Genetic Epidemiology 35:755–765 (2011)

Multilocus Association Testing With Penalized Regression

Saonli Basu,[1] Wei Pan,[1]* Xiaotong Shen,[2] and William S. Oetting[3]

[1]*Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, Minnesota*
[2]*School of Statistics, University of Minnesota, Minneapolis, Minnesota*
[3]*Department of Experimental and Clinical Pharmacology, Institute of Human Genetics, University of Minnesota, Minneapolis, Minnesota*

In multilocus association analysis, since some markers may not be associated with a trait, it seems attractive to use penalized regression with the capability of automatic variable selection. On the other hand, in spite of a rapidly growing body of literature on penalized regression, most focus on variable selection and outcome prediction, for which penalized methods are generally more effective than their nonpenalized counterparts. However, for statistical inference, i.e. hypothesis testing and interval estimation, it is less clear how penalized methods would perform, or even how to best apply them, largely due to lack of studies on this topic. In our motivating data for a cohort of kidney transplant recipients, it is of primary interest to assess whether a group of genetic variants are associated with a binary clinical outcome, acute rejection at 6 months. In this article, we study some technical issues and alternative implementations of hypothesis testing in Lasso penalized logistic regression, and compare their performance with each other and with several existing global tests, some of which are specifically designed as variance component tests for high-dimensional data. The most interesting, and perhaps surprising, conclusion of this study is that, for low to moderately high-dimensional data, statistical tests based on Lasso penalized regression are not necessarily more powerful than some existing global tests. In addition, in penalized regression, rather than building a test based on a single selected ''best'' model, combining multiple tests, each of which is built on a candidate model, might be more promising. *Genet. Epidemiol.* 35:755–765, 2011.

Key words: Lasso; logistic kernel machine regression; logistic regression; random-effects model; score test; sum of squared score (SSU) test

Contract grant sponsor: NIH; Contract grant numbers: R21DK089351; R01HL105397; R01HL65462; R01GM081535.
*Correspondence to: Wei Pan, Division of Biostatistics, MMC 303, School of Public Health, University of Minnesota, Minneapolis, MN 55455-0392. E-mail: weip@biostat.umn.edu
Received 20 April 2011; Revised 9 June 2011; Accepted 4 July 2011
Published online 15 September 2011 in Wiley Online Library (wileyonlinelibrary.com/journal/gepi).
DOI: 10.1002/gepi.20625

INTRODUCTION

There has been an intensive research effort devoted to developing and applying penalized regression methods, especially for high-dimensional data. The main motivation is that penalized methods, closely related to Bayesian and shrinkage methods, generally lead to better point estimates of parameters, e.g. measured by the mean squared error (MSE), thus improve the *predictive* performance over their nonpenalized counterparts. Furthermore, some penalized methods, e.g. Lasso [Tibshirani, 1996], possess the ability for variable selection, especially for high-dimensional data, facilitating the interpretation of the final selected and often largely simplified model. While the majority of research on penalized methods focus on prediction and variable selection [Kooperberg et al., 2010; Ayers and Cordell, 2010], it is somewhat surprising that little attention has been paid to inference with only a few exceptions in methodology [Meinshausen et al., 2009; Wasserman and Roeder, 2009; Zou and Qiu, 2010] and applications [Malo et al., 2008; Guo and Lin, 2009; Tzeng and Bondell, 2010], in which there is still a lack of comparisons with other approaches. In many applications, e.g. in genetic association analysis of genotypes [e.g. Wu et al., 2010a,b] or gene set analysis of expression data [e.g. Goeman et al., 2004; Liu et al., 2008; Nettleton et al., 2008], one can argue that a primary statistical task is inference: we are not only interested in selecting a subset of important variables, but more so in assessing their statistical significance. In this article, we focus on global testing on a group of variables. In our motivating example, we are interested in validating whether a group of about 20 genetic variants, mostly single nucleotide polymorphisms (SNPs), are associated with a binary outcome, acute rejection (AR), in a study of kidney transplant patients. All these genetic variants were reported to be associated with acute rejection or related clinical outcomes in the previous, though often much smaller, studies. With a much larger sample size here, a univariate analysis identified two SNPs with *P*-values between 0.02 and 0.05 while there were several between 0.05 and 0.1, and none would be significant after adjusting for multiple testing. Since it is well known that typical effect sizes of common genetic variants on complex phenotypes, as AR here, are expected to be small, a global test on the whole group of SNPs might be more powerful than single-SNP analysis or testing each SNP separately [Pan, 2009]. As to be shown, some powerful global tests did indicate marginal significance. On the other hand, due to possible interactions among SNPs (i.e. epistasis), it might be more powerful to consider interactions. Although the number of SNPs, k,

relative to the sample size of $n = 550$ is not large, adding interaction terms into a model would lead to a much larger number of parameters, thus motivating variable selection by penalized regression. Nevertheless, it is unclear whether a variable selection-based approach would be more powerful than some existing global tests developed specifically for high-dimensional data. Furthermore, it is not clear how to most effectively construct tests in the framework of penalized regression. These are key issues to be addressed here. Motivated by the kidney data, we focus on situations with $k < n$, though k/n may be relatively large.

Malo et al. [2008] showed an application of ridge regression to association analysis of multiple SNPs in strong linkage disequilibrium. Since in our motivating example and in many other applications, one does not expect all the predictors (e.g. SNPs) to be significant, it may be reasonable to assume that a penalized method with the capability of variable selection, such as Lasso, is preferred. Since Lasso is perhaps most widely used with fast computational algorithms [Efron et al., 2004; Friedman et al., 2007], treating Lasso as a representative for penalized methods, we restrict our attention to Lasso throughout this article. Although Lasso is most widely used for variable selection based on its nonzero parameter estimates, such a use does not control the Type I error rate, and more importantly, often introduces too many false positives [Devlin et al., 2003]. A typical approach to hypothesis testing with Lasso (or other penalized methods) is to first select the tuning parameter based on cross-validation (CV) or some model selection criteria, then conduct a likelihood ratio test (LRT) and use permutations to estimate its P-value [Guo and Lin, 2009]. There are two potential issues. First, since tuning parameter selection is well known to be unstable [Meinshausen and Buhlmann, 2010], it is possible for such a procedure to end up with a suboptimal tuning parameter, leading to loss of power. Recognizing this limitation, Zou and Qiu [2010] considered using multiple tuning parameters and then combining them. The above two approaches correspond to ''model selection'' and ''model averaging'', respectively, in the well-studied literature of variable selection for prediction. Second, by default the standard LRT or Wald (or score) statistic is used, which however is well known to be nonoptimal for high-dimensional data [Goeman et al., 2006; Chen et al., 2010]. Pan [2009] proposed a modified score (or Wald) test statistic, called sum of squared score (SSU) (or sum of squared betas, SSB), while ignoring the nondiagonal elements of its covariance matrix, which is closely related to Goeman et al.'s [2006] test for high-dimensional data. Treating the parameters as random effects from a distribution, Goeman's test is a score test on the variance component of the random effects, reminiscent of homogeneity tests [Neyman and Scott, 1966; Zelterman and Chen, 1988]. As an approach to gene set analysis, Goeman et al.'s [2004] test is powerful in analyzing high-dimensional microarray data. Even for low-dimensional SNP data, Goeman's test and SSU test have been shown empirically to be often more powerful than the usual score test [Chapman and Whittaker, 2008; Pan, 2009]. Hence, in addition to the standard score statistic, we also consider the use of the SSU statistic. Chen et al. [2010] used a test statistic similar to the SSB, which is asymptotically equivalent to SSU.

We study the performance of various methods with simulated data that mimic the real kidney transplant data. The main conclusion, perhaps surprisingly, is that tests based on model selection or penalized regression do not necessarily outperform some existing global tests proposed for high-dimensional data, which is true across all our simulation setups for low to moderately high dimensional data. Furthermore, for Lasso, tests based on combining multiple SSU statistics corresponding to multiple tuning parameters generally perform better than those based on a single selected tuning parameter.

METHODS

To be concrete, we consider conducting a global test on a set of predictors to assess their effects on a binary outcome. Specifically, suppose that we have n iid observations (Y_i, X_i) for $i = 1, \ldots, n$, where $Y_i = 0$ or 1 is a binary outcome/response variable while $X_i = (X_{i1}, \ldots, X_{ik})'$ is a k-dimensional vector of predictors. We assess the effects of the predictors on the outcome based on logistic regression:

$$\text{Logit Pr}(Y_i = 1) = \beta_0 + \sum_{j=1}^{k} X_{ij}\beta_j. \quad (1)$$

We aim to conduct a global test on the null hypothesis H_0: $\beta = (\beta_1, \ldots, \beta_k)' = 0$ vs.a general alternative H_1: $\beta \neq 0$. Our *primary goal* is to find a test such that it has as high power as possible to reject H_0 when H_0 does not hold, while of course controlling the Type I error rate within a specified significance level α; we use the usual $\alpha = 0.05$ throughout.

TESTS BASED ON (UNPENALIZED) LOGISTIC REGRESSION

The most widely used statistical tests are three asymptotically equivalent ones: the score test, Wald's test, and LRT, all based on maximum likelihood. Since the score test is computationally simplest, we adopt the score test and its modifications throughout. For model (1), under H_0, the score vector and its covariance matrix are

$$U = \sum_{i=1}^{n} (Y_i - \bar{Y}) X_i,$$

$$V = \text{Cov}(U) = \bar{Y}(1 - \bar{Y}) \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})',$$

where $\bar{Y} = \sum_{i=1}^{n} Y_i/n$ and $\bar{X} = \sum_{i=1}^{n} X_i/n$. The (multivariate) score test statistic is

$$T_{\text{Score}} = U' V^{-1} U,$$

which has an asymptotic χ^2 distribution with degrees of freedom (DF) k (or more generally DF = rank(V) and V^{-1} is possibly a generalized inverse). If k is large, the score test may not have high power. Note that the score test is asymptotically equivalent to Hotelling's T^2 test [Clayton et al., 2004].

Another potential problem with the score test is that for high-dimensional data, it will be problematic to estimate its large covariance matrix V. An alternative is to conduct a univariate (i.e. marginal) test on each individual predictor, and then combine the univariate tests by taking the minimum of their P-values. This is the so-called (univariate) min P (Umin$P$) test, most popular in genome-wide association studies). The corresponding Umin$P$ score test

statistic is

$$T_{U\min P} = \max_{j=1,\ldots,k} U_j^2/v_j,$$

where U_j is the jth element of U and v_j is the (j,j)th diagonal element of V. To obtain its P-value, the Bonferroni adjustment or a permutation method is most commonly used, which however is conservative or computationally demanding. An asymptotically ''exact'' method based on the asymptotic normality of the score vector is to calculate the P-value by numerical integration with respect to a multivariate normal density [Conneely and Boehnke, 2007], which we use throughout.

Pan [2009] proposed two tests, called the SSU and sum of weighted squared score (SSUw) tests:

$$T_{\text{SSU}} = U'U, \quad T_{\text{SSUw}} = U'V_d^{-1}U \quad \text{with } V_d = \text{Diag}(V).$$

Under H_0, each of the two test statistics has an asymptotic distribution of a mixture of χ_1^2's, which can be approximated by a scaled and shifted χ^2 distributions [Pan, 2009]. Compared to the score test, the SSU and SSUw tests ignore the nondiagonal elements of V, i.e. correlations among the components of U, which is known to be advantageous for high-dimensional data [Chen and Qin, 2010]. More importantly, as shown in Pan [2009], the SSU test is equivalent to the permutation-based version of Goeman et al.'s [2006] test, which is derived as a score test on a variance component for a random-effects logistic regression model. Specifically, in model (1), if we assume β_j's to be random effects drawn from a distribution with $E(\beta) = 0$ and $\text{Cov}(\beta) = \tau I$, then Goeman's permutation-based score test on H_0: $\tau = 0$ is equivalent to the SSU test. Interestingly, though derived for high-dimensional data, the good performance of SSU and SSUw for low-dimensional SNP data has also been empirically confirmed [Chapman and Whittaker, 2008; Pan, 2009]. Goeman et al. [2006] showed that their test has the highest local power *averaged* over the alternative space of $\beta \neq 0$ satisfying the conditions of $E(\beta) = 0$ and $\text{Cov}(\beta) = \tau I$, implying that the SSU test is also nearly optimal in the above sense. Pan [2009] also showed that SSUw can be regarded as an *estimated* most powerful test. In particular, Pan [2009, Fig. 1] showed that when the components of β are close to each other in absolute values, the SSU (or SSB) tends to be more powerful than the score and $U\min P$ tests.

A class of nonparametric regression techniques called logistic kernel machine regression (LKMR) are closely related to the SSU test. The LKMR model is

$$\text{Logit Pr}(Y_i = 1) = \beta_0 + h(X_{i1}, \ldots, X_{ik}), \tag{2}$$

where $h(.)$ is an unknown nonparametric function, which is determined by a user-specified positive and semi-definite kernel function $K(X_i, X_j)$ [Liu et al., 2008]. $K(X_i, X_j)$ measures the similarity of the predictors for subjects i and j. Some commonly used kernels include linear and quadratic kernels. By the representer theorem of Kimeldorf and Wahba [1971], $h_i = h(X_i) = \sum_{j=1}^n \gamma_j K(X_i, X_j)$ for some $\gamma_1, \ldots, \gamma_n$. To test the null hypothesis of no association between the predictors and the outcome, one can simply test H_0: $h = (h_1, \ldots, h_n)' = 0$. Denote K as the $n \times n$ matrix with the (i,j)th element as $K(X_i, X_j)$ and $\gamma = (\gamma_1, \ldots, \gamma_n)'$, then we have $h = K\gamma$. Treating h as subject-specific random effects with mean 0 and covariance matrix τK, testing H_0: $h = 0$ is equivalent to testing H_0: $\tau = 0$. The corresponding score test on the variance component has a statistic of

$$Q = (Y - \bar{Y}\mathbf{1})'K(Y - \bar{Y}\mathbf{1}),$$

whose asymptotic null distribution is a mixture of χ_1^2's, which can be approximated by a scaled χ^2 distributions [Wu et al., 2010b].

As shown in Pan [2011], LKMR can be formulated as a SSU test on H_0: $b = 0$ in a new logistic regression model:

$$\text{Logit Pr}(Y = 1) = b_0 + Zb, \tag{3}$$

where $K = ZZ'$. A special case is that, if the linear kernel is used, then $Z = X$ and thus the SSU and LKMR test statistics are equal, but there is a minor difference in approximating their (common) asymptotic distribution: the SSU is based on a shifted-scaled χ^2 distribution, whereas LKMR is based on a scaled χ^2 distribution. In general, the difference between the SSU test for model (1) and LKMR is only in the functional forms of the predictors being used; both tests are actually an SSU test applied to two different regression models. Pan [2011] showed that the above SSU and LKMR are closely related to other genomic-distance-based regression methods in genetic association analysis [Wessel and Schork, 2006; Schaid, 2010a,b].

TESTS BASED ON LASSO LOGISTIC REGRESSION

The Lasso estimate $\beta_L(\lambda)$ is based on a penalized log-likelihood

$$\beta_L(\lambda) = \arg \max_{\beta} \{\log L(\beta) - \lambda \|\beta\|_1\},$$

where the penalty is the l_1-norm of the regression coefficients β with a penalization parameter λ. A useful property of Lasso is that, if λ is large enough, some or all components of $\beta_L(\lambda)$ will be exactly zero, automatically realizing variable selection.

The tuning parameter λ is typically chosen based on CV or some model selection criterion, e.g. Akaike's [1973] information criterion (AIC) [Guo and Lin, 2009], by searching in a set of its candidate values, say Λ; throughout this article, we used five fold CV with 21 grid points in Λ, whose values were default from R function glmnet(). Denote the selected tuning parameter as $\hat{\lambda}$. Guo and Lin [2009] proposed the LRT to test for disease-haplotype association based on the Lasso estimate. Following the same line, we can construct a LRT statistic as

$$T_{\text{Sel,LRT}} = 2 \log L(\beta(\hat{\lambda})) - 2 \log L(0).$$

To assess its statistical significance, we use permutations: we permute the original phenotypes to obtain a permuted data set, say bth data set, then apply the Lasso method to obtain a new test statistic $T_{\text{Sel,LRT}}^{(b)}$; the abobe process is repeated for $b = 1, \ldots, B$. Then the permutation P-value is $\sum_{b=1}^B I(T_{\text{Sel,LRT}} > T_{\text{Sel,LRT}}^{(b)})/B$. Depending on whether we fix $\lambda = \hat{\lambda}$ or use CV to re-select λ for each permuted data set, we have two versions of the LRT, called $T_{\text{Sel,LRT,fixed}}$ and $T_{\text{Sel,LRT,tuning}}$, respectively. The latter is expected to better control the Type I error rate, but may be computationally too demanding for large or high-dimensional data.

Alternatively, to mimic the global tests based on the score vector, we do not use $\beta_L(\hat{\lambda})$ directly; rather, we use Lasso for variable selection and then construct the corresponding test statistics. Specifically, suppose that

the component of the score vector corresponding to the nonzero components of $\beta_L(\lambda)$ is $U(\lambda)$, then

$$T_{\mathrm{SSU}}(\lambda) = U(\lambda)'U(\lambda), \quad T_{\mathrm{Sco}}(\lambda) = U(\lambda)'V^{-1}(\lambda)U(\lambda),$$

where $V(\lambda) = (U(\lambda))$, a submatrix of $V = (U)$. We use fivefold CV to minimize the deviance as the model selection criterion to select $\hat{\lambda}$, based on which we obtain the test statistics for the selection approaches:

$$T_{\mathrm{Sel,SSU}} = T_{\mathrm{SSU}}(\hat{\lambda}), \quad T_{\mathrm{Sel,Sco}} = T_{\mathrm{Sco}}(\hat{\lambda}).$$

Now we investigate how to combine test statistics for multiple tuning parameters. The basic idea is to construct a test statistic for each tuning parameter, say $T(\lambda)$, then combine them. There are two technical issues. First, since the distributions of $T(\lambda_i)$ may vary with λ_i, it may be a good idea to standardize $T(\lambda_i)$'s before combining them. For example, for any given λ_i, if we ignore the effects of variable selection, $T_{\mathrm{Sco}}(\lambda_i)$ is approximately distributed as $\chi^2_{d_i}$ with DF $d_i = \dim(U(\lambda_i))$, which in general is a decreasing function of λ_i. If we simply combine $T(\lambda_i)$'s, the combined statistic may be dominated or unduly influenced by a few ones with larger values of d_i. Hence, we standardize $T(\lambda_i)$'s by using their P-values based on their approximate $\chi^2_{d_i}$ distributions. Note that the validity of our procedure does not depend on whether the approximate distribution of $T(\lambda_i)$ holds, since we will use permutations to derive a final P-value. Second, there are various methods of combining P-values, though no single one is expected to be uniformly best. We consider three representative ones based on taking the minimum (Min) P-value, Fisher's [1932] method, and truncated product method (TPM) [Zaykin et al., 2002], respectively. The Min method is similar to that of Zou and Qiu [2010] and is in the same spirit of model selection. Here, we argue that more than one λ_i is informative, motivating the use of Fisher's method. On the other hand, depending on the choice of candidate set Λ, some λ_i's, e.g. those corresponding to shrink all β's to be or close to be 0 for a true non-null model, may not be informative. Thus, it may be a good idea to use multiple, but not necessarily all, λ_i's in combining; TPM is such an approach, though other approaches, e.g. selecting a few most significant components, are also possible. More generally, we may want to assign different weights to different P-values based on the performance of their corresponding models (indexed by their λ_i's), as proven useful in the context of prediction with model averaging [Yang, 2001; Shen and Huang, 2006], though we do not pursue it here.

Specifically, if the SSU (or score) statistic is used, for any $\lambda_i \in \Lambda$, we construct $T_{\mathrm{SSU}}(\lambda_i)$ and derive its P-value, say $P_{\mathrm{SSU}}(\lambda_i)$. Then

$$\begin{aligned} T_{\mathrm{Ave,SSU,Min}} &= \min_{\lambda_i \in \Lambda} P_{\mathrm{SSU}}(\lambda_i), \\ T_{\mathrm{Ave,SSU,Fisher}} &= \prod_{\lambda_i \in \Lambda} P_{\mathrm{SSU}}(\lambda_i), \\ T_{\mathrm{Ave,SSU,TPM}} &= \prod_{\lambda_i \in \Lambda} P_{\mathrm{SSU}}(\lambda_i)^{I(p_{\mathrm{SSU}}(\lambda_i) \le \alpha_0)}, \end{aligned}$$

where we used $\alpha_0 = 0.05$ throughout in TPM [Zaykin et al., 2002]. Similarly, we construct the tests based on the score (or any other test) statistic.

We use permutations to obtain the P-value for each of the above tests. For example, for $T_{\mathrm{Ave,SSU,Min}}$, we permute the outcomes Y to obtain $Y^{(b)}$, then we apply the same procedure to the new data $(X, Y^{(b)})$ to obtain a new test statistic $T^{(b)}_{\mathrm{Ave,SSU,Min}}$. We repeat the above process for $b = 1, 2, \ldots, B$. The P-value for $T_{\mathrm{Ave,SSU,Min}}$ is simply $\sum_{b=1}^{B} I(T_{\mathrm{Ave,SSU,Min}} > T^{(b)}_{\mathrm{Ave,SSU,Min}})/B$. To save computing time, we used a relatively small $B = 100$ in simulations, though we used larger $B = 500$ for real data.

For the selection approaches, we tried both fixing $\lambda = \hat{\lambda}$ in permuted data and choosing λ for each permuted data set, denoted as $T_{\mathrm{Sel,SSU,fixed}}$ and $T_{\mathrm{Sel,SSU,tuning}}$ for the SSU. While the second is computationally more demanding, there may be concerns on possibly inflated Type I error rates for the former. It turned that the former (with our chosen score or SSU statistic) could control the Type I error rates in our simulations, as shown in previous studies [Chen et al., 2010]. Hence, we skip the discussion of the latter.

As a comparison, we also consider a two-stage procedure called screening and cleaning (SC) [Wasserman and Roeder, 2009]. In the SC test, one first splits the data into (almost) equally sized two parts, uses one part to select a final model (with a selected $\hat{\lambda}$ by CV), say $\hat{M}$, then applies the selected model $\hat{M}$ to the second part of the data to obtain a P-value for each covariate included in $\hat{M}$. To be consistent with our aim of global testing, we apply the LRT to $\hat{M}$ with the second part of the data. The SC test is attractive for its nice theoretical properties, low computing cost, and its unique ability to assess statistical significance of each individual parameter; here, we only restrict to global testing. In addition, an improvement based on multiple splitting has been proposed [Meinshausen et al., 2009]. Nevertheless, as demonstrated in statistical inference after model selection [Faraway, 1992] and genetic association analysis [Skol et al., 2006], we suspect that the two-step procedure based on data-splitting as adopted by SC may be too costly with much reduced sample sizes (i.e. only a half of the original sample size) for model selection and significance testing, respectively, leading to reduced power as to be confirmed.

RESULTS

EXAMPLE

Data. The identification genetic variants that predispose individuals to adverse outcomes associated with kidney allograft transplantation, including acute rejection (AR), could help personalized treatment of kidney allograft recipients. A number of genetic variants associated with risk of AR have been identified. The protein products of the identified genes are often involved in the regulation and responsiveness of the immune system. However, there is a lack of reproducibility of identified genetic variants associated with AR. It could be due to typically small sample sizes, often less than 150, and also heterogeneous study populations. It is also expected that, as for other complex traits, the effect sizes of associated genetic variants for AR are small. Hence, a validation study was conducted with a much larger sample size of more than 550 patients transplanted at the University of Minnesota Transplant Center. All the genetic variants, mostly SNPs and a few insertions/deletions (In/Del) (all called SNPs for simplicity in this article), are candidate variants suggested from previous studies to be associated with AR in kidney allografts or with poor outcomes after transplantation [Marder et al., 2003; Pavarino-Bertelli et al.,

2004; Goldfarb-Rumyantzev and Naiman, 2008; Kruger et al., 2008; Nickerson, 2008]. After removing patients with missing genotypes and SNPs either with minor allele frequency (MAF) lower than 1% or for which Hardy-Weinberg equilibrium did not hold, we had $n = 550$ patients and 23 SNPs. Among the 550 patients, 69 patients experienced AR at 6 months. Three SNPs had MAFs between 1.1 and 3.7%, whereas others had between 10.3 and 48.8%. We used an additive genetic model to code each SNP: SNP i is coded as $X_i = 0$, 1 or 2, representing the number of its minor alleles. Our *primary goal* is to test whether these SNPs, either individually or collectively, and if latter, either additively or interactively, are associated with AR at 6 months.

Analysis. First, we consider only main effects. Although the dimension $k = 23$ is much smaller than the sample size $n = 550$, testing on the 23 regression coefficients for the 23 SNPs simultaneously may not be as simple as it appears, partly because of the challenge in estimating a 23×23 covariance matrix for the score test. Hence, it may be appealing to conduct variable selection first. We adopted a commonly used stepwise procedure based on the AIC for variable selection. It selected a model with eight SNPs with the corresponding MLEs and their standard errors shown in Table I. Some individual SNPs as well as the whole group based on a global LRT were statistically significant, but we may not want to trust the given P-values since they did not take account of the effect of model selection (or equivalently, multiple testing). On the other hand, if various global tests were directly applied to the group of the 23 SNPs (without model selection), we obtained the P-values ranging from marginally significant to nonsignificant (Table IV). In particular, the SSU, SSUw and LKMR, and Lasso-based tests of Ave-SSU-TPM and Sel-SSU, all yielded marginally significant P-values around 0.05. A natural question is which tests should be trusted more. As to be shown in our simulation studies, since the SSU, SSUw, LKMR, and Lasso-Ave-SSU-TPM tended to have higher power than other tests, we believe that there was some, albeit not highly significant, statistical evidence to support an overall association between the group of the SNPs and the outcome, acute rejection. It is noted that neither the $U\min P$ nor the multivariate score test gave a significant P-value.

TABLE I. MLEs of the regression coefficients β for the main-effects model ($k = 23$) selected by a stepwise variable selection procedure based on AIC for the kidney data

	Predictor				
Variable	Gene	SNP	$\hat{\beta}_M$	SE	P-value
Intercept	–	–	−1.91	0.34	2.62e−8
CCR5	CCR5	rs333	−0.84	0.41	0.039
IL10819	IL10	rs1800871	0.41	0.21	0.048
END1198	END1	rs5370	−0.43	0.25	0.081
MTHFR677	MTHFR	rs1801133	0.35	0.20	0.075
F7353	Factor VII (F7)	rs6046	−0.55	0.35	0.116
GNB3825	GNB3	rs5443	−0.36	0.20	0.079
AGT235	AGT	rs699	0.28	0.19	0.148
TNFA308	TNF-α	rs1800629	−0.33	0.24	0.163
Global test:					5.93e−5

MLE, maximum likelihood estimates; AIC, Akaike information criterion; TNF, tumor necrosis factor; MTHFR, methylenetetrahydrofolate; GNB3, G protein B3 subunit; AGT, angiotensinogen; IL-10, interleukin-10; END1, endothelin-1; CCR5, chemokine (C-C motif) receptor 5; SE, standard error. AIC = 408.2.

Since the main-effects model did not include any possible interactions among the SNPs (i.e. epistatic effects), it might fail to capture some complex association between the genotypes and trait [e.g. Zhang et al., 2003; Zhang and Liu, 2007; Zheng et al., 2006]. Thus, it is tempting to consider both the main effects and some interaction terms. However, adding all interaction terms will dramatically increase the number of parameters to be tested, leading to possible loss of power. Given the sample size $n = 550$, it is perhaps unwise to consider all two-way interactions (and other high-order interactions). Hence, we made a compromise by considering only 28 two-way interactions among the eight "significant" SNPs selected by the stepwise procedure in the main-effects model. We acknowledge that, albeit popular in practice, this approach may give too optimistic (i.e. more significant) results than the (unknown) truth. By considering both 23 main effects and 28 two-way interaction terms, a stepwise procedure based on the AIC selected a model with eight terms: four main effects and four 2-way interactions (Table II), for which we did not impose a hierarchical principle. Compared to the final main-effects model, the majority of the SNPs selected in the current model also appeared in the former. Note that the most significant SNP IL10592 (rs1800872 in gene IL10) was highly correlated with IL10819 selected in the previous model: their Pearson correlation was 0.8, and both were in the promoter region of the interleukin 10 gene (IL10). Based on the MLEs, multiple terms are statistically significant; however, since the MLEs were based on the selected model and did not take account of any model selection effects, the conclusion based on the MLEs might be misleading. Hence, alternatively, we applied the global tests to the full model with all the 23 main effects and 28 two-way interaction terms (i.e. without model selection); their results are shown in Table IV. The more powerful SSU and LKMR seemed to give more significant P-values, lending some, but not conclusive, evidence to support the association between the SNPs and the outcome.

Finally, to avoid missing some important interactions between the SNPs, we considered a model with a large number of two-way interactions. Excluding three SNPs with MAF less than 5%, we used the remaining 20 SNPs to form their pairwise two-way interactions. Hence, the model contained 23 main effects and 190 two-way interactions with a moderately high dimension of $k = 213$. A stepwise procedure selected a model with seven main-effects and three interactions, for which the MLEs are shown in Table III. All seven main-effects appeared in Model 1 in Table I, and the three interactions were formed by three of the seven main-effects, in which both CCR5 and GNB3825 appeared twice. There were some significant individual SNPs, especially much significant IL10819 as before (or as its highly correlated IL10592). When the global tests were applied to the full model with all $k = 213$ parameters (i.e. no variable selection), the P-values were generally less significant than those in the previous two full models (i.e. the main-effects only model and the model with the main-effects and 28 two-way interactions), possibly due to the cost of the large number of parameters

TABLE II. MLEs of the regression coefficients β for a model with both main-effects and a subset of two-way interactions ($k = 51$) selected by a stepwise variable selection procedure based on AIC for the kidney data

Predictor	$\hat{\beta}_M$	SE	*P*-value
Intercept	−2.22	0.27	4.04e−16
IL10592	0.84	0.28	0.003
END1198	−0.43	0.25	0.080
MTHFR677	0.49	0.21	0.023
F7353	−0.52	0.36	0.147
CCR5 × EBD1198	−1.90	0.95	0.046
IL10592 × END1198	−0.79	0.31	0.011
MTHFR677 × AGT235	−0.36	0.24	0.128
END1198 × TNFA308	0.35	0.16	0.024
Global test:			0.0014

MLE, maximum likelihood estimates; AIC, Akaike information criterion; TNF, tumor necrosis factor; MTHFR, methylenetetrahydrofolate; GNB3, G protein B3 subunit; AGT, angiotensinogen; IL-10, interleukin-10; END1, endothelin-1; CCR5, chemokine (C-C motif) receptor 5; SE, standard error. AIC = 400.3.

TABLE III. MLEs of the regression coefficients β for a model with all main-effects and *all* two-way interactions ($k = 213$) selected by a stepwise variable selection procedure based on AIC for the kidney data

Predictor	$\hat{\beta}_M$	SE	*P*-value
Intercept	−1.94	0.34	1.10e−8
CCR5	−0.60	0.56	0.280
IL10819	0.78	0.29	0.008
END1198	−0.44	0.25	0.070
MTHFR677	0.36	0.19	0.065
F7353	−0.55	0.36	0.121
GNB3825	0.13	0.29	0.649
TNFA308	−0.48	0.26	0.065
CCR5 × GNB3825	−1.76	1.04	0.091
CCR5 × TNFA308	1.07	0.63	0.092
IL10819 × GNB3825	−0.53	0.33	0.113
Global test:			3.24e−4

MLE, maximum likelihood estimates; AIC, Akaike information criterion; TNF, tumor necrosis factor; MTHFR, methylenetetrahydrofolate; GNB3, G protein B3 subunit; AGT, angiotensinogen; IL-10, interleukin-10; END1, endothelin-1; CCR5, chemokine (C-C motif) receptor 5; SE, standard error. AIC = 404.9.

and perhaps a sparse true model. However, in agreement with the previous results, the SSU, SSUw, and LKMR gave marginally significant *P*-values, suggesting possible association between the SNPs and the outcome.

SIMULATIONS WITH THE KIDNEY DATA

Simulation setups. To mimic real data and to be as practical as possible, we used the genotypes (i.e. X_i's) from the kidney data to generate a binary outcome. We considered three scenarios with small to moderately large dimension k, corresponding to the three selected models for real data as shown in Tables I–III. Specifically, in a data-generating logistic regression model (1), the true regression coefficients were chosen to be proportional to the MLE in the corresponding selected model in Tables I–III: $\beta = c\hat{\beta}_M$; that is, $\beta_j = c\hat{\beta}_{j,M}$ if predictor j was chosen in the selected model, and otherwise $\beta_j = 0$. Hence, we had $n = 550$ and $k = 23$, 51, and 213 for the three scenarios, respectively. We used $c = 0$ and $c > 0$ to assess the test size and power properties of various procedures. For each simulation setup, we generated 1,000 simulated data sets to estimate the Type I error rates and power.

TABLE IV. *P*-values of the global tests on H_0: $\beta = 0$ for the main-effects model ($k = 23$), a model with both main-effects and a subset of two-way interactions ($k = 51$), and a model with both main-effects and all two-way interactions ($k = 213$) for the kidney data

Var selection	Test	Main-effects	Some two-way int.	All two-way int.
No	Sco	0.215	0.497	0.362
No	Sco-P	0.224	0.496	0.335
No	*U*min*P*	0.346	0.366	0.708
No	SSU	**0.069**	**0.086**	**0.095**
No	SSU-P	**0.062**	**0.087**	**0.105**
No	SSUw	**0.068**	0.105	0.100
No	LKMR-Linear	**0.064**	**0.081**	**0.088**
No	LKMR-Quad	0.092	0.408	0.204
Lasso	Ave, SSU, Min	0.190	0.336	0.120
Lasso	Ave, SSU, Fisher	0.094	0.172	0.220
Lasso	Ave, SSU, TPM	**0.059**	0.296	0.225
Lasso	Ave, Sco, Min	0.310	0.401	0.805
Lasso	Ave, Sco, Fisher	0.239	0.467	0.720
Lasso	Ave, Sco, TPM	0.284	0.471	0.705
Lasso	Sel, SSU	**0.048**	0.322	0.715
Lasso	Sel, Sco	0.208	0.333	0.715
Lasso	2-stage, SC	0.842	0.934	0.965

Bold characters refer to more significant *P* values.
SSU, sum of squared score; LKMR, logistic kernel machine regression; SSUw, sum of weighted squared score; SC, screening and cleaning.

For larger k, the asymptotic distributions for the score and SSU tests may not apply, so we also calculated their *P*-values using $B = 100$ permutations; the two tests are denoted as Sco-P and SSU-P.

Type I error and power. As shown in Table V, for the low-dimensional case with $k = 23$, there seemed to be not much difference between using the score test and SSU, though SSUw had a slight edge. More importantly, the model selection-based approaches were *not* more powerful than the global SSU and LKMR methods that were not based on model selection. Among the Lasso-based tests, the averaging approaches seemed to be more powerful than the selection methods, though the differences were small; and among the averaging methods, those based on Fisher's method and TPM seemed to be more powerful than the Min method. The Lasso-selection-based LRT with a fixed tuning parameter for permuted data sets seemed to have a slightly inflated Type I error rate; both versions of the Lasso-selection-based LRT performed similarly to other Lasso-based methods.

As shown in Table VI, for the intermediate dimension of $k = 51$, the global SSU test was clearly more powerful than the usual score test: their absolute power difference was as large as 15%. Overall, the SSUw, SSU, and LKMR were most powerful, outperforming model selection-based methods. Among the model selection-based method, the

TABLE V. Empirical Type I error rates ($c = 0$) and power ($c > 0$) based on 1,000 replicates for the main-effects model with $k = 23$ SNPs

Variable selection	Test statistics	C					
		0	0.75	0.9	1	1.1	1.25
No	Sco	0.049	0.489	0.699	0.805	0.884	0.963
No	Sco-P	0.051	0.483	0.682	0.797	0.873	0.952
No	*U*min*P*	0.059	0.302	0.463	0.553	0.657	0.780
No	SSU	0.048	0.519	0.714	0.821	0.865	0.972
No	SSU-P	0.056	0.509	0.713	0.813	0.882	0.962
No	SSUw	0.052	**0.546**	**0.739**	**0.839**	**0.902**	**0.976**
No	LKMR-Linear	0.053	0.532	0.723	0.826	0.888	0.975
No	LKMR-Quad	0.057	0.486	0.689	0.789	0.858	0.948
Lasso	Ave, SSU, Min	0.055	0.444	0.655	0.768	0.844	0.944
Lasso	Ave, SSU, Fisher	0.040	0.469	0.672	0.783	0.861	0.952
Lasso	Ave, SSU, TPM	0.054	0.485	0.686	0.797	0.864	0.955
Lasso	Ave, Sco, Min	0.051	0.445	0.624	0.761	0.843	0.946
Lasso	Ave, Sco, Fisher	0.041	0.440	0.630	0.761	0.853	0.952
Lasso	Ave, Sco, TPM	0.047	0.462	0.653	0.780	0.870	0.957
Lasso	Sel, SSU	0.057	0.451	0.642	0.756	0.849	0.934
Lasso	Sel, Sco	0.052	0.415	0.635	0.752	0.845	0.942
Lasso	Sel, LRT, fixed	0.072	0.487	0.689	0.805	0.874	0.948
Lasso	Sel, LRT, tuning	0.057	0.454	0.645	0.749	0.841	0.922
Lasso	2-stage, SC	0.018	0.121	0.204	0.304	0.386	0.520

Bold characters refer to highest power.
SSU, sum of squared score; LKMR, logistic kernel machine regression; SSUw, sum of weighted squared score; TPM, truncated product method; LRT, likelihood ratio test; SC, screening and cleaning.

TABLE VI. Empirical Type I error rates ($c = 0$) and power ($c > 0$) based on 1,000 replicates for the model with the main effects and some two-way interactions ($k = 51$)

Variable selection	Test statistics	c					
		0	0.75	0.9	1	1.1	1.25
No	Sco	0.044	0.281	0.487	0.617	0.751	0.907
No	Sco-P	0.042	0.282	0.480	0.608	0.741	0.908
No	*U*min*P*	0.081	0.236	0.354	0.424	0.520	0.677
No	SSU	0.057	0.464	0.656	0.772	0.874	0.961
No	SSU-P	0.056	0.461	0.647	0.763	0.870	0.957
No	SSUw	0.049	0.477	**0.680**	**0.808**	**0.898**	**0.973**
No	LKMR-Linear	0.067	**0.483**	0.672	0.787	0.891	0.965
No	LKMR-Quad	0.060	0.304	0.480	0.571	0.687	0.862
Lasso	Ave, SSU, Min	0.051	0.338	0.488	0.606	0.718	0.877
Lasso	Ave, SSU, Fisher	0.046	0.395	0.579	0.695	0.813	0.933
Lasso	Ave, SSU, TPM	0.051	0.425	0.605	0.723	0.833	0.944
Lasso	Ave, Sco, Min	0.049	0.311	0.515	0.613	0.778	0.916
Lasso	Ave, Sco, Fisher	0.036	0.287	0.488	0.610	0.764	0.910
Lasso	Ave, Sco, TPM	0.035	0.289	0.489	0.615	0.759	0.912
Lasso	Sel, SSU	0.054	0.355	0.501	0.580	0.706	0.870
Lasso	Sel, Sco	0.057	0.321	0.500	0.610	0.760	0.905
Lasso	2-stage, SC	0.011	0.147	0.239	0.347	0.444	0.610

Bold characters refer to highest power.
SSU, sum of squared score; LKMR, logistic kernel machine regression; SSUw, sum of weighted squared score; TPM, truncated product method; LRT, likelihood ratio test; SC, screening and cleaning.

TABLE VII. Empirical Type I error rates ($c = 0$) and power ($c > 0$) based on 1,000 replicates for the model with the main effects and all two-way interactions ($k = 213$)

Variable selection	Test statistics	c					
		0	0.75	0.9	1	1.1	1.25
No	Sco	0.021	0.039	0.056	0.079	0.109	0.181
No	Sco-P	0.039	0.073	0.100	0.135	0.186	0.260
No	*U*min*P*	0.075	0.233	0.313	0.396	0.462	0.595
No	SSU	0.048	0.411	0.564	0.699	0.802	0.907
No	SSU-P	0.047	0.399	0.571	0.692	0.792	0.899
No	SSUw	0.047	**0.466**	**0.649**	**0.774**	**0.847**	**0.941**
No	LKMR-Linear	0.053	0.441	0.597	0.734	0.825	0.917
No	LKMR-Quad	0.063	0.199	0.286	0.362	0.444	0.594
Lasso	Ave, SSU, Min	0.053	0.304	0.466	0.582	0.687	0.814
Lasso	Ave, SSU, Fisher	0.033	0.286	0.455	0.556	0.664	0.805
Lasso	Ave, SSU, TPM	0.034	0.310	0.469	0.579	0.687	0.829
Lasso	Ave, Sco, Min	0.047	0.088	0.150	0.219	0.291	0.426
Lasso	Ave, Sco, Fisher	0.030	0.105	0.171	0.239	0.325	0.491
Lasso	Ave, Sco, TPM	0.030	0.105	0.171	0.239	0.324	0.491
Lasso	Sel, SSU	0.046	0.280	0.454	0.562	0.658	0.797
Lasso	Sel, Sco	0.041	0.178	0.267	0.316	0.431	0.576
Lasso	2-stage, SC	0.018	0.095	0.144	0.183	0.255	0.342

Bold characters refer to highest power.
SSU, sum of squared score; LKMR, logistic kernel machine regression; SSUw, sum of weighted squared score; TPM, truncated product method; LRT, likelihood ratio test; SC, screening and cleaning.

averaging methods based on Fisher's or TPM were more powerful than others, especially the selection methods. For the averaging methods, using SSU statistic gained over using the traditional score statistic, but they performed similarly in the selection method; note possibly a large difference between Ave-SSU-TPM and Ave-Score-TPM.

We draw similar conclusions for the moderately high-dimensional case of $k = 213$, though the overall trends manifest more clearly, as shown in Table VII. We note that the score test was low-powered, partly because it was too conservative; even if its permutation distribution, not its asymptotic one, was used to calculate its *P*-value, its power was still much lower than others. For the SSU test, its asymptotic version and permutation-based version gave similar results. The *U*min*P* had an inflated Type I error rate (perhaps due to the poor asymptotic approximation here), and its power was still lower than that of SSU. Again the SSUw test seemed to have a slight edge over SSU. The SSU and LKMR with a linear kernel performed similarly, but LKMR with a quadratic kernel worked less well. For the Lasso-based tests, using the SSU statistic gave much higher power than using the score statistic.

For the third scenario, we also considered mis-specified candidate models. Although the data-generating model included some interaction terms, we only considered candidate models with the 23 main-effects. This represents a common strategy adopted in practice: even if a true model is believed to contain some complex high-order terms, it may be more practical to consider some much simpler candidate models. It is confirmed that indeed such a strategy yielded much higher power for *every* test (Table VIII). In particular, the tests based on the score statistic performed much better for this low-dimensional scenario.

Parameter estimation. We compared the performance of the three methods for parameter estimation: the

TABLE VIII. Empirical Type I error rates ($c = 0$) and power ($c > 0$) based on 1,000 replicates for a mis-specified model with only the main effects ($k = 23$), though the true model contained both main-effects and two-way interactions as shown in Table III

Variable selection	Test statistics	c					
		0	0.75	0.9	1	1.1	1.25
No	Sco	0.046	0.481	0.664	0.766	0.850	0.942
No	Sco-P	0.045	0.480	0.648	0.753	0.837	0.940
No	*U*min*P*	0.062	0.284	0.416	0.502	0.594	0.740
No	SSU	0.044	0.486	0.643	0.767	0.856	0.936
No	SSU-P	0.050	0.465	0.646	0.756	0.849	0.930
No	SSUw	0.045	0.548	0.714	0.820	0.884	0.959
No	LKMR-Linear	0.046	0.499	0.664	0.779	0.859	0.939
No	LKMR-Quad	0.054	0.432	0.578	0.717	0.809	0.906
Lasso	Ave, SSU, Min	0.050	0.424	0.593	0.710	0.797	0.909
Lasso	Ave, SSU, Fisher	0.041	0.423	0.612	0.715	0.812	0.918
Lasso	Ave, SSU, TPM	0.049	0.444	0.617	0.732	0.823	0.923
Lasso	Ave, Sco, Min	0.052	0.440	0.608	0.724	0.809	0.924
Lasso	Ave, Sco, Fisher	0.041	0.429	0.607	0.725	0.818	0.927
Lasso	Ave, Sco, TPM	0.040	0.445	0.622	0.744	0.828	0.934
Lasso	Sel, SSU	0.058	0.418	0.590	0.702	0.791	0.900
Lasso	Sel, Sco	0.052	0.442	0.606	0.717	0.795	0.910
Lasso	2-stage, SC	0.016	0.120	0.198	0.267	0.323	0.474

SSU, sum of squared score; LKMR, logistic kernel machine regression; SSUw, sum of weighted squared score; TPM, truncated product method; LRT, likelihood ratio test; SC, screening and cleaning.

TABLE IX. The MSEs of the parameter estimates based on 1,000 replicates for the model with the main effects and some two-way interactions ($k = 51$) as in Table VI

Method	$c = 0$	0.75	0.9	1	1.1	1.25
Lasso	20.9	293.4	367.5	413.5	459.5	527.7
Ridge	18.2	303.1	406.4	483.1	562.1	689.8
MLE	7,394	99,680	98,700	116,700	77,850	75,100

MLE, maximum likelihood estimate; MSE, mean squared error.

Lasso, the ridge, and the MLE applied to the full model with $k = 51$ regression coefficients for each simulated data set. The application of the penalized methods was the same as before except that we generated a separate tuning data set to select the tuning parameters. We measured the performance of each method based on the MSE of its parameter estimates in the linear predictor scale. That is, for a method, suppose $\beta^{(s)}$ is its estimate of true β from data set s, and X is the design matrix (i.e. genotypes), then its MSE is defined as

$$\text{MSE} = \sum_{s=1}^{1,000} (X\beta^{(s)} - X\beta)'(X\beta^{(s)} - X\beta)/1,000$$

$$= \sum_{s=1}^{1,000} (\beta^{(s)} - \beta)'(\beta^{(s)} - \beta)/1,000.$$

As shown in Table IX, for all $c > 0$, the Lasso performed best with the smallest MSEs, and the MLE did not work well at all. This confirms the advantage of penalized regression, especially variable selection by Lasso, for parameter estimation and thus outcome prediction.

TABLE X. Empirical Type I error rates (OR = 1) and power (OR = 1.25) with various correlations ($\rho > 0$) among $k = 40$ SNPs; $k_0 = 4$ causal SNPs were correlated with the other 36 noncausal ones if $\rho > 0$

Variable selection	Test statistics	$\rho = 0$		$\rho = 0.4$		$\rho = 0.8$	
		OR = 1	1.25	1	1.25	1	1.25
No	Sco	0.045	0.165	0.048	0.197	0.043	0.440
No	Sco-P	0.054	0.174	0.056	0.207	0.059	0.451
No	*U*min*P*	0.060	0.155	0.046	0.256	0.054	0.785
No	SSU	0.048	0.184	0.054	0.342	0.050	0.863
No	SSU-P	0.046	0.203	0.054	0.343	0.057	0.856
No	SSUw	0.043	0.175	0.052	0.339	0.049	0.865
No	LKMR-Linear	0.058	0.202	0.055	0.371	0.053	**0.872**
No	LKMR-Quad	0.054	**0.231**	0.065	**0.389**	0.054	0.861
Lasso	Ave, SSU, Min	0.054	0.211	0.050	0.353	0.057	0.837
Lasso	Ave, SSU, Fisher	0.046	0.184	0.039	0.333	0.044	0.846
Lasso	Ave, SSU, TPM	0.045	0.190	0.048	0.333	0.048	0.849
Lasso	Ave, Sco, Min	0.051	0.185	0.054	0.266	0.065	0.606
Lasso	Ave, Sco, Fisher	0.046	0.170	0.051	0.225	0.055	0.565
Lasso	Ave, Sco, TPM	0.049	0.172	0.049	0.223	0.055	0.540
Lasso	Sel, SSU	0.063	0.204	0.053	0.319	0.049	0.828
Lasso	Sel, Sco	0.056	0.189	0.058	0.283	0.068	0.684

Bold characters refer to highest power.
Each simulation setup was based on 1,000 replicates. SSU, sum of squared score; LKMR, logistic kernel machine regression; SSUw, sum of weighted squared score; TPM, truncated product method; LRT, likelihood ratio test; SC, screening and cleaning; SNP, single nucleotide polymorphism.

SIMULATIONS WITH SIMULATED GENOTYPES

One may wonder whether the main conclusion depends on the correlations among the predictors. In the kidney data, some SNPs were highly correlated while others were not. For example, for the main-effects model, the Pearson correlation coefficients were distributed as the following: the minimum, first quantile (Q1), median (Q2), third quantile (Q3), and maximum were −0.498, −0.033, −0.007, 0.031, and 0.831, respectively; for the full model with $k = 213$ predictors, the Pearson correlation coefficients had the minimum, Q1, Q2, Q3, and maximum as −0.498, −0.028, 0.006, 0.057, and 0.927, respectively. To further assess the effects of the correlations among the predictors, we did more simulation studies with simulated predictors.

We generated genotypes based on a latent multivariate Normal model as in Wang and Elston [2007]. Specifically, we simulated a k-dimensional latent variate, say $Z = (Z_1, \ldots, Z_k)'$, from a multivariate Normal distribution with mean 0 and covariance matrix $AR1(\rho)$(i.e. $\text{Corr}(Z_i, Z_j) = \rho^{|i-j|}$). Then we generate the k minor allele frequencies, say $\text{MAF}_1, \ldots, \text{MAF}_k$, from a uniform distribution $U(0.1,0.4)$. We dichotomized Z into a haplotype, say $H_1 = (h_{11}, \ldots, h_{1k})'$, with $h_{1j} = I(Z_j < MAF_j)$. Similarly, we generated another independent haplotype H_2. Combining the two haplotypes, we obtained an individual's genotype $X = H_1 + H_2$.

We considered two setups. In the first, we took $k = 40$, randomly chose $k_0 = 4$ of the SNPs in X as causal and generated the disease status Y of each individual based on a main-effects logistic regression model (with a common

TABLE XI. Empirical Type I error rates (OR = 1) and power (OR = 1.25) with various correlations (ρ) among $k_0 = 4$ causal SNPs, and among 36 non-causal SNPs, but there was no correlation between any causal and noncausal ones

Variable selection	Test statistics	ρ = 0		ρ = 0.4		ρ = 0.8	
		OR = 1	1.25	1	1.25	1	1.25
No	Sco	0.048	0.168	0.042	0.366	0.047	0.749
No	Sco-P	0.056	0.184	0.052	0.376	0.058	0.755
No	*U*min*P*	0.049	0.148	0.060	0.682	0.060	**0.992**
No	SSU	0.050	0.200	0.053	0.652	0.057	0.867
No	SSU-P	0.058	0.208	0.059	0.657	0.052	0.858
No	SSUw	0.055	0.193	0.053	0.631	0.058	0.876
No	LKMR-Linear	0.059	0.217	0.067	0.672	0.060	0.882
No	LKMR-Quad	0.058	**0.228**	0.065	0.598	0.061	0.610
Lasso	Ave, SSU, Min	0.063	0.208	0.066	**0.700**	0.062	0.986
Lasso	Ave, SSU, Fisher	0.050	0.197	0.053	0.676	0.040	0.955
Lasso	Ave, SSU, TPM	0.053	0.194	0.058	0.663	0.046	0.925
Lasso	Ave, Sco, Min	0.061	0.188	0.050	0.545	0.060	0.934
Lasso	Ave, Sco, Fisher	0.048	0.172	0.044	0.481	0.051	0.913
Lasso	Ave, Sco, TPM	0.050	0.175	0.047	0.440	0.056	0.880
Lasso	Sel, SSU	0.062	0.192	0.053	0.692	0.055	0.984
Lasso	Sel, Sco	0.071	0.181	0.048	0.574	0.057	0.964

Bold characters refer to highest power.
SSU, sum of squared score; LKMR, logistic kernel machine regression; SSUw, sum of weighted squared score; TPM, truncated product method; LRT, likelihood ratio test; SC, screening and cleaning; SNP, single nucleotide polymorphism. Each simulation setup was based on 1,000 replicates.

odds ratio, OR, for each causal SNP). In the second, we generated two independent genotype blocks, say X_0 and X_1, with $k_0 = 4$ and $k_1 = 36$ SNPs, respectively; we then used X_0 as causal SNPs to generate Y as in the first case. We supplied the combined genotypes $X = (X_0, X_1)$ to each data set. In both cases, we had $k = 40$ SNPs, but in the first, the causal and noncausal ones were correlated for $\rho > 0$, whereas in the second they were independent. We followed the typical case-control design: in each data set, we had $n = 200$ cases and $n = 200$ controls.

The simulation results are shown in Tables X and XI. In the first case (Table X), LKMR-Quad and LKMR-Linear were winners, closely followed by SSU, SSUw, and the Lasso-Ave-SSU-Min; the other Lasso methods with the SSU statistic also performed well. The score test and the Lasso methods with the score statistic were all low powered for $\rho > 0$. In the second case, since the true model was sparse with the causal SNPs independent of the noncausal ones, two minP methods, *U*min*P* and Lasso-Ave-SSU-Min were the winners. In both cases, it was confirmed that Lasso-based methods did not outperform some global tests.

DISCUSSION

The most interesting, and perhaps surprising, conclusion of this study is that, for our experimental data with small to moderately high dimensions, the tests based on Lasso for variable selection did not perform better than some global tests, i.e. the SSU, SSUw, and LKMR. This is not the first negative report on penalized regression for genetic association analysis; see Croiseau and Cordell [2009] for a case study and Martinez et al. [2010] for disappointing performance of penalized regression in a different context. Note that in our simulations, the data-generating models were indeed sparse, favoring variable selection by Lasso while the random-effects assumption utilized by the SSU, SSUw, and LKMR was violated; if the true models contained many more nonzero and small coefficients, the SSU, SSUw, and LKMR methods would be expected to perform even better. Possible reasons are the following. First, model selection is difficult. For our examples, there was no or only weak marginal effect of any single predictor, rendering low accuracy of model selection, thus degrading the performance of any test based on model selection. Second, although intuitively it is beneficial to eliminate noninformative variables to reduce the number of parameters to be tested (i.e. reduced DF), e.g. by variable selection, in addition to possibly low selection accuracy, there is always some cost associated with model selection: any test statistic after variable selection is expected to have a null distribution with heavier tails than that without model selection, leading to possible loss of power [Han and Pan, 2010]. In other words, there is always a trade-off between a gain with eliminating noninformative variables (i.e. reduced DF) and a loss due to model searching as measured by inflated generalized DF [Shen and Ye, 2002]. Hence, at the end, the gain may not outweigh the loss. Third, the winning global tests were all developed for high-dimensional data based on testing some variance component in a random-effects model; hence, they are robust to large numbers of parameters to be tested. In fact, there is a close connection between penalized methods and random-effects models: first, a random-effects model can be regarded as a Bayesian model, whose posterior distribution can be interpreted as a penalized likelihood; second, the marginal quasi-likelihood of a generalized linear mixed model can be approximated as a penalized quasi-likelihood based on Laplace's method [Breslow and Clayton, 1993]. Hence, in this sense the global tests such as SSU and LKMR can be also regarded as penalized methods.

One of our main motivations for this study was to combine the strengths of variable selection and powerful test statistics for high-dimensional data. We have proposed and studied such approaches, e.g. "Ave-SSU-Fisher" and "Ave-SSU-TPM." Although the proposed methods performed better than the standard methods based on selecting a single penalization parameter and/or the usual score statistic, they did not outperform the global tests of SSU, SSUw, and LKMR. Of course, we do not claim that it is impossible for model selection-based methods to outperform the global tests, but further studies are needed. The most important message of this report is that, although penalized regression (via variable selection and parameter shrinkage) can often improve parameter estimation and outcome prediction over its nonpenalized counter-parts, it is not clear whether, if yes how, penalized regression can also improve power in hypothesis testing. Even only within the framework of penalized regression, in addition to the choice of the test statistic, there is another critical issue of choosing between averaging over multiple penalization parameters and selecting a single "best" penalization parameter. Our numerical study here seemed to indicate better performance of the averaging approaches. Nevertheless, there may not be a single

uniform winner, in analogous to model averaging vs. model selection studied in the context of variable selection and prediction [Yang, 2003; Shen and Huang, 2006]. More work is warranted.

ACKNOWLEDGMENTS

We thank a reviewer for many helpful and constructive comments.

REFERENCES

Akaike H. 1973. Information theory and the maximum likelihood principle. In: Petrov V, Csáki F, editors. International Symposium on Information Theory. Budapest: Akademiai Kiádo. p 267–281.

Ayers KA, Cordell HJ. 2010. SNP selection in genome-wide and candidate gene studies via penalized logistic regression. Genet Epidemiol 34:879–891.

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

Chapman JM, Whittaker J. 2008. Analysis of multiple SNPs in a candidate gene or region. Genet Epidemiol 32:560–566.

Chen SX, Qin Y-L. 2010. A two-sample test for high-dimensional data with applications to gene-set testing. Ann Statist 38:808–835.

Chen LS, Hutter CM, Potter JD, Liu Y, Prentice RL, Peters U, Hsu L. 2010. Insights into colon cancer etiology via a regularized approach to gene set analysis of GWAS data. Am J Hum Genet 86:860–871.

Clayton D, Chapman J, Cooper J. 2004. Use of unphased multilocus genotype data in indirect association studies. Genet Epidemiol 27: 415–428.

Conneely KN, Boehnke M. 2007. So many correlated tests, so little time! Rapid adjustment of p values for multiple correlated tests. Am J Hum Genet 81:1158–1168.

Croiseau P, Cordell HJ. 2009. Analysis of North American Rheumatoid Arthritis Consortium data using a penalized logistic regression approach. BMC Proc 3:S61.

Devlin B, Roeder K, Wasserman L. 2003. Analysis of multilocus models of association. Genet Epidemiol 25:36–47.

Efron B, Hastie T, Johnstone I, Tibshirani R. 2004. Least angle regression (with discussion). Ann Statist 32:407–499.

Faraway JJ. 1992. On the cost of data analysis. J Comp Graph Stat 1: 213–229.

Fisher RA. 1932. Statistical Methods for Research Workers, 4th edition. London: Oliver & Boyd.

Friedman J, Hastie T, Hoefling H, Tibshirani R. 2007. Pathwise coordinate optimization. Ann Appl Stat 1:302–332.

Goeman JJ, van de Geer S, de Kort F, van Houwelingen HC. 2004. A global test for groups of genes: testing association with a clinical outcome. Bioinformatics 20:93–99.

Goeman JJ, van de Geer S, van Houwelingen HC. 2006. Testing against a high dimensional alternative. J R Stat Soc B 68:477–493.

Goldfarb-Rumyantzev AS, Naiman N. 2008. Genetic prediction of renal transplant outcome. Curr Opin Nephrol Hypertens 17: 573–579.

Guo W, Lin S. 2009. Generalized linear modeling with regularization for detecting common disease rare haplotype association. Genet Epidemiol 33:308–316.

Han F, Pan W. 2010. A data-adaptive sum test for disease association with multiple common or rare variants. Hum Hered 70:42–54.

Kimeldorf GS, Wahba G. 1971. Some results on Tchebycheffian spline function. J Math Anal Appl 33:82–95.

Kooperberg C, LeBlanc ML, Obenchain V. 2010. Risk prediction using genome-wide association studies. Genet Epidemiol 34: 643–652.

Kruger B, Schroppel B, Murphy BT. 2008. Genetic polymorphisms and the fate of the transplanted organ. Transplant Rev 22:131–140.

Liu D, Ghosh D, Lin X. 2008. Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. BMC Bioinformatics 9:292.

Malo N, Libiger O, Schork NJ. 2008. Accommodating linkage disequilibrium in genetic-association analyses via ridge regression. Am J Hum Genet 82:375–385.

Marder B, Schroppel B, Murphy B. 2003. Genetic variability and transplantation. Curr Opin Urol 13:81–89.

Martinez JG, Carroll RJ, Muller S, Sampson JN, Chatterjee N. 2010. A note on the effect on power of score tests via dimension reduction by penalized regression under the null. Int J Biostat 6:12.

Meinshausen N, Buhlmann P. 2010. Stability selection. J R Stat SocB 72: 417–473.

Meinshausen N, Meier L, Buhlmann P. 2009. *P*-values for high-dimensional regression. J Am Stat Assoc 104:1671–1681.

Nettleton D, Recknor J, Reecy JM. 2008. Identification of differentially expressed gene categories in microarray studies using nonparametric multivariate analysis. Bioinformatics 24:192–201.

Neyman J, Scott E. 1966. On the use of c-alpha optimal tests of composite hypothesis. Bull Int Stat Inst 41:477–497.

Nickerson P. 2008. The impact of immune gene polymorphisms in kidney and liver transplantation. Clin Lab Med 28:455–468.

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.

Pavarino-Bertelli EC, Sanches de Alvarenga MP, Goloni-Bertollo EM, Baptista MA, Haddad R, Hoerh NF, Eberlin MN, Abbud-Filho M. 2004. Hyperhomocysteinemia and MTHFR C677T and A1298C polymorphisms are associated with chronic allograft nephropathy in renal transplant recipients. Transplant Proc 36:2979–2981.

Schaid DJ. 2010a. Genomic similarity and kernel methods I: advancements by building on mathematical and statistical foundations. Hum Hered 70:109–131.

Schaid DJ. 2010b. Genomic similarity and kernel methods I: methods for genomic information. Hum Hered 70:132–140.

Shen X, Huang H. 2006. Optimal model assessment, selection and combination. J Am Stat Assoc 101:554–568.

Shen X, Ye J. 2002. Adaptive model selection. J Am Stat Assoc 97: 210–221.

Skol AD, Scott LJ, Abecasis GR, Boehnke M. 2006. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. Nat Genet 38:209–213.

Tibshirani R. 1996. Regression shrinkage and selection via the LASSO. JRSS-B 58:267–288.

Tzeng JY, Bondell H. 2010. A comprehensive approach to haplotype specific analysis via penalized likelihood. Eur J Hum Genet 18: 95–103.

Wang T, Elston RC. 2007. Improved power by use of a weighted score test for linkage disequilibrium mapping. Am J Hum Genet 80: 353–360.

Wasserman L, Roeder K. 2009. High-dimensional variable selection. Ann Stat 37:2178–2201.

Wessel J, Schork NJ. 2006. Generalized genomic distance-based regression methodology for multilocus association analysis. Am J Hum Genet 79:792–806.

Wu J, Devlin B, Ringquist S, Trucco M, Roeder K. 2010a. Screen and clean: a tool for identifying interactions in genome-wide association studies. Genet Epidemiol 34:275–285.

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

Yang Y. 2001. Adaptive regression by mixing. J Am Stat Assoc 9: 574–588.

Yang Y. 2003. Regression with multiple candidate models: selecting or mixing? Stat Sin 13:783–809.

Zaykin DV, Zhivotovsky LA, Westfall PH, Weir BS. 2002. Truncated product method for combing *P*-values. Genet Epidemiol 22: 170–185.

Zelterman D, Chen C-F. 1988. Homogeneity tests against central-mixture alternative. J Am Stat Assoc 83:179–182.

Zhang Y, Liu JS. 2007. Bayesian inference of epistatic interactions in case-control studies. Nat Genet 39:1167–1173.

Zhang J, Liang F, Dassen WR, Veldman BA, Doevendans PA, De Gunst M. 2003. Search for haplotype interactions that influence susceptibility to type 1 diabetes, through use of unphased genotype data. Am J Hum Genet 73:1385–1401.

Zheng T, Wang H, Lo SH. 2006. Backward genotype-trait Association (BGTA) - based dissection of complex traits in case-control design. Hum Hered 62:196–212.

Zou C, Qiu P. 2010. Multivariate statistical process control using LASSO. J Am Stat Assoc 104:1586–1596.