Statistics in Biosciences
https://doi.org/10.1007/s12561-024-09472-3

ORIGINAL PAPER

Functional Varying-Index Coefficients Model for Dynamic Synergistic Gene–Environment Interactions

Jingyi Zhang[1,2] · **Xu Liu**[3] · **Honglang Wang**[4] · **Yuehua Cui**[1]

Received: 28 February 2024 / Revised: 5 October 2024 / Accepted: 15 December 2024

Abstract
Human complex diseases are affected by both genetic and environmental factors. When multiple environmental risk factors are present, the interaction effect between a gene and the environmental mixture can be larger than the addition of individual interactions, resulting in the so-called synergistic gene–environment (G×E) interactions. Existing literature has shown the power of synergistic gene-environment interaction analysis with cross-sectional traits. In this work, we propose a functional varying index coefficient model for longitudinal traits together with multiple longitudinal environmental risk factors 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 function and penalized spline framework. We evaluate some theoretical properties such as estimation consistency and asymptotic normality of the estimates. We further propose a hypothesis testing procedure to assess the significance of the synergistic G×E effect. The performance of the estimation and testing procedure is evaluated through Monte Carlo simulation studies. Finally, the utility of the method is illustrated by a real dataset from a pain sensitivity study in which SNP effects are nonlinearly modulated by a mixture of drug dosages and other environmental variables to affect patients' blood pressure and heart rate.

Keywords Genetic association · Longitudinal data · Mixture exposures · Nonlinear G×E interaction · Quadratic inference function · Varying-index coefficients model

1 Introduction

It has been broadly recognized that gene–environment (G×E) interaction plays important roles in human complex diseases. A growing number of scientific researches have confirmed the role of G×E interaction in many human diseases,

Extended author information available on the last page of the article

Published online: 08 January 2025

such as Parkinson's disease [1] and type 2 diabetes [2]. G×E interaction is defined as how genotypes influence phenotypes differently under different environmental conditions [3]. 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 exposures to multiple environmental factors. The effect of simultaneous exposure is larger than the simple addition of the effects of factors acting alone (e.g., [4, 5]). Under the G×E context, this is the so-called synergistic G×E interaction. This motivated us to assess the combined effect of environmental mixtures, and how they as a whole, interact with genes to affect a disease risk [6]. In our previous models, we proposed a varying multi-index coefficient model (VMICM) to capture the nonlinear interaction between a gene and environmental mixtures with a cross-sectional trait [6, 7]. To our best knowledge, no study has been conducted to assess synergistic G×E effects on a longitudinal trait, and further dissect the interaction mechanism.

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 [8–10]. With longitudinal disease traits, one can study the dynamic gene effect over time. Coupled with longitudinal measures of environmental exposures, one can study how genes respond to the dynamic change of environmental factors to affect a disease trait. Thus, the purpose of this study is to develop a new statistical model to evaluate synergistic G×E effects on a longitudinal trait.

Some nonparametric and semi-parametric models such as varying coefficient models have been proposed to explore time-dependent effects in longitudinal data analysis, for example, [11–17]. However, these methods do not fit our purpose. 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 response, i.e.,

$$Y_{ij} = m_0(\boldsymbol{\beta}_0^T \mathbf{X}_{ij}) + m_1(\boldsymbol{\beta}_1^T \mathbf{X}_{ij}) G_i + \varepsilon_{ij}, \tag{1}$$

where Y_{ij} is the response variable which measures the phenotype 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; and $\boldsymbol{\beta}_0$ 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 restrict the first elements of $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$ be positive.

[18] proposed the quadratic inference function (QIF) for longitudinal data analysis, as an improvement of the generalized estimation equation (GEE) approach introduced by [19]. 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. [18] found that the QIF estimator could be generally more efficient than the GEE estimator. [18] applied the QIF method to the

varying coefficient model for longitudinal data. [20] developed an estimating procedure for single-index models with longitudinal data based on the 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 procedures for model (1). We first approximate the varying index coefficient function by penalized splines [21] 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 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 [22, 23], 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 work is organized as follows: in Sect. 2, we propose an estimation procedure under the FVICM model and further establish the consistency and asymptotic normality of the proposed estimator in Sect. 2.1. In Sect. 3, we discuss some practical issues to implement the proposed estimation procedures. In Sect. 4, 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 simulations in Sect. 5 and illustrate the proposed method by an analysis of a pain sensitivity dataset in Sect. 6, followed by discussions in Sect. 7. Additional simulation and real data analysis results and the proof of the theorems are rendered in the supplemental file.

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.,

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

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

where $\mathbf{B}(u) = (1, u, u^2, \cdots, u^q, (u - \kappa_1)_+^q, \cdots, (u - \kappa_K)_+^q)^T$ 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 $\boldsymbol{\beta} = (\boldsymbol{\beta}_0^T, \boldsymbol{\beta}_1^T)^T$ and $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_0^T, \boldsymbol{\gamma}_1^T)^T$.

For longitudinal data, the conditional variance-covariance matrix of the response needs to be modeled. 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 $\mathbf{y}_i$ and $\mathbf{y}_i = (y_{i1}, ..., y_{in_i})^T$, $\boldsymbol{\mu}_i = E(\mathbf{y}_i)$ 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

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

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, [19] 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, [18] 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 follows:

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

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. [18] introduced the Quadratic Inference Function (QIF) based on the generalized method of moments [24]. Thus, we can estimate the parameters by minimizing the QIF, which is defined as

$$Q_N(\boldsymbol{\theta}) = N\bar{g}_N^T \bar{C}_N^{-1} \bar{g}_N, \tag{3}$$

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

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

To overcome the well-known over-parameterization issue, [18] further proposed the penalized quadratic inference function

$$N^{-1}Q_N(\boldsymbol{\theta}) + \lambda\boldsymbol{\theta}^T\mathbf{D}\boldsymbol{\theta}, \tag{4}$$

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

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}}(N^{-1}Q_N(\boldsymbol{\theta}) + \lambda\boldsymbol{\theta}^T\mathbf{D}\boldsymbol{\theta}). \tag{5}$$

2.1 Asymptotic Properties

In this section, we establish the asymptotic properties of 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 by [25]. 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 [17, 26, 27]. Then, the parameter 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

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

be the Jacobian matrix of dimension $p \times (p-1)$. Denote $\boldsymbol{\beta}_{-1} = (\boldsymbol{\beta}^T_{0,-1}, \boldsymbol{\beta}^T_{1,-1})^T$ 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 supplemental file 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),*exists and converges to the true parameters* $\boldsymbol{\theta}_0$ *in probability*.

Theorem 2 *Suppose the assumptions* (A1)-(A6) *in the supplemental file 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) *is asymptotically normally distributed, i.e.*,

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

where $\mathbf{G}_0$ *and* $\mathbf{C}_0$ *are given in the supplemental file.*

The proofs can be found in the supplemental file.

3 Practical Implementation

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

3.1 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.

3.2 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) for a single observation) and the number of estimating equations is larger than the number of parameters, we have $Q(\widehat{\theta}) \to \chi^2_{r-k}$ in distribution [24], where r is the dimension of $\bar{g}_N(\boldsymbol{\theta})$, k is the dimension of $\boldsymbol{\theta}$, $\widehat{\theta}$ 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 $Q(\widehat{\theta})$ is asymptotically chi-square distributed, we can use BIC to penalize $Q(\widehat{\theta})$ 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

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

The model with minimum BIC would be considered better. If we choose h basis matrices in (2), then $r - k = hk - k = (h - 1)k$. As will be discussed in Section 3.3, we usually use h=2 in our setting. Thus, the BIC criterion is actually

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

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 $u(= \boldsymbol{\beta}^T \mathbf{X})$.

3.3 Choice of the Basis for the Inverse of the Correlation Matrix

[25] offered several choices of basis matrixes. For exchangeable working correlation, $\mathbf{M}_1$ is an 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 [25] through simulation studies. [28] also proposed an adaptive estimation method to approximate the correlation empirically when there is no prior information available.

3.4 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 [25] suggested, we can extend the generalized cross-validation [29] 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 (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 to reduce the computational time.

4 Hypothesis Test

4.1 Linear Mixed Model Representation for FVICM Model

In our proposed FVICM model (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 $u_0 = \boldsymbol{\beta}_0^T\mathbf{X}$, $u_1 = \boldsymbol{\beta}_1^T\mathbf{X}$, with the truncated power spline basis, the coefficient function can be modeled by

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

Note that under the current model setup, we cannot assess the zero effect of the nonparametric function $m_1(\cdot)$ since under the null hypothesis of $m_1(\cdot) = 0$, the index loading parameters $\boldsymbol{\beta}_1$ are not identifiable, unless we impose the constraint that $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_0 = \boldsymbol{\beta}$. This constraint, however, is practically unrealistic. Thus, our goal is to test the linearity of $m_1(u_1)$, which is equivalent to test

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

If the above H_0 is rejected, we conclude there exists a nonlinear relationship. Otherwise, we assume a linear relationship and fit $m_1(\cdot)$ with a linear function and further test the zero effect of the linear relationship. Let $\mathbf{w}_{0ij} = (1, u_{0ij}, \cdots, u_{0ij}^q)^T$, $\mathbf{z}_{0ij} = \left((u_{0ij} - \kappa_1)_+^q, \cdots, (u_{0ij} - \kappa_K)_+^q\right)^T$, $\tilde{\boldsymbol{\gamma}}_0 = (\gamma_{00}, \cdots, \gamma_{0q})^T$, $\mathbf{b}_0 = (b_{01}, \cdots, b_{0K})^T$, $\mathbf{w}_{1ij} = (1, u_{1ij}, \cdots, u_{1ij}^q)^T$, $\mathbf{z}_{1ij} = \left((u_{1ij} - \kappa_{11})_+^q, \cdots, (u_{1ij} - \kappa_{1K})_+^q\right)^T$,$\mathbf{b}_1 = (b_{11}, \cdots, b_{1K})^T$, and $\tilde{\boldsymbol{\gamma}}_1 = (\gamma_{10}, \cdots, \gamma_{1q})^T$, then we have $m_0(u_{0ij}) = \mathbf{w}_{0ij}^T\tilde{\boldsymbol{\gamma}}_0 + \mathbf{z}_{0ij}^T\mathbf{b}_0$ and $m_1(u_{1ij}) = \mathbf{w}_{1ij}^T\tilde{\boldsymbol{\gamma}}_1 + \mathbf{z}_{1ij}^T\mathbf{b}_1$.

We further define $\mathbf{Y}_i = (y_{i1}, \cdots, y_{in_i})^T$, $\mathbf{W}_{0i} = (\mathbf{w}_{0i1}, \cdots, \mathbf{w}_{0in_i})^T$, $\mathbf{W}_{1i} = (\mathbf{w}_{1i1}G_i, \cdots, \mathbf{w}_{1in_i}G_i)^T$, $\mathbf{Z}_{0i} = (\mathbf{z}_{0i1}, \cdots, \mathbf{z}_{0in_i})^T$, and $\mathbf{Z}_{1i} = (\mathbf{z}_{1i1}G_i, \cdots, \mathbf{z}_{1in_i}G_i)^T$, then a linear mixed model (LMM) representation [30] can be obtained as

$$\mathbf{Y}_i = \mathbf{1}_i a_i + \mathbf{W}_{0i}\tilde{\boldsymbol{\gamma}}_0 + \mathbf{W}_{1i}\tilde{\boldsymbol{\gamma}}_1 + \mathbf{Z}_{0i}\mathbf{b}_0 + \mathbf{Z}_{1i}\mathbf{b}_1 + \boldsymbol{\epsilon}_i, \quad i = 1, \cdots, n, \tag{6}$$

where $\mathbf{b}_l \sim N(\mathbf{0}, \sigma^2_{\mathbf{b}_l}\mathbf{I}_K)$, $l = 0, 1$, $\boldsymbol{\epsilon}_i \sim N(\mathbf{0}, \sigma^2_\epsilon \mathbf{I})$, and the random intercept effects a_i are assumed to be independent as $N(0, \sigma^2_a)$.

With the LMM representation, testing the linearity of the varying index coefficients is equivalent to testing some fixed effects and a variance component in model (6). To be specific, we want to test

$$H_0 : \gamma_{12} = \cdots = \gamma_{1q} = 0 \text{ and } \sigma^2_{\mathbf{b}_1} = 0. \tag{7}$$

4.2 Likelihood Ratio Test (LRT) and Pseudo-LRT in LMM

4.2.1 LRT for One Variance Component

[31] proposed the likelihood ratio test in linear mixed effect models with one variance component. Consider an LMM with one variance component

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon}, \quad E\begin{bmatrix}\mathbf{b}\\ \boldsymbol{\epsilon}\end{bmatrix} = \begin{bmatrix}\mathbf{0}_L\\ \mathbf{0}_n\end{bmatrix}, \quad \text{Cov}\begin{bmatrix}\mathbf{b}\\ \boldsymbol{\epsilon}\end{bmatrix} = \begin{bmatrix}\left\{ cc\sigma^2_b\boldsymbol{\Sigma} & \mathbf{0}\right.\\ \mathbf{0} & \sigma^2_\epsilon\mathbf{I}_n\end{bmatrix}, \tag{8}$$

where $\boldsymbol{\beta}$ is a p-dimensional vector of fixed effect coefficients, $\mathbf{b}$ is a L-dimensional vector of random effects, $\mathbf{0}_L$ is a L-dimensional vector of zeros, $\boldsymbol{\Sigma}$ is a known $L \times L$ symmetric positive definite matrix. Let $\lambda = \sigma^2_b/\sigma^2_\epsilon$ be the signal-to-noise ratio and then the covariance matrix of $\mathbf{Y}$ can be written as $\text{Cov}(\mathbf{Y}) = \sigma^2_\epsilon \mathbf{V}_\lambda$, where $\mathbf{V}_\lambda = \mathbf{I}_n + \lambda\mathbf{Z}\boldsymbol{\Sigma}\mathbf{Z}^T$. Consider testing for the null hypothesis

$$H_0 : \beta_{p+1-p'} = 0, \cdots, \beta_p = 0, \ \sigma^2_b = 0 \tag{9}$$

for $p' > 0$.

The LRT statistic is defined as

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

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

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

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

for fixed λ, we obtain the LRT statistic

$$LRT_n = \sup_{\lambda \geq 0}\{n\log(\mathbf{Y}^T\mathbf{S}_0\mathbf{Y}) - n\log(\mathbf{Y}^T\mathbf{P}^T_\lambda\mathbf{V}^{-1}_\lambda\mathbf{P}_\lambda\mathbf{Y}) - \log|\mathbf{V}_\lambda|\}, \tag{10}$$

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, and $\mathbf{S}_0 = \mathbf{I}_n - \mathbf{X}_0(\mathbf{X}_0^T\mathbf{X}_0)^{-1}\mathbf{X}_0^T$.

Theorem 1 in [31] provides the distribution of LRT statistic (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 $\mathbf{\Sigma}^{1/2}\mathbf{Z}^T\mathbf{Z}\mathbf{\Sigma}^{1/2}$, $s = 1, \cdots, L$, then

$$LRT_n \overset{d}{=} n\left(1 + \frac{\sum_1^{p'} u_s^2}{\sum_1^{n-p} w_s^2}\right) + \sup_{\lambda \geq 0} f_n(\lambda), \tag{11}$$

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

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

with

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

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

The distribution in (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. A detailed algorithm for this simulation can be found in [31].

4.2.2 Pseudo-LRT for Multiple Variance Components

For an LMM with multiple variance components

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}_s + \cdots + \mathbf{Z}\mathbf{b}_S + \boldsymbol{\epsilon}, \tag{12}$$

$$\mathbf{b}_s \sim N(\mathbf{0}, \sigma_s^2\mathbf{\Sigma}_s),\ s = 1, \cdots, S,\quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma_\epsilon^2\mathbf{I}_n),$$

where $\mathbf{b}_s$, $s = 1, \cdots, S$ are random effects and $S > 1$. Suppose we are interested in testing

$$H_0 : \beta_{p+1-p'} = 0, \cdots, \beta_p = 0,\ \sigma_s^2 = 0.$$

[32] proposed to approximate the distribution of LRT for the model (12) based on the pseudo-likelihood ratio test theory [33] using a pseudo-outcome. In the framework of model (12), $\mathbf{b}_l, l \neq s$ are the nuisance random parameters. We can define the pseudo-outcome as

$$\widetilde{\mathbf{Y}} = \mathbf{Y} - \sum_{l \neq s} \mathbf{Z}_l \widehat{\mathbf{b}}_l,$$

where $\widehat{\mathbf{b}}_l$ are the best linear unbiased predictors (BLUP) of nuisance random effects $\mathbf{b}_l, l \neq s$. Then, the model (12) can be reduced to

$$\widetilde{\mathbf{Y}} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_s \mathbf{b}_s + \boldsymbol{\epsilon}. \tag{13}$$

Then, the method for testing one variance component introduced by [31] can be applied to the model in (13).

4.3 Pseudo-LRT in FVICM Model

For the model in (6), we can use the idea of [32] and define the pseudo-outcome

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

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 (6) can be written as

$$\widetilde{\mathbf{Y}}_i = \mathbf{W}_{0i}\tilde{\boldsymbol{\gamma}}_0 + \mathbf{W}_{1i}\tilde{\boldsymbol{\gamma}}_1 + \mathbf{Z}_{1i}\mathbf{b}_1 + \boldsymbol{\epsilon}_i. \ i = 1, \cdots, n. \tag{14}$$

For the new model (14) using pseudo-response, we can apply the method for the single variance component model introduced in Sect. 10 to test hypothesis (7). Statistical significance can be assessed through the resampling approach described in section 4.2.1.

5 Simulation Study

5.1 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 terms $\epsilon_i = (\epsilon_{i1}, \cdots, \epsilon_{in_i})$ 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 $\rho = 0.5$ 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 datasets 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 using the BIC method.

5.2 Performance of Estimation

Table 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 the 95% confidence level are reported. Note that the estimation of the loading parameter $\boldsymbol{\beta}_1$ improves (smaller Bias, SD, and SE and CP closer to 95%), as MAF p_A increases, while the estimation of $\boldsymbol{\beta}_0$ show an 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.

The plots for the estimations of $m_0(u_0)$ and $m_1(u_1)$ under different sample sizes and MAFs are shown in Figs. 1 and 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 $m_0(u_0)$ shows an opposite direction, which coincides with the results for the parametric estimation in Table 1.

Simulation results for the case with $\rho = 0.8$ are shown in the supplemental file (See Table S1, Figures S1 and S2). 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.

5.3 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 $m_1^0(u_1) = \delta_0 + \delta_1 u_1$ and 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

Table 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

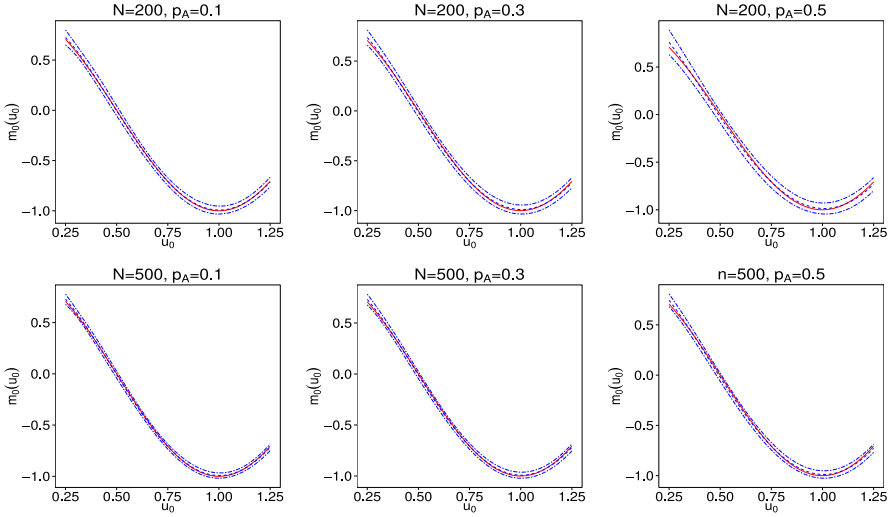

Fig. 1 The estimation of function $m_0(\cdot)$ under different MAFs when N=200, 500, and ρ=0.5. The estimated and true functions are denoted by the solid and dashed lines, respectively. The 95% confidence band is denoted by the dotted dash line

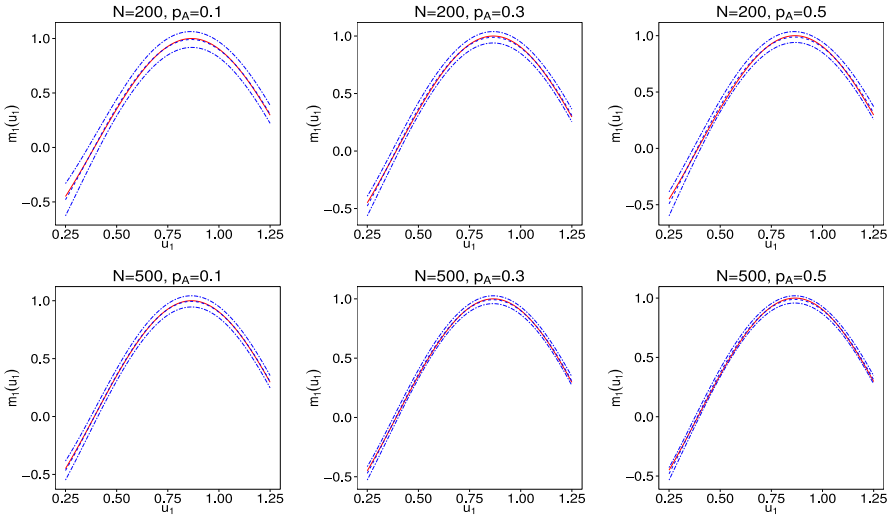

Fig. 2 The estimation of function $m_1(\cdot)$ under different MAFs when N=200, 500 and ρ=0.5. The estimated and true functions are denoted by the solid and dashed lines, respectively. The 95% confidence band is denoted by the dotted dash line

$H_1^\tau : m_1^\tau(\cdot) = m_1^0(\cdot) + \tau\{m_1(\cdot) - m_1^0(\cdot)\}$. When $\tau = 0$, the corresponding power is the false-positive rate.

Figure 3 shows the size (when $\tau = 0$) and power (when $\tau > 0$) at the 0.05 significance level. We obtain 1000 Monte Carlo simulations each with 5000 replications

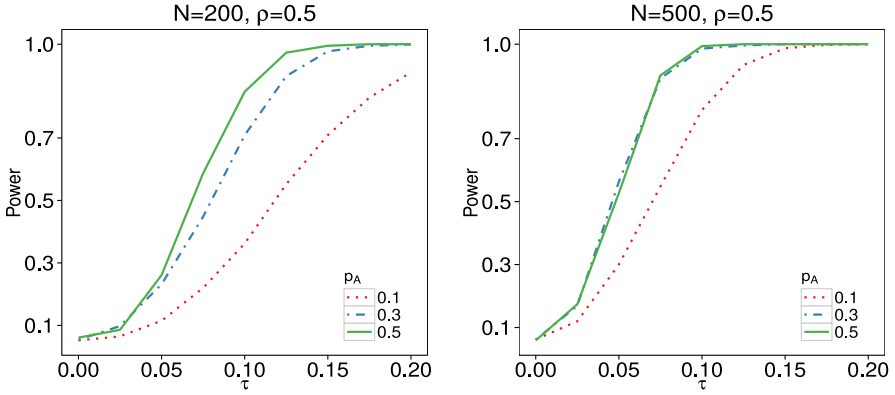

Fig. 3 The empirical size and power of testing the linearity of the nonparametric function m_1 under different MAFs when N=200 and 500 and ρ=0.5

to access the null distribution of test statistic under sample sizes $N = 200$ and 500 with $\rho = 0.5$. The empirical type I error under three MAFs is very close to the nominal level of 0.05 and the power increases dramatically when MAF increases from 0.1 to 0.3. Results for $\rho = 0.8$ are presented in Figure S3 in the supplemental file. 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 Fig. 3, the power increases a little bit slower when $\rho = 0.8$. The numerical empirical type I error rates across different MAFs, correlations, and sample sizes are presented in Table S2 of the supplemental file. We also included the testing results for the case with $N = 150$, which mimics the real data (see Figure S4 in the supplemental file), with the empirical type I error rates detailed in Table S2. We observed a slight conservativeness for the type I error at $P_A = 0.1$, with improvement as the MAF increases to 0.3 and 0.5. The difference between $P_A = 0.3$ and 0.5 is minimal. The power pattern is quite similar to the case with $N = 200$. The results indicate that our method can reasonably control the false-positive rates and has appropriate power to detect genetic variation.

6 Real Data Application

We applied the proposed FVICM model to a real dataset from a pharmacogenetic study of cardiovascular disease [34]. Cardiovascular disease, including heart disease and stroke, is the top cause of death for men and women across all racial and ethnic backgrounds. Dobutamine, a medication that stimulates the heart, is used to manage congestive heart failure by enhancing heart rate and the strength of heart contractions via β-adrenergic receptors (βARs). A group of 163 men and women aged 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 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 were treated as "time" and measurements at different dosage levels were considered as longitudinal measurements. In addition to that age and body mass index (BMI) were also recorded.

Five SNPs in genes β_1AR and β_2AR were genotyped, namely *codon16*, *codon27*, *codon49*, *codon389*, and *codon492*. We chose X_1= dosage level as the "time-varying" variable, and X_2 = age and X_3= BMI as the "time-invariant" variables. Our goal was 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 tested whether any SNP was associated with the drug response in a linear fashion 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 Tables 2 and S3-S4 in the supplemental file. We also reported the p-values for testing the significance of the index loading coefficients β_{11}, β_{12}, and β_{13}, which were labeled by $p_{\beta_{11}}$, $p_{\beta_{12}}$, and $p_{\beta_{13}}$, based on the asymptotic normality of the estimates. We also compared 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 the unknown functions of X_1. To see the relative gain by integrative analysis, we calculated 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 summarizes the performance of our method for response SBP. In the table, p_{m_1} for all 5 SNPs is 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 show 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 4 shows the fitted nonlinear functions for each SNP, along with the 95% confidence bands.

The tables and figures for DBP and HR are presented in the supplemental file. Table S3 presents similar results for response DBP. The values of p_{m_1} show that

Table 2 List of SNPs with MAF, alleles, and p-values under different hypotheses and MSE for SBP

			p-value					MSE	
SNP ID	MAF	Alleles	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

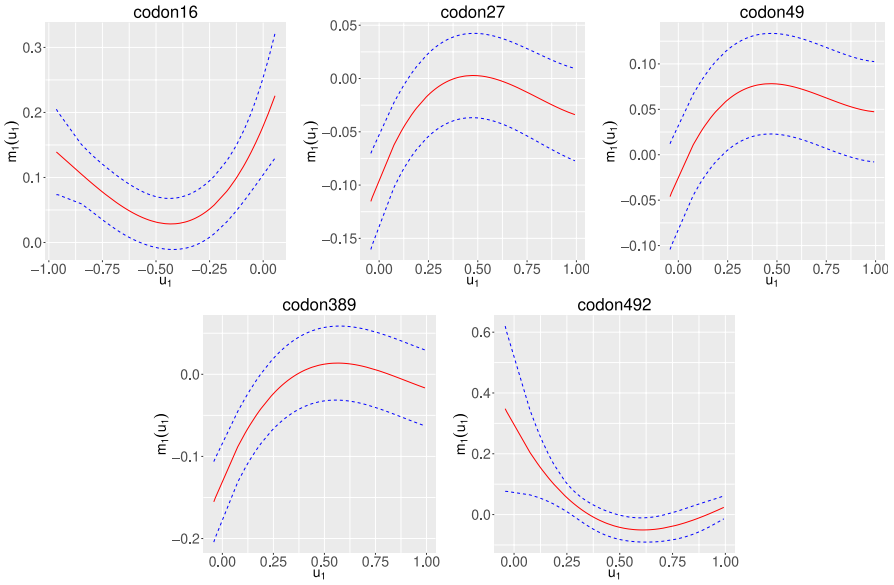

Fig. 4 Plot of the estimate (solid curve) of the nonparametric function $m_1(u_1)$ for SNPs codon16, codon27, codon49, codon389, and codon492. The 95% confidence band is denoted by the dashed line. The response is SBP

the test results for all 5 SNPs are significant, indicating nonlinear interactions for all 5 SNPs, while no significance is shown for the AVCM model. MSEs further support our method by showing a smaller value for FVICM compared with AVCM. The estimated interaction curves with 95% confidence bands are shown in Figure S5.

In Table S4, the performance of our method for trait HR also leads to a similar conclusion except for SNP *codon16*, which shows (marginal) significant test results for both models. For all the other SNPs, FVICM outperforms AVCM in terms of MSE. Figure S6 displays the corresponding estimated nonlinear interaction curves.

7 Discussion

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

Our FVICM model is different from 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 reduces to 1. FVICM can 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 burdens by treating multiple environmental variables as a single-index variable. The advantage of modeling multiple variables as a single index and further assessing its effect via a nonparametric function has also been demonstrated by [6, 17] in a cross-sectional design. Our real data analysis results further confirmed the advantage under a longitudinal design.

We applied the model to a pharmacogenetic study of cardiovascular disease [34]. Testing results indicated 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 and offers additional insights that otherwise cannot be revealed by models with cross-sectional data. By checking the nonlinear effect function together with the confidence band, people can get a sense of how genes respond to the combined change of the environmental factors over time to affect a response variable. Although the method was demonstrated using a candidate gene study, it is capable of analyzing a large number of SNPs, limited only by computational constraints.

As noted by [35], misspecifying environmental main effects can lead to false-positive interaction findings, particularly when gene–environment correlations exist. In our study, we assume gene–environment independence, so any significant interactions identified in our analysis are likely to represent true functional interactions rather than spurious associations driven by gene–environment correlations. Additionally, we model the intercept term of the joint effect of multiple environmental mixtures using a flexible nonparametric approach. This adaptability allows the model to better capture the underlying data structure, reducing the risk of misspecification of the environmental main effect.

Our 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.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12561-024-09472-3.

Acknowledgements The authors wish to thank the associate editor and two anonymous reviewers for their insightful comments and suggestions that greatly improved the presentation of the manuscript.

Funding This work was supported in part by the National Human Genome Research Institute of the National Institutes of Health (NIH) under award number R21HG010073 and by the American Heart Association under award number 24TPA1288424 (to Y. Cui) and by the National Science Foundation under award number DMS-2212928 (to H. Wang). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data Availability Additional results from the simulations and real data analyses, as well as the proofs of the theorems, are available in the online supplemental file. The R code used to implement the method can be accessed at https://github.com/Honglang/FVICM.

Declarations

Conflict of interest The authors have no conflict of interest to declare that are relevant to the content of this article.

References

1. Ross CA, Smith WW (2007) Gene-enviroment interactions in Parkinson's disease. Parkinsonism Relat Disord 13:S309–S315
2. Zimmet P, Alberti K, Shaw J (2001) Global and societal implications of the diabetes epidemic. Nature 414:782–787
3. Falconer DS (1952) The problem of environment and selection. Am Nat 86:293–298
4. Carpenter DO, Arcaro K, Spink DC (2002) Understanding the human health effects of chemical mixtures. Environ Health Perspect 110(suppl 1):25–42
5. Sexton K, Hattis D (2007) Asymptotic properties of maximum likelihood estimators and likelihood ratio under non-standard conditions. Environ Health Perspect 115:825–832
6. Liu X, Cui Y, Li R (2016) Partial linear varying multi-index coefficient model for integrative gene-environment interactions. Stat Sin 26:1037–1060
7. Liu X, Gao B, Cui Y (2017) Generalized partial linear varying multi-index coefficient model for gene-environment interactions. Stat Appl Genet Mol Biol 16:59–74
8. Sitlani CM, Rice KM, Lumley T et al (2015) Generalized estimating equations for genome-wide association studies using longitudinal phenotype data. Stat Med 34:118–130
9. Furlotte NA, Eskin E, Eyheramendy S (2014) Genome-wide association mapping with longitudinal data. Genet Epidemiol 36:463–471
10. 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
11. Hoover DR, Rice JA, Wu CO, Yang L-P (1998) Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. Biometrika 85:809–822
12. Wu CO, Chiang C-T, Hoover DR (1998) Asymptotic confidence regions for kernel smoothing of a varying-coefficient model with longitudinal data. J. Amer. Statist. Assoc. 93:1388–1402
13. Fan J, Zhang JT (2000) Functional linear models for longitudinal data. J Roy Stat Soc B 62:303–322
14. Martinussen T, Scheike T (2001) Sampling adjusted analysis of dynamic additive regression models for longitudinal data. Scand J Stat 28:303–323
15. Chiang CT, Rice JA, Wu CO (2001) Smoothing Spline Estimation for Varying Coefficient Models with Repeatedly Measured Dependent Variables. J Am Stat Assoc 96:605–619
16. Huang JZ, Wu CO, Zhou L (2002) Varying-coefficient models and basis function approximations for the analysis of repeated measurements. Biometrika 89:111–128
17. Ma S, Song P (2015) Varying Index Coefficient Models. J Am Stat Assoc 110:341–356
18. Qu A, Lindsay BG, Li B (2000) Improving generalised estimation equations using quadratic inference fucntions. Biometrika 87:823–836
19. Liang KY, Zeger SL (1986) Longitudinal data analysis using generalised linear models. Biometrika 73:12–22
20. Bai Y, Fung WK, Zhu Z (2009) Penalized quadratic inference functions for single-index models with longitudinal data. J Multivar Anal 100:152–161
21. Ruppert D, Carroll RJ (2000) Spatially-adaptive penalties for spline fitting. Aust N Z J Stat 42:205–223
22. Ruppert D, Wand M, Carroll R (2003) Semiparametric Regression. Cambridge University Press, Cambridge
23. Wand M (2003) Smoothing and mixed models. Comput Statistics 18:223–249
24. Hansen LP (1982) Large sample properties of generalized method of moments estimators. Econometrica 50:1029–1054

25. Qu A, Li R (2006) Quadratic inference functions for varying coefficient models with longitudinal data. Biometrics 62:379–391
26. Yu Y, Ruppert D (2002) Penalized spline estimation for partially linear single-index models. J Am Stat Assoc 97:1042–1054
27. Cui X, Härdle W, Zhu L (2011) The EFM approach for single-index models. Ann Stat 39:1658–1688
28. Qu A, Lindsay BG (2003) Building adaptive estimating equations when inverse of covariance estimation is difficult. J Roy Stat Soc B 65:127–142
29. Ruppert D (2002) Selecting the number of knots for penalized splines. J Comput Graph Stat 11:735–757
30. Wang Y, Chen H (2012) On testing an unspecified function through a linear mixed effects model with multiple variance compnents. Biometrics 68:1113–1125
31. Crainiceanu C, Ruppert D (2004) Likelihood ratio tests in linear mixed models with one variance component. J Roy Stat Soc B 65:165–185
32. Greven S, Crainiceanu C, Kühenhoff H, Peters A (2008) Restricted likelihood ratio testing for zero variance components in linear mixed models. J Comput Graph Stat 17:870–891
33. Liang KY, Self SG (1996) On the asymptotic behaviour of the pseudolikelihood ratio test statistic. J Roy Stat Soc B 58:785–796
34. Johnson JA, Terra SG (2002) Beta-adrenergic receptor polymorphisms: cardiovascular disease associations and pharmacogenetics. Pharm Res 19:1779–1787
35. Sun R, Carroll RJ, Christiani DC, Lin X (2018) Testing for gene-environment interaction under exposure misspecification. Biometrics 74:653–662

Authors and Affiliations

Jingyi Zhang[1,2] · Xu Liu[3] · Honglang Wang[4] · Yuehua Cui[1]

✉ Yuehua Cui
cuiy@msu.edu

[1] Department of Statistics and Probability, Michigan State University, East Lansing, MI 48824, USA

[2] Amazon Lab126, Sunnyvale, CA 94089, USA

[3] School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai 200433, China

[4] Department of Mathematical Sciences, Indiana University Indianapolis, Indianapolis, IN 46202, USA