CAHIERS DU BURO

P. CAZES
A. BAUMERDER
S. BONNEFOUS
J. P. PAGÈS

Codage et analyse des tableaux logiques Introduction à la pratique des variables qualitatives

Cahiers du Bureau universitaire de recherche opérationnelle. Série Recherche, tome 27 (1977), p. 3-47

<http://www.numdam.org/item?id=BURO_1977__27__3_0>

CODAGE ET ANALYSE DES TABLEAUX LOGIQUES
INTRODUCTION A LA PRATIQUE DES VARIABLES QUALITATIVES

P. CAZES [1], A. BAUMERDER [2], S. BONNEFOUS [2] et J.P. PAGÈS [2]

On se propose de faire le point sur la pratique des variables qualitatives en analyse des données en se limitant aux méthodes ne faisant appel qu'à l'algèbre linéaire.

Les idées développées ici sont présentées soit de façon géométrique, on raisonne alors dans l'espace des variables muni de la métrique des poids (distance en moyenne quadratique), soit de façon algébrique, on recherche les codages simultanés de p variables qualitatives rendant maximum un certain indice.

Après avoir défini au § 1 les notations et fait des rappels sur le codage d'une variable qualitative, on donne au § 2 une généralisation possible de l'analyse canonique, généralisation qui revient à une analyse en composantes principales.

Au § 3, on traite de l'analyse en composantes principales sur variables qualitatives ; on rappelle que l'analyse factorielle des correspondances (A F C) de deux variables qualitatives x et y est équivalente à d'autres analyses comme par exemple l'A F C du tableau Z des indicatrices de x et de y ou comme l'A F C du tableau de Burt $V = ZZ'/n$ associé. Ces analyses ont l'avantage de se généraliser immédiatement quand on a plus de deux variables qualitatives. Cette généralisation est équivalente à la généralisation de l'analyse canonique donnée au § 2 ; elle correspond à la pratique usuelle qui est de faire l'A F C du tableau Z de l'ensemble des indicatrices de toutes les variables qualitatives.

---

(1) Laboratoire de Statistique. Université Pierre et Marie-Curie (Paris VI).

(2) Laboratoire de Statistiques et d'Etudes Economiques et Sociales – Département de Protection (C.E.A.).

Au § 4, on traite de "l'analyse canonique" entre deux paquets de variables qualitatives ; on compare en particulier l'analyse canonique des deux sous-espaces respectivement associés aux indicatrices des deux paquets, avec l'A F C du tableau obtenu en croisant l'ensemble des modalités du premier paquet avec l'ensemble des modalités du second.

Enfin, au § 5, on montre que faire un test en analyse de variance multidimensionnelle est équivalent soit à rechercher le premier facteur et la valeur propre associée d'une certaine analyse en composantes principales, soit à rechercher le premier couple de directions canoniques et la valeur propre associée d'une certaine analyse canonique.

1 – NOTATIONS – CODAGE – DUALITE.

Les variables considérées permettent de décrire un ensemble d'"individus" $K = \{k \mid k = 1, \ldots, n\}$ dont les éléments sont munis de poids p_k avec :

$$p_k > 0 \quad ; \quad \sum_{k \in K} p_k = 1.$$

Si on note $I = \{i \mid i = 1, \ldots, p\}$ l'ensemble des modalités de la variable qualitative x, on sait que toutes les variables quantitatives ξ que l'on sait reconstruire à partir de x sont de la forme :

$$\xi = a \circ x = \sum_{i=1}^{p} a^i \, x^i$$

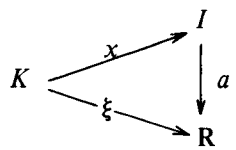
où :

- a est l'application codage : $I \xrightarrow{a} \mathbf{R}$
- x^i est la variable indicatrice associée à la $i^{\text{ème}}$ modalité de x :

$$x^i(k) = 1 \quad \text{si } x(k) = i$$

$$x^i(k) = 0 \quad \text{sinon.}$$

- $a^i = a(i)$ est le codage numérique de la $i^{\text{ème}}$ modalité de x



Si l'application a est injective les variables x et ξ qui induisent sur K la même partition sont équivalentes ; les codages a^i sont alors tous différents.

Les variables quantitatives définies sur K sont représentées comme des points dans le vectoriel des caractères $F = \mathbf{R}^n$; aux variables indicatrices x^i sont associés en particulier les points $\underline{x}^i$. A l'ensemble des variables indicatrices est alors associé le tableau logique :

$$X' = (\underline{x}^1, \underline{x}^2, \ldots, \underline{x}^p).$$

L'espace F étant muni de la distance en moyenne quadratique D_p (métrique des poids), à laquelle est associée la matrice diagonale D_p des poids p_k, au couple (X, D_p) correspond le schéma de dualité :

$$\begin{array}{ccc} E = \mathbf{R}^p & \xleftarrow{\quad X \quad} & F^* \\ \Big\downarrow D_{1/p_I} \quad \Big\uparrow V = D_{p_I} & & \Big\uparrow D_p \\ E^* & \xrightarrow[\quad X' \quad]{} & F = \mathbf{R}^n \end{array}$$

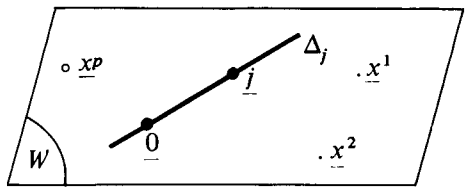
Le vectoriel E de dimension p est assimilé ici à l'ensemble des mesures définies sur $(I, \mathcal{P}(I))$; le simplexe des lois de probabilité définies sur $(I, \mathcal{P}(I))$ est situé dans un sous-espace affine parallèle au sous-espace vectoriel des mesures dont la masse totale est nulle.

- p_I désigne la loi de probabilité sur $(I, \mathcal{P}(I))$ définie par les probabilités p_i de prendre les différentes modalités i de I ;
- le dual E^* de E est considéré comme l'espace des variables définies sur I ;
- la forme quadratique induite sur E^* par D_p, représentée dans le schéma par l'application $V = X \circ D_p \circ X'$, admet pour matrice D_{p_I}, matrice diagonale dont les éléments diagonaux sont les poids p_i ;
- restreinte au simplexe des lois de probabilités, la métrique D_{1/p_I} induite sur E par D_{p_I}, n'est autre que la distance du chi-deux de centre p_I.

L'ensemble des variables quantitatives ξ que l'on sait reconstruire à partir de x est alors représenté dans F par le sous-espace vectoriel :

$$W = \{\underline{\xi} \mid \underline{\xi} = X'(\underline{a}) \; ; \underline{a} \in E^*\}.$$

Remarque :



W contient la droite des constantes Δ_j engendrée par le vecteur $\underline{j}$ dont toutes les coordonnées sont égales à 1.

$$W = \Delta_j \oplus W^-$$

où W^- est le supplémentaire D_p – orthogonal de Δ_j dans W.

2 – ANALYSE CANONIQUE ET ANALYSE EN COMPOSANTES PRINCIPALES GENERALISEES.

On sait que l'analyse canonique est une technique permettant de décrire les positions relatives occupées dans un espace euclidien F par deux sous-espaces W_1 et W_2 ; si $W_1^\perp$ et $W_2^\perp$ désignent les supplémentaires orthogonaux dans F de W_1 et W_2 on aboutit en analyse canonique aux deux décompositions en sommes directes simultanées :

$$\begin{cases} W_1 = W_1 \cap W_2 \oplus W_{11} \oplus W_{111} \\ W_2 = W_1 \cap W_2 \oplus W_{22} \oplus W_{222} \end{cases} \qquad (1)$$

où $\qquad W_{111} = W_1 \cap W_2^\perp \qquad W_{222} = W_1^\perp \cap W_2$

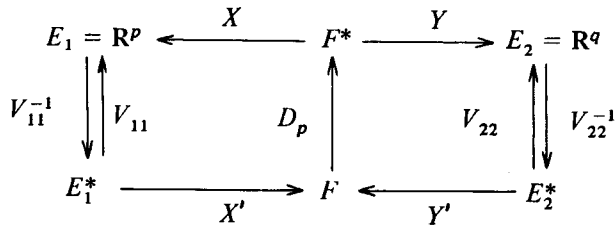
où W_{11} et W_{22} sont les supplémentaires orthogonaux respectivement dans W_1 du sous-espace $W_1 \cap W_2 \oplus W_{111}$ et dans W_2 du sous-espace $W_1 \cap W_2 \oplus W_{222}$.

En général on se place dans le vectoriel des caractères $F = \mathbf{R}^n$ muni de la métrique des poids D_p, les sous-espaces W_1 et W_2 étant engendrés respectivement par les deux paquets de caractères $\{\underline{x}^i \mid i = 1, \ldots, p\}$ et $\{\underline{y}^j \mid j = 1, \ldots, q\}$.

A ces deux paquets de caractères correspondent les tableaux de données :

$$X' = (\underline{x}^1, \underline{x}^2, \ldots, \underline{x}^p) \quad ; \quad Y' = (\underline{y}^1, \underline{y}^2, \ldots, \underline{y}^q)$$

et on est amené à considérer le schéma de dualité :



- $W_1 = X'(E_1^*)$ $\qquad W_2 = Y'(E_2^*)$
- $V_{11} = X \circ D_p \circ X'$ $\qquad V_{22} = Y \circ D_p \circ Y'$

$$V_{12} = V'_{21} = X \circ D_p \circ Y'.$$

Si tous les caractères sont centrés, les matrices associées aux applications V_{11}, V_{22} et V_{12} ne sont autres que des matrices de covariance.

Si on note A_1 et A_2 les opérateurs de D_p – projection sur W_1 et W_2, les vecteurs propres normés de $A_1 \circ A_2$ (les valeurs propres sont rangées par valeurs décroissantes) fournissent une base orthonormée de W_1 ;

$$A_1 \circ A_2 \, \underline{\xi}^i = \lambda_i \, \underline{\xi}^i \; ; \; \| \underline{\xi}^i \| = 1 \; ; i = 1, \ldots, p.$$

$\underline{\xi}^1$ est le vecteur normé de W_1 rendant maximum la quantité $\| A_2 \, \underline{\xi} \|^2$; $\underline{\xi}^1$ est donc le vecteur normé de W_1 le plus proche de W_2.

A cette base de W_1 correspond la base orthonormée de W_2 composée des vecteurs propres de $A_2 \circ A_1$:

$$A_2 \circ A_1 \, \underline{\eta}^i = \lambda_i \, \underline{\eta}^i \; ; \; \| \underline{\eta}^i \| = 1 \; ; i = 1, \ldots, q.$$

Si p est inférieur à q, la valeur propre λ_i ($i = 1, \ldots, p$) n'est autre que le carré du cosinus (cos) de l'angle entre les caractères canoniques $\underline{\xi}^i$ et $\underline{\eta}^i$; le couple $(\underline{\xi}^1, \underline{\eta}^1)$ est le couple de vecteurs normés de $W_1 \times W_2$ rendant maximum la quantité $D_p(\underline{\xi}, \underline{\eta}) = \cos(\underline{\xi}, \underline{\eta})$ ou ce qui revient au même, rendant minimum la quantité $\| \underline{\xi} - \underline{\eta} \|^2$.

Les vecteur $\underline{\xi}^i$ et $\underline{\eta}^i$, s'ils ne sont pas à angle droit, vérifient les relations :

$$\underline{\eta}^i = \frac{A_2 \, \underline{\xi}^i}{\| A_2 \, \underline{\xi}^i \|} = \frac{1}{\sqrt{\lambda_i}} A_2 \, \underline{\xi}^i \; ; \; \underline{\xi}^i = \frac{A_1 \, \underline{\eta}^i}{\| A_1 \, \underline{\eta}^i \|} = \frac{1}{\sqrt{\lambda_i}} A_1 \, \underline{\eta}^i.$$

Si les applications X' et Y' sont injectives, les opérateurs de D_p – projection s'écrivent :

$$A_1 = X' \circ V_{11}^{-1} \circ X \circ D_p \; , \; A_2 = Y' \circ V_{22}^{-1} \circ Y \circ D_p.$$

Aux caractères canoniques $\underline{\xi}^i$ et $\underline{\eta}^i$ correspondent respectivement les facteurs canoniques $\underline{a}_i$ et $\underline{b}_i$.

$$\underline{\xi}^i = X'(\underline{a}_i) \quad ; \quad \underline{a}_i \in E_1^*$$
$$\underline{\eta}^i = Y'(\underline{b}_i) \quad ; \quad \underline{b}_i \in E_2^*.$$

Les facteurs canoniques, si les applications X' et Y' sont injectives, sont obtenus directement par les équations :

$$B_1 \circ B_2 \, \underline{a}_i = \lambda_i \, \underline{a}_i$$
$$\| \underline{a}_i \|_{V_{11}} = 1$$
$$B_2 \circ B_1 \, \underline{b}_i = \lambda_i \, \underline{b}_i$$
$$\| \underline{b}_i \|_{V_{22}} = 1$$

avec : $B_1 = V_{11}^{-1} \circ V_{12} \qquad B_2 = V_{22}^{-1} \circ V_{21}.$

Si la valeur propre λ_i est non nulle :

$$\underline{a}_i = \frac{1}{\sqrt{\lambda_i}} B_1 \, \underline{b}_i \qquad \underline{b}_i = \frac{1}{\sqrt{\lambda_i}} B_2 \, \underline{a}_i.$$

2.1 – A propos d'une autre façon de décrire les positions relatives de deux sous-espaces vectoriels.

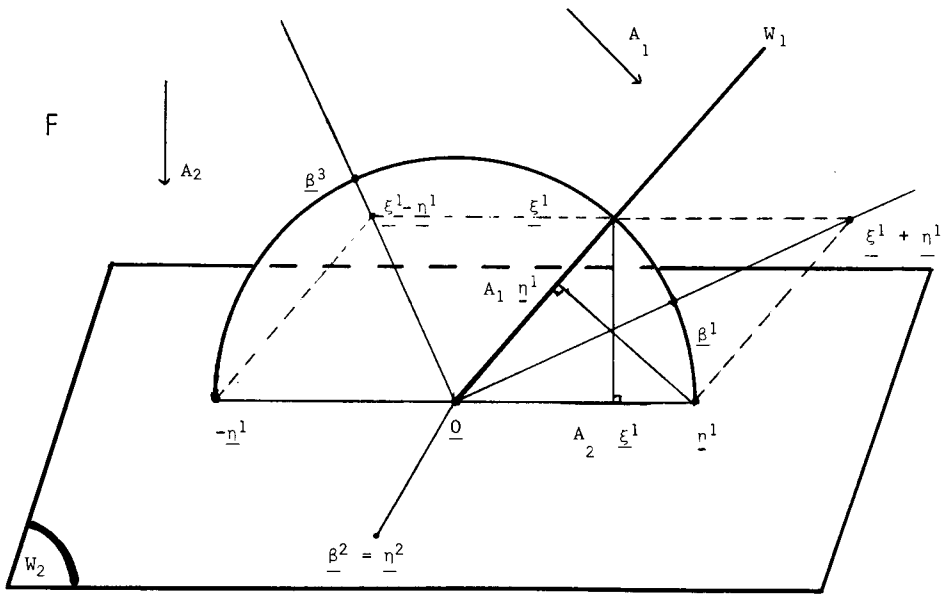


Figure 1

Si l'on considère l'exemple de la figure 1, les positions relatives de W_1 et W_2 peuvent être décrites à l'aide d'un système orthonormé de vecteurs situés soit à égale distance de W_1 et W_2, soit dans l'un de ces sous-espaces :

- le vecteur $\underline{\beta}^1$ est le vecteur normé de F, équidistant de W_1 et W_2, à distance minimum de W_1 (ou de W_2) ;
- le vecteur $\underline{\beta}^2$ est le vecteur normé de F situé dans W_2 et orthogonal à $\underline{\beta}^1$;
- le vecteur $\underline{\beta}^3$ est le vecteur normé de F, orthogonal au plan engendré par le couple $(\underline{\beta}^1, \underline{\beta}^2)$ et équidistant de W_1 et W_2, à distance minimum de W_1 ;

Ces vecteurs ne sont autres que les vecteurs propres normés de $A_1 + A_2$ associés aux valeurs propres μ_i non nulles rangées par valeurs décroissantes :

$$(A_1 + A_2)\,\underline{\beta}^i = \mu_i\,\underline{\beta}^i \qquad i = 1, 2, 3$$

$$\| \underline{\beta}^i \| = 1$$

- $\underline{\beta}^1$ est homothétique à $(\underline{\xi}^1 + \underline{\eta}^1)$:

$$(A_1 + A_2)\,(\underline{\xi}^1 + \underline{\eta}^1) = (1 + \sqrt{\lambda_1})\,(\underline{\xi}^1 + \underline{\eta}^1)\,;$$

- $\underline{\beta}^2$ est identique à $\underline{\eta}^2$:

$$(A_1 + A_2)\,\underline{\beta}^2 = \underline{\beta}^2\,;$$

- β^3 est homothétique à $(\underline{\xi}^1 - \underline{\eta}^1)$

$$(A_1 + A_2)\,(\underline{\xi}^1 - \underline{\eta}^1) = (1 - \sqrt{\lambda_1})\,(\underline{\xi}^1 - \underline{\eta}^1).$$

On a :

$$\mu_1 = 1 + \sqrt{\lambda_1}\ ,\ \mu_2 = 1\ ,\ \mu_3 = 1 - \sqrt{\lambda_1},$$

$\frac{\mu_1}{2}$ et $\frac{\mu_3}{2}$ sont les carrés des cosinus des angles entre W_1 et $\underline{\beta}^1$ et $\underline{\beta}^3$ respectivement.

$$\underline{\xi}^1 = \sqrt{\frac{2}{\mu_1}}\,A_1\,\underline{\beta}^1 = \sqrt{\frac{2}{\mu_3}}\,A_1\,\underline{\beta}^3$$

$$\underline{\eta}^1 = \sqrt{\frac{2}{\mu_1}}\,A_2\,\underline{\beta}^1 = -\sqrt{\frac{2}{\mu_3}}\,A_2\,\underline{\beta}^3.$$

Dans le cas général on obtient, en considérant les vecteurs propres $\underline{\beta}$ de $A_1 + A_2$ associés aux valeurs propres non nulles, un système orthonormé de vecteurs de F composé de vecteurs qui sont situés :

- soit dans $W_1 \cap W_2$; la valeur propre μ associée est alors égale à 2.
- soit à égale distance de W_1 et W_2, l'angle formé avec ces sous-espaces étant inférieur à 45 degrés ; homothétiques aux vecteurs $(\underline{\xi}^i + \underline{\eta}^i)$, il leur est associé une valeur propre μ comprise entre 1 et 2.
- soit dans $W_1 \cap W_2^{\perp}$ ou $W_1^{\perp} \cap W_2$, la valeur propre μ associée est alors égale à 1.
- soit à égale distance de W_1 et W_2, l'angle formé avec ces sous-espaces étant supérieur à 45° ; homothétiques aux vecteurs $(\underline{\xi}^i - \underline{\eta}^i)$, il leur est associé une valeur propre μ comprise entre 0 et 1.

Le nombre de vecteurs propres $\underline{\beta}$ considérés est alors égal à :

$$\dim W_1 + \dim W_2 - \dim (W_1 \cap W_2).$$

Rechercher les vecteurs propres de $A_1 + A_2$ revient à effectuer une analyse en composantes principales particulière ; considérons en effet le schéma de dualité associé au tableau de données $Z = \begin{pmatrix} X \\ Y \end{pmatrix}$:

$$M = \begin{pmatrix} V_{11}^{-1} & 0 \\ 0 & V_{22}^{-1} \end{pmatrix}$$

$$\begin{array}{ccc} & Z = \begin{pmatrix} X \\ Y \end{pmatrix} & \\ E = \mathbf{R}^{p+q} & \longleftarrow & F^* \\ \Big\uparrow\Big\downarrow V & & W \Big\uparrow\Big\downarrow D_p \\ E^* & \xrightarrow[Z' = (X' \,,\, Y')]{} & F = \mathbf{R}^n \end{array}$$

- à la forme quadratique V induite par D_p sur E^* est associée la matrice :

$$V = Z D_p Z' = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}$$

- la matrice $V_{12} = V_{21}'$ décrit les covariances entre les caractères $\underline{x}^i$ et les caractères $\underline{y}^j$ si les caractères sont centrés.
- les sous-espaces W_1 et W_2 engendrés par les vecteurs $\underline{x}^i$ et les vecteurs $\underline{y}^j$ étant supposés de dimension p et q respectivement, l'espace des

individus E est muni de la métrique de matrice :

$$M = \begin{bmatrix} V_{11}^{-1} & 0 \\ 0 & V_{22}^{-1} \end{bmatrix}$$

- $W = Z' \circ M \circ Z$.

Les coordonnées des n individus dans le système des axes principaux (en toute rigueur Z doit être ici supposé centré), l'espace E étant muni de la métrique M, sont fournies par les composantes principales vecteurs propres de $W \circ D_p$; or :

$$W \circ D_p = (X' \circ V_{11}^{-1} \circ X + Y' \circ V_{22}^{-1} \circ Y) \circ D_p = A_1 + A_2.$$

Les composantes principales $\underline{c}^i$ sont donc homothétiques aux vecteurs $\underline{\beta}^i$ introduits précédemment :

$$(A_1 + A_2)\, \underline{c}^i = \mu\, \underline{c}^i$$

$$\| \underline{c}^i \| = \sqrt{\mu_i}$$

$$\underline{c}^i = \sqrt{\mu_i}\, \underline{\beta}^i \ .$$

Les vecteurs $\underline{c}^i = \sqrt{\mu_i}\, \underline{\beta}^i$, vecteurs propres de $A_1 + A_2$, ne sont donc autres que les composantes principales du nuage des individus repérés par les $(p + q)$ caractères considérés, l'espace à $(p + q)$ dimensions E étant muni de la métrique particulière M. Cette métrique reflète le souci dans le calcul des distances entre individus d'accorder aux deux paquets de caractères des importances proportionnelles aux nombres de variables considérées.

Si ce sont les positions relatives des deux sous-espaces vectoriels W_1 et W_2 qui intéressent l'analyste, (optique de l'analyse canonique classique) il lui suffira de considérer dans l'exemple cité les caractères $\underline{\beta}^1$ ou $\underline{c}^1$ ou le couple de caractères $(\underline{\xi}^1, \underline{\eta}^1)$; ces caractères suffisent pour décrire la liaison entre les deux paquets de caractères. Si par contre il s'agit de dresser un bilan des proximités entre individus (optique : analyse en composantes principales), en accordant à chaque paquet de caractères des importances similaires (en réalité proportionnelles à l'effectif du paquet), on retiendra un nombre de composantes principales suffisant ; ici par exemple, les composantes principales $\underline{c}^1$ et $\underline{c}^2$ sont nécessaires pour bien décrire les proximités entre individus.

2.2 – Une généralisation possible.

La façon de procéder décrite en 2.1 se généralise immédiatement.

Si k paquets de variables sont considérés, on peut penser décrire les positions relatives des sous-espaces $W_1, W_2, \ldots, W_k$ qui leur sont associés dans $F = \mathbf{R}^n$ en tirant les vecteurs propres de l'opérateur :

$$U = \sum_{j=1}^{k} A_j$$

où A_j désigne le D_p – projecteur associé au sous-espace W_j.

Procéder ainsi, si p_j est le nombre de variables considérées dans le $j^{\text{ème}}$ paquet, auquel correspond le tableau "individus x caractères" X_j, revient à effectuer une analyse en composantes principales sur le nuage des individus repérés dans $E = \mathbf{R}^p$ par les $p = \sum_{j=1}^{k} p_j$ caractères (en toute rigueur les tableaux X_j devraient être ici centrés), l'espace E étant muni de la métrique M de matrice :

$$M = \begin{bmatrix} V_{11}^{-1} & & 0 \\ & V_{22}^{-1} & \\ & \ddots & \\ 0 & & V_{kk}^{-1} \end{bmatrix}$$

où V_{jj} désigne la matrice de covariance associée au $j^{\text{ème}}$ paquet lorsque les variables sont centrées.

Le schéma de dualité considéré est alors le suivant :

$$\begin{array}{ccc} E = \mathbf{R}^p & \xleftarrow{\quad Z \quad} & F^* \\ M \updownarrow & V & W \updownarrow \quad D_p \\ E^* & \xrightarrow[\quad Z' \quad]{} & F = \mathbf{R}^n \end{array}$$

avec :

- $Z = \begin{bmatrix} X_1 \\ X_2 \\ . \\ . \\ . \\ X_k \end{bmatrix}$

- $V = \begin{bmatrix} V_{11} & V_{12} \ldots\ldots\ldots & V_{1k} \\ V_{21} & V_{22} \ldots\ldots\ldots & \\ \ldots\ldots\ldots\ldots\ldots & & \\ V_{k1} \ldots\ldots\ldots\ldots & & V_{kk} \end{bmatrix}$

- $W = Z' \circ M \circ Z,$

on a comme précédemment :

$$W \circ D_p = \sum_{i=1}^{k} A_i = U.$$

Rappelons que les axes principaux sont engendrés par les vecteurs propres M- normés de $V \circ M$, que les facteurs principaux sont les vecteurs propres M^{-1} normés de $M \circ V$, et que les composantes principales sont les vecteurs propres de $U = W \circ D_p$ normés à la racine carrée des valeurs propres correspondantes.

Tout vecteur $\underline{c}$ de Z' (E^*) s'écrit :

$$\underline{c} = Z' (\underline{a}) = \sum_{j=1}^{k} X'_j (\underline{a}_j) = \sum_{j=1}^{k} \underline{\xi}^j$$

où $\underline{\xi}^j = X'_j (\underline{a}_j)$ est un point de W_j.

$$\| \underline{c} \|_{D_p} = \| \sum_{j=1}^{k} \underline{\xi}^j \|_{D_p}$$

$$\| \underline{a} \|^2_{M^{-1}} = \sum_{j=1}^{k} V_{jj} (\underline{a}_j) = \sum_{j=1}^{k} \| \underline{\xi}^j \|^2_{D_p} .$$

L'analyse en composantes principales précédente conduit donc à rechercher simultanément dans les différents sous-espaces W_j les vecteurs $\underline{\xi}^j$ rendant maximum la D_p – norme de $\sum_{j=1}^{k} \underline{\xi}^j$ sous la contrainte de normalisation : $\sum_{j=1}^{k} \| \underline{\xi}^j \|^2_{D_p} = 1.$

Les résultats obtenus ne dépendent pas du choix des unités de mesure et il est équivalent en particulier de travailler sur les caractères "centrés" ou sur les caractères "centrés – réduits".

La représentation simultanée dans le plan principal (représentation dans $\mathbf{R}^p$), à l'aide de leurs projections, des vecteurs-individus et des axes de coordonnées associés aux caractères initiaux, permet d'interpréter les groupes d'individus considérés comme semblables ; les corrélations (cor) entre les caractères et les composantes principales conduisent à une description des caractères dans le plan des deux premières composantes principales (description dans $\mathbf{R}^n$) ; le "cercle des corrélations" obtenu permet parfois de donner une interprétation intéressante à ces deux composantes principales.

Remarques :

- les valeurs propres de $U = W \circ D_p$ sont comprises entre 0 et k :
  - $\underline{c} \in \cap \{W_j \mid j = 1, \ldots, k\} \Rightarrow U\underline{c} = k\underline{c}$
  - $\underline{c} \in W_1 \cap \ldots \cap W_\ell \cap W_{\ell+1}^\perp \cap \ldots \cap W_k^\perp \Rightarrow U\underline{c} = \ell\underline{c}$
  - $\underline{c} \in W_1 \cap W_2^\perp \cap \ldots \cap W_k^\perp \Rightarrow U\underline{c} = \underline{c}$.

- si chacun des paquets de caractères ne comprend qu'une seule variable, l'analyse précédente revient à l'analyse en composantes principales "réduite" où l'espace E à $p = k$ dimensions est muni de la métrique diagonale $M = D_{1/\sigma^2}$ des inverses des carrés des écarts-types.

- si chacun des paquets de caractères est composé de variables non corrélées entre elles, l'analyse précédente revient encore à une analyse en composantes principales "réduite" ; la première composante principale $\underline{c}$ est donc homothétique au vecteur normé de F rendant maximum l'indice :

$$J_c = \sum_{j=1}^{k} \sum_{j'=1}^{pj} \mathrm{cor}^2 (\underline{c}_j^{j'}, \underline{c})$$

où $\underline{c}_j^{j'}$ est la $(j')^{\text{ème}}$ composante principale obtenue dans l'analyse en composantes principales réduite du $j^{\text{ème}}$ paquet.

3 – ANALYSE EN COMPOSANTES PRINCIPALES SUR VARIABLES QUALITATIVES.

La façon de procéder décrite au § 2 se transpose immédiatement au cas où les variables considérées sont qualitatives. On utilise ici les notations introduites au § 1.

3.1 – Différentes façons d'effectuer une analyse factorielle des correspondances.

Le tableau de contingence N dont l'élément (i, j) est noté n_{ij} décrit l'imbrication des partitions associées aux deux caractères qualitatifs x et y dont les ensembles de modalités sont notés respectivement I et J.

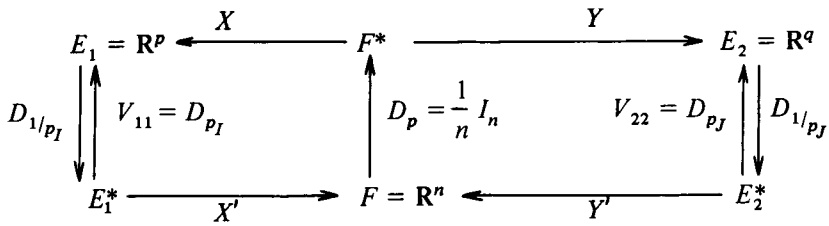
$$p_{ij} = \frac{n_{ij}}{n}$$

$$p_{i.} = \sum_{j \in J} p_{ij} \quad ; \quad p_{.j} = \sum_{i \in I} p_{ij}.$$

Le tableau $\frac{1}{n} N$ des probabilités p_{ij} est noté P.

3.1.1 *Analyse des correspondances du tableau N.*

On sait qu'effectuer une analyse factorielle des correspondances sur le tableau de contingence N (nous dirons analyse des correspondances du tableau N) revient à effectuer l'analyse canonique des deux paquets de variables indicatrices associées aux modalités respectivement de la variable x (variables $\underline{x^i}$) et de la variable y (variables $\underline{y^j}$) ; le schéma de dualité est alors le suivant :



Les facteurs canoniques sont donnés par exemple par les équations :

$$B_1 \circ B_2 \, \underline{a_i} = \lambda_i \, \underline{a_i}$$

$$\underline{b_i} = \frac{1}{\sqrt{\lambda_i}} B_2 \, \underline{a_i} \qquad \text{si} \qquad \lambda_i \neq 0.$$

Les matrices associées aux applications B_1 et B_2 ne sont autres ici que les matrices :

$$B_1 = V_{11}^{-1} V_{12} = D_{1/p_I} P \; ; \; B_2 = V_{22}^{-1} V_{21} = D_{1/p_J} P'.$$

Aux colonnes de B_1' et B_2' correspondent les vecteurs $\underline{p_J^i}$ et $\underline{p_I^j}$ qui représentent dans E_2 et E_1 respectivement les lois conditionnelles "sachant i" et "sachant j".

Ayant éliminé le couple de facteurs trivial $(\underline{a}_0, \underline{b}_0)$ de valeur propre 1, correspondant au vecteur $\underline{j} = \sum_{i \in I} \underline{x}^i = \sum_{j \in J} \underline{y}^j$ situé dans l'intersection des sous-espaces W_1 et W_2 engendrés par les variables indicatrices, le $\ell^{\text{ème}}$ couple de facteurs canoniques $(\underline{a}_\ell, \underline{b}_\ell)$ fournit le $\ell^{\text{ème}}$ codage simultané des modalités de x et y rendant maximum le produit scalaire (corrélation) entre les caractères quantitatifs $\underline{\xi} = X'(\underline{a})$ et $\underline{\eta} = Y'(\underline{b})$.

Ayant affecté les lois conditionnelles "sachant j" des masses $p_{.j}$ et muni l'espace E_1 de la métrique D_{1/p_I} (métrique du chi-deux), les facteurs canoniques $\underline{a}_\ell$ ne sont autres que les facteurs principaux du nuage $\{\underline{p}_I^j \mid j \in J\}$; la coordonnée de la loi $\underline{p}_I^j$ par rapport au $\ell^{\text{ème}}$ axe principal s'écrit :

$$< \underline{a}_\ell , \underline{p}_I^j > = \sum_{i \in I} p_i^j \, a_\ell^i = \sqrt{\lambda_\ell} \, b_\ell^j .$$

La représentation simultanée dans le plan principal des lois $\underline{p}_I^j$ et des vecteurs de base (lois certaines) correspond à l'une des deux représentations barycentriques usuelles ; si les valeurs propres λ_ℓ sont petites, on préférera représenter simultanément les modalités "i" de x et "j" de y à l'aide des codages a_ℓ^i et b_ℓ^j (représentation simultanée non barycentrique).

Remarque :

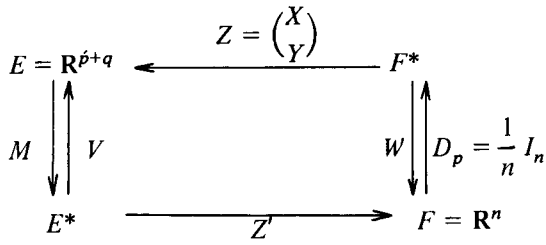
- trace $(B_1 B_2) = \Phi^2 + 1$ où Φ^2 est le "phi-deux" associé au tableau P des probabilités.
- la part d'inertie expliquée par le plan principal est égale à $\dfrac{\lambda_1 + \lambda_2}{\Phi^2}$.

3.1.2 *Analyse sur le tableau Z des indicatrices.*

Le tableau $Z = \begin{pmatrix} X \\ Y \end{pmatrix}$ désigne ici le tableau obtenu en rangeant les unes sous les autres les variables indicatrices associées respectivement aux modalités de x et de y (tableau "disjonctif complet").

3.1.2.1 Analyse en composantes principales.

On a vu au § 2.1 que l'on pouvait obtenir les caractères canoniques en tirant les vecteurs propres de l'opérateur $U = A_1 + A_2$; le schéma de dualité considéré est alors le suivant :



$$V = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix} = \begin{bmatrix} D_{p_I} & P \\ P' & D_{p_J} \end{bmatrix} ; \; M = \begin{bmatrix} V_{11}^{-1} & 0 \\ 0 & V_{22}^{-1} \end{bmatrix} = \begin{bmatrix} D_{1/p_I} & 0 \\ 0 & D_{1/p_J} \end{bmatrix}$$

Nous appelerons tableau de Burt le tableau V qui est au facteur n près celui que considérait Burt en 1950 [3].

$$U = W \circ D_p = A_1 + A_2$$

A la composante principale $\underline{c}$ vecteur propre de U correspond le facteur principal $\begin{pmatrix} \underline{a} \\ \underline{b} \end{pmatrix}$ vecteur propre de $M \circ V$:

$$(A_1 + A_2)\,\underline{c} = \mu\,\underline{c}$$

$$\underline{c} = \underline{\xi} + \underline{\eta} = X'(\underline{a}) + Y'(\underline{b}).$$

Si μ est différent de 1, $\underline{c}$ est équidistant des sous-espaces W_1 et W_2, et :

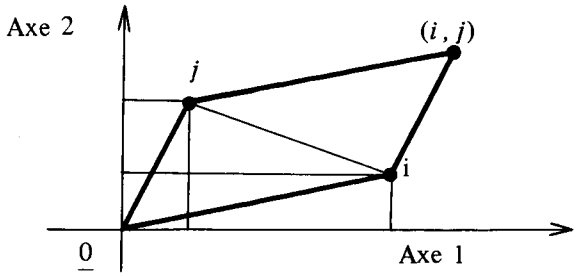
$$D_{P_I}(\underline{a}) = D_{P_J}(\underline{b}) = \frac{1}{2}\cdot$$

Les composantes principales qui sont orthogonales au vecteur $\underline{j}$, vecteur propre trivial de U de valeur propre 2, sont centrées ; il en est de même des codages $\underline{a}$ et $\underline{b}$ correspondants :

$$\sum_{i \in I} p_{i.}\; a^i = \sum_{j \in J} p_{.j}\; b^j = 0.$$

Aussi l'analyse que l'on effectue ici sur un tableau de caractères non centrés est équivalente à celle que l'on pourrait effectuer sur le tableau des indicatrices centrées.

L'individu (i, j) prenant la modalité "i" du caractère x et la modalité "j" du caractère y est repéré dans le système des axes principaux par un point dont la $\ell^{ème}$ coordonnée est $a_\ell^i + b_\ell^j$; les vecteurs de base représentatifs des modalités "i" de x et "j" de y ont respectivement pour $\ell^{ème}$ coordonnée a_ℓ^i et b_ℓ^j :



Au facteur $\frac{1}{\sqrt{2}}$ près, on retrouve, si les deux plus grandes valeurs propres sont plus grandes que 1, la représentation simultanée non barycentrique des modalités rencontrée au § 3.1.1.

Les valeurs propres μ associées à des composantes $\underline{c}$ équidistantes de W_1 et W_2 sont liées aux racines canoniques λ par la formule :

$$\mu = 1 \pm \sqrt{\lambda}.$$

Si les deux premières composantes font des angles inférieurs à 45° avec les sous-espaces W_1 et W_2 ($\mu_1 \geqslant \mu_2 > 1$), on évaluera la part d'inertie expliquée par le plan principal, pour retrouver le même résultat que dans l'analyse des correspondances du tableau de contingence, à l'aide de l'expression :

$$\frac{(\mu_1 - 1)^2 + (\mu_2 - 1)^2}{\Phi^2}.$$

Cette analyse en composantes principales effectuée sur le tableau Z est un cas particulier de l'analyse que proposait Burt [3] en 1950.

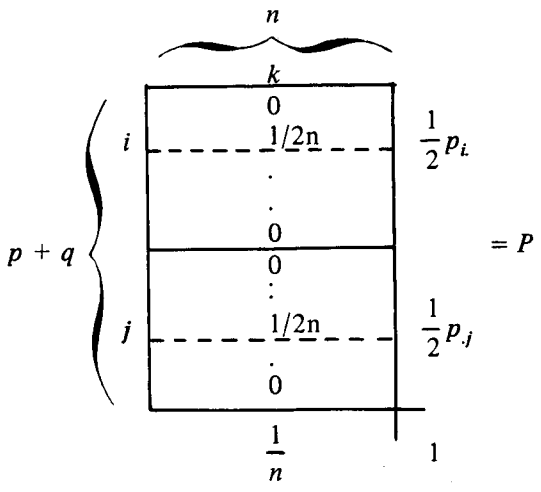
Remarque :

$(A_1 + A_2)\,\underline{c} = \mu\,\underline{c} \Rightarrow (A_1 + A_2 - I_n)^2\,\underline{c} = (\mu - 1)^2\,\underline{c} = \lambda\,\underline{c}$, les composantes principales $\underline{c}$ sont vecteurs propres de $(A_1 + A_2 - I_n)^2$ de valeurs propres λ ; les sous-espaces propres associés à la valeur propre λ de $(A_1 + A_2 - I_n)^2$ sont de dimension au moins égale à 2.

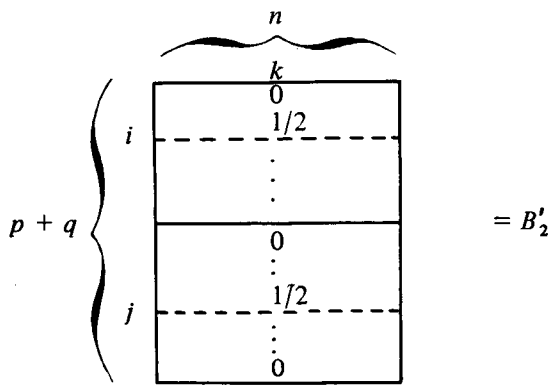
3.1.2.2 Analyse des correspondances du tableau Z.

Si l'on effectue l'analyse factorielle des correspondances du tableau Z considéré comme un tableau de contingence, on retrouve, à peu de chose

près, les résultats obtenus en 3.1.2.1 :



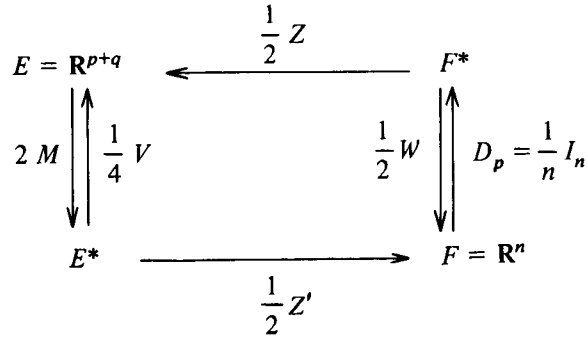
Le tableau des probabilités déduit du tableau Z n'est composé que de valeurs nulles ou égales à 1/2n.



Les lois conditionnelles associées aux individus ne sont autres que les colonnes du tableau B'_2.

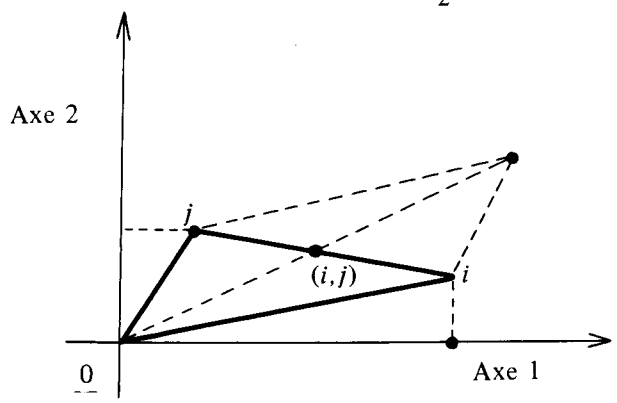
$$B'_2 = \frac{1}{2} Z.$$

Effectuer l'analyse des correspondances du tableau Z revient à effectuer une analyse en composantes principales sur le nuage des lois conditionnelles associées aux n individus. Le schéma de dualité considéré est alors le suivant :



où V et M désignent respectivement le tableau de Burt et la métrique sur E introduits en 3.1.2.1.

Les composantes principales sont à $\frac{1}{\sqrt{2}}$ près identiques à celles obtenues en 3.1.2.1 ; les facteurs principaux $\underline{a}$ et $\underline{b}$ sont identiques aux précédents au facteur $\sqrt{2}$ près ; compte tenu que des résultats identiques à une homothétie près fournissent des représentations graphiques équivalentes, on peut considérer que la représentation des modalités des deux caractères est identique à celle obtenue en 3.1.2.1, les individus étant représentés par des vecteurs identiques aux précédents au facteur $\frac{1}{2}$ près :



Remarque :

Si l'on veut retrouver les mêmes parts d'inertie que dans le cas de l'analyse des correspondances du tableau de contingence, il faut tenir compte du fait que les valeurs propres sont ici deux fois plus petites que celles trouvées en 3.1.2.1.

3.1.3. *Analyse des correspondances du tableau de Burt* ([2], [7]).

Montrons que si on effectue l'analyse factorielle des correspondances sur le tableau symétrique de Burt :

$$V = \begin{bmatrix} D_{p_I} & P \\ P' & D_{p_J} \end{bmatrix}$$

considéré comme s'il était un tableau de contingence, on retrouve encore, à peu de chose près, les résultats décrits en 3.1.2.

Ici le tableau des probabilités n'est autre que le tableau symétrique :

$$P = \frac{1}{4} V.$$

Les matrices B_1 et B_2, dont les lignes correspondent aux lois conditionnelles associées aux lignes et aux colonnes du tableau P, ont pour expression, en utilisant les mêmes notations qu'en 3.1.2. :

$$B_1 = B_2 = \frac{1}{2} M V.$$

Les facteurs canoniques (cf. 3.1.1) étant vecteurs propres de $B_1 \circ B_2$ sont ici vecteurs propres de $\frac{1}{4} MVMV$. Ces facteurs sont donc bien les homothétiques des facteurs principaux $\begin{pmatrix} a \\ \underline{b} \end{pmatrix}$ vecteurs propres de MV, trouvés en 3.1.2.1 ; la métrique considérée étant ici 2 M, les facteurs trouvés sont identiques à ces facteurs $\begin{pmatrix} a \\ \underline{b} \end{pmatrix}$ au coefficient $\sqrt{2}$ près.

Les valeurs propres dans l'analyse des correspondances du tableau de Burt sont, au facteur $\frac{1}{4}$ près, égales aux carrés des valeurs propres $\mu = 1 \pm \sqrt{\lambda}$ que l'on obtient dans l'analyse en composantes principales du tableau Z.

La représentation graphique simultanée des modalités i et des modalités j coïncide avec celle donnée en 3.1.2.2., mais ici on ne représente pas les "individus" (les couples (i, j)) dans les plans principaux ; si les deux plus grandes valeurs propres θ_1 et θ_2 obtenues sont plus grandes que 1/4, la part d'inertie expliquée par le plan principal, pour retrouver le même résultat que dans l'analyse des correspondances du tableau de contingence, sera calculée à l'aide de l'expression :

$$\frac{(2\sqrt{\theta_1} - 1)^2 + (2\sqrt{\theta_2} - 1)^2}{\Phi^2}.$$

Remarque :

$$\text{tr}\left(\frac{1}{4} MVMV\right) = \text{tr}\left(\left(\frac{A_1 + A_2}{2}\right)^2\right) = \frac{p + q + 2(\Phi^2 + 1)}{4}.$$

3.1.4 *Analyse sur le tableau des indicatrices déduit du tableau Z.*

Effectuer l'analyse des correspondances du tableau Z considéré comme un tableau de contingence (*i.e.* tableau $(p + q) \times n$, qui correspondrait à

2 n individus) (cf. 3.1.1) revient à décrire, soit par l'analyse en composantes principales (cf. 3.1.2.1), soit par l'analyse des correspondances (cf. 3.1.2.2), le tableau U des indicatrices dont Z peut être considéré comme issu.

Le tableau U peut se mettre sous la forme :

$$U = \begin{pmatrix} X & 0 \\ 0 & Y \\ I_n & I_n \end{pmatrix}.$$

Si on effectue l'analyse des correspondances de U, on obtient pour valeur propre $\frac{1}{2}\left(1 \pm \sqrt{\frac{\mu}{2}}\right)$, μ désignant toujours la valeur propre obtenue dans l'analyse en composantes principales de Z.

Si $\begin{pmatrix} \underline{a} \\ \underline{b} \\ \underline{c} \end{pmatrix}$ désigne un facteur principal de U, il est alors normé pour la métrique :

$$\frac{1}{4}\begin{bmatrix} D_{p_I} & & 0 \\ & D_{p_J} & \\ 0 & & 2\,D_p \end{bmatrix}$$

on a : $D_{p_I}(\underline{a}) = D_{p_J}(\underline{b}) = D_p(\underline{c}) = 1$.

On retrouve pour $\underline{a}$ et $\underline{b}$ les mêmes facteurs principaux qu'en 3.1.1, 3.1.2.2 et 3.1.3. Par contre, $\underline{c}$ est normé à 1, alors qu'en 3.1.2.2 considéré comme composante principale, il était normé à $\sqrt{\frac{\mu}{2}}$.

Remarques :

1) On pourrait continuer le processus en considérant U comme un tableau de contingence et en recherchant le tableau U_1 (de dimensions $(p + q + 3n) \times 4n$) des indicatrices associé à U.

On pourrait de même chercher le tableau U_2 des indicatrices, associé à U_1 considéré comme un tableau de contingence, etc. Les analyses des correspondances de tous ces tableaux fournissent les mêmes codages $\underline{a}_\varrho$ et $\underline{b}_\varrho$ de x et y que l'analyse des correspondances du tableau initial N.

2) Si l'on considère le tableau $n\ V$ comme un tableau de contingence, il est facile de voir qu'il correspond au tableau des indicatrices suivant :

X	X	0	0
0	0	Y	Y
X	0	X	0
0	Y	0	Y

tableau dont l'analyse des correspondances est équivalente à celle de V (valeurs propres égales à $(2 \pm \mu)/4$).

3.1.5 *Analyse factorielle sur tableau de distance.*

3.1.5.1 Distance du chi-deux entre parties et entre relations binaires.

Soit K un ensemble, dont les éléments sont munis des poids p_k. Cailliez (cf. [4]) a introduit la distance suivante entre parties de K :

$$\delta^2\ (A\ ,\ B) = \frac{P\ (A\ \Delta\ B)}{P\ (A)\ P\ (B)} \tag{1}$$

où $P\ (A)$, $P\ (B)$, $P\ (A\ \Delta\ B)$ désignent respectivement les masses des parties A et B et de la différence symétrique $A\ \Delta\ B$.

Dans le cas où R_1 et R_2 sont deux relations binaires considérées comme des applications de K dans $\mathfrak{P}\ (K)$, on peut définir une distance entre ces deux relations à l'aide de la formule :

$$d^2\ (R_1\ ,\ R_2) = \Sigma\ \{p_k\ \delta^2\ (R_1\ (k),\ R_2\ (k))\ |\ k \in K\}. \tag{2}$$

Si R_1 et R_2 sont des relations d'équivalence, et si $A_i\ (1 \leqslant i \leqslant p)$ et $B_j\ (1 \leqslant j \leqslant q)$ désignent respectivement les classes d'équivalence de R_1 et de R_2, on a, en posant :

$$\left.\begin{aligned} p_{ij} &= P\ (A_i \cap B_j) \\ p_{i.} &= P\ (A_i) = \Sigma\ \{p_{ij}\ |\ j = 1, \ldots, q\} \\ p_{.j} &= P\ (B_j) = \Sigma\ \{p_{ij}\ |\ i = 1, \ldots, p\} \end{aligned}\right\} \tag{3}$$

$$\left.\begin{aligned} d^2\ (R_1\ , R_2) &= p + q - 2 \sum_{i,j} p_{ij}^2 / p_{i.}\ p_{.j} \\ &= p + q - 2\ (\Phi^2 + \quad 1) \end{aligned}\right\} \tag{4}$$

où Φ^2 est le phi-deux associé au tableau P de terme général p_{ij}.

De la formule (1), on tire :

$$\left.\begin{aligned} \delta^2(A_i, A_{i'}) &= \frac{1}{p_{i.}} + \frac{1}{p_{i'.}} \\ \delta^2(B_j, B_{j'}) &= \frac{1}{p_{.j}} + \frac{1}{p_{.j'}} \\ \delta^2(A_i, B_j) &= \frac{1}{p_{i.}} + \frac{1}{p_{.j}} - 2\,\frac{p_{ij}}{p_{i.}\,p_{.j}} \end{aligned}\right\} \qquad (5)$$

3.1.5.2 Analyse factorielle sur tableau de distances.

A toute modalité i ($i \in I$) (resp. j, $j \in J$) de la variable qualitative x (resp. y) correspond l'ensemble A_i (resp. B_j) des individus possédant cette modalité. On peut donc, à partir des formules (5) du § 3.1.5.1 (K étant toujours supposé muni de la mesure uniforme donnant même poids $\frac{1}{n}$ à chaque individu) calculer la distance soit entre deux modalités i et i' de I, soit entre deux modalités j et j' de J, soit entre une modalité i de I et une modalité j de J.

On obtient ainsi un tableau de distances Δ sur l'ensemble $I \cup J$ des $p + q$ modalités de x et y. Munissant i et j des masses $p_{i.}/2$ et $p_{.j}/2$ ($p_{i.} = P(A_i)$, $p_{.j} = P(B_j)$; $\sum_i p_{i.} + \sum_j p_{.j} = 2$) on peut chercher une représentation euclidienne de l'ensemble $I \cup J$ en effectuant une analyse factorielle sur le tableau de distances Δ.

Il est immédiat de constater que cette analyse fournit les mêmes résultats que l'analyse des correspondances du tableau Z des indicatrices de x et y (cf. 3.1.2.2). En effet, si dans l'analyse des correspondances de Z, on calcule respectivement les distances entre les profils des lignes i et i', j et j', i et j, on retrouve les mêmes distances que celles données par les formules (5) ; de plus, on a le même système de masses sur I dans les deux analyses.

3.1.6 *Bilan.*

Les différentes façons d'effectuer l'analyse factorielle des correspondances de deux variables qualitatives sont résumées dans le tableau n° 1.

3.2– Généralisation.

La pratique usuelle consistant à décrire simultanément, l'ensemble K des n individus et les ensembles $I_1, \ldots, I_k$ des modalités associées aux k variables qualitatives $x_1, \ldots, x_k$ en effectuant l'analyse des correspondances du tableau des indicatrices (tableau disjontif complet),

$$Z = \begin{bmatrix} X_1 \\ \cdot \\ \cdot \\ X_k \end{bmatrix}$$

où X_i désigne le tableau des indicatrices associées à x_i, s'explique aisément compte tenu des remarques faites au § 3.1.

Procéder ainsi revient à effectuer l'analyse en composantes principales de Z (au facteur $\frac{1}{k}$ près), suivant le schéma de dualité :

$$\begin{array}{ccc} \mathbf{R}^{\Sigma p_i} = E & \xleftarrow{\quad Z \quad} & F^* \\ \Big\downarrow M \quad \Big\uparrow & V & \Big\downarrow W \quad \Big\uparrow \quad D_p = \frac{1}{n} I_n \\ E^* & \xrightarrow[\quad Z' \quad]{} & F = \mathbf{R}^n \end{array}$$

avec :

$$M = \begin{bmatrix} D_{1/p_{I_1}} & & 0 \\ & \ddots & \\ 0 & & D_{1/p_{I_k}} \end{bmatrix}$$

$$V = \begin{bmatrix} D_{p_{I_1}} & P_{12} & \cdots & P_{1k} \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ P_{k1} & \cdots & & D_{p_{I_k}} \end{bmatrix} \qquad \text{(Tableau de Burt)}$$

$$W \circ D_p = \sum_{\ell=1}^{k} A_\ell$$

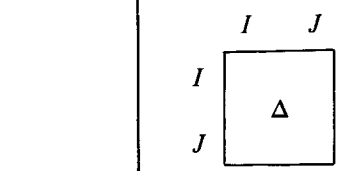
où $D_{p_{I_\ell}}$ $(1 \leqslant \ell \leqslant k)$ est la matrice diagonale des probabilités associée à la variable x_ℓ, $D_{1/p_{I_\ell}}$ son inverse, P_{ij} le tableau des probabilités correspondant au couple (x_i, x_j) et A_ℓ l'opérateur de projection associé au sous-espace de F engendré par les indicatrices des modalités de x_ℓ.

Pour éviter d'introduire le tableau Z, si n est grand, on peut faire l'analyse des correspondances du tableau de Burt V. Si on calcule à l'aide

Tableau N° 1

Différentes façons de faire l'analyse factorielle des correspondances

Tableau	Analyse	Valeur propre	Remarques
1) $I \begin{matrix} \\ \boxed{P} \end{matrix}$ J	AFC	λ	
2) $Z = \begin{matrix} I \\ J \end{matrix} \begin{bmatrix} X \\ Y \end{bmatrix}$ K	A canonique	λ	On diagonalise $A_1 \circ A_2$ ou $A_2 \circ A_1$
3) $Z = \begin{matrix} I \\ J \end{matrix} \begin{bmatrix} X \\ Y \end{bmatrix}$ K	ACP	$\mu = 1 \pm \sqrt{\lambda}$	Métrique dans $\mathbf{R}^{p+q}$: $M = \begin{bmatrix} D_1/p_I & 0 \\ 0 & D_1/p_J \end{bmatrix}$ On diagonalise : $A_1 + A_2$ ou MV
4) $Z = \begin{matrix} I \\ J \end{matrix} \begin{bmatrix} X \\ Y \end{bmatrix}$ K	AFC	$\frac{\mu}{2} = \frac{1 \pm \sqrt{\lambda}}{2}$	Métrique dans $\mathbf{R}^{p+q}$: $2M$ On diagonalise $(A_1 + A_2)/2$ ou $MV/2$ $d^2(i, i') = 1/p_{i.} + 1/p_{i'.}$; $d^2(j, j') = 1/p_{.j} + 1/p_{.j'}$ $d^2(i, j) = 1/p_{i.} + 1/p_{.j} - 2\, p_{ij}/p_{i.}\, p_{.j}$

5) $V = \begin{matrix} \\ I \\ J \end{matrix} \begin{matrix} \begin{matrix} I & J \end{matrix} \\ \begin{bmatrix} D_{p_I} & P \\ P' & D_{p_J} \end{bmatrix} \end{matrix}$	AFC	$\frac{\mu^2}{4}$	Métrique dans $\mathbf{R}^{p+q}$: $2M$ On diagonalise $(MV/2)^2$
6) $U = \begin{matrix} \\ I \\ J \\ K \end{matrix} \begin{matrix} \begin{matrix} K & K \end{matrix} \\ \begin{bmatrix} X & 0 \\ 0 & Y \\ I_n & I_n \end{bmatrix} \end{matrix}$	AFC	$\frac{1 \pm \sqrt{\frac{\mu}{2}}}{2}$	Métrique dans $\mathbf{R}^{p+q+n}$: $4 \begin{bmatrix} D_{1/p_I} & & \\ & D_{1/p_J} & \\ & & \frac{n}{2} I_n \end{bmatrix}$
7) 	AFTD	$\frac{1 \pm \sqrt{\lambda}}{2}$	$\delta^2_{ij} = P(A_i \Delta B_j) / P(A_i) P(B_j) = d^2(i, j)$ $\delta^2_{ii'} = P(A_i \Delta A_{i'}) / P(A_i) P(A_{i'}) = d^2(i, i')$ $\delta^2_{jj'} = P(B_j \Delta B_{j'}) / P(B_j) P(B_{j'}) = d^2(j, j')$

N B. Sauf dans l'analyse 3) où $D_{p_I}(\underline{a}) = D_{p_J}(\underline{b}) = \frac{1}{2}$, on a : $D_{p_I}(\underline{a}) = D_{p_J}(\underline{b}) = 1$.

des formules (5) du § 3.1.5.1 les distances entre deux modalités quelconques de l'ensemble $I = \cup \{I_\ell \mid \ell = 1, k\}$ de toutes les modalités, on peut encore, ayant muni ces modalités des bonnes masses (probabilités divisées par k de prendre la modalité), obtenir l'image euclidienne de I que donne l'analyse des correspondances de Z en effectuant une analyse factorielle sur le tableau de distances Δ ainsi construit. Dans tous les cas on représentera simultanément sur les graphiques les individus et les modalités des différents caractères.

Le tableau n° 2 résume les résultats précédents.

Tableau N° 2

Généralisation de l'Analyse des correspondances

Tableau	Analyse	Valeur propre	Remarques
$Z = \begin{bmatrix} X_1 \\ . \\ . \\ . \\ X_k \end{bmatrix}$	AFC	λ	on diagonalise $(MV)/k$ ou $(\Sigma A_i)/k$
$Z = \begin{bmatrix} X_1 \\ . \\ . \\ . \\ X_k \end{bmatrix}$	ACP	$\mu = k\lambda$	on diagonalise MV ou ΣA_i
$V = \frac{ZZ'}{n}$	AFC	λ^2	on diagonalise $(MV/k)^2$
Δ	AFTD	λ	$\delta_{ij}^2 = P(A_i \Delta B_j)/P(A_i)\,P(B_j)$

Remarque.

Effectuer l'analyse des correspondances de Z (resp. $n\,V$) considéré comme un tableau de contingence revient à faire l'analyse des correspondances du tableau U (resp. T) des indicatrices dont Z (resp. $n\,V$) peut être considéré comme issu. Les tableaux U et T s'écrivent :

$U =$

X_1	0	...	0
0	X_2		
⋮		⋱	
0		...	X_k
I_n	I_n		I_n

$T =$

X_1	X_1		X_1	0	0		0	0 0	0	0		0
0	0		0	X_2	X_2		X_2	0 0	0	0		0
0	0	0	0	0	0	0	0		X_k	X_k		X_k
X_1	0		0	X_1	0		0		X_1	0		0
0	X_2			0	X_2				0	X_2		
0			X_k	0			X_k		0			X_k

4 – ANALYSE CANONIQUE SUR VARIABLES QUALITATIVES.

4.1 Etude de la liaison entre une variable qualitative y et un paquet de variables qualitatives $x_1, x_2, \ldots, x_p$.

Le p-uple de variables qualitatives $(x_1, x_2, \ldots, x_p)$ est équivalent à la variable qualitative z dont l'ensemble des modalités est obtenu en croisant les modalités des variables x_i. Aussi la méthode consistant à étudier la

liaison entre y et les variables x_i en effectuant une analyse des correspondances entre y et z est-elle théoriquement la meilleure.

Cette façon de procéder est impraticable dès que p est supérieur à quelques unités ; elle a de plus le défaut de fournir des graphiques souvent difficilement interprétables, chaque modalité de la variable z faisant intervenir simultanément plusieurs dimensions.

Aussi dans la pratique (ici nous discutons de méthodes ne faisant appel qu'à l'algèbre linéaire) on adopte une stratégie, qui, espère-t-on, réalise un compromis entre la facilité (utilisation de techniques connues auxquelles correspondent des programmes performants et souples), la fidélité (description de la liaison avec un maximum de précision), l'interprétabilité (obtention de documents où on a des chances de noter les faits qui s'y trouvent présents) et la stabilité.

Les deux méthodes que l'on rappelle ici sont classiques :

• La première résulte de considérations géométriques : les positions relatives des sous-espaces W_0 et W engendrés respectivement par les indicatrices des modalités de y et par l'ensemble des indicatrices des modalités des x_i reflètent dans une certaine mesure la liaison qu'il s'agit d'appréhender.

• Les constats faits au § 3.1 à propos de l'analyse des correspondances ont inspiré la seconde méthode : à quoi correspond la pratique usuelle consistant à effectuer une analyse des correspondances sur le tableau, considéré comme tableau de contingence, obtenu en empilant les tableaux de probabilités associés aux p couples (y, x_i) ?

4.1.1 *Etude des positions relatives de W_0 et de W par l'analyse canonique.*

Effectuer une analyse canonique pour décrire les positions relatives des sous-espaces W_0 et W conduit à diagonaliser $A_0 \circ A$, $A \circ A_0$ ou $A + A_0$, A_0 et A désigant les opérateurs de projection associés respectivement à W_0 et W.

Notons Z le tableau où sont empilées successivement les r indicatrices associées aux modalités des variables $x_1, x_2, \ldots, x_p$ (tableau X), puis les r_0 indicatrices associées aux modalités de y (tableau Y) :

$$Z = \underbrace{\begin{pmatrix} X \\ Y \end{pmatrix}}_{n} \begin{matrix} \}r \\ \}r_0 \end{matrix} = \underbrace{\begin{pmatrix} X_1 \\ \cdot \\ \cdot \\ \cdot \\ X_p \\ Y \end{pmatrix}}_{n} \begin{matrix} \}r_1 \\ \\ \\ \\ \}r_p \\ \}r_0 \end{matrix}$$

X_i désignant le tableau $(r_i \times n)$ des indicatrices, rangées en ligne, des r_i modalités de x_i.

Le tableau de Burt V associé à l'ensemble des variables s'écrit :

$$V = Z\, D_p\, Z' = \frac{1}{n} Z\, Z' = \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix}$$

avec

$$V_{11} = X\, D_p\, X' = \begin{pmatrix} P_{11} & P_{12} & \dots & P_{1p} \\ . & & & \\ . & & & \\ . & & & \\ P_{p1} & P_{p2} & \dots & P_{pp} \end{pmatrix}$$

$$V_{21} = Y\, D_p\, X' = V'_{12} = (P_{01}, P_{02}, \dots, P_{0p})$$

$$V_{22} = Y\, D_p\, Y' = P_{00}.$$

Les tableaux P_{ij} sont les tableaux des probabilités associées aux différents couples de variables (l'indice "0" correspond à la variable y Les éléments diagonaux du tableau diagonal P_{ii} ne sont autres que les probabilités de prendre les différentes modalités de la variable "i".

A la racine canonique λ correspondent les caractères canoniques $\underline{\xi}$ et $\underline{\eta}$ vecteurs propres de $A \circ A_0$ et $A_0 \circ A$ respectivement :

$$\underline{\xi} = X'(\underline{a}) = \sum_{i=1}^{p} X'_i(\underline{a}_i)$$

$$\eta = Y'(\underline{b}).$$

Les facteurs canoniques $\underline{a}$ et $\underline{b}$ vérifient les équations :

$$\left.\begin{aligned} V_{21}\,\underline{a} &= \sqrt{\lambda}\, V_{22}\,\underline{b} \\ V_{12}\,\underline{b} &= \sqrt{\lambda}\, V_{11}\,\underline{a} \end{aligned}\right\} \qquad (1)$$

d'où l'on déduit en particulier que

$$V_{12}\, V_{22}^{-1}\, V_{21}\,\underline{a} = \lambda\, V_{11}\,\underline{a}$$

cette équation est encore équivalente aux p équations $(1 \leqslant i \leqslant p)$:

$$\sum_{j=1}^{p} (P'_{0i}\, P_{00}^{-1}\, P_{0j} - \lambda\, P_{ij})\,\underline{a}_j = 0$$

Décrire ainsi les positions relatives de W_0 et W généralise au cas qualitatif ce que l'on fait en analyse de la variance ([1]) quand on pose un modèle sans interaction.

La méthode dont on a parlé dans l'introduction, et qui consiste à étudier la liaison entre la variable y et la variable $z = (x_1, x_2, \ldots, x_p)$ au moyen de l'analyse des correspondances (équivalente à une analyse canonique, cf. § 3.1.1) revient à effectuer une analyse de la variance généralisée au sens précédent, sur un modèle avec interaction d'ordre p.

Cette double remarque suggère tout un ensemble de méthodes "intermédiaires" revenant à faire des analyses de variance généralisées sur un modèle où l'on introduit des interactions de différents ordres.

Ces méthodes conduisent à décrire par l'analyse canonique les positions relatives de W_0 et du sous-espace engendré par les indicatrices des modalités des variables x_i d'une part, et par les indicatrices associées aux modalités obtenues en faisant les croisements correspondant aux interactions choisies d'autre part.

Exemple :

Trois variables "explicatives" x_1, x_2, x_3 sont considérées ; on décide de retenir l'interaction d'ordre 2, entre x_1 et x_2. Effectuer l'analyse de la variance "généralisée" précédente revient ici à effectuer une analyse canonique sur le tableau $Z = \begin{pmatrix} X \\ Y \end{pmatrix}$ où X est obtenu en empilant les indicatrices des modalités de x_1, x_2, x_3 et (x_1, x_2). Notons qu'il est inutile d'ailleurs ici, notre propos n'étant pas de "tester" les interactions, d'introduire dans le tableau X les indicatrices des modalités de x_1 et x_2. L'analyse canonique faite ici est donc équivalente à une analyse "sans interaction" des liaisons entre y d'une part et les variables (x_1, x_2) et x_3 d'autre part.

A-t -on réalisé par un choix approprié des interactions un bon compromis entre facilité, fidélité, interprétabilité et stabilité ?

Si ce choix n'est pas "dirigé", il semble que l'on ne sera satisfait qu'en ce qui concerne la facilité et l'interprétabilité.

Pour diriger le choix, les deux façons de procéder suivantes viennent à l'esprit :

---

(1) Nous préférons parler en termes d'analyse de la variance plutôt qu'en termes d'analyse discriminante, car l'analyse de la variance fait référence à toute une pratique où l'un des objectifs fondamentaux est d'obtenir un modèle aussi simple que possible ; c'est ici évidemment et obligatoirement l'un de nos objectifs (cf. début du § 4.1).

a) Procéder comme en analyse de la variance.

On part du modèle le plus complet, c'est-à-dire en supposant une interaction d'ordre p, et on procède pas à pas en utilisant les indices dont on se sert pour faire les tests en analyse de la variance multidimensionnelle (cf. § 5). Ces indices ne donnent qu'une indication sur les interactions qu'il faut retenir ; en effet les hypothèses qui doivent être faites pour que ces indices aient un sens précis, c'est-à-dire pour que l'on puisse calculer les probabilités de dépasser les valeurs de l'indice, ne sont pas réalisées ici ; en effet, dans l'optique de l'analyse de la variance multidimensionnelle, la variable y, qui a été considérée comme la variable vectorielle des indicatrices de ses modalités, n'est évidemment pas gaussienne. Les valeurs des indices obtenues seront utilisées "par analogie" comme de simples éléments d'appréciation. On mesure donc, compte tenu du nombre d'interactions qu'il faut considérer, les difficultés de cette façon de procéder qui ne sera retenue que dans des cas très particuliers (faible nombre de variables explicatives et de modalités, nombre suffisant d'individus, etc.).

b) Décrire les liaisons entre les variables x_i avant de mettre ces variables en relation avec y.

Cette description se fera (cf. § 3) en effectuant l'analyse des correspondances du tableau X, ou du tableau de Burt V_{11} associé aux variables x_i. On pourra ainsi regrouper, au vu des représentations graphiques, les individus en classes homogènes ; ces classes interprétées permettent de définir une nouvelle variable qualitative tenant compte de l'ensemble des variables x_i, variable dont on étudiera la liaison avec y à l'aide de l'analyse des correspondances. Cette manière de procéder tient compte manifestement des interactions d'ordre 2.

4.1.2 *Analyse des correspondances du tableau* V_{21}

Effectuer l'analyse des correspondances du tableau V_{21} revient à étudier la liaison de y avec les x_i en croisant l'ensemble des modalités de y avec l'ensemble des modalités de toutes les variables explicatives.

Si V_1 désigne la matrice diagonale ayant mêmes éléments diagonaux que V_{11} ($V_1 = \text{diag}\,(V_{11})$), et si l'on pose $V_2 = p\,P_{00} = p\,V_{22}$, tout couple $(\underline{a}, \underline{b})$ de facteurs associés issus de l'analyse des correspondances de V_{21} et relatif à la valeur propre μ, vérifie les équations :

$$\left.\begin{aligned} V_{21}\,\underline{a} &= \sqrt{\mu}\; V_2\,\underline{b} \\ V_{12}\,\underline{b} &= \sqrt{\mu}\; V_1\,\underline{a} \end{aligned}\right\} \qquad (2)$$

Remarque :

On peut noter que les équations (2) de l'analyse des correspondances de V_{21} s'obtiennent à partir des équations (1) de l'analyse canonique de W_0 et W, en remplaçant V_{11}, V_{22} et λ respectivement par V_1, V_2 et μ.

Posant

$$\underline{\xi} = X'(\underline{a}) = \sum_{i=1}^{p} X_i'(\underline{a_i}) = \sum_{i=1}^{p} \underline{\xi_i}, \qquad \text{et} \qquad \underline{\eta} = Y'(\underline{b})$$

et désignant par A_i le projecteur sur l'espace W_i engendré par les indicatrices des modalités de x_i, les équations (2) sont équivalentes aux $p+1$ équations $(1 \leqslant i \leqslant p)$:

$$A_0\,\underline{\xi} = p\sqrt{\mu}\,\underline{\eta}$$

$$A_i\,\underline{\eta} = \sqrt{\mu}\,\underline{\xi_i}$$

d'où l'on déduit que

$$\Big(\sum_{i=1}^{p} A_i\Big)\underline{\eta} = \sqrt{\mu}\,\underline{\xi}$$

d'où

$$\left.\begin{aligned} \Big(\frac{1}{p}\sum_{i=1}^{p} A_i\Big) \circ A_0\,\underline{\xi} &= \mu\,\underline{\xi} \\ A_0 \circ \Big(\frac{1}{p}\sum_{i=1}^{p} A_i\Big)\underline{\eta} &= \mu\,\underline{\eta} \end{aligned}\right\} \qquad (3)$$

L'analyse des correspondances effectuée sur le tableau V_{21} donne les mêmes résultats que la même analyse effectuée sur le tableau des indicatrices dont V_{21} peut être considéré comme issu (cf. § 3) ; ce tableau s'écrit :

X_1	0	. . . .	0
0	X_2	0 . . .	0
0	0		X_p
Y	Y		Y

Ce tableau fait songer à un plan factoriel à deux facteurs.

4.1.3 *Comparaison des deux techniques précédentes.*

L'analyse canonique de W_0 et de W revient à rechercher

$$\underline{\xi} = X'(\underline{a}) = \sum_{i=1}^{p} X'_i \underline{a}_i = \sum_{i=1}^{p} \underline{\xi}_i, \qquad \text{si} \qquad \underline{\xi}_i = X'_i(\underline{a}_i)$$

et

$$\underline{\eta} = Y'(\underline{b})$$

de telle sorte que $D_p(\underline{\xi}, \underline{\eta})$ soit maximum, sous les conditions de normalisation :

$$\| \underline{\xi} \|^2_{D_p} = \| \sum_{i=1}^{p} \underline{\xi}_i \|^2_{D_p} = 1 \tag{4}$$

$$\| \underline{\eta} \|^2_{D_p} = 1. \tag{5}$$

Le maximum de $D_p(\xi, \eta)$ est égal à $\sqrt{\lambda}$ où λ désigne la valeur propre associée au couple $(\underline{\xi}, \underline{\eta})$.

Dans le cas de l'analyse des correspondances du tableau V_{21}, on peut montrer (cf. [6]) que l'on recherche toujours le couple $(\underline{\xi}, \underline{\eta})$ rendant maximum $D_p(\underline{\xi}, \underline{\eta})$ (ou plus exactement $\frac{1}{p} D_p(\underline{\xi}, \underline{\eta})$) mais sous les conditions de normalisation :

$$\frac{1}{p} \sum_{i=1}^{p} \| \underline{\xi}_i \|^2_{D_p} = 1 \tag{6}$$

$$\| \underline{\eta} \|^2_{D_p} = 1. \tag{7}$$

Le maximum de $\frac{1}{p} D_p(\underline{\xi}, \underline{\eta})$ est égal à $\sqrt{\mu}$, μ étant la valeur propre associée au couple $(\underline{\xi}, \underline{\eta})$.

La différence entre l'analyse canonique de W_0 et W, et l'analyse des correspondances de V_{21} réside donc dans les conditions de normalisation (4) et (6).

Si les variables explicatives x_i sont deux à deux indépendantes les sous-espaces W_i $(1 \leqslant i \leqslant p)$ se coupent orthogonalement suivant la droite des constantes Δ_j.

On déduit alors des équations (3) et du fait que $\underline{\xi} = \underline{\eta} = \underline{j}$ est solution des deux analyses précédentes (la solution triviale relative a la valeur

propre $\lambda = \mu = 1$) qu'effectuer l'analyse des correspondances de V_{21} revient à effectuer l'analyse canonique de W_0 et de W.

Notons que l'on aurait pu aussi obtenir ce résultat à partir des conditions de normalisation (4) et (6) qui pour des facteurs non triviaux, i.e. centrés ($\xi \in W^-$), sont équivalentes, au coefficient p près.

Dans la pratique, on se trouve dans cette situation (où pour un facteur centré $\mu = \lambda/p$) quand on a un plan d'expérience orthogonal, la variable à expliquer étant qualitative (cf. [5]).

4.2 – Etude de la liaison entre deux paquets de variables qualitatives.

Supposons que l'on veuille expliquer les liaisons existant entre deux paquets de variables qualitatives, $x_1, x_2, \ldots, x_p$ d'une part, $y_1, y_2, \ldots, y_q$ d'autre part. On va généraliser la façon de procéder décrite en 4.1.

Comme précédemment la méthode consistant à étudier à l'aide de l'analyse des correspondances la liaison entre $t = (x_1, x_2, \ldots, x_p)$ et $z = (y_1, y_2, \ldots, y_q)$ doit être considérée comme impraticable.

4.2.1 *Etude par l'analyse canonique.*

L'analyse canonique permet d'étudier la position relative du sous-espace W engendré par l'ensemble des indicatrices des modalités associées à toutes les variables x_i, et du sous-espace W' engendré par l'ensemble des indicatrices des modalités associées à toutes les variables y_j. On diagonalisera alors $A \circ B$ et $B \circ A$ ou $A + B$, A et B désignant les opérateurs de projection associés à W et W' respectivement.

Comme précédemment, entre les deux techniques extrêmes consistant à exploiter aucune interaction et les interactions d'ordre le plus élevé, il existe tout un ensemble de techniques intermédiaires où l'on fait intervenir des interactions de divers ordres.

On a intérêt à décrire d'une part les liaisons entre les variables x_i, d'autre part les liaisons entre les variables y_j, en effectuant l'analyse des correspondances des tableaux disjonctifs complets X et Y respectivement associés aux variables x_i et aux variables y_j, avant d'étudier les liaisons entre ces deux paquets de variables. Comme précédemment, l'imbrication des deux partitions de l'ensemble des individus obtenues dans ces deux analyses, sera décrite par l'analyse des correspondances.

4.2.2 *Etude par l'analyse des correspondances*

Pour étudier la liaison entre les deux paquets de variables $x_1, \ldots, x_p$ et $y_1, \ldots, y_q$, on croise l'ensemble des modalités des variables x_i avec

l'ensemble des modalités des variables y_j, et l'on effectue l'analyse des correspondances du tableau $V_{12} = X D_p Y' = X Y'/n$ ainsi obtenu. Le tableau V_{12} est donc un tableau comportant $p\,q$ blocs, et dont le bloc (i, j) $(1 \leqslant i \leqslant p\ ;\ 1 \leqslant j \leqslant q)$ est le tableau P_{ij} des probabilités associées au couple $(x_i\ y_j)$.

L'analyse des correspondances de V_{12} revient à extraire valeurs propres et vecteurs propres de

$$\left(\frac{1}{p}\sum_{i=1}^{p} A_i\right) \circ \left(\frac{1}{q}\sum_{j=1}^{q} B_j\right) \quad \text{et de} \quad \left(\frac{1}{q}\sum_{j=1}^{q} B_j\right) \circ \left(\frac{1}{p}\sum_{i=1}^{p} A_i\right)$$

où A_i et B_j désignent les opérateurs de projection associés aux sous-espaces respectivement engendrés par les indicatrices des modalités des variables x_i, et par les indicatrices des modalités des variables y_j.

Effectuer l'analyse des correspondances de V_{12} est équivalent à effectuer l'analyse des correspondances du tableau des indicatrices dont V_{12} peut être considéré comme issu (cf. § 3). Ce tableau s'écrit, si X_i et Y_j désignent respectivement les tableaux associés aux indicatrices des modalités de x_i et de y_j :

X_1	X_1		X_1	0	0		0	0	0		0	0	0		0
0	0		0	X_2	X_2		X_2	0	0		0	0	0		0
0	0		0	0	0		0	0	0		0	X_p	X_p		X_p
Y_1	0		0	Y_1	0		0					Y_1	0		0
0	Y_2		.		Y_2								Y_2		
0			Y_q	0			Y_q					0			Y_q

Ce tableau fait penser de façon encore plus évidente que précédemment à un plan d'expérience généralisé (plan factoriel), les paquets d'indicatrices jouant ici le rôle que tiennent les "1" en analyse de la variance.

4.2.4 – *Comparaison des deux techniques précédentes.*

L'analyse canonique de W et W' revient à rechercher le couple $(\underline{\xi}, \underline{\eta})$, avec

$$\left.\begin{aligned} \underline{\xi} = X'(\underline{a}) = \sum_{i=1}^{p} X'_i \underline{a}_i = \sum_{i=1}^{p} \underline{\xi}_i, \qquad \text{si} \qquad \underline{\xi}_i = X'_i \underline{a}_i \\ \underline{\eta} = Y'(\underline{b}) = \sum_{j=1}^{q} Y'_j \underline{b}_j = \sum_{j=1}^{q} \underline{\eta}_j, \qquad \text{si} \qquad \underline{\eta}_j = Y'_j \underline{b}_j \end{aligned}\right\} \tag{8}$$

maximisant

$$D_p(\underline{\xi}, \underline{\eta}) \tag{9}$$

sous les conditions de normalisation :

$$\left.\begin{aligned} \| \underline{\xi} \|^2_{D_p} = \| \sum_{i=1}^{p} \underline{\xi}_i \|^2_{D_p} = 1 \\ \| \underline{\eta} \|^2_{D_p} = \| \sum_{j=1}^{q} \underline{\eta}_j \|^2_{D_p} = 1 \end{aligned}\right\} \tag{10}$$

tandis que l'analyse des correspondances de V_{12} revient (cf. [6]) à maximiser la même quantité (ou plus exactement $D_p(\underline{\xi}, \underline{\eta})/p\,q$) mais sous les conditions de normalisation :

$$\left.\begin{aligned} \frac{1}{p} \sum_{i=1}^{p} \| \underline{\xi}_i \|^2_{D_p} = 1 \\ \frac{1}{q} \sum_{j=1}^{q} \| \underline{\eta}_j \|^2_{D_p} = 1 \end{aligned}\right\} \tag{11}$$

Dans le cas où les variables $x_i (1 \leqslant i \leqslant p)$ sont deux à deux indépendantes, et où de même les variables y_j sont deux à deux indépendantes, on peut montrer comme précédemment (cf. 4.1.3), qu'analyse des correspondances de V_{12} et analyse canonique de W et de W' sont équivalentes. Dans ce cas, on a (cf. [6]) $\mu = \lambda/p\,q$, μ désignant toujours (pour un facteur centré) la valeur propre de l'analyse des correspondances, et λ celle de l'analyse canonique.

Dans le cas général, on peut dire que la différence entre analyse canonique de W et de W' et analyse des correspondances de V_{12} réside dans le fait que

cette dernière adopte pour métrique induite dans W la métrique correspondant à "l'indépendance" des variables $x_i (1 \leqslant i \leqslant p)$ et pour métrique induite dans W' celle associée à "l'indépendance" des variables $y_j (1 \leqslant j \leqslant q)$.

4.3 – Conclusion.

Les différentes remarques qui ont été faites à propos des deux techniques qui ont été proposées montrent qu'elles sont toutes les deux également justifiées. Compte tenu des approximations faites dans les deux cas (absence de fidélité au niveau de la description des liaisons), on ne saurait se prononcer pour l'une ou pour l'autre. L'avantage de la seconde réside dans la facilité d'emploi (utilisation d'un programme d'analyse des correspondances).

5 – A PROPOS DE L'ANALYSE DE VARIANCE MULTIDIMENSIONNELLE.

5.1 – Rappel sur l'analyse de variance usuelle.

5.1.1 – *Equivalence entre analyse de variance et analyse en composantes principales.*

Un modèle d'analyse de variance peut se mettre sous la forme :

$$\underset{(n,1)}{\underline{y}} = \underset{(n,p)}{X'} \underset{(p,1)}{(\underline{\beta})} + \underline{\epsilon}$$

avec

$$E(\underline{y}) = X'(\underline{\beta})$$

$$\text{Var } \underline{y} = \sigma^2 I_n.$$

X' n'est rien d'autre que le tableau des indicatrices des modalités des facteurs intervenant, avec en plus le vecteur constant $\underline{j}$, et le cas échéant les indicatrices associées aux variables produit de deux facteurs, s'il y a interaction, ou les vecteurs associés aux n réalisations de variables quantitatives si on est en analyse de covariance.

Pour estimer $X'(\beta)$, on projette $\underline{y}$, $\mathbf{R}^n$ étant muni de la métrique usuelle $N = I_n$ (1), sur l'espace W engendré par les colonnes de X'.

Soit $\underline{y}^* = A\ \underline{y}$ la projection de $\underline{y}$ sur W.

On peut encore considérer que la recherche de $\underline{y}^*$ est équivalente à l'analyse canonique entre W et le sous-espace W_0 de dimension 1 engendré

(1) Nous supposerons, dans tout le § 5, $\mathbf{R}^n$ muni de cette métrique.

par $\underline{y}$, puisque cela revient à chercher deux droites, l'une dans W_0 (i.e. W_0 lui-même) et l'autre dans W (la droite engendrée par $\underline{y}^*$), faisant entre elles un angle minimum.

Si A_0 désigne le projecteur sur W_0, on est donc ramené à diagonaliser $A_0 \circ A$ ou $A \circ A_0$ ou $A + A_0$, ce qui revient encore à faire l'analyse en composantes principales du tableau (1) :

$$\underset{(n\,,\,p+1)}{Z'} = (\underline{y}\,,\,X'),$$

$\mathbf{R}^n$ étant muni de la métrique I_n, et $\mathbf{R}^{p+1}$ de la métrique M telle que :

$$M = \begin{pmatrix} (y'\,y)^{-1} & 0 \\ 0 & (XX')^- \end{pmatrix}$$

où $(X\,X')^-$ désigne une inverse généralisée de $X\,X'$. L'analyse de variance se ramène donc, comme toute régression, à une analyse en composantes principales

5.1.2 – *Tests.*

Supposons que l'on veuille tester l'hypothèse K :

$$\underset{(s\,,\,p)}{D}\ \underset{(p\,,\,1)}{\underline{\beta}} = \underline{0}$$

où D est de rang s.

Cela revient encore à tester l'hypothèse que $E(\underline{y}) = X'\,\underline{\beta}$ appartient au sous-espace de $\mathbf{R}^n$, W_1, défini par :

$$W_1 = \{\underline{z} \mid \underline{z} = X'\,\underline{\beta},\ D\,\underline{\beta} = \underline{0}\}.$$

Notons que si W est de dimension r, W_1 est de dimension $t = r - s$.

Soit $\underline{y}^{**}$ la projection de $\underline{y}$ sur W_1 ; alors le test usuel de l'hypothèse $K : E(\underline{y}) \in W_1$ contre l'hypothèse : $E(\underline{y}) \in W$ est construit à partir du rapport :

$$F_1 = \frac{\| \underline{y}^* - \underline{y}^{**} \|^2}{\| \underline{y} - \underline{y}^* \|^2}.$$

(1) Nous entendons ici ACP au sens de la recherche du tenseur d'ordre s le plus proche du tenseur de $\mathbf{R}^{p+1} \otimes \mathbf{R}^n$ défini par Z' ($s <$ rang Z'), $\mathbf{R}^{p+1}$ (resp. $\mathbf{R}^n$) étant muni de la métrique M (resp. $N = I_n$).

On rejettera l'hypothèse K si F_1 est supérieur à une certaine constante, facile à déterminer si $\underline{y}$ est un vecteur gaussien, et si l'on se fixe un seuil de première espèce égal à α. En effet dans le cas gaussien $F = \frac{n-r}{s} F_1$ suit, si l'hypothèse K est vérifiée, une loi de Fisher-Snedecor à s et $n-r$ degrés de liberté respectivement.

N.B. Au lieu du rapport F_1, on considère souvent le rapport

$$G = \frac{\| \underline{y} - \underline{y}^{**} \|^2}{\| \underline{y} - \underline{y}^{*} \|^2},$$

qui est égal, d'après le théorème des trois perpendulaires, à $1 + F_1$.

Supposons que l'on veuille tester le modèle lui même, et posons :

$$X' = (X'_1 , \underline{j})$$

$$\underline{\beta} = \begin{pmatrix} \beta_1 \\ \mu \end{pmatrix}.$$

On a $$X' \underline{\beta} = X'_1 \underline{\beta}_1 + \mu \underline{j}$$

et l'hypothèse K s'écrit alors :

$$X'_1 \underline{\beta}_1 = \underline{0},$$

ce qui revient à dire que W_1 est la droite des constantes.

Si $\bar{y}$ désigne la moyenne empirique des composantes y_i de $\underline{y}$, $y^{**} = \bar{y}\, \underline{j}$ et le test portera donc sur le rapport

$$F = \frac{(n-r) \| \underline{y}^{*} - \bar{y}\, \underline{j} \|^2}{(r-1) \| \underline{y} - \underline{y}^{*} \|^2} = \frac{n-r}{r-1} F_1.$$

Notons que le rapport G introduit plus haut, et égal à $1 + F_1$, s'écrit :

$$G = \frac{\| \underline{y} - \bar{y}\, \underline{j} \|^2}{\| \underline{y} - \underline{y}^{*} \|^2}.$$

5.1.3 – *Cas où la variable y est qualitative.*

Dans ce cas, si Y' désigne le tableau des indicatrices de y, et W_0 le sous-espace engendré par ces indicatrices, on fera l'analyse canonique de W_0, qui n'est plus ici de dimension 1, et de W (cf. 4.1.1).

Si en particulier on a un plan orthogonal (i.e. si les différents facteurs d'analyse de variance sont indépendants, dans le sens que les sous-espaces

vectoriels associés aux indicatrices de chaque facteur se coupent orthogonalement suivant la droite des constantes), l'analyse canonique précédente est équivalente à faire l'analyse des correspondances du tableau croisant les modalités de y avec l'ensemble des modalités de tous les facteurs (cf. 4.1.3).

5.2 – Analyse de variance multidimensionnelle.

5.2.1 – *Le modèle.*

Au lieu d'avoir pour chaque observation $i\,(1 \leqslant i \leqslant n)$ une valeur y_i d'une variable y à expliquer, on a un vecteur $\underline{y}_i\,(\underline{y}_i' = (y_i^1, y_i^2, \ldots, y_i^q))$ de q valeurs à expliquer.

On pose alors le modèle suivant :

$$\underset{(n\,,\,q)}{Y'} = (\underline{y}^1, \underline{y}^2, \ldots, \underline{y}^q) = \underset{(n\,,\,p)}{X'}\ \underset{(p\,,\,q)}{B} + \underset{(n\,,\,q)}{E'}$$

$$E(Y') = X'B$$

avec

$$B = (\underline{\beta}^1, \underline{\beta}^2, \ldots, \underline{\beta}^q)$$

$$E' = (\underline{\epsilon}^1, \underline{\epsilon}^2, \ldots, \underline{\epsilon}^q)$$

$\underline{y}^j\,(\underline{y}'^j = (y_1^j, y_2^j, \ldots, y_n^j))$ désigne le vecteur des n observations de la $j^{\text{ème}}$ $(1 \leqslant j \leqslant q)$ variable à expliquer, et $\underline{\beta}^j$ le vecteur $(p \times 1)$ des paramètres associés.

Le modèle précédent est équivalent aux q modèles :

$$\underline{y}^j = X'\underline{\beta}^j + \underline{\epsilon}^j$$

$$E(\underline{y}^j) = X'\underline{\beta}^j,$$

On supposera de plus que :

$$\text{Var}\ \underline{y}^j = \sigma_j^2\, I_n.$$

Pour estimer les paramètres du modèle précédent (i.e. B) ou $E(Y') = X'B$, il suffit de faire q analyses de variance unidimensionnelle (cf. 5.1) i.e. de projeter chaque $\underline{y}^j$ en $\underline{y}^{j*}$ sur l'espace W engendré par les colonnes de X'.

5.2.2 – *Tests.*

5.2.2.1 Etude d'un cas particulier ; lien avec l'analyse en composantes principales.

Supposons que l'on veuille tester l'hypothèse K :

$$\underset{(s\,,\,p)}{D}\ \ \underset{(p\,,\,q)}{B} = 0$$

où D est de rang s.

Cette hypothèse est encore équivalente à :

$$D\ \underline{\beta}^1 = 0$$
$$D\ \underline{\beta}^2 = 0$$
$$\vdots$$
$$D\ \underline{\beta}^q = 0.$$

Soit W_1 le sous-espace de W défini par :

$$W_1 = \{\underline{z} \mid \underline{z} = X'\,\underline{\beta}\,,\, D\,\underline{\beta} = 0\}.$$

Alors l'hypothèse K est encore équivalente à :

$$E\,(\underline{y}^1) \in W_1$$
$$E\,(\underline{y}^2) \in W_1$$
$$\vdots$$
$$E\,(\underline{y}^q) \in W_1.$$

Si $\underline{u} = \begin{pmatrix} u_1 \\ \vdots \\ u_q \end{pmatrix} \in (\mathbf{R}^q)^*$ et si $\underline{z} = \Sigma\ \{u_j\,\underline{y}^j \mid j = 1\,,\,q\} = Y'\,\underline{u}$, l'hypothèse

K peut aussi s'écrire

$$\forall\ \underline{u} \in (\mathbf{R}^q)^* : E\,(Y'\underline{u}) \in W_1.$$

Supposons que $\underline{u}$ soit fixé et que l'on veuille tester si $\underline{z} = Y'\,\underline{u}$, l'hypothèse $K\,(\underline{u})$: $E\,(\underline{z}) \in W_1$.

Posant $\underline{\beta} = B\,\underline{u}$, $\underline{\epsilon} = E'\,\underline{u}$, on a le modèle :

$$\underline{z} = X'\,\underline{\beta} + \underline{\epsilon}$$

avec

$$E\,(\underline{z}) = X'\,\underline{\beta} \in W.$$

Pour tester l'hypothèse $K(\underline{u})$ contre l'hypothèse $E(\underline{z}) \in W$, on formera donc d'après 5.1.2 le rapport :

$$F_1(\underline{u}) = \frac{\| \underline{z}^* - \underline{z}^{**} \|^2}{\| \underline{z} - \underline{z}^* \|^2} = \frac{\| (A - A_1)\, \underline{z} \|^2}{\| (I_n - A)\, \underline{z} \|^2}$$

où $\underline{z}^* = A\, \underline{z}$ (resp. $\underline{z}^{**} = A_1\, \underline{z}$) désigne la projection de $\underline{z}$ sur W (resp. W_1) et A (resp. A_1) l'opérateur de projection sur W (resp. W_1).

Compte tenu de ce que $\underline{z} = Y'\, \underline{u}$, $F_1(\underline{u})$ s'écrit encore, en notation matricielle :

$$F_1(\underline{u}) = \frac{\underline{u}'\, Y\, (A - A_1)\, Y'\, \underline{u}}{\underline{u}'\, Y\, (I_n - A)\, Y'\, \underline{u}}$$

et on acceptera l'hypothèse $K(\underline{u})$ si $F_1(\underline{u})$ est inférieur à un certain seuil, facile à calculer sous des hypothèses de normalité (et pour un risque de première espèce α fixé) puisqu'alors comme on l'a déjà dit $\frac{n-r}{s} F_1(\underline{u})$ suit une loi de Fisher-Snedecor à s et $n - r$ degrés de liberté respectivement.

Si maintenant on veut tester l'hypothèse K, on recherchera la combinaison linéaire $\underline{u}$ assurant le maximum de $F_1(\underline{u})$ de façon à avoir la valeur de $F_1(\underline{u})$ la plus défavorable au choix de l'hypothèse K.

Si l'on pose :

$$L(\underline{u}, \underline{u}) = \underline{u}'\, Y\, (A - A_1)\, Y'\, \underline{u}$$
$$H(\underline{u}, \underline{u}) = \underline{u}'\, Y\, (I_n - A)\, Y'\, \underline{u}$$

on est ramené à maximiser $L(\underline{u}, \underline{u})$ sous la contrainte $H(\underline{u}, \underline{u}) = 1$, ce qui revient à rechercher le premier facteur de l'analyse en composantes principales du tableau $(A - A_1)\, Y'$, $\mathbf{R}^q$ étant muni de la métrique H^{-1} (1), (et $\mathbf{R}^n$ de la métrique $N = I_n$) ; $\underline{u}$ est le vecteur propre normé ($H(\underline{u}, \underline{u}) = 1$) associé à la plus grande valeur propre λ_1 de $H^{-1} L$ (2), $\underline{c} = (A - A_1)\, Y'\, u$ est la composante principale associée, tandis que la valeur maximale de $F_1(\underline{u})$ est égale à λ_1.

5.2.2.2 Lien avec l'analyse canonique.

Au lieu de considérer le rapport $F_1(\underline{u})$, nous considérerons ici la quantité $G(\underline{u}) = 1 + F_1(\underline{u})$.

---

(1) H étant supposée régulière.

(2) Si H n'est pas régulière, u sera tel que $L\, \underline{u} = \lambda_1\, H\, \underline{u}$, λ_1 étant la valeur la plus élevée vérifiant $L\, u = \lambda\, H\, u$.

Posant $\underline{v} = \underline{z} - \underline{z}^{**}$, $\underline{t}^j = \underline{y}^j - \underline{y}^{j**}$ ($1 \leqslant j \leqslant q$; $\underline{y}^{j**}$ désignant la projection de $\underline{y}^j$ sur W_1) , $G(\underline{u})$ peut s'écrire d'après 5.1.2 :

$$G(\underline{u}) = \frac{\| \underline{z} - \underline{z}^{**} \|^2}{\| \underline{z} - \underline{z}^{*} \|^2} = \frac{\| \underline{v} \|^2}{\| \underline{v} - \underline{v}^{*} \|^2} = \frac{\| \Sigma \{u_j \underline{t}^j \mid j = 1 , q\} \|^2}{\| \Sigma \{u_j (\underline{t}^j - \underline{t}^{j*}) \mid j = 1 , q\} \|^2}$$

où $\underline{v}^*$ (resp. $\underline{t}^{j*}$) désigne la projection de $\underline{v}$ (resp. $\underline{t}^j$) sur W.

Maximiser $F_1(\underline{u})$ est encore équivalent à maximiser $G(\underline{u})$, ce qui revient à :

$$\text{minimiser} \qquad \| \Sigma \{u_j (\underline{t}^j - \underline{t}^{j*}) \mid j = 1 , q\} \|^2$$

sous la contrainte : $\| \Sigma \{u_j \underline{t}^j \mid j = 1 , q\} \|^2 = 1$.

On est donc ramené à rechercher la combinaison linéaire

$$\underline{v} = \Sigma \{u_j \underline{t}^j \mid j = 1 , q\}$$

de norme 1 la plus proche de W. Si W_T désigne le sous-espace de $\mathbf{R}^n$ engendré par $\underline{t}^1, \ldots, \underline{t}^q$ le problème précédent revient à trouver l'élément normé de W_T le plus proche de W, ce qui revient à faire l'analyse canonique de W et W_T.

On diagonalisera donc, si A_T (resp. A) désigne l'opérateur de projection sur W_T (resp. W), $A_T \circ A$ ou $A \circ A_T$ ou encore $A + A_T$ .Si λ_1 est la valeur propre la plus élevée de $A_T \circ A$ (ou $A \circ A_T$) la valeur maximale de $G(\underline{u})$ est $\dfrac{1}{1 - \lambda_1}$ tandis que celle de $F_1(\underline{u})$ est $\dfrac{\lambda_1}{1 - \lambda_1}$.

Remarque.

Dans le cas où l'on teste le modèle lui-même, W_1 est la droite des constantes, et l'espace W_T n'est rien d'autre que l'espace engendré par les $\underline{y}^j$ après centrage. Si de plus on a une analyse de variance à un seul facteur (W est alors engendré par les indicatrices de ce facteur) on retrouve les équations de l'analyse factorielle discriminante classique.

5.2.2.3 Cas général.

Supposons que l'on veuille tester l'hypothèse

$$\underset{(s\,,\,p)}{D} \quad \underset{(p\,,\,q)}{B} \quad \underset{(q\,,\,t)}{M} = 0$$

où D et M sont connus, M étant de rang t ($t \leqslant q$).

Posons

$$\underset{(n\,,\,t)}{S'} = \underset{(n\,,\,q)}{Y'}\ \underset{(q\,,\,t)}{M}$$

$$\underset{(p\,,\,t)}{C} = \underset{(p\,,\,q)}{B}\ \underset{(q\,,\,t)}{M}$$

On a alors

$$S' = X' C + E' M$$

avec $E(S') = X' C$

et l'on désire tester l'hypothèse

$$D C = 0.$$

On est donc ramené au problème précédent (cf. 5.2.2.1 et 5.2.2.2) si l'on raisonne sur S' et C à la place de Y' et B. On effectuera donc soit une analyse en composantes principales, soit une analyse canonique.

BIBLIOGRAPHIE

[1] BENZECRI J.P. – "L'analyse des données" – Tome 2 "L'analyse des correspondances" – Dunod, 1973.

[2] BENZECRI J.P. – "Sur l'analyse des tableaux binaires associés à une correspondance multiple" (Bin. Mult.) Publication du Laboratoire de de Statistique (1972).

[3] BURT G. – "The factorial analysis of qualitative data" *British Journal of Psychology,* Stat. Sec. III, 1950.

[4] CAILLIEZ F., PAGES J.P. – "Introduction à l'analyse des données" SMASH-BURO, 1976.

[5] CARROLL J.D. – "Categorical conjoint measurement" Ann Arbor, Michigan ; Meeting of Mathematical Psychology, Août 1969.

[6] CAZES P. – "Etude de quelques propriétés extrêmales des facteurs issus d'un sous-tableau d'un tableau de Burt" (extr. Fac.), Publication du Laboratoire de Statistique (1975).

[7] LEBART L. – "L'orientation du dépouillement de certaines enquêtes par l'analyse des correspondances multiples" *Consommation* n° 2 (1975).

[8] MORRISON – "Multivariate Statistical Methods" Mac Graw Hill company (1967)

[9] PAGES J.P., ESCOUFIER Y., CAZES P. – "Opérateurs et analyse des tableaux à plus de deux dimensions". Cahiers du B.U.R.O., n° 25 (1976).

[10] SAPORTA G. – "Liaison entre plusieurs ensembles de variables et codage des données qualitatives" Thèse de 3e cycle, Université de Paris VI (1975).