

Analyse des données avancée

Chapitre 6 : Dualité et relations de transition



Jaouad Madkour

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

Les deux nuages

Le théorème de dualité

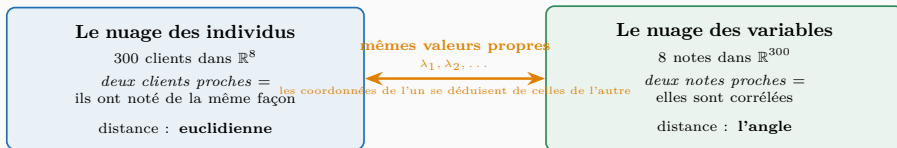
Les relations de transition

La vérification sur nos données

Ce qu'il faut retenir

Les deux nuages

Un tableau, deux lectures



Une seule ACP, deux lectures — c'est ce qu'on appelle la dualité.

On ne fait pas deux calculs : la décomposition de la matrice des corrélations donne *en même temps* les axes des individus et ceux des variables. Le pourcentage d'inertie est donc le même des deux côtés.

D'où la méthode de lecture, invariable : on interprète d'abord le *cercle des corrélations* — il donne le *sens* des axes — puis le *plan des individus* — il dit *qui* se trouve où. Jamais l'inverse.

Ce qui distingue les deux nuages

Le nuage des individus

n points dans $\mathbb{R}^p$. Deux points sont proches quand les individus ont **répondu de la même façon**.

La métrique est euclidienne — sur données réduites, $d^2(i, \ell) = \sum_j (z_{ij} - z_{\ell j})^2$.

Le nuage des variables

p points dans $\mathbb{R}^n$. Sur données réduites, chaque variable est un vecteur de **norme 1**, et

$$\cos(\mathbf{z}_j, \mathbf{z}_m) = r(x_j, x_m)$$

Ce qui compte n'est donc pas la distance, mais **l'angle** : la corrélation est un cosinus.

Le point à ne pas manquer

Ces deux nuages ne vivent pas dans le même espace, ne se mesurent pas de la même façon, et ne répondent pas à la même question.

Rien n'annonce qu'ils aient un rapport. Et pourtant ils ont **les mêmes valeurs propres**.

Le théorème de dualité

Les deux matrices

$$\mathbf{R} = \frac{1}{n} \mathbf{Z}' \mathbf{Z} \quad (p \times p) \qquad \mathbf{W} = \frac{1}{n} \mathbf{Z} \mathbf{Z}' \quad (n \times n)$$

Elles ont les mêmes valeurs propres non nulles, en nombre égal au rang de $\mathbf{Z}$.

La démonstration, en trois lignes

Soit $\mathbf{R}\mathbf{u} = \lambda\mathbf{u}$ avec $\lambda \neq 0$. Multiplions à gauche par $\mathbf{Z}$:

$$\frac{1}{n} \mathbf{Z} \mathbf{Z}' \mathbf{Z} \mathbf{u} = \lambda \mathbf{Z} \mathbf{u} \quad \text{soit} \quad \mathbf{W} (\mathbf{Z} \mathbf{u}) = \lambda (\mathbf{Z} \mathbf{u})$$

Donc $\mathbf{Z} \mathbf{u}$ est vecteur propre de $\mathbf{W}$ pour la même valeur propre. La réciproque s'obtient en multipliant par $\mathbf{Z}'$.

Intuition

Le vecteur propre de $\mathbf{W}$ n'est pas seulement associé à la même valeur propre : c'est $\mathbf{Zu}$, c'est-à-dire la composante principale elle-même.

Autrement dit, l'analyse des individus ne fournit rien de neuf. Elle **est** l'analyse des variables, lue de l'autre côté.

Pourquoi les logiciels diagonalisent toujours la plus petite

R est 8×8 . **W** est 300×300 .

Le coût d'une diagonalisation croît comme le cube de la taille : $8^3 = 512$ contre $300^3 = 27$ millions.

On diagonalise **R**, puis on **déduit** le nuage des individus.

Sur un fichier de 100 000 clients et 20 variables, la dualité n'est plus une élégance : c'est la seule façon de calculer quoi que ce soit.

Les relations de transition

Les deux formules

$$\mathbf{F}_k = \frac{1}{\sqrt{\lambda_k}} \mathbf{Z} \mathbf{G}_k \quad \mathbf{G}_k = \frac{1}{\sqrt{\lambda_k}} \frac{1}{n} \mathbf{Z}' \mathbf{F}_k$$

où $\mathbf{F}_k$ porte les coordonnées des individus et $\mathbf{G}_k$ celles des variables.

Ce qu'elles disent en français

Un individu se place là où le tirent les variables sur lesquelles il note haut. Un client qui donne de bonnes notes à tout est tiré du côté où toutes les variables se trouvent.

Une variable se place là où sont les individus qui y notent haut. Chaque formule est une moyenne pondérée de l'autre, à un facteur $1/\sqrt{\lambda_k}$ près.

Ce que le facteur d'échelle empêche

Le facteur $1/\sqrt{\lambda_k}$

Sans lui, les deux nuages ne seraient pas à la même échelle et la transition serait fausse.

Avec lui, on peut superposer les deux représentations — c'est le **biplot** — mais **jamais lire une distance entre un individu et une variable**.

L'erreur de lecture la plus fréquente

Sur un biplot, un client tombe près de la variable *accueil*. On conclut qu'il a bien noté l'accueil.

C'est illégitime. Les deux nuages sont dans des espaces différents; leur superposition est une commodité graphique. Seule la **direction** commune se lit : le client est du même côté de l'origine que la variable.

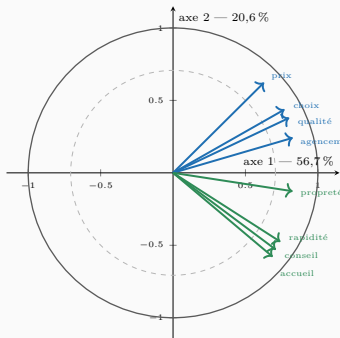
En AFC, au chapitre 10, la superposition sera au contraire **pleinement justifiée** — et c'est la seule méthode du cours où elle le soit.

Ce que le cercle des corrélations est vraiment

La coordonnée d'une variable

$$G_{jk} = r(x_j, \mathbf{F}_k)$$

La coordonnée d'une variable sur l'axe k est sa corrélation avec la composante k .



Trois règles de lecture, et elles ne se discutent pas.

1. La longueur mesure la qualité de représentation. Toutes les flèches approchent le cercle : les huit notes sont bien restituées par le plan.

2. L'angle mesure la corrélation. *Prix* et *accueil* forment un angle proche de 90° : ces deux jugements sont indépendants.

3. Le côté donne le sens. Toutes les flèches pointent à droite : **l'axe 1 est la satisfaction d'ensemble**. L'axe 2 sépare l'offre du **service**.

Une flèche courte ne s'interprète pas : elle est ailleurs.

La vérification sur nos données

Les deux calculs coïncident

Diagonaliser $\mathbf{R}$, puis transiter

$\mathbf{R}$ donne $\lambda_1 = 4,535$ et le vecteur propre $\mathbf{u}_1$.

La transition $\mathbf{F}_1 = \mathbf{Z}\mathbf{u}_1$ place les 300 clients. On vérifie alors

$$\frac{1}{300} \sum_{i=1}^{300} F_{i1}^2 = 4,535 = \lambda_1$$

La variance des coordonnées des individus est la valeur propre issue des variables. Les deux nuages sont bien deux faces d'un seul objet.

Et dans l'autre sens

$$r(\text{prix}, \mathbf{F}_1) = 0,629 = G_{\text{prix},1}$$

La corrélation calculée directement entre la variable brute et la composante retombe sur la coordonnée lue dans le cercle — à la quatrième décimale.

Quatre individus, deux variables

Côté variables : $\mathbf{R}$ est 2×2 , valeurs propres 1,6 et 0,4.

Côté individus : $\mathbf{W} = \frac{1}{4}\mathbf{ZZ}'$ est 4×4 . Elle a **quatre** valeurs propres, mais seules deux sont non nulles — et ce sont 1,6 et 0,4.

Pourquoi deux valeurs propres nulles

Le rang de $\mathbf{Z}$ vaut 2 : quatre points de $\mathbb{R}^2$ ne peuvent engendrer qu'un plan.

En règle générale, le nombre de valeurs propres non nulles vaut $\min(n - 1, p)$ — le -1 venant du centrage, qui consomme une dimension.

Ce qu'il faut retenir

Six points

1. Un tableau définit **deux nuages** : les individus dans $\mathbb{R}^p$, les variables dans $\mathbb{R}^n$.
2. **$\mathbf{R}$ et $\mathbf{W}$ ont les mêmes valeurs propres non nulles** — c'est la dualité.
3. On diagonalise **la plus petite des deux**, puis on déduit l'autre.
4. Les **relations de transition** font passer d'un nuage à l'autre, au facteur $1/\sqrt{\lambda_k}$ près.
5. La coordonnée d'une variable **est** sa corrélation avec la composante : d'où le cercle.
6. Sur un biplot d'ACP, **on ne lit pas une distance individu-variable** : seulement une direction commune.

La suite

Les axes existent, en nombre égal au rang du tableau. Reste la question qu'aucun théorème ne tranche : **combien en garder ?** C'est l'objet du chapitre 7.