CH1 : Introduction à l'Analyse Des Données (ADD)

A- Introduction

B- Les données et leurs caractéristiques

C- Grandeurs associées aux données

A-1 Les méthodes

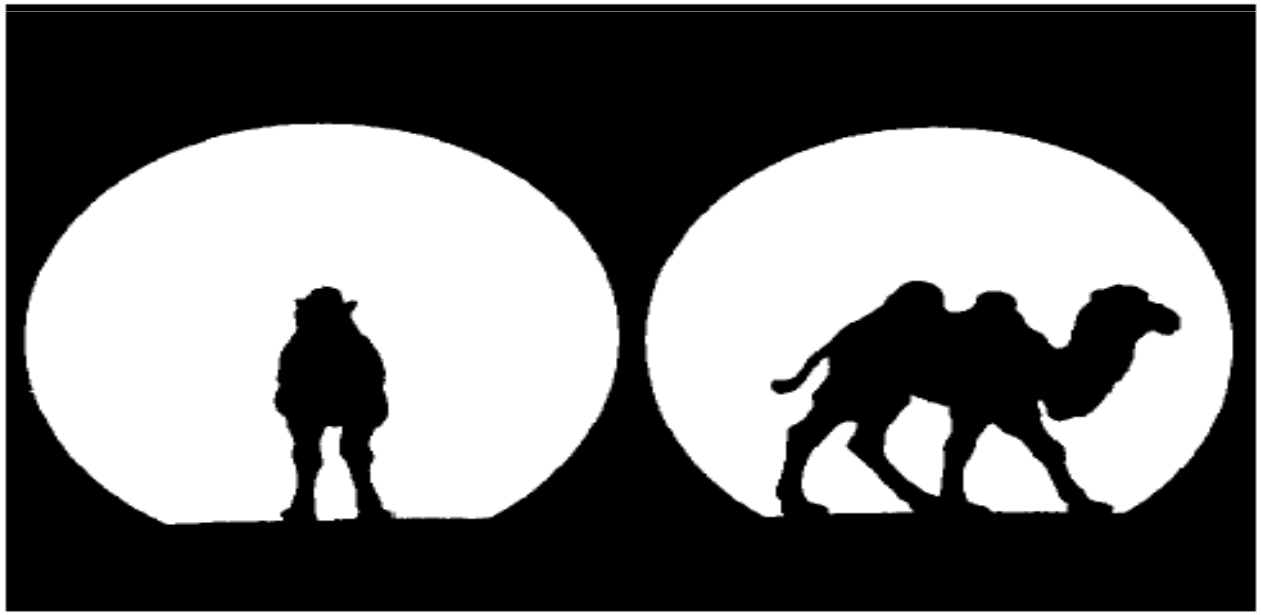
- ✓ Lors de toute étude statistique, il est nécessaire de *décrire* et *explorer* les données avant d'en tirer de quelconques lois ou modèles prédictifs.

- ✓ 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 -1 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 -1 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,..)

A-2 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

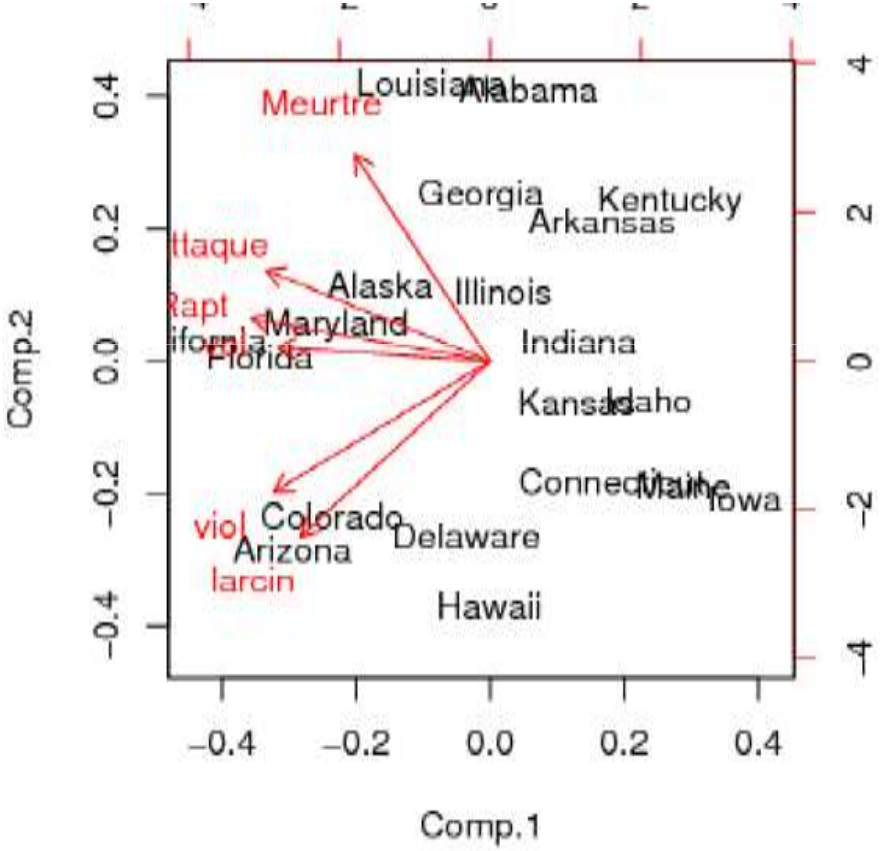
A-2 Exemples

- **ACP:**

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'opposent les états ayant un fort taux de délits mineurs (Colorado, Arizona) aux états concernés par des délits majeurs (Alabama, Louisiane).

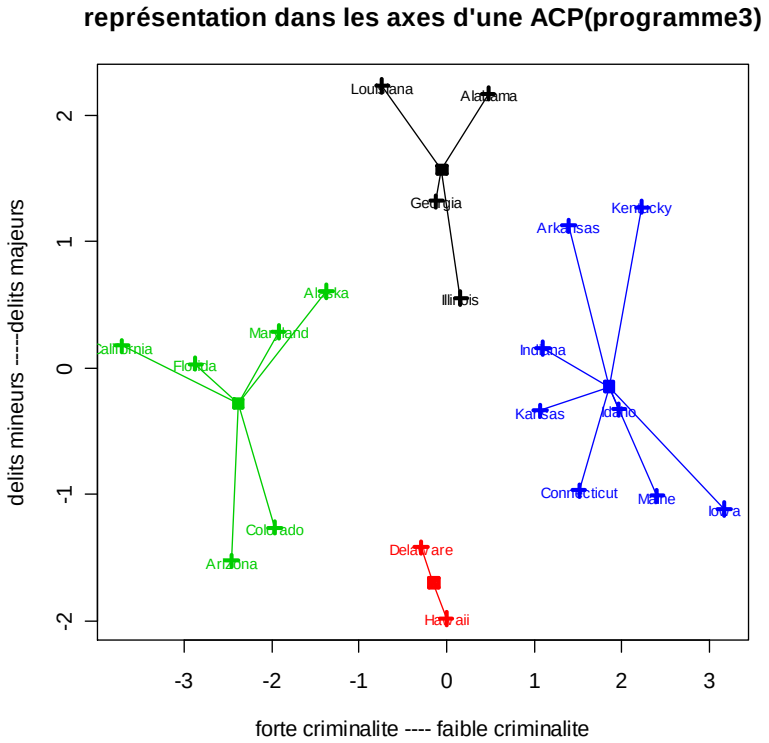


A-2 Exemples

- **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



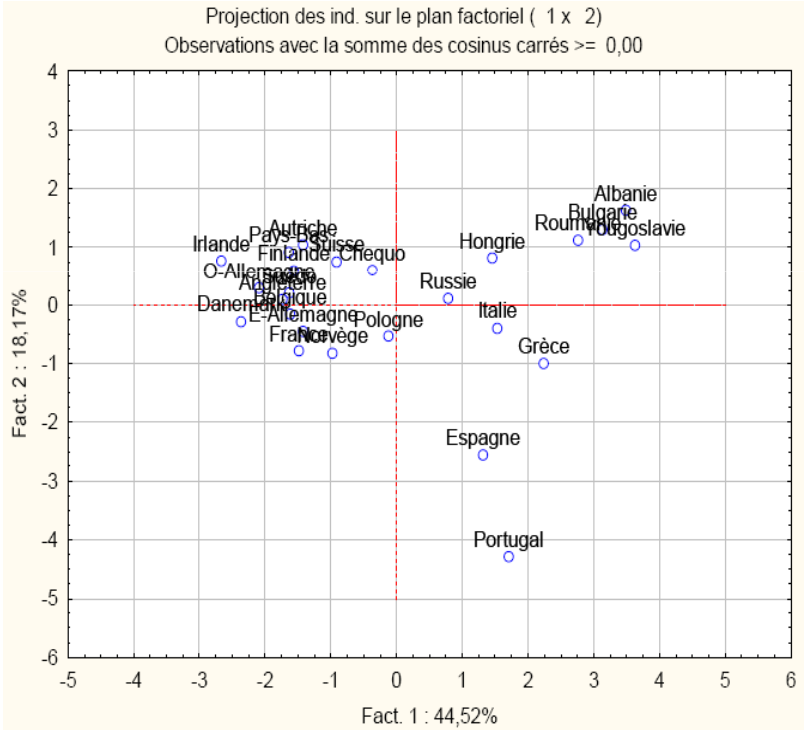
A-2- Exemples

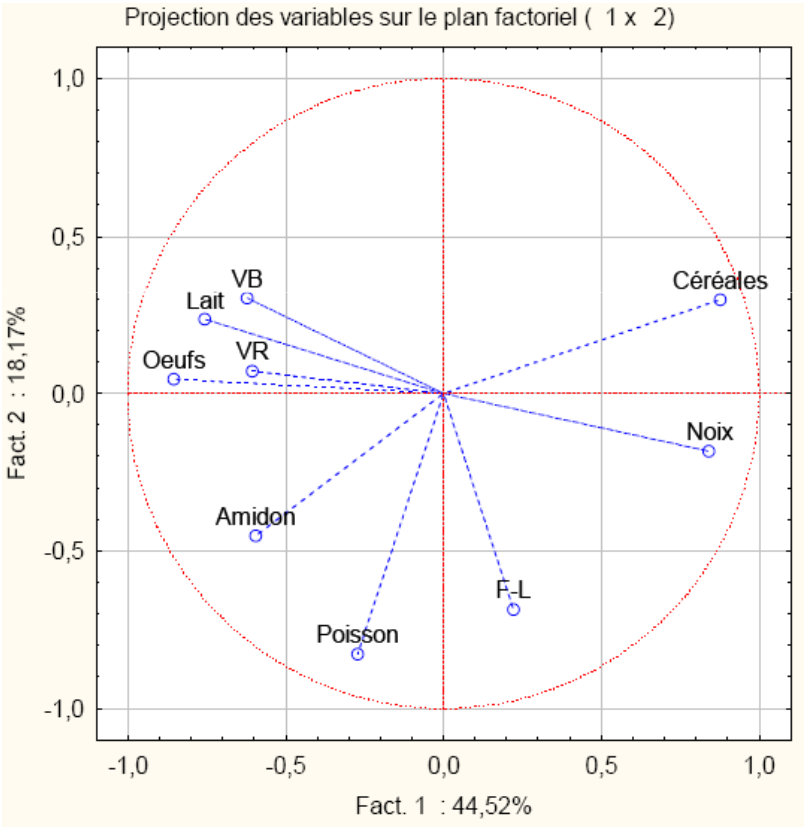
Les données mesurent la consommation de protéines dans 25 pays européens par rapport à 9 groupes d'aliments.

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

A-2 – Exemples (ACP sous statistica)

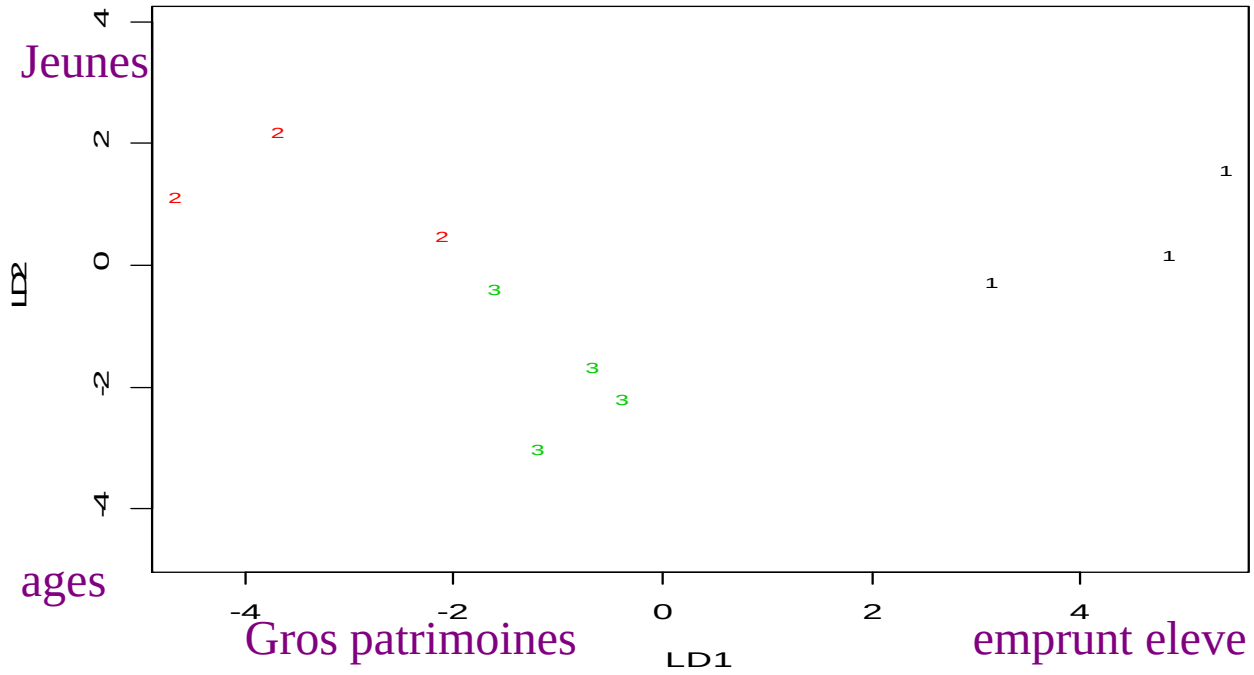




A-2 Exemples

- **Analyse discriminante**

3 groupes de personnes repérées par 4 variables : age revenu patrimoine emprunt



A-2 Exemples

Le groupe 1 est un groupe de gens assez jeunes à revenus plus faibles que la moyenne dont le patrimoine est nettement plus faible que dans les autres classes et le taux d'emprunt plus élevé que la moyenne

Le groupe 2 est caractérisé par des gens jeunes de revenus moyens, mais dont le patrimoine est très important et le taux d'emprunt très faible

Le groupe 3 est caractérisé par des gens plus agés de revenus confortables et de patrimoine assez important, ayant un taux d'emprunt plus élevé que dans les autres classes

B –1 Tableau individu*variables

- ✓ On observe p caractéristiques $X_1,.....X_p$ quantitatives sur n individus $e_1,....,e_i,....e_n$
- ✓ On note 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}

B –1 Tableau individu*variables

✓ Le tableau peut être mis sous forme matricielle

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

B –1 Tableau individu*variables

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

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

B –1 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.

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

B –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

B.2- 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) :

$$P=\begin{bmatrix} p_1 & 0 & 0 & 0 & 0 \\ 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & p_i & 0 & 0 \\ 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & p_n \end{bmatrix}$$

$$0 \le p_i \le 1, \quad i=1,\dots n$$

$$\sum_{i=1}^{n} p_i = 1$$

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

B-3 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.

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

- ✓ *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.

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

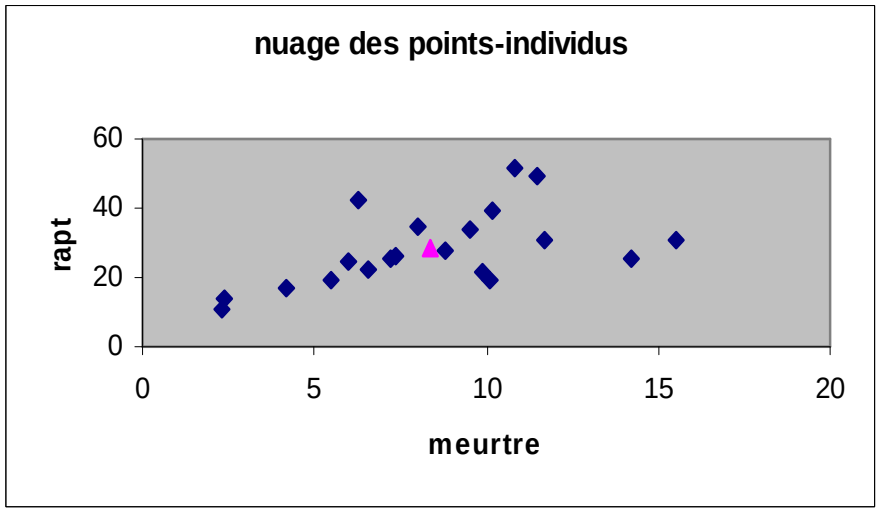
B-3 Nuages de points

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

B-3 Nuages de points

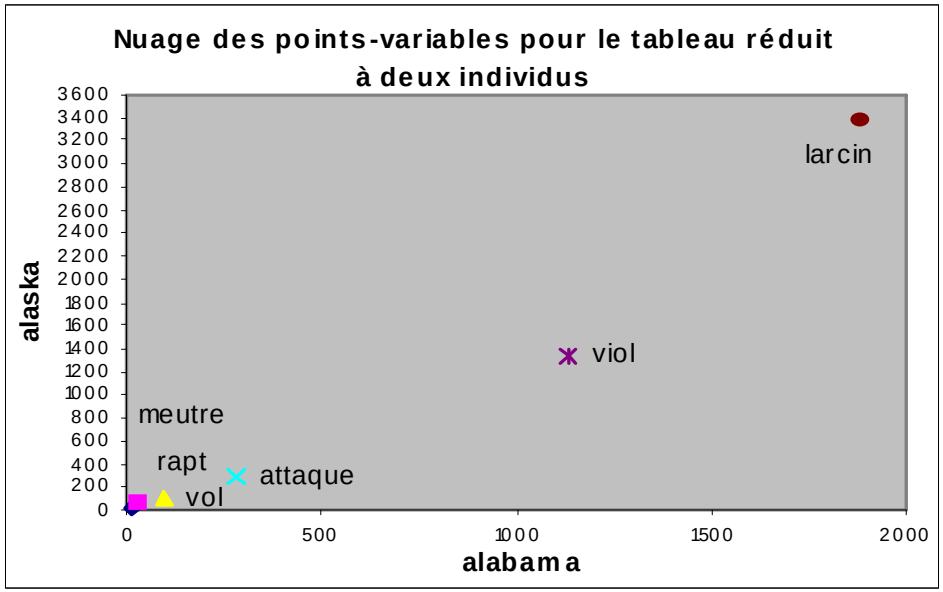
- ✓ Les n individus forment un nuage de points dans le sous-espace de R^p défini par les variables, appelé *nuage des points-individus*



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.

B-3 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.

B-4 Centre de gravité

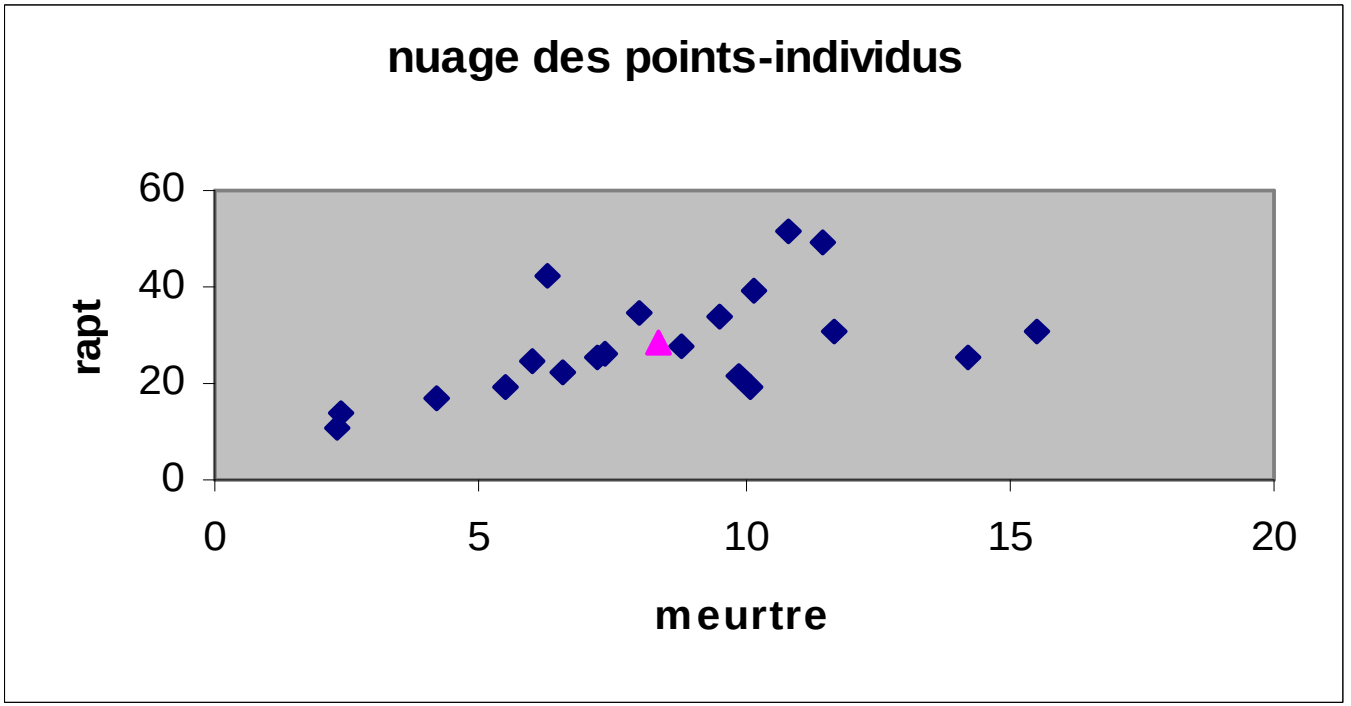
- ✓ Le centre *de gravité du nuage de points individus* G caractérise la position globale de nuage (individu) dans le repère défini par les variables. C'est le point autour duquel « gravitent » les individus du nuage.

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

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.

B-4 Centre de gravité



B-4 Centre de gravité

- ✓ Centre de gravité du tableau des protéine

```
>mean(proteine)

      VR       VB     Oeufs      Lait  Poisson Céréales   Amidon      Noix     FL
   9.828    7.896    2.936   17.112    4.284   32.248    4.276    3.072  4.136
```

B-5 Inertie

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

B-5 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 :

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

B-5 Inertie

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

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

- ✓ 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

B-5 Inertie

```
> cov=cov(crime2)
> c=as.matrix(cov); c
```

	Meutre	Rapt	Vol	Attaque	Viol	Larcin
Meutre	14.95190	25.01378	165.2459	251.4141	645.1653	286.0809
Rapt	25.01378	115.76964	562.6393	798.5073	3313.5864	4795.5602
Vol	165.24587	562.63926	7805.4693	4934.1608	24347.0033	28650.7691
Attaque	251.41408	798.50735	4934.1608	10050.6739	27006.2014	29427.3639
Viol	645.16533	3313.58639	24347.0033	27006.2014	187017.9416	248665.3015
Larcin	286.08095	4795.56021	28650.7691	29427.3639	248665.3015	526943.4505

```
> I=sum(diag(c));I

[1] 731948.3
```

C-1 Tableau 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 :

$$X_c = X - EG'$$

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

C-2 Tableau centré-réduit associé à X

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

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

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

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

C-2 Tableau centré-réduit associé à X

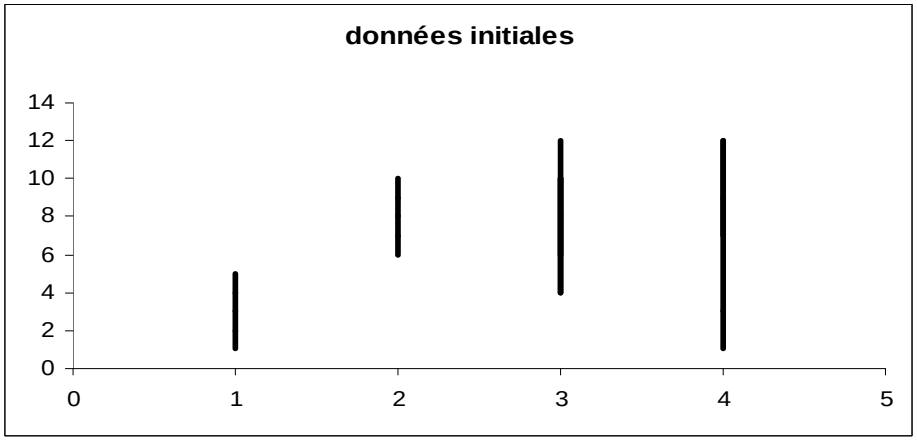
$$X_r = \begin{bmatrix} \frac{x_{11}-\bar{x}_1}{\sigma(X_1)} & \frac{x_{12}-\bar{x}_2}{\sigma(X_2)} & \dots & \frac{x_{1j}-\bar{x}_j}{\sigma(X_j)} & \dots & \frac{x_{1p}-\bar{x}_p}{\sigma(X_p)} \\ \frac{x_{21}-\bar{x}_1}{\sigma(X_1)} & \frac{x_{22}-\bar{x}_2}{\sigma(X_2)} & \dots & \frac{x_{2j}-\bar{x}_j}{\sigma(X_j)}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

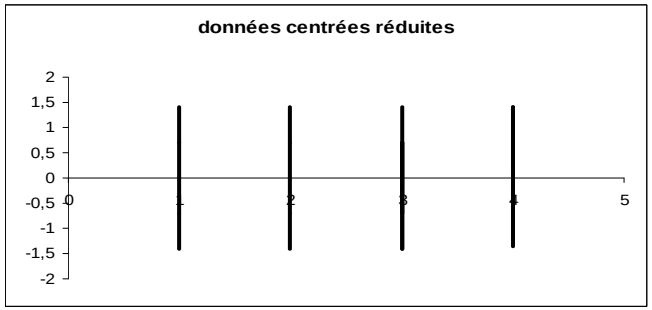
C-2 Tableau centré-réduit associé à X

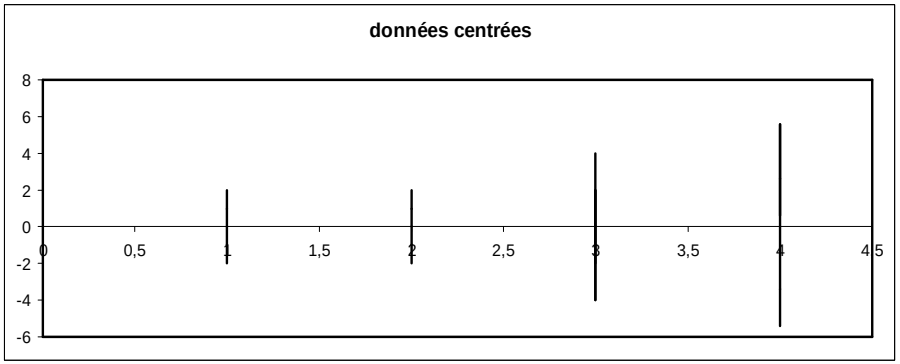
```
> crimer=scale(crime2)*sqrt(20/19); round(crimer, digit=3)
```

	Meutre	Rapt	Vol	Attaque	Viol	Larcin
Alabama	1.793	-0.051	-0.317	0.686	-0.371	-1.116
Alaska	0.890	2.466	-0.317	0.744	0.094	0.987
Arizona	0.546	0.807	0.164	1.034	2.501	2.539
Arkansas	0.360	0.178	-0.475	-0.081	-0.758	-1.144
California	1.076	2.257	1.892	1.501	2.011	1.171
Colorado	-0.304	1.551	0.541	0.835	1.526	1.741
Connecticut	-0.861	-0.852	0.063	-0.814	0.128	-0.071

C-2 Tableau centré-réduit associé à X







C-3 Matrice de variance-covariance associée à X

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

$$V=X_c' P X_c$$

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

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

C-3 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:

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

$$r(X_j, X_k) = X_j^{r\prime} P X_k^r$$

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

C-3 Matrice de corrélation associée à X

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

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

C-3 Matrice de corrélation associée à X

```
> cor=cor(crime2)    #=cov(crimer)
> cor
          Meutre      Rapt       Vol   Attaque      Viol    Larcin
Meutre  1.0000000 0.6012205 0.4837076 0.6485505 0.3858168 0.1019198
Rapt    0.6012205 1.0000000 0.5918793 0.7402595 0.7121301 0.6139882
Vol     0.4837076 0.5918793 1.0000000 0.5570782 0.6372420 0.4467399
Attaque 0.6485505 0.7402595 0.5570782 1.0000000 0.6229085 0.4043633
Viol    0.3858168 0.7121301 0.6372420 0.6229085 1.0000000 0.7921210
Larcin  0.1019198 0.6139882 0.4467399 0.4043633 0.7921210 1.0000000
```

C.4- 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

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