IV. Group Lasso

(Ch. 4 in PB&vdG (2011))

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

estimated (or true) β is a high-dim. parameter vector with a group structure

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

$$\bigcup_{j=1}^{q} g_j = \{1, \dots, p\}, \quad g_i \cap g_j = \emptyset \quad (i \neq j)$$

Example: (Ch 4.3 in PB&vdG)

$Y \in \mathbb{R}$

$X = (X^{(1)}, \ldots, X^{(p)})$

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

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

for example $p = 2$

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

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

overparameterized $\rightsquigarrow$ $\sum_{k=0}^{3} \gamma_k = 0$, $\sum_k \delta_k = 0$

$$\forall \ell \quad \sum_k \kappa_{k,\ell} = 0; \quad \sum_\ell \kappa_{k,\ell} = 0 \quad \forall k$$

$$\left(\rightsquigarrow \sum_{k,\ell} \kappa_{k,\ell} = 0 \right)$$

$\underbrace{1 + 1 + 4 + (4-1)}_{9 \text{ constraints}}$ constraints

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

can write this as

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

group parameters: $\beta_1 = \mu$ — main eff. of $X^{(3)}$

corresp. to main eff. of $X^{(1)}$ — $g_1 = \{2,3,4\}$, $g_2 = \{5,6,7\}$

$g_3 = \{8, 9, \ldots, 16\}$ — interaction effects

aim: we want sparsity in terms of whole groups

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

this can be achieved by the Group Lasso penalty

(Yuan and Lin, 2006)

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

$$m_j = \sqrt{|g_j|}$$

Lasso: $g_j = \{j\}$

$m_j = 1$; $\|\beta_{g_j}\|_2 = \sqrt{\beta_j^2} = |\beta_j|$

$\rightsquigarrow$ Grp lasso equals lasso for $g_j = \{j\}$

why $m_j = \sqrt{|g_j|}$?

for lasso: $(\beta_{g_j})_r \neq 0 \quad \forall r \in g_j$

$\rightsquigarrow \sum_{r} |(\beta_{g_j})_r| \simeq |g_j|$

with Grp lasso: $\|\beta_{g_j}\|_2 \simeq \sqrt{|g_j|} \rightsquigarrow m_j \|\beta_{g_j}\|_2 \simeq |g_j|$

why does sparsity hold in terms of

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

reason:

sparsity happens at points of non-differentiability

(if derivative is ~~undefined~~ "not determined" it pays off to "set to zero")

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

$\frac{\partial}{\partial \beta_{g_j}}$: if $\|\beta_{g_j}\|_2 \neq 0$

$\rightsquigarrow = \frac{1}{2}\left(\|\beta_{g_j}\|_2^2\right)^{-1/2} \cdot 2\,\beta_{g_j}$

$= \frac{\beta_{g_j}}{\|\beta_{g_j}\|_2}$

if $\|\beta_{g_j}\|_2 = 0$ $\rightsquigarrow$ not differentiable

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

because Ridge regression is not sparse

V. Additive models and many smooth univariate functions

(Ch. 5 in PB&vdG)

V.1. Model and estimation (Ch. 5.2 in PB&vdG)

additive model:

$$Y_i = \mu + \sum_{j=1}^{p} f_j^0(X_i^{(j)}) + \varepsilon_i$$

$\mathbb{E}[\varepsilon_i] = 0$, ε_i indep. of X_i (if X_i is random)

identification: $\sum_{i=1}^{n} f_j^0(X_i^{(j)}) \equiv 0 \quad \forall j$

for $p < n$: e.g. Hastie & Tibshirani (1990)

f_j^0: smooth $\mathbb{R} \longrightarrow \mathbb{R}$

regularize w.r.t. smoothness

for $p \gg n$: need to regularize further

approach: use basis expansion

each f_j^0 is approximated with K basis functions

$\rightsquigarrow$ # param: $1 + pK$

still OK if problem is sparse since

$\log(\dim)/n = \log(1+pK)/n$ small