1

Group Variable Selection Methods and Their Applications in Analysis of Genomic Data

Jun Xie[1] and Lingmin Zeng[2]

[1]Department of Statistics, Purdue University, 250 N. University Street, West Lafayette, IN 47907, USA
[2]MedImmune, 1 MedImmune Way, Gaithersburg, MD 20878, USA

1.1 Introduction

Regression is a simple but the most useful statistical method in data analysis. The goal of regression analysis is to discover the relationship between a response y and a set of predictors $x_1, x_2, \ldots, x_p$. When fitting a regression model, besides prediction accuracy, parsimony is another important criterion of goodness. Simpler models are preferred by researchers for easier interpretation of the relationship between x and y. Moreover, discarding irrelevant predictors often improves prediction accuracy [13]. Variable selection methods have long been used in regression analysis, for example forward selection, backward elimination, best subset regression. The number of variables p in the traditional setting is typically 10 or at most a few dozens. Modern scientific technology, led by the microarray, has produced data dramatically above the conventional scale. We have $p = 1,000$ to 10,000 in gene expression microarray data, and p up to 500,000 in single nucleotide polymorphism (SNP) data.

To make things more complicated, the large number of variables in the biological data are dependent. For example, it is well known that for genes that share a common biological function or participate in the same metabolic pathway, the pairwise correlations among them can be very high [14]. Traditional variable selection methods that select variables one by one may miss important group effects on pathways. Consequently, when traditional variable selection methods are applied in multiple data sets from a common biological system, the selected variables from the multiple studies may show little overlap. To overcome the challenges, we have developed a series of group variable selection methods, which construct highly correlated genes into a group and select the whole group once one gene among them is in the model. In this chapter, we introduce the idea of group variable selection and illustrate its utility by applying the methods to genomic data analysis.

1.2 Background

1.2.1 Existing variable selection methods

We consider a linear regression model

$$\mathbf{y} = \mathbf{x}_1\beta_1 + \mathbf{x}_2\beta_2 + \cdots + \mathbf{x}_p\beta_p + \epsilon$$

where the response $\mathbf{y}$ is predicted by p predictors $\mathbf{x}_1, \ldots, \mathbf{x}_p$. Without loss of generality, the response and the predictors are all centered so that there is no intercept in the model. Assume n observations and the error term $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$ are i.i.d. with mean 0 and variance σ^2. The regression model is often expressed in a matrix format

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

where $\mathbf{y} \in R^n$, $\mathbf{X} \in R^{n\times p}$, and $\beta \in R^p$.

A traditional variable selection method is known as the best subset selection. The procedure first determines a criterion of model goodness, for example, residual sum of squares, adjusted R^2, Mallow's C_p, the Akaike information criterion (AIC), or the Bayesian information criterion (BIC). Then all possible subsets of variables are evaluated by the criterion and the subset that optimizes the criterion is selected. However, when the number of variables p is large, the best subset selection is computationally intensive. Huo and Ni [5] prove that the best subset selection is an NP-hard (nondeterministic polynomial-time hard) problem. That is, the best subset solution cannot be obtained in computation times as a polynomial of the number of variables. Alternatively, sequential approaches can be used, including forward selection, backward elimination, and stepwise regression. The sequential approaches are computationally less demanding than the best subset selection. However, their heuristic searches of variables cannot guarantee an optimal solution to the regression model.

More recently, penalized least squares methods have been used for variable selection. The most popular one is Lasso (Least absolute shrinkage and selection operator) proposed by Tibshirani [17]. The Lasso estimators are defined by

$$\hat{\beta}_{Lasso} = \text{argmin}_\beta ||\mathbf{y} - \mathbf{X}\beta||^2 + \lambda \sum_{j=1}^{p} |\beta_j|,$$

where λ is a nonnegative regularization parameter. The second term of the sum of the absolute regression coefficients is usually called L_1 penalty. Equivalently, Lasso is a constrained ordinary least squares that minimizes

$$||\mathbf{y} - \mathbf{X}\beta||^2 \qquad \text{subject to } \sum_{j=1}^{p} |\beta_j| \le s,$$

where s is a corresponding regularization parameter. Due to the nature of the L_1 penalty, Lasso shrinks the regression coefficients toward 0 and produces some coefficients that are exactly 0, and hence implements variable selection. Many researchers have studied properties of Lasso [8, 9, 24, 25]. Under certain conditions, Lasso is shown to select the right set of variables with a probability going to 1. However, Lasso's conditions are violated for a group of highly correlated variables. In this situation, Lasso tends to select only one variable from the group and does not care which one is selected.

Efron et al. [3] propose a new variable selection algorithm, Least Angle Regression (LARS), which is a less greedy version of traditional forward selection methods. A special feature of LARS is that a simple modification of the LARS algorithm calculates all possible Lasso estimators but uses computer times an order of magnitude less than Lasso. The efficiency of the LARS algorithm makes it an attractive variable selection method.

Besides Lasso, other penalized least squares approaches have been proposed, using penalty functions more general than the L_1 penalty. Fan and Li [4] define a special penalty function that is singular at the origin to produce sparse coefficient estimators; satisfies certain conditions to produce continuous models; is bounded by a constant to produce nearly unbiased estimators for large coefficients. Their penalty function is called SCAD. Fan and Li [4] show that, with a proper choice of the regularization parameter, SCAD possesses oracle properties, which are referred to that the probability of selecting the right set of variables (with nonzero coefficients) converges to 1 and that the estimators of the nonzero coefficients are asymptotically normal with the same means and covariances as if the zero coefficients were known in advance. Kim et al. [7] also apply SCAD in certain high dimensional data.

The traditional variable selection methods and the later additions Lasso, LARS, SCAD, do not select variable groups. In fact, they all ignore the correlation between the variables. Elastic net proposed by Zou and Hastie [27] is the first variable selection method that works for groups of predictors. Elastic net is also a member of penalized least squares. The penalty function is a linear combination of L_1 and L_2 penalties. By introducing a L_2 penalty term, elastic net encourages strongly correlated variables to be in or out of the model at the same time. This phenomenon is termed "grouping effect". In theory, a strictly convex penalty function provides a sufficient condition for the grouping effect. The L_2 penalty guarantees strict convexity. On the other hand, elastic net does not reveal the underlying group structure in its solution and does not possess the properties introduced by Fan and Li [4] in SCAD.

When people have prior knowledge on variable groups, group Lasso proposed by Yuan and Lin [22] is designed to select pre-defined groups of predictors. Suppose p predictors are divided into J groups with sizes $k_1, \ldots, k_J$. The group Lasso estimators are obtained by minimizing

$$||\mathbf{y} - \sum_{j=1}^{J} \mathbf{X}_j \beta_j||^2 + \lambda \sum_{j=1}^{J} ||\beta_j||_{K_j}$$

where λ is the regularization parameter and $||z||_K = (z'Kz)^{1/2}$ with a symmetric $k \times k$ positive definite matrix K. The positive definite matrices $K_1, \ldots, K_J$ for the J groups can be chosen as identity matrices of sizes $k_1, \ldots, k_J$ respectively. Yuan and Lin [22] also propose group LARS as an extension of LARS. Group Lasso and group LARS have been used in multi-factor ANOVA models, in which each factor may have several levels and can be expressed through a group of dummy variables.

In addition to the frequentist methods, Bayesian interpretations of the penalized regression Ridge and Lasso have been proposed. It can be shown that the Bayesian estimators of the coefficients $\beta_1, \ldots, \beta_p$ are equivalent to Ridge when we assume normal prior distributions for β's, and are equivalent to Lasso when we assume Laplace prior distributions [11, 21]. Bayesian approach offers an alternative framework of variable selection. Theoretically, Bayesian methods can deal with high dimensional inter-correlated variables through generalized prior distributions. In practice, Bayesian methods will encounter the same difficulty as its frequentist counterpart.

1.2.2 Large scale genomic data

High-throughput gene expression microarray techniques have now been routinely used in biological applications. An array measures expression levels of thousands of genes simultaneously. Differences between experiment conditions (treatments) are implied by expression variations of a large number of genes. In medical research, microarray is used to detect associations between gene expression profiles and clinical outcomes, for example cancer types or stages. Consider a clinical outcome as the response variable y and all genes measured in the microarray as the predictor variables $x_1, \ldots, x_p$. Then the variables are of high dimension, with complicated dependent structures. Identifying a subset of significant genes that affect the clinical outcome will be a good application of our proposed group selection methods.

In one of our previous projects, we have developed a suite of statistical methods [23] for inferring cis-regulatory modules, which are groups of transcription factors binding in the promoter regions to regulate gene expression. Our approach is an integrative analysis that combines information from multiple types of biological data, including genomic DNA sequences, genome-wide location analysis (ChIP-chip experiments), and gene expression microarray. We first use a hidden Markov model by Wu and Xie [19] to predict a cluster of transcription factor binding sites in DNA sequences. The predictions are refined by regression analysis on gene expression microarray data and/or ChIP-chip binding experiments. We have constructed a regression model that describes a gene of interest as a function of its TFs. The response variable is the gene expression level. The predictor variables are the TF binding levels approximated by the TF gene expression values. We view a combinatorial effect of multiple TFs on the gene through multiple regression analysis. How-

ever, the difficulty is to select an appropriate set of TFs which has significant effects on the gene.

Due to complicated dependence among TFs, the problem of selecting TF covariates for a gene posts a challenge to the standard variable selection procedures. Consider a regression example of gene ACE2 versus a set of TFs consisting of Fkh1, Fkh2, Mcm1, Ndd1, Swi4, and Swi6 using Spellman et al.'s [16] yeast cell cycle microarray data. It is known that ACE2 was bound by Fkh1, Fkh2, and the complex Mcm1/Fkh2/Ndd1 [15]. Hence, a good variable selection method is to select the corresponding four TFs as much as possible. The expression levels of FKH1 and FKH2 are highly correlated with a correlation coefficient of 0.63. Using forward selection, Lasso, and elastic net, Fkh2 always enters the model first. However, the standardized regression coefficient of Fkh2 is 0.677 and that of Fkh1 is -0.045, when both Fkh2 and Fkh1 are in the model. The available methods fail to select both Fkh2 and Fkh1 as a group of covariates. In fact, variable selection in regression analysis tends to keep only one variable in the model, whenever there is a group of highly correlated covariates. The group variable selection methods attempt to solve this problem.

Another application of the proposed group variable selection methods is SNP data analysis. SNPs are the most common genetic variations in the human genome and occur once in several hundred base pairs. A SNP is a position at which two alternative bases occur at appreciable frequency ($> 1\%$) in the human population. The NCBI dbSNP database currently stores 5 million human SNPs identified by comparing the DNA of different individuals, making it possible to use them for genome-wide SNP genotyping. SNPs can serve as genetic markers for identifying disease genes by linkage studies in families, linkage disequilibrium in isolated populations, association analysis of patients and controls, and loss-of-heterozygosity studies in tumors [18]. Oligonucleotide SNP microarrays have been developed for high-throughput genotyping of human SNPs with marker number ranging from 10,000 (Mapping 10K array) to 500,000 (Mapping 500K array set). With the technique advances, genome-wide association studies become popular to detect specific DNA variants that contribute to human phenotypes and particularly human diseases. In SNP data analysis, we assume a phenotype of interest as the response variable y, and a large number of SNPs as the predictor variables. The proposed group variable selection methods will be used to identify genetic variants that associate with variation in the phenotype.

1.3 gLars and gRidge algorithms

Consider a linear regression model

$$\mathbf{y} = \mathbf{X}\beta + \epsilon,$$

where $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$ is the response variable, $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p)$ is the predictor matrix, and ϵ is a vector of independent and identically distributed random errors with mean 0 and variance σ^2. There are n observations and p predictors. We center the response variable and standardize the column vectors of the predictor matrix. Hence, there is no intercept in our model.

$$\sum_{i=1}^{n} y_i = 0, \qquad \sum_{i=1}^{n} x_{ij} = 0, \qquad \sum_{i=1}^{n} x_{ij}^2 = 1 \qquad \text{for } j = 1, 2, \ldots, p$$

The LARS algorithm proposed by Efron et al. [3] is a less greedy forward model selection procedure. At the beginning of LARS, a predictor enters the model if its absolute correlation with the response is the largest one among all the predictors. The coefficient of this predictor grows in its ordinary least squares direction until another predictor has the same correlation with the current residual (i.e. equal-angle). Next, both coefficients of the two selected predictors begin to move along their ordinary least squares directions until a third predictor has the same correlation with the current residual as the first two. The whole process continues until all predictors enter the model. In each step, one variable adds into the model and the solution paths, which are the coefficient estimators as functions of the tuning parameter (defined later in Formula 1.1), are extended in a piecewise linear fashion. After all variables enter the model, the whole LARS solution paths complete.

For data with dependent structures, we propose gLars and gRidge algorithms that construct groups simultaneously along the variable selection process. We first give a grouping definition. Predictors form a group if they satisfy both of the two criterions:

- They are highly correlated with the response variable (or current residual);
- They are highly correlated with each other.

The correlation thresholds for the two criterions will be determined from the data. For instance, the threshold of the first criterion is suggested to be the 75th percentile of all correlations (in absolute value) between the current residual and unselected predictors. The correlation threshold (absolute value) for the second criterion is either the 75th percentile of all pairwise correlations among the predictors or chosen from a set of grids, for example 0.9, 0.8, 0.7, 0.6. An important difference of the proposed method from the standard forward selection procedures is that our variable selection criterion has two components hence is defined by a region in the 2-dimensional space, (t_1, t_2) in Step 3 in the following algorithm. In addition, the first requirement of selecting a variable highly correlated with the response variable is not affacted by collinearity among predictors.

In the gLars algorithm, we start as LARS to select a predictor which has the largest correlation with the response. We call this predictor a "leader element". We then build a group based on this leader element and the current residual according to the two grouping criterions. Note that both criterions

have to be satisfied when selecting a variable into a group. Once a group has been constructed, it will be represented by a unique direction in R^n as the linear combination of the ordinary least squares directions of all variables in the group. Next, we choose another leader element, analogous to the equal-angle requirement of the LARS algorithm. A new group is formed again following the grouping definition. We refine the solution paths in a piecewise linear format. The whole process continues until all predictors enter the model. The detailed algorithm is described below.

1. Initialization: Set the step index $k = 1$, $\beta^{[0]} = 0$, residual $r^{[0]} = Y$, active set $A_0 = \emptyset$, inactive set $A_0^C = \{X_1, X_2, \ldots, X_p\}$.
2. Identify the leader predictor $\boldsymbol{x}$ for the first group, where $\boldsymbol{x} = \text{argmax}_{\boldsymbol{x}_i} |\boldsymbol{x}_i' r^{[k-1]}|$, $\boldsymbol{x}_i \in A_{k-1}^C$.
3. Construct the group G_k with the leader predictor $\boldsymbol{x}$ from Step 2 according to the two criterions: $\boldsymbol{x}_j' r^{[k-1]} > t_1$ and $\boldsymbol{x}_j' \boldsymbol{x} > t_2$, $\boldsymbol{x}_j \in A_{k-1}^C$, where $t_1 = 0.75$th percentile of all correlations between $\boldsymbol{x}_j$ and $r^{[k-1]}$, and $t_2 = t \in \{0.9, 0.8, 0.7, 0.6\}$.
   Set $A_k = A_{k-1} \cup G_k$, $A_k^C = A_{k-1}^C \setminus G_k$.
4. Compute the current direction γ with components
$$\gamma_{A_k} = (X_{A_k}' X_{A_k})^{-1} X_{A_k}' r^{[k-1]}, \quad \gamma_{A_k^C} = 0$$
   where X_{A_k} denotes the matrix comprised of the columns of $\mathbf{X}$ corresponding to A_k.
5. Calculate how far the gLars algorithm progresses in direction γ. It divides into two small steps:
   Find $\boldsymbol{x}_{j'}$ in A_k^C which corresponds to the smallest $\alpha \in (0, 1]$ such that
$$\frac{||\boldsymbol{X}_{G_j}' (r^{[k-1]} - \alpha \boldsymbol{X}\gamma)||_{L_1}}{p_j} = |\boldsymbol{x}_{j'}' (r^{[k-1]} - \alpha \boldsymbol{X}\gamma)|,$$
   where G_j is a group from A_k, p_j is the number of variables in group G_j, and $||.||_{L_1}$ represents the sum of absolute values.
   Justification. As in Step 3, find the group with the leader predictor $\boldsymbol{x}_{j'}$ selected above and denote the group as $\boldsymbol{X}_{G_{j'}}$. Recalculate $\alpha \in (0, 1]$ for this selected new group such that
$$\frac{||\boldsymbol{X}_{G_j}' (r^{[k-1]} - \alpha \boldsymbol{X}\gamma)||_{L_1}}{p_j} = \frac{||\boldsymbol{X}_{G_{j'}}' (r^{[k-1]} - \alpha \boldsymbol{X}\gamma)||_{L_1}}{p_{j'}}$$
   Update $\beta^{[k]} = \beta^{[k-1]} + \alpha\gamma$, $r^{[k]} = Y - X\beta^{[k]}$.
6. Update k to $k + 1$, and $A_k = A_{k-1} \cup G_{j'}$, $A_k^C = A_{k-1}^C \setminus G_{j'}$.
7. If $A_k^C \neq \emptyset$, return to Step 4. Otherwise, set γ, β and r to be the OLS solutions and stop.

Ordinary least squares would perform poorly when the correlations among the predictors are high and/or the noise level is high. Since both LARS and

gLars move towards ordinary least squares direction in each step, they face the same shortage. Ridge estimators, on the other hand, perform better in this situation. We propose a gRidge algorithm, which moves towards ridge estimator direction in each step. The relationship between ridge estimator $\hat{\beta}(\lambda)$ and ordinary least squares estimator $\hat{\beta}$ can be shown as

$$\begin{aligned}\hat{\beta}(\lambda) &= (X'X + \lambda I)^{-1}X'Y = (X'X(I + \lambda(X'X)^{-1}))^{-1}X'Y \\ &= (I + \lambda(X'X)^{-1})^{-1}\hat{\beta} = C\hat{\beta},\end{aligned}$$

where $C = (I + \lambda(X'X)^{-1})^{-1}$ and λ is the ridge parameter. The gRidge algorithm is thus a simple modification of the gLars algorithm. When a group is constructed, gRidge represents the group by a unique direction from the linear combination of the ridge directions of all variables in the group. The variable coefficients are moving towards the ridge directions.

As we run simulations, we notice that gRidge outperforms other methods in terms of relative prediction errors (RPEs, defined below in the simulations). However, this method is limited by its comparably larger false positives due to an over-grouping effect. We propose to add a hard threshold δ to gRidge estimators so that small (but nonzero) coefficients will be removed, i.e. $\tilde{\beta_j} = \hat{\beta_j} I(\hat{\beta_j} > \delta)$. Based on simulations, we define a threshold $\delta = \sqrt{\sigma \log(p)/n}$. Hence smaller error term, smaller number of predictors, or larger sample size give smaller threshold. We name the modified gRidge algorithm gRidge_new, with this hard threshold filtering. Simulation studies show that gRidge_new not only preserves low RPE but also greatly reduces false positives.

Both gLars and gRidge produce the entire piecewise linear solution paths as LARS does. Groups of variables are selected when we stop the paths after a certain number of steps. The number of step k is the tuning parameter. Equivalently, we may use a tuning parameter as the fraction of the L_1 norm of the coefficients

$$s = \Sigma_{j\_selected} ||\hat{\beta}_j||_{L_1} / \Sigma_j ||\hat{\beta}_j||_{L_1}. \tag{1.1}$$

For gLars, s (or k) is the only tuning parameter. It is determined by a standard five-fold cross-validation (CV). For gRidge, there are two tuning parameters, the ridge parameter λ in addition to s (or k). Similar to elastic net, we cross-validate on two dimensions. First, we choose a grid for λ, say $\{0.01, 0.1, 1, 10, 100, 1000\}$. Then for each λ, gRidge produces the entire solution path. The parameter s (or k) is selected by five-fold CV. At the end, we choose the λ value which gives the smallest CV error.

Simulation studies

Simulation studies are used to compare the proposed gLars and gRidge with ordinary least squares, ridge regression, LARS and elastic net. The simulated data are generated from the true model $\mathbf{y} = \mathbf{X}\beta + \sigma\epsilon$, $\epsilon \sim N(0, 1)$. We have studied many examples for different scenarios but only present four here

due to the space limit. For each example, we simulate 100 data sets. Each data set consists of a training set and a test set. The tuning parameters are selected on the training set by five-fold cross-validation. The variable selection methods are compared in terms of relative prediction error (RPE) [26] and selection accuracy on the test set. The relative prediction error is defined as $RPE = (\hat{\beta} - \beta)^T \Sigma (\hat{\beta} - \beta)/\sigma^2$, where Σ is the population covariance matrix of $\mathbf{X}$. The four scenarios are given by:

1. Example 1 (adopted from [27]), there are 100 and 200 observations in the training and test sets respectively. The true parameter $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)$ and $\sigma = 3$. The pairwise correlation between $\mathbf{x}_i$ and $\mathbf{x}_j$ is set to be $corr(\mathbf{x}_i, \mathbf{x}_j) = 0.5^{|i-j|}$. This example creates a sparse model with a few large effects and the covariates have first-order autoregressive correlation.
2. Example 2 (adopted from Daye and Jeng unpublished), we simulate 100 and 400 observations in the training and test sets respectively. We set the true parameters as

$$\beta = (\underbrace{3, \ldots, 3}_{15}, \underbrace{1.5, \ldots, 1.5}_{5}, \underbrace{0, \ldots, 0}_{20})$$

   and $\sigma = 6$. The predictors are generated as:

$$\begin{aligned} \mathbf{x}_i = Z + \epsilon_i^x,\ Z \sim N(0,1),\ i = 1, \ldots, 15, \\ \mathbf{x}_i \sim N(0,1), i.i.d.,\ i = 16, \ldots, 40, \end{aligned}$$

   where ϵ_i^x are independent identically distributed $N(0, 0.01)$, $i = 1, \ldots, 15$. This example creates one group from the first 15 highly correlated covariates. The next five covariates are independent but provide signals on the response variable.
3. Example 3 (adopted from [27]), we simulate 100 and 400 observations in the training and test sets respectively. We set the true parameters as

$$\beta = (\underbrace{3, \ldots, 3}_{15}, \underbrace{0, \ldots, 0}_{25})$$

   and $\sigma = 15$. The predictors are generated as:

$$\begin{aligned} &\mathbf{x}_i = Z_1 + \epsilon_i^x, && Z_1 \sim N(0,1), && i = 1, \ldots, 5, \\ &\mathbf{x}_i = Z_2 + \epsilon_i^x, && Z_2 \sim N(0,1), && i = 6, \ldots, 10, \\ &\mathbf{x}_i = Z_3 + \epsilon_i^x, && Z_3 \sim N(0,1), && i = 11, \ldots, 15, \\ &\mathbf{x}_i \sim N(0,1), i.i.d., && i = 16, \ldots, 40, \end{aligned}$$

   where ϵ_i^x are independent identically distributed $N(0, 0.01)$, $i = 1, \ldots, 15$. There are three equally important groups with five members in each. There are also 25 noise variables.

4. Example 4, we simulate 100 and 200 observations in the training and test sets respectively. We set the true parameters as

$$\beta = (\underbrace{3,3,3,0,0}_{5},\underbrace{3,3,3,0,0}_{5},\underbrace{3,3,3,0,0}_{5},\underbrace{0,\ldots,0}_{25})$$

The predictors and the error terms are the same as in Example 3. There are also three equally important groups with five members in each of them. However, in each group, there are two noise variables, which have no effect on the response variable but are highly correlated with the other three important variables. There are totally 31 noise variables.

Table 1.1 summarizes the prediction results. The median RPE from 100 simulations is reported. The smallest RPE is emphasized in italic font, which indicates the most accurate method for each example. We also report the median number of nonzero coefficients versus the median number of zero coefficients mis-specified as nonzero, which imply the true positive and false positive of a method. The simulation results indicate that LARS tends to produce very sparse models but does not work for collinearity. Elastic net improves LARS when predictors are correlated. But elastic net misses the five true signals with the small coefficients 1.5 in Example 2. The first proposed method gLars improves elastic net in terms of true positives, especially in Example 2 and 4. gRidge and gRidge_new produce the smallest RPEs in all the examples and therefore are the most accurate models in terms of prediction. We also notice that while preserving the large coefficients close to the true coefficients, gRidge tends to select more variables than elastic net, due to its over grouping effect. After we add a hard threshold to gRidge, the gRidge_new estimators achieves the best performance.

Table 1.1. *Median relative prediction errors (RPE) and median number of nonzero coefficients / median number of zero coefficients mis-specified as nonzero coefficients for the four examples based on 100 replications. The best results are emphasized in italic fonts.*

Methods	Example 1		Example 2		Example 3		Example 4	
OLS	0.5843	3/5	0.6364	20/20	0.6390	15/25	0.6458	9/31
Ridge	0.2832	3/5	0.2519	20/20	0.0993	15/25	0.1971	9/31
LARS	0.4640	***3/0***	0.3208	12/1	0.1620	6/2	0.1200	3/6
Elastic net	***0.1714***	3/1	0.2587	15/2	0.0800	***15/1***	0.1110	*7/8*
gLars	0.2616	3/2	0.4235	***20/3***	0.2220	***15/3***	0.2121	***9/8***
gRidge	0.1806	3/3	***0.1963***	20/12	***0.0700***	15/10	***0.0700***	9/13
gRidge_new	0.1816	3/1	***0.1988***	***19/3***	***0.0700***	***15/2***	***0.0690***	***9/8***

1.4 Unbiased variable selection via SCAD_ℓ2

SCAD is proposed by Fan and Li [4] as a variable selection method via penalized least squares. The SCAD penalty function is specially defined to satisfy three properties for the coefficient estimators: unbiasedness, sparsity and continuity. To address the challenges of genomic data analysis, we add another property of grouping effect and propose a new penalty function named SCAD_ℓ2. Instead of defining grouping criterions as we have proposed in gLars and gRidge, we achieve the grouping effect in SCAD_ℓ2 through a strictly convex penalty function, which is a linear combination of the L_2 norm and the SCAD function. More specifically, we propose a naive SCAD_ℓ2 estimator $\hat{\boldsymbol{\beta}}_{naive}$ as the minimizer of the penalized least squares function

$$Q(\boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \sum_{j=1}^{p} f_{\lambda_1}(\beta_j) + \lambda_2\|\boldsymbol{\beta}\|^2 \tag{1.2}$$

where $||\beta||^2 = \sum_{j=1}^{p} \beta_j^2$ and $f_\lambda(\theta)$ is the SCAD function defined as

$$f_\lambda(\theta) = \begin{cases} \lambda|\theta|, & \text{if } 0 \le |\theta| < \lambda, \\ -\dfrac{\theta^2 - 2a\lambda|\theta| + \lambda^2}{2(a-1)}, & \text{if } \lambda \le |\theta| < a\lambda, \\ (a+1)\lambda^2/2 & \text{otherwise.} \end{cases}$$

Here a is a real number larger than 2. Under the condition that the columns of $\mathbf{X}$ are orthonormal, we can obtain the explicit expression of the naive SCAD_ℓ2 estimator $\hat{\boldsymbol{\beta}}_{naive}$. Specifically, $\hat{\boldsymbol{\beta}}_{naive} = \hat{\boldsymbol{\beta}}_{OLS}/(1+2\lambda_2)$ for large $|\hat{\boldsymbol{\beta}}_{OLS}|$ and hence a biased estimator. The true SCAD_ℓ2 estimator $\hat{\boldsymbol{\beta}}_{SCAD\_\ell2}$ is defined as $\hat{\boldsymbol{\beta}}_{SCAD\_\ell2} = (1+2\lambda_2)\hat{\boldsymbol{\beta}}_{naive}$, to attain unbiasedness.

For a general predictor matrix $\mathbf{X}$ not orthonormal, including situations with correlated predictors, SCAD_ℓ2 estimator is defined by the naive SCAD_ℓ2 estimator multiplying a matrix depending on λ_2 and the covariance matrix of $\mathbf{X}$. We can show that the SCAD_ℓ2 estimator satisfies four properties.

1. *Unbiasedness*: $\hat{\beta}_{j,SCAD\_\ell2} = \hat{\beta}_{j,OLS}$ for large components of $|\hat{\beta}_{j,OLS}|$;
2. *Sparsity*: $\hat{\beta}_{j,SCAD\_\ell2} = 0$ when $|\hat{\beta}_{j,OLS}|$ is small;
3. *Continuity*: $\hat{\boldsymbol{\beta}}_{SCAD\_\ell2}$ is a continuous function with respect to $\hat{\boldsymbol{\beta}}_{OLS}$;
4. *Grouping effect*: Two coefficients $\hat{\beta}_{j,SCAD\_\ell2}$ and $\hat{\beta}_{i,SCAD\_\ell2}$ tend to be equal if the two respective variables $\mathbf{x}_j$ and $\mathbf{x}_i$ are highly correlated.

Following Fan and Li's [4] discussion, for sparsity, it is sufficient to prove that $\min_{\theta\neq0}\{(|\theta| + p'_\lambda(|\theta|)\} > 0$; and for continuity, it is sufficient to prove that $\text{argmin}_\theta\{|\theta| + p'_\lambda(|\theta|)\} = 0$, where $p_\lambda(|\theta|)$ is the penalty function of SCAD_ℓ2 as defined by the last two terms in Formula (1.2). To prove the grouping effect, we use the fact that the penalty function is strictly convex. In addition, let $\hat{\beta}_{i,naive}$ and $\hat{\beta}_{j,naive}$ denote the i-th and j-th elements of $\hat{\boldsymbol{\beta}}_{naive}$ respectively. Define

$D(i,j) = |\hat{\beta}_{i,naive} - \hat{\beta}_{j,naive}|/|\mathbf{y}|$. The following theorem implies that strongly correlated variables will be in or out of model together through SCAD_ℓ2.

Theorem 1. *Assume $\hat{\beta}_{i,naive} \cdot \hat{\beta}_{j,naive} > 0$ and a regularity condition for λ_2, we have*

$$D(i,j) \leq C \cdot \sqrt{2(1-\rho)}$$

where C is a constant that may depend on λ_2, and the sample correlation $\rho = \mathbf{x}_i^T \mathbf{x}_j$.

The quantity $D(i,j)$ measures the difference between the coefficients of two predictors $\mathbf{x}_i$ and $\mathbf{x}_j$. In an extreme case when the absolute value of the correlation between the two predictors are close to 1, Theorem 1 guarantees that the coefficients of the two predictors will be almost identical except the sign difference. In other words, naive SCAD_ℓ2 has the group effect. The true SCAD_ℓ2 estimator equals a scalar multiplying the naive SCAD_ℓ2 estimator. Therefore, SCAD_ℓ2 has the grouping effect as well.

We establish asymptotic theories for SCAD_ℓ2, when the number of variables p is fixed and the sample size n goes to infinity. Note that the larger n becomes, the heavier the least squares part in Formula (1.2) weighs. As an adjustment, we consider the following penalized least squares function

$$Q(\boldsymbol{\beta}) = \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + n\sum_{j=1}^{p} f_{\lambda_1}(\beta_j) + n\lambda_2||\boldsymbol{\beta}||^2.$$

Let $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_p^*)^T$ denote the true value of $\boldsymbol{\beta}$ in the linear regression problem. Without loss of generality, we assume the first p_1 elements $\beta_1^*, \ldots, \beta_{p_1}^*$ are nonzeros and the rest $p - p_1$ elements are zeros. Denote $\boldsymbol{\beta}_N^* = (\beta_1^*, \ldots, \beta_{p_1}^*)^T$ and $\boldsymbol{\beta}_Z^* = (\beta_{p_1+1}^*, \ldots, \beta_p^*)^T$. We use $\hat{\boldsymbol{\beta}}(n) = (\hat{\beta}_1(n), \ldots, \hat{\beta}_p(n))^T$ to denote the minimizer of $Q(\boldsymbol{\beta})$, and denote $\hat{\boldsymbol{\beta}}_N(n) = (\hat{\beta}_1(n), \ldots, \hat{\beta}_{p_1}(n))^T$ and $\hat{\boldsymbol{\beta}}_Z = (\hat{\beta}_{p_1+1}(n), \ldots, \hat{\beta}_p(n))^T$ as the estimators of the nonzero and zero coefficients respectively. We rewrite λ_1 and λ_2 as $\lambda_1(n)$ and $\lambda_2(n)$ to emphasize that they vary as n changes. The following asymptotic theorems hold.

Theorem 2. *(Estimation consistency) If $\lambda_1(n) \to 0$ and $\sqrt{n}\lambda_2(n) \to 0$ when $n \to \infty$, then there exists a local minimizer $\hat{\boldsymbol{\beta}}(n)$ of $Q(\boldsymbol{\beta})$ such that $\|\hat{\boldsymbol{\beta}}(n) - \boldsymbol{\beta}^*\| = \mathcal{O}_p(n^{-1/2})$.*

Theorem 3. *(Selection consistency) If $\lambda_1(n) \to 0$, $\sqrt{n}\lambda_1(n) \to +\infty$, and $\sqrt{n}\lambda_2(n) \to 0$ as $n \to +\infty$, then $\lim_{n\to\infty} Prob\left\{\hat{\boldsymbol{\beta}}_Z(n) = 0\right\} = 1$.*

Theorem 4. *(Oracle property) If $\lambda_1(n) \to 0$, $\sqrt{n}\lambda_1(n) \to +\infty$, and $\lambda_2(n) \to 0$, $\sqrt{n}\lambda_2(n) \to 0$ as $n \to +\infty$, then the root-n consistent local minimizer $\hat{\boldsymbol{\beta}}(n) = \begin{pmatrix} \hat{\boldsymbol{\beta}}_N(n) \\ \hat{\boldsymbol{\beta}}_Z(n) \end{pmatrix}$ satisfies the following with probability tending to 1:*

1. *Sparsity:* $\hat{\boldsymbol{\beta}}_Z(n) = \mathbf{0}$,
2. *Asymptotic normality:* $\sqrt{n}\left(\hat{\boldsymbol{\beta}}_N(n) - \boldsymbol{\beta}_N^*\right) \xrightarrow{D} \mathcal{N}(\mathbf{0}, \sigma^2 \Sigma_N^{-1})$,

where Σ_N *is the covariance matrix of the first* p_1 *predictors.*

The basic ideas in the proofs of these asymptotic theorems include applying Taylor expansion of the penalized least squares function $Q(\boldsymbol{\beta})$ and the law of large number or the central limit theorem. These results build strong theoretical backgrounds for the proposed group variable selection method. To implement SCAD_ℓ2, we use local quadratic approximation similar to the algorithm of SCAD.

1.5 Applications in genomic data analysis

1.5.1 SNP data analysis

The proposed group variable selection methods are particularly useful in high dimensional data with dependent structures, for instance, gene expression microarray data, genetic variation SNP data, and transcription factor binding ChIP-chip data. Statisticians have been playing important roles in gene expression microarray data analyses in the past decade. With the advance of SNP techniques and the stride in SNP detections and the international HapMap project, SNP data analysis becomes another interesting field for statisticians to explore.

As an initial example, we study genetic variation (SNPs) for human gene expression. Natural variation in the baseline expression of many genes can be considered as heritable traits. Morley et al. [10] have collected microarray and SNP data to localize the DNA variants that contribute to the expression phenotypes. The data consists of 14 families with 56 unrelated individuals (the grandparents). There are $\sim 8,500$ genes on the array and 2,756 SNP markers genotyped for each individual. The expression level of a given gene is the response variable and the 2,756 SNP markers are the predictors in our model. We apply the proposed group selection methods to search for optimal set of SNPs for gene ICAP-1A, which is the top gene in Morley et al.'s [10] Table 1 with the strongest linkage evidence.

We code each SNP as 0, 1, 2 for wild type homozygous, heterozygous, and mutation (rare) homozygous genotypes respectively according to the genotype frequency. We first screen data to exclude SNPs that have a call rate less than 95% or minor allele frequency less than 2.5%. The number of SNPs is reduced to 1,739 after screening. Then we select 500 most "variable" SNPs as the potential predictors. The variability of a SNP is measured by its sample variance.

We split data into the training set with 42 observations and the test set with 14 observations. Model fitting and choices of the tuning parameters are

based on the training set. The first grouping criterion is set up to be the 75th percentile of all correlations between x and y. The second grouping criterion requires the correlation with the leader element to be greater than 0.6. The prediction error (residual sum of squares) is evaluated on the test data. Table 1.2 shows that gLars and gRidge have lower prediction errors than LARS and elastic net with about 30 SNPs selected for gene ICAP-1A. We notice that R^2 of the LARS fitted model with 24 SNP covariates is 0.779, which supports the hypothesis of Morley et al.'s [10] study that gene expression phenotypes are controlled by genetic variants. On the other hand, the coefficient estimators of all SNP covariates are very small, in the scale of 0.01, suggesting small additive effects of multiple SNPs.

Table 1.3 lists the first 12 steps that the predictors are selected in each algorithm. The numbers in the table are the indices of the variables. For instance, 458 in Step 1 means that variable x_{458} enters the model at the first step. At Step 3, gLars and gRidge depart from LARS and elastic net due to the grouping effect. The first group consists of two SNPs, x_{321} SNP rs1004620 and x_{131} SNP rs1868237. The correlation of these two variables is 0.65. The two SNPs are in chromosome 3 with 14K base pairs apart. They are in an intergenic region. The two closest genes have no functional annotation. Another group consisting of two SNPs x_{30} rs1882600 and x_{27} rs1001396 are selected by gLars and gRidge at Step 11. These two SNPs are in chromosome 7 with over 2.5 million base pairs apart. SNP x_{30} rs1882600 is in an intergenic region, whereas x_{27} rs1001396 resides in the gene FOXK1. According to Swiss-Prot functional annotation, FOXK1 is a transcriptional regulator that binds to the upstream of myoglobin gene.

Table 1.2. Test prediction errors of Lasso, elastic net, gLars and gRidge for the SNP data.

Methods	Test prediction error	number of genes	tuning parameter s
LARS	1.569	24	0.4343
Elastic net ($\lambda_2 = 0.01$)	1.872	23	0.2323
gLars	1.360	33	0.3838
gRidge ($\lambda_2 = 0.01$)	1.531	28	0.4141

Our results suggest more SNP associations with a gene expression phenotype than the simple linkage analysis. For example, variables x_{240} SNP rs1446297 and x_{76} SNP rs933602 are jointly selected as important covariates for the expression of ICAP-1A according to all four variable selection methods. However, they are not significant in simple regression analysis with a p-value cutoff 0.01. These two SNPs locate in the same chromosome as ICAP-1A (chromosome 2) with nearly 27 million base pairs and 220 million base pairs respectively away from ICAP-1A. Fig. 1.1 shows their locations and re-

Table 1.3. The first 12 steps of predictors selected by Lasso, elastic net, gLars and gRidge for the SNP data.

Methods	LASSO	elastic net	gLars	gRidge
Step 1	458	458	458	458
Step 2	481	481	481	481
Step 3	321	321	321,131	321,131
Step 4	287	287	287	287
Step 5	240	240	240	240
Step 6	76	76	406	406
Step 7	406	131	76	76
Step 8	131	406	102	102
Step 9	345	345	345	345
Step 10	102	102	498	498
Step 11	498	498	30,27	30,27
Step 12	30	30	167	167

gression plots. SNP rs1446297 is in the promotor region (about 200 base pairs upstream) of gene FAM82A1. SNP rs933602 is in the promotor region (about 300 base pairs upstream) of gene DNER. The significant effects of these two SNPs on gene ICAP-1A may suggest associations among the corresponding genes.

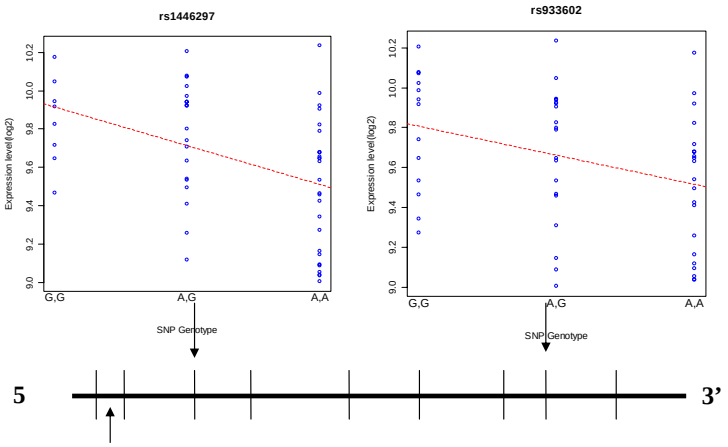

Fig. 1.1. Regression of the expression phenotype of ICAP-1A on the two nearby SNPs.

1.5.2 Gene expression data analysis

We apply SCAD_ℓ2 to a genomic data set in a study of rat eye disease by Scheetz et. al [12]. The data set consists of 120 rats generated from two highly imbred parental rat strains. Among 31,000 genes that express in eyes, we are interested in finding relevant genes which are correlated with gene TRIM32 known to cause the eye disease Bardet-Biedl syndrome.

We first exclude genes that lack sufficient variation to result in 3,000 most variable genes. Then we order the 3,000 genes based on their absolute correlations with gene TRIM32 from the largest to the smallest. The top 90 genes are selected as the potential preditors for the response variable gene TRIM32.

Next, we apply SCAD_ℓ2 to select groups of genes that may influence the expression of TRIM32. The 120 rats samples are randomly split into a training set with 100 samples and a test set with 20 samples. A regression model is fitted in the training set. A generalized cross validation method is used to decide the tuning parameters, $(a, \lambda_1, \lambda_2)$, on the training data. Model prediction accuracy is measured by the mean squared error (MSE) on the test set. We compare the prediction accuracy of SCAD_ℓ2 with those of SCAD, Lasso, and elastic net. The whole processes are repeated 100 times. The median MSE and the median number of selected genes are shown in Table 1.4. Elastic net and Lasso produce more sparse models with fewer numbers of predictors than those of SCAD and SCAD_ℓ2. Elastic net and Lasso also provide similar results, without an obvious group effect in elastic net. Although elastic net and Lasso give small MSEs, their sparse set of variables may miss important signals, due to the fact that the methods may only select one variable from a group. On the other hand, SCAD_ℓ2 performs better than its non-group effect counterpart SCAD. Specifically, SCAD_ℓ2 outperforms SCAD by offering a moderate size model (with 25 predictors) and 13% reduction of MSE.

Table 1.4. Comparison of SCAD_ℓ2 with SCAD, Lasso, and elastic net based on 100 simulations in the analysis of gene expression data of rat eye disease.

Methods	Median MSE (SE)	Median nonzero
Lasso	0.017 (0.0012)	13
Elastic net	0.016 (0.0015)	10
SCAD	0.0439 (0.0043)	39
SCAD_ℓ2	0.0383 (0.0043)	25

1.6 Discussion

Although large scale genomic data have been routinely created in biomedical research, extracting useful information from the data remains a challenge. Available statistical and computational tools encounter major difficulties of high dimensionality and complicated dependence in the data. This chapter discusses variable selection approaches for high dimensions, and more importantly new ideas of group variable selection. The group information naturally embedded in biological systems or pathways helps to enhance signals in analysis of genomic data.

Traditional forward selection is a heuristic approach, not guaranteeing an optimal solution. LARS a less greedy version of traditional forward selection method, however, is shown by Efron et al. [3] to be closely related to Lasso, which possesses optimal properties under appropriate conditions [8, 24, 25]. Our proposed gLars and gRidge take advantage of the LARS procedure while aiming at group selections for dependent data. The methods do not require prior information on the underlying group structures but construct groups along the selection procedure. Our grouping criterions consider the joint information of x and y therefore better fit the context of variable selection than standard clustering on x alone. On the other hand, any prior information on the model or groups can be easily incorporated into the algorithms of gLars and gRidge by manually selecting certain variables at specific steps. The current methods may be improved by exploring different thresholds (t_1, t_2) in the grouping definition.

SCAD_$\ell 2$ is a combination of the unbiased approach SCAD and the ridge regression. It is not computationally efficient as the forward procedure but possess good properties in terms of coefficient estimation. One of our future works is to extend the proposed group selection methods to general regression models, where y may depend on x through any nonlinear function. The proposed methods are more appropriate than other variable selection algorithms for data with complicated dependent structures.

Acknowledgments

This work is supported by National Science Foundation Grant DMS-0604776.

References

1. T.L. Bailey and C. Elkan. The value of prior knowledge in discovering motifs with MEME. *Proc Int Conf Intell Syst Mol Biol*, **3**: 21–9, 1995.
2. D. Donoho and I. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, **81**: 425–455, 1994.
3. B. Efron, I. Johnstone, T. Hastie, and R. Tibshirani. Least angle regression. *Annals of Statistics*, **32**: 407–499, 2004.

4. J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, **96**: 1348–1360, 2001.
5. X. Huo and X. Ni. When do stepwise algorithms meet subset selection crieteria? *The Annals of Statistics*, **35**(2): 870–887, 2006.
6. N-K. Kim and J. Xie. Protein Multiple Alignment Incorporating Primary and Secondary Sequence Information. *Journal of Computational Biology*, **13**(10): 1735–1748, 2006.
7. Y. Kim, H. Choi, and H-S. Oh. Smoothly Clipped Absolute Deviation on High Dimensions. *Journal of the American Statistical Association*, **103**(484): 1665–1673, 2008.
8. K. Knight and W. Fu. Asymptotics for lasso-type estimators. *Technometrics*, **12**(1): 69–82, 2000.
9. C. Leng, Y. Lin, and G. Wahba. A note on the lasso and related procedures in model selection. *Statisica Sinica*, **16**: 1273–1284, 2006.
10. M. Morley, C.M. Molony, T.M. Weber, J.L. Devlin, K.G. Ewens, R.S. Spielman, and V.G. Cheung. Genetic analysis of genome-wide variation in human gene expression. *Nature*, **430**: 743–747, 2004.
11. T. Park and G. Casella. The Bayesian Lasso. *Technical report, University of Florida, Gainesville, FL*, 2008.
12. T.E. Scheetz, K.Y. Kim, R.E. Swiderski, A.R. Philp, T.A. Braun, K.L. Knudtson, A.M. Dorrance, G.F. DiBona, J. Huang, T.L. Casavant, V.C. Sheffield, and E.M. Stone. Regulation of Gene Expression in the Mammalian Eye and Its Relevance to Eye Disease. *Proceedings of the National Academy of Sciences*, **103**: 14429–14434, 2006.
13. G.A. Seber and J. Alan. Linear regression analysis (2rd ed). *Wiley-interscience*, 2003.
14. M. Segal, K. Dahlquist and B. Conklin. Regression approach for microarray data analysis. *Journal of Computational Biology*, **10**: 961–980, 2003.
15. I. Simon, J. Barnett, N. Hannett, C.T. Harbison, N.J. Rinaldi, T.L. Volkert, J.J. Wyrick, J. Zeitlinger, D.K. Gifford, T.S. Jaakkola, and R.A. Young. Serial regulation of transcriptional regulators in the yeast cell cycle. *Cell*, **106**: 697–708, 2001.
16. P.T. Spellman, G. Sherlock, M.Q. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein, B. Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell*, **9**(12): 3273–97, 1998.
17. R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal Royal Statistical Society B*, **58**: 267–288, 1996.
18. D.G. Wang, J. Fan, C. Siao, A. Berno, P. Young, R. Sapolsky, G. Ghandour, N. Perkins, E. Winchester, J. Spencer, L. Kruglyak, L. Stein, L. Hsie, T. Topaloglou, E. Hubbell, E. Robinson, M.S. Morris, N. Shen, D. Kilburn, J. Rioux, C. Nusbaum, S. Rozen, T.J. Hudson, R. Lipshutz, M. Chee, and E.S. Lander. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science*, **280**: 1077–1082, 1998.
19. J. Wu and J. Xie. Computation-Based Discovery of Cis-Regulatory Modules by Hidden Markov Models. *Journal of Computational Biology*, **15**(3): 279–290, 2008.
20. J. Xie and N-K. Kim. Bayesian Models and Markov Chain Monte Carlo Methods for Protein Motifs with the Secondary Characteristics. *Journal of Computational Biology*, **12**(7): 952–970, 2005.

21. N. Yi and S. Xu. Bayesian Lasso for quantitative trait loci mapping. *Genetics*, **179**: 1045–1055, 2008.
22. M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of Royal Statistical Society B*, **68**(1): 49–67, 2006.
23. L. Zeng, J. Wu, and J. Xie. Statistical methods in integrative analysis for gene regulatory modules. *Statistical Applications in Genetics and Molecular Biology*, **7**(1): Article 28, 2008, http://www.bepress.com/sagmb/vol7/iss1/art28 .
24. C. Zhang and J. Huang. The sparsity and bias of the lasso selection in high-dimensional linear regression. *The Annals of Statistics*, **36**(4): 1567–1594, 2008.
25. P. Zhao and B. Yu. On model seleciton consistency of lasso. *Journal of Machine Learning Research*, **7**: 2541–2563, 2006.
26. H. Zou. The adpative lasso and its oracle properties. *Journal of the American Statistical Association*, **101**(476): 1418–1429, 2006.
27. H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of Royal Statistical Society B*, **67**: 301–320, 2005.