Statistica Sinica **26** (2016), 1037-1060
doi:http://dx.doi.org/10.5705/ss.202015.0114

PARTIAL LINEAR VARYING MULTI-INDEX COEFFICIENT MODEL FOR INTEGRATIVE GENE-ENVIRONMENT INTERACTIONS

Xu Liu[1], Yuehua Cui[1] and Runze Li[2]

[1]*Michigan State University and* [2]*Pennsylvania State University*

Abstract: Gene-environment (G×E) interactions play key roles in many complex diseases. An increasing number of epidemiological studies have shown the combined effect of multiple environmental exposures on disease risk. However, no appropriate statistical models have been developed to conduct a rigorous assessment of such combined effects when G×E interactions are considered. In this paper, we propose a partial linear varying multi-index coefficient model (PLVMICM) to assess how multiple environmental factors act jointly to modify individual genetic risk on complex disease. Our model includes the varying-index coefficient model as a special case, where discrete variables are admitted as the linear part. Thus PLVMICM allows one to study nonlinear interaction effects between genes and continuous environments as well as linear interactions between genes and discrete environments, simultaneously. We derive a profile method to estimate parametric parameters and a B-spline backfitted kernel method to estimate nonlinear interaction functions. Consistency and asymptotic normality of the parametric and nonparametric estimates are established under some regularity conditions. Hypothesis testing for the parametric coefficients and nonparametric functions are conducted. Results show that the statistics for testing the parametric coefficients and the non-parametric functions asymptotically follow a χ^2-distribution with different degrees of freedom. The utility of the method is demonstrated through extensive simulations and a case study.

Key words and phrases: Association study, backfitting, B-spline, single index model, varying coefficient model.

1. Introduction

There has been great interest in identifying gene-environment (G×E) interaction in the scientific literature. G×E interaction is defined as how genotypes influence phenotypes differently under different environmental conditions (Falconer (1952)), a phenomenon also termed as genetic sensitivity to environmental stimulus. A growing number of reports have confirmed the role of G×E interaction in many diseases, such as Parkinson disease (Ross and Smith (2007)) and type 2 diabetes (Zimmet, Alberti, and Shaw (2001)). G×E interaction has traditionally been pursued based on a single environment exposure model. Evidence from epidemiological studies has clearly indicated that disease risk can be modified by simultaneous exposure to multiple environmental factors, higher than what would be expected from simple addition of the effects of factors acting alone (Carpenter, Arcaro, and Spink (2002); Sexton and Hattis (2007)). Thus, assessing the combined effect of environmental mixtures and the mechanism in which they, as a whole, interact with genes to affect disease risk could shed novel insight into disease etiology. Suppose that Y is the trait response of primary interest. In many genetic studies, one collects a p-dimensional continuous covariate vector $\mathbf{X}$, and a q-dimensional discrete covariate vector $\mathbf{Z}$. Motivated by an empirical analysis to study G×E interaction, see Section 5, we propose a partial linear varying multi-index coefficient model (PLVMICM):

$$Y = m_0(\boldsymbol{\beta}_0^T\mathbf{X}) + \boldsymbol{\alpha}_0^T\mathbf{Z} + \sum_{l=1}^{L}\{m_l(\boldsymbol{\beta}_l^T\mathbf{X})G_l + \boldsymbol{\alpha}_l^T\mathbf{Z}G_l\} + \varepsilon, \tag{1.1}$$

where G_l, $l = 1,\ldots,L$ are genetic variables (e.g., single nucleotide polymorphisms (SNPs)) of interest, ε is an error term with mean 0 and finite variance; $m_l(\cdot), l = 0,1,\ldots,L$ are unknown index functions; $\boldsymbol{\alpha}_0,\ldots,\boldsymbol{\alpha}_L$ and $\boldsymbol{\beta}_0,\ldots,\boldsymbol{\beta}_L$ are parameters of interest, where the index coefficients $\boldsymbol{\beta}_l$ are the index loadings or the loading parameters. The SNP variable G_l can be coded as 2, 1, and 0 for genotype AA, Aa, and aa, assuming an additive model. Note that the main genetic effect for each G_l is captured by the function $m_l(\boldsymbol{\beta}_l^T\mathbf{X})$ $(l = 1,\ldots,L)$. Thus we do not need to have a separate term to model the main genetic effect for each SNP. Model (1.1) provides a unified model framework for many existing models used for studying $G \times E$ interaction. Specifically, the model proposed in Ma et al. (2011) can be viewed as a special case with $p = 1$ (the dimension of $\boldsymbol{\beta}_l$), $q = 0$ (the dimension of $\boldsymbol{\alpha}_l$), and $L = 1$. Model (1.1) also include the semiparametric varying-index coefficient model proposed by Ma and Xu (2015), studying G×E interaction, as a special case with $\boldsymbol{\beta}_0 = \boldsymbol{\beta}_l = \boldsymbol{\beta}, l = 1,\ldots,L$, i.e., assuming the same index loading parameter. Our empirical analysis in the data example in Section 5 clearly shows that this assumption is not realistic, making it necessary to allow different loading parameters in the model.

Model (1.1) also includes many other existing models as special cases. It reduces to the partial linear single-index model (Carroll et al. (1997); Xia and Li (1999); Xia and Härdle (2006); Liang et al. (2010); Cui, Härdle, and Zhu (2011)), in which the discrete variable in the linear part is admitted if all $G_l = 0$; it reduces to VICM proposed by Ma and Song (2015) if $\mathbf{Z} = 0$.

This paper aims to develop a set of statistical estimation and hypothesis procedures for model (1.1). We employ a B-spline backfitted kernel smoothing

(BSBK) procedure to estimate the parametric parameters and the nonparametric functions (Wang and Yang (2007)). We first develop a profile least squares method to estimate the index coefficients $\boldsymbol{\beta}_l$ and the linear coefficients $\boldsymbol{\alpha}_l$ by approximating unknown function $m_l(\cdot)$ with B-spline basis functions. The parametric estimates can be shown to be $n^{1/2}$-consistent and asymptotically normal. We also obtain uniformly consistent estimators of the nonparametric functions. Given the $n^{1/2}$-consistent parametric estimators and the consistent estimators of the nonparametric functions, the kernel estimators of nonparametric functions can be obtained from which we establish the asymptotic normality.

Under model (1.1), it is natural to ask whether there is an interaction between discrete/continuous environments and genes, and whether the interaction with the combined environmental exposures is linear or nonlinear. Cai, Fan, and Li (2000) studied the nonparametric testing problem for varying coefficient models based on the generalized likelihood ratio test. Nonparametric inferences for additive models were previously discussed by employing the generalized likelihood ratio (GLR) statistic (Fan and Jiang (2005)). We propose a parametric likelihood ratio test to test for the linear interaction term and a nonparametric GLR test to test for the nonparametric interaction functions (Fan, Zhang, and Zhang (2001)). We further show that the proposed nonparametric GLR statistic is asymptotically χ^2. We conduct rigorous theoretical evaluation of the proposed estimators and test statistics and show the utility of the model through extensive simulations and a case study.

The paper is organized as follows. In Section 2.2, we formulate the model and describe the BSBK procedure and the parametric estimators for the continuous and discrete parts based on a profile least squares method. The nonparametric kernel estimators for index functions are given in Section 2.3. The consistency and normality of parametric and nonparametric estimators are given in Section 2.4. Section 3 gives the parametric likelihood ratio statistic and several nonparametric GLR statistics, as well as their theoretical properties. In Section 4, we report on simulation studies that illustrate the finite sample performance of the proposed estimators and test statistics. In Section 5 we show the utility of the method by applying it to a baby birthweight data set. Some concluding remarks are given in Section 6. The proofs of the main results are relegated to the Appendix.

2. Estimation Procedures

2.1. Estimation Procedures

We focus on the situation with $L = 1$ for ease of presentation, and rewrite (1.1) as

$$Y = m_0(\boldsymbol{\beta}_0^T\mathbf{X}) + \boldsymbol{\alpha}_0^T\mathbf{Z} + m_1(\boldsymbol{\beta}_1^T\mathbf{X})G + \boldsymbol{\alpha}_1^T\mathbf{Z}G + \varepsilon. \tag{2.1}$$

The proposed procedure for model (2.1) can be easily extend to model (1.1) with multiple G's (i.e., multiple SNPs), and it is still more general than the existing ones used for G×E interaction. It is motivated by a recent genome-wide association study to identify genetic risk factors interacting with maternal uterine environments to increase the risk of low and high birth weight (HAPO Study Cooperative Research Group (2009)). The underlying hypothesis is that the variation of birth weight can be explained by complex G×E interactions in the context of the maternal-fetal unit. As a fetus resides inside its mother's womb, there is intensive signalling and chemical exchanges between the two. The effects of fetal genes could be modified by simultaneous exposure to multiple stimuli from the mother's side such as mother's glucose level and blood pressure. For continuously measured environmental variables, we propose to model the joint effect of environment variables as a whole through an unknown index function $m(\cdot)$. The index function can be linear or nonlinear. That is determined by the data, with flexibility to capture the underlying mechanism of environmental mixtures modifying genetic influences on disease risk. For such discrete environmental variables as smoking status and family disease history, their interaction effects with genes can be modeled through a parametric function.

The motivation for assessing nonlinear G×E interaction in complex disease has been discussed extensively in Ma et al. (2011) and Wu and Cui (2013). The model for testing nonlinear G×E interactions in Ma et al. (2011) can be viewed as a special case of (2.1) with $p = 1$ (the dimension of $\boldsymbol{\beta}_l$) and $q = 0$ (the dimension of $\boldsymbol{\alpha}_l$). We assume the index loading parameters $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$ to be different; this differs from the single index model assuming common loading parameters for different index functions proposed by Xia and Li (1999). Li, Wang, and Carroll (2010) studied the generalized functional linear models with semi-parametric single index interaction, but did not allow dissimilar loading parameters in different index functions. Although the varying-index coefficient model (VICM) proposed by Ma and Song (2015) could consider the joint interaction of multiple environments with genes, it does not admit discrete variables $\mathbf{Z}$. Such discrete environmental variables are common in G×E studies and the inclusion of these variables is crucial to assess the discrete G×E interactions, as implemented in most partial linear single index models (Carroll et al. (1997); Xia and Li (1999); Xia and Härdle (2006);Liang et al. (2010)). Nevertheless, including both parametric and nonparametric terms into the same model poses computational and theoretical challenges. As discussed earlier, our model differs from that proposed by Ma and Xu (2015) in which they assumed the same loading parameters for different index functions. This assumption is too strong in reality, the modulation effect of environmental variables may differ from gene to gene. Our data analysis results in Section 5 indicate that such an assumption is invalid there.

Theoretical and practical considerations thus motivate us to consider a more flexible model that can incorporate both linear and nonlinear interactions, and without too many assumptions on the model parameters, as in (2.1).

2.2. Parameter estimation

Consider the PLVMICM model given in (2.1). Let $\boldsymbol{\theta} = (\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T)^T$, where $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_0^T, \boldsymbol{\alpha}_1^T)^T$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_0^T, \boldsymbol{\beta}_1^T)^T$. Let $\mathbf{V}_i = (\mathbf{X}_i, \mathbf{Z}_i, G_i)$, $i = 1, \ldots, n$, be the observations, and $\boldsymbol{\Theta}_\alpha$ and $\boldsymbol{\Theta}_\beta$ be the parameter spaces for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, respectively. In this section, we derive the detailed estimation procedure employing the BSBK method proposed by Wang and Yang (2007). Let $\mathcal{F}_n$ be the space of B-spline basis functions of order r ($r \geq 2$) (de Boor (2001)) with the B-spline basis $\mathbf{B}_r(u) = (B_{s,r}(u) : 1 \leq s \leq J_n)^T$, $u \in [a, b]$, where $J_n = N + r$ and $N = N_n$ is the number of interior knots for a knot sequence $\xi_1 = \cdots = 0 = \xi_r < \xi_{r+1} < \cdots < \xi_{r+N_n} < 1 = \xi_{r+N_n+1} = \cdots = \xi_{N_n+2r}$ in which N_n increases along with the sample size n. Then $m_l(u_l)$ with $u_l = u_l(\boldsymbol{\beta}_l) = \boldsymbol{\beta}_l^T\mathbf{X}$, $l = 0, 1$, can be approximated by a spline function,

$$\tilde{m}_l(u_l) \equiv \tilde{m}_l(u_l, \boldsymbol{\beta}) \approx \sum_{s=1}^{J_n} B_{s,r}(u_l)\lambda_{s,l}(\boldsymbol{\beta}) = \mathbf{B}_r^T(u_l)\lambda_l(\boldsymbol{\beta}),$$

where $\lambda_l(\boldsymbol{\beta}) = (\lambda_{s,l}(\boldsymbol{\beta}), 1 \leq s \leq J_n)^T$ and $\lambda(\boldsymbol{\beta}) = (\lambda_0(\boldsymbol{\beta})^T, \lambda_1(\boldsymbol{\beta})^T)$. For given $\boldsymbol{\beta}$, the B-spline coefficients $\lambda(\boldsymbol{\beta})$ and $\boldsymbol{\alpha}$ can be estimated as

$$(\widehat{\boldsymbol{\alpha}}^T, \widehat{\lambda}(\boldsymbol{\beta})^T)^T = \underset{\boldsymbol{\alpha} \in \boldsymbol{\Theta}_\alpha, \lambda(\boldsymbol{\beta}) \in \mathbb{R}^{2J_n}}{\operatorname{argmin}} \tilde{R}\left((\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T)^T, \lambda(\boldsymbol{\beta})\right),$$

where $\tilde{R}((\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T)^T, \lambda(\boldsymbol{\beta})) = \sum_{i=1}^n [Y_i - \tilde{m}_0(\boldsymbol{\beta}_0^T\mathbf{X}_i) - \boldsymbol{\alpha}_0^T\mathbf{Z}_i - (\tilde{m}_1(\boldsymbol{\beta}_1^T\mathbf{X}_i) - \boldsymbol{\alpha}_1^T\mathbf{Z}_i)G_i]^2$. Let $D_i(\tilde{\mathbf{Z}}_i, \boldsymbol{\beta}) = [\tilde{\mathbf{Z}}_i^T, (D_{i,sl}(\boldsymbol{\beta}_l), 1 \leq s \leq J_n, l = 0, 1)^T]^T$, where $\tilde{\mathbf{Z}}_i = (\mathbf{Z}_i^T, \mathbf{Z}_i^T G_i)^T$, $D_{i,s0}(\boldsymbol{\beta}_0) = B_{s,r}(\boldsymbol{\beta}_0^T\mathbf{X}_i)$ and $D_{i,s1}(\boldsymbol{\beta}_1) = B_{s,r}(\boldsymbol{\beta}_1^T\mathbf{X}_i)G_i$. Let $\mathbf{D}(\tilde{\mathbf{Z}}, \boldsymbol{\beta}) = (D_1(\tilde{\mathbf{Z}}_1, \boldsymbol{\beta}), \ldots, D_n(\tilde{\mathbf{Z}}_n, \boldsymbol{\beta}))^T$, an $n \times 2(q + J_n)$ matrix, and $Y = (Y_1, \ldots, Y_n)^T$, where $\tilde{\mathbf{Z}} = (\tilde{\mathbf{Z}}_1, \ldots, \tilde{\mathbf{Z}}_n)^T$ is an $n \times 2q$ matrix. Then the least squares estimators of $\boldsymbol{\alpha}$ and $\lambda(\boldsymbol{\beta})$ is

$$(\widehat{\boldsymbol{\alpha}}^T, \widehat{\lambda}(\boldsymbol{\beta})^T)^T = (\mathbf{D}(\tilde{\mathbf{Z}}, \boldsymbol{\beta})^T\mathbf{D}(\tilde{\mathbf{Z}}, \boldsymbol{\beta}))^{-1}\mathbf{D}(\tilde{\mathbf{Z}}, \boldsymbol{\beta})^T Y. \tag{2.2}$$

Once the B-spline coefficients $\lambda(\boldsymbol{\beta})$ are estimated, we can obtain the first derivative of the spline approximation of the nonparametric function as $\tilde{m}_l'(u_l) \equiv \tilde{m}_l'(u_l, \boldsymbol{\beta}) \approx \mathbf{B}_r'(u_l)^T\widehat{\lambda}_l(\boldsymbol{\beta})$, where $\mathbf{B}_r'(u_l)^T$ is the first derivative of $\mathbf{B}_r(u_l)$. Given the estimator $\widehat{\lambda}_l(\boldsymbol{\beta})$ in (2.2), we can estimate the loading parameters $\boldsymbol{\beta}$ by

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \boldsymbol{\Theta}_\beta}{\operatorname{argmin}} \tilde{R}\left((\widehat{\boldsymbol{\alpha}}^T, \boldsymbol{\beta}^T)^T, \widehat{\lambda}(\boldsymbol{\beta})\right).$$

Let $\widehat{\lambda}_l(\widehat{\boldsymbol{\beta}})$ be the estimators of the spline coefficients obtained by replacing $\mathbf{D}(\tilde{\mathbf{Z}}, \boldsymbol{\beta})$ with $\mathbf{D}(\tilde{\mathbf{Z}}, \widehat{\boldsymbol{\beta}})$ in (2.2). Based on the parametric estimator $\widehat{\boldsymbol{\theta}}$, it is easy to obtain the estimator of the nonparametric function $m_l(u_l)$ as

$$\tilde{m}_l(u_l, \widehat{\boldsymbol{\beta}}) = \mathbf{B}_r(u_l)^T \widehat{\lambda}_l(\widehat{\boldsymbol{\beta}}), l = 0, 1. \tag{2.3}$$

A detailed estimation algorithm is given in Supplementary Materials.

2.3. Kernel estimator of nonparametric functions

To obtain the asymptotic normality of the spline estimators for the nonparametric functions $m_l(u_l)$, $l = 0, 1$, as in Wang and Yang (2007), we use the BSBK estimator to establish their asymptotic normality. Define $\tilde{Y}_l = (\tilde{Y}_{1l}, \ldots, \tilde{Y}_{nl})^T$ as the new pseudo-responses, and their corresponding "oracle" responses as $Y_l^O = (Y_{1l}^O, \ldots, Y_{nl}^O)^T$, $l = 0, 1$. By using the B-spline estimators $\tilde{m}_l(\cdot)$ and the parametric estimators $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\alpha}}_0^T, \widehat{\boldsymbol{\alpha}}_1^T, \widehat{\boldsymbol{\beta}}_0^T, \widehat{\boldsymbol{\beta}}_1^T)^T$ of Section 2.2, we have

$$\tilde{Y}_{i1} = Y_i - \widehat{\boldsymbol{\alpha}}_0^T \mathbf{Z}_i - \tilde{m}_0(\widehat{\boldsymbol{\beta}}_0^T \mathbf{X}_i, \widehat{\boldsymbol{\beta}}) - \widehat{\boldsymbol{\alpha}}_1^T \mathbf{Z}_i G_i,$$

and

$$Y_{i1}^O = Y_i - \widehat{\boldsymbol{\alpha}}_0^T \mathbf{Z}_i - m_0(\widehat{\boldsymbol{\beta}}_0^T \mathbf{X}_i) - \widehat{\boldsymbol{\alpha}}_1^T \mathbf{Z}_i G_i.$$

Similarly, $\tilde{Y}_{i0}$ and Y_{i0}^O can be defined. In the "oracle" responses, the functions $m_l(\cdot)$ are assumed to be known.

Based on the new responses $\tilde{Y}_1$, we can obtain the BSBK estimator of $m_1(u_1)$ as $\widehat{m}_1(u_1, \widehat{\boldsymbol{\beta}}) = \widehat{a} + \widehat{b} u_1$ by local linear fitting, in which

$$(\widehat{a}, \widehat{b}) = \underset{a,b}{\operatorname{argmin}} \sum_{i=1}^{n} \left\{ \tilde{Y}_{i1} - a G_i - b(\widehat{\boldsymbol{\beta}}_1^T \mathbf{X}_i - u_1) G_i \right\}^2 K_{h_1}(\widehat{\boldsymbol{\beta}}_1^T \mathbf{X}_i - u_1),$$

where $K_h(t) = K(t/h)/h$ and $K(\cdot)$ is a kernel function and h is a bandwidth. By minimizing the weighted least squares, the estimator $\widehat{m}_1(u_1, \widehat{\boldsymbol{\beta}})$ has a closed form

$$\widehat{m}_1(u_1, \widehat{\boldsymbol{\beta}}) = (1, 0)[\tilde{\mathbf{X}}^T \mathbf{W} \tilde{\mathbf{X}}]^{-1} \tilde{\mathbf{X}}^T \mathbf{W} \tilde{Y}_1, \tag{2.4}$$

where

$$\tilde{\mathbf{X}} \equiv \tilde{\mathbf{X}}(u_1, \widehat{\boldsymbol{\beta}}_1) = \begin{pmatrix} G_1 & \cdots & G_n \\ (\widehat{\boldsymbol{\beta}}_1^T \mathbf{X}_1 - u_1)\frac{G_i}{h_1} & \cdots & (\widehat{\boldsymbol{\beta}}_1^T \mathbf{X}_n - u_1)\frac{G_n}{h_1} \end{pmatrix}^T,$$

$$\mathbf{W} \equiv \mathbf{W}(u_1, \widehat{\boldsymbol{\beta}}_1) = \operatorname{diag}\left\{ K_{h_1}(\widehat{\boldsymbol{\beta}}_1^T \mathbf{X}_1 - u_1), \ldots, K_{h_1}(\widehat{\boldsymbol{\beta}}_1^T \mathbf{X}_n - u_1) \right\}.$$

Similarly, we can also obtain the "oracle" kernel estimator of $m_1(u_1)$ as $\widehat{m}_1^O(u_1, \widehat{\beta}_1)$ based on new data Y_1^O by local linear fitting

$$\widehat{m}_1^O(u_1, \widehat{\boldsymbol{\beta}}) = (1, 0)[\tilde{\mathbf{X}}^T \mathbf{W} \tilde{\mathbf{X}}]^{-1} \tilde{\mathbf{X}}^T \mathbf{W} Y_1^O. \tag{2.5}$$

An outline of the algorithm can be found in Supplementary Materials. We use the BIC criterion to select the number of interior knots, while fixing the order of basis function as cubic to approximate the unknown functions, as described in Ma and Song (2015). The positions of interior knots are chosen as the uniform quantiles of $u_l^{(k)} = \mathbf{X}^T\widehat{\boldsymbol{\beta}}_l^{(k)}$ in the $(k+1)$-th step $(l = 0, 1, \ldots, L)$. Thus they change at each step while the number of knots remain fixed. This, however, does not affect the convergence of the algorithm in practice. To prove convergence of the algorithm with changes in knots is beyond the scope of this work. The BSBK estimator $\widehat{m}_l(u_l, \widehat{\boldsymbol{\theta}})$ is sensitive to the choice of bandwidth h_l, $l = 0, 1$. Bandwidth selection has been intensively studied, see Sepanski, Knickerbocker, and Carroll (1994) and Ruppert, Sheathers, and Wand (1995) for good discussions. To avoid the estimation of high order derivatives, we employ a bandwidth selector based on the mean squared error (MSE) criterion, called empirical bias bandwidth selection (EBBS) (Ruppert (1997); Carroll, Ruppert, and Welsh (1998); Liu, Jiang, and Zhou (2014(@)). The details of EBBS are provided in Supplementary Materials.

Remark 1. Cui, Härdle, and Zhu (2011) and Ma and Song (2015) relaxed the constraints $\|\boldsymbol{\beta}_l\|_2 = 1$ to $\|\boldsymbol{\beta}_{l,-1}\| < 1$ with $\boldsymbol{\beta}_{l,-1} = (\beta_{l2}, \ldots, \beta_{lp})^T$, $l = 0, 1$. We work directly on the equality constraints $\|\boldsymbol{\beta}_l\|_2 = 1$ which allows us to easily develop a Newton-Raphson algorithm. We can then test $H_0 : \beta_{lk} = 0$ for all $k = 1, \ldots, p$ (see Section 5 for a demonstration). In addition, the Newton-Raphson algorithm is faster than the nonlinear optimization method adopted in Ma and Song (2015), especially under nonlinear constraints.

2.4. Theoretical results

We need some additional notation to show the asymptotic normality of the estimator. Let $\boldsymbol{\theta}^0 = ((\boldsymbol{\alpha}^0)^T, (\boldsymbol{\beta}^0)^T)^T$ be the true parameter $\boldsymbol{\theta}$, where $\boldsymbol{\alpha}^0 = ((\boldsymbol{\alpha}_0^0)^T, (\boldsymbol{\alpha}_1^0)^T)^T$ and $\boldsymbol{\beta}^0 = ((\boldsymbol{\beta}_0^0)^T, (\boldsymbol{\beta}_1^0)^T)^T$. Let the space $\mathcal{M}$ be a collection of functions with finite L_2 norm on $[a_0, b_0] \times [a_1, b_1] \times \mathcal{R}$ with $\mathcal{M} = \{g(\mathbf{u}) = g_0(u_0) + g_1(u_1)G, Eg_l(u_l)^2 \leq \infty\}$, where $\mathbf{u} = (u_0, u_1)^T$. For $1 \leq k \leq q$, let $g_{Z_k}^0(\mathbf{u})$ be a maximizer in $\mathcal{M}$ for the optimization problem,

$$g_{Z_k}^0(U(\boldsymbol{\beta}^0)) = g_0^0(\mathbf{X}^T\boldsymbol{\beta}_0^0) + g_1^0(\mathbf{X}^T\boldsymbol{\beta}_1^0)G = \underset{g \in \mathcal{M}}{\operatorname{argmin}}\ E\{Z_k - g(U(\boldsymbol{\beta}^0))\}^2,$$

where $U(\boldsymbol{\beta}^0) = (\mathbf{X}^T\boldsymbol{\beta}_0^0, \mathbf{X}^T\boldsymbol{\beta}_1^0)^T$. Let $P_k(Z_k) = g_{Z_k}^0(U(\boldsymbol{\beta}^0))$ and $\mathbf{P}(\mathbf{Z}) = (P_1(Z_1), \ldots, P_q(Z_q))^T$. We take $\mathbf{P}(\mathbf{X}) = (P_1(X_1), \ldots, P_p(X_p))^T$ with $P_k(X_k) = g_{X_k}^0(U(\boldsymbol{\beta}^0))$. Let $\widehat{\mathbf{Z}} = \mathbf{Z} - \mathbf{P}(\mathbf{Z})$, $\widehat{\mathbf{X}} = \mathbf{X} - \mathbf{P}(\mathbf{X})$ and $\boldsymbol{\phi}(\mathbf{V}, \boldsymbol{\beta}^0) = (\boldsymbol{\phi}_1(\mathbf{V}, \boldsymbol{\beta}^0)^T, \boldsymbol{\phi}_2(\mathbf{V}, \boldsymbol{\beta}^0)^T)^T$, where $\boldsymbol{\phi}_1(\mathbf{V}, \boldsymbol{\beta}^0) = (\widehat{\mathbf{Z}}^T, \widehat{\mathbf{Z}}^T G)^T$ and $\boldsymbol{\phi}_2(\mathbf{V}, \boldsymbol{\beta}^0) = ([m_0'(\mathbf{X}^T\boldsymbol{\beta}^0)\widehat{\mathbf{X}}]^T, [m_1'(\mathbf{X}^T\boldsymbol{\beta}^0)\widehat{\mathbf{X}}G]^T)^T$. Define the covariance matrix of $\boldsymbol{\theta}^0$ as

$$\Sigma = \left\{E[\boldsymbol{\phi}(\mathbf{V}, \boldsymbol{\beta}^0)^{\otimes 2}]\right\}^{-1} \left\{E[\sigma(\mathbf{V})^2\boldsymbol{\phi}(\mathbf{V}, \boldsymbol{\beta}^0)^{\otimes 2}]\right\} \left\{E[\boldsymbol{\phi}(\mathbf{V}, \boldsymbol{\beta}^0)^{\otimes 2}]\right\}^{-1},$$

where $\boldsymbol{\zeta}^{\otimes 2} = \boldsymbol{\zeta}\boldsymbol{\zeta}^T$ for any vector $\boldsymbol{\zeta}$. Σ can be simplified as $\Sigma = \sigma_0^2 \left\{E[\boldsymbol{\phi}(\mathbf{V}, \boldsymbol{\beta}^0)^{\otimes 2}]\right\}^{-1}$ if the error variance $\sigma(\mathbf{V})$ is a constant σ_0^2.

Theorem 1. *If assumptions* (A.1)$-$(A.4) *in the Appendix hold, and* $nN^{-4} \to \infty$ *and* $nN^{-2r-2} \to 0$, *then* $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\|_2 = O_p(n^{-1/2})$, *and as* $n \to \infty$, $n^{1/2}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\right) \xrightarrow{\mathcal{L}} N(\mathbf{0}, \Sigma)$.

Theorem 2. *If assumptions* (A.1)$-$(A.4) *in the Appendix hold, and* $nN^{-4} \to \infty$ *and* $nN^{-2r-2} \to 0$, *then for* $l = 0, 1$,

$$\sup_{u_l \in [a_l, b_l]} |\tilde{m}_l(u_l, \widehat{\boldsymbol{\beta}}) - m_l(u_l)| = O_p((\frac{N}{n})^{1/2} + N^{-r}),$$

where $\tilde{m}_l(u_l, \widehat{\boldsymbol{\beta}})$ *is given in* (2.3), *and* $m_l(\cdot)$ *is the true function.*

Next we show that the order of the asymptotic uniform magnitude of the difference between the BSBK estimator $\widehat{m}_l(u_l, \widehat{\boldsymbol{\beta}})$ and its "oracle" version $\widehat{m}_l^O(u_l, \widehat{\boldsymbol{\beta}})$ is $o_p(n^{-2/5})$, so $\widehat{m}_l(u_l, \widehat{\boldsymbol{\beta}})$ and $\widehat{m}_l^O(u_l, \widehat{\boldsymbol{\beta}})$ share the same asymptotic distribution.

Theorem 3. *If assumptions* (A.1)$-$(A.6) *in the Appendix hold, and* $nN^{-4} \to \infty$ *and* $nN^{-\delta} \to 0$ *with* $\delta = \min(2r+2, 5r/2)$, *then for* $l = 0, 1$,

$$\sup_{u_l \in [a_l, b_l]} |\widehat{m}_l(u_l, \widehat{\boldsymbol{\beta}}) - \widehat{m}_l^O(u_l, \widehat{\boldsymbol{\beta}})| = o_p(n^{-2/5}).$$

Set $\mu_k = \int t^k K(t)dt$, $\nu_k = \int t^k K^2(t)dt$. The consistency and asymptotic normality of the unknown functions $m_0(\cdot)$ and $m_1(\cdot)$ now follow.

Theorem 4. *If assumptions* (A.1)$-$(A.6) *in the Appendix hold, and* $nN^{-4} \to \infty$ *and* $nN^{-2r-2} \to 0$, *then, for* $l = 0, 1$,

$$(nh_l)^{1/2}\left\{\widehat{m}_l(u_l, \widehat{\boldsymbol{\beta}}) - m_l(u_l) - b_l(u_l)h_l^2\right\} \xrightarrow{\mathcal{L}} N(\mathbf{0}, v_l(u_l)), \ \textit{as } n \to \infty,$$

where $b_l(u_l) = \mu_1 m_l''(u_l)/2, l = 0, 1$, $v_0(u_0) = f_0(u_0)^{-1}\nu_0 E\left[\sigma^2(\mathbf{V})|\mathbf{X}^T\boldsymbol{\beta}_0^0 = u_0\right]$, *and* $v_1(u_1) = f_1(u_1)^{-1}\nu_0 E\left[G^2\sigma^2(\mathbf{V})|\mathbf{X}^T\boldsymbol{\beta}_1^0 = u_1\right] / \left(E[G^2|\mathbf{X}^T\boldsymbol{\beta}_1^0 = u_1]\right)^2$.

If $\sigma^2(\mathbf{V}) = \sigma_0^2$, the variance $v_l(u_l)$ can be simplified as $f_l(u_l)^{-1}\nu_0\sigma_0^2$ for $l = 0, 1$.

3. Hypothesis Tests

3.1. Testing for nonparametric components

Our model can assess the interaction of the combined effect of multiple environmental exposures with genes. This can be achieved by testing the nonparametric component $m_1(\cdot)$ to discover the change trend of the interaction of the combined environmental effect. We consider a test to detect whether $m_1(u_1)$ is a linear function $m_1^0(u_1) = \delta_0 + \delta_1 u_1$,

$$H_0 : m_1(\cdot) = m_1^0(\cdot) \text{ v.s. } H_1 : m_1(\cdot) \neq m_1^0(\cdot), \tag{3.1}$$

via a generalized likelihood ratio (GLR) test (Fan, Zhang, and Zhang (2001); Liang et al. (2010); Ma and Song (2015)). Rejecting H_0 indicates statistical evidence of nonlinear interaction between G and multiple environmental mixtures. If we fail to reject H_0, we can further assess whether there exists a genetic effect as well as linear interaction effect between a gene and multiple environmental exposures by fitting a parametric linear interaction model.

Remark 2. In addition to the linear hypothesis, we are interested in testing $H_0 : m_1(\cdot) = 0$ or $H_0 : m_1(\cdot) = c$ where c is a constant. Testing for zero or constant effect can be done under the varying-coefficient model proposed in Ma et al. (2011), this cannot be done in the current model setup due to the fact that the loading parameters $\boldsymbol{\beta}_1$ are not identifiable under the above nulls. If we fail to reject the null in hypotheses (3.1), we can fit a linear interaction model as $Y = m_0(\boldsymbol{\beta}_0^T\mathbf{X}) + \boldsymbol{\alpha}_0^T\mathbf{Z} + (\delta_0 + \boldsymbol{\beta}_1^T\mathbf{X} + \boldsymbol{\alpha}_1^T\mathbf{Z})G + \varepsilon$, where no constraints on $\boldsymbol{\beta}_1$ are imposed. Then one can proceed to test $H_0^L : \delta_0 = \boldsymbol{\beta}_1 = \boldsymbol{\alpha}_1 = 0$ to assess the overall effect of G on Y. One can continue to assess the marginal effect of G on Y and the interaction effect between G and $\mathbf{X}$ or $\mathbf{Z}$ if H_0^L is rejected.

Consider (3.1). Let $\widehat{\boldsymbol{\theta}}$ be the BSBK estimate of $\boldsymbol{\theta}$ proposed in Section 2.2. Let $\widehat{m}_{l,H_0}(u_l)$ and $\widehat{m}_{l,H_1}(u_l)$ be the estimators under H_0 and H_1, respectively. Let the residual sums of squares under H_0 and H_1 in (3.1) be $\mathrm{RSS}_1(H_0) = \sum_{i=1}^n \left\{\widehat{Y}_i - \widehat{m}_{0,H_0}(\widehat{\boldsymbol{\beta}}_0^T\mathbf{X}_i) - \widehat{m}_{1,H_0}(\widehat{\boldsymbol{\beta}}_1^T\mathbf{X}_i)G_i\right\}^2$ and
$\mathrm{RSS}_1(H_1) = \sum_{i=1}^n \left\{\widehat{Y}_i - \widehat{m}_{0,H_1}(\widehat{\boldsymbol{\beta}}_0^T\mathbf{X}_i) - \widehat{m}_{1,H_1}(\widehat{\boldsymbol{\beta}}_1^T\mathbf{X}_i)G_i\right\}^2$, where $\widehat{Y}_i = Y_i - \widehat{\boldsymbol{\alpha}}^T\tilde{\mathbf{Z}}_i$. We define the generalized likelihood ratio (GLR) test statistic as

$$T_1 = \frac{n}{2}\frac{\mathrm{RSS}_1(H_0) - \mathrm{RSS}_1(H_1)}{\mathrm{RSS}_1(H_1)}. \tag{3.2}$$

Let $a_K = \{K(0) - 1/2\int K^2(u)du\}\left[\int\{K(u) - 1/2K * K(u)\}du\right]^{-1}$, where $K * K(u)$ denotes the convolution of K. Denote by Ω_l the support of $\boldsymbol{\beta}_l^T\mathbf{x}$, and by $|\Omega_l|$ the length of Ω_l, $l = 0, 1$.

Theorem 5. *If assumptions* (A.1)−(A.6) *in the Appendix hold, and* $nN^{-4} \to \infty$ *and* $nN^{-2r-2} \to 0$*, then under* H_0 *in* (3.1)*, when* $m_1^0(u_1)$ *is a linear function of* u_1*,*

$$\sigma_{1n}^{-1}(T_1 - \mu_{1n}) \xrightarrow{\mathcal{L}} N(0,1),$$

where $\sigma_{1n}^2 = (2/h_1)|\Omega_1|\int\{K(u) - 1/2K * K(u)\}^2du$, *and* $\mu_{1n} = (1/h_1)|\Omega_1|\{K(0) - 1/2\int K^2(u)du\}$. *Furthermore,* $a_K T_1 \overset{a}{\sim} \chi^2_{d_1}$, *where* $d_1 = a_K\mu_{1n}$.

When assessing the linear form of the function, $RSS_1(H_0)$ and $RSS_1(H_1)$ can be calculated by first getting the estimators of $m_0(\cdot)$ and $m_1(\cdot)$ using the B-spline method under the null and alternative hypotheses. The B-spline estimators under H_0 are given by $\tilde{m}_{0,H_0}(u_0) = \mathbf{B}_r^T(u_0)\widehat{\lambda}_0$ and $\widehat{m}_{1,H_0}(u_1) = \widehat{\delta}_0 + \widehat{\delta}_1 u_1$, where $\widehat{\delta}_0$, $\widehat{\delta}_1$, and $\widehat{\lambda}_0$ are the ordinary least squares estimators of δ_0, δ_1, and λ_0. Then, we can obtain the kernel estimator $\widehat{m}_{0,H_0}(u_0)$ based on the new data $(\widehat{Y}_{H_0}, \mathbf{X}, \mathbf{Z}, G)$ and $\widehat{u}_0 = \widehat{\boldsymbol{\beta}}_0^T\mathbf{X}$, using the arguments in Section 2.3, where $\widehat{Y}_{H_0} = (\widehat{Y}_{1,H_0}, \ldots, \widehat{Y}_{n,H_0})^T$ and $\widehat{Y}_{i,H_0} = Y_i - \boldsymbol{\alpha}^T\tilde{\mathbf{Z}}_i - \widehat{m}_{1,H_0}(\widehat{\boldsymbol{\beta}}_1^T\mathbf{X}_i)$. Here $\widehat{m}_{0,H_1}(\cdot)$ and $\widehat{m}_{1,H_1}(\cdot)$ are the BSBK estimators which can be obtained as in (2.4).

To illustrate the testing for the case with $l > 1$, we consider a model with two genetic variables G_1 and G_2,

$$Y = m_0(\boldsymbol{\beta}_0^T\mathbf{X}) + \boldsymbol{\alpha}_0^T\mathbf{Z} + \{m_1(\boldsymbol{\beta}_1^T\mathbf{X}) + \boldsymbol{\alpha}_1^T\mathbf{Z}\}G_1 + \{m_2(\boldsymbol{\beta}_2^T\mathbf{X}) + \boldsymbol{\alpha}_2^T\mathbf{Z}\}G_2 + \varepsilon. \tag{3.3}$$

One can simultaneously test $m_1(\cdot)$ and $m_2(\cdot)$, for example, testing

$$H_0 : m_1(\cdot) = m_1^0(\cdot), m_2(\cdot) = m_2^0(\cdot) \text{ v.s. } H_1 : m_1(\cdot) \neq m_1^0(\cdot) \text{ or } m_2(\cdot) \neq m_2^0(\cdot), \tag{3.4}$$

where $m_1^0(\cdot)$ and $m_2^0(\cdot)$ are linear functions. Similarly, we can construct the corresponding GLR test statistic

$$T_2 = \frac{n}{2}\frac{\{\mathrm{RSS}_2(H_0) - \mathrm{RSS}_2(H_1)\}}{\mathrm{RSS}_2(H_1)}, \tag{3.5}$$

where

$$\mathrm{RSS}_2(H_0) = \sum_{i=1}^{n}\left\{\widehat{Y}_i - \widehat{m}_{0,H_0}(\widehat{\boldsymbol{\beta}}_0^T\mathbf{X}_i) - \widehat{m}_{1,H_0}(\widehat{\boldsymbol{\beta}}_1^T\mathbf{X}_i)G_{i1} - \widehat{m}_{2,H_0}(\widehat{\boldsymbol{\beta}}_2^T\mathbf{X}_i)G_{i2}\right\}^2,$$

$$\mathrm{RSS}_2(H_1) = \sum_{i=1}^{n}\left\{\widehat{Y}_i - \widehat{m}_{0,H_1}(\widehat{\boldsymbol{\beta}}_0^T\mathbf{X}_i) - \widehat{m}_{1,H_1}(\widehat{\boldsymbol{\beta}}_1^T\mathbf{X}_i)G_{i1} - \widehat{m}_{2,H_1}(\widehat{\boldsymbol{\beta}}_2^T\mathbf{X}_i)G_{i2}\right\}^2,$$

and $\widehat{Y}_i = Y_i - \widehat{\boldsymbol{\alpha}}_0^T\mathbf{Z}_i - \widehat{\boldsymbol{\alpha}}_1^T\mathbf{Z}_iG_{i1} - \widehat{\boldsymbol{\alpha}}_2^T\mathbf{Z}_iG_{i2}$. Note that $\widehat{m}_{l,H_0}(\widehat{\boldsymbol{\beta}}_l^T\mathbf{X}_i)$, $l = 0, 1, 2$, are different from those in T_1, but the estimation is similar.

Theorem 6. *If assumptions* (A.1)−(A.6) *in the Appendix hold,* $nN^{-4} \to \infty$ *and* $nN^{-2r-2} \to 0$*, then under* H_0 *in* (3.4)*, when* $m_1^0(u_1)$ *and* $m_2^0(u_2)$ *are linear functions,*

$$\sigma_{2n}^{-1}(T_2 - \mu_{2n}) \xrightarrow{\mathcal{L}} N(0,1),$$

where $\sigma_{2n}^2 = 2b_n\int\{K(u) - 1/2K * K(u)\}^2du$, $\mu_{2n} = b_n\left\{K(0) - 1/2\int K^2(u)du\right\}$ *and* $b_n = \sum_{l=1,2}|\Omega_l|/h_l$. *Furthermore,* $a_K^*T_2 \overset{a}{\sim} \chi^2_{d_2}$*, where* $d_2 = a_K^*\mu_{2n}$ *with* $a_K^* = 2\mu_{2n}/\sigma_{2n}^2$.

Remark 3. The formulation of asymptotic normality in Theorem 6 is that in Fan and Jiang (2005). Theorem 6 can be generalized to cases where three or more genetic variables can be fitted and tested ($l \geq 3$). One can apply Theorem 6 for simultaneous inference on the functions of some components of varying index coefficients. While the asymptotic results for T_1 and T_2 are available, they may not perform well when sample sizes are small. We recommend the conditional bootstrap method (Cai, Fan, and Li (2000); Fan, Zhang, and Zhang (2001)) in applications.

3.2. Testing parametric components

We are also interested in assessing the interaction effects of genes with discrete environments. This can be addressed via parametric hypothesis testing. Furthermore, if there is G×E interaction, one may be interested in testing which index coefficients contribute to the joint effect. This results in another parametric hypothesis testing problem. We consider a class of general hypothesis testing problems with

$$H_0 : \mathbf{A}\boldsymbol{\zeta} = \gamma \ \text{ v.s. } \ H_1 : \mathbf{A}\boldsymbol{\zeta} \neq \gamma, \tag{3.6}$$

where $\mathbf{A}$ is a known $k \times (q+s)$ full-rank matrix, s is the number of elements in $S \subset \{1, \ldots, p\}$, $\boldsymbol{\zeta} = (\boldsymbol{\alpha}_1^T, \boldsymbol{\beta}_S^T)^T$ with $\boldsymbol{\beta}_S = (\beta_{j_1}, \ldots, \beta_{j_s})^T$, $j_l \in S$, and γ is a k-dimensional vector. For a special case, we can detect whether $\boldsymbol{\alpha}_1$ and $\boldsymbol{\beta}_S$ are zeros by taking

$$H_0 : \boldsymbol{\alpha}_1 = \mathbf{0}, \boldsymbol{\beta}_S = \mathbf{0} \ \text{ v.s. } H_1 : \boldsymbol{\alpha}_1 \neq \mathbf{0} \ \text{ or } \ \boldsymbol{\beta}_S \neq \mathbf{0}. \tag{3.7}$$

Let $\boldsymbol{\theta}_{H_0} = (\boldsymbol{\alpha}_{0,H_0}^T, \boldsymbol{\alpha}_{1,H_0}^T, \boldsymbol{\beta}_{0,H_0}^T, \boldsymbol{\beta}_{1,H_0}^T)^T$ be the parameters corresponding to $\boldsymbol{\theta}$ under H_0 in (3.7) and $\boldsymbol{\theta}_{H_1} = (\boldsymbol{\alpha}_{0,H_1}^T, \boldsymbol{\alpha}_{1,H_1}^T, \boldsymbol{\beta}_{0,H_1}^T, \boldsymbol{\beta}_{1,H_1}^T)^T$ be the counterparts under H_1. Define the residual sums of squares under H_0 and H_1 as

$$\begin{aligned} R_{H_0} = & \sum_{i=1}^n \Big\{ Y_i - \widehat{m}_{0,H_0}(\widehat{\boldsymbol{\beta}}_{0,H_0}^T \mathbf{X}_i, \widehat{\boldsymbol{\beta}}_{H_0}) - \widehat{\boldsymbol{\alpha}}_{0,H_0}^T \mathbf{Z}_i - (\widehat{m}_{1,H_0}(\widehat{\boldsymbol{\beta}}_{1,H_0}^T \mathbf{X}_i, \widehat{\boldsymbol{\beta}}_{H_0}) \\ & - \widehat{\boldsymbol{\alpha}}_{1,H_0}^T \mathbf{Z}_i) G_i \Big\}^2, \\ R_{H_1} = & \sum_{i=1}^n \Big\{ Y_i - \widehat{m}_{0,H_1}(\widehat{\boldsymbol{\beta}}_{0,H_1}^T \mathbf{X}_i, \widehat{\boldsymbol{\beta}}_{H_1}) - \widehat{\boldsymbol{\alpha}}_{0,H_1}^T \mathbf{Z}_i - (\widehat{m}_{1,H_1}(\widehat{\boldsymbol{\beta}}_{1,H_1}^T \mathbf{X}_i, \widehat{\boldsymbol{\beta}}_{H_1}) \\ & - \widehat{\boldsymbol{\alpha}}_{1,H_1}^T \mathbf{Z}_i) G_i \Big\}^2, \end{aligned}$$

where $\widehat{\boldsymbol{\theta}}_{H_0}$ and $\widehat{\boldsymbol{\theta}}_{H_1}$ are the estimators of $\boldsymbol{\theta}$ under H_0 and H_1 proposed in Section 2.2, and $\widehat{m}_{l,H_0}(\cdot)$ and $\widehat{m}_{l,H_1}$ are estimators of $m_l(\cdot)$ proposed in (2.4) under H_0 and H_1, $l = 0, 1$, respectively. We take the test statistic

$$T_3 = \frac{n\{R_{H_0} - R_{H_1}\}}{R_{H_1}}. \tag{3.8}$$

Theorem 7. *If assumptions* (A.1)−(A.6) *in the Appendix hold,* $nN^{-4} \to \infty$ *and* $nN^{-2r-2} \to 0$*, then when* $\sigma(\mathbf{V})$ *is a constant* σ_0^2*,*

(i) *under* H_0 *in* (3.6)*,* $T_3 \xrightarrow{\mathcal{L}} \chi_k^2$*;*

(ii) *under* H_1 *in* (3.6)*,* T_3 *converges to a noncentral* χ^2 *distribution with* k *degrees of freedom with noncentrality parameter* $\phi = \lim_{n\to\infty} n\sigma^2(\mathbf{A}\zeta-\gamma)^T(\mathbf{A}\Sigma^{-1}\mathbf{A})^{-1}(\mathbf{A}\zeta-\gamma)$*, where* Σ *is defined as in Theorem* 1*.*

4. Monte Carlo Simulation

The finite sample performance of the proposed method was evaluated by simulation studies. Under model (2.1), we generated continuous X variables X_1, X_2, X_3 as independent uniform $U(0,1)$ and discrete Z variables Z_1, Z_2 as independent Bernoulli $Ber(1,0.5)$. The genetic variable G was coded as $(2,1,0)$ corresponding to genotypes (AA, Aa, aa). We set the minor allele frequency (MAF) $p_A = (0.1, 0.3, 0.5)$ and assumed Hardy-Weinberg equilibrium. SNP genotypes AA, Aa, and aa were simulated from a multinomial distribution with frequencies p_A^2, $2p_A(1-p_A)$ and $(1-p_A)^2$, respectively. The error term ε was normal $N(0, 0.1)$.

We set $m_0(u) = \cos(\pi u)$ and $m_1(u) = \sin\{\pi(u-A)/(B-A)\}$ with $A = \sqrt{3}/2 - 1.645/\sqrt{12}$ and $B = \sqrt{3}/2 + 1.645/\sqrt{12}$, and $\boldsymbol{\beta}_0 = (\sqrt{5}, \sqrt{4}, \sqrt{4})/\sqrt{13}$, $\boldsymbol{\beta}_1 = (1,1,1)/\sqrt{3}$, $\boldsymbol{\alpha}_0 = (0.5, 0.5)^T$, and $\boldsymbol{\alpha}_1 = (0.3, 0.3)^T$. We drew 1,000 data sets with sample size $n = 200, 500$. The Epanechnikov kernel $K(t) = 0.75(1-t^2)_+$ was chosen to localize the unknown functions $m_0(\cdot)$ and $m_1(\cdot)$. The suitable smoothing bandwidths for estimating both functions were selected using the EBBS method described in Section 2.3. The number of interior knots N_k was selected by the BIC method.

4.1. Performance of estimation

Table 1 summarizes the average bias of the estimators (Bias), the standard deviation of the 1,000 estimators (SD), the average of the estimated standard errors (SE) based on the theoretical calculation, and the estimated coverage probability (CP) at the nominal 95% confidence level for the parameters. In general, the coverage probability for all the parameters was close to 95% and reasonably controlled. As the sample size increased, the performance of the parameter estimators improved. We observed consistently smaller SD and SE when n increased from 200 to 500. The same trend was observed when n increased to 1,000 (see Supplementary Materials for more details). The parameter estimators for the interaction effects $(\boldsymbol{\beta}_1, \boldsymbol{\alpha}_1)$ improved as MAF increased. For example, the SD of $\widehat{\beta}_{11}$ went from 0.028 to 0.012 when MAF increased from 0.1 to 0.5 under a

Table 1. Simulation results for $p_A = 0.1, 0.3, 0.5$ with sample size $n = 200, 500$.

			$p_A = 0.1$				$p_A = 0.3$				$p_A = 0.5$			
n	Param	True	Bias	SD	SE	CP	Bias	SD	SE	CP	Bias	SD	SE	CP
200	α_{01}	0.500	4.4E-04	0.016	0.016	95.2	3.1E-04	0.020	0.020	95.2	9.9E-04	0.026	0.026	95.1
	α_{02}	0.500	-1.6E-04	0.016	0.016	95.3	4.1E-04	0.020	0.020	95.3	5.6E-04	0.026	0.026	95.8
	α_{11}	0.300	9.4E-05	0.040	0.039	94.1	6.0E-04	0.024	0.024	94.1	6.7E-05	0.022	0.022	95.2
	α_{12}	0.300	-1.1E-03	0.040	0.039	95.0	-1.1E-03	0.023	0.024	95.9	-4.4E-04	0.021	0.022	96.3
	β_{01}	0.620	-3.7E-04	0.011	0.011	94.7	-1.7E-03	0.012	0.013	94.8	-2.1E-03	0.014	0.014	94.5
	β_{02}	0.555	3.3E-04	0.012	0.012	95.3	1.0E-03	0.013	0.013	96.4	1.5E-03	0.014	0.015	96.6
	β_{03}	0.555	-2.7E-04	0.012	0.012	94.0	4.2E-04	0.013	0.013	95.3	3.1E-04	0.015	0.015	95.4
	β_{11}	0.577	1.4E-03	0.028	0.027	92.9	-4.0E-04	0.015	0.015	95.5	-7.5E-05	0.012	0.012	95.1
	β_{12}	0.577	-3.4E-04	0.029	0.028	93.5	9.5E-05	0.015	0.015	95.3	2.9E-04	0.011	0.012	96.2
	β_{13}	0.577	-3.2E-03	0.028	0.027	94.3	-2.6E-04	0.015	0.015	96.1	-5.7E-04	0.012	0.012	96.0
500	α_{01}	0.500	-3.2E-04	0.010	0.010	95.8	-5.5E-04	0.012	0.012	95.2	-4.0E-04	0.016	0.016	96.1
	α_{02}	0.500	1.9E-04	0.010	0.010	94.1	2.0E-04	0.013	0.012	94.2	3.8E-04	0.016	0.016	94.6
	α_{11}	0.300	5.6E-04	0.023	0.022	93.7	9.9E-04	0.015	0.014	93.8	6.5E-04	0.013	0.013	94.5
	α_{12}	0.300	1.2E-05	0.023	0.022	94.0	2.6E-04	0.015	0.014	93.8	2.0E-04	0.013	0.013	94.1
	β_{01}	0.620	-4.6E-04	0.007	0.007	95.2	-1.0E-03	0.008	0.008	95.7	-1.2E-03	0.009	0.009	94.9
	β_{02}	0.555	1.2E-04	0.007	0.007	95.5	4.3E-04	0.008	0.008	95.1	5.5E-04	0.009	0.009	95.1
	β_{03}	0.555	2.6E-04	0.007	0.007	94.2	5.2E-04	0.008	0.008	94.1	5.2E-04	0.009	0.009	94.4
	β_{11}	0.577	5.2E-04	0.015	0.016	95.0	3.0E-05	0.009	0.009	96.6	-8.5E-06	0.007	0.007	95.9
	β_{12}	0.577	-3.4E-04	0.016	0.016	94.0	-8.0E-06	0.009	0.009	95.6	1.0E-04	0.007	0.007	96.3
	β_{13}	0.577	-8.3E-04	0.016	0.016	94.5	-2.3E-04	0.009	0.009	95.2	-2.3E-04	0.007	0.007	94.8

fixed sample size $n = 200$. However, the estimators for the main effects $(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0)$ showed an opposite direction due to limited data information to estimate these parameters when MAF increased. This is due to the fact that the amount of data used to estimate these parameters is proportional to $(1 - p_A)^2$.

Figure 1 shows the plot of the estimators of $m_1(u_1)$, and its corresponding confidence bands under different sample sizes and MAFs in the interval of u_1 from 0.25 to 1.25. It can be there seen that the estimated curves almost overlap with the corresponding true curves, and the confidence bands are very tight, especially under large MAF and sample size. We also plotted the estimate of $m_0(\cdot)$, see the Supplementary Materials.

4.2. Performance of hypothesis tests

We first evaluated the performance of the test for the nonparametric function under the 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. Power was evaluated under a sequence of alternative models indexed by τ, $H_1^\tau : m_1^\tau(\cdot) = m_1^0(\cdot) + \tau\{m_1(\cdot) - m_1^0(\cdot)\}$. When $\tau = 0$, the test results provide the false positive rates. The null model corresponds to a linear G×E effect.

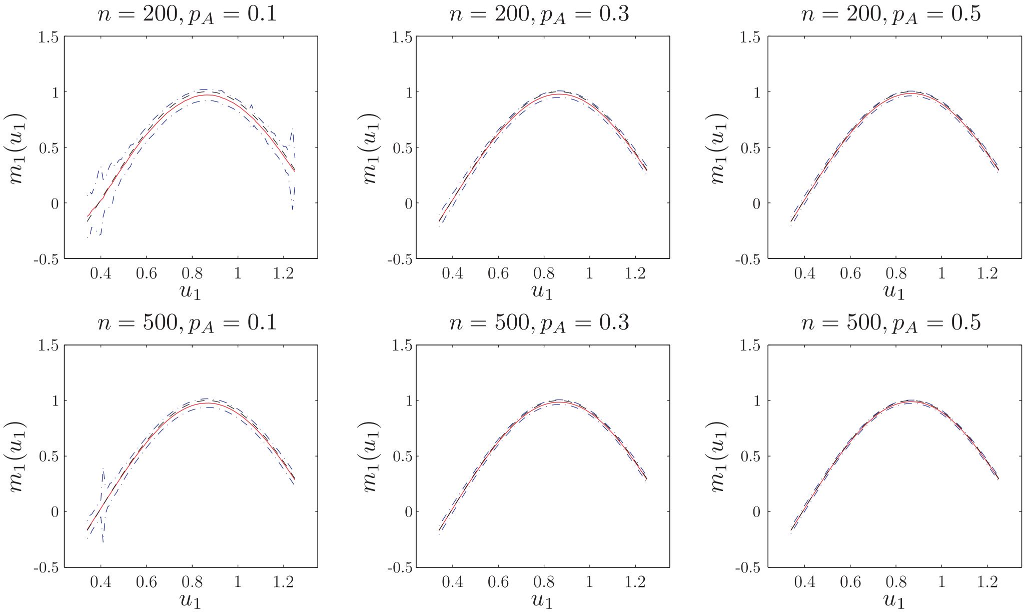

Figure 1. The estimation of function $m_1(\cdot)$ under different MAFs and sample sizes. The estimated and true functions are denoted by the solid and dashed lines, respectively. The 95% confidence band is denoted by the dotted-dash line.

Figure 2 shows the size ($\tau = 0$) and power function ($\tau > 0$) at significance level 0.05 based on 500 Monte Carlo simulations each with 500 bootstrap samples under sample sizes $n = 200, 500$, 1,000. The empirical type I errors under the three scenarios are very close to the nominal level 0.05. We observed dramatic power increase when MAF increased from 0.1 to 0.3 in all scenarios. The results indicate that our method can reasonably control the false positives and has appropriate power to detect genetic difference. We also considered the PLVMICM model in (3.3) with two genetic components and tested if both $m_1(\cdot)$ and $m_2(\cdot)$ are simultaneously linear, following Theorem 6. The results are in the Supplemental Materials.

To check the performance of the interaction test between G and discrete variable $\mathbf{Z}$, under model (2.1), we considered the hypothesis $H_0 : \boldsymbol{\alpha}_1 = 0$. The power of the test was evaluated under a sequence of alternatives indexed by τ, $H_1^\tau : \boldsymbol{\alpha}_1^\tau = \tau\boldsymbol{\alpha}_1$. Data were simulated as in the previous section. Figure 3 depicts the empirical size ($\tau = 0$) and power functions ($\tau > 0$) under different sample sizes and MAFs at the 0.05 significance level. As expected, the power and size improve as MAF and sample size increase. Under low MAF ($p_A = 0.1$), the size is a little inflated when n is small (200 and 500), but is well controlled when n increases to 1,000. As tith the nonparametric test, dramatic power improvement

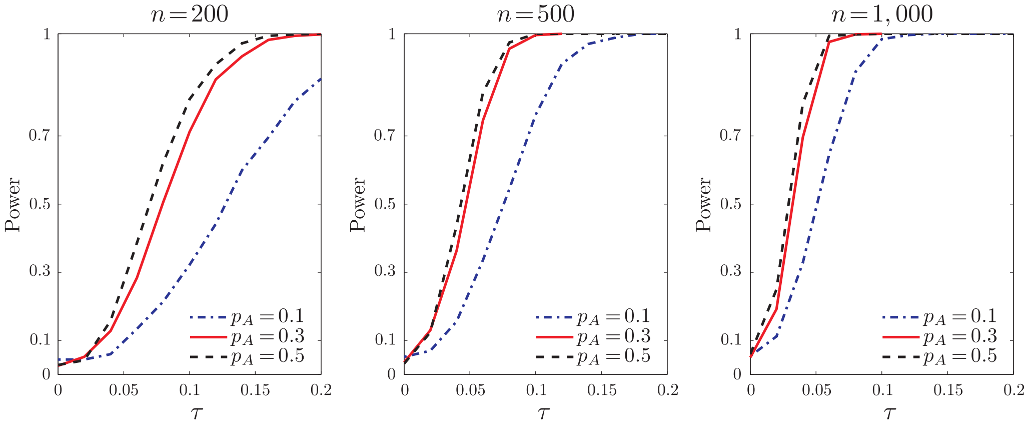

Figure 2. The empirical size and power function of testing nonparametric function $m_1(\cdot)$ under different sample sizes and MAFs.

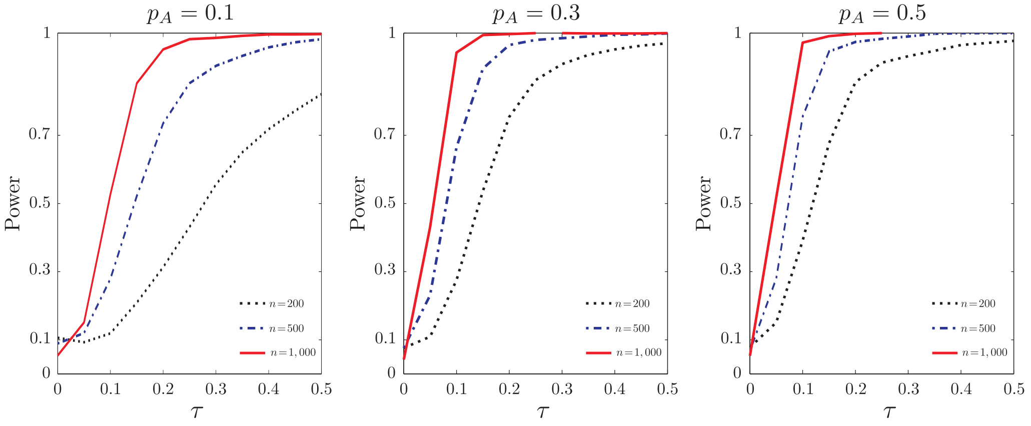

Figure 3. The empirical size and power functions of testing $H_0 : \boldsymbol{\alpha}_1 = 0$ under differen sample sizes and MAFs.

is observed when MAF increases from 0.1 to 0.3. The power difference between MAF=0.3 and MAF=0.5 is small indicating good performance of the test.

5. A Case Study

We applied the proposed PLVMICM model to a data set from the Gene Environment Association Studies initiative (GENEVA, `http://www.genevastudy.org`) funded by the trans-NIH Genes, Environment, and Health Initiative (GEI), to show the utility of the method. Low and high birth weights are not only the major causes of neonatal morbidity and mortality, but are also related to increased risk of metabolic diseases later in life. Fetal growth is determined by fetal genes as well as complex interactions between fetal genes and the maternal uterine environment. We focused on the Thai population with 1,126 subjects genotyped with the Omni1-Quad_v1-0_B platform after removing outliers. After regressing the baby's body weight on twelve environmental variables, including nine continuous and three discrete variables, five continuous variables and one discrete variable remained significant at the 0.0001 significance level. Three of the five continuous variables were chosen, including mother's mean OGTT diastolic blood pressure (denoted as X_1), mother's one hour OGTT glucose level (denoted as X_2), and mother's mean OGTT systolic blood pressure (denoted as X_3). The discrete variable, denoted as Z, is baby's gender. To show the utility of the method, we picked one candidate gene *CDKAL1* for a demonstration. The gene is located on chromosome 6 and contains 192 SNPs after removing those with MAF$<$ 0.05. Low birth weight has been shown to be associated with high risk in type 2 diabetes later in life. Evidence of genetic studies on type 2 diabetes loci suggests that this gene is associated with reduced birth weight in Caucasian populations (Zhao et al. (2009); Andersson et al. (2010)). Our goal is to evaluate whether this gene also functions in the Thai population and, if so, how SNPs in the gene interact with mother's condition (considered as environment) to affect birth weight and further determine the interaction mechanism.

We first tested whether any SNP is associated with birth weight based on the nonparametric test of $H_0 : m_1(u_1) = \delta_0 + \delta_1 u_1$ with p-value denoted by p_{m_1}. Since we tested each SNP individually, we applied a simple multiple testing correction method. We first calculated the effective number of tests E_0 by using the Cheverud estimation method, given by $E_0 = 1 + L^{-1}\sum_{i,j=1}^{L}(1 - r_{ij}^2)$, where $L = 192$ is the total number of SNPs and r_{ij} are the pairwise correlation coefficients of SNPs (Cheverud (2001)). The estimated $E_0 = 188.09$, which yields a gene-wide significance level of $\alpha = 0.01/E_0 = 5.3 \times 10^{-5}$. Figure 4 depicts the $-log_{10}$(p-values). Clearly, six SNPs *rs16884481 rs10946428*, *rs6904348*, *rs10806925*, *rs9465873*, and *rs12662218* passed the significance level based on 10^5 bootstrap samples.

The testing results for the six SNPs are reported in Table 2. We report SNP ID, MAF, allele information with bold font letter as the minor allele, p-values for the nonparametric test (described in Section 4.2). We also report the p-value of the test $H_0 : \boldsymbol{\beta}_0 = \boldsymbol{\beta}_1$ v.s. $H_1 : \boldsymbol{\beta}_0 \neq \boldsymbol{\beta}_1$ in the column labeled by p_β as opposed to the model by Ma and Xu (2015) based on the generalized likelihood ratio test in Section 3.2. The p-value of the parametric test $H_0 : \boldsymbol{\alpha}_1 = 0$ is reported in the column labeled by p_{α_1} following the procedure described in Section 4.2. To compare the goodness of fit for PLVMICM with an additive varying-coefficient model (AVCM), $E(Y|\mathbf{X}, \mathbf{Z}, G) = \sum_{j=1}^{3} m_{0j}(X_j) + \boldsymbol{\alpha}_0^T\mathbf{Z} + \sum_{j=1}^{3} m_{1j}(X_j)G + \boldsymbol{\alpha}_1^T\mathbf{Z}G$, and to see the relative gain by the integrative analysis, we calculated the MSEs of both models; they are given in the last two columns of Table 2. The p-values for

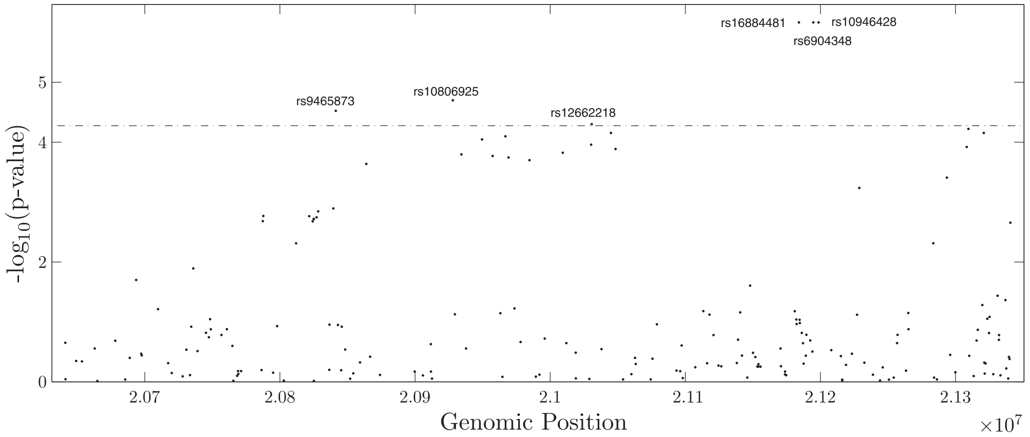

Figure 4. Plot of $-log_{10}$(p-value) for SNPs within gene *CDKAL1*.

Table 2. List of SNPs with MAF, allele, p-values under different hypothesis testing and MSE.

			p-value				MSE	
SNP ID	MAF	Alleles	p_{m_1}	p_β	p_{α_1}	p_{AVCM}	PLVMICM	AVCM
rs16884481	0.1960	**C**/T	≤1.0E-05	5.1E-04	0.2517	0.1799	0.1342	0.1402
rs10946428	0.2744	**A**/G	≤1.0E-05	1.5E-05	0.0960	0.1227	0.1333	0.1399
rs6904348	0.2766	**A**/C	≤1.0E-05	1.9E-05	0.0869	0.1358	0.1334	0.1399
rs10806925	0.4761	**C**/T	2.0E-05	2.2E-06	0.3671	0.2733	0.1340	0.1405
rs9465873	0.4503	**A**/G	3.0E-05	6.5E-06	0.4911	0.2562	0.1340	0.1403
rs12662218	0.2719	**A**/G	5.0E-05	5.4E-06	0.2802	0.4616	0.1345	0.1408

testing $H_0 : m_{11}(X_1) = m_{12}(X_2) = m_{13}(X_3) = 0$ when fitting the AVCM model are reported in the column labeled by p_{AVCM}.

The p-values in column p_β for the comparison of different model assumptions clearly show that the loading parameters are different for different index functions, indicating the necessity of the proposed model vs the one proposed by Ma and Xu (2015). The p-values in column p_{α_1} indicate that SNP×gender interactions are not significant for these six SNPs. The goodness of fit measure in the last two columns shows that the PLVMICM model fits the data better than the AVCM model, indicating the potential benefit of integrative G×E analysis. Furthermore, the testing p-values for the AVCM model do not show significance. The results imply that the genetic effects of these six SNPs are modified by the mixture effect of the three X variables, rather than separately, which further indicate the power of the integrative analysis.

For the 186 SNPs that were rejected, we fitted the model assuming $m_1(u_1) = \delta_0 + \delta_1 u_1$, assuming linear G×E interaction, then testing $H_0 : \delta_0 = \delta_1 = 0$. No SNPs showed signs of significance at the 5.3E-05 significant level. The most

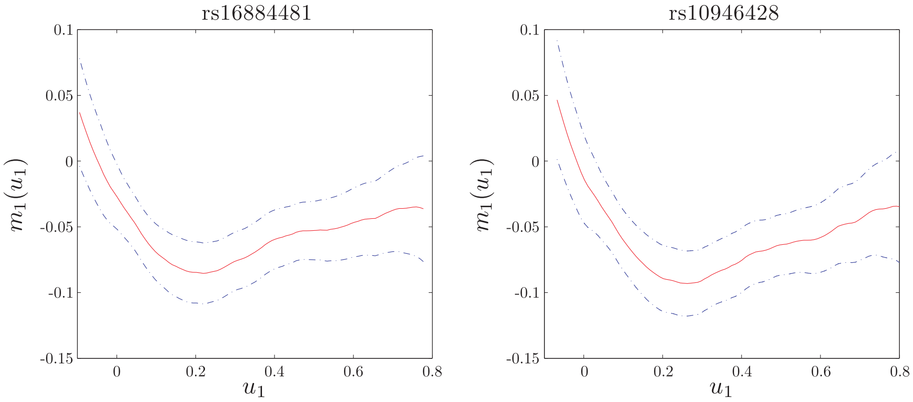

Figure 5. Plot of the estimate (solid curve) of the nonparametric function $m_1(u_1)$ for SNPs *rs16884481* and *rs10946428* along with their 95% confidence band (dash-dotted line).

significant SNP was *rs12209806* with a p-value of 6.72E-05. This indicates that there is no linear interaction between these SNPs and the three environmental variables. However, there are four SNPs, *rs12196595*, *rs6908425*, *rs6917599*, and *rs7773189* showing interactions with gender based on p_{α_1} for the 186 SNPs; the p-values were 6.12E-08, 1.89E-07, 3.69E-07, and 1.61E-05, respectively.

We tested the significance of the individual X variable that contributes to the joint effect following the procedure given in Section 3.2. The results showed that X_1 and X_2 contribute significantly to the joint effect in these six SNPs, but not X_3 (see Table S2 in Supplementary Materials). The estimators of the nonparametric function $m_1(u_1)$ for the first two SNPs, *rs16884481* and *rs10946428* along with their 95% confidence band are given in Figure 5. The estimators for the other four SNPs are shown in Section 3 in Supplementary Materials due to space limit. The estimated function shows a decreasing pattern then slightly increases as the index value u_1 increases. Our model clearly reveals the nonlinear modulating effect of environmental mixtures on genetic effect of birth weight. Such dynamic effects can be helpful in designing prevention strategy when the model is applied to other complex diseases such as diabetes.

6. Discussion

G×E interaction has been studied intensively in the literature and many statistical methods have been proposed. In this paper, we developed a partially linear varying multi-index coefficient model (PLVMICM) to conduct a rigorous assessment of the combined effect of multiple environmental exposures on the risk of disease under the paradigm of G×E interaction. Our model can be interpreted as a systems genetics approach to modeling the joint effect of environmental mixtures as a whole, then assessing how the integrative effect modifies genetic influence on disease risk. Our model is biologically attractive in that it addresses a long-term question on G×E interaction from a systems genetics perspective and is well supported by epidemiological studies (Carpenter, Arcaro, and Spink (2002); Monosson (2005); Powers et al. (2008)); and it has the flexibility to detect nonlinear interactions, and therefore, is more powerful when genetic effects are nonlinearly modified by simultaneous exposure to multiple environments.

From a statistical point of view, the index coefficient function treats multiple environmental variables $\mathbf{X}$ as a single index variable, and therefore can reduce multiple testing burden when interactions between the $\mathbf{X}$ variables and G are modelled separately. In addition, when there exist interactions between the $\mathbf{X}$ variables, our model has the flexibility to incorporate such interactions by adding interaction terms to the index function. PLVMICM is flexible and includes several existing models as special cases, for example, the partially linear single-index model (Carroll et al. (1997); Xia and Li (1999); Xia and Härdle (2006); Liang et al. (2010); Cui, Härdle, and Zhu (2011)) and the nonparametric additive model discussed by Fan and Jiang (2005).

In a typical G×E study, there are usually a large number of genetic variables (e.g., SNPs), and it is important to fit multiple SNPs in a single model and to select important players that interact with environmental mixtures to affect disease risk in a high dimensional model setup. In addition, many human diseases are measured on a binary scale. It is natural to extend the current PLVMICM model to a generalized PLVMICM model framework. This will be considered in a future investigation.

Supplementary Materials

Proofs of theorems and lemmas, additional simulation, and data analysis results can be found in the Supplementary Materials.

Acknowledgements

This work was partially supported by grants from NSF (IOS-1237969, DMS-1209112 and DMS-1512422), from NIDA/NIH (P50 DA10075 and P50 DA039838), and from NSFC (31371336). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NSF, the NIDA/NIH and the NSFC. The authors thank an associate editor and two reviewers for their constructive and helpful comments. Funding support for the GWA mapping: Maternal Metabolism-Birth Weight Interactions study was provided through the NIH Genes, Environment and Health Initiative [GEI] (U01HG004415). The datasets used for the analyses described in this manuscript were obtained from

dbGaP at `http://www.ncbi.nlm.nih.gov/sites/entrez?db=gap` through dbGaP accession number phs000096.v4.p1. Code for implementing the method was written in Matlab and C, and is available for free download at `http://www.stt.msu.edu/~cui/software.html`.

Appendix: Proofs

Notations: For any vector $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_s)^T \in \mathcal{R}^s$, let $||\boldsymbol{\xi}||_\infty = \max_{1 \leq l \leq s} |\xi_l|$. For any nonzero matrix $\mathbf{A}_{s\times s}$, denotes its L_r norm as $||\mathbf{A}||_r = \max_{\boldsymbol{\xi}\in\mathbb{R}^s, \boldsymbol{\xi}\neq 0} ||\mathbf{A}||_r ||\boldsymbol{\xi}||_r^{-1}$. For any matrix $\mathbf{A} = (A_{ij})_{i,j=1}^{s,t}$, let $||\mathbf{A}||_\infty = \max_{i\leq i\leq s} \sum_{j=1}^{t} |A_{ij}$. Let $C^{(p)}[a,b] = \{\psi : \psi^{(p)} \in C[a,b]\}$ be the space of the pth-order smooth functions. Denote the space of Lipschitz continuous functions for any fixed constant c_0 as $\mathrm{Lib}([a,b],c_0) = \{\psi : |\psi(x_1) - \psi(x_2)| \leq c_0|x_1 - x_2|, \forall x_1, x_2 \in [a,b]\}$. The following assumptions are required.

A.1 For each $l = 0, 1$, the density function $f_{U(\boldsymbol{\beta}_l)}(\cdot)$ of random variable $U(\boldsymbol{\beta}_l) = \boldsymbol{\beta}_l^T \mathbf{X}$ is bounded away from 0 on Ω_l, and there exists a constant $0 < c_0 < \infty$ such that $f_{U(\boldsymbol{\beta}_l)}(\cdot) \in \mathrm{Lib}([a,b],c_0)$ for $\boldsymbol{\beta}_l$ in the neighborhood of $\boldsymbol{\beta}_l^0$, where $\Omega_l = \{\boldsymbol{\beta}_l^T \mathbf{X}, \mathbf{X} \in \mathcal{X}\}$ and $\mathcal{X}$ is a compact support of $\mathbf{X}$.

A.2 The nonparametric function $m_l \in C^{(r)}[a_l, b_l]$, $l = 0, 1$.

A.3 The noise ε satisfies $E(\varepsilon|\mathbf{V}) = 0$, $E(|\varepsilon|^4) < \infty$ and $\sigma(\mathbf{v}) = \mathrm{var}(\varepsilon|\mathbf{V} = \mathbf{v}) < c_1$ for some $0 < c_1 < \infty$.

A.4 There exist constants $0 < c_z \leq C_z < \infty$ such that $c_z \leq Q(\mathbf{x}) = E(\tilde{Z}\tilde{Z}^T|\mathbf{X} = \mathbf{x}) \leq C_z$ for all $\mathbf{x} \in \mathcal{X}$.

A.5 The kernel function $K(\cdot)$ is a symmetric density function with compact support $[-1, 1]$ and $K \in \mathrm{Lib}([a,b], c_K)$ for some constant c_K. The bandwidth $h_l = O(n^{-1/5})$, $l = 0, 1$.

A.6 The function $u^3K(u)$ and $u^3K'(u)$ are bounded and $\int u^4K(u)du < \infty$.

Let $Y_{z,i} = Y_i - \mathbf{Z}_i^T\boldsymbol{\alpha}_0^0 - \mathbf{Z}_i^T\boldsymbol{\alpha}_1^0 G_i$, $Y_z = (Y_{z,1}, \ldots, Y_{z,n})^T$, $\mathbf{e} = (\varepsilon_1, \ldots, \varepsilon_n)^T$, $\mathbb{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)^T$, $\mathbb{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_n)^T$, $\tilde{\mathbb{Z}} = (\mathbf{1}_n, \mathbb{Z})$, and $\mathbf{G} = (G_1, \ldots, G_n)^T$. Define

$$\begin{aligned} \mathbf{U}(\boldsymbol{\beta}) &= E[D_i(\boldsymbol{\beta})D_i(\boldsymbol{\beta})^T], \quad \widehat{\mathbf{U}}(\boldsymbol{\beta}) = \frac{1}{n}\mathbf{D}(\boldsymbol{\beta})^T\mathbf{D}(\boldsymbol{\beta}), \\ \mathbf{U}(\tilde{\mathbf{Z}},\boldsymbol{\beta}) &= E[D_i(\tilde{\mathbf{Z}},\boldsymbol{\beta})D_i(\tilde{\mathbf{Z}},\boldsymbol{\beta})^T], \quad \widehat{\mathbf{U}}(\tilde{\mathbf{Z}},\boldsymbol{\beta}) = \frac{1}{n}\mathbf{D}(\tilde{\mathbf{Z}},\boldsymbol{\beta})^T\mathbf{D}(\tilde{\mathbf{Z}},\boldsymbol{\beta}), \end{aligned} \tag{A.1}$$

where $D_i(\boldsymbol{\beta}) = (D_{i,sl}(\boldsymbol{\beta}_l), 1 \leq s \leq J_n, l = 0, 1)^T$ and $\mathbf{D}(\boldsymbol{\beta}) = (D_1(\boldsymbol{\beta}), \ldots, D_n(\boldsymbol{\beta}))^T$, an $n \times 2J_n$ matrix.

Proof of Theorem 1. This is a straightforward result of Lemma S.6 in the Supplementary Materials.

Proof of Theorem 2. For simplicity, we assume $[a_l, b_l] = [a, b]$ for $l = 0, 1$. Since for any $u_l \in [a_l, b_l]$, $B_{s,l}(u_l)$, $s = 1, \dots, J_n, l = 0, 1$, have bounded first derivatives, by Lemmas S.4 and S.5 in the Supplementary Materials and Theorem 1, we have for any $u_l \in [a, b]$,

$$\begin{aligned}|\tilde{m}_l(u_l, \widehat{\boldsymbol{\beta}}) - \tilde{m}_l(u_l, \boldsymbol{\beta}^0)| =&|\mathbf{D}(\widehat{\boldsymbol{\beta}})^T \widehat{\lambda}(\widehat{\boldsymbol{\beta}}) - \mathbf{D}(\boldsymbol{\beta}^0)^T \lambda(\boldsymbol{\beta}^0)| \\ \le&|\mathbf{D}(\boldsymbol{\beta}^0)^T \{\widehat{\lambda}(\widehat{\boldsymbol{\beta}}) - \lambda(\boldsymbol{\beta}^0)\}| + |\{\mathbf{D}(\widehat{\boldsymbol{\beta}}) - \mathbf{D}(\boldsymbol{\beta}^0)\}^T \widehat{\lambda}(\widehat{\boldsymbol{\beta}})| \\ \le&|n^{-1}\mathbf{D}(\boldsymbol{\beta}^0)^T \widehat{\mathbf{U}}(\boldsymbol{\beta}^0)^{-1}\mathbf{D}(\boldsymbol{\beta}^0)^T \mathbf{e}| + O_p(n^{-1/2}) \\ =&O_p\left((\frac{N}{n})^{1/2}\right).\end{aligned}$$

Then, combined with Lemma S.4, we have

$$\begin{aligned}\sup_{u_l \in [a,b]} |\tilde{m}_l(u_l, \widehat{\boldsymbol{\beta}}) - m_l(u_l)| \le& \sup_{u_l \in [a,b]} |\tilde{m}_l(u_l, \widehat{\boldsymbol{\beta}}) - \tilde{m}_l(u_l, \boldsymbol{\beta}^0)| \\ &+ \sup_{u_l \in [a,b]} |\tilde{m}_l(u_l, \boldsymbol{\beta}^0) - m_l(u_l)| \\ =&O_p\left((\frac{N}{n})^{1/2} + N^{-r}\right).\end{aligned}$$

This completes the proof of Theorem 2.

Proof of Theorem 4. As $nh^5 = O(1)$, we have $(nh_l)^{1/2} n^{-2/5} = o(1)$. By Theorem 3, we have

$$\begin{aligned}&(nh_l)^{1/2}\left\{\widehat{m}_l(u_l, \widehat{\boldsymbol{\beta}}) - m_l(u_l) - b_l(u_l)h_l^2\right\} \\ &= (nh_l)^{1/2}\left\{\widehat{m}_l^O(u_l, \widehat{\boldsymbol{\beta}}) - m_l(u_l) - b_l(u_l)h_l^2\right\} + o_p(1).\end{aligned}$$

Thus Theorem 4 can be shown straightforwardly following Lemma S.7 in the Supplementary Materials.

Proof of Theorem 7. This proof is similar to that of Liang et al. (2010). Accordingly, we only provide a sketch of the proof here, more details can be found in the Supplementary Materials. We first prove $n^{-1}R(H_1) = E\{\sigma(\mathbf{V})\} + o_p(1)$. Let $\widehat{m}(\mathbf{X}, \boldsymbol{\beta}) = \widehat{m}_0(\mathbf{X}^T\boldsymbol{\beta}_0, \boldsymbol{\beta}) + \widehat{m}_1(\mathbf{X}^T\boldsymbol{\beta}_1, \boldsymbol{\beta})G$ and, correspondingly, $\widehat{m}^O(\mathbf{X}, \boldsymbol{\theta}) = \widehat{m}_0^O(\mathbf{X}^T\boldsymbol{\beta}_0, \boldsymbol{\beta}) + \widehat{m}_1^O(\mathbf{X}^T\boldsymbol{\beta}_1, \boldsymbol{\beta})G$. By Theorem 3 and Lemma S.7 in the Supplementary Materials, $n^{-1}R(H_1)$ can be decomposed as

$$\begin{aligned}n^{-1}R(H_1) =&\frac{1}{n}\sum_{i=1}^n \left\{y_i - \tilde{\mathbf{Z}}^T\widehat{\boldsymbol{\alpha}} - \widehat{m}(\mathbf{X}_i, \widehat{\boldsymbol{\beta}})\right\}^2 \\ =&\frac{1}{n}\sum_{i=1}^n \left\{y_i - \tilde{\mathbf{Z}}^T\boldsymbol{\alpha}^0 - \widehat{m}^O(\mathbf{X}_i, \boldsymbol{\beta}^0)\right\}^2 + o_p(n^{-2/5}) + O_p(n^{-1/2})\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{n}\sum_{i=1}^{n}\left\{\varepsilon_i - (\widehat{m}^O(\mathbf{X}_i,\boldsymbol{\beta}^0) - m(\mathbf{X}_i,\boldsymbol{\beta}^0)\right\}^2 + o_p(n^{-2/5})\\
&\equiv \mathbb{I}_1 + \mathbb{I}_2 + \mathbb{I}_3 + o_p(n^{-2/5}),
\end{aligned}
$$

where $\mathbb{I}_3 = (1/n)\sum_{i=1}^{n}\left\{\widehat{m}^O(\mathbf{X}_i,\boldsymbol{\beta}^0) - m(\mathbf{X}_i,\boldsymbol{\beta}^0\right\}^2$, $\mathbb{I}_2 = -2(1/n)\sum_{i=1}^{n}\{\widehat{m}^O(\mathbf{X}_i, \boldsymbol{\beta}^0) - m(\mathbf{X}_i,\boldsymbol{\beta}^0)\}\varepsilon_i$, and $\mathbb{I}_1 = (1/n)\sum_{i=1}^{n}\varepsilon_i^2$. It is easy to see by the Law of Large Numbers that $\mathbb{I}_1 = E\{\sigma(\mathbf{V})\} + O_p(n^{-1/2})$. By Theorem 2.6 in Li and Racine (2007), we have $\max_i |\widehat{m}^O(\mathbf{X}_i,\boldsymbol{\beta}^0) - m(\mathbf{X}_i,\boldsymbol{\beta}^0)| = O_p((\log(n)/(nh))^{1/2})$, which results in $\mathbb{I}_2 = O_p((\log(n)/(n^2h))^{1/2})$ and $\mathbb{I}_3 = O_p(\log(n)/(nh))$. This leads to $n^{-1}R(H_1) = E\{\sigma(\mathbf{V})\} + o_p(1)$.

The difference $R(H_0) - R(H_1)$ can be decomposed as

$$
\begin{aligned}
R(H_0) - R(H_1) = &\sum_{i=1}^{n}\left\{\tilde{\mathbf{Z}}^T(\widehat{\boldsymbol{\alpha}}_{H_0} - \widehat{\boldsymbol{\alpha}}_{H_1}) + (\widehat{m}(\mathbf{X}_i,\widehat{\boldsymbol{\beta}}_{H_0}) - \widehat{m}(\mathbf{X}_i,\widehat{\boldsymbol{\beta}}_{H_1}))\right\}^2\\
&+ 2\sum_{i=1}^{n}\left\{\tilde{\mathbf{Z}}^T(\widehat{\boldsymbol{\alpha}}_{H_0} - \widehat{\boldsymbol{\alpha}}_{H_1}) + (\widehat{m}(\mathbf{X}_i,\widehat{\boldsymbol{\beta}}_{H_0}) - \widehat{m}(\mathbf{X}_i,\widehat{\boldsymbol{\beta}}_{H_1}))\right\}\\
&\times\left\{y_i - \tilde{\mathbf{Z}}^T\widehat{\boldsymbol{\alpha}}_{H_1} - \widehat{m}(\mathbf{X}_i,\widehat{\boldsymbol{\beta}}_{H_1})\right\} \equiv \mathbb{I}_4 + \mathbb{I}_5.
\end{aligned}
$$

Under the null, we have $\sigma^{-2}\mathbb{I}_4 \xrightarrow{\mathcal{L}} \chi_k^2$, and under the alternative $\sigma^{-2}\mathbb{I}_4$ asymptotically follows a noncentral Chi-squared distribution with k degrees of freedom and noncentrality parameter ϕ. It remains to show that $\mathbb{I}_5 = o_p(1)$. This can be shown along the same lines as $\mathbb{I}_4$. This completes the proof of Theorem 7.

The proofs of Theorem 3, 5, and 6 are in the Supplementary Materials.

References

Andersson, E. A., Pilgaard, K., Pisinger, C., Harder, M. N., Grarup, N., Faerch, K., Poulsen, P., Witte, D. R., Jrgensen, T., Vaag, A., Hansen, T. and Pedersen, O. (2010). Type 2 diabetes risk alleles near ADCY5, CDKAL1 and HHEX-IDE are associated with reduced birthweight. *Diabetologia* **53**, 1908-1916.

Cai, Z., Fan, J. and Li, R. (2000). Efficient estimation and inferences for varying-coefficient models. *J. Amer. Statist. Assoc.* **95**, 888-902.

Carroll, R. J., Fan, J., Gijbels, I. and Wand, M. P. (1997). Generalized partially linear single-index models. *J. Amer. Statist. Assoc.* **92**, 477-489.

Carroll, R. J., Ruppert, D. and Welsh, A. H. (1998). Local estimating equations. *J. Amer. Statist. Assoc.* **93**, 214-227.

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

Cheverud, J. (2001). A simple correction for multiple comparisons in interval mapping genome scans. *Heredity* **87**, 52-58.

Cui, X., Härdle, W. and Zhu, L. (2011). The EFM approach for single-index models. *Ann. Statist.* **39**, 1658-1688.

de Boor, C. (2001), *A Practical Guide to Splines*, Springer, New York.

Falconer, D. S. (1952). The problem of environment and selection. *Amer. Natural.* **86**, 293-299.

Fan, J. and Jiang, J. (2005). Nonparametric inferences for additive models. *J. Amer. Statist. Assoc.* **100**, 890-907.

Fan, J., Zhang, C. and Zhang, J. (2001). Generalized likelihood ratio statistics and Wilks phenomenon. *Ann. Stat.* **29**, 153-193.

HAPO Study Cooperative Research Group. (2009). Hyperglycemia and Adverse Pregnancy Outcome (HAPO) Study: associations with neonatal anthropometrics. *Diabetes* **58**, 453-459.

Li, Q. and Racine, R. S. (2007). *Nonparametric Econometrics: Theory and Practice*. Princeton University Press, Princeton, N. J.

Li, Y., Wang, N. and Carroll, R. J. (2010). Generalized functional linear models with semiparametric single-index interactions. *J. Amer. Statist. Assoc.* **105**, 621-633.

Liang, H., Liu, X., Li, R. and Tsai, C. L. (2010). Estimation and testing for partially linear single index models. *Ann. Statist.* **38**, 3811-3836.

Liu, X., Jiang, H. and Zhou, Y. (2014). Local empirical likelihood inference for varying-coefficient density-ratio models based on case-control data. *J. Amer. Statist. Assoc.* **109**, 635-646.

Ma, S. and Song, P. X. (2015). Varying index coefficient models. *J. Amer. Statist. Assoc.* **110**, 341-356.

Ma, S. and Xu, S. (2015). Semiparametric nonlinear regression for detecting gene and environment interactions. *J. Statist. Plann. Inference* **156**, 31-47.

Ma, S., Yang, L., Romero, R. and Cui, Y. (2011). Varying coefficient model for gene-environment interaction: a non-linear look. *Bioinformatics* **27**, 2119-2126.

Monosson, E. (2005). Chemical mixtures: considering the evolution of toxicology and chemical assessment. *Environ. Health. Perspect.* **113**, 383-390.

Powers, K. M., Kay, D. M., Factor, S. A., Zabetian, C. P., Higgins, D. S., Samii, A., Nutt, J. G., Griffith, A., Leis, B., Roberts, J. W., Martinez, E. D., Montimurro, J. S., Checkoway, H. and Payami, H. (2008). Combined effects of smoking, coffee, and NSAIDs on Parkinson's disease risk. *Mov. Disord.* **23**, 88-95.

Ross, C. A. and Smith, W. W. (2007). Geneenvironment interactions in Parkinson's disease. *Parkins. Rel. Dis.* **13**, S309-S315.

Ruppert, D. (1997). Empirical-bias bandwidths for lcoal polynomial nonparametric regression and density estimation. *J. Amer. Statist. Assoc.* **92**, 1049-1062.

Ruppert, D., Sheathers, S. J. and Wand, M. P. (1995). An effective bandwidth selector for local least squares regression. *J. Amer. Statist. Assoc.* **90**, 1257-1270.

Sepanski, J. H., Knickerbocker, R. and Carroll, R. J. (1994). A semiparametric correction for attenuation. *J. Amer. Statist. Assoc.* **89**, 1366-1373.

Sexton, K. and Hattis, D. (2007). Assessing cumulative health risks from exposure to environmental mixtures - three fundamental questions. *Environ. Health. Perspect.* **115**, 825-832.

Wang, L. and Yang, L. (2007). Spline-backfitted kernel smoothing of nonlinear additive autoregression model. *Ann. Stat.* **35**, 2474-2503.

Wu, C. and Cui, Y. (2013). A novel method for identifying nonlinear gene-environment interactions in case-control association studies. *Hum. Genet.* **132**, 1413-1425.

Xia, Y. and Härdle, W. (2006). Semi-parametric estimation of partially linear single-index models. *J. Multiv. Anal.* **97**, 1162-1184.

Xia, Y. C. and Li, W. K. (1999). On single-index coefficient regression models. *J. Amer. Statist. Assoc.* **94**, 1275-1285.

Zhao, J., Li, M., Bradfield, J. P., Wang, K., Zhang, H., Sleiman, P., Kim, C. E., Annaiah, K., Glaberson, W., Glessner, J. T., Otieno, F. G., Thomas, K. A., Garris, M., Hou, C., Frackelton, E. C., Chiavacci, R. M., Berkowitz, R. I., Hakonarson, H. and Grant, S. F. (2009). Examination of type 2 diabetes loci implicates CDKAL1 as a birth weight gene. *Diabetes* **58**, 2414-8.

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

Department of Statistics and Probability, Michigan State University, 619 Red Cedar Road, C413 Wells Hall East Lansing, MI 48824-1027, USA.

E-mail: xuliu@stt.msu.edu

Department of Statistics and Probability, Michigan State University, 619 Red Cedar Road, C413 Wells Hall East Lansing, MI 48824-1027, USA.

E-mail: cui@stt.msu.edu

Department of Statistics, Pennsylvania State University, University Park, PA 16802, USA.

E-mail: rzli@psu.edu

(Received March 2015; accepted October 2015)