

Analyse des données avancée

Chapitre 1 : Reprendre l'analyse des données par l'algèbre



Jaouad Madkour

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

Prérequis

Avant de commencer

Explorer et confirmer

Les deux fils rouges

Ce qu'il faut retenir

Prérequis

Ce que ce cours suppose acquis

Les cours qui précèdent

Introduction à l'analyse des données et *Algèbre linéaire*.

Les notions mobilisées

De l'analyse des données : la variable centrée réduite, la covariance, l'inertie, l'idée de l'ACP et de l'AFC.

De l'algèbre linéaire, cette fois exigée : la **diagonalisation** d'une matrice symétrique, les **formes quadratiques**, la **projection orthogonale**.

Repris depuis le début

La décomposition en valeurs singulières est établie dans le cours : elle n'est pas supposée connue.

Ce cours reprend l'intégralité de l'analyse des données, mais avec les démonstrations que la Licence avait admises. Il ne se contente pas d'ajouter des méthodes : il refait les premières autrement.

Avant de commencer

Ce que ce cours reprend, et pourquoi

Le même objet, un autre niveau

Le cours *Introduction à l'analyse des données*, en Licence, a montré **ce que font** les méthodes factorielles : réduire la dimension en perdant le moins d'information possible.

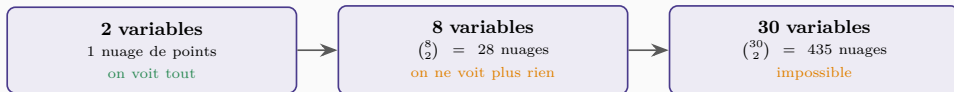
Ce cours reprend l'ensemble depuis le début pour montrer **pourquoi elles le font**, et **comment** — puis le prolonge jusqu'à la classification, la discrimination et le positionnement.

La différence tient en une phrase

En Licence, on a admis que les axes factoriels étaient les vecteurs propres d'une certaine matrice.

Ici, on le **démontre** — et l'on découvre au passage que l'analyse en composantes principales, l'analyse des correspondances et le positionnement multidimensionnel sont **le même calcul**, appliqué à trois métriques différentes.

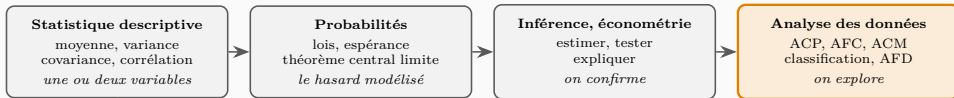
Le problème, rappelé en une image



Et le problème n'est pas seulement le nombre de graphiques.

Vingt-huit nuages pris deux à deux ne montrent que des liaisons *par paires*. Une structure qui n'apparaît qu'en combinant quatre variables reste invisible sur les vingt-huit.

L'analyse des données ne fait pas plus de graphiques : elle en fait un seul, et le bon.



Ce cours prolonge la statistique descriptive, il ne prolonge pas l'inférence.

Aucune loi de probabilité n'y intervient, aucun test, aucune p -value. On ne parle jamais de population ni d'échantillon : **tout se passe dans le tableau que l'on a sous les yeux**, exactement comme en descriptive.

Ce qui change est la dimension. La descriptive s'arrêtait à deux variables — la covariance, le nuage, le coefficient de corrélation. L'analyse des données reprend ces trois objets et les fait travailler sur trente.

Explorer et confirmer

Deux démarches, deux logiques

Démarche confirmatoire

Statistique inférentielle, économétrie

une hypothèse **posée d'avance**



les données tranchent

on sait ce qu'on cherche

Démarche exploratoire

Analyse des données

aucune hypothèse préalable



les données suggèrent

on cherche ce qu'on ne sait pas

Les deux démarches ne s'opposent pas : elles se suivent.

L'exploration fait *naître* les hypothèses ; la confirmation les *juge*. Une segmentation obtenue par classification n'est pas un résultat prouvé : c'est une hypothèse de travail, qu'il faudra tester sur d'autres données.

Le danger est de présenter comme démontré ce qui n'a été qu'observé. Aucune *p*-value ne sortira de ce cours ; ce n'est pas un oubli.

Aucun test, aucune probabilité

Rien de ce cours n'est un test statistique. Il n'y a ni hypothèse nulle, ni p -valeur, ni intervalle de confiance.

Les méthodes décrivent **le tableau qu'on leur donne**. Elles ne disent rien d'une population dont ce tableau serait un échantillon.

La conséquence pratique, souvent oubliée

Un axe factoriel n'est pas « significatif ». Une segmentation n'est pas « vraie ». Ce sont des **résumés**, dont la seule justification est d'être **utiles et reproductibles**.

D'où la règle qui gouverne tout le cours : **une structure qu'une seule méthode voit n'est pas une structure**. On la confirmera par une seconde méthode, ou on ne la retiendra pas.

La carte des méthodes

	Variables quantitatives	Variables qualitatives	
Décrire	ACP chapitres 5 à 8	AFC et ACM chapitres 9 à 11	
Classer	CAH et <i>k</i>-means chapitres 12 à 15	les mêmes, après une AFC ou une ACM	MDS à partir de simples distances chapitre 18
Prédire	AFD chapitres 16 et 17	régression logistique <i>hors programme</i>	

Une seule question commande le choix : quelle est la nature des variables, et que veut-on en faire ?

Décrire un nuage, y former des groupes, ou séparer des groupes déjà connus. Les six méthodes du cours remplissent ce tableau — il n'y a pas d'autre case.

Les deux fils rouges

Le fichier qui traverse tout le cours

Une enseigne interroge **300** clients. Chacun note **huit critères** de 0 à 20 : prix, choix, qualité, agencement, propreté, rapidité, conseil, accueil.

S'y ajoutent des variables **signalétiques** — sexe, âge, catégorie socioprofessionnelle, fréquence de visite, ancienneté, panier moyen, motif — qui ne serviront pas à construire les axes, mais à les interpréter.

Ce qu'on en attend

Aucune question n'est posée d'avance. On demande au tableau **ce qu'il contient**, et l'on verra qu'il contient deux choses seulement : un niveau général de satisfaction, et une opposition entre deux familles de critères.

Tout le reste — les $300 \times 8 = 2\,400$ nombres — est du détail individuel.

Le tableau qui se calcule à la main

Le fichier de 300 clients ne se diagonalise pas au tableau. Un second exemple, minuscule, servira à écrire les calculs en entier :

Individu	x_1	x_2
1	-2	-2
2	-1	1
3	1	-1
4	2	2

Les deux colonnes sont **déjà centrées** : leurs sommes valent zéro.

Intuition

Elles ont été choisies pour que tout tombe juste. On verra apparaître $\mathbf{X}'\mathbf{X} = \begin{pmatrix} 10 & 6 \\ 6 & 10 \end{pmatrix}$, de valeurs propres 16 et 4, donc de **valeurs singulières 4 et 2**.

Aucune décimale, aucune calculatrice : c'est la seule façon de voir la mécanique.

Ce qu'il faut retenir

Cinq points

1. Ce cours reprend l'analyse des données **depuis le début**, en donnant les démonstrations que la Licence avait admises.
2. La démarche est **exploratoire** : aucune hypothèse préalable, donc aucun test.
3. Une structure vue par une seule méthode ne se retient pas : **il faut la recouper**.
4. Deux fils rouges : **300** clients et 8 notes pour les résultats, 4 individus et 2 variables pour les calculs.
5. Le fichier des **300** clients est réel et fourni : **tous les chiffres du cours sont reproductibles**.

La suite

Avant de calculer quoi que ce soit, il faut dire précisément ce sur quoi l'on calcule : un tableau, des **poids** sur les lignes, une **métrique** sur les colonnes. C'est l'objet du chapitre 2, et c'est là que se jouent la plupart des choix qu'un logiciel fait à notre place.