Within Group Variable Selection through the Exclusive Lasso

Genevera I. Allen

Dobelman Family Junior Chair,
Department of Statistics and Electrical and Computer Engineering, Rice University,
Department of Pediatrics-Neurology, Baylor College of Medicine,
Jan and Dan Duncan Neurological Research Institute, Texas Children’s Hospital.

October 9, 2015

Joint Work with Frederick Campbell.
Motivation

Problems with a predefined group structure.

- Genomics.
  - Genes grouped by function / pathways.
- Finance.
  - Stocks grouped by sector.
- Chemistry.
  - Molecules grouped by chemical structure.
Motivation

Motivating Example - Index Fund.

Can we choose the best diverse Index Fund?

- Technology.
  - Apple, Intel, Microsoft, etc.
- Financial Services.
  - Goldman Sachs, UBS, etc.
- Energy.
  - Exxon, Halliburton, Philips 66, etc.

Objective

Sparse Regression: Select at least one variable per group.
Why not the Lasso?

Objective

Sparse Regression: Select at least one variable per group.


- Fails to respect group structure.
- Fails when strong correlations between and/or within groups.
Our Objective

Sparse Regression: Select at least one variable per group.

Others suggested an **Exclusive Lasso** penalty for selecting one variable per group.

- Zhang et al. (2009) - proposed penalty for multi-task learning (Application).
- Obozinski and Bach (2012) - showed penalty was tightest convex relaxation of NP-hard problem (Optimization Theory).
- Halabi and Cevher (2014) - showed when penalty recovers NP-hard global solution (Optimization Theory).

**Our Goal:** Statistically study Exclusive Lasso for sparse regression.
1 Motivation

2 Exclusive Lasso
   - Problem & Solution
     - Estimation
     - Theory
     - Model Selection

3 Results
   - Simulation Studies
   - NMR Spectroscopy
Exclusive Lasso for Sparse Regression

Set-up:

- Linear Model: \( y = X\beta + \epsilon, \ \epsilon \sim \text{Gaussian} \).
- Known, disjoint groups: \( G \).

Exclusive Lasso Estimate

\[
\hat{\beta} = \underset{\beta}{\text{argmin}} \quad \|y - X\beta\|_2^2 + \frac{\lambda}{2} \sum_{g \in G} \|\beta_g\|_1
\]
Exclusive Lasso Penalty

\[ P(\beta) = \frac{1}{2} \sum_{g \in G} \| \beta_g \|_1^2 \]

- $\ell_1$-norm within groups.
  - Sparsity within groups (exclusivity).
- $\ell_2$-norm between groups.
  - Density between groups.

Opposite of Group Lasso ($\ell_2$ within and $\ell_1$ between).
Exclusive Lasso Penalty

3-dimensional, 2-group Example:

\[ \beta = (\beta_{1,1}, \beta_{1,2}, \beta_{2,1}) \]

\[ \ell_1 \text{ Within: } \beta_{1,1} \text{ vs. } \beta_{1,2}. \]
Exclusive Lasso Penalty

3-dimensional, 2-group Example:

\[ \beta = (\beta_{1,1}, \beta_{1,2}, \beta_{2,1}) \]

\( \ell_2 \) Between: \( \beta_{1,1} \) vs. \( \beta_{2,1} \).
Exclusive Lasso Penalty

3-dimensional, 2-group Example:

$$\beta = (\beta_{1,1}, \beta_{1,2}, \beta_{2,1})$$
Exclusive Lasso Behavior

- Always selects at least one variable per group.
- Never sends entire groups to zero.
- Behaves like an adaptively shrunken ridge regression problem.
  - Group-wise adaptive shrinkage to achieve sparsity.
  - Solution characterized in terms of the active set.
Exclusive Lasso Behavior

- Always selects at least one variable per group.
- Never sends entire groups to zero.
- Behaves like an adaptively shrunken ridge regression problem.
  - Group-wise adaptive shrinkage to achieve sparsity.
  - Solution characterized in terms of the active set.

Illustrative Simulation:

- $p = 30$ variables & $n = 100$ observations.
- 5 groups; 1 true variable per group.
- Correlation both within and between groups.
Exclusive Lasso Behavior

Regularization Paths: Exclusive Lasso.

Coefficient values vs. Log(lambda) value for different groups.
Exclusive Lasso Behavior

Regularization Paths: Exclusive Lasso - Group 1.
Exclusive Lasso Behavior

Regularization Paths: Lasso.

Coefficient values vs Log(lambda) value
1 Motivation

2 Exclusive Lasso
   - Problem & Solution
   - Estimation
     - Theory
     - Model Selection

3 Results
   - Simulation Studies
   - NMR Spectroscopy
Exclusive Lasso Estimate

\[ \hat{\beta} = \arg\min_{\beta} \| y - X\beta \|^2_2 + \frac{\lambda}{2} \sum_{g \in G} \| \beta_g \|^2_1 \]
Exclusive Lasso Estimation

Proximal Gradient Descent Scheme:

- Usual gradient update for sparse regression.
- **Problem**: Proximal operator has no closed form solution.
- **Solution**: Group-wise coordinate descent scheme!

\[
\beta_{i,g}^{k+1} = S \left( \frac{1}{1 + \lambda} z_{i,g}, \frac{\lambda}{1 + \lambda \| \beta_{-i} \|_1} \right)
\]

- **Theorem**: Converges to proximal operator.

Advantages:

- Parallelizable over groups.
- **Theorem**: Converges at similar rates to comparable algorithms for the Lasso.
Algorithm 1: Exclusive Lasso Algorithm to fit the Exclusive Lasso

Input: $\beta^0 \in \mathbb{R}^p$, $\epsilon \in \mathbb{R}$, $\delta \in \mathbb{R}$

Output: $\hat{\beta} \in \mathbb{R}^p$

1. while $\|\beta^{k+1} - \beta^k\| > \epsilon$ do
2.   \[ z_g = \beta^k_g - \frac{1}{L} (X^T X \beta^k - X^T y) \]
3.   In parallel for each $g$:
4.   Initialize $\tilde{\beta}_g \in \mathbb{R}^{p_g}$
5.   while $\|\tilde{\beta}_g^{t+1} - \tilde{\beta}_g^t\| > \delta$ do
6.     for $i \leftarrow 1$ to $p_g$ do
7.       $\beta_{g,i}^{t+1} = S(\frac{1}{\lambda + 1} z_{g,i}, \frac{\lambda}{\lambda + 1} \|\tilde{\beta}_g^{-i}\|_1)$
8.     $\beta_{g}^{k+1} = \tilde{\beta}_g$
9. return $\beta$
1 Motivation

2 Exclusive Lasso
   - Problem & Solution
   - Estimation
   - Theory
   - Model Selection

3 Results
   - Simulation Studies
   - NMR Spectroscopy
Statistical Theory

Assumptions:
- $|X_{i,j}| \leq M$, distribution of $X$ has a covariance matrix
- $P(\beta^*) \leq K$
- $y = X\beta^* + \epsilon$ such that $\epsilon \sim N(0, \sigma^2 I)$

Theorem (Prediction Consistency)

$$\mathbb{E}(X\beta^* - X\hat{\beta}) \leq 2(K + |G|)M\sigma \sqrt{\frac{2\log(2p)}{n}} + 8(K + |G|)^2 M^2 \sqrt{\frac{2p \log(2p^2)}{n}}$$

which goes to 0 as

$$n \to \infty$$
Additional Assumption:

- $\lambda_{\text{min}}(\Sigma) > c$ for some $c > 0$.

**Theorem (Consistency)**

\[
\|\beta^* - \hat{\beta}\|_2^2 \leq \frac{2}{c} (K + |G|) M \sigma \sqrt{\frac{2 \log(2p)}{n}} + \frac{8}{c} (K + |G|)^2 M^2 \sqrt{\frac{2p \log(2p^2)}{n}}
\]

which goes to 0 as

\[n \to \infty\]
1 Motivation

2 Exclusive Lasso
   • Problem & Solution
   • Estimation
   • Theory
   • Model Selection

3 Results
   • Simulation Studies
   • NMR Spectroscopy
Model Selection

Data-driven way to select $\lambda$?

- Cross-validation / Stability Selection problematic.
- BIC / EBIC.
  - Need degrees of freedom.
Model Selection

Data-driven way to select $\lambda$?

- Cross-validation / Stability Selection problematic.
- BIC / EBIC.
  - Need degrees of freedom.

**Theorem: Degrees of Freedom**

\[
\text{Df}(\hat{y}) = \mathbb{E} \left[ \text{trace} (X_S (X_S^T X_S + M)^\dagger X_S^T) \right]
\]

($M$ is a block diagonal matrix where each block is $\text{sign}(\beta_g) \text{sign}(\beta_g)^T$.)
Model Selection
1 Motivation

2 Exclusive Lasso
   • Problem & Solution
   • Estimation
   • Theory
   • Model Selection

3 Results
   • Simulation Studies
   • NMR Spectroscopy
Simulation Studies

Set-up:
- Vary within and between group correlation.
- Comparison Methods: Lasso, Marginal Regression, Group-wise variations of these.

Summary of Results:
- All methods do well in limited correlation settings.
- Exclusive Lasso better performance when either high within group or between group correlations.
  - Prediction Error.
  - Variable Selection Accuracy.
  - Respecting Group Structure.
## Simulation Studies

<table>
<thead>
<tr>
<th></th>
<th>Exclusive Lasso</th>
<th>Lasso</th>
<th>Marginal Regression</th>
<th>Group-wise Marginal Regression</th>
<th>Thresholded Exclusive Lasso</th>
<th>Thresholded Lasso</th>
<th>Thresholded Regularization Path</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>w = .9, b = .9</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>True Vars</td>
<td>2.180 (1.02)</td>
<td>2.160 (0.82)</td>
<td>1.340 (0.63)</td>
<td>1.500 (0.84)</td>
<td>3.760 (0.96)</td>
<td>1.760 (1.06)</td>
<td>2.080 (0.97)</td>
</tr>
<tr>
<td>False Vars</td>
<td>2.820 (1.02)</td>
<td>2.840 (0.82)</td>
<td>3.660 (0.63)</td>
<td>3.500 (0.84)</td>
<td>1.240 (0.96)</td>
<td>3.240 (1.06)</td>
<td>2.920 (0.97)</td>
</tr>
<tr>
<td>Pred Err</td>
<td>1.351 (0.15)</td>
<td>1.433 (0.13)</td>
<td>1.608 (0.14)</td>
<td>1.411 (0.12)</td>
<td><strong>1.115</strong> (0.13)</td>
<td>1.411 (0.17)</td>
<td>1.325 (0.15)</td>
</tr>
<tr>
<td><strong>w = .9, b = .6</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>True Vars</td>
<td>3.86 (0.88)</td>
<td>3.700 (0.81)</td>
<td>2.10 (0.74)</td>
<td>4.020 (0.82)</td>
<td>4.480 (0.68)</td>
<td>4.060 (1.10)</td>
<td>3.96 (0.90)</td>
</tr>
<tr>
<td>False Vars</td>
<td>1.14 (0.88)</td>
<td>1.300 (0.81)</td>
<td>2.90 (0.74)</td>
<td>0.980 (0.82)</td>
<td>0.520 (0.68)</td>
<td>0.940 (1.10)</td>
<td>1.04 (0.90)</td>
</tr>
<tr>
<td>Pred Err</td>
<td>1.11 (0.10)</td>
<td>1.236 (0.17)</td>
<td>1.55 (0.16)</td>
<td>1.102 (0.11)</td>
<td><strong>1.064</strong> (0.09)</td>
<td>1.129 (0.15)</td>
<td>1.10 (0.11)</td>
</tr>
<tr>
<td><strong>w = .6, b = .6</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>True Vars</td>
<td>4.720 (0.50)</td>
<td>4.600 (0.53)</td>
<td>3.620 (0.53)</td>
<td>4.200 (0.49)</td>
<td>4.940 (0.24)</td>
<td>4.720 (0.45)</td>
<td>4.740 (0.44)</td>
</tr>
<tr>
<td>False Vars</td>
<td>0.280 (0.50)</td>
<td>0.400 (0.53)</td>
<td>1.380 (0.53)</td>
<td>0.800 (0.49)</td>
<td>0.060 (0.24)</td>
<td>0.280 (0.45)</td>
<td>0.260 (0.44)</td>
</tr>
<tr>
<td>Pred Err</td>
<td>1.066 (0.15)</td>
<td>1.094 (0.15)</td>
<td>1.304 (0.15)</td>
<td>1.162 (0.15)</td>
<td><strong>1.022</strong> (0.10)</td>
<td>1.062 (0.13)</td>
<td>1.057 (0.13)</td>
</tr>
</tbody>
</table>
1 Motivation

2 Exclusive Lasso
   - Problem & Solution
   - Estimation
   - Theory
   - Model Selection

3 Results
   - Simulation Studies
   - NMR Spectroscopy
NMR Spectroscopy

- High-throughput technology to measure small molecules (metabolites) in a sample.
- Spectra is a linear combination of molecule chemical signatures.

Figure 1. Left: Spectra for biological sample. These samples can be composed of up to 5000 unique molecules. We believe the spectra is a linear combination of the component molecules respective spectra. The neuron sample spectra is a linear combination of sucrose and acetaminophen among other possible component molecules. Right: A chemical shift for the molecule carnosine. Chemical shifts occur due to the chemical environment of the molecule being measured. The positional uncertainty introduced by chemical shifts complicates the identification and quantification problems.

Figure 2. The unit ball for the Exclusive Lasso Penalty from 3 different perspectives. Consider the following example. Let $\mathbf{x} = (x_1, 1, x_1, 2, x_2, 1)$ be parameters such that the first index denotes group membership and the second denotes the element within group. Left: Figure considers only $x_1, 1$ and $x_1, 2$ and it is equivalent to the $\ell_1$ unit ball. Middle: Figure considers only $x_1, 1$ and $x_2, 1$ showing that it is equivalent to the $\ell_2$ unit ball in these dimensions. Right: Figure shows the ball in all 3 dimensions. We know the structure enforced in the estimate is connected to the extreme points of the constraint space. In the Exclusive Lasso's case, we have sparsity within group and no structure between groups because the Exclusive Lasso is the Lasso in some dimensions and Ridge Regression in other dimensions.
NMR Spectroscopy

Objective: Quantify concentrations of known molecules in sample.

Problem: Spectra of known molecular subject to random positional shifts.
NMR Spectroscopy

**Solution:** For each molecule, create a group of lagged possible shifts to represent position uncertainty. (Expanded Dictionary)

![Covariance of Expanded Dictionary](image)

Covariance of Expanded Dictionary
Objective

Out of a group of positional shifts for each molecule, use the Exclusive Lasso to select the best shift for quantifying molecule concentrations.
NMR Simulation

Simulate spectra using 33 molecules with different concentrations.

NMR Signal Recovery Results:

![Graph showing simulated and estimated signals](image-url)
NMR Simulation

NMR Metabolite Concentration Results:

<table>
<thead>
<tr>
<th>Method</th>
<th>Mean Squared Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exclusive Lasso</td>
<td>1.11</td>
</tr>
<tr>
<td>OLS Regression</td>
<td>2.34</td>
</tr>
<tr>
<td>Lasso</td>
<td>2.41</td>
</tr>
<tr>
<td>Marginal Regression</td>
<td>3.26</td>
</tr>
</tbody>
</table>

MSE = average squared quantification error.
Summary

- Exclusive Lasso effective for selecting at least one variable per group.
- Major advantages when correlations within or between groups.

Future Work

- Overlapping / Hierarchical group structures.
- Group exclusivity AND selection.
- Study other loss functions.
Acknowledgments:

Joint work with Frederick Campbell, PhD Candidate, Rice University.

- NSF Graduate Research Fellowship.
- NSF DMS-1209017.

Reference: