FUNCTIONAL VARYING INDEX COEFFICIENT MODEL FOR DYNAMIC GENE-ENVIRONMENT INTERACTIONS WITH LONGITUDINAL DATA

By

Jingyi Zhang

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Statistics—Doctor of Philosophy

2017

ABSTRACT

FUNCTIONAL VARYING INDEX COEFFICIENT MODEL FOR DYNAMIC GENE-ENVIRONMENT INTERACTIONS WITH LONGITUDINAL DATA

By

Jingyi Zhang

Rooted in genetics, human complex diseases are largely influenced by environmental factors. Existing literature has shown the power of integrative gene-environment interaction analysis by considering the joint effect of environmental mixtures on disease risk. Motivated by that, we propose a functional varying index coefficient model for longitudinal measurements of a phenotypic trait and multiple environmental variables, and assess how the genetic effects on a longitudinal disease trait are nonlinearly modified by a mixture of environmental influences. We derive an estimation procedure for the nonparametric functional varying index coefficients under the quadratic inference functions and penalized splines framework. Theoretical results such as estimation consistency and asymptotic normality of the estimates are established. In addition, we propose a hypothesis testing procedure to assess the significance of the nonparametric index coefficient function. We evaluate the performance of our estimation and testing procedure through Monte Carlo simulation studies. The proposed method is illustrated by applying to a real data set from a pain sensitivity study in which SNP effects are nonlinearly modulated by the combination of dosage levels and other environmental variables to affect blood pressure and heart rate of patients.

In order to deal with discrete measurements for risk of disease, we further extend our proposed FVICM to a generalized varying index coefficient model (gFVICM) to binary longitudinal traits. We apply penalized splines to approximate the nonparametric varying index coefficients and develop an estimation procedure based on the quadratic inference functions. The asymptotic normality established in the theoretical results enables us to develop a model selection criteria and construct a test statistic based on the quadratic inference function. In hypothesis test, we investigate the linearity of G$\times$E interactions using the proposed testing procedure. The utility of the method is

further demonstrated through a pain sensitivity case study in which SNP effects are nonlinearly modulated by the combination of environmental mixtures to affect high blood pressure.

Genetic pleiotropy refers to the situation in which a single gene influences multiple traits and so it is considered as a major factor that underlies genetic correlation among traits. For some complex diseases, there are multiple phenotypes that can used to diagnose or to quantify the risk of diseases and usually they have shared genetic determinations. In multivariate longitudinal data, multiple response variables are jointly measured over time from the same individual. It is appropriate to take into account the correlation between multivariate longitudinal responses. Therefore, we propose the joint partially linear varying coefficient models and the testing framework to jointly test the association of genetic factors with bivariate phenotypic values adjusting for environmental factors. We extended the quadratic inference functions to deal with the longitudinal correlations and used penalized splines for the approximation of nonparametric coefficients. The proposed method is illustrated by applying to a real data set from a pain sensitivity study, in which systolic blood pressure (SBP) and diastolic blood pressure (DBP) were correlated longitudinal quantified phenotypes of SNP effects.

ACKNOWLEDGEMENTS

I would like to express the deepest appreciation to my advisor and committee chair Professor Yuehua Cui for his excellent guidance and continuous support in my Ph.D. research. Without his guidance and persistent help this dissertation would not have been possible. I also would like to thank my committee members, Professor Ping-Shou Zhong, Professor Lyudmila Sakhanenko and Professor Qing Lu, for their insightful comments and suggestions in my research.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Gene-environment (G×E) interaction is defined as how genotypes influence phenotypes differently under different environmental conditions (Falconer, 1952). An increasing number of studies have confirmed the role of G×E interaction in many human diseases. One classic example is Phenylketonuria (PKU). PKU is caused by a defect in the gene coding for a particular enzyme which is needed to break down phenylalanine. Newborns found to have high levels of phenylalanine in their blood can be put on a special, phenylalanine-free diet to avoid the severe effects of PKU (Baker, 2004). This example confirms the role of gene-environment interaction by showing that a change in environment can affect the phenotype of a particular trait.

In genetic epidemiology, G×E interactions are useful for understanding the risk of some complex human diseases. Famous studies such as Parkinson disease (Ross and Smith (2007)) and type 2 diabetes (Zimmet, Alberti, and Shaw (2001)) both indicate the importance of G×E interaction in complex human diseases. However, the underlying mechanism of G×E interaction is still poorly understood due to the lack of powerful statistical methods. The traditional way to investigate G×E interaction is based on a single environment exposure model. Parametric models such as additive linear models, use the products of two variables to denote the interaction effects. However, this product may not capture the true interaction effect of gene and environment, since it could not be able to capture the possible nonlinear G×E interactions.

In order to assess possible nonlinear G×E interactions, some nonparametric and semiparametric models have been developed, such as varying coefficient models (VCM) proposed by Hastie and Tibishirani (1993). In a VCM model, the coefficients of covariates are allowed to change with some other variables through smooth functions, so nonlinear interactions can be assessed. A VCM has the form

$$Y = \sum_{l=1}^{L} m_l(X) Z_l + \varepsilon, \tag{1.1}$$

where Y is the response variable, $(X, \mathbf{Z}^T)^T$ is a vector of predictors consisting of a scalar X and a

L-dimensional vector $\mathbf{Z} = (Z_1, Z_2, ..., Z_L)^T$. $m_l(\cdot)$, $l = 1, ..., L$, are unknown nonparametric smooth functions. In particular, dealing with G×E interaction problems, one can replace the predicts $\mathbf{Z}$ to be genetic variants, for example, single nucleotide polymorphisms (SNPs).

Epidemiological evidences suggested that a disease risk can be modified by simultaneous exposure to multiple environmental agents with effect larger than simple addition of individual factor acting along. The specification of VCM in (1.1) may be limited in dealing with simultaneous exposure to multiple environmental factors. It will result in difficulties in the estimation of the coefficient function $m_l(\mathbf{X})$ when variable $\mathbf{X} = (X_1, ..., X_p)^T$ in Model (1.1) is multidimensional. To overcome such challenge, Ma and Song (2015) proposed the varying index coefficient model (VICM) with a form

$$Y = \sum_{l=1}^{L} m_l(\beta_l^T \mathbf{X}) Z_l + \varepsilon, \tag{1.2}$$

where $\boldsymbol{\beta}_l = (\beta_{l1}, ..., \beta_{lp})^T$ are the coefficients for covariate vector $\mathbf{X}$ with β_{lk} be the loading weight for the k-th covariate of $\mathbf{X}$, i.e. X_k associated with Z_l. The single index $\boldsymbol{\beta}^T \mathbf{X}$ is actually a linear combination of several environmental effects. The VICM model is able to pursue the interaction between a set of environmental factors as a whole and genetic variables on the disease risk, if the covariate Z is specified to be some genetic effect, such as a SNP variable. Liu et al. (2016) extended the model to a partial linear varying multi-index coefficient model (PLVMICM):

$$Y = m_0(\beta_0^T \mathbf{X}) + \boldsymbol{\alpha}_0^T \mathbf{Z} + \sum_{l=1}^{L} \{m_l(\beta_l^T \mathbf{X}) G_l + \boldsymbol{\alpha}_l^T \mathbf{Z} G_l\} + \varepsilon, \tag{1.3}$$

where G_l, l = 1, . . . , L are genetic variables of interest, $m_l(\cdot)$, $l = 0, 1, ..., L$ are unknown index functions, $\boldsymbol{\alpha}_0, ..., \boldsymbol{\alpha}_L$ and $\boldsymbol{\beta}_0, ..., \boldsymbol{\beta}_L$ are parametric parameters. The main genetic effect for each G_l is captured by the function $m_l(\beta_l^T \mathbf{X})$ when the function is approximated by some nonparametric techniques such as B-spline approximation. Model (1.3) is an extension of Model (1.2) by considering partial linear covariates $\mathbf{Z}$ with application in capturing nonlinear G×E interactions. However, the above mentioned medoels can not be used directly for longitudinal data because of the assumption of independence among observations. Little work has been done to deal with

nonlinear G×E interaction in longitudinal studies. This motivates us to extend the varying index coefficient model to longitudinal traits.

Longitudinal studies play an important role in epidemiology, biological and clinical research. Longitudinal studies are used to characterize normal growth and aging, to assess the effect of risk factors on human health, and to evaluate the effectiveness of treatments. Some researches, such as Sitlani et al. (2015), Furlotte et al. (2015) and Xu et al. (2014) demonstrate the longitudinal design is more powerful in detecting genetic associations than cross-sectional designs.

The traditional way to analyze longitudinal data is using the regression models (Belle et al., 2004). However, standard regression methods assume that all observations are independent. The assumption of independence would result in invalid inferences. One approach to deal with the issue is using a complete model which specify the correlation structure among observations for each subject. In linear mixed effect models, we can make specific assumptions for the covariance structure in observations. Based on the parametric assumptions for the covariance components, we can apply the maximum likelihood methods to estimate the regression parameters. Another regression approach for inference with longitudinal data is known as generalized estimating equation (GEE) approach proposed by Liang and Zeger in 1986. The GEE method is very popular in recent decades and has been widely used in longitudinal data analysis. However, the application of the GEE method has several disadvantages. One of those disadvantages is that GEE may fail to produce consistent estimators if the nuisance correlation parameters are not consistently estimated (Crowder 1986, 1995). In addition, model selection and hypothesis testing can be complicated since there is no objective function in the estimation procedure of the GEE approach.

The quadratic inference function (QIF) approach proposed by Qu et al. (2000) is one of the improvements of the GEE method. The benefits of using QIF approach in longitudinal analysis have been discussed in some researches, such as Qu et al. (2000), Qu and Li (2006) and Song et al. (2009). The QIF method avoids estimating the nuisance correlation parameters by using a linear combination of several basis matrices to approximate the inverse of correlation matrix. When the working correlation structure is correctly specified, both the QIF and GEE are equally efficient.

However, when the working correlation structure is misspecified, the QIF is more efficient than the GEE. In addition, the QIF has an asymptotic χ^2 distribution, based on which we can implement the model selection criteria like BIC, and construct testing statistic. These advantages of using QIF in longitudinal analysis motivates us to extend the QIF method to the varying coefficient model in detecting G×E interactions on risk of diseases.

We organise the rest of this dissertation in the following way: In Chapter 2, in order to capture the dynamic nonlinear G×E interaction with the combined effect of environmental factors for longitudinal data, we propose a functional varying index coefficient model (FVICM) for correlated responses, i.e.,

$$Y_{ij} = m_0(\boldsymbol{\beta}_0^T \mathbf{X}_{ij}) + m_1(\boldsymbol{\beta}_1^T \mathbf{X}_{ij}) G_i + \varepsilon_{ij}, \tag{1.4}$$

where Y_{ij} is the response variable which measures the risk of certain disease on the ith subject at the jth time point, where $i = 1, \cdots, N$, $j = 1, \cdots, n_i$; $\mathbf{X}_{ij}$ is a p-dimensional vector of environmental variables, which can be either time-dependent or time invariant; G_i denotes the genetic variable; ε_{ij} is an error term with mean 0 and some correlation structure; $m_0(\cdot)$ and $m_1(\cdot)$ are unknown functions; $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$ are p-dimensional vectors of index loading coefficients. Compared with Model (1.2) and (1.3), Model (1.4) assumes correlations among observations from the same subject, which can be applied to capture the longitudinal correlation among different time points for a subject. We use penalized splines to approximate the nonparametric functions in the model and then develop an estimation procedure based on QIF method. In addition, we are interested to see whether the interaction function $m_1(\cdot)$ is significantly nonlinear or not. This is a natural concern in our model since if a linear interaction function is sufficient, a varying coefficient model would not be necessary. We develop the testing procedure by representing our model in a linear mixed model form and then apply the pseudo-likelihood ratio test. On the other hand, discrete longitudinal traits, for example, the binary traits are very common in clinical researches. To deal with binary responses, in Chapter 3, we extend Model (1.4) to a generalized functional varying-index

coefficient model (gFVICM) with a form

$$g\{E(Y_{ij}|\mathbf{X}_{ij},G_{ij})\} = m_0(\boldsymbol{\beta}_0^T\mathbf{X}_{ij}) + m_1(\boldsymbol{\beta}_1^T\mathbf{X}_{ij})G_{ij}, \tag{1.5}$$

where $g(\cdot)$ is the logit link function, $m_0(\cdot)$ and $m_1(\cdot)$ are unknown nonparametric functions, $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$ are p-dimensional vectors of index coefficients. The QIF method is modified to accommodate binary response. To test the linearity of interaction function $m_1(\cdot)$, we develop a hypothesis testing procedure which is built on the asymptotical property of the QIF objective function. In Chapter 4, we consider G×E interactions when multiple longitudinal traits are measured to improve the power of association test, especially to identify any pleiotropic effect (i.e., one gene can affect multiple traits). For this purpose, we propose a multivariate partially linear varying coefficients model and derive a testing framework to jointly test the association of genetic factors with multivariate phenotypic values adjusting for environmental factors. The joint models are written as

$$Y_{lij} = Y_{li}(t_{ij}) = \beta_{0l}\{X(t_{ij})\} + \beta_{1l}\{X(t_{ij})\}G_i + \boldsymbol{\alpha}_l\mathbf{Z}_{ij} + \varepsilon_{lij}, \tag{1.6}$$

where Y_{lij} is the response variable which measures the l-th phenotype on the i-th subject at the j-th time point, $X(t_{ij})$ is a time-varying covariate, $\mathbf{Z}_{ij}$ is a p-dimensional vector of covariates, which can either depend or be independent of time, G_i denotes the genetic variable within subject, ε_{lij} is an error term and

$$\boldsymbol{\varepsilon}_i = \begin{pmatrix} \boldsymbol{\varepsilon}_{1i} \\ \vdots \\ \boldsymbol{\varepsilon}_{Li} \end{pmatrix} \sim N(\mathbf{0},\boldsymbol{\Sigma})$$

with $\boldsymbol{\Sigma}$ be some covariance structure, $\beta_{0l}(\cdot)$ and $\beta_{1l}(\cdot)$ are unknown functions. The robustness of QIF method in the variance of the estimators helps us to build the estimation and hypothesis testing procedure. In Chapter 5, we conclude the thesis with a brief conclusion about the contributions of this thesis and point out some future research directions. Proofs are provided in the Appendix.

CHAPTER 2

FUNCTIONAL VARYING INDEX COEFFICIENT MODEL FOR DYNAMIC GENE-ENVIRONMENT INTERACTIONS

2.1 Introduction

It has been broadly recognized that gene-environment (G×E) interaction plays important role in human complex diseases. A growing number of scientific researches have confirmed the role of G×E interaction in many human diseases, such as Parkinson disease (Ross and Smith, 2007) and type 2 diabetes (Zimmet et al., 2001). G×E interaction is defined as how genotypes influence phenotypes differently under different environmental conditions (Falconer, 1952). It also refers to the genetic sensitivity to environmental changes. Usually, G×E has been investigated based on a single environment exposure model. Evidence from epidemiological studies has suggested that disease risk can be modified by simultaneous exposure to multiple environmental factors. The effect of simultaneous exposure is higher than the simple addition of the effects of factors acting alone (Carpenter et al., 2002; Sexton and Hattis, 2007). This motivated us to assess the combined effect of environmental mixtures, and how they, as a whole, interact with genes to affect disease risk (Liu et al. 2016). In our previous model, we proposed a varying-index coefficient model to capture the nonlinear interaction between a gene and environmental mixtures (Liu et al. 2016). The method was extended for any univariate trait distribution in a generalized linear model framework (Liu et al. 2017).

In biomedical studies, longitudinal traits are often observed, with repeated measures of the same subject over time. The increased power of a longitudinal design to detect genetic associations over cross-sectional designs has been evaluated (Sitlani et al. 2015; Furlotte et al. 2015; Xu et al. 2014). With longitudinal disease traits, one can study the dynamic gene effect over time. Coupling with longitudinal measure of environmental exposures, one can study how genes respond to the dynamic change of environmental factors to affect a disease trait. This motivates us to extend the

varying index coefficient model to longitudinal traits.

To explore time-dependent effects in longitudinal data analysis, some nonparametric and semiparametric models such as varying coefficient models have been proposed, for example, Hoover et al. (1998), Wu, Chiang, and Hoover (1998), Fan and Zhang (2000), Martinussen and Scheike (2001), Chiang, Rice, and Wu (2001), Huang, Wu, and Zhou (2002), Ma and Song (2015). However, these methods do not fit to our purpose. In order to capture the dynamic nonlinear G×E interaction with combined effect of environmental factors for longitudinal data, we propose a functional varying index coefficient model (FVICM) for correlated response, i.e.,

$$Y_{ij} = m_0(\boldsymbol{\beta}_0^T \mathbf{X}_{ij}) + m_1(\boldsymbol{\beta}_1^T \mathbf{X}_{ij}) G_i + \varepsilon_{ij}, \tag{2.1}$$

where Y_{ij} is the response variable which measures the risk of certain disease on the ith subject at the jth time point, where $i = 1, \cdots, N$, $j = 1, \cdots, n_i$; $\mathbf{X}_{ij}$ is a p-dimensional vector of environmental variables, which can be either time-dependent or time invariant; G_i denotes the genetic variable; ε_{ij} is an error term with mean 0 and some correlation structure; $m_0(\cdot)$ and $m_1(\cdot)$ are unknown functions; $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$ are p-dimensional vectors of index loading coefficients. For model identifiability, we have the constraints $\|\boldsymbol{\beta}_0\| = \|\boldsymbol{\beta}_1\| = 1$ and the first elements of $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$ are positive.

Qu et al. (2000) proposed the quadratic inference function (QIF) for longitudinal data analysis, as an improvement of the generalized estimation equation (GEE) approach introduced by Liang and Zeger (1986). The QIF approach avoids estimating the nuisance correlation parameters by assuming that the inverse of the correlation matrix can be approximated by a linear combination of several basis matrices. Qu et al. (2000) found that the QIF estimator could be generally more efficient than the GEE estimator. Qu and Li (2006) applied the QIF method to the varying coefficient model for longitudinal data. Bai et al. (2009) developed an estimating procedure for single index models with longitudinal data also based on QIF method. Motivated by that, in this paper, we extend the QIF method to the FVICM model for dynamic G×E interactions.

Our goal in this work is to develop a set of statistical estimation and hypothesis testing procedure for model (2.1). We first approximate the varying index coefficient function by the penalized

splines (Ruppert and Carroll, 2000) and then extend the QIF approach to our model in order to estimate the index loading coefficients and the penalized spline coefficients. Under certain regularity conditions, we establish the consistency and asymptotic normality of the resulting estimators.

Another goal of this work is to test the linearity of the G×E interaction effect. This is of particular interest in our model setting since if the G×E interaction is linear, a simple linear regression model should be fit, and fitting any higher order nonlinear functions would be unnecessary. With a mixed effects model representation of the penalized spline approximations (Speed, 1991; Ruppert, Wand, and Carroll, 2003; Wand, 2003), we can transform the problem of testing an unknown function into testing some fixed effects and a variance component in a linear mixed effects model setup with multiple variance components, which will be evaluated in this study.

This chapter is organized as follows: in Section 2.2, we propose an estimation procedure under the FVICM model, and further establish the consistency and asymptotic normality of the proposed estimator in Section 2.3. In Section 2.4, we discuss some practical issues to implement the proposed estimation procedures. In Section 2.5, a pseudo-likelihood ratio test procedure with a linear mixed effects model representation is illustrated. We assess the finite sample performance of the proposed procedure with Monte Carlo simulation in Section 2.6 and illustrate the proposed method by an analysis of a pain sensitivity data set in Section 2.7, followed by discussions in Section 2.8. Technical details are rendered in Appendix.

2.2 Quadratic inference function for FVICM with longitudinal data

For longitudinal data, suppose the response y_{ij}, p-dimensional covariate vector $\boldsymbol{x}_{ij}$, and SNP variable G_i are observed from the ith observation at the jth time point. SNP variable $\{G_i, i = 1,...,N\}$ does not change over time. Assume the model satisfies

$$E(y_{ij}|\boldsymbol{x}_{ij}, G_i) = m_0(\boldsymbol{\beta}_0^T \boldsymbol{x}_{ij}) + m_1(\boldsymbol{\beta}_1^T \boldsymbol{x}_{ij}) G_i,$$

We can approximate the unknown coefficient functions $m_0(u_0)$ and $m_1(u_1)$ by a q-degree truncated power spline basis, i.e.

$$m_0(u_0) = m_0(u_0, \boldsymbol{\beta}) \approx \mathbf{B}(u_0)^T \boldsymbol{\gamma}_0,$$

$$m_1(u_1) = m_1(u_1, \boldsymbol{\beta}) \approx \mathbf{B}(u_1)^T \boldsymbol{\gamma}_1,$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}_0^T, \boldsymbol{\beta}_1^T)^T$, $\mathbf{B}(u) = (1, u, u^2, \cdots, u^q, (u-\kappa_1)_+^q, \cdots, (u-\kappa_K)_+^q)^T$ is a q-degree truncated power spline basis with K knots $\kappa_1, \cdots, \kappa_K$. $\boldsymbol{\gamma}_0$ and $\boldsymbol{\gamma}_1$ are $(q+K+1)$-dimensional vectors of spline coefficients. Let $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_0^T, \boldsymbol{\gamma}_1^T)^T$.

For longitudinal data, the conditional variance-covariance matrix of the response need to be modelled. The method of generalized estimation equation (GEE) is often applied to estimate the unknowns. The GEE is defined as,

$$\sum_{i=1}^{N} \dot{\boldsymbol{\mu}}_i^T \mathbf{V}_i^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i) = 0,$$

where $\mathbf{V}_i$ is the covariance matrix of $\mathbf{y}_i$, $\mathbf{y}_i = (y_{i1}, ..., y_{in_i})^T$, $\boldsymbol{\mu}_i = E(\mathbf{y}_i)$ is the mean function and $\dot{\boldsymbol{\mu}}_i$ is the first derivative of $\boldsymbol{\mu}_i$ with respect to the parameters. Based on the spline approximation, the mean function can be written as

$$\boldsymbol{\mu}_i = \boldsymbol{\mu}_i(\boldsymbol{\theta}) = \begin{bmatrix} \boldsymbol{\mu}_{i1}(\boldsymbol{\theta}) \\ \vdots \\ \boldsymbol{\mu}_{in_i}(\boldsymbol{\theta}) \end{bmatrix} = \begin{bmatrix} \mathbf{B}^T(\boldsymbol{\beta}_0^T \mathbf{x}_{i1})\boldsymbol{\gamma}_0 + \mathbf{B}^T(\boldsymbol{\beta}_1^T \mathbf{x}_{i1})\boldsymbol{\gamma}_1 G_i \\ \vdots \\ \mathbf{B}^T(\boldsymbol{\beta}_0^T \mathbf{x}_{in_i})\boldsymbol{\gamma}_0 + \mathbf{B}^T(\boldsymbol{\beta}_1^T \mathbf{x}_{in_i})\boldsymbol{\gamma}_1 G_i \end{bmatrix},$$

and the first derivative of $\boldsymbol{\mu}_i$ is

$$\dot{\boldsymbol{\mu}}_i = \begin{bmatrix} \mathbf{B}_d^T(\boldsymbol{\beta}_0^T \mathbf{x}_{i1})\boldsymbol{\gamma}_0 \mathbf{x}_{i1}^T & \mathbf{B}_d^T(\boldsymbol{\beta}_1^T \mathbf{x}_{i1})\boldsymbol{\gamma}_1 G_i \mathbf{x}_{i1}^T & \mathbf{B}^T(\boldsymbol{\beta}_0^T \mathbf{x}_{i1}) & \mathbf{B}^T(\boldsymbol{\beta}_1^T \mathbf{x}_{i1}) G_i \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{B}_d^T(\boldsymbol{\beta}_0^T \mathbf{x}_{in_i})\boldsymbol{\gamma}_0 \mathbf{x}_{in_i}^T & \mathbf{B}_d^T(\boldsymbol{\beta}_1^T \mathbf{x}_{in_i})\boldsymbol{\gamma}_1 G_i \mathbf{x}_{in_i}^T & \mathbf{B}^T(\boldsymbol{\beta}_0^T \mathbf{x}_{in_i}) & \mathbf{B}^T(\boldsymbol{\beta}_1^T \mathbf{x}_{in_i}) G_i \end{bmatrix},$$

where $\mathbf{B}_d(u) = \frac{\partial \mathbf{B}(u)}{\partial u} = (0, 1, 2u, \cdots, qu^{q-1}, q(u-\kappa_1)_+^{q-1}, \cdots, q(u-\kappa_K)_+^{q-1})$, $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T)^T$.

When $\mathbf{V}_i$ is unknown, Liang and Zeger (1986) suggested that $\mathbf{V}_i$ can be simplified as $\mathbf{V}_i = \mathbf{A}_i^{1/2} \mathbf{R}(\rho) \mathbf{A}_i^{1/2}$ with $\mathbf{A}_i$ being a diagonal matrix of marginal variances and $\mathbf{R}(\rho)$ being a common

working correlation matrix with a small number of nuisance parameters ρ. When ρ is consistently estimated, the GEE estimators of the regression coefficients are consistent. When such consistent estimators for the nuisance parameters do not exist, Qu et al. (2000) suggested that the inverse of $\mathbf{R}(\rho)$ can be represented by a linear combination of a class of basis matrices such as $\mathbf{R}^{-1}(\rho) \approx a_1\mathbf{M}_1 + a_2\mathbf{M}_2 \cdots + a_h\mathbf{M}_h$, where $\mathbf{M}_1$ is the identity matrix and $\mathbf{M}_2, \cdots, \mathbf{M}_h$ are symmetric matrices. The advantage of this method is that the estimation of nuisance parameters $a_1, \cdots, a_h$ are not required. Following this idea, we define the estimation function as,

$$\bar{g}_N(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N} g_i(\boldsymbol{\theta}) = \frac{1}{N}\begin{bmatrix} \sum_{i=1}^{N} \dot{\boldsymbol{\mu}}_i^T \mathbf{A}_i^{-1/2}\mathbf{M}_1\mathbf{A}_i^{-1/2}(\mathbf{y}_i - \boldsymbol{\mu}_i) \\ \vdots \\ \sum_{i=1}^{N} \dot{\boldsymbol{\mu}}_i^T \mathbf{A}_i^{-1/2}\mathbf{M}_h\mathbf{A}_i^{-1/2}(\mathbf{y}_i - \boldsymbol{\mu}_i) \end{bmatrix} \tag{2.2}$$

Because the dimension of the estimation equation $\bar{g}_N$ is greater than the number of parameters, we cannot obtain the estimators by simply setting each element in $\bar{g}_N$ to be zero. Qu et al. (2000) introduced the Quadratic Inference Function (QIF) based on the generalized method of moments (Hansen, 1982). Thus, we can estimate the parameters by minimizing the QIF, which is defined as

$$Q_N(\boldsymbol{\theta}) = N\bar{g}_N^T\bar{C}_N^{-1}\bar{g}_N, \tag{2.3}$$

where $\bar{C}_N = \frac{1}{N}\sum_{i=1}^{N} g_i g_i^T$ is a consistent estimator for $\text{var}(g_i)$. By minimizing the quadratic inference function, we can obtain the estimation of the parameters

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} Q_N(\boldsymbol{\theta}).$$

To overcome the well known over-parameterization issue, Qu et al. (2000) further proposed the penalized quadratic inference function

$$N^{-1}Q_N(\boldsymbol{\theta}) + \lambda\boldsymbol{\theta}^T\mathbf{D}\boldsymbol{\theta}, \tag{2.4}$$

where $\mathbf{D}$ is a diagonal matrix with element 1 if the corresponding parameters are spline coefficients associated with the knots and 0 otherwise, i.e., $\mathbf{D} = \text{diag}(\mathbf{0}^T_{(2p+q+1)\times 1}, \mathbf{1}^T_{K\times 1}, \mathbf{0}^T_{(q+1)\times 1}, \mathbf{1}^T_{K\times 1})$. Then the estimator is given by

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}}(N^{-1}Q_N(\boldsymbol{\theta}) + \lambda\boldsymbol{\theta}^T\mathbf{D}\boldsymbol{\theta}). \tag{2.5}$$

2.3 Asymptotic properties

In this section, we establish the asymptotic properties for the penalized quadratic inference function estimators with fixed knots. Assume $\boldsymbol{\theta}_0$ is the parameter satisfying $E_{\boldsymbol{\theta}_0}(g_i) = 0$. Theorem 1 provides the consistency of the resulting estimators. We show the asymptotic normality of the estimators in Theorem 2. The theoretical results are similar to those provided in Qu and Li (2006). The difference is that we have constraints for the index loading parameters in our model, i.e. $\|\boldsymbol{\beta}_0\|=\|\boldsymbol{\beta}_1\|=1$, and $\beta_{01} > 0$, $\beta_{11} > 0$. To handle the constraints, we do the reparameterization as $\beta_{l1} = \sqrt{1 - \|\boldsymbol{\beta}_{l,-1}\|_2^2}$ with $\boldsymbol{\beta}_{l,-1} = (\beta_{l2}, ..., \beta_{lp})^T$ for l=1, 2 (Yu and Ruppert, 2002; Cui et al., 2011; Ma and Song, 2015). Then the parameters space of $\boldsymbol{\beta}_l$, l=1,2, becomes

$$[\{(\sqrt{1 - \|\boldsymbol{\beta}_{l,-1}\|_2^2}, \beta_{l2}, ..., \beta_{lp})^T\} : \|\boldsymbol{\beta}_{l,-1}\|_2^2 < 1].$$

Let

$$\mathbf{J}_l = \frac{\partial \boldsymbol{\beta}_l}{\partial \boldsymbol{\beta}_{l,-1}^T} = \begin{pmatrix} -\boldsymbol{\beta}_{l,-1}^T / \sqrt{1 - \|\boldsymbol{\beta}_{l,-1}\|_2^2} \\ \mathbf{I}_{p-1} \end{pmatrix}$$

be the Jacobian matrix of dimension $p \times (p-1)$. Denote $\boldsymbol{\beta}_{-1} = (\boldsymbol{\beta}_{0,-1}^T, \boldsymbol{\beta}_{1,-1}^T)^T$, and $\boldsymbol{\theta}^* = (\boldsymbol{\beta}_{-1}, \boldsymbol{\gamma})^T$. From $\boldsymbol{\theta}$ to $\boldsymbol{\theta}^*$, we have Jacobian matrix $\mathbf{J} = \text{diag}(\mathbf{J}_0, \mathbf{J}_1, \mathbf{I}_{q+K+1}, \mathbf{I}_{q+K+1})$.

Theorem 1 *Suppose the assumptions (A1)-(A6) in the Appendix are satisfied, and the smoothing parameter $\lambda_N = o(1)$, then the estimator $\widehat{\boldsymbol{\theta}}$, which is obtained by minimizing the penalized quadratic function in (2.4), exists and converges to $\boldsymbol{\theta}_0$ in probability.*

Theorem 2 *Suppose the assumptions (A1)-(A6) in the Appendix are satisfied, and the smoothing parameter $\lambda_N = o(N^{-1/2})$, then the estimator $\widehat{\boldsymbol{\theta}}$ obtained by minimizing the penalized quadratic function in (2.4) is asymptotically normally distributed, i.e.,*

$$\sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N(\boldsymbol{0}, \boldsymbol{J}(\boldsymbol{G}_0^T \boldsymbol{C}_0^{-1} \boldsymbol{G}_0)^{-1} \boldsymbol{J}^T),$$

where $\boldsymbol{G}_0$ and $\boldsymbol{C}_0$ are given in the Appendix.

2.4 Practical issues

In this section, we discuss some practical issues when we implement the proposed method.

2.4.1 Algorithm for estimation

A two-step iterative Newton-Raphson algorithm is applied when we estimate the index loading parameters and the varying spline coefficients. The algorithm of the estimation procedure can be summarized in the following steps.

Step 0 Choose initial values for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. Denote them by $\boldsymbol{\beta}^{(old)}$ and $\boldsymbol{\gamma}^{(old)}$.

Step 1 Estimate $\boldsymbol{\gamma}^{(new)}$ by

$$\boldsymbol{\gamma}^{(new)} = \arg\min_{\boldsymbol{\gamma}}(N^{-1}Q_N(\boldsymbol{\gamma}, \boldsymbol{\beta}^{(old)}) + \lambda \boldsymbol{\gamma}^T \mathbf{D}\boldsymbol{\gamma}.$$

The Newton-Raphson algorithm is used for the minimization.

Step 2 Estimate $\boldsymbol{\beta}^{(new)}$ by

$$\boldsymbol{\beta}^{(new)} = \arg\min_{\boldsymbol{\beta}} Q_N(\boldsymbol{\beta}, \boldsymbol{\gamma}^{(new)}).$$

Also use Newton-Raphson for minimization.

Step 3 Update $\boldsymbol{\beta}_l^{(old)}$ by $\boldsymbol{\beta}_l^{(old)} = \text{sign}(\beta_{l1}^{(new)})\boldsymbol{\beta}_l^{(new)}/\|\boldsymbol{\beta}_l^{(new)}\|_2$, $l = 1,2$. Update $\boldsymbol{\gamma}^{(old)}$ by setting $\boldsymbol{\gamma}^{(old)} = \boldsymbol{\gamma}^{(new)}$.

Step 4 Repeat Steps 1-3 until convergence.

2.4.2 Model selection

It is important to determine the order and number of knots in the spline approximation since too many knots in the model might overfit the data. Under the assumption $E(g) = 0$ (g is the estimation function in (2.2) for a single observation) and the number of estimating equations is larger than the number of parameters, we have $Q(\widehat{\boldsymbol{\theta}}) \to \chi^2_{r-k}$ in distribution (Hansen, 1982), where r is the

dimension of $\bar{g}_N(\boldsymbol{\theta})$, k is the dimension of $\boldsymbol{\theta}$, $\widehat{\boldsymbol{\theta}}$ is the estimator by minimizing the QIF when certain order and number of knots are chosen. This asymptotic property of the QIF provides a goodness-of-fit test, which can be useful to determine the order and number of knots to be selected in our model.

However, it is also possible that the goodness-of-fit tests fail to reject several different models which may not be nested. Since $Q(\widehat{\boldsymbol{\theta}})$ is asymptotically chi-square distributed, we can use BIC to penalize $Q(\widehat{\boldsymbol{\theta}})$ for the difference of the numbers of estimating equations and parameters. In particular, the BIC criterion for a model with r estimating equation and k parameters is defined as

$$Q(\widehat{\boldsymbol{\theta}}) + (r-k)\ln N,$$

The model with minimum BIC would be considered better. If we choose h basis matrices in (2.2), then $r-k = hk-k = (h-1)k$. As we discussed in Section 2.4.3, we usually use h=2 in our setting. Thus, the BIC criterion is actually

$$Q(\widehat{\boldsymbol{\theta}}) + k\ln N,$$

where k is the number of parameters in the model.

In our simulation and real data application, we search the optimal order and the number of knots over a set of combinations of q and K using BIC. Knots are evenly distributed in the range of $u(=\boldsymbol{\beta}^T\mathbf{X})$.

2.4.3 Choice of the basis for the inverse of the correlation matrix

Qu and Li (2006) offered several choices of basis matrixes. For exchangeable working correlation, $\mathbf{M}_1$ is identity matrix and $\mathbf{M}_2$ has 0 on the diagonal and 1 off-diagonal. If the working correlation is AR(1), we can set $\mathbf{M}_2$ to have 1 on its two subdiagonals and 0 elsewhere. Prior information on correlation can help us to determine the choice of appropriate basis matrices. The effect of choosing different basis matrices is discussed in Qu and Li (2006) through simulation studies. Qu and Lindsay (2003) also proposed an adaptive estimation method to approximate the correlation empirically when there is no prior information available.

2.4.4 Choice of the tuning parameter

Since the penalized spline is used to approximate the unknown functions, we need to determine the tuning parameter λ involved in the method. As Qu and Li (2006) suggested, we can extend the generalized cross-validation (Ruppert, 2002) to the penalized QIF and define the generalized cross-validation statistic as

$$\mathrm{GCV}(\lambda) = \frac{N^{-1}Q_N}{(1-N^{-1}\mathrm{df})^2}$$

where $\mathrm{df} = \mathrm{tr}[(\ddot{Q}_N + 2N\lambda D)^{-1}\ddot{Q}_N]$ is the effective degree of freedom, Q_N is defined in (2.3) and $\ddot{Q}_N$ is the second derivative of Q_N. The desirable choice of tuning parameter λ is

$$\widehat{\lambda} = \arg\min_{\lambda} \mathrm{GCV}(\lambda).$$

In the implementation of GCV, the golden search method can be applied in order to reduce the computational time.

2.5 Hypothesis test

2.5.1 Linear mixed model representation for FVICM model

In our proposed FVICM model (2.1), it is of interest to test the unspecified coefficient function. In particular, we are interested in testing whether a linear function is good enough to describe the G×E interaction. Given $\boldsymbol{\beta}$, let $u_0 = \boldsymbol{\beta}_1^T\mathbf{X}$, $u_1 = \boldsymbol{\beta}_0^T\mathbf{X}$, with the truncated power spline basis, the coefficient function can be modeled by

$$m_1(u_1) = \gamma_{10} + \gamma_{11}u_1 + \gamma_{12}u_1^2 + \cdots + \gamma_{1q}u_1^q + \sum_{k=1}^{K} b_{1k}(u_1 - \kappa_k)_+^q.$$

Our goal is to test the linearity of $m_1(u_1)$, which is equivalent to test

$$H_0 : \gamma_{12} = \cdots = \gamma_{1q} = 0, b_{11} = \cdots = b_{1K} = 0.$$

Let $\mathbf{w}_{0ij} = (1, u_{0ij}, \cdots, u_{0ij}^q)^T$, $\mathbf{z}_{0ij} = \left((u_{0ij}-\kappa_1)_+^q, \cdots, (u_{0ij}-\kappa_K)_+^q\right)^T$, $\tilde{\boldsymbol{\gamma}}_0 = (\gamma_{00}, \cdots, \gamma_{0q})^T$, $\mathbf{b}_0 = (b_{01}, \cdots, b_{0K})^T$, $\mathbf{w}_{1ij} = (1, u_{1ij}, \cdots, u_{1ij}^q)^T$, $\mathbf{z}_{1ij} = \left((u_{1ij}-\kappa_{11})_+^q, \cdots, (u_{1ij}-\kappa_{1K})_+^q\right)^T$,$\mathbf{b}_1 = (b_{11}, \cdots, b_{1K})^T$, $\tilde{\boldsymbol{\gamma}}_1 = (\gamma_{10}, \cdots, \gamma_{1q})^T$,

$$m_0(u_{0ij}) = \mathbf{w}_{0ij}^T \tilde{\boldsymbol{\gamma}}_0 + \mathbf{z}_{0ij}^T \mathbf{b}_0,$$

$$m_1(u_{1ij}) = \mathbf{w}_{1ij}^T \tilde{\boldsymbol{\gamma}}_1 + \mathbf{z}_{1ij}^T \mathbf{b}_1.$$

We further define $\mathbf{Y}_i = (y_{i1}, \cdots, y_{in_i})^T$, $\mathbf{W}_{0i} = (\mathbf{w}_{0i1}, \cdots, \mathbf{w}_{0in_i})^T$, $\mathbf{W}_{1i} = (\mathbf{w}_{1i1}G_i, \cdots, \mathbf{w}_{1in_i}G_i)^T$, $\mathbf{Z}_{0i} = (\mathbf{z}_{0i1}, \cdots, \mathbf{z}_{0in_i})^T$, $\mathbf{Z}_{1i} = (\mathbf{z}_{1i1}G_i, \cdots, \mathbf{z}_{1in_i}G_i)^T$, then a linear mixed model (LMM) representation (Wang and Chen, 2012) can be obtained as,

$$\mathbf{Y}_i = \mathbf{W}_{0i}\tilde{\boldsymbol{\gamma}}_0 + \mathbf{W}_{1i}\tilde{\boldsymbol{\gamma}}_1 + \mathbf{Z}_{0i}\mathbf{b}_0 + \mathbf{Z}_{1i}\mathbf{b}_1 + \mathbf{1}_i a_i + \boldsymbol{\varepsilon}_i, \;\; i = 1, \cdots, n, \tag{2.6}$$

$$\mathbf{b}_l \sim N(\mathbf{0}, \sigma^2_{\mathbf{b}_l}\mathbf{I}_K), \; l = 0, 1, \;\; \boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \sigma^2_{\varepsilon}\mathbf{I}),$$

where the random incept effects a_i are assumed to be independent as $N(0, \sigma_a^2)$ which model the correlation in the response.

With the LMM representation, testing the linearity of the varying index coefficients is equivalent to test some fixed effects and a variance component in model (2.6). To be specific, we want to test

$$H_0 : \gamma_{12} = \cdots = \gamma_{1q} = 0 \text{ and } \sigma^2_{\mathbf{b}_1} = 0. \tag{2.7}$$

2.5.2 LRT and pseudo-LRT in LMM

2.5.2.1 LRT for one variance component

Crainiceanu and Ruppert (2004) proposed the likelihood ratio test in linear mixed effect models with one variance component. Consider a LMM with one variance component

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\varepsilon}, \;\; E\begin{bmatrix} \mathbf{b} \\ \boldsymbol{\varepsilon} \end{bmatrix} = \begin{bmatrix} \mathbf{0}_K \\ \mathbf{0}_n \end{bmatrix}, \;\; \text{Cov}\begin{bmatrix} \mathbf{b} \\ \boldsymbol{\varepsilon} \end{bmatrix} = \begin{bmatrix} \sigma_b^2\boldsymbol{\Sigma} & \mathbf{0} \\ \mathbf{0} & \sigma_{\varepsilon}^2\mathbf{I}_n \end{bmatrix}, \tag{2.8}$$

where $\boldsymbol{\beta}$ is a p-dimensional vector of fixed effect coefficients, $\mathbf{b}$ is a K-dimensional vector of random effects, $\mathbf{0}_K$ is a K-dimensional vector of zeros, $\boldsymbol{\Sigma}$ is a known $K \times K$ symmetric positive definite matrix. Let $\lambda = \sigma_b^2/\sigma_{\varepsilon}^2$ be the signal-to-noise ratio and then the covariance matrix of

$\mathbf{Y}$ cab be written as $\text{Cov}(\mathbf{Y}) = \sigma_\varepsilon^2 \mathbf{V}_\lambda$, where $\mathbf{V}_\lambda = \mathbf{I}_n + \lambda \mathbf{Z}\mathbf{\Sigma}\mathbf{Z}^T$. Consider testing for the null hypothesis

$$H_0 : \beta_{p+1-p'} = 0, \cdots, \beta_p = 0,\ \sigma_b^2 = 0 \tag{2.9}$$

for $p' > 0$.

The LRT statistic is defined as

$$LRT_n \propto 2\Big\{ \sup_{H_A} L(\boldsymbol{\beta}, \lambda, \sigma_\varepsilon^2) - \sup_{H_0} L(\boldsymbol{\beta}, \lambda, \sigma_\varepsilon^2) \Big\}.$$

If we substitute the parameters $\boldsymbol{\beta}$ and σ_ε^2 with their profile estimators

$$\widehat{\boldsymbol{\beta}}(\lambda) = (\mathbf{X}^T \mathbf{V}_\lambda^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}_\lambda^{-1} \mathbf{Y},$$

$$\widehat{\sigma}_\varepsilon^2(\lambda) = \frac{\{\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}(\lambda)\}^T V_\lambda^{-1} \{\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}(\lambda)\}}{n},$$

for fixed λ, we obtain the LRT statistic

$$LRT_n = \sup_{\lambda \geq 0} \{ n \log(\mathbf{Y}^T \mathbf{S}_0 \mathbf{Y}) - n \log(\mathbf{Y}^T \mathbf{P}_\lambda^T \mathbf{V}_\lambda^{-1} \mathbf{P}_\lambda \mathbf{Y}) - \log |\mathbf{V}_\lambda| \}, \tag{2.10}$$

where $\mathbf{P}_\lambda = \mathbf{I}_n - \mathbf{X}(\mathbf{X}^T \mathbf{V}_\lambda^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}_\lambda^{-1}$, $\mathbf{X}_0$ denotes the design matrix of fixed effects under the null hypothesis, $\mathbf{S}_0 = \mathbf{I}_n - \mathbf{X}_0 (\mathbf{X}_0^T \mathbf{X}_0)^{-1} \mathbf{X}_0^T$.

Theorem 1 in Crainiceanu and Ruppert (2004) provides the distribution of LRT statistic (2.10). Let μ_s be the eigenvalues of $\mathbf{\Sigma}^{1/2} \mathbf{Z}^T \mathbf{P}_0 \mathbf{Z} \mathbf{\Sigma}^{1/2}$, ξ_s be the eigenvalues of $\mathbf{\Sigma}^{1/2} \mathbf{Z}^T \mathbf{Z} \mathbf{\Sigma}^{1/2}$, $s = 1, \cdots, K$, then

$$LRT_n \stackrel{d}{=} n \left(1 + \frac{\sum_1^{p'} u_s^2}{\sum_1^{n-p} w_s^2} \right) + \sup_{\lambda \geq 0} f_n(\lambda), \tag{2.11}$$

where $u_s \stackrel{iid}{\sim} N(0,1)$ for $s = 1, \cdots, K$, $w_s \stackrel{iid}{\sim} N(0,1)$ for $s = 1, \cdots, n-p$, and

$$f_n(\lambda) = n \log \left\{ 1 + \frac{N_n(\lambda)}{D_n(\lambda)} \right\} - \sum_{s=1}^{K} \log(1 + \lambda \mu_s),$$

with

$$N_n(\lambda) = \sum_{s=1}^{K} \frac{\lambda \mu_s}{1 + \lambda \mu_s} w_s^2,$$

$$D_n(\lambda) = \sum_{s=1}^{K} \frac{w_s^2}{1+\lambda\mu_s} + \sum_{s=K+1}^{n-p} w_s^2.$$

The distribution in (2.11) only depends on the eigenvalues μ_s and ξ_s. Based on the spectral decomposition, simulation from this distribution can be done very rapidly. Detailed algorithm for this simulation can be found in Crainiceanu and Ruppert (2004).

2.5.2.2 Pseudo-LRT for multiple variance components

For a LMM with multiple variance components

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}_1 + \cdots + \mathbf{Z}\mathbf{b}_L + \boldsymbol{\varepsilon}, \tag{2.12}$$

$$\mathbf{b}_l \sim N(\mathbf{0}, \sigma^2_{\mathbf{b}_l}\boldsymbol{\Sigma}_l),\ l = 1, \cdots, L,\ \ \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2_{\varepsilon}\mathbf{I}_n),$$

where $\mathbf{b}_l$, $l = 1, \cdots, L$ are random effects and $L > 1$. Suppose we are interested in testing

$$H_0 : \beta_{p+1-p'} = 0, \cdots, \beta_p = 0,\ \sigma^2_{\mathbf{b}_L} = 0.$$

Greven et al. (2008) proposed to approximate the distribution of LRT for the model (2.12) based on the pseudo-likelihood ratio test theory (Liang and Self, 1996) by using a pseudo-outcome. In the framework of model (2.12), $\mathbf{b}_i, i \neq L$, are nuisance random parameters. We can define the pseudo-outcome as

$$\widetilde{\mathbf{Y}} = \mathbf{Y} - \sum_{i \neq L} \mathbf{Z}_i \widehat{\mathbf{b}}_i,$$

where $\widehat{\mathbf{b}}_i$ are the best linear unbiased predictors (BLUP) of nuisance random effects $\mathbf{b}_i, i \neq L$. The the model (2.12) can be reduced to

$$\widetilde{\mathbf{Y}} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_L\mathbf{b}_L + \boldsymbol{\varepsilon}. \tag{2.13}$$

Then the method for testing one variance component introduced by Crainiceanu and Ruppert (2004) can be applied to the model in (2.13).

2.5.3 Pseudo-LRT in FVICM model

For the model in (2.6), we can use the idea of Greven et al. (2008) and define the pseudo-outcome

$$\widetilde{\mathbf{Y}}_i = \mathbf{Y}_i - \mathbf{Z}_{0i}\widehat{\mathbf{b}}_0 - \mathbf{U}_i\widehat{a}_i, \;\; i = 1, \cdots, n,$$

where $\widehat{\mathbf{b}}_0$ and $\widehat{a}_i$ are BLUPs of $\mathbf{b}_0$ and a_i, respectively. The reduced model using pseudo-outcome for model (2.6) can be written as

$$\widetilde{\mathbf{Y}}_i = \mathbf{W}_{0i}\tilde{\boldsymbol{\gamma}}_0 + \mathbf{W}_{1i}\tilde{\boldsymbol{\gamma}}_1 + \mathbf{Z}_{1i}\mathbf{b}_1 + \boldsymbol{\varepsilon}_i. \;\; i = 1, \cdots, n. \tag{2.14}$$

For the new model (2.14) using pseudo-response, we can apply the method for the single variance component model introduced in Section 2.10 to test hypothesis (2.7). Statistical significance can be assessed through the resampling approach described in section 2.5.2.1.

2.6 Simulation study

2.6.1 Simulation

In this section, the finite sample performance of the proposed method is evaluated through Monte Carlo simulation studies. We generate three covariates X_1, X_2, X_3. For each subject i, $X_{1ij}, X_{2ij}, X_{3ij}$ are generated independently from uniform distribution $U(0,1)$. We set the minor allele frequency (MAF) as p_A=(0.1, 0.3, 0.5) and assume Hardy-Weinberg equilibrium. We use AA, Aa and aa to denote three different SNP genotypes, where allele A is the minor allele. These genotypes are simulated from a multinomial distribution with frequencies p_A^2, $2p_A(1-p_A)$ and $(1-p_A)^2$, respectively. Variable G takes value in the set {0,1,2}, corresponding to genotypes $\{aa, Aa, AA\}$ respectively. The error term $\varepsilon_i = (\varepsilon_{i1}, \cdots, \varepsilon_{in_i})$ are independently generated from the multivariate normal distribution $N(\mathbf{0}, 0.1\mathbf{R}(\rho))$. The true correlation structure $\mathbf{R}(\rho)$ is assumed to be exchangeable with ρ=0.5 and 0.8.

We set $m_0(u_0) = \cos(\pi u_0)$ and $m_1(u_1) = \sin[\pi(u_1 - A)/(B - A)]$ with $A = \sqrt{3}/2 - 1.645/\sqrt{12}$ and $B = \sqrt{3}/2 + 1.645/\sqrt{12}$. The true parameters are $\boldsymbol{\beta}_0 = (\sqrt{5}, \sqrt{4}, \sqrt{4})/\sqrt{13}$ and $\boldsymbol{\beta}_1 = (1,1,1)/\sqrt{3}$.

To simplify the simulation and save computational time, we consider the balanced case, which means each observation has the same number of time points. We draw 1000 data sets with sample size $N = 200, 500$ and time points $n_i = T = 10$. Since the true correlation structure is exchangeable, we set $\mathbf{M}_1$ to be the identity matrix and $\mathbf{M}_2$ to be 0 on the diagonal and 1 off-diagonal. The order and number of knots of the splines are chosen by using the BIC method.

2.6.2 Performance of estimation

Table 2.1 summarizes the results based on 1000 replications. In this table, the average bias (Bias), the standard deviation of the 1000 estimates (SD), the average of the estimated standard error (SE) based on the theoretical results, and the estimated coverage probability (CP) at 95% confidence level are reported. Note that the estimation of the loading parameter $\boldsymbol{\beta}_1$ improves as MAF p_A increases, while the estimation of $\boldsymbol{\beta}_0$ show a opposite direction. This is because we have limited data information to estimate the marginal effects $m_0(\cdot)$ when p_A increases. As the sample size increases, the performance of the estimation improves by showing smaller bias, SD and SE.

Table 2.1 Simulation results for $p_A = 0.1, 0.3, 0.5$ with sample size $N = 200, 500$ and correlation ρ=0.5.

			$p_A = 0.1$				$p_A = 0.3$				$p_A = 0.5$			
N	Param	True	Bias	SD	SE	CP	Bias	SD	SE	CP	Bias	SD	SE	CP
200	β_{01}	0.620	7.3E-04	0.008	0.008	95.6	1.7E-03	0.009	0.010	96.2	1.5E-03	0.011	0.011	95.0
	β_{02}	0.555	-3.9E-04	0.008	0.009	93.2	-1.0E-03	0.010	0.010	92.5	-1.2E-03	0.012	0.011	92.4
	β_{03}	0.555	-6.2E-04	0.008	0.008	94.4	-1.2E-03	0.010	0.010	94.2	-8.5E-04	0.012	0.011	93.0
	β_{11}	0.577	-2.3E-05	0.018	0.020	91.0	-3.1E-04	0.011	0.011	93.7	-8.6E-04	0.009	0.009	94.7
	β_{12}	0.577	-6.3E-04	0.018	0.020	91.3	-3.0E-04	0.011	0.011	94.3	-6.8E-05	0.009	0.009	93.8
	β_{13}	0.577	-3.9E-04	0.018	0.020	91.0	2.8E-04	0.011	0.011	94.8	7.1E-04	0.009	0.009	93.1
500	β_{01}	0.620	7.5E-04	0.005	0.005	95.5	1.7E-03	0.006	0.006	95.1	1.6E-03	0.007	0.007	95.8
	β_{02}	0.555	-5.7E-04	0.005	0.005	94.4	-1.1E-03	0.006	0.006	94.6	-8.8E-04	0.007	0.007	95.2
	β_{03}	0.555	-3.4E-04	0.005	0.005	93.9	-8.7E-04	0.006	0.006	94.1	-1.1E-03	0.007	0.007	94.7
	β_{11}	0.577	6.4E-04	0.012	0.012	93.8	-1.7E-04	0.007	0.007	95.6	-7.3E-04	0.006	0.006	95.1
	β_{12}	0.577	-6.0E-04	0.012	0.012	93.6	-1.5E-05	0.007	0.007	96.1	5.3E-04	0.006	0.006	94.7
	β_{13}	0.577	-4.1E-04	0.012	0.012	94.6	6.0E-05	0.007	0.007	95.0	1.1E-04	0.006	0.006	95.6

The plots for the estimations of $m_0(u_0)$ and $m_1(u_1)$ under different sample size and MAFs are shown in Figure 2.1 and Figure 2.2. The estimated and true functions are denoted by the solid

and dashed lines, respectively. The 95% confidence band is denoted by the dotted-dash line. The estimated curves almost overlap with the corresponding true curves as shown in the plots. The confidence bands are tight, especially under a large sample size. Note that the estimation for the interaction effects $m_1(u_1)$ improves as MAF p_A increases, while the estimation for the marginal effects $m_0(u_0)$ show a opposite direction, which coincides with the results for the parametric estimation in Table 2.1.

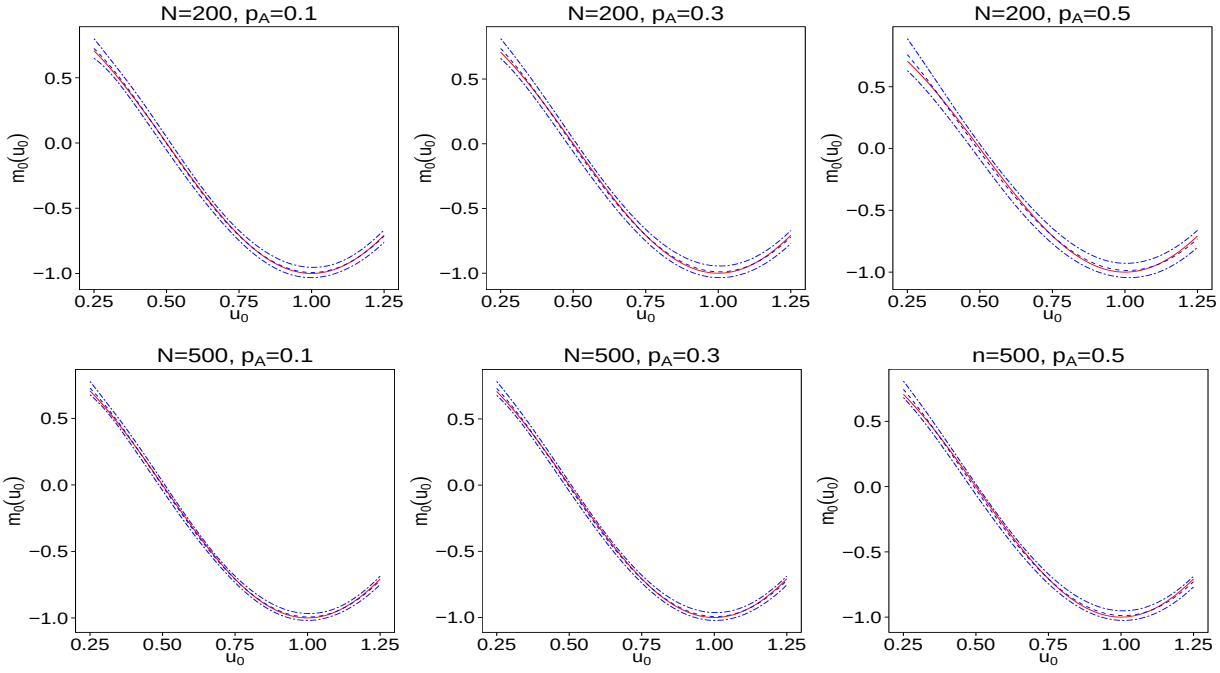

Figure 2.1 The estimation of function $m_0(\cdot)$ under different MAFs when N=200, 500 and ρ=0.5. The estimated and true functions are denoted by the solid and dashed lines respectively. The 95% confidence band is denoted by the dotted-dash line.

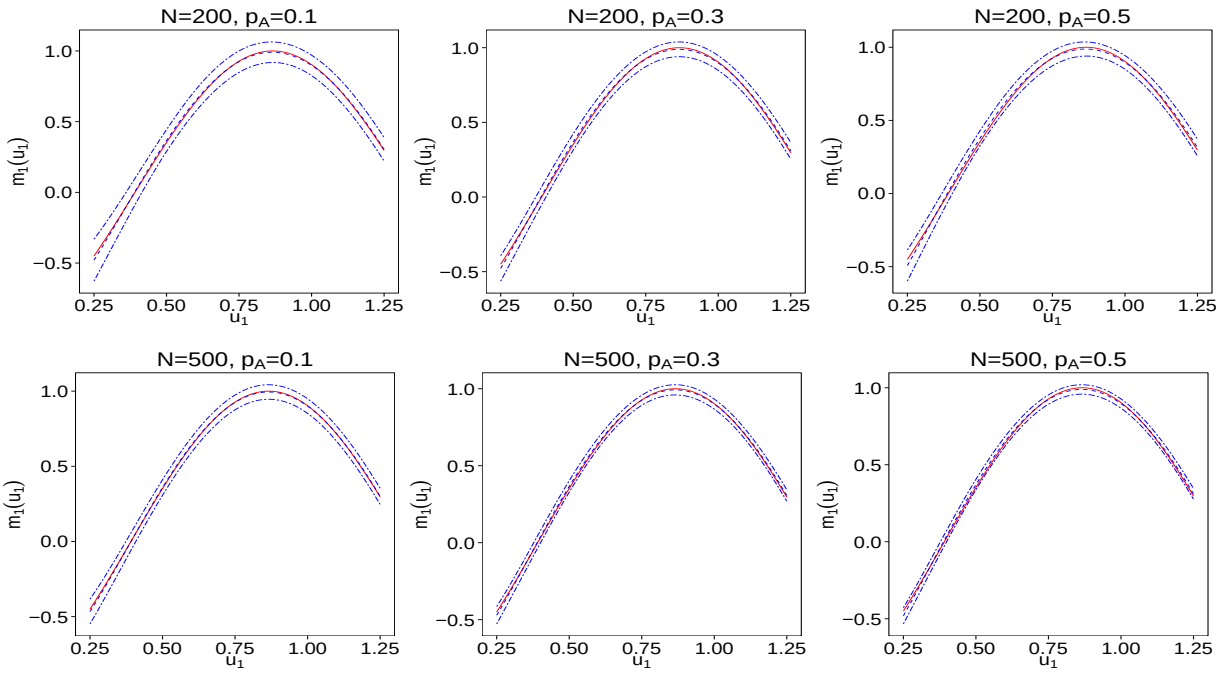

Figure 2.2 The estimation of function $m_1(\cdot)$ under different MAFs when N=200, 500 and ρ=0.5. The estimated and true functions are denoted by the solid and dashed lines respectively. The 95% confidence band is denoted by the dotted-dash line.

The performance of the estimation for $\rho = 0.8$ is shown in Table 2.2, Figure 2.3 and Figure 2.4. It is seen that the SD and SE are smaller when ρ is larger compared to the results when $\rho = 0.5$. The confidence bands are a little bit wider, especially for m_0 when p_A=0.5 and for m_1 when p_A=0.1 for larger ρ. In summary, the simulation results show that the estimation method performs reasonably well under different simulation settings in finite samples.

Table 2.2 Simulation results for $p_A = 0.1, 0.3, 0.5$ with sample size $N = 200, 500$ and correlation ρ=0.8

			$p_A = 0.1$				$p_A = 0.3$				$p_A = 0.5$			
N	Param	True	Bias	SD	SE	CP	Bias	SD	SE	CP	Bias	SD	SE	CP
200	β_{01}	0.620	4.4E-04	0.005	0.005	95.8	5.5E-04	0.006	0.006	95.8	-5.0E-06	0.007	0.007	95.3
	β_{02}	0.555	-2.2E-04	0.006	0.005	92.3	-3.2E-04	0.007	0.006	91.8	-2.8E-04	0.008	0.007	92.7
	β_{03}	0.555	-3.6E-04	0.006	0.005	94.1	-4.0E-04	0.006	0.006	94.6	1.4E-04	0.007	0.007	93.7
	β_{11}	0.577	-6.7E-05	0.014	0.012	90.3	-2.4E-04	0.007	0.007	94.3	-7.7E-04	0.006	0.006	92.7
	β_{12}	0.577	-2.4E-04	0.014	0.012	91.8	-1.3E-04	0.007	0.007	94.2	3.3E-05	0.006	0.006	93.4
	β_{13}	0.577	-1.8E-04	0.014	0.012	89.9	2.4E-04	0.007	0.007	93.7	6.4E-04	0.006	0.006	93.5
500	β_{01}	0.620	5.3E-04	0.004	0.003	94.0	5.8E-04	0.004	0.004	95.4	3.3E-04	0.004	0.005	95.6
	β_{02}	0.555	-4.0E-04	0.003	0.003	93.8	-4.2E-04	0.004	0.004	94.8	-1.3E-04	0.005	0.004	95.0
	β_{03}	0.555	-2.3E-04	0.004	0.003	93.1	-2.8E-04	0.004	0.004	94.2	-2.9E-04	0.004	0.004	94.7
	β_{11}	0.577	2.5E-04	0.008	0.007	94.0	-1.5E-04	0.004	0.004	95.3	-6.8E-04	0.004	0.004	93.9
	β_{12}	0.577	-2.9E-04	0.008	0.007	93.5	4.4E-05	0.004	0.004	95.7	4.5E-04	0.004	0.004	93.5
	β_{13}	0.577	-1.1E-04	0.007	0.007	95.4	6.1E-05	0.004	0.004	95.4	1.9E-04	0.004	0.004	95.2

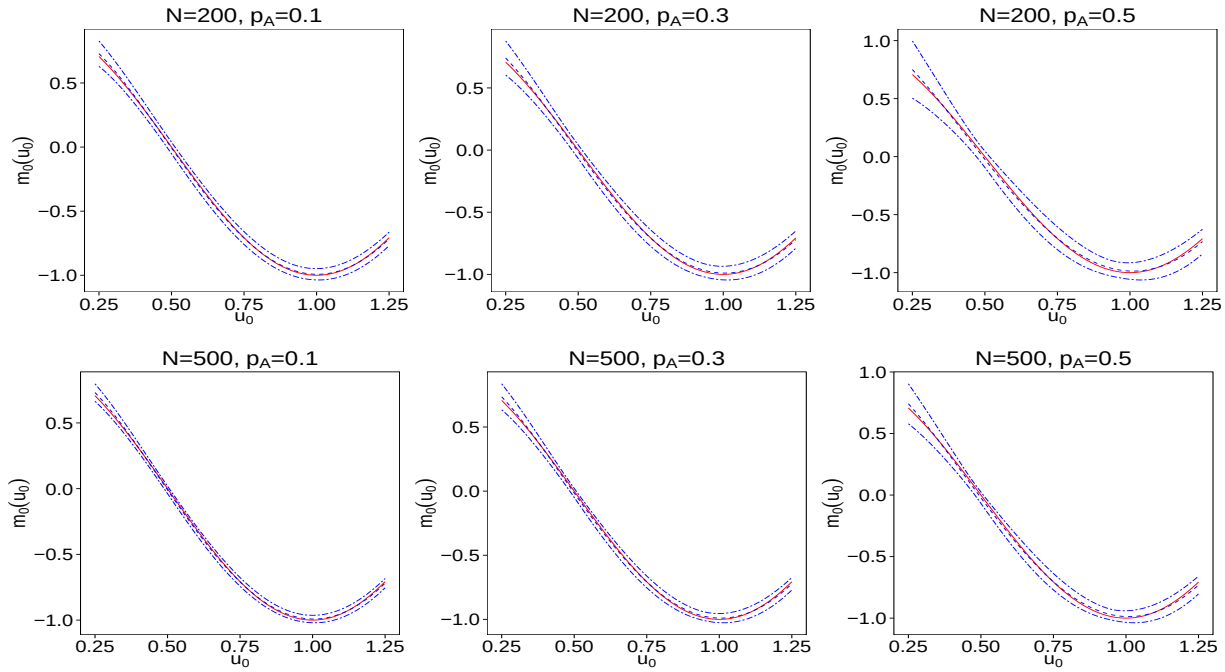

Figure 2.3 The estimation of function $m_0(\cdot)$ under different MAFs when N=200, 500 and ρ=0.8. The estimated and true functions are denoted by the solid and dashed lines respectively. The 95% confidence band is denoted by the dotted-dash line.

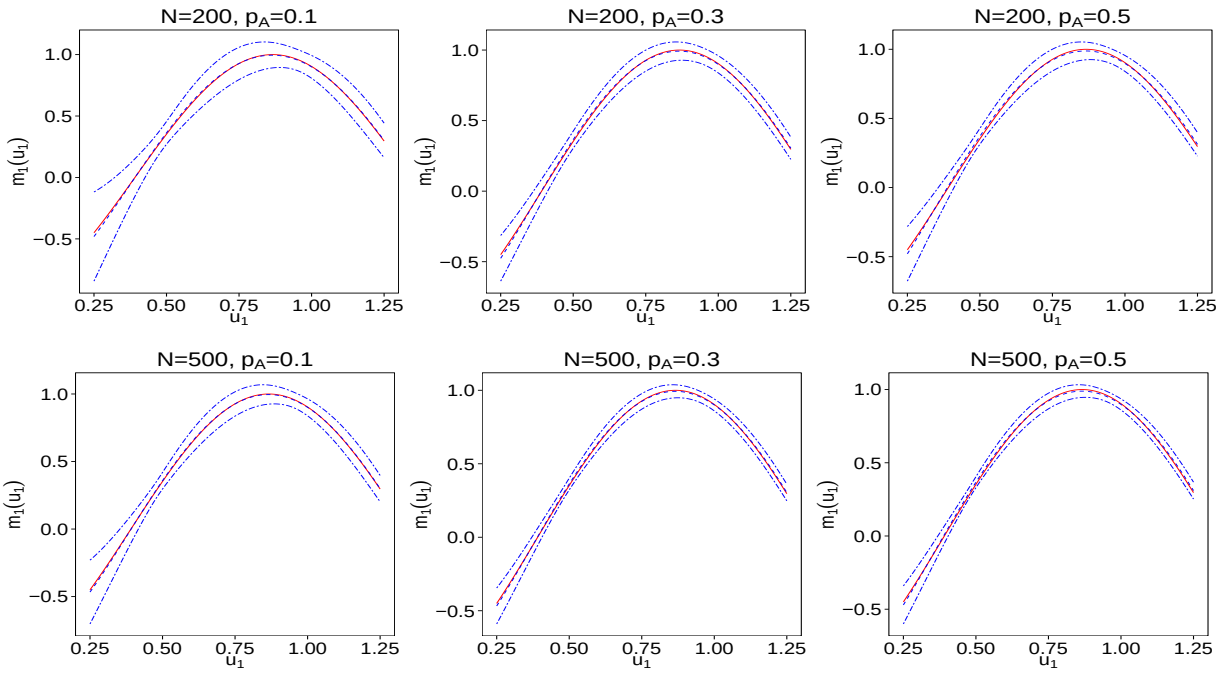

Figure 2.4 The estimation of function $m_1(\cdot)$ under different MAFs when N=200, 500 and ρ=0.8 The estimated and true functions are denoted by the solid and dashed lines respectively. The 95% confidence band is denoted by the dotted-dash line.

2.6.3 Performance of hypothesis tests

We evaluate the performance of the test for the nonparametric function under the null hypothesis $H_0 : m_1(\cdot) = m_1^0(\cdot)$, where $m_1^0(u_1) = \delta_0 + \delta_1 u_1$, δ_0 and δ_1 are some constants, which corresponds to a linear G×E interaction. If we fail to reject the null, then a linear model can be fit to further assess the linear G×E interaction. Otherwise, we conclude nonlinear G×E interaction. Power is evaluated under a sequence of alternative models with different values of τ, which is denoted by $H_1^\tau : m_1^\tau(\cdot) = m_1^0(\cdot) + \tau\{m_1(\cdot) - m_1^0(\cdot)\}$. When $\tau = 0$, the corresponding power is the false positive rate.

Figure 2.5 shows the size (when $\tau = 0$) and power (when $\tau > 0$) at significance level 0.05. We obtain 1000 Monte Carlo simulations each with 5000 replications to access the null distribution of test statistic under sample sizes $N = 200, 500$ with $\rho = 0.5$. The empirical Type I error under three MAFs are very close to the nominal level 0.05 and the power increases dramatically when MAF

increases from 0.1 to 0.3. Results for $\rho = 0.8$ is presented in Figure 2.6. Similarly, the empirical Type I error is close to 0.05 and the power increases rapidly when MAF increases from 0.1 to 0.3. Compared to the performance when $\rho = 0.5$ shown in Figure 2.5, the power increases a little bit slower when $\rho = 0.8$. The results indicate that our method can reasonably control the false positive rates and has appropriate power to detect the genetic variation.

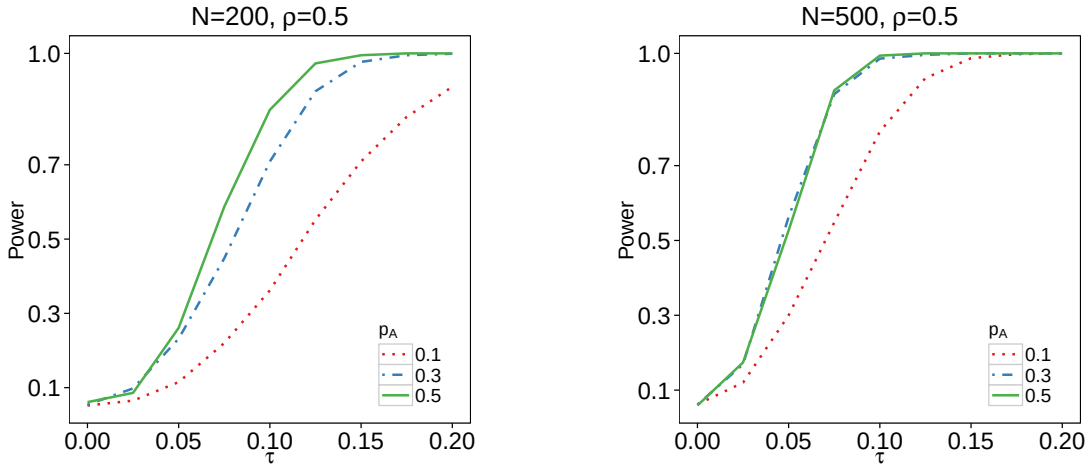

Figure 2.5 The empirical size and power of testing the linearity of nonparametric function m_1 under different MAFs when N=200, 500 and ρ=0.5.

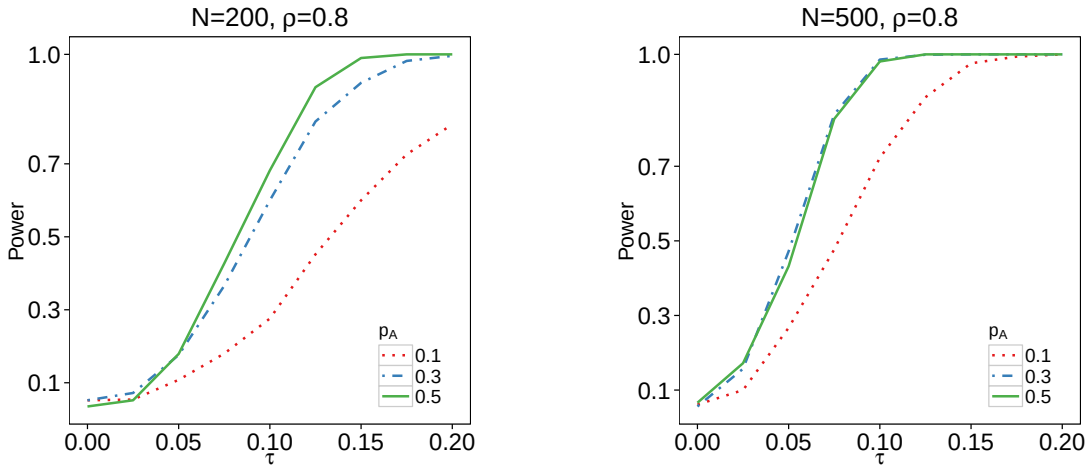

Figure 2.6 The empirical size and power of testing the linearity of nonparametric function m_1 under different MAFs when N=200, 500 and ρ=0.8.

2.7 Real data application

We applied the proposed FVICM model to a real data set from a study examining the association of the A118G SNP in OPRM1 to experimental pain sensitivity (Jonson and Terra, 2002). A group of 163 men and women in ages from 32 to 86 years participated in the study. Systolic blood pressure (SBP), diastolic blood pressure (DBP) and heart rate (HR) were measured at 6 Dobutamine dosage levels for each subject. Dobutamine is a medication that is used to treat congestive heart failure by increasing heart rate and cardiac contractility. Dobutamine was injected into these subjects to investigate their response in heart rate and blood pressure to this drug, at different dosage levels: 0 (baseline), 5, 10, 20, 30 and 40 mcg/min. In this study, dosage levels are treated as time and measurements at different dosage levels are considered as longitudinal measures. In addition to that, age and body mass index (BMI) were also recorded.

Total five SNPs in genes $Beta_1AR$ and $Beta_2AR$ were genotyped, namely, *codon16*, *codon27*, *codon49*, *codon389*, and *codon492*. We choose X_1= dosage level as the "time-varying" variable, and X_2 = age and X_3= BMI as the "time-invariant" variable. Our goal is to evaluate how the SNPs interact with age, BMI and dose level to affect SBP, DBP and HR. With the proposed FVICM model, we can model the dynamic gene effect on drug response under different dosage levels.

In this analysis, we test whether any SNP is associated with the drug response based on the hypothesis test $H_0 : m_1(u_1) = \delta_0 + \delta_1 u_1$ with p-value denoted by p_{m_1} in Table 2.3 - 2.5. We also reported the p-values for testing the significance of coefficients β_{11}, β_{12} and β_{13}, which are labeled by $p_{\beta_{11}}$, $p_{\beta_{12}}$ and $p_{\beta_{13}}$, based on the asymptotic normality of the estimates. We also compare our proposed model to an additive varying-coefficient model (AVCM) $E(Y|\mathbf{X},G) = \beta^*_{01}(X_1) + \beta^*_{02}X_2 + \beta^*_{03}X_3 + \{\beta^*_{11}(X_1) + \beta^*_{12}X_2 + \beta^*_{13}X_3\}G$, where $\beta^*_{01}(\cdot)$ and $\beta^*_{11}(\cdot)$ are unknown functions of X_1. To see the relative gain by integrative analysis, we calculate the MSEs of both models. The p-values for testing $H_0 : \beta^*_{11}(\cdot) = \beta^*_{12} = \beta^*_{13} = 0$ for AVCM is also reported in the tables and denoted by p_{AVCM}.

Table 2.3 summarizes the performance of our method for response SBP. In the table, p_{m_1} for all

the 5 SNPs are smaller than the significance level 0.05, which implies the nonlinear function of the SNPs on SBP in response to the dosage level, age and BMI as a whole. The MSEs in the last two columns shows that FVICM fits the data better than AVCM, indicating the benefit of integrative analysis. Besides, the testing results for AVCM do not show significance of the coefficients, which further implies that the genetic effects of SNPs are nonlinearly modified by the mixture of these three variables. Figure 2.7 shows the fitted nonlinear functions for each SNP, along with the 95% confidence bands.

Table 2.3 List of SNPs with MAF, allele, p-values under different hypothesis testing and MSE for SBP.

SNP ID	MAF	Alleles	p-value					MSE	
			p_{m_1}	$p_{\beta_{11}}$	$p_{\beta_{12}}$	$p_{\beta_{13}}$	p_{AVCM}	FVICM	AVCM
codon16	0.3990	A/G	<1.0E-04	0.0011	<1.0E-04	0.0917	0.5308	0.0403	0.0421
codon27	0.4160	G/C	<1.0E-04	<1.0E-04	0.0027	0.1675	0.6748	0.0388	0.0415
codon49	0.1387	G/A	<1.0E-04	<1.0E-04	0.3614	0.8668	0.2910	0.0398	0.0410
codon389	0.3045	G/C	<1.0E-04	<1.0E-04	<1.0E-04	0.7552	0.3927	0.0397	0.0431
codon492	0.4250	T/C	<1.0E-04	0.4102	<1.0E-04	0.0182	0.2990	0.0392	0.0409

Table 2.4 presents similar results for response DBP. The values of p_{m_1} shows that the test results for all 5 SNPs are significant, indicating nonlinear interactions for all 5 SNPs, while no significance is shown for AVCM model. MSEs further support our method by showing smaller value for FVICM comparing with AVCM. The estimated interaction curves with 95% confidence bands are shown in Figure 2.8.

Table 2.4 List of SNPs with MAF, allele, p-values under different hypothesis testing and MSE for DBP.

SNP ID	MAF	Alleles	p-value					MSE	
			p_{m_1}	$p_{\beta_{11}}$	$p_{\beta_{12}}$	$p_{\beta_{13}}$	p_{AVCM}	VICM	AVCM
codon16	0.3990	A/G	0.0066	<1.0E-04	0.2834	0.0007	0.3160	0.0366	0.0372
codon27	0.4160	G/C	0.0004	0.8431	<1.0E-04	<1.0E-04	0.0946	0.0360	0.0386
codon49	0.1387	G/A	0.0003	0.5750	<1.0E-04	0.0042	0.7986	0.0369	0.0395
codon389	0.3045	G/C	0.0001	<1.0E-04	0.9675	<1.0E-04	0.2615	0.0369	0.0377
codon492	0.4250	T/C	0.0001	0.7934	<1.0E-04	<1.0E-04	0.5837	0.0369	0.0389

In Table 2.5, the performance of our method for trait HR also leads to similar conclusion expect

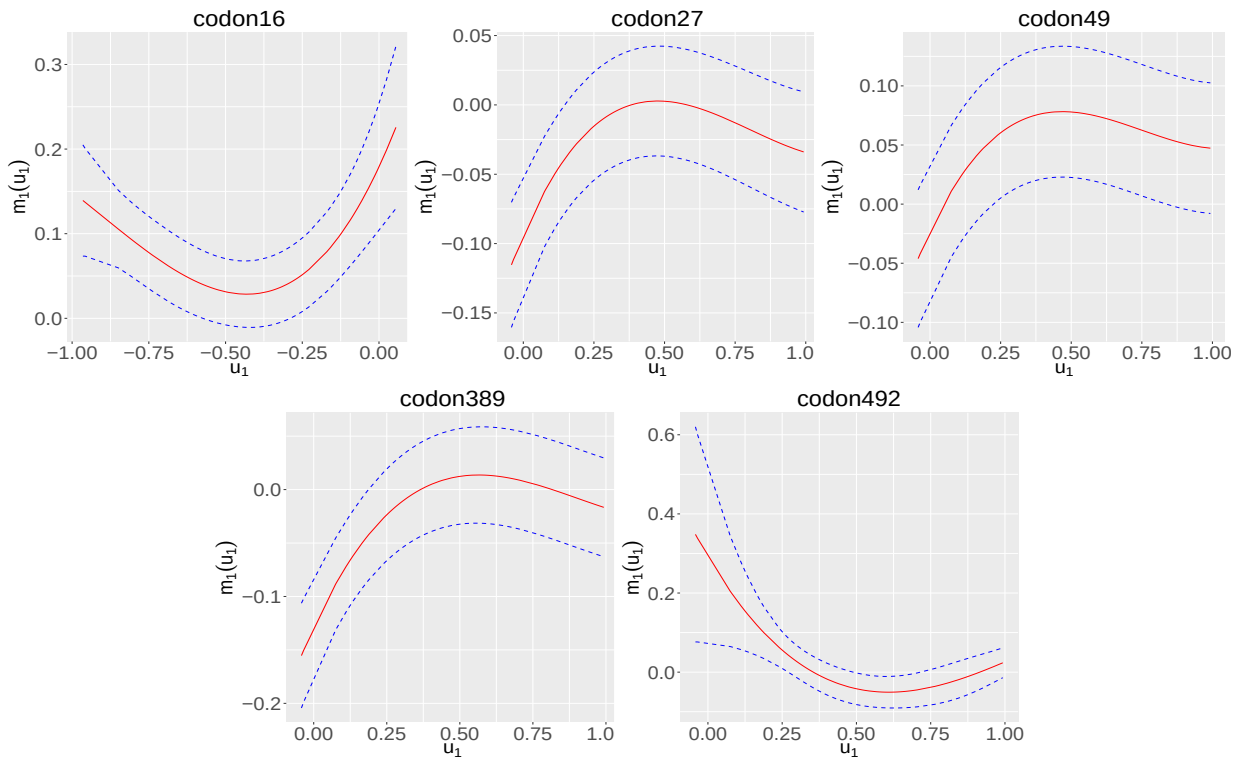

Figure 2.7 Plot of the estimate (solid curve) of the nonparametric function $m_1(u_1)$ for SNPs codon16, codon27, codon49, codon389 and codon492. The 95% confidence band is denoted by the dashed line. Response is SBP.

for SNP *codon16*, which shows significant test results for both models. For all the other SNPs, FVICM outperforms AVCM in terms of MSE. Figure 2.9 displays the corresponding estimated nonlinear interaction curves.

Table 2.5 List of SNPs with MAF, allele, p-values under different hypothesis testing and MSE for HR.

SNP ID	MAF	Alleles	p-value					MSE	
			p_{m_1}	$p_{\beta_{11}}$	$p_{\beta_{12}}$	$p_{\beta_{13}}$	p_{AVCM}	VICM	AVCM
codon16	0.3990	A/G	<1.0E-04	<1.0E-04	0.1158	0.0028	0.0328	0.0309	0.0308
codon27	0.4160	G/C	<1.0E-04	0.0007	0.6434	0.0001	0.9620	0.0320	0.0325
codon49	0.1387	G/A	0.0001	0.0147	0.0172	0.0133	0.8371	0.0298	0.0300
codon389	0.3045	G/C	<1.0E-04	<1.0E-04	0.0024	0.0021	0.8959	0.0311	0.0313
codon492	0.4250	T/C	0.0002	<1.0E-04	0.0011	0.0582	0.3732	0.0315	0.0316

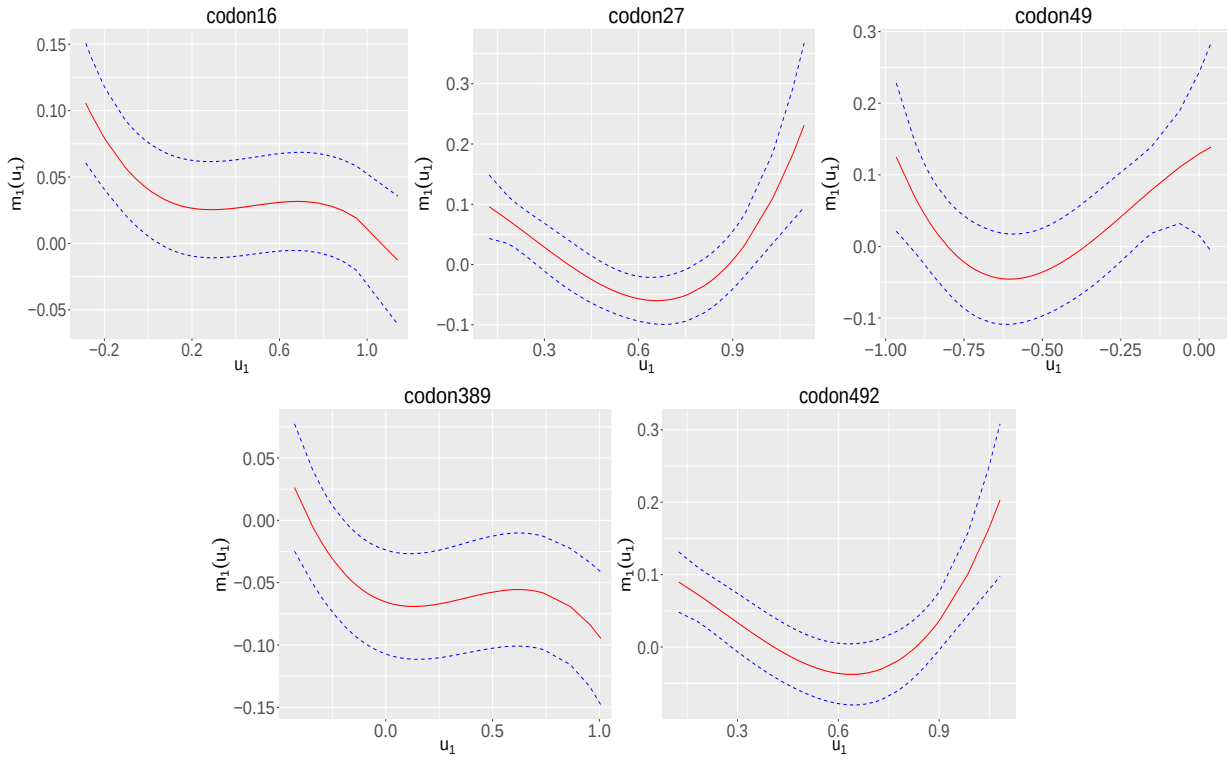

Figure 2.8 Plot of the estimate (solid curve) of the nonparametric function $m_1(u_1)$ for SNPs codon16, codon27, codon49, codon389 and codon492. The 95% confidence band is denoted by the dashed line. Response is DBP.

2.8 Discussion

In this paper, we propose a functional varying index coefficient modeling procedure to study gene effects nonlinearly modified by a mixture of environmental variables in a longitudinal design. We implement the quadratic inference function (QIF) method to estimate the index loading and spline coefficients. Furthermore, we apply the pseudo likelihood ratio test in a linear mixed model representation to test the linearity of the nonparametric coefficient function. Simulation study has been conducted to illustrate the estimation and testing procedures and confirm the asymptotical property. Real analysis shows that our model outperforms the additive varying coefficient model, which considers the G×E effect for each single environmental factor separately.

Our FVICM model distinguishes the varying coefficient model for longitudinal data. In fact, the varying coefficient model is a special case of our model when the dimension of the X variable

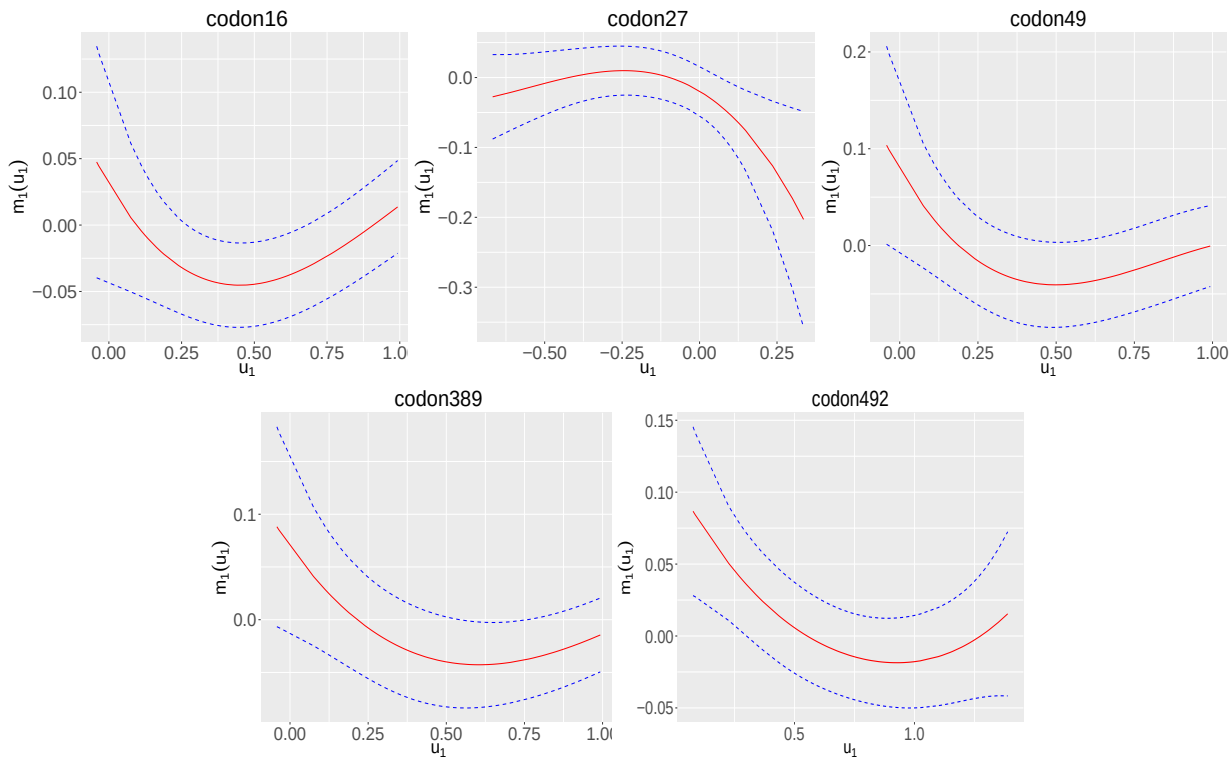

Figure 2.9 Plot of the estimate (solid curve) of the nonparametric function $m_1(u_1)$ for SNPs codon16, codon27, codon49, codon389 and codon492. The 95% confidence band is denoted by the dashed line. Response is HR.

reduces to one. FVICM is able to capture the effect of genes nonlinearly modified by the joint effect of multiple environmental variables as a whole. In addition, it can reduce multiple testing burden by treating multiple environmental variables as a single index variable.

We apply the model to a pain sensitivity study. Testing results indicate that all five SNPs have significant nonlinear interaction effects with environmental factors, which makes practical sense since these SNPs were genotyped from candidate genes. Our model was motivated by a practical need in G×E study. However, the method can be applied to any longitudinal data in which the purpose is to model nonlinear interaction effects. For example, we can consider gene expressions in a pathway (denoted as $\boldsymbol{X}$) and model how they regulate downstream genes (G) to affect a disease trait. Both the trait and gene expressions can be measured over time. Thus, one can understand the dynamic effect of genes nonlinearly regulated by a pathway to affect a disease trait.

CHAPTER 3

GENERALIZED FUNCTIONAL VARYING INDEX COEFFICIENT MODEL FOR DYNAMIC GENE-ENVIRONMENT INTERACTIONS

3.1 Introduction

Longitudinal data analysis is very common in epidemiological studies when the response variables are measured over time on objectives. Many studies demonstrated the increased power of a longitudinal design in detecting genetic associations over cross-sectional designs (Sitlani et al. 2015; Furlotte et al. 2015; Xu et al. 2014). On the other hand, there has been growing interest in the role of G×E interaction in many human diseases, such as Parkinson disease (Ross and Smith, 2007) and type 2 diabetes (Zimmet et al., 2001). In many studies, G×E has been traditionally investigated based on a single environment exposure model. However, evidence from an increasing number of studies has shown that risk of disease can be modified by simultaneous exposure to multiple environmental factors, which might be higher than what would be expected from simple addition of the single effects of environmental factors (Carpenter et al., 2002; Sexton and Hattis, 2007). Thus, of particular interest and complexity are assessing the combined effect of environmental mixtures and the mechanism in which they interact with genes to affect disease risk. Some researches have been done to assess nonlinear interactions between environmental mixtures and genes by applying some nonparametric or semiparametric models, such as the varying index coefficients model (VICM) proposed by Ma and Song (2015) and the partial linear multi-varying index coefficients model (PLMVICM) by Liu et al. (2016) and the generalized PLMVICM by Liu et al. (2017). However, these methods were developed for cross-sectional data and they can not be used for longitudinal data. This motivates us to extend the varying index coefficient model to longitudinal traits.

In our previous work, we proposed a functional varying index coefficient model to capture the nonlinear G×E interaction for continuous longitudinal traits. However, in practice, it is possible

that the response measured over time is a discrete variable, for example, a binary measure of a disease status. In human genetics, many disease traits are binary in nature, being affected vs unaffected (or cases vs controls). In order to investigate the dynamic nonlinear G×E interaction with environmental mixtures as a whole for a binary longitudinal trait, we propose the following generalized functional varying index coefficient model (gFVICM):

$$g\{E(Y_{ij}|\mathbf{X}_{ij}, G_{ij})\} = m_0(\boldsymbol{\beta}_0^T \mathbf{X}_{ij}) + m_1(\boldsymbol{\beta}_1^T \mathbf{X}_{ij}) G_{ij}, \tag{3.1}$$

where Y_{ij} (= 0 or 1) denotes the binary longitudinal response variable observed for the ith subject at the jth time point; $\mathbf{X}_{ij}$ is a p-dimensional vector of environmental variables, which can be either time-variant or time-invariant variables; G_i denotes the SNP variable which does not depend on time; $g(\cdot)$ is a known link function; $m_0(\cdot)$ and $m_1(\cdot)$ are unknown nonparametric smooth functions which depend on the data; and $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$ are p-dimensional vectors of index loading parameters. In this model, the function $m_1(\boldsymbol{\beta}_1^T \mathbf{X})$ captures the interaction effect between environmental mixtures and the genetic variable (e.g., a single nucleotide polymorphism (SNP)) on the risk of disease.

The aim of this paper is to develop a set of statistical estimation and hypothesis testing procedure for model (3.1). The Generalized Estimation Equation (GEE) method, proposed by Liang and Zeger in 1986, has been widely used in longitudinal data analysis. However, there are several disadvantages of GEE method due to some of its critical assumptions (Song et al., 2009). One disadvantage is that the consistency of GEE estimators are based on the consistency of estimators for the nuisance correlation parameter (Crowder, 1986, 1995). Another shortcoming of the GEE method is that model selection and hypothesis testing are complicated. This is because the estimation procedure of the GEE method does not involve an objective function. The quadratic inference function (QIF) approach proposed by Qu et al. (2000) is one of the improvements of the GEE method. The QIF avoids estimating the nuisance correlation parameters and has been confirmed by Qu et al. (2000) to be generally more efficient than the GEE. In addition, since the QIF is built upon an objective function which is asymptotically chi-square distributed, we can naturally

implement the model selection criterion such as BIC to the QIF. The asymptotic property can also allows us to conduct hypothesis tests. This motivates us to extend the QIF method to our model for estimation and hypothesis testing.

In our proposed estimation procedure, we first use penalized splines (Ruppert and Carroll, 2000) to approximate the nonparametric smooth functions $m_l(\cdot)$, l=0, 1. Then we develop a profile estimation procedure to estimate the index loading parameters and spline coefficients iteratively based on the QIF approach. In order to avoid overfitting and reduce the number of parameters in spline approximation, we use BIC method by adding a penalty to the objective function. Under certain regularity conditions we establish the asymptotic normality of the resulting estimators. In addition, we are interested in testing the linearity of G×E interaction, i.e. the linearity of function $m_1(\cdot)$. The QIF can be regarded as an inference function which has properties similar to the likelihood ratio test. Based on that, we construct a testing procedure for linearity of nonparametric interaction function, where the test statistic asymptotically follows a χ^2 distribution.

The rest of this chapter is organized in the following way: In Section 3.2, we propose an estimation procedure for model (3.1) and also provide the consistency and asymptotic normality of the proposed estimator; In Section 3.3, we derive a testing procedure for the linearity of nonparametric interaction function based on the goodness-of-fit test of QIF. The finite sample performance of the proposed procedure are accessed by Monte Carlo simulations illustrated in Section 3.4; In Section 3.5, the application of the proposed methodology is shown through the analysis of a pain sensitivity data with a binary response variable indicating whether a subject has hypertension or not (Yes=1, No=0); Some discussions are given in Section 3.6; The proofs of are rendered in Appendix.

3.2 The model and estimation methods

3.2.1 The model

For a longitudinal disease trait, suppose the binary response y_{ij}, the p-dimensional covariate vector $\boldsymbol{x}_{ij}$, and the SNP variable G_i are observed for the ith observation at the jth time point, where

i=1,...,N, j=1,...,n_i. Assume that the observations from different subjects are independent, but those within the same subject are correlated. We also assume the model satisfies the first moment assumption:

$$\mu_{ij}(\boldsymbol{x}_{ij}, G_i) = E(y_{ij}|\boldsymbol{x}_{ij}, G_i) = g^{-1}\{m_0(\boldsymbol{\beta}_0^T \boldsymbol{x}_{ij}) + m_1(\boldsymbol{\beta}_1^T \boldsymbol{x}_{ij})G_i\},$$

where $g^{-1}(\cdot)$ is a given inverse link function. If we use a logit link function for binary response, then the model can be written as

$$\mu_{ij}(\boldsymbol{x}_{ij}, G_i) = P(y_{ij} = 1|\boldsymbol{x}_{ij}, G_i) = \frac{\exp\{m_0(\boldsymbol{\beta}_0^T \boldsymbol{x}_{ij}) + m_1(\boldsymbol{\beta}_1^T \boldsymbol{x}_{ij})G_i\}}{1 + \exp\{m_0(\boldsymbol{\beta}_0^T \boldsymbol{x}_{ij}) + m_1(\boldsymbol{\beta}_1^T \boldsymbol{x}_{ij})G_i\}}.$$

For model identifiability, we have the constraints $\|\boldsymbol{\beta}_0\| = \|\boldsymbol{\beta}_1\| = 1$ and the first elements of $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$ are positive.

3.2.2 Quadratic inference function for gFVICM

First, we approximate the unknown coefficient functions $m_0(u_0)$ and $m_1(u_1)$ by truncated power spline basis as

$$m_l(u_l) = m_l(u_l, \boldsymbol{\beta}) \approx \mathbf{B}(u_l)^T \boldsymbol{\gamma}_l, \text{ for } l = 0, 1, \tag{3.2}$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}_0^T, \boldsymbol{\beta}_1^T)^T$, $\mathbf{B}(u) = (1, u, u^2, ..., u^q, (u-\kappa_1)_+^q, ..., (u-\kappa_K)_+^q)^T$ is a q-degree truncated power spline basis with K knots $\kappa_1, ..., \kappa_K$; $\boldsymbol{\gamma}_0$ and $\boldsymbol{\gamma}_1$ are $(q+K+1)$-dimensional vectors of spline coefficients.

A marginal approach such as the GEE assumes that the marginal mean μ_{ij} is a function of the covariates through a link function and the variance of y_{ij} is a function of the mean $\text{var}(y_{ij}) = V(\mu_i)$. The generalized estimation equation for longitudinal data is

$$\sum_{i=1}^{N} \dot{\boldsymbol{\mu}}_i^T \mathbf{V}_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i) = 0,$$

where $\mathbf{y}_i = (y_{i1}, ..., y_{in_i})^T$, $\boldsymbol{\mu}_i = E(\mathbf{y}_i)$ is the mean function and $\dot{\boldsymbol{\mu}}_i = \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\theta}}$ is the first derivative of $\boldsymbol{\mu}_i$ with respect to parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T)^T$, with $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_0^T, \boldsymbol{\gamma}_1^T)^T$. The covariance matrix $\mathbf{V}_i$ can be decomposed as $\mathbf{V}_i = \mathbf{A}_i^{1/2} \mathbf{R}(\rho) \mathbf{A}_i^{1/2}$ with $\mathbf{A}_i$ being a diagonal matrix of marginal variances and

$\mathbf{R}(\rho)$ being a common working correlation matrix with a small number of nuisance parameters ρ. Using the spline approximation in (3.2), the mean function can be written as

$$\boldsymbol{\mu}_i = \boldsymbol{\mu}_i(\boldsymbol{\theta}) = \begin{bmatrix} \mu_{i1}(\boldsymbol{\theta}) \\ \vdots \\ \mu_{in_i}(\boldsymbol{\theta}) \end{bmatrix} = \begin{bmatrix} g^{-1}\{\mathbf{B}^T(\boldsymbol{\beta}_0^T\mathbf{x}_{i1})\boldsymbol{\gamma}_0 + \mathbf{B}^T(\boldsymbol{\beta}_1^T\mathbf{x}_{i1})\boldsymbol{\gamma}_1 G_i\} \\ \vdots \\ g^{-1}\{\mathbf{B}^T(\boldsymbol{\beta}_0^T\mathbf{x}_{in_i})\boldsymbol{\gamma}_0 + \mathbf{B}^T(\boldsymbol{\beta}_1^T\mathbf{x}_{in_i})\boldsymbol{\gamma}_1 G_i\} \end{bmatrix},$$

and the first derivative of $\boldsymbol{\mu}_i$ is

$$\dot{\boldsymbol{\mu}}_i = \begin{bmatrix} (g^{-1})'\mathbf{B}_d^T(\boldsymbol{\beta}_0^T\mathbf{x}_{i1})\boldsymbol{\gamma}_0\mathbf{x}_{i1}^T & (g^{-1})'\mathbf{B}_d^T(\boldsymbol{\beta}_1^T\mathbf{x}_{i1})\boldsymbol{\gamma}_1 G_i\mathbf{x}_{i1}^T & (g^{-1})'\mathbf{B}^T(\boldsymbol{\beta}_0^T\mathbf{x}_{i1}) & (g^{-1})'\mathbf{B}^T(\boldsymbol{\beta}_1^T\mathbf{x}_{i1})G_i \\ \vdots & \vdots & \vdots & \vdots \\ (g^{-1})'\mathbf{B}_d^T(\boldsymbol{\beta}_0^T\mathbf{x}_{in_i})\boldsymbol{\gamma}_0\mathbf{x}_{in_i}^T & (g^{-1})'\mathbf{B}_d^T(\boldsymbol{\beta}_1^T\mathbf{x}_{in_i})\boldsymbol{\gamma}_1 G_i\mathbf{x}_{in_i}^T & (g^{-1})'\mathbf{B}^T(\boldsymbol{\beta}_0^T\mathbf{x}_{in_i}) & (g^{-1})'\mathbf{B}^T(\boldsymbol{\beta}_1^T\mathbf{x}_{in_i})G_i \end{bmatrix},$$

where $\mathbf{B}_d(u) = \frac{\partial \mathbf{B}(u)}{\partial u} = (0, 1, 2u, ..., qu^{q-1}, q(u-\kappa_1)_+^{q-1}, ..., q(u-\kappa_K)_+^{q-1})$.

In the QIF method, the inverse of the working correlation matrix can be approximated by a linear combination of several basis matrices, i.e.

$$\mathbf{R}^{-1}(\rho) \approx a_1\mathbf{M}_1 + ... + a_h\mathbf{M}_h,$$

where $\mathbf{M}_1$ is the identity matrix and $\mathbf{M}_2, ..., \mathbf{M}_h$ are known basis matrixes. For example, if the working correlation is exchangeable, $\mathbf{R}^{-1} \approx a_1\mathbf{M}_1 + a_2\mathbf{M}_2$ with $\mathbf{M}_2$ having 0 on the diagonal and 1 off-diagonal. If the working correlation is AR(1), then $\mathbf{R}^{-1} \approx a_1^*\mathbf{M}_1 + a_2^*\mathbf{M}_2^*$ and we can set $\mathbf{M}_2^*$ to have 1 on its two subdiagonals and 0 elsewhere. The advantage of this method is that the estimation of nuisance parameters $a_1, ..., a_h$ are not required.

Following this idea, we can derive the estimation function

$$\bar{g}_N(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N} g_i(\boldsymbol{\theta}) = \frac{1}{N}\begin{bmatrix} \sum_{i=1}^N \dot{\boldsymbol{\mu}}_i^T\mathbf{A}_i^{-1/2}\mathbf{M}_1\mathbf{A}_i^{-1/2}(\mathbf{y}_i - \boldsymbol{\mu}_i) \\ \vdots \\ \sum_{i=1}^N \dot{\boldsymbol{\mu}}_i^T\mathbf{A}_i^{-1/2}\mathbf{M}_h\mathbf{A}_i^{-1/2}(\mathbf{y}_i - \boldsymbol{\mu}_i) \end{bmatrix} \tag{3.3}$$

We cannot obtain the estimators by simply setting each element in $\bar{g}_N$ to be zero since the number of equations is more than the number of unknown parameters. To deal with this issue, we can estimate the parameters by minimizing the following quadratic inference function,

$$Q_N(\boldsymbol{\theta}) = N\bar{g}_N^T\bar{C}_N^{-1}\bar{g}_N,$$

where $\bar{C}_N = \frac{1}{N}\sum_{i=1}^{N} g_i g_i^T$ is a consistent estimator for var(g_i). By minimizing the quadratic inference function, we can obtain the estimation of the parameters as,

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} Q_N(\boldsymbol{\theta}).$$

In order to overcome the issue of over-parameterization, we can add a penalty term to QIF to penalize the number of knots in the approximation. The penalized QIF is written as

$$N^{-1}Q_N(\boldsymbol{\theta}) + \lambda \boldsymbol{\theta}^T \mathbf{D} \boldsymbol{\theta}, \tag{3.4}$$

where $\mathbf{D}$ is a diagonal matrix with 1 if the corresponding parameter is the spline coefficient associated with knots and 0 otherwise, that is, $\mathbf{D} = \text{diag}(\mathbf{0}^T_{(2p+q+1)\times 1}, \mathbf{1}^T_{K\times 1}, \mathbf{0}^T_{(q+1)\times 1}, \mathbf{1}^T_{K\times 1})$. Then the estimator is given by

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \left\{ N^{-1}Q_N(\boldsymbol{\theta}) + \lambda \boldsymbol{\theta}^T \mathbf{D} \boldsymbol{\theta} \right\}. \tag{3.5}$$

To determine the tuning parameter λ, we can extend the generalized cross-validation (Ruppert, 2002; Qu and Li, 2006; Bai et al., 2009) to the penalized QIF. The generalized cross-validation statistic is defined as

$$\text{GCV}(\lambda) = \frac{N^{-1}Q_N}{(1 - N^{-1}\text{df})^2}$$

with the effective degree of freedom $\text{df} = \text{tr}\left\{ (\ddot{Q}_N + 2N\lambda D)^{-1} \ddot{Q}_N \right\}$, where $\ddot{Q}_N$ is the second derivative of Q_N. The desirable choice of tuning parameter λ is which minimize the GCV(λ). In the implementation of GCV, the desired value of λ can be found using a grid search by predefining a set of values for λ .

3.2.3 Theoretical results

To establish the asymptotic properties for the estimators of the index loading parameters and the penalized spline regression coefficients, we assume $\boldsymbol{\theta}_0$ is the parameter satisfying $E_{\boldsymbol{\theta}_0}(g_i) = 0$. Theorem 3 provides the consistency of the resulting estimators. We show the $\sqrt{N}$-consistency and asymptotic normality of the estimators in Theorem 4. The theoretical results are similar to Theorem 1 and 2 in Chapter 2.

First, to handle the constraints $\|\boldsymbol{\beta}_0\|=\|\boldsymbol{\beta}_1\|=1$, and $\beta_{01}>0$, $\beta_{11}>0$, we set $\beta_{l1}=\sqrt{1-\|\boldsymbol{\beta}_{l,-1}\|_2^2}$ with $\boldsymbol{\beta}_{l,-1}=(\beta_{l2},...,\beta_{lp})^T$ for l=1, 2. Then the parameters space of $\boldsymbol{\beta}_l$, l=1,2, becomes

$$[\{(\sqrt{1-\|\boldsymbol{\beta}_{l,-1}\|_2^2},\beta_{l2},...,\beta_{lp})^T\}:\|\boldsymbol{\beta}_{l,-1}\|_2^2<1].$$

Let

$$\mathbf{J}_l=\frac{\partial\boldsymbol{\beta}_l}{\partial\boldsymbol{\beta}_{l,-1}^T}=\begin{pmatrix}-\boldsymbol{\beta}_{l,-1}^T/\sqrt{1-\|\boldsymbol{\beta}_{l,-1}\|_2^2}\\ \mathbf{I}_{p-1}\end{pmatrix}$$

be the Jacobian matrix of dimension $p\times(p-1)$. Denote $\boldsymbol{\beta}_{-1}=(\boldsymbol{\beta}_{0,-1}^T,\boldsymbol{\beta}_{1,-1}^T)^T$, and $\boldsymbol{\theta}^*=(\boldsymbol{\beta}_{-1},\boldsymbol{\gamma})^T$. From $\boldsymbol{\theta}$ to $\boldsymbol{\theta}^*$, we have Jacobian matrix $\mathbf{J}=\text{diag}(\mathbf{J}_0,\mathbf{J}_1,\mathbf{I}_{q+K+1},\mathbf{I}_{q+K+1})$.

Theorem 3 *Suppose the assumptions (A1)-(A6) in the Appendix are satisfied, and the smoothing parameter $\lambda_N=o(1)$, then the estimator $\widehat{\boldsymbol{\theta}}$, which is obtained by minimizing the penalized quadratic function in (3.4), exists and converges to $\boldsymbol{\theta}_0$ in probability.*

Theorem 4 *Suppose the assumptions (A1)-(A6) in the Appendix are satisfied, and the smoothing parameter $\lambda_N=o(N^{-1/2})$, then the estimator $\widehat{\boldsymbol{\theta}}$ obtained by minimizing the penalized quadratic function in (3.4) is asymptotically normally distributed, i.e.,*

$$\sqrt{N}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\xrightarrow{d}N(\boldsymbol{0},\boldsymbol{J}(\boldsymbol{G}_0^T\boldsymbol{C}_0^{-1}\boldsymbol{G}_0)^{-1}\boldsymbol{J}^T),$$

where the detailed calculation of $\boldsymbol{G}_0$ and $\boldsymbol{C}_0$ are given in the Appendix.

3.3 Model selection and hypothesis test

3.3.1 Model selection

Model selection is important in the spline approximation since too many parameters in the model might result in the overfitting issue. According to the theocratical property of generalized method of moments estimator (Hansen, 1982), under the assumption $E(g_1)=0$ and also the number of estimating equations is larger than the number of parameters, we have $Q(\widehat{\boldsymbol{\theta}})\to\chi_{r-k}^2$ in distribution,

where r is the dimension of $\bar{g}_N(\boldsymbol{\theta})$, k is the dimension of $\boldsymbol{\theta}$ and $\widehat{\boldsymbol{\theta}}$ is the estimator by minimizing the QIF when certain order and number of knots are chosen. This asymptotic property of the QIF provides a goodness-of-fit test, which can be useful to determine the order and number of knots to be selected in our model. However, it is also possible that the goodness-of-fit tests fail to reject several different models which may not be nested. Since $Q(\widehat{\boldsymbol{\theta}})$ is asymptotically chi-square distributed, it is natural to extend the BIC to the QIF approach, by replacing twice the negative log-likelihood function by the QIF objective function. In particular, the BIC criterion for a model with r estimating equation and k parameters is given as,

$$Q(\widehat{\boldsymbol{\theta}}) + (r-k)\ln N,$$

The model with the minimum BIC would be considered the optimal one. If we choose h basis matrices in (3.3), then $r-k = hk-k = (h-1)k$.

In our simulation and real data application, we search the optimal order and number of knots over a set of combinations of q and K using the BIC criterion. Knots are evenly distributed in the range of the single index $\boldsymbol{\beta}^T\mathbf{X}$.

3.3.2 Nonparametric goodness-of-fit test based on QIF

The QIF can also be regarded as an inference function since it has properties similar to the likelihood ratio test. Suppose that the d-dimensional parameter vector $\boldsymbol{\gamma}$ is partitioned into $(\boldsymbol{\psi}, \boldsymbol{\zeta})$, where $\boldsymbol{\psi}$ is the parameter of interest with dimension d_1, and $\boldsymbol{\zeta}$ is the nuisance parameter with dimension $d_2 = d - d_1$. If we are interested in testing

$$H_0 : \boldsymbol{\psi} = \boldsymbol{\psi}_0,$$

the test statistic

$$Q(\boldsymbol{\psi}_0, \widetilde{\boldsymbol{\zeta}}) - Q(\widehat{\boldsymbol{\psi}}, \widehat{\boldsymbol{\zeta}})$$

follows an asymptotically chi-square distribution with d_1 degrees of freedom. The following theorem introduced by Qu et al. (2000) provided a way to conduct hypothesis testing in the QIF framework.

Theorem 5 *(Qu et al., 2000) Suppose that all required regularity conditions are satisfied and* $\boldsymbol{\psi}$ *has dimension* d_1*. Under the null hypothesis,* $Q(\boldsymbol{\psi}_0, \widetilde{\boldsymbol{\zeta}}) - Q(\widehat{\boldsymbol{\psi}}, \widehat{\boldsymbol{\zeta}})$ *is asymptotically chi-square distribution with* d_1 *degrees of freedom, where*

$$\widetilde{\boldsymbol{\zeta}} = \arg\min Q(\boldsymbol{\psi}_0, \boldsymbol{\zeta}), \quad (\widehat{\boldsymbol{\psi}}, \widehat{\boldsymbol{\zeta}}) = \arg\min Q(\boldsymbol{\psi}, \boldsymbol{\zeta}). \tag{3.6}$$

When there is no nuisance parameter, which is a special case of the condition in Theorem 5, $Q(\boldsymbol{\gamma}_0) - Q(\widehat{\boldsymbol{\gamma}})$ has an asymptotical chi-square distribution with d degree of freedom under the null hypothesis.

3.3.3 Test for linearity of interaction function in gFVICM

In our proposed gFVICM model (3.1), it is of interest to test the unspecified coefficient function. In particular, we are interested in testing whether a linear function is good enough to describe the G×E interaction. If the we fail to reject the linearity of the coefficient function, then a parametric linear interaction function should be fitted to further assess if there exists linear G×E interaction; otherwise, we conclude there exists nonlinear G×E interaction. Let $u_1 = \boldsymbol{\beta}_1^T \mathbf{X}$. With the truncated power spline basis, the coefficient function can be modeled by

$$m_1(u_1) \approx \gamma_{10} + \gamma_{11} u_1 + \gamma_{12} u_1^2 + \cdots + \gamma_{1q} u_1^q + \sum_{k=q+1}^{K+q+1} \gamma_{1k} (u_1 - \kappa_k)_+^q.$$

Our goal is to test the linearity of $m_1(u_1)$, which is equivalent to test

$$H_0 : \gamma_{12} = \cdots = \gamma_{1,K+q+1} = 0.$$

Let $\widetilde{\boldsymbol{\theta}}$ be the estimator of the full parameter $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T)^T$ under the null hypothesis with

$$\widetilde{\boldsymbol{\theta}} = (\widetilde{\boldsymbol{\beta}}^T, \widetilde{\boldsymbol{\gamma}}_0^T, \widetilde{\gamma}_{10}, \widetilde{\gamma}_{11}, \mathbf{0}^T)^T = \underset{\gamma_{12}=\cdots=\gamma_{1,K+q+1}=0}{\arg\min} Q_N(\boldsymbol{\theta}),$$

and the estimator of $\boldsymbol{\theta}$ under the alternative as

$$\widehat{\boldsymbol{\theta}} = \arg\min Q_N(\boldsymbol{\theta}).$$

Then the test statistic

$$T_N = Q_N(\widetilde{\boldsymbol{\theta}}) - Q_N(\widehat{\boldsymbol{\theta}}),$$

asymptotically follows a chi-square distribution with $K+q-1$ degrees of freedom, following Theorem 5.

3.4 Simulation study

The finite sample performance of the proposed method was evaluated through Monte Carlo simulation studies. We considered the following logistic regression model

$$P(y_{ij}=1|\mathbf{X}_{ij},G_i,\boldsymbol{\beta}) = \frac{\exp\{\eta(\mathbf{X}_{ij},G_i,\boldsymbol{\beta})\}}{1+\exp\{\eta(\mathbf{X}_{ij},G_i,\boldsymbol{\beta})\}},$$

where

$$\eta(\mathbf{X}_{ij},G_i,\boldsymbol{\beta}) = m_0(\boldsymbol{\beta}_0^T\mathbf{X}_{ij}) + m_1(\boldsymbol{\beta}_1^T\mathbf{X}_{ij})G_i.$$

We simulated a three-dimensional environmental variables $\mathbf{X} = (X_1, X_2, X_3)$. For the ith subject, $X_{1ij}, X_{2ij}, X_{3ij}$ are independently generated from a uniform distribution $U(0,1)$. We set the minor allele frequency (MAF) as p_A = 0.1, 0.3, 0.5 and assumed Hardy-Weinberg equilibrium. We used AA, Aa and aa to denote three different SNP genotypes. These genotypes were simulated from a multinomial distribution with frequencies p_A^2, $2p_A(1-p_A)$ and $(1-p_A)^2$, respectively. Variable G was coded as {0,1,2}, corresponding to genotypes $\{aa, Aa, AA\}$ respectively. To create correlated responses, we implemented the R package 'bindata' developed by Leisch et al. (1998) under an AR(1) correlation structure with correlation parameter ρ=0.5. When implementing the function 'rmvbin' to generate the correlated binary data, one should specify the marginal probabilities and the correlation structure.

We set $m_0(u_0) = \cos(\pi u_0)$ and $m_1(u_1) = \sin[\pi(u_1 - A)/(B-A)]$ with $A = \sqrt{3}/2 - 1.645/\sqrt{12}$ and $B = \sqrt{3}/2 + 1.645/\sqrt{12}$. The true parameters were $\boldsymbol{\beta}_0 = (\sqrt{5}, \sqrt{4}, \sqrt{4})/\sqrt{13}$ and $\boldsymbol{\beta}_1 =$

$(1,1,1)/\sqrt{3}$. We drew 500 data sets with sample size $N = 200, 500$ and time points $n_i = T = 10, 20$, respectively. The basis matrix $\mathbf{M}_2$ was set to have 1 on its two subdiagonals and 0 elsewhere. The order and number of knots of the splines were selected through the BIC method.

3.4.1 Performance of estimation

Table 3.1 and Table 3.2 summarize the parameters estimation results under different sample sizes and measurement times respectively. In these two tables, the average bias (Bias), the standard deviation of the 500 estimates (SD), the average of the estimated standard error (SE) based on the theoretical results, and the estimated coverage probability (CP) at 95% confidence level are reported. It is shown from each table that, as the sample size increases, the performance of the estimation improves by showing smaller bias, SD and SE. More repeated measurement for each subject also results in improvement in estimations, which can be shown when we compare Table 3.1 and Table 3.2. For example, the CP for β_{01} improves from 86.8% to 90% when the number of measurement time increases from 10 to 20, under a sample size of 200. The estimation of the loading parameter $\boldsymbol{\beta}_1$ improves as MAF p_A increases, while the estimation of $\boldsymbol{\beta}_0$ show a opposite direction. This is because we have limited data information to estimate the marginal effects $m_0(\cdot)$ when p_A increases.

Table 3.1 Simulation results under different MAFs $p_A = 0.1, 0.3, 0.5$ with sample size $N = 200, 500$, $T = 10$ and correlation ρ=0.5.

			$p_A = 0.1$				$p_A = 0.3$				$p_A = 0.5$			
N	Param	True	Bias	SD	SE	CP	Bias	SD	SE	CP	Bias	SD	SE	CP
200	β_{01}	0.620	-0.008	0.058	0.057	93.6	-0.003	0.074	0.064	90.6	-0.007	0.084	0.068	86.8
	β_{02}	0.555	-0.002	0.066	0.056	90.0	-0.004	0.080	0.062	87.2	-0.008	0.093	0.065	84.5
	β_{03}	0.555	-4.1E-05	0.064	0.056	91.8	-0.008	0.074	0.062	88.8	-0.006	0.094	0.066	86.6
	β_{11}	0.577	-0.013	0.134	0.091	82.2	-0.003	0.092	0.072	87.2	0.007	0.088	0.063	84.9
	β_{12}	0.577	-0.024	0.139	0.090	79.2	-0.10	0.096	0.071	85.2	-0.013	0.085	0.062	87.2
	β_{13}	0.577	-0.013	0.140	0.091	82.4	-0.11	0.095	0.071	86.0	-0.014	0.088	0.062	85.4
500	β_{01}	0.620	0.002	0.038	0.038	94.8	-0.002	0.043	0.043	95.4	-0.003	0.047	0.048	95.0
	β_{02}	0.555	0.003	0.039	0.037	93.0	-0.002	0.043	0.042	93.4	-2.1E-04	0.052	0.046	92.4
	β_{03}	0.555	-0.003	0.038	0.037	93.8	-6.3E-04	0.045	0.042	93.0	-0.004	0.050	0.046	92.5
	β_{11}	0.577	-0.007	0.078	0.065	89.6	-0.002	0.055	0.049	92.0	0.002	0.045	0.044	94.2
	β_{12}	0.577	-0.003	0.079	0.066	88.0	-0.001	0.052	0.049	92.8	-0.003	0.049	0.043	91.4
	β_{13}	0.577	-0.005	0.075	0.066	90.6	-0.004	0.054	0.049	92.8	-0.005	0.047	0.043	92.0

Table 3.2 Simulation results under different MAFs $p_A = 0.1, 0.3, 0.5$ with sample size $N = 200, 500$, $T = 20$ and correlation ρ=0.5.

			$p_A = 0.1$				$p_A = 0.3$				$p_A = 0.5$			
N	Param	True	Bias	SD	SE	CP	Bias	SD	SE	CP	Bias	SD	SE	CP
200	β_{01}	0.620	-0.002	0.044	0.043	94.4	-0.002	0.053	0.049	94.2	-0.011	0.062	0.052	90.0
	β_{02}	0.555	-0.004	0.048	0.042	90.2	-0.003	0.056	0.048	90.2	-0.003	0.065	0.050	88.3
	β_{03}	0.555	2.3E-04	0.048	0.042	91.2	-0.002	0.057	0.048	90.2	0.005	0.063	0.050	87.4
	β_{11}	0.577	-0.006	0.074	0.064	90.0	0.008	0.062	0.054	91.8	0.004	0.063	0.050	88.2
	β_{12}	0.577	-0.004	0.075	0.064	91.4	-0.009	0.063	0.053	89.8	-0.009	0.064	0.050	87.1
	β_{13}	0.577	-0.005	0.072	0.064	91.8	-0.009	0.063	0.054	90.0	-0.006	0.064	0.050	86.8
500	β_{01}	0.620	-0.002	0.028	0.028	97.2	-0.002	0.033	0.033	95.0	-0.004	0.037	0.036	93.2
	β_{02}	0.555	-3.2E04	0.028	0.027	94.8	0.001	0.033	0.032	94.0	7.9E-05	0.038	0.035	92.8
	β_{03}	0.555	3.5E04	0.029	0.027	93.2	-0.002	0.034	0.032	92.8	6.3E-04	0.037	0.035	94.2
	β_{11}	0.577	-0.005	0.044	0.044	92.2	0.004	0.039	0.036	92.4	0.006	0.038	0.035	93.0
	β_{12}	0.577	-0.002	0.045	0.045	91.0	-0.003	0.037	0.036	93.0	-0.004	0.036	0.035	94.2
	β_{13}	0.577	0.003	0.042	0.042	96.4	-0.005	0.038	0.036	93.4	-0.005	0.035	0.035	93.6

The plots for the estimations of $m_0(u_0)$ and $m_1(u_1)$ under different sample sizes and number of replications are shown in Figure 3.1–3.4. The estimated and true functions are denoted by the solid and dashed lines, respectively. The 95% confidence bands are denoted by the dotted-dash line. The estimated curves almost overlap with the corresponding true curves as shown in the plots, indicating the estimation accuracy of the method. Also the confidence bands are tight, especially for large sample size and large number of measurement times. Note that the estimation for the interaction effects $m_1(u_1)$ improves as MAF p_A increases, while the estimation for the marginal effects $m_0(u_0)$ show a opposite direction, which coincides with the results for the parametric estimation in Table 3.1 and Table 3.2.

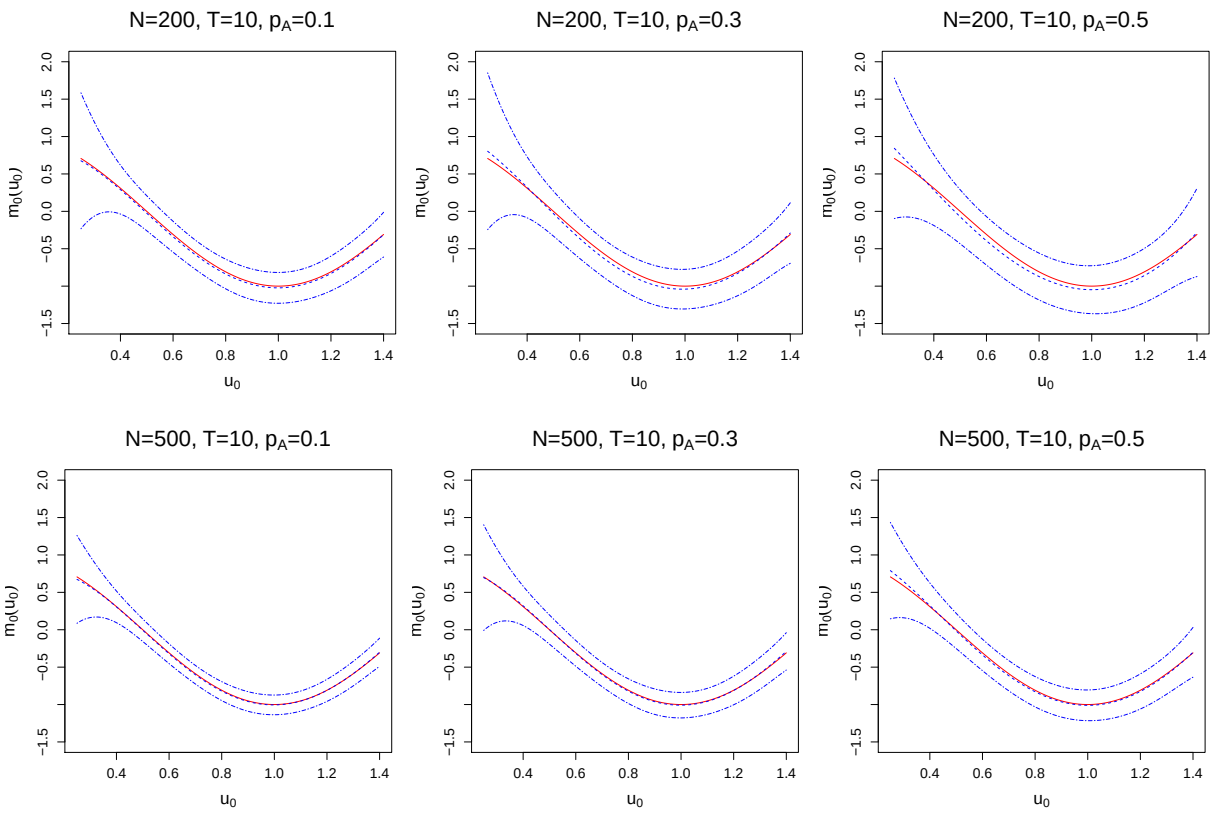

Figure 3.1 The estimation of function $m_0(\cdot)$ when N=200, 500 and T=10. The estimated and true functions are denoted by the solid and dashed lines respectively. The 95% confidence bands are denoted by the dotted-dash lines.

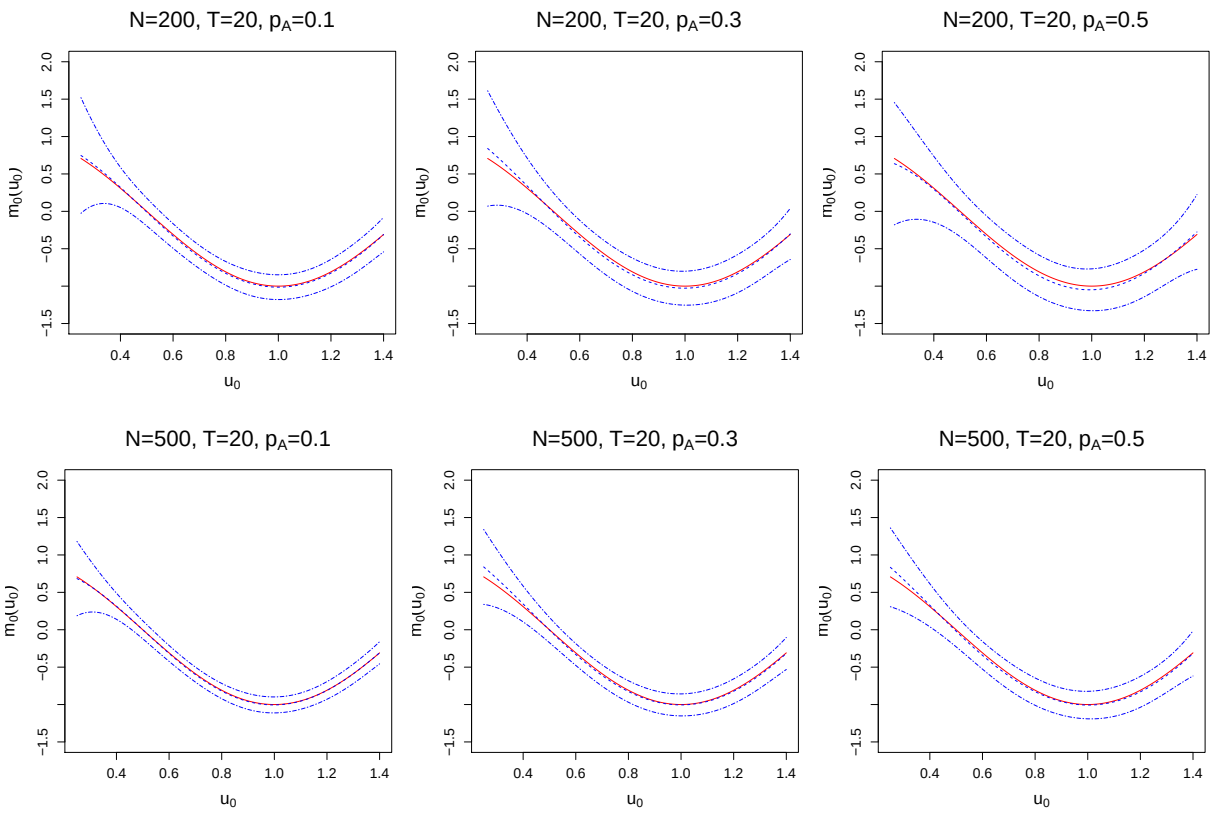

Figure 3.2 The estimation of function $m_0(\cdot)$ when N=200, 500 and T=20. The estimated and true functions are denoted by the solid and dashed lines respectively. The 95% confidence bands are denoted by the dotted-dash lines.

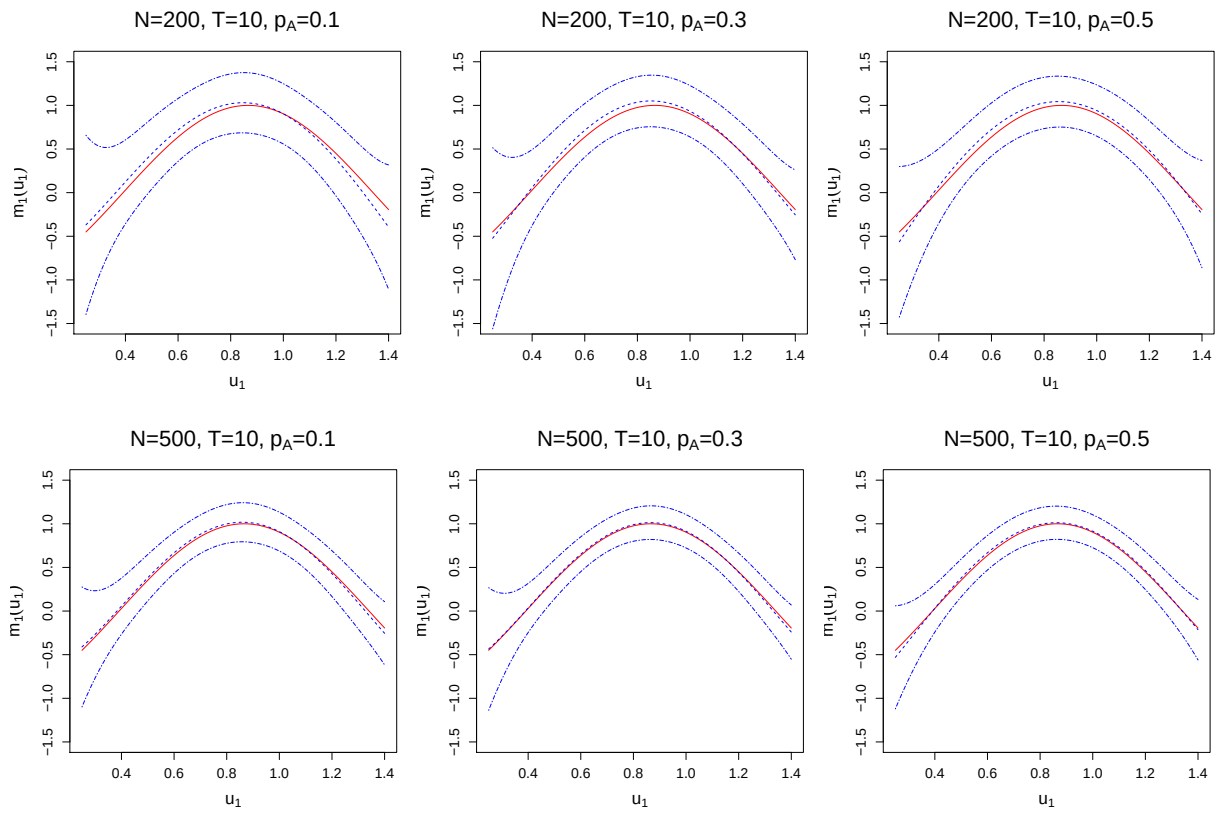

Figure 3.3 The estimation of function $m_1(\cdot)$ when N=200, 500 and T=10. The estimated and true functions are denoted by the solid and dashed lines respectively. The 95% confidence bands are denoted by the dotted-dash lines.

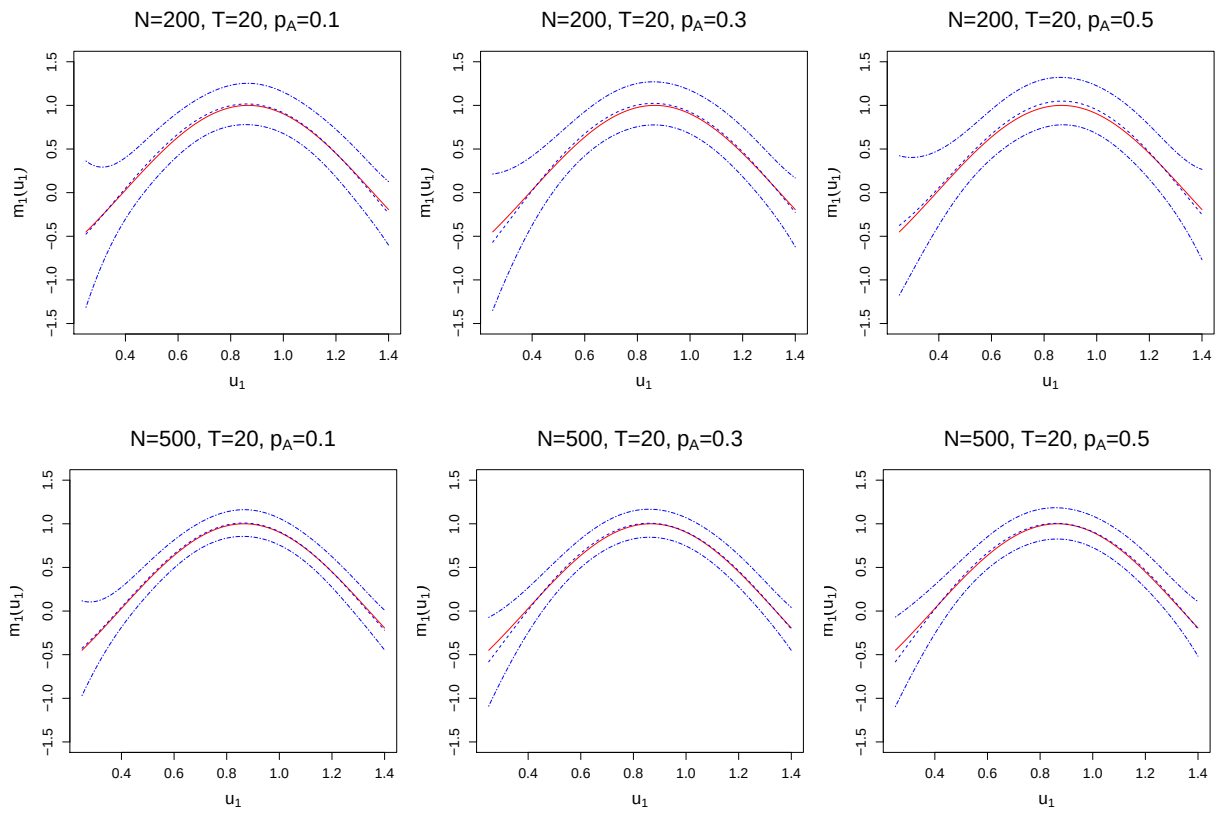

Figure 3.4 The estimation of function $m_1(\cdot)$ when N=200, 500 and T=20. The estimated and true functions are denoted by the solid and dashed lines respectively. The 95% confidence bands are denoted by the dotted-dash lines.

3.4.2 Performance of hypothesis tests

We evaluated the performance of the test for the nonparametric function under the null hypothesis $H_0 : m_1(\cdot) = m_1^0(\cdot)$, where $m_1^0(u_1) = \delta_0 + \delta_1 u_1$, δ_0 and δ_1 are some constants, which corresponds to a linear G×E interaction. Power is evaluated under a sequence of alternative models with different values of τ, which is denoted by $H_1^\tau : m_1^\tau(\cdot) = m_1^0(\cdot) + \tau\{m_1(\cdot) - m_1^0(\cdot)\}$. When τ = 0, the corresponding power is the false positive rate.

Figure 3.5 shows the size (when τ = 0) and power (when $\tau > 0$) at significance level 0.05 based on 500 Monte Carlo simulations for N=200, 500 under different measurement times T=10 (left panel) and T=20 (right panel). The empirical Type I error is large when $N = 200$, which decreases dramatically when the sample size increases to N=500. The power increases when the

sample size increases from 200 to 500. The results indicate that our method can reasonably control the false positive rates and has appropriate power to detect the linearity function under a relatively large sample size. Comparing the results for T=10 and T=20, we can see that the testing power improves when the number of measurement time increases.

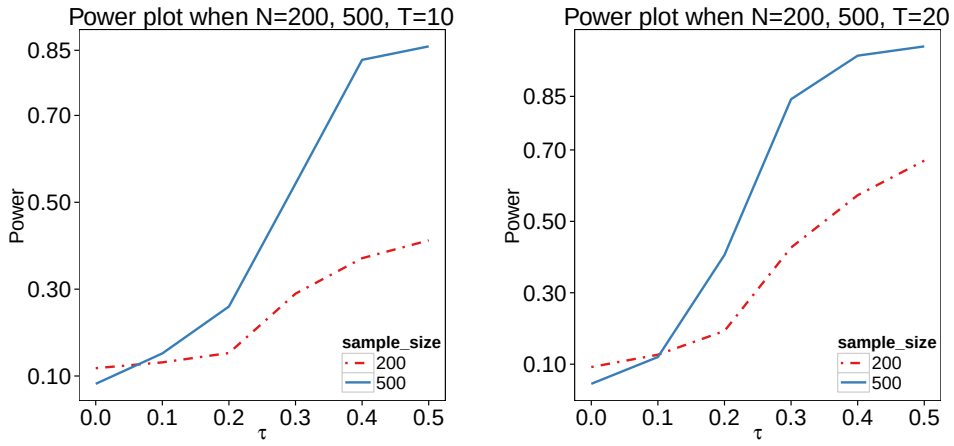

Figure 3.5 The empirical size and power of testing the linearity of nonparametric function m_1 when N=200, 500 and T=10, 20.

To assess how the values of MAF affect the testing performance, we plot the power plot under different MAFs p_A=0.1, 0.3, 0.5 when N=500, T=10, which is shown in Figure 3.6. Note that the power increases dramatically when MAF increases from 0.1 to 0.3. The values of power are very close when p_A=0.3 and 0.5.

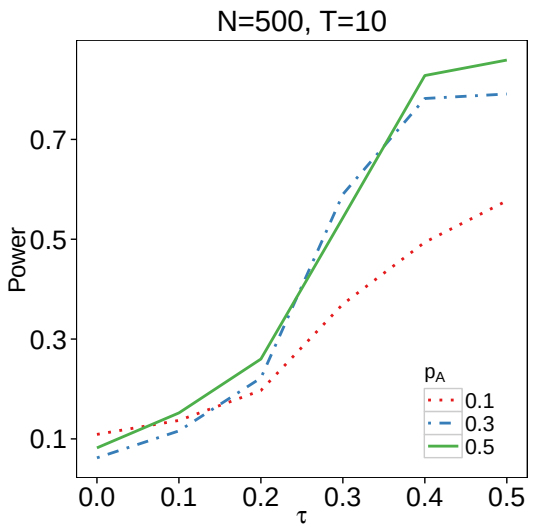

Figure 3.6 The empirical size and power of testing the linearity of nonparametric function m_1 under different MAFs when N=500 and T=10.

3.5 Real data application

We applied the proposed gFVICM model to a data set from a study examining the association of the A118G SNP of OPRM1 to experimental pain sensitivity. A sample of 163 healthy volunteers were evolved in this study. For each volunteer, Systolic Blood Pressure (SBP) and Diastolic Blood Pressure (DBP) were measured at 6 Dobutamine dosage levels: 0 (baseline), 5, 10, 20, 30 and 40mcg/min. Clinically, a person is said to be hypertensive if the individuals SBP is greater than 140mm Hg or DBP is greater than 90mm Hg (Choi et al. 2014). Thus, the response variable Y is a binary variable indicating whether a person has hypertension or not, i.e. $Y = 1$ for hypertension and $Y = 0$ for non-hypertension.

One longitudinal covariate X_1= dosage, two time-invariant covariates X_2 = age and X_3= BMI are included as the environmental factors in the model. The genetic variables are five SNPs located at *codon16*, *codon27*, *codon49*, *codon389*, and *codon492* in the gene. Our purpose is to evaluate how the mixture of age, BMI and dosage modifies the SNP effect on the risk of hypertension. In particular, we test the hypothesis $H_0 : m_1(u_1) = \delta_0 + \delta_1 u_1$ with p-value denoted by p_{m_1} in Ta-

ble 2.7. We also reported the p-values for testing the significance of three components of index loading coefficients $\boldsymbol{\beta}_1 = (\beta_{11}, \beta_{12}, \beta_{13})$, which are labeled by $p_{\beta_{11}}$, $p_{\beta_{12}}$ and $p_{\beta_{13}}$, based on the asymptotic property of the estimations. We also compared our proposed model to a generalized additive varying-coefficient model (gAVCM) $E(Y|\mathbf{X}, G) = \eta\{\beta^*_{01}(X_1) + \beta^*_{02}X_2 + \beta^*_{03}X_3 + (\beta^*_{11}(X_1) + \beta^*_{12}X_2 + \beta^*_{13}X_3)G\}$, where $\beta^*_{01}(\cdot)$ and $\beta^*_{11}(\cdot)$ are unknown functions of X_1. To see the relative gain by integrative analysis, we calculated the objective function Q_N in both models. The p-values for testing $H_0 : \beta^*_{11}(\cdot) = \beta^*_{12} = \beta^*_{13} = 0$ for gAVCM are also reported in the tables and denoted by p_{gAVCM}.

In Table 3.3, p_{m_1} for all the 5 SNPs are smaller than the significance level 0.05, which means the functions capturing the G×E interactions are nonlinear for all these 5 SNPs. The objective function Q_N in the last two columns shows that gFVICM fits the data better than gAVCM, indicating the benefit of integrative analysis. Besides, the testing results for gAVCM do not show significance of the coefficients for interactions. The results imply that the genetic effects of SNPs are modified by the mixture of environmental variables, rather than separately. Figure 3.7 exhibits the fitted nonlinear curves indicating G×E interactions for each SNP, along with the 95% confidence bands.

Table 3.3 List of SNPs with MAF, allele, p-values under different hypothesis testing and values of objective function Q_N.

SNP ID	MAF	Alleles	p-value					Q_N	
			p_{m_1}	$p_{\beta_{11}}$	$p_{\beta_{12}}$	$p_{\beta_{13}}$	p_{gAVCM}	gFVICM	gAVCM
codon16	0.3990	A/G	<1.0E-04	<1.0E-04	0.0207	0.3475	0.2960	3.9240	11.2082
codon27	0.4160	G/C	<1.0E-04	0.2329	<1.0E-04	0.0014	0.6982	6.9502	9.3500
codon49	0.1387	G/A	<1.0E-04	<1.0E-04	<1.0E-04	0.6325	0.1777	6.6303	12.2648
codon389	0.3045	G/C	<1.0E-04	<1.0E-04	<1.0E-04	0.3329	0.8436	3.3678	10.5593
codon492	0.4250	T/C	<1.0E-04	0.6731	<1.0E-04	0.0008	0.5001	6.0766	7.4877

Table 3.4 displays the estimated odds for different genotypes at different dosage levels. The changes in the values of odds demonstrate the interaction between SNP and environmental mixtures at different dosage levels. For example, we noted that the odds for genotype AA in SNP *codon16* does not change too much as the dosage level increases, which means that the genetic

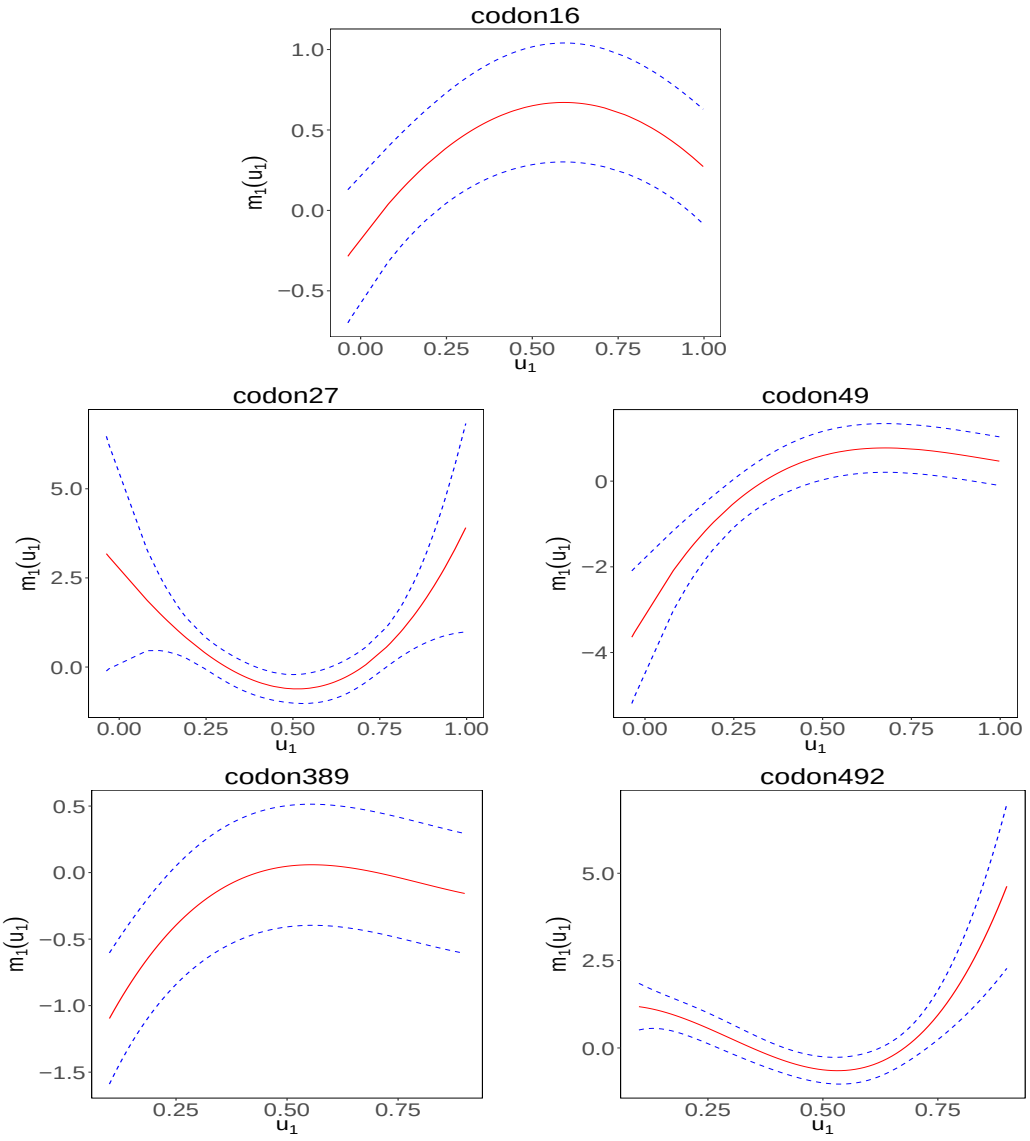

Figure 3.7 Plot of the estimate (solid curve) of the nonparametric function $m_1(u_1)$ for SNPs codon16, codon27, codon49, codon389 and codon492. The 95% confidence bands are denoted by the dashed lines.

effect of this genotype remains the same when subjects are exposed to different environments at different doses. While for the other two genotypes, there is an increase in the value of odds until dosage level four, indicating increased blood pressure as dosage level increases from 0mcg/min to 20mcg/min.

Table 3.4 Estimated odds for different genotypes at each dosage levels.

		Dosage level					
SNP ID	Genotyoe	1	2	3	4	5	6
codon16	AA	0.92	0.92	0.92	0.92	0.92	0.93
	GG	1.22	1.89	2.63	3.66	3.17	1.73
	GA	1.35	1.66	1.94	2.28	2.14	1.63
codon27	GG	1.25	2.07	2.75	2.96	2.32	1.96
	CC	1.06	1.75	2.34	2.58	2.06	1.73
	CG	1.08	1.63	2.07	2.21	1.81	1.56
codon49	GG	4.19	3.56	3.05	2.28	1.76	1.38
	AA	1.13	1.50	1.79	1.98	1.79	1.52
	GA	1.12	2.17	3.41	4.66	3.88	2.80
codon389	GG	1.78	1.78	1.78	1.77	1.77	1.77
	CC	1.41	1.87	2.21	2.35	2.09	1.91
	CG	0.91	1.57	2.14	2.33	1.86	1.65
codon492	TT	1.11	1.99	2.69	2.77	2.20	2.33
	CC	1.12	1.97	2.61	2.63	2.05	2.14
	CT	1.03	1.67	2.12	2.12	1.70	1.78

3.6 Discussion

In this paper, we proposed a generalized varying index coefficient modeling procedure to assess the interaction effect of multiple environmental factors as a whole with a genetic factor. The model was motivated by empirical evidence and was developed under an longitudinal design with a binary disease response. We developed a profile estimation procedure to estimate the index coefficients and nonparametric interaction functions iteratively. The estimation was conducted under the QIF framework. To estimate the nonparametric functions, we first approximated the function using truncated power spline basis, then estimated the spline coefficients based on QIF. Furthermore, we proposed a nonparametric hypothesis test to assess the linearity of the nonparametric interaction function. Simulation study has been conducted to illustrate the estimation and testing procedures

to evaluate the finite sample performance. The results indicate reasonable estimation performance of the method under different sample sizes and measurement times.

Our method was proposed to evaluate the joint interaction effect between multiple environmental variables as a whole with genetic variables. Compared to the generalized additive varying coefficient model (gAVCM), which considers the G×E effect for each single environmental factor separately, our model presents two advantages: 1) it is biologically more attractive if there are synergistic effects between multiple exposures; and 2) it can potentially increase the testing power for detecting interactions since it can reduce multiple testing burden by treating multiple exposures as a single index variable. Although our method was motivated by a genetic association study, the developed model and inference procedures can be applied to other disciplines with the purpose to model the synergistic effect of multiple variables as a whole.

We applied our method to a real data set from a pain sensitivity study. Testing results indicate that all of the five SNPs are nonlinear moderated, by the synergistic effect of the three variables with dosage as a “time"-varying variable, to affect the risk of high blood pressure. These five SNPs were genotyped from a candidate gene which has been shown to be related to blood pressure changes (Johnson and Terra, 2002). Although the purpose of the data was not generated to evaluate the genetic effect on “hypertension", we simply applied the method to this data set to demonstrate the utility of the method. The estimated odds of different genotypes for a particular SNP at different dosage levels does give insights into the effect of the SNPs nonlinearly modulated by dosages. Of particular interest is SNP condon49 in which individuals carrying genotype GG show a decreasing risk of developing high blood pressure as the dosage level increases, indicating a protective effect of this genotype. For the same SNP, individuals carrying genotype GA show a different pattern of developing high blood pressure as the dosage level increases. Such a dynamic change of genetic effect over different dosage levels cannot be revealed by a cross-sectional study, indicating the relative merit of a longitudinal design.

CHAPTER 4

DETECTING GENETIC ASSOCIATIONS WITH MULTIVARIATE PARTIALLY LINEAR VARYING-COEFFICIENTS MODELS FOR MULTIPLE LONGITUDINAL TRAITS

4.1 Introduction

Cross-sectional disease traits have been the primary focus in genetic association studies. Given the improved power to identify disease genes with phenotypic data measured over time, longitudinal designs are becoming popular in genetic association studies (Sitlani et al. 2015; Macgregor et al. 2005; Furlotte et al. 2012; Xu et al. 2014). Most statistical methods developed so far focus on a single outcome of interest. When multiple outcomes are measured over time, for example, multiple measures of heart function in a longitudinal study of cardiac function, methods focusing on just a single outcome over time may not provide a complete picture of cardiac function.

In genetics, the phenomenon that a single gene or locus influences more than one trait is known as pleiotropy (Wang et al., 2014; Gratten and Visscher, 2016). Genetic pleiotropy plays a crucial role in many complex diseases. One of the most well-known examples is the phenylketonuria (PKU) disease (Lobo, 2008). The conventional approach to identify genetic pleiotropic effects on multiple traits is to test the association between a gene and each trait individually and then determine whether the genetic effect is significantly associated with more than one trait. The disadvantages of this approach, such as the inflation in the family-wise Type I error and incomplete information in individual tests compared to a combined analysis for multiple traits, have been discussed in some studies (e.g. Wang et al., 2014). Therefore, a joint genetic association test on multiple traits is more desirable to control the family-wise Type I error and enhance the power of tests.

In real life, timing is a very important factor in the development of a disease. Genetic effects on a disease trait vary during the life span of an individual. The function of a gene depends largely on when it turns on and off, which could show a temporal pattern. In order to capture the dynamic

effect of a gene on a disease trait over time, it is natural to model the dynamic effect as a potential nonlinear function over time. Considering multiple longitudinal traits, we proposed the following partially linear varying coefficients model,

$$Y_{lij} = Y_{li}(t_{ij}) = \beta_{0l}(t_{ij}) + \beta_{1l}(t_{ij})G_i + \boldsymbol{\alpha}_l \mathbf{Z}_{ij} + \varepsilon_{lij}, \tag{4.1}$$

where Y_{lij} is the response variable which measures the l-th phenotype on the i-th subject at the j-th time point; $\mathbf{Z}_{ij}$ is a p-dimensional vector of covariates, which can be either dependent or independent of time; G_i denotes the time-invariant genetic variable within subject; $\beta_{0l}(\cdot)$ and $\beta_{1l}(\cdot)$ are unknown functions; and ε_{lij} is an error term which is assumed to following the following joint distribution,

$$\boldsymbol{\varepsilon}_i = \begin{pmatrix} \boldsymbol{\varepsilon}_{1i} \\ \vdots \\ \boldsymbol{\varepsilon}_{Li} \end{pmatrix} \sim N\big(\mathbf{0}, \boldsymbol{\Sigma}\big)$$

with $\boldsymbol{\Sigma}$ be some covariance structure. If we use a time-varying environmental factor X_{ij} instead of t_{ij} in the model, i.e.

$$Y_{lij} = \beta_{0l}(X_{ij}) + \beta_{1l}(X_{ij})G_i + \boldsymbol{\alpha}_l \mathbf{Z}_{ij} + \varepsilon_{lij},$$

then the model can be used for jointly modeling dynamic G×E interactions for multiple longitudinal traits.

Models for multivariate longitudinal traits are necessarily complex, because they must consider different types of correlations for each independent subject: correlation between measurements for the same trait at different time points, correlation between measurements at the same time point on different traits, and correlation between measurements at different time points and on different traits.

Qu and Li (2006) applied the method of quadratic inference functions (QIF) to the varying coefficients models for longitudinal data. One important advantage is that the QIF method only requires correct specification of the mean structure and does not require any likelihood or approximation of the likelihood in hypothesis testing. In addition, when the working correlation structure

is misspecified, the QIF is more efficient than the generalized estimation equation (GEE) approach. Another advantage of QIF approach is that the inference function has an asymptotic form, which provides a model selection criteria similar to AIC and BIC. It also allows us to test whether coefficients are significantly time-varying based on the asymptotic results.

The purpose of this paper is to develop a set of hypothesis testing procedure, including joint testing for multiple traits and marginal testing for each individual trait, for model (4.1). We first use penalized splines (Ruppert and Carroll, 2000) to approximate the nonparametric functions in the model. Then we develop a 2-step testing procedure to first jointly test the genetic effect on multiple traits and then separately test marginal genetic effect on each trait based on the QIF approach. Estimation of the parametric coefficients and nonparametric spline coefficients are obtained under the QIF framework.

This chapter is organized as follows: we state our proposed model in Section 4.2.1, and define the objective function in a QIF method in Section 4.2.2. Estimation procedure and asymptotical properties of estimators are provided in Section 4.2.3. A model selection criteria using BIC is provided in Section 4.2.4. A theorem for goodness-of-fit test in QIF approach is established in Section 4.2.5 and we propose a 2-step testing procedure based on that in Section 4.2.6. We assess the finite sample performance of the proposed procedure with Monte Carlo simulation in Section 4.3 and illustrate the proposed methodology by the analysis of a pain sensitivity data set in Section 4.4. Conclusions and discussion are made in Section 4.5. Proofs are included in Appendix.

4.2 Joint models and statistical methods

4.2.1 Joint multivariate models

In multivariate longitudinal studies, suppose y_{lij} is the l-th continuous outcome collected on the i-th observation at the time point t_{ij}, where $l = 1, \ldots, L$, $i = 1, \ldots, N$, $j = 1, \ldots, n_i$. The joint

partially linear varying coefficient models are defined as

$$y_{lij} = y_{li}(t_{ij}) = \beta_{0l}(t_{ij}) + \beta_{1l}(t_{ij})G_i + \boldsymbol{\alpha}_l \mathbf{Z}(t_{ij}) + \varepsilon_{lij},$$

where G_i is the SNP variable which is not depend on time and type of measurement, $\boldsymbol{Z}(t_{ij})$ is the p-dimensional covariate vector, which can be either time-dependent or time-independent. ε_{lij} is an error term and

$$\boldsymbol{\varepsilon}_i = \begin{pmatrix} \boldsymbol{\varepsilon}_{1i} \\ \vdots \\ \boldsymbol{\varepsilon}_{Li} \end{pmatrix} \sim N\big(\mathbf{0}, \boldsymbol{\Sigma}\big)$$

with $\boldsymbol{\Sigma}$ be some covariance structure. $\beta_{0l}(\cdot)$ and $\beta_{1l}(\cdot)$ are unknown smooth nonparametric functions. To illustrate the idea, in the following we demonstrate the methods assuming L=2. For the situation where there are more than two traits ($L > 2$), the technique can be easily extended. For the case when L=2, the joint models can be written as

$$y_{1ij} = y_{1i}(t_{ij}) = \beta_{01}(t_{ij}) + \beta_{11}(t_{ij})G_i + \boldsymbol{\alpha}_1 \mathbf{Z}(t_{ij}) + \varepsilon_{1ij},$$

$$y_{2ij} = y_{2i}(t_{ij}) = \beta_{02}(t_{ij}) + \beta_{12}(t_{ij})G_i + \boldsymbol{\alpha}_2 \mathbf{Z}(t_{ij}) + \varepsilon_{2ij},$$

where

$$\boldsymbol{\varepsilon}_i = \begin{pmatrix} \boldsymbol{\varepsilon}_{1i} \\ \boldsymbol{\varepsilon}_{2i} \end{pmatrix} \sim N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \sigma_1^2 \boldsymbol{\Sigma}_{11} & \rho_{12}\sigma_1\sigma_2 \boldsymbol{\Sigma}_{12} \\ \rho_{12}\sigma_1\sigma_2 \boldsymbol{\Sigma}_{21} & \sigma_2^2 \boldsymbol{\Sigma}_{22} \end{bmatrix} \right)$$

4.2.2 Objective function based on QIF

To construct the objective function using the QIF approach, we first approximate the unknown functions β_{01}, β_{11}, β_{02} and β_{12} by a q-degree truncated power spline basis, i.e.

$$\beta_{sl}(t) \approx \mathbf{B}_{sl}(t)^T \boldsymbol{\gamma}_{sl}, \text{ for } s = 0,1 \text{ and } l = 1,2, \tag{4.2}$$

where $\mathbf{B}_{sl}(t) = (1, t, t^2, ..., t^{q_{sl}}, (t-\kappa_1)_+^{q_{sl}}, ..., (t-\kappa_{K_{sl}})_+^{q_{sl}})^T$ is a truncated power spline basis with degree q_{sl} and K_{sl} knots $\kappa_1, ..., \kappa_{K_{sl}}$. $\boldsymbol{\gamma}_{sl}$ is a $(q_{sl} + K_{sl} + 1)$-dimensional vector of spline coefficients.

In a GEE approach we solve

$$\sum_{i=1}^{N} \dot{\boldsymbol{\mu}}_i^T \mathbf{V}_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i) = 0, \tag{4.3}$$

where $\mathbf{y}_i = (\mathbf{y}_{1i}^T, \mathbf{y}_{2i}^T)^T$, $\mathbf{y}_{li} = (y_{li1}, ..., y_{lin_i})^T$; $\boldsymbol{\mu}_i = E(\mathbf{y}_i)$ is the mean function and $\dot{\boldsymbol{\mu}}_i$ is the first derivative of $\boldsymbol{\mu}_i$ with respect to the parameters; $\mathbf{V}_i$ is the covariance matrix of $\mathbf{y}_i$ and can be decomposed as $\mathbf{V}_i = \mathbf{A}_i^{1/2}\mathbf{R}(\boldsymbol{\rho})\mathbf{A}_i^{1/2}$ with $\mathbf{A}_i$ being a diagonal matrix of marginal variances and $\mathbf{R}(\boldsymbol{\rho})$ being a correlation matrix with nuisance parameters $\boldsymbol{\rho}$. QIF approach considers the inverse of the correlation matrix $\mathbf{R}$ as a linear combination of several known basis matrices in a form

$$\mathbf{R}^{-1} \approx a_1\mathbf{M}_1 + a_2\mathbf{M}_2 + ... + a_h\mathbf{M}_h, \tag{4.4}$$

where $\mathbf{M}_1$ is the identity matrix and $\mathbf{M}_2, ..., \mathbf{M}_h$ are symmetric basis matrixes. For exchangeable working correlation, $\mathbf{M}_2$ has 0 on the diagonal and 1 elsewhere. If the working correlation is AR(1), we can set $\mathbf{M}_2$ to have 1 on its two subdiagonals and 0 elsewhere. Plugging the expression of $\mathbf{R}^{-1}$ (4.4) into the GEE stated in (4.3), we define the estimation function as

$$\bar{g}_N(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N} g_i(\boldsymbol{\theta}) = \frac{1}{N}\begin{bmatrix} \sum_{i=1}^{N} \dot{\boldsymbol{\mu}}_i^T \mathbf{A}_i^{-1/2}\mathbf{M}_1\mathbf{A}_i^{-1/2}(\mathbf{y}_i - \boldsymbol{\mu}_i) \\ \vdots \\ \sum_{i=1}^{N} \dot{\boldsymbol{\mu}}_i^T \mathbf{A}_i^{-1/2}\mathbf{M}_h\mathbf{A}_i^{-1/2}(\mathbf{y}_i - \boldsymbol{\mu}_i) \end{bmatrix} \tag{4.5}$$

Using the spline approximation, the mean function $\boldsymbol{\mu}_i$ can be written as

$$\boldsymbol{\mu}_i(\boldsymbol{\theta}) = \begin{bmatrix} \boldsymbol{\mu}_{1i}(\boldsymbol{\theta}) \\ \boldsymbol{\mu}_{2i}(\boldsymbol{\theta}) \end{bmatrix} = \begin{bmatrix} \mu_{1i1}(\boldsymbol{\theta}) \\ \vdots \\ \mu_{1in_i}(\boldsymbol{\theta}) \\ \mu_{2i1}(\boldsymbol{\theta}) \\ \vdots \\ \mu_{2in_i}(\boldsymbol{\theta}) \end{bmatrix} = \begin{bmatrix} \mathbf{B}_{01}^T(t_{i1})\boldsymbol{\gamma}_{01} + \mathbf{B}_{11}^T(t_{i1})\boldsymbol{\gamma}_{11}G_i + \boldsymbol{\alpha}_1\mathbf{Z}(t_{i1}) \\ \vdots \\ \mathbf{B}_{01}^T(t_{in_i})\boldsymbol{\gamma}_{01} + \mathbf{B}_{11}^T(t_{in_i})\boldsymbol{\gamma}_{11}G_i + \boldsymbol{\alpha}_1\mathbf{Z}(t_{in_i}) \\ \mathbf{B}_{02}^T(t_{i1})\boldsymbol{\gamma}_{02} + \mathbf{B}_{12}^T(t_{i1})\boldsymbol{\gamma}_{12}G_i + \boldsymbol{\alpha}_2\mathbf{Z}(t_{i1}) \\ \vdots \\ \mathbf{B}_{02}^T(t_{in_i})\boldsymbol{\gamma}_{02} + \mathbf{B}_{12}^T(t_{in_i})\boldsymbol{\gamma}_{12}G_i + \boldsymbol{\alpha}_2\mathbf{Z}(t_{in_i}) \end{bmatrix},$$

and the first derivative of $\boldsymbol{\mu}_i$ is

$$\dot{\boldsymbol{\mu}}_i = \begin{bmatrix} \mathbf{B}_{01}^T(t_{i1}) & \mathbf{B}_{11}^T(t_{i1})G_i & \mathbf{Z}(t_{i1}) & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{B}_{01}^T(t_{in_i}) & \mathbf{B}_{11}^T(t_{in_i})G_i & \mathbf{Z}(t_{in_i}) & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathbf{B}_{02}^T(t_{i1}) & \mathbf{B}_{12}^T(t_{i1})G_i & \mathbf{Z}(t_{i1}) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \mathbf{B}_{02}^T(t_{in_i}) & \mathbf{B}_{12}^T(t_{in_i})G_i & \mathbf{Z}(t_{in_i}) \end{bmatrix},$$

where $\boldsymbol{\theta} = (\boldsymbol{\gamma}_{01}^T, \boldsymbol{\gamma}_{11}^T, \boldsymbol{\alpha}_1^T, \boldsymbol{\gamma}_{02}^T, \boldsymbol{\gamma}_{12}^T, \boldsymbol{\alpha}_2^T)^T$.

Setting each component in (4.5) to be zero will result in more equations than unknown parameters. Following the idea of generalized method of moments (Hansen, 1982), the QIF is defined as

$$Q_N(\boldsymbol{\theta}) = N\bar{g}_N^T \bar{C}_N^{-1} \bar{g}_N, \tag{4.6}$$

where $\bar{C}_N = \frac{1}{N}\sum_{i=1}^N g_i g_i^T$ is a consistent estimator for $\text{var}(g_i)$. Minimizing the objective function (4.6) provides the estimations of parameters.

4.2.3 Estimation

The estimation of the parameters can be obtained through minimizing the objective function, i.e.

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} Q_N(\boldsymbol{\theta}).$$

To avoid over-fitting, we can define a penalized QIF in a form

$$N^{-1}Q_N(\boldsymbol{\theta}) + \lambda \boldsymbol{\theta}^T \mathbf{D} \boldsymbol{\theta}, \tag{4.7}$$

where $\mathbf{D}$ is a diagonal matrix with 1 if the corresponding parameter is the spline coefficient associated with knots and 0 otherwise. Minimizing the penalized QIF provides

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} (N^{-1}Q_N(\boldsymbol{\theta}) + \lambda \boldsymbol{\theta}^T \mathbf{D} \boldsymbol{\theta}). \tag{4.8}$$

To estimate the tuning parameter λ, we can extend the generalized cross-validation (Ruppert, 2002; Qu and Li, 2006; Bai et al., 2009) to the penalized QIF and define the generalized cross-validation statistic as

$$\mathrm{GCV}(\lambda) = \frac{N^{-1}Q_N}{(1-N^{-1}df)^2}$$

with the effective degree of freedom

$$\mathrm{df} = \mathrm{tr}\big[(\ddot{Q}_N + 2N\lambda D)^{-1}\ddot{Q}_N\big],$$

where $\ddot{Q}_N$ is the second derivative of Q_N. The optimized tuning parameter λ is given as

$$\widehat{\lambda} = \arg\min_{\lambda} \mathrm{GCV}(\lambda).$$

To establish the asymptotic properties for the penalized quadratic inference function estimators with fixed knots, we assume $\boldsymbol{\theta}_0$ is the parameter satisfying $E_{\boldsymbol{\theta}_0}(g_i) = 0$. Similar theoretical results are provided in Qu and Li (2006). Following their idea and extend those results to the estimators in our model, we get the strong consistency of the resulting estimators in Theorem 6. The $\sqrt{N}$-consistency and asymptotic normality of the estimators are given in Theorem 7 .

Theorem 6 *Suppose conditions (B1)-(B6) in the Appendix hold and the smoothing parameter $\lambda_N = o(1)$, then the estimator $\widehat{\boldsymbol{\theta}}$, which is obtained by minimizing the penalized quadratic function in (4.7), exists and converges to $\boldsymbol{\theta}_0$ almost surely.*

Theorem 7 *Suppose conditions (B1)-(B6) in the Appendix hold and the smoothing parameter $\lambda_N = o(N^{-1/2})$, then the estimator $\widehat{\boldsymbol{\theta}}$ obtained by minimizing the penalized quadratic function in (4.7) is asymptotically normally distributed with the limiting distribution,*

$$\sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N(\boldsymbol{0}, (\boldsymbol{G}_0^T \boldsymbol{C}_0^{-1} \boldsymbol{G}_0)^{-1}),$$

where the calculation of $\boldsymbol{G}_0$ and $\boldsymbol{C}_0$ can be found in Appendix.

4.2.4 Model selection

In contrast to the complicated model selection in GEE method due to the lack of an objective function in its estimation procedure, it is natural to extend the BIC method to the QIF approach since the QIF objective function and twice the negative log-likelihood function have similar asymptotic properties. Under the assumption $E(g) = 0$ and the number of estimating equations is larger than the number of parameters, we have $Q(\widehat{\boldsymbol{\theta}}) \xrightarrow{d} \chi^2_{r-k}$ (Hansen, 1982), where r is the dimension of $\bar{g}_N(\boldsymbol{\theta})$, k is the dimension of $\boldsymbol{\theta}$, $\widehat{\boldsymbol{\theta}}$ is the estimator by minimizing the QIF when certain order and number of knots are chosen. Based on the asymptotic property of QIF, the BIC criterion for a model with r estimating equations and k parameters is

$$Q(\widehat{\boldsymbol{\theta}}) + (r-k)\ln N.$$

The model with minimum BIC would be considered optimal.

4.2.5 Nonparametric goodness-of-fit test

Compared to GEE, an advantage of QIF approach is that QIF provides a goodness-of-fit test without estimations for second moment parameters. In Model (4.1), it is of interest to test whether the spline approximations for the varying coefficient functions are appropriate.

Qu et al. (2000) constructed a test statistic based on QIF. Suppose that the d-dimension parameter vector $\boldsymbol{\gamma}$ is partitioned into $(\boldsymbol{\psi}, \boldsymbol{\zeta})$, where $\boldsymbol{\psi}$ is the parameter of interest with dimension d_1, and $\boldsymbol{\zeta}$ is a nuisance parameter with dimension $d_2 = d - d_1$. If we are interested in testing

$$H_0 : \boldsymbol{\psi} = \boldsymbol{\psi}_0,$$

the test statistic

$$Q(\boldsymbol{\psi}_0, \widetilde{\boldsymbol{\zeta}}) - Q(\widehat{\boldsymbol{\psi}}, \widehat{\boldsymbol{\zeta}})$$

follows an asymptotically chi-square distribution with d_1 degrees of freedom.

Theorem 8 *(Qu et al., 2000) Suppose that all required regularity conditions are satisfied and* $\boldsymbol{\psi}$ *has dimension* d_1*. Under the null hypothesis,* $Q_N(\boldsymbol{\psi}_0, \widetilde{\boldsymbol{\zeta}}) - Q_N(\widehat{\boldsymbol{\psi}}, \widehat{\boldsymbol{\zeta}})$ *is asymptotically chi-square distribution with* d_1 *degrees of freedom, where*

$$\widetilde{\boldsymbol{\zeta}} = \arg\min Q_N(\boldsymbol{\psi}_0, \boldsymbol{\zeta}), \quad (\widehat{\boldsymbol{\psi}}, \widehat{\boldsymbol{\zeta}}) = \arg\min Q_N(\boldsymbol{\psi}, \boldsymbol{\zeta}). \tag{4.9}$$

4.2.6 Two-step hypothesis testing procedure

In Model (4.1), it is of interest to test whether the genetic effects on multiple traits are significant or not. Based on Theorem 8, we develop a 2-step testing procedure for testing the significance of the varying coefficient functions. In the first step, the joint test is performed to see whether a genetic factor has significant effect on at least one longitudinal trait. If the testing result in the first step is significant, we will further conduct the marginal tests in the second step to assess if the genetic effect is significant on both traits or just one trait. So the first step is a joint test of significance followed by a marginal test to assess individual significance.

4.2.6.1 Step 1: Joint test

First, we are interested in testing whether the genetic factor G has effect on at least one longitudinal trait. The hypothesis is stated as

$$H_0 : \beta_{11}(\cdot) = \beta_{12}(\cdot) = 0 \text{ v.s. } H_1 : \beta_{11}(\cdot) \neq 0 \text{ or } \beta_{12}(\cdot) \neq 0.$$

This can be handled through the truncated power spline approximation of the nonparametric functions stated in (4.2). In particular, test this hypothesis is equivalent to test the following null hypothesis

$$H_0 : \boldsymbol{\gamma}_{11} = \boldsymbol{\gamma}_{12} = \mathbf{0}.$$

According to Theorem 8, we can construct a test statistic

$$T_N = Q_N(\widetilde{\boldsymbol{\theta}}) - Q_N(\widehat{\boldsymbol{\theta}}),$$

where

$$\widetilde{\boldsymbol{\theta}} = \underset{\boldsymbol{\gamma}_{11}=\boldsymbol{\gamma}_{12}=\mathbf{0}}{\arg\min}\, Q_N\{\boldsymbol{\gamma}_{01},\boldsymbol{\gamma}_{11},\boldsymbol{\alpha}_1,\boldsymbol{\gamma}_{02},\boldsymbol{\gamma}_{12},\boldsymbol{\alpha}_2 \mid \mathbf{y}_1,\mathbf{y}_2,\mathbf{G},\mathbf{Z}\},$$

and

$$\widehat{\boldsymbol{\theta}} = \arg\min Q_N\{\boldsymbol{\gamma}_{01},\boldsymbol{\gamma}_{11},\boldsymbol{\alpha}_1,\boldsymbol{\gamma}_{02},\boldsymbol{\gamma}_{12},\boldsymbol{\alpha}_2 \mid \mathbf{y}_1,\mathbf{y}_2,\mathbf{G},\mathbf{Z}\}.$$

The test statistic T_N has an asymptotical χ^2 distribution with degree of freedom equals the number of constraints under H_0, according to Theorem 8.

4.2.6.2 Step 2: Marginal tests

From the joint test, if there exist a significant genetic effect on at least one longitudinal trait, then we can further test the marginal effects:

$$H_0 : \beta_{1l}(\cdot) = 0 \text{ v.s. } H_1 : \beta_{1l}(\cdot) \neq 0, \;\; l = 1,2.$$

Based on (4.2), this is equivalent to test the following two hypotheses

$$H_0 : \boldsymbol{\gamma}_{11} = \mathbf{0}$$

and

$$H_0 : \boldsymbol{\gamma}_{12} = \mathbf{0}$$

separately.

For testing $H_0 : \boldsymbol{\gamma}_{11} = \mathbf{0}$, we use test statistic

$$T_{N1} = Q_N(\widetilde{\boldsymbol{\gamma}}_{01},\mathbf{0},\widetilde{\boldsymbol{\alpha}}_1) - Q_N(\widehat{\boldsymbol{\gamma}}_{01},\widehat{\boldsymbol{\gamma}}_{11},\widehat{\boldsymbol{\alpha}}_1),$$

where

$$(\widetilde{\boldsymbol{\gamma}}_{01},\mathbf{0},\widetilde{\boldsymbol{\alpha}}_1) = \underset{\boldsymbol{\gamma}_{11}=\mathbf{0}}{\arg\min}\, Q_N\{\boldsymbol{\gamma}_{01},\boldsymbol{\gamma}_{11},\boldsymbol{\alpha}_1 \mid \mathbf{y}_1,\mathbf{G},\mathbf{Z}\},$$

and

$$(\widehat{\boldsymbol{\gamma}}_{01},\widehat{\boldsymbol{\gamma}}_{11},\widehat{\boldsymbol{\alpha}}_1) = \arg\min Q_N\{\boldsymbol{\gamma}_{01},\boldsymbol{\gamma}_{11},\boldsymbol{\alpha}_1 \mid \mathbf{y}_1,\mathbf{G},\mathbf{Z}\}.$$

We can also construct another test statistic

$$T_{N2} = Q_N(\widetilde{\boldsymbol{\gamma}}_{02}, \mathbf{0}, \widetilde{\boldsymbol{\alpha}}_2) - Q_N(\widehat{\boldsymbol{\gamma}}_{02}, \widehat{\boldsymbol{\gamma}}_{12}, \widehat{\boldsymbol{\alpha}}_2)$$

for testing $H_0 : \boldsymbol{\gamma}_{12} = \mathbf{0}$, where

$$(\widetilde{\boldsymbol{\gamma}}_{02}, \mathbf{0}, \widetilde{\boldsymbol{\alpha}}_2) = \underset{\boldsymbol{\gamma}_{12}=\mathbf{0}}{\arg\min}\, Q_N\{\boldsymbol{\gamma}_{02}, \boldsymbol{\gamma}_{12}, \boldsymbol{\alpha}_2 \mid \mathbf{y}_2, \mathbf{G}, \mathbf{Z}\},$$

and

$$(\widehat{\boldsymbol{\gamma}}_{02}, \widehat{\boldsymbol{\gamma}}_{12}, \widehat{\boldsymbol{\alpha}}_2) = \arg\min Q_N\{\boldsymbol{\gamma}_{01}, \boldsymbol{\gamma}_{12}, \boldsymbol{\alpha}_2 \mid \mathbf{y}_2, \mathbf{G}, \mathbf{Z}\}.$$

The asymptotical distribution of test statistics T_{N1} and T_{N2} can be obtained from Theorem 8.

4.3 Simulation study

4.3.1 Simulation setup

In this section, the finite sample performance of the proposed method is evaluated through Monte Carlo simulation studies. The two continuous variables are generated from the models

$$y_{1i}(t_{ij}) = \beta_{01}(t_{ij}) + \beta_{11}(t_{ij})G_i + \alpha_1 z(t_{ij}) + \varepsilon_{1ij},$$

$$y_{2i}(t_{ij}) = \beta_{02}(t_{ij}) + \beta_{12}(t_{ij})G_i + \alpha_2 z(t_{ij}) + \varepsilon_{2ij},$$

where $\beta_{01}(t_{ij}) = 0.5\cos(2\pi t_{ij})$, $\beta_{11} = \sin(\pi(t_{ij} - 0.2))$, $\beta_{02}(t_{ij}) = \sin(\pi t_{ij}) - 0.5$, $\beta_{12}(t_{ij}) = \cos(\pi t_{ij} - 0.8)$, $\alpha_1 = 0.2$ and $\alpha_2 = 0.3$. We generate T time points $\mathbf{t}_i = (t_{i1}, \ldots, t_{iT})$ from a uniform distribution $U(0,1)$. The predictor variable $z(t_{ij})$ is also generated from $U(0,1)$. We set the minor allele frequency (MAF) as p_A = 0.5 and assume Hardy-Weinberg equilibrium. Three different SNP genotypes AA, Aa and aa are simulated from a multinomial distribution with frequencies p_A^2, $2p_A(1-p_A)$ and $(1-p_A)^2$, respectively. Variable G takes value in the set {0,1,2}, corresponding to genotypes $\{aa, Aa, AA\}$. We assume ε_{1ij} and ε_{2ij} are jointly normally distributed with the correlation $\text{corr}(\varepsilon_{1ij}, \varepsilon_{2ij}) = 0.5$. Then we generate the error terms from a multivariate

normal distribution

$$\begin{pmatrix} \boldsymbol{\varepsilon}_{1i} \\ \boldsymbol{\varepsilon}_{2i} \end{pmatrix} \sim N \left[\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \sigma_1^2 \Sigma_{11} & 0.5\sigma_1\sigma_2\Sigma_{12} \\ 0.5\sigma_1\sigma_2\Sigma_{12} & \sigma_2^2\Sigma_{22} \end{pmatrix} \right]$$

We set the marginal variances $\sigma_1^2 = \sigma_2^2 = 0.1$. The true correlation structure of Σ_{11}, Σ_{22}, and Σ_{12} are all exchangeable with $\rho_1 = 0.5$, $\rho_2 = 0.5$ and $\rho_{12} = 0.2$, respectively.

We draw 1000 data sets with sample size $N = 200, 500$ and time points $n_i = T = 10$, in order to compare the performances of our proposed method under different sample sizes. We set $\mathbf{M}_1$ to be the identity matrix and $\mathbf{M}_2$ to has 1 on subdiagonals and 0 elsewhere. The order and number of knots of the splines are chosen by the BIC method.

4.3.2 Performance of estimation

Table 4.1 summarizes the parameter estimation for unknown coefficients. In this table, the average bias (Bias), the standard deviation of the 1000 estimates (SD), the average of the estimated standard error (SE) based on the theoretical results, and the estimated coverage probability (CP) at 95% confidence level are reported. In general, the biases for all parameter estimations are very small, the coverage probabilities are very close to the confidence level 95%, which indicate good performance of our proposed estimation procedure. As the sample size increases, the performance of the estimation improves by showing smaller bias, SD and SE.

Table 4.1 Estimation results for parameters α_1 and α_2 with sample size $N = 200, 500$.

N	Parameter	True	Bias	SD	SE	CP
200	α_1	0.2	0.0004	0.018	0.018	94.6
	α_2	0.3	0.0006	0.018	0.018	94.2
500	α_1	0.2	-5.2E-05	0.012	0.012	95.2
	α_2	0.3	-0.0003	0.012	0.012	94.8

The plots for the estimations of nonparametric functions $\beta_{01}(\cdot)$ and $\beta_{11}(\cdot)$ under different sample sizes are shown in Figure 4.1. The estimated and true functions are denoted by the solid and dashed lines, respectively. The 95% confidence bands are denoted by the dotted-dash line. The

estimated curves almost overlap with the corresponding true curves as shown in the plots, indicating good estimate of the function. Larger sample size leads to tighter confidence bands. Figure 4.2 displays the estimations for functions $\beta_{02}(\cdot)$ and $\beta_{12}(\cdot)$, which are included in the model corresponding to response variable y_2. The results are similar to those in Figure 4.1 and further demonstrate the good performance of our estimation methods.

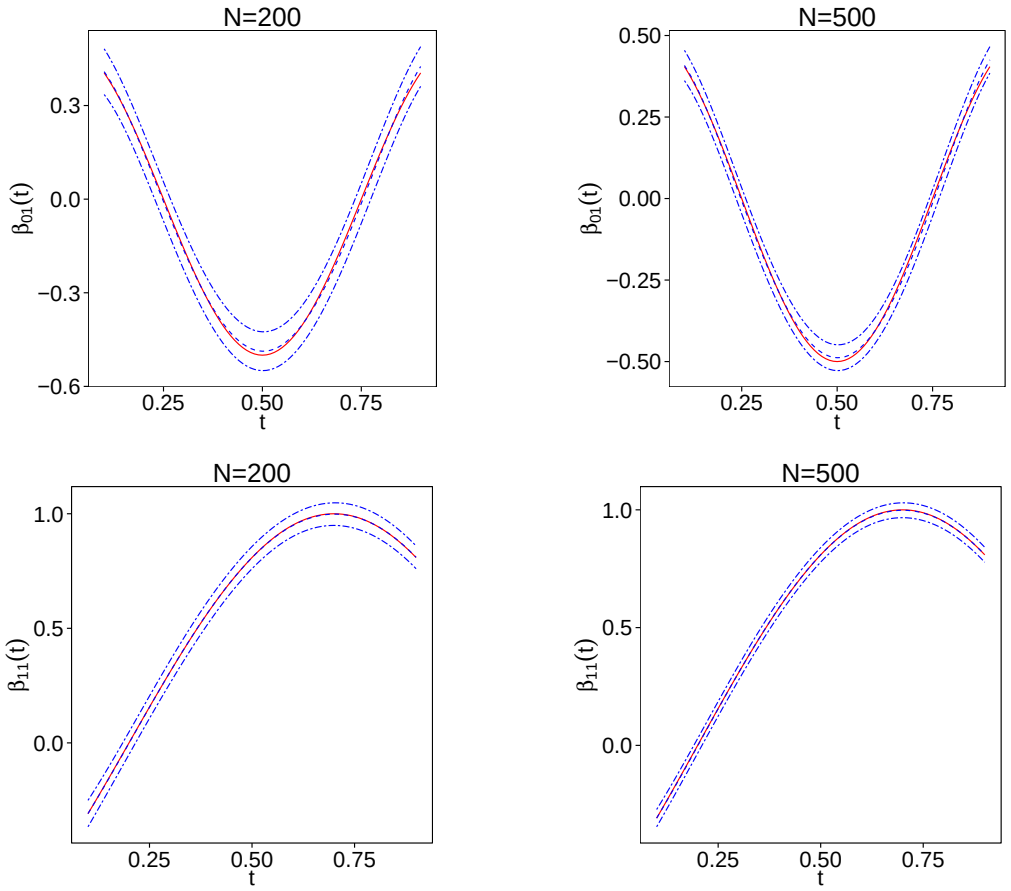

Figure 4.1 The estimation of nonparametric functions $\beta_{01}(\cdot)$ and $\beta_{11}(\cdot)$ when N=200, 500. The estimated and true functions are denoted by the solid and dashed lines respectively. The 95% confidence bands are denoted by the dotted-dash line.

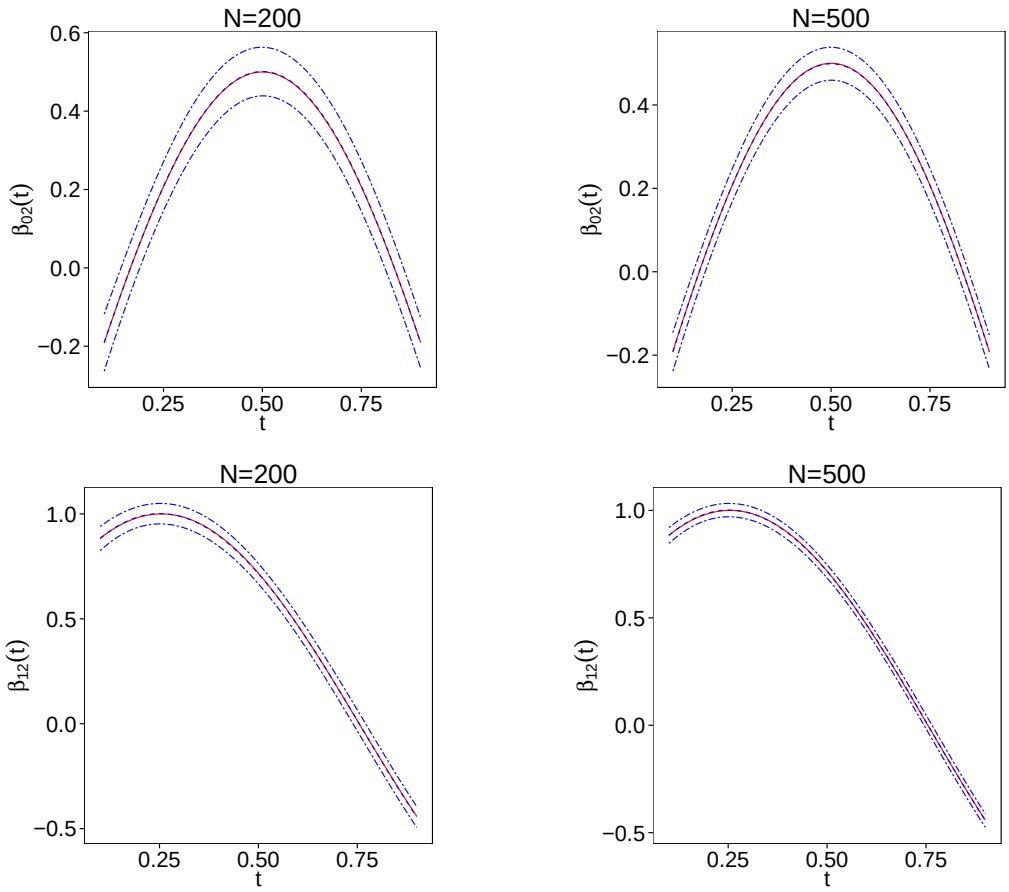

Figure 4.2 The estimation of nonparametric functions $\beta_{02}(\cdot)$ and $\beta_{12}(\cdot)$ when N=200, 500. The estimated and true functions are denoted by the solid and dashed lines respectively. The 95% confidence bands are denoted by the dotted-dash line.

4.3.3 Performance of hypothesis tests

4.3.3.1 Performance for joint test

We evaluate the performance of the joint test under the null hypothesis $H_0 : \beta_{11}(\cdot) = \beta_{12}(\cdot) = 0$. Power is evaluated under a sequence of alternative models with different values of τ, which is denoted by $H_1^\tau : \beta_{11}^\tau(\cdot) = \tau\beta_{11}(\cdot)$ and $\beta_{12}^\tau(\cdot) = \tau\beta_{12}(\cdot)$.

Figure 4.3 shows the empirical size (when $\tau = 0$) and power function (when $\tau > 0$) at significance level 0.05. We obtain 1000 Monte Carlo simulations to assess the null distribution of test statistic under sample sizes N = 200, 500. The empirical Type I error under both sample sizes are

close to the nominal level 0.05 and the power increases dramatically when τ increases from 0 to 0.05. To see the effect of sample size, we compare the performances under N=200 and N=500. As expected, the Type I error is closer to 0.05 and the power increase faster for larger sample size when N=500. For a relatively small sample size N=200, the Type I error is a little inflated and the power function increases slower compared to the case with $N = 500$. Overall, the results indicate good performance of the proposed joint testing procedure.

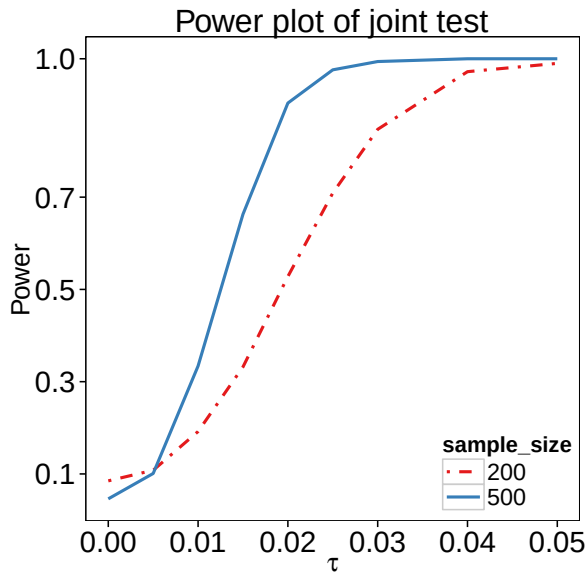

Figure 4.3 The power plot for the joint test under different sample sizes N=200, 500 when T=10.

4.3.3.2 Performance for marginal tests

The performance of the marginal tests for the nonparametric functions corresponding to different traits is evaluated through simulations. Two null hypotheses $H_0 : \beta_{11}(\cdot) = 0$ and $H_0 : \beta_{12}(\cdot) = 0$ were considered separately. For each test, power is evaluated under a sequence of alternative models, denoted by $H_a^\tau : \beta_{1l}^\tau(\cdot) = \tau\beta_{1l}(\cdot)$, $l = 1,2$, correspondingly.

Figure 4.4 displays the power for both marginal tests under different sample sizes N=200 and 500.The empirical Type I error under both sample sizes are very close to the nominal level 0.05 and the power increases dramatically when τ increases from 0 to 0.05. It is obvious that the power

increases more rapidly with larger sample size, while the overall performances under N=500 and N=200 are close, which indicates that our method performances well and does not require very large sample size.

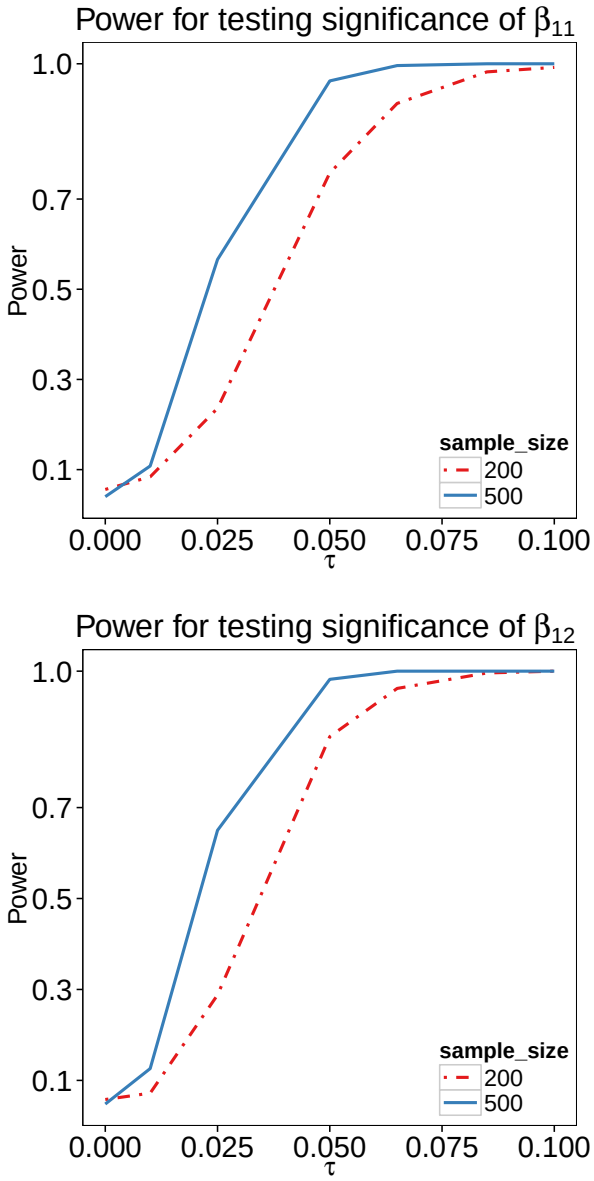

Figure 4.4 The power plots for the marginal test for N=200, 500 and T=10.

In summary, the simulation results indicate that our proposed estimation method works well. The test results indicate that the asymptotic χ^2 distribution for the proposed joint and marginal test works reasonably well under a finite sample size.

4.4 Real data application

We applied the proposed multivariate partially linear varying coefficients model and the two-step hypothesis testing procedure to a data set from a study examining the association of the A118G SNP of OPRM1 to experimental pain sensitivity. A sample of 163 healthy volunteers were evolved in this study. For each subject, Systolic Blood Pressure (SBP) and Diastolic Blood Pressure (DBP) were measured at 6 dosage levels of Dobutamine to assess the genetic effect on drug response. The 6 dosage levels are: 0 (baseline), 5, 10, 20, 30 and 40mcg/min. We treat the dosage levels as time in this analysis.

We consider the partially linear varying coefficient model with two longitudinal traits, with the form

$$Y_{ij}^{SBP} = \beta_0^{SBP}(X_{ij}) + \beta_1^{SBP}(X_{ij})G_i + \alpha^{SBP}Z_i + \varepsilon_{ij}^{SBP},$$
$$Y_{ij}^{DBP} = \beta_0^{DBP}(X_{ij}) + \beta_1^{DBP}(X_{ij})G_i + \alpha^{DBP}Z_i + \varepsilon_{ij}^{DBP}.$$

The two longitudinal traits are SBP and DBP. One time-invariant covariate $Z =$ age are included in the model. Five SNPs *codon16*, *codon27*, *codon49*, *codon389*, *codon492* are considered.

Table 4.2 displays the results of the joint and marginal testing for responses SBP and DBP, respectively. We note that *condon49* and *condon389* show significant result with p-values smaller than the significance level 0.05. Further marginal tests tell us that SNP *condon49* has significant association with SBP but not DBP, while SNP *condon389* has significant association with DBP but not SBP. The results indicate no pleiotropic effect of the two SNPs. In addition, we note that for SNP *condon16*, the joint test does not show significant result but the p-value of the marginal test for SBP is smaller than the significance level. If we choose $\alpha = 0.1$ significance level, then the SNP shows a potential pleiotropic effect on the two traits.

Table 4.2 List of SNPs with MAF, allele, p-values under the joint and marginal testing for SBP and DBP.

			p-values under different null hypotheses				
SNP ID	MAF	Alleles	$\beta_1^{SBP} = \beta_1^{DBP} = 0$	$\beta_1^{SBP} = 0$	$\beta_1^{DBP} = 0$	β_1^{SBP} is linear	β_1^{DBP} is linear
codon16	0.3990	A/G	0.0866	0.0428	0.0514	0.8311	0.3094
codon27	0.4160	G/C	0.3048	0.3819	0.1018	0.5938	0.1012
codon49	0.1387	G/A	0.0229	0.0410	0.8349	0.7846	0.4439
codon389	0.3045	G/C	0.0343	0.3550	0.0242	0.2150	0.2938
codon492	0.4250	T/C	0.3234	0.5779	0.6957	0.2611	0.7875

Table 4.3 shows the estimation results for the age effect α_1 and α_2. The results show that age does not have significant contribution to the two blood traits in this data set.

Table 4.3 List of SNPs with MAF, allele, estimation of coefficients, p-values of significance for coefficients corresponding to SBP and DBP, respectively.

					p-value	
SNP ID	MAF	Alleles	$\widehat{\alpha}^{SBP}$	$\widehat{\alpha}^{DBP}$	$H_0 : \alpha^{SBP} = 0$	$H_0 : \alpha^{DBP} = 0$
codon16	0.3990	A/G	0.0109	-0.0435	0.8562	0.5005
codon27	0.4160	G/C	0.0235	-0.0365	0.7078	0.5902
codon49	0.1387	G/A	0.0356	-0.0336	0.5637	0.6079
codon389	0.3045	G/C	0.0103	-0.0644	0.8743	0.3409
codon492	0.4250	T/C	0.0213	-0.0494	0.7366	0.4679

Figure 4.5 illustrates the estimated shapes for nonparametric coefficient functions for different SNPs for trait SBP. The 95% confidence bands cover the zero line for SNPs condon27, condon389 and condon492, which agrees with the testing results that these SNPs do not show significance. For SNP condon49 which shows significance at $\alpha = 0.05$ level, the estimated increasing effect function as dosage increases suggests that this SNP positively responds to dosage increase to affect SBP.

Figure 4.6 illustrates the estimated shapes for nonparametric coefficient functions for different SNPs for trait DBP. Again, we observe that the 95% confidence bands cover the zero line for SNPs condon27, condon49 and condon429 which agrees with the testing results that these SNPs are not significant. For SNP condon389, the estimated effect function shows a marginal significance, indicating the role of this SNP to drug response on DBP.

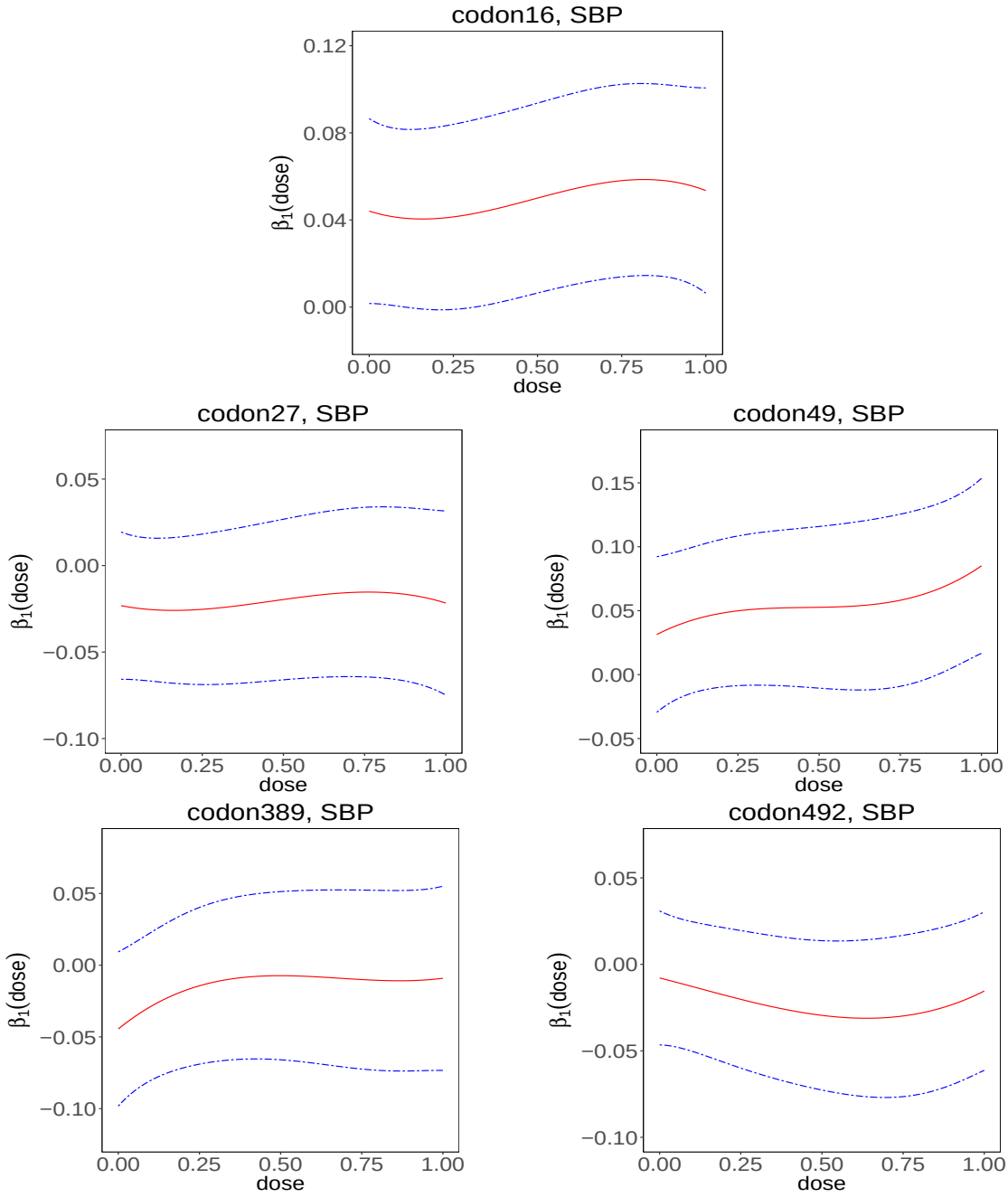

Figure 4.5 Plot of the estimate (solid curve) of the nonparametric function $\beta_1^{SBP}(\cdot)$ for SNPs *codon16*, *codon27*, *codon49*, *codon389* and *codon492*. The 95% confidence bands are denoted by the dashed line.

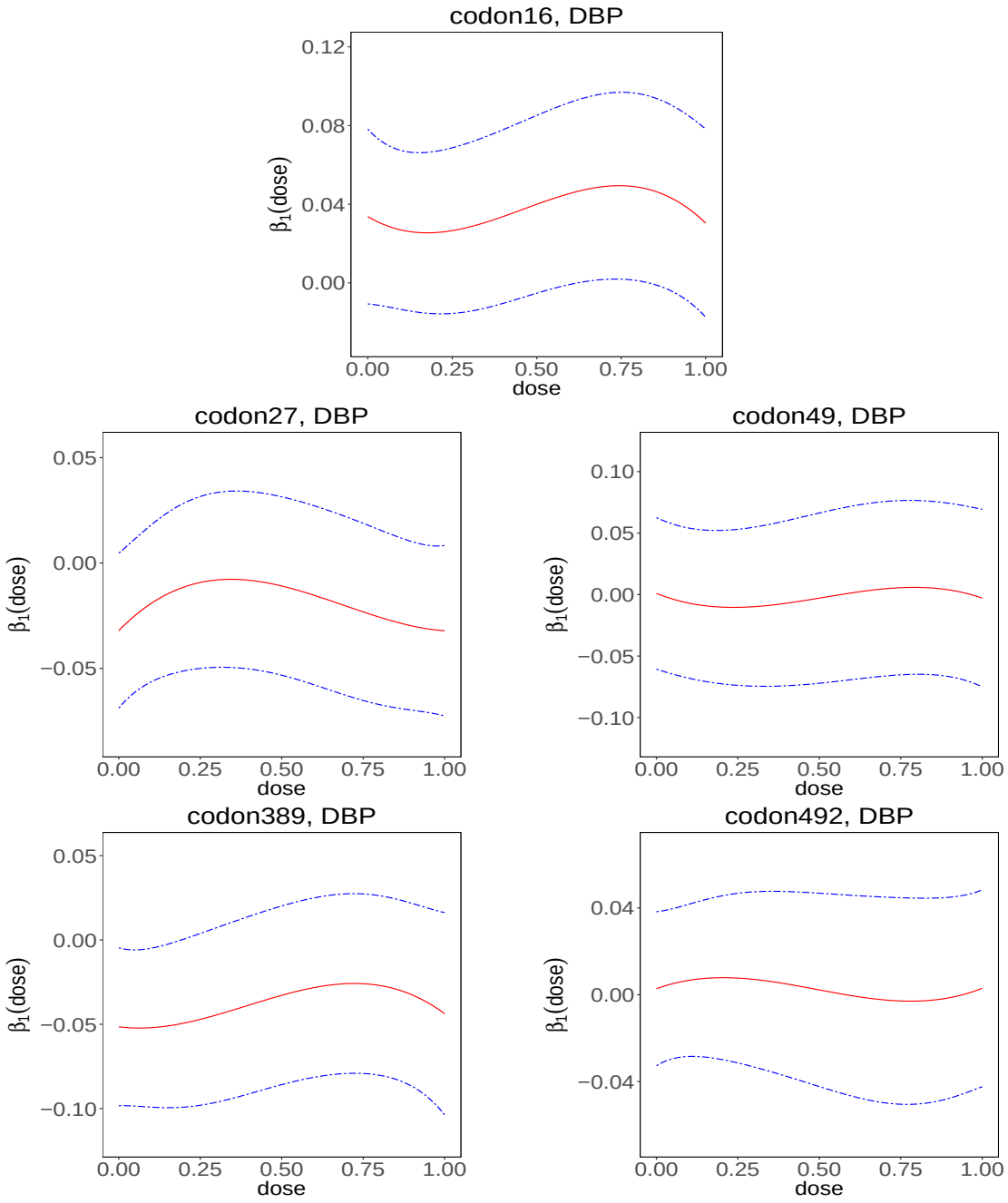

Figure 4.6 Plot of the estimate (solid curve) of the nonparametric function $\beta_1^{DBP}(\cdot)$ for SNPs *codon16*, *codon27*, *codon49*, *codon389* and *codon492*. The 95% confidence bands are denoted by the dashed line.

4.5 Discussion

Identification of genetic pleiotropy effects has been an important task in genetic association studies. If one gene is associated with multiple traits, special attention should be paid to such genes

when designing drug target on those genes. In this paper, we propose a joint multivariate varying coefficient modeling procedure to accommodate correlated longitudinal traits and propose a testing procedure to find the dynamic genetic association between SNPs and multiple traits. Both simulation and real data analysis demonstrate the utility of the proposed method.

One difficulty in jointly modeling multiple longitudinal traits is to model the complex correlation structure. For each subject, we should consider correlation between measurements for the same trait at different time points, correlation between measurements at the same time point on different traits, and correlation between measurements at different time points and on different traits. We implement the QIF approach in estimation and testing procedures. There are several advantages for QIF approach. First, the QIF approach only requires correct specification of the mean structure and does not require any joint likelihood in hypothesis testing. Second, it avoids estimating the nuisance correlation structure parameters by assuming that the inverse of working correlation matrix can be approximated by a linear combination of several known basis matrices. Third, when the working correlation structure is misspecified, the QIF is more efficient than the GEE approach. Forth, the inference function of the QIF approach has an explicit asymptotic form, which provides a model selection criteria and allows us to test whether coefficients are significant or time varying based on the asymptotic results.

Our method was demonstrated with the $L = 2$ case. The proposed method can be extended to multiple longitudinal traits with $L > 2$, although the computational cost might increase. In the real application, we investigate association of SNPs in a candidate gene with two longitudinal traits SBP and DBP. Although the data were not longitudinal in terms of time measurement, the increasing dosage levels can be treated in a time scale. So we can apply the proposed method. The results indicate a weakly pleiotropic effect for SNP condon16 to affect both SBP and DBP. As the application shows, our method is not restricted to a longitudinal study. It also applies to other studies where a certain trait can be measured in a linear scale. Therefore, our method is directly applicable to neuro-genetics studies in which the purpose is to identify SNPs associated with spatial distribution of neuroimages in brain.

CHAPTER 5

CONCLUSION AND FUTURE WORK

5.1 Conclusion

The originalities and contributions of our work can be summarized in two respects. From the respect of statistical methodology, we propose a functional varying index coefficient model (FVICM) to capture the nonlinear G×E interactions under a longitudinal design. In the development of the estimation procedure, penalized spline method is implemented to approximate the nonparametric unctions in the model. A profile estimation procedure is proposed to estimate two sets of parameters: the index loading coefficients and spline coefficients. Then the quadratic inference function approach for analysing longitudinal data is extended to estimate the index loading coefficients and spline coefficients in the profile estimation method. To test the linearity of nonparametric G×E interaction function, we apply the pseudo-likelihood ratio test using a linear mixed model representation of our proposed model. Consistency and asymptotic normality of the estimators are established.

To deal with binary longitudinal traits, it is a natural extension of the FVICM to a generalized functional varying index coefficient model (gFVICM) for investigating nonlinear G×E interactions. We modify the profile estimation procedure with QIF approach to the gFVICM and establish theoretical results of the estimators. Then we proposed a testing procedure based on the asymptotic χ^2 distribution of the objective function in QIF approach. The testing procedure can be used to assess the linearity of the interaction function.

For some complex diseases, there are multiple phenotypes that can be used to quantify the risk of diseases and sometimes they have shared genetic determinations and this phenomenon is termed genetic pleiotropy. A joint modeling for multiple longitudinal traits using varying index coefficient model is proposed in our work to deal with correlated longitudinal traits in G×E interaction problems. One difficulty of the joint model is the specification of the complicated correlation structure.

The important advantage of QIF approach is that a misspecified working correlation does not affect the consistency of the regression parameter estimation, and the QIF provides a robust sandwich estimator for the variance of the regression parameter estimator. When the working correlation structure is misspecified, the QIF is more efficient than the GEE.

From the application perspective, the varying index coefficient modeling is a powerful tool when we consider the joint effect of environmental mixtures and how they interact with genes to affect disease risk. Our methods are well motivated by epidemiological studies with the hope to identify any synergistic G$\times$E interaction effects. Real data application shows that, compared to the additive varying coefficient model, which consider the G$\times$E for each single environmental factor separately, our models outperform in detecting the significant interaction effect since it can reduce multiple testing burden by treating the serval environmental variables as a single index variable. Also, the assumption of a nonparametric interaction is flexible for possible nonlinear interactions in practice. We also provide a framework to assess the simultaneous genetic effect on multiple phenotypes with longitudinal data.

5.2 Future work

In the future, we plan to extend the functional varying index coefficient model to joint modeling of binary and continuous longitudinal traits. This is practically important for some diseases. For example, over-weighted people will have a higher chance to develop hypertension. Both obese and hypertension might share some common genetic determinants. Jointly modeling the binary hypertension and continuous body weight or BMI could shed novel insights into the genetic etiology of the disease. The main difficulty of joint modeling for binary and continuous longitudinal traits is the lack of a joint distribution. To overcome this difficulty, many researchers introduce a continuous latent variable underlying the binary response and assuming a joint normal distribution for the latent variable and the continuous variable. Catalano and Ryan (1992) suggested to decompose the joint distribution into two components: a marginal distribution for the continuous response and a

conditional distribution for the binary distribution given the continuous response. The first component can be easily modeled. The conditional distribution for binary response can be modeled using the underlying latent continuous variable through a probit link function. Kürüm et al. (2016) proposed the time-varying coefficient models for joint modeling binary and continuous longitudinal responses based on the above idea. However, they only focus on the estimation part and did not provide a testing method for the nonparametric functions in the model. Motivated by their work, we can extend their method to varying index coefficient models for joint modeling binary and continuous longitudinal traits and develop a testing method for joint testing of the significance or linearity of the interaction functions. This will be investigated in our future work.

APPENDIX

Regularity conditions

To establish the asymptotic properties for the estimator of $\boldsymbol{\theta}$, we need the following regularity conditions.

(A1) $\{n_i\}$ is a bounded sequence of integers.

(A2) The parameter space Ω is compact and $\boldsymbol{\theta}_0^*$ is an interior point of Ω.

(A3) The parameter $\boldsymbol{\theta}^*$ is identified, that is, there is a unique $\boldsymbol{\theta}_0^* \in \Omega$ such that the mean zero model assumption $E[g(\boldsymbol{\theta}_0^*)] = 0$.

(A4) $E[g(\boldsymbol{\theta})]$ is continuous in $\boldsymbol{\theta}$.

(A5) $\bar{C}_N(\widehat{\boldsymbol{\theta}}^*) = \frac{1}{N}\sum_{i=1}^{N} g_i(\widehat{\boldsymbol{\theta}}^*)g_i(\widehat{\boldsymbol{\theta}}^*)^T$ converges almost surely to $\mathbf{C}_0$, which is a constant and invertible matrix.

(A6) The first derivative of $\bar{g}_N$ exists and is continuous. $\frac{\partial \bar{g}_N}{\partial \boldsymbol{\theta}^*}(\widehat{\boldsymbol{\theta}}^*)$ converges in probability to $\mathbf{G}_0$ if $\widehat{\boldsymbol{\theta}}^*$ converges in probability to $\boldsymbol{\theta}_0^*$.

(B1) $\{n_i\}$ is a bounded sequence of integers.

(B2) The parameter space $\Omega_{\boldsymbol{\theta}}$ is compact and $\boldsymbol{\theta}_0$ is an interior point of $\Omega_{\boldsymbol{\theta}}$.

(B3) The parameter $\boldsymbol{\theta}$ is identified, that is, there is a unique $\boldsymbol{\theta}_0 \in \Omega_{\boldsymbol{\theta}}$ such that the first moment assumption $E[g_i(\boldsymbol{\theta}_0)] = 0$ holds for $i = 1, ..., N$, and $E[g_i(\boldsymbol{\theta})]$ is continuous.

(B4) $E[g(\boldsymbol{\theta})]$ is continuous in $\boldsymbol{\theta}$.

(B5) $\bar{C}_N(\widehat{\boldsymbol{\theta}}) = \frac{1}{N}\sum_{i=1}^{N} g_i(\widehat{\boldsymbol{\theta}})g_i(\widehat{\boldsymbol{\theta}})^T$ converges almost surely to $\mathbf{C}_0$, which is a constant and invertible matrix.

(B6) The first derivative of $\bar{g}_N$ exists and is continuous. $\frac{\partial \bar{g}_N}{\partial \boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}})$ converges in probability to $\mathbf{G}_0$ if $\widehat{\boldsymbol{\theta}}$ converges in probability to $\boldsymbol{\theta}_0$.

Proof of Theorem 1:

If we can prove that $\widehat{\boldsymbol{\theta}}^*$ exist and converges to $\boldsymbol{\theta}_0^*$ almost surely, then we can prove the consistency of $\boldsymbol{\theta}$ directly. $\widehat{\boldsymbol{\theta}}^* = \arg\min(N^{-1}Q_N(\boldsymbol{\theta}^*) + \lambda\boldsymbol{\theta}^{*T}\mathbf{D}\boldsymbol{\theta}^*)$ exists because (2.4) has zero as a lower bound and the global minimum exists. To prove the consistency, first, the estimator $\widehat{\boldsymbol{\theta}}^*$ is obtained by minimizing $N^{-1}Q_N(\boldsymbol{\theta}^*) + \lambda\boldsymbol{\theta}^{*T}\mathbf{D}\boldsymbol{\theta}^*$, then we have

$$\frac{1}{N}Q_N(\widehat{\boldsymbol{\theta}}^*) + \lambda_N\widehat{\boldsymbol{\theta}}^{*T}D\widehat{\boldsymbol{\theta}}^* \leq \frac{1}{N}Q_N(\boldsymbol{\theta}_0^*) + \lambda_N\boldsymbol{\theta}_0^{*T}D\boldsymbol{\theta}_0^*. \tag{1}$$

Since

$$\frac{1}{N}Q_N(\boldsymbol{\theta}_0^*) = \bar{g}_N^T(\boldsymbol{\theta}_0^*)\bar{C}_N^{-1}(\boldsymbol{\theta}_0^*)\bar{g}_N(\boldsymbol{\theta}_0^*) = o(1)$$

by the strong law of large number and (A5), and $\lambda_N = o(1)$,

$$\frac{1}{N}Q_N(\boldsymbol{\theta}_0^*) + \lambda_N\boldsymbol{\theta}_0^{*T}D\boldsymbol{\theta}_0^* \xrightarrow{a.s.} 0.$$

Thus, we can obtain from (1) that

$$\frac{1}{N}Q_N(\widehat{\boldsymbol{\theta}}^*) = \bar{g}_N^T(\widehat{\boldsymbol{\theta}}^*)\bar{C}_N^{-1}(\widehat{\boldsymbol{\theta}}^*)\bar{g}_N(\widehat{\boldsymbol{\theta}}^*) \xrightarrow{a.s.} 0. \tag{2}$$

Since the parameter space Ω is compact, by Glvenko-Cantelli theorem,

$$\sup_{\theta^*\in\Omega}\left|\bar{g}_N(\boldsymbol{\theta}^*) - E[g(\boldsymbol{\theta}^*)]\right| \xrightarrow{a.s.} 0.$$

Hence, by (A5) and the continuous mapping theorem,

$$\left|\bar{g}_N^T(\widehat{\boldsymbol{\theta}}^*)\bar{C}_N^{-1}(\widehat{\boldsymbol{\theta}}^*)\bar{g}_N(\widehat{\boldsymbol{\theta}}^*) - E[g(\widehat{\boldsymbol{\theta}}^*)]^T\mathbf{C}_0^{-1}E[g(\widehat{\boldsymbol{\theta}}^*)]\right| \xrightarrow{a.s.} 0.$$

Combined with (2), we get

$$E[g(\widehat{\boldsymbol{\theta}}^*)]^T\mathbf{C}_0^{-1}E[g(\widehat{\boldsymbol{\theta}}^*)] \xrightarrow{a.s.} 0. \tag{3}$$

Then we will show that it is impossible that $\widehat{\boldsymbol{\theta}}^*$ remains outside of U, where U is any neighborhood of the true parameter $\boldsymbol{\theta}_0^*$. Suppose there exist a neighborhood U such that $\widehat{\boldsymbol{\theta}}^* \in U^c$. Since $E[g(\boldsymbol{\theta}^*)]^T\mathbf{C}_0^{-1}E[g(\boldsymbol{\theta}^*)]$ is a continuous function and U^c is compact, there exists a point $\widetilde{\boldsymbol{\theta}}^* \in U^c$ such that

$$E[g(\widetilde{\boldsymbol{\theta}}^*)]^T\mathbf{C}_0^{-1}E[g(\widetilde{\boldsymbol{\theta}}^*)]$$

achieve its minimum in U^c. By the identification of $\boldsymbol{\theta}^*$ in (A3), there is a unique $\boldsymbol{\theta}_0^* \in \Omega$ satisfying $E[g(\boldsymbol{\theta}_0)] = 0$, we have

$$E[g(\boldsymbol{\theta}^*)]^T \mathbf{C}_0^{-1} E[g(\boldsymbol{\theta}^*)] > 0,$$

which contradicts (3). Then we can prove that $\widehat{\boldsymbol{\theta}}^*$ converges almost surely to $\boldsymbol{\theta}^*$. Thus, $\widehat{\boldsymbol{\theta}}$ is a consistent estimator of $\boldsymbol{\theta}$.

Proof of Theorem 2

The estimate of $\boldsymbol{\theta}$ satisfies

$$0 = \frac{1}{N}\frac{\partial Q_N}{\partial \boldsymbol{\theta}^*}(\widehat{\boldsymbol{\theta}}^*) + 2\lambda_N D\widehat{\boldsymbol{\theta}}^*.$$

By Taylor expansion, we obtain

$$0 = \frac{1}{N}\frac{\partial Q_N}{\partial \boldsymbol{\theta}}^*(\boldsymbol{\theta}_0^*) + 2\lambda_N D\boldsymbol{\theta}_0^* + \Big(\frac{1}{N}\frac{\partial^2 Q_N}{\partial \boldsymbol{\theta}^{*2}}(\widetilde{\boldsymbol{\theta}}^*) + 2\lambda_N D\Big)(\widehat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}_0^*),$$

where $\widetilde{\boldsymbol{\theta}}^*$ is some value between $\widehat{\boldsymbol{\theta}}^*$ and $\boldsymbol{\theta}_0^*$. Thus, we can have

$$\widehat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}_0^* = -\Big(\frac{1}{N}\frac{\partial^2 Q_N}{\partial \boldsymbol{\theta}^{*2}}(\widetilde{\boldsymbol{\theta}}^*) + 2\lambda_N D\Big)^{-1}\Big(\frac{1}{N}\frac{\partial Q_N}{\partial \boldsymbol{\theta}^*}(\boldsymbol{\theta}_0^*) + 2\lambda_N D\boldsymbol{\theta}_0^*\Big). \tag{4}$$

Since $\widehat{\boldsymbol{\theta}}^*$ converges to $\boldsymbol{\theta}_0^*$ in probability and $\widetilde{\boldsymbol{\theta}}^*$ is between $\widehat{\boldsymbol{\theta}}^*$ and $\boldsymbol{\theta}_0^*$, by (A5) and (A6) we can get

$$\begin{aligned}\frac{1}{N}\frac{\partial^2 Q_N}{\partial \boldsymbol{\theta}^{*2}}(\widetilde{\boldsymbol{\theta}}^*) &= 2\frac{\partial \bar{g}_N}{\partial \boldsymbol{\theta}^*}^T(\widetilde{\boldsymbol{\theta}}^*)\bar{C}_N^{-1}(\widetilde{\boldsymbol{\theta}}^*)\frac{\partial \bar{g}_N}{\partial \boldsymbol{\theta}^*}(\widetilde{\boldsymbol{\theta}}^*) + o_p(1) \\ &\xrightarrow{p} 2\mathbf{G}_0^T\mathbf{C}_0^{-1}\mathbf{G}_0\end{aligned}$$

When $\lambda_N = o(N^{-1/2})$,

$$\Big(\frac{1}{N}\frac{\partial^2 Q_N}{\partial \boldsymbol{\theta}^{*2}}(\widetilde{\boldsymbol{\theta}}^*) + 2\lambda_N D\Big)^{-1} = \frac{1}{2}(\mathbf{G}_0^T\mathbf{C}_0^{-1}\mathbf{G}_0)^{-1} + o_p(N^{-1/2}).$$

Similarly, since

$$\frac{1}{N}\frac{\partial Q_N}{\partial \boldsymbol{\theta}^*}(\boldsymbol{\theta}_0^*) = \frac{\partial \bar{g}_N}{\partial \boldsymbol{\theta}^*}^T(\boldsymbol{\theta}_0^*)\bar{C}_N^{-1}(\boldsymbol{\theta}_0^*)\bar{g}_N(\boldsymbol{\theta}_0^*)$$

and $\lambda_N = o(N^{-1/2})$, we have

$$\frac{1}{N}\frac{\partial Q_N}{\partial \boldsymbol{\theta}^*}(\boldsymbol{\theta}_0^*) + 2\lambda_N D\boldsymbol{\theta}_0^* = G_0^T C_0^{-1}\bar{g}_N(\boldsymbol{\theta}_0^*) + o(N^{-1/2}).$$

Therefore, (4) can be written as

$$\sqrt{N}(\widehat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}_0^*) = -\sqrt{N}(\mathbf{G}_0^T\mathbf{C}_0^{-1}\mathbf{G}_0)^{-1}\mathbf{G}_0^T\mathbf{C}_0^{-1}\bar{g}_N(\boldsymbol{\theta}_0^*) + o_p(1). \tag{5}$$

By Central Limit Theorem,

$$\sqrt{N}\bar{g}_N(\boldsymbol{\theta}_0^*) \xrightarrow{d} N(\mathbf{0}, \mathbf{C}_0). \tag{6}$$

Using (5) and (6), we obtain

$$\sqrt{N}(\widehat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}_0^*) \xrightarrow{d} N(\mathbf{0}, (\mathbf{G}_0^T\mathbf{C}_0^{-1}\mathbf{G}_0)^{-1}),$$

and directly,

$$\sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N(\mathbf{0}, \mathbf{J}(\mathbf{G}_0^T\mathbf{C}_0^{-1}\mathbf{G}_0)^{-1}\mathbf{J}^T).$$

BIBLIOGRAPHY

BIBLIOGRAPHY

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

[2] Baker, C. (2004). "Chapter 3. Environment Illustrated". *Behavioral Genetics* AAAS. ISBN 978-0871686978.

[3] Belle, G., Fisher, L. D., Heagerty, P. J. and Lumley, T. (2004). "Chapter 18: Longitudinal data analysis". *Biostatistics: A Methodology For the Health Sciences, 2nd Edition*.

[4] Carpenter, D. O., Arcaro, K., and Spink, D. C. (2002). Understanding the human health effects of chemical mixtures. *Environmental Health Perspectives* **110**(suppl 1), 25-42.

[5] Catalano, P. J. and Ryan, L. M. (1992) Bivariate latent variable models for clustered discrete and continuous outcomes. *Journal of the American Statistical Association* **87(419)**, 651-658.

[6] Choi, Y., Chowdhury, R. and Swaminathan, B. (2014). Prediction of hypertension based on the genetic analysis of longitudinal phenotypes: a comparison of different modeling approaches for the binary trait of hypertension. *BMC Proceedings* **8(Suppl 1)**:S78. doi:10.1186/1753-6561-8-S1-S78.

[7] Crainiceanu, C. and Ruppert, D. (2004). Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society, Series B* **65**, 165-185.

[8] Crainiceanu, C., Ruppert, D., Claeskens, G., and Wand, P. (2005). Exact likelihood ratio tests for penalized splines. *Biometrica* **92**, 91-103.

[9] Crowder M. (1986). On consistency and inconsistency of estimating equations. *Econometric Theory* **2**, 305-330.

[10] Crowder M. (1995). On the use of a working correlation matrix in using generalized linear models for repeated measures. *Biometrika* **82(2)**, 407-410.

[11] Cui, X., Härdle, W., and Zhu, L. (2011). The EFM approach for single-index models. *The Annals of Statistics* **39**, 16581688.

[12] Falconer, D. S. (1952). The problem of environment and selection. *The American Naturalist* **86**, 293-298.

[13] Furlotte, N. A., Eskin, E. and Eyheramendy, S. (2014). Genome-wide association mapping with longitudinal data. *Genet Epidemiol* **36**, 463-471.

[14] Gratten, J. and Visscher, P. M. (2016). Genetic pleiotropy in complex traits and diseases: implications for genomic medicine. *Genome Medicine* **8**:78, doi: 10.1186/s13073-016-0332-x.

[15] Greven, S., Crainiceanu, C., Kühenhoff, H., and Peters, A. (2008). Restricted likelihood ratio testing for zero variance components in linear mixed models. *Journal of Computational and Graphical Statistics* **17**, 870-891.

[16] Johnson, J. A. and Terra, S. G. (2002). Beta-adrenergic receptor polymorphisms: cardiovascular disease associations and pharmacogenetics. *Pharm Res* **19**, 1779-1787.

[17] Kürüm, E., Li, R., Shiffman, S. and Yao, W. (2016). Time-varying coefficient models for joint modeling binary and continuous outcomes in longitudinal data. *Statistica Sinica* **26**, 979-1000.

[18] Leisch, F., Weingessel, A., and Hornik, K. (1998). On the generation of correlated artificial binary data. *Working Paper Series, SFB Adaptive Information Systems and Modelling in Economics and Management Science, Vienna University of Economics*.

[19] Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalised linear models. *Biometrika* **73**, 12-22.

[20] Liu, X., Cui, Y. and Li, R. (2016). Partial linear varying multi-index coefficient model for integrative gene-environment interactions. *Statistica Sinica* **26**, 1037-1060.

[21] Liu, X., Gao, B. and Cui Y. (2017) Generalized partial linear varying multi-index coefficient model for gene-environment interactions. *Statistical Applications in Genetics and Molecular Biology* **16(1)**, 59-74.

[22] Lobo, I. (2008). Pleiotropy: one gene can affect multiple traits. *Nature Education* **1**, 10.

[23] Ma, S. and Song, P. (2015). Varying Index Coefficient Models. *Journal of the American Statistical Association* **110**, 341-356.

[24] Macgregor, S., Knott, S. A., White, I. and Visscher, P. M. (2005) Quantitative trait locus analysis of longitudinal quantitative trait data in complex pedigrees. *Genetics* **171**, 1365-1376.

[25] Song, P., Jiang, Z., Park, E. and Qu A. (2009) Quadratic inferance functions in marginal models for longitudinal data. *Statistics in Medicine* **28**, 3683-3696.

[26] Qu, A. and Li, R. (2006). Quadratic inference functions for varying coefficient models with longitudinal data. *Biometrics* **62**, 379-391.

[27] Qu, A., Lindsay, B. G. and Li, B. (2000). Improving generalised estimation equations using quadratic inference fucntions. *Biometrika* **87**, 823-836.

[28] Ross, C. A., and Smith, W. W. (2007). Gene-enviroment interactions in Parkinson's disease. *Parkinsonism and Related Disorders* **13**, S309-S315.

[29] Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics* **11**, 735-757.

[30] Ruppert, D. and Carroll, R. J. (2000). Spatially-adaptive penalties for spline fitting. *Australian and New Zealand Journal of Statistics* **42**, 205-223.

[31] Self, S. G. and Liang, K. Y. (1987). Assessing cumulative health risks from exposure to environmental mixtures - three fundamental questions. *Journal of the American Statistical Association* **82**, 605-610.

[32] Sexton, K. and Hattis, D. (2007). Asymptotic properties of maximum likelihood estimators and likelihood ratio under non-standard conditions. *Environmental Health Perspectives* **115**, 825-832.

[33] Sitlani, C. M., Rice, K. M., Lumley, T. et al. (2015). Generalized estimating equations for genome-wide association studies using longitudinal phenotype data, *Stat Med* **34**, 118-130.

[34] Stram, D. O. and Lee, J. W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics* **50**, 1171-1177.

[35] Wang, Y. and Chen, H. (2012). On testing an unspecified function through a linear mixed effects model with multiple variance compnents. *Biometrics* **68**, 1113-1125.

[36] Wang, W., Feng, Z., Bull, S. B. and Wang Z. (2014). A 2-step strategy for detecting pleiotropic effects on multiple longitudinal traits. *Front. Genet.* doi.org/10.3389/fgene.2014.00357.

[37] Xu, Z., Shen, X., Pan, W. (2014) Longitudinal analysis is more powerful than cross-sectional analysis in detecting genetic association with neuroimaging phenotypes. *PLoS One* **9(8)**, e102312.

[38] Yu, Y. and Ruppert, D. (2002). Penalized spline estimation for partially linear single-index models. *Journal of the American Statistical Association* **97(460)**, 10421054.

[39] Zhang, D. (2004). Genaralized linear mixed models with varying coefficients for longitudinal data. *Biometrics* **60**, 8-15.

[40] Zimmet, P., Alberti, K., and Shaw, J. (2001). Global and societal implications of the diabetes epidemic. *Nature* **414**, 782-787.