Comprendre l'ACP et l'AFC à travers la décomposition en valeurs singulières (SVD)

L. B.

1 Introduction

Cette note a pour objectif de comprendre comment la décomposition en valeurs singulières (SVD) est utilisée dans les deux méthodes d'analyse de données : l'analyse en composantes principales (ACP) et l'analyse des correspondances (AFC). La SVD est avant tout un outil algébrique, qui révèle la structure fondamentale d'une matrice de données. En s'appuyant sur cette structure, l'ACP et l'AFC construisent des représentations simplifiées qui permettent d'interpréter efficacement la variabilité ou les relations dans les données. Pour mettre en évidence ces liens, nous partirons d'une matrice de données simple de dimensions $n \times p$. Par des manipulations progressives, nous verrons comment apparaissent naturellement les concepts de valeurs propres, vecteurs propres, axes principaux et projections factorielles.

Exemple : on va considérer cette matrice 4×3

$$X = \begin{bmatrix} 2 & 0 & 1 \\ 0 & 1 & 1 \\ 2 & 1 & 2 \\ 0 & 1 & 0 \end{bmatrix}$$

1.1 1er concept clé: Le rang de X

Le rang r de X est le nombre de dimensions effectives dans lesquelles les données vivent. Le rang est une notion fondamentale pour les raisons suivantes :

- Il détermine le nombre d'axes principaux non triviaux, correspondant aux directions significatives conservant de l'information.
- Le rang est égal au nombre de valeurs singulières non nulles dans la décomposition en valeurs singulières (SVD) de la matrice.

1.2 Application à l'exemple X

Observons les dépendances linéaires dans l'exemple:

- La troisième ligne est dépendante parce que c'est la somme des deux premières :

$$(2, 0, 1) + (0, 1, 1) = (2, 1, 2)$$

- La quatrième ligne $(0, 1, 0)$ n'est pas une combinaison linéaire des deux premières lignes. Elle est donc indépendante.

En conclusion, trois lignes indépendantes existent, donc :

$$\boxed{\text{rang}(X) = 3}$$

2 Construction des matrices carrées et diagonalisation

La matrice X est en général rectangulaire $(n \times p)$, et il n'est pas possible de rechercher directement ses vecteurs propres. Pour étudier ses directions principales (axes factoriels), il est nécessaire de construire une matrice carrée, afin de pouvoir appliquer la diagonalisation. Deux options sont possibles:

- Construire $X'X$ (de dimension $p \times p$),
- Ou construire XX' (de dimension $n \times n$).

Ces deux matrices sont symétriques, positives semi-définies, et permettent d'analyser la structure de X à travers leurs vecteurs et valeurs propres.

2.1 Liens entre $X'X$ et XX'

- Les valeurs propres non nulles de $X'X$ et XX' sont identiques.
- Les vecteurs propres v_k de $X'X$ et u_k de XX' sont liés par:

$$u_k = \frac{1}{\sqrt{\lambda_k}} X v_k \quad \text{et} \quad v_k = \frac{1}{\sqrt{\lambda_k}} X' u_k$$

2.1.1 Démonstration

Soit la matrice carrée $X'X$, ses vecteurs propres *normalisés* correspondent à v_k, d'où

$$X'X v_k = \lambda_k v_k$$

On pré-multiplie par X:

$$XX'X v_k = X \lambda_k v_k$$

D'où:

$$XX'(X v_k) = \lambda_k (X v_k)$$

Cette expression indique que $(X v_k)$ correspond aux vecteurs propres de XX'. Néanmoins, il faut vérifier la norme de ces vecteurs. On calcule

$$||X v_k||^2 = (X v_k)'(X v_k) = v_k' X' X v_k$$

Puisque $X'X v_k = \lambda_k v_k$,

$$||X v_k||^2 = v_k' X' X v_k = v_k' \lambda_k v_k = \lambda_k v_k' v_k$$

Puisque la norme de v_k est unitaire, on a $v_k' v_k = ||v_k||^2 = 1$. Donc

$$||Xv_k||^2 = \lambda_k$$

Donc, pour obtenir des vecteurs propres normalisés à partir de (Xv_k), il suffit de diviser par $\sqrt{\lambda_k}$. On obtient les vecteurs propres normalisés de XX' qu'on va désigner par u_k tels que:

$$u_k = \frac{1}{\sqrt{\lambda_k}} X v_k$$

De la même manière, on peut partir de $XX'u_k = \lambda_k u_k$ pour démontrer que

$$v_k = \frac{1}{\sqrt{\lambda_k}} X' u_k$$

2.1.2 Illustration par l'exemple chiffré

Soit la matrice :

$$X = \begin{bmatrix} 2 & 0 & 1 \\ 0 & 1 & 1 \\ 2 & 1 & 2 \\ 0 & 1 & 0 \end{bmatrix}$$

On commence par construire les deux matrices carrées associées :

- La matrice $X'X$:

$$X'X = \begin{bmatrix} 2 & 0 & 2 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 1 & 2 & 0 \end{bmatrix}' \times \begin{bmatrix} 2 & 0 & 1 \\ 0 & 1 & 1 \\ 2 & 1 & 2 \\ 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 8 & 2 & 6 \\ 2 & 3 & 3 \\ 6 & 3 & 6 \end{bmatrix}$$

- La matrice XX' :

$$XX' = \begin{bmatrix} 2 & 0 & 2 & 0 \\ 0 & 1 & 1 & 1 \\ 2 & 1 & 2 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} 2 & 0 & 2 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 1 & 2 & 0 \end{bmatrix}' = \begin{bmatrix} 5 & 1 & 6 & 0 \\ 1 & 2 & 3 & 1 \\ 6 & 3 & 9 & 1 \\ 0 & 1 & 1 & 1 \end{bmatrix}$$

On obtient :

- **Valeurs propres non nulles :** (obtenues avec la fonction eigen() sur R)

$$\lambda_1 \approx 14.17, \quad \lambda_2 \approx 2.50, \quad \lambda_3 \approx 0.34$$

On peut vérifier que pour XX', les valeurs propres sont bien

$$\lambda_1 \approx 14.17, \quad \lambda_2 \approx 2.50, \quad \lambda_3 \approx 0.34, \quad \lambda_4 \approx 0$$

- **Vecteurs propres à droite** (colonnes de V, associés à $X'X$) (obtenus avec la fonction eigen() sur R):

$$v_1 = \begin{bmatrix} 0.71 \\ 0.30 \\ 0.63 \end{bmatrix}, \quad v_2 = \begin{bmatrix} 0.54 \\ -0.81 \\ -0.22 \end{bmatrix}, \quad v_3 = \begin{bmatrix} -0.45 \\ -0.50 \\ 0.74 \end{bmatrix}$$

- **Vecteurs propres à gauche** (colonnes de U, associés à XX') sont approximativement :

$$u_1 = \begin{bmatrix} 0.55 \\ 0.25 \\ 0.80 \\ 0.08 \end{bmatrix}, \quad u_2 = \begin{bmatrix} 0.54 \\ -0.66 \\ -0.12 \\ -0.52 \end{bmatrix}, \quad u_3 = \begin{bmatrix} -0.28 \\ 0.42 \\ 0.14 \\ -0.85 \end{bmatrix}, \quad u_4 = \begin{bmatrix} -0.58 \\ -0.58 \\ 0.58 \\ 0.00 \end{bmatrix}$$

  Ces deux ensembles de vecteurs sont liés. On peut vérifier que pour chaque k :

$$u_k = \frac{1}{\sqrt{\lambda_k}} X v_k$$

  ce qui confirme la relation fondamentale entre vecteurs propres de $X'X$ et de XX' via la matrice X.

3 Reconstruction de la matrice originale X et lien entre diagonalisation et SVD

Nous pouvons reconstruire X via la formule suivante :

$$X = \sum_{k=1}^{r} \sqrt{\lambda_k} u_k v_k'$$

Chaque terme $\sqrt{\lambda_k} u_k v_k'$ est une matrice de rang 1.

3.1 Démonstration

Puisque

$$u_k = \frac{1}{\sqrt{\lambda_k}} X v_k$$

En multipliant par v_k', on obtient

$$u_k v_k' = \frac{1}{\sqrt{\lambda_k}} X v_k v_k'$$

D'où

$$X v_k v_k' = \sqrt{\lambda_k} u_k v_k'$$

Que représente $v_k v_k'$?
Le produit $v_k v_k'$ est une matrice carrée de taille $p \times p$ (si v_k est un vecteur colonne de taille $p \times 1$). Ce produit correspond à un projecteur orthogonal de rang 1 sur la direction définie par v_k :

- Il projette n'importe quel vecteur de l'espace sur l'axe engendré par v_k.
- Sa trace est égale à 1, car il projette sur une seule dimension.
- C'est une matrice symétrique ($v_k v_k'$ est égal à sa transposée).

- C'est une matrice idempotente : $(v_k v_k')(v_k v_k') = v_k v_k'$.

Reconstruction de X
Partant de l'expression :

$$X v_k v_k' = \sqrt{\lambda_k} u_k v_k'$$

en faisant la somme sur $k = 1, \ldots, r$, on obtient :

$$\sum_{k=1}^{r} \sqrt{\lambda_k} u_k v_k' = X \sum_{k=1}^{r} v_k v_k'$$

On va utiliser un principe important:

La somme des projecteurs $v_k v_k'$ sur une base orthonormale reconstruisant l'espace engendré par X donne une matrice identité sur cet espace réduit. Autrement dit:

$$\sum_{k=1}^{r} v_k v_k' = I_r$$

En utilisant ce résultat, on a :

$$\sum_{k=1}^{r} \sqrt{\lambda_k} u_k v_k' = X$$

On en déduit que :

$$X = \sum_{k=1}^{r} \sqrt{\lambda_k} u_k v_k' = \sum_{k=1}^{r} X_k$$

où chaque $X_k = \sqrt{\lambda_k} u_k v_k'$ représente une *composante de rang 1* de la matrice X.

3.2 Illustration avec l'exemple

Prenons la matrice de données :

$$X = \begin{bmatrix} 2 & 0 & 1 \\ 0 & 1 & 1 \\ 2 & 1 & 2 \\ 0 & 1 & 0 \end{bmatrix}$$

À partir des valeurs propres $\lambda_1 \approx 14.17$, $\lambda_2 \approx 2.50$, $\lambda_3 \approx 0.34$, et des vecteurs propres u_k, v_k trouvés précédemment, on peut calculer les matrices élémentaires :

$$X_k = \sqrt{\lambda_k} u_k v_k' \quad \text{pour} \quad k = 1, 2, 3$$

Explicitement :

- $X_1 = \sqrt{14.17} \times u_1 v_1'$
- $X_2 = \sqrt{2.50} \times u_2 v_2'$
- $X_3 = \sqrt{0.34} \times u_3 v_3'$

Les matrices élémentaires obtenues sont :

$$X_1 = \begin{bmatrix} 1.47 & 0.61 & 1.31 \\ 0.67 & 0.28 & 0.59 \\ 2.14 & 0.89 & 1.90 \\ 0.21 & 0.09 & 0.19 \end{bmatrix}, \quad X_2 = \begin{bmatrix} 0.46 & -0.69 & -0.19 \\ -0.55 & 0.84 & 0.23 \\ -0.10 & 0.15 & 0.04 \\ -0.44 & 0.66 & 0.18 \end{bmatrix}, \quad X_3 = \begin{bmatrix} 0.07 & 0.08 & -0.12 \\ -0.11 & -0.12 & 0.18 \\ -0.04 & -0.04 & 0.06 \\ 0.22 & 0.25 & -0.37 \end{bmatrix}$$

En sommant ces trois matrices, on retrouve la matrice originale X (à des arrondis près).

$$X = X_1 + X_2 + X_3$$

Quelques remarques:

1. La somme des carrés des éléments de chaque X_k est égale à λ_k, ce qui confirme que chaque axe principal porte une quantité d'information proportionnelle à sa valeur propre.

2. L'importance des composantes décroît : X_1 capture l'essentiel de l'information (environ 83% de l'énergie totale), suivi par X_2 et X_3.

Axe k	Valeur propre λ_k	Pourcentage de l'énergie totale
1	14.17	83%
2	2.50	15%
3	0.34	2%

   On voit que la quasi-totalité de l'information est concentrée dans les deux premiers axes, le troisième étant plus marginal.

3. Ce résultat est cohérent avec le fait que le rang de la matrice X est $r = 3$: il existe exactement trois axes principaux portant l'intégralité de l'information, chacun associé à une valeur propre non nulle.

3.3 Quel lien avec la décomposition en valeurs singulières (SVD)?

Nous avons obtenu l'expression :

$$X = \sum_{k=1}^{r} \sqrt{\lambda_k} u_k v_k'$$

Chaque terme $\sqrt{\lambda_k} u_k v_k'$ est une matrice de rang 1, qui capture la contribution de l'axe principal k. En regroupant les vecteurs colonnes u_k dans une matrice U, les vecteurs colonnes v_k dans une matrice V, et les racines carrées des valeurs propres $\sqrt{\lambda_k}$ dans une matrice diagonale Σ, on peut écrire cette somme sous forme matricielle compacte :

$$X = U \Sigma V'$$

où :

- U est une matrice $n \times r$ dont les colonnes sont les vecteurs propres normalisés u_k de XX',
- V est une matrice $p \times r$ dont les colonnes sont les vecteurs propres normalisés v_k de $X'X$,
- Σ est une matrice diagonale $r \times r$ contenant les valeurs singulières $\sqrt{\lambda_k}$.

Cette écriture matricielle est exactement la décomposition en valeurs singulières (SVD) de X.

3.4 Illustration de la SVD sur l'exemple

Nous appliquons directement la décomposition en valeurs singulières (SVD) à la matrice X :

$$X = \begin{bmatrix} 2 & 0 & 1 \\ 0 & 1 & 1 \\ 2 & 1 & 2 \\ 0 & 1 & 0 \end{bmatrix}$$

La SVD de X nous donne :

$$X = U\Sigma V'$$

avec :

$$U = \begin{bmatrix} 0.55 & 0.54 & -0.28 \\ 0.25 & -0.66 & 0.42 \\ 0.80 & -0.12 & 0.14 \\ 0.08 & -0.52 & -0.85 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 3.76 & 0 & 0 \\ 0 & 1.58 & 0 \\ 0 & 0 & 0.58 \end{bmatrix}, \quad V = \begin{bmatrix} 0.71 & 0.54 & -0.45 \\ 0.30 & -0.81 & -0.50 \\ 0.63 & -0.22 & 0.74 \end{bmatrix}$$

Par comparaison avec les résultats précédemment obtenus :

- Les colonnes de V sont les vecteurs propres normalisés de $X'X$,
- Les colonnes de U sont les vecteurs propres normalisés de XX',
- Les valeurs singulières $\sqrt{\lambda_k}$ sont les racines carrées des valeurs propres de $X'X$ (ou XX').

Ainsi, tout ce qui a été construit pas à pas (valeurs propres, vecteurs propres, ...) correspond parfaitement aux éléments de la décomposition en valeurs singulières de X. Cela montre que la diagonalisation et la SVD sont très liées, et que la SVD est simplement une façon compacte de résumer toute la structure principale de la matrice X.

4 Analyse en Composantes Principales (ACP) : Projection des individus

4.1 Présentation de la matrice de données

Considérons que dans la matrice de données X, chaque ligne correspond à un individu (A, B, C, D) et chaque colonne à une variable (V1, V2, V3) :

$$X = \begin{array}{c|ccc} & \text{V1} & \text{V2} & \text{V3} \\ \hline \text{Individu A} & 2 & 0 & 1 \\ \text{Individu B} & 0 & 1 & 1 \\ \text{Individu C} & 2 & 1 & 2 \\ \text{Individu D} & 0 & 1 & 0 \end{array}$$

Cette matrice X contient :

- $n = 4$ individus (A, B, C, D),
- $p = 3$ variables (V1, V2, V3).

Chaque individu est décrit par ses valeurs sur les trois variables.

4.2 Projection des individus

L'Analyse en Composantes Principales (ACP) vise en premier lieu à étudier la structure des individus. Pour cela :

- On forme la matrice carrée $X'X$ (produit de X transposée par X),
- On diagonalise $X'X$ pour obtenir les vecteurs propres v_k et les valeurs propres λ_k,
- Les axes principaux correspondent aux directions données par les v_k.

Les nouvelles coordonnées des individus dans l'espace factoriel sont obtenues par le produit :

$$\text{Coordonnées individus} = D = XV$$

(en considérant que X a déjà été centrée ou bien centrée-réduite).

4.3 Projection des variables

Il est aussi possible de représenter les variables. Les coordonnées des variables sur les axes principaux peuvent être obtenues de deux manières :

- En utilisant directement :

  $$\text{Coordonnées variables} = C = V\Lambda^{1/2}$$

  où Λ est la matrice diagonale des valeurs propres,

- Ou de manière équivalente :

  $$\text{Coordonnées variables} = C = X'U$$

  en utilisant les vecteurs propres u_k de XX' (parce que $v_k = \frac{1}{\sqrt{\lambda_k}} X'u_k$).

Explication: $X = U\Sigma V'$ d'où $X' = V\Sigma' U'$. On multiplie par U: $X'U = V\Sigma' U'U$. On obtient: $X'U = V\Sigma'$. Puisque $\Sigma' = \Sigma$ et $C = V\Lambda^{1/2} = V\Sigma$, on obtient: $C = X'U$

Remarque : Ces formules données pour l'ACP ($D = XV$, $C = X'U$) supposent que la matrice X est préalablement centrée par variables (et éventuellement réduite), et que les individus sont équipondérés (pas de pondération spécifique). En présence de pondérations ou de normalisations supplémentaires, des ajustements seraient nécessaires.

Enfin, en ACP, les variables peuvent être projetées pour mieux comprendre les axes, mais elles ne jouent pas un rôle symétrique avec les individus. Ceci contraste avec l'Analyse des Correspondances (AFC), où lignes et colonnes sont traitées de manière équivalente.

5 Analyse des correspondances (AFC) : Symétrie lignes-colonnes

5.1 De la matrice de contingence à la matrice de travail

En analyse factorielle des correspondances (AFC), on commence traditionnellement par une matrice de contingence notée Y, de dimensions $n \times p$, croisant deux caractères qualitatifs (par exemple, individus et catégories).

On construit ensuite :

- La matrice des fréquences P, dont les éléments correspondent aux f_{ij}
- Les profils lignes r (somme des lignes de P) et profils colonnes c (somme des colonnes de P).
- Les matrices diagonales $D_r = \text{diag}(r)$ et $D_c = \text{diag}(c)$.

On définit alors la matrice des résidus normalisés :

$$Z = D_r^{-1/2}(P - rc')D_c^{-1/2}$$

5.2 Matrice de contingence de l'exemple

Supposons une enquête portant sur les préférences de consommation entre trois produits (P1, P2 et P3), classées par catégorie de clients.

La matrice de contingence Y est la suivante :

$Y =$

	P1	P2	P3
Catégorie (A)	...	...	...
Catégorie (B)	...	...	...
Catégorie (C)	...	...	...
Catégorie (D)	...	...	...

Chaque cellule indique le nombre de personnes appartenant à la catégorie de clients ayant exprimé une certaine préférence pour un produit donné. À partir de la matrice de contingence Y, nous construisons Z. Pour simplifier l'illustration, nous supposons que Z obtenu après normalisation coïncide avec la matrice X utilisée précédemment.

$$Z = \begin{array}{c|ccc} & \text{P1} & \text{P2} & \text{P3} \\ \hline \text{Catégorie (A)} & 2 & 0 & 1 \\ \text{Catégorie (B)} & 0 & 1 & 1 \\ \text{Catégorie (C)} & 2 & 1 & 2 \\ \text{Catégorie (D)} & 0 & 1 & 0 \end{array}$$

5.3 Projection des lignes et des colonnes et différence avec ACP

Dans l'AFC, les lignes (individus) et les colonnes (catégories) sont traitées de manière symétrique. Les coordonnées factorielles sont obtenues comme suit :

- Coordonnées des lignes (appelées F) :

$$F = D_r^{-1/2} U \Sigma$$

- Coordonnées des colonnes (appelées G) :

$$G = D_c^{-1/2} V \Sigma$$

Ces formules diffèrent de celles utilisées en ACP parce qu'en AFC, la géométrie sous-jacente n'est pas euclidienne standard, mais pondérée par les masses des profils lignes et colonnes. Les projections directes ZV et $Z'U$ ne respecteraient pas cette pondération. Il est nécessaire de repondérer les vecteurs propres V et U par $D_r^{-1/2}$ et $D_c^{-1/2}$ respectivement, ce qui conduit aux coordonnées factorielles F et G.

5.4 Résumé: ACP et AFC

Méthode	**Objet**	**Expression fondamentale**	**Formule équivalente**	**Pondération**
ACP	Coordonnées individus (D)	$D = XV$	$D = U\Lambda^{1/2}$	Non
	Coordonnées variables (C)	$C = V\Lambda^{1/2}$	$C = X'U$	Non
AFC	Coordonnées lignes (F)	$F = D_r^{-1/2} U \Lambda^{1/2}$	$F = D_r^{-1/2} ZV$	Oui, par masse lignes
	Coordonnées colonnes (G)	$G = D_c^{-1/2} V \Lambda^{1/2}$	$G = D_c^{-1/2} Z'U$	Oui, par masses colonnes