Analyse de données

Cours 2

ANALYSE EN COMPOSANTES PRINCIPALES (ACP)

Introduction

•ACP est l'une des méthodes d'analyse de données **Multivariées**. Permettant d'explorer des données **Multidimensionnels** constitués de variables **Quantitatives**.

- L'ACP 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

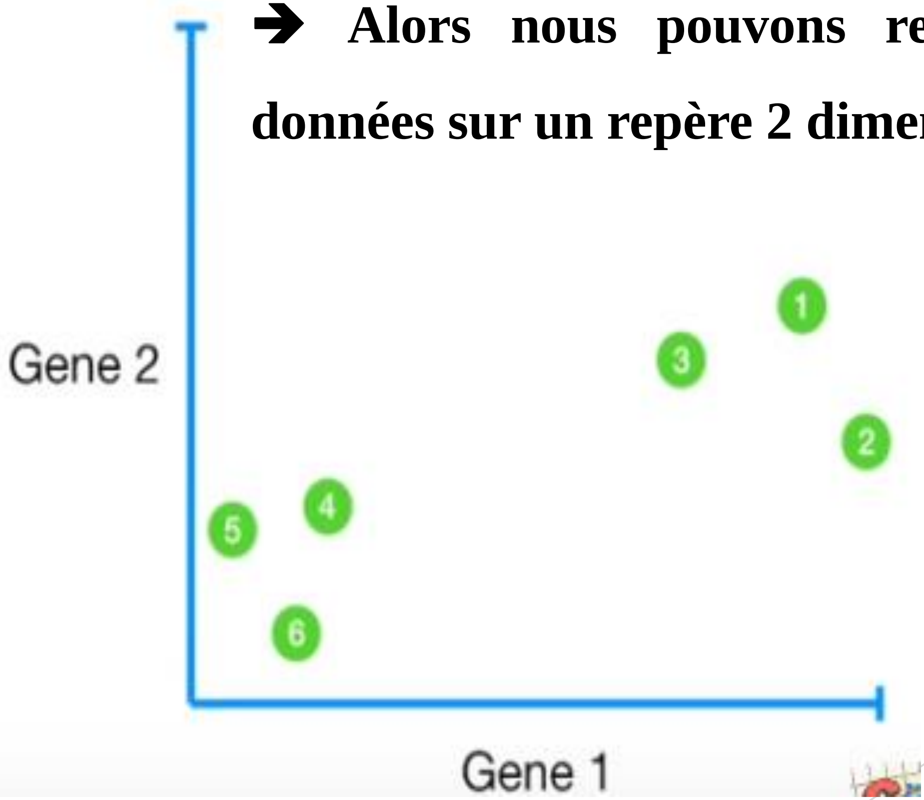
ACP est une procédure statistique utilisée pour réduire la dimensionnalité.

Illustration graphique de l'ACP

- Si nous mesurons 2 gènes ?

	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1

➔ **Alors nous pouvons représenter les données sur un repère 2 dimensions X/Y**



- L'ACP s'intéresse à des tableaux de données rectangulaires avec des **Individus** en lignes et des **variables Quantitatives** en colonnes.

$$\underset{(n,p)}{X} = \begin{array}{c|c} & \begin{array}{ccccc} 1 & \cdots & j & \cdots & p \end{array} \\ \begin{array}{c} 1 \\ \vdots \\ i \\ \vdots \\ n \end{array} & \begin{bmatrix} & & \vdots & & \\ & & \vdots & & \\ \cdots & \cdots & \boxed{x_{i,j}} & \cdots & \cdots \\ & & \vdots & & \\ & & \vdots & & \end{bmatrix} \end{array}$$

p variables **métriques**

n individus

valeur de la variable *j* prise par l'individu *i*

Projeter les observations depuis l'espace à P dimensions des P variables vers un espace à K dimensions (K < P) tel qu'un maximum d'information soit conservée

Détails des Calculs de l'ACP

Variables

$$\underset{(n,p)}{X} = \text{Individus} \begin{array}{cc} & \begin{array}{ccccc} 1 & \dots & j & \dots & p \end{array} \\ \begin{array}{c} 1 \\ \vdots \\ i \\ \vdots \\ n \end{array} & \begin{bmatrix} & & \vdots & & \\ & & \vdots & & \\ \dots & \dots & x_{i,j} & \dots & \dots \\ & & \vdots & & \\ & & \vdots & & \end{bmatrix} \end{array}$$

Composantes

$$\begin{array}{cc} & \begin{array}{ccccc} 1 & \dots & j & \dots & p \end{array} \\ \begin{array}{c} 1 \\ \vdots \\ i \\ \vdots \\ n \end{array} & \begin{bmatrix} & & \vdots & & \\ & & \vdots & & \\ \dots & \dots & u_{i,j} & \dots & \dots \\ & & \vdots & & \\ & & \vdots & & \end{bmatrix} \end{array}$$

Projeter les observations depuis l'espace à P dimensions des P variables vers un espace à K dimensions (K < P) tel qu'un maximum d'information soit conservée

Détails des Calculs de l'ACP(1)

Étape 1:

standardisez l'ensemble de données

Étape 2:

Calculez la matrice de covariance des entités du jeu de données

Étape 3:

Calculez les valeurs propres et les vecteurs propres de la matrice de covariance

Étape 4:

Trier les valeurs propres et leurs vecteurs propres correspondants

Étape 5:

Choisissez k valeurs propres et formez une matrice de vecteurs propres

Etape 6:

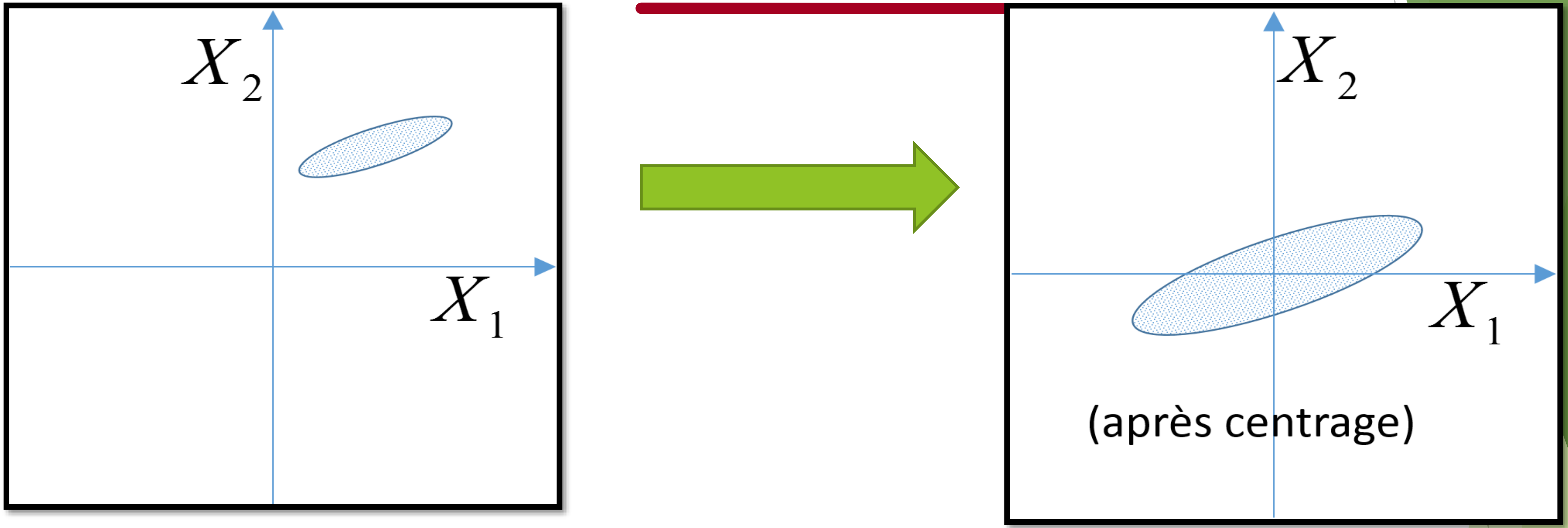
Calculer les composants principales

Détails des Calculs de l'ACP

Soit le tableau suivant :

f1	f2	f3	f4
1	2	3	4
5	5	6	7
1	4	2	3
5	3	2	1
8	1	2	2

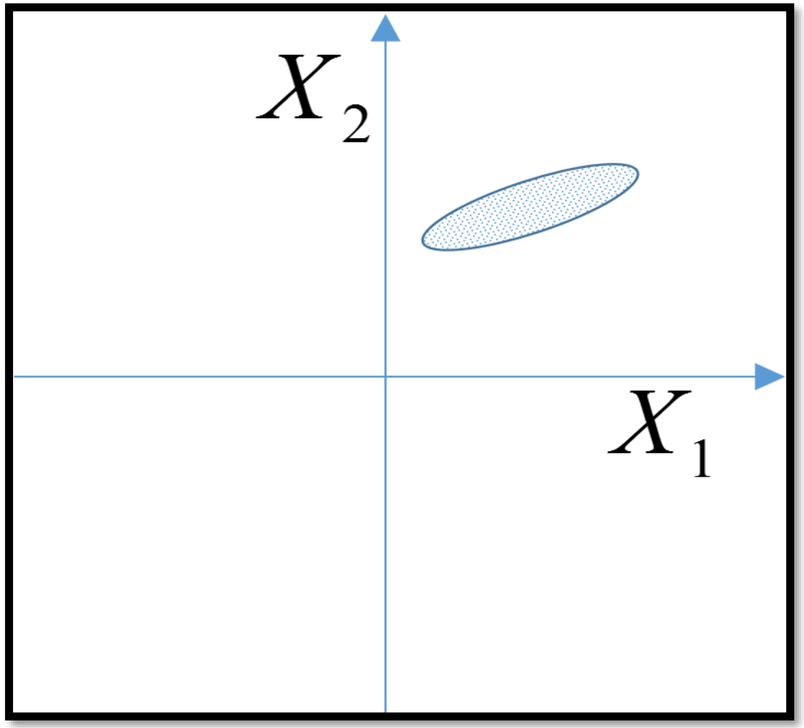
Illustration graphique de l’ACP

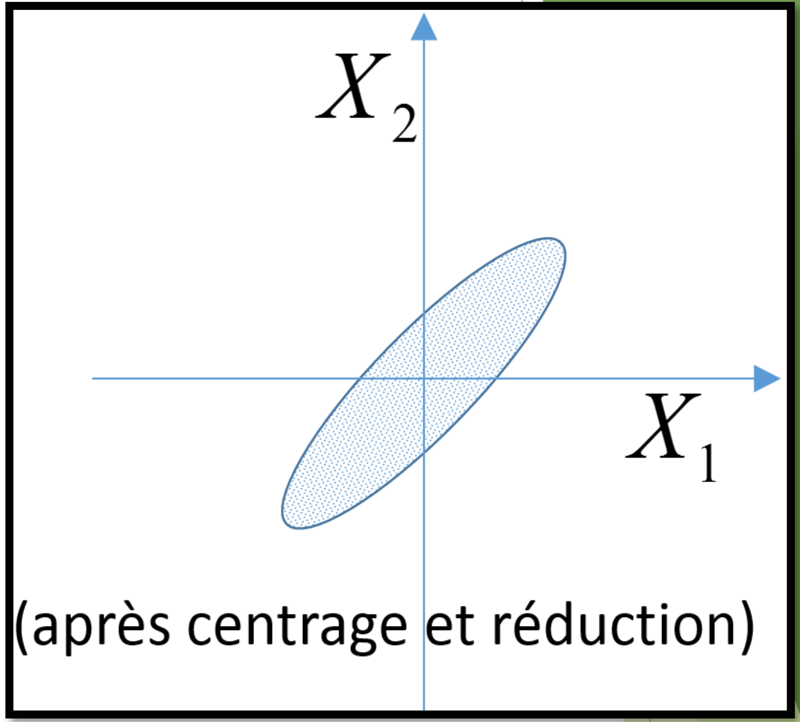


ACP centrée

Centrage préalable des variables. Cela revient à s’intéresser à la *forme* du nuage d’individus par rapport à son centre de gravité. Cette variante est utilisée lorsque les variables initiales sont directement comparables (de même nature, intervalles de variation comparables).

Illustration graphique de l'ACP





ACP centrée et réduite

Réduction préalable des variables. On s'intéresse donc à la forme du nuage d'individus après centrage et réduction des variables. Cette normalisation est employée lorsque les variables (toutes quantitatives) sont de nature différente ou présentent des intervalles de variation très différents.

Détails des Calculs de l'ACP

Étape 1: standardisez l'ensemble de données

ACP centrée

$$X_{new} = X - \bar{x}$$

ACP centrée et réduite

$$X_{new} = \frac{X - \bar{x}}{\sigma}$$

Matrice centrée

$$X_c = \begin{bmatrix} x_{11} - \overline{x}_{.1} & \cdots & x_{1j} - \overline{x}_{.j} & \cdots & x_{1p} - \overline{x}_{.p} \\ \vdots & & \vdots & & \vdots \\ x_{i1} - \overline{x}_{.1} & \cdots & x_{ij} - \overline{x}_{.j} & \cdots & x_{1ip} - \overline{x}_{.p} \\ \vdots & & \vdots & & \vdots \\ x_{n1} - \overline{x}_{.1} & \cdots & x_{nj} - \overline{x}_{.j} & \cdots & x_{np} - \overline{x}_{.p} \end{bmatrix}$$

Matrice centrée réduite

$$Y = \begin{bmatrix} (x_{11} - \bar{x}_{.1})/\sigma_1 & \cdots & (x_{1j} - \bar{x}_{.j})/\sigma_j & \cdots & (x_{1p} - \bar{x}_{.p})/\sigma_p \\ \vdots & & \vdots & & \vdots \\ (x_{i1} - \bar{x}_{.1})/\sigma_1 & \cdots & (x_{ij} - \bar{x}_{.j})/\sigma_j & \cdots & (x_{1ip} - \bar{x}_{.p})/\sigma_p \\ \vdots & & \vdots & & \vdots \\ (x_{n1} - \bar{x}_{.1})/\sigma_1 & \cdots & (x_{nj} - \bar{x}_{.j})/\sigma_j & \cdots & (x_{np} - \bar{x}_{.p})/\sigma_p \end{bmatrix}$$

Détails des Calculs de l'ACP

Étape 1: standardisez l'ensemble de données

	f1	f2	f3	f4
$\bar{x}$ =	4	3	3	3.4
σ =	3	1.58114	1.73205	2.30217

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

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{N}$$

X=

f1	f2	f3	f4
1	2	3	4
5	5	6	7
1	4	2	3
5	3	2	1
8	1	2	2

$Y=X_{centré}$,reduit

f1	f2	f3	f4
-1	-0.63246	0	0.26062
0.33333	1.26491	1.73205	1.56374
-1	0.63246	-0.57735	-0.17375
0.33333	0	-0.57735	-1.04249
1.33333	-1.26491	-0.57735	-0.60812

$Y_{11}=(1-4)/3 \ =-1$

Détails des Calculs de l'ACP

Étape 2: Calculez la matrice de covariance/Corrélation des entités du jeu de données

Les données initiales

Matrice de covariance

$$C=\begin{pmatrix} var(X_1) & cov(X_1,X_2) & \dots & cov(X_1,X_j) & \dots & cov(X_1,X_p) \\ cov(X_2,X_1) & var(X_2) & \dots & cov(X_2,X_j) & \dots & cov(X_2,X_p) \\ \dots & \dots & \dots & \dots & \dots & \dots \\ cov(X_j,X_1) & cov(X_j,X_2) & \dots & var(X_j) & \dots & cov(X_j,X_p) \\ \dots & \dots & \dots & \dots & \dots & \dots \\ cov(X_p,X_1) & cov(X_p,X_2) & \dots & cov(X_p,X_j) & \dots & var(X_p) \end{pmatrix}$$

$$\operatorname{cov}(x_1,x_2)=\frac{1}{n-1}\left(\sum_{i=1}^{n}(x_{i,1}-\bar{X}_1)(x_{i,2}-\bar{X}_2)\right)$$

Les données centrées et réduites

Matrice de corrélation

$$C=\begin{pmatrix} 1 & \rho(X_1,X_2) & \dots & \rho(X_1,X_j) & \dots & \rho(X_1,X_p) \\ \rho(X_2,X_1) & 1 & \dots & \rho(X_2,X_j) & \dots & \rho(X_2,X_p) \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \rho(X_j,X_1) & \rho(X_j,X_2) & \dots & 1 & \dots & \rho(X_j,X_p) \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \rho(X_p,X_1) & \rho(X_p,X_2) & \dots & \rho(X_p,X_j) & \dots & 1 \end{pmatrix}$$

$$corr(x_1,x_2)=\rho(x_1,x_2)=\frac{\operatorname{cov}(X_1,X_2)}{\sigma(X_1)\sigma(X_2)}$$

Détails des Calculs de l'ACP

Étape 2: Calculez la matrice de covariance des entités du jeu de données avec ()

	f1	f2	f3	f4
f1	var(f1)	cov(f1,f2)	cov(f1,f3)	cov(f1,f4)
f2	cov(f2,f1)	var(f2)	cov(f2,f3)	cov(f2,f4)
f3	cov(f3,f1)	cov(f3,f2)	var(f3)	cov(f3,f4)
f4	cov(f4,f1)	cov(f4,f2)	cov(f4,f3)	var(f4)

A=

	f1	f2	f3	f4
f1	0.8	-0.25298	0.03849	-0.14479
f2	-0.25298	0.8	0.51121	0.4945
f3	0.03849	0.51121	0.8	0.75236
f4	-0.14479	0.4945	0.75236	0.8

Étape 3: Calculez les valeurs propres et les vecteurs propres de la matrice de covariance A

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

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

Les valeurs propres
c'est le facteur multiplicatif associe a une direction donnée.

Un vecteur **v** de taille P est un vecteur propre d'une matrice **A** de taille p x p s'il existe **λ Є C** telle que

$$Av = \lambda v$$

Détails des Calculs de l'ACP

Étape 3: Calculez les valeurs propres et les vecteurs propres de la matrice de covariance

$$Av = \lambda v$$

$$Av - \lambda v = 0$$

$$(A - \lambda I)v = 0$$

$$\det(A - \lambda I) = 0$$

Détails des Calculs de l’ACP

Étape 3: Calculez les valeurs propres et les vecteurs propres de la matrice de covariance

$$\det(A - \lambda I) = 0$$

	f1	f2	f3	f4
f1	0.8 - λ	-0.25298	0.03849	-0.14479
f2	-0.25298	0.8- λ	0.51121	0.4945
f3	0.03849	0.51121	0.8 - λ	0.75236
f4	-0.14479	0.4945	0.75236	0.8 - λ

=0

$$\lambda = 2.51579324 \,,\ 1.0652885 \,,\ 0.39388704 \,,\ 0.02503121$$

Valeurs propres et vecteurs propres

- 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).
- • U1 est le vecteur propre unitaire associé à la plus grande valeur propre $\lambda 1$, il vérifie $X'X\ U1 = \lambda 1\ U1$ et $||U1\ || = 1$
- • 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)

Détails des Calculs de l'ACP

Étape 3: Calculez les valeurs propres et les vecteurs propres de la matrice de covariance

$$(A-\lambda I)v=0$$

Ou

A*v= λ *v

$$\begin{pmatrix} 0.800000-\lambda & -(0.252982) & 0.038490 & -(0.144791) \\ -(0.252982) & 0.800000-\lambda & 0.511208 & 0.494498 \\ 0.038490 & 0.511208 & 0.800000-\lambda & 0.752355 \\ -(0.144791) & 0.494498 & 0.752355 & 0.800000-\lambda \end{pmatrix} \times \begin{pmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{pmatrix} = 0$$

Détails des Calculs de l'ACP

Étape 3: Calculez les valeurs propres et les vecteurs propres de la matrice de covariance

$$(A - \lambda I)v = 0$$

Pour $\lambda = 2.51579324$, *les valeurs du vecteur V obtenu sont*

$v1 = 0.16195986$

$v2 = -0.52404813$

$v3 = -0.58589647$

$v4 = -0.59654663$

Étape 4: Trier les valeurs propres

$\lambda 1 = 2.51579324$, $\lambda 2 = 1.0652885$, $\lambda 3 = 0.39388704$, $\lambda 4 = 0.02503121$

Étape 5: Choisissez k valeurs propres et formez une matrice de vecteurs propres

- En pratique, on arrête l'extraction des valeurs propres lorsque la somme des **k** valeurs propres que l'on a déterminés représente un pourcentage satisfaisant de la variance.

Pourcentage de variation expliqué par: $$\frac{\lambda_i}{\sum_{1}^{p} \lambda_i}$$

% de $\lambda 1$ **:** 0,63=63%

% de $\lambda 2$ 0,27=27%

63+27=100%

$\rightarrow$ K=2

Calculs de vecteurs propres

Vp1 vecteur propre de λ_1

Vp2 vecteur propre de $\lambda 2$

Vp1	Vp2
0.161960	-0.917059
-0.524048	0.206922
-0.585896	-0.320539
-0.596547	-0.115935

Etape 6: Calculer les composants principales

(Transformez la matrice d'origine)

X:matrice de depart
Uj: vecteur propre

$C_j=XU_j$

Variables

$$\underset{(n,p)}{X} = \text{Individus} \begin{array}{cc} & \begin{array}{ccccc} 1 & \cdots & j & \cdots & p \end{array} \\ \begin{array}{c} 1 \\ \vdots \\ i \\ \vdots \\ n \end{array} & \begin{bmatrix} & & \vdots & & \\ & & \vdots & & \\ \cdots & \cdots & x_{i,j} & \cdots & \cdots \\ & & \vdots & & \\ & & \vdots & & \end{bmatrix} \end{array}$$

Composantes

$$\begin{array}{cc} & \begin{array}{ccccc} 1 & \cdots & j & \cdots & k \end{array} \\ \begin{array}{c} 1 \\ \vdots \\ i \\ \vdots \\ n \end{array} & \begin{bmatrix} & & \vdots & & \\ & & \vdots & & \\ \cdots & \cdots & u_{i,j} & \cdots & \cdots \\ & & \vdots & & \\ & & \vdots & & \end{bmatrix} \end{array}$$

Dans notre exemple: on a deux composantes principales (car k=2)

$C_1=XU_1$ et $C2=XU_2$

Étape 6: ou on calcule :

(Matrice de caractéristiques * k principaux vecteurs propres = données transformées)

f1	f2	f3	f4
-1.000000	-0.632456	0.000000	0.260623
0.333333	1.264911	1.732051	1.563740
-1.000000	0.632456	-0.577350	-0.173749
0.333333	0.000000	-0.577350	-1.042493
1.333333	-1.264911	-0.577350	-0.608121

(5,4)

*

Vp1	Vp2
0.161960	-0.917059
-0.524048	0.206922
-0.585896	-0.320539
-0.596547	-0.115935

(4,2)

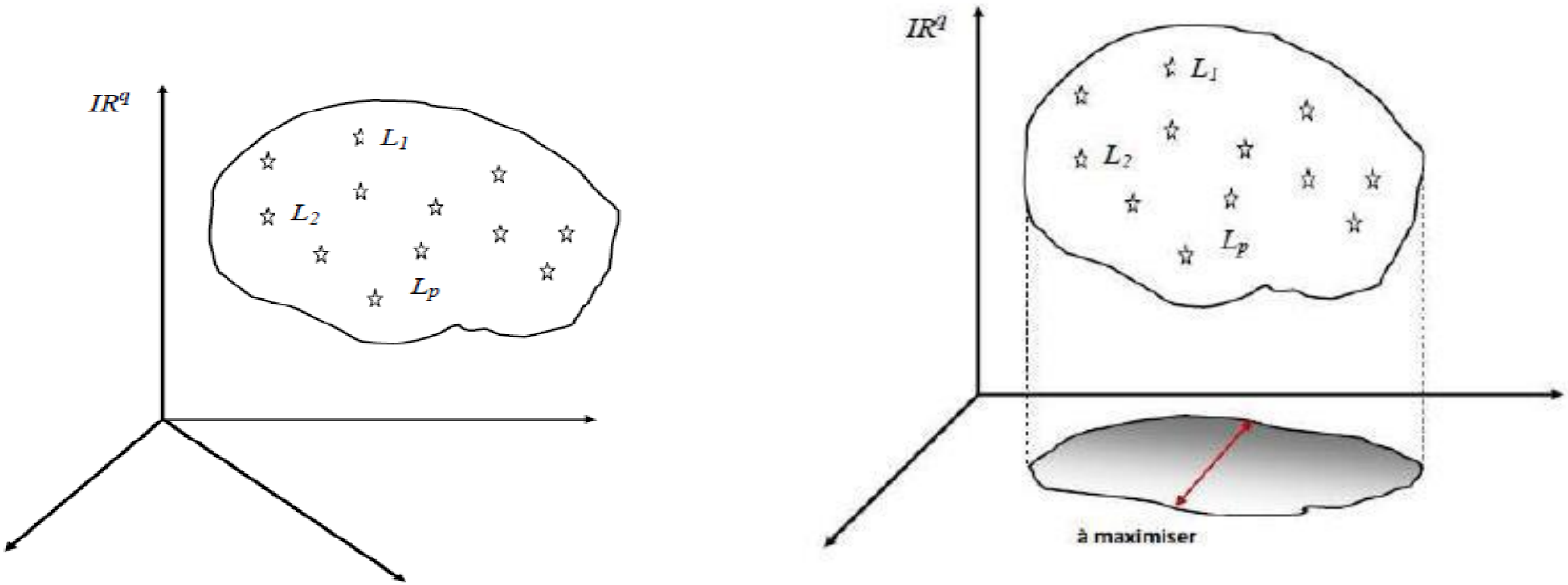
=

C1	C2
0.014003	0.755975
-2.556534	-0.780432
-0.051480	1.253135
1.014150	0.000239
1.579861	-1.228917

(5,2)

L'ACP (Géométriquement)

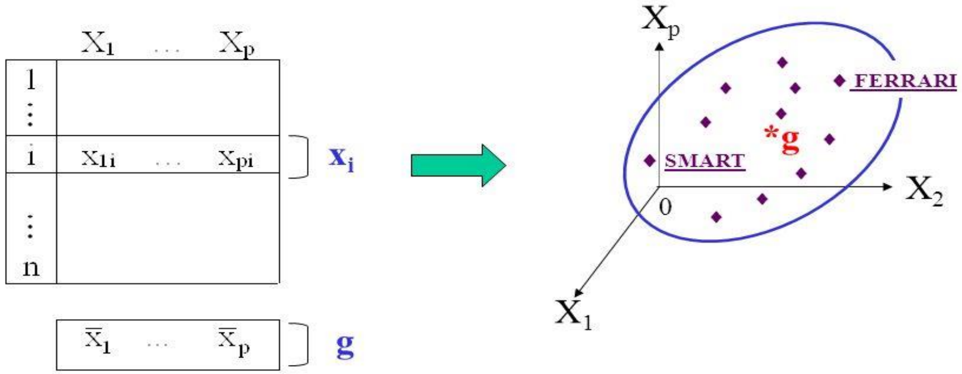
- Projette un nuage de point sur un sous espace de dimension inferieur
- Lors de la projection, le nuage peut être déformé est donc serait différent de réel, alors les méthodes d'ajustement consistent en minimiser cette possible déformation et ce en maximisant les distances projetées



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.

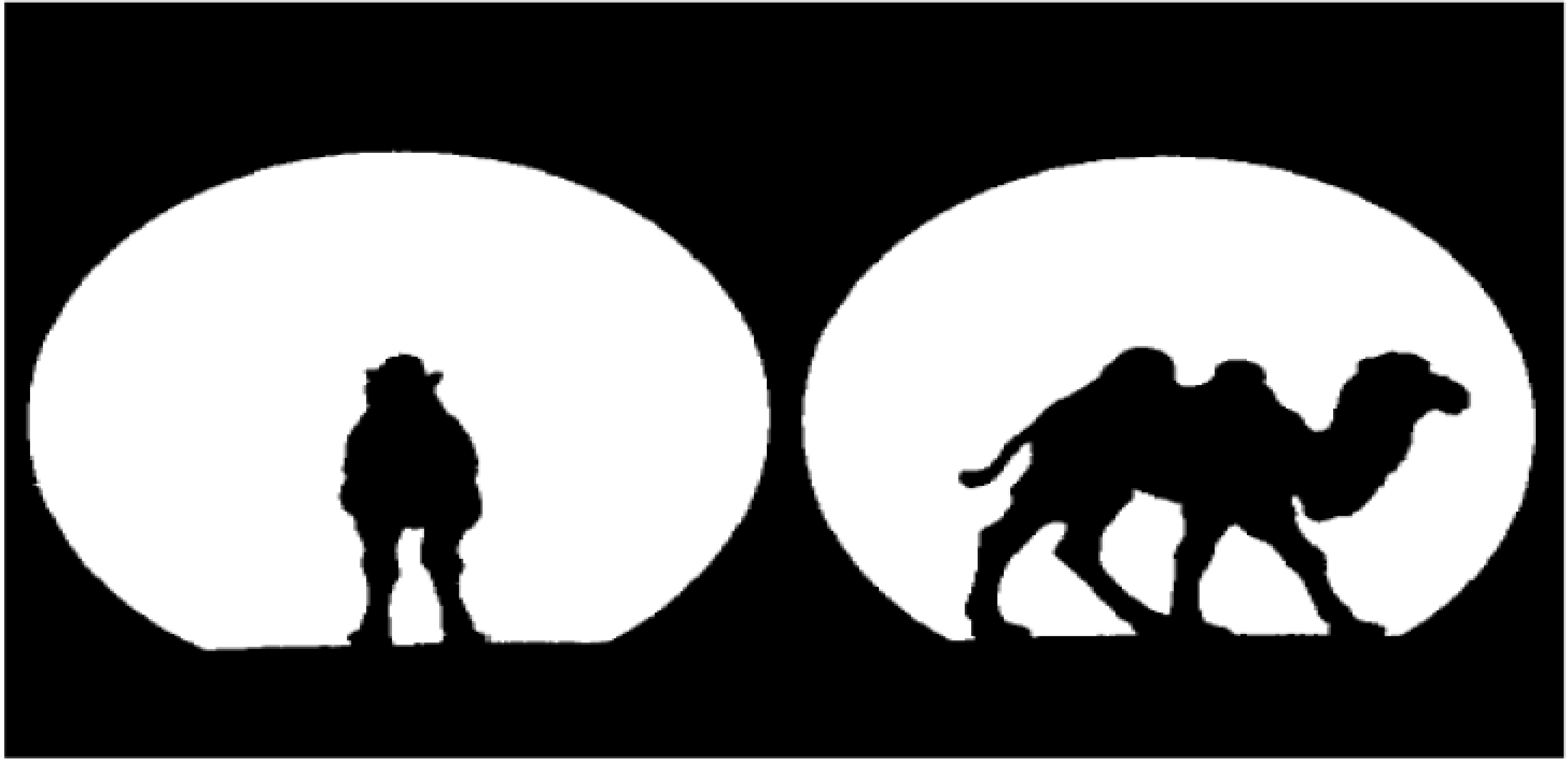
$$G=\begin{pmatrix}\bar{x}_1\\ \bar{x}_2\\ \ldots\\ \bar{x}_p\end{pmatrix}$$



Les données centrées réduites

G*=0

L'ACP (Géométriquement)



Distance ou métrique utilisée

•Soient Lm et Ln deux points de IR_q:

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

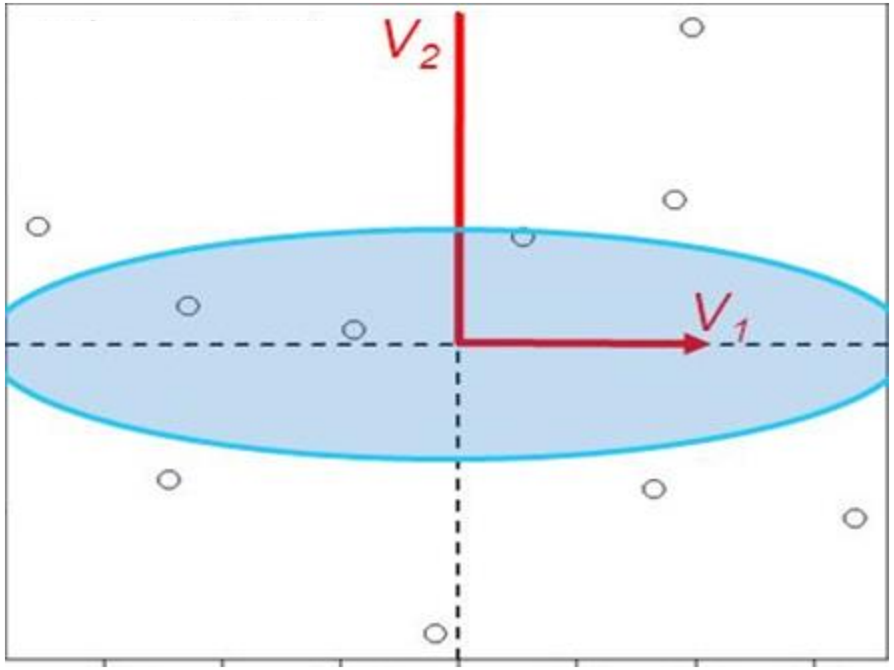
$$L_n = \left(x_{n1}, x_{n2}, \cdots, x_{nj}, \cdots, x_{nq}\right)$$

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

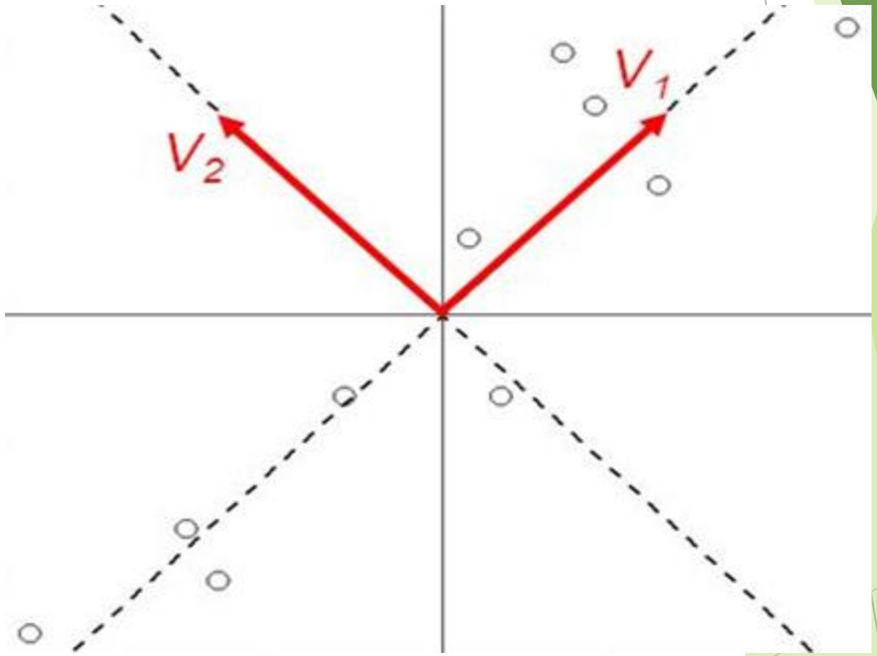
$$d(L_m, L_n) = \sqrt{\sum_{j=1}^{q} \left(x_{mj} - x_{nj}\right)^2}$$

L'ACP (Géométriquement)

V1 et V2 sont les vecteurs propres



ACP



Nuage de points initiale

V1 et V2 sont des axes factoriels

L'ACP (Géométriquement)

•

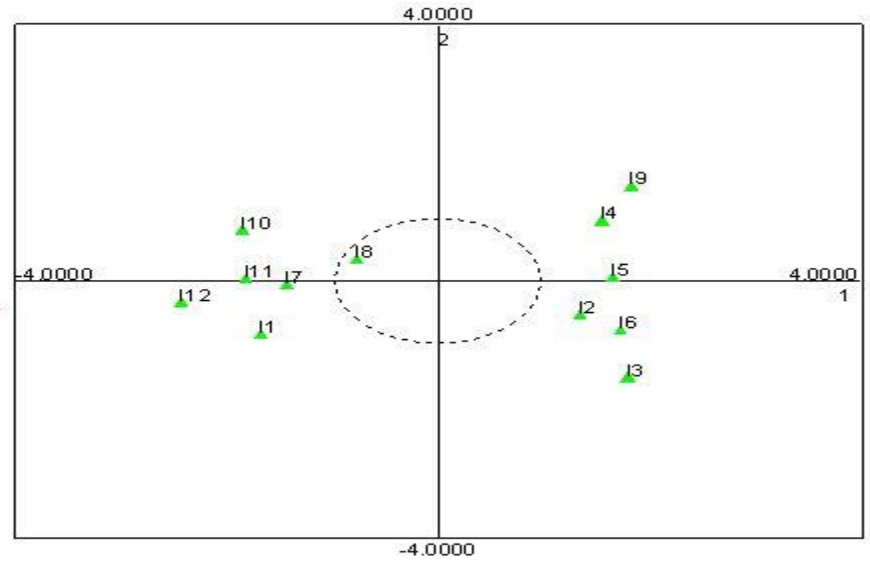
NOUVELLES COORDONNEES

SCORES

ind	w300	w350	w400	w450
I1	16.0	62.0	67.0	27.0
I2	15.0	60.0	69.0	31.0
I3	14.0	59.0	68.0	31.0
I4	15.0	61.0	71.0	31.0
I5	14.0	60.0	70.0	30.0
I6	14.0	59.0	69.0	30.0
I7	17.0	63.0	68.0	29.0
I8	16.0	62.0	69.0	28.0
I9	15.0	60.0	72.0	30.0
I10	17.0	63.0	69.0	27.0
I11	18.0	62.0	68.0	28.0
I12	18.0	64.0	67.0	29.0

	Axe : 1	Axe : 2
	Coord	Coord
I1	-1.666	-0.801
I2	1.348	-0.487
I3	1.800	-1.476
I4	1.559	0.968
I5	1.664	0.082
I6	1.730	-0.740
I7	-1.424	-0.017
I8	-0.765	0.364
I9	1.834	1.516
I10	-1.840	0.821
I11	-1.811	0.065
I12	-2.430	-0.295

ACP



PLAN FACTORIEL 1,2

ACP Normée   et ACP non normée

- Nous parlerons d'ACP non normée lorsque les données sont seulement centrées,

- Nous parlerons d'ACP normée lorsque les données sont centrées et réduites