Genome-wide association analysis by lasso penalized logistic regression
Tong Tong Wu1, Yi Fang Chen2, Trevor Hastie2,3, Eric Sobel4 and Kenneth Lange4,5,*
1Department of Epidemiology and Biostatistics, University of Maryland, College Park, MD 20742, 2Department of Statistics, 3Department of Biostatistics, Stanford University, Stanford, CA 94305, 4Department of Human Genetics and 5Department of Biomathematics, University of California, Los Angeles, CA 90095

Received on September 29, 2008; revised on December 11, 2008; accepted on January 18, 2009
Advance Access publication January 28, 2009
Associate Editor: Alex Bateman

ABSTRACT
Motivation: In ordinary regression, imposition of a lasso penalty makes continuous model selection straightforward. Lasso penalized regression is particularly advantageous when the number of predictors far exceeds the number of observations.
Method: The present article evaluates the performance of lasso penalized logistic regression in case–control disease gene mapping with a large number of SNPs (single nucleotide polymorphisms) predictors. The strength of the lasso penalty can be tuned to select a predetermined number of the most relevant SNPs and other predictors. For a given value of the tuning constant, the penalized likelihood is quickly maximized by cyclic coordinate ascent. Once the most potent marginal predictors are identified, their two-way and higher order interactions can also be examined by lasso penalized logistic regression.
Results: This strategy is tested on both simulated and real data. Our findings on coeliac disease replicate the previous SNP results and shed light on possible interactions among the SNPs.
Availability: The software discussed is available in Mendel 9.0 at the UCLA Human Genetics web site.

1 INTRODUCTION
The recent successes in association mapping of disease genes have been propelled by logistic regression using cases and controls. In most ways this represents a step down from the computational complexities of linkage analysis performed on large pedigrees. The most novel feature of these genome-wide association studies is their sheer scale. Hundreds of thousands of SNPs (single nucleotide polymorphisms) are now being typed on samples involving thousands of individuals. This avalanche of data creates new problems in data storage, manipulation and analysis. Size does matter. For instance, with hundreds of thousands of predictors, the standard methods of multivariate regression break down. These methods involve matrix inversion or the solution of linear equations for a very large number of predictors \( p \). Since these operations scale as \( p^3 \), it is hardly surprising that geneticists have opted for univariate linear regression SNP by SNP. This simplification goes against the grain of most statisticians, who are trained to consider predictors in concert. In this article, we explore an intermediate strategy that permits fast computation while preserving the spirit of multivariate regression.

The lasso penalty is an effective device for continuous model selection, especially in problems where the number of predictors \( p \) far exceeds the number of observations \( n \) (Chen et al., 1998; Claerbout and Muir, 1973; Santos and Symes, 1986; Taylor et al., 1979; Tibshirani, 1996). Several authors have explored lasso penalized ordinary regression (Daubechies et al., 2004; Friedman et al., 2007; Fu, 1998; Wu and Lange, 2008) in both the \( \ell_1 \) and \( \ell_2 \) settings. Let \( y_i \) be the response for case \( i \), \( x_{ij} \) the \( j \)-th predictor for case \( i \), \( \beta_j \) the regression coefficient corresponding to the \( j \)-th predictor and \( \mu \) the intercept. For notational convenience also let \( \theta = (\mu, \beta_1, \ldots, \beta_p)^T \) and \( x_i = (x_{i1}, \ldots, x_{ip})^T \). In ordinary linear regression, the objective function is \( f(\theta) = \sum_{i=1}^{n} (y_i - \mu - x_i^T \beta)^2 \). In \( \ell_1 \) regression one replaces squares by absolute values. Lasso penalized regression is implemented by minimizing the modified objective function

\[
g(\theta) = f(\theta) + \lambda \sum_{j=1}^{p} |\beta_j|, \tag{1}\]

Note that the intercept \( \mu \) is ignored in the lasso penalty \( \lambda \sum_{j=1}^{p} |\beta_j| \). The tuning constant \( \lambda \) controls the strength of the penalty, which shrinks each \( \beta_j \) toward the origin and enforces sparse solutions. A ridge penalty \( \lambda \sum_{j=1}^{p} \beta_j^2 \) also shrinks parameter estimates, but it is not as effective in actually forcing many estimates to vanish. This defect of the ridge penalty reflects the fact the \( |\beta| \) is much larger than \( \beta^2 \) for small \( b \).

Many diseases are believed to stem from the interaction of risk factors. This further complication can also be handled by lasso penalization if we proceed in two stages. In the first stage, we select the important marginal predictors; in the second stage, we look for interactions among the supported predictors. In both stages, we adjust the penalty constant to give a fixed number of supported predictors. In most genetic studies, researchers have a general idea of how many true predictors to expect. Our software encourages experimentation and asks the user to decide

*To whom correspondence should be addressed.
on the right balance between model completeness and quick computation.

This article, like most papers, has its antecedents. In particular, Shi et al. (2006, 2007, 2008); Uh et al. (2007) and Park and Hastie (2008) make substantial progress in adapting the lasso to logistic regression and to the discovery of interactions. Malo et al. (2008) apply ridge regression to distinguish causative from non-causative SNPs in a small region. Schwender and Ickstadt (2008) and Kooperberg and Ruczinski (2005) identify interactions using logic regression. These and other relevant papers are reviewed by Liang and Kelemen (2008). We focus on a coordinate descent algorithm because it appears to be the fastest available. Competing algorithms for lasso penalized logistic regression include non-negative quadratic programming (Sha et al., 2007), quadratic approximations (Lee et al., 2006) and interior point methods (Koh et al., 2007). Friedman et al. (2008) compared coordinate descent with several competing algorithms and concluded that it performs the best.

The specific contributions made in this article include (i) the consistent use of the lasso penalty for both marginal and interaction predictors, (ii) selection of the tuning constant to give a fixed number of predictors, (iii) application of cyclic coordinate ascent in maximizing the lasso penalized loglikelihood, (iv) rigorous pre-selection of a working set of predictors and (v) application of false discovery rates for global significance. Our overall strategy pre-selection of a working set of predictors and (v) application of variance 1.

2.2 Selection of the tuning constant

For a given value of the tuning constant \( \lambda \), maximizing the penalized loglikelihood singles out a certain number of predictors with non-zero regression coefficients. Let \( r(\lambda) \) denote the number of predictors selected. If we reduce \( \lambda \) and relax the penalty, then more predictors can enter the model. Although minor exceptions occasionally occur, \( r(\lambda) \) is basically a decreasing function of \( \lambda \) with jumps of size 1. Hence, once a predictor enters the model, it usually remains in the model as \( \lambda \) decreases. Although a predictor’s order of entry tends to be correlated with its marginal significance, violations of this rule of thumb occur with correlated predictors. For every integer \( s \leq p \), we assume that there is an interval \( I_s \) on which \( r(\lambda) = s \). One can quickly find a point in \( I_s \) by a combination of bracketing and bisection. In bracketing, we start with a guess \( \lambda \). If \( r(\lambda) = s \), we are done. If \( r(\lambda) < s \) and \( a \in (0, 1) \), then there is a positive integer \( j \) such that \( r(\lambda a^j) \geq s \). If \( r(\lambda) > s \) and \( b > 1 \), then there is a positive integer \( k \) such that \( r(b^k \lambda) \leq s \). In practice, \( b = 1/2 \) and \( b = 2 \) and take the smallest integer \( k \) or \( \lambda \) yielding the second bracketing point. Once we have a bracketing interval \([\lambda_1, \lambda_2]\), we employ bisection. This involves testing the midpoint \( \lambda_{m} = \lambda_1 + \lambda_2 \). There are three possibilities: if \( r(\lambda_m) = s \), we are done; if \( r(\lambda_m) < s \), we replace \( \lambda_2 \) by \( \lambda_m \); and if \( r(\lambda_m) > s \), we replace \( \lambda_1 \) by \( \lambda_m \). In either of the latter two cases, we bisect again and continue. As soon as we hit a point in \( I_s \), we halt the process.

The primary danger in bracketing is visiting a \( \lambda \) with \( r(\lambda) \) very large. To limit the damage from a poor choice of \( \lambda \), we abort optimization of the objective function whenever the search process encounters too many non-zero predictors. Since predictors can enter and leave the model repeatedly prior to convergence, this check is delayed for several iterations, say 10. As a further safeguard, we set the maximum number of non-zero predictors allowed well above the desired number of predictors \( s \). In practice we use \( s + 10 \).

In simpler settings, cross-validation is used to find the best value of \( \lambda \). Recall that in \( k \)-fold cross-validation, one divides the data into \( k \) equal batches (subsamples) and estimates parameters \( k \) times, leaving one batch out per time. The testing error for each omitted batch is computed using the estimates derived from the remaining batches, and the cross-validation curve \( c(\lambda) \) is computed by averaging testing error across the \( k \) batches. The curve \( c(\lambda) \) can be quite ragged, and many values of \( \lambda \) must be tried to find its minimum. To avoid this time-consuming process, we let the desired number of predictors drive statistical analysis. In actual gene mapping studies, geneticists would be thrilled to map even 5 or 10 genes. In our coeliac disease example, it is necessary to consider a larger number of predictors to uncover the full biological truth.

2 METHODS

2.1 Lasso penalized logistic regression

In case–control studies, the dichotomous response variable \( y_i \) is typically coded as 1 for cases and 0 for controls. By analogy to ordinary linear regression, in linear logistic regression we write the probability

\[ p_i = \Pr(y_i = 1) \text{ of case } i \text{ given the predictor vector } x_i \text{ as} \]

\[ p_i = \frac{e^{\beta_0 + \beta' x_i}}{1 + e^{\beta_0 + \beta' x_i}}. \]  

(2)

The parameter vector \( \theta = (\mu, \beta_1, \ldots, \beta_p)' \) is usually estimated by maximizing the loglikelihood

\[ L(\theta) = \sum_{i=1}^{n} \left[ y_i \log p_i + (1 - y_i) \log (1 - p_i) \right]. \]  

(3)

To encourage sparse solutions, we subtract a lasso penalty from the loglikelihood as just suggested. For the purposes of this article, we consider only additive models where the range of the predictors \( x_{ij} \) is restricted to the three values \(-1, 0 \) and 1, corresponding to the three SNPs genotypes 1/1, 1/2 and 2/2, respectively. A dominant model can be achieved by collapsing the genotypes 1/1 and 1/2, and a recessive model can be achieved by collapsing genotypes 1/2 and 2/2. In both models the assigned quantitative values are \(-1 \) and 1. In our experience, the set of markers entering the model is relatively insensitive to the genetic model assumptions. We recommend standardizing all non-SNP quantitative predictors to have mean 0 and variance 1.
2.3 Assessing significance

When SNPs are tested one by one, it is easy to assign a $P$-value to a SNP by conducting a likelihood ratio test. If we ignore non-genetic predictors such as age, sex and diet, then the only relevant parameters are the intercept $\mu$ and the slope $\beta$ of the SNP. The null hypothesis $\beta=0$ can be tested by maximizing the loglikelihood under the null and alternative hypotheses and forming twice the difference in maximum loglikelihoods. This statistic is asymptotically distributed as a $\chi^2$-distribution with 1 degree of freedom. Collectively, the $P$-values must be corrected for multiple testing, either by a Bonferroni correction or some version of a FDR correction. The latter choice is more appropriate when we anticipate a fairly large number of true positives. We will say more about FDR corrections later. A more compelling concern is that proceeding SNP by SNP omits the impact of other SNPs. Most statisticians prefer to assess significance in the context of multiple linear regression rather than simple linear regression. They resist this natural impulse in association studies because of the computational barriers and the mismatch between numbers of observations and predictors.

In our multivariate setting, we compare the standard SNP by SNP $P$-values with alternative $P$-values generated by considering the $s$ selected predictors as a whole. Once we have selected the $s$ model predictors, we discard the non-selected predictors and re-estimate parameters for the selected predictors with $\lambda=0$. Since $s$ is small, say 10 to 20 in our numerical studies, re-estimation is now a fully determined problem. We then undertake $s$ further rounds of estimation, omitting each of the selected predictors in turn. These actions put us into position to conduct likelihood ratio tests by leaving one predictor out at a time. It is tempting to assign $P$-values by comparing the resulting likelihood ratio statistics to the percentile points of a $\chi^2$-distribution with 1 degree of freedom. This is invalid because it neglects the complex selection procedure for defining the reduced model in the first place. Nonetheless, these LOO $P$-values are helpful in assessing the correlations between the retained predictors in the reduced model. To avoid confusion, we will refer to the LOO $P$-values as LOO indices. The contrast between the univariate $P$-values and the LOO indices is instructive. Although both of these measures are defective and should not be taken too seriously, they are defective in different ways and together give a better idea of the truth.

2.4 Interaction effects

As mentioned previously, we advocate testing for interactions after identifying main effects. This strategy is prompted by the sobering number of interactions possible. With $p$ predictors, there are $\binom{p}{2}$ k-way interactions, and $2^p$ interactions in all. With hundreds of thousands of SNPs, it is impossible even to examine all two-way interactions. These problems disappear once we focus on a handful of interesting marginal predictors. However, our commitment to a two-stage strategy brings in its wake certain technical problems.

First, there is the combinatorial question of how to generate all subsets of $\{1, \ldots, r\}$ up to a given size. Fortunately, good algorithms for this task already exist. Minor changes to the NEKSB code in Nijenhuis and Wilf (1978) permit one to generate one subset after another, with smaller subsets coming before larger subsets. Thus, when the number of predictors $r$ retained from stage one is too large to generate all subsets, one can easily visit all lower order interactions and bypass higher order interactions. Second, there is the problem of storing the interaction predictors. We finesse this problem by computing interaction products on the fly. Third, there is the question of how to integrate SNP predictors with other predictors such as sex, age and environmental exposures. Since this is largely a programming problem, we omit further discussion of it. Fourth, our interactions do not involve any self-interactions. Inclusion of self-interactions would force us to pass from subsets to multisets. For SNPs the potential gain seems worth less than the bother. Other predictors such as age have a richer range of values, so it may be useful to add predictors such as age squared, age cubed and so forth to the original list of predictors. Finally, there are the problems of model selection and hypothesis testing for the interaction effects. Here, again it seems reasonable to rely on lasso penalized estimation and LOO indices.

2.5 Cyclic coordinate ascent algorithm

In linear logistic regression, maximum likelihood estimates are usually found by the scoring algorithm. This requires the score and observed information

$$\nabla L(\theta) = \sum_{i=1}^{n} [y_i - p_i(\theta)] x_i \tag{4}$$

of the loglikelihood (3). Because scoring coincides with Newton’s method, it is fast and reliable, and most statisticians would agree that it is the method of choice for low-dimensional problems. Its Achilles heel is the need to invert the observed information at each iteration. If we add to this drawback the complication of dealing with the non-differentiable lasso penalty, then it becomes abundantly clear that competing algorithms should be considered in association analysis.

The oldest and simplest alternative, coordinate ascent, updates one parameter one at a time. Coordinate ascent comes in two flavors, cyclic and greedy (Wu and Lange, 2008). In cyclic coordinate ascent, each parameter is updated in turn; in greedy coordinate ascent, the parameter leading to the greatest increase in the objective function is updated. Although greedy coordinate ascent makes faster initial progress in logistic regression, it suffers from excess overhead. For this reason we will confine our attention to cyclic coordinate ascent.

Although the logistic loglikelihood (3) is non-linear, it has the compensating property of concavity. Concavity fortunately carries over to the lasso penalized loglikelihood

$$g(\theta) = L(\theta) - \alpha \sum_{j=1}^{p} |\beta_j|,$$

because the sum of two concave functions is concave. The objective function $g(\theta)$ is non-differentiable, but it does possess a directional derivative along each forward or backward coordinate direction. For instance, if $\theta_j$ is the coordinate direction along which $\beta_j$ varies, then

$$d_{\theta_j} g(\theta) = \lim_{t \downarrow 0} \frac{g(\theta + tu_j) - g(\theta)}{t} = d_{\theta_j} L(\theta) + \left\{ \begin{array}{ll} -\lambda \beta_j & \text{if } \beta_j > 0 \\ -\lambda & \text{if } \beta_j < 0 \\ \lambda \beta_j & \text{if } \beta_j < 0 \\ \lambda & \text{if } \beta_j > 0. \end{array} \right.$$ 

When a function such as $L(\theta)$ is differentiable, its directional derivative along $u_j$ coincides with its ordinary partial derivative, and its directional derivative along $v_j = -u_j$ coincides with the negative of its ordinary partial derivative.

To update a single parameter of the objective function $g(\theta)$, we use one-dimensional scoring. This works well for the intercept parameter $\mu$ because there is no lasso penalty. For a slope parameter $\beta_j$, the lasso penalty intervenes, and particular care must be exercised near the origin. In fact, it simplifies matters to start scoring at the origin. Here, we test the directional derivatives $d_{\theta_j} g(\theta)$ and $d_{\beta_j} g(\theta)$. If both are non-positive, then $g(\theta)$ cannot be increased by moving away from the origin. This claim follows from the concavity of $g(\theta)$. If one of the directional derivatives $d_{\beta_j} g(\theta)$ and $d_{\theta_j} g(\theta)$ is positive and the other is non-positive, then progress can be made along the corresponding arm of $g(\theta)$, and scoring is commenced until convergence is achieved along that arm. Concavity rules out the possibility that both directional derivatives are positive. A simple sketch of a concave function will convince the reader of this assertion.

In practice, we start all parameters at the origin. In overdetermined problems, the vast majority of slopes $\beta_j$ are permanently parked there. Only those with considerable evidence in their favor can overcome the pressure of the lasso pushing them toward the origin. Even those that escape this pressure can be forced back to the origin as other more potent predictors enter the model. It is clearly computationally beneficial to organize parameter updates
by tracking the linear predictor $\mu + x^T \beta$ of each case. These start at 0, and when a single component of $\theta$ is updated, it is trivial to update the linear predictors.

2.6 The score criterion and efficient computations

In Section 2.2, we demonstrated that the lasso penalty can be tuned to select a predetermined number of the most relevant SNPs. Once the value of the tuning constant $\lambda$ is fixed, the penalized likelihood is quickly maximized by cyclic coordinate ascent to give us the desired number of non-zero coefficients. However, since we face a very large number of SNP predictors, it would be much more efficient if we could start our search procedure by focusing on a substantially smaller set of features that are more likely to be associated with the response. We accomplish this by a ‘swindle’ that screens the predictors according to a simple score criterion.

The score equations of the loglikelihood (4) for linear logistic regression define part of the Karush–Kuhn–Tucker (KKT) conditions (Lange, 2004)

$$\sum_{i=1}^{n} (y_i - p_i(\lambda))x_{ij} = \lambda \text{ if } \beta_j \neq 0 \quad (5)$$

$$\sum_{i=1}^{n} (y_i - p_i(\lambda))x_{ij} \leq \lambda \text{ if } \beta_j = 0 \quad (6)$$

for optimality in the penalized model. Here $p_i(\lambda)$ is the fitted probability for observation $i$, fit using the indicated value of $\lambda$. For very large $\lambda$, all the $\beta_j$ are estimated as zero, and the only non-trivial parameter is the intercept $\mu$, which is unpenalized. If $p_i(\lambda)$ is the overall proportion of cases in the data, then the intercept is estimated as $\hat{\mu} = \log[p_i/(1-p_i)]$ for large $\lambda$.

We accordingly define the following initial absolute score:

$$a_j = \sum_{i=1}^{n} (y_i - p_i(\lambda))x_{ij} \quad (7)$$

for each predictor. Note that $a_j$ determines the standard score statistic for testing the null model $\beta_j = 0$ with $\mu$ fixed at $\hat{\mu}$. The first predictor to enter the lasso penalized model as $\lambda$ decreases is the predictor with the largest value of $a_j$.

These considerations suggest a screening device for models with large numbers of SNPs. Because we insist on tuning the lasso penalty to select just a handful of predictors, the final absolute scores are apt to correlate strongly with the precomputed absolute scores. Thus, if we desire $s$ predictors, we take $k$ to be a reasonably large multiple of $s$, say $k=10s$, sort the $a_j$, and extract the $k$ predictors with the largest values of $a_j$. Call this subset $S_1$. We now subject $S_1$ to our estimation procedure and choose a value $\lambda_1$ to give us exactly $k$ predictors. The selected predictors satisfy the KKT conditions (5) and (6). If the predictors omitted from $S_1$ also satisfy the KKT condition (6), then we have found the global minimum for the given value $\lambda_1$ and stop. If one of the omitted predictors fails the KKT condition (6), we replace $k$ by $2k$, say, and repeat the process. Eventually, the KKT conditions are satisfied by all predictors. Since the KKT conditions are sufficient as well as necessary for a global maximum, this process legalizes the swindle. Often the value 10s works. When it does not, usually just a few doublings suffice. For example, if the desired number of predictors is $s=10$, in stage one we fit a model with 100 predictors. When stage two is needed, we fit a model with 200 predictors, and so forth. If there are hundreds of thousands of SNPs, our swindle saves an enormous amount of computing with no loss in rigor.

Of course, the swindle sets $S_1$ may contain highly correlated features with redundant information. This turns out to be the case with the HLA SNPs in our coeliac example. Fortunately, most of the redundant features are discarded by the lasso penalty. Our numerical results, for instance those displayed in Table 1, confirm that the swindle vastly speeds up computation while preserving model selection results.

2.7 Computation of FDR

The score swindle also has implications for the assessment of the FDR for the univariate $P$-values. We will not pursue these delicate connections here because in practice most geneticists demand that all univariate tests be done. Fortunately, it takes just a few minutes of computing time to carry out the univariate logistic regressions encountered in a modern association study. Even substituting likelihood ratio tests for score tests does not change this fact.

In the Simes procedure highlighted by Benjamini and Hochberg (1995) in their analysis of FDR, there are $n$ null hypotheses $H_1, \ldots, H_n$ and $n$ corresponding $P$-values $P_1, \ldots, P_n$. The latter are replaced by their order statistics $P_{(1)}, \ldots, P_{(n)}$. If for a given $\alpha \geq 0$, we choose the largest integer $j$ such that $P_{(i)} \leq (i/n)\alpha$ for all $i \leq j$, then we can reject the hypotheses $H_{(1)}, \ldots, H_{(j)}$ at an FDR of $\alpha$ or better. This procedure is justified in theory when the tests are independent or positively correlated. In the presence of linkage equilibrium, association tests are independent; in the presence of linkage disequilibrium, they are positively correlated. For a more detailed discussion of the multiple testing issues in SNP studies, see Nyholt (2004).

3 ANALYSIS OF SIMULATED DATA

To evaluate the performance of lasso penalized regression in association testing, we focus on underdetermined problems where the number of predictors $p$ far exceeds the number of observations $n$. Our simulation model

$$\log\left(\frac{p_i}{1-p_i}\right) = \mu + \sum_{j=1}^{p} x_{ij}\beta_j + \sum_{k=1}^{p} \sum_{l=1}^{p} x_{ik}x_{jl}\eta_{kl}. \quad (8)$$

involves both marginal effects and two-way interactions. For ease of simulation, we assume that each predictor vector $x_j$ is derived from a realization of a multivariate normal vector $Y_j$ whose marginals are standard normal and whose covariances are

$$\text{Cov}(Y_{ij}, Y_{ik}) = \begin{cases} 1 & j=k \\ \rho & j,k \leq 10, j \neq k \\ 0 & \text{otherwise} \end{cases}$$

Thus, only the first 10 predictors are correlated. To mimic a SNP with equal allele frequencies, we set $x_{ij}$ equal to $-1$, 0 or 1 according to whether $Y_{ij} < -c$, $-c \leq Y_{ij} \leq c$, or $Y_{ij} > c$. The cutoff $-c$ is the first quartile of a standard normal distribution. In every simulation, we set $\mu = 1$, $\beta_j = 1$ for $1 \leq j \leq 5$, and $\beta_j = 0$ for $j > 5$. We also set $\eta_{12} = 0$ except for the special cases $\eta_{12} = 0.5$. These substantial effect sizes allow us to discern signal from noise in fairly small samples.

To ameliorate the shrinkage of the non-zero estimates for a particular $\lambda$, we always re-estimate the selected parameters in the final model, omitting the non-selected parameters and the lasso penalty. This yields better parameter estimates for testing purposes. We compute LOO indices as mentioned earlier and contrast them to univariate $P$-values based on estimating the impact of each predictor without reference to the other predictors.

We analyzed the simulated data in two stages. In stage one, we considered only main effects and selected $s_1$ predictors. In stage two, we discarded the non-selected predictors and sought $s_2$ marginal effects or interactions among the selected predictors. The sensible choice $s_2 \geq s_1$ permits all predictors singled out in stage one to remain in contention as marginal effects in stage two. Because virtually all association studies yield only a handful of predictors that can be replicated, we took $s_1$ and $s_2$ small and considered the specific pairs $(s_1, s_2) = (10, 10), (10, 20), (20, 10), (20, 20)$. Table 1 summarizes our results over 50 random replicates for various choices of the number of predictors $p$, the number of subjects $n$ and the correlation coefficient $\rho$. Table 1 reports the average values of the
Table 1. Simulation results based on 50 random samples

<table>
<thead>
<tr>
<th>$(p,n)$</th>
<th>$\rho$ (s1, s2)</th>
<th>$\lambda_1$</th>
<th>$K_{\text{true,1}}$</th>
<th>$\lambda_2$</th>
<th>$K_{\text{true,2}}$</th>
<th>Time no swindle</th>
<th>Time swindle</th>
</tr>
</thead>
<tbody>
<tr>
<td>(5000, 500)</td>
<td>0.0 (10,10)</td>
<td>29.43 (1.50)</td>
<td>5.00 (0.00)</td>
<td>29.64 (1.90)</td>
<td>5.84 (0.67)</td>
<td>1.36 (0.34)</td>
<td>0.68 (0.11)</td>
</tr>
<tr>
<td>(5000, 500)</td>
<td>0.0 (10,20)</td>
<td>29.43 (1.50)</td>
<td>5.00 (0.00)</td>
<td>10.86 (1.71)</td>
<td>6.98 (0.14)</td>
<td>2.18 (0.39)</td>
<td>1.57 (0.26)</td>
</tr>
<tr>
<td>(5000, 500)</td>
<td>0.0 (20,10)</td>
<td>25.46 (1.06)</td>
<td>5.00 (0.00)</td>
<td>30.06 (1.65)</td>
<td>5.84 (0.67)</td>
<td>2.67 (0.40)</td>
<td>1.10 (0.17)</td>
</tr>
<tr>
<td>(5000, 500)</td>
<td>0.0 (20,20)</td>
<td>25.46 (1.06)</td>
<td>5.00 (0.00)</td>
<td>25.49 (1.25)</td>
<td>6.24 (0.65)</td>
<td>2.75 (0.36)</td>
<td>2.17 (0.39)</td>
</tr>
<tr>
<td>(5000, 500)</td>
<td>0.8 (10,10)</td>
<td>19.51 (1.94)</td>
<td>5.00 (0.00)</td>
<td>17.62 (3.24)</td>
<td>5.04 (0.20)</td>
<td>3.06 (0.52)</td>
<td>1.76 (0.46)</td>
</tr>
<tr>
<td>(5000, 500)</td>
<td>0.8 (10,20)</td>
<td>19.51 (1.94)</td>
<td>5.00 (0.00)</td>
<td>6.16 (1.12)</td>
<td>6.58 (0.57)</td>
<td>5.91 (5.74)</td>
<td>4.61 (5.63)</td>
</tr>
<tr>
<td>(5000, 500)</td>
<td>0.8 (20,10)</td>
<td>16.40 (1.50)</td>
<td>5.00 (0.00)</td>
<td>19.79 (2.01)</td>
<td>5.04 (0.20)</td>
<td>6.40 (2.94)</td>
<td>3.08 (0.94)</td>
</tr>
<tr>
<td>(5000, 500)</td>
<td>0.8 (20,20)</td>
<td>16.40 (1.50)</td>
<td>5.00 (0.00)</td>
<td>16.28 (1.65)</td>
<td>5.12 (0.38)</td>
<td>6.50 (4.32)</td>
<td>5.14 (3.33)</td>
</tr>
<tr>
<td>(50000, 2000)</td>
<td>0.0 (10,20)</td>
<td>67.39 (2.21)</td>
<td>5.00 (0.00)</td>
<td>21.83 (3.18)</td>
<td>7.00 (0.00)</td>
<td>39.17 (11.45)</td>
<td>10.09 (10.81)</td>
</tr>
<tr>
<td>(50000, 2000)</td>
<td>0.8 (10,20)</td>
<td>45.99 (2.12)</td>
<td>5.00 (0.00)</td>
<td>15.09 (2.39)</td>
<td>7.00 (0.00)</td>
<td>102.31 (33.92)</td>
<td>14.59 (10.37)</td>
</tr>
<tr>
<td>(100000, 2000)</td>
<td>0.0 (10,20)</td>
<td>69.77 (2.13)</td>
<td>5.00 (0.00)</td>
<td>23.62 (3.24)</td>
<td>7.00 (0.00)</td>
<td>110.24 (22.59)</td>
<td>8.94 (11.27)</td>
</tr>
<tr>
<td>(100000, 2000)</td>
<td>0.8 (10,20)</td>
<td>47.71 (2.30)</td>
<td>5.00 (0.00)</td>
<td>14.66 (2.54)</td>
<td>7.00 (0.00)</td>
<td>197.20 (53.17)</td>
<td>10.81 (1.69)</td>
</tr>
</tbody>
</table>

Fig. 1. Plots of the stage one penalty constant $\lambda_1$ versus the number of selected predictors, the number of true predictors and FDR. The stage two penalty constant $\lambda_2 = 25$.

The last two columns of Table 1 summarize computing times with and without our computational swindle. Forgoing the swindle inflates all times in Table 1. For $p = 5000$ the differences are not too noticeable, but for $p = 100000$ it takes 10 to 20 times longer to reach the lasso solution without the swindle. The results Table 1 for the choice $(s_1, s_2) = (10, 20)$ appear best. In general, we recommend using a substantially larger $s_2$ than $s_1$. Performance degrades as we pass from uncorrelated to highly correlated predictors. More iterations are needed for convergence, and the fraction of true predictors captured falls. With a large enough sample size, performance is perfect. Table 1 in our submitted Supplementary Materials displays our results for a single representative sample with $p = 50000$, $n = 2000$, $\rho = 0$ and $(s_1, s_2) = (10, 20)$. At stage one, all five true predictors are correctly selected with impressive univariate $P$-values and LOO indices. At stage two, all five main effects and both interaction effects are selected. In both instances, the univariate $P$-values and LOO indices of the true predictors are much smaller than the corresponding values for the false predictors.

It is also instructive to consider what happens in the simulated data with $p = 5000$, $n = 500$ and $\rho = 0$ when the stage one tuning constant $\lambda_1$ varies. Figure 1 plots six things as a function of $\lambda_1$: (i) the number of predictors selected at stage one, (ii) the number of predictors selected at stage two, (iii) the number of true predictors selected at stage one, (iv) the number of true predictors selected at stage two, (v) the FDR at stage one and (vi) the FDR at stage two. In stage two, we set the tuning constant $\lambda_2 = 25$. In counting true predictors, we consider only marginal predictors at stage one and marginal plus interaction predictors at stage two. When we know the true predictors, estimating FDR is trivial, and the Simes procedure can be ignored. Inspection of the six plots shows that all true predictors are recovered for a fairly broad range of $\lambda_1$ values. As $\lambda_1$ decreases, more predictors enter the model, and FDR increases.

4 ANALYSIS OF COELIAC DATA

4.1 Data description

In the British coeliac data of van Heel et al. (2007), $p = 310,637$ SNPs are typed on $n = 2200$ subjects (938 males and 1262 females).
Controls outnumber cases 1422 to 778. Across the sample, an impressive 99.875% of all genotypes are assigned; no individual has more than 10% missing data. We impute missing genotypes at a SNP by the method sketched in Ayers and Lange (2008). Only 32 SNPs show a minor allele frequency below 1%; these are dropped from further analysis.

4.2 Simulation study based on coeliac data

We also tested our method by conducting a simulation study based on the coeliac data. Here in model (8), we took \( \mu = -3 \), \( \beta_1 = 1 \) for gender, \( \text{s3737728 (SNP2)} \), \( \text{rs9651273 (SNP4)} \) and \( \text{rs4970362 (SNP9)} \), and \( \beta_2 = 0 \) for the remaining SNPs. We also set \( \eta_{ij} = 2 \) for the interaction of gender and \( \text{rs3934834 (SNP1)} \) and the interaction of SNP4 and SNP9; all other \( \eta_{ij} \) we set to 0. Notice that SNP1 has no marginal effect even though it interacts with gender in determining the response. The lower right-hand block of the correlation matrix

\[
\begin{bmatrix}
gender & SNP1 & SNP2 & SNP4 & SNP9 \\
0.0009 & -0.0106 & -0.0178 & -0.0307 & 0.0009 \\
-0.0106 & 1.0000 & -0.2249 & 0.0991 & -0.0207 \\
-0.0178 & -0.2249 & 1.0000 & 0.5289 & 0.3892 \\
-0.0307 & 0.0991 & 0.5289 & 1.0000 & 0.2894 \\
0.0009 & -0.0207 & 0.3892 & 0.2894 & 1.0000 \\
\end{bmatrix}
\]

indicates fairly strong linkage disequilibrium among the three marginally important SNPs. Table 2 summarizes Fisher’s exact test indicating fairly strong linkage disequilibrium among the three SNPs.

<table>
<thead>
<tr>
<th>Locus name</th>
<th>Est. P-value</th>
<th>Range</th>
<th>Expected homozygotes</th>
<th>Observed homozygotes</th>
</tr>
</thead>
<tbody>
<tr>
<td>s3934834</td>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>s3737728</td>
<td>0.7101</td>
<td>±</td>
<td>0.0009743</td>
<td>1303.39</td>
</tr>
<tr>
<td>s9651273</td>
<td>0.6445</td>
<td>±</td>
<td>0.0095733</td>
<td>1304.25</td>
</tr>
<tr>
<td>s4970362</td>
<td>0.7412</td>
<td>±</td>
<td>0.0087595</td>
<td>1189.35</td>
</tr>
</tbody>
</table>

4.3 Results of real data analysis

Revisiting earlier results with antigenic markers, van Heel et al. (2007) find overwhelming evidence for association in the human leukocyte antigen (HLA) region of chromosome 6. SNP rs2187668 in the first intron of HLA-DQA1 has the strongest association, followed by SNPs rs9357152 and rs9275141 within or adjacent to HLA-DQB1. van Heel et al. also identify a more weakly associated region on chromosome 4 centered on SNPs rs13119723 and rs6822844 in the KIAA1109-TENR-IL2-IL21 linkage disequilibrium block. Their results are summarized in Table 3 of our submitted Supplementary Materials.

<table>
<thead>
<tr>
<th>((s_1,s_2))</th>
<th>(\lambda_1)</th>
<th>(K_{true,1})</th>
<th>(\lambda_2)</th>
<th>(K_{true,2})</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Additive model</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(10,10)</td>
<td>48.78</td>
<td>4.00</td>
<td>52.19</td>
<td>4.46</td>
<td>45.33</td>
</tr>
<tr>
<td></td>
<td>(1.57)</td>
<td>(0.00)</td>
<td>(4.11)</td>
<td>(0.50)</td>
<td>(13.48)</td>
</tr>
<tr>
<td>(10,20)</td>
<td>48.78</td>
<td>4.00</td>
<td>18.24</td>
<td>5.70</td>
<td>66.36</td>
</tr>
<tr>
<td></td>
<td>(1.57)</td>
<td>(0.00)</td>
<td>(4.05)</td>
<td>(0.61)</td>
<td>(12.82)</td>
</tr>
<tr>
<td>(20,10)</td>
<td>45.10</td>
<td>4.00</td>
<td>53.93</td>
<td>4.44</td>
<td>74.22</td>
</tr>
<tr>
<td></td>
<td>(1.24)</td>
<td>(0.00)</td>
<td>(4.00)</td>
<td>(0.50)</td>
<td>(29.64)</td>
</tr>
<tr>
<td>(20,20)</td>
<td>45.10</td>
<td>4.00</td>
<td>45.11</td>
<td>4.54</td>
<td>137.16</td>
</tr>
<tr>
<td></td>
<td>(1.24)</td>
<td>(0.00)</td>
<td>(1.63)</td>
<td>(0.50)</td>
<td>(51.70)</td>
</tr>
<tr>
<td>Dominant model</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(10,20)</td>
<td>85.18</td>
<td>3.96</td>
<td>26.68</td>
<td>5.70</td>
<td>182.98</td>
</tr>
<tr>
<td></td>
<td>(4.57)</td>
<td>(0.20)</td>
<td>(5.69)</td>
<td>(0.54)</td>
<td>(20.62)</td>
</tr>
<tr>
<td>(10,20)</td>
<td>62.53</td>
<td>3.00</td>
<td>20.76</td>
<td>3.14</td>
<td>83.76</td>
</tr>
<tr>
<td></td>
<td>(3.93)</td>
<td>(0.00)</td>
<td>(8.32)</td>
<td>(0.35)</td>
<td>(66.12)</td>
</tr>
</tbody>
</table>

Table 3. Results for 50 random replicates using the coeliac genotypes

Downloaded from https://academic.oup.com/bioinformatics/article/25/6/714/251234 by guest on 22 October 2023
Table 4. The non-HLA predictors found under model 3 with 50 mandated predictors

<table>
<thead>
<tr>
<th>SNP</th>
<th>Chr Position in BP</th>
<th>Univariate P-value</th>
<th>LOO index</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>gender</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs1888176</td>
<td>1</td>
<td>0.00156</td>
<td>0.000234</td>
<td>0.36746</td>
</tr>
<tr>
<td>rs13397583</td>
<td>2</td>
<td>1.60268e-05</td>
<td>0.001518</td>
<td>0.36265</td>
</tr>
<tr>
<td>rs6735141</td>
<td>2</td>
<td>0.060684</td>
<td>0.28418e-05</td>
<td>0.44583</td>
</tr>
<tr>
<td>rs2836985</td>
<td>21</td>
<td>0.000149</td>
<td>0.014373</td>
<td>0.24478</td>
</tr>
<tr>
<td>rs2879414</td>
<td>18</td>
<td>0.000272</td>
<td>2.78969e-05</td>
<td>0.41328</td>
</tr>
<tr>
<td>rs7320671</td>
<td>13</td>
<td>5.33772e-05</td>
<td>0.849038</td>
<td>0.30108</td>
</tr>
<tr>
<td>rs10501723</td>
<td>11</td>
<td>8.56749e-05</td>
<td>0.004005</td>
<td>0.28470</td>
</tr>
<tr>
<td>rs7320671</td>
<td>13</td>
<td>0.001174</td>
<td>1.64655e-05</td>
<td>0.43494</td>
</tr>
<tr>
<td>rs13357969</td>
<td>15</td>
<td>0.00047</td>
<td>0.007989</td>
<td>0.28059</td>
</tr>
<tr>
<td>rs1064981</td>
<td>10</td>
<td>3.41871e-06</td>
<td>0.089192</td>
<td>0.35675</td>
</tr>
<tr>
<td>rs1539234</td>
<td>10</td>
<td>5.33772e-06</td>
<td>0.849038</td>
<td>0.30108</td>
</tr>
<tr>
<td>rs10505604</td>
<td>8</td>
<td>1.43831e-05</td>
<td>3.47134e-05</td>
<td>0.47647</td>
</tr>
<tr>
<td>rs1499447</td>
<td>8</td>
<td>0.000804</td>
<td>0.00174</td>
<td>0.37033</td>
</tr>
<tr>
<td>rs1901633</td>
<td>10</td>
<td>0.000047</td>
<td>0.007989</td>
<td>0.28059</td>
</tr>
<tr>
<td>rs5762743</td>
<td>18</td>
<td>0.000272</td>
<td>2.78969e-05</td>
<td>0.41487</td>
</tr>
<tr>
<td>rs7320671</td>
<td>13</td>
<td>1.64655e-05</td>
<td>0.43494</td>
<td>0.28829</td>
</tr>
<tr>
<td>rs6735141</td>
<td>2</td>
<td>0.000272</td>
<td>2.78969e-05</td>
<td>0.41487</td>
</tr>
<tr>
<td>rs12642902</td>
<td>4</td>
<td>0.000286</td>
<td>5.86295e-05</td>
<td>0.38997</td>
</tr>
<tr>
<td>rs736191</td>
<td>8</td>
<td>9.9264380</td>
<td>0.000049</td>
<td>0.35143</td>
</tr>
<tr>
<td>rs10505604</td>
<td>8</td>
<td>1.39406770</td>
<td>3.47134e-05</td>
<td>0.47647</td>
</tr>
<tr>
<td>rs1499447</td>
<td>8</td>
<td>3.909901e-11</td>
<td>6.99199e-05</td>
<td>0.42571</td>
</tr>
<tr>
<td>rs1901633</td>
<td>10</td>
<td>4.800561</td>
<td>0.007989</td>
<td>0.28059</td>
</tr>
<tr>
<td>rs10503018</td>
<td>18</td>
<td>0.000247</td>
<td>0.000925</td>
<td>0.36819</td>
</tr>
<tr>
<td>rs23689685</td>
<td>21</td>
<td>0.000149</td>
<td>0.014373</td>
<td>0.24478</td>
</tr>
<tr>
<td>rs6517581</td>
<td>21</td>
<td>0.001141</td>
<td>0.000959</td>
<td>0.33649</td>
</tr>
<tr>
<td>rs5764419</td>
<td>22</td>
<td>0.000411</td>
<td>0.003744</td>
<td>0.28829</td>
</tr>
<tr>
<td>rs2283693</td>
<td>X</td>
<td>0.000248</td>
<td>0.007031</td>
<td>0.28610</td>
</tr>
<tr>
<td>rs5394725</td>
<td>X</td>
<td>0.000757</td>
<td>0.010604</td>
<td>0.22593</td>
</tr>
<tr>
<td>rs4335267</td>
<td>X</td>
<td>0.000084</td>
<td>0.012583</td>
<td>0.21747</td>
</tr>
</tbody>
</table>

Dutch and Irish samples. Our failure to identify the same SNPs in the KIAA1109-TENR-IL2-IL21 block is hardly a disaster; the region and ultimately the underlying gene are more important than the individual SNPs. It is noteworthy that among the 1000 most significant SNPs listed by van Heel et al. (2007), 979 are in the HLA region. Since SNPs in the HLA region on chromosome 6 are highly correlated with coeliac disease, model 4 with 10 mandated predictors removes the HLA SNPs, with the aim of finding associated SNPs outside the HLA region. Table 4 in our submitted Supplementary Materials now picks up SNPs on chromosomes 9, 11, 14 and 18 that do not appear in Table 4. Removing all chromosome 6 SNPs rather than just HLA SNPs leads to virtually the same results as displayed in Supplementary Table 4.

To test for interactions, we take the s1 = 50 predictors selected in model 3 and examine all marginal and two-way effects. The total number of predictors is 50 + (50 \choose 2) = 1275, and we keep s2 = 50 predictors in the model. Most of the 50 selected predictors have LOO indices close to one. Table 5 lists the marginal and interaction predictors with LOO indices less than 0.01. Several of these interactions are interesting. Given the predominance of female patients, the interaction between gender and one of the HLA SNPs is credible. The interactions between two HLA SNPs and SNPs on chromosomes 2, 3 and 2 are more surprising. It is particularly noteworthy that the univariate P-values for these three SNPs as marginal effects (Table 4) are far less impressive than their univariate P-values as interaction effects (Table 5).

5 DISCUSSION

Our analysis of simulated data demonstrates that lasso penalized regression is easily capable of identifying pertinent predictors in grossly underdetermined problems. Computational speed is

Table 5. Strongest predictors under model 5 with all main effects and two-ways interactions included

<table>
<thead>
<tr>
<th>SNP</th>
<th>Chr Position in BP</th>
<th>Univariate P-value</th>
<th>LOO index</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>gender</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs1888176</td>
<td>1</td>
<td>0.00156</td>
<td>0.000234</td>
<td>0.36746</td>
</tr>
<tr>
<td>rs1836577</td>
<td>3</td>
<td>0.000955</td>
<td>8.69904e-05</td>
<td>0.38923</td>
</tr>
<tr>
<td>rs1559810</td>
<td>3</td>
<td>6.24178e-05</td>
<td>9.71542e-05</td>
<td>0.40263</td>
</tr>
<tr>
<td>rs991316</td>
<td>4</td>
<td>0.000026</td>
<td>0.000514</td>
<td>0.34159</td>
</tr>
<tr>
<td>rs13357969</td>
<td>5</td>
<td>8.89698e-05</td>
<td>4.56935e-05</td>
<td>0.43473</td>
</tr>
<tr>
<td>rs2187668</td>
<td>6</td>
<td>9.48302e-19</td>
<td>1.65234e-09</td>
<td>1.30307</td>
</tr>
<tr>
<td>rs916786</td>
<td>7</td>
<td>0.000804</td>
<td>3.13293e-05</td>
<td>0.41684</td>
</tr>
<tr>
<td>rs736191</td>
<td>8</td>
<td>0.000026</td>
<td>0.001485</td>
<td>0.33010</td>
</tr>
<tr>
<td>rs10503604</td>
<td>8</td>
<td>4.38312e-06</td>
<td>0.001069</td>
<td>0.38067</td>
</tr>
<tr>
<td>rs10501723</td>
<td>11</td>
<td>8.56749e-05</td>
<td>0.005633</td>
<td>0.28523</td>
</tr>
<tr>
<td>rs2387414</td>
<td>18</td>
<td>4.796289</td>
<td>0.000272</td>
<td>0.38128</td>
</tr>
<tr>
<td>rs10503018</td>
<td>18</td>
<td>0.000024</td>
<td>0.003744</td>
<td>0.28829</td>
</tr>
<tr>
<td>rs5934725</td>
<td>23</td>
<td>6.24178e-05</td>
<td>3.13293e-05</td>
<td>0.41684</td>
</tr>
<tr>
<td>gender, rs2586997</td>
<td>6</td>
<td>0.000084</td>
<td>0.012583</td>
<td>0.21747</td>
</tr>
</tbody>
</table>

Here, we take s1 = 50 and s2 = 50 and list an effect when its LOO index falls below 0.01.
impressive. If predictors are uncorrelated, then interaction effects
can be found readily as well. As one might expect, correlations
among important predictors degrade computational speed and the
recognition of interactions. For very large datasets involving more
than, say, $10^9$ total SNP genotypes, data compression is mandatory.
Repeated decompression of chunks of the data then slows
computation. Our computational swindle circumvents this problem
because all of the working predictors easily fit within memory.

The coeliac dataset of van Heel et al. (2007) is challenging
for two reasons. First, the overwhelming HLA signal masks
the weaker signals coming from other chromosome regions.
Second, the HLA SNPs are in strong linkage disequilibrium and
hence highly correlated. Linkage disequilibrium manifests itself
as increased LOO indices and significant two-way interactions.
Despite these handicaps, lasso penalized regression identifies several
promising non-HLA regions and interaction effects. Our results for
chromosome 4 differ slightly from those of van Heel et al. (2007)
because we impute missing genotypes differently. Ayers and Lange (2008)
introduce a new penalized method of haplotype frequency
estimation that enforces parsimony and achieves both speed and
accuracy. When phase can be deduced from relatives, this extra
information can be included in estimation. Finally, it is noteworthy
that van Heel et al. have validated the chromosome 4 association on
two further datasets.

One can quibble with our method of picking candidate predictors
for interaction modeling. An obvious alternative would be to look
for two-way interactions between the top $s$ predictors and all other
predictors. This tactic requires little change in our numerical
methods.

Readers may want to compare our approach with the approach of
Shi et al. (2006, 2007, 2008). One major difference is our application of
cyclic coordinate ascent. A second major difference is that we
always select a fixed number of predictors. These choices allow us
to quickly process a very large numbers of SNPs or interactions among
SNPs. The path following algorithm of Park and Hastie (2008) has
the advantage of revealing the exact sequence in which predictors
enter the model. Path following is more computationally demanding
than simply finding the best $r$ predictors, but note that their software
[glmpath in R, Park and Hastie (2007)] can quickly post-process the
best $r$ predictors discovered.

We have featured univariate $P$-values and LOO indices in this
article, but neither measure is ideal. Although FDR analysis is valuable, no one has said the last word on multiple testing (Balding,
2006; Kimmel and Shamir, 2006). For instance, some form of
generalized cross-validation may ultimately prove useful. As a
matter of principle, most geneticists would not accept a single study
as definitive. All important findings are subject to replication. This
attitude, whether justified or not, puts the onus on finding the most
important SNPs rather than on declaring their global significance.
Our approach to data analysis is motivated by this consideration. The
software discussed here will be made available in the next release of
Mendel.

Funding: USPHS (GM53275, MH59490 to K.L. in part).

Conflict of Interest: none declared.

REFERENCES

Bioinformatics, 24, 1596–1602.
20, 33–61.
Claribou,J.F. and Mouri,F. (1973) Robust modeling with erratic data. Geophysics, 38,
826–844.
Friedman,J. et al. (2008) Regularized Paths for Generalized Linear Models Via
Coordinate Descent. Department of Statistics, Stanford University.
Koh,K. et al. (2007) An interior-point method for large-scale l1-regularized logistic
21th National Conference on Artificial Intelligence (AAAI-06). Available at http://
citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.64.1993
correlated high dimensional SNP data in genomic study for complex diseases. Stat.
Surv., 2, 43–60.
Nyholt,D.R. (2004) A simple correction for multiple testing for SNPs in linkage
and Cox Proportional Hazards Model. R package. Available at http://biostats.bcm.tmc.edu/
R/Rpackages/lasso.
BMC Proc., 1, Suppl. 1, S60.
Shi,W. et al. (2008) LASSO-Patternsearch Algorithm with Applications to
Ophthalmology and Genomic Data. Technical Report 1131, University of Wisconsin,
Madison.
Taylor,H.L. et al. (1979) Deconvolution with the $\ell_1$ norm. Geophysics, 44, 39–52.
Uh,H.-W. et al. (2007) Model selection based on logistic regression in a highly correlated
candidate gene region. BMC Proc., 1 (Suppl. 1), S60.
van Heel,D. et al. (2007) A genome-wide association study for celiac disease identifies