High dimensional variable selection for gene-environment interactions

Cen Wu[†], Ping-Shou Zhong[†] and Yuehua Cui[†,*]

[†]*Department of Statistics and Probability, Michigan State University, East Lansing, Michigan, 48824*

{August 28, 2013}

Abstract

Gene-environment (G×E) interaction plays a pivotal role in understanding the genetic basis of complex disease. When environment factors are measured in a continuous scale, one can assess the genetic sensitivity over different environmental conditions on a disease phenotype. Motivated by the increasing awareness of the power of gene set based association analysis over single variant based approach, 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 set are mediated by one or multiple environment factors to affect a disease phenotype. We approached the problem from a high dimensional variable selection perspective. In particular, we can 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 (SCAD) 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.

Key words: Nonlinear gene-environment interaction; SCAD penalty; Local quadratic approximation; Varying-coefficient model

1 Introduction

Human complex diseases are not only determined by genetic variants, but also affected by the environmental factors, as well as the interplay between them. Gene expression changes under different environmental conditions reveals the interaction between genes and environment. The expression changes are less likely due to the change of gene sequence itself, but rather due to the structural changes such as DNA methylation or histone modification which consequently play a regulatory rule to moderate gene expressions. Such epigenetic changes has been increasing recognized as the epigenetic basis of gene-environment (G×E) interaction (Liu et al. 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 in a continuous scale, a more clear 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. [1]; Wu and Cui. [2]). As illustrated in Wu and Cui [2], one can assess the nonlinear G×E interaction when an environment factor is measured in a continuous scale. For example, individual obese condition 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 the disease heterogeneity.

When assessing G×E interactions, investigators are predominantly focused on the single variant based analysis, such as the parametric methods in Guo [3], semi-parametric methods in Chatterjee et al [4] and Maity et al [5], and non-parametric methods in Ma et al [1] and Wu and Cui [2]. Recently, there is a large wave of genetic association studies focusing on a set of variants, namely the set-based association studies, for example, the gene-centric analysis in Cui et al [6], the gene-set analysis in Schaid et al [7] and Efron and Tibshirani [8], and the pathway-based analysis in Wang et al. [9]. By assessing the joint function of multiple variants in a set, one can obtain better interpretation of the disease signals and gain novel

insight into the disease etiology. Motivated by the set-based association studies, we propose a set-based framework to investigate how variants in a gene-set mediated by one or multiple environment factors to affect the disease responses. This framework could shed novel insight into the elucidation of the regulation mechanism of a genetic set (e.g., a pathway), triggered by environment factors.

In a typical set-based association study, the number of variants p within a genetic system could be large, which makes the regular regression fail, especially when p is close to or larger than the sample size n. The problem can be approached from the perspective of high dimensional variable selection. In this work, we extend our previous work on nonlinear gene-environment interaction study from a single variant based analysis to a multiple variant based analysis under a penalized regression framework. We include variants that belongs 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 which corresponds to cases with G×E interactions, no G×E interactions and no genetic effects, respectively. Our approach enjoys the power and merits of high-dimensional variable selection by simultaneously fitting all the variants in a genetic system into a regression model, therefore avoids 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. In Section 3, we present the theoretical results including consistency in variable selection and show the optimal convergence rates of the estimates of varying effect. We show that the estimates of non-zero constant coefficients enjoy the oracle property, that is, the asymptotic distribution of the non-zero constant coefficient function is the same as that when the true model is known in priori. The merit of the proposed approaches is demonstrated through extensive simulation studies in Section 4 and real data analysis in Section 5. The technical proofs are relegated to Appendix.

2 Statistical Method

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 $(\mathbf{X}_i, Y_i, Z_i)$, $i = 1, \ldots, n$ be independent and identically distributed (i.i.d.) random vectors, then the varying coefficient (VC) model, initially proposed by Hastie and Tibshirani [10], 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 total d genetic variants which are moderated by a common environment 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

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

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 continuous 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 $j = 0, \ldots, d$, 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 of 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 [18]. 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 and moreover, 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,

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

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.

2.2 Computational Algorithm

The SCAD penalty function is singular at the origin, and do 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 following the idea of Tang

et al. [11]. Following Fan and Li (2001) [12], in a neighbourhood 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 $j = 1, ..., d$. 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

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

The sets of predictors with varying, non-zero constant, and zero effects are termed as $\mathcal{V}$, $\mathcal{C}$ and $\mathcal{Z}$ respectively. We implement the iterative algorithm in the following two-stage procedure. At 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, $\boldsymbol{\Omega}_{\lambda_1}(\boldsymbol{\gamma}_0)$=diag$\{\boldsymbol{\Omega}_0, \boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_d\}$, where $\boldsymbol{\Omega}_0 = \mathbf{0}_L$, $\boldsymbol{\Omega}_j = \left\{0, \frac{p_{\lambda 1}^T(\|\boldsymbol{\gamma}_{j*}^0\|_2)}{\|\boldsymbol{\gamma}_{j*}^0\|_2}, \ldots, \frac{p_{\lambda 1}^T(\|\boldsymbol{\gamma}_{j*}^0\|_2)}{\|\boldsymbol{\gamma}_{j*}^0\|_2}\right\}_L$ for $j = 1, \ldots, d$. Hence the estimator can be iteratively obtained as

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

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

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

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

where $\boldsymbol{\Omega}_{\lambda_2}(\hat{\boldsymbol{\gamma}}^{\mathcal{VC}})$=diag$\{\boldsymbol{\Omega}_0, \boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_d\}$ with $\boldsymbol{\Omega}_0 = \mathbf{0}_L$, $\boldsymbol{\Omega}_j = \left\{\frac{p_{\lambda 2}^T(|\hat{\gamma}_{j,1}^{\mathcal{VC}}|)}{|\hat{\gamma}_{j,1}^{\mathcal{VC}}|} I(\|\hat{\boldsymbol{\gamma}}_{j*}^{\mathcal{VC}}\|_{L_2} = 0), 0, \ldots, 0\right\}_L$. The estimator can be iteratively obtained as

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

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

We can obtain the estimator $\hat{\boldsymbol{\gamma}}$ at convergence from the iterative procedure between the above two stages, 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 $\mathcal{V}$, $\mathcal{C}$ and $\mathcal{Z}$ 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 from 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. We adopt the Schwarz BIC criterion [13] 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 $\mathcal{V}$ and $\mathcal{C}$.

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 (8), 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 $\mathcal{H}_r$ 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., $|h^{r_1}(z_1) - h^{r_1}(z_2)| \leq C_0|z_1 - z_2|^{r_2}$ where $0 \leq z_1, z_2 \leq 1$ and C_0 is a finite positive constant. Then $\beta_j(z) \in \mathcal{H}_r$, $j = 0, 1, \ldots, v$, for some $r \geq \frac{3}{2}$.

(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 $C_1 \leq f(z) \leq C_2$ for all $z \in [0, 1]$.

(A3) Let $\tilde{\lambda}_0 \leq \ldots \leq \tilde{\lambda}_d$ 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 finite positive constant C_3 such that

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

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

(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 literature of polynomial splines and variable selections. The assumption similar to (A1) could be found in Kim [14] and Tang et al [11]. (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 et al [15, 16] and Wang et al [17]. (A4) suggests that the knot sequence is quasi-uniform on [0,1], by Schumaker [18]. (A5-A7) are conditions on tuning parameters, of which (A5) could be found in Tang et al [11]; (A6) and (A7) are similar to those in Fan and Li [12] and Wang et al [17].

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 the Appendix. 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 the estimator.

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

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

where Σ is defined as (B.12) in the Appendix.

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.

3.1 Case when predictors are continuous

In example 1, we simulated data from the following VC model,

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

where the index variable $Z_i \sim unif(0,1)$, and the predictors X_i were generated from a multivariate normal distribution with mean $\mathbf{0}$ and $Cov(X_j, X_{j'}) = 0.5^{|j-j'|}$ for $1 \leq j, j' \leq d$. The performance was evaluated under both d=10 and 50. We let the coefficients of X_j, $j = 0, 1, 2$ be of varying effects, X_j, $j = 3, 4$ be of non-zero constant effects, and the rest be zeros. The random error ε_i were generated from a standard normal distribution and t distribution with 3 degrees of freedom respectively. The coefficients were set as: $\beta_0(z) = \sin(2\pi z)$, $\beta_1(z) = 2 - 3\cos\{(6z-5)\pi/3\}$, $\beta_2(z) = 3(2z-1)^3$, $\beta_3(z) = 2$, $\beta_4(z) = 2.5$, and $\beta_j(z) = 0$ for $j > 4$.

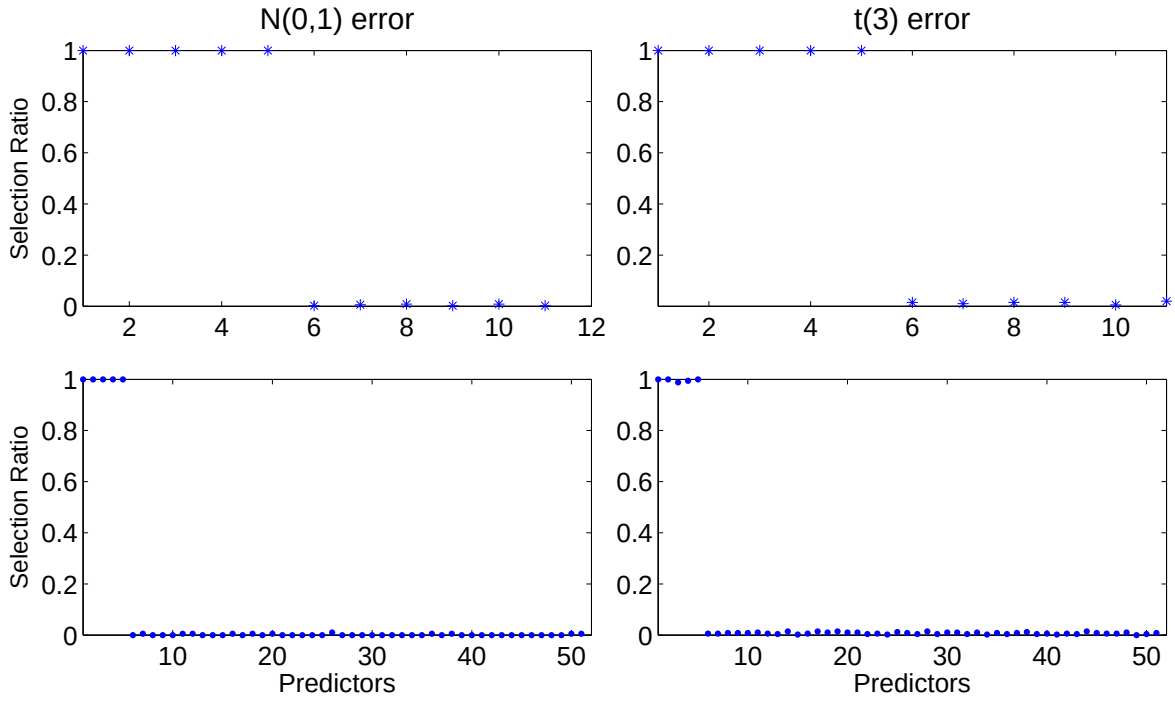

Figure 1: The selection ratio under different error distributions for different coefficient functions.

Figure 1 shows the selection ratio for predictors under different error distributions. The top panel denotes the result for $d = 10$ and the bottom panel for $d = 50$. Under different data

dimensions, the SCAD penalty performs relatively stable with consistently high selection ratio for true functions under both error distributions. For the false selection ratio, higher error rate is observed under the $t(3)$ error distribution. The oracle percentage and parameter estimation results were summarized in Table 1. Here we computed IMSEs for all predictors, including β_4 and β_5 to reflect the overall estimation precision. When β_j $(j = 4, 5)$ was selected as non-zero constant, IMSE reduces to MSE. The IMSEs was calculated if β_j $(j = 4, 5)$ was incorrectly identified as varying effect. In all the cases and under different error distributions, the SCAD approach demonstrates stable performance in comparison to the result obtained when fitting the true model (the oracle model). Under the $d = 10$ case, the penalized approach has smaller TIMSE than the oracle model does due to shrinkage. When data dimension increases to $d = 50$, the penalized method has larger TIMSE. This is expected since relatively larger noises are introduced when the data dimension is increased.

Table 1: List of IMSE, TIMSE, and Oracle Percentage under $\mathcal{N}(0, 1)$ and $t(3)$ error distributions with SCAD and ALASSO penalty functions.

	$d = 10$				$d = 50$			
	$\mathcal{N}(0,1)$ error		$t(3)$ error		$\mathcal{N}(0,1)$ error		$t(3)$ error	
	SCAD	Oracle	SCAD	Oracle	SCAD	Oracle	SCAD	Oracle
Oracle Perc.	0.972	1	0.92	1	0.945	1	0.8	1
$\beta_0(u)$	0.0214	0.0216	0.0398	0.1929	0.0221	0.0219	0.0431	0.0426
$\beta_1(u)$	0.0902	0.0951	0.1166	0.3392	0.0878	0.0927	0.1230	0.1253
$\beta_2(u)$	0.0365	0.0431	0.0764	0.5859	0.0404	0.0428	0.1042	0.0751
$\beta_3(u)$	0.0122	0.0032	0.0753	0.1775	0.0478	0.0027	0.1727	0.0105
$\beta_4(u)$	0.0045	0.0031	0.0183	0.1100	0.0101	0.0029	0.0239	0.0083
TIMSE	0.1648	0.1661	0.3282	0.4017	0.2086	0.1631	0.5146	0.2619

3.2 Cases when predictors are discrete genetic variables

Since the paper deals with G×E interaction, in example 2, we simulated genetic predictors which are discrete in nature. 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 with the largest probability can be generated. 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 $\delta = r\sqrt{(}p_A(1-p_A)p_B(1-p_B))$. Detailed information about the simulation can be found at Cui et al. (2008) [6]. The non-zero coefficient functions were assumed to be the same as those given in example 1. 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 2 shows the selection ratio when d=10, under different combinations of MAF and error distribution. The height of 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, 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 Fig. 3, which shows a very similar pattern as the $d = 10$ case. The results demonstrate the stable performance of the proposed

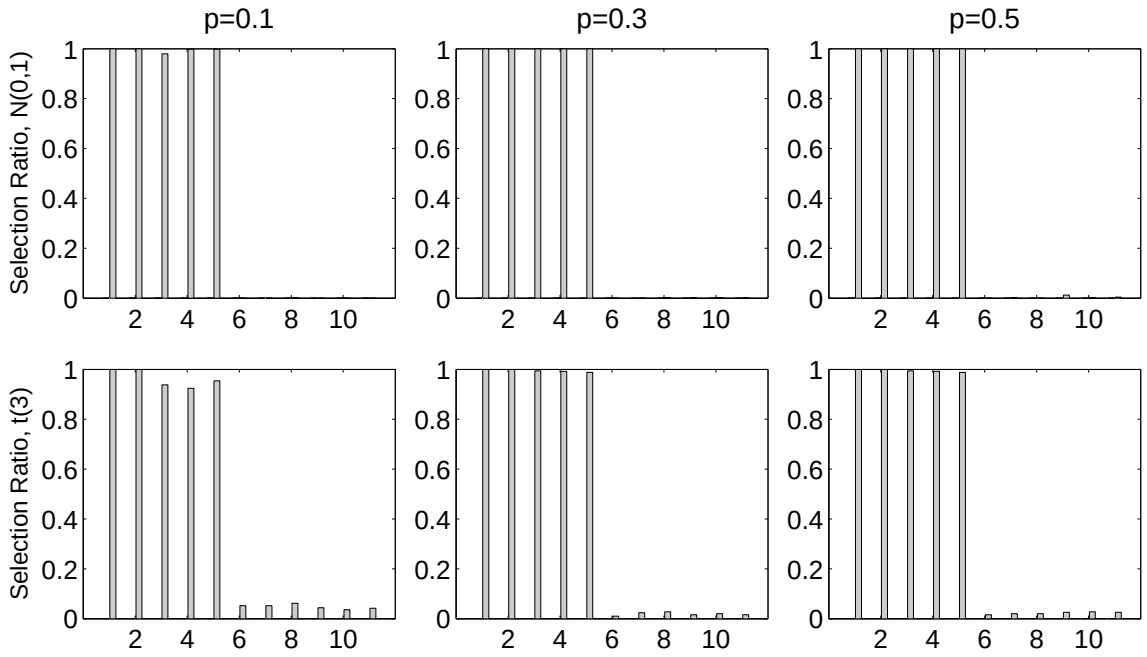

Figure 2: The selection ratio under different error distributions for different coefficient functions when $d = 10$.

variable selection method.

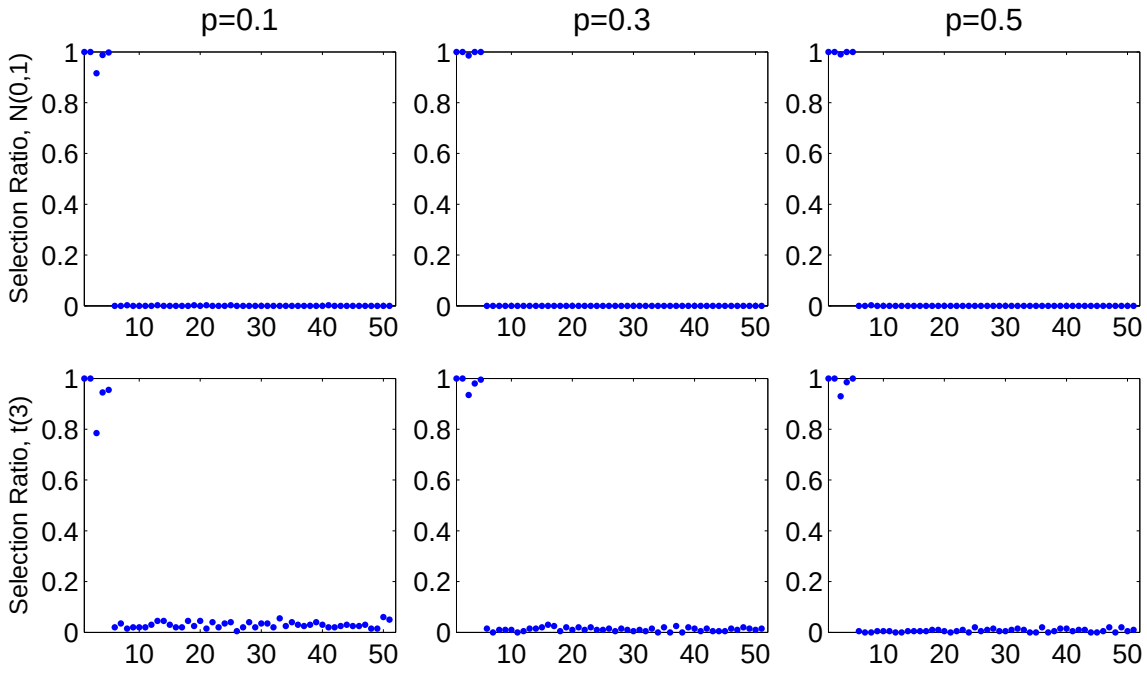

Figure 3: The selection ratio under different error distributions for different coefficient functions when $d = 50$.

Table 2 lists the oracle proportions, the IMSE and TIMSE for d=10. In general, the

Table 2: 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	SCAD	Oracle	SCAD	Oracle	SCAD	Oracle	SCAD	Oracle	SCAD	Oracle
Oracle perc.	0.976	1	0.72	1	0.992	1	0.91	1	0.98	1	0.894	1
$\beta_0(u)$	0.0863	0.0891	0.3078	0.2247	0.0268	0.0273	0.0607	0.0601	0.0213	0.0214	0.0431	0.0451
$\beta_1(u)$	0.1611	0.1667	0.3285	0.3557	0.1071	0.1174	0.1600	0.1746	0.1044	0.1106	0.1581	0.1725
$\beta_2(u)$	0.1264	0.1238	0.4890	0.2932	0.0561	0.0637	0.1360	0.1320	0.0497	0.0604	0.1101	0.1170
$\beta_3(u)$	0.0270	0.0192	1.3307	0.0643	0.0086	0.0084	0.1111	0.0237	0.0077	0.0077	0.0439	0.0192
$\beta_4(u)$	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

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 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 $\beta_1(u)$ reduces from 0.3285 to 0.1600, a 51% reduction when p increases from 0.1 to 0.3, while it only has 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.

Another observation from the simulation is that the model performs better under the normal error than the $t(3)$ error does. We observed larger oracle percentage, smaller IMSE and TIMSE for the coefficient functions under the normal error compared to the results under 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 we know the truth (the oracle) in all cases, demonstrating the stable selection performance of the SCAD penalty.

Table 3: 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	
	SCAD	Oracle	SCAD	Oracle	SCAD	Oracle	SCAD	Oracle	SCAD	Oracle	SCAD	Oracle
Oracle Perc.	0.908	1	0.435	1	0.986	1	0.745	1	0.988	1	0.87	1
$\beta_0(u)$	0.1929	0.0884	0.5687	0.2209	0.0289	0.0278	0.0860	0.0599	0.0215	0.0216	0.0450	0.0434
$\beta_1(u)$	0.2064	0.1684	0.3851	0.3340	0.1107	0.1137	0.1858	0.1742	0.1048	0.1123	0.1551	0.1608
$\beta_2(u)$	0.5235	0.1218	0.6934	0.2614	0.0817	0.0646	0.2205	0.1301	0.0608	0.0579	0.1754	0.1085
$\beta_3(u)$	2.0918	0.0196	2.4522	0.0484	0.1083	0.0075	0.3865	0.0254	0.0470	0.0078	0.1681	0.0167
$\beta_4(u)$	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

A similar pattern was observed when the data dimension increases to 50 (Table 3). 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. Coupling with the results displayed in Fig. 2 and 3, the proposed variable selection method shows relatively stable performance to assess gene-environment interactions.

4 A Case Study

We applied the method to a real dataset from a study conducted at 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) depending on new born babies' weight and mothers' gestational age. After data cleaning by removing SNPs with MAF less

than 0.05 or deviation from Hardy-Weinberg equilibrium, the dataset contains 1536 new born babies with 189 genes covering 660 single nucleotide polymorphisms (SNPs).

Mother's body mass index (MBMI), defined as mother's body mass (kg) divided by the square of their height (m^2), is a measure for mothers' body shape and obesity condition. Since a baby resides inside its mother's womb, the environment factor for a baby is defined through its mother, such as mother's obesity condition (MBMI) or age. Increasing evidence has indicated that both pre-pregnant weight (BMI) and weight gain in pregnancy have big influence on babies' birth weight (Stamnes Koepp et al. [19]. Due to the complicated interaction between fetus' genes and mother's obesity level, baby's birth weight might be different for a fetus with the same gene but under different environment conditions. The variation in birth weight could be explained or partially by the underlying genetic machinery and how genes respond to mother's obesity condition to affect birth weight.

Janus kinase/signal transducers and activators of transcription (JAK/STAT) signaling pathway is the main signaling mechanism for a broad range of cytokines and growth factors in mammals [20]. Total 68 SNPs covering 24 genes in the data were extracted for this pathway. We applied the SCAD penalty method to the pathway and selected one SNP (2069762) located in the exon region in gene Interleukin 9 with constant effect. This means that the SNP is associated with birth weight but is not sensitive to mother's BMI condition. All the other SNPs have no effect and the intercept term shows varying effect.

We also conducted the single SNP based analysis as shown in Ma et al [1] for this SNP by fitting the following model

$$Y = \beta_0(X) + \beta_1(X)G + \varepsilon,$$

We first tested $H_0 : \beta_1(X) = \beta$ and obtained a p-value of 0.0913. This implies that the coefficient is a constant. Then we fitted a partial linear model $Y = \beta_0(X) + \beta G + \varepsilon$ without G×E interaction, and tested $H_0 : \beta = 0$ and obtained a p-value of 7.32×10^{-5}, which gives strong evidence of association of the SNP with birth weight. We did the same analysis for all other SNPs in the same pathway and found no SNPs with p-value less than 0.001. The single SNP-based analysis confirms the variable selection result by the SCAD penalty approach.

5 Discussion

The significance of G×E interactions in complex disease traits has stimulated waves of discussion. A diversity of statistical models have been proposed to assess the gene effect under different environmental exposures, as reviewed in Cornelis et al [20]. The success of gene set based association analysis, as shown in Wang et al [9], Cui et al [6], Wu and Cui [21] and Schaid et al [7], motivates 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.

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. We can separate the varying-coefficients and constants 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. A two-stage iterative estimation procedure with double SCAD penalty functions was developed following Tang et al. [11]. Asymptotic properties of the two-stage estimator were established under suitable regularity conditions.

The current work only demonstrates the case with one environment factor. It is broadly recognized that the etiology of many complex disease is less likely to be affected by one environment factor, but is rather heterogeneous. When multiple continuously measured environment factors (say K_1) are measured (denoted as $\boldsymbol{Z}_1$), 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 $X_0 = 1$. This model is called the additive varying-coefficient model. The same estimation and variable selection framework can be applied to select important genetic players that

show sensitivity to different environmental stimuli. When discrete environment 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

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

which is termed as 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 $\boldsymbol{Z}_2$. We will evaluate this in a future study. In addition, we will extend the current framework to a generalized linear model framework to consider cases when a disease trait is measured as a binary response. In this case, we are interested in modeling $E[Y|\boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{X}] = \sum_{j=0}^{d} \left\{ \sum_{k=1}^{K_1} \beta_{kj}(Z_{1k}) + \sum_{l=1}^{K_2} \alpha_{lj} Z_{2l} \right\} X_j$.

In this study, we implemented the estimation through the local quadratic approximation method. It is known that LQA may suffer from the efficiency loss caused by repeated factorizations of large matrices, especially when the dimension of the predictors gets large. Thus, the LQA method may limit the power of the proposed framework to dissect G×E interactions. An efficient alternative is to use the group coordinate descent (GCD) approach. We will investigate this in our future work to improve the computational efficiency.

Acknowledgements

The authors wish to thank Dr. Lifeng Wang for insightful discussions on parameter estimation. This work was partially supported by NSF grant DMS-1209112 and by National Natural Science Foundation of China 31371336.

Appendix: 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 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 et al [15], and Lemma 2 follows from Corollary 6.21 of Schumaker [18].

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) $\|\tilde{\gamma}_{j*}\|_{L_2} > C_3$, $j = 0, \ldots, v$; $\tilde{\gamma}_{j1} = \beta_j$, $\|\tilde{\boldsymbol{\gamma}}_{j*}\|_{L_2} = 0$, $j = v+1, \ldots, c$; $\tilde{\gamma}_j = \mathbf{0}$, $j = c+1, \ldots, d$;

(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.

(A) 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 $\hat{\boldsymbol{\gamma}}^{\mathcal{VC}}$ of $Q_1(\boldsymbol{\gamma})$ exists in a neighborhood of $\tilde{\boldsymbol{\gamma}}$ where

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

Let $\alpha_n = n^{-\frac{1}{2}}k_n + a_n$, where $a_n := \max_j\{|p'_{\lambda_1}(\|\tilde{\boldsymbol{\gamma}}_{j*}\|)|, |p'_{\lambda_2}(|\tilde{\gamma}_{j,1}|)| : \tilde{\boldsymbol{\gamma}}_{j*} \neq \mathbf{0}, \tilde{\gamma}_{j,1} \neq 0\}$. 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

$$P\left\{\inf_{\|\boldsymbol{\delta}\|=C} Q_1(\hat{\boldsymbol{\gamma}}^{\mathcal{VC}}) \geq Q_1(\tilde{\boldsymbol{\gamma}})\right\} \geq 1 - \varepsilon, \tag{B.2}$$

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

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

With assumption (A6), we can prove that Δ_2 dominates Δ_3 uniformly in $\|\boldsymbol{\delta}\| = C$. Therefore, (B.2) 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 $j = v+1, \ldots, d$ in probability, it is sufficient to demonstrate that $\hat{\boldsymbol{\gamma}}_{j*}^{vc} = \mathbf{0}$, $j = v+1, \ldots, d$. 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

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

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 $\hat{\boldsymbol{\gamma}}_{j*}^{vc} = \mathbf{0}$, $j = v+1, \ldots, d$. This completes the proof of Theorem 1(1), part 1. □

(B) 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

$$Q_2(\boldsymbol{\gamma}_{(v)}, \boldsymbol{\gamma}_{(d)}) = \sum_{i=1}^{n}\left(Y_i - \boldsymbol{U}_{(v)i}^T\boldsymbol{\gamma}_{(v)} - \boldsymbol{U}_{(d)i}^T\boldsymbol{\gamma}_{(d)}\right)^2 + n\sum_{j=v+1}^{d} p_{\lambda_2}(|\gamma_{j,1}|). \tag{B.3}$$

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

$$P\left\{\inf_{\|\boldsymbol{\delta}\|=C} Q_2(\hat{\boldsymbol{\gamma}}_{(v)}, \hat{\boldsymbol{\gamma}}_{(d)}) \geq Q_2(\tilde{\boldsymbol{\gamma}}_{(v)}, \tilde{\boldsymbol{\gamma}}_{(d)})\right\} \geq 1 - \varepsilon. \tag{B.4}$$

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

$$\begin{aligned} D_n(\boldsymbol{\delta}_v, \boldsymbol{\delta}_d) &= Q_2(\hat{\boldsymbol{\gamma}}_{(v)}, \hat{\boldsymbol{\gamma}}_{(d)}) - Q_2(\tilde{\boldsymbol{\gamma}}_{(v)}, \tilde{\boldsymbol{\gamma}}_{(d)}) \\ &= -2\alpha_n \sum_{i=1}^{n} \left[\varepsilon_i + X_1^T R(Z_i)\right] \left[\boldsymbol{U}_{(v)i}^T \boldsymbol{\delta}_{(v)} + \boldsymbol{U}_{(d)i}^T \boldsymbol{\delta}_{(d)}\right] \\ &\quad + \alpha_n^2 \sum_{i=1}^{n} \left[\boldsymbol{U}_{(v)i}^T \boldsymbol{\delta}_{(v)} + \boldsymbol{U}_{(d)i}^T \boldsymbol{\delta}_{(d)}\right]^2 + n \sum_{j=v+1}^{d} \left[p_{\lambda_2}(|\hat{\gamma}_{j,1}|) - p_{\lambda_2}(|\tilde{\gamma}_{j,1})|\right] \\ &:= \Delta_1 + \Delta_2 + \Delta_3, \end{aligned}$$

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

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i [\boldsymbol{U}_{(v)i}^T \boldsymbol{\delta}_{(v)} + \boldsymbol{U}_{(d)i}^T \boldsymbol{\delta}_{(d)}] = O_p(\|\boldsymbol{\delta}\|). \tag{B.5}$$

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,

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

Recall assumption A6, then it follows that, by choosing an enough large C, Δ_2 dominates Δ_1 uniformly in $\|\boldsymbol{\delta}\| = C$. Consequently (B.4) 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 $j = 0, \ldots, v$

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

Recall Lemma 1, 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). □

(C) 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 (A). 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

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

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). □

Proofs 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

$$\begin{aligned}Q_2(\boldsymbol{\gamma}_{(v)}, \boldsymbol{\gamma}_{(d)}) &= \sum_{i=1}^{n}\left(Y_i - \boldsymbol{U}_{(v)i}^T\boldsymbol{\gamma}_{(v)} - \boldsymbol{U}_{(c)i}^T\boldsymbol{\gamma}_{(c)}\right)^2 + n\sum_{j=v+1}^{c} p_{\lambda_2}(|\gamma_{j,1}|) \\ &:= Q_2(\boldsymbol{\gamma}_{(v)}, \boldsymbol{\gamma}_{(c)}).\end{aligned} \tag{B.6}$$

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

$$\begin{aligned}
\frac{\partial Q_2(\hat{\boldsymbol{\gamma}}_{(v)}, \hat{\boldsymbol{\gamma}}_{(c)})}{\partial \hat{\boldsymbol{\gamma}}_{(v)}} &= -2\sum_{i=1}^{n} \boldsymbol{U}_{(v)i}\left[Y_i - \boldsymbol{U}_{(v)i}^T\hat{\boldsymbol{\gamma}}_{(v)} - \boldsymbol{U}_{(d)i}^T\hat{\boldsymbol{\gamma}}_{(d)}\right] = 0; \\
\frac{\partial Q_2(\hat{\boldsymbol{\gamma}}_{(v)}, \hat{\boldsymbol{\gamma}}_{(c)})}{\partial \hat{\boldsymbol{\gamma}}_{(c)}} &= -2\sum_{i=1}^{n} \boldsymbol{U}_{(c)i}\left[Y_i - \boldsymbol{U}_{(v)i}^T\hat{\boldsymbol{\gamma}}_{(v)} - \boldsymbol{U}_{(c)i}^T\hat{\boldsymbol{\gamma}}_{(c)}\right] \\
&\quad + n\sum_{j=v+1}^{c} p_{\lambda 2}^{'}(|\hat{\gamma}_{j,1}|)\mathrm{sgn}(\hat{\gamma}_{j,1}) = 0.
\end{aligned} \tag{B.7}$$

By applying Taylor expansion on $p_{\lambda 2}^{'}(|\hat{\gamma}_{j,1}|)$ in (B.7), 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

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{U}_{(c)i}\left[\varepsilon_i + X_i^T r(z_i) + \boldsymbol{U}_{(v)i}^T(\boldsymbol{\gamma}_{(v)} - \hat{\boldsymbol{\gamma}}_{(v)}) + \boldsymbol{U}_{(c)i}^T(\boldsymbol{\gamma}_{(c)} - \hat{\boldsymbol{\gamma}}_{(c)})\right] + o_p(\hat{\boldsymbol{\gamma}}_{(c)} - \boldsymbol{\gamma}_{(c)}) = 0. \tag{B.8}$$

Meanwhile, a straightforward calculation yields

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{U}_{(v)i}\left[\varepsilon_i + \boldsymbol{X}_i^T r(u_i) + \boldsymbol{U}_{(v)i}^T(\boldsymbol{\gamma}_{(v)} - \hat{\boldsymbol{\gamma}}_{(v)}) + \boldsymbol{U}_{(c)i}^T(\boldsymbol{\gamma}_{(c)} - \hat{\boldsymbol{\gamma}}_{(c)})\right] = 0. \tag{B.9}$$

Recall the definition of Φ_n and Ψ_n, (B.9) is equivalent to

$$\hat{\boldsymbol{\gamma}}_{(v)} - \boldsymbol{\gamma}_{(v)} = \Phi_n^{-1}\left\{\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{U}_{(v)i}\left[\varepsilon_i + \boldsymbol{X}_i^T r(z_i)\right] + \Psi_n[\boldsymbol{\gamma}_{(c)} - \hat{\boldsymbol{\gamma}}_{(c)}]\right\}. \tag{B.10}$$

Plugging (B.10) into (B.8) results in

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{U}_{(c)i}\left\{\varepsilon_i + X_i^T r(z_i) - \boldsymbol{U}_{(v)i}^T\Phi_n^{-1}\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{U}_{(v)i}\left[\varepsilon_i + \boldsymbol{X}_i^T r(z_i)\right]\right\} \\
&= \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{U}_{(c)i}\left[\boldsymbol{U}_{(c)i} - \Psi_n^T\Phi_n^{-1}\boldsymbol{U}_{(v)i}\right]^T(\hat{\boldsymbol{\gamma}}_{(c)} - \boldsymbol{\gamma}_{(c)}) + o_p(\hat{\boldsymbol{\gamma}}_{(c)} - \boldsymbol{\gamma}_{(c)}).
\end{aligned} \tag{B.11}$$

Together with the facts that

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

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 (B.11) 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

$$\Sigma = E\left(\boldsymbol{U}_{(c)}\boldsymbol{U}_{(c)}^T\right) - E\left\{E(\Psi_n^T|Z)E(\Phi_n|Z)^{-1}E(\Psi_n|Z)\right\}. \tag{B.12}$$

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

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

[2] Wu, C. and Cui, Y. H. (2013). A novel method for identifying nonlinear gene-environment interactions in case-control association studies. *Hum. Genet.* (Accepted).

[3] Guo, S. W.(2000). Gene-environment interaction and the mapping of complex traits: some statistical models and their implications. *Hum. Hered.* **50** 286–303.

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

[5] Maity, A., Carrol, R. J., Mammen, E. et al. (2009). Testing in semiparametric models with interaction, with applications to gene-environment interactions. *J. Roy. Stat. Soc. B* **71** 75–96.

[6] Cui, Y. H., Kang, G. L., Sun, K. L. et al. (2008). Gene-centric genomewide association study via entropy. *Genetics* **179** 637–650.

[7] Schaid, D. J., Sinnwell, J. P., Jenkins, G. D. et al. (2012). Using the gene ontology to scan multilevel gene sets for associations in genome wide association studies. *Genet. Epidemiol.* **36** 3-16.

[8] Efron, B. and Tibshirani, R. (2007). On testing the significance of sets of genes. *Ann. Appl. Stat.* **1** 107-129.

[9] Wang, K., Li, M. and Hakonarson, H. (2011). Analysing biological pathways in genome-wide association studies. *Nat. Rev. Genet.* **11** 843-54.

[10] T. Hastie and R. Tibshirani. (1993). Varying-coefficient models. *J. R. Statist. Soc. B* **55** 757–796.

[11] Tang, Y. L., Wang, H. X., Zhu, Z. Y. et al. (2012). A unified variable selection approach for varying coefficient models. *Stat. Sinica.* **22** 601–628.

[12] Fan, J.Q. and Li, R.Z. (2001). Variable selection via nonconcave penzlied likelihood and its oracle properties. *J. Amer. Stat. Assoc.* **96** 1348–1360.

[13] Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat.* **6** 461–464.

[14] Kim, M. O. (2007). Quantile regression with varying coefficients. *Ann. Stat.* **35** 92–108.

[15] Huang, J. H., Wu, C. O. and Zhou, L. (2004). Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Stat. Sinica.* **14** 763–788.

[16] Huang, J. Z., Wu, C. O. and Zhou, L. (2002). Varying-coefficient models and basis function approximation for the analysis of repeated measurements. *Biometrika* **89** 111–128.

[17] L.F. Wang, H.Z. Li and J.Z. Huang. (2008). Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *J. Amer. Stat. Assoc.* **103** 1556-1569.

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

[19] Stamnes Koepp, U. M., Andersen, L.F., Dahl-Joergensen, K., et al. (2012) Maternal pre-pregnant body mass index, maternal weight change and offspring birthweight. *Acta Obstet. Gynecol. Scand.* **91** 243–249.

[20] Rawlings, J. S., Rosler, K. M. and Harrison, D. A. (2004) The JAK/STAT signaling pathway. *J. Cell Sci.* **117** 1281–1283.

[21] Wu, C and Cui, Y. H. (2013). Boosting signals in gene-based association studies via efficient SNP selection. *Brief. Bioinform.* (In Press).

[22] Cornelis, M. C., Tchetgen, E. J., Liang, L. et al. (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.