Data Analysis & Visualization (ADDA71)

Stefani EL KALAMOUNI

Partie II

Méthodes d’Analyse Factorielle

Introduction aux méthodes factorielles

Introduction aux méthodes factorielles - Objectifs

- ✓ Comprendre le rôle des méthodes factorielles en analyse de données.
- ✓ Différencier réduction de dimension (ACP) et analyse exploratoire (AFC, ACM).
- ✓ Poser les bases statistiques nécessaires.

1. Qu’est-ce qu’un jeu de données ?

Définition

Un *jeu de données* (ou *tableau de données*) est généralement un tableau :

- **Lignes** = individus / observations (ex. un client, un pays, un élève, une image…).
- **Colonnes** = variables / caractéristiques (ex. âge, revenu, taille, poids, note, catégorie…).

Exemple. 100 étudiants (100 lignes) × 5 variables (taille, poids, âge, spécialité, réussite).

Types de variables.

- **Quantitatives** : nombres mesurables (taille, revenu).
- **Qualitatives** (catégorielles) : étiquettes/valeurs discrètes (sexe, spécialité, pays).

Vocabulaire minimal.

- **Dimension p** : nombre de variables utilisées en même temps. Si on travaille avec 7 variables → dimension = 7.
- **Matrice de données** : notée X, de taille n×p (n lignes, p colonnes).
- **Centre d’une variable** : sa moyenne
- **écart-type d’une variable** : sa dispersion.

Qu’est-ce qu’une AFC ?

L’**AFC** résume et **visualise les relations** entre **deux variables qualitatives** (catégorielles) à partir d’un **tableau croisé** (tableau de contingence).

•**Lignes** = modalités d’une variable A (ex. *Pays*).

•**Colonnes** = modalités d’une variable B (ex. *Plat préféré*).

•**Cellules** = **comptages** (effectifs observés).

L’AFC construit une **carte 2D** (ou 3D) où :

•des **lignes proches** ont des **profils de colonne** semblables,

•des **colonnes proches** ont des **profils de ligne** semblables,

•une **ligne proche d’une colonne** indique une **association** plus forte que prévu si A et B étaient **indépendantes**.

En français courant : l’AFC **rapproche ce qui va ensemble** et **éloigne ce qui s’oppose**.

À quoi ça sert ?

- ✓ **Explorer des liaisons** entre catégories (marketing, sociologie, santé, éducation…).
- ✓ **Comprendre des profils** : “quels clients achètent par quel canal ?”, “quels pays consomment quels produits ?”…
- ✓ **Communiquer visuellement** : une carte facile à commenter (qui va avec qui).

Quand l’utilise-t-on (et quand non) ?

✅ Deux variables **qualitatives** (catégorielles) ⇒ **AFC**.

✅ Un **tableau de contingence** (comptages) est disponible ou facile à construire (pandas crosstab).

❌ Plusieurs variables qualitatives (≥3) ⇒ préférer **ACM** (Analyse des Correspondances Multiples).

❌ Variables **numériques** ⇒ **ACP** (Analyse en Composantes Principales).

De quoi a-t-on besoin ?

1.Un **tableau croisé** (I×J) d'**effectifs** N_{ij}

2.Les **totaux** de lignes/colonnes et le **total général** n.

3.(Option) Des **étiquettes** claires pour lignes/colonnes (pour commenter).

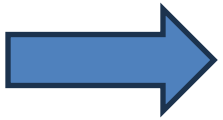
Rappels utiles :

- Profil de ligne i : proportions de la ligne i sur les colonnes.
- Profil de colonne j : proportions de la colonne j sur les lignes.
- Masses (poids) : part d'une ligne (ou colonne) dans le total (totaux/n).

Comment l'appliquer (étapes, logique)

1. **Construire** le tableau croisé d'effectifs N.

ID	sexe	Pays
123A	homme	FR
124B	femme	FR
125D	homme	BE
456F	femme	BE
788Y	femme	BE



	FR	BE	total
homme	1	1	2
femme	1	2	3
total	2	3	5

1. **Lancer** l'AFC (logiciel/Python) → obtenir :
    1. **Inertie par axe** (analogue de "variance expliquée"),
    2. **Coordonnées** des **lignes** et des **colonnes** sur Dim1, Dim2 (et plus).

    Pour calculer l'inertie il nous faut les valeurs propres, ensuite la somme des valeurs propres,

    l'inertie de chaque valeur propre $\textbf{Inertie}_{\lambda i}=\frac{\lambda_i}{\Sigma\lambda}$ (la valeur propre/la somme des valeurs propres*100)

Comment l'appliquer (étapes, logique)

3. **Choisir** combien d'axes regarder :
    1. **Scree** (courbe d'inertie par axe) : chercher le **coude** ;
    2. **Inertie cumulée** : viser un **compromis** (en AFC, 2–3 axes suffisent souvent pour la lecture).
4. **Tracer** la **carte lignes/colonnes (Dim1–Dim2)**.

Si nous avons une seule valeur propre, l'inertie est égale à 100% nous avons donc une dimension et un seul axe (pour dessiner ce sera des points)

5. **Lire/interpréter** :
    1. **Proximité** (ligne–ligne, colonne–colonne),
    2. **Ligne–colonne proche** ⇒ **association**,
    3. **Oppositions** de part et d'autre d'un axe,
    4. **cos²** (qualité de représentation), **contributions** (qui façonne l'axe).
6. **Conclure** en langage simple (qui va avec qui, qui s'oppose à qui, message métier).

Exemple concret très simple

	Pâtes	Sushi	Burger
FR	30	5	15
IT	35	3	12
JP	2	40	8

Lecture intuitive avant les calculs :

✓ FR et IT mangent surtout Pâtes → profils similaires → proches sur la carte.

✓ JP mange surtout Sushi → JP proche de Sushi.

✓ Burger est “au milieu” (un peu partout) → plus central, moins discriminant.

Ce que l’AFC va montrer :

✓ Dim1 opposera probablement *(FR/IT & Pâtes)* à *(JP & Sushi)*.

✓ Dim2 servira d’axe fin (ex. position relative de *Burger*).

Les graphiques et comment les lire

(A) Scree plot (inertie par axe)

- But : décider combien d'axes sont utiles.
- Lecture : chercher le coude (après, chaque axe apporte peu).
- Remarque : en AFC, l'inertie est souvent répartie (ne pas s'attendre à 90% en 2D).

(B) Carte lignes/colonnes (Dim1–Dim2)

- Points ronds = lignes (modalités de A), carrés = colonnes (modalités de B).
- Proximité ligne–ligne / colonne–colonne : profils similaires.
- Proximité ligne–colonne : association forte (observé > attendu).
- Opposition : de part et d'autre d'un axe → profils qui s'opposent.
- Origine (0,0) ≈ profil moyen (modalité peu discriminante si proche de l'origine).

Étape 1 — Construire le tableau croisé

•Lignes = **A** (ex. Pays), colonnes = **B** (ex. Plat), cellules = **comptages**.

Étape 2 — Profils & indépendance

•**Profil de ligne** : “dans ce pays, comment se répartissent les plats ?”

•**Indépendance** = ce qu’on verrait **sans lien** entre A et B.

•**AFC** regarde **où** l’observé **s’éloigne** de l’attendu (pondéré correctement).

Étape 3 — Axes (Dim1, Dim2)

•**Dim1** : grande **opposition** de profils (l’histoire principale).

•**Dim2** : **nuance** secondaire.

•**Carte** : on place **lignes** et **colonnes** ensemble pour voir **qui va avec qui**.

Étape 4 — Lire la carte

- **Proximité** ligne–ligne / colonne–colonne ⇒ profils **semblables**.
- **Ligne près d'une colonne** ⇒ **association** forte.
- **Oppositions** : de part et d'autre d'un axe.
- **Origine** ⇒ modalités **peu discriminantes**.

Étape 5 — Décider combien d'axes

- **Scree** : coude après Dim2 ? → **2 axes**.
- **Cumul** : la 2D **suffit pour raconter** ? sinon, citer **Dim3**.
- **Interprétabilité** : si un axe ne raconte **rien de clair**, on **l'ignore**.

Étape 6 — Conclure en langage simple

- **Qui va avec qui**, **qui s'oppose**, **quelle histoire métier** (ex. *FR/IT = Pâtes, JP = Sushi, Burger = moins discriminant*).