Cen Wu[1] / Ping-Shou Zhong[2] / Yuehua Cui[2]

Additive varying-coefficient model for nonlinear gene-environment interactions

[1] Department of Statistics, Kansas State University, Manhattan, KS 66506, USA
[2] Department of Statistics and Probability, Michigan State University, East Lansing, MI 48824, USA, E-mail: cui@stt.msu.edu

Abstract:
Gene-environment (G×E) interaction plays a pivotal role in understanding the genetic basis of complex disease. When environmental factors are measured continuously, one can assess the genetic sensitivity over different environmental conditions on a disease trait. Motivated by the increasing awareness of gene set based association analysis over single variant based approaches, we proposed an additive varying-coefficient model to jointly model variants in a genetic system. The model allows us to examine how variants in a gene set are moderated by an environment factor to affect a disease phenotype. We approached the problem from a variable selection perspective. In particular, we select variants with varying, constant and zero coefficients, which correspond to cases of G×E interaction, no G×E interaction and no genetic effect, respectively. The procedure was implemented through a two-stage iterative estimation algorithm via the smoothly clipped absolute deviation penalty function. Under certain regularity conditions, we established the consistency property in variable selection as well as effect separation of the two stage iterative estimators, and showed the optimal convergence rates of the estimates for varying effects. In addition, we showed that the estimate of non-zero constant coefficients enjoy the oracle property. The utility of our procedure was demonstrated through simulation studies and real data analysis.

Keywords: B-spline, gene-set analysis, local quadratic approximation, SCAD penalty, variable selection
DOI: 10.1515/sagmb-2017-0008

1 Introduction

Complex human diseases are determined not only by genetic variants, but may also be affected by environmental factors and the interplay between them. Changes in gene expression under different environmental conditions reveal the interaction between genes and the environment. These changes are less likely due to changes in the gene sequence itself, but to structural changes such as DNA methylation or histone modification that consequently play a regulatory role and modulate gene expression. Such epigenetic changes have been increasing recognized as the epigenetic basis of gene-environment (G×E) interaction (Liu, Li & Tollefsbol, 2008). Identification of G×E interaction could shed novel insights into the phenotypic plasticity of complex disease phenotypes (Feinberg, 2004).

In a typical G×E interaction study, the environmental factor can be either discrete or continuous. For example, smoking can be a discrete variable when evaluating the risk of asthma. When environmental variables are measured on a continuous scale, a clearer picture of the interaction can be assessed since the varying patterns of genetic effects responsive to environmental changes can be traced, leading to a better understanding of the genetic heterogeneity under different environmental stimuli (Ma et al., 2011; Wu & Cui, 2013). As illustrated in Wu and Cui (2013), one can assess the nonlinear G×E interaction when an environmental factor is measured in a continuous scale. For example, individual obesity can be a factor when evaluating the risk of hypertension. One can assess the nonlinear effect of a genetic factor on the risk of hypertension considering the heterogeneity of individual obese conditions in a population, leading to a better understanding of disease heterogeneity.

When assessing G×E interactions, investigators have focused predominantly on single variant based analysis, such as the parametric methods in Guo (2000), the semi-parametric methods in Chatterjee and Carroll (2005), Chen, Chatterjee, and Carroll (2013), and Maity et al. (2009), and the non-parametric methods in Ma et al. (2011) and Wu and Cui (2013). Recently, there has been a significant increase in set-based genetic association studies focusing on a set of variants, for example, the gene-centric analysis of Cui et al. (2008), the gene-set analysis of Schaid et al. (2012) and Efron and Tibshirani (2007), and the pathway-based analysis of Wang, Li, and Hakonarson (2011). By assessing the joint function of multiple variants in a set, one can obtain a better interpretation of the disease signals and gain novel insights into disease etiology. Motivated by these set-based

Yuehua Cui is the corresponding author.

association studies, we propose a set-based framework to investigate how variants in a gene-set moderated by an environment factor affect disease and in what form.

In a typical set-based association study, the number of variants d within a genetic system can be relatively large compared to the sample size, which makes the regression coefficients estimation instable. The problem can be approached from the perspective of variable selection. In this work, we extend our previous work on nonlinear gene-environment interaction from a single variant based analysis to a multiple variant based analysis under a penalized regression framework. We include variants that belong to a particular gene-set or pathway which potentially interact with one or multiple environment factors through an additive varying-coefficient model. We propose to select genetic variants with coefficient functions that are varying, non-zero constant and zero corresponding to cases with G×E interactions, no G×E interactions and no genetic effects, respectively. Our approach employs the power and merits of variable selection by simultaneously fitting all the variants in a genetic system into a regression model, therefore avoiding the limitation of multiple testing corrections, especially when the data dimension is large.

This paper is organized as follows. In Section 2, we describe the penalized least square estimation procedure via B-spline basis expansion and smoothly clipped absolute deviation (SCAD) penalty, as well as the computational algorithms. We also present the theoretical results including consistency in variable selection and show the optimal convergence rates of the estimates of varying effects. We show that the estimates of non-zero constant coefficients enjoy the oracle property in the sense that the asymptotic distribution of the non-zero constant coefficient function is the same as that when the true model is known a priori. The merit of the proposed method is demonstrated through extensive simulation studies in Section 3 and real data analysis in Section 4. The technical proofs are relegated to the A.

2 Methods

2.1 Additive varying-coefficient model with SCAD penalty

Throughout this paper, we assume an environment variable (Z) is continuously measured through which we can model the nonlinear interaction effect. For simplicity, we start the presentation with one environmental factor. Extension to multiple environmental factors are given in the end. Let $(\mathbf{X}_i, Y_i, Z_i)$, $i = 1, ..., n$ be independent and identically distributed (i.i.d.) random vectors, then the varying coefficient (VC) model, initially proposed by Hastie and Tibshirani (1993), has the form

$$Y_i = \sum_{j=0}^{d} \beta_j(Z_i) X_{ij} + \varepsilon_i, \tag{1}$$

where X_{ij} is the jth component of (d+1)-dimensional genetic vector $\mathbf{X}_i$ with the first component X_{i0} being 1, $\beta_j(\cdot)$'s are unknown varying-coefficient functions, Z_i is the environmental variable, and ε_i is the random error such that $E(\varepsilon_i|X, Z) = 0$ and $Var(\varepsilon_i|X, Z) = \sigma^2 < \infty$. In the model, we assume there are a total of d genetic variants which are moderated by a common environmental factor Z.

The smooth functions $\{\beta_j(\cdot)\}_{j=0}^{d}$ in (1) can be approximated by polynomial splines. Without loss of generality, suppose that $Z \in [0, 1]$. Let w_k be a partition of the interval [0,1], with k_n uniform interior knots

$$w_k = \{0 = w_{k,0} < w_{k,1} < ... < w_{k,k_n} < w_{k,k_n+1} = 1\}, \text{for } k = 0, \cdots, d.$$

Let $\mathcal{F}_n$ be a collection of functions on [0,1] satisfying: (1) the function is a polynomial of degree p or less on subintervals $I_s = [w_{k,s}, w_{k,s+1}), s = 0, \ldots, k_n - 1$ and $I_{k_n} = [w_{j,k_n}, w_{j,k_n+1})$; and (2) the functions are $p - 1$ times continuously differentiable on [0,1]. Let $\bar{B}(\cdot) = \{\bar{B}_{jl}(\cdot)\}_{l=1}^{L_j}$ be a set of normalized B spline basis of $\mathcal{F}_n$. Then for $j = 0, ..., d$, the VC functions can be approximated by basis functions $\beta_j(Z) \approx \sum_{l=1}^{L_j} \bar{\gamma}_{jl} \bar{B}_{jl}(Z)$, where L_j is the number of basis functions in approximating the function $\beta_j(Z)$. By changing the equivalent basis, the basis expansion can be reexpressed as

$$\beta_j(\cdot) \approx \sum_{l=1}^{L_j} \gamma_{j,l} B_{j,l}(\cdot) \doteq \gamma_{j,1} + \tilde{B}_j^T(\cdot)\gamma_{j,*},$$

where the spline coefficient vector $\boldsymbol{\gamma}_j = (\gamma_{j,1}, \boldsymbol{\gamma}_{j*}^T)^T$, and $\tilde{B}_j(\cdot) = (B_{j2}(\cdot), \ldots, B_{jL_j}(\cdot))^T$; $\gamma_{j,1}$ and $\boldsymbol{\gamma}_{j*}$ correspond to the constant and varying part of the coefficient function, respectively (Schumaker, 1981). We treat $\boldsymbol{\gamma}_{j*}$ as a group. If $\|\boldsymbol{\gamma}_{j*}\|_2 = 0$, then the jth predictor only has a non-zero constant effect; if $\gamma_{j,1} = 0$, then the predictor is redundant.

To carry out variable selection separating the varying, non-zero constant, and zero effects, we minimize the penalized least square function,

$$\begin{aligned} Q(\boldsymbol{\gamma}) =& \frac{1}{n}\sum_{i=1}^{n}\left[Y_i - \sum_{j=0}^{d}\sum_{l=1}^{L}\gamma_{j,l}X_{ij}B_{jl}(Z_i)\right]^2 + \sum_{j=1}^{d} p_{\lambda_1}(\|\boldsymbol{\gamma}_{j*}\|_2) \\ &+ \sum_{j=1}^{d} p_{\lambda_2}(|\gamma_{j,1}|)I(\|\boldsymbol{\gamma}_{j*}\|_2 = 0), \end{aligned} \tag{2}$$

where λ_1 and λ_2 are the penalization parameters, $p_\lambda(\cdot)$ is the SCAD penalty function, defined as

$$p_\lambda(u) = \begin{cases} \lambda u & \text{if } 0 \le u \le \lambda \\ -\frac{(u^2 - 2a\lambda u + \lambda^2)}{2(a-1)} & \text{if } \lambda < u \le a\lambda \\ \frac{(a+1)\lambda^2}{2} & \text{if } u > a\lambda. \end{cases} . \tag{3}$$

In matrix notation, (2) can be reexpressed as,

$$\begin{aligned} Q(\boldsymbol{\gamma}) =& (\boldsymbol{Y} - \boldsymbol{U}\boldsymbol{\gamma})^T(\boldsymbol{Y} - \boldsymbol{U}\boldsymbol{\gamma})/n + \sum_{j=1}^{d} p_{\lambda 1}(\|\boldsymbol{\gamma}_{j*}\|_2) \\ &+ \sum_{j=1}^{d} p_{\lambda 2}(|\gamma_{j,1}|)I(\|\boldsymbol{\gamma}_{j*}\|_2 = 0), \end{aligned} \tag{4}$$

where $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$, $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_0^T, \ldots, \boldsymbol{\gamma}_d^T)^T$, and $\boldsymbol{U} := \boldsymbol{U}(\boldsymbol{X}, Z) = (U_1^T, \ldots, U_n^T)^T$ with $U_i = (X_{i0}B(Z_i)^T, \ldots, X_{id}B(Z_i)^T)^T$. The first penalty function in (4) is to separate the varying and constant effects by penalizing the L_2 norm of the varying part of the coefficient functions. The indicator function in the 2nd penalty term helps to penalize the variables of the constant effects. Both $\gamma_{j,1}$ and $\boldsymbol{\gamma}_{j*}$ will be shrunk to zero if predictor X_j has no genetic effect. Since the indicator function in $Q(\boldsymbol{\gamma})$ leads to much difficulty in optimizing the penalized loss function, we resort to the two stage iterative framework of great computational convenience described in 2.2. It can be shown that the estimator from the iterative procedure is asymptotically equivalent to the minimizer in (2) by the arguments in the proof of Theorem 1 and 2 in the A.

2.2 Computation algorithm

The SCAD penalty function is singular at the origin, and does not have continuous 2nd order derivatives, therefore the regular gradient-based optimization cannot be applied. In this section, we develop an iterative two-stage algorithm to minimize the penalized loss function using local quadratic approximation (LQA) to the SCAD penalty. The two-stage strategy was adopted in Tang et al. (2012) for penalized quantile regression with adaptive LASSO penalty. As in Fan and Li (2001), in a neighborhood of a given positive $u_0 \in \mathbb{R}^+$,

$$p_\lambda(u) \approx p_\lambda(u_0) + \frac{p_\lambda'(u_0)}{2u_0}(u^2 - u_0^2),$$

where $p_\lambda'(u) = \lambda\{I(u \leqslant \lambda) + \frac{(a\lambda - u)_+}{(a-1)\lambda}I(u > \lambda)\}$ for $u > 0$ and $a = 3.7$. Here we use a similar quadratic approximation by substituting u with $\|\boldsymbol{\gamma}_{j*}\|_2$ and $|\gamma_{j1}|$ in LQA, for $j = 1, \ldots, d$. Given an initial value of $\boldsymbol{\gamma}_j^0$ such that $\|\boldsymbol{\gamma}_{j*}\|_2 \neq 0$ and $|\gamma_{j1}| \neq 0$, we have

$$p_\lambda(\|\boldsymbol{\gamma}_{j*}\|_2) \approx p_\lambda(\|\boldsymbol{\gamma}_{j*}^0\|_2) + \frac{p_\lambda'(\|\boldsymbol{\gamma}_{j*}^0\|_2)}{2\|\boldsymbol{\gamma}_{j*}^0\|_2}(\|\boldsymbol{\gamma}_{j*}\|_2^2 - \|\boldsymbol{\gamma}_{j*}^0\|_2^2) \tag{5}$$

and

$$p_\lambda(|\boldsymbol{\gamma}_{j,1}|) \approx p_\lambda(|\boldsymbol{\gamma}^0_{j,1}|) + \frac{p'_\lambda(|\boldsymbol{\gamma}^0_{j,1}|)}{2|\boldsymbol{\gamma}^0_{j,1}|}(|\boldsymbol{\gamma}_{j,1}|^2 - |\boldsymbol{\gamma}^0_{j,1}|^2). \tag{6}$$

The sets of predictors with varying, non-zero constant, and zero effects are denoted by $\mathscr{V}$, $\mathscr{C}$ and $\mathscr{Z}$ respectively. We implement the iterative algorithm in the following two-stage procedure. In stage 1, using the LQA (5) and dropping the irrelevant constant terms, we minimize

$$Q_1(\boldsymbol{\gamma}) = (\boldsymbol{Y} - \boldsymbol{U}\boldsymbol{\gamma})^T(\boldsymbol{Y} - \boldsymbol{U}\boldsymbol{\gamma}) + \frac{n}{2}\boldsymbol{\gamma}^T\boldsymbol{\Omega}_{\lambda_1}(\boldsymbol{\gamma}_0)\boldsymbol{\gamma}, \tag{7}$$

where the initial spline vector $\boldsymbol{\gamma}_0$ is the unpenalized estimator, $\boldsymbol{\Omega}_{\lambda_1}(\boldsymbol{\gamma}_0) = \mathrm{diag}\{\boldsymbol{\Omega}_0, \boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_d\}$, where $\boldsymbol{\Omega}_0 = \mathbf{0}_L$, $\boldsymbol{\Omega}_j = \left\{0, \frac{p^T_{\lambda 1}(\|\boldsymbol{\gamma}^0_{j*}\|_2)}{\|\boldsymbol{\gamma}^0_{j*}\|_2}, \ldots, \frac{p^T_{\lambda 1}(\|\boldsymbol{\gamma}^0_{j*}\|_2)}{\|\boldsymbol{\gamma}^0_{j*}\|_2}\right\}_L$ for $j = 1, \ldots, d$. Hence the estimator can be iteratively obtained as

$$\hat{\boldsymbol{\gamma}}^{\mathscr{V}\mathscr{C}(m)} = \left\{\boldsymbol{U}^T\boldsymbol{U} + \frac{n}{2}\boldsymbol{\Omega}_{\lambda_1}(\hat{\boldsymbol{\gamma}}^{\mathscr{V}\mathscr{C}(m-1)})\right\}^{-1}\boldsymbol{U}^T\boldsymbol{Y}. \tag{8}$$

If all the predictors are in $\mathscr{V}$ at the beginning, then the jth predictor will be moved to $\mathscr{C}$ if $\|\hat{\boldsymbol{\gamma}}^{\mathscr{V}\mathscr{C}(m)}_{j*}\|_2 = 0$, otherwise it will stay in $\mathscr{V}$.

In stage 2, using the LQA (6) and dropping the irrelevant constant terms, we minimize the following penalized loss only for the predictors in $\mathscr{C}$,

$$Q_2(\boldsymbol{\gamma}) = (\boldsymbol{Y} - \boldsymbol{U}\boldsymbol{\gamma})^T(\boldsymbol{Y} - \boldsymbol{U}\boldsymbol{\gamma}) + \frac{n}{2}\boldsymbol{\gamma}^T\boldsymbol{\Omega}_{\lambda_2}(\hat{\boldsymbol{\gamma}}^{\mathscr{V}\mathscr{C}})\boldsymbol{\gamma}, \tag{9}$$

where $\boldsymbol{\Omega}_{\lambda_2}(\hat{\boldsymbol{\gamma}}^{\mathscr{V}\mathscr{C}}) = \mathrm{diag}\{\boldsymbol{\Omega}_0, \boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_d\}$ with $\boldsymbol{\Omega}_0 = \mathbf{0}_L$, $\boldsymbol{\Omega}_j = \left\{\frac{p^T_{\lambda 2}(|\hat{\gamma}^{\mathscr{V}\mathscr{C}}_{j,1}|)}{|\hat{\gamma}^{\mathscr{V}\mathscr{C}}_{j,1}|}I(\|\hat{\boldsymbol{\gamma}}^{\mathscr{V}\mathscr{C}}_{j*}\|_2 = 0), 0, \ldots, 0\right\}_L$. The estimator can be iteratively obtained as

$$\hat{\boldsymbol{\gamma}}^{\mathscr{C}\mathscr{Z}(m)} = \left\{\boldsymbol{U}^T\boldsymbol{U} + \frac{n}{2}\boldsymbol{\Omega}_{\lambda_2}(\hat{\boldsymbol{\gamma}}^{\mathscr{C}\mathscr{Z}(m-1)})\right\}^{-1}\boldsymbol{U}^T\boldsymbol{Y}. \tag{10}$$

If the jth predictor is in $\mathscr{C}$, then it will be moved to $\mathscr{Z}$ if $|\hat{\gamma}^{\mathscr{C}\mathscr{Z}}_{k,1}|=0$, otherwise it stays in $\mathscr{C}$.

We can obtain the estimator $\hat{\boldsymbol{\gamma}}$ at convergence from the iterative procedure between the two stages above, and the estimated coefficient function in (1) as $\hat{\beta}_j(z) = B^T(z)\hat{\boldsymbol{\gamma}}_j$. $\hat{\boldsymbol{\beta}}_j(z)$ will be a varying function, non-zero constant and zero if $\hat{\boldsymbol{\gamma}}_j$ is in $\mathscr{V}$, $\mathscr{C}$ and $\mathscr{Z}$ correspondingly.

2.3 Choosing the tuning parameters

We choose the number of interior knots k_n, the degree of the spline basis p, and the tuning parameters λ_1 and λ_2 by a data driven procedure. Here p and k_n control the smoothness of the coefficient functions, while λ_1 and λ_2 determine the threshold for variable selection. The Schwarz BIC criterion (1978) was used to choose k_n and p. Due to heavy computational costs, it becomes infeasible to simultaneously select p and k_n for each varying-coefficient function. Thus, we assume the same p and k_n for the varying-coefficient functions. The range for k_n is $[\max(\lfloor 0.5n^{\frac{1}{(2p+3)}}\rfloor, 1), \lfloor 1.5n^{\frac{1}{(2p+3)}}\rfloor]$, where $\lfloor x\rfloor$ denotes the integer part of x. The optimal pair of k_n and p can be selected via a two-dimensional grid search, according to the following criterion:

$$\mathrm{BIC}_{k_n,p} = \log(\mathrm{RSS}_{k_n,p}) + \frac{(k_n + p + 1)}{n}\log(n),$$

where $\mathrm{RSS}_{k_n,p} = (\boldsymbol{Y} - \boldsymbol{U}\hat{\boldsymbol{\gamma}})^T(\boldsymbol{Y} - \boldsymbol{U}\hat{\boldsymbol{\gamma}})$, $\hat{\boldsymbol{\gamma}} = (\hat{\boldsymbol{\gamma}}_0^T, \mathbf{0}^T, \ldots, \mathbf{0}^T)^T$. Conditional on the selected k_n and p, λ_1 is the minimizer of

$$\mathrm{BIC}_{\lambda_1} = \log(\mathrm{RSS}_{\lambda_1}) + \frac{df_{\lambda_1}}{n}\log(n),$$

where $\mathrm{RSS}_{\lambda_1} = (\boldsymbol{Y} - \boldsymbol{U}\hat{\boldsymbol{\gamma}}_{\lambda_1})^T(\boldsymbol{Y} - \boldsymbol{U}\,\hat{\boldsymbol{\gamma}}_{\lambda_1})$, $\hat{\boldsymbol{\gamma}}_{\lambda 1}$ is the minimizer of (7), and df_{λ_1} is the effective degree of freedom, defined as the total number of predictors in $\mathscr{V}$ and $\mathscr{C}$.

Conditional on $\hat{\boldsymbol{\gamma}}_{\lambda_1}$, λ_2 is the minimizer of

$$\mathrm{BIC}_{\lambda_2} = \log(\mathrm{RSS}_{\lambda_2}) + \frac{df_{\lambda_2}}{n}\log(n),$$

where $\mathrm{RSS}_{\lambda_2} = (\boldsymbol{Y} - \boldsymbol{U}\hat{\boldsymbol{\gamma}}_{\lambda_2})^T(\boldsymbol{Y} - \boldsymbol{U}\hat{\boldsymbol{\gamma}}_{\lambda_2})$, $\hat{\boldsymbol{\gamma}}_{\lambda_2}$ is the minimizer of (9), and df_{λ_2} is the effective degree of freedom, defined similarly as df_{λ_1}.

2.4 Asymptotic results

Here we establish the asymptotic properties of the penalized least square estimators. Without loss of generality, we assume there are v varying coefficients as $\beta_j(\cdot) \equiv \beta_j(z), j = 1, \ldots, v$, $(c - v)$ non-zero constant coefficients as $\beta_j(\cdot) \equiv \beta_j > 0, j = v + 1, \ldots, c$, and $(d - c)$ zero coefficients as $\beta_j(\cdot) \equiv 0, j = (c + 1), \ldots, d$. Our asymptotic results are based on the following assumptions.

(A1) Let $\mathscr{H}_r$ be the collection of all functions on the compact support [0,1] such that the r_1th order derivatives of the functions are Hölder of order r_2 with $r = r_1 + r_2$, i.e. $|h^{r_1}(z_1) - h^{r_1}(z_2)| \le C_0|z_1 - z_2|^{r_2}$ where $0 \le z_1, z_2 \le 1$ and C_0 is a finite positive constant. Then $\beta_j(z) \in \mathscr{H}_r, j = 0, 1, \ldots, v$, for some $r \ge \frac{3}{2}$.

(A2) The density function of the index variable Z, $f(z)$, is continuous and bounded away from 0 and infinity on $[0, 1]$, i.e. there exist finite positive constants C_1 and C_2 such that $C_1 \le f(z) \le C_2$ for all $z \in [0, 1]$.

(A3) Let $\tilde{\lambda}_0 \le \ldots \le \tilde{\lambda}_d$ be the eigenvalues of $E[\mathbf{X}\mathbf{X}^T|Z = z]$. Assume that $\tilde{\lambda}_j$ $(k = 0, \ldots, d)$ are uniformly bounded away from 0 and infinity in probability. In addition, the random design vectors are bounded in probability.

(A4) For w_j, the partition of the compact interval [0,1] defined as $\{0 = w_{j,0} < w_{j,1} < \ldots < w_{j,k_n} < w_{j,k_n+1} = 1\}$, $j = 0, \ldots, d$, there exists a finite positive constant C_3 such that

$$\frac{\max(w_{j,k+1} - w_{j,k}, k = 0, \ldots, k_n)}{\min(w_{j,k+1} - w_{j,k}, k = 0, \ldots, k_n)} \le C_3.$$

(A5) The tuning parameters satisfy $k_n^{\frac{1}{2}}\max\{\lambda_1, \lambda_2\} \to 0$ and $n^{\frac{1}{2}}k_n^{-1}\min\{\lambda_1, \lambda_2\} \to \infty$.

(A6) $b_n := \max_j\{|p''_{\lambda_1}(\|\tilde{\boldsymbol{\gamma}}_{j*}\|)|, |p''_{\lambda_2}(|\tilde{\gamma}_{j,1}|)| : \tilde{\boldsymbol{\gamma}}_{j*} \ne \mathbf{0}, \tilde{\gamma}_{j,1} \ne 0\} \to 0$ as $n \to \infty$, where $\tilde{\boldsymbol{\gamma}}_j$ is defined in the A.

(A7) $\liminf_{n\to\infty}\liminf_{\theta\to 0^+}\lambda_1^{-1}p'_{\lambda_1}(\theta) > 0$ and $\liminf_{n\to\infty}\liminf_{\theta\to 0^+}\lambda_2^{-1}p'_{\lambda_2}(\theta) > 0$

The above assumptions are commonly used in the literature on polynomial splines and variable selections. An assumption similar to (A1) is found in Kim (2007) and Tang et al. (2012). (A1) guarantees certain degrees of smoothness of the true coefficient function in order to improve goodness of approximation. (A2) and (A3) are similar to those in Huang, Wu, and Zhou (2002, 2004)) and Wang, Li, and Huang (2008). (A4) suggests that the knot sequence is quasi-uniform on [0,1], as in Schumaker (1981). (A5–A7) are conditions on tuning parameters, of which (A5) was reported by Tang et al. (2012) while (A6) and (A7) are similar to those in Fan and Li (2001) and Wang, Li, and Huang (2008).

Theorem 1

Under the assumptions (A1–A7) and suppose $k_n = O\left(n^{\frac{1}{2r+1}}\right)$, then we have

(1) $\hat{\beta}_j(z)$ are nonzero constant, $j = v + 1, \ldots, c$ and $\hat{\beta}_j(z) = 0, j = c + 1, \ldots, d$, with probability approaching 1;

(2) $|\hat{\boldsymbol{\beta}}_j(z) - \boldsymbol{\beta}_j(z)| = O_p(n^{\frac{-r}{2r+1}}), j = 0, \ldots, v$ for any fixed z.

The proof can be found in A. Denote $\beta^* = (\beta_{v+1}, \ldots, \beta_c)^T$ as the vector of true nonzero constant coefficients. The following theorem establishes the asymptotic normality of $\hat{\beta}^*$.

Theorem 2

Under the assumptions (A1–A7) and suppose $k_n = O(n^{\frac{1}{2r+1}})$, then as $n \to \infty$,

$$\sqrt{n}\Sigma^{\frac{1}{2}}(\hat{\beta}^* - \beta^*) \xrightarrow{d} \mathcal{N}(0, \sigma^2 I_{c-v}),$$

where Σ is defined as 22 in A, and $\sigma^2 = E(\varepsilon_i^2)$.

3 Simulation

The performance of the proposed method was demonstrated through extensive simulation studies. We used the percentage of choosing the true model out of total R replicates, defined as the oracle percentage, to evaluate the accuracy of variable selection by identifying varying, non-zero constant and zero effects. The precision of estimation was assessed by integrated mean squared error (IMSE). Let $\hat{\beta}_j^{(r)}$ be the estimator of a nonparametric function β_j in the rth ($1 \leqslant r \leqslant R$) replication, and $\{z_m\}_{m=1}^{n_{\text{grid}}}$ be the grid points where $\hat{\beta}_j^{(r)}$ was evaluated. We used the integrated mean squared error (IMSE) of $\hat{\beta}_k(z)$, defined as

$$\text{IMSE}(\hat{\beta}_j(z)) = \frac{1}{R}\sum_{r=1}^{R}\frac{1}{n_{\text{grid}}}\sum_{m=1}^{n_{\text{grid}}}\{\hat{\beta}_k^{(r)}(z_m) - \beta_j(z_m)\}^2,$$

to evaluate the estimation accuracy of coefficient β_j, and the total integrated mean squared error (TIMSE) of all the d coefficients, defined as TIMSE=$\sum_{j=1}^{d}\hat{\beta}_j(z)$, to evaluate the overall estimation accuracy. Note that IMSE($\hat{\beta}_j$) is reduced to MSE($\hat{\beta}_j$) when $\hat{\beta}_j$ is a constant. The percentage of correctly selecting each individual true functions (defined as the selection ratio) was used to evaluate the selection performance.

We considered multiple genetic factors X obtained from a gene-set or pathway, with the following additive VC model,

$$Y_i = \beta_0(Z_i) + \sum_{j=1}^{d}\beta_j(Z_i)X_{ij} + \varepsilon_i,$$

where SNP X_i's were coded with 3 categories $(1, 0, -1)$ for genotypes (AA, Aa, aa) respectively. We simulated the SNP genotype data based on the pairwise linkage disequilibrium(LD) structure. Suppose the two risk alleles A and B of two adjacent SNPs have the minor allele frequencies (MAFs) p_A and p_B, respectively, with LD denoted as δ. Then the frequencies of four haplotypes can be expressed as $p_{ab} = (1-p_A)(1-p_B)+\delta, p_{Ab} = p_A(1-p_B)-\delta, p_{aB} = (1-p_A)p_B - \delta$, and $p_{AB} = p_A p_B + \delta$. Assuming Hardy-Weinberg equilibrium, the SNP genotype at locus 1 can be simulated assuming a multinomial distribution with frequencies p_A^2, $2p_A(1-p_A)$ and $(1-p_A)^2$ for genotypes AA, Aa, aa, respectively. We can then simulate genotype for locus 2 based on the conditional probability. For example, $P(BB|AA) = p_{AB}^2/p_{AA}$, $P(Bb|AA) = p_{AB}p_{Ab}/p_{AA}$ and $P(bb|AA) = p_{ab}^2/p_{AA}$. So conditional on genotype AA at locus 1, the genotype at locus 2 can be generated according to a multinomial distribution with the derived probabilities. The advantage of this simulation is that we can control the pairwise LD structure between adjacent SNPs. We assumed pairwise correlation of $r = 0.5$ which leads to $\delta = r\sqrt{(p_A(1-p_A)p_B(1-p_B))}$. To save space, we omitted the detailed simulation information which can be found in Cui et al. (2008). The coefficient functions were set as: $\beta_1(z) = \sin(2\pi z)$, $\beta_2(z) = 2 - 3\cos\{(6z-5)\pi/3\}$, $\beta_3(z) = 3(2z-1)^3$, $\beta_4(z) = 2$, $\beta_5(z) = 2.5$, and $\beta_j(z) = 0$ for $j > 5$. We evaluated the performance under $n = 500$ with 500 replicates. Better performance results for large samples ($n > 500$) were observed, but were omitted to save space.

Figure 1 shows the selection ratio when $d = 10$, under different combinations of MAF and error distribution. The height of the bars represents the selection percentage out of 500 replicates. The selection performance is better under the normal error distribution, with relatively higher selection rate for the first five true functions and lower false selection ratio for the rest, compared to the results obtained under the $t(3)$ error. In genetic association studies, model performance generally improves as the MAF increases. The same trend is observed under our variable selection framework. For example, a higher false selection ratio was observed under the $t(3)$ error when $p = 0.1$. The false selection ratio decreases as MAF increases to 0.3. The result for $d = 50$ is presented in Figure 2, which shows a very similar pattern. The results demonstrate the stable performance of the proposed variable selection method.

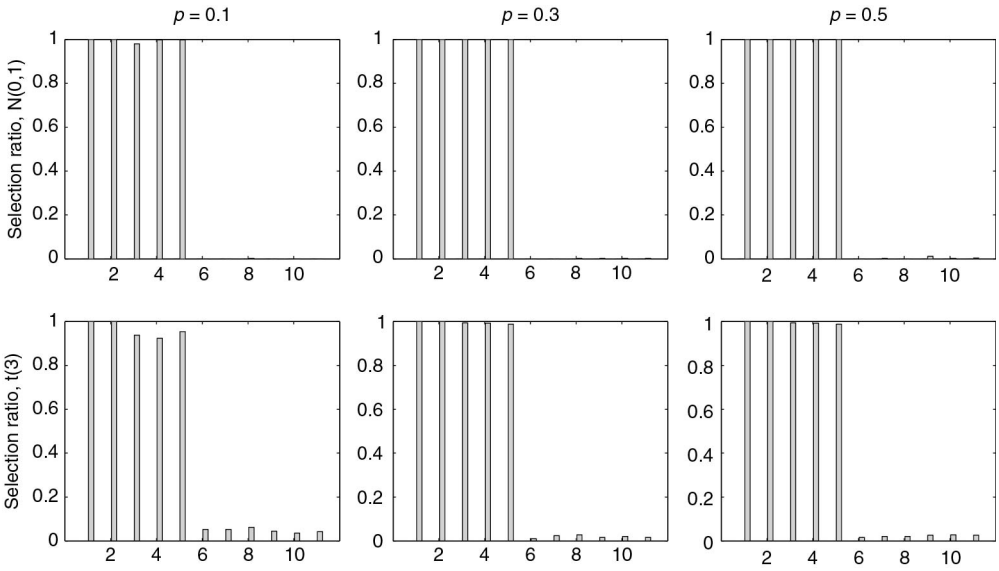

Figure 1: The selection ratio under different error distributions for different coefficient functions when $d = 10$. The horizontal axis represents the SNPs.

Table 1 lists the oracle percentage (%) of choosing the true model out of all the simulation replicates, the IMSE (inside the panel), and TIMSE (the last row) for the case with $d = 10$. In general, the model selection performance improves as the MAF increases from 0.1 to 0.5. For example, the oracle percentage increases from 0.72 to 0.91 under the $t(3)$ error with SCAD penalty, when the MAF increases from 0.1 to 0.3. We observed dramatic reduction on the IMSE and TIMSE as the MAF increases. Under the normal error, the TIMSE is 0.4205 which reduces to 0.2007 when the MAF increases to 0.3 and further reduces to 0.1895 when $p = 0.5$. This result is consistent with the general observation in a genetic association study in which typically a model performs better as the MAF increases. It is worth mentioning that we observed dramatic improvement in model performance when the MAF increases from 0.1 to 0.3, compared to the improvement when the MAF increases from 0.3 to 0.5. For example, the IMSE for $\beta_2(u)$ reduces from 0.3285 to 0.1600, a 51% reduction when p increases from 0.1 to 0.3, while there is only a 1% reduction when p increases from 0.3 to 0.5 under the $t(3)$ error distribution for the SCAD penalty. This empirical observation shows the stable performance of the model under moderate allele frequency.

Table 1: List of IMSE, TIMSE, and Oracle percentage (%) under $\mathcal{N}(0, 1)$ and $t(3)$ error distributions when $d = 10$.

	$p = 0.1$				$p = 0.3$				$p = 0.5$			
	$\mathcal{N}(0,1)$ error		$t(3)$ error		$\mathcal{N}(0,1)$ error		$t(3)$ error		$\mathcal{N}(0,1)$ error		$t(3)$ error	
	SCAD	Oracle[2]	SCAD	Oracle	SCAD	Oracle	SCAD	Oracle	SCAD	Oracle	SCAD	Oracle
Oracle %[1]	0.976	1	0.72	1	0.992	1	0.91	1	0.98	1	0.894	1
$\beta_1(u)$	0.0863	0.0891	0.3078	0.2247	0.0268	0.0273	0.0607	0.0601	0.0213	0.0214	0.0431	0.0451
$\beta_2(u)$	0.1611	0.1667	0.3285	0.3557	0.1071	0.1174	0.1600	0.1746	0.1044	0.1106	0.1581	0.1725
$\beta_3(u)$	0.1264	0.1238	0.4890	0.2932	0.0561	0.0637	0.1360	0.1320	0.0497	0.0604	0.1101	0.1170
$\beta_4(u)$	0.0270	0.0192	1.3307	0.0643	0.0086	0.0084	0.1111	0.0237	0.0077	0.0077	0.0439	0.0192
$\beta_5(u)$	0.0191	0.0174	0.2943	0.0475	0.0066	0.0065	0.0443	0.0222	0.0063	0.0063	0.0240	0.0135
TIMSE	0.4205	0.4162	2.9342	0.9855	0.2007	0.2233	0.5311	0.4126	0.1895	0.206	0.4072	0.3673

[1]Oracle % refers to the percentage of selecting all variables that are used to generate the phenotype Y;
[2]Oracle refers to the oracle IMSE, that is, the IMSE calculated assuming that we know the true regression model.

Another observation from the simulation is that the model performs better under the normal error than under the $t(3)$ error. We observed a larger oracle percentage, smaller IMSE and TIMSE for the coefficient functions under the normal error compared to the $t(3)$ error. For example, the TIMSE for the SCAD penalty is 0.4205 under the normal error, while it is 2.9342 under the $t(3)$ error for fixed $p = 0.1$. In addition, the oracle percentage, IMSE and TIMSE under the normal error are all quite similar as those obtained as if the truth were known (the oracle) in all cases, demonstrating the stable selection performance of the SCAD penalty.

A similar pattern was observed when the data dimension increases to 50 (Table 2). As the MAF increases from 0.1 to 0.3, we observed sharply decreased IMSE and TIMSE. Compared to the low dimensional case when $d = 10$, the performance under $p = 0.1$ is relatively unstable. For example, the TIMSE for the SCAD method is 3.3644 when $d = 50$, compared to 0.4205 when $d = 10$ under the normal error and $p = 0.1$. However, we observed dramatic reduction in TIMSE when the MAF increases to 0.3 under $d = 50$. Thus, one has to be very careful about the interpretation of the selection result under low MAF in real data analysis. We did additional simulations when the sample size increases to 1000 and observed consistently improved results under different scenarios (data not shown). In summary, the SCAD penalty function shows consistently good performance and can separate varying, constant and zero effects under moderate allele frequencies. Coupled with the results shown in Figure 1 and Figure 2, the proposed variable selection method shows relatively stable performance to assess gene-environment interactions.

Table 2: List of IMSE, TIMSE, and Oracle percentage (%) under $\mathcal{N}(0, 1)$ and $t(3)$ error distributions when $d = 50$.

	$p = 0.1$				$p = 0.3$				$p = 0.5$			
	$\mathcal{N}(0,1)$ error		$t(3)$ error		$\mathcal{N}(0,1)$ error		$t(3)$ error		$\mathcal{N}(0,1)$ error		$t(3)$ error	
Oracle %	**SCAD 0.908**	**Oracle 1**	**SCAD 0.435**	**Oracle 1**	**SCAD 0.986**	**Oracle 1**	**SCAD 0.745**	**Oracle 1**	**SCAD 0.988**	**Oracle 1**	**SCAD 0.87**	**Oracle 1**
$\beta_1(u)$	0.1929	0.0884	0.5687	0.2209	0.0289	0.0278	0.0860	0.0599	0.0215	0.0216	0.0450	0.0434
$\beta_2(u)$	0.2064	0.1684	0.3851	0.3340	0.1107	0.1137	0.1858	0.1742	0.1048	0.1123	0.1551	0.1608
$\beta_3(u)$	0.5235	0.1218	0.6934	0.2614	0.0817	0.0646	0.2205	0.1301	0.0608	0.0579	0.1754	0.1085
$\beta_4(u)$	2.0918	0.0196	2.4522	0.0484	0.1083	0.0075	0.3865	0.0254	0.0470	0.0078	0.1681	0.0167
$\beta_5(u)$	0.3475	0.0158	0.5996	0.0445	0.0229	0.0068	0.0840	0.0220	0.0120	0.0053	0.0480	0.0190
TIMSE	3.3644	0.4140	5.7021	0.9092	0.3526	0.2204	1.2288	0.4117	0.2461	0.2050	0.6492	0.3484

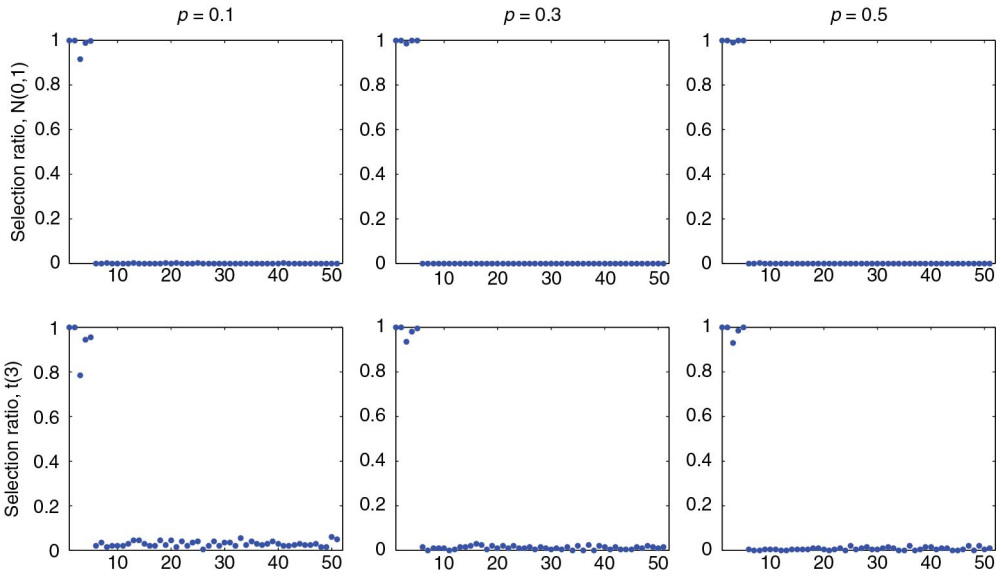

Figure 2: The selection ratio under different error distributions for different coefficient functions when $d = 50$. The horizontal axis represents the SNPs.

To further assess the false positive controls of the proposed method, we generated the response from the intercept only model, i.e. $Y_i = \beta_0(Z_i) + \varepsilon_i$. There are no main and interaction effects associated with the disease phenotype. The average number of false positive effects for (d,error)= $(10, N(0,1)), (10, t(3)), (50, N(0,1)), (50, t(3))$ setups are 0.004, 0.042, 0.002 and 0.036, respectively. Overall, the proposed method achieves satisfactory false positive controls under the null model.

4 Case study

The body mass index of the mother (MBMI) is often used as a measure of the mothers' body shape and degree of obesity. Since the baby resides inside its mother's womb, its environment is defined through its mother. Increasing evidence indicates that both pre-pregnant weight (BMI) and weight gain in pregnancy have a major influence on babies' birth weight (Stamnes Koepp et al., 2012). Due to the complicated interaction between the genes of the fetus and the mother's level of obesity, the birth weight might be different for a fetus with the same genes but under different environment conditions. Thus, variation in birth weight could be totally or partially explained by the underlying genetic machinery and how those genes respond to the mother's obesity to affect birth weight.

We applied the method to a real dataset from a study conducted in the Department of Obstetrics and Gynecology at Sotero del Rio Hospital in Puente Alto, Chile. The initial objective of the study was to pinpoint genetic variants associated with a binary response indicating large for gestational age (LGA) or small for gestational age (SGA) infants based on the birthweight of new born babies. After data cleaning by removing SNPs with MAF less than 0.05 or deviation from the Hardy-Weinberg equilibrium, the dataset contains 1536 new born babies genotyped with 189 candidate genes covering 660 single nucleotide polymorphisms (SNPs).

Genes were mapped to the KEGG pathway using the GATHER software which can be accessed at http://-gather.genome.duke.edu. A total 30 pathways based on 189 candidate genes were retrieved. We treated the mother's BMI as the environmental factor and the baby's birth weight as the response variable; this was standardized before fitting to the model. Since some genes were mapped to multiple pathways, we did the variable selection for each pathway separately. Table 3 shows the selection results with SNP ID, the gene and pathway name the SNP(s) belong(s) to and the selected effect. Two SNPs in gene IL2 were mapped to two pathways and both SNPs consistently show varying effects in the two pathways. SNP rs2069762 in gene IL2 was previously reported to be associated with preterm birth and low birthweight in a Japanese population study (Sata F et al., 2009). Several other SNPs in gene IL1B were also reported to be associated with low birthweight in that paper. In addition, one SNP in gene IL1B in the Toll-like receptor signalling pathway was selected as a varying effect. Two SNPs in gene COL1A2 were mapped to two pathways and both were selected as varying effects. SNP rs997049 in gene IL1R1 was selected as a constant effect in two different pathways.

Table 3: List of selected SNPs in each pathway with constant and varying coefficients.

Pathway (# of genes)(# of SNPs)	SNP ID	Gene	Selected Effect
Cytokine-cytokine receptor interaction(45)(123)	rs2069762	IL2	varying
	rs2069772	IL2	varying
	rs997049	IL1R1	constant
Complement and coagulation cascades(18)(53)	rs2053044	ADRB2	constant
Jak-STAT signaling pathway(24)(65)	rs2069762	IL2	varying
	rs2069772	IL2	varying
ECM-receptor interaction pathway(15)(95)	rs2301643	COL1A2	varying
	rs13240759	COL1A2	varying
Toll-like receptor signaling pathway(15)(21)	rs3136558	IL1B	constant
Focal adhesion(21)(109)	rs2301643	COL1A2	varying
	rs13240759	COL1A2	varying
Apoptosis(8)(20)	rs997049	IL1R1	constant
Glycolysis/Gluconeogenesis(1)(2)	rs10891315	DLAT	constant
Pyruvate metabolism(1)(2)	rs10891315	DLAT	constant

Figure 3 plots the varying coefficient function for two SNPs, SNP rs2039762 in the Cytokine-cytokine receptor interaction pathway and SNP rs2301643 in the ECM-receptor interaction pathway. The varying pattern of the function over mother's BMI indicates the nonlinear interaction of the SNPs with mother's BMI condition to affect birth weight. When fitting a linear interaction model, no SNPs show significant interaction with mother's BMI (data not shown).

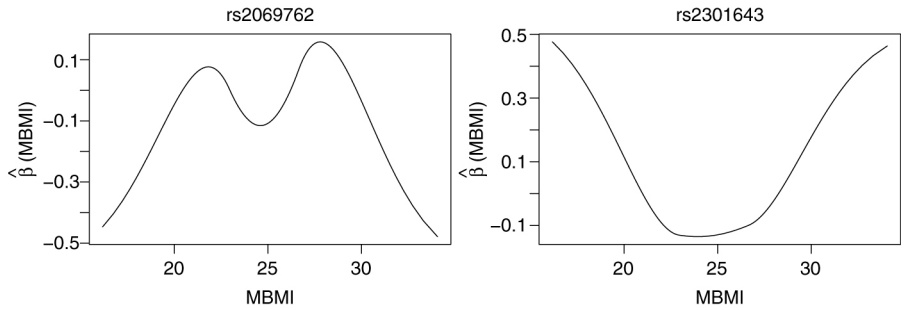

Figure 3: The estimated varying coefficient function for SNP rs2039762 in the Cytokine-cytokine receptor interaction pathway and SNP rs2301643 in the ECM-receptor interaction pathway.

5 Discussion

The significance of G×E interactions in complex human disease traits has stimulated widespread discussion. As reviewed in Cornelis et al. (2011), a number of statistical models have been proposed to assess gene effect under different environmental exposures. The success of gene set based association analysis, as shown in Wang, Li, and Hakonarson (2011), Cui et al. (2008), Wu and Cui (2013), and Schaid et al. (2012), motivated us to propose a high dimensional variable selection approach to understand the mechanism of G×E interactions associated with complex diseases. We adopted a penalized regression method within the VC model framework to investigate how multiple variants within a genetic system are moderated by environmental factors to influence the phenotypic response.

Within the model-based regression framework, most G×E interactions are modeled via a product term between a G and an E variable (Hutter et al., 2013), so the contribution of a genetic variant to the phenotypic variation is considered as a linear function in the environmental factor. Any non-linear interaction can be pursued to relax the linearity assumption (Ma et al., 2011; Wu & Cui, 2013). As pointed out by one reviewer, statistical interactions introduced by R.A. Fisher, are defined as deviation from a generalized linear model, which implicitly suggests a nonlinear relationship and is more general. To avoid confusion with the nonlinear G×E interaction presented in this work, we make it clear that our nonlinear G×E interaction refers to the effect of a genetic variant assessed as a nonlinear function of an environment variable.

In a G×E study, people are typically interested in assessing variants which are sensitive to environment changes and those that are not. We can determine if a particular genetic variant is sensitive to environmental stimuli by examining the status of the coefficient function. Varying-coefficients and constants can be separated through B-spline basis expansions under a penalized framework. The varying coefficients correspond to G×E effects and the constant effects correspond to no interaction effects. Through another penalty function, we can further shrink the constant effect into zero if the corresponding SNP has no genetic effect. We developed a two-stage iterative estimation procedure with double SCAD penalty functions. Although the two-stage strategy has been adopted for regularized quantile regression with adaptive LASSO in Tang et al. (2012), our work significantly differs in that we focused on the regularized least square regression and rigorously establish the asymptotic properties of the nonconvex double SCAD estimator under suitable regularity conditions. The potential of non-convex penalty functions in investigating G×E interactions is far from fully understood or explored. As a representative non-convex penalty function, the SCAD is adopted mainly due to its nice oracle properties as stepping stones for building statistically sound and practically useful models to accommodate more complex data structures. It is worth to mention that our method is fundamentally different from the work of Xue and Qu (2012) and Antoniadis, Gijbels, and Lambert-Lacroix (2014) in which the authors developed a variable selection framework under the additive VC model to distinguish zero vs varying coefficients. They did not distinguish non-zero constant vs varying coefficients which is one of the key objects in understanding the mechanisms of G×E interaction. Identification of the constant coefficients in the varying coefficient models is closely related to the estimation of linear part in additive models. This line of work includes Hu and Xia (2012) and Zhang, Cheng, and Liu (2011). None of the existing studies closely explore the automatic structure identification and separation of different effects under the G×E framework.

The current work only demonstrates the case with one environmental factor. It is broadly recognized that the etiology of many complex disease is less likely to be affected by one environmental factor but is more likely to be heterogeneous. When multiple continuously measured environmental factors (say K_1) are measured (denoted as Z_1), we can extend the current model to a more general case formulated as follows,

$$Y = \sum_{j=0}^{d} \left\{ \sum_{k=1}^{K_1} \beta_{kj}(Z_{1k}) \right\} X_j + \varepsilon,$$

where $X_0 = 1$. The same estimation and variable selection framework can be applied to select important genetic players that show sensitivity to different environmental stimuli. When discrete environmental variables such as smoking status are also available, denote $\boldsymbol{Z}_2$ as a collection of K_2 such variables, then we can fit the following model

$$Y = \sum_{j=0}^{d} \left\{ \sum_{k=1}^{K_1} \beta_{kj}(Z_{1k}) + \sum_{l=1}^{K_2} \alpha_{lj} Z_{2l} \right\} X_j + \varepsilon,$$

the partial linear varying-coefficient model. In addition to the two penalty functions specified in this work, an additional penalty function should be imposed for $\{\alpha\}_{lj}$ to select important variants showing interaction with $\boldsymbol{Z}_2$. In case of a binary response, we are interested in modeling $E[Y|\boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{X}] = \sum_{j=0}^{d} \left\{ \sum_{k=1}^{K_1} \beta_{kj}(Z_{1k}) + \sum_{l=1}^{K_2} \alpha_{lj} Z_{2l} \right\} X_j$. We will investigate this in future studies.

The proposed method is not only restricted to quantitative phenotypes and can be extended to other types of phenotypes. For example, in cancer prognostic studies, it can be modified as the Accelerated failure time (AFT) model to accommodate the survival outcomes. Binary phenotypes significantly differ from quantitative and survival outcomes in that they contain much less information, hence the accuracy of estimating nonlinear interactions might be sacrificed. Nevertheless, extension to the binary case can be done by developing a coordinate descent (CD) based iteratively reweighted least squares (IRLS) algorithm under the regularized logistic regression framework. The CD based IRLS algorithm have been extensively used to extend regularized variable selection methods from continuous phenotypes to binary phenotypes such as in case control studies.

In the model, we did not include any covariates. However, the proposed varying coefficient model can be readily modified to allow for covariate effects. Typically, the covariates included in the model are predetermined as important ones and are in low dimension, so their effects are not subject to penalization. Assuming there are no interactions between genes and those covariates (those with interactions will be included in the model), one can fit a regression model by regressing Y against those covariates only, assuming either linear on nonlinear effects, Then focusing the obtained residuals (after removing the covariates effects) to do the rest of the analysis by fitting the models described in this paper. It is also worth mentioning that the real data analysis in this work does not take other covariates (e.g. gender and mother's gestational age) effects into account, which may lead to biased results. Due to this limitation, readers should be cautious when interpreting the real data analysis results.

Acknowledgements

The authors wish to thank two anonymous referees for their constructive comments that greatly improved the manuscript. We would also like to thank Dr. L. Wang for insightful discussions on parameter estimation and Dr. R. Romero for sharing the birthweight data. This work was supported in part by grants from National Natural Science Foundation of China (31371336) and National Science Foundation (IOS-1237969), and by an Innovative Research Award from the Johnson Cancer Research Center at Kansas State University.

A Technical Proofs

Useful notations and lemmas

For convenience, the following notations are adopted :

$\boldsymbol{\gamma}_{(v)} = (\boldsymbol{\gamma}_0^T, \ldots, \boldsymbol{\gamma}_v^T)^T, \boldsymbol{\gamma}_{(c)} = (\boldsymbol{\gamma}_{v+1}^T, \ldots, \boldsymbol{\gamma}_c^T)^T,$
$\boldsymbol{\gamma}_{(d)} = (\gamma_{v+1,1}^T, \ldots, \gamma_{d,1}^T)^T, \tilde{\boldsymbol{\gamma}}_{(v)} = (\tilde{\boldsymbol{\gamma}}_0^T, \ldots, \tilde{\boldsymbol{\gamma}}_v^T)^T,$
$\tilde{\boldsymbol{\gamma}}_{(c)} = (\tilde{\boldsymbol{\gamma}}_{v+1}^T, \ldots, \tilde{\boldsymbol{\gamma}}_c^T)^T, \tilde{\boldsymbol{\gamma}}_{(d)} = (\gamma_{v+1,1}, \ldots, \gamma_{d,1})^T,$
$\boldsymbol{G}_n = (B(z_1), \ldots, B(z_n))(B(z_1), \ldots, B(z_n))^T,$
$\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T, \Phi_n = n^{-1} \sum_{i=1}^{n} \boldsymbol{U}_{(v)i} \boldsymbol{U}_{(v)i}^T,$
$\Psi_n = n^{-1} \sum_{i=1}^{n} \boldsymbol{U}_{(v)i} \boldsymbol{U}_{(c)i}^T, \Lambda_i = \boldsymbol{U}_{(c)i} - \Psi_n^T \Phi_n^{-1} \boldsymbol{U}_{(v)i},$

where $\boldsymbol{U}_{(v)}$ and $\boldsymbol{U}_{(c)}$ are the sub design matrices corresponding to the predictors with varying and nonzero constant coefficients respectively. We use $\| \bullet \|$ to denote the L_2 norm $\| \bullet \|_2$ in the A.

We first provide several lemmas necessary for the proofs of Theorems 1 and 2. Lemma 1 follows directly from the proof of Lemma A.3 in Huang, Wu, and Zhou (2004), and Lemma 2 follows from Corollary 6.21 of Schumaker (1981).

Lemma 1

Under assumptions (A1–A3), there exists finite positive constants C_1 and C_2 such that all the eigenvalues of $(k_n/n)\boldsymbol{G}_n$ fall between C_1 and C_2, and therefore, $\boldsymbol{G}_n$ is invertible.

Lemma 2

Under assumptions (A1–A3), for some finite constant C_3, there exists $\tilde{\boldsymbol{\gamma}} = (\tilde{\boldsymbol{\gamma}}_0^T, \ldots, \tilde{\boldsymbol{\gamma}}_d^T)^T$ satisfying

1. $\|\tilde{\gamma}_{j*}\| > C_3, j = 0, \ldots, v; \tilde{\gamma}_{j1} = \beta_j, \|\tilde{\boldsymbol{\gamma}}_{j*}\| = 0, j = v+1, \ldots, c; \tilde{\gamma}_j = \boldsymbol{0}, j = c+1, \ldots, d;$
2. $\sup_{z\in[0,1]} |\beta_j(z) - B(z)^T \tilde{\boldsymbol{\gamma}}_j| = O(k_n^{-r}), j = 0, \ldots, d,$ where $\tilde{\boldsymbol{\gamma}}_j = (\tilde{\gamma}_{j,1}, \tilde{\boldsymbol{\gamma}}_{j*}^T)^T;$
3. $\sup_{(z,\boldsymbol{x})\in[0,1]\times R^{d+1}} |\boldsymbol{x}^T\beta(z) - \boldsymbol{U}(\boldsymbol{x}, z)^T \tilde{\boldsymbol{\gamma}}| = O(k_n^{-r}).$

Proofs of Theorem 1

(I) Proof of Theorem 1(1), part 1

Here we first show $\hat{\beta}_j(z)$ is constant for $j = v+1, \ldots, d$ with probability approaching 1 as $n \to \infty$, which amounts to demonstrating $\|\hat{\boldsymbol{\gamma}}_{j*}^{vc}\| = 0, j = v+1, \ldots, d$ with probability tending to 1, as $n \to \infty$. To this end, we first show that a minimizer $\hat{\boldsymbol{\gamma}}^{\mathscr{VC}}$ of $Q_1(\boldsymbol{\gamma})$ exists in a neighborhood of $\tilde{\boldsymbol{\gamma}}$ where

$$Q_1(\boldsymbol{\gamma}) = \sum_{i=1}^{n} \left(Y_i - \boldsymbol{U}_i^T\boldsymbol{\gamma}\right)^2 + n\sum_{j=1}^{d} p_{\lambda_1}(\|\gamma_{j*}\|). \tag{11}$$

Let $\alpha_n = n^{-\frac{1}{2}}k_n + a_n$, where $a_n := \max_j\{|p'_{\lambda_1}(\|\tilde{\boldsymbol{\gamma}}_{j*}\|)|, |p'_{\lambda_2}(|\tilde{\gamma}_{j,1}|)| : \tilde{\boldsymbol{\gamma}}_{j*} \neq \boldsymbol{0}, \tilde{\gamma}_{j,1} \neq 0\}$. The property of SCAD penalty function implies that if $\max\{\lambda_1, \lambda_2\} \to 0$, $a_n = 0$. We show that for any given $\varepsilon > 0$, there exists a large constant C such that

$$P\left\{\inf_{\|\boldsymbol{\delta}\|=C} Q_1(\hat{\boldsymbol{\gamma}}^{\mathscr{VC}}) \geq Q_1(\tilde{\boldsymbol{\gamma}})\right\} \geq 1 - \varepsilon, \tag{12}$$

where $\hat{\boldsymbol{\gamma}}^{vc} = \tilde{\boldsymbol{\gamma}} + \alpha_n\boldsymbol{\delta}$. This suggests that with probability at least $1-\varepsilon$ there exists a local minimum in the ball $\{\tilde{\boldsymbol{\gamma}} + \alpha_n\boldsymbol{\delta} : \|\boldsymbol{\delta}\| \leq C\}$. Hence, there exists a local minimizer such that $\|\hat{\boldsymbol{\gamma}}^{vc} - \tilde{\boldsymbol{\gamma}}\| = O_p(\alpha_n)$. A direct computation yields

$$\begin{aligned} D_n(\boldsymbol{\delta}) &= Q_1(\hat{\boldsymbol{\gamma}}^{vc}) - Q_1(\tilde{\boldsymbol{\gamma}}) \\ &= -2\alpha_n \sum_{i=1}^{n} \left[\varepsilon_i + X_i^T r(z_i)\right] \boldsymbol{U}_i^T\boldsymbol{\delta} + \alpha_n^2 \sum_{i=1}^{n} \boldsymbol{U}_i^T\boldsymbol{\delta}\boldsymbol{\delta}^T\boldsymbol{U}_i \\ &\quad + n\sum_{j=1}^{d} \left[p_{\lambda_1}(\|\hat{\gamma}_{j*}^{vc}\|) - p_{\lambda_1}(\|\tilde{\gamma}_{j*}\|)\right] \\ &:= \Delta_1 + \Delta_2 + \Delta_3 \end{aligned}$$

where $r_j(z) = B(z)^T\tilde{\gamma}_j - \beta_j(z), j = 1, \ldots, d$ and $r(z) = (r_1(z), \ldots, r_d(z))^T$. By the fact $E(\varepsilon_i|\boldsymbol{U}_i, z_i) = 0$, we obtain that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i \boldsymbol{U}_i^T\boldsymbol{\delta} = O_p(\|\boldsymbol{\delta}\|).$$

Recall Lemma 2, then

$$\frac{1}{n}\sum_{i=1}^{n} X_i^T r(z_i)\boldsymbol{U}_i^T\boldsymbol{\delta} = O_p(k_n^{-r}\|\boldsymbol{\delta}\|).$$

Therefore

$$\Delta_1 = O_p(\sqrt{n}\alpha_n\|\boldsymbol{\delta}\|) + O_p(nk_n^{-r}\alpha_n\|\boldsymbol{\delta}\|) = O_p(nk_n^{-r}\alpha_n)\|\boldsymbol{\delta}\|.$$

We can also show that $\Delta_2 = O_p(n\alpha_n^2)\|\boldsymbol{\delta}\|^2$. Then, by choosing a sufficiently large C, Δ_1 is dominated by Δ_2 uniformly in $\|\boldsymbol{\delta}\| = C$. It follows from Taylor expansion that

$$\begin{aligned}\Delta_3 &\le n\sum_{j=1}^{d}\Big[\alpha_n p_{\lambda 1}^{'}(\|\tilde{\gamma}_{j*}\|)\frac{\tilde{\gamma}_{j*}}{\|\tilde{\gamma}_{j*}\|}\|\boldsymbol{\delta}_{j*}\| \\ &\quad + \alpha_n^2 p_{\lambda 1}^{''}(\|\tilde{\gamma}_{j*}\|)\|\boldsymbol{\delta}_{j*}\|^2(1+o_p(1))\Big] \\ &\le n\sqrt{d}\alpha_n a_n\|\boldsymbol{\delta}\| + nb_n\alpha_n^2\|\boldsymbol{\delta}\|^2.\end{aligned}$$

With assumption (A6), we can prove that Δ_2 dominates Δ_3 uniformly in $\|\boldsymbol{\delta}\| = C$. Therefore, (12) holds for sufficiently large C, and we have $\|\hat{\boldsymbol{\gamma}}^{vc} - \tilde{\boldsymbol{\gamma}}\| = O_p(\alpha_n)$.

In order to prove $\hat{\beta}_j(z)$ is constant for $j = v+1,\dots,d$ in probability, it is sufficient to demonstrate that $\hat{\boldsymbol{\gamma}}_{j*}^{vc} = \mathbf{0}$, $j = v+1,\dots,d$. Note that when $\max\{\lambda_1,\lambda_2\} \to 0$, $a_n = 0$ for large n. Then we need to show that with probability approaching 1 as $n \to \infty$, for any $\hat{\boldsymbol{\gamma}}^{vc}$ satisfying $\|\hat{\boldsymbol{\gamma}}^{vc} - \tilde{\boldsymbol{\gamma}}\| = O_p(n^{-\frac{1}{2}}k_n)$ and some small $\varepsilon_n = Cn^{-\frac{1}{2}}k_n$, we have

$$\begin{aligned}\frac{\partial Q_1(\boldsymbol{\gamma})}{\partial\gamma_{j,*}} &< 0, \quad \text{for} \quad -\varepsilon_n < \gamma_{j,*} < 0, \quad j = v+1,\dots,d; \\ &> 0, \quad \text{for} \quad 0 < \gamma_{j,*} < \varepsilon_n, \qquad j = v+1,\dots,d.\end{aligned}$$

where $\gamma_{j,*}$ denotes the individual component of $\boldsymbol{\gamma}_{j*}$. It can be shown that,

$$\begin{aligned}\frac{\partial Q_1(\hat{\boldsymbol{\gamma}}^{vc})}{\partial\hat{\gamma}_{j,*}^{vc}} &= -2\sum_{i=1}^{n}\boldsymbol{U}_{ij}\left[Y_i - \boldsymbol{U}_i^T\hat{\boldsymbol{\gamma}}^{vc}\right] + np_{\lambda_1}^{'}(|\hat{\gamma}_{j,*}|)\mathrm{sgn}(\hat{\gamma}_{j,*}) \\ &= -2\sum_{i=1}^{n}\boldsymbol{U}_{ij}[\varepsilon_i + \boldsymbol{X}_i^T r(z_i)] - 2\sum_{i=1}^{n}\boldsymbol{U}_{ij}\boldsymbol{U}_i^T[\tilde{\boldsymbol{\gamma}} - \hat{\boldsymbol{\gamma}}^{vc}] \\ &\quad + np_{\lambda_1}^{'}(|\hat{\gamma}_{j,*}|)\mathrm{sgn}(\hat{\gamma}_{j,*}^{vc}) \\ &= n\lambda_1\left[O_p(\lambda_1^{-1}n^{\frac{-r+1/2}{2r+1}}) + \lambda_1^{-1}p_{\lambda}^{'}(|\hat{\gamma}_{j,*}|)\mathrm{sgn}(\hat{\gamma}_{j,*}^{vc})\right].\end{aligned}$$

By assumption (A5), $\lambda_1^{-1}n^{\frac{-r+1/2}{2r+1}} \to 0$. Then it follows from assumption (A7) that the sign of the derivative is completely determined by that of $\hat{\gamma}_{j,*}^{vc}$. Therefore, $\hat{\boldsymbol{\gamma}}^{vc}$, the minimizer of Q_1, is achieved at $\hat{\boldsymbol{\gamma}}_{j*}^{vc} = \mathbf{0}$, $j = v+1,\dots,d$. This completes the proof of Theorem 1(1), part 1. □

(II) Proof of Theorem 1(2)

Next we establish the consistency of the varying coefficient estimators. Let $\alpha_n = n^{-\frac{1}{2}}k_n + a_n$, $\hat{\boldsymbol{\gamma}}_{(v)} = \tilde{\boldsymbol{\gamma}}_{(v)} + \alpha_n\boldsymbol{\delta}_v$, $\hat{\boldsymbol{\gamma}}_{(d)} = \tilde{\boldsymbol{\gamma}}_{(d)} + \alpha_n\boldsymbol{\delta}_d$, $\boldsymbol{\delta} = (\boldsymbol{\delta}_v^T, \boldsymbol{\delta}_d^T)^T$, and

$$\begin{aligned}Q_2(\boldsymbol{\gamma}_{(v)}, \boldsymbol{\gamma}_{(d)}) &= \sum_{i=1}^{n}\left(Y_i - \boldsymbol{U}_{(v)i}^T\boldsymbol{\gamma}_{(v)} - \boldsymbol{U}_{(d)i}^T\boldsymbol{\gamma}_{(d)}\right)^2 \\ &\quad + n\sum_{j=v+1}^{d} p_{\lambda_2}(|\gamma_{j,1}|).\end{aligned} \tag{13}$$

We first show that there exists a local minimizer of $Q_2(\boldsymbol{\gamma}_{(v)}, \boldsymbol{\gamma}_{(d)})$. It suffices to show that for any given $\varepsilon > 0$, there exists a large constant C such that

$$P\left\{\inf_{\|\boldsymbol{\delta}\|=C} Q_2(\hat{\boldsymbol{\gamma}}_{(v)}, \hat{\boldsymbol{\gamma}}_{(d)}) \geq Q_2(\tilde{\boldsymbol{\gamma}}_{(v)}, \tilde{\boldsymbol{\gamma}}_{(d)})\right\} \geq 1 - \varepsilon. \tag{14}$$

which implies that with probability at least $1 - \varepsilon$ there exists a local minimum in the ball $\{\tilde{\boldsymbol{\gamma}}_{(v)} + \alpha_n \boldsymbol{\delta}_v : \|\boldsymbol{\delta}_v\| \leq C\}$ and $\{\tilde{\boldsymbol{\gamma}}_{(d)} + \alpha_n \boldsymbol{\delta}_d : \|\boldsymbol{\delta}_d\| \leq C\}$, respectively. Therefore, there exists local minimizers such that $\|\hat{\boldsymbol{\gamma}}_{(v)} - \tilde{\boldsymbol{\gamma}}_{(v)}\| = O_p(\alpha_n)$ and $\|\hat{\boldsymbol{\gamma}}_{(d)} - \tilde{\boldsymbol{\gamma}}_{(d)}\| = O_p(\alpha_n)$. We have

$$\begin{aligned} D_n(\boldsymbol{\delta}_v, \boldsymbol{\delta}_d) &= Q_2(\hat{\boldsymbol{\gamma}}_{(v)}, \hat{\boldsymbol{\gamma}}_{(d)}) - Q_2(\tilde{\boldsymbol{\gamma}}_{(v)}, \tilde{\boldsymbol{\gamma}}_{(d)}) \\ &= -2\alpha_n \sum_{i=1}^{n} \left[\varepsilon_i + X_i^T r(z_i)\right] \left[\boldsymbol{U}_{(v)i}^T \boldsymbol{\delta}_{(v)} + \boldsymbol{U}_{(d)i}^T \boldsymbol{\delta}_{(d)}\right] \\ &+ \alpha_n^2 \sum_{i=1}^{n} \left[\boldsymbol{U}_{(v)i}^T \boldsymbol{\delta}_{(v)} + \boldsymbol{U}_{(d)i}^T \boldsymbol{\delta}_{(d)}\right]^2 \\ &+ n \sum_{j=v+1}^{d} \left[p_{\lambda_2}(|\hat{\gamma}_{j,1}|) - p_{\lambda_2}(|\tilde{\gamma}_{j,1}|)\right] \\ &:= \Delta_1 + \Delta_2 + \Delta_3, \end{aligned}$$

where $r(z) = (r_1(z), \ldots, r_d(z))^T$ and $r_j(z) = B(z)^T \tilde{\gamma}_j - \beta_j(z)$, $j = 1, \ldots, d$. Since $E(\varepsilon_i | \boldsymbol{U}_{(v)}, \boldsymbol{U}_{(d)}, z_i) = 0$, we have

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i [\boldsymbol{U}_{(v)i}^T \boldsymbol{\delta}_{(v)} + \boldsymbol{U}_{(d)i}^T \boldsymbol{\delta}_{(d)}] = O_p(\|\boldsymbol{\delta}\|). \tag{15}$$

With Lemma 2 we can show

$$\frac{1}{n} \sum_{i=1}^{n} X_i^T r(z_i) \left[\boldsymbol{U}_{(v)i}^T \boldsymbol{\delta}_{(v)} + \boldsymbol{U}_{(d)i}^T \boldsymbol{\delta}_{(d)}\right] = O_p\left(k_n^{-r} \|\boldsymbol{\delta}\|\right).$$

Combine the above two equations, we can obtain that

$$\Delta_1 = O_p(n^{\frac{1}{2}} \alpha_n \|\boldsymbol{\delta}\|) + O_p(n k_n^{-r} \alpha_n \|\boldsymbol{\delta}\|) = O_p(n k_n^{-r} \alpha_n) \|\boldsymbol{\delta}\|.$$

Since $\Delta_2 = O_p(n\alpha_n^2)\|\boldsymbol{\delta}\|^2$, it can be shown that by choosing a sufficiently large C, Δ_1 is dominated by Δ_2 uniformly in $\|\boldsymbol{\delta}\| = C$. By Taylor expansion,

$$\begin{aligned} \Delta_3 &\leq n \sum_{j=v+1}^{d} \left[\alpha_n p'_{\lambda 2}(|\tilde{\gamma}_{j,1}|) \mathrm{sgn}(\tilde{\gamma}_{j,1}) |\delta_{j1}| \right. \\ &\quad \left. + \alpha_n^2 p''_{\lambda 2}(|\tilde{\gamma}_{j,1}|) \delta_{j1}^2 (1 + o(1))\right] \\ &\leq (d - v)^{\frac{1}{2}} n \alpha_n a_n \|\boldsymbol{\delta}\| + n b_n \alpha_n^2 \|\boldsymbol{\delta}\|^2. \end{aligned}$$

Recall assumption A6, then it follows that, by choosing an enough large C, Δ_2 dominates Δ_1 uniformly in $\|\boldsymbol{\delta}\| = C$. Consequently (14) holds for sufficiently large C, and we have $\|\hat{\boldsymbol{\gamma}}_v - \tilde{\boldsymbol{\gamma}}_v\| = O_p(\alpha_n)$ and $\|\hat{\boldsymbol{\gamma}}_d - \tilde{\boldsymbol{\gamma}}_d\| = O_p(\alpha_n)$. By the definition of $\boldsymbol{\gamma}^{cz}$, we have $\hat{\boldsymbol{\gamma}}_{(d)}^{cz} - \tilde{\boldsymbol{\gamma}}_{(d)} = O_p(\alpha_n)$. Then for $j = 0, \ldots, v$

$$\begin{aligned} \|\hat{\beta}_j(z_i) - \beta_j(z)\|^2 &= \int_0^1 \left[\hat{\beta}_j(z) - \beta_j(z)\right]^2 dz \\ &\leq 2 \int_0^1 \left[\boldsymbol{B}(z)^T \hat{\boldsymbol{\gamma}}_j^{cz}(z) - \boldsymbol{B}(z)^T \tilde{\boldsymbol{\gamma}}_j\right]^2 dz \\ &\quad + 2 \int_0^1 r_j^2(z) dz \\ &= \frac{2}{n} (\hat{\boldsymbol{\gamma}}_j^{cz} - \tilde{\boldsymbol{\gamma}}_j)^T \boldsymbol{G}_n (\hat{\boldsymbol{\gamma}}_j^{cz} - \tilde{\boldsymbol{\gamma}}_j) \\ &\quad + 2 \int_0^1 r_j^2(z) dz \\ &:= \Delta_1 + \Delta_2. \end{aligned}$$

Recall Lemma 1, Lemma 2 and $k_n = O\left(n^{\frac{1}{2r+1}}\right)$, we can demonstrate that $\Delta_1 = O_p\left(k_n^{-1}\alpha_n^2\right)$, $\Delta_2 = O_p\left(k_n^{-2r}\right)$. Δ_1 is dominated by Δ_2, thus we finish the proof of Theorem 1(2). □

(III) Proof of Theorem 1(1), part 2

To show $\hat{\beta}_j(z) = 0$ for $j = c+1, \ldots, d$, it is sufficient to demonstrate that $\hat{\boldsymbol{\gamma}}_{j,1}^{cz} = 0$, since the constancy of $\beta_j(z)$, $j = v+1, \ldots, d$ was already established in (B). By definition, when $\max\{\lambda_1, \lambda_2\} \to 0$, $a_n = 0$ for large n. Then we need to prove that with probability approaching 1 as $n \to \infty$, for any $\hat{\boldsymbol{\gamma}}_{(v)}$ and $\hat{\boldsymbol{\gamma}}_{(d)}$ satisfying $\|\hat{\boldsymbol{\gamma}}_{(v)} - \tilde{\boldsymbol{\gamma}}_{(v)}\| = O_p(n^{-\frac{1}{2}}k_n)$, and $\|\hat{\boldsymbol{\gamma}}_{(d)} - \tilde{\boldsymbol{\gamma}}_{(d)}\| = O_p(n^{-\frac{1}{2}}k_n)$, as well as some small $\varepsilon_n = Cn^{-\frac{1}{2}}k_n$, we have

$$\begin{aligned}\frac{\partial Q_2(\boldsymbol{\gamma}_{(v)}, \boldsymbol{\gamma}_{(d)})}{\partial \gamma_{j,1}} &< 0, \quad \text{for} \quad -\varepsilon_n < \gamma_{j,1} < 0, \quad j = c+1, \ldots, d;\\ &> 0, \quad \text{for} \quad 0 < \gamma_{j,1} < \varepsilon_n, \quad j = c+1, \ldots, d.\end{aligned}$$

It can be shown that

$$\begin{aligned}\frac{\partial Q_2(\hat{\boldsymbol{\gamma}}_{(v)}, \hat{\boldsymbol{\gamma}}_{(d)})}{\partial \hat{\gamma}_{j,1}} =& -2\sum_{i=1}^{n} \boldsymbol{U}_{(d)ij}\left[Y_i - \boldsymbol{U}_{(v)i}^T\hat{\boldsymbol{\gamma}}_{(v)} - \boldsymbol{U}_{(d)i}^T\hat{\boldsymbol{\gamma}}_{(d)}\right]\\ &+ np_\lambda^{'}(|\hat{\gamma}_{j,1}|)\mathrm{sgn}(\hat{\gamma}_{j,1})\\ =& -2\sum_{i=1}^{n} \boldsymbol{U}_{(d)ij}\left[\varepsilon_i + \mathbf{X}_i^T r(z_i)\right]\\ &-2\sum_{i=1}^{n} \boldsymbol{U}_{(d)ij}\boldsymbol{U}_{(v)i}^T\left[\tilde{\boldsymbol{\gamma}}_v - \hat{\boldsymbol{\gamma}}_v\right]\\ &-2\sum_{i=1}^{n} \boldsymbol{U}_{(d)ij}\boldsymbol{U}_{(d)i}^T\left[\tilde{\boldsymbol{\gamma}}_d - \hat{\boldsymbol{\gamma}}_d\right] + np_\lambda^{'}(|\hat{\gamma}_{j,1}|)\mathrm{sgn}(\hat{\gamma}_{j,1})\\ =& n\lambda_2\left[O_p\left(\lambda_2^{-1}n^{\frac{-r+1/2}{2r+1}}\right) + \lambda_2^{-1}p_\lambda^{'}(|\hat{\gamma}_{j,1}|)\mathrm{sgn}(\hat{\gamma}_{j,1})\right].\end{aligned}$$

By assumption (A5), $\lambda_2^{-1}n^{\frac{-r+1/2}{2r+1}} \to 0$. Then it follows from assumption (A7) that the sign of the derivative is completely determined by that of $\hat{\gamma}_{j,1}$. Therefore, $\hat{\boldsymbol{\gamma}}^{cz}$, the minimizer of Q_2, is achieved at $\hat{\boldsymbol{\gamma}}_{j,1}^{cz} = 0, j = c+1, \ldots, d$. This completes the proof of Theorem 1(1). □

Proof of Theorem 2

In Theorem 1, we showed that both $\hat{\boldsymbol{\gamma}}_{j*} = \mathbf{0}, j = v+1, \ldots, c$ and $\hat{\gamma}_j = 0, j = c+1, \ldots, d$, hold with probability approaching 1. Then Q_2 reduces to

$$\begin{aligned}Q_2(\boldsymbol{\gamma}_{(v)}, \boldsymbol{\gamma}_{(d)}) =& \sum_{i=1}^{n}\left(Y_i - \boldsymbol{U}_{(v)i}^T\boldsymbol{\gamma}_{(v)} - \boldsymbol{U}_{(c)i}^T\boldsymbol{\gamma}_{(c)}\right)^2\\ &+ n\sum_{j=v+1}^{c} p_{\lambda_2}(|\gamma_{j,1}|)\\ :=& Q_2(\boldsymbol{\gamma}_{(v)}, \boldsymbol{\gamma}_{(c)}).\end{aligned} \tag{16}$$

Since $(\hat{\boldsymbol{\gamma}}_{(v)}, \hat{\boldsymbol{\gamma}}_{(c)})$ is the minimizer of $Q_2(\boldsymbol{\gamma}_{(v)}, \boldsymbol{\gamma}_{(c)})$, we obtain

$$\begin{aligned}\frac{\partial Q_2(\hat{\boldsymbol{\gamma}}_{(v)}, \hat{\boldsymbol{\gamma}}_{(c)})}{\partial \hat{\boldsymbol{\gamma}}_{(v)}} &= -2\sum_{i=1}^{n} \boldsymbol{U}_{(v)i}\left[Y_i - \boldsymbol{U}_{(v)i}^T\hat{\boldsymbol{\gamma}}_{(v)} - \boldsymbol{U}_{(d)i}^T\hat{\boldsymbol{\gamma}}_{(d)}\right]\\ &= 0;\end{aligned}$$

$$\frac{\partial Q_2(\hat{\boldsymbol{\gamma}}_{(v)}, \hat{\boldsymbol{\gamma}}_{(c)})}{\partial \hat{\boldsymbol{\gamma}}_{(c)}} = -2\sum_{i=1}^{n} \boldsymbol{U}_{(c)i}\left[Y_i - \boldsymbol{U}_{(v)i}^T\hat{\boldsymbol{\gamma}}_{(v)} - \boldsymbol{U}_{(c)i}^T\hat{\boldsymbol{\gamma}}_{(c)}\right] + n\sum_{j=v+1}^{c} p_{\lambda 2}^{'}(|\hat{\gamma}_{j,1}|)\mathrm{sgn}(\hat{\gamma}_{j,1}) = 0. \tag{17}$$

By applying Taylor expansion on $p_{\lambda 2}^{'}(|\hat{\gamma}_{j,1}|)$ in (17), we have

$$p_{\lambda 2}^{'}(|\hat{\gamma}_{j,1}|) = p_{\lambda 2}^{'}(|\gamma_{j,1}|) + p_{\lambda 2}^{''}(|\gamma_{j,1}|)(\hat{\gamma}_{j,1} - \gamma_{j,1})[1 + o_p(1)].$$

By the fact that $p_{\lambda 2}^{'}(|\hat{\gamma}_{j,1}|) = 0$ as $\lambda_2 \to 0$, and $p_{\lambda 2}^{''}(|\gamma_{j,1}|) = o_p(1)$ from the assumption, it follows that

$$\sum_{j=v+1}^{c} p_{\lambda 2}^{'}(|\hat{\gamma}_{j,1}|)\mathrm{sgn}(\hat{\gamma}_{j,1}) = o_p(\hat{\gamma}_{j,1} - \gamma_{j,1}) = o_p(\hat{\boldsymbol{\gamma}}_{(c)} - \boldsymbol{\gamma}_{(c)}).$$

Consequently, we have

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{U}_{(c)i}\left[Y_i - \boldsymbol{U}_{(v)i}^T\hat{\boldsymbol{\gamma}}_{(v)} - \boldsymbol{U}_{(c)i}^T\hat{\boldsymbol{\gamma}}_{(c)}\right] + o_p(\hat{\boldsymbol{\gamma}}_{(c)} - \boldsymbol{\gamma}_{(c)}) = 0.$$

Following similar lines of arguments in Theorem 1, we can show

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{U}_{(c)i}\left[\varepsilon_i + X_i^T r(z_i) + \boldsymbol{U}_{(v)i}^T(\boldsymbol{\gamma}_{(v)} - \hat{\boldsymbol{\gamma}}_{(v)}) + \boldsymbol{U}_{(c)i}^T(\boldsymbol{\gamma}_{(c)} - \hat{\boldsymbol{\gamma}}_{(c)})\right] + o_p(\hat{\boldsymbol{\gamma}}_{(c)} - \boldsymbol{\gamma}_{(c)}) = 0. \tag{18}$$

Meanwhile, a straightforward calculation yields

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{U}_{(v)i}\left[\varepsilon_i + \boldsymbol{X}_i^T r(u_i) + \boldsymbol{U}_{(v)i}^T(\boldsymbol{\gamma}_{(v)} - \hat{\boldsymbol{\gamma}}_{(v)}) + \boldsymbol{U}_{(c)i}^T(\boldsymbol{\gamma}_{(c)} - \hat{\boldsymbol{\gamma}}_{(c)})\right] = 0. \tag{19}$$

Recall the definition of Φ_n and Ψ_n, (19) is equivalent to

$$\hat{\boldsymbol{\gamma}}_{(v)} - \boldsymbol{\gamma}_{(v)} = \Phi_n^{-1}\left\{\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{U}_{(v)i}\left[\varepsilon_i + \boldsymbol{X}_i^T r(z_i)\right] + \Psi_n[\boldsymbol{\gamma}_{(c)} - \hat{\boldsymbol{\gamma}}_{(c)}]\right\}. \tag{20}$$

Plugging (20) into (18) results in

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{U}_{(c)i}\left\{\varepsilon_i + X_i^T r(z_i) - \boldsymbol{U}_{(v)i}^T\Phi_n^{-1}\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{U}_{(v)i} \times \left[\varepsilon_i + \boldsymbol{X}_i^T r(z_i)\right]\right\} \\ &= \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{U}_{(c)i}\left[\boldsymbol{U}_{(c)i} - \Psi_n^T\Phi_n^{-1}\boldsymbol{U}_{(v)i}\right]^T(\hat{\boldsymbol{\gamma}}_{(c)} - \boldsymbol{\gamma}_{(c)}) + o_p(\hat{\boldsymbol{\gamma}}_{(c)} - \boldsymbol{\gamma}_{(c)}).\end{aligned} \tag{21}$$

Together with the facts that

$$\frac{1}{n}\sum_{i=1}^{n}\Psi_n^T\Phi_n^{-1}\boldsymbol{U}_{(v)i}\Big[\varepsilon_i + X_i^T r(z_i) - \boldsymbol{U}_{(v)i}^T\Phi_n^{-1} \times \frac{1}{n}\sum_{j=1}^{n}\boldsymbol{U}_{(v)j}[\varepsilon_j + \boldsymbol{X}_j^T r(z_j)]\Big] = 0$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\Psi_n^T\Phi_n^{-1}\boldsymbol{U}_{(v)i}\left[\boldsymbol{U}_{(c)i} - \Psi_n^T\Phi_n^{-1}\boldsymbol{U}_{(v)i}\right]^T = 0.$$

and recall the definition of Λ_i, a direct computation from (21) leads to

$$\begin{aligned}&\left[\frac{1}{n}\sum_{i=1}^{n}\Lambda_i\Lambda_i^T + o_p(1)\right]\sqrt{n}(\boldsymbol{\gamma}_{(c)} - \hat{\boldsymbol{\gamma}}_{(c)})\\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Lambda_i\varepsilon_i + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Lambda_i\boldsymbol{X}_i^T r(z_i)\\ &\quad + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Lambda_i\boldsymbol{U}_{(v)i}^T\Phi_n^{-1}\frac{1}{n}\sum_{j=1}^{n}\boldsymbol{U}_{(v)j}\left[\varepsilon_j + \boldsymbol{X}_j^T r(z_j)\right]\\ &:= \Delta_1 + \Delta_2 + \Delta_3.\end{aligned}$$

It follows from the law of large numbers that

$$\frac{1}{n}\sum_{i=1}^{n}\Lambda_i\Lambda_i^T \xrightarrow{p} \Sigma$$

where

$$\Sigma = E\left[\boldsymbol{U}_{(c)}\left(I - \boldsymbol{U}_{(v)}(\boldsymbol{U}_{(v)}\boldsymbol{U}_{(v)}^T)^{-1}\boldsymbol{U}_{(v)}^T\right)\boldsymbol{U}_{(c)}^T\right] \tag{22}$$

Consequently,

$$\Delta_1 \xrightarrow{d} \mathcal{N}(0, \sigma^2\Sigma)$$

follows from central limit theorem. Because $\boldsymbol{X}_i$ is bounded and $\|r(z)\| = o_p(1)$, we have $\Delta_2 = o_p(1)$. Besides, $\sum_{i=1}^{n}\Lambda_i\boldsymbol{U}_{(v)i}^T$=0 implies that $\Delta_3 = 0$. Therefore, by Slutsky theorem, we complete the proof of Theorem 2. □

References

Antoniadis, A., I. Gijbels and S. Lambert-Lacroix (2014): "Penalized estimation in additive varying coefficient models using grouped regularization," Stat. Pap., 55, 727–750.

Chatterjee, N. and R. J. Carroll (2005): "Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies," Biometrika, 92, 399–418.

Chen, Y.-H., N. Chatterjee and R. J. Carroll (2013): "Using shared genetic controls in studies of gene-environment interactions," Biometrika, 100, 319–338.

Cornelis, M. C., E. J. Tchetgen, L. Liang, L. Qi, N. Chatterjee, F. B. Hu and P. Kraft (2011): "Gene-environment interactions in genome-wide association studies: a comparative study of tests applied to empirical studies of type 2 diabetes," Am. J. Epidemiol., 175, 191–202.

Cui, Y. H., G. L. Kang, K.L. Sun, M. Qian, R. Romero and W. Fu (2008): "Gene-centric genomewide association study via entropy," Genetics, 179, 637–650.

Efron, B. and R. Tibshirani (2007): "On testing the significance of sets of genes," Ann. Appl. Stat., 1, 107–129.

Feinberg, A. P. (2004): "Phenotypic plasticity and the epigenetics of human disease," Nature, 447, 433–440.

Fan, J. Q. and R. Z. Li (2001): "Variable selection via nonconcave penzlied likelihood and its oracle properties," J. Am. Stat. Assoc., 96, 1348–1360.

Guo, S. W. (2000): "Gene-environment interaction and the mapping of complex traits: some statistical models and their implications," Hum. Hered., 50, 286–303.
Hastie, T. and R. Tibshirani (1993): "Varying-coefficient models," J. R. Stat. Soc. B, 55, 757–796.
Hu, T. and Y. Xia (2012): "Adaptive semi-varying coefficient model selection," Stat. Sin., 22, 575–599.
Huang, J. Z., Wu, C. O., and Zhou, L. (2002): "Varying-coefficient models and basis function approximations for the analysis of repeated measurements." Biometrika, 89, 111–128.
Huang, J. H., Wu, C. O., and Zhou L. (2004): "Polynomial spline estimation and inference for varying coefficient models with longitudinal data," Stat. Sin., 14, 763–788.
Hutter, C. M., L. E. Mechanic, N. Chatterjee, P. Kraft and E. M. Gillanders. (2013): "Gene-environment interactions in cancer epidemiology: a national cancer institute think tank report," Genet. Epidemiol., 37, 643–657.
Kim, M. O. (2007): "Quantile regression with varying coefficients," Ann. Stat., 35, 92–108.
Liu, L., Y. Li and T. O. Tollefsbol (2008): "Gene-environment interactions and epigenetic basis of human diseases," Curr. Issues Mol. Biol., 10, 25–36.
Ma, S., L. Yang, R. Romero and Y. Cui (2011): "Varying coefficient model for gene-environment interaction: a non-linear look," Bioinformatics, 27, 2119–2126.
Maity, A., R. J. Carrol, E. Mammen and N. Chatterjee (2009): "Testing in semiparametric models with interaction, with applications to gene-environment interactions," J. R. Stat. Soc. B, 71, 75–96.
Rawlings, J. S., K. M. Rosler and D. A. Harrison (2004): "The JAK/STAT signaling pathway," J. Cell Sci., 117, 1281–1283.
Sata F, S. Toya, H. Yamada, K. Suzuki, Y. Saijo, A. Yamazaki, H. Minakami and R. Kishi (2009): "Proinflammatory cytokine polymorphisms and the risk of preterm birth and low birthweight in a Japanese population," Mol. Hum. Reprod., 15, 121–130.
Schwarz, G. (1978): "Estimating the dimension of a model," Ann. Stat., 6, 461–464.
Schaid, D. J., J. P. Sinnwell, G. D. Jenkins, S. K. McDonnell, J. N. Ingle, M. Kubo, P. E. Goss, J. P. Costantino, D. L. Wickerham, and R. M. Weinshilboum (2012): "Using the gene ontology to scan multilevel gene sets for associations in genome wide association studies," Genet. Epidemiol., 36, 3–16.
Schumaker, L. L. (1981): Spline Functions: basic theory, Wiley, New York.
Stamnes Koepp, U. M., L. F. Andersen, K. Dahl-Joergensen, H. Stigum, O. Nass and W. Nystad (2012): "Maternal pre-pregnant body mass index, maternal weight change and offspring birthweight," Acta Obstet. Gynecol. Scand., 91, 243–249.
Tang, Y. L., H. X. Wang, Z. Y. Zhu, X. Song (2012): "A unified variable selection approach for varying coefficient models," Stat. Sin., 22, 601–628.
Wang, K., M. Li and H. Hakonarson. (2011): "Analysing biological pathways in genome-wide association studies," Nat. Rev. Genet., 11, 843–854.
Wang, L. F., H. Z. Li and J. Z. Huang. (2008): "Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements," J. Am. Stat. Assoc., 103, 1556–1569.
Wu, C. and Y. Cui (2013): "A novel method for identifying nonlinear gene-environment interactions in case-control association studies," Hum. Genet., 132, 1413–1425.
Wu, C. and Y. Cui (2014): "Boosting signals in gene-based association studies via efficient SNP selection," Brief. Bioinform., 15, 279–291.
Xue, L. and A. Qu (2012): "Variable selection in high-dimensional varying coefficient models with global optimality," J. Mach. Learn. Res., 13, 1973–1998.
Zhang, H. H., G. Cheng and Y. Liu (2011): "Linear or nonlinear? Automatic structure discovery for partially linear models," J. Am. Stat. Assoc., 106, 1099–1112.