HMMA 307

Modéle linéaire avancé.

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

Rappels et Introduction

1 Moindres Carrés

$$Y = X\beta^* + \epsilon, \quad \epsilon \sim \mathbb{N}(0, \sigma^2 Id_p).$$

Y : observation, $\quad Y \in \mathbb{R}^n$.
X : Covariables("features"), $\quad X \in \mathbb{R}^{n\times p}$.
$\beta^* \in \mathbb{R}^p$: vecteurs des coefficents.
$\epsilon \in \mathbb{R}^n$; bruits (d'observations).

L'estimateurs des moindres carrés (MCO) (LS) :

$$\hat{\beta}^{obs} \in argmin_{\beta\in\mathbb{R}^n} \frac{1}{2}\|y - X\beta\|_2^2$$

$$f(\beta) = \beta^T \frac{X^T X}{2}\beta + \frac{1}{2}\|y\|^2 - \langle y; X\beta\rangle, \quad avec\ \langle y; X\beta\rangle = y^T X\beta = \beta^T X^T y = \langle \beta; X^T y\rangle$$

$X^T X$ est appelée matrice de Gram.

$$X^T X = \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix} (x_1 \cdots x_p) = \begin{pmatrix} \|x_1\|^2 & & \langle X_i; X_j\rangle \\ & \ddots & \\ & & \|x_p\|^2 \end{pmatrix}$$

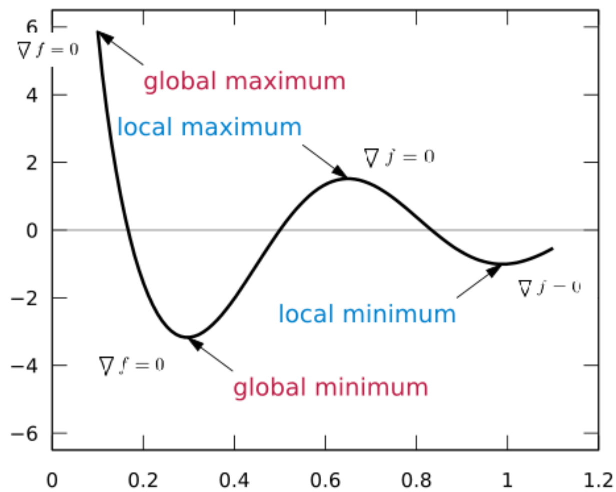
$$X^T X = [\langle X_i; X_j\rangle]_{(i,j)\in[1,p]^2}$$

Remarque : Souvent $X_1 = \mathbb{1}_n = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \in \mathbb{R}^n$

Conditions du premier ordre :

$$\hat{\beta}^{OLS} \in argmin_{\beta \in \mathbb{R}^n} \Longrightarrow \nabla f(\hat{\beta}^{obs}) = 0$$

f est c^∞ et donc $\nabla f(\hat{\beta}^{OLS}) = 0$ est nécessaire en un minimum local :



Si besoin, calculer $f(\beta + h) - f(\beta)$ et trouver le gradient :

$$\nabla f(\beta) = X^T X\beta - X^T y$$

$$\nabla f(\hat{\beta}^{OLS}) = 0 \Leftrightarrow X^T(y - X\hat{\beta}^{OLS}),$$

où $(y - X\hat{\beta}^{obs})$ sont des résidus.

$$\Leftrightarrow \forall j \in [1, p], \langle X_j, y - X\hat{\beta}^{obs} \rangle = 0$$

Remarques :

- **existence :** On a besoin que $XX^T > 0$ (Semi-defini positif).
- Si $X^T X$ est inversible alors elle reste semi défini posistive et il y a encore l'existence d'une solution. Pour s'en convaincre : f est continue et coercive.

$$f(\beta) \longrightarrow +\infty$$

$$\|\beta\| \longrightarrow +\infty$$

— **Unicité :** Avec les conditions normales de premier ordre (C.N.O) on a : $X^T X \hat{\beta}^{OLS} = X^T y$, qui donne un système linéaire.
L'unicité est garantie si $X^T X$ est inversible, et on a :

$$X^T X \; est \; inversible \Leftrightarrow rang(X) = p \; (hypothese \; de \; plein \; rang \; en \; colonne)$$

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

$$(unique) \; \hat{\beta}^{OLS} = (X^T X)^{-1} X^T y$$

Sinon : Une solution possible s'écrit :

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

où X^+est la pseudo-inverse de X.

2 La décomposition en valeurs singulières (En anglais : Singular Value Decomposition, SVD)

2.1 Rappels

1- Une matrice $U \in \mathbb{R}^{n \times n}$ est dite orthogonale si elle vérifie
$U^\top U = UU^\top = \mathrm{Id}_n$
ou $\forall (i,j) \in \{1, \ldots, n\}^2 \quad \mathbf{u}_i^\top \mathbf{u}_j = \langle \mathbf{u}_i, \mathbf{u}_j \rangle = \delta_{i,j}$.
2- Théorème spectral :
Une matrice symétrique $A \in \mathbb{R}^{n \times n}$ est diagonalisable en base orthonormée, i.e., il existe $\lambda_1 \geqslant \ldots \geqslant \lambda_n$ et une matrice orthogonale $U \in \mathbb{R}^{n \times n}$ telle que :

$$A = U \operatorname{diag}(\lambda_1, \ldots, \lambda_n) U^\top \text{ ou } AU = U \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$$

Les $\lambda_i, \quad i = 1, \ldots, n$ sont les valeurs propres de A et les $\mathbf{u}_i \in \mathbb{R}^n$ sont les vecteurs propres associés.
Remarque : Si l'on écrit $U = [\mathbf{u}_1, \ldots, \mathbf{u}_n]$ cela signifie que :

$$A = \sum_{i=1}^{n} \lambda_i \mathbf{u}_i \mathbf{u}_i^\top, \quad avec \; \forall i \in \{1, \ldots, n\}, \quad A\mathbf{u}_i = \lambda_i \mathbf{u}_i$$

2.2 La décomposition en valeurs singulières

Théorème :
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

$$M = U \operatorname{diag}(s_1, \ldots, s_r) V^T$$

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

$$M = \sum_{i=1}^{r} s_i u_i v_i^T$$

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^TM et MM^T.

(ii) U est la matrice des vecteurs propres de M^TM .

(iii) V est la matrice des vecteurs propres de MM^T.

2 - $\sum_{i=1}^{r} s_i u_i v_i^T$ somme des termes de rang 1 ($rg(u_i, v_i^T) = 1$).

3- On peut aussi forcer les $u_1, \ldots, u_r$ à être orthogonaux deux à deux (de même pour les v_i), ie : $U^TU = \boldsymbol{Id}_r$ et $V^TV = \boldsymbol{Id}_r$.

4- Les $\mathbf{u}_i$ (resp. les $\mathbf{v}_i^T$) sont orthonormés et engendrent le même espace que celui engendré par les colonnes (resp. les lignes) de M

$$\text{vect}\,(\mathbf{m}_1, \ldots, \mathbf{m}_{m_2}) = \text{vect}\,(\mathbf{u}_1, \ldots, \mathbf{u}_r)$$

5- SVD généralise aux matrices non carées la décomposition spéctrale.

Preuve :

M^TM est une matrice $m_2 \times m_2$ symétrique, semi définie positive, et de rang r ($rg(M^TM) = rg(M)$), donc elle admet des valeurs propres réelles positives $(\lambda_1, \ldots, \lambda_r, \ldots)$ et une base orthonormée de vecteurs propres associés $(v_1, \ldots, v_r, \ldots)$:

$$M^TMv_i = \lambda_i v_i$$

De plus :

$$v_j^TM^TMv_i = \lambda_i v_j^T v_i = \lambda_i \delta_{ij}$$

Définissons pour tout $i = 1, \ldots, r$:

$$s_i = \sqrt{\lambda_i} \quad et \quad u_i = \frac{Mv_i}{\sqrt{\lambda_i}}$$

Les $(u_1, \ldots u_r)$ forment une famille orthonormée , en effet :

$\langle u_i, u_j \rangle = \left\langle \frac{Mv_i}{\sqrt{\lambda_i}}, \frac{Mv_j}{\sqrt{\lambda_j}} \right\rangle = \frac{1}{\sqrt{\lambda_i \lambda_j}} \langle Mv_i, Mv_j \rangle = \frac{\lambda_i}{\sqrt{\lambda_i \lambda_j}} \delta_{ij}$.

Si $i \neq j \quad \langle u_i, u_j \rangle = 0$.

Si $i = j \quad \langle u_i, u_j \rangle = \frac{\lambda_i}{\sqrt{\lambda_i \lambda_i}} = 1$.

Alors, 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 :

$$U^TMV = \begin{pmatrix} \frac{1}{s_1} v_1^T M^T M v_1 & 0 & \ldots & 0 \\ 0 & \frac{1}{s_2} v_2^T M^T M v_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & \frac{1}{s_r} v_r^T M^T M v_r \end{pmatrix}$$

$$= \begin{pmatrix} \frac{\lambda_1}{s_1} v_1^T v_1 & 0 & \ldots & 0 \\ 0 & \frac{\lambda_2}{s_2} v_2^T v_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & \frac{\lambda_r}{s_r} v_r^T v_r \end{pmatrix} = \begin{pmatrix} s_1 & 0 & \ldots & 0 \\ 0 & s_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & s_r \end{pmatrix} = diag(s_1, \ldots, s_r).$$

Proposition :

UU^T est la projection orthogonale sur l'espace engendré par les $(\mathbf{u}_1, \ldots, \mathbf{u}_r)$.

Preuve :

Rappel : la projection orthogonale d'un vecteur $\mathbf{y} \in \mathbb{R}^{m_2}$ sur le sous espace vectoriel de $\mathbb{R}^{m_2}$ engendré par $(\mathbf{u}_1, \ldots, \mathbf{u}_r)$. Cette projection, notée $\mathbb{P}_U$ est caractérisée par les propriétés :

(i) $\mathbb{P}_U(\mathbf{y})$est une combinaison linéaire des colonnes de U.

(ii) $\mathbf{y} - \mathbb{P}_U(\mathbf{y})$ est orthogonal aux colonnes de U.

- Commençons à prouver les deux propriétés caractéristique de la projection orthogonale.

(i) Le produit $U^T\mathbf{y}$ st un vecteur ligne de taille r. Notons le $(\mathbf{x}_1, \ldots, \mathbf{x}_r)$. On peut alors écrire le vecteur $UU^T\mathbf{y}$ de la façon suivante :

$$UU^T\mathbf{y} = \sum_{i=1}^{r} \mathbf{x}_i \mathbf{u}_i$$

C'est-à-dire que la propriété caractéristique (i) est vérifiée.
(ii) Soit $\mathbf{w}$ la matrice ligne de taille r dont les éléments sont les produits scalaires $\langle \mathbf{y} - UU^T\mathbf{y}, w_i \rangle$. On veut montrer que chacun de ces produits scalaires est nul, c'est-à-dire que $\mathbf{w}$ est le vecteur nul. Or

$$\mathbf{w} = U^T\left(\mathbf{y} - UU^T\mathbf{y}\right) = U^T\mathbf{y} - U^TUU^T\mathbf{y} = U^T\mathbf{y} - U^T\mathbf{y} = 0$$

Corollaire :
Comme $\mathrm{vect}\,(\mathbf{m}_1, \ldots, \mathbf{m}_{m_2}) = \mathrm{vect}\,(\mathbf{u}_1, \ldots, \mathbf{u}_r)$ (d'après la remarque 4), On peut déduire que UU^T est la projection orthogonale sur l'espace engendré par les colonnes de M. Ainsi $U^TUM = M$.
Identiquement, on peut montrer que $MVV^T = M$, ceci donne : $UU^TMVV^T = UU^TM = M$.

3 Pseudo-inverse (Moore-Penrose (1955), inverse généralisée)

Définition :
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 :

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

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

$$\begin{aligned} XX^+ &= \sum_{j=1}^{n} s_j \mathbf{u}_j \mathbf{v}_j^\top \sum_{i=1}^{n} \frac{1}{s_i} \mathbf{v}_i \mathbf{u}_i^\top \\ &= \sum_{j=1}^{n} \sum_{i=1}^{n} s_j \frac{1}{s_i} \mathbf{u}_j \mathbf{v}_j^\top \mathbf{v}_i \mathbf{u}_i^\top \\ &= \sum_{j=1}^{n} \sum_{i=1}^{n} s_j \frac{1}{s_i} \delta_{i,j} \mathbf{u}_j \mathbf{u}_i^\top = \sum_{i=1}^{n} \mathbf{u}_i \mathbf{u}_i^\top = \mathrm{Id}_n \end{aligned}$$

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 *cf.*
full-matrices=True/False.
• Pseudo-inverse : ***Xinv = np.linalg.pinv(X)***.
Exemple :
Soit une matrice A et sa décomposition sigulière obtenu avec la commande

U, s, V = np.linalg.svd(A)

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

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

Le pseudo inverse peut être obtenu aussi plus simplement avec la commande A^+ = ***np.linalg.pinv(A)***

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

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.