Genetic Epidemiology 34: 892–911 (2010)

# Evaluating Haplotype Effects in Case-Control Studies via Penalized-Likelihood Approaches: Prospective or Retrospective Analysis?

**Megan L. Koehler,[1] Howard D. Bondell,[1] and Jung-Ying Tzeng[1,2]***

[1]*Department of Statistics, North Carolina State University, Raleigh, North Carolina*
[2]*Bioinformatics Research Center, North Carolina State University, Raleigh, North Carolina*

Penalized likelihood methods have become increasingly popular in recent years for evaluating haplotype-phenotype association in case-control studies. Although a retrospective likelihood is dictated by the sampling scheme, these penalized methods are typically built on prospective likelihoods due to their modeling simplicity and computational feasibility. It has been well documented that for unpenalized methods, prospective analyses of case-control data can be valid but less efficient than their retrospective counterparts when testing for association, and result in substantial bias when estimating the haplotype effects. For penalized methods, which combine effect estimation and testing in one step, the impact of using a prospective likelihood is not clear. In this work, we examine the consequences of ignoring the sampling scheme for haplotype-based penalized likelihood methods. Our results suggest that the impact of prospective analyses depends on (1) the underlying genetic mode and (2) the genetic model adopted in the analysis. When the correct genetic model is used, the difference between the two analyses is negligible for additive and slight for dominant haplotype effects. For recessive haplotype effects, the more appropriate retrospective likelihood clearly outperforms the prospective likelihood. If an additive model is incorrectly used, as the true underlying genetic mode is unknown a priori, both retrospective and prospective penalized methods suffer from a sizeable power loss and increase in bias. The impact of using the incorrect genetic model is much bigger on retrospective analyses than prospective analyses, and results in comparable performances for both methods. An application of these methods to the Genetic Analysis Workshop 15 rheumatoid arthritis data is provided. *Genet. Epidemiol.* 34:892–911, 2010. 

**Key words: haplotype-based association analysis; variable selection; regularized regression; prospective likelihood; retrospective likelihood**

Contract grant sponsor: NIH; Contract grant numbers: 5RO1-HL049609-14; 1R01-AG021917-01A1; T32GM081057; R01 MH084022-01; R01 MH084022-01; 1P01-CA142538-0; R01 MH084022-01; 1P01-CA142538-01; Contract grant sponsors: University of Minnesota; Minnesota Supercomputing Institute; GAW; Contract grant numbers: R01-GM031575; AR44422; Contract grant sponsor: NSF; Contract grant number: DMS-0705968.
*Correspondence to: Jung-Ying Tzeng, Department of Statistics, Campus Box 7566, North Carolina State University, Raleigh, NC 27695. E-mail: jytzeng@stat.ncsu.edu
Received 29 April 2010; Revised 3 September 2010; Accepted 15 September 2010
Published online 18 November 2010 in Wiley Online Library (www.wileyonlinelibrary.com).
DOI: 10.1002/gepi.20545

## INTRODUCTION

Haplotype-based association analysis evaluates the joint effects of closely linked genetic markers on a trait of interest. When compared to its single-marker counterparts, this multi-marker approach can be more powerful to detect associations when the causal variants are not genotyped [de Bakker et al., 2005; Zaitlen et al., 2007], have low frequency [de Bakker et al., 2005; Schaid, 2004], or exhibit cis-acting effects [Clark, 2004; Schaid, 2004]. A standard approach for performing haplotype-based analysis is to regress the trait value on the haplotypes and to test the significance of the regression parameters [Balding, 2006]. In recent years, applying penalized likelihood methods to identify important haplotypic factors has become increasingly popular in the literature. For example, Li et al. [2007] use the least absolute shrinkage and selection operator (LASSO) [Tibshirani, 1996] to perform selection among numerous possible haplotypes resulting from different haplotype window lengths. Guo and Lin [2009] use LASSO regression to evaluate the effects of rare haplotypes and high-dimension haplotype-environment interactions. Tzeng et al. [2010] use adaptive LASSO regression [Zou, 2006] to study high-dimensional gene-treatment interactions in a haplotype-based pharmacogenetic analysis. These methods introduce a penalty on the regression coefficients and shrink the coefficient estimates of non-important covariates toward zero. The motivation behind using penalized methods in haplotype-based analysis is that while the model space under consideration may be large (e.g. 6–16 haplotypic predictors with a sample size of 500–1,000 [Chen and Kao 2006; Epstein and Satten, 2003; French et al., 2006; Stram et al., 2003], which can yield $2^6$ to $2^{16}$ possible models), many of the haplotypic predictors are not likely to be associated

with the phenotype. In this case, it is more efficient to shrink these effect estimates to zero than to estimate them purely. This shrinkage leads to a reduction in variance and can increase the power to detect important haplotypic predictors [Guo and Lin, 2009].

Modifications of classic penalized methods have also been developed to perform haplotype-based analysis and attempt to address issues specific to this type of analysis. Tanck et al. [Souverein et al., 2006, 2008; Tanck et al., 2003] use a modified version of Ridge regression to stabilize inference for rare haplotypes. By constructing an $L_2$-norm penalty term on the differences in coefficients of similar haplotypes, the coefficients of rare haplotypes are smoothed toward that of a similar common haplotype. Chen et al. [2009] develop an adaptive penalized likelihood framework to address the precision-efficiency tradeoff encountered in retrospective methods. Motivated by the fact that typical retrospective case-control estimates of haplotype effects are efficient but sensitive to violations of underlying assumptions (e.g. Hardy-Weinberg equilibrium and gene-environment independence), they construct a penalized estimator based on either an $L_1$-norm or an $L_2$-norm penalty that combines the merits of assumption-free estimators (i.e. robust) and assumption-dependent estimators (i.e. efficient). Tzeng and Bondell [2010] modify traditional adaptive LASSO regression by placing an $L_1$-norm penalty on pair-wise differences of the regression coefficients. This allows for effect comparisons between all pairs of distinct haplotypes, rather than with respect to an arbitrary baseline haplotype, during the estimation process. As a result, the approach is able to sort haplotypes into different groups according to their effect sizes and eliminates the need for a post-hoc pair-wise analysis of haplotype effects. In general, the key of a penalized regression method lies in the form of the penalty—by carefully designing the form of the penalty, one can gear the penalized likelihood approach toward accomplishing various desired tasks.

Penalized regression methods rely on the underlying data likelihood. When analyzing data from case-control studies, one can implement methods based on a prospective likelihood (modeling the probability of disease status conditional on exposure) or a retrospective likelihood (modeling the probability of exposure conditional on disease status). Under a case-control design, a retrospective likelihood should be used because data are collected based on disease status. However, in practice, it is common for researchers to use a prospective likelihood, as it does not require specifying a model for the joint distribution of the genetic and environmental effects. Bypassing this step makes implementing prospective methods much easier than retrospective methods [Lin et al., 2005]. This approach seems congruent with the well-known result that optimizing the prospective likelihood yields the same inference on the disease model parameters as optimizing the retrospective likelihood [Prentice and Pyke, 1979]. This result requires that the distribution of the covariates be free of restrictions, which does not generally hold in haplotype-based analysis. Haplotypes are not directly observed from unphased genotype data. In order to reconstruct the haplotypes and estimate their effects, some assumptions must be placed on their frequency distribution (typically Hardy-Weinberg equilibrium).

Most of the penalized regression approaches mentioned above utilized a prospective likelihood. It has been well documented that when using non-penalized regression methods in haplotype-based analysis of case-control data, ignoring the ascertainment scheme can be detrimental. A prospective analysis can lead to a loss of efficiency and severe bias when assessing the haplotype effects [Cordell, 2006; Satten and Epstein, 2004; Stram et al., 2003]. The aim of this work is to determine whether similar consequences occur when using penalized regression for case-control studies. Specifically, we consider the adaptive LASSO penalty, and use simulation studies to examine the relative performance in parameter estimation and model selection between the penalized method using a prospective likelihood and using a retrospective likelihood. Our results suggest that the impact of using a prospective likelihood in place of a retrospective likelihood depends on (1) the underlying genetic mode of the causal variants, and (2) the genetic model used in the analysis. If the correct genetic model is used, then the difference between the two analyses is negligible for additive and slight for dominant haplotype effects. For recessive haplotype effects, the more appropriate retrospective likelihood clearly outperforms the prospective likelihood. If an additive model is used regardless of the underlying genetic mode, then both retrospective and prospective penalized methods suffer from a sizeable power loss and increase in bias. The impact of using the incorrect genetic model is much bigger on retrospective analyses than prospective analyses, and results in comparable performances for both methods. In addition to extensive simulation studies, we present an application to the Genetic Analysis Workshop 15 rheumatoid arthritis data.

# METHODS

## PROSPECTIVE AND RETROSPECTIVE LIKELIHOODS

Let the vector $(Y_i,G_i,E_i)$ represent the observed data for individual $i$ in a case-control sample of size $n$. Let $Y_i$ be a binary indicator of disease status, where $Y_i = 1$, if individual $i$ is a case and 0 otherwise. Let $G_i$ denote the unphased genotype of individual $i$ at $m$ biallelic SNPs and $E_i$ denote any environmental covariates measured on individual $i$. Let $H_i$ represent the vector of haplotype counts for individual $i$. Although researchers want to investigate the relationship between $Y_i$ and $H_i$, they only have access to $G_i$; therefore, the individual's haplotype set must be inferred from their unphased genotypes.

The relationship between the disease phenotype and the covariates can be characterized by the conditional density function $P(Y \mid H,E)$. A standard approach for binary trait values is logistic regression, which models the conditional probability as

$$P(Y = y|H, E) = \frac{\exp\{y \cdot (\beta_0 + \mathcal{Z}(H, E)^{\mathrm{T}}\beta)\}}{1+\exp\{\beta_0 + \mathcal{Z}(H, E)^{\mathrm{T}}\beta\}},$$

where $\beta_0$ is an intercept, $\beta$ is the vector of disease model parameters representing the log-odds ratios, and $\mathcal{Z}(H,E)$ is a specified vector-valued function of the vector of haplotype counts $H$ and the vector of environmental covariates $E$. For example, one can use an identity function for $\mathcal{Z}(H,E)$ so that $\mathcal{Z}(H,E)^{\mathrm{T}} = [H^{*\mathrm{T}}, E^{\mathrm{T}}]$, where $H^*$ is the vector $H$ with baseline haplotype element removed. Other examples of $\mathcal{Z}(H,E)$ are described in the data generation

section and the choice depends on the genetic model of the haplotype effects adopted in the analysis. The dimension of β is determined by the dimension of $[H^*,E]$; it is the sum of the number of haplotypes (excluding the baseline) and the number of environmental covariates included in the model, along with any interaction terms that may be used. The dimension of β is denoted by $p$. Throughout this work, we assume that the sample size is greater than the dimension of β, i.e. $n>p$. This is consistent with case-control data sets, which typically have sample sizes in the order of $10^2$ to $10^3$ and the number of potential predictors in the order of 10 [Chen and Kao 2006; Epstein and Satten, 2003; French et al., 2006; Stram et al., 2003].

Various likelihood models have been developed to conduct inference about the disease model parameters in haplotype-based analyses while properly accounting for phase uncertainty. The inference can be based either on a prospective likelihood or on a retrospective likelihood. In this work, we consider maximum likelihood methods developed by two groups—one focusing on a prospective approach and the other on a retrospective approach. We implement the prospective method developed in Lake et al. [2003]. Their prospective likelihood models $P(Y_i \mid G_i,E_i)$ and is expressed as

$$L_P = \prod_{i=1}^{n} P(Y_i|G_i,E_i) = \prod_{i=1}^{n} \sum_{H_i \in S(G_i)} P(H_i,Y_i|G_i,E_i)$$
$$= \prod_{i=1}^{n} \sum_{H_i \in S(G_i)} P(Y_i|H_i,E_i)P(H_i),$$

where $S(G)$ is the set of all haplotype pairs consistent with $G$, $P(H=h) = 2\prod_{k=1}^{l} \pi_k^{n_k}/n_k$ under the assumption of Hardy-Weinberg equilibrium, $n_k$ is the number of copies of the $k$th haplotype in $h$, $\pi_k$ is the population frequency of the $k$th haplotype, and $l$ is the number of haplotypes included in the disease model. We implement the retrospective method developed in Lin and Zeng [2006]. Their retrospective likelihood models $P(G_i,E_i \mid Y_i)$ and is expressed as

$$L_R = \prod_{i=1}^{n} P(G_i,E_i|Y_i) = \prod_{i=1}^{n} \sum_{H_i \in S(G_i)} P(H_i,G_i,E_i|Y_i)$$
$$\propto \prod_{i=1}^{n} \sum_{H_i \in S(G_i)} P(Y_i|H_i,E_i)P(H_i)P(E_i|G_i).$$

The only difference between the two likelihoods is the conditional density function $P(E_i \mid G_i)$ found in the retrospective likelihood. The parameters in this model are of no interest to researchers performing haplotype-based association analysis, but they must be estimated in order to make proper inference when using a retrospective design. Specifying a model for this conditional density function and the subsequent maximum likelihood estimation are computationally intensive. As a result, researchers often rely on prospective methods when analyzing case-control data even though retrospective methods are dictated by the ascertainment scheme [Lin and Zeng, 2006].

## HAPLOTYPE ANALYSIS VIA PENALIZED LIKELIHOOD METHODS

While many different penalized likelihood methods can be used in haplotype-based association analysis, we consider the adaptive LASSO (ALASSO) penalty in this work. This approach achieves simultaneous variable selection and parameter estimation and is an oracle procedure. This refers to the fact that the approach asymptotically selects the correct model, and the resulting estimator is root-n consistent and asymptotically normal with the same variance as if the true model were known beforehand [Zou, 2006].

The ALASSO effect estimates are obtained by minimizing a penalized negative log-likelihood. These estimates are expressed as

$$\hat{\beta}_\lambda = \arg\min_\beta - \ell_n(\beta,\phi) + \lambda \sum_{j=1}^{p} w_j|\beta_j|,$$

where $\ell_n(\beta,\phi)$ denotes the log-likelihood, $\phi$ is a (possible) set of nuisance parameters (e.g. the haplotype frequencies, $\pi_k$), $\lambda$ is the non-negative regularization parameter that controls the amount of shrinkage, and $w_j$ are data-dependent weights. By placing an $L_1$-norm penalty on the regression coefficients, the ALASSO can set their estimates to exactly zero if the value of $\lambda$ is large enough. It is this feature that allows the procedure to perform simultaneous variable selection and parameter estimation. Unlike its predecessor the LASSO, the ALASSO places a different penalty on each coefficient through the use of adaptive weights that are inversely proportional to their relative importance. Consequently, haplotypes with negligible effects receive larger penalties and are more readily shrunk to zero. This allows the effects of associated haplotypes to be estimated more efficiently. Zou [2006] proposed to set the weights as $w_j = |\tilde{\beta}_j|^{-\gamma}$, where $\tilde{\beta}_j$ is an initial root-n consistent estimator of $\beta_j$ and $\gamma>0$ is an additional tuning parameter. In our analysis, we chose $\gamma=1$ and let $\tilde{\beta}_j$ be the maximum likelihood estimate of the haplotype effect computed by haplo.glm in R and HAPSTAT in Linux for the prospective and retrospective likelihoods, respectively [Lake et al., 2003; Lin et al., 2005].

When performing penalized likelihood methods, it is typical to center and scale the design matrix. Scaling assures that each column of the design matrix has the same variance and the resulting estimator is scale-equivariant (i.e. multiplication of any predictor by any constant will simply divide the resulting slope estimate by the identical constant; hence the linear predictor remains unchanged). This is desirable so that if, for example, the units of a predictor are changed, such as feet to inches, the resulting predicted values will remain unchanged. Often the predictors are also centered, so that in the normal linear regression setting, the intercept can be omitted and the slope parameter estimates are orthogonal to the intercept estimate. However, in the generalized linear models as considered here, this is not the case; hence, the design matrix is typically not centered. Furthermore, in the ALASSO analysis, we also do not scale the imputed haplotype design matrix because the adaptive weights we set (i.e. $|\tilde{\beta}_j|^{-1}$) are scale-equivariant. The use of scale-equivariant weights automatically forces the resulting estimator to be scale-equivariant.

The ALASSO solution ($\hat{\beta}_\lambda$) also depends on the value of $\lambda$. The regularization parameter controls the tradeoff between model fit and model sparsity. By including more predictors, one can continually improve the fit on the training data at the expense of interpretability and over fitting. Many model selection criteria, such as Mallow's

$C_P$, Akaike information criterion (AIC), Bayesian information criterion (BIC), and cross validation [Arlot and Celisse, 2010; Hastie et al., 2009; Shao, 1997], can be used to determine the appropriate value of $\lambda$ from an exhaustive grid search. Because the goal of haplotype-based association analysis is more aligned with selecting the true model than minimizing prediction error, we use the BIC for tuning, which can achieve consistent model selection [Yang, 2005]. BIC is defined as

$$\text{BIC} = -2\ell_n(\hat{\beta}_\lambda, \hat{\phi}) + \text{d}f_\lambda \cdot \log(n),$$

where $\ell_n(\hat{\beta}_\lambda, \hat{\phi})$ is the log-likelihood evaluated at the estimated regression coefficients and maximized over $\phi$ for a given $\lambda$, and $\text{d}f_\lambda$ is the degrees of freedom, which equals the number of non-zero elements in $(\hat{\beta}_\lambda, \hat{\phi})$. The $\lambda$ that minimizes the BIC is chosen as the regularization parameter, and its corresponding $\hat{\beta}_\lambda$ is the ALASSO estimate. For comparison, we also present some of the results using AIC as a tuning method. In the definition of AIC, the penalty on the degrees of freedom is changed from $\log(n)$ to 2. As a result, models selected using AIC incur less shrinkage, and the chosen ALASSO estimate will be closer to the unpenalized MLE estimate than those found using BIC.

For computational convenience, the objective function created via the least squares approximation (LSA) method was used to calculate the ALASSO solution. The LSA method replaces the objective function of the original ALASSO problem with a least squares objective function [Wang and Leng, 2007]. The method is motivated by a standard Taylor series expansion of $-\ell_n(\beta, \phi)$ about $(\tilde{\beta}, \tilde{\phi})$, the function's unpenalized minimizer, and shows that the ALASSO estimate has the exact same asymptotic distribution as the estimator given by

$$\hat{\beta}_\lambda = \arg\min_\beta (\beta - \tilde{\beta})^{\mathrm{T}} \tilde{\Sigma}^{-1} (\beta - \tilde{\beta}) + \lambda \sum_{j=1}^{p} w_j |\beta_j|,$$

where $\tilde{\Sigma}$ is the estimated covariance matrix of $\tilde{\beta}$. Note that the minimizer of the unpenalized least squares objective function is exactly the maximum likelihood estimator. Hence, as with the penalized likelihood, varying the tuning parameter yields a continuous solution path from the MLE to the solution with all coefficients equal to zero. Because the underlying data likelihoods are not quadratic in the regression coefficients, using the alternative least squares objective function greatly reduces the computational costs for finding the ALASSO solution [Wang and Leng, 2007]. Using the LSA method eliminates the need for an iterative procedure to perform optimization; it only requires one unpenalized fit of the original objective function and then a grid search to determine $\lambda$. The final estimate is again chosen by minimizing the BIC (and AIC for some results).

# SIMULATION STUDIES

We performed simulation studies to examine the performance of the ALASSO method under two competing data likelihoods when analyzing case-control data. Specifically, we wanted to determine if using a prospective likelihood in place of the more appropriate retrospective likelihood was detrimental when performing haplotype-based analyses using a penalized likelihood method. To answer this question, we compared the parameter estimation and model selection properties of each approach. For ease of discussion, let *aPro* refer to ALASSO coupled with a prospective likelihood and *aRetro* refer to ALASSO coupled with a retrospective likelihood.

## SIMULATION SETTINGS

Our simulation studies were based on two haplotype distributions (given in Table I) studied by Lin and Huang [2008]. These distributions are based on the common haplotypes formed by five SNPs on chromosome 18 in the CEU sample of the HapMap data. The SNPs used to build the first haplotype distribution were in strong linkage disequilibrium, while those used to build the second haplotype distribution were not. Distribution 1 represents a haplotype distribution with a few high frequency haplotypes, while the haplotype frequencies in Distribution 2 are more uniform. Each distribution was normalized so that the haplotype frequencies summed to 1. Because eight haplotypes define Distribution 1 and 11 haplotypes define Distribution 2, the specific dimension of $\beta$ is $p = 7$ and $p = 10$, respectively.

For each haplotype distribution, we considered two simulation studies—one in which a single haplotype was associated with the disease (Simulation I) and one in which two haplotypes were associated with the disease (Simulation II). Because our focus was on identifying and estimating disease-haplotype associations, only genetic covariates were considered in our simulation studies (i.e. $E$ is taken to be $\emptyset$). We took the sample size to be $n = 1{,}000$ with an equal number of cases and controls. In both simulation studies, we examined the effect of varying the genetic mode of the associated haplotype(s) on the performance of aPro and aRetro. We allowed the associated haplotype(s) to act additively, dominantly, or recessively with respect to disease risk. In practice, the genetic mode of a risk haplotype is unknown a priori, and researchers typically analyze the data additively regardless of the true genetic model. To mimic this scenario, we analyzed each data set using the correct genetic model and again using an additive model. We use the term "genetic mode" to refer to the true underlying architecture of the relationship between the disease and the risk haplotypes, and the term "genetic model" to refer to the assumed architecture used in the analysis.

**TABLE I. Haplotype distributions used in simulations**

| | Distribution 1 | | Distribution 2 | |
|---|---|---|---|---|
| Hap ID | Haplotype | Frequency | Haplotype | Frequency |
| 1 | 00000 | 0.406 | 00010 | 0.131 |
| 2 | 00001 | 0.213 | 00001 | 0.105 |
| 3 | 01111 | 0.141 | 10010 | 0.103 |
| 4 | 10000 | 0.132 | 10101 | 0.100 |
| 5 | 10001 | 0.055 | 00100 | 0.088 |
| 6 | 01000 | 0.021 | 10100 | 0.088 |
| 7 | 01100 | 0.018 | 00101 | 0.086 |
| 8 | 01001 | 0.014 | 01101 | 0.084 |
| 9 | | | 10001 | 0.081 |
| 10 | | | 10000 | 0.079 |
| 11 | | | 00000 | 0.055 |

**TABLE II. Settings for SIM I and SIM II (odds ratios)**

| | Distribution 1 | | | | | | Distribution 2 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Sim I | | | Sim II | | | Sim I | | | Sim II | | |
| Hap ID | Freq | R | C | R/R | R/C | C/C | Freq | R | C | R/R | R/C | C/C |
| 1 | 0.406 | | | | | | 0.131 | | | | | |
| 2 | 0.213 | 1 | 1 | 1 | 1 | 1 | 0.105 | 1 | θ | 1 | θ | θ |
| 3 | 0.141 | 1 | 1 | 1 | 1 | θ | 0.103 | 1 | 1 | 1 | 1 | θ |
| 4 | 0.132 | 1 | θ | 1 | θ | θ | 0.100 | 1 | 1 | 1 | 1 | 1 |
| 5 | 0.055 | θ | 1 | θ | θ | 1 | 0.088 | 1 | 1 | 1 | 1 | 1 |
| 6 | 0.021 | 1 | 1 | θ | 1 | 1 | 0.088 | 1 | 1 | 1 | 1 | 1 |
| 7 | 0.018 | 1 | 1 | 1 | 1 | 1 | 0.086 | 1 | 1 | 1 | 1 | 1 |
| 8 | 0.014 | 1 | 1 | 1 | 1 | 1 | 0.084 | 1 | 1 | 1 | 1 | 1 |
| 9 | | | | | | | 0.081 | 1 | 1 | 1 | 1 | 1 |
| 10 | | | | | | | 0.079 | 1 | 1 | θ | 1 | 1 |
| 11 | | | | | | | 0.055 | θ | 1 | θ | θ | 1 |

In addition to genetic mode, we varied the frequency and effect size of the associated haplotype. In Simulation I (SIM I), a rare or a common haplotype was chosen to be the associated genetic variant for each haplotype distribution. A haplotype with a frequency less than 0.10 was considered rare; otherwise, it was considered common. We set the effect sizes of the associated haplotype (in terms of the odds ratio θ) so that the power of finding the effect fell in a reasonable range. Under additive and dominant modes, we set $\theta = \{1.0, 1.3, 1.5, 1.7, 2.0\}$, and under a recessive mode, we set $\theta = \{1.0, 2.0, 2.5, 3.0, 3.5\}$. We let $\theta = 1$ to examine the performance of the approaches under a null model. In Simulation II (SIM II), we allowed two haplotypes to be associated with the disease, where the associated haplotypes were both rare, one rare and one common, or both common. The odds ratios of both associated haplotypes were set to $\theta = 1.7$ for additive and dominant modes and $\theta = 3.0$ for a recessive mode. The settings for Simulations I and II can be found in Table II. In all, 78 different simulation settings were studied.

## DATA GENERATION

We generated the haplotype pair of an individual conditional on their disease status and then dissolved the haplotype pair into its unphased genotypes. Let $P(H = h \mid Y = y)$ denote the probability of having a particular haplotype pair conditional on disease status. This probability can be expressed as

$$P(H = h|Y = y) = \frac{P(Y = y|H = h) \cdot P(H = h)}{\sum_h P(Y = y|H = h) \cdot P(H = h)}.$$

For a case individual, $P(Y = 1 \mid H = h)$ was found using the logistic regression model

$$P(Y = 1|H) = \frac{\exp\{\beta_0 + \mathcal{Z}(H)^{\mathrm{T}}\beta\}}{1 + \exp\{\beta_0 + \mathcal{Z}(H)^{\mathrm{T}}\beta\}}.$$

For a control individual, $P(Y=0|H=h) = 1 - P(Y=1|H=h)$. The function $\mathcal{Z}(\cdot)$ depends on the genetic mode of the haplotype(s) associated with the disease. If the haplotype acts additively with respect to disease risk, then $\mathcal{Z}(H) = H^*$ where $H^*$ is the haplotype-count vector $H$ with the baseline haplotype element removed. If the haplotype acts dominantly, then $\mathcal{Z}(H) = I\{H^* \geq 1\}$, where the inequality is taken component wise, and $I\{A\} = 1$ if $A$ is true. If the haplotype acts recessively, then $\mathcal{Z}(H) = I\{H^* = 2\}$. The vector β was taken to be the log of the vectors given in Table II for each simulation setting. The value of $\beta_0$ was set to maintain a disease prevalence between 3% and 5%. Once $P(Y = y \mid H = h)$ was calculated for each haplotype pair formed from the haplotype distributions given in Table I, the vectors $P_{H|Y=y} = (P(H=h_1|Y=y) \cdots P(H=h_q|Y=y))$ were calculated for $Y = 0$ and $Y = 1$ where $q$ is total number of haplotype pairs. The sample was generated by taking $n/2$ draws from the multinomial distribution parameterized by $P_{H|Y=0}$ to determine the haplotype pairs of the controls, and by taking $n/2$ draws from the multinomial distribution parameterized by $P_{H|Y=1}$ to determine the haplotype pairs of the cases. The haplotype pair of each individual was then dissolved into its unphased genotype.

## COMPUTATIONAL DETAILS

For each simulation setting, 1,000 replicate data sets were generated, except for simulation under the null (i.e. $\theta = 1$) where 2,000 replicated data sets were generated. We doubled the number of simulated data sets in the null simulation to obtain more stable estimates of the false-positive counts (FP counts). For each data set, analysis began by calculating the unpenalized MLEs of the haplotype log-odds ratios. Prospective MLEs were obtained using haplo.glm in R [Lake et al., 2003] and retrospective MLEs were obtained using HAPSTAT in Linux [Lin et al., 2005]. The estimated covariance matrix of the MLEs was also obtained from each program. The final aPro and aRetro estimates were calculated by using the MLEs and their covariance matrix to compute the ALASSO solution via LSA. Based on these final estimates, estimation and model selection measures were calculated to compare the performance of the aPro and aRetro approaches. The estimation measures provided in this analysis are the bias and mean square error (MSE) of haplotype effect estimates. The model selection measures provided are the FP count (i.e. the number of non-risk haplotypes retained in the model) and the false-negative count (FN count, i.e. the number of risk haplotypes not retained in the model), for which a smaller FN count indicates better power to identify risk haplotypes. For each measure, we report the mean across the replicated data sets and the corresponding standard error of the mean. The functions used to obtain the results of this simulation study are available at the corresponding author's website http://www4.stat.ncsu.edu/~tzeng/publications.php.

# SIMULATION RESULTS

We present the results of Null Simulation (i.e. no-risk haplotypes) for both haplotype distributions in Table III. For the simulations involving risk haplotypes, because the pattern of results was similar across both haplotype distributions, for brevity we focus the discussion on the results for the first haplotype distribution. The results of Simulation I (single-risk haplotype) for this setting are found in Tables IV and V. The results of Simulation II (two-risk haplotypes) for this setting are found in Tables VI and VII. The discussion generalizes to the second haplotype distribution, and specific results are shown in

**TABLE III. BIC-penalized results of null simulation (no-risk haplotypes)**

| | Model selection results | | Parameter estimation results | | | |
|---|---|---|---|---|---|---|
| | False positives[a] | | Bias | | MSE | |
| Model[b] | Pro | Retro | Pro | Retro | Pro | Retro |
| Haplotype Distribution 1 (Correct analysis) | | | | | | |
| Additive | 0.025[c] | 0.028 | 0.000 | −0.001 | 0.002 | 0.002 |
| | (0.005) | (0.006) | (0.002) | (0.002) | (0.001) | (0.001) |
| Dominant | 0.025 | 0.033 | 0.000 | 0.000 | 0.002 | 0.003 |
| | (0.006) | (0.006) | (0.002) | (0.002) | (0.001) | (0.001) |
| Recessive | 0.026 | 0.029 | −0.001 | 0.003 | 0.001 | 0.003 |
| | (0.008) | (0.011) | (0.001) | (0.002) | (0.001) | (0.002) |
| Haplotype Distribution 2 (Correct analysis) | | | | | | |
| Additive | 0.062 | 0.055 | 0.000 | 0.000 | 0.002 | 0.002 |
| | (0.020) | (0.017) | (0.002) | (0.002) | (0.001) | (0.001) |
| Dominant | 0.059 | 0.054 | 0.001 | 0.001 | 0.001 | 0.001 |
| | (0.018) | (0.014) | (0.002) | (0.001) | (0.001) | (0.001) |
| Recessive | 0.057 | 0.054 | 0.000 | 0.007 | 0.000 | 0.007 |
| | (0.001) | (0.017) | (0.000) | (0.005) | (0.000) | (0.005) |

[a]False positives are counts out of 7 for Distribution 1 and out of 10 for Distribution 2.
[b]Model refers to genetic model adopted in the analysis.
[c]Measures in this table are found by averaging over 2,000 simulated data sets; standard errors (SE) are shown in parentheses. (Additionally, the bias and MSE were first averaged across all null haplotypes in the model)

Tables VIII–XI. For each simulation, the results are broken down into two broad categories—correct analysis versus additive analysis. Correct analysis refers to specifying the correct genetic model when analyzing the data using haplo.glm or HAPSTAT, while additive analysis refers to analyzing non-additive data additively.

## NULL SIMULATION

For each haplotype distribution, both aPro and aRetro have desirable and similar performances under the null model (Table III). The FP count is low for both methods (at most 0.033 FPs out of 7 variables for Distribution 1, and at most 0.062 FPs out of 10 variables for Distribution 2). The effect estimations are also very similar, with the bias from both methods ranging from −0.001 to 0.007, and MSE ranging from 0.000 to 0.007. When comparing results between the two haplotype distributions, the FP count is lower for Distribution 1 than Distribution 2. This result is not unexpected, as the dimension of Distribution 2 is larger.

## SIMULATION I FOR HAPLOTYPE DISTRIBUTION 1

**Additive genetic mode.** When the risk haplotype acts additively, the FN and FP counts are very close for aPro and aRetro (Table IV). The FN count decreases when the effect size of the risk haplotype increases or if the frequency of the risk haplotype increases, but the relative performance of the two methods stays the same. When comparing the bias and MSE of aPro and aRetro, the measures from the two procedures are also very close, which indicates that aPro and aRetro perform similarly with respect to effect estimation. For both procedures, the bias on the effect estimates is negative and the magnitudes are larger than what have been reported for an unpenalized likelihood analysis (e.g. Lin and Zeng [2006] and Tables XIV and XV). These results are not unexpected when using a penalized likelihood approach. When a haplotype is not included in the model, its effect estimate is shrunk toward zero or set to exactly zero. Shrinkage can cause a large bias on the effect estimates. The impact of using a penalized method on the bias is greatest when the effect size is large and the power to detect the risk haplotype is low. As a result, a decrease in effect size or an increase in power does not necessarily guarantee a reduction in bias; the magnitude of the bias is a compromise between these two factors. This phenomenon is seen when examining the biases in Table IV. First, the bias for rare-risk haplotypes is larger than the corresponding common-risk haplotypes with the same effect size because the power to detect rare-risk haplotype is smaller. Second, for a given risk haplotype, the bias on the effect estimate increases as the effect size increases until the power to detect the risk haplotype becomes large enough to overcome the shrinkage, and the bias on the effect estimate begins to decrease.

A similar pattern is observed when examining the MSE of the two procedures. Again, the MSE of aPRo and aRetro is larger than what has been found in an unpenalized likelihood analysis. MSE is an estimation measure that incorporates both the variance of an estimator and its bias. Because the effect estimates obtained from penalized methods are typically more efficient than those obtained from the corresponding unpenalized methods, it appears that the MSE of the effect estimates from aPro or aRetro could be dominated by their biases.

**Dominant genetic mode.** Under a dominant genetic mode, aRetro performs slightly better than aPro when the data are analyzed under the correct genetic model

**TABLE IV. BIC-penalized results of Sim I (one-risk haplotype) for Distribution 1—correct analysis[a]**

| | | | Model selection results | | | | Parameter estimation results | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | False negatives[b] | | False positives | | Bias | | MSE | |
| Mode[c] | Freq | OR | Pro | Retro | Pro | Retro | Pro | Retro | Pro | Retro |
| Additive | Rare | 1.3 | 0.943[d] | 0.945 | 0.019 | 0.024 | −0.230 | −0.231 | 0.071 | 0.071 |
| | | | (0.007) | (0.007) | (0.004) | (0.005) | (0.004) | (0.004) | (0.001) | (0.001) |
| | | 1.5 | 0.820 | 0.805 | 0.010 | 0.015 | −0.304 | −0.298 | 0.142 | 0.140 |
| | | | (0.012) | (0.013) | (0.007) | (0.006) | (0.005) | (0.006) | (0.003) | (0.004) |
| | | 1.7 | 0.623 | 0.614 | 0.051 | 0.050 | −0.305 | −0.304 | 0.186 | 0.183 |
| | | | (0.015) | (0.015) | (0.008) | (0.008) | (0.01) | (0.01) | (0.004) | (0.004) |
| | | 2.0 | 0.244 | 0.239 | 0.059 | 0.064 | −0.201 | −0.200 | 0.144 | 0.142 |
| | | | (0.014) | (0.013) | (0.008) | (0.008) | (0.010) | (0.010) | (0.006) | (0.006) |
| | Common | 1.3 | 0.801 | 0.792 | 0.030 | 0.027 | −0.190 | −0.188 | 0.059 | 0.058 |
| | | | (0.013) | (0.013) | (0.005) | (0.005) | (0.005) | (0.005) | (0.001) | (0.001) |
| | | 1.5 | 0.465 | 0.460 | 0.075 | 0.075 | −0.189 | −0.187 | 0.082 | 0.081 |
| | | | (0.016) | (0.016) | (0.006) | (0.007) | (0.005) | (0.007) | (0.002) | (0.001) |
| | | 1.7 | 0.120 | 0.112 | 0.066 | 0.073 | −0.103 | −0.101 | 0.050 | 0.048 |
| | | | (0.010) | (0.010) | (0.009) | (0.009) | (0.006) | (0.006) | (0.003) | (0.003) |
| | | 2.0 | 0.003 | 0.003 | 0.054 | 0.056 | −0.071 | −0.072 | 0.027 | 0.027 |
| | | | (0.002) | (0.002) | (0.008) | (0.008) | (0.005) | (0.005) | (0.001) | (0.001) |
| Dominant | Rare | 1.3 | 0.946 | 0.941 | 0.018 | 0.028 | −0.230 | −0.228 | 0.072 | 0.071 |
| | | | (0.007) | (0.007) | (0.004) | (0.005) | (0.004) | (0.004) | (0.001) | (0.001) |
| | | 1.5 | 0.870 | 0.840 | 0.025 | 0.015 | −0.324 | −0.311 | 0.151 | 0.146 |
| | | | (0.011) | (0.012) | (0.008) | (0.007) | (0.006) | (0.007) | (0.002) | (0.002) |
| | | 1.7 | 0.659 | 0.632 | 0.034 | 0.031 | −0.317 | −0.307 | 0.196 | 0.188 |
| | | | (0.015) | (0.015) | (0.006) | (0.006) | (0.010) | (0.010) | (0.004) | (0.004) |
| | | 2.0 | 0.332 | 0.303 | 0.064 | 0.069 | −0.251 | −0.239 | 0.184 | 0.170 |
| | | | (0.015) | (0.015) | (0.009) | (0.009) | (0.011) | (0.011) | (0.007) | (0.007) |
| | Common | 1.3 | 0.864 | 0.849 | 0.024 | 0.025 | −0.207 | −0.203 | 0.063 | 0.062 |
| | | | (0.011) | (0.011) | (0.006) | (0.005) | (0.004) | (0.005) | (0.001) | (0.001) |
| | | 1.5 | 0.590 | 0.555 | 0.060 | 0.055 | −0.231 | −0.224 | 0.101 | 0.096 |
| | | | (0.016) | (0.016) | (0.007) | (0.007) | (0.006) | (0.007) | (0.002) | (0.002) |
| | | 1.7 | **0.263**[e] | **0.206** | 0.064 | 0.067 | −0.161 | −0.145 | **0.088** | **0.073** |
| | | | (0.014) | (0.013) | (0.008) | (0.009) | (0.008) | (0.007) | (0.004) | (0.003) |
| | | 2.0 | 0.035 | 0.025 | 0.062 | 0.065 | −0.090 | −0.091 | 0.046 | 0.041 |
| | | | (0.006) | (0.005) | (0.008) | (0.008) | (0.006) | (0.006) | (0.003) | (0.003) |
| Recessive | Rare | 2.0 | **0.990** | **0.910** | 0.010 | 0.035 | **−0.693** | **−0.583** | 0.480 | 0.474 |
| | | | (0.003) | (0.011) | (0.007) | (0.013) | 0.000 | (0.026) | (0.000) | (0.005) |
| | | 2.5 | **0.990** | **0.870** | 0.010 | 0.010 | **−0.916** | **−0.753** | **0.840** | **0.753** |
| | | | (0.003) | (0.013) | (0.007) | (0.007) | (0.000) | (0.031) | (0.000) | (0.016) |
| | | 3.0 | **0.990** | **0.770** | 0.010 | 0.015 | **−1.099** | **−0.764** | **1.207** | **0.987** |
| | | | (0.002) | (0.014) | (0.007) | (0.009) | (0.000) | (0.045) | (0.000) | (0.030) |
| | | 3.5 | **0.995** | **0.720** | 0.005 | 0.005 | **−1.253** | **−0.865** | **1.569** | **1.163** |
| | | | (0.001) | (0.009) | (0.005) | (0.005) | (0.000) | (0.046) | (0.000) | (0.047) |
| | Common | 2.0 | **0.925** | **0.695** | 0.000 | 0.010 | **−0.603** | **−0.441** | **0.467** | **0.347** |
| | | | (0.008) | (0.015) | (0.000) | (0.007) | (0.023) | (0.028) | (0.006) | (0.014) |
| | | 2.5 | **0.735** | **0.350** | 0.010 | 0.030 | **−0.549** | **−0.329** | **0.711** | **0.321** |
| | | | (0.014) | (0.015) | (0.007) | (0.012) | (0.045) | (0.033) | (0.023) | (0.027) |
| | | 3.0 | **0.480** | **0.110** | 0.015 | 0.045 | **−0.440** | **−0.226** | **0.631** | **0.188** |
| | | | (0.016) | (0.010) | (0.009) | (0.016) | (0.047) | (0.026) | (0.040) | (0.026) |
| | | 3.5 | **0.355** | **0.060** | 0.030 | 0.020 | **−0.367** | **−0.242** | **0.671** | **0.176** |
| | | | (0.015) | (0.008) | (0.012) | (0.010) | (0.052) | (0.024) | (0.055) | (0.026) |

[a]Correct analysis means the genetic model adopted in the analysis is the same as the underlying genetic mode.
[b]False negatives are counts out of 1 and false positives are counts out of 6.
[c]Mode refers to underlying genetic mode of data; Freq refers to frequency of risk haplotype; OR refers to the odds ratio of the risk haplotype.
[d]Measures in this table are found by averaging over 1,000 simulated data sets; standard errors (SE) are shown in parentheses.
[e]Bolded measures (pro vs. retro) are statistically significantly different at 0.05 level (i.e. the $\pm 2$SE intervals do NOT overlap).

**TABLE V. BIC-penalized results of Sim I (one-risk haplotype) for Distribution 1—additive analysis[a]**

| | | | Model selection results | | | | Parameter estimation results | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | False negatives[b] | | False positives | | Bias | | MSE | |
| Mode[c] | Freq | OR | Pro | Retro | Pro | Retro | Pro | Retro | Pro | Retro |
| Dominant | Rare | 1.3 | 0.949[d] | 0.956 | 0.023 | 0.020 | −0.234 | −0.237 | 0.070 | 0.070 |
| | | | (0.007) | (0.006) | (0.005) | (0.004) | (0.004) | (0.004) | (0.001) | (0.001) |
| | | 1.5 | 0.865 | 0.860 | 0.025 | 0.025 | −0.326 | −0.324 | 0.149 | 0.148 |
| | | | (0.011) | (0.011) | (0.007) | (0.006) | (0.006) | (0.006) | (0.003) | (0.004) |
| | | 1.7 | 0.665 | 0.677 | 0.040 | 0.040 | −0.331 | −0.342 | 0.195 | 0.197 |
| | | | (0.015) | (0.015) | (0.006) | (0.006) | (0.009) | (0.009) | (0.004) | (0.004) |
| | | 2.0 | 0.356 | 0.366 | 0.074 | 0.083 | −0.286 | −0.304 | 0.194 | 0.198 |
| | | | (0.015) | (0.015) | (0.009) | (0.010) | (0.011) | (0.010) | (0.007) | (0.007) |
| | Common | 1.3 | 0.879 | 0.883 | 0.024 | 0.025 | −0.219 | −0.221 | 0.062 | 0.062 |
| | | | (0.010) | (0.010) | (0.005) | (0.006) | (0.004) | (0.004) | (0.001) | (0.001) |
| | | 1.5 | 0.650 | 0.660 | 0.060 | 0.050 | −0.273 | −0.281 | 0.110 | 0.111 |
| | | | (0.015) | (0.015) | (0.007) | (0.007) | (0.004) | (0.005) | (0.004) | (0.003) |
| | | 1.7 | 0.291 | 0.301 | 0.064 | 0.068 | −0.225 | −0.242 | 0.099 | 0.103 |
| | | | (0.014) | (0.015) | (0.008) | (−0.009) | (0.007) | (0.007) | (0.004) | (0.004) |
| | | 2.0 | 0.049 | 0.051 | 0.081 | 0.090 | **−0.186** | **−0.216** | 0.068 | 0.077 |
| | | | (0.007) | (0.007) | (0.009) | (0.010) | (0.006) | (0.006) | (0.003) | (0.003) |
| Recessive | Rare | 2.0 | 0.985 | 0.985 | 0.025 | 0.035 | −0.688 | −0.689 | 0.481 | 0.481 |
| | | | (0.004) | (0.004) | (0.011) | (0.013) | (0.006) | (0.005) | (0.006) | (0.006) |
| | | 2.5 | **1.000**[e] | **0.995** | 0.025 | 0.025 | −0.916 | −0.914 | 0.840 | 0.836 |
| | | | (0.000) | (0.002) | (0.011) | (0.011) | (0.000) | (0.002) | (0.000) | (0.003) |
| | | 3.0 | 0.985 | 0.980 | 0.025 | 0.030 | −1.089 | −1.088 | 1.192 | 1.189 |
| | | | (0.004) | (0.004) | (0.011) | (0.012) | (0.005) | (0.006) | (0.009) | (0.009) |
| | | 3.5 | **1.000** | **0.990** | 0.000 | 0.005 | −1.253 | −1.248 | 1.569 | 1.561 |
| | | | (0.000) | (0.003) | (0.000) | (0.005) | (0.000) | (0.003) | (0.000) | (0.006) |
| | Common | 2.0 | **0.975** | **0.950** | 0.005 | 0.020 | −0.685 | −0.684 | 0.472 | 0.470 |
| | | | (0.005) | (0.007) | (0.005) | (0.010) | (0.004) | (0.004) | (0.004) | (0.004) |
| | | 2.5 | 0.890 | 0.890 | 0.015 | 0.025 | −0.876 | −0.877 | 0.781 | 0.782 |
| | | | (0.010) | (0.010) | (0.009) | (0.013) | (0.008) | (0.008) | (0.012) | (0.012) |
| | | 3.0 | 0.860 | 0.845 | 0.015 | 0.030 | −1.051 | −1.044 | 1.120 | 1.107 |
| | | | (0.011) | (0.011) | (0.009) | (0.012) | (0.008) | (0.009) | (0.016) | (0.017) |
| | | 3.5 | 0.740 | 0.715 | 0.010 | 0.020 | −1.156 | −1.146 | 1.365 | 1.343 |
| | | | (0.014) | (0.014) | (0.007) | (0.010) | (0.012) | (0.012) | (0.025) | (0.026) |

[a]Additive analysis means an additive model is adopted in the analysis when the underlying genetic mode is not.
[b]False negatives are counts out of 1 and false positives are counts out of 6.
[c]Mode refers to underlying genetic mode of data; Freq refers to frequency of risk haplotype; OR refers to the odds ratio of the risk haplotype.
[d]Measures in this table are found by averaging over 1,000 simulated data sets; standard errors (SE) are shown in parentheses.
[e]Bolded measures (pro vs. retro) are statistically significantly different at 0.05 level (i.e. the ±2SE intervals do NOT overlap).

(Table IV). While having comparable FP counts, the FN count of aRetro is less than that of aPro although the differences are not always significant (i.e. the intervals of mean ± 2SEs overlap). Similar results are observed for the estimation measures. The bias and MSE of aRetro are similar to or lower than that of aPro.

When the data are incorrectly analyzed additively (Table V), the FP count is similar for both methods, and stays roughly at the same level as the correct analysis. However, both methods suffer from an increase in FNs, bias, and MSE when compared to the performance of the correct analysis. It appears that the impact of using an incorrect genetic model is larger on aRetro than on aPro. For example, the aRetro FN count of retaining a common dominant risk haplotype with OR = 1.7 increased from 0.206 to 0.301 under the additive analysis, while the FN count of aPro increased from 0.263 to 0.291, with the SEs of all figures around 0.014. The aRetro bias in this setting increases from $|-0.145|$ to $|-0.242|$, while the bias of aPro increases from $|-0.161|$ to $|-0.225|$. As a result, aPro performs worse than aRetro under correct analysis (Table IV) but is comparable or slightly better than aRetro under the additive analysis (Table V).

**Recessive genetic mode.** Under a recessive genetic mode, aRetro clearly outperforms aPro when the data are analyzed under the correct model (Table IV). Both methods have comparable FP counts, but the FN count of aRetro is significantly less than that of aPro. When the risk haplotype is rare, the FN count of aPro is almost 1, indicating that aPro has almost no power to retain the risk haplotype in the model under this scenario. When the risk haplotype is common, the FN count of aRetro is 75% to 17% smaller than the FN count of aPro. The lack of power of aPro also manifests in significantly more bias on the effect estimates.

The bias on effect estimates from aRetro is 84% to 50% of the bias from aPro. The MSE on the effect estimates from aRetro is 90% to 26% of the MSE from aPro in most cases.

When this data are incorrectly analyzed additively (Table V), the performance of each method suffers from an increase in FNs, bias, and MSE. The magnitude of the performance loss due to incorrect modeling is more severe than what was observed under a dominant mode and is much more severe for aRetro than aPro. For example, the aRetro FN count of retaining a common recessive risk haplotype with OR = 2.5 increased from 0.350 to 0.890 under the additive analysis, while the FN count of aPro increased from 0.735 to 0.890, with the SEs of all figures around 0.014. The aRetro bias in this setting increases from $|-0.329|$ to $|-0.877|$, while the bias of aPro increases from $|-0.549|$ to $|-0.876|$. Consequently, while aRetro exhibits absolute superiority over the aPro method (Table IV) under a correct analysis, it becomes comparable to or slightly better than aPro under the additive analysis (Table V).

## SIMULATION II FOR HAPLOTYPE DISTRIBUTION 1

Simulation II examines the performance of aPro and aRetro when two haplotypes are associated with the disease (Tables VI and VII). Under each genetic mode, the patterns

**TABLE VI. BIC-penalized results of Sim II (two-risk haplotypes) for Distribution 1—correct analysis[a]**

| | | | Model selection results | | | | Parameter estimation results | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | False negatives[b] | | False positives | | Bias | | MSE | |
| Mode[c] | Freq | | Pro | Retro | Pro | Retro | Pro | Retro | Pro | Retro |
| Additive | R/R | R1[d] | 1.548[e] | 1.536 | 0.029 | 0.031 | −0.443 | −0.444 | 0.267 | 0.266 |
| | | | (0.020) | (0.021) | (0.006) | (0.006) | (0.008) | (0.008) | (0.002) | (0.003) |
| | | R2 | | | | | −0.327 | −0.324 | 0.193 | 0.190 |
| | | | | | | | (0.009) | (0.009) | (0.004) | (0.004) |
| | R/C | R1 | 0.596 | 0.588 | 0.089 | 0.098 | −0.250 | −0.249 | 0.156 | 0.155 |
| | | | (0.020) | (0.020) | (0.010) | (0.011) | (0.006) | (0.006) | (0.003) | (0.002) |
| | | C1 | | | | | −0.109 | −0.109 | 0.048 | 0.046 |
| | | | | | | | (0.010) | (0.010) | (0.004) | (0.004) |
| | C/C | C1 | 0.204 | 0.205 | 0.095 | 0.090 | −0.113 | −0.115 | 0.053 | 0.054 |
| | | | (0.017) | (0.017) | (0.010) | (0.010) | (0.006) | (0.006) | (0.002) | (0.002) |
| | | C2 | | | | | −0.101 | −0.103 | 0.042 | 0.042 |
| | | | | | | | (0.006) | (0.006) | (0.003) | (0.003) |
| Dominant | R/R | R1 | 1.628 | 1.566 | 0.036 | 0.039 | −0.441 | −0.433 | 0.270 | 0.261 |
| | | | (0.020) | (0.021) | (0.006) | (0.006) | (0.009) | (0.009) | (0.003) | (0.003) |
| | | R2 | | | | | −0.371 | −0.352 | 0.214 | 0.202 |
| | | | | | | | (0.009) | (0.009) | (0.004) | (0.004) |
| | R/C | R1 | 0.741 | 0.704 | 0.077 | 0.084 | −0.253 | −0.252 | 0.167 | 0.162 |
| | | | (0.022) | (0.022) | (0.009) | (0.010) | (0.008) | (0.007) | (0.003) | (0.003) |
| | | C1 | | | | | −0.142 | −0.133 | 0.077 | 0.071 |
| | | | | | | | (0.010) | (0.010) | (0.004) | (0.004) |
| | C/C | C1 | 0.477 | 0.433 | 0.087 | 0.114 | −0.170 | −0.149 | **0.092** | **0.074** |
| | | | (0.024) | (0.021) | (0.010) | (0.011) | (0.007) | (0.006) | (0.003) | (0.003) |
| | | C2 | | | | | −0.152 | −0.130 | **0.076** | **0.059** |
| | | | | | | | (0.008) | (0.007) | (0.004) | (0.003) |
| Recessive | R/R | R1 | **2.000**[f] | **1.865** | 0.005 | 0.020 | −0.916 | −0.916 | 0.840 | 0.840 |
| | | | (0.000) | (0.024) | (0.005) | (0.010) | (0.000) | (0.000) | (0.000) | (0.000) |
| | | R2 | | | | | **−0.916** | **−0.731** | **0.840** | **0.766** |
| | | | | | | | (0.000) | (0.034) | (0.000) | (0.015) |
| | R/C | R1 | **1.775** | **1.215** | 0.010 | 0.030 | **−0.916** | **−0.684** | 0.840 | 0.749 |
| | | | (0.030) | (0.046) | (0.007) | (0.014) | (0.000) | (0.038) | (0.000) | (0.019) |
| | | C1 | | | | | **−0.618** | **−0.367** | **0.720** | **0.354** |
| | | | | | | | (0.041) | (0.033) | (0.026) | (0.027) |
| | C/C | C1 | **1.490** | **0.595** | **0.000** | **0.050** | **−0.662** | **−0.319** | **0.717** | **0.322** |
| | | | (0.054) | (0.053) | (0.000) | (0.015) | (0.043) | (0.028) | (0.029) | (0.025) |
| | | C2 | | | | | **−0.547** | **−0.292** | **0.664** | **0.244** |
| | | | | | | | (0.037) | (0.033) | (0.022) | (0.027) |

[a]Correct analysis means the genetic model adopted in the analysis is the same as the underlying genetic mode.
[b]False negatives are counts out of 2 and false positives are counts out of 5.
[c]Mode refers to underlying genetic mode of data; Freq refers to frequency of risk haplotype.
[d]R refers to a rare risk haplotype and C refers to a common risk haplotype.
[e]Measures in this table are found by averaging over 1,000 simulated data sets; standard errors (SE) are shown in parentheses.
[f]Bolded measures (pro vs. retro) are statistically significantly different at 0.05 level (i.e. the $\pm 2$SE intervals do NOT overlap).

**TABLE VII. BIC-penalized results of Sim II (two-risk haplotypes) for Distribution 1—additive analysis[a]**

| | | | Model selection results | | | | Parameter estimation results | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | False negatives[b] | | False positives | | Bias | | MSE | |
| Mode[c] | Freq | | Pro | Retro | Pro | Retro | Pro | Retro | Pro | Retro |
| Dominant | R/R | R1[d] | 1.653[e] | 1.660 | 0.028 | 0.029 | −0.448 | −0.452 | 0.268 | 0.265 |
| | | | (0.019) | (0.019) | (0.005) | (0.005) | (0.008) | (0.008) | (0.002) | (0.002) |
| | | R2 | | | | | −0.391 | −0.397 | 0.218 | 0.219 |
| | | | | | | | (0.008) | (0.008) | (0.004) | (0.004) |
| | R/C | R1 | 0.790 | 0.789 | 0.079 | 0.083 | −0.261 | −0.265 | 0.163 | 0.160 |
| | | | (0.024) | (0.024) | (0.009) | (0.010) | (0.007) | (0.007) | (0.004) | (0.004) |
| | | C1 | | | | | −0.216 | −0.233 | 0.094 | 0.098 |
| | | | | | | | (0.01) | (0.009) | (0.004) | (0.004) |
| | C/C | C1 | 0.594 | 0.603 | 0.085 | 0.093 | −0.243 | −0.255 | 0.110 | 0.113 |
| | | | (0.026) | (0.026) | (0.010) | (0.010) | (0.007) | (0.006) | (0.004) | (0.004) |
| | | C2 | | | | | −0.225 | −0.237 | 0.096 | 0.098 |
| | | | | | | | (0.007) | (0.007) | (0.004) | (0.004) |
| Recessive | R/R | R1 | 2.000 | 2.000 | 0.015 | 0.025 | −0.916 | −0.916 | 0.840 | 0.840 |
| | | | (0.000) | (0.000) | (0.009) | (0.011) | (0.000) | (0.000) | (0.000) | (0.000) |
| | | R2 | | | | | −0.916 | −0.916 | 0.840 | 0.840 |
| | | | | | | | (0.000) | (0.000) | (0.000) | (0.000) |
| | R/C | R1 | 1.910 | 1.910 | 0.025 | 0.030 | −0.916 | −0.916 | 0.843 | 0.844 |
| | | | (0.021) | (0.021) | (0.011) | (0.012) | (0.000) | (0.000) | (0.011) | (0.011) |
| | | C1 | | | | | −0.885 | −0.884 | 0.795 | 0.794 |
| | | | | | | | (0.005) | (0.005) | (0.009) | (0.009) |
| | C/C | C1 | 1.875 | 1.865 | 0.015 | 0.025 | −0.895 | −0.894 | 0.809 | 0.807 |
| | | | (0.028) | (0.029) | (0.009) | (0.011) | (0.006) | (0.006) | (0.009) | (0.009) |
| | | C2 | | | | | −0.895 | −0.894 | 0.808 | 0.806 |
| | | | | | | | (0.006) | (0.006) | (0.009) | (0.009) |

[a]Additive analysis means an additive model is adopted in the analysis when the underlying genetic mode is not.
[b]False negatives are counts out of 2 and false positives are counts out of 5.
[c]Mode refers to underlying genetic mode of data; Freq refers to frequency of risk haplotype.
[d]R refers to a rare risk haplotype and C refers to a common risk haplotype.
[e]Measures in this table are found by averaging over 1,000 simulated data sets; standard errors (SE) are shown in parentheses.

of results observed in Simulation I remain the same in Simulation II. First, when the risk haplotypes act additively on disease susceptibility, the performances of aPro and aRetro are comparable for FNs, FPs, bias and MSE. Second, the performance of aRetro is better than that of aPro under a dominant mode with correct analysis. The gain brought by aRetro is similar across all three simulation scenarios but is rarely significantly different from aPro. However, when analyzing the data with an additive model, all measures increase for both methods. The performance loss is more severe in aRetro, resulting in a comparable performance of aPro and aRetro. Finally, under a recessive mode analyzed correctly, aRetro has significantly fewer FNs, slightly more FPs, and significantly smaller bias/MSE when compared to aPro for almost all simulation scenarios. However, when recessively acting haplotypes are analyzed using an additive genetic model, the performance of each procedure suffers, especially aRetro. Both methods essentially lose their power to detect the two-risk haplotypes (e.g. having FN counts close to 2 as no-risk haplotypes were retained in the model) and yield sizable biases/MSE.

## SIMULATION I AND II FOR HAPLOTYPE DISTRIBUTION 2

When comparing results between the two haplotype distributions for Simulations I and II, all measures were typically higher for Distribution 2. Like for the Null Simulation, these results are not unexpected because Distribution 2 has a larger dimension, which means more parameters needed to be estimated in the analysis. Increasing the number of parameters and using the same amount of data for estimation can decrease power (i.e. increase the FN count) and increase bias. Although the magnitudes of model selection and estimation measures differ between the two haplotype distributions, when comparing the relative performance of aPro and aRetro, the pattern of results observed in the first haplotype distribution is similar to that in the second haplotype distribution for both Simulations I and II (Tables VIII–XI). The relative performance of aPro and aRetro depends on both the underlying genetic mode of the risk haplotypes and the genetic model adopted in the analysis. When the haplotypes associated with disease risk act additively, the two procedures perform comparably with respect to model selection and estimation measures. Under a dominant mode, aRetro performs slightly better than aPro and substantially better under a recessive mode. When these data are analyzed using an additive model, both procedures suffer from a loss in power and an increase in bias/MSE. The impact of imposing the incorrect genetic model is more severe for aRetro than for aPro, and the performance gain of aRetro is lost.

**TABLE VIII. BIC-penalized results of Sim I (one-risk haplotype) for Distribution 2—correct analysis[a]**

| | | | Model selection results | | | | Parameter estimation results | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | False Negatives[b] | | False Positives | | Bias | | MSE | |
| Mode[c] | Freq | OR | Pro | Retro | Pro | Retro | Pro | Retro | Pro | Retro |
| Additive | Rare | 1.3 | 0.965[d] | 0.960 | 0.065 | 0.035 | −0.245 | −0.244 | 0.068 | 0.068 |
| | | | (0.006) | (0.006) | (0.017) | (0.013) | (0.006) | (0.006) | (0.000) | (0.000) |
| | | 1.5 | 0.885 | 0.865 | 0.090 | 0.045 | −0.338 | −0.329 | 0.151 | 0.148 |
| | | | (0.010) | (0.011) | (0.021) | (0.015) | (0.014) | (0.014) | (0.003) | (0.003) |
| | | 1.7 | 0.675 | 0.685 | 0.080 | 0.070 | −0.349 | −0.351 | 0.198 | 0.201 |
| | | | (0.015) | (0.015) | (0.02) | (0.021) | (0.020) | (0.020) | (0.009) | (0.009) |
| | | 2.0 | 0.290 | 0.290 | 0.140 | 0.080 | −0.250 | −0.250 | 0.169 | 0.168 |
| | | | (0.014) | (0.014) | (0.027) | (0.022) | (0.023) | (0.023) | (0.015) | (0.015) |
| | Common | 1.3 | 0.715 | 0.710 | 0.085 | 0.075 | −0.229 | −0.221 | 0.067 | 0.067 |
| | | | (0.014) | (0.014) | (0.021) | (0.020) | (0.009) | (0.010) | (0.001) | (0.001) |
| | | 1.5 | **0.385**[e] | **0.315** | 0.050 | 0.050 | −0.311 | −0.286 | 0.130 | 0.121 |
| | | | (0.015) | (0.015) | (0.017) | (0.017) | (0.013) | (0.014) | (0.005) | (0.005) |
| | | 1.7 | 0.125 | 0.120 | 0.065 | 0.080 | **−0.443** | **−0.382** | 0.233 | 0.199 |
| | | | (0.010) | (0.010) | (0.020) | (0.023) | (0.013) | (0.016) | (0.008) | (0.009) |
| | | 2.0 | 0.005 | 0.010 | 0.090 | 0.160 | **−0.646** | **−0.583** | **0.450** | **0.404** |
| | | | (0.002) | (0.003) | (0.025) | (0.034) | (0.013) | (0.018) | (0.009) | (0.012) |
| Dominant | Rare | 1.3 | 0.955 | 0.930 | 0.085 | 0.040 | −0.234 | −0.219 | 0.073 | 0.076 |
| | | | (0.007) | (0.008) | (0.021) | (0.016) | (0.010) | (0.012) | (0.003) | (0.004) |
| | | 1.5 | 0.895 | 0.880 | 0.085 | 0.045 | −0.343 | −0.336 | 0.152 | 0.151 |
| | | | (0.010) | (0.010) | (0.023) | (0.016) | (0.013) | (0.014) | (0.003) | (0.003) |
| | | 1.7 | **0.770** | **0.690** | 0.125 | 0.110 | −0.388 | −0.346 | 0.224 | 0.203 |
| | | | (0.013) | (0.015) | (0.023) | (0.024) | (0.019) | (0.021) | (0.008) | (0.009) |
| | | 2.0 | **0.510** | **0.425** | 0.125 | 0.080 | −0.361 | −0.322 | 0.260 | 0.225 |
| | | | (0.016) | (0.016) | (0.023) | (0.020) | (0.026) | (0.025) | (0.016) | (0.016) |
| | Common | 1.3 | 0.815 | 0.800 | 0.030 | 0.035 | −0.240 | −0.234 | 0.067 | 0.068 |
| | | | (0.012) | (0.013) | (0.012) | (0.013) | (0.007) | (0.008) | (0.001) | (0.001) |
| | | 1.5 | 0.545 | 0.520 | 0.070 | 0.110 | −0.296 | −0.273 | 0.132 | 0.123 |
| | | | (0.016) | (0.016) | (0.019) | (0.025) | (0.015) | (0.016) | (0.004) | (0.005) |
| | | 1.7 | 0.270 | 0.235 | 0.075 | 0.100 | −0.376 | −0.339 | 0.205 | 0.185 |
| | | | (0.014) | (0.013) | (0.019) | (0.025) | (0.018) | (0.019) | (0.008) | (0.009) |
| | | 2.0 | 0.020 | 0.010 | 0.110 | 0.130 | −0.509 | −0.441 | 0.349 | 0.304 |
| | | | (0.004) | (0.003) | (0.025) | (0.026) | (0.021) | (0.024) | (0.014) | (0.016) |
| Recessive | Rare | 2.0 | **1.000** | **0.930** | 0.050 | 0.035 | **−0.693** | **−0.610** | **0.480** | **0.467** |
| | | | (0.000) | (0.008) | (0.015) | (0.013) | (0.000) | (0.022) | (0.000) | (0.005) |
| | | 2.5 | **1.000** | **0.810** | 0.050 | 0.075 | **−0.916** | **−0.660** | **0.840** | **0.728** |
| | | | (0.000) | (0.012) | (0.015) | (0.019) | (0.000) | (0.038) | (0.000) | (0.018) |
| | | 3.0 | **1.000** | **0.810** | 0.070 | 0.025 | **−1.099** | **−0.859** | **1.207** | **0.995** |
| | | | (0.000) | (0.012) | (0.018) | (0.011) | (0.000) | (0.036) | (0.000) | (0.031) |
| | | 3.5 | **1.000** | **0.685** | 0.045 | 0.055 | **−1.253** | **−0.841** | **1.569** | **1.103** |
| | | | (0.000) | (0.015) | (0.015) | (0.018) | (0.000) | (0.045) | (0.000) | (0.049) |
| | Common | 2.0 | **0.935** | **0.820** | **0.025** | **0.060** | **−0.693** | **−0.539** | **0.481** | **0.404** |
| | | | (0.008) | (0.012) | (0.011) | (0.020) | (0.000) | (0.024) | (0.000) | (0.012) |
| | | 2.5 | **0.880** | **0.610** | **0.010** | **0.070** | **−0.856** | **−0.555** | **0.824** | **0.531** |
| | | | (0.010) | (0.015) | (0.007) | (0.019) | (0.021) | (0.034) | (0.007) | (0.028) |
| | | 3.0 | **0.700** | **0.465** | **0.020** | **0.060** | **−0.932** | **−0.548** | **1.087** | **0.590** |
| | | | (0.014) | (0.016) | (0.010) | (0.018) | (0.033) | (0.038) | (0.024) | (0.041) |
| | | 3.5 | **0.585** | **0.405** | **0.000** | **0.100** | **−0.974** | **−0.590** | **1.284** | **0.690** |
| | | | (0.016) | (0.016) | (0.000) | (0.024) | (0.041) | (0.041) | (0.041) | (0.052) |

[a]Correct analysis means the genetic model adopted in the analysis is the same as the underlying genetic mode.
[b]False negatives are counts out of 1 and false positives are counts out of 9.
[c]Mode refers to underlying genetic mode of data; Freq refers to frequency of risk haplotype; OR refers to the odds ratio of the risk haplotype.
[d]Measures in this table are found by averaging over 1,000 simulated data sets; standard errors (SE)are shown in parentheses.
[e]Bolded measures (pro vs. retro) are statistically significantly different at 0.05 level (i.e. the $\pm 2$SE intervals do NOT overlap).

**TABLE IX. BIC-penalized results of Sim I (one-risk haplotype) for Distribution 2—additive analysis[a]**

| | | | Model selection results | | | | Parameter estimation results | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | False negatives[b] | | False positives | | Bias | | MSE | |
| Mode[c] | Freq | OR | Pro | Retro | Pro | Retro | Pro | Retro | Pro | Retro |
| Dominant | Rare | 1.3 | 0.950[d] | 0.955 | 0.095 | 0.045 | −0.232 | −0.235 | 0.073 | 0.072 |
| | | | (0.007) | (0.007) | (0.023) | (0.015) | (0.010) | (0.009) | (0.003) | (0.002) |
| | | 1.5 | 0.900 | 0.900 | 0.070 | 0.045 | −0.351 | −0.351 | 0.151 | 0.152 |
| | | | (0.009) | (0.009) | (0.022) | (0.016) | (0.012) | (0.012) | (0.003) | (0.003) |
| | | 1.7 | 0.795 | 0.790 | 0.130 | 0.075 | −0.406 | −0.408 | 0.229 | 0.227 |
| | | | (0.013) | (0.013) | (0.024) | (0.019) | (0.018) | (0.017) | (0.008) | (0.008) |
| | | 2.0 | 0.505 | 0.550 | 0.100 | 0.070 | −0.390 | −0.414 | 0.263 | 0.281 |
| | | | (0.016) | (0.016) | (0.021) | (0.019) | (0.024) | (0.023) | (0.016) | (0.016) |
| | Common | 1.3 | 0.820 | 0.805 | 0.035 | 0.025 | −0.243 | −0.245 | 0.067 | 0.067 |
| | | | (0.012) | (0.013) | (0.013) | (0.011) | (0.006) | (0.006) | (0.001) | (0.001) |
| | | 1.5 | 0.565 | 0.570 | 0.075 | 0.075 | −0.314 | −0.308 | 0.133 | 0.129 |
| | | | (0.016) | (0.016) | (0.019) | (0.019) | (0.013) | (0.013) | (0.004) | (0.005) |
| | | 1.7 | 0.275 | 0.270 | 0.075 | 0.100 | −0.397 | −0.385 | 0.208 | 0.200 |
| | | | (0.014) | (0.014) | (0.020) | (0.024) | (0.016) | (0.016) | (0.008) | (0.009) |
| | | 2.0 | 0.015 | 0.020 | 0.090 | 0.100 | −0.539 | −0.478 | 0.358 | 0.304 |
| | | | (0.004) | (0.004) | (0.025) | (0.026) | (0.018) | (0.019) | (0.014) | (0.015) |
| Recessive | Rare | 2.0 | 1.000 | 1.000 | 0.075 | 0.025 | −0.693 | −0.693 | 0.480 | 0.480 |
| | | | (0.000) | (0.000) | (0.020) | (0.011) | (0.000) | (0.000) | (0.000) | (0.000) |
| | | 2.5 | 0.995 | 0.995 | 0.105 | 0.055 | −0.912 | −0.912 | 0.835 | 0.835 |
| | | | (0.002) | (0.002) | (0.026) | (0.020) | (0.004) | (0.004) | (0.004) | (0.004) |
| | | 3.0 | 0.995 | 0.990 | **0.080**[e] | **0.010** | −1.096 | −1.094 | 1.203 | 1.199 |
| | | | (0.002) | (0.003) | (0.019) | (0.007) | (0.002) | (0.003) | (0.004) | (0.006) |
| | | 3.5 | 0.985 | 0.985 | 0.075 | 0.035 | −1.244 | −1.242 | 1.552 | 1.551 |
| | | | (0.004) | (0.004) | (0.020) | (0.015) | (0.005) | (0.006) | (0.010) | (0.011) |
| | Common | 2.0 | 0.930 | 0.950 | 0.035 | 0.015 | −0.691 | −0.691 | 0.478 | 0.478 |
| | | | (0.008) | (0.007) | (0.013) | (0.009) | (0.002) | (0.002) | (0.002) | (0.002) |
| | | 2.5 | 0.890 | 0.890 | 0.045 | 0.055 | −0.903 | −0.895 | 0.821 | 0.809 |
| | | | (0.010) | (0.010) | (0.015) | (0.016) | (0.005) | (0.007) | (0.007) | (0.009) |
| | | 3.0 | 0.790 | 0.805 | 0.040 | 0.020 | −1.087 | −1.078 | 1.186 | 1.171 |
| | | | (0.013) | (0.013) | (0.014) | (0.010) | (0.005) | (0.006) | (0.009) | (0.011) |
| | | 3.5 | 0.730 | 0.715 | 0.020 | 0.030 | −1.230 | −1.216 | 1.523 | 1.495 |
| | | | (0.014) | (0.014) | (0.010) | (0.014) | (0.007) | (0.009) | (0.014) | (0.018) |

[a]Additive analysis means an additive model is adopted in the analysis when the underlying genetic mode is not.
[b]False negatives are counts out of 1 and false positives are counts out of 9.
[c]Mode refers to underlying genetic mode of data; Freq refers to frequency of risk haplotype; OR refers to the odds ratio of the risk haplotype.
[d]Measures in this table are found by averaging over 1,000 simulated data sets; standard errors (SE) are shown in parentheses.
[e]Bolded measures (pro vs. retro) are statistically significantly different at 0.05 level (i.e. the $\pm 2$SE intervals do NOT overlap).

# ANALYSIS OF GAW15 RA DATA

Rheumatoid arthritis (RA) is a condition that is believed to be influenced by a number of genetic and environmental factors. Genetic Analysis Workshop (GAW) 15 designed a series of studies to investigate the genetic risk factors of this condition. One of these studies focused on a dense panel of 2,300 SNPs from a 10-megabase region of chromosome 18q. This region of the genome was chosen because it had shown prior evidence of linkage to RA [Browning and Thomas, 2007]. The panel of SNPs was genotyped for 460 cases and 460 controls collected by the North American Rheumatoid Arthritis Consortium. Before implementing the prospective and retrospective analyses discussed in this article, the quality control filters described by Browning and Thomas [2007] were applied to the data, and five SNPs in Hardy-Weinberg disequilibrium and one case with no genotype information were removed from the data set prior to analysis. We analyzed this data using the four methods discussed in this article—unpenalized prospective likelihood, prospective likelihood coupled with the ALASSO penalty, unpenalized retrospective likelihood, and retrospective likelihood coupled with the ALASSO penalty. The unpenalized methods were fit using haplo.glm in R and HAPSTAT in Linux for the prospective and retrospective likelihoods, respectively. The penalized methods were fit using the LSA code posted on the corresponding author's website.

Based on the findings of Browning and Thomas, we analyzed two blocks of linkage disequilibrium (LD) around the region that showed significant association (SNPs 1631 and 1632). The two regions we investigated were (a) SNPs 1626 to 1632, a block of SNPs that are in

**TABLE X. BIC-penalized results of Sim II (two-risk haplotypes) for Distribution 2—correct analysis[a]**

| | | | Model selection results | | | | Parameter estimation results | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | False negatives[b] | | False positives | | Bias | | MSE | |
| Mode[c] | Freq | | Pro | Retro | Pro | Retro | Pro | Retro | Pro | Retro |
| Additive | R/R | R1[d] | 1.040[e] | 1.025 | 0.195 | 0.140 | −0.318 | −0.301 | 0.190 | 0.184 |
| | | | (0.056) | (0.056) | (0.037) | (0.028) | (0.021) | (0.022) | (0.009) | (0.009) |
| | | R2 | | | | | −0.227 | −0.219 | 0.125 | 0.122 |
| | | | | | | | (0.019) | (0.019) | (0.009) | (0.009) |
| | R/C | R1 | 1.030 | 0.985 | 0.120 | 0.120 | −0.470 | −0.422 | 0.257 | 0.234 |
| | | | (0.033) | (0.054) | (0.028) | (0.028) | (0.013) | (0.017) | (0.006) | (0.007) |
| | | C1 | | | | | **−0.417** | **−0.333** | **0.219** | **0.176** |
| | | | | | | | (0.015) | (0.018) | (0.008) | (0.009) |
| | C/C | C1 | 0.960 | 0.970 | 0.095 | 0.130 | −0.467 | −0.425 | 0.251 | 0.228 |
| | | | (0.037) | (0.055) | (0.024) | (0.029) | (0.013) | (0.015) | (0.006) | (0.007) |
| | | C2 | | | | | −0.452 | −0.416 | 0.235 | 0.217 |
| | | | | | | | (0.013) | (0.015) | (0.007) | (0.008) |
| Dominant | R/R | R1 | 1.120 | 0.975 | 0.170 | 0.130 | −0.319 | −0.279 | 0.188 | 0.166 |
| | | | (0.055) | (0.054) | (0.028) | (0.026) | (0.021) | (0.021) | (0.009) | (0.009) |
| | | R2 | | | | | −0.248 | −0.229 | 0.151 | 0.135 |
| | | | | | | | (0.021) | (0.020) | (0.009) | (0.009) |
| | R/C | R1 | **1.290**[f] | **1.060** | 0.165 | 0.210 | −0.414 | −0.337 | 0.240 | 0.197 |
| | | | (0.042) | (0.052) | (0.029) | (0.037) | (0.019) | (0.021) | (0.007) | (0.009) |
| | | C1 | | | | | −0.391 | −0.349 | 0.210 | 0.184 |
| | | | | | | | (0.017) | (0.018) | (0.009) | (0.009) |
| | C/C | C1 | 1.180 | 1.130 | 0.075 | 0.095 | −0.456 | −0.411 | **0.243** | **0.212** |
| | | | (0.046) | (0.056) | (0.020) | (0.023) | (0.013) | (0.015) | (0.007) | (0.008) |
| | | C2 | | | | | −0.434 | −0.379 | **0.229** | **0.194** |
| | | | | | | | (0.014) | (0.016) | (0.008) | (0.009) |
| Recessive | R/R | R1 | **2.000** | **1.610** | 0.030 | 0.050 | **−0.916** | **−0.749** | **0.840** | **0.763** |
| | | | (0.000) | (0.044) | (0.012) | (0.015) | (0.000) | (0.032) | (0.000) | (0.018) |
| | | R2 | | | | | **−0.916** | **−0.642** | **0.840** | **0.642** |
| | | | | | | | (0.000) | (0.034) | (0.000) | (0.025) |
| | R/C | R1 | **1.880** | **1.400** | 0.020 | 0.045 | **−0.916** | **−0.736** | **0.840** | **0.746** |
| | | | (0.023) | (0.045) | (0.010) | (0.015) | (0.000) | (0.032) | (0.000) | (0.017) |
| | | C1 | | | | | **−0.846** | **−0.517** | **0.827** | **0.487** |
| | | | | | | | (0.024) | (0.033) | (0.009) | (0.028) |
| | C/C | C1 | **1.725** | **1.330** | **0.010** | **0.085** | **−0.885** | **−0.632** | **0.854** | **0.593** |
| | | | (0.034) | (0.052) | (0.007) | (0.022) | (0.019) | (0.031) | (0.014) | (0.026) |
| | | C2 | | | | | **−0.874** | **−0.604** | **0.824** | **0.560** |
| | | | | | | | (0.017) | (0.031) | (0.007) | (0.027) |

[a]Correct analysis means the genetic model adopted in the analysis is the same as the underlying genetic mode.
[b]False negatives are counts out of 2 and false positives are counts out of 8.
[c]Mode refers to underlying genetic mode of data; Freq refers to frequency of risk haplotype.
[d]R refers to a rare risk haplotype and C refers to a common risk haplotype.
[e]Measures in this table are found by averaging over 1,000 simulated data sets; standard errors (SE) are shown in parentheses.
[f]Bolded measures (pro vs. retro) are statistically significantly different at 0.05 level (i.e. the ±2SE intervals do NOT overlap).

extremely high LD with SNPs 1631 and 1632 and (b) SNPs 1621 to 1630, a haplotype block [Barrett et al., 2005] that is next to SNPs 1631 and 1632 and located in gene CCBE1. High LD was defined as having a pair-wise $D' \geq 0.99$; $D'$ is a measure of LD that scales the raw difference between the expected and observed haplotype frequencies of a pair of SNPs so that it is between −1 and 1. For each block of SNPs, we considered additive, dominant, and recessive models for haplotype effects. Examining the values of the fitted likelihoods for each method, it appears that the haplotypes in these blocks do not act recessively; so only the results of the additive and dominant analyses are shown. The results of our analysis (coefficient estimates for all methods and unadjusted $p$-values for the unpenalized methods) can be found in Tables XII and XIII.

In their analysis, Browning and Thomas located a significant haplotype sequence "21" at SNPs 1631 and 1632. When analyzing the LD block containing these SNPs (a), all four methods considered in this article were able to find the same signal. This result can be seen when examining the small $p$-value and the non-zero coefficient of the haplotype "1121121" for the unpenalized and penalized methods, respectively (Table XII). All the four methods finding the same result is not surprising and is supported by the results of our simulation studies. Based

**TABLE XI. BIC-penalized results of Sim II (two-risk haplotypes) for Distribution 2—additive analysis[a]**

| | | | Model selection results | | | | Parameter estimation results | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | False negatives[b] | | False positives | | Bias | | MSE | |
| Mode[c] | Freq | | Pro | Retro | Pro | Retro | Pro | Retro | Pro | Retro |
| Dominant | R/R | R1[d] | 1.140[e] | 1.135 | 0.160 | 0.110 | −0.337 | −0.348 | 0.192 | 0.195 |
| | | | (0.055) | (0.056) | (0.028) | (0.025) | (0.020) | (0.019) | (0.009) | (0.009) |
| | | R2 | | | | | −0.286 | −0.289 | 0.153 | 0.148 |
| | | | | | | | (0.019) | (0.018) | (0.009) | (0.009) |
| | R/C | R1 | **1.425**[f] | **1.120** | 0.155 | 0.195 | −0.443 | −0.402 | 0.249 | 0.225 |
| | | | (0.040) | (0.055) | (0.032) | (0.039) | (0.016) | (0.018) | (0.006) | (0.008) |
| | | C1 | | | | | −0.413 | −0.381 | 0.213 | 0.192 |
| | | | | | | | (0.015) | (0.015) | (0.008) | (0.009) |
| | C/C | C1 | 1.135 | 1.140 | 0.085 | 0.125 | −0.469 | −0.458 | 0.246 | 0.238 |
| | | | (0.044) | (0.050) | (0.020) | (0.025) | (0.011) | (0.012) | (0.006) | (0.007) |
| | | C2 | | | | | −0.458 | −0.453 | 0.241 | 0.236 |
| | | | | | | | (0.012) | (0.012) | (0.007) | (0.007) |
| Recessive | R/R | R1 | 1.980 | 1.960 | 0.050 | 0.025 | −0.911 | −0.911 | 0.833 | 0.833 |
| | | | (0.010) | (0.014) | (0.015) | (0.011) | (0.004) | (0.004) | (0.005) | (0.005) |
| | | R2 | | | | | −0.913 | −0.905 | 0.834 | 0.823 |
| | | | | | | | (0.003) | (0.005) | (0.004) | (0.007) |
| | R/C | R1 | 1.890 | 1.910 | 0.045 | 0.025 | −0.913 | −0.913 | 0.836 | 0.836 |
| | | | (0.023) | (0.017) | (0.015) | (0.011) | (0.003) | (0.003) | (0.004) | (0.004) |
| | | C1 | | | | | −0.902 | −0.895 | 0.819 | 0.809 |
| | | | | | | | (0.005) | (0.006) | (0.007) | (0.009) |
| | C/C | C1 | 1.730 | 1.740 | 0.045 | 0.050 | −0.910 | −0.907 | 0.830 | 0.826 |
| | | | (0.032) | (0.020) | (0.016) | (0.017) | (0.003) | (0.004) | (0.005) | (0.006) |
| | | C2 | | | | | −0.906 | −0.901 | 0.824 | 0.818 |
| | | | | | | | (0.005) | (0.006) | (0.007) | (0.008) |

[a]Additive analysis means an additive model is adopted in the analysis when the underlying genetic mode is not.
[b]False negatives are counts out of 2 and false positives are counts out of 8.
[c]Mode refers to underlying genetic mode of data; Freq refers to frequency of risk haplotype.
[d]R refers to a rare risk haplotype and C refers to a common risk haplotype.
[e]Measures in this table are found by averaging over 1,000 simulated datas ets; standard errors (SE) are shown in parentheses.
[f]Bolded measures (pro vs. retro) are statistically significantly different at 0.05 level (i.e. the $\pm 2$SE intervals do NOT overlap).

**TABLE XII. Analysis of GAW15-RA data SNPs 1626 to 1632[a]**

| | | Prospective analysis | | | Retrospective analysis | | |
|---|---|---|---|---|---|---|---|
| Model[b] | Haplotype | Unpenalized ($p$-value) | BIC-penalized | AIC-penalized | Unpenalized ($p$-value) | BIC-penalized | AIC-penalized |
| Additive | 1121111 | 0.016 | 0.000 | 0.000 | 0.028 | 0.000 | 0.000 |
| | | (0.883) | | | (0.784) | | |
| | 1121112 | 0.069 | 0.000 | 0.000 | 0.076 | 0.000 | 0.000 |
| | | (0.883) | | | (0.871) | | |
| | 1121121 | −2.028 | −2.023 | −2.023 | −1.979 | −1.978 | −1.978 |
| | | (0.000) | | | (0.000) | | |
| | 2212222 | −0.016 | 0.000 | 0.000 | −0.016 | 0.000 | 0.000 |
| | | (0.919) | | | (0.922) | | |
| Dominant | 1121111 | −0.039 | 0.000 | 0.000 | 0.126 | 0.000 | 0.000 |
| | | (0.797) | | | (0.335) | | |
| | 1121112 | 0.043 | 0.000 | 0.000 | 0.110 | 0.000 | 0.000 |
| | | (0.927) | | | (0.814) | | |
| | 1121121 | −2.022 | −2.018 | −2.018 | −1.967 | −1.861 | −1.861 |
| | | (0.000) | | | (0.000) | | |
| | 2212222 | −0.026 | 0.000 | 0.000 | 0.018 | 0.000 | 0.000 |
| | | (0.880) | | | (0.914) | | |

[a]SNPs in high LD (i.e. $D' \geq 0.99$) with SNPs 1631 and 1632, a significant region identified by Browning and Thomas [2007].
[b]Model refers to genetic model adopted in the analysis.

**TABLE XIII. Analysis of GAW15-RA data SNPs 1621 to 1630[a]**

| | | Prospective analysis | | | Retrospective analysis | | |
|---|---|---|---|---|---|---|---|
| Model[b] | Haplotype | Unpenalized (*p*-value) | BIC-penalized | AIC-penalized | Unpenalized (*p*-value) | BIC-penalized | AIC-penalized |
| Additive | 1221211211 | 0.175 (0.096) | 0.000 | 0.120 | 0.173 (0.098) | 0.000 | 0.152 |
| | 2112122122 | 0.098 (0.539) | 0.000 | 0.000 | 0.070 (0.656) | 0.000 | 0.000 |
| | 2211211211 | −0.368 (0.355) | 0.000 | 0.000 | −0.451 (0.224) | 0.000 | −0.407 |
| | 2221211211 | −0.762 (0.100) | 0.000 | −0.522 | −0.767 (0.094) | 0.000 | −0.729 |
| Dominant | 1221211211 | 0.283 (0.042) | 0.000 | 0.225 | 0.246 (0.052) | 0.000 | 0.226 |
| | 2112122122 | 0.128 (0.461) | 0.000 | 0.000 | 0.072 (0.660) | 0.000 | 0.000 |
| | 2211211211 | −0.356 (0.371) | 0.000 | 0.000 | −0.446 (0.231) | 0.000 | −0.398 |
| | 2221211211 | −0.739 (0.111) | 0.000 | −0.506 | −0.768 (0.094) | 0.000 | −0.727 |

[a]A haplotype block that is next to SNPs 1631 and 1632 and is in gene CCBE1.
[b]Model refers to genetic model adopted in the analysis.

on our simulations, we found that the difference between a prospective and retrospective approach is often negligible when the haplotype effect size is large under additive and dominant genetic modes.

Browning and Thomas also found that individuals with the haplotype ''21'' at SNPs 1631 and 1632 also share the sequence ''11211'' at SNPs 1626 to 1630. When analyzing the LD block (a) defined by SNPs 1626 to 1632, we replicated the exact haplotype using the article's four methods. Motivated by this finding, we analyzed the second LD block (b), which is a 10-SNP haplotype block covering SNPs 1626 to 1630 and is located in gene CCBE1. The unpenalized methods (at significance level 0.01) and penalized methods tuned with BIC did not find any significant signals. However, in practice, the significance level is often set more liberally at 0.05; at these levels, one haplotype (''1221211211'') sits at the border of significance. More interestingly, the penalized method tuned with AIC identified some of the haplotypes in the sample with the shared sequence ''11211'' under a prospective analysis and all of them under a retrospective analysis. Although it is unknown whether these findings are true signals or false positives, it is promising that penalized methods are able to detect other haplotypes containing the sequence ''11211'', and that a retrospective penalized approach is able to find all haplotypes with the sequence ''11211'' significant.

# DISCUSSION

Like other haplotype-based methods developed to assess haplotype-phenotype association, the success of penalized regression methods depends on the underlying data likelihood. For unpenalized methods, prospective analyses are valid but less efficient than their retrospective counterparts for hypothesis testing, and can result in substantial bias when estimating the haplotype effects. Based on our simulation studies, the same can be said for penalized methods, which combine testing and estimation into one procedure. We found that the impact of using a prospective likelihood in the analysis depends on the underlying genetic mode of the associated genetic variant and the genetic model adopted in the analysis. When the genetic mode of the haplotypes is known and the correct inheritance model is imposed, using a prospective analysis in place of the more appropriate retrospective analysis is detrimental when the associated haplotypes act dominantly or recessively with respect to disease risk. These results agree with the findings for non-penalized likelihood methods [Satten and Epstein, 2004]. Because the genetic mode of a genetic variant is usually unknown, researchers often analyze the data additively. When dominant or recessive data are analyzed under an additive genetic model, the performance of the prospective and retrospective analyses become comparable. Both methods suffer from an increase in FNs and bias for using an incorrect genetic model. The retrospective analysis appears to be more sensitive to model misspecification and exhibits a larger degree of performance loss, thus making its performance gain over the prospective analysis negligible or slight.

While our simulations focused on penalized methods using the ALASSO penalty with the prospective likelihood of haplo.glm and the retrospective likelihood of HAPSTAT, we hope our findings can provide insight when coupling other penalized approaches with a prospective or a retrospective likelihood for case-control studies. If the main consideration is the relative performance of the retrospective versus prospective penalized method, then our results suggest that the negative impact of developing haplotype-based penalized methods based on a prospective likelihood for case-control data is non-trivial only when the risk haplotypes act non-additively and the correct genetic model is adopted in the analysis. However, we think a more appropriate way to summarize our findings is to note that a careful haplotype-based penalized analysis of case-control data requires the use

**TABLE XIV. Unpenalized results of Sim I (one-risk haplotype) for Distribution 1—correct analysis[a]**

| | | | Model selection results | | | | Parameter estimation results | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | False negatives[b] | | False positives | | Bias | | MSE | |
| Mode[c] | Freq | OR | Pro | Retro | Pro | Retro | Pro | Retro | Pro | Retro |
| Additive | Rare | 1.3 | 0.934[d] | 0.934 | 0.030 | 0.028 | 0.013 | 0.010 | 0.049 | 0.048 |
| | | | (0.008) | (0.008) | (0.006) | (0.005) | (0.007) | (0.007) | (0.002) | (0.002) |
| | | 1.5 | **0.865**[e] | **0.785** | 0.050 | 0.050 | 0.012 | 0.005 | 0.050 | 0.048 |
| | | | (0.012) | (0.011) | (0.008) | (0.007) | (0.007) | (0.008) | (0.001) | (0.002) |
| | | 1.7 | 0.677 | 0.680 | 0.040 | 0.042 | 0.005 | −0.003 | 0.047 | 0.046 |
| | | | (0.016) | (0.016) | (0.007) | (0.007) | (0.007) | (0.007) | (0.002) | (0.002) |
| | | 2.0 | 0.249 | 0.253 | 0.032 | 0.028 | 0.021 | 0.013 | 0.045 | 0.044 |
| | | | (0.013) | (0.013) | (0.006) | (0.005) | (0.007) | (0.007) | (0.002) | (0.002) |
| | Common | 1.3 | 0.823 | 0.820 | 0.026 | 0.024 | 0.011 | 0.009 | 0.024 | 0.023 |
| | | | (0.012) | (0.012) | (0.005) | (0.005) | (0.005) | (0.005) | (0.001) | (0.001) |
| | | 1.5 | 0.575 | 0.550 | 0.035 | 0.035 | −0.004 | −0.009 | 0.022 | 0.022 |
| | | | (0.01) | (0.011) | (0.006) | (0.005) | (0.005) | (0.005) | (0.002) | (0.001) |
| | | 1.7 | 0.168 | 0.167 | 0.033 | 0.032 | 0.009 | 0.003 | 0.023 | 0.022 |
| | | | (0.012) | (0.012) | (0.006) | (0.006 | (0.005) | (0.005) | (0.001) | (0.001) |
| | | 2.0 | 0.012 | 0.011 | 0.028 | 0.029 | 0.001 | −0.009 | 0.021 | 0.020 |
| | | | (0.003) | (0.003) | (0.005) | (0.005) | (0.005) | (0.005) | (0.001) | (0.001) |
| Dominant | Rare | 1.3 | 0.938 | 0.930 | 0.034 | 0.030 | 0.011 | 0.007 | 0.062 | 0.058 |
| | | | (0.008) | (0.008) | (0.006) | (0.005) | (0.008) | (0.008) | (0.003) | (0.003) |
| | | 1.5 | 0.850 | 0.820 | 0.020 | 0.025 | 0.024 | 0.008 | 0.054 | 0.048 |
| | | | (0.011) | (0.012) | (0.006) | (0.007) | (0.007) | (0.008) | (0.003) | (0.002) |
| | | 1.7 | 0.678 | 0.656 | 0.028 | 0.025 | 0.016 | 0.003 | 0.052 | 0.047 |
| | | | (0.015) | (0.015) | (0.005) | (0.005) | (0.007) | (0.007) | (0.002) | (0.002) |
| | | 2.0 | 0.332 | 0.292 | 0.036 | 0.035 | 0.013 | −0.005 | 0.053 | 0.048 |
| | | | (0.015) | (0.014) | (0.006) | (0.006) | (0.007) | (0.007) | (0.003) | (0.002) |
| | Common | 1.3 | 0.888 | 0.857 | 0.026 | 0.025 | −0.004 | −0.01 | 0.032 | 0.028 |
| | | | (0.010) | (0.011) | (0.005) | (0.005) | (0.006) | (0.005) | (0.001) | (0.001) |
| | | 1.5 | **0.670** | **0.600** | 0.045 | 0.030 | −0.011 | −0.022 | **0.028** | **0.023** |
| | | | (0.011) | (0.011) | (0.005) | (0.006) | (0.007) | (0.006) | (0.001) | (0.001) |
| | | 1.7 | **0.328** | **0.262** | 0.024 | 0.023 | 0.012 | −0.002 | **0.031** | **0.026** |
| | | | (0.015) | (0.014) | (0.005) | (0.005) | (0.006) | (0.005) | (0.001) | (0.001) |
| | | 2.0 | 0.080 | 0.052 | 0.036 | 0.040 | 0.016 | −0.006 | 0.031 | 0.025 |
| | | | (0.009) | (0.007) | (0.007) | (0.007) | (0.006) | (0.005) | (0.002) | (0.001) |
| Recessive | Rare | 2.0 | **1.000** | **0.920** | 0.010 | 0.020 | −0.256 | −0.328 | 0.686 | 0.537 |
| | | | (0.000) | (0.019) | (0.007) | (0.010) | (0.056) | (0.106) | (0.066) | (1.128) |
| | | 2.5 | **1.000** | **0.865** | 0.010 | 0.005 | −0.270 | −0.333 | 0.622 | 0.868 |
| | | | (0.000) | (0.024) | (0.007) | (0.005) | (0.053) | (0.094) | (0.062) | (1.019) |
| | | 3.0 | **1.000** | **0.800** | 0.010 | 0.015 | −0.240 | −0.218 | 0.641 | 0.457 |
| | | | (0.000) | (0.030) | (0.007) | (0.009) | (0.054) | (0.045) | (0.071) | (0.049) |
| | | 3.5 | **1.000** | **0.730** | 0.010 | 0.010 | −0.386 | −0.271 | **0.711** | **0.404** |
| | | | (0.000) | (0.032) | (0.007) | (0.007) | (0.053) | (0.041) | (0.075) | (0.046) |
| | Common | 2.0 | **0.945** | **0.780** | **0.000** | **0.025** | 0.016 | −0.080 | **0.165** | **0.092** |
| | | | (0.016) | (0.033) | (0.000) | (0.011) | (0.029) | (0.021) | (0.017) | (0.010) |
| | | 2.5 | **0.755** | **0.395** | 0.000 | 0.015 | **0.096** | **−0.086** | **0.281** | **0.095** |
| | | | (0.032) | (0.034) | (0.000) | (0.009) | (0.037) | (0.021) | (0.033) | (0.010) |
| | | 3.0 | **0.485** | **0.105** | 0.010 | 0.015 | **0.083** | **−0.076** | **0.163** | **0.065** |
| | | | (0.035) | (0.022) | (0.007) | (0.009) | (0.028) | (0.017) | (0.015) | (0.008) |
| | | 3.5 | **0.340** | **0.055** | 0.015 | 0.010 | **0.061** | **−0.122** | **0.229** | **0.073** |
| | | | (0.034) | (0.016) | (0.009) | (0.007) | (0.034) | (0.017) | (0.034) | (0.008) |

[a]Correct analysis means the genetic model adopted in the analysis is the same as the underlying genetic mode.
[b]False negatives are counts out of 1 and false positives are counts out of 6.
[c]Mode refers to underlying genetic mode of data; Freq refers to frequency of risk haplotype; OR refers to the odds ratio of the risk haplotype.
[d]Measures in this table are found by averaging over 1,000 simulated data sets; standard errors (SE) are shown in parentheses.
[e]Bolded measures (pro vs. retro) are statistically significantly different at 0.05 level (i.e. the $\pm 2$SE intervals do NOT overlap).

of a retrospective likelihood and the correct genetic mode. In practice, a major concern about using retrospective likelihoods is that they are difficult to implement. When using penalized likelihood methods, optimizing the retrospective likelihood can become even more intractable when the penalty term is incorporated. Use of the exact likelihood coupled with the EM algorithm due to the unknown phase, with the penalization, and tuning on a dense grid of tuning parameters, requires a computationally intensive iterative procedure.

To overcome the computational burden, the least squares objective function provides a promising alternative for implementing penalized retrospective methods. By using the least squares objective function, the need to directly optimize the penalized retrospective likelihood is bypassed. Instead, the unpenalized likelihood is optimized once for the centering value in the objective function; hence, implementing prospective and retrospective penalized methods have similar computational costs. The spared computational efforts can be put into exploring and identifying the correct genetic mode for potential risk haplotypes.

Penalized likelihood methods can have higher power than unpenalized methods in detecting important haplotypic factors [Guo and Lin, 2009]. Our simulations agree with their findings and reveal that the methods considered here can have better power to identify risk haplotypes (i.e. fewer FNs) than the unpenalized version (e.g. Tables XIV and XV vs. Tables IV and V, respectively). However, the power enjoyed by penalized likelihood methods comes at the expense of obtaining effect estimates with relatively larger bias than their unpenalized likelihood counterparts. As observed in our simulations, the bias on the effect estimates obtained by the penalized method can remain

**TABLE XV. Unpenalized results of Sim I (one-risk haplotype) for Distribution 1—additive analysis[a]**

| | | | Model selection results | | | | Parameter estimation results | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | False negatives[b] | | False positives | | Bias | | MSE | |
| Mode[c] | Freq | OR | Pro | Retro | Pro | Retro | Pro | Retro | Pro | Retro |
| Dominant | Rare | 1.3 | 0.954[d] | 0.953 | 0.036 | 0.035 | −0.008 | −0.014 | 0.055 | 0.052 |
| | | | (0.007) | (0.007) | (0.006) | (0.006) | (0.007) | (0.007) | (0.002) | (0.002) |
| | | 1.5 | 0.850 | 0.855 | 0.025 | 0.030 | −0.013 | −0.023 | 0.048 | 0.046 |
| | | | (0.010) | (0.009) | (0.006) | (0.006) | (0.008) | (0.007) | (0.002) | (0.003) |
| | | 1.7 | 0.700 | 0.687 | 0.029 | 0.032 | −0.019 | −0.036 | 0.048 | 0.046 |
| | | | (0.015) | (0.015) | (0.005) | (0.006) | (0.007) | (0.007) | (0.002) | (0.002) |
| | | 2.0 | 0.367 | 0.361 | 0.031 | 0.033 | −0.032 | −0.059 | 0.049 | 0.047 |
| | | | (0.015) | (0.015) | (0.006) | (0.006) | (0.007) | (0.007) | (0.002) | (0.002) |
| | Common | 1.3 | 0.904 | 0.908 | 0.022 | 0.022 | −0.047 | −0.052 | 0.028 | 0.027 |
| | | | (0.009) | (0.009) | (0.005) | (0.005) | (0.005) | (0.005) | (0.001) | (0.001) |
| | | 1.5 | 0.735 | 0.750 | 0.035 | 0.030 | −0.074 | −0.088 | 0.028 | 0.028 |
| | | | (0.010) | (0.009) | (0.005) | (0.006) | (0.006) | (0.005) | (0.001) | (0.002) |
| | | 1.7 | 0.386 | 0.414 | 0.026 | 0.031 | **−0.076**[e] | **−0.098** | 0.031 | 0.032 |
| | | | (0.015) | (0.016) | (0.006) | (0.006) | (0.005) | (0.005) | (0.001) | (0.001) |
| | | 2.0 | 0.120 | 0.135 | 0.038 | 0.037 | **−0.096** | **−0.134** | 0.035 | 0.040 |
| | | | (0.010) | (0.011) | (0.007) | (0.006) | (0.005) | (0.005) | (0.002) | (0.002) |
| Recessive | Rare | 2.0 | 0.990 | 0.990 | 0.040 | 0.050 | −0.671 | −0.665 | 0.497 | 0.491 |
| | | | (0.007) | (0.007) | (0.014) | (0.015) | (0.015) | (0.016) | (0.021) | (0.021) |
| | | 2.5 | 0.995 | 0.990 | 0.020 | 0.030 | −0.871 | −0.865 | 0.810 | 0.805 |
| | | | (0.005) | (0.007) | (0.010) | (0.012) | (0.016) | (0.017) | (0.028) | (0.028) |
| | | 3.0 | 0.970 | 0.965 | 0.040 | 0.040 | −0.995 | −0.989 | 1.045 | 1.039 |
| | | | (0.012) | (0.013) | (0.014) | (0.014) | (0.017) | (0.018) | (0.034) | (0.036) |
| | | 3.5 | 1.000 | 0.985 | 0.010 | 0.015 | −1.148 | −1.14 | 1.357 | 1.345 |
| | | | (0.000) | −0.009 | −0.007 | −0.011 | −0.014 | −0.015 | −0.032 | −0.035 |
| | Common | 2.0 | 0.985 | 0.980 | 0.010 | 0.030 | −0.560 | −0.557 | 0.335 | 0.332 |
| | | | (0.009) | (0.010) | (0.007) | (0.012) | (0.010) | (0.011) | (0.012) | (0.013) |
| | | 2.5 | 0.910 | 0.890 | 0.015 | 0.025 | −0.722 | −0.708 | 0.542 | 0.526 |
| | | | (0.020) | (0.022) | (0.009) | (0.011) | (0.010) | (0.011) | (0.015) | (0.016) |
| | | 3.0 | 0.865 | 0.835 | 0.025 | 0.040 | −0.867 | −0.845 | 0.771 | 0.739 |
| | | | (0.024) | (0.026) | (0.011) | (0.014) | (0.010) | (0.011) | (0.018) | (0.020) |
| | | 3.5 | 0.775 | 0.725 | 0.025 | 0.025 | −0.979 | −0.947 | 0.978 | 0.920 |
| | | | (0.030) | (0.032) | (0.011) | (0.013) | (0.010) | (0.011) | (0.020) | (0.021) |

[a]Additive analysis means an additive model is adopted in the analysis when the underlying genetic mode is not.
[b]False negatives are counts out of 1 and false positives are counts out of 6.
[c]Mode refers to underlying genetic mode of data; Freq refers to frequency of risk haplotype; OR refers to the odds ratio of the risk haplotype.
[d]Measures in this table are found by averaging over 1,000 simulated data sets; standard errors (SE) are shown in parentheses.
[e]Bolded measures (pro vs. retro) are statistically significantly different at 0.05 level (i.e. the $\pm 2$SE intervals do NOT overlap).

sizeable, particularly when there is very little power to detect the effect. Part of this is due to the fact that non-detection of the effect implies that the estimated coefficient is zero. In contrast, for the MLE, the estimate is still non-zero even though it is not significantly different from zero.

**TABLE XVI. AIC-penalized results of Sim I (one-risk haplotype) for Distribution 1—correct analysis[a]**

| | | | Model selection results | | | | Parameter estimation results | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | False negatives[b] | | False positives | | Bias | | MSE | |
| Mode[c] | Freq | OR | Pro | Retro | Pro | Retro | Pro | Retro | Pro | Retro |
| Additive | Rare | 1.3 | 0.605[d] | 0.600 | 1.010 | 1.090 | −0.110 | −0.106 | 0.059 | 0.058 |
| | | | (0.015) | (0.015) | (0.089) | (0.094) | (0.015) | (0.015) | (0.003) | (0.003) |
| | | 1.5 | 0.320 | 0.330 | 1.095 | 1.145 | −0.106 | −0.109 | 0.074 | 0.074 |
| | | | (0.015) | (0.015) | (0.088) | (0.089) | (0.018) | (0.018) | (0.005) | (0.005) |
| | | 1.7 | 0.100 | 0.105 | 0.975 | 0.990 | −0.079 | −0.082 | 0.060 | 0.060 |
| | | | (0.009) | (0.010) | (0.078) | (0.077) | (0.016) | (0.016) | (0.006) | (0.006) |
| | | 2.0 | 0.025 | 0.025 | 1.000 | 1.000 | −0.070 | −0.067 | 0.055 | 0.055 |
| | | | (0.005) | (0.005) | (0.082) | (0.081) | (0.016) | (0.016) | (0.006) | (0.006) |
| | Common | 1.3 | 0.350 | 0.325 | 1.010 | 1.050 | −0.077 | −0.073 | 0.032 | 0.030 |
| | | | (0.015) | (0.015) | (0.082) | (0.082) | (0.011) | (0.011) | (0.002) | (0.002) |
| | | 1.5 | 0.080 | 0.070 | 1.125 | 1.130 | −0.060 | −0.062 | 0.030 | 0.030 |
| | | | (0.009) | (0.008) | (0.085) | (0.082) | (0.011) | (0.011) | (0.003) | (0.003) |
| | | 1.7 | 0.015 | 0.005 | 1.060 | 1.035 | −0.016 | −0.018 | 0.022 | 0.020 |
| | | | (0.004) | (0.002) | (0.084) | (0.083) | (0.010) | (0.010) | (0.003) | (0.002) |
| | | 2.0 | 0.000 | 0.000 | 0.910 | 0.900 | −0.041 | −0.046 | 0.022 | 0.021 |
| | | | (0.000) | (0.000) | −0.077 | −0.076 | −0.010 | −0.010 | −0.002 | −0.002 |
| Dominant | Rare | 1.3 | 0.630 | 0.625 | 0.885 | 0.940 | −0.116 | −0.117 | 0.057 | 0.057 |
| | | | (0.015) | (0.015) | (0.082) | (0.086) | (0.015) | (0.015) | (0.003) | (0.003) |
| | | 1.5 | 0.355 | 0.345 | 0.855 | 0.855 | −0.128 | −0.122 | 0.078 | 0.076 |
| | | | (0.015) | (0.015) | (0.080) | (0.079) | (0.018) | (0.018) | (0.006) | (0.005) |
| | | 1.7 | 0.175 | 0.170 | 0.905 | 0.945 | −0.102 | −0.109 | 0.088 | 0.083 |
| | | | (0.012) | (0.012) | (0.072) | (0.079) | (0.020) | (0.019) | (0.008) | (0.007) |
| | | 2.0 | 0.025 | 0.020 | 1.080 | 1.075 | −0.070 | −0.070 | 0.064 | 0.058 |
| | | | (0.005) | (0.004) | (0.083) | (0.086) | (0.017) | (0.016) | (0.007) | (0.006) |
| | Common | 1.3 | 0.435 | 0.425 | 0.950 | 0.875 | −0.087 | −0.097 | 0.038 | 0.036 |
| | | | (0.016) | (0.016) | (0.078) | (0.075) | (0.012) | (0.012) | (0.002) | (0.002) |
| | | 1.5 | 0.105 | 0.115 | 1.045 | 1.065 | −0.085 | −0.089 | 0.040 | 0.038 |
| | | | (0.010) | (0.010) | (0.080) | (0.081) | (0.013) | (0.012) | (0.004) | (0.004) |
| | | 1.7 | 0.015 | 0.015 | 1.010 | 0.940 | −0.033 | −0.037 | 0.034 | 0.029 |
| | | | (0.004) | (0.004) | (0.084) | (0.079) | (0.013) | (0.012) | (0.004) | (0.003) |
| | | 2.0 | 0.000 | 0.000 | 0.935 | 0.915 | −0.018 | −0.027 | 0.026 | 0.024 |
| | | | (0.000) | (0.000) | (0.083) | (0.088) | (0.011) | (0.011) | (0.002) | (0.002) |
| Recessive | Rare | 2.0 | **0.900**[e] | **0.570** | 0.450 | 0.545 | **−0.563** | **−0.233** | **0.506** | **0.383** |
| | | | (0.009) | (0.016) | (0.049) | (0.057) | (0.031) | (0.041) | (0.015) | (0.017) |
| | | 2.5 | **0.835** | **0.505** | 0.515 | 0.565 | **−0.717** | **−0.386** | **0.771** | **0.507** |
| | | | (0.012) | (0.016) | (0.049) | (0.052) | (0.036) | (0.042) | (0.031) | (0.027) |
| | | 3.0 | **0.765** | **0.360** | 0.500 | 0.480 | **−0.779** | **−0.381** | **0.965** | **0.511** |
| | | | (0.013) | (0.015) | (0.052) | (0.048) | (0.042) | (0.043) | (0.032) | (0.038) |
| | | 3.5 | **0.670** | **0.330** | 0.500 | 0.565 | **−0.804** | **−0.403** | **1.101** | **0.617** |
| | | | (0.015) | (0.015) | (0.050) | (0.056) | (0.048) | (0.048) | (0.048) | (0.049) |
| | Common | 2.0 | **0.435** | **0.270** | 0.380 | 0.450 | −0.202 | −0.176 | **0.290** | **0.162** |
| | | | (0.016) | (0.014) | (0.043) | (0.047) | (0.035) | (0.026) | (0.019) | (0.014) |
| | | 2.5 | **0.180** | **0.060** | 0.425 | 0.430 | −0.130 | −0.121 | **0.259** | **0.117** |
| | | | (0.012) | (0.008) | (0.047) | (0.048) | (0.035) | (0.023) | (0.023) | (0.014) |
| | | 3.0 | **0.095** | **0.025** | 0.505 | 0.435 | −0.135 | −0.164 | **0.273** | **0.107** |
| | | | (0.009) | (0.005) | (0.051) | (0.047) | (0.036) | (0.020) | (0.026) | (0.015) |
| | | 3.5 | **0.020** | **0.000** | 0.375 | 0.400 | −0.056 | −0.101 | **0.193** | **0.072** |
| | | | (0.004) | (0.000) | −0.041 | −0.041 | −0.031 | −0.018 | −0.024 | −0.008 |

[a]Correct analysis means the genetic model adopted in the analysis is the same as the underlying genetic mode.
[b]False negatives are counts out of 1 and false positives are counts out of 6.
[c]Mode refers to underlying genetic mode of data; Freq refers to frequency of risk haplotype; OR refers to the odds ratio of the risk haplotype.
[d]Measures in this table are found by averaging over 1,000 simulated data sets; standard errors (SE) are shown in parentheses.
[e]Bolded measures (pro vs. retro) are statistically significantly different at 0.05 level (i.e. the $\pm 2$SE intervals do NOT overlap).

The power and bias is also affected by the model selection criterion used to select the tuning parameter. In our analysis, we used BIC to choose the final model because it can achieve consistent model selection [Shao, 1997; Yang, 2005], and although estimation accuracy was of interest, our primary goal was to identify the true model structure. To achieve selection consistency, BIC penalizes degrees of freedom more heavily, which can place a larger amount of shrinkage on the effect estimates and increase their bias. Alternatively, AIC could be used to select the final model. (See Tables XVI and XVII for the AIC-based results for Tables IV and V, respectively.) This selection criterion targets prediction error rather than finding the true model structure [Shao, 1997; Yang, 2005] and imposes a smaller penalty on the degrees of freedom. As a result, models selected via AIC will incur less shrinkage on the effect estimates, which can decrease their bias, but also increases the chance of including non-important predictors in the final model. In our analyses, when tuning with AIC instead of BIC, the FN count stays relatively similar, while the FP count often increases and the bias/MSE decreases. This is a direct result of placing less shrinkage/penalty on the final estimates.

Throughout this paper, we assume that the sample size is larger than the number of predictors ($n > p$), which is consistent with typical case-control data sets. However, if the investigator wishes to consider the situation in which the number of candidate haplotypes exceeds the sample size ($n < p$), the analysis can be extended by replacing the initial estimates $\tilde{\beta}_j$ in the adaptive LASSO penalty with an $L_2$ penalized solution, i.e. ridge regression. The initial ridge step would then need to be tuned; we suggest the use of AIC, so that less shrinkage occurs in the initial estimate. However, the use of the least squares objective function

**TABLE XVII. AIC-penalized results of Sim I (one-risk haplotype) for Distribution 1—additive analysis[a]**

| | | | Model selection results | | | | Parameter estimation results | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | False negatives[b] | | False positives | | Bias | | MSE | |
| Mode[c] | Freq | OR | Pro | Retro | Pro | Retro | Pro | Retro | Pro | Retro |
| Dominant | Rare | 1.3 | 0.640[d] | 0.635 | 0.870 | 0.860 | −0.128 | −0.132 | 0.055 | 0.053 |
| | | | (0.015) | (0.015) | (0.078) | (0.079) | (0.014) | (0.013) | (0.003) | (0.003) |
| | | 1.5 | 0.385 | 0.395 | 0.815 | 0.840 | −0.145 | −0.154 | 0.081 | 0.081 |
| | | | (0.015) | (0.015) | (0.081) | (0.077) | (0.017) | (0.017) | (0.005) | (0.005) |
| | | 1.7 | 0.215 | 0.210 | 0.895 | 0.920 | −0.141 | −0.155 | 0.092 | 0.092 |
| | | | (0.013) | (0.013) | (0.069) | (0.072) | (0.019) | (0.019) | (0.008) | (0.008) |
| | | 2 | 0.030 | 0.035 | 1.115 | 1.095 | −0.106 | −0.132 | 0.065 | 0.067 |
| | | | (0.005) | (0.006) | (0.082) | (0.082) | (0.017) | (0.016) | (0.007) | (0.007) |
| | Common | 1.3 | 0.460 | 0.470 | 0.840 | 0.815 | −0.121 | −0.128 | 0.037 | 0.038 |
| | | | (0.016) | (0.016) | (0.072) | (0.071) | (0.011) | (0.010) | (0.002) | (0.002) |
| | | 1.5 | 0.170 | 0.170 | 1.120 | 1.100 | −0.142 | −0.153 | 0.049 | 0.049 |
| | | | (0.012) | (0.012) | (0.084) | (0.085) | (0.012) | (0.011) | (0.004) | (0.004) |
| | | 1.7 | 0.025 | 0.025 | 0.980 | 1.040 | −0.116 | −0.134 | 0.039 | 0.041 |
| | | | (0.005) | (0.005) | (0.081) | (0.088) | (0.011) | (0.011) | (0.004) | (0.004) |
| | | 2 | **0.005**[e] | **0.000** | 0.885 | 1.015 | −0.118 | −0.153 | 0.037 | 0.043 |
| | | | (0.002) | (0.000) | (0.083) | (0.092) | (0.011) | (0.010) | (0.004) | (0.004) |
| Recessive | Rare | 2 | 0.795 | 0.800 | **0.405** | **0.775** | −0.653 | −0.657 | 0.449 | 0.453 |
| | | | (0.013) | (0.013) | (0.047) | (0.072) | (0.011) | (0.010) | (0.013) | (0.013) |
| | | 2.5 | 0.845 | 0.840 | **0.455** | **0.765** | −0.883 | −0.887 | 0.796 | 0.803 |
| | | | (0.011) | (0.012) | (0.047) | (0.069) | (0.009) | (0.009) | (0.015) | (0.015) |
| | | 3 | 0.805 | 0.780 | **0.555** | **0.930** | −1.047 | −1.041 | 1.121 | 1.108 |
| | | | (0.013) | (0.013) | (0.056) | (0.085) | (0.011) | (0.011) | (0.022) | (0.021) |
| | | 3.5 | 0.780 | 0.785 | **0.455** | **0.815** | −1.198 | −1.195 | 1.462 | 1.456 |
| | | | (0.013) | (0.013) | (0.050) | (0.080) | (0.012) | (0.012) | (0.026) | (0.026) |
| | Common | 2 | 0.695 | 0.700 | **0.480** | **0.900** | −0.623 | −0.622 | 0.405 | 0.405 |
| | | | (0.015) | (0.014) | (0.050) | (0.078) | (0.009) | (0.009) | (0.011) | (0.010) |
| | | 2.5 | 0.500 | 0.460 | **0.540** | **1.050** | −0.783 | −0.776 | 0.635 | 0.624 |
| | | | (0.016) | (0.016) | (0.053) | (0.082) | (0.011) | (0.011) | (0.016) | (0.016) |
| | | 3 | 0.430 | 0.420 | **0.535** | **1.020** | −0.944 | −0.933 | 0.915 | 0.897 |
| | | | (0.016) | (0.016) | (0.049) | (0.085) | (0.011) | (0.012) | (0.020) | (0.021) |
| | | 3.5 | 0.240 | 0.210 | **0.590** | **1.070** | −1.036 | −1.015 | 1.095 | 1.055 |
| | | | (0.014) | (0.013) | (0.052) | (0.080) | (0.011) | (0.011) | (0.022) | (0.023) |

[a]Additive analysis means an additive model is adopted in the analysis when the underlying genetic mode is not.
[b]False negatives are counts out of 1 and false positives are counts out of 6.
[c]Mode refers to underlying genetic mode of data; Freq refers to frequency of risk haplotype; OR refers to the odds ratio of the risk haplotype.
[d]Measures in this table are found by averaging over 1,000 simulated data sets; standard errors (SE) are shown in parentheses.
[e]Bolded measures (pro vs. retro) are statistically significantly different at 0.05 level (i.e. the $\pm 2$SE intervals do NOT overlap).

would no longer be possible, as the maximum likelihood estimator is no longer necessarily unique.

## ACKNOWLEDGMENTS

The authors appreciate the GAW15 RA data set from the Genetic Analysis Workshop. Support for GAW15 data was provided from NIH grants 5RO1-HL049609-14, 1R01-AG021917-01A1, the University of Minnesota, and the Minnesota Supercomputing Institute, and the GAW grant, R01-GM031575 and AR44422. The authors thank the reviewers for the helpful and constructive feedback that greatly improved the manuscript. They also thank Dr. Danyu Lin, Tammy Bailey, and Chris Smith for their generous help with HAPSTAT and perl. M.L.K. was supported by NIH T32GM081057 and NIH R01 MH084022-01. H.D.B. was supported by NSF DMS-0705968, NIH R01 MH084022-01, and NIH 1P01-CA142538-01. J.Y.T. was supported by NIH R01 MH084022-01 and NIH 1P01-CA142538-01.

## REFERENCES

Arlot S, Celisse A. 2010. A survey of cross-validation procedures for model selection. Stat Surv 4:40–79.

Balding DJ. 2006. A tutorial on statistical methods for population association studies. Nat Rev Genet 7:781–791.

Barrett JC, Fry B, Maller J, Daly MJ. 2005. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21: 263–265.

Browning S, Thomas J. 2007. Multilocus analysis of GAW15 NARAC chromosome 18 case-control data. BMC Proc 1:S1–S11.

Chen YH, Kao JT. 2006. Multinomial logistic regression approach to haplotype association analysis in population-based case-control studies. BMC Genet 7:43.

Chen YH, Chatterjee N, Carroll RJ. 2009. Shrinkage estimators for robust and efficient inference in haplotype-based case-control studies. J Am Stat Assoc 104:220–233.

Clark AG. 2004. The role of haplotypes in candidate-gene studies. Genet Epidemiol 27:321–333.

Cordell HJ. 2006. Estimation and testing of genotype and haplotype effects in case-control studies: comparison of weighted regression and multiple imputation procedures. Genet Epidemiol 30:259–275.

de Bakker PW, Yelensky R, Pe'er I, Gabriel S, Daly MJ, Altshuler D. 2005. Efficiency and power in genetic association studies. Nat Genet 37:1217–1223.

Epstein MP, Satten GA. 2003. Inference on haplotype effects in case-control studies using unphased genotype data. Am J Hum Genet. 73:1316–1329.

French B, Lumley T, Monks SA, Rice KM, Hindorff LA, Reiner AP, Psaty BM. 2006. Simple estimates of haplotype relative risks in case-control data. Genet Epidemiol 30:485–494.

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

Hastie T, Tibshirani R, Friedman J. 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edition. Berlin: Springer.

Lake SL, Lyon H, Tantisira K, Silverman EK, Weiss ST, Laird NM, Schaid DJ. 2003. Estimation and tests of haplotype-environment interaction when linkage phase is ambiguous. Hum Hered 55: 56–65.

Li Y, Sung WK, Liu JJ. 2007. Association mapping via regularized regression analysis of single nucleotide polymorphism haplotypes in variable-sized sliding windows. Am J Hum Genet 80: 705–715.

Lin DY, Huang BE. 2008. The use of inferred haplotypes in downstream analysis. Am J Hum Genet 80:577–579.

Lin DY, Zeng D. 2006. Likelihood-based inference on haplotype effects in genetic association studies. J Am Stat Assoc 101:89–104.

Lin DY, Zeng D, Milikan R. 2005. Maximum likelihood estimation of haplotype effects and haplotype-environment interactions in association studies. Genet Epidemiol 29:299–312.

Prentice RL, Pyke R. 1979. Logistic disease incidence models and case-control studies. Biometrika 66:403–412.

Satten GA, Epstein MP. 2004. Comparison of prospective and retrospective methods for haplotype inference in case-control studies. Genet Epidemiol 27:192–201.

Schaid DJ. 2004. Evaluating associations of haplotypes with traits. Genet Epidemiol 27:348–364.

Shao J. 1997. An asymptotic theory for linear model selection. Stat Sin 7:221–264.

Souverein OW, Zwinderman AH, Tanck MW. 2006. Estimating haplotype effects on dichotomous outcome for unphased genotype data using a weighted penalize log-likelihood approach. Hum Hered 61:104–110.

Souverein OW, Zwinderman AH, Jukema JW, Tanck MW. 2008. Estimating effects of rare haplotypes on failure times using a penalized Cox proportional hazards regression model. BMC Genet 9:9.

Stram DO, Leigh Pearce C, Bretsky P, Freedman M, Hirschhorn JN, Altshuler D, Kolonel LN, Henderson BE, Thomas DC. 2003. Modeling and E-M estimation of haplotype specific relative risks from genotype data for a case-control study of unrelated individuals. Hum Hered 55:179–190.

Tanck MW, Klerkx AH, Jukema JW, De Knijff P, Kastelein JJ, Zwinderman AH. 2003. Estimation of multilocus haplotype effects using weighted penalized log-likelihood: analysis of five sequence variations at the choleteryl ester transfer protein gene locus. Ann Hum Genet 67:175–184.

Tibshirani R. 1996. Regression shrinkage and selection via the LASSO. J R Stat Soc 58:267–288.

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

Tzeng JY, Lu W, Farmen MW, Liu Y, Sullivan PF. 2010. Haplotype-based pharmacogenetic analysis for longitudinal quantitative traits in the presence of dropout. J Biopharm Stat 20:334–350.

Wang H, Leng C. 2007. Unified LASSO estimation by least squares approximation. J Am Stat Assoc 102:1039–1048.

Yang Y. 2005. Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. Biometrika 92:937–950.

Zaitlen M, Kang H, Eskin E, Halperin E. 2007. Leveraging the HapMap correlation structure in association studies. Am J Hum Genet 80: 683–691.

Zou H. 2006. The adaptive LASSO and its oracle properties. J Am Stat Assoc 101:1418–1429.