An Efficient Algorithm For Weak Hierarchical Lasso

Yashu Liu, Jie Wang, Jieping Ye

Arizona State University
Outline

• Regression with Interactions
• Problems and Challenges
• Weak Hierarchical Lasso
• The Proposed Method
• Experimental Results
• Conclusion
• Future Work
Linear Regression

- Linear Regression is a widely used tool in data mining and machine learning areas.
- A linear regression model is:

$$y = x^T w + \epsilon$$

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Feature</th>
<th>Coefficients</th>
<th>Noise</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y$</td>
<td>$x^T$</td>
<td>$w$</td>
<td>$\epsilon$</td>
</tr>
</tbody>
</table>
Regression with Interactions

- Regression with only linear effects may not be sufficient
- One effective way to capture the non-linearity is to include interactions

\[ y = x^T w + \frac{1}{2} x^T Q x + \epsilon \]
Issues of Including Interactions

• ISSUE 1:
  Including interactions leads to ultra high dimensionality
  – e.g., 100 features will result in 10,000 interactions

• ISSUE 2:
  Hierarchical structural estimator is typically difficult to obtain
Related Work

- Ling Yan et al. 2014 used low-rank interaction model for click through rate prediction
- Peter Radchenko et al. 2010 proposed a Lasso Type penalty VANISH to achieve strong hierarchical structure
- Ming Yuan et al. 2009 proposed a type of non-negative garrote method to achieve hierarchical structures
- Peng Zhao et al. 2009 proposed Composite Absolute Penalties for hierarchical interaction models
- Nam Hee Choi et al. 2010 formulated the interaction coefficients as the product of main effects coefficients to achieve strong hierarchy
Weak Hierarchical Lasso
(Jacob et al. 2013 Annals of Statistics)

- Weak Hierarchy
  - An interaction is selected only if at least one of the main effects is included in the model
  - $Q_{ij} \neq 0$ only if $w_i \neq 0$ OR $w_j \neq 0$. 

![Diagram of interactions between Age, Gender, and MMSE]

- Age × Gender
- Age × MMSE
- Gender × MMSE
Weak Hierarchical Lasso

- Weak Hierarchical Constraints
  - \[ \|Q_j\|_1 \leq |w_j| \text{ for } j = 1, \ldots, d \]

When \( j = 1 \),

\[ w \]

\[ Q \]
Weak Hierarchical Lasso

- Weak Hierarchical Constraints
  - \( \|Q_j\|_1 \leq |w_j| \) for \( j = 1, ..., d \)

When \( j = 2 \),
Weak Hierarchical Lasso

- Let $L(w, Q)$ be the loss function for interaction regression.
- The Weak Hierarchical Lasso formulation:
  $$\min_{w,Q} L(w, Q) + \lambda \|w\|_1 + \frac{\lambda}{2} \|Q\|_1$$
  s.t. $\|Q_{.,j}\|_1 \leq |w_j|$, \quad \forall j = 1, \ldots, d.$

- Constraint $\|Q_{.,j}\|_1 \leq |w_j|$ makes the optimization problem non-Convex.
Weak Hierarchical Lasso

- Jacob et al. 2013 propose to solve a relaxed version:

\[
\min_{w,Q} L(w^+ - w^-, Q) + \lambda 1^T(w^+ + w^-) + \frac{\lambda}{2} \|Q\|_1
\]

\[
\|Q_{.j}\|_1 \leq w^+_j + w^-_j
\]

s.t. \quad \begin{align*}
    w^+_j &\geq 0 \\
    w^-_j &\geq 0
\end{align*}

\begin{align*}
    j &= 1, \ldots, d
\end{align*}

- \(\|w\|_1\) is relaxed to \(w^+ + w^-\) such that the optimization problem is convex

- The relaxed formulation guarantees a weak hierarchical solution only if a ridge penalty exists
Main Contributions

• We propose to directly solve the non-convex Weak Hierarchical Lasso by a proximal algorithm

• We show that the associated proximal operator admits a closed-form solution

• We accelerate the computation of each sub-problem of the proximal operator from quadratic to linearithmic
Proximal Operator

Weak Hierarchical Lasso:

\[
\min_{w,Q} L(w, Q) + \lambda \|w\|_1 + \frac{\lambda}{2} \|Q\|_1
\]

\[
s.t. \quad \|Q_{:,j}\|_1 \leq |w_j|, \quad j = 1, \ldots, d
\]

Proximal Operator:

\[
\arg\min_{w,Q} \frac{1}{2} \|w - v\|^2_F + \frac{1}{2} \|Q - U\|^2_F + \frac{\lambda}{t} \|w\|_1 + \frac{\lambda}{2t} \|Q\|_1
\]

\[
s.t. \quad \|Q_{:,j}\|_1 \leq |w_j|, \quad j = 1, \ldots, d
\]
Factorizing Unknown Variables

• The **KEY IDEA** is to factorize unknown coefficients \( w \) and \( Q \) by their signs \( S_w, S_Q \) and magnitudes \( \tilde{w} \) and \( \tilde{Q} \)

• Unknown coefficients \( w \) and \( Q \) are factorized as

\[
\begin{align*}
    w &= S_w \odot \tilde{w} \\
    Q &= S_Q \odot \tilde{Q}
\end{align*}
\]
Proximal Operator Subproblem

• The proximal operator problem can be decoupled

\[
\begin{align*}
    w_j &= S_{w_j} \tilde{w}_j \\
    Q_{.j} &= S_{Q.j} \tilde{Q}_{.j}
\end{align*}
\]

• By factorization, the sub-problem is:

\[
\min_{s_{w_j}, s_{Q_j}, \tilde{w}_j, \tilde{Q}_{.j}} \frac{1}{2} \left\| S_{w_j} \tilde{w}_j - v_j \right\|^2_2 + \frac{1}{2} \left\| S_{Q_j} \tilde{Q}_{.j} - U_{.j} \right\|^2_F + \frac{\lambda}{t} \tilde{w}_j + \frac{\lambda}{2t} 1^T \tilde{Q}_{.j}
\]

s.t.

\[
\begin{align*}
    1^T \tilde{Q}_{.j} &\leq \tilde{w}_j \\
    \tilde{Q}_{.j} &\geq 0
\end{align*}
\]

\[j = 1, \ldots, d\]
Solving Sign Variables

• We solve sign variables $S_w, S_Q$ and magnitude variables $\tilde{w}, \tilde{Q}$ together

• Since

$\frac{1}{2} (w_j - v_j)^2 = \frac{1}{2} (S_{w,j}\tilde{w}_j - v_j)^2$

$= \frac{1}{2} (S_{w,j} (S_{w,j}\tilde{w}_j - v_j))^2$

$= \frac{1}{2} (\tilde{w}_j - S_{w,j}v_j)^2$

then $S_w = \text{sign}(v)$

• $S_{Q,j}$ must be the sign of $U_{,j}$, i.e., $S_{Q,j} = \text{sign}(U_{,j})$
Solving Magnitude Variables

• The magnitude variables can thus be obtained by solving

\[
\min_{\tilde{Q}, \tilde{w}} \frac{1}{2} \| \tilde{w}_j - \tilde{v}_j \|_2^2 + \frac{1}{2} \| \tilde{Q} - \tilde{U}_j \|_2^2 + \frac{\lambda}{t} \tilde{w}_j + \frac{\lambda}{2t} 1^T \tilde{Q}_j
\]

\[
\text{s.t. } \begin{cases} 
1^T \tilde{Q}_j \leq \tilde{w}_j \\
\tilde{Q}_j \geq 0
\end{cases}
\]

where \( \tilde{v}_j = S_{w_j} v_j - \frac{\lambda}{t} \cdot 1 \) and \( \tilde{U}_j = S_{Q_j} U_j - \frac{\lambda}{2t} \cdot 1 \)

• The above subproblem has closed form solutions
Accelerating Computation

- A straightforward implementation of solving the magnitude variables will take quadratic time
  - Not scalable for large scale data

- We further accelerate the time complexity of computing each proximal operator subproblem to linearithmic
  - Exploit special structure of the subproblem
Comparisons of Efficiency

- We compare the efficiency of our algorithm with the convex Weak Hierarchical Lasso in terms of running time and numbers of iterations.
Comparisons of Recovery

- We compare the recovery performances of the two algorithms.
### Performance on Real Data

- We compare the original Weak Hierarchical Lasso with the convex version and other baseline methods on ADNI data.

<table>
<thead>
<tr>
<th></th>
<th>Main Effects Only</th>
<th>Main Effects + Interactions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RF</td>
<td>SVM</td>
</tr>
<tr>
<td>Accuracy (%)</td>
<td>74.23 ± 8.67</td>
<td>75.22 ± 8.72</td>
</tr>
<tr>
<td>Sensitivity (%)</td>
<td>78.75 ± 14.00</td>
<td>80.18 ± 13.89</td>
</tr>
<tr>
<td>Specificity (%)</td>
<td>69.29 ± 11.63</td>
<td>69.76 ± 12.80</td>
</tr>
<tr>
<td>Accuracy (%)</td>
<td>71.26 ± 10.22</td>
<td>59.45 ± 14.43</td>
</tr>
<tr>
<td>Sensitivity (%)</td>
<td>83.04 ± 13.18</td>
<td>59.29 ± 17.83</td>
</tr>
<tr>
<td>Specificity (%)</td>
<td>58.10 ± 23.23</td>
<td>60.00 ± 15.42</td>
</tr>
</tbody>
</table>
Conclusion

• We study Weak Hierarchical Lasso which is popular for building non-linear regression models with interactions
  – non-convex, challenging to solve

• We propose a proximal algorithm to directly solve the Weak Hierarchical Lasso
  – We show that the associated proximal operator admits a closed form solution
  – We further accelerate the computation of each proximal operator subproblem from quadratic to linearithmic

• Empirical study demonstrated superior efficiency and effectiveness of the proposed algorithm
Future Work

• Apply the non-convex Weak Hierarchical Lasso to other challenging applications such as depression study

• Extend our proposed algorithm to solve Strong Hierarchical Lasso problems
Thank You!