Introduction à l’Analyse Des Données (ADD)

A- Les méthodes

B- Exemples

C- Les données

A- Les méthodes

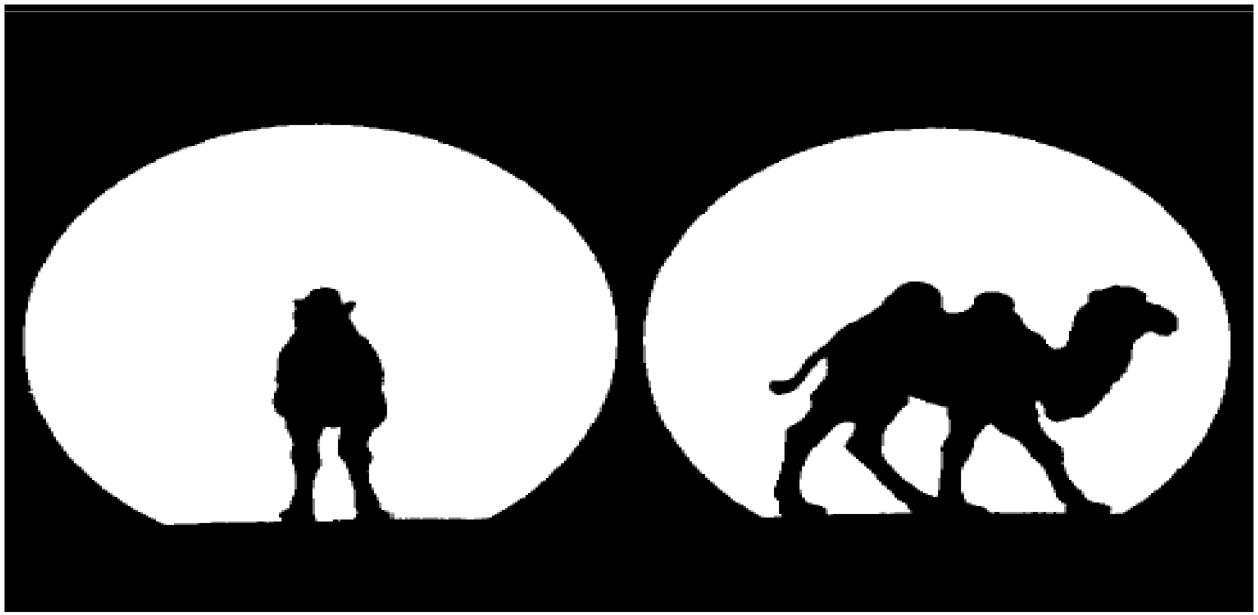
- ✓ Lors de toute étude statistique, il est nécessaire de *décrire* et *explorer* les données avant d’en tirer de quelconques lois ou modèles prédictifs.

- ✓ Dans beaucoup de situations, les données sont trop nombreuses pour pouvoir être visualisables (nombre de caractéristiques trop élevées)

- ✓ Il est alors nécessaire d’extraire l’information pertinente qu’elles contiennent ; Les techniques d’ADD répondent à ce besoin.

A - Les méthodes

- ✓ **ADD** = ensemble de méthodes descriptives ayant pour objectif de ***résumer*** et ***visualiser*** *l'information* ***pertinente*** *contenue dans un grand tableau de données*



A - Les méthodes

✓ **Trois grandes familles de méthodes:**

Objectif	Variables quanti	Variables quali/mixtes
Repérer et visualiser les corrélations multiples entre variables et/ou les ressemblances entre individus	**Analyse en composantes principales (ACP)**	Analyse factorielle des correspondances (AFC AFCM)
Réaliser une typologie des individus	**Methodes de classification (CAH,..)**	AFC ou AFCM et classification
Caractériser de groupes d'individus à l'aide de variables	Analyse discriminante (AFD,..)	Analyse discriminante (AFD,..)

B- Exemples

On dispose de 6 variables représentant les taux de différents délits commis pour 100000 habitants dans 20 Etats des Etats-unis. Ces données peuvent être mises dans un tableau individu*variable

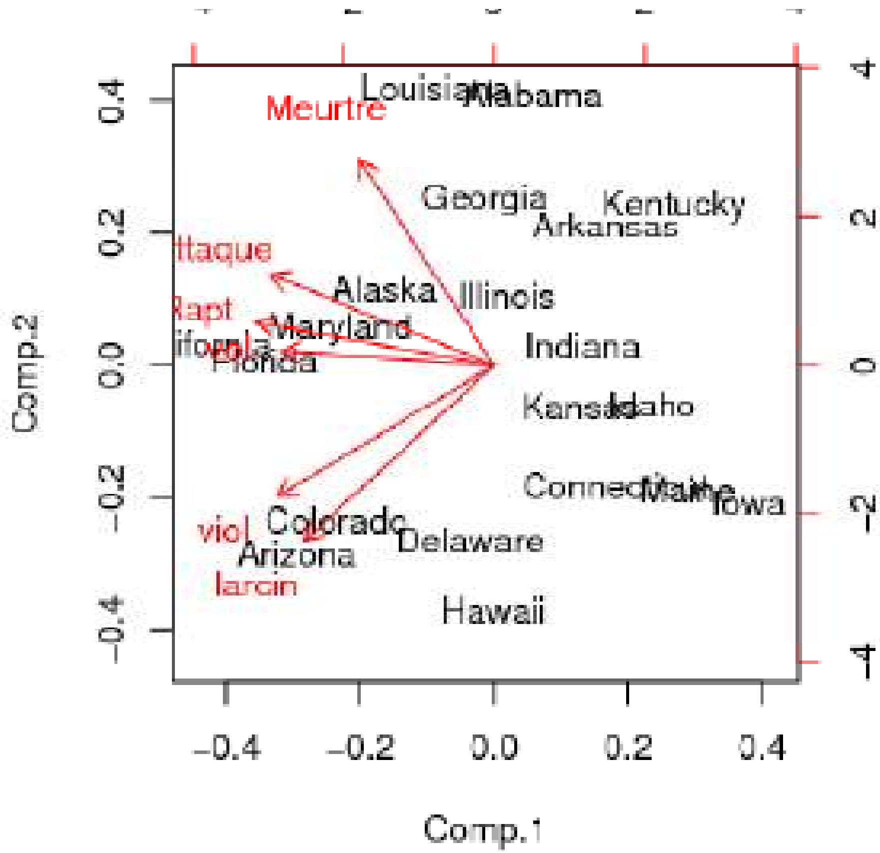
ETAT	Meurtre	Rapt	vol	attaque	viol	larcin
Alabama	14.2	25.2	96.8	278.3	1135.5	1881.9
Alaska	10.8	51.6	96.8	284.0	1331.7	3369.8
Arizona	9.5	34.2	138.2	312.3	2346.1	4467.4
Arkansas	8.8	27.6	83.2	203.4	972.6	1862.1
California	11.5	49.4	287.0	358.0	2139.4	3499.8
Colorado	6.3	42.0	170.7	292.9	1935.2	3903.2
Connecticut	4.2	16.8	129.5	131.8	1346.0	2620.7
Delaware	6.0	24.9	157.0	194.2	1682.6	3678.4
Florida	10.2	39.6	187.9	449.1	1859.9	3840.5
Georgia	11.7	31.1	140.5	256.5	1351.1	2170.2
Hawaii	7.2	25.5	128.0	64.1	1911.5	3920.4
Idaho	5.5	19.4	39.6	172.5	1050.8	2599.6
Illinois	9.9	21.8	211.3	209.0	1085.0	2828.5
Indiana	7.4	26.5	123.2	153.5	1086.2	2498.7
Iowa	2.3	10.6	41.2	89.8	812.5	2685.1
Kansas	6.6	22.0	100.7	180.5	1270.4	2739.3
Kentucky	10.1	19.1	81.1	123.3	872.2	1662.1
Louisiana	15.5	30.9	142.9	335.5	1165.5	2469.9
Maine	2.4	13.5	38.7	170.0	1253.1	2350.7
Maryland	8.0	34.8	292.1	358.9	1400.0	3177.7

B- Exemples (ACP sous R)

- Deux grandes tendances :

✓ L’axe 1 distingue les états de Floride, Colorado, Arizona, Californie, Maryland caractérisés par un fort taux de délits en tous genres aux autres états.

✓ L’axe 2 est un axe de gravité des délits : s’oppose les états ayant un fort taux de délits mineurs (Colorado, Arizona) aux états concernés par des délits majeurs (Alabama, Louisiane).



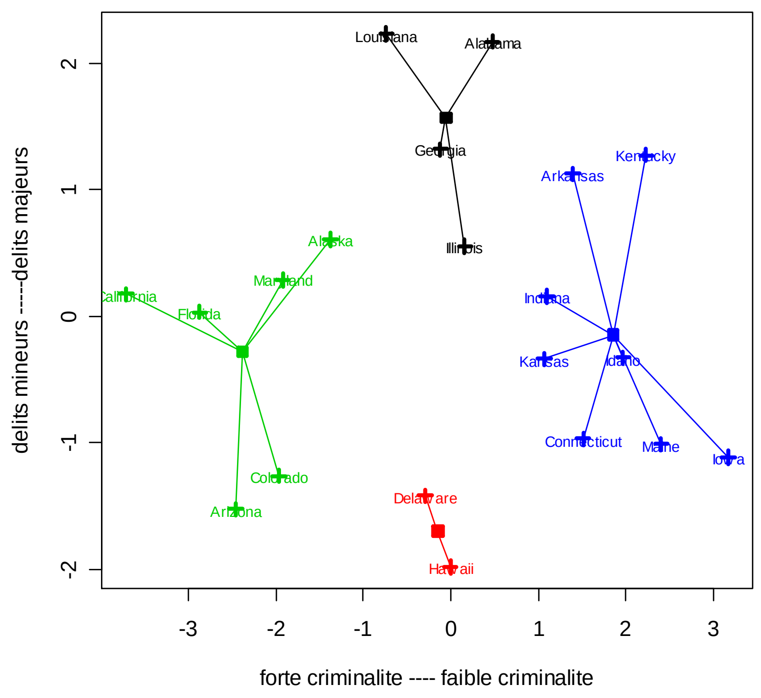
B- Exemples (classif sous R)

- Classification

On distingue 4 groupes d'états :

- ✓ le groupe vert , caractérisé par un taux de délits en tous genres inférieur à la moyenne
- ✓ Le groupe bleu caractérisé par un taux de délits en tous genres supérieur à la moyenne
- ✓ Le groupe noir caractérisé par un taux de délits graves supérieur à la moyenne
- ✓ Le groupe rouge caractérisé par un taux de délits mineurs supérieur à la moyenne

représentation dans les axes d'une ACP(programme3)



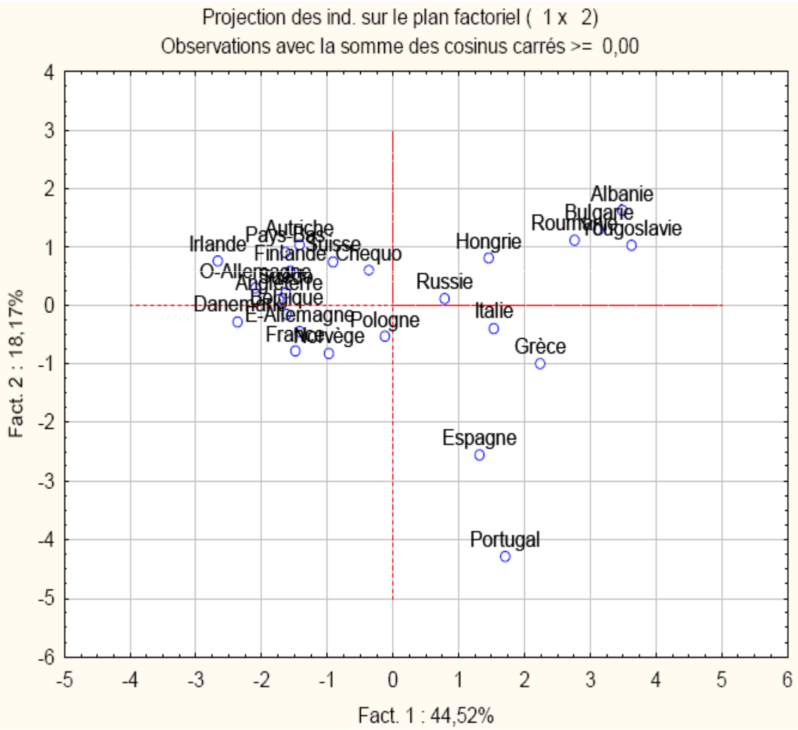
B- Exemples

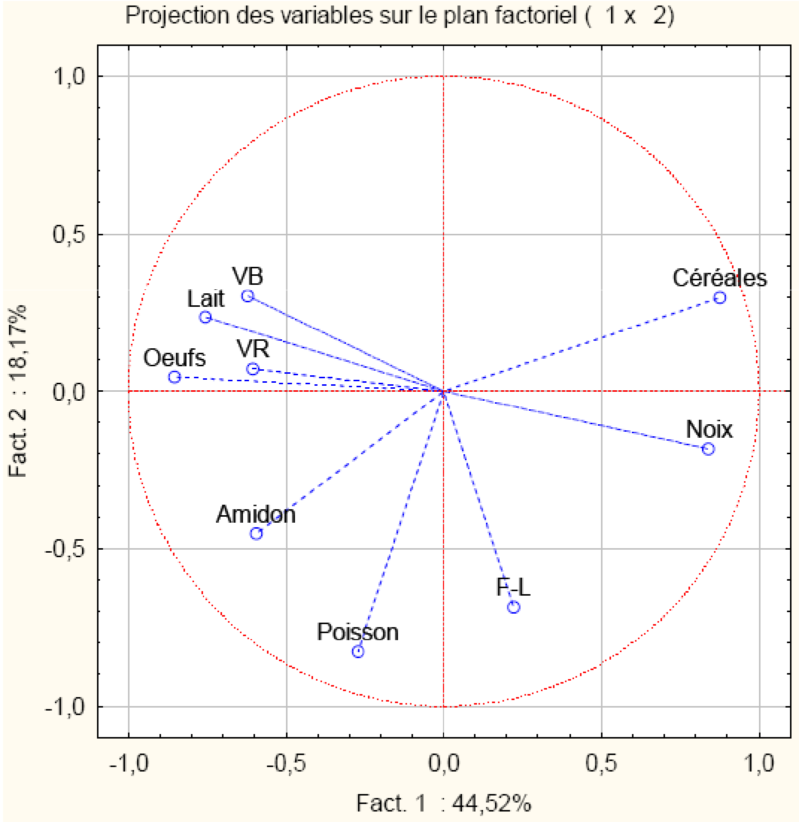
Les données mesurent la consommation de protéines dans 25 pays européens par rapport à 9 groupes d'aliments.

VR: Viande rouge ; VB: Viande blanche ; Starch: Starchy foods ; FV: Fruits et légumes

Pays	VR	VB	Oeufs	Lait	Poisson	Céréales	Starch	Noix	FL
Albanie	10.1	1.4	0.5	8.9	0.2	42.3	0.6	5.5	1.7
Autriche	8.9	14.0	4.3	19.9	2.1	28.0	3.6	1.3	4.3
Belgique	13.5	9.3	4.1	17.5	4.5	26.6	5.7	2.1	4.0
Bulgarie	7.8	6.0	1.6	8.3	1.2	56.7	1.1	3.7	4.2
Cheko.	9.7	11.4	2.8	12.5	2.0	34.3	5.0	1.1	4.0
Danemark	10.6	10.8	3.7	25.0	9.9	21.9	4.8	0.7	2.4
Allemagne-E	8.4	11.6	3.7	11.1	5.4	24.6	6.5	0.8	3.6
Finlande	9.5	4.9	2.7	33.7	5.8	26.3	5.1	1.0	1.4
France	18.0	9.9	3.3	19.5	5.7	28.1	4.8	2.4	6.5
Grèce	10.2	3.0	2.8	17.6	5.9	41.7	2.2	7.8	6.5
Hongrie	5.3	12.4	2.9	9.7	0.3	40.1	4.0	5.4	4.2
Irlande	13.9	10.0	4.7	25.8	2.2	24.0	6.2	1.6	2.9
Italie	9.0	5.1	2.9	13.7	3.4	36.8	2.1	4.3	6.7
Pays-bas	9.5	13.6	3.6	23.4	2.5	22.4	4.2	1.8	3.7
Norvège	9.4	4.7	2.7	23.3	9.7	23.0	4.6	1.6	2.7
Pologne	6.9	10.2	2.7	19.3	3.0	36.1	5.9	2.0	6.6
Portugal	6.2	3.7	1.1	4.9	14.2	27.0	5.9	4.7	7.9
Roumanie	6.2	6.3	1.5	11.1	1.0	49.6	3.1	5.3	2.8
Espagne	7.1	3.4	3.1	8.6	7.0	29.2	5.7	5.9	7.2
Suède	9.9	7.8	3.5	24.7	7.5	19.5	3.7	1.4	2.0
Suisse	13.1	10.1	3.1	23.8	2.3	25.6	2.8	2.4	4.9
Angleterre	17.4	5.7	4.7	20.6	4.3	24.3	4.7	3.4	3.3
Russie	9.3	4.6	2.1	16.6	3.0	43.6	6.4	3.4	2.9
Allemagne-O	11.4	12.5	4.1	18.8	3.4	18.6	5.2	1.5	3.8
Yougoslavie	4.4	5.0	1.2	9.5	0.6	55.9	3.0	5.7	3.2

B – Exemples
(ACP sous statistica)





C –1 Les données: tableau individu*variables

- ✓ On observe p caractéristiques $\boxed{X_1,\ldots..X_p}$ quantitatives sur n individus $\boxed{e_1,\ldots.,e_i,\ldots.e_n}$

- ✓ On note $\boxed{x_{ij}}$ la valeur de la variable X_j observée sur l'individu e_i

Individu	X_1	X_2		X_j		X_p
e1	x_{11}	x_{12}		x_{1j}		x_{1p}
e2	x_{21}	x_{22}		x_{2j}		x_{2p}
ei	x_{i1}	x_{i2}		x_{ij}		x_{ip}
en	x_{n1}	x_{n2}		x_{nj}		x_{np}

C –1 Les données: tableau individu*variables

- ✓ Le tableau peut être mis sous forme matricielle

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{np} \end{bmatrix}$$

C –1 Les données : tableau individu*variables

- ✓ Chaque individu est décrit par p variables, formant un vecteur de dimension p, appelé *vecteur individu*.

$$e_i = \begin{pmatrix} x_{i1} \\ \ldots \\ x_{ij} \\ \ldots \\ x_{ip} \end{pmatrix} \in R^p$$

C –1 Les données : tableau individu*variables

- ✓ Chaque variable peut être représentée par un vecteur de dimension n , appelé *vecteur variable*, correspondant aux valeurs prises par cette variable sur les n individus.

$$x_{.j} = \begin{pmatrix} x_{1j} \\ \dots \\ x_{ij} \\ \dots \\ x_{nj} \end{pmatrix} \in R^n$$

C –1 Les données: tableau individu*variables

Les données mesurent la consommation de protéines dans 25 pays européens par rapport à 9 groupes d'aliments.

VR: Viande rouge ; VB: Viande blanche ; Starch: Starchy foods ; FV: Fruits et légumes

Pays	VR	VB	Oeufs	Lait	Poisson	Céréales	Starch	Noix	FL
Albanie	10.1	1.4	0.5	8.9	0.2	42.3	0.6	5.5	1.7
Autriche	8.9	14.0	4.3	19.9	2.1	28.0	3.6	1.3	4.3
Belgique	13.5	9.3	4.1	17.5	4.5	26.6	5.7	2.1	4.0
Bulgarie	7.8	6.0	1.6	8.3	1.2	56.7	1.1	3.7	4.2
Cheko.	9.7	11.4	2.8	12.5	2.0	34.3	5.0	1.1	4.0
Danemark	10.6	10.8	3.7	25.0	9.9	21.9	4.8	0.7	2.4
Allemagne-E	8.4	11.6	3.7	11.1	5.4	24.6	6.5	0.8	3.6
Finlande	9.5	4.9	2.7	33.7	5.8	26.3	5.1	1.0	1.4
France	18.0	9.9	3.3	19.5	5.7	28.1	4.8	2.4	6.5
Grèce	10.2	3.0	2.8	17.6	5.9	41.7	2.2	7.8	6.5
Hongrie	5.3	12.4	2.9	9.7	0.3	40.1	4.0	5.4	4.2
Irlande	13.9	10.0	4.7	25.8	2.2	24.0	6.2	1.6	2.9
Italie	9.0	5.1	2.9	13.7	3.4	36.8	2.1	4.3	6.7
Pays-bas	9.5	13.6	3.6	23.4	2.5	22.4	4.2	1.8	3.7
Norvège	9.4	4.7	2.7	23.3	9.7	23.0	4.6	1.6	2.7
Pologne	6.9	10.2	2.7	19.3	3.0	36.1	5.9	2.0	6.6
Portugal	6.2	3.7	1.1	4.9	14.2	27.0	5.9	4.7	7.9
Roumanie	6.2	6.3	1.5	11.1	1.0	49.6	3.1	5.3	2.8
Espagne	7.1	3.4	3.1	8.6	7.0	29.2	5.7	5.9	7.2
Suède	9.9	7.8	3.5	24.7	7.5	19.5	3.7	1.4	2.0
Suisse	13.1	10.1	3.1	23.8	2.3	25.6	2.8	2.4	4.9
Angleterre	17.4	5.7	4.7	20.6	4.3	24.3	4.7	3.4	3.3
Russie	9.3	4.6	2.1	16.6	3.0	43.6	6.4	3.4	2.9
Allemagne-O	11.4	12.5	4.1	18.8	3.4	18.6	5.2	1.5	3.8
Yougoslavie	4.4	5.0	1.2	9.5	0.6	55.9	3.0	5.7	3.2

C.2- Les données : Grandeurs associées au tableau de données

a- *Matrice des poids associés aux individus*

- ✓ *Les données peuvent être pondérées : Le poids attribué à chaque individu* exprime l'importance que l'on désire lui accorder dans l'étude (représentativité de l'échantillon étudié dans la population) :

$$P=\begin{bmatrix} p_1 & 0 & 0 & 0 & 0 \\ 0 & \ldots & 0 & 0 & 0 \\ 0 & 0 & p_i & 0 & 0 \\ 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & 0 & p_n \end{bmatrix} \qquad 0 \leq p_i \leq 1, \quad i=1,\ldots n \qquad \sum_{i=1}^{n} p_i = 1$$

- ✓ Généralement $P=\frac{1}{n}I_n$ (même poids pour tous les individus)

C.2- Les données : Grandeurs associées au tableau de données

b- *Nuages de points*

Ils permettent de visualiser les liens entre les variables ou les ressemblances/dissemblances entre individus contenus dans le tableau de données X.

- ✓ *Nuage des points-individus* = coordonnées des n vecteurs individus e_i dans le repère de R^p dont les axes sont les p variables du tableau.

$$e_i = \left[x_{i1}, \dots, x_{ij}, \dots, x_{ip} \right]'$$

- ✓ *Nuage des points-variables* = coordonnées des p vecteurs variables X_j dans le repère de R^n dont les axes sont déterminés par les n individus.

$$X_j = [x_{1j}, \dots, x_{ij}, \dots x_{nj}]'$$

C.2- Les données : Grandeurs associées au tableau de données

b- *Nuages de points*

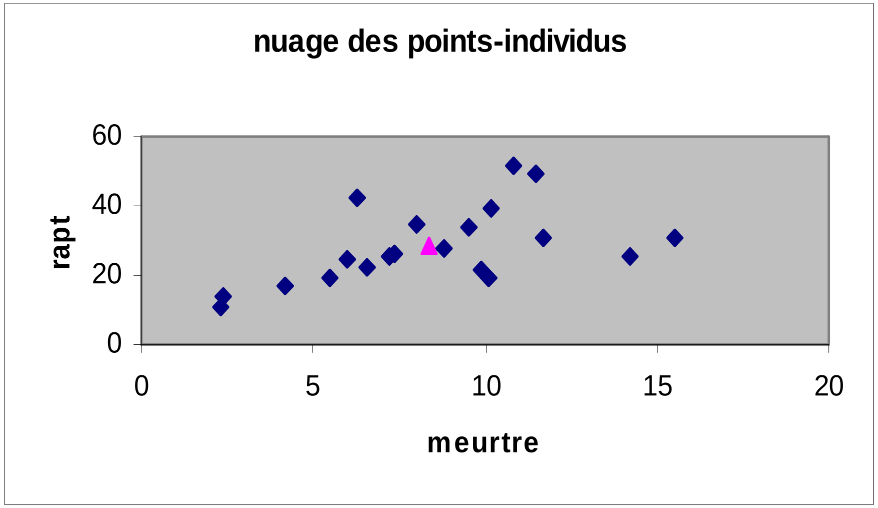
On dispose de 6 variables représentant les taux de différents délits commis pour 100000 habitants dans 20 Etats des Etats-unis. Ces données peuvent être mises dans un tableau individu*variable

ETAT	Meurtre	Rapt	vol	attaque	viol	larcin
Alabama	14.2	25.2	96.8	278.3	1135.5	1881.9
Alaska	10.8	51.6	96.8	284.0	1331.7	3369.8
Arizona	9.5	34.2	138.2	312.3	2346.1	4467.4
Arkansas	8.8	27.6	83.2	203.4	972.6	1862.1
California	11.5	49.4	287.0	358.0	2139.4	3499.8
Colorado	6.3	42.0	170.7	292.9	1935.2	3903.2
Connecticut	4.2	16.8	129.5	131.8	1346.0	2620.7
Delaware	6.0	24.9	157.0	194.2	1682.6	3678.4
Florida	10.2	39.6	187.9	449.1	1859.9	3840.5
Georgia	11.7	31.1	140.5	256.5	1351.1	2170.2
Hawaii	7.2	25.5	128.0	64.1	1911.5	3920.4
Idaho	5.5	19.4	39.6	172.5	1050.8	2599.6
Illinois	9.9	21.8	211.3	209.0	1085.0	2828.5
Indiana	7.4	26.5	123.2	153.5	1086.2	2498.7
Iowa	2.3	10.6	41.2	89.8	812.5	2685.1
Kansas	6.6	22.0	100.7	180.5	1270.4	2739.3
Kentucky	10.1	19.1	81.1	123.3	872.2	1662.1
Louisiana	15.5	30.9	142.9	335.5	1165.5	2469.9
Maine	2.4	13.5	38.7	170.0	1253.1	2350.7
Maryland	8.0	34.8	292.1	358.9	1400.0	3177.7

C.2- Les données : Grandeurs associées au tableau de données

b- *Nuages de points*

✓ Les n individus forment un nuage de points dans le sous-espace de R^p défini par les variables, appelé *nuage des points-individus*

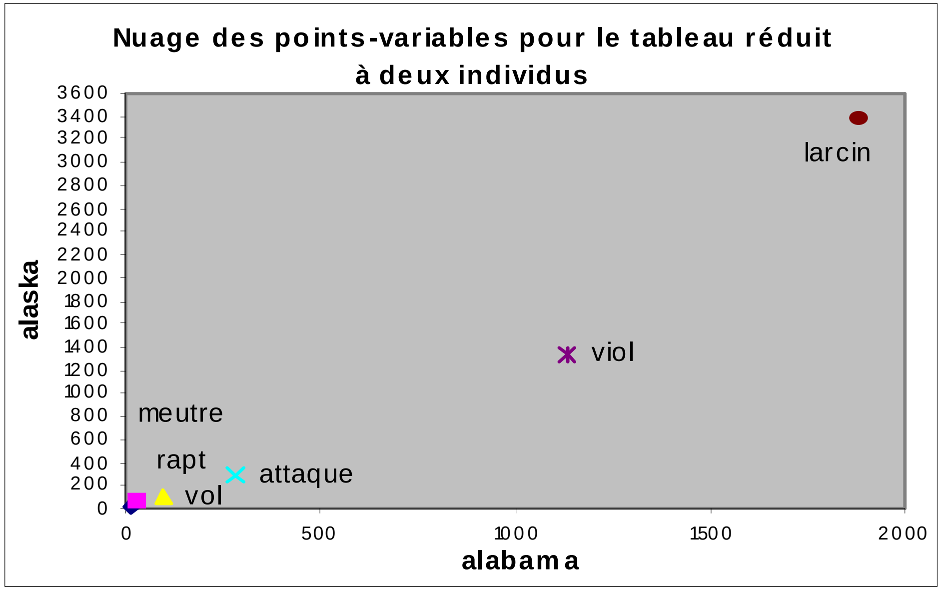


Le taux de meurtre et le taux de rapt sont corrélés positivement, ce qui signifie que les états où il y a beaucoup de meurtres sont généralement des états où il y a beaucoup de rapt, et inversement.

C.2- Les données : Grandeurs associées au tableau de données

b- *Nuages de points*

✓ Les p variables forment un nuage de points dans le sous-espace de R^n défini par les individus, *appelé nuage des points-variables.*



on peut comparer par rapport à la première bissectrice les valeurs prises par les variables sur les différents individus afin d'identifier des individus proches en terme de valeurs prises par les variables.

Ainsi, l'Alaska se distingue par un nombre relativement important de larcins.

C.2- Les données : Grandeurs associées au tableau de données

c- *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.

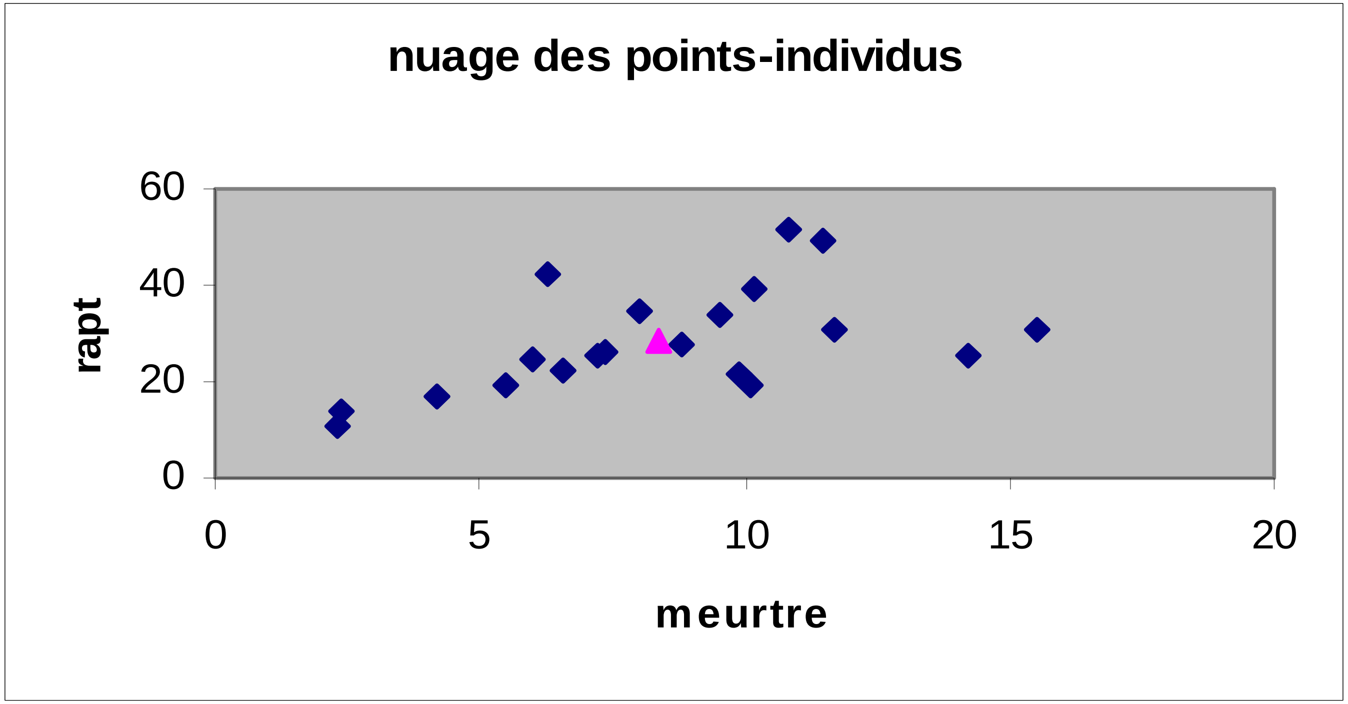
$$G = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \dots \\ \bar{x}_p \end{pmatrix} \qquad \boxed{\bar{x}_j = \sum_{i=1}^{n} p_i x_{ij}}$$

Au plus G est loin de l’origine, au moins le nuage est centré.

RQ : lorsque les poids sont égaux, G est le vecteur des moyennes.

C.2- Les données : Grandeurs associées au tableau de données

c- *Centre de gravité*



C.2- Les données : Grandeurs associées au tableau de données
c- *Centre de gravité*

✓ Centre de gravité du tableau des protéines

Variable	Statistiques Descriptives (proteines2)				
	N Actifs	Moyenne	Minimum	Maximum	Ecart-type
VR	25	9,82800	4,40000	18,00000	3,34708
VB	25	7,89600	1,40000	14,00000	3,69408
Oeufs	25	2,93600	0,50000	4,70000	1,11762
Lait	25	17,11200	4,90000	33,70000	7,10542
Poisson	25	4,28400	0,20000	14,20000	3,40253
Céréales	25	32,24800	18,60000	56,70000	10,97479
Amidon	25	4,27600	0,60000	6,50000	1,63408
Noix	25	3,07200	0,70000	7,80000	1,98568
F-L	25	4,13600	1,40000	7,90000	1,80390

Statist.exe

C.2- Les données : Grandeurs associées au tableau de données

d- *Matrice de variance-covariance associée à X*

$$V = \begin{bmatrix} Var(X_1) & Cov(X_1,X_2) & \ldots\ldots\ldots & Cov(X_1,X_j) & \ldots\ldots\ldots & Cov(X_1,X_p) \\ Cov(X_1,X_2) & Var(X_2) & \ldots\ldots\ldots & Cov(X_2,X_j) & \ldots\ldots\ldots & Cov(X_2,X_p) \\ \ldots\ldots\ldots & \ldots\ldots\ldots & \ldots\ldots\ldots & \ldots\ldots\ldots & \ldots\ldots\ldots & \ldots\ldots\ldots \\ Cov(X_1,X_j) & Cov(X_2,X_j) & \ldots\ldots\ldots & Var(X_j) & \ldots\ldots\ldots & Cov(X_j,X_p) \\ \ldots\ldots\ldots & \ldots\ldots\ldots & \ldots\ldots\ldots & \ldots\ldots\ldots & \ldots\ldots\ldots & \ldots\ldots\ldots \\ Cov(X_1,X_p) & Cov(X_2,X_p) & \ldots\ldots\ldots & Cov(X_j,X_p) & \ldots\ldots\ldots & Var(X_p) \end{bmatrix}$$

$$\boxed{V = X_c^{'} P X_c}$$

$$\boxed{\operatorname{cov}(X_j,X_l) = \sum_{i=1}^{n} p_i (x_{ij} - \overline{x}_j)(x_{il} - \overline{x}_l) = X_j^{c}{}' P X_l^{c}}$$

$$Var(X_j) = \operatorname{cov}(X_j,X_j); \sigma(X_j) = \sqrt{Var(X_j)}$$

Statist.exe

C.2- Les données : Grandeurs associées au tableau de données

e- *Inertie*

- ✓ On peut définir une distance ou éloignement entre individus :

$$d^2(e_i, e_k) = \|e_i - e_k\|^2 = \sum_{j=1}^{p} (x_{ij} - x_{kj})^2 = (e_i - e_k)'(e_i - e_k)$$

- ✓ Application : Eloignement d'un point du nuage par rapport au centre de gravité :

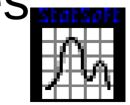
$$d^2(e_i, G) = \sum_{j=1}^{p} (x_{ij} - \bar{x}_j)^2$$

C.2- Les données : Grandeurs associées au tableau de données
e- *Inertie*

- ✓ *Inertie du nuage de points par rapport à son centre de gravité* = somme pondérée des éloignements au centre de gravité

$$\boxed{I = \sum_{i=1}^{n} p_i d^2(e_i, G) = \sum_{j=1}^{p} Var(X_j) = Tr(V)}$$

- ✓ I caractérise la *dispersion* ou la *forme* du nuage par rapport à son centre. : au plus *I* est élevée, au plus le nuage est dispersé autour de son centre de gravité.
- ✓ Une inertie nulle signifie que tous les individus sont identiques
- ✓ Lorsque les variables sont centrées et réduites I=p
- ✓ L'inertie mesure la quantité d'information contenue dans X


Statist.exe

C.2- Les données : Grandeurs associées au tableau de données
e- *Inertie*

I=218,47

Variable	Covariances (proteines2)								
	VR	VB	Oeufs	Lait	Poisson	Céréales	Amidon	Noix	F-L
VR	11,2029	1,89178	2,19062	11,9609	0,69422	-18,362	0,7407	-2,3225	-0,4481
VB	1,89178	13,6462	2,5614	7,38838	-2,9413	-16,776	1,89407	-4,6576	-0,4086
Oeufs	2,19062	2,5614	1,24907	4,57038	0,24935	-8,7385	0,8259	-1,2423	-0,0918
Lait	11,9609	7,38838	4,57038	50,4869	3,33353	-46,222	2,58238	-8,763	-5,2342
Poisson	0,69422	-2,9413	0,24935	3,33353	11,5772	-19,576	2,24543	-0,9942	1,63352
Céréales	-18,362	-16,776	-8,7385	-46,222	-19,576	120,446	-9,5634	14,1868	0,92153
Amidon	0,7407	1,89407	0,8259	2,58238	2,24543	-9,5634	2,67023	-1,539	0,24882
Noix	-2,3225	-4,6576	-1,2423	-8,763	-0,9942	14,1868	-1,539	3,94293	1,34313
F-L	-0,4481	-0,4086	-0,0918	-5,2342	1,63352	0,92153	0,24882	1,34313	3,25407

C.3- Les données : Transformations du tableau

a- Tableau (matrice) centré associé à X

Centrage : permet de ramener toutes les colonnes de X a la même origine, zero:

$$x_{ij} \rightarrow x_{ij} - \bar{x}_j$$

Matrice centrée :

$$\boxed{X_c = X - EG'}$$

$$X_c = \begin{bmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \dots & x_{1j} - \bar{x}_j & \dots & x_{1p} - \bar{x}_p \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & \dots & x_{2j} - \bar{x}_j & \dots & x_{2p} - \bar{x}_p \\ \dots & \dots & \dots & \dots & \dots & \dots \\ x_{i1} - \bar{x}_1 & x_{i2} - \bar{x}_2 & \dots & x_{ij} - \bar{x}_j & \dots & x_{ip} - \bar{x}_p \\ \dots & \dots & \dots & \dots & \dots & \dots \\ x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & \dots & x_{nj} - \bar{x}_j & \dots & x_{np} - \bar{x}_p \end{bmatrix}$$

C.3- Les données : Transformations du tableau

b- Tableau centré-réduit associé à X

Réduction = ramener toutes les variables à une même origine 0 et un même écart-type 1.

Centrage + réduction = $x_{ij} \rightarrow \dfrac{x_{ij} - \bar{x}_j}{\sigma(X_j)}$

$$\boxed{X_r = X_c D_s^{-1}}$$

$$D_s = \begin{bmatrix} \sigma(X_1) & 0 & 0 & 0 & 0 \\ 0 & \ldots. & 0 & 0 & 0 \\ 0 & 0 & \sigma(X_j) & 0 & 0 \\ 0 & 0 & 0 & \ldots. & 0 \\ 0 & 0 & 0 & 0 & \sigma(X_p) \end{bmatrix}$$

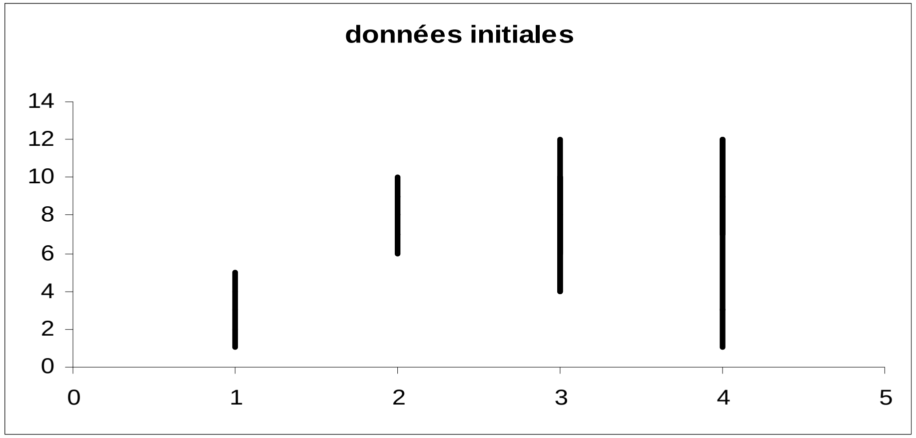
C.3- Les données : Transformations du tableau

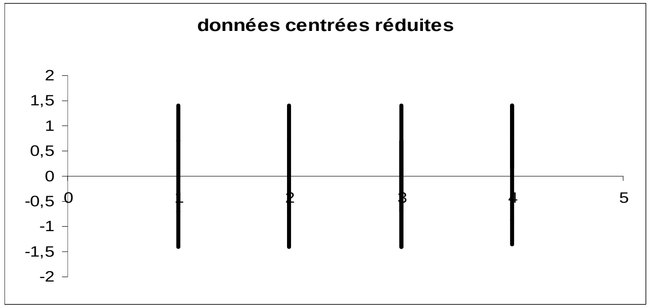
b- Tableau centré-réduit associé à X

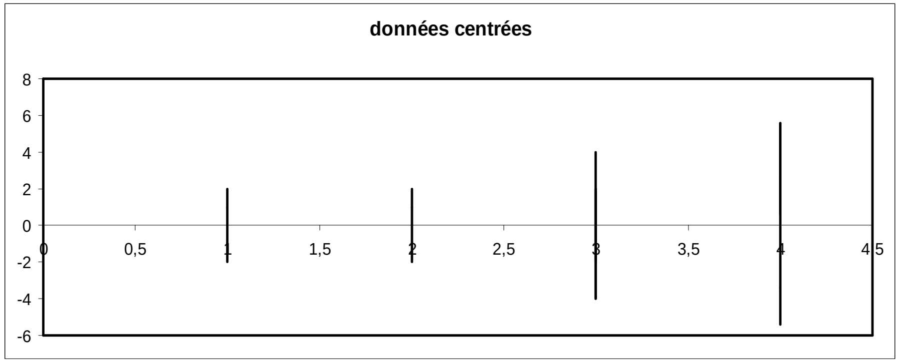
$$X_r = \begin{bmatrix} \frac{x_{11}-\bar{x}_1}{\sigma(X_1)} & \frac{x_{12}-\bar{x}_2}{\sigma(X_2)} & \cdots & \frac{x_{1j}-\bar{x}_j}{\sigma(X_j)} & \cdots & \frac{x_{1p}-\bar{x}_p}{\sigma(X_p)} \\ \frac{x_{21}-\bar{x}_1}{\sigma(X_1)} & \frac{x_{22}-\bar{x}_2}{\sigma(X_2)} & \cdots & \frac{x_{2j}-\bar{x}_j}{\sigma(X_j)} & \cdots & \frac{x_{2p}-\bar{x}_p}{\sigma(X_p)} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \frac{x_{i1}-\bar{x}_1}{\sigma(X_1)} & \frac{x_{i2}-\bar{x}_2}{\sigma(X_2)} & \cdots & \frac{x_{ij}-\bar{x}_j}{\sigma(X_j)} & \cdots & \frac{x_{ip}-\bar{x}_p}{\sigma(X_p)} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \frac{x_{n1}-\bar{x}_1}{\sigma(X_1)} & \frac{x_{n2}-\bar{x}_2}{\sigma(X_2)} & \cdots & \frac{x_{nj}-\bar{x}_j}{\sigma(X_j)} & \cdots & \frac{x_{np}-\bar{x}_p}{\sigma(X_p)} \end{bmatrix}$$

C.3- Les données : Transformations du tableau

b- Tableau centré-réduit associé à X







C.3- Les données : Transformations du tableau

d- Matrice de corrélation associée à X

- ✓ Le coefficient de corrélation linéaire entre deux variables quantitatives permet de mesurer le lien linéaire entre ces deux variables:

$$r(X_j, X_k) = \frac{Cov(X_j, X_k)}{\sigma(X_j)\sigma(X_k)}$$

$$\boxed{r(X_j, X_k) = \overset{r}{X}_j{}' P \overset{r}{X}_k}$$

$\boxed{-1 \leq r(X_j, X_k) \leq 1}$, d'autant plus grand en valeur absolue que le lien linéaire est grand. Nul si absence de lien linéaire.

C.3- Les données : Transformations du tableau

d- Matrice de corrélation associée à X

$$R = \begin{bmatrix} 1 & r(X_1, X_2) & \dots\dots\dots & r(X_1, X_j) & \dots\dots\dots & r(X_1, X_p) \\ r(X_1, X_2) & 1 & \dots\dots\dots & r(X_2, X_j) & \dots\dots\dots & r(X_2, X_p) \\ \dots\dots\dots & \dots\dots\dots & \dots\dots\dots & \dots\dots\dots & \dots\dots\dots & \dots\dots\dots \\ r(X_1, X_j) & r(X_2, X_j) & \dots\dots\dots & 1 & \dots\dots\dots & r(X_j, X_p) \\ \dots\dots\dots & \dots\dots\dots & \dots\dots\dots & \dots\dots\dots & \dots\dots\dots & \dots\dots\dots \\ r(X_1, X_p) & r(X_2, X_p) & \dots\dots\dots & r(X_j, X_p) & \dots\dots\dots & 1 \end{bmatrix}$$

$$\boxed{R = X_r' P X_r = D_S^{-1} V D_S^{-1}}$$

Variable	Corrélations (proteines2)								
	VR	VB	Oeufs	Lait	Poisson	Céréales	Amidon	Noix	F-L
VR	1,000000	0,153003	0,585609	0,502931	0,060957	-0,499877	0,135426	-0,349449	-0,074221
VB	0,153003	1,000000	0,620409	0,281484	-0,234009	-0,413797	0,313772	-0,634962	-0,061317
Oeufs	0,585609	0,620409	1,000000	0,575533	0,065571	-0,712437	0,452231	-0,559781	-0,045518
Lait	0,502931	0,281484	0,575533	1,000000	0,137884	-0,592737	0,222411	-0,621087	-0,408364
Poisson	0,060957	-0,234009	0,065571	0,137884	1,000000	-0,524231	0,403853	-0,147153	0,266139
Céréales	-0,499877	-0,413797	-0,712437	-0,592737	-0,524231	1,000000	-0,533262	0,650997	0,046548
Amidon	0,135426	0,313772	0,452231	0,222411	0,403853	-0,533262	1,000000	-0,474312	0,084410
Noix	-0,349449	-0,634962	-0,559781	-0,621087	-0,147153	0,650997	-0,474312	1,000000	0,374970
F-L	-0,074221	-0,061317	-0,045518	-0,408364	0,266139	0,046548	0,084410	0,374970	1,000000

C.4- Les données : Ecriture matricielles importantes

- Le carré de la P-norme d'une variable centrée Xj est sa variance

$$\left\| X_j \right\|_P^2 = X'_j P X_j = \sigma^2(X_j)$$

- Le carré de la P-norme d'une variable centrée réduite Xj est égal à 1
- Le P-produit scalaire entre deux variables centrées est leur covariance

$$\left\langle X_j, X_k \right\rangle_P = X'_j P X_k = Cov(X_j, X_k)$$

- Le P-produit scalaire entre deux variables centrées réduites est leur coefficient de corrélation

$$X'_j P X_k = r(X_j, X_k)$$

Ch2 : Analyse en Composantes Principales (ACP)

A- Objectifs

B- construction d’un espace factoriel

C- Les étapes d’une ACP

D- Interprétation

E- Limites

F- Exemple

A- Objectifs

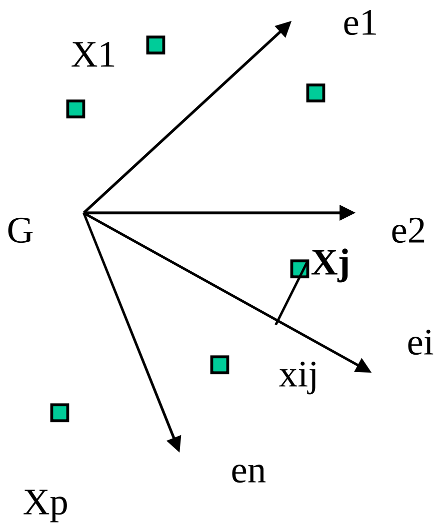
On dispose d'un tableau de données X. Ce tableau définit deux nuages de points :

- ✓ Nuage de points-variables = coordonnées des vecteurs variables tracées dans le repère dont les axes représentent les individus (espace de dimension n)

- ✓ Nuage de points-individus = coordonnées des vecteurs individus tracées dans le repère dont les axes représentent les variables (espace de dimension p)

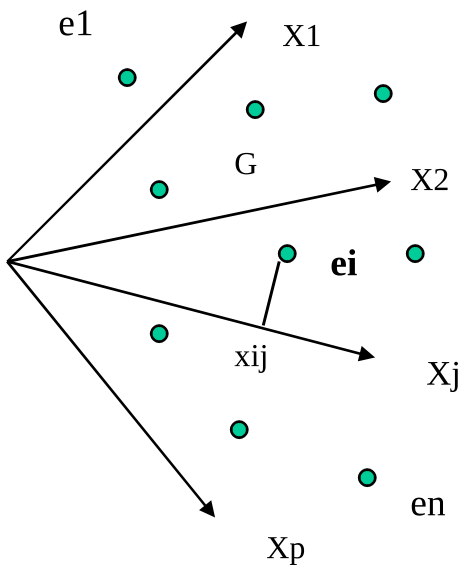
A- Objectifs

$$X_j = (x_{1j}, \ldots, x_{ij}, \ldots, x_{nj})'$$



Le nuage des points variables représenté dans l'espace de dim n défini par les individus

A- Objectifs



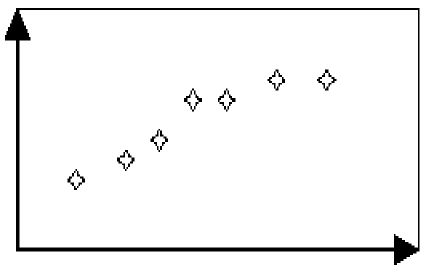
$$e_i = (x_{i1}, \ldots x_{ij}, \ldots x_{ip})'$$

Le nuage des points individus représenté dans l'espace de dim p défini par les variables

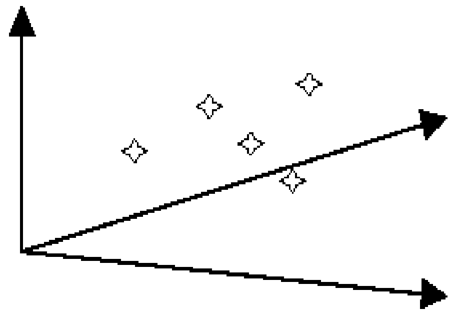
A - Objectifs

- Difficulté à mettre en évidence les relations globales existant entre les variables dès que $p>3$, car impossibles à visualiser.

Lorsqu'il n'y a que deux dimensions (largeur et longueur par exemple), il est facile de représenter les données sur un plan :



Avec trois dimensions (largeur, hauteur et profondeur par ex.), c'est déjà plus difficile :



A - Objectifs

- On veut **condenser** l'information du tableau de manière à retirer les relations vraiment caractéristiques (proximités entre variables et individus), ceci **en limitant la perte d'information**.

$$\Leftrightarrow$$

Déterminer un sous-espace de **dimension** q<p (q nouveaux axes), sur lequel **projeter** les nuages de points relatifs au tableau de données. Ce sous-espace doit être:

- **« compréhensible » par l'œil**: q faible, de préférence q=1 ou q=2,

- **le moins déformant possible**

Ce sous-espace est appelé ***espace factoriel*** du nuage.

A - Objectifs

- Définir un nouveau sous-espace de dimension q (q nouvelles directions de l'espace) revient à

✓ Définir q nouvelles variables comme axes du repère du nuage de points-individus : on les appelle axes factoriels ou composantes principales

✓ Définir q nouveaux individus comme axes du repère du nuage de points-variables

B- construction d’un espace factoriel

- Un espace factoriel est défini par un repère de dimension q, dont les axes sont construits de la façon suivante (ex: nuage de points-individus):

✓ On effectue un changement de repère, passant du repère défini par les p variables à un repère de dimension p le moins déformant possible pour le nuage. Il sera défini par p nouveaux axes, appelés *axes factoriels*.

✓ On retient ensuite les q premiers axes du nouveau repère, ce qui nous donnera l’espace factoriel de dimension q.

B- construction d’un espace factoriel

- Les p axes factoriels sont définis séquentiellement : On détermine l’axe (premier axe factoriel) sur lequel le nuage se déforme le moins possible en projection, On cherche un second axe, sur lequel le nuage se déforme le moins en projection, après le premier axe, tout en étant orthogonal au premier, On réitère jusqu’à l’obtention de p axes.

- Dans le second repère, les axes ne véhiculent pas la même information selon leur rang : leur capacité à « résumer » le nuage se détériore au fur et à mesure que l’on observe des axes de rang élevé.

B- construction d’un espace factoriel

- Interprétation en termes statistiques :

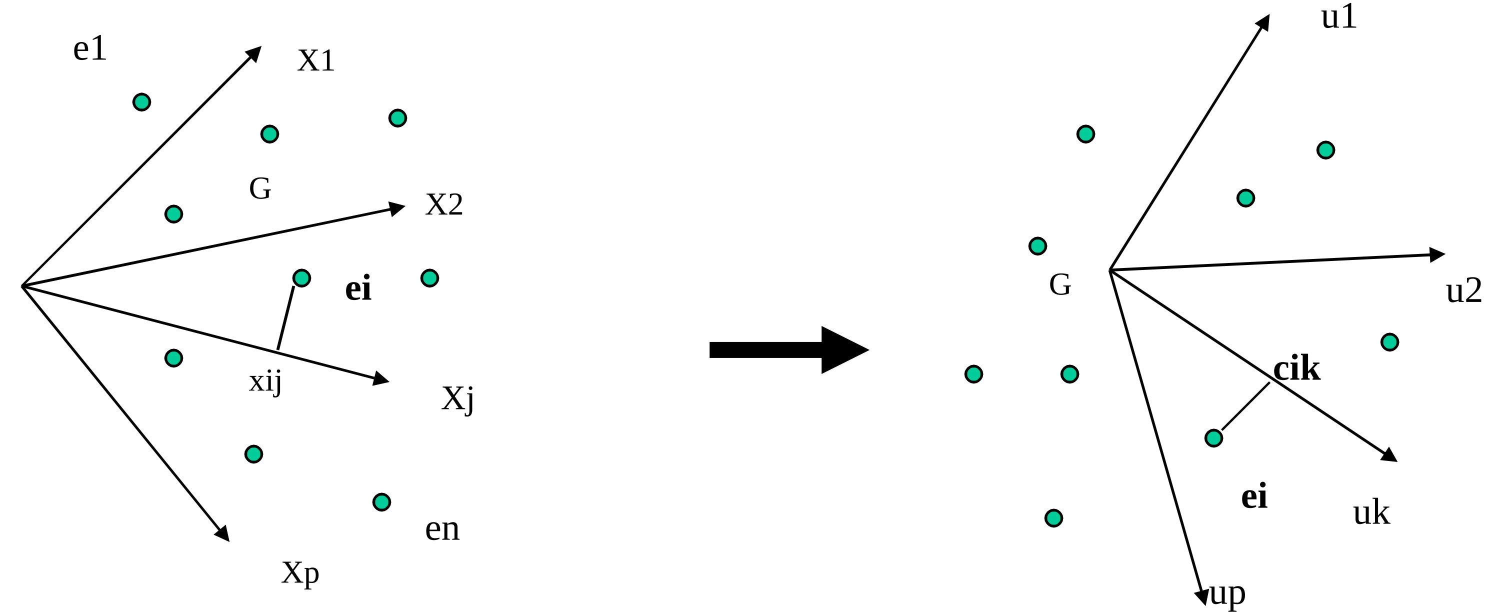
✓ Chaque axe factoriel représente une nouvelle variable, construite comme combinaison linéaire des variables (axes) de départ, appelée *composante principale.* La coordonnée d’un individu donné sur cet axe correspond à la valeur de la composante principale prise par cet individu.

✓ Les composantes principales sont construites de manière à restituer la majeure partie de l’information du tableau . Elles déforment le moins possible l’information)

✓ Les composantes principales sont non corrélées (les axes sont orthogonaux)

B- construction d’un espace factoriel

$$\mathbf{e_i = (c_{i1}, ..., c_{ik}, ...c_{ip})'}$$

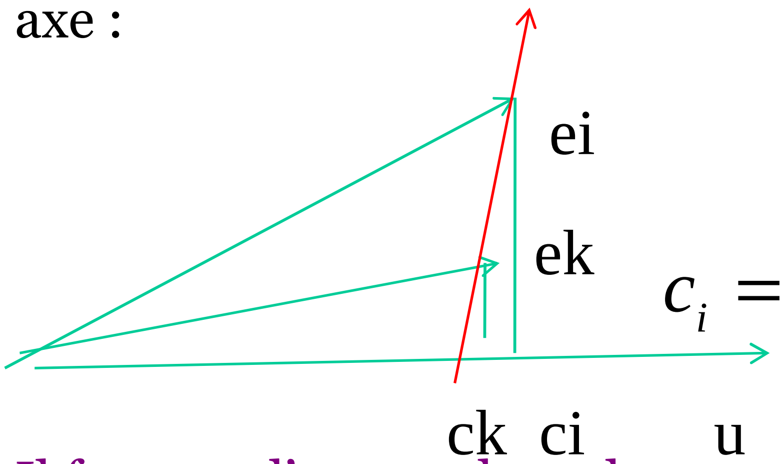


$$e_i = (x_{i1}, \ldots x_{ij}, \ldots x_{ip})'$$

B- construction d'un espace factoriel

- le rôle de l'ACP est de trouver le sous-espace sur lequel projeter le nuage tel que la déformation subie soit la plus faible possible et permette ainsi de récupérer les liens les plus significatifs contenus dans le tableau.
- Comment obtenir une déformation minimale ? Projection sur un axe :

$$d(c_i, c_k) \le d(e_i, e_k)$$

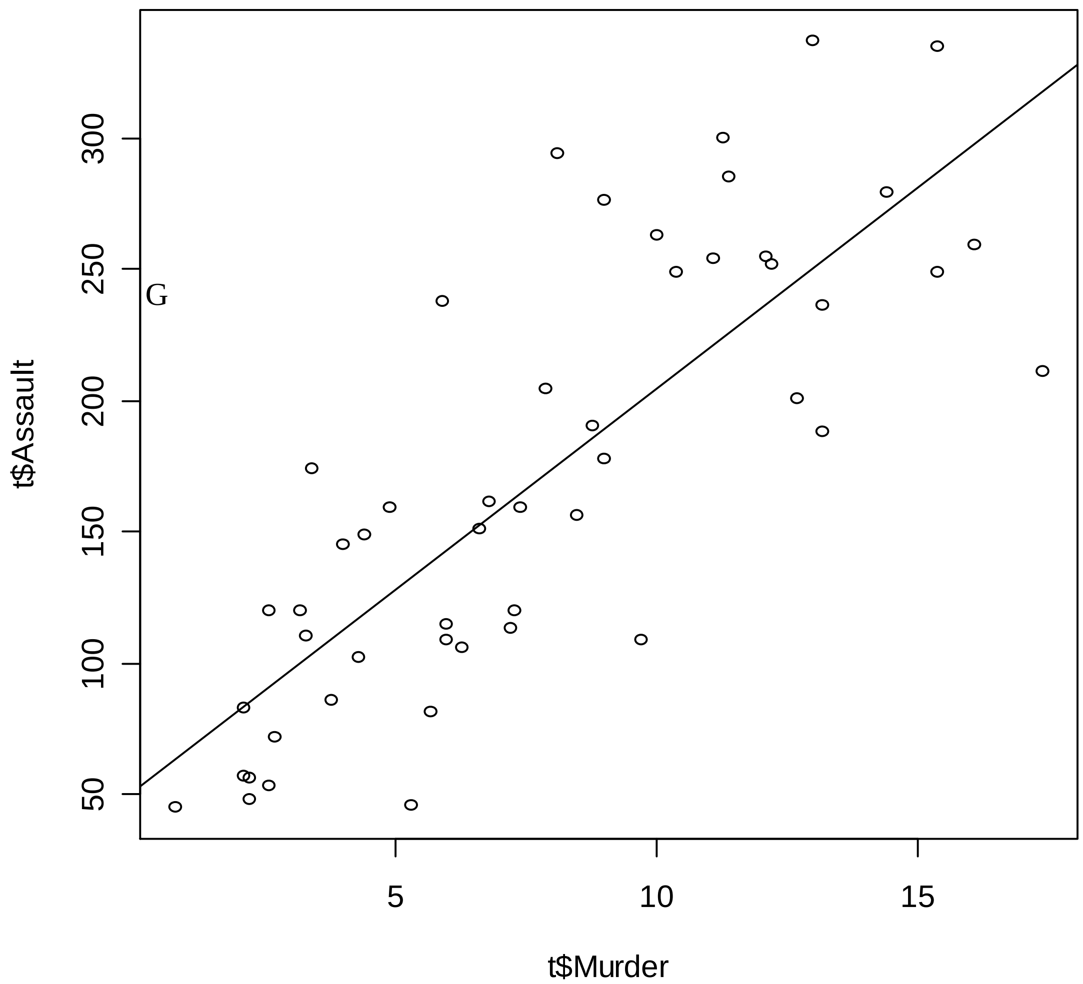


$$c_i = \langle e_i, u \rangle = x_{i1}u_1 + \ldots + x_{ip}u_p$$

- Il faut que l'axe sur lequel on projette permette la dispersion maximale

$$d(c_i, c_k) \simeq d(e_i, e_k)$$

B- construction d'un espace factoriel



B- construction d’un espace factoriel

- Conclusion : le meilleur axe (premier axe factoriel) sera celui sur lequel le nuage de points projeté est de dispersion, **d’inertie maximale**. La première composantes principale sera une CL des variables de départ de dispersion (de variance) maximale.

C- Les étapes d'une ACP

- ***Choix du tableau X***

- *Analyse directe* : Construction de l'espace factoriel du nuage de points-individus associé au tableau . On garde pour l'instant les p axes factoriels

- *Analyse duale :* : Construction de l'espace factoriel du nuage de points-variables : elle est *d*éduite de la première

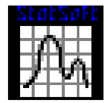
- **Interprétation de ces analyses : choix du nombre d'axes q à retenir, construction des nuages de points projetés sur ces axes, interprétation des axes principaux et étude des proximités entre points.**

- **Synthèse des résultats, construction éventuelle du tableau C réduit (tableau des composantes principales) et visualisation des nuages de points associés.**

C-1 Choix du tableau X

- Travailler sur le tableau brut (centré par défaut) ou centré réduit?

- X centré non réduit : L'importance que prendront le variables dans le calcul des composantes principales est fonction de leur ordre de grandeur; une variable d'écart-type important aura plus de poids qu'une variable d'écart-type faible. Des variables de fort écart-type construiront les premières composantes principales : les calculs ne sont pas faux, et conduisent aux même interprétation mais la lecture des résultats risque d'être brouillée.
- $\Rightarrow$ On commence souvent par centrer et réduire X



Statist.exe

C-2 Analyse directe
a- Origine du repère

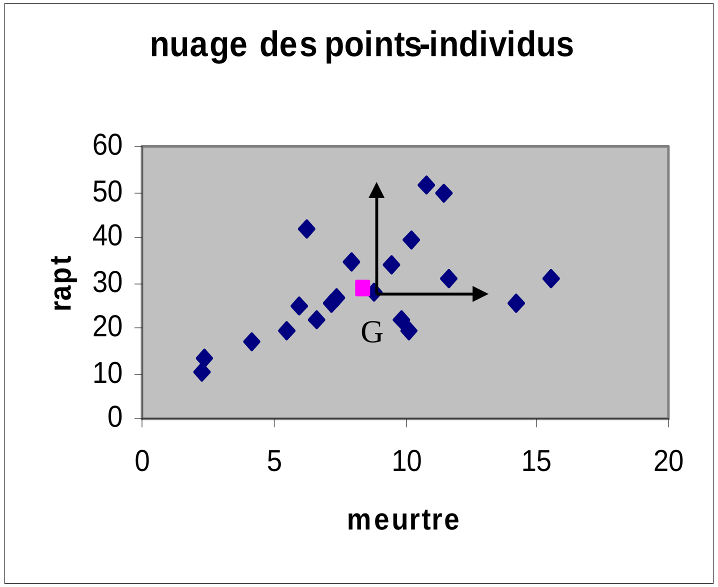
Recherche des axes factoriels du nuage de points-individus

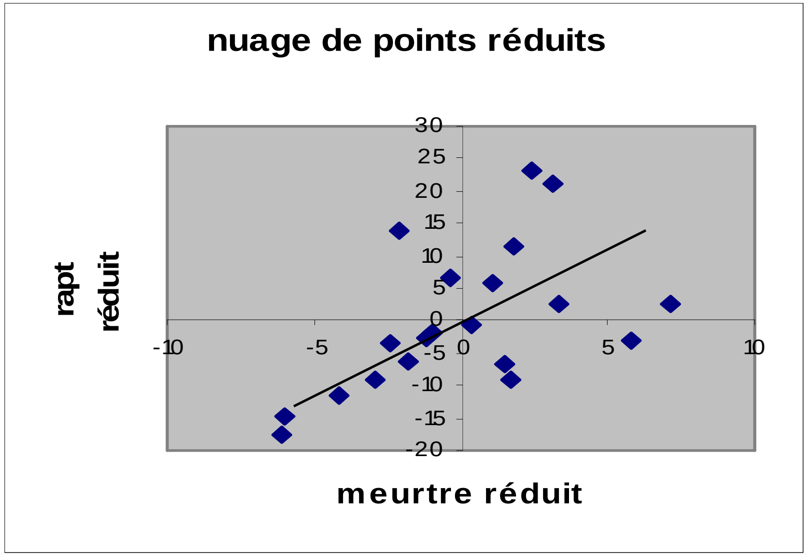
- Détermination de l'origine : On MQ les axes « les plus informatifs » passent forcement par le centre d'inertie du nuage de points. $\Rightarrow$ Le nouveau repère aura pour origine G. On travaille toujours sur le nuage centré.

- Un axe étant déterminé par un point et un vecteur directeur (une direction de l'espace), il suffit dès lors de rechercher les directions des p axes factoriels.

- Dès à présent, On note X le tableau centré, e_i ses vecteurs individus et X_j ses vecteurs variables.

C-2 Analyse directe
a- Origine du repère





C-2 Analyse directe
b- Recherche du premier axe factoriel

- ✓ Il passe par G
- ✓ Vecteur directeur : u_1 normé t.q. le nuage de points projeté sur u_1 est d'inertie maximale

(P) $\boxed{\begin{array}{l} I_1 \text{ est maximale} \\ \text{sous la contrainte : } \|u_1\| = 1 \end{array}}$

Où I_1 est l'inertie du nuage projeté $I_1 \leq I$

C-2 Analyse directe
b- Recherche du premier axe factoriel

- **Calcul de I1 :**

- Soit $C_1 = (c_{11}, \ldots, c_{i1}, \ldots, c_{n1})$ le vecteurs des coordonnées de la projection orthogonale des individus du tableau X sur l'axe $u_1 = (u_{11}, \ldots, u_{j1}, \ldots, u_{p1})$

$$c_{i1} = \langle e_i, u_1 \rangle = e_i' u_1 = x_{i1} u_{11} + \ldots + x_{ip} u_{1p} \qquad C_1 = X u_1$$

$$I_1 = \sum_{i=1}^{n} p_i d^2(c_{i1}, G) = \sum_{i=1}^{n} p_i c_{i1}^2 = C'_1 P C_1 = Var(C_1) = u_1' X' P X u_1 = u_1' S u_1$$

- S=X'PX= matrice d'inertie
  - Lorsque X est centré S=V
  - Lorsque X est centré-réduit, S=R

C-2 Analyse directe
b- Recherche du premier axe factoriel

- On a $$I_1 = u_1' S u_1$$

(P) $$\boxed{\begin{array}{l} u_1' S u_1 \text{ est maximale} \\ \text{sous la contrainte : } \|u_1\| = 1 \end{array}}$$

Solution : u_1 est le vecteur propre unitaire de S associé à la plus grande valeur propre λ_1. Il vérifie : $$S u_1 = \lambda_1 u_1$$

- le vecteur des coordonnées des n points du nuage sur le premier axe est $C_1 = X u_1$
  C'est le vecteur des valeurs prises par la première composante principale sur les n individus.

C-2 Analyse directe
b- Recherche du premier axe factoriel

✓ Propriétés du premier axe :

- Information véhiculée par l'axe :

$$I_1 = u'_1 S u_1 = \lambda_1 u_1' u_1 = \lambda_1$$

L'axe 1 restitue une information égale à λ_1

- La première composante principale est
  - ✓ centrée.
  - ✓ De variance $Var(C)_1 = C'_1 P C_1 = u'_1 S u_1 = \lambda_1$

C-2 Analyse directe
c- Recherche des axes de rang supérieur

- Même méthode : le deuxième axe factoriel est l'axe associé à la *valeur propre de rang 2* (2° plus grande valeur propre de *S*), que l'on pourra choisir *orthogonal au premier axe* (car S est une matrice orthogonale), et ainsi de suite, jusqu'au p° axe.

- L'inertie de l'axe k (information véhiculée par l'axe) est $I_k = \lambda_k$

- la k° composante $C_k = Xu_k$ est centrée, de variance $Var(C_k) = I_k = \lambda_k$, non corrélée avec les autres $Cov(C_k, C_{k'}) = 0$

C-2 Analyse directe
c- Recherche des axes de rang supérieur

Inertie d'un sous-espace factoriel (espace constitué de q<p premiers axes factoriels) :

- Soit IE(q) l'inertie du nuage de points sur le sous-espace factoriel de dimension q. On montre que

$$\boxed{I_{E(q)} = \sum_{k=1}^{q} I_k = \sum_{k=1}^{q} \lambda_k}$$

- Dans une ACP normée, $\boxed{I = I_{E(p)} = p}$

C-2 Analyse directe
c- Conclusion

L'analyse directe passe par les étapes suivantes :

- Diagonalisation de S (S est définie positive d'ordre p, elle n'a pas de valeurs propres nulles et il y a donc p directions).

- Classement des valeurs propres par ordre décroissant (elles sont toutes <=1). Les vecteurs propres associés déterminent les axes du nouveau repère.

- Les valeurs prises par la projection des individus sur ces axes sont les valeurs des composantes principales (les nouvelles variables créées, CL des variables de départ de variance max)

C-3 Analyse duale

On peut montrer qu'il n'y a pas lieu de réitérer l'ensemble des calculs faits précédemment et que :

- les axes factoriels dans l'analyse duale se déduisent des axes factoriels trouvés lors de l'analyse directe (par symétrie, ce sont les vecteurs propres de XX'P). Il y en a seulement p d'informatifs

- l'inertie (représentant l'information restituée) est identique pour des axes de même rang dans les deux analyses.

C-3 Analyse duale
Relation entre les axes factoriels

- Pour des raisons de symétrie, les axes factoriels du nuage de points-variables passent par l'origine (il n'y a donc pas lieu de centrer) et ont pour vecteurs directeurs les vecteurs propres P-unitaires de la matrice *XX'P*.
- On montre que *XX'P* a p valeurs propres non nulles et n-p nulles donc seulement p axes sont informatifs. Les valeurs propres non nulles sont les mêmes que celles de R. Les valeurs propres non nulles et donc l'inertie sont identiques pour des axes de rang homologues.

$$I_k = \lambda_k$$

- Les vecteurs propres satisfont

$$u_k = \frac{X'Pv_k}{\sqrt{\lambda_k}} \qquad v_k = \frac{Xu_k}{\sqrt{\lambda_k}}$$

C-3 Analyse duale
Coordonnées des points-variables sur les axes

- ✓ Le vecteur de coordonnées $D_k = (d_{1k}, \ldots, d_{pk})'$ des variables sur le k° axe factoriel du nuage de points variables est donné par

$$D_k = \sqrt{\lambda_k}\, u_k = \frac{X' P C_k}{\sqrt{\lambda_k}} \qquad d_{jk} = \sqrt{\lambda_k}\, u_{jk} = \frac{X_j' P C_k}{\sqrt{\lambda_k}}$$

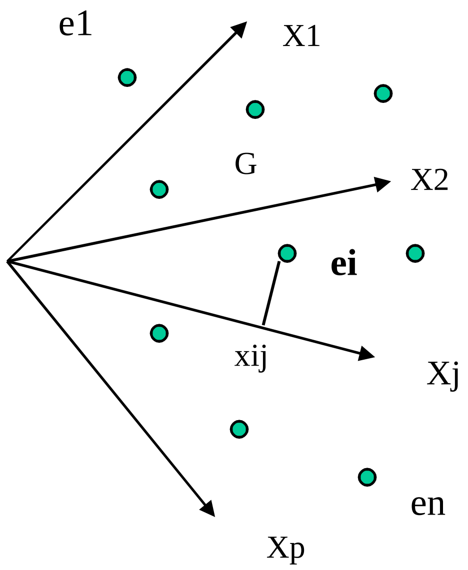
- ✓ Lorsque l'ACP est normée (X tableau centré réduit), la deuxième formule ci-dessus permet de montrer que :

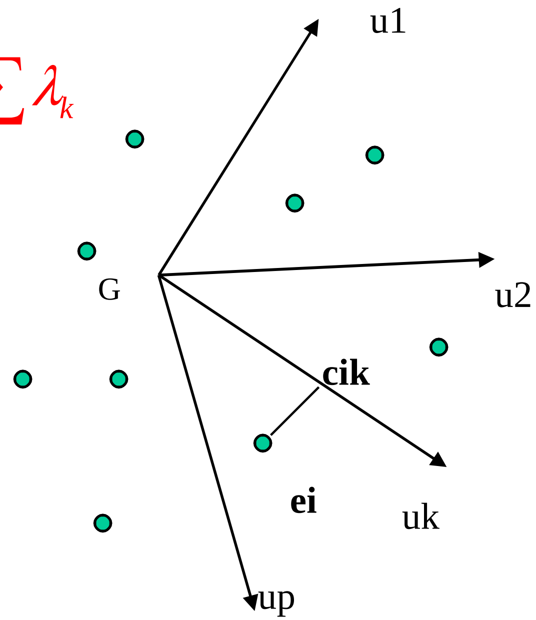
$$d_{jk} = r(X_j, C_k)$$

C-4 Résumé de la décomposition factorielle
Analyse directe

$$\mathbf{e_i = (c_{i1}, ..., c_{ik}, ...c_{ip})'}$$

$$I = \sum p_i d^2(e_i, G) = \sum \lambda_k$$





$$e_i = (x_{i1}, ...x_{ij}, ...x_{ip})'$$

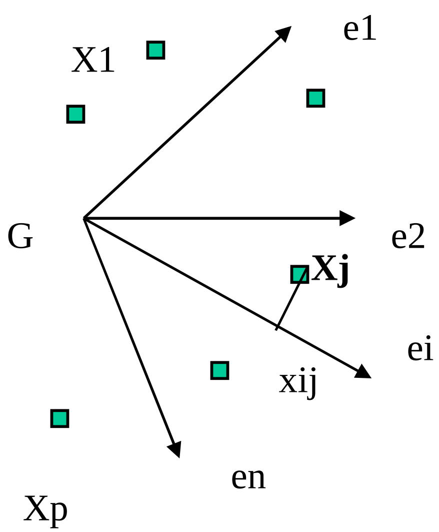
$$Ru_k = \lambda_k u_k, \quad \|u_k\| = 1, \lambda_1 \geq .. \geq \lambda_p$$

$$C_k = Xu_k, \quad C_k = (c_{1k}, .., c_{ik}, ..., c_{nk})'$$

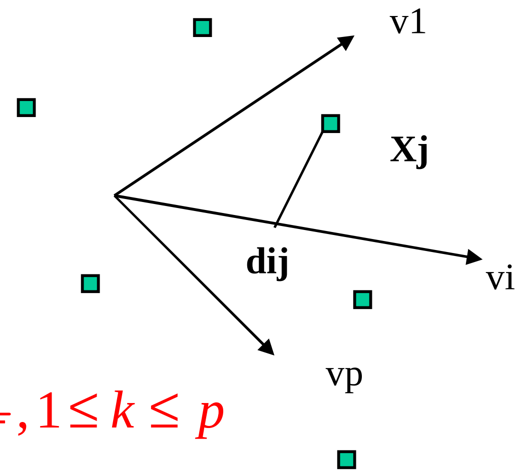
$$I_k = \lambda_k = Var(C_k); Cov(C_k, C_l) = 0$$

C-4 Résumé de la décomposition factorielle
Analyse duale

$$X_j = (x_{1j},...,x_{ij},..,x_{nj})'$$



$$X_j = (d_{1j},...,d_{ij},...d_{pj})$$



$$v_k = \frac{Xu_k}{\sqrt{\lambda_k}}, 1 \leq k \leq p$$

$$D_k = \sqrt{\lambda_k}\, u_k = \frac{X'PC_k}{\sqrt{\lambda_k}}, \; D_k = (d_{1k},..d_{pk})'$$

$$I_k = \lambda_k = \sum d_{jk}^{\;2}$$

C-5 Conclusion de la décomposition factorielle

L'ACP permet donc de construire de nouvelles variables (les composantes principales), $C_k = Xu_k$ combinaison linéaire des variables d'origine. On montre facilement qu'elles sont

- ✓ centrées (les variables d'origine le sont)
- ✓ non corrélées $Cov(C_k, C_l) = C_k' P C_l = 0$
- ✓ de variance maximale. $\|C_k\|_p^2 = Var(C_k) = \lambda_k$

Nous pouvons en sélectionner une partie pour construire le tableau C, résumant l'information contenue dans le tableau initial, et tenter de leur donner une signification.

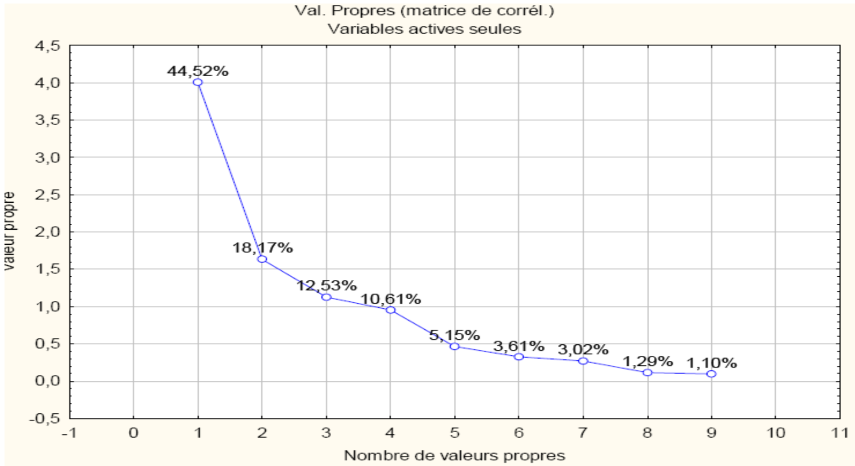
ACP normée avec Statistica

Statist.exe

- ✓ Valeurs propres de R
- ✓ Col1 : rang de l'axe
- Col2 : inertie de l'axe
- Col3 : proportion d'inertie expliquée par l'axe
- Col4: inertie expliquée par le sous-espace déterminé par les axes de rang inférieurs ou égaux,
- Col5 : % inertie cumulée

$$I = 9 = \sum \lambda_k$$

Valeur numéro	Val. Propres (matrice de corrél.) & stat. associées (proteines2) Variables actives seules Val Propre	% Total variance	Cumul Val Propre	Cumul %
1	4,006438	44,51597	4,006438	44,5160
2	1,634999	18,16666	5,641437	62,6826
3	1,127920	12,53244	6,769357	75,2151
4	0,954664	10,60738	7,724020	85,8224
5	0,463838	5,15376	8,187859	90,9762
6	0,325131	3,61257	8,512990	94,5888
7	0,271606	3,01785	8,784596	97,6066
8	0,116292	1,29213	8,900888	98,8988
9	0,099112	1,10124	9,000000	100,0000

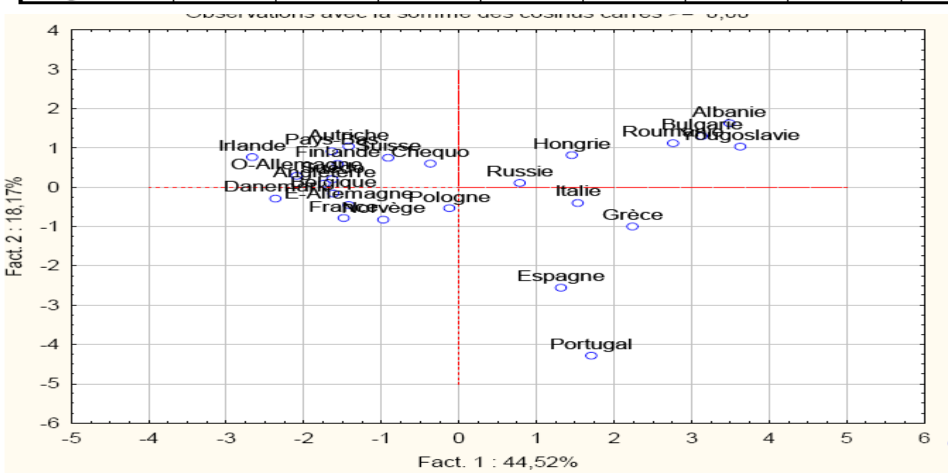


ACP normée avec Statistica

- ✓ Coordonnées des individus sur les axes factoriels du nuage de points-individus:

$$c_{ik} = e_i{}'u_k = x_{i1}u_{1k} + \ldots + x_{ip}u_{pk}$$

Individus	Coordonnées factorielles des ind., basées sur les corrélations (proteines2)							
	Fact. 1	Fact. 2	Fact. 3	Fact. 4	Fact. 5	Fact. 6	Fact. 7	Fact. 8
Albanie	3,48537	1,63048	1,76123	0,22966	0,02325	1,03426	0,471742	0,761551
Autriche	1,42267	1,04123	1,33780	0,16810	0,93345	0,21843	0,181154	0,251002
Belgique	1,62203	0,15950	0,21653	0,52073	0,75509	0,28981	0,195597	0,203312
Bulgarie	3,13408	1,30107	0,15129	0,21419	0,48475	0,69558	0,464782	0,808245
Chequo	0,37046	0,60267	1,19594	0,46398	0,25682	0,82309	0,314948	0,012298
Danemark	2,36527	0,28545	0,75226	0,96734	0,75243	0,17033	0,225816	0,621021
E-Allemagne	1,42221	0,45030	1,30254	1,13596	0,42294	0,64831	0,554783	0,163177
Finlande	1,56386	0,59600	2,04951	1,41531	0,03720	0,83420	0,726230	0,225917
France	1,48798	0,78537	0,00188	1,95746	0,25046	0,89895	0,946475	0,022220
Grèce	2,23970	1,00106	0,88260	1,79432	0,40498	1,14448	0,147391	0,305831
Hongrie	1,45744	0,81595	1,91417	0,21739	0,04140	0,53911	0,768102	0,145618
Irlande	2,66348	0,76371	0,01988	0,43473	1,01439	0,48233	0,028669	0,022999
Italie	1,53457	0,39899	0,12609	1,22246	0,80354	0,21409	0,149992	0,080406
Pays-Bas	1,64145	0,91199	0,76649	0,12615	0,76128	0,29752	0,062096	0,459926
Norvège	0,97470	0,82203	1,70408	1,13762	0,41487	0,05645	0,042788	0,107346
Pologne	0,12187	0,53174	1,47479	0,45822	0,02322	0,58830	1,260723	0,191596
Portugal	1,70585	4,28893	0,04363	0,89356	0,38529	0,69710	0,046500	0,205022
Roumanie	2,75681	1,11879	0,07008	0,61501	0,31710	0,13052	0,133079	0,026894
Espagne	1,31181	2,55352	0,51528	0,35920	0,51590	0,66929	0,597211	0,235328
Suède	1,63373	0,20738	1,28037	0,73410	0,81982	0,04408	0,541162	0,072218
Suisse	0,91232	0,75106	0,15425	1,17044	0,83096	0,09024	0,512291	0,529297
Angleterre	1,73537	0,09398	1,15268	1,73369	1,08394	0,09656	0,650969	0,239209
Russie	0,78260	0,11077	0,36968	0,92757	1,66956	0,18543	0,574102	0,052027
O-Allemagne	2,09384	0,29378	0,80398	0,10880	0,06836	0,20099	0,456777	0,356629
Yougoslavie	3,62301	1,03803	0,20605	0,82155	0,37769	0,35392	0,061291	0,193276



ACP normée avec Statistica

Statist.exe

- ✓ Coordonnées des variables sur les axes factoriels du nuage de points-variables:

$$d_{jk} = r(X_j, C_k)$$

- ✓ On les représente sur le cercle des corrélations

Variable	Coord. factorielles des var., basées sur les corrélations (proteines2)							
	Fact. 1	Fact. 2	Fact. 3	Fact. 4	Fact. 5	Fact. 6	Fact. 7	Fact. 8
VR	0,605706	0,071927	0,316040	0,631652	0,219409	0,262219	0,078348	0,006772
VB	0,621612	0,302857	0,662601	0,036144	0,204429	0,068999	0,010248	0,009506
Oeufs	0,854043	0,045183	0,192789	0,305983	0,053879	0,205985	0,231015	0,167507
Lait	0,756062	0,236028	0,409582	0,003242	0,136493	0,352635	0,240825	0,027766
Poisson	0,271518	0,827070	0,341205	0,211003	0,197527	0,077998	0,055448	0,153025
Céréales	0,876191	0,298551	0,101868	0,006062	0,162206	0,046049	0,211051	0,239733
Amidon	0,594974	0,451148	0,258048	0,328964	0,501240	0,084200	0,079609	0,039060
Noix	0,841345	0,183247	0,057762	0,322714	0,102524	0,254886	0,212248	0,062679
F-L	0,221017	0,685611	0,432839	0,451460	0,159038	0,067597	0,234510	0,031361