Communications in Information and Systems
Vol. 10, No. 1, pp. 39-52, 2010

© 2010 International Press

003

MODEL SELECTION METHODS FOR GENOME WIDE ASSOCIATION STUDIES*

SUDEEP SRIVASTAVA[†] and LIANG CHEN[‡]

Abstract. Due to the multiple loci control nature of complex phenotypes, there is great interest to test markers simultaneously instead of one by one. In this paper, we compare three model selection methods for genome wide association studies using simulations: the Stochastic Search Variable Selection (SSVS), the Least Absolute Shrinkage and Selection Operator (LASSO) and the Elastic Net. We also apply the three methods to identify genetic variants that are associated with daunorubicin-induced cytotoxicity. The simulation studies were performed by using the genotype data of 60 unrelated individuals from the CEU population in the Hapmap project. For the cytotoxicity data, we used 3,967,790 markers across the whole genome for 56 unrelated individuals from the CEU population. Using Sure Independence Screening as the pre-screening procedure, the SSVS gives a small model while the LASSO gives an intermediate sized model and the Elastic Net provides a large model. The three models share many common markers although the model sizes are different. The model sizes are subject to various cutoffs and parameters. The SSVS outperforms the LASSO and the Elastic Net in simulation studies. We also demonstrate the ability of the SSVS, the LASSO, and the Elastic Net to handle the situation when the number of markers is larger than the number of samples.

1. Introduction.

With the advances in genotyping technology, it has become feasible to perform large-scale, high-density genome wide association (GWA) studies to search for common genetic variants underlying complex phenotypes ([1, 2]). However, due to lack of computing power, single-marker tests remain the primary tools in the analysis of GWA data. Most quantitative phenotypes are complex in nature and are caused by multiple genetic variants, each of them having varying degree of effects. The possible interactions among genetic variants and the interactions between genes and the environment present additional challenges for Quantitative Trait Loci (QTL) mapping. Due to the multiple loci control nature, testing markers simultaneously instead of one by one may increase statistical power. In order to identify the correct set of genetic variants from millions of markers, efficient and reasonable model selection algorithms are in urgent need. Three popular model selection methods have been proposed : the Stochastic Search Variable Selection (SSVS) [3], the Least Absolute Shrinkage and Selection Operator (LASSO) [4], and the Elastic Net [5].

In the SSVS, a latent variable γ is introduced to perform the variable selection for the regression mode. $\gamma_i = 1$ implies that the i^{th} variable is included in the model and

[*] Dedicated to Michael Waterman on the occasion of his 67th birthday.

[†] Molecular and Computational Biology, Department of Biological Sciences, University of Southern California, Los Angeles CA, USA. E-mail: sudeep.srivastava@usc.edu

[‡] Molecular and Computational Biology, Department of Biological Sciences, University of Southern California, Los Angeles CA, USA. To whom correspondence should be addressed. E-mail: liangche@usc.edu

$\gamma_i = 0$ implies that the i^{th} variable is excluded from the model. A homogenous ergodic Markov Chain can be generated by the Gibbs Sampler. The empirical distribution of γ based on the Markov chain will converge to the actual posterior distribution of γ [6].

The LASSO proposed by Tibshirani is a shrinkage based selection method for linear regression. The LASSO minimizes the residual sum of squares subject to the constraint on the sum of absolute value of coefficients. This L^1-Norm constraint produces shrunk coefficients with some of them exactly equal to zero, which leads to interpretable models. In 2004, Efron et al. proposed the Least Angle Regression(LARS) [7] which is a computationally efficient model selection algorithm. There is a close connection between the LARS and the LASSO. A simple modification of the LARS algorithm can yield all the LASSO solutions. Due to their popularity and usefulness, the LASSO and the LARS have drawn intensive research interest in the statistical field.

The Elastic Net proposed by Zou and Hastie [5] uses a novel regularization penalty. The naive Elastic Net uses a combination of the LASSO and the Ridge regression penalty. However, the Elastic Net uses a scaled version of the naive Elastic Net estimate to reduce the overshrinking of parameters. It has been shown that the Elastic Net outperforms the LASSO [5]. In addition, it has a grouping effect in which correlated predictors group together. Thus, they are included together or excluded together from the model. This is advantageous in association studies as many markers are highly correlated via high linkage disequilibrium (LD). Another modification to the LARS algorithm gives all the solutions to the Elastic Net for a given value of the parameter. This enables a fast implementation of the Elastic Net algorithm. Due to efficient implementation and the grouping effect, the Elastic Net is very commonly used and gives more insight into the LASSO.

The LASSO and the Elastic Net are two of the most popular model selection methods which involve the minimization of the mean square error with respect to some constraints. SSVS on the other hand is based on the Gibbs Sampler which belongs to the broader class of Markov Chain Monte Carlo methods. Hence, a comparison of the three methods would be of great interest.

Daunorubicin is an anthracycline chemotherapeutic agent, which is used in the treatment of various cancers including leukemia, lymphoma, and advanced HIV-associated Kaposi's sarcoma [8, 9]. Daunorubicin has also been shown to be toxic and is associated with myelosuppression and cardiac toxicity [10, 11, 12]. It has been reported that ~ 29 % of variation in susceptibility to daunorubicin-induced toxicity is due to genetics [13]. Therefore, it is important to conduct GWA studies to identify genetic variants which are responsible for increased susceptibility to daunorubicin-induced toxicity. We used the phenotype data provided in [14]. The authors used a cell growth inhibition assay to measure variations in the cytotoxicity of daunorubicin. We applied the SSVS, the LASSO and the Elastic Net to the 3,967,790 SNPs to identify genetic variants associated with daunorubicin-induced cytotoxicity.

2. Methods.

2.1. Sure Independence Screening.

The genotype was coded as 0,1 or 2 for homozygous rare alleles, heterozygous alleles, and homozygous common alleles respectively (i.e., assuming additive effect). The missing alleles were imputed according to the genotype frequency calculated from the available data. As a prescreening step, markers with a minor allele frequency less than 0.01 were discarded. For the simulation studies, the phenotype data was simulated with 5 causal markers and each of the causal markers had an equal effect on the phenotype. The marker effect and the linkage disequilibrium among markers were varied. For the real data set, the phenotype data was transformed using an inverse normalization of percentile ranks.

We assume that the high dimensional data is sparse. That means most of the markers are not associated with the output phenotype. This is a reasonable assumption because of the huge number of markers. The biggest challenge with a large number of predictors is that there might be spurious correlations among different predictors which might lead to confounding correlations with the output. This is accentuated in association studies when markers are highly correlated among themselves due to linkage disequilibrium. The Dantzig Selector which is the solution to a L^1 regularization problem [15] has been shown to have the ideal risk up to a logarithmic factor $log(p)$ where p is the number of predictors. However $log(p)$ can also grow very fast and this bound is no longer adequate. Hence, we need to apply an effective dimensionality reduction method in the first stage before we apply a model selection method in the second stage as the efficiency of most model selection algorithms decreases dramatically with the increase in the number of predictors. This dimensionality reduction also helps in reducing computational time required for the model selection methods.

To decrease the dimensionality of the marker data for the real data set, a method called Sure Independence Screening (SIS) was performed [16]. The SIS has been shown to have the Sure Screening property which is that all the important variables survive in the model with a probability close to 1 under some conditions. The SIS method assumes a linear model.

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\mathbf{X}$ denotes the genotype data which are columnwise standardized, $\mathbf{Y}$ denotes the phenotype data, $\boldsymbol{\beta}$ is the regression coefficient, and $\boldsymbol{\epsilon}$ are i.i.d $N(0,1)$ random variables independent of the rest of the parameters in the model. The method uses correlation learning to detect predictors in the true model. $\mathbf{w}$ is defined as the vector obtained by

$$\mathbf{w} = \mathbf{X}^{\mathbf{T}}\mathbf{Y}.$$

The SIS selects the largest componentwise magnitudes of the vector $\mathbf{w}$. For a given α, M_α denotes the marker set output by SIS as

$$M_\alpha = \{1 \leq i \leq p :: |w_i| \text{ is amongst the } [\alpha n] \text{ largest of all } \},$$

where $[\alpha n]$ denotes the integer part of αn. The authors in [16] suggest that α should be chosen such that $[\alpha n] < n$. However, since we want to demonstrate that the SSVS and the LASSO can handle more markers than the number of samples, we select 200 markers with the largest correlations. Therefore, for our data, we used $\alpha = 10/3$. The SIS reduces the number of markers to $o(n)$ number of markers.

2.2. Stochastic Search Variable Selection. The SSVS was proposed by George et al. [3]. The SSVS uses a hierarchical Bayes model to identify the associated variables. Here, we assumed that the phenotype follows a multiple regression model of a subset of the markers. The canonical regression setup is given by

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1}$$

where the phenotype data $\mathbf{Y}$ is $n \times 1$, genotype data $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_p]$ is $n \times p$, $\mathbf{X}_i$ is the genotype data for marker i, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{'}$, and $\boldsymbol{\epsilon} \sim N(0, \theta^2)$ where θ^2 is scalar. The number of samples in the population is given by n and the number of markers by p.

A latent variable γ_i is defined as the indicator whether marker i is selected in the model or not. The $\beta_i^{\prime}s$ follow a mixture model of the form:

$$\beta_i|\gamma_i \sim (1 - \gamma_i)N(0, \sigma^2) + \gamma_i N(0, \tau^2). \tag{2}$$

And any prior information about the γ_i's can be incorporated by setting a prior on the γ_i's. Let $f(\boldsymbol{\gamma})$ denote the prior. In our model, we assumed that the γ_i's are independent with marginal distributions as below:

$$P(\gamma_i = 1) = 1 - P(\gamma_i = 0) = 1/p. \tag{3}$$

To obtain (2) as the prior for $\beta_i|\gamma_i$, a multivariate normal prior is used as follows:

$$\boldsymbol{\beta}|\boldsymbol{\gamma} \sim N(\mathbf{0}, \mathbf{D}_\gamma \mathbf{R} \mathbf{D}_\gamma). \tag{4}$$

$\mathbf{R}$ is the prior correlation matrix and $\mathbf{D}_\gamma$ is defined as

$$\mathbf{D}_\gamma = diag[\phi_1, \phi_2, \ldots, \phi_p], \tag{5}$$

where $\phi_i = \sigma$ if $\gamma_i = 0$ and $\phi_i = \tau$ if $\gamma_i = 1$. We used the prior correlation matrix $\mathbf{R}$ as the identity matrix, but correlations between the markers can be incorporated in this prior correlation matrix. The prior on the residual variance is given by

$$\theta^2|\boldsymbol{\gamma} \sim InverseGamma(n/2, \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2/2). \tag{6}$$

The posterior probabilities of $\boldsymbol{\gamma}$ can be estimated from the Markov chain generated by the Gibbs Sampler. We run the Gibbs Sampler for 2000 iterations to achieve stationarity and then run it for an additional 8000 iterations to estimate the posterior probabilities.

2.3. LASSO. The LASSO, or the "least absolute shrinkage and selection operator", tries to shrink some coefficients and set most of the coefficients exactly equal to 0 to achieve a model with a small number of variables and a small mean square error.

Using the linear model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\mathbf{Y}, \mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\epsilon}$ are the same as above. The LASSO tries to minimize $\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2$ subject to the L^1 norm $\sum_j |\beta_j| \leq t$. Here $t \geq 0$ is the tuning or shrinkage parameter. This statement can be rephrased as

$$\hat{\boldsymbol{\beta}} = argmin_{\boldsymbol{\beta}}(\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \sum_j |\beta_j|).$$

By using the L^1 norm, the LASSO ensures that a subset of the predictors are exactly 0. Some studies have been done on the consistency of the LASSO. Zhao and Yu [17] proved that when a condition known as the *Strong Irrepresentable Condition* is satisfied and when the error terms have some finite moments, the LASSO is strongly sign consistent, i.e. $\exists \lambda = f(n)$ which is independent of $\mathbf{Y}$ or $\mathbf{X}$ such that

$$\lim_{n \to \infty} P(sign(\hat{\boldsymbol{\beta}}(\lambda)) = sign(\boldsymbol{\beta})) \longrightarrow 1.$$

for large p and q where q is the number of markers which are not associated with the phenotype. However, we only have 56 samples, hence we cannot rely on this condition for the LASSO to pick up the correct markers. The LASSO is a quadratic programming problem but can be solved by a simple modification to the Least Angle Regression algorithm by Efron et al.[7].

2.4. Elastic Net. The Elastic Net uses the linear model and tries to minimize the least square error using a novel regularization penalty. It uses two regularization parameters (λ_1, λ_2). The linear model is given below :

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\mathbf{Y}$, $\mathbf{X}$, $\boldsymbol{\beta}$ and $\boldsymbol{\epsilon}$ have their usual meanings. The definition for the naive Elastic Net estimator is given below. For any fixed non-negative λ_1 and λ_2, the naive Elastic Net criterion is defined as

$$L(\lambda_1, \lambda_2, \boldsymbol{\beta}) = |\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}|^2 + \lambda_2 \sum_{i=1}^{p} \beta_i^2 + \lambda_1 \sum_{i=1}^{p} |\beta_i|.$$

The naive Elastic Net estimator $\hat{\boldsymbol{\beta}}$ is the minimizer of the above equation:

$$\hat{\boldsymbol{\beta}} = \arg\ \min_{\boldsymbol{\beta}}\{L(\lambda_1, \lambda_2, \boldsymbol{\beta})\}.$$

An artificial data set $(\mathbf{Y}^*, \mathbf{X}^*)$ is defined as follows:

$$\mathbf{X}^* = (1+\lambda_2)^{-1/2} \begin{pmatrix} \mathbf{X} \\ \sqrt{\lambda_2}\mathbf{I_p} \end{pmatrix},$$

where $\mathbf{I_p}$ is the $p \times p$ identity matrix and p is the number of markers. And

$$\mathbf{Y}^* = \begin{pmatrix} \mathbf{Y} \\ \mathbf{0} \end{pmatrix},$$

where $\mathbf{0}$ is the $p \times 1$ 0-vector.
It can be shown that the naive Elastic Net solves a lasso-type problem given by

$$\hat{\boldsymbol{\beta}}^* = \arg\ \min_{\hat{\boldsymbol{\beta}}^*} |\mathbf{Y}^* - \mathbf{X}^*\boldsymbol{\beta}^*|^2 + \frac{\lambda_1}{\sqrt{1+\lambda_2}}|\boldsymbol{\beta}^*|_1.$$

The Elastic Net estimates are given as

$$\hat{\boldsymbol{\beta}}(\text{elastic net}) = \sqrt{1+\lambda_2}\hat{\boldsymbol{\beta}}^*.$$

This kind of scaling undoes the overshrinking effect when we combine the L^1 and L^2 penalties. To choose the coefficients, a two dimensional cross validation is performed. The LARS extension which implements the Elastic Net algorithm, called the LARS-EN, outputs a sequence of variables corresponding to a given λ_2. λ_1 has a one to one correspondence with the number of iterations that the LARS-EN algorithm was run for. Therefore selecting the active model at a given iteration for a particular λ_2 would give us the Elastic Net solution corresponding to particular value of (λ_1, λ_2). To estimate the model parameters, different values of λ_2 are chosen (i.e. $(0, 0.01, 0.1, 1, 10, 100)$) and the other tuning parameter is chosen using 10 fold cross validation. The chosen λ_2 is the one giving the minimum cross-validation error. The Elastic Net has been shown to have a grouping effect, i.e. variables highly correlated with each other are included and excluded in the model together. This is extremely useful in association mapping as many markers are in high LD with each other.

2.5. Area Under the Curve. For simulation studies, the Area Under the Curve (AUC) statistic was used to assess the power of the method. The True Positive Rates(TP) and the False Positive Rates(FP) are defined as follows:

$$TP(c) = \frac{\text{Number of Markers correctly classified as causal markers}}{\text{Number of Causal Markers}},$$

$$FP(c) = \frac{\text{Number of Markers wrongly classified as causal markers}}{\text{Number of Non-Causal Markers}},$$

where c is the cutoff used in the method. For the SSVS c is the cutoff for the posterior probability for the γ_j's. For the LASSO, c is the number of iterations that the LASSO is run. The Receiver Operating Characteristic (ROC) curve is the two-dimensional plot of TP(c) vs. FP(c) (or sensitivity vs. (1-specificity)) for $-\infty < c < \infty$. The overall performance of a classifier can be measured by the area under the ROC curve. This quantity is called the AUC. An AUC of 0.5 represents a complete random guess.

For the SSVS, the AUC is calculated using the following formula modified from [18]

$$AUC = \frac{1}{n_C n_{C^c}} \sum_{i \in C, j \in C^c} I\{\gamma_i > \gamma_j\},$$

where C is the set of indices of the causal markers and C^c is the set of indices of the non-causal markers. n_C and n_{C^c} are the number of causal and non-causal markers respectively. For the LASSO, the following formula is used to calculate the AUC

$$AUC = \frac{1}{n_C n_{C^c}} \sum_{i \in C, j \in C^c} I\{\delta_i < \delta_j\},$$

where δ_i represents the first iteration at which the i^{th} marker enters the model. This corresponds to a cutoff of iterations used to select the model for the LASSO.
The AUC for the Elastic Net is calculated using the same formula as the LASSO, using the λ_2 from the set $\{0, 0.01, 0.1, 1, 10, 100\}$ as given in [5].

3. Results.

3.1. Simulation Studies. For simulation studies, we considered the phenotype data for 60 individuals. The genotype data was simulated as follows: we selected a set of markers from chromosome 1 of the Hapmap CEU population data at different marker densities. The density of markers was varied according to the average number of markers selected from every 1,000 markers of the Hapmap Phase I data. We used 60 markers per sample in the simulations. Among them, five markers were selected to be associated with the phenotype. These markers were labeled as the causal markers. The phenotype was simulated from a linear model with different coefficients according to the following equation:

$$Y_j = \beta X_{1j} + \beta X_{2j} + \ldots + \beta X_{5j} + \epsilon_j,$$

where Y_j is the simulated phenotype for the j^{th} individual, $X_{1j}, X_{2j}, \ldots, X_{5j}$ denotes the genotypes of the five causal markers for the j^{th} individual, β denotes the coefficient of the causal markers and $\epsilon'_j s$ are simulated as i.i.d $N(0, 1)$. The AUC statistic was calculated for the three methods and are summarized in Table 1. AUC values are shown for the SSVS with $\sigma = 0.05$ and different values of τ, the LASSO and the Elastic Net. The AUC values for a single marker F-test are also shown in Table 1. The SSVS consistently has a higher AUC value than the other two methods. The Elastic Net has a smaller AUC than the LASSO when the effect size is small and LD is high, but is consistently higher than the LASSO in other cases. The single marker F-test has a higher AUC than the LASSO and the Elastic Net when the marker effect is small. However, the LASSO and the Elastic Net are better in other cases.

Figures 1 and 2 show the ROC curves for the LASSO and the SSVS. Figure 1 shows the ROC curves for the LASSO with the marker data set at different marker densities for $\beta = 1$. These marker densities are measured as selecting 100,10 and 1 marker on an average per 1000 markers in the Hapmap data. We can see that the area under the ROC curve increases as the marker density decreases. This is because when the marker density decreases, the independence between the markers "increases" and hence the method can detect the causal markers more effectively. The same is seen in Figure 2 for the SSVS ($\beta = 1, \sigma = 0.05$ and $\tau = 1$). However, since the values are very close to each other, the curves nearly overlap with one another. Also, the ROC Curves intersect with each other, making it difficult to judge which one has the higher AUC. More detailed values are listed in Table 1. It clearly shows that the AUC increases when marker density decreases. The AUC values also change as the parameter τ changes for the SSVS. When coefficient $\beta = 1$, $\tau = 1$ gives the maximum AUC score. However, the differences are small. The ROC has a dip at the end for the lasso in Figure 1. The LASSO is implemented as a modified LARS algorithm which can remove predictors after they are added into the model. Since the algorithm with different iteration cutoffs corresponds to a LASSO solution for a particular λ, we used the iteration number as the cutoff to make the ROC curve. However, as we increase the number of iterations in the cutoff, more false positives might be included and the true positive rate might also decrease as true causal markers are removed from the model. Fortunately, these mainly happen when the false positive rate is high (e.g., > 0.7). And we are interested in the performance of the method when the false positive rate is reasonably low. If we assume that a variable is never removed from the model and calculate the sensitivity and specificity, we will get a ROC curve which is monotone increasing.

3.2. Daunorubicin-Induced Cytotoxicity Data. We considered 3,967,790 SNPs in the real data analysis. Markers with a minor allele frequency less than 0.01 were screened out. After this step, 2,598,208 markers remained. A total of

TABLE 1

Simulation results of QTL mapping. Simulation results are shown for different coefficients, marker densities (measured in the average number of markers selected from every 1,000 markers of the Hapmap Phase I data) and different values of τ for SSVS. $\sigma = 0.05$ for the SSVS. The AUC values are the average values across 1,000 simulations.

Coefficient(β)	Marker Density	SSVS AUC ($\tau = 1$)	SSVS AUC ($\tau = 2$)	SSVS AUC ($\tau = 3$)	LASSO AUC	Elastic Net AUC	F-test AUC
0.5	200	0.898	0.891	0.884	0.684	0.679	0.729
0.5	100	0.923	0.918	0.913	0.7	0.718	0.753
0.5	10	0.949	0.944	0.942	0.783	0.824	0.800
0.5	1	0.950	0.947	0.942	0.798	0.830	0.814
1	200	0.969	0.955	0.947	0.873	0.894	0.814
1	100	0.987	0.982	0.978	0.886	0.912	0.854
1	10	0.997	0.996	0.995	0.965	0.977	0.922
1	1	0.999	0.998	0.997	0.979	0.982	0.932
1.5	200	0.989	0.985	0.976	0.926	0.940	0.838
1.5	100	0.998	0.998	0.996	0.939	0.961	0.887
1.5	10	1	1	0.999	0.991	0.993	0.951
1.5	1	1	1	1	0.995	0.996	0.951
2	200	0.991	0.988	0.986	0.952	0.962	0.823
2	100	0.999	0.997	0.997	0.962	0.973	0.874
2	10	1	1	0.999	0.995	0.997	0.963
2	1	1	1	1	0.998	0.999	0.966

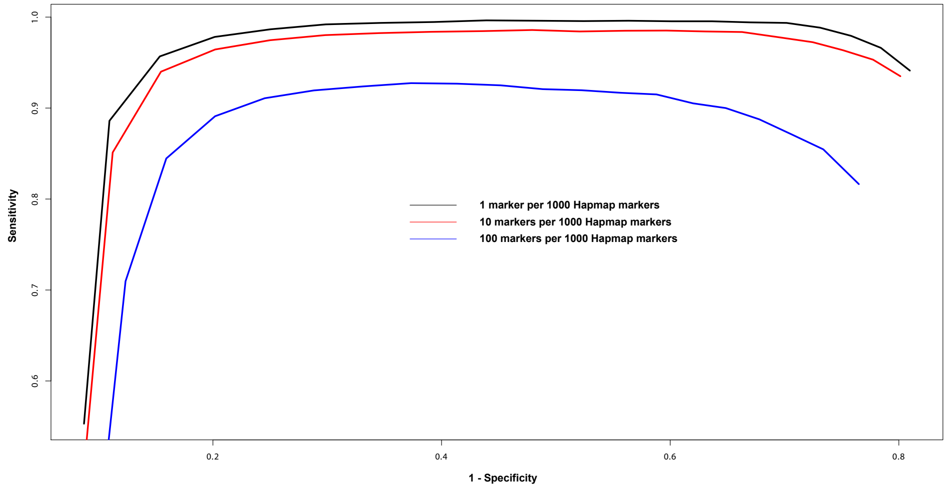


FIG. 1. *ROC Curves for the LASSO for different marker densities ($\beta = 1$).*

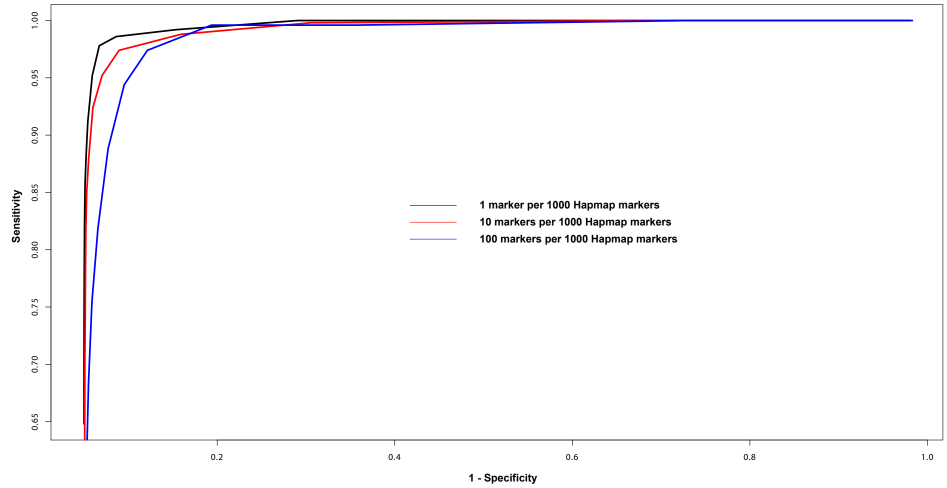


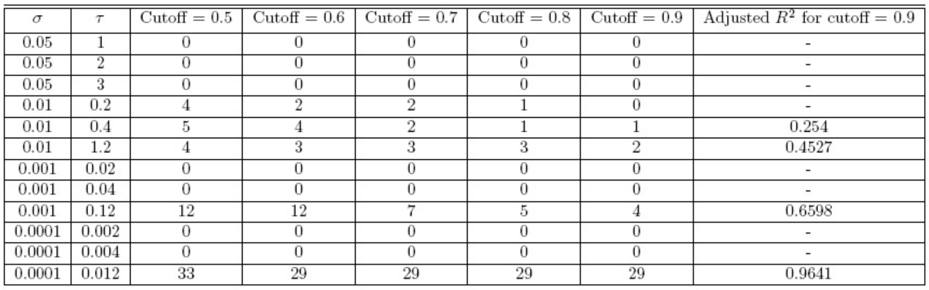
FIG. 2. *ROC Curves for the SSVS for different marker densities ($\beta = 1, \sigma = 0.05, \tau = 1$).*

56 unrelated CEU individual have the phenotype data available. Using SIS, the correlation was calculated between the phenotype and the genotype. The top 200 markers were chosen for the SSVS, the LASSO, and the Elastic Net.

For the SSVS, the results can vary dramatically by changing the initial parameters of τ and σ. We use $\sigma = 0.05, 0.01, 0.001, 0.0001$ and corresponding values of τ based on the values of $\frac{\tau^2}{\sigma^2} = 400, 1600, 3600$. Table 2 shows the number of variables selected for the models with different parameters and posterior probability cutoffs. The results are robust to the posterior probability cutoffs, which also indicates the convergence of the Chain. Using the model with the largest adjusted R^2, we select $\sigma = 0.0001$ and $\tau = 0.012$ giving an adjusted R^2 of 0.9641. Figure 3 shows the positions (green triangles) of the markers which were selected by the SSVS with model parameters $\sigma = 0.0001, \tau = 0.012$.

TABLE 2

Number of markers selected by the SSVS for different posterior probability cutoffs.



σ	τ	Cutoff = 0.5	Cutoff = 0.6	Cutoff = 0.7	Cutoff = 0.8	Cutoff = 0.9	Adjusted R^2 for cutoff = 0.9
0.05	1	0	0	0	0	0	-
0.05	2	0	0	0	0	0	-
0.05	3	0	0	0	0	0	-
0.01	0.2	4	2	2	1	0	-
0.01	0.4	5	4	2	1	1	0.254
0.01	1.2	4	3	3	3	2	0.4527
0.001	0.02	0	0	0	0	0	-
0.001	0.04	0	0	0	0	0	-
0.001	0.12	12	12	7	5	4	0.6598
0.0001	0.002	0	0	0	0	0	-
0.0001	0.004	0	0	0	0	0	-
0.0001	0.012	33	29	29	29	29	0.9641

The LASSO algorithm was run till the residual was below a certain level . The optimum number of iterations was selected using a 10 fold cross validation. The minimum mean square error was achieved at the 115^{th} iteration. The LASSO selected 56 markers in the final model. The positions of the markers are also shown in Figure 3 (blue diamonds). The Venn diagram in Figure 4 shows that the SSVS and the LASSO identify 10 common markers.

The Elastic Net algorithm was also used to select significant markers. The λ_2 parameter was selected from possible values of 0,0.01,0.1,1,10 and 100 using a 10 fold cross validation. A $\lambda_2 = 0.1$ was used in the final model as it gave the minimum cross validation score. The number of iterations were further chosen using a 10 fold cross validation and a model of size 160 was chosen (red circles in Figure 3). Among these selected markers, 23 were also identified by the SSVS and 55 were identified by the LASSO. The 10 markers common between the SSVS and the LASSO were also selected by the Elastic Net. From Figure 3, we can see that the models chosen by the different methods are quite consistent, except for the size of the model.

4. Discussion. We show that the SSVS outperforms the LASSO and the Elastic Net in simulation studies. All the three methods have similar trends in power as we change marker density and coefficients of the markers. For the daunorubicin data set, the SSVS, the LASSO and the Elastic Net select a significantly common model, however the size of the SSVS model is small. The Elastic Net selects a very large model. None of the common markers selected have been previously reported to be associated with daunorubicin induced cytotoxicity. However, in the prescreening, the SIS only selects two markers which have been previously reported to be associated

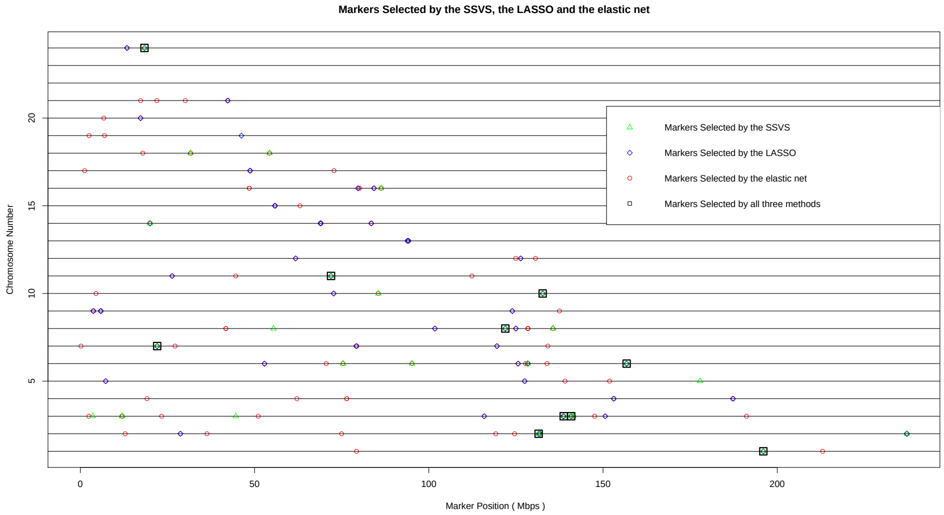


FIG. 3. *Positions of markers selected by the SSVS, the LASSO and the Elastic Net.*

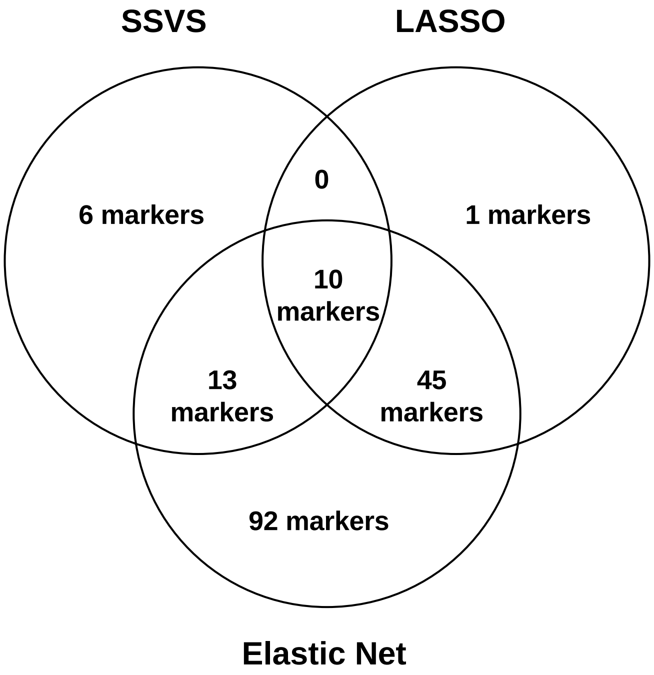


FIG. 4. *Comparison of the outputs of the three methods for the Daunorubicin-induced cytotoxicity data.*

with daunorubicin induced cytotoxicity, rs220200 and rs10142144.

The methods have their advantages and disadvantages in their application to model selection. The SSVS is computationally intensive as it requires computation of a matrix inverse at each iteration. The LARS requires the inverse computation of a matrix with the size of the active set. Therefore as the number of iterations increases, the time required for an iteration would increase. Both the LASSO and the Elastic Net use the LARS algorithm and hence have the same problem. However, since the SSVS requires the Markov chain achieve stationarity, it requires much more time than the LASSO and the Elastic Net. Therefore, we need to carry out a marker screening step (e.g., SIS procedure) before running any of the three methods. The Gibbs Sampler in the SSVS requires prior parameter selection. The parameters for the SSVS have been selected according to [3] for different values of σ. However, it is impossible to explore the whole space. The final results may be very sensitive to parameter estimation. Biological information can be incorporated using the prior parameters. For the LASSO on the other hand, the cutoff needs to be selected. A cross-validation scheme as used in this paper could be biased by the data set. The Elastic Net involves a two fold cross validation where one of the parameters are again chosen from a specific set. As the size of this set increases, the cross validation procedure can become very computationally intensive. In real data, it would be beneficial to use multiple methods and weight the results accordingly.

Neither of the methods are able to find any association with rs120525235 and rs3750518 mentioned in [14]. These markers are not detected by the SIS. However [14] uses gene expression data along with the genotype-phenotype association study. Thus, they used additional data resources. More methods need to be developed to integrate gene expression into GWA studies.

Prescreening methods have become of utmost importance with the advances in technology. We used the SIS to reduce the number of markers from 2598208 to 200. The authors in the SIS paper [16] suggested to use $p = n - 1$ or $p = \frac{n}{logn}$ which are much smaller than 200 in our case. We wanted to demonstrate the ability of these methods to use more markers than samples. So we chose 200 (much larger than $n - 1$ and $\frac{n}{logn}$) instead. The dimensionality reduction would ideally be method specific. We need to develop a method-specific algorithm to choose the number of retained markers after dimensionality reduction. If the causal markers are discarded in the prescreening step, then it is impossible for any method to identify them. We also demonstrated the ability of the SSVS, the LASSO and the Elastic Net to handle more markers than the sample size of 56. However, increasing the number of markers can dramatically reduce the power. It would be of great interest to address the relationship between the number of markers chosen in the prescreening step and the statistical power.

The SSVS and the LASSO both rely on the assumption that the predictors are independent. However, the markers are dependent on each other due to linkage disequilibrium, which would need to be considered to make an accurate statistical inference. The Elastic Net with its grouping property is a better choice when there are highly correlated variables.

REFERENCES

[1] M. I. McCarthy et al., *Genome-wide association studies for complex traits: consensus, uncertainty and challenges.* Nature Rev. Genet., 9(2008), pp. 356–369.

[2] L. Kruglyak, *The road to genome-wide association studies.* Nature Rev. Genet., 9(2008), pp. 314–318.

[3] E. I. George and R. E. McCulloch, *Variable selection via gibbs sampling.* Journal of the American Statistical Association, 88(1993), pp. 881–889.

[4] R. Tibshirani, *Regression shrinkage and selection via the lasso.* Journal of the Royal Statistical Society, 58(1996), pp. 267–288.

[5] H. Zou and T. Hastie, *Regularization and variable selection via the elastic net.* JRST, 67(2005), pp. 301–320.

[6] G. Casella and E. I. George, *Explaining the gibbs sampler.* The American Statistician, 46(1992), pp. 167–174.

[7] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. *Least angle regression.* Annals of Statistics, 32(2004), pp. 407–499.

[8] H. Davis and T. Davis, *Daunorubicin and adriamycin in cancer treatment: an analysis of their roles and limitations.* Cancer Treat Rep, 63(1979), pp. 809–815.

[9] D. Schurmann, A. Dormann, T. Grunewald, and B. Ruf, *Successful treatment of aids-related pulmonary kaposi's sarcoma with liposomal daunorubicin.* Eur Respir J, 7(1994), pp. 824–825.

[10] S. Lipschultz, *Exposure to anthracyclines during childhood causes cardiac injury.* Semin Oncol, 33(2006), pp. S8–14.

[11] K. Seiter, *Toxicity of the topoisomerase ii inhibitors.* Expert Opin Drug Saf, 4(2005), pp. 219–34.

[12] R. Young, R. Ozols, and C. Myers, *The anthracycline antineoplastic drugs.* N Engl J Med, 305(1981), pp. 139–153.

[13] S. Duan, W. K. Bleibel, R. S. Huang, S. J. Shukla, X. Wu, J. A. Badner, and M. E. Dolan, *Mapping genes that contribute to daunorubicin-induced cytotoxicity.* Cancer Res, 67(2007), pp. 5425–33.

[14] R. S. Huang, S. Duan, E. O. Kistner, W. K. Bleibel, S. M. Delaney, D. L. Fackenthal, S. Das, and M. E. Dolan, *Genetic variants contributing to daunorubicin-induced cytotoxicity.* Cancer Res, 68(2008), pp. 3161–3168.

[15] E. Candes and T. Tao, *The dantzig selector : Statistical estimation when p is much larger than n.* Annals of Statistics, 35(2007), pp. 2313–2351.

[16] J. Fan and J. Lv. *Sure independence screening for ultrahigh dimensional feature space.* Journal of the Royal Statistical Society Series B, 70(2008), pp. 849–911.

[17] P. Zhao and B. Yu, *On model selection consistency of lasso.* Journal of Machine Learning Research, 7(2006), pp. 2541–2563.

[18] S. Ma and J. Huang, *Combining multiple markers for classification using roc.* Biometrics, 63(2007), pp. 751–757.