L'ANALYSE DES CORRESPONDANCES : SES PROPRIETES ET SES EXTENSIONS

Y. ESCOUFIER
Unité de Biométrie
9, place P. Viala
34060 MONTPELLIER CEDEX
France

I - INTRODUCTION : RAPPEL SUR L'ANALYSE DES CORRESPONDANCES.

Le but de ce travail est de rappeler ou de construire un certain nombre d'extensions de l'Analyse des Correspondances (A.C.) d'un tableau de contingence. Le mécanisme de l'obtention de ces extensions consiste à expliciter les propriétés reconnues de l'A.C. en mettant clairement en évidence les propriétés mathématiques qui les sous-tendent. Il est alors possible d'utiliser des constructions analogues qui conduiront à des propriétés voisines susceptibles d'être plus intéressantes dans certaines situations pratiques.

Une présentation rapide de l'A.C. adaptée à l'objectif poursuivi est faite. D'autres présentations plus riches en détails pourront être trouvées soit dans l'un ou l'autre des nombreux livres écrits en Français sur l'Analyse des Données soit dans les ouvrages de S. NISHISATO (1980), M.J. GREENACRE (1984) ou encore l'article de M.O. HILL (1974). On retiendra également la version anglaise du livre de L. LEBART (1984).

I.1. Considérons tout d'abord un tableau de données quantitatives X, nxp, représentant les mesures faites sur n individus pour p variables. Associons à X une matrice symétrique définie positive Q, pxp, nécessaire au calcul des distances entre individus dans $\mathbb{R}^p$ et une matrice diagonale positive D dont les éléments diagonaux représentent les poids attachés à chacun des individus. Admettons de noter ${}^tX$ la transposée de la matrice X et soit $\delta_{\alpha\alpha'}$ le symbole de Kronecker qui vaut 1 si $\alpha = \alpha'$ et 0 sinon.

<u>Définition 1</u> : On appelle Analyse en Composantes Principales (ACP) d'ordre $k \leqslant v = \min(n,p)$ du triplet (X, Q, D), la recherche des triplets $((\lambda_\alpha, \psi_\alpha, \varphi_\alpha)\ ;\ \alpha = 1, \ldots, k)$ définis par :

1) $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_k$

2) $X\,Q\,{}^tX\,D\,\psi_\alpha = \lambda_\alpha\,\psi_\alpha$ avec ${}^t\psi_\alpha\,D\,\psi_{\alpha'} = \delta_{\alpha\alpha'}$

3) ${}^tX\,D\,X\,Q\,\varphi_\alpha = \lambda_\alpha\,\varphi_\alpha$ avec ${}^t\varphi_\alpha\,Q\,\varphi_{\alpha'} = \delta_{\alpha\alpha'}$

Remarque : les présentations courantes de l'ACP, telles qu'on les trouve par exemple dans les logiciels de grande diffusion, font des choix pour X, Q, D. L'ACP dite "sur matrice de variance" donne à tous les éléments diagonaux de D la même valeur 1/n ; X est alors la matrice des données centrées pour ces poids ; Q est la matrice identité de $\mathbb{R}^p$. Dans l'ACP "sur matrice de corrélation", X est la matrice des données centrées et réduites ; Q et D sont les mêmes que dans l'ACP sur matrice de variance. Ces choix sont justifiés pour les interprétations en termes de covariances et de corrélations qu'ils permettent ; le centrage des données qui place l'origine des axes de $\mathbb{R}^p$ au point moyen du nuage des individus, assure à ce dernier une inertie minimale : il n'en reste pas moins que de nombreuses propriétés des solutions ne sont pas liées à ces choix. Une seule sera explicitée ici pour l'usage qu'il en sera fait ultérieurement. Un exposé plus complet sur ce sujet peut être trouvé dans M. OKAMOTO (1969) ou R. SABATIER et al. (1984).

Propriété 1 (Formule de reconstitution des données)

Pour $k \leqslant v$, posons $\tilde{X}_{(k)} = \sum_{\alpha=1}^{k} \sqrt{\lambda_\alpha}\, \psi_\alpha\, {}^t\varphi_\alpha$ alors pour toute matrice $U_{(k)}$, nxp, on a :

$$\mathrm{Tr}[(X - U_{(k)})\, Q\, {}^t(X - U_{(k)})\, D] \geqslant \mathrm{Tr}[(X - \tilde{X}_{(k)})\, Q\, {}^t(X - \tilde{X}_{(k)})\, D] = \sum_{\alpha=k+1}^{v} \lambda_\alpha$$

La démonstration de cette propriété n'est pas immédiate. Elle demande de considérer l'application linéaire X de l'espace euclidien $\mathbb{R}^p$ muni de la forme bilinéaire Q dans l'espace euclidien $\mathbb{R}^n$ muni de D. On peut alors montrer (R. SABATIER et al. (1984)) que pour $i \leqslant v-k$ la valeur singulière de rang i de $X - U_{(k)}$ qui est la racine carrée de la valeur propre de rang i de $(X - U_{(k)})\, Q\, {}^t(X - U_{(k)})\, D$ est inférieure ou égale à $\sqrt{\lambda_{i+k}}$ qui s'obtient pour $U_{(k)} = \tilde{X}_{(k)}$ d'où le résultat.

Remarque. Notons $Q_{jj'}$ les éléments de Q, D_{ii} les éléments diagonaux de D et $(X - \tilde{X}_{(k)})_{ij}$ ceux de $X - \tilde{X}_{(k)}$. On a :

$$\sum_{\alpha=k+1}^{p} \lambda_\alpha = \sum_{i=1}^{n} \sum_{j=1}^{p} \sum_{j'=1}^{p} Q_{jj'}\, (X - \tilde{X}_{(k)})_{ij}\, (X - \tilde{X}_{(k)})_{ij'}\, D_{ii}$$

Si Q est diagonale alors :

$$4) \quad \sum_{\alpha=k+1}^{p} \lambda_\alpha = \sum_{i=1}^{n} \sum_{j=1}^{p} Q_{jj}\, (X - \tilde{X}_{(k)})_{ij}^2\, D_{ii}$$

Sous cette forme, on voit que la propriété 1 s'interprète comme une propriété d'approximation de X par $\tilde{X}_{(k)}$ au sens des moindres carrés pour des pondérations définies par Q_{jj} et D_{ii}.

I.2. Considérons maintenant un tableau de contingence P, nxp, $\left(\sum_{i=1}^{n}\sum_{j=1}^{p} P_{ij} = 1\right)$. P croise les n modalités d'une variable I avec les p modalités d'une variable J. Soit D_J la matrice diagonale pxp des poids marginaux des colonnes et D_I, nxn, la matrice diagonale des poids marginaux des lignes. Notons $\underset{\sim}{1}_n$ et $\underset{\sim}{1}_p$ les vecteurs de $\mathbb{R}^n$ et $\mathbb{R}^p$ dont toutes les composantes sont égales à l'unité. Soit $k \leqslant v$ et $X = D_I^{-1} (P - D_I \underset{\sim}{1}_n \, {}^t\underset{\sim}{1}_p \, D_J) \, D_J$

Définition 2 : On appelle analyse des correspondances d'ordre k de P, l'ACP d'ordre k de (X, D_J, D_I).

Enumérons quelques unes des propriétés de cette analyse.

Propriété 2

$$d^2(i,i') = \sum_{j=1}^{p} \frac{(P(j/i) - P(j/i'))^2}{P_{.j}}$$

$$d^2(j,j') = \sum_{i=1}^{n} \frac{(P(i/j) - P(i/j'))^2}{P_{i.}}$$

Ainsi le carré de la distance entre deux lignes de X calculé pour la matrice définie positive D_J revient à calculer entre les lois conditionnelles de J la distance du χ^2 de centre la loi marginale. On a un résultat symétrique pour les colonnes de X.

Propriété 3

$$\sum_{\alpha=1}^{v} \lambda_\alpha = \sum_{i=1}^{n}\sum_{j=1}^{p} \frac{(P_{ij} - P_{i.} P_{.j})^2}{P_{i.} P_{.j}}$$

La démonstration se déduit de l'énoncé 2) de la Définition 1 en écrivant, dans ce cas particulier, que

$$\sum_{\alpha=1}^{v} \lambda_\alpha = Tr(X \, D_J \, {}^tX \, D_I)$$

On est donc conduit à considérer l'AC comme une méthode d'exploration de l'écart entre le tableau observé et celui qui découlerait de l'indépendance.

Propriété 4

$${}^t\underset{\sim}{1}_n \, D_I \, \psi_\alpha = {}^t\underset{\sim}{1}_p \, D_J \, \varphi_\alpha = 0$$

Le calcul permet de vérifier que $\underset{\sim}{1}_n$ et $\underset{\sim}{1}_p$ sont respectivement vecteur propre de $XD_J\,{}^tX\,D_I$ et ${}^tX\,D_I\,X\,D_J$ pour la valeur propre O. Les vecteurs propres ϕ_α de $X\,D_J\,{}^tX\,D_I$ sont donc orthogonaux pour D_I à $\underset{\sim}{1}_n$ et de même les φ_α, vecteurs propres de ${}^tX\,D_I\,X\,D_J$, pour D_J à $\underset{\sim}{1}_p$. Les ϕ_α fournissent donc une représentation centrée des modalités de I pour les poids D_I. La représentation des modalités de J fournie pour les φ_α est centrée pour D_J.

Propriété 5 (Propriété du centre de gravité)

Pour $\lambda_\alpha \neq 0$

$$\phi_\alpha = \frac{D_I^{-1}\,P\,\varphi_\alpha}{\sqrt{\lambda_\alpha}} \qquad \text{et} \qquad \varphi_\alpha = \frac{D_J^{-1}\,{}^tP\,\phi_\alpha}{\sqrt{\lambda_\alpha}}$$

Le rapprochement des équations 2) et 3) de la Définition 1 permet d'obtenir le résultat général :

$$\sqrt{\lambda_\alpha}\,\phi_\alpha = X\,Q\,\varphi_\alpha \qquad \text{et} \qquad \sqrt{\lambda_\alpha}\,\varphi_\alpha = {}^tX\,D\,\phi_\alpha .$$

La propriété 5 en est la particularisation à l'A.C. On voit que la $i^{\text{ième}}$ coordonnée de ϕ_α est, au facteur $\sqrt{\lambda_\alpha}$ près, le centre de gravité des $\varphi_{\alpha j}$ affectés des poids $P_{ij}/P_{i.}$. On a un résultat symétrique pour $\varphi_{\alpha j}$.

Propriété 6 (Reconstitution de P)

Soit $$\tilde{P}_{(k)} = D_I \left(\underset{\sim}{1}_n\,{}^t\underset{\sim}{1}_p + \sum_{\alpha=1}^{k} \sqrt{\lambda_\alpha}\,\phi_\alpha\,\varphi_\alpha\right) D_J$$

Alors (5) $$\tilde{P}_{(k)}\,\underset{\sim}{1}_p = D_I\,{}^t\underset{\sim}{1}_n$$

$${}^t\tilde{P}_{(k)}\,\underset{\sim}{1}_n = D_J\,{}^t\underset{\sim}{1}_p$$

et (6) $$\sum_{i=1}^{n}\sum_{j=1}^{p} \frac{(P - \tilde{P}_{(k)})^2_{ij}}{P_{i.}\,P_{.j}} = \sum_{\alpha=k+1}^{v} \lambda_\alpha$$

Les résultats (5) se déduisent simplement de la propriété 4, ils signifient que $\tilde{P}_{(k)}$ a les mêmes marges que P.

Pour démontrer (6) remarquons que

$$D_I^{-1}\,(P - \tilde{P}_{(k)})\,D_J^{-1} = X - \tilde{X}_{(k)}$$

$$\text{or} \quad \sum_{i=1}^{n}\sum_{j=1}^{p} \frac{(P - \tilde{P}_{(k)})^2_{ij}}{P_{i.}\, P_{.j}} = \sum_{i=1}^{n}\sum_{j=1}^{p} P_{i.} \left(\frac{(P - \tilde{P}_{(k)})_{ij}}{P_{i.}\, P_{.j}} \right)^2 P_{.j}$$

si bien que (6) découle de la propriété 1.

$\tilde{P}_{(k)}$ est donc une approximation au sens des moindres carrés de P pour des pondérations définies par $\frac{1}{P_{i.}}$ et $\frac{1}{P_{.j}}$

Propriété 7 (Equivalence distributionnelle)

On ne change pas les résultats concernant les lignes de P en remplaçant dans P deux colonnes proportionnelles par leur somme.

Il résulte de la propriété 4 que pour tout $\psi_\alpha \neq \underset{\sim}{1}_n$

$$\lambda_\alpha \psi_\alpha = X\, D_J\, {}^tX\, D_I\, \psi_\alpha = D_I^{-1}\, P\, D_J^{-1}\, {}^tP\, \psi_\alpha$$

$$\text{or} \quad (P\, D_J^{-1}\, {}^tP)_{ii'} = \sum_{j=1}^{p} \frac{P(i,j)\; P(i',j)}{P_{.j}} = \sum_{j=1}^{p} P_{.j}\, \frac{P(i,j)}{P_{.j}} \times \frac{P(i',j)}{P_{.j}}$$

Supposons alors que $\forall i = 1,\ldots,n \quad \frac{P(i,j_1)}{P_{.j_1}} = \frac{P(i,j_2)}{P_{.j_2}}$

La contribution des colonnes j_1 et j_2 à $(P\, D_J^{-1}\, {}^tP)_{ii'}$ est :

$$(P_{.j_1} + P_{.j_2}) \left(\frac{P(i,j_1)}{P_{.j_1}} \times \frac{P(i',j_1)}{P_{.j_1}} \right) = (P_{.j_1} + P_{.j_2}) \left(\frac{P(i,j_1)+P(i,j_2)}{P_{.j_1} + P_{.j_2}} \times \frac{P(i',j_1)+P(i',j_2)}{P_{.j_1} + P_{.j_2}} \right)$$

ce qui établit le résultat pour ψ_α.

$$\text{On en déduit : } X\, D_J\, {}^tX = \sum_{\alpha=1}^{v-1} \lambda_\alpha\, \psi_\alpha\, {}^t\psi_\alpha + 0 \times \underset{\sim}{1}_n\, {}^t\underset{\sim}{1}_n$$

$$= \sum_{\alpha=1}^{v-1} \lambda_\alpha\, \psi_\alpha\, {}^t\psi_\alpha$$

Les produits scalaires entre lignes ne dépendent donc pas de la modification ce qui entraîne la même propriété pour les distances entre lignes.

Remarque : on a bien sûr un résultat analogue pour les colonnes.

II - MODIFICATIONS DES CHOIX FAITS DANS L'ANALYSE DES CORRESPONDANCES.

Les propriétés que le chapitre I a reconnues à l'A.C. peuvent être considérées soit comme les énoncés des qualités de la méthode soit comme les énoncés de ses limites. Le second point de vue a encouragé des chercheurs à altérer les éléments de la définition 2 de façon à définir des méthodes proches de l'AC mais différentes par certains de leurs résultats.

Dans la suite, pour faciliter les rapprochements avec le paragraphe précédent nous avons adopté une numérotation des propriétés dont le premier chiffre est celui de la propriété correspondante de l'AC et le second le numéro donné à la définition. Les démonstrations sont omises quand elles sont calquées sur celles faites pour l'AC.

II.1. Altérations basées sur la signification de la somme des valeurs propres

II.1.1. Définition 3 On appellera Décomposition du critère de Belson, l'ACP du triplet (Y, I_{pxp}, I_{nxn}) où $Y = P - D_I \underset{\sim}{1}_n {}^t\underset{\sim}{1}_p D_J$.

Cette définition est justifiée par la propriété suivante qui fait le lien avec le critère proposé par W. BELSON (1959) cité par F. MARCOTORCHINO (1984).

Propriété 3.3 Dans l'ACP du triplet (Y, I_{pxp}, I_{nxn}), on a :

$$\sum_{\alpha=1}^{v} \lambda_\alpha = \sum_{i=1}^{n} \sum_{j=1}^{p} (P_{ij} - P_{i.}\ P_{.j})^2$$

On obtient le résultat en écrivant la trace des opérateurs $Y\,{}^tY$ et ${}^tYY$ à diagonaliser pour calculer les ψ_α et les φ_α. Remarquant que $Y \underset{\sim}{1}_p = {}^tY \underset{\sim}{1}_n = 0$, on vérifie aisément les deux propriétés suivantes :

Propriété 4.3 $${}^t\underset{\sim}{1}_n\ \psi_\alpha = {}^t\underset{\sim}{1}_p\ \varphi_\alpha = 0$$

Propriété 6.3

$$\text{Soit } \widetilde{P}_{(k)} = D_I \underset{\sim}{1}_n\ {}^t\underset{\sim}{1}_p\ D_J + \sum_{\alpha=1}^{k} \sqrt{\lambda_\alpha}\ \psi_\alpha\ \varphi_\alpha$$

$$\text{alors} \quad \widetilde{P}_{(k)} \underset{\sim}{1}_p = D_I \underset{\sim}{1}_n$$

$${}^t\widetilde{P}_{(k)} \underset{\sim}{1}_n = D_J \underset{\sim}{1}_p$$

$$\text{et} \quad \sum_{i=1}^{n} \sum_{j=1}^{p} (P - \widetilde{P}_{(k)})^2_{ij} = \sum_{\alpha=k+1}^{v} \lambda_\alpha$$

La décomposition du critère de Belson fournit donc des représentations centrées des lignes et des colonnes ainsi que des tableaux $\tilde{P}_{(k)}$ de mêmes marges que P.

II.1.2. Définition 4 Suivant LAURO et D'AMBRA (1983), on appelle Analyse non symétrique des correspondances l'ACP du triplet (Z, I_{pxp}, D_I) où $Z = D_I^{-1} (P - D_I \underset{\sim}{1}_n {}^t\underset{\sim}{1}_p D_J)$

Ces auteurs ont introduit cette analyse avec la volonté de substituer aux χ^2 le critère de Goodman-Kruskal.

La propriété suivante énonce ce résultat.

Propriété 3.4 Dans l'ACP du triplet (Z, I_{pxp}, D_I)

$$\sum_{\alpha=1}^{v} \lambda_\alpha = \sum_{i=1}^{n} \sum_{j=1}^{p} \frac{(P_{ij} - P_{i.} P_{.j})^2}{P_{i.}}$$

Elle s'obtient en écrivant la trace de $Z {}^tZ D_I$ ou de ${}^tZ D_I Z$. On remarque de plus que : $Z {}^tZ D_I \underset{\sim}{1}_n = 0$ et ${}^tZ D_I Z \underset{\sim}{1}_p = 0$. Il en découle les résultats suivants :

Propriété 4.4 $${}^t\underset{\sim}{1}_n D_I \phi_\alpha = {}^t\underset{\sim}{1}_p \varphi_\alpha = 0$$

Propriété 5.4 Pour $\lambda_\alpha \neq 0$ $$\varphi_\alpha = \frac{{}^tZ D_I \phi_\alpha}{\sqrt{\lambda_\alpha}} = \frac{{}^tP \phi_\alpha}{\sqrt{\lambda_\alpha}}$$

Propriété 6.4 Soit $$\tilde{P}_{(k)} = D_I \underset{\sim}{1}_n {}^t\underset{\sim}{1}_p D_J + D_I \left(\sum_{\alpha=1}^{k} \sqrt{\lambda_\alpha} \phi_\alpha \varphi_\alpha \right)$$

alors $$\tilde{P}_{(k)} \underset{\sim}{1}_p = D_I \underset{\sim}{1}_n$$

$${}^t\tilde{P}_{(k)} \underset{\sim}{1}_n = D_J \underset{\sim}{1}_p$$

et $$\sum_{i=1}^{n} \sum_{j=1}^{p} \frac{(P - \tilde{P}_{(k)})^2_{ij}}{P_{i.}} = \sum_{\alpha=k+1}^{v} \lambda_\alpha$$

F. MARCOTORCHINO (1984) relie le critère de Goodman Kruskal à des propositions de R.J. LIGHT et B.H. MARGOLIN (1971). On peut voir dans ces multiples intérêts, l'attrait de ce critère. L'Analyse non symétrique des Correspondances permettra une exploration des écarts des P_{ij} aux $P_{i.} P_{.j}$ tels qu'ils sont pris en compte dans ce critère.

II.2. Analyse des correspondances par rapport à un modèle

Les propriétés 3, 3.3 et 3.4 ont mis en évidence que l'A.C., la décomposition du critère de Belson et l'Analyse non symétrique des Correspondances ont pour but d'explorer les écarts des P_{ij} observés aux $P_{i.} P_{.j}$ que donnerait l'hypothèse d'indépendance. Les trois méthodes diffèrent par le choix du critère utilisé pour mesurer cet écart. B. ESCOFIER (1984) envisage des situations pratiques dans lesquelles l'objectif peut être de comparer le tableau P à un tableau Π donné.

Soient alors S_J et R_I deux matrices diagonales positives. S_J est pxp, d'éléments diagonaux $(s_j, j = 1,\ldots,p)$ et R_I, nxn, d'éléments diagonaux $(r_i, i=1,\ldots,n)$.

Définition 5 On appelle Analyse des Correspondances du tableau de contingence P par rapport au modèle Π, l'ACP du triplet $(R_I^{-1} (P - \Pi) S_J^{-1}, S_J, R_I)$

On vérifie aisément que les ϕ_α qui permettront de représenter les lignes de P-Π sont les vecteurs propres de $R_I^{-1}(P - \Pi) S_J^{-1}(P - \Pi)$ tandis que les φ_α sont les vecteurs propres de $S_J^{-1}\, {}^t(P - \Pi) R_I^{-1}(P - \Pi)$. On a alors :

Propriété 3.5 $$\sum_{\alpha=1}^{v} \lambda_\alpha = \frac{(P_{ij} - \Pi_{ij})^2}{r_i\, s_j}$$

Propriété 5.5 Si $\lambda_\alpha \neq 0$ $\quad \phi_\alpha = \dfrac{R_I^{-1} (P - \Pi) \varphi_\alpha}{\sqrt{\lambda_\alpha}}$ et $\varphi_\alpha = \dfrac{S_J^{-1} (P - \Pi) \phi_\alpha}{\sqrt{\lambda_\alpha}}$

On voit en particulier que $\sqrt{\lambda_\alpha}\, \phi_{\alpha i} = \sum_{j=1}^{p} \dfrac{(P_{ij} - \Pi_{ij})\, \varphi_{\alpha j}}{r_i}$. Le point représentant la ligne i aura donc tendance à se trouver du même côté que les points représentant les colonnes j pour lesquelles $P_{ij} > \Pi_{ij}$.

Précisons la nature de Π en supposant que P et Π ont les mêmes marges. Ce sera le cas par exemple si $\Pi_{ij} = \frac{P_{i.}}{p} + \frac{P_{.j}}{n} - \frac{1}{pxn}$.

On a alors $(P - \Pi)\, \underset{\sim}{1}_p = {}^t(P - \Pi)\, \underset{\sim}{1}_n = 0$ d'où :

<u>Propriété 4.5</u> Si P et Π ont les mêmes marges

$$ {}^t\underset{\sim}{1}_n \, R_I \, \phi_\alpha = {}^t\underset{\sim}{1}_p \, S_J \, \varphi_\alpha = 0 $$

<u>Propriété 6.5</u> Soit $\tilde{P}_{(k)} = \Pi + R_I \left(\sum_{\alpha=1}^{k} \sqrt{\lambda_\alpha}\, \phi_\alpha \, \varphi_\alpha \right) S_J$

$$ \text{Alors} \quad \sum_{i=1}^{n} \sum_{j=1}^{p} \frac{(P - \tilde{P}_{(k)})^2_{ij}}{r_i \, s_j} = \sum_{\alpha=k+1}^{v} \lambda_\alpha $$

et si Π et P ont les mêmes marges

$$ \tilde{P}_{(k)}\, \underset{\sim}{1}_p = P\, \underset{\sim}{1}_p $$

$$ {}^t\tilde{P}_{(k)}\, \underset{\sim}{1}_n = {}^tP\, \underset{\sim}{1}_n $$

II.3. Altérations basées sur le choix de la distance

La propriété 2 a rappelé que l'AC compare deux lignes de P au moyen de la distance du χ^2 entre les lois conditionnelles. Ce choix peut être mis en question lorsqu'on veut éviter la sensibilité de cette distance à certaines configurations particulières du tableau P.

II.3.1. Pour éviter les problèmes pratiques que peuvent poser des $P_{ij} = 0$, B. ESCOFIER (1978) introduit les quantités

$$ s_j = \sum_{\substack{i=1 \\ P_{ij} \neq 0}}^{n} P_{i.} \qquad \text{et} \qquad r_i = \sum_{\substack{j=1 \\ P_{ij} \neq 0}}^{p} P_{.j} $$

Soient alors S_J la matrice diagonale pxp d'éléments s_j et R_I la matrice diagonale nxn d'éléments r_i. Parmi les deux possibilités envisagées par B. ESCOFIER, nous retiendrons la plus riche en propriétés qui peut conduire à l'énoncé suivant :

Définition 6 On appellera Analyse Pondérée des Correspondances, l'ACP du triplet $(X, D_J S_J, D_I R_I)$ où $X = D_I^{-1}(P - D_I \underset{\sim}{1}_n \, {}^t\underset{\sim}{1}_p \, D_J) \, D_J^{-1}$.

On vérifie aisément par le calcul la propriété suivante :

Propriété 2.6

$$d^2(i,i') = \sum_{j=1}^{p} \frac{s_j}{P_{.j}} (P(j/i) - P(j/i'))^2$$

$$d^2(j,j') = \sum_{i=1}^{n} \frac{r_i}{P_{i.}} (P(i/j) - P(i/j'))^2$$

Propriété 3.6

$$\sum_{\alpha=1}^{v} \lambda_\alpha = \sum_{i=1}^{n} \sum_{j=1}^{p} r_i \, s_j \, \frac{(P_{ij} - P_{i.} P_{.j})^2}{P_{i.} \, P_{.j}}$$

Pour démontrer cette propriété, il suffit de calculer les traces des opérateurs $X \, D_J \, S_J \, {}^tX \, D_I \, R_I$ ou ${}^tX \, D_I \, R_I \, X \, D_J \, S_J$. On voit de plus $R_I^{-1} \underset{\sim}{1}_n$ est vecteur propre du premier pour la valeur propre 0, tandis que $S_J^{-1} \underset{\sim}{1}_p$ l'est du second pour la même valeur propre. L'orthogonalité des ϕ_α pour la métrique $D_I R_I$ et celles des φ_α pour $D_J S_J$ donnent alors :

Propriété 4.6 $\underset{\sim}{1}_n \, D_I \, \phi_\alpha = \underset{\sim}{1}_p \, D_J \, \varphi_\alpha = 0$

Propriété 5.6

Pour $\lambda_\alpha \neq 0$ $\quad \phi_\alpha = \dfrac{D_I^{-1} \, (P - D_I \underset{\sim}{1}_n \, {}^t\underset{\sim}{1}_p \, D_J) \, S_J \, \varphi_\alpha}{\sqrt{\lambda_\alpha}}$

et $\quad \varphi_\alpha = \dfrac{D_J^{-1} \, {}^t(P - D_I \underset{\sim}{1}_n \, {}^t\underset{\sim}{1}_p \, D_J) \, R_I \, \phi_\alpha}{\sqrt{\lambda_\alpha}}$

Pour des R_I et S_J quelconques on n'obtient pas une propriété de centre de gravité aussi simple que pour l'A.C.

On peut remarquer toutefois que les propriétés obtenues ne dépendent en rien des choix faits par B. ESCOFIER pour R_I et S_J. Elles sont valables pour tout système de pondérations. Le choix $S_J = I_{pxp}$, $R_I \neq I_{nxn}$ donnera alors partiellement la propriété du centre de gravité.

Propriété 6.6

$$\text{Soit } \tilde{P}_{(k)} = D_I \left(\underset{\sim}{1}_n \; {}^t\underset{\sim}{1}_p + \sum_{\alpha=1}^{k} \sqrt{\lambda_\alpha}\, \phi_\alpha\, \varphi_\alpha \right) D_J$$

$$\text{Alors } \tilde{P}_{(k)} \underset{\sim}{1}_p = P \underset{\sim}{1}_p = D_J \underset{\sim}{1}_p$$

$${}^t\tilde{P}_{(k)} \underset{\sim}{1}_n = {}^t P \underset{\sim}{1}_n = D_I \underset{\sim}{1}_n$$

$$\text{et } \sum_{i=1}^{n} \sum_{j=1}^{p} r_i\, s_j \frac{(P - P_{(k)})^2_{ij}}{P_{i.}\, P_{.j}} = \sum_{\alpha=k+1}^{v} \lambda_\alpha$$

Propriété 7.6

Si deux colonnes j_1 et j_2 sont proportionnelles et si $s_{j_1} = s_{j_2}$, on ne change pas les résultats concernant les lignes de P en remplaçant ces deux colonnes par leur somme.

Remarquons d'abord que dans le choix de S_J fait par B. ESCOFIER, la proportionnalité des colonnes j_1 et j_2 entraîne l'égalité de s_{j_1} et s_{j_2}.

Pour la démonstration, on écrira l'élément (i, i') de l'opérateur $X\, D_J\, S_J\, {}^tX\, D_I\, R_I$ à diagonaliser pour trouver les ϕ_α. On trouve :

$$\frac{1}{P_{i.}} \left(\sum_{j=1}^{p} s_j\, P_{.j}\, \frac{(P_{ij} - P_{i.}\, P_{.j})}{P_{.j}}\, \frac{(P_{i'j} - P_{i'.}\, P_{.j})}{P_{.j}} \right) r_i$$

et les arguments utilisés pour démontrer la propriété 7 sont applicables à la quantité entre parenthèses.

II.3.2. D. DOMENGES et M. VOLLE (1979) proposent de comparer deux tableaux de contingence P et Π par la distance de Hellinger

$$d^2(P, \Pi) = \sum_{i=1}^{n} \sum_{j=1}^{p} (\sqrt{P_{ij}} - \sqrt{\Pi_{ij}})^2.$$

B. ESCOFIER (1978) reprend cette idée en la modifiant légèrement. On peut faire une synthèse de ces deux approches en considérant M_I une matrice diagonale positive, nxn, d'éléments m_i et M_J une matrice diagonale positive, pxp, d'éléments m_j. On peut alors proposer l'énoncé suivant :

Définition 7 — On appelle Analyse des Correspondances sphériques du tableau de contingence P par rapport au tableau de contingence Π, l'ACP du triplet (X, M_I, M_J) où

$$X_{ij} = \sqrt{\frac{P_{ij}}{M_i M_j}} - \sqrt{\frac{\Pi_{ij}}{M_i M_j}}$$

Propriété 3.7

$$\sum_{\alpha=1}^{v} \lambda_\alpha = \sum_{i=1}^{n} \sum_{j=1}^{n} (\sqrt{P_{ij}} - \sqrt{\Pi_{ij}})^2$$

Propriété 5.7 — Pour tout $\lambda_\alpha \neq 0$

$$\psi_{\alpha i} = \sum_{j=1}^{p} (\sqrt{P_{ij}} - \sqrt{\Pi_{ij}}) \sqrt{\frac{M_j}{M_i}}\ \varphi_{\alpha j}$$

et

$$\varphi_{\alpha j} = \sum_{i=1}^{n} (\sqrt{P_{ij}} - \sqrt{\Pi_{ij}}) \sqrt{\frac{M_i}{M_j}}\ \varphi_{\alpha j}$$

Propriété 6.7 — Soit $\widetilde{P}_{(k)}$ tel que :

$$\sqrt{(\widetilde{P}_{(k)})_{ij}} = \sqrt{\Pi_{ij}} + \sqrt{M_i M_j} \sum_{\alpha=1}^{k} \lambda_\alpha \psi_{\alpha i} \varphi_{\alpha j}$$

$$\text{alors} \sum_{i=1}^{n} \sum_{j=1}^{n} (\sqrt{P_{ij}} - \sqrt{\widetilde{P}_{(k)ij}})^2 = \sum_{\alpha=k+1}^{v} \lambda_\alpha$$

Le résultat découle de la propriété 1 en remarquant que $(\sqrt{P_{ij}} - \sqrt{(\widetilde{P}_{(k)})_{ij}})^2 = M_i (X - (\widetilde{X}_{(k)})^2_{ij} M_j$

B. ESCOFIER se place dans le cas particulier où pour tout i et tout j $\Pi_{ij} = 0$, $M_i = P_{i.}$ et $M_j = P_{.j}$. On a alors :

Propriété 2.7

$$d^2(i,i') = \sum_{j=1}^{p} (\sqrt{P(j/i)} - \sqrt{P(j/i')})^2$$

$$\text{et} \quad d^2(j,j') = \sum_{i=1}^{p} (\sqrt{P(i/j)} - \sqrt{P(i/j')})^2$$

Propriété 7.7 Dans le cas particulier étudié par B. ESCOFIER, on ne change pas la distance entre deux lignes en remplaçant deux colonnes proportionnelles par leur somme.

Dans ce cas, l'opérateur à diagonaliser pour calculer ϕ_α a pour élément (i,i') $\sum_{j=1}^{p} \sqrt{P(j/i)\ P(j/i')}$ qu'on peut écrire

$$\sum_{j=1}^{p} P_{.j} \sqrt{\frac{P(i,j)}{P_{.j}} \frac{P(i',j)}{P_{.j}}} \sqrt{\frac{1}{P_{i.}\ P_{i'.}}}$$

Le résultat s'obtient comme il a été obtenu pour la propriété 7.

II.4. Analyse des Correspondances Multiples

L'Analyse des Correspondances Multiples (A.C.M.) peut être considérée comme une altération de l'A.C. basée sur le changement des données à traiter. Au lieu d'étudier conjointement 2 variables qualitatives, on s'engage dans l'analyse q > 2 variables. Cette extension de l'A.C. est ancienne, ce qui nous permettra de ne pas la reprendre ici. On en trouvera des présentations dans les ouvrages cités en bibliographie ou dans les articles de ESCOUFIER (1984), J.C. DEVILLE et al. (1983). Plusieurs auteurs sont allés encore plus loin dans l'analyse de plus de 2 variables qualitatives : on pourra lire par exemple P. CAZES et al. (1976) ; P. CAZES et al. (1977).

III - LA PORTEE DES RESULTATS D'UNE ANALYSE DES CORRESPONDANCES

Pour terminer cet exposé sur l'A.C. et ses extensions, il paraît important d'évoquer une évolution récente du comportement des praticiens de l'A.C. vis-à-vis des données qu'ils traitent. Cette évolution concerne d'ailleurs très généralement toute la pratique de l'Analyse des Données. Dans les années précédentes l'A.C. était présentée comme une méthode d'exploration de données et se défendait de toutes préoccupations inférentielles, allant même jusqu'à en nier le bien fondé (BENZECRI, 1973). Aujourd'hui, sous les termes d'études de validité et de stabilité des résultats obtenus (B. ESCOFIER (1979), J. BENASSENI (1984)) apparaissent des préoccupations parfois mal situées par rapport aux deux motivations fondamentales justifiant l'exploration d'un phénomène réel observé :

1) l'obtention d'une description pertinente des observations faites susceptibles d'aider à la compréhension du phénomène,

2) la volonté de généraliser au non-observé les propriétés reconnues sur l'observé.

III.1. Envisageons dans une premier temps le cas d'un ensemble de données qui soit exhaustif du phénomène qu'on veut étudier.

Le but de l'analyse de telles données est de substituer aux données brutes trop nombreuses pour être facilement appréhendées dans leur totalité, quelques descripteurs synthétiques aptes à exprimer les propriétés que ces données possèdent. Dans l'A.C. ce sont les représentations graphiques fournies par les ϕ_α, les φ_α et les λ_α qui sont les descripteurs. Il est nécessaire d'attacher à chaque descripteur une appréciation d'importance qui rejaillira sur la propriété qu'il traduit. Pour le faire, il semble alors assez naturel de lier l'importance reconnue à un descripteur à la stabilité dont ce descripteur fait preuve par rapport à des perturbations des données compatibles avec le problème envisagé. Dans certains cas, il sera possible de donner une spécification mathématique des perturbations envisagées dont on pourra éventuellement déduire une spécification mathématique des fluctuations du descripteur. Cette tentative peut être refusée soit parce que les perturbations qu'on envisage conduiraient à une formulation mathématique inextricable soit parce que dans le cadre d'une formulation mathématique des perturbations, le descripteur est trop complexe pour qu'une spécification mathématique efficace de ses fluctuations soit possible. S'ouvre alors la voie de techniques basées sur des altérations artificielles des données compatibles avec le phénomène étudié. "Par exemple dans le cas classique des réponses ordonnées du type : pas du tout d'accord ; pas très d'accord ; assez d'accord ; tout-à-fait d'accord ; on peut supposer que l'individu enquêté a une chance sur deux d'avoir exprimé ce qu'il ressentait, une chance sur quatre (sauf aux extrémités) de répondre à une modalité immédiatement contigüe" (in L. LEBART (1977) p. 240). Seuls les descripteurs qui resteront stables par rapport à ces altérations mériteront d'être pris en compte.

III.2. Le second point de vue qui peut présider à une analyse des données consiste à considérer la collection des données disponibles comme représentative d'autres collections qui pourraient lui être substituées ou l'englober. Dans l'A.C., on pensera bien sûr à la représentativité des individus sur lesquels les deux variables qualitatives ont été observées. Comme dans toutes situations analogues, cette représentativité est fondée sur l'honnêteté de la procédure qui a présidé aux choix des individus observés. Mais dans certaines situations on devra aussi penser aux procédures qui ont conduit à énumérer les modalités des variables ; d'autres énumérations peuvent être possibles. Enfin dans l'A.C.M. c'est le choix des variables lui-même qui peut être mis en cause.

Sur ces bases fragiles, un descripteur mérite qu'on lui attache de l'importance s'il est reproductible d'un ensemble de données observé à un autre qui aurait pu lui être substitué. La permanence par rapport aux substitutions des données observées apparaît alors comme le gage de la reproductibilité. Comment la vérifier ?

La démarche statistique classique consiste à formuler une hypothèse sur la population hypothétique dont les données disponibles seraient issues (par exemple l'hypothèse d'indépendance pour deux variables qualitatives dont les modalités sont clairement spécifiées). Cette hypothèse prend la forme d'une spécification mathématique des fluctuations des unités statistiques concernées, spécification qui induit la forme des fluctuations d'un critère qui va permettre d'apprécier l'adéquation des données à l'hypothèse (la loi du χ^2 pour notre exemple). Les données disponibles permettront d'estimer des paramètres intervenant dans la spécification mathématique (les $P_{i.}$ et $P_{.j}$ pour notre exemple), si bien que ce sont elles qui fournissent une estimation de l'intensité des fluctuations possibles.

Forme supposée et intensité estimée permettront de mettre à l'épreuve l'hypothèse faite. Le rejet de l'hypothèse montre la spécificité des observations ; aux données disponibles on ne peut pas substituer n'importe quoi en particulier l'hypothèse faite. La conservation de l'hypothèse permet d'énumérer des résultats qui auraient pu se substituer à ceux qui ont été observés.

En imposant une hypothèse à mettre à l'épreuve cette démarche force la forme des fluctuations à envisager. On peut alors s'interroger sur la capacité des données disponibles à fournir elles-mêmes des informations sur la forme des fluctuations à attendre de leur remplacement par des données substituables. Quand les ensembles de données sont grands, ce qui est souvent le cas en Analyse des Données, on peut accepter l'idée que la fluctuation interne des données disponibles est la meilleure approximation de la fluctuation externe qui résulterait de la prise en compte d'un autre ensemble de données. Il s'agit de substituer à une fluctuation hypothétique, une fluctuation observée : reconnaissons que ce choix est d'ordre philosophique.

On est alors conduit à mettre en oeuvre des procédures de ré-échantillonnage inspirées des techniques de Bootstrap et de Jackknife.

Il faudra apprendre à construire les procédures efficaces de leurs mises en oeuvre et à choisir les modes de présentation des résultats. Il est clair toutefois que les travaux qui débutent dans ce sens mériteront d'être suivis avec attention.

BIBLIOGRAPHIE

BELSON W. (1959). Matching and Prediction in the Principle of Biological Classification. Applied Statistics, Vol. II.

BENASSENI J. (1984). Une contribution à l'étude de la stabilité en analyse factorielle. Thèse de 3e cycle. U.S.T.L. Montpellier.

BENZECRI J.P. (1973). L'analyse des Données. 1 - La Taxonomie. 2 - Analyse des Correspondances. Dunod.

CAZES P., BAUMERDER A., BONNEFOUS S., PAGES J.P. (1976). Description cohérente des variables qualitatives prises globalement et de leurs modalités. S.A.D., 2-3, p. 48-62.

CAZES P., BAUMERDER A., BONNEFOUS S., PAGES J.P. (1977). Codage et Analyse des tableaux logiques : introduction à la pratique des variables qualitatives. BURO, Cahier n° 27, Université Pierre et Marie Curie.Paris.

DEVILLE J.C., SAPORTA G. (1983). Correspondence analysis : with an extension towards nominal time series. Journal of Econometrics n° 22, p. 169-189.

DOMENGES D., VOLLE M. (1979). Analyse factorielle sphérique : une exploration. Annales de 1'INSEE n° 35.

ESCOFIER B. (1978). Analyse factorielle et distances répondant au principe d'équivalence distributionnelle. RSA, Vol. XXVI n° 4, p. 29-37.

ESCOFIER B. (1979). Stabilité et approximation en analyse factorielle. Thèse de Doctorat d'Etat. Université Paris VI.

ESCOFIER B. (1984). Analyse factorielle en référence à un modèle : application à l'analyse de tableaux d'échanges. R.T. 337. INRIA, Domaine de Voluceau - Rocquencourt. B.P. 105. 78153 Le Chesnay Cedex. France.

ESCOUFIER Y. (1982). L'analyse des tableaux de contingence simples et multiples. Metron Vol. XL, n° 1-2, p. 53-77.

GOODMAN L., KRUSKAL W.H. (1979). Measures of association for cross classification. Springer Verlag, Berlin, New-York.

GREENACRE M.J. (1984). Theory and applications of correspondence analysis. Academic Press.

HILL M.O. (1974). Correspondence Analysis : A neglected multivariate method. Applied statistics n° 23, p. 340-354.

LAURO N., D'AMBRA L. (1983). L'analyse non symétrique des correspondances. In Data Analysis and Informatics III, ed. by E. DIDAY et al. Elsevia Science Publishers BV (North-Holland), p. 433-446.

LEBART L., MORINEAU A., TABARD N. (1977). Techniques de la description statistique. Dunod.

LEBART L., MORINEAU A., WARWICK (1984). Multivariate descriptive statistical analysis. John Wiley.

LIGHT R.J., MARGOLIN B.H. (1971). An analysis of variance for categorical data. JASA, Vol. 66, p. 534-544.

MARCOTORCHINO F. (1984). Utilisation des comparaisons par paires en statistique des contingences (Partie I). Etude F069. Centre Scientifique IBM-France, 36, av. Raymond Poincaré, 75016 Paris France.

NISHISATO S. (1980). Analysis of categorical data : dual scaling and its applications. University of Toronto Press-Toronto.

OKAMOTO M. (1969). Optimality of principal components. In Multivariate Analysis II. Ed. by P.K. Krishnaiah, Academic Press,p. 673-685.

SABATIER R., JAN Y., ESCOUFIER Y. (1984). Approximations d'applications linéaires et analyse en composantes principales. In Data Analysis and Informatics III, ed. by E. DIDAY et al. Elsevia Science Publishers B.V. (North-Holland), p. 569-580.

RESUME

Le premier paragraphe rappelle les choix qui conduisent à développer l'Analyse des Correspondances (A.C.) et les propriétés des résultats que cette méthode fournit.

Le second paragraphe construit des méthodes voisines de l'A.C. en modifiant certains des choix initiaux. Il explore les propriétés qui résultent de ces altérations.

Le troisième paragraphe est une réflexion sur la signification des travaux portant sur la stabilité et la validité des résultats en A.C.

SUMMARY

A first paragraph recalls the reasons that have lead to development of correspondence analysis and this method's properties.

The second paragraph sets up methods similar to correspondence analysis by altering some of the initial options. It investigates the properties resulting from these changes.

The third paragraph contains considerations concerning significance of work done on stability and validity of correspondence analysis'results.