Published in final edited form as:

Biometrics. 2012 March ; 68(1): 307–315. doi:10.1111/j.1541-0420.2011.01649.x.

A Composite Likelihood Approach to Latent Multivariate Gaussian Modeling of SNP Data with Application to Genetic Association Testing

Fang Han and Wei Pan*

Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN 55455, USA

SUMMARY

Many statistical tests have been proposed for case-control data to detect disease association with multiple single nucleotide polymorphisms (SNPs) in linkage disequilibrium (LD). The main reason for the existence of so many tests is that each test aims to detect one or two aspects of many possible distributional differences between cases and controls, largely due to the lack of a general and yet simple model for discrete genotype data. Here we propose a latent variable model to represent SNP data: the observed SNP data are assumed to be obtained by discretizing a latent multivariate Gaussian variate. Since the latent variate is multivariate Gaussian, its distribution is completely characterized by its mean vector and covariance matrix, in contrast to much more complex forms of a general distribution for discrete multivariate SNP data. We propose a composite likelihood approach for parameter estimation. A direct application of this latent variable model is to association testing with multiple SNPs in a candidate gene or region. In contrast to many existing tests that aim to detect only one or two aspects of many possible distributional differences of discrete SNP data, we can exclusively focus on testing the mean and covariance parameters of the latent Gaussian distributions for cases and controls. Our simulation results demonstrate potential power gains of the proposed approach over some existing methods.

Keywords

Genome-wide association study; GWAS; latent model; logistic regression; multi-marker analysis; multivariate discrete distribution

1. Introduction

Our motivating application is to testing association between a binary phenotype and multiple single nucleotide polymorphisms (SNPs) as arising in case-control candidate gene or genome-wide association studies. Specifically, given n independent observations (Y_i, X_i') with $Y_i = 0$ or 1 as disease status and $X_i' = (X_{i1}, \dots, X_{ik})$ as genotypes at k SNPs for subject i , $i = 1, \dots, n$, we would like to test whether Y_i and X_i are associated. The k SNPs may be correlated (i.e., in linkage disequilibrium, LD). Many methods have been proposed for this purpose, including various tests based on logistic regression, haplotype-based tests, and genomic similarity-based methods. A general problem is that, due to the discrete genotype

*weip@biostat.umn.edu.

Supplementary Materials

Web Appendices A–E referenced in Sections 2–4 are available under the Paper Information link at the Biometrics website <http://www.biometrics.tibs.org>.

data X_i and its unknown distributions for cases and controls, many possible models can be used. For example, with the dosage coding for SNPs, one can focus on testing the difference of the means of X_i , leading to Hotelling's T^2 test, which is related to a score test in logistic regression. Alternatively, one can test on the difference of the means of pairwise interactions/correlations $X_{ij}X_{il}$ (for $j \neq l$) as in the LD contrast tests (Wang et al 2007), or difference of the means of X_i^2 as in contrasting the Hardy-Weinberg equilibrium (HWE) parameters (Kim et al 2010), or contrasting some complex higher-order interactions among the components of X_i 's as aimed by haplotype-based tests and similarity-based tests (Schaid et al 2005; Wessel and Schork 2006; Wei et al 2008). More generally, a combination of the above tests can be applied (Kim et al 2010; Pan 2010). A natural question is why there are so many different types of tests. An explanation lies in the lack of a simple representation for multivariate discrete data, as is the case for multiple SNPs in LD. For example, for k -dimensional multivariate binary data, a general distribution contains as many as 2^k parameters (Cox 1972).

In this paper, we propose a latent Gaussian variable model: we assume that there is a latent multivariate Gaussian variate that can be discretized to yield the observed discrete SNP data. An advantage of our proposed latent Gaussian model is that, if the model holds, rather than trying to model and test on a large number of main effects and higher-order interactions for the discrete genotypes at multiple loci, we can exclusively focus on the mean and covariance parameters of the latent Gaussian model since the mean and covariance parameters completely determine a Gaussian distribution. Although a latent Gaussian model has been used previously to simulate multiple SNPs in LD (Wang et al 2007; Pan 2009), it has never been used explicitly for modeling and inference for SNP data as discussed here.

2. Methods

2.1 SNP Data and a Latent Gaussian Model

Given n independent observations (Y_i, X_i') with $Y_i = 0$ or 1 as disease status and $X_i = (X_{i1}, \dots, X_{ik})'$ as genotypes at k SNPs for subject i , $i = 1, \dots, n$, we would like to test for any possible association between the disease and genotypes. The k SNPs are taken from a candidate region or an LD block. As usual, we use the dosage coding for X_{ij} under an additive genetic model: $X_{ij} = x_{ij} = 0, 1$ or 2, representing the copy number of one of the two alleles present in SNP j of subject i , though other genetic models can be adopted.

We assume that there is a latent Gaussian variable $Z_i = (Z_{i1}, \dots, Z_{ik})'$ for each X_i : $Z_i \sim MV N(0, R)$ with $R = (r_{jl})$ as a $k \times k$ correlation matrix; for any j , there exists two constants c_{j1} and c_{j2} such that $Pr(X_{ij} = 0) = Pr(Z_{ij} \leq c_{j1})$ and $Pr(X_{ij} = 1) = Pr(c_{j1} < Z_{ij} \leq c_{j2})$. For simplicity of notation, we define intervals $T_{ij} = (-\infty, c_{j1}]$, $(c_{j1}, c_{j2}]$ and (c_{j2}, ∞) for $X_{ij} = 0, 1$ and 2, respectively.

The above model reflects the ordinal nature of genotypes. However, we do not assume the HWE. If the HWE holds, we do not need the parameters c_{j2} 's since we have one more constraint: $Pr(X_{ij} = 2) = [Pr(X_{ij} = 1)]^2$. Without loss of generality, for identifiability we assume the marginal mean zero and marginal variance one for each latent Z_{ij} ; we do not impose any other structural assumption on the covariance matrix R .

2.2 Composite Likelihood for Parameter Estimation

Given a sample of genotypes $\{X_i : i = 1, \dots, n\}$, we would like to estimate the unknown parameters in R and $C = \{(c_{j1}, c_{j2}) : j = 1, \dots, k\}$. The (full) likelihood is

$$L_F((X_1, \dots, X_n); R, C) = \prod_{i=1}^n Pr(X_i = x_i) = \prod_{i=1}^n \int_{T_{i1}} \dots \int_{T_{ik}} f(z; 0, R) dz,$$

which involves multiple integrations with respect to $f(z; 0, R)$, the density function of a multivariate Normal distribution $MVN(0, R)$. With no simple closed-form, it is difficult to directly maximize L_F with respect to the parameters in the integral limits (C) and in the covariance matrix R . Hence, it is computationally challenging to obtain the maximum likelihood estimate (MLE) of (R, C) . As an alternative, we recourse to composite likelihoods (Lindsay 1988), which include the pseudo-likelihood as a member. Composite likelihoods are often used for multivariate data for which the full likelihood is too complicated. It is often based on the marginal or conditional distributions of some subcomponents of the multivariate data under some working, though generally incorrect, assumption, such as the independence among the subcomponents of the multivariate data (Varin et al 2011). The pseudo-likelihood and composite likelihood are often used exchangeably in the literature (Cox and Reid 2004), though the former is perhaps used more loosely in referring to conditional likelihoods of some subcomponents (Besag 1974) or something else (Gong and Samaniego 1981). Although composite likelihoods have been used in genetics, mainly in population genetics and fine mapping (Devlin et al 1996), we are not aware of their use in association analysis; see a recent review (Larribe and Fearnhead 2011).

Overall, our use of composite likelihoods for parameter estimation is similar to that of Anderson and Pemberton (1985). Since the full multivariate data Z_i is multivariate Normal, any of its subcomponents is Normal too. For example, the marginal distribution of Z_{ij} is a univariate $N(0, 1)$, while the marginal likelihood for X_{ij} only depends on the known distribution of Z_{ij} with unknown parameter (c_{j1}, c_{j2}) . Hence, for any fixed j , it will be simple to use the marginal likelihood of X_{ij} 's to estimate (c_{j1}, c_{j2}) . Formally, to estimate C , we propose using a composite marginal likelihood

$$CL_1((X_1, \dots, X_n); C) = \prod_{i=1}^n \prod_{j=1}^k Pr(X_{ij} = x_{ij}) = \prod_{i=1}^n \prod_{j=1}^k \int_{T_{ij}} f(z_j; 0, I) dz_j.$$

Maximizing CL_1 leads to the maximum composite likelihood estimate (MCLE) of C , $\hat{C} = \{(\hat{c}_{j1}, \hat{c}_{j2}) : j = 1, \dots, k\}$, determined by

$\hat{Pr}(X_{ij}=0) = \sum_{i=1}^n I(X_{ij}=0)/n = Pr(Z_{ij} \leq \hat{c}_{j1})$ and $\hat{Pr}(X_{ij}=1) = \sum_{i=1}^n I(X_{ij}=1)/n = Pr(\hat{c}_{j1} < Z_{ij} \leq \hat{c}_{j2})$. Noting that $Z_{ij} \sim N(0, 1)$, we obtain the MCLE $(\hat{c}_{j1}, \hat{c}_{j2})$ from the corresponding quantiles of $N(0, 1)$. Accordingly, we define intervals $\hat{T}_{ij} = (-\infty, \hat{c}_{j1}]$, $(\hat{c}_{j1}, \hat{c}_{j2}]$ and $(\hat{c}_{j2}, \infty)$ for $X_{ij} = 0, 1$ and 2, respectively.

To estimate R , we use a composite pairwise likelihood $CL_2(\hat{C}, R)$ with

$$CL_2(C, R) = CL_2((X_1, \dots, X_n); C, R) = \prod_{i=1}^n \prod_{j>l} Pr(X_{ij}=x_{ij}, X_{il}=x_{il}) = \prod_{i=1}^n \prod_{j>l} \int_{T_{ij}} \int_{T_{il}} f(z_{jl}; 0, R_{(jl)}) dz_{jl},$$

where $R_{(jl)}$ is a 2×2 correlation matrix with off-diagonal elements r_{jl} , and $z_{jl} = (z_j, z_l)'$ is bivariate. It is easy to see that $CL_2(\hat{C}, R)$ is a product of $CL_2^{jl}(r_{jl})$'s with

$$CL_2^{jl}(r_{ij}) = \prod_{i=1}^n \int_{\hat{r}_{ij}} \int_{\hat{r}_{il}} f(z_{jl}; 0, R_{(jl)}) dz_{jl}.$$

Hence, maximizing $CL_2(\hat{C}, R)$ with respect to the unknown elements in R is equivalent to separately maximizing each $CL_2^{jl}(r_{jl})$'s with respect to r_{jl} for each pair of (j, l) for $j > l$. Since each $CL_2^{jl}(r_{jl})$ is a univariate function of r_{jl} that is bounded between -1 and 1 , many of the existing methods can be applied to maximize $CL_2^{jl}(r_{jl})$, obtaining the MCLE $\hat{r}_{jl}$. In our implementation, we used the R function `nllminb()`.

It has been shown generally that MCLE is consistent and asymptotically Normal, and a corresponding composite likelihood ratio test has a mixture of χ_1^2 distributions (Zhao and Joe 2005). However, the covariance matrix estimator of the MCLE (as a sandwich estimator) does not appear easy to calculate here. We will propose a permutation-based approach to estimating the covariance matrix under a null hypothesis. For finite samples, the MCLE $\hat{R}$ may not be positive-definite, though we never encountered such a problem in a simulation study with 1000 simulated datasets. Nevertheless, as a constraint on its condition number (Won et al 2009), we regularize $\hat{R}$ in the following way: we first apply a singular value decomposition (SVD): $\hat{R} = U_1 \Lambda_1 U_1'$; if any of its eigen values (i.e. the elements of the diagonal matrix Λ_1) is less than a small number e (e.g. $e = 0.001$), we replace it by e , and keep it otherwise, to obtain a possibly new Λ_2 ; finally we have a positive definite $\hat{R} = U_1 \Lambda_2 U_1'$.

2.3 New Association Tests Based on the Latent Variable Model

Now given case-control data (Y_i, X_i) , we would like to test whether Y_i and X_i are associated. Many methods have been proposed. The challenge is due to possibly complicated distributions for discrete genotype data X_i . For example, one can test on the mean difference of X_i as targeted by Hotelling's T^2 test, which can be also done by regressing Y on X_i in a

logistic regression model: $\text{LogitPr}(Y_i=1|X_i) = \beta_0 + \sum_{j=1}^k X_{ij}\beta_j$, and testing the null hypothesis $H_0: \beta_1 = \dots = \beta_k = 0$. However, a potential problem is that the above logistic regression model may be mis-specified, leading to reduced power. For example, why not to add

squared terms X_{ij}^2 , and/or two-way interactions $X_{ij}X_{il}$ into the above main-effects model since these terms may be significant predictors of the disease status Y_i ? On the other hand, there is a trade-off: adding too many terms may largely diminish the power of the resulting test due to a high number of degrees of freedom (DF). An advantage of our proposed latent Gaussian model is that, rather than trying to model and test on a large number of main effects and higher-order interactions of the discrete genotypes at multiple loci, we simply focus on the mean and covariance parameters in the latent Gaussian model since the mean vector and covariance matrix completely determine a multivariate Normal distribution. However, as usual, care needs to be taken to guard against model-misspecification; future studies need to address whether the latent Gaussian model is more robust and parsimonious in practice.

Without loss of generality, we assume that for the given case-control data, the first n_1 observations are cases while the remaining $n_0 = n - n_1$ are controls. Denote Ω_1 and Ω_0 as the unknown parameters for the latent distributions for the case and control groups respectively. We first use only the cases to obtain the MCLE $\hat{\Omega}_1$ for cases, then use only the controls to

obtain $\hat{\Omega}_0$. Finally, we pool the cases and controls together, and use their genotype data to obtain the MCLE $\hat{\Omega}$.

To test the null hypothesis H_0 that the genotype distributions for the case and control groups are the same, we can simply test on the equality of the two latent Gaussian distributions.

That is, we test $H_0': \Omega_1 = \Omega_0$. We propose using some Wald-type tests, including the Wald test (shortened as L-Wald) and its two modifications called L-SSB and L-SSBw test, and the popular UminP (L-UminP) test (Pan 2009). We propose using permutations to obtain the covariance matrix V of MCLE $\hat{\Omega}_1 - \hat{\Omega}_0$ under H_0' : 1) we randomly shuffle the case and control labels, yielding permuted dataset b ; 2) we obtain MCLE $\hat{\Omega}_1^b - \hat{\Omega}_0^b$ from permuted dataset b ; 3) repeat the above steps for $b = 1, \dots, B$ times. Then the sample covariance matrix of $\{\hat{\Omega}_1^b - \hat{\Omega}_0^b\}$ is $\hat{V}$. The test statistics for H_0' are

$$T_{L-Wald} = (\hat{\Omega}_1 - \hat{\Omega}_0)' \hat{V}^{-1} (\hat{\Omega}_1 - \hat{\Omega}_0), T_{L-SSB} = (\hat{\Omega}_1 - \hat{\Omega}_0)' (\hat{\Omega}_1 - \hat{\Omega}_0),$$

$$T_{L-SSBw} = (\hat{\Omega}_1 - \hat{\Omega}_0)' \text{Diag}(\hat{V})^{-1} (\hat{\Omega}_1 - \hat{\Omega}_0), T_{L-UminP} = \max_{1 \leq j \leq d} (\hat{\Omega}_1 - \hat{\Omega}_0)_j / \sqrt{\hat{V}_{jj}},$$

where for a vector a and matrix M , a_j is the j th element of a and M_{jj} is the j th diagonal element of M . The asymptotic null distribution of the L-Wald test is χ_d^2 with $d = 2k + k(k - 1)/2$, while that for the L-SSB or L-SSBw test is $\sum_{j=1}^d \tau_j \chi_1^2$ with τ_j 's as eigenvalues of $\hat{V}$ or $\hat{V} \text{Diag}(\hat{V})^{-1}$. We approximate $\sum_{j=1}^d \tau_j \chi_1^2$ by a scaled-shifted chi-squared distribution $a\chi_r^2 + b$, where the three constants a , b and r , as proposed by Zhang (2005) and studied in the context of association testing by Pan (2009), can be easily obtained:

$$a = \frac{\sum_{j=1}^d \tau_j^3}{\sum_{j=1}^d \tau_j^2}, b = \sum_{j=1}^d \tau_j - \frac{\left(\sum_{j=1}^d \tau_j^2\right)^2}{\sum_{j=1}^d \tau_j^3}, r = \frac{\left(\sum_{j=1}^d \tau_j^2\right)^3}{\left(\sum_{j=1}^d \tau_j^3\right)^2}.$$

Based on the (asymptotic) normality of $\hat{\Omega}_1 - \hat{\Omega}_0$, numerical methods can be applied to obtain a p-value for the L-UminP test.

We can also apply a composite likelihood ratio test (CLRT) to test H_0' :

$$CLRT = 2 \log CL_2((X_1, \dots, X_{n_1}); \hat{\Omega}_1) + 2 \log CL_2((X_{n_1+1}, \dots, X_n); \hat{\Omega}_0) - 2 \log CL_2((X_1, \dots, X_{n_1}, X_{n_1+1}, \dots, X_n); \hat{\Omega}),$$

whose asymptotic null distribution is a mixture of χ_1^2 's with weights depending on the covariance matrix of the MCLE, which is not easy to calculate here. We propose two methods to estimate its null distribution. First, we use permutations to estimate its p-value directly, as described briefly below: we randomly shuffle the case-control label to obtain a permuted dataset b , then the corresponding CLRT statistic $CLRT_{(b)}$ for $b = 1, \dots, B$. The p-value is calculated as $\sum_{b=1}^B I(CLRT_{(b)} > CLRT_0) / B$, where $CLRT_0$ is the CLRT statistic based

on the original data. We denote the resulting latent model-based CLRT as L-CLRT-P. A downside of the above procedure is that, in order to achieve a highly significant p-value, we need to run a larger number of permutations, which may be time-consuming. Alternatively, in the second method, we use permutations to estimate its null distribution. Since a shifted-scaled chi-squared distribution, say $a\chi_r^2 + b$, has been effectively used to approximate a mixture of χ_1^2 's (Zhang 2005), we propose using the Satterthwaite (1946) method to estimate a , b and r as the following

$$a_c = (\tau_c^3 - 3\mu_c\sigma_c^2)/(4\sigma_c^2), \quad r_c = \sigma_c^2/(2a_c^2), \quad b_c = \mu_c - a_c r_c,$$

where μ_c , σ_c^2 and τ_c^3 are the sample mean, sample variance and (centered) third moment of $CLRT_{(b)}; \mu_c = \sum_{b=1}^B CLRT_{(b)}/B$, $\sigma_c^2 = \sum_{b=1}^B CLRT_{(b)}^2/B - \mu_c^2$ and $\tau_c^3 = \sum_{b=1}^B CLRT_{(b)}^3/B - \mu_c^3$. We call the resulting test L-CLRT-S. Although the CLRT is asymptotically equivalent to the Wald test based on the MCLE, since the former is a scalar, it can be easier to estimate its variance accurately, compared to estimating possibly a high-dimensional covariance matrix of the MCLE in the Wald test. This point will be confirmed by our simulation studies, in which it is shown that the CLRT had better-controlled Type I error rates while the Wald test might have inflated Type I error rates with a small B . On the other hand, the SSBw test, a modification to the Wald test, will also be shown to perform well.

As an alternative to the CLRT, we propose using the usual log-likelihood with the MCLE plugged-in to approximate the usual MLE. Specifically, to test $H_0: \hat{\Omega}_1 = \hat{\Omega}_0$, we propose using the (full) likelihood ratio test (LRT):

$$LRT = 2 \log L_F((X_1, \dots, X_{n_l}); \hat{\Omega}_1) + 2 \log L_F((X_{n_l+1}, \dots, X_n); \hat{\Omega}_0) - 2 \log L_F((X_1, \dots, X_{n_l}, X_{n_l+1}, \dots, X_n); \hat{\Omega}),$$

whose asymptotic null distribution is difficult to calculate. As for CLRT, we propose two methods based on permutations to 1) estimate its p-value and 2) to estimate its null distribution as a shifted-scaled chi-squared distribution, denoted as the latent model-based LRT-P (L-LRT-P) and LRT-S (L-LRT-S) tests respectively. Note that, although it is difficult to maximize the full likelihood L_F to obtain the MLE, it is much easier to calculate L_F with any estimate $\hat{\Omega}$, e.g. using the function `pmvnorm()` from R package `mvtnorm`.

2.4 Other Association Tests

For comparison, we also consider two logistic regression models:

$$M1: \text{LogitPr}(Y_i=1|X_i) = \beta_0 + X_i' \beta,$$

$$M2: \text{LogitPr}(Y_i=1|X_i) = \beta_0 + X_i' \beta_1 + X_i^2 \beta_2,$$

where X_i^2 includes both quadratic terms and two-way interactions for X_i . For model M1 or M2, we applied the score, SSU, SSUw and UminP tests (Pan 2009) to test $H_0: \beta_1 = 0$, or $H_0: \beta_1 = 0$ and $\beta_2 = 0$ respectively, and denote the resulting tests as M1-score, M1-SSU, and so on.

In addition, we also applied the haplotype-based score test and its max version, denoted as H-score and H-max respectively. The two tests were implemented in the function `haplo.score()` in R package `haplo.stats` (Schaid et al 2005). We used their default settings to derive their p-values with 1000 permutations. Finally, we applied the Normal-

based likelihood ratio test (NLRT) of Wang et al (2008), which is obtained under the incorrect Normality assumption on observed genotypes X_i , rather than on a latent variable Z_i as in our proposed tests. A variation of the NLRT was to first extract the first few principle components (PCs) of genotypes X that could explain at least 85% of the total variation, then apply the NLRT to the resulting first few PCs, denoted as NLRT_{pc}. The p-values of both NLRT and NLRT_{pc} were obtained from 1000 permutations.

2.5 Comparison with Multivariate Probit Models and Related Methods

The proposed latent Gaussian model is related to multivariate probit models for correlated binary or ordinal data (Ashford and Sowden 1970; Anderson and Pemberton 1985). However, in existing multivariate probit models, the focus is on regression coefficients related to the mean parameters, and thus a common and covariate-independent correlation matrix is usually assumed; in contrast, here we are *more* interested in the group-specific correlation parameters; see Appendix A for details. Due to this key difference, a direct application of a usual multivariate probit model may be low powered in the current context. In addition, maximum (full) likelihood is typically used for multivariate probit models with a small k , as implemented in R package `mprobit`, whereas we adopt a composite likelihood approach to avoid the complicated (full) likelihood. As discussed before, a severe limitation of the full likelihood approach is its prohibitively high computational cost for high-dimensional data, though future computational advance will ease the bottleneck. Currently R package `mprobit` restricts that “The maximum number of parameters (regression and latent correlation parameters combined) is 23”, which implies $k \leq 6$ with an unstructured correlation matrix; furthermore, as to be shown, the full maximum likelihood approach did not perform well even for $k = 6$.

For computational simplicity, we have proposed a two-step procedure for parameter estimation: we first maximize the composite marginal likelihood CL_1 to obtain MCLE $\hat{C}$ for the location or cut-off parameters C due to their closed-form solutions, then we maximize the composite pairwise likelihood CL_2 to obtain MCLE $\hat{R}$ for the correlation parameters R , which can be conveniently decomposed into a series of univariate optimization problems. Alternatively, one can maximize CL_2 jointly with both C and R , which obviously is computationally more challenging. As to be shown, this one-step procedure is much slower than the two-step procedure; importantly, there was barely any difference between the estimates from the two procedures, which is in agreement with the result of Fiocco et al (2009).

Finally, rather than using an unstructured correlation matrix R , one may want to use a simpler parametric model to reduce the number of parameters. For SNP data, it is unclear what parametric models are suitable. We will show in Appendix E that for our real data, as typical for SNP data, the correlation structures are complex to model. As to be shown, an incorrect model for R may lead to loss of power, in which case, even the maximum full likelihood approach may be less powerful than our proposed method.

3. Simulations

3.1 Simulated Genotypes with Latent Gaussian Models

We first considered simulating genotypes from latent Gaussian models, as done by Wang and Elston (2007). We used $k = 3, 6$ and 9 marker SNPs, and sample size $n = 200$ with $n_1 = 100$ cases and $n_0 = 100$ controls. We considered three covariance structures for the latent Gaussian model: a compound symmetry $CS(\rho)$ with a constant off-diagonal element ρ ; an AR-1(ρ) with its (i, j) th element as $\rho^{|i-j|}$; and a random correlation matrix $Rand(\rho)$ with its (i, j) th element randomly chosen from a Uniform distribution $U(\rho - 0.2, \rho + 0.2)$. We discarded any $Rand$ correlation matrix that was not positive definite.

We followed the below steps to generate each simulated dataset. First, we generated two independent latent vectors Z_1 and Z_2 from $MV N(0, R_0)$ or $MV N(0, R_1)$, corresponding to a control or a case respectively. The two covariance matrices shared one of the three structures with parameter ρ_0 and ρ_1 respectively. Second, each element of Z_1 was dichotomized to yield an allele according to its specified allele frequency, which was randomly chosen between 0.4 and 0.6 for a control, or for a case was the allele frequency of the control plus a small number randomly chosen between $-|\rho_1 - \rho_0|/10$ and $|\rho_1 - \rho_0|/10$. We similarly discretized Z_2 . Third, summing up the discretized Z_1 and Z_2 , we obtained genotype X_i . We repeated the above steps for $i = 1, \dots, n_1$ cases and $i = 1, \dots, n_0$ controls to obtain a simulated dataset.

For our proposed tests, we used $B = 200$ permutations. For each of the simulation set-ups, we used 1000 replicates to estimate Type I error rates and power of the various tests.

For $k = 6$ SNPs (Table 1), it is somewhat surprising to see that L-Wald had dramatically inflated Type I error rates while other tests seemed to have well-controlled Type I error rates. We conjectured that it was likely due to the relatively small number ($B = 200$) of permutations in estimating a relatively large (27×27) covariance matrix V . We explored this conjecture in two ways: first, the Type I error rates of L-Wald were even larger, near 30%, if $B = 50$ was used; second, although it seemed to suffice to use $B = 200$ for $k = 3$ SNPs (for a 9×9 covariance matrix), using $B = 50$ also led to inflated Type I error rates at about 15%. Note that, in contrast to the use of $\hat{V}^{-1}$ by L-Wald, the other three L-Wald-type tests either do not use $\hat{V}$ or use only the diagonal elements of $\hat{V}$ or eigenvalues of $\hat{V}$ $\text{Diag}(\hat{V})^{-1}$. Although we only used a small $B = 200$ to save computing time for simulations, it is practical to use a much larger B for real data. Hence, we conclude that a larger k requires a larger B for the L-Wald-type tests, especially L-Wald, and we would skip the discussion on the power of L-Wald under this simulation set-up for its inflated Type I error rates.

For power with $k = 6$ (Table 1), L-SSBw was the most powerful, followed by the L-LRT and NLRT/NLRTpc tests. L-SSB performed well except for AR-1. In the next tier we have L-UminP, then followed by H-score, H-max, M2-Score and L-CLRT. Again other tests based on the two logistic regression models had minimal power due to little differences in the marginal means of the SNPs between the two groups. We highlight substantial power gains of L-SSBw over H-score and NLRT: for example, for $\rho_1 = 0.6$ and $\rho_0 = 0.2$ with the Rand correlation structure, the power of L-SSBw was 92.8%, much higher than 55.6% and 79.0% of the latter two tests.

For $k = 9$ with a Rand correlation structure (Table 2), the L-Wald test had dramatically inflated Type I error rates while other tests seemed to have satisfactory Type I error rates. For power, it appears that our proposed L-SSBw was most powerful, closely followed by L-SSB, then by NLRTpc. Next came the NLRT, L-UminP, L-LRT-P and L-LRT-S, followed by L-CLRT-P, L-CLRT-S, H-score and H-max. Except M2-score, other tests based on logistic regression had minimal power.

3.1.1 Other comparisons—Here we summarize other simulation results included in Appendices A–C. First, we consider various approaches using maximum full likelihood implemented in R package `mprob` (Appendix A). As expected, if we treat our two-sample test as a multivariate probit regression problem with a group indicator (i.e. case vs control) as a covariate, and with or without interactions between group and SNP indicators, the power was very low, similar to that with a main-effects logistic regression model M1. The reason is that, in such a multivariate probit model, as usual, a common correlation structure independent of any covariates was assumed. On the other hand, if we explicitly modeled

group-specific correlation matrices, the full likelihood ratio test (F-LRT) was much more powerful: for $k = 3$ SNPs (with Rand correlation structure), it maintained a correct Type I error rate, and was as powerful as the SSBw test. However, for $k = 6$, it had an inflated Type I error rate, under which it was no more powerful than the SSBw test. As noted earlier, the `mprobit` package could not deal with more parameters than $k = 6$ SNPs; if we ran it, it gave either some error messages or strange results. In addition, for a dataset with $k = 3, 6$ and 9 SNPs respectively, running `mprobit` on a Linux workstation took 2.2, 352.5 and 1852.1 seconds, compared to 67.9, 310.7 and 876.8 seconds for our proposed methods (including running 200 permutations).

Second, we compared the performance of our proposed two-step estimation procedure with the one-step procedure. To obtain the MCLE for a dataset with $k = 3$ and 6 SNPs (without permutations), the two-step procedure took 0.27 and 1.03 second respectively, while the one-step procedure took 8.16 and 120.64 seconds respectively. Thus, it is not feasible to use the one-step procedure for inference with permutations. More importantly, as shown in Appendix B, the two procedures gave almost the same parameter estimates.

Third, we implemented both our proposed composite likelihood methods and the full likelihood method (in R package `mprobit`) by assuming a CS correlation matrix for each of the case and control groups. As shown in Appendix C, for $k = 3$ SNPs, with the true correlation structure as either CS or Rand, the full likelihood based F-LRT was most powerful, while our composite likelihood-based tests performed similarly as before. For $k = 6$ SNPs, with the correctly specified CS structure, F-LRT was slightly more powerful than our proposed L-SSBw; however, with the incorrectly specified Rand structure, F-LRT lost to our L-SSBw with an unrestricted correlation structure. Again there was not much difference in power between the composite likelihood-based tests with a CS or an unrestricted correlation structure.

3.2 HapMap Data: Chromosome 5

To mimic real human LD structures, we used the genotypes of the 60 unrelated CEU samples (i.e. parents of the 30 trios) in the HapMap data (Thorisson et al 2005). We considered a region on chromosome 5 studied by Lin and Schaid (2009). The region of 70Kb contained 8 SNPs. For each simulated dataset, we randomly chosen four causal SNPs, say i_1, i_2, i_3 and i_4 , then the disease status was generated from a complex logistic regression model:

$$\text{LogitPr}(Y_i=1)=\beta_0+\log(OR)[I(X_{i_1}>0) \times I(X_{i_2}>0) \times (X_{i_3}+X_{i_4})^2].$$

The causal SNPs were removed and only the remaining $k = 4$ SNPs were retained in each dataset. The sample size was $n_1 = n_0 = 100$. For each value of OR , 1000 replicates were used.

As shown in Table 3, the L-Wald and L-UminP had slightly inflated Type I error rates, while other tests had well-controlled Type I error rates. It appeared that M2-Score was most powerful, closely followed by L-CLRT, H-score and H-max, then by M2-UminP and L-SSBw. NLRTpc and NLRT were more powerful than only L-SSB, the one with the lowest power.

3.3 Summary of Simulation Results

Taking together the simulation results (including those in Appendix D), we recommend the use of L-SSBw and L-CLRT, one of which was either more powerful than or almost as

powerful as the haplotype-based tests or Normal-based LRT. The L-LRT also performed well. A caution should be raised in using the L-Wald and L-UminP tests for their possibly inflated Type I error rates with a small permutation number B . Hence, we do not recommend the use of L-Wald and L-UminP. We also do not recommend L-SSB, which ignores differential variations of the estimated parameters, especially those between C and R . The tests based on logistic regression with or without interaction terms, except M2-Score, may not be able to fully capture information in LD or higher-order interactions, leading to reduced power. Therefore, our proposed tests, along with haplotype-based tests, are competitive alternatives to logistic regression.

4. Example: ALS Data

We applied the proposed tests along with several other popular tests to the data drawn from a genome-wide association study on amyotrophic lateral sclerosis (ALS) (Schymick et al 2007). The study sample consisted of $n_1 = 276$ patients with sporadic ALS and $n_0 = 268$ controls. Schymick et al (2007) took a single-marker approach by testing SNP by SNP under both the dosage (1-DF) and the general (2-DF) codings for each SNP, identifying 34 most significant SNPs, none of which however reached a genome-wide significance level after a Bonferroni adjustment. We randomly picked up 9 SNPs from Schymick's list of the 34 most significant common SNPs. As before, we used the dosage coding for each SNP. Table 4 gives the p-value of the univariate score test on each SNP. For each of the 9 SNPs, we extracted 10 neighboring SNPs upstream and another 10 downstream, then applied the default LD blocking algorithm implemented in Haploview (v4.1) (Barrett et al 2005) to each 21-SNP region for the control group. We only consider the nine LD blocks surrounding the nine SNPs with each block size given in Table 4. Appendix E illustrates complex correlation structures of two of the nine LD blocks.

The p-values of various tests are listed in Table 4. It is noted that for the last four SNP blocks, the L-CLRT or L-SSBw gave the most significant p-values, possibly due to their high power in these cases.

5. Discussion

We have proposed and studied a latent multivariate Gaussian model for multiple SNPs in a candidate region to facilitate association testing. When the latent model (approximately) holds, we can exclusively focus on the mean and covariance parameters of the latent Gaussian distributions, eliminating the need to consider a large number of other high-order interactions among the SNPs. Compared to some commonly used tests, such as those based on logistic regression and inferred haplotypes, our numerical results demonstrate the potential power gain of the proposed tests under some situations. Specifically, we recommend our latent model-based L-SSBw test, L-LRT and L-CLRT. Note that, since there is no uniformly most powerful test for multiple SNPs (Pan 2009), our proposed tests may be useful alternatives and complements to other tests in practice.

We note that, as shown by Cox and Reid (2004), a composite pairwise likelihood for a latent Gaussian model is equal to the full likelihood for a quadratic model for multivariate *binary* data. In the current context, one may speculate that our proposed model might be equivalent to a logistic regression model with two-way interactions. However, in our simulations (e.g. Table 1), they did not appear to be equivalent: the tests based on model M2 were often much less powerful than the tests based on the latent Gaussian model, though a theory lacks.

It is straightforward to apply the proposed latent Gaussian model to ordinal data with more than 3 categories: we can simply add more cut-off parameters in C , and adopt the same two-step estimation procedure. In addition, we can extend the proposed model to include

covariates. First, as the usual probit model (e.g. in R package `mprobit`), rather than assuming mean $\mu = 0$ in the latent multivariate Gaussian distribution, we assume $\mu = A\alpha$ for some covariates A (including group indicator) and unknown regression parameter α . Second, we may model the dependence of latent correlations on covariates A : $F(r_{jl}) = A\gamma$, where F is Fisher's (or other suitable) transformation, and r_{jl} is the correlation between latent components j and l . Again to avoid the use of complicated full likelihood, composite marginal and pairwise likelihoods can be used for parameter estimation and inference (Fiocco et al 2009). Furthermore, it is of interest to investigate whether the CLRT can be adjusted to have the usual chi-squared null distribution and improved power (Chandler and Bate 2007; Pace et al 2010) in the current context; a challenge lies in how to calculate information matrices not only needed for such an adjusted CLRT, but also intractable with the latent Gaussian model.

A potential problem with the proposed methods is the loss of power due to the large number of parameters in the covariance matrix if the number of SNPs is large. A similar issue exists with haplotype-based approaches. In theory, for k SNPs, compared to 2^k , the maximum number of possible haplotypes, the number of DF in the L-Wald test, $2k + k(k-1)/2$, can be much smaller. Their ratio is about 1 for $k \leq 4$, but equals to 1.6, 2.4, 3.7, 5.8, 9.5 and 15.8 for $k = 5, \dots, 10$; for $k = 20$, the ratio is as large as 4559. Hence, compared to the haplotype-based approaches, the proposed method has a far smaller number of DF, though both suffer from large DF for moderate to large k . On the other hand, in haplotype-based methods, to reduce the number of DF, it is a common practice to combine with other more common haplotypes any rare haplotype with an estimated frequency less than a specified small number, say α_0 . In our numerical examples we used the default value of $\alpha_0 = 0.01$ in the R function `haplo.score()`. However, such a combining procedure is somewhat arbitrary and may have unknown negative consequences; a more rigorous approach is through parameter regularization via penalized regression (Guo and Lin 2009; Tzeng and Bondell 2010). For our proposed latent model, some parametric models or regularization on the covariance matrix may be productive in further reducing the number of DF; more work is needed.

Software in R implementing the proposed new tests will be posted on our web site at <http://www.biostat.umn.edu/~weip/prog.html>.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported by NIH grants GM081535 and HL65462. The authors thank Dr Melanie Wall for helpful discussions, and thank the reviewers for constructive comments.

REFERENCES

- Anderson JA, Pemberton JD. The grouped continuous model for multivariate ordered categorical variables and covariate adjustment. *Biometrics*. 1985; 41:875–885. [PubMed: 3830255]
- Ashford JR, Sowden RR. Multivariate probit analysis. *Biometrics*. 1970; 26:535–546. [PubMed: 5480663]
- Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*. 2005; 21:263–265. [PubMed: 15297300]
- Besag J. Spatial interaction and the statistical analysis of lattice systems (with Discussion). *J R Statist Soc B*. 1974; 36:192–236.
- Chandler RE, Bates S. Inference for clustered data using the independence log-likelihood. *Biometrika*. 2007; 94:167–183.

- Cox DR. The analysis of multivariate binary data. *Applied Statistics*. 1972; 21:113–120.
- Cox DR, Reid N. A note on pseudolikelihood constructed from marginal densities. *Biometrika*. 2004; 91:729–737.
- Devlin B, Risch N, Roeder K. Disequilibrium mapping: composite likelihood for pairwise disequilibrium. *Genomics*. 1996; 36:1–16. [PubMed: 8812410]
- Fiocco M, Putter H, van Houwelingen JC. A new serially correlated gamma-frailty process for longitudinal count data. *Biostatistics*. 2009; 10:245–257. [PubMed: 18796463]
- Gong G, Samaniego FJ. Pseudo maximum likelihood estimation: theory and applications. *Ann Statistic*. 1981; 9:861–869.
- Guo W, Lin SL. Generalized linear modeling with regularization for detecting common disease rare haplotype association. *Genet Epidemiol*. 2009; 33:308–316. [PubMed: 19025789]
- Kim S, Morris NJ, Won S, Elston RC. Single-marker and two-marker association tests for unphased case-control genotype data, with a power comparison. *Genetic Epidemiology*. 2010; 34:67–77. [PubMed: 19557751]
- Larribe F, Fearnhead P. On composite likelihood in statistical genetics. *Statistica Sinica*. 2011; 21:43–69.
- Lin WY, Schaid DJ. Power comparisons between similarity-based multilocus association methods, logistic regression, and score tests for haplotypes. *Genet Epidemiol*. 2009; 33:183–197. [PubMed: 18814307]
- Lindsay BG. Composite likelihood methods. *Contemporary Mathematics*. 1988; 80:221–239.
- Pace L, Salvan A, Sartori N. Adjusting composite likelihood ratio statistics. *Statistica Sinica*. 2011; 21:129–148.
- Pan W. Asymptotic tests of association with multiple SNPs in linkage disequilibrium. *Genetic Epidemiology*. 2009; 33:497–507. [PubMed: 19170135]
- Pan W. A Unified Framework for Detecting Genetic Association with Multiple SNPs in a Candidate Gene or Region: Contrasting Genotype Scores and LD Patterns between Cases and Controls. *Human Heredity*. 2010; 69:1–13. [PubMed: 19797904]
- Satterthwaite FE. An Approximate Distribution of Estimates of Variance Components. *Biometrics Bulletin*. 1946; 2:110–114. [PubMed: 20287815]
- Schaid DJ, McDonnell SK, Hebring SJ, Cunningham JM, Thibodeau SN. Nonparametric tests of association of multiple genes with human disease. *Am J Hum Genet*. 2005; 76:780–793. [PubMed: 15786018]
- Schymick JC, Scholz SW, Fung HC, et al. Genome-wide genotyping in amyotrophic lateral sclerosis and neurologically normal controls: first stage analysis and public release of data. *Lancet Neurol*. 2007; 6:322–328. [PubMed: 17362836]
- Thorisson GA, Smith AV, Krishnan L, Stein LD. The international HapMap project web site. *Genome Res*. 2005; 15:1591–1593.
- Tzeng J-Y, Bondell HD. A comprehensive approach to haplotype-specific analysis by penalized likelihood. *European Journal of Human Genetics*. 2010; 18:95–103. [PubMed: 19584902]
- Varin C, Reid N, Firth D. An overview of composite likelihood methods. *Statistica Sinica*. 2011; 21:5–42.
- Wang T, Zhu X, Elston RC. Improving power in contrasting linkage-disequilibrium patterns between cases and controls. *Am J Hum Genet*. 2007; 80:911–920. [PubMed: 17436245]
- Wang X, Zhang S, Sha Q. A new association test to test multiple-marker association. *Genetic Epidemiology*. 2009; 33:164–171. [PubMed: 18720476]
- Wei Z, Li M, Rebbeck T, Li H. U-statistics-based tests for multiple genes in genetic association studies. *Annals of Human Genetics*. 2008; 72:821–833. [PubMed: 18691161]
- Wessel J, Schork NJ. Generalized genomic distance-based regression methodology for multilocus association analysis. *Am J Hum Genet*. 2006; 79:792–806. [PubMed: 17033957]
- Won, J-H.; Lim, J.; Kim, S.J.; Rajaratnam, B. Technical Report 2009-10. Department of Statistics, Stanford University; 2009. Maximum likelihood covariance estimation with a condition number constraint.

- Zhang J-T. Approximate and asymptotic distributions of Chi-squared-type mixtures with applications. *Journal of the American Statistical Association*. 2005; 100:273–285.
- Zhao Y, Joe H. Composite likelihood estimation in multivariate data analysis. *Canadian Journal of Statistics*. 2005; 33:335–356.

Table 1

Empirical test size (for $\rho_1 = \rho_0$) and power (for $\rho_1 \neq \rho_0$) from 1000 simulations with $k = 6$ SNPs with CS, AR-1 and Rand correlation structures. Statistical significance level is set to be $\alpha = 0.05$.

			M1: Logistic, Main-effects				M2: Logistic, 2-way Interactions					
Corr	ρ_1	ρ_0	Score	SSU	SSUw	UminP	Score	SSU	SSUw	UminP	H-score	H-max
CS	0.2	0.2	0.066	0.044	0.040	0.040	0.034	0.032	0.028	0.038	0.048	0.046
	0.3	0.2	0.044	0.060	0.060	0.064	0.058	0.050	0.05	0.054	0.062	0.064
	0.4	0.2	0.070	0.058	0.062	0.052	0.136	0.078	0.072	0.078	0.148	0.128
	0.5	0.2	0.100	0.066	0.072	0.088	0.400	0.106	0.106	0.130	0.358	0.356
	0.6	0.2	0.134	0.084	0.082	0.112	0.728	0.162	0.160	0.194	0.700	0.714
	0.7	0.2	0.182	0.102	0.108	0.124	0.954	0.218	0.214	0.300	0.938	0.980
	ρ_1	ρ_0	Wald	SSB	SSBw	UminP	CLRT-P	CLRT-S	LRT-P	LRT-S	NLRT	NLRTpc
	0.2	0.2	0.242	0.026	0.032	0.062	0.042	0.040	0.055	0.053	0.062	0.050
	0.3	0.2	0.304	0.054	0.064	0.104	0.056	0.066	0.088	0.091	0.070	0.100
	0.4	0.2	0.466	0.270	0.352	0.194	0.123	0.137	0.223	0.211	0.184	0.278
	0.5	0.2	0.720	0.704	0.764	0.492	0.342	0.368	0.562	0.571	0.552	0.618
	0.6	0.2	0.824	0.958	0.976	0.804	0.744	0.754	0.889	0.891	0.894	0.894
	0.7	0.2	0.888	0.996	0.996	0.976	0.962	0.974	0.992	0.992	0.994	0.988
	ρ_1	ρ_0	Score	SSU	SSUw	UminP	Score	SSU	SSUw	UminP	H-score	H-max
AR-1	0.2	0.2	0.040	0.036	0.044	0.046	0.034	0.040	0.046	0.048	0.048	0.062
	0.3	0.2	0.064	0.062	0.062	0.056	0.050	0.056	0.056	0.058	0.068	0.068
	0.4	0.2	0.068	0.076	0.072	0.054	0.106	0.078	0.074	0.064	0.128	0.086
	0.5	0.2	0.088	0.090	0.082	0.088	0.294	0.096	0.090	0.114	0.254	0.180
	0.6	0.2	0.114	0.106	0.102	0.112	0.666	0.134	0.140	0.182	0.564	0.406
	0.7	0.2	0.158	0.118	0.122	0.134	0.934	0.194	0.186	0.240	0.896	0.784

Corr	ρ_1	ρ_0	M1: Logistic, Main-effects					M2: Logistic, 2-way Interactions					H-max	
			Score	SSU	SSUw	UminP	UminP	Score	SSU	SSUw	UminP	H-score		
L: Latent, Wald-type										L: Latent, CLRT or LRT				
ρ_1	ρ_0	Wald	SSB	SSBw	UminP	UminP	CLRT-P	CLRT-S	LRT-P	LRT-S	NLRT	NLRTpc		
RAND	0.2	0.2	0.210	0.028	0.036	0.064	0.054	0.052	0.054	0.052	0.056	0.028		
	0.3	0.2	0.322	0.030	0.050	0.076	0.051	0.053	0.095	0.097	0.080	0.098		
	0.4	0.2	0.428	0.112	0.164	0.134	0.075	0.094	0.191	0.195	0.174	0.160		
	0.5	0.2	0.662	0.364	0.466	0.318	0.172	0.193	0.475	0.469	0.468	0.336		
	0.6	0.2	0.906	0.766	0.840	0.610	0.395	0.434	0.852	0.869	0.826	0.644		
	0.7	0.2	0.984	0.974	0.994	0.908	0.775	0.800	0.978	0.994	0.992	0.904		
	M1: Logistic, Main-effects										M2: Logistic, 2-way Interactions			
Corr	ρ_1	ρ_0	Score	SSU	SSUw	UminP	UminP	Score	SSU	SSUw	UminP	H-score	H-max	
RAND	0.2	0.2	0.048	0.052	0.054	0.048	0.036	0.062	0.052	0.052	0.040	0.058		
	0.3	0.2	0.054	0.054	0.058	0.064	0.036	0.060	0.058	0.072	0.048	0.038		
	0.4	0.2	0.052	0.062	0.062	0.066	0.118	0.062	0.064	0.080	0.116	0.122		
	0.5	0.2	0.069	0.067	0.069	0.059	0.218	0.072	0.068	0.079	0.248	0.276		
	0.6	0.2	0.080	0.064	0.066	0.074	0.278	0.082	0.086	0.112	0.556	0.546		
	0.7	0.2	0.108	0.082	0.088	0.106	0.610	0.142	0.134	0.172	0.898	0.920		
	L: Latent, Wald-type										L: Latent, CLRT or LRT			
ρ_1	ρ_0	Wald	SSB	SSBw	UminP	UminP	CLRT-P	CLRT-S	LRT-P	LRT-S	NLRT	NLRTpc		
RAND	0.2	0.2	0.244	0.020	0.038	0.060	0.037	0.041	0.051	0.049	0.056	0.054		
	0.3	0.2	0.244	0.042	0.046	0.076	0.051	0.054	0.065	0.063	0.044	0.074		
	0.4	0.2	0.412	0.194	0.252	0.158	0.092	0.103	0.189	0.185	0.168	0.236		
	0.5	0.2	0.620	0.552	0.634	0.384	0.242	0.256	0.441	0.439	0.420	0.506		
	0.6	0.2	0.796	0.894	0.928	0.720	0.546	0.582	0.807	0.802	0.790	0.818		
	0.7	0.2	0.942	0.994	0.998	0.940	0.902	0.912	0.981	0.990	0.980	0.974		

Empirical test size (for $\rho_1 = \rho_0$) and power (for $\rho_1 \neq \rho_0$) from 1000 simulations with $k = 6$ SNPs with a Rand correlation structure. Statistical significance level is set to be $\alpha = 0.05$.

Table 2

Corr	ρ_1	ρ_0	M1: Logistic, Main-effects				M2: Logistic, 2-way Interactions					
			Score	SSU	SSUw	UminP	Score	SSU	SSUw	UminP	H-score	H-max
Rand	0.2	0.2	0.069	0.063	0.063	0.058	0.049	0.041	0.041	0.049	0.047	0.051
	0.4	0.2	0.064	0.061	0.055	0.055	0.055	0.072	0.072	0.099	0.127	0.116
	0.5	0.2	0.081	0.081	0.076	0.079	0.153	0.110	0.107	0.143	0.346	0.363
	0.6	0.2	0.132	0.087	0.087	0.112	0.690	0.171	0.171	0.228	0.736	0.755
L: Latent, CLRT or LRT												
	ρ_1	ρ_0	L: Latent, Wald-type				L: Latent, CLRT or LRT					NLRTpc
			Wald	SSB	SSBw	UminP	CLRT-P	CLRT-S	LRT-P	LRT-S	NLRT	
	0.2	0.2	1.000	0.004	0.011	0.058	0.047	0.039	0.047	0.045	0.041	0.042
	0.4	0.2	1.000	0.209	0.301	0.221	0.118	0.112	0.191	0.202	0.164	0.261
	0.5	0.2	1.000	0.745	0.831	0.516	0.348	0.291	0.490	0.465	0.506	0.679
	0.6	0.2	1.000	0.976	0.987	0.842	0.759	0.729	0.812	0.794	0.869	0.947

Empirical test size (for OR=1) and power (for OR $\neq$ 1) from 1000 simulations with k = 4 SNPs in a region of chromosome 5. Statistical significance level is set to be $\alpha = 0.05$.

M1: Logistic, Main-effects					M2: Logistic, 2-way Interactions					
OR	Score	SSU	SSUw	UminP	Score	SSU	SSUw	UminP	H-score	H-max
1.0	0.056	0.05	0.042	0.046	0.056	0.036	0.038	0.046	0.052	0.046
1.5	0.404	0.370	0.364	0.390	0.412	0.366	0.356	0.408	0.434	0.420
2.0	0.632	0.584	0.578	0.590	0.692	0.580	0.572	0.630	0.646	0.648
2.5	0.674	0.632	0.622	0.640	0.716	0.628	0.620	0.668	0.688	0.678
L: Latent, CLRT or LRT										
OR	Wald	SSB	SSBw	UminP	CLRT-P	CLRT-S	LRT-P	LRT-S	NLRT	NLRTpc
1.0	0.084	0.070	0.052	0.070	0.056	0.057	0.059	0.057	0.076	0.046
1.5	0.436	0.116	0.370	0.402	0.410	0.428	0.342	0.339	0.364	0.350
2.0	0.670	0.140	0.600	0.634	0.646	0.656	0.557	0.561	0.524	0.550
2.5	0.726	0.156	0.654	0.686	0.691	0.695	0.607	0.611	0.558	0.594

Table 4

P-values for the ALS data. Chromosome location (chrom), LD block size (blk size; number of SNPs inside the block), and p-value for a univariate (Univ; single SNP-based) score test are also given.

SNP	chrom	blk size	Univ	L: Latent							
				SSBw				LRT-S			
				CLRT-P	CLRT-S	LRT-P	LRT-S	NLRT	NLRT-S	NLRT-P	NLRTpc
rs4363506	10q26.13	3	3.6e-6	6.2e-5	5.5e-6	<1e-3	4.7e-5	<1e-3	<1e-3	<1e-3	<1e-3
rs16984239	2p24	16	1.0e-5	0.46	0.034	0.019	0.022	0.48	0.48	0.58	0.58
rs12680546	8q24.2	11	5.4e-3	0.13	0.011	7.0e-3	4.5e-3	0.24	0.24	0.065	0.065
rs6013382	20q13.2	5	9.6e-3	5.3e-3	<1e-3	0.035	0.032	0.45	0.45	0.014	0.014
rs2782931	9q31.3	19	0.31	0.22	0.018	0.025	0.028	0.28	0.28	3.0e-3	3.0e-3
rs7976059	12q13	4	8.0e-4	2.0e-4	<1e-3	9.9e-5	0.838	7.0e-3	7.0e-3	0.873	0.873
rs10773543	12q24.32	2	3.0e-4	2.5e-6	<1e-3	1.1e-4	1.1e-4	1e-3	1e-3	<1e-3	<1e-3
rs332389	3p14.1	4	0.81	5.8e-4	<1e-3	<1e-3	8.1e-4	0.086	0.086	0.038	0.038
rs2767584	6p21	7	0.21	0.015	<1e-3	4.0e-3	2.7e-3	0.58	0.58	0.011	0.011

SNP	M1: Logistic, Main-effects				M2: Logistic, 2-way Interactions			
	Score	SSU	SSUw	UminP	Score	SSU	SSUw	UminP
	H-score	H-max	H-score	H-max	H-score	H-max	H-score	H-max
rs4363506	2.6e-5	2.5e-6	6.9e-6	9.3e-6	1.8e-4	9.7e-7	1.5e-6	1.4e-6
rs16984239	0.018	0.016	0.016	3.6e-5	0.023	0.032	0.020	3.7e-5
rs12680546	4.0e-3	7.4e-3	8.0e-3	6.8e-3	4.9e-3	8.1e-3	1.9e-3	0.080
rs6013382	0.095	0.013	0.018	0.035	0.027	0.012	0.010	2.5e-3
rs2782931	0.82	0.67	0.69	0.84	0.37	0.74	0.71	0.26
rs7976059	2.3e-3	6.9e-3	4.4e-3	2.3e-3	7.0e-3	3.3e-3	1.7e-3	1.6e-3
rs10773543	1.3e-4	2.3e-5	2.3e-5	2.2e-4	1.4e-4	2.3e-6	4.2e-6	1.0e-5
rs332389	0.74	0.74	0.81	0.75	3.2e-3	0.17	0.030	1.5e-5
rs2767584	0.13	0.47	0.60	0.57	0.027	0.16	0.10	5.2e-3