Bi-level feature selection with applications to genetic association studies

Patrick Breheny & Jian Huang

October 15, 2008
Motivation

- In many applications, biological features possess a grouping structure
  - Categorical variables may be represented by a group of indicator functions
  - Continuous variables may be represented by a group of basis functions
  - Genes may be grouped into pathways
  - Genetic markers may be grouped by the gene or haplotype that they belong to

- In high-dimensional regression problems, incorporating grouping into the analysis helps reduce the dimensionality and produce more interpretable results
Grouped penalization

- Penalized regression methods such as the lasso, bridge, and SCAD are an attractive framework for variable selection.
- Penalized methods have also been developed that incorporate grouping structure:
  - Yuan and Lin (2006): group lasso
  - Huang et al. (2007): group bridge
- For genetic association studies, we would like to be able to select groups (genes) as well as individual variables (SNPs) within a group, which we call *bi-level selection*. 
Overview

In this talk, I will:

- Introduce a new method for bi-level selection, group SCAD
- Develop a new algorithm for fitting models with grouped penalties that is very fast even when the number of variables is much larger than the sample size
- Apply the idea of group penalization to genetic association studies
Suppose we have data \( \{(x_i, y_i)_{i=1}^{n}\} \), where \( y_i \) is the response variable and \( x_i \) is a \( p \)-dimensional grouped predictor.

We denote \( x_i \) as being composed of \( J \) groups \( x_{ij} \), with \( K_j \) denoting the size of factor \( j \).

The problem of interest is to estimate a sparse vector of coefficients \( \beta \) using a loss function \( L \) which quantifies the discrepancy between \( y_i \) and the linear predictor \( \eta_i = x_i' \beta = \beta_0 + \sum_{j=1}^{J} x_{ij}' \beta_j \).

To ensure that the penalty is applied equally, covariates are standardized prior to fitting such that \( \sum_{i=1}^{n} x_{jk} = 0 \) and \( \frac{1}{n} \sum_{i=1}^{n} x_{jk}^2 = 1 \) \( \forall j, k \).
Fan & Li (2001) propose a nonconvex penalty called SCAD defined on $[0, \infty)$ by

$$f_{\lambda,a}(\theta) = \begin{cases} 
\lambda \theta & \text{if } \theta \leq \lambda \\
a\lambda \theta - \frac{1}{2}(\theta^2 + \lambda^2) & \text{if } \lambda < \theta < a\lambda \\
\frac{\lambda^2(a^2-1)}{2(a-1)} & \text{if } \theta \geq a\lambda 
\end{cases}$$

for $\lambda \geq 0$, $a > 2$
The group SCAD estimate minimizes

\[ Q(\beta) = \frac{1}{2n} \| y - X\beta \|^2 + \sum_{j=1}^{J} f_{\lambda,a} \left( \phi \sum_{k=1}^{K_j} f_{\lambda,a}(|\beta_{jk}|) \right) \]

where

\[ \phi = \frac{2a(a - 1)}{K_j \lambda(a^2 - 1)} \]
The group SCAD estimate minimizes

\[ Q(\beta) = \frac{1}{2n} \|y - X\beta\|^2 + \sum_{j=1}^{J} f_{\lambda,a} \left( \phi \sum_{k=1}^{K_j} f_{\lambda,a}(|\beta_{jk}|) \right) \]

where \[ \phi = \frac{2a(a - 1)}{K_j \lambda (a^2 - 1)} \]

For comparison,

**gLasso:** \[ Q(\beta) = \frac{1}{2n} \|y - X\beta\|^2 + \lambda \sum_{j=1}^{J} \sqrt{K_j} \|\beta_j\| \]

**gBridge:** \[ Q(\beta) = \frac{1}{2n} \|y - X\beta\|^2 + \lambda \sum_{j=1}^{J} K_j^{\gamma} \|\beta_j\|^{\gamma} \]
Illustration of the three group penalties

Group Lasso

Group Bridge

Group SCAD
In generalized linear models, the negative log-likelihood is used as the loss function.

The usual approach to model fitting is to make a quadratic approximation to the loss function using current estimates:

$$L(\beta) \approx \frac{1}{2} (z - X\beta)'W(z - X\beta)$$

For the sake of clarity, we will present the algorithms from the perspective of squared error loss; the modifications are straightforward for logistic regression.

However, the tuning parameter of the SCAD penalty is affected.
Overview

- Coordinate descent algorithms optimize a target function with respect to a single parameter at a time, iteratively cycling through all parameters until convergence is reached.
- Coordinate descent algorithms are ideal for problems like the lasso where deriving the solution is complicated in high dimensions but simple in one dimension.
- Group penalties usually do not have a simple solution even with respect to a single parameter; however, we may approximate these penalties to obtain a locally accurate representation that does.
Letting $\tilde{\beta}$ represent the current estimate of $\beta$, the overall structure of the local coordinate descent (LCD) algorithm is as follows:

1. Choose an initial estimate $\tilde{\beta} = \beta^{(0)}$
2. Approximate loss function, if necessary
3. Update covariates:
   a. Update $\tilde{\beta}_0$
   b. For $j \in \{1, \ldots, J\}$, update $\tilde{\beta}_j$
4. Repeat steps 2 and 3 until convergence
The intercept

- The partial residual for updating $\tilde{\beta}_0$ is $\tilde{y} = y - X_{-0}\tilde{\beta}_{-0}$, where the $-0$ subscript refers to what remains of $X$ or $\beta$ after the intercept’s column or element has been removed, respectively.

- The updated value of $\tilde{\beta}_0$ is therefore the simple linear regression solution:

$$
\tilde{\beta}_0 \leftarrow \frac{x'_0\tilde{y}}{x'_0x_0} = \frac{1}{n}x'_0\tilde{y}
$$
Solution to the univariate lasso

When the penalty being applied to a single parameter is $\lambda |\beta|$, the solution is

$$
\hat{\beta} = \frac{S\left(\frac{1}{n}x'y, \lambda\right)}{\frac{1}{n}x'x} = S\left(\frac{1}{n}x'y, \lambda\right)
$$

where $S(z, c)$ is the soft-thresholding operator defined for positive $c$ by

$$
S(z, c) = \begin{cases} 
  z - c & \text{if } z > c \\
  0 & \text{if } |z| \leq c \\
  z + c & \text{if } z < -c 
\end{cases}
$$
Taking the first order Taylor series approximation about $\tilde{\beta}_j$, the penalty applied to $|\beta_{jk}|$ is approximately proportional to $\tilde{\lambda}_{jk} |\beta_{jk}|$, where

$$\tilde{\lambda}_{jk} = \phi f'_{\lambda,a} \left( \phi \sum_{m=1}^{K_j} f_{\lambda,a}(|\tilde{\beta}_{jm}|) \right) f'_{\lambda,a}(|\tilde{\beta}_{jk}|)$$

leading to the simple updating step

$$\tilde{\beta}_{jk} \leftarrow S \left( \frac{1}{n} x'_{jk} \tilde{y}, \tilde{\lambda}_{jk} \right)$$
The LCD algorithm for group bridge is rather similar to that for group SCAD, only with

\[
\tilde{\lambda}_{jk} = \lambda \gamma K_j \| \tilde{\beta}_j \|_1^{\gamma-1}
\]

The difficulty posed by group bridge is that for \( \gamma < 1 \), the derivative of the penalty function is undefined at \( \tilde{\beta}_j = 0 \).
Group lasso sparsity condition

- One-at-a-time updating is more complicated in the group lasso because of its sparsity properties: group members go to 0 all at once or not at all.
- Thus, we must update $\tilde{\beta}_j$ in two steps: first, check whether $\tilde{\beta}_j = 0$ and second, if $\tilde{\beta}_j \neq 0$, update $\tilde{\beta}_{jk}$ for $k \in \{1, \ldots, K_j\}$.
- The first step is performed by noting that $\tilde{\beta}_j \neq 0$ if and only if
  \[
  \frac{1}{n} \|X_j'\tilde{y}\| > \sqrt{K_j} \lambda
  \]
- If this condition does not hold, then we can set $\tilde{\beta}_j = 0$ and move on...
...if not, we once again make a local approximation to the penalty and update the members of group $j$

However, instead of approximating the penalty as a function of $|\beta_{jk}|$, for group lasso we can obtain a more accurate approximation by considering the penalty as a function of $\beta_{jk}^2$

Now, the penalty applied to $\beta_{jk}$ may be approximated by

$$\frac{1}{2} \tilde{\lambda}_{jk}/\beta_{jk}^2$$

This approach yields a shrinkage updating step instead of a soft-thresholding step:

$$\tilde{\beta}_{jk} \leftarrow \frac{1}{n} x_j^T \tilde{y} \frac{1 + \tilde{\lambda}_{jk}}{1 + \tilde{\lambda}_{jk}}$$
Usually, we are interested in obtaining $\hat{\beta}$ not just for a single value of $\lambda$, but for a range of values and then applying some criterion to choose an optimal $\lambda$.
The LCD algorithm requires an initial value $\beta^{(0)}$.

Because the paths are continuous, a reasonable approach to choosing initial values is to start at one extreme of the path and use the estimate $\hat{\beta}$ from the previous value of $\lambda$ as the initial value for the next value of $\lambda$. 
We consider the average time to fit the entire path of solutions for group lasso, group bridge, and group SCAD, as well as the lasso as a benchmark.

Besides LCD, we consider:

- lars, the most widely used algorithm for fitting lasso paths
- glmnet, a very efficient coordinate descent algorithm for computing lasso paths
- glmpath, predecessor of glmnet, an approach to fitting lasso paths for GLMs not based on coordinate descent
- LQA, algorithm proposed by Fan & Li in 2001
- LLA, algorithm proposed by Zou & Li in 2008
Overview (cont’d)

- We look at three situations:
  - $n = 500$, $p = 200$ for linear regression
  - $n = 1000$, $p = 200$ for logistic regression loss
  - $n = 500$, $p = 2000$ for linear regression

- For the $n > p$ regressions, paths were computed down to $\lambda = 0$

- For the $p > n$ regressions, paths were computed down to 5% of the maximum lambda

- All times averaged over $N = 100$ simulated data sets; all groups of size 10
Linear regression with $n = 500$, $p = 200$

<table>
<thead>
<tr>
<th>Penalty</th>
<th>Algorithm</th>
<th>Average Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lasso</td>
<td>glmnet</td>
<td>.05</td>
</tr>
<tr>
<td>Lasso</td>
<td>lars</td>
<td>.56</td>
</tr>
<tr>
<td>Group lasso</td>
<td>LCD</td>
<td>.37</td>
</tr>
<tr>
<td>Group bridge</td>
<td>LCD</td>
<td>.62</td>
</tr>
<tr>
<td>Group SCAD</td>
<td>LCD</td>
<td>.23</td>
</tr>
<tr>
<td>Group lasso</td>
<td>LQA</td>
<td>6.93</td>
</tr>
<tr>
<td>Group bridge</td>
<td>LLA</td>
<td>6.37</td>
</tr>
<tr>
<td>Group SCAD</td>
<td>LLA</td>
<td>5.02</td>
</tr>
</tbody>
</table>
Logistic regression with $n = 1000, p = 200$

<table>
<thead>
<tr>
<th>Penalty</th>
<th>Algorithm</th>
<th>Average Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lasso</td>
<td>glmnet</td>
<td>.27</td>
</tr>
<tr>
<td>Lasso</td>
<td>glmpath</td>
<td>16.77</td>
</tr>
<tr>
<td>Group lasso</td>
<td>LCD</td>
<td>1.15</td>
</tr>
<tr>
<td>Group bridge</td>
<td>LCD</td>
<td>2.06</td>
</tr>
<tr>
<td>Group SCAD</td>
<td>LCD</td>
<td>.55</td>
</tr>
<tr>
<td>Group lasso</td>
<td>LQA</td>
<td>31.22</td>
</tr>
<tr>
<td>Group bridge</td>
<td>LLA</td>
<td>43.77</td>
</tr>
<tr>
<td>Group SCAD</td>
<td>LLA</td>
<td>18.00</td>
</tr>
</tbody>
</table>
### Linear regression with $n = 500, p = 2000$

<table>
<thead>
<tr>
<th>Penalty</th>
<th>Algorithm</th>
<th>Average Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lasso</td>
<td>glmnet</td>
<td>1.26</td>
</tr>
<tr>
<td>Lasso</td>
<td>lars</td>
<td>16.54</td>
</tr>
<tr>
<td>Group lasso</td>
<td>LCD</td>
<td>14.01</td>
</tr>
<tr>
<td>Group bridge</td>
<td>LCD</td>
<td>9.46</td>
</tr>
<tr>
<td>Group SCAD</td>
<td>LCD</td>
<td>2.87</td>
</tr>
<tr>
<td>Group lasso</td>
<td>LQA</td>
<td>–</td>
</tr>
<tr>
<td>Group bridge</td>
<td>LLA</td>
<td>–</td>
</tr>
<tr>
<td>Group SCAD</td>
<td>LLA</td>
<td>–</td>
</tr>
</tbody>
</table>
Data generated according to

\[ y_i = \mathbf{x}_i' \beta_1 + \ldots + \mathbf{x}_i' \beta_{10} + \epsilon_i \quad \epsilon_i \sim \text{iid} \mathcal{N}(0, 1) \]

with

- 10 groups, 10 variables per group \((p = 100)\)
- Generating models present a variety of group and variable/group sparsity
- 500 replications per model
- \(n = 100\) per replication
- \(\beta\) chosen such that the signal to noise ratio (SNR) is approximately 1:

\[
\frac{\beta_0'E[\mathbf{x}_i\mathbf{x}_i']\beta_0}{\sigma^2} \approx 1
\]
To study the selection properties of the methods, we introduce two measures of selection accuracy:

- **Group misclassification loss (GML):**

\[
\text{GML} = \sum_{j=1}^{J} S_j \% I \left\{ \hat{\beta}_j = 0 \right\} + \sum_{j=1}^{J} .05I \left\{ \hat{\beta}_j \neq 0 \right\} I \left\{ \beta_j^0 = 0 \right\}
\]

where \( S_j \% = \frac{\beta_j^0' \mathbb{E}[x_{ij}x_{ij}'] \beta_j^0}{\sum_{l} \beta_l^0' \mathbb{E}[x_{il}x_{il}'] \beta_l^0} \)

- **Individual misclassification loss (IML):**

\[
\text{IML} = \sum_{j=1}^{J} \sum_{k=1}^{K_j} S_{jk} \% I \left\{ \hat{\beta}_{jk} = 0 \right\} + \sum_{j=1}^{J} \sum_{k=1}^{K_j} .01I \left\{ \hat{\beta}_{jk} \neq 0 \right\} I \left\{ \beta_{jk}^0 = 0 \right\}
\]

where \( S_{jk} \% = \frac{\beta_{jk}^0' \mathbb{E}[x_{ijk}x_{ijk}'] \beta_{jk}^0}{\sum_{l} \sum_{m} \beta_{lm}^0' \mathbb{E}[x_{ilm}x_{ilm}'] \beta_{lm}^0} \)
Model error

![Graph showing model error comparisons for gBridge, gLasso, and gSCAD. The x-axis represents the number of nonzero members per group, ranging from 2 to 10. The y-axis represents ME (model error), ranging from 0.1 to 0.3. The graph is divided into three sections, each showing trends for different numbers of members per group.]
# of variables selected

<table>
<thead>
<tr>
<th>n.var</th>
<th>gBridge</th>
<th>gLasso</th>
<th>gSCAD</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>2</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>20</td>
<td>3</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>30</td>
<td>4</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>40</td>
<td>5</td>
<td>5</td>
<td>6</td>
</tr>
</tbody>
</table>

Patrick Breheny & Jian Huang
Bi-level feature selection with applications to genetic association studies
# of groups selected

![Graph showing the number of groups selected for different numbers of nonzero members per group for gBridge, gLasso, and gSCAD.]
Group misclassification loss

The image shows a graph comparing the group misclassification loss (GML) for three different bi-level feature selection methods: gBridge, gLasso, and gSCAD. The x-axis represents the number of nonzero members per group, ranging from 2 to 10, and the y-axis represents the GML values ranging from 0.05 to 0.25.

- gBridge is represented by red lines.
- gLasso is represented by blue lines.
- gSCAD is represented by green lines.

The graph illustrates how the GML decreases as the number of nonzero members per group increases for all three methods.
Individual misclassification loss

![Graph showing individual misclassification loss for gBridge, gLasso, and gSCAD. The x-axis represents the number of nonzero members per group (2, 3, 4, 5, 6, 7, 8, 9, 10), and the y-axis represents IML (0.5, 1.0, 1.5). The lines for each method show different trends as the number of nonzero members increases.]
Genetic association studies are an increasingly important tool for detecting links between genetic markers and diseases.

The data I will analyze here come from a case-control study of age-related macular degeneration consisting of 400 cases and 400 controls (Sheffield, Stone, Cassavant, Mullins).

We confined our analysis to 30 genes that previous biological studies have suggested may be related to the disease.

These genes contained 532 markers with acceptably low rates of missing data (< 20% no call rate) and high minor allele frequency (≥ 10%).
We analyzed the data with the group lasso, group bridge, and group SCAD methods by considering markers to be grouped by the gene they belong to.

Logistic regression models were fit assuming an additive effect for all markers.

Missing data was imputed from the nearest non-missing marker for that subject.

We compared the group penalization methods to a traditional one-at-a-time approach:

- Markers were selected using a $p < .05$ cutoff for the likelihood ratio test of the univariate model.
- To assess error rate, an unpenalized model was then fit using the selected markers.
## Results

<table>
<thead>
<tr>
<th>Method</th>
<th># of groups</th>
<th># of covariates</th>
<th>Error rate</th>
<th>Test error rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>One-at-a-time</td>
<td>19</td>
<td>49</td>
<td>.312</td>
<td>.445</td>
</tr>
<tr>
<td>Group lasso</td>
<td>12</td>
<td>221</td>
<td>.324</td>
<td>.436</td>
</tr>
<tr>
<td>Group bridge</td>
<td>4</td>
<td>38</td>
<td>.342</td>
<td>.418</td>
</tr>
<tr>
<td>Group SCAD</td>
<td>7</td>
<td>23</td>
<td>.346</td>
<td>.424</td>
</tr>
</tbody>
</table>
Conclusions

- Using the LCD algorithm, it is quite feasible to apply complex group penalties to very large datasets.
- Group SCAD has promising properties for both group and individual variable selection.
- Grouped penalties are easily adapted to genetic association studies and yield results that are both more accurate and more interpretable than one-at-a-time approaches.