Paul
Fayard.

Theoretical questions

OLS

- On a $\tilde{\beta} = Cy = Hy + Dy$ et $\beta^* = Hy$

Donc $Var(\tilde{\beta}) = Var(Cy)$

$= C\,Var(y)\,C^T$

$= \sigma^2 CC^T$

$= \sigma^2 \left((X^TX)^{-1}X^T + D\right)\left(X(X^TX)^{-1} + D^T\right)$

$= \sigma^2 (X^TX)^{-1} + \sigma^2 (X^TX)^{-1}(DX)^T + \sigma^2 DX(X^TX)^{-1} + DD^T$

$= \sigma^2 (X^TX)^{-1} + \sigma^2 DD^T$ car $DX = 0$

car $\tilde{\beta}$ est non biaisée

$= Var(\beta^*) + \underbrace{\sigma^2 DD^T}_{\geq 0}$ car DD^T est semi-def positive

$E(\tilde{\beta}) = E[Cy]$

$= (I_d + DX)\beta$ donc non biaisé $\Leftrightarrow DX = 0$.

Donc l'OLS est l'estimateur qui a la plus petite variance.

Ridge Regression

- On a $\beta^*_{ridge} = (X_c^TX_c + \lambda I_d)^{-1} X_c^T y_c$

Donc $E[\beta^*_{ridge}] = (X_c^TX_c + \lambda I_d)^{-1} X_c^T \beta \neq \beta$

et β^*_{ridge} est biaisé.

- $\beta^*_{ridge} = (X_c^T X_c + \lambda I_p)^{-1} X_c^T y_c$

$=$

on utilise la décomposition SVP de X_c

$X_c = U D V^T$ où U, V sont orthogonales

D est diagonale. $= \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \\ & & \ddots \end{pmatrix}$

$$\beta^*_{ridge} = (V D^2 V^T + \lambda I_p)^{-1} V D U^T y_c$$

$$= \left(V \begin{pmatrix} \sigma_1^2+\lambda & & & \\ & \ddots & & \\ & & \sigma_r^2+\lambda & \\ & & & \lambda \ddots \lambda \end{pmatrix} V^T \right) V D U^T y_c$$

$$= V \begin{pmatrix} \frac{1}{\sigma_1^2+\lambda} & & & \\ & \ddots & & \\ & & \frac{1}{\sigma_r^2+\lambda} & \\ & & & \frac{1}{\lambda} \ddots \frac{1}{\lambda} \end{pmatrix} \begin{pmatrix} \sigma_1 & & & \\ & \ddots & & \\ & & \sigma_r & \\ & & & 0 \ddots 0 \end{pmatrix} U^T y_c$$

$$= V \begin{pmatrix} \frac{\sigma_1}{\sigma_1^2+\lambda} & & & \\ & \ddots & & \\ & & \frac{\sigma_r}{\sigma_r^2+\lambda} & \\ & & & 0 \ddots 0 \end{pmatrix} U^T y_c$$

Utiliser cette décomposition est utile pour vérifier qu'il n'y ait pas de problème de colinéarité

- On a

$$(X_c^T X_c + \lambda I_d)\, \beta^*_{ridge} = X_c^T y_c = X_c^T X_c \beta + X_c^T \varepsilon$$

Donc $(X_c^T X_c + \lambda_d)\, Var(\beta^*_{ridge}) (X_c^T X_c + \lambda I_d)^T = \sigma^2 X_c^T X_c$

et $Var(\beta^*_{ridge}) = \sigma^2 (X_c^T X_c + \lambda I_d)^{-1} X_c^T X_c (X_c^T X_c + \lambda I_d)^{-1}$

On note $X_c^T X_c = P D P^T$

Alors $Var(\beta^*_{ridge}) = \sigma^2 P^T (D + \lambda I_d)^{-1} D (D + \lambda I_d)^{-1} P$

$$= \sigma^2 P^T \begin{pmatrix} \frac{\sigma_A}{(\sigma_A+\lambda)^2} & & \\ & \ddots & \\ & & \frac{\sigma_r}{(\sigma_r+\lambda)^2} \\ & & & 0 \\ & & & & \ddots \\ & & & & & 0 \end{pmatrix} P$$

$$\leq \sigma^2 P^T \begin{pmatrix} \frac{1}{\sigma_1} & & \\ & \ddots & \\ & & \frac{1}{\sigma_r} \\ & & & 0 \\ & & & & \ddots \\ & & & & & 0 \end{pmatrix} P$$

$$= Var(\beta^*_{OLS})$$

β^*_{ridge} est biaisé mais a une plus petite variance !

- $E(\beta^*_{ridge}) = (X_c^T X_c + \lambda I_d)^{-1} X_c^T \beta$

donc $\underset{\lambda}{\text{biais}}(\beta^*_{ridge}) = -\rho^T \begin{pmatrix} \frac{\lambda}{\sigma_1+\lambda} & & \\ & \ddots & \\ & & \frac{\lambda}{\sigma_r+\lambda} \\ & & & 1 & \\ & & & & \ddots \\ & & & & & 1 \end{pmatrix} \rho \beta$

$\underset{\lambda \to +\infty}{\to} -\rho^T \begin{pmatrix} 1 & & \\ & \ddots & \\ & & 1 \end{pmatrix} \rho \beta = \beta$

et quand $\beta \to +\infty$

$\text{var}(\beta^*_{ridge}) = \sigma^2 \rho^T \begin{pmatrix} \frac{\sigma_1}{(\sigma_1+\lambda)^2} & & \\ & \ddots & \\ & & 0 \\ & & & \ddots \\ & & & & 0 \end{pmatrix} \rho$

$\to 0$

C'est logique, quand on régularise beaucoup, $E(\beta^*_{ridge})$ tend vers 0, donc le biais vers β. On fait des concessions sur l'écart à β mais on obtient une variance de plus en plus petite, qui tend vers 0.

- $\beta^*_{ridge} = (X_c^T X_c + \lambda I_d)^{-1} X_c^T y_c$

Quand $X_c^T X_c = I_d$,

$\beta^*_{ridge} = ((1+\lambda) I_d)^{-1} X_c^T y_c$ $\quad$ $\beta^*_{OLS} = X_c^T y_c$

$= \frac{1}{1+\lambda} X_c^T y_c$

$\beta^*_{ridge} = \frac{\beta^*_{OLS}}{1+\lambda}$

Elastic Net.

- Si chaque classe possède sa propre matrice de covariance Σ_ℓ, on a

$$f^*(x_j) = \arg\max_{c_\ell} p_{c_\ell}(x_j)\,\pi_{c_\ell}$$

$$= \arg\min_{c_\ell} (x-\mu_\ell)^T \Sigma_\ell^{-1} (x-\mu_\ell) + \log(|\Sigma_\ell|) - 2\log(\pi_{c_\ell})$$

$$= \arg\min_{c_\ell} x^T \Sigma_\ell^{-1} x - 2x^T \Sigma_\ell^{-1} \mu_\ell + \mu_\ell^T \Sigma_\ell^{-1} \mu_\ell + \log|\Sigma_\ell| - 2\log(\pi_{c_\ell})$$

$$= \arg\min_{c_\ell} x^T a_\ell x + x^T b_\ell + c_\ell$$

où $a_\ell = \Sigma_\ell^{-1}$

$b_\ell = -2\Sigma_\ell^{-1}\mu_\ell$

$c_\ell = \mu_\ell^T \Sigma_\ell^{-1} \mu_\ell - 2\log(\pi_{c_\ell}) + \log|\Sigma_\ell|$

ce qui est bien une fonction quadratique de x.