Optimization of the Sparse Group Lasso Problem using Block-wise Coordinate Descent

Jingyiran Li, Xiaoxuan Liang

1 Introduction

When performing variable selection, sometimes it is desirable to attain sparsity both in a group level and in coefficients within each group. The sparse group lasso problem is an extension of the group lasso problem in the sense that it combines features from the regular lasso and that from the group lasso to achieve within-group sparsity [2]. More specifically, its penalty term is a weighted sum of the l_1 and l_2 penalty. Let y be a vector of observations with dimension $n \times 1$. X is a $n \times p$ feature matrix where p is divided into L groups with size g_l for each group l. Denote X_l as the features corresponding to group l and β_l as coefficients for group l. The sparse group lasso problem can be formulated as follows [1]:

$$\min_{\beta \in \mathbb{R}^p} \left(\frac{1}{2n} \|y - \sum_{k=1}^{L} X^{(k)} \beta^{(k)}\|_2^2 + (1-\alpha)\lambda \sum_{k=1}^{L} \|\beta^{(k)}\|_2 + \alpha\lambda \|\beta\|_1 \right)$$

where $\alpha \in [0, 1]$. Note that when $\alpha = 0$, the above program reduces to the group lasso problem and when $\alpha = 1$, it becomes the regular lasso problem. The sparse group lasso problem is convex since it is a linear combination of convex functions. Its optimum can be solved using subdifferential techniques.

2 Methodology

2.1 Setup

Since the sparse group lasso problem is convex, for each group k, if all coefficients in other groups are fixed, and the penalties corresponding to those coefficients can be ignored, the optimal solution should satisfy [3]

$$\frac{1}{2n} \|r_{(-k)} - X^{(k)} \beta^{(k)}\|_2^2 + (1-\alpha)\lambda \|\beta^{(k)}\|_2 + \alpha\lambda \|\beta^{(k)}\|_1$$

where $r_{(-k)} = y - \sum_{l \neq k} X^{(k)} \beta^{(k)}$.
Let u, v be the subgradients of $\beta_2^{(k)}$ and $\beta_1^{(k)}$,

$$u = \begin{cases} \frac{\beta^{(k)}}{\|\beta^{(k)}\|_2} & \beta^{(k)} \neq \mathbf{0} \\ \in \{u : \|u\|_2 \leq 1\} & \beta^{(k)} = \mathbf{0} \end{cases} \qquad v_j = \begin{cases} 1 & \beta_j^{(k)} > 0 \\ -1 & \beta_j^{(k)} < 0 \\ \in \{v_j : |v_j| \leq 1\} & \beta_j^{(k)} = 0 \end{cases}$$

Denote $l(r_{(-k)}, \beta) = \frac{1}{2n} \|r_{-k)} - X^{(k)} \beta\|_2^2$, then the update is (given $\beta_0^{(k)}$ from previous iteration)

$$\begin{aligned} \beta^{(k)} &= \arg\min_z \; l(r_{(-k)}, \beta_0^{(k)}) + \nabla l(r_{(-k)}, \beta_0^{(k)})^\top (z - \beta_0^{(k)}) + \frac{1}{2\gamma} \|z - \beta_0^{(k)}\|_2^2 + (1-\alpha)\lambda \|z\|_2 + \alpha\lambda \|z\|_1 \\ &= \arg\min_z \frac{1}{2\gamma} \|z - (\beta_0^{(k)} - \gamma \nabla l(r_{(-k)}, \beta_0^{(k)})\|_2^2 + (1-\alpha)\lambda \|z\|_2 + \alpha\lambda \|z\|_1 \end{aligned}$$

The coefficient $\beta^{(k)} = \mathbf{0}$ if and only if $\|S(\beta_0^{(k)} - \gamma \nabla l(r_{(-k)}, \beta_0^{(k)}), \gamma\alpha\lambda)\|_2 \leq \gamma(1-\alpha)\lambda$ with

$$S(\beta_0^{(k)} - \gamma \nabla l(r_{(-k)}, \beta_0^{(k)}), \gamma\alpha\lambda) = \text{sign}(\beta_0^{(k)} - \gamma \nabla l(r_{(-k)}))(|\beta_0^{(k)} - \gamma \nabla l(r_{(-k)}, \beta_0^{(k)})| - \gamma\alpha\lambda)_+$$

where $\beta_0^{(k)}$ is the initial value, and $S(\cdot)$ is the soft-thresholding operatorator applied to each component. Using singular value decomposition, $X^{(k)}$ is orthonormal ($X^{(k)\top} X^{(k)} = I$). The $\beta^{(k)}$ could be found using

$$\beta^{(k)} = \left(1 - \frac{\gamma(1-\alpha)\lambda}{\|S(\beta_0^{(k)} - \gamma \nabla l(r_{(-k)}, \beta_0^{(k)}), \gamma\alpha\lambda)\|_2}\right)_+ S(\beta_0^{(k)} - \gamma \nabla l(r_{(-k)}, \beta_0^{(k)}), \gamma\alpha\lambda) \tag{1}$$

We will iterate (1) by assigning $\beta^{(k)}$ to β_0 and updating $\beta^{(k)}$ by [3] until $\beta^{(k)}$ converges. For convenience, denote the update formula by $U(\beta_0, \gamma)$:

$$U(\beta_0, \gamma) = \left(1 - \frac{\gamma(1-\alpha)\lambda}{\|S(\beta_0^{(k)} - \gamma \nabla l(r_{(-k)}, \beta_0^{(k)}), \gamma\alpha\lambda)\|_2}\right)_+ S(\beta_0^{(k)} - \gamma \nabla l(r_{(-k)}, \beta_0^{(k)}), \gamma\alpha\lambda) \tag{2}$$

until $\beta^{(k)}$ converges in each block. Here is the complete algorithm:

Algorithm 1: Blockwise Descent Algorithm

Let p be the number of covariates;

Let g be the number of covariates in each group;

Let q be the number of groups, which is $\frac{p}{g}$;

$\beta \leftarrow$ rnorm(p);

while *Not converge* **do**

  for k *in 1,2,*$\cdots, q$ **do**

    if $\|S(X^{(k)^\top} r_{(-k)}, \alpha\lambda\|_2 \leq (1-\alpha)\lambda$ **then**

      $\beta^{(k)} \leftarrow \mathbf{0}$;

    else

      Update $\beta^{(k)}$ *using formula (2) until* $\beta^{(k)}$ *converges;*

    end

  end

end

return *updated* β

2.2 Simulation

X is a $n \times p$ matrix consisted of n examples and p predictors where $n = 200$, $p = 100$. The predictors are further divided into 10 groups and group sizes are uniformly 10. The within-group correlation for the predictors is set to be 0.2. Additionally, the last 4 groups have corresponding coefficient of 0. However, the first 6 groups have n non-zero coefficient(s) with $n = 10, 8, 6, 4, 2, 1$ for groups $1, \cdots, 6$, respectively. The non-zero coefficients are either -1 or 1 based on a random sampling with replacement. Furthermore, the singular value decomposition was performed on X to ensure $X^{(k)^\top} X^{(k)} = I$. Lastly, a $n \times 1$ observation vector y is generated by $y = X\beta + \epsilon$ where $\epsilon \sim N(200, 16)$.

3 Results

The values of the objective function generated from block-wise coordinate descent (BCD) for each problem are shown in figure 1. It could be discerned that all three problems behaved identically around the 97^{th} iteration and the sparse group lasso (SgLasso) demonstrated a reasonable compromise between the lasso and the group lasso (GLasso) starting from the 98^{th} iteration by yielding sparseness on both the group and individual predictor levels. Overall, the BCD has the fastest convergence rate for the group lasso and the slowest for the sparse group lasso. This could be attributed to the fact that the group lasso only performs shrinkage in between groups; it does not yield sparsity within the group. On the other hand, the sparse group lasso not only checks between-group sparsity but also that in within-group.

Furthermore, the accuracy of BCD for each problem based on the simulated data could be examined using figure 2. The blue dots are the predicted coefficients and the hollow circles are the actual coefficients. It is clearly that BCD

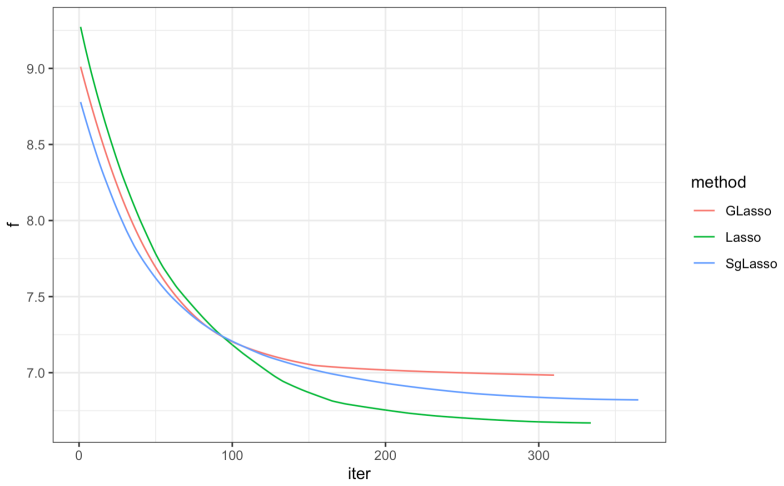

Figure 1: Objective function value for each problem based on BCD

can more accurately obtained the coefficients when they are equal to zero for all three problems. Overall, BCD on lasso has the lowest classification rate and the highest on the sparse group lasso.

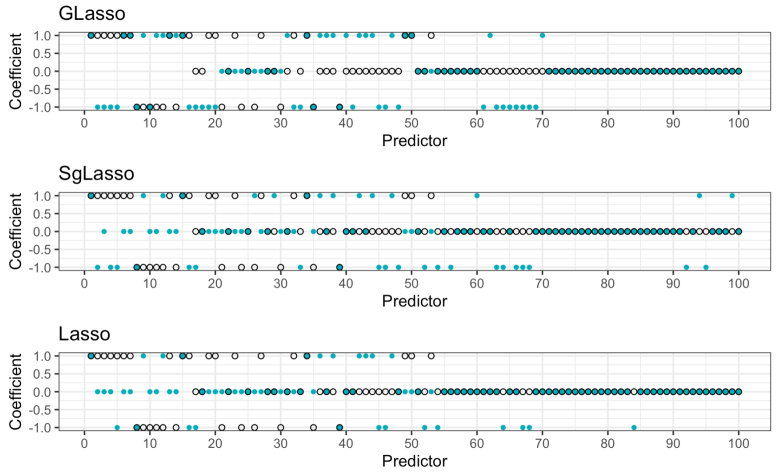

Figure 2: Predicted vs. Actual Coefficient

4 Conclusion and Future Directions

The sparse group lasso effectively combines the benefits from both the lasso and the group lasso problems. In particular, not only it guarantees between-group sparsity but also within-group sparsity. Sometimes, the group lasso fails to shrink the unimportant coefficients that are in the same group as the significant ones since it conducts shrinkage on a group level. Fortunately, the sparse lasso can resolve this issue by checking the coefficients in the non-zero groups and can shrink coefficients within that particular group if there exist insignificant coefficients. In the future, the BCD on sparse group lasso could be optimized by including momentum to increase the convergence rate to $O(1/\sqrt{\epsilon})$.

5 Math Proof

The objective sparse group lasso function is

$$\min_{\beta\in\mathbb{R}^p}\left(\frac{1}{2n}\|y-\sum_{k=1}^{L}X^{(k)}\beta^{(k)}\|_2^2+(1-\alpha)\lambda\sum_{k=1}^{L}\|\beta^{(k)}\|_2+\alpha\lambda\|\beta\|_1\right)$$

where $\lambda>0$, and $0\leq\alpha\leq 0$, and the target is to find $\beta^{(k)}$ to minimize it. Let $r_{(-k)}=y-\sum_{l\neq k}X^{(k)}\beta^{(k)}$, and we choose a group k to minimize over, keeping all coefficients in other groups are fixed and ignoring the corresponding penalties to those coefficients, the objective function can be written as

$$\frac{1}{2n}\|r_{(-k)}-X^{(k)}\beta^{(k)}\|_2^2+(1-\alpha)\lambda\|\beta^{(k)}\|_2+\alpha\lambda\|\beta^{(k)}\|_1$$

Denote $l(r_{(-k)},\beta)=\frac{1}{2n}\|r_{-k)}-X^{(k)}\beta\|_2^2$, then the update is (given $\beta_0^{(k)}$ from previous iteration)

$$\begin{aligned}\beta^{(k)}&=\arg\min_z\ l(r_{(-k)},\beta_0^{(k)})+\nabla l(r_{(-k)},\beta_0^{(k)})^\top(z-\beta_0^{(k)})+\frac{1}{2\gamma}\|z-\beta_0^{(k)}\|_2^2+(1-\alpha)\lambda\|z\|_2+\alpha\lambda\|z\|_1\\&=\arg\min_z\frac{1}{2\gamma}\|z-(\beta_0^{(k)}-\gamma\nabla l(r_{(-k)},\beta_0^{(k)})\|_2^2+(1-\alpha)\lambda\|z\|_2+\alpha\lambda\|z\|_1\end{aligned}$$

where

$$\nabla l(r_{(-k)},\beta_0)=-\frac{1}{n}{X^{(k)}}^\top r_{(-k)}.$$

Let u,v be subgradients of $\|z\|_2$ and $\|z\|_1$, and using conjugate trick

$$u=\begin{cases}\frac{z}{\|z\|_2} & z\neq\mathbf{0}\\ \in\{u:\|u\|_2\leq 1\} & z=\mathbf{0}\end{cases}\qquad v_j=\begin{cases}1 & z_j>0\\ -1 & z_j<0\\ \in\{v_j:|v_j|\leq 1\} & z_j=0\end{cases}$$

In order to minimize $h(z)=\frac{1}{2\gamma}\|z-(\beta_0^{(k)}-\gamma\nabla l(r_{(-k)},\beta_0^{(k)})\|_2^2+(1-\alpha)\lambda\|z\|_2+\alpha\lambda\|z\|_1$ over z, since $h(z)$ is convex, the optimal solution is obtained by the subgradient equationsand set it to zero:

$$\begin{aligned}\partial h(z)&=\frac{1}{\gamma}\left(z-(\beta_0^{(k)}-\gamma\nabla l(r_{(-k)},\beta_0^{(k)}))\right)+(1-\alpha)\lambda\,\partial\|z\|_2+\alpha\lambda\,\partial\|z\|_1\\&=\frac{1}{\gamma}\left(z-(\beta_0^{(k)}-\gamma\nabla l(r_{(-k)},\beta_0^{(k)}))+\gamma(1-\alpha)\lambda\,\partial\|z\|_2+\gamma\alpha\lambda\,\partial\|z\|_1\right)\\&=0\end{aligned}$$

If $z=\mathbf{0}$, then $\|\ \partial\|z\|_2\ \|_2\leq 1,-1\leq\partial\|z\|_1\leq 1$, and take the norm on the both sides,

$$\begin{aligned}\Rightarrow\|(\beta_0^{(k)}-\gamma\nabla l(r_{(-k)},\beta_0^{(k)}))-\gamma\alpha\lambda\,\partial\|z\|_1\|_2=\|\gamma(1-\alpha)\lambda\|_2\,\|\partial\|z\|_2\|_2\leq\gamma(1-\alpha)\lambda\\\Rightarrow\|(\beta_0^{(k)}-\gamma\nabla l(r_{(-k)},\beta_0^{(k)}))-\gamma\alpha\lambda\,\partial\|z\|_1\|_2\leq\gamma(1-\alpha)\lambda\end{aligned}$$

- When $\gamma\alpha\lambda \leq |(\beta_0^{(k)} - \gamma\nabla l(r_{(-k)}, \beta_0^{(k)}))|$,

$$\begin{aligned}\gamma(1-\alpha)\lambda &\geq \|(\beta_0^{(k)} - \gamma\nabla l(r_{(-k)}, \beta_0^{(k)})) - \gamma\alpha\lambda\|_2 = \|\text{sign}\left(\beta_0^{(k)} - \gamma\nabla l(r_{(-k)}, \beta_0^{(k)})\right)\left(|(\beta_0^{(k)} - \gamma\nabla l(r_{(-k)}, \beta_0^{(k)}))| - \gamma\alpha\lambda\right)_+\|_2 \\ &= \|S(\beta_0^{(k)} - \gamma\nabla l(r_{(-k)}, \beta_0^{(k)}), \gamma\alpha\lambda)\|_2\end{aligned}$$

- When $\gamma\alpha\lambda \geq |(\beta_0^{(k)} - \gamma\nabla l(r_{(-k)}, \beta_0^{(k)}))|$,

$$\begin{aligned}\gamma(1-\alpha)\lambda &\geq \|(\beta_0^{(k)} - \gamma\nabla l(r_{(-k)}, \beta_0^{(k)})) - \gamma\alpha\lambda\|_2 \geq 0 \\ &= \|\text{sign}\left(\beta_0^{(k)} - \gamma\nabla l(r_{(-k)}, \beta_0^{(k)})\right)\left(|(\beta_0^{(k)} - \gamma\nabla l(r_{(-k)}, \beta_0^{(k)}))| - \gamma\alpha\lambda\right)_+\|_2 \\ &= \|S(\beta_0^{(k)} - \gamma\nabla l(r_{(-k)}, \beta_0^{(k)}), \gamma\alpha\lambda)\|_2\end{aligned}$$

Therefore, $\beta^{(k)} = z = \mathbf{0}$ if $\|S(\beta_0^{(k)} - \gamma\nabla l(r_{(-k)}, \beta_0^{(k)}), \gamma\alpha\lambda)\|_2 \leq \gamma(1-\alpha)\lambda$.
If $\beta^{(k)} \neq \mathbf{0}$, assuming that we have orthogonalized each group of $X^{(k)}$ such that ${X^{(k)}}^\top X^{(k)} = I$, then we can simplify $\partial h(z) = 0$ as:

$$\begin{aligned}(\beta_0^{(k)} - \gamma\nabla l(r_{(-k)}, \beta_0^{(k)}) - z &= \gamma(1-\alpha)\lambda\frac{z}{\|z\|_2} + \gamma\alpha\lambda\,\text{sign}(z) \\ \Rightarrow z &= \left(1 + \frac{\gamma(1-\alpha)\lambda}{\|z\|_2}\right)^{-1}((\beta_0^{(k)} - \gamma\nabla l(r_{(-k)}, \beta_0^{(k)})) - \gamma\alpha\lambda\,\text{sign}(z)) \\ &= \left(\frac{\|z\|_2}{\|z\|_2 + \gamma(1-\alpha)\lambda}\right)((\beta_0^{(k)} - \gamma\nabla l(r_{(-k)}, \beta_0^{(k)})) - \gamma\alpha\lambda\,\text{sign}(z))\end{aligned}$$

Then we take norm on the both sides,

$$\begin{aligned}\|z\|_2 &= \left(\frac{\|z\|_2}{\|z\|_2 + \gamma(1-\alpha)\lambda}\right)\|(\beta_0^{(k)} - \gamma\nabla l(r_{(-k)}, \beta_0^{(k)})) - \gamma\alpha\lambda\,\text{sign}(z)\|_2 \\ \Rightarrow \|z\|_2 &= \|(\beta_0^{(k)} - \gamma\nabla l(r_{(-k)}, \beta_0^{(k)})) - \gamma\alpha\lambda\,\text{sign}(z)\|_2 - \gamma(1-\alpha)\lambda \\ \Rightarrow z &= \left(1 - \frac{\gamma(1-\alpha)\lambda}{\|(\beta_0^{(k)} - \gamma\nabla l(r_{(-k)}, \beta_0^{(k)})) - \gamma\alpha\lambda\,\text{sign}(z)\|_2}\right)((\beta_0^{(k)} - \gamma\nabla l(r_{(-k)}, \beta_0^{(k)})) - \gamma\alpha\lambda\,\text{sign}(z))\end{aligned}$$

- When $z > \mathbf{0}$, then

$$\begin{aligned}&\|(\beta_0^{(k)} - \gamma\nabla l(r_{(-k)}, \beta_0^{(k)})) - \gamma\alpha\lambda\,\text{sign}(z)\|_2 \\ &= \|(\beta_0^{(k)} - \gamma\nabla l(r_{(-k)}, \beta_0^{(k)})) - \gamma\alpha\lambda\| \\ &= \|\text{sign}(\beta_0^{(k)} - \gamma\nabla l(r_{(-k)}, \beta_0^{(k)}))(|(\beta_0^{(k)} - \gamma\nabla l(r_{(-k)}, \beta_0^{(k)}))| - \gamma\alpha\lambda)\|_2 \\ &= \|S(\beta_0^{(k)} - \gamma\nabla l(r_{(-k)}, \beta_0^{(k)}), \gamma\alpha\lambda)\|_2\end{aligned}$$

- When $z < \mathbf{0}$, then

$$\begin{aligned}
&\|(\beta_0^{(k)} - \gamma\nabla l(r_{(-k)}, \beta_0^{(k)})) - \gamma\alpha\lambda\,\mathrm{sign}(z)\|_2 \\
&= \|(\beta_0^{(k)} - \gamma\nabla l(r_{(-k)}, \beta_0^{(k)})) + \gamma\alpha\lambda\| \\
&= \|\mathrm{sign}(\beta_0^{(k)} - \gamma\nabla l(r_{(-k)}, \beta_0^{(k)}))(|(\beta_0^{(k)} - \gamma\nabla l(r_{(-k)}, \beta_0^{(k)}))| + \gamma\alpha\lambda)\|_2 \\
&= \|S(\beta_0^{(k)} - \gamma\nabla l(r_{(-k)}, \beta_0^{(k)})), \gamma\alpha\lambda)\|_2
\end{aligned}$$

Therefore, we have

$$\beta^{(k)} = z = \left(1 - \frac{\gamma(1-\alpha)\lambda}{\|S(\beta_0^{(k)} - \gamma\nabla l(r_{(-k)}, \beta_0^{(k)}), \gamma\alpha\lambda)\|_2}\right)_+ S(\beta_0^{(k)} - \gamma\nabla l(r_{(-k)}, \beta_0^{(k)}), \gamma\alpha\lambda)$$

6 References

[1] J. Friedman, T. Hastie, and R. Tibshirani. *A note on the group lasso and a sparse group lasso*. 2010. arXiv: 1001.0736 [math.ST].

[2] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. CRC press, 2015.

[3] Noah Simon et al. "A Sparse-Group Lasso". In: *Journal of Computational and Graphical Statistics* 22.2 (2013), pp. 231–245. DOI: 10.1080/10618600.2012.681250. eprint: https://doi.org/10.1080/10618600.2012.681250. URL: https://doi.org/10.1080/10618600.2012.681250.