Qualités et défauts de l'analyse en composantes principales :

L'analyse en composantes principales est essentiellement une méthode de description et d'exploration qui permet de révéler des regroupements de faits et suggérer des idées. C'est un outil confortable pour résumer un vaste tableau de données difficilement accessible à l'analyse descriptive habituelle. Les facteurs nés de l'analyse ont la mission de proposer des variables permettant d'élaborer des modèles économétriques de sens traditionnel.

D'un point de vue technique, ce procédé a pour objet l'étude de la structure de la matrice des variances-covariances ou de la matrice des corrélations (des variables). Cette prospection se fait par l'utilisation des ordinateurs et des logiciels de statistique. 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 convaincant es pour l'analyse économétrique postérieure. Néanmoins, l'analyse des données a toujours un rôle essentiel à jouer dans certains problèmes dans certaines limites.

Nombre d'axes à retenir :

L'analyse en composantes principales a pour objet de réduire le nombre de données du phénomène à étudier et de conserver ainsi le moins d'axes possibles. Il faut pour cela que les variables de départ soient raisonnablement corrélées entre elles.

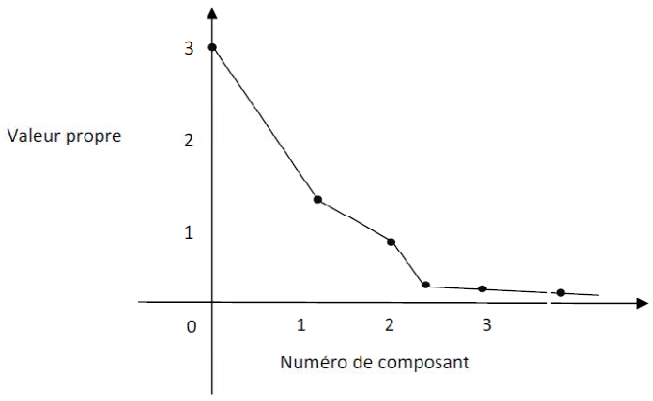
Les critères les plus utilisables sont les suivantes :

1°) **Interprétation des axes :** On retient que les axes que l'on peut attribuer une forme d'interprétation économique, par exemple, soit directement, soit en terme des variables avec lesquelles ils sont très corrélés.

2°) **Critère de Kaiser** (variables centrées et réduites) : On ne retient que les axes associés à valeurs propres supérieurs à 1, c'est-à-dire dont la variance est supérieure à celle des variables d'origine.

Une autre interprétation est que la moyenne des valeurs propres étant 1, on ne garde que celles qui sont supérieures à cette moyenne.

3°) **Éboulis des valeurs propres** :



On cherche un « coude » dans le graphe des valeurs propres et on ne conserve que les valeurs jusqu'au ce « coude ».

Compléments du cours :

Multiplicateurs de Lagrange :

Optimisation classique avec contraintes: Cas de deux variables.

Soit une fonction à deux variables $f(x, y)$ soumise à une seule contrainte de la forme $g(x, y) = b$, avec b une constante réelle.

La méthode des multiplicateurs de Lagrange consiste à construire une fonction auxiliaire $L(x, y, \lambda)$, appelée Lagrangien, définie ainsi :

$$L(x, y,\lambda) = f(x, y)+\lambda[g(x, y)-b]$$

Où λ appelé multiplicateur de Lagrange est une inconnue.

Il faut ensuite annuler ses premières dérivées partielles (condition nécessaire) :

$$\begin{cases} \dfrac{\partial L}{\partial x} = \dfrac{\partial f}{\partial x} + \lambda \dfrac{\partial g}{\partial x} = 0 \\ \dfrac{\partial L}{\partial y} = \dfrac{\partial f}{\partial y} + \lambda \dfrac{\partial g}{\partial y} = 0 \\ \dfrac{\partial L}{\partial \lambda} = g(x, y) - b = 0 \end{cases}$$

Les points candidats s'obtiennent en résolvant ce système de trois équations à trois inconnues (x, y, λ).

Mentionnons que la troisième équation de ce système $\partial L/\partial\lambda = g(x, y) - b=0$ n'est rien d'autre que la contrainte ! Les points candidats satisfont par conséquent cette contrainte.

La solution des trois équations ci-dessus fournit les points candidats de la fonction sous contrainte. Ces points candidats satisfont la contrainte mais il reste à déterminer leur nature ;

Condition suffisante:

On pose:
$$\Delta = \frac{\partial^2 L}{\partial x^2} \cdot \frac{\partial^2 L}{\partial y^2} - \left(\frac{\partial^2 L}{\partial x \partial y} \right)^2$$

1. Si $\Delta>0$, $\dfrac{\partial^2 L}{\partial x^2} > 0$ *et* $\dfrac{\partial^2 L}{\partial y^2} > 0$, on a un minimum
2. Si $\Delta>0$, $\dfrac{\partial^2 L}{\partial x^2} < 0$ *et* $\dfrac{\partial^2 L}{\partial y^2} < 0$, on a un maximum
3. Si $\Delta<0$, pas d'extremum.
4. Si $\Delta=0$, on ne peut pas conclure.

Rappel sur la distance :

Définition d'une distance :

Soit E un sous-ensemble de IR^n.

Une distance sur E est une application $d : E \times E \to IR+$ possédant les propriétés suivantes :

i. $\forall x, y \in E; \quad d(x,y) = 0 \Rightarrow x = y$
ii. $\forall x, y \in E; \quad d(x,y) = d(y,x)$
iii. $\forall x, y, z \in E; \quad d(x,y) \le d(x,z) + d(z,y)$

Exemple : « La distance euclidienne »

Pour $x = (x_1, x_2, \cdots, x_n), y = (y_1, y_2, \cdots, y_n) \in E \subset IR^n$, la distance euclidienne entre x et y est définie par : $d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2}$.

On peut vérifier facilement les propriétés i, ii, et iii précédentes pour la distance euclidienne.

Rappel sur la matrice des variances-covariances et la matrice des corrélations :

1) La matrice des variances-covariances V de $X=(x_1,x_2,\ldots, x_q)$ est définie par :

$$V = \begin{pmatrix} \sigma_1^2 & Cov(x_1,x_2) & \cdots & Cov(x_1,x_q) \\ Cov(x_2,x_1) & \sigma_2^2 & \cdots & Cov(x_2,x_q) \\ \vdots & & \ddots & \vdots \\ Cov(x_q,x_1) & \cdots & \cdots & \sigma_q^2 \end{pmatrix} = E(XX') - E(X)E(X)'$$

C'est une matrice carrée symétrique d'ordre q.

Si les variables x_i sont réduites, V s'identifie avec la matrice des corrélations :

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

2) Lorsque l'on observe les valeurs numériques de q variables sur p individus, on se trouve en présence d'un tableau X à p lignes et q colonnes :

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1q} \\ x_{21} & x_{22} & \cdots & x_{2q} \\ \vdots & & \ddots & \vdots \\ x_{p1} & x_{p2} & \cdots & x_{pq} \end{pmatrix}$$

x_{ij} est la valeur prise par la variable n° j sur l'ième individu.

Le tableau des données centrés Y est :

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

La matrice des variances-covariances des q variables est :

$$V = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1q} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2q} \\ \vdots & & \ddots & \vdots \\ \sigma_{q1} & \cdots & \cdots & \sigma_q^2 \end{pmatrix}$$

où $\sigma_{kl} = \frac{1}{p}\sum_{i=1}^{p}\left(x_{ik}x_{il} - \bar{x}_k\bar{x}_l\right)$ est telle que $V = \frac{1}{p}Y'Y$

La matrice des corrélations entre les q variables prises deux à deux est :

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

Γ est identique à V des données centrées et réduites.

Γ résume la structure des dépendances linéaires entre les q variables.

Le tableau des données centrées et réduites Z est :

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

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

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

Si $\sigma_j = 1$, alors $V = \frac{1}{p}Y'Y = \frac{1}{p}Z'Z = \Gamma$

Exercices de TD :

Exercice 1 :

On considère la matrice X de type (2,3) suivante :

$$X = \begin{pmatrix} -1 & 0 & 1 \\ 0 & -1 & 1 \end{pmatrix}.$$

1. Calculer le produit matriciel. $X' \times X$.
   s'assurer que c'est une matrice carrée et symétrique
2. Chercher les valeurs propres λ_i et les sous-espaces propres associés F_i. Donner le vecteur unitaire u_i de chaque sous-espace. Ecrire la matrice diagonale Λ semblable à $X'X$ et sa matrice de passage A
3. Calculer et vérifier que $tr(X'X) = tr(\Lambda)$..

Exercice 2 :

Soit la matrice des données suivante :

$$X = \begin{pmatrix} 4 & 5 \\ 6 & 7 \\ 8 & 0 \end{pmatrix}$$

1. On note C1 et C2 les vecteurs colonnes de X. Centrer et normer les variables C1 et C2.
2. Déterminer la matrice V des variances-covariances et la matrice Γ des corrélations.
3. Diagonaliser ces matrices. On note λ_i leurs valeurs propres.
4. Déterminer les espaces propres F_i associés aux valeurs propres λ_i .

Exercice 3 :

Réaliser l'ACP de la matrice suivante, à partir de sa matrice de dispersion (données centrées mais non réduites) :

$$\begin{pmatrix} 2 & 2 \\ 6 & 2 \\ 6 & 4 \\ 10 & 4 \end{pmatrix}$$