Published in final edited form as:

Stat Appl Genet Mol Biol. 2015 June 1; 14(3): 265–277. doi:10.1515/sagmb-2014-0073.

Modeling gene-covariate interactions in sparse regression with group structure for genome-wide association studies

Yun Li,
Department of Mathematics and Statistics, Boston University, MA 02215, USA; and Department of Biostatistics, Boston University School of Public Health, MA 02118, USA

George T. O'Connor,
Pulmonary Center, Department of Medicine, Boston University School of Medicine, MA 02118, USA

Josée Dupuis, and
Department of Biostatistics, Boston University School of Public Health, MA 02118, USA

Eric Kolaczyk
Department of Mathematics and Statistics, Boston University, MA 02215, USA

Yun Li: yrlee@bu.edu

Abstract

In genome-wide association studies (GWAS), it is of interest to identify genetic variants associated with phenotypes. For a given phenotype, the associated genetic variants are usually a sparse subset of all possible variants. Traditional Lasso-type estimation methods can therefore be used to detect important genes. But the relationship between genotypes at one variant and a phenotype may be influenced by other variables, such as sex and life style. Hence it is important to be able to incorporate gene-covariate interactions into the sparse regression model. In addition, because there is biological knowledge on the manner in which genes work together in structured groups, it is desirable to incorporate this information as well. In this paper, we present a novel sparse regression methodology for gene-covariate models in association studies that not only allows such interactions but also considers biological group structure. Simulation results show that our method substantially outperforms another method, in which interaction is considered, but group structure is ignored. Application to data on total plasma immunoglobulin E (IgE) concentrations in the Framingham Heart Study (FHS), using sex and smoking status as covariates, yields several potentially interesting gene-covariate interactions.

Keywords

gene-environment/covariate interaction; genome-wide association studies; sparse regression

Correspondence to: Yun Li, yrlee@bu.edu.

1 Introduction

Earlier genetic studies focused on Mendelian traits which are, according to Mendel's law, typically triggered through a single mutated gene. More recently, advancement in genotyping technology has made genome-wide association studies (GWAS) possible, and has led to the discovery of multiple loci affecting complex diseases that do not exhibit a Mendelian inheritance pattern. However, most complex diseases are affected by both genetics and covariates, such as lifestyle variables. In order to better understand the etiology of disease, both genetics and environmental variables must be taken into consideration. For example, genetics factors may have different effects on diseases smokers and non-smokers. The multiple regression model with gene-environment interactions (G×E) or more generally gene-covariate interactions is therefore likely more suitable to find associations between diseases and different genetic factors.

In GWAS, single nucleotide polymorphisms (SNPs) are measured on a large collection of participants, and association between SNPs and trait of interest is tested one SNP at a time. The number of SNPs measured is usually in the order of millions, and can be even larger when imputation approaches are utilized to estimate the SNPs at ungenotyped loci, creating an ultra-high-dimensional problem that increases with the number of participants enrolled in a study. The classical variable selection method Lasso (Tibshirani, 1996) with L_1 penalty on the coefficients can help to select the important genetic factors. Numerous follow-up work has been done in the area with different penalties including the smoothly clipped absolute deviation (SCAD, Fan and Li, 2001), the elastic net (Zou and Hastie, 2005), the Adaptive Lasso (Zou, 2006), the Dantzig selector (Candes and Tao, 2007), the relaxed Lasso (Meinshausen, 2007), among others. Due to the presence of interactions, some special methods, such as the strong heredity interaction model (SHIM, Choi et al., 2010), the composite absolute penalties (CAP, Zhao et al., 2009) and the Variable selection using Adaptive Non-linear Interaction Structures in High dimensions (VANISH, Radchenko and James, 2010) are proposed to solve the selection problems by considering both main and interaction effects together. Naturally, all those models enforce a hierarchical structure where main effects are automatically added to a model simultaneously with the corresponding interaction term. This is considered as the marginality in generalized linear models (McCullagh and Nelder, 1989; Nelder, 1994) or the strong heredity in the study of designed experiments (Hamada and Wu, 1992). Justifications of the effects of heredity can be found in Chipman (1996) and Joseph (2006).

But current biological understanding is that genetic variables can be formed into certain groups according to biological information, such as biological pathways or gene functions. Even ignoring interactions in the model, the prior biological group information can play a crucial role in the variable selection for the main effects (Yuan and Lin, 2006; Huang et al., 2009; Zhou and Zhu, 2010; Friedman et al., 2010a, and Simon et al., 2013). Chen and Thomas (2010) proposed an approach to incorporate such biological knowledge, e.g., a Bayesian stochastic search algorithm was applied to identify gene-gene interactions. But none of the existing Lasso-like methodologies for selection of interactions incorporate prior group structure. In this paper, we design a special grouped interaction selection penalty (GISP) which not only enforces the interaction with the strong heredity property in the model, but also considers the prior biological group information in the study. For the study of the gene-covariate interactions, the interactions between genetic variables and risk factor variables are considered in the model, and by adding the genetic group information, our designed penalty can greatly affect the variable selection efficiency. Simulation studies show that our proposed GISP method performs much better than the existing SHIM model without considering group structure.

We apply our method on allergy disease studies with the long-term and ongoing Framingham Heart Study (FHS) data (Granada et al., 2012). The total plasma immunoglobulin E (IgE) concentrations, which is a biomarker related to allergy to environmental allergens, is used as the phenotype, and the genetic SNP variables are genotypes. The covariates, such as, sex, smoking status and age, are also considered in the study. The Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways are employed to group the genetic variables. The gene-covariate interactions are evaluated using the proposed method.

The rest of this paper is organized as follows: In Section 2, we describe our proposed method. We introduce the general model for gene-covariate interaction study, display the designed penalties and explain the specific roles the penalties play in the estimation procedures. We then present the algorithm to solve our estimation criteria in detail, and also show one way to reduce the high-dimensional computation cost. The proposed model is examined through extensive simulation studies in Section 3. The real data analysis of the IgE concentration data is provided in Section 4. Finally a short discussion is included in Section 5.

2 Methodology

In this section we present our proposed estimation method by considering both interaction and group structure in the model. The model and the optimization criterion are described in Section 2.1. A coordinate descent algorithm is then detailed in Section 2.2.

2.1 Optimization criterion

Suppose that there are p predictors in a multiple regression model, $X_1, \dots, X_p$, which may be collected into K groups, $G_1, \dots, G_K$. The groups are usually not disjoint and typically have very complex overlapping structures when defined by biological pathways. This means that for a given genetic predictor X_j, it may belong to more than one group. Denote the phenotype vector of response for n subjects as $Y = (Y_1, \dots, Y_n)^T$. Suppose that besides the p genetic variables, L risk factor variables, $E_1, \dots, E_L$, are considered in our study. For example, in genetic studies, sex and smoking status can be treated as important risk factors/ covariates related to phenotypes. The interaction terms between genetic variables and covariates are included in our analysis to gain a better understanding of the association between genotypes and phenotypes. Denote $I_{GE} = \{(j, l)\}: j \in G_g, 1 \le g \le K \text{ and } 1 \le l \le L\}$ as the two-way interaction set generated between gene and risk factor effects. Also naturally, we will insist that the strong heredity property is kept when interaction terms are included in the model, i.e., if the interaction term X_jE_l is in the model, then the main terms X_j and E_l must both be in the model.

In introducing the interaction terms between genetic variables X_j and risk factor variables E_l, we write the regression model as

$$Y=\sum_{j=1}^{p}\beta_j X_j+\sum_{l=1}^{L}\alpha_l E_l+\sum_{(j,l)\in I_{GE}}\gamma_{jl}\beta_j\alpha_l X_j E_l+\varepsilon$$

with a normal error vector $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T$.

In order to perform the variable selection not only including the group information but also keeping the heredity property, we design the following penalized estimation method:

$$\begin{aligned}(\hat{\alpha},\hat{\beta},\hat{\gamma})=&\arg\min_{(\alpha,\beta,\gamma)}\frac{1}{2}\left\|Y-\sum_{j=1}^{p}\beta_j X_j-\sum_{l=1}^{L}\alpha_l E_l-\sum_{(j,l)\in I_{GE}}\gamma_{jl}\beta_j\alpha_l X_j E_l\right\|^2\\&+\lambda_1\left(\sum_{j=1}^{p}|W_j^G\beta_j|+\sum_{l=1}^{L}|W_l^E\alpha_l|\right)+\lambda_2\sum_{g=1}^{K}\sqrt{p_g\sum_{j\in G_g}W_j^G w_j^2\beta_j^2}\\&+\lambda_3\sum_{(j,l)\in I_{GE}}|\gamma_{jl}|.\end{aligned}\qquad(1)$$

where λ_1, λ_2 and λ_3 are tuning parameters, p_g and w_j are pre-chosen weights for genetic groups and individual genetic predictors, and W_j^G and W_l^E are the indicator functions for the genetic and risk factor variables. If we don't penalize one particular jth genetic or lth risk factor variable and force it in the model, we can set W_j^G or W_l^E at 0. Otherwise, they are taken at the value of 1. The first Lasso penalty with tuning parameter λ_1 controls the sparsity of all main effects including genetic variables and risk factor variables. The second group Lasso penalty with tuning parameter λ_2 controls the sparsity of groups. The third penalty is applied to select important interaction terms. Because the size of each group may vary, p_g is used to avoid over-penalizing the groups with small size (Yuan and Lin, 2006). Moreover, because some of predictors may exist in more than two groups, we use the weights w_j to avoid over-penalization for individual predictors that exist in more than two groups. Typically, one can choose p_g to equal the size of the g-th group, and w_j is chosen as the reciprocal of the number of groups which contain the jth main effect. From the above regularized penalty, analogous to the SHIM model (Choi et al., 2010), we can easily find that the interaction coefficient $\gamma_{jl}\beta_j\alpha_l$ will shrink to zero if either β_j or α_l goes to zero. Therefore, the heredity property is automatically enforced in the optimized solution.

Note that, due to the generality with which our estimation criterion and notion of interactions and group structure are defined, our method is not restricted only to work with certain biological pathways, but can be applied as well to more general biological units with group structure, such as the functional units recently produced by the ENCODE study (The ENCODE Project Consortium, 2012).

In our simulation and real data analysis, there are only a handful of risk factor variables. We want to fully recover the interaction between genetic variable and all risk factor variables, and we set all W_l^E equal to 0. This means that the risk factor variables are all included in our estimation. In particular, when $\lambda_2 = 0$, the estimation criterion will reduce to the SHIM model (Choi et al., 2010), and when no interaction terms are involved and $\lambda_3 = 0$, this becomes similar to the method in Friedman et al. (2010a). But the difference is that the groups in our study have complex overlapping structure. In Friedman et al. (2010a), they only consider equal-size and non-overlapping groups.

2.2 Algorithm

In this subsection, we develop a unified shooting algorithm (Fu, 1998; Friedman et al., 2010b) for solving (1). The shooting algorithm is essentially a “coordinate descent” algorithm. In short, in each iteration we fix all but one coefficient, say, β_j, at their current values, then optimize (1) to solve for β_j. Because this optimization only involves one parameter, it is often easy to achieve a solution. Both simulation and theoretical results in Fu (1998) and Friedman et al. (2010b) show that this is a very stable and fast algorithm to solve L_1-type regularization problem. Moreover, similar to Friedman et al. (2010b), we can run iterations around the active set of variables with nonzero coefficients until convergence after a full cycle through all the variables. This active strategy significantly speeds up the convergence, specially, for large genetic datasets.

We first introduce some mathematical notations to better describe the algorithm. For the β coefficient vector, let β_{-j} be the same as the coefficient vector β except that the jth element is equal to 0, and for the α coefficient vector, α_{-l} holds the same meaning as β_{-j}. We denote $\beta_{(k)}$ as the coefficient vector for the group G_k. If $j \in G_k$, let $\beta_{(k),-j}$ be the same as the coefficient vector $\beta_{(k)}$ except that the jth element is equal to 0. Denote $G_{(j)} = \{k{:}j \in G_k\}$, $G_{(j)}^o = \{k{:}\|\beta_{(k),-j}\| = 0, k \in G_{(j)}\}$ and $G_{(j)}^N = G_{(j)} \backslash G_{(j)}^o$. The algorithm can be formulated as follows:

1. (Standardization): Center Y, center and normalize each X_j, E_l and X_jE_l
2. (Initialization): Initialize $\hat{\alpha}_l^{(0)}, \hat{\beta}_j^{(0)}$ and $\hat{\gamma}_{jl}^{(0)}$ with possible values. For example, use the least square regression results or simple regression results by regressing Y on each term.
3. (Update $\hat{\gamma}_{jl}$) For each $(j, l) \in I_{GE}$, update $\hat{\gamma}_{jl}$ with $\hat{\alpha}_l$, $\hat{\beta}_j$ and $\hat{\gamma}_{j_o l_o}$ $((j_0, l_0) \in I_{GE}/(j, l))$ fixed at the previous s-th step. Let

$$\tilde{Y} = Y - \sum_{j=1}^{p} \hat{\beta}_j^{(s)} X_j - \sum_{l=1}^{L} \hat{\alpha}_l^{(s)} E_l - \sum_{(j_o, l_o) \in I_{GE} \backslash (j,l)} \hat{\gamma}_{j_o l_o}^{(s)} \hat{\beta}_{j_o}^{(s)} \hat{\alpha}_{l_o}^{(s)} X_{j_o} E_{l_o};$$

$$\tilde{X}_{jl} = \hat{\beta}_j^{(s)} \hat{\alpha}_l^{(s)} X_j E_l.$$

Then update γ_{jl} with

$$\hat{\gamma}_{jl}^{(S+1)}=\frac{\left(|\tilde{X}_{jl}^T\tilde{Y}|-\lambda_3\right)_+\text{sign}(\tilde{Y}^T\tilde{X}_{jl})}{\tilde{X}_{jl}^T\tilde{X}_{jl}}.$$

4. (Update $\hat{\beta}$) For each $j\in\{1, \dots, p\}$, update $\hat{\beta}_j$ with $\hat{\alpha}_l$, $\hat{\beta}_{j_o}$ ($j_o \neq j$) and $\hat{\gamma}_{j_o l_o}$ ((j_o, l_o)$\in I_{GE}$, $j_o \neq j$) fixed at the previous s-th step. Let

$$\tilde{Y}=Y-\sum_{j_o\neq j}\hat{\beta}_{j_o}^{(s)}X_j-\sum_{l=1}^{L}\hat{\alpha}_l^{(s)}E_l-\sum_{j_o\neq j,(j_o,l_o)\in I_{GE}}\hat{\gamma}_{j_o l_o}^{(s)}\hat{\beta}_{j_o}^{(s)}\hat{\alpha}_{l_o}^{(s)}X_{j_o}E_{l_o};$$

$$\tilde{X}_j=X_j+\sum_{(j,l_o)\in I_{GE}}\hat{\gamma}_{jl_o}^{(s)}\hat{\alpha}_{l_o}^{(s)}X_jE_{l_o}.$$

If $G_{(j)}^N$ is the empty set φ, then

$$\hat{\beta}_j^{(s+1)}=\frac{\left(|S_j^{(s)}|-W_j^G\left(\lambda_1+\lambda_2\omega_j\sum_{k\in G_{(j)}}\sqrt{p_k}\right)\right)_+\text{sign}(S_j^{(s)})}{\tilde{X}_j^T\tilde{X}_j}.$$

where $S_j^{(s)}=\tilde{X}_j^T(\tilde{Y}-\tilde{X}\hat{\beta}_{-j}^{(s)})$ with $\tilde{X}=[\tilde{X}_1, \cdots, \tilde{X}_p]$ else if $G_{(j)}^N\neq\phi$ then

$$\hat{\beta}_j^{(s+1)}=\frac{\left(|S_j^{(s)}|W_j^G\left(\lambda_1+\lambda_2\omega_j\sum_{k\in G_{(j)}^o}\sqrt{p_k}\right)\right)_+}{\tilde{X}_j^T\tilde{X}_j+\lambda_2W_j^G\omega_j^2\sum_{k\in G_{(j)}^N}\sqrt{p_k}\left(\omega_j^2\hat{\beta}_j^{(s+1)2}+\sum_{j'\neq j}\omega_j^2\hat{\beta}_{j'}^{(s)2}\right)^{-1/2}}\text{sign}(S_j^{(s)}). \quad (2)$$

Note that both sides of (2) involve $\hat{\beta}_j^{(s+1)}$, thus the solution $\hat{\beta}_j^{(s+1)}$ can be achieved by iterating between the two sides of (2).

5. (Update $\hat{\alpha}$) For each $l\in\{1, \dots, L\}$, update $\hat{\alpha}_l$ with $\hat{\alpha}_{l_o}$($l_o \neq l$), $\hat{\beta}_j$ and $\hat{\gamma}_{j_o l_o}$((j_o, l_o)$\in I_{GE}$, $l_o \neq l$) fixed at the previous s-th step. Let

$$\tilde{Y}=Y-\sum_{j=1}^{p}\hat{\beta}_j^{(s)}X_j-\sum_{l_o\neq l}\hat{\alpha}_{l_o}^{(s)}E_{l_o}-\sum_{l_o\neq l,(j_o,l_o)\in I_{GE}}\hat{\gamma}_{j_o l_o}^{(s)}\hat{\beta}_{j_o}^{(s)}\hat{\alpha}_{l_o}^{(s)}X_{j_o}E_{l_o};$$

$$\tilde{E}_l=E_l+\sum_{(j_o,l)\in I_{GE}}\hat{\gamma}_{j_o l}^{(s)}\hat{\beta}_{j_o}^{(s)}X_{j_o}E_l.$$

Estimate $\hat{\alpha}_l$ by

$$\hat{\alpha}_l^{(s+1)}=\frac{\left(|S_l^{(s)}|-W_l^E(\lambda_1+\lambda_2)\right)_+}{\tilde{E}_l^T\tilde{E}_l}\text{sign}(S_l^{(s)});$$

where $S_l^{(s)}=\tilde{E}_l^T(\tilde{Y}-\tilde{E}\hat{\alpha}_{-l}^{(s)})$ with $\tilde{E}=[\tilde{E}_1,\cdots,\tilde{E}_L]$,

6. Calculate the difference $\Delta^{(s+1)} = \|\alpha^{\widehat{(s+1)}}\alpha^{\widehat{(s)}}\|+\|\beta^{\widehat{(s+1)}}\|+\|\lambda^{\widehat{(s+1)}}-\lambda^{\widehat{(s)}}\|$. If $\Delta^{(s+1)}$ is small enough, stop the algorithm. Otherwise, let $s = s+1$, go to 3.

In the above algorithm, the element-wise coordinate method is applied due to complicated overlapping group structure and interaction terms. In Yuan and Lin (2006) and Simon et al. (2013), they used the group-wise coordinate descent algorithm for simple non-overlapping group penalties. But the group effect can also be found when updating β_j in Step 4 of our algorithm. When $G_{(j)}^N$ is empty, meaning that all other β elements in the same group as β_j are shrunk to 0, and if the whole group(s) is/are not important, β_j should be shrunk to 0. In this situation, the threshold of β_j is $W_j^G(\lambda_1+\lambda_2\omega_j\sum_{k\in G_{(j)}}\sqrt{p_k})$ which is larger than $W_j^G(\lambda_1+\lambda_2\omega_j\sum_{k\in G_{(j)}^o}\sqrt{p_k})$ due to empty $G_{(j)}^N$. This means β_j would be shrunk to 0 more easily and the whole non-important group(s) would tend to be knocked out. Also using the same argument, if the important group has several important variables, the threshold when updating β_j is always smaller because of non-empty $G_{(j)}^N$. As a result, the important group will be kept during the iteration.

There are three tuning parameters in our estimation criteria. In order to reduce the computation cost, we set the three tuning parameters at reasonable ratios informed by carefully consideration. First, note that each X_j and X_jE_l are standardized in our estimation. But in order to maintain the heredity property, we add $\beta_j\alpha_l$ in the regressor term X_jE_l. The additional $\beta_j\alpha_l$ affect the threshold in the algorithm, and we cannot simply take λ_3 to be equal to λ_1. We can absorb the $\beta_j\alpha_l$ into the tuning parameter by setting $\lambda_3 = c_3\lambda_1$ where $c_3=\overline{|\beta_j\alpha_l|}$ the average value of absolute values of $\beta_j\alpha_l$ for all interaction pairs $(j, l)\in I_{GE}$. Since the true values of β_j and α_l are unknown, we use a rough approximation in the form of the least square estimates or ridge regression estimates for $p>n$ to find c_3.

Second, from examination of the penalties we observed previously that λ_1 controls the sparsity of main effects and λ_2 controls the sparsity of groups. In real biological data, the number of genetic predictors are typically much larger than the number of groups. The ratio of true predictors over all predictors is smaller than the ratio of true important groups over all groups. The ratio of λ_2 over λ_1 should likely be smaller than 1. We can find a simple justification from some theoretical results about the ratio of λ_2 over λ_1. In Nardi and Rinaldo (2008), if the groups have no overlapping structure and group sizes are equal, the tuning parameter for controlling the sparsity of groups with group Lasso method is around $K_1\sqrt{\log K/n}$ where K_1 is a constant related to the restricted eigenvalues of the design matrix with group structure constraint, and from Bickel et al. (2009), we know that the tuning parameter for controlling the sparsity of individual predictors with Lasso method is around $K_2\sqrt{\log p/n}$ where K_2 is also a constant related to the restricted eigenvalues of the design matrix. With the design matrix from the same data X, we might assume that $K_1 \approx K_2$, and since $K<p$, we have $\lambda_2/\lambda_1<1$. In our simulation study, we consider different values of the ratio $c_2 = \lambda_2/\lambda_1$ and find that the simulation results are not particularly sensitive to the value of c_2. We can therefore reduce three tuning parameters into one justified by the above analysis.

The number of predictors is usually very large. We first select a moderate number of nonzero main effects by ignoring the interaction terms. Within the selected main effects, the estimation criterion (1) is considered. Due to the extremely low signal-to-noise ratio (SNR) in real biological data, the estimates of β and λ could have very large standard errors and the traditional information criteria, such as BIC and AIC, may not work well in the presence of low SNR. Also, our goal is to select important associated genetic variables and possible gene-covariate interaction terms, not to predict the disease response variable *Y*. We are most interested in the subset of nonzero regression coefficients. Therefore, in analogy to Wu et al. (2009), instead of selecting the tuning parameters for each data with information criteria or cross validation, we choose a certain fixed number of predictors with gradually decreasing tuning parameters. The estimation procedure can be formulated in the following two steps.

- Step 1: We apply the double penalized group LASSO penalty on the main effects only and select *n*/4 main predictors with *n* samples. This step is similar to the relaxed lasso method in Meinshausen (2007). The main effects for both participating and non-participating interactions will be selected with high probability in this step. The optimization criterion is written as:

$$(\hat{\alpha},\hat{\beta})=\underset{(\alpha,\beta)}{\arg\min}\frac{1}{2}\left\|Y-\sum_{j=1}^{p}\beta_jX_j-\sum_{l=1}^{L}\alpha_lE_l\right\|^2+\lambda_1\left(\sum_{j=1}^{p}|W_j^G\beta_j|+\sum_{j=1}^{L}|W_l^E\alpha_l|\right)+\lambda_2\sum_{g=1}^{K}\sqrt{p_g\sum_{j\in G_g}W_j^Gw_j^2\beta_j^2}$$

  Because the effect of the group penalty, the unimportant groups tend to be shrunk simultaneously, and we cannot select nonzero main effects with exact numbers, for example, 250 if sample size is equal to 1000. Therefore we restrict the number of nonzero main effects in the range of [*n*/4−*n*/100, *n*/4+*n*/100], which is [240, 260] if *n* = 1000.

- Step 2: Within the selected main effects, we apply (1) to re-select *n*/20 nonzero main effects and also associated nonzero important interaction terms. Again, due to the effect of the group penalty, we pick up main effects in the range of [*n*/20−*n*/200, *n*/20+*n*/200], which is [45, 55] if *n* = 1000.

The proportion 1/4 and 1/20 in the two steps can be adjusted on a case by case basis. For our simulation study, Steps 1 and 2 can be done over 100 times with simulated data. We can rank the selection frequencies to find the important predictors and also associated interaction terms. For the real data analysis, one can apply Steps 1 and 2 on bootstrapped data or sub-sampled data for each analysis, and then rank the corresponding selection frequencies to detect important main effects and interaction terms. Also, we must assume that the true signals are sparse because the number of true predictors is unknown in real data.

3 Simulation study

In this section, the proposed method is evaluated using simulated data sets for modeling gene-covariate interactions. We use two scenarios to generate the genotype SNPs datasets. In the first scenario, the SNPs are simulated through independent binomial random generators. Due to complicated dependence structure of real SNPs, we also randomly subsample the real Framingham Heart Study dataset to generate the other SNPs datasets in the second scenario. In detail, for the first scenario, we simulate 1000 subjects with 1000 SNPs (SNP1 to SNP1000) in the model, i.e., $n=p=1000$. Within 1000 SNPs, there are 20 true SNPs with nonzero regression coefficients. The allele frequencies for the 20 true predictors are 0.3 and 0.5 alternatively, i.e, SNP1 is generated from *Binomial*(2, 0.3), SNP2 is generated from *Binomial*(2, 0.5), etc. In addition to 1000 SNPs, we generate 2 covariates E_1 and E_2 into the models, where E_1 is one binary random variable from *Bernoulli*(0.5) and E_2 is one normal distributed random variable for $N(0, 0.5^2)$. In our setting, we treat each SNP as a random variable with a binomial distribution, not a category variable with 3 levels. In some genetic studies, SNPs are considered as 3-level category variables and 2 dummy variables are used to represent each SNP. But in that setting, we will consider more restriction, e.g., the estimation coefficients for the 2 dummy variables should be either both not equal to 0 or both equal to 0. Also, the interactions term has similar restriction. The algorithm would be more complicated and the scale of predictors including both main and interactions are doubled. The computing load would be heavier. After considering SNPs as random binomial variables, similar to Choi et al. (2010), we standardize the main and interaction terms before applying the coordinate descent algorithm.

About the interactions, we specify two gene-covariate interaction settings as follows:

- Case I: We add the interactions (SNP1×E_1, SNP3×E_1, …, SNP9×E_1) and (SNP1×E_2, SNP3×E_2, …, SNP9×E_2) in the model.
- Case II: We add the interactions (SNP1 ×E_1, SNP3×E_1, …, SNP9×E_1) and (SNP11×E_2, SNP13×E_2, …, SNP19×E_2) in the model.

In Case I, both covariates E_1 and E_2 interact with the same set of true active SNPs, but in Case II, E_1 and E_2 interact with different sets of true active SNPs. The coefficients for both main and interaction effects are set for 80% power with 5% significance level under standard single-SNP GWAS models with additive-trait structure. In detail, the true coefficients for SNPs with 0.3 and 0.5 allele frequencies are set at 0.15 and 0.13 respectively. The true coefficient for E_1 and E_2 are equal to 0.21 and 0.15, respectively. The interaction coefficients are set at 0.24 between SNP and E_1, and 0.20 between SNP and E_2.

For both cases, we simulate a high level of normal observation noise with SNR equal to 0.1 to mimic similar real weak genetic signals. Since the simulated SNPs have independent correlation structure, in order to show the efficiency of our proposed method on real genetic data, we randomly take 100 subsamples of 1000 SNPs in our real data example in the second scenario of SNP dataset generation. We also assign the true SNPs with the same coefficients, and use the same interaction settings (Case I and Case II) as in the first scenario. The normal observation noises are still applied to guarantee SNR equal to 0.1 for the second scenario.

The group structures are simulated from the most popular and interesting real KEGG biological pathway. Since about one-third of genes are found in the KEGG pathways in our real biological dataset, we first randomly sample 300 genes from KEGG pathways, and 700 genes are not from KEGG. We can consider these other 700 genes as 700 groups of size 1. First the 300 genes in the pathways are randomly selected. We then use our first 300 SNPs to represent the 300 selected genes, one SNP per gene. In our simulation, 159 pathways are formed to group 300 genes with the KEGG pathway information. The total number of pathways in KEGG is 186. Therefore around 85% KEGG groups are represented in our simulation. Among the selected 159 pathways, only 16 of them do not overlap with others.

We design two strategies to assign the true SNPs into the formed groups. In the first strategy (Group I), the true SNPs are assigned to guarantee that the true active SNPs percentages are lower than 10% in their groups. This is one more realistic scenario comparing with real genetic data. We put six true SNPs (SNP1 to SNP6) in the pathway group with the largest size. The 12 SNPs (SNP7 to SNP18) are randomly put into 6 groups which include at least 20 variables. SNP19 and SNP20 are randomly distributed into 2 additional groups which include at least 10 variables. In this setting, the situation that one group may contain only one true active SNP is also simulated. In the second strategy (Group II), all of the true SNPs are put in the largest group, and don't overlap with other groups. This is one extreme case. We use this special group assignment to find how the performance of our method is affected by the groups containing true SNPs.

The two covariates are not penalized and are forced in our model. We run the simulation 100 times for each simulation setting, and record the selection frequencies for main effects in Step 1, and the selection frequencies for both main effects and interaction terms in Step 2. For the competing SHIM model, we simply run Lasso without considering the group structure in Step 1. We also rank the selection frequency of main and interaction effect terms, and use the top 20 main SNPs to calculate the false discovery rate (FDR) for main effects FDR_M and the top 10 interaction terms to calculate the corresponding FDRI for interaction terms.

The simulation results of our proposed method are shown and compared with the results of the SHIM model in Tables 1 and 2. From the results of various simulation outcome, due to the additional group Lasso penalty, our proposed method tends to have much higher selection frequencies for true active main effects and lower selection frequencies for non-active main effects comparing with the SHIM method. This means that the power performance of our method is much better for main effects. The performance of selection frequencies is also very consistent when the ratio c_2 between λ_2 and λ_1 ranges from 0.1 to 0.9. The selection frequencies at $c_2 = 0.9$ are slightly higher than the results at $c_2 = 0.1$, especially, for simulation using random subsamples of real SNPs. Within the same interaction case, the performance of our method for the second true SNP assignment Group II is slightly better comparing with Group I under the same set of SNP variables. This is due to the special true SNP group structure. Since all true SNPs are in the same group, the group Lasso penalty is more efficient to knock out all nuisance groups. Because SHIM method does not consider the group structure, the performance of SHIM for Group I and Group II is similar. Also within the same interaction case, due to the complicated correlated structure of real SNPs, we can find that the results of real SNPs are always worse than the ones of independent simulated SNPs. The selection frequencies for interaction terms are comparable between our proposed method and SHIM method. In most situations, our method selects the true interaction terms with slightly higher frequencies, and selects the non-true interaction terms with slightly lower frequencies. Moreover, because the interaction effects of Case II are much stronger than the ones of Case I, both the true main variables involved in interaction and the true interactions have higher selection frequencies in Case II. In terms of FDR, our method tends to have better performance for main effects. For interaction effects, most simulation results indicate that our method performs better than SHIM. This means that our methods generally tend to have smaller Type I error comparing with SHIM.

In Figure 1, we plot the histograms of main effect selection frequencies for the two different interaction cases (Case I and Case II) of SHIM method and our GISP method at $c_2 = 0.5$. Since the performance patterns of our proposed method and SHIM for different true SNP assignment strategies and different datasets using in the simulation are similar, we only display the simulation results of Group I with random subsamples of real SNPs. We can find the frequencies at low main effect selection frequencies in our method are larger than the ones in SHIM method. Also, the frequency bars from the true active main effect are further apart from the histogram peak from non-active main effects. The dotted line in Figure 1 represents the minimum value of true SNP selection frequencies (f_d), while the solid one represents the 20-th value of the ordered selection frequencies for all SNPs (f_s). The smaller the relative distance, which is defined as $(f_s - f_d)/f_d$, the better performance of the estimation method. Our method has smaller relative distances in all simulation situations comparing to SHIM. Moreover, if there are fewer SNPs between the solid line and dotted line (δ_N), one can improve FDR result by just lowering the cutoff number of chosen SNPs. Comparing with SHIM, our method always has a smaller δ_N for both interaction cases.

4 Real data analysis

In this section, we use the Framingham Heart Study (FHS) data in illustrate the performance of our proposed method in real data. Participants from the town of Framingham, Massachusetts have been recruited in the studies from 1948, and have been followed over the years for the development of heart disease and related traits, including pulmonary function and allergic response measured by IgE concentration. We use the log transformed plasma IgE concentration (logIgE), which is a biomarker that is often elevated in individuals with allergy to environmental allergens, as the response phenotype. The plasma IgE concentration is associated with allergic diseases, for example, asthma, allergic rhino conjunctivitis, atopic dermatitis, and food allergy. In Granada et al. 2012, some genes associated with IgE are identified, but the gene-covariate interaction has not yet been carefully studied. In our analysis, we consider the risk factor variables *Sex*, *Former Smoker*, *Current Smoker* and *Age*, and apply our method to detect possible gene-covariate interactions using the logIgE concentration response variable.

The genotype SNP data are from Affymetrix 500 K and MIPS 50 K arrays, with imputation performed using HapMap 2 European reference panel (Li and Abecasis, 2006). The expected number of minor alleles, i.e., dosage genotypes, are used in our analysis. Some pre-processing was applied to select a set of SNPs for the final analysis. We first attempt to map each of 2,411,590 genotyped and imputed SNPs in the dataset to a reference gene containing it. If no such gene is available, we map the SNP to the closest reference gene within 60 kilobases of the SNP, if available. Because this example focus on gene groups, SNPs that are not within 60 kilobases of a gene are excluded. After mapping SNPs to genes, some genes are found to include multiple SNPs. In this situation, we select one SNP, which is most significantly associated with the phenotype logIgE, to represent the gene using a linear mixed effect regression. Finally, we get 17,025 SNPs and construct a unique SNP-to-gene correspondence.

There are 6918 participants (3183 men and 3735 women) included in our analysis. Among the participants, there are 6674 related individuals from 991 families and 244 persons who have no relatives in the dataset. We first reduce the number of SNPs to 1000 by ranking the correlations with the response variable logIgE. This type of univariate screening process is justified, for example, by theory by Fan and Lv (2008). Then we take 100 random subsamples of 1000 participants from all of participants. All existing theoretical works about Lasso-type variable selection methods are based on homoscedastic random noises, such as, Fan and Li (2001) and Bickel et al. (2009). Due to family structures in our data, the noise errors within each family might be heteroscedastic. We take certain steps to avoid the heteroscedasticity when random samples are from the same family, such as, siblings with same mother or father cannot be sampled together, and parents and offsprings cannot be sampled together. The KEGG pathways are used to group the genes in our analysis. For those genes which are not in the KEGG pathway, we simply treat them as individual groups with size 1. In the pre-selected 1000 genes, there are 291 genes found in the KEGG pathways. These 291 genes form 152 groups, and among those groups, only 16 groups do not overlap with others. The group structure is similar to our simulation study.

We apply both our proposed method and the SHIM method to this real data. We set $c_2 = 0.5$ in our method. The real data are more noisy than the simulated data. The interaction selection frequencies are very low when we take the c_3 value suggested in the simulation. We lower c_3 to some extent, say, $c_3/50$, to allow the weaker interaction terms into the model. We rank the selection frequencies for both the main effects and interaction terms, pick the top 20 main effects and top 10 interaction terms and list them in Table 3 for the gene-covariate interaction outcomes.

In general, our proposed method has slightly higher selection frequencies for both main and interaction effects comparing to the SHIM method. The gene-covariate results show that the interactions between genetic variable and *Sex* has high selection frequencies comparing to other interactions from both our method and SHIM method. Some of genes may have weak interaction with smoking status, such as LRP1 and OSBPL3 from our proposed method, and EMID2 from the SHIM method. Since most of gene-covariate interaction studies are observational studies, further study using other data sets is recommended to confirm our results.

5 Discussion

In this paper, we have proposed a new method,which we call "GISP," to model the interactions with strong heredity property and simultaneously incorporate the prior biological group structure during the estimation. We also implement a unified fast "coordinate descent" algorithm to implement the proposed new method for gene-covariate interaction studies. The numerical simulation results show that the new designed penalty has much better selection performance compared to the SHIM model, in which the group structure is not considered. These results suggest substantial promise for the use of this method to detect the gene-covariate interactions for the genome-wide association studies (GWAS).

Due to the difficulty in choosing the three turning parameters, we use multiple simulation replicates in our simulation study, and bootstrap samples in the real data analysis, and treat variables selected with high frequency as important variables. This is very computationally expensive. Moreover, because the Lasso-type regularization cannot provide standard error estimates, it is difficult to set up a proper hypothesis test to evaluate our results. But similar to Nardi and Rinaldo (2008) and Bickel et al. (2009), the theoretical nonasymptotical bounds of our estimators could be derived and used to justify the results.

Because the relationship between genotypes at a variant and a phenotype may also be influenced by other genetic variants, in addition to studying gene-covariate interactions, it is straightforward to extend our gene-covariate estimation criterion to study gene-gene (G×G) interactions. With biological pathway information, one can assume that two-way interactions between genetic variables to be allowed only within the same group. Then, similar to the interaction set I_{GE}, one can define the interaction set $I_{GG} = \{(j, j'):\text{both } j \text{ and } j' \in G_g, g = 1, \cdots, K\}$ for the gene-gene study. To consider possible interaction across groups, the interaction set I_{GG} can be revised according to other reasonable requirement. But the estimation criteria and algorithm for genegene is similar to the algorithm presented for gene-covariate study. Moreover, it is worth mentioning that our method can select gene-covariate and gene-gene interactions simultaneously within one criteria if we modify the interaction set to the union set of I_{GE} and I_{GG}.

In the real data analysis, we apply our method on the FHS data to find important genes related to the plasma IgE concentration. To minimize the high correlation due to the linkage disequilibrium (LD) with each gene, we select one SNP per genes. However, one could potentially select multiple SNPs per genes, or use the first principal component (PC) to represent a gene (Gauderman et al., 2007). For multiple SNP approach, it is difficult to find a standard criterion to select useful SNPs which have no high correlation structure. Other approaches are worth investigating in future work.

Acknowledgments

This research is supported by National Institute Health grants ES020827, DK078616, N01 HC25195 and P01 AI050516 (in part). A portion of this research was conducted using the Linux Clusters for Genetic Analysis (LinGA) computing resources at Boston University Medical Campus.

References

Bickel P, Ritov Y, Tsybakov A. Simultaneous analysis of lasso and dantzig selector. Ann. Stat. 2009; 37:1705–1732.

Candes E, Tao ET. The dantzig selector: Statistical estimation when *p* is much larger than *n* (with discussion). Ann. Stat. 2007; 35:2313–2351.

Chen G, Thomas D. Using biological knowledge to discover higher order interactions in genetic association studies. Genet. Epidemiol. 2010; 34:863–878. [PubMed: 21104889]

Chipman H. Bayesian variable selection with related predictors. Can. J. Stat. 1996; 24:17–36.

Choi N, Li W, Zhu J. Variable selection with the strong heredity constraint and its oracle property. J. Am. Stat. Assoc. 2010; 105:354–364.

Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. J. Am. Stat. Assoc. 2001; 96:1348–1360.

Fan J, Lv J. Sure independence screening for ultra-high dimensional feature space. J. R. Stat. Soc., Series B. 2008; 70:849–911.

Friedman J, Hastie T, Tibshirani R. A note on the group lasso and sparse group lasso. arXiv: 1001.0736v1. 2010a (http://arxiv.org/pdf/1001.0736v1.pdf).

Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J. Stat. Software. 2010b; 33:1–22.

Fu W. Penalized regression: the bridge versus the lasso. J. Comput. Graph. Stat. 1998; 7:397–416.

Gauderman W, Murcray C, Gilliland F, Conti D. Testing association between disease and multiple SNPs in a candidate gene. Genet. Epidemiol. 2007; 31:383–395. [PubMed: 17410554]

Granada M, Wilk J, Tuzova M, Strachan D, Weiding S, Albrecht E, Gieger C, Heinrish J, Himes B, Hunninghake G, Celedn J, Weiss S, Cruikshank W, Farrer L, Center D, O'Connor G. A genome-wide association study of plasma total IgE concentration in the Framingham Heart Study. J. Allergy Clin. Immun. 2012; 129:840–845. [PubMed: 22075330]

Hamada M, Wu C. Analysis of designed experiments with complex aliasing. J. Qual. Technol. 1992; 24:130–137.

Huang J, Ma S, Xie H, Zhang C. A group bridge approach for variable selection. Biometrika. 2009; 96:339–355. [PubMed: 20037673]

Joseph V. A Bayesian approach to the design and analysis of fractionated experiments. Technometrics. 2006; 48:219–229.

Li Y, Abecasis G. Mach 1.0: rapid haplotype reconstruction and missing genotype inference. Am. J. Hum. Genet. S. 2006; 79:2290.

McCullagh, P.; Nelder, J. Generalized linear models. London: Chapman & Hall/CRC; 1989.

Meinshausen N. Relaxed lasso. Comput. Stat. Data Anal. 2007; 52:374–393.

Nardi Y, Rinaldo A. On the asymptotic properties of the group lasso estimator for linear models. Electron. J. Stat. 2008; 2:605–633.

Nelder J. The statistics of linear models: Back to basics. Stat. Comput. 1994; 4:221–234.

Radchenko P, James G. Variable selection using adaptive nonlinear interaction structures in high dimensions. J. Am. Stat. Assoc. 2010; 105:1541–1553.

Simon N, Friedman J, Hastie T, Tibshirani R. A sparse-group lasso. J. Comput. Graph. Stat. 2013; 22.2:231–245.

The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489:57–74. [PubMed: 22955616]

Tibshirani R. Regression shrinkage and selection via the lasso. J. R. Stat. Soc., Series B. 1996; 58:267–288.

Wu T, Chen Y, Hastie T, Sobel E, Lange K. Genomewide association analysis by lasso penalized logistic regression. Bioinformatics. 2009; 25:714–721. [PubMed: 19176549]

Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. J. R. Stat. Soc., Series B. 2006; 68:4967.

Zhao R, Rocha G, Yu B. The composite absolute penalties family for grouped and hierarchical variable selection. The Annals of Stat. 2009; 6A:3468–3497.

Zhou N, Zhu J. Group variable selection via a hierarchical lasso and its oracle property. Stat. Interface. 2010; 3:574.

Zou H. The adaptive lasso and its oracle properties. J. Am. Stat. Assoc. 2006; 101:1418–1429.

Zou H, Hastie T. Regularization and variable selection via the elastic net. J. R. Stat. Soc., Series B. 2005; 67:301–320.

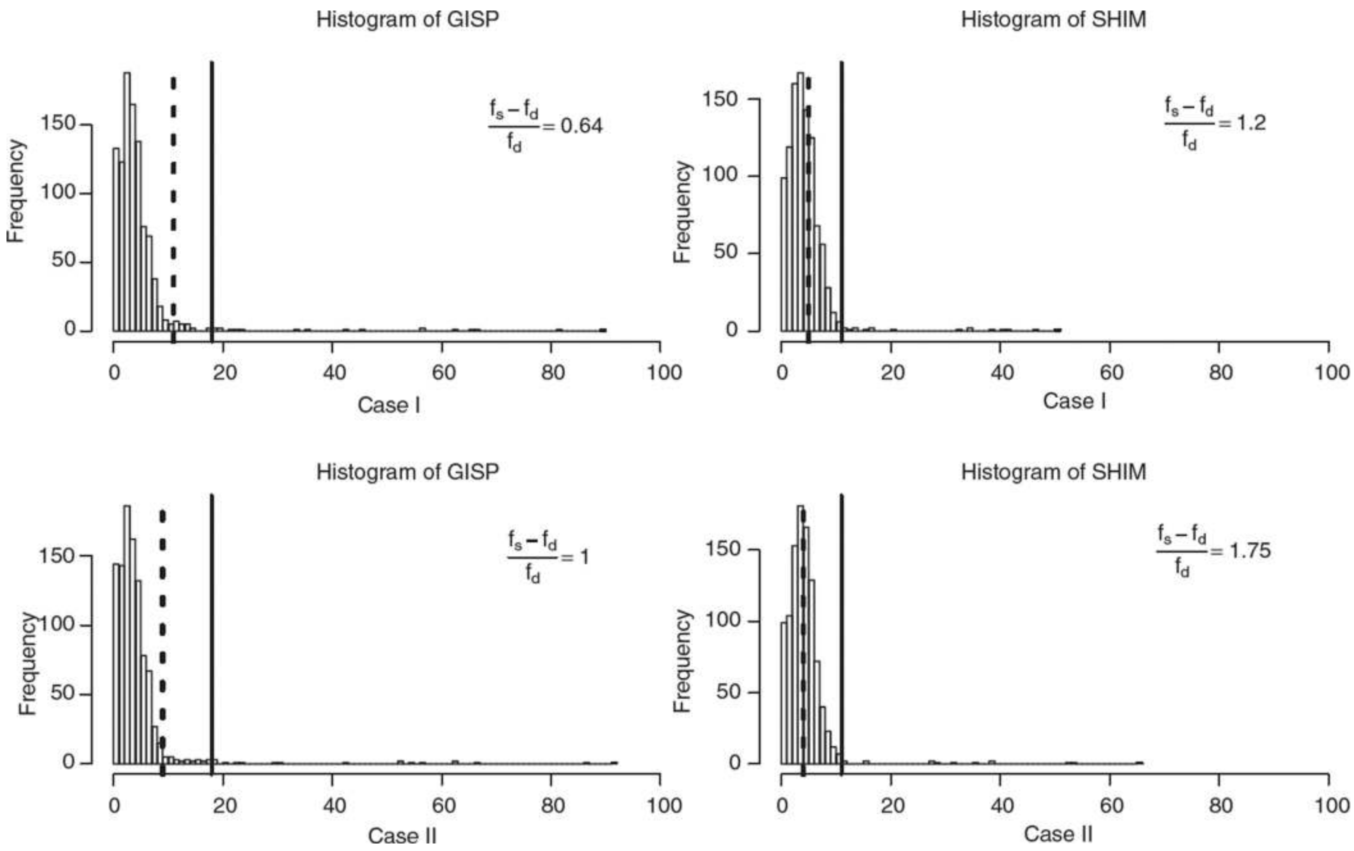

Figure 1.

Histograms for SNP selection frequencies of two different interaction settings (Case I and Case II) with the same true SNP assignment strategy (Group I) using random subsamples of real SNPs dataset. Dotted lines: the minimum value of true SNP selection frequencies. Solid lines: the 20th value of the ordered selection frequencies for all SNPs.

Table 1

Simulation results for Case I with two different true SNP assignment strategies (Group I and Group II) and two different ways to generate SNP datasets (simulated SNPs and real SNPs).

	Step 1				Step 2						
	S_w^A	S_o^A	S^N	FDR_M	S_w^A	S_o^A	S^N	I^A	I^N	FDR_M	FDR_I
Group I with simulated SNPs											
GISP(0.1)	85.40	60.47	24.02	0.00	57.80	24.47	4.29	13.10	0.49	0.00	0.10
GISP(0.3)	85.60	61.73	23.88	0.00	62.40	27.13	4.26	13.50	0.69	0.00	0.10
GISP(0.5)	86.80	62.60	23.88	0.00	64.00	29.13	4.24	15.20	1.04	0.05	0.10
GISP(0.7)	87.00	63.20	23.92	0.00	65.40	32.80	4.10	13.80	0.95	0.05	0.10
GISP(0.9)	87.00	64.07	23.86	0.00	66.80	35.87	4.02	12.30	0.96	0.05	0.20
SHIM	73.60	43.73	24.27	0.05	42.00	15.27	4.47	11.80	1.00	0.10	0.20
Group II with simulated SNPs											
GISP(0.1)	87.00	60.87	24.01	0.00	61.80	28.13	4.25	14.90	0.62	0.00	0.00
GISP(0.3)	89.20	63.00	23.90	0.00	68.80	36.33	3.99	15.40	0.55	0.00	0.00
GISP(0.5)	90.00	65.00	23.87	0.00	73.20	40.73	3.90	17.20	0.81	0.00	0.10
GISP(0.7)	90.60	66.53	23.79	0.00	75.60	44.13	3.84	17.10	0.79	0.00	0.10
GISP(0.9)	91.00	67.47	23.80	0.00	74.80	45.60	3.81	17.70	0.90	0.00	0.10
SHIM	74.00	44.40	24.29	0.05	43.60	15.80	4.41	12.30	1.20	0.05	0.20
Group I with real SNPs											
GISP(0.1)	64.40	53.40	24.12	0.2	49.20	24.27	4.40	10.50	0.58	0.25	0.30
GISP(0.3)	69.60	55.93	24.02	0.2	55.00	28.00	4.24	10.60	0.72	0.35	0.20
GISP(0.5)	71.40	57.40	24.09	0.15	56.80	31.60	4.19	8.70	0.74	0.25	0.30
GISP(0.7)	72.40	58.60	23.97	0.2	61.60	34.13	4.11	11.10	0.92	0.25	0.30
GISP(0.9)	74.20	59.73	24.07	0.3	64.60	38.93	4.04	9.50	0.92	0.15	0.30
SHIM	52.60	36.67	23.79	0.45	36.60	18.47	4.37	8.00	0.81	0.30	0.50
Group II with real SNPs											
GISP(0.1)	66.60	53.60	24.18	0.15	50.80	28.20	4.28	9.60	0.42	0.20	0.30
GISP(0.3)	70.40	58.20	24.11	0.10	59.80	37.53	4.09	7.80	0.64	0.10	0.30
GISP(0.5)	73.60	60.73	23.91	0.10	64.40	41.47	4.05	12.60	0.83	0.05	0.30

	Step 1				Step 2						
	S_w^A	S_o^A	S^N	FDR_M	S_w^A	S_o^A	S^N	I^A	I^N	FDR_M	FDR_I
GISP(0.7)	74.80	62.27	23.94	0.10	67.60	45.53	3.93	10.20	0.88	0.00	0.20
GISP(0.9)	76.80	63.53	23.88	0.05	69.40	48.07	3.90	12.50	1.01	0.05	0.30
SHIM	55.80	38.73	23.96	0.35	38.80	17.87	4.40	8.60	0.87	0.35	0.40

"GISP(c_2)" means our proposed estimation method with $\lambda_2 = c_2\lambda_1$. "SHIM" refers to the method in Choi et al., 2010. S_w^A is the selection frequency(%) for true active SNPs involved in interaction; S_o^A is the selection frequency(%) for true active SNPs not involved in interaction; S^N is the selection frequency(%) for non-active SNPs. I^A is the selection frequency(%) for active interaction terms; I^N is the selection frequency(%) for non-active interaction terms.

Table 2

Simulation results for Case II with two different true SNP assignment strategies (Group I and Group II) and two different ways to generate SNP datasets (simulated SNPs and real SNPs). Notations have the same meanings as in Table 1.

	Step 1				Step 2						
	S_w^A	S_o^A	S^N	FDR_M	S_w^A	S_o^A	S^N	I^A	I^N	FDR_M	FDR_I
Group I with simulated SNPs											
GISP(0.1)	71.30	56.40	24.07	0.00	39.30	22.40	4.34	8.20	0.52	0.00	0.40
GISP(0.3)	72.60	58.90	23.91	0.00	41.90	25.90	4.26	7.80	0.71	0.00	0.40
GISP(0.5)	73.60	60.10	23.96	0.00	43.60	27.90	4.22	8.30	0.80	0.00	0.50
GISP(0.7)	74.00	60.40	23.94	0.00	45.30	31.70	4.11	9.50	0.93	0.00	0.50
GISP(0.9)	74.50	60.60	23.85	0.00	48.80	34.40	4.12	9.60	1.02	0.00	0.50
SHIM	56.90	40.00	23.98	0.15	29.10	12.60	4.62	8.10	1.07	0.15	0.50
Group II with simulated SNPs											
GISP(0.1)	74.40	58.40	24.00	0.00	46.10	27.10	4.27	11.00	0.63	0.00	0.40
GISP(0.3)	76.10	61.70	23.81	0.00	52.50	33.80	4.10	9.80	0.75	0.00	0.40
GISP(0.5)	77.00	63.90	23.84	0.00	57.80	38.90	3.98	10.60	0.84	0.00	0.50
GISP(0.7)	77.80	65.80	23.88	0.00	59.60	42.30	3.89	13.10	0.94	0.00	0.50
GISP(0.9)	78.30	66.60	23.68	0.00	61.90	45.20	3.88	13.40	1.02	0.00	0.40
SHIM	58.60	41.50	23.95	0.15	29.10	14.60	4.46	8.20	1.20	0.15	0.50
Group I with real SNPs											
GISP(0.1)	59.30	54.20	24.03	0.2	37.50	30.30	4.27	7.10	0.36	0.15	0.50
GISP(0.3)	62.40	57.10	24.08	0.25	40.60	34.30	4.16	7.70	0.46	0.30	0.50
GISP(0.5)	64.70	58.90	24.15	0.25	43.90	37.50	4.00	7.10	0.38	0.25	0.60
GISP(0.7)	64.90	59.80	24.10	0.25	45.50	39.40	4.01	6.60	0.51	0.15	0.50
GISP(0.9)	65.70	61.00	24.07	0.25	47.80	42.90	4.04	9.50	0.92	0.15	0.50
SHIM	45.70	40.60	24.06	0.45	28.20	20.40	4.35	6.60	0.60	0.40	0.60
Group II with real SNPs											
GISP(0.1)	60.00	54.80	24.04	0.15	39.40	31.00	4.21	6.50	0.49	0.10	0.50
GISP(0.3)	63.50	58.50	24.06	0.15	47.20	38.40	4.02	7.20	0.64	0.10	0.60
GISP(0.5)	66.10	60.70	23.91	0.10	51.70	42.60	4.01	7.90	0.89	0.05	0.60

	Step 1				Step 2						
	S_w^A	S_o^A	S^N	FDR_M	S_w^A	S_o^A	S^N	I^A	I^N	FDR_M	FDR_I
GISP(0.7)	69.40	62.40	23.96	0.10	55.40	46.30	3.85	8.60	0.84	0.05	0.60
GISP(0.9)	70.80	64.00	24.00	0.10	59.40	50.40	3.77	8.30	0.82	0.05	0.50
SHIM	47.00	40.50	23.86	0.40	26.80	20.50	4.38	5.40	0.58	0.35	0.60

Table 3

Gene-covariate results.

GISP				SHIM			
Main	**Freq.**	**Interaction**	**Freq.**	**Main**	**Freq.**	**Interaction**	**Freq.**
LRP1	34	LRP1×Sex*	14	EMID2	24	FCER1A×Sex*	11
ANKS4B	33	ANKS4B×Sex*	14	STAT6	22	EMID2×Sex*	10
OSBPL3	31	RAD50×Sex*	12	FCER1A	22	STAT6×Sex*	9
RAD50	29	TRAF3×Sex*	12	OSBPL3	21	SLC13A3×Sex*	8
FCER1A	28	IL13×Sex*	11	HAND1	18	LRP1×Sex*	8
EDAR	26	PPP2R2B×Sex*	10	RAD50	17	OSBPL3×Sex*	8
LOC441108	26	FBLN1×Sex*	10	SNFT	17	EMID2×Current Smoker*	8
IL13	24	LOC441108×Sex*	9	IL13	17	HAND1×Sex*	7
PPP2R2B	24	LRP1×Former Smoker*	9	GPR98	16	SNFT×Sex*	7
HLA-DPB1	23	OSBPL3×Current Smoker*	9	LRP1	16	IL13×Sex*	7
TRAF3	23			ANKS4B	16		
BDH1	22			HLA-DPB1	16		
SRL	21			KIAA1609	15		
TNFSF4	21			PXDNL	15		
RAB22A	21			TCTEX1D1	15		
IL6	21			STT3B	15		
LIN28B	19			TLL1	15		
KIAA1609	19			BDH1	15		
FRY	19			SLC13A3	15		
FBLN1	19			SRL	15		

* in the "Interaction" column represents that the gene which participates the interaction is in the "Main" column.