ANALYSE EN COMPOSANTES PRINCIPALES

Mr Z.Bouafia

Master RID

2016 / 2017

Introduction

- L'ACP (Hotelling, 1933) a pour objectif de réduire le nombre de données, souvent très élevé, d'un tableau de données représenté algébriquement comme une matrice et géométriquement comme un nuage de points.
- L'ACP consiste en l'étude des projections des points de ce nuage sur un axe (axe factoriel ou principal), un plan ou un hyperplan judicieusement déterminé.
- Mathématiquement, on obtiendrait le meilleur ajustement du nuage par des sous-espaces vectoriels.

Données

- Soit un tableau de données ayant p lignes et q colonnes:

colonnes / *lignes*	1	....	J	...	*q*
1	X_{11}	...	X_{1j}	...	X_{1q}
i	X_{i1}	...	X_{ij}	...	X_{iq}
p	X_{p1}	...	X_{pj}	...	X_{pq}

Matrice de données

- On représente ce tableau sous forme d'une matrice notée X de type (p,q).

$$X = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1j} & \cdots & X_{1q} \\ X_{21} & X_{22} & \cdots & X_{2j} & \cdots & X_{2q} \\ \vdots & & \vdots & & & \vdots \\ X_{i1} & X_{i2} & \cdots & X_{ij} & \cdots & X_{iq} \\ \vdots & & & \vdots & & \vdots \\ X_{p1} & X_{p2} & \cdots & X_{pj} & \cdots & X_{pq} \end{pmatrix}$$

Procédure de l'ACP

- Algébriquement, il s'agit de chercher les valeurs propres maximales de la matrice des données et par conséquent ses vecteurs propres associes qui représenteront ces sous espaces vectoriels (axes factoriels ou principales).

Étapes :

- On cherche X' la transposée de la matrice X.
- On détermine les valeurs propres de la matrice symétrique X'X.
- Soient λ1, λ2, …, λq ces valeurs propres.
- On les classe λ1>λ2>λ3> λ4>….
- Alors $X'X=A\Lambda A^{-1}$ où

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_q \end{pmatrix}$$

Procédure de l'ACP

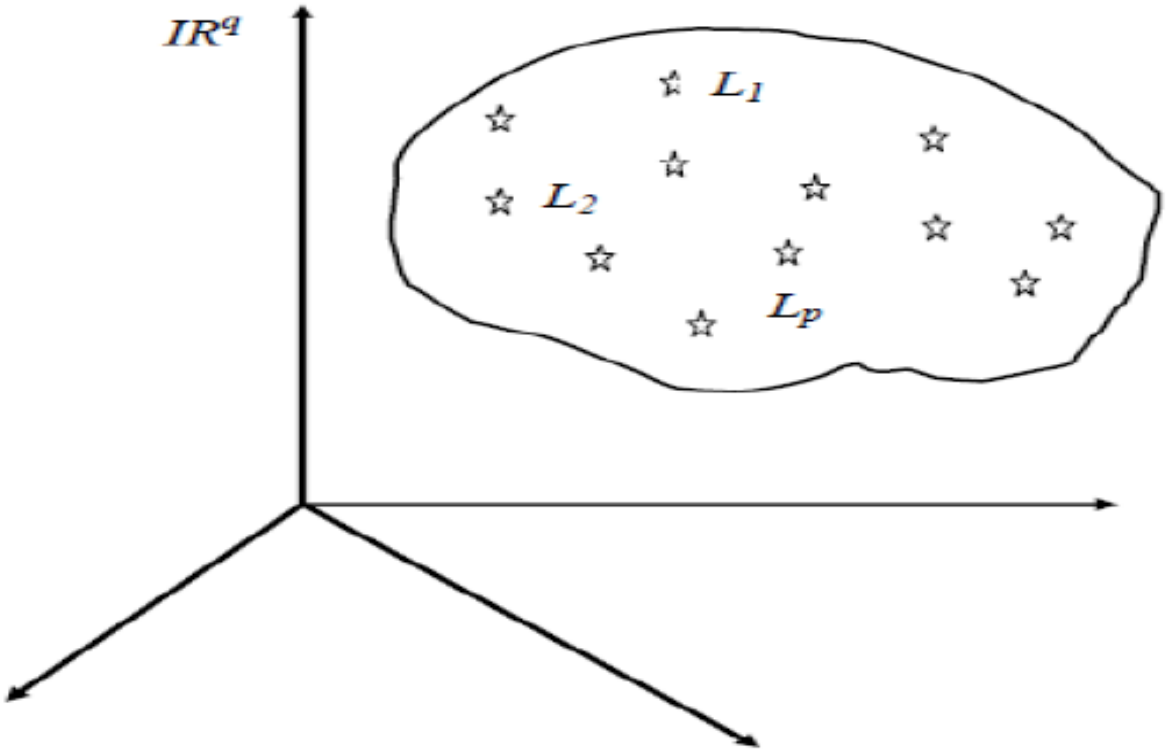
- D'après les propriétés de la trace des matrices; on a:
- $tr(X\,'\,X) = tr(A\,\Lambda\,A^{-1}) = tr(AA\text{-}1\Lambda) = tr(\Lambda)$
- Soit $tr(X\,'\,X) = \lambda 1+\lambda 2+\ldots\ldots\ldots\ldots+\lambda q$
- En raison des valeurs numériques décroissantes de λ1, λ2, …, la somme des premiers valeurs propres représente souvent une proportion importante de la trace de X'X.
- Diagonalisation de X'X : $\det(X'X- \lambda I) = 0$
- Ainsi, dans la pratique on peut se limiter à trouver les premiers valeurs propres λ1, λ2, …,λs avec s assez inferieur a q.
- L'information perdue est alors relativement faible.
- On pratique s=3 (trois premiers valeurs propres les plus grands)

Procédure de l'ACP

- Les valeurs propres trouves étant simples, les espaces propres associes aux vecteurs propres seront des droites vectorielles (on les appelles des axes factoriels ou des facteurs).
- U_1 est le vecteur propre unitaire associé à la plus grande valeur propre λ1 , il vérifie $X'X\ U_1 = \lambda_1\ U_1$ et $||U_1|| = 1$
- D'un point de vue général, L'ACP nous a permit de traiter un très grand nombre de données (matrice) pour identifier un nombre relativement restreint de données (axes factoriels)

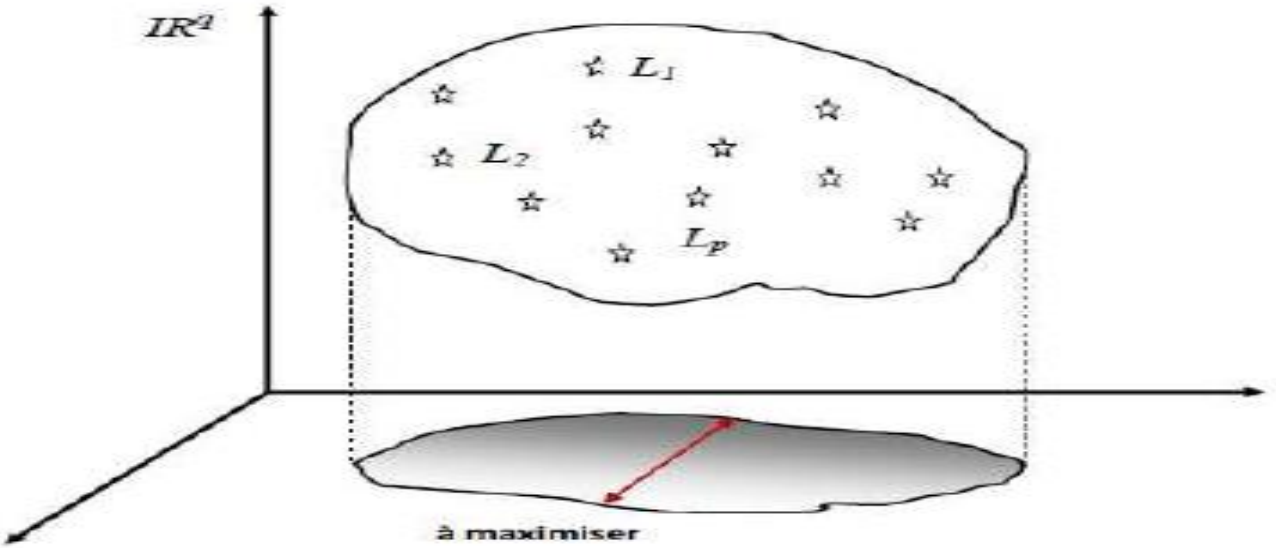
L'ACP (Géométriquement)

- on représente le tableau comme un nuage de points.

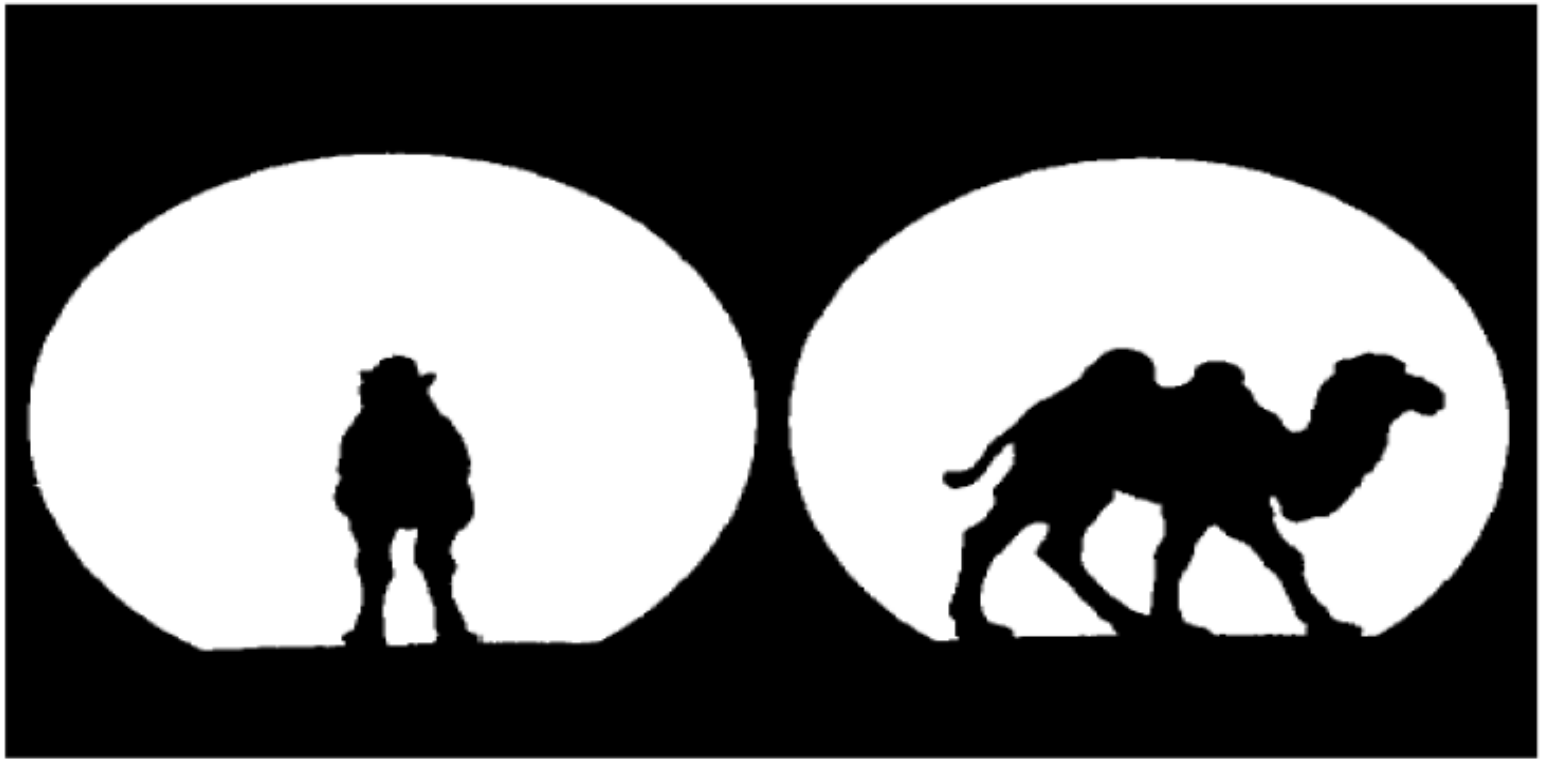


L'ACP (Géométriquement)

- Lors de la projection, le nuage peut être déformé est donc serait différent de réel, alors les méthodes d'ajustement consistent en minimiser cette possible déformation et ce en maximisant les distances projetées.



L’ACP (Géométriquement)



Distance ou métrique utilisée

- Soient Lm et Ln deux points de IR^q:

$$L_m = (x_{m1}, x_{m2}, \cdots, x_{mj}, \cdots, x_{mq})$$

$$L_n = (x_{n1}, x_{n2}, \cdots, x_{nj}, \cdots, x_{nq})$$

- La distance euclidienne (classique) entre ces points est:

$$d(L_m, L_n) = \sqrt{\sum_{j=1}^{q} (x_{mj} - x_{nj})^2}$$

Distance ou métrique utilisée

- Les points $\mathbf{L_m}$ et $\mathbf{L_n}$ sont encore plus proches lorsque la somme précédente est plus petite.
- Si les différents coordonnées des points **L** ne se mesurent pas avec les mêmes unités, la distance **d** sera la somme des termes de ≪ poids ≫ très différents.
- Pour éviter ce problème des unités, on va centrer auparavant les vecteurs colonnes de la matrice X.
- Le tableau des données centrés *Y* est :

$$Y = \begin{pmatrix} X_{11} - \overline{X}_1 & X_{12} - \overline{X}_2 & \cdots & X_{1q} - \overline{X}_q \\ X_{21} - \overline{X}_1 & X_{22} - \overline{X}_2 & \cdots & X_{2q} - \overline{X}_q \\ \vdots & & \ddots & \vdots \\ X_{p1} - \overline{X}_1 & X_{p2} - \overline{X}_2 & \cdots & X_{pq} - \overline{X}_q \end{pmatrix}$$

Centre de gravité

- Le centre de gravité du nuage de points individus G caractérise la position globale de nuage (individu) dans le repère défini par les variables. C'est le point autour duquel « gravitent » les individus du nuage.

$$G = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \dots \\ \bar{x}_p \end{pmatrix}$$

Inertie

- Inertie du nuage de points par rapport à son centre de gravité = somme pondérée des éloignements au centre de gravité.

  - $I = \sum_{i=1}^{n} d2(ei, G) = \sum_{j=1}^{p} V(X_j)$ = Tr(V)

- **I** caractérise la dispersion ou la forme du nuage par rapport à son centre : au plus **I** est élevée, au plus le nuage est dispersé autour de son centre de gravité.
- Dans le cas où les variables sont centrées réduites, la variance de chaque variable vaut 1. L'inertie totale est alors égale à p (nombre de variables).

L'ACP normé

- On s'intéresse à étudier la matrice des variances-covariances **V** au lieu de la matrice **X** de départ.
- La matrice **V** est une matrice de type carrée d'ordre **q** de terme général $\mathbf{v_{kl}}$ égal à:

$$v_{kl} = \frac{1}{p}\sum_{i=1}^{p}(y_{ik} - \bar{y}_k)(y_{il} - \bar{y}_l) = \frac{1}{p}\sum_{i=1}^{p}(x_{ik} - \bar{x}_k)(x_{il} - \bar{x}_l)$$

$$v_{kl} = \frac{1}{p}\sum_{i=1}^{p}(x_{ik}\, x_{il} - \bar{x}_k\, \bar{x}_l)$$

L'ACP normé

- La matrice V des variances-covariances est telle que

$$V = \frac{1}{p} Y'Y$$

- On peut aussi considérer la matrice Z des données centrées et normés d'éléments z_{ij}

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j}$$

- Avec :

$$\bar{x}_j = \frac{\sum_{i=1}^{p} x_{ij}}{p} \quad ; \quad \sigma_j = \sqrt{\frac{1}{p} \sum_{i=1}^{p} \left(x_{ij} - \bar{x}_j\right)^2}$$

Matrice centrée normée

- Donc, la matrice des données centrées et normées sera:

$$Z = \begin{pmatrix} \frac{X_{11} - \overline{X}_1}{\sigma_1} & \frac{X_{12} - \overline{X}_2}{\sigma_2} & \cdots & \frac{X_{1q} - \overline{X}_q}{\sigma_q} \\ \frac{X_{21} - \overline{X}_1}{\sigma_1} & \frac{X_{22} - \overline{X}_2}{\sigma_2} & \cdots & \frac{X_{2q} - \overline{X}_q}{\sigma_q} \\ \vdots & & \ddots & \vdots \\ \frac{X_{p1} - \overline{X}_1}{\sigma_1} & \frac{X_{p2} - \overline{X}_2}{\sigma_2} & \cdots & \frac{X_{pq} - \overline{X}_q}{\sigma_q} \end{pmatrix}$$

Matrice des corrélations

- A partir de cette matrice, on définit la matrice Γ des corrélations entre les q variables prises deux a deux:

$$\Gamma = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1q} \\ \rho_{21} & 1 & \cdots & \rho_{2q} \\ \vdots & & \ddots & \vdots \\ \rho_{q1} & \cdots & \cdots & 1 \end{pmatrix}$$

- Γ résume la structure des dépendances linéaires entre les q variables et on a

$$\Gamma = \frac{1}{p} Z' Z$$

Procédure de l'ACP normé

- On extrait les valeurs propres les plus grands λ1, λ2, …, de la matrice V des variances covariances ou de la matrice Γ des corrélations.
- En pratique, on arrête l'extraction des valeurs propres lorsque la somme des **s** valeurs propres que l'on a déterminés représente un pourcentage satisfaisant de la variance.
- On détermine les vecteurs propres associés aux valeurs propres λ1, λ2, …, λs.
- Ce sont les axes factoriels
- Dans la majorité des cas, ne sont prise en considération que les deux, les trois, ou les quatre premiers axes factoriels.
- Les axes factoriels sont perpendiculaires et ne sont pas corrélés entre eux.

Qualités et défauts de l'ACP

- D'un point de vue technique, ce procédé à pour objet l'étude de la structure de la matrice de variances covariances ou de la matrice des corrélations.
- Mais, le procédé est imparfait dans la mesure que le nuage est déformé par la projection, même si cette dernière est la plus idéale possible. Certains points sont plus altérés que d'autres par la transformation.
- L'inconvénient majeur réside dans l'interprétation des axes. Parfois, l'explication est évidente et fait que l'analyse en composantes principales soit redondante ; ou bien elle est contingente pour l'analyste et dans ce dernier cas elle n'apporte pas des renseignements très convaincantes pour l'analyse économétrique postérieure.