Group exponential penalties for bi-level variable selection

Patrick Breheny

Department of Biostatistics
Department of Statistics
University of Kentucky

July 31, 2011
In regression, variables can often be thought of as grouped:
- Indicator variables for a single categorical variable
- Basis functions derived from a single continuous variable
- Grouped by exterior knowledge (e.g. genetic variants by gene, genes by pathway)

In the penalized regression framework, this issue was first addressed by Yuan and Lin (2006), who proposed the group lasso:

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

Patrick Breheny
One limitation of the group lasso is that it selects entire groups at once – *i.e.*, either an entire group of coefficients is zero or they are all nonzero.

To extend the group lasso idea and allow for bi-level selection, Breheny and Huang (2009) proposed a general framework of group penalization, in which the penalty applied to the $j$th group consists of an outer penalty $f_O$ and an inner penalty $f_I$:

$$f_O \left( \sum_{k=1}^{K_j} f_I(|\beta_{jk}|) \right)$$

This framework encompassed existing approaches such as the group lasso, group bridge (Huang et al. 2009), and group SCAD (Wang, Chen, and Li 2007), as well as suggesting a new approach, group MCP.
In this talk, I will propose a new class of group penalties (also fitting into the aforementioned framework) based on exponential penalties, and

- Illustrate why they are attractive in theory
- Demonstrate that they perform well in simulation
- Apply them to the problem of genetic association studies of rare variants
To begin, let us define the exponential penalty:

\[ p(x|\lambda, \tau) = \frac{\lambda^2}{\tau} \left\{ 1 - \exp \left( -\frac{\tau}{\lambda} x \right) \right\} \]
Note that the partial derivative with respect to the $j_k$th covariate is

$$f_O' \left\{ \sum_{k=1}^{K_j} f_I(|\beta_{jk}|) \right\} f_I'(\beta_{jk})$$

Thus, if we choose the outer penalty to be exponential, we can control the rate of penalty relaxation applied to a coefficient if it is located in an important group.

We refer to this notion (coefficients are more likely to be selected if they are in an important group) as *coupling*, and consequently refer to $\tau$ as the *coupling parameter*, as it controls the magnitude of this phenomenon.

In principle, any penalty could be chosen as the inner penalty – we focus here on the lasso (*geLasso*) and MCP (*geMCP*)
As an illustration for a group of two covariates, consider the effect of increasing $\beta_1$ on the selection threshold for $\beta_2$: 

![Coupling plots](image)

- **Lasso**
- **Group lasso**
- **Group MCP**
- **geLasso**

Coupling plots
It can be shown that

\[
\lambda^2 \frac{1}{\tau} \left\{ 1 - \exp \left( -\frac{\tau}{\lambda^2} \sum_{k=1}^{K_j} f_I(|\beta_{jk}|) \right) \right\} = \sum_{k=1}^{K_j} f_I(|\beta_{jk}|) + O(\tau)
\]

i.e., as \( \tau \to 0 \), geLasso is equivalent to the Lasso, geMCP is equivalent to MCP, and so on.

Furthermore, the objective function is to remain convex in every coordinate dimension, there are also upper bounds on \( \tau \):

\[
\begin{cases}
\tau \leq 1 & \text{geLasso} \\
\tau + \frac{1}{\gamma} \leq 1 & \text{geMCP}
\end{cases}
\]
Group exponential penalties can be fit via the local coordinate descent algorithm, introduced in Breheny and Huang (2009); this approach has the following attractive property:

**Proposition**

At every step of the LCD algorithm for both geLasso and geMCP,

$$Q(\beta^{(m+1)}) \leq Q(\beta^{(m)})$$  \hspace{1cm} (1)

i.e., the algorithms decrease the objective function at every iteration.
Simulation study

To examine the statistical properties of group exponential penalties, we carried out a simulation study with the following design:

- \( n = p = 100 \), with 10 groups, and 10 covariates in each group
- True coefficients were heterogeneous so that there were strongly predictive and weakly predictive groups, as well as strongly and weakly predictive covariates within a group, but the overall SNR was always 1
- There were three nonzero groups, but the number of nonzero covariates within a group varied
- \( \gamma = 4 \) for gMCP and geMCP; \( \tau = 1/3 \) for geLasso and geMCP; BIC was used to select \( \lambda \)
Simulation results – MSE and model size

Simulation GAW 2010 data

Group exponential penalties

Nonzero covariates/group

MSE

2 4 6 8 10

gLasso  gMCP  geLasso  geMCP

Nonzero covariates/group

Model size

0 5 10 15 20 25 30

2 4 6 8 10
Simulation results – variable selection

- **gLasso**
- **gMCP**
- **geLasso**
- **geMCP**

<table>
<thead>
<tr>
<th>Nonzero covariates/group</th>
<th>FDR (covariates)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.8</td>
</tr>
<tr>
<td>4</td>
<td>0.6</td>
</tr>
<tr>
<td>6</td>
<td>0.4</td>
</tr>
<tr>
<td>8</td>
<td>0.2</td>
</tr>
<tr>
<td>10</td>
<td>0.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Nonzero covariates/group</th>
<th>FDR (groups)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.8</td>
</tr>
<tr>
<td>4</td>
<td>0.6</td>
</tr>
<tr>
<td>6</td>
<td>0.4</td>
</tr>
<tr>
<td>8</td>
<td>0.2</td>
</tr>
<tr>
<td>10</td>
<td>0.0</td>
</tr>
</tbody>
</table>
One important potential application of this method is to the problem of rare variants in genetic association studies. Briefly, the issue in such studies is that, while the collective effect of multiple rare genetic variants may be large, each variant is sufficiently rare that association studies have low power to detect them. Currently, most methods for addressing this problem involve “collapsing” the rare variants into a single test. This approach has the considerable disadvantage of assuming that every variant has precisely the same effect on the outcome.
An alternative approach is to use a group penalization approach in which the variants are grouped by the gene they belong to.

To test this approach on real(istic) data, we analyzed the data set from the 2010 Genetic Analysis Workshop (GAW).

The data set contains real data from the 1000 genomes project on 697 unrelated individuals and 24,487 genetic variants, grouped into 3,205 genes.

200 independent sets of outcomes were simulated by the organizers of the workshop according to a plausible genetic model of variant-disease association.
GAW results

There were 39 causal variants; each method was allowed to select 39 covariates

<table>
<thead>
<tr>
<th>Method</th>
<th>Variants selected</th>
<th>Genes selected</th>
<th>Variants correct (%)</th>
<th>Genes correct (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lasso</td>
<td>39.2</td>
<td>35.9</td>
<td>10.4</td>
<td>5.7</td>
</tr>
<tr>
<td>gLasso</td>
<td>248.8</td>
<td>2.2</td>
<td>0.1</td>
<td>1.1</td>
</tr>
<tr>
<td>gLasso($\sqrt{K_j}$)</td>
<td>41.0</td>
<td>34.8</td>
<td>1.1</td>
<td>1.3</td>
</tr>
<tr>
<td>gMCP</td>
<td>42.6</td>
<td>39.5</td>
<td>9.1</td>
<td>5.0</td>
</tr>
<tr>
<td>geLasso</td>
<td>39.6</td>
<td>6.9</td>
<td>28.3</td>
<td>21.6</td>
</tr>
<tr>
<td>geMCP</td>
<td>44.1</td>
<td>7.7</td>
<td>25.0</td>
<td>19.2</td>
</tr>
</tbody>
</table>
Conclusions

- Group exponential penalties are both theoretically attractive and practically useful.
- The framework allows control over the degree of grouping (coupling).
- The idea extends readily to GLMs.
- The method is publicly available via the grpreg package.