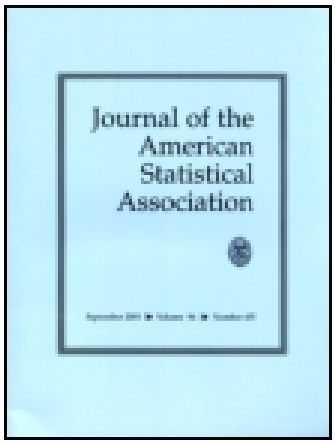

Journal of the American Statistical Association
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/uasa20

Biomarker Detection in Association Studies: Modeling SNPs Simultaneously via Logistic ANOVA

Yoonsuh Jung[a], Jianhua Z. Huang[bc] & Jianhua Hu[d]

[a] Department of Statistics, Univerisity of Waikato, Private Bag 3105, Hamilton, 3240, New Zealand

[b] Department of Statistics, , Texas A&M University, College Station, TX, USA

[c] ISEM, Captial University of Economics and Business, Beijing, China

[d] Department of Biostatistics, , The University of Texas MD Anderson Cancer Center, Houston, TXUSA

Accepted author version posted online: 12 Jun 2014.

To cite this article: Yoonsuh Jung, Jianhua Z. Huang & Jianhua Hu (2014): Biomarker Detection in Association Studies: Modeling SNPs Simultaneously via Logistic ANOVA, Journal of the American Statistical Association, DOI: 10.1080/01621459.2014.928217

To link to this article: http://dx.doi.org/10.1080/01621459.2014.928217

Biomarker Detection in Association Studies: Modeling SNPs Simultaneously via Logistic ANOVA

Yoonsuh Jung, Jianhua Z. Huang, Jianhua Hu

Abstract

In genome-wide association studies, the primary task is to detect biomarkers in the form of Single Nucleotide Polymorphisms (SNPs) that have nontrivial associations with a disease phenotype and some other important clinical/environmental factors. However, the extremely large number of SNPs comparing to the sample size inhibits application of classical methods such as the multiple logistic regression. Currently the most commonly used approach is still to analyze one SNP at a time. In this paper, we propose to consider the genotypes of the SNPs simultaneously via a logistic analysis of variance (ANOVA) model, which expresses the logit transformed mean of SNP genotypes as the summation of the SNP effects, effects of the disease phenotype and/or other clinical variables, and the interaction effects. We use a reduced-rank representation of the interaction-effect matrix for dimensionality reduction, and employ the L_1-penalty in a penalized likelihood framework to filter out the SNPs that have no associations. We develop a Majorization-Minimization algorithm for computational implementation. In addition, we propose a modified BIC criterion to select the penalty parameters and determine the rank number. The proposed method is applied to a Multiple Sclerosis data set and simulated data sets and shows promise in biomarker detection.

KEYWORDS: BIC, GWAS, MM Algorithm, L_1-penalty, Penalized Bernoulli Likelihood, Simultaneous Modeling of SNPs

Author's Footnote:

Yoonsuh Jung is Lecturer (Email: yoonsuh@waikato.ac.nz), Department of Statistics, Univerisity of Waikato, Private Bag 3105, Hamilton 3240, New Zealand. Jianhua Z. Huang is Professor (Email: jianhua@stat.tamu.edu), Department of Statistics, Texas A&M University, College Station, TX, USA, and Special Term Professor at ISEM, Captial University of Economics and Business, Beijing, China. Jianhua Hu is Associate Professor (Email: jhu@mdanderson.org), Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. Hu's work was partially supported by the National Institute of Health Grants R21CA129671, R01GM080503, R01CA158113, and CGSG P30 CA016672. Huang's work was partially supported by grants from NSF (DMS-0907170, DMS-1007618, DMS-1208952), and Award Number KUS-CI-016-04 and GRP-CF-2011-19-P-Gao-Huang, made by King Abdullah University of Science and Technology (KAUST). The authors would like to thank the editor, the associate editor and reviewers for many constructive comments. Corresponding author: Jianhua Hu.

1 Introduction

The area of genome-wide association studies (GWAS) has been growing rapidly in recent years. This type of study is designed to explore the associations between genetic markers, disease phenotypes, and other clinical/environmental factors. Particularly, GWAS focuses on identification of susceptible Single Nucleotide Polymorphisms (SNPs) as biomarkers of a disease. The promise of this technology has been shown in various biomedical studies (Egan et al., 2011; Festen et al., 2011; Shete et al., 2009). For example, Egan et al. (2011) identified susceptible variants in malignant gliomas-deadly brain tumors in a case-control study. They used multinomial logistic regression to evaluate genotype associations for glioma subtypes.

One of the challenges in analyzing GWAS data is the presence of extremely large number of SNPs (often at the magnitude of hundreds of thousands), which hinders applying systematic approaches to search for nontrivial associations. Currently, a commonly taken approach is to examine the association between a phenotype and one SNP at a time (Shete et al., 2009; Chen et al., 2011) by using some classical statistical methods (e.g., Pearson's chi-squared test and logistic regression). To consider several SNPs jointly, one usually takes a traditional regression approach (e.g., Li et al., 2012), in which the disease phenotype are treated as the response variable and the SNPs as the covariates. To further study the interactions among disease, SNPs, and other clinical factors, the corresponding interaction terms have been included as additional covariates (Chatterjee et al., 2006; Maity et al., 2009; Kooperberg et al., 2009), which introduces even more parameters. However, the contrast between the number of SNPs (or the number of the covariates) and that of the subjects (at most several thousands) makes it infeasible to interrogate the association between the phenotype and many SNPs simultaneously via a single regression model. A common strategy is to disregard many SNPs through a screening process prior to conducting the formal analysis, which could filter out important SNPs that only jointly reveal the genetic risk of a disease (Philips, 2008). Alternatively, multiple logistic regression with the L_1-penalty (Friedman et al., 2010) can be considered, which carries the same challenge in the case of sample size much smaller than the

number of parameters.

In this paper, we develop a novel logistic Analysis of Variance (ANOVA) model for association study that treats the genetic locations, the disease phenotype, and other clinical variables as factors and uses these factors to explain the variability in the logit transformation of the minor allele frequency. Our model expresses the logit transformation as the summation of the SNP effects, effects of the disease phenotype and/or other clinical variables, and the interaction effects. To reduce the number of parameters, we adopt a reduced-rank representation of the matrix of interaction effects. A further reduction of the number of parameters is achieved by assuming that, among the extremely large number of SNPs, only a relatively small number of SNPs have true association with the disease phenotype. Since this is a sensible assumption, it is reasonable to make most of the interaction parameters to be zero when fitting model. We achieve this desired sparsity of parameter vector by employing an L_1-penalty on the SNP association parameters. For determination of the penalty parameters, we propose a modified Bayesian information criterion (BIC) with an extra term included to reflect importance of the interactions that point to associations between SNPs and disease phenotype. The classical BIC (Schwarz, 1978) usually selects a model that is too parsimonious and fails to detect any association under our logistic ANOVA model in GWAS studies (see Figure 1).

In practice, researchers often use an analytical tool to identify several SNPs as the potential biomarkers for further study in biological and clinical validation experiments. The estimated interaction parameters from our logistic ANOVA model can be used to rank the SNPs and the top-ranked SNPs are identified as potential biomarkers. In simulation studies to be reported in Section 3, we found that our method can detect more true biomarkers than the logistic regression. The logistic ANOVA model is general enough to incorporate multi-category phenotype. It can also be used to study associations of several categorical phenotypes with SNP genotypes through forming one multi-category phenotype by considering all combinations of these phenotypes (see Section 4.2).

Our logistic ANOVA model provides a framework for study a phenotype and a large number

of SNPs simultaneously. The reduced-rank representation of the interaction effects in the model can substantially reduce the number of parameters and thus improve statistical efficiency. The idea of dimensionality reduction through a low-rank matrix has been used in the literature in different context for modeling interactions; see e.g., Snee (1982) and Hu et al. (2009). Our proposed model also shares some similarity with the bilinear model described in Hoff (2005). However, fundamental distinctions exist. The goal of Hoff is to model pairs of objects corresponding to a common variable (e.g., measurements of similarity between two units) with the bilinear term modeling the errors, while our goal is to model how two sets of different variables (phenotype and SNP locations) influence the frequency of a binary variable (SNP genotype).

The rest of paper is organized as follows. In Section 2, we introduce the proposed logistic ANOVA model and present details of method. In particular, we define the penalized likelihood and discuss several implementation issues including computational algorithm, selection of the penalty parameters and rank number, and missing data handling. Results of a simulation study are presented in Section 3. In Section 4 we present application of the proposed method to a Multiple Sclerosis data set. Section 5 concludes the paper. The Appendix gives the details of the computational algorithm.

2 Methodology

2.1 The logistic ANOVA model for simultaneously modeling SNPs

We dichotomize the SNP genotype as typically done in the literature (e.g., Cantor et al., 2010). Specifically, we code the genotype as 0 if the original genotype contains only the minor allele; and 1 otherwise. Consider I categories for a discrete phenotype and J SNPs. Let y_{ijk_i} denote the genotype of the SNP at the j^{th} position ($j = 1, \cdots, J$) on a chromosome in the k_i^{th} subject ($k_i = 1, \cdots, K_i$) of the i^{th} phenotype ($i = 1, \cdots, I$). Note that the subscript of K_i indicates that there may be different number of observations for different phenotypes. The mean of the binary variable

y_{ijk_i} is written as

$$E(y_{ijk_i}) = p(\eta_{ij}), \tag{1}$$

where η_{ij} is the canonical parameter of the Bernoulli distribution, and $p(\eta) = (1 + \exp(-\eta))^{-1}$ is the inverse logit link function. The canonical parameter η_{ij} has the following Analysis of Variance (ANOVA) decomposition,

$$\eta_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}, \tag{2}$$

where μ is the grand mean, α_i is the main effect of the i^{th} phenotype, β_j is the main effect of the j^{th} SNP, and γ_{ij} corresponds to the interaction between the i^{th} phenotype and the j^{th} SNP. For identifiability, we impose the following constraints on the parameters $\sum_{i=1}^{I} \alpha_i = 0$, $\sum_{j=1}^{J} \beta_j = 0$, $\sum_{i=1}^{I} \gamma_{ij} = \sum_{j=1}^{J} \gamma_{ij} = 0$. We use the interaction terms γ_{ij} to study the association between the phenotypes and SNPs.

The interaction degrees of freedom, $(I - 1)(J - 1)$, becomes very large when the number of phenotype categories I gets large. To reduce the interaction degrees of freedom, we employ a reduced-rank representation of the matrix of interaction terms (e.g., Johnson and Graybill, 1972; Hu et al., 2009) so that $\gamma_{ij} = \sum_{d=1}^{D} u_{id} v_{jd}$ for $D \leq min(I-1, J-1)$. This reduced-rank representation is directly related to the singular value decomposition of the matrix. The ANOVA decomposition (2) then becomes

$$\eta_{ij} = \mu + \alpha_i + \beta_j + \sum_{d=1}^{D} u_{id} v_{jd}. \tag{3}$$

For model identifiability, we impose the restrictions on the parameters as $\sum_{i=1}^{I} \alpha_i = 0$, $\sum_{j=1}^{J} \beta_j = 0$, $\sum_{i=1}^{I} u_{id} = 0$, $\sum_{j=1}^{J} v_{jd} = 0$, and $\sum_{j=1}^{J} v_{jd}^2 = J - 1$. For $D > 1$, the additional restrictions of $\sum_{i=1}^{I} u_{id} u_{id'} = 0$ and $\sum_{j=1}^{J} v_{jd} v_{jd'} = 0$ are required for $d \neq d'$. In (3), the interaction effect is decomposed as the summation of D multiplicative terms $u_{id} v_{jd}$, where u_{id} and v_{jd} can be interpreted as the contributions to the interaction effect from phenotype i and SNP j, respectively.

We use v_{jd} to identify the SNPs that contribute the most to the phenotype-SNP interaction effect. Specifically, we define the SNP associate index for SNP j as the largest absolute value of $\{v_{jd}\}$, $d = 1, \ldots, D$, and rank the SNPs according to this index.

We refer to our model specified by (1) and (3) as the logistic ANOVA model. When $I = 2$, there are only two choices of D, $D = 0$ (corresponding to no interaction) and $D = 1$. When $I > 2$, use of $D < I - 1$ reduces the number of parameters from $(I - 1)(J - 1)$ of the full model (2) to $D(I + J - D - 2)$. This is a substantial reduction when the number of phenotypes I is large and D is small, making it possible to simultaneously explore the association of thousands of SNPs with the phenotypes.

Under model (1), the individual data generating probability is

$$Pr(Y_{ijk_i} = y_{ijk_i}) = p(\eta_{ij})^{y_{ijk_i}}(1 - p(\eta_{ij}))^{1-y_{ijk_i}} = p(q_{ijk_i}\eta_{ij}), \tag{4}$$

with $q_{ijk_i} = 2y_{ijk_i} - 1$, which equals -1 if $y_{ijk_i} = 0$ and 1 $y_{ijk_i} = 1$. Notice that $p(-\eta) = 1 - p(\eta)$. The log likelihood can be expressed as

$$l = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k_i=1}^{K_i} \log p(q_{ijk_i}\eta_{ij}).$$

Denote $n_{ij}^{+} = \#\{k_i : q_{ijk_i} = 1\}$ and $n_{ij}^{-} = \#\{k_i : q_{ijk_i} = -1\}$. The log likelihood then can be rewritten as

$$l = \sum_{i=1}^{I}\sum_{j=1}^{J}\{n_{ij}^{+} \log p(\eta_{ij}) + n_{ij}^{-} \log p(-\eta_{ij})\}. \tag{5}$$

Denote the intercept matrix $\boldsymbol{\mu} = \mu \mathbf{1}_I \mathbf{1}_J^{\top}$, $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_I)^{\top}$, $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_J)^{\top}$, $\mathbf{u}_i = (u_{i1}, \cdots, u_{iD})^{\top}$, and $\mathbf{v}_j = (v_{j1}, \cdots, v_{jD})^{\top}$. According to (3), the canonical parameter matrix $H = (\eta_{ij})$ has the representation

$$H = \boldsymbol{\mu} + A + B + UV^{\top}, \tag{6}$$

where $A_{I\times J} = \boldsymbol{\alpha}\mathbf{1}_J^{\top}$, $B_{I\times J} = \mathbf{1}_I\boldsymbol{\beta}^{\top}$, $U_{I\times D} = (u_{id})$, and $V_{J\times D} = (v_{jd})$. Using these notations, we write the log likelihood as

$$l(\boldsymbol{\mu}, A, B, U, V) = \sum_{i=1}^{I}\sum_{j=1}^{J}[n_{ij}^{+} \log p(\mu + \alpha_i + \beta_j + \mathbf{u}_i^{\top}\mathbf{v}_j) + n_{ij}^{-} \log p\{-(\mu + \alpha_i + \beta_j + \mathbf{u}_i^{\top}\mathbf{v}_j)\}]. \tag{7}$$

Given the large number of SNPs, it is reasonable to assume that only a relatively small number of SNPs have true association with the phenotype. In our model (3), if v_{jd} is zero for $d = 1, \ldots, D$,

then SNP j plays no role in explaining the phenotype-SNP interaction. We thus borrow the idea from the LASSO regression (Tibshirani, 1996) and introduce a sparsity-inducing penalty to help identify those zeros when maximizing the likelihood function. Similar to LASSO, the L_1-penalty has the form

$$P_\lambda(V) = \sum_{d=1}^{D} \lambda_d \parallel \tilde{\mathbf{v}}_d \parallel_1 = \lambda_1 \sum_{j=1}^{J} |v_{j1}| + \cdots + \lambda_D \sum_{j=1}^{J} |v_{jD}|, \tag{8}$$

where $\tilde{\mathbf{v}}_d$ is the d^{th} column of V and λ_d is the corresponding penalization parameter. We estimate the model parameters via maximizing the penalized log likelihood $l(\boldsymbol{\mu}, A, B, U, V) - n \cdot P_\lambda(V)$ with $n = J \sum_i K_i$. Equivalently, we minimize the following objective function

$$S(\boldsymbol{\mu}, A, B, U, V) = -l(\boldsymbol{\mu}, A, B, U, V) + n \cdot P_\lambda(V) \tag{9}$$

subject to the identifiability constraints.

The logistic ANOVA model specified by (1) and (3) is more general than it appears. The phenotype can be a binary or multi-category phenotype. The multi-category phenotype can be constructed by considering all combinations of possible levels of several factors. In Section 4.2, we present an example that the phenotype is formed by combining a disease status (MS) and a clinical variable (hypertension). When we consider the combination of several factors as the phenotype, the number of categories easily gets large. In such cases, the dimensionality reduction feature of (3) helps significantly reduce the number of parameters to be estimated and makes the model estimation feasible.

2.2 Comparison with logistic regression for the case-control study

The case-control study can be handled as a special case of our logistic ANOVA model. In this case, $I = 2$ and $D = 1$. Taking into account the identifiability constraints, our model (3) reduces to

$$\text{logit}(p_{1j}) = \mu + \alpha_1 + \beta_j + u_1 v_j$$

$$\text{logit}(p_{2j}) = \mu - \alpha_1 + \beta_j - u_1 v_j,$$

which implies that the log-odds ratio at the j^{th} SNP is $\text{logit}(p_{1j}) - \text{logit}(p_{2j}) = 2(\alpha_1 + u_1 v_j)$. A typical logistic regression model of the SNP effect on a phenotype for a case-control study has the form

$$\text{logit}(p_{1j}) = \mu + 2v'_j,$$

$$\text{logit}(p_{2j}) = \mu$$

(the constant 2 is used to simplify notations in later presentation), which gives the log-odds ratio $\text{logit}(p_{1j}) - \text{logit}(p_{2j}) = 2v'_j$. Recall that the v_j's have the sample mean 0 and the sample standard deviation 1, since they satisfy the constraints $\sum_j v_j = 0$ and $\sum_j v_j^2 = J - 1$. When no penalty is used for the maximum likelihood estimation, the two models give the same log-odds ratio when v_js are the standardized versions of v'_js. Indeed, if we let α_1 and u_1 be the sample mean and standard deviation of the v'_js, standardization of v'_js by centering at α_1 and scaling with u_1 gives v_j.

There are two main differences of the two methods: 1). Different quantities are used to rank the SNPs: While the logistic ANOVA uses the standardized log-odds ratios, the logistic regression uses the raw log-odds ratios. Note that standardization does change the ordering of the SNPs. For example, assume there are three SNPs with $v'_1 = 1$, $v'_2 = 2$, $v'_3 = 4$, then $\alpha_1 = 2.33$, $u_1 = 1.53$, and the standardized values are $v_1 = -0.87$, $v_2 = -0.22$, $v_3 = 1.09$. The ordering of the original numbers is $|v'_1| < |v'_2| < |v'_3|$, but after standardization is $|v_2| < |v_1| < |v_3|$. The standardization can be considered as a kind of background adjustment. 2). By using the same penalty parameter for simultaneously estimation of the SNP effects, our logistic ANOVA model allows direct comparison of all SNPs in one framework. These two differences help the logistic ANOVA detect more true biomarkers than the simple logistic regression, as confirmed by the simulation results in Section 3. However, more significant advantage of the logistic ANOVA method over logistic regression is observed in the case of multi-category phenotype, which can be attributed to the dimensionality reduction feature of the reduced-rank model.

2.3 Computational algorithm

Since the penalized log likelihood criterion (9) is not differentiable, the gradient-based methods are not applicable for its minimization. We propose to apply the Majorization-Minimization (MM) algorithm (Hunter and Lange, 2004) to sequentially minimize a quadratic surrogate objective function. A function $g(x|y)$ is said to *majorize* a function $f(x)$ at y if

$$g(x|y) \geq f(x) \quad \text{for all } x \quad \text{and} \quad g(y|y) = f(y). \tag{10}$$

To minimize $f(x)$, the MM algorithm starts from an initial guess $x^{(0)}$ of x and iteratively minimizes $g(x|x^{(m)})$ until convergence, where $x^{(m)}$ is the estimate of x at the m^{th} iteration. The theory for the MM algorithm suggests that the objective function decreases along the iterations and the algorithm is guaranteed to converge to a local minimum.

We develop two functions which majorize the log likelihood term and the penalty term separately as follows. For the log likelihood term $-\log p(x)$, we use the following result of Jaakkola and Jordan (2000) and de Leeuw (2006):

$$-\log p(x) \leq -\log p(y) + \frac{1}{8}[x - y - 4\{1 - p(y)\}]^2 - 2(1 - p(y))^2. \tag{11}$$

where the equality holds when $x = y$. Substituting x and y with $q_{ij}\eta_{ij}$ and $q_{ij}\eta_{ij}^{(m)}$ respectively in (11) yields

$$-\log p(q_{ij}\eta_{ij}) \leq -\log p(q_{ij}\eta_{ij}^{(m)}) + \frac{1}{8}(\eta_{ij} - x_{ij}^{(m)})^2 - 2\{1 - p(q_{ij}\eta_{ij}^{(m)})\}^2, \tag{12}$$

where $q_{ij} = \pm 1$ and

$$\begin{aligned} x_{ij}^{(m)} &= \eta_{ij}^{(m)} + 4q_{ij}\{1 - p(q_{ij}\eta_{ij}^{(m)})\} \\ &= \begin{cases} \eta_{ij}^{(m)} + 4(1 - p(\eta_{ij}^{(m)})) \equiv x_{ij}^{+(m)} & \text{for } q_{ij} = 1 \\ \eta_{ij}^{(m)} - 4(1 - p(-\eta_{ij}^{(m)})) \equiv x_{ij}^{-(m)} & \text{for } q_{ij} = -1. \end{cases} \end{aligned} \tag{13}$$

Ignoring a constant term which does not depend on unknown parameters, the quadratic upper bound of the negative log likelihood is

$$\frac{1}{8}\sum_{i=1}^{I}\sum_{j=1}^{J}\{n_{ij}^{+}(\eta_{ij} - x_{ij}^{+(m)})^2 + n_{ij}^{-}(\eta_{ij} - x_{ij}^{-(m)})^2\} \tag{14}$$

where $\eta_{ij} = \mu + \alpha_i + \beta_j + \mathbf{u}_i^\top \mathbf{v}_j$. To find a majorizing function of the penalty term, we utilize the following majorization relation (Hunter and Li, 2005)

$$|x| \leq \frac{x^2 + y^2}{2|y|}, \quad y \neq 0, \tag{15}$$

from which we obtain

$$P_\lambda(V) \leq \lambda_1 \sum_{j=1}^{J} \frac{v_{j1}^2 + v_{j1}^{(m)2}}{2|v_{j1}^{(m)}|} + \cdots + \lambda_D \sum_{j=1}^{J} \frac{v_{jD}^2 + v_{jD}^{(m)2}}{2|v_{jD}^{(m)}|} \tag{16}$$

Combing (14) and (16) yields the following quadratic upper bound of (9) (up to a constant)

$$\begin{aligned} & g(\boldsymbol{\mu}, A, B, U, V | \boldsymbol{\mu}^{(m)}, A^{(m)}, B^{(m)}, U^{(m)}, V^{(m)}) \\ &= \frac{1}{8} \sum_{i=1}^{I} \sum_{j=1}^{J} [n_{ij}^{+} \{x_{ij}^{+(m)} - (\mu + \alpha_i + \beta_j + \mathbf{u}_i^\top \mathbf{v}_j)\}^2 \\ &\quad + n_{ij}^{-} \{x_{ij}^{-(m)} - (\mu + \alpha_i + \beta_j + \mathbf{u}_i^\top \mathbf{v}_j)\}^2] + \sum_{j=1}^{J} \mathbf{v}_j^\top W_{\lambda,j}^{(m)} \mathbf{v}_j, \end{aligned} \tag{17}$$

where $W_{\lambda,j}^{(m)}$ is a diagonal matrix with the diagonal elements $n\lambda_d / (2|v_{jd}^{(m)}|)$ for $d = 1, \cdots, D$. The above quantity majorizes (9) at $(\boldsymbol{\mu}^{(m)}, A^{(m)}, B^{(m)}, U^{(m)}, V^{(m)})$. The MM algorithm iteratively minimizes the majorizing function until convergence. At the $(m+1)^{th}$ iteration of the MM algorithm, we sequentially minimize the majorizing function given in (17) with respect to μ, α_i, β_j, $\mathbf{u}_i$ and $\mathbf{v}_j$. Luckily, each step is a quadratic optimization and has a closed-form solution. Details of the complete MM algorithm are given in the Appendix.

2.4 Choice of the penalty parameters

The penalty parameters λ_d's are used to control the degrees of penalization. We choose these parameters by minimizing the following modified BIC criterion

$$\mathrm{BIC}_m(\{\lambda_d\}) = -2\, l(\boldsymbol{\mu}, A, B, U, V) + m(\{\lambda_d\})\ \log n - J \cdot m_D(UV^\top), \tag{18}$$

where $m(\lambda_D)$ is a measure of the model degrees of freedom. Following Zou et al. (2007), we define $m(\lambda_D) = (I + J - D - 1)(D + 1) - |V(\lambda_D)|$, where $|V(\lambda_D)|$ is the number of zero elements in V.

Finally, $m_D(UV^\top)$ is the summation of the singular values of $UV^\top$, also called the nuclear norm. The modified BIC criterion is also used for selecting D, the number of multiplicative components in (3).

Note that the ordinary BIC criterion is defined without the rightmost term in (18). Shen and Ye (2002) indicates that the BIC criterion performs poorly for models with a large number of parameters. Our experiments also suggest that the BIC does not work well in our setting. Specifically, we observed that the BIC often chooses the no-association model ($D = 0$), as opposed to the existing results of presence of strong association between disease and SNPs in some published association studies. For this reason, we modify the BIC criterion by introducing the additional term of the nuclear norm $m_D(UV^\top)$. It is designed to stress the importance of the interaction terms which point to possible associations between the phenotype and SNPs. Our empirical results shown later demonstrate the good performance of this modified BIC criterion.

2.5 Handling missing data

In real application, missing observations of SNP genotypes are often encountered. We take the following approach to handle a missing genotype observation y_{ijk_i}. Let $\mathcal{M} = \{(i, j, k_i)|y_{ijk_i} \text{ is missing}\}$ be an index set of missing values. In the presence of missing data, we minimize the following modified version of (9)

$$S_{obs}(\boldsymbol{\mu}, A, B, U, V) = -l_{obs}(\boldsymbol{\mu}, A, B, U, V) + n \cdot P_\lambda(V), \tag{19}$$

where

$$l_{obs}(\boldsymbol{\mu}, A, B, U, V) = \sum_{(i,j,k_i)\notin\mathcal{M}} \log p\{q_{ijk_i}(\mu + \alpha_i + \beta_j + \mathbf{u}_i^\top \mathbf{v}_j)\}. \tag{20}$$

With some slight modifications, the MM algorithm developed in Section 2.3 still applies to minimize the penalized likelihood criterion (19). When the data are complete, we have $n_{ij}^+ + n_{ij}^- = K_i$ for all j; this does not hold any more when there are missing data. Details of the modification of the MM algorithm are presented in the Appendix.

3 Simulation studies

We conducted some simulation studies to assess performance of the proposed method in terms of accuracy of biomarker detection. We report the results for binary phenotypes and multi-category phenotypes separately in two subsections.

3.1 Binary phenotype

We considered two simulation setups with binary phenotypes. The first setup generated the data from the logistic ANOVA model, the second setup generated the data by sub-sampling the MS data set in Section 4.

Simulation Setup 1. To allow correlation in the generated data, we first generate random numbers from multivariate normal distribution and then dichotomize them to get binary phenotype data. The steps for generating data are

1. Set the canonical parameters as $\eta_{ij} = \mu + \alpha_i + \beta_j + u_i v_j$.
2. For $i = 1, \ldots, I$ and $k = 1, \ldots, K$, independently generate the vector $(X_{ijk}, j = 1, \ldots, J)$ from $MVN(\mathbf{0}, \Sigma_{I\times I})$.
3. Assign $y_{ijk} = 1$ if $\Phi(X_{ijk}) < e^{\eta_{ij}}/(1 + e^{\eta_{ij}})$, and 0 otherwise, for $i = 1, \ldots, I$, $k = 1, \ldots, K$, where $\Phi(\cdot)$ is the c.d.f. of a standard normal distribution.

We used $I = 2$, $J = 10,000$, $K = 500$, and set the parameters as $\mu = 0$, $(\alpha_1, \alpha_2) = (-1, 1)$, and $(u_1, u_2) = (-0.3, 0.3)$; β_j were randomly drawn from the standard normal distribution and v_j were randomly drawn from $N(0, 1/3)$ where the 50 largest values of $|v_j|$ were regarded as significant biomarkers. The correlation matrix Σ was set to mimic the real data in Section 4. Specifically, we set the diagonal elements of Σ to be all 1's, and set the off-diagonal element at (i, j) position to be $\rho^{|i-j|}$ if $|i - j| \leq 7$, and 0 if otherwise. We considered two settings of ρ, $\rho = 0.3$ and 0.5.

To each simulated data set, we applied our logistic ANOVA method, fitting with and without penalization. The modified BIC criterion suggested selection of $D = 1$. We used the absolute val-

ues of the fitted v_{j1} to rank the SNPs and count the number of times that the significant biomarkers are identified. We also applied the simple logistic regression, which deals with the biomarkers one at a time, and the L_1-penalized logistic regression (Friedman et al., 2010). For both regression approaches, we ranked the SNPs based on the estimated regression coefficients. We checked the top 50 SNPs selected by each method and counted how many of them are the significant biomarkers. The summary statistics of the results based on 500 simulation runs is given in Table 1. We observe that the logistic ANOVA outperforms the logistic regression approach for both settings of the correlation coefficient. The reason that L_1-penalized logistic regression does not work better than the simple logistic regression is that it tends to select models that are too parsimonious.

Simulation Setup 2. In this setup we generated data by suitably subsetting the MS data in Section 4.1, which contains 1,803 subjects with 34,282 SNP locations. We randomly selected 1,000 out of all subjects as the samples in the simulation study. We used the values of $|\hat{v}_j|$ obtained by fitting the logistic ANOVA as in Section 4.1 to decide which SNPs to include in the data set. We fixed 10,000 "null" SNPs which correspond to the small values of $|\hat{v}_j|$. We then randomly selected 50 "null" SNPs and replaced them with the top 50 SNPs (corresponding to the largest values of $|\hat{v}_j|$), which was treated as the true biomarkers in this simulation study. This sub-sampling scheme from the MS data set generated a "simulated" data set containing 1,000 subjects with 10,000 SNP locations. We noticed that many adjacent SNPs are contained in the set of 10,000 SNPs, thus to some extent the dependency structure among the SNPs is preserved. For example, in one simulated data set, the first 10 among the 10,000 SNPs are located at the positions of 1, 2, 5, 9, 10, 11, 12, 13, 14, and 16 in the MS data set, which are in close vicinity of each other.

The four methods considered in Setup 1 were applied to each of the 500 data sets simulated according to the scheme described in the previous paragraph. The BIC_m criterion suggested $D = 1$ for the logistic ANOVA model for all data sets. An example of the BIC_m curve for a simulated data set is shown in the lower right panel of Figure 1. It is obvious that a minimum value of BIC_m shown in this panel is achieved at $\log(\lambda) = -14$ and $D = 1$, as opposed to BIC, shown on the lower left panel, which fails to identify any associations since it suggests $D = 0$. We

identified top 50 SNPs based on each of the four methods, and recorded the number of detected true biomarkers. The ranking of the SNPs was determined in the same way as in simulation Setup 1. The summary statistics of the correct detection based on 500 simulation runs are summarized in Table 2. The logistic ANOVA with penalization performs the best, while the logistic ANOVA without penalization and the simple logistic regression have similar performance. The L_1-penalized logistic regression with the penalty parameter determined by the commonly used 10-fold cross validation yields the least accuracy of biomarker detection. We also considered the BIC criterion for the L_1-penalized logistic regression and obtained even worse results than the cross validation.

For this simulation setup, the estimated α_1 value from the real data is 0.00085, which is close to 0. According to the discussion giving in Section 2.2, we expect that the results by the logistic ANOVA without penalty and the simple logistic regression should be similar. This is indeed the case: For 500 simulated data sets, the average number of overlapping SNPs selected by both methods is 46.78 among top 50 SNPs with a standard error of 0.07.

3.2 Multi-category phenotype

We modify the two simulation setups in Section 3.1 to generate data sets with multi-category phenotypes.

Simulation Setup 3. This setup is the same as Setup 1 with some slight modifications to make $I = 6$ categories. The parameters $\boldsymbol{\alpha}$ and $\mathbf{u}$ are not six-dimensional vectors. We set their true values to be $\boldsymbol{\alpha} = (-1, -0.6, -0.3, 0.3, 0.6, 1)$, and $\mathbf{u} = (-0.5, -0.3, -0.1, 0.1, 0.3, 0.5)$. We also changed the number of subject for each category to be $K = 300$. The other parameters are unchanged. We generated 500 data sets for this setup.

To each simulated data set, we applied our logistic ANOVA method, fitting with and without penalization. We used the absolute values of the fitted v_{j1}'s to rank the SNPs and count the number of times that the significant biomarkers are identified. We also applied simple multinomial logistic regression, which deals with the biomarkers one at a time, and the L_1-penalized multinomial

logistic regression. The six-category multinomial logistic regression yields five coefficients per SNP location, and we used the maximum of the absolute values of the coefficients as a measure for that location when ranking the SNPs. We also applied the L_1-penalized multinomial logistic regression but found that it frequently failed to fit the data due to many fitted values of exact 0's or 1's and thus we did not include it in our comparison. We checked the top-ranked 50 SNPs by each method and counted how many of them are the significant biomarkers. The summary statistics of the results based on 500 simulation runs is given in Table 3. We observe that the logistic ANOVA outperforms the logistic regression approach for both settings of the correlation coefficient.

Simulation Setup 4. In this setup we generated data by suitably subsetting the MS data as we did in simulation Setup 2. We created a phenotype using combinations of hypertension status and MS status. For hypertension status, we merged the "normal" and "pre-hypertension" into a common category because they were shown to be alike according to Section 4.2. This leads to a 4-category phenotype. For each simulation run, we randomly selected 500 subjects consisting of 150 "non-MS and non-hypertension", 100 "non-MS and hypertension", 150 "MS and non-hypertension", and 100 "MS and hypertension" cases. Similar to simulation Setup 2, we considered 10,000 SNPs, among which 50 are the true biomarkers.

We applied four methods used in simulation Setup 3 to each data set and identified top-ranked 50 SNPs by each method. Table 4 reports the summary statistics of the number of detected true biomarkers by each of the methods, based on 500 simulation runs. We observe that the logistic ANOVA with penalty detects the true biomarkers substantially more than the other methods. It is also clear from the table that penalization indeed benefits biomarker detection, since the logistic ANOVA without penalty performs much inferior to its penalized version.

It is interesting to compare the results in Setup 2 and the current simulation setup to find out differences when a binary phenotype is changed to a multi-category phenotype. We observe that the improvement of the logistic ANOVA over the simple logistic regression is more substantial for the 4-category phenotype case (Table 4) than the binary phenotype case (Table 3). This is not surprising: In the binary phenotype case, we show in Section 2.2 that the two methods are

closely related to each other and thus we do not expect they give substantial different results. For the multi-category case, the dimensionality reduction introduced by the reduced-rank model can make a difference. Compared with the binary phenotype case, the fitted logistic ANOVA model for the 4-category case needs only four additional parameters—two for α_i's and two for u_i's. In contrast, the simple multinomial logistic regression requires triple additional number of parameters in comparison to the simple logistic regression for Setup 2. We also observe that the L_1-penalized multinomial logistic regression in Setup 4 suffers less from the problem of low detection as the L_1-penalized binary logistic regression in Setup 2. This may be due to using the maximum absolute value of multiple coefficients for each SNP as the final association index. However, its number of detections is still less than half of that by penalized logistic ANOVA.

3.3 Zero-Mean log odds ratios

The observed significant difference in performance of the logistic ANOVA (without penalization) and the simple logistic regression for Setup 1 may be explained by the fact that the log odds ratios do not have mean zero in the simulation model. We ran a simulation study that modifies Setup 1 by letting $\alpha_i = 0$ so that the log odds ratios have mean zero. Similar to Table 1, Table 5 reports the summary of results for detecting the true biomarkers by three different methods.

As explained in Section 2.2, the logistic ANOVA model (without penalization) and the simple logistic regression model are identical but use different parametrizations. The simple logistic regression ranks the SNPs using the absolute values of the log odds ratios, while the logistic ANOVA using the absolute values of the standardized log odds ratios. Since $\alpha_i = 0$, it seems that standardization of the log odds ratios is not necessary and one wonders whether it may hurt the performance of logistic ANOVA. Table 5 shows that the logistic ANOVA without penalization performs slightly better than the simple logistic regression (i.e., standardization does not have a negative effect), but the difference is very small and statistically not significant. On the other hand, introducing sparsity regularization in the logistic ANOVA has a bigger effect than the standardization of the log odds

ratios.

Table 5 also gives the results for a simulation study that modifies Setup 3 by letting all $\alpha_i = 0$. We observe that for this case of multi-category phenotype, logistic ANOVA model clearly shows better performance than the simple multinomial logistic regression. This significant improvement can be contributed to the dimensionality reduction feature of the logistic ANOVA.

4 A Multiple Sclerosis study

We studied a Multiple Sclerosis (MS) data set obtained from the National Institute of Neurological Disorders and Stroke (NINDS) to demonstrate the applicability of the proposed methodology. The data set contains 1,803 subjects (864 controls and 939 cases) and is available through dbGaP accession number phs000171.v1.p1. The genotyping data were generated using Illumina HumanHap 550 BeadChip platform. MS is a disease that affects the brain and spinal cord. This disease is commonly triggered by the attack of a virus or due to genetic defects. Some existing studies (Ramagopalan et al., 2009; McElroy et al., 2010) pointed out that a region known as the Major Histocompatibility Complex (MHC) on chromosome 6 is believed to be highly susceptible to MS. Thus, we focus our study on chromosome 6 which contains 34,282 SNPs to search for the biomarkers that are possibly associated with the phenotype of MS.

4.1 The association of SNPs with Multiple Sclerosis status

We first investigated the association between SNPs and the MS disease using our proposed method. For this case-control study, $I = 2$ in our logistic ANOVA model and there are two possible choices of D, $D = 0$ and $D = 1$. To deal with the missing values in the data set, we used the procedure described in Section 2.5. The values of the modified BIC defined in (18), BIC_m, are plotted against the log-transformed values of the penalty parameter λ for $D = 0, 1$ in the upper right panel of Figure 1. Since there is no interaction term for $D = 0$, the values of BIC_m is a constant in this case. It is clear from the plot that $D = 1$ is preferred to $D = 0$. For comparison, we also show a

similar plot of BIC in the upper left panel. We see that BIC_m selects $D = 1$ with a clear minimum at $\log(\lambda) = -15$. In contrast, BIC identifies $D = 0$ which points to no interactions between the phenotype and SNPs. The bottom two panels of Figure 1 present the BIC and BIC_m plot for two simulated data sets generated using the logistic ANOVA model specified by (1) and (3) with an interaction term. We observe that the BIC and BIC_m for the simulated data behave similarly as for the real data. Our method with $D = 1$ obtained the u_i estimates of the controls and the cases as $(\hat{u}_1, \hat{u}_2) = (-1.1603, 1.1603)$.

In the ANOVA model (3), v_{jd} represents the association of SNP j with the phenotype, which is the MS status in this study. We ranked the SNPs according to the absolute values of v_{j1} ($D = 1$). The top-ranked SNPs are identified by our method as highly associated with the MS status. We compared our results with existing results of MS related association studies available in a web-based NIH database, Phenotype-Genotype Integrator (*PheGenI*) (http://www.ncbi.nlm.nih.gov/gap/PheGenI). This facility merges the results from GWAS catalog data with several databases housed at the National Center for Biotechnology Information, including Gene, dbGaP, OMIM, GTEx and dbSNP. According to this database, 138 SNPs were identified to be significantly related to MS by ten research papers using the p-value threshold of 10^{-6}, but only 3 SNPs were commonly detected by any two publications, and none were commonly detected by more than two publications. Moreover, six out of ten publications detected only one SNP, while one other paper detected about 100 SNPs. This large variation among published work indicates the difficulty of the problem and lack of stability across existing methods. Nonetheless, we tried to position our top candidate SNPs in the database for comparison purpose. The top 50 SNPs ranked by our method have 9 in common with the significant SNPs (p-value $< 10^{-6}$) chosen by the database *PheGenI*, while the top 100 and 150 SNPs by our method have respectively 16 and 25 in common with the significant SNPs in *PheGenI*. Table 6 shows that the minor allele frequencies of the top 50 SNPs spread out between 15% and 50% except for one lower than 15%.

We compared our logistic ANOVA method with the simple logistic regression that examines one SNP at a time. We obtained two sets of SNPs, each of which consists of the top 50 ranked

SNPs by one method. For the simple logistic regression, there are typically two ways to define the association index for ranking SNPs, one is the absolute value of the regression coefficient, the other is the p-value for testing if the coefficient is zero or not. The top and middle panels of Figure 2 show respectively the $\hat{v}_{j1}$ produced by the logistic ANOVA and the negative log-10 transformed p-values obtained by the simple logistic regression, against the SNP locations. The detected 50 SNPs are represented by the circles in blue for both methods. We observe that 41 SNPs are shared by the two sets. On the other hand, if the regression coefficient from the logistic regression is used to rank the SNPs, the set of top 50 SNPs is identical to that selected by the logistic ANOVA.—This is only a coincidence, the ranking of these 50 SNPs are not exactly the same by the two methods, indicating that the two sets are different if a different number (say, 40) of top SNPs are considered. A clear common feature of the two panels is that most of the top biomarker candidates cluster around the 8000^{th} location index. As the most distinction, the logistic ANOVA identifies 6 SNPs to the right side of this cluster. To see if these 6 SNPs have any real interpretations or implications of the biological association with MS, further biological validations are required.

A primary goal of biomarker detection is to construct a classifier for making predictions of the disease phenotype that can be potentially useful for facilitating medicine prevention and treatment. We evaluated the performance of using detected SNPs to predict MS status. The details of our evaluation procedure is as follows. We divided all the subjects into two sets, one training set and one test set. The training set consisted of n_t randomly chosen subjects and was used for SNP selection and for training a classification method. The test set consisted of the remaining $(1803-n_t)$ subjects and was used for calculating the prediction error. On the training set, we first selected 50 SNPs that are most associated with MS status using either the logistic ANOVA or the simple logistic regression (using the absolute value of the regression coefficient), then used the genotype of the selected 50 SNPs as the predictors for a classification method. We used the "random forest" method implemented in R package randomForest as our classification method. The prediction error rate is defined as the proportion of the test subjects that have incorrect prediction of the MS status comparing to the known true disease status.

We considered three settings of random split of data into training and test sets, consisting respectively $n_t = 600$, 1000, 1400 training subjects. For each setting, we repeated the above described procedure 300 times. Figure 3 shows the scatterplot of prediction error rates when the simple logistic regression and logistic ANOVA are used for SNP selection. For all three settings of the training set size, the average prediction error of the logistic ANOVA is smaller than that of the simple logistic regression; the paired-t test gives a p-value less than 10^{-15} when the training set size is 600 and 1000, and less than 10^{-5} when the training set size is 1400.

We notice that the average prediction error rate for the logistic ANOVA model is about 0.38 for all three training set size. The fairly high error rate is in concordant with the current medical finding that the disease of MS is difficult to diagnose and test, according to the statement posted by National Multiple Sclerosis Society at their website (www.nationalmssociety.org) and by Poser and Brinar (2001).

4.2 Association among SNPs, MS, and hypertension

Recent work (e.g., Platten et al., 2009) made some interesting discovery of the linkage of hypertension and MS such that treatment of the former could obstruct the development of the latter. Thus, we include the hypertension information in our study to detect possible interactions among MS, hypertension, and SNPs, and to investigate if the identified biomarkers via incorporating hypertension information can potentially improve disease diagnosis.

To apply the proposed method, we need to categorize the numerical valued blood pressure measurements. Following the standard guideline of classifying hypertension stages based on systolic blood pressure, we defined a nominal variable taking the values of “normal”, “pre-hypertension” (with a systolic pressure from 120 to 139), and “hypertension”. We then created a categorical phenotype with 6 categories consisting of all possible combinations of the MS status and 3 hypertension stages. With this categorization of the MS status and hypertension stages, the data set contains 379 cases of non-MS and normal blood pressure (NN), 354 cases of non-MS and pre-hypertension

(NP), 131 cases of non-MS and hypertension (NH), 436 cases of MS and normal blood pressure (MN), 378 cases of MS and pre-hypertension (MP), and 125 cases of MS and hypertension (MH).

When applying the proposed method, the missing values were handled using the method described in Section 2.5. The modified BIC defined in (18) was used to select the penalty parameters and it also suggested $D = 1$. We display the plot of the SNP association index v_{j1}'s along with the location indexes in the bottom panel of Figure 2, in which the top-ranked 50 SNPs are again represented by the blue circles. The locations of the top-ranked 50 SNPs associated with the 6-category phenotype are quite distinct from those for binary phenotype of MS status; only two SNPs are shared by the two sets of 50 SNPs. We also applied the proposed method to study the association of SNPs with hypertension status (treated as a 3-category phenotype) and identified a quite different set of SNPs. There are four common SNPs shared by the SNP set for hypertension and the set for MS and hypertension. These results are not necessarily surprising because MS and hypertension are two different diseases, wherein the former is much more difficult to diagnose than the latter.

For comparison purpose, we also ranked the SNPs using the regression coefficients from fitting the simple multinomial logistic regression for the 6-category phenotype. The ranking is based on the maximum of the absolute values of the 5 regression coefficients for the 6-category multinomial distribution. We found that the set of top-ranked 50 SNPs using multinomial logistic regression is very different from that using the logistic ANOVA—there are only 12 SNPs shared by the two sets.

Table 7 shows the estimated values of u_{i1}'s, contributions of the 6 categories to the interaction effect. We observe that large absolute values of u_{i1} are present only for NH and MH, while the trivial values are shown for all the other four categories of the phenotype. The results immediately suggest possible three-way interactions between SNPs, MS, and hypertension stages as elaborated below. Note that the interaction effect in the fitted model is represented by the multiplicative term $u_{i1}v_{j1}$, which is trivial when one of the two multiplicative terms is close to 0. From Table 7, the estimates of u_{i1} corresponding to NN, NP, MN, and MP are rather small (in absolute value),

suggesting that there is weak association between SNPs and MS in the presence of normal blood pressure and pre-hypertension. On the contrary, the clear contrast between NH (with large negative u_{i1}) and MH (with large positive u_{i1}) indicates the existence of nontrivial association between SNPs and MS only in the presence of hypertension. Interestingly, this finding is consistent with the biological results obtained by Platten et al. (2009), which point out that the association between autoimmunity (exemplified by MS) and genetic factors tends to more likely occur for hypertension patients.

We also did a prediction exercise that uses only the genotype information of the top-ranked 50 SNPs. We created a binary response variable whose value is set to 1 if a patient has both MS and hypertension, and 0 otherwise. We used both the logistic ANOVA and the simple multinomial logistic regression to select the SNPs. The genotypes corresponding to the 50 selected SNPs were used as predictors. As in Section 4.1, we randomly split the data into a training and a test set, used the training set for SNP selection and to train the "random forest", and used the test set for computing the prediction error. We considered three settings of the training set size, namely 600, 1000, and 1400. For each setting, we repeated the procedure 300 times. The scatterplots of the prediction error rates by the two methods for the 300 random splits are presented in Figure 4. We observed that using the logistic ANOVA to select SNPs gives smaller average prediction error rate than using the simple multinomial logistic regression; the paired-t test has a p-value less than 10^{-15} for all three settings of the training set size.

The error rate for predicting MS and hypertension is smaller than that for predicting MS alone (comparing previous subsection). However, this smaller error rate might be explained by the small proportion of cases of MS and hypertension in the data set. We did another experiment in which we have balanced samples from each of the six phenotype. Specifically, we subsampled 150 cases from each of the six phenotypes to compose a data set of 750 subjects. Then, we randomly took 600 subjects as training data, the rest 150 subjects as test data, and repeated the same prediction exercise described in the previous paragraph. For 500 runs, the average test set prediction error for predicting MS and hypertension is 12.9% with a standard error less than 0.15%. This error rate

cannot be entirely explained by the proportion of MS and hypertension subjects in the data, which is 16.7%.

5 Concluding remarks

In this paper we propose a novel method for studying the association between a categorical phenotype and genotypes of a large number of SNPs. The core of our framework is an ANOVA decomposition of the logit transformation of the minor allele frequency of the SNPs, where we treat the categorical phenotype (such as disease status and other important clinical variables) and the SNP locations as two factors for explaining the variability present in the logit transformation. By permitting a reduced-rank representation of the interaction effect for dimensionality reduction and SNP selection through the L_1-penalized maximum likelihood estimation, our logistic ANOVA model can simultaneously deal with a large number of SNPs in one framework. Our simulation studies and real data analysis demonstrated that the logistic ANOVA outperforms alternative methods in capturing the true associations between SNPs and a phenotype, and it has more promise in the multi-category case than in case-control studies.

The proposed logistic ANOVA method can handle multi-category disease phenotype in addition to a binary phenotype in case-control studies. In this paper, our focus has been on ranking the SNPs for the association of their genotype with a disease phenotype and some other important clinical factors. One future research direction is to develop a formal procedure to assess the statistical significance of the associations under the proposed model. Another direction is to consider extensions of the proposed method to handle the continuous phenotype such as survival time which is often of interest in real applications. Moreover, the proposed method can be modified to handle three-categorical genotype data by using the multinomial logit link function instead of the binary logit link, for which model formulation and estimation require substantial further investigation.

Appendix: The MM algorithm

Define $\eta_{ij}^{(m)} = \mu^{(m)} + \alpha_i^{(m)} + \beta_j^{(m)} + \mathbf{u}_i^{(m)\top}\mathbf{v}_j^{(m)}$. The updates of μ, α_i, and β_j at the $(m+1)^{th}$ iteration of the MM algorithm are given by certain weighted averages:

$$\mu^{(m+1)} = \sum_{i=1}^{I}\sum_{j=1}^{J}[n_{ij}^{+}\{x_{ij}^{+(m)} - \eta_{ij}^{(m)} + \mu^{(m)}\} + n_{ij}^{-}\{x_{ij}^{-(m)} - \eta_{ij}^{(m)} + \mu^{(m)}\}]\Big/\Big(J\sum_{i=1}^{I}K_i\Big) \tag{21}$$

$$\alpha_i^{(m+1)} = \sum_{j=1}^{J}[n_{ij}^{+}\{x_{ij}^{+(m)} - \eta_{ij}^{(m)} + \alpha_i^{(m)}\} + n_{ij}^{-}\{x_{ij}^{-(m)} - \eta_{ij}^{(m)} + \alpha_i^{(m)}\}]\Big/(JK_i) \tag{22}$$

$$\beta_j^{(m+1)} = \sum_{i=1}^{I}[n_{ij}^{+}\{x_{ij}^{+(m)} - \eta_{ij}^{(m)} + \beta_j^{(m)}\} + n_{ij}^{-}\{x_{ij}^{-(m)} - \eta_{ij}^{(m)} + \beta_j^{(m)}\}]\Big/\sum_{i=1}^{I}K_i. \tag{23}$$

When there are missing data, the denominators in (21)–(23) should be replaced respectively by $\sum_{i=1}^{I}\sum_{j=1}^{J}(n_{ij}^{+} + n_{ij}^{-})$, $\sum_{j=1}^{J}(n_{ij}^{+} + n_{ij}^{-})$, and $\sum_{i=1}^{I}(n_{ij}^{+} + n_{ij}^{-})$.

Let $x_{ij}^{*\pm(m)} = x_{ij}^{\pm(m)} - \mu^{(m)} - \alpha_i^{(m)} - \beta_j^{(m)}$ and let $\mathbf{x}_i^{*\pm(m)} = (n_{i1}^{\pm}x_{i1}^{*\pm(m)}, \cdots, n_{iJ}^{\pm}x_{iJ}^{*\pm(m)})^{\top}$. From (17), $\mathbf{u}_i^{(m+1)}$ minimizes w.r.t. $\mathbf{u}_i$ the following

$$\sum_{j=1}^{J}[n_{ij}^{+}\{x_{ij}^{*+(m)} - \mathbf{u}_i^{\top}\mathbf{v}_j^{(m)}\}^2 + n_{ij}^{-}\{x_{ij}^{*+(m)} - \mathbf{u}_i^{\top}\mathbf{v}_j^{(m)}\}^2]. \tag{24}$$

Setting to zero the derivatives w.r.t. $\mathbf{u}_i$ yields

$$\mathbf{u}_i^{(m+1)} = (V^{(m)\top}V^{(m)})^{-1}V^{(m)\top}(\mathbf{x}_i^{*+(m)} + \mathbf{x}_i^{*-(m)})/K_i, \tag{25}$$

where $V^{(m)}$ is the matrix whose columns are $\mathbf{v}_j^{(m)}$'s. When there are missing data, the updating formula for $\mathbf{u}_i^{(m+1)}$ is

$$\mathbf{u}_i^{(m+1)} = (V^{(m)\top}\Omega_i V^{(m)})^{-1}V^{(m)\top}(\mathbf{x}_i^{*+(m)} + \mathbf{x}_i^{*-(m)}), \tag{26}$$

where Ω_i is the diagonal matrix whose j^{th} diagonal element is $n_{ij}^{+} + n_{ij}^{-}$, $j = 1, \ldots, J$.

Let $\mathbf{x}_j^{\dagger\pm(m)} = (n_{1j}^{\pm}x_{1j}^{*\pm(m)}, \cdots, n_{Ij}^{\pm}x_{Ij}^{*\pm(m)})^{\top}$. Then, $\mathbf{v}_j^{(m+1)}$ minimizes w.r.t. $\mathbf{v}_j$

$$\frac{1}{8}\sum_{i=1}^{I}\{n_{ij}^{+}(x_{ij}^{\dagger+(m)} - \mathbf{u}_i^{(m)\top}\mathbf{v}_j)^2 + n_{ij}^{-}(x_{ij}^{\dagger-(m)} - \mathbf{u}_i^{(m)\top}\mathbf{v}_j)^2\} + n\,\lambda_d\sum_{j=1}^{J}\frac{v_{jd}^2}{2|v_{jd}^{(m)}|}. \tag{27}$$

Setting to zero the derivatives w.r.t. $\mathbf{v}_j$ gives

$$\mathbf{v}_j^{(m+1)} = \left(U^{(m)\top}\Lambda U^{(m)} + W_{\lambda,j}^{(m)}\right)^{-1} U^{(m)\top}(\mathbf{x}_j^{+\dagger(m)} + \mathbf{x}_j^{-\dagger(m)}), \tag{28}$$

where $U^{(m)}$ is the matrix whose columns are $\mathbf{u}_i^{(m)}$, Λ is the diagonal matrix with the diagonal elements K_i, and $W_{\lambda,j}^{(m)}$ is the diagonal matrix with the diagonal elements $4n\lambda_d/(|v_{jd}^{(m)}|)$ for $d = 1, \cdots, D$. When there are missing data, the i^{th} diagonal element of Λ should replaced by $n_{ij}^{+} + n_{ij}^{-}$, $i = 1, \ldots, I$.

The steps of the complete MM algorithm are given below.

1. Initialize $\mu^{(1)}$, $\boldsymbol{\alpha}^{(1)} = (\alpha_1^{(1)}, \cdots, \alpha_I^{(1)})^\top$, $\boldsymbol{\beta}^{(1)} = (\beta_1^{(1)}, \cdots, \beta_J^{(1)})^\top$, $\mathbf{u}_i^{(1)} = (u_{i1}^{(1)}, \cdots, u_{iD}^{(1)})^\top$, and $\mathbf{v}_j^{(1)} = (v_{j1}^{(1)}, \cdots, v_{jD}^{(1)})^\top$. Set $m = 1$.

2. Compute n_{ij}^{+} and n_{ij}^{-} as in (5) and compute $x_{ij}^{+(m)}$ and $x_{ij}^{-(m)}$ from (13).

3. Compute $\mathbf{x}_i^{+*(m)}$, $\mathbf{x}_i^{-*(m)}$, $\mathbf{x}_j^{+\dagger(m)}$, and $\mathbf{x}_j^{-\dagger(m)}$ in (25) and (28).

4. Update μ to $\mu^{(m+1)}$ using (21).

5. Update $\boldsymbol{\alpha}_i$ to $\boldsymbol{\alpha}_i^{(m+1)}$ using (22). Then $\boldsymbol{\alpha}_i$'s are centered to 0 for $i = 1, \cdots, I$.

6. Update $\boldsymbol{\beta}_j$ to $\boldsymbol{\beta}_j^{(m+1)}$ using (23). Then $\boldsymbol{\beta}_j$'s are centered to 0 for $j = 1, \cdots, J$.

7. Update $\mathbf{u}_i$ to $\mathbf{u}_i^{(m+1)}$ using (25). Then each column of $U^{(m+1)} = (\tilde{\mathbf{u}}_d^{(m+1)})$ is centered to 0 for $d = 1, \cdots, D$.

8. Compute the QR decomposition of $U^{(m+1)} = \mathbf{QR}$. Reset $U^{(m+1)}$ to $\mathbf{Q}$ so that the orthogonal constraints are satisfied.

9. Compute $W_{\lambda,j}^{(m)}$ and update $\mathbf{v}_j$ to $\mathbf{v}_j^{(m+1)}$ using (27). Then each column of $V^{(m+1)} = (\tilde{\mathbf{v}}_d^{(m+1)})$ is centered to 0 for $d = 1, \cdots, D$, and $\|\tilde{\mathbf{v}}_d^{(m+1)}\|_2 = \sqrt{\sum_j \{v_{jd}^{(m+1)}\}^2}$ is computed.

10. Compute the QR decomposition of $V^{(m+1)} = \mathbf{QR}$. To satisfy the orthogonal constraints, $V^{(m+1)}$ is replaced by $\mathbf{Q}$, and then each of the d^{th} column of $\mathbf{Q}$ is multiplied by $\|\tilde{\mathbf{v}}_d^{(m+1)}\|_2$.

11. Repeat steps 2 through 10 with m replaced by $m + 1$ until convergence.

12. After convergence, multiply $u_{id}^{(m+1)}$ by the scaling factor $\|\tilde{\mathbf{u}}_{id}^{(m+1)}\|_2 = \sqrt{\sum_i \{u_{id}^{(m+1)}\}^2}$, and also multiply $v_{jd}^{(m+1)}$ by $\sqrt{J-1}/\|\tilde{\mathbf{v}}_d^{(m+1)}\|_2$.

In this algorithm, we use the QR decomposition to rotate U and V such that the orthogonal constraints are satisfied when $D > 1$. Steps 7 and 9 are used to satisfy $\sum_i u_{id} = 0$ and $\sum_j v_{jd} = 0$. Step 12 is used to re-normalize u_{id}'s and v_{jd}'s such that $\sum_j v_{jd}^2 = J - 1$. Notice that we choose to center and scale v_{jd}'s while only center u_{id}'s. This strategy enables us to obtain data-dependent estimates of u_{id}s.

We suggest the following scheme for obtaining the initial values for the algorithm. Use the logit transformation of the overall mean of the observations $\{y_{ijk}\}$'s for $\mu^{(1)}$. Define by $\mathbf{Y}$ the $I \times J$ matrix whose (i, j)-element is the average of the genotype observations for the i^{th} phenotype and the j^{th} SNP. We let $\boldsymbol{\alpha}^{(1)}$ and $\boldsymbol{\beta}^{(1)}$ to be the logit transformation of the row means and column means of $\mathbf{Y}$. The initial values of $U^{(1)}$ and $V^{(1)}$ are obtained respectively as the left and right singular vectors obtained from the singular value decomposition of the element-wise logit transformed $\mathbf{Y}$ after subtracting the main effects. Our experiments showed that using this initialization scheme requires fewer iterations and the algorithm converges to an empirically more reasonable local minimum of the objective function than using random initialization.

Our algorithm implementation using the programming language R is reasonably fast. For example, applying the algorithm to a simulated data containing 1,000 subjects and 10,000 SNPs takes about 5 minutes to obtain parameter estimates in the standard system of Dell CPU 3.0 GHz, RAM 4GB.

References

Cantor, R. M., Lange, K. and Sinsheimer, J. S. (2010). Prioritizing GWAS results: A review of statistical methods and recommendations for their application, *American Journal of Human*

Genetics **86**(1): 6 – 22.

Chatterjee, N., Kalaylioglu, Z., Moslehi, R., Peters, U. and Wacholder, S. (2006). Powerful multilocus tests of genetic association in the presence of gene-gene and gene-environment interactions, *The American Journal of Human Genetics* **79**: 1002 – 1016.

Chen, H., Chen, Y., Zhao, Y., Fan, W., Zhou, K., Liu, Y., Zhou, L., Mao, Y., Wei, Q., Xu, J. and Lu, D. (2011). Association of sequence variants on chromosomes 20, 11, and 5 (20q13.33, 11q23.3, and 5p15.33) with glioma susceptibility in a Chinese population, *American Journal of Epidemiology* **173**(8): 915 – 922.

de Leeuw, J. (2006). Principal component analysis of binary data by iterated singular value decomposition, *Computational Statistics & Data Analysis* **50**(1): 21 – 39.

Egan, K. M., Thompson, R. C., Nabors, L. B., Olson, J. J., Brat, D. J., LaRocca, R. V., Brem, S., Moots, P. L., Madden, M. H., Browning, J. E. and Chen, Y. A. (2011). Cancer susceptibility variants and the risk of adult glioma in a US case-control study, *Journal of Neurooncology* **104**(2): 535 – 542.

Festen, E. A. M., Goyette, P., Green, T., Boucher, G., Beauchamp, C., Trynka, G., Dubois, P. C., Lagace, C., Stokkers, P. C. F., Hommes, D. W., Barisani, D., Palmieri, O., Annese, V., van Heel, D. A., Weersma, R. K., Daly, M. J., Wijmenga, C. and Rioux, J. D. (2011). A meta-analysis of genome-wide association scans identifies il18rap, ptpn2, tagap, and pus10 as shared risk loci for crohn's disease and celiac disease, *PLoS Genetics* **7**(1): e1001283.

Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent, *Journal of Statistical Software* **33**(1): 1 – 22.

Hoff, P. D. (2005). Bilinear mixed-effects models for dyadic data, *Journal of the American Statistical Association* **100**(469): 286 – 295.

Hu, J., He, X., Cote, G. J. and Krahe, R. (2009). Singular value decomposition based alternative splicing detection, *Journal of the American Statistical Association* **104**(487): 944 – 953.

Hunter, D. R. and Lange, K. (2004). A tutorial on MM algorithms, *American Statistician* **1**(1): 30 – 37.

Hunter, D. R. and Li, R. (2005). Variable selection using MM algorithms, *The Annals of Statistics* **33**(4): 1617 – 1642.

Jaakkola, T. and Jordan, M. (2000). Bayesian parameter estimation via variational methods, *Statistics and Computing* **10**: 25 – 37.

Johnson, D. E. and Graybill, F. A. (1972). An analysis of a two-way model with interaction and no replication, *Journal of the American Statistical Association* **67**(340): 862 – 868.

Kooperberg, C., LeBlanc, M., Dai, J. Y. and Rajapakse, I. (2009). Structures and assumptions: Strategies to harness gene × gene and gene × environment interactions in GWAS, *Statistical Science* **24**(4): 472 – 488.

Li, Z., Gopal, V., Li, X., Davis, J. M. and Casella, G. (2012). Simultaneous SNP identification in association studies with missing data, to appear.

Maity, A., Carroll, R. J., Mammen, E. and Chatterjee, N. (2009). Testing in semiparametric models with interaction, with applications to gene-environment interactions, *Journal of the Royal Statistical Society, Series B (Methodological)* **71**: 75 – 96.

McElroy, J., Cree, B., Caillier, S., Gregersen, P., Herbert, J., Khan, O., Freudenberg, J., Lee, A., Bridges, S. J., Hauser, S., Oksenberg, J. and Gourraud, P. (2010). Refining the association of MHC with multiple sclerosis in African Americans, *Human Molecular Genetics* **19**(15): 3080 – 3088.

Philips, P. C. (2008). Epistasis – the essential role of gene interactions in the structure and evolution of genetic systems, *Nature Reviews* **9**: 855 – 867.

Platten, M., Youssef, S., Hur, E. M., Ho, P. P., Han, M. H., Lanz, T. V., Phillips, L. K., Goldstein, M. J., Bhat, R., Raine, C. S., Sobel, R. A. and Steinmana, L. (2009). Blocking angiotensin-converting enzyme induces potent regulatory t cells and modulates TH1- and TH17-mediated autoimmunity, *Proceedings of the National Academy of Sciences, U.S.A.* **106**(35): 14948 – 14953.

Poser, C. M. and Brinar, V. V. (2001). Diagnostic criteria for multiple sclerosis, *Clinical Neurology and Neurosurgery* **103**(1): 1 – 11.

Ramagopalan, S., Maugeri, N., Handunnetthi, L., Lincoln, M., Orton, S., Dyment, D., Deluca, G., Herrera, B., Chao, M., Sadovnick, A., Ebers, G. and JC, K. (2009). Expression of the multiple sclerosis-associated mhc class ii allele hla-drb1*1501 is regulated by vitamin d, *PLoS Genetics* **5**(2): e1000369.

Schwarz, G. (1978). Estimating the dimension of a model, *Annals of Statistics* **6**(2): 461 – 464.

Shen, X. and Ye, J. (2002). Adaptive model selection, *Journal of the American Statistical Association* **97**(459): 210 – 221.

Shete, S., Hosking, F., Robertson, L., Dobbins, S., Sanson, M., Malmer, B., Simon, M., Marie, Y., Boisselier, B., Delattre, J., Hoang-Xuan, K., El Hallani, S., Idbaih, A., Zelenika, D., Andersson, U., Henriksson, R., Bergenheim, A., Feychting, M., Lonn, S., Ahlbom, A., Schramm, J., Linnebank, M., Hemminki, K., Kumar, R., Hepworth, S., Price, A., Armstrong, G., Liu, Y., Gu, X., Yu, R., Lau, C., Schoemaker, M., Muir, K., Swerdlow, A., Lathrop, M., Bondy, M. and Houlston, R. (2009). Genome-wide association study identifies five susceptibility loci for glioma, *Nature Genetics* **41**(8): 899 – 904.

Snee, R. D. (1982). Nonadditivity in a two-way classification: Is it interaction or nonhomogeneous variance?, *Journal of the American Statistical Association* **77**(379): 515 – 519.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society. Series B (Methodological)* **58**(1): 267 – 288.

Zou, H., Hastie, T. and Tibshirani, R. (2007). On the "degrees of freedom" of the lasso, *Biometrics* **35**(5): 2173 – 2192.

Table 1: Simulation Setup 1. The average number (its standard error) of detected true biomarkers by the logistic ANOVA (*ANOVA*) with and without penalization, simple logistic regression (*regression*), and L_1-penalized logistic regression (*penalized regression*). Results are based on 500 simulated data sets.

	ANOVA	*ANOVA* w/o penalty	*regression*	*penalized regression*
$\rho = 0.3$	9.29 (0.11)	9.22 (0.12)	6.39 (0.09)	6.54 (0.10)
$\rho = 0.5$	9.31 (0.12)	9.21 (0.12)	6.23 (0.09)	6.31 (0.09)

Table 2: Simulation Setup 2. The average number (its standard error) of detected true biomarkers by the logistic ANOVA (*ANOVA*) with and without penalization, simple logistic regression (*regression*), and L_1-penalized logistic regression (*penalized regression*). Results are based on 500 simulated data sets.

ANOVA	*ANOVA* w/o penalty	*regression*	*penalized regression*
17.72 (0.120)	16.90 (0.117)	16.76 (0.114)	5.59 (0.07)

Table 3: Simulation Setup 3. The average number (its standard error) of detected true biomarkers by the logistic ANOVA (*ANOVA*) with and without penalization, and simple multinomial logistic regression (*regression*). Results are based on 500 simulated data sets.

	ANOVA	*ANOVA* w/o penalty	*regression*
$\rho = 0.3$	19.17 (0.135)	18.95 (0.132)	8.18 (0.10)
$\rho = 0.5$	18.90 (0.129)	18.78 (0.129)	7.99 (0.10)

Table 4: Simulation Setup 4. The average number (its standard error) of detected true biomarkers by the logistic ANOVA (*ANOVA*) with and without penalization, simple multinomial logistic regression (*regression*), and L_1-penalized multinomial logistic regression (*penalized regression*). Results are based on 500 simulated data sets.

ANOVA	*ANOVA* w/o penalty	*regression*	*penalized regression*
8.87 (0.27)	1.65 (0.16)	2.56 (0.13)	4.03 (0.14)

Table 5: The average number (its standard error) of detected true biomarkers by the logistic ANOVA (*ANOVA*) with and without penalization, and simple logistic regression (*regression*) in the cases of zero-mean log odds ratios. Results are based on 500 simulated data sets.

	Binary phenotype		
	ANOVA	*ANOVA* w/o penalty	*regression*
$\rho = 0.5$	12.536 (0.137)	12.142 (0.127)	12.082 (0.133)
	Multi-category phenotype		
	ANOVA	*ANOVA* w/o penalty	*regression*
$\rho = 0.5$	20.612 (0.141)	20.552 (0.140)	15.066 (0.132)

Table 6: Summary of minor allele frequencies (MAF) of the top 50 SNPs in the the association study of SNPs with MS status.

MAF	5% to 15%	15% to 25%	25% to 35%	35% to 45%	45% to 50%
count	1	10	8	24	7

Table 7: Estimate of u_{i1}'s (multiplied by 10) corresponding to the 6-category phenotype for the MS study. NN stands for non-MS and normal blood pressure, NP for non-MS and pre-hypertension, NH for non-MS and hypertension, MN for MS and normal blood pressure, MP for MS and pre-hypertension, and MH for MS and hypertension.

NN	NP	NH	MN	MP	MH
-0.7691	-1.1666	-9.5172	0.3718	0.4301	10.6511

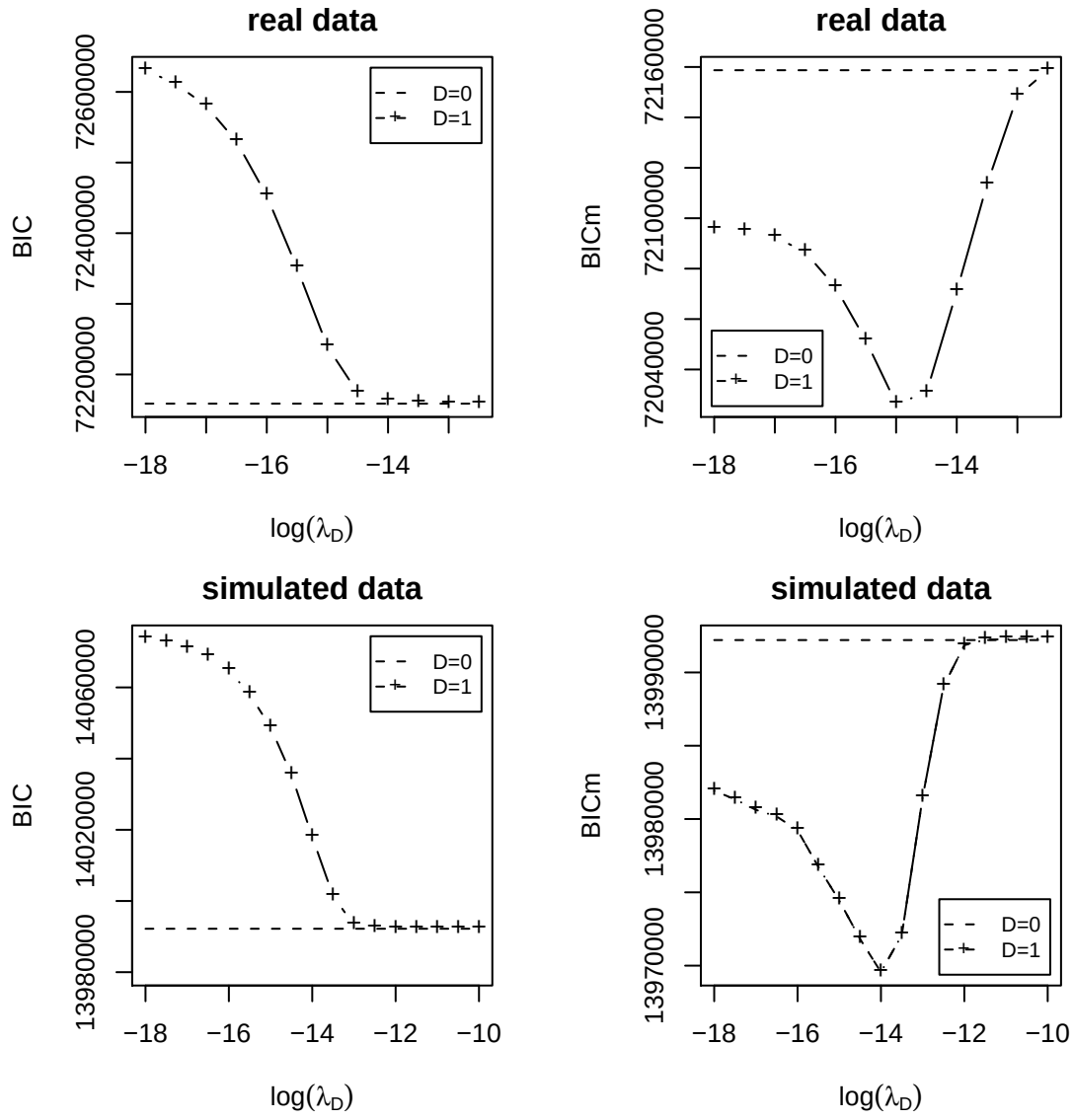

Figure 1: Plots of BIC and BIC_m against log-transformed values of the penalty parameter λ in a Multiple Sclerosis study data set and a simulated data set.

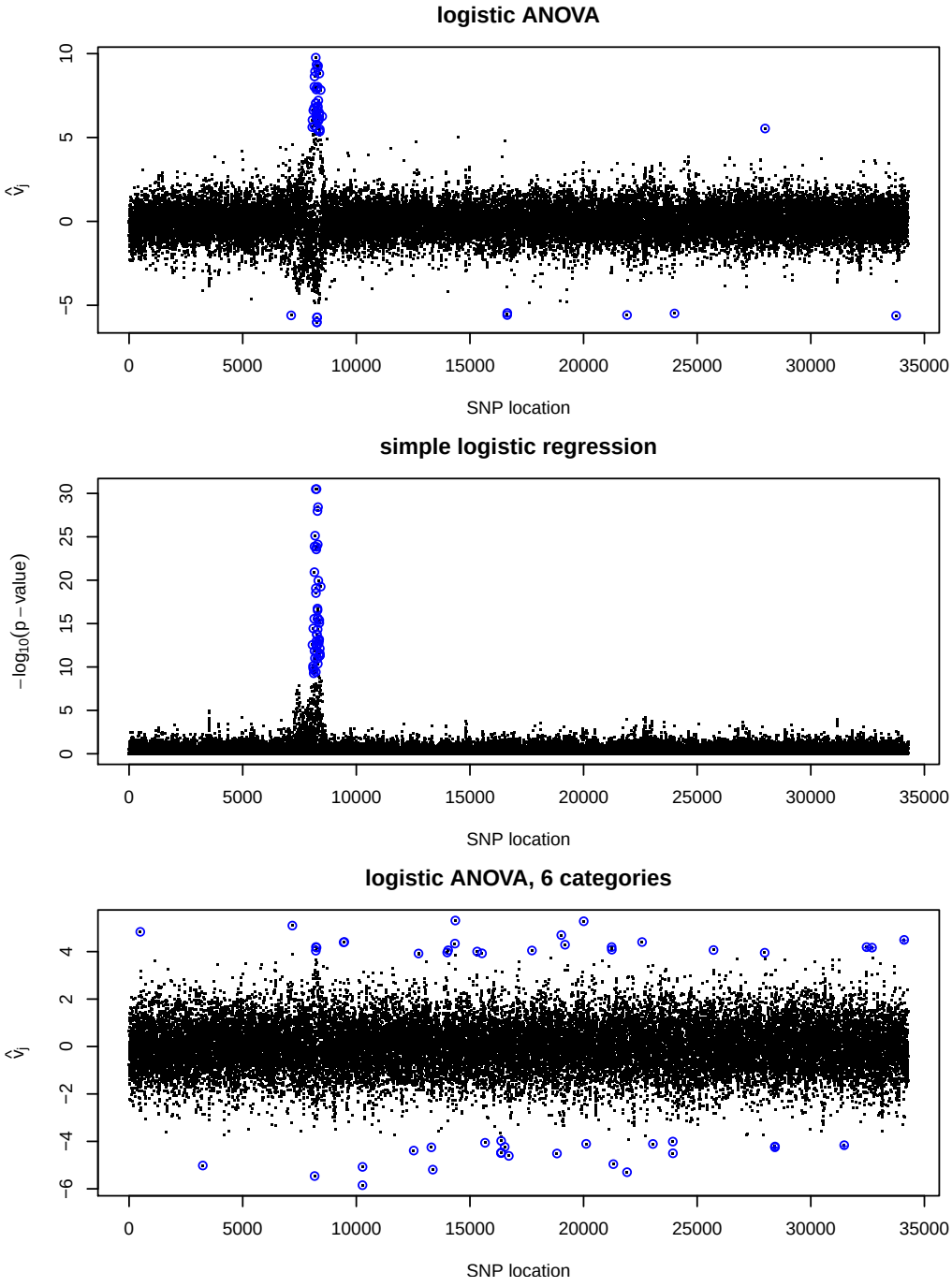

Figure 2: Plots of the SNP association indexes obtained by *logistic ANOVA* (top) and *simple logistic regression* (middle) for the MS study. The plot in the bottom is for the association among SNPs, MS and Hypertension presented in Section 4.2. The detected 50 SNPs are indicated by the circles in blue.

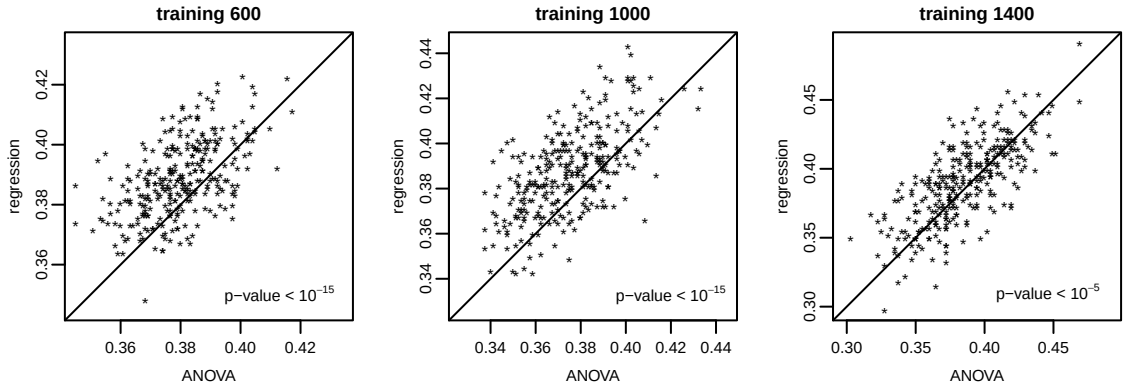

Figure 3: Scatterplot of the prediction error rates for predicting the MS status when the logistic ANOVA (*ANOVA*) and the simple logistic regression (*regression*) are used for SNP selection. Based on 300 random splits of data into training and test sets. The training set size is 600, 1000, and 1400, from the left to the right.

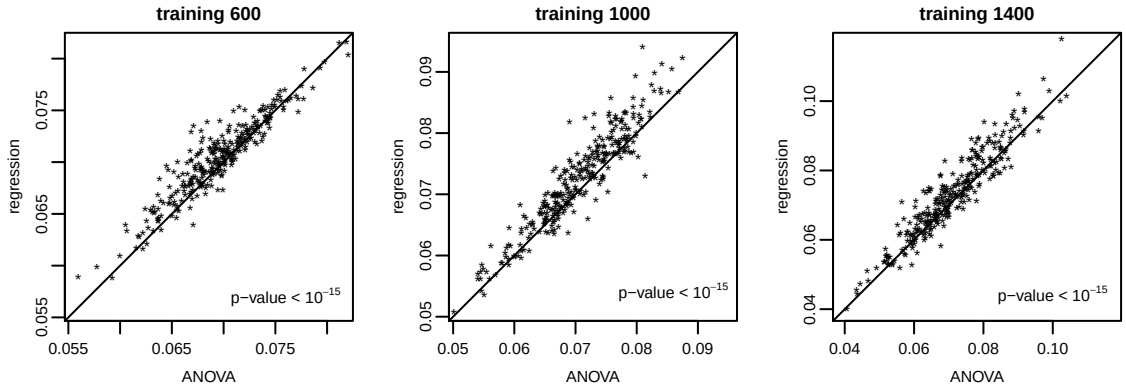

Figure 4: Scatterplot of the prediction error rates for predicting "both MS and Hypertension" and "otherwise", when the logistic ANOVA (*ANOVA*) and the simple logistic regression (*regression*) are used for SNP selection. Based on 300 random splits of data into training and test sets. The training set size is 600, 1000, and 1400, from the left to the right.