HMMA237: Séries temporelles avancées

Rappels: moindre carrés, complément de Schur et applications

Cours: Joseph Salmon *Scribes: Luana Timofte et Anas El Benna*

1 Introduction

1.1 L'estimateur des moindres carrés et Ridge

Soit $X = [X_1, X_2, \dots, X_p] \in \mathbb{R}^{n \times p}$ une matrice de taille $n \times p$, avec pour chaque $j \in [\![1, p]\!]$, la colonne X_j est un vecteur de taille n, avec n le nombre d'observations et p le nombre de variables qualitatives (ou features,covariables).

Soit $y \in \mathbb{R}^n$ un vecteur de taille n qui est le vecteur que l'on observe.

On souhaite prédire ou représenter y à partir des variables explicatives $X_1, X_2, \dots, X_p$. Le but est donc d'estimer les coefficients β du modèle linéaire de nos observations.

Definition. *La méthode des* ***moindres carrées ordinaires (OLS : ordinary least squares)*** *permet d'estimer les coefficients β par :*

$$\hat{\beta}^{(OLS)} \in \underset{\beta \in \mathbb{R}^p}{\arg\min} \underbrace{\frac{1}{2}\|y - X\beta\|^2}_{f(\beta)} \quad . \tag{1}$$

avec $\hat{\beta}^{(OLS)} \in \mathbb{R}^p$ et $f(\beta)$ la fonction qu'on va optimiser, par la suite. ***Attention!*** *$\hat{\beta}^{(OLS)}$ n'est pas forcément unique ! L'estimateur $\hat{\beta}^{(OLS)}$ est donc extrémum de $f(\beta)$. Avec :*

$$f(\beta) = \frac{1}{2} y^\top y + \frac{1}{2} \beta^\top X^\top X \beta - \underbrace{< y, X\beta >}_{= y^\top X\beta = \beta^\top X^\top y} \tag{2}$$

Représentation graphique:

Notations:

$\rightarrow i$: observations, $i \in [\![1, n]\!] \rightarrow j$: covariables, $j \in [\![1, p]\!]$

Remark. *Soit on traite les "constantes" (intercept en anglais ou ordonnée à l'origine, en français) de la manière suivante :* ***(1)****:*

$$\min_{\beta, \beta_0} \|y - X\beta - \beta_0 \mathbb{1}\|^2 \tag{3}$$

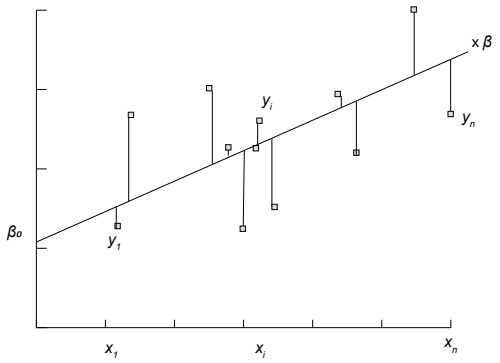

Figure 1: Exemple de graphique de régression linéaire

avec $\beta_0 \in \mathbb{R}$ *Soit : (2): On ajoute une variable explicative :* $X_0 = \underbrace{[1, \ldots, 1]^\top}_{\mathbb{1}} \in \mathbb{R}^n$

$$\tilde{X} = [X_0, X_1, \ldots, X_p] \tag{4}$$

et on fait :

$$\min_{\tilde{\beta} \in \mathbb{R}^{p+1}} \|y - \tilde{X}\tilde{\beta}\|^2 \tag{5}$$

1.2 Phénomènes de normalisation et standardisation

Par la suite, nous allons parler de deux phénomènes qui apparaissent.

1.2.1 Normalisation

On essaie de normaliser les variables explicatives $X = [X_1, \ldots, X_p]$

Exemple:

→ On met tout à la même échelle: kg, cm ...

→ Soit $X_j = \frac{X_j}{\|X_j\|_2}$ (avec $\|X_j\|_2$ la norme euclidienne classique). Pour tout j : $\|X_j\|_2 = 1$.

→ On peut aussi l'écrire de façon matricielle : $X \leftarrow XD$ (transformation par une matrice) et $D = diag(\frac{1}{\|X_1\|}, \ldots, \frac{1}{\|X_p\|})$.

1.2.2 Standardisation

Soit $X_j \leftarrow \frac{X_j - \overline{X_j}}{\|X_j - \overline{X_j}\|}$ qu'on va normaliser et ensuite on va centrer les variables.

Suite à cela :

$\rightarrow \overline{X_j} = 0$ (c'est à dire que c'est centré) $\rightarrow \|X_j\| = 1$ (c'est normalisé, donc c'est à dire que c'est sans unité ou adimensionnel)

Conclusion : nous avons obtenu quelque chose de centré et réduit.

Modèle Gaussien:

$$y = X \underbrace{\beta^*}_{\text{"vrai" paramètre}} + \underbrace{\varepsilon}_{\text{bruit, taille } n}$$

$$\varepsilon \sim \mathcal{N}(0, \sigma^2 Id_n), \quad \varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^\top \ . \tag{6}$$

Theorem. *Le maximum de vraisemblance dans ce modèle est un estimateur des moindres carrés.*

1.3 Formulation exacte/résolution

$\rightarrow f$ est C^∞

$\rightarrow f$ est convexe (admis) (pour cela on calcule la dérivée deuxième et on regarde si la hessienne $X^\top X$ est positive).

Hypothèse: "rang plein" $\rightarrow$ "plein rang colonne".

i.e., $\text{vect}(X_1, \dots, X_p) = \mathbb{R}^p$ et $\text{Im}(X) = \mathbb{R}^p$.

Theorem. ***Théorème du rang:***

$$\underbrace{\dim(\text{Ker}(X))}_{=0} + \underbrace{\dim(\text{Im}(X))}_{=p} = p \iff \text{Ker}(X) = \{0\} \ . \tag{7}$$

Et c'est aussi équivalent à $X^\top X$, qui est inversible.

Remark. *Si n (i.e., le nombre d'observations) est inférieur strictement à p (i.e., ce n'est pas de plein rang !), comme $rg(X) \leqslant \min(n, p)$, donc $rg(X) \leqslant n < p$. Ce contexte "$n < p$" s'appelle "grande dimension".*

Exemple: $\rightarrow$ dans le domaine de la médecine ou de la génétique, nous pouvons avoir $n = 100$ patients et $p = 50000$ gènes, environ.

Remark. *Si $Ker(X) \neq \{0\}$ (i.e., $\exists \underbrace{\beta_0}_{\neq 0} \in \mathbb{R}^p$ tel que $X\beta_0 = 0$). Alors*

$$\beta_1 \in \hat{\beta}^{OLS} + Ker(X)$$

Donc, par exemple :

$$\beta_1 = \hat{\beta}^{OLS} + \beta_0 \in \arg\min \|y - X\beta\|^2 \tag{8}$$

$$y - X\beta_1 = y - X(\hat{\beta}^{OLS} + \beta_0) \tag{9}$$

$$\|y - X\beta_1\| = \|y - X\hat{\beta}^{OLS}\| \tag{10}$$

$$\tag{11}$$

(et $X\beta_0 = 0$).

Donc pour ce type de modèle, il suffit de trouver une solution puis faire des translations $\mathrm{Ker}(X)$ *pour trouver d'autres solutions.*

Rappel: Une fonction convexe est utile dans l'optimisation puisque on a qu'un seul minimum GLOBAL et non LOCAL.

Voici une représentation graphique:

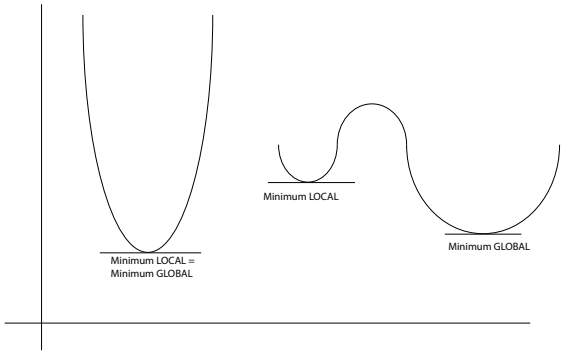

Figure 2: Minimum local vs global

C'est donc un critère nécessaire pour que $\hat{\beta}$ soit un minimum local de f. On a une pente égale à 0 (tangente), c'est-à-dire en dimension 1 et donc que la dérivée s'annule.

$$\nabla f(\hat{\beta}) = 0$$

Rappel:

$$f(\beta) = \frac{y^\top y}{2} + \frac{\beta^\top X^\top X\beta}{2} - \underbrace{< y; X\beta >}_{< X^\top y; \beta >} \tag{12}$$

$$\nabla f(\hat{\beta}) = -X^\top y + X^\top X\beta \tag{13}$$

$$\underbrace{\nabla(\beta+h)^\top X^\top X(\beta+h)}_{\in\mathbb{R}} = \beta^\top X^\top X\beta + h^\top X^\top X\beta + \beta^\top X^\top Xh + \underbrace{h^\top X^\top Xh}_{O(h^2)} \tag{14}$$

$$= \beta^\top X^\top X\beta + 2\underbrace{\beta^\top X^\top Xh}_{\langle X^\top X\beta, h\rangle} \tag{15}$$

$$= X^\top(X\beta - y) \ . \tag{16}$$

Donc

$$\nabla f(\hat{\beta}) = -X^\top y + X^\top X\beta$$

$$= X^\top (X\beta - y) = 0$$

$$[X_1, \dots, X_p]^\top (X\hat{\beta} - y) = 0$$

$$\forall j \in [\![1, p]\!], \quad \langle X_j, X\hat{\beta} - y\rangle \iff \langle X_j, \underbrace{y - X\hat{\beta}}_{\text{résidus}}\rangle = 0$$

Ainsi

$$\nabla f(\hat{\beta}) = 0 \iff (X^\top X)\hat{\beta} = X^\top y \ .$$

C'est un système linéaire. On peut le résoudre en utilisant la méthode du pivot de Gauss.

1.3.1 Pivot de Gauss

But du pivot de Gauss: résoudre

$$Ax = b, x \in \mathbb{R}^p$$

Remark.

$$X^\top X \in \mathbb{R}^{p\times p}$$

$A = X^\top X$ inversible, avec X de plein rang.

$$\begin{cases} \overbrace{a_{11}}^{\neq 0,\ \text{pivot}} \ x_1 + a_{12}x_2 + \cdots + a_{1p}x_p & = & b_1 (L_1) \\ \vdots & & \\ a_{p1}x_1 + a_{p2} + \cdots + a_{pp}x_p & = & b_p (L_p) \end{cases}$$

Itération :

$$L_p \leftarrow L_p - L_1 \left(\frac{a_{p1}}{a_{11}}\right)$$

On prend à chaque fois un pivot et on répète l'opération jusqu'à ce qu'on arrive à un système triangulaire, où on n'aura qu'une seule équation : $a_{pp}x_p = b_p$ et ensuite on remonte et on calcule les autres.

1.3.2 Algorithme itératif d'optimisation

<u>"Descente de gradient"</u> → c'est un algorithme de point fixe.

β_0: initialisation; choix $\beta_0 = 0$ ou $X^\top y = \beta_0$ α: pas de descente

$$\begin{cases} \beta_0 \in \mathbb{R}^p \\ \beta_t = \beta_{t-1} - \alpha \nabla f(\beta_{t-1}), \text{ pour tout } t \geqslant 1 \end{cases} \tag{17}$$

Attention ! On a besoin que le pas (de descente) soit suffisamment petit: $\alpha < \frac{1}{L}$ pour assurer la convergence, avec L la/une constante de Lipschitz du gradient :

$$\|\nabla f(\beta) - \nabla f(\tilde{\beta})\| \leqslant L\|\beta - \tilde{\beta}\|, \forall \beta, \forall \tilde{\beta}$$
$$\iff |||\nabla^2 f||| \leqslant L$$

Theorem. *Si f est convexe et ∇f est L-Lipschitz, alors (β_t) converge vers la valeur minimum de f.*

$$\nabla f(\beta) = X^\top(X\beta - y)$$
$$\beta_{t+1} = \beta_t - \alpha X^\top X\beta + \alpha X^\top y$$
$$= \underbrace{(Id - \alpha X^\top X)\beta_t + \alpha X^\top y}_{g(\beta_t)}$$

$$\|g(\tilde{\beta}) - g(\beta)\| = \|Id - \alpha(X^\top X)(\tilde{\beta} - \beta)\|$$
$$|||Q||| = \lambda_{\max}(Q)$$

Avec Q symétrique et $\geqslant 0$.

Remark.

$$g(\beta) = \beta \iff \beta - \alpha X^\top X\beta + \alpha X^\top y = \beta (\alpha \neq O!) \tag{18}$$
$$\iff X^\top X\beta = X^\top y \tag{19}$$
$$\iff \nabla f(\beta) = 0 \tag{20}$$

Trouver un point fixe de g c'est de trouver β tel que $\nabla f(\beta) = 0$. Si $\alpha < 1$ alors $0 < \lambda_{\max}(Id - \alpha X^\top X) < 1$. (Il faut donc que la plus grande valeur propre soit comprise entre 0 et 1.) Pour $X^\top X$ symétrique : $X^\top X$ est semi-définie positive ($\geqslant 0$).

$$\forall \beta, \beta^\top X^\top X\beta \geqslant 0 \iff S_p(X^\top X) \geqslant 0 \tag{21}$$

Avec S_p le spectre de $X^\top X$ (ou les valeurs propres de la matrice $X^\top X$).

Remark.

$$\|X\beta\|^2 \geqslant 0 \iff \beta^\top X^\top X\beta \geqslant 0 \tag{22}$$

Donc la matrice X n'a que des valeurs propres positives!

Si

$$\lambda \in S_p(X^\top X) \Rightarrow X^\top X\beta = \lambda\beta$$
$$\beta - \alpha X^\top X\beta = (1 - \alpha\lambda)\beta$$
$$1 - \alpha\lambda \in S_p(Id - \alpha X^\top X)$$

Si

$$\alpha < \frac{1}{\lambda_{\max}(X^\top X)} \Rightarrow \alpha\lambda < 1, \forall \lambda \in S_p(X^\top X)$$
$$\Rightarrow 1 - \alpha\lambda > 0, \forall \lambda \in S_p(X^\top X)$$

Conséquence: Le théorème de Picard (point fixe) s'applique.