ANALYSE EN COMPOSANTES PRINCIPALES (ACP)

PLAN

Rappel	Transparent
Vecteur, espace vectoriel	3 à 8
Distance	9 à 12
Produit scalaire	13 à 16
Géométrie	17 à 19
Décomposition	
Notion de rang	20 à 22
Décomposition d'une matrice de rang 2 et généralisation	23
Application du théorème d'Eckart-Young à l'approximation d'une matrice et interprétation géométrique	24 à 30
Analyse générale d'un tableau	
Inertie, approximation globale	31 à 34
Décomposition selon les lignes et les colonnes	35 à 42
Les mesure de qualité	43
Illustration numérique	44 à 46
Analyse en composantes principales	
Particularité de l'ACP par rapport à l'analyse générale	47 à 53
Pratique de l'ACP	54 à 63
Biplot	64 à 68

Notations du cours

Un vecteur est noté par une lettre latine en caractère gras. Il est toujours représenté en colonne.

$$\mathbf{x} = \begin{pmatrix} x_1 \\ \dots \\ x_i \\ \dots \\ x_n \end{pmatrix} \in \mathbb{R}^n.$$

Si l'on veut représenter un vecteur ligne, on écrit ${}^t\mathbf{x}$ (transposé de $\mathbf{x}$).

Une matrice est en général notée par une lettre capitale, par exemple X.

$$X = \begin{pmatrix} x_1^1 & \dots & x_1^j & \dots & x_1^p \\ \dots & \dots & \dots & \dots & \dots \\ x_i^1 & \dots & x_i^j & \dots & x_i^p \\ \dots & \dots & \dots & \dots & \dots \\ x_n^1 & \dots & x_n^j & \dots & x_n^p \end{pmatrix}$$

$\mathbf{x}_i^j$ valeur de la colonne j pour la ligne i.

Dimensions de X notées $n \times p$ soit $\underset{n \times p}{X}$

On note $\mathbf{x}^j$ la j$ème$ colonne de X : $\mathbf{x}^j = \begin{pmatrix} x_1^j \\ \dots \\ x_i^j \\ \dots \\ x_n^j \end{pmatrix}$, on peut écrire X par colonnes :

$$X = \left(\begin{array}{c|c|c|c|c} x_1^1 & \dots & x_1^j & \dots & x_1^p \\ \dots & \dots & \dots & \dots & \dots \\ x_i^1 & \dots & x_i^j & \dots & x_i^p \\ \dots & \dots & \dots & \dots & \dots \\ x_n^1 & \dots & x_n^j & \dots & x_n^p \end{array} \right) = \left(\mathbf{x}^1 \,\middle|\, \dots \,\middle|\, \mathbf{x}^j \,\middle|\, \dots \,\middle|\, \mathbf{x}^p \right)$$

En notant ${}^t\mathbf{x}_i$ la i$ème$ ligne de X, on a :

$$X = \left(\begin{array}{ccccc} x_1^1 & \dots & x_1^j & \dots & x_1^p \\ \hline \dots & \dots & \dots & \dots & \dots \\ \hline x_i^1 & \dots & x_i^j & \dots & x_i^p \\ \hline \dots & \dots & \dots & \dots & \dots \\ \hline x_n^1 & \dots & x_n^j & \dots & x_n^p \end{array} \right) = \left(\begin{array}{c} {}^t\mathbf{x}_1 \\ \hline \dots \\ \hline {}^t\mathbf{x}_i \\ \hline \dots \\ \hline {}^t\mathbf{x}_n \end{array} \right)$$

Exemple : notes scolaires

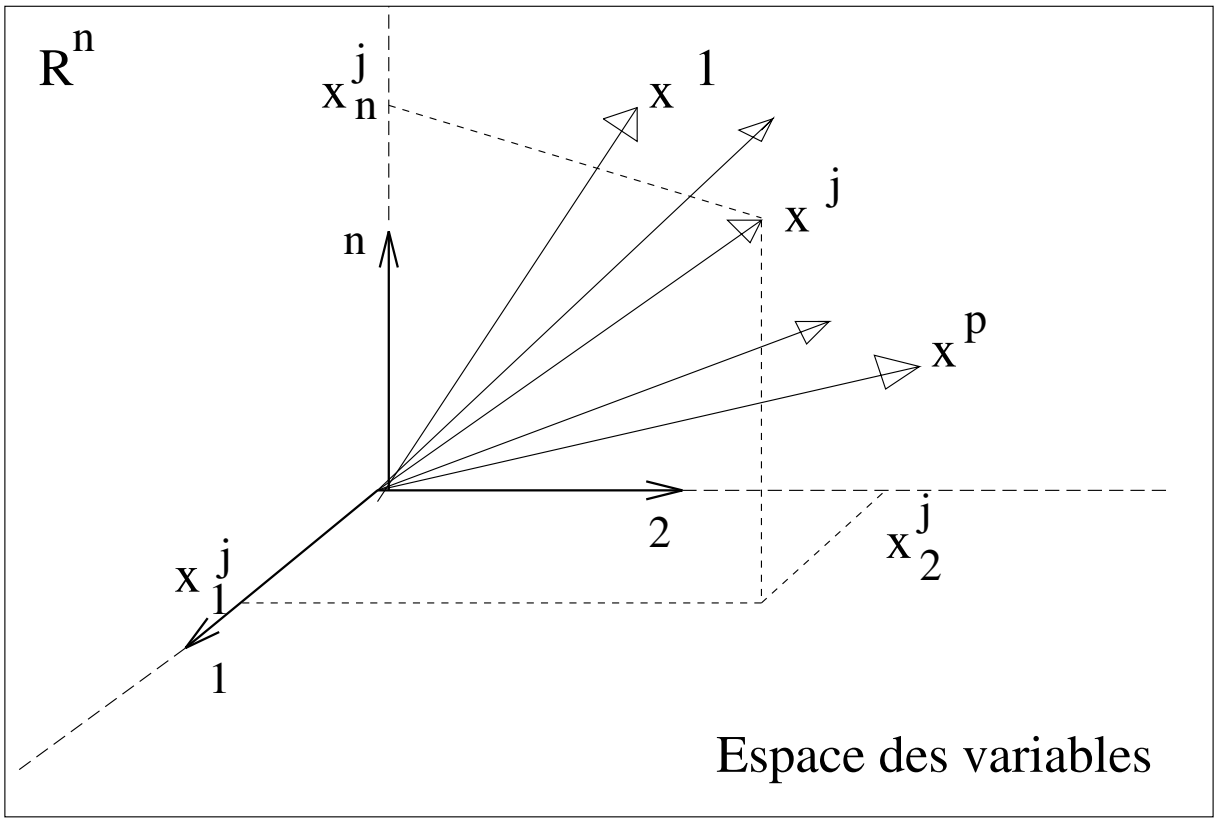
$$\underset{9\times 5}{X} =$$

nom	m	s	f	l	dm
JE	6.0	6.0	5.0	5.5	8.0
AL	8.0	8.0	8.0	8.0	9.0
AN	6.0	7.0	11.0	9.5	11.0
MO	14.5	14.5	15.5	15.0	8.0
DI	14.0	14.0	12.0	12.5	10.0
AD	11.0	10.0	5.5	7.0	13.0
PI	5.5	7.0	14.0	11.5	10.0
BR	13.0	12.5	8.5	9.5	12.0
EV	9.0	9.5	12.5	12.0	18.0

Les lignes du tableau sont associées aux élèves ou *individus*.
Les colonnes du tableau sont associées aux notes ou *variables*.

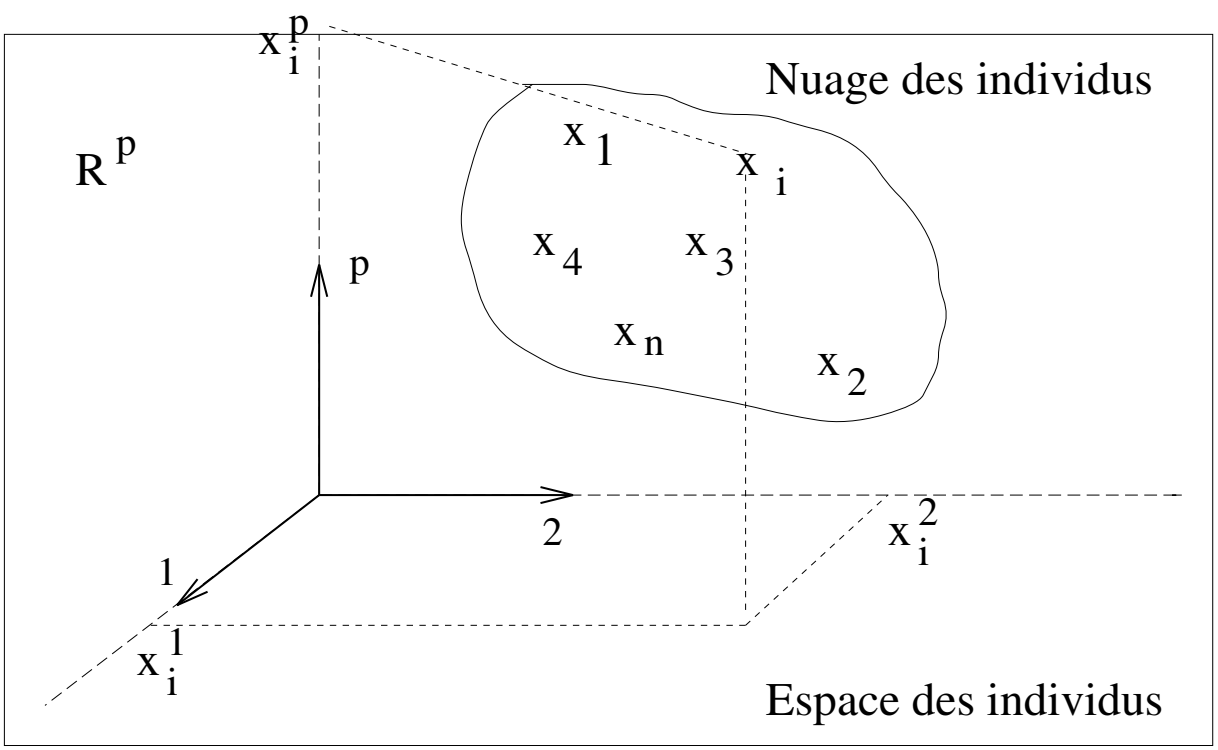
Nuage des colonnes

$$X = \begin{pmatrix} & & \mathbf{x}^j & & \\ x_1^1 & \dots & x_1^j & \dots & x_1^p \\ \dots & \dots & \dots & \dots & \dots \\ x_i^1 & \dots & x_i^j & \dots & x_i^p \\ \dots & \dots & \dots & \dots & \dots \\ x_n^1 & \dots & x_n^j & \dots & x_n^p \end{pmatrix}$$



Nuage des lignes

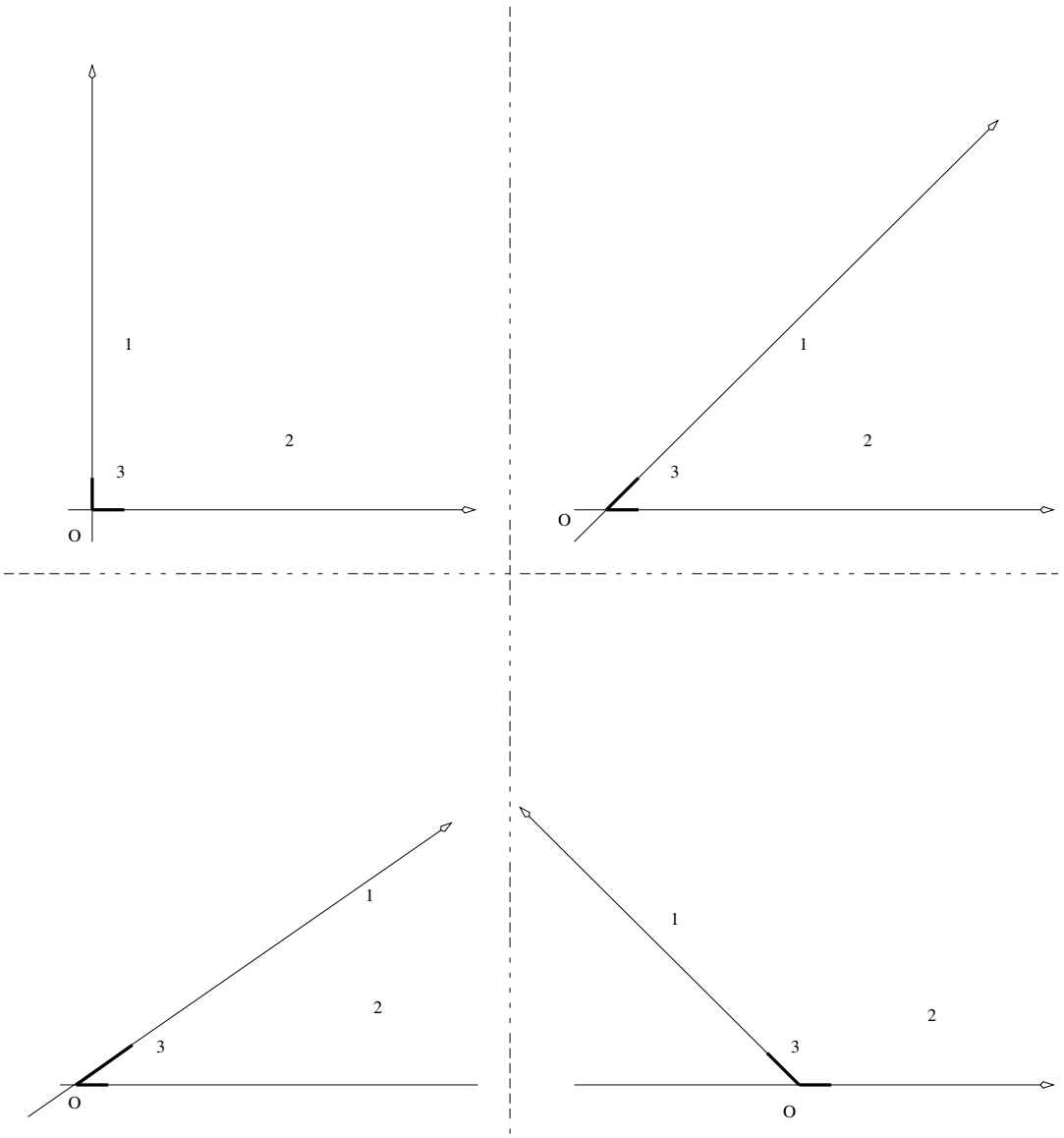
$$X = \left(\begin{array}{ccccc} x_1^1 & \ldots & x_1^j & \ldots & x_1^p \\ \hline \ldots & \ldots & \ldots & \ldots & \ldots \\ \hline x_i^1 & \ldots & x_i^j & \ldots & x_i^p \\ \hline \ldots & \ldots & \ldots & \ldots & \ldots \\ \hline x_n^1 & \ldots & x_n^j & \ldots & x_n^p \end{array} \right)$$



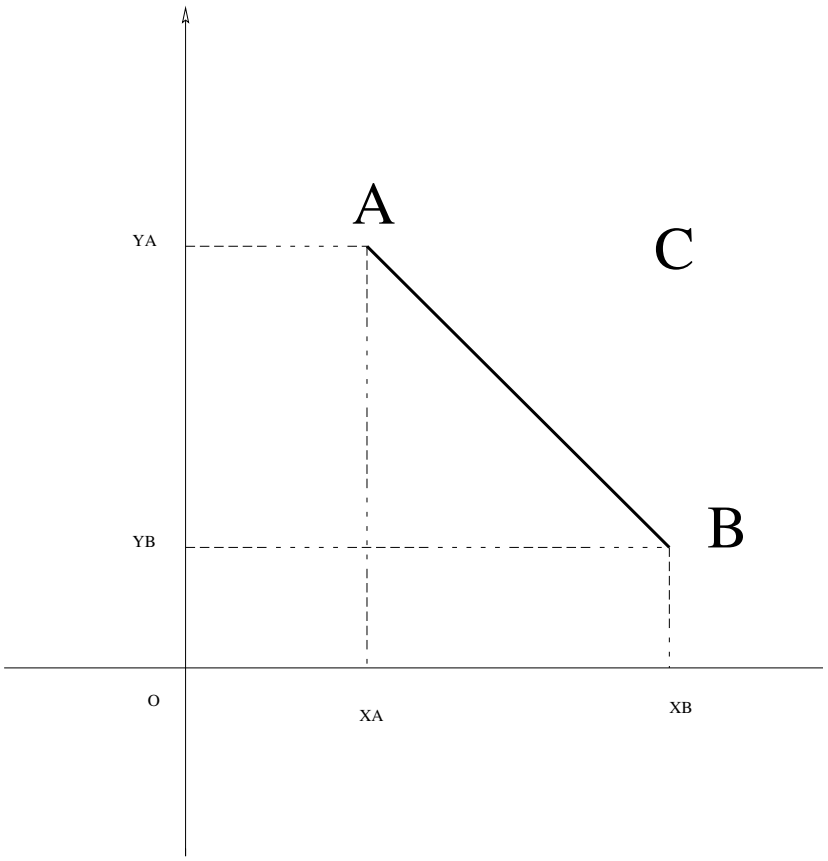
Incidence du choix de la base

Soient les points :

$$1 = \begin{pmatrix} 1 \\ 5 \end{pmatrix}, 2 = \begin{pmatrix} 6 \\ 2 \end{pmatrix} \text{ et } 3 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$



Distance euclidienne usuelle dans $\mathbb{R}^2$



- $d(A,B) = \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2}$

- Propriétés d'une distance :

  $d(A,B) > 0$ si $A \neq B$, $d(A,B) = 0$ si $A = B$

  $d(A,B) = d(B,A)$ symétrie

  $d(A,C) \leq d(A,B) + d(B,C)$ inégalité triangulaire

Distance euclidienne dans $\mathbb{R}^p$

- $\mathbf{x}_i$ et $\mathbf{x}_k$ deux points de $\mathbb{R}^p$
- $d_M(\mathbf{x}_i, \mathbf{x}_k) = \sqrt{\sum_{j=1}^{p} m_j (x_i^j - x_k^j)^2}$
- Les poids m_j sont positifs et définissent une métrique M dans l'espace des lignes de X.

Distance euclidienne dans $\mathbb{R}^n$

- $\mathbf{x}^j$ et $\mathbf{x}^k$ deux points de $\mathbb{R}^n$ (2 variables)
- $d_D(\mathbf{x}^j, \mathbf{x}^k) = \sqrt{\sum_{i=1}^{n} p_i (x_i^j - x_i^k)^2}$
- Les poids p_i sont positifs **et** vérifient $\sum_{i=1}^{n} p_i = 1$
- d_D définit la distance de 2 colonnes de X.

Distance de deux matrices

- $\underset{n\times p}{X}$ et $\underset{n\times p}{Y}$ deux matrices de même dimension.

- $d(X,Y)^2 = \sum_{j=1}^{p} m_j d_D(\mathbf{x}^j, \mathbf{y}^j)^2$

  (somme des carrés des distances entre les variables)

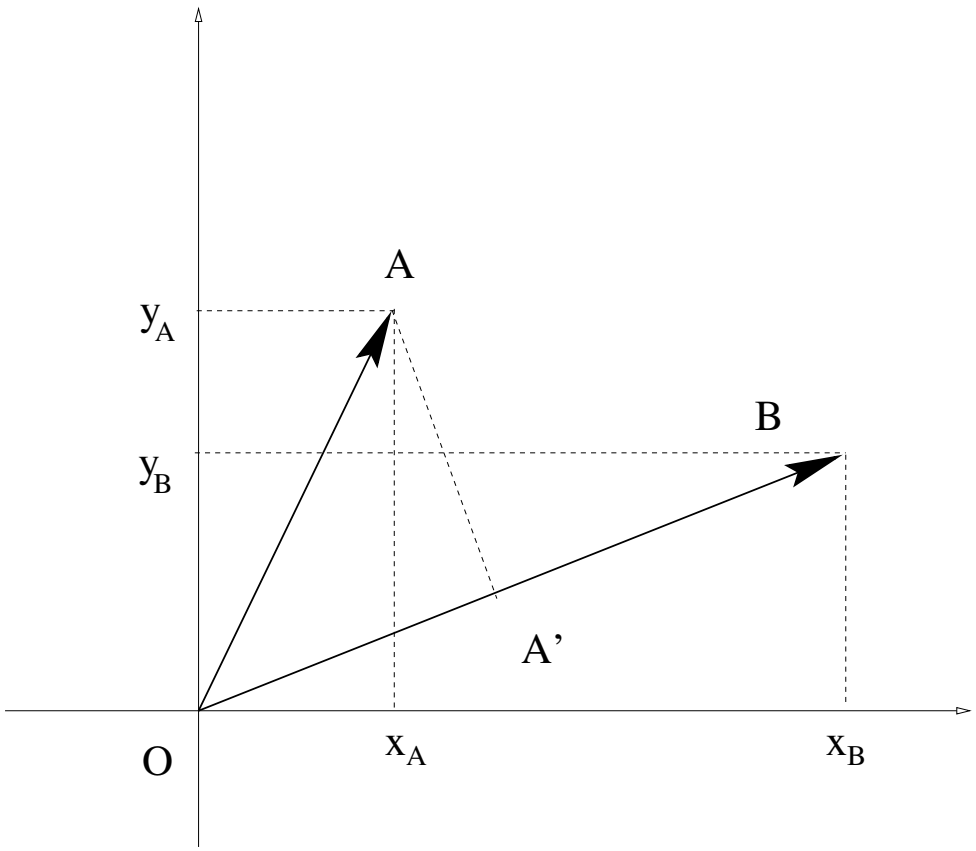
  $= \sum_{i=1}^{n} p_i d_M(\mathbf{x}_i, \mathbf{y}_i)^2$

  (somme des carrés des distances entre les individus)

  $= \sum_{i=1}^{n}\sum_{j=1}^{p} p_i m_j (x_i^j - y_i^j)^2$

  (somme des carrés des distances entre les éléments)

- **Produit scalaire des vecteurs OA et OB :** $< OA, OB >$



$$< OA, OB > = \; x_A x_B + y_A y_B$$

- **Norme de OA** : $\|OA\| = \sqrt{< OA, OA >}$
- $\cos\theta = \dfrac{\|OA'\|}{\|OA\|} = \dfrac{< OA, OB >}{\|OA\|\|OB\|}$
- De manière plus générale, $< OA, OB >_Q = q_1 x_A x_B + q_2 y_A y_B$
  Q, matrice diagonale 2×2 formée avec q_1 et q_2 **positifs**, est la métrique qui définit le produit scalaire.

Produit scalaire dans $\mathbb{R}^p$, espace des individus

Pour tout $\mathbf{x} = \begin{pmatrix} x^1 \\ \vdots \\ x^p \end{pmatrix} \in \mathbb{R}^p$, et pour tout $\mathbf{y} = \begin{pmatrix} y^1 \\ \vdots \\ y^p \end{pmatrix} \in \mathbb{R}^p$,

on note $< \mathbf{x}, \mathbf{y} >_M = \sum_{j=1}^{p} m_j x^j y^j$

$\sqrt{< \mathbf{x}, \mathbf{x} >_M}$ est une norme dite **euclidienne**, notée $\|\mathbf{x}\|_M$:
$\|\mathbf{x}\|_M = \sqrt{\sum_{j=1}^{p} m_j (x^j)^2}$

- Écriture matricielle : $< \mathbf{x}, \mathbf{y} >_M = {}^t \mathbf{x} M \mathbf{y}$

- Propriétés : $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z}$ vecteurs de $\mathbb{R}^p$, α réel

  $< \alpha\mathbf{x}, \mathbf{y} >_M = \alpha < \mathbf{x}, \mathbf{y} >_M$

  $< \mathbf{x}, \mathbf{y} >_M = < \mathbf{y}, \mathbf{x} >_M$ (symétrie)

  $< \mathbf{x} + \mathbf{y}, \mathbf{z} >_M = < \mathbf{x}, \mathbf{z} >_M + < \mathbf{y}, \mathbf{z} >_M$

Produit scalaire dans $\mathbb{R}^n$, espace des variables

$\mathbf{x}$, $\mathbf{y}$ vecteurs de $\mathbb{R}^n$.

Les n poids p_i positifs vérifiant $\sum_{i=1}^{n} p_i = 1$ définissent une matrice diagonale appelée métrique D.

$$< \mathbf{x}, \mathbf{y} >_D = \sum_{i=1}^{n} p_i x_i y_i = {}^t\mathbf{x} D \mathbf{y}$$

Norme associée : $\|\mathbf{x}\|_D = \sqrt{\sum_{i=1}^{n} p_i (x_i)^2}$

Distance associée : $d_D(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{n} p_i (x_i - y_i)^2}$

Exemple : notes en Math et sciences et notes centrées :

nom	m	s
JE	6.0	6.0
AL	8.0	8.0
AN	6.0	7.0
MO	14.5	14.5
DI	14.0	14.0
AD	11.0	10.0
PI	5.5	7.0
BR	13.0	12.5
EV	9.0	9.5

nom	m.c	s.c
JE	-3.7	-3.8
AL	-1.7	-1.8
AN	-3.7	-2.8
MO	4.8	4.7
DI	4.3	4.2
AD	1.3	0.2
PI	-4.2	-2.8
BR	3.3	2.7
EV	-0.7	-0.3

Si $p_i = \frac{1}{9}$, i = 1, ..., 9

$$< m, s >_D = \frac{1}{9}(6 \times 6 + 8 \times 8 + 6 \times 7 + 14.5 \times 14.5 + 14 \times 14 + 11 \times 10 + 5.5 \times 7 + 13 \times 12.5 + 9 \times 9.5) = 104.97$$

$$< m.c, s.c >_D = \frac{1}{9}((-3.7) \times (-3.8) + (-1.7) \times (-1.8) + (-3.7) \times (-2.8) + (4.8) \times (4.7) + (4.3) \times (4.2) + (1.3) \times (0.2) + (-4.2) \times (-2.8) + (3.3) \times (2.7) + (-0.7) \times (-0.3)) = 9.92$$

C'est la covariance de ces deux variables (car les variables sont centrées).

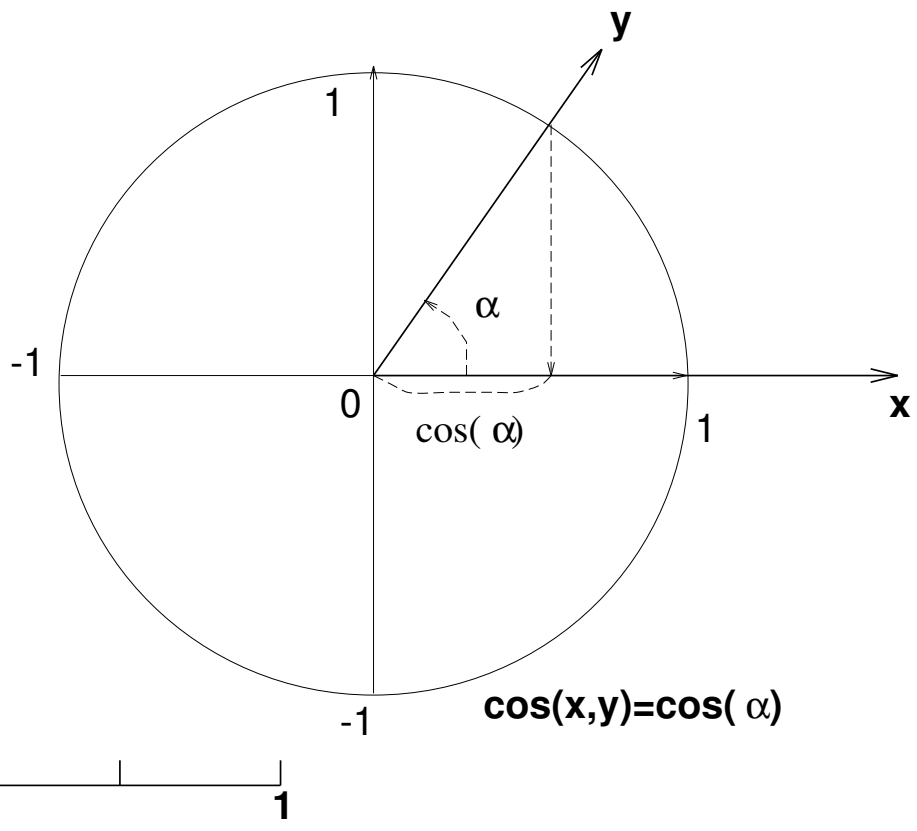
Le cosinus

On note $\cos(\mathbf{x}, \mathbf{y})$ la quantité

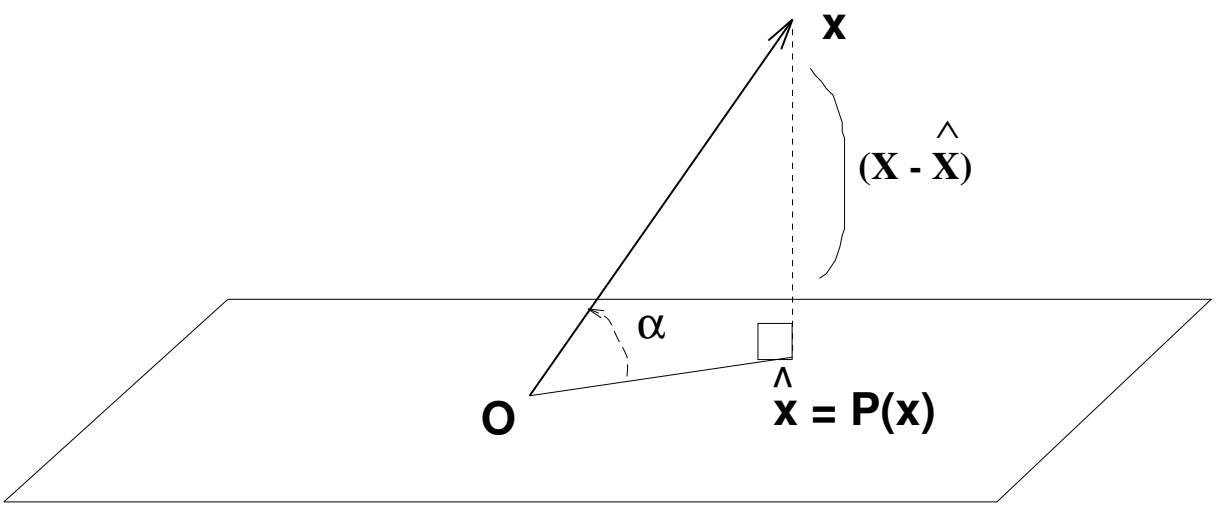
$$\cos(\mathbf{x}, \mathbf{y}) = \frac{< \mathbf{x}, \mathbf{y} >_D}{\|\mathbf{x}\|_D \|\mathbf{y}\|_D}$$

Propriétés

- $\cos(\mathbf{x}, \mathbf{y})$ varie entre -1 et +1.
- $\cos(\mathbf{x}, \mathbf{y}) = +1 \Leftrightarrow \mathbf{x}$ et $\mathbf{y}$ colinéaires et de même sens.
- $\cos(\mathbf{x}, \mathbf{y}) = -1 \Leftrightarrow \mathbf{x}$ et $\mathbf{y}$ colinéaires et de sens contraires.
- $\cos(\mathbf{x}, \mathbf{y}) = 0 \Leftrightarrow x \perp \mathbf{y}$.



Projection orthogonale



Pythagore :

$$\|\mathbf{x}\|^2 = \|\hat{\mathbf{x}}\|^2 + \|\mathbf{x} - \hat{\mathbf{x}}\|^2$$

$$\cos^2 \alpha = \frac{\|\hat{\mathbf{x}}\|^2}{\|\mathbf{x}\|^2}$$

représente le carré de la proportion de la longueur de $\mathbf{x}$ représenté par la projection.

$\Longrightarrow$ Coefficient de qualité de la représentation de $\mathbf{x}$ sur le plan.

Interprétation géométrique d'indices statistiques

- **moyenne pondérée** $\bar{x} = \sum_{i=1}^{n} p_i x_i$
- on note $\tilde{\mathbf{x}}$ le vecteur $\mathbf{x}$ centré avec sa moyenne $\bar{x}$
- **variance** $\sigma_x^2 = \sum_{i=1}^{n} p_i (x_i - \bar{x})^2 = \|\tilde{\mathbf{x}}\|_D^2$
- **covariance** $cov(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} p_i (x_i - \bar{x})(y_i - \bar{y}) = < \tilde{\mathbf{x}}, \tilde{\mathbf{y}} >_D$
- **coefficient de corrélation linéaire** $\rho_{xy} = \dfrac{< \tilde{\mathbf{x}}, \tilde{\mathbf{y}} >_D}{\|\tilde{\mathbf{x}}\|_D \|\tilde{\mathbf{y}}\|_D}$

  c'est le cosinus de l'angle que font $\tilde{\mathbf{x}}$ et $\tilde{\mathbf{y}}$

Décomposition aux valeurs singulières généralisées (DVSG)

Notion de rang

- Soit la matrice $\underset{n\times p}{X} = \left(\underset{n\times 1}{\mathbf{x}^1} \middle| \dots \middle| \underset{n\times 1}{\mathbf{x}^p} \right)$
- Les combinaisons linéaires des p vecteurs $\mathbf{x}^j$ (colonnes de X) dans $\mathbb{R}^n$:

  $X\alpha = \sum_{j=1}^{p} \alpha_j \mathbf{x}^j$ avec $\alpha = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_p \end{pmatrix}$

  sont les éléments de l'espace engendré par les colonnes de X.
- La dimension du sous-espace de $\mathbb{R}^n$ engendré par les p vecteurs colonnes de X est égale à celle du sous-espace de $\mathbb{R}^p$ engendré par ses n vecteurs lignes. Par définition, c'est le **rang** de X.

Exemple : matrice 5×3 de rang 1

$$X = \begin{pmatrix} 1 & 2 & -2 \\ 4 & 8 & -8 \\ 8 & 16 & -16 \\ 5 & 10 & -10 \\ 4 & 8 & -6 \end{pmatrix}$$

Les colonnes de X sont les vecteurs $\mathbf{x}$, $2\mathbf{x}$ et $-2\mathbf{x}$ avec $\mathbf{x} = \begin{pmatrix} 1 \\ 4 \\ 8 \\ 5 \\ 3 \end{pmatrix}$

En effet, $\begin{pmatrix} 1 \\ 4 \\ 8 \\ 5 \\ 3 \end{pmatrix} \begin{pmatrix} 1 & 2 & -2 \end{pmatrix} = \begin{pmatrix} 1 & 2 & -2 \\ 4 & 8 & -8 \\ 8 & 16 & -16 \\ 5 & 10 & -10 \\ 4 & 8 & -6 \end{pmatrix} = X$

- Construction d'une matrice de rang 2

  Soient $\mathbf{u}^1$ et $\mathbf{u}^2$ deux vecteurs quelconques de $\mathbb{R}^n$
- $\mathbf{x}^j = v_1^j \mathbf{u}^1 + v_2^j \mathbf{u}^2 \qquad\qquad j = 1, \ldots, p$

  engendrés par $\mathbf{u}^1$ et $\mathbf{u}^2$
- $\underset{n \times p}{X} = \begin{pmatrix} u_1^1 & u_1^2 \\ \vdots & \vdots \\ u_n^1 & u_n^2 \end{pmatrix} \begin{pmatrix} v_1^1 & \ldots & v_1^p \\ v_2^1 & \ldots & v_2^p \end{pmatrix} = \mathbf{u}^1 \, {}^t\mathbf{v}_1 + \mathbf{u}^2 \, {}^t\mathbf{v}_2$

Décomposition aux valeurs singulières généralisées (DVSG) d'une matrice de rang 2 et généralisation

- Soit $\underset{n\times p}{X}$ de rang 2, on sait construire :
  - $\mathbf{u}^1$ et $\mathbf{u}^2$ de $\mathbb{R}^n$, **orthonormés** pour la métrique D
  - $\mathbf{v}_1$ et $\mathbf{v}_2$ de $\mathbb{R}^p$, **orthonormés** pour la métrique M
  - deux scalaires positifs $\lambda_1 \geq \lambda_2$

tels que

$$X = \begin{pmatrix} \mathbf{u}^1 & \mathbf{u}^2 \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \begin{pmatrix} {}^t\mathbf{v}_1 \\ {}^t\mathbf{v}_2 \end{pmatrix} = U\Lambda\, {}^tV = \sum_{s=1}^{2} \lambda_s \mathbf{u}^s\, {}^t\mathbf{v}_s$$

- $\mathbf{u}^1$ et $\mathbf{u}^2$ sont les **vecteurs singuliers à gauche**
- $\mathbf{v}_1$ et $\mathbf{v}_2$ sont les **vecteurs singuliers à droite**
- λ_1 et λ_2 sont les **valeurs singulières**

- La DVSG générale d'une matrice de rang r s'écrit :
  $X = \sum_{s=1}^{r} \lambda_s \mathbf{u}^s\, {}^t\mathbf{v}_s$
  avec des propriétés analogues.

Application du théorème d'Eckart-Young (1936) à l'approximation d'une matrice et interprétation géométrique

But : construction d'une matrice de faible rang (fixé) la plus proche d'une matrice X.

Soit X, de rang r, dont la DVSG s'écrit :

$$X = \sum_{s=1}^{r} \lambda_s \mathbf{u}^s \, {}^t\mathbf{v}_s$$

alors la matrice, notée $\hat{X}^k$, de rang k $(k \leq r)$ la plus proche de X est la matrice

$$\widehat{X}^k = \sum_{s=1}^{k} \lambda_s \mathbf{u}^s \, {}^t\mathbf{v}_s$$

$$\widehat{X}^k = \begin{pmatrix} u_1^1 & \dots & u_1^k \\ \dots & \dots & \dots \\ u_i^1 & \dots & u_i^k \\ \dots & \dots & \dots \\ u_n^1 & \dots & u_n^k \end{pmatrix} \begin{pmatrix} \lambda_1 & \dots & 0 \\ \dots & \dots & \dots \\ 0 & \dots & \lambda_k \end{pmatrix} \begin{pmatrix} v_1^1 & \dots & v_1^j & \dots & v_1^p \\ \dots & \dots & \dots & \dots & \dots \\ v_k^1 & \dots & v_k^j & \dots & v_k^p \end{pmatrix}$$

Pour obtenir l'expression de l'approximation de rang k de X, il suffit de supprimer dans la DVSG de X les termes de rang supérieurs à k.

Exemple “note” :

$$X = \begin{array}{c} \\ JE \\ AL \\ AN \\ MO \\ DI \\ AD \\ PI \\ BR \\ EV \end{array} \begin{pmatrix} m & s & f & l & dm \\ 6 & 6 & 5 & 5.5 & 8 \\ 8 & 8 & 8 & 8 & 9 \\ 6 & 7 & 11 & 9.5 & 11 \\ 14.5 & 14.5 & 15.5 & 15 & 8 \\ 14 & 14 & 12 & 12.5 & 10 \\ 11 & 10 & 5.5 & 7 & 13 \\ 5.5 & 7 & 14 & 11.5 & 10 \\ 13 & 12.5 & 8.5 & 9.5 & 12 \\ 9 & 9.5 & 12.5 & 12 & 18 \end{pmatrix}$$

$$= \begin{pmatrix} \mathbf{u}^1 & \mathbf{u}^2 & \mathbf{u}^3 & \mathbf{u}^4 & \mathbf{u}^5 \\ -0.59 & -0.25 & 0.54 & 0.26 & -1.45 \\ -0.79 & -0.05 & 0.15 & 0.21 & 0.27 \\ -0.86 & -0.96 & -0.60 & -1.29 & 0.21 \\ -1.29 & 1.57 & -1.17 & 1.27 & 0.96 \\ -1.20 & 1.16 & -0.01 & -0.38 & -1.80 \\ -0.89 & -0.11 & 1.75 & 0.14 & 1.45 \\ -0.93 & -0.94 & -1.60 & -0.90 & 0.43 \\ -1.06 & 0.56 & 1.10 & -1.46 & 0.25 \\ -1.18 & -1.72 & 0.30 & 1.58 & -0.49 \end{pmatrix} \begin{pmatrix} 23.26 & 0 & 0 & 0 & 0 \\ 0 & 3.61 & 0 & 0 & 0 \\ 0 & 0 & 3.41 & 0 & 0 \\ 0 & 0 & 0 & .456 & 0 \\ 0 & 0 & 0 & 0 & .1 \end{pmatrix}$$

$$\begin{pmatrix} -.429 & -.436 & -.453 & -.446 & -.471 \\ .518 & .414 & -.135 & .01 & -.736 \\ .365 & .196 & -.675 & -.37 & .486 \\ .318 & -.575 & -.371 & .656 & -.022 \\ .559 & -.519 & .428 & -.484 & .017 \end{pmatrix} \begin{matrix} {}^t\mathbf{v}_1 \\ {}^t\mathbf{v}_2 \\ {}^t\mathbf{v}_3 \\ {}^t\mathbf{v}_4 \\ {}^t\mathbf{v}_5 \end{matrix}$$

$= U\Lambda {}^tV$

U est la matrice dont les colonnes sont $\mathbf{u}^1, \ldots, \mathbf{u}^5$ normés par D,
V est la matrice dont les colonnes sont $\mathbf{v}_1, \ldots, \mathbf{v}_5$ normés par M,
X est de rang 5.

$$\hat{X}^k = \sum_{s=1}^{k} \lambda_s u^s \, {}^t v_s$$

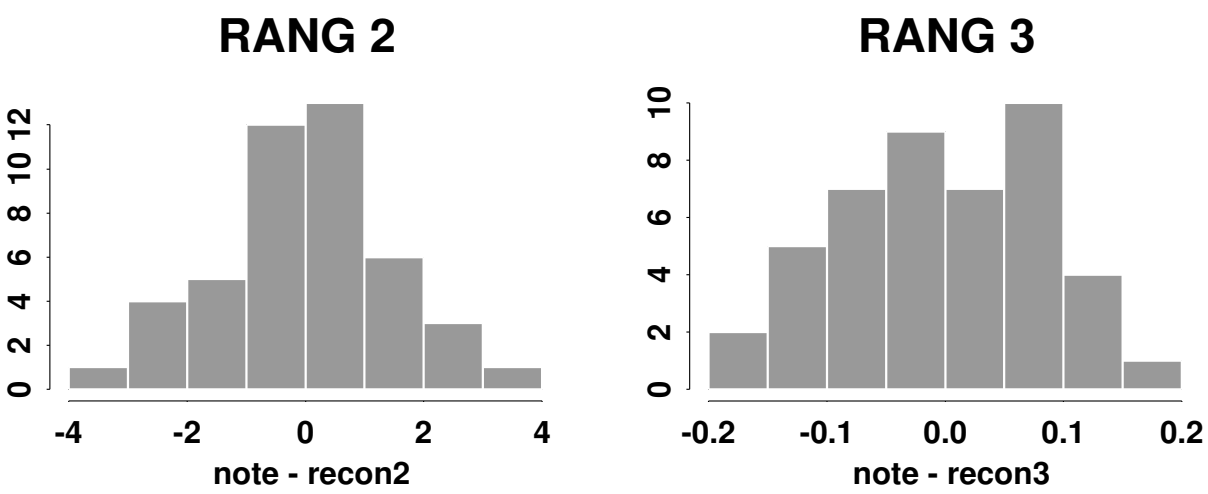
Approximation de rang 2 du tableau "note"

	m	s	f	l	dm
JE	5.4	5.6	6.3	6.1	7.1
AL	7.8	7.9	8.3	8.2	8.8
AN	6.8	7.3	9.5	8.9	12.0
MO	15.8	15.4	12.8	13.4	9.9
DI	14.1	13.9	12.0	12.4	10.0
AD	8.7	8.9	9.5	9.3	10.1
PI	7.5	8.0	10.2	9.6	12.7
BR	11.7	11.6	10.9	11.1	10.2
EV	8.6	9.4	13.3	12.2	17.5

Approximation de rang 3 du tableau "note"

	m	s	f	l	dm
JE	6.1	5.9	5.1	5.4	8
AL	8	8	8	8	9
AN	6.1	6.9	10.9	9.6	11
MO	14.4	14.7	15.5	14.9	8
DI	14.1	13.9	12.1	12.5	10
AD	10.9	10.1	5.4	7.1	13
PI	5.5	6.9	13.9	11.6	10
BR	13.1	12.4	8.4	9.7	12
EV	8.9	9.6	12.6	11.8	18

Analyse des résidus $X - \hat{X}$



Conséquences du théorème d'Eckart-Young

Décomposition exacte	Approximation de rang k $(k \leq r)$
Elément par élément	
$x_i^j = \sum_{s=1}^{r} \lambda_s u_i^s v_s^j$	$\widehat{x_i^j}^k = \sum_{s=1}^{k} \lambda_s u_i^s v_s^j$
Par ligne (dans $\mathbb{R}^p$)	
$\mathbf{x}_i = \sum_{s=1}^{r} (\lambda_s u_i^s)\mathbf{v}_s$	$\widehat{\mathbf{x}_i}^k = \sum_{s=1}^{k} (\lambda_s u_i^s)\mathbf{v}_s$
Par colonne (dans $\mathbb{R}^n$)	
$\mathbf{x}^j = \sum_{s=1}^{r} (\lambda_s v_s^j)\mathbf{u}^s$	$\widehat{\mathbf{x}^j}^k = \sum_{s=1}^{k} (\lambda_s v_s^j)\mathbf{u}^s$

(1)	(2)
(3)	(4)
(5)	(6)

Interprétation géométrique

Soit une base orthogonale $(\mathbf{u}^s)$, $s = 1, \ldots, r$ de $\mathbb{R}^n$.

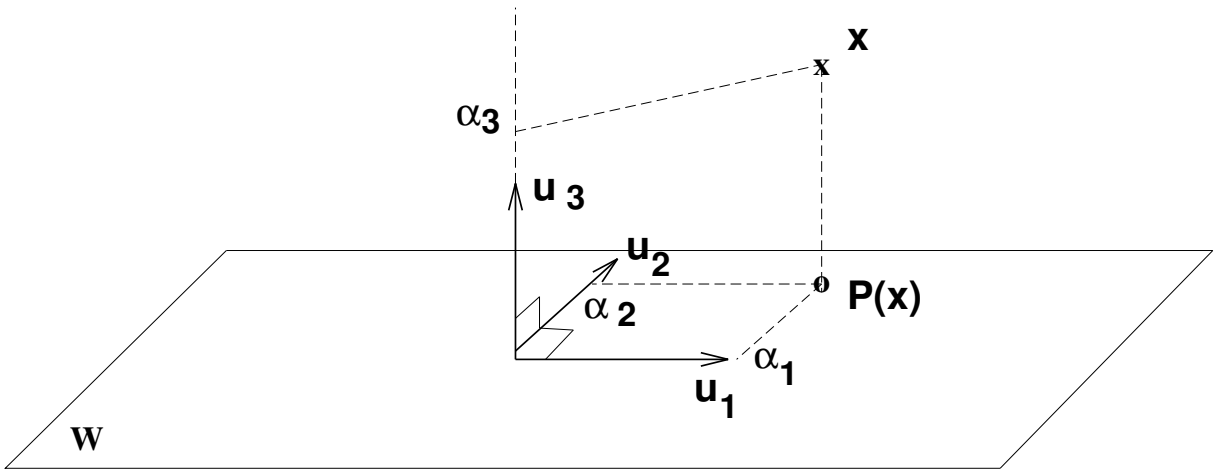
Projeter sur le sous-espace W engendré par certains vecteurs de la base $(\mathbf{u}^s)$, $s = 1, \ldots, k$ $(k \leq r)$, revient à annuler les coordonnées des vecteurs qui n'engendrent pas W.

Exemple avec $r = 3$:

$\mathbf{x} = \alpha_1 \mathbf{u}^1 + \alpha_2 \mathbf{u}^2 + \alpha_3 \mathbf{u}^3$.

P, projecteur orthogonal sur le plan engendré par $(\mathbf{u}^1, \mathbf{u}^2)$.

$P(\mathbf{x}) = \alpha_1 \mathbf{u}^1 + \alpha_2 \mathbf{u}^2$



Formules de transition

De la DVSG de X,

$$X = \sum_{s=1}^{r} \lambda_s \mathbf{u}^s \, {}^t\mathbf{v}_s$$

on déduit :

$$XM\mathbf{v}_{s'} = (\sum_{s=1}^{r} \lambda_s \mathbf{u}^s \, {}^t\mathbf{v}_s) M\mathbf{v}_{s'} = \sum_{s=1}^{r} \lambda_s \mathbf{u}^s \, ({}^t\mathbf{v}_s M\mathbf{v}_{s'})$$

De ${}^t\mathbf{v}_s M\mathbf{v}_{s'} = 0$ si $s \neq s'$ et ${}^t\mathbf{v}_s M\mathbf{v}_{s'} = 1$ si $s = s'$ on déduit :

$$XM\mathbf{v}_s = \lambda_s \mathbf{u}^s \Longrightarrow \mathbf{u}^s = XM\mathbf{v}_s/\lambda_s$$

De même, l'expression de $\mathbf{v}_s$ en fonction de $\mathbf{u}^s$ est :

$$\mathbf{v}_s = {}^tXD\mathbf{u}^s/\lambda_s$$

Analyse générale d'un tableau X

- Construction d'indices de qualité de l'approximation du tableau.
- Représentation graphique de l'approximation de faible dimension :
  - du nuage de ses lignes
  - du nuage de ses colonnes

L'inertie totale de la matrice X

Définie comme le carré de la norme de X, elle mesure globalement l'importance des éléments de la matrice :

$$\begin{aligned}
\text{In(X)} &= \|X\|^2 \\
&= \sum_i p_i \|\mathbf{x}_i\|_M^2 \\
&\quad \text{(somme pondérée des carrés des longueurs des lignes)} \\
&= \sum_j m_j \|\mathbf{x}^j\|_D^2 \\
&\quad \text{(somme pondérée des carrés des longueurs des colonnes)} \\
&= \sum_i \sum_j p_i m_j (x_i^j)^2 \\
&\quad \text{(somme pondérée des carrés des éléments de la matrice)}
\end{aligned}$$

$\dfrac{p_i \|\mathbf{x}_i\|_M^2}{\|X\|^2}$ = **contribution** d'un individu à l'inertie totale.

$\dfrac{m_j \|\mathbf{x}^j\|_D^2}{\|X\|^2}$ = **contribution** d'une variable à l'inertie totale.

Approximation globale de X

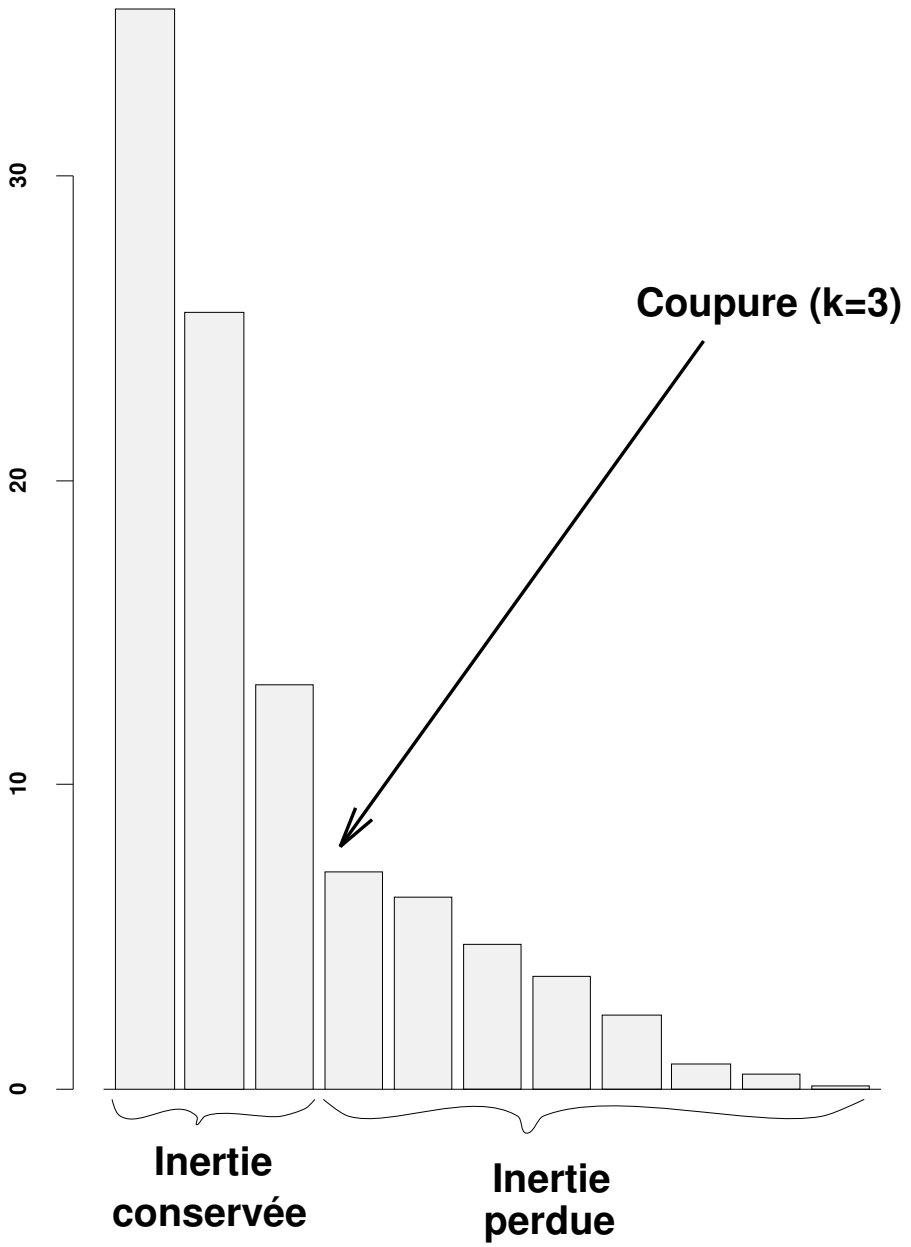
$$\|X\|^2 = \sum_{i=1}^{n} p_i \|\mathbf{x}_i\|_M^2 = \sum_{j=1}^{p} m_j \|\mathbf{x}^j\|_D^2 = \sum_{s=1}^{r} (\lambda_s)^2$$

$$\|\hat{X}\|^2 = \sum_{i=1}^{n} p_i \|\hat{\mathbf{x}}_i\|_M^2 = \sum_{j=1}^{p} m_j \|\hat{\mathbf{x}}^j\|_D^2 = \sum_{s=1}^{k} (\lambda_s)^2$$

$$\frac{\sum_{s=1}^{k} (\lambda_s)^2}{\sum_{s=1}^{r} (\lambda_s)^2} \times 100 = \text{ pourcentage d'inertie expliquée}$$

par les k premières dimensions.

Éboulis des valeurs propres : diagramme des $\{ \lambda_s^2,\ s = 1, \ldots, r \}$ (ordonnées par construction)



Décomposition selon les lignes et les colonnes

Exemple

note	m	s	f	l	dm	inertie des lignes
JE	6.0	6.0	5.0	5.5	8.0	21.3
AL	8.0	8.0	8.0	8.0	9.0	37.4
AN	6.0	7.0	11.0	9.5	11.0	46.4
MO	14.5	14.5	15.5	15.0	8.0	105.5
DI	14.0	14.0	12.0	12.5	10.0	88.0
AD	11.0	10.0	5.5	7.0	13.0	52.1
PI	5.5	7.0	14.0	11.5	10.0	56.4
BR	13.0	12.5	8.5	9.5	12.0	70.2
EV	9.0	9.5	12.5	12.0	18.0	88.4
inertie des colonnes	104.8	105.6	116.6	109.0	129.7	565.7

Contribution des lignes (en millièmes)

JE	AL	AN	MO	DI	AD	PI	BR	EV
38	66	82	187	156	92	100	124	156

Contribution des colonnes (en millièmes)

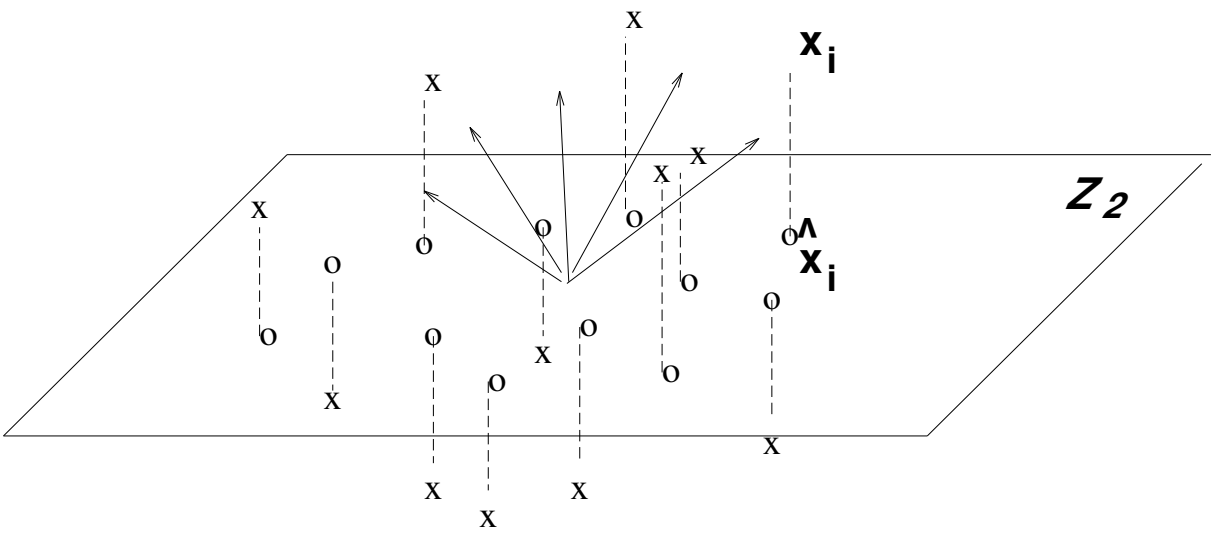
m	s	f	l	dm
185	187	206	193	229

Décomposition par ligne :

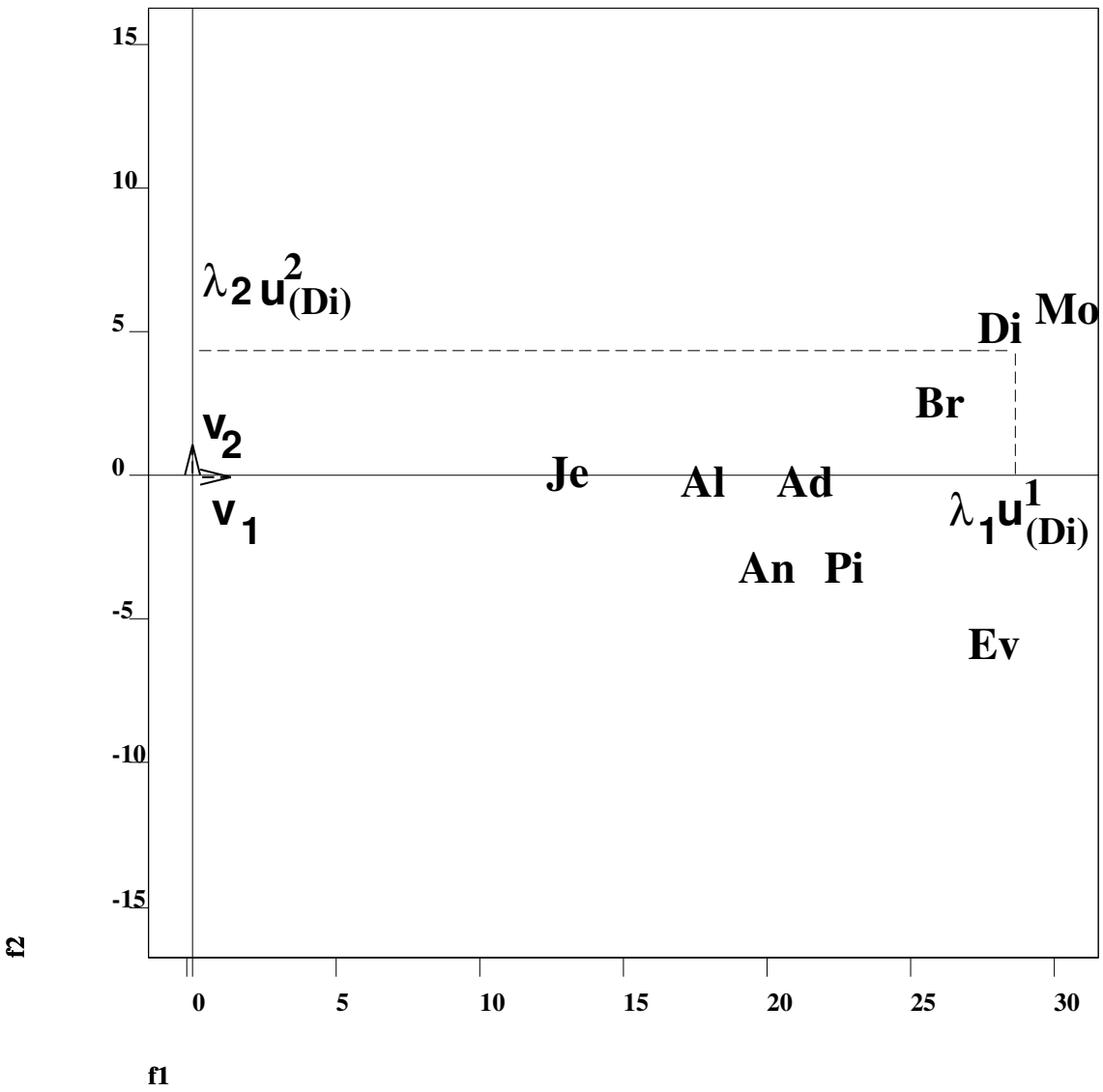
$$\mathbf{x}_i = \sum_{s=1}^{r} (\lambda_s u_i^s)\mathbf{v}_s \qquad \widehat{x_i}^k = \sum_{s=1}^{k} (\lambda_s u_i^s)\mathbf{v}_s$$

$\lambda_s u_i^s$ représente les coordonnées de $\mathbf{x}_i$ dans la base $(\mathbf{v}_s)$.

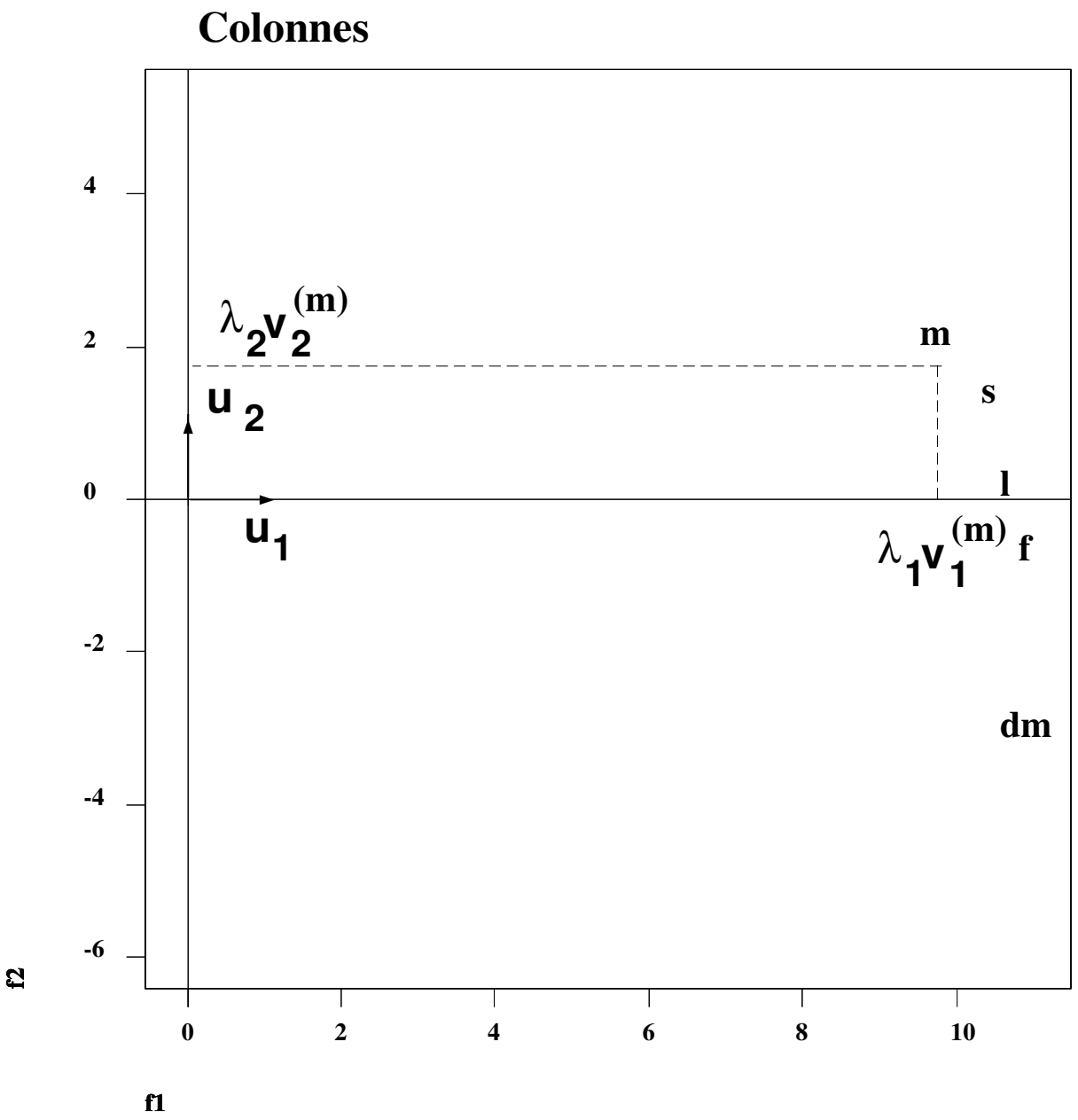
Nuage des lignes projeté sur le sous-espace $Z_2 = vect(\mathbf{v}_1, \mathbf{v}_2)$ de $\mathbb{R}^p$ (approximation de rang 2) :


x
x
x
x
x
x
x
x
x
x
x
x
x
x
x
x
x
x
x
x
o
o
o

Exemple
Nuage des lignes projeté sur Z_2



Nuage des colonnes projeté sur $W_2 = vect(\mathbf{u}^1, \mathbf{u}^2)$



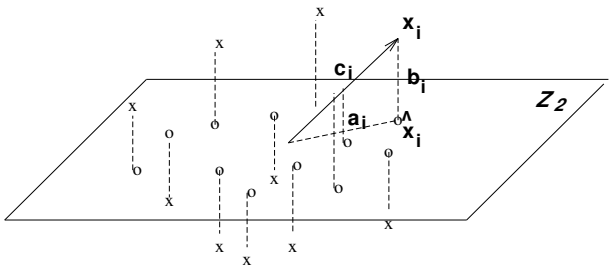
Critères équivalents (lignes de X)

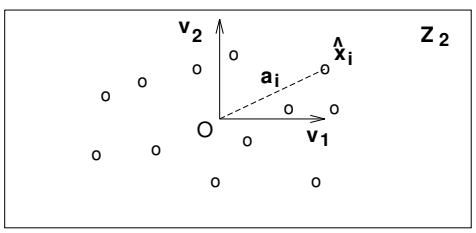
$\sum_{i=1}^{n} p_i \|\widehat{\mathbf{x}}_i^k\|_M^2$ est l'inertie du nuage projeté des lignes de X sur $Z_k \subset \mathbb{R}^p$, $(Z_k = vect(\mathbf{v}_1, \ldots, \mathbf{v}_k))$.

On a les équivalences :

- la matrice approchée $\widehat{X}^k$ est la matrice de rang k la plus proche de X.
- Le sous-espace Z_k est le sous-espace de $\mathbb{R}^p$ qui, parmi les sous-espaces de dimension k, rend maximum l'inertie du nuage projeté des lignes de X ($\sum_i p_i \|\hat{\mathbf{x}}_i\|_M^2$ maximum).

Th. de Pythagore $\Rightarrow \sum_i p_i \|x_i\|_M^2 = \sum_i p_i \|\hat{\mathbf{x}}_i\|_M^2 + \sum_i p_i \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|_M^2$





Inertie totale du nuage des lignes = Inertie du nuage projeté des lignes + somme pondérée des distances des lignes x_i au sous-espace Z_k.

Critères équivalents (colonnes de X)

De même $\sum_{j=1}^{p} m_j \|\widehat{\mathbf{x}^j}^k\|_D^2$ est l'inertie du nuage projeté des colonnes de X sur le sous-espace $W_k \subset \mathbb{R}^n$ ($W_k = vect(\mathbf{u}^1, \ldots, \mathbf{u}^k)$). On a les équivalences suivantes :

- la matrice approchée $\widehat{X}^k$ est la matrice de rang k la plus proche de X.
- W_k est le sous-espace de dimension k de $\mathbb{R}^n$ tel qu'en projection l'inertie du nuage des colonnes est maximum ($\sum_j m_j \|\widehat{\mathbf{x}^j}^k\|_D^2$ maximum).

Th. de Pythagore $\Rightarrow \sum_j m_j \|\mathbf{x}^j\|_D^2 = \sum_j m_j \|\hat{\mathbf{x}}^j\|_D^2 + \sum_j m_j \|\mathbf{x}^j - \hat{\mathbf{x}}^j\|_D^2$

L'inertie totale est la somme de

– l'inertie du nuage des projetés des colonnes

– la somme pondérée des distances des $\mathbf{x}^j$ au sous-espace W_k.

Double décomposition de l'inertie (colonnes × facteurs)

Facteur	1	...	s	...	k	Sous-total	...	r	Total
Colonne									
1	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...
j	...	...	$m_j(\lambda_s v_s^j)^2$	...	...	$m_j\|\widehat{\mathbf{x}^j}^k\|_D^2$	...	...	$m_j\|\mathbf{x}^j\|_D^2$
...	...	...	...	...	...	...	...	...	...
J	...	...	...	...	...	...	...	...	...
Total	λ_1^2	...	λ_s^2	...	λ_k^2	$\|\widehat{X}^k\|^2$	...	λ_r^2	$\|X\|^2$

Table 1 : la décomposition "colonne × facteur" de l'inertie.

– contribution du facteur s à l'inertie totale : $\dfrac{\lambda_s^2}{\|X\|^2}$

– contribution de la colonne j à l'inertie totale : $\dfrac{m_j\|\mathbf{x}^j\|_D^2}{\|X\|^2}$

– contribution du facteur s à la colonne j : $\dfrac{(\lambda_s v_s^j)^2}{\|\mathbf{x}^j\|_D^2}$

– contribution de la colonne j au facteur s : $m_j(v_s^j)^2$

Double décomposition de l'inertie (lignes × facteurs)

Facteur	1	...	s	...	k	Sous-total	...	r	Total
Individu									
1	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...
i	...	...	$p_i(\lambda_s u_i^s)^2$	...	...	$p_i\|\widehat{\mathbf{x}_i}^k\|_M^2$	...	...	$p_i\|\mathbf{x}_i\|_M^2$
...	...	...	...	...	...	...	...	...	...
n	...	...	...	...	...	...	...	...	...
Total	λ_1^2	...	λ_s^2	...	λ_k^2	$\|\widehat{X}^k\|^2$	...	λ_r^2	$\|X\|^2$

Table 2 : la décomposition "individu × facteur" de l'inertie.

Les mesures de qualité

1. Qualité globale de l'approximation d'ordre k :

$$\frac{\|\widehat{X}^k\|^2}{\|X\|^2} = \frac{\sum_{s=1}^{k} (\lambda_s)^2}{\sum_{s=1}^{r} (\lambda_s)^2}$$

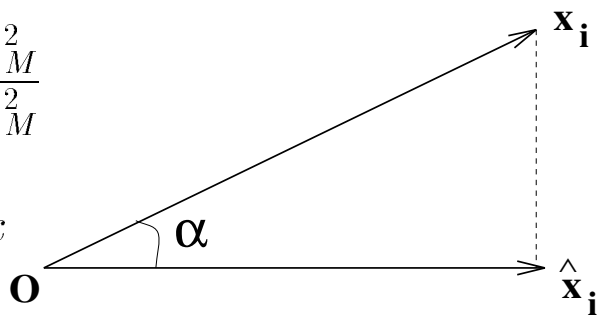
2. Qualité de l'approximation des lignes et des colonnes :

Approximation des lignes

$$\|\mathbf{x}_i\|_M^2 = \|\hat{\mathbf{x}}_i\|_M^2 + \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|_M^2$$

$$\cos(\alpha)^2 = \cos(\mathbf{x}_i, \hat{\mathbf{x}}_i)^2 = \frac{\|\hat{x}_i\|_M^2}{\|x_i\|_M^2}$$

Carré de la proportion de la longueur de $\mathbf{x}_i$ expliquée en k dimensions.



Approximation des colonnes

$$\|\mathbf{x}^j\|_D^2 = \|\hat{\mathbf{x}}^j\|_D^2 + \|\mathbf{x}^j - \hat{\mathbf{x}}^j\|_D^2$$

$$\cos(\mathbf{x}^j, \hat{\mathbf{x}}^j)^2 = \frac{\|\hat{\mathbf{x}}^j\|_D^2}{\|\mathbf{x}^j\|_D^2}$$

Carré de la proportion de la longueur de $\mathbf{x}^j$ expliquée en k dimensions.

Illustration numérique

Édition des contributions

Contribution des colonnes à l'inertie

colonne	facteur 1	facteur 2	facteur 3	facteur 4	facteur 5	total	vari
m	184	269	133	101	313	185	233
s	190	171	39	330	270	187	183
f	205	18	456	138	184	206	246
l	199	0	137	430	234	193	162
dm	222	542	236	0	0	229	177
total	1000	1000	1000	1000	1000	1000	1000

Contribution des lignes à l'inertie

ligne	facteur 1	facteur 2	facteur 3	facteur 4	facteur 5	total	Norme^2
JE	38	7	32	7	235	38	38
AL	69	0	2	5	8	66	66
AN	82	103	40	185	5	82	82
MO	185	276	153	178	103	187	187
DI	159	150	0	16	361	156	156
AD	89	1	342	2	234	92	92
PI	96	98	286	89	20	100	100
BR	126	35	135	238	7	124	124
EV	155	330	10	279	27	156	156
total	1000	1000	1000	1000	1000	1000	1000

Remarque : l'avant dernière colonne donne (en millième) la contribution de la ligne à l'inertie totale ($p_i \|\mathbf{x}_i\|^2_M / \|X\|^2$).
La dernière est égale à $\|\mathbf{x}_i\|^2_M / \sum \|\mathbf{x}_i\|^2_M$.

Edition des contributions

Contribution des facteurs aux colonnes

colonne	facteur 1	facteur 2	facteur 3	facteur 4	facteur 5	total
m	951.78	33.41	14.74	2.35e-02	2.98e-02	1000
s	974.53	21.12	4.24	7.63e-02	2.55e-02	1000
f	952.45	2.03	45.46	2.87e-02	1.57e-02	1000
l	985.23	0.01	14.63	9.62e-02	2.144e-02	1000
dm	924.36	54.47	21.16	8.95e-05	2.159e-05	1000
total	956.33	23.04	20.56	4.31e-02	1.76e-02	1000

Contribution des facteurs aux lignes

ligne	facteur 1	facteur 2	facteur 3	facteur 4	facteur 5	total
JE	977.82	4.38	1.76e+01	0.008	0.110	1000
AL	999.18	0.08	7.27e-01	0.003	0.002	1000
AN	960.92	29.00	9.96e+00	0.097	0.001	1000
MO	949.05	34.04	1.68e+01	0.041	0.009	1000
DI	977.73	22.21	5.46e-04	0.004	0.041	1000
AD	923.36	0.34	7.62e+01	0.001	0.044	1000
PI	918.34	22.60	5.90e+01	0.038	0.003	1000
BR	970.97	6.49	2.24e+01	0.082	0.0011	1000
EV	950.02	48.62	1.27e+00	0.076	0.003	1000
total	956.33	23.04	2.05e+01	0.043	0.017	1000

Qualité de la représentation des colonnes ($\cos^2$)

colonne	facteur 1	facteur 2	facteur 3	facteur 4	facteur 5
m	952	985	1000	1000	1000
s	975	996	1000	1000	1000
f	952	954	1000	1000	1000
l	985	985	1000	1000	1000
dm	924	979	1000	1000	1000
total	956	979	1000	1000	1000

Qualité de la représentation des lignes ($\cos^2$)

ligne	facteur 1	facteur 2	facteur 3	facteur 4	facteur 5
JE	978	982	1000	1000	1000
AL	999	999	1000	1000	1000
AN	961	990	1000	1000	1000
MO	949	983	1000	1000	1000
DI	978	1000	1000	1000	1000
AD	923	924	1000	1000	1000
PI	918	941	1000	1000	1000
BR	971	977	1000	1000	1000
EV	950	999	1000	1000	1000
total	956	979	1000	1000	1000

L'Analyse en Composantes Principales (ACP)

Particularité de l'ACP par rapport à l'analyse générale

C'est une analyse générale d'un tableau X individus $\times$ variables centrées.

Objectifs

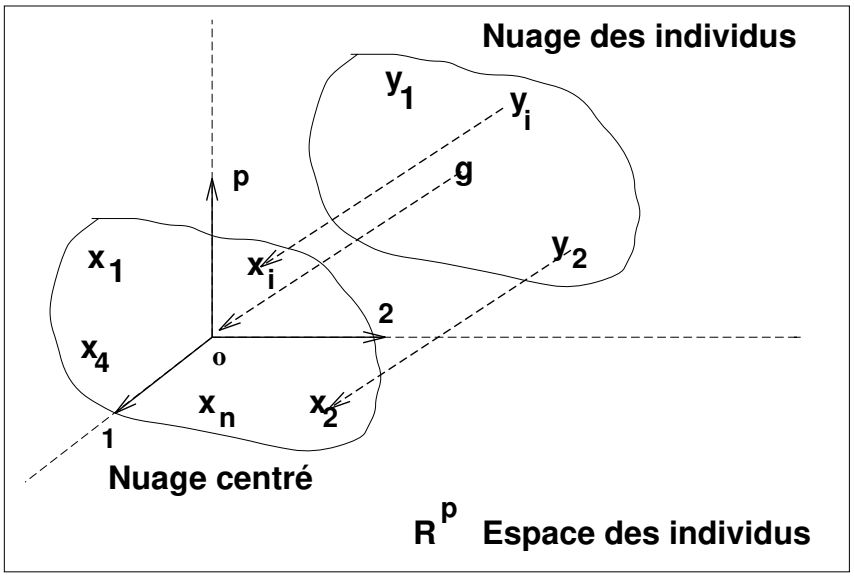
- Analyse des proximités des vecteurs individus.
- Analyse des corrélations entre les variables.
- Analyse simultanée : lorsque deux groupes d'individus se distinguent l'un de l'autre, pour quelles variables prennent-ils des valeurs différentes ?

Barycentre et nuage centré des individus

On note $Y = (y_i^j)$ (i individu, j variable) le tableau de données initial. Le nuage des individus est le nuage des vecteurs lignes $\mathbf{y}_i$ de Y dans $\mathbb{R}^p$.

Barycentre du nuage des individus

Le vecteur $\mathbf{g} = \sum_{i=1}^{n} p_i \mathbf{y}_i$ est le *barycentre* du nuage des individus.
Sa j*ième* coordonnée est $g_j = \sum_{i=1}^{n} p_i y_i^j$, moyenne pondérée de la variable $\mathbf{y}^j$.
$\mathbf{g}$ est encore l'*individu moyen* (les variables ont pour cet individu leurs valeurs moyennes).



Nuage des individus centrés

C'est le nuage des vecteurs $\mathbf{x}_i = \mathbf{y}_i - \mathbf{g}$, résultat de la translation du nuage initial par le vecteur $-\mathbf{g}$.

Le tableau centré

Tableau centré : tableau des écarts à la moyenne, ayant en position (i, j) $x_i^j = (y_i^j - \bar{y^j})$

$$Y = \begin{pmatrix} y_1^1 & \dots & y_1^j & \dots & y_1^p \\ \dots & \dots & \dots & \dots & \dots \\ y_i^1 & \dots & y_i^j & \dots & y_i^p \\ \dots & \dots & \dots & \dots & \dots \\ y_n^1 & \dots & y_n^j & \dots & y_n^p \end{pmatrix}$$ Tableau non centré

$$X = \begin{pmatrix} y_1^1 - \bar{y}^1 & \dots & y_1^j - \bar{y}^j & \dots & y_1^p - \bar{y}^p \\ \dots & \dots & \dots & \dots & \dots \\ y_i^1 - \bar{y}^1 & \dots & y_i^j - \bar{y}^j & \dots & y_i^p - \bar{y}^p \\ \dots & \dots & \dots & \dots & \dots \\ y_n^1 - \bar{y}^1 & \dots & y_n^j - \bar{y}^j & \dots & y_n^p - \bar{y}^p \end{pmatrix}$$ Tableau centré

Exemple

nom	m	s	f	l	dm
JE	6.0	6.0	5.0	5.5	8.0
AL	8.0	8.0	8.0	8.0	9.0
AN	6.0	7.0	11.0	9.5	11.0
MO	14.5	14.5	15.5	15.0	8.0
DI	14.0	14.0	12.0	12.5	10.0
AD	11.0	10.0	5.5	7.0	13.0
PI	5.5	7.0	14.0	11.5	10.0
BR	13.0	12.5	8.5	9.5	12.0
EV	9.0	9.5	12.5	12.0	18.0
g	9.7	9.8	10.2	10.1	11.0

tableau initial

← barycentre

nom	m	s	f	l	dm
JE	-3.7	-3.8	-5.2	-4.6	-3.0
AL	-1.7	-1.8	-2.2	-2.1	-2.0
AN	-3.7	-2.8	0.8	-0.6	0.0
MO	4.8	4.7	5.3	4.9	-3.0
DI	4.3	4.2	1.8	2.4	-1.0
AD	1.3	0.2	-4.7	-3.1	2.0
PI	-4.2	-2.8	3.8	1.4	-1.0
BR	3.3	2.7	-1.7	-0.6	1.0
EV	-0.7	-0.3	2.3	1.9	7.0

tableau centré

Le centrage dans l'ACP (rappel)

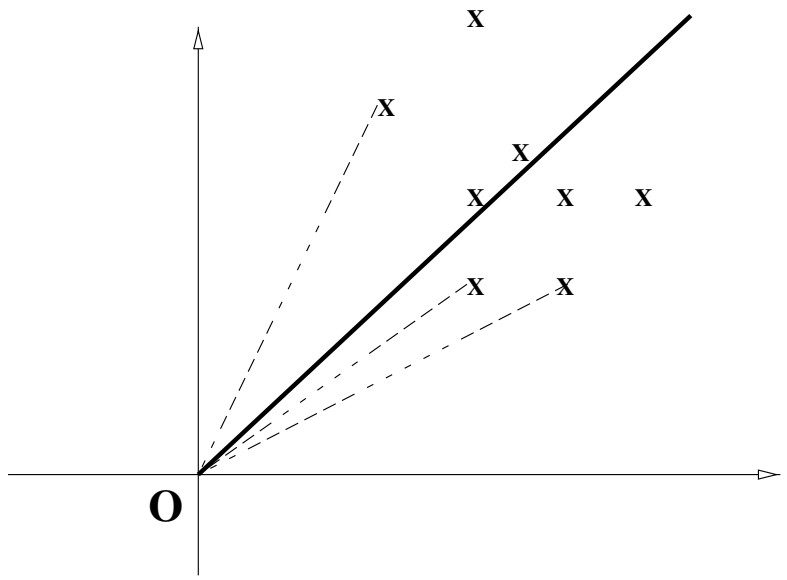
- centrage avec la moyenne
- interprétation du nuage des variables (covariance, corrélation, variance)
- le centrage ne modifie pas la distance entre lignes du tableau initial

INRA © 1998 - Formation statistique - Module 7/8 (ACP)-52

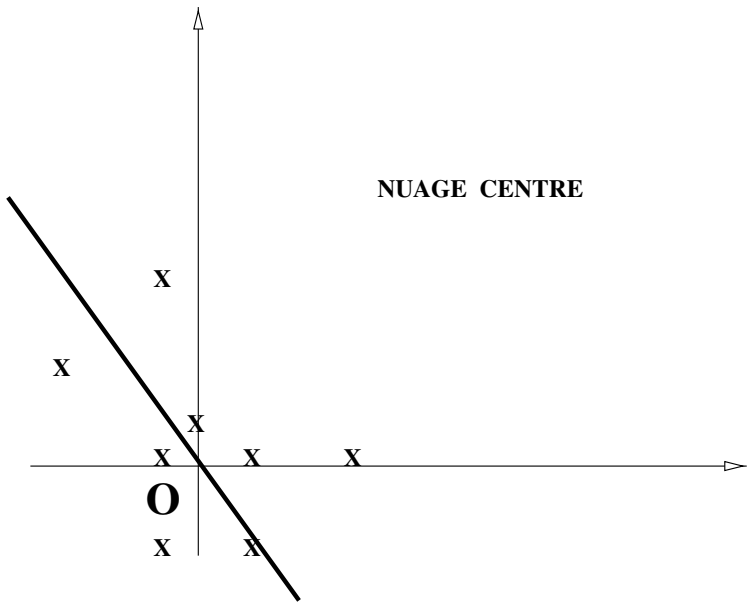
Nuage de points et valeur du coefficient de corrélation

Corrélation <u>linéaire</u> de 0.70 dans chaque cas

Effet du centrage



NUAGE
NON CENTRE



NUAGE CENTRE

$$\text{Inertie totale} = \sum_{i=1}^{n} p_i \|\mathbf{y}_i - \mathbf{g}\|^2 = \sum_{j=1}^{p} m_j \|\mathbf{x}^j\|^2$$

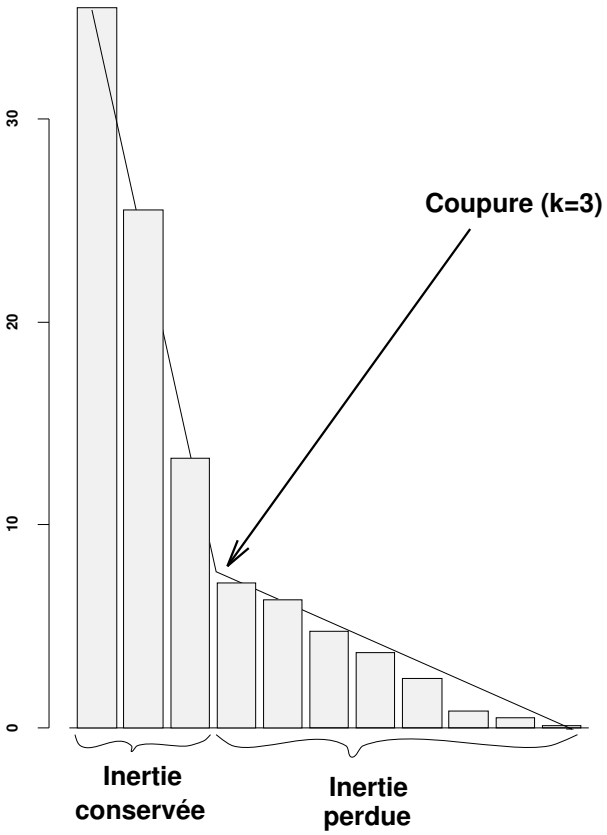
Pratique de l'ACP : bilan des choix initiaux

- X tableau de variables quantitatives ?
- unités identiques ?
- homogénéiser les variances ?
- définition de la métrique des variables
- définition de la métrique des individus
- choix les plus fréquents :
  ACP centrée ou **ACP centrée réduite**

Analyse des sorties de l'ACP

Choix du nombre de facteurs

But : obtenir le maximum d'inertie conservée avec le minimum de facteurs.
Règle du "coude" : on coupe l'éboulis des valeurs propres à l'endroit où celui-ci possède un "coude".

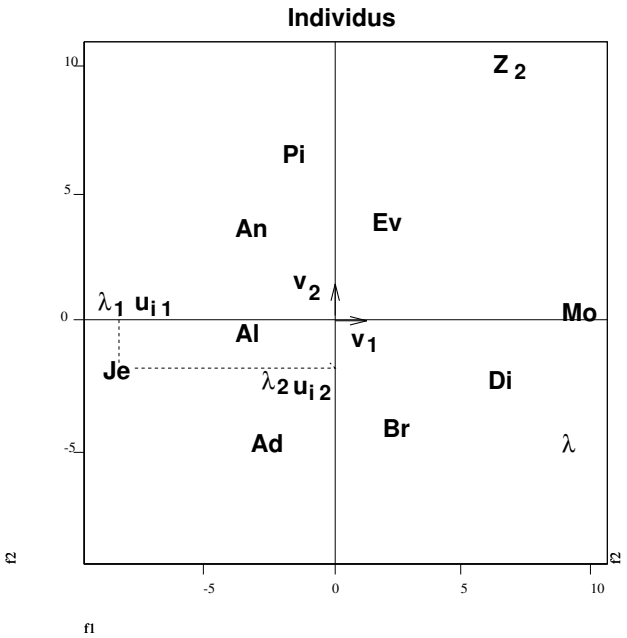


Le choix dépend des objectifs de l'analyse :
description des données $\longrightarrow$ en général au plus 4 ou 5 facteurs (difficultés d'interprétation).
Compression ou recodage $\longrightarrow$ ce nombre peut-être très grand.

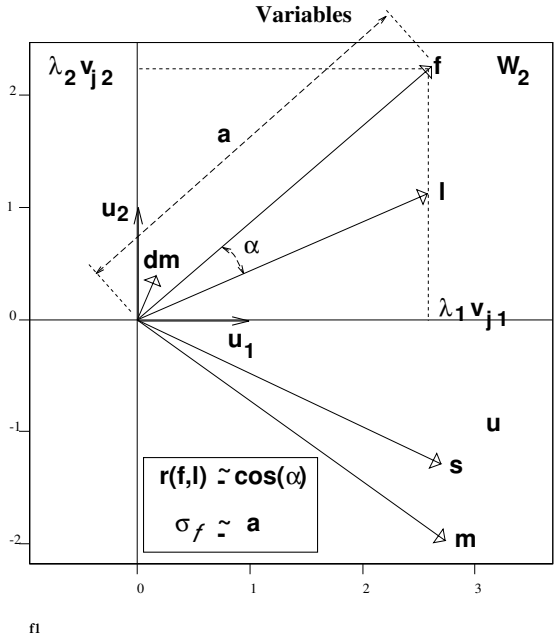
ACP non réduite

Les représentations graphiques

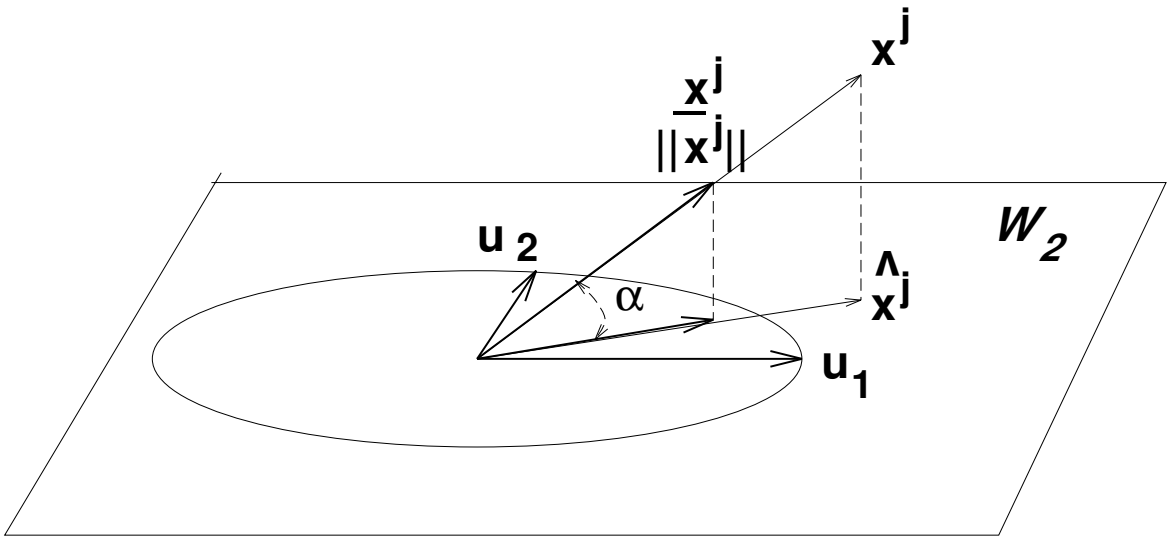
Représentation, la plus fidèle possible, du nuage centré des individus :

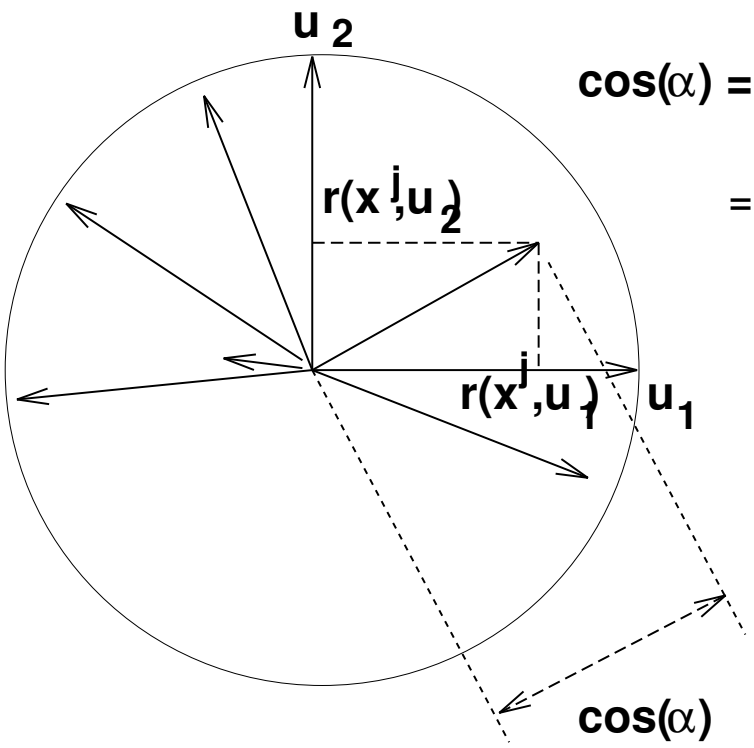


Représentation, la plus fidèle possible, du nuage des variables centrées :



Le cercle des corrélations



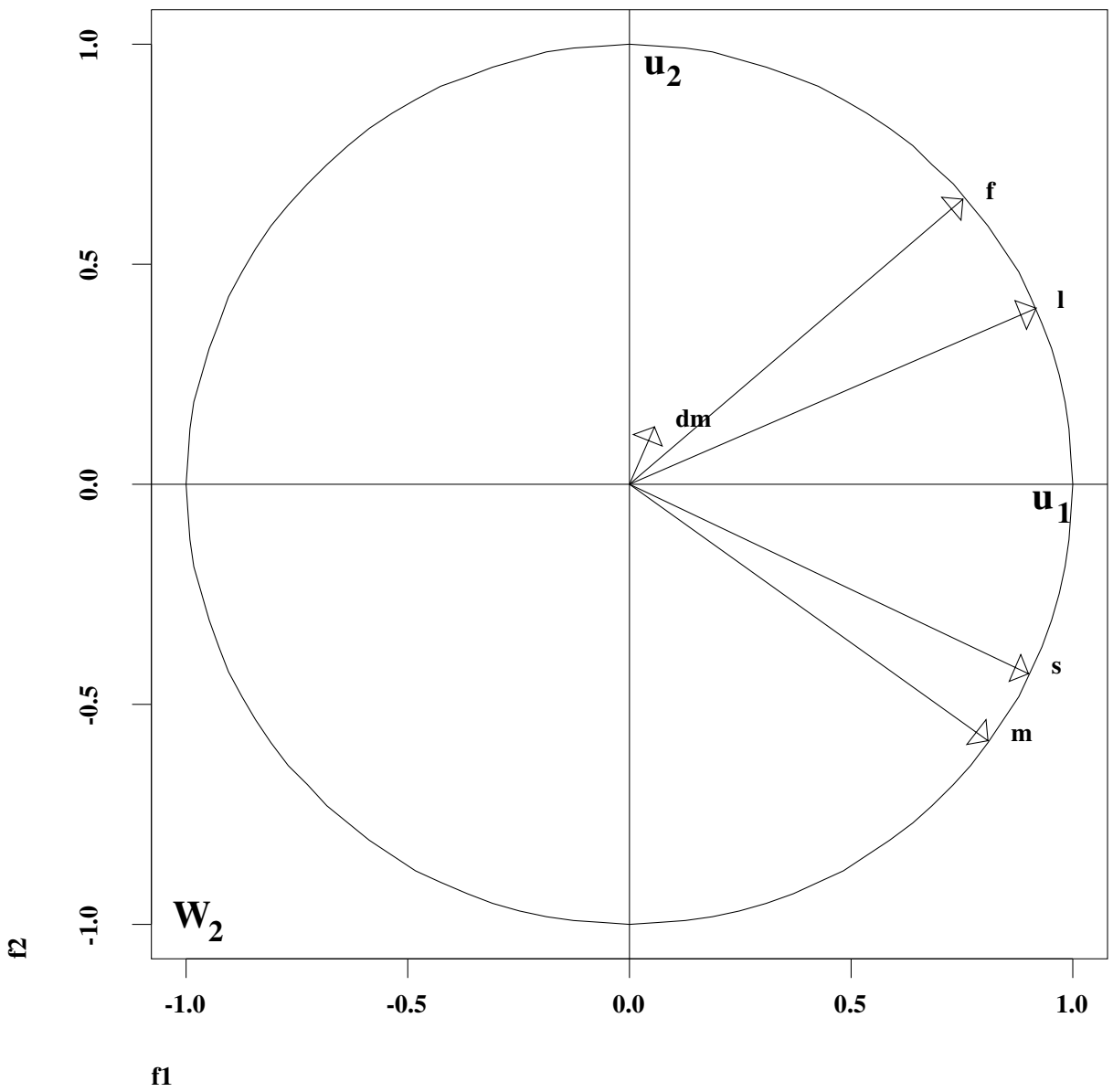


$$\cos(\alpha) = \frac{\|\hat{x}^j\|}{\|\overline{x}^j\|}$$

$$= r(x^j, (u_1, u_2))$$

Exemple

Cercle des correlations



Pratique de l'ACP : interprétation des facteurs

- Variables qui contribuent le plus au facteur.
- Fortes corrélations des variables au facteur.
- Fortes contributions des individus.
- Facteur **taille** : les corrélations de toutes les variables sont de même signe.
- Facteur **forme** : les variables avec des corrélations positives avec le facteur sont opposées à des variables avec des corrélations négatives avec le facteur.

Double décomposition de l'inertie (colonnes × facteurs)

Facteur	1	...	s	...	k	Sous-total	...	r	Total
Colonne									
1	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...
j	...	...	$m_j(\lambda_s v_s^j)^2$	...	...	$m_j\|\widehat{\mathbf{x}^j}^k\|_D^2$	...	...	$m_j\|\mathbf{x}^j\|_D^2$
...	...	...	...	...	...	...	...	...	...
J	...	...	...	...	...	...	...	...	...
Total	λ_1^2	...	λ_s^2	...	λ_k^2	$\|\widehat{X}^k\|^2$	...	λ_r^2	$\|X\|^2$

Table 1 : la décomposition "colonne × facteur" de l'inertie.

Double décomposition de l'inertie (lignes × facteurs)

Facteur	1	...	s	...	k	Sous-total	...	r	Total
Individu									
1	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...
i	...	...	$p_i(\lambda_s u_i^s)^2$	...	...	$p_i\|\widehat{x_i}^k\|_M^2$	...	...	$p_i\|x_i\|_M^2$
...	...	...	...	...	...	...	...	...	...
n	...	...	...	...	...	...	...	...	...
Total	λ_1^2	...	λ_s^2	...	λ_k^2	$\|\widehat{X}^k\|^2$	...	λ_r^2	$\|X\|^2$

Table 2 : la décomposition "individu × facteur" de l'inertie.

Pratique de l'ACP : individus et variables supplémentaires

- Représenter individus et variables non utilisés dans l'analyse.

  Exemple : individus mis à l'écart après une $1^{ère}$ analyse, point moyen d'un ensemble d'individus, variables illustratives, ...

Méthode :

- Individu supplémentaire (le vecteur associé a pour coordonnées les valeurs des p variables pour cet individu). On fait subir à ce vecteur les mêmes transformations que les vecteurs initiaux (centrage : $y^j_{sup} \rightarrow x^j_{sup} = y^j_{sup} - \bar{y^j}$, éventuellement réduction). Puis on projette la ligne transformée $\mathbf{x}_{sup}$ sur le sous-espace principal.

  Les coordonnées de $\mathbf{x}_{sup}$ sur $\mathbf{v}_s$ sont ${}^t\mathbf{v}_s M \mathbf{x}_{sup} = \sum_j m_j v^j_s x_{sup}$.

  ${}^t\mathbf{v}_s M$ est le projecteur orthogonal sur v_s.

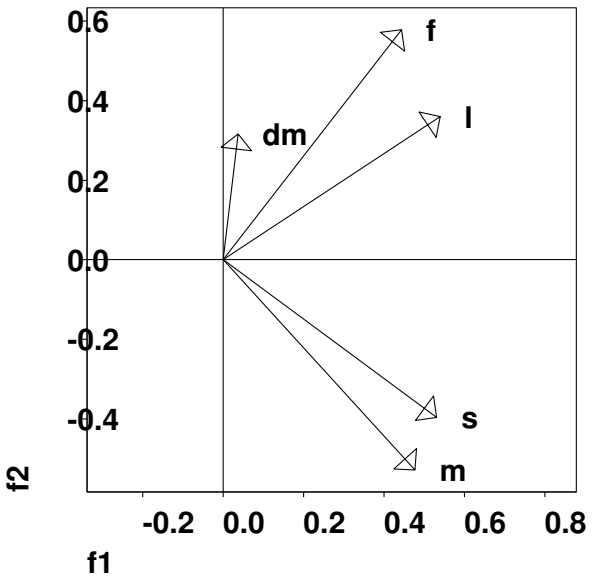
- Variable continue supplémentaire : même démarche que pour un vecteur individu : centrage et éventuellement réduction, puis projection sur le sous-espace principal W_k.

  La projection de $\mathbf{x}^{sup}$ sur W_k s'écrit $\sum_{s=1}^{k} ({}^t\mathbf{u}^s D \mathbf{x}^{sup})\mathbf{u}^s$.

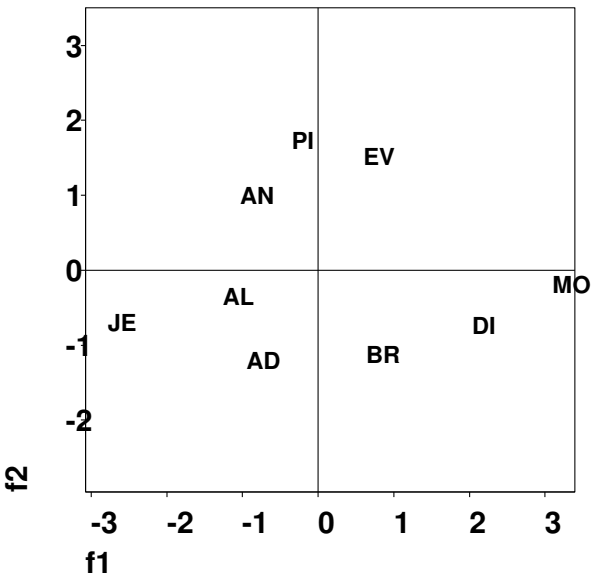
- Variable catégorielle illustrative : elle définit des sous-populations (ensemble d'individus ayant la même modalité). On représente en individu supplémentaire les barycentres de ces groupes, ou encore les contours des individus d'une même sous-population.

ACP **centrée réduite** sur le tableau "note" **complet**

Variables

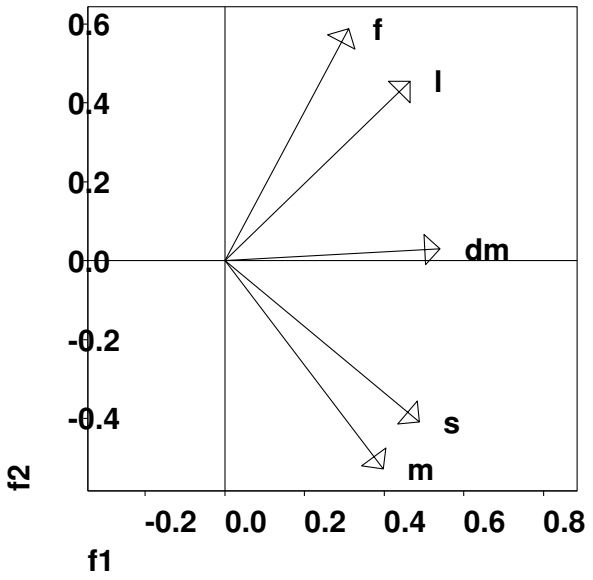


Individus

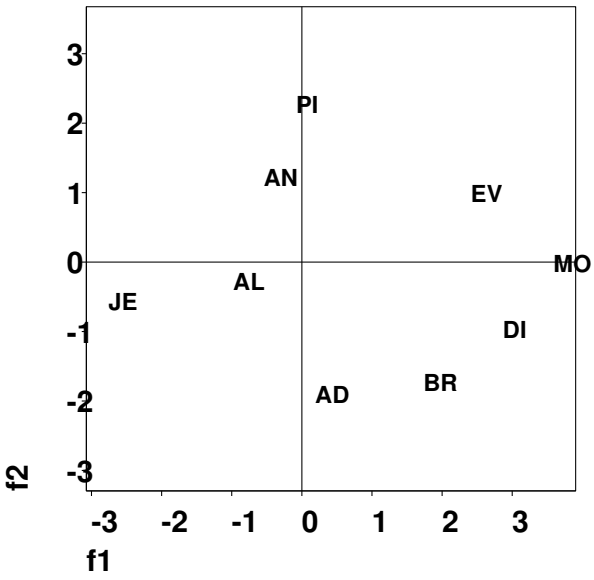


ACP centrée réduite sur le tableau "note" avec MO et DI individus supplémentaires

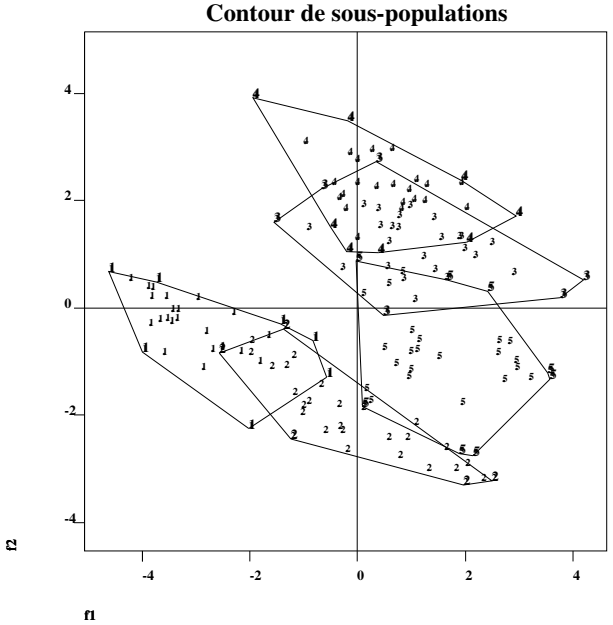
Variables



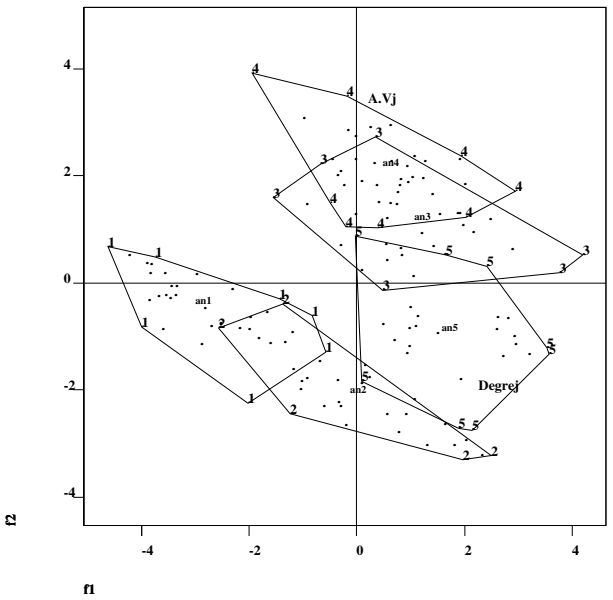
Individus



Contour de sous-populations



Contour et barycentres



Le BIPLOT d'une matrice de rang 2

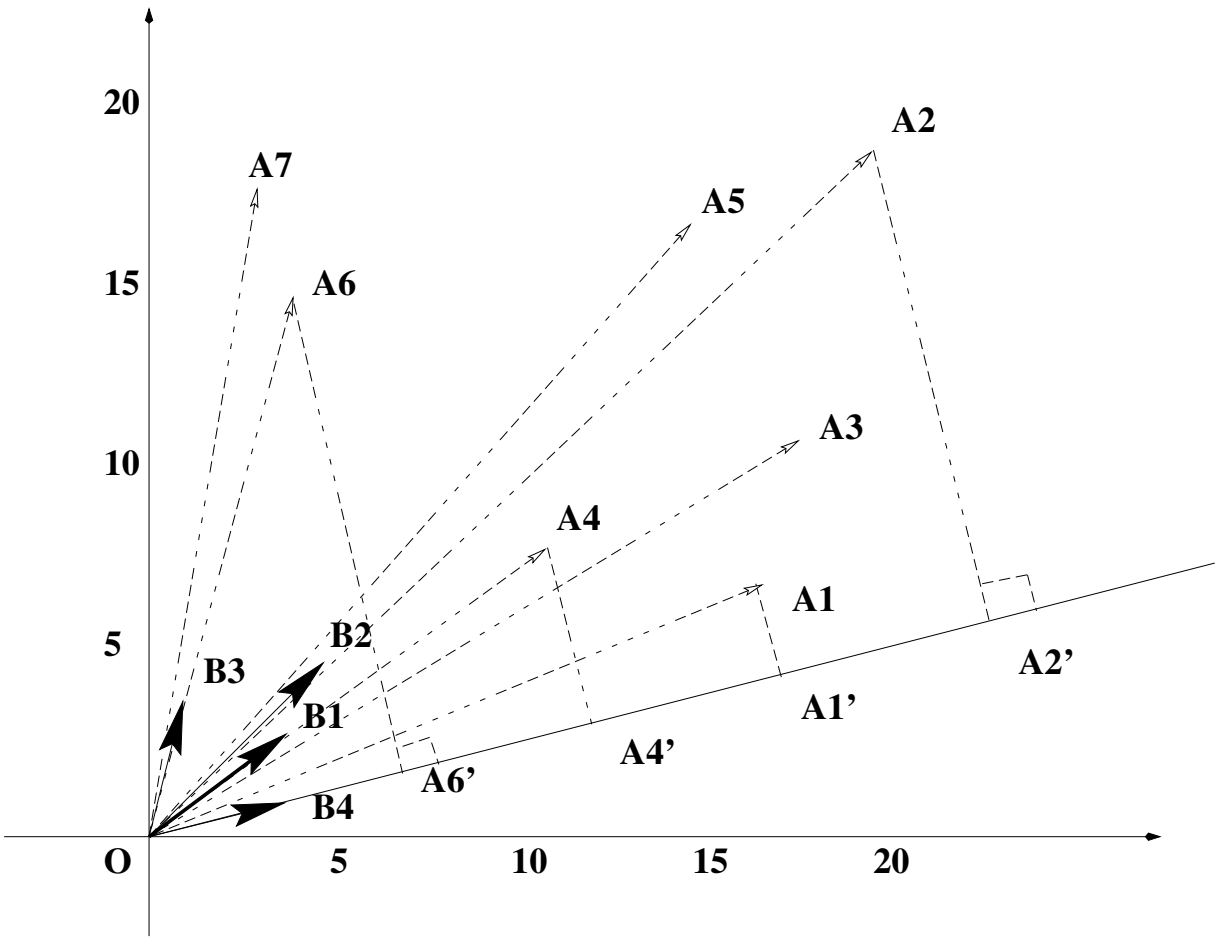
- Soit X de rang 2 telle que $X = \mathbf{u}^1 \, {}^t\mathbf{v}_1 + \mathbf{u}^2 \, {}^t\mathbf{v}_2 = U \, {}^tV$
- Par exemple :

$$X = \begin{pmatrix} 89 & 120 & 45 & 75 \\ 137 & 195 & 96 & 99 \\ 105 & 145 & 62 & 83 \\ 68 & 95 & 43 & 52 \\ 111 & 160 & 83 & 77 \\ 61 & 95 & 64 & 31 \\ 66 & 105 & 75 & 30 \\ 81 & 120 & 69 & 51 \\ 98 & 145 & 83 & 62 \\ 57 & 85 & 50 & 35 \end{pmatrix} = \begin{pmatrix} 17 & 7 \\ 20 & 19 \\ 18 & 11 \\ 11 & 8 \\ 15 & 17 \\ 4 & 15 \\ 3 & 18 \\ 9 & 15 \\ 11 & 18 \\ 6 & 11 \end{pmatrix} \begin{pmatrix} 4 & 5 & 1 & 4 \\ 3 & 5 & 4 & 1 \end{pmatrix}$$

$$x_3^2 = 145 = 18 \times 5 + 11 \times 5$$

Le BIPLOT d'une matrice de rang 2 (suite)

- A_i points ayant la ligne i de U pour coordonnées
- B_j points ayant la ligne j de V pour coordonnées



- si $A_i^{'}$ projection orthogonale de A_i sur l'axe OB_j

$$x_i^j =< OA_i, OB_j >= \|OA_i\|\|OB_j\| \cos\alpha$$
$$= \|OA_i^{'}\|\|OB_j\| signe(\cos\alpha)$$

Le BIPLOT : application à l'ACP

- soit l'approximation de rang 2 du tableau X
- 3 écritures de la décomposition :

$$\hat{x}_i^j{}^2 = \sum_{s=1}^{2} (\lambda_s u_s^i)(v_j^s) \qquad (1)$$

$$= \sum_{s=1}^{2} (u_s^i)(\lambda_s v_j^s) \qquad (2)$$

$$= \sum_{s=1}^{2} (\sqrt{\lambda_s} u_s^i)(\sqrt{\lambda_s} v_j^s) \quad (3)$$

(1) $A_i = \begin{pmatrix} \lambda_1 u_1^i \\ \lambda_2 u_2^i \end{pmatrix}, B_j = \begin{pmatrix} v_j^1 \\ v_j^2 \end{pmatrix}$, biplot **isométrique lignes**

(2) $A_i = \begin{pmatrix} u_1^i \\ u_2^i \end{pmatrix}, B_j = \begin{pmatrix} \lambda_1 v_j^1 \\ \lambda_2 v_j^2 \end{pmatrix}$, biplot **isométrique colonnes**

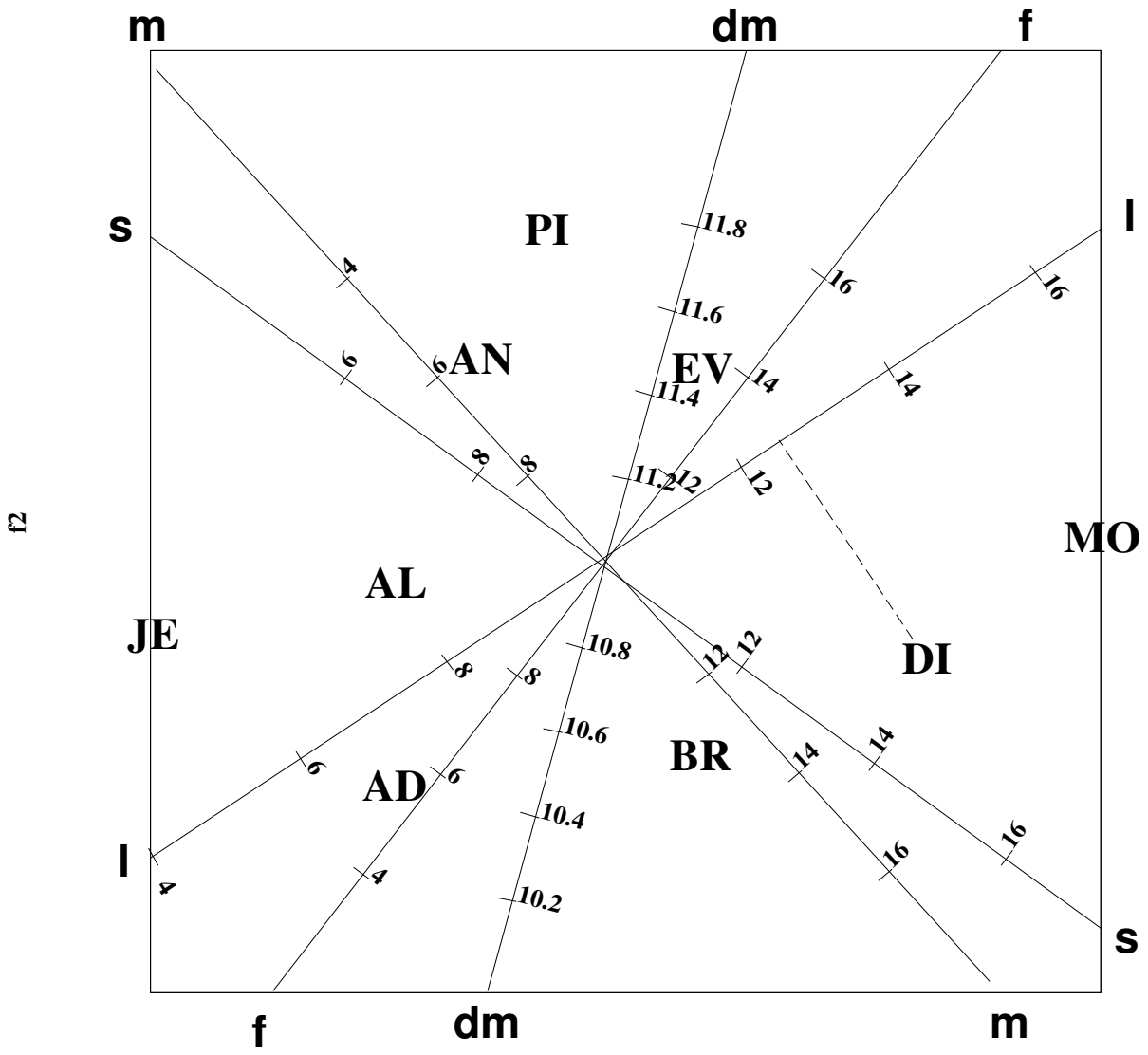
(3) $A_i = \begin{pmatrix} \sqrt{\lambda_1} u_1^i \\ \sqrt{\lambda_2} u_2^i \end{pmatrix}, B_j = \begin{pmatrix} \sqrt{\lambda_1} v_j^1 \\ \sqrt{\lambda_2} v_j^2 \end{pmatrix}$, biplot **symétrique**

Exemple

Notes centrées

nom	m	s	f	l	dm
JE	-3.7	-3.8	-5.2	-4.6	-3.0
AL	-1.7	-1.8	-2.2	-2.1	-2.0
AN	-3.7	-2.8	0.8	-0.6	0.0
MO	4.8	4.7	5.3	4.9	-3.0
DI	4.3	4.2	1.8	2.4	-1.0
AD	1.3	0.2	-4.7	-3.1	2.0
PI	-4.2	-2.8	3.8	1.4	-1.0
BR	3.3	2.7	-1.7	-0.6	1.0
EV	-0.7	-0.3	2.3	1.9	7.0

Approximation de rang 2

nom	m	s	f	l	dm
JE	-3.5	-3.8	-5.4	-4.8	-0.5
AL	-1.6	-1.7	-2.4	-2.1	-0.2
AN	-3.6	-2.9	0.7	-0.4	0.3
MO	4.9	4.9	4.9	4.8	0.3
DI	4.5	4.0	1.7	2.4	-0.1
AD	1.1	0.2	-4.5	-2.9	-0.6
PI	-4.1	-2.8	3.5	1.5	0.7
BR	3.3	2.5	-1.7	-0.4	-0.4
EV	-1.0	-0.4	3.1	2.0	0.4

Biplot gradué isométrique ligne