L'Analyse en Composantes Principales (ACP)

Un exemple élémentaire

On considère la population constituée par 17 pays (ou ***individus***) sur lesquels on a relevé les valeurs de deux ***caractères***: l'espérance de vie (EVI), et le taux d'analphabétisme (ANA) en 1970.

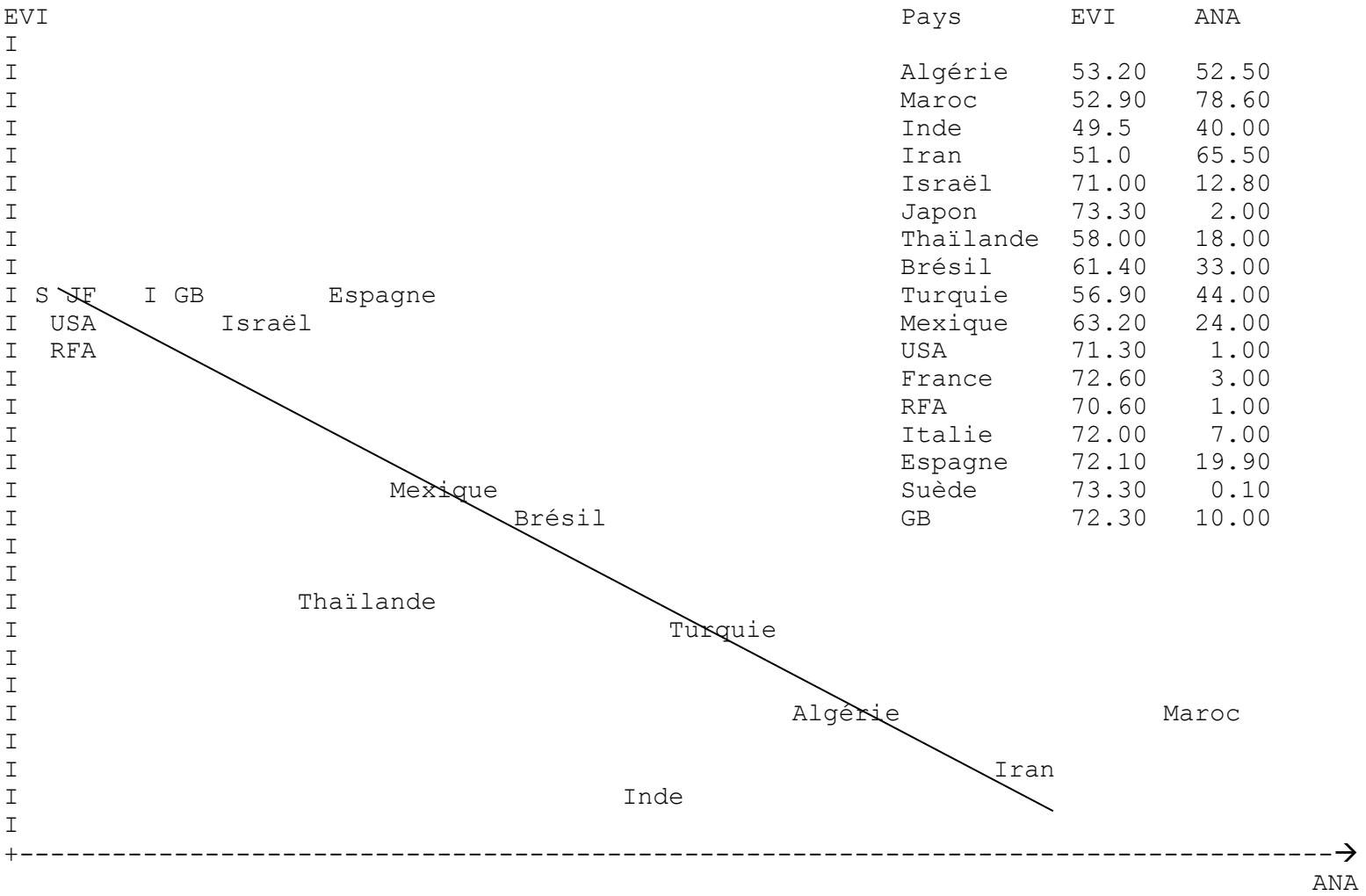
Le graphique 1 donne la représentation de ces données dans un diagramme EVIxANA.

On remarque que l'ensemble des points figurant les pays (ou ***nuage des individus***) présente une direction privilégiée, approximativement tracée sur le graphique.

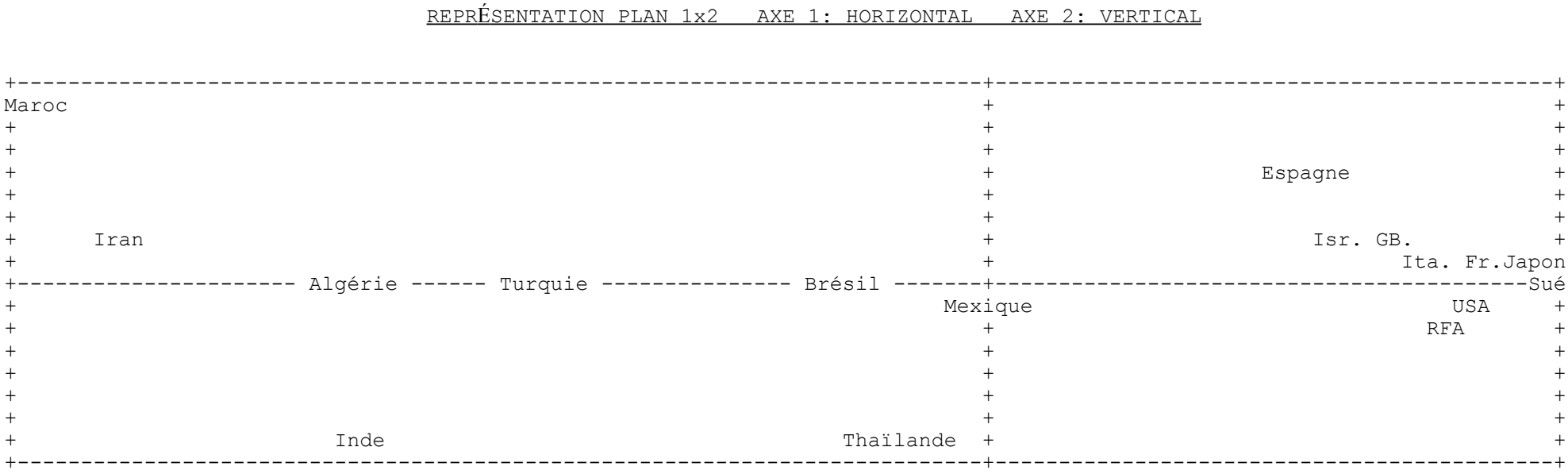
On peut songer à faire un nouveau graphique dans lequel cette direction serait l'un des axes de coordonnées, le second étant naturellement perpendiculaire. Par commodité, on place l'origine au centre de gravité du nuage. A quelques conventions supplémentaires non explicitées pour l'instant près, c'est là l'idée de l'ACP.

On obtient le graphique 2. On y voit les pays s'égrener suivant l'axe 1, des plus développés aux plus arriérés. Cet axe peut s'interpréter comme l'axe du progrès. L'éloignement selon l'axe 2, transversal, note au contraire un développement différent: d'un côté l'Espagne et le Maroc, où l'alphabétisation est en retard sur l'espérance de vie par rapport au comportement général, de l'autre côté l'Inde et la Thaïlande, en situation inverse.

Exemple: espérance de vie x taux d'analphabétisme en 1970



Pays	EVI	ANA
Algérie	53.20	52.50
Maroc	52.90	78.60
Inde	49.5	40.00
Iran	51.0	65.50
Israël	71.00	12.80
Japon	73.30	2.00
Thaïlande	58.00	18.00
Brésil	61.40	33.00
Turquie	56.90	44.00
Mexique	63.20	24.00
USA	71.30	1.00
France	72.60	3.00
RFA	70.60	1.00
Italie	72.00	7.00
Espagne	72.10	19.90
Suède	73.30	0.10
GB	72.30	10.00



Tableaux de données, notations

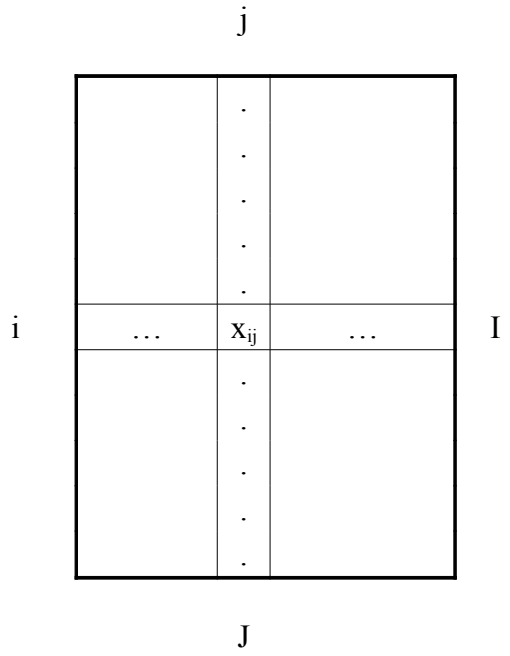
L'ACP traite des tableaux rectangulaires donnant les valeurs d'un ensemble de caractères quantitatifs relevés sur un ensemble d'individus.

Dans l'exemple préliminaire, les individus étaient les 17 pays retenus, les caractères, au nombre de deux, l'espérance de vie et le taux d'analphabétisme.

L'usage dominant - que nous adoptons ici - est de placer les individus en ligne, et les caractères, ou *variables*, en colonne.

Soit X un tableau de données, notons I l'ensemble des individus (lignes), en nombre n, et J, l'ensemble des variables (colonnes), en nombre p, et x_{ij} est la valeur prise par l'individu i pour le caractère j.

La ligne $x_{i.}$ est un vecteur de R^p donnant les valeurs prises par l'individu i pour les p caractères de J. On parlera indifféremment de ligne $x_{i.}$, de ligne i ou d'individu i.



La colonne $x_{.j}$ est un vecteur de R^n donnant les valeurs du caractère j relevées sur les n individus de I. On parlera indifféremment de colonne $x_{.j}$, de colonne j, de variable ou de caractère j.

Exemples de tableaux de données

- I = Ensemble de personnes, J = Ensemble de caractères biologiques (taille, poids, rythme cardiaque, capacité thoracique, etc.).
- I = Ensemble d'étudiants, J = Ensemble de matières, x_{ij} étant la note obtenue par l'étudiant i dans la matière j.
- I = Ensemble de pays, J = Ensemble de postes de dépenses publiques (éducation, police, culture, etc.), x_{ij} étant la dépense du pays i pour le poste j en 1988.
- I = J = Ensemble de pays, x_{ij} étant le total des exportations de i vers j en 1912.

Dans certains cas, le choix entre ce qui sera l'ensemble des individus et celui des variables peut sembler indifférent (dernier exemple), il faut toutefois le préciser clairement car, en ACP, les individus et les variables ne sont pas traités de manière équivalente.

On appelle nuage (des individus), l'ensemble des lignes i considérées comme points de l'espace vectoriel R^p. On note que la coordonnée de l'individu i sur l'axe canonique j de R^p est la valeur x_{ij} prise par le caractère j pour cet individu; en ce sens les axes canoniques correspondent aux variables.

Principes de l'ACP

L'idée de l'ACP est de déterminer un nouveau repère de R^p associé de manière naturelle à la structure du nuage considéré, de façon à pouvoir l'y examiner plus commodément.

Pour s'affranchir des effets d'échelle dus à l'hétérogénéité éventuelle des variables, ces dernières sont en général normalisées, c'est à dire que chaque colonne est divisée par son écart-type; toutes sont dès lors exprimées dans la même échelle standard.

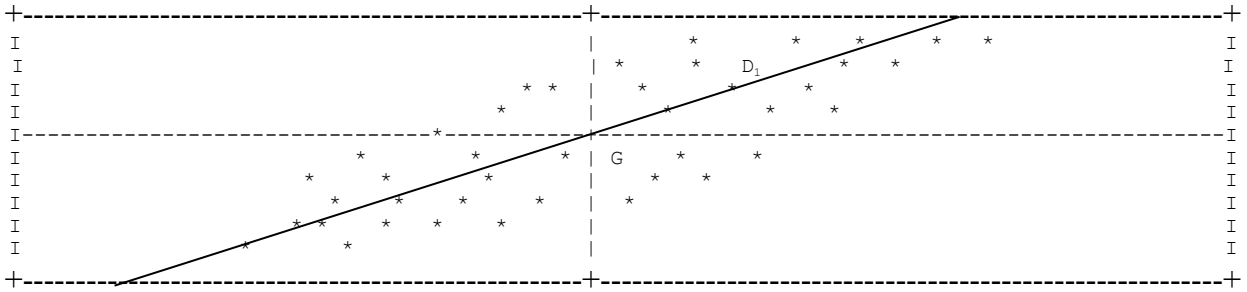
D'autre part, l'origine est placée au centre de gravité du nuage.

C'est le nuage ainsi transformé qui est en fait considéré; l'utilisateur n'a cependant pas à se préoccuper de ces transformations préalables, sauf demande contraire elles sont exécutées automatiquement par les logiciels d'ACP.

Directions principales - plans principaux - représentation des individus

Le nuage présente généralement des directions d'allongement privilégiées, celle d'allongement maximal D_1 est dite ***première direction principale*** (du nuage), la suivante D_2 parmi toutes celles perpendiculaires à D_1 est la ***seconde direction principale***, la suivante D_3 parmi toutes celles perpendiculaires à D_1 et D_2 est la ***troisième direction principale***, etc.

On choisit un vecteur unitaire u_k sur chaque direction D_k (le choix du sens est libre et décidé arbitrairement par le logiciel utilisé) et on obtient une base orthonormée de R^p, c'est la ***base principale*** du nuage.



On appelle ***plan principal*** ixj le plan vectoriel déterminé par les directions D_i et D_j. En général, le nuage est approximativement situé dans un sous-espace de R^p de faible dimension, engendré par les premières directions principales; l'examen de ses projections sur quelques plans principaux bien choisis (1x2, 1x3, etc.) permet alors de découvrir ses particularités et de décrire sa structure assez précisément.

Composantes principales - représentation des variables

De même que les variables initiales sont associées aux axes canoniques de R^p, de nouvelles variables appelées ***composantes principales*** sont associées aux axes principaux: la composante principale c_k est le vecteur de R^n qui donne les coordonnées des individus sur l'axe principal D_k muni du vecteur unitaire u_k.

Les composantes principales sont naturellement des combinaisons linéaires des variables initiales, on montre qu'elles sont centrées et non corrélées.

L'examen des corrélations entre les variables initiales et les composantes principales permet d'interpréter ces dernières et les axes principaux correspondants.

Les programmes usuels permettent de représenter ces quantités dans le ***cercle des corrélations***. Cette représentation n'est pas de même nature que celle des individus sur les plans principaux. Et si certains logiciels superposent les deux sur les mêmes graphiques, il faut garder à l'esprit que la position des points-variables par rapport aux points-individus n'y est pas directement interprétable!

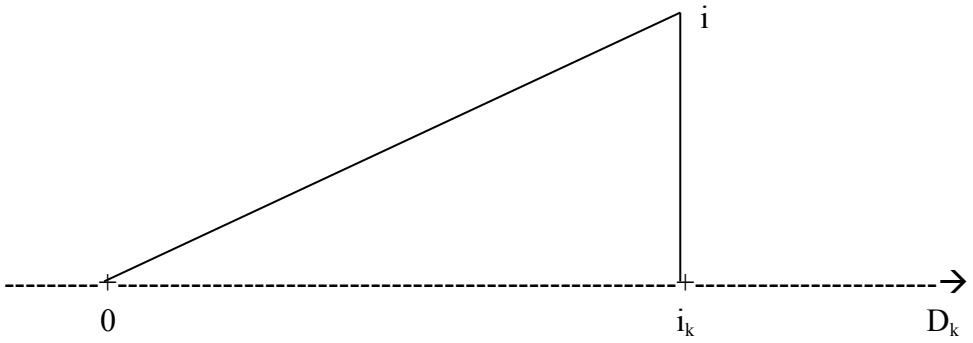
Une présentation alternative de l'ACP, privilégiant les variables mais équivalente, l'introduit comme la recherche de nouvelles variables (les composantes principales) non corrélées entre elles, et les plus corrélées avec l'ensemble des variables initiales.

Les composantes principales sont parfois vues comme des variables cachées non-observables, que la méthode permet donc de mettre en évidence derrière les variables initiales, seules observables. Elles permettent par ailleurs de résumer, par les premières d'entre elles, une information répartie sur un grand nombre de variables (cela est parfois utilisé en régression linéaire pour échapper à la multicolinéarité).

Aides à l'interprétation

Même si tout n'a pas été détaillé, on aura compris que les bases mathématiques de l'ACP sont réduites. L'art de l'analyste est celui de l'interprétation des résultats, cela nécessite la fois la compréhension des méthodes employées et la connaissance du domaine des données étudiées.

L'interprétation s'appuie sur l'examen de différentes quantités calculées et éditées par les logiciels d'ACP.



Inertie

Un individu i du nuage (supposé muni des poids uniformes $p_i = 1$) a une ***inertie*** I(i):

$$I(i) = p_i \, Oi^2 = Oi^2$$

Si i_k est la projection de i sur l'axe principal D_k, l'inertie de i suivant cet axe est:

$$I_k(i) = p_i \, Oi_k^2 = Oi_k^2$$

L'inertie de i se décompose en la somme de ses inerties suivant les différents axes principaux D_k (perpendiculaires):

$$I(i) = \sum_k I_k(i)$$

L'inertie totale suivant l'axe D_k, est:

$$I_k = \sum_i I_k(i)$$

Et l'inertie totale du nuage est:

$$I = \sum_i I(i) = \sum_k I_k$$

Les directions principales d'allongement du nuage sont en fait les directions perpendiculaires successives d'inertie maximum du nuage.

Taux d'inertie

Il s'agit des inerties successives I_1, I_2, I_3, etc. suivant les axes principaux D_1, D_2, D_3, etc. du nuage. Leurs valeurs relatives traduisent l'importance de l'allongement suivant ces directions successives.

On édite les taux relatifs I_1/I, I_2/I, I_3/I, etc., ainsi que les taux relatifs cumulés. Lorsque ces derniers approchent 100%, on considère que l'on a assez d'axes principaux pour représenter convenablement le nuage.

Contributions des axes aux individus (COR)

Il s'agit des ratios tels que:

$$COR(k, i) = I_k(i)/I(i)$$

qui mesure la qualité de la représentation de l'individu i sur l'axe principal D_k.

On a: $\sum_k COR(k, i) = 1$

Il n'est licite de commenter la position de l'individu i sur le plan principal kxh que si le ratio:

$$[I_k(i)+I_h(i)]/I(i)$$

n'est pas trop faible.

La considération de ces ratios, qui sont des *cosinus carrés*, n'est pertinente que pour les points pas trop proches de l'origine. Pour ceux-ci, c'est plus leur position, centrale, que la direction dans laquelle se manifeste leur faible éloignement, qui les caractérise.

Contributions des individus aux axes (CTR)

Il s'agit des ratios tels que:

$$CTR(i, k) = I_k(i)/I_k$$

qui mesure la part prise par l'individu i dans la détermination de l'axe principal D_k.

On a: $\sum_i CTR(i, k) = 1$

Contributions des axes aux variables (COR)

Il s'agit des coefficients de corrélation au carré tels que:

$$COR(k, j) = corr^2(c_k, x_{.j})$$

entre la variable initiale $x_{.j}$ et la composante principale c_k. Elles permettent comme on l'a vu de dégager la signification des axes.

On a: $\sum_k COR(k, j) = 1$

Ces quantités sont les carrés de celles figurées dans le cercle des corrélations utilisé pour représenter graphiquement les variables.

Contributions des variables aux axes (CTR)

Il s'agit des ratios tels que:

$$CTR(j, k) = corr^2(c_k, x_{.j}) / \sum_i corr^2(c_k, x_{.i})$$

On a: $\sum_j CTR(j, k) = 1$

L'observation des premiers plans principaux ne permet aucune conclusion, et peut même être source de contresens, si elle ne s'accompagne pas de l'examen des quantités précédentes. Il faut donc toujours les faire éditer par le logiciel utilisé et les consulter.

Éléments supplémentaires

Si on craint que l'influence de certains individus ne soit excessive pour la détermination des axes principaux, il est possible de les placer en ***éléments supplémentaires***, c'est à dire qu'ils ne font pas partie du nuage dont on cherche les directions principales, mais on peut figurer leur position sur les plans principaux obtenus.

On traite de la même manière des variables en éléments supplémentaires, elles ne font pas partie de l'ensemble des variables de base mais on peut examiner leurs corrélations avec les composantes principales obtenues.

Après une première ACP des données étudiées, il est recommandé d'éprouver la stabilité des configurations observées en effectuant de nouvelles analyses laissant en éléments supplémentaires les individus ou variables d'importance trop marquée, ou encore les données douteuses.

Rotations

Si globalement l'ACP détermine via les premières directions principales des sous-espaces de faibles dimensions dans lequel l'essentiel de l'information portée par le nuage des individus se manifeste, il est fréquent que l'interprétation des nouveaux axes, c'est à dire encore des nouvelles variables ou composantes principales, soit malaisée. Les choses seraient plus simples si chaque nouvelle variable était bien corrélée avec un groupe de variables initiales (elles-mêmes plus ou moins liées) et peu avec les autres, ces groupes étant naturellement exclusifs. C'est l'idée qui a inspiré la méthode dite des ***rotations***.

On fixe d'abord le nombre de directions propres retenues (3, 4, 5...) selon la pratique habituelle par l'examen des valeurs propres, ou taux d'inertie, successives, puis on cherche la rotation des axes principaux, conservant donc leur orthogonalité, qui approche au mieux la situation désirée précédente.

Le critère mathématique généralement retenu est celui dit du ***varimax***, qui cherche à maximiser la variance de la série des corrélations au carré avec les variables initiales. Comme on le conçoit, celles-ci sont entrainées soit vers 1 soit vers 0, valeurs les plus éloignées, et permettent donc une interprétation plus aisée de ces nouveaux axes après rotation. Comme les composantes principales, les nouvelles variables sont non corrélées.

Le calcul effectif mis en œuvre par les logiciels offrant ces option est une classique optimisation sous contraintes.

Conclusion

L'ACP est une technique de statistique descriptive dont le principe est simple mais qui met en oeuvre des calculs numériques importants, pour cette raison elle n'a pu se développer qu'avec l'apparition des ordinateurs.

L'ACP est à conseiller pour un premier examen, une mise en forme ou une présentation synthétique de données abondantes croisant des individus avec des variables quantitatives. On n'omettra cependant pas d'examiner préalablement les données par les méthodes statistiques usuelles (moyenne, écart-type, graphiques, corrélation, etc.).

Un reproche fréquemment adressé à l'ACP et aux techniques connexes est qu'elles ne révéleraient que des évidences. Le propos est injuste, mais il est rassurant que souvent les premiers axes retrouvent et confirment ce qui était déjà connu.

Comme avec les autres méthodes descriptives, il faut être très prudent pour inférer des modèles explicatifs ou causals à partir des configurations obtenues.

----===oo**Ω**oo===----

Appendice mathématique

Formalisation de l'ACP

On note X la matrice n.p des données (ie portant les observations en ligne, éléments de R^p, et les variables, quantitatives, en colonnes, éléments de R^n), on suppose les colonnes de X préalablement centrées et réduites si nécessaire.

Soit u un vecteur (en colonne) unitaire de R^p, le vecteur X.u de R^n a pour composantes les produits scalaires des observations avec u, c'est à dire encore, les distances à l'origine des projections des observations selon la direction de u, tandis que l'inertie totale du nuage dans cette direction est donnée par le produit matriciel : u'.X'.X.u .

La matrice symétrique X'.X est la ***matrice d'inertie*** du nuage, tandis que le produit u'.X'.X.u , qui donne l'inertie dans cette direction, est l'application de la forme bilinéaire symétrique de matrice X'.X au vecteur unitaire u. On remarque que X'.X est simplement, au facteur 1/n près, la matrice des corrélations entre les variables-colonnes initiales (ou des covariances si on effectue une ACP non normée).

La recherche des directions principales, c'est à dire des directions successives d'inertie maximale du nuage, se traduit donc par le problème de maximisation sous contrainte :

$$\max_{u_k} (u_k'.X'.X.u_k) \quad \text{avec} \quad u_k'.u_k = 1$$

les vecteurs u_k successifs devant en outre être orthogonaux.

L'algébre linéaire enseigne que les vecteurs propres normés successifs : u_k , associés à la suite décroissante des valeurs propres (positives) de X'.X : λ_k , apportent la solution du problème, la valeur propre λ_k mesurant l'inertie dans la k-ième direction principale u_k :

$$u_k'.X'.X.u_k = \lambda_k.u_k'.u_k = \lambda_k$$

Les vecteurs $c_k = X.u_k$ de R^n sont les *composantes principales* successives du nuage, centrées, de variances respectives λ_k/n et non corrélées (de covariances : $c_k'.c_h/n = \lambda_h.u_k'.u_h/n$, nulles), ce sont les « nouvelles variables », dont les composantes donnent les coordonnées des points du nuage sur les axes factoriels.

Les diverses contributions, corrélations et autres aides à l'interprétation, enfin, sont aisées à écrire, en fonction des λ_k, u_i et c_j . Ainsi, par exemple, la contribution de l'observation i à l'axe k est : $c_k(i)^2/\lambda_k$, où $c_k(i)$ désigne la i-ème composante de c_k…

La présentation de l 'ACP par les variables conduit par une autre voie au même problème mathématique : on cherche de nouvelles variables combinaisons linéaires des anciennes variables, non corrélées entre elles, et les plus corrélées possible avec l'ensemble de ces variables initiales, plus exactement telles que la somme des carrés des corrélations avec les anciennes variables soit maximale.

Soit $y = X.v$ une telle nouvelle variable supposée normalisée, c'est à dire telle que $y'.y = 1$, le produit $X'.y$ est alors le vecteur des corrélations avec les anciennes variables, tandis que le scalaire $y'.X.X'.y$ est la somme des corrélations au carré que l'on veut maximiser.

L'algèbre linéaire dit encore que les vecteurs propres normés successifs : y_k , associés à la suite décroissante des valeurs propres (positives) de X.X' : μ_k , apportent la solution du problème, la valeur propre μ_k mesurant la corrélation totale maximisée pour la variable y_k.

De plus, les valeurs propres de $X'.X$ et de $X.X'$ sont les mêmes : $\lambda_k = \mu_k$, tandis que les vecteurs propres se déduisent par application de la matrice X (ou X'), ainsi les nouvelles variables y_k sont simplement les composantes principales $c_k = X.u_k$ normalisées.

Une autre présentation encore équivalente est par la recherche de nouvelles variables combinaisons linéaires normalisées des anciennes, non corrélées et de dispersion ou de variance maximale.

----===**O**===----

(21.07.2009)