RESEARCH ARTICLE

Genetic Epidemiology
OFFICIAL JOURNAL
INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY
www.geneticepi.org

Approximate score-based testing with application to multivariate trait association analysis

Zhiyuan Xu, Wei Pan,* and for the Alzheimer's Disease Neuroimaging Initiative†

Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, Minnesota, United States of America

Received 14 February 2015; Revised 4 May 2015; accepted revised manuscript 11 June 2015.
Published online 22 July 2015 in Wiley Online Library (wileyonlinelibrary.com). DOI 10.1002/gepi.21911

ABSTRACT: **For genome-wide association studies and DNA sequencing studies, several powerful score-based tests, such as kernel machine regression and sum of powered score tests, have been proposed in the last few years. However, extensions of these score-based tests to more complex models, such as mixed-effects models for analysis of multiple and correlated traits, have been hindered by the unavailability of the score vector, due to either no output from statistical software or no closed-form solution at all. We propose a simple and general method to asymptotically approximate the score vector based on an asymptotically normal and consistent estimate of a parameter vector to be tested and its (consistent) covariance matrix. The proposed method is applicable to both maximum-likelihood estimation and estimating function-based approaches. We use the derived approximate score vector to extend several score-based tests to mixed-effects models. We demonstrate the feasibility and possible power gains of these tests in association analysis of multiple and correlated quantitative or binary traits with both real and simulated data. The proposed method is easy to implement with a wide applicability.**
Genet Epidemiol 39:469–479, 2015. © 2015 Wiley Periodicals, Inc.

KEY WORDS: **GWAS; kernel machine regression; mixed-effects models; multiple traits; SNP; sum of powered score (SPU) tests**

Introduction

To detect genetic associations in genome-wide association studies (GWASs) and DNA sequencing studies, in addition to the popular univariate minimum *P*-value (UminP) test, many multivariate methods have been proposed to improve statistical power. Several competitive ones are score based, such as the classic score test, a variance-component score test in kernel machine regression (KMR) [Kwee et al., 2008; Wu et al., 2010, 2011], an adaptive score test [Lin and Tang, 2011], and an adaptive sum of powered score (aSPU) test [Pan et al., 2014]. A challenge is how to extend these score-based tests to more complex models beyond the generalized linear models (GLMs) for independent data. There are several reasons to consider more complex mixed-effects models in genetic association studies. First, even for a single-trait analysis, to properly account for some complex and hidden relatedness among the study subjects, or more generally for population structure or population stratification, mixed-effects models have been proposed as a general and effective approach [e.g., Yu et al., 2011; Zhang et al., 2010; Zhou and Stephens, 2014]. These mixed-effects models differ from the standard ones in that a random effect is introduced to induce correlations among *all* the subjects, thus requiring some special and fast algorithms for model fitting as implemented in several recent software packages. These packages do not directly output the score vector. Second, there has been increasing interest in association analysis of multiple traits, which may help gain power and shed light on pleiotropy. In addition, one may encounter correlated traits as arising from familial studies. To account for correlations among multiple traits, either marginal models (based on generalized estimating equations, GEE) [Liang and Zeger, 1986] or mixed-effects models [Breslow and Clayton, 1993] can be applied. For quantitative traits, a linear mixed-effects model (LMM) can be used, from which the score vector can be derived. Accordingly the KMR test has been extended to LMMs [Maity et al., 2012; Schifano et al., 2012]. However, it is unclear how to extend the KMR and other score-based tests to GLM models (GLMMs) and Cox mixed-effects models, for which there is no closed form for the score vector (because the marginal likelihood involves an integral with random effects and in general has no closed form) [Breslow and Clayton, 1993]. Although as an alternative to GLMM, marginal models/GEE can be used, from which (generalized) score-based tests can be derived [Wang et al., 2013; Zhang et al., 2014], there may be substantial differences between the two in terms of modeling assumptions, interpretation, and thus their choices [Diggle et al., 2013]. More importantly, in genetic association studies, as discussed earlier, random effects may be necessary to effectively account for population structure, prompting the use of non-LMMs. In these situations, due to

†Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

*Correspondence to: Wei Pan, A460 Mayo, MMC 303, 420 Delaware St SE, Minneapolis, MN 55455. E-mail: weip@biostat.umn.edu

the lack of computer output or closed-form solution for the score vector, it is challenging to implement score-based tests.

In this paper, we propose a simple and yet general method to approximate the score vector for any model. It is based on the asymptotics of an estimator of the parameters to be tested. It applies to both the maximum-likelihood estimates (MLEs) and estimating function-based estimates. Its implementation involves only a few lines of R code. We demonstrate its use in two types of mixed-effects models, a multivariate LMM (mvLMM) proposed very recently for genetic association analysis of multiple quantitative traits while correcting for cryptic relatedness and population stratification [Zhou and Stephens, 2014], and a GLMM with correlated binary traits. Among others, we use both real and simulated data to illustrate possible power gains of some approximate score-based tests over the standard Wald and UminP tests.

Methods

Review: Some Score-Based Tests

Suppose $U = (U_1, \ldots, U_k)^T$ is the score vector for a set of k parameters to be tested with H_0: $\psi = 0$. The classic score test is

$$T_{Sco} = U^T \widehat{\mathrm{Cov}}(U)^{-1} U,$$

which is asymptotically equivalent to the Wald test and likelihood ratio test (LRT). The UminP test that has been widely used in GWASs can be written as

$$T_{UminP} = \max_{j=1}^{k} U_j^2 / V_{jj},$$

where V_{jj} is the jth diagonal element of $V = \widehat{\mathrm{Cov}}(U)$. Recently, a variance-component score test in KMR has been proposed for GLMs and shown to be powerful for analysis of single nucleotide polymorphism (SNP) sets [Kwee et al., 2008; Wu et al., 2010, 2011]. As discussed by Pan (2011), with a linear kernel it is equivalent to the sum of squared score (SSU) test:

$$T_{SSU} = U^T U = \sum_{j=1}^{k} U_j^2.$$

Pan et al. [2014] proposed a family of the so-called SPU tests:

$$T_{SPU(\gamma)} = \sum_{j=1}^{p} U_j^{\gamma} \qquad (1)$$

for a set of integers $\gamma \geq 1$. It is easy to see that SPU(1) and SPU(2) are exactly the same as the Sum test and the SSU test, respectively [Pan, 2009]; the sum test, an example of so-called burden tests, has been shown to perform well for genetic association testing, especially for rare variants, under some situations [Li and Leal, 2008; Pan, 2009]. In addition, for an even integer $\gamma \to \infty$, we have

$$T_{SPU(\gamma)} \propto \left(\sum_{j=1}^{p} |U_j|^{\gamma} \right)^{1/\gamma} \to \max_j |U_j| = T_{SPU(\infty)},$$

the SPU(∞) is closely related to the UminP test (but ignoring possibly varying variances of U_j's). Alternatively, an SPU(γ) test can be regarded as a weighted score test [Lin and Tang, 2011] with adaptive weights $U_j^{\gamma-1}$ on each component j. In practice, because it is unknown which γ value would yield high power, we use an adaptive SPU (aSPU) test to combine the evidence across the SPU tests:

$$T_{aSPU} = \min_{\gamma \in \Gamma} P_{SPU(\gamma)}, \qquad (2)$$

where $P_{SPU(\gamma)}$ is the P-value of the SPU(γ) test, and Γ contains a set of candidate values γ. Pan et al. [2014] found that in many situations $\Gamma = \{1, 2, 3, \ldots, 8, \infty\}$ appeared to perform well, which will be used here.

In general, resampling methods can be used to obtain P-values for the SPU and aSPU tests. In this paper, we assume that the asymptotic null distribution of the score vector $U \sim N(0, V)$ holds (under H_0). Accordingly, we can generate B-independent copies of the null score vector $U^{(b)}$, for which the B copies of the SPU test statistics can be calculated. Then the P-value of each SPU(γ) test is calculated as $P_{SPU(\gamma)} = \sum_{b=1}^{B} I(|T_{SPU(\gamma)}^{(b)}| \geq |T_{SPU(\gamma)}|)/B$. Furthermore, based on the same B copies of the simulated score vector, we calculate the P-value for the aSPU test as $P_{aSPU} = \sum_{b=1}^{B} I(T_{aSPU}^{(b)} \leq T_{aSPU})/B$ with $T_{aSPU}^{(b)} = \min_{\gamma \in \Gamma} p_{\gamma}^{(b)}$ and $p_{\gamma}^{(b_1)} = \sum_{b \neq b_1} I(|T_{SPU(\gamma)}^{(b)}| \geq |T_{SPU(\gamma)}^{(b_1)}|)/(B-1)$.

In this paper, we use the SPU and aSPU tests as examples, though other score-based tests [e.g., Lin and Tang, 2011; Sun et al., 2013; Wu et al., 2010] can be equally applied. Our main point does not depend on the choice of a specific score-based test; rather, we aim to show how to extend a score-based test to cases where there is no easy access to the score vector, as arising in below two important applications. To be concrete, we focus on detecting genetic association with SNPs, but the proposed method is generally applicable to other problems of interest.

Two Example Models

Multivariate Linear Mixed Model

A multivariate linear mixed model (mvLMM) was proposed by Zhou and Stephens [2014] to test for association with multiple phenotypes while correcting for possible population stratification. Specifically, suppose we would like to test for association between a multivariate trait and a single SNP. We first combine the n trait vectors $Y_i^T = (Y_{i1}, \ldots, Y_{ik})$ by row such that the resulting trait matrix Y is of dimension $n \times k$, and the jth column of Y corresponds to phenotype j while the ith row of Y corresponds to the multiple traits from the ith subject; W is an $n \times q$ design matrix for covariates (including a column of 1's for the intercept); $x = (x_1, \ldots, x_n)^T$ is an $n \times 1$ vector of genotype scores (i.e., the counts of the minor allele) for the SNP.

The mvLMM can be written as

$$Y = W\lambda + x\psi^T + G + E,\ G \sim \mathrm{MN}_{n\times k}(0, K, V_g),$$
$$E \sim \mathrm{MN}_{n\times k}(0, I_{n\times n}, V_\epsilon), \qquad (3)$$

where λ is a $q \times k$ matrix of regression coefficients for covariates; ψ is a $k \times 1$ vector of the SNP effect sizes for the k phenotypes; G is an $n \times k$ matrix of random effects; E is an $n \times k$ matrix of random errors; K is an $n \times n$ known kinship matrix, or more generally, a genetic relatedness matrix (GRM) estimated from whole-genome genotype data; $I_{n\times n}$ is an $n \times n$ identity matrix; V_g is a $k \times k$ symmetric matrix of genetic variance components; V_ϵ is a $k \times k$ symmetric matrix of environmental variance components; and $\mathrm{MN}_{n\times k}(0, V_1, V_2)$ denotes the $n \times k$ matrix normal distribution with mean 0, a column covariance matrix V_1 of dimension $n \times n$, and a row covariance matrix V_2 of dimension $k \times k$. The goal is to test H_0: $\psi = 0$.

A mvLMM differs from a standard LMM in that an $n \times n$ matrix K is used to account for possible genetic relatedness among all the subjects. Because the kinship matrix K may be full and may not be block diagonal, it means that *all* the subjects may be possibly correlated. However, as discussed by Zhou and Stephen [2014], the mvLMM can be rewritten more like a standard LMM in the following way. An eigendecomposition of the relatedness matrix K can be performed as $K = U_k D_k U_k^T$, where U_k is a $n \times n$ matrix of eigenvectors and D_k is a diagonal $n \times n$ matrix with diagonal elements corresponding to eigenvalues (i.e., $\mathrm{diag}(\delta_1, \ldots, \delta_n)$). Then one can obtain the transformed phenotype matrix $\tilde{Y} = U_k Y$, transformed covariate matrix $\tilde{W} = U_k W$, transformed SNP vector $\tilde{x} = U_k x$, transformed random effect matrix $\tilde{G} = U_k G$, and transformed residual error matrix $\tilde{E} = U_k E$. After transformation, for each individual i, the transformed phenotypes given the transformed covariates and SNP follow independent (but not identical) multivariate normal distributions:

$$\tilde{y}_i = \lambda^T \tilde{w}_i + \psi \tilde{x}_i + \tilde{g}_i + \tilde{e}_i,\ \tilde{g}_i \sim N(0, \delta_i V_g),\ \tilde{e}_i \sim N(0, V_e), \tag{4}$$

where for $i = 1, \ldots, n$, $\tilde{y}_i{}^T$ is the ith row vector of $\tilde{Y}$, $\tilde{w}_i{}^T$ is the ith row vector of $\tilde{W}$, $\tilde{x}_i$ is the ith element of vector $\tilde{x}$, $\tilde{g}_i{}^T$ is the ith row vector of $\tilde{G}$, and $\tilde{e}_i{}^T$ is the ith row vector of $\tilde{E}$; $\mathrm{Var}(\tilde{y}_i) = \delta_i V_g + V_e \equiv V_i$.

Based on model (4), one can write down the score vector:

$$U = \sum_{i=1}^{n} (\tilde{W}_i, \tilde{X}_i)^T \hat{V}_{i,0}^{-1} (\tilde{y}_i - \hat{\lambda}_0^T \tilde{w}_i), \tag{5}$$

where $\hat{\lambda}_0$ and $\hat{V}_{i,0}$ are obtained by fitting the null model under $H_0 : \tilde{y}_i = \lambda^T \tilde{w}_i + \tilde{g}_i + \tilde{e}_i$.

It is quite challenging to develop a fast algorithm to fit a mvLMM. Now such an algorithm is implemented in software package `GEMMA` [Zhou and Stephens, 2014]. However, as for most software packages, one is not able to obtain the score vector directly from the output. A simple and practical way to obtain the score vector is, as proposed earlier, to approximate it by the MLE and its covariance estimate, both available directly from the output of `GEMMA`; accordingly a score-based test can be simply constructed and applied.

Generalized Linear Mixed Model

In a familial study, we observe that in each family i, subject j has a univariate trait Y_{ij}, q covariates $W_{ij} = (W_{ij1}, \ldots, W_{ijq})^T$ and p SNPs $X_{ij} = (X_{ij1}, \ldots, X_{ijp})^T$. We would like to test for association between the trait and the SNPs through a GLMM:

$$g(\mu_{ij}) = W_{ij}^T \lambda + X_{ij}^T \psi + b_i,\ b_i \sim N(0, \sigma_b^2), \tag{6}$$

where $g()$ is a link function, $\mu_{ij} = E(Y_{ij} | X_{ij}, W_{ij}, b_i)$ is the conditional mean of the trait for subject j in family i, $\lambda = (\lambda_1, \ldots, \lambda_q)^T$ is a $q \times 1$ vector of regression coefficients for covariates W_{ij}, $\psi = (\psi_1, \ldots, \psi_p)^T$ a $p \times 1$ vector of regression coefficients for SNP set X_{ij}, and b_i is a random effect inducing correlations among the traits of the subjects from the same family.

The goal is to test H_0: $\psi = 0$. However, in general, due to the lack of the closed form for the marginal likelihood, there is no closed-form expression for the score vector for ψ either [Breslow and Clayton, 1993]. Hence, it is not easy to develop a score-based test for such a model. Below we propose a new method to approximate the score vector, based on which it is straightforward to construct a score-based test.

New Method: Approximating the Score Vector

Estimation via Maximum Likelihood

Suppose that we would like to test H_0: $\psi = \psi_0$ in the presence of nuisance parameter λ. Denote $\hat{\theta}_0 = (\psi_0, \hat{\lambda}_0)^T$ as the restricted MLE of $\theta = (\psi, \lambda)'$ under H_0, while $\hat{\theta} = (\hat{\psi}, \hat{\lambda})^T$ as the unrestricted MLE of $\theta = (\psi, \lambda)^T$ (i.e., under H_1). Partition the Fisher's information matrix H accordingly as

$$H = \begin{pmatrix} H_{\psi\psi} & H_{\psi\lambda} \\ H_{\lambda\psi} & H_{\lambda\lambda} \end{pmatrix},\ H^{-1} = \begin{pmatrix} H^{\psi\psi} & H^{\psi\lambda} \\ H^{\lambda\psi} & H^{\lambda\lambda} \end{pmatrix}.$$

Denote the whole score vector for θ as $U_\theta(\theta) = (U_\psi(\theta)^T, U_\lambda(\theta)^T)^T$. As shown by Kent [1982, the equation following (4.1)],

$$U \equiv U_\psi(\hat{\theta}_0) = (H^{\psi\psi})^{-1}(\hat{\psi} - \psi_0) + o_p(1). \tag{7}$$

Because the consistent estimator $\widehat{\mathrm{Cov}}(\hat{\psi}) = H^{\hat{\psi}\hat{\psi}}$, we have

$$U \approx \widehat{\mathrm{Cov}}(\hat{\psi})^{-1}(\hat{\psi} - \psi_0), \quad \widehat{\mathrm{Cov}}(U) = \widehat{\mathrm{Cov}}(\hat{\psi})^{-1}.$$

Thus, we first fit a full model (under H_1) to obtain the MLE $\hat{\psi}$ and its covariance estimate $\widehat{\mathrm{Cov}}(\hat{\psi})$, then we can approximate the score vector U and its covariance matrix accordingly. In this way, we can construct (approximate) score-based tests such as the score test, the SPU, and aSPU tests. In particular, it is easy to see that the approximate score-based score test is the same as the Wald test:

$$U^T \widehat{\mathrm{Cov}}(U)^{-1} U = (\hat{\psi} - \psi_0)^T \widehat{\mathrm{Cov}}(\hat{\psi})^{-1} (\hat{\psi} - \psi_0).$$

Estimation via Estimating Functions

For estimating function-based approaches, although a generalized score test [e.g., Boos, 1992; Kent, 1982; Rotnitzky and

Jewell, 1990] can be constructed, the popular statistical software may not provide direct output of such (generalized) score vectors. For easy implementation, it may be useful to approximate the (generalized) score vector by the parameter estimate and its covariance matrix. Specifically, by treating an unbiased mean 0 estimating function as a (generalized) score function and by a Taylor expansion, we still have Equation (7). However, $\widehat{\mathrm{Cov}}(\hat{\psi})_M = H^{\hat{\psi}\hat{\psi}}$ is the model-based covariance estimator, which is not consistent unless all working assumptions hold (essentially assuming that the estimating function is indeed a score function). More generally, a consistent sandwich estimator $\widehat{\mathrm{Cov}}(\hat{\psi})_S$ is used. Hence, we can modify the score vector approximation as

$$U \approx \widehat{\mathrm{Cov}}(\hat{\psi})_M^{-1}(\hat{\psi} - \psi_0),$$
$$\mathrm{Cov}(U) \approx \widehat{\mathrm{Cov}}(\hat{\psi})_M^{-1}\widehat{\mathrm{Cov}}(\hat{\psi})_S\widehat{\mathrm{Cov}}(\hat{\psi})_M^{-1}.$$

Accordingly, once we obtain the point estimate $\hat{\psi}$, its model-based covariance estimate $\widehat{\mathrm{Cov}}(\hat{\psi})_M$ and its sandwich estimate $\widehat{\mathrm{Cov}}(\hat{\psi})_S$, we can obtain an approximation to the score vector U, based on which we can construct a score-based test. Again it is easy to verify that the score test based on the approximate score is exactly the same as the Wald test.

We explored the use of such tests for marginal approaches to GLMMs for correlated binary data (i.e., GEE) [Liang and Zeger, 1986] (and to Cox regression for correlated survival data; not shown). Note that in general our proposed approximate score vector is derived based on an asymptotically normal point estimator, and thus is only asymptotically unbiased, while the generalized score vector is simply the estimating function being used and is often unbiased for finite samples; this difference leads to varying performances of an approximate score-based test and an exact generalized score test [Boos, 1992] for finite samples, though their difference diminishes as the sample size increases, as to be shown later for GEE.

Results

Example

ADNI Data

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies, and nonprofit organizations, as a $60 million, 5-year public–private partnership. The primary goal of ADNI has been to test whether serial magnetic resonance imaging, positron emission tomography, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials.

The principal investigator of this initiative is Michael W. Weiner, MD, VA Medical Center and University of California-San Francisco. ADNI is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the United States and Canada. The initial goal of ADNI was to recruit 800 subjects but ADNI has been followed by ADNI-GO and ADNI-2. To date these three protocols have recruited over 1,500 adults, ages 55–90, to participate in the research, consisting of cognitively normal older individuals, people with early or late MCI, and people with early AD. The follow-up duration of each group is specified in the protocols for ADNI-1, ADNI-2, and ADNI-GO. Subjects originally recruited for ADNI-1 and ADNI-GO had the option to be followed in ADNI-2. For up-to-date information, see www.adni-info.org.

We applied the methods to the ADNI-1 data consisting of 681 non-Hispanic Caucasians with both genotypic and phenotypic data. The phenotypes were cortical thickness measures of some regions of interest (ROIs) in the brain; they were cross-sectionally processed using FreeSurfer by UCSF researchers [Hartig et al., 2012]. We tested on about 20 SNPs and several multivariate traits as considered in Shen et al. [2010] and Zhang et al. [2014]. For the purpose of illustration, we only show the results for two SNPs and four multivariate traits: APOE-ϵ4 in gene APOE that is well known to be associated with AD, and rs7526034 on chromosome 1 (LOC199897), both were associated with multiple neuroimaging phenotypes [Shen et al., 2010]; the four multivariate traits were left and right sides of "Par" (denoted as LPar and RPar), each with four ROIs (inferior and superior parietal gyri, supramarginal gyrus, and precuneus), right side of "Front" with six ROIs (caudal midfrontal, rostral midfrontal, superior frontal, lateral orbitofrontal, medial orbitofrontal gyri, and frontal pole), right side of "LatTemp" with three ROIs (inferior temporal, middle temporal, and superior temporal gyri). Given a large number of parameters to be estimated ($> k^2$ with k traits) in a mvLMM, as pointed out by Zhou and Stephen [2014], only a small to moderate number of phenotypes ($\sim$ 2–10) were recommended to be used for a typical sample size for GWAS (i.e., n in thousands). Hence, with only a moderate sample size $n = 681$ here, we only considered a few multivariate traits containing no more than six univariate traits (otherwise, in addition to the questionable asymptotics, we also encountered some numerical convergence problems).

For each subject i, five covariates (W_i) were included: baseline age, gender, baseline education (in years), handedness (left or right), and baseline intracranial volume, plus an intercept term; X_i was the genotypic score of one of the above two SNPs; $Y_i = (Y_{i1}, \ldots, Y_{ik})^T$ was a vector of k quantitative traits (i.e., cortical thickness measures of k ROIs). For each pair of the SNP-multivariate trait, a mvLMM (3) was fitted using software GEMMA; we applied the proposed method to approximate the score vector based on the MLE and its

Table 1. *P*-values of the various mvLMM-based tests in analysis of the ADNI data

SNP	Test	LPar(4)	RFront(6)	RLatTemp(3)	RPar(4)
APOE-ϵ4	SPU(1)	1.2×10^{-5}	2.0×10^{-7}	2.0×10^{-5}	1.3×10^{-6}
	SPU(2)	8.3×10^{-2}	2.4×10^{-2}	1.3×10^{-1}	3.8×10^{-4}
	SPU(3)	1.9×10^{-1}	2.4×10^{-3}	1.9×10^{-2}	1.3×10^{-2}
	SPU(4)	1.6×10^{-1}	7.7×10^{-3}	6.8×10^{-2}	1.1×10^{-3}
	SPU(5)	3.6×10^{-1}	4.3×10^{-3}	3.5×10^{-2}	2.1×10^{-2}
	SPU(∞)	2.9×10^{-1}	5.2×10^{-3}	4.8×10^{-2}	7.8×10^{-3}
	aSPU	3.5×10^{-5}	5.0×10^{-7}	6.0×10^{-5}	3.8×10^{-6}
	Wald	1.2×10^{-5}	1.4×10^{-5}	4.6×10^{-5}	6.0×10^{-8}
	Score	1.8×10^{-5}	2.3×10^{-5}	5.9×10^{-5}	1.4×10^{-7}
rs7526034	SPU(1)	5.7×10^{-2}	9.0×10^{-4}	$<1.0\times10^{-7}$	6.2×10^{-2}
	SPU(2)	8.2×10^{-1}	5.7×10^{-2}	2.0×10^{-1}	1.4×10^{-2}
	SPU(3)	6.5×10^{-1}	5.8×10^{-2}	6.8×10^{-2}	4.8×10^{-2}
	SPU(4)	8.4×10^{-1}	8.3×10^{-2}	2.1×10^{-1}	2.6×10^{-2}
	SPU(5)	7.7×10^{-1}	7.5×10^{-2}	1.5×10^{-1}	4.8×10^{-2}
	SPU(∞)	8.4×10^{-1}	8.4×10^{-2}	2.7×10^{-1}	4.0×10^{-2}
	aSPU	1.3×10^{-1}	2.3×10^{-3}	$<1.0\times10^{-7}$	3.5×10^{-2}
	Wald	2.6×10^{-1}	1.4×10^{-5}	2.0×10^{-6}	3.5×10^{-4}
	Score	2.6×10^{-1}	1.6×10^{-5}	3.1×10^{-6}	4.3×10^{-4}

covariance estimate, from which the approximate score-based SPU and aSPU tests were conducted. The K matrix in the mvLMM was estimated based on nearly a half million SNPs in the data (prior to fitting the model) in `GEMMA`. We used a step-up procedure to gradually increase the value of the simulation number B to calculate the P-values: starting from $B = 10^4$, if a P-value was no more than $5/B$, we increased B to 10 times of its previous value and then repeated the test until either its P-value was more than $5/B$ or we reached $B = 10^7$. For comparison, we also showed the results of the asymptotic Wald and Score tests, directly available from the output of `GEMMA`.

Analysis Results

As shown in Table 1, for most SNP-multivariate trait pairs, the aSPU test gave similar results as those of the classic Wald and Score tests. However, there were a few differences. Notably, for rs7526034-RLatTemp, the aSPU test gave a more significant P-value than those of the Wald and Score tests; on the other hand, for APOE-ϵ4-RPar, it was the reverse. Among the SPU tests, SPU(1) usually gave most significant results, presumably because of the smaller number of univariate traits (k) and the same direction of the SNP-univariate trait associations. In summary, our results demonstrate the feasibility of using our proposed method to approximate the score vector for a complex mvLMM, and accordingly construct the score-based aSPU test, which might be more powerful in some situations (to be shown in simulations) and thus can be complementary to the standard Wald and Score tests.

Simulations

Simulation I: mvLMM

To mimic real data, we used the ADNI data to generate multivariate phenotypes according to the fitted the mvLMM models (3) while using the covariates and SNPs in the ADNI data too. Specifically, two SNP-phenotype pairs, rs7526034-RLatTemp and APOE-ϵ4-RPar were chosen; from their corresponding fitted models (3), we obtained the parameter estimates such as $\hat{\lambda}$, $\hat{\psi}$, $\hat{V}_g$, and $\hat{V}_\epsilon$. Those parameter estimates except $\hat{\psi}$, which was either 0 for the null model or was scaled by a factor $1/2$ to reduce the effect size of the SNP for the non-null model (because we were using a nominal significance level at 0.05), were then used to simulate the phenotypes by model (3). For each simulated dataset, as before, the MLE of ψ and its covariance estimate were obtained from `GEMMA` to approximate its score vector so that the SPU and aSPU tests could be applied. We used $B = 1,000$ to calculate their P-values. As a comparison, we also used the Wald test, Score test, and LRT directly output by `GEMMA`. Based on 5,000 replicates for each setup, we obtained the empirical Type I error and power estimates. However, we note that there were some convergence problems when running GEMMA for about 1% and 2% of simulated datasets for the two SNP-phenotype pairs, respectively; our results were based on the remaining ones without any convergence problems.

Table 2. Simulation I: empirical Type I error and power for two SNP-phenotype pairs, rs7526034-RLatTemp (pair 1) and APOE-ϵ4-RPar (pair 2), based on fitting mvLMM

		Approximate score vector									
		SPU(γ)									
Pair		$\gamma = 1$	2	3	4	5	∞	aSPU	Wald	Score	LRT
1	Type I	0.062	0.067	0.067	0.066	0.069	0.068	0.066	0.067	0.067	0.178
	Power	0.722	0.103	0.195	0.112	0.139	0.120	0.621	0.599	0.591	0.604
2	Type I	0.058	0.065	0.063	0.066	0.065	0.067	0.068	0.064	0.058	0.188
	Power	0.644	0.389	0.335	0.381	0.336	0.361	0.643	0.689	0.674	0.641

As shown in Table 2, the Type I error rates for both pairs were only slightly inflated for all the tests except the LRT, which had largely inflated Type I error rates. The inflation could be due to a large number of parameters to be estimated in an mvLMM with a moderate sample size.

Because the Type I error rates based on fitting the mvLMM were slightly inflated, while it was known that there was barely any population stratification in the ADNI data [Xu et al., 2014], we fitted the corresponding model after treating $K = I$; however, we experienced some numerical convergence problems in fitting the mvLMM, likely due to that the two unstructured covariance matrices V_g and V_ϵ were not identifiable (after forcing $K = I$). Thus we simply used function `gls()` in R package `nlme` to fit a corresponding marginal linear model with or without top 10 principal components (PCs); the PCs were extracted using Plink [Purcell et al., 2007] based on almost a half million SNPs of the 757 subjects in the ADNI data. As shown in Table 3, the Type I error rates were better controlled. Note that because the simulated data were generated with a $K \neq I$, some slight inflation of a Type I error rate was expected under the incorrect assumption $K = I$.

The two SNP-phenotype pairs were chosen partly because in the ADNI data analysis the aSPU test gave a more significant P-value than those of the Wald and Score tests for

Table 3. Simulation I: empirical Type I error and power for two SNP-phenotype pairs, rs7526034-RLatTemp (pair 1) and APOE-ϵ4-RPar (pair 2) based on fitting marginal linear models (LMs) with or without the top 10 PCs

			Approximate score vector							
			SPU(γ)							
Model	Pair		$\gamma = 1$	2	3	4	5	∞	aSPU	Wald
LM, 10 PCs	1	Type I	0.049	0.057	0.063	0.057	0.061	0.058	0.061	0.059
		Power	0.742	0.092	0.212	0.117	0.158	0.134	0.648	0.611
	2	Type I	0.050	0.052	0.050	0.051	0.048	0.049	0.052	0.052
		Power	0.616	0.347	0.306	0.347	0.316	0.339	0.610	0.664
LM, no PCs	1	Type I	0.049	0.059	0.062	0.057	0.061	0.057	0.060	0.057
		Power	0.749	0.095	0.214	0.116	0.160	0.136	0.645	0.609
	2	Type I	0.052	0.056	0.054	0.057	0.052	0.056	0.054	0.053
		Power	0.633	0.360	0.317	0.361	0.327	0.352	0.634	0.683

pair 1, while it was the opposite for pair 2 (Table 1). It was confirmed that in the simulations the aSPU test was indeed slightly more (or less) powerful than the Wald and Score tests for pair 1 (or pair 2).

Simulation II: LMM

We considered a case with unrelated individuals and multiple quantitative traits similar to those in Zhang et al. [2014], for which the exact score vector could be derived. We would compare the performance of an exact score-based test with that of its approximate score-based one. For each subject $i = 1, \ldots, n$, we generated his/her genotype data as in Pan [2009]. Specifically, for each subject i, we first generated a latent vector $G_i = (G_{j1}, \ldots, G_{j,p+1})'$ from a multivariate normal distribution with a first-order autoregressive (AR-1) covariance structure with parameter $\rho = 0.5$: $\text{Cov}(G_{is}, G_{it}) = \rho^{|s-t|}$. Second, each latent element G_{is} was dichotomized to 0 or 1 with probability $\text{Prob}(G_{is} = 1)$ as its minor allele frequency (MAF), randomly drawn from a uniform distribution. Third, we independently generated another haplotype for subject i, then combined the two haplotypes to form the genotypes for subject j. In this way, we obtained the genotypes of all the subjects.

The first SNP was chosen as the causal one with MAF randomly drawn from a uniform distribution $U(0.3, 0.4)$, while the MAFs of the other SNPs were independently drawn from $U(0.1, 0.5)$. For each subject i, we simulated k traits $Y_i = (Y_{i1}, \ldots, Y_{ik})$ from a linear model:

$$Y_i = \lambda + x_i\psi + \epsilon_i, \tag{8}$$

where $\lambda = (\lambda_1, \ldots, \lambda_k)^T$, $\psi = (\psi_1, \ldots, \psi_k)^T$; $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2 R)$ with $\sigma = 1$ and R as a compound symmetry (CS) correlation matrix with the correlation parameter $r = 0.3$; x_i is the genotype dosage of the causal SNP. Under H_0 we had $\psi = 0$; under H_1 we had $\psi_m \neq 0$ for $1 \leq m \leq 5$ and $\psi_m = 0$ for $5 < m \leq k$. The nonzero ψ_j's were simulated from a uniform distribution $U(0.2, 0.3)$. In this way, under H_1, only the first five traits were associated with the causal SNP; that is, as $k = 5, 10, 20, 30, 40$, we gradually increased the number of the nonassociated traits from 0 to 5, 15, 25, and 35.

The simulated data were fitted by an LMM:

$$Y_{ij} = \lambda_j + x_i\psi_j + b_i + \epsilon_{ij},\ b_i \sim N(0, \sigma_b^2),\ \epsilon_{ij} \sim N(0, \sigma_e^2), \tag{9}$$

where b_i was a normal random effect to model the correlations among multiple traits, ϵ_{ij} was the random error, x_i was a scalar corresponding to the genetic score of the SNP nearest to the causal SNP.

We implemented both the approximate score-based and exact score-based tests. For an approximate score-based test, we fitted model (9) and used the MLE of ψ to approximate its score vector and its variance–covariance matrix. For an exact score-based test, we fitted the reduced LMM modeled under H_0, $Y_{ij} = \lambda_j + b_i + \epsilon_{ij}$, to obtain the MLEs $\hat{\lambda} = (\hat{\lambda}_1, \ldots, \hat{\lambda}_k)^T$. Denote I as the $k \times k$ identity matrix. The exact score vector and its variance–covariance matrix can be written as

$$U = \sum_{i=1}^{n} (I, X_i)^T \hat{V}_i^{-1} (Y_i - \hat{\lambda}),$$

$$Cov(U) = \sum_{i=1}^{n} (I, X_i)^T \hat{V}_i^{-1} (I, X_i), \tag{10}$$

where $\hat{V}_i$ was the MLE of V_i with its diagonal elements $\hat{\sigma}_b^2 + \hat{\sigma}_\epsilon^2$ and off-diagonal elements $\hat{\sigma}_b^2$. Partition the score vector and its covariance into two parts corresponding to the intercept and ψ parameters, respectively,

$$U = \begin{pmatrix} U_1 \\ U_2 \end{pmatrix}, \quad Cov(U) = \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix},$$

then we have the exact score vector for ψ as U_2 and its covariance matrix $Cov(U_2) = V_{22} - V_{21} V_{11}^{-1} V_{12}$.

To estimate the Type I error and power, 1,000 datasets were independently simulated and analyzed. Each of the 1,000 datasets consisted of 1,000 subjects. We used $B = 1,000$ to obtain P-values for any permutation based methods. As a comparison, we also showed the results from the UminP.

As shown in Table 4, first, regardless of the test being examined, its version based on the approximate score vector and that based on the exact score vector gave almost the same results, suggesting the high accuracy of the asymptotic approximation in this case. Second, we note that the Type I error rates were satisfactorily controlled, even for 40 traits. Third, in agreement with Zhang et al. [2014], the aSPU test was more powerful than the score test for five traits, but not for other numbers of traits; both were much more powerful than single trait based UminP test, presumably due to the former two's combining information across the five associated traits.

Simulation III: GLMM

We considered a familial/trio study design with a single binary trait; because there were multiple subjects in each family, their traits might be correlated. For each of the two parents in

Table 4. Simulation II: empirical type I error rates and power of the approximate (approx) score- and exact score-based tests when multiple traits were correlated with a CS structure with correlation $r = 0.3$

				SPU(γ)							
	Score vector	No. of traits	UminP	$\gamma = 1$	2	3	4	5	∞	aSPU	Score
Type I	Approx	5	0.041	0.042	0.040	0.048	0.038	0.044	0.043	0.042	0.034
		10	0.050	0.051	0.056	0.056	0.057	0.059	0.053	0.052	0.059
		20	0.048	0.057	0.046	0.048	0.049	0.047	0.050	0.045	0.047
		30	0.048	0.048	0.051	0.050	0.049	0.051	0.050	0.053	0.053
		40	0.047	0.052	0.042	0.046	0.048	0.045	0.051	0.042	0.046
	Exact	5	0.043	0.045	0.034	0.042	0.038	0.043	0.043	0.039	0.034
		10	0.052	0.051	0.056	0.057	0.054	0.052	0.050	0.052	0.059
		20	0.046	0.053	0.051	0.050	0.049	0.048	0.049	0.046	0.047
		30	0.047	0.047	0.053	0.056	0.045	0.053	0.050	0.053	0.053
		40	0.051	0.055	0.039	0.047	0.045	0.042	0.049	0.046	0.046
Power	Approx	5	0.140	0.686	0.139	0.288	0.146	0.188	0.144	0.549	0.435
		10	0.324	0.274	0.528	0.341	0.454	0.331	0.324	0.486	0.590
		20	0.394	0.097	0.597	0.433	0.544	0.403	0.392	0.512	0.606
		30	0.378	0.088	0.581	0.395	0.540	0.399	0.375	0.493	0.570
		40	0.364	0.071	0.538	0.416	0.532	0.404	0.368	0.474	0.543
	Exact	5	0.138	0.685	0.133	0.285	0.141	0.182	0.141	0.543	0.434
		10	0.323	0.274	0.536	0.334	0.458	0.315	0.322	0.479	0.588
		20	0.384	0.099	0.594	0.426	0.544	0.401	0.397	0.519	0.606
		30	0.384	0.080	0.571	0.399	0.543	0.398	0.378	0.500	0.569
		40	0.369	0.068	0.534	0.419	0.531	0.410	0.368	0.470	0.542

The first five traits were associated with a causal SNP with effect size $\beta_j \sim U(0.2, 0.3)$; the SNP nearest to the causal SNP was tested.

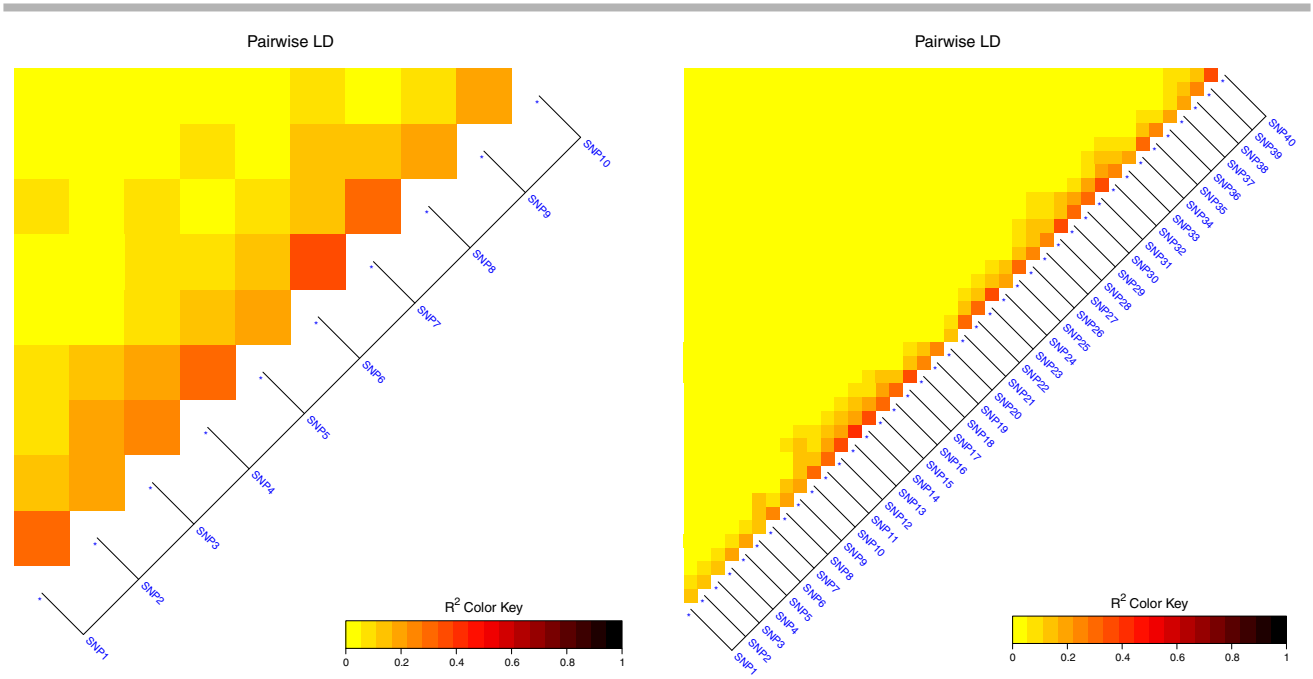

Figure 1. Linkage disequilibrium plots for simulated genotypes with ($n = 200$, $p = 10$) (left panel) and ($n = 400$, $p = 20$) (right panel).

each family $i = 1, \ldots, n$, we generated their haplotypes and thus genotypes as described in the previous section (with $\rho = 0.8$); then their offspring's haplotype and thus genotype data were obtained according to the Mendelian transmission. In this way, we obtained the genotype data with $p + 1$ SNPs for each subjects. The SNP at the center (i.e., at position $p/2 + 1$) was chosen as the causal one with MAF 0.3, while the MAFs of the other SNPs were independently drawn from $U(0.1, 0.4)$. Figure 1 shows linkage disequilibrium plots for the generated SNPs (after the causal SNP was removed) based on one parent from each family.

For subject j in family i, denote x_{ij} as the genotypic score for the causal SNP, and $X_{ij} = (X_{ij1}, \ldots, X_{ijp})^T$ as a vector of the genotypic scores for the p noncausal SNPs. The disease indicator $Y_{ij} = 0$ or 1 was generated from the below GLMM: Assuming λ_0 is the background log odds ratio, ψ_0 is the effect

sizes, the resulting GLMM model can be written as

$$\text{Logit}(E[Y_{ij}|b_i]) = \lambda_0 + x_{ij}\psi_0 + b_i; b_i \sim N(0, \sigma_b^2), \quad (11)$$

where $\lambda_0 = -\log(4)$ was chosen to have a 20% background disease prevalence, while ψ_0 at varying effect sizes was used to investigate the empirical Type I error ($\psi_0 = \log(1)$) and power ($\psi_0 > \log(1)$). We fixed $\sigma_b = 1$, and considered two cases with ($p = 10$, $n = 200$) and ($p = 20$, $n = 400$). For each simulation setup, we generated 1,000 simulated datasets to estimate the empirical Type I error or power.

We fitted both a GLMM (6) and a corresponding marginal model to test H_0: $\psi = 0$. A GLMM was fitted using function `glmer()` from R package `lme4`. Either the Laplace approximation (LA) or adaptive Gaussian quadrature (AGQ) was used to approximate the (marginal) log-likelihood. For AGQ, we specified the number of points per axis nAGQ = 25; in the manual of `lme4`, it was mentioned that "a model with a single, scalar random-effects term could reasonably use up to 25 quadrature points per scalar integral" [Bates et al., 2014]. For a marginal model, both a working independence and a working CS correlation structures were used in GEE. We then applied our method to approximate the score vector for the fitted GLMM and GEE models, respectively, based on which we applied the SPU and aSPU tests; we used $B = 1,000$ to calculate their P-values. As a comparison, we also showed the results from the UminP and Wald tests; recall that the Wald test is equivalent to the approximate score test.

As shown in Table 5, it seems that the Type I error rates were appropriately controlled by all the tests except the Wald test, which gave slightly inflated Type I error rates in GEE as well known in the literature [e.g., Zhang et al., 2014]. It is clear that, for the same fitted model, the aSPU test was much more powerful than the popular UminP and Wald tests. Among the SPU tests, the SPU(1) was nearly as powerful as SPU(2) and SPU(3) for a smaller number of parameters to be tested with $p = 10$, but it was less powerful than the latter two for $p = 20$. This is in agreement with the analysis and motivation of the SPU tests: for $p = 10$, because all the 10 SNPs were correlated with the causal SNP, the SPU(1) test was expected to be powerful, as more generally known for the burden tests; on the other hand, for $p = 20$, because some SNPs were barely correlated with the causal SNP, to minimize the effects on power of the nonassociated SNPs, a larger γ (here $\gamma = 3$) would yield higher power for the SPU(γ) test. Among the three fitted models, the aSPU test based on GEE(CS) was more powerful than that based on the other two models, the latter of which gave similar results; the lower power of GEE(Ind) might be due to the use of the working independence correlation structure, while the GLMM was fitted by an approximate maximum likelihood (using either the LA by default, or ACQ in the function `glme`), leading to loss of efficiency. It is noted that the results from the LA and ACQ approximations were very close; they seemed to be a little conservative with Type I rates much lower than the nominal 0.05.

Because the proposed method of approximating the score vector is asymptotic, to further evaluate its finite-sample performance, we also applied the score-based tests based on the *exact*, not approximate, GEE (generalized) score formulas, as implemented in Zhang et al. [2014]. Comparing the test results in Tables 5 and 6, we can see that the approximate score-based tests had slight losses of power for the smaller sample size $n = 200$, but performed equally well as the exact score-based tests for the larger sample size $n = 400$.

Discussion

We have described an asymptotic approach to approximating the score vector in some complex models, such as a mvLMM or GLMM. The approximate score vector can then be used to construct any score-based tests, including KMR and aSPU tests for multiple traits or familial data. Using both real and simulated data, we have demonstrated such approximate score-based tests can improve power over (and control the Type I error rate better than) the standard Wald test and the UminP test that has been widely used in GWAS. The proposed approximate score vector offers a simple and general way to extend many score-based tests to other complex models, in which the score vector is unavailable from the statistical package being used. Although we have focused on the mvLMM and GLMM, we also considered the LMM, the Cox frailty model [Therneau et al., 2003], and the Cox mixed-effects model [Therneau, 2012], and reached similar conclusions (results not shown to save space); the difference between the two Cox models is that the latter includes a random effect to account for genetic relatedness across *all* subjects, similar to that adopted in the mvLMM. We envision the use of the proposed approximate score vector in other models.

Our proposed general approach is a two-step procedure. In the first step, a full model including a set of parameters to be tested is fitted, then in the second step the score vector for the parameter and its covariance matrix are approximated based on the parameter estimates and their covariance matrix. In this way, a score-based test can be applied without directly calculating the score vector (and its covariance matrix), which may not be easy to derive based on existing software packages, such as for mvLMM and GLMM. Due to the nature of the proposed two-step approach, the validity of the approach depends on both the first step and the asymptotics. For example, if we have a familial dataset with trait-ascertained samples, then it is necessary to appropriately account for the sample ascertainment in step one, e.g., based on some family-based association testing procedures [Moerkerke et al., 2010; Zhang et al., 2012]. Furthermore, because the proposed approximation to the score vector is based on the asymptotics of the parameters to be tested, it has some limitations. First, if the sample size is too small or more generally, if the conditions for the asymptotics do not hold, e.g., in analysis of rare variants [Chen et al., 2013; Jiang et al., 2014], then it may not perform well with inflated false positives and false negatives. Second, in order to obtain a point estimate of the parameters to be tested, a full model has to be fitted, which may not be even computationally feasible if the number of the parameters to be estimated is too large relative to the sample size. Nevertheless,

Table 5. Simulation III: empirical Type I error (for OR = 1) and power (for OR > 1) of the *approximate* score-based tests with correlated binary traits

				SPU(γ)							
Model	Case	OR	UminP	$\gamma = 1$	2	3	4	5	∞	aSPU	Wald
GLMM	1	1	0.018	0.044	0.015	0.016	0.011	0.012	0.014	0.020	0.021
(LA)		1.4	0.078	0.125	0.100	0.105	0.088	0.090	0.074	0.100	0.057
		1.8	0.216	0.323	0.296	0.294	0.266	0.257	0.195	0.300	0.141
		2.2	0.406	0.514	0.494	0.511	0.461	0.452	0.368	0.470	0.280
		2.6	0.580	0.702	0.681	0.680	0.645	0.641	0.568	0.667	0.427
		3	0.736	0.825	0.822	0.826	0.799	0.794	0.713	0.807	0.587
	2	1	0.028	0.030	0.018	0.024	0.017	0.019	0.024	0.023	0.032
		1.4	0.096	0.157	0.116	0.138	0.106	0.108	0.090	0.131	0.067
		1.8	0.344	0.424	0.422	0.467	0.425	0.431	0.353	0.414	0.187
		2.2	0.677	0.668	0.750	0.776	0.750	0.755	0.667	0.748	0.419
		2.6	0.862	0.834	0.897	0.914	0.902	0.904	0.848	0.893	0.666
		3	0.950	0.918	0.964	0.969	0.968	0.972	0.938	0.963	0.825
GLMM	1	1	0.015	0.047	0.014	0.011	0.008	0.010	0.011	0.022	0.018
(AGQ)		1.4	0.073	0.118	0.084	0.096	0.085	0.083	0.070	0.092	0.039
		1.8	0.198	0.310	0.279	0.281	0.248	0.242	0.187	0.281	0.122
		2.2	0.390	0.505	0.472	0.489	0.439	0.431	0.353	0.463	0.255
		2.6	0.565	0.688	0.669	0.664	0.624	0.616	0.536	0.654	0.397
		3	0.724	0.816	0.817	0.817	0.792	0.783	0.695	0.794	0.557
	2	1	0.024	0.029	0.016	0.023	0.016	0.015	0.022	0.024	0.025
		1.4	0.085	0.149	0.110	0.133	0.096	0.103	0.083	0.122	0.056
		1.8	0.329	0.407	0.408	0.450	0.396	0.415	0.332	0.412	0.166
		2.2	0.665	0.667	0.727	0.764	0.730	0.737	0.657	0.744	0.374
		2.6	0.851	0.833	0.892	0.910	0.895	0.901	0.834	0.888	0.637
		3	0.947	0.916	0.962	0.970	0.968	0.970	0.928	0.958	0.810
GEE	1	1	0.030	0.040	0.024	0.023	0.026	0.024	0.026	0.032	0.057
(Ind)		1.4	0.093	0.134	0.108	0.111	0.100	0.099	0.088	0.116	0.106
		1.8	0.215	0.291	0.284	0.288	0.253	0.248	0.202	0.272	0.170
		2.2	0.375	0.478	0.459	0.467	0.428	0.422	0.366	0.445	0.308
		2.6	0.519	0.625	0.625	0.628	0.596	0.588	0.519	0.603	0.438
		3	0.672	0.754	0.755	0.757	0.736	0.729	0.656	0.740	0.558
	2	1	0.036	0.037	0.035	0.033	0.032	0.034	0.042	0.033	0.081
		1.4	0.113	0.146	0.128	0.149	0.134	0.137	0.110	0.146	0.125
		1.8	0.323	0.368	0.368	0.410	0.377	0.394	0.320	0.385	0.229
		2.2	0.589	0.593	0.662	0.709	0.677	0.680	0.579	0.672	0.427
		2.6	0.775	0.761	0.829	0.847	0.826	0.830	0.770	0.825	0.619
		3	0.902	0.865	0.926	0.939	0.932	0.928	0.881	0.925	0.755
GEE	1	1	0.044	0.061	0.041	0.037	0.035	0.035	0.034	0.040	0.065
(CS)		1.4	0.116	0.141	0.133	0.136	0.115	0.117	0.111	0.133	0.104
		1.8	0.271	0.366	0.349	0.358	0.323	0.319	0.261	0.337	0.222
		2.2	0.459	0.576	0.573	0.577	0.541	0.538	0.459	0.547	0.385
		2.6	0.653	0.730	0.734	0.740	0.704	0.698	0.632	0.738	0.526
		3	0.800	0.845	0.862	0.862	0.847	0.840	0.780	0.851	0.670
	2	1	0.039	0.041	0.032	0.035	0.035	0.038	0.041	0.038	0.070
		1.4	0.124	0.183	0.159	0.180	0.151	0.163	0.124	0.167	0.137
		1.8	0.433	0.476	0.500	0.543	0.509	0.508	0.438	0.506	0.309
		2.2	0.750	0.712	0.805	0.832	0.806	0.809	0.755	0.791	0.544
		2.6	0.895	0.881	0.925	0.943	0.933	0.931	0.883	0.925	0.749
		3	0.969	0.934	0.971	0.982	0.978	0.978	0.962	0.976	0.881

The two cases were for ($n = 200$, $p = 10$) and ($n = 400$, $p = 20$). A GLMM was fitted using either the Laplace (LA) or adaptive Gaussian quadrature (AGQ) approximation; the working correlation structure in GEE was assumed to be either independent(Ind) or compound symmetry (CS).

the proposed method offers a simple and practical way to extend many score-based tests to more complex models, for which the score vector is either unavailable from software or has no closed-form solution. In addition to testing for main effects as considered here, it will also be interesting to explore the use of the proposed method to detect gene-gene and gene–environment interactions [Tzeng et al., 2011].

Acknowledgments

This research was supported by NIH grants R01GM113250, R01HL116720, and R01DA033958, and by the Minnesota Supercomputing Institute (MSI). The authors thank Dr. Xiang Zhou for help with the use of mvLMM and the reviewers for constructive comments. Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Alzheimer's Association; Alzheimer's Drug Discovery Foundation; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; GE Healthcare; Innogenetics, N.V.; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Synarc Inc.; and Takeda

Table 6. Simulation III: empirical Type I error (for OR = 1) and power (for OR > 1) of the *GEE score-based tests* with correlated binary traits

				SPU(γ)										
Model	Case	OR	UminP	$\gamma = 1$	2	3	4	5	6	7	8	∞	aSPU	Score
GEE	1	1	0.036	0.047	0.037	0.038	0.031	0.027	0.031	0.033	0.033	0.034	0.038	0.042
(Ind)		1.4	0.090	0.129	0.120	0.117	0.108	0.106	0.099	0.097	0.093	0.090	0.117	0.074
		1.8	0.228	0.295	0.305	0.303	0.277	0.264	0.254	0.248	0.245	0.220	0.284	0.148
		2.2	0.404	0.474	0.480	0.489	0.459	0.450	0.433	0.428	0.425	0.399	0.462	0.280
		2.6	0.567	0.630	0.654	0.658	0.623	0.614	0.603	0.591	0.577	0.555	0.625	0.406
		3	0.698	0.766	0.786	0.779	0.759	0.752	0.740	0.736	0.727	0.694	0.765	0.544
	2	1	0.049	0.046	0.044	0.038	0.039	0.042	0.045	0.044	0.044	0.048	0.051	0.054
		1.4	0.124	0.141	0.152	0.158	0.148	0.151	0.137	0.134	0.125	0.115	0.153	0.092
		1.8	0.355	0.369	0.403	0.429	0.413	0.415	0.392	0.385	0.368	0.344	0.413	0.198
		2.2	0.627	0.595	0.690	0.725	0.696	0.699	0.672	0.669	0.658	0.613	0.693	0.399
		2.6	0.803	0.770	0.852	0.865	0.843	0.841	0.827	0.825	0.820	0.787	0.841	0.597
		3	0.915	0.875	0.941	0.949	0.941	0.942	0.933	0.930	0.924	0.895	0.934	0.745
GEE	1	1	0.040	0.066	0.047	0.048	0.044	0.043	0.042	0.041	0.041	0.043	0.050	0.041
(CS)		1.4	0.119	0.140	0.135	0.134	0.127	0.122	0.120	0.116	0.114	0.114	0.130	0.082
		1.8	0.282	0.361	0.370	0.369	0.342	0.329	0.315	0.308	0.300	0.276	0.349	0.192
		2.2	0.487	0.572	0.582	0.585	0.554	0.546	0.533	0.525	0.510	0.474	0.554	0.357
		2.6	0.651	0.738	0.763	0.747	0.727	0.719	0.700	0.693	0.688	0.647	0.737	0.506
		3	0.823	0.847	0.865	0.870	0.850	0.845	0.837	0.834	0.829	0.797	0.865	0.644
	2	1	0.047	0.044	0.041	0.040	0.045	0.048	0.038	0.041	0.039	0.044	0.049	0.054
		1.4	0.150	0.172	0.162	0.183	0.159	0.168	0.154	0.159	0.149	0.140	0.168	0.098
		1.8	0.446	0.465	0.519	0.543	0.527	0.525	0.502	0.502	0.486	0.465	0.512	0.258
		2.2	0.766	0.712	0.825	0.832	0.823	0.823	0.805	0.807	0.793	0.770	0.816	0.507
		2.6	0.908	0.879	0.928	0.947	0.937	0.942	0.931	0.928	0.922	0.896	0.931	0.739
		3	0.973	0.936	0.975	0.982	0.981	0.979	0.979	0.979	0.978	0.967	0.978	0.866

The two cases were for ($n = 200$, $p = 10$) and ($n = 400$, $p = 20$).

Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

References

Bates D, Maechler M, Bolker B, Walker S. 2014. lme4: linear mixed-effects models using Eigen and S4. R package version 1.1-6. http://CRAN.R-project.org/package=lme4

Boos DD. 1992. On generalized score test. *Am Stat* 46:327–333.

Breslow NE, Clayton DG. 1993. Approximate inference in generalized linear mixed models. *J Am Stat Assoc* 88:9–25.

Chen H, Meigs JB, Dupuis J. 2013. Sequence kernel association test for quantitative traits in family samples. *Genet Epidemiol* 37:196–204.

Diggle P, Heagerty P, Liang K-Y, Zeger S. 2013. *Analysis of Longitudinal Data.* Oxford University Press.

Hartig M, Truran-Sacrey D, Raptentsetsang S, Simonson A, Mezher A, Schuff N, Weiner M. 2012. *UCSF FreeSurfer Overview and QC Ratings.* San Francisco: Alzheimer's Disease Neuroimaging Initiative (ADNI).

Jiang Y, Conneely KN, Epstein MP. 2014. Flexible and robust methods for rare variant testing of quantitative traits in trios and nuclear families. *Genet Epidemiol* 38:542–551.

Kent JT. 1982. Robust properties of likelihood ratio tests. *Biometrika* 69:19–27.

Klei L, Luca D, Devlin B, Roeder K. 2008. Pleiotropy and principal components of heritability combine to increase power for association analysis. *Genet Epidemiol* 32:9–19.

Korte A, Vilhjálmsson BJ, Segura V, Platt A, Long Q, Nordborg M. 2012. A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nat Genet* 44:1066–1071.

Kwee LC, Liu D, Lin X, Ghosh D, Epstein MP. 2008. A powerful and flexible multilocus association test for quantitative traits. *Am J Hum Genet* 82:386–397.

Li B, Leal S M. 2008. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet* 83:311–321.

Liang K, Zeger S. 1986. Longitudinal data analysis using generalized linear models. *Biometrika* 73:13–22.

Lin DY, Tang ZZ. 2011. A general framework for detecting disease associations with rare variants in sequencing studies. *Am J Hum Genet* 89:354–367.

Lin J, Zhu H, Knickmeyer R, Styner M, Gilmore J, Ibrahim JG. 2012. Projection regression models for multivariate imaging phenotype. *Genet Epidemiol* 36:631–641.

Maity A, Sullivan PF, Tzeng J. 2012. Multivariate phenotype association analysis by marker-set kernel machine regression. *Genet Epidemiol* 36:686–695.

Moerkerke B, Vansteelandt S, Lange C. 2010. A doubly robust test for gene-environment interaction in family-based studies of affected offspring. *Biostatistics* 11:213–225.

Pan W. 2009. Asymptotic tests of association with multiple SNPs in linkage disequilibrium. *Genet Epidemiol* 33:497–507.

Pan W. 2011. Relationship between genomic distance-based regression and kernel machine regression for multi-marker association testing. *Genet Epidemiol* 35:211–216.

Pan W, Kim J, Zhang Y, Shen X, Wei P. 2014. A powerful and adaptive association test for rare variants. *Genetics* 197:1081–1095.

Purcell SM, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ and others. 2007. Plink: a toolset for whole-genome association and population-based linkage analysis. *Am J Hum Genet* 81:559–575.

Rotnitzky A, Jewell NP. 1990. Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. *Biometrika* 77:485–497.

Schifano ED, Epstein MP, Bielak LF, Jhun MA, Kardia SL, Peyser PA, Lin X. 2012. SNP set association analysis for familial data. *Genet Epidemiol* 36:797–810.

Shen L, Kim S, Risacher SL, Nho K, Swaminathan S, West JD, Foroud T, Pankratz N, Moore JH, Sloan CD and others. 2010. Whole genome association study of brain-wide imaging phenotypes for identifying quantitative trait loci in MCI and AD: a study of the ADNI cohort. *Neuroimage* 53:1051–1063.

Sun J, Zheng Y, Hsu L. 2013. A unified mixed-effects model for rare-variant association in sequencing studies. *Genet Epidemiol* 37:334–344.

Therneau TM. 2012. Mixed effects Cox models. R-package description. URL: http://cran.r-project.org/web/packages/coxme/vignettes/coxme.pdf.

Therneau TM, Grambsch PM, Pankratz VS. 2003. Penalized survival models and frailty. *J Comput Graph Stat* 12:156–175.

Tzeng JY, Zhang D, Pongpanich M, Smith C, McCarthy MI, Sale MM, Worrall BB, Hsu FC, Thomas DC, Sullivan PF. 2011. Studying gene and gene-environment effects of uncommon and common variants on continuous traits: a marker-set approach using gene-trait similarity regression. *Am J Hum Genet* 89:277–288.

Wang X, Lee S, Zhu X, Redline S, Lin X. 2013. GEE-based SNP set association test for continuous and discrete traits in family-based association studies. *Genet Epidemiol* 37:778–786.

Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, Lin X. 2010. Powerful SNP-set analysis for case-control genome-wide association studies. *Am J Hum Genet* 86:929–942.

Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. 2011. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* 89:82–93.

Xu Z, Shen X, Pan W; Alzheimer's Disease Neuroimaging Initiative. 2014. Longitudinal analysis is more powerful than cross-sectional analysis in detecting genetic association with neuroimaging phenotypes. *PLoS One* 9:e102312.

Yang Q, Wang Y. 2012. Methods for analyzing multivariate phenotypes in genetic association studies. *J Probab Stat* 2012:652569.

Yang Q, Wu H, Guo CY, Fox CS. 2010. Analyze multivariate phenotypes in genetic association studies by combining univariate association tests. *Genet Epidemiol* 34:444–454.

Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB and others. 2005. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet* 38:203–208.

Zhang W, Langefeld CD, Grunwald GK, Fingerlin TE. 2014. Testing gene-environment interactions in family-based association studies using trait-based ascertained samples. *Stat Med* 33:304–318.

Zhang Y, Xu Z, Shen X, et al. 2014. Testing for association with multiple traits in generalized estimation equations, with application to neuroimaging data. *NeuroImage* 96:309–325.

Zhang Z, Ersoz E, Lai CQ, Todhunter RJ, Tiwari HK, Gore MA, Bradbury PJ, Yu J, Arnett DK, Ordovas JM, Buckler ES. 2010. Mixed linear model approach adapted for genome-wide association studies. *Nat Genet* 42:355–360.

Zhou X, Stephens M. 2012. Genome-wide efficient mixed-model analysis for association studies. *Nat Genet* 44:821–824.

Zhou X, Stephens M. 2014. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nat Methods* 11:407–409.