HMMA237

Retour sur les moindres carrés et la régression Ridge

Cours: SALMON Joseph *Scribes: ABOUQATEB Mouad et KANDOUCI Walid*

1 Ridge (Régression)

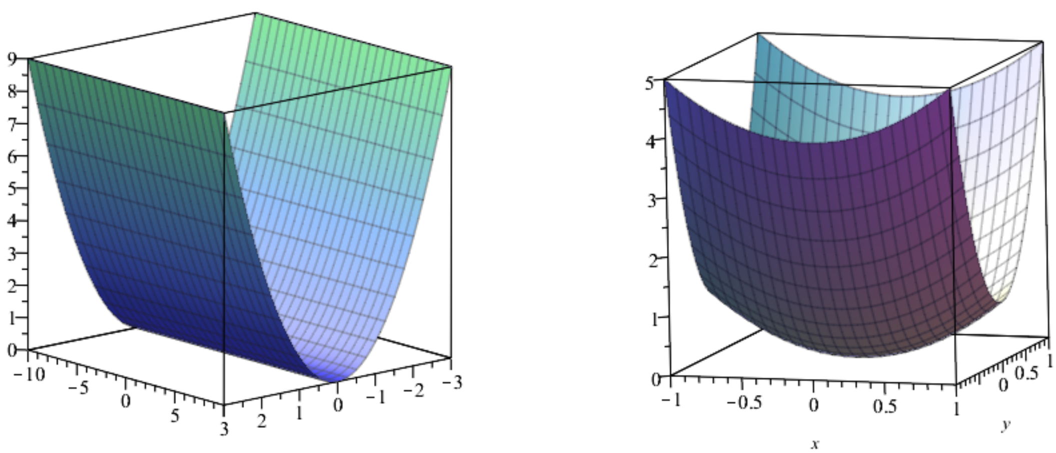
Considérons le modèle linéaire :

$$Y = X\beta^* + \varepsilon$$

Où Y est la matrice observations définies sur $\mathbb{R}^n$, X la matrice des covariables définie sur $\mathbb{R}^{n\times p}$, B^* dans $\mathbb{R}^p$ est l'opérateur linéaire et ε le bruit.
Considérons la matrice (X^TX), c'est la matrice de covariance des covariables, nommée **matrice de Gram**.

Le but est d'exprimer Y en fonction des variables explicatives X sous forme de l'opérateur linéaire β dans les paramètres inconnus du modèle.

L'approche naturelle pour traiter ce problème est d'estimer β en utilisant la méthode des moindres carrées. (LS)



(a) $(x_1, x_2) \to x_1^2$ de forme matricielle donnée par $f_1(x) = x^TAx = x^T\begin{pmatrix}1 & 0\\ 0 & 0\end{pmatrix}x$ avec A non-inversible

(b) $(x_1, x_2) \to x_1^2 + x_2^2$ de forme matricielle $f_2(x) = x^TBx = x^T = x^T\begin{pmatrix}1+\frac{1}{2} & 0\\ 0 & \frac{1}{2}\end{pmatrix}x$ avec B inversible.

FIGURE 1 – Moindres carrés : cas non unique, avant et après régularisation

On cherche β minimisant l'erreur :

$$\hat{\beta}^{LS} \in Argmin\|y - x\beta\|^2 \tag{1}$$

Pour cela il suffit de trouver β vérifiant l'équation du gradient au minimum, qui s'écrit :

$$X^T(X\hat{\beta}^{LS} - y) \tag{2}$$

Notre $\hat{\beta}$ vérifie donc : $(X^T X)\hat{\beta}^{LS} = X^T y$
Remarquons que notre **matrice de Gram** $(X^T X)$ est symétrique donc diagonalisable dans une base orthonormée, et semi-définie, (i.e., $v^T X^T X v = \|Xv\|^2 \geqslant 0$), pas forcément définie, ce qui pourra nous empêcher dans un premier temps d'utiliser l'estimateur de la méthode de moindre carrées,qui correspond à $(X^T X)^{-1} X^T y$ néanmoins toutes ses valeurs propres sont donc positives, ou nulles.
Il existe donc (U) matrice orthogonale (son inverse est sa transposé), et $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_p \geqslant 0$ dans $\mathbb{R}^+$, tel que :

$$(X^T X) = U^T \operatorname{diag}(\lambda_1, \ldots, \lambda_p)U \tag{3}$$

C'est la décomposition spectrale de $X^T X$, pour résoudre notre problème d'inversibilité, on en ajoute une pénalité sous forme λId avec $\lambda > 0$, notre nouvelle matrice $\operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ sera définie positive et par suite inversible, d'inverse $U^T.\operatorname{diag}(\frac{1}{\lambda_1+\lambda}, \ldots, \frac{1}{\lambda_p+\lambda}).U$.
Nous définissons ainsi notre nouvel estimateur de l'opérateur linéaire : $\hat{B}^{ridge} = (x^T x + \lambda Id_p)^{-1} x^T y$

Remark. *Remarquons que notre estimateur Ridge $\hat{\beta}^{ridge(\lambda)}$ tends vers l'estimateur des moindres carrés $\hat{\beta}^{LS}$ quant $\lambda \to 0$. En effet pour un modèle de plein rang ($X^T X$ inversible) l'estimateur des moindres carrées correspond à un $\hat{\beta}^{ridge(0)}$*

Theorem. $\hat{\beta}^{ridge(\lambda)} \in \underset{\lambda_1,\ldots,\lambda_T}{Argmin}||y - x\beta||^2 + \lambda||\beta||^2$

Démonstration. On définit $g :\to \|y - xa\|^2 + \lambda\|\beta\|^2$, $f :\to \|y - xa\|^2$ et $\nabla g(\hat{\beta}^\lambda) = 0$
Notons $\hat{\beta}^\lambda$ l'argument minimal de g, et montrons que $\hat{\beta}^\lambda = \hat{\beta}^{ridge}$.
Notre $\hat{\beta}^\lambda$ vérifie l'équation du gradient au minimum de g (condition du premier ordre) est :

$$\nabla g(\hat{\beta}^\lambda) = \nabla f(\hat{\beta}^\lambda) + 2\lambda\beta = 0$$
$$\text{(avec)} \quad \nabla f(\hat{\beta}^\lambda) = 2x^T x\beta - 2(x^T y)^T \beta$$

En sommant nos deux équations, on obtient :

$$x^T x\beta\lambda - x^T y + \lambda\beta^\lambda = 0$$

Et puisque $(X^T X + \lambda Id_p)$ est inversible comme on l'a éjà vue ($1^{re} partie$), on déduit le résultat :

$$\hat{\beta}^{ridge(\lambda)} = (x^T x + \lambda Id_p)^{-1} x^T y$$

□

Remark. *On peut constater à travers notre théorème que le rôle du coefficient λ, sera de privilégier la minimisation des erreurs au profit du biais (régularisation) et vice versa. En effet un λ de plus en plus grand impliquera des prédictions de moins en moins sensibles à la variation des variables, mathématiquement* *parlant, un opérateur linéaire de plus en plus négligeable* $\hat{\beta} \underset{\lambda \to \infty}{\to} 0$ *. Respectivement un* λ *négligeable nous donnera un estimateur proche des moindres carrées* ***(Remarque 1)****.*

La question qui se pose est quelle sera le choix optimale de λ*.*

$$Sp(x^T x + \lambda Id_p)^{-1} = (\frac{1}{\lambda_1 + \lambda} \cdots \frac{1}{\lambda_p + \lambda}) \tag{4}$$

2 Validation croisé ("Hold out")

Cette méthode consiste à décomposer notre matrice d'observations de taille n en deux sous-matrice, n_1 observations d'"apprentissage", et n_2 observations de "validation", (tel que $n_1 + n_2 = n$)
$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}; x = \begin{pmatrix} x^A \\ \vdots \\ x^V \end{pmatrix}; y = \begin{pmatrix} y^A \\ \vdots \\ y^V \end{pmatrix}$
et d'appliquer notre régression Ridge, pour différent choix de λ_i, c sous $\lambda = (\lambda_1, \ldots, \lambda_T)$

Pour des raisons pratiques, on considère une grille géométrique
Le choix du λ optimal est :

$$\hat{\lambda}^{cv} = \underset{\lambda_1, \ldots, \lambda_T}{\arg\min} \|y_v - X_v \hat{\beta}^{\lambda}\|^2 \tag{5}$$

Et pour des raisons computationnelles, on procède par la méthode "Leave on out", qui comme son nom l'indique, consiste a valider le modèle sur la n^{e} observation après avoir appris sur n-1 observations et l'on répète cette opération n fois. $n_1 = n - 1$ $n_2 = 1$

(Paramètre par défaut courant,$n_1 \approx 80\% n; n_2 \approx \%20n$)

Proposition. $(x^T + \lambda Id_p)$ *est symétrique définie positif*

Démonstration. Notre matrice est symétrique comme somme de deux matrices symétriques.
Soit β non-nul dans $\mathbb{R}^P$, d'après **(1)**, et en conservant les mêmes notations :

$$X^T X + \lambda Id_p = U^T \operatorname{diag}(\lambda_1 + \lambda, \ldots, \lambda_n + \lambda)\beta^T (X^T X + \lambda Id_p)U + \lambda U^T U = \operatorname{diag}(\lambda + \lambda_1, \ldots, \lambda + \lambda_p)$$

Notre matrice est donc diagonalisable dans une base orthonormée avec des valeurs propres **strictement** positives, par conséquent, elle est définie positive.

Le risque(estimation) quadratique moyen est donnée par $\mathbb{E}\|\hat{\beta} - \beta^*\|^2$ et le risque (prédiction) quadratique moyen avec $\hat{\beta}$ vecteur de prédiction est donnée par $\mathbb{E}\|X\hat{\beta} - X\beta^*\|^2$

(On rappel que : $\hat{\beta}^{LS} = (X^TX)^{-1}X^Ty$)

$$\begin{aligned}
R &= \mathbb{E}\|(X^TX)^{-1}X^Ty - \beta^*\|^2 \qquad \text{(Définie et de plein rang)}\\
&= \mathbb{E}\|(X^TX)^{-1}X^T(X\beta^* + \sigma\varepsilon) - \beta^*\|^2\\
&= \mathbb{E}\|(X^TX)^{-1}X^TX\beta^* + \sigma(X^TX)^{-1}X^T\varepsilon - \beta^*\|^2\\
&= \mathbb{E}\|\Sigma(X^TX)^{-1}X^T\varepsilon\|^2\\
&= \sigma^2\mathbb{E}\|(X^TX)^{-1}X^T\varepsilon\|^2\\
&= \sigma^2\mathbb{E}[\varepsilon^TX(X^TX)^{-1}(X^TX)^{-1}X^T\varepsilon]\\
&= \sigma^2 tr(\varepsilon^TX(X^TX)^{-1}(X^TX)^{-1}X^T\varepsilon)\\
R &= \sigma^2\mathbb{E}(tr[\varepsilon^TX(X^TX)^{-2}X^T\varepsilon])
\end{aligned}$$

Or on a pour tout vecteur v de taille finie, $v^Tx = \|v\|^2 \in \mathbb{R}$ et pour tout réel a, $a = tr(a)$ en particulier $\|v\|^2 = a$,d'où :

$$R = \sigma^2 tr(\varepsilon^TX(X^TX)^{-1}(X^TX)^{-1}X^T\varepsilon) \tag{6}$$

Donc :

$$R = \sigma^2\mathbb{E}(tr[\varepsilon^TX(X^TX)^{-2}X^T\varepsilon])$$

□

Remark. *$tr(AB) = tr(BA)$ même si A et B ne sont pas de taille identique. Par exemple : $tr(xx^T) = tr(x^Tx)$, mais il faut que AB et BA existe.*

$$\begin{aligned}
R &= \sigma^2 tr[\mathbb{E}((X^TX)^{-1}\varepsilon^T\varepsilon)]\\
&= \sigma^2 tr(X(X^TX)^{-2}X^T\mathbb{E}(\varepsilon\varepsilon^T))\\
&= \sigma tr(X(X^TX)^{-2}X^T)\\
&= \sigma^2 tr((X^TX)^{-1})\\
R &= \sigma^2\sum_{j=1}^{P}\frac{1}{\lambda_j}
\end{aligned}$$

(Si $Sp(X^TX) = (\lambda_1, \ldots, \lambda_p)$)

Remark. *Le risque est proportionnel à σ^2, le risque se détérioré si la structure de corrélation devient singulière.*

3 Pour aller plus loin sur ce thème

Parmi les lectures intéressantes on notera : [Del15] [1]

Références

[Del15] B. Delyon. Régression, 2015. 4

1. https://perso.univ-rennes1.fr/bernard.delyon/regression.pdf