Published in final edited form as:

Stat Med. 2014 December 10; 33(28): 4988–4998. doi:10.1002/sim.6287.

Integrative Analysis of Gene-Environment Interactions under a Multi-response Partially Linear Varying Coefficient Model

Cen Wu^a, Yuehua Cui^b, and Shuangge Ma^{a,c,d,*,†}

^aDepartment of Biostatistics, School of Public Health, Yale University, 60 College Street, New Haven, CT 06520

^bDepartment of Statistics and Probability, Michigan State University, 619 Red Cedar Rd, East Lansing, MI 48824

^cVA Cooperative Studies Program Coordinating Center, West Haven, CT 06516

^dData Mining Research Center, School of Economics, Xiamen University, Xiamen, China

Abstract

Consider the integrative analysis of genetic data with multiple correlated response variables. The goal is to identify important gene-environment (G×E) interactions along with main gene and environment effects that are associated with the responses. The homogeneity and heterogeneity models can be adopted to describe the genetic basis of multiple responses. To accommodate possible nonlinear effects of some environment effects, a multi-response partially linear varying coefficient (MPLVC) model is assumed. Penalization is adopted for marker selection. The proposed penalization method can select genetic variants with G×E interactions, no G×E interactions, and no main effects simultaneously. It adopts different penalties to accommodate the homogeneity and heterogeneity models. The proposed method can be effectively computed using a coordinate descent algorithm. Simulation study and the analysis of Health Professional's Follow-up Study (HPFS), which has two correlated continuous traits, SNP measurements, and multiple environment effects, show superior performance of the proposed method over its competitors.

Keywords

Integrative analysis; Gene-environment interactions; Multi-response partially linear varying coefficient model; Marker selection

1. Introduction

This study has been partly motivated by the analysis of the Health Professionals Follow-up Study (HPFS) and studies alike. HPFS was launched in 1986 as part of the Gene Environment Association Studies Initiative (GENEVA), in order to assess a series of hypotheses on the predisposition of complex diseases in male, such as vascular diseases,

type 2 diabetes, and cancer [1]. Two continuous traits measured in the HPFS are body mass index (BMI) and weight, both of which can quantify the obesity level, have been extensively used in epidemiologic studies, but have different implications, and cannot replace each other. The two traits are correlated, as has been observed in a large number of studies, suggesting that they may share common genetic basis. On the other hand, they do not have a perfect correlation and should satisfy different data generating models. For the two traits, both “G” factors (SNPs), “E” factors (age and family history of diabetes among first degree relatives – FAMDB), and their interactions may have contributions.

The first challenge comes from the need to analyze multiple correlated response variables. Recently, the joint analysis of multiple correlated responses has been recognized as a competitive alternative to individual-response analysis to enhance statistical power [2, 3], which is mainly due to pleiotropy, the phenomenon that correlated traits share common genetic basis [4]. For example, in the genome-wide association studies (GWAS) of schizophrenia and cardiovascular disease, the pleiotropy-informed approach leads to a significant improvement in the discovery of schizophrenia susceptible loci by taking advantage of the two pleiotropic phenotypes [5]. Moreover, the integrative analysis of five correlated psychiatric disorders in a recent GWAS demonstrates superiority over the individual-disorder analysis (Psychiatric Genomics Consortium [6]). Converging evidence has suggested the particular strength gained by taking the pleiotropy effect into consideration.

With a large number of SNPs, most of which are not associated with the responses, marker selection is needed. The genetic basis of multiple correlated responses can be described using the homogeneity or heterogeneity models [7]. Under the homogeneity model, the response-associated genetic variants have consistent penetrance effects across all responses. For example, if a variant has $G \times E$ interactions (or no interaction) for one response, then it also has $G \times E$ interactions (or no interaction) for the other responses. Thus under the homogeneity model, only one level of selection is required. That is, we only need to determine whether a variant is associated with any response at all. In contrast, the set of susceptible variants may not be the same for different responses under the heterogeneity model, and the response-associated variants can have distinct penetrance effects across responses. Thus two-level selection is required. For a variant, we need to determine if it is associated with any response at all. In addition, for an important variant, we need to determine which response(s) it is associated with. With different marker selection goals, the homogeneity and heterogeneity models demand different methods.

For complex diseases and traits, $G \times E$ interactions can have important implications beyond the additive gene and environment effects. In the literature, there are multiple approaches for accommodating $G \times E$ interactions. We adopt the statistical modeling approach, where the $G \times E$ interactions are described using product terms in statistical models. A lot of methods have been developed to detect important $G \times E$ interactions. For a review, we refer to [8, 9] and others. However, most of the existing methods assume linear G and E effects. A few recent studies have shown that such an assumption not necessarily holds. The varying coefficient (VC) model framework [10] makes it possible to study the nonlinear effects of

continuous E (or G) factors and has been considered for both continuous responses [11] and binary disease phenotypes [12].

In this study, we conduct the integrative analysis of data with multiple correlated responses. As the G factors are categorical (SNPs), we only consider possible nonlinear effects of E factors. With both continuous and categorical E factors, a partially linear VC model is considered. To further accommodate multiple responses, a multi-response partially linear VC (MPLVC) model is adopted. In marker selection, we need to determine whether a genetic variant has G×E interactions, or a main effect but no interaction, or no effect at all. For marker selection, we adopt the penalization technique, which has been used in quite a few recent studies. For example, the hiernet method [13] adopts Lasso penalties, respects the “main effects, interactions” hierarchical structure, and has attracted a lot of attention. Liu *et al.* [14] develop a “group MCP+MCP” penalty to identify important interactions. Shi *et al.* [15] propose a smoothed penalized rank estimation based method for marginal analysis of gene effects, which is robust to diverse types of error distributions. The aforementioned and most other existing penalization methods are limited by analyzing a single response variable and by assuming additive effects.

Some of the recent publications are relevant to this study, although their methods are not directly applicable to our problem. Liu *et al.* [16] propose a two-stage feature selection method for ultrahigh-dimensional VC models and select SNPs with age-dependent effects on BMI. Li *et al.* [17] develop a Bayesian group Lasso approach for time varying model to identify SNPs with significant effects on BMI, when the effects can be viewed as functions of age. Lian and Ma [18] explore G×E interactions via a semiparametric dimension reduction and variable selection method in multivariate VC models.

This study differs from the published ones along the following aspects. First, the integrative analysis of multiple correlated responses is conducted, which differs from the individual-response analysis. Integrative analysis can borrow information across responses to increase power. Second, partially linear modeling of the E effects is considered, allowing more flexibility than linear modeling. Third, the proposed penalization method can simultaneously determine whether a SNP has interactions, or main effect only, or no effect at all. It can be more powerful and more convenient than some of the existing methods. The rest of the paper is organized as follows. In Section 2, we formulate the MPLVC model and develop the computational algorithm. The convergence of the algorithm and selection of tuning parameters are also discussed. We carry out simulation studies to demonstrate the utility of our method in Section 3. A case study of HPFS data is conducted in Section 4. This paper concludes with discussion in Section 5. Some numerical results are provided in Appendix.

2. Statistical Methods

2.1. Multi-response Partially Linear Varying Coefficient Model

Use subscript i to denote the i th subject. Let (X_i, Y_i, Z_i, E_i) , $i = 1, \dots, n$ be independent and identically distributed (i.i.d.) random vectors. $Y_i = (Y_i^1, \dots, Y_i^D)^T$ is the D -dimensional response vector. In the HPFS data, $D = 2$, and the response variables are BMI and weight,

both of which are measured at a single time point, i.e., cross-sectional data. X_i is the $(p + 1)$ -dimensional vector of G factors (SNPs). Its first component is 1. E_i is the discrete environment factor, which is FAMDB, and Z_i is the continuous environment factor, which is age in our data analysis. For simplicity of notation, we consider only two environment factors. Extension to multiple environment factors is trivial. Suppose that the d th ($d = 1, \dots, D$) response variable is associated with X , Z , and E through the PLVC models

$$Y_i^d = \sum_{j=0}^p \beta_j^d(Z_i) X_{ij} + \sum_{j=0}^p \delta_j^d X_{ij} E_i + \varepsilon_i^d. \quad (1)$$

$\beta_j^d(\cdot)$ is a smooth VC function. X_{ij} is the j th component of X_i . ε_i^d is the random error. Here as the response variables are continuous, linear regression models are assumed. Identification of important G×E interactions amounts to conducting variable selection and parameter estimation for model (1), which can be achieved through a penalized polynomial method.

2.2. Basis expansion and penalized estimation

Suppose that the smooth coefficient function $\beta_j^d(\cdot)$ in (1) can be approximated by basis expansion

$$\beta_j^d(Z_i) \approx \sum_{l=1}^L \tilde{\gamma}_{jl}^d \tilde{B}_{jl}^d(Z_i) = (\tilde{\gamma}_j^d)^T \tilde{B}_j^d(Z_i),$$

where L is the number of basis functions, $\tilde{\gamma}_j^d = (\tilde{\gamma}_{j1}^d, \dots, \tilde{\gamma}_{jL}^d)^T$ is the spline coefficient vector, and $\tilde{B}_j^d(Z_i) = (\tilde{B}_{j1}^d(Z_i), \dots, \tilde{B}_{jL}^d(Z_i))^T$ is the set of normalized B-spline basis. Rewrite the above basis expansion as

$$\beta_j^d(Z_i) \approx \sum_{l=1}^L \tilde{\gamma}_{jl}^d \tilde{B}_{jl}^d(Z_i) \doteq \gamma_{j1}^d + (\gamma_{j*}^d)^T \tilde{B}_{j*}^d(Z_i),$$

where $\tilde{B}_{j*}^d(Z_i) = (\tilde{B}_{j2}^d(Z_i), \dots, \tilde{B}_{jL}^d(Z_i))^T \cdot \gamma_{j1}^d$ and $\gamma_{j*}^d = (\gamma_{j2}^d, \dots, \gamma_{jL}^d)^T$ correspond to the constant and varying parts of the coefficient functional respectively. Define

$\gamma_j^d = (\gamma_{j1}^d, (\gamma_{j*}^d)^T)^T$ and $B_j^d(Z_i) = (1, (\tilde{B}_{j*}^d(Z_i))^T)^T \doteq (B_{j1}^d(Z_i), \dots, B_{jL}^d(Z_i))^T$. With the basis expansion, model (1) changes to

$$\begin{aligned} Y_i^d &= \sum_{j=0}^p \sum_{l=1}^L \gamma_{jl}^d B_{jl}^d(Z_i) X_{ij} + \sum_{j=0}^p \delta_j^d X_{ij} E_i + \varepsilon_i^d \\ &= \sum_{j=0}^p (\gamma_j^d)^T W_{ij}^d + \sum_{j=0}^p \delta_j^d X_{ij} E_i + \varepsilon_i^d, \end{aligned} \quad (2)$$

where $W_{ij}^d = B_j^d(Z_i) X_{ij} = \left(B_{j1}^d(Z_i) X_{ij}, \dots, B_{jL}^d(Z_i) X_{ij} \right)^T$ for $d = 1, \dots, D$. Let $U_i = (U_{i0}, \dots, U_{ip})$, where $U_{ij} = \text{diag} \left\{ \left(W_{ij}^1, X_{ij} E_i \right), \dots, \left(W_{ij}^D, X_{ij} E_i \right) \right\}$ ($i = 1, \dots, n$ and $j = 0, \dots, p$). The coefficient vector is $\alpha = \left(\alpha_0^T, \dots, \alpha_p^T \right)^T$ with $\alpha_j = \left(\alpha_j^1, \dots, \alpha_j^D \right)^T$ and $\alpha_j^d = \left(\left(\gamma_j^d \right)^T, \delta_j^d \right)^T$ for $j = 0, \dots, p$ and $d = 1, \dots, D$.

The penalized least square objective function is

$$\begin{aligned}
& \frac{1}{2n} \sum_{i=1}^n [Y_i - U_i \alpha]^T \Sigma^{-1} [Y_i - U_i \alpha] + P_\lambda(\alpha) \\
= & \frac{1}{2n} \|\tilde{Y} - \sum_{j=0}^p \tilde{U}_j \tilde{\alpha}_j\|^2 + P_\lambda(\alpha) = L(\alpha) + P_\lambda(\alpha), \quad (3)
\end{aligned}$$

where $\tilde{Y} = (\tilde{Y}_1^T, \dots, \tilde{Y}_n^T)^T$, $\tilde{Y}_i = \sum^{-\frac{1}{2}} Y_i$, $\tilde{U}_{ij} = \sum^{-\frac{1}{2}} U_{ij}$ with $\tilde{U}_j = (\tilde{U}_{1j}, \dots, \tilde{U}_{nj})^T$, and $\tilde{\alpha}_j = \sum^{-\frac{1}{2}} \alpha_j$ ($i = 1, \dots, n; j = 0, \dots, p$). is the residual covariance matrix defined as $dd = \sigma_d$ and $dk = \sigma_{dk}(d \quad k)$ for $d, k = 1, \dots, D$. $P_\lambda(\alpha)$ is the penalty function, which takes different forms under the homogeneity and heterogeneity models.

Homogeneity model—First we consider the penalized estimate

$$\hat{\alpha} = \underset{\alpha}{\operatorname{argmin}} \left\{ L(\alpha) + \lambda \sum_{j=0}^p \|\alpha_j\|_{\Sigma_j} \right\}. \quad (4)$$

Denote $a_{j,1}$, a_{j*} , and $a_{j,(L+1)}$ as the components of a_j corresponding to the constant and varying parts of the coefficient functional and the discrete environment factor, respectively.

$\tilde{\alpha}_{j,1}$, $\tilde{\alpha}_{j*}$, and $\tilde{\alpha}_{j,(L+1)}$ can be defined consequently.

$$\|\alpha_{j,1}\|_{\Sigma_{j,1}} = (\alpha_{j,1}^T \Sigma_{j,1} \alpha_{j,1})^{1/2}, \|\alpha_{j*}\|_{\Sigma_{j*}} = (\alpha_{j*}^T \Sigma_{j*} \alpha_{j*})^{1/2}, \text{ and}$$

$$\|\alpha_{j,(L+1)}\|_{\Sigma_{j,(L+1)}} = (\alpha_{j,(L+1)}^T \Sigma_{j,(L+1)} \alpha_{j,(L+1)})^{1/2}, \text{ with } \Sigma_{j,1} = n^{-1} \tilde{U}_{j,1}^T \tilde{U}_{j,1}, \Sigma_{j*} = n^{-1} \tilde{U}_{j*}^T \tilde{U}_{j*},$$

and $\Sigma_{j,(L+1)} = n^{-1} \tilde{U}_{j,(L+1)}^T \tilde{U}_{j,(L+1)}$, respectively. Here $\tilde{U}_{j,1}$, $\tilde{U}_{j*}$ and $\tilde{U}_{j,(L+1)}$ are the components of $\tilde{U}_j$ corresponding to $\tilde{\alpha}_{j,1}$, $\tilde{\alpha}_{j*}$, and $\tilde{\alpha}_{j,(L+1)}$ respectively. $\hat{\alpha}$ defined in (4) is a group Lasso type estimate. Denote $\hat{\alpha}_{j,1}$, $\hat{\alpha}_{j*}$, and $\hat{\alpha}_{j,(L+1)}$ ($j = 0, \dots, p$) as the corresponding components of the estimate.

Under the homogeneity model, the same set of susceptible SNPs is associated with multiple responses. Such an assumption is sensible when common genetic basis is shared among the responses. Consider the penalized objective function

$$Q(\alpha) = L(\alpha) + \lambda \sum_{j=0}^p \left[w_{j,1} \|\alpha_{j,1}\|_{\Sigma_{j,1}} + w_{j*} \|\alpha_{j*}\|_{\Sigma_{j*}} + w_{j,(L+1)} \|\alpha_{j,(L+1)}\|_{\Sigma_{j,(L+1)}} \right], \quad (5)$$

where the weights are $w_{j,1} = \|\hat{\alpha}_{j,1}\|_{\Sigma_{j,1}}^{-1}$, $w_{j*} = (L-1) \|\hat{\alpha}_{j*}\|_{\Sigma_{j*}}^{-1}$, and $w_{j,(L+1)} = \|\hat{\alpha}_{j,(L+1)}\|_{\Sigma_{j,(L+1)}}^{-1}$. The development of penalty in (5) has been driven by the following considerations. To separate the nonzero constant and varying effect at a SNP locus, we penalize $\|\alpha_{j*}\|_{j*}$, the norm of spline coefficients corresponding to the varying parts for D responses, as a group. If $\|\alpha_{j*}\|_{j*} = 0$, the j th SNP only has a constant (or main) effect. Moreover, if $\|\alpha_{j,1}\|_{j,1} > 0$, then this SNP has no main effect. Similarly, whether the interaction between the SNP and discrete environment factor E exists can be determined by if $\|\alpha_{j,(L+1)}\|_{j,(L+1)}$ is shrunk to 0 or not. Here the coefficients $\alpha_{j,1}$, α_{j*} , and $\alpha_{j,(L+1)}$ of all the responses are treated as a group, respectively. In particular, each member in α_{j*} is also a group of spline coefficients, corresponding to the varying part for one of the D responses. Once a group is selected, all members within the group are selected with nonzero estimates. Otherwise, all the group members are estimated as 0, and the goal of both variable selection and separation of the varying/constant effects is achieved.

Heterogeneity model—First consider the penalized estimate

$$\hat{\alpha} = \underset{\alpha}{\operatorname{argmin}} \left\{ L(\alpha) + \lambda \sum_{j=0}^p \sum_{d=1}^D \|\alpha_j^d\|_{\Sigma_j^d} \right\}. \quad (6)$$

The components in α_j^d corresponding to the constant and varying parts of the coefficient functional and the discrete environment factor are denoted as $\alpha_{j,1}^d$, α_{j*}^d , and $\alpha_{j,(L+1)}^d$, respectively. Then $\tilde{\alpha}_{j,1}^d$, $\tilde{\alpha}_{j*}^d$, and $\tilde{\alpha}_{j,(L+1)}^d$ can be defined accordingly. Denote $\tilde{\alpha}_{j,1}^d$, $\tilde{\alpha}_{j*}^d$, and $\tilde{\alpha}_{j,(L+1)}^d$ ($j = 0, \dots, p$) as the corresponding components of the estimate.

Under the heterogeneity model, a SNP can be associated with some but not all the responses. The penalized objective function is

$$Q(\alpha) = L(\alpha) + \lambda \sum_{j=0}^p \sum_{d=1}^D \left[w_{j,1}^d \|\alpha_{j,1}^d\|_{\Sigma_{j,1}^d} + w_{j*}^d \|\alpha_{j*}^d\|_{\Sigma_{j*}^d} + w_{j,(L+1)}^d \|\alpha_{j,(L+1)}^d\|_{\Sigma_{j,(L+1)}^d} \right], \quad (7)$$

where $\|\alpha_{j,1}^d\|_{\Sigma_{j,1}^d} = \left((\alpha_{j,1}^d)^T \Sigma_{j,1}^d \alpha_{j,1}^d \right)^{1/2}$, $\|\alpha_{j*}^d\|_{\Sigma_{j*}^d} = \left((\alpha_{j*}^d)^T \Sigma_{j*}^d \alpha_{j*}^d \right)^{1/2}$, and

$\|\alpha_{j,(L+1)}^d\|_{\Sigma_{j,(L+1)}^d} = \left((\alpha_{j,(L+1)}^d)^T \Sigma_{j,(L+1)}^d \alpha_{j,(L+1)}^d \right)^{1/2}$, with

$\Sigma_{j,1}^d = n^{-1} (\tilde{U}_{j,1}^d)^T \tilde{U}_{j,1}^d$, $\Sigma_{j*}^d = n^{-1} (\tilde{U}_{j*}^d)^T \tilde{U}_{j*}^d$, and $\Sigma_{j,(L+1)}^d = n^{-1} (\tilde{U}_{j,(L+1)}^d)^T \tilde{U}_{j,(L+1)}^d$. The components in $\tilde{U}_j^d$ corresponding to $\tilde{\alpha}_{j,1}^d$, $\tilde{\alpha}_{j*}^d$, and $\tilde{\alpha}_{j,(L+1)}^d$ are denoted as $\tilde{U}_{j,1}^d$, $\tilde{U}_{j*}^d$ and

$\tilde{U}_{j,(L+1)}^d$, respectively. The weights are set as $w_{j,1}^d = \|\hat{\alpha}_{j,1}^d\|_{\Sigma_{j,1}^d}^{-1}$, $w_{j*}^d = (L-1) \|\hat{\alpha}_{j*}^d\|_{\Sigma_{j*}^d}^{-1}$ and

$w_{j,(L+1)}^d = \|\hat{\alpha}_{j,(L+1)}^d\|_{\Sigma_{j,(L+1)}^d}^{-1}$.

Here as the D responses may have different sets of susceptibility SNPs, the weights and hence penalties are response-dependent. The homogeneity and heterogeneity models have been assumed in the literature [3, 7]. However, as the present data and model settings are much more complicated than those in the existing studies, new penalization methods are needed. In practical data analysis, as the homogeneity model is a special case of the heterogeneity model, it is “safer” to implement the heterogeneity model first.

2.3. Computational Algorithms

The group coordinate descent (GCD) approach is adopted. In the GCD algorithm, the objective function is minimized with respect to a single predictor group at a time, and the minimization is cycled through all the predictor groups iteratively till convergence.

To improve computational effectiveness, we orthogonalize the predictor groups block-wisely. For the homogeneity model, let $V_{j,1}, V_{j,(L+1)}$ be invertible, and $V_{j,1}, V_{j,*}, V_{j,(L+1)}$ be the lower triangular matrices in the Cholesky decomposition such that

$\Sigma_{j,1} = V_{j,1} V_{j,1}^{-1}$, $\Sigma_{j,*} = V_{j,*} V_{j,*}^{-1}$, and $\Sigma_{j,(L+1)} = V_{j,(L+1)} V_{j,(L+1)}^{-1}$. Let $a_{j,1} = V_{j,1} a_{j,1}$, $a_{j,*} = V_{j,*} a_{j,*}$, and $a_{j,(L+1)} = V_{j,(L+1)} a_{j,(L+1)}$. Denote $T_{j,1} = \tilde{U}_{j,1} V_{j,1}^{-1}$, $T_{j,*} = \tilde{U}_{j,*} V_{j,*}^{-1}$, and $T_{j,(L+1)} = \tilde{U}_{j,(L+1)} V_{j,(L+1)}^{-1}$. Then the penalized objective function (5) changes to 1

$$Q(a) = \frac{1}{2n} \|\tilde{Y} - \sum_{j=0}^p T_j a_j\|^2 + \lambda \sum_{j=0}^p \left[w_{j,1} \|a_{j,1}\| + w_{j,*} \|a_{j,*}\| + w_{j,(L+1)} \|a_{j,(L+1)}\| \right]. \quad (8)$$

Denote $a = (a_0^T, \dots, a_p^T)^T$ with $a_j = (a_{j,1}^T, a_{j,*}^T, a_{j,(L+1)}^T)^T$ for $j = 0, \dots, p$.

$w_{j,1} = \|\hat{a}_{j,1}\|^{-1}$, $w_{j,*} = (L-1) \|\hat{a}_{j,*}\|^{-1}$, and $w_{j,(L+1)} = \|\hat{a}_{j,(L+1)}\|^{-1}$.

For a fixed tuning parameter λ , the GCD algorithm proceeds as follows:

- (1) Initialize $a^{(0)}$, the vector of residuals $r = \tilde{Y} - \sum_{j=0}^p T_j a_j^{(0)}$, and $m = 0$.
- (2) For $j = 0, 1, \dots, p$, iterate the following steps: Denote $a_{j,\text{sub},1} = a_{j,1}^{(m)}$, $a_{j,\text{sub},2} = a_{j,*}^{(m)}$, and $a_{j,\text{sub},3} = a_{j,(L+1)}^{(m)}$. Let $T_{j,\text{sub},i}$ be the submatrix of T_j corresponding to $a_{j,\text{sub},i}$ and $w_{j,\text{sub},i}$ be the corresponding weight calculated as in (5) ($i=1, 2, 3$). For $i = 1, 2, 3$, repeat
 - (a) Calculate $t_{j,\text{sub},i} = n^{-1} T_{j,\text{sub},i}^T r + a_{j,\text{sub},i}^{(m)}$
 - (b) Calculate $a_{j,\text{sub},i}^{(m+1)} = \left(1 - \frac{w_{j,\text{sub},i} \lambda}{\|t_{j,\text{sub},i}\|} \right)_+ t_{j,\text{sub},i}$
 - (c) Update $r = r - T_{j,\text{sub},i} (a_{j,\text{sub},i}^{(m+1)} - a_{j,\text{sub},i}^{(m)})$
 Update $m = m + 1$.
- (3) Iterate step (2) until convergence.

At each cycle of the GCD algorithm, the objective function is minimized with respect to a as defined in (8), and we have the descent property for the sequence $\{a^{(m)} : m = 0, 1, \dots\}$ as follows:

Proposition 1—For $m = 0, 1, \dots$,

$$Q(a^{(m)}) \leq Q(a^{(m-1)}).$$

Moreover, the sequence $\{a^{(m)} : m = 0, 1, \dots\}$ converges to the global minimum of Q .

The proof follows from arguments similar to those in [19] and the application of Theorem 4.1 of [20]. The computational algorithm for the heterogeneity model can be derived similarly due to the close connection between (5) and (7). We can also establish the convergence property like Proposition 1 following similar lines of argument, and the details are omitted here.

2.4. Selection of Tuning Parameters

We need to choose proper tuning parameters O , K_L , and λ , where the degree of the B spline basis O and the number of interior knots K_L govern the smoothness of VC functions, while λ controls sparsity of the estimator.

First, we use BIC to select O and K_L . The computational cost of choosing O and K_L for each VC function separately is prohibitively expensive. Therefore, we take that all the VC functions are approximated with the same O and K_L in basis expansion. The range of K_L is determined by $\{\max([0.5n^{1/(2O+3)}], 1), [1.5n^{1/(2O+3)}]\}$, with $[x]$ denoting the integer part of x . The optimal combination of O and K_L can be found by a two-dimensional grid search for the minimum BIC, according to the following criteria

$$BIC_{O, K_L} = \log(RSS_1/n) + (O + K_L + 1) \log(n)/n,$$

where RSS_1 is the residual sum of squares of the intercept model.

Conditional on the selected O and K_L , we choose λ using the extended Bayesian Information Criteria (EBIC) [21] with a proper effective degree of freedom as follows:

$$EBIC_\lambda = \log(RSS_\lambda/n) + edf(\log n + 2\gamma \log p)/n, \gamma \geq 0,$$

where RSS_λ is the residual sum of squares with a given λ . The effective degree of freedom is defined as the number of predictor groups with non-zero estimates [22].

3. Simulation

In simulation, we assess the proposed method in terms of selection, estimation, and prediction accuracy. Specifically, selection accuracy is evaluated using the number of true/false positives. Estimation accuracy is evaluated using the integrated mean squared error

(IMSE). Prediction performance is evaluated using the mean prediction errors on an independent testing dataset generated under the same settings.

Let $\hat{\beta}_j^{(r)}$ be the estimator of a nonparametric function β_j in the r th ($1 \leq r \leq R$) replicate, and $\{z_m\}_{m=1}^{n_{\text{grid}}}$ be the grid points where $\hat{\beta}_j^{(r)}$ is assessed. The IMSE of $\hat{\beta}_j(z)$, defined as $\text{IMSE}(\hat{\beta}_j(z)) = \frac{1}{R} \sum_{r=1}^R \frac{1}{n_{\text{grid}}} \sum_{m=1}^{n_{\text{grid}}} \left\{ \hat{\beta}_j^{(r)}(z_m) - \beta_j(z_m) \right\}^2$, evaluates the estimation accuracy for β_j . Note that when $\hat{\beta}_j$ is a constant, $\text{IMSE}(\hat{\beta}_j)$ reduces to $\text{MSE}(\hat{\beta}_j)$. We consider the following two simulation examples.

Example 1

We generate the phenotypic responses under the homogeneity model. The PLVC model for the d th ($1 \leq d \leq D$) outcome is $Y_i^d = \sum_{j=0}^p \beta_j^d(Z_i) X_{ij} + \sum_{j=0}^p \delta_j^d E_i X_{ij} + \varepsilon_i^d$. Consider the bivariate case ($D=2$) which is consistent with our data analysis. With $n=200$ and $p=500$, two response variables are simulated with correlation coefficient 0.4 and 0.8. The two random errors are simulated from a multivariate normal distribution with mean vector 0 and covariance $\begin{pmatrix} 2.25 & 0.90 \\ 0.90 & 2.25 \end{pmatrix}$. We generate the continuous environment variable $Z_i \sim \text{Unif}(0,1)$, and the discrete environment variable E_i from a binomial distribution.

X_j 's are simulated as SNPs coded with 3 categories (2, 1, 0) for genotypes (AA, Aa, aa) respectively. The SNP genotype data are simulated based on a pairwise linkage disequilibrium (LD) structure. Let q_A and q_B be the minor allele frequencies (MAFs) of two alleles A and B for two adjacent SNPs. Then with LD denoted as η , the frequencies of four haplotypes are $p_{AB} = q_A q_B + \eta$, $p_{Ab} = q_A(1 - q_B) - \eta$, $p_{aB} = (1 - q_A)q_B - \eta$, and $p_{ab} = (1 - q_A)(1 - q_B) + \eta$. We generate the SNP genotype (AA, Aa, aa) at locus 1 from a multinomial distribution with frequencies $(q_A^2, 2q_A(1 - q_A), (1 - q_A)^2)$, under the assumption of Hardy-Weinberg equilibrium. Then the genotypes for locus 2 can be simulated based on the conditional probability [23]. We set MAFs as 0.3 and pairwise correlation $r = 0.5$, which results in $\eta = r \sqrt{q_A(1 - q_A)q_B(1 - q_B)}$.

The coefficients are set as: $\beta_0^1(z) = 2\sin(2\pi z)$, $\beta_1^1(z) = 2\exp(2z - 1)$, $\beta_1^2(z) = 0.8$, $\delta_0^1 = 0.6$ and $\delta_1^1 = 0.6$; $\beta_0^2(z) = 9(z - 0.3)^2 - 1.11$, $\beta_1^2(z) = -4z^3$, $\beta_2^2(z) = -0.6$, $\delta_0^2 = 0.7$ and $\delta_2^2 = -0.7$. All the rest coefficients are set as 0.

The estimation results are summarized in Table 1. Here ALASSO denotes the proposed method. ALASSO_M is the marginal version of the ALASSO method, where each response is analyzed individually under a PLVC model. LASSO_N serves as the benchmark under a full parametric model where the responses are analyzed with Lasso separately. For the alternatives, the tuning parameters are selected by EBIC. The results of the oracle model, which is under the true data-generating model, are also provided. We observe that ALASSO

has smaller prediction errors and estimation errors on regression coefficients than $ALASSO_M$. For instance, under correlation coefficient 0.4, the mean prediction errors for $ALASSO$ and $ALASSO_M$ are 4.33 (sd 0.41) and 5.56 (sd 0.45) respectively, while that for $LASSO_N$ is 10.39 (sd 1.32). $LASSO_N$ cannot accommodate the nonparametric effects of the environment factor, and thus has the worst estimation accuracy on the varying coefficients and the largest prediction errors. The identification results in Table 2 indicate that compared with the marginal analysis, the proposed method can identify more true positives, though occasionally at the cost of slightly more false positives.

Figure 1 and 2 in Appendix show the averages of estimated varying coefficients with correlation coefficient 0.4 and 0.8, respectively. The upper and lower panels of both figures show the estimated coefficient curves of SNPs associated with the first and second outcomes, correspondingly. Note that the dimension of regression coefficients to be estimated after basis expansion is much larger than the sample size ($n=200$). For example, if the optimal $O=2$ and $K_L=2$, then the effective dimension of regression coefficients is 3006. The plots indicate that the average estimates match the true coefficient functions well, and the curves of ± 2 pointwise standard deviations cover the true coefficient functions.

Example 2

The phenotypic responses are generated under the heterogeneity model with coefficients

$$\beta_0^1(z) = 2\sin(2\pi z), \beta_1^1(z) = 0.8, \beta_2^1(z) = 2\exp(2z - 1), \delta_0^1 = 0.6 \text{ and } \delta_3^1 = 0.6; \beta_0^2(z) = 9(z - 0.3)^2 - 1.11, \beta_1^2(z) = -1.6z^3, \beta_4^2(z) = -0.6, \delta_0^2 = 0.7 \text{ and } \delta_5^2 = -0.7. \text{ All the other coefficients are set as 0. The random errors are simulated from a multivariate}$$

normal distribution with mean vector 0 and covariance $\begin{pmatrix} 2.25 & 1.80 \\ 1.80 & 2.25 \end{pmatrix}$ Tables 5 and 6 in Appendix show the results of parameter estimation and identification. Under correlation coefficient 0.4, the proposed method has prediction error 4.40 (sd 0.36), which is closer to that of the oracle model (mean prediction error 4.22, sd 0.31), compared to the alternatives. Our method identifies more true positives as shown in Table 6 (Appendix). The mean curves of estimated varying coefficients under correlation coefficient 0.4 and 0.8 are in Figure 3 and 4 in Appendix. Again, the true coefficient functions are well approximated by the mean estimates, and are enveloped by the ± 2 pointwise standard deviations. Overall, the simulations demonstrate that the proposed method has satisfactory performance and outperforms alternatives in terms of variable selection, parameter estimation, and prediction.

Although the proposed method has been motivated by data with correlated responses, to be thorough, we also conduct simulation with uncorrelated responses. Details are provided in Appendix, which also show better performance of the proposed method over the alternatives. In addition, we also conduct sensitivity analysis on how the smoothness specification of O and K_L affects variable selection. Results in Appendix show that performance of the proposed method is relatively insensitive to the smoothness specification, as long as K_L is not too small.

4. Analysis of the HPFS data

As described in Section 1, we analyze two correlated traits: BMI and weight. We focus on SNPs on chromosome 4, which is found to harbor “interesting” SNPs in preliminary analysis. After data cleaning by keeping subjects with matched phenotypes and genotypes, removing SNPs with $MAF < 0.05$ or deviation from Hardy-Weinberg equilibrium, the working dataset contains 2568 subjects with 40,568 SNPs. To ensure the feasibility of joint analysis on the two correlated responses, we impute missing data using fastPHASE [24].

As the obesity level varies with age [25], we treat age as a continuous environment factor. Besides, the binary indicator of family history of diabetes among first degree relatives (FAMDB) is a sensible candidate for a discrete environment factor. For computational convenience, prescreening can be conducted so that variable selection on the reduced feature space is more attainable. For example, in [17], the single SNP analysis is conducted to filter SNPs in a GWA study before downstream analysis. Here we implement the procedure described in [11] to screen SNPs. Specifically in [11], three statistical tests are proposed to evaluate if the penetrance effect of a variant under the environmental stimuli is non-linear, linear, constant, or zero. The SNPs with p values less than a certain cutoff (for instance, 0.005) from any of the tests are kept under either BMI or weight. Besides, we also keep SNPs with p values less than the cutoff in single SNP analysis with respect to FAMDB. 388 SNPs pass the screening.

The numbers of identified SNPs and overlapping SNPs among all the methods under comparison are summarized in Table 3. $LASSO_N$, the LASSO method on separate responses, only shares a limited number of commonly selected SNPs with the other methods, which is not surprising since LASSO tends to randomly select one from correlated SNPs, and we find that some highly correlated SNPs are identified by $LASSO_N$ and the other methods respectively, for instance, rs2616382 ($LASSO_N$) and rs2616443 (heterogeneity model), or rs1502744 ($LASSO_N$) and rs2221312 (both homogeneity and heterogeneity models). Table 4 provides details on SNPs identified by the proposed method. Under the homogeneity model, two SNPs (rs2293595 and rs905902) are selected due to their interactions with age under both responses. Two other SNPs (rs1502744 and rs2221312) are selected with no $G \times E$ interactions (main effects only). Under the heterogeneity model, we identify three SNPs interacting with age and four with only constant effects. Table 7 and 8 in Appendix show the results of $ALASSO_M$ and $LASSO_N$, correspondingly. None of the interactions between SNPs and FAMDB are identified.

It is difficult to objectively evaluate the selection performance with real data. We evaluate the prediction performance, which may provide partial information on the relative performance of different methods. It is noted that selection and prediction are two different analysis goals. The prediction evaluation is conducted to complement selection analysis. Following [26], for each method, we refit the selected model without penalty and obtain the prediction mean squared errors (PMSE) as 192.37, 193.62, 193.72 and 194.28 for the proposed method under the homogeneity and heterogeneity models, $ALASSO_M$, and $LASSO_N$, respectively. The proposed method under the homogeneity model slightly outperforms its competitors. Similar results have been observed in the literature [3]. By

conducting integrative analysis, the proposed method can outperform the alternatives. With the two highly correlated responses (which are likely to share common genetic basis), the homogeneity model can be sensible and may outperform the heterogeneity model.

Under the homogeneity model, the proposed method identifies the main effect of age, as well as two SNPs, rs2293595 and rs905902, interacting with age. We plot the estimated varying coefficients along with the pointwise bootstrap standard deviations in Figure 5 (Appendix). The plot clearly shows the nonlinear genetic effects corresponding to the two SNPs across age. Another observation is that the shapes of coefficient functions of one SNP on the two responses are similar and mainly differ in magnitude.

5. Discussion

This study has conducted the integrative analysis of multiple correlated responses and advances from the individual-response analysis by borrowing information across responses. Different from the existing integrative analysis, this study has focused on the $G \times E$ interactions. Especially, it accommodates nonlinear effects of E factors, allowing for more flexibility. A Lasso-based penalization method has been implemented. It uses a group Lasso-type estimate as the initial estimate and then conducts a weighted group Lasso-type estimation. It has roots in [27] but significantly advances from the existing studies by analyzing much more complicated data and models. Even though the proposed method may seem conceptually simple, this is the very first time such a method is implemented under the present data and model settings. Lasso-type estimation is adopted mainly because of their computational simplicity. In addition, under simpler settings, it has been proved that such a “Lasso + weighted Lasso” strategy may achieve consistency. We conjecture that other penalties can be used to replace the Lasso-type penalties. Simulation studies show that the proposed method can outperform alternatives in marker selection, estimation, and prediction. In the analysis of HPFS data, the proposed method suggests the possible nonlinear effects of age interacting with SNPs. The identified SNPs differ from those using the alternatives.

The responses in the HPFS are continuous. Thus we have assumed the linear regression models. With other types of response variables, we can use the negative log-likelihood as the loss function and then adopt the proposed penalties. Coupled with the MM (Majorize-minimization) approach or Taylor expansion, the proposed algorithm is potentially applicable. The proposed method involves the correlation coefficients between the residuals. In simulation we have found that using the correlation coefficients between the responses provides a reasonable approximation. A more rigorous approach is to first ignore the correlation and estimate the regression coefficients. Here the estimates can be consistent although lack efficiency. Then the estimates can be used to estimate the residuals and their correlations. This procedure can be iterated multiple times. We have focused on methodological development. Theoretical study is postponed to the future.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank the Associate Editor and three reviewers for careful review and insightful comments, which have led to a significant improvement of this article. The authors have been supported by awards CA142774, CA165923 and CA182984 from NCI/NIH, DMS-1209112 from NSF, the VA Cooperative Studies Program of the Department of Veterans Affairs, Office of Research and Development, 13CTJ001 and 13&ZD148 from National Social Science Foundation of China, and 31371336 from National Natural Science Foundation of China. Funding support for the GWAS of Gene and Environment Initiatives in Type 2 Diabetes was provided through the NIH Genes, Environment and Health Initiative [GEI] (U01HG004399). The datasets used for the analyses described in this manuscript were obtained from dbGaP through accession number phs000091.v2.p1.

References

1. Cornelis MC, Agrawal A, Cole JW, Hansel NN, Barnes KC, et al. The Gene, Environment Association Studies consortium (GENEVA): maximizing the knowledge obtained from GWAS by collaboration across studies of multiple conditions. *Genet Epidemiol.* 2010; 34(4):364–372. [PubMed: 20091798]
2. Korte A, Vilhjalmsdottir BJ, Segura V, Platt A, Long Q, Nordborg M. A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nat. Genet.* 2012; 44(9):1066–1071. [PubMed: 22902788]
3. Liu J, Huang J, Ma S. Analysis of genome-wide association studies with multiple outcomes using penalization. *PLoS ONE.* 2012; 7(12):e51198. doi:10.1371/journal.pone.0051198. [PubMed: 23272092]
4. Solovieff N, Cotsapas C, Lee PH, Purcell SM, Smoller JW. Pleiotropy in complex traits: challenges and strategies. *Nat. Rev. Genet.* 2013; 14(7):483–495. [PubMed: 23752797]
5. Andreassen OA, Djurovic S, Thompson WK, Schork AJ, et al. Improved detection of common variants associated with schizophrenia by leveraging pleiotropy with cardiovascular-disease risk factors. *Am. J. Hum. Genet.* 2013; 92(2):197–209. [PubMed: 23375658]
6. Psychiatric Genomics Consortium. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *The Lancet.* 2013; 381:1371–1379.
7. Liu J, Huang J, Ma S. Integrative analysis of multiple cancer genomic datasets under the heterogeneity model. *Statistics in Medicine.* 2013; 32:3509–3521.
8. Sokolowski, MB. Gene-Environment Interplay. Vol. Volume 77 (Advances in Genetics). Academic Press; 2012.
9. Schaefer, GB.; Thompson, JN. Medical Genetics McGraw-Hill Professional. 2013.
10. Hastie T, Tibshirani R. Varying-coefficient models. *J. R. Statist. Soc. B.* 1993; 55:757–796.
11. Ma SJ, Yang LJ, Romero R, Cui YH. Varying coefficient model for gene-environment interaction: a non-linear look. *Bioinformatics.* 2011; 27:2119–2126. [PubMed: 21690105]
12. Wu C, Cui YH. A novel method for identifying nonlinear gene-environment interactions in case-control association studies. *Human Genetics.* 2013; 132(12):1413–1425. [PubMed: 23974428]
13. Bien J, Taylor J, Tibshirani R. A Lasso for hierarchical interactions. *Annals of Statistics.* 2013; 41:1111–1141.
14. Liu J, Huang J, Zhang Y, Lan Q, Rothman N, Zheng T, Ma S. Identification of gene-environment interactions in cancer studies using penalization. *Genomics.* 2013; 102(4):189–194. [PubMed: 23994599]
15. Shi XJ, Liu J, Huang J, Zhou Y, Xie Y, Ma SG. A penalized robust method for identifying gene-environment interactions. *Genetic Epidemiology.* 2013; 38(3):220–230. [PubMed: 24616063]
16. Liu JY, Li RZ, Wu RL. Feature selection for varying coefficient models with ultrahigh dimensional covariates. *J. Amer. Stat. Assoc.* 109(505):266–274. [PubMed: 24678135]
17. Li, JH.; Wang, Z.; Li, RZ.; Wu, RL. Bayesian group LASSO for nonparametric varying coefficient models. 2013. (Under review) http://www.personal.psu.edu/jzl185/jonathan/paper/bayesian_lasso_fi.pdf
18. Lian, H.; Ma, S. Reduced-rank regression in sparse multivariate varying-coefficient models with high-dimensional covariates. 2013. (Under review) <http://arxiv.org/pdf/1309.6058v1.pdf>

19. Breheny P, Huang J. Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Statistics and Computing*. 2014 (To appear).
20. Tseng P. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*. 2001; 109:475–494.
21. Chen J, Chen Z. Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*. 2008; 95(3):759–771.
22. Wang HS, Xia YC. Shrinkage estimation of the varying coefficient model. *J. Amer. Stat. Assoc.* 2008; 104:747–757.
23. Cui YH, Kang GL, Sun KL, et al. Gene-centric genomewide association study via entropy. *Genetics*. 2008; 179:637–650. [PubMed: 18458106]
24. Scheet P, Stephens M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.* 2006; 78(4):629–644. [PubMed: 16532393]
25. Qi L, Cho YA. Gene-environment interaction and obesity. *Nutr Rev.* 2008; 66(12):684–694. [PubMed: 19019037]
26. Yan J, Huang J. Model selection for Cox models with time-varying coefficients. *Biometrics*. 2012; 68:419–428. [PubMed: 22506825]
27. Huang J, Ma S. Variable selection in the accelerated failure time model via the bridge method. *Lifetime Data Analysis*. 2010; 16:176–195. [PubMed: 20013308]

Table 1

Simulation results under the homogeneity model: mean(sd) based on 100 replicates, with $\rho=0.4$ and 0.8 , $n=200$, and $p=500$.

	$\rho=0.4$					$\rho=0.8$				
	ALASSO	ALASSO _M	LASSO _N	Oracle	ALASSO	ALASSO _M	LASSO _N	Oracle	ALASSO	ALASSO _M
IMSE										
$\beta_0^1(z)$	0.02(0.05)	0.03(0.07)	0.82(1.69)	0.02(0.05)	0.02(0.05)	0.03(0.07)	0.84(1.73)	0.02(0.05)	0.03(0.07)	0.03(0.07)
$\beta_1^1(z)$	0.01(0.03)	0.02(0.05)	1.00(2.06)	0.01(0.03)	0.01(0.03)	0.02(0.04)	1.00(2.07)	0.01(0.04)	0.02(0.04)	0.02(0.04)
$\beta_2^1(z)$	0.00(0.02)	0.02(0.06)	0.00(0.00)	0.01(0.02)	0.00(0.02)	0.02(0.07)	0.00(0.00)	0.02(0.02)	0.02(0.07)	0.02(0.07)
$\beta_0^2(z)$	0.01(0.03)	0.03(0.07)	0.56(1.15)	0.01(0.03)	0.01(0.03)	0.03(0.10)	0.56(1.16)	0.01(0.03)	0.03(0.10)	0.03(0.10)
$\beta_1^2(z)$	0.01(0.03)	0.02(0.05)	0.40(0.83)	0.01(0.03)	0.01(0.03)	0.02(0.06)	0.40(0.82)	0.01(0.03)	0.02(0.06)	0.02(0.06)
$\beta_2^2(z)$	0.00(0.01)	0.03(0.08)	0.00(0.00)	0.01(0.02)	0.00(0.01)	0.03(0.09)	0.00(0.00)	0.02(0.02)	0.03(0.09)	0.03(0.09)
MSE										
δ_0^1	0.01(0.04)	0.06(0.12)	0.07(0.14)	0.01(0.02)	0.02(0.06)	0.06(0.12)	0.07(0.14)	0.02(0.02)	0.02(0.06)	0.02(0.06)
δ_2^1	0.00(0.01)	0.06(0.13)	0.07(0.15)	0.01(0.02)	0.00(0.01)	0.06(0.13)	0.07(0.14)	0.01(0.02)	0.06(0.13)	0.06(0.13)
δ_0^2	0.01(0.06)	0.05(0.13)	0.10(0.20)	0.01(0.02)	0.02(0.09)	0.05(0.13)	0.10(0.20)	0.01(0.02)	0.02(0.09)	0.02(0.09)
δ_2^2	0.00(0.01)	0.06(0.15)	0.10(0.20)	0.01(0.02)	0.00(0.02)	0.05(0.14)	0.10(0.20)	0.02(0.02)	0.05(0.14)	0.05(0.14)
Pred>Error	4.33(0.41)	5.56(0.45)	10.39(1.32)	4.22(0.38)	4.43(0.50)	5.63(0.57)	10.47(1.35)	4.28(0.45)	4.43(0.50)	5.63(0.57)

Table 2

Simulation results under the homogeneity model: mean(sd) based on 100 replicates, $n=200$, and $p=500$.
 TP/FP: true/false positives. Varying and constant effects correspond to the continuous environment factor, and
 nonzero effect corresponds to the discrete environment factor.

	ρ	Varying Effect		Constant Effect		Nonzero Effect	
		TP	FP	TP	FP	TP	FP
ALASSO	0.4	4.00(0.00)	0.02(0.20)	1.98(0.20)	0.02(0.20)	3.90(0.44)	0.00(0.00)
	0.8	4.00(0.00)	0.20(0.60)	1.80(0.60)	0.00(0.00)	3.74(0.68)	0.04(0.28)
ALASSO _M	0.4	4.00(0.00)	0.03(0.17)	1.79(0.41)	0.04(0.20)	2.12(0.94)	0.00(0.00)
	0.8	3.99(0.10)	0.02(0.14)	1.77(0.42)	0.03(0.17)	2.14(0.94)	0.00(0.00)
LASSO _N	0.4	5.95(0.22)		23.21(9.49)		3.96(0.24)	21.87(10.60)
	0.8	5.96(0.20)		22.95(9.65)		3.90(0.33)	21.34(11.93)

Table 3

Number of selected SNPs and overlap of SNPs among the proposed method assuming the homogeneity (A_1) and heterogeneity models (A_2), ALASSO_M on BMI (A_3) and weight (A_4), and LASSO_N on BMI (A_5) and weight (A_6).

Method	# of SNPs	# of Overlapping SNPs				
		A_2	A_3	A_4	A_5	A_6
A_1	4	2	2	0	1	0
A_2	7		2	2	1	0
A_3	5			0	0	0
A_4	1				0	0
A_5	7					2
A_6	5					

Table 4

SNPs identified under the homogeneity model (A_1) and heterogeneity model (A_2).

	SNP	Position	Gene*	Varying Effect (Age)			Constant Effect			Nonzero Effect(FAMDB)		
				BMI	Weight		BMI	Weight		BMI	Weight	
A_1	rs2293595	69178919	YTHDC1	varying	varying							
	rs1502744	160505274	RAPGEF2				-0.22	-1.54				
	rs2221312	160508402	RAPGEF2				0.03	-0.93				
	rs905902	188449086	RP11-565A3.2	varying	varying							
A_2	rs2616443	54554929	LNX1		varying							
	rs11935423	100477461	TRMT10A		varying							
	rs17048768	137511440	TERFIP3	varying								
	rs1502744	160505274	RAPGEF2					-0.56				
	rs2221312	160508402	RAPGEF2					-0.03				
	rs7657391	175460402	LOC100419741					-0.34				
	rs6827557	188932293	ZFP42					0.06				

* Genes that SNPs belong to or are closest to.