Analyse de données
Analyse en composantes principales

Farida Zehraoui
Email: zehraoui@ibisc.univ-evry.fr

Rappels – Notions élémentaires

Moyenne

Définition On note

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

ou pour des données pondérés

$$\bar{x} = \sum_{i=1}^{n} p_i x_i.$$

Propriétés la moyenne arithmétique est une mesure de *tendance centrale* qui dépend de toutes les observations et est sensible aux valeurs extrêmes. Elle est très utilisée à cause de ses bonnes propriétés mathématiques.

Variance / Ecart type

Définition la *variance* de $\mathbf{x}$ est définie par

$$s_{\mathbf{x}}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \text{ ou } s_{\mathbf{x}}^2 = \sum_{i=1}^{n} p_i(x_i - \bar{x})^2$$

L'*écart-type* $s_{\mathbf{x}}$ est la racine carrée de la variance.

Propriétés La variance satisfait la formule suivante

$$s_{\mathbf{x}}^2 = \sum_{i=1}^{n} p_i x_i^2 - (\bar{x})^2$$

La variance est « la moyenne des carrés moins le carré de la moyenne ».
L'écart-type, qui a la même unité que $\mathbf{x}$, est une mesure de *dispersion*.

Mesure de liaison entre deux variables

Définitions la covariance observée entre deux variables x et y est

$$s_{xy} = \sum_{i=1}^{n} p_i(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} p_i x_i y_i - \bar{x}\bar{y}.$$

et le *coefficient de r de Bravais-Pearson* ou coefficient de corrélation est donné par

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^{n} p_i(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i-1}^{n} p_i(x_i - \bar{x})^2}\sqrt{\sum_{i-1}^{n} p_i(y_i - \bar{y})^2}}.$$

Propriétés du coefficient de corrélation

Borne On a toujours (inégalité de Cauchy-Schwarz)

$$-1 \leq r_{xy} \leq 1.$$

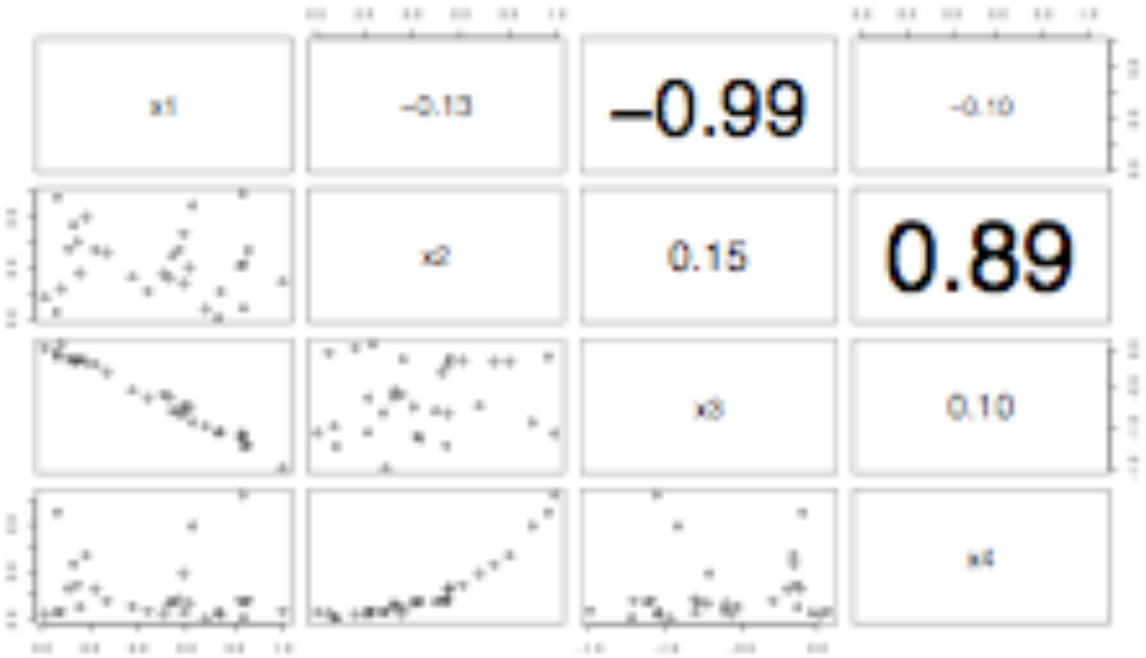
Variables liées $|r_{xy}| = 1$ si et seulement si $\mathbf{x}$ et $\mathbf{y}$ sont linéairement liées :

$$ax_i + by_i = c, \text{ pour tout } 1 \leq i \leq n.$$

En particulier, $r_{xx} = 1$.

Variables décorrélées si $r_{xy} = 0$, on dit que les variables sont *décorrélées*. Cela ne veut pas dire qu'elles sont indépendantes !

Exemple : coefficient de corrélation



Interprétation on a 4 variables numériques avec 30 individus. Les variables 1 et 2 sont indépendantes ; les variables 1 et 3 ont une relation linéaire ; les variables 2 et 4 ont une relation non-linéaire.

Tableau de données

Tableau des données : matrice X de dimension n.p

On a n individus décrits par les valeurs respectives de p variables ou attributs

$$X = \begin{bmatrix} x_{11} \ldots\ldots & x_{1p} \\ x_{21} \ldots\ldots & x_{2p} \\ \ldots\ldots\ldots & \\ & \\ x_{n1} \ldots\ldots & x_{np} \end{bmatrix}$$

x_{ij} est un réel ou une valeur qualitative

Vecteur variable et individu

Variable Une colonne du tableau

$$\mathbf{x}^j = \begin{bmatrix} x_1^j \\ \vdots \\ x_i^j \\ \vdots \\ x_n^j \end{bmatrix}$$

Individu Une ligne du tableau

$$\mathbf{e}_i' = (x_i^1, \ldots, x_i^j, \ldots, x_i^p)$$

La matrice des poids

Définition on associe aux individus un poids p_i tel que

$$p_1 + \cdots + p_n = 1$$

et on représente ces poids dans la matrice diagonale de taille n

$$\mathbf{D} = \begin{bmatrix} p_1 & & & 0 \\ & p_2 & & \\ & & \ddots & \\ 0 & & & p_n \end{bmatrix}.$$

Cas uniforme tous les individus ont le même poids $p_i = 1/n$ et $\mathbf{D} = \frac{1}{n}\mathbf{I}_n$.

Point moyen et tableau centré

Point moyen c'est le vecteur $\mathbf{g}$ des moyennes arithmétiques de chaque variable :

$$\mathbf{g}' = (\bar{x}^1, \ldots, \bar{x}^p),$$

où

$$\bar{x}^j = \sum_{i=1}^{n} p_i x_i^j.$$

On peut aussi écrire $\mathbf{g} = \mathbf{X}'\mathbf{D}\mathbf{1}_n$.

Tableau centré il est obtenu en centrant les variables autour de leur moyenne

$$y_i^j = x_i^j - \bar{x}^j$$

ou, en notation matricielle,

$$\mathbf{Y} = \mathbf{X} - \mathbf{1}_n\mathbf{g}' = (\mathbf{I} - \mathbf{1}_n\mathbf{1}_n'\mathbf{D})\mathbf{X}$$

Matrice de variance-covariance

Définition c'est une matrice *carrée* de dimension p

$$\mathbf{V} = \begin{bmatrix} s_1^2 & s_{12} & \cdots & s_{1p} \\ s_{21} & & & \\ \vdots & & \ddots & \\ s_{p1} & & & s_p^2 \end{bmatrix},$$

où s_{kl} est la covariance des variables $\mathbf{x}^k$ et $\mathbf{x}^\ell$ et s_j^2 est la variance de la variable $\mathbf{x}^j$

Formule matricielle

$$\mathbf{V} = \mathbf{X}'\mathbf{D}\mathbf{X} - \mathbf{g}\mathbf{g}' = \mathbf{Y}'\mathbf{D}\mathbf{Y}.$$

Matrice de corrélations

Définition Si l'on note $r_{k\ell} = s_{k\ell}/s_k s_\ell$, c'est la matrice $p \times p$

$$\mathbf{R} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & & \\ \vdots & & \ddots & \\ r_{p1} & & & 1 \end{bmatrix},$$

Formule matricielle $\mathbf{R} = \mathbf{D}_{1/s}\mathbf{V}\mathbf{D}_{1/s}$, où

$$\mathbf{D}_{1/s} = \begin{bmatrix} \frac{1}{s_1} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{s_p} \end{bmatrix}$$

Vecteurs propres / valeurs propres

Définition un vecteur $\mathbf{v} \neq \mathbf{0}$ de taille p est un *vecteur propre* d'une matrice $\mathbf{A}$ de taille $p \times p$ s'il existe $\lambda \in \mathbb{C}$ telle que

$$\mathbf{A}\mathbf{v} = \lambda \mathbf{v}.$$

λ est une *valeur propre* de $\mathbf{A}$ associée à $\mathbf{v}$.

Domaine En général, les vecteurs propres et valeurs propres sont complexes ; dans tous les cas qui nous intéressent, ils seront réels.

Interprétation des vecteurs propres ce sont les directions dans lesquelles la matrice agit.

Interprétation des valeurs propres c'est le facteur multiplicatif associé à une direction donnée.

Exemple : valeurs et vecteurs propres

- Soit la matrice :

$$\begin{pmatrix} 5 & 1 & -1 \\ 2 & 4 & -2 \\ 1 & -1 & 3 \end{pmatrix}$$

- Calculer les vecteurs et valeurs propres de cette matrice

a pour vecteurs propres

$$\mathbf{v}_1 = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}, \ \mathbf{v}_2 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \ \mathbf{v}_3 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}.$$

On vérifie facilement que les valeurs propres associées sont

$$\lambda_1 = 2, \ \lambda_2 = 4, \ \lambda_3 = 6.$$

Méthodes factorielles

Les méthodes factorielles visent à fournir des représentations synthétiques des tableaux de données multidimensionnelles en les représentant dans des espaces euclidiens de dimension faible (en général, plans)

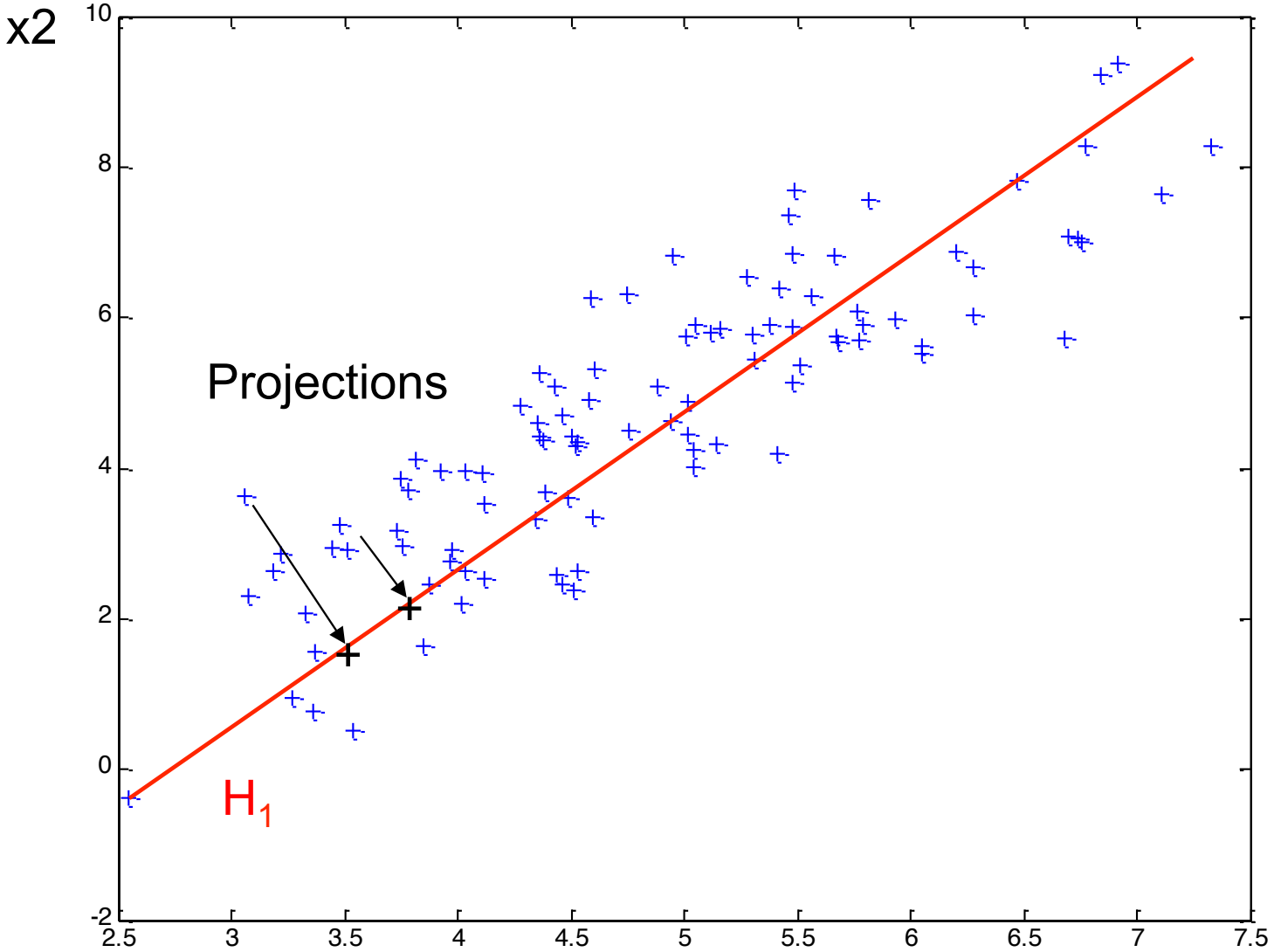
Analyse en composantes principales (ACP)

Analyse en composantes principales

Technique de réduction de dimension d'un tableau de données multidimensionnelles numériques

Objectif : projeter le nuage de points (définis par le tableau de données) sur une droite (plan) telle que la ***dispersion totale*** des points projetés soit maximale

Exemple : données en 2D



2. Les objectifs de l'analyse en composantes principales

- Résumer un tableau individus×variables à l'aide d'un petit nombre de facteurs.
- Visualiser le positionnement des individus les uns par rapport aux autres.
- Visualiser les corrélations entre les variables.
- Interpréter les facteurs.

ACP (1) : Interprétation géométrique

Les lignes et les colonnes du tableau X sont représentées par des points dans deux espaces différents : l'espace des variables et l'espace des individus.

- **Espace des individus**

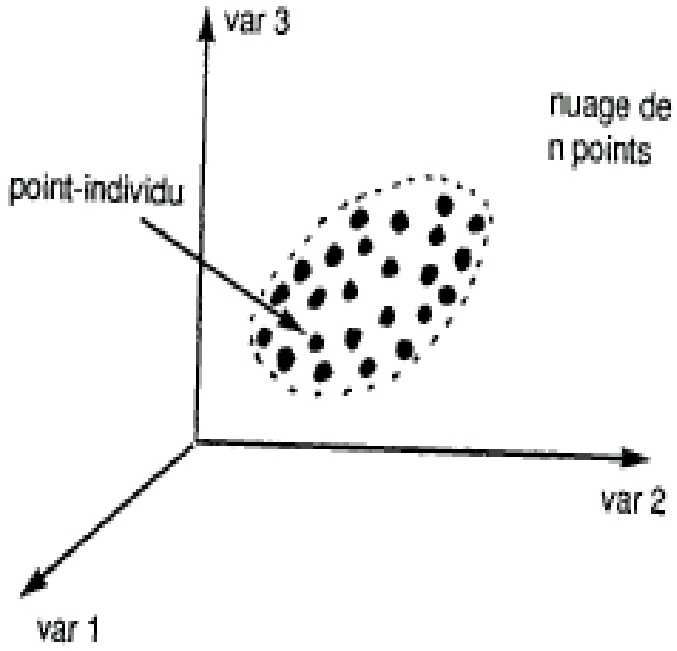
  Les n lignes peuvent être considérées comme n points de l'espace des individus à p dimensions.

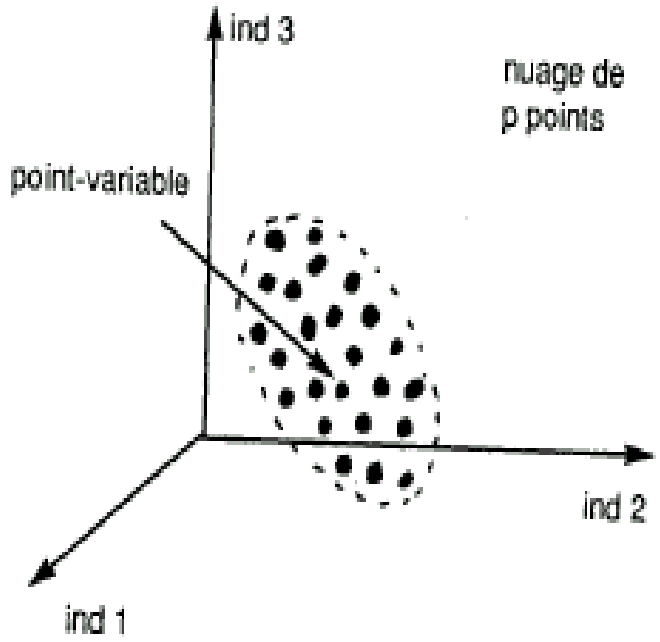
- **Espace des variables**

  Les p colonnes peuvent être considérées comme p points dans un espace à n dimensions. Cet espace est appelé l'espace des variables.

ACP (1) : Interprétation géométrique

Les lignes et les colonnes du tableau X sont représentées par des points dans deux espaces différents : l'espace des variables et l'espace des individus.





Exemple

Le tableau représente des notes obtenues par des élèves dans diverses matières. Nous notons qu'il contient 45 valeurs numériques (n = 9, p = 5).

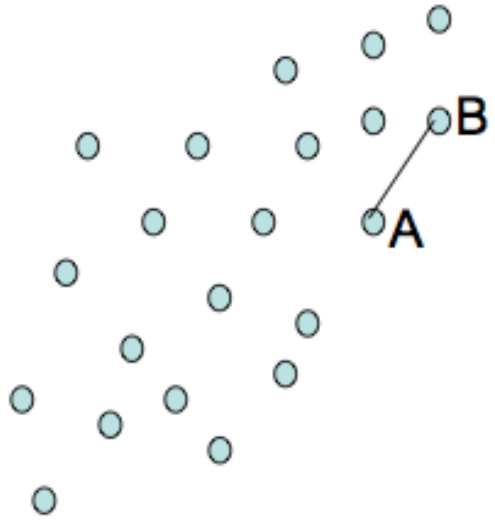
Sujet	Math	Sciences	Français	Latin	Musique
Jean	6	6	5	5,5	8
Aline	8	8	8	8	9
Annie	6	7	11	9,5	11
Monique	14,5	14,5	15,5	15	8
Didier	14	14	12	12	10
André	11	10	5,5	7	13
Pierre	5,5	7	14	11,5	10
Brigitte	13	12,5	8,5	9,5	12
Evelyne	9	9,5	12,5	12	18

ACP : suite

Espace des individus

L'espace des individus est muni de la distance euclidienne classique.

$$d^2(X_{A.}, X_{B.}) = \sum_{j=1}^{p} \left(X_A^j - X_B^j\right)^2$$



ACP : suite

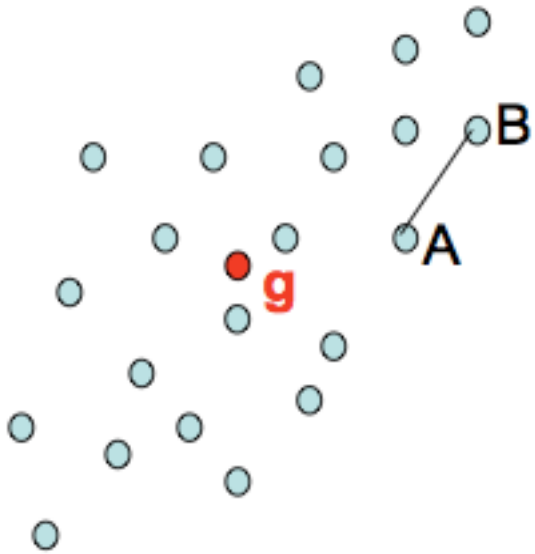
Espace des individus

L'espace des individus est muni de la distance euclidienne classique.

$$d^2(X_{A.}, X_{B.}) = \sum_{j=1}^{p} \left(X_A^j - X_B^j \right)^2$$

Soit *g* le centre de gravité (moyenne) du nuage de points :

$$g = \left(\frac{1}{n} \right) \sum_{i=1}^{n} X_i$$



ACP : suite

Espace des individus

L'espace des individus est muni de la distance euclidienne classique.

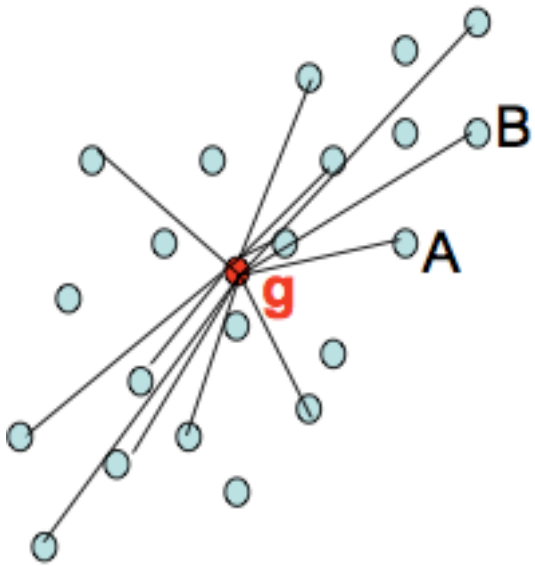
$$d^2(X_{A.}, X_{B.}) = \sum_{j=1}^{p} \left(X_A^j - X_B^j \right)^2$$

Soit *g* le centre de gravité (moyenne) du nuage de points :

$$g = \left(\frac{1}{n} \right) \sum_{i=1}^{n} X_i$$

La variance appelée « inertie totale » du nuage :

$$I_g = \sum_{i=1}^{n} \frac{1}{n} d^2(g, X_i)$$

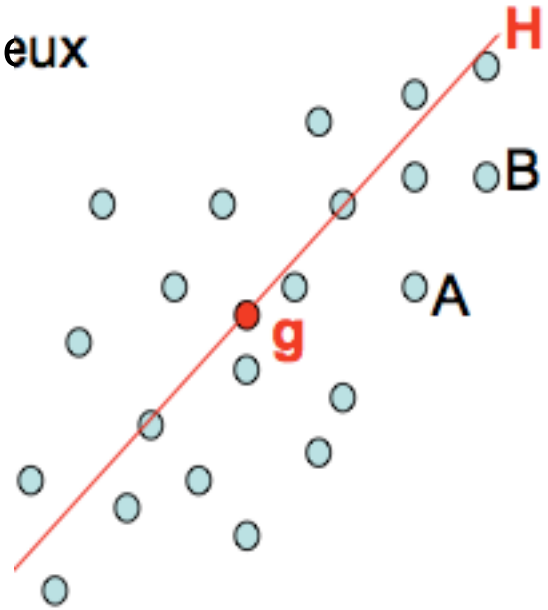


ACP : suite

Espace des individus

On recherche un sous-espace représentant au mieux ce nuage de points en respectant la dispersion des points. Ce sous espace par le centre de gravité g.

Soit H le sous-espace passant par g, on distingue deux types d'inertie :



ACP : suite

Espace des individus

On recherche un sous-espace représentant au mieux ce nuage de points en respectant la dispersion des points. Ce sous espace par le centre de gravité g.

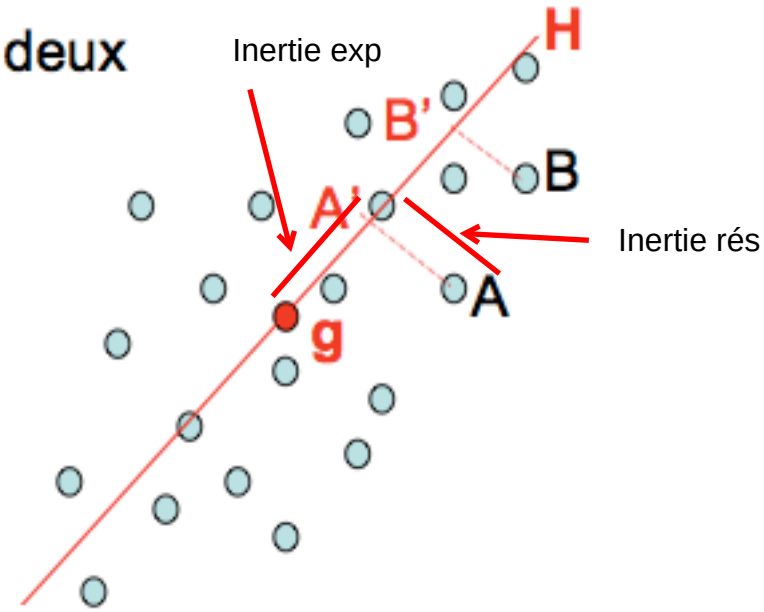
Soit H le sous-espace passant par g, on distingue deux types d'inertie :

Inertie expliquée par H :

$$I_{\text{exp}}(H) = \sum_{i=1}^{n} \frac{1}{n} d^2\left(g, \hat{X}_i\right)$$

Inertie résiduelle autour de H :

$$I_{rés}(H) = \sum_{i=1}^{n} \frac{1}{n} d^2\left(\hat{X}_i, X_i\right)$$



ACP : suite

Espace des individus

On recherche des sous-espaces représentant au mieux ce nuage de points en respectant la dispersion des points. Ces sous espaces passent par le centre de gravité g.

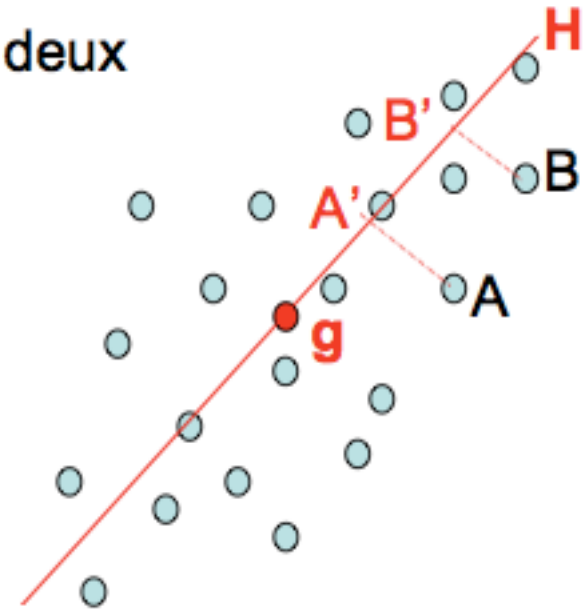
Soit H le sous-espace passant par g, on distingue deux types d'inertie :

Inertie expliquée par H :

$$I_{\text{exp}}(H) = \sum_{i=1}^{n} \frac{1}{n} d^2\left(g, \hat{X}_i\right)$$

Inertie résiduelle autour de H :

$$I_{rés}(H) = \sum_{i=1}^{n} \frac{1}{n} d^2\left(\hat{X}_i, X_i\right)$$



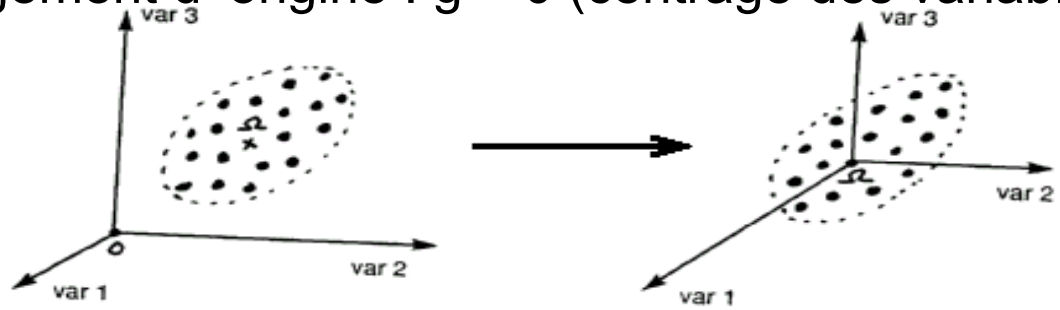
Inertie totale = inertie expliquée + inertie résiduelle

Pour choisir H : maximiser I(exp) ou minimiser I(rés)

ACP : suite

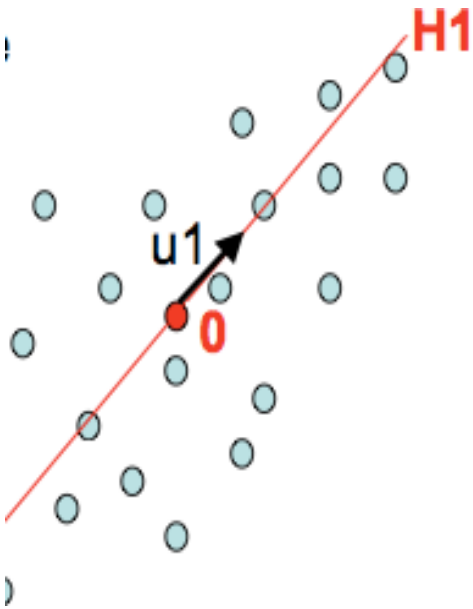
Espace des variables

Changement d'origine : g = 0 (centrage des variables)



La recherche des sous-espaces Hk se fait de k=1 à p :

- La détermination de H1 revient à chercher une droite passant par l'origine qui maximise l'inertie expliquée.
- Pour trouver cette droite, il faut déterminer un vecteur u1 porté par cette droite avec norme(u1)=1.

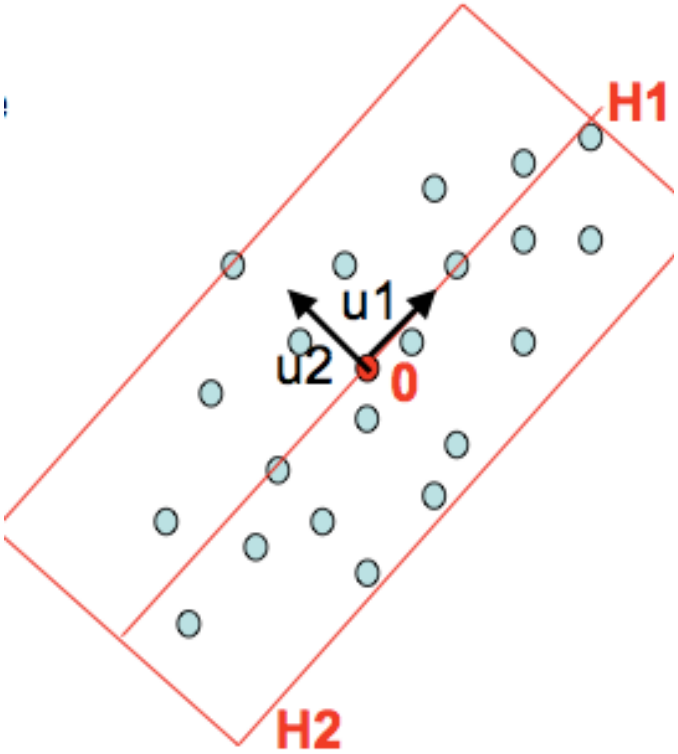


ACP : suite

Changement d'origine : g = 0 (centrage des variables)

La recherche des sous-espaces Hk se fait de k=1 à p :

- La détermination de H1 revient à chercher une droite passant par l'origine qui maximise l'inertie expliquée. Pour trouver cette droite, il faut déterminer un vecteur u1 porté par cette droite avec norme(u1)=1.



- Afin de déterminer le sous-espace H2, on recherche un vecteur u2 perpendiculaire à u1 et tel que la droite portée par u2, passant par 0, ait une inertie maximale.

- On peut démontrer que le sous-espace H3 contient nécessairement u1 et u2. etc.

ACP : suite

Espace des variables

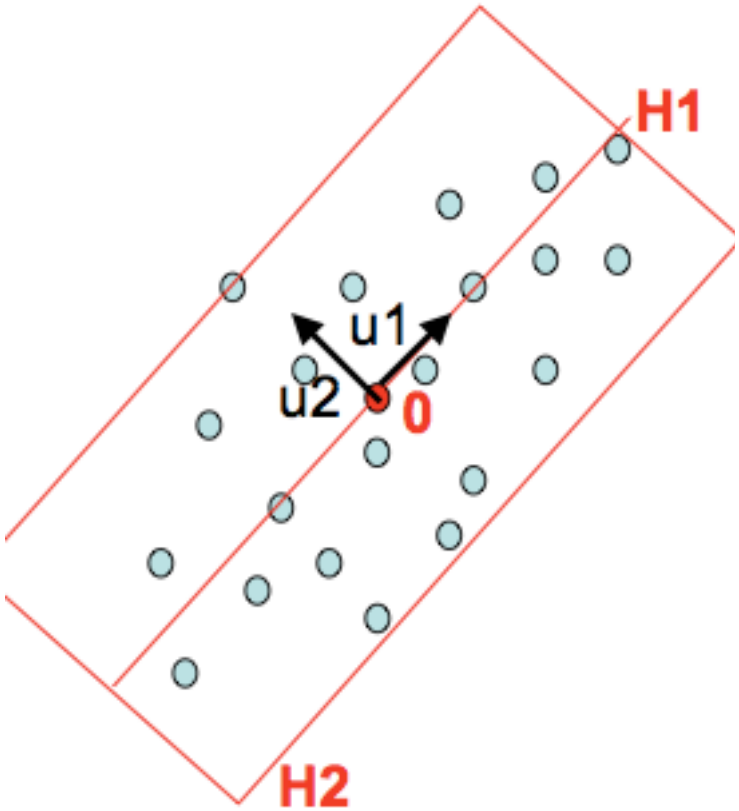
Les vecteurs u1,u2,...,up peuvent s'obtenir à partir de la matrice d'inertie C (matrice de covariance ou corrélation) entre les variables du tableau.

Il existe p vecteurs et p constantes λ qui vérifient l'équation matricielle suivante :

$C.v = \lambda.v$

Les p vecteurs v sont les vecteurs propres de C et les constantes associées sont les valeurs propres de C.

Ces vecteurs sont orthogonaux deux à deux et unitaires (norme = 1). Ils peuvent être rangés par ordre décroissant des valeurs propres associées : le premier vecteur propre v1 est associé à la valeur propre la plus élevé λ1. Ces vecteurs sont les vecteurs u1 à up recherchés (solution du problème de maximisation de l'inertie expliquée).



Exemple (suite)

$Y =$

-1,09	-1,28	-1,50	-1,63	-1,02
-0,49	-0,61	-0,64	-0,72	-0,68
-1,09	-0,95	0,22	-0,18	0,00
1,43	1,56	1,52	1,81	-1,02
1,28	1,39	0,51	0,72	-0,34
0,40	0,06	-1,36	-1,08	0,68
-1,23	-0,95	1,09	0,54	-0,34
0,99	0,89	-0,50	-0,18	0,34
-0,20	-0,11	0,66	0,72	2,38

Calculons la matrice des corrélations :

$c = Y^T Y = (1/9)$

-1,09	-0,49	-1,09	1,43	1,28	0,40	-1,23	0,99	-0,20
-1,28	-0,61	-0,95	1,56	1,39	0,06	-0,95	0,89	-0,11
-1,50	-0,64	0,22	1,52	0,51	-1,36	1,09	-0,50	0,66
-1,63	-0,72	-0,18	1,81	0,72	-1,08	0,54	-0,18	0,72
-1,02	-0,68	0,00	-1,02	-0,34	0,68	-0,34	0,34	2,38

-1,09	-1,28	-1,50	-1,63	-1,02
-0,49	-0,61	-0,64	-0,72	-0,68
-1,09	-0,95	0,22	-0,18	0,00
1,43	1,56	1,52	1,81	-1,02
1,28	1,39	0,51	0,72	-0,34
0,40	0,06	-1,36	-1,08	0,68
-1,23	-0,95	1,09	0,54	-0,34
0,99	0,89	-0,50	-0,18	0,34
-0,20	-0,11	0,66	0,72	2,38

=

1,00	0,98	0,23	0,49	0,01
0,98	1,00	0,40	0,63	0,01
0,23	0,40	1,00	0,96	0,04
0,49	0,63	0,96	1,00	0,09
0,01	0,01	0,04	0,09	1,00

Exemple (suite)

Les valeurs propres de la matrice de correlations :

$\lambda_1 = 2{,}86$
$\lambda_2 = 1{,}15$
$\lambda_3 = 0{,}98$
$\lambda_4 = 0{,}01$
$\lambda_5 = 0{,}00$

Les vecteurs propres associés :

$$u_1 = \begin{bmatrix} 0.4763860 \\ 0.5302490 \\ 0.4481194 \\ 0.5380781 \\ 0.0394130 \end{bmatrix} \quad u_2 = \begin{bmatrix} -0.5326614 \\ -0.4015910 \\ 0.5696672 \\ 0.3705534 \\ 0.3052311 \end{bmatrix} \quad u_3 = \begin{bmatrix} -0.1547149 \\ -0.0935850 \\ 0.2275909 \\ 0.1092850 \\ -0.9505600 \end{bmatrix} \quad u_4 = \begin{bmatrix} -0.3040957 \\ 0.5177931 \\ 0.4766606 \\ -0.6408529 \\ 0.0389649 \end{bmatrix} \quad u_5 = \begin{bmatrix} 0.6106696 \\ -0.5298045 \\ 0.4423408 \\ -0.3879775 \\ 0.0140703 \end{bmatrix}$$

ACP : suite

L'espace des variables

Les droites engendrées par ces vecteurs propres sont appelées respectivement le 1er, 2ème, et pième axe principal d'inertie du nuage.

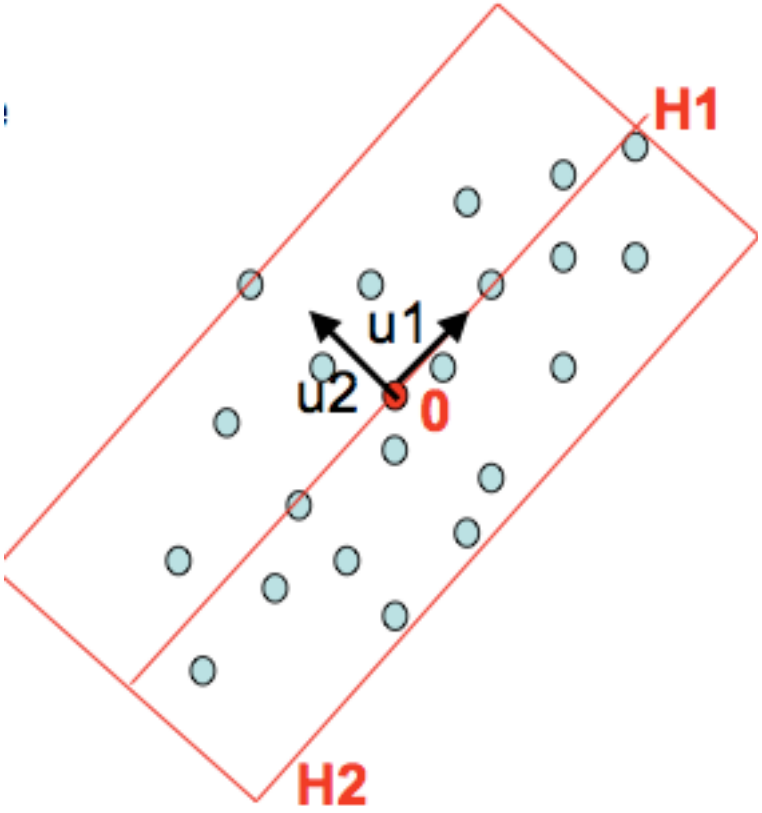
L'inertie expliquée par H1, le premier axe principal engendré par v1 est égale à :

I(H1)= λ1

L'inertie expliqué par H2, le plan engendré par v1 et v2 est égale à :

I(H2)= λ1+ λ2

Les valeurs propres de C représentent donc les parts d'inertie expliquée par chacun des axes principaux du nuage des individus.



Interprétation des résultats de l'ACP

On appelle *axe factoriel* ou *composante principale* $\mathbf{v_j}$, le j-ième vecteur propre de C, de norme 1 , associée à λ_j.

Soit Z la matrice des données centrées et réduites.
Les coordonnées des n points (individus) sur l'axe factoriel v_j sont les n composantes du vecteur, appelé facteur:

$$\mathbf{c_j} = Z\mathbf{v_j}$$

Le point i a pour coordonnée sur l'axe $\mathbf{u_j}$:

$$c_{ji} = \Sigma\, v_{jk}\, z_{ik}$$

Exemple (suite)

		Y		
-1,09	-1,28	-1,50	-1,63	-1,02
-0,49	-0,61	-0,64	-0,72	-0,68
-1,09	-0,95	0,22	-0,18	0,00
1,43	1,56	1,52	1,81	-1,02
1,28	1,39	0,51	0,72	-0,34
0,40	0,06	-1,36	-1,08	0,68
-1,23	-0,95	1,09	0,54	-0,34
0,99	0,89	-0,50	-0,18	0,34
-0,20	-0,11	0,66	0,72	2,38

u_1	u_2
0,48	-0,53
0,53	-0,40
0,45	0,56
0,54	0,37
0,04	0,31

=

axe1	axe2	
-2,79	-0,67	
-1,26	-0,33	
-1,01	1,02	← individu
3,12	-0,18	
1,95	-0,79	
-0,95	-1,19	
-0,32	1,75	
0,63	-1,13	
0,63	1,53	

Evaluer l'ACP

On a : $\lambda_j = v_j^T C v_j$

$\sum_{j=1}^{h} \lambda_j$ est *l'inertie totale* ou *dispersion totale* liée au sous–espace engendré par les h premières composantes principales

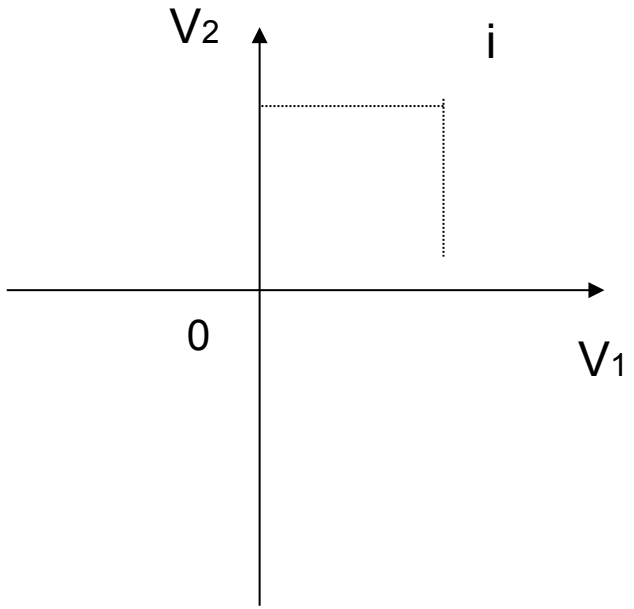
$$\frac{\sum_{j=1}^{h} \lambda_j}{\sum_{i=1}^{p} \lambda_i}$$

est le taux d'inertie expliqué par le sous-espace H engendré par les h premières composantes principales.

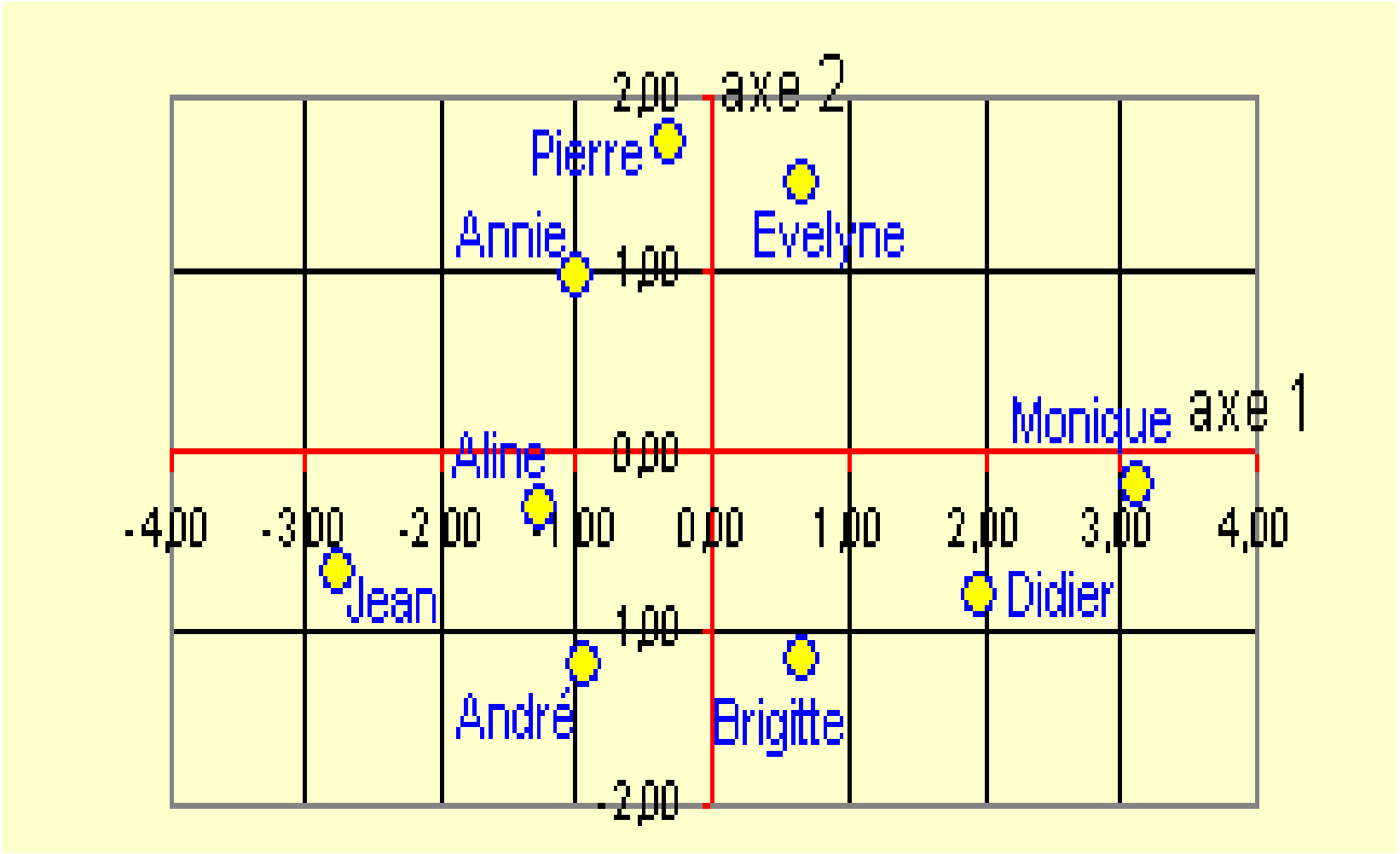
Représentation des individus dans un plan principal

Qu'est-ce que c' est ?

Chaque individu *i* est représenté par un point dans un plan représenté par les deux composantes principales V_1 et V_2



Exemple (suite)

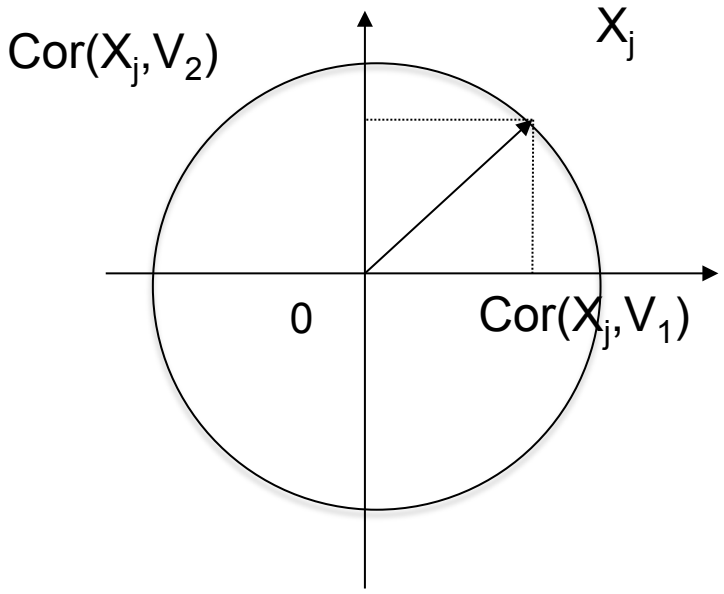


Représentation des variables dans un plan principal

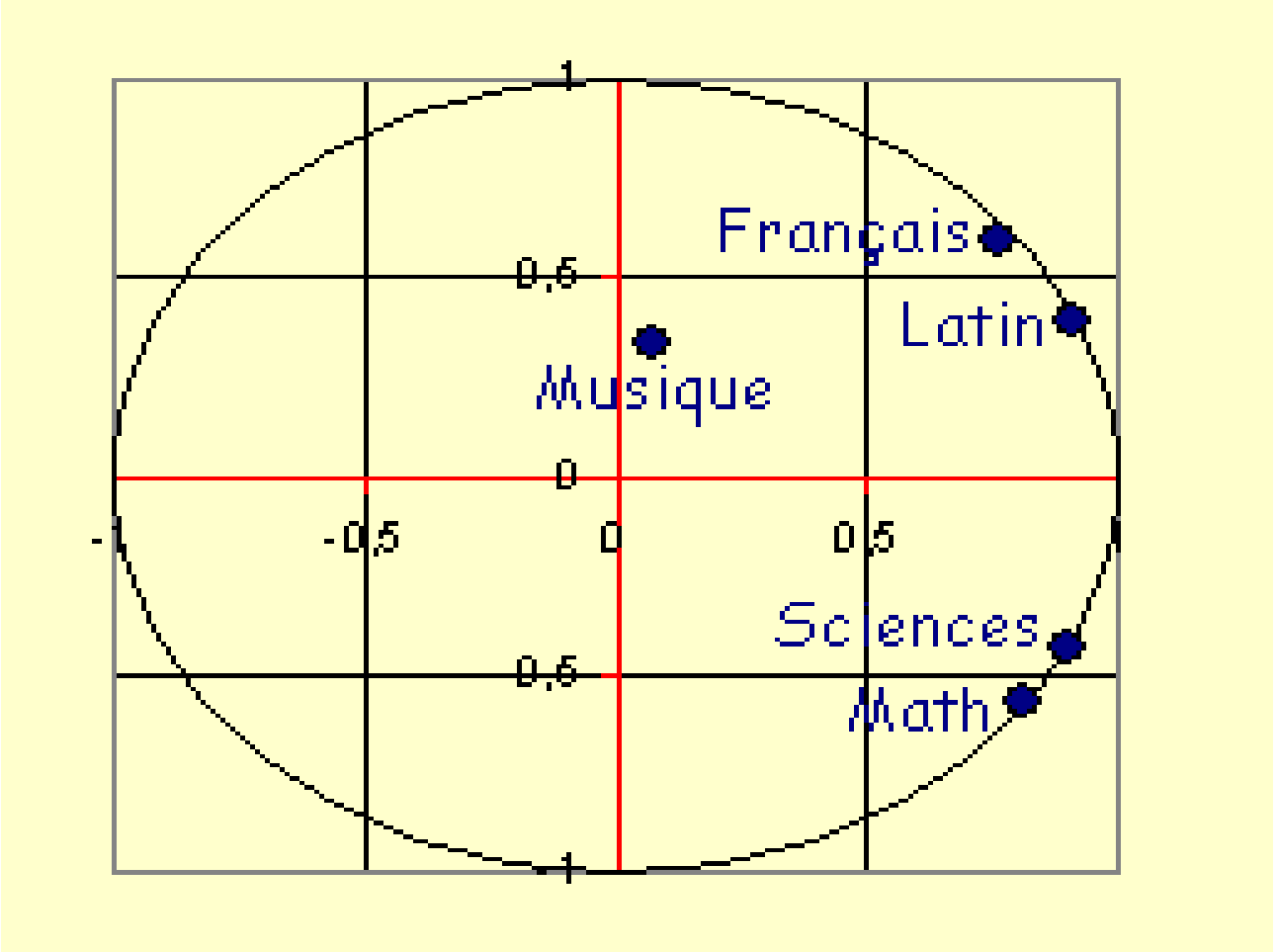
Qu'est-ce que c'est ?

Les variables sont représentées dans un cercle de rayon 1 défini par deux composantes principales v1 et v2.

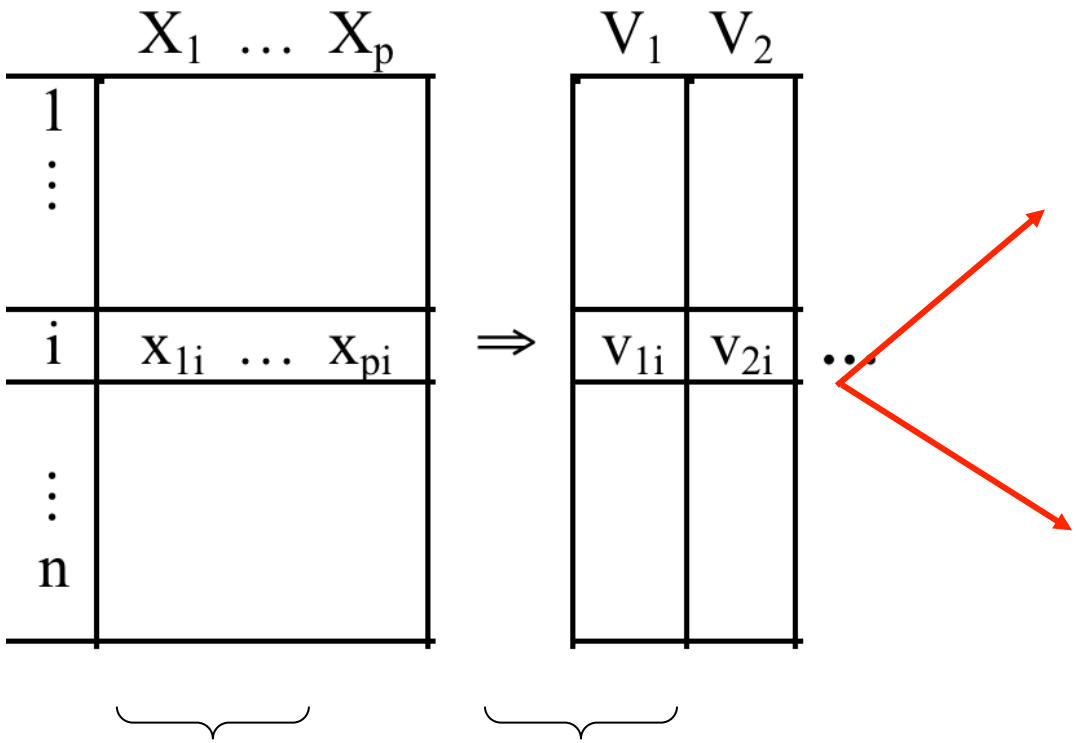
Ce cercle est appelé cercle des corrélations. Les coordonnées des variables sont les coefficients de corrélation de ces variables avec les composantes principales.



Exemple (suite)



Résumé de l'ACP

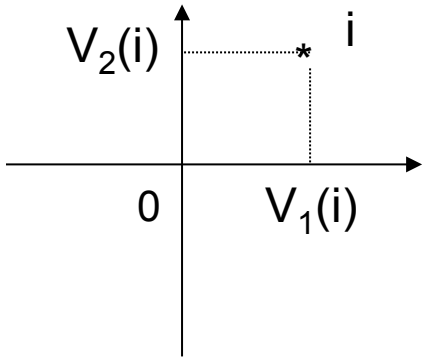


Le tableau
Des données

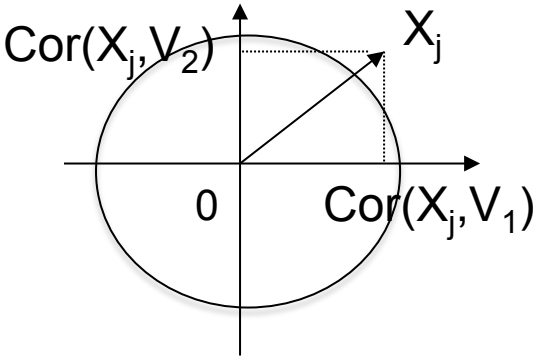
Les composantes principales

$$v_h = \sum_{j=1}^{p} u_{hj} X_j$$

(non corrélées entre elles)



Le premier plan principal



Le carte des variables

Bibliographie

Saporta G. (1990) - *Probabilités, analyse des données et statistique. Dunod, Paris.*

Escoffier B., Pagès J. (1990) - *Analyses fatorielles simples et multiples. Dunod, Paris.*

Escoffier B., Pagès J. (1997) - Initiation aux traitements statistiques, méthodes, méthodologies. Dida Stat, PUR, Rennes.

Lebart L., Morineau A., Piron M. (1995) - Statistique exploratoire multidimensionnelle. Dunod, Paris