4.) Odabir varijabli u lin. modelu

U lin. modelu, ako je p blizu n (ili čak $p >> n$), ali

$$\|\beta_*\|_0 := \sum_{j=1}^{p} \mathbb{1}_{\{\beta_j \neq 0\}} << p, \quad (\text{"sparsity"})$$

korištenjem samo podskupa kovarijata možemo

(i) smanjiti testnu grešku [t.d. smanjimo varijancu]

(ii) dobiti interpretabilnije rješenje.

4.1 "Klasičan" pristup

↳ S. svakom modelu (tj. $A \subseteq \{1, \dots, p\}$) pridružimo "skor"

$$s(A) := \underbrace{\frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \sum_{j \in A} \hat{\beta}_j x_{ij}\right)^2}_{L_I(\hat{\beta}_A) = \frac{RSS(\hat{\beta}_A)}{n}} + \underbrace{p(A)}_{\text{kazna za kompleksnost}} \qquad (4.1)$$

$\left(\hat{\beta}_A = \underset{\beta_A \in \mathbb{R}^{|A|+1}}{\operatorname{argmin}} RSS(\beta_A)\right)$

npr. • $p(A) = \frac{2\hat{\sigma}^2}{n} \cdot |A|$ → C_p statistika (vidi Nap. 2.8),
→ ili što kao tzv. **AIC** kriterij

• $p(A) = \log n \cdot \frac{\hat{\sigma}^2}{n} \cdot |A|$ → tzv. **BIC** kriterij

[motivacija : $s(A) = \hat{L}^{in}$]

↳ uočimo, za A, A' t.d. $|A| = |A'|$ imamo zbog $\boxed{p(A) = p(A')}$,

$$s(A) \leq s(A') \iff RSS(\hat{\beta}_A) \leq RSS(\hat{\beta}_{A'}).$$

1° Nekim algoritmom nađemo "predstavnike" $A_0, A_1, \ldots, A_p$ t.d. $|A_k| = k, \forall k = 1, \ldots, p$.

samo slobodni član

npr. • "best subset": $\boxed{A_k = \underset{|A|=k}{\operatorname{argmin}} RSS(\hat{\beta}_A)}$

→ gledamo sve modele

⇒ izvedivo za $\boxed{p \leq 50}$ [otprilike]

• "forward stepwise":

$\forall k = 1, 2, \ldots, p$, $\boxed{A_k = \underset{\substack{|A|=k: \\ A_{k-1} \subseteq A_k}}{\operatorname{argmin}} RSS(\hat{\beta}_A)}$

• "backward stepwise":

$\forall k = p-1, \ldots, 1$, $\boxed{A_k = \underset{\substack{|A|=k, \\ A_k \subseteq A_{k+1}}}{\operatorname{argmin}} RSS(\hat{\beta}_A)}$

[$p > n$?]

2° Na temelju CV metode ili skora u (4.1) izaberemo jedan od modela $A_0, \ldots, A_p$.

[Problem "varijance" za best subset (čak i kada je $p \ll 50$)?]

4.2 Lasso

Predp. da je $\bar{x}_j = 0, \forall j = 1, \ldots, p$, $\bar{y} = 0$ (a tipično $\widehat{Var}(x_j) = 1, \forall j$)

Za $\lambda \geq 0$,

$$\hat{\beta}^{lass}_\lambda = \hat{\beta}^{lass} := \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{n} \| y - X\beta \|_2^2 + \lambda \cdot \underbrace{\|\beta\|_1}_{= \sum_{j=1}^p |\beta_j|} \quad (4.2)$$

$$= \underset{\beta}{\operatorname{argmin}} \underbrace{\| y - X\beta \|_2^2}_{= RSS(\beta)} + n\lambda \|\beta\|_1$$

$\hookrightarrow$ tipično $\boxed{\hat{\beta}^{lass}_j = 0}$ za neke (mnoge) $j \in \{1, \ldots, p\}$!

[$\Rightarrow$ lasso naredi odabir varijabli!]

Predp. da je $\boxed{\hat{\Sigma} = \frac{1}{n} X^T X = I_p}$: [predp. $p \leq n$]

(1°) $\forall \beta \in \mathbb{R}^p$, $\| y - X\beta \|_2^2 = \| \underbrace{(y - X\hat{\beta}^{ls})}_{\perp \text{ na } x_1, \ldots, x_p} + \underbrace{(X\hat{\beta}^{ls} - X\beta)}_{\in \operatorname{span}\{x_1, \ldots, x_p\}} \|_2^2$

$$= \| y - X\hat{\beta}^{ls} \|_2^2 + \| X\hat{\beta}^{ls} - X\beta \|_2^2 .$$

(2°) $\hat{\beta}^{lass} \underset{(1°)}{=} \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{n} \| X\hat{\beta}^{ls} - X\beta \| + \lambda \|\beta\|_1$

$$= \underset{\beta}{\operatorname{argmin}} \frac{1}{n} (\hat{\beta}^{ls} - \beta)^T X^T X (\hat{\beta}^{ls} - \beta) + \lambda \|\beta\|_1$$

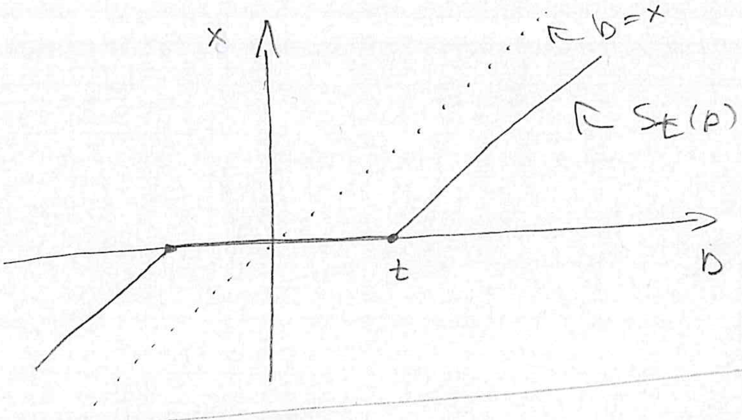
$$\underset{\hat{\Sigma} = I_p}{=} \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{j=1}^p (\beta_j - \hat{\beta}_j^{ls})^2 + \lambda |\beta_j| \right\}$$

$$\Rightarrow \boxed{\hat{\beta}_j^{lass} = \underset{x \in \mathbb{R}}{\operatorname{argmin}} (x - \hat{\beta}_j^{ls})^2 + \lambda |x| \quad , \forall j = 1, \ldots, p} \quad (4.3)$$

(3°) (Z2) $\forall b \in \mathbb{R}, \quad \underset{x \in \mathbb{R}}{\operatorname{argmin}} (x-b)^2 + \lambda |x| = S_{\lambda/2}(b)$, gdje je

$$S_t(b) = \begin{cases} b - t, & \text{za } b > t \\ 0, & \text{za } |b| \leq t \\ b + t, & \text{za } b < -t \end{cases} \tag{4.4}$$

$$= \operatorname{sgn}(b)\left(|b| - \lambda/2\right)_+, \quad t \geq 0$$



$$\Rightarrow \boxed{\hat{\beta}_j^{las} = \operatorname{sgn}(\hat{\beta}^{ls})\left(|\hat{\beta}_j^{ls}| - \lambda/2\right)_+} \tag{4.5}$$

↳ lasso radi <u>odabir</u> varijabli i <u>smanjuje</u> koeficijente!

[neprekidna verzija best subset selec.!]

(4.2.1) Alternativna formulacija za lasso i ridge

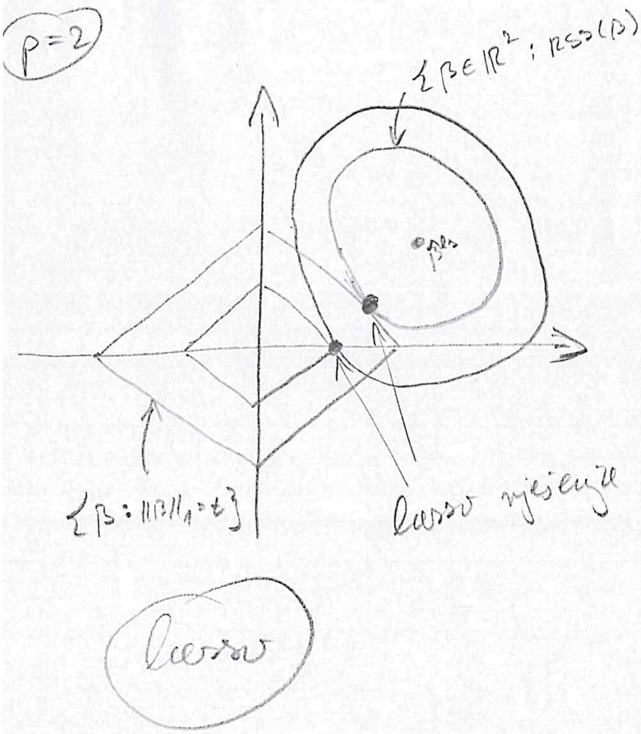
Može se pokazati da $\forall \lambda \geq 0, \exists t_\lambda, t'_\lambda \in [0, +\infty]$ t.d.

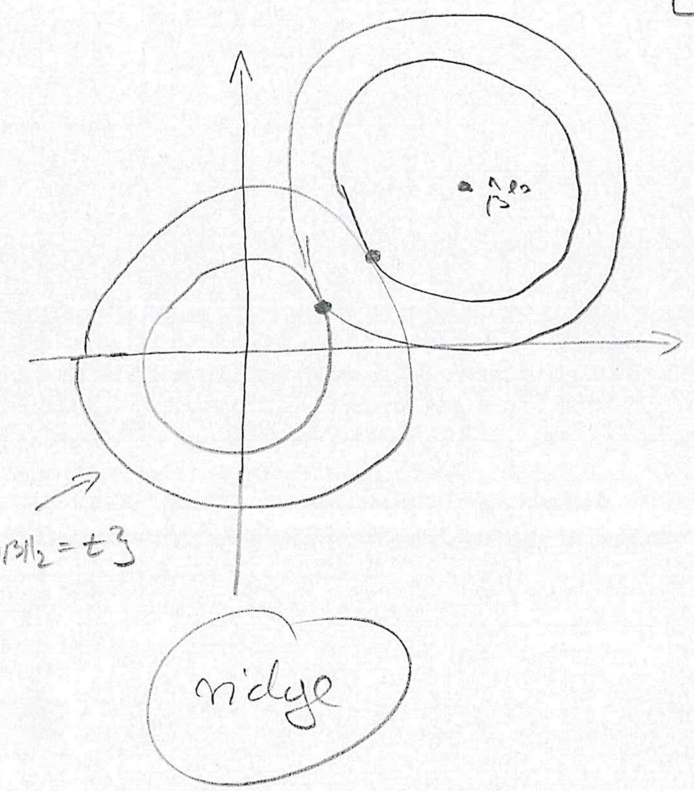
$$\hat{\beta}^r = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \tfrac{1}{n} \| y - X\beta \|_2^2 \right\} \text{ uz uvjet } \boxed{\|\beta\|_2 \leq t_\lambda}, \tag{4.6}$$

$$\hat{\beta}^{las} = \text{———} \; '' \; \text{———} \; \boxed{\|\beta\|_1 \leq t'_\lambda}, \tag{4.7}$$

pri čemu $t_\lambda, t_\lambda' \to 0$ kada $\lambda \to +\infty$.

p=2





U slučaju koreliranih koenficijenata lasso tipično bira samo jednu (ili neke) od njih. [za razliku od ridge regresije!]

↳ kompromis između lasso i ridge → "elastic net"

$$\hat{\beta}^{net} = \underset{\beta}{\operatorname{argmin}} \frac{1}{2}\|y - X\beta\|_2^2 + \boxed{\lambda \sum_{j=1}^{p} \left(\alpha \beta_j^2 + (1-\alpha)|\beta_j|\right)}. \quad (4.8)$$

za $\lambda \geq 0$, $\alpha \in [0,1]$.

→ ridge_vs_lasso_shrinkage.pdf
→ ——"—— MSE.pdf
→ ISLR_ridge_lasso_bss.R

[Pop.] • Praktično računanje lasso rješenja je netrivijalna stvar
→ nećemo se time baviti

• kada je $p > n$, rješenje ne mora biti jedinstveno

• može se pokazati da vrijedi

$$df(\text{lasso}_\lambda) = \mathbb{E}\left[\|\hat{\beta}_\lambda^{las}\|_0\right]$$

nepristrani procjenitelj za df!