Maryse Raffestin – octobre 2005

Pratiquer une analyse en composantes principales

I - Objectif de l'ACP

1 - LA PLACE DE L'ACP DANS LES METHODES STATISTIQUES

a) Lorsqu'on recueille des informations sur des *individus* ou *unités statistiques* (un individu, au sens statistiques du terme, peut être une personne physique, une entreprise, un pays ,etc.) , on aboutit à la constitution d'un *tableau individus-variables* du type suivant :

individus	V1	V2	V3	,	,	Vp
I1						
I2						
I3						
I4						
I5						
,						
,						
,						
In						

Ex. : les individus sont les 200 étudiants de 1e année de DEUG et les variables sont : l'âge, le sexe, le redoublement (oui/non), la série du bac, les notes en maths, stats, économie, etc.

Pour décrire ces données, si elles sont nombreuses, le statisticien traitera d'abord les variables une par une (*traitements univariés*), puis il s'intéressera aux éventuelles interactions entre deux variables (*traitement bivariés*) voire plus (*traitements multivariés*). Après *l'analyse descriptive* des données (où toutes les variables sont placées sur le même plan), il poursuivra dans certains cas par une *analyse explicative* (il y a alors d'une part la variable expliquée, d'autre part les variables explicatives).

Les variables (ou *caractères*) auxquels on s'intéresse sont essentiellement de deux natures :

- les *variables quantitatives* (en abrégé VQT) sont *mesurées* par un nombre.
ex. : âge, chiffre d'affaires, note en stats, température, poids

- les *variables qualitatives* (en abrégé VQL) peuvent prendre plusieurs *modalités* :
ex. : sexe, série du bac, code APE, jour de la semaine, profession

L'ACP sert à ***décrire*** des tableaux "individus-variables ***quantitatives***" de grande dimension (beaucoup de variables - c'est un traitement ***multivarié***- et beaucoup d'individus -s'il y an a peu, inutile de faire appel à des outils statistiques pour résumer).

Remarques :

- dans les enquêtes d'opinion, on utilise souvent des variables d'un 3e type : les variables ordinales qui indiquent un rang de classement (ex : classer des produits par ordre de préférence, se situer sur une échelle allant de "très favorable" à "très défavorable", etc.)
- on peut passer d'une variable quantitative (total des points obtenus à un concours) à une variable ordinale (rang de classement au concours) puis à une variable qualitative (reçu / collé). Mais l'inverse n'est pas possible. En effectuant cette transformation, on perd de l'information. C'est le cas lorsqu'on passe d'une variable QT (ex. : "effectif salarié") à une variable classifiée ("tranche d'effectif" avec par ex. 3 classes : petites, moyennes et grandes entreprises) qui devient alors une variable QL.

b) Les traitements statistiques des variables qualitalives et des variables quantitatives sont fondamentalement différents. C'est vrai notamment pour les méthodes descriptives :

- caractères qualitatifs :

Traitement univarié : on calculera la distribution des effectifs n_i ou des fréquences $f_i = {}^{n_i}/_{n}$ selon les différentes modalités. On pourra éventuellement noter la modalité dominante (celle qui a la plus forte fréquence), parfois appelée "mode" par analogie aux variables quantitatives.

Traitement bivarié : lorsqu'on répartit une population selon 2 caractères qualitatifs, on constitue un *tableau de contingence*. Ce tableau peut être très grand (donc illisible) si les caractères étudiés comportent beaucoup de modalités(ex. : répartition de la population française par régions et classes d'âges). Une technique d'analyse factorielle, l'analyse factorielle des correspondances (AFC) sert à décrire les grands tableaux de contingence. Lorsque l'observation porte sur un échantillon, le test du χ^2 sert à juger de l'indépendance de ces 2 caractères

- caractères quantitatifs :

Traitement univarié : on peut calculer la moyenne et l'écart-type, ainsi que les quantiles (médiane, quartiles, déciles , centiles...). Lorsque l'observation porte sur un échantillon, on peut estimer ces paramètres par intervalle de confiance.

Traitement bivarié : lorsqu'on s'intéresse à la liaison entre deux variables QT, on peut représenter le nuage des points $M_i(x_i, y_i)$ et examiner sa forme. La covariance et le coef de corrélation linéaire sont des indicateurs de l'intensité de la liaison linéaire éventuelle de ces deux variables.

Traitements multivariés : lorsqu'on s'intéresse à la liaison entre plus de deux ou trois variables QT, on ne peut plus représenter graphiquement le nuage des points M_i. L'ACP nous permet de l'observer sous ses angles les plus intéressants, en examinant les projections du nuage sur des plans qui en conserve le mieux la forme. Elle permet également de repérer les groupes de variables fortement corrélées entre elles, et éventuellement de détecter des caractères complexes sous-jacents à ces groupes.

2 - LES DONNÉES TRAITEES EN ACP

Soit X un tableau à n lignes et m colonnes. La ligne i décrit la valeur prise par m *variables quantitatives* pour l'individu i . Avant toutes choses, les données sont centrées et réduites, c'est-à-dire que chaque variable a une moyenne nulle et une variance égale à 1.

On note X_j le vecteur-colonne constitué par les éléments de la colonne j ; x_{ij} désigne l'élément situé à l'intersection de la ligne i et de la colonne j, c'est-à-dire la valeur de la variable x_j pour l'individu i .

3 - LE PROBLÈME

Pour observer sous un angle plus favorable les données contenues dans le tableau X, on remplace les anciens axes (donc les anciennes variables x_k) par de nouveaux axes (donc par des variables nouvelles C_k). Ces nouvelles variables C_k sont appelées *composantes principales*; elles s'expriment comme combinaisons linéaires des anciennes variables $x_1,....x_m$.

$$C_k = a_{k1}x_1 + a_{k2}x_2 .....+ \ a_{km}x_m$$

Les nouveaux axes, appelés *axes factoriels*, sont choisis de la façon suivante :

- le 1er axe factoriel , ou *axe principal d'inertie*, est la direction de "plus grand allongement" du nuage (en statistiques on dit : "de plus grande dispersion" ou "de plus grande inertie" du nuage).
Lorsque on projette les points P_i du nuage sur cet axe, leurs projections H_i sont plus dispersées qu'elles ne le seraient sur n'importe quel autre axe . L'axe factoriel F1 est donc l'axe selon lequel est préservé, par projection, le maximum de la dispersion initiale des points du nuage.
Le fait que le nuage soit allongé précisément dans cette direction doit trouver une explication. La nouvelle variable C1 (la *composante principale n°1*) est le caractère selon lequel les individus se différencient le plus. Pourquoi ? Quelle signification peut bien avoir cette variable qui combine avec des

poids plus ou moins importants (les coefficients a_i) les variables initiales mesurées sur les individus? Une étape fondamentale de l'ACP est l'interprétation de cette composante principale, qui se fera par l'examen de sa combinaison avec les variables de départ. On espère toujours pouvoir détecter dans cette nouvelle variable un *caractère complexe*, qui n'est pas directement mesurable par une seule quantité, mais bien réel, comme par exemple la *santé* (pour des individus, pour des entreprises...), l'i*ndustrialisation* (d'une région...), la qualité du *jeu d'attaque* (pour un joueur de football, de tennis...), la *compétence dans les matières quantitatives* (pour un étudiant), etc.

- le 2e axe factoriel est la 2e direction d'allongement du nuage, c'est-à-dire celle qui explique, après le 1er axe, le maximum de l'inertie résiduelle. De plus le 2e axe est choisi orthogonal au 1er , ce qui traduit - comme nous le verrons- le fait que la 2e composante principale est non corrélée à la 1e (les vecteurs directeurs des 2 premiers axes ont un produit scalaire nul $\Leftrightarrow$ les 2 premières composantes principales ont une covariance nulle). Comme précédemment, on cherchera à donner un sens à cette 2e composante principale, en observant comment elle combine les variables de départ.

- et ainsi de suite, jusqu'à avoir remplacé les m anciens axes par m nouveaux axes (les axes factoriels), portant des part décroissantes de la dispersion initiale et dont les 2, 3 ou 4 premiers suffisent souvent à donner une image à peine déformée du nuage initial. C'est cette image **réduite donc beaucoup plus accessible à notre observation** que nous examinerons pour décrire et analyser les données du tableau initial.

Mathématiquement, la détermination des axes factoriels se fait par diagonalisation de la matrice de variances-covariances, d'où le vocabulaire utilisé (valeurs propres, vecteurs propres)

II - Interpréter une ACP

1 - Le % de l'inertie expliquée par les premiers axes factoriels

Un facteur est une variable composite fabriquée à partir des variables d'origine; il s'exprime comme combinaison linéaire des anciennes variables. Le 1er axe factoriel correspond à la variable composite qui différencie le mieux les individus.

Le % d'inertie (ou "variance" du nuage ou "dispersion") expliquée par un axe factoriel permet d'évaluer en quelque sorte la quantité d'information recueillie par cet axe. Notons que l'inertie expliquée par un axe est égale à la *valeur propre* correspondante et que l'inertie totale (somme des valeurs propres) est égale au nombre de variables de départ dans le cas d'une analyse sur données centrées-réduites (qui est l'option par défaut dans la plupart des logiciels) .

La qualité de la représentation des données par un plan factoriel s'évalue en ajoutant les % d'inertie expliquée par les 2 axes. Si les 2 premiers axes factoriels expliquaient 100% de l'inertie du nuage , tous les points-individus seraient situés dans le plan factoriel 1-2 . Ceci n'arrive jamais... Il faut en général plusieurs facteurs pour expliquer une part significative de la dispersion.

S'il n'y avait pas de direction privilégiée d'allongement du nuage, chaque axe factoriel porterait une part identique de la dispersion : 100% divisé par le nombre p de variables. Dans le cas d'une analyse sur données centrées-réduites, chaque valeur propre serait égale à 1. Ainsi , s'il y a au départ 5 variables , un % d'inertie expliquée par le 1er axe factoriel qui serait de 25% montre que le nuage n'a pas de véritable axe d'allongement remarquable (25% comparé à 20%, c'est peu), alors que ce serait tout à fait remarquable s'il y a au départ 50 variables (25% comparé à 2% , c'est énorme).

Le cas le plus intéressant est évidemment celui où avec un petit nombre d'axes on arrive à bien résumer un nuage d'un espace de grande dimension.

L'analyse est pertinente si, avec un petit nombre d'axes, on explique une part importante de l'inertie.

Il est difficile de donner une règle pour savoir combien d'axes on va retenir. Certains critères peuvent être proposés :
- retenir autant d'axes qu'il le faut pour atteindre le seuil de variance expliquée désiré (80% par ex.)
- observer le changement de concavité de la courbe des valeurs propres (*cf* Market – Nathan -p.373)
- retenir les valeurs propres supérieures à 1 (dans le cas d'une analyse sur données centrées-réduites)

En pratique, on pourra difficilement interpréter plus de 3 axes, parfois 4. Donc concrètement l'analyse mérite d'être poursuivie si avec 3 ou 4 axes, on conserve une part importante de l'inertie initiale.

2 - La démarche d'interprétation d'une ACP

1 - Tenter de donner une signification aux nouveaux axes retenus pour l'analyse (les 2 ou 3 premiers, parfois 4), en les interprétant à partir des variables de départ. Pour cela , on examine le nuage des points-variables, inscrit dans le cercle des corrélations.
2 - Etudier (éventuellement) le nuage des individus par référence aux nouveaux axes dont l'interprétation vient d'être donnée. Attention aux effets de perspective !

Les points-variables

- Les nouvelles variables, associées aux axes factoriels, sont appelées facteurs ou composantes principales. Elles s'expriment comme combinaisons linéaires des anciennes variables .
- Les coefficients de ces combinaisons linéaires sont fournis par le logiciel; c'est eux qui définissent les nouveaux axes :
    - ils permettent de calculer les nouvelles coordonnées d'un point-individu à partir des anciennes
    - ils permettent également de voir le poids d'une ancienne variable dans la définition d'un facteur. Le repérage des variable d'origine correspondant aux **coefficients les plus élevés**

en valeur absolu permet de dégager une interprétation des facteurs. Cette interprétation est facilitée par l'examen des corrélations "anciennes- nouvelles" variables (qui sont d'ailleurs proportionnelles aux coefficients) représentées dans le cercle des corrélations...

Le cercle des corrélations

A chaque **point-variable**, on associe un point dont la **coordonnée** sur un axe factoriel est une mesure de la **corrélation** entre cette variable et le facteur. Dans l'espace de dimension p la distance des points-variables à l'origine est égale à 1. Donc par projection sur un plan factoriel les points-variables s'inscrivent dans un cercle de rayon 1 - le cercle des corrélations - et sont **d'autant plus proche du bord du cercle** que le point-variable est bien représenté par le plan factoriel, c'est-à-dire **que la variable est bien corrélée avec les deux facteurs** constituant ce plan.
Attention ! Les variables qui ne sont pas situées au bord du cercle dans un plan factoriel ne sont pas corrélées avec les deux facteurs représentées. Elles ne servent pas à l'interprétation et l'effet de perspective empêche d'interpréter la proximité de deux variables (voir d'autres plans factoriels, où la corrélation sera plus forte)

- L'angle entre 2 point-variables, mesuré par son cosinus est égal au coefficient de corrélation linéaire entre les 2 variables: $\cos \alpha = r(X_1,X_2)$

Ainsi :

- si les points sont très proches (α peu différent de 0) : $\cos \alpha = r(X_1,X_2) = 1$ donc X_1 et X_2 sont très fortement corrélés positivement
- si α est égal à 90°, $\cos \alpha = r(X_1,X_2) = 0$ alors pas de corrélation linéaire entre X_1 et X_2
- si les points sont opposés, α vaut 180°, $\cos \alpha = r(X_1,X_2) = -1$: X_1 et X_2 sont très fortement corrélés négativement

Le cercle des corrélations permet de voir, parmi les anciennes variables, les groupes de variables très corrélées entre elles.

Pour interpréter un axe, on examine les coefficients de la combinaison linéaire qui le définissent ou bien – si on préfère - on examine sa corrélation avec les anciennes variables en observant le cercle des corrélations (ou le tableau donnant ces corrélations).
Une variable qui a une coordonnée faible, donc un coefficient faible, ne sert pas pour l'interprétation d'un facteur. Une variable (ou un groupe de variables) ayant un coefficient fort -positif ou négatif- servira d'abord par elle-même (les "forts" en facteur i sont les "forts" en x_k, x_l..(ou les "faibles"en cas de corrélation négative) mais également par opposition à d'autres variables diamétralement opposées.

Les points-individus

La qualité de la représentation d'un point M par un axe U dépend de sa distance à l'axe dans le nuage, mesurée par l'angle (OM, U), ou plus exactement par son cosinus ou son $\cos^2$. (s'il est proche de 1 le point est bien représenté).

La qualité de la représentation d'un point M par un plan factoriel constitué de 2 axes est mesurée par la somme des $\cos^2$ avec 2 axes (Pythagore!).

La position d'un point-individu par rapport à un axe factoriel , ainsi que les proximités entre les individus, peuvent être interprétées dès lors que ces points sont bien représentés par le plan factoriel observé. Certains individus seront bien représentés par le plan 1-2 (les "très forts" ou "très faibles " en facteur 1 et 2 surtout), d'autres par le plan 1-3 s'ils sont mieux décrits par l'axe 3, etc.

III - Exemple

Pendant une semaine, 2000 femmes de 30 à 40 ans ont noté leur emploi du temps quart d'heure par quart d'heure. On a ainsi calculé la durée hebdomadaire qu'elles ont consacrée aux 10 activités quotidiennes ci-dessous :

profess	travail professionnel
transp	transport
sommeil	sommeil
sport	activités physiques et sportives
courses	shopping - courses
enfants	enfants
toilette	toilette
cuisine	préparation des repas
menage	travail ménager
tele	télévision

Une ACP a été effectuée sur le tableau individus x variables de dimension (2000 x 10) ainsi constitué. Le logiciel SAS a fourni les résultats suivants, donnant respectivement les valeurs propres, les coordonnées des vecteurs propres et les corrélations des composantes principales avec les anciennes variables (pour ces 2 derniers tableaux, on a retenu seulement 3 composantes). L'étude des individus n'est pas réalisée ici.

Interpréter les résultats ci-dessous.

Eigenvalues (CORR)

Component	Eigenvalue	Difference	Proportion	Cumulative
1	6.811855	5.331191	0.6812	0.6812
2	1.480664	0.520142	0.1481	0.8293
3	0.960522	0.687585	0.0961	0.9253
4	0.272937	0.080137	0.0273	0.9526
5	0.192800	0.087564	0.0193	0.9719
6	0.105236	0.015654	0.0105	0.9824
7	0.089582	0.048620	0.0090	0.9914
8	0.040961	0.007344	0.0041	0.9955
9	0.033617	0.021790	0.0034	0.9988
10	0.011827	_	0.0012	1.0000

Eigenvectors (CORR)

	Component		
Variable	1	2	3
profess	-0.357484	0.186237	-0.059064
transp	-0.342792	0.165548	0.098082
sommeil	0.096610	0.122027	0.972935
sport_	0.272695	-0.555467	0.062051
courses	0.273525	-0.562038	0.029820
enfants	0.344824	0.265692	-0.025668
toilette	0.347996	0.268115	-0.016325
cuisine	0.336030	0.268342	-0.149160
menage	0.350519	0.232441	-0.009884
tele	0.346991	0.173419	-0.110799

Correlations (Structure)

Variable	PCR1_2	PCR2_2	PCR3_2
profess	-0.9330	0.2266	-0.0579
transp	-0.8947	0.2014	0.0961
sommeil	0.2521	0.1485	0.9535
sport_	0.7117	-0.6759	0.0608
courses	0.7139	-0.6839	0.0292
enfants	0.9000	0.3233	-0.0252
toilette	0.9083	0.3262	-0.0160
cuisine	0.8770	0.3265	-0.1462
menage	0.9148	0.2828	-0.0097
tele	0.9056	0.2110	-0.1086