Ch2 : Analyse en Composantes Principales (ACP)

A- Objectifs

B- construction d'un espace factoriel

C- Les étapes d'une ACP

D- Interprétation

E- Limites

A- Objectifs

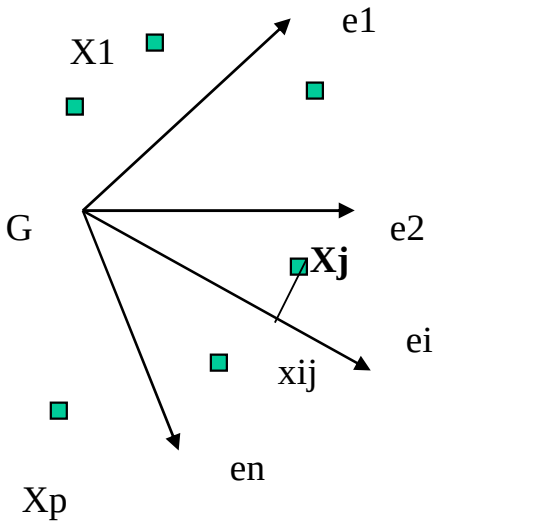
On dispose d'un tableau de données X. Ce tableau définit deux nuages de points :

- ✓ 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)

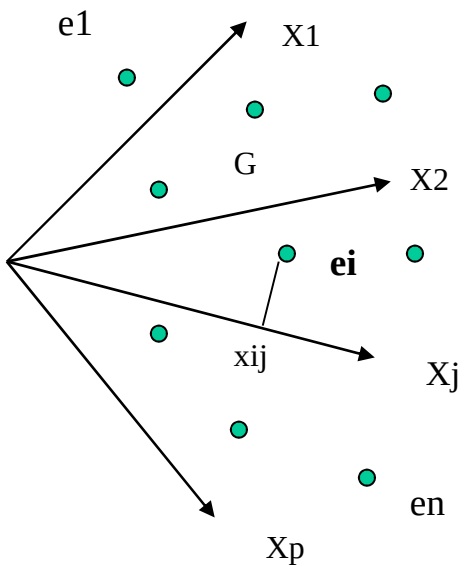
A- Objectifs

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



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

A- Objectifs



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

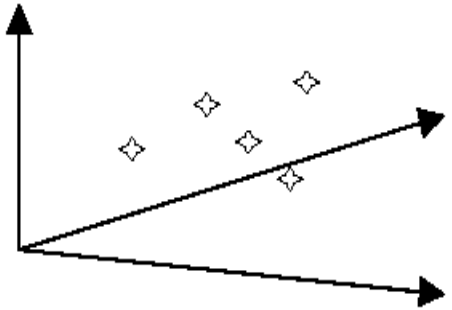
Le nuage des points individus représenté dans l'espace de dim p défini par les variables

A - Objectifs

- Problème : 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

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

  ⇔

  Déterminer un sous-espace de **dimension** q<p (q nouveaux axes) (ou q<n), sur lequel **projeter** les nuages de points relatifs au tableau de données qui soit :

  - **« compréhensible » par l'œil**: q faible, de préférence q=1,2 ou 3
  - **le moins déformant possible (projection la plus fidèle possible)**

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

B- construction d'un espace factoriel

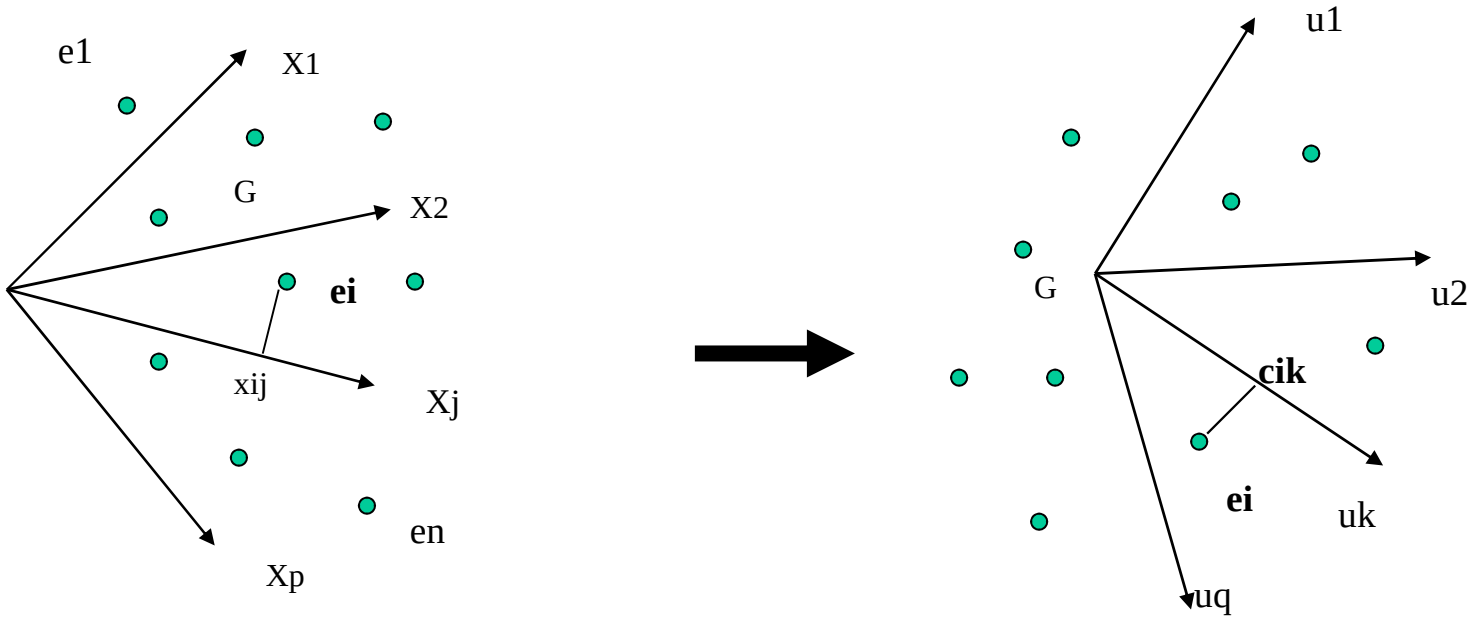
- Principe de construction de l'espace factoriel (ex : individus) :

✓ 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. Il permet de récupérer les liens les plus significatifs contenus dans le tableau

B- construction d'un espace factoriel

$$u_k = a_{k1}X_1 + ...a_{kp}X_p$$



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

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.

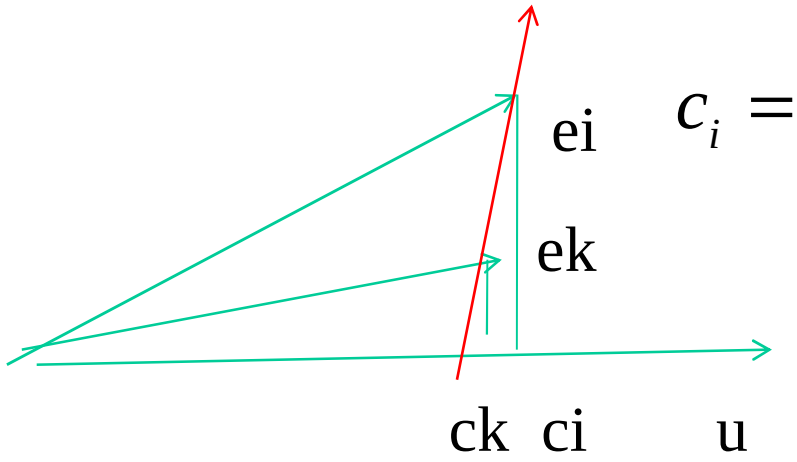
Rq : Dans le second repère, les axes ne véhiculent pas la même information selon leur rang : leur capacité à « résumer » le nuage se détériore au fur et à mesure que l'on observe des axes de rang élevé.

B- construction d'un espace factoriel

- Comment obtenir une déformation minimale ?

Projection sur un axe :

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



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

- 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

- Conclusion :

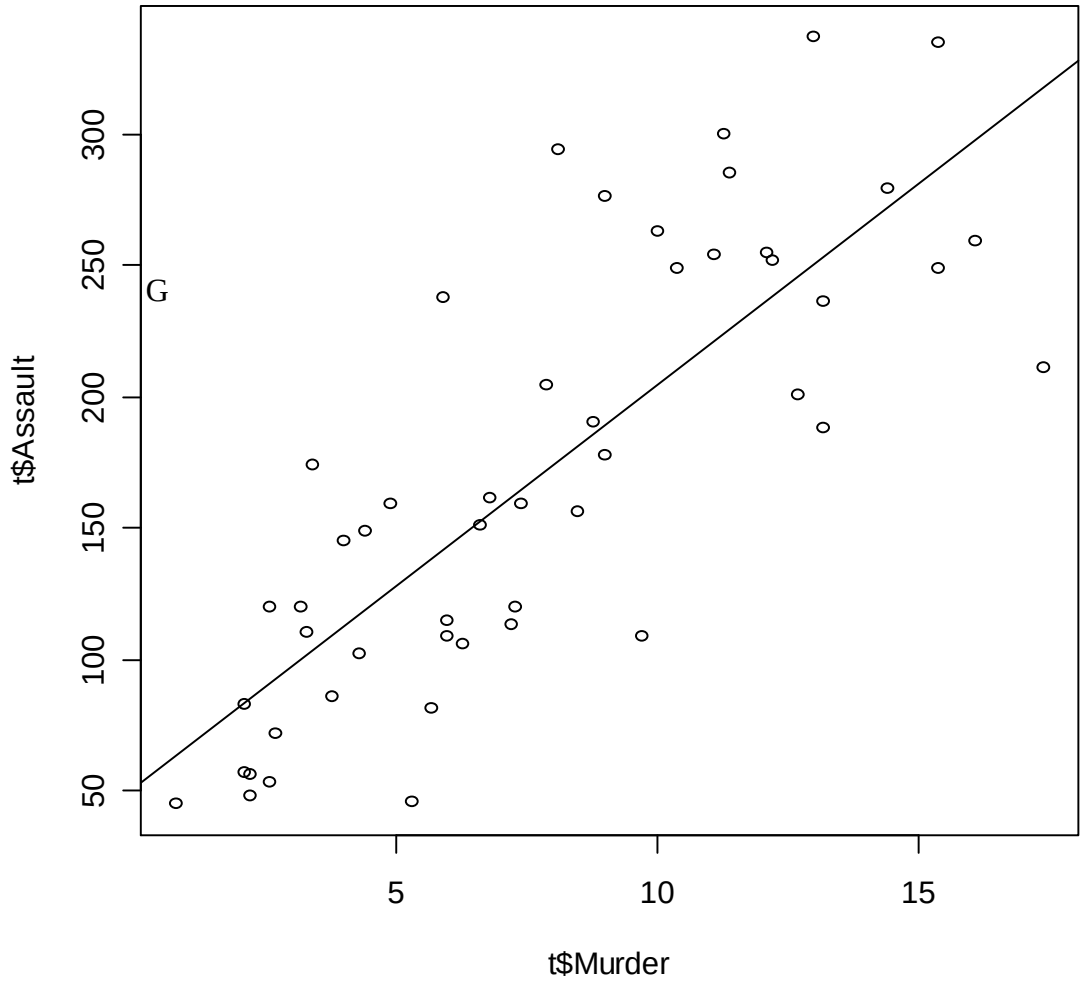
➢ le meilleur axe (premier axe factoriel) sera celui sur lequel le nuage de points projeté est de dispersion, ie tel que le nuage projeté est **d'inertie maximale**.

➢ Le second axe sera celui qui, après le premier est tel que le nuage projeté est **d'inertie maximale**, tout en étant orthogonal au premier

➢ .....

➢ Idem pour le nuage de points variables

B- construction d'un espace factoriel



B- construction d'un espace factoriel

- Interprétation en termes statistiques de l'espace factoriel du nuage de points individus:

✓ Chaque axe factoriel k, de vecteur directeur uk, représente une nouvelle variable Ck de dimension n, construite comme combinaison linéaire des variables (axes) de départ, appelée *composante principale*. La coordonnée cik d'un individu i 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). La première composantes principale sera une CL des variables de départ de dispersion (de variance) maximale.

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

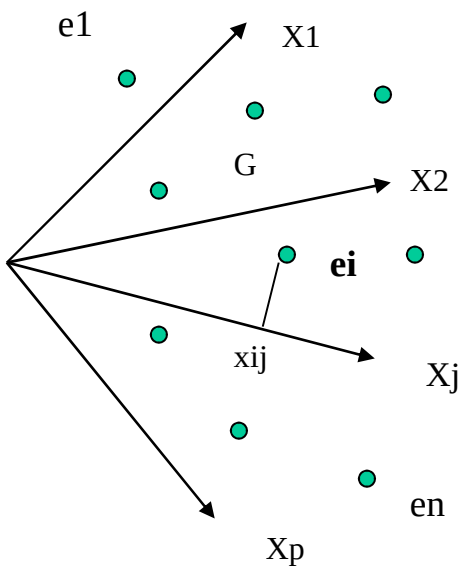
A - Objectifs

Rq : Définir l'espace factoriel revient à

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

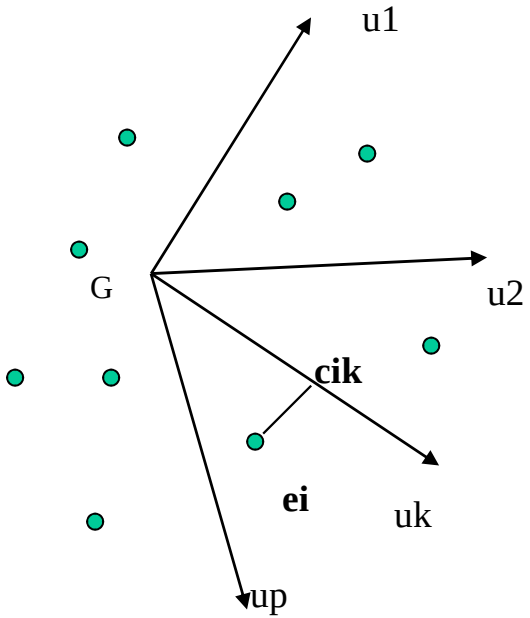
- ✓ Définir q nouveaux individus comme axes du repère du nuage de points-variables

B- construction d'un espace factoriel



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

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



$$c_{ik} = \langle e_i, u_k \rangle = x_{i1}u_{k1} + \ldots + x_{ip}u_{kp}$$

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

- On travaille toujours sur le tableau centré : On montre que tout axe factoriel passe par le centre de gravité : le nouveau repère est centré en G.

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

  Si X n'est pas 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

C-2 Analyse directe
a- Origine du repère

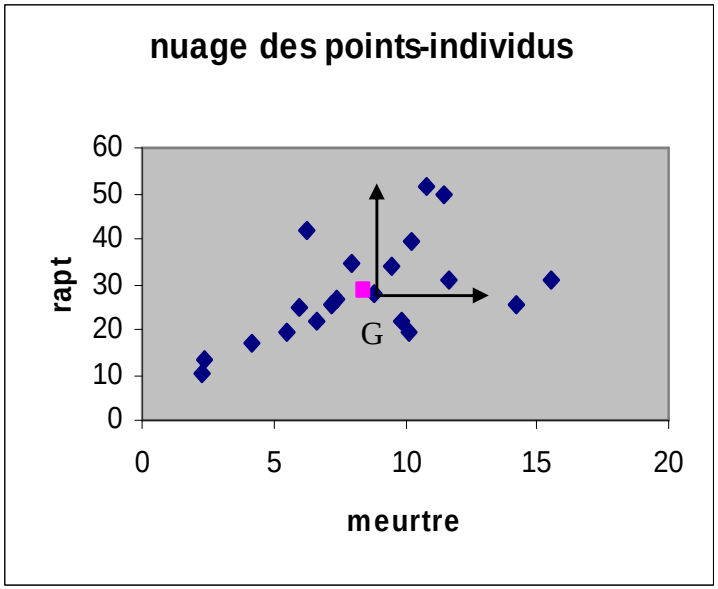
Recherche des axes factoriels du nuage de points-individus

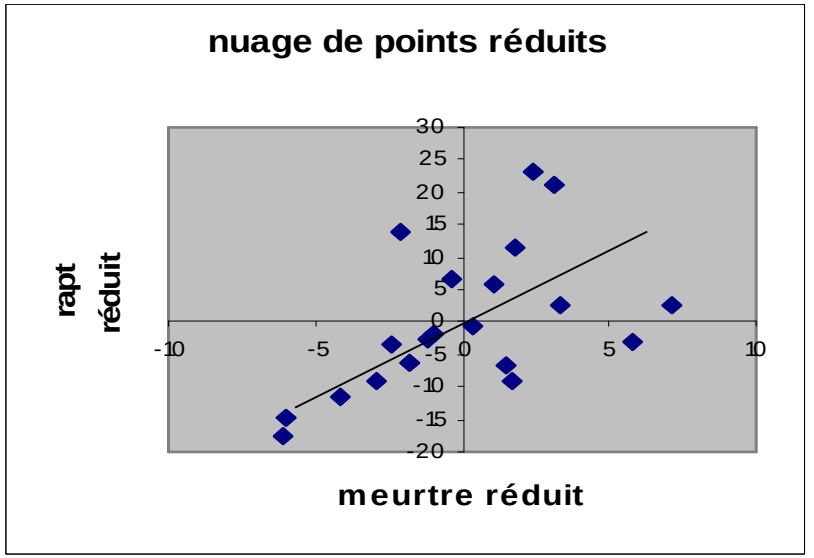
- Détermination de l'origine : on MQ les axes « les plus informatifs » passent forcement par le centre d'inertie du nuage de points. $\Rightarrow$ Le nouveau repère aura pour origine G. On travaille toujours sur le nuage (tableau) centré.

- 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)

I_1 est maximale

sous la contrainte : $\|u_1\| = 1$

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 $u_1 = (u_{11}, \ldots, u_{j1}, \ldots, u_{p1})$

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

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

✓ 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

(P)

$u_1' S u_1$ est maximale

sous la contrainte : $\|u_1\| = 1$

✓ Solution de (P):

u_1 est le vecteur propre unitaire de S associé à la plus grande valeur propre λ_1 . Il vérifie : $Su_1 = \lambda_1 u_1$

.

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

- 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.

- La première composante principale est
  - ✓ Centrée (le nouveau repère a pour origine G).
  - ✓ 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

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

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

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 coordonnées 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

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

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

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

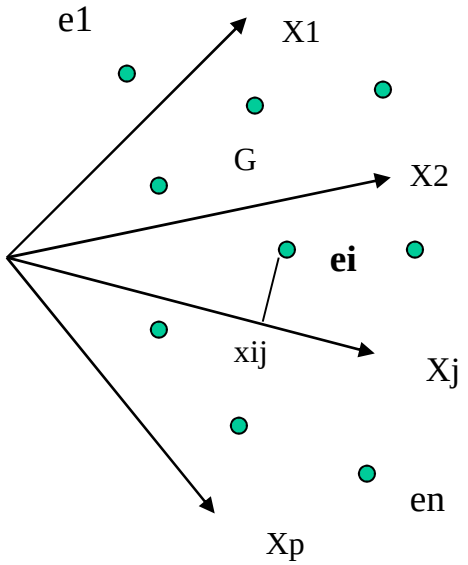
$$D_k = \sqrt{\lambda_k}\, u_k = \frac{X'PC_k}{\sqrt{\lambda_k}} \qquad d_{jk} = \sqrt{\lambda_k}\, u_{jk} = \frac{X_j'PC_k}{\sqrt{\lambda_k}}$$

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

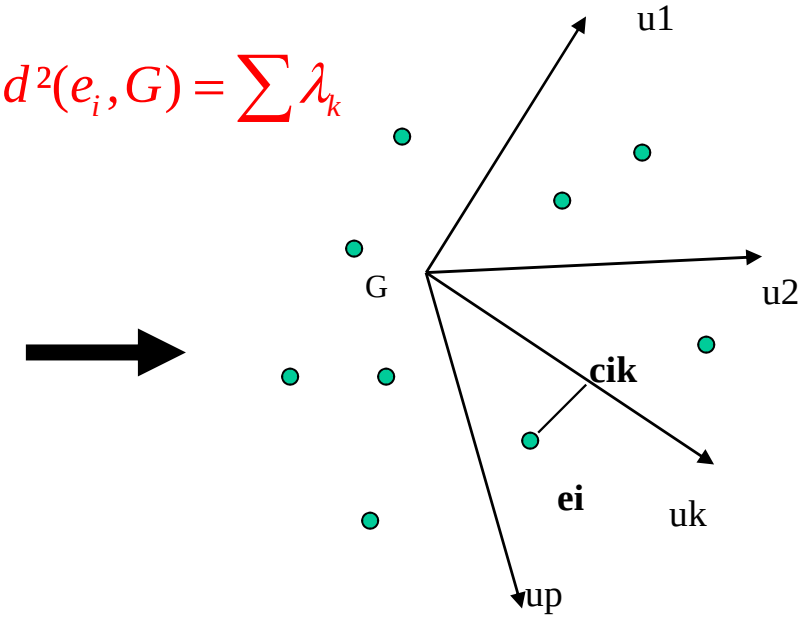
$$d_{jk} = r(X_j, C_k)$$

C-4 Résumé de la décomposition factorielle
Analyse directe

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



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



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

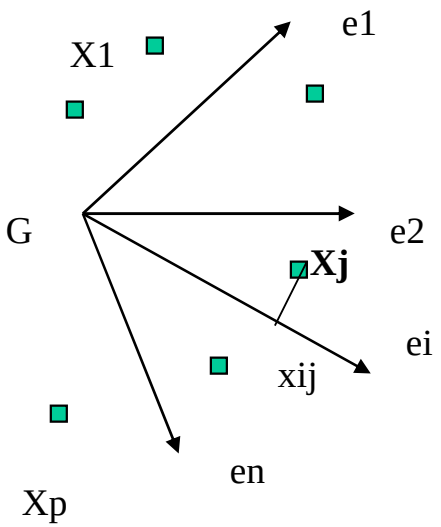
$$Ru_k = \lambda_k u_k, \quad \|u_k\| = 1, \lambda_1 \geq .. \geq \lambda_p$$

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

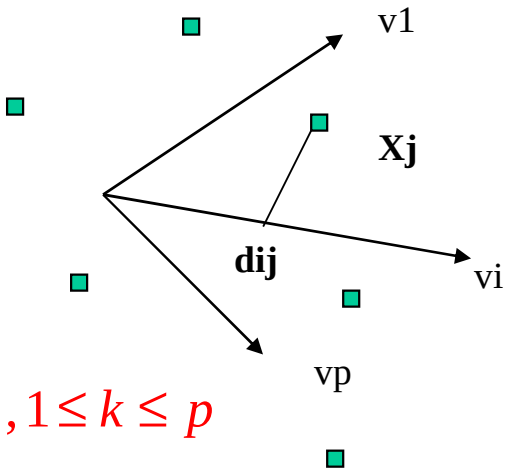
$$C_k = Xu_k, \quad C_k = (c_{1k},..,c_{ik},...,c_{nk})'$$

C-4 Résumé de la décomposition factorielle
Analyse duale

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



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



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

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

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

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 $\boldsymbol{Cov(C_k, C_l) = C_k' P C_l = 0}$
- ✓ de variance maximale. $\boldsymbol{\|C_k\|_p^2 = Var(C_k) = \lambda_k}$

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 R

procedure princomp du package stat

>princomp(x=crime, cor = TRUE)

Call:
Call:
princomp(x = crime3, cor = T)

Standard deviations: $\sqrt{\lambda_k}$

Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6
1.8670647	1.1924148	0.6875928	0.5425280	0.4676170	0.3262364

6 variables and 20 observations.

ACP normée avec R

✓ Valeurs propres de R

>summary(acp)

Importance of components:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
Standard deviation	1.8670647	1.1924148	0.68759281	0.54252803	0.46761698
Proportion of Variance	0.5809884	0.2369755	0.07879731	0.04905611	0.03644427
Cumulative Proportion	0.5809884	0.8179639	0.89676125	0.94581736	0.98226164

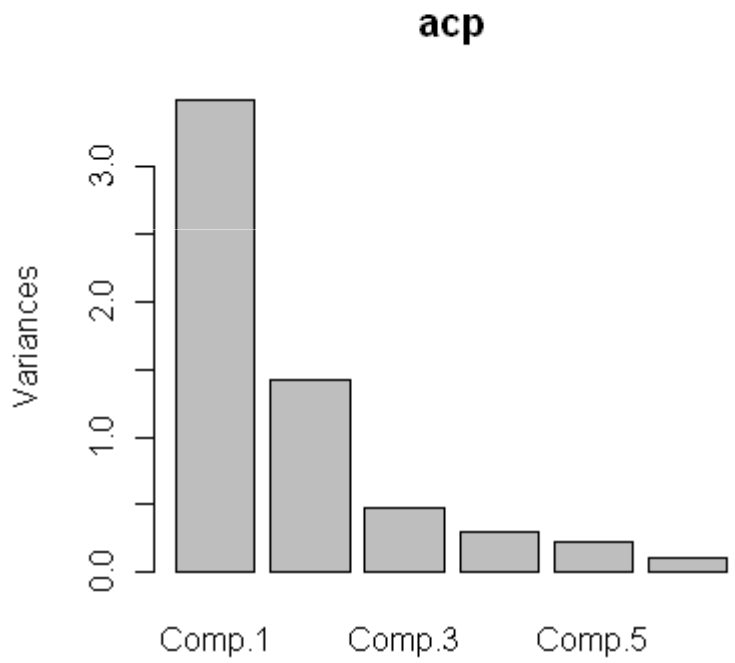
	Comp.6
Standard deviation	0.32623640
Proportion of Variance	0.01773836
Cumulative Proportion	1.00000000

$\sqrt{\lambda_k}$

I_k / I

$$I = 6 = \sum \lambda_k$$

>plot(acp)



ACP normée avec R

✓ Coordonnées des vecteurs propres

>loadings(acp)

u_k

Loadings:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6
Meutre	0.268	0.649	0.269	0.599	0.150	-0.232
Rapt	0.474	0.135	0.301	-0.245	-0.764	0.149
Vol	0.422		-0.876	0.185	-0.137	
Attaque	0.446	0.288		-0.656	0.537	
Viol	0.430	-0.412	0.204	0.336	0.291	0.637
Larcin	0.377	-0.553	0.169		-0.719	

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6
SS loadings	1.000	1.000	1.000	1.000	1.000	1.000
Proportion Var	0.167	0.167	0.167	0.167	0.167	0.167
Cumulative Var	0.167	0.333	0.500	0.667	0.833	1.000

>

ACP normée avec R

✓ Coordonnées des individus sur les axes

> acp$scores

c_{ik}

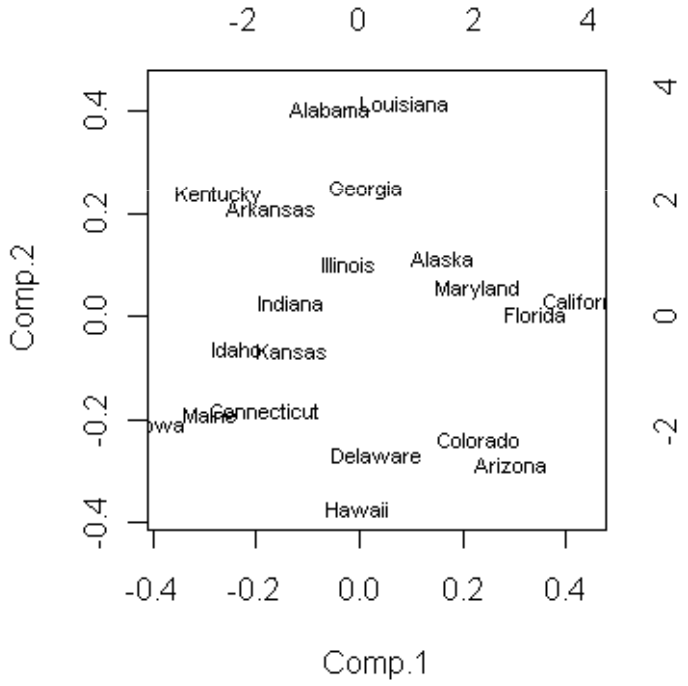
	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
Alabama	-0.47421533	2.17292554	0.491274280	0.36690798	0.569339519
Alaska	1.37443010	0.60952764	1.383696674	-0.57728208	-1.199624829
Arizona	2.46115288	-1.52470179	1.005513852	0.41303753	0.802039585
Arkansas	-1.38815961	1.12678123	0.219480169	-0.29979717	-0.330233705
California	3.71367458	0.17439369	-0.611213633	0.26601608	-0.425546920
Colorado	1.96872562	-1.26030699	0.236035848	-0.48098019	-0.372267715
Connecticut	-1.50957496	-0.96866341	-0.686613377	0.10404155	0.048142778
Delaware	0.29867735	-1.41908466	-0.268773295	0.25293681	0.127685148
Florida	2.87206179	0.03328554	0.217223412	-0.80305229	0.728682057

ACP normée avec R

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

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

>biplot(acp, cex=0.7,col=c(1,0))

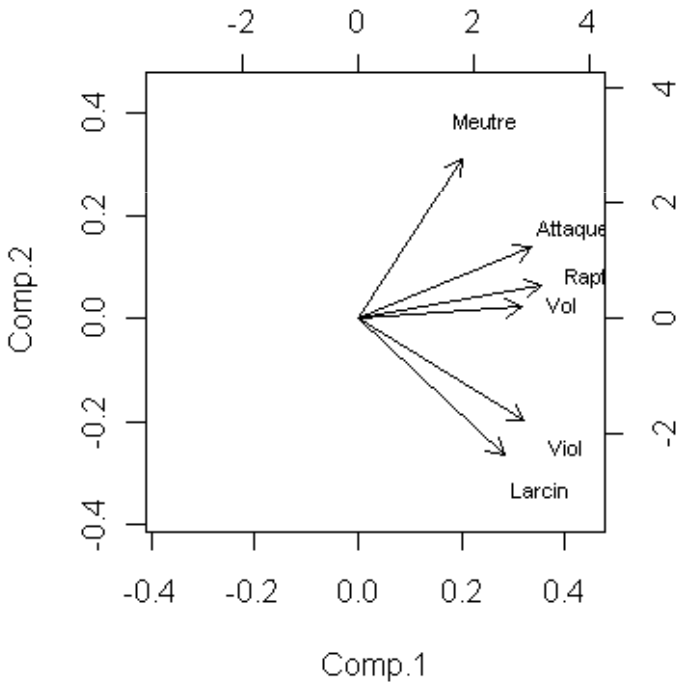


ACP normée avec R

- ✓ 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




```
>biplot(acp, cex=0.7,col=c(0,1))
```

ACP normée avec R

```
> cor(crime2, acp$scores)
           Comp.1     Comp.2      Comp.3     Comp.4     Comp.5      Comp.6
Meutre  0.5010412  0.77373375  0.184695091  0.32487464  0.07017138 -0.075591071
Rapt    0.8851265  0.16088028  0.207216500 -0.13277559 -0.35741465  0.048532584
Vol     0.7876267  0.05377445 -0.602100074  0.10059265 -0.06402387 -0.003181126
Attaque 0.8321579  0.34336607  0.004782874 -0.35573941  0.25102092 -0.005303311
Viol    0.8024956 -0.49122078  0.140518975  0.18219932  0.13607035  0.207955881
Larcin  0.7032765 -0.65970966  0.115930861  0.03602475  0.01925336 -0.234684626
```