Support de cours

# ANALYSE DES DONNEES I

## CHAPITRE 2 : L'ANALYSE EN COMPOSANTES PRINCIPALES

Iphygénie SARR
*Ingénieure statisticienne économiste*

Ecole Nationale de la Statistique et de l'Analyse Economique Pierre Ndiaye
Mars 2024

# OBJECTIF DU CHAPITRE

L’objectif de ce chapitre est de décrire l’Analyse en Composantes Principales (ACP). Elle est la plus ancienne des méthodes d’analyses factorielles et s’applique à des bases de données comprenant exclusivement des variables quantitatives.

# PLAN DU CHAPITRE

# 01.

# Principe d’une ACP et notion d’ACP réduite

# Principe d'une ACP

L'ACP est une méthode d'analyse factorielle qui cherche à ressortir graphiquement les ressemblances entre individus et les liaisons entre variables. De façon spécifique, elle vise à :

- repérer les individus qui se ressemblent le plus vis-à-vis des variables ;
- relever les différences entre individus et mettre en évidence ceux dont le comportement est atypique par rapport à l'ensemble des variables ;
- savoir si l'information brute ne pourrait pas être obtenue à partir d'un nombre restreint de variables ;
- décrire simultanément les liaisons multiples entre les variables.

# Principe d'une ACP

L'ACP s'applique sur un tableau de type « individus-variables ». On note alors $X_{ij}$ la valeur pour l'individu $i$ de la variable j.

| | $X_1$ | ..... | $X_j$ | ..... | $X_K$ |
|---|---|---|---|---|---|
| 1 | $X_{11}$ | ..... | $X_{1j}$ | ..... | $X_{1p}$ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| $i$ | $X_{i1}$ | ..... | $X_{ij}$ | ..... | $X_{ip}$ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| N | $X_{n1}$ | ..... | $X_{nj}$ | ..... | $X_{np}$ |

# ACP réduite

Précisons, avant de présenter l'analyse en composantes principales, que celle-ci comporte de nombreuses variantes selon les modifications que l'on apporte aux données brutes.

Nous n'en présenterons ici qu'une seule, la plus utilisée, celle où l'on suppose que le nuage des points représentant les individus est centré et réduit : on dit que l'on opte pour une **ACP réduite**.

# ACP réduite

- Cela consiste a opéré sur chaque variable une transformation en ramenant sa moyenne à zéro et sa variance à l'unité. Mais, bien évidemment, on peut aussi construire l'analyse en composantes principales en supposant le nuage non centré et/ou non réduit.
- Réduire et centrer les données permet de les mettre à la même échelle et de les recentrer autour de zéro, ce qui facilite l'analyse et l'interprétation des données.

# 02.

# Construction et analyse du nuage des individus

# Nuage des individus

Considérons $N$ individus repérés par un indice $i$ $(i = 1, \dots, N)$. Sur chacun de ces individus, nous observons les valeurs de $K$ variables repérés par un indice $j$ $(j = 1, \dots, K)$.

Associons à ces données le nuage des points $P^i$ de $\mathbb{R}^K$ de coordonnées $k_{ij}$. Nous allons centrer et réduire ce nuage. C'est-à-dire substituer à chaque point $P^i$ le point $X^i$ dont la $j^{ème}$ coordonnée est :

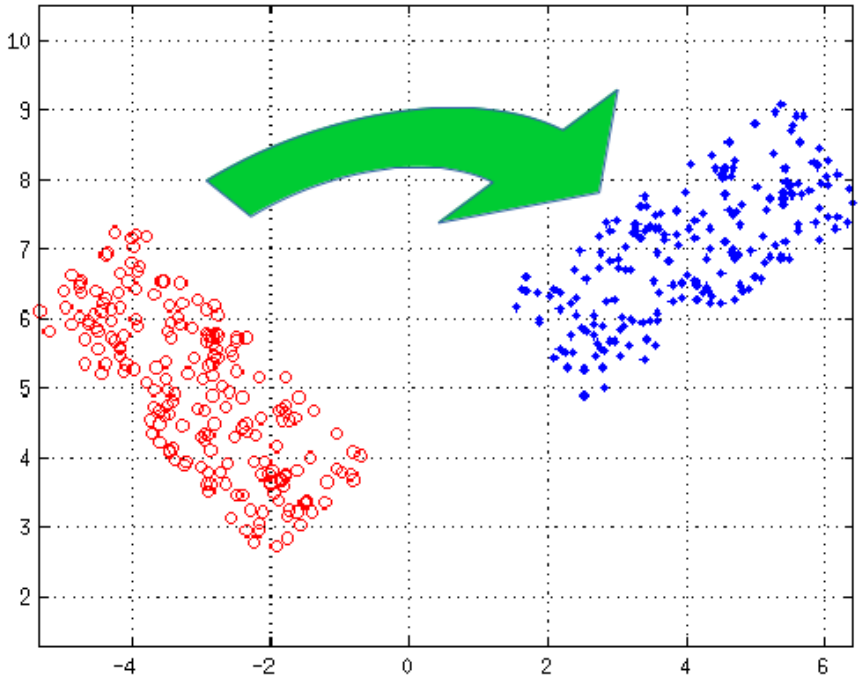
$$x_{ij} = \frac{k_{ij} - \bar{k_j}}{\sigma_j}$$

Avec :

$$\bar{k_j} = \frac{1}{N}\sum_i k_{ij} \;\; et \; {\sigma_j}^2 = \frac{1}{N}\sum_i \left(k_{ij} - \bar{k_j}\right)^2$$

# Nuage des individus

Nous venons de substituer au nuage des points $P^i$ le nuage de points $X^i$ qui n'étant qu'une transformation de ce dernier conserve la même inertie.



# Nuage des individus

De plus, associons à chacun des points $X^i$, une masse unité (c'est-à-dire égale à 1).

On peut maintenant procéder à l'analyse en composantes principales de $\mathcal{N}(I)$. Nous ferons cette analyse à partir de son centre de gravité qui se trouve à l'origine du repère de $\mathbb{R}^K$.

En effet : $G(I) = (g^1, \dots, g^j, \dots g^K)$ est le vecteur nul.

# Nuage des individus

Nous avons vu dans le précédent chapitre que l'analyse factorielle d'un tel nuage de points nécessite que l'on diagonalise la matrice :

$$V = \sum_{i=1}^{N} m_i X^i X^{i^T}$$

En considérant $m_i = 1$, on a :

$$V = \sum_{i=1}^{N} X^i X^{i^T}$$

$$V = XX^T$$

# Nuage des individus

- Pour cela, on commence par identifier les valeurs et vecteurs propres de $V$ et à projeter les points $X^i$ sur ces vecteurs propres.
- On pourra ainsi aisément visualiser le nuage $\mathcal{N}(I)$ par projection sur les plans définis par des couples de vecteurs propres.
- L'intérêt premier de cette représentation sera de mettre en relief les **similarités** ou **ressemblances** entre les individus.

# Nuage des individus

- Deux individus se ressemblent d'autant plus qu'ils possèdent des valeurs proches pour l'ensemble des variables. Il existe donc une faible distance entre eux au niveau du nuage des points.
- Mais n'oublions pas qu'à l'origine un individu i est représenté par un point $X^i$, mais dans un espace où il nous est souvent impossible d'**observer** la proximité ou la similarité de cet individu avec d'autres.

# Nuage des individus

- A cet effet, l'analyse factoriel a été mené pour considérer à la place de $X^i$ le point $Z^i$, qui est son projeté orthogonal sur les p premiers axes factoriels considérés.
- Mais dans ce nouveau sous-espace considéré est-ce que les proximités entre les points sont conservées ?
- En d'autres termes, si $X^i$ et $X^{i\prime}$ étaient proches en est-il de même pour $Z^i$ et $Z^{i\prime}$ ?

# Nuage des individus

$$\left\|Z^i - Z^{i\prime}\right\|^2 = \sum_{j=1}^{K} \left(z_{ij} - z_{i\prime j}\right)^2$$

$$\left\|Z^i - Z^{i\prime}\right\|^2 = \sum_{j=1}^{K} \left(\left\langle X^i \middle| U^j \right\rangle - \left\langle X^{i\prime} \middle| U^j \right\rangle\right)^2$$

$$\left\|Z^i - Z^{i\prime}\right\|^2 = \sum_{j=1}^{K} \left(\left\langle X^i - X^{i\prime} \middle| U^j \right\rangle\right)^2$$

# Nuage des individus

$$\left\|Z^i - Z^{i\prime}\right\|^2 = \sum_{j=1}^{K} \left[U^{j^T}\left(X^i - X^{i\prime}\right)\right]^2$$

$$\left\|Z^i - Z^{i\prime}\right\|^2 = \sum_{j=1}^{K} \left(U^{j^T}\right)^2 \left(X^i - X^{i\prime}\right)^2$$

$$\left\|Z^i - Z^{i\prime}\right\|^2 = \left[\sum_{j=1}^{K} \left(U^{j^T}\right)^2\right] * \left(X^i - X^{i\prime}\right)^2$$

$Z^i$ et $Z^{i\prime}$ sont d'autant plus proches que $X^i$ et $X^{i\prime}$ le sont.

# 03.

# Construction et analyse du nuage des variables

# Nuage des variables

Nous associerons à chaque variable $j$ un point $Y^j$ de $\mathbb{R}^N$ dont la $i^{ème}$ coordonnée est :

$$y_{ij} = \frac{1}{\sqrt{N}} x_{ij}$$

Ainsi :

$$y_{ij} = \frac{k_{ij} - \overline{k_j}}{\sigma_j \sqrt{N}}$$

Nous attribuerons aussi à chaque point $Y^j$ une masse égale à l'unité.

# Nuage des variables

Nous avons vu dans le précédent chapitre qu'on pouvait effectuer l'analyse factoriel d'un tel nuage de deux manières :

- Soit en diagonalisant directement sa matrice d'inertie :

$$\Gamma = X^T X$$

- Soit à travers une analyse duale en diagonalisant la matrice d'inertie :

$$V = X X^T$$

# Nuage des variables

- Dans les deux cas, on obtient les valeurs et les vecteurs propres de $\Gamma$ sur lesquels on peut projeter les points $Y^j$.
- On pourra ainsi aisément visualiser le nuage $\mathcal{N}(J)$ par projection sur les plans définis par des couples de vecteurs propres.
- L'intérêt premier de cette représentation sera de mettre en relief les **liaisons** ou **corrélations** entre les variables.

# Nuage des variables

- Or, deux variables sont liées d'autant plus qu'elles prennent des valeurs proches ou distinctes chez un grand nombre d'individus.
- Mais comment ce phénomène peut-il être observé graphiquement à partir du nuage de point ?
- En d'autres termes, quand est-ce qu'en observant deux points du nuage des variables peut-on en conclure que ces deux variables sont liées ?

# Nuage des variables

Pour cela établissons d'abord deux (02) résultats préliminaires :

1. $\left\|Y^j\right\| = 1$ pour tout j $(j = 1, \ldots, K)$.
2. ${Y^j}^T Y^{j'} = \rho_{jj'}$ où $\rho_{jj'}$ est le coefficient des corrélations des variables j et j'.

# Nuage des variables

**Résultat 1 :**

$$\left\|Y^j\right\|^2 = \left\langle Y^j \middle| Y^j \right\rangle$$

$$\left\|Y^j\right\|^2 = \sum_{i=1}^{N} {y_{ij}}^2$$

$$\left\|Y^j\right\|^2 = \sum_{i=1}^{N} \left( \frac{k_{ij} - \bar{k_j}}{\sigma_j \sqrt{N}} \right)^2$$

# Nuage des variables

**Résultat 1:**

$$\left\|Y^j\right\|^2 = \sum_{i=1}^{N} \frac{1}{{\sigma_j}^2 N} \left(k_{ij} - \bar{k_j}\right)^2$$

$$\left\|Y^j\right\|^2 = \sum_{i=1}^{N} \frac{N}{\left(k_{ij} - \bar{k_j}\right)^2 N} \left(k_{ij} - \bar{k_j}\right)^2$$

$$\left\|Y^j\right\|^2 = 1$$

$$\left\|Y^j\right\| = 1$$

# Nuage des variables

**Résultat 2 :**

$$\rho_{jj'} = \frac{cov(j,j')}{\sigma_j \sigma_{j'}}$$

$$\rho_{jj'} = \frac{\frac{1}{N}\sum_{i=1}^{N}\left(k_{ij} - \overline{k_j}\right)\left(k_{ij'} - \overline{k_{j'}}\right)}{\sigma_j \sigma_{j'}}$$

$$\rho_{jj'} = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{\left(k_{ij} - \overline{k_j}\right)}{\sigma_j} * \frac{\left(k_{ij'} - \overline{k_{j'}}\right)}{\sigma_{j'}}\right]$$

# Nuage des variables

**Résultat 2 :**

$$\rho_{jj'} = \sum_{i=1}^{N} \left[ \frac{\left(k_{ij} - \overline{k_j}\right)}{\sqrt{N}\sigma_j} * \frac{\left(k_{ij'} - \overline{k_{j'}}\right)}{\sqrt{N}\sigma_{j'}} \right]$$

$$\rho_{jj'} = {Y^j}^T Y^{j'}$$

# Nuage des variables

Ces deux résultats préliminaires nous permettent d'aboutir au résultat suivant concernant la distance en les points $Y^j$ et $Y^{j'}$:

$$\left\|Y^j - Y^{j'}\right\|^2 = \left\|Y^j\right\|^2 + \left\|Y^{j'}\right\|^2 - 2Y^{j^T}Y^{j'}$$

Or : $\left\|Y^j\right\|^2 = \left\|Y^{j'}\right\|^2 = 1$ et $\rho_{jj'} = Y^{j^T}Y^{j'}$

On trouve alors que :

$$\left\|Y^j - Y^{j'}\right\|^2 = 2\left(1 - \rho_{jj'}\right)$$

$$\left\|Y^j - Y^{j'}\right\| = \sqrt{2\left(1 - \rho_{jj'}\right)}$$

# Nuage des variables

- La distance entre $Y^j$ et $Y^{j'}$ dépend donc du coefficient de corrélation $\rho_{jj'}$ entre les variables j et j'. On peut dès lors mettre en évidence les liaisons entre les variables en observant leur configuration dans l'espace.
- Nous venons de prouver que le nuage $\mathcal{N}(J)$ des variables doit être interprété en termes de corrélations entre les variables, ainsi que les images de ce nuage que nous obtiendrons par projection.

# Nuage des variables

**Ainsi :**

1. Si les variables $j$ et $j'$ sont liées par une corrélation égale à +1, $\left\| Y^j - Y^{j'} \right\| = 0$ les deux points sont confondus.
2. Si les variables $j$ et $j'$ sont indépendantes, ou de façon plus générale, si $\rho_{jj'} = 0$, alors $Y^j$ et $Y^{j'}$ sont orthogonaux. Dans ce cas $\left\| Y^j - Y^{j'} \right\| = \sqrt{2}$.
3. Si les variables $j$ et $j'$ sont liées par une corrélation égale à -1, $\left\| Y^j - Y^{j'} \right\| = 2$, les deux points $Y^j$ et $Y^{j'}$ sont diamétralement opposés sur la sphère $S(0,1)$.

# 04.

# Cas pratique d’une ACP sur le logiciel R studio

# Présentation de la base de données

**Population :** un ensemble de pays

**Individu statistique :** chaque pays

**Nombre d'individus :** 25

**Nombre de variables :** 23

**Nature des variables :** quantitatives continues

# SYNTHESE

## Quelques points essentiels à retenir :

- l'ACP est une méthode d'analyses factorielles qui ne s'applique qu'à des bases de données de variables quantitatives.
- on opte généralement pour une **ACP réduite** ;
- la proximité entre deux variables s'interprète en termes de liaison et celles entre deux individus en termes de similitude ;
- Cette lecture des nuages doit toujours être accompagnée de l'utilisation des aides à l'interprétation.

# FIN DU CHAPITRE

MERCI POUR VOTRE ATTENTION !