Efficient Methods for Overlapping Group Lasso

Presented by Xingguo Li

CSCI 8980, UMN

Authors : Lei Yuan, Jun Liu, and Jieping Ye

March 31, 2014

Outline

- The Overlapping Group Lasso
- The Proximal Operator and Efficient Computation
    - Key Properties of the Proximal Operator
    - Reformulation as a Smooth Convex Problem
    - Proximal Splitting Methods
- Extensions
    - ℓ_q Norm Overlapping Group Lasso
    - Capped Norm Overlapping Group Lasso
- Numerical Experiments
    - Synthetic Data
    - Gene Expression Data

Problem

Overlapping Group Lasso

$$\min_{\mathbf{x}\in\mathbb{R}^p} f(\mathbf{x}) = l(\mathbf{x}) + \phi_{\lambda_2}^{\lambda_1}(\mathbf{x}) \tag{1}$$

- $l(\mathbf{x})$ is a smooth convex loss function, *e.g.* $l(\mathbf{x}) = \sum_{i=1}^{n}(y_i - \mathbf{a}_i^T\mathbf{x})^2$
- $\phi_{\lambda_2}^{\lambda_1}(\mathbf{x}) = \lambda_1\|\mathbf{x}\|_1 + \lambda_2\sum_{i=1}^{g} w_i\|\mathbf{x}_{G_i}\|$ is the overlapping group Lasso penalty
  - $\lambda_1 = 0, \lambda_2 > 0$: group Lasso (Yuan *et al.*, 2006)
  - $\lambda_1 > 0, \lambda_2 = 0$: Lasso (Tibshirani, 1996)

Algorithm

"FoGLasso", *F*ast *o*verlapping *G*roup *Lasso*, based on accelerated gradient descent (AGD) (Beck *et al.*, 2009).

- Approximation (Linearization) of $f(\mathbf{x})$ as

$$f_{L,\mathbf{x}}(\mathbf{y}) = \left[l(\mathbf{x}) + \langle l'(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2} \|\mathbf{y} - \mathbf{x}\|^2 \right] + \phi_{\lambda_2}^{\lambda_1}(\mathbf{y}) \quad (2)$$

- A sequence of approximate solutions $\{\mathbf{x}_i\}$ by proximal operator,

$$\mathbf{x}_{i+1} = \arg\min_{\mathbf{y}} f_{L_i,\mathbf{s}_i}(\mathbf{y}) = \pi_{\lambda_2/L_i}^{\lambda_1/L_i}(\mathbf{s}_i - \frac{1}{L_i} l'(\mathbf{s}_i)), \quad (3)$$

where $\mathbf{s}_i = \mathbf{x}_i + \beta_i(\mathbf{x}_i - \mathbf{x}_{i-1})$ and L_i can be determined by line search.

Algorithm 1: "FoGLasso"

Input: $L_0 > 0, \mathbf{x}_0, k$

Output: $\mathbf{x}_{k+1}$

1: Initialize $\mathbf{x}_1 = \mathbf{x}_0, \alpha_{-1} = 0, \alpha = 1$, and $L = L_0$.

2: **for** $i = 1$ to k **do**

3: Set $\beta_i = \frac{\alpha_{i-2}-1}{\alpha_{i-1}}, \mathbf{s}_i = \mathbf{x}_i + \beta_i(\mathbf{x}_i - \mathbf{x}_{i-1})$

4: Find the smallest $L = 2^j L_{i-1}, j = 0, 1, \cdots$ such that $f(\mathbf{x}_{i+1}) \leq f_{L,\mathbf{s}_i}(\mathbf{x}_{i+1})$ holds, where $\mathbf{x}_{i+1} = \pi_{\lambda_2/L_i}^{\lambda_1/L_i}(\mathbf{s}_i - \frac{1}{L_i} l'(\mathbf{s}_i))$

5: Set $L_i = L$ and $\alpha_{i+1} = \frac{1+\sqrt{1+4\alpha_i^2}}{2}$

6: **end for**

Proximal Operator and Efficient Computation

The proximal operator:

$$\mathbf{x}_{i+1} = \pi_{\lambda_2/L_i}^{\lambda_1/L_i}(\mathbf{s}_i - \frac{1}{L_i} l'(\mathbf{s}_i)),$$

Definition (recall $\phi_{\lambda_2}^{\lambda_1}(\mathbf{x}) = \lambda_1 \|\mathbf{x}\|_1 + \lambda_2 \sum_{i=1}^{g} w_i \|\mathbf{x}_{G_i}\|$):

$$\pi_{\lambda_2}^{\lambda_1}(\mathbf{v}) = \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ g_{\lambda_2}^{\lambda_1}(\mathbf{x}) \equiv \frac{1}{2}\|\mathbf{x}-\mathbf{v}\|^2 + \phi_{\lambda_2}^{\lambda_1}(\mathbf{x}) \right\} \qquad (4)$$

- Many groups are zero (identify $\mathbf{x}_{G_i} = \mathbf{0}$)
- $g_{\lambda_2}^{\lambda_1}(\mathbf{x})$ is nonsmooth (smooth reformulation)
- More proximal operator solver (Dykstra-like, ADMM)

Key Properties of the Proximal Operator

Lemma 1

Suppose $\lambda_1, \lambda_2 \geq 0$ *and* $w_i > 0, i = 1, 2, \cdots, g$. *Let* $\mathbf{x}^* = \pi_{\lambda_2}^{\lambda_1}(\mathbf{v})$ *and* $\odot$ *be point-wise product, then the following holds:*

1. *if* $v_i > 0$, *then* $0 \leq x_i^* \leq v_i$;
2. *if* $v_i < 0$, *then* $v_i \leq x_i^* \leq 0$;
3. *if* $v_i = 0$, *then* $x_i^* = 0$;
4. $SGN(\mathbf{v}) \subseteq SGN(\mathbf{x}^*)$; *and*
5. $\pi_{\lambda_2}^{\lambda_1}(\mathbf{v}) = sgn(\mathbf{v}) \odot \pi_{\lambda_2}^{\lambda_1}(|\mathbf{v}|)$.

$$\mathrm{SGN}(t) = \begin{cases} \{1\}, & t > 0 \\ \{-1\}, & t < 0 \\ [-1, 1], & t = 0 \end{cases}, \quad \mathrm{sgn}(t) = \begin{cases} 1, & t > 0 \\ -1, & t < 0 \\ 0, & t = 0 \end{cases}$$

Key Properties of the Proximal Operator

Theorem 1

Let $\mathbf{u} = sgn(\mathbf{v}) \odot \max(|\mathbf{v}| - \lambda_1, 0)$, *and*

$$\pi^0_{\lambda_2}(\mathbf{u}) = arg \min_{\mathbf{x} \in \mathbb{R}^p} \left\{ h_{\lambda_2}(\mathbf{x}) \equiv \frac{1}{2}\|\mathbf{x} - \mathbf{u}\|^2 + \lambda_2 \sum_{i=1}^{g} w_i \|\mathbf{x}_{G_i}\| \right\}. \quad (5)$$

Then, the following holds: $\pi^{\lambda_1}_{\lambda_2}(\mathbf{v}) = \pi^0_{\lambda_2}(\mathbf{u})$.

- Nice! $\pi^{\lambda_1}_{\lambda_2}(\mathbf{v})$ reduces to (5).
- Difficulty: groups may overlap.
- Many groups are zero (sparse solution solution desired), how to identify?

Key Properties of the Proximal Operator

- Sufficient condition for a group to be zero:

Lemma 2

Let $\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathbb{R}^p} h_{\lambda_2}(\mathbf{x})$. *If the i-th group satisfies* $\|\mathbf{u}_{G_i}\| \leq \lambda_2 w_i$, *then* $\mathbf{x}^*_{G_i} = \mathbf{0}$, i.e. *the i-th group is zero.*

- Given $S_i = \bigcup_{j \neq i, \mathbf{x}^*_{G_j}=\mathbf{0}} (G_j \cap G_i)$, a much weaker condition (much more zero groups can be identified):

Lemma 3

Let $\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathbb{R}^p} h_{\lambda_2}(\mathbf{x})$. *If the i-th group satisfies* $\|\mathbf{u}_{G_i - S_i}\| \leq \lambda_2 w_i$, *then* $\mathbf{x}^*_{G_i} = \mathbf{0}$, i.e. *the i-th group is zero.*

- Iterative procedure to identify the zero groups.

Reformulation as a Smooth Convex Problem

Focus on reduced problem $\mathbf{u} \succ \mathbf{0}$. Rewrite $\pi^0_{\lambda_2}(\mathbf{u})$ as:

$$\pi^0_{\lambda_2}(\mathbf{u}) = \arg \min_{\substack{\mathbf{x}\in\mathbb{R}^p \\ \mathbf{0}\preceq\mathbf{x}\preceq\mathbf{u}}} \left\{ h_{\lambda_2}(\mathbf{x}) \equiv \frac{1}{2}\|\mathbf{x}-\mathbf{u}\|^2 + \lambda_2 \sum_{i=1}^{g} w_i \|\mathbf{x}_{G_i}\| \right\}.$$

Use dual norm of $\|\cdot\|$, rewrite $h_{\lambda_2}(\mathbf{x})$ as:

$$h_{\lambda_2}(\mathbf{x}) = \max_{Y\in\Omega} \frac{1}{2}\|\mathbf{x}-\mathbf{u}\|^2 + \sum_{i=1}^{g} \langle \mathbf{x}, Y^i \rangle, \qquad (6)$$

where $\Omega = \left\{ Y \in \mathbb{R}^{p\times g} : Y^i_{G_i^c} = \mathbf{0}, \|Y^i\| \leq \lambda_2 w_i, i = 1, \cdots, g \right\}$.
Reformulation as a min-max problem:

$$\pi^0_{\lambda_2}(\mathbf{u}) = \arg \min_{\substack{\mathbf{x}\in\mathbb{R}^p \\ \mathbf{0}\preceq\mathbf{x}\preceq\mathbf{u}}} \max_{Y\in\Omega} \left\{ \psi(\mathbf{x}, Y) \equiv \frac{1}{2}\|\mathbf{x}-\mathbf{u}\|^2 + \langle \mathbf{x}, Y\mathbf{e} \rangle \right\} \qquad (7)$$

Reformulation as a Smooth Convex Problem (continue....)

$\psi(\mathbf{x}, Y)$ is convex in $\mathbf{x}$, concave in Y. Methodology for min $h_{\lambda_2}(\cdot)$:

- w.r.t. Y, $\mathrm{argmin}_{Y \in \Omega}\{w(Y) = -\psi(\max(\mathbf{u} - Y\mathbf{e}, \mathbf{0}), Y)\}$
- w.r.t. $\mathbf{x}$, $\mathbf{x} = \max(\mathbf{u} - Y\mathbf{e}, \mathbf{0}) \Rightarrow$ construct solution to $h_{\lambda_2}(\cdot)$

Theorem 2

The function $w(Y)$ is convex and continuously differentiable with

$$w'(Y) = -\max(\mathbf{u} - Y\mathbf{e}, \mathbf{0})\mathbf{e}^T \tag{8}$$

In addition, $w'(Y)$ is Lipschitz continuous with constant g, i.e.,

$$\|w'(Y_1) - w'(Y_2)\|_F \leq g\|Y_1 - Y_2\|_F, \forall Y_1, Y_2 \in \mathbb{R}^{p \times g}. \tag{9}$$

Use accelerated gradient descent (AGD) method to solve $\psi(\mathbf{x}, Y)$.

Duality Gap

Theorem 3

Let gap$\widetilde{Y} = \max_{Y \in \Omega} \psi(\widetilde{\mathbf{x}}, Y) - \min_{\mathbf{x} \in \mathbb{R}^p, \mathbf{0} \preceq \mathbf{x} \preceq \mathbf{u}} \psi(\mathbf{x}, \widetilde{Y})$ *be the duality gap. Then, the following holds:*

$$\mathrm{gap}(\widetilde{Y}) = \sum_{i=1}^{g} (\lambda_2 w_i \|\widetilde{\mathbf{x}}_{G_i}\| - \langle \widetilde{\mathbf{x}}_{G_i}, \widetilde{Y}^i_{G_i} \rangle). \tag{10}$$

In addition, we have

$$w(\widetilde{Y}) - w(Y^*) \leq \mathrm{gap}(\widetilde{Y}), \tag{11}$$

$$h(\widetilde{\mathbf{x}}) - h(\mathbf{x}^*) \leq \mathrm{gap}(\widetilde{Y}). \tag{12}$$

Serve as the stopping criteria (*e.g.* $< 10^{-10}$).

Proximal Splitting Methods

- Dykstra-like Proximal Splitting Method (Combettes *et al.*, 2009)
- ADMM (Boyd *et al.*, 2011)

Dykstra-like Proximal Splitting Method: *convex feasibility problem*

$$\text{find } x \in \{\textstyle\bigcap_{i=1}^{m} C_i \mid C_i \text{ is a convex set}\}$$

- Iterative scheme by cycling through all convex sets
- Convergence guarantee under certain conditions

Consider $\pi_{\lambda_2}^{0}(\mathbf{u}) = \arg\min_{\mathbf{x}\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{x}-\mathbf{u}\|^2 + \lambda_2 \sum_{i=1}^{g} w_i\|\mathbf{x}_{G_i}\|$ as the projection of $\mathbf{u}$ onto a collection of convex sets $\{w_i\|\mathbf{x}_{G_i}\|\}$.

Algorithm 2: Dykstra-like Proximal Splitting Methods

Set $\mathbf{x}_0 = \mathbf{u}, \mathbf{q}_{1,0}, \cdots, \mathbf{q}_{g,0} = \mathbf{x}_0, n = 0$
repeat $n = n + 1$
  for $i = 1$ to g **do**
    $\mathbf{p}_{i,n} = \operatorname{prox}_{\lambda\|\mathbf{x}_{G_i}\|}\mathbf{q}_{i,n}$
  $\mathbf{x}_{n+1} = \sum_{i=1}^{g} w_i \mathbf{q}_{i,n}$
  for $i = 1$ to g **do**
    $\mathbf{q}_{i,n+1} = \mathbf{x}_{n+1} + \mathbf{q}_{i,n} - \mathbf{p}_{i,n}$
until Convergence

$$\mathbf{p} = \operatorname{prox}_{\lambda\|\mathbf{x}_{G_i}\|}\mathbf{q} = \underset{\mathbf{x}\in\mathbb{R}^p}{\operatorname{argmin}} \|\mathbf{x} - \mathbf{q}\|^2/2 + \lambda\|\mathbf{x}_{G_i}\|$$

$$\Rightarrow \mathbf{p}_{G_i} = \frac{\max(\|\mathbf{q}_{G_i}\| - \lambda, 0)}{\|\mathbf{q}_{G_i}\|}\mathbf{q}_{G_i} \text{ (closed form)}$$

ADMM

- Reformulation with auxiliary variables:

$$\begin{array}{c} \min_{\mathbf{x},\mathbf{z}} \ \frac{1}{2}\|\mathbf{x}-\mathbf{u}\|^2 + \lambda \sum_{i=1}^{g} w_i \|z_i\| \\ \text{s.t. } \mathbf{z}_i = \mathbf{x}_{G_i}, \ i = 1, \cdots, g. \end{array}$$

- Augmented Lagrangian:

$$\begin{aligned} L_\rho(\mathbf{x},\mathbf{z},\mathbf{y}) = & \ \frac{1}{2}\|\mathbf{x}-\mathbf{u}\|^2 + \lambda \sum_{i=1}^{g} w_i \|z_i\| \\ & + \sum_{i=1}^{g} \mathbf{y}_i^T (\mathbf{z}_i - \mathbf{x}_{G_i}) + \frac{\rho}{2} \sum_{i=1}^{g} \|\mathbf{z}_i - \mathbf{x}_{G_i}\|^2 \end{aligned}$$

- ADMM iterations:

$$\begin{aligned} \mathbf{x}^{k+1} &:= \mathrm{argmin}_{\mathbf{x}} \, L_\rho(\mathbf{x},\mathbf{z}^k,\mathbf{y}^k) \\ \mathbf{z}^{k+1} &:= \mathrm{argmin}_{\mathbf{x}} \, L_\rho(\mathbf{x},\mathbf{z}^k,\mathbf{y}^k) \\ \mathbf{y}^{k+1} &:= \mathrm{argmin}_{\mathbf{x}} \, L_\rho(\mathbf{x},\mathbf{z}^k,\mathbf{y}^k) \end{aligned}$$

ADMM

- For $\mathbf{x}$, $\frac{\partial}{\partial \mathbf{x}} L_\rho(\mathbf{x}, \mathbf{z}^k, \mathbf{y}^k) = \mathbf{x} - \mathbf{u} - \sum_{i=1}^g \widetilde{\mathbf{y}}_i^k$
$$+\rho \sum_{i=1}^g \widetilde{\mathbf{e}}_i \odot \mathbf{x} - \rho \sum_{i=1}^g \widetilde{\mathbf{z}}_i^k$$
$$\Rightarrow \mathbf{x}^{k+1} = \left(\mathbf{u} + \sum_{i=1}^g \widetilde{\mathbf{y}}_i^k + \rho \sum_{i=1}^g \widetilde{\mathbf{z}}_i^k\right) \oslash \left(\mathbf{e} + \rho \sum_{i=1}^g \widetilde{\mathbf{e}}_i\right)$$
- For $\mathbf{z}$, use subdifferential,
$$0 \in \mathbf{z}_i^{k+1} - \mathbf{x}_{G_i}^{k+1} + \frac{1}{\rho}\mathbf{y}_i^k + \frac{\lambda w_i}{\rho} \partial \|\mathbf{z}_i^{k+1}\|,$$
where $\partial \|\mathbf{z}_i^{k+1}\| = \begin{cases} \frac{\mathbf{z}_i^{k+1}}{\|\mathbf{z}_i^{k+1}\|} & \|\mathbf{z}_i^{k+1}\| \neq 0 \\ \{\mathbf{t} | \mathbf{t} \in \mathbb{R}^{|G_i|}, \|\mathbf{t}\| < 1\} & \|\mathbf{z}_i^{k+1}\| = 0 \end{cases}$
$$\Rightarrow \mathbf{z}_i^{k+1} = \frac{\max\{\|\widetilde{\mathbf{x}}_{G_i}^{k+1}\| - \widetilde{\lambda}_i, 0\}}{\|\widetilde{\mathbf{x}}_{G_i}^{k+1}\|} \widetilde{\mathbf{x}}_{G_i}^{k+1},$$
where $\widetilde{\mathbf{x}}_{G_i}^{k+1} = \mathbf{x}_{G_i}^{k+1} - \frac{1}{\rho}\mathbf{y}_i^k, \widetilde{\lambda}_i = \frac{\lambda w_i}{\rho}$

ℓ_q Norm Overlapping Group Lasso

- Generalize $\psi_{\lambda_2}^{\lambda_1}(\mathbf{x})$ and $\pi_{\lambda_2}^{\lambda_1}(\mathbf{v})$ to

$$\psi_{q,\lambda_2}^{\lambda_1}(\mathbf{x}) = \lambda_1 \|\mathbf{x}\|_1 + \lambda_2 \sum_{i=1}^{g} w_i \|\mathbf{x}_{G_i}\|_q \tag{13}$$

$$\pi_{q,\lambda_2}^{\lambda_1}(\mathbf{v}) = \arg\min_{\mathbf{x} \in \mathbb{R}^p} \left\{ g_{q,\lambda_2}^{\lambda_1}(\mathbf{x}) \equiv \frac{1}{2}\|\mathbf{x} - \mathbf{v}\|^2 + \phi_{q,\lambda_2}^{\lambda_1}(\mathbf{x}) \right\} \tag{14}$$

- Same properties hold for ℓ_q proximal operator: $1/q + 1/\bar{q} = 1$,
  Necessary condition: If $\|\mathbf{u}_{G_i}\|_{\bar{q}} \leq \lambda_2 w_i$, then $\mathbf{x}^*_{G_i} = \mathbf{0}$.
  A weaker condition: If $\|\mathbf{u}_{G_i - S_i}\|_{\bar{q}} \leq \lambda_2 w_i$, then $\mathbf{x}^*_{G_i} = \mathbf{0}$.

ℓ_q Norm Overlapping Group Lasso

- Same result holds for the duality gap for smooth reformulation:

  $$\mathrm{gap}(\widetilde{Y}) = \sum_{i=1}^{g}(\lambda_2 w_i \|\widetilde{\mathbf{x}}_{G_i}\|_q - \langle \widetilde{\mathbf{x}}_{G_i}, \widetilde{Y}^i_{G_i} \rangle).$$

  Feasible region of the dual variable Y:

  $$\Omega = \left\{ Y \in \mathbb{R}^{p \times g} : Y^i_{G_i^c} = \mathbf{0}, \|Y^i\|_{\bar{q}} \leq \lambda_2 w_i, i = 1, \cdots, g \right\}$$

  Efficient bisection root-finding based ℓ_q-norm projection (Liu *et al.*, 2010)

Capped Norm Overlapping Group Lasso

- Consider the problem:

$$\min_{\mathbf{x}\in\mathbb{R}^p} l(\mathbf{x}) + \lambda_1\|\mathbf{x}\|_0 + \lambda_2 \sum_{i=1}^{g} w_i I(\|\mathbf{x}_{G_i}\| \neq 0) \tag{15}$$

- ℓ_1-norm regularization introduces bias.
- Nonconvex capped norms: closer to ℓ_0-norm than ℓ_1-norm (Zhang 2011, Shen *et al.* 2012): for some small $\theta_1, \theta_2 > 0$,

$$\|\mathbf{x}\|_0 \approx \sum_{j=1}^{p} \min\left(1, \frac{|x_j|}{\theta_1}\right)$$

$$\sum_{i=1}^{g} w_i I(\|\mathbf{x}_{G_i}\| \neq 0) \approx \sum_{i=1}^{g} w_i \min\left(1, \frac{|x_{G_i}|}{\theta_2}\right)$$

Capped Norm Overlapping Group Lasso

- Decompose $\sum_{j=1}^{p} \min\left(1, \frac{|x_j|}{\theta_1}\right)$ and $\sum_{i=1}^{g} w_i \min\left(1, \frac{|x_{G_i}|}{\theta_2}\right)$, approximate the problem 15 as:

$$\min_{\mathbf{x}\in\mathbb{R}^p} l(\mathbf{x}) + \frac{\lambda_1}{\theta_1}\|\mathbf{x}\|_1 + \frac{\lambda_2}{\theta_2}\sum_{i=1}^{p}\|\mathbf{x}_{G_i}\| - P(\mathbf{x}) - D(\mathbf{x}) \tag{16}$$

$$P(\mathbf{x}) = \frac{\lambda_1}{\theta_1}\sum_{i=1}^{p}\max(|\mathbf{x}_j| - \theta_1, 0) \text{ convex in } \mathbf{x}$$

$$D(\mathbf{x}) = \frac{\lambda_2}{\theta_2}\sum_{i=1}^{p} w_i \max(\|\mathbf{x}_{G_i}\| - \theta_2, 0) \text{ convex in } \mathbf{x}$$

- "Difference of two convex functions" (DC) programming

Algorithm 3: DC Programming for Overlapping Group Lasso with the Capped Norm

$$\frac{\partial}{\partial \mathbf{x}_j} P(\mathbf{x}) \ni \begin{cases} \frac{\lambda_1}{\theta_1}\mathrm{sgn}(\mathbf{x}_j) & |\mathbf{x}_j| > \theta_1 \\ 0 & |\mathbf{x}_j| \leq \theta_1 \end{cases} \quad \frac{\partial}{\partial \mathbf{x}_{G_i}} D(\mathbf{x}_{G_i}) \ni \begin{cases} \frac{\mathbf{x}_{G_i}}{\|\mathbf{x}_{G_i}\|} & \|\mathbf{x}_{G_i}\| > \theta_2 \\ \mathbf{0} & \|\mathbf{x}_{G_i}\| \leq \theta_2 \end{cases}$$

Input: $\theta_0, \theta_1 > 0, \mathbf{x}_0, k$

Output: $\mathbf{x}_{k+1}$

1: Initialize $\mathbf{x}_1 = \mathbf{x}_0$

2: **for** $i = 1$ to k **do**

3: Choose $U^i \in \partial P(\mathbf{x}^i)$ and $V^i \in \partial D(\mathbf{x}^i)$

4: Solve $\mathbf{x}^{i+1} = \mathrm{argmin}_{\mathbf{x}\in\mathbb{R}^p}\, l(\mathbf{x}) + \frac{\lambda_1}{\theta_1}\|\mathbf{x}\|_1 + \frac{\lambda_2}{\theta_2}\sum_{i=1}^p \|\mathbf{x}_{G_i}\| - \langle U^k + V^k, \mathbf{x}\rangle$ (via "FoGLasso")

5: **end for**

Experiments: Efficiency of Calculating the Proximal Operator

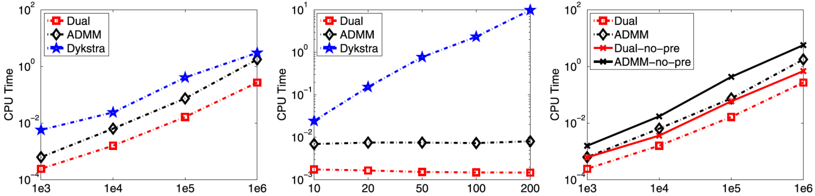

Figure 1 : Time comparison for computing the proximal operators. The group number g is fixed in the left figure and the problem size p is fixed in the middle figure. The right figure illustrates the effectiveness of the preprocessing.

Sparse Pattern Recovery

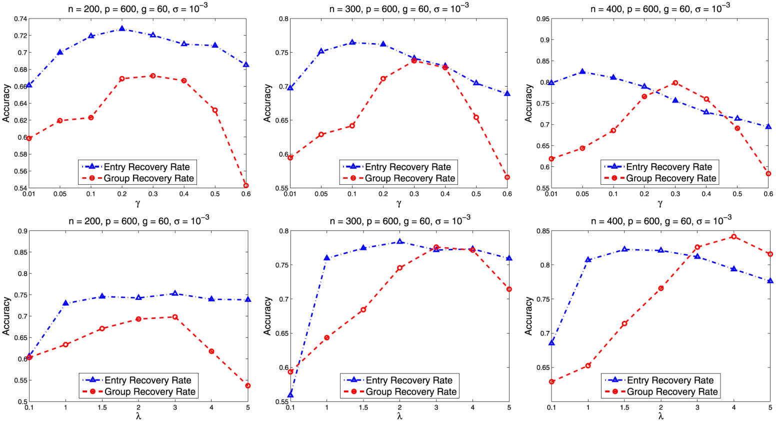

Figure 2 : Results of the convex overlapping group Lasso formulation (top row) and the nonconvex overlapping group Lasso with the capped norm (bottom row).

Sparse Pattern Recovery

TABLE 1
Cross-Validation Performance of Sparse Pattern Recovery of the Convex Overlapping Group Lasso Formulation and the Nonconvex Overlapping Group Lasso Formulation Based on the Capped Norm on Synthetic Data with Different Problem Sizes

	Convex		Non-convex	
n	Entry Rate	Group Rate	Entry Rate	Group Rate
300	0.71	0.60	0.77	0.71
400	0.80	0.61	0.82	0.70

- Nonconvex formulation outperforms convex formulation.

Comparison with SLasso, Prox-Grad, and ADMM

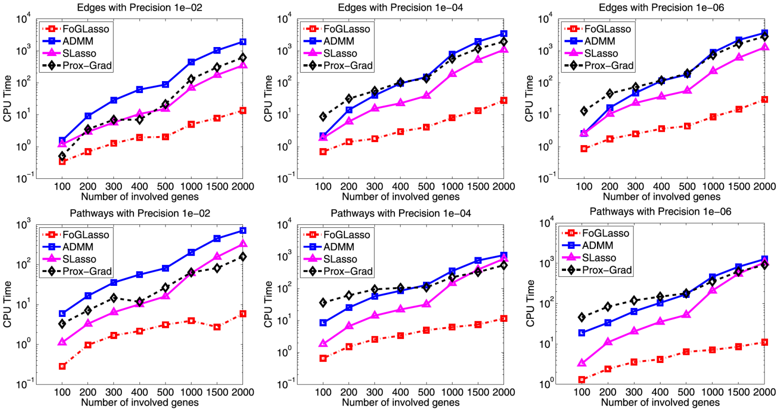

Figure 3 : Comparison of SLasso (Jenatton *et al.* 2009), ADMM (Boyd *et al.* 2010), Prox-Grad (Chen *et al.* 2012), and "FoGLasso" in terms of computational time (in seconds and in the log scale).

Comparison with Picard-Nesterov

TABLE 3
Comparison of FoGLasso, Picard-Nesterov, and Picard-Nesterov with Our Proposed Preprocessing Technique Using Different Numbers (p) of Genes and Various Precision Levels

Precision Level	10^{-2}			10^{-4}			10^{-6}		
p	100	200	400	100	200	400	100	200	400
FoGLasso	81	189	353	192	371	1299	334	507	1796
	288	401	921	404	590	1912	547	727	2387
Picard-Nesterov	78	176	325	181	304	1028	318	504	1431
	8271	6.8e4	2.2e5	2.6e4	1.0e5	7.8e5	5.1e4	1.3e5	1.1e6
Picard-Nesterov-PreProc	78	176	325	181	304	1028	318	504	1431
	2683	3.8e4	1.1e5	8427	6.4e4	4.9e5	1.9e4	8.2e4	7.3e5

- For each particular method, the first row denotes the number of outer iterations required for convergence, while the second row represents the total number of inner iterations.
- Same complexity of $\mathcal{O}(pg)$ for inner iteration.

Computation of the Proximal Operator

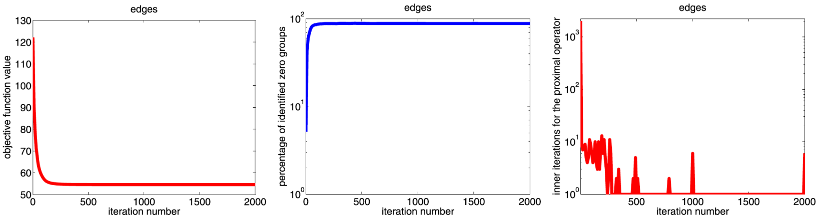

Figure 4 : Performance of the computation of the proximal operator in FoGLasso. The left plot shows the objective function value during the FoGLasso iteration. The middle plot shows the percentage of the identified zero groups. The right plot shows the number of inner iterations for achieving the duality gap less than 10^{-10} when one solves the proximal operator via the dual reformulation.

- Most zero groups are identified after $\sim$ 100 steps.

Convergence with Inexact Proximal Operator

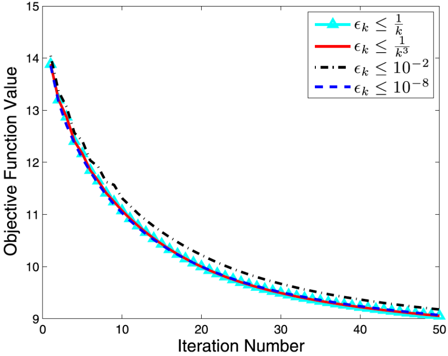

Figure 5 : Illustration of the objective function values of the first 50 iterations with different stopping criteria used for computing the proximal operator.

- No dramatic change w.r.t different termination conditions.

Thank you!