Analyse de Données – ID
ACP - le retour de la pratique

Philippe LERAY
philippe.leray@univ-nantes.fr

Equipe COnnaissances et Décision
Laboratoire d'Informatique de Nantes Atlantique – FRE 2729
Site de l'Ecole Polytechnique de l'université de Nantes

Polytech'Nantes
École d'ingénieurs de l'université de Nantes

UNIVERSITÉ DE NANTES

lina LABORATOIRE D'INFORMATIQUE DE NANTES ATLANTIQUE

Choix de la dimension

Composante	Valeur propre « Eigenvalue »	Pourcentage de variance	Pourcentage de variance cumulée
C_1	3.31217	47.3	47.3
C_2	2.61662	37.4	84.7
C_3	.57780	8.3	93.0
C_4	.23840	3.4	96.4
C_5	.13526	1.9	98.3
C_6	.08297	1.2	99.5
C_7	.03678	.5	100.0
Total :	7.00000	100.00	

- ici il faut les 7 composantes pour expliquer toute la variance
- mais avec $k = 2, 3, 4$, la perte d'information semble faible
- plusieurs méthodes ont été proposées pour choisir k ...

Pourcentage d'inertie expliquée

Définition

Le pourcentage d'inertie expliquée est défini par : $\frac{I_{F_k}}{I_F} = \frac{\sum_{j=1}^{k} \lambda_j}{\sum_{j=1}^{p} \lambda_j}$

Choix de k

$\frac{I_{F_k}}{I_F} < \alpha$ (α seuil fixé par l'utilisateur, par ex. 0.95)

Composante	Valeur propre « Eigenvalue »	Pourcentage de variance	Pourcentage de variance cumulée
C_1	3.31217	47.3	47.3
C_2	2.61160	37.4	84.7
C_3	.57700	8.3	93.0
[illegible]	[illegible]	[illegible]	[illegible]
C_5	.13626	1.9	98.3
C_6	.08207	1.2	99.6
C_7	.03678	.5	100.0
Total :	7.00000	100.00	

Règle de Kaiser

Choix de k

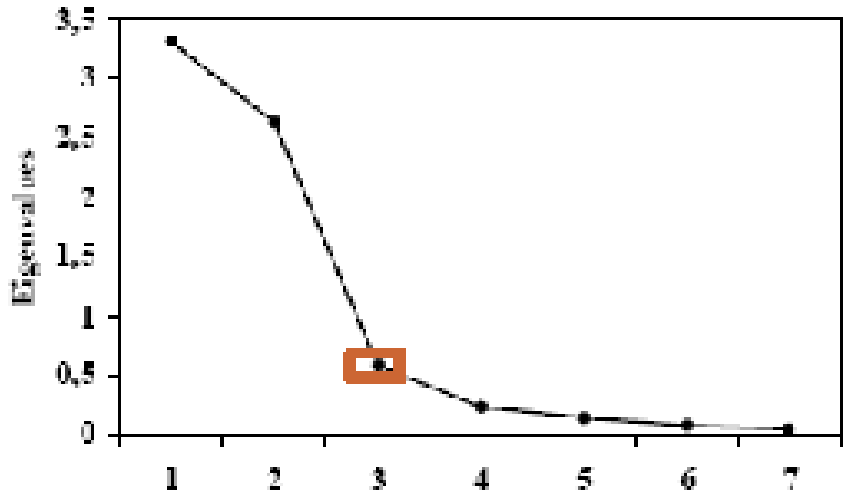
- ne conserver que les valeurs propres supérieures à leur moyenne car seules jugées plus "informatives" que les variables initiales
- dans le cas d'une ACP centrée-réduite, $\sum \lambda_i = p$, donc ne sont donc retenues que celles plus grandes que 1.
- tendance à surestimer le nombre de composantes pertinentes

Composante	Valeur propre « Eigenvalue »	Pourcentage de variance	Pourcentage de variance cumulée
C_1	3.31217	47.3	47.3
C_2	2.61802	37.4	84.7
C_3	[illegible]	[illegible]	[illegible]
C_4	.23040	3.4	93.4
C_5	.13628	1.9	98.3
C_6	.09207	1.2	99.5
C_7	.03678	.5	100.0
Total :	7.00000	100.00	

Éboulis des valeurs propres

Choix de k

- graphique de décroissance des valeurs propres
- rechercher d'un "coude" (changement de signe dans la suite des différences d'ordre 2) dans le graphe
- ne conserver que les valeurs propres jusqu'à ce coude



Interprétation de l'ACP

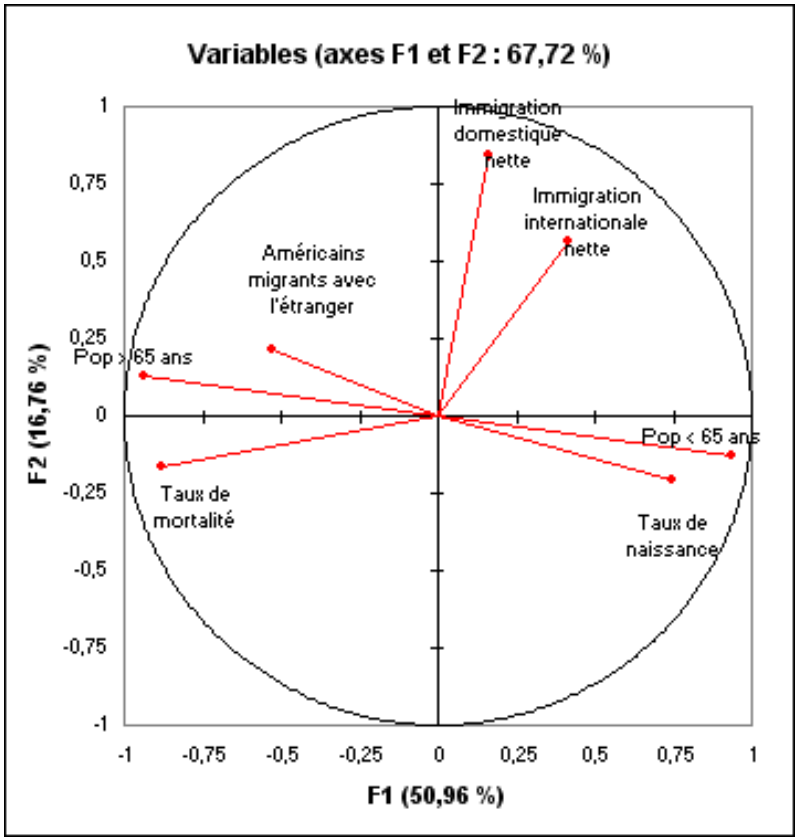
- L'ACP construit de nouvelles variables, artificielles, les composantes principales.
- Des représentations graphiques permettent de visualiser :
  - les relations intra-variables
  - les relations variables / composantes principales,
  - l'existence éventuelle de groupes d'individus et de groupes de variables.

Corrélations composantes-var. initiales

- comment relier une composante principale c aux v_j ?
- il suffit de regarder $|r(c, v_j)|$, val. absolue du coef. de corrélation linéaire...

cercle des corrélations

- pour une paire de composantes principales c_1 et c_2
- v_j = point de coordonnées $r(c_1, v_j)$ et $r(c_2, v_j)$



Place et importance des individus

contribution d'un individu à une composante

Soit c_{jk} la valeur de la composante c_k pour le j-ème individu. On a:

$$\sum_{j=1}^{n} p_i c_{jk}^2 = \lambda_k.$$

La contribution de l'individu j à la composante c_k est définie par :

$$\frac{p_j c_{jk}^2}{\lambda_k}$$

En pratique, on considère comme importante une contribution qui excède le poids p_i de l'individu concerné.

Qualité des représentations

Mesure globale : le pourcentage d'inertie expliquée

Si par exemple $(\lambda_1 + \lambda_2)/I = 0.9$,

- le nuage de points est presque aplati sur un plan
- la représentation du nuage dans le plan est satisfaisante

Mesures individuelles

- pour comparer deux points dans F_k, il faut être sûr que ces points soient bien représentés lors de la projection...
- qualité de représentation d'un individu $i_j = cos^2$ de l'angle entre le plan principal et le vecteur i_j
  - si cos^2 grand, la projection de i_j dans le plan a du sens
  - sinon, on évite d'utiliser la projection de ce point dans les interprétations.