

Analyse des données avancée

Chapitre 4 : Diagonalisation et décomposition en valeurs singulières



Jaouad Madkour

Faculté d'Économie et de Gestion de Tanger · 27 août 2026

Diagonaliser une matrice symétrique

La décomposition en valeurs singulières

Approcher un tableau par un tableau plus simple

Ce qu'il faut retenir

Diagonaliser une matrice symétrique

L'énoncé

Toute matrice **symétrique réelle** $\mathbf{A}$ d'ordre p s'écrit

$$\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}', \quad \mathbf{U}'\mathbf{U} = \mathbf{I}$$

avec $\mathbf{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_p)$ réelle. Les colonnes de $\mathbf{U}$ sont **orthonormées**.

Ce que le théorème promet, et qui n'allait pas de soi

Trois choses, dont aucune n'est vraie pour une matrice quelconque : les valeurs propres sont **réelles**, il y en a **toujours** p , et les vecteurs propres sont **orthogonaux**.

C'est cette orthogonalité qui donnera aux axes factoriels leur propriété la plus utile : **le deuxième axe n'a rien à ajouter à ce que le premier a dit**.

Le cas qui nous concerne

Une matrice de variance-covariance $\mathbf{V} = \mathbf{X}'\mathbf{D}\mathbf{X}$ est **symétrique semi-définie positive** :

$$\mathbf{a}'\mathbf{V}\mathbf{a} = \sum_i p_i (\mathbf{x}'_i \mathbf{a})^2 \geq 0 \quad \text{pour tout } \mathbf{a}$$

Ses valeurs propres sont donc **toutes positives ou nulles**.

La lecture statistique

$\mathbf{a}'\mathbf{V}\mathbf{a}$ est la **variance de la combinaison linéaire** $\mathbf{X}\mathbf{a}$. Une variance ne peut pas être négative : la positivité n'est pas un détail technique, c'est la même chose dite en algèbre.

Une valeur propre nulle signale une **variable redondante** — une colonne combinaison exacte des autres.

La matrice

$$\mathbf{A} = \mathbf{X}'\mathbf{X} = \begin{pmatrix} 10 & 6 \\ 6 & 10 \end{pmatrix}$$

Le polynôme caractéristique :

$$(10 - \lambda)^2 - 36 = 0 \iff 10 - \lambda = \pm 6$$

$$\lambda_1 = 16, \quad \lambda_2 = 4$$

Exemple

Pour $\lambda_1 = 16$: $-6a_1 + 6a_2 = 0$, donc $a_1 = a_2$.

$$\mathbf{u}_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \mathbf{u}_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

Vérifications : $\mathbf{u}_1' \mathbf{u}_2 = 0$, $\lambda_1 + \lambda_2 = 20 = \text{trace}(\mathbf{A})$, $\lambda_1 \lambda_2 = 64 = \det(\mathbf{A})$.

Intuition

$\mathbf{u}_1$ est la **bissectrice** : la direction où les deux variables varient ensemble. Elle porte $16/20 = 80\%$ de l'information.

$\mathbf{u}_2$ est la direction où elles s'opposent. Sans avoir prononcé le mot ACP, on vient de la faire.

La décomposition en valeurs singulières

Ce que la diagonalisation ne sait pas faire

Le problème

$\mathbf{X}$ est $n \times p$ — **rectangulaire**. Elle n'a ni valeurs propres, ni vecteurs propres.

On peut diagonaliser $\mathbf{X}'\mathbf{X}$, qui est $p \times p$, ou $\mathbf{X}\mathbf{X}'$, qui est $n \times n$ — mais rien ne dit encore comment les deux se relient.

La décomposition en valeurs singulières

Toute matrice $\mathbf{X}$ de rang r s'écrit

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}'$$

- $\mathbf{U}$, $n \times r$, colonnes orthonormées — les vecteurs singuliers à **gauche**;
- $\mathbf{V}$, $p \times r$, colonnes orthonormées — les vecteurs singuliers à **droite**;
- $\mathbf{\Sigma} = \text{diag}(\sigma_1 \geq \dots \geq \sigma_r > 0)$, les **valeurs singulières**.

Les trois identités

$$\mathbf{X}'\mathbf{X} = \mathbf{V}\Sigma^2\mathbf{V}'$$

$$\mathbf{X}\mathbf{X}' = \mathbf{U}\Sigma^2\mathbf{U}'$$

$$\mathbf{X}\mathbf{v}_k = \sigma_k\mathbf{u}_k$$

$$\mathbf{X}'\mathbf{u}_k = \sigma_k\mathbf{v}_k$$

Le résultat qui va tout gouverner

$\mathbf{X}'\mathbf{X}$ et $\mathbf{X}\mathbf{X}'$ n'ont pas la même taille, mais **exactement les mêmes valeurs propres non nulles** : les σ_k^2 .

C'est la **dualité**. Analyser le nuage des individus ou celui des variables donnera les mêmes valeurs propres — non par coïncidence, mais parce que ce sont les deux faces d'une seule décomposition. Le chapitre 6 en tire toutes les conséquences.

Les valeurs singulières

$\mathbf{X}'\mathbf{X}$ a pour valeurs propres 16 et 4, donc

$$\sigma_1 = \sqrt{16} = 4, \quad \sigma_2 = \sqrt{4} = 2$$

Les vecteurs à gauche, par $\mathbf{X}\mathbf{v}_k = \sigma_k \mathbf{u}_k$

$$\mathbf{X}\mathbf{v}_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} -4 \\ 0 \\ 0 \\ 4 \end{pmatrix} \Rightarrow \mathbf{u}_1 = \frac{\mathbf{X}\mathbf{v}_1}{\sigma_1} = \frac{1}{4\sqrt{2}} \begin{pmatrix} -4 \\ 0 \\ 0 \\ 4 \end{pmatrix} = \begin{pmatrix} -0,707 \\ 0 \\ 0 \\ 0,707 \end{pmatrix}$$

Les individus 2 et 3 ont une coordonnée **nulle** sur le premier axe : ils sont exactement sur le second.

Intuition

$$\sum_i (\mathbf{X}\mathbf{v}_1)_i^2 = \frac{16 + 0 + 0 + 16}{2} = 16 = \lambda_1$$

La somme des carrés des coordonnées sur un axe est sa valeur propre. C'est le contrôle à faire sur toute sortie de logiciel.

Approcher un tableau par un
tableau plus simple

La meilleure approximation de rang k

$$\mathbf{X}_{(k)} = \sum_{m=1}^k \sigma_m \mathbf{u}_m \mathbf{v}_m'$$

est, parmi **toutes** les matrices de rang k , celle qui minimise $\|\mathbf{X} - \mathbf{B}\|^2$.

L'erreur commise vaut $\sigma_{k+1}^2 + \dots + \sigma_r^2$.

Pourquoi ce théorème est le vrai fondement du cours

Il ne dit pas seulement que la solution factorielle est bonne : il dit qu'**aucune autre méthode ne peut faire mieux**, à rang égal.

Retenir deux axes, c'est remplacer un tableau 300×8 par le meilleur tableau de rang 2 qui existe. Et l'erreur est **connue d'avance** : la somme des valeurs propres abandonnées.

Le coût du résumé

$$\lambda_1 + \lambda_2 = 4,535 + 1,645 = 6,180 \quad \text{sur } 8$$

Le plan retient 77,3 % de l'inertie; l'erreur d'approximation est

$$0,382 + 0,349 + 0,327 + 0,283 + 0,249 + 0,229 = 1,820$$

Ce que cela veut dire pour l'enseigne

Deux nombres par client — au lieu de huit — restituent plus des trois quarts de ce que le questionnaire contenait.

Le quart perdu n'est pas du signal : ce sont six valeurs propres presque égales, entre 0,23 et 0,38, c'est-à-dire **la signature du bruit**. Le chapitre 7 en fait un critère.

Ce qu'il faut retenir

Six points

1. Une matrice **symétrique réelle** se diagonalise en base orthonormée, à valeurs propres réelles.
2. Une matrice de variance-covariance est **semi-définie positive** : ses valeurs propres sont positives.
3. Toute matrice rectangulaire admet une **décomposition en valeurs singulières** $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}'$.
4. $\mathbf{X}'\mathbf{X}$ et $\mathbf{X}\mathbf{X}'$ ont **les mêmes valeurs propres non nulles** — c'est la dualité.
5. $\mathbf{X}\mathbf{v}_k = \sigma_k \mathbf{u}_k$: la relation qui fera passer des variables aux individus.
6. **Eckart et Young** : la troncature de la décomposition est la meilleure approximation de rang donné, et l'erreur vaut la somme des valeurs propres écartées.

La suite

L'outil est en place. Le chapitre 5 pose le problème de l'ACP et le résout — en trois lignes, désormais.