Technical Report RM 701, Department of Statistics and Probability, Michigan State University

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

Running head: Variable selection for nonlinear G×E interactions

***Corresponding author**

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. 2011; Wu and Cui, 2013). As illustrated in Wu and Cui (2013), 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 [1], semi-parametric methods in Chatterjee et al [4] and Maity et al [5], and non-parametric methods in Ma et al [2] and Wu and Cui [3]. 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 [9] and Efron and Tibshirani [8], and the pathway-based analysis in Wang et al. (2007). 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 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. Readers are referred to the work of Ma et al. (2011) and Wu and Cui (2013) for the motivation of modeling nonlinear interactions. 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, proposed by Hastie and Tibshirani [11], 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 vector $\mathbf{X}_i$ with the first component X_{i0} being 1, $\beta_j(\cdot)$'s are unknown varying-coefficient functions, Z_i's are the scalar index variable, and ε_i is the random error such that $E(\varepsilon|X, Z) = 0$ and $Var(\varepsilon|X, Z) = \sigma^2 < \infty$. 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

$$w_k = \{0 = w_{k,0} < w_{k,1} < \ldots < w_{k,k_n} < w_{k,k_n+1} = 1\}$$

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 $I_s = [w_{k,s}, w_{k,s+1}), s = 0, \ldots, N_n - 1$ and $I_{N_n} = [w_{j,N_n}, w_{j,N_n+1})$; 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 functions $\beta_j(Z)$. By changing of equivalent basis, the basis expansion can be reexpressed as

$$\beta_j(\cdot) \approx \sum_{l=1}^{L_j} \gamma_{jl} B_{jl}(\cdot) \doteq \gamma_{j1} + \tilde{B}_j^T(\cdot)\gamma_{j,*}$$

where the spline coefficient vector $\boldsymbol{\gamma}_j \doteq (\gamma_{j,1}, \boldsymbol{\gamma}_{j*}^T)^T$, 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. 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,

$$\begin{aligned} Q(\boldsymbol{\gamma}) =& \frac{1}{n}\sum_{i=1}^{n}\left[Y_i - \sum_{j=0}^{d}\sum_{l=1}^{L}\gamma_{jl}X_{ij}B_{jl}(Z_i)\right]^2 + \sum_{j=1}^{d} p_{\lambda_1}(\|\boldsymbol{\gamma}_{j*}\|_2) \\ &+ \sum_{j=1}^{d} p_{\lambda_2}(|\gamma_{j1}|)I(\|\boldsymbol{\gamma}_{j*}\|_2 = 0) \end{aligned} \tag{2}$$

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,

$$\begin{aligned} Q(\boldsymbol{\gamma}) =& (\boldsymbol{Y} - \boldsymbol{U}\boldsymbol{\gamma})^T(\boldsymbol{Y} - \boldsymbol{U}\boldsymbol{\gamma}) + n\sum_{j=1}^{d} p_{\lambda 1}(\|\gamma_{j*}\|_2) \\ &+ n\sum_{j=1}^{d} p_{\lambda 2}(|\gamma_{j1}|)I(\|\boldsymbol{\gamma}_{j*}\|_2 = 0) \end{aligned} \tag{4}$$

where $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$, $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_0^T, \ldots, \boldsymbol{\gamma}_d^T)^T$, and $\boldsymbol{U} = (U_1^T, \ldots, U_n^T)^T$ with $U_i = (X_{i0}B(Z_i)^T, \ldots, X_{id}B(Z_i)^T)^T$. The first penalty function in (2) 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 Fan and Li (2001) [15], 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 a=3.7 and u >0. Here we use a similar quadratic approximation by substituting u with $\|\gamma_{j*}\|_2$ and $|\gamma_{k1}|$ in LQA, for $k = 0, ..., d$. Therefore 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

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

The sets of predictors with varying, non-zero constant, and zero effects are termed as $\mathcal{V}$, $\mathcal{C}$ and $\mathcal{Z}$ respectively. Following Tang et al. (2012), 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 $\|\hat{\gamma}_{j*}^{\mathcal{VC}}\|_2$=0, otherwise it will stay in $\mathcal{V}$.

At stage 2, using the LQA (6) and dropping the irrelevant constant terms, we minimize the 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 $|\hat{\gamma}_{k,1}^{\mathcal{CZ}}|$=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 Selection of tuning parameters

In this section, we choose the tuning parameters N,p, λ_1 and λ_2 from a data driven procedure where N is the number of interior knots uniformly spaced on [0,1]; p is the degree of the spline basis. Here p and 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 [16] to choose N and p. The range for 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 N and p can be achieved via a two-dimensional grid search, according to the following criterion:

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

where $\mathrm{RSS}_{N,p} = (\boldsymbol{Y} - \boldsymbol{U}\hat{\boldsymbol{\gamma}})^T(\boldsymbol{Y} - \boldsymbol{U}\hat{\boldsymbol{\gamma}})/n$, $\hat{\boldsymbol{\gamma}} = (\hat{\boldsymbol{\gamma}}_0^T, \mathbf{0}^T, \ldots, \mathbf{0}^T)^T$. Conditional on the selected 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 $\mathrm{RSS}_{\lambda_1} = (\boldsymbol{Y} - \boldsymbol{U}\hat{\boldsymbol{\gamma}}_{\lambda_1})^T(\boldsymbol{Y} - \boldsymbol{U}\hat{\boldsymbol{\gamma}}_{\lambda_1})/n$, $\hat{\boldsymbol{\gamma}}_{\lambda 1}$ 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 $\mathrm{RSS}_{\lambda_2} = (\boldsymbol{Y} - \boldsymbol{U}\hat{\boldsymbol{\gamma}}_{\lambda_2})^T(\boldsymbol{Y} - \boldsymbol{U}\hat{\boldsymbol{\gamma}}_{\lambda_2})/n$, , $\hat{\boldsymbol{\gamma}}_{\lambda_2}$ 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 $\mathscr{H}_r$ 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 b with $r = r_1 + r_2$, i.e., $|h^{r_1}(z_1) - h^{z_1}(z_2)| \leq C_0|z_1 - z_2|^{r_2}$ where $0 \leq z_1, z_2 \leq 1$ and C_0 is a finite positive constant. Then $\beta_j(z) \in \mathscr{H}_r$, $j = 0, 1, \ldots, v$, 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 $\lambda_0 \leq \ldots \leq \lambda_d$ be the eigenvalues of $E[\mathbf{X}\mathbf{X}^T|Z = z]$. Then λ_j $(k = 0, \ldots, d)$ are uniformly bounded away from 0 and infinity in probability. In addition, the random design vector 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) $\max_j\{|p''_{\lambda_1}(|\gamma_{j*}|)| : \gamma_{j*} \neq 0\} \to 0$ as $n \to \infty$ and $\max_j\{|p''_{\lambda_2}(|\gamma_{j1}|)| : \gamma_{j1} \neq 0\} \to 0$ as $n \to \infty$

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

The above assumptions are commonly used in literature of polynomial splines and variable selections. The assumption similar to (A1) could be found in Kim [17] and Tang et al [13]. (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 [14, 18] and Wang et al [19]. (A4) suggests that the knot sequence is quasi-uniform on [0,1], by Schumaker [12]. (A5-A7) are conditions on tuning parameters, of which (A5) could be found in Tang et al [13]; (A6) and (A7) are similar to those in Fan and Li [15] and Wang et al [19].

Theorem 1. Under the assumptions (A1-A7) and suppose $k_n = O_p\left(n^{\frac{1}{2r+1}}\right)$, 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) $\|\hat{\beta}_j - \beta_j\|_{L_2} = O_p(n^{\frac{-r}{2r+1}})$, $j = 0, \ldots, v$.

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 $k_n = O_p(n^{\frac{1}{2r+1}})$, then as $n \to \infty$,

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

where Σ is defined in the Appendix.

3 Simulation

The performance of our proposed approach is demonstrated through extensive simulation studies. We use the percentage of choosing the true model out of total R replicates, defined as oracle percentage, to evaluate the accuracy of variable selection by identifying varying, non-zero constant and zero effects. The precision of estimation is 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 $\hat{\beta}_j^{(r)}$ is evaluated. We use the integrated mean squared error (IMSE) of $\hat{\beta}_k(x)$, 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 (IMSE) of all the d coefficients (TIMSE), defined as TIMSE=$\sum_{j=1}^{d} \hat{\beta}_j(z)$, to evaluate the overall estimation accuracy. Note that IMSE$(\hat{\beta}_j)$ will be reduced to MSE$(\hat{\beta}_j)$ when $\hat{\beta}_j$ is a constant. The percentage of correctly selecting true functions (defined as the selection ratio) is used to evaluate the selection performance. Tang et al. (2012) proposed to use the adaptive LASSO penalty. We compare the performance of the SCAD penalty with the ALASSO penalty in the simulation studies.

In example 1, we simulate 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 ~Unif(0,1), and the predictors X_i are generated from a multivariate normal distribution with mean $\mathbf{0}$ and $Cov(X_j, X_{j'}) = 0.5^{|j-j^{'}|}$ for $0 \leq j, j^{'} \leq d$. The performance is 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 $j > 4$. The results are listed in Figure 1 and Table 1.

Figure 1 shows the selection ratio for predictors under different error distributions with SCAD and ALASSO penalty, for the first 5 predictors and false positives for the rest predictors. The top panel denotes the result for $d = 10$ and the bottom panel for $d = 50$. Under the $N(0,1)$ error, the performance of the SCAD and ALASSO penalization methods performs very similarly. However, under the $t(3)$ error scenario, the SCAD penalty is capable of correctly selecting true effect with high percentages, while maintaining a very small percentages of choosing false positives, in comparison to the results by ALASSO penalty. In addition, the SCAD penalty performs relatively stable when the data dimension increases.

The oracle percentage and parameter estimation results are summarized in Table 1. Here We compute IMSEs for all predictors, including β_4 and β_5 to reflect the overall estimation precision. When β_j $(j = 4, 5)$ is selected as non-zero constant, IMSE reduces to MSE. The IMSEs will be calculated if β_j $(j = 4, 5)$ is incorrectly identified as varying effect. In all the cases and under different error distributions, the SCAD approach demonstrates superior performance over the ALASSO approach. The SCAD approach has better oracle percentage and smaller IMSE and TIMSE compared to the ALASSO approach.

The above simulation considers continuous predictor variables to compare the performance of the SCAD penalty with the ALASSO penalty. Since the paper deals with gene×environment

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

		N(0,1) error			t(3) error		
		SCAD	ALASSO	Oracle	SCAD	ALASSO	Oracle
d=10	Oracle Perc.	0.972	0.82	1	0.92	0.315	1
	IMSE						
	$\beta_0(u)$	0.0214	0.0243	0.0216	0.0398	0.0448	0.1929
	$\beta_1(u)$	0.0902	0.0930	0.0951	0.1166	0.1254	0.3392
	$\beta_2(u)$	0.0365	0.1018	0.0431	0.0764	0.2211	0.5859
	$\beta_3(u)$	0.0122	0.2405	0.0032	0.0753	0.6248	0.1775
	$\beta_4(u)$	0.0045	0.0405	0.0031	0.0183	0.1713	0.1100
	TIMSE	0.1648	0.5075	0.1661	0.3282	1.3000	0.4017
d=50	Oracle Perc.	0.945	0.635	1	0.8	0.012	1
	IMSE						
	$\beta_0(u)$	0.0221	0.0230	0.0219	0.0431	0.0612	0.0426
	$\beta_1(u)$	0.0878	0.0896	0.0927	0.1230	0.1477	0.1253
	$\beta_2(u)$	0.0404	0.0551	0.0428	0.1042	0.0969	0.0751
	$\beta_3(u)$	0.0478	0.0776	0.0027	0.1727	0.0771	0.0105
	$\beta_4(u)$	0.0101	0.0165	0.0029	0.0239	0.0608	0.0083
	TIMSE	0.2086	0.2966	0.1631	0.5146	2.4926	0.2619

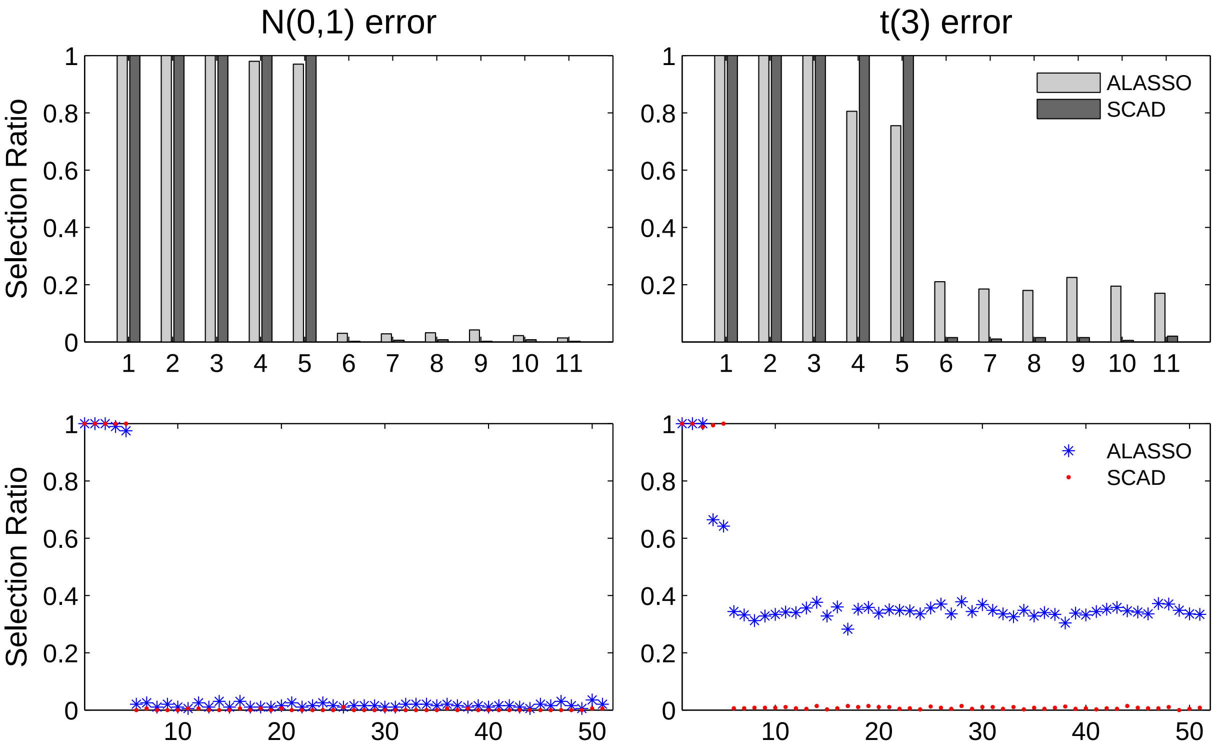

Figure 1: The selection ratio under different error distributions for different coefficient functions.

interaction studies, in example 2 we simulate genetic predictors which are discrete in nature. We consider a quantitative phenotypic measure Y and multiple genetic factors X 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 the SNP X_i was coded with 3 categories (1,0,-1) for genotypes (AA,Aa,aa) respectively. We simulate 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 are assumed the same as those given in example 1. We evaluate the performance under $n = 500$ with 500 replicates. Better performance results for large samples ($n > 500$) are observed, hence are omitted to save

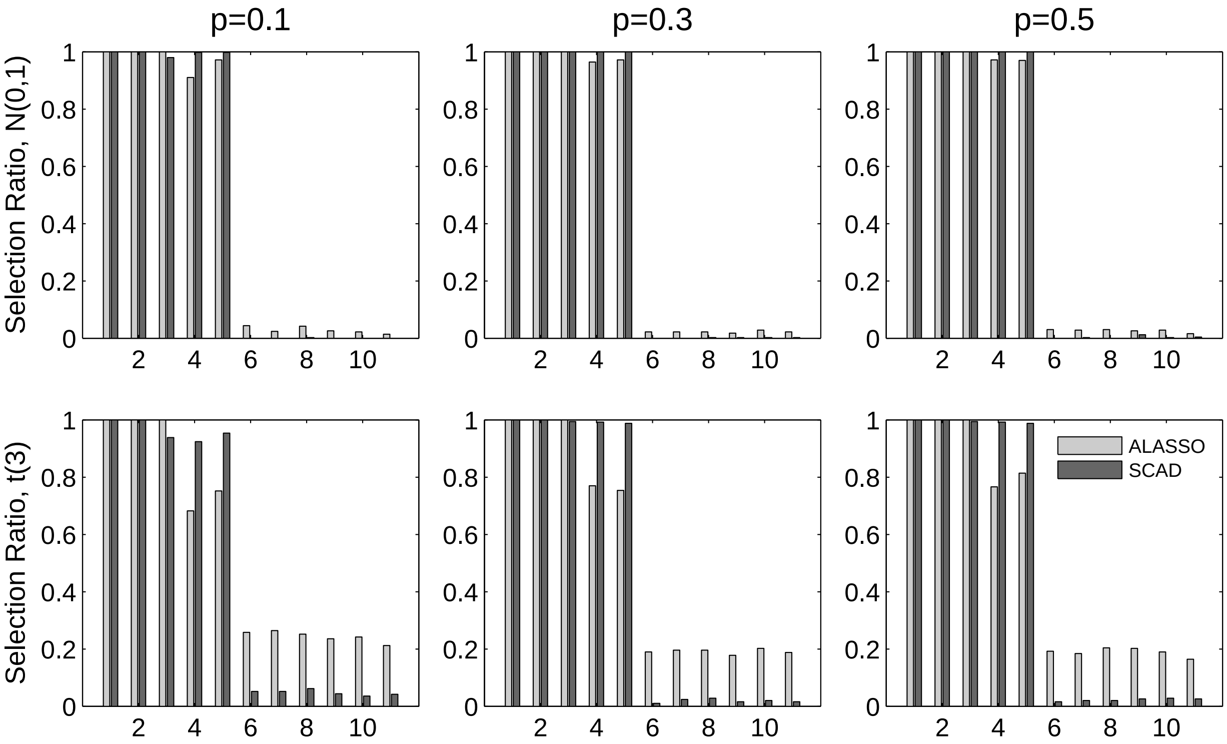

Figure 2: The selection ratio under different error distributions for different coefficient functions when $d = 10$.

Figure 2 shows the selection ratio when d=10, under different combinations of MAF and error distributions. The height of bars represent the selection percentage out of 500 replicates. Under both error distributions, the SCAD penalty has higher percentage of choosing true positive SNPs and lower percentages of choosing false positive SNPs in comparison to the results by the ALASSO penalty. As MAF increases, both approaches lead to higher selection ratios for true positive SNPs and lower selection ratios for false positive SNPs. The SCAD

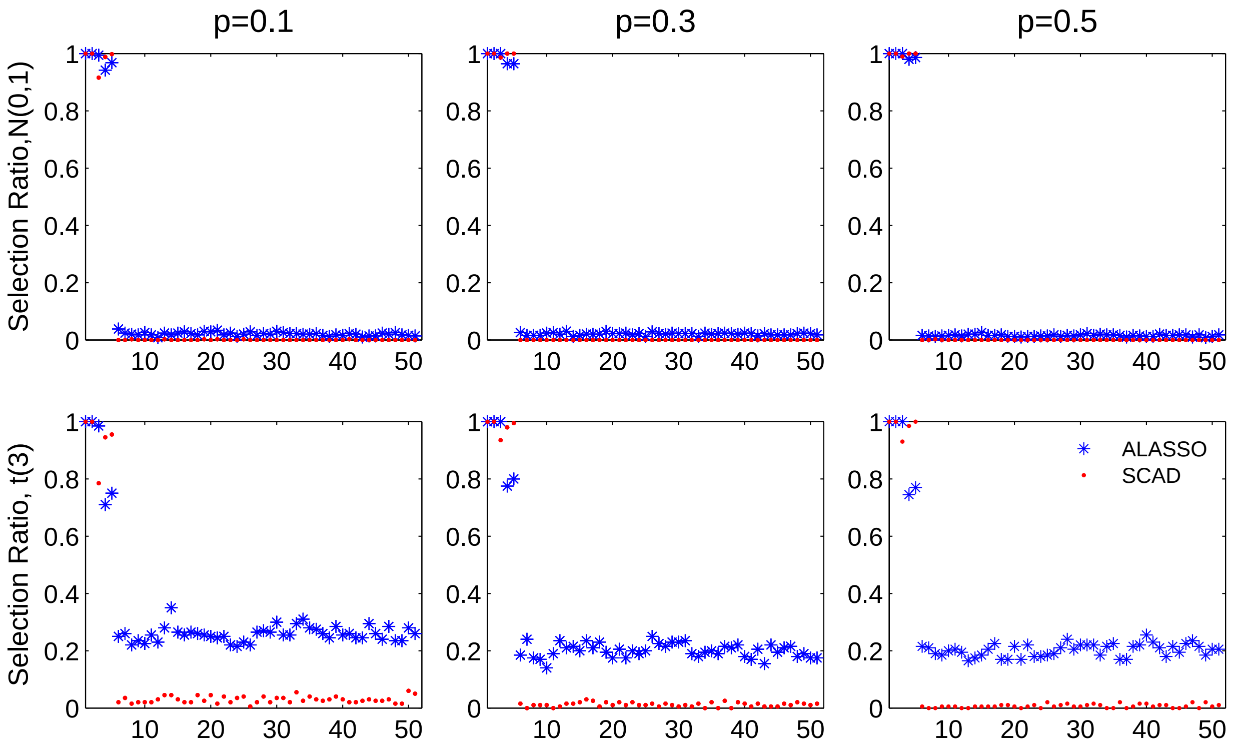

Figure 3: The selection ratio under different error distributions for different coefficient functions when $d = 50$.

penalty performs much better than the ALASSO penalty under the $t(3)$ errors. A similar pattern can be observed when d increases to 50 (Fig. 3). The ALASSO penalty gives constantly high false positive selection ratios for those zero effects under the $t(3)$ errors. The results demonstrate the stable performance of the SCAD penalty, especially under the $t(3)$ errors.

Table 2 presents the oracle proportions and estimation results for d=10. We observe superior performance of the SCAD penalty function over the ALASSO penalty with higher oracle percentage, smaller IMSE for all the coefficient functions and smaller TIMSE under different MAFs and error distributions. The performance of the ALASSO penalty is very unstable under the $t(3)$ error distribution. For example, the TIMSE for SCAD approach is 0.4072 compared to 1.8768 with the ALASSO method under $p_A = 0.5$. A similar pattern can be observed for the high dimensional case (d=50) in Table 3. Even the IMSE and TIMSE are increased in all the cases, the SCAD penalty approach still performs better than the ALASSO method. As the MAF increases from 0.1 to 0.3, we observe sharply decreased IMSE and TIMSE. Under the $t(3)$ error distribution, the ALASSO penalty method barely select the true model with extremely low oracle percentage. In summary, the SCAD penalty function shows consistently better performance over the ALASSO penalty method and should be recommended in real data analysis.

Table 2: List of IMSE, TIMSE, and Oracle Percentage under $N(0,\ 1)$ and $t(3)$ error distributions with SCAD and ALASSO penalty functions when $d = 10$.

		N(0,1) error			t(3) error		
		SCAD	ALASSO	Oracle	SCAD	ALASSO	Oracle
pA=0.1	Oracle Perc.	0.976	0.784	1	0.72	0.268	1
IMSE	$\beta_0(u)$	0.0863	0.1250	0.0891	0.3078	1.6608	0.2247
	$\beta_1(u)$	0.1611	0.1601	0.1667	0.3285	0.3947	0.3557
	$\beta_2(u)$	0.1264	0.1358	0.1238	0.4890	1.2776	0.2932
	$\beta_3(u)$	0.0270	0.1183	0.0192	1.3307	2.8155	0.0643
	$\beta_4(u)$	0.0191	0.0433	0.0174	0.2943	2.1633	0.0475
TIMSE		0.4205	0.6106	0.4162	2.9342	9.2044	0.9855
pA=0.3	Oracle Perc.	0.992	0.84	1	0.91	0.33	1
IMSE	$\beta_0(u)$	0.0268	0.0297	0.0273	0.0607	0.0975	0.0601
	$\beta_1(u)$	0.1071	0.1074	0.1174	0.1600	0.2065	0.1746
	$\beta_2(u)$	0.0561	0.0551	0.0637	0.1360	0.1373	0.1320
	$\beta_3(u)$	0.0086	0.0271	0.0084	0.1111	0.1216	0.0237
	$\beta_4(u)$	0.0066	0.0118	0.0065	0.0443	0.1125	0.0222
TIMSE		0.2007	0.2404	0.2233	0.5311	1.3069	0.4126
pA=0.5	Oracle Perc.	0.98	0.846	1	0.894	0.34	1
IMSE	$\beta_0(u)$	0.0213	0.0214	0.0214	0.0431	0.0485	0.0451
	$\beta_1(u)$	0.1044	0.1043	0.1106	0.1581	0.1721	0.1725
	$\beta_2(u)$	0.0497	0.0507	0.0604	0.1101	0.3270	0.1170
	$\beta_3(u)$	0.0077	0.0210	0.0077	0.0439	0.7984	0.0192
	$\beta_4(u)$	0.0063	0.0103	0.0063	0.0240	0.3082	0.0135
TIMSE		0.1895	0.2177	0.2065	0.4072	1.8768	0.3673

Table 3: List of IMSE, TIMSE, and Oracle Percentage under $N(0,\ 1)$ and $t(3)$ error distributions with SCAD and ALASSO penalty functions when $d = 50$.

		$N(0,1)$ error			$t(3)$ error		
		SCAD	ALASSO	Oracle	SCAD	ALASSO	Oracle
pA=0.1	Oracle Perc.	0.908	0.542	1	0.435	0.025	1
	$\beta_0(u)$	0.1929	0.9911	0.0884	0.5687	1.2335	0.2209
	$\beta_1(u)$	0.2064	0.1988	0.1684	0.3851	0.3484	0.3340
IMSE	$\beta_2(u)$	0.5235	0.8382	0.1218	0.6934	0.4432	0.2614
	$\beta_3(u)$	2.0918	2.0345	0.0196	2.4522	0.7892	0.0484
	$\beta_4(u)$	0.3475	0.4798	0.0158	0.5996	0.4671	0.0445
TIMSE		3.3644	4.7239	0.4140	5.7021	8.9145	0.9092
pA=0.3	Oracle Perc.	0.986	0.642	1	0.745	0.06	1
	$\beta_0(u)$	0.0289	0.0732	0.0278	0.0860	0.1970	0.0599
	$\beta_1(u)$	0.1107	0.1124	0.1137	0.1858	0.1974	0.1742
IMSE	$\beta_2(u)$	0.0817	0.1834	0.0646	0.2205	0.1768	0.1301
	$\beta_3(u)$	0.1083	0.4072	0.0075	0.3865	0.2018	0.0254
	$\beta_4(u)$	0.0229	0.0748	0.0068	0.0840	0.1099	0.0220
TIMSE		0.3526	0.9334	0.2204	1.2288	3.3013	0.4117
pA=0.5	Oracle Perc.	0.988	0.706	1	0.8	0.07	1
	$\beta_0(u)$	0.0215	0.0232	0.0216	0.0450	0.0560	0.0434
	$\beta_1(u)$	0.1048	0.1073	0.1123	0.1551	0.1716	0.1608
IMSE	$\beta_2(u)$	0.0608	0.1269	0.0579	0.1754	0.1525	0.1085
	$\beta_3(u)$	0.0470	0.2846	0.0078	0.1681	0.1501	0.0167
	$\beta_4(u)$	0.0120	0.0444	0.0053	0.0480	0.0889	0.0190
TIMSE		0.2461	0.6426	0.2050	0.6492	2.8755	0.3484

4 Real Data Analysis

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.

The environment factor for a baby inside mother's body is defined through the mother, such as mother's obesity condition (MBMI) or age. Due to the complicated interaction between fetus' genes and mother's obesity level, the birth weight might be different for a fetus with the same gene but under different environment conditions. The phenomenon of regular variation in birth weight could be explained by corresponding genetic variants and how they respond to mother's obesity condition.

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 [21]. Total 68 SNPs covering 24 genes in the data were extracted for this pathway. We applied both the SCAD and ALASSO penalty method to the pathway. Using the SCAD penalty approach, we 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. The ALASSO penalty method only identified the varying-coefficient intercept term and the others were all zero.

To further validate the result, we conducted the single SNP based analysis as shown in Ma et al [2] 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 $7.32{\times}10^{-5}$, 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 [21]. The success of gene set based association analysis, as shown in Wang et al [10], Cui et al [6], Wu and Cui [7] and Schaid et al [9], 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. (2012) [13]. Asymptotic properties of the two-stage estimator were established under suitable regularity conditions.

A comprehensive comparison between the SCAD and ALASSO penalty methods was evaluated. Simulation studies show that the SCAD penalty function performs better than the ALASSO penalty approach under various settings. In the simulations, the estimation accuracy was evaluated via IMSE for varying coefficients and via MSE for constant coefficients. In Tang et al [13], MSE was calculated for the predictors with non-zero constant coefficients when the predictors are corrected identified. This does not reveal the error caused by failure to classify the coefficient as non-zero constant. Thus, we suggest calculating IMSE for all the predictors since IMSE reduces to MSE when the coefficient is a constant. This can lead to a much more accurate assessment on the performance of different methods.

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 environment factors are measured, e.g., K_1 are continuously distributed (denoted as Z_1) and K_2 are discrete (denoted as Z_2), we can extend the current model to a more general case formulated as follows,

$$Y = \sum_{j=0}^{d} \left\{ \sum_{k=1}^{K_1} \beta_{kj}(Z_1) + \sum_{l=1}^{K_2} \alpha_{lj} Z_2 \right\} X_j + \varepsilon$$

This model is called the partial linear varying-coefficient model. The same estimation and variable selection framework can be applied.

In this study, we implemented the estimation through the LQA method. It is well known that LQA suffers from the efficiency loss caused by repeated factorizations of large matrices, especially when the dimension of the predictors gets large. In this case, the LQA method will greatly limit the power of the 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 Lifeng Wang for insightful discussions on parameter estimations. This work was partially supported by NSF grant DMS-1209112.

Appendix: Technical Proofs

Useful notations and lemmas

For convenience, the following notations are adopted :

$\bar{Y} = E(Y|\boldsymbol{X},T)$, $\bar{\boldsymbol{\gamma}} = (\boldsymbol{U}^T\boldsymbol{U})^{-1}\boldsymbol{U}^T\bar{Y}$, $\bar{\boldsymbol{\beta}} = \boldsymbol{B}\bar{\boldsymbol{\gamma}}$

$\boldsymbol{\gamma}_{(v)} = (\boldsymbol{\gamma}_0^T,\ldots,\boldsymbol{\gamma}_v^T)^T$, $\boldsymbol{\gamma}_{(c)} = (\boldsymbol{\gamma}_{v+1}^T,\ldots,\boldsymbol{\gamma}_c^T)^T$, $\boldsymbol{\gamma}_{(d)} = (\gamma_{v+1,1}^T,\ldots,\gamma_{d,1}^T)^T$,

$\tilde{\boldsymbol{\gamma}}_{(v)} = (\tilde{\boldsymbol{\gamma}}_0^T,\ldots,\tilde{\boldsymbol{\gamma}}_v^T)^T$, $\tilde{\boldsymbol{\gamma}}_{(c)} = (\tilde{\boldsymbol{\gamma}}_{v+1}^T,\ldots,\tilde{\boldsymbol{\gamma}}_c^T)^T$, $\tilde{\boldsymbol{\gamma}}_{(d)} = (\gamma_{v+1,1},\ldots,\gamma_{d,1})^T$,

$\boldsymbol{G}_n = (B(z_1),\ldots,B(z_n))(B(z_1),\ldots,B(z_n))^T$, $\boldsymbol{\varepsilon} = (\varepsilon_1,\ldots,\varepsilon_n)^T$

$\Phi_n = n^{-1}\sum_{i=1}^n \boldsymbol{U}_{(v)i}\boldsymbol{U}_{(v)i}^T$, $\Psi_n = n^{-1}\sum_{i=1}^n \boldsymbol{U}_{(v)i}\boldsymbol{U}_{(c)i}^T$, $\Lambda_i = \boldsymbol{U}_{(c)i} - \Psi_n^T\Phi_n^{-1}\boldsymbol{U}_{(c)i}$

We first provide several lemmas to facilitate the proofs of Theorems 1 and 2.

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_0, there exists $\tilde{\boldsymbol{\gamma}} = (\tilde{\boldsymbol{\gamma}}_0^T,\ldots,\tilde{\boldsymbol{\gamma}}_d^T)^T$ satisfying

(1) $\|\tilde{\gamma}_{j*}\|_{L_2} > C_0$, $j = 0,\ldots,v$; $\tilde{\gamma}_{j1} = \beta_j$, $\|\tilde{\boldsymbol{\gamma}}_{j*}\|_{L_2} = 0$, $j = v+1,\ldots,c$; $\tilde{\gamma}_j = \mathbf{0}$, $j = c+1,\ldots,d$

(2) $\sup_{\mathrm{t}\in[0,1]}|\beta_j(z) - B(z)^T\tilde{\boldsymbol{\gamma}}_j| = O_p(k_n^{-r})$, $j = 0,\ldots,d$, where $\tilde{\boldsymbol{\gamma}}_j = (\tilde{\gamma}_{j,1},\tilde{\boldsymbol{\gamma}}_{j*}^T)^T$

(3) $\sup_{(t,x)\in[0,1]\times R^{d+1}}|\boldsymbol{X}^T\beta(z) - \boldsymbol{U}(\boldsymbol{X})'\tilde{\boldsymbol{\gamma}}| = O_p(k_n^{-r})$

Proofs of Theorem 1.

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

Here we first show $\hat{\beta}_j(z)$ is constant for $j = v+1,\ldots,d$ in probability, which amounts to demonstrating $\|\hat{\boldsymbol{\gamma}}_{j*}^{vc}\|_j = \mathbf{0}$, $j = v+1,\ldots,d$ with probability tending to 1, as $n\to\infty$. For

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

let $\alpha_n = n^{-\frac{1}{2}}k_n + a_n$ and $\hat{\boldsymbol{\gamma}}^{vc} = \tilde{\boldsymbol{\gamma}} + \alpha_n\boldsymbol{\delta}$. We want to show that for any given $\varepsilon > 0$, there exists a large constant C such that

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

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

$$\begin{aligned}
D_n(\boldsymbol{\delta}) &= Q_1(\hat{\boldsymbol{\gamma}}^{vc}) - Q_1(\tilde{\boldsymbol{\gamma}}) \\
&= -2\alpha_n\boldsymbol{\delta}\sum_{i=1}^{n}\left[\varepsilon_i + X_1^T r(z_i)\right]\boldsymbol{U}_i^T + \alpha_n^2\boldsymbol{\delta}^2\sum_{i=1}^{n}\boldsymbol{U}_i^T\boldsymbol{U}_i \\
&\quad + n\sum_{j=1}^{d}\left[p_{\lambda_1}(\|\hat{\gamma}_{j*}^{vc}\|) - p_{\lambda_1}(\|\tilde{\gamma}_{j*}\|)\right] \\
&\equiv \Delta_1 + \Delta_2 + \Delta_3
\end{aligned}$$

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 $E(\varepsilon_i|\boldsymbol{U}, z_i) = 0$, 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 1, then

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

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

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

where $f_n = \max_j\{|\tilde{\gamma}_{j*}| : \tilde{\gamma}_{j*} \neq 0\}$. 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 $\hat{\beta}_j(z) = 0$ for $j = v+1, \ldots, d$ in probability, it is sufficient to demonstrate that $\hat{\boldsymbol{\gamma}}_{j*}^{vc} = \mathbf{0}$, $j = v+1, \ldots, d$. It follows from the definition 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 γ_{j*}. It can be shown that,

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

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 $\hat{\gamma}^{vc}_{j,*}$. Therefore, $\hat{\boldsymbol{\gamma}}^{vc}$, the minimizer of Q_1, is achieved at $\hat{\boldsymbol{\gamma}}^{vc}_{j*} = \mathbf{0}$, $j = v+1, \ldots, d$. This completes the proof of Theorem 1(1), part 1. □

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

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

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

We need to show that for any given $\varepsilon > 0$, there exists a large constant C_ε 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

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

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 1 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 is easy to show that by choosing a sufficiently large C, Δ_1 is dominated by Δ_2 uniformly in $\|\boldsymbol{\delta}\| = C$. By Taylor expansion,

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

where $f_n = \max_j\{|\tilde{\gamma}_{j,1}| : \tilde{\gamma}_{j,1} \neq 0\}$. 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 $\hat{\boldsymbol{\gamma}}^{cz}_{(d)} - \tilde{\boldsymbol{\gamma}}_{(d)} = O_p(\alpha_n)$. Then for $j = 0, \ldots, d$

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

Recall Lemma 1, 2 and $k_n = O_p\left(n^{\frac{1}{2r+1}}\right)$, 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(1). □

(C) Proof of Theorem 1(2)

To show $\hat{\beta}_j(z) = 0$ for $j = c+1, \dots, d$, it is sufficient to demonstrate that $\hat{\boldsymbol{\gamma}}^{cz}_{j,1} = 0$, since the constancy of $\beta_j(z)$, $j = v+1, \dots, d$ was already established in (A). It follows from the definition that 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

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

We can prove that

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

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 $\hat{\boldsymbol{\gamma}}^{cz}_{j,1} = 0$, $j = c+1, \dots, d$. This completes the proof of Theorem 1(2). □

Proofs of Theorem 2.

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

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

Since $(\hat{\boldsymbol{\gamma}}_{(v)}, \hat{\boldsymbol{\gamma}}_{(c)})$ is the minimal value 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

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

and

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

and recall the definition of Λ_i, a direct computation from (B.11) leads to

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

It follows from law of large numbers that

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

where $\Sigma = E\left(\boldsymbol{U}_{(c)}\boldsymbol{U}_{(c)}^T\right) - E\left\{E(\Psi_n^T|T)E(\Phi_n|T)^{-1}E(\Psi_n|T)\right\}$. Consequently,

$$\Delta_2 \xrightarrow{d} N(0, \sigma^2\Sigma)$$

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

[2] S.J. Ma, L.J. Yang, R. Romero and Y.H. Cui. Varying coefficient model for gene-environment interaction: a non-linear look. *Bioinformatics* 2012; 27: 2119-2126.

[3] C. Wu and Y.H. Cui. A novel method for identifying nonlinear gene-environment interactions in case-control association studies. (Submitted)

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

[5] A. Maity, R.J. Carrol, E. Mammen and N. Chatterjee. Testing in semiparametric models with interaction, with applications to gene-environment interactions. *J. Roy. Stat. Soc. B* 71:75-96. 2009

[6] Y.H. Cui, G.L. Kang,K.L. Sun,R. Romero, M.P. Qian, W.J. Fu. Gene-centric genomewide association study via entropy. *Genetics* 179: 637-650, 2008

[7] C. Wu and Y.H. Cui. Boosting signals in gene-based association studies via efficient SNP selection. *Briefings in Bioinformatics* (in press) 2013

[8] B. Efron, R. Tibshirani. On testing the significance of sets of genes. *Ann Appl Stat* 1:10729. 2007

[9] D.J. Schaid, J.P. Sinnwell, G.D. Jenkins, et al. Using the gene ontology to scan multilevel gene sets for associations in genome wide association studies. *Genet Epidemiol* 36:316. 2012

[10] K. Wang, M. Li and H. Hakonarson. Analysing biological pathways in genome-wide association studies. *Nat Rev Genet* 11:84354. 2011

[11] T. Hastie and R. Tibshirani. Varying-coefficient models. *J.R.Statist.Soc.B* 1993; 55: 757-796.

[12] L.L. Schumaker. Spline Functions: basic theory. Wiley, New York. 1981

[13] Y.L. Tang, H.X. Wang, Z.Y. Zhu and X.Y. Song. A unified variable selection approach for varying coefficient models. *Statistica Sinica* 2010; 22: 601-628.

[14] J.H. Huang, C. Wu and L. Zhou. Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statistica Sinica* 14: 763-788. 2004

[15] J.Q. Fan and R.Z. Li. Variable selection via nonconcave penzlied likelihood and its oracle properties. *J. Amer. Stat. Assoc.* 2001; 96: 1348-1360.

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

[17] M.O. Kim. Quantile regression with varying coefficients. *Ann. Stat.* 35: 92-108. 2007

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

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

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

[21] M.C. Cornelis, E.J. Tchetgen, L. Liang, L. Qi, N. Chatterjee, F.B. Hu and P. Kraft. Gene-environment interactions in genome-wide association studies: A comparative study of tests applied to empirical studies of type 2 diabetes. *Am. J. of Epidemiology* 175 (3): 191-202. 2011