3) Regularizacija linearnog modela [ridge regresija]

[Notacija]

↳ [prvo želimo pokazati da LS procjena može biti loša i kada je Lin. model točan !?]

- vektore u $\mathbb{R}^k$ ($k \in \mathbb{N}$) shvaćamo kao vektor-stupce

- $$\mathbb{X} = \begin{bmatrix} — & X^{(1)T} & — \\ & \vdots & \\ — & X^{(n)T} & — \end{bmatrix} = \begin{bmatrix} x_{11}, \dots, x_{1p} \\ \vdots \\ x_{n1}, \dots, x_{np} \end{bmatrix} \in \mathbb{R}^{n \times p}$$

matrica dizajna

- $y = (y_1, \dots, y_n)^T \in \mathbb{R}^{n \times 1}$

- $\mathbb{X}_j = (x_{1j}, \dots, x_{nj})^T \in \mathbb{R}^{n \times 1}$, $j = 1, \dots, p$

↳ j-ti stupac od $\mathbb{X}$ (umjesto kovarijate X_j u svim $X^{(i)}, i = 1, \dots, n$)

(često u $\mathbb{X}$ dodajemo $\mathbb{X}_0 = (1, \dots, 1)^T \in \mathbb{R}^{n \times 1}$, $x_{10}, \dots, x_{n0}$)

Lin. model: biramo $\hat{f} = f(\tau)$ iz $\mathcal{F} = \{ f_\beta : \beta \in \Theta \}$ uz $\Theta \subseteq \mathbb{R}^{p+1}$;

$$f_\beta(x) = X^T \cdot \beta, \quad X = (x_1, \dots, x_p) \in \mathbb{R}^{p+1}. \qquad (3.1)$$

[dopuštamo (i) $X_2 := f(X_1)$, (ii) interakcije $X_3 := X_1 \cdot X_2$, (ii) kvalitativne kovarijate.]

1) Metoda najmanjih kvadrata → (LS) ["least squares"]

$$\hat{\beta}^{ls} = (\hat{\beta}^{ls}_0, \dots, \hat{\beta}^{ls}_p) := \underset{\beta \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \underbrace{\frac{1}{n} \sum_{i=1}^{n} (y_i - f_\beta(x^{(i)}))^2}_{=: RSS(\beta)}$$

$$L_I(f_\beta) \text{ uz } L = (y - \hat{y})^2 \qquad (3.2)$$

$$= \underset{\beta}{\operatorname{argmin}} \| y - \mathbb{X}\beta \|_2^2$$

Pretpostavljamo da su $X_0, \dots, X_p \in \mathbb{R}^n$ lin. nezavisni, tj. da je $X^T X \in \mathbb{R}^{(p+1)\times(p+1)}$ poz. definitna (simetrična) matrica.

(mora je $\boxed{n \geq p+1}$)

- $$\hat{\beta}^{OLS} = (X^T X)^{-1} X^T y = \frac{1}{n} \hat{\Sigma}^{-1} X^T y, \qquad (3.3)$$

gdje je $\boxed{\hat{\Sigma} := \frac{1}{n} X^T X}$. [pazi se stupci centrirani, onda je točno uzorčka kovarijacijska matrica]

- $$\hat{y} := \left(f_{\hat{\beta}^{OLS}}(x^{(1)}), \dots, f_{\hat{\beta}^{OLS}}(x^{(n)}) \right)^T = X \cdot \hat{\beta}^{OLS} = H y, \qquad (3.4)$$

gdje $H = X (X^T X)^{-1} X^T$ ["hat matrix"],

te je $\hat{y}$ ort. projekcija vektora y na $\text{span}(X_0, \dots, X_p)$.

Nap.1 U slučaju da $X_0, \dots, X_p$ nisu lin. nezavisni (npr. $n < p+1$ ili $X_j = 0 \cdot X_i$) $\hat{\beta}^{OLS}$ nije jedinstven. Ipak, $\hat{y}$ je i dalje ortogonalna projekcija od y na $\text{span}(X_0, \dots, X_p)$. //

(3.2) Statistička svojstva od $\hat{\beta}^{OLS}$

[moramo pretpostaviti model za (X, Y)]

Pretpostavimo da $\boxed{\exists \beta_* \in \mathbb{R}^{p+1}}$ t.d.

$$Y_i = (X^{(i)})^T \beta_* + \varepsilon_i, \quad i = 1, \dots, n \qquad (3.5)$$

pri čemu su [nadi jednostavnosti]

- $X^{(i)}$, $i = 1, \dots, n$ neslučajni (tzv. "fiksni dizajn")
- ε_i, $i = 1, \dots, n$, nezavisni t.d. $\boxed{\mathbb{E}[\varepsilon_i] = 0}$, $\boxed{\text{Var}(\varepsilon_i) = \sigma^2}$, $\forall i = 1, \dots, n$. za neki $\sigma^2 > 0$.

Nap. 1 u (3.5) pretpostavljamo da je lin. model *točan*, tj. da nema *greške aproksimacije* ("bias").

(3.5) možemo zapisati kao [tikovi]

$$Y := \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \mathbb{X}\beta_* + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix} =: \mathbb{X}\beta_* + \mathcal{E}, \qquad (3.6)$$

uz $\mathbb{E}[\mathcal{E}] = 0$, $\mathrm{Var}(\mathcal{E}) = \sigma^2 I_n$.

$$\Rightarrow \quad \hat{\beta}^{LS} = \beta_* + \frac{1}{n}\hat{\Sigma}^{-1}\mathbb{X}^T\mathcal{E} \quad \text{(jedina slučajna)}$$

(3.3)

Lako →

- $\mathbb{E}[\hat{\beta}^{LS}] = \beta_*$
- $\mathrm{Var}(\hat{\beta}^{LS}) = \frac{\sigma^2}{n}\hat{\Sigma}^{-1}$ (3.7)

Tipično

$$\hat{\sigma}^2 := \frac{1}{n-p-1}\|Y - \hat{Y}\|_2^2 = \frac{RSS(\hat{\beta})}{n-p-1}. \qquad (3.8)$$

3.2.1 *Testna greška* → F. Bach, 3.5

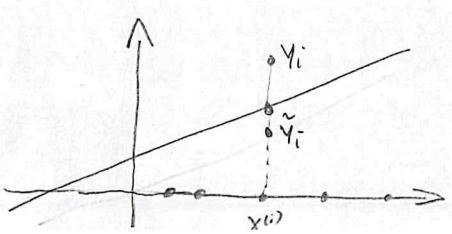
Neka je $\tilde{Y} := \mathbb{X}\beta_* + \tilde{\mathcal{E}}$, pri čemu su $\mathcal{E}$ i $\tilde{\mathcal{E}}$ (tj. Y i $\tilde{Y}$) nezavisni, testni skup.

Promatramo tzv. *testnu grešku unutar uzorka* ("in-sample")

$$L^{in}(f_\beta) := \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(\tilde{Y}_i - f_\beta(x^{(i)}))^2\right]$$

$$= \mathbb{E}\left[\frac{1}{n}\|\tilde{Y} - \mathbb{X}\beta\|_2^2\right], \qquad (3.9)$$

za $\forall \beta \in \mathbb{R}^{p+1}$.



Prop. 3.1 (BVT) Ako je $\hat{\beta} = \hat{\beta}(\mathbb{Y}, \mathbb{X})$ procjenitelj za β_*,

$$\mathbb{E}\left[L^{in}(f_{\hat{\beta}})\right] = \sigma^2 + \underbrace{\|\mathbb{E}[\hat{\beta}] - \beta_*\|^2_{\hat{\Sigma}}}_{\text{"pristranost"}} + \underbrace{\mathbb{E}\left[\|\hat{\beta} - \mathbb{E}[\hat{\beta}]\|^2_{\hat{\Sigma}}\right]}_{\text{"varijanca"}}, \tag{3.10}$$

gdje $\boxed{\mathbb{E} = \mathbb{E}_{\mathbb{Y}}}$ te $\boxed{\|\beta\|^2_{\hat{\Sigma}} := \beta^T \hat{\Sigma} \beta}$, $\beta \in \mathbb{R}^{p+1}$.

[uzima u obzir distribuciju podataka]

Dokaz:

(1°) Iz (3.6) za $\tilde{\mathbb{Y}}$ imamo, $\forall \beta \in \mathbb{R}^{p+1}$,

$$n \cdot L^{in}(f_\beta) = \mathbb{E}\left[\|\tilde{\mathbb{Y}} - \mathbb{X}\beta\|_2^2\right] = \mathbb{E}\left[\|\mathbb{X}(\beta_* - \beta) + \tilde{\mathcal{E}}\|_2^2\right]$$

$$= \left[\|a+b\|_2^2 = \langle a+b, a+b\rangle = \|a\|_2^2 + \|b\|_2^2 + 2 \cdot a^T b \;,\; a,b \in \mathbb{R}^{p+1}\right]$$

$$= \mathbb{E}\Big[\|\mathbb{X}(\beta_* - \beta)\|_2^2 + \|\tilde{\mathcal{E}}\|_2^2 + \underbrace{2\left(\mathbb{X}(\beta_* - \beta)\right)^T \tilde{\mathcal{E}}}_{\text{nije slučajno}}\Big]$$

$$= \left[\mathbb{E}\left[\|\tilde{\mathcal{E}}\|_2^2\right] = n \cdot \sigma^2,\; \mathbb{E}[\tilde{\mathcal{E}}] = 0 \in \mathbb{R}^n\right]$$

$$= n \cdot \sigma^2 + (\beta_* - \beta)^T \mathbb{X}^T \mathbb{X} (\beta_* - \beta)$$

min za β_K je $:= \sigma^2$

$$\Rightarrow \boxed{L^{in}(f_\beta) = \sigma^2 + \|\beta_* - \beta\|^2_{\hat{\Sigma}}} \tag{3.11}$$

(2°) (OZ)

$$\mathbb{E}\left[L^{in}(f_{\hat{\beta}})\right] \overset{(3.11)}{=} \sigma^2 + \mathbb{E}\left[\|\beta_* - \hat{\beta}\|^2_{\hat{\Sigma}}\right]$$

← ovisi o $\mathbb{Y}$

$$\hookrightarrow \mathbb{E}\left[\|\beta_* - \hat{\beta}\|^2_{\hat{\Sigma}}\right] = \mathbb{E}\left[\|(\beta_* - \mathbb{E}[\hat{\beta}]) + (\mathbb{E}[\hat{\beta}] - \hat{\beta})\|^2_{\hat{\Sigma}}\right]$$

$$= \left[\|a+b\|^2_{\hat{\Sigma}} \overset{OZ}{=} \|a\|^2_{\hat{\Sigma}} + \|b\|^2_{\hat{\Sigma}} + 2\, a^T \hat{\Sigma} b\right]$$

$$= \mathbb{E}\left[\|\beta_* - \mathbb{E}[\hat{\beta}]\|^2_{\hat{\Sigma}}\right] + \mathbb{E}\left[\|\mathbb{E}[\hat{\beta}] - \hat{\beta}\|^2_{\hat{\Sigma}}\right] +$$

$$+ 2\, \mathbb{E}\Big[\underbrace{(\beta_* - \mathbb{E}[\hat\beta])^T \hat\Sigma}_{\text{nije slučajno}} (\mathbb{E}[\hat\beta] - \hat\beta) \Big]$$

$$= (\beta_* - \mathbb{E}[\hat\beta])^T \hat\Sigma \, \underbrace{\mathbb{E}\big[\mathbb{E}[\hat\beta] - \hat\beta\big]}_{=0} \qquad \square$$

Prop. 3.2 (!)

$$\mathbb{E}\big[L^{in}(\hat\beta^{ols}) \big] = \sigma^2 + 0 + \boxed{\frac{\sigma^2(p+1)}{n}} \qquad (3.12)$$

Dokaz. Slijedi iz Prop. 3.1 jer $\mathbb{E}[\hat\beta^{ols}] = \beta_*$ te

$$\mathbb{E}\Big[\| \hat\beta^{ols} - \mathbb{E}[\hat\beta^{ols}] \|_{\hat\Sigma}^2 \Big] = \mathbb{E}\Big[\underbrace{(\hat\beta^{ols} - \mathbb{E}[\hat\beta^{ols}])^T \hat\Sigma (\hat\beta^{ols} - \mathbb{E}[\hat\beta^{ols}])}_{\in \mathbb{R}} \Big]$$

$$= \mathbb{E}[\, tr(\,\cdot\,) \,] = \mathbb{E}\Big[tr\big((\hat\beta^{ols} - \mathbb{E}[\hat\beta^{ols}]) (\hat\beta^{ols} - \mathbb{E}[\hat\beta^{ols}])^T \hat\Sigma \big) \Big]$$

($\hat\Sigma$: nije slučajno)

$$= tr\Big(\underbrace{Var(\hat\beta^{ols})}_{\overset{(3.7)}{=} \frac{\sigma^2}{n} \hat\Sigma^{-1}} \cdot \hat\Sigma \Big) = \frac{\sigma^2}{n}\, tr(I_{p+1}) = \frac{\sigma^2(p+1)}{n}$$

$\square$

Alternativno,

$$\mathbb{E}\Big[\| \underbrace{\hat\beta^{ols} - \mathbb{E}[\hat\beta^{ols}]}_{= \hat\beta^{ols} - \beta_*} \|_{\hat\Sigma}^2 \Big] = \mathbb{E}\Big[\| (X^T X)^{-1} X^T \varepsilon \|_{\hat\Sigma}^2 \Big]$$

$$= \dots = \frac{1}{n} \mathbb{E}\Big[\underbrace{\varepsilon^T H \varepsilon}_{= tr(\varepsilon^T H \varepsilon)} \Big] = \frac{1}{n} \mathbb{E}\big[tr(\varepsilon \varepsilon^T H) \big] \overset{\mathbb{E}[\varepsilon\varepsilon^T] = \sigma^2 I_n}{=} \frac{\sigma^2}{n} \underbrace{tr(H)}_{= p+1} = \frac{\sigma^2}{n}(p+1)$$

$\square$

Nap. 3.3 |

(i) Iz $\hat{y} = Hy$ te Prop. 2.7 slijedi da

$$df(LS) = tr(H) = p+1 = \#\text{ procijenjenih parametara.}$$

H ort. projektor na $\underbrace{\text{span}\{x_0, \dots, x_p\}}_{\text{lin. nez.}}$

(Dz) Nadalje, iz Prop (2.5) i (3.2) imamo $\left(\mathbb{E}[L_T(\hat{f})] = \mathbb{E}_y[L^{in}(\hat{f})]\right)$,

$$\mathbb{E}\left[L_T(f_{\hat{\beta}^{LS}})\right] = \sigma^2 \ominus \frac{\sigma^2(p+1)}{n}$$

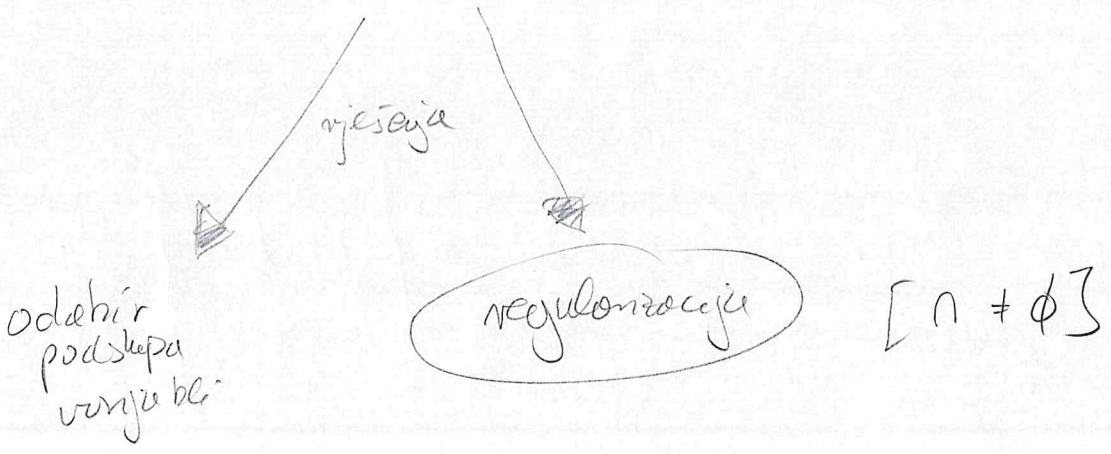
[trening greška]

$$= \sigma^2 \frac{n-p-1}{n} \cdot \left[\overset{p+1 \to n}{\longrightarrow} 0\right]$$

[Specijalno, za $\hat{\sigma}^2$ iz (3.8) imamo $\mathbb{E}[\hat{\sigma}^2] = \sigma^2$.]

(ii) Dakle, čak i kad je lin. model točan (pa je pristranost = 0), varijanca [a zovnim time i ukupna greška] je velika [u odnosu na σ^2] ako je $\frac{p+1}{n}$ velik!

[čak i da svih p kovarijata znatno utječu na odziv !]



(3.3) Ridge regresija

Pr. (Motivacija) [Zašto ne uvijek nepristranost?]

Pretp. da $z_1, \dots, z_n$ njd t.d. $z_i \sim N(\Theta, d^2)$.

- $MSE(\bar{z}) := \mathbb{E}\left[(\bar{z} - \Theta)^2\right] = Var(\bar{z}) + 0 = \frac{d^2}{n} + 0$
- $MSE(\bar{z}/2) := \mathbb{E}\left[(\frac{\bar{z}}{2} - \Theta)^2\right] \underset{\text{Dz}}{=} Var(\bar{z}/2) + (\mathbb{E}[\bar{z}/2] - \Theta)^2 = \frac{d^2}{4n} + \boxed{\frac{\Theta^2}{4}}$

$\hookrightarrow MSE(\bar{z}) \leq MSE(\bar{z}/2)$

$\Leftrightarrow$

$\boxed{n \geq \frac{3d^2}{\Theta^2}}$ ← $\boxed{\text{velike ako } \Theta \approx 0.}$

$\blacksquare$

Za $\lambda \geq 0$ definiramo

$$\hat{\beta}^r_\lambda = \hat{\beta}^r = (\hat{\beta}^r_0, \dots, \hat{\beta}^r_p) = \underset{\beta \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - f_\beta(x_{(i)}))^2 + \boxed{\lambda \sum_{j=1}^{p} \beta_j^2} \right\}$$

$$= \underset{\beta \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \left\{ RSS(f_\beta) + n\lambda \sum_{j=1}^{p} \beta_j^2 \right\}.$$

(3.13)

"L_2 regularizacija"

ne uključuje β_0!

3.3.4 $\forall \lambda \geq 0, \forall \beta_1, \dots, \beta_p \in \mathbb{R},$

$$\underset{\beta_0 \in \mathbb{R}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + n\lambda \sum_{j=1}^{p} \beta_j^2 \right\} = \bar{y} - \sum_{j=1}^{p} \beta_j \bar{x}_j,$$

je je $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$, $j = 1, \dots, p$.

$$(\hat{\beta}^r_1, \dots, \hat{\beta}^r_p) = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} (y_i - \bar{y} - \sum_{j=1}^{p} (x_{ij} - \bar{x}_j)\beta_j)^2 + n\lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

↑ $y_i - \bar{y}$ ↑ $x_{ij} - \bar{x}_j$

Zbog Nap. 3.4 u nastavku pretpostavljamo da je

$$\bar{x}_j = 0, \ \forall j = 1, \dots, p, \quad \bar{y} = 0, \qquad (3.14)$$

te rješavamo

$$\hat{\beta}^r = (\hat{\beta}^r_1, \dots, \hat{\beta}^r_p) = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \|y - X\beta\|_2^2 + n\lambda \|\beta\|_2^2 \right\} \qquad (3.15)$$

$X \in \mathbb{R}^{n \times p}$

(vidi (3.24))

Nap. 3.5 | Uz centriranje, kovarijate tipično još i skalirano t.d. imaju uzoračku varijancu 1 jer je kazna $\|\beta\|_2^2$ "nepoštena" ako kovarijate nisu u istim mjernim jedinicama.

$$\left[\left(\frac{Y - \bar{y}}{s_y} \right) = \hat{\beta}^r_1 \cdot \left(\frac{X_1 - \bar{x}_1}{s_1} \right) + \dots + \hat{\beta}^r_p \cdot \left(\frac{X_p - \bar{x}_p}{s_p} \right) \Rightarrow Y = \hat{\beta}^r_0 + \hat{\beta}^r_1 \frac{s_y}{s_1} X_1 + \dots + \hat{\beta}^r_p \frac{s_y}{s_p} X_p \right] \blacksquare$$

$\hat{\beta}^r_0 = \bar{y} - \sum_{j=1}^{p} \bar{x}_j \frac{s_y}{s_j} \cdot \hat{\beta}^r_j$

Prp 3.6 | $\forall \lambda \geq 0$,

$$\hat{\beta}^r = \frac{1}{n} \left(\hat{\Sigma} + \lambda I \right)^{-1} X^T y, \qquad (3.16)$$

Dokaz | (DZ). → analogno kao za $\hat{\beta}^{ls}$.

Nap. | (i) za $\lambda > 0$, $\hat{\beta}^r$ je uvijek dobro definirano (čak i u slučaju $p > n$!)

(ii) • $\hat{\beta}^r = \hat{\beta}^{ls}$ za $\lambda = 0$,
• $\hat{\beta}^r \xrightarrow{\lambda \to \infty} 0$.

} hiperparametar λ kontrolira efekt smanjenja (engl. shrinkage) koeficijenata prema 0.

→ ridge – primjer – smanjenja – koef

Prp. 3.7 | Pod pretpostavkama u (3.5), $\forall \lambda \geq 0$,

$$\mathbb{E}\left[L^{(r)}\left(\hat{\beta}^r_\lambda \right) \right] = \sigma^2 + \underbrace{\lambda^2 \beta_*^T \left(\hat{\Sigma} + \lambda I_p \right)^{-2} \hat{\Sigma} \beta_*}_{\text{pristranost}} + \underbrace{\frac{\sigma^2}{n} \operatorname{tr}\left[\hat{\Sigma}^2 \left(\hat{\Sigma} + \lambda I \right)^{-2} \right]}_{\text{varijanca}}. \qquad (3.17)$$

Dokaz Imamo [Sami!]

$$\hat\beta^r_\lambda = \frac{1}{n}(\hat\Sigma + \lambda I_p)^{-1} X^T X \beta_* + \frac{1}{n}(\hat\Sigma + \lambda I_p)^{-1} X^T \varepsilon . \quad (*)$$

$Y = X\beta_* + \varepsilon$

Iz Prop. 3.1, $\mathbb{E}[L^{in}(f_{\hat\beta^r_\lambda})] - \sigma^2 = B + V$, gdje

- $B = \| \mathbb{E}[\hat\beta^r_\lambda] - \beta_* \|^2_{\hat\Sigma}$

$\hookrightarrow$ $\mathbb{E}[\hat\beta^r_\lambda] = \frac{1}{n}(\hat\Sigma + \lambda I_p)^{-1} X^T X \beta_* = (\hat\Sigma + \lambda I_p)^{-1}[\hat\Sigma \beta_* \pm \lambda I_p \beta_*]$

$\uparrow$ $\mathbb{E}[\varepsilon] = 0$

$$= \beta_* - \lambda(\hat\Sigma + \lambda I_p)^{-1}\beta_* \quad (**)$$

$$\Rightarrow B = \|\lambda(\hat\Sigma + \lambda I_p)^{-1}\beta_*\|^2_{\hat\Sigma}$$

$$= \lambda^2 \beta_*^T (\hat\Sigma + \lambda I_p)^{-1} \underbrace{\hat\Sigma (\hat\Sigma + \lambda I_p)^{-1}}_{\overset{PZ}{=} (\hat\Sigma + \lambda I_p)^{-1}\hat\Sigma} \beta_*$$

$\uparrow$ $\hat\Sigma^T = \hat\Sigma$

$$= \lambda^2 \beta_*^T (\hat\Sigma + \lambda I_p)^{-2} \hat\Sigma \beta_* .$$

- $V = \mathbb{E}[\|\hat\beta^r_\lambda - \mathbb{E}[\hat\beta^r_\lambda]\|^2_{\hat\Sigma}]$

$$\overset{(*),(**)}{=} \mathbb{E}\left[\left\|\frac{1}{n}(\hat\Sigma + \lambda I_p)^{-1} X^T \varepsilon\right\|^2_{\hat\Sigma}\right]$$

$$= \mathbb{E}\left[tr\left(\frac{1}{n^2} \varepsilon^T X^T (\hat\Sigma + \lambda I_p)^{-1} \hat\Sigma (\hat\Sigma + \lambda I_p)^{-1} X^T \varepsilon\right)\right]$$

$$= \frac{\sigma^2}{n} tr\left(\hat\Sigma (\hat\Sigma + \lambda I_p)^{-1} \hat\Sigma (\hat\Sigma + \lambda I_p)^{-1}\right)$$

$\mathbb{E}[\varepsilon\varepsilon^T] = \sigma^2 I_n$ — kao gore

$$= \frac{\sigma^2}{n} tr\left(\hat\Sigma^2 (\hat\Sigma + \lambda I_p)^{-2}\right)$$

$\boxtimes$

Nap. 1 (i) Pokažat ćemo da je

$$\operatorname{tr}\left(\hat{\Sigma}^2(\hat{\Sigma}+\lambda I)^{-2}\right) = \sum_{j=1}^{p} \frac{\lambda_j^2}{(\lambda_j+\lambda)^2}$$

gdje su $\lambda_1, \dots, \lambda_p \geq 0$ svoj. vrijednosti od $\hat{\Sigma}$.

$\Rightarrow$ $\forall \lambda > 0$

varijanca $< \frac{d^2}{n} \cdot p$, te

varijanca $\searrow 0$, za $\lambda \nearrow +\infty$.

[vidi (3.11)]

(ii) pristranost $= \beta_\alpha^T \left(\frac{\hat{\Sigma}}{\lambda} + I_p\right)^{-2} \hat{\Sigma} \beta_\alpha \nearrow \beta_\alpha^T \hat{\Sigma} \beta_\alpha = \|\beta_\alpha\|_{\hat{\Sigma}}^2$,
$= \|\beta_\alpha - 0\|_{\hat{\Sigma}}^2$

$(\neq 0!)$

kada $\lambda \nearrow +\infty$. ∎

(iii) Može se pokazati da postoji $\lambda_\alpha > 0$ t.d.

$$\mathbb{E}\left[L^{in}\left(t_{\hat{\beta}^{LS}}\right)\right] \geq \mathbb{E}\left[L^{in}\left(t_{\hat{\beta}^{r}_{\lambda_\alpha}}\right)\right].$$

↪ u praksi, λ određujemo CV metodom.

[λ_α ne možemo eksplicitno izračunati jer ovisi o nepoznatim veličinama]

(3.3.1) Kako ridge smanjuje procjene → nekorelirani slučaj:

↳ Pretp. da je $\widehat{Cov}(\mathbb{X}_j, \mathbb{X}_k) = \frac{1}{n}\mathbb{X}_j^T \mathbb{X}_k = 0, \ \forall j \neq k$

Zbog $\bar{\mathbb{X}}_j = 0, \ \forall j$, $\hat{\Sigma} = \frac{1}{n}\mathbb{X}^T\mathbb{X} \in \mathbb{R}^{p\times p}$ = uzoračka kovarijacijska matrica od $X^{(1)}, \dots, X^{(n)}$

te

$$\hat{\Sigma} = \text{diag}\left(\widehat{Var}(\mathbb{X}_1), \dots, \widehat{Var}(\mathbb{X}_p)\right) =: \text{diag}(s_1, \dots, s_p)$$

$$\widehat{Var}(\mathbb{X}_1) = \frac{\|\mathbb{X}_1\|_2^2}{n}$$

Sada imamo [isto kao kod regresije y samo na $\mathbb{X}_j$!]

- $$\hat{\beta}_j^{ls} = \frac{\mathbb{X}_j^T y}{\|\mathbb{X}_j\|_2^2}, \ j = 1, \dots, p \qquad (3.18)$$

  ↑ $\hat{y}^{ls} = \text{Proj}_{\mathbb{X}_1,\dots,\mathbb{X}_p}(y)$

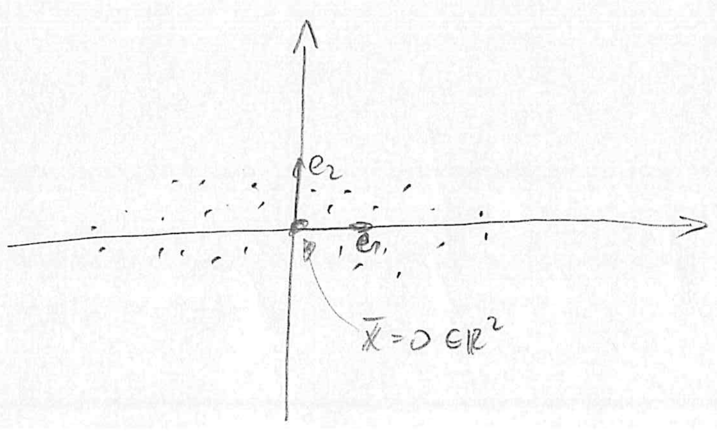
- $$\hat{\beta}_j^{r} = \frac{\|\mathbb{X}_j\|_2^2}{\|\mathbb{X}_j\|_2^2 + n\lambda} \cdot \hat{\beta}_j^{ls} = \boxed{\frac{s_j}{s_j + \lambda} \cdot \hat{\beta}^{ls}}, \ j = 1, \dots, p$$

  ↑ (3.16) DZ

$\frac{s_j}{s_j+\lambda} < 1$ za $\lambda > 0$, te

$$\frac{s_j}{s_j + \lambda} \searrow 0 \text{ kako } s_j = \widehat{Var}(\mathbb{X}_j) \searrow 0$$

npr. $p = 2$, $Var(X_1) \gg Var(X_2)$, n nije velik



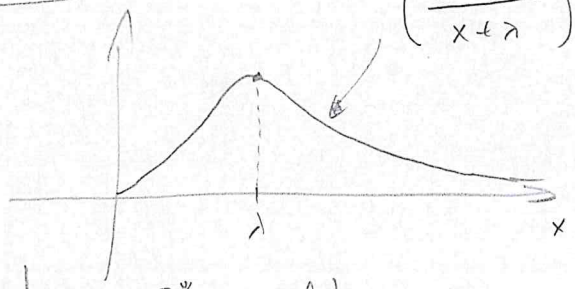
• Intuitivno, ako prilagođavamo lin. plohu (Y-os je treća koordinata), $Var(\hat{\beta}^{ls})$ bit će velika je procjena nagiba u smjeru e_2 jako ovisi o T

• Ridge smanjuje te varijance tako da smanjuje ("skrinka") procjene, i to više u smjeru e_2.

→ dobivamo pristranost čija veličina ovisi o tome je li stvarni nagib u smjeru e_2 (tj. β_2^*), velik li ne!

• Iz (3.17) dobivamo da je

$$bias = \underset{\text{Dz}}{\uparrow} \sum_{j=1}^{p} \left(\frac{\sqrt{s_j} \cdot \lambda}{s_j + \lambda} \right)^2 \cdot (\beta_j^*)^2$$



• dakle veliku pristranost ako β_j^* velik te s_j nije previdiku (niti premalen)

?

→ ridge_simulacijski_primjer.pdf

(3.3.2) Kako ridge smanjuje koeficijente? - opčenito slučaj - [vidi Gl. Strang.]

SVD od $X \in \mathbb{R}^{n \times p}$ je

$$\underset{n\times p}{X} = \underset{n\times p}{U} \cdot \underset{p\times p}{D} \cdot \underset{p\times p}{V^T}, \qquad \text{(SVD)}$$

pri čemu

- $U^T U = I_p$
- $V V^T = V^T V = I_p$

} stupci $u_1, \dots, u_p \in \mathbb{R}^{n\times 1}$ od U i $v_1, \dots, v_p \in \mathbb{R}^{p\times 1}$ su ortonormirani.

- $D = \operatorname{diag}(d_1, \dots, d_p)$ za $d_1 \geq d_2 \geq \dots \geq d_p \geq 0$

["singularne vrijednosti"]

$$\hat{\Sigma} = \frac{1}{n} X^T X = V \frac{D^2}{n} V^T \qquad \text{(SVD)}$$

tj. $\lambda_j := \frac{d_j^2}{n}$, $j = 1, \dots, p$ su svojstv. vrijednosti od $\hat{\Sigma}$, a

v_j, $j = 1, \dots, p$ odgovarajući svojstv. vektori.

tzv. smjerovi glavnih komponenti (od X)

↳ pretp. da $\operatorname{rang}(X) = p$ (te $n > p$) $\Rightarrow$ $d_p > 0$.

Nadalje, uvedimo

$$Z_j := X v_j = \begin{bmatrix} v_j^T x^{(1)} \\ \vdots \\ v_j^T x^{(n)} \end{bmatrix} \in \mathbb{R}^{n \times 1}, \quad j = 1, \dots, p \qquad (3.15)$$

zovemo glavne komponente (od X).

($\text{Proj}_{v_j}(x) = (v_j^T \cdot x) \cdot v_j$, za $x \in \mathbb{R}^p$)

Nađimo 1) Vrijedi, $\forall j = 1, \dots, p$,

$$v_j = \underset{\substack{v \in \mathbb{R}^p, \\ \|v\|_2 = 1}}{\text{argmax}} \widehat{\text{Var}}(Xv),$$

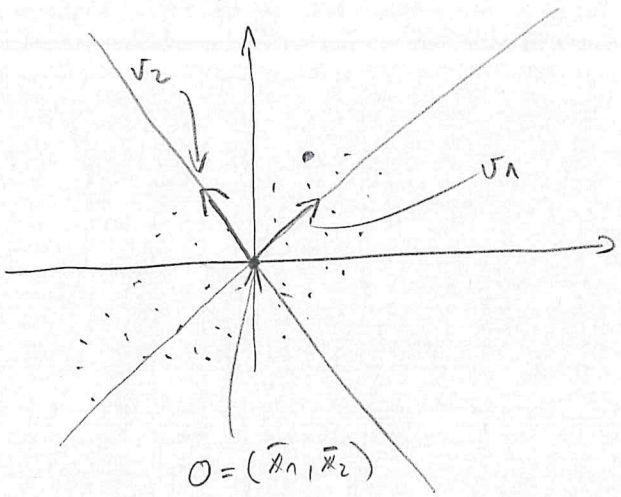
↖ uzoračka varijanca od $v^T x^{(i)}$, $i = 1, \dots, n$

uz uvjet

$$v_j^T v_\ell = 0, \quad \forall \ell = 1, \dots, j-1.$$

[bez dokaza → Primijenjena statistika]

npr. $p = 2$



[v_1 je smjer v koji maksimizira varijancu projekcija od $x^{(1)}, \dots, x^{(n)}$ na v, v_2 je smjer $v \perp v_1$ koji maksimizira

———— " ———— , itd.]

Neka je

$$Z := XV = \left[\begin{array}{ccc} | & & | \\ z_1 & \cdots & z_p \\ | & & | \end{array} \right] \in \mathbb{R}^{n \times p}, \quad (3.20)$$

zapis podataka u <u>bazi $v_1, \dots, v_p$</u>.

(SVD) $\Rightarrow$ $Z = U \cdot D$, tj.

$$\boxed{z_j = d_j \cdot u_j, \; j = 1, \dots, p} \quad (3.21)$$

Specijalno, $\left[\bar{z}_j = (\underbrace{\bar{x}}_{0})^T \cdot v_j = 0, \forall j \right]$

$$\|z_j\|_2^2 = z_j^T z_j = d_j^2, \forall j, \quad (3.22)$$

- $\boxed{\widehat{Var}(z_j) = \frac{d_j^2}{n} = \underline{\underline{\lambda_j}}, \forall j}$

- $\boxed{\widehat{Cov}(z_j, z_h) = \frac{1}{n} z_j^T z_h = 0}, \forall j \neq h \quad (3.23)$

Ridge i glavne komponente

$X = UDV^T = ZV^T$

$\forall \beta \in \mathbb{R}^p$,

$$\|y - X\beta\|_2^2 + \lambda n \|\beta\|_2^2 = \boxed{\|y - Z\cdot\alpha\| + \lambda n \|\alpha\|_2^2)}$$

gdje je $\boxed{\alpha = V^T\beta}$ ($V^2V = I \Rightarrow \|\alpha\|_2 = \|\beta\|_2$).

$\alpha \overset{1-1}{\Longleftrightarrow} \beta$

Iz $\frac{1}{n} Z^T Z = \text{diag}(\lambda_1, \dots, \lambda_p)$,

(3.22), (3.23)

[rekurzivnim izrazi]

$\Rightarrow$

(3.18)

$$\boxed{\hat{\alpha}_j^r = \frac{\lambda_j}{\lambda_j + \lambda} \cdot \hat{\alpha}_j^{ls}}, \quad j = 1, \dots, p$$

$(\hat{\beta}^r = V \cdot \hat{\alpha}^r)$

Općenito, ridge smanjuje ls procjene, ali nekim projekcijama podataka na $v_1, \dots, v_p$, te je smanjuje više za $\hat{\alpha}_j^{ls}$ za koje je $\widehat{Var}(z_j) = \lambda_j$ manje.

Nap. 3.2] $df(\text{ridge}) = ?$

Imamo $\forall \lambda \geq 0$,

$$\hat{y}_\lambda^r = X\hat{\beta}_\lambda^r = Z\cdot\hat{\alpha}_\lambda^r \overset{(3.16)}{=} Z\left(Z^TZ + n\lambda I_p\right)^{-1} Z^T y$$

($Z^TZ = \text{diag}(\lambda_1, \dots, \lambda_p)$)

$$= U \cdot \text{diag}\left(\frac{\lambda_1}{\lambda_1+\lambda}, \dots, \frac{\lambda_p}{\lambda_p+\lambda}\right) U^T \cdot y$$

U: $n \times p$, U^T: $p \times n$

$Z = UD$

$=: S_\lambda$

$$\Rightarrow df(\text{ridge}_\lambda) = tr(S_\lambda) = \boxed{\sum_{j=1}^{p} \frac{\lambda_j}{\lambda_j + \lambda}} \qquad (3.25)$$

$= df(L^S)$

↳ $df(\text{ridge}_\lambda) < p$, za $\lambda > 0$, te

$df(\text{ridge}_\lambda) \searrow 0$, za $\lambda \to +\infty$.

[ideja: projekcijom na p parametara]

(DZ) Pokažite da je $tr\left(\hat{\Sigma}^2 (\hat{\Sigma} + \lambda I)^{-2}\right) = \sum_{j=1}^{p} \frac{\lambda_j^2}{(\lambda_j + \lambda)^2}$, $\forall \lambda \geq 0$.
(vidi Prop. 3.7).

3.3.2 Multikolinearnost

Uz model (3.5),

$\hat{\Sigma}$ = kovarijacijska matrica od $(X^{(1)}, \dots, X^{(n)}) \in \mathbb{R}^{p \times p}$

$\bar{X}_j = 0, \forall j$

$$Var(\hat{\beta}_j^{LS}) = \frac{\sigma^2}{n} \left(\hat{\Sigma}^{-1}\right)_{jj}.$$

$\Rightarrow$ ako je $\hat{\Sigma} = diag(s_1, \dots, s_p)$, $Var(\hat{\beta}_j^{LS}) = \frac{\sigma^2}{n} \cdot (\lambda_j)^{-1}$.

$\widehat{Var}(X_1)$

Općenito, vrijedi [varijanca reziduala]

$$Var(\hat{\beta}_j^{LS}) = \frac{\sigma^2}{n} \boxed{\widehat{Var}\left(X_j - \hat{X}_{j|-j}\right)^{-1}} \qquad (3.26)$$

gdje je $\boxed{\hat{X}_{j|-j} := Proj_{span(X_k; k \neq j)}(X_j)}$.

$\Rightarrow$ ako $\boxed{X_j \approx \sum_{k \in I} a_k X_k}$ za neke $I \subseteq \{1, \dots, p\} \setminus \{j\}$ te $a_k \in \mathbb{R}$,

$Var(\hat{\beta}_j^{LS})$ će biti veliko. — "problem multikolinearnosti"

S druge strane, ako $p \ll n$, testna greška neće biti velika (ako je lin. model točan, nemamo).

mpr.: Pretp. da $X_1 \approx X_2$ (dakle, $\rho(X_1, X_2) \approx 1$) te vrijedi:

$$Y = \underbrace{X_1 + X_2}_{\approx c X_1 + (2-c) X_2, \ \forall c \in \mathbb{R}!} + \varepsilon \qquad (\text{dakle, } \beta_0 = (1,1))$$

Intuitivno
$\Rightarrow$ $RSS((c, 2-c)) \approx RSS((1,1))$, $\forall c \in \mathbb{R}$, a očekujemo $RSS(\beta)$ $\beta \in \mathbb{R}^2$ jako ovisi o X i y, tj. T!

↳ ridge regresija penalizira više vektore $(c, 2-c)$ za $|c| \gg 0$, pa $Var(\hat{\beta}_j^r) \overset{(<)}{<} Var(\hat{\beta}_j^{LS})$. ∎

→ ridge_kolinearnost.R

Kako prepoznati multikolinearnost

(1°) pogledati korelacije među kovarijatama

(2°) Variance Inflation Factor:

$$VIF(\beta_j) := \frac{(\hat{\Sigma}^{-1})_{jj}}{\lambda_j^{-1}} \qquad (3.27)$$

$$= \frac{\text{varijanca od } \hat{\beta}_j^{LS}}{\text{---''--- u slučaju } \forall j^T x_k = 0, \ \forall k \neq j}$$

↳ gledamo je li $VIF(\beta_j) \geq 5$ ili 10. (heuristika)

(3°) Ako λ_k = svojstv. vrijedn. od $\hat{\Sigma} \approx 0$, [k-ta po veličini]

$$\Rightarrow 0 \approx \lambda_k = \widehat{Var}\underbrace{(z_k)}_{= X v_k} = \widehat{Var}\left(\sum_{j=1}^p X_j \cdot v_{jk}\right)$$

$$\Rightarrow \sum_{j=1}^p X_j v_{jk} \approx \text{const!}$$

∎