IMA205 - TP1 - Theoretical Questions

Théo Rouvet

5 février 2020

J'ai parfois noté p la taille de l'espace des features au lieu de d. On peut donc considérer dans ce qui suit que $d = p$.

1 Ordinary Least Squares

1.1 Espérance de $\tilde{\beta}$

$\mathbb{E}(\tilde{\beta}) = \mathbb{E}(Cy) = \mathbb{E}((H + D)y) = \mathbb{E}(\beta^*) + \mathbb{E}(Dy) = \beta + \mathbb{E}(Dy)$

On peut considérer un **design du type fixe** $y = x\beta + \epsilon$, tel que ϵ est un bruit blanc additif (d'espérance nulle). En injectant cette expression, on obtient :

$\mathbb{E}(\tilde{\beta}) = \beta + \mathbb{E}(D(x\beta + \epsilon)) = \beta + \mathbb{E}(D)x\beta + \mathbb{E}(D\epsilon)$

On peut supposer de plus que D **est déterministe** pour pouvoir continuer le calcul. On obtient par la suite :

$\mathbb{E}(\tilde{\beta}) = \beta + Dx\beta$

Ainsi, compte-tenu des hypothèses précédentes, $\tilde{\beta}$ **est non biaisé si et seulement si** $Dx\beta = 0$.

1.2 Variance de $\tilde{\beta}$

Calculons à présent la variance de $\tilde{\beta}$ pour montrer que l'OLS est l'estimateur non biaisé qui donne la plus petite variance.

En supposant que β^* **et** Dy **sont indépendants**, on écrit :

$$\begin{aligned} Var(\tilde{\beta}) &= Var(\beta^* + Dy) \\ &= Var(\beta^*) + DVar(y)D^T \\ &= (x^Tx)^{-1}x^TVar(y)x(x^Tx)^{-1} + DVar(y)D^T \end{aligned}$$

Or, on peut écrire que $Var(y) = Var(x\beta + \epsilon) = Var(\epsilon) = \sigma^2 I_p$ en **reprenant l'hypothèse du bruit blanc** précédente.

On a ensuite $Var(\tilde{\beta}) = \sigma^2(x^Tx)^{-1}I_p + \sigma^2DD^T$ i.e. $Var(\tilde{\beta}) = Var(\beta^*) + \sigma^2DD^T$

Or, DD^T est une matrice symétrique et positive, et $\sigma^2 \geqslant 0$, **donc** $Var(\tilde{\beta}) \geqslant Var(\beta^*)$.

L'inégalité est stricte si et seulement si $DD^T \neq 0$ et $\sigma \neq 0$.

On obtient bien que **l'OLS est l'estimateur non biaisé qui donne la plus petite variance**.

2 Ridge regression

2.1 Question 1

On sait que l'estimateur Ridge est tel que $\beta^*_{\text{ridge}} = (x^T x + \lambda I_p)^{-1} x^T y$, donc par passage à l'espérance, $\mathbb{E}(\beta^*_{\text{ridge}}) = \mathbb{E}((x^T x + \lambda I_p)^{-1} x^T y) = (x^T x + \lambda I_p)^{-1} x^T x \beta$ en reprenant le design d'OLS, avec x déterministe et $y = x\beta + \epsilon$ avec ϵ centré.

Si $\lambda = 0$, on trouve une OLS et $\mathbb{E}(\beta^*_{\text{ridge}}) = \beta$.

Si $\lambda > 0$, il s'agit bien d'un Ridge et on a $(x^T x + \lambda I_p)^{-1} x^T x \neq I_n$ ce qui entraîne $\mathbb{E}(\beta^*_{\text{ridge}}) \neq \beta$ i.e. **$\beta^*_{\textbf{ridge}}$ est biaisé**.

2.2 Question 2

On résout en utilisant la décomposition SVD :

$$\begin{aligned}
\beta^*_{\text{ridge}} &= (x^T x + \lambda I_p)^{-1} x^T y \\
&= (V D U^T U D V^T + \lambda I_p)^{-1} x^T y \\
&= (V D^2 V^T + \lambda V V^T)^{-1} x^T y \\
&= V (D^2 + \lambda I_p)^{-1} V^T V D U^T y \\
\beta^*_{\text{ridge}} &= V (D^2 + \lambda I_p)^{-1} D U^T y
\end{aligned}$$

Comme $D = diag(d_1, d_2, ..., d_p)$, on a $D^2 + \lambda I_p = diag(d_1^2 + \lambda, d_2^2 + \lambda, ..., d_p^2 + \lambda)$. Cette matrice est clairement inversible comme les d_i sont positifs et $\lambda > 0$. On a donc $(D^2 + \lambda I_p)^{-1} = diag(\frac{1}{d_1^2+\lambda}, \frac{1}{d_2^2+\lambda}, ..., \frac{1}{d_p^2+\lambda})$

Cette décomposition est utile quand on souhaite **accélerer le calcul de l'inversion de la matrice $x^T x + \lambda I_p$**.

2.3 Question 3

$Var(\beta^*_{\text{ridge}}) = Var((x^T x + \lambda I_p)^{-1} x^T y) = Var((x^T x + \lambda I_p)^{-1} x^T (x\beta + \epsilon))$

Comme supposé au-dessus, x est déterministe donc $Var(x) = 0$. Par ailleurs, en supposant que $Var(\epsilon) = \sigma^2 I_n$, on obtient :

$$
\begin{aligned}
Var(\beta^*_{\text{ridge}}) &= Var((x^Tx + \lambda I_p)^{-1}x^T\epsilon)\\
Var(\beta^*_{\text{ridge}}) &= \sigma^2(x^Tx + \lambda I_p)^{-1}x^Tx(x^Tx + \lambda I_p)^{-1}\\
Var(\beta^*_{\text{ridge}}) &= \sigma^2V(D^2 + \lambda I_p)^{-1}V^TVDU^TUDV^TV(D^2 + \lambda I_p)^{-1}V^T\\
Var(\beta^*_{\text{ridge}}) &= \sigma^2V(D^2 + \lambda I_p)^{-1}D^2(D^2 + \lambda I_p)^{-1}V^T\\
Var(\beta^*_{\text{ridge}}) &= \sigma^2\sum_{k=1}^{p}\frac{d_k^2}{(d_k^2 + \lambda)^2}v_kv_k^T
\end{aligned}
$$

Or, $Var(\beta^*_{\text{OLS}}) = \sigma^2VD^{-2}V^T = \sigma^2\sum_{k=1}^{p}\frac{1}{d_k^2}v_kv_k^T$

Comme $\frac{d_k^2}{(d_k^2+\lambda)^2} < \frac{1}{d_k^2}$, on a $Var(\beta^*_{\text{ridge}}) < Var(\beta^*_{\text{OLS}})$

2.4 Question 4

Plus λ augmente, plus $Var(\beta^*_{\text{ridge}})$ diminue. En effet, $Var(\beta^*_{\text{ridge}}) = \sigma^2\sum_{k=1}^{p}\frac{d_k^2}{(d_k^2+\lambda)^2}v_kv_k^T$ **est une fonction décroissante de λ.**

Regardons à présent le biais :

$$
\begin{aligned}
\mathbb{E}[\beta^*_{Ridge}] &= \mathbb{E}[(x^Tx + \lambda I_p)^{-1}x^Ty]\\
&= \mathbb{E}[(x^Tx + \lambda I_p)^{-1}x^T(x\beta + \epsilon)]\\
&= \mathbb{E}[(x^Tx + \lambda I_p)^{-1}(x^Tx\beta + x^T\epsilon + \lambda\beta - \lambda\beta)]\\
&= \beta - \lambda(x^Tx + \lambda I_p)^{-1}\beta\\
\Rightarrow biais_\beta(\beta^*_{Ridge}) &= -\lambda(x^Tx + \lambda I_p)^{-1}\beta
\end{aligned}
$$

En diagonalisant la matrice $x^Tx + \lambda I_p = P(\Delta + \lambda I_p)P^T$, on obtient que $biais_\beta(\beta^*_{Ridge}) = -P^T(\frac{\Delta}{\lambda} + I_p)^{-1}P\beta$, donc lorsque λ augmente, le biais tend vers une certaine constante de la forme $-P^T\Gamma_\infty P\beta$.

2.5 Question 5

En supposant que $x^Tx = I_p$, on obtient :

$$
\begin{aligned}
\beta^*_{\text{ridge}} &= (x^Tx + \lambda I_p)^{-1}x^Ty\\
&= ((\lambda + 1)I_p)^{-1}x^Ty\\
&= \frac{1}{\lambda + 1}x^Ty\\
\beta^*_{\text{OLS}} &= (x^Tx)^{-1}x^Ty = x^Ty
\end{aligned}
$$

On a donc bien $\beta^*_{\text{ridge}} = \frac{\beta^*_{\text{OLS}}}{\lambda+1}$ lorsque $x^Tx = I_p$.

3 Elastic Net

Par définition, on a :

$$
\begin{aligned}
\beta^*_{ElNet} &= \arg\min_\beta ||y_c - x_c\beta||_2^2 + \lambda_2||\beta||_2^2 + \lambda_1||\beta||_1 \\
&= \arg\min_\beta y_c^T y_c + \beta^T x_c^T x_c \beta - 2\beta^T x_c^T y_c + \lambda_2||\beta||_2^2 + \lambda_1 \sum_{j=1}^{d} |\beta_i| \\
&= \arg\min_\beta \beta^T\beta - 2\beta^T (x_c^T x_c)^{-1} x_c^T y_c + \lambda_2||\beta||_2^2 + \lambda_1 \sum_{j=1}^{d} |\beta_i| \\
&= \arg\min_\beta ||\beta||_2^2 - 2\beta^T \beta^*_{OLS} + \lambda_2||\beta||_2^2 + \lambda_1 \sum_{j=1}^{d} |\beta_i| \\
&= \arg\min_\beta - 2\beta^T \beta^*_{OLS} + (\lambda_2+1)||\beta||_2^2 + \lambda_1 \sum_{j=1}^{d} |\beta_i| \\
\beta^*_{ElNet} &= \arg\min_\beta \sum_{j=1}^{d} (-2\beta_j\beta^*_{OLS,j} + (\lambda_2+1)\beta_j^2 + \lambda_1|\beta_j|)
\end{aligned}
$$

On peut chercher l'argmin pour chaque terme. On a pour tout j :

$$
\begin{aligned}
\beta^*_{ElNet,j} &= \arg\min_{\beta_j} - 2\beta_j\beta^*_{OLS,j} + (\lambda_2+1)\beta_j^2 + \lambda_1|\beta_j| \\
&= \arg\min_{\beta_j} \begin{cases} -2\beta_j\beta^*_{OLS,j} + (\lambda_2+1)\beta_j^2 + \lambda_1\beta_j & \text{si } \beta \geq 0 \\ 2\beta_j\beta^*_{OLS,j} + (\lambda_2+1)\beta_j^2 - \lambda_1\beta_j & \text{si } \beta < 0 \end{cases} \\
&= \begin{cases} \frac{2\beta^*_{OLS,j} - \lambda_1}{2(\lambda_2+1)} & \text{si } \beta \geq 0 \\ \frac{2\beta^*_{OLS,j} + \lambda_1}{2(\lambda_2+1)} & \text{si } \beta < 0 \end{cases} \\
\beta^*_{ElNet,j} &= \frac{\beta^*_{\text{OLS}} \pm \frac{\lambda_1}{2}}{1-\lambda_2}
\end{aligned}
$$

4 LDA

Si la covariance de chaque classe est différente, on ne peut plus simplifier par $\frac{1}{|\Sigma|^{1/2}}$ pour toutes les classes. Ainsi, toute la LDA ne peut plus s'appliquer car f^* n'est plus une "**linear** discriminant function". Dans ce cas, il faut utiliser la **QDA**.

$$
\begin{aligned}
f^*(x_j) &= argmax_{C_k} P_{C_k}(x_j)\pi_{C_k} \\
&= argmin_{C_k} - 2log(P_{C_k}(x_j)\pi_{C_k} \\
&\propto argmin_{C_k}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) + log(|\Sigma_k|) - 2log(\pi_{C_k}) \\
&= argmin_{C_k} x^T\Sigma_k^{-1}x - x^T \cdot 2\Sigma_k^{-1}\mu_k + (\mu_k^T\Sigma_k^{-1}\mu_k - 2log(\pi_{C_k}) + log(|\Sigma_k|)) \\
&= argmin_{C_k} x^T c_k x + x^T b_k + a_k
\end{aligned}
$$

On voit qu'ici, f^* est une fonction discriminante quadratique (et donc non linéaire). En effet, le terme $x^T c_k x = x^T\Sigma_k^{-1}x$ intervient, i.e. la matrice de covariance de chaque classe intervient.