotic properties. Following the general results, the conditions on our particular longitudinal covariate model, as that in model (3.4,3.5), are specified accordingly, and lead to the asymptotic properties for the estimated population profile function $\hat{x}_0(\cdot)$.

4.1.1 General Results for GLMMs with Finite Parameters

We start from a GLMM involving clustered random effects, when the variance component $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ is known. Typically, for $i = 1, \ldots, n$, $j = 1, \ldots, n_i$, y_{ij}'s are conditionally independent with

$$y_{ij} \sim f(\cdot; \mu_{ij}),$$

and

$$h(\mu_{ij}) = \mathbf{x}_{ij}^T\boldsymbol{\alpha} + \mathbf{z}_{ij}^T\mathbf{b}_i, \tag{4.1}$$

where $\mathbf{x}_{ij}$ and $\mathbf{z}_{ij}$ are vectors of covariates, $\boldsymbol{\alpha}$ is a p-dimensional vector of fixed effects, and $\mathbf{b}_i$ is a q-dimensional vector of random effects following $\mathbf{b}_i \sim q(\cdot;\boldsymbol{\theta}_0)$. We denote $\mathbf{X}_i = (\mathbf{x}_{i1},\ldots,\mathbf{x}_{in_i})^T$, $\mathbf{Z}_i = (\mathbf{z}_{i1},\ldots,\mathbf{z}_{in_i})^T$, $\mathbf{Y}_i = (y_{i1},\ldots,y_{in_i})^T$, and $V_i = \{\mathbf{X}_i, \mathbf{Z}_i, \mathbf{Y}_i\}$, $\mathbf{V} = \{V_1,\ldots,V_n\}$. The likelihood is

$$L(\boldsymbol{\alpha}) = \prod_i L_i(V_i,\boldsymbol{\alpha}) = \prod_i \int \prod_j f(y_{ij}; h^{-1}(\mathbf{x}_{ij}^T\boldsymbol{\alpha} + \mathbf{z}_{ij}^T\mathbf{b}_i))q(\mathbf{b}_i;\boldsymbol{\theta}_0)d\mathbf{b}_i, \tag{4.2}$$

and the log-likelihood is $l(\boldsymbol{\alpha}) = \sum_i l_i(\boldsymbol{\alpha})$, where

$$l_i(\boldsymbol{\alpha}) = \log(L_i(V_i,\boldsymbol{\alpha})) = \log(\int \prod_j f(y_{ij}; h^{-1}(\mathbf{x}_{ij}^T\boldsymbol{\alpha} + \mathbf{z}_{ij}^T\mathbf{b}_i))q(\mathbf{b}_i;\boldsymbol{\theta}_0)d\mathbf{b}_i).$$

The coefficient vector $\boldsymbol{\alpha}$ is estimated by maximizing $l(\boldsymbol{\alpha})$. In order to derive asymptotic properties for $\hat{\boldsymbol{\alpha}}$, we impose following conditions on each of $l_i(\boldsymbol{\alpha})$'s.

(A1) The likelihood $L_i(V_i,\boldsymbol{\alpha})$ has a common support and the model is identifiable; moreover, $\mathrm{E}(\nabla_{\boldsymbol{\alpha}} l_i) = \mathbf{0}$, $\mathrm{E}(\nabla_{\boldsymbol{\alpha}} l_i \nabla_{\boldsymbol{\alpha}}^T l_i) = \mathrm{E}(-\nabla_{\boldsymbol{\alpha}}^2 l_i)$.

(A2) The information matrix $\mathbf{I}_i(\boldsymbol{\alpha}) = \mathrm{E}(\nabla_{\boldsymbol{\alpha}} l_i \nabla_{\boldsymbol{\alpha}}^T l_i)$ is finite, and

$$0 < \lambda_{\min}(\mathbf{I}_i(\boldsymbol{\alpha}_0)) \le \lambda_{\max}(\mathbf{I}_i(\boldsymbol{\alpha}_0)) \le c_1 < \infty,$$

where $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ denote the smallest and largest eigenvalues, respectively; in addition, $n^{-1}\sum_i \mathbf{I}_i(\boldsymbol{\alpha}_0) \to \bar{\mathbf{I}}_0$, where $\bar{\mathbf{I}}_0$ is positive definite.

(A3) For every (r,s),

$$\mathrm{E}((\frac{\partial}{\partial\alpha_r} l_i(\boldsymbol{\alpha}_0)\frac{\partial}{\partial\alpha_s} l_i(\boldsymbol{\alpha}_0))^2) \le c_2 < \infty.$$

(A4) For every (r, s),

$$\mathrm{E}((\frac{\partial^2}{\partial\alpha_r\partial\alpha_s}l_i(\boldsymbol{\alpha}_0))^2) \leq c_3 < \infty.$$

(A5) There exists an open subset ω of $\boldsymbol{\alpha}_0 \in \omega \subset \Omega$, for almost all V_i, there exist functions $M_{rst}(\cdot)$'s such that

$$|\frac{\partial^3}{\partial\alpha_r\partial\alpha_s\partial\alpha_t}l_i(V_i, \boldsymbol{\alpha})| \leq M_{rst}(V_i),$$

for all $\boldsymbol{\alpha} \in \omega$ and every (r, s, t), where $\mathrm{E}(M_{rst}(V_i)) < \infty$.

Conditions (A1) - (A4) are imposed on the first and second derivatives of each $l_i(\boldsymbol{\alpha})$ with respect to $\boldsymbol{\alpha}$, and lead to properties regarding $\nabla_{\boldsymbol{\alpha}} l(\boldsymbol{\alpha}_0)$ and $\nabla^2_{\boldsymbol{\alpha}} l(\boldsymbol{\alpha}_0)$ as stated in Lemma 4.1. These conditions, along with an additional one on the third derivative of $l_i(\boldsymbol{\alpha})$, guarantee the consistency and asymptotic normality of the local maximizer $\hat{\boldsymbol{\alpha}}$ of $l(\boldsymbol{\alpha})$.

Under these conditions, we first have following conclusions regarding $\nabla_{\boldsymbol{\alpha}} l(\boldsymbol{\alpha}_0)$ and $\nabla^2_{\boldsymbol{\alpha}} l(\boldsymbol{\alpha}_0)$.

Lemma 4.1. *If conditions (A1) - (A4) hold, the gradient vector and Hessian matrix of $l(\boldsymbol{\alpha}_0)$ satisfy*

$$\frac{1}{\sqrt{n}}\nabla_{\boldsymbol{\alpha}} l(\boldsymbol{\alpha}_0) \xrightarrow{d} N(\mathbf{0}, \bar{\mathbf{I}}_0),$$

and

$$-\frac{1}{n}\nabla^2_{\boldsymbol{\alpha}} l(\boldsymbol{\alpha}_0) \xrightarrow{P} \bar{\mathbf{I}}_0.$$

Proof. A proof is given in Section 4.1.3. □

With notation $\mathbf{A} = O_P(\mathbf{B})$ such that $\mathbf{A}_{(r,s)} = O_P(\mathbf{B}_{(r,s)})$ for every (r, s), and $\mathbf{A} = o_P(\mathbf{B})$ similarly defined, we have equivalent conclusions that $n^{-\frac{1}{2}}\nabla_{\boldsymbol{\alpha}} l(\boldsymbol{\alpha}_0) = O_P(\mathbf{1})$ and $-n^{-1}\nabla^2_{\boldsymbol{\alpha}} l(\boldsymbol{\alpha}_0) = \bar{\mathbf{I}}_0 + o_P(\mathbf{1}\mathbf{1}^T)$.

We then derive asymptotic properties of the local maximizer $\hat{\boldsymbol{\alpha}}$.

Theorem 4.2. *Suppose that a GLMM involving clustered random effect has likelihood as equation (4.2). If $l_i(\boldsymbol{\alpha})$'s satisfy conditions (A1) - (A5), there exists a local maximizer $\hat{\boldsymbol{\alpha}}$ of $l(\boldsymbol{\alpha})$ such that*

a) (consistency)

$$\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0\| = O_P(\frac{1}{\sqrt{n}});$$

b) (asymptotic normality)

$$\sqrt{n}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0) \xrightarrow{d} N(\mathbf{0}, \bar{\mathbf{I}}_0^{-1}).$$

Proof. A proof is given in Section 4.1.3. □

Remark 1. When the variance component $\boldsymbol{\theta}$ is unknown and needs to be estimated from the data $\mathbf{V}$, we instead have parameter $\boldsymbol{\zeta} = (\boldsymbol{\alpha}^T, \boldsymbol{\theta}^T)^T$, and log-likelihood $l(\boldsymbol{\zeta}) = \sum_i l_i(\boldsymbol{\zeta})$ where

$$l_i(\boldsymbol{\zeta}) = \log(\int \prod_j f(y_{ij}; h^{-1}(\mathbf{x}_{ij}^T\boldsymbol{\alpha} + \mathbf{z}_{ij}^T\mathbf{b}_i))q(\mathbf{b}_i; \boldsymbol{\theta})d\mathbf{b}_i).$$

The estimate $\hat{\boldsymbol{\zeta}}$ can be obtained by maximizing $l(\boldsymbol{\zeta})$, and has properties as stated in Corollary 4.3.

Corollary 4.3. *If conditions (A1) - (A5) still hold with respect to $l_i(\boldsymbol{\zeta})$'s, there exists a local maximizer $\hat{\boldsymbol{\zeta}}$ of $l(\boldsymbol{\zeta})$ such that*

a) (consistency)

$$\|\hat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}_0\| = O_P(\frac{1}{\sqrt{n}});$$

b) (asymptotic normality)

$$\sqrt{n}(\hat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}_0) \xrightarrow{d} N(\mathbf{0}, \bar{\mathbf{I}}_0^{-1}),$$

here $n^{-1}\sum_i \mathbf{I}_i(\boldsymbol{\zeta}_0) \to \bar{\mathbf{I}}_0$.

Remark 2. In particular, if V_i's are independent and identically distributed with $L_*(V_*, \boldsymbol{\alpha})$, which requires $n_i = n_*$ for each i, we will instead consider following conditions.

(A1a) The likelihood $L_*(V_*, \boldsymbol{\alpha})$ has a common support and the model is identifiable; moreover, $\mathrm{E}(\nabla_{\boldsymbol{\alpha}} l_*) = \mathbf{0}$, $\mathrm{E}(\nabla_{\boldsymbol{\alpha}} l_* \nabla_{\boldsymbol{\alpha}}^T l_*) = \mathrm{E}(-\nabla_{\boldsymbol{\alpha}}^2 l_*)$.

(A2a) The Fisher information matrix $\mathbf{I}(\boldsymbol{\alpha}) = \mathrm{E}(\nabla_{\boldsymbol{\alpha}} l_* \nabla_{\boldsymbol{\alpha}}^T l_*)$ is finite, and $\mathbf{I}_0 = \mathbf{I}(\boldsymbol{\alpha}_0)$ is positive definite.

(A3a) There exists an open subset ω of $\boldsymbol{\alpha}_0 \in \omega \subset \Omega$, for almost all V_*, there exist functions $M_{rst}(\cdot)$'s such that

$$\left|\frac{\partial^3}{\partial\alpha_r \partial\alpha_s \partial\alpha_t} l_*(V_*, \boldsymbol{\alpha})\right| \leq M_{rst}(V_*),$$

for all $\boldsymbol{\alpha} \in \omega$ and every (r, s, t), where $\mathrm{E}(M_{rst}(V_*)) < \infty$.

Conditions (A1a) - (A3a) are the usual regularity conditions imposed on the likelihood (compared with Lehmann and Casella (1998)), and lead to following conclusions.

Theorem 4.4. *For a GLMM with likelihood as equation (4.2) when* $n_i = n_*$ *for each* i *(i.e.,* V_i*'s are independent and identically distributed with* $L_*(V_*, \boldsymbol{\alpha})$*), if* $l_*(\boldsymbol{\alpha}) = \log(L_*(V_*, \boldsymbol{\alpha}))$ *satisfies conditions (A1a) - (A3a), there exists a local maximizer* $\hat{\boldsymbol{\alpha}}$ *of* $l(\boldsymbol{\alpha})$ *such that*

a) (consistency)

$$\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0\| = O_P(\frac{1}{\sqrt{n}});$$

b) (asymptotic normality)

$$\sqrt{n}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0) \xrightarrow{d} N(\mathbf{0}, \mathbf{I}_0^{-1}).$$

Remark 3. Breslow and Clayton (1993) propose an approximate procedure to estimate $\boldsymbol{\alpha}$ by maximizing PQL $l_B(\boldsymbol{\alpha})$ instead of $l(\boldsymbol{\alpha})$; and Lin and Breslow (1996) point out that this approximation brings a difference of $B(\boldsymbol{\alpha})$ to $l(\boldsymbol{\alpha})$ such that

$$l_B(\boldsymbol{\alpha}) = l(\boldsymbol{\alpha}) - B(\boldsymbol{\alpha}).$$

For a small $\boldsymbol{\theta}_0$, a quadratic expansion of $l(\boldsymbol{\alpha})$ gives

$$B(\boldsymbol{\alpha}) = \mathbf{l}_{12}^T\boldsymbol{\theta}_0 + \frac{1}{2}\boldsymbol{\theta}_0^T(\mathbf{l}_{22} + \mathbf{l}_{23} + \mathbf{l}_{24})\boldsymbol{\theta}_0 + o(\|\boldsymbol{\theta}_0\|^2),$$

where vector $\mathbf{l}_{12}$ and matrices $\mathbf{l}_{22}$, $\mathbf{l}_{23}$, $\mathbf{l}_{24}$ all depend on $\boldsymbol{\alpha}$, with the explicit forms defined in Section 3 of Lin and Breslow (1996).

In order to derive the asymptotic properties of $\hat{\boldsymbol{\alpha}}_B$, the maximizer of $l_B(\boldsymbol{\alpha})$, we additionally consider following conditions on $B(\boldsymbol{\alpha})$.

(B1) $v_{Bn}^{-1}\nabla_{\boldsymbol{\alpha}}B(\boldsymbol{\alpha}_0) = O_P(\mathbf{1})$.

(B2) $n^{-1}\nabla_{\boldsymbol{\alpha}}^2 B(\boldsymbol{\alpha}_0) = o_P(\mathbf{1}\mathbf{1}^T)$.

(B3) There exists an open subset ω_B of $\omega \subset \omega_B \subset \Omega$, for almost all $\mathbf{V}$, there exist functions $M_{Brst}(\cdot)$'s such that

$$\left|\frac{\partial^3}{\partial\alpha_r\partial\alpha_s\partial\alpha_t}B(\mathbf{V}, \boldsymbol{\alpha})\right| \leq M_{Brst}(\mathbf{V}),$$

for all $\boldsymbol{\alpha} \in \omega_B$ and every (r, s, t), where $\mathrm{E}(M_{Brst}(\mathbf{V})) = o(n^2(n^{\frac{1}{2}} + v_{Bn})^{-1})$.

Conditions (B1) - (B3) are imposed on the first, second and third derivatives of $B(\boldsymbol{\alpha})$ respectively, in a similar way as those on $l(\boldsymbol{\alpha})$ by conditions (A1) - (A5) or (A1a) - (A3a). These conditions will guarantee the consistency and asymptotic normality of the local maximizer $\hat{\boldsymbol{\alpha}}_B$ of $l_B(\boldsymbol{\alpha})$. Moreover, v_{Bn} will influence the consistency rate of $\hat{\boldsymbol{\alpha}}_B$.

We then have Corollary 4.5 when dealing with $l_B(\boldsymbol{\alpha})$ instead of $l(\boldsymbol{\alpha})$.

Corollary 4.5. *Suppose that a GLMM problem is solved through an approximate procedure, that is, maximizing the PQL $l_B(\boldsymbol{\alpha})$, if $l_i(\boldsymbol{\alpha})$'s satisfy conditions (A1) - (A5) and $B(\boldsymbol{\alpha})$ satisfies conditions (B1) - (B3), there exists a local maximizer $\hat{\boldsymbol{\alpha}}_B$ of $l_B(\boldsymbol{\alpha})$ such that*

a) (consistency)

$$\|\hat{\boldsymbol{\alpha}}_B - \boldsymbol{\alpha}_0\| = O_P(\frac{1}{\sqrt{n}} + \frac{v_{Bn}}{n});$$

b) (asymptotic normality)

$$\sqrt{n}(\bar{\mathbf{I}}_0 + \frac{1}{n}\nabla^2_{\boldsymbol{\alpha}} B(\boldsymbol{\alpha}_0))(\hat{\boldsymbol{\alpha}}_B - \boldsymbol{\alpha}_0^*) \xrightarrow{d} N(\mathbf{0}, \bar{\mathbf{I}}_0),$$

where $\boldsymbol{\alpha}_0^* = \boldsymbol{\alpha}_0 - (\bar{\mathbf{I}}_0 + n^{-1}\nabla^2_{\boldsymbol{\alpha}} B(\boldsymbol{\alpha}_0))^{-1}(n^{-1}\nabla_{\boldsymbol{\alpha}} B(\boldsymbol{\alpha}_0))$.

Proof. A proof is given in Section 4.1.3. □

We then consider a GLMM as model (4.1), but a penalty term is imposed on $l(\boldsymbol{\alpha})$ to characterize its "smoothness". It is noted that, a special form of such model gives the GAMM studied by Lin and Zhang (1999). When a quadratic penalty term is adopted, $\boldsymbol{\alpha}$ is estimated by maximizing the penalized log-likelihood

$$l_P(\boldsymbol{\alpha}) = l(\boldsymbol{\alpha}) - \lambda \boldsymbol{\alpha}^T \mathbf{K} \boldsymbol{\alpha}, \tag{4.3}$$

where λ is a smoothing parameter assumed to be fixed for now, and $\mathbf{K}$ is a smooth-

ing matrix similarly defined as in Section 3.1.2.

We have the asymptotic properties for the local maximizer $\hat{\boldsymbol{\alpha}}_P$ of $l_P(\boldsymbol{\alpha})$.

Corollary 4.6. *For a GLMM problem with a quadratic penalty, which has the penalized log-likelihood as equation (4.3), if $l_i(\boldsymbol{\alpha})$'s satisfy conditions (A1) - (A5), there exists a local maximizer $\hat{\boldsymbol{\alpha}}_P$ of $l_P(\boldsymbol{\alpha})$ such that*

a) (consistency)

$$\|\hat{\boldsymbol{\alpha}}_P - \boldsymbol{\alpha}_0\| = O_P(n^{-\frac{1}{2}});$$

b) (asymptotic normality)

$$\sqrt{n}(\bar{\mathbf{I}}_0 + \frac{2}{n}\lambda\mathbf{K})(\hat{\boldsymbol{\alpha}}_P - \boldsymbol{\alpha}_0^*) \xrightarrow{d} N(\mathbf{0}, \bar{\mathbf{I}}_0),$$

where $\boldsymbol{\alpha}_0^* = \boldsymbol{\alpha}_0 - (\bar{\mathbf{I}}_0 + 2n^{-1}\lambda\mathbf{K})^{-1}(2n^{-1}\lambda\mathbf{K}\boldsymbol{\alpha}_0)$.

Proof. A proof is given in Section 4.1.3. □

4.1.2 Results for Longitudinal Covariate Model with Fixed Observation Time Points

For the problem in Section 3.2.1, the population profile $x_0(\cdot)$ may be any continuous function on the interval $[T_1, T_2]$; in particular, we assume that $x_0(\cdot)$ belongs to $W^2([T_1, T_2])$, a 2-order Sobolev space such that $x_0^{'}(\cdot)$ is absolutely continuous, and $\|x_0^{''}\|^2 = \int_{T_1}^{T_2}(x_0^{''}(t))^2 dt < \infty$. It is usually practicable to approximate $x_0(\cdot)$ by a linear combination of a set of appropriate basis functions. Typically, natural cubic spline (NCS) is adopted for the approximation in a way that $x_{00}(\cdot) \approx x_{00}^{\mathrm{NCS}}(\cdot) = \mathbf{x}_{00}^T\mathbf{c}(\cdot)$, where $x_{00}(\cdot)$ is the true function, $\mathbf{c}(\cdot)$ is a set of NCS basis functions as defined in Section 3.1.2, and $\mathbf{x}_{00} = x_{00}^{\mathrm{NCS}}(\mathbf{t}^0) = x_{00}(\mathbf{t}^0)$. Under this approximation, the asymptotic properties of $\hat{x}_0(\cdot)$ can be studied through $\hat{\mathbf{x}}_0$.

In particular, we have the log-quasi-likelihood

$$ql(\mathbf{x}_0) = \sum_{i=1}^{n} ql_i(\mathbf{x}_0) = \sum_{i=1}^{n} \log(\int \exp(-\frac{1}{2}\sum_{j=1}^{n_i} \tilde{d}_{ij} - \frac{1}{2}\mathbf{b}_i^T \mathbf{D}^{-1}(\boldsymbol{\theta})\mathbf{b}_i) d\mathbf{b}_i),$$

where $\tilde{d}_{ij}$ is the conditional deviance defined as equation (3.13). A smoothing spline technique is applied on $x_0(\cdot)$ by including a roughness penalty term

$$\frac{1}{2}\lambda_x \int_{T_1}^{T_2} (x_0''(t))^2 dt = \frac{1}{2}\lambda_x \mathbf{x}_0^T \mathbf{K} \mathbf{x}_0,$$

where λ_x is a smoothing parameter and $\mathbf{K}$ is a smoothing matrix. Thus, $\mathbf{x}_0$ is estimated by maximizing the penalized log-quasi-likelihood

$$ql_P(\mathbf{x}_0) = ql(\mathbf{x}_0) - \frac{1}{2}\lambda_x \mathbf{x}_0^T \mathbf{K} \mathbf{x}_0.$$

In order to derive asymptotic properties for $\hat{\mathbf{x}}_0$, we impose following conditions on $ql_i(\mathbf{x}_0)$'s.

(H1) The distribution has a common support and the model is identifiable; moreover, $\mathrm{E}(\nabla_{\mathbf{x}} ql_i) = \mathbf{0}$, $\mathrm{E}(\nabla_{\mathbf{x}} ql_i \nabla_{\mathbf{x}}^T ql_i) = \mathrm{E}(-\nabla_{\mathbf{x}}^2 ql_i)$.

(H2) The information matrix $\mathbf{I}_i(\mathbf{x}_0) = \mathrm{E}(\nabla_{\mathbf{x}} ql_i \nabla_{\mathbf{x}}^T ql_i)$ is finite, and in particular,

$$0 < \lambda_{\min}(\mathbf{I}_i(x_{00}(\mathbf{t}^0))) \leq \lambda_{\max}(\mathbf{I}_i(x_{00}(\mathbf{t}^0))) \leq c_1 < \infty;$$

in addition, $n^{-1}\sum_i \mathbf{I}_i(x_{00}(\mathbf{t}^0)) \to \bar{\mathbf{I}}_0$, where $\bar{\mathbf{I}}_0$ is positive definite.

(H3) For every (r, s),

$$\mathrm{E}((\frac{\partial}{\partial x_r} l_i(x_{00}(\mathbf{t}^0)) \frac{\partial}{\partial x_s} l_i(x_{00}(\mathbf{t}^0)))^2) \leq c_2 < \infty.$$

(H4) For every (r, s),

$$\mathrm{E}((\frac{\partial^2}{\partial x_r \partial x_s} l_i(x_{00}(\mathbf{t}^0)))^2) \leq c_3 < \infty.$$

(H5) There exists an open subset ω of $x_{00}(\mathbf{t}^0) \in \omega \subset \Omega \subset \mathbb{R}^r$, for almost all V_i, there exist functions $M_{rst}(\cdot)$'s such that

$$|\frac{\partial^3}{\partial x_r \partial x_s \partial x_t} q l_i(V_i, \mathbf{x}_0)| \leq M_{rst}(V_i),$$

for all $\mathbf{x}_0 \in \omega$ and every (r, s, t), where $\mathrm{E}(M_{rst}(V_i)) < \infty$.

These conditions are comparable to conditions (A1) - (A5). Followed by the arguments in Section 4.1.1, we have conclusions on the asymptotic properties of $\hat{\mathbf{x}}_0$.

Theorem 4.7. *For a GAMM estimation problem as model (3.4,3.5), the true population profile function $x_{00}(\cdot)$ belongs to $W^2([T_1, T_2])$, and the properties of its projection in $\mathcal{S}_{\mathrm{NCS}}(\mathbf{t}^0)$ are studied. If conditions (H1) - (H5) are satisfied, there exists a local maximizer $\hat{\mathbf{x}}_0$ of $ql_P(\mathbf{x}_0)$ such that*

a) (consistency)

$$\|\hat{\mathbf{x}}_0 - x_{00}(\mathbf{t}^0)\| = O_P(\frac{1}{\sqrt{n}});$$

b) (asymptotic normality)

$$\sqrt{n}(\bar{\mathbf{I}}_0 + \frac{2}{n}\lambda_x \mathbf{K})(\hat{\mathbf{x}}_0 - \mathbf{x}_{00}^*) \xrightarrow{d} N(\mathbf{0}, \bar{\mathbf{I}}_0),$$

where $\mathbf{x}_{00}^* = x_{00}(\mathbf{t}^0) - (\bar{\mathbf{I}}_0 + 2n^{-1}\lambda_x \mathbf{K})^{-1}(2n^{-1}\lambda_x \mathbf{K} x_{00}(\mathbf{t}^0))$.

4.1.3 Proofs

In this section, we provide rigorous proofs of Lemma 4.1, Theorem 4.2, Corollary 4.5, and Corollary 4.6. The Euclidean norm and Frobenius norm are adopted for vectors and matrices, respectively.

Proof of Lemma 4.1. First, for any $\varepsilon > 0$, we have

$$\begin{aligned}
\mathrm{E}(\|\frac{1}{\sqrt{n}}\nabla_{\boldsymbol{\alpha}} l_i(\boldsymbol{\alpha}_0)\|^4) &= \frac{1}{n^2}\mathrm{tr}(\mathrm{E}(\nabla^T_{\boldsymbol{\alpha}} l_i(\boldsymbol{\alpha}_0)\nabla_{\boldsymbol{\alpha}} l_i(\boldsymbol{\alpha}_0)\nabla^T_{\boldsymbol{\alpha}} l_i(\boldsymbol{\alpha}_0)\nabla_{\boldsymbol{\alpha}} l_i(\boldsymbol{\alpha}_0))) \\
&= \frac{1}{n^2}\mathrm{tr}(\mathrm{E}(\nabla_{\boldsymbol{\alpha}} l_i(\boldsymbol{\alpha}_0)\nabla^T_{\boldsymbol{\alpha}} l_i(\boldsymbol{\alpha}_0)\nabla_{\boldsymbol{\alpha}} l_i(\boldsymbol{\alpha}_0)\nabla^T_{\boldsymbol{\alpha}} l_i(\boldsymbol{\alpha}_0))) \\
&= \frac{1}{n^2}\sum_r \mathrm{E}(\sum_s (\frac{\partial}{\partial \alpha_r} l_i(\boldsymbol{\alpha}_0)\frac{\partial}{\partial \alpha_s} l_i(\boldsymbol{\alpha}_0))^2) \\
&= \frac{1}{n^2}\sum_r \sum_s \mathrm{E}((\frac{\partial}{\partial \alpha_r} l_i(\boldsymbol{\alpha}_0)\frac{\partial}{\partial \alpha_s} l_i(\boldsymbol{\alpha}_0))^2) \\
&\leq \frac{c_2 p^2}{n^2},
\end{aligned}$$

and

$$\mathrm{P}(\|\frac{1}{\sqrt{n}}\nabla_{\boldsymbol{\alpha}} l_i(\boldsymbol{\alpha}_0)\| > \varepsilon) \leq \frac{\mathrm{E}(\|\nabla_{\boldsymbol{\alpha}} l_i(\boldsymbol{\alpha}_0)\|^2)}{\varepsilon^2 n} = \frac{\mathrm{tr}(\mathbf{I}_i(\boldsymbol{\alpha}_0))}{\varepsilon^2 n} \leq \frac{c_1 p}{\varepsilon^2 n},$$

then

$$\begin{aligned}
&\sum_i \mathrm{E}(\|\frac{1}{\sqrt{n}}\nabla_{\boldsymbol{\alpha}} l_i(\boldsymbol{\alpha}_0)\|^2 \mathrm{I}(\|\frac{1}{\sqrt{n}}\nabla_{\boldsymbol{\alpha}} l_i(\boldsymbol{\alpha}_0)\| > \varepsilon)) \\
&\leq \sum_i \sqrt{\mathrm{E}(\|\frac{1}{\sqrt{n}}\nabla_{\boldsymbol{\alpha}} l_i(\boldsymbol{\alpha}_0)\|^4)\mathrm{P}(\|\frac{1}{\sqrt{n}}\nabla_{\boldsymbol{\alpha}} l_i(\boldsymbol{\alpha}_0)\| > \varepsilon)} \\
&\leq \sum_i \sqrt{\frac{c_2 p^2}{n^2}\frac{c_1 p}{\varepsilon^2 n}} \\
&= o(1).
\end{aligned}$$

Thus, $n^{-\frac{1}{2}}\nabla_{\boldsymbol{\alpha}} l_i(\boldsymbol{\alpha}_0)$'s satisfy the conditions of multivariate Lindeberg-Feller central

limit theorem (van der Vaart, 1998), which yields

$$\frac{1}{\sqrt{n}}\nabla_{\boldsymbol{\alpha}} l(\boldsymbol{\alpha}_0) = \frac{1}{\sqrt{n}}\sum_i \nabla_{\boldsymbol{\alpha}} l_i(\boldsymbol{\alpha}_0) \xrightarrow{d} \mathrm{N}(\mathbf{0}, \bar{\mathbf{I}}_0).$$

Next, with every (r, s), we have that, for any $\varepsilon > 0$,

$$\begin{aligned}
&\mathrm{P}(|\frac{1}{n}\frac{\partial^2}{\partial\alpha_r\partial\alpha_s} l(\boldsymbol{\alpha}_0) + \frac{1}{n}\sum_i \mathbf{I}_{i(r,s)}(\boldsymbol{\alpha}_0)| > \varepsilon) \\
&\leq \frac{1}{n^2\varepsilon^2}\mathrm{E}((\frac{\partial^2}{\partial\alpha_r\partial\alpha_s} l(\boldsymbol{\alpha}_0) + \sum_i \mathbf{I}_{i(r,s)}(\boldsymbol{\alpha}_0))^2) \\
&= \frac{1}{n^2\varepsilon^2}\mathrm{E}(\{\sum_i (\frac{\partial^2}{\partial\alpha_r\partial\alpha_s} l_i(\boldsymbol{\alpha}_0) + \mathbf{I}_{i(r,s)}(\boldsymbol{\alpha}_0))\}^2) \\
&= \frac{1}{n^2\varepsilon^2}\sum_i \{\mathrm{E}((\frac{\partial^2}{\partial\alpha_r\partial\alpha_s} l_i(\boldsymbol{\alpha}_0))^2) - (\mathrm{E}(\frac{\partial^2}{\partial\alpha_r\partial\alpha_s} l_i(\boldsymbol{\alpha}_0)))^2\} \\
&\to 0,
\end{aligned}$$

which means

$$\frac{1}{n}\frac{\partial^2}{\partial\alpha_r\partial\alpha_s} l(\boldsymbol{\alpha}_0) + \frac{1}{n}\sum_i \mathbf{I}_{i(r,s)}(\boldsymbol{\alpha}_0) \xrightarrow{P} 0,$$

and

$$-\frac{1}{n}\frac{\partial^2}{\partial\alpha_r\partial\alpha_s} l(\boldsymbol{\alpha}_0) \xrightarrow{P} \bar{\mathbf{I}}_{0(r,s)},$$

thus,

$$-\frac{1}{n}\nabla^2_{\boldsymbol{\alpha}} l(\boldsymbol{\alpha}_0) \xrightarrow{P} \bar{\mathbf{I}}_0.$$

□

Proof of Theorem 4.2. First, we show that, for any given $\epsilon > 0$, there exists a constant C such that

$$\mathrm{P}(\sup_{\|\mathbf{u}\|=C} l(\boldsymbol{\alpha}_0 + \frac{1}{\sqrt{n}}\mathbf{u}) < l(\boldsymbol{\alpha}_0)) \geq 1 - \epsilon, \tag{4.4}$$

which implies that, with probability at least $1-\epsilon$, there exists a local maximizer in the ball $\{\boldsymbol{\alpha}_0 + n^{-\frac{1}{2}}\mathbf{u} : \|\mathbf{u}\| \leq C\}$, or $\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0\| = O_P(n^{-\frac{1}{2}})$.

By the Taylor expansion, we have

$$\begin{aligned} D_n(\mathbf{u}) &= l(\boldsymbol{\alpha}_0 + \frac{1}{\sqrt{n}}\mathbf{u}) - l(\boldsymbol{\alpha}_0) \\ &= \frac{1}{\sqrt{n}}\nabla^T_{\boldsymbol{\alpha}} l(\boldsymbol{\alpha}_0)\mathbf{u} + \frac{1}{2n}\mathbf{u}^T\nabla^2_{\boldsymbol{\alpha}} l(\boldsymbol{\alpha}_0)\mathbf{u} + \frac{1}{6n^{\frac{3}{2}}}\nabla^T_{\boldsymbol{\alpha}}(\mathbf{u}^T\nabla^2_{\boldsymbol{\alpha}} l(\boldsymbol{\alpha}^*)\mathbf{u})\mathbf{u}, \end{aligned}$$

where $\boldsymbol{\alpha}^*$ is between $\boldsymbol{\alpha}_0$ and $\boldsymbol{\alpha}_0 + n^{-\frac{1}{2}}\mathbf{u}$. From Lemma 4.1, we have

$$\frac{1}{\sqrt{n}}\nabla^T_{\boldsymbol{\alpha}} l(\boldsymbol{\alpha}_0)\mathbf{u} = O_P(1),$$

and

$$-\frac{1}{2n}\mathbf{u}^T\nabla^2_{\boldsymbol{\alpha}} l(\boldsymbol{\alpha}_0)\mathbf{u} = \frac{1}{2}\mathbf{u}^T\bar{\mathbf{I}}_0\mathbf{u}(1 + o_P(1)) \asymp_P 1.$$

We also have

$$\begin{aligned} &|\frac{1}{6n^{\frac{3}{2}}}\nabla^T_{\boldsymbol{\alpha}}(\mathbf{u}^T\nabla^2_{\boldsymbol{\alpha}} l(\boldsymbol{\alpha}^*)\mathbf{u})\mathbf{u}| \\ &= |\frac{1}{6n^{\frac{3}{2}}}\sum_r\sum_s\sum_t\sum_i(\frac{\partial^3}{\partial\alpha_r\partial\alpha_s\partial\alpha_t}l_i(\boldsymbol{\alpha}^*)u_r u_s u_t)| \\ &\leq \frac{1}{6n^{\frac{3}{2}}}\|\mathbf{u}\|^3\sum_r\sum_s\sum_t\sum_i M_{rst}(V_i) \\ &= O(\frac{1}{\sqrt{n}}) = o(1), \end{aligned}$$

then for any $\varepsilon > 0$,

$$\mathrm{P}(|\frac{1}{6n^{\frac{3}{2}}}\nabla^T_{\boldsymbol{\alpha}}(\mathbf{u}^T\nabla^2_{\boldsymbol{\alpha}} l(\boldsymbol{\alpha}^*)\mathbf{u})\mathbf{u}| > \varepsilon) \leq \frac{1}{\varepsilon}\mathrm{E}(|\frac{1}{6n^{\frac{3}{2}}}\nabla^T_{\boldsymbol{\alpha}}(\mathbf{u}^T\nabla^2_{\boldsymbol{\alpha}} l(\boldsymbol{\alpha}^*)\mathbf{u})\mathbf{u}|) \to 0,$$

which means

$$\frac{1}{6}n^{-\frac{3}{2}}\nabla^T_{\boldsymbol{\alpha}}(\mathbf{u}^T\nabla^2_{\boldsymbol{\alpha}} l(\boldsymbol{\alpha}^*)\mathbf{u})\mathbf{u} = o_P(1).$$

Thus, the second term of $D_n(\mathbf{u})$ dominates all others uniformly in $\|\mathbf{u}\| = C$, which means, for a sufficiently large C, $D_n(\mathbf{u}) < 0$ and inequality (4.4) holds.

Next, since $\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0\| = O_P(n^{-\frac{1}{2}})$, we have, by the Taylor expansion,

$$\begin{aligned}
\mathbf{0} &= \nabla_{\boldsymbol{\alpha}} l(\hat{\boldsymbol{\alpha}}) \\
&= \nabla_{\boldsymbol{\alpha}} l(\boldsymbol{\alpha}_0) + \nabla^2_{\boldsymbol{\alpha}} l(\boldsymbol{\alpha}_0)(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0) + \frac{1}{2}\nabla_{\boldsymbol{\alpha}}\{(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0)^T \nabla^2_{\boldsymbol{\alpha}} l(\boldsymbol{\alpha}^*)(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0)\},
\end{aligned}$$

where $\boldsymbol{\alpha}^*$ is between $\hat{\boldsymbol{\alpha}}$ and $\boldsymbol{\alpha}_0$. For each r,

$$\begin{aligned}
&|\frac{1}{2\sqrt{n}} \sum_s \sum_t \{\frac{\partial^3}{\partial\alpha_r \partial\alpha_s \partial\alpha_t} l(\boldsymbol{\alpha}^*)(\hat{\alpha}_s - \alpha_{s0})(\hat{\alpha}_t - \alpha_{t0})\}| \\
&\leq \frac{1}{2\sqrt{n}} \|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0\|^2 \sum_s \sum_t |\sum_i \frac{\partial^3}{\partial\alpha_r \partial\alpha_s \partial\alpha_t} l_i(\boldsymbol{\alpha}^*)| \\
&\leq \frac{1}{2\sqrt{n}} \|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0\|^2 \sum_s \sum_t \sum_i M_{rst}(V_i),
\end{aligned}$$

then for any $\varepsilon > 0$,

$$\begin{aligned}
&\mathrm{P}(|\frac{1}{2\sqrt{n}} \sum_s \sum_t \{\frac{\partial^3}{\partial\alpha_r \partial\alpha_s \partial\alpha_t} l(\boldsymbol{\alpha}^*)(\hat{\alpha}_s - \alpha_{s0})(\hat{\alpha}_t - \alpha_{t0})\}| > \varepsilon) \\
&\leq \frac{1}{\varepsilon}\mathrm{E}(|\frac{1}{2\sqrt{n}} \sum_s \sum_t \{\frac{\partial^3}{\partial\alpha_r \partial\alpha_s \partial\alpha_t} l(\boldsymbol{\alpha}^*)(\hat{\alpha}_s - \alpha_{s0})(\hat{\alpha}_t - \alpha_{t0})\}|) \\
&\to 0,
\end{aligned}$$

which means,

$$\frac{1}{2\sqrt{n}} \sum_s \sum_t \{\frac{\partial^3}{\partial\alpha_r \partial\alpha_s \partial\alpha_t} l(\boldsymbol{\alpha}^*)(\hat{\alpha}_s - \alpha_{s0})(\hat{\alpha}_t - \alpha_{t0})\} \xrightarrow{P} 0,$$

and

$$\mathbf{r} = \frac{1}{2\sqrt{n}} \nabla_{\boldsymbol{\alpha}}\{(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0)^T \nabla^2_{\boldsymbol{\alpha}} l(\boldsymbol{\alpha}^*)(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0)\} \xrightarrow{P} \mathbf{0}.$$

Hence, with Slutsky's theorem and the central limit theorem,

$$\sqrt{n}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0) = (-\frac{1}{n}\nabla^2_{\boldsymbol{\alpha}} l(\boldsymbol{\alpha}_0))^{-1}(\frac{1}{\sqrt{n}}\nabla_{\boldsymbol{\alpha}} l(\boldsymbol{\alpha}_0) + \mathbf{r}) \xrightarrow{d} \mathrm{N}(\mathbf{0}, \bar{\mathbf{I}}_0^{-1}).$$

[Note 1] For the consistency, we may alternatively have

$$D_n(\mathbf{u}) = \frac{1}{\sqrt{n}}\nabla^T_{\boldsymbol{\alpha}} l(\boldsymbol{\alpha}_0)\mathbf{u} + \frac{1}{2n}\mathbf{u}^T\nabla^2_{\boldsymbol{\alpha}} l(\boldsymbol{\alpha}^*)\mathbf{u},$$

where $\boldsymbol{\alpha}^*$ is between $\boldsymbol{\alpha}_0$ and $\boldsymbol{\alpha}_0 + n^{-\frac{1}{2}}\mathbf{u}$. Particularly, for every (r, s),

$$\begin{aligned}
&|\frac{1}{n}\frac{\partial^2}{\partial\alpha_r\partial\alpha_s} l(\boldsymbol{\alpha}^*) - \frac{1}{n}\frac{\partial^2}{\partial\alpha_r\partial\alpha_s} l(\boldsymbol{\alpha}_0)| \\
&= \frac{1}{n}|\sum_t \{\frac{\partial^3}{\partial\alpha_r\partial\alpha_s\partial\alpha_t} l(\boldsymbol{\alpha}^{**})(\alpha_t^* - \alpha_{t0})\}| \\
&\leq \frac{1}{n}\|\frac{1}{\sqrt{n}}\mathbf{u}\| \sum_t \sum_i M_{rst}(V_i) \\
&= O(\frac{1}{\sqrt{n}}) \to 0,
\end{aligned}$$

where $\boldsymbol{\alpha}^{**}$ is between $\boldsymbol{\alpha}^*$ and $\boldsymbol{\alpha}_0$, then

$$\frac{1}{n}\nabla^2_{\boldsymbol{\alpha}} l(\boldsymbol{\alpha}^*) - \frac{1}{n}\nabla^2_{\boldsymbol{\alpha}} l(\boldsymbol{\alpha}_0) \to \mathbf{O}.$$

Thus,

$$\begin{aligned}
&\frac{1}{2n}\mathbf{u}^T\nabla^2_{\boldsymbol{\alpha}} l(\boldsymbol{\alpha}^*)\mathbf{u} \\
&= \frac{1}{2n}\mathbf{u}^T(\nabla^2_{\boldsymbol{\alpha}} l(\boldsymbol{\alpha}^*) - \nabla^2_{\boldsymbol{\alpha}} l(\boldsymbol{\alpha}_0) + \nabla^2_{\boldsymbol{\alpha}} l(\boldsymbol{\alpha}_0) + \bar{\mathbf{I}}_0 - \bar{\mathbf{I}}_0)\mathbf{u} \\
&= -\frac{1}{2n}\mathbf{u}^T\bar{\mathbf{I}}_0\mathbf{u}(1 + o_P(1)),
\end{aligned}$$

which also leads to the $n^{\frac{1}{2}}$-consistency of $\hat{\boldsymbol{\alpha}}$.

[Note 2] Alternatively, for asymptotic normality, since

$$
\begin{aligned}
&|\frac{1}{2n}\sum_t\{\frac{\partial^3}{\partial\alpha_r\partial\alpha_s\partial\alpha_t}l(\boldsymbol{\alpha}^*)(\hat{\alpha}_t-\alpha_{t0})\}|\\
&=|\frac{1}{2n}\sum_t\sum_i\{\frac{\partial^3}{\partial\alpha_r\partial\alpha_s\partial\alpha_t}l_i(\boldsymbol{\alpha}^*)(\hat{\alpha}_t-\alpha_{t0})\}|\\
&\le\frac{1}{2n}\|\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}_0\|\sum_t\sum_i M_{rst}(V_i),
\end{aligned}
$$

which implies that

$$
\frac{1}{2n}\sum_t\frac{\partial^3}{\partial\alpha_r\partial\alpha_s\partial\alpha_t}l(\boldsymbol{\alpha}^*)(\hat{\alpha}_t-\alpha_{t0})\xrightarrow{P}0,
$$

holds for every (r,s), we have

$$
\frac{1}{n}\mathbf{R}=\frac{1}{2n}\nabla^2_{\boldsymbol{\alpha}}\{(\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}_0)^T\nabla_{\boldsymbol{\alpha}}l(\boldsymbol{\alpha}^*)\}\xrightarrow{P}\mathbf{O},
$$

and

$$
\sqrt{n}(\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}_0)=\{-\frac{1}{n}(\nabla^2_{\boldsymbol{\alpha}}l(\boldsymbol{\alpha}_0)+\mathbf{R})\}^{-1}(\frac{1}{\sqrt{n}}\nabla_{\boldsymbol{\alpha}}l(\boldsymbol{\alpha}_0))\xrightarrow{d}\mathrm{N}(\mathbf{0},\bar{\mathbf{I}}_0^{-1}).
$$

□

Proof of Corollary 4.5. First, we show that, for any given $\epsilon>0$, there exists a constant C_B such that

$$
\mathrm{P}(\sup_{\|\mathbf{u}\|=C_B}l_B(\boldsymbol{\alpha}_0+a_n\mathbf{u})<l_B(\boldsymbol{\alpha}_0))\ge1-\epsilon, \tag{4.5}
$$

where $a_n=n^{-\frac{1}{2}}+n^{-1}v_{Bn}$.

By the Taylor expansion, we have

$$\begin{aligned} D_{Bn}(\mathbf{u}) &= l_B(\boldsymbol{\alpha}_0 + a_n\mathbf{u}) - l_B(\boldsymbol{\alpha}_0) \\ &= a_n\nabla_{\boldsymbol{\alpha}}^T l(\boldsymbol{\alpha}_0)\mathbf{u} - \frac{1}{2}a_n^2 n\mathbf{u}^T\bar{\mathbf{I}}_0\mathbf{u}(1+o_P(1)) \\ &\quad - a_n\nabla_{\boldsymbol{\alpha}}^T B(\boldsymbol{\alpha}_0)\mathbf{u} - \frac{1}{2}a_n^2\mathbf{u}^T\nabla_{\boldsymbol{\alpha}}^2 B(\boldsymbol{\alpha}_0)\mathbf{u} - \frac{1}{6}a_n^3\nabla_{\boldsymbol{\alpha}}^T(\mathbf{u}^T\nabla_{\boldsymbol{\alpha}}^2 B(\boldsymbol{\alpha}^*)\mathbf{u})\mathbf{u}, \end{aligned}$$

where $\boldsymbol{\alpha}^*$ is between $\boldsymbol{\alpha}_0$ and $\boldsymbol{\alpha}_0 + a_n\mathbf{u}$. Similarly to the proof of Theorem 4.2, we have

$$a_n\nabla_{\boldsymbol{\alpha}}^T l(\boldsymbol{\alpha}_0)\mathbf{u} = O_P(a_n n^{\frac{1}{2}}) = O_P(a_n^2 n),$$

and

$$\frac{1}{2}a_n^2 n\mathbf{u}^T\bar{\mathbf{I}}_0\mathbf{u}(1+o_P(1)) \asymp_P a_n^2 n.$$

In addition, we have

$$a_n\nabla_{\boldsymbol{\alpha}}^T B(\boldsymbol{\alpha}_0)\mathbf{u} = O_P(a_n v_{Bn}) = O_P(a_n^2 n),$$

$$\frac{1}{2}a_n^2\mathbf{u}^T\nabla_{\boldsymbol{\alpha}}^2 B(\boldsymbol{\alpha}_0)\mathbf{u} = o_P(a_n^2 n);$$

also, since

$$\begin{aligned} &|\frac{1}{6}a_n^3\nabla_{\boldsymbol{\alpha}}^T(\mathbf{u}^T\nabla_{\boldsymbol{\alpha}}^2 B(\boldsymbol{\alpha}^*)\mathbf{u})\mathbf{u}| \\ &= |\frac{1}{6}a_n^3\sum_r\sum_s\sum_t(\frac{\partial^3}{\partial\alpha_r\partial\alpha_s\partial\alpha_t}B(\boldsymbol{\alpha}^*)u_r u_s u_t)| \\ &\leq \frac{1}{6}a_n^3\|\mathbf{u}\|^3\sum_r\sum_s\sum_t M_{Brst}(\mathbf{V}), \end{aligned}$$

we have that, for any $\varepsilon > 0$,

$$\begin{aligned}
&\mathrm{P}(\frac{1}{a_n^2 n}|\frac{1}{6}a_n^3 \nabla_{\boldsymbol{\alpha}}^T(\mathbf{u}^T\nabla_{\boldsymbol{\alpha}}^2 B(\boldsymbol{\alpha}^*)\mathbf{u})\mathbf{u}| > \varepsilon) \\
&\leq \frac{1}{a_n^2 n\varepsilon}\mathrm{E}(|\frac{1}{6}a_n^3 \nabla_{\boldsymbol{\alpha}}^T(\mathbf{u}^T\nabla_{\boldsymbol{\alpha}}^2 B(\boldsymbol{\alpha}^*)\mathbf{u})\mathbf{u}|) \\
&= o(1),
\end{aligned}$$

then

$$\frac{1}{6}a_n^3 \nabla_{\boldsymbol{\alpha}}^T(\mathbf{u}^T\nabla_{\boldsymbol{\alpha}}^2 B(\boldsymbol{\alpha}^*)\mathbf{u})\mathbf{u} = o_P(a_n^2 n).$$

Thus, the second term of $D_{Bn}(\mathbf{u})$ dominates all others uniformly in $\|\mathbf{u}\| = C_B$. Hence, by choosing a sufficiently large C_B, $D_{Bn}(\mathbf{u}) < 0$, and inequality (4.5) holds.

Next, since $\|\hat{\boldsymbol{\alpha}}_B - \boldsymbol{\alpha}_0\| = O_P(a_n)$, we have, by the Taylor expansion,

$$\begin{aligned}
\mathbf{0} =& \nabla_{\boldsymbol{\alpha}} l_B(\hat{\boldsymbol{\alpha}}_B) \\
=& \nabla_{\boldsymbol{\alpha}} l(\boldsymbol{\alpha}_0) + \nabla_{\boldsymbol{\alpha}}^2 l(\boldsymbol{\alpha}_0)(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0) + \frac{1}{2}\nabla_{\boldsymbol{\alpha}}\{(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0)^T\nabla_{\boldsymbol{\alpha}}^2 l(\boldsymbol{\alpha}^*)(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0)\} \\
&- \nabla_{\boldsymbol{\alpha}} B(\boldsymbol{\alpha}_0) - \nabla_{\boldsymbol{\alpha}}^2 B(\boldsymbol{\alpha}_0)(\hat{\boldsymbol{\alpha}}_B - \boldsymbol{\alpha}_0) - \frac{1}{2}\nabla_{\boldsymbol{\alpha}}\{(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0)^T\nabla_{\boldsymbol{\alpha}}^2 B(\boldsymbol{\alpha}^*)(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0)\},
\end{aligned}$$

where $\boldsymbol{\alpha}^*$ is between $\boldsymbol{\alpha}_0$ and $\hat{\boldsymbol{\alpha}}_B$. Since, for every (r, s),

$$|\frac{1}{2n}\sum_t\{\frac{\partial^3}{\partial\alpha_r\partial\alpha_s\partial\alpha_t}B(\boldsymbol{\alpha}^*)(\hat{\alpha}_{Bt} - \alpha_{t0})\}| \leq \frac{1}{2n}\|\hat{\boldsymbol{\alpha}}_B - \boldsymbol{\alpha}_0\|\sum_t M_{Brst}(\mathbf{V}),$$

which implies that, for any $\varepsilon > 0$,

$$\begin{aligned}
&\mathrm{P}(|\frac{1}{2n}\sum_t\{\frac{\partial^3}{\partial\alpha_r\partial\alpha_s\partial\alpha_t}B(\boldsymbol{\alpha}^*)(\hat{\alpha}_{Bt} - \alpha_{t0})\}| > \varepsilon) \\
&\leq \frac{1}{\varepsilon}\mathrm{E}(|\frac{1}{2n}\sum_t\{\frac{\partial^3}{\partial\alpha_r\partial\alpha_s\partial\alpha_t}B(\boldsymbol{\alpha}^*)(\hat{\alpha}_{Bt} - \alpha_{t0})\}|) \\
&= o(1),
\end{aligned}$$

and

$$\frac{1}{2n}\sum_t \frac{\partial^3}{\partial\alpha_r\partial\alpha_s\partial\alpha_t}B(\boldsymbol{\alpha}^*)(\hat{\alpha}_{Bt}-\alpha_{t0}) \xrightarrow{P} 0,$$

we have

$$\frac{1}{n}\mathbf{R}_B = \frac{1}{2n}\nabla^2_{\boldsymbol{\alpha}}\{(\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}_0)^T\nabla_{\boldsymbol{\alpha}}B(\boldsymbol{\alpha}^*)\} \xrightarrow{P} \mathbf{O}.$$

Thus,

$$\mathbf{0} = \nabla_{\boldsymbol{\alpha}}l(\boldsymbol{\alpha}_0) - \nabla_{\boldsymbol{\alpha}}B(\boldsymbol{\alpha}_0) + (\nabla^2_{\boldsymbol{\alpha}}l(\boldsymbol{\alpha}_0) - \nabla^2_{\boldsymbol{\alpha}}B(\boldsymbol{\alpha}_0) + \mathbf{R} + \mathbf{R}_B)(\hat{\boldsymbol{\alpha}}_B - \boldsymbol{\alpha}_0),$$

and

$$\sqrt{n}(\bar{\mathbf{I}}_0 + \boldsymbol{\Sigma}_{B0})\{\hat{\boldsymbol{\alpha}}_P - \boldsymbol{\alpha}_0 + (\bar{\mathbf{I}}_0 + \boldsymbol{\Sigma}_{B0})^{-1}\mathbf{b}_{B0}\} \xrightarrow{d} \mathrm{N}(\mathbf{0}, \bar{\mathbf{I}}_0),$$

by Slutsky's theorem and the central limit theorem, where $\boldsymbol{\Sigma}_{B0} = n^{-1}\nabla_{\boldsymbol{\alpha}}B(\boldsymbol{\alpha}_0)$ and $\mathbf{b}_{B0} = n^{-1}\nabla_{\boldsymbol{\alpha}}B(\boldsymbol{\alpha}_0)$.

[Note 1] Rather than conditions (B2) and (B3), we may instead assume that $n^{-1}\nabla^2_{\boldsymbol{\alpha}}B(\boldsymbol{\alpha}^*) = o_P(\mathbf{1}\mathbf{1}^T)$ holds for $\boldsymbol{\alpha}^*$ within a neighborhood of $\boldsymbol{\alpha}_0$. We then have

$$B(\boldsymbol{\alpha}_0 + a_n\mathbf{u}) - B(\boldsymbol{\alpha}_0) = a_n\nabla^T_{\boldsymbol{\alpha}}B(\boldsymbol{\alpha}_0)\mathbf{u} + \frac{1}{2}a_n^2\mathbf{u}^T\nabla^2_{\boldsymbol{\alpha}}B(\boldsymbol{\alpha}^*)\mathbf{u},$$

and

$$\frac{1}{2}a_n^2\mathbf{u}^T\nabla^2_{\boldsymbol{\alpha}}B(\boldsymbol{\alpha}^*)\mathbf{u} = o_P(a_n^2 n),$$

which also leads to the a_n^{-1}-consistency of $\hat{\boldsymbol{\alpha}}_B$. Furthermore, we have

$$\begin{aligned}\mathbf{0} &= \nabla_{\boldsymbol{\alpha}}l_B(\hat{\boldsymbol{\alpha}}_B)\\ &= \nabla_{\boldsymbol{\alpha}}l(\boldsymbol{\alpha}_0) - \nabla_{\boldsymbol{\alpha}}B(\boldsymbol{\alpha}_0) + (\nabla^2_{\boldsymbol{\alpha}}l(\boldsymbol{\alpha}_0) - \nabla^2_{\boldsymbol{\alpha}}B(\boldsymbol{\alpha}^*) + \mathbf{R})(\hat{\boldsymbol{\alpha}}_B - \boldsymbol{\alpha}_0),\end{aligned}$$

which gives

$$\sqrt{n}(\hat{\boldsymbol{\alpha}}_P - \boldsymbol{\alpha}_0 + \bar{\mathbf{I}}_0^{-1}\mathbf{b}_{B0}) \xrightarrow{d} \mathrm{N}(\mathbf{0}, \bar{\mathbf{I}}_0^{-1}).$$

□

Proof of Corollary 4.6. First, we show that, for any given $\epsilon > 0$, there exists a large constant C_P such that

$$\mathrm{P}(\sup_{\|\mathbf{u}\|=C_P} l_P(\boldsymbol{\alpha}_0 + \frac{1}{\sqrt{n}}\mathbf{u}) < l_P(\boldsymbol{\alpha}_0)) \geq 1 - \epsilon. \tag{4.6}$$

Similar to the proof of Theorem 4.2, we have

$$\begin{aligned} D_{Pn}(\mathbf{u}) &= l_P(\boldsymbol{\alpha}_0 + \frac{1}{\sqrt{n}}\mathbf{u}) - l_P(\boldsymbol{\alpha}_0) \\ &= \frac{1}{\sqrt{n}}\nabla_{\boldsymbol{\alpha}}^T l(\boldsymbol{\alpha}_0)\mathbf{u} - \frac{1}{2}\mathbf{u}^T\bar{\mathbf{I}}_0\mathbf{u}(1 + o_P(1)) - \frac{2}{\sqrt{n}}\lambda\boldsymbol{\alpha}_0^T\mathbf{K}\mathbf{u} - \frac{1}{n}\lambda\mathbf{u}^T\mathbf{K}\mathbf{u}, \end{aligned}$$

where

$$\frac{1}{\sqrt{n}}\nabla_{\boldsymbol{\alpha}}^T l(\boldsymbol{\alpha}_0)\mathbf{u} = O_P(1),$$

$$\frac{1}{2}\mathbf{u}^T\bar{\mathbf{I}}_0\mathbf{u}(1 + o_P(1)) \asymp_P 1,$$

and

$$-\frac{2}{\sqrt{n}}\lambda\boldsymbol{\alpha}_0^T\mathbf{K}\mathbf{u} = O_P(\frac{1}{\sqrt{n}}) = O_P(1),$$

$$\frac{1}{n}\lambda\mathbf{u}^T\mathbf{K}\mathbf{u} = O_P(\frac{1}{n}) = o_P(1).$$

Thus, the second term of $D_{Pn}(\mathbf{u})$ dominates all others, which means, by choosing a sufficiently large C_P, $D_{Pn}(\mathbf{u}) < 0$, and inequality (4.6) holds.

Next, since $\|\hat{\boldsymbol{\alpha}}_P - \boldsymbol{\alpha}_0\| = O_P(n^{-\frac{1}{2}})$, we have

$$\mathbf{0} = \nabla_{\boldsymbol{\alpha}} l_P(\hat{\boldsymbol{\alpha}}_P) = \nabla_{\boldsymbol{\alpha}} l(\boldsymbol{\alpha}_0) - 2\lambda\mathbf{K}\boldsymbol{\alpha}_0 + (\nabla_{\boldsymbol{\alpha}}^2 l(\boldsymbol{\alpha}_0) - 2\lambda\mathbf{K} + \mathbf{R})(\hat{\boldsymbol{\alpha}}_P - \boldsymbol{\alpha}_0),$$

which leads to

$$\sqrt{n}(\bar{\mathbf{I}}_0 + \frac{2}{n}\lambda\mathbf{K})\{\hat{\boldsymbol{\alpha}}_P - \boldsymbol{\alpha}_0 + (\bar{\mathbf{I}}_0 + \frac{2}{n}\lambda\mathbf{K})^{-1}(\frac{2}{n}\lambda\mathbf{K}\boldsymbol{\alpha}_0)\} \xrightarrow{d} \mathrm{N}(\mathbf{0}, \bar{\mathbf{I}}_0),$$

using Slutsky's theorem and the central limit theorem.

□

4.2 Asymptotic Properties for Diverging Observation Time Points

In this section, we take a further step by allowing the number of total observation time points to diverge as the sample size increases. In practice, it is very likely that one or more of the observation time points for a newly entered subject will be different from the existing ones. Under such settings, stronger conditions on the likelihood are desired, and additional conditions are required, especially, how the observation time points change with the sample size. Similarly to Section 4.1, we first obtain general results regarding the GLMMs with a diverging number of parameters, then specify the conditions and reach conclusions within our longitudinal covariate model framework.

4.2.1 General Results for GLMMs with Infinite Parameters

In practice, the number of introduced covariates possibly depends on the sample size. For our GLMM problems, we assume

$$h(\mu_{ij}) = \mathbf{x}_{ij}^T\boldsymbol{\alpha}_n + \mathbf{z}_{ij}^T\mathbf{b}_i,$$

and let the dimension $p_n = \dim(\boldsymbol{\alpha}_n)$ diverge. Under this setting, the log-likelihood is $l_{(n)}(\boldsymbol{\alpha}_n) = \sum_i l_{(n)i}(\boldsymbol{\alpha}_n)$, where

$$l_{(n)i}(\boldsymbol{\alpha}_n) = \log(L_{(n)i}(V_{(n)i}, \boldsymbol{\alpha}_n)) = \log(\int \prod_j f(y_{ij}; h^{-1}(\mathbf{x}_{ij}^T\boldsymbol{\alpha}_n + \mathbf{z}_{ij}^T\mathbf{b}_i))q(\mathbf{b}_i; \boldsymbol{\theta}_0)d\mathbf{b}_i).$$

When a quadratic penalty is considered, the coefficient vector $\boldsymbol{\alpha}_n$ can be estimated by maximizing the penalized log-likelihood

$$l_{P(n)}(\boldsymbol{\alpha}_n) = l_{(n)}(\boldsymbol{\alpha}_n) - \lambda_n \boldsymbol{\alpha}_n^T \mathbf{K}_n \boldsymbol{\alpha}_n,$$

where λ_n and $\mathbf{K}_n$ are smoothing parameter and matrix respectively, depending on the sample size n.

In order to derive asymptotic properties for $\hat{\boldsymbol{\alpha}}_{Pn}$, the maximizer of $l_{P(n)}(\boldsymbol{\alpha}_n)$, we impose following conditions on $l_{(n)i}(\boldsymbol{\alpha}_n)$'s.

(C1) The likelihood $L_{(n)i}(V_{(n)i}, \boldsymbol{\alpha}_n)$ has a common support and the model is identifiable; moreover, $\mathrm{E}(\nabla_{\boldsymbol{\alpha}} l_{(n)i}) = \mathbf{0}$, $\mathrm{E}(\nabla_{\boldsymbol{\alpha}} l_{(n)i} \nabla_{\boldsymbol{\alpha}}^T l_{(n)i}) = \mathrm{E}(-\nabla_{\boldsymbol{\alpha}}^2 l_{(n)i})$.

(C2) The information matrix $\mathbf{I}_{(n)i}(\boldsymbol{\alpha}_n) = \mathrm{E}(\nabla_{\boldsymbol{\alpha}} l_{(n)i} \nabla_{\boldsymbol{\alpha}}^T l_{(n)i})$ satisfies

$$0 < c_1 \leq \lambda_{\min}(\mathbf{I}_{(n)i}) \leq \lambda_{\max}(\mathbf{I}_{(n)i}) \leq c_2 < \infty;$$

in addition, $\bar{\mathbf{I}}_{(n)0} = n^{-1} \sum_i \mathbf{I}_{(n)i}(\boldsymbol{\alpha}_{n0})$ satisfies

$$0 < c_{10} \leq \lambda_{\min}(\bar{\mathbf{I}}_{(n)0}) \leq \lambda_{\max}(\bar{\mathbf{I}}_{(n)0}) \leq c_{20} < \infty.$$

(C3) For every (r, s),

$$\mathrm{E}((\frac{\partial}{\partial \alpha_r} l_{(n)i}(\boldsymbol{\alpha}_{n0}) \frac{\partial}{\partial \alpha_s} l_{(n)i}(\boldsymbol{\alpha}_{n0}))^2) \leq c_3 < \infty.$$

(C4) For every (r, s),

$$\mathrm{E}((\frac{\partial^2}{\partial\alpha_r\partial\alpha_s}l_{(n)i}(\boldsymbol{\alpha}_{n0}))^2) \leq c_4 < \infty.$$

(C5) There exists a large enough open subset ω_n of $\boldsymbol{\alpha}_{n0} \in \omega_n \subset \Omega_n \subset \mathbb{R}^{p_n}$, for almost all $V_{(n)i}$, there exist functions $M_{nrst}(\cdot)$'s such that

$$|\frac{\partial^3}{\partial\alpha_r\partial\alpha_s\partial\alpha_t}l_{(n)i}(V_{(n)i}, \boldsymbol{\alpha}_n)| \leq M_{nrst}(V_{(n)i}),$$

for all $\boldsymbol{\alpha}_n \in \omega_n$ and every (r, s, t), where $\mathrm{E}(M^2_{nrst}(V_{(n)i})) \leq c_5 < \infty$.

In addition, we make following assumptions regarding p_n, λ_n, $\mathbf{K}_n$, and $\boldsymbol{\alpha}_{n0}$.

(D1) $n^{-1}p_n^5 = o(1)$.

(D2) $n^{\frac{1}{2}}\lambda_n = O(1)$.

(D3) $\mathbf{K}_n = p_n^3 O(\mathbf{1}\mathbf{1}^T)$.

(D4) $\boldsymbol{\alpha}_{n0} = O(\mathbf{1})$.

These conditions are inspired from Fan and Peng (2004). Actually, conditions (C1) - (C5) are stronger than their counterparts of conditions (A1) - (A5) or (A1a) - (A3a), but facilitate the derivations of the asymptotic properties. In particular, condition (C2) assumes the information matrix to be positive definite, and its eigenvalues to be uniformly bounded; conditions (C3) and (C4) are imposed on the fourth moment of the likelihood function. Additionally, condition (D1) bounds the increasing rate of number of parameters; when accounting for "smoothness" of parameters, conditions on the smoothing parameter λ_n and quadratic smoothing

matrix $\mathbf{K}_n$ are further imposed. These conditions altogether guarantee the consistency and asymptotic normality of the local maximizer $\hat{\boldsymbol{\alpha}}_{Pn}$ of $l_{P(n)}(\boldsymbol{\alpha}_n)$ under a setting of infinite parameters.

We then derive asymptotic properties for the local maximizer $\hat{\boldsymbol{\alpha}}_{Pn}$.

Theorem 4.8. *For a GLMM involving clustered random effect, where the dimension of parameters is diverging, if $l_{(n)i}(\boldsymbol{\alpha}_n)$'s satisfy conditions (C1) - (C5), as well as conditions (D1) - (D4) hold, there exists a local maximizer $\hat{\boldsymbol{\alpha}}_{Pn}$ of $l_{P(n)}(\boldsymbol{\alpha}_n)$ such that*

a) (consistency)

$$\|\hat{\boldsymbol{\alpha}}_{Pn} - \boldsymbol{\alpha}_{n0}\| = O_P(\sqrt{\frac{p_n}{n}});$$

b) (asymptotic normality) for a $p_0 \times p_n$ matrix $\mathbf{A}_n$ and a $p_0 \times p_0$ nonnegative symmetric matrix $\mathbf{G}$ satisfying $\mathbf{A}_n\mathbf{A}_n^T \to \mathbf{G}$,

$$\sqrt{n}\mathbf{A}_n\bar{\mathbf{I}}_{(n)0}^{-\frac{1}{2}}(\bar{\mathbf{I}}_{(n)0} + \frac{2}{n}\lambda_n\mathbf{K}_n)(\hat{\boldsymbol{\alpha}}_{Pn} - \boldsymbol{\alpha}_{n0}^*) \xrightarrow{d} N(\mathbf{0}, \mathbf{G}),$$

where $\boldsymbol{\alpha}_{n0}^* = \boldsymbol{\alpha}_{n0} - (\bar{\mathbf{I}}_{(n)0} + 2n^{-1}\lambda_n\mathbf{K}_n)^{-1}(2n^{-1}\lambda_n\mathbf{K}_n\boldsymbol{\alpha}_{n0})$.

Proof. A proof is given in Section 4.2.3. □

4.2.2 Results for Longitudinal Covariate Model with Diverging Observation Time Points

For the problem discussed in Section 3.2.1, we further consider a more realistic situation that more observation time points are introduced as more subjects are included in the study. We again assume that $x_0(\cdot)$ belongs to $W^2([T_1, T_2])$, and use NCS functions for the approximation; that is, $x_{00}(\cdot) \approx x_{00n}^{\mathrm{NCS}}(\cdot) = \mathbf{x}_{00n}^T\mathbf{c}(\cdot)$ for every n, where $x_{00n}^{\mathrm{NCS}}(\cdot) \in \mathcal{S}_{\mathrm{NCS}}(\mathbf{t}_n^0)$ and $\mathbf{x}_{00n} = x_{00n}^{\mathrm{NCS}}(\mathbf{t}_n^0) = x_{00}(\mathbf{t}_n^0)$. In particular, when

$r_n = \dim(\mathbf{t}_n^0)$ increases, the number of basis functions, as well as the dimension of $\mathbf{x}_{0n}$, will increase correspondingly. We estimate $\mathbf{x}_{0n}$ by maximizing the penalized log-quasi-likelihood

$$ql_{P(n)}(\mathbf{x}_{0n}) = ql_{(n)}(\mathbf{x}_{0n}) - \frac{1}{2}\lambda_{xn}\mathbf{x}_{0n}^T\mathbf{K}_n\mathbf{x}_{0n},$$

where

$$ql_{(n)}(\mathbf{x}_{0n}) = \sum_{i=1}^{n} ql_{(n)i}(\mathbf{x}_{0n}) = \sum_{i=1}^{n} \log(\int \exp(-\frac{1}{2}\sum_{j=1}^{n_i}\tilde{d}_{ij} - \frac{1}{2}\mathbf{b}_i^T\mathbf{D}^{-1}(\boldsymbol{\theta})\mathbf{b}_i)d\mathbf{b}_i),$$

and λ_{xn} is a smoothing parameter, $\mathbf{K}_n$ is a smoothing matrix, both determined by $\mathbf{t}_n^0$.

In order to derive asymptotic properties for $\hat{\mathbf{x}}_{0n}$, we impose following conditions on $ql_{(n)i}(\mathbf{x}_{0n})$'s.

(I1) The distribution has a common support and the model is identifiable; moreover, $\mathrm{E}(\nabla_{\mathbf{x}} ql_{(n)i}) = \mathbf{0}$, $\mathrm{E}(\nabla_{\mathbf{x}} ql_{(n)i}\nabla_{\mathbf{x}}^T ql_{(n)i}) = \mathrm{E}(-\nabla_{\mathbf{x}}^2 ql_{(n)i})$.

(I2) The information matrix $\mathbf{I}_{(n)i}(\mathbf{x}_{0n}) = \mathrm{E}(\nabla_{\mathbf{x}} ql_{(n)i}\nabla_{\mathbf{x}}^T ql_{(n)i})$ satisfies

$$0 < c_1 \leq \lambda_{\min}(\mathbf{I}_{(n)i}) \leq \lambda_{\max}(\mathbf{I}_{(n)i}) \leq c_2 < \infty;$$

in addition, $\bar{\mathbf{I}}_{(n)0} = n^{-1}\sum_i \mathbf{I}_{(n)i}(x_{00}(\mathbf{t}_n^0))$ satisfies

$$0 < c_{10} \leq \lambda_{\min}(\bar{\mathbf{I}}_{(n)0}) \leq \lambda_{\max}(\bar{\mathbf{I}}_{(n)0}) \leq c_{20} < \infty.$$

(I3) For every (r, s),

$$\mathrm{E}((\frac{\partial}{\partial x_r} ql_{(n)i}(x_{00}(\mathbf{t}_n^0))\frac{\partial}{\partial x_s} ql_{(n)i}(x_{00}(\mathbf{t}_n^0)))^2) \leq c_3 < \infty.$$

(I4) For every (r, s),

$$\mathrm{E}((\frac{\partial^2}{\partial x_r \partial x_s} l_{(n)i}(x_{00}(\mathbf{t}_n^0)))^2) \leq c_4 < \infty.$$

(I5) There exists an open subset ω_n of $x_{00}(\mathbf{t}_n^0) \in \omega_n \subset \Omega_n \subset \mathbb{R}^{r_n}$, for almost all $V_{(n)i}$, there exist functions $M_{nrst}(\cdot)$'s such that

$$|\frac{\partial^3}{\partial x_r \partial x_s \partial x_t} q l_{(n)i}(V_{(n)i}, \mathbf{x}_{0n})| \leq M_{nrst}(V_{(n)i}),$$

for all $\mathbf{x}_{0n} \in \omega_n$ and every (r, s, t), where $\mathrm{E}(M_{nrst}^2(V_{(n)i})) \leq c_5 < \infty$.

In addition, we make following assumptions.

(J1) $n^{-1} r_n^5 = o(1)$.

(J2) $n^{\frac{1}{2}} \lambda_n = O(1)$.

(J3) The observation time points of longitudinal covariates are "nearly" evenly spaced, in particular, there exists $c_6 \in (0, \infty)$ such that

$$\frac{\max(s_{ni}^0)}{\min(s_{ni}^0)} \leq c_6$$

holds for every n, where $s_{ni}^0 = |t_{n(i+1)}^0 - t_{ni}^0|$ for $i = 1, \ldots, (r_n - 1)$.

The conditions above are comparable to conditions (C1) - (C5) and (D1) - (D4). In particular, condition (J3) indicates that $\mathbf{K}_n = r_n^3 O(\mathbf{1}\mathbf{1}^T)$, and $x_{00}(\cdot) \in W^2([T_1, T_2])$ implies that $x_{00}(\mathbf{t}_n^0) = O(\mathbf{1})$.

Followed by the arguments in Section 4.2.1, we reach the conclusions on the estimator $\hat{\mathbf{x}}_{0n}$.

Theorem 4.9. *For a GAMM estimation problem as model (3.4,3.5) when the number of observation time points increases as more subjects are included, the true population profile function $x_{00}(\cdot)$ belongs to $W^2([T_1, T_2])$, and the properties of its projection in $\mathcal{S}_{\mathrm{NCS}}(\mathbf{t}_n^0)$ are studied. If conditions (I1) - (I5) and (J1) - (J3) are satisfied, there exists a local maximizer $\hat{\mathbf{x}}_{0n}$ of $ql_{(n)}(\mathbf{x}_{0n})$ such that*

a) (consistency)

$$\|\hat{\mathbf{x}}_{0n} - x_{00}(\mathbf{t}_n^0)\| = O_P(\sqrt{\frac{r_n}{n}});$$

b) (asymptotic normality) for a $p_0 \times r_n$ matrix $\mathbf{A}_n$ and a $p_0 \times p_0$ nonnegative symmetric matrix $\mathbf{G}$ satisfying $\mathbf{A}_n\mathbf{A}_n^T \to \mathbf{G}$,

$$\sqrt{n}\mathbf{A}_n\bar{\mathbf{I}}_{(n)0}^{-\frac{1}{2}}(\bar{\mathbf{I}}_{(n)0} + \frac{2}{n}\lambda_n\mathbf{K}_n)(\hat{\mathbf{x}}_{0n} - \mathbf{x}_{00n}^*) \xrightarrow{d} N(\mathbf{0}, \mathbf{G}),$$

where $\mathbf{x}_{00n}^* = x_{00}(\mathbf{t}_n^0) - (\bar{\mathbf{I}}_{(n)0} + 2n^{-1}\lambda_n\mathbf{K}_n)^{-1}(2n^{-1}\lambda_n\mathbf{K}_n x_{00}(\mathbf{t}_n^0))$.

4.2.3 Proofs

In this section, we provide rigorous proof of Theorem 4.8. Particularly, the Euclidean norm and Frobenius norm are adopted for vectors and matrices, respectively.

Proof of Theorem 4.8. First, we show that, for any given $\epsilon > 0$, there exists a large constant C_P such that

$$\mathrm{P}(\sup_{\|\mathbf{u}\|=C_P} l_{P(n)}(\boldsymbol{\alpha}_{n0} + a_n\mathbf{u}) < l_{P(n)}(\boldsymbol{\alpha}_{n0})) \geq 1 - \epsilon, \tag{4.7}$$

where $a_n = p_n^{\frac{1}{2}}n^{-\frac{1}{2}}$.

By the Taylor expansion, we have

$$\begin{aligned} D_{Pn}(\mathbf{u}) &= l_{P(n)}(\boldsymbol{\alpha}_{n0} + a_n\mathbf{u}) - l_{P(n)}(\boldsymbol{\alpha}_{n0}) \\ &= a_n\nabla_{\boldsymbol{\alpha}}^T l_{(n)}(\boldsymbol{\alpha}_{n0})\mathbf{u} + \frac{1}{2}a_n^2\mathbf{u}^T\nabla_{\boldsymbol{\alpha}}^2 l_{(n)}(\boldsymbol{\alpha}_{n0})\mathbf{u} + \frac{1}{6}a_n^3\nabla_{\boldsymbol{\alpha}}^T(\mathbf{u}^T\nabla_{\boldsymbol{\alpha}}^2 l_{(n)}(\boldsymbol{\alpha}_n^*)\mathbf{u})\mathbf{u} \\ &\quad - 2a_n\lambda_n\boldsymbol{\alpha}_{n0}^T\mathbf{K}_n\mathbf{u} - a_n^2\lambda_n\mathbf{u}^T\mathbf{K}_n\mathbf{u}, \end{aligned}$$

where $\boldsymbol{\alpha}_n^*$ is between $\boldsymbol{\alpha}_{n0}$ and $\boldsymbol{\alpha}_{n0} + a_n\mathbf{u}$.

For each of the five terms, we have that:
(1) for any $\varepsilon > 0$, there exists $M_\varepsilon = c_2^{\frac{1}{2}}\varepsilon^{-\frac{1}{2}}$ such that

$$\begin{aligned} &\mathrm{P}\Big(\frac{\|\nabla_{\boldsymbol{\alpha}} l_{(n)}(\boldsymbol{\alpha}_{n0})\|}{\sqrt{np_n}} > M_\varepsilon\Big) \\ &\le \frac{\mathrm{E}(\|\nabla_{\boldsymbol{\alpha}} l_{(n)}(\boldsymbol{\alpha}_{n0})\|^2)}{M_\varepsilon^2 np_n} \\ &= \frac{\mathrm{tr}(\mathrm{E}(\nabla_{\boldsymbol{\alpha}} l_{(n)}(\boldsymbol{\alpha}_{n0})\nabla_{\boldsymbol{\alpha}} l_{(n)}^T(\boldsymbol{\alpha}_{n0})))}{M_\varepsilon^2 np_n} \\ &= \frac{\sum_i \mathrm{tr}(\mathbf{I}_{(n)i}(\boldsymbol{\alpha}_{n0}))}{M_\varepsilon^2 np_n} \\ &\le \frac{np_n c_2}{M_\varepsilon^2 np_n} = \varepsilon, \end{aligned}$$

which means

$$\|\nabla_{\boldsymbol{\alpha}} l_{(n)}(\boldsymbol{\alpha}_{n0})\| = O_P(\sqrt{np_n}),$$

then

$$|a_n\nabla_{\boldsymbol{\alpha}}^T l_{(n)}(\boldsymbol{\alpha}_{n0})\mathbf{u}| \le a_n\|\nabla_{\boldsymbol{\alpha}} l_{(n)}(\boldsymbol{\alpha}_{n0})\|\|\mathbf{u}\| = O_P(a_n\sqrt{p_n n}) = O_P(a_n^2 n);$$

(2) for any $\varepsilon > 0$,

$$\begin{aligned}
&\mathrm{P}\Big(\frac{\|\frac{1}{n}\nabla^2_{\boldsymbol{\alpha}} l_{(n)}(\boldsymbol{\alpha}_{n0}) + \bar{\mathbf{I}}_{(n)0}\|}{p_n^{-1}} > \varepsilon\Big) \\
&\leq \frac{p_n^2}{n^2\varepsilon^2}\mathrm{E}(\|\nabla^2_{\boldsymbol{\alpha}} l_{(n)}(\boldsymbol{\alpha}_{n0}) + n\bar{\mathbf{I}}_{(n)0}\|^2) \\
&= \frac{p_n^2}{n^2\varepsilon^2}\sum_r\sum_s \mathrm{E}(\{\sum_i(\frac{\partial^2}{\partial\alpha_r\partial\alpha_s}l_{(n)i}(\boldsymbol{\alpha}_{n0}) - \mathrm{E}(\frac{\partial^2}{\partial\alpha_r\partial\alpha_s}l_{(n)i}(\boldsymbol{\alpha}_{n0})))\}^2) \\
&= O(\frac{p_n^4}{n}) = o(1),
\end{aligned}$$

which means

$$\|\frac{1}{n}\nabla^2_{\boldsymbol{\alpha}} l_{(n)}(\boldsymbol{\alpha}_{n0}) + \bar{\mathbf{I}}_{(n)0}\| = o_P(\frac{1}{p_n}), \tag{4.8}$$

then

$$\mathbf{u}^T(\frac{1}{n}\nabla^2_{\boldsymbol{\alpha}} l_{(n)}(\boldsymbol{\alpha}_{n0}) + \bar{\mathbf{I}}_{(n)0})\mathbf{u} \leq \|\frac{1}{n}\nabla^2_{\boldsymbol{\alpha}} l_{(n)}(\boldsymbol{\alpha}_{n0}) + \bar{\mathbf{I}}_{(n)0}\|\|\mathbf{u}\|^2 = o_P(\frac{1}{p_n}),$$

and

$$\begin{aligned}
&-\frac{1}{2}a_n^2\mathbf{u}^T\nabla^2_{\boldsymbol{\alpha}} l_{(n)}(\boldsymbol{\alpha}_{n0})\mathbf{u} \\
&= -\frac{1}{2}a_n^2 n\mathbf{u}^T(\frac{1}{n}\nabla^2_{\boldsymbol{\alpha}} l_{(n)}(\boldsymbol{\alpha}_{n0}) + \bar{\mathbf{I}}_{(n)0} - \bar{\mathbf{I}}_{(n)0})\mathbf{u} \\
&= -\frac{1}{2}a_n^2 n\mathbf{u}^T(\frac{1}{n}\nabla^2_{\boldsymbol{\alpha}} l_{(n)}(\boldsymbol{\alpha}_{n0}) + \bar{\mathbf{I}}_{(n)0})\mathbf{u} + \frac{1}{2}a_n^2 n\mathbf{u}^T\bar{\mathbf{I}}_{(n)0}\mathbf{u} \\
&= o_P(a_n^2 n) + \frac{1}{2}a_n^2 n\mathbf{u}^T\bar{\mathbf{I}}_{(n)0}\mathbf{u} \\
&\asymp_P a_n^2 n;
\end{aligned}$$

(3)

$$
\begin{aligned}
&\{\frac{1}{6}a_n^3\nabla_{\boldsymbol{\alpha}}^T(\mathbf{u}^T\nabla_{\boldsymbol{\alpha}}^2 l_{(n)}(\boldsymbol{\alpha}_n^*)\mathbf{u})\mathbf{u}\}^2 \\
&= \frac{1}{36}a_n^6\{\sum_r\sum_s\sum_t(\frac{\partial^3}{\partial\alpha_r\partial\alpha_s\partial\alpha_t}l_{(n)}(\boldsymbol{\alpha}_n^*)u_ru_su_t)\}^2 \\
&\le \frac{1}{36}a_n^6\sum_r\sum_s\sum_t\{(\frac{\partial^3}{\partial\alpha_r\partial\alpha_s\partial\alpha_t}l_{(n)}(\boldsymbol{\alpha}_n^*))^2\}\sum_r\sum_s\sum_t(u_r^2u_s^2u_t^2) \\
&= \frac{1}{36}a_n^6\|\mathbf{u}\|^6\sum_r\sum_s\sum_t\{(\sum_i\frac{\partial^3}{\partial\alpha_r\partial\alpha_s\partial\alpha_t}l_{(n)i}(\boldsymbol{\alpha}_n^*))^2\} \\
&\le \frac{1}{36}a_n^6 n\|\mathbf{u}\|^6\sum_r\sum_s\sum_t\sum_i M_{nrst}^2(V_{(n)i}),
\end{aligned}
$$

then for any $\varepsilon > 0$,

$$
\begin{aligned}
&\mathrm{P}(\frac{|\frac{1}{6}a_n^3\nabla_{\boldsymbol{\alpha}}^T(\mathbf{u}^T\nabla_{\boldsymbol{\alpha}}^2 l_{(n)}(\boldsymbol{\alpha}_n^*)\mathbf{u})\mathbf{u}|}{a_n^2 n} > \varepsilon) \\
&\le \frac{1}{\varepsilon^2 a_n^4 n^2}\mathrm{E}(\{\frac{1}{6}a_n^3\nabla_{\boldsymbol{\alpha}}^T(\mathbf{u}^T\nabla_{\boldsymbol{\alpha}}^2 l_{(n)}(\boldsymbol{\alpha}_n^*)\mathbf{u})\mathbf{u}\}^2) \\
&= O(\frac{p_n^4}{n}) \to 0,
\end{aligned}
$$

and

$$
|\frac{1}{6}a_n^3\nabla_{\boldsymbol{\alpha}}^T(\mathbf{u}^T\nabla_{\boldsymbol{\alpha}}^2 l_{(n)}(\boldsymbol{\alpha}_n^*)\mathbf{u})\mathbf{u}| = o_P(a_n^2 n);
$$

(4)

$$
|2a_n\lambda_n\boldsymbol{\alpha}_{n0}^T\mathbf{K}_n\mathbf{u}| \le 2a_n\lambda_n\|\boldsymbol{\alpha}_{n0}\|\|\mathbf{K}_n\|\|\mathbf{u}\| = O_P(a_n\lambda_n p_n^{\frac{9}{2}}) = o_P(a_n^2 n);
$$

(5)

$$
|a_n^2\lambda_n\mathbf{u}^T\mathbf{K}_n\mathbf{u}| \le a_n^2\lambda_n\|\mathbf{K}_n\|\|\mathbf{u}\|^2 = O_P(a_n^2\lambda_n p_n^4) = o_P(a_n^2 n).
$$

Thus, the second term of $D_{Pn}(\mathbf{u})$ dominates all others uniformly in $\|\mathbf{u}\| = C_P$, which means, by choosing a sufficiently large C_P, $D_{Pn}(\mathbf{u}) < 0$, and inequality (4.7) holds.

Next, since $\|\hat{\boldsymbol{\alpha}}_{Pn} - \boldsymbol{\alpha}_{n0}\| = O_P(a_n)$, we have

$$\begin{aligned}
\mathbf{0} &= \nabla_{\boldsymbol{\alpha}} l_{P(n)}(\hat{\boldsymbol{\alpha}}_{Pn}) \\
&= \nabla_{\boldsymbol{\alpha}} l_{(n)}(\boldsymbol{\alpha}_{n0}) + \nabla^2_{\boldsymbol{\alpha}} l_{(n)}(\boldsymbol{\alpha}_{n0})(\hat{\boldsymbol{\alpha}}_{Pn} - \boldsymbol{\alpha}_{n0}) \\
&\quad + \frac{1}{2}\nabla_{\boldsymbol{\alpha}}\{(\hat{\boldsymbol{\alpha}}_{Pn} - \boldsymbol{\alpha}_{n0})^T \nabla^2_{\boldsymbol{\alpha}} l_{(n)}(\boldsymbol{\alpha}^*_n)(\hat{\boldsymbol{\alpha}}_{Pn} - \boldsymbol{\alpha}_{n0})\} \\
&\quad - 2\lambda_n \mathbf{K}_n \boldsymbol{\alpha}_{n0} - 2\lambda_n \mathbf{K}_n(\hat{\boldsymbol{\alpha}}_{Pn} - \boldsymbol{\alpha}_{n0}),
\end{aligned}$$

where $\boldsymbol{\alpha}^*_n$ is between $\boldsymbol{\alpha}_{n0}$ and $\hat{\boldsymbol{\alpha}}_{Pn}$, or equivalently,

$$\begin{aligned}
&-\frac{1}{n}\{(\nabla^2_{\boldsymbol{\alpha}} l_{(n)}(\boldsymbol{\alpha}_{n0}) - 2\lambda_n \mathbf{K}_n)(\hat{\boldsymbol{\alpha}}_{Pn} - \boldsymbol{\alpha}_{n0}) - 2\lambda_n \mathbf{K}_n \boldsymbol{\alpha}_{n0}\} \\
&= \frac{1}{n}[\nabla_{\boldsymbol{\alpha}} l_{(n)}(\boldsymbol{\alpha}_{n0}) + \frac{1}{2}\nabla_{\boldsymbol{\alpha}}\{(\hat{\boldsymbol{\alpha}}_{Pn} - \boldsymbol{\alpha}_{n0})^T \nabla^2_{\boldsymbol{\alpha}} l_{(n)}(\boldsymbol{\alpha}^*_n)(\hat{\boldsymbol{\alpha}}_{Pn} - \boldsymbol{\alpha}_{n0})\}],
\end{aligned}$$

furthermore,

$$\begin{aligned}
&\sqrt{n}\mathbf{A}_n \bar{\mathbf{I}}^{-\frac{1}{2}}_{(n)0}(\bar{\mathbf{I}}_{(n)0} + \frac{2}{n}\lambda_n \mathbf{K}_n)\{(\hat{\boldsymbol{\alpha}}_{Pn} - \boldsymbol{\alpha}_{n0}) + (\bar{\mathbf{I}}_{(n)0} + \frac{2}{n}\lambda_n \mathbf{K}_n)^{-1}(2\lambda_n \mathbf{K}_n \boldsymbol{\alpha}_{n0})\} \\
&= \frac{1}{\sqrt{n}}\mathbf{A}_n \bar{\mathbf{I}}^{-\frac{1}{2}}_{(n)0}\nabla_{\boldsymbol{\alpha}} l_{(n)}(\boldsymbol{\alpha}_{n0}) + \sqrt{n}\mathbf{A}_n \bar{\mathbf{I}}^{-\frac{1}{2}}_{(n)0}(\mathbf{r}_{n1} + \mathbf{r}_{n2}),
\end{aligned} \tag{4.9}$$

where

$$\mathbf{r}_{n1} = (\frac{1}{n}\nabla^2_{\boldsymbol{\alpha}} l_{(n)}(\boldsymbol{\alpha}_{n0}) + \bar{\mathbf{I}}_{(n)0})(\hat{\boldsymbol{\alpha}}_{Pn} - \boldsymbol{\alpha}_{n0}),$$

and

$$\mathbf{r}_{n2} = \frac{1}{2n}\nabla_{\boldsymbol{\alpha}}\{(\hat{\boldsymbol{\alpha}}_{Pn} - \boldsymbol{\alpha}_{n0})^T \nabla^2_{\boldsymbol{\alpha}} l_{(n)}(\boldsymbol{\alpha}^*_n)(\hat{\boldsymbol{\alpha}}_{Pn} - \boldsymbol{\alpha}_{n0})\}.$$

With equation (4.8), we have

$$\begin{aligned}
\|\mathbf{r}_{n1}\| &\leq \|\frac{1}{n}\nabla^2_{\boldsymbol{\alpha}} l_{(n)}(\boldsymbol{\alpha}_{n0}) + \bar{\mathbf{I}}_{(n)0}\|\|\hat{\boldsymbol{\alpha}}_{Pn} - \boldsymbol{\alpha}_{n0}\| \\
&= o_P(\frac{1}{\sqrt{p_n n}}) = o_P(\frac{1}{\sqrt{n}});
\end{aligned}$$

also,

$$
\begin{aligned}
&\|\mathbf{r}_{n2}\|^2 \\
&= \frac{1}{4n^2}\sum_r[\sum_s\sum_t\{\frac{\partial^3}{\partial\alpha_r\partial\alpha_s\partial\alpha_t}l_{(n)}(\boldsymbol{\alpha}_n^*)(\hat{\alpha}_{Pns}-\alpha_{ns0})(\hat{\alpha}_{Pnt}-\alpha_{nt0})\}]^2 \\
&\le \frac{1}{4n^2}\sum_r[\sum_s\sum_t(\frac{\partial^3}{\partial\alpha_r\partial\alpha_s\partial\alpha_t}l_{(n)}(\boldsymbol{\alpha}_n^*))^2\sum_s\sum_t\{(\hat{\alpha}_{Pns}-\alpha_{ns0})^2(\hat{\alpha}_{Pnt}-\alpha_{nt0})^2\}] \\
&= \frac{1}{4n^2}\|\hat{\boldsymbol{\alpha}}_{Pn}-\boldsymbol{\alpha}_{n0}\|^4\sum_r\sum_s\sum_t(\sum_i\frac{\partial^3}{\partial\alpha_r\partial\alpha_s\partial\alpha_t}l_{(n)i}(\boldsymbol{\alpha}_n^*))^2 \\
&\le \frac{1}{4n^2}\|\hat{\boldsymbol{\alpha}}_{Pn}-\boldsymbol{\alpha}_{n0}\|^4\sum_r\sum_s\sum_t(n\sum_i M_{nrst}^2(V_{(n)i})) \\
&= O_P(\frac{p_n^5}{n^2}) = o_P(\frac{1}{n}),
\end{aligned}
$$

then for any $\varepsilon > 0$,

$$
\mathrm{P}(\sqrt{n}\|\mathbf{r}_{n2}\| > \varepsilon) \le \frac{n}{\varepsilon^2}\mathrm{E}(\|\mathbf{r}_{n2}\|^2) = O(\frac{p_n^5}{n}) \to 0,
$$

and

$$
\|\mathbf{r}_{n2}\| = o_P(\frac{1}{\sqrt{n}}),
$$

thus,

$$
\|\mathbf{r}_n\| = \|\mathbf{r}_{n1}+\mathbf{r}_{n2}\| \le \|\mathbf{r}_{n1}\| + \|\mathbf{r}_{n2}\| = o_P(\frac{1}{\sqrt{n}}).
$$

Since

$$
\begin{aligned}
\|\sqrt{n}\mathbf{A}_n\bar{\mathbf{I}}_{(n)0}^{-\frac{1}{2}}\mathbf{r}_n\|^2 &= n\mathbf{r}_n^T\bar{\mathbf{I}}_{(n)0}^{-\frac{1}{2}}\mathbf{A}_n^T\mathbf{A}_n\bar{\mathbf{I}}_{(n)0}^{-\frac{1}{2}}n\mathbf{r}_n \\
&\le n\lambda_{\max}(\mathbf{A}_n^T\mathbf{A}_n)\lambda_{\max}(\bar{\mathbf{I}}_{(n)0}^{-1})\|\mathbf{r}_n\|^2 \\
&= n\|\mathbf{r}_n\|^2\lambda_{\max}(\mathbf{A}_n^T\mathbf{A}_n)\lambda_{\min}^{-1}(\bar{\mathbf{I}}_{(n)0}) \\
&= o_P(1),
\end{aligned}
$$

which implies

$$
\sqrt{n}\mathbf{A}_n\bar{\mathbf{I}}_{(n)0}^{-\frac{1}{2}}\mathbf{r}_n = o_P(\mathbf{1}),
$$

equation (4.9) then yields

$$\sqrt{n}\mathbf{A}_n\bar{\mathbf{I}}_{(n)0}^{-\frac{1}{2}}(\bar{\mathbf{I}}_{(n)0}+\frac{2}{n}\lambda_n\mathbf{K}_n)\{(\hat{\boldsymbol{\alpha}}_{Pn}-\boldsymbol{\alpha}_{n0})+(\bar{\mathbf{I}}_{(n)0}+\frac{2}{n}\lambda_n\mathbf{K}_n)^{-1}(2\lambda_n\mathbf{K}_n\boldsymbol{\alpha}_{n0})\}$$
$$=\frac{1}{\sqrt{n}}\mathbf{A}_n\bar{\mathbf{I}}_{(n)0}^{-\frac{1}{2}}\nabla_{\boldsymbol{\alpha}}l_{(n)}(\boldsymbol{\alpha}_{n0})+o_P(\mathbf{1}).$$

By letting

$$\mathbf{h}_{ni}=\frac{1}{\sqrt{n}}\mathbf{A}_n\bar{\mathbf{I}}_{(n)0}^{-\frac{1}{2}}\nabla_{\boldsymbol{\alpha}}l_{(n)i}(\boldsymbol{\alpha}_{n0}),$$

we have

$$\mathrm{E}(\mathbf{h}_{ni})=\mathbf{0},$$

$$\mathrm{Var}(\mathbf{h}_{ni})=\frac{1}{n}\mathbf{A}_n\bar{\mathbf{I}}_{(n)0}^{-\frac{1}{2}}\mathbf{I}_{(n)i}\bar{\mathbf{I}}_{(n)0}^{-\frac{1}{2}}\mathbf{A}_n^T,$$

and

$$\sum_i\mathrm{Var}(\mathbf{h}_{ni})=\mathbf{A}_n\mathbf{A}_n^T\to\mathbf{G}.$$

We have that, for any $\varepsilon>0$,

$$\mathrm{P}(\|\mathbf{h}_{ni}\|>\varepsilon)\le\frac{1}{\varepsilon^2}\mathrm{E}(\|\mathbf{h}_{ni}\|^2)=\frac{1}{\varepsilon^2}\mathrm{tr}(\frac{1}{n}\mathbf{A}_n\bar{\mathbf{I}}_{(n)0}^{-\frac{1}{2}}\mathbf{I}_{(n)i}\bar{\mathbf{I}}_{(n)0}^{-\frac{1}{2}}\mathbf{A}_n^T),$$

then

$$\sum_i\mathrm{P}(\|\mathbf{h}_{ni}\|>\varepsilon)\le\frac{1}{\varepsilon^2}\mathrm{tr}(\mathbf{A}_n\mathbf{A}_n^T)=O(1).$$

Also, since

$$\begin{aligned}\|\mathbf{h}_{ni}\|^2&=\frac{1}{n}\nabla_{\boldsymbol{\alpha}}^Tl_{(n)i}(\boldsymbol{\alpha}_{n0})\bar{\mathbf{I}}_{(n)0}^{-\frac{1}{2}}\mathbf{A}_n^T\mathbf{A}_n\bar{\mathbf{I}}_{(n)0}^{-\frac{1}{2}}\nabla_{\boldsymbol{\alpha}}l_{(n)i}(\boldsymbol{\alpha}_{n0})\\&\le\frac{1}{n}\lambda_{\max}(\mathbf{A}_n^T\mathbf{A}_n)\lambda_{\max}(\bar{\mathbf{I}}_{(n)0}^{-1})\|\nabla_{\boldsymbol{\alpha}}^Tl_{(n)i}(\boldsymbol{\alpha}_{n0})\|^2\\&=\frac{1}{n}\lambda_{\max}(\mathbf{A}_n\mathbf{A}_n^T)\lambda_{\min}^{-1}(\bar{\mathbf{I}}_{(n)0})\|\nabla_{\boldsymbol{\alpha}}^Tl_{(n)i}(\boldsymbol{\alpha}_{n0})\|^2,\end{aligned}$$

we have

$$\begin{aligned}\mathrm{E}(\|\mathbf{h}_{ni}\|^4) &\leq \frac{1}{n^2}\lambda^2_{\max}(\mathbf{A}_n\mathbf{A}_n^T)\lambda^{-2}_{\min}(\bar{\mathbf{I}}_{(n)0})\mathrm{tr}(\mathrm{E}((\nabla_{\boldsymbol{\alpha}} l_{(n)i}(\boldsymbol{\alpha}_{n0})\nabla^T_{\boldsymbol{\alpha}} l_{(n)i}(\boldsymbol{\alpha}_{n0}))^2)) \\ &= O(\frac{p_n^2}{n^2}).\end{aligned}$$

Hence,

$$\begin{aligned}&\sum_i \mathrm{E}(\|\mathbf{h}_{ni}\|^2 \mathrm{I}(\|\mathbf{h}_{ni}\| > \varepsilon)) \\ &\leq \sum_i \sqrt{\mathrm{E}(\|\mathbf{h}_{ni}\|^4)\mathrm{P}(\|\mathbf{h}_{ni}\| > \varepsilon)} \\ &\leq \sqrt{\sum_i \mathrm{E}(\|\mathbf{h}_{ni}\|^4) \sum_i \mathrm{P}(\|\mathbf{h}_{ni}\| > \varepsilon)} \\ &= O(\frac{p_n}{\sqrt{n}}) \to 0,\end{aligned}$$

which implies that $\mathbf{h}_{ni}$'s satisfy the conditions of the multivariate Lindeberg-Feller central limit theorem, and

$$\frac{1}{\sqrt{n}}\mathbf{A}_n\bar{\mathbf{I}}^{-\frac{1}{2}}_{(n)0}\nabla_{\boldsymbol{\alpha}} l_{(n)}(\boldsymbol{\alpha}_{n0}) = \sum_i \mathbf{h}_{ni} \xrightarrow{d} \mathrm{N}(\mathbf{0}, \mathbf{G}).$$

[Note 1] For condition (D1), it is only required for $n^{-1}p_n^4 = o(1)$ in order to prove the consistency.

□

Chapter 5

Modeling Scalar Response with Zero-Inflated Longitudinal Covariates

Count-valued data are often modeled through discrete distributions, such as Poisson, binomial and negative binomial distributions. However, in many psychology and sociology studies, it is common to encounter data with an abundance of zeros, where the proportion of observations with zero is much higher than standard distributions could allow. In such problems, the data are considered to be zero inflated relative to ordinary count distributions, and special treatments that account for zero-inflation are required in modeling.

In particular, for a dataset $\{(Y_i, \mathbf{Z}_i, \mathbf{W}_i)\}$ studied in Chapter 3, the longitudinal covariates W_{ij}'s might suggest a significant abundance of zeros. For example, in the youth alcohol abuse data from the Michigan Longitudinal Study (MLS), the longitudinal covariate, number of drinking days in one typical month, has a large proportion of zero values. In this chapter, we extend the model in Chapter 3 for

longitudinal covariate processes of zero-inflated count values.

This chapter is organized as follows. In Section 5.1, we introduce the particular model with zero-inflated longitudinal covariates. Section 5.2 presents the inference procedure with details. We provide simulation studies to evaluate the proposed estimation in Section 5.3. An application to an alcohol study of the MLS is included in Section 5.4, to further demonstrate the proposed procedure.

5.1 Model and Notation

Suppose that a longitudinal dataset of $\{(Y_i, \mathbf{Z}_i, \mathbf{W}_i)\}$ is collected from n subjects. For i-th subject, we have a response variable Y_i, a p-dimensional vector $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{ip})^T$ of time-invariant covariates, and an observed longitudinal covariate process $\mathbf{W}_i = (W_{i1}, \ldots, W_{in_i})^T$ consisting of count-valued measurements at time points $\{t_{ij}\} \in [T_1, T_2]$.

In Chapter 3, we model the count-valued longitudinal covariates W_{ij}'s through

$$W_{ij} \sim f_W(\cdot; \mu_{ij}),$$

where $f_W(\cdot)$ is a distribution from the exponential family. However, when W_{ij}'s suggest an abundance of zeros, such $f_W(\cdot)$ is often unable to correctly specify the distribution. Thus, a mixture model is usually considered for zero-inflated data, in order to address both the abundance of zeros and the distribution of non-zero values. Various two-part models have been developed; the hurdle model (Mullahy, 1986) and the zero-inflated model (Lambert, 1992) are most commonly used in practice.

Considering its flexibility in accommodating both zero inflation and zero deflation situations, and computational efficiency under certain settings, the hurdle

model will be adopted in our estimation. The hurdle model is actually a mixture of a degenerate zero distribution and a zero-truncated count distribution F^+, and gives the probability function as

$$\begin{aligned} &\mathrm{P}(X=0) = 1-\pi, \\ &\mathrm{P}(X=x) = \pi \frac{f(x)}{1-f(0)}, \text{ for } x>0, \end{aligned}$$

where $f(\cdot)$ is the probability function of the corresponding untruncated distribution F.

In particular, we will model the observed longitudinal covariate processes $\mathbf{W}_i$'s by a hurdle model with zero-truncated Poisson distribution, that is,

$$\begin{aligned} &\mathrm{P}(W_{ij}=0) = 1-\pi_i, \\ &\mathrm{P}(W_{ij}=w) = \pi_i \frac{e^{-\mu_{ij}}\mu_{ij}^w}{w!(1-e^{-\mu_{ij}})}, \text{ for } w>0, \end{aligned} \tag{5.1}$$

and

$$\begin{aligned} \mathrm{logit}(\pi_i) &= u_i, \\ \log(\mu_{ij}) &= x_i(t_{ij}). \end{aligned} \tag{5.2}$$

Here, u_i is a zero component, $x_i(\cdot)$ characterizes the zero-truncated Poisson effect, and u_i's and $x_i(\cdot)$'s are assumed to be mutually independent; in addition, canonical link functions of $\mathrm{logit}(\cdot)$ and $\log(\cdot)$ are employed.

Correspondingly, the response Y_i will be related to u_i, $x_i(\cdot)$ and $\mathbf{Z}_i$ through a generalized functional linear model as

$$Y_i \sim f_Y(\cdot;\eta_i), \tag{5.3}$$

and

$$g(\eta_i) = \mathbf{Z}_i^T \boldsymbol{\delta} + u_i\beta + \int_{T_1}^{T_2} x_i(t)\gamma(t)dt, \tag{5.4}$$

where $f_Y(\cdot)$ is a distribution belonging to an exponential family, $g(\cdot)$ is a known link function, $\boldsymbol{\delta}$ is a p-dimensional vector of regression coefficients for the time-invariant covariate $\mathbf{Z}_i$, β is a coefficient of u_i characterizing the effect of zero component, and $\gamma(\cdot)$ is an effect function of $x_i(\cdot)$ that delineates the time-varying effect of covariate process. If Y_i is Gaussian with the identity link function, an equivalent partial functional linear model is instead assumed as

$$Y_i = \mathbf{Z}_i^T \boldsymbol{\delta} + u_i\beta + \int_{T_1}^{T_2} x_i(t)\gamma(t)dt + \epsilon_i, \tag{5.5}$$

where ϵ_i's are independent and identically distributed as $\mathrm{N}(0, \sigma_\epsilon^2)$.

5.2 Estimation Procedure

We assume that $x_i(\cdot)$'s and $\gamma(\cdot)$ belong to the function space of $\mathcal{S}_{\mathrm{NCS}}$ as defined in Chapter 3, and propose to estimate $\boldsymbol{\delta}$, β and $\gamma(\cdot)$ with a two-stage calibration regression procedure: we first obtain $\hat{u}_i$'s and $\hat{\mathbf{x}}_i$'s from model (5.1,5.2), then estimate $\boldsymbol{\delta}$, β and $\gamma(\cdot)$ by fitting a calibration model of

$$g(\eta_i) = \mathbf{Z}_i^T \boldsymbol{\delta} + \hat{u}_i\beta + \hat{\mathbf{x}}_i^T \mathbf{C}\boldsymbol{\gamma},$$

or

$$g(\boldsymbol{\eta}) = \mathbf{Z}\boldsymbol{\delta} + \hat{\mathbf{u}}\beta + \hat{\mathbf{X}}\boldsymbol{\gamma}, \tag{5.6}$$

where $g(\boldsymbol{\eta}) = (g(\eta_1), \ldots, g(\eta_n))^T$, $\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_n)^T$, $\hat{\mathbf{u}} = (\hat{u}_1, \ldots, \hat{u}_n)^T$ and $\hat{\mathbf{X}} = (\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_n)^T\mathbf{C}$ are stacked over i.

5.2.1 Stage-I

We assume that, both the zero component u_i and the positive component $x_i(\cdot)$ can be decomposed into a population part and a random subject-specific part, that is,

$$u_i = u_0 + v_i,$$

where v_i's are independent and identically distributed as $\mathrm{N}(0, \sigma_v^2)$, and

$$x_i(\cdot) = x_0(\cdot) + d_i(\cdot),$$

where $x_0(\cdot)$ and $d_i(\cdot)$'s are natural cubic spline functions, and $d_i(\cdot)$'s are in addition independent mean-zero Gaussian processes. Thus, model (5.2) has an equivalent expression as

$$\begin{aligned} \mathrm{logit}(\pi_i) &= u_0 + v_i, \\ \log(\mu_{ij}) &= x_0(t_{ij}) + d_i(t_{ij}). \end{aligned} \tag{5.7}$$

Provided that u_i's and $x_i(\cdot)$'s are mutually independent, as well as v_i's and $d_i(\cdot)$'s, the two components corresponding to π_i's and μ_{ij}'s can be fitted separately within the hurdle model framework. To be specific, by letting $W_{ij}^* = \mathrm{I}(W_{ij} > 0)$, we estimate u_i's with a generalized linear mixed model (GLMM) as

$$W_{ij}^* \sim \text{Bernoulli}(\pi_i), \tag{5.8}$$

and

$$\mathrm{logit}(\pi_i) = u_0 + v_i; \tag{5.9}$$

simultaneously, $x_i(\cdot)$'s are estimated from all positive W_{ij}'s, which are denoted by

Table 5.1: Example Data from a Youth Alcohol Abuse Study: Illustration for W^*_{ij}'s and $W^+_{i^+j^+}$'s

	13	14	15	16	17	18	19	20
i				W_{ij}				
4			2				4	
14	0	0	1	1	4		6	4
15	0	0	0	0	10	1	6	
70		0		0	0			
i				W^*_{ij}				
4			1				1	
14	0	0	1	1	1		1	1
15	0	0	0	0	1	1	1	
70		0		0	0			
i				$W^+_{i^+j^+}$				
4			2				4	
14			1	1	4		6	4
15					10	1	6	
70								

$W^+_{i^+j^+}$'s, with a model of

$$W^+_{i^+j^+} \sim \text{ZeroTruncatedPoisson}(\mu^+_{i^+j^+}), \tag{5.10}$$

and

$$\log(\mu^+_{i^+j^+}) = x_0(t^+_{i^+j^+}) + d_{i^+}(t^+_{i^+j^+}). \tag{5.11}$$

It is worth to note that, in model (5.10,5.11), some observations are dropped out from the dataset due to $W_{ij} = 0$, so we have indices $i^+ = 1, \ldots, n^+$ and $j^+ = 1, \ldots, n^+_{i^+}$. It is often the case that $n^+ < n$, since some certain subject may have the observed longitudinal covariate process of all 0's. An illustration for W^*_{ij}'s and $W^+_{i^+j^+}$'s, based on the example data in Table 3.1, is provided as Table 5.1.

Thus, $\mathbf{t}^{+0} = (t^{+0}_1, \ldots, t^{+0}_{r^+})^T$ is defined as an r^+-dimensional vector of ordered distinct values of all time points $t^+_{i^+j^+}$'s, and the set of basis functions $\mathbf{c}^+(\cdot)$, as well as matrices $\mathbf{K}^+$, $\mathbf{T}^+$, $\mathbf{B}^+$, $\mathbf{B}^+_*$, $\mathbf{C}^+$, are correspondingly defined.

The details for the estimation of zero and non-zero components are provided in Sections 5.2.1.1 and 5.2.1.2, respectively. Section 5.2.1.3 then summarizes the procedures for a quick review.

5.2.1.1 Estimation of Zero Components

We apply the penalized quasi-likelihood (PQL) method of Breslow and Clayton (1993) to make inference for the GLMM of (5.8,5.9). Specifically, by stacking over j and i, equation (5.9) can be rewritten as

$$\text{logit}(\boldsymbol{\pi}) = \mathbf{1}^{\pi} u_0 + \mathbf{I}^{\pi}\mathbf{v},$$

where $\boldsymbol{\pi} = (\boldsymbol{\pi}_1^T, \ldots, \boldsymbol{\pi}_n^T)^T$ with $\boldsymbol{\pi}_i = \pi_i \mathbf{1}_{n_i}$, $\mathbf{v} = (v_1, \ldots, v_n)^T$, and $\mathbf{1}^{\pi} = \mathbf{1}_{\sum_{i=1}^n n_i}$, $\mathbf{I}^{\pi} = \text{diag}\{\mathbf{1}_{n_1}, \ldots, \mathbf{1}_{n_n}\}$. By the Laplace approximation, the log-quasi-likelihood is then calculated as

$$ql_{I\pi} = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n_i} \tilde{d}_{ij}^{\pi} - \frac{1}{2\sigma_v^2}\mathbf{v}^T\mathbf{v},$$

where

$$\tilde{d}_{ij}^{\pi} = -2\int_{W_{ij}^*}^{\pi_i} \frac{w_{ij}(W_{ij}^* - s)}{s(1-s)} ds,$$

for known prior weights w_{ij}'s. The estimates $\hat{u}_0$ and $\hat{\mathbf{v}}$ will maximize the log-quasi-likelihood $ql_{I\pi}$, or equivalently, solve the estimating equations of

$$\begin{aligned} &\mathbf{1}^{\pi T}\mathcal{W}^{\pi}\boldsymbol{\Delta}^{\pi}(\mathbf{W}^* - \boldsymbol{\pi}) = 0, \\ &\mathbf{I}^{\pi T}\mathcal{W}^{\pi}\boldsymbol{\Delta}^{\pi}(\mathbf{W}^* - \boldsymbol{\pi}) - \frac{1}{\sigma_v^2}\mathbf{v} = \mathbf{0}, \end{aligned} \tag{5.12}$$

where $\mathbf{W}^* = (\mathbf{W}_1^{*T}, \ldots, \mathbf{W}_n^{*T})^T$ with $\mathbf{W}_i^* = (W_{i1}^*, \ldots, W_{in_i}^*)^T$, and $\boldsymbol{\Delta}^{\pi}$ and $\mathcal{W}^{\pi}$ are diagonal matrices such that $\boldsymbol{\Delta}^{\pi} = \text{diag}\{\boldsymbol{\Delta}_1^{\pi}, \ldots, \boldsymbol{\Delta}_n^{\pi}\}$ with $\boldsymbol{\Delta}_i^{\pi} = \frac{1}{\pi_i(1-\pi_i)}\mathbf{I}_{n_i \times n_i}$,

$\mathcal{W}^{\pi} = \text{diag}\{\mathcal{W}_1^{\pi}, \ldots, \mathcal{W}_n^{\pi}\}$ with $\mathcal{W}_i^{\pi} = \pi_i(1-\pi_i)\text{diag}\{w_{ij}\}$. Considering the mutual independence of v_i's, the estimating equation for $\mathbf{v}$ may be separated as

$$\mathbf{1}_{n_i}^T \mathcal{W}_i^{\pi} \boldsymbol{\Delta}_i^{\pi} (\mathbf{W}_i^* - \pi_i \mathbf{1}_{n_i}) - \frac{1}{\sigma_v^2} v_i = 0,$$

for $i = 1, \ldots, n$.

In estimating u_0 and $\mathbf{v}$, the corresponding linear mixed model (LMM) will be

$$\tilde{\mathbf{W}}^* = \mathbf{1}^{\pi} u_0 + \mathbf{I}^{\pi} \mathbf{v} + \mathbf{e}_{\varepsilon}^*,$$

where $\mathbf{e}_{\varepsilon}^* \sim \text{N}(\mathbf{0}, \mathcal{W}^{\pi -1})$. Thus, we obtain an approximate variance of $\hat{u}_0$ as

$$\text{var}(\hat{u}_0) = \mathbf{1}^{\pi T} \mathbf{R}^{\pi -1} \mathbf{1}^{\pi}, \tag{5.13}$$

where $\mathbf{R}^{\pi} = \sigma_v^2 \mathbf{I}^{\pi} \mathbf{I}^{\pi T} + \mathcal{W}^{\pi -1}$.

Moreover, by defining $\mathbf{P}^{\pi} = \mathbf{R}^{\pi -1} - \mathbf{R}^{\pi -1} \mathbf{1}^{\pi} (\mathbf{1}^{\pi T} \mathbf{R}^{\pi -1} \mathbf{1}^{\pi})^{-1} \mathbf{1}^{\pi T} \mathbf{R}^{\pi -1}$, an estimate of σ_v^2 can be obtained through a resticted maximum likelihood (REML) estimating equation of

$$-\frac{1}{2} \text{tr}(\mathbf{P}^{\pi} \mathbf{I}^{\pi} \mathbf{I}^{\pi T}) + \frac{1}{2} (\tilde{\mathbf{W}}^* - \mathbf{1}^{\pi} \hat{u}_0)^T \mathbf{R}^{\pi -1} \mathbf{I}^{\pi} \mathbf{I}^{\pi T} \mathbf{R}^{\pi -1} (\tilde{\mathbf{W}}^* - \mathbf{1}^{\pi} \hat{u}_0) = 0.$$

5.2.1.2 Estimation of Non-Zero Components

By stacking over j^+, we rewrite equation (5.11) as

$$\log(\boldsymbol{\mu}_{i^+}^+) = \mathbf{N}_{i^+}^+ \mathbf{x}_0^+ + \mathbf{N}_{i^+}^+ \mathbf{d}_{i^+}^+,$$

where $\log(\boldsymbol{\mu}_{i^+}^+) = (\log(\mu_{i^+1}^+), \ldots, \log(\mu_{i^+ n_{i^+}^+}^+))^T$, $\mathbf{x}_0^+ = x_0(\mathbf{t}^{+0})$, $\mathbf{d}_{i^+}^+ = d_{i^+}(\mathbf{t}^{+0})$, and $\mathbf{N}_{i^+}^+$ is an $n_{i^+}^+ \times r^+$ incidence matrix mapping $(t_{i^+1}^+, \ldots, t_{i^+ n_{i^+}^+}^+)^T$ to $\mathbf{t}^{+0}$ such that

(j^+, l)-th element is 1 if $t^+_{i^+j^+} = t^{+0}_l$ and 0 otherwise. A transformation of $\mathbf{d}^+_{i^+} = \mathbf{B}^+_* \mathbf{b}^+_{i^+}$ is then taken to incorporate the smoothness of the random Gaussian process $d_{i^+}(\cdot)$, where $\mathbf{b}^+_{i^+}$'s are independent and identically distributed as $\mathrm{N}(\mathbf{0}, \mathbf{D}^+(\boldsymbol{\theta}))$, for a covariance matrix

$$\mathbf{D}^+(\boldsymbol{\theta}) = \begin{pmatrix} \sigma_1^2 & & \\ & \sigma_2^2 & \\ & & \tau_d \mathbf{I}_{(r^+-2)\times(r^+-2)} \end{pmatrix}$$

with $\boldsymbol{\theta} = (\sigma_1^2, \sigma_2^2, \tau_d)^T$.

With further stacking over i^+, that is, $\log(\boldsymbol{\mu}^+) = (\log(\boldsymbol{\mu}^+_1)^T, \ldots, \log(\boldsymbol{\mu}^+_{n^+})^T)^T$, $\mathbf{N}^+ = ({\mathbf{N}^+_1}^T, \ldots, {\mathbf{N}^+_{n^+}}^T)^T$, and $\mathbf{Q}^+ = \mathrm{diag}\{\mathbf{Q}^+_1, \ldots, \mathbf{Q}^+_{n^+}\}$ with $\mathbf{Q}^+_{i^+} = \mathbf{N}^+_{i^+}\mathbf{B}^+_*$, we have a generalized additive mixed model as

$$\log(\boldsymbol{\mu}^+) = \mathbf{N}^+\mathbf{x}^+_0 + \mathbf{Q}^+\mathbf{b}^+, \tag{5.14}$$

where $\mathbf{b}^+ = ({\mathbf{b}^+_1}^T, \ldots, {\mathbf{b}^+_{n^+}}^T)^T \sim \mathrm{N}(\mathbf{0}, \mathcal{D}^+(\boldsymbol{\theta}))$ with $\mathcal{D}^+ = \mathrm{diag}\{\mathbf{D}^+, \ldots, \mathbf{D}^+\}$.

Generally, $\mathbf{x}^+_0$ and $\mathbf{b}^+$ can be achieved by maximizing the corresponding double penalized quasi-likelihood (DPQL) derived from model (5.10,5.14), in the same fashion as that in Chapter 3. However, $\log(\cdot)$ in equation (5.14) is in fact a pseudo-link function, since it models $\mu^+_{i^+j^+}$ rather than the mean of $W^+_{i^+j^+}$ from the zero-truncated Poisson distribution. In particular, both mean and variance functions can be specified as

$$u_\mu(\mu) = \frac{\mu}{1 - \mathrm{e}^{-\mu}},$$

and

$$v_\mu(\mu) = \frac{\mu(1 - \mathrm{e}^{-\mu} - \mu\mathrm{e}^{-\mu})}{(1 - \mathrm{e}^{-\mu})^2},$$

respectively. Thus, all terms in Chapter 3 that are related with either of the

functions should be modified accordingly.

In particular, based on model (5.10,5.14), we have the DPQL as

$$-\frac{1}{2}\sum_{i^+=1}^{n^+}\sum_{j^+=1}^{n_{i^+}^+}\tilde{d}_{i^+j^+}^{\mu}-\frac{1}{2}\mathbf{b}^{+T}\mathcal{D}^{+-1}\mathbf{b}^{+}-\frac{1}{2}\lambda_x\mathbf{x}_0^{+T}\mathbf{K}^{+}\mathbf{x}_0^{+}, \tag{5.15}$$

where the deviance is

$$\tilde{d}_{i^+j^+}^{\mu}=-2\int_{u_\mu^{-1}(W_{i^+j^+}^+)}^{\mu_{i^+j^+}^+}\frac{w_{i^+j^+}(W_{i^+j^+}^+-u_\mu(s))}{s}ds.$$

By maximizing the DPQL, $\hat{\mathbf{x}}_0^+$ and $\hat{\mathbf{b}}^+$ solve the estimating equations of

$$\mathbf{N}^{+T}\mathcal{W}^{\mu}\boldsymbol{\Delta}^{\mu}(\mathbf{W}^{+}-u_\mu(\boldsymbol{\mu}^{+}))-\lambda_x\mathbf{K}^{+}\mathbf{x}_0^{+}=\mathbf{0},$$
$$\mathbf{Q}^{+T}\mathcal{W}^{\mu}\boldsymbol{\Delta}^{\mu}(\mathbf{W}^{+}-u_\mu(\boldsymbol{\mu}^{+}))-\mathcal{D}^{+-1}\mathbf{b}^{+}=\mathbf{0},$$

where $\mathbf{W}^+=(\mathbf{W}_1^{+T},\ldots,\mathbf{W}_{n^+}^{+T})^T$ with $\mathbf{W}_{i^+}^+=(W_{i^+1}^+,\ldots,W_{i^+n_{i^+}^+}^+)^T$, and $\boldsymbol{\Delta}^{\mu}$ and $\mathcal{W}^{\mu}$ are diagonal matrices such that $\boldsymbol{\Delta}^{\mu}=\mathrm{diag}\{\boldsymbol{\Delta}_1^{\mu},\ldots,\boldsymbol{\Delta}_{n^+}^{\mu}\}$ with $\boldsymbol{\Delta}_{i^+}^{\mu}=\mathrm{diag}\{\frac{1}{\mu_{i^+j^+}^+}\}$, $\mathcal{W}^{\mu}=\mathrm{diag}\{\mathcal{W}_1^{\mu},\ldots,\mathcal{W}_{n^+}^{\mu}\}$ with $\mathcal{W}_{i^+}^{\mu}=\mathrm{diag}\{w_{i^+j^+}\mu_{i^+j^+}^+\}$. The independence of $\mathbf{b}_{i^+}^+$'s then suggests an equivalent set of estimating equations as

$$\mathbf{Q}_{i^+}^{+T}\mathcal{W}_{i^+}^{\mu}\boldsymbol{\Delta}_{i^+}^{\mu}(\mathbf{W}_{i^+}^{+}-u_\mu(\boldsymbol{\mu}_{i^+}^{+}))-\mathbf{D}^{+-1}\mathbf{b}_{i^+}^{+}=\mathbf{0},$$

for $i^+=1,\ldots,n^+$.

In addition, a mixed effects transformation of

$$\mathbf{x}_0^+=\mathbf{T}^+\boldsymbol{\alpha}_x^++\mathbf{B}^+\mathbf{a}_x^+,$$

where $\mathbf{a}_x^+ \sim \mathrm{N}(\mathbf{0}, \lambda_x^{-1}\mathbf{I})$, will indicate an equivalent GLMM of

$$\log(\boldsymbol{\mu}^+) = \mathbf{N}^+\mathbf{T}^+\boldsymbol{\alpha}_x^+ + \mathbf{N}^+\mathbf{B}^+\mathbf{a}_x^+ + \mathbf{Q}^+\mathbf{b}^+.$$

Further, the estimation problem corresponds to an LMM of

$$\tilde{\mathbf{W}}^+ = \mathbf{N}^+\mathbf{T}^+\boldsymbol{\alpha}_x^+ + \mathbf{N}^+\mathbf{B}^+\mathbf{a}_x^+ + \mathbf{Q}^+\mathbf{b}^+ + \mathbf{e}_\varepsilon^+,$$

where $\tilde{\mathbf{W}}^+ = \mathbf{N}^+\mathbf{x}_0^+ + \mathbf{Q}^+\mathbf{b}^+ + \boldsymbol{\Delta}^\mu(\mathbf{W}^+ - u_\mu(\boldsymbol{\mu}^+))$ is a working vector, and $\mathbf{e}_\varepsilon^+ \sim \mathrm{N}(\mathbf{0}, \mathcal{W}^{\mu-1})$. By some calculation, the LMM representation yields an approximate covariance matrix of $\hat{\mathbf{x}}_0^+$ as

$$\mathrm{Cov}(\hat{\mathbf{x}}_0^+) = \mathbf{B}_*^+\mathbf{H}^{+-1}\mathbf{H}_0^+\mathbf{H}^{+-1}\mathbf{B}_*^{+T}, \tag{5.16}$$

where $\mathbf{R}^+ = \mathbf{Q}^+\mathcal{D}^+\mathbf{Q}^{+T} + \mathcal{W}^{\mu-1}$, $\mathbf{H}_0^+ = (\mathbf{N}^+\mathbf{B}_*^+)^T\mathbf{R}^{+-1}(\mathbf{N}^+\mathbf{B}_*^+)$, and

$$\mathbf{H}^+ = \begin{pmatrix} (\mathbf{N}^+\mathbf{T}^+)^T\mathbf{R}^{+-1}(\mathbf{N}^+\mathbf{T}^+) & (\mathbf{N}^+\mathbf{T}^+)^T\mathbf{R}^{+-1}(\mathbf{N}^+\mathbf{B}^+) \\ (\mathbf{N}^+\mathbf{B}^+)^T\mathbf{R}^{+-1}(\mathbf{N}^+\mathbf{T}^+) & (\mathbf{N}^+\mathbf{B}^+)^T\mathbf{R}^{+-1}(\mathbf{N}^+\mathbf{B}^+) + \lambda_x\mathbf{I} \end{pmatrix}.$$

The variance component $\boldsymbol{\theta}$, on the other hand, is estimated by maximizing the approximate marginal log-quasi-likelihood of

$$\begin{aligned} ql_{\mathrm{M}I}^\mu \approx & -\frac{1}{2}\log|\mathcal{V}^\mu| - \frac{1}{2}\log|(\mathbf{N}^+\mathbf{T}^+)^T\mathcal{V}^{\mu-1}(\mathbf{N}^+\mathbf{T}^+)| \\ & -\frac{1}{2}\{\tilde{\mathbf{W}}^+ - (\mathbf{N}^+\mathbf{T}^+)\hat{\boldsymbol{\alpha}}_x^+\}^T\mathcal{V}^{\mu-1}\{\tilde{\mathbf{W}}^+ - (\mathbf{N}^+\mathbf{T}^+)\hat{\boldsymbol{\alpha}}_x^+\}, \end{aligned} \tag{5.17}$$

where $\mathcal{V}^\mu = \frac{1}{\lambda_x}(\mathbf{N}^+\mathbf{B}^+)(\mathbf{N}^+\mathbf{B}^+)^T + \mathbf{R}^+$. Thus, for each $\boldsymbol{\theta}_l$ of $\boldsymbol{\theta} = (\sigma_1^2, \sigma_2^2, \tau_d)^T$, $\hat{\boldsymbol{\theta}}_l$ solves the estimating equation of

$$-\frac{1}{2}\mathrm{tr}(\mathbf{P}^+\frac{\partial\mathbf{R}^+}{\partial\boldsymbol{\theta}_l}) + \frac{1}{2}(\tilde{\mathbf{W}}^+ - \mathbf{N}^+\hat{\mathbf{x}}_0^+)^T\mathbf{R}^{+-1}\frac{\partial\mathbf{R}^+}{\partial\boldsymbol{\theta}_l}\mathbf{R}^{+-1}(\tilde{\mathbf{W}}^+ - \mathbf{N}^+\hat{\mathbf{x}}_0^+) = 0,$$

where $\mathbf{P}^+ = \mathbf{R}^{+-1} - \mathbf{R}^{+-1}(\mathbf{N}^+\mathbf{B}_*^+)\mathbf{H}^{+-1}(\mathbf{N}^+\mathbf{B}_*^+)^T\mathbf{R}^{+-1}$.

The smoothing parameter, λ_x, is selected by GCV. Explicitly, from $\hat{\boldsymbol{\mu}}^+$ of equation (5.14) evaluated at $\hat{\mathbf{x}}_0^+$ and $\hat{\mathbf{b}}^+$, we have

$$\mathbf{A}^+ = \frac{\partial^2}{\partial \mathbf{x}_0^+ \partial \mathbf{x}_0^{+T}}\left(\frac{1}{2}\sum_{i^+=1}^{n^+}\sum_{j^+=1}^{n_{i^+}^+}\tilde{d}_{i^+j^+}^{\mu}\right) = \mathbf{N}^{+T}\mathcal{A}^+\mathbf{N}^+,$$

where $\mathcal{A}^+ = \text{diag}\{\mathcal{A}_1^+, \ldots, \mathcal{A}_{n^+}^+\}$ with $\mathcal{A}_{i^+}^+ = \text{diag}\{w_{i^+j^+}\hat{\mu}_{i^+j^+}^+ u_{\mu}'(\hat{\mu}_{i^+j^+}^+)\}$. Thus, the effective degrees of freedom are

$$edf_x = \text{tr}((\mathbf{A}^+ + \lambda_x\mathbf{K}^+)^{-1}\mathbf{A}^+),$$

and the GCV statistic is

$$GCV(\lambda_x) = \frac{\frac{1}{2}\sum_{i^+=1}^{n^+}\sum_{j^+=1}^{n_{i^+}^+}\tilde{d}_{i^+j^+}^{\mu}}{\dot{n}^+(1-\frac{edf_x}{N^+})^2}, \tag{5.18}$$

where $\dot{n}^+ = \sum_{i^+=1}^{n^+} n_{i^+}^+$ is the total number of observations of $W_{ij} > 0$ (i.e., $W_{i^+j^+}^+$'s). The optimal λ_x will be searched by minimizing $GCV(\lambda_x)$ over a grid of points.

Following the above procedure, we have $\hat{\mathbf{x}}_{i^+}^+ = \hat{\mathbf{x}}_0^+ + \hat{\mathbf{b}}_{i^+}^+$, for $i^+ = 1, \ldots, n^+$. In addition, it is not possible to figure out $\mathbf{d}_i = \mathbf{B}_*\mathbf{b}_i$ for some subject i who has the observed longitudinal covariate process of all 0's, so we set $\hat{\mathbf{x}}_i$ to take the value of the estimated population profile $\hat{\mathbf{x}}_0^+$ as the best estimation. Hence, we obtain $\hat{\mathbf{x}}_i$ for all subjects.

5.2.1.3 Summary of Stage-I

In the first stage, we obtain $\hat{u}_i$'s and $\hat{x}_i(\cdot)$'s for model (5.1,5.2) through two independent estimation procedures. For the zero component, $\hat{u}_i$'s are obtained from a GLMM of (5.8,5.9) using W_{ij}^*'s. To be specific, by the PQL method, $\hat{u}_0$ and $\hat{v}_i$'s

solve the estimating equations of (5.12), and an approximate variance of $\hat{u}_0$ is given as equation (5.13). In the estimation of non-zero component, $\hat{x}_i(\cdot)$'s are achieved by fitting $W^+_{i^+j^+}$'s in a GAMM of (5.10,5.11). Explicitly, $\hat{\mathbf{x}}^+_0$ and $\hat{\mathbf{b}}^+$ maximize the DPQL of (5.15); equation (5.16) gives an approximate covariance matrix for $\hat{\mathbf{x}}^+_0$; the variance component is estimated by the approximate marginal log-quasi-likelihood of (5.17); and equation (5.18) defines GCV for selecting the smoothing parameter λ_x. Finally, $\hat{\mathbf{x}}_i$'s are obtained by $\hat{\mathbf{x}}^+_{i^+} = \hat{\mathbf{x}}^+_0 + \hat{\mathbf{b}}^+_{i^+}$ for $i^+ = 1, \ldots, n^+$, or $\hat{\mathbf{x}}^+_0$ as the best estimation if i-th subject has the longitudinal covariate process of all 0's.

5.2.2 Stage-II

The estimates $\hat{u}_i$'s and $\hat{\mathbf{x}}_i$'s achieved in the first stage will allow us to estimate $\boldsymbol{\delta}$, β and $\gamma(\cdot)$ based on the calibration model of (5.3,5.6). Particularly, by denoting $\mathbf{Z}^{\text{new}} = (\mathbf{Z}, \hat{\mathbf{u}})$ and $\boldsymbol{\delta}^{\text{new}} = (\boldsymbol{\delta}^T, \beta)^T$, we rewrite equation (5.6) as

$$g(\boldsymbol{\eta}) = \mathbf{Z}^{\text{new}}\boldsymbol{\delta}^{\text{new}} + \hat{\mathbf{X}}\boldsymbol{\gamma}.$$

When responses Y_i's are considered as Gaussian with the identity link function, and model (5.5) is assumed, the calibration model then becomes

$$\mathbf{Y} = \mathbf{Z}^{\text{new}}\boldsymbol{\delta}^{\text{new}} + \hat{\mathbf{X}}\boldsymbol{\gamma} + \boldsymbol{\epsilon}^*,$$

where ϵ^*_i's are treated as independent with $\boldsymbol{\epsilon}^* \sim \text{N}(\mathbf{0}, \sigma^2_{\epsilon^*}\mathbf{I})$. These representations coincide with the models discussed in Chapter 3. Hence, same inference procedures described in Section 3.2.2 will be adopted to obtain $\hat{\boldsymbol{\delta}}$, $\hat{\beta}$ and $\hat{\gamma}(\cdot)$.

5.3 Simulation Studies

In this section, we conduct two sets of simulation experiments to evaluate the proposed estimation method for models with longitudinal covariates of zero-modified count values. Section 5.3.1 is designed to assess the estimation performance with different settings of sample size, number of knots, and proportion of zeros. In Section 5.3.2, we evaluate the potential negative consequence of ignoring zero-inflation, by fitting a Poisson model on longitudinal covariate data generated according to a zero-inflated model.

5.3.1 Simulation 1: Evaluating Performance of Proposed Model

In this experiment, three sample sizes, small, medium and large, are considered, for n being 250, 500 and 750, respectively. We set $\mathbf{t}^0$ to be r equally spaced time points in $[-1.5, 1.5]$, where r is 8, 16 or 32. In order to simulate unbalanced observation time points t_{ij}'s among n subjects, we allow the longitudinal covariate process $\mathbf{W}_i$ for i-th subject to be observed at $\{t_{ij}\}$ for $j = 1, \ldots, n_i$, where n_i is an integer randomly chosen from $\{\lceil \frac{3}{4}r \rceil, \ldots, r\}$, and t_{ij}'s are n_i distinct points randomly drawn $\mathbf{t}^0$.

The longitudinal covariate data W_{ij}'s are zero-inflated counts generated by model (5.1,5.7). Here, we set u_0 to be -0.75 or 0.75, $\sigma_v^2 = 1^2$, and $x_0(t) = 0.5 + 0.8 \arctan(2t)$; $d_i(\cdot)$ is determined by $\mathbf{d}_i = \mathbf{B}_* \mathbf{b}_i$ with $\mathbf{b}_i$'s being a random sample from $\mathrm{N}(\mathbf{0}, \mathrm{diag}\{\sigma_1^2, \sigma_2^2, \tau_d \mathbf{I}_{(r-2)\times(r-2)}\})$, where $\sigma_1^2 = 0.6^2$, $\sigma_2^2 = 0.3^2$ and $\tau_d = 1^2$. Particularly, when $u_0 = -0.75$, approximately 65% of W_{ij}'s are 0's, while $u_0 = 0.75$ corresponds to around 35% zero proportion, which is a situation close to a regular Poisson distribution.

The response data Y_i's are then generated from

$$Y_i \sim \text{Bernoulli}(\eta_i),$$

and

$$\text{logit}(\eta_i) = Z_i\delta + u_i\beta + \int_{-1.5}^{1.5} x_i(t)\gamma(t)dt,$$

where $\beta = 1.25$, $\delta = 1.5$, and $\gamma(t) = -1.5t^3 + t^2 + 1.5t - 0.5$. In addition, Z_i's are randomly drawn from $\{0, 1\}$ with $\text{P}(Z_i = 1) = \frac{2}{3}$ for each subject.

In summary, we manipulate three factors, the sample size (n), the number of time points (r) and the coefficient for zero proportion (u_0), and consider 18 situations in total. For each combination, we generate $N = 100$ datasets and apply the proposed two-stage method in estimation. For each parameter, mean squared error (MSE) and its empirical standard error are calculated from N replications; for nonparametric functions $x_0(\cdot)$ and $\gamma(\cdot)$, mean integrated squared error (MISE) is instead used to summarize the estimation results.

Table 5.2: Estimation Results of Simulation 1: varying n ($r = 16$, $u_0 = -0.75$)

	$n = 250$	$n = 500$	$n = 750$
	MSE/MISE (se)		
u_0	0.0071 (0.0086)	0.0048 (0.0058)	0.0039 (0.0040)
β	0.0446 (0.0566)	0.0271 (0.0378)	0.0245 (0.0318)
δ	0.1152 (0.1590)	0.0708 (0.0918)	0.0540 (0.0647)
σ_v^2	0.0248 (0.0311)	0.0204 (0.0212)	0.0150 (0.0157)
σ_1^2	0.0061 (0.0058)	0.0047 (0.0038)	0.0042 (0.0031)
σ_2^2	0.0015 (0.0015)	0.0014 (0.0012)	0.0013 (0.0009)
τ_d	0.1437 (0.1198)	0.1188 (0.0877)	0.1032 (0.0696)
$x_0(\cdot)$	0.0712 (0.0447)	0.0621 (0.0306)	0.0594 (0.0249)
$\gamma(\cdot)$	5.0362 (6.5024)	3.9944 (4.2980)	2.7841 (2.8647)

Table 5.2 presents the results of similations with varied sample size of n, and fixed number of time knots ($r = 16$) and zero coefficient ($u_0 = -0.75$). When the

Table 5.3: Estimation Results of Simulation 1: varing r ($n = 500$, $u_0 = -0.75$)

	$r = 8$	$r = 16$	$r = 32$
	MSE/MISE (se)		
u_0	0.0076 (0.0080)	0.0048 (0.0058)	0.0042 (0.0057)
β	0.0269 (0.0471)	0.0271 (0.0378)	0.0274 (0.0328)
δ	0.0950 (0.0975)	0.0708 (0.0918)	0.0719 (0.1013)
σ_v^2	0.0292 (0.0286)	0.0204 (0.0212)	0.0106 (0.0174)
σ_1^2	0.0075 (0.0060)	0.0047 (0.0038)	0.0031 (0.0027)
σ_2^2	0.0015 (0.0015)	0.0014 (0.0012)	0.0009 (0.0006)
τ_d	0.1683 (0.1369)	0.1188 (0.0877)	0.0797 (0.0515)
$x_0(\cdot)$	0.1084 (0.0496)	0.0621 (0.0306)	0.0405 (0.0233)
$\gamma(\cdot)$	4.1672 (6.8168)	3.9944 (4.2980)	2.5494 (2.8868)

Table 5.4: Estimation Results of Simulation 1: varing u_0 ($n = 500$, $r = 16$)

	$u_0 = -0.75$	$u_0 = 0.75$
	MSE/MISE (se)	
u_0	0.0048 (0.0058)	0.0041 (0.0054)
β	0.0271 (0.0378)	0.0310 (0.0430)
δ	0.0708 (0.0918)	0.0851 (0.0951)
σ_v^2	0.0204 (0.0212)	0.0183 (0.0214)
σ_1^2	0.0047 (0.0038)	0.0034 (0.0028)
σ_2^2	0.0014 (0.0012)	0.0010 (0.0007)
τ_d	0.1188 (0.0877)	0.0773 (0.0501)
$x_0(\cdot)$	0.0621 (0.0306)	0.0418 (0.0209)
$\gamma(\cdot)$	3.9944 (4.2980)	3.3347 (3.7341)

sample size increases, the performance of the proposed model improves in terms of MSE. In Table 5.3, the number of time points r is varing while holding the other two factors constant ($n = 500$, $u_0 = -0.75$). The performance of the proposed estimation improves with more time points. The effect of the zero coefficient u_0 is summarized in Table 5.4 with the other two factors fixed at their medium values ($n = 500$, $r = 16$). By varying the value of u_0 from -0.75 to 0.75, we manipulate the proportion of zeros in the data from a zero-inflated situation to one close to regular Poisson. The result shows that the performance of the proposed model is only slightly better in the regular Poisson situation, which means that our model

can handle the zero-inflation as well as regular settings.

5.3.2 Simulation 2: Evaluating Consequence of Ignoring Zero-Inflation

In this experiment, longitudinal covariates are generated from zero-inflated model (5.1,5.2), but fitted by a Poisson model, that is, model (3.4,3.5) with $f_W(\cdot)$ being a Poisson probability function. The simulation settings are similar as the experiment in Section 5.3.1, except that we only manipulate the proportion of zeros through varying the value of u_0, while holding the sample size and the number of time points at their medium values ($n = 500$, $r = 16$).

Table 5.5: Estimation Results of Simulation 2: varing u_0 ($n = 500$, $r = 16$)

	$u_0 = -0.75$	$u_0 = 0.75$
	MSE/MISE (se)	
u_0	NA	NA
β	NA	NA
δ	0.0981 (0.0993)	0.0502 (0.0605)
σ_v^2	NA	NA
σ_1^2	0.1613 (0.0447)	0.0031 (0.0028)
σ_2^2	0.0090 (0.0051)	0.0001 (0.0001)
τ_d	358.6152 (184.6555)	2.2420 (1.6657)
$x_0(\cdot)$	3.5962 (0.3174)	0.3112 (0.0355)
$\gamma(\cdot)$	3.5142 (1.3032)	4.3391 (3.6713)

Table 5.5 summarizes the results of the simulation. When the longitudinal covariates contain excess zeros ($u_0 = -0.75$), the regular Poisson does not perform as well as in the situation without zero-inflation ($u_0 = 0.75$). In Figure 5.1, we present the population profile $x_0(\cdot)$ and the time-varying effect $\gamma(\cdot)$ under the two situations. The estimated curves are derived from evaluating the fitted functions of the $N = 100$ replications at a set of grid points and connecting means at all grid points. As demonstrated in the left panel, the regular Poisson model estimates

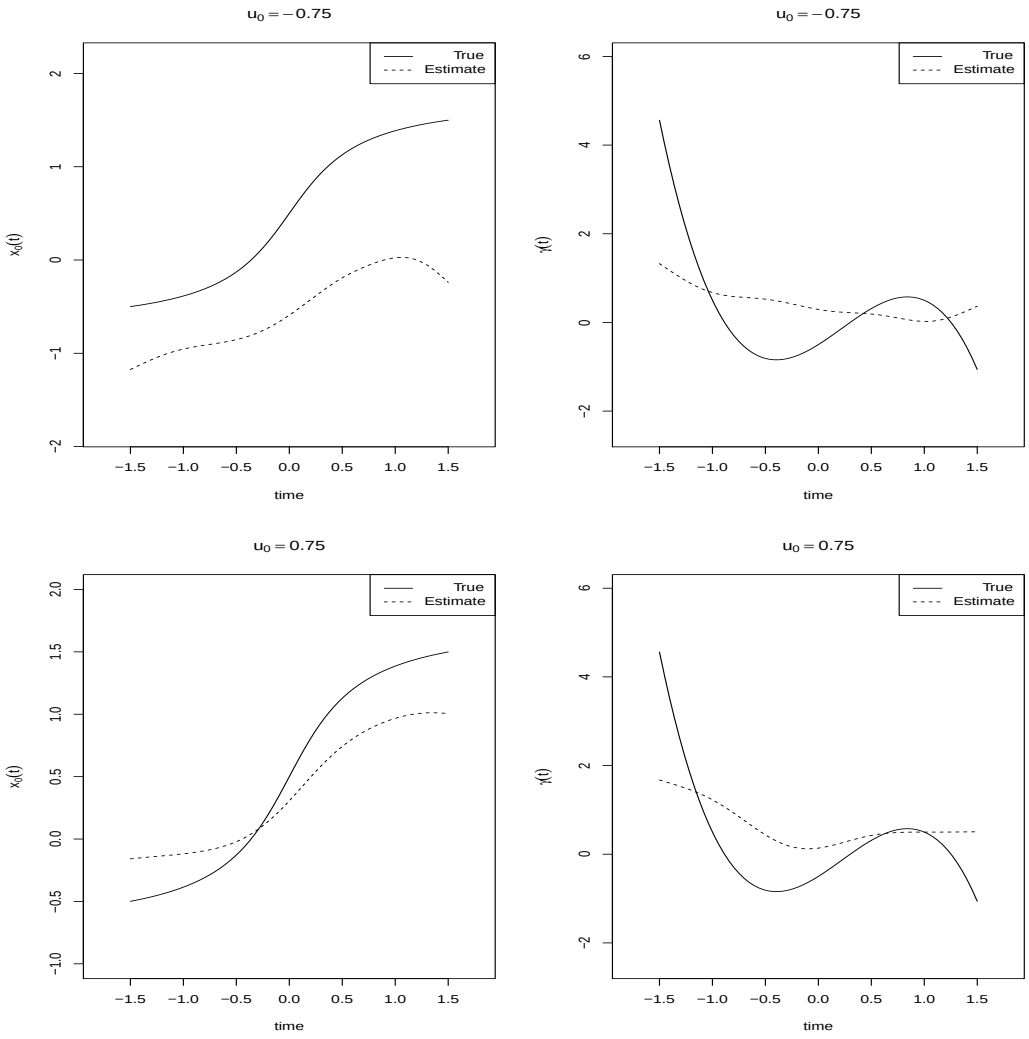

Figure 5.1: Estimated Functions when Ignoring Zero-Inflation: left, population profile function $x_0(\cdot)$; right, effect function $\gamma(\cdot)$

$x_0(\cdot)$ reasonably when there is no significant zero-inflation ($u_0 = 0.75$), whereas the model does poorly in the situation of zero-inflation ($u_0 = -0.75$). On the other hand, the right panel of Figure 5.1 shows that the performance of the regular Poisson does not vary much with the existence of zero-inflation in the data. The regular Poisson model estimates the curve of $\gamma(\cdot)$ to be flat in both situations, whereas the true effect changes developmentally across time.

5.4 Application

In this section, we apply the proposed two-stage procedure to a study of youth at high risk for alcohol abuse from the MLS.

The MLS is an ongoing multi-wave prospective study of people at high risk for substance use disorders (Zucker et al., 1996). This study recruited participating families using fathers' drunk driving conviction records and door-to-door community canvassing in a four-county area in mid-Michigan. All participants received extensive in-home assessments of their substance use, related risk factors and consequences at baseline, and thereafter at a 3-year interval. The children of participating families were followed from early childhood to adulthood. Particularly, during the critical developmental period of alcohol use of ages 13 - 20, these children were assessed annually in order to measure drinking onset and patterns more accurately. In this study, we analyze the longitudinal data from a sample consisting of 508 children, of which 362 are males.

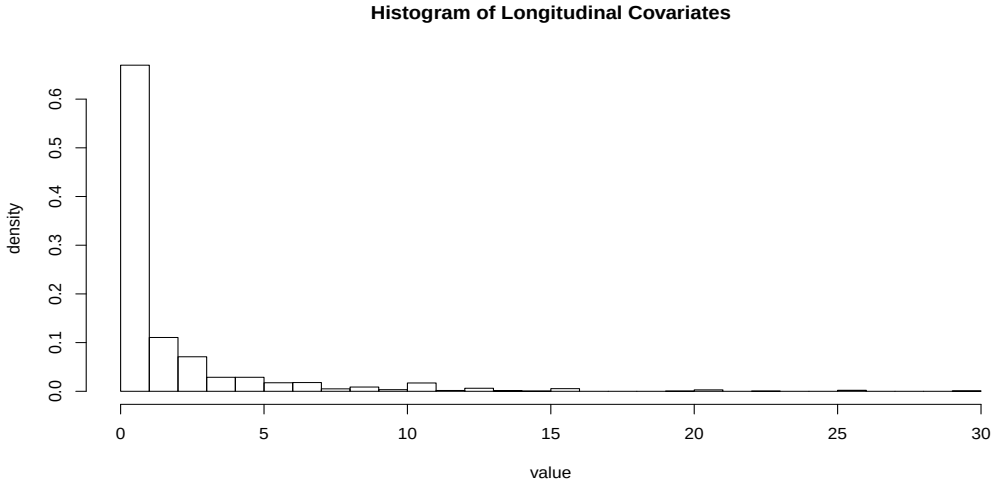

Figure 5.2: Histogram of the Longitudinal Covariates: Number of Drinking Days in One Month from a Youth Alcohol Abuse Study

Our study aims to characterize alcohol use behavior developmentally during

adolescence, and delineate the time-varying effect of adolescent alcohol use patterns on alcohol dependence diagnosis in adulthood. In the analysis, the time-varying covariate (W_{ij}) is the number of drinking days in the past month of assessment. As shown in Figure 5.2, the proportion of zero values is about 67% across waves, which suggests a pattern of zero inflation, thus statistical models for zero-inflated count data are required. The response variable (Y_i) is the DSM-IV alcohol dependence diagnosis (American Psychiatric Association, 1994) during early adulthood, which is a binary variable with 1 for positive diagnosis and 0 otherwise. Since both positive family history of alcoholism and male gender possibly contribute to higher risk for meeting alcohol dependence diagnosis (Buu et al., 2012b), we also include gender (Z_{i1}, 1 for male) and parental lifetime alcohol use disorder diagnosis (Z_{i2}, 1 for positive diagnosis from either biological parent) as two binary time-invariant covariates in the model. Therefore, we analyze the data through model

$$\begin{aligned}
&\mathrm{P}(W_{ij}=0) = 1-\pi_i,\\
&\mathrm{P}(W_{ij}=w) = \pi_i \frac{e^{-\mu_{ij}}\mu_{ij}^{w}}{w!(1-e^{-\mu_{ij}})},\ \text{for } w>0,
\end{aligned}$$

$$\begin{aligned}
\mathrm{logit}(\pi_i) &= u_i,\\
\log(\mu_{ij}) &= x_i(t_{ij}),
\end{aligned}$$

and

$$\begin{aligned}
Y_i &\sim \text{Bernoulli}(\eta_i),\\
\mathrm{logit}(\eta_i) &= Z_{i1}\delta_1 + Z_{i2}\delta_2 + u_i\beta + \int_{13}^{20} x_i(t)\gamma(t)dt,
\end{aligned}$$

where δ_1 and δ_2 are coefficients for gender and parental alcohol use, β characterizes the effect from individual alcohol use risk, and $\gamma(\cdot)$ delineates the time-varying effect of adolescent alcohol use.

In the analysis, we first standardize the original time points $\{13, \ldots, 20\}$ into $\mathbf{t}^0 = (-1.4289, -1.0206, \ldots, 1.4289)^T$, and conduct estimation under this scale, then convert results, particularly $\hat{x}_0(\cdot)$ and $\hat{\gamma}(\cdot)$, to the original scale. We fit the proposed model on the MLS data, have the overall developmental trajectory of alcohol use behavior during adolescence $x_0(\cdot)$, and the time-varying effect of adolescent alcohol use on adult alcohol dependence diagnosis $\gamma(\cdot)$ in Figure 5.3. As shown by the left panel, in general, the frequency of alcohol use increases during adolescence, and the curve becomes flat when the development approaches early adulthood. The right panel shows that drinking behavior at early adolescence (i.e., before age 15) has the highest effect on adult alcohol dependence, while the time-varying effect has a small peak around the time when most of the youth enter college (i.e., age 18).

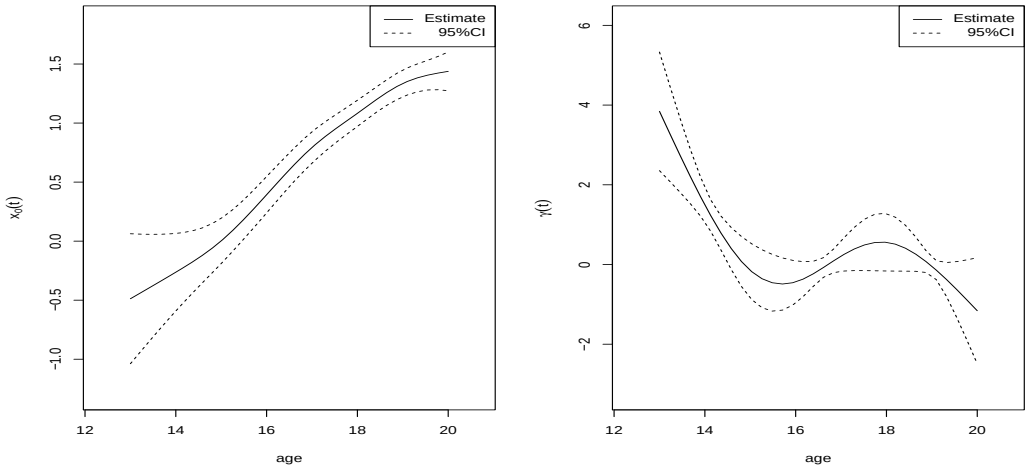

Figure 5.3: Estimated Functions for the Youth Alcohol Abuse Study Dataset: left, the overall developmental trajectory of alcohol use behavior during adolescence, $x_0(\cdot)$; right, the time-varying effect of adolescent alcohol use on adult alcohol dependence diagnosis, $\gamma(\cdot)$

In terms of the effects for time-invariant covariates, our model indicates that males are at higher risk for meeting alcohol dependence criteria ($\hat{\delta}_1 = 1.4353$

with 95% confidence interval (CI) of $[0.8586, 2.0120]$); however, parental alcohol use disorder does not have a significant effect ($\hat{\delta}_2 = 0.3256$ with 95% CI of $[-0.2194, 0.8706]$). Individual risk for alcohol use contributes to greater likelihood of adult alcohol dependence ($\hat{\beta} = 1.2449$ with 95% CI of $[0.8973, 1.5924]$). Our analysis also estimates the population risk level u_0 to be -0.7310 with 95% CI of $[-0.8637, -0.5983]$, which indicates an abundance of zeros. Moreover, the variance components in the model are estimated as $\hat{\sigma}_v^2 = 1.0630^2$, $\hat{\sigma}_1^2 = 0.6198^2$, $\hat{\sigma}_2^2 = 0.3057^2$, $\hat{\tau}_d = 1.7353^2$, and the smoothing parameters are selected as $\lambda_x = 1.3$, and $\lambda_\gamma = 0.003$.

Chapter 6

Future Research

Some possible future research topics are discussed in this chapter. We will further explore the theoretical properties of the estimates of regression coefficient and time-varying effect function in the calibration model framework, improve the existing estimation procedures, as well as extend the current models for more general settings.

6.1 Properties of Estimated Regression Coefficient and Effect Function

Following Section 3.2.1, one may obtain $\tilde{x}_i(\cdot)$, or $\tilde{\mathbf{x}}_i = \mathrm{E}(\mathbf{x}_i|\mathbf{W})$, from $\mathbf{W}$, then estimate $\boldsymbol{\delta}$ and $\gamma(\cdot)$ by fitting the calibration model (3.9). However, from equation (3.7), the true model should be

$$g(\eta_i) = \mathbf{Z}_i^T\boldsymbol{\delta} + \tilde{\mathbf{x}}_i^T\mathbf{C}\boldsymbol{\gamma} + \mathbf{u}_i^T\mathbf{C}\boldsymbol{\gamma},$$

where $\mathbf{u}_i = \mathbf{x}_i - \tilde{\mathbf{x}}_i$ with $\mathrm{E}(\mathbf{u}_i) = \mathbf{0}$. This difference of $\mathbf{u}_i^T\mathbf{C}\boldsymbol{\gamma}$ raises our concern to quantify the potential biases of $\hat{\boldsymbol{\delta}}$ and $\hat{\gamma}(\cdot)$ from the estimation.

We first consider the continuous situation such that $\mathbf{W}_i$ is related to $\mathbf{x}_i$ through model (3.6), which gives

$$\mathbf{W}_i = \mathbf{N}_i\mathbf{x}_i + \boldsymbol{\varepsilon}_i,$$

where $\mathbf{x}_i = \mathbf{T}\boldsymbol{\alpha}_x + \mathbf{B}\mathbf{a}_x + \mathbf{d}_i$ with $\mathbf{T}$, $\mathbf{B}$, $\boldsymbol{\alpha}_x$, $\mathbf{a}_x$, $\mathbf{d}_i$ as defined in Section 3.2.1, and $\boldsymbol{\varepsilon}_i \sim \mathrm{N}(\mathbf{0}, \sigma_\varepsilon^2\mathbf{I})$. Following Zhang et al. (2007), we have

$$\mathbf{x} = (\mathbf{x}_1^T, \ldots, \mathbf{x}_n^T)^T \sim \mathrm{N}(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x),$$

and

$$\mathbf{W} = (\mathbf{W}_1^T, \ldots, \mathbf{W}_n^T)^T \sim \mathrm{N}(\boldsymbol{\mu}_W, \boldsymbol{\Sigma}_W),$$

where $\boldsymbol{\mu}_x = (\boldsymbol{\alpha}_x^T\mathbf{T}^T, \ldots, \boldsymbol{\alpha}_x^T\mathbf{T}^T)^T$, $\boldsymbol{\Sigma}_x = \mathrm{diag}\{\mathbf{B}_*\mathbf{D}\mathbf{B}_*^T, \ldots, \mathbf{B}_*\mathbf{D}\mathbf{B}_*^T\} + \tilde{\mathbf{B}}\tilde{\mathbf{B}}^T$ with $\tilde{\mathbf{B}} = \lambda_x^{-\frac{1}{2}}(\mathbf{B}^T, \ldots, \mathbf{B}^T)^T$, and $\boldsymbol{\mu}_W = \mathcal{N}\boldsymbol{\mu}_x$, $\boldsymbol{\Sigma}_W = \mathcal{N}\boldsymbol{\Sigma}_x\mathcal{N}^T + \sigma_\varepsilon^2\mathbf{I} = \mathbf{Q}\mathcal{D}\mathbf{Q}^T + \mathbf{N}\mathbf{B}\mathbf{B}^T\mathbf{N}^T + \sigma_\varepsilon^2\mathbf{I}$ with $\mathcal{N} = \mathrm{diag}\{\mathbf{N}_1, \ldots, \mathbf{N}_n\}$. Thus, $\mathbf{x}$ and $\mathbf{W}$ follow a joint normal distribution, and $\mathrm{Cov}(\mathbf{x}, \mathbf{W}) = \boldsymbol{\Sigma}_x\mathcal{N}^T$, $\mathrm{Var}(\mathbf{x}|\mathbf{W}) = \boldsymbol{\Sigma}_x - \boldsymbol{\Sigma}_x\mathcal{N}^T\boldsymbol{\Sigma}_W^{-1}\mathcal{N}\boldsymbol{\Sigma}_x$. In particular,

$$\mathbf{u} = (\mathbf{u}_1^T, \ldots, \mathbf{u}_n^T)^T = \mathbf{x} - \tilde{\mathbf{x}} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_x - \boldsymbol{\Sigma}_x\mathcal{N}^T\boldsymbol{\Sigma}_W^{-1}\mathcal{N}\boldsymbol{\Sigma}_x),$$

and $\mathbf{u}_i$'s are correlated since the covariance matrix is not block diagonal.

With the specified distribution of $\mathbf{u}_i$, the mean structure of the response model will be further studied. For model (3.3), a discussion of

$$Y_i = \mathbf{Z}_i^T\boldsymbol{\delta} + \tilde{\mathbf{x}}_i^T\mathbf{C}\boldsymbol{\gamma} + \mathbf{u}_i^T\mathbf{C}\boldsymbol{\gamma} + \epsilon_i \tag{6.1}$$

is provided in Section 4 of Zhang et al. (2007). If model (3.1,3.7) is instead assumed,

one needs to consider the mean structure as

$$\eta_i = \int g^{-1}(\mathbf{Z}_i^T\boldsymbol{\delta} + \tilde{\mathbf{x}}_i^T\mathbf{C}\boldsymbol{\gamma} + \mathbf{u}_i^T\mathbf{C}\boldsymbol{\gamma})\phi(\mathbf{u}_i)d\mathbf{u}_i,$$

where $\phi(\cdot)$ is the density function of normal distribution for $\mathbf{u}_i$. Discussions are deliberated case by case in Section 3 of Wang et al. (1999). In particular, certain link functions may lead to estimation bias for some of the coefficients, which are then subject to corrections.

However, in our model settings, W_{ij}'s are count values and $\mathbf{W}_i$ is related to $\mathbf{x}_i$ through model (3.4,3.5), then the distribution of $\mathbf{u}_i$ has a complex form. Suppose that $\mathbf{u}_i \sim f_{\mathbf{u}}(\cdot)$ can be figured out, we would need to study the mean structure of equation (6.1) for model (3.3), or

$$\eta_i = \int g^{-1}(\mathbf{Z}_i^T\boldsymbol{\delta} + \tilde{\mathbf{x}}_i^T\mathbf{C}\boldsymbol{\gamma} + \mathbf{u}_i^T\mathbf{C}\boldsymbol{\gamma})f_{\mathbf{u}}(\mathbf{u}_i)d\mathbf{u}_i, \tag{6.2}$$

for model (3.1,3.2). It is found that, when model (3.3) is assumed, equation (6.1) correctly specifies the mean structure since $\mathrm{E}(\mathbf{u}_i) = \mathbf{0}$. However, when the responses are of count values, further considerations on equation (6.2) should be taken in the similar fashion to those of Wang et al. (1999). This indicates that, both of the fixed and random effects can be misspecified with certain link functions, thus some further corrections are desired. These findings also suggest the advantages and limitations of our proposed estimation procedures.

6.2 Improvements in Estimation Procedures

The results from the simulation studies in Sections 3.3 and 5.3 suggest that, our inference procedures are able to provide stable and reliable estimators for the model

with longitudinal covariates when either or both of the covariates and responses are in the generalized linear model framework. However, some bias may exist theoretically, which is negligible in most cases but likely to have impact under certain extreme situations. In this section, we discuss several possible improvements for our current estimation procedures.

As discussed in Section 6.1, the fact that $\mathbf{u}_i$'s are correlated suggests that the responses Y_i's become correlated in the calibration model (3.9) or (3.25), even though they are assumed to be independent as in model (3.1,3.2) or (3.3). Thus, a similar justification as that in Section 4 of Zhang et al. (2007) should be studied. Moreover, when working on model (3.3), we need to derive a bias correction procedure to properly estimate σ_ϵ^2 from $\hat{\sigma}_{\epsilon^*}^2$.

The use of the Laplace approximation in quasi-likelihood may bring bias in estimating population profile function, coefficient parameters and effect function as well. Such phenomena are also suggested from the findings of the asymptotic properties in Corollary 4.5. With ideas from Lin and Breslow (1996), further correction procedures on the mean components will be studied to improve the estimation performance.

6.3 Extensions of Model Scope

The model in Zhang et al. (2007) is designed for countinuous longitudinal covariates and continuous responses; the general model settings proposed in Chapter 3 will accommodate discrete longitudinal covariates and either continuous or discrete scalar responses; in addition, the particular model developed in Chapter 5 allows zero inflation of count values for the longitudinal covariates. However, there exist many possibilities to extend the current model scope to accommodate other situations, for example, the responses are count values exhibiting an excess of zeros,

and the predictor consists of multiple longitudinal covariate processes instead of a single one as in current models. We summarize our thoughts for these extensions in this section.

In many circumstances, Y_i's may also suggest an abundance of zeros, thus we would extend our models for zero-inflated count responses. Typically, we assume that longitudinal covariates are generated according to model (3.4,3.5), and zero-inflated responses are from a hurdle model with zero-truncated Poisson distribution, that is,

$$\begin{aligned} \mathrm{P}(Y_i = 0) &= 1 - \pi, \\ \mathrm{P}(Y_i = y) &= \pi \frac{e^{-\eta_i}\eta_i^y}{y!(1 - e^{-\eta_i})}, \text{ for } y > 0, \end{aligned}$$

and

$$\begin{aligned} \mathrm{logit}(\pi) &= u_0, \\ \log(\eta_i) &= \mathbf{Z}_i^T \boldsymbol{\delta} + \int_{T_1}^{T_2} x_i(t)\gamma(t)dt. \end{aligned}$$

Hence, our proposed models, along with the model in Zhang et al. (2007), will be adapted to fit a wider range of data.

The proposed models can also be extended to account for multiple longitudinal covariate processes, each of which can take either discrete or continuous values. To be specific, for each of the C independent processes, we assume

$$W_{ij}^{(c)} \sim f_{W^{(c)}}(\cdot; \mu_{ij}^{(c)}),$$

and

$$h_c(\mu_{ij}^{(c)}) = x_i^{(c)}(t_{ij}),$$

where $f_{W^{(c)}}(\cdot)$ is a distribution function of an exponential family, and $h_c(\cdot)$ is a

known link function; or

$$W_{ij}^{(c)} = x_i^{(c)}(t_{ij}) + \varepsilon_{ij}^{(c)},$$

where $\varepsilon_{ij}^{(c)}$'s are random measurement errors with $\mathrm{N}(0, \sigma^2_{\varepsilon^{(c)}})$. Meanwhile, the response Y_i is related to the latent profiles $x_i^{(c)}(\cdot)$'s and a vector of time-invariant covariates $\mathbf{Z}_i$ through

$$Y_i \sim f_Y(\cdot; \eta_i),$$

and

$$g(\eta_i) = \mathbf{Z}_i^T \boldsymbol{\delta} + \sum_{c=1}^{C} \int_{T_1}^{T_2} x_i^{(c)}(t)\gamma^{(c)}(t)dt,$$

where $f_Y(\cdot)$ is a distribution belonging to an exponential family, and $g(\cdot)$ is a link function; or

$$Y_i = \mathbf{Z}_i^T \boldsymbol{\delta} + \sum_{c=1}^{C} \int_{T_1}^{T_2} x_i^{(c)}(t)\gamma^{(c)}(t)dt + \epsilon_i,$$

where ϵ_i's are assumed to be independent and identically distributed as $\mathrm{N}(0, \sigma^2_\epsilon)$.

Bibliography

American Psychiatric Association (1994). *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition*. American Psychiatric Association, Washington.

Bardone, A. M., Krahn, D. D., Goodman, B. M., and Searles, J. S. (2000). Using Interactive Voice Response Technology and Timeline Follow-Back Methodology in Studying Binge Eating and Drinking Behavior: Different Answers to Different Forms of the Same Question? *Addictive Behaviors*, 25:1–11.

Bartlett, M. S. (1953a). Approximate Confidence Intervals. *Biometrika*, 40:12–19.

Bartlett, M. S. (1953b). Approximate Confidence Intervals II More than One Unknown Parameter. *Biometrika*, 40:306–317.

Beck, A. T., Steer, R. A., and Brown, G. K. (1996). *Manual for the Beck Depression Inventory-II*. Psychological Corporation, San Antonio.

Berhane, K. and Tibshirani, R. (1998). Generalized Additive Models for Longitudinal Data. *The Canadian Journal of Statistics*, 26:517–535.

Bhadra, D., Daniels, M. J., Kim, S., Ghosh, M., and Mukherjee, B. (2012). A Bayesian Semiparametric Approach for Incorporating Longitudinal Information

on Exposure History for Inference in Case-Control Studies. *Biometrics*, 68:361–370.

Breslow, N. E. and Clayton, D. G. (1993). Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association*, 88:9–25.

Buu, A., Li, R., Tan, X., and Zucker, R. A. (2012a). Statistical Models for Longitudinal Zero-Inflated Count Data with Applications to the Substance Abuse Field. *Statistics in Medicine*, 31:4074–4086.

Buu, A., Wang, W., Schroder, S. A., Kalaida, N. L., Puttler, L. I., and Zucker, R. A. (2012b). Developmental Emergence of Alcohol Use Disorder Symptoms and Their Potential as Early Indicators for Progression to Alcohol Dependence in a High Risk Sample: a Longitudinal Study from Childhood to Early Adulthood. *Journal of Abnormal Psychology*, 121:897–908.

Chen, K., Fan, J., and Jin, Z. (2008). Design-Adaptive Minimax Local Linear Regression for Longitudinal Clustered Data. *Statistica Sinica*, 18:515–534.

Chen, K. and Jin, Z. (2005). Local Polynomial Regression Analysis of Clustered Data. *Biometrika*, 92:59–74.

Cranford, J. A., Tennen, H., and Zucker, R. A. (2010). Feasibility of Using Interactive Voice Response to Monitor Daily Drinking, Moods, and Relationship Processes on a Daily Basis in Alcoholic Couples. *Alcoholism: Clinical and Experimental Research*, 34:499–508.

de Boor, C. (1978). *A Practical Guide to Splines*. Springer-Verlag, New York.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood

from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38.

Diggle, P. J., Heagerty, P. J., Liang, K.-Y., and Zeger, S. L. (2002). *Analysis of Longitudinal Data, Second Edition*. Oxford University Press, Oxford.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. Chapman and Hall, London.

Fan, J., Huang, T., and Li, R. (2007). Analysis of Longitudinal Data with Semiparametric Estimation of Covariance Function. *Journal of the American Statistical Association*, 102:632–641.

Fan, J. and Li, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. *Journal of the American Statistical Association*, 96:1348–1360.

Fan, J. and Li, R. (2004). New Estimation and Model Selection Procedures for Semiparametric Modeling in Longitudinal Data Analysis. *Journal of the American Statistical Association*, 99:710–723.

Fan, J. and Peng, H. (2004). Nonconcave Penalized Likelihood with a Diverging Number of Parameters. *The Annals of Statistics*, 32:928–961.

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

Gasser, T. and Muller, H.-G. (1979). Kernel Estimation of Regression Functions. In Gasser, T. and Rosenblatt, M., editors, *Smoothing Techniques for Curve Estimation*, pages 23–68. Springer-Verlag, Heidelberg.

Goluba, G. H., Heathb, M., and Wahba, G. (1979). Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter. *Technometrics*, 21:215–224.

Green, P. J. (1987). Penalized Likelihood for General Semi-Parametric Regression Models. *International Statistical Review*, 55:245–259.

Green, P. J. (1995). Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination. *Biometrika*, 82:711–732.

Green, P. J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models*. Chapman and Hall, London.

Gwaltney, C. J., Magill, M., Barnett, N. P., Apodaca, T. R., Colby, S. M., and Monti, P. M. (2011). Using Daily Drinking Data to Characterize the Effects of a Brief Alcohol Intervention in an Emergency Room. *Addictive Behaviors*, 36:248–250.

Harville, D. A. (1977). Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems. *Journal of the American Statistical Association*, 72:320–338.

Hastie, T. and Tibshirani, R. (1986). Generalized Additive Models. *Statistical Science*, 1:297–318.

Hastie, T. and Tibshirani, R. (1993). Varying-Coefficient Models. *Journal of the Royal Statistical Society, Series B*, 55:757–796.

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

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

James, G. M. (2002). Generalized Linear Models with Functional Predictors. *Journal of the Royal Statistical Society, Series B*, 64:411–432.

Laird, N. M. and Ware, J. H. (1982). Random-Effects Models for Longitudinal Data. *Biometrics*, 38:963–974.

Lambert, D. (1992). Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing. *Technometrics*, 34:1–14.

Lehmann, E. L. and Casella, G. (1998). *Theory of Point Estimation, Second Edition*. Springer-Verlag, New York.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika*, 73:13–22.

Lin, X. and Breslow, N. E. (1996). Bias Correction in Generalized Linear Mixed Models with Multiple Components of Dispersion. *Journal of the American Statistical Association*, 91:1007–1016.

Lin, X. and Carroll, R. J. (2000). Nonparametric Function Estimation for Clustered Data When the Predictor is Measured without/with Error. *Journal of the American Statistical Association*, 95:520–534.

Lin, X. and Zhang, D. (1999). Inference in Generalized Additive Mixed Models by Using Smoothing Splines. *Journal of the Royal Statistical Society, Series B*, 61:381–400.

Linton, O. B., Mammen, E., Lin, X., and Carroll, R. J. (2003). Accounting for Correlation in Marginal Longitudinal Nonparametric Regression. In Lin, D.-Y. and Heagerty, P. J., editors, *Proceedings of the Second Seattle Symposium in Biostatistics: Analysis of Correlated Data*, pages 23–33. Springer, New York.

Marron, J. S. and Nolan, D. (1988). Canonical Kernels for Density Estimation. *Statistics and Probability Letters*, 7:195–199.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models, Second Edition.* Chapman and Hall, London.

Mullahy, J. (1986). Specification and Testing of Some Modified Count Data Models. *Journal of Econometrics*, 33:341–365.

Muller, H.-G. and Stadtmuller, U. (2005). Generalized Functional Linear Models. *The Annals of Statistics*, 33:774–805.

Mundt, J. C., Perrine, M. W., Searles, J. S., and Walter, D. (1995). An Application of Interactive Voice Response (IVR) Technology to Longitudinal Studies of Daily Behavior. *Behavior Research Methods, Instruments, and Computers*, 27:351–357.

Nadaraya, E. A. (1964). On Estimating Regression. *Theory of Probability and Its Applications*, 9:141–142.

Qu, A. and Li, R. (2006). Quadratic Inference Functions for Varying-Coefficient Models with Longitudinal Data. *Biometrics*, 62:379–391.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis, Second Editon.* Springer, New York.

Rose, C. E., Martin, S. W., Wannemuehler, K. A., and Plikaytis, B. D. (2006). On the Use of Zero-Inflated and Hurdle Models for Modeling Vaccine Adverse Event Count Data. *Journal of Biopharmaceutical Statistics*, 16:463–481.

Ruppert, D. (1997). Empirical-Bias Bandwidths for Local Polynomial Nonparametric Regression and Density Estimation. *Journal of the American Statistical Association*, 92:1049–1062.

Ruppert, D., Sheather, S. J., and Wand, M. P. (1995). An Effective Bandwidth Selector for Local Least Squares Regression. *Journal of the American Statistical Association*, 90:1257–1270.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

Stone, M. (1974). Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society, Series B*, 36:111–147.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge.

Wahba, G. (1980). Spline Bases, Regularization, and Generalized Cross-Validation for Solving Approximation Problems with Large Quantities of Noisy Data. In Cheney, E. W., editor, *Approximation Theory III*, pages 905–912. Academic Press, New York.

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

Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing*. Chapman and Hall, London.

Wang, N. (2003). Marginal Nonparametric Kernel Regression Accounting for within-Subject Correlation. *Biometrika*, 90:43–52.

Wang, N., Lin, X., and Guttierrez, R. G. (1999). A Bias Correction Regression Calibration Approach in Generalized Linear Mixed Measurement Error Models. *Communications in Statistics, Theory and Methods*, 28:217–232.

Watson, G. S. (1964). Smooth Regression Analysis. *Sankhya, Series A*, 26:359–372.

Wedderburn, R. W. M. (1974). Quasi-Likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method. *Biometrika*, 61:439–447.

Yao, W. and Li, R. (2013). New Local Estimation Procedure for Nonparametric Regression Function of Longitudinal Data. *Journal of the Royal Statistical Society, Series B*, 75:123–138.

Zhang, D., Lin, X., Raz, J., and Sowers, M. F. (1998). Semiparametric Stochastic Mixed Models for Longitudinal Data. *Journal of the American Statistical Association*, 93:710–719.

Zhang, D., Lin, X., and Sowers, M. F. (2007). Two-Stage Functional Mixed Models for Evaluating the Effect of Longitudinal Covariate Profiles on a Scalar Outcome. *Biometrics*, 63:351–362.

Zucker, R. A., Ellis, D. A., Fitzgerald, H. E., Bingham, C. R., and Sanford, K. (1996). Other Evidence for at Least Two Alcoholisms II: Life Course Variation in Antisociality and Heterogeneity of Alcoholic Outcome. *Development and Psychopathology*, 8:831–848.

Vita

Hanyu Yang

EDUCATION

Department of Statistics, The Pennsylvania State University (2010 – 2014)
Ph.D. in Statistics

Department of Mathematics, Zhejiang University (2006 – 2010)
B.S. in Statistics

RESEARCH INTEREST

Longitudinal Analysis, Functional Data, Nonparametric Models, Generalized Linear Mixed Models, MCMC Methods

PUBLICATION

Buu, A., Li, R., Walton, M. A., Yang, H., Zimmerman, M. A., and Cunningham, R. M. (2014), "Changes in Substance Use-Related Health Risk Behaviors on the Timeline Follow-Back Interview as a Function of Length of Recall Period", *Substance Use and Misuse*, 49:1259-1269.

Yang, H., Cranford, J., Li, R., and Buu, A., "Two-Stage Model for Time-Varying Effects of Discrete Longitudinal Covariates with Applications in Analysis of Daily Process Data", *Statistics in Medicine*, accepted.

Yang, H., Li, R., Zucker, R. A., and Buu, A., "Generalized Time-Varying Effect Model with Zero-Inflated Count Longitudinal Covariates", submitted.

ACADEMIC EXPERIENCE

Department of Statistics (2010 – 2014)
Instructor, Teaching Assistant

Department of Statistics (2013)
Research Assistant

Statistical Consulting Center (2012 – 2013)
Graduate Consultant

CONFERENCE PARTICIPATION

Joint Statistical Meetings (2014)
Paper Presentation: "Modeling Distal Scalar Response with Zero-Inflated Longitudinal Covariates"

PROFESSIONAL EXPERIENCE

Corporate Model Risk, Wells Fargo (2014)
Risk Analytics Consultant (summer intern)