A Powerful Method To Test Associations Between Ordinal Traits and Genotypes

Jinjuan Wang,*1 Juan Ding,‡ Shouyou Huang,§ Qizhai Li,*1 and Dongdong Pan**††1

*LSC, NCMIS, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, 100190, China, ‡University of Chinese Academy of Sciences, Beijing, 100049, China, §School of Mathematics and Statistics, Guangxi Normal University, Guilin, 541004, China, **School of Mathematics and Statistics, Hubei Normal University, Huangshi, 435002, China, ††Yunnan Key Laboratory of Statistical Modeling and Data Analysis, Yunnan University, Kunming, 650500, China, and †††Department of Biostatistics, School of Public Health, Yale University, New Haven, CT 06511

ABSTRACT The methods commonly used to test the associations between ordinal phenotypes and genotypes often treat either the ordinal phenotype or the genotype as continuous variables. To address limitations of these approaches, we propose a model where both the ordinal phenotype and the genotype are viewed as manifestations of an underlying multivariate normal random variable. The proposed method allows modeling the ordinal phenotype, the genotype and covariates jointly. We employ the generalized estimating equation technique and M-estimation theory to estimate the model parameters and deduce the corresponding asymptotic distribution. Numerical simulations and real data applications are also conducted to compare the performance of the proposed method with those of methods based on the logit and probit models. Even though there may be potential limitations in Type I error rate control for our method, the gains in power can prove its practical value in case of exactly ordinal phenotypes.

KEYWORDS
association study ordinal phenotype latent normal variate generalized estimating equation M-estimation

Research in the field of genetic epidemiology suggests that some genetic variants play important roles in the etiology of human diseases. On the one hand, genetic variants are defined as genotypes, which are often treated as ordinal variables. On the other hand, there are multiple data types for diseases, i.e., phenotypes, which can be continuous, binary, or ordinal (Li et al. 2006; Kim et al. 2010). Note that both binary and ordinal variables are categorical variables, but the latter can describe the disease state of a patient more precisely in many circumstances. For example, four levels—normal liver, light steatosis, moderate steatosis, and severe steatosis—have been utilized to describe the severity of liver steatosis (Bedogni et al. 2010).

With the development of high throughput biologic technology, increasingly more genotypes and data with complex traits have been generated and deposited in public databases. It is urgently required to develop new statistical testing methods to investigate the associations between these and extract useful information to understand the underlying occurrence and development mechanisms of diseases and traits.

Genome-wide association studies aim to identify associations between phenotypes and genotypes. In these studies, genotypes are often treated as predictors and phenotypes as outcomes. If the phenotype of interest is continuous, then the classic linear regression model is commonly employed. When the phenotype is ordinal, the multinomial logit model (McCullagh 1980; Zhang et al. 2015) or ordered probit model (Daykin and Moffatt 2002; Wang 2014) should be recommended. All these models regress phenotype values or their distribution-based transformations on genotypes, with the assumptions that genotype values are continuous (Korste et al. 2009; Bedogni et al. 2010) and the probability of having a disease increases linearly with the genotype value. However, the continuity assumption on genotype values and the linearity assumption between a phenotype and genotype are difficult to verify in practice. If these two assumptions are violated, the corresponding Wald testing statistics may severely decrease in power.

To overcome this, some researchers treated genotypes as ordinal variables and reversed the regression process by regressing genotypes on phenotypes (O’Reilly et al. 2012). When a phenotype is a continuous
variable, this new method is indeed useful for removing or relaxing the continuity and linearity assumptions. However, this does not work when a phenotype is exactly ordinal, such as in the above-mentioned example of liver steatosis. Therefore, we propose a new method to deal with this problem.

In this work, we treat genotypes as ordinal variables and propose a new procedure to assess the association between an ordinal phenotype and ordinal genotype after adjusting for covariates. Rather than regressing the phenotype on the genotype or regressing the genotype on the phenotype using existing methods, we jointly model the phenotype and genotype by introducing a latent variable following a multivariate normal distribution. The phenotype and genotype are regarded as manifestation values of the latent variable. The relationships between phenotypes, genotypes, and covariates of interest are elaborately described by the covariance matrix. Taking advantage of the framework of generalized estimation equations (Hanley et al. 2003; Zhang et al. 2014) and M-estimation theory (Huber 1981; Stefanski and Boos 2002), we construct a Wald test statistic for an equivalent transformation of the original null hypothesis, and prove that it asymptotically follows the standard normal distribution under the null hypothesis. Numerical simulations are conducted to compare the proposed method with other methods. Our simulation results show that the proposed method can suitably maintain Type I error control and may achieve considerable statistical power compared to existing methods in various scenarios. Finally, we apply the proposed method to anti-cyclic citrullinated protein antibody data for rheumatoid arthritis studies, to further demonstrate its performance.

MATERIALS AND METHODS

Notations

Let random variables $G$ and $Y$ denote a genotype and ordinal phenotype, respectively, and $Z = (Z_1, Z_2, \ldots, Z_{k-2})^T$ be a $(k-2)$-dimensional continuous covariate of interest, with $k \geq 3$, where $\tau$ denotes the transpose of a matrix or vector. Without loss of generality, we assume that there are two alleles at a genetic locus, with one being the risk allele and the other being the reference allele. The value of the random variable $G$ represents the number of the risk allele at a locus, which means that $G$ can take values: 0, 1, or 2. Suppose that $n$ subjects are enrolled in the genetic association study.

Equivalent statement of $H_{01}$ by introducing a latent variable $U$

We assume that the combined vector $(G, Y, Z')^T$ is generated from a $k$-dimensional random variable $U = (U_1, U_2, \ldots, U_k)^T$ following a multivariate normal distribution, with mean vector $\Theta_U = (0, 0, \ldots, 0)^T_{k \times 1}$ (the $k$-dimensional column vector with all units being zero) and covariance matrix $\Delta$, where $\Delta = (\delta_{ikl})_{k \times k}$ and all its diagonal elements are equal to one. That is, $(G, Y)$ represents the manifestation of $U_{01} = (U_1, U_2)$ and $Z = (U_3, \ldots, U_k)^T$. We rewrite $\Delta$ in the partitioned form

$$\Delta = \begin{pmatrix} \Delta_{11} & \Delta_{12} \\ \Delta_{21} & \Delta_{22} \end{pmatrix},$$

where $\Delta_{11}$ is a $2 \times 2$ matrix. Then, $Z$ follows a multivariate normal distribution with mean vector $\Theta_{Z} = (0, 0, \ldots, 0)^T_{(k-2) \times 1}$ and covariance matrix $\Delta_{22}$. Then, $G$ and $Y$ can be obtained as follows:

$$G = \begin{cases} 0, & -\infty < U_1 \leq \xi_1; \\ 1, & \xi_1 < U_1 \leq \xi_2; \\ 2, & \xi_2 < U_1 < +\infty; \end{cases}$$

and

$$Y = \begin{cases} 1, & -\infty < U_2 \leq \eta_1; \\ 2, & \eta_1 < U_2 \leq \eta_2; \\ \ldots \\ m, & \eta_{m-1} < U_2 < +\infty, \end{cases}$$

where $-\infty < \xi_1 < \xi_2 < +\infty$ and $-\infty < \eta_1 < \eta_2 < \ldots < \eta_{m-1} < +\infty$.

Based on the theory of the conditional normal distribution, we have that

$$(U_1, U_2)^T | Z \sim N(\Delta_{12} \Delta_{22}^{-1} Z, \Delta_{11} - \Delta_{12} \Delta_{22}^{-1} \Delta_{21}).$$

We define $\xi_1 = (\delta_{13}, \delta_{14}, \ldots, \delta_{1k})^T, \xi_2 = (\delta_{23}, \delta_{24}, \ldots, \delta_{2k})^T$. Then, the conditional covariance matrix above can be further expressed as

$$\Delta_{11} - \Delta_{12} \Delta_{22}^{-1} \Delta_{21} = \begin{pmatrix} 1 & \delta_{12} \\ \delta_{12} & 1 \end{pmatrix} - \begin{pmatrix} \xi_1 \xi_2^T \\ \xi_2 \xi_1^T \end{pmatrix} \Delta_{22}^{-1} \begin{pmatrix} \xi_1 \xi_2^T \\ \xi_2 \xi_1^T \end{pmatrix}.$$

Now, denote $\delta_0 = (\delta_{12}, \xi_1, \xi_2)^T, \delta(h) = \delta_{12} - \xi_1 \xi_2$. By introducing the latent variable $U$ and taking advantage of its distribution property, we can state that the original hypothesis $H_{01}$ is exactly equivalent to

$$H_{02} : \delta(h) = 0.$$

Proposed statistical test

In this subsection, we construct a test statistic to test $H_{02}$ based on the generalized estimating equation technique and M-estimation theory. Recall that the joint distribution of $U_1$ and $Z'$ is

$$(U_1, Z')^T \sim N(\Theta, \begin{pmatrix} 1 & \xi_1^T \\ \xi_1 & \Delta_{22} \end{pmatrix}).$$

Hence, the conditional distribution of $U_1$ given $Z$ is

$$U_1 | Z \sim N(\xi_1^T \Delta_{22}^{-1} Z, 1 - \xi_1^T \Delta_{22}^{-1} \xi_1).$$

Similarly, the conditional distribution of $U_2$ given $Z$ is

$$U_2 | Z \sim N(\xi_2^T \Delta_{22}^{-1} Z, 1 - \xi_2^T \Delta_{22}^{-1} \xi_2).$$

Denote $\xi_0 = -\infty, \xi_3 = +\infty, \eta_0 = -\infty, \eta_m = +\infty$. It should be noted that the marginal density function of $Z$ and the joint density functions of each pair of variables among $G, Y, Z$ are as follows:

$$f_Z(z; \Delta_{22}) = \frac{1}{(2\pi)^{d/2} |\Delta_{22}|^{1/2}} \exp \left( -\frac{z^T \Delta_{22}^{-1} z}{2} \right)$$
\[
f_{G,Y}(g,Y;\delta_{12}) = \prod_{1 \leq j \leq 3} \prod_{1 \leq r \leq m} \left[ \int_{h_{j-1}}^{\eta_j} \int_{\delta_{12}}^{1} \frac{1}{2\pi \sqrt{1-\delta_{12}}} \exp \left( -\frac{u_j^2 - 2\delta_{12}u_j u_2 + u_2^2}{2(1-\delta_{12})} \right) du_1 du_2 \right]^{I(g=y=r)}(12)
\]

\[
f_{G,Z}(g,z;\xi_1) = f_2(z;\Delta_{22}) \prod_{j=1}^{3} \int_{\eta_{j-1}}^{\eta_j} \frac{1}{\sqrt{2\pi(1 - \xi_1^2 \Delta_{22}^2 \xi_1)}} \exp \left( -\frac{(u_2^2 + \xi_1^2 \Delta_{22}^2 \xi_1 - 2u_2)^2}{2(1 - \xi_1^2 \Delta_{22}^2 \xi_1)} \right) du_2 \tag{13}
\]

\[
f_{Y,Z}(y,z;\xi_2) = f_2(z;\Delta_{22}) \prod_{j=1}^{m} \int_{\eta_{j-1}}^{\eta_j} \frac{1}{\sqrt{2\pi(1 - \xi_2^2 \Delta_{22}^2 \xi_2)}} \exp \left( -\frac{(u_2^2 + \xi_2^2 \Delta_{22}^2 \xi_2 - 2u_2)^2}{2(1 - \xi_2^2 \Delta_{22}^2 \xi_2)} \right) du_2 \tag{14}
\]

where \(I(E)\) is an indicator function, i.e., \(I(E)\) is one if the event \(E\) holds and zero otherwise. The unknown parameters \(\Delta_{22}, \{\xi_j, j = 1, 2\}, \{\eta_j, r = 1, \ldots, m-1\}\) and \(\delta_0\) can be estimated via the following procedures.

First, based on the marginal density function (11) of \(Z\), we have the likelihood function

\[
L(\Delta_{22}) = \prod_{i=1}^{n} f_2(z_i;\Delta_{22}) = \prod_{i=1}^{n} \frac{1}{(2\pi)^{m/2} \Delta_{22}^m} \exp \left( -\frac{\eta_i^2 \Delta_{22}^{-1} z_i}{2} \right).
\]

By maximizing \(L(\Delta_{22})\) on \(\Delta_{22}\), we can obtain the MLE of \(\Delta_{22}\) (denoted by \(\hat{\Delta}_{22}\)), which is the sample covariance matrix of the observed data \(\{z_1, z_2, \ldots, z_n\}\).

Second, in our model both \(G\) and \(Y\) are ordinal variables, whose realizations are determined by the intervals in which the values of two standard normal random variables \(U_1\) and \(U_2\) may fall in, respectively. Specifically, we employ the distribution properties of \(G\) and \(Y\) to intuitively estimate \(\{\xi_j, j = 1, 2\}\) and \(\{\eta_j, r = 1, \ldots, m-1\}\), respectively. We define \(\lambda_1 = \#(g=0, j=1, \ldots, n)\), \(\lambda_2 = \#(g=1, j=1, \ldots, n)\), \(\lambda_3 = \#(g=2, j=1, \ldots, n)\), and \(\omega_r = \#(y=r, j=1, \ldots, n)\) for \(r = 1, 2, \ldots, m\) based on the observed data \(\{g_1, g_2, \ldots, g_n\}\) and \(\{y_1, y_2, \ldots, y_n\}\) of \(G\) and \(Y\), respectively. Recall that \(\xi_0 = \eta_0 = -\infty\). Then, we can estimate \(\xi_j\) as \(\xi_j = 1, 2\) and \(\eta_r\) as \(\eta_r = 1, \ldots, m-1\) by solving the following equations:

\[
\lambda_j = \Phi(\xi_j) - \Phi(\xi_{j-1}), j = 1, 2;
\]

\[
\omega_r = \Phi(\eta_r) - \Phi(\eta_{r-1}), r = 1, \ldots, m;
\]

where \(\Phi(\cdot)\) is the cumulative distribution function of the standard normal variable.

The parameter \(\delta_0\) can be estimated using the generalized estimating equation technique and M-estimation theory. First, let

\[
h(\delta_0) = \sum_{i=1}^{n} \log f_{G,Y}(g_i, y_i; \delta_{12}) + \sum_{i=1}^{n} \log f_{G,Z}(g_i, z_i; \xi_1) + \sum_{i=1}^{n} \log f_{Y,Z}(y_i, z_i; \xi_2).
\]

The function vector consisting of the first-order partial derivatives of \(h(\delta_0)\) with respect to each parameter in \(\delta_0\) is

\[
\phi(G, Y, Z; \xi, \delta_0) = \left[ \frac{\partial \log f_{G,Y}(G, Y; \delta_{12})}{\partial \delta_{12}}, \left( \frac{\partial \log f_{G,Z}(G, Z; \xi_1)}{\partial \xi_1} \right)^T, \left( \frac{\partial \log f_{Y,Z}(Y, Z; \xi_2)}{\partial \xi_2} \right)^T \right]^T.
\]

Then, the estimator \(\hat{\delta}_0 = (\hat{\delta}_{12}, \hat{\xi}_1, \hat{\xi}_2)^T\) of \(\delta_0\) is the root of the following generalized estimating equation

\[
\sum_{i=1}^{n} \phi(g_i, y_i, z_i; \delta_0) = 0.
\]

After estimating all the unknown parameters, the estimate of \(\varphi(\hat{\delta}_0)\) can be expressed as \(\varphi(\hat{\delta}_0) = \hat{\delta}_{12} - \hat{\xi}_1 \hat{\Delta}_{22} \hat{\xi}_2\). To construct a Wald-type test statistic, we need to derive the asymptotic distribution of \(\varphi(\hat{\delta}_0)\). Based on the classical M-estimation theory, \(\hat{\delta}_0\) asymptotically follows a multivariate normal distribution. That is,

\[
\sqrt{n}(\hat{\delta}_0 - \delta_0) \xrightarrow{d} N(0_{2k-3}, V(\delta_0)) \text{ as } n \to \infty,
\]

where \(V(\delta_0) = f'(\delta_0)^{-1} f(\delta_0)^{-1} f'(\delta_0)f(\delta_0)\) and \(f(\delta_0) = -\frac{1}{n} \sum_{i=1}^{n} \varphi(g_i, y_i, z_i; \delta_0)\) is the second derivative of \(\varphi(\delta_0)\). According to the delta method, \(\varphi(\hat{\delta}_0)\) is also asymptotically multivariate normal. Namely,

\[
\sqrt{n}(\varphi(\hat{\delta}_0) - \varphi(\delta_0)) \xrightarrow{d} N(0, V(\delta_0)\varphi(\delta_0)) \text{ as } n \to \infty,
\]

where \(\varphi(\delta_0) = \frac{\varphi(\delta_0)}{\delta_0} = (1 - (\Delta_{22}^2 \xi_2)^T, -\Delta_{22}^2 \xi_1)^T\).

Now, we can propose a new Wald test statistic for the null hypothesis \(H_{02}\) as follows:

\[
T_W = \frac{\varphi(\hat{\delta}_0)^T V(\delta_0) \varphi(\hat{\delta}_0)}{n} \xrightarrow{d} N(0, 1) \text{ as } n \to \infty.
\]

**Data availability**
The authors affirm that all data necessary for confirming the conclusions of the article are present within the article, figures, and tables. Supplemental material available at FigShare: https://doi.org/10.25387/g3.8226650.

**RESULTS**

**Simulation results**
In this subsection, we present a series of simulation studies to investigate the performance of our proposed latent variable model (abbreviated as lvm), and compare it with the probit and logit models, which both regressing an ordinal phenotype on a genotype. We
compared these under multiple simulation scenarios, so that different modeling assumptions would be favored. Two types of data generation mechanism were considered throughout our simulation studies: (i) generating data from a multivariate normal random variable (named the ND mechanism) and (ii) generating data under the proportional odds model (named the PO mechanism). In addition, three genetic models (co-dominant, dominant, and recessive models) were considered.

The specifics of our simulation data generation scenarios are as follows. For simplicity, the dimension of the covariate \(Z\) is one, and the number of levels for a phenotype \(Y\) is set to five. Under the ND mechanism, the three-dimensional latent variable \(U\) was generated from a multivariate normal distribution, with the genotype \(G\) being a manifestation of \(U_1\), the ordinal phenotype \(Y\) being a manifestation of \(U_2\), and the covariate \(Z\) being equal to \(U_3\). Each marginal distribution of \(U_1, U_2,\) and \(U_3\) followed a standard normal distribution. Note that the distribution of \(G\) varied according to the type of the true genetic model. Let \(A\) (major allele) and \(a\) (minor allele) denote two alleles at the single biallelic locus corresponding to the genotype \(G\). Under the Hardy–Weinberg equilibrium (HWE) conditions, the expected genotype frequencies of \(G\) being \(AA, Aa,\) and \(aa\) would be \((1–p)^2, 2p(1–p),\) and \(p^2\), respectively, with the minor allele frequency (MAF) \(p\) taking on values from \((0.1, 0.2, 0.3, 0.4, 0.5)\). If the HWE assumption did not hold, we directly set three kinds of multinomial distribution (P) for \((AA, Aa, aa)\) with different parameter structures \(P_1 = (0.2, 0.3, 0.5), P_2 = (0.2, 0.5, 0.3),\) and \(P_3 = (0.2, 0.7, 0.1)\). When a co-dominant model was assumed, the three genotypes \(AA, Aa,\) and \(aa\) were coded as 0, 1, and 2, respectively. In a model where a dominant effect was assumed, the genotype \(AA\) was coded as 0 while both \(Aa\) and \(aa\) were coded as 1. Accordingly, scores of 0 for both \(AA\) and \(Aa\) and 1 for \(aa\) were employed in a recessive model. In addition, under each genetic model the probabilities of \(Y\) being 1, 2, 3, 4, and 5 were always 0.7, 0.1, 0.08, 0.07, and 0.05, respectively. We set the covariance matrix \(\Sigma\) of \(U\) as

\[
\Sigma_1 = \begin{pmatrix} 1 & 0.4 & 0.5 \\ 0.4 & 1 & 0.8 \\ 0.5 & 0.8 & 1 \end{pmatrix}
\]

(24)

to investigate the Type I error rate, and let \(\Sigma\) be

\[
\Sigma_2 = \begin{pmatrix} 1 & \theta & 0.1 \\ \theta & 1 & 0.1 \\ 0.1 & 0.1 & 1 \end{pmatrix}
\]

(25)
to compare the statistical power under different alternatives, depending on the parameter \(\theta\), whose range was the set \((-0.2, -0.1, 0.1, 0.2)\).

Under the PO mechanism, the ordinal phenotype \(Y\) was related to the genotype \(G\) and covariate \(Z\) through the proportional odds model. Specifically, the distribution of \(G\) under different genetic models would remain the same as that under the ND mechanism, regardless of whether the HWE held. In this case, \(Z\) still followed the standard normal distribution, and \(Y\) was generated with five levels \(\{1, 2, 3, 4, 5\}\) using the proportional odds model

\[
Pr(Y \leq j) = \left[1 + \exp(-\alpha_j + Z + \beta_j(I[G = 0]) + \beta_0, I[G = 1] + \beta_3, I[G = 2])\right]^{-1},
\]

(26)

where \(j = 1, \ldots, 4\), \((\alpha_1, \alpha_2, \alpha_3, \alpha_4) = (0.85, 1.39, 1.99, 2.94)\), \((\beta_{12}, \beta_{23}, \beta_{34}) = v \times (0.15, 0.3), v = 0.5\). It should be noted that \(Pr(Y \leq 5) = 1\). When \(v = 0\), the null hypothesis was true, and the derived simulation data were used to compare the Type I error rates for the three models of interest, *i.e.*, the lvm, probit, and logit models. When \(v = 5\), we explored the power of these three models under different alternatives.
When the true genetic model is recessive and $p = 0.503$, 0.459, and 0.544 under the dominant genetic model, respectively, the empirical powers for the probit, logit, and lvm models are 0.481, and 0.557 under the co-dominant genetic model, and are 0.126, 0.094, and 0.166, respectively.

The results of the three methods under the ND mechanism when the HWE does not hold are presented in Table 2. We can observe that the proposed method achieves a greater power than the other two methods, even though the distribution of the genotype $G$ does not satisfy the HWE conditions. Specifically, the power gain can be as high as 0.07 when $\theta = 0.2$ and the distribution of $G$ is $P_2$.

The corresponding results when the data are generated under the PO mechanism and the HWE is assumed are displayed in Table 3. It follows that all three models can control the Type I error rates under the null hypothesis. Even though the data are generated using the proportional odds model, the lvm method still performs better than the other two methods in detecting an alternative hypothesis, and can achieve a power gain of up to 0.058 in some cases, such as the setting with $p = 0.3$ for the recessive genetic model. It is worth noting that the advantage of the proposed lvm method is more obvious under the recessive model.

As previously described, we considered $5 \times 5 \times 3 + 2 \times 5 \times 3 + 2 \times 3 \times 3 = 168$ scenarios. For each simulation scenario, we generated 1000 datasets, each consisting of 300 subjects. P-values were calculated for each dataset using the three respective models. The nominal level of the tests was set to 0.05, and all simulations were performed using the R language (https://www.r-project.org/).

The empirical Type I error rates and power estimates were calculated using the percentage of rejection in each scenario. The results were presented side-by-side in Tables 1–4. Note that the blanks (marked with —) in these four tables are a result of unavailability of the simulation data under the corresponding scenarios. The reason for this is that in these parameter setting conditions, the mean number for $G$ amounting to 1 is three, which can easily lead to samples with all the $G$ values being 0, such that none of the three considered models can be applied.

Table 1 presents the results for the ND mechanism when the HWE holds. The first five rows suggest that all three methods can control the Type I error rate at the nominal level of 0.05 under the three different genetic models. Furthermore, the remaining rows show that the lvm model enhances the statistical power over the probit and logit models. In some cases, the power gain can be as high as 0.07 to 0.1. For example, when $\theta = 0.2$ and $p = 0.1$, the empirical powers for the probit, logit, and lvm models are 0.514, 0.481, and 0.557 under the co-dominant genetic model, and are 0.503, 0.459, and 0.544 under the dominant genetic model, respectively. When the true genetic model is recessive and $\theta = −0.1$ and $p = 0.3$, the power estimates for the probit, logit, and lvm models are 0.126, 0.094, and 0.166, respectively.

The simulation results under the PO mechanism when the HWE does not hold are presented in Table 4. We observe that the advantage of the proposed lvm method is more obvious under the recessive model.
From all of the four tables, it can be seen that the proposed method might have potential limitations in controlling Type I error rates in a few situations, while the power gains in almost all of simulation scenarios indeed indicate its efficiency for practical applications.

**Application to anticyclic citrullinated protein antibody data for rheumatoid arthritis study**

It is well known that rheumatoid arthritis (RA) is significantly associated with some genetic variants (Carlton et al. 2005; Ruiz-Larraga et al. 2016). The anticyclic citrullinated protein antibody (anti-CCP) can be an auxiliary diagnosis indicator for RA, and the specificity of anti-CCP lies between 87.8% and 96.4% (Coenen et al. 2007). Besides, the genomic region of 6p21.33 has been reported to be associated with RA (Zhang et al. 2009; Zhang and Li 2015). The aim here is to check whether the single nucleotide polymorphisms (SNPs) in the 6p21.33 region are associated with anti-CCP, taking advantage of the proposed lvm test.

Note that there are a total of 45 SNPs in the region of 6p21.33 according to the Genetic Analysis Workshop 16 Data, all of which meet the quality control rule of the MAF being more than 5%, the missing rate being smaller than 15%, and the least genotype frequency being no less than five. The anti-CCP measure takes four values, 1, 2, 3, and 4, and the number of subjects who have these four values were 1195, 103, 66, and 698, respectively. The total number of subjects was 2062. Five principal components coordinate by applying the multi-dimensional scaling method (Li and Yu 2008) to the 127 accurate population structure information SNPs (Yu et al. 2008) were used to adjust for population stratification effects. Before conducting association analysis, we run chi-square tests to check whether HWE holds for each of these 45 SNPs in controls. The P-value results are summarized in Table 5. At a 0.05 level of significance, we can state that the HWE law holds for all SNPs in controls on the basis of their Bonferroni-corrected P-values. Then we apply the probit model, the logit model, and the lvm model to these 45 SNPs to test their association with anti-CCP in sequence. The results are presented in Table 5. It shows that after Bonferroni correction, the SNPs rs2246986, rs3093998, rs2071596, and rs2844490 were found to be significant under the probit and logit models, while the SNPs rs2516398, rs2844494, rs3130637, rs3093993, rs3095227, rs2259435, and rs2844450 were identified as significant using the lvm method. Though these two groups of SNPs overlap at only one SNP rs2844490, each of other SNPs found by the probit (logit) model is physically close to one or two SNPs found by the lvm model. For example, the SNP rs2246986 of the first group is 677 kb away from rs2516398 on one side and is 1212 kb away from rs2844494 on the other side, while both of these two SNPs rs2516398 and rs2844490 are included in the second group. In addition, the SNP rs3093998 (in the first group) is 2971 kb away from rs3130637 (in the second group). The distances are so short that it is reasonable to infer that the SNPs rs2246986, rs2516398, and rs2844494 contain similar information. So do another two SNPs rs3093998 and rs3130637. In short, for detecting the association between anti-CCP and the genomic region 6p21.33, the proposed lvm method might have potential limitations in controlling Type I error rates in a few situations, while the power gains in almost all of simulation scenarios indeed indicate its efficiency for practical applications.
Type I error rate control in some situations, the power gains prove its usefulness. Even though the proposed method might have potential limitations in certain situations, it can also be applied to other situations where the outcome is continuous. In this regard, Tables 1-4. It is worth mentioning that our proposed test model can reasonably control Type I error rates and achieve high power. The important strength of our method is that it does not require the assumption of linearity and genotype; nor does it treat the genotype as a continuous variable. Rather, our approach only introduces a latent multivariate normal variable to characterise the relationship between the two, which is very reasonable, and generally considerably more useful.

Besides the simulations with respect to significance level of 0.05, we also conducted simulation studies with a lower significance level 0.005. The results are given in Supplementary Table S1. We found that the proposed method can reasonably control Type I error rates and achieve power gains at this lower significance level, similar to those results in Tables 1-4. It is worth mentioning that our proposed test model can also be applied to other situations where the outcome is continuous. In such a situation, we can still apply a joint multivariate normal distribution to model outcomes, genotypes, and covariates simultaneously. Even though the proposed method might have potential limitations in Type I error rate control in some situations, the power gains prove its efficiency in practical applications.

In population-based genetic association studies, hundreds of thousands of subjects are often enrolled to achieve optimal power. It is inevitable that there exists a population stratification effect in such large-scale studies. Not considering the effect of population stratification could lead to many false positive findings, and therefore adjusting for its effect represents the basis for conducting a genetic association analysis (Price et al. 2006; Li and Yu 2008). In this study, to characterize the influence of population stratification when investigating the relationships between ordinal traits and genotypes, we treat these effects as covariates in the proposed lvm method. The numerical results of our simulation studies and real data applications have demonstrated that the strategy is feasible and effective.

ACKNOWLEDGMENTS

We would like to thank the associate editor and the two anonymous reviewers for their insightful comments, which helped us to improve the manuscript. This work is partially supported by the Beijing Natural Science Foundation (Z1800006), National Nature Science Foundation of China (11722113, 11661080, 11731011, 11501134), Applied Basic Research Project of Yunnan Province (2017FB002), and Science and Technology Research Project of the Education Department in Hubei Province (Q20172505).

LITERATURE CITED

Bedogni, G., H. S. Kahn, S. Bellentani, and C. Tiritelli, 2010 A simple index of lipid overaccumulation is a good marker of liver steatosis. BMC Gastroenterol. 10: 98. https://doi.org/10.1186/1471-230X-10-98
Carlton, V. E., X. Hu, A. P. Chokkalingam, S. J. Schrodi, R. Brandon et al., 2005 Ptpn22 genetic variation: evidence for multiple variants associated with rheumatoid arthritis. Am. J. Hum. Genet. 77: 567–581. https://doi.org/10.1086/468189
Coenen, D., P. Verschuuren, R. Westhovens, and X. Bossuyt, 2007 Technical and diagnostic performance of 6 assays for the measurement of citrullinated protein/peptide antibodies in the diagnosis of rheumatoid arthritis. Clin. Chem. 53: 498–504. https://doi.org/10.1373/clinchem.2006.078063
Daykin, A. R., and P. G. Moffatt, 2002 Analyzing ordered responses: A review of the ordered probit model. Underst. Stat. 1: 157–166. https://doi.org/10.1207/S15328031US1003_02
Hanley, J. A., A. Negassa, M. D. Edwards, and J. E. Forrester, 2003 Statistical analysis of correlated data using generalized estimating equations: an orientation. Am. J. Epidemiol. 157: 364–375. https://doi.org/10.1093/aje/kwf215
Huber, P. J., 1981 Robust statistics, John Wiley & Sons, Inc., Hoboken, NJ. https://doi.org/10.1002/0471722520
Kim, K., M. S. Kwon, S. Oh, and T. Park, 2013 Identification of multiple gene-gene interactions for ordinal phenotypes. BMC Med. Genomics 6: 59. https://doi.org/10.1186/1755-8794-6-52-59
Korse, C. M., R. G. Taal, C. A. De Groot, R. H. Bakker, and J. M. Bonfrer, 2009. Chromogranin-a and n-terminal pro-brain natriuretic peptide: an excellent pair of biomarkers for diagnostics in patients with neuroendocrine tumor. J. Clin. Oncol. 27: 4293–4299. https://doi.org/10.1200/JCO.2008.18.7047
Li, J., S. Wang, and Z. B. Zeng, 2006 Multiple-interval mapping for ordinal traits. Genetics 173: 1649–1663. https://doi.org/10.1534/genetics.105.054619
Li, Q., and K. Yu, 2008 Improved correction for population stratification in genome-wide association studies by identifying hidden population structures. Genet. Epidemiol. 32: 215–226. https://doi.org/10.1002/gepi.20296
McCullagh, P., 1980 Regression models for ordinal data (with discussion). J. R. Stat. Soc. B 42: 109–142.
O’Reilly, P. F., C. J. Hoggart, Y. Pomyen, F. C. Calboli, P. Elliott et al., 2012 Multiphen: Joint model of multiple phenotypes can increase discovery in gwas. PLoS One 7: e34861. https://doi.org/10.1371/journal.pone.0034861
Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick et al., 2006 Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. 38: 904–909. https://doi.org/10.1038/ng1847
Ruiz-Larrañaga, O., M. Uribarri, M. C. Alcaro, S. Escorza-Treviño, J. Del Amo et al., 2016 Genetic variants associated with rheumatoid arthritis patients and serotypes in european populations. Clin. Exp. Rheumatol. 34: 236–241.
Stefanski, L. A., and D. D. Boos, 2002 The calculus of m-estimation. Am. Stat. 56: 29–38. https://doi.org/10.1198/000313002753631330
Wang, K., 2014 Testing genetic association by regressing genotype over multiple phenotypes. PLoS One 9: e106918. https://doi.org/10.1371/journal.pone.0106918
Yu, K., Z. Wang, Q. Li, S. Wacholder, D. J. Hunter et al., 2008 Population substructure and control selection in genome-wide association studies. PLoS One 3: e2551. https://doi.org/10.1371/journal.pone.0002551
Zhang, M., Y. Lin, L. Wang, V. Pungapong, J. C. Fleet et al., 2009 Case-control genome-wide association study of rheumatoid arthritis from genetic analysis workshop 16 using penalized orthogonal-components regression-linear discriminant analysis. BMC Proc. 3: S17. https://doi.org/10.1186/1753-5463-3-S17
Zhang, W., and Q. Li, 2015 Nonparametric risk and nonparametric odds in quantitative genetic association studies. Sci. Rep. 5: 12105. https://doi.org/10.1038/srep12105
Zhang, W., Z. Zhang, X. Li, and Q. Li, 2015 Fitting proportional odds model to case-control data with incorporating hardy-weinberg equilibrium. Sci. Rep. 5: 12105. https://doi.org/10.1038/srep12105
Zhang, Y., Z. Xu, X. Shen, and W. Pan, 2014 Testing for association with multiple traits in generalized estimation equations, with application to neuroimaging data. Neuroimage 96: 309–325. https://doi.org/10.1016/j.neuroimage.2014.03.061

Communicating editor: G. de los Campos