Theorical Questions

OLS:

$(X^T X)^{-1} X^T X = Id$ donc $((X^T X)^{-1} X^T)^{-1} = X$

- $E[\tilde{\beta}] = E[Cy] = E[(H + D)y] = E[((X^T X)^{-1} X^T + D)y]$
$= E[(Id + DX)(X^T X)^{-1} X^T y]$
$= (Id + DX) E[\beta^*]$ comme β^* non biaisé
$= (Id + DX)\beta$

Si on veut que $\tilde{\beta}$ soit non biaisé, Il faut $DX = 0$.

- $var(\tilde{\beta}) = var(Cy) = C\, var(y)\, C^T = \sigma^2 CC^T$
$= \sigma^2 (X^T X)^{-1} + \sigma^2 (X^T X)^{-1} (DX)^T + \sigma^2 DX (X^T X)^{-1} + \sigma DD$

OR $DX = 0$ donc $var(\tilde{\beta}) = \sigma^2 (X^T X)^{-1} + \sigma^2 DD^T$
$= var(\beta^*) + \underbrace{\sigma^2 DD^T}$

$\sigma^2 DD^T = \sigma^2 \|D\|_2^2 > 0$ car D non-nulle

donc $var(\tilde{\beta}) > var(\beta^*)$.

On doit utiliser l'assomption que $\tilde{\beta}$ et β^* sont non biaisés c'est-à-dire que X est est de rang complet et $Y = \beta X$.
Donc l'OLS est l'estimateur de plus petite variance.

Ridge Regression:

- On a
$$\beta^*_{ridge} = (X_c^T X_c + \lambda Id)^{-1} X_c^T y_c$$
car $f : z \mapsto (y_c - x_c z)^T (y_c - x_c z) + \lambda \|z\|_2^2$
est differenciable et $f'(\beta^*_{ridge}) = 0$

avec $f'(z) = X_c^T (y_c - X_c z) + \lambda z$

$(X_c^T X_c + \lambda Id)$ inversible pour $\lambda > 0$

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

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

$$= \beta - \lambda (X_c^T X_c + \lambda I_d)^{-1} \beta \neq \beta$$

L'estimateur est biaisé

- $X_c = U D V^T$ avec $\begin{cases} U, V \text{ orthogonales} \\ D = \begin{pmatrix} d_1 & & & \\ & \ddots & & 0 \\ & 0 & d_R & \\ & & & 0 \ddots 0 \end{pmatrix} \end{cases}$

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

$$= V (D^2 + \lambda I_d)^{-1} D U^T y_c$$

$$= V \begin{pmatrix} \frac{d_1}{d_1^2+\lambda} & & & \\ & \ddots & & \\ & & \frac{d_R}{d_R^2+\lambda} & \\ & & & 0 \ddots 0 \end{pmatrix} U^T y_c$$

Cette décomposition est utile pour calculer le pseudo-inverse

- $V(\beta^*_{ridge}) = (X_c^T X_c + \lambda I_d)^{-1} X_c^T \mathrm{var}(X_c \beta - \varepsilon)$ avec $y_c = X_c \beta - \varepsilon$.

$$= (X_c^T X_c + \lambda I_d)^{-1} X_c^T \mathrm{var}(y_c) ((X^T X_c + \lambda I_d)^{-1} X_c^T)^T$$

$$= \sigma^2 (X_c^T X_c + \lambda I_d) X_c^T X_c (X^T X_c + \lambda I_d)^{-1}$$

On utilise la décomposition SVD vue à la question précédente $X_c = U D V^T$

$$\mathrm{Var}(\beta_{ridge}^*) = \sigma^2 V (D^2 + \lambda I_d)^{-1} D^2 (D + \lambda I_d)^{-1} V^T$$

$$= \sigma^2 V \begin{pmatrix} \frac{d_1}{(d_1^2+\lambda)^2} & & & \\ & \ddots & & \\ & & \frac{d_r}{(d_r^2+\lambda)^2} & \\ & & & 0 \ddots 0 \end{pmatrix} V^T$$

$$\leq \sigma^2 V \begin{pmatrix} \frac{1}{d_1} & & & \\ & \ddots & & \\ & & \frac{1}{d_r} & \\ & & & 0 \ddots 0 \end{pmatrix} V^T = \mathrm{var}(\beta^*_{OLS})$$

$$= \sigma^2 X_c^T X_c$$

- $\text{Biais}(\beta^*_{ridge}) = \lambda (X_c^T X_c + \lambda Id)^{-1} \beta$

$$= V \begin{pmatrix} \frac{\lambda}{d_1^2+\lambda} & & & & \\ & \ddots & & & \\ & & \frac{\lambda}{d_r^2+\lambda} & & \\ & & & 1 & \\ & & & & \ddots \; 1 \end{pmatrix} V^T \beta$$

$$\xrightarrow[\lambda \to +\infty]{} V \begin{pmatrix} 1 & & & \\ & \ddots 1 & & \\ & & 1 & \\ & & & \ddots \; 1 \end{pmatrix} V^T \beta = \beta$$

$$\text{var}(\beta^*_{ridge}) = \sigma^2 V \begin{pmatrix} \frac{d_1}{(d_1^2+\lambda)^2} & & & & \\ & \ddots & & & \\ & & \frac{d_r}{(d_r^2+\lambda)^2} & & \\ & & & 0 & \\ & & & & \ddots \; 0 \end{pmatrix} V^T \xrightarrow[\lambda \to +\infty]{} 0$$

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

$= \frac{1}{(1+\lambda)} X_c^T y_c$ avec $X_c^T y_c = \beta^*_{OLS}$

$= \frac{\beta^*_{OLS}}{1+\lambda}$ car $((1+\lambda) Id)^{-1} = \frac{1}{1+\lambda} Id$.

Elastic Net

Avec $X_c^T X_c = Id$, $\beta^*_{elNet} = \underset{\beta}{\arg\min} (y_c - X_c\beta)^T (y_c - X_c\beta) + \lambda_2 \|\beta\|_2^2 + \lambda_1 \|\beta\|_1$

$= \underset{\beta}{\arg\min} - y_c^T X_c \beta - \beta^T X_c^T y_c + \beta^T\beta + \lambda_2 \|\beta\|_2^2 + \lambda_1 \|\beta\|_1$

$= \underset{\beta}{\arg\min} - 2\beta^T \beta^*_{OLS} + \beta^T\beta + \lambda_2 \|\beta\|_2^2 + \lambda_1 \|\beta\|_1$

avec $\beta^*_{OLS} = X_c^T y_c$,

on a donc $\beta^*_{elNet} = \underset{\beta}{\arg\min} - 2\beta^T \beta^*_{OLS} + (\lambda_2 + 1) \|\beta\|_2^2 + \lambda_1 \|\beta\|_1$

$= \underset{\beta}{\arg\min} - 2\beta^T \beta^*_{OLS} + (\lambda_2 + 1) \|\beta\|_2^2 + \begin{cases} \lambda_1 \beta & \text{if } \beta \geq 0 \\ -\lambda_1 \beta & \text{else} \end{cases}$

On pose $f : \beta \mapsto -2\beta^T \beta^*_{OLS} + (\lambda_2 + 1)\|\beta\|_2^2 + \begin{cases} \lambda_1 \beta & \text{if } \beta \geq 0 \\ -\lambda_1 \beta & \text{if } \beta < 0 \end{cases}$

f est dérivable sur $\mathbb{R}^+$ et $\mathbb{R}^-$.

$f'(\beta_{elNet}) = 0 \iff 0 = -2\beta^*_{OLS} + 2(\lambda_2 + 1)\beta_{elNet} + \begin{cases} \lambda_1 & \text{if } \beta_{elNet} \geq 0 \\ -\lambda_1 & \text{if } \beta_{elNet} < 0 \end{cases}$

Donc $\beta_{ElNet} = \dfrac{\beta^*_{OLS} \pm \frac{\lambda_1}{2}}{\lambda_2 + 1}$

LDA:

- Supposons que chaque classe possède sa propre matrice de covariance Σ_K.

$$f^*(x_j) = \underset{C_K}{\arg\max}\ P_{C_K}(x_j)\,\pi_{C_K}$$

$$= \underset{C_K}{\arg\min}\ (x-\mu_K)^T \Sigma_K^{-1}(x-\mu_K) + \log(|\Sigma_K|) - 2\log(\pi_{C_K})$$

$$= \underset{C_K}{\arg\min}\ x^T\Sigma_K^{-1}x - \mu_K^T\Sigma_K^{-1}x - x^T\Sigma_K^{-1}\mu_K + \mu_K^T\Sigma_K^{-1}\mu_K + \log|\Sigma_K| - 2\log(\pi_{C_K})$$

$$= \underset{C_K}{\arg\min}\ x^T C_K x - x^T b_K + a_K$$

avec
$$\begin{cases} a_K = \mu_K^T \Sigma_K^{-1}\mu_K + \log|\Sigma_K| - 2\log(\pi_{C_K}) \\ b_K = -2\Sigma_K^{-1}\mu_K \\ C_K = \Sigma_K^{-1} \end{cases}$$

- On obtient bien une solution quadratique de x, à cause du terme en $x^T C_K x$.