Statistica Sinica Preprint No: SS-2017-0039	
Title	Joint Test of Parametric and Nonparametric Effects in Partial Linear Models for Gene-environment Interaction
Manuscript ID	SS-2017-0039
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202017.0039
Complete List of Authors	Xu Liu Ping-Shou Zhong and Yuehua Cui
Corresponding Author	Yuehua Cui
E-mail	cui@stt.msu.edu
Notice: Accepted version subject to English editing.	

Joint Test of Parametric and Nonparametric Effects in Partial Linear Models for Gene-Environment Interaction

Xu Liu[1], Ping-Shou Zhong[2], and Yuehua Cui[2*]

[1]School of Statistics and Management,

Shanghai University of Finance and Economics, Shanghai 200433 China

[2]Department of Statistics and Probability,

Michigan State University, East Lansing, Michigan 48824 U.S.A.

Abstract: Gene-environment (G×E) interactions play a crucial role in many complex diseases. Many studies have highlighted the importance of the linear and nonlinear effects of G×E interactions to the risk of contracting diseases. Linear effects can be modeled parametrically, whereas nonlinear effects are typically modeled and estimated by using nonparametric functions under the framework of a partial linear model. Because of the difference in the rates of convergence of the parametric and nonparametric parts, few statistical studies have been devoted to assessing the simultaneous effects of the linear and nonlinear effects of G×E interactions in the context of a partial linear model. In this paper, we consider a hypothesis test to simultaneously detect the linear and nonlinear effects in a generalized partial linear varying-coefficient model. We propose a B-spline backfitted kernel method to estimate the effect of nonlinear interaction. A Wald-type statistic is constructed for the joint testing problem based on the nonparametric generalized likelihood ratio statistic. We show that the joint test statistic asymptotically follows a χ^2-distribution under the null hypothesis of no effect of G×E interaction and a non-central χ^2-distribution under the alternative. Moreover, the proposed test can simultaneously detect alternatives at optimal rates for both parame-

*Corresponding author: cui@stt.msu.edu

tric and nonparametric components. The utility of the method was demonstrated through extensive simulations and a case study.

Keywords and phrases: Associate study, Backfitting, B-spline, Joint test, GLM, Partially linear.

1 Introduction

Gene-environment (G×E) interaction is defined as phenomenon of the variation in the influence of genotypes on phenotypes under different environmental conditions (Falconer 1952). It is a key driver of epigenetic effects. A growing number of reports have confirmed the role of G×E interaction in many diseases, such as Parkinsons disease (Ross and Smith 2007) and type 2 diabetes (Zimmet et al. 2001). While linear G×E effects have been commonly assumed in many statistical analyses, there is increasing evidence of nonlinear effects of the G×E interaction on the risk of disease (Martinez et al. 2003; Sparrow et al. 2012). Certain methods for detecting nonlinear G×E effects (e.g., Ma et al. (2011); Wu and Cui (2013)) involve the application of varying-coefficient (VC) models to model the effects of nonlinear interaction. A key feature of the VC model is its flexibility in capturing the dynamics of gene effects over a spectrum of environmental changes.

In a typical genetic study on G×E interaction, both continuously and discretely measured environmental variables can be collected. For example, a mother's nutrition intake can be considered a continuously measured environmental variable whereas gender or a person's smoking status can be regarded as a discretely measured environmental variable when studying the effects of G×E interactions on birth weight. Discrete environmental variables, such as smoking status, do not interact with genes nonlinearly whereas continuous environmental variables can (Ma et al. 2011). When both types of interactions are considered in a model, a partial linear VC model (PLVCM) can be applied to study the linear and nonlinear effects of G×E interactions, where the former are typically modeled and estimated nonparametrically. Zhang et al. (2002) considered the PLVCM by introducing a two-step estimation procedure and gave a root n-consistent estimator for the parametric component. Fan and Huang (2005)

proposed a profile likelihood method to estimate parameters based on a local linear method. The PLVCM has been extended to a generalized partially linear VC model (GPLVCM) by Lu (2008) when the response is discrete. Fan and Zhang (2008) reviewed statistical methods for VCM and PLVCM, including the applications of VCM to survival, longitudinal, and functional data as well as time series data.

Let Y be the disease trait that can be quantitative or qualitative. We consider the following GPLVCM:

$$\eta(\mathbf{V};\boldsymbol{\alpha},\boldsymbol{\beta}) = g\{\mu(\mathbf{V})\} = \mathbf{Z}^T\boldsymbol{\alpha}_0 + \beta_0(U) + \{\mathbf{X}^T\boldsymbol{\alpha}_1 + \beta_1(U)\}G, \tag{1}$$

where $g\{\cdot\}$ is a given link function, and $\mu(\mathbf{V})$ is the conditional mean regression function of Y given by $\mathbf{V} = (\mathbf{Z}, U, G)$, where $\mathbf{Z} = (Z_1, \cdots, Z_q)^T$ is a q-dimensional covariate vector containing both discrete and continuous variables. U is a continuously measured environmental variable of interest, and $\mathbf{X}$ is a p-dimensional vector that is a subset of $\mathbf{Z}$, and contains environmental variables that show the linear interactions with G to affect Y, where G is a gene variable (e.g., SNP). $\boldsymbol{\alpha}_0$ and $\boldsymbol{\alpha}_1$ are parametric coefficients, and $\beta_0(u)$ and $\beta_1(u)$ are nonparametric functions. In particular, $\boldsymbol{\alpha}_1$ models the linear $G \times X$ interaction effect and $\beta_1(u)$ models the nonlinear $G \times U$ effect. They form our primary focus here.

Several methods have been developed to estimate parameters in model (1) based on local likelihood, such as the two-step estimation procedure (Zhang et al. 2002) and the profile likelihood method (Fan and Huang 2005), which iteratively estimate the linear and nonlinear parts, but this involves time consuming computation. These procedures are needed to smoothen the bandwidth when the nonparametric component is estimated by local likelihood. In this study, we propose estimating the linear and the nonparametric parts in two stages using the B-spline backfitted kernel smoothing (BSBK) procedure introduced by Wang and Yang (2007). BSBK is a two-stage method. In the first stage, we estimate the linear part by approximating nonparametric functions with B-spline bases. Then, the nonparametric functions can be obtained component wise based on local likelihood when the linear part and the other components of the nonparametric functions from B-spline estimation are

plugged in. Liu et al. (2013) studied a generalized additive model using BSBK, which was also extended by Ma and Song (2015) to estimate the varying-index coefficient model. Liu et. al. (2016) applied BSBK to study the partial linear varying multi-index coefficient model for G×E interactions. BSBK is considerably faster than the existing two-step method and the profile likelihood method without the requirement of under-smoothing.

Statistical inference has been extensively studied for the VCM and the PLVCM, but all relevant work has separately dealt with parametric and nonparametric components. In our study, the testing problem for the parametric component is to detect if $\boldsymbol{\alpha}_1 = \mathbf{0}$; that is,

$$H_0^P : \boldsymbol{\alpha}_1 = \mathbf{0} \text{ v.s. } H_1^P : \boldsymbol{\alpha}_1 \neq \mathbf{0}. \tag{2}$$

The likelihood ratio test statistic (LRT) can be applied (see Fan and Huang (2005); Fan and Zhang (2008)). It has been shown to be asymptotically χ^2-distributed with p degrees of freedom. It is also interesting to assess the interaction between U and G and determine if there exists any nonlinear interaction. This results in the nonparametric component test problem, i.e.,

$$H_0^{NP} : \beta_1(\cdot) = \mathbf{0} \text{ v.s. } H_1^{NP} : \beta_1(\cdot) \neq \mathbf{0}. \tag{3}$$

Fan and Huang (2005) proposed a generalized likelihood ratio statistic (GLRT) that extended the GLRT for the VC model (Fan et al. 2001; Cai et al. 2000). They proved that the GLRT under the null hypothesis converges to a normal distribution, and can be asymptotically approximated by a χ^2-distribution that reveals Wilk's phenomenon for nonparametric and semiparametric models. Because of the difference in the rates of convergence between the parametric ($\boldsymbol{\alpha}_1$) and the nonparametric parts ($\beta_1(\cdot)$), simultaneous inference of both the linear and effects of nonlinear interactions has not been studied thus far.

In this work, we are interested in assessing the overall effects of G×E interaction: that is, we simultaneously determine whether $\boldsymbol{\alpha} = \mathbf{0}$ and $\beta_1(\cdot) = \mathbf{0}$. We frame this joint test

problem as follows:

$$H_0^{PNP} : \boldsymbol{\alpha}_1 = \mathbf{0}, \beta_1(\cdot) = \mathbf{0} \text{ v.s. } H_1^{PNP} : \boldsymbol{\alpha}_1 \neq \mathbf{0} \text{ or } \beta_1(\cdot) \neq 0. \tag{4}$$

The challenge posed by the joint test is that the parametric and nonparametric components have different convergence rates. We can easily obtain the $\sqrt{n}$-consistent parametric estimator $\widehat{\boldsymbol{\alpha}}_1$ but not for the nonparametric estimator $\widehat{\boldsymbol{\beta}}_1(\cdot)$. Recently, Cheng and Shang (2015) proposed jointly testing the parametric and the nonparametric functions at fixed points, instead of assessing the functions entirely. Their testing problem was defined as

$$H_0 : \boldsymbol{\alpha}_1 = \mathbf{0}, \beta_1(u_0) = \beta_1^0(u_0) \text{ v.s. } H_1 : \boldsymbol{\alpha}_1 \neq \mathbf{0} \text{ or } \beta_1(\mathrm{u}_0) \neq \beta_1^0(\mathrm{u}_0).$$

where u_0 is a given fixed point and $\beta_1^0(\cdot)$ is a given function. In the setting in this study, the test for the function at some given points is not meaningful. We are more interested in testing whether the entire function is zero. This motivates our joint hypothesis test defined in (4).

The remainder of this paper is organized as follows: We introduce the model, the method of estimation, and the statistical properties of the estimators in Section 2. In Section 3, we lay out the hypothesis testing framework and derive the asymptotic distribution of the test statistic. Simulations and real data analysis are detailed in Sections 4 and 5, followed by a discussion in Section 6. The technical details are rendered in the Appendix and the online supplementary file.

2 Model and estimation

We denote the conditional mean of Y, given $\mathbf{V} = (\mathbf{Z}, U, G)$, by $\mu(\mathbf{v}) = E(Y|\mathbf{V} = \mathbf{v})$. For an ordinary generalized linear model (GLM), the conditional density of Y given $\mathbf{V} = \mathbf{v}$ belongs

to an exponential family

$$f_{Y|V}(y|\mathbf{v}) = \exp[y\xi(\mathbf{v}) - b\{\xi(\mathbf{v})\} + c(y)]$$

for known functions $b\{\cdot\}$ and $c(\cdot)$, where $\xi(\mathbf{v})$ is the canonical parameter. In this paper, we consider the model defined in (1).

Under the quasi-likelihood framework where only the relationship between the mean and the variance is specified, we can estimate the conditional mean by replacing the conditional log-likelihood $\log\{f_{Y|V}(y|\mathbf{v})\}$ by a quasi-likelihood function $Q\{\mu(\mathbf{v}), y\}$. Let the conditional variance of Y given $\mathbf{V}$ be $\mathrm{Var}(Y|\mathbf{V} = \mathbf{v}) = \sigma^2 V(\mu(\mathbf{v}))$, with an unknown function $V(\cdot)$. Thus, the quasi-score function can be given by (see Carroll et al. (1997) and Cai et al. (2000))

$$\frac{\partial}{\partial u} Q(u, y) = \frac{y - u}{V(u)}. \tag{5}$$

2.1 Parameter estimation in GPLVCM

We define $\tilde{\mathbf{Z}} = (\mathbf{Z}^T, \mathbf{X}^T G)^T$, $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_0^T, \boldsymbol{\alpha}_1^T)^T$, and $\boldsymbol{\beta}(u) = (\beta_0(u), \beta_1(u))^T$. The $\mathbf{X}$ variables contain the environmental variables that linearly interact with G. Thus, $\mathbf{X}$ is a subset of $\mathbf{Z}$, and the dimensionality of $\mathbf{X}$ is smaller than that of $\mathbf{Z}$. Model (1) can then be simplified as follows:

$$\eta(\mathbf{V}; \boldsymbol{\alpha}, \lambda) = \beta_0(U) + \beta_1(U)G + \tilde{\mathbf{Z}}^T \boldsymbol{\alpha}. \tag{6}$$

where λ contains parameters used to estimate the nonparametric functions $\beta_0(U)$ and $\beta_0(U)$. Consider the knot sequence $\xi_1 = \cdots = 0 = \xi_r < \xi_{r+1} < \cdots < \xi_{r+N_n} < 1 = \xi_{r+N_n+1} = \cdots = \xi_{N_n+2r}$, where the number of interior knots $N = N_n$ increases along with sample size n. Let $J_n = N + r$. We denote by $\mathcal{J}$ the space of the B-spline basis function of order r ($r \geq 3$) (de Boor 2001) with the B-spline basis $\mathbf{B}_r(u) = (B_{s,r}(u) : 1 \leq s \leq J_n)^T$, $u \in [a_u, b_u]$, where $[a_u, b_u]$ is the support of U. Then, $\beta_l(u)$, $l = 0, 1$, are approximated by the following spline

functions:

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

where $\lambda = (\lambda_1^T, \lambda_2^T)$, with $\lambda_l = (\lambda_{s,l}, 1 \le s \le J_n)^T$. Then, $\boldsymbol{\alpha}$ and the B-spline coefficients λ are estimated by

$$(\widehat{\boldsymbol{\alpha}}^T, \widehat{\lambda}^T)^T = \underset{\boldsymbol{\alpha}\in\boldsymbol{\Theta}_\alpha, \lambda\in\mathbb{R}^{2J_n}}{\arg\min}\ \ell_n(\boldsymbol{\alpha}, \lambda), \tag{7}$$

with the log-likelihood function given as

$$\ell_n(\boldsymbol{\alpha}, \lambda) = \sum_{i=1}^{n} Q(g^{-1}\{\tilde{\eta}(\mathbf{V}_i; \boldsymbol{\alpha}, \lambda)\}, Y_i), \tag{8}$$

where $\tilde{\eta}(\mathbf{V}; \boldsymbol{\alpha}, \lambda) = \tilde{\mathbf{Z}}^T\boldsymbol{\alpha} + \tilde{\beta}_0(U) + \tilde{\beta}_1(U)G$, $\tilde{\boldsymbol{\beta}}(u) = (\tilde{\beta}_0(u), \tilde{\beta}_1(u))^T$, and $\boldsymbol{\Theta}_\alpha$ is the parametric space of $\boldsymbol{\alpha}$.

The consistency of the spline estimators for the nonparametric functions $\tilde{\beta}_l(u)$, $l = 0, 1$ can be established. As in Wang and Yang (2007) and Liu et. al. (2016), we use the BSBK estimator to establish the asymptotic normality. We define $\widehat{\eta}_{-0}(\mathbf{V}_i; a, b) = \tilde{\mathbf{Z}}_i^T\widehat{\boldsymbol{\alpha}} + \tilde{\beta}_1(u_i)G_i + a + b(u_i - u)$, $\widehat{\eta}_{-1}(\mathbf{V}_i; a, b) = \tilde{\mathbf{Z}}_i^T\widehat{\boldsymbol{\alpha}} + \tilde{\beta}_0(u_i) + aG_i + b(u_i - u)G_i$ and $\widehat{\ell}_l(a, b) = \sum_{i=1}^n Q(g^{-1}\{\widehat{\eta}_{-l}(\mathbf{V}_i; a, b)\}, Y_i)K_{h_l}(u_i - u)$, where $K(\cdot)$ is a kernel function and h_l is bandwidth, $l = 0, 1$. We can obtain the BSBK estimator of $\beta_l(u)$ as $\widehat{\beta}_l(u) = \widehat{a}$ by local linear fitting:

$$(\widehat{a}, \widehat{b}) = \underset{(a,b)\in\mathcal{A}}{\arg\min}\, \widehat{\ell}_l(a, b), \tag{9}$$

where $\mathcal{A} \subset \mathcal{R}^2$ is a subset.

We set the space $\mathcal{M}$ as a collection of functions with finite L_2 norm on $[a_u, b_u] \times \mathcal{R}$ by $\mathcal{M} = \{\kappa(u, g) = \beta_0(u) + \beta_1(u)g, E\beta_l(U)^2 < \infty, l = 0, 1\}$. For $1 \le j \le p + q$, let $g_0(u, g)$ be a

minimizer in $\mathcal{M}$ for the following optimization problem:

$$\kappa_0(\tilde{Z}_j) = g_0(u, g) = \arg\min_{\kappa \in \mathcal{M}} E\{\tilde{Z}_j - \kappa(U, G)\}^2,$$

where E represents the conditional expectation of Z_j given (U, G). Let $P_j(\tilde{Z}_j) = \kappa_0(\tilde{Z}_j)$ and $\mathbf{P}(\tilde{\mathbf{Z}}) = (P_1(\tilde{Z}_1), \cdots, P_{p+q}(\tilde{Z}_{p+q}))^T$. Let $\widehat{\mathbf{Z}} = \tilde{\mathbf{Z}} - \mathbf{P}(\tilde{\mathbf{Z}})$. Let $q_j(x) = (\partial^j/\partial x^j)Q\{g^{-1}(x), y\}$, $j = 1, 2, 3$. Then, $q_1(x) = \{y - g^{-1}(x)\}\rho_1(x)$ and $q_2(x) = \{y - g^{-1}(x)\}\rho_1'(x) - \rho_2(x)$, and $\rho_j(x) = \{dg^{-1}(x)/dx^j\}/V\{g^{-1}(x)\}$. We define the covariance matrix of $\boldsymbol{\alpha}$ as

$$\Sigma_{\alpha^0} = E\left\{\rho_2(\mathbf{V})^{-1}\widehat{\mathbf{Z}}\widehat{\mathbf{Z}}^T\right\}^{-1}.$$

Σ_{α^0} can be simplified as $\Sigma_{\alpha^0} = \rho_0 E\left\{\widehat{\mathbf{Z}}\widehat{\mathbf{Z}}^T\right\}^{-1}$ if the error variance $\rho(\mathbf{V})$ is a constant ρ_0. Let $\mu_k = \int t^k K(t)dt$, $\nu_k = \int t^k K^2(t)dt$. Then, we can establish the asymptotic normality for the parametric estimator $\widehat{\boldsymbol{\alpha}}$ and the nonparametric estimator $\widehat{\beta}_l(u)$. Theorems 1 and 2 below are special cases in Liu et al. (2016), when the dimension of the loading parameter is one. We omit the proofs of these theorems in this paper.

Theorem 1 *Suppose that assumptions* (A.1)-(A.5) *in the Appendix hold,* $nN^{-4} \to \infty$ *and* $nN^{-2r-2} \to 0$*; then,*

$$\|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\|_2 = O_p(n^{-1/2}).$$

Furthermore, as $n \to \infty$*,*

$$n^{1/2}\left(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\right) \xrightarrow{\mathcal{L}} N(\mathbf{0}, \Sigma_{\alpha^0}),$$

where $\boldsymbol{\alpha}$ *is the true parameter of* $\boldsymbol{\alpha}$*.*

Theorem 2 *Suppose that assumptions* (A.1)-(A.5) *in the Appendix hold,* $nN^{-4} \to \infty$ *and*

$nN^{-2r-2} \to 0$; *then, for* $l = 0, 1$,

$$(nh_l)^{1/2} \left\{ \widehat{\beta}_l(u) - \beta_l(u) - b_l(u) h_l^2 \right\} \xrightarrow{\mathcal{L}} N(0, v_l(u)), \ \ as\ n \to \infty,$$

where $v_0(u) = \nu_0 \{ E\left[\rho_2(\mathbf{V}) | U = u\right] f(u) \}^{-1}$, $v_1(u) = \nu_0 \{ E\left[\rho_2(\mathbf{V}) G | U = u\right] f(u) \}^{-1}$, *and* $b_l(u) = \mu_2 \beta_l''(u)/2$.

3 Hypothesis test

Our model can simultaneously assess the effects of both the linear and nonlinear G×E interactions. This can be achieved by simultaneously testing the parametric and nonparametric components $\boldsymbol{\alpha}_1$ and $\beta_1(\cdot)$, which allows us to jointly discover the change in the trends of the interactions of the linear and nonlinear environmental effects. We consider a more general hypothesis test to detect if both $\boldsymbol{\alpha}_1$ and $\beta_1(u)$ are simultaneously equal to specific parametric forms, i.e.,

$$H_0 : \boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_1^*, \beta_1(\cdot) = \beta_1^*(\cdot) \text{ v.s. } H_1 : \boldsymbol{\alpha}_1 \neq \boldsymbol{\alpha}_1^* \text{ or } \beta_1(\cdot) \neq \beta_1^*(\cdot), \tag{10}$$

where $\boldsymbol{\alpha}_1^*$ are given constants, and $\beta_1^*(\cdot)$ is a given parametric form with unknown parameters, such as a linear form $\beta_1^*(u) = \delta_0 + \delta_1 u$. Note that hypothesis (4), i.e., $H_0^{PNP} : \boldsymbol{\alpha}_1 = \mathbf{0}, \beta_1(\cdot) = \mathbf{0}$, is a special case of hypothesis (10). To make the work more general, we develop the testing procedure following this general setup.

3.1 Generalized likelihood ratio test

To test the nonparametric function $\beta_1(\cdot)$, i.e.,

$$H_0^{NP} : \beta_1(\cdot) = \beta_1^*(\cdot) \text{ v.s. } H_1^{NP} : \beta_1(\cdot) \neq \beta_1^*(\cdot), \tag{11}$$

we can construct a generalized likelihood ratio (GLR) test. Let $\widehat{\boldsymbol{\alpha}} = (\widehat{\boldsymbol{\alpha}}_0^T, \widehat{\boldsymbol{\alpha}}_1^T)^T$ be the BSBK estimate of $\boldsymbol{\alpha}$ proposed in Section 2.1. Let $\widehat{\boldsymbol{\beta}}_{H_0}(u)$ and $\widehat{\boldsymbol{\beta}}_{H_1}(u)$ be the estimates

of $\boldsymbol{\beta}_1(u)$ under H_0 and H_1, respectively. Let the log-likelihood functions under H_0 and H_1 in the hypothesis test (11) be $\ell_n(H_0) = \sum_{i=1}^n Q(g^{-1}\{\widehat{\eta}_{H_0}(\mathbf{V}_i; \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}})\}, Y_i)$ and $\ell_n(H_1) = \sum_{i=1}^n Q(g^{-1}\{\widehat{\eta}_{H_1}(\mathbf{V}_i; \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}})\}, Y_i)$, where $\widehat{\eta}_{H_0}(\mathbf{V}_i; \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}) = \widehat{\beta}_{0,H_0}(U_i) + \widehat{\boldsymbol{\alpha}}_{H_0}^T \tilde{\mathbf{Z}}_i + \widehat{\beta}_{1,H_0}(U_i)G_i$, and $\widehat{\eta}_{H_1}(\mathbf{V}_i; \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}) = \widehat{\beta}_{0,H_1}(U_i) + \widehat{\boldsymbol{\alpha}}_{H_1}^T \tilde{\mathbf{Z}}_i + \widehat{\beta}_{1,H_1}(U_i)G_i$. We define the following GLR test statistic:

$$T_{NP} = -2(\ell_n(H_0) - \ell_n(H_1)). \tag{12}$$

To facilitate expression, we use the same bandwidth h for all coefficients. We denote by Ω the support of U, and by $|\Omega|$ the length of Ω. $\sigma_n^2 = 2h^{-1}|\Omega| \int \{K(u) - 0.5K * K(u)\}^2 du$ and $\mu_n = h^{-1}|\Omega|(K(0) - 0.5\nu_0)$, where $K * K(u)$ denotes the convolution of $K(u)$. Following the same arguments as in Fan et al. (2001), we can show that under some regular conditions, $\sigma_n^{-1}(T_{NP} - \mu_n)$ is asymptotically normally distributed.

Theorem 3 *If assumptions* (A.1)-(A.5) *in the Appendix hold,* $nN^{-4} \to \infty$ *and* $nN^{-2r-2} \to 0$*, then, under* H_0 *in* (11)*, when* $nh^{9/2} \to 0$*, then*

$$\sigma_n^{-1}(T_{NP} - \mu_n) \xrightarrow{\mathcal{L}} N(0, 1),$$

where $\sigma_n^2 = 2h^{-1}|\Omega| \int \{K(u) - 1/2K * K(u)\}^2 du$ *and* $\mu_n = h^{-1}|\Omega| \{K(0) - 1/2\nu_0\}$.

Let $\boldsymbol{\xi}_1 = \sqrt{n}\Sigma_{\alpha_1^*}^{-1/2}(\widehat{\boldsymbol{\alpha}}_1 - \boldsymbol{\alpha}_1^*)$, $\boldsymbol{\xi}_2 = \sigma_n^{-1}(T_{NP} - \mu_n)$, and $\boldsymbol{\xi} = (\boldsymbol{\xi}_1^T, \boldsymbol{\xi}_2)^T$, where $\Sigma_{\alpha_1^*}$ is the asymptotical covariance of $\widehat{\boldsymbol{\alpha}}_1$. $\Sigma_{\alpha_1^*}$ is the bottom-right block diagonal matrix of Σ_{α^*} with dimension $p \times p$. This motivates us to construct the following test statistic to simultaneously assess both the parametric and the nonparametric parts:

$$T_n = \|\boldsymbol{\xi}\|_2^2. \tag{13}$$

Lemma 1 *Suppose that assumptions* (A.1)-(A.5) *in the Appendix hold; then, under* H_0 *in* (10)*,*

$$\mathrm{COV}(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2) \xrightarrow{P} 0.$$

Lemma 1 states that $\boldsymbol{\xi}_1$ is asymptotically unrelated to $\boldsymbol{\xi}_2$.

Theorem 4 *If assumptions* (A.1)-(A.5) *in the Appendix hold, under* H_0 *in* (10),

$$T_n \xrightarrow{\mathcal{L}} \chi^2_{p+1}.$$

Theorem 4 states that T_n has an asymptotic χ^2-distribution with $p+1$ degrees of freedom. Note that Cheng and Shang (2015) had a similar result at fixed points. However, we can test the entire function instead of testing it at fixed points.

3.2 Power approximation

In this section, we consider the power of the joint test according to a sequence of local alternatives as follows:

$$H_{1n} : \boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_1^* + \boldsymbol{\alpha}_{1n} \text{ or } \beta_1(\cdot) = \beta_1^*(\cdot) + \beta_{1n}(\cdot), \tag{14}$$

where $\beta_{1n}(\cdot)$ is a vector-valued function. Before discussing the above alternative (14), we first consider the alternative of testing the nonparametric component:

$$H_{1n}^{NP} : \beta_1(\cdot) = \beta_1^*(\cdot) + \beta_{1n}(\cdot). \tag{15}$$

Theorem 5 *If assumptions* (A.1)-(A.5) *in the Appendix hold,* $nN^{-4} \to \infty$ *and* $nN^{-2r-2} \to 0$*, and if* $nh^4 \to 0$ *and* $nh^{1/2}E[\rho(\mathbf{V})\beta_{1n}(U)^2G^2] \to C(\beta)$*, where* $C(\beta)$ *is a constant, then, under* H_{1n}^{NP} *in* (15),

$$(T_{NP} - \mu_n - d_{2n})/\sigma_n \xrightarrow{\mathcal{L}} N(0,1),$$

where $d_{2n} = nE[\rho(V)\beta_{1n}(U)^2G^2]$.

Let $\phi = \phi_\alpha + \phi_\beta$, where $\phi_\alpha = \lim_{n\to\infty} n\boldsymbol{\alpha}_{1n}^T\Sigma_\alpha^{-1}\boldsymbol{\alpha}_{1n}$ and $\phi_\beta = \lim_{n\to\infty} d_{2n}$. The following theorem states the asymptotic distribution of statistic T_n under H_{1n} in (14);

Theorem 6 *Suppose that the assumptions in Theorem 5 hold, and* $n^{-1/2}\|\boldsymbol{\alpha}_{1n}\| \to C$, *where* C *is a constant. Then, under* H_{1n} *in* (14), *statistic* T_n *in (13) converges to a noncentral* χ^2*-distribution with degrees of freedom* $p+1$ *and noncentrality* ϕ.

Theorem 6 implies that the test can detect the alternatives with orders $\boldsymbol{\alpha}_{1n} = n^{-1/2}C$ and $\beta_{1n}(u) = n^{-1/2}h^{-1/4}\beta_c(u)$ simultaneously with a given constant C and a given function $\beta_c(u)$. This simultaneously yields the parametric and nonparametric convergence rates (see Hardle and Mammen (1993); Gao and Gijbels (2008)).

4 Monte Carlo simulation

The finite-sample performance of the proposed method was evaluated by simulations. In example 1, we assumed a quantitative trait, whereas a binary disease trait was assumed in example 2.

Example 1 (continuous response). Consider the following PLVCM model:

$$Y = \boldsymbol{\alpha}_0^T\mathbf{Z} + \beta_0(U) + \{\boldsymbol{\alpha}_1^T\mathbf{X} + \beta_1(U)\}G + \varepsilon,$$

where $\mathbf{Z} = (Z_0, Z_1, Z_2, Z_3)^T$ and $\mathbf{X} = (Z_1, Z_2, Z_3)^T$. We generated $Z_0, Z_1, and Z_2$ from independent normal distribution $N(0,1)$, Z_3 from Bernoulli $Ber(1, 0.5)$, and U from a uniform distribution $U(0,1)$. 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. Single nucleotide polymorphism (SNP) genotypes AA, Aa, and aa were simulated from a multinomial distribution with frequencies p_A^2, $2p_A(1-p_A)$, and $(1-p_A)^2$ for the three genotypes, respectively. The error ε came from normal distribution $N(0, \sigma^2)$. We set $\boldsymbol{\alpha}_0 = (0.7, 0.6, 0.8, 0.5)^T$, $\boldsymbol{\alpha}_1 = (0.6, 0.8, 0.5)^T$, $\beta_0(u) = \cos(\pi u)$, and $\beta_1(u) = \sin(\pi u)$. We assessed the performance of the joint test under $H_0 : \boldsymbol{\alpha}_1 = \mathbf{0}, \beta_1(\cdot) = 0$ in (4). Note that we first tested if both terms are zero, as researchers are typically interested in whether there exists an overall interaction effect. We also evaluated the power under a sequence of alternative models indexed by τ, i.e., $H_1^\tau : \boldsymbol{\alpha}_1^\tau = \tau\boldsymbol{\alpha}_1, \beta_1^\tau(\cdot) = \tau\beta_1(\cdot)$.

We used the BIC criterion to select the number of interior knots while fixing the order of the basis function as cubic to approximate the unknown functions, as described in Ma and Song (2015). The BSBK estimator $\widehat{\beta}_l(u)$ is sensitive to the choice of bandwidth h_l, $l = 0, 1$. Bandwidth selection has been extensively studied (see Sepanski et al. (1994); Ruppert et al. (1995)). To avoid the estimation of high-order derivatives, we employed a bandwidth selector based on the MSE criterion called empirical bias bandwidth selection (EBBS) (Ruppert 1997; Carroll et al. 1998; Liu et al. 2014).

Table 1: Testing size with $\sigma = 0.1, 0.5, 1.0$, $p_A = 0.1, 0.3, 0.5$, and $n = 200, 500, 1000$

	$n = 200$			$n = 500$			$n = 1000$		
σ	$p_A = 0.1$	$p_A = 0.3$	$p_A = 0.5$	$p_A = 0.1$	$p_A = 0.3$	$p_A = 0.5$	$p_A = 0.1$	$p_A = 0.3$	$p_A = 0.5$
1.0	0.063	0.060	0.057	0.056	0.061	0.054	0.057	0.053	0.052
0.5	0.063	0.063	0.060	0.058	0.065	0.055	0.055	0.052	0.052
0.1	0.067	0.058	0.059	0.057	0.060	0.054	0.056	0.053	0.051

Table 1 reports the size ($\tau = 0$) for standard deviation $\sigma = 0.1, 0.5, 1.0$, MAF $p_A = 0.1, 0.3, 0.5$, and sample size $n = 200, 500, 1000$. We can see that the sizes were closer to 0.05 as sample size n increased, and the same phenomenon was observed as the MAF approached 0.5 and the standard deviation increased. Figure 1 shows the size and power function ($\tau > 0$) at significance level 0.05 based on 500 Monte Carlo simulations with different sample sizes and MAFs. Empirical type I errors in the three scenarios were very close to the nominal level 0.05. We observed a drastic power increase when the MAF increased from 0.1 to 0.3 in all scenarios. The sample size and error variance also had a significant impact on testing power as shown in the figure. Overall, the results indicate that our method can reasonably control the false positive rate and has appropriate power to detect the joint association signal.

Example 2 (Binary response). This simulation design assumed an underlying GPLVCM for binary responses with the logistic regression model given as

$$\text{logit}\{P(Y = 1|\mathbf{Z}, U, G)\} = \boldsymbol{\alpha}_0^T\mathbf{Z} + \beta_0(U) + \{\boldsymbol{\alpha}_1^T\mathbf{X} + \beta_1(U)\}G, \tag{16}$$

where U and G were generated in the same manner as in Example 1, $\mathbf{Z} = (Z_0, Z_1, Z_2, Z_3)^T$

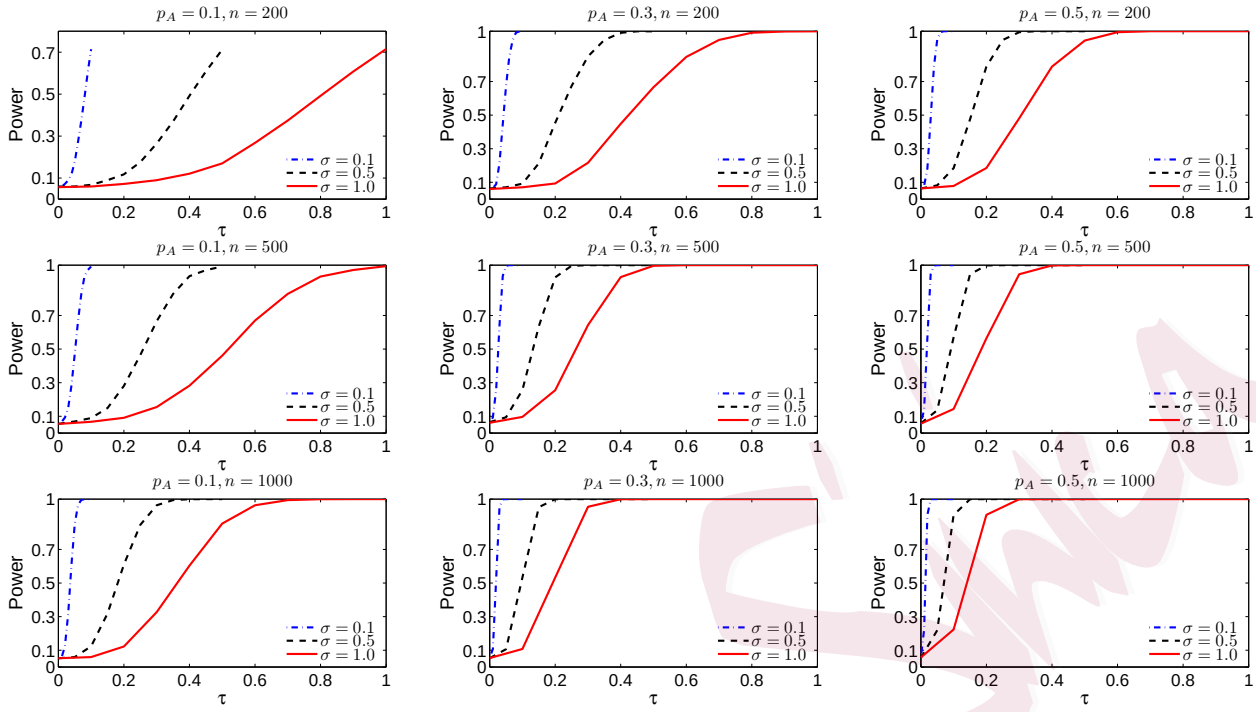

Figure 1: *The empirical size and power function of test statistic T_n for the simultaneous inference of parametric and nonparametric parameters under different simulation settings.*

were generated from independent normal distribution $N(0,1)$, and $\mathbf{X} = (Z_1, Z_2, Z_3)^T$, $\boldsymbol{\alpha}_0 = (0.7, 0.6, 0.8, 0.5)^T$, and $\boldsymbol{\alpha}_1 = (0.6, 0.8, 0.5)^T$. We set $\beta_0(u) = \cos(\pi u)$ and $\beta_1(u) = \sin(\pi u)$. We reported in Figure 2 the size ($\tau = 0$) and power function ($\tau > 0$) at a significance level of 0.05 based on 1,000 Monte Carlo simulations with different sample sizes and $P_A = 0.3$. Similar results to those in example 1 were observed for $P_A = 0.1$ and $P_A = 0.5$, and hence are omitted. The results demonstrated the finite sample performance of the proposed joint test statistic.

Intuitively, we expect a power gain for the joint test when both the parametric and nonparametric components contribute something, as pointed out by one reviewer. When one component has weak signal, the joint test signal could be diluted by the weak one. To demonstrate this, we conducted further simulations. We simulated data assuming three scenarios. In scenario 1 (denoted as S1), both the parametric and nonparametric components were assumed to have an effect. In scenario 2 (S2), there was only parametric interaction effect while the nonparametric effect was assumed to be zero. In S3, no parametric effect

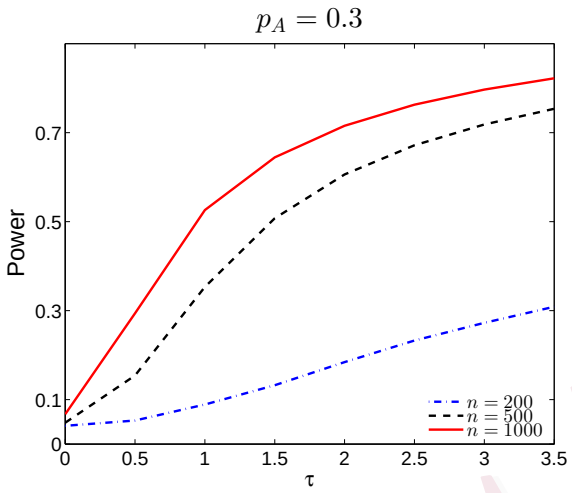

Figure 2: *The empirical size and power function of test statistic T_n for the simultaneous inference of both parametric and nonparametric parameters with binary response with different sample sizes.*

was assumed and only nonparametric effect was included. Scenarios S2 and S3 are extreme cases. The corresponding data generating model under the alternative in the three scenarios were given as follows, while each component was described in the setting given in Example 1 of the paper.

(S1). $H_1^\tau : Y = \boldsymbol{\alpha}_0^T\mathbf{Z} + \beta_0(U) + \tau\{\boldsymbol{\alpha}_1^T\mathbf{X} + \beta_1(U)\}G + \varepsilon$

(S2). $H_1^\tau : Y = \boldsymbol{\alpha}_0^T\mathbf{Z} + \beta_0(U) + \tau\boldsymbol{\alpha}_1^T\mathbf{X}G + \varepsilon$

(S3). $H_1^\tau : Y = \boldsymbol{\alpha}_0^T\mathbf{Z} + \beta_0(U) + \tau\beta_1(U)G + \varepsilon$

where data $(\mathbf{Z}, U, G)$ are generated the same as described in Example 1 of the paper; $\mathbf{X}$ is a subset of $\mathbf{Z}$; and the model in H_1^τ is a sequence of alternative models indexed by τ, $\tau = 0, 0.01, \cdots, 0.1$. For this simulation, we only focused on cases with sample size $n = 500$, MAF=0.3, and error variance $\sigma^2 = 1$ in these three scenarios. Similar performance was observed under other settings, hence was omitted.

We considered three hypothesis testing as follows:

(1). Joint test, denoted by "JointTest", i.e., $H_0 : \boldsymbol{\alpha}_1 = 0, \beta(\cdot) = 0$;

(2). Partial parametric test, denoted by "ParTest", i.e., $H_0 : \boldsymbol{\alpha}_1 = 0$;

(3). Partial nonparametric test, denoted by "NonparTest", i.e., $H_0 : \beta(\cdot) = 0$.

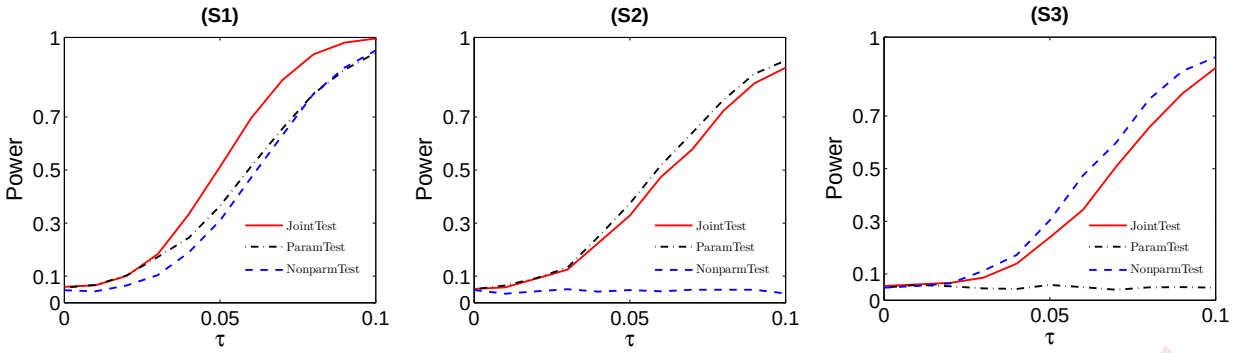

Figure 3: *Plot of testing size and power for "JointTest" (red solid line), "ParamTest" (blue dotted line) and "NonparamTest" (black dot-dashed line) under scenarios S1, S2 and S3.*

Figure 3 shows the power functions of the three tests under three different scenarios. In all the cases, the size ($\tau = 0$) for the three tests can be reasonably controlled. In S1 where both the parametric and nonparametric effects are present, we observed better power of the joint test than that of the two partial tests, highlighting the power gain of the joint test compared to the individual test when both components contribute something. In S2 where there only exists parametric effect, the partial size for the nonparametric test is well controlled (dotted blue line). The power of JointTest is sightly smaller than that of ParTest, due to the signal dilution from the nonparametric component. In S3 where no parametric effect exists, the partial size for the parametric test is well controlled (dot-dashed black line). The power of JointTest is sightly smaller than that of NonparTest, due to the signal dilution from the parametric component.

From this simulation, we see that the joint test achieves power gain when both the parametric and nonparametric components contribute something. In extreme cases where one component does not show any effect or show weak effect, the joint test will have power loss due to potential signal dilution by the weaker one. Although we cannot theoretically show conditions under which the joint test has larger power than individual tests, this simulation result does illustrate the power gain of the joint test and gives us some practical insight about the proposed test.

5 Case study

We applied the proposed GPLVCM model to a dataset 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 proposed method. Birth weight was the primary variable of interest of the trait. Fetal growth is not only determined by fetal genes, but is also controlled by complex interactions between fetal genes and the maternal uterine environment. In this example, we focused on Thai population, with 1,126 subjects genotyped with the Omni1-Quad_v1-0_B platform. Since the mother's glucose level can have a significant impact on fetal growth, we chose *b_CordPGC_mg* (the fetuss cord glucose level) as the varying environmental variable U to try to understand whether fetal genes respond to the mother's glucose level to influence birth weight. The discrete variable, denoted by Z_1, contains the gender of the fetus. The continuous variables, denoted by Z_2 and Z_3, respectively, contain *m_HtM_OGTT* (the mother's mean OGTT height) and *m_OneHrPG_CLC_mg* (the mother's one-hour OGTT glucose). We set $\mathbf{Z} = (Z_1, Z_2, Z_3)^T$ and $\mathbf{X} = (Z_2, Z_3)^T$. To show the utility of the method, we picked chromosome 6 for demonstration. There are 43,261 SNPs following the removal of those with minor allele frequency (MAF) less than 0.05 or p-value < 0.001 for testing Hardy-weinberg equilibrium. Our goal was to determine if there are any SNPs associated with birth weight in Thai population and if, so, how the SNPs responded to the mother's glucose level (considered the environment) to influence birth weight, and to further determine the mechanism of interaction.

We tested to determine if any SNP was associated with birth weight based on the joint test, i.e., $H_0 : \boldsymbol{\alpha}_1 = 0$ and $\beta_1(\cdot) = 0$. We individually tested each SNP and applied the local false discovery rate (LFDR) (see Efron et al. (2001); Storey (2002); Storey and Tibshirani (2003)) to adjust the multiple testing. We used the R package "fdrtool" with "bootstrap" method (Strimmer 2008) to calculate local FDR, and then estimated the proportion of null p-values for the joint test which is calculated as $\eta_0 = 0.7237$. The bandwidth constant was chosen as $c = 0.4125$ in the bandwidth calculation formula $h = c \times sd(U) \times n^{-2/9}$. We used the same notation as given in Section 4: the proposed joint test (denoted by "JointTest");

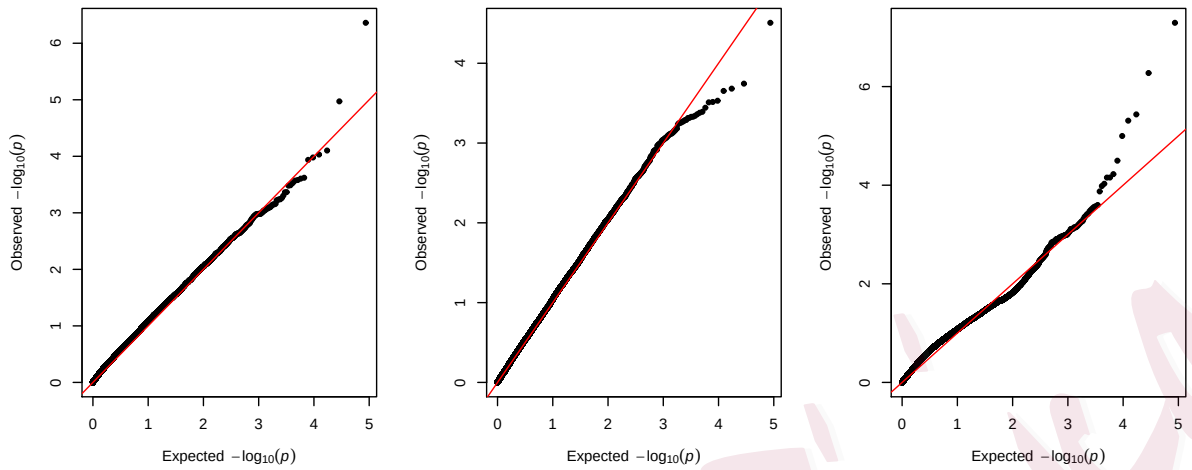

Figure 4: *The QQ-plot of the -log10(p-values) for the "JointTest" (left), "ParTest" (middle) and "NonparTest" (right) with a chosen bandwidth constant* $c = 0.4125$.

partial parametric test (denoted by "ParTest"); and partial nonparametric test (denoted by "NonparTest"). The Q-Q plot of the $-\log_{10}$(p-values) for these three tests are depicted in Figure 4. It can be seen that the chosen bandwidth does not lead to inflated p-values. Based on the proportion of null p-value estimate $\eta_0 = 0.7237$, there is only one SNP (*rs1490352*) showing statistical significance.

Table 2: List of SNP ID, gene to which the SNP belongs, MAF, alleles (minor allele is shown as bold font), and p-values for SNP *rs1490352* on chromosome 6, under the marginal and joint tests.

SNP ID	Gene	MAF	Alleles	p-values		
				JointTest	ParTest	NonparTest
rs1490352	NKAIN2	0.4082	**G**/A	4.349E-07	0.259	5.087E-08

Table 2 shows the SNP *rs1490352* with the SNP ID, MAF, alleles, and the p-value of each for the joint and separate tests. Alleles in bold represent minor alleles. We also separately tested the two interaction effects. We see that the joint test yielded p-values closer to that of the nonparametric testing. The parametric component is not significant. The weak effect of the parametric component may dilute the joint test signal, leading to slightly larger p-value of the joint test than that of the nonparametric test. This result is consistent with and

supported by our simulation study.

6 Discussion

The evaluation of G×E interaction is an important topic in research on genetic association studies. With the development of statistical models, e.g., the partially linear varying-coefficient model, we can assess the nonlinear G×E interaction in a model-based framework. In this study, we proposed and verified a joint testing framework to assess the effects of G×E interaction including linear and nonlinear interactions. Note that the joint test is equivalent to assessing the total genetic effect (the main genetic effect is embedded into the nonparametric function. See below for further explanation). In a genetic association study, the natural choice is to assess the total genetic effect first, then assess the effects of the interaction. This is another motivation, in addition to the gain in power offered by the proposed joint testing framework. Linear and nonlinear interactions can be assessed separately if the null hypothesis of the joint test is rejected.

We theoretically assessed the distribution of the joint test statistic under the proposed estimation framework. Both simulation and real data analysis demonstrated the utility of the method. Novel genetic insight can be obtained by the joint test. In contrast to the work of Cheng and Shang (2015), where the parametric and nonparametric functions were jointly assessed at fixed points, we assess the two components globally. Although the parametric and nonparametric components have different convergence rates, the proposed test can simultaneously yield their respective optimal rates. In addition to the application of G×E interactions, our work also contributes to the theory of semiparametric inference.

Under the proposed GPLVCM model, the joint test of the effects of the G×E interaction is equivalent to testing the total genetic effect. If we take $\beta_1(u) = \beta_1 + f(u)$, where $f(u)$ can be linear or nonlinear in u, $\beta_1(u)G = \beta_1 G + f(u)G$. It can be seen that the term $\beta_1(u)G$ contains the marginal effect of G on Y. If we use B-spline to approximate the basis functions of $\beta_1(u)$, a change in the normalized basis functions can be obtained with the first column of the basis functions containing all ones (Schumaker 1981). Such a transformation

does not change the nature of the spline functions, but allows us to separate the marginal and interaction-related effects. Thus, the main genetic effects and those of nonlinear G×E interaction can be separately tested under the proposed framework.

Our method was motivated and demonstrated by a genetic association study. It can be applied to other studies, where a partial linear structure can be fitted. Partial linear models have been extensively studied in the literature. While most studies focus on the estimation problem, little research has been dedicated to testing the significance of the joint parametric and nonparametric effects. Our work fills this gap beyond the application of assessing G×E interactions. In addition, it can be extended to generalized partially linear additive models (e.g., Zhang and Liang (2011) and Ma and Yang (2011)) and partially linear varying multi-index coefficient models (Liu et. al. 2016). The extension will allow us to assess the nonlinear G×E effect when more than one continuous environmental variable of interest is considered.

Supplementary Materials

The technical details, including proofs of the major theorems and lemmas used in this paper, can be found in the Supplementary Materials.

Acknowledgments

This work was partially supported by grants from the NSFC (11771267), the Program for Innovative Research Team of Shanghai University of Finance and Economics, and the NSF (IOS-1237969, DMS-1209112 and DMS-1309156). We thank Saad Anis, PhD, from Liwen Bianji, Edanz Editing China (www.liwenbianji.cn/ac), for editing the English text of a draft of this manuscript.

Appendix: Proofs

Notations: For any vector $\boldsymbol{\xi} = (\xi_1, \cdots, \xi_s)^T \in \mathcal{R}^s$, $||\boldsymbol{\xi}||_\infty = \max_{1\le l\le s} |\xi_l|$. For any nonzero matrix $\mathbf{A}_{s\times s}$, denote its L_r norm by $||\mathbf{A}||_r = \max_{\boldsymbol{\xi}\in\mathbb{R}^s, \boldsymbol{\xi}\neq 0} ||\mathbf{A}||_r ||\boldsymbol{\xi}||_r^{-1}$. For any matrix $\mathbf{A} = (A_{ij})_{i,j=1}^{s,t}$, $||\mathbf{A}||_\infty = \max_{i\le i\le s} \sum_{j=1}^t |A_{ij}|$. Let $C^{(p)}[a_u, b_u] = \{\psi : \psi^{(p)} \in C[a_u, b_u]\}$ be the space of pth-order smooth functions. Denote the space of Lipschitz continuous functions for any fixed constant c_0 by $\text{Lib}([a_u, b_u], c_0) = \{\psi : |\psi(x_1) - \psi(x_2)| \le c_0|x_1 - x_2|, \forall x_1, x_2 \in [a_u, b_u]\}$. The following assumptions are required to show the consistency and asymptotic normality of our estimators:

Assumptions:

A.1 The random variable U has compact support $[a_u, b_u]$. The density function $f_u(\cdot)$ of random variable U is bounded away from zero on Ω, and there exists a constant $0 < c_0 < \infty$ such that $f_u(\cdot) \in \text{Lib}([a_u, b_u], c_0)$.

A.2 The nonparametric function $m_l \in C^{(p)}[a_u, b_u]$, $l = 0, 1$.

A.3 $c_x \le \|E\{\mathbf{Z}^T\mathbf{Z}|U = u\}\|_2 \le C_x$.

A.4 The kernel function $K(\cdot)$ is a symmetric density function with compact support $[-1, 1]$ and $K \in \text{Lib}([a_u, b_u], c_K)$ for some constant c_K.

A.5 The functions $u^3K(u)$ and $u^3K'(u)$ are bounded, and $\int u^4K(u)du < \infty$.

Denote $q_k(\tilde{\eta}_i)$ by $q_k\{\tilde{\eta}(\mathbf{V}_i; \boldsymbol{\alpha}_0, \lambda)\}$, $k = 1, 2$, $i = 1, \cdots, n$. Let $\mathbf{q}_k = (q_k(\tilde{\eta}_1), \cdots, q_k(\tilde{\eta}_n))^T$ and $\mathbf{W}_{q_2}$ be a diagonal matrix with diagonal elements $\mathbf{q}_2\{\tilde{\eta}\}$. Define

$$\begin{aligned} \mathbf{U} = E[q_2(\tilde{\eta}_i)D_iD_i^T], \quad \widehat{\mathbf{U}} = \frac{1}{n}\mathbf{D}^T\mathbf{W}_{q_2}\mathbf{D}, \\ \mathbf{U}(\mathbf{Z}) = E[q_2(\tilde{\eta}_i)D_i(\mathbf{Z})D_i(\mathbf{Z})^T], \quad \widehat{\mathbf{U}}(\mathbf{Z}) = \frac{1}{n}\mathbf{D}(\mathbf{Z})^T\mathbf{W}_{q_2}\mathbf{D}(\mathbf{Z}), \end{aligned} \tag{A.1}$$

where $D_i = (\mathbf{B}_r(U_i)^T\tilde{X}_{i,l}, l = 1, \cdots, 2p)^T$, $D_i(\mathbf{Z}) = (\mathbf{Z}_i^T, D_i^T)^T$, $\mathbf{D} = (D_1, \cdots, D_n)^T$ which is an $n \times pJ_n$ matrix, and $\mathbf{D}(\mathbf{Z}) = (D_1(\mathbf{Z}), \cdots, D_n(\mathbf{Z}))^T$ which is an $n \times 2(q + pJ_n)$ matrix.

The proofs of Theorems 1 and 2 are omitted here; they are special cases in Liu et al. (2016). The details are shown in the Supplementary Materials. To prove Theorem 3, we

define "oracle" estimation. Similar to (9), we can obtain the "Oracle" kernel estimator of $\beta_l(u)$ as $\widehat{\beta}_l^O(u) = \widehat{a}^O + \widehat{b}^O u$ by local linear fitting:

$$(\widehat{a}^O, \widehat{b}^O) = \arg\min_{(a,b)\in\mathcal{A}} \tilde{\ell}(a,b), \tag{A.2}$$

where $\tilde{\ell}(a,b) = \sum_{i=1}^n Q(g^{-1}\{\widehat{\eta}_{-l}^O(\mathbf{V}_i; a^O, b^O)\}, Y_i)K_{h_l}(u_i - u)$, $\widehat{\eta}_{-0}^O(\mathbf{V}_i; a, b) = \widehat{\boldsymbol{\alpha}}^T\tilde{\mathbf{Z}}_i + \beta_1(u_i)G_i + a + b(u_i - u)$ and $\widehat{\eta}_{-1}^O(\mathbf{V}_i; a, b) = \widehat{\boldsymbol{\alpha}}^T\tilde{\mathbf{Z}}_i + \beta_0(u_i) + aG_i + b(u_i - u)G_i$. The "Oracle" means that we already know the true functional structure before estimating function $\beta_l(u)$.

As in Liu et. al. (2016), assuming that the nonparametric functions $\boldsymbol{\beta}(u)$ are known, we can construct a GLR statistic based on the "Oracle" estimator $\widehat{\boldsymbol{\beta}}^O(u)$. Consider hypothesis test (3.14). Let $\widehat{\beta}_{l,H_0}^O(u)$ and $\widehat{\beta}_{l,H_1}^O(u)$ be the "Oracle" estimates under H_0 and H_1, the same as in Section 2.1, respectively. The resulting likelihoods under H_0 and H_1 in hypothesis test (3.14) are

$$\ell_n^O(H_0) = \sum_{i=1}^n Q(g^{-1}\{\widehat{\eta}_{H_0}^O(\mathbf{V}_i; \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\theta}})\}, Y_i),$$
$$\ell_n^O(H_1) = \sum_{i=1}^n Q(g^{-1}\{\widehat{\eta}_{H_1}^O(\mathbf{V}_i; \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\theta}})\}, Y_i),$$

where $\widehat{\eta}_{H_0}^O(\mathbf{V}_i; \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\theta}}) = \mathbf{Z}_i^T\widehat{\boldsymbol{\alpha}}_{0,H_0} + \mathbf{X}_i^T\widehat{\boldsymbol{\theta}}_{0,H_0}^O(U_i) + \mathbf{Z}_i^T\widehat{\boldsymbol{\alpha}}_{1,H_0}G_i$ and $\widehat{\eta}_{H_1}^O(\mathbf{V}_i; \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\theta}}) = \mathbf{Z}_i^T\widehat{\boldsymbol{\alpha}}_{0,H_1} + \mathbf{X}_i^T\widehat{\boldsymbol{\theta}}_{0,H_1}^O(U_i) + \{\mathbf{Z}_i^T\widehat{\boldsymbol{\alpha}}_{1,H_1} + \mathbf{X}_i^T\widehat{\boldsymbol{\theta}}_{1,H_1}^O(U_i)\}G_i$. We define the following Oracle-version of the GLR test statistic as

$$T_{NP}^O = 2(\ell_n^O(H_1) - \ell_n^O(H_0)). \tag{A.3}$$

Let $a_K = \{K(0) - 1/2\int K^2(u)du\}[\int\{K(u) - 1/2K * K(u)\}du]^{-1}$, where $K * K(u)$ denotes the convolution of K.

Proof of Theorem 3: According to Lemma S.9 in the Supplementary Materials,

$$\ell_n^O(H_0) - \ell_n(H_0) = O_p(\log n),$$
$$\ell_n^O(H_1) - \ell_n(H_1) = O_p(\log n),$$

which implies that

$$T_{NP} = T_{NP}^{O} + O_p(\log n).$$

Lemma S.10 in the Supplementary Materials states that under the assumptions of Theorem 3,

$$\sigma_n^{-1}(T_{NP}^{O} - \mu_n) \xrightarrow{\mathcal{L}} N(0,1),$$

where $\sigma_n^2 = 2h^{-1}|\Omega| \int \{K(u) - 1/2K * K(u)\}^2 du$ and $\mu_n = h^{-1}|\Omega| \left\{K(0) - 1/2 \int K^2(u)du\right\}$. This results directly in Theorem 3.

Proof of Lemma 1: Invoking the proof of Theorem 1 and 3,

$$\begin{aligned} n^{1/2}\Sigma_\alpha^{-1/2}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*) =& n^{-1/2}\Sigma_\alpha^{-1/2} \sum_{i=1}^{n} (\tilde{\mathbf{Z}}_i - \mathbf{P}(\tilde{\mathbf{Z}}_i))\varepsilon_i + o_p(1), \\ \sigma_n^{-1}(T_{NP} - \mu_n) =& v^{-1}\Upsilon(n) + o_p(1), \end{aligned}$$

where $\varepsilon_i = q_1(\eta^*_{i,H_0})$, $\Upsilon(n) = \frac{1}{n}h^{-1/2} \sum_{i\neq j}^{n} \varepsilon_i\varepsilon_j \check{\mathbf{X}}_i^T \Gamma(U_i)\check{\mathbf{X}}_j\{2K((U_i - U_j)/h) - \tilde{K}((U_i - U_j)/h)\}$, and $v^2 = 2|\Omega| \int \{K(t) - 1/2\tilde{K}(t)\}^2 dt$ are defined in the proof of Lemma S.10 in the Supplementary Materials. Let

$$\begin{aligned} \mathbb{I}_{1n} =& \sum_{k\neq i,j}^{n} (\tilde{\mathbf{Z}}_k - \mathbf{P}(\tilde{\mathbf{Z}}_k))\varepsilon_k \sum_{i\neq j}^{n} \varepsilon_i\varepsilon_j \check{\mathbf{X}}_i^T \Gamma(U_i)\check{\mathbf{X}}_j\{2K((U_i - U_j)/h) - \tilde{K}((U_i - U_j)/h)\}, \\ \mathbb{I}_{2n} =& \sum_{i\neq j}^{n} \varepsilon_i^2\varepsilon_j (\tilde{\mathbf{Z}}_i - \mathbf{P}(\tilde{\mathbf{Z}}_i))\check{\mathbf{X}}_i^T \Gamma(U_i)\check{\mathbf{X}}_j\{2K((U_i - U_j)/h) - \tilde{K}((U_i - U_j)/h)\}. \end{aligned}$$

It is easy to see that $E[\mathbb{I}_{1n}] = 0$ and $E[\mathbb{I}_{2n}] = 0$. Therefore,

$$\mathrm{COV}(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2) = n^{-1/2}v^{-1}\Sigma_\alpha^{-1/2}(\mathbb{I}_{1n} + \mathbb{I}_{2n}) + o_p(1),$$

which results directly in $\mathrm{COV}(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2) \xrightarrow{P} 0$.

Proof of Theorem 4: Theorem 1 and Theorem 3 imply that

$$\|\boldsymbol{\xi}_1\|_2^2 \xrightarrow{\mathcal{L}} \chi_p^2, \text{ and } \boldsymbol{\xi}_2^2 \xrightarrow{\mathcal{L}} \chi^2.$$

Theorem 4 follows from Lemma 1.

Proof of Theorem 5: We proved in Lemma S.11 that under H_1^{NP} in (16),

$$\sigma_n^{-1}(T_{NP}^O - \mu_n - d_n) \xrightarrow{\mathcal{L}} N(0,1), \tag{A.4}$$

where $d_{2n} = nE[\rho(V)\boldsymbol{\theta}_n(U)^T\check{\mathbf{X}}\check{\mathbf{X}}^T\boldsymbol{\theta}_n(U)]$. According to Lemma S.9 in the Supplementary Materials,

$$\begin{aligned}\ell_n^O(H_0) - \ell_n(H_0) =& O_p(\log n),\\ \ell_n^O(H_1) - \ell_n(H_1) =& O_p(\log n),\end{aligned}$$

which implies that

$$T_{NP} = T_{NP}^O + O_p(\log n). \tag{A.5}$$

Thus, Theorem 5 can be shown by (A.4) and (A.5).

Proof of Theorem 6: From Theorem 1,

$$\begin{aligned}\boldsymbol{\xi}_1 =& \sqrt{n}\Sigma_{\alpha_1}^{-1/2}(\widehat{\boldsymbol{\alpha}}_1 - \boldsymbol{\alpha}_1^*)\\ =& \sqrt{n}\Sigma_{\alpha_1}^{-1/2}(\widehat{\boldsymbol{\alpha}}_1 - \boldsymbol{\alpha}_1) + \sqrt{n}\Sigma_{\alpha}^{-1/2}(\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_1^*),\end{aligned}$$

which implies that $\boldsymbol{\xi}_1$ is asymptotically normally distributed with mean $\sqrt{n}\Sigma_{\alpha_1}^{-1/2}\boldsymbol{\alpha}_{1n}$ and variance one. Along the lines of the proof of Lemma 1, we can prove that $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$ are asymptotically uncorrelated under H_{1n}. It is easy to see that $\|\boldsymbol{\xi}_1\|_2^2$ converges to a noncentral chi-squared distribution with q degrees of freedom and noncentrality parameter $\phi_\alpha = \lim_{n\to\infty} n\boldsymbol{\alpha}_{1n}^T\Sigma_{\alpha_1}^{-1}\boldsymbol{\alpha}_{1n}$. This implies that T converges to a noncentral chi-squared dis-

tribution with $q+1$ degrees of freedom and noncentrality parameter $\phi = \phi_\theta + \phi_\theta$, where $\phi_\theta = \lim_{n\to\infty} d_{2n}$.

References

Barrett, J. C., Fry, B., Maller, J. and Daly, M. J. (2005). "Haploview: analysis and visualization of LD and haplotype maps". *Bioinformatics*, **21** 263-265.

Cai, Z., Fan, J. and Li, R. (2000), "Efficient estimation and inferences for varying-coefficient models", *Journal of the American Statistical Association*, **95**, 888-902.

Carroll, R. J., Fan, J., Gijbels, I. and Wand, M. P. (1997), "Generalized partially linear single-index models", *Journal of the American Statistical Association*, **92**, 477-489.

Carroll, R. J., Ruppert, D. and Welsh, A. H. (1998), "Local estimating equations," *Journal of the American Statistical Association*, **93**, 214-227.

Cheng, G. and Shang, Z. (2015). "Joint asymptotics for semi-nonparametric regression models with partially linear structure", *The Annals of Statistics*, **43**, 1351-1390.

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

de Boor, C. *A practical guide to splines*, Springer, New York.

Efron, B., Tibshirani, R. Storey, J. D. and Tusher, V. (2001). Empirical bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, **96**, 1151-1160.

Falconer, D. S. (1952). The Problem of Environment and Selection. *Am. Natural.* **86**, 293-299.

Fan, J. and Huang, T. (2005), "Profile likelihood inferences on semiparametric varying-coefficient partially linear models", *Bernoulli*, **11**, 1031-1057.

Fan, J., Zhang, C. and Zhang, J. (2001), "Generalized likelihood ratio statistics and Wilks phenomenon", *The Annals of Statistics*, **29**, 153-193.

Fan, J. and Zhang, W. (2008), "Statistical mothods with varying coefficient models", *Statistics and Its Interface*, **1**, 179-195.

Gao, J. and Gijbels, I. (2008), "Bandwidth selection in nonparametric kernel testing", *Journal of the American Statistical Association*, **103**, 1584-1594.

Härdle, W. and Mammen, E. (1993), "Comparing nonparametric versus parametric regression fits", *The Annals of Statistics*, **21**, 1926-1947.

Liu, R., Yang, L. and Härdle, W. K. (2013), "Oracally efficient two-step estimation of generalized additive model", *Journal of the American Statistical Association*, **108**, 619-631.

Liu, X., Gao B. and Cui, Y. (2016). "Generalized Partially linear varying multi-index coefficient model for gene-environment interactions". *Statistical Applications in Genetics and Molecular Biology*, **16**, 59-74.

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

Liu, X., Jiang, H. and Zhou, Y. (2014), "Local empirical likelihood inference for varying-coefficient density-ratio models based on case-control data", *Journal American statistical Association*, **109**, 635-646.

Lu, Y. (2008), "Generalized partially linear varying-coefficient models", *Journal of Statistical Planning and Inference*, **138**, 901-904.

Ma, S. and Song, P. X. (2015). Varying index coefficient models. *Journal American statistical Association*, **110**, 341-356.

Ma, S. and Yang, L. (2011). Spline-backfitted kernel smoothing of partially linear additive model. *Journal of Statistical Planning and Inference* **141**, 204-219.

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.

Martinez, J. A., Corbalán, M. S. , Sánchez-Villegas, A., Forga, L., Marti, A. and Martinez-González, M. A. (2003). "Obesity risk is associated with carbohydrate intake in women carrying the Gln27Glu β_2-adrenoceptor polymorphism." *J. Nutrition.* **133**, 2549-2554.

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," *Journal of the American statistical Association*, **92**, 1049-1062.

Ruppert, D., Sheathers, S. J. and Wand, M. P. (1995), "An effective bandwidth selector for local least squares regression," *Journal of the American Statistical Association*, **90**, 1257-1270.

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

Sepanski, J. H., Knickerbocker, R. and Carroll, R. J. (1994), "A semiparametric correction for attenuation," *Journal of the American Statistical Association*, **89**, 1366-1373.

Sparrow, D. B., Chapman, G., Smith, A. J., Mattar, M. Z., Major, J. A., O'Reilly, V. C., Saga, Y., Zackai, E. H., Dormans, J. P., Alman, B. A., Mc Gregor, L., Kageyama, R., Kusumi, K. and Dunwoodie, S. L. (2012). "A mechanism for gene-environment interaction in the etiology of congenital scoliosis." *Cell*, **149** 295-306.

Strimmer, K. (2008). fdrtool: a versatile R package for estimating local and tail area-based false discovery rates. *Bioinformatrics*, **24**, 1461-1462.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **64**, 479-498.

Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, **100**, 9440-9445.

Wang, L. and Yang, L. (2007), "Spline-backfitted kernel smoothing of nonlinear additive autoregression model", *The Annals of Statistics*, **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.

Zhang, W., Lee, S. Y. and Song, X. (2002), "Local polynomial fitting in semivarying coefficient model", *Journal of Multivariate Analysis*, **82**, 166-188.

Zhang, X. and Liang, H. (2011) Focused information criterion and model averaging for generalized additive partial linear models *The Annals of Statistics*, **39**, 174-200.

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