

Analyse des données avancée

Chapitre 5 : L'ACP en écriture matricielle



Jaouad Madkour

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

Le problème et sa solution

Les composantes principales

Le résultat sur nos données

Ce qu'il faut retenir

Le problème et sa solution

Le programme d'optimisation

Chercher la direction $\mathbf{a}$ qui maximise l'inertie du nuage projeté :

$$\max_{\mathbf{a}} \mathbf{a}'\mathbf{V}\mathbf{a} \quad \text{sous la contrainte} \quad \mathbf{a}'\mathbf{a} = 1$$

où $\mathbf{V} = \mathbf{X}'\mathbf{D}\mathbf{X}$, sur données centrées.

Pourquoi la contrainte est indispensable

Sans elle, le problème n'a pas de solution : doubler $\mathbf{a}$ quadruple $\mathbf{a}'\mathbf{V}\mathbf{a}$, et l'on peut monter indéfiniment.

La contrainte dit qu'on cherche une **direction**, non une longueur. C'est exactement ce qu'on veut : l'axe, pas l'échelle.

Le lagrangien

$$\mathcal{L}(\mathbf{a}, \lambda) = \mathbf{a}'\mathbf{V}\mathbf{a} - \lambda(\mathbf{a}'\mathbf{a} - 1)$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{a}} = 2\mathbf{V}\mathbf{a} - 2\lambda\mathbf{a} = \mathbf{0} \quad \Longleftrightarrow \quad \boxed{\mathbf{V}\mathbf{a} = \lambda\mathbf{a}}$$

Ce que la condition du premier ordre vient de dire

La direction cherchée est un **vecteur propre** de $\mathbf{V}$. Ce n'était pas une convention de calcul : c'est la conclusion d'une maximisation.

Et laquelle prendre ? En multipliant à gauche par $\mathbf{a}'$:

$$\mathbf{a}'\mathbf{V}\mathbf{a} = \lambda\mathbf{a}'\mathbf{a} = \lambda$$

L'inertie projetée sur l'axe est la valeur propre elle-même. On prend donc la plus grande.

La récurrence

Le deuxième axe maximise la même quantité sous **deux** contraintes : $\mathbf{a}'\mathbf{a} = 1$ et $\mathbf{a}'\mathbf{a}_1 = 0$.

La solution est le vecteur propre associé à λ_2 , la deuxième valeur propre par ordre décroissant. Et ainsi de suite.

L'orthogonalité était offerte

On aurait pu craindre que la contrainte d'orthogonalité complique le problème. Il n'en est rien : le théorème spectral garantit que les vecteurs propres d'une matrice symétrique sont **déjà** orthogonaux.

La contrainte est donc automatiquement satisfaite. C'est pourquoi l'ACP se calcule d'un coup, et non axe après axe.

Les composantes principales

Définition et propriétés

La composante principale de rang k

$$\mathbf{F}_k = \mathbf{Z} \mathbf{u}_k$$

où $\mathbf{Z}$ est le tableau centré réduit et $\mathbf{u}_k$ le k -ième vecteur propre de $\mathbf{R}$.

Ses trois propriétés

$$\bar{F}_k = 0 \quad V(\mathbf{F}_k) = \lambda_k \quad \text{cov}(\mathbf{F}_k, \mathbf{F}_m) = 0 \quad (k \neq m)$$

Ce que la troisième propriété change

Les huit notes de départ sont fortement corrélées entre elles — c'est même tout le problème.

Les composantes, elles, sont **rigoureusement non corrélées**. On a remplacé huit variables redondantes par huit variables indépendantes deux à deux, rangées par ordre d'importance décroissante. **C'est cela, changer de point de vue.**

La matrice des corrélations de départ

	prix	choix	qual.	agen.	prop.	rapid.	cons.	accu.
prix	1	0,67	0,67	0,59	0,43	0,22	0,15	0,09
choix	0,67	1	0,73	0,69	0,56	0,32	0,34	0,31
qualité	0,67	0,73	1	0,71	0,56	0,41	0,36	0,35
agencement	0,59	0,69	0,71	1	0,59	0,50	0,44	0,41
propreté	0,43	0,56	0,56	0,59	1	0,64	0,59	0,58
rapidité	0,22	0,32	0,41	0,50	0,64	1	0,69	0,71
conseil	0,15	0,34	0,36	0,44	0,59	0,69	1	0,73
accueil	0,09	0,31	0,35	0,41	0,58	0,71	0,73	1

L'ACP est déjà là, dans ce tableau.

Deux blocs sombres se détachent : **prix, choix, qualité, agencement** d'un côté — ce que le magasin *propose* — et **propreté, rapidité, conseil, accueil** de l'autre — ce que le personnel *fait*. Entre les deux, les corrélations tombent à 0,09.

L'ACP ne fera que rendre visible cette structure — mais sur 30 variables, aucun œil ne l'aurait vue dans le tableau.

Intuition

Toutes les corrélations sont **positives** : un client sévère sur un critère l'est sur tous. Cela donnera un premier axe où les huit notes seront du même côté.

Deux blocs apparaissent pourtant : prix-choix-qualité-agencement d'un côté, propreté-rapidité-conseil-accueil de l'autre, moins liés entre blocs qu'à l'intérieur. Cela donnera le second axe. **Les deux axes se lisaient déjà dans le tableau des corrélations.**

Le résultat sur nos données

Les huit valeurs propres de $\mathbf{R}$

Rang	λ_k	Part	Cumul
1	4,535	56,69	56,69
2	1,645	20,57	77,26
3	0,382	4,77	82,03
4	0,349	4,37	86,40
5	0,327	4,09	90,49
6	0,283	3,54	94,02
7	0,249	3,11	97,14
8	0,229	2,86	100,00

Somme : 8,000, comme le chapitre 3 l'exigeait.

Les corrélations avec les axes

note	r axe 1	r axe 2	$\cos^2$ plan	ctr 1	ctr 2
prix	0,629	0,624	0,784	8,7 %	23,6 %
choix	0,771	0,438	0,787	13,1 %	11,7 %
qualité	0,803	0,380	0,789	14,2 %	8,8 %
agencement	0,828	0,242	0,744	15,1 %	3,6 %
propreté	0,827	-0,127	0,701	15,1 %	1,0 %
rapidité	0,742	-0,476	0,778	12,2 %	13,8 %
conseil	0,711	-0,531	0,787	11,1 %	17,1 %
accueil	0,689	-0,580	0,811	10,5 %	20,5 %
total				100 %	100 %

Deux colonnes, deux questions différentes — et on les confond sans cesse.

Le $\cos^2$ répond à : *cette variable est-elle bien représentée sur le plan ?* Il se lit par ligne, et vaut au plus 1.

La **contribution** répond à : *cette variable a-t-elle fabriqué cet axe ?* Elle se lit par colonne, et les colonnes somment à 100 %.

Une variable peut être très bien représentée sans avoir construit l'axe — et une contribution moyenne vaut ici $100/8 = 12,5\%$: au-delà, la variable compte.

Ce que les deux axes signifient

Axe 1 : le niveau général de satisfaction

Les **huit** corrélations sont positives et comprises entre **0,63** et **0,83**.

Un tel axe s'appelle un **effet taille** : il oppose les clients globalement satisfaits aux clients globalement mécontents, sans distinguer les critères. Il porte à lui seul **56,7 %** de l'information — autrement dit, **la première chose que le questionnaire mesure est l'humeur du répondant**.

Axe 2 : l'offre contre le service

Il oppose **prix, choix, qualité, agencement** à **rapidité, conseil, accueil**.

Ce sont **les produits contre les personnes**. À satisfaction générale égale, certains clients viennent pour l'assortiment, d'autres pour la relation. Voilà une distinction que huit moyennes n'auraient jamais montrée — et sur laquelle une politique commerciale peut se construire.

ACP normée sur quatre individus

$$\mathbf{R} = \begin{pmatrix} 1 & 0,6 \\ 0,6 & 1 \end{pmatrix} \quad \lambda_1 = 1,6, \quad \lambda_2 = 0,4$$

Somme : $2 = p$. Premier axe : $1,6/2 = 80\%$.

Exemple

Sur $\mathbf{u}_1 = (1, 1)/\sqrt{2}$, avec $z_{ij} = x_{ij}/\sqrt{2,5}$:

Individu	F_1	F_2
1	-1,789	0
2	0	-0,894
3	0	0,894
4	1,789	0

Contrôle : $V(F_1) = \frac{1}{4}(3,2 + 0 + 0 + 3,2) = 1,6 = \lambda_1$.

Ce qu'il faut retenir

Six points

1. L'ACP maximise $\mathbf{a}'\mathbf{V}\mathbf{a}$ sous $\mathbf{a}'\mathbf{a} = 1$: c'est un lagrangien, rien de plus.
2. La condition du premier ordre **est** l'équation aux valeurs propres $\mathbf{V}\mathbf{a} = \lambda\mathbf{a}$.
3. **L'inertie portée par un axe est sa valeur propre** : $\mathbf{a}'\mathbf{V}\mathbf{a} = \lambda$.
4. Les composantes sont centrées, de variance λ_k , et **non corrélées entre elles**.
5. Sur nos données : $\lambda_1 = 4,535$ et $\lambda_2 = 1,645$, soit 77,3 % sur deux axes.
6. Axe 1 : la satisfaction générale. Axe 2 : l'offre contre le service.

La suite

Tout a été fait du côté des variables. Le chapitre 6 montre que le côté des individus **n'exige aucun calcul supplémentaire** — et pourquoi.