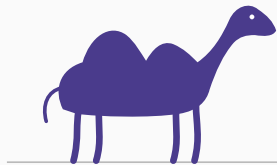


Introduction à l'analyse des données

Chapitre 5 : La corrélation linéaire



Jaouad Madkour

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

Rendre la covariance lisible

Ce que r vaut, et ce qu'il dit

La lecture géométrique

Les trois pièges

Ce qu'il faut retenir

Rendre la covariance lisible

Le coefficient de corrélation linéaire

$$r_{xy} = \frac{\text{cov}(x, y)}{s_x s_y}$$

On divise la covariance par le produit des deux écarts-types.

Pourquoi cette division précisément

La covariance s'exprime en « unité de x × unité de y », et le produit $s_x s_y$ **exactement** dans la même **unité**. Le quotient est donc **sans unité** — la condition pour comparer, établie au chapitre 3.

Sur le fil rouge

$$s_x s_y = \sqrt{8} \times \sqrt{32} = \sqrt{8 \times 32} = \sqrt{256} = 16$$

$$r_{xy} = \frac{14,4}{16} = \mathbf{0,9}$$

Le produit des écarts-types tombe **exactement** sur 16, et la corrélation vaut **exactement** 0,9.

Ce que r vaut, et ce qu'il dit

Le résultat central

$$-1 \leq r_{xy} \leq +1$$

Cette borne ne dépend d'aucune hypothèse : elle est une propriété algébrique du quotient.

La lecture

Valeur de r	Situation
+1	les points sont exactement alignés , pente positive
proche de +1	liaison positive forte
0	aucune liaison linéaire
proche de -1	liaison négative forte
-1	alignement exact, pente négative

Application économique

$r = 0,9$ est une liaison **très forte**. Accueil et fidélité vont ensemble de façon presque parfaitement linéaire.

Sur des données réelles, un tel r entre deux critères est d'ailleurs suspect : il suggère que les deux questions **mesurent la même chose**. L'analyse en composantes principales exploitera précisément ces redondances.

La lecture géométrique

La corrélation est une covariance de variables réduites

Une écriture équivalente

Si z_x et z_y désignent les variables centrées réduites du chapitre 3 :

$$r_{xy} = \text{COV}(z_x, z_y) = \frac{1}{n} \sum_{i=1}^n z_{x,i} z_{y,i}$$

Ce que cela révèle

Corréler, c'est **réduire d'abord, covarier ensuite**.

Les trois chapitres de rappel se referment ici : centrer et réduire (chapitre 3), covarier (chapitre 4), et le quotient du chapitre 5 n'est que la combinaison des deux.

Une fois les variables réduites, covariance et corrélation sont le même nombre. C'est pourquoi l'analyse en composantes principales travaillera systématiquement sur données réduites : elle n'aura plus à choisir entre les deux.

La matrice des corrélations

La matrice

$$R = \begin{pmatrix} 1 & 0,9 \\ 0,9 & 1 \end{pmatrix}$$

Sa diagonale ne contient que des 1 : $\text{cov}(z, z) = V(z) = 1$.

	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.

Application économique

Sur un fichier réel, la matrice des corrélations se regarde **avant** de lancer quoi que ce soit :

- des blocs de variables fortement corrélées entre elles signalent des **groupes de critères redondants** ;
- une variable corrélée à aucune autre n'apportera rien aux premiers axes ;
- une corrélation proche de 1 entre deux variables indique qu'on en mesure une deux fois.

C'est déjà de l'analyse des données, faite à l'œil.

Les trois pièges

Premier piège : la liaison non linéaire

r ne mesure que la liaison **linéaire**.

Deux variables liées par une relation en cloche parfaitement déterministe peuvent avoir une corrélation **nulle**. Un r de zéro signifie « pas de liaison **linéaire** », jamais « pas de liaison ».

Il faut donc toujours regarder le nuage de points, et ne jamais se contenter du coefficient.

Deuxième piège : les points aberrants

Un seul point très éloigné peut créer une corrélation forte là où il n'y en a aucune, ou détruire une corrélation réelle.

Le coefficient est une moyenne : il est sensible aux valeurs extrêmes, comme toute moyenne.

Application économique

Une corrélation forte n'établit **aucun** lien de cause à effet. Trois explications restent ouvertes :

1. x agit sur y ;
2. y agit sur x ;
3. une **troisième variable** agit sur les deux.

Le troisième est le plus fréquent : l'ancienneté du client, par exemple, peut produire à elle seule la corrélation entre accueil et fidélité.

Les distinguer est l'objet du cours d'**inférence causale**, au Master. Celui-ci décrit des associations, et s'y tient.

Ce qu'il faut retenir

Cinq points

1. $r_{xy} = \text{cov}(x, y) / (s_x s_y)$: la covariance débarrassée de ses unités.
2. Sur le fil rouge : $s_x s_y = \sqrt{256} = 16$, donc $r = 14,4 / 16 = \mathbf{0,9}$ exactement.
3. r est toujours compris entre -1 et $+1$.
4. La corrélation est la **covariance des variables réduites** — les trois rappels se referment ici.
5. Trois pièges : liaison **non linéaire**, points **aberrants**, et **causalité**.

La suite

Les rappels sont terminés. Le chapitre 6 change de regard : il cesse de voir un tableau de chiffres et se met à voir un **nuage de points** dans un espace à p dimensions.