An Efficient Algorithm For Weak Hierarchical Lasso

Yashu Liu, Jie Wang, Jieping Ye

Arizona State University

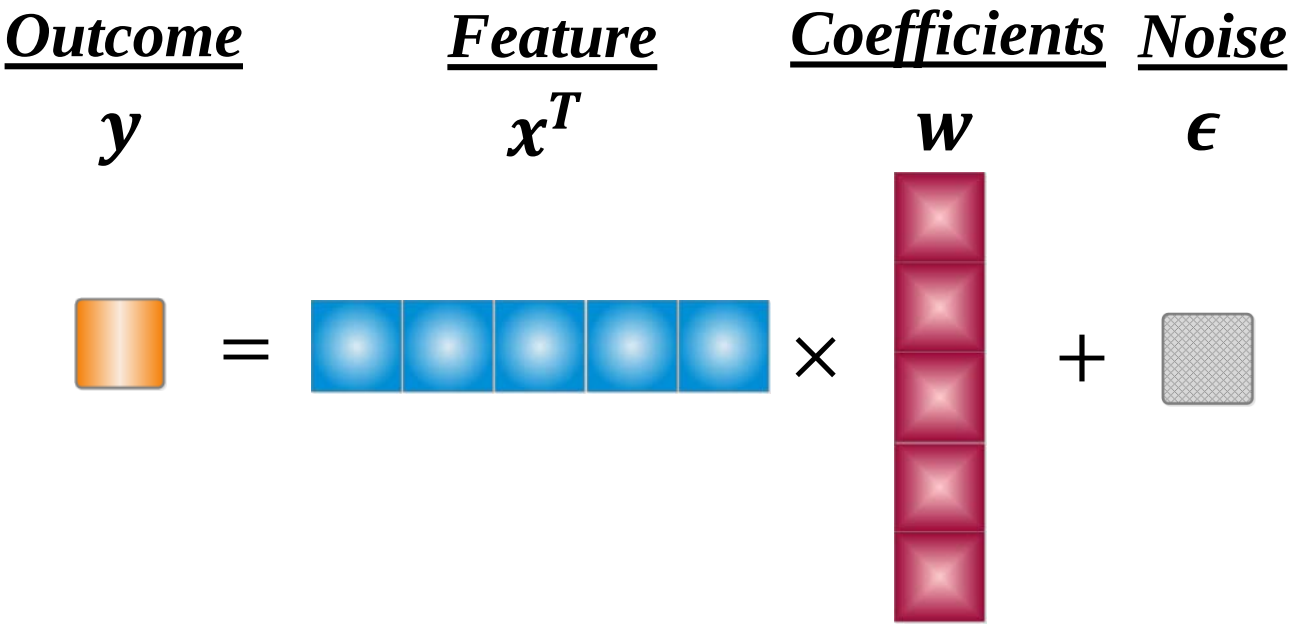
Outline

- 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

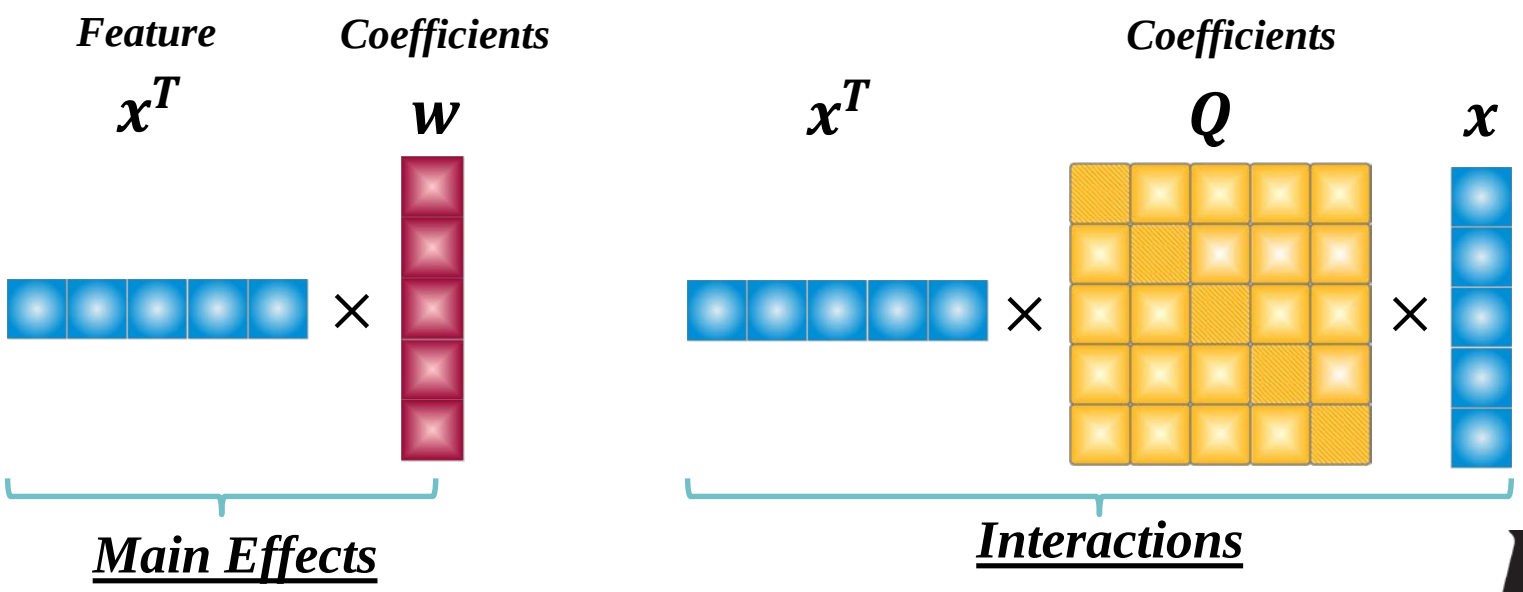
$$y = x^T w + \epsilon$$



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

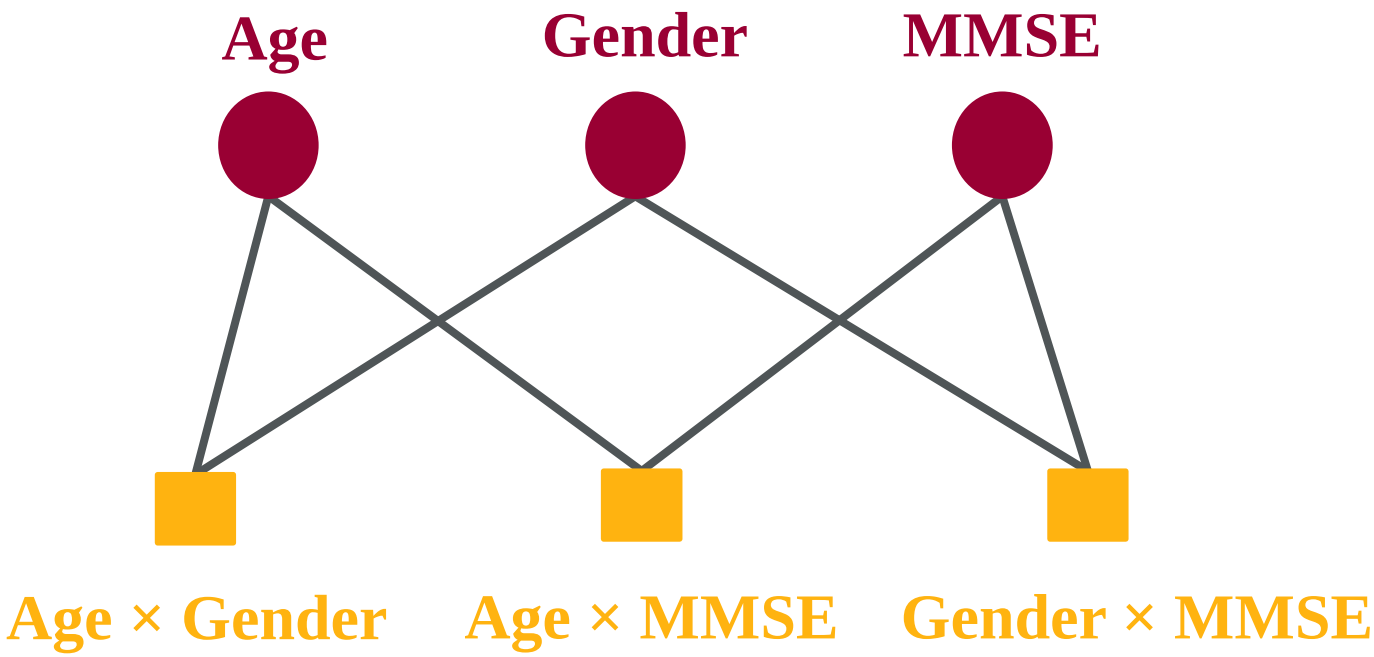
Related Work

- 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

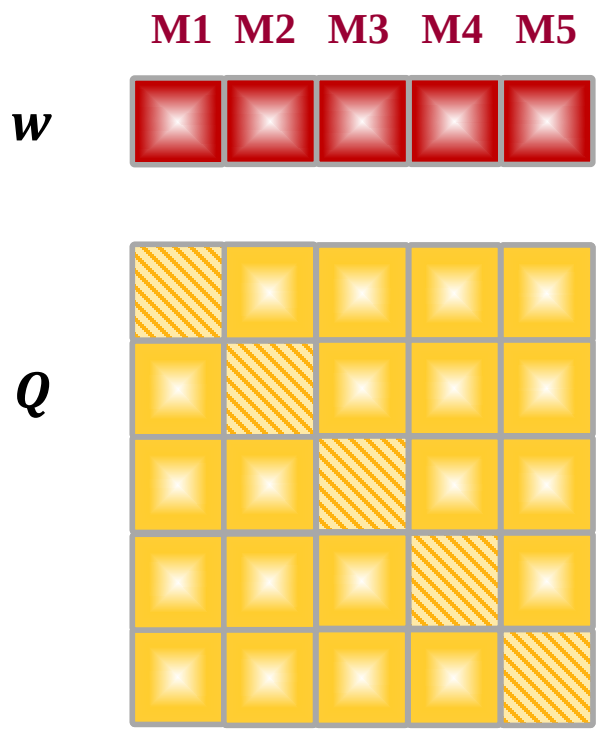
(Jacob *et al.* 2013 *Annals of Statistics*)

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

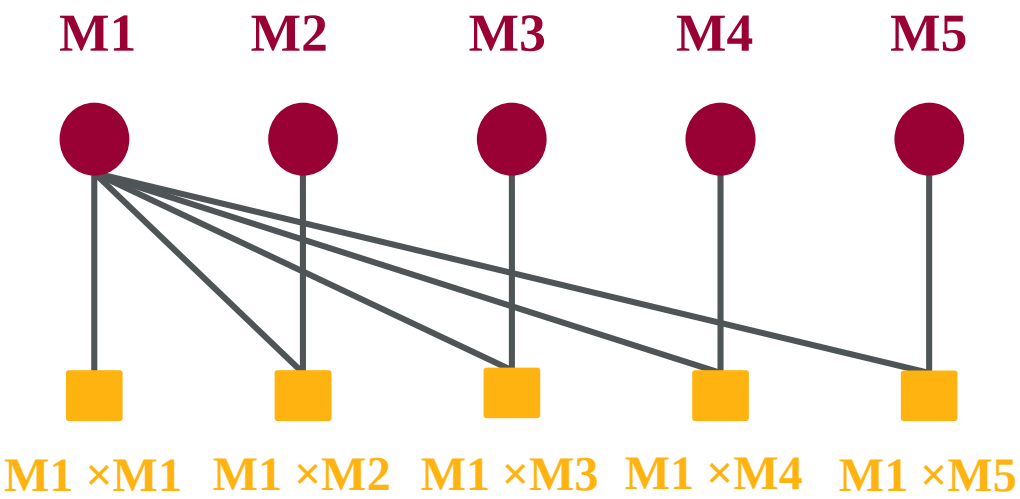


Weak Hierarchical Lasso

- Weak Hierarchical Constraints
  - $||\boldsymbol{Q}_{\cdot j}||_1 \leq |\boldsymbol{w}_j|$ **for** $\boldsymbol{j} = \mathbf{1}, \ldots, \boldsymbol{d}$

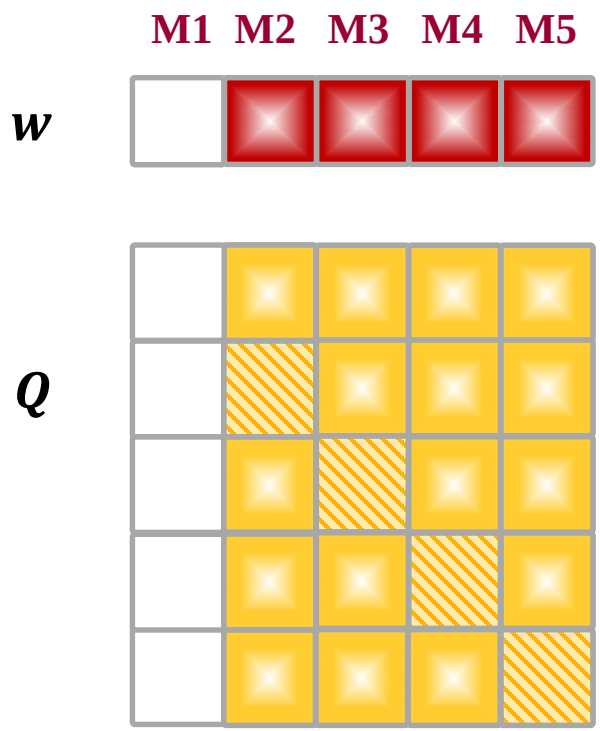


When $j = 1$,

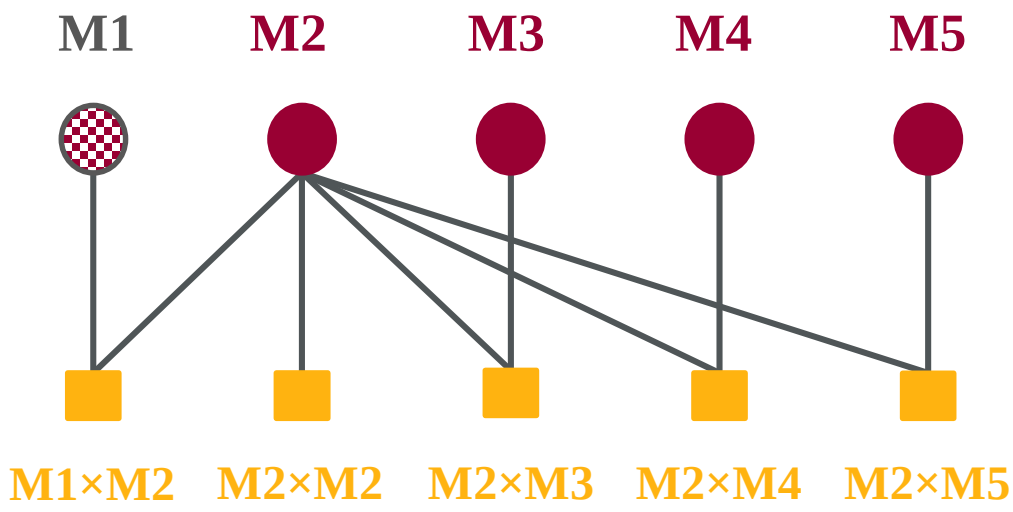


Weak Hierarchical Lasso

- Weak Hierarchical Constraints
  - $||\boldsymbol{Q}_{\cdot j}||_1 \leq |\boldsymbol{w}_j|$ **for** $\boldsymbol{j} = \mathbf{1}, \ldots, \boldsymbol{d}$



When $j = 2$,



Weak Hierarchical Lasso

- Let $L(w, Q)$ be the loss function for interaction regression
- The Weak Hierarchical Lasso formulation:

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

$$s.t. \ \left\|Q_{\cdot,j}\right\|_1 \leq \left|w_j\right|, \qquad j = 1, \dots, d.$$

- Constraint $\left\|Q_{\cdot,j}\right\|_1 \leq \left|w_j\right|$ makes the optimization problem ***non-Convex***

Weak Hierarchical Lasso

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

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

$$s.t. \quad \left.\begin{aligned} \|Q_{\cdot,j}\|_1 &\le w_j^+ + w_j^- \\ w_j^+ &\ge 0 \\ w_j^- &\ge 0 \end{aligned}\right\}, \qquad j = 1, \dots, d$$

- $\|w\|_1$ 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 :

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

$$s.t.\ \left\|Q_{\cdot,j}\right\|_1 \leq |w_j|, \qquad j = 1,\dots,d$$

Proximal Operator:

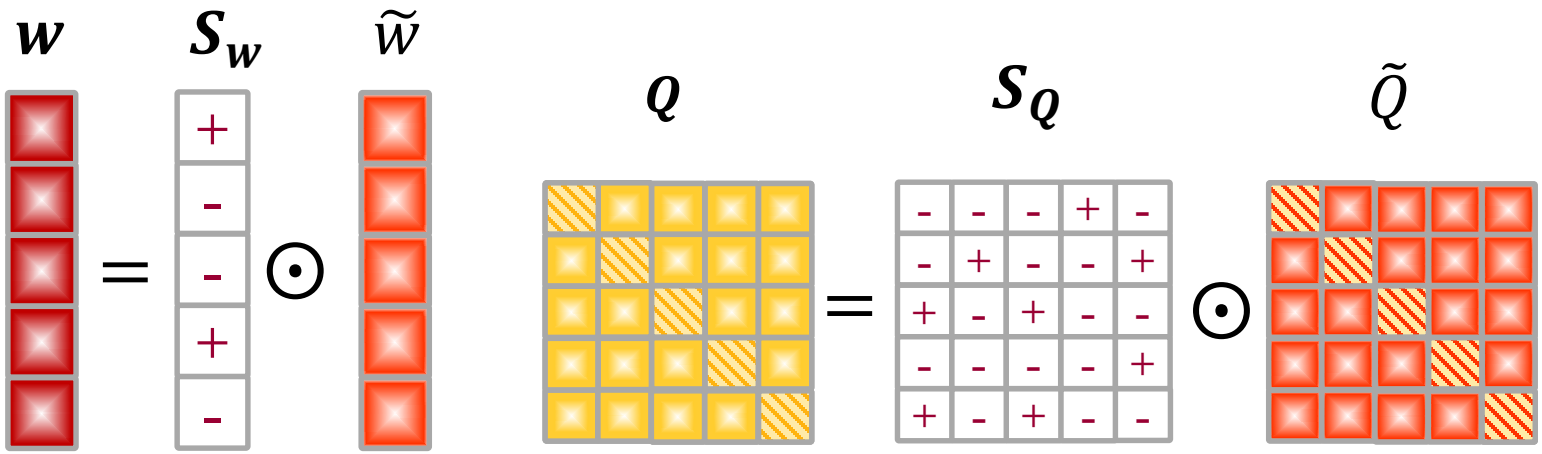
$$\arg\min_{w,Q} \frac{1}{2}\|w - v\|_2^2 + \frac{1}{2}\|Q - U\|_F^2 + \frac{\lambda}{t}\|w\|_1 + \frac{\lambda}{2t}\|Q\|_1$$

$$s.t.\ \left\|Q_{\cdot,j}\right\|_1 \leq |w_j|, \qquad j = 1,\dots,d$$

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



Proximal Operator Subproblem

- The proximal operator problem can be decoupled
- By factorization $\begin{cases} w_j = S_{w\,j}\widetilde{w}_j \\ Q_{\cdot,j} = S_{Q\,j}\tilde{Q}_{\cdot,j} \\ \widetilde{w}_j \geq 0 \\ \tilde{Q}_{\cdot,j} \geq 0 \end{cases}$, the sub-problem is:

$$\min_{S_{w\,j},S_{Q\,j},\widetilde{w}_j,\tilde{Q}_{\cdot,j}} \frac{1}{2}\left\|S_{w\,j}\widetilde{w}_j - v_j\right\|_2^2 + \frac{1}{2}\left\|S_{Q\,j}\tilde{Q}_{\cdot j} - U_{\cdot,j}\right\|_F^2 + \frac{\lambda}{t}\widetilde{w}_j + \frac{\lambda}{2t}1^T\tilde{Q}_{\cdot j}$$

$$s.t. \quad \begin{cases} 1^T\tilde{Q}_{\cdot j} \leq \widetilde{w}_j \\ \tilde{Q}_{\cdot,j} \geq 0 \end{cases}, \qquad j = 1, \dots, d$$

Solving Sign Variables

- We solve sign variables S_w, S_Q and magnitude variables $\widetilde{w}$, $\tilde{Q}$ together
- Since
$$\begin{aligned}\frac{1}{2}\left(w_j - v_j\right)^2 &= \frac{1}{2}\left(S_{w\,j}\widetilde{w}_j - v_j\right)^2 \\ &= \frac{1}{2}\left(S_{w\,j}(S_{w\,j}\widetilde{w}_j - v_j)\right)^2 \\ &= \frac{1}{2}\left(\widetilde{w}_j - S_{w\,j}v_j\right)^2,\end{aligned}$$
then $S_w = \mathrm{sign}(v)$
- $S_{Q\,j}$ must be the sign of $U_{\cdot,j}$, *i.e.*, $S_{Q\,j} = \mathrm{sign}(U_{\cdot,j})$

Solving Magnitude Variables

- The magnitude variables can thus be obtained by solving

$$\min_{\tilde{Q}_{\cdot,j},\tilde{w}_j} \frac{1}{2}\left\|\tilde{w}_j - \check{v}_j\right\|_2^2 + \frac{1}{2}\left\|\tilde{Q}_{\cdot j} - \check{U}_{\cdot,j}\right\|_2^2 + \frac{\lambda}{t}\tilde{w}_j + \frac{\lambda}{2t}1^T\tilde{Q}_{\cdot j}$$

$$s.t. \begin{cases} 1^T\tilde{Q}_{\cdot j} \leq \tilde{w}_j \\ \tilde{Q}_{\cdot,j} \geq 0 \end{cases},$$

where $\check{v}_j = S_{w\,j}v_j - \frac{\lambda}{t}\cdot\mathbf{1}$ and $\check{U}_{\cdot,j} = S_{Q\,j}U_{\cdot,j} - \frac{\lambda}{2t}\cdot\mathbf{1}$

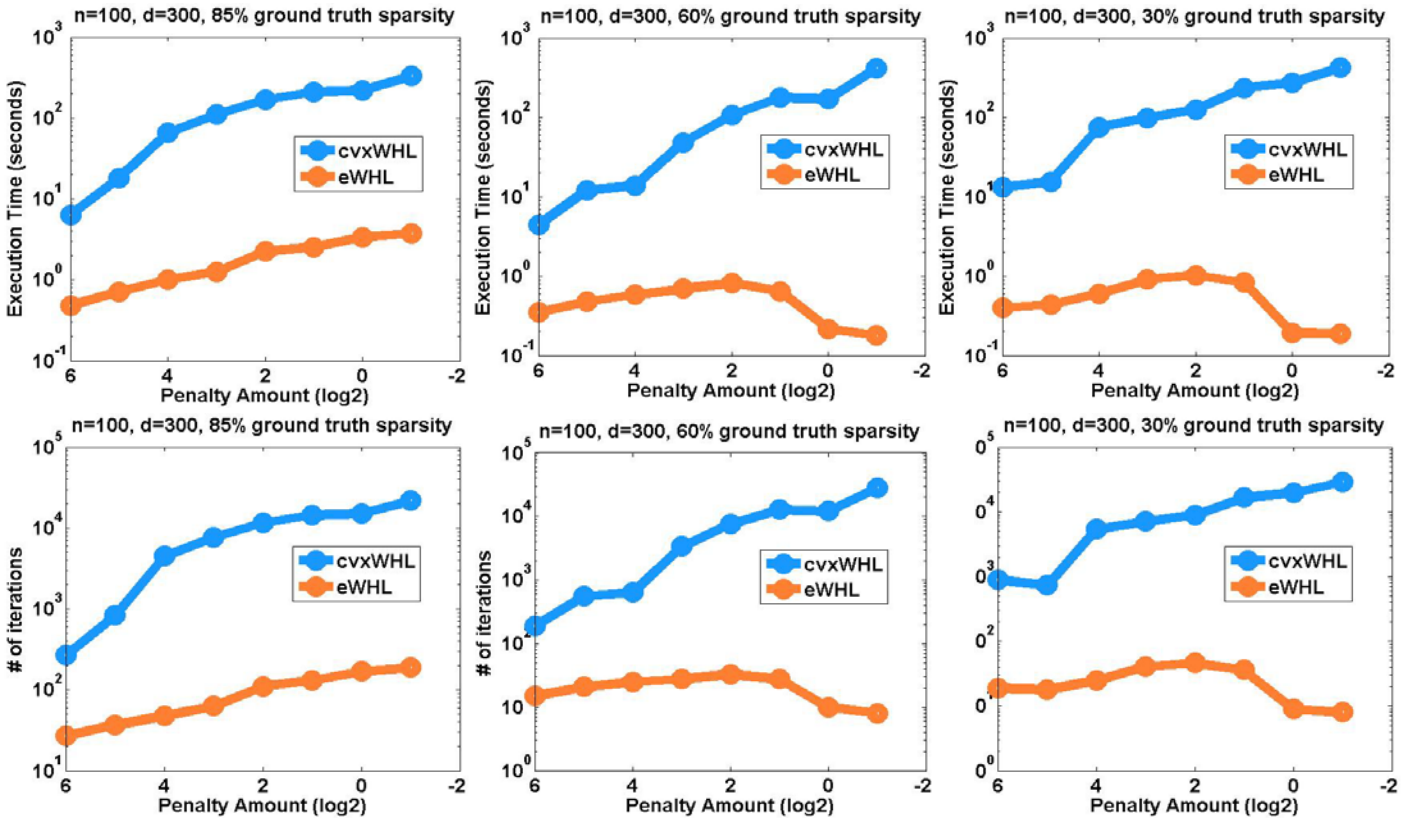
- The above subproblem has closed form solutions

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

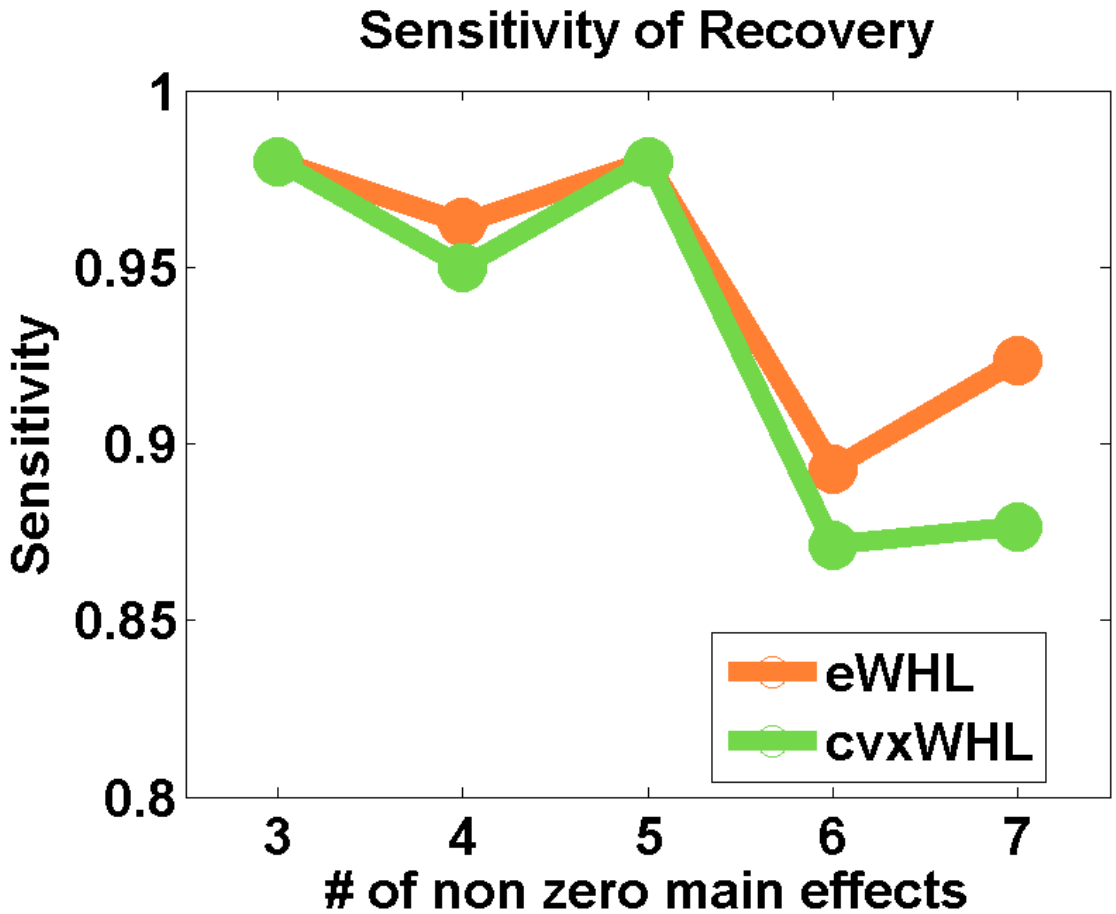
Comparisons of Efficiency

- 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

Main Effects Only					
	RF	SVM	spsLog	cvxWHL	eWHL
Accuracy (%)	74.23 ± 8.67	75.22 ± 8.72	74.34 ± 9.56	NA	NA
Sensitivity (%)	78.75 ± 14.00	80.18 ± 13.89	80.18 ± 13.88	NA	NA
Specificity (%)	69.29 ± 11.63	69.76 ± 12.80	69.52 ± 13.74	NA	NA
Main Effects + Interactions					
	RF	SVM	spsLog	cvxWHL	eWHL
Accuracy (%)	71.26 ± 10.22	59.45 ± 14.43	73.57 ± 10.30	75.22 ± 11.02	77.42 ± 8.50
Sensitivity (%)	83.04 ± 13.18	59.29 ± 17.83	74.29 ± 16.22	75.71 ± 19.11	77.14 ± 12.05
Specificity (%)	58.10 ± 23.23	60.00 ± 15.42	72.86 ± 12.46	74.52 ± 16.84	77.62 ± 15.02

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!