

Feuille 2 : applications des moindres carrés, système de recommandation

Exercice 1. Compléments d'algèbre linéaire sur le rang de matrice.

Soit une matrice $A \in \mathbb{R}^{m \times n}$. Le rang de A noté $\text{rang}(A)$ est la dimension du sous-espace engendré par ses vecteurs colonnes et c'est aussi la dimension de son espace image :

$$\text{rang}(A) = \dim(\text{vect}(A^1, \dots, A^n)) = \dim(\text{Im}(A)) \quad (1)$$

où A^i désigne le i -ème vecteur colonne de A et $\text{Im}(A) = \{y \in \mathbb{R}^m, \exists x \in \mathbb{R}^n \text{ tel que } Ax = y\}$. Le rang de A est aussi le nombre maximal de vecteurs colonnes linéairement indépendants. On a $\text{rang}(A) \leq n$. Enfin, on rappelle le théorème du rang :

$$\text{rang}(A) + \dim(\ker(A)) = n \quad (2)$$

où $\ker(A)$ désigne le noyau de A .

1. Soit $M \in \mathbb{R}^{n \times n}$ une matrice inversible. Montrer les résultats suivants.

(i) La matrice $D = BM$ avec $B \in \mathbb{R}^{m \times n}$ vérifie $\text{rang}(D) = \text{rang}(B)$

(ii) La matrice $E = MB$ avec $B \in \mathbb{R}^{n \times m}$ vérifie $\text{rang}(E) = \text{rang}(B)$

Indications. Pour (i), montrer que $\text{Im}(D) = \text{Im}(B)$. Pour (ii), montrer que $\ker(E) = \ker(B)$ et utiliser le théorème du rang.

2. *Décomposition SVD et rang.* On considère la décomposition SVD de A sous la forme

$$A = U \Sigma V^T = \sum_{i=1}^p \sigma_i U^i V^{i^T} \quad (3)$$

où $p = \min(n, m)$, $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_p \geq 0$ sont les valeurs singulières de A et $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_p) \in \mathbb{R}^{m \times n}$. Les matrices $U \in \mathbb{R}^{m \times m}$ et $V \in \mathbb{R}^{n \times n}$ sont orthogonales. Montrer que

$$\text{rang}(A) \leq k \Leftrightarrow A = U \Sigma_k V^T = \sum_{i=1}^k \sigma_i U^i V^{i^T} \quad (4)$$

où $\Sigma_k = \text{diag}(\underbrace{\sigma_1, \dots, \sigma_k}_p, 0, \dots, 0) \in \mathbb{R}^{m \times n}$.

3. *Autre caractérisation du rang.* Montrer que

$$\text{rang}(A) \leq k \Leftrightarrow \exists X \in \mathbb{R}^{m \times k} \text{ et } Y \in \mathbb{R}^{n \times k} \text{ telles que } A = XY^T \quad (5)$$

Exercice 2*. Approximation de rang faible : théorème d'Eckart-Young (1936)

On rappelle que la trace d'une matrice est la somme de ses éléments diagonaux : $\text{tr}(M) = \sum_{i=1}^n M_{ii}$ pour $M \in \mathbb{R}^{n \times n}$. Soit une matrice $A \in \mathbb{R}^{m \times n}$. On définit la norme de Frobenius* de A par

$$\|A\|_F = \sqrt{\text{tr}(A^T A)} = \left(\sum_{i=1}^m \sum_{j=1}^n a_{ij}^2 \right)^{1/2}$$

1. Soit Q une matrice orthogonale ($Q^{-1} = Q^T$). Montrer que $\|QA\|_F = \|A\|_F$ c'est-à-dire qu'une transformation orthogonale laisse inchangé la norme de Frobenius.

*. On peut montrer que la norme de Frobenius n'est pas une norme subordonnée à une norme vectorielle mais qu'elle vérifie quand même la propriété de sous-multiplicativité $\|AB\|_F \leq \|A\|_F \|B\|_F$

2. Pour simplifier, on suppose dans toute la suite de l'exercice que $m < n$ (le cas $m > n$ se traite de façon analogue). Soit $r = \text{rang}(A)$ et on considère la décomposition SVD de A donnée par (4) à l'exercice 1. Soit aussi une matrice $X \in \mathbb{R}^{m \times n}$ avec $\text{rang}(X) \leq k \leq r$. On cherche à déterminer la meilleure approximation X de A au sens de la norme de Frobenius :

$$\min_{\text{rang}(X) \leq k} \|A - X\|_F^2 \quad (6)$$

On va montrer que la matrice X^* réalisant le minimum est donnée par (théorème d'Eckart-Young)

$$X^* = U \Sigma_k V^T = \sum_{i=1}^k \sigma_i U^i V^{i^T}. \quad (7)$$

- (a) Soit $X \in \mathbb{R}^{m \times n}$ avec $\text{rang}(X) \leq k \leq r$. On pose $Y = U^T X V$. Montrer que $\|A - X\|_F = \|\Sigma_r - Y\|_F$ et

$$\min_{\text{rang}(X) \leq k} \|A - X\|_F^2 = \min_{\text{rang}(Y) \leq k} \|\Sigma_r - Y\|_F^2 \quad (8)$$

- (b) Pour tout $Y \in \mathbb{R}^{m \times n}$, montrer que

$$\|\Sigma_r - Y\|_F^2 = \sum_{i=1}^m \sum_{\substack{j=1 \\ j \neq i}}^n y_{ij}^2 + \sum_{i=1}^r (\sigma_i - y_{ii})^2 + \sum_{i=r+1}^m y_{ii}^2 \quad (9)$$

- (c) En déduire que la matrice Y^* réalisant le minimum de (8) est nécessairement *diagonale*, puis que Y^* est de la forme

$$Y^* = \text{diag}(\underbrace{s_1, \dots, s_k}_{p(=m)}, 0, \dots, 0) \in \mathbb{R}^{m \times n}$$

- (d) On a montré ainsi que

$$\min_{\text{rang}(X) \leq k} \|A - X\|_F^2 = \min_{Y \in D_k} \|\Sigma_r - Y\|_F^2 \quad (10)$$

où D_k est l'ensemble des matrices diagonales de rang $\leq k$:

$$D_k = \left\{ Y \in \mathbb{R}^{m \times n}, Y = \text{diag}(\underbrace{s_1, \dots, s_k}_p, 0, \dots, 0) \right\}.$$

Montrer que pour tout $Y \in D_k$, on a

$$\|\Sigma_r - Y\|_F^2 = \sum_{i=1}^k (\sigma_i - s_i)^2 + \sum_{i=k+1}^r \sigma_i^2 \quad (11)$$

En déduire la matrice Y^* optimale minimisant (8) et que la matrice X^* optimale correspondante est donnée par (7). Donner la valeur minimale $\|A - X^*\|_F^2$.

Exercice 3. Système de recommandation, moindres carrés et méthode ALS

Les systèmes de recommandation reposent souvent sur l'idée que les préférences peuvent être expliquées par un petit nombre de facteurs dits latents. Dans un contexte de systèmes de recommandation tels qu'ils peuvent être proposés par des plateformes de streaming (Youtube, Spotify, Netflix...), les utilisateurs évaluent certains items (une note) et le système de recommandation cherche à calculer une note aux items qui n'en ont pas reçus des utilisateurs.

Dans l'exemple ci-dessous, 4 utilisateurs expriment leurs préférences parmi 6 films (cf. tableau 1. ci-dessous). Les points d'interrogations indiquent l'absence de notes. Un système de recommandation va évaluer ces notes manquantes en déterminant une approximation de rang faible k de la matrice incomplète des notes.

Modélisation. On considère m utilisateurs donnant des évaluations (notes) à certains objets d'un ensemble de n objets (ou items). Cela définit alors une matrice des notes $A \in \mathbb{R}^{m \times n}$, où A_{ij} représente la note donnée par

	Star Wars épisode VIII	Pulp Fiction	Le fabuleux destin d'Amélie Poulain	Matrix	Il était une fois dans l'Ouest	Les Bronzés à TELECOM Nancy
Cathy	5	?	3	2	3	?
Stéphane	4	5	4	1	2	1
Sophie	1	?	?	?	?	?
Anna	?	3	5	4	2	?

TABLE 1 – Un exemple de matrice de notes incomplète.

l'utilisateur i à l'objet j . Certaines entrées peuvent être manquantes. L'objectif est de construire un modèle de recommandation basé sur une factorisation de rang faible $k \ll p := \min(m, n)$. En suivant la propriété (5) (cf. Exercice 1, 3.), on cherche une approximation

$$A \simeq XY^T$$

avec $X \in \mathbb{R}^{m \times k}$, $Y \in \mathbb{R}^{n \times k}$ et $k \ll p = \min(m, n)$. Ainsi, on voudrait déterminer les matrices X et Y qui minimisent la norme

$$\|A - XY^T\|_F^2 = \sum_{i=1}^m \sum_{j=1}^n (A_{ij} - X_i Y_j^T)^2$$

(X_i désigne la i -ème ligne de X) mais puisque A est incomplètement connue, on va minimiser uniquement sur les quantités connues dans la matrice A . On note $\Omega \subset \llbracket 1, m \rrbracket \times \llbracket 1, n \rrbracket$ l'ensemble des indices observés i.e. $(i, j) \in \Omega$ si et seulement si le coefficient A_{ij} de A est connu. On définit alors

$$\|A - XY^T\|_{F, \Omega} := \left(\sum_{(i, j) \in \Omega} (A_{ij} - X_i Y_j^T)^2 \right)^{1/2} \quad (12)$$

et on cherche les matrices X et Y solution de

$$\min_{X, Y} \|A - XY^T\|_{F, \Omega}^2 \quad (13)$$

Le but de l'approximation $\hat{A} := XY^T$ de A est de prédire les valeurs $\hat{A}_{ij}$ pour $(i, j) \notin \Omega$ c'est-à-dire prédire les valeurs de A_{ij} inconnues.

Pour résoudre (13), on utilise la méthode ALS (Alternating Least Square) qui consiste à résoudre itérativement et alternativement (13) pour Y fixé puis pour X fixé

1. On fixe $Y \in \mathbb{R}^{n \times k}$. Pour résoudre (13) avec Y fixé, montrer qu'il est suffisant de résoudre le problème suivant, pour chaque ligne i de X :

$$\min_{x_i \in \mathbb{R}^k} \|A_i - X_i Y^T\|_{2, \Omega_i}^2 = \min_{x_i \in \mathbb{R}^k} \|a_i - \mathcal{Y}_i x_i\|_2^2 \quad (14)$$

où $x_i = X_i^T$ (vecteur colonne), a_i est le vecteur (colonne) formé des items notés par l'utilisateur i c'est-à-dire que le vecteur a_i est formé de tous les coefficients A_{ij} pour $(i, j) \in \Omega$ et enfin $\mathcal{Y}_i$ est la matrice formée des vecteurs lignes $Y_j \in \mathbb{R}^k$ tels que $(i, j) \in \Omega$. On a noté $\|u\|_{2, \Omega_i} = (\sum_{j \in \Omega_i} u_j^2)^{1/2}$ avec $j \in \Omega_i$ ssi $(i, j) \in \Omega$.

En déduire les systèmes linéaires vérifiés par la solution optimale X^* de (14) en fonction de Y .

2. On fixe $X \in \mathbb{R}^{m \times k}$. Pour résoudre (13) avec X fixé, montrer qu'il est suffisant de résoudre le problème suivant, pour chaque ligne j de Y :

$$\min_{y_j \in \mathbb{R}^k} \|A^j - XY_j^T\|_{2, \Omega^j}^2 = \min_{y_j \in \mathbb{R}^k} \|a_j - X_j y_j\|_2^2 \quad (15)$$

où $y_j = Y_j^T$ (vecteur colonne), et X_j est la matrice formée des vecteurs lignes $X_i \in \mathbb{R}^k$ tels que $(i, j) \in \Omega$. On a noté $\|u\|_{2, \Omega^j} = (\sum_{i \in \Omega^j} u_i^2)^{1/2}$ avec $i \in \Omega^j$ ssi $(i, j) \in \Omega$.

En déduire les systèmes linéaires vérifiés par la solution optimale Y^* de (15) en fonction de X .

3. L'algorithme ALS pour résoudre (13) consiste à résoudre itérativement et alternativement (14) pour tout i puis (15) pour tout j , en utilisant le Y et le X trouvés à l'itération précédente jusqu'à obtenir une convergence. On définit ainsi une suite de matrices $(X^{(l)}, Y^{(l)})$ selon l'algorithme ALS suivant.

On note $F(X, Y) = \|A - XY^\top\|_{F, \Omega}^2$.

ALGORITHME ALS

INITIALISER $(X^{(0)}, Y^{(0)})$

POUR $l = 1, 2, \dots$

- Détermination de $X^{(l)}$: résoudre $F(X^{(l)}, Y^{(l-1)}) = \min_X F(X, Y^{(l-1)})$
- Détermination de $Y^{(l)}$: résoudre $F(X^{(l)}, Y^{(l)}) = \min_Y F(X^{(l)}, Y)$

FIN POUR

- (a) Montrer que la fonction objective $F(X^{(l)}, Y^{(l)})$ est décroissante et en déduire la convergence.

Remarque. On n'a pas de garantie d'un minimum global mais seulement local, F n'étant pas convexe en (X, Y) .

- (b) Ecrire un script python pour résoudre (13) par ALS.

— Pour l'initialisation de $X^{(0)}$ et $Y^{(0)}$, on prendra $\hat{A} := X^{(0)}Y^{(0)\top} \in \mathbb{R}^{m \times n}$ avec

$$\hat{A}_{ij} = \begin{cases} A_{ij} & \text{si } (i, j) \in \Omega \\ \mu & \text{sinon} \end{cases}$$

avec $\mu = (\sum_{(i,j) \in \Omega} A_{ij}) / \text{card}(\Omega)$ i.e. μ est la moyenne globale des données.

On effectue alors une décomposition SVD de $\hat{A} = U\Sigma V^\top$ comme à l'exercice 1. avec (4) et on prend $X = U_k(\hat{\Sigma}_k)^{1/2}$ et $Y = V_k(\hat{\Sigma}_k)^{1/2}$ où $U_k \in \mathbb{R}^{m \times k}$ et $V_k \in \mathbb{R}^{n \times k}$ sont les matrices obtenues en sélectionnant les k premières colonnes de U et V respectivement et $\hat{\Sigma}_k = \text{diag}(\sigma_1, \dots, \sigma_k) \in \mathbb{R}^{k \times k}$.

— Le test d'arrêt (test de convergence à la place de la boucle "Pour/Fin pour") peut s'écrire

$$\frac{|e^{(l)} - e^{(l-1)}|}{|e^{(l-1)}|} \leq \epsilon$$

où $e^{(l)} = \|A - X^{(l)}Y^{(l)\top}\|_{F, \Omega}$ est l'erreur entre la matrice A initiale et la matrice complétée à l'itération l . Il est judicieux aussi de mettre un "garde-fou" en limitant le nombre d'itérations à `maxiter`.

— *Régularisation.* On peut régulariser les erreurs à minimiser dans (14) et (15) en ajoutant les termes $\lambda\|X_i\|_2^2$ et $\lambda\|Y_j\|_2^2$ respectivement, avec un petit paramètre $\lambda > 0$:

$$\min_{X_i \in \mathbb{R}^k} (\|A_i - X_i Y^\top\|_{2, \Omega}^2 + \lambda\|X_i\|_2^2) \quad \text{et} \quad \min_{Y_j \in \mathbb{R}^k} (\|A^j - X Y_j^\top\|_{2, \Omega}^2 + \lambda\|Y_j\|_2^2)$$

Les équations normales s'écrivent alors (le vérifier !)

$$(\mathcal{Y}_i^\top \mathcal{Y}_i + \lambda I_k) x_i = \mathcal{Y}_i^\top a_i \quad (\forall i)$$

$$(\mathcal{X}_j^\top \mathcal{X}_j + \lambda I_k) y_j = \mathcal{X}_j^\top a_j \quad (\forall j)$$

- (c) Résoudre l'exemple donné par le tableau 1. On prendra $k = 2$ et $\lambda = 10^{-5}$.