Regularization Path Algorithms for Detecting Gene Interactions

Mee Young Park * Trevor Hastie †

July 16, 2006

Abstract

In this study, we consider several regularization path algorithms with grouped variable selection for modeling gene-interactions. When fitting with categorical factors, including the genotype measurements, we often define a set of dummy variables that represent a single factor/interaction of factors. Yuan & Lin (2006) proposed the group-Lars and the group-Lasso methods through which these groups of indicators can be selected simultaneously. Here we introduce another version of group-Lars. In addition, we propose a path-following algorithm for the group-Lasso method applied to generalized linear models. We then use all these path algorithms, which select the grouped variables in a smooth way, to identify gene-interactions affecting disease status in an example. We further compare their performances to that of L_2 penalized logistic regression with forward stepwise variable selection discussed in Park & Hastie (2006*b*).

1 Introduction

In this paper, we propose using regularization path algorithms with grouped variable selection for fitting a binary classification model with genotype data and for identifying significant interaction effects among the genes. We implement the group-Lars and the group-Lasso methods introduced in Yuan & Lin (2006), and we also introduce a different version of the group-Lars method. To fit the nonlinear regularization path for group-Lasso, we develop an algorithm based on the *predictor-corrector* scheme as in Park & Hastie (2006*a*). Our group-Lasso algorithm can use any loss function in the family of generalized linear models. We regard this strategy of using path algorithms as a compromise between our two earlier studies, described next.

*Ph.D. candidate, Department of Statistics, Stanford University, CA 94305. mypark@stat.stanford.edu, *tel* 16507042581

†Professor, Department of Statistics and Department of Health Research & Policy, Stanford University, CA 94305. hastie@stat.stanford.edu

In Park & Hastie (2006*b*), we proposed using forward stepwise logistic regression to fit gene-interaction models. The forward stepwise procedure is a traditional variable selection mechanism; we made a set of dummy variables for each factor/interaction of factors and added/deleted a group at a time. In many studies dealing with gene-interactions, logistic regression has been criticized for its limited applicability: a small sample size prohibits high-order interaction terms, and a sparse distribution of the genotypes (for a factor/interaction of factors) is not tolerated as it results in zero column inputs. However, by modifying logistic regression with a slight penalty on the L_2 norm of the coefficients, we could fit a stable gene-interaction model. Although the forward stepwise method is a greedy approach, we showed that it successfully selected significant terms and achieved a reasonable prediction accuracy when combined with the L_2 penalized logistic regression.

A smoother way to select the features that has been explored extensively in many regression/classification settings is to incorporate an L_1 regularization constraint. Tibshirani (1996) first introduced Lasso, a regression method that minimizes the sum of squared error loss subject to an L_1 norm constraint on the coefficients. Various applications of the L_1 norm penalty can be found for example in Genkin, Lewis & Madigan (2004), Tibshirani (1997), or Zhu, Rosset, Hastie & Tibshirani (2003). Efron, Hastie, Johnstone & Tibshirani (2004) proposed the Lars algorithm, a slight modification of which gave a fast way to fit the entire regularization path for Lasso. Motivated by this algorithm, we proposed a path-following algorithm for L_1 regularized generalized linear models, which generates piecewise-smooth paths (Park & Hastie 2006*a*). Rosset (2004), and Zhao & Yu (2004) also developed algorithms that serve the same purpose. While these algorithms are limited to selecting a single term at a time, the group-Lars and the group-Lasso methods mentioned earlier select features as a group, among the predefined sets of variables.

To fit gene-interaction models with the data consisting of genotype measurements and a binary response, we first construct sets of indicators representing all the available factors and all possible two-way interactions. We then provide these grouped variables to the path algorithms. Although we expected an improvement in terms of correct feature selection and prediction accuracy over our L_2 penalized stepwise logistic regression approach, which selects variables in a greedy manner, this was not always the case. We showed that these smoother methods perform no better than stepwise logistic regression, mainly because they tend to select large groups of variables too easily.

In the following sections, we illustrate several regularization schemes for grouped variable selection in detail and compare their performance with that of stepwise logistic regression with L_2 penalization. In Section 2, we describe the group-Lars and the group-Lasso methods; in addition, we propose a modified group-Lars algorithm and a path-following procedure for group-Lasso. We present detailed simulation results in Section 3 and a real data example in Section 4. We conclude with a summary and further thoughts in Section 5.

2 Regularization Methods for Grouped Variable Selection

In this section, we review the group-Lars and the group-Lasso methods proposed by Yuan & Lin (2006) and propose another version of group-Lars. We call these two group-Lars methods Type I and Type II, respectively. We describe a path-following algorithm for group-Lasso, which uses the *predictor-corrector* convex optimization scheme as in Park & Hastie (2006*a*).

2.1 Group-Lars: Type I

Efron et al. (2004) proposed least angle regression (LARS) as a procedure that is closely related to Lasso and that provides a fast way to fit the entire regularization path for Lasso. At the beginning of the Lars algorithm, a predictor that is most strongly correlated with the response enters the model. The coefficient of the chosen predictor grows in the direction of the sign of its correlation. The single coefficient grows until another predictor achieves the same absolute correlation with the current residuals. At this point, both coefficients start moving in the least squares direction of the two predictors; this is also the direction that keeps their correlations with the current residuals equal. At each subsequent step, a new variable is added to the model and the path extends in a piecewise-linear fashion. The path is completed either when the size of the active set reaches the sample size, or when all the variables are active and have attained the ordinary least squares fit. As the Lars path proceeds, all the active variables carry the same, and the largest, amount of correlation with the current residuals. Yuan & Lin's group-Lars algorithm operates in a similar way: a group is included in the model if and only if the *average squared correlation* of the variables in the group is the largest, and thus, the same as other active groups.

The group-Lars algorithm proceeds as follows:

1. The algorithm begins by computing the average squared correlation of the elements in each group, with the response.

2. The group of variables with the largest average squared correlation enters the model, and the coefficients of all its components move in the least squares direction.

3. The first segment of the path extends linearly until the average squared correlation of another group meets that of the active group.

4. Once the second group joins, the coefficients of all the variables in the two groups again start moving in their least squares direction.

5. Analogously, a new group enters the model at each step until all the groups are added, and all the individual predictors are orthogonal to the residuals, with zero correlations.

Once a group has entered the active set, its average squared correlation with the residuals stays the same as other active groups because all the individual correlations (their absolute

values) decrease proportionally to their current sizes. We can compute how long a segment of a path extends until the next group joins the active set by solving a set of quadratic equations. If the total number of the predictors would exceed the sample size with the addition of the next group, then the algorithm must stop without adding the new candidate group.

2.2 Group-Lars: Type II

The group-Lars Type I algorithm controls groups of different sizes by tracing their *average squared correlations;* however, squaring the individual correlations sometimes makes a few strongest predictors in the group dominate the average score. We propose another variant of the Lars algorithm for group variable selection that lessens such effect; our algorithm traces the *average absolute correlation* for each group instead of the average squared correlation. The most important difference in effect is that our strategy requires more components of a group to be strongly correlated with the residuals (in a relative sense, when compared to the previous Type I algorithm). Therefore, our algorithm tracking the average absolute correlation is more robust against false positives of selecting large groups when only a few of the elements are strongly correlated with the residuals.

The Type II algorithm is identical to the enumerated description of the Type I algorithm in Section 2.1, except that the *average squared correlation* is replaced by the *average absolute correlation.* Here we illustrate the algorithm using the same mathematical notation as in Efron et al. (2004):

- The given data are $\{(\mathbf{X}, \mathbf{y}) : \mathbf{X} \in \mathcal{R}^{n \times p}, \mathbf{y} \in \mathcal{R}^n\}$. n is the sample size. Denote the columns of $\mathbf{X}$ as $\mathbf{x}_j,\ j = 1, \ldots, p$, and assume that each column has been centered to have mean zero. The p variables have been partitioned into K disjoint groups $G_1, \ldots, G_K$.

- The initial residuals are $\mathbf{r} = \mathbf{y} - \bar{\mathbf{y}}$.

- For each group, compute the average absolute correlation of the variables with the residuals. Select the group with the largest average absolute correlation:

$$M \quad = \quad \underset{k \in \{1, \ldots, K\}}{\operatorname{argmax}} \sum_{j \in G_k} |\mathbf{x}_j' \mathbf{r}| / |G_k|. \tag{1}$$

  The corresponding group forms the active set: $\mathcal{A} = G_M$ and $\mathcal{A}^c = \{1, \ldots, p\} \setminus G_M$.

- Repeat the following while $|\mathcal{A}| \leq n$ and $\sum_{j \in \mathcal{A}} |\mathbf{x}_j' \mathbf{r}| > 0$.

  1. Let $\mathbf{u}_{\mathcal{A}}$ be the unit vector toward the least squares direction of $\{\mathbf{x}_j : j \in \mathcal{A}\}$.
  2. If $\mathcal{A}^c = \emptyset$, then find how much to extend $\mathbf{u}_{\mathcal{A}}$ so that all the average absolute correlations for the active groups decrease to zero. That is, solve the following

equation for $\gamma > 0$ for any $G_k \subset \mathcal{A}$:

$$\sum_{j \in G_k} |\mathbf{x}_j'(\mathbf{r} - \gamma \mathbf{u}_{\mathcal{A}})|/|G_k| \;\; = \;\; 0. \tag{2}$$

3. If $\mathcal{A}^c \neq \emptyset$, then for every group in $\mathcal{A}^c$, find how much to extend $\mathbf{u}_{\mathcal{A}}$ so that the average absolute correlation for the group is the same as those in $\mathcal{A}$. That is, for all $G_l \subset \mathcal{A}^c$, solve the following equation for $\gamma > 0$:

$$\sum_{j \in G_l} |\mathbf{x}_j'(\mathbf{r} - \gamma \mathbf{u}_{\mathcal{A}})|/|G_l| \;\; = \;\; \sum_{j \in G_k} |\mathbf{x}_j'(\mathbf{r} - \gamma \mathbf{u}_{\mathcal{A}})|/|G_k|, \tag{3}$$

where G_k is any group in $\mathcal{A}$. Among the solution $\gamma' s$, choose the smallest positive value and call it $\hat{\gamma}$. Letting G_M be the corresponding group index, enlarge the active set: $\mathcal{A} = \mathcal{A} \cup G_M$, and $\mathcal{A}^c = \mathcal{A}^c \setminus G_M$.

4. Compute the residuals: $\mathbf{r} = \mathbf{r} - \hat{\gamma} \mathbf{u}_{\mathcal{A}}$.

For each run of the above enumerated steps, the unit vector $\mathbf{u}_{\mathcal{A}}$ for a linear segment of the path can be expressed in a simple form. Denote the sub-matrix of $\mathbf{X}$ for the active variables as $\mathbf{X}_{\mathcal{A}}$, and define the following:

$$\mathbf{c}_{\mathcal{A}} = \mathbf{X}_{\mathcal{A}}'\mathbf{r}, \;\; \mathbf{G}_{\mathcal{A}} = \mathbf{X}_{\mathcal{A}}'\mathbf{X}_{\mathcal{A}}, \;\; A_{\mathcal{A}} = (\mathbf{c}_{\mathcal{A}}'\mathbf{G}_{\mathcal{A}}^{-1}\mathbf{c}_{\mathcal{A}})^{-1/2}. \tag{4}$$

Then $\mathbf{u}_{\mathcal{A}} = A_{\mathcal{A}}\mathbf{X}_{\mathcal{A}}\mathbf{G}_{\mathcal{A}}^{-1}\mathbf{c}_{\mathcal{A}}$ is the unit vector in the direction of $\hat{\boldsymbol{\mu}}_{\mathcal{A}} = \mathbf{X}_{\mathcal{A}}\mathbf{G}_{\mathcal{A}}^{-1}\mathbf{X}_{\mathcal{A}}'\mathbf{y}$, the full least squares fit using the current active set. It can be easily shown that in every run, $\hat{\gamma} \in [0, 1/A_{\mathcal{A}}]$, and it equals the upper bound $1/A_{\mathcal{A}}$ when there is no additional variable to enter the model as in Step 2.

The following lemma ensures that the average absolute correlations stay identical across all the groups in the active set as the path proceeds, given that the average absolute correlation of each group had been the same as the rest (the ones who joined $\mathcal{A}$ earlier) at its entry. We omit the proof.

Lemma 2.1. *Let $\hat{\boldsymbol{\mu}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ be the fitted values with some coefficient estimates $\hat{\boldsymbol{\beta}}$. Let $\mathbf{r} = \mathbf{y} - \hat{\boldsymbol{\mu}}$ and $\mathbf{c} = \mathbf{X}'\mathbf{r}$ denote the residuals and their correlations with the predictors, respectively. If $\hat{\boldsymbol{\beta}}$ extends in the least squares direction for $\mathbf{r}$, which is $\tilde{\boldsymbol{\mu}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{r}$, then the entries of $|\mathbf{c}|$ decrease at the rate proportional to their current sizes. That is, for any $\alpha \in [0, 1]$,*

$$|\mathbf{c}(\alpha)| = |\mathbf{X}'(\mathbf{r} - \alpha\tilde{\boldsymbol{\mu}})| = (1 - \alpha)|\mathbf{c}|.$$

If we apply the group-Lars methods to over-represented groups of dummy variables (dummy variables summing up to 1), the Lars paths may not be unique or suffer from a singularity issue. To avoid such problems, we add a slight L_2 norm penalty to the sum of squared error loss and formulate the Lars algorithms the same way. This modification simply amounts to extending the LARS-EN algorithm, a variant of the Lars algorithm proposed by Zou & Hastie (2005), to a version that performs grouped variable selection.

2.3 Group-Lasso

2.3.1 Criterion

Yuan & Lin (2006) introduced the group-Lasso method, which finds the coefficients that minimize the following criterion:

$$L(\boldsymbol{\beta};\lambda) \;=\; \frac{1}{2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2+\lambda\sum_{k=1}^{K}\sqrt{|G_k|}\|\boldsymbol{\beta}_k\|_2, \tag{5}$$

where λ is a positive regularization parameter, and $\boldsymbol{\beta}_k$ denotes the elements of $\boldsymbol{\beta}$ corresponding to the group G_k. We can replace the loss function (sum of squared error loss above) with that of generalized linear models. The criterion (5) is now written in this general form:

$$L(\boldsymbol{\beta};\lambda) \;=\; -l(\mathbf{y};\boldsymbol{\beta})+\lambda\sum_{k=1}^{K}\sqrt{|G_k|}\|\boldsymbol{\beta}_k\|_2, \tag{6}$$

where $\mathbf{y}$ is the response vector that follows a distribution in exponential family, and l is the corresponding log-likelihood. Meier, van de Geer & Buhlmann (2006) studied the properties of this criterion for the binomial case and proposed an algorithm.

When the response $\mathbf{y}$ is Gaussian, the criterion of minimizing (5) may be written in this equivalent form:

$$\begin{aligned}
&\text{Minimize} && t && (7)\\
&\text{subject to} && \|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2\le t, && (8)\\
& && \|\boldsymbol{\beta}_k\|_2\le a_k \text{ for } k=1,\ldots,K, && (9)\\
& && \sum_{k=1}^{K}\sqrt{|G_k|}\;a_k\le s, && (10)
\end{aligned}$$

where $\boldsymbol{\beta}$, $\mathbf{a}$, and t are the variables, and s is a fixed value that replaces λ in (5). This formulation suggests that the minimization can be solved as a *second-order cone programming* (SOCP) problem. For all the other distributions in exponential family, the problem cannot be treated as a standard SOCP, but as a convex optimization problem with second-order cone (SOC) constraints. In our algorithm, we choose to use the form of the criterion as in (6) for any distribution and solve a convex optimization problem with SOC constraints as follows:

$$\begin{aligned}
&\text{Minimize} && -l(\mathbf{y};\boldsymbol{\beta})+\lambda\sum_{k=1}^{K}\sqrt{|G_k|}\;a_k && (11)\\
&\text{subject to} && \|\boldsymbol{\beta}_k\|_2\le a_k \text{ for } k=1,\ldots,K. && (12)
\end{aligned}$$

According to the Karush-Kuhn-Tucker conditions (also shown in Proposition 1 of Yuan & Lin (2006)), the group G_k is in the active set, and thus, all the elements of $\boldsymbol{\beta}_k$ are nonzero

at a given λ if and only if the following holds:

$$\sum_{j \in G_k} |\mathbf{x}_j' \mathbf{W} \mathbf{r}|^2 / |G_k| \quad = \quad \lambda^2, \tag{13}$$

where $\mathbf{W}$ is a diagonal matrix with n diagonal elements V_i, the variance estimate for the *i-th* observation, and $\mathbf{r}$ denotes the current residuals. The residual for the *i-th* observation is $(y_i - \mu_i)(\frac{\delta \eta}{\delta \mu})_i$, where μ and η denote the mean of the response and the linear predictor $\mathbf{x}'\boldsymbol{\beta}$, respectively.

2.3.2 Path-following algorithm

We introduce an algorithm for finding the entire regularization path for criterion (6). That is, we trace the coefficient paths as the regularization parameter λ ranges from zero to a value large enough to force all the coefficients to be zero. Analogous to the algorithm presented in Park & Hastie (2006*a*), we propose another version of the *predictor-corrector* scheme. We repeat the following steps, for each iteration decreasing λ.

- Predictor step: (1) Estimate the direction of $\boldsymbol{\beta}$ for the next segment of the path. (2) Assuming the coefficients will move in that direction, compute the (smallest) decrement in λ that would change the active set. (3) Estimate the solution for the new λ, by extending the previous solution in the estimated direction.

- Corrector step: Using the estimate from the predictor step as the initial value, compute the exact solution corresponding to the decreased λ.

- Active set: Check if the active set has been changed with the new λ, and if that is true, repeat the corrector step with the updated active set.

We now describe the steps in the *m-th* iteration in detail, for the Gaussian case.

1. Predictor step

   In the *m-th* predictor step, we estimate the solution $\hat{\boldsymbol{\beta}}^{m+}$ for a decreased regularization parameter $\lambda = \lambda_m$. To determine λ_m, we first estimate the direction in which $\boldsymbol{\beta}$ extends from the previous solution $\hat{\boldsymbol{\beta}}^{m-1}$; denote the vector in this direction as $\mathbf{b}_m$, scaled such that $\mathbf{X}_{\mathcal{A}} \mathbf{b}_m$ is a unit vector. Then we compute the smallest, positive constant γ_m such that

   $$\hat{\boldsymbol{\beta}}^{m+} \quad = \quad \hat{\boldsymbol{\beta}}^{m-1} + \gamma_m \mathbf{b}_m \tag{14}$$

   would change the active set.

   As used in Park & Hastie (2006*a*), the natural and most accurate choice for $\mathbf{b}_m$ would be $\delta\boldsymbol{\beta}/\delta\lambda$, the tangent slope of the curved path along with the change in λ. However, for simplicity of the algorithm, we approximate the direction as follows:

   $$\mathbf{b}_m \quad = \quad A_{\mathcal{A}} \mathbf{G}_{\mathcal{A}}^{-1} \mathbf{c}_{\mathcal{A}}, \tag{15}$$

using the notation (4). Using this choice of $\mathbf{b}_m$, the fitted responses move in the direction of $\mathbf{u}_{\mathcal{A}} = \mathbf{X}_{\mathcal{A}}\mathbf{b}_m$, which is exactly the same as a step in the group-Lars algorithms in Section 2.1 and Section 2.2. The approximation (15) is desirable also because it makes the correlations of the active variables with the current residuals change proportionally to the current sizes as in Lemma 2.1. This fact makes it possible to estimate the decrement in λ that would change the active set. From our experience, (15) is a reasonable approximation of $\delta\boldsymbol{\beta}/\delta\lambda$. An example in Section 3 demonstrates the fact.

As γ_m increases from zero to $\gamma > 0$, the correlations of the variables with the current residuals change as follows:

$$\mathbf{c}(\gamma) = \hat{\mathbf{c}} - \gamma \mathbf{X}'\mathbf{u}_{\mathcal{A}}, \tag{16}$$

where $\hat{\mathbf{c}}$ is the vector of correlations for the previous coefficient estimates $\hat{\boldsymbol{\beta}}^{m-1}$. Note that $\mathbf{c}(\gamma) = (1 - \gamma A_{\mathcal{A}})\hat{\mathbf{c}}$ for the variables in $\mathcal{A}$, and thus, their group correlation measure (the average squared correlation as in (13)) decreases from λ_{m-1}^2 by a factor of $(1 - \gamma A_{\mathcal{A}})^2$. By solving a quadratic set of equations, we can compute the smallest $\gamma \in (0, A_{\mathcal{A}}^{-1}]$ with which the average squared correlation of a group that is currently not active will be the same as that of currently active groups, satisfying

$$\sum_{j \in G_l} c_j(\gamma)^2/|G_l| = (1 - \gamma A_{\mathcal{A}})^2 \lambda_{m-1}^2 \tag{17}$$

for some $G_l \subset \mathcal{A}^c$. We also compute the smallest $\gamma > 0$ for which $\hat{\boldsymbol{\beta}}^{m-1} + \gamma \mathbf{b}_m$ will be zero, in which case we suppose the corresponding variable will drop out of the active set. We then let the smaller one of these two values of γ be γ_m, and using this constant, $\lambda_m = (1 - \gamma_m A_{\mathcal{A}})\lambda_{m-1}$. $\hat{\boldsymbol{\beta}}^{m+}$ computed as in (14) is our estimate of the coefficients at λ_m.

If we let $\lambda_m = \lambda_{m-1} - h$ for a small constant $h > 0$, then we can generate the exact path, by computing the solutions at many values of λ. However, selecting the step sizes in λ adaptively adds efficiency and accuracy to the algorithm as demonstrated in Park & Hastie (2006*a*).

2. Corrector step
   Having estimated the *m-th* set of the coefficients $\hat{\boldsymbol{\beta}}^{m+}$ and the corresponding value for the regularization parameter λ, we can now solve the optimization problem of minimizing (6) with $\lambda = \lambda_m$. As in (11) - (12), it is formulated as a convex optimization problem with SOC constraints. Using $\hat{\boldsymbol{\beta}}^{m+}$ as a warm starting value, we expect that the cost of solving for the exact solution $\hat{\boldsymbol{\beta}}^{m}$ is low.

3. Active set
   We first complete the *m-th* predictor and corrector steps using the active set from the

previous iteration. After the corrector step, we check if the active set must have been modified. As was done in Park & Hastie (2006*a*), we augment the active set $\mathcal{A}$ with G_l if

$$\sum_{j \in G_l} |\mathbf{x}_j' \mathbf{r}|^2 / |G_l| \;\; \geq \;\; \max_{G_k \in \mathcal{A}} \sum_{j \in G_k} |\mathbf{x}_j' \mathbf{r}|^2 / |G_k| = \lambda_m^2 \tag{18}$$

for any $G_l \subset \mathcal{A}^c$. We repeat this check, followed by another corrector step with the updated active set, until no more group needs to be added.

We then check whether the active set must be reduced. If $\|\hat{\boldsymbol{\beta}}_k^m\|_2 = 0$ for any $G_k \subset \mathcal{A}$, we eliminate the group G_k from the active set.

We iterate this set of steps until $\lambda = 0$, at which point all the correlations $\hat{\mathbf{c}}$ are zero.

When $\mathbf{y}$ follows a distribution other than Gaussian, this algorithm still applies the same way. $\mathbf{G}_\mathcal{A}$ in (15) is replaced by $\mathbf{X}_\mathcal{A}' \mathbf{W} \mathbf{X}_\mathcal{A}$, and thus, the predictor step amounts to taking a step in the weighted group-Lars direction. In other words, for a predictor step, we approximate the log-likelihood as a quadratic function of $\boldsymbol{\beta}$ and compute the group-Lars direction as in the case of Gaussian distribution. When checking to see if the active set is augmented, the correlation $\mathbf{x}_j' \mathbf{r}$ in (18) should be replaced by the weighted correlation $\mathbf{x}_j' \mathbf{W} \mathbf{r}$.

3 Simulations

In this section, we compare different regression/classification methods for group variable selection through three sets of simulations. To imitate data with genotype measurements at multiple loci, we generate six categorical variables, each with three levels, and a binary response variable. For every factor and every two-way interaction, we define a set of indicators, assigning a dummy variable for each level. These sets form the groups that are selected simultaneously. Among the six factors, only the first two affect the response. As in Park & Hastie (2006*b*), we assign balanced class labels with the following conditional probabilities of belonging to class 1. (AA,Aa,aa) and (BB,Bb,bb) are the level sets for the first two factors.

Additive Model

$P(A) = P(B) = 0.5$

	BB	Bb	bb
AA	0.845	0.206	0.206
Aa	0.206	0.012	0.012
aa	0.206	0.012	0.012

Interaction Model I

$P(A) = P(B) = 0.5$

	BB	Bb	bb
AA	0.145	0.206	0.206
Aa	0.206	0.012	0.012
aa	0.206	0.012	0.012

Interaction Model II

$P(A) = P(B) = 0.5$

	BB	Bb	bb
AA	0.045	0.206	0.206
Aa	0.206	0.012	0.012
aa	0.206	0.012	0.012

As can be seen in Figure 1, in the first scenario, the log-odds for class 1 is additive in the effects from the first two factors. For the next two, weak and strong interaction effects are present between the two factors. Therefore, $\mathbf{A} + \mathbf{B}$ is the true model for the first, while $\mathbf{A} * \mathbf{B}$ is appropriate for the other two settings.

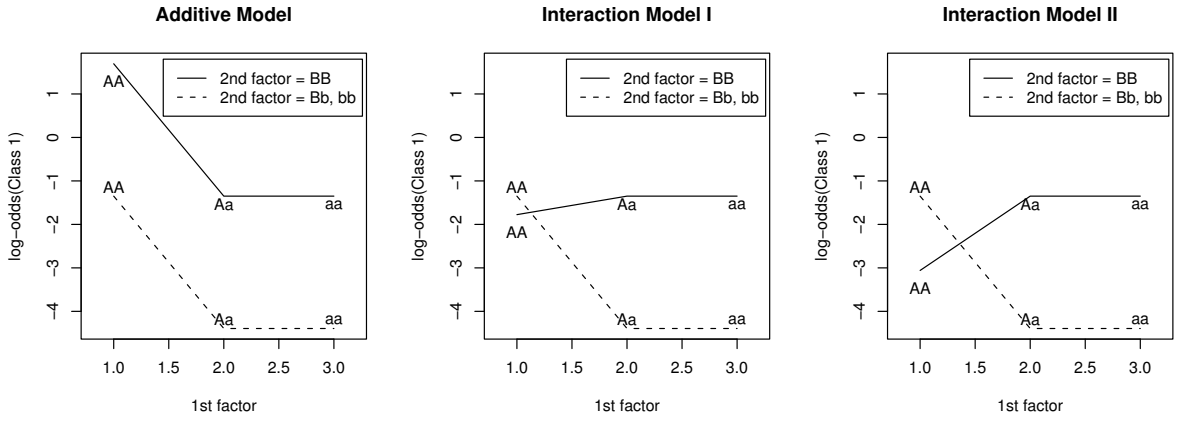

Figure 1: *The patterns of log-odds for class 1, for different levels of the first two factors*

Under these settings, we generated 50 independent datasets consisting of six factors and 200 training and another 200 test observations. Each training and test dataset consisted of 100 cases and 100 controls. For all 50 repetitions, we fit group Lars Type I and II, group-Lasso assuming Gaussian and binomial distributions for the response, and stepwise penalized logistic regression with L_2 penalization illustrated in Park & Hastie (2006*b*). We estimated the prediction error for each method using the test data (by averaging the 50); the results are summarized in Table 1. The standard errors for the estimates are parenthesized. Although the error estimates were similar across all the methods we presented, stepwise logistic regression was significantly more accurate than other methods for the additive model.

In Table 2, we present a further comparison by counting the number of runs (out of 50) for which the correct model was identified. For the additive model, group-Lars Type II selected $\mathbf{A} + \mathbf{B}$ (the true model) more often than Type I; the Type II method too easily let the interaction terms of size 9 enter the model. Stepwise logistic regression with L_2 penalization scored the highest for the additive and interaction model II. Forward stepwise selection used in penalized logistic regression is a greedy approach; however, it found the true model more frequently than the path-following procedures, which more aggressively allowed terms to join the active set. In general, group-Lasso with binomial log-likelihood selected noisy terms more frequently than the Gaussian case.

Methods	Additive Model	Interaction Model I	Interaction Model II
Group-Lars I	0.2311(0.005)	0.2451(0.006)	0.2228(0.005)
Group-Lars II	0.2306(0.005)	0.2389(0.004)	0.2203(0.005)
Group-Lasso (Gaussian)	0.2355(0.005)	0.2456(0.006)	0.2229(0.005)
Group-Lasso (Binomial)	0.2237(0.005)	0.2453(0.005)	0.2249(0.005)
step PLR	0.2180(0.004)	0.2369(0.004)	0.2244(0.005)

Table 1: *Comparison of prediction performances*

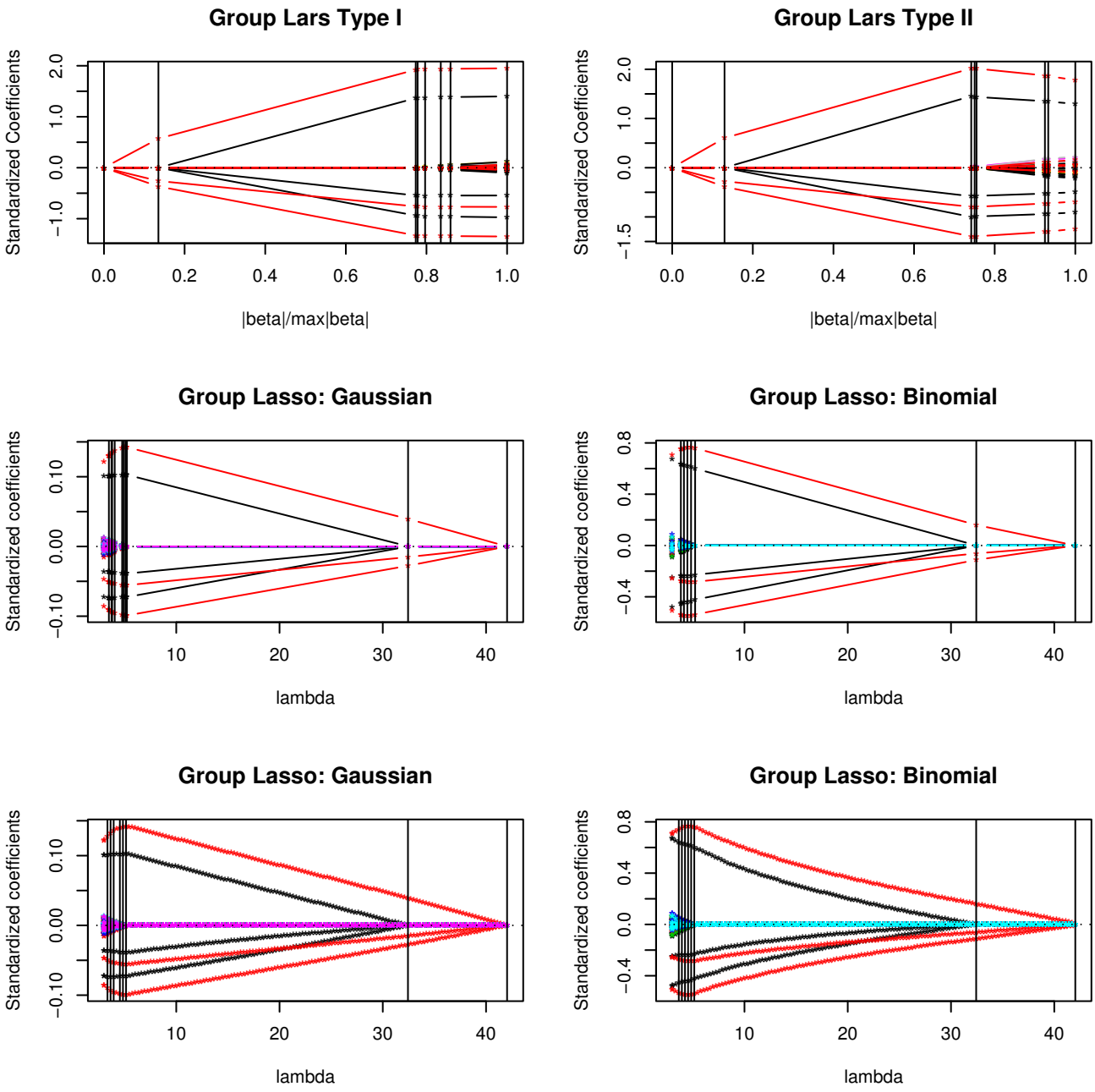

Figure 2: *Comparison of the coefficient paths for the group-Lars (the first row) and the group-Lasso (the rest) methods. The step sizes in λ are adaptively selected for the plots in the second row, while they were fixed at 0.3 for the last two plots.*

Methods	Additive Model	Interaction Model I	Interaction Model II
Group-Lars I	46/50	33/50	38/50
Group-Lars II	34/50	42/50	35/50
Group-Lasso (Gaussian)	46/50	36/50	37/50
Group-Lasso (Binomial)	17/50	20/50	27/50
step PLR	49/50	31/50	39/50

Table 2: *Counts for correct term selection*

In Figure 2, we compared the coefficient paths for the group-Lars and the group-Lasso methods for one of the datasets in which the first two factors were additive. The first two factors are marked black and red in the figure. The first two plots show the paths for group-Lars Type I and group-Lars Type II; both are piecewise-linear. The next two plots are from the group-Lasso methods, using negative log-likelihoods for the Gaussian and binomial distributions as loss functions, respectively. The step sizes in λ were determined adaptively in these two runs of group-Lasso. For the last two plots, we computed the solutions decreasing λ by a small constant (0.3) in every iteration. Nonlinearity of the paths is visible in the last plot, which used the negative binomial log-likelihood. For both binomial and Gaussian cases, we approximated the exact paths with a reasonable accuracy by adjusting the step lengths as illustrated in Section 2.3.2, thereby significantly reducing the total number of iterations (132 to 16 for the Gaussian, and 132 to 9 for the binomial case).

4 Real Data Example

We applied the path-following procedures and stepwise logistic regression with L_2 penalization to a real dataset with genotype measurements on 14 loci and a binary response indicating the presence of bladder cancer (201 cases and 214 controls). The dataset was first introduced in Hung et al. (2004).

Table 3 summarizes the cross-validated prediction error, sensitivity, and specificity from a five-fold cross-validation. For each fold, we ran an internal cross-validation to choose the level of regularization. The negative log-likelihood was used as the criterion in the internal cross-validations. Overall (classification) error rate was the lowest for stepwise logistic regression, the specificity being especially high compared to other methods.

Methods	Prediction error	Sensitivity	Specificity
Group-Lars I	156/415=0.376	128/201	131/214
Group-Lars II	155/415=0.373	127/201	133/214
Group-Lasso (Gaussian)	154/415=0.371	126/201	135/214
Group-Lasso (Binomial)	157/415=0.378	128/201	130/214
step PLR	147/415=0.354	122/201	146/214

Table 3: *Comparison of prediction performances*

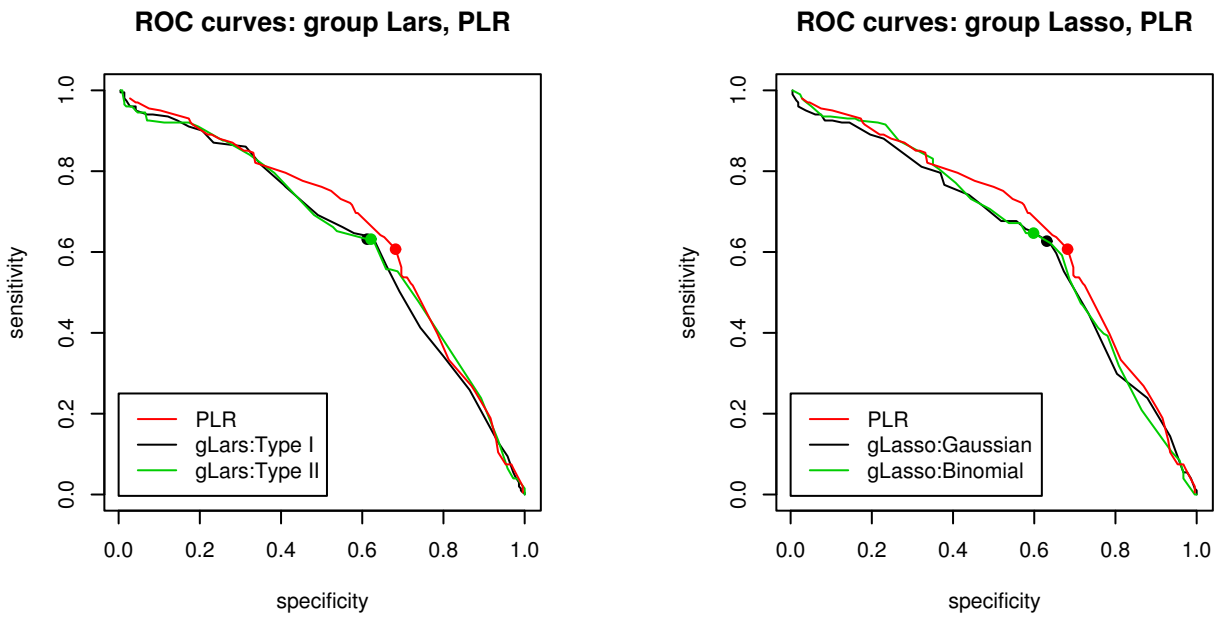

Figure 3: *Comparison of ROC curves*

One would expect an improvement by applying a smooth variable selection mechanism as in group-Lars or group-Lasso. However, such smoothness may turn out to be a disadvantage. As in Section 3, the group-Lars and the group-Lasso methods tend to select irrelevant terms more easily than stepwise logistic regression. Even when these path-following methods identify a correct subset of features, the nonzero coefficients are shrunken fits. On the other hand, the L_2 regularization in our stepwise logistic regression is meaningful as a technical device (Park & Hastie 2006*b*) rather than as a smoothing tool, and thus, we often apply only a slight penalization to the size of the coefficients.

We further extended the prediction error analysis by plotting the receiver operating characteristic (ROC) curves for all the methods compared in Table 3. For each method, we generated multiple sets of classification results by applying different cut-off values to the cross-validated responses from the previous analysis. The left and right panels of Figure 3 compare the ROC curves of the group-Lars methods and the group-Lasso methods to stepwise logistic regression, respectively. The ROC curve for stepwise logistic regression lies slightly more toward the upper right-hand corner than all the other curves, although the difference is not statistically significant.

5 Discussion

In this paper, we studied the use of various regularization path algorithms for grouped variable selection to fit gene-interaction models. We first considered two types of group-Lars algorithms. Group-Lars Type I, proposed by Yuan & Lin (2006), kept the groups with the

largest average squared correlation with the current residuals in the active set. In group-Lars Type II, the active groups were the ones with the largest average absolute correlation. We showed some simulation results in which the Type II algorithm was preferred because the Type I algorithm selected large groups too easily. We then studied the group-Lasso method and suggested a general path-following algorithm that can be implemented with the log-likelihood of any distribution in the exponential family. Although the path-algorithm for group-Lasso is more complex than that of group-Lars, group-Lasso is more informative in that we have the explicit criterion as in (6).

Group-Lasso yields a stable fit even when highly correlated variables are input simultaneously. When some variables are perfectly correlated, as in the case of over-represented groups of indicators for categorical factors, the solution for group-Lasso is still uniquely determined. This can be seen from the Karush-Kuhn-Tucker conditions (13), and we omit the details. This property of the group-Lasso method makes it more attractive as a way to fit with categorical factors coded in dummy variables. On the other hand, the group-Lars paths are not unique in this situation, and as a remedy, we used the LARS-EN algorithm with a slight L_2 penalization instead of Lars. This modification adds a quadratic feature to the group-Lars method, as in the group-Lasso criterion.

We compared the performances of the group-Lars and the group-Lasso methods to that of the forward stepwise approach, a more conventional variable selection strategy, implemented with L_2 penalized logistic regression. The group-Lars and the group-Lasso methods can be preferred for being smooth in selecting terms and being faster. However, based on our experiments, we learned that L_2 penalized logistic regression with the forward stepwise variable selection scheme is still comparable to those alternatives.

References

Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. (2004), 'Least angle regression', *Annals of Statistics* **32**, 407–499.

Genkin, A., Lewis, D. & Madigan, D. (2004), Large-scale bayesian logistic regression for text categorization, Technical report, Rutgers University, New Jersey.

Hung, R., Brennan, P., Malaveille, C., Porru, S., Donato, F., Boffetta, P. & Witte, J. (2004), 'Using hierarchical modeling in genetic association studies with multiple markers: Application to a case-control study of bladder cancer', *Cancer Epidemiology, Biomarkers & Prevention* **13**, 1013–1021.

Meier, L., van de Geer, S. & Buhlmann, P. (2006), The group lasso for logistic regression, Technical report, Eidgenossische Technische Hochschule Zurich, Zurich.

Park, M. & Hastie, T. (2006*a*), l_1 regularization path algorithm for generalized linear models, Technical report, Stanford University, Stanford.

Park, M. & Hastie, T. (2006*b*), Penalized logistic regression for detecting gene interactions, Technical report, Stanford University, Stanford.

Rosset, S. (2004), Tracking curved regularized optimization solution paths, *in* 'Neural Information Processing Systems'.

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society. Series B* **58**, 267–288.

Tibshirani, R. (1997), 'The lasso method for variable selection in the cox model', *Statistics in Medicine* **16**, 385–395.

Yuan, M. & Lin, Y. (2006), 'Model selection and estimation in regression with grouped variables', *Journal of the Royal Statistical Society. Series B* **68**, 49–67.

Zhao, P. & Yu, B. (2004), Boosted lasso, Technical report, University of California, Berkeley, USA.

Zhu, J., Rosset, S., Hastie, T. & Tibshirani, R. (2003), 1-norm support vector machines, *in* 'Neural Information Processing Systems'.

Zou, H. & Hastie, T. (2005), 'Regularization and variable selection via the elastic net', *Journal of the Royal Statistical Society. Series B* **67**, 301–320.