Recap

Group Lasso (Yuan and Lin, 2006)

groups $\mathcal{G}_1, \ldots, \mathcal{G}_q$ which build a partition of $\{1, \ldots, p\}$
write the (high-dimensional) parameter vector as

$$\beta = (\beta_{\mathcal{G}_1}, \beta_{\mathcal{G}_2}, \ldots, \beta_{\mathcal{G}_q})^T$$

goal: an estimator which is "group-sparse", i.e.:
for all $j = 1, \ldots, p$,

$$\text{either } \hat{\beta}_{\mathcal{G}_j} \equiv 0$$
$$\text{or } (\hat{\beta}_{\mathcal{G}_j})_r \neq 0 \ \forall r \in \mathcal{G}_j$$

Group Lasso:

$$\hat{\beta}(\lambda) = \text{argmin}_{\beta} \left(\|Y - X\beta\|_2^2/n + \lambda \sum_{j=1}^{q} m_j \|\beta_{\mathcal{G}_j}\|_2 \right)$$

where typically $m_j = \sqrt{|\mathcal{G}_j|}$

group sparsity because objective function is non-differentiable at $\|\beta_{\mathcal{G}_j}\|_2 = 0 \iff \beta_{\mathcal{G}_j} \equiv 0$ $(j = 1, \ldots, q)$

objective function is non-differentiable at $\|\beta_{\mathcal{G}_j}\|_2 = 0$
sub-differential:

$$\frac{\partial}{\partial \beta_{\mathcal{G}_j}} \left(\|Y - X\beta\|_2^2/n + \lambda \sum_{j=1}^{q} m_j \|\beta_{\mathcal{G}_j}\|_2 \right)$$

$$= \quad G(\beta)_{\mathcal{G}_j} + \lambda m_j E(\beta_{\mathcal{G}_j})$$

$$G(\beta) = -2X^T(Y - X\beta)/n \quad (= \text{gradient of } \|Y - X\beta\|_2^2/n)$$

$$E(\beta_{\mathcal{G}_j} = \{e \in \mathbb{R}^{|\mathcal{G}_j|};\ e = \frac{\beta_{\mathcal{G}_j}}{\|\beta_{\mathcal{G}_j}\|_2} \text{ if } \beta_{\mathcal{G}_j} \neq 0,\ \|e\|_2 \leq 1 \text{ if } \beta_{\mathcal{G}_j} = 0\}$$

KKT conditions: solution is characterized by

$$0_{p\times 1} \in \text{ sub-differential}$$

either $\underbrace{\hat{\beta}_{\mathcal{G}_j} \equiv 0}_{\text{point of non-differentiability}}$ or $(\hat{\beta}_{\mathcal{G}_j})_r \neq 0 \; \forall r \in \mathcal{G}_j$

why the second "or $(\hat{\beta}_{\mathcal{G}_j})_r \neq 0 \; \forall r \in \mathcal{G}_j$?" (when $\|\hat{\beta}_{\mathcal{G}_j}\|_2 \neq 0$)

$\rightsquigarrow$

$$0 = G(\hat{\beta})_{\mathcal{G}_j} + \lambda m_j \frac{\hat{\beta}_{\mathcal{G}_j}}{\|\hat{\beta}_{\mathcal{G}_j}\|_2}$$

suppose $X^T X/n = I$ (orthonormal design) and $\exists r \; (\hat{\beta}_{\mathcal{G}_j})_r = 0$:

$$0 = (-2X^T Y/n)_{\mathcal{G}_j} + 2\hat{\beta}_{\mathcal{G}_j} + \lambda m_j \frac{\hat{\beta}_{\mathcal{G}_j}}{\|\hat{\beta}_{\mathcal{G}_j}\|_2}$$

rth component $\quad 0 = -2((X^T Y/n)_{\mathcal{G}_j})_r + 0 + 0$

but it will not happen that $X^T Y$ is zero (random noise in Y)

Sparse Group Lasso

(Simon, Friedman, Hastie & Tibshirani, 2013)

$$\hat{\beta}(\lambda, \alpha)$$

$$= \text{argmin}_\beta(\|Y - X\beta\|_2^2/n + (1-\alpha)\lambda \sum_{j=1}^{q} m_j \|\beta_{\mathcal{G}_j}\|_2 + \alpha\lambda \underbrace{\|\beta\|_1}_{=\sum_{j=1}^{q} \|\beta_{\mathcal{G}_j}\|_1})$$

convex combination of Group Lasso and Lasso penalties

$\rightsquigarrow$ may also lead to sparsity within groups for $\alpha > 0$

The generalized Group Lasso penalty

$$\text{pen}(\beta) = \lambda \sum_{j=1}^{q} m_j \sqrt{\beta_{\mathcal{G}_j}^T A_j \beta_{\mathcal{G}_j}}, \quad A_j \text{ positive definite}$$

important example: groupwise prediction penalty

$$\text{pen}(\beta) = \lambda \sum_{j=1}^{q} m_j \|X_{\mathcal{G}_j} \beta_{\mathcal{G}_j}\|_2 = \lambda \sum_{j=1}^{q} m_j \sqrt{\beta_{\mathcal{G}_j}^T X_{\mathcal{G}_j}^T X_{\mathcal{G}_j} \beta_{\mathcal{G}_j}}$$

$X_{\mathcal{G}_j}^T X_{\mathcal{G}_j}$ typically positive definite for $|\mathcal{G}_j| < n$

- penalty is invariant under arbitrary reparameterizations (bijective mappings) within every group $\mathcal{G}_j$: important!
- when using an orthogonal parameterization such that $X_{\mathcal{G}_j}^T X_{\mathcal{G}_j} = I$: it is the standard Group Lasso