

# I. Méthodes factorielles

## 1. Rappels

### a) Objectif

L'objectif est de représenter dans  $\mathbb{R}$ ,  $\mathbb{R}^2$  ou  $\mathbb{R}^3$  des données en grande dimension. La matrice des données  $X$  est de dimensions  $n$  (nombre d'individus, en ligne)  $p$  (nombre de variables, en colonne). L'espace des individus est donc  $\mathbb{R}^n$ , celui des variables  $\mathbb{R}^p$ .

### b) Distances et matrices

Propriétés d'un indice de distance :

- $d(x,y) = d(y,x) \geq 0$ .
- $d(x,x) = 0$ .
- $d(x,y) \leq d(x,z) + d(z,y)$ .

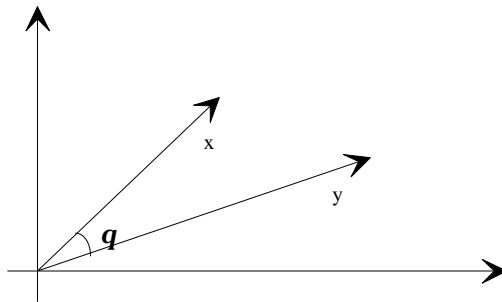
Distance euclidienne :

$$d^2(x,y) = (x-y)^t Q (x-y)$$

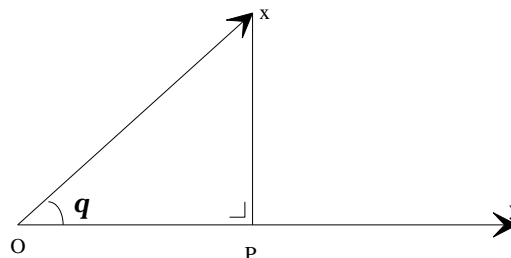
avec  $Q$  est une métrique (matrice définie positive symétrique).

Norme :  $\|x\| = d(0,x) = \sqrt{x^t x}$ .

Angle entre deux vecteurs  $x$  et  $y$  de  $\mathbb{R}^p$  :  $\cos(\mathbf{q}) = \frac{x^t y}{\|x\| \|y\|}$



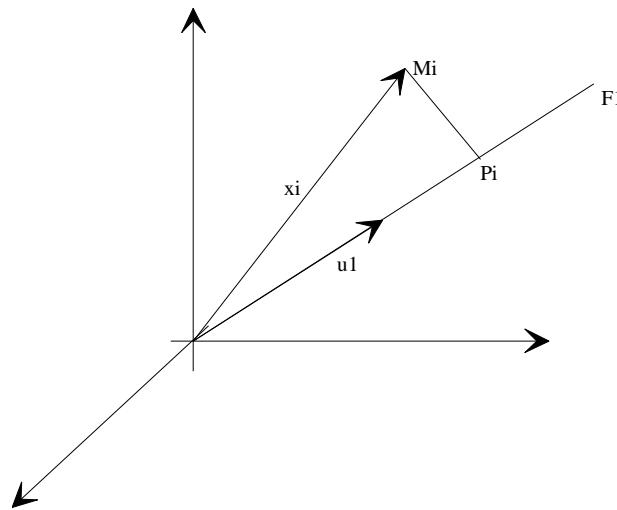
### c) Projection



→ projection de  $x$  sur  $y$  :  $\text{dist}(OP) = \|OP\|$  :  $\cos(\mathbf{q}) = \frac{x^t y}{\|x\| \|y\|} = \frac{\|OP\|}{\|OX\|}$  car il s'agit d'un triangle rectangle. Si  $\|y\| = 1$ , alors  $\text{dist}(OP) = x^t y$ .

### d) Analyse dans l'espace des individus

L'objectif poursuivi est de trouver une droite  $F_1$ , définie par un vecteur  $u_1$  de  $\mathbb{R}^p$ , de taille 1, qui "ajuste" le mieux possible le nuage de  $n$  points (individus).



On obtient alors la représentation de  $x_i \in \mathbb{R}^p$  en le projetant sur  $F_1$  :  $\|OP_i\| = x_i^t u_1$ . Le meilleur ajustement est celui qui minimise la somme des carrés des résidus :

→ trouver  $u_1 \in \mathbb{R}^p$  tel que :

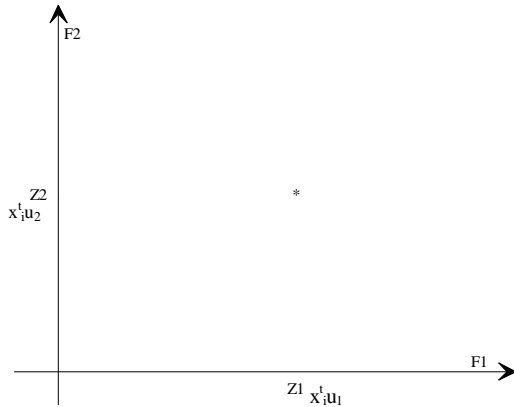
$$\|u_1\| = 1.$$

$$u_1 \text{ minimise } \sum_{i=1}^n \|M_i P_i\|^2.$$

Comme il s'agit d'un triangle rectangle,  $\|M_i P_i\|^2 = \|OM_i\|^2 - \|OP_i\|^2$ . Minimiser la somme des carrés des résidus revient donc à la variance de la projection :  $\sum_{i=1}^n \|OP_i\|^2$ , toujours sous la contrainte que  $\|u_1\| = 1$ . La solution est obtenue à l'aide d'un lagrangien. Le résultat est le suivant :

**$u_1$  est le vecteur propre de la  $X^t X$  associé à sa plus grande valeur propre  $\lambda_1$**

On obtient donc graphiquement la solution suivante :



Remarques :

→  $z^1 = X^t u_1$  : première variable factorielle, projection des  $n$  individus sur le vecteur  $u_1$ .

→  $u_1$  : premier facteur = coefficient de la première variable factorielle ou, en d'autres termes, les coefficients de la combinaison

linéaires  $z_{i1} = \sum_{j=1}^p x_{ij} u_{j1}$ .

→  $\lambda_1$  : inertie (variance) e la première variable factorielle =  $(z^1)^t z^1$  : somme des carrés de la projection.

### e) Analyse dans l'espace des variables

L'approche suivie est strictement symétrique à celle suivie dans l'espace des individus. On peut donc projeter les variables dans un espace de dimensions réduites. Le vecteur  $v^1$  sera le vecteur propre qui correspond à la plus grande valeur propre  $\mu_1$  de la matrice  $X X^t$ .

### f) Relation entre l'espace des individus et l'espace des variables

Il y a bien entendu une relation directe entre l'espace des individus et l'espace des variables. Si  $r$  est le rang de la matrice  $X$  ( $\leq \min(p,n)$ ),  $\forall k \leq r$ , on peut affirmer que :

→ les valeurs propres des variables sont celles des individus :  $\lambda_k = \mu_k$ .

$$\rightarrow u_k = \frac{1}{\sqrt{I_k}} X^t v_k \text{ et } v_k = \frac{1}{\sqrt{I_k}} X u_k .$$

### g) Qualité de la projection

Comme on peut démontrer que la somme des valeurs propres est égale à l'inertie total du nuage de

points, le rapport  $\frac{\sum_{i=1}^q I_i}{\sum_{j=1}^p I_j}$  ( $q$  étant le nombre de dimensions retenues pour effectuer la projection)

fournit une mesure claire de la représentativité du sous-espace par rapport à l'espace d'entrée des données.

## 2. Analyse en composantes principales (ACP)

### a) Représentation dans l'espace des variables

La première étape consiste à centrer les données variables par variables. Dans une ACP, l'origine des axes n'a en effet pas de signification particulière, aussi paraît-il "raisonnable" (et, en fait, surtout pratique) de la placer au centre gravité du nuage de points.

La métrique choisie repose sur la distance euclidienne. Cependant, comme cela a été mentionné ci-dessus, la distance euclidienne est sensible au choix des unités. Pour éviter ce problème, les données sont centrées (chaque variable est divisée par son écart type).

Enfin, en pratique, on divise également chaque variable par  $\frac{1}{\sqrt{n}}$ .

→ la matrice des données originales Y est donc transformées en  $X = \frac{1}{\sqrt{n}} MYD_{1/s}$  où M est la

matrice de centrage et  $D = \begin{bmatrix} \frac{1}{\sqrt{s_1^2}} & & \\ & \ddots & \\ & & \frac{1}{\sqrt{s_p^2}} \end{bmatrix}$ .

## b) Représentation dans $R^n$

Le passage de la matrice Y à la matrice X possède un certain nombre de propriétés intéressantes :

→ produit scalaire : le produit scalaire entre deux variables  $x^j$  et  $x^k$

$$((x^j)^t x^k = \frac{1}{n} \sum_{i=1}^n \left( \frac{y_{ij} - \bar{y}_j}{s_j} \right) \left( \frac{y_{ik} - \bar{y}_k}{s_k} \right)) \text{ est égal à leur coefficient de corrélation } r_{jk}.$$

→ norme : la norme d'une variable  $x^n = 1$  ( $\|x^j\|^2 = (x^j)^t (x^j) = 1$ ).

→ distance : la distance entre deux variables est fonction de leur coefficient de corrélation :

$$d^2(j, k) = (x^j - x^k)^t (x^j - x^k) = (x^j)^t (x^j) + (x^k)^t (x^k) - 2(x^j)^t x^k = 2(1 - r_{jk})$$

$$\rightarrow \text{si } r_{jk} = +1 : d^2 = 0 \rightarrow d=0.$$

$$\rightarrow \text{si } r_{jk} = 0 : d^2 = 2 \rightarrow d=\sqrt{2}.$$

$$\rightarrow \text{si } r_{jk} = -1 : d^2 = 4 \rightarrow d=2.$$

→ angle entre deux variables : l'angle entre deux variables est égal à leur coefficient de corrélation :

$$\cos(\mathbf{q}) = \frac{x^t y}{\|x\| \|y\|} = r_{xy}$$

→ matrice variance/covariance :  $X^t X$  est la matrice variance/covariance et la matrice des coefficients de corrélation. Il s'agit d'une matrice toujours diagonalisable puisqu'elle est symétrique.

## c) Application

Pour réaliser une ACP, on suit la procédure suivante :

- standardisation des variables : passage de Y à X

- décomposition spectrale de  $R_X$  (matrice des coefficients de corrélation) :  $R_X = \Gamma D_\lambda \Gamma^t$  qui produit les valeurs propres, les vecteurs propres, les facteurs et les variables factorielles.
- représentation graphique de la projection des individus et analyse des proximités.
- représentation graphique de la projection des variables et interprétation des axes.