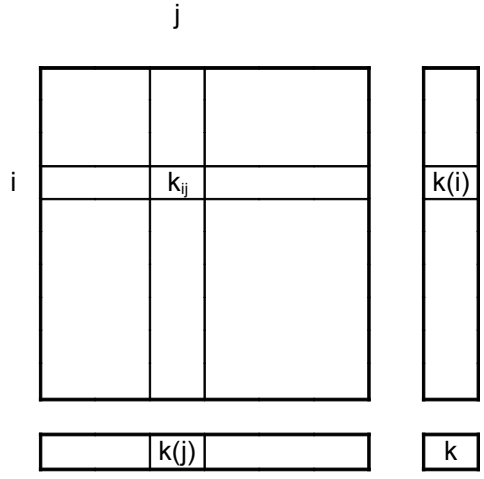
L'Analyse Factorielle des Correspondances (AFC)

Les tableaux de contingence

L'AFC est une forme particulière de l'ACP, adaptée au traitement de certain types de tableaux rectangulaires de données: les ***tableaux de contingence***.

Exemple: la population des reçus au baccalauréat en 1976 est classée selon les diverses combinaisons possibles de modalités des deux caractères: série et région.
Notant par exemple la région en ligne, à l'intersection de la ligne: "Limousin", et de la colonne: "série C", figure le nombre: 386, de reçus au bac C dans le Limousin cette année-là.

Définition: plus généralement, soit P, une population examinée suivant deux caractères, l'un prenant un ensemble: I, de n modalités exclusives: i, et l'autre un ensemble: J, de p modalités exclusives: j, le tableau de contingence: K (ou ***tableau croisé***), associé à ces données est le tableau rectangulaire de dimensions nxp et de terme général k_{ij} : le nombre d'individus présentant simultanément la modalité i pour le premier caractère, et j pour le second.



On a coutume d'indiquer dans une colonne et une ligne supplémentaires, les totaux des lignes et des colonnes correspondantes, ou ***totaux marginaux***, donnant la distribution en effectifs de la population selon les items de chacune des variables considérée isolément :

$$k(i) = \sum_j k_{ij}$$

et

$$k(j) = \sum_i k_{ij}$$

par suite

$$\sum_i k(i) = \sum_j k(j) = k$$

k étant l'effectif total de la population observée.

Autres exemples de tableaux de contingence:

- Chômeurs ventilés selon le département d'inscription croisé avec la dernière profession déclarée.
- Citoyens répartis suivant l'appartenance politique avouée croisée avec la tranche d'age.
- Militaires en fin de carrière classés selon le grade croisé avec la CSP du père.

Les principes de l'AFC

Analyse du nuage des profils en ligne

À partir du tableau de contingence: K, on passe au tableau: X, des ***profils en ligne***, c'est à dire des ***fréquences conditionnelles***, obtenu en divisant chaque élément de la ligne i, par son total k(i):

$$x_{ij} = k_{ij}/k(i)$$

Ainsi la ligne i du tableau X, donne la répartition en proportions selon les p modalités du second caractère des k(i) individus qui prennent la modalité i du premier caractère.

Le terme de *profil* fait référence à l'aspect visuel des diagrammes en rectangles fréquemment utilisés pour représenter les distributions, ici celles des fréquences conditionnelles indiquées.

Les lignes de X étant donc des fréquences, ou encore des ***probabilités empiriques***, l'ACP du nuage de ses n lignes, éléments de R^p, est effectuée non selon la distance usuelle, mais selon la distance, dite du ***chi-deux***, adaptée à la comparaison de fréquences ou de probabilités:

$$\text{distance } (i, i') = \sum_j (x_{ij} - x_{i'j})^2/k(j) \qquad \text{(à un facteur multiplicatif près)}$$

D'autre part, pour tenir compte de l'importance relative des différentes modalités en ligne, la ligne-individu i est affectée du poids: k(i).

L'AFC du tableau de contingence: K, est l'ACP du tableau: X, des profils en ligne, avec la distance du chi-deux et les poids précédents.

Les différentes notions présentées en ACP:

- les taux d'inertie;
- les coordonnées et plans factoriels (ou principaux);
- les contributions;
- les éléments supplémentaires.

se retrouvent ici; elles conservent la même signification et s'utilisent de la même manière qu'en ACP.

Analyse des profils en colonne, représentation simultanée

À l'inverse de ce qui se passe en ACP, lignes et colonnes d'un tableau de contingence: K, jouent des rôles comparables.

À l'analyse du tableau: X, des profils en ligne correspond celle du tableau: Y, des ***profils en colonnes*** - avec la distance et les poids convenables.

Du fait de relations algébriques particulières, les deux analyses peuvent être menées conjointement. Elles présentent en outre une propriété remarquable: sans entrer dans le détail mathématique, il est loisible de représenter les deux ensembles: I et J, simultanément sur les plans factoriels, de telle sorte que la position d'un point de l'ensemble I (resp. J) y est interprétable par rapport à l'ensemble de tous les points de l'ensemble J (resp. I).

Pour cette raison, la plupart des logiciels éditent la représentation simultanée des deux ensembles dans les plans principaux demandés.

On retiendra impérativement que cette interprétation reste délicate, et que la proximité entre un point-ligne et un point-colonne considérés isolément n'a aucune signification.

Extensions de la notion de tableau de contingence, questionnaires

On a présenté l'AFC d'un tableau de contingence, mais il est d'autres tableaux que peut traiter l'AFC.

Correspondances multiples

Les modalités d'un caractère peuvent être croisées avec celles de deux autres caractères, ou davantage.

Exemples:

- Chômeurs selon le département, croisé avec la profession et la classe d'age;
- Baccalauréat selon la région croisée avec la section et la mention obtenue;
- Militaires selon le grade croisé avec la CSP paternelle et le niveau d'étude.

On peut aussi mettre "bout-à-bout" des tableaux de contingence observés à des périodes différentes, et faire l'AFC du tableau obtenu.

Ex: Chômeurs par région croisée avec la profession en 1975, en 1981, en 1988 et en 1994.

« Questionnaires »

Les résultats d'une enquête par questionnaire peuvent se condenser en un tableau logique (c'est à dire composé de 0 et de 1): K, du type suivant:

J_q

i		0	1	0	0	0	
i'		0	0	0	0	1	

Les lignes désignent l'ensemble: I, des individus interrogés.

À chaque question: q, correspond un ensemble J_q de colonnes associées aux différentes réponses possibles à q.

La ligne i porte un 1 dans chaque groupe J_q marquant la réponse de l'individu i à la question q, et des 0 ailleurs.

On exige en général que les réponses à chaque question soient exclusives, et incluent tous les cas possibles (quitte à prévoir une colonne "non réponse"), à ce moment dans une ligne donnée, il y a un 1 unique par groupe J_q de colonnes, et le tableau K est dit ***disjonctif complet***.

L'analyse de tableaux disjonctifs complets est commune en sociologie et en sciences humaines.

On montre que l'AFC du tableau disjonctif complet: K, donne les mêmes résultats que celle du tableau de contingence multiple: K'.K , qui croise les caractères entre eux. Mais cette dernière analyse perd l'information sur les individus.

Dans l'analyse des questionnaires, il est souvent commode de distinguer deux groupes de questions: les unes, propres au domaine étudié, sont placées en éléments principaux et servent à structurer les données, les autres, d'ordre général (ex: les caractéristiques socio-professionnelles de la population interrogée), sont mises en supplémentaires et permettent d'identifier les déterminants des variables étudiées.

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

Appendice mathématique

Formalisation de l'AFC

L'AFC, comme il a été dit, est une forme particulière de l'ACP appliquée aux tableaux de contingence : non centrée-réduite, avec pondérations, et utilisant la *métrique* dite du *chi-deux* (ie des inverses des fréquences marginales) au lieu de la métrique euclidienne usuelle.

On note K le tableau de contingence, ou tableau croisé initial, de dimension n.p, $F_{J/I}$ le tableau des profils en ligne (fréquences conditionnelles, conditionnées par les items en ligne) et $F_{I/J}$ celui des profils en colonne. D_I désigne la matrice diagonale portant sur sa diagonale les totaux en ligne (ou totaux marginaux) et D_J celle des totaux en colonne.

Les différentes matrices précédentes sont naturellement liées:

$$F_{J/I} = D_I^{-1}.K \quad \text{et} \quad F_{I/J} = K.D_J^{-1}$$

Le produit scalaire de deux vecteurs u et v dans R^p pour la métrique du chi-deux est donné par le produit matriciel $u'.D_J^{-1}.v$ à un facteur multiplicatif près, par suite l'inertie dans la direction du vecteur D_J^{-1}-unitaire u du nuage des profils en ligne, pour la métrique précédente avec pour pondérations les totaux en ligne, est donnée, au même facteur près au carré, par le produit matriciel:

$$u'.D_J^{-1}.F_{J/I}'.D_I.F_{J/I}.D_J^{-1}.u = u'.D_J^{-1}.K'.D_I^{-1}.K.D_J^{-1}.u$$

Les directions principales d'inertie sont obtenues en maximisant la quantité précédente sous la contrainte $u'.D_J^{-1}.u = 1$ dans des directions D_J^{-1}-orthogonales successives. La théorie indique que la solution est la suite des vecteurs propres D_J^{-1}-normés u_k associée à la suite décroissante des valeurs propres λ_k de la matrice (non symétrique):

$$F_{J/I}'.D_I.F_{J/I}.D_J^{-1} = K'.D_I^{-1}.K.D_J^{-1}$$

En fait, on obtient une première valeur propre *triviale* égale à 1 associée au vecteur des totaux marginaux en colonnes, elle disparaît si on traite le nuage pondéré centré et on ne la retient pas.

Les composantes principales :

$$c_k = F_{J/I}.D_J^{-1}.u_k = D_I^{-1}.K.D_J^{-1}.u_k$$

donnent à nouveau les coordonnées des profils en ligne sur les axes factoriels, tandis que les différentes aides à l'interprétations s'obtiennent aisément en tenant compte de la métrique D_J^{-1} et des pondérations données par D_I.

L'analyse des profils en colonne est étroitement liée à la précédente, du fait des relations entre $F_{I/J}$ et $F_{J/I}$. Les directions principales de cette analyse sont données par les vecteurs propres D_I^{-1}-normés et orthogonaux:

$$v_k = \lambda_k^{-1/2}.K.D_J^{-1}.u_k$$

de la matrice:

$$F_{I/J}.D_J.F_{I/J}'.D_I^{-1} = K.D_J^{-1}.K'.D_I^{-1}$$

et les composantes principales par:

$$d_k = F_{I/J}'.D_I^{-1}.v_k = \lambda_k^{-1/2}.F_{I/J}'.c_k$$

La j-ième composante: $d_k(j)$, de d_k est donc:

$$d_k(j) = \lambda_k^{-1/2}. \sum_i (n_{ij}/n_{.j}).c_k(i)$$

relation barycentrique, au facteur $\lambda_k^{-1/2}$ près, qui relie les deux analyses et justifie la représentation simultanée.

----==ooOoo==----

(14.03.2009)