IV. Group Lasso (Ch. 4 in Bühlm. & van de Geer)

$$Y = X\beta^0 + \varepsilon$$

estimated and $\underset{\beta^0}{\underset{\|}{\text{true}}}$ parameter are high-dim. with group structure

$$\beta = (\beta_{g_1}, \beta_{g_2}, \ldots, \beta_{g_q}), \quad \beta_{g_j} = \{\beta_r \,;\, r \in g_j\}$$

$$g_j \subseteq \{1, \ldots, p\} \qquad \bigcup_{j=1}^{q} g_j = \{1, \ldots, p\} \,;\; g_j \cap g_k = \emptyset \;(j \neq k)$$

__Example:__ (Ch. 4.3 in Book)

$Y \in \mathbb{R}$, $X = (X^{(1)}, \dots, X^{(p)})$

$\forall j$: $X^{(j)} \in \{0, 1, 2, 3\}$ a factor with 4 levels

e.g. the labels A, C, G, T in DNA sequences

for example $p = 2$:

$$Y_i = \mu + \underbrace{\sum_{k=0}^{3} \gamma_k I(X_i^{(1)} = k) + \sum_{k=0}^{3} \delta_k I(X_i^{(2)} = k)}_{\text{main effects}}$$

$$+ \underbrace{\sum_{k,\ell=0}^{3} \kappa_{k,\ell} I(X_i^{(1)} = k, X_i^{(2)} = \ell)}_{\text{interaction terms}} + \varepsilon_i$$

can write this as

$$Y = X\beta + \varepsilon$$

with groups / parameters: $\beta_1 = \mu$

$$\underbrace{g_1 = \{2,3,4\}}_{\text{corresponds to } X^{(1)} \text{ main effect}} \; ; \; \beta_{g_1}$$

$$\underbrace{g_2 = \{5,6,7\}}_{\text{corresponds to } X^{(2)} \text{ main effect}}$$

$$\underbrace{g_3 = \{8,9,\ldots,16\}}_{\text{corresponds to interaction } X^{(1)},\, X^{(2)}}$$

overparametrized:

for example

$$\underbrace{\begin{aligned} &\sum_{k=0}^{3} \gamma_k = 0 = \sum_{k=0}^{3} \delta_k \\ &\sum_{k=0}^{3} u_{k,l} = 0 = \sum_{l=0}^{3} u_{k,l} \qquad \forall k,l \\ &\rightsquigarrow \sum_{k,l=0}^{3} u_{k,l} = 0 \end{aligned}}_{\text{sum constraints}}$$

~~free p~~ constraints: $1 + 1 + 4 + \cancel{4} \; 3 = 9$

free parameters: $1 + 4 + 4 + 16 - 9 = 16$

aim: we want sparsity in terms of whole groups

either $\hat{\beta}_{g_j} \equiv 0$ or $(\hat{\beta}_{g_j})_r \neq 0 \ \forall r \in g_j$

this can be achieved by Group Lasso (Yuan and Lin, 2006)

even if $p \gg n$

$$\hat{\beta}(\lambda) = \underset{\beta}{\operatorname{argmin}} \left(\|Y - X\beta\|_2^2/n + \lambda \sum_{j=1}^{q} \underbrace{m_j}_{\text{weight}} \|\beta_{g_j}\|_{\cancel{2}} \right)$$

$$m_j \underset{\text{typically}}{=} \sqrt{|g_j|}$$

then $m_j \|\beta_{g_j}\|_2 \overset{\text{if } \|\beta_{g_j}\|_2 \neq 0}{\asymp} \sqrt{|g_j|} \cdot \sqrt{|g_j|} = |g_j|$

for lasso: jth group if $|\beta_k| \neq 0 \ \forall k$

$$\sum_{k \in g_j} |\beta_k| \simeq |g_j|$$

heuristics:

Sparsity happens at points of non-differentiability

here: $\|\beta_{g_j}\|_2^{\cancel{2}} = \sqrt{\|\beta_{g_j}\|_2^2}$

but $\sqrt{\cdot}$ is not differentiable at zero

$\rightsquigarrow$ sparsity happens at $\underbrace{\|\beta_{g_j}\|_2^2 = 0}$

$\Leftrightarrow \beta_{g_j} \equiv 0$

if not sparse where $\|\beta_{g_j}\|_2^2 > 0 \rightsquigarrow$ no inside sparsity in β_{g_j}

$\rightarrow (\beta_{g_j})_r \neq 0 \quad \forall r \in g_j$

formally: sub-differential

$$\frac{\partial}{\partial \beta_{g_j}} \quad \ldots$$

$$\frac{\partial}{\partial \beta_{g_j}} \|\beta_{g_j}\|_2 = \frac{\beta_{g_j}}{\|\beta_{g_j}\|_2} \qquad \text{if } \beta_{g_j} \neq 0$$