

Analyse des données avancée

Chapitre 2 : Le tableau, les poids et la métrique



Jaouad Madkour

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

Le triplet d'une analyse factorielle

Les poids des individus

La métrique des variables

Actives et illustratives

Ce qu'il faut retenir

Le triplet d'une analyse factorielle

Le triplet

Une analyse factorielle n'est pas définie par un tableau seul, mais par **trois objets** :

$$(\mathbf{X}, \mathbf{M}, \mathbf{D})$$

- $\mathbf{X}$, le tableau $n \times p$ des données;
- $\mathbf{M}$, la **métrique** sur les colonnes, matrice $p \times p$;
- $\mathbf{D}$, les **poids** des lignes, matrice diagonale $n \times n$.

Pourquoi trois et pas un

Parler de « la distance entre deux clients » n'a aucun sens tant qu'on n'a pas dit **dans quelle unité**.

Parler de « l'individu moyen » n'en a aucun tant qu'on n'a pas dit **quel poids** chaque individu porte.

$\mathbf{M}$ et $\mathbf{D}$ ne sont pas des raffinements : sans eux, la question n'est pas posée.

Ce que le triplet permet de voir

Une seule mécanique, trois méthodes

Méthode	$\mathbf{X}$	$\mathbf{M}$	$\mathbf{D}$
ACP non normée	données centrées	$\mathbf{I}$	$\frac{1}{n}\mathbf{I}$
ACP normée	données centrées	$\mathbf{D}_{1/s^2}$	$\frac{1}{n}\mathbf{I}$
AFC	écarts à l'indépendance	$\mathbf{D}_c^{-1}$	$\mathbf{D}_r$

Le gain, et il est considérable

Les trois lignes du tableau ne diffèrent que par $\mathbf{M}$ et $\mathbf{D}$. Le calcul qui suit est rigoureusement le même.

Ce que la Licence présentait comme trois méthodes à apprendre séparément est **une méthode et trois réglages**. C'est le résultat central de ce cours, et il sera établi au chapitre 5 pour l'ACP, au chapitre 9 pour l'AFC.

Les poids des individus

Pourquoi ils ne sont pas toujours égaux

Le poids d'un individu

$$\mathbf{D} = \text{diag}(p_1, \dots, p_n), \quad p_i > 0, \quad \sum_{i=1}^n p_i = 1$$

Le cas usuel est le **poids uniforme** $p_i = 1/n$.

Trois situations où l'uniforme est faux

Un sondage redressé. Si les cadres sont sous-représentés dans l'échantillon, on leur donne un poids supérieur pour rétablir la structure de la population.

Des données agrégées. Une ligne qui résume 40 magasins ne doit pas peser autant qu'une ligne qui en résume 2.

L'analyse des correspondances. Le poids d'une ligne y est sa fréquence marginale $f_{i\cdot}$. — jamais $1/n$. C'est ce qui la distingue de l'ACP.

L'individu moyen

$$\mathbf{g} = \sum_{i=1}^n p_i \mathbf{x}_i = \mathbf{X}'\mathbf{D}\mathbf{1}$$

Centrer, c'est remplacer $\mathbf{X}$ par $\mathbf{X} - \mathbf{1}\mathbf{g}'$, ce qui place l'origine en $\mathbf{g}$.

Ce que le centrage change, et ce qu'il ne change pas

Il ne change **aucune distance entre individus** : une translation les conserve toutes.

Il change **tout le reste** du calcul, parce que les axes factoriels passeront par l'origine. Un axe qui passe par le centre du nuage n'a rien à voir avec un axe qui passe à côté : c'est la différence entre une droite de régression et une droite arbitraire.

La métrique des variables

La distance générale

$$d_{\mathbf{M}}^2(i, \ell) = (\mathbf{x}_i - \mathbf{x}_\ell)' \mathbf{M} (\mathbf{x}_i - \mathbf{x}_\ell)$$

Avec $\mathbf{M} = \mathbf{I}$, on retrouve la distance euclidienne ordinaire.

Pourquoi l'identité est presque toujours un mauvais choix

Un fichier client mêle un panier moyen en dirhams — de 200 à 50 000 — et une note entre 0 et 20.

Avec $\mathbf{M} = \mathbf{I}$, la variance du panier se compte en millions, celle de la note en dizaines. **Le panier écrase tout** : le premier axe serait le panier, et le résultat changerait si on comptait en milliers de dirhams.

Un résultat qui dépend du choix de l'unité n'est pas un résultat.

Diviser par les écarts-types

$$\mathbf{M} = \mathbf{D}_{1/s^2} = \text{diag}\left(\frac{1}{s_1^2}, \dots, \frac{1}{s_p^2}\right)$$

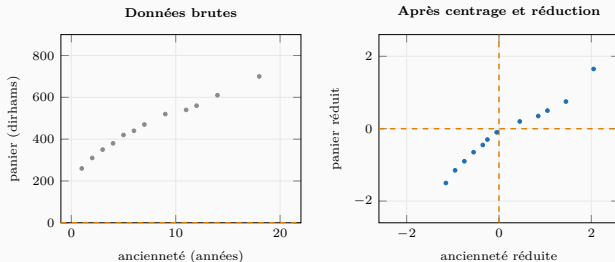
$$d^2(i, \ell) = \sum_{j=1}^p \frac{(x_{ij} - x_{\ell j})^2}{s_j^2}$$

Deux lectures de la même opération

Intuition

On peut dire qu'on **change la métrique** en gardant les données brutes, ou qu'on **réduit les données** en gardant la métrique identité. Les deux donnent exactement le même résultat.

La seconde formulation est plus commode à calculer; la première dit mieux ce qui se passe. **Réduire**, c'est décider que chaque variable pèse un.



La règle

On **norme** dès que les variables n'ont pas la même unité, ou que leurs dispersions sont sans commune mesure.

On ne norme **pas** quand les variables sont déjà homogènes et que les écarts de dispersion sont eux-mêmes une information.

Sur nos données, le cas est limite

Les huit notes sont sur la même échelle 0–20, et leurs écarts-types sont pratiquement égaux — tous compris entre 2,38 et 2,40.

Normer ou non **ne changera donc presque rien**, et c'est précisément ce qui rend le choix instructif : on norme malgré tout, parce que l'inertie totale vaut alors exactement 8, le nombre de variables. Toute lecture en devient immédiate — un axe à 4,54 porte $4,54/8 = 56,7\%$ de l'information.

Intuition

Cette quasi-égalité est une commodité de nos données, non une règle. Sur un fichier où le panier moyen côtoie une note de satisfaction, l'écart entre les deux analyses est **total** : ne pas normer y revient à n'analyser que le panier.

Actives et illustratives

La décision la plus lourde de conséquences

Les deux statuts

Une variable **active** construit les axes. Une variable **illustrative** ne participe à aucun calcul : elle est projetée après coup sur des axes construits sans elle.

client	8 notes de satisfaction				signalétique	
	prix	choix	...	accueil	CSP	fréquence
1	13,2	15,4	...	16,1	Cadre	Hebdo.
2	9,8	11,0	...	10,4	Employé	Rare
⋮	⋮	⋮		⋮	⋮	⋮
300	12,5	13,9	...	14,8	Retraité	Mensuelle

variables **actives** : elles construisent les axes variables **illustratives** : on les projette ensuite

Le choix des variables actives est la seule décision vraiment engageante de toute l'analyse.

Elles fixent la géométrie du nuage — donc les axes, donc l'interprétation. Les illustratives n'y participent pas : on les superpose après coup pour *éclairer* les axes, jamais pour les construire.

Mettre le chiffre d'affaires en actif quand on étudie la satisfaction, c'est laisser la taille des clients écraser leur opinion.

Pourquoi cette séparation est indispensable

L'argument, en une phrase

Si la fréquence de visite participait à la construction des axes, il serait **circulaire** de constater ensuite qu'elle les explique.

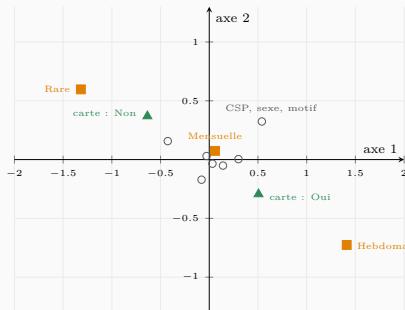
En la gardant illustrative, on construit les axes sur la seule satisfaction — puis on découvre que la fréquence s'y ordonne. Ce n'est plus une tautologie : **c'est un résultat**.

Le principe de choix

Sont **actives** les variables qui définissent le point de vue : ici, la satisfaction.

Sont **illustratives** celles qu'on veut **expliquer** par ce point de vue, et les variables qualitatives, qu'une ACP ne sait pas traiter autrement.

Ce que donne la projection



Deux variables parlent, cinq se taisent.

La **fréquences** s'étire le long de l'axe 1, de $-1,32$ à $+1,41$: *le client satisfait revient*. La **carte de fidélité** raconte la même histoire.

La CSP, le sexe et le motif de visite s'agglutinent au centre. Leurs coordonnées ne dépassent pas 0,5.

C'est un résultat, pas un échec. La satisfaction ne dépend pas de qui l'on est. Un service marketing qui segmente par âge et par CSP passe à côté de ce qui structure réellement sa clientèle.

Exemple

$$\mathbf{g}_q = \frac{1}{n_q} \sum_{i \in q} \mathbf{F}_i$$

La modalité *Hebdomadaire* (101 clients) tombe en $F_1 = 1,41$, la modalité *Rare* (112 clients) en $F_1 = -1,32$. Les deux sexes tombent en 0,03 et -0,03 : **confondus avec l'origine.**

Ce qu'il faut retenir

Cinq points

1. Une analyse factorielle est définie par un **triplet** $(\mathbf{X}, \mathbf{M}, \mathbf{D})$ – jamais par le tableau seul.
2. **D** porte les **poids des lignes**, **M** la **métrique des colonnes**.
3. ACP et AFC ne diffèrent que par **M** et **D** : **le calcul qui suit est identique**.
4. **Normer**, c'est poser $\mathbf{M} = \mathbf{D}_{1/s^2}$, donc décider que chaque variable pèse un.
5. Le partage **actives** / **illustratives** décide de ce que l'analyse peut démontrer sans se contredire.

La suite

Le triplet étant posé, on peut enfin définir la quantité que toutes les méthodes du cours maximisent : **l'inertie**. C'est l'objet du chapitre 3.