# TP3 : Intro Supervised Machine Learning

### Theorical Questions

Maëliss de Beaumont

### OLS

On se place sous le modèle fixé suivant :
$Y = X\beta + \epsilon$ avec $\epsilon \sim \mathcal{N}(0, \sigma^2)$

On a d'une part :

$$\begin{aligned} \beta^* &= (X^T X)^{-1} X^T (X\beta + \epsilon) \\ &= \beta + (X^T X)^{-1} X^T \epsilon \end{aligned}$$

D'autre part :

$$\begin{aligned} \mathbb{E}(\tilde{\beta}) &= \mathbb{E}((H + D)Y) \\ &= \mathbb{E}(\beta^*) + \mathbb{E}(DY) \\ &= \mathbb{E}(\beta^*) + DX\beta \\ \mathbb{E}(\tilde{\beta}) &= (I_d + DX)\beta \end{aligned}$$

Donc, $\tilde{\beta}$ est un estimateur non biaisé si $DX = 0$.
Puis :

$$\begin{aligned} \mathbb{V}(\beta^*) &= \mathbb{V}(Hy) \\ &= H\mathbb{V}(y)H^T \\ &= \sigma^2 H H^T \end{aligned}$$

$$\begin{aligned}
\mathbb{V}(\tilde{\beta}) &= \mathbb{V}(Cy) \\
&= \mathbb{V}(Cy) \\
&= C\mathbb{V}(y)C^T \\
&= \sigma^2 CC^T \\
&= \sigma^2 (H+C)(H+C)^T \\
&= \sigma^2 (HH^T + HD^T + DH^T + DD^T) \\
&= \mathbb{V}(\beta^*) + (HD^T + DH^T)\sigma^2 + \sigma^2 DD^T
\end{aligned}$$

Or, on a $DX = 0$ car $\tilde{\beta}$ est non biaisé ; donc $X^T D^T = 0$ et donc $HD^T = 0$ (car $H = (X^T X)^{-1} X^T$).
Donc : $\mathbb{V}(\tilde{\beta}) = \mathbb{V}(\beta^*) + \sigma^2 DD^T$
Or $D$ n'est pas la matrice nulle, donc $DD^T$ est positive et donc, on a bien :
$\mathbb{V}(\beta^*) < \mathbb{V}(\tilde{\beta})$

**Ridge Regression**

1.
On pose : $f(\beta) = ||Y_c - X_c\beta||_2^2 + \lambda||\beta||_2^2$
On dérive : $f'(\beta) = 2X_c^T X_c\beta - 2X_c^T Y_c + 2\lambda\beta = 0$
Alors : $\beta^*_{ridge} = (X_c^T X_c + \lambda I_d)^{-1} X_c^T Y_c$

$$\begin{aligned}
\mathbb{E}(\beta^*_{ridge}) &= \mathbb{E}((X_c^T X_c + \lambda I_d)^{-1} X_c^T Y_c) \\
&= ((X_c^T X_c + \lambda I_d)^{-1} X_c^T)\mathbb{E}(Y_c) \\
&= (X_c^T X_c + \lambda I_d)^{-1} X_c^T X_c \beta
\end{aligned}$$

Donc on a un estimateur non biaisé si $\lambda = 0$ c'est à dire avec un estimateur OLS. Donc, dans le cas du ridge, l'estimateur est biaisé.

2.

On pose $X_c = UDV^T$ avec la décomposition SVD. Avec $UU^T = I_d$ et $VV^T = I_d$ et $D$ diagonale.
Alors :

$$\begin{aligned}
\beta^*_{ridge} &= ((UDV)^T UDV + \lambda I_d)^{-1} (UDV)^T Y_c \\
&= (VD^2V^T + \lambda I_d)^{-1} VDU^T Y_c \\
&= V(D^2 + \lambda I_d)^{-1} DU^T Y_c
\end{aligned}$$

C'est donc utile d'utiliser une telle décomposition car elle permet d'éviter d'inverser une matrice, mais simplement les coefficients diagonaux de la matrice $D^2 + \lambda I_d$ qui est diagonale.

3.

On rappelle que :

$$\begin{aligned}
\mathbb{V}(\beta^*_{OLS}) &= \mathbb{V}((X^TX)^{-1}X^T(X\beta + \epsilon)) \\
&= (X^TX)^{-1}X^T\mathbb{V}(X\beta + \epsilon)X((X^TX)^{-1})^T \\
&= (X^TX)^{-1}X^TX\sigma^2(X^TX)^{-1} \\
&= \sigma^2(X^TX)^{-1}
\end{aligned}$$

En utilisant la décomposition SVD :

$$\begin{aligned}
\mathbb{V}(\beta^*_{\text{ridge}}) &= \mathbb{V}((X^TX + \lambda I)^{-1}X^Ty) \\
&= (X^TX + \lambda I)^{-1}X^T\mathbb{V}(y)((X^TX + \lambda I)^{-1}X^T)^T \\
&= (X^TX + \lambda I)^{-1}X^T\mathbb{V}(\varepsilon)X(X^TX + \lambda I)^{-1} \\
&= \sigma^2(X^TX + \lambda I)^{-1}X^TX(X^TX + \lambda I)^{-1} \\
&= \sigma^2((UDV^T)^TUDV^T + \lambda I)^{-1}(UDV^T)^TUDV^T((UDV^T)^TUDV^T + \lambda I)^{-1} \\
&= \sigma^2((VDU^T)UDV^T + \lambda I)^{-1}(VDU^T)UDV^T((VDU^T)UDV^T + \lambda I)^{-1} \\
&= \sigma^2V(D^2 + \lambda I)^{-1}D^2(D^2 + \lambda I)^{-1}V^T \\
&= \sum_{i=1}^{rg(X)} \frac{d_i^2\sigma^2}{(d_i^2 + \lambda)^2}v_iv_i^T
\end{aligned}$$

avec $d_i$ les éléments diagonaux de $D$ et $v_i$ les vecteurs de $V$ associés.
Or :

$$\mathbb{V}(\beta^*_{\text{OLS}}) = \sum_{i=1}^{\text{rank}(X)} \frac{\sigma^2}{d_i^2}v_iv_i^T$$

Donc, pour $\lambda > 0$, on a : $\mathbb{V}(\beta^*_{\text{ridge}}) \leq \mathbb{V}(\beta^*_{\text{OLS}})$

4.

$$\begin{aligned}
\mathbb{E}(\beta^*_{ridge}) &= (X^TX+\lambda I_d)^{-1}X^TX\beta \\
&= ((UDV^T)^TUDV^T+\lambda I)^{-1}(UDV^T)^TUDV^T\beta \\
&= ((VDU^T)UDV^T+\lambda I)^{-1}(VDU^T)UDV^T\beta \\
&= (VD^2V^T+\lambda I)^{-1}VD^2V^T\beta \\
&= V(D^2+\lambda I)^{-1}D^2V^T\beta \\
&= \sum_{i=1}^{rg(X)} \frac{d_i^2}{d_i^2+\lambda} v_i v_i^T \beta
\end{aligned}$$

On a alors pour le biais ; sachant qu'on a l'expression de la variance par la question précédente :

$$\begin{aligned}
biais &= \mathbb{E}(\beta^*_{\text{ridge}}) - \beta \\
&= \sum_{i=1}^{rg(X)} \frac{d_i^2}{d_i^2+\lambda} v_i v_i^T \beta - \beta
\end{aligned}$$

Donc :

- Pour des valeurs de $\lambda$ qui augmentent, la variance diminue, mais le biais augmente.
- Pour des valeurs de $\lambda$ qui diminuent, la variance augmente, mais le biais diminue ; et quand $\lambda = 0$, on a un estimateur OLS.

5.

Quand $X_c^TX_c = I_d$, on a :

$$\begin{aligned}
\beta^*_{OLS} &= (X_c^TX_c)^{-1}X_c^TY_c \\
&= X_c^TY_c
\end{aligned}$$

et

$$\begin{aligned}
\beta^*_{ridge} &= (X_c^TX_c+\lambda I_d)^{-1}X_c^TY_c \\
&= (\lambda+1)^{-1}X_c^TY_c \\
&= (\lambda+1)^{-1}\beta^*_{OLS}
\end{aligned}$$

d'où : $\beta^*_{ridge} = \frac{\beta^*_{OLS}}{\lambda+1}$ quand $X_c^TX_c = I_d$

**Elastic Net**

On suppose : $X_c^TX_c = I_d$ donc : $\beta^*_{OLS} = X_c^TY_c$

On pose $f$ la fonction à minimiser :
$f(\beta) = (y - x\beta)^T(y - x\beta) + \lambda_2||\beta||_2^2 + \lambda_1||\beta||_1$

alors :
$\partial f(\beta^*) = -2x^T(y - x\beta^*) + 2\lambda_2\beta^* \pm \lambda_1 = 0$
$0 = -2x^T y + 2\beta^* + 2\lambda_2\beta^* \pm \lambda_1$
Ainsi :
$\beta^*_{ElNet} = \frac{x_c^T y_c \pm \frac{\lambda_1}{2}}{1+\lambda_2}$
En remplaçant dans l'expression, on a bien :
$\beta^*_{ElNet} = \frac{\beta^*_{OLS} \pm \frac{\lambda_1}{2}}{1+\lambda_2}$