HMMA 307

Modéle linéaire avancé : introduction

Cours: Joseph Salmon *Scribes: BOUZALMAT Ibrahim et Ibrahima GUEyE*

1 Introduction et rappels

1.1 Moindres Carrés

$$y = X\beta^* + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 Id_n) \ . \tag{1}$$

- $y \in \mathbb{R}^n$: vecteur des observations,
- $X = [\mathbf{x}_1, \dots, \mathbf{x}_p] \in \mathbb{R}^{n \times p}$: Covariables("features").
- $\beta^* \in \mathbb{R}^p$: vecteurs des coefficients.
- $\varepsilon \in \mathbb{R}^n$: vecteur des bruits.

L'estimateur des moindres carrés ordinaires (MCO) (: *Ordinary Least Squares*) :

$$\hat{\beta}^{\text{OLS}} \in \operatorname*{arg\,min}_{\beta \in \mathbb{R}^n} \frac{1}{2}\|y - X\beta\|_2^2 \ , \tag{2}$$

avec

$$f(\beta) = \beta^\top \frac{X^\top X}{2}\beta + \frac{1}{2}\|y\|^2 - \langle y, X\beta \rangle \ , \tag{3}$$

où l'on note $\langle y, X\beta \rangle = y^\top X\beta = \beta^\top X^\top y = \langle \beta, X^\top y \rangle$.

<u>Vocabulaire</u> : la matrice $X^\top X$ est appelée matrice de Gram [1].

$$X^\top X = \begin{pmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_p^\top \end{pmatrix} \begin{pmatrix} \mathbf{x}_1, \dots, \mathbf{x}_p \end{pmatrix} \ ,$$

ce qui est équivalent à :

$$\left[X^\top X\right]_{j,j'} = \left[\langle \mathbf{x}_j, \mathbf{x}_{j'} \rangle\right]_{(j,j') \in [\![1,p]\!]^2} \ .$$

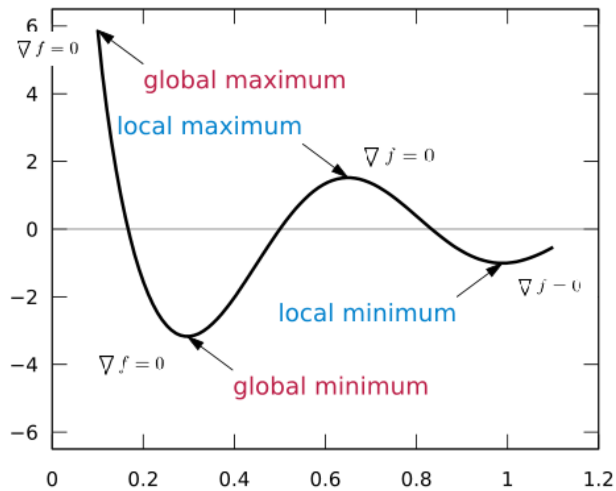
1. parfois on appelle plutôt $X^\top X/n$ cette matrice.

Remarque : Il arrive souvent de rajouter le vecteur des constantes, c'est-à-dire de prendre $\mathbf{x}_1 = \mathbb{1}_n = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \in \mathbb{R}^n$.

Conditions du premier ordre :

$$\hat{\beta}^{\text{OLS}} \in \operatorname*{arg\,min}_{\beta \in \mathbb{R}^n} \implies \nabla f(\hat{\beta}^{\text{OLS}}) = 0 \ . \tag{4}$$

La fonction f est différentiable et donc la condition $\nabla f(\hat{\beta}^{\text{OLS}}) = 0$ est nécessaire en un minimum local :



Remarque 1.1. *Pour retrouver le gradient de la fonction f : il suffit de calculer $f(\beta + h) - f(\beta)$ et ainsi on obtient :*

$$\nabla f(\beta) = X^\top X\beta - X^\top y \ . \tag{5}$$

On peut alors obtenir les conditions du premier ordre avec

$$\nabla f(\hat{\beta}^{\text{OLS}}) = 0 \implies X^\top(y - X\hat{\beta}^{\text{OLS}}) = 0 \ . \tag{6}$$

où $(y - X\hat{\beta}^{\text{OLS}})$ est le vecteur des ***résidus****. Enfin les conditions du premier ordre s'écrivent aussi sous la forme suivante , que l'on nome parfois "équations normales" :*

$$\forall j \in [\![1, p]\!], \quad \langle \mathbf{x}_j, y - X\hat{\beta}^{\text{OLS}} \rangle = 0 \ . \tag{7}$$

Existence : L'existence d'une solution pour le problème des moindres carrés $\hat{\beta}^{\text{OLS}}$ est assurée quand la matrice $X^\top X \succ 0$ (c'est-à-dire que la matrice de Gram est définie positive). C'est condition est satisfaite dès que $X^\top X$ est inversible, car une matrice de Gramm est toujours semi-définie positive.

En effet dans ce cas f est continue (car différentiable) et coercive :

$$\lim_{\|\beta\|\to+\infty} f(\beta) = +\infty \tag{8}$$

Unicité : Avec les conditions normales de premier ordre (C.N.O) on a : $X^\top X\hat{\beta}^{\mathrm{OLS}} = X^\top y$, qui donne un système linéaire.

L'unicité est garantie si $X^\top X$ est inversible, et on a :

$$X^\top X \text{ inversible } \iff \operatorname{rang}(X) = p \ . \tag{9}$$

On appelle souvent cette hypothèse l'*hypothèse de plein colonne*.

: pour cette hypothèse (de plein rang colonne), p doit être plus petit ou égal à n ($p \leqslant n$).

Sous cette hypothèse, la solution des moindre carrés est unique et vaut :

$$\hat{\beta}^{\mathrm{OLS}} = (X^\top X)^{-1} X^\top y \tag{10}$$

Remarque 1.2. *Sans cette hypothèse, une solution existe encore et peut s'écrire :*

$$\hat{\beta}^{l_2} = X^+ y \ , \tag{11}$$

où X^+est la pseudo-inverse de X.

2 Décomposition en valeurs singulières

(: *Singular Value Decomposition, SVD*)

2.1 Rappels

Définition 2.1. *Une matrice $U \in \mathbb{R}^{n\times n}$ est dite* orthogonale *si elle vérifie la propriété suivante :*

$$U^\top U = UU^\top = \mathrm{Id}_n \tag{12}$$

ou de manière équivalente :

$$\forall (i,j) \in \{1,\ldots,n\}^2, \quad \mathbf{u}_i^\top \mathbf{u}_j = \langle \mathbf{u}_i, \mathbf{u}_j \rangle = \delta_{i,j} \ . \tag{13}$$

Dans la cas d'une matrice non carrée on dit que $U \in \mathbb{R}^{m_1\times m_2}$ est orthogonale si ses colonnes le sont, ou ce qui est équivalent :

$$U^\top U = \mathrm{Id}_{m_1} \ . \tag{14}$$

Théorème 2.1. *Une matrice symétrique $A \in \mathbb{R}^{n\times n}$ est diagonalisable en base orthonormée,* i.e., *il existe $\lambda_1 \geqslant \cdots \geqslant \lambda_n$ et une matrice orthogonale $U \in \mathbb{R}^{n\times n}$ telle que :*

$$A = U \operatorname{diag}(\lambda_1,\ldots,\lambda_n) U^\top \iff AU = U \operatorname{diag}(\lambda_1,\ldots,\lambda_n) \ . \tag{15}$$

Les réels λ_i pour $i = 1,\ldots,n$ sont les valeurs propres *de A et les $\mathbf{u}_i \in \mathbb{R}^n$ sont les* vecteurs propres *associés.*

On peut interpréter la décomposition spectrale comme une décomposition en somme de matrice de rang un : si l'on écrit $U = [\mathbf{u}_1,\ldots,\mathbf{u}_n]$ cela signifie que :

$$A = \sum_{i=1}^n \lambda_i \mathbf{u}_i \mathbf{u}_i^\top, \quad \text{avec} \quad \forall i \in [\![1,n]\!], \quad A\mathbf{u}_i = \lambda_i \mathbf{u}_i \ . \tag{16}$$

2.2 La décomposition en valeurs singulières

Théorème 2.2. *Pour toute matrice $M \in \mathbb{R}^{m_1 \times m_2}$ et de rang r, il existe une matrice orthogonale $U \in \mathbb{R}^{m_1 \times r}$ et une matrice orthogonale $V \in \mathbb{R}^{m_2 \times r}$, telles que*

$$M = U \operatorname{diag}(s_1, \ldots, s_r) V^\top \ . \tag{17}$$

avec $s_1 \geqslant s_2 \geqslant \cdots \geqslant s_r \geqslant 0$ sont les valeurs singulières de M, ou encore :

$$M = \sum_{i=1}^{r} s_i \mathbf{u}_i \mathbf{v}_i^\top \ . \tag{18}$$

avec $U = [\mathbf{u}_1, \ldots, \mathbf{u}_r]$ *et* $V = [\mathbf{v}_1, \ldots, \mathbf{v}_r]$.

Remarques :

1. Les matrices sont obtenues comme suit :
   (i) les valeurs singulières sont les racines carrées des valeurs propres à la fois de $M^\top M$ et $MM^\top$.
   (ii) U est la matrice des vecteurs propres de $M^\top M$.
   (iii) V est la matrice des vecteurs propres de $MM^\top$.
2. $\sum_{i=1}^{r} s_i \mathbf{u}_i \mathbf{v}_i^\top$ est une somme de termes de rang 1 (*i.e.*, $\operatorname{rang}(\mathbf{u}_i \mathbf{v}_i^\top) = 1$).
3. On peut aussi forcer les $\mathbf{u}_1, \ldots, \mathbf{u}_r$ à être orthogonaux deux à deux (de même pour les $\mathbf{v}_i$), *i.e.*, : $U^\top U = \mathrm{Id}_r$ et $V^\top V = \mathrm{Id}_r$.
4. Les $\mathbf{u}_i$ (resp. les $\mathbf{v}_i^\top$) sont orthonormés et engendrent le même espace que celui engendré par les colonnes (resp. les lignes) de M.

$$\operatorname{vect}(M_{1,:}, \ldots, M_{m_2,:}) = \operatorname{vect}(\mathbf{u}_1, \ldots, \mathbf{u}_r) \ . \tag{19}$$

5. La SVD généralise aux matrices non carrées la décomposition spectrale.

Démonstration. La matrice $M^\top M$ est une matrice $m_2 \times m_2$ symétrique, semi définie-positive, et de rang r ($\operatorname{rang}(M^\top M) = \operatorname{rang}(M)$), donc elle admet des valeurs propres réelles positives $(\lambda_1, \ldots, \lambda_r, \ldots)$ (on suppose que au delà de r, elles sont nulles) et une base orthonormée

$$M^\top M \mathbf{v}_i = \lambda_i \mathbf{v}_i \ . \tag{20}$$

De plus :

$$\mathbf{v}_j^\top M^\top M \mathbf{v}_i = \lambda_i \mathbf{v}_j^\top \mathbf{v}_i = \lambda_i \delta_{ij} \ . \tag{21}$$

Définissons pour tout $i \in [\![1, r]\!]$:

$$s_i = \sqrt{\lambda_i} \text{ et } \mathbf{u}_i = \frac{M\mathbf{v}_i}{\sqrt{\lambda_i}} \ . \tag{22}$$

Les $(\mathbf{u}_1, \ldots, \mathbf{u}_r)$ forment une famille orthonormée. En effet :

$$\langle \mathbf{u}_i, \mathbf{u}_j \rangle = \left\langle \frac{M\mathbf{v}_i}{\sqrt{\lambda_i}}, \frac{M\mathbf{v}_j}{\sqrt{\lambda_j}} \right\rangle = \frac{1}{\sqrt{\lambda_i \lambda_j}} \langle M\mathbf{v}_i, M\mathbf{v}_j \rangle = \frac{\lambda_i}{\sqrt{\lambda_i \lambda_j}} \delta_{i,j} = \delta_{i,j} \ . \tag{23}$$

Ainsi, pour $U = [\mathbf{u}_1, \ldots, \mathbf{u}_r] \in \mathbb{R}^{m_1 \times r}$ et $V = [\mathbf{v}_1, \ldots, \mathbf{v}_r] \in \mathbb{R}^{m_2 \times r}$, on a :

$$U^\top M V = \begin{pmatrix} \frac{1}{s_1}\mathbf{v}_1^\top M^\top M \mathbf{v}_1 & 0 & \ldots & 0 \\ 0 & \frac{1}{s_2}\mathbf{v}_2^\top M^\top M \mathbf{v}_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & \frac{1}{s_r}\mathbf{v}_r^\top M^\top M \mathbf{v}_r \end{pmatrix} \tag{24}$$

$$= \begin{pmatrix} \frac{\lambda_1}{s_1}\mathbf{v}_1^\top \mathbf{v}_1 & 0 & \ldots & 0 \\ 0 & \frac{\lambda_2}{s_2}\mathbf{v}_2^\top \mathbf{v}_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & \frac{\lambda_r}{s_r}\mathbf{v}_r^\top \mathbf{v}_r \end{pmatrix} \tag{25}$$

$$= \begin{pmatrix} s_1 & 0 & \ldots & 0 \\ 0 & s_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & s_r \end{pmatrix} \tag{26}$$

$$= \mathrm{diag}(s_1, \ldots, s_r) \ . \tag{27}$$

□

Proposition 2.1. *La matrice $UU^\top \in \mathbb{R}^{m_1 \times m_1}$ est la matrice de projection orthogonale sur l'espace engendré par $\mathbf{u}_1, \ldots, \mathbf{u}_r$.*

Démonstration. Soit Π_U la projection orthogonale sur le sous espace vectoriel de $\mathbb{R}^{m_1}$ engendré par $(\mathbf{u}_1, \ldots, \mathbf{u}_r)$ et $\mathbf{y} \in \mathbb{R}^{m_1}$ Cette projection est caractérisée par les propriétés :

(i) Π_U est une combinaison linéaire des colonnes de U.

(ii) $\mathrm{Id}_{m_1} - \Pi_U$ est orthogonale aux colonnes de U.

Commençons par prouver les deux propriétés caractéristique de la projection orthogonale.
Propriété (i) : Le produit $U^\top \mathbf{y}$ est un vecteur de taille r, et notons le $(\mathbf{x}_1, \ldots, \mathbf{x}_r)$. On peut alors écrire le vecteur $UU^\top \mathbf{y}$ de la façon suivante :

$$UU^\top \mathbf{y} = \sum_{i=1}^{r} \mathbf{x}_i \mathbf{u}_i \ . \tag{28}$$

C'est-à-dire que la propriété caractéristique (i) est vérifiée.

Propriété (ii) : Soit $\mathbf{w}$ la matrice ligne de taille r dont les éléments sont les produits scalaires $\langle \mathbf{y} - UU^\top \mathbf{y}, \mathbf{u}_i \rangle$. On veut montrer que chacun de ces produits scalaires est nul, c'est-à-dire que $\mathbf{w}$ est le vecteur nul. Or

$$\mathbf{w} = U^\top \left(\mathbf{y} - UU^\top \mathbf{y}\right) = U^\top \mathbf{y} - \underbrace{U^\top U}_{\mathrm{Id}_r} U^\top \mathbf{y} = U^\top \mathbf{y} - U^\top \mathbf{y} = 0$$

□

Corollaire 2.2.1. *La matrice $UU^\top$ est la projection orthogonale sur l'espace engendré par les colonnes de M et de plus $U^\top U M = M$.*
Identiquement, on peut montrer que $MVV^\top = M$, ceci donne : $UU^\top M V V^\top = UU^\top M = M$.

Démonstration. Comme $\text{vect}(M_{:,1}, \ldots, M_{:,m_1}) = \text{vect}(\mathbf{u}_1, \ldots, \mathbf{u}_r)$ (d'après la remarque XXX), On peut déduire que $UU^\top$ est la projection orthogonale sur l'espace engendré par les colonnes de M. Ainsi $U^\top UM = M$. □

3 Pseudo-inverse, inverse de Moore-Penrose, inverse généralisée

Définition 3.1. *Si $X \in \mathbb{R}^{m_1 \times m_2}$, admet pour SVD $X = \sum_{i=1}^r s_i \mathbf{u}_i \mathbf{v}_i^\top$ avec $r = rang(X)$, alors sa* pseudo-inverse *$X^+ \in \mathbb{R}^{m_2 \times m_1}$ est définie par :*

$$X^+ = \sum_{i=1}^r \frac{1}{s_i} \mathbf{v}_i \mathbf{u}_i^\top \ . \tag{29}$$

Remarques :

1. X^+X et XX^+ existant.

2. Si $X = \sum_{i=1}^n s_i \mathbf{u}_i \mathbf{v}_i^\top \in \mathbb{R}^{n \times n}$ est inversible alors $X^+ = X^{-1}$, en effet :

$$XX^+ = \sum_{j=1}^n s_j \mathbf{u}_j \mathbf{v}_j^\top \sum_{i=1}^n \frac{1}{s_i} \mathbf{v}_i \mathbf{u}_i^\top \tag{30}$$

$$= \sum_{j=1}^n \sum_{i=1}^n s_j \frac{1}{s_i} \mathbf{u}_j \mathbf{v}_j^\top \mathbf{v}_i \mathbf{u}_i^\top \tag{31}$$

$$= \sum_{j=1}^n \sum_{i=1}^n s_j \frac{1}{s_i} \delta_{i,j} \mathbf{u}_j \mathbf{u}_i^\top \tag{32}$$

$$= \sum_{i=1}^n \mathbf{u}_i \mathbf{u}_i^\top = \text{Id}_n \tag{33}$$

3.1 SVD et numérique

Les fonctions SVD et pseudo-inverse sont disponibles dans les librairies numériques classiques, par exemple `numpy`

- SVD : `U, s, V = np.linalg.svd(X)`
  Attention dans ce cas : `X = U.dot(np.diag(s).dot(V))`. On accède aux variantes compactes ou non par l'option `full-matrices=True/False`.

— Pseudo-inverse : `Xinv = np.linalg.pinv(X)`.

Exemple :
Soit une matrice A et sa SVD obtenue avec la commande `U, s, V = np.linalg.svd(A)` :

$$A = \begin{pmatrix} 2 & 1 & 6 & 1 \\ 1 & 1 & 1 & 2 \\ 0 & 1 & 2 & 3 \end{pmatrix} \quad U = \begin{pmatrix} 0.863 & 0.505 & -0.002 \\ 0.286 & -0.485 & 0.827 \\ 0.416 & -0.714 & -0.563 \end{pmatrix} \tag{34}$$

$$S = \begin{pmatrix} 7.304 & 0 & 0 & 0 \\ 0 & 2.967 & 0 & 0 \\ 0 & 0 & 0.918 & 0 \end{pmatrix}, \quad V = \begin{pmatrix} 0.275 & 0.177 & 0.896 & -0.302 \\ 0.214 & -0.234 & 0.285 & 0.905 \\ 0.862 & 0.376 & -0.339 & 0.000 \\ 0.367 & -0.879 & -0.041 & -0.302 \end{pmatrix}. \tag{35}$$

Le pseudo inverse peut être obtenu aussi plus simplement avec la commande `np.linalg.pinv(A)`

$$A^+ = \begin{pmatrix} 0.061 & 0.788 & -0.576 \\ -0.015 & 0.303 & -0.106 \\ 0.167 & -0.333 & 0.167 \\ -0.106 & 0.121 & 0.258 \end{pmatrix} \tag{36}$$

Références

[1] SVD, Nicolas Verzelen, Joseph Salmon,INRA / Université de Montpellier.

[2] METHODES NUMERIQUES, Manfred GILLI.

[3] La Décomposition en Valeurs Singulières, Analyse numérique et Application à la Vision,Valérie Perrier, Roger Mohr,Ensimag et Laboratoire Jean Kuntzmann.

[4] Cours de Mathématiques II, Chapitre 1. Algèbre linéaire,Université de Paris X Nanterre ,U.F.R. Segmi.