Liaisons entre plusieurs ensembles de variables et codage de données qualitatives

Gilbert Saporta

▶ **To cite this version:**

Gilbert Saporta. Liaisons entre plusieurs ensembles de variables et codage de données qualitatives. Statistiques [math.ST]. Université Pierre et Marie Curie - Paris VI, 1975. Français. NNT : . tel-00649127

HAL Id: tel-00649127
https://theses.hal.science/tel-00649127

Submitted on 7 Dec 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

THESE

présentée à

L' UNIVERSITE PIERRE ET MARIE CURIE

(PARIS VI)

pour l'obtention du

DOCTORAT TROISIEME CYCLE

Spécialité: MATHEMATIQUES APPLIQUEES

Mention: STATISTIQUE MATHEMATIQUE

par M. Gilbert SAPORTA

Sujet de la thèse: Liaisons entre plusieurs ensembles de variables et codage de données qualitatives

Soutenue le 15 mai 1975 devant la Commission

composée de: Président M. D. DUGUE

Examinateurs: MM. J.GEFFROY

P. KREE

Invité : M. J.P. PAGES

Je remercie Monsieur le Professeur D. DUGUE de m'avoir fait l'honneur de présider le jury de cette thèse.

Que Monsieur le Professeur P. KREE trouve ici l'expression de ma reconnaissance pour les conseils et les critiques utiles qu'il m'a prodigués.

Je remercie Monsieur le Professeur J. GEFFROY de l'intérêt qu'il a bien voulu porter à mon travail en acceptant d'être membre du jury.

Je remercie Monsieur J.P. PAGES qui m'a encouragé tout au long de ce travail et m'a permis d'utiliser la bibliothèque de programmes mise au point par son équipe au C.E.A.

Je remercie l'I.U.T. de Paris qui a pris en charge la confection de cette thèse. Madame LE LOUET a assuré la dactylographie et Monsieur DUTOURNIER le tirage, je les remercie du soin qu'ils y ont apporté.

TABLE DES MATIERES

Chapitre IV : L'ETUDE DIRECTE DES PROXIMITES ENTRE ENSEMBLES D'OBSERVATIONS ET ENTRE VARIABLES QUALITATIVES AU MOYEN D'OPERATEURS LINEAIRES.

RESUME SCIENTIFIQUE

Notre recherche portait initialement sur les méthodes de codage des variables qualitatives afin de les intégrer aux analyses statistiques usuelles. En étudiant le cas de deux variables [22], nous avions donné une interprétation géométrique de l'analyse canonique de LANCASTER [14] et indiqué que cette dernière permettait de retrouver l'analyse des correspondances comme cas particulier. Ces résultats étaient obtenus parallèlement par M. MASSON [16] [17] qui leur donnait une portée bien plus vaste en montrant que le codage des variables qualitatives débouchait sur une méthode générale d'analyse non linéaire des données et traitait le cas de plus de deux variables. Nous nous orientâmes alors vers l'approfondissement de ce point de vue ainsi que des méthodes d'analyses de plus de deux groupes de variables car l'équivalence entre une variable qualitative et l'ensemble des indicatrices de ses modalités conduit nécessairement à l'analyse des liaisons entre plusieurs ensembles de variables.

Dans le premier chapitre nous rappelons les propriétés de l'analyse canonique en proposant une rédaction différente de celle de M. MASSON, ainsi que les principes du codage et de l'analyse non linéaire des données. Nous donnons, au passage, des formules de résolution pour le cas d'un couple constitué d'une variable discrète et d'une variable continue.

Après un rappel sommaire des difficultés posées par l'analyse de trois variables aléatoires et plus, nous proposons une méthode originale de généralisation de l'analyse canonique : X, Y, Z, étant trois v. a. r. et L^2_x, L^2_y, L^2_z, L^2_{xy}, L^2_{yz} ... les espaces de Hilbert qu'elles engendrent, nous définissons la première variable canonique associée à X comme l'élément de L^2_x le plus proche de L^2_{yz} etc..., ce qui revient à maximiser le rapport de corrélation multiple $\eta^2_{\varphi(x)/yz}$.

Nous appliquons cette méthode au codage des marges d'un tableau de contingence cubique.

Notre travail se place ensuite dans le cadre non-probabiliste de l'analyse des données.

Les liaisons entre plus de deux ensembles de données numériques font l'objet du deuxième chapitre où nous développons une idée (inexploitée par son auteur) de J.D. CARROLL [5] pour généraliser l'analyse canonique, qui avait été reprise par R.J. KETTENRING [13] et M. MASSON [17] mais parmi d'autres points de vue également préférables selon eux.

Nous montrons que cette méthode qui consiste à définir des variables auxiliaires ou "facteurs" ayant une somme des carrés des corrélations multiples avec chaque ensemble, maximale, et à les projeter ensuite orthogonalement sur les espaces engendrés par ces ensembles pour obtenir des variables canoniques, n'est autre qu'une analyse en composantes principales où chaque nuage d'individus est muni de sa métrique de MAHALANOBIS. Cette méthode permet de retrouver comme cas particuliers, toutes les analyses multidimensionnelles (canonique, discriminante ...). Cette propriété assure, à notre avis personnel, sa supériorité sur les autres tentatives de généralisation de l'analyse canonique dont la plus connue est celle de HORST [10] que nous lui comparons théoriquement.

Nous établissons que la somme des carrés des valeurs propres de l'analyse proposée est égale à la somme des carrés des coefficients de corrélation canoniques de toutes les p^2 analyses canoniques possibles des p ensembles pris deux à deux. Nous montrons aussi que les valeurs propres ont une interprétation en terme d'analyse de variance.

Nous traitons en détail le cas de deux ensembles de données où l'application de la méthode générale permet de proposer une représentation graphique des proximités entre variables des deux ensembles, qui peut être véritablement qualifiée de simultanée, sur le plan bissecteur des deux plans associés aux deux premières variables canoniques de chaque ensemble, levant ainsi l'ambiguïté classique du choix entre ces deux plans.

Le troisième chapitre traite de l'application de cette méthode aux données qualitatives et des codages descriptifs et explicatifs. On rappelle le résultat mis en évidence par M. MASSON que l'analyse de plusieurs variables qualitatives par cette méthode est équivalente à une analyse des correspondances formelle sur le tableau logique complet individus - modalités, ce qui justifie une pratique courante mal fondée. L'application de la relation concernant les

valeurs propres, prouve que la somme des carrés des valeurs propres de l'analyse des correspondances formelle est égale à la somme de tous les ϕ^2 de contingences possibles.

Pour résoudre le problème de l'explication d'un critère quantitatif par p variables qualitatives (régréssion qualitative) nous proposons une méthode qui consiste à effectuer une régréssion sur les facteurs de l'analyse descriptive des p variables ce qui est une généralisation de la régression sur composantes principales. Les codages explicatifs obtenus sont alors une combinaison linéaire des codages descriptifs. Un paragraphe est consacré à la méthode numérique de S. WOLD [26] qui aboutit au même résultat mais est avantageuse lorsque la taille des tableaux interdit d'utiliser la méthode que nous proposons.

Dans le cas où la variable à expliquer est-elle même qualitative (analyse discriminante) la même technique est proposée qui amène à des simplifications notables pour une discrimination entre deux groupes. Revenant sur des résultats établis par M. MASSON nous étudions une méthode de codage qui minimise la somme des pouvoirs discriminants de chaque variable et qui est équivalente à l'analyse directe si les variables explicatives sont deux à deux indépendantes, ainsi que des améliorations de la méthode séquentielle de M. MASSON.

Un exemple numérique est alors traité.

Ayant montré au deuxième chapitre que la méthode générale ne permet pas de représenter de façon satisfaisante les proximités entre groupes de variables directement nous reprenons cette question au dernier chapitre en utilisant les opérateurs introduits par Y. ESCOUFIER [8] et qui permettent de représenter un ensemble de variables par une seule entité. Suivant en cela J.P. PAGES [19] nous montrons que la distance euclidienne que l'on peut définir entre opérateurs permet de définir des distances et des indices de proximités entre ensemble de variables que l'on peut alors représenter au moyen d'une analyse factorielle sur tableaux de distances. Le choix de la métrique utilisée sur l'espace des individus découle simplement de la nature des équivalences recherchées entre tableaux de données.

Nous montrons que l'analyse du nuage des projecteurs D_p-orthogonaux associés à p tableaux de données est reliée par une formule simple à la méthode

valeurs propres, prouve que la somme des carrés des valeurs propres de l'analyse des correspondances formelle est égale à la somme de tous les ϕ^2 de contingences possibles.

Pour résoudre le problème de l'explication d'un critère quantitatif par p variables qualitatives (régression qualitative) nous proposons une méthode qui consiste à effectuer une régression sur les facteurs de l'analyse descriptive des p variables ce qui est une généralisation de la régression sur composantes principales. Les codages explicatifs obtenus sont alors une combinaison linéaire des codages descriptifs. Un paragraphe est consacré à la méthode numérique de S. WOLD [26] qui aboutit au même résultat mais est avantageuse lorsque la taille des tableaux interdit d'utiliser la méthode que nous proposons.

Dans le cas où la variable à expliquer est-elle même qualitative (analyse discriminante) la même technique est proposée qui amène à des simplifications notables pour une discrimination entre deux groupes. Revenant sur des résultats établis par M. MASSON nous étudions une méthode de codage qui minimise la somme des pouvoirs discriminants de chaque variable et qui est équivalente à l'analyse directe si les variables explicatives sont deux à deux indépendantes, ainsi que des améliorations de la méthode séquentielle de M. MASSON.

Un exemple numérique est alors traité.

Ayant montré au deuxième chapitre que la méthode générale ne permet pas de représenter de façon satisfaisante les proximités entre groupes de variables directement nous reprenons cette question au dernier chapitre en utilisant les opérateurs introduits par Y. ESCOUFIER [8] et qui permettent de représenter un ensemble de variables par une seule entité. Suivant en cela J.P. PAGES [19] nous montrons que la distance euclidienne que l'on peut définir entre opérateurs permet de définir des distances et des indices de proximités entre ensemble de variables que l'on peut alors représenter au moyen d'une analyse factorielle sur tableaux de distances. Le choix de la métrique utilisée sur l'espace des individus découle simplement de la nature des équivalences recherchées entre tableaux de données.

Nous montrons que l'analyse du nuage des projecteurs D_p-orthogonaux associés à p tableaux de données est reliée par une formule simple à la méthode

du chapitre II : $\sum_{i=1}^{p} \mu_i = 2 \left[p\,m - \sum_{i=1}^{m-p} \lambda_i^2 \right]$

m nombre total de variables. μ_i valeur propre de l'analyse du nuage des projecteurs. p nombre de groupes. λ_i valeur propre de l'analyse du chapitre II.

Les indices de proximité entre tableaux de données ne sont autres que les cosinus d'angle entre les opérateurs associés munis d'une structure euclidienne.

L'application inédite, à des variables qualitatives montre alors que le cosinus d'angle entre projecteurs centrés n'est autre que le coefficient T^2 de TSCHUPROW et en faisant un choix différent d'opérateurs on aboutit à une nouvelle mesure de dépendance entre variables qualitatives.

CHAPITRE I

L'ANALYSE DES LIAISONS ENTRE VARIABLES ALEATOIRES

Nous considérerons tout d'abord les relations entre deux variables aléatoires dans l'optique de l'analyse canonique (voir [16], [7] et [14]). Notre rédaction diffère de celle de MASSON de façon à présenter de façon unitaire les divers cas possibles. On utilisera pour cela une hypothèse concernant l'absolue continuité de la loi conjointe par rapport à la loi produit.

Le cas d'un couple de variables discrètes sera l'occasion d'exposer la notion de codage et ses applications dans l'esprit de MASSON.

Après un bref exposé des problèmes posés par l'étude des liaisons entre plus de trois variables aléatoires, nous proposerons une méthode originale pour l'obtention de variables canoniques que nous illustrerons sur un triplet de variables qualitatives.

Notations :

(Ω, a, P) étant un espace probabilisé, L^2 désignera l'espace de Hilbert des v. a. r. du second ordre définies sur (Ω, a, P), muni du produit scalaire habituel $< X, Y > = E(XY)$ et de la norme $|| X||^2 = E(X^2)$.

On identifiera les v.a. presque partout égales. On notera L^2_o le sous-espace de Hilbert des variables orthogonales à la droite des constantes, c'est-à-dire le sous-ensemble de L^2 des variables d'espérance nulle. Dans L^2_o le produit scalaire et la norme sont alors respectivement la covariance et l'écart-

type. L'angle de deux éléments de L_o^2 a pour cosinus le coefficient de corrélation linéaire des deux variables correspondantes.

X étant un élément de L_o^2 de loi μ_x, on notera L_x^2 le sous-espace de Hilbert de L_o^2 engendré par X, c'est-à-dire l'ensemble des variables aléatoires centrées U définies par $U = \varphi(X)$ où φ est de carré sommable par rapport à μ_x. Sauf mention contraire toutes les variables considérées seront des éléments de L_o^2.

I. ANALYSE CANONIQUE D'UN COUPLE DE VARIABLES ALEATOIRES.

1. Résultats généraux.

Nous nous contenterons de rappeler ici les principales propriétés de l'analyse canonique, renvoyant pour des démonstrations à [14] et [22].

L'analyse canonique d'un couple (X, Y) de variables appartenant à L_o^2 consiste à chercher dans un premier temps les éléments normés de L_x^2 et L_y^2 notés $\varphi_1(X)$ et $\psi_1(Y)$ respectivement présentant entre eux un angle minimal, c'est-à-dire ayant un coefficient de corrélation linéaire maximal ρ_1. $\varphi_1(X)$ et $\psi_1(Y)$ constituent alors le premier couple de variables canoniques ; on cherche ensuite le couple suivant $\varphi_2(X)$, $\psi_2(Y)$ avec $\varphi_2(X)$ dans L_x^2 orthogonal à $\varphi_1(X)$ et $\psi_2(Y)$ dans L_y^2 orthogonal à $\psi_1(Y)$, ayant un coefficient de corrélation linéaire maximal ρ_2 et ainsi de suite.

Afin de traiter ce problème par la théorie spectrale nous supposerons que les variables X et Y possèdent la propriété suivante :

Propriété du ϕ^2 - borné.

μ, μ_x, μ_y désignant les lois conjointes et marginales de X, Y la fonction f définie par :

$$d\mu(x, y) = f(x, y)\ d\mu_x(x)\ d\mu_y(y)$$

est de carré sommable par rapport aux produits des mesures marginales $\mu_x\ \mu_y$ et on note $(1 + \phi^2)$ cette intégrale.

On reconnaît dans $f(x, y)$ la généralisation à des variables quelconques du $\frac{p_{ij}}{p_{i.}\ p_{.j}}$ de variables discrètes et dans l'intégrale de $\left[f(x, y)\right]^2$ la généralisation de $\sum_i \sum_j \left(\frac{p_{.j}}{p_{i.}\ p_{.j}}\right)^2 p_{i.}\ p_{.j}$ qui vaut $\sum_i \sum_j \frac{(p_{ij})^2}{p_{i.}\ p_{.j}} = 1 + \phi^2$ où ϕ^2 désigne la mesure d'association de K. PEARSON, d'où le nom de la proprieté.

Il est possible de faire une théorie des variables canoniques sans exiger cette hypothèse, comme l'a démontré LANCASTER [15], mais nous n'aborderons pas ce point de vue.

Selon un résultat classique les variables canoniques $\varphi_1(X)$ et $\psi_1(Y)$ sont telles que :

$$\begin{cases} \Pi_x(\psi_1(Y)) = \rho_1\ \varphi_1(X) \\ \Pi_y(\varphi_1(X)) = \rho_1\ \psi_1(Y) \end{cases}$$

où Π_x et Π_y désignent les projecteurs orthogonaux sur L^2_x et L^2_y respectivement.

On a alors :

$$\boxed{\begin{aligned} \Pi_x \circ \Pi_y\ (\varphi_1(X)) &= \rho_1^2\ \varphi_1(X) \\ \Pi_y \circ \Pi_x\ (\psi_1(Y)) &= \rho_1^2\ \psi_1(Y) \end{aligned}}$$

Or les opérateurs de projection orthogonale ne sont autres que les opérateurs d'espérance conditionnelle :

$$\Pi_x(Z) = E[Z \mid X]$$

Comme ici seules importent les restrictions de Π_x et Π_y à L^2_y et L^2_x respectivement, ces opérateurs sont les opérateurs intégraux de noyau $f(x, y)$ définis par :

$$\Pi_x(\psi_1(Y)) = \int_{\mathbb{R}} \psi_1(Y)\ f(x, y)\ d\mu_y(y)$$

$$\Pi_y(\varphi_1(x)) = \int_{\mathbb{R}} \varphi_1(x)\ f(x, y)\ d\mu_x(x)$$

où $f(x, y)$ est la fonction définie dans la propriété du ϕ^2-borné.

Les restrictions de Π_x et Π_y à L_y^2 et L_x^2 respectivement sont alors des opérateurs de HILBERT-SCHMIDT si X et Y sont de ϕ^2-bornés, voir [17], leur produit est alors un opérateur nucléaire ; la recherche des variables canoniques et des corrélations canoniques n'est donc que l'analyse spectrale de ces opérateurs.

Les variables canoniques $(\varphi_1(X), \varphi_2(X), \dots)$ $(\psi_1(Y), \psi_2(Y), \dots)$ forment alors des systèmes orthonormés totaux de L_x^2 et L_y^2 .

On voit que $\Pi_x \circ \Pi_y$ et $\Pi_y \circ \Pi_x$ exprimés sous forme d'intégrales admettent la valeur propre parasite $\rho_o = 1$ associée aux fonctions constantes $\varphi_o = \psi_o = 1$ qui sont à éliminer puisque nous recherchons des solutions dans L_o^2. Cette remarque montre au passage que les solutions non triviales de

$$\Pi_x \circ \Pi_y(\varphi(X)) = \rho^2\ \varphi(X) \quad \text{et} \quad \Pi_y \circ \Pi_x\ (\psi(Y)) = \rho^2\ \psi(Y)$$

sont automatiquement centrées.

On démontre que $\boxed{\phi^2 = \sum_{i=1}^{\infty} \rho_i^2}$

et donc, en tenant compte de la valeur propre ρ_o que :

Trace $(\Pi_x \circ \Pi_y) = 1 + \phi^2$. La propriété du ϕ^2- borné est donc équivalente à la nucléarité de $\Pi_x \circ \Pi_y$.

La fonction $f(x, y)$ se décompose alors selon :

$$f(x, y) = \sum_{i=o}^{\infty} \rho_i\ \varphi_i(x)\ \psi_i(y)$$

ce qui aboutit à la formule de reconstitution de la loi conjointe :

$$\boxed{d\mu(x, y) = \left[\sum_{i=o}^{\infty} \rho_i\ \varphi_i(x)\ \psi_i(y)\right] d\mu_x(x)\ d\mu_y(y)}$$

On en déduit que $\rho_1 = 0$, qui entraîne $\rho_i = 0 \quad \forall i > 1$, est équivalent à l'indépendance de X et Y puisqu'alors $d\mu(x, y) = d\mu_x(x)\ d\mu_y(y)$.

2. Cas particulier de deux variables absolument continues.

On peut alors écrire :

$$\begin{cases} d\mu(x, y) = h(x, y)\, dx\, dy \\ d\mu_x(x) = f(x)\, dx \\ d\mu_y(y) = g(y)\, dy \end{cases}$$

où h, f, g sont la densité du couple et lesdensités marginales.

L'expression des opérateurs de projection donne alors les formules :

$$\rho^2\, \varphi(x) = \iint_{\mathbb{R}^2} \varphi(u)\, \frac{h(u,y)\ h(x,y)}{f(x)\ g(y)}\, du\, dy$$

$$\rho^2\, \psi(y) = \iint_{\mathbb{R}^2} \psi(v)\, \frac{h(x,v)\ h(x,y)}{f(x)\ g(y)}\, dx\, dv$$

et on a :

$$\sum_{i=1}^{\infty} \rho_i^2 = \phi^2 = \iint_{\mathbb{R}^2} \frac{(h(x,y))^2}{f(x)\ g(y)}\, dx\, dy - 1$$

car la fonction $f(x,y)$ vaut $\dfrac{h(x,y)}{f(x)\ g(y)}$

De plus $h(x, y) = \left[\sum_{i=o}^{\infty} \rho_i\, \varphi_i(x)\ \psi_i(y)\right] f(x)\, g(y)$

L'application de la théorie des variables canoniques au cas d'un couple binormal de coefficient de corrélation ρ, centré réduit conduit aux résultats bien connus suivants :

Les variables canoniques sont les polynomes d'Hermite - Tchebyshev normalisés :

$$\varphi_i(x) = H_i(x) \qquad \psi_i(y) = H_i(y)$$

Les coefficients de corrélation canoniques sont en progression géométrique :

$$\rho_i = \rho^i$$

l'application de la relation $\phi^2 = \sum_{i=1}^{\infty} \rho_i^2$ donne alors ici la formule de K. PEARSON $\rho^2 = \frac{\phi^2}{1+\phi^2}$ utilisée quelquefois pour estimer ρ^2 dans une loi normale lorsque l'échantillon de (X, Y) est mis sous forme de table de contingence.

Le fait que $H_1(x) = x$ constitue le théorème de LANCASTER [14] établissant qu'on ne peut trouver de transformations séparées des variables marginales d'une distribution binormale, augmentant leur coefficient de corrélation.

3. Cas particulier de deux variables discrètes ou qualitatives.

$$\mu(x_i\ y_j) = p_{ij} \qquad \mu_x(x_i) = p_{i.} \qquad \mu_y(y_j) = p_{.j}$$

Toute variable qualitative à p modalités pouvant être considérée comme équivalente à une variable à valeurs dans l'ensemble des entiers de 1 à p, l'analyse canonique de deux variables qualitatives est un cas particulier de l'analyse canonique de deux variables aléatoires réelles.

Nous supposerons que X et Y prennent des nombres de valeurs finies p et q respectivement ; le cas infini s'en déduisant aisément.

Les espaces L_x^2 et L_y^2 sont alors de dimension finie et les opérateurs linéaires peuvent être représentés par des matrices.

Se définir une fonction $\psi(X)$ équivaut à se donner une suite finie de p nombres $(a_1, a_2, \ldots, a_p)$ muni des probabilités $p_{1.}, p_{2.}, \ldots, p_{p.}$ où ce qui revient au même un vecteur $\underline{a}$ de $\mathbb{R}^p$ si on identifie L_x^2 à $\mathbb{R}^p$.

Posons : $\underline{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{pmatrix}$ et $\underline{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_q \end{pmatrix}$

Dans ces conditions l'analyse canonique est la recherche d'éléments $\underline{a}$ de $\mathbb{R}^p$ et $\underline{b}$ de $\mathbb{R}^q$ satisfaisant le critère de corrélation maximale.

Puisque Π_x est l'opérateur d'espérance conditionnelle par rapport à X, si on identifie un élément de L_y^2 à un vecteur $\underline{b}$ de $\mathbb{R}^q$, la restriction de Π_x à L_y^2 sera telle que :

$$\Pi_{x=x_i}(\underline{b}) = \sum_{j=1}^{q} \frac{p_{ij}}{p_{i.}} b_j = E\left[\underline{b} \mid X = x_i\right]$$

En identifiant les variables aléatoires fonction de X à des éléments de $\mathbb{R}^p$ on voit que la restriction de Π_x à L_y^2 peut être représentée comme l'opérateur linéaire dont la matrice, a pour terme général (i, j) $\frac{p_{ij}}{p_{i.}}$, en prenant les bases ordinaires de $\mathbb{R}^p$ et $\mathbb{R}^q$.

Si on introduit les notations suivantes :

P matrice (p,q) d'éléments p_{ij}

D_x matrice (p,p) diagonale d'éléments $p_{i.}$

D_y matrice (q,q) diagonale d'éléments $p_{.j}$

la restriction de Π_x à L_x^2 se représente par la matrice $D_x^{-1} P$,
la restriction de Π_y à L_y^2 se représente par la matrice $D_y^{-1} P'$ où P' est la transposée de P.

Remarque :

Les matrices $D_x^{-1} P$, $D_y^{-1} P'$ (pas plus d'ailleurs que les opérateurs intégraux du paragraphe 1) ne sont pas idempotentes ni symétriques ce qui peut sembler paradoxal pour des projecteurs, mais est en fait parfaitement normal car il ne s'agit pas des projecteurs sur L_x^2 et L_y^2 dans L^2, mais de la représentation de leurs restrictions à L_y^2 et L_x^2 respectivement.

Les variables canoniques $\underline{a}$ et $\underline{b}$ s'obtiennent comme vecteurs propres de matrices selon les équations suivantes :

$$\boxed{\begin{aligned} D_x^{-1}\, P\, D_y^{-1}\, P'\, \underline{a} &= \rho^2\, \underline{a} \\ D_y^{-1}\, P' D_x^{-1}\, P\, \underline{b} &= \rho^2\, \underline{b} \end{aligned}}$$

qui développées donnent :

$$\sum_{k=1}^{p} \sum_{j=1}^{q} \frac{p_{kj}\, p_{ij}}{p_{i.}\, p_{.j}}\, a_k = \rho^2\, a_i$$

$$\sum_{i=1}^{p} \sum_{l=1}^{q} \frac{p_{il}\, p_{ij}}{p_{i.}\, p_{.j}}\, b_l = \rho^2\, b_j$$

On vérifie que les vecteurs $\underline{1}_p$ et $\underline{1}_q$, dont toutes les composantes sont égales à 1, sont solutions triviales de ces équations associées à la valeur propre $\rho_o = 1$.

Les matrices $D_x^{-1}\, P\, D_y^{-1}\, P'$ et $D_y^{-1}\, P'\, D_x^{-1}\, P$ sont respectivement D_x et D_y - symétriques, ce qui entraîne que leurs vecteurs propres sont D_x et D_y orthogonaux, ce qui est la traduction de l'orthogonalité des variables canoniques dans L^2 (non corrélation).

Ces deux matrices ont mêmes valeurs propres non nulles et si nous supposons $p < q$, il existe alors $(p-1)$ couples de variables canoniques non triviales.

Les formules générales s'écrivent donc :

$$\phi^2 = \sum_{i=1}^{p-1} \rho_i^2 = \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{(p_{ij} - p_{i.}\, p_{.j})^2}{p_{i.}\, p_{.j}}$$

et :

$$p_{ij} = \left[\sum_{k=o}^{p-1} \rho_i\, a_i^{(k)}\, b_j^{(k)} \right] p_{i.}\, p_{.j}$$

formule découverte par FISHER où $\underline{a}^{(k)}$ et $\underline{b}^{(k)}$ désignent les vecteurs propres ρ_k^2 des matrices précédentes.

Pour résumer :

L'analyse canonique de deux variables discrètes revient à l'obtention des valeurs propres et vecteurs propres du produit des deux matrices de probabilités conditionnelles.

Lien avec l'analyse des correspondances de J.P. BENZECRI :

Si on pose $\underline{a} = D_x^{-\frac{1}{2}} \underline{\alpha}$ et $\underline{b} = D_y^{-\frac{1}{2}} \underline{\beta}$ il vient :

$D_x^{-1} P D_y^{-1} P' D_x^{-\frac{1}{2}} \underline{\alpha} = \rho^2 D_x^{-\frac{1}{2}} \underline{\alpha}$. $\underline{\alpha}$ est donc vecteur propre de la matrice $D_x^{-\frac{1}{2}} P D_y^{-1} P' D_x^{-\frac{1}{2}}$, c'est-à-dire du produit $R R'$ où R désigne $D_x^{-\frac{1}{2}} P D_y^{-\frac{1}{2}}$ qui est la matrice de terme général $\dfrac{p_{ij}}{\sqrt{p_{i.}\, p_{.j}}}$.

On reconnaît ici la forme classique des équations fournissant les facteurs de la correspondance définie par le tableau des p_{ij} . Les formules de "transition" sont alors la traduction de la propriété de double régression linéaire $E[\varphi_i / \psi_i] = \rho_i \psi_i$. L'inertie totale est le ϕ^2 du tableau de contingence et les inerties portées par les axes factoriels les carrés des coefficients de corrélation canonique. La représentation "simultanée" des deux ensembles de modalités sur un plan consiste à choisir par exemple les deux premières variables canoniques $\varphi_1(X)$, $\varphi_2(X)$ de L_x^2 et à représenter sur le repère cartésien qu'elles engendrent la modalité n° i de X par les deux coordonnées $a_i^{(1)}$ et $a_i^{(2)}$ et la modalité j de Y par les deux coordonnées :

$$\rho_1 b_j^{(1)} \quad \text{et} \quad \rho_2 b_j^{(2)}.$$

Un cas très particulier : $p = 2 \quad q > 2$.

Il n'existe alors qu'un seul couple de variables canoniques et le coefficient de corrélation canonique ρ_1 est tel que :

$$\rho_1^2 = \phi^2$$

De plus l'unique variable canonique associée à X est alors définie sans référence à la distribution conjointe de X et Y mais uniquement en fonction des probabilités marginales $p_{1.}$ et $1 - p_{1.}$.

En effet, si $\underline{a} = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$ est le vecteur cherché, les nombres a_1 et a_2 sont entièrement déterminés par le fait que la variable canonique doit être centrée et réduite.

$$\begin{cases} p_{1.}\ a_1 + (1-p_{1.})\ a_2 = 0 \\ p_{1.}\ a_1^2 + (1-p_{1.})\ a_2^2 = 1 \end{cases}$$

ce qui donne : $a_1 = \sqrt{\dfrac{1-p_{1.}}{p_{1.}}} \qquad a_2 = -\sqrt{\dfrac{p_{1.}}{1-p_{1.}}}$

La variable canonique associée à Y s'obtient alors par application de la formule de régression :

$$E\,(\varphi(X) \mid Y) = \rho\ \psi\ (Y)$$

c'est-à-dire :

$$\sqrt{\phi^2}\ b_j = \frac{p_{1j}}{p_{.j}}\ a_1 + \frac{p_{2j}}{p_{.j}}\ a_2$$

ce qui donne à un facteur de normalisation près :

$$\boxed{b_j = \frac{1}{p_{.j}} \left(\frac{p_{1j}}{p_{1.}} - \frac{p_{2j}}{1-p_{1.}} \right) \quad j = 1, 2, \ldots, q}$$

On trouve dans WILLIAMS [25] des développements sur ce cas particulier, notamment en ce qui concerne la distribution d'échantillonnage des b_j.

4. Un cas particulier non classique : X variable discrète, X variable absolument continue.

$$d\,\mu_x(x) = f(x)\,dx$$

$$\mu_y(y_j) = p_j$$

Ce cas théorique peut constituer un modèle de situations concrètes du type suivant : distribution d'une grandeur continue selon q sous-populations ;

influence d'un facteur aléatoire à q modalités sur une variable ...

Nous supposerons que Y prend un nombre fini de valeurs et que l'on connaît les densités conditionnelles de X à $Y = y_i$ fixé que l'on note $f(x \mid i)$.

On a alors :

$$f(x) = \sum_{i=1}^{q} f(x/i)\ p_i$$

La loi de probabilité conditionnelle de Y à X fixé est donné alors par :

$$P(Y = y_i / X = x) = \frac{p_i\ f(x/i)}{f(x)}$$

Si on écrit le système d'équations :

$$\begin{cases} \Pi_x(\psi_1(Y)) = \rho_1\ \varphi_1(X) \\ \Pi_y(\varphi_1(X)) = \rho_1\ \psi_1(Y) \end{cases}$$

en identifiant la variable canonique $\psi_1(Y)$ à un vecteur $\underline{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_q \end{pmatrix}$ de $\mathbb{R}^q$

il vient, puisque les projecteurs sont les opérateurs d'espérance conditionnelle :

$$E[\underline{b} / X = x] = \Pi_{X=x}\ (\underline{b}) = \sum_{i=1}^{q} \frac{p_i\ f(x / i)}{f(x)}\ b_i = \rho\ \varphi(x)$$

$$E[\varphi(X) / Y = y_j] = \Pi_{y=y_j}\ (\varphi(x)) = \int_{\mathbb{R}} \varphi(x)\ f(x/j)\, dx = \rho\, b_j$$

$\underline{b}$ est donc solution de :

$$\rho^2\ b_j = \int_{\mathbb{R}} \sum_{i=1}^{q} \frac{p_i\ f(x/i)\ f(x/j)}{f(x)}\ b_i\ dx$$

et $\varphi(x)$ de :

$$\rho^2\ \varphi(x) = \sum_{i=1}^{q} \left[\frac{p_i\ f(x/i)}{f(x)} \int_{\mathbb{R}} \varphi(u)\ f(u/j)\, du. \right]$$

Il suffit de résoudre l'équation en $\underline{b}$ pour obtenir la solution de l'équation intégrale en φ par application des formules de régression.

En supposant la permutation des signes sommes légitimes on trouve :

$$\rho^2 b_j = \sum_{i=1}^{q} \left[\int_{\mathbb{R}} \frac{p_i \, f(x/i) \, f(x/j)}{f(x)} \, dx \right] b_i$$

Le vecteur $\underline{b}$ est donc vecteur propre de la matrice A dont le terme général a_{ij} est donné par :

$$a_{ij} = p_i \int_{\mathbb{R}} \frac{f(x/i) \, f(x/j)}{f(x)} \, dx$$

Dans ces conditions on trouve :

$$\varphi(x) = \frac{1}{\rho_1 \, f(x)} \sum_{i=1}^{q} p_i \, f(x/i) \, b_i$$

En particulier si $q = 2$ on peut résoudre complètement les équations puisqu'alors on sait que :

$$b_1 = \sqrt{\frac{1-p}{p}} \qquad b_2 = -\sqrt{\frac{p}{1-p}}$$

On peut interpréter le résultat trouvé en disant que $\varphi(x)$ définit la meilleure manière de mesurer les individus d'une population continue afin d'expliquer une séparation de cette population en deux sous-populations.

5. La notion de codage et l'analyse non-linéaire de M. MASSON [17].

On aura remarqué qu'en ce qui concerne les variables discrètes en particulier, les résultats de l'analyse canonique ne dépendent nullement des valeurs numériques prises par les variables initiales X et Y, mais dépendent uniquement du tableau des p_{ij} ; cette propriété est générale et vient de ce que l'analyse canonique d'un couple de variables est invariante par toute transformation bijective séparée de X et Y.

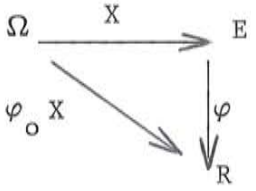
Cette remarque permet le traitement des variables qualitatives sans avoir

besoin de recourir à l'artifice indiqué au paragraphe 3 qui consistait à dire qu'une variable qualitative est équivalente à une variable discrète sur $\{1, 2, \ldots, p\}$. Les variables canoniques sont, elles, à valeurs dans une partie finie de $\mathbb{R}$ puisqu'à chaque couple de modalités (i,j) elles associent le couple réel $(a_i^{(k)}, b_j^{(k)})$ où k est le numéro d'ordre du couple canonique considéré.

Les variables canoniques permettent donc, dans une certaine mesure, de transformer le qualitatif en quantitatif, en opérant un "codage" simultané des variables, adapté à leur liaison.

Définition :

Soit X une v.a. à valeurs dans E, ensemble fini, on dit que l'application φ de E dans $\mathbb{R}$ réalise un codage numérique équivalent à X si φ est injective, un codage compatible avec X si φ est quelconque.



$\varphi \circ X$ est la variable codée.

Pour une variable qualitative X, il existe une infinité de codages possibles ; la recherche d'un codage n'a de sens que si on précise le but poursuivi : la plupart du temps il s'agira de conserver l'"information" apportée par une ou plusieurs variables sur X ; il est certain que le codage d'une variable peut faire perdre une partie de l'information apportée par les autres, on cherchera alors à minimiser cette perte en maximisant un critère de dépendance entre la variable codée $\varphi \circ X$ et les variables associées $Y_1, Y_2, \ldots$ qui sera en général un critère de variance expliquée.

Dans le cas particulier de deux variables qualitatives un codage simultané sera dit "adapté" si les variables codées sont en corrélation maximale.

La définition d'un seul codage risquant d'être insuffisante, on pourra recourir à la notion de codage multidimensionnel selon l'exemple fourni par l'analyse canonique si on considère tous les couples de variables canoniques.

Les $(a_i^{(b)}, b_j^{(k)})$ forment le codage $(p-1)$ dimensionnel adapté des variables X et Y. La formule de reconstitution :

$$p_{ij} = p_{i.}\, p_{.j} \left[\sum_{k=o}^{p-1} \rho_k \, a_i^{(k)} \, b_j^{(k)} \right]$$

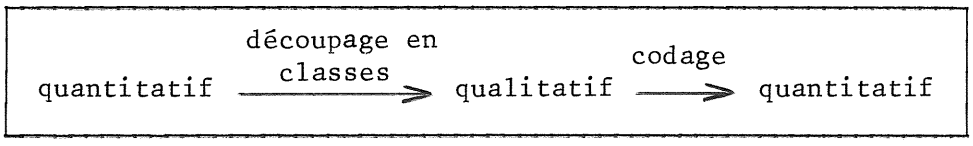
traduit le fait que le codage $(p-1)$-dimensionnel de X et Y reconstitue la totalité de l'information apportée par le tableau de dépendance des p_{ij}.

Sans entrer dans les problèmes mathématiques de convergence, qui n'entrent pas dans le cadre de ce travail, il est clair que l'analyse canonique d'un couple de variables continues peut être approximée par l'analyse d'un tableau de probabilités p_{ij} obtenue par discrétisation de ces variables si on procède à un découpage en classes de valeurs suffisamment fin.

Indiquons ici simplement le résultat cité par LANCASTER [15] p. 99 : les valeurs propres de l'analyse canonique de variables discrétisées sont toujours inférieures ou égales à celles de l'analyse des variables de départ et convergent vers elles si on utilise des partitions de plus en plus fines.

Outre un avantage pratique certain (il est plus facile de diagonaliser une matrice que de faire l'analyse spectrale d'un opérateur intégral), ceci a une conséquence d'une importance capitale, mise en évidence par M. MASSON [17] :

la transformation de variables continues en variables discrètes par un découpage en classes permet d'approcher les relations non linéaires entre ces variables, puisqu'elle aboutit à une approximation de l'analyse canonique complète dont le but est de rechercher les liaisons fonctionnelles approchées entre variables quelconques. On peut alors procèder véritablement à une analyse des données non-linéaires en adoptant le schéma de travail :



Ainsi, loin d'être une simple technique consistant à rendre numérique ce qui ne l'est pas pour faciliter le travail du statisticien, le codage est la clé permettant de sortir du cadre trop strict des analyses linéaires.

6. Analyse non-linéaire et analyse linéaire des variables indicatrices.

Cependant, ainsi que nous allons le voir, l'analyse des données non linéaires se ramène d'une certaine façon à l'analyse linéaire, tant il est vrai que les relations linéaires entre fonctions quelconques de v.a. ne sont autres que les relations fonctionnelles entre ces v.a.

Considérons une variable qualitative à p modalités et un codage $(a_1, a_2, \ldots, a_p)$ des modalités de cette variable obtenu par une application φ de E dans $\mathbb{R}$. Notons $e_1, e_2, \ldots, e_p$ les modalités de X on a donc : $a_i = \varphi(e_i)$.

Introduisons alors les p variables indicatrices des modalités de X : $1_{e_1}, 1_{e_2}, \ldots, 1_{e_p}$. Il est alors immédiat que :

$$\varphi(X) = \sum_{i=1}^{p} a_i \, 1_{e_i}$$

Définir un codage, c'est donc spécifier une combinaison linéaire des indicatrices des modalités.

Ce résultat exprime alors mathématiquement l'équivalence existant entre l'analyse canonique, au sens de LANCASTER, de deux variables discrètes et l'analyse canonique linéaire des deux ensembles d'indicatrices associées aux modalités respectives de X et Y.

Etant donné deux variables qualitatives X et Y, la recherche des variables canoniques $\varphi_1(X)$ et $\psi_1(Y)$ présentant un coefficient de corrélation ρ_1 maximal est alors équivalente à la détermination simultanée des deux combinaisons linéaires des indicatrices de X et de Y, respectivement, présentant une corrélation maximale.

A titre d'exercice, établissons l'identité formelle entre ces deux analyses canoniques :

L'analyse canonique linéaire de deux groupes de variables $(X_1, Y_2, \ldots X_p)$; $(Y_1, Y_2, \ldots, Y_q)$ consiste à rechercher les paires $\xi = \sum_{i=1}^{p} a_i X_i$ $\eta = \sum_{j=1}^{q} b_j Y_j$ telles que $\rho(\xi, \eta)$ soit maximal.

D'après la théorie classique, voir par exemple RAO [20], le vecteur

$\underline{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{pmatrix}$ est alors vecteur propre de la matrice $\Sigma_{11}^{-1} \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}$

où Σ_{11} est la matrice des variances - covariances du premier groupe, Σ_{22} celle du deuxième, $\Sigma_{12} = \Sigma_{21}'$ celle des covariances croisées du premier groupe avec le deuxième.

Si les X_i et les Y_j désignent les variables indicatrices des modalités de X et Y, variables qualitatives à p et q modalités, les deux groupes ne sont pas linéairement indépendants puisque :

$$\sum_{i=1}^{p} X_i = \sum_{j=1}^{q} Y_j = 1$$

Il est alors équivalent de procéder à une analyse <u>non - centrée</u> c'est-à-dire de prendre pour Σ_{ij}, au lieu des covariances, les espérances des produits : en effet, l'analyse canonique linéaire non centrée donnera les mêmes résultats que l'analyse centrée à la solution triviale près $\rho_o = 1 \quad \xi = \eta = 1$, puisque les couples suivants orthogonaux à 1 seront alors centrés.

Les X_i et les Y_j étant des variables indicatrices on a :

$$E(X_i) = p_{i.} \qquad E(Y_j) = p_{.j}$$
$$E(X_i^2) = p_{i.} \qquad E(Y_j^2) = p_{.j}$$
$$E(X_i\, X_j) = E(Y_i\, Y_j) = 0 \quad \text{si} \quad i \neq j$$
$$E(X_i\, Y_j) = p_{ij} .$$

En reprenant les notations du paragraphe 3 on trouve alors

$$\Sigma_{11} = D_x \qquad \Sigma_{22} = D_y \qquad \Sigma_{12} = P \qquad \Sigma_{21} = P'$$

et l'équation $\Sigma_{11}^{-1} \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{12} \underline{a} = \rho^2 \underline{a}$ s'écrit :

$$D_x^{-1} P D_y^{-1} P' \underline{a} = \rho^2 \underline{a} .$$

La recherche des liaisons fonctionnelles entre variables discrètes est donc équivalente à la recherche des liaisons linéaires entre les variables indicatrices des modalités.

En rapprochant ce résultat de l'analyse non - linéaire de MASSON, nous voyons que toute variable continue se trouve transformée en un groupe de p variables indicatrices par discrétisation et que l'analyse des liaisons non - linéaires débouche naturellement sur l'analyse des relations entre k groupes de variables qui fera l'objet du deuxième chapitre.

II. LES LIAISONS ENTRE PLUS DE DEUX VARIABLES ALEATOIRES.

La généralisation des résultats du I se heurte à des difficultés qui tiennent d'une part à l'absence d'indicateurs de liaisons convenables entre plus de deux variables et d'autre part, à l'apparition de la notion de dépendance partielle qui entraîne que la considération des liaisons deux à deux n'épuise pas l'étude des dépendances.

1. Quelques mesures de liaison entre p variables.

La matrice des corrélations linéaires deux à deux regroupe l'ensemble des informations concernant les liaisons marginales entre les p variables et permet de déduire au moyen des mineurs de son déterminant les coefficients de corrélation linéaire partiels.

Déduire d'une matrice une mesure unique est problèmatique : certains auteurs (HORST) ont proposé de prendre comme mesure globale de la corrélation la somme des termes de la matrice ; on peut aussi considérer la somme des carrés des termes ce qui a l'avantage d'être le carré de la norme euclidienne de la matrice de corrélation ...

En remarquant que la matrice est singulière si une des variables est combinaison linéaire des autres, on peut utiliser le déterminant de la matrice comme mesure de dépendance et en cas de nullité, examiner les divers mineurs pour détecter d'autres liaisons linéaires. Ce point de vue est à l'origine de la notion de variance généralisée ou volume engendré par les variables.

La dépendance d'une variable par rapport aux $p-1$ autres peut s'exprimer au moyen du coefficient de corrélation linéaire multiple mais tous les indices précédents ont le défaut de ne tenir compte que des liaisons linéaires.

Afin de mesurer des dépendances non linéaires il est préférable d'utiliser les rapports de corrélation multiple :

$$\eta^2_{X_1/X_2\ X_3 \ldots\ X_p} = \frac{V(X_1/X_2\ X_3 \ldots\ X_p)}{V(X_1)}$$

qui mesurent l'angle formé par une variable et la variété engendrée par les autres dans L^2 et généralisent commodément les rapports de corrélation simple.

Le ϕ^2 de K. PEARSON se généralise aussi, aisément à plus de 2 variables et se calcule par exemple pour 3 variables absolument continues par la formule :

$$1 + \phi^2_{xyz} = \int_{\mathbb{R}^3} \frac{(h(x,y,z))^2}{f(x)\ g(y)\ k(z)}\ dx\ dy\ dz$$

Les distributions à 3 dimensions peuvent alors être classées en huit types selon que chacun des ϕ^2_{xy} ϕ^2_{xz} ϕ^2_{yz} est nul ou non.

La relation $\phi^2_{xyz} = 0$ entraîne comme dans le cas bidimensionnel l'indépendance mutuelle des 3 variables.

Rechercher des variables canoniques suppose alors que l'on a fait choix d'une mesure de liaison entre variables aléatoires que ces variables rendent maximale. Or, selon le choix fait, somme des corrélations, somme des carrés des corrélations etc... on obtient des résultats différents, voir [17].

LANCASTER [15] p. 100 a tenté une solution ne faisant pas appel implicitement à une mesure de liaison unique et aboutissant aux résultats suivants pour 3 variables X, Y, 2 absolument continues.

Il est possible de définir 3 ensembles orthonormés de fonctions sur L^2_x, L^2_y et L^2_z notées φ_i, ψ_j, ξ_k vérifiant la formule de reconstitution

$$h(x, y, z) = f(x)\, g(y)\, h(z) \left[1 + \sum_{j,k} \rho_{ojk}\, \psi_j\, \xi_k + \sum_{i,j} \rho_{ijo}\, \varphi_i\, \psi_j \right.$$

$$\left. + \sum_{i,k} \rho_{iok}\, \varphi_i\, \xi_k + \sum_{i,j,k} \rho_{ijk}\, \varphi_i\, \psi_j\, \xi_k \right]$$

où les ρ_{ijk} sont égaux à :

$$\rho_{ijk} = \int_{\mathbb{R}^3} \varphi_i\, \psi_j\, \xi_k\, h(x, y, z)\, dx\, dy\, dz$$

en prenant par convention $\varphi_o = \psi_o = \xi_o = 1$.

On a alors : $\phi^2_{xyz} = \sum_i \sum_j \sum_k (\rho_{ijk})^2$

Cependant les ρ_{ijk} qui apparaissent comme des coefficients de corrélation généralisés, peuvent être supérieurs à 1 ce qui en limite l'interprétation. De plus la résolution effective même dans le cas fini est quasiment impossible et les formules deviennent inextricables lorsque $p > 3$.

2. Une proposition de méthode destinée à construire des variables canoniques.

Dans l'analyse canonique de deux variables aléatoires il est équivalent de rechercher un couple $\psi_1(Y)$, $\varphi_1(X)$ ayant un coefficient de corrélation maximale ρ_1 et de chercher séparément deux variables $\psi_1(Y)$ et $\varphi_1(X)$ faisant un angle minimal avec L^2_x et L^2_y respectivement : en d'autres termes $\varphi_1(X)$ et $\psi_1(Y)$ sont tels qu'ils réalisent séparément des maxima de :

$$\| \Pi_x(\psi_1(Y)) \|^2 \quad \text{et} \quad \| \Pi_y(\varphi_1(Y)) \|^2$$

φ_1 et ψ_1 étant de norme unité.

Puisque toutes les variables sont centrées ceci équivaut à l'obtention du maxima de :

$$\eta^2_{\psi(Y)/X} \qquad \text{ou de} \qquad \eta^2_{\varphi(X)/Y}$$

η^2 désignant le rapport de corrélation, les deux maxima étant d'ailleurs égaux à ρ_1^2 .

Par analogie avec ce qui précède, nous proposons de rechercher dans L^2_x ,

L_y^2, L_z^2, ... séparément les éléments les plus proches des sous-espaces de Hilbert engendré par les autres variables.

Si il n'y a que 3 variables, $\varphi(X)$ sera l'élément de L_x^2 le plus proche de L_{yz}^2 où L_{yz}^2 désigne l'ensemble des variables centrées de la forme $U(Y, Z)$ où U est de carré sommable par rapport à la mesure conjointe de Y et Z.

Ceci revient à chercher les maxima de :

$$\eta^2_{\varphi(X)/YZ} \qquad \eta^2_{\psi(Y)/XZ} \qquad \eta^2_{\xi(Z)/XY}$$

Les variables canoniques que nous définissons ainsi ont la propriété d'être les plus liées possible avec les deux autres au sens de la variance expliquée.

Si nous introduisons les projecteurs Π_{xy} Π_{xz} Π_{yz} sur L^2_{xy}, L^2_{yz}, L^2_{yz} les variables ainsi définies vérifient les équations :

$$\begin{cases} \Pi_x \circ \Pi_{yz} \ (\varphi(X)) = \eta^2_{\varphi/YZ} \ \varphi(X) \\ \Pi_y \circ \Pi_{xz} \ (\psi(Y)) = \eta^2_{\psi/XZ} \ \psi(Y) \\ \Pi_z \circ \Pi_{xy} \ (\xi(Z)) = \eta^2_{\xi/XY} \ \xi(Z) \end{cases}$$

Sous l'hypothèse de ϕ^2 - borné pour (X, Y, Z) la théorie spectrale s'applique et ce problème revient à étudier le spectre de 3 opérateurs nucléaires.

Etudions le cas particulier d'un tableau de contingence cubique de dimension $p \times q \times r$, défini par les probabilités p_{ijk}. Le problème est alors de déterminer trois vecteurs $\underline{a}$, $\underline{b}$, $\underline{c}$ à p, q et r composantes respectivement. Ecrivons les équations fournissant $\underline{a}$, celles donnant $\underline{b}$ et $\underline{c}$ en déduisant par permutation des indices.

Si $\underline{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{pmatrix}$ le projecteur Π_{yz} a pour effet de transformer $\underline{a}$ en

un élement de $\mathbb{R}^{rq}$ que nous pouvons représenter par un tableau à deux indices d'éléments b_{jk} tels que :

$$b_{jk} = \sum_{i=1}^{p} \frac{p_{ijk}}{p_{.jk}} \ a_i$$

Le projecteur Π_x, étant l'opérateur d'espérance conditionnelle par rapport à X, transforme alors le tableau b_{jk} en un élément $\underline{\alpha}$ de $\mathbb{R}^p$ tel que :

$$\alpha_1 = \sum_{j=1}^{q} \sum_{k=1}^{r} \frac{p_{1jk}}{p_{1..}} b_{jk} .$$

L'équation $\Pi_x \circ \Pi_{yz} (\varphi(X)) = \eta^2 \varphi(X)$ s'écrit donc :

$$\sum_{j=1}^{q} \sum_{k=1}^{r} \sum_{i=1}^{p} \frac{p_{1jk}\, p_{ijk}}{p_{1..}\, p_{.jk}} a_i = \eta^2 a_1 \qquad 1 = 1, 2, \ldots, p$$

Le vecteur $\underline{a}$ est donc vecteur propre de la matrice M dont les éléments m_{1i} valent :

$$m_{1i} = \sum_{j=1}^{q} \sum_{k=1}^{r} \frac{p_{ijk}\, p_{1jk}}{p_{.jk}\, p_{1..}} = \sum_{j=1}^{q} \sum_{k=1}^{r} p_{i/jk}\, p_{jk/1} .$$

Cette méthode peut s'étendre à plus de trois variables, cependant nous ne la développerons pas plus avant en raison des difficultés d'application qui surgissent quand les variables possèdent de nombreuses modalités : le nombre de cases du tableau multiple qui s'en déduit devient alors souvent trop élevé vis à vis du nombre total d'observations dans le cas d'un échantillon.

CHAPITRE II

L'ANALYSE FACTORIELLE DE PLUSIEURS ENSEMBLES D'OBSERVATIONS

Nous adoptons désormais le point de vue expérimental de l'analyse des données.

I. GENERALITES.

1. Les données.

Elles consistent en p tableaux rectangulaires X_i de dimension $n \times m_i$ $(i = 1, 2, \ldots, p)$ représentant chacun les observations de m_i variables numériques sur n individus de masses positives p_i avec $\sum_{i=1}^{n} p_i = 1$.

Ces tableaux seront considérés par la suite comme les matrices d'applications linéaires de $\mathbb{R}^{m_i}$ dans $\mathbb{R}^n$ chaque espace étant muni de sa base usuelle. On définira aussi une matrice X de dimension $n \times \left(\sum_{i=1}^{p} m_i \right)$ regroupant les blocs X_i :

$$X = \begin{array}{c} \\ 1 \\ 2 \\ \vdots \\ n \end{array} \begin{array}{c} \begin{array}{cccccccc} 1 & 2 & .. & m_1 & ..... & m_1+m_2 & ..... & ..... \end{array} \\ \left(\begin{array}{c|c|c|c} & & & \\ X_1 & X_2 & & X_p \\ & & & \end{array} \right) \end{array}$$

$\mathbb{R}^n$ sera dit "espace des variables" et $\mathbb{R}^{\sum_{i=1}^{p} m_i}$ "espace des individus". On identifiera une variable à un vecteur $\underline{x}$ de $\mathbb{R}^n$, $\mathbb{R}^n$ sera muni du produit scalaire défini par la matrice diagonale $D_p = \begin{pmatrix} p_1 & & & \\ & p_2 & & \\ & & \ddots & \\ & & & p_n \end{pmatrix}$ c'est-à-dire :

$$< \underline{x}, \underline{y} > = \sum_{i=1}^{n} p_i x_i y_i = \underline{x}' D_p \underline{y}$$

Dans le sous-espace de dimension $n-1$ des variables centrées (c'est-à-dire D_p - orthogonales ou vecteur $\underline{1}$ dont toutes les composantes sont égales à l'unité) ce produit scalaire se confond avec la covariance empirique. Le cosinus de l'angle formé par deux vecteurs D_p - orthogonaux à $\underline{1}$ est alors égal au coefficient de corrélation linéaire empirique entre les variables correspondantes.

Sauf indication contraire, nous nous placerons dans le sous-espace des variables centrées.

Les matrices $V_{ii} = X_i' D_p X_i$ (i = 1, 2,..., p) sont alors les matrices de variance - covariance du $i^{ème}$ ensemble de variables ; les matrices $V_{ij} = X_i' D_p X_j$ sont les matrices des covariances croisées du $i^{ème}$ ensemble avec le $j^{ème}$.

La matrice de variance - covariance totale $H = X' D_p X$ est donc partitionnée selon les blocs V_{ij} :

$$H = X' D_p X = \begin{pmatrix} V_{11} & V_{12} & \cdots & V_{1p} \\ V_{21} & V_{22} & & \vdots \\ \vdots & & \ddots & \vdots \\ V_{p1} & \cdots & \cdots & V_{pp} \end{pmatrix}$$

Nous noterons W_i le sous-espace vectoriel de $\mathbb{R}^n$ engendré par X_i, c'est-à-dire l'ensemble de toutes les variables que l'on peut obtenir par combinaisons linéaires des m_i colonnes de X_i :

$$W_i = \left\{ X_i \underline{b} \;\middle|\; \underline{b} \in \mathbb{R}^{m_i} \right\}$$

Nous supposerons que les X_i sont de rang m_i donc que $\dim W_i = m_i$.

$$W = \left\{ X\,\underline{b} \;\middle|\; \underline{b} \in R^{\sum_{i=1}^{m} m_i} \right\} \text{ est la somme directe des } W_i$$

$$W = \bigoplus_{i=1}^{p} W_i \quad \text{et} \quad \dim W \leqslant \sum_{i=1}^{p} m_i$$

Si on désigne par P_i l'opérateur de projection D_p-orthogonale sur W_i, sa représentation matricielle sera :

$$P_i = X_i (X_i' D_p X_i)^{-1} X_i' D_p = X_i V_{ii}^{-1} X_i' D_p .$$

2. Position du problème.

Lorsqu'il n'y a que deux groupes de variables, l'analyse des liaisons existant entre ces deux groupes, s'effectue au moyen de l'analyse canonique classique qui détermine les couples $(\underline{\xi}_i, \underline{\eta}_i)$ et leurs coefficients de corrélation canonique r_i , $(i = 1, 2, \ldots \inf\{m_1, m_2\})$ de variables D_p-normées de W_1 et W_2 respectivement qui sont les vecteurs propres et les valeurs propres des opérateurs $P_1 P_2$ et $P_2 P_1$.

Généraliser l'analyse canonique classique à plus de deux ensembles débouche sur les difficultés, exposées au chapitre précédent, de mesurer la dépendance entre plus de deux variables aléatoires : en effet, si on veut déterminer un p-uple de "variables canoniques" D_p-normées $(\underline{\xi}_1, \underline{\xi}_2, \ldots, \underline{\xi}_p)$ où $\underline{\xi}_j \in W_j$ le choix du critère à maximiser n'est pas indifférent et les résultats varient selon que l'on cherche des variables présentant une somme des corrélations simples maximale (critère de HORST [10]) ou une somme des corrélations carrées maximales (critère de KETTENRING [13]) ou encore une variance généralisée minimale.

Tous ces critères qui ont été étudiés en détail par M. MASSON [17] ont pour trait commun de redonner l'analyse canonique usuelle lorsque $p = 2$.

- critère de HORST :

$\underline{\xi}_i$ étant de la forme $X_i \underline{b}_i$, le coefficient de corrélation entre $\underline{\xi}_i$ et $\underline{\xi}_j$ s'écrit :

$$r_{ij} = \frac{\underline{\xi}_i' D_p \underline{\xi}_j}{\|\underline{\xi}_i\| \|\underline{\xi}_j\|} = \frac{\underline{b}_i' X_i' D_p X_j \underline{b}_j}{\|X_i \underline{b}_i\| \|X_j \underline{b}_j\|} = \frac{\underline{b}_i' V_{ij} \underline{b}_j}{\|X_i \underline{b}_i\| \|X_j \underline{b}_j\|}$$

Maximiser $\sum_i \sum_j r_{ij}$ équivaut alors à maximiser

$\sum_i \sum_j \underline{b}_i' V_{ij} \underline{b}_j$ sous les contraintes $\underline{b}_i' V_{ii} \underline{b}_i = 1$

ou, ce qui revient au même, à maximiser sans contraintes :

$$\sum_i \sum_j \underline{b}_i' V_{ij} \underline{b}_j - \sum_{i=1}^{p} \lambda_i \underline{b}_i' V_{ii} \underline{b}_i$$

la solution est donnée par le système d'équations obtenues par annulation des dérivées en $\underline{b}_i$:

$$\sum_{j=1}^{p} V_{ij} \underline{b}_j = \lambda_i V_{ii} \underline{b}_i \quad (i = 1, 2, \ldots, p)$$

Un algorithme itératif de résolution a été programmé par M. MASSON [17].

- critère de KETTENRING :

Les corrélations sont alors élevées au carré et il s'agit de maximiser :

$$\sum_i \sum_j (\underline{b}_i' V_{ij} \underline{b}_j)^2 - \sum_i \lambda_i \underline{b}_i' V_{ii} \underline{b}_i$$

Aucun algorithme de résolution n'est actuellement disponible à notre connaissance pour résoudre ce problème.

- critère de volume minimal ou de la variance généralisée :

Considérons les $\underline{\xi}_i$ comme les colonnes d'une matrice K, alors on définit leur variance généralisée d'après ANDERSON [1] comme étant égale au déterminant de $K' D_p K$ ou ce qui est pareil :

$$V = (n-1)^p \det \left[\frac{1}{n-1} \sum_{i=1}^{p} \underline{\xi}_i D_p \underline{\xi}_i' \right]$$

On trouve dans STEEL [24] une tentative de minimisation de ce déterminant par une méthode différentielle.

- L'idée de J.D. CARROLL :

J.D. CARROLL proposa en 1968 [5] de remplacer le problème de la recherche des p variables canoniques par celui qui consiste à trouver une variable auxiliaire unique la plus liée en un certain sens aux ensembles X_i et de la

décomposer ensuite sur les W_i pour obtenir des variables canoniques. C'est cette idée que nous développons maintenant et que M. MASSON avait considérée pour des variables qualitatives sous le nom de méthode de la "première composante principale maximale", sans toutefois en pousser l'étude.

II. LA METHODE FACTORIELLE.

1. Le principe.

Nous allons rechercher une variable unique $\underline{z}$ qui soit la plus liée en moyenne aux p ensembles de variables au sens ou la somme des carrés des coefficients de corrélation multiple de $\underline{z}$ avec les X_i soit maximale.

Nous appellerons "facteur" cette variable et "variables canoniques associées" les projections D_p-orthogonalesde $\underline{z}$ sur les W_i.

Le carré du coefficient de corrélation multiple R_i de $\underline{z}$ avec X_i n'étant autre que le carré du cosinus de l'angle formé par $\underline{z}$ et W_i, c'est-à-dire par $\underline{z}$ et sa projection D_p-orthogonale sur W_i on a :

$$R_i^2 = \frac{\underline{z}' D_p P_i \underline{z}}{\underline{z}' D_p \underline{z}}$$

Il s'agit donc de trouver une variable $\underline{z}$ définie à une constante multiplicative près qui maximise $\sum_{i=1}^{p} R_i^2$ c'est-à-dire :

$$\max \frac{\underline{z}' D_p \left(\sum_{i=1}^{p} P_i \right) \underline{z}}{\underline{z}' D_p \underline{z}}$$

Ce problème classique de maximisation du quotient de deux formes quadratiques admet alors la solution suivante :

$\underline{z}$ est vecteur propre de $\sum_{i=1}^{p} P_i$ associe à sa plus grande valeur propre λ_1 qui est alors la valeur maximale de $\sum_{i=1}^{p} R_i^2$.

Les variables canoniques $\underline{\xi}_i = P_i\,\underline{z}$ sont telles que $\lambda\,\underline{z} = \sum_{i=1}^{p} \underline{\xi}_i$.

On remarque que si $\underline{z}$ est D_p-normé à 1 , les $\underline{\xi}_i$ ne le sont pas et ont généralement des normes différentes puisque $\left\| \underline{\xi}_i \right\|^2 = R_i^2$ à moins que $R_1^2 = R_2^2 = \ldots = R_p^2$ ce qui n'a aucune raison d'être vrai.

Ayant obtenu un premier facteur $\underline{z}$ on peut poursuivre la diagonalisation de $\sum_{i=1}^{p} P_i$ pour obtenir des facteurs D_p-orthogonaux au premier, et entre eux, puisque $\sum_{i=1}^{p} P_i$ somme de matrices D_p-symétriques est elle-même D_p-symétrique.

On vérifie alors que les facteurs $\underline{z}$ sont indépendants de toute transformation linéaire régulière de chaque tableau de données car les projecteurs P_i ne dépendent que des espaces W_i et non d'une base particulière.

Tels sont les résultats obtenus par J.D. CARROLL et qui conduisent à la diagonalisation d'une matrice carrée $n \times n$ ce qui aboutit à des impossibilités pratiques dès que le nombre des observations dépasse la centaine ce qui est quand même courant.

Nous avons alors poursuivi l'étude de cette méthode pour obtenir une résolution dans l'espace des individus, de dimension $\sum_{i=1}^{p} m_i$, guidés par le fait que lesfacteurs $\underline{z}$, appartenant à $\bigoplus_{i=1}^{p} W_i$, se trouvaient en fait dans un sous-espace de $\mathbb{R}^n$ de dimension au plus égale au nombre total de variables.

2. Résolution dans $\mathbb{R}^{\sum_{i=1}^{p} m_i}$.

Posons $\underline{\xi}_i = P_i\,\underline{z} = X_i\,\underline{b}_i$ avec $\underline{b}_i \in \mathbb{R}^{m_i}$

on a : $\lambda\,\underline{z} = \sum_{j=1}^{p} X_j\,\underline{b}_j$ puisque $\lambda\,\underline{z} = \left(\sum_{i=1}^{p} P_i \right) \underline{z}$

d'où : $P_j \left(\sum_{i=1}^{p} X_i\,\underline{b}_i \right) = \lambda\, X_j\,\underline{b}_j$

soit : $X_j (X_j' D_p X_j)^{-1} X_j' D_p \sum_{i=1}^{p} X_i \underline{b}_i = \lambda X_j \underline{b}_j$

Les X_j étant de rang m_j ceci s'écrit :

$$(X_j' D_p X_j)^{-1} X_j' D_p \sum_{i=1}^{p} X_i \underline{b}_i = \lambda \underline{b}_j$$

Soit alors le vecteur $\underline{b} \in \mathbb{R}^{\Sigma m_i}$ dont les composantes sont les $\underline{b}_j$ considérés comme blocs :

$$\underline{b} = \begin{pmatrix} \underline{b}_1 \\ \hline \underline{b}_2 \\ \hline \\ \underline{b}_p \end{pmatrix}$$

On reconnaît dans la relation précédente une équation aux valeurs propres concernant $\underline{b}$, écrite par blocs :

$\underline{b}$ est vecteur propre associé à la valeur propre λ d'une matrice M dont les blocs M_{ij} sont définis par :

$$M_{ij} = (X_i' D_p X_i)^{-1} X_i' D_p X_j = V_{ii}^{-1} V_{ij}$$

$$M = \begin{bmatrix} I_{m_1} & V_{11}^{-1} V_{12} & \cdots & V_{11}^{-1} V_{1p} \\ V_{22}^{-1} V_{21} & I_{m_2} & & \vdots \\ \vdots & & \ddots & \vdots \\ \vdots & & & \vdots \\ \vdots & & & \vdots \\ \cdots & \cdots & \cdots & I_{m_p} \end{bmatrix}$$

M peut également s'écrire :

$$M = \begin{pmatrix} V_{11}^{-1} X_1' D_p \\ \hline V_{22}^{-1} X_2' D_p \\ \hline \vdots \\ \hline V_{pp}^{-1} X_p' D_p \end{pmatrix} \left(X_1 \mid X_2 \mid \cdots \mid X_p \right)$$

ou encore :

$$M = \begin{pmatrix} V_{11}^{-1} & & & \\ & V_{22}^{-1} & & \\ & & \ddots & \\ & & & V_{pp}^{-1} \end{pmatrix} \begin{pmatrix} V_{11} & V_{12} & \cdots & V_{1p} \\ V_{21} & V_{22} & & \vdots \\ \vdots & & \ddots & \vdots \\ \vdots & & & \vdots \\ \vdots & & & \vdots \\ \vdots & & & V_{pp} \end{pmatrix}$$

C'est-à-dire $M = V^{-1} H$ si V désigne la matrice de blocs diagonaux V_{ii}.

On voit que M est une matrice V et H-symétrique, ses vecteurs propres sont alors V-orthogonaux et H-orthogonaux.

Nous avons donc établi le résultat suivant :

> Les variables canoniques $\underline{\xi}_i = X_i \underline{b}_i$ sont telles que les $\underline{b}_i$ sont les blocs du premier vecteur propre $\underline{b}$ de la matrice $V^{-1}H$ qui a mêmes valeurs propres non nulles que $\sum_{i=1}^{p} P_i$.

Comme on a $\lambda \underline{z} = X \underline{b}$ nous avons les relations suivantes entre les normes des vecteurs propres de $\sum_{i=1}^{p} P_i$ et de $V^{-1} H$:

Si $\underline{z}$ est D_p - normé à 1 alors $\underline{b}$ est H - normé à λ et V - normé à $\sqrt{\lambda}$.

Si $\underline{b}$ est H - normé à 1 alors $\underline{z}$ est D_p - normé à $\frac{1}{\lambda}$

Si $\underline{b}$ est V - normé à 1 alors $\underline{z}$ est D_p - normé à $\frac{1}{\sqrt{\lambda}}$

3. Remarques sur les valeurs propres et vecteurs propres de $\sum_{i=1}^{p} P_i$

$\sum_{i=1}^{p} P_i$ est une matrice $(n \times n)$ diagonalisable parce-que D_p - symétrique. Ses valeurs propres sont positives et il vient immédiatement que :

$$\text{Trace}\left(\sum_{i=1}^{p} P_i\right) = \sum_{i=1}^{p} m_i$$

par linéarité de la trace, les P_i étant des projecteurs de rang m_i par hypothèse.

Parmi les n valeurs propres de $\sum_{i=1}^{p} P_i$ il y en a toujours $n - \sum_{i=1}^{p} m_i$ nulles, car $\sum_{i=1}^{p} P_i$ est de rang au plus égal à $\sum_{i=1}^{p} m_i$.

Si $\dim\left(\bigoplus_{i=1}^{p} W_i\right) = s < \sum_{i=1}^{p} m_i$ il y a alors $n - s$ valeurs propres nulles correspondant à des vecteurs propres orthogonaux à tous les W_i.

Les valeurs propres de $\sum_{i=1}^{p} P_i$ sont toutes inférieures à p puisque $\lambda = \sum_{i=1}^{p} R_i^2$.

Si p est valeur propre, cela prouve que les vecteurs propres correspondant à p appartiennent à tous les W_i puisqu'alors $R_i^2 = 1 \; \forall i$. L'ordre de multiplicité de la valeur propre p est donc égal à la dimension de l'intersection des W_i.

Puisque $0 \leqslant \lambda_1 \leqslant p$ on peut utiliser le rapport $\frac{\lambda_1}{p}$ qui est alors la moyenne des carrés des coefficients de corrélation multiple de $\underline{z}$ avec les X_i comme une mesure globale de la dépendance entre les p ensembles de variables. Le cas $\lambda_1 = 1$ indique que les W_i sont orthogonaux deux à deux.

III. LA METHODE FACTORIELLE ET LES METHODES CLASSIQUES.

1. Réduction à l'analyse canonique classique lorsque $p = 2$.

Comme il se doit, les variables canoniques obtenues par projection des facteurs $\underline{z}$ sur W_1 et W_2 ne sont autres que les variables canoniques $\underline{\xi}$ et $\underline{\eta}$ de l'analyse classique.

En effet, en partant de $(P_1 + P_2)\,\underline{z} = \lambda\,\underline{z}$ on trouve aisément :

$$P_1\,\underline{z} + P_1\,P_2\,\underline{z} = \lambda\,P_1\,\underline{z}$$

$$P_2 \underline{z} + P_2 P_1 \underline{z} = \lambda P_2 \underline{z}$$

à cause de l'idempotence des projecteurs et ou en posant $P_1 \underline{z} = \underline{\xi}$ $P_2 \underline{z} = \underline{\eta}$ il vient alors :

$$P_1 \underline{\eta} = (\lambda - 1) \underline{\xi}$$

$$P_2 \underline{\xi} = (\lambda - 1) \underline{\eta}$$

soit :

$$\begin{cases} P_2 P_1 \underline{\eta} = (\lambda - 1)^2 \underline{\eta} \\ P_1 P_2 \underline{\xi} = (\lambda - 1)^2 \underline{\xi} \end{cases}$$

On reconnaît bien ici les équations fondamentales de l'analyse canonique qui fournissent $\inf\{m_1, m_2\}$ couples de variables canoniques. Les $\lambda_i - 1$ étant les coefficients de corrélation canonique si $1 \leqslant i \leqslant \inf\{m_1, m_2\}$.

Cependant, comme on peut extraire $m_1 + m_2$ facteurs $\underline{z}$ non triviaux, c'est-à-dire correspondant à des valeurs propres non identiquement nulles, il est intéressant d'étudier la structure complète des valeurs propres et des vecteurs propres de $P_1 + P_2$.

Proposition.

Si λ est valeur propre, de rang inférieur à $m_1 + m_2$, alors $2 - \lambda$ est aussi valeur propre de $P_1 + P_2$.

Preuve : les valeurs propres de $P_1 + P_2$ de rang $\leqslant m_1 + m_2$ sont les mêmes que celles de M qui est égal à :

$$M = \begin{pmatrix} I_{m_1} & V_{11}^{-1} V_{12} \\ V_{22}^{-1} V_{21} & I_{m_2} \end{pmatrix}$$

Il est facile de voir que si $\underline{b} = \begin{pmatrix} \underline{b}_1 \\ \underline{b}_2 \end{pmatrix}$ est vecteur propre de M pour la valeur propre λ alors $\begin{pmatrix} \underline{b}_1 \\ -\underline{b}_2 \end{pmatrix}$ est vecteur propre pour la valeur $2 - \lambda$.

$$M\underline{b} = \begin{pmatrix} \underline{b}_1 + V_{11}^{-1} V_{12} \underline{b}_1 \\ \\ \underline{b}_2 + V_{22}^{-1} V_{21} \underline{b}_1 \end{pmatrix} = \begin{pmatrix} \lambda \underline{b}_1 \\ \\ \lambda \underline{b}_2 \end{pmatrix}$$

soit : $V_{11}^{-1} V_{12} \underline{b}_2 = (\lambda - 1) \underline{b}_1$ $\qquad$ $V_{22}^{-1} V_{21} \underline{b}_1 = (\lambda - 1) \underline{b}_2$

d'où :

$$(1 - \lambda) \underline{b}_1 = V_{11}^{-1} V_{12} (-\underline{b}_2) \qquad (1 - \lambda) (-\underline{b}_2) = V_{22}^{-1} V_{21} \underline{b}_1$$

soit encore :

$$\underline{b}_1 + V_{11}^{-1} V_{12} (-\underline{b}_2) = (2 - \lambda) \underline{b}_1 \qquad -\underline{b}_2 + V_{22}^{-1} V_{21} \underline{b}_1 = (2 - \lambda) \underline{b}_1$$

C.Q.F.D.

On retrouve au passage que $\underline{b}_1$ et $\underline{b}_2$ sont vecteurs propres respectivement de matrices $V_{11}^{-1} V_{12} V_{22}^{-1} V_{21}$ et $V_{22}^{-1} V_{21} V_{11}^{-1} V_{12}$ car on a :

$$\begin{cases} (\lambda - 1)^2 \underline{b}_1 = V_{11}^{-1} V_{12} V_{22}^{-1} V_{21} \underline{b}_1 \\ (\lambda - 1)^2 \underline{b}_2 = V_{22}^{-1} V_{21} V_{11}^{-1} V_{12} \underline{b}_2 \end{cases}$$

Les deux matrices précédentes étant respectivement de taille m_1 et m_2 si on suppose que $m_1 < m_2$ alors la deuxième possède $m_2 - m_1$ valeurs propres nulles donc $\lambda = 1$ est valeur propre de $P_1 + P_2$ avec une multiplicité au moins égale à $m_2 - m_1$.

On peut donc écrire la structure complète des valeurs propres de $P_1 + P_2$ où les λ_i sont rangées en ordre décroissant :

$$\lambda_i = \begin{cases} 2 \geqslant \lambda_i \geqslant 1 & \text{si} \quad 1 \leqslant i \leqslant m_1 \\ \lambda_i = 1 & \text{si} \quad m_1 < i \leqslant m_2 \\ 2 - \lambda_{m_1 + m_2 - i + 1} & \text{si} \quad m_2 < i \leqslant m_1 + m_2 \\ \lambda_i = 0 & \text{si} \quad m_1 + m_2 < i \leqslant n \end{cases} \qquad m_1 < m_2$$

On vérifie que $\sum_{i=1}^{n} \lambda_i = m_1 + m_2$ car $\sum_{i=1}^{n} \lambda_i = 2 m_1 + (m_2 - m_1)$

Lorsque $\lambda_i = 1$ avec $m_1 < i \leqslant m_2$ ceci prouve que $\underline{b}_1 = 0$ car $V_{11}^{-1} V_{12} V_{22}^{-1} V_{21}$ est de rang m_1. On en déduit que la variable $\underline{z}_i$ correspondante appartient à W_2.

Si on considère maintenant les facteurs $\underline{z}$ on remarque l'intéressante propriété que $\underline{z}_1, \underline{z}_2, \ldots, \underline{z}_m$, qui sont D_p - orthogonaux se projettent sur W_1 et W_2 selon des systèmes eux-mêmes D_p - orthogonaux. Cette propriété ne se conserve pas pour plus de deux ensembles.

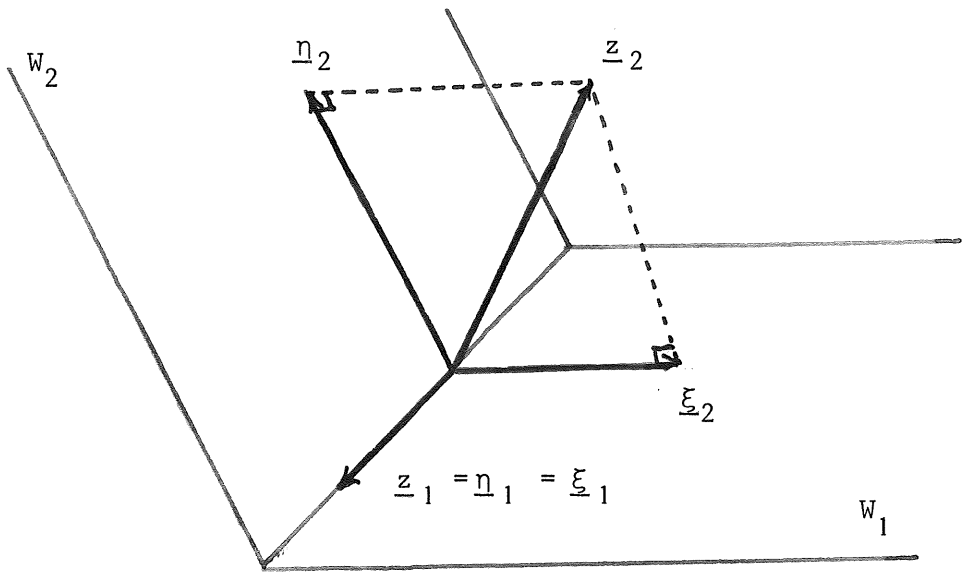
Il est facile de voir que $R_{i1}^2 = R_{i2}^2 = \frac{\lambda_i}{2}$ si R_{ij} désigne le coefficient de corrélation multiple du $i^{\text{ème}}$ facteur $\underline{z}_i$ avec X_j $(j = 1, 2)$. En effet on a :

$$(\lambda_i - 1)\ \underline{\xi}_i = P_1\ \underline{\eta}_i$$

d'où $R_{i1}^2 = \underline{z}_i' P_1 \underline{z}_i = \underline{z}_i' \underline{\xi}_i = \frac{1}{\lambda_i - 1}\ \underline{z}_i' P_1 \underline{\eta}_i = \frac{1}{\lambda_i - 1}\ \underline{\xi}_i' \underline{\eta}_i$

ce qui prouve que $R_{i1}^2 = R_{i2}^2$ $\forall i$, la relation étant symétrique.

A un coefficient multiplicateur près on a donc $\underline{z}_i = \underline{\eta}_i + \underline{\xi}_i$ ce qui peut se traduire par la figure suivante.



Le plan défini par $(\underline{z}_1, \underline{z}_2)$ s'interprète ici comme le bissecteur du dièdre formé par W_1 et W_2 ou plan médian.

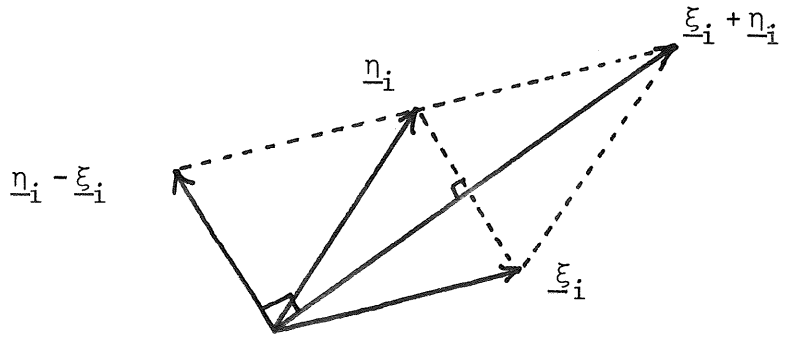
Les m_1 premiers facteurs $\underline{z}_i$ se projettent en formant une base D_p - orthogonale de W_1, les $m_2 - m_1$ facteurs suivants qui appartiennent à W_2 complètent la base de W_2 formée des projections D_p - orthogonales de $\underline{z}_1, \underline{z}_2, \ldots, \underline{z}_{m_1}$ sur W_2.

Les facteurs $\underline{z}_i$ d'ordre supérieur correspondent à des valeurs propres inférieures à l'unité et peuvent s'interpréter géométriquement de la manière suivante :

Les $\underline{z}_i$ pour $1 \leqslant i \leqslant m_1$ sont les bissectrices intérieures des angles aigüs formés par $\underline{\xi}_i$ et $\underline{\eta}_i$ tandis que les $\underline{z}_{m_1 + m_2 - i + 1}$ (relatifs aux valeurs propres $2 - \lambda_i$) sont les bissectrices extérieures de ces mêmes angles.

En effet si $\underline{z}_i$ s'écrit $\underline{\xi}_i + \underline{\eta}_i$ on a :

$$\underline{z}_{m_1 + m_2 - i + 1} = \underline{\xi}_i - \underline{\eta}_i \quad \text{puisqu'à} \quad \begin{pmatrix} \underline{b}_1 \\ \underline{b}_2 \end{pmatrix} \quad \text{correspond} \quad \begin{pmatrix} \underline{b}_1 \\ -\underline{b}_2 \end{pmatrix}.$$



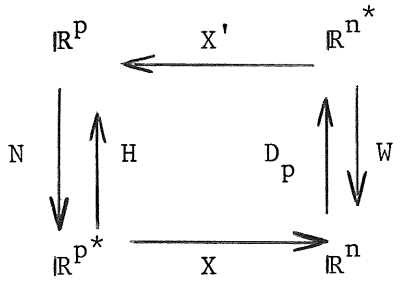
Comme $\underline{\xi}_i$ a même norme que $\underline{\eta}_i$ la figure est un losange et les diagonales sont les bissectrices des angles.

Pour résumer, la méthode factorielle fournit, lorsque $p = 2$, m_1 facteurs $\underline{z}$ qui sont les sommes des variables canoniques de l'analyse usuelle, les facteurs suivants complétant les premiers en une base de $W_1 \oplus W_2$ puis de $\mathbb{R}^n$ mais n'ont pas d'interprétation directe.

2. Identité avec l'analyse en composantes principales.

La méthode factorielle proposée n'est en fait qu'une analyse en composantes principales du tableau de données complet X dans une métrique particulière ainsi qu'on pouvait le constater sur les résultats obtenus en II.2. puisque les coefficients définissant le facteur $\underline{z}$ comme combinaison linéaire des variables sont donnés par le vecteur $\underline{b}$, vecteur propre de $V^{-1}H$. On reconnaît ici l'équation aux valeurs propres fournissant ce qu'on appelle les "facteurs principaux" (voir CEEE [6])lorsque l'espace des individus est muni de la métrique V^{-1}.

Rappelons en effet que si l'espace des variables est muni de la métrique D_p, l'espace des individus de la métrique N les "facteurs principaux" $\underline{b}$ sont les solutions de $NH\underline{b} = \lambda \underline{b}$ et les composantes principales de $WD_p\underline{z} = \lambda \underline{z}$ où V désigne la matrice de variance - covariance $H = X'D_pX$ et W la matrice XNX'. Ces relations entre ces divers opérateurs étant synthétisées dans le schéma de dualité suivant :



On peut donc énoncer le résultat suivant :

> La méthode factorielle proposée pour généraliser l'analyse canonique, est identique à une analyse en composantes principales dans laquelle chaque espace d'observation E_i relatif à un tableau X_i est muni de sa métrique de MAHALANOBIS V_{ii}^{-1}

Dans cette analyse en composantes principales WD_p n'est autre que la somme des projecteurs D_p - orthogonaux sur les espaces W_i.

En effet, d'après l'A.C.P. il vient :

$$X\, V^{-1}\, X'\, D_p\, \underline{z} = \lambda\, z$$

et il est facile de voir en décomposant $X\, V^{-1}\, X'$ par blocs que $X\, V^{-1}\, X'\, D_p$ n'est autre que $\sum_{i=1}^{p} P_i$:

$$X\, V^{-1}\, X' = (X_1 \mid X_2 \mid \ldots \mid X_p) \begin{pmatrix} V_{11}^{-1} & & & \\ & V_{22}^{-1} & & \\ & & \ddots & \\ & & & V_{pp}^{-1} \end{pmatrix} \begin{pmatrix} \underline{X_1'} \\ \underline{X_2'} \\ \vdots \\ \underline{X_p'} \end{pmatrix}$$

$$= \sum_{i=1}^{p} X_i\, V_{ii}^{-1}\, X_i'$$

d'où :

$$\boxed{X\, V^{-1}\, X'\, D_p = \sum_{i=1}^{p} P_i}$$

En particulier si $m_i = 1 \quad \forall\, i$, c'est-à-dire si il n'y a qu'une variable par groupe, l'analyse se réduit alors à l'analyse en composantes principales sur variables réduites du tableau X qui est alors un tableau $n \times p$:

Les matrices V_{ii} sont alors des matrices à un seul élément σ_i^2 si σ_i désigne l'écart-type de la $i^{\text{ème}}$ variable. La matrice V^{-1} est alors égale à la matrice diagonale des $\frac{1}{\sigma_i^2}$, ce qui revient à travailler sur des variables D_p-normées.

Les facteurs $\underline{z}$ ne sont autres alors que les composantes principales dans l'A.C.P. normée du tableau X et diagonaliser $V^{-1}\, H$ revient alors à diagonaliser la matrice des corrélations.

Dans ce cas, les coefficients de corrélation multiple de $\underline{z}$ avec les X_i ne sont que des coefficients de corrélation ordinaire ce qui démontre une propriété peu citée de l'analyse en composantes principales normée :

$\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_p$ étant p variables réduites, leur première composante

principale $\underline{z}$ rend maximale la quantité $\sum_{i=1}^{p} r(\underline{z}, \underline{x}_i)^2$.

Dans le cas général où $m_i \neq 1$ nous avons un résultat analogue liant $\underline{z}$ à ses projections D_p-orthogonales sur les W_i puisqu'alors les coefficients de corrélation multiple R_i de $\underline{z}$ avec X_i ne sont autres que les coefficients de corrélation simples de $\underline{z}$ avec ses projections d'où :

Proposition.

Tout facteur $\underline{z}_i$ est composante principale de ses projections D_p-normées et $\underline{z}_1$ est la première composante principale des variables canoniques D_p-normées associées à la valeur propre λ_1.

On ne peut dire en effet que $\underline{z}_i$ soit toujours première composante principale si $i \neq 1$ en raison du contre-exemple suivant :

Si $p = 2$ on a vu précédemment que $\underline{z}_i$ et $\underline{z}_{m_1 + m_2 - i+1}$ si $1 \leqslant i \leqslant m_1$ avaient pour projections respectives sur W_1 et W_2 les couples $(\underline{\xi}_i, \underline{\eta}_i)$ et $(\underline{\xi}_i, -\underline{\eta}_i)$, c'est-à-dire qu'ils définissent en fait les mêmes variables canoniques et on voit immédiatement que si $\underline{z}_i$ est première composante principale de $(\underline{\xi}_i, \underline{\eta}_i)$ tandis que $\underline{z}_{m_1 + m_2 - i+1}$ est deuxième composante principale de $(\underline{\xi}_i, -\underline{\eta}_i)$.

C'est le résultat précédent qui avait conduit M. MASSON à appeler cette méthode du nom de "méthode de la première composante principale optimale".

3. Analyse canonique généralisée et analyses canoniques partielles.

Nous établissons ici une propriété reliant les valeurs propres de $M = V^{-1}H$ aux coefficients de corrélation canonique que l'on obtiendrait en effectuant la $\frac{p(p-1)}{2}$ analyses canoniques possibles de X_i avec X_j.

La somme des carrés des valeurs propres de M est égale à la

trace de M^2

La trace de M^2 n'est autre que la somme des traces de ses blocs diagonaux. Or, si M est écrite sous forme de blocs M_{ij} avec $M_{ij} = V_{ii}^{-1} V_{ij}$ les

blocs diagonaux B_i de M^2 sont de la forme :

$$B_i = \sum_{j=1}^{p} M_{ij} M_{ji} = \sum_{j=1}^{p} V_{ii}^{-1} V_{ij} V_{jj}^{-1} V_{ji}$$

La trace de $V_{ii}^{-1} V_{ij} V_{jj}^{-1} V_{ji}$ si $i \neq j$ est égale à la somme des carrés des coefficients de corrélation canonique de l'analyse de X_i avec X_j. Si $i = j$ la trace vaut m_i car il s'agit alors de la matrice unité I_{m_i}.

La trace d'un bloc diagonal B_i de M^2 vaut donc :

$$m_i + \sum_{j=1}^{p} \sum_{k} \left[r_{ij}^{(k)}\right]^2 \quad \text{où} \quad r_{ij}^{(k)} \text{ désigne le } k^{\text{ème}} \text{ coefficient}$$

de corrélation canonique de l'analyse de X_i avec X_j.

La somme des traces des blocs diagonaux vaut donc :

$$\sum_{i=1}^{m} \lambda_i^2 = m + 2 \sum_{i=1}^{p-1} \sum_{j=i+1}^{p} \sum_{k} \left[r_{ij}^{(k)}\right]^2$$

si $m = \sum_{i=1}^{p} m_i$

Comme on sait que $\sum_{i=1}^{p} \lambda_i = m$ on a alors le résultat suivant :

> $\sum_{i=1}^{m} (\lambda_i - 1)^2$ est égal à deux fois la somme de tous les carrés des coefficients de corrélation canoniques de toutes les analyses canoniques distinctes de X_i avec X_j.

On peut aussi dire que la somme des carrés des valeurs propres est égale à la somme des carrés de tous les coefficients de corrélation canonique de toutes les analyses canoniques possibles, (y compris X_i avec X_i) sans distinction.

Pour résumer, nous avons étudié une méthode d'analyse des liaisons entre plusieurs ensembles d'observations, qui est une généralisation simple de l'analyse en composantes principales normée où tout se passe comme si on remplaçait une variable par un groupe de variables, une variance par une matrice de

variance - covariance, un coefficient de corrélation simple par un coefficient de corrélation multiple.

Nous avons établi au passage l'équivalence entre l'analyse canonique classique et l'analyse en composantes principales, montrant ainsi que toutes les analyses de données linéaires sont issues d'un même modèle.

IV. CONSEQUENCES PRATIQUES.

1. Nombre de facteurs utiles et reconstitution des données.

Les propriétés usuelles de l'analyse en composantes principales sont valables pour la méthode factorielle proposée. On peut interpréter ainsi les valeurs propres λ_i comme des inerties portées par des directions de $\mathbb{R}^{\sum m_i}$ muni de la métrique V^{-1} et considérer pour un ensemble de k facteurs $\underline{z}_1, \underline{z}_2, \ldots, \underline{z}_k$ la qualité de la représentation au moyen du rapport habituel.

$$\frac{\lambda_1 + \lambda_2 + \ldots + \lambda_k}{\text{Trace}(V^{-1}H)} = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{p} m_i}$$

Pour retrouver exactement les propriétés de l'A.C.P. il suffit de poser $\underline{z} = X\underline{b}$ (au lieu de $\lambda \underline{z} = X\underline{b}$) et alors si $\underline{z}$ est D_p-normé à 1 $\underline{b}$ est H-normé à 1 et V-normé à $\frac{1}{\sqrt{\lambda}}$.

Dans l'A.C.P. de métrique N il est usuel de N^{-1}-normer les "facteurs principaux" c'est-à-dire le vecteur $\underline{b}$:

ici on conviendra de V-normer $\underline{b}$ à 1 ce qui entraîne que $\underline{z}$ est D_p-normé à $\sqrt{\lambda}$; dans ces conditions λ_i est la variance du facteur $\underline{z}_i$, et on retrouve le fait que l'A.C.P. de métrique N revient à chercher une variable de la forme $X\underline{b}$ de variance maximale sous la contrainte que $\underline{b}$ est de N^{-1}-norme égale à l'unité.

Il est clair que, si on extrait tous les facteurs correspondant à des valeurs propres non nulles, on pourra reconstituer complètement le tableau de

données puisque les $\underline{z}_i$ correspondant formeront une base de $\bigoplus_{i=1}^{p} W_i$.

Remarquons que dans le cas $p = 2$, si $W_1 \cap W_2 = \{0\}$, les m_1 premiers facteurs $\underline{z}$ n'engendrent pas W_1 car ils se trouvent dans un espace "intermédiaire" entre W_1 et W_2, ce sont leurs projections D_p-orthogonales qui forment une base de W_1 . (m_1 désigne ici le plus petit des m_i).

Si $p > 2$, les projections D_p-orthogonales des $\underline{z}_1, \underline{z}_2, \ldots, \underline{z}_m$, ne forment pas en général une base D_p-orthogonale de W_1 ; on peut supposer qu'elles forment une base de W_1 mais c'est une conjecture que nous n'avons pu encore établir ni réfuter.

Etablissons maintenant la formule de reconstitution du tableau de données X au moyen des vecteurs propres D_p-normés de $XV^{-1}X'D_p$ et des vecteurs propres V-normés de $V^{-1}H$, associés aux mêmes valeurs propres entre lesquels on a donc la relation :

$$X\underline{b} = \sqrt{\lambda}\,\underline{z}$$

Si nous notons $\underline{b}^i$ le $i^{ème}$ vecteur propre V-normé de $V^{-1}H$ pour le différencier de la notation utilisée pour les blocs de $\underline{b}$, on tire de la relation précédente :

$$X\,\underline{b}^i\,(V\,\underline{b}^i)' = \sqrt{\lambda_i}\,\underline{z}_i\,(V\,\underline{b}^i)'$$

En sommant membre à membre pour $i = 1, 2, \ldots, m = \sum_{i=1}^{p} m_i$, il vient :

$$X \sum_{i=1}^{m} \underline{b}^i\,(V\underline{b}^i)' = \sum_{i=1}^{m} \sqrt{\lambda_i}\,\underline{z}_i\,(V\underline{b}^i)'$$

Les $\underline{b}^i$ formant une base V-orthonormée de $\mathbb{R}^n$, la matrice $\sum_{i=1}^{m} \underline{b}^i(V\underline{b}^i)'$ n'est autre que la matrice unité I_m comme on le constate aisément en l'appliquant aux divers vecteurs $\underline{b}^j$ de la base :

$$\sum_{i=1}^{m} \underline{b}^i(V\underline{b}^i)'\underline{b}^j = \sum_{i=1}^{p} \underline{b}^i\,\underline{b}^{i\prime}\,V\,\underline{b}^j = \delta_{ij}\,\underline{b}^j$$

où δ_{ij} est le symbole de Kronecker.

On a alors la formule de reconstitution du tableau global :

$$X = \sum_{i=1}^{m} \sqrt{\lambda_i}\ \underline{z}_i\ (V\,\underline{b}^i)'$$

et en écrivant par blocs avec $\underline{b}^i = \begin{pmatrix} \underline{b}_1^i \\ \underline{b}_2^i \\ \\ \underline{b}_p^i \end{pmatrix}$

on a la formule de reconstitution de chaque tableau X_j :

$$\boxed{X_j = \sum_{i=1}^{m} \sqrt{\lambda_i}\ \underline{z}_i\ (V_{jj}\ \underline{b}_j^i)'}$$

2. Représentation graphique des variables.

Les facteurs $\underline{z}_1, \underline{z}_2, \ldots, \underline{z}_m$ constituent une base D_p-orthonormée de $\bigoplus_{i=1}^{p} W_i$. Sur cette base les variables $\underline{x}_k$ D_p-normées (ou colonnes réduites de X) ont pour coordonnées :

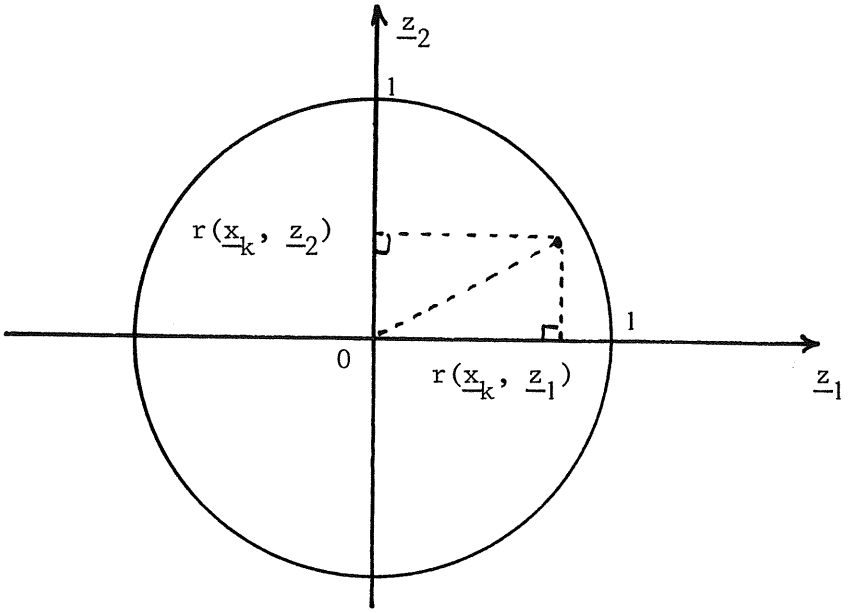
$$\alpha_{ik} = r(\underline{x}_k\ ,\ \underline{z}_i)$$

et

$$\underline{x}_k = \sum_{i=1}^{m} \alpha_{ik}\ \underline{z}_i$$

C'est-à-dire que les projections de $\underline{x}_k$ sur les $\underline{z}_i$ ont pour mesure algébrique les coefficients de corrélation linéaire de $\underline{x}_k$ avec $\underline{z}_i$.

Si on se limite à une représentation plane des proximités entre variables on fera figurer les projections des $\underline{x}_k$ sur le plan $(\underline{z}_1, \underline{z}_2)$ selon la figure dite du cercle des corrélations. La distance de l'origine à la projection de $\underline{x}_k$ est égale au coefficient de corrélation multiple de $\underline{x}_k$ avec $(\underline{z}_1, \underline{z}_2)$: une variable est d'autant mieux représentée que sa projection est voisine de la circonférence du cercle.



Cette figure aide à l'interprétation des facteurs $\underline{z}_i$.

En particulier lorsque $p = 2$, la considération du plan $(\underline{z}_1, \underline{z}_2)$ permet de lever l'ambiguïté courante en analyse canonique où deux représentations planes sont possibles : l'une dans le plan $(\underline{\xi}_1, \underline{\xi}_2)$; l'autre dans le plan $(\underline{\eta}_1, \underline{\eta}_2)$ c'est-à-dire qu'il faut choisir entre une représentation dans W_1 ou dans W_2 .

Nous proposons donc de projeter dans le plan intermédiaire défini par les axes $(\underline{\xi}_1 + \underline{\eta}_1 \; ; \; \underline{\xi}_2 + \underline{\eta}_2)$ ce qui respecte alors la symétrie entre les deux ensembles de données. On peut alors parler véritablement de représentation simultanée des deux ensembles de variables, cette représentation s'effectuant sur le sous-espace de dimension 2 le plus proche de W_1 et W_2 .

3. Représentation des groupes de variables.

On peut chercher à représenter directement les p groupes de variables afin d'étudier des proximités globales. Plusieurs approches sont possibles dont aucune ne sous semble réellement satisfaisante.

On peut tout d'abord résumer chaque groupe de variables par sa première variable canonique D_p-normée et étudier leurs proximités en projection dans le plan $(\underline{z}_1, \underline{z}_2)$. Cependant si $\underline{z}_1$ est bien la première composante principale des p variables canoniques $\underline{z}_2$ n'en est pas la deuxième et la considération

du plan ($\underline{z}_1$, $\underline{z}_2$) n'est pas aussi naturelle que l'on pourrait le croire. On peut envisager alors de remplacer $\underline{z}_2$ par la deuxième composante principale des p premières variables canoniques.

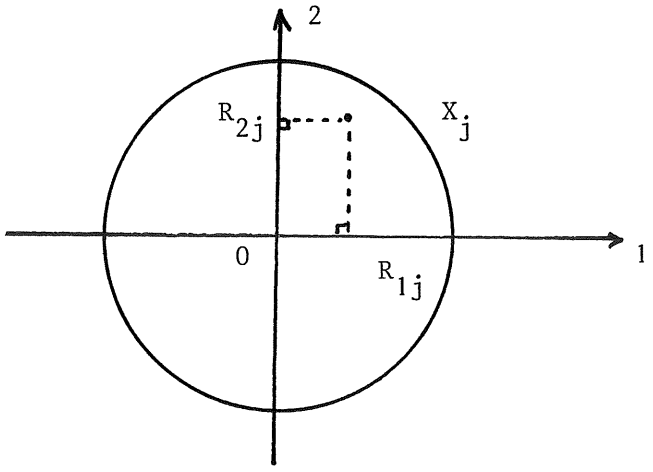
Il demeure cependant que la considération des p premières variables canoniques ne reflète qu'un aspect limité des proximités entre ensembles.

Nous proposons la méthode suivante qui semble une extension naturelle de l'analyse en composantes principales puisqu'ici tout se passe comme si on effectuait une A.C.P. normée en remplaçant des variables par des groupes de variables :

décrivons chaque groupe de variables par ses coefficients de corrélation multiple avec les $\underline{z}_i$ et effectuons la représentation graphique suivante dans un plan.

Un groupe de variables X_j sera représenté dans un système orthonormé par le point de coordonnées R_{1j} et R_{2j} qui sera toujours dans le premier quadrant à moins d'affecter à R_{1j} le signe du coefficient de corrélation simple entre $\underline{z}_1$ et $P_1\,\underline{z}_1$.

Cette représentation permet d'interpréter globalement les facteurs $\underline{z}_i$ en examinant les groupes de variables dont ils sont proches.



Cependant cette représentation peut conduire à des proximités illusoires entre groupes car il est très possible d'observer deux points confondus sans que les espaces engendrés le soient. La preuve en est fournie par l'analyse canonique ordinaire où les points représentatifs de X_1 et X_2 sont toujours confondus puisque $R^2_{i1} = R^2_{i2}$ $\forall\, i$ comme nous l'avons montré précédemment.

De plus si en analyse en composantes principales on a la relation :

$$\sum_{i=1}^{p} r^2(\underline{x}_j, \underline{z}_i) = 1 \quad \forall j .$$

car les composantes principales forment une base D_p-orthonormée de $\mathbb{R}^n$, il n'existe pas de relation équivalente pour les corrélations multiples.

Une méthode à l'abri de ces difficultés sera développée au chapitre IV en faisant appel à un point de vue très différent.

4. Pour mémoire : représentation des individus. (Cf. [6])

Comme dans toute analyse en composantes principales on peut représenter les proximités entre individus en utilisant la description qui en est faite par les composantes principales.

Moyennant certaines précautions d'échelle on peut alors effectuer une représentation plane simultanée où figurent tous les individus et toutes les variables.

V. UNE INTERPRETATION EN TERMES D'ANALYSE DE VARIANCE.

L'analyse précédente permet de savoir s'il existe une intersection non vide aux W_i, c'est-à-dire si on peut trouver des vecteurs de coefficients, $\underline{b}_1, \underline{b}_2, \ldots, \underline{b}_p$ tels que $X_1 \underline{b}_1 = X_2 \underline{b}_2 = \ldots = X_p \underline{b}_p$.

S'il en est ainsi on définit donc une mesure commune de tous les individus au moyen du premier facteur $\underline{z}$.

Il est clair que rechercher des $\underline{b}_i$ tels que les $X_i \underline{b}_i$ soient aussi voisins que possible est un problème d'analyse canonique généralisée. En nous fixant un critère dérivé de considérations d'analyse de variance nous montrerons que nous aboutissons à la méthode précédente et non à une des autres variantes proposées (Horst, Kettenring, Masson ...).

Nous supposerons que $D_p = D_{1/n} = \frac{1}{n} I_n$.

Posons $\underline{y}_j = X_j \underline{b}_j$ et appellons y_{ij} la $i^{ème}$ composante de $\underline{y}_j$.

Considérons alors le tableau Y à n lignes dont les p colonnes sont les $\underline{y}_j$.

Nous allons chercher à maximiser l'homogénéité de ce tableau au sens où nous désirons que les valeurs y_{ij} $j = 1, 2, \ldots, p$ pour un même individu soient peu différentes, nous inspirant en cela d'une idée de P. HORST [10] reprise dans [11].

Remarquons au passage que si X est un tableau de variables centrées alors Y aussi et la somme de tous les éléments de Y est nulle.

Il est équivalent de rechercher à rendre maximale la dispersion des moyennes de lignes car les individus sont d'autant mieux séparés que l'on peut les mesurer sur une échelle unique.

La dispersion des moyennes de ligne est mesurée par :

$\sum_{i=1}^{n} y_{i.}^2$ où $y_{i.}$ désigne la moyenne de la ligne n° i.

Comme $y_{ij} = y_{i.} + (y_{ij} - y_{i.})$ et que $y_{.J} = y_{..} = 0$

la relation classique d'analyse de variance s'écrit ici :

$$\sum_{i=1}^{n} \sum_{j=1}^{p} (y_{ij})^2 = p \sum_{i=1}^{n} (y_{i.})^2 + \sum_{i=1}^{n} \sum_{j=1}^{p} (y_{ij} - y_{i.})^2$$

et le problème posé est de maximiser le rapport :

$$\frac{p \sum_{i=1}^{n} (y_{i.})^2}{\sum_{i=1}^{n} \sum_{j=1}^{p} (y_{ij})^2}$$

or $\underline{y}_{\cdot} = \frac{1}{p} \sum_{j=1}^{p} X_j \underline{b}_j$ et $\Sigma\, y_{i.}^2 = \|\underline{y}_{\cdot}\|^2 = \frac{1}{p^2} \sum_j \sum_k \underline{b}_j' X_j' X_k \underline{b}_k$

si on désigne par $\underline{y}_{\cdot}$ le vecteur de $\mathbb{R}^n$ des moyennes de lignes.

d'où $\sum_{i=1}^{n} (y_{i.})^2 = \frac{n}{p^2} \underline{b}' H \underline{b}$ où $\underline{b} = \begin{pmatrix} \underline{b}_1 \\ \underline{b}_2 \\ \vdots \\ \underline{b}_p \end{pmatrix}$

Par ailleurs $\sum_i \sum_j y_{ij}^2 = \sum_{j=1}^{p} (X_j \underline{b}_j)' (X_j \underline{b}_j) = \sum_{j=1}^{p} \underline{b}_j' X_j' X_j \underline{b}_j = n \underline{b}' V \underline{b}$

Le rapport à maximiser est donc égal à $\frac{\underline{b}' H \underline{b}}{p \underline{b}' V \underline{b}}$

Le vecteur $\underline{b}$ assurant le maximum est le premier vecteur propre de $V^{-1} H$ et le maximum vaut alors $\frac{\lambda_1}{p}$, ce qui nous ramène bien à la méthode générale de ce chapitre.

Si on dresse le tableau usuel d'analyse de la variance on a :

Source de variation	degré de liberté
dûe aux observations $p \sum z_{i.}^2 = \frac{n}{p} \underline{b}' H \underline{b}$	$n - 1$
résiduelle $\sum_i \sum_j (z_{ij} - z_{i.})^2 = n (\underline{b}' V \underline{b} - \frac{\underline{b}' H \underline{b}}{p})$	$np - n$
Variation totale $\sum_i \sum_j (z_{ij})^2 = n \underline{b}' V \underline{b}$	$np - 1$

$\frac{\lambda_1}{p}$ a donc le sens d'une part de variance expliquée (pour le vecteur $\underline{b}$ optimal), c'est-à-dire d'un rapport de corrélation mesurant la variance expliquée par les différences entre individus, par rapport à la variance totale.

Si on considère le rapport F usuel on a :

$F = \frac{\underline{b}' H \underline{b} \ / \ p}{\underline{b}' V \underline{b} - \frac{\underline{b}' H \underline{b}}{p}} \ \frac{np - n}{n - 1}$ qui vaut lorsque $\underline{b}$ est vecteur propre de $V^{-1} H$ associé à λ_1 :

$$F = \frac{\lambda_1}{p - \lambda_1} \ \frac{np - 1}{n - 1}$$

Si l'utilisation du test de Fisher - Snedecor nous semble hasardeuse pour savoir si une valeur propre est significative, cette valeur de F n'en reste pas moins un indice descriptif intéressant pour mesurer la qualité d'une valeur propre et qui est complémentaire de la part de trace $\frac{\lambda_1}{m}$ habituellement considérée.

Si on remarque que H peut s'écrire sous la forme $V + C$ où C est une matrice de "covariances seules" :

$$C = \begin{pmatrix} 0 & V_{12} & \cdots & V_{1p} \\ V_{21} & 0 & & \vdots \\ \vdots & & \ddots & \vdots \\ \vdots & & & 0 \end{pmatrix}$$

La maximisation du rapport $\frac{\underline{b}' H \underline{b}}{\underline{b}' V \underline{b}}$ équivaut alors à la maximisation de $\frac{\underline{b}' C \underline{b}}{\underline{b}' V \underline{b}}$ ce qui prouve que $\underline{b}$ est ainsi vecteur propre de $V^{-1}C$ mais avec la valeur propre $\lambda - 1$. Si on développe $\underline{b}' C \underline{b}$, on voit alors que $\underline{b}$ a la propriété de rendre maximale la somme des covariances des variables canoniques $X_i \underline{b}_i$ sous la contrainte de normalisation $\underline{b}' V \underline{b} = 1$:

$$\underline{b}' C \underline{b} = \sum_{i \neq j} \sum \underline{b}_i' V_{ij} \underline{b}_j = \sum_{i \neq j} \sum \operatorname{cov} (X_i \underline{b}_i , X_j \underline{b}_j).$$

VI. COMPARAISON AVEC LA METHODE DE HORST.

Rappelons que cette méthode cherche des variables canoniques réalisant le maximum de la somme de tous leurs coefficients de corrélation linéaire deux à deux, et que la solution en est donnée par le système d'équations :

$$\sum_{j=1}^{p} V_{ii} \underline{b}_j = \lambda_i V_{ii} \underline{b}_i \qquad i = 1, 2, \ldots, p .$$

Pour connaître la signification des multiplicateurs λ_i, multiplions de chaque côté à gauche par $\underline{b}_i$, il vient alors :

$$\lambda_i = \underline{b}_i' \sum_j V_{ij} \underline{b}_j \quad \text{car la contrainte} \quad \underline{b}_i' V_{ii} \underline{b}_i = 1$$

doit être satisfaite à l'optimum - on trouve alors par sommation que $\sum_{i=1}^{p} \lambda_i$ est égal à la somme des corrélations deux à deux des $X_i \underline{b}_i$ c'est-à-dire au critère à maximiser.

Si nous écrivons $V_{ij} = X_i' D_p X_j$ les équations deviennent :

$$X_i' D_p \sum_{j=1}^{p} X_j \underline{b}_j = \lambda_i V_{ii} \underline{b}_i$$

soit $$X_i V_{ii}^{-1} X_i' D_p \sum_{j=1}^{p} X_j \underline{b}_j = \lambda_i X_i \underline{b}_i$$

ou encore $$P_i \underline{z} = \lambda_i X_i \underline{b}_i \qquad i = 1, 2, \ldots, p$$

en posant $$\underline{z} = \sum_{j=1}^{p} X_j \underline{b}_j .$$

Ces équations sont très voisines de celle de l'analyse factorielle de X dans la métrique V^{-1} qui sont $P_i \underline{z} = \lambda X_i \underline{b}_i \quad i = 1, 2, \ldots, p$.
La différence essentielle résidant dans les multiplicateurs de Lagrange qui correspondent à des contraintes de normalisation différentes.

On peut dresser le tableau suivant qui fait apparaître les ressemblances entre les formules des deux méthodes.

Analyse factorielle	Méthode de HORST
$P_i \underline{z} = \lambda X_i \underline{b}_i$	$P_i \underline{z} = \lambda_i X_i \underline{b}_i$
$\underline{z} = \sum_{i=1}^{p} X_i \underline{b}_i$	$\underline{z} = \sum_{i=1}^{p} X_i \underline{b}_i$
$\underline{z}' D_p \underline{z} = 1$	$\underline{b}_i' V_{ii} \underline{b}_i = 1 \quad \forall i$
λ maximal	$\sum_{i=1}^{p} \lambda_i$ maximal

De plus on peut dire que si la méthode de HORST recherche à maximiser la somme des corrélations des variables canoniques, l'analyse factorielle proposée maximise la somme des covariances de ces composantes, leur somme étant D_p-normée ainsi que nous l'avons établi au paragraphe précédent.

Dans l'analyse factorielle la variable $\underline{z}$ est telle que la somme des carrés de ses corrélations avec ses projections est maximale tandis que dans la méthode de HORST $\underline{z}$ est telle que la somme de ses corrélations est maximale.

En effet, puisque dans cette méthode $X_i \underline{b}_i$ est D_p-normé on a :

$$r\ (\underline{z}\ ;\ X_i\ \underline{b}_i) = \frac{\underline{z}'\ D_p\ X_i\ \underline{b}_i}{\sqrt{\underline{z}'\ D_p\ \underline{z}}} = \frac{\sum_{j=1}^{p} \underline{b}_j'\ X_j'\ D_p\ X_i\ \underline{b}_i}{\sqrt{\underline{z}'\ D_p\ \underline{z}}} = \frac{\sum_{j=1}^{p} \underline{b}_j'\ V_{ji}\ \underline{b}_i}{\sqrt{\underline{z}'\ D_p\ \underline{z}}}$$

d'où $r\ (\underline{z}\ ;\ X_i\ \underline{b}_i) = \frac{\lambda_i}{\|\underline{z}\|} = \frac{\lambda_i}{\sqrt{\sum_{j=1}^{p} \lambda_j}}$

car $\|\underline{z}\|^2 = \sum_i \sum_j \underline{b}_i'\ X_i'\ D_p\ X_j\ \underline{b}_j = \sum_{j=1}^{p} \lambda_j$.

et alors $\sum_{i=1}^{p} r\ (\underline{z}\ ;\ X_i\ \underline{b}_i) = \sqrt{\sum_{i=1}^{p} \lambda_i}$, or $\sum_{i=1}^{p} \lambda_i$ est précisément le critère à maximiser.

Les propriétés précédentes montrent la grande parenté des deux méthodes. Cependant la méthode factorielle nous semble préférable à celle de HORST car elle constitue une extension naturelle des méthodes usuelles d'analyse des données (indépendamment du fait qu'elle conduit à des solutions plus aisément calculables) et que ses résultats s'interprètent simplement. Notons pour finir que si $m_i = 1 \quad \forall i$, la méthode factorielle n'est autre que l'A.C.P. normée tandis que la méthode de HORST ne peut s'appliquer (son facteur $\underline{z}$ n'étant autre que $\sum_{i=1}^{p} \frac{x_i}{\sigma_i}$ ce qui est sans intérêt).

CHAPITRE III

LE TRAITEMENT DES VARIABLES QUALITATIVES PAR CODAGE

Ainsi que nous l'avons vu au chapitre I, toute variable qualitative à m modalités est équivalente à l'ensemble des m variables indicatrices de ses modalités. La description des relations entre p variables qualitatives peut donc s'effectuer en appliquant la méthode factorielle du chapitre II dans ce cas particulier. Nous traiterons ensuite deux problèmes prévisionnels où les variables explicatives sont qualitatives : celui de la régression (critère quantitatif), celui de la discrimination (critère qualitatif).

I. L'ANALYSE CONJOINTE DE p VARIABLES QUALITATIVES.

Le tableau de données X consiste alors en p tableaux logiques X_j où dans une ligne i de X_j on trouve $m_j - 1$ zéros et une fois 1 dans la colonne correspondant à la modalité prise par l'individu n° i .

La somme des vecteurs colonnes de X_j est donc toujours égale à $\underline{1}$, la somme des vecteurs colonnes de X vaut $p\ \underline{1}$.

$$X = \begin{array}{c} \\ 1 \\ 2 \\ \vdots \\ \\ n \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & \cdots & \cdots & \cdots & \cdots & n_1 \end{array} \\ \left(\begin{array}{cccccc} 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ & & & & & \\ & & & X_1 & & \\ & & & & & \\ & & & & & \end{array}\right. \end{array} \begin{array}{c} \begin{array}{cccccccc} 1 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & n_2 \end{array} \\ \left.\begin{array}{cccccccc} 0 & 0 & 1 & 0 & 0 & 0 & 0 & \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & \\ & & & & & & & \\ & & & X_2 & & & & \\ & & & & & & & \\ & & & & & & & \end{array}\right| \end{array} \cdots\cdots \begin{array}{c} \begin{array}{cccc} 1 & \cdots & \cdots & n_p \end{array} \\ \left|\begin{array}{cccc} & & & \\ & & & \\ & & & \\ & X_p & & \\ & & & \\ & & & \end{array}\right) \end{array}$$

La somme des éléments d'une colonne est égale à l'effectif de la modalité correspondante.

Nous considérerons dans ce chapitre que les individus sont affectés de poids tous égaux, c'est-à-dire que $D_p = \frac{1}{n} I_n$.

Afin d'allèger les écritures nous omettrons la plupart du temps d'effectuer la division par n ce qui ne change rien aux résultats.

Contrairement à ce qui existait au chapitre II, les données ne sont pas centrées, car d'une part, centrer des variables logiques est dénué d'intérêt et d'autre part, les analyses centrées et non centrées conduisent aux mêmes résultats au facteur trivial $\underline{1}$ près.

Dans ces conditions les matrices V_{ii} et V_{ij} introduites au chapitre précédent sont telles que :

$$V_{ii} = X_i' X_i$$
$$V_{ij} = X_i' X_j$$

V_{ii} est la matrice diagonale des effectifs des modalités de la $i^{ème}$ variable qualitative.

V_{ij} est le tableau de contingence obtenu par croisement des variables qualitatives n° i et n° j.

1. Application de la méthode factorielle.

Les variables canoniques vont alors fournir les codages descriptifs successifs des diverses variables qualitatives.

En conservant nos notations, le vecteur $\underline{b}$ regroupant tous les codages est vecteur propre de la matrice M d'éléments blocs $M_{ij} = V_{ii}^{-1} V_{ij}$. Ou ce qui revient au même ici de la matrice $V^{-1}X'X$ car $D_p = I$.

$$M = \begin{pmatrix} (X_1' X_1)^{-1} & & & \\ & (X_2' X_2)^{-1} & & \\ & & \ddots & \\ & & & (X_p' X_p)^{-1} \end{pmatrix} X' X$$

M s'exprime donc sous forme du produit de la matrice X par sa transposée X' dans laquelle on a divisé chaque élément par le total de sa colonne puisque $(X_i' X_i)^{-1}$ est la matrice diagonale des inverses des effectifs des modalités de la $i^{ème}$ variable.

Comme les sommes de ligne sont constantes et égales à p on ne change pas l'analyse en divisant les termes de X par p et $\underline{b}$ est vecteur propre de $V^{-1} X' \frac{1}{p} X$.

Ainsi que l'a remarqué MASSON [17] tout se passe donc comme si on faisait une analyse canonique formelle sur le tableau X considéré comme un tableau de contingence P selon les formules du chapitre I § I.3. puisque diagonaliser $D_x^{-1} P D_y^{-1} P'$ revient à diagonaliser la matrice obtenue en divisant les lignes de P par les totaux de lignes, multipliée par la transposée de P où on a divisé les colonnes par les sommes de colonnes.

L'analyse canonique de 2 variables discrètes n'étant autre que l'analyse des correspondances du tableau de contingence, faire l'analyse conjointe de p variables qualitatives au moyen de la méthode exposée au chapitre précédent revient à effectuer l'analyse des correspondances du tableau X.

Ce résultat très important justifie la pratique courante de l'analyse des correspondances sur ce qu'on appelle des variables mises sous forme disjonctive complète, pratique apparaissant jusque là comme une recette sans fondements mathématiques.

La matrice $X'X$ qui regroupe les divers tableaux de contingences possibles obtenues en croisant deux à deux avec répétition les p variables qualitatives avait déjà été considéré par BURT [4] qui recommandait d'en extraire les vecteurs propres pour obtenir des codages des p variables qualitatives.

En d'autres termes BURT effectuait une analyse en composantes principales sur X avec la métrique $I_{\sum_{i=1}^{p} m_i}$, au lieu de prendre comme ici la métrique V^{-1}.

Ce faisant BURT ne pondérait pas les diverses modalités des variables ce qui entraîne que les composantes principales (i.e. les facteurs $\underline{z}$ du chapitre précédent) sont "attirées" par les modalités à fort effectif, ce qui est en particulier le cas pour une variable qualitative à 2 modalités, alors que l'analyse factorielle avec la métrique V^{-1} revient à donner des poids égaux aux diverses modalités en les divisant par leurs effectifs.

Comme l'effectif d'une modalité est égal au moment d'ordre 2 non centré de la variable correspondante (à $\frac{1}{n}$ près) on peut dire que :

"L'analyse factorielle de p variables qualitatives n'est autre que l'analyse en composantes principales non centrée et normée du tableau X".

2. Propriétés des valeurs propres de M.

Le fait de travailler sur des variables indicatrices entraîne les propriétés particulières suivantes déduites des propriétés générales étudiées au chapitre II.

Tous les espaces W_i ayant en commun le vecteur $\underline{1}$, celui-ci est vecteur propre trivial de $\sum_{i=1}^{p} P_i$ associée à la valeur propre p, ce qui entraîne que $\underline{b} = \frac{\underline{1}}{\sum m_i}$ est vecteur propre de M associé à la même valeur propre.

Comme il existe alors $p-1$ relations linéaires indépendantes entre les colonnes de X, la dimension de $\bigoplus_{i=1}^{p} W_i$ est ou plus égale à $\sum_{i=1}^{p} m_i - (p-1)$ ce qui entraîne que zéro est valeur propre de M avec une multiplicité au moins égale à $p-1$.

M possède donc $(m-p)$ valeurs propres non triviales (si $m = \sum_{i=1}^{p} m_i$).

La somme de toutes les valeurs propres de M étant égale à m d'après un résultat précédent on en déduit que les (m - p) valeurs propres non triviales ont une somme égale à (m - p).

L'application du théorème concernant la somme des $(\lambda_i - 1)^2$ fait apparaître alors une relation intéressante entre les valeurs propres et les ϕ^2 de contingence des divers tableaux possibles puisque l'on sait que la somme des carrés des coefficients de corrélation canonique dans l'analyse centrée de deux variables discrètes est égale au ϕ^2.

La relation citée donne donc :

$$\sum_{i=1}^{m-p} (\lambda_i - 1)^2 = 2 \sum_{i=1}^{p-1} \sum_{j=i+1}^{p} \phi^2_{ij}$$

ou encore :

$$\boxed{\sum_{i=1}^{m-p} \lambda_i^2 = (m-p) + 2 \sum_{i=1}^{p-1} \sum_{j=i+1}^{p} \phi^2_{ij}}$$

puisque $\sum_{i=1}^{m-p} \lambda_i = m - p$.

3. Codages et représentations graphiques.

A chaque vecteur propre $\underline{b}$ de M correspond un codage $\underline{b}_i$ des m_i modalités de la variable n° i puisque :

$$\underline{b} = \begin{pmatrix} \underline{b}_1 \\ \underline{b}_2 \\ \\ \underline{b}_p \end{pmatrix}$$ et un codage correspondant $X_i \underline{b}_i$ des individus.

Le codage résultant $X\underline{b} = \sum_{i=1}^{n} X_i \underline{b}_i$ n'est autre, à une constante multiplicative près que le facteur $\underline{z}$ réalisant la meilleure description de tous les individus au sens de la somme des carrés des corrélations avec les différents codages.

Le carré du cosinus de l'angle entre $\underline{z}$ et W_i n'est autre que le carré

du rapport de corrélation η^2 entre $\underline{z}$ et la $i^{ème}$ variable qualitative comme le montre un calcul évident. La variable $\underline{z}$ possède donc la propriété d'être la plus liée à toutes les variables qualitatives au sens de la somme des carrés des rapports de corrélation.

Ainsi que nous l'avons vu au chapitre II, il est possible de faire une représentation simultanée des modalités de toutes les variables par projection orthogonale sur le plan défini par $(\underline{z}_1, \underline{z}_2)$. Ce procédé est particulièrement recommandé pour le cas où $p = 2$ (analyse des correspondances ordinaire) ce qui revient à se placer dans le plan "médian" et permet de résoudre élégamment le problème du choix de l'espace de représentation (facteurs de W_1 ou de W_2) et de se passer des principes "barycentriques" délicats à manipuler.

Ces sujets faisant l'objet d'une abondante littérature, nous ne nous étendrons pas plus.

II. LA REGRESSION SUR VARIABLES QUALITATIVES.

1. Généralités sur le modèle linéaire.

a) Le cas ordinaire.

Soit un tableau de données numériques à n lignes et $m+1$ colonnes de la forme suivante :

$$\begin{array}{c} \\ 1 \\ 2 \\ \cdot \\ \cdot \\ n \end{array} \begin{array}{|c|c|} \hline 1 & 1\ 2\ .\ .\ .\ .\ m \\ \hline \underline{y} & X_1 \\ \hline \end{array}$$

La régression linéaire peut être alors définie comme une technique permettant d'ajuster le vecteur $\underline{y}$ par une combinaison linéaire $X_1 \underline{b}$ des m variables "explicatives".

Si les individus sont munis de poids p_i positifs de somme 1, le critère à minimiser est :

$$\|\underline{y} - X_1 \underline{b}\|^2_{D_p}$$

et la relation est donnée par $\underline{b} = (X_1' D_p X_1)^{-1} X_1' D_p \underline{y}$ si X_1 est de rang m.

Le vecteur $X\underline{b}$ est noté y^* et on a $y^* = P_1 \underline{y}$ où P_1 est le projecteur D_p-orthogonal associé à X_1.

$\underline{y}^*$ est la projection D_p-orthogonale de $\underline{y}$ sur W_1.

On peut remarquer que la régression linéaire, considérée sous cet aspect algébrique ou géométrique, est un cas particulier de la méthode factorielle du chapitre précédent où $p = 2$ $m_1 = 1$ $m_2 = m$, dans laquelle on impose que la composante canonique relative à $\underline{y}$ soit égale à $\underline{y}$.

Lorsque les données ne sont pas centrées on sait qu'il suffit d'introduire une variable supplémentaire dans X_1 représentée par le vecteur colonne $\underline{1}$ pour obtenir l'analyse usuelle centrée.

Pour la suite de cette partie nous supposerons que $D_p = \frac{1}{n} I_n$ ce qui donne les formules usuelles :

$$\underline{b} = (X_1' X_1)^{-1} X_1' \underline{y} \qquad \underline{y}^* = X_1 (X_1' X_1)^{-1} X_1' \underline{y}$$

b) Le modèle d'analyse de variance.

Si nous supposons que X_1 est un tableau logique associé à une variable qualitative, nous sommes dans le cadre de l'analyse de variance à un facteur : en effet, expliquer $\underline{y}$ au moyen des m indicatrices des modalités revient à évaluer l'influence de chaque modalité sur le caractère $\underline{y}$.

Si le tableau de données se présente sous la forme :

$$\left(\underline{y} \;\|\; X_1 \;|\; X_2 \;|\; \ldots \;|\; X_p \right) = \left(\underline{y} \;\|\; X \right)$$

où les X_i sont les tableaux logiques associés à p variables qualitatives, la recherche de $\underline{y}^*$ rentre alors dans le cadre du modèle d'analyse de variance à p facteurs sans interaction (modèle additif).

Dans ces conditions la matrice X n'est plus de plein rang sauf dans le cas d'une seule variable qualitative, où on trouve, si $\underline{y}$ est centrée, $b_i = \bar{y}_i$

où $\bar{y}_i$ représente la moyenne des valeurs de la variable à expliquer pour la modalité n° i .

c) Le modèle d'analyse de covariance.

C'est un modèle mixte où X se compose d'une part d'un bloc X_1 de caractères quantitatifs et d'autre part de (p - 1) blocs X_i (i = 2,..., p) représentant des variables qualitatives.

Nous nous préoccuperons uniquement du cas b) en l'envisageant exclusivement sous son aspect géométrique, c'est-à-dire dans l'optique de l'analyse des données, sans entrer dans le domaine des tests qui appartient à la théorie classique de l'analyse de variance. Le modèle mixte c) ne sera pas abordé car d'une part sa résolution ne présente aucune difficulté une fois connue celle de b) , et dans l'optique de prévision qui est la nôtre, d'autre part, il peut y avoir intérêt à rendre qualitatives toutes les variables explicatives ce qui permet alors de trouver des formules d'approximation non linéaires.

Envisagée sous l'aspect de l'analyse des données, l'analyse de variance, est alors une technique de codage de variables qualitatives à but prévisionnel.

Cependant le fait que X soit de rang ou plus égal à $\Sigma m_i - p + 1$ introduit un élément nouveau car il y a alors dans une certaine mesure indétermination du codage des variables qui est fourni par le vecteur $\underline{b}$ des coefficients de régression.

2. Quelques techniques classiques de résolution du modèle linéaire de rang $< m$.

L'équation normale $X'X\underline{b} = X'\underline{y}$ a une infinité de solutions fournissant le même minimum de $\| \underline{y} - X\underline{b} \|^2$ lorsque $X'X$ n'est pas inversible.

Cependant on a le théorème suivant [20], [23] :

Théorème.

Toute solution $\underline{b}$ de $X'X\underline{b} = X'\underline{y}$ fournit le même vecteur $\underline{y}^* = X\underline{b}$.

Si il y a indétermination pour les coefficients de régression, il y a donc unicité de la prévision, ce qui est le plus important la plupart du temps.

Toutes les méthodes d'obtention d'une solution des équations normales sont alors équivalentes de ce point de vue, ce qui équivaut à choisir un inverse généralisé $(X'X)^-$ quelconque de $X'X$.

Citons parmi les diverses méthodes d'obtention d'un inverse généralisé de $X'X$ [20] :

- la méthode d'augmentation.

Si le rang de X est égal à $r < m$ on cherche une matrice H à $m-r$ lignes et r colonnes telles que :

$$\text{rang} \left(\frac{X}{H} \right) = m$$

Alors $X'X + H'H$ est inversible et on montre aisément que $(X'X + H'H)^{-1}$ est un inverse généralisé de $X'X$.

Une variante consiste à inverser $\left(\begin{array}{c|c} X'X & H' \\ \hline H & 0 \end{array} \right)$ qui est alors de plein rang et si $\left(\begin{array}{c|c} X'X & H' \\ \hline H & 0 \end{array} \right)^{-1} = \left(\begin{array}{c|c} C_1 & C_2' \\ \hline C_2 & C_3 \end{array} \right)$.

On vérifie que C_1 est aussi un inverse généralisé de $X'X$.

- méthode d'élimination.

Cette méthode qui s'adapte bien au cas des variables qualitatives consiste à isoler les colonnes de X non indépendantes des autres, ici une colonne quelconque par sous-tableau X_i, de telle sorte que leur élimination rende X de plein rang. On inverse la matrice $X_1' X_1$ obtenue par cette élimination et on augmente cette matrice jusqu'à obtenir une matrice de taille $m \times m$ en incorporant des lignes et les colonnes de zéro aux emplacements des variables précédemment supprimées. La matrice ainsi obtenue est un inverse généralisé de X'X.

- méthode de contraintes.

Analogue à la précédente cette méthode consiste à imposer $m-r$ contraintes linéaires indépendantes aux composantes du vecteur de régression $\underline{b}$. Un des jeux de contraintes les plus couramment utilisés est le suivant :

On exige que la somme des coefficients soit nulle pour chaque variable qualitative.

Ainsi pour X_1 on doit avoir $b_1 + b_2 + \ldots + b_{m_1} = 0$

donc : $X_1 \underline{b}_1 = \underline{x}_1 b_1 + \underline{x}_2 \underline{b}_2 + \ldots + \underline{x}_{m_1} b_{m_1}$

$$= \underline{x}_1 b_1 + \underline{x}_2 b_2 + \ldots - \underline{x}_{m_1} \left(\sum_{i=1}^{m_i - 1} b_i \right)$$

$$= \sum_{i=1}^{m_1 - 1} b_i \, (\underline{x}_i - \underline{x}_{m_1})$$

Tout se passe donc comme si on supprimait une des variables, $\underline{x}_{m_1}$ par exemple et que l'on effectue une régression sur les variables obtenues par différence avec $\underline{x}_{m_1}$, ceci pour chaque tableau X_i.

Les techniques de contraintes sont théoriquement intéressantes, mais on peut juger arbitraire d'exiger une somme des coefficients nulles et il nous semble plus logique d'imposer des contraintes de centrage :

$X_i \underline{b}_i$ doit être une variable centrée $\quad \forall i$.

ce qui s'accorde avec les méthodes de codage exposées au chapitre I.

Par ailleurs, plutôt qu'un inverse généralisé quelconque, on peut souhaiter utiliser celui de MOORE - PENROSE qui présente les propriétés les plus intéressantes.

Nous montrerons au paragraphe suivant que les contraintes de centrage sont équivalentes, par construction, au choix de l'inverse de MOORE - PENROSE.

3. Une technique nouvelle : la régression sur facteurs $\underline{z}$.

Le problème traité est celui de la régression sur p variables qualitatives.

Par analogie avec la régression sur composantes principales, utilisée dans le cas ordinaire a), lorsque les variables explicatives sont quasi-colinéaires, nous proposons de remplacer le tableau X des variables explicatives par le tableau Z des facteurs que l'on peut extraire en effectuant l'analyse conjointe des p variables qualitatives ; Z sera alors une matrice à n lignes et $(m - p)$ colonnes puisqu'il existe $(m - p)$ facteurs non triviaux. Les diverses colonnes de Z sont centrées. Si nous choisissons des facteurs normés à 1 la matrice Z est alors orthogonale puisque les facteurs sont orthogonaux deux à

deux. Nous supposerons que la variable $\underline{y}$ est centrée.

Si on effectue la régression de $\underline{y}$ sur $\underline{z}_1, \underline{z}_2, \ldots, \underline{z}_{m-p}$ on trouve en minimisant $\|\underline{y} - Z\,\underline{c}\|^2$ que :

$$\boxed{\underline{c} = Z'\,\underline{y} \qquad \underline{y}^* = Z\,Z'\,\underline{y}}$$

car la solution classique de la régression $\underline{c} = (Z'Z)^{-1}\,Z'\,\underline{y}$ se simplifie puisque Z étant une matrice orthogonale $Z'Z = I$.

A chaque facteur $\underline{z}_j$ correspond un vecteur $\underline{b}^j$, $j^{\text{ième}}$ codage descriptif des variables qualitatives, qui est vecteur propre de $V^{-1}H$ associé à la valeur propre λ_j.

Si $\underline{z}_j$ est de norme unité $\underline{b}^j$ doit être tel que :

$$\underline{b}^{j\prime}\,X'X\,\underline{b}^j = 1 \qquad \text{avec} \qquad \underline{z}_j = X\,\underline{b}^j$$

$\underline{b}^j$ pouvant s'écrire $\underline{b}^j = \begin{pmatrix} \underline{b}_1^j \\ \underline{b}_2^j \\ \vdots \\ \underline{b}_p^j \end{pmatrix}$

on a donc $\underline{y}^* = \sum_{j=1}^{m-p} c_j\,\underline{z}_j = \sum_{j=1}^{m-p} c_j \sum_{i=1}^{p} X_i\,\underline{b}_i^j = \sum_{j=1}^{m-p} c_j\,X\,\underline{b}^j$

Or, nous cherchons une formule de régression de la forme :

$\underline{y}^* = X\,\underline{a}$ où $\underline{a}$ est solution de l'équation normale $X'X\,\underline{a} = X'\,\underline{y}$

$\underline{a}$ est le vecteur des codages explicatifs des variables qualitatives.

$$\underline{a} = \begin{pmatrix} \underline{a}_1 \\ \underline{a}_2 \\ \vdots \\ \underline{a}_p \end{pmatrix} \qquad \underline{y}^* = \sum_{i=1}^{p} X_i\,\underline{a}_i$$

il suffit alors de comparer les deux formules :

$$\underline{y}^* = \sum_{j=1}^{n-p} c_j\,X\,\underline{b}^j \quad \text{et} \quad \underline{y}^* = X\,\underline{a} \quad \text{pour identifier} \quad \underline{a}.$$

On trouve immédiatement $\underline{a} = \sum_{j=1}^{m-p} c_j \underline{b}^j$

soit encore

$$\boxed{\underline{a}_i = \sum_{j=1}^{m-p} c_j \underline{b}_i^j}$$

Les codages explicatifs sont les combinaisons linéaires des codages descriptifs avec pour coefficients, les coefficients de régression de $\underline{y}$ sur les facteurs $\underline{z}$ correspondant : ils sont donc centrés. On peut en fait se dispenser d'effectuer la régression de $\underline{y}$ sur Z :

Comme $c_j = \underline{z}_j' \underline{y}$ puisque $\underline{c} = Z' \underline{y}$ on peut aussi écrire :

$$\underline{a}_i = \sum_{j=1}^{m-p} (\underline{z}_j' \underline{y}) \underline{b}_i^j \quad \text{soit} \quad \underline{a} = \sum_{j=1}^{m-p} (\underline{z}_j' \underline{y}) \underline{b}^j$$

et comme $\underline{z}_j = X \underline{b}^j$ il vient :

$$\underline{a} = \sum_{j=1}^{m-p} (\underline{z}_j' \underline{y}) \underline{b}^j = \sum_{j=1}^{m-p} (X \underline{b}^j)' \underline{y} \underline{b}^j = \sum_{j=1}^{m-p} \underline{b}^j (X \underline{b}^j)' \underline{y}$$

car $(X \underline{b}^j)' \underline{y}$ est un scalaire.

d'où :

$$\boxed{\underline{a} = \left(\sum_{j=1}^{m-p} \underline{b}^j \underline{b}^{j'} \right) X' \underline{y}}$$

Or les $\underline{b}^j$ sont vecteurs propres de $V^{-1} X'X$ et sont tels que : $\underline{b}^{j'} X'X \underline{b}^j = 1$. Ils constituent donc un système orthonormé maximal pour la métrique de matrice semi-définie positive $X'X$. Dans ces conditions la matrice $\sum_{j=1}^{m-p} \underline{b}^j \underline{b}^{j'}$ est égale à l'inverse de MOORE - PENROSE de $X'X$ selon un résultat classique généralisant une propriété bien connue des bases M-orthonormées où M est une matrice définie positive d'ordre m :

Si $\underline{u}_1, \underline{u}_2, \ldots, \underline{u}_m$ sont tels que $\underline{u}_i M \underline{u}_j = \delta_{ij}$ alors $\sum_{i=1}^{m} \underline{u}_i \underline{u}_i' = M^{-1}$;

(il suffit de vérifier que $\sum_{i=1}^{m} \underline{u}_i \underline{u}_i' M \underline{u}_j = \delta_{ij} \underline{u}_j$ donc que $\sum_{i=1}^{m} \underline{u}_i \underline{u}_i' M = I$)

Ceci prouve que le vecteur $\underline{a}$ obtenu est bien solution des équations normales et que, puisque les codages obtenus sont centrés, choisir les contraintes

de centrage équivaut à choisir l'inverse généralisé de MOORE - PENROSE, comme nous l'avions annoncé.

Pour résumer, notre méthode revient à :

1) Diagonaliser la matrice $V^{-1} H$

2) Normer les vecteurs propres non triviaux de telle sorte que : $\underline{b}^{j'} X'X \underline{b}^{j} = 1$.

3) Former la matrice $(X'X)^{+} = \sum_{j=1}^{m-p} \underline{b}^{j} \underline{b}^{j'}$

4) Les codages explicatifs centrés sont alors donné par $\underline{a} = (X'X)^{+} X' \underline{y}$ et la prévision $\underline{y}^{*} = X \underline{a}$.

Remarques :

1) Cette méthode n'est évidemment pas la seule à fournir l'inverse de MOORE - PENROSE de $X'X$ puisque n'importe quel système de vecteurs orthonormés pour $X'X$ aboutit au même résultat, en particulier celui des vecteurs propres de $X'X$. Cependant comme nous l'avons exposé précédemment, les vecteurs propres de $V^{-1} X'X$ ont une interprétation statistique que n'ont pas ceux de $X'X$ puisque ce sont eux qui fournissent la meilleure description des individus au sens de la somme des carrés des rapports de corrélation.

Par ailleurs dans un tel type d'analyse il y a toujours intérêt à procéder d'abord à une description des relations entre variables qualitatives, par analyse factorielle, avant de procéder à des calculs de régression.

2) On pourrait songer, afin d'allèger les calculs, à ne retenir que quelques facteurs $\underline{z}$ et non tous les $(m - p)$ possibles ce qui revient par exemple si on ne retient que $\underline{z}_1$ et $\underline{z}_2$ à ajuster $\underline{y}$ par projection sur le plan défini par $\underline{z}_1$, $\underline{z}_2$.

Cette pratique nous semble dangereuse car, d'une part la formule de régression ne pourra évidemment qu'être approximative et ne réalisera pas la solution des moindres carrés, mais surtout d'autre part il se peut que les derniers facteurs que l'on a négligé soient très corrélés avec la variable à expliquer.

4. Un algorithme spécial pour les grands tableaux : la méthode itérative de S. WOLD [26].

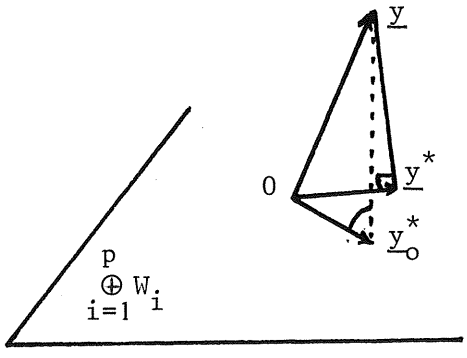
La méthode proposée précédemment implique la diagonalisation d'une matrice d'ordre m, ce qui risque de poser des problèmes pratiques pour les grandes valeurs de m.

Certes les méthodes classiques d'élimination ou de contraintes de somme nulle ramènent le problème à celui de l'inversion d'une matrice de taille $m-p$, ce qui est quelquefois avantageux, mais tout dépend des valeurs respectives de m et p.

L'algorithme suivant, dû à S. WOLD, permet d'atteindre la solution à codages centrés de la régression sur p variables qualitatives, par une méthode de point fixe ne nécessitant qu'un faible encombrement de mémoire sur un calculateur automatique, en contrepartie d'un grand nombre d'opérations élémentaires.

Le principe consiste à partir d'un codage centré réduit quelconque de chaque variable qualitative ; on remplace donc le tableau X à n lignes et m colonnes par le tableau Z_o à n lignes et p colonnes des codages des individus ainsi réalisés.

On effectue alors la régression de $\underline{y}$ sur les p variables codées ce qui fournit $\underline{y}_o^*$, en même temps que les codages centrés explicatifs des variables qui sont alors égaux aux codages arbitraires de départ multipliés par le coefficient de régression correspondant. Soit $\underline{a}_o$ le vecteur des codages ainsi obtenus. Si on considère alors le résidu $\underline{y} - \underline{y}_o^*$, la théorie usuelle de la régression nous indique qu'une solution des équations normales est atteinte si $\underline{y} - \underline{y}_o$ est orthogonal à l'espace engendré par toutes les colonnes de X, c'est-à-dire à $\bigoplus_{i=1}^{p} W_i$, donc si $\underline{y} - \underline{y}_o^*$ est orthogonal à toutes les colonnes de X.



S'il n'en est pas ainsi on modifie alors les codages pour se rapprocher de l'orthogonalité en procédant par des projections successives sur chacun des W_i.

Posons $\underline{e}_o = \underline{y} - \underline{y}_o^*$. Soit $P_1 \underline{e}_o$ la projection du résidu sur W_1, alors :

$\underline{y}_1^* = \underline{y}_o^* + P_1 \underline{e}_o$ est une meilleure approximation de $\underline{y}$ que $\underline{y}_o^*$ au sens des moindres carrés ; en effet, on a :

$$\|\underline{y} - \underline{y}_o^*\|^2 = \|\underline{y} - \underline{y}_1^* + P_1 \underline{e}_o\|^2 = \|\underline{y} - \underline{y}_1^*\|^2 + \|P_1 \underline{e}_o\|^2 + 2(\underline{y} - \underline{y}_1^*)' P_1 \underline{e}_o$$

et comme $\underline{y} - \underline{y}_1^*$ est orthogonal à $P_1 \underline{e}_o$ car il est égal à $\underline{y} - \underline{y}_o^* - P_1 \underline{e}_o$ on en déduit que :

$$\|\underline{y} - \underline{y}_1^*\|^2 \leqslant \|\underline{y} - \underline{y}_o^*\|^2 .$$

l'égalité ne s'obtenant que si le résidu $\underline{e}_o$ était orthogonal à W_1. La nouvelle approximation $\underline{y}_1^*$ est telle que le résidu $\underline{e}_1 = \underline{y} - \underline{y}_1^*$ est orthogonal à W_1 par construction. On améliore alors l'approximation de $\underline{y}$ en considérant W_2 : on projette $\underline{y} - \underline{y}_1^*$ sur W_2 et alors $\underline{y}_2^* = \underline{y}_1^* + P_2 \underline{e}_1$ est une meilleure approximation de $\underline{y}$ que $\underline{y}_1^*$ et ainsi de suite jusqu'à la projection sur W_p

$\underline{y}_p^* = \underline{y}_{p-1}^* + P_p \underline{e}_{p-1}$ ce qui achève la première itération.

Pendant tout ce processus on a modifié séparément chacun des p codages : ainsi passer de $\underline{y}_o^*$ à $\underline{y}_1^*$ revient à modifier exclusivement le codage de la première variable qualitative de la manière suivante : $P_1 \underline{e}_o$ n'est autre que la somme des projections de $\underline{e}_o$ sur chaque indicatrice $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_{m_1}$ car celles-ci sont orthogonales entre elles. Si $r_{o_1}, r_{o_2}, \ldots, r_{o_{m_1}}$ désignent les produits scalaires de $\underline{e}_o$ avec les $\underline{x}_i$ on a :

$$P_1 \underline{e}_o = \sum_{k=1}^{m_1} r_{ok} \underline{x}_k$$

ceci correspond à modifier les codages de départ a_{o_k} $(k = 1, 2, \ldots, m_1)$ en ajoutant à a_{ok} la quantité $\frac{r_{ok}}{n_k}$ où n_k est l'effectif de la $k^{ième}$ modalité.

Cette modification ne change pas le fait que les codages sont centrés puisque $\underline{e}_o$ étant centré $P_1 \underline{e}_o$ l'est aussi.

Après avoir effectué successivement les p projections, on recommence en projetant $\underline{e}_p$ sur W_1 etc. car le résidu n'est plus orthogonal en général aux $W_1, W_2, \ldots, W_{p-1}$.

L'algorithme converge car il y a décroissance stricte de la norme du résidu à chaque itération; lorsque le vecteur des résidus ne peut plus être modifié c'est qu'il est orthogonal à tous les W_i donc que la solution des moindres carrés est atteinte.

En pratique on pourra se contenter d'une solution approchée si tous les produits scalaires du résidu avec les diverses indicatrices sont assez petits.

5 . Remarques diverses.

a) Qualité de l'approximation obtenue.

Ainsi qu'en régression ordinaire le coefficient de corrélation multiple R défini par $R^2 = \frac{\|\underline{y}^*\|^2}{\|\underline{y}\|^2}$ mesure la qualité de l'ajustement obtenu.

C'est le carré du cosinus de l'angle formé par $\underline{y}$ et $\bigoplus_{i=1}^{p} W_i$.

En l'absence d'étude précise sur la distribution d'échantillonnage de R^2 pour le cas de variables qualitatives l'utilisation du test usuel de FISHER-SNEDECOR nous semble hasardeuse et ne peut être recommandée.

Cependant si on peut admettre que la variable à expliquer a une distribution gaussienne conditionnellement à chaque modalité, on peut tester l'influence marginale de chaque variable qualitative par le test usuel du rapport de corrélation.

Si η^2_{Y/X_i} est le rapport de corrélation de Y avec la $i^{ème}$ variable qualitative on sait qu'en cas d'indépendance le rapport $\frac{\eta^2}{1-\eta^2} \quad \frac{n-m_i}{m_i-1}$ suit une loi $F\ (m_i - 1\ ;\ n - m_i)$.

b) Méthodes progressives.

Il est parfaitement possible d'adapter à la régression sur variables qualitatives les méthodes de pas à pas usuelles : on peut par exemple sélectionner à chaque pas la variable à adjoindre aux précédentes qui provoque le plus fort accroissement de R^2 (on pourrait qualifier cette méthode d'analyse de variance pas à pas). Cependant, là aussi, l'absence de tests statistiques constitue un

obstacle dans l'établissement de critères d'arrêt.

c) Le cas des interactions.

Nous nous sommes bornés au modèle purement additif sans interaction. Il est toutefois possible d'introduire des interactions dans le modèle en ajoutant de nouvelles variables dans la matrice X qui s'obtiennent en considérant les produits des indicatrices ; ainsi pour 2 variables à respectivement 2 et 3 modalités, il faudrait introduire 6 variables supplémentaires soit au total 11 variables.

La multiplication des dimensions de la matrice X devient vite prohibitive et il est en fait impossible de prendre en compte toutes les interactions dès que p dépasse quelques unités.

III. L'ANALYSE DISCRIMINANTE SUR VARIABLES QUALITATIVES.

Les observations se répartissant en un certain nombre de catégories définies à l'avance, l'analyse discriminante ordinaire se propose, en tenant compte des valeurs prises par des variables numériques, de prévoir au mieux l'appartenance des observations à ces catégories.

Il s'agit donc en quelque sorte d'une régression où la variable à expliquer est qualitative. De nombreuses publications traitent de ce problème, voir par exemple [21], qui peut être abordé par la théorie classique de la décision en faisant des hypothèses de normalité, ou par l'analyse factorielle. C'est cette dernière optique que nous retiendrons.

Le cas de variables explicatives qualitatives a été peu traité jusqu'aux travaux récents de M. MASSON [17] qui a mis au point une méthode séquentielle sur laquelle nous reviendrons.

Après avoir rappelé brièvement quelques propriétés utiles pour la suite, de l'analyse factorielle discriminante classique, nous proposerons une technique dérivée de celle développée pour la régression : la discrimination sur facteurs $\underline{z}$; puis nous étudierons le cas particulier où les variables explicatives sont indépendantes deux à deux, ainsi que les propriétés de la méthode non-optimale, de discrimination qu'on peut en déduire lorsque les variables explicatives ne sont pas indépendantes. Une mention spéciale sera faite dans le

le cas pratique important d'une discrimination entre deux catégories qui conduit à des simplifications notables.

1. Rappels d'analyse factorielle discriminante ordinaire.

n individus affectés de masses p_i positives de somme 1 sont répartis en m classes disjointes d'effectifs $n_1, n_2, \ldots, n_m$ ce qui définit ainsi une variable qualitative à m modalités.

Ils sont décrits d'autre part, par p variables numériques centrées. Le tableau de données a donc la forme suivante où A désigne le tableau logique associé à la partition et X le tableau des mesures numériques.

$$\begin{array}{c} \\ 1 \\ 2 \\ \vdots \\ \vdots \\ n \end{array} \left(\begin{array}{c|c} 1\ 2\ \ldots\ m & 1\ 2\ \ldots\ p \\ \\ A & X \\ \\ \\ \end{array} \right)$$

On considère alors les m centres de gravité des classes qui sont des points affectés de masses définies par les éléments de la matrice diagonale $A'D_p A$. Le tableau à m lignes et p colonnes des coordonnées de ces centres de gravité est alors égal à $(A'D_p A)^{-1} (A'D_p X)$.

La matrice des variances - covariances de ces m centres vaut donc :

$$((A'D_p A)^{-1} (A'D_p X))' A'D_p A \quad ((A'D_p A)^{-1} (A'D_p X)) \quad \text{soit :}$$

$$(X'D_p A) (A'D_p A)^{-1} (A'D_p X) = B$$

B s'appelle la matrice d'inertie inter-classes.

L'analyse factorielle discriminante se propose de définir de nouvelles variables, les facteurs discriminants, obtenus par combinaisons linéaires des p variables explicatives de telle sorte que les projections sur les axes de $\mathbb{R}^p$ ainsi définis des divers centres de gravité soient le mieux séparées possibles dans la métrique de $\mathbb{R}^p$ définies par H^{-1}. Si $\underline{u}$ représente le vecteur de $\mathbb{R}^p$ définissant la combinaison linéaire discriminante ce critère revient à maximiser le rapport $\dfrac{\underline{u}'B\,\underline{u}}{\underline{u}'H\,\underline{u}}$ où $H = X'D_p X$.

$\underline{u}$ est donc vecteur propre de $H^{-1} B$ associé à sa plus grande valeur propre qui est appelée pouvoir discriminant du premier axe factoriel. Les vecteurs propres suivants définissent les axes discriminants suivants qui sont D_p-orthogonaux entre eux et sont au nombre de $m-1$ si $m < p$.

On voit aisément que l'analyse factorielle discriminante n'est autre que l'analyse canonique des deux tableaux A et X puisque l'équation fournissant les variables canoniques de X s'écrit d'après le chapitre II :

$$(X'D_p X)^{-1} (X'D_p A) (A'D_p A)^{-1} (X'D_p A) \underline{u} = \lambda \underline{u}$$

ce qui est identique à $H^{-1} B \underline{u} = \lambda \underline{u}$.

Si on désigne par $A\underline{a}$ la variable canonique associée à A solution de :

$$(A'D_p A)^{-1} A'D_p X (X'D_p X)^{-1} X'D_p A \underline{a} = \lambda \underline{a}$$

normée de telle façon que :

$$P_x A \underline{a} = X \underline{u}$$

on peut dire alors que $\underline{a}$ définit le codage des modalités de la variable qualitative tel que $\underline{u}$ est le vecteur des coefficients de régression de $A\underline{a}$ par les p variables explicatives.

Le pouvoir discriminant λ est alors égal au carré du coefficient de corrélation linéaire entre le codage des individus $A\underline{a}$ résultant de la variable qualitative et le facteur discriminant $X\underline{u}$.

Dans le cas particulier d'une discrimination entre deux groupes, il n'existe qu'un seul facteur discriminant.

Si on désigne par μ_1 et μ_2 les centres de gravité des deux classes, représentés par des éléments de $\mathbb{R}^p$, la forme linéaire discriminante $\underline{u}$ est donnée par :

$$\underline{u} = H^{-1}(\underline{\mu}_1 - \underline{\mu}_2)$$

c'est la "fonction linéaire discriminante de FISHER". Si $p_i = \frac{1}{n}$ le pouvoir discriminant est $\frac{n_1 n_2}{n^2} (\underline{\mu}_1 - \underline{\mu}_2)' H^{-1} (\underline{\mu}_1 - \underline{\mu}_2)$ où n_1 et n_2 sont les effectifs des deux groupes, c'est-à-dire le D^2 de MAHALANOBIS au coefficient $\frac{n_1 n_2}{n^2}$ près.

2. La discrimination sur variables qualitatives au moyen des facteurs $\underline{z}$.

Pour les mêmes raisons qu'en régression, nous ne traiterons pas le cas d'un mélange de variables qualitatives et quantitatives. Si nous disposons de p variables explicatives qualitatives le tableau de données se présente sous la forme suivante :

	$1\ 2\ \ldots\ k$	$1\ 2\ \ldots\ m_1$	m_1+m_2	$\ldots\ldots$ m
1				
2				
.	A	X_1	X_2	. . . X_p
.				
.				
n				

où A et les X_i sont les tableaux logiques habituels.

Le problème est de déterminer $\underline{a}$ et $\underline{u} = \begin{pmatrix} \underline{u}_1 \\ \underline{u}_2 \\ \vdots \\ \underline{u}_p \end{pmatrix}$ réalisant le minimum de

$\| A\,\underline{a} - \sum_{i=1}^{p} X_i\,\underline{u}_i \|^2_{D_p}$ sous une condition de normalisation qui peut être soit

$\| A\,\underline{a} \|_{D_p} = 1$, soit $\| X\,\underline{u} \|_{D_p} = 1$.

Les X_i étant des tableaux logiques on ne peut inverser la matrice $X'D_pX$ où $X = (X_1 | X_2 | \ldots | X_p)$ qui est de rang $m-p+1$ (de rang utile $m-p$ si on se borne comme à l'ordinaire aux codages centrés).

Utilisant à nouveau la méthode proposée pour la régression, nous proposons de remplacer X par le tableau Z des $(m-p)$ facteurs de l'analyse conjointe des p variables qualitatives et d'effectuer alors une analyse discriminante ordinaire puisque les observations sont alors décrites par $m-p$ caractères quantitatifs.

Ceci revient à effectuer l'analyse canonique de A et Z . Si nous notons $\underline{c}$ le vecteur à $(m-p)$ composantes définissant la combinaison linéaire discriminante des $\underline{z}_j$ on a :

$$(Z'D_pZ)^{-1}\,(Z'D_pA)\,(A'D_pA)^{-1}\,(A'D_pZ)\,\underline{c} = \lambda\,\underline{c}$$

Or les $\underline{z}_j$ formant un système D_p-orthonormé si on a pris le soin de D_p-normer à 1 tous les facteurs, la matrice $Z'D_p Z$ est égale à I_{m-p} et l'équation devient :

$$(Z'D_p A)\ (A'D_p A)^{-1}\ (A'D_p Z)\ \underline{c} = \lambda\ \underline{c}$$

c'est-à-dire que $\underline{c}$ est vecteur propre de la matrice d'inertie inter-classe B_z des centres de gravité dans le système des axes définis par les facteurs $\underline{z}$

$$\boxed{B_z\ \underline{c} = \lambda\ \underline{c}}$$

Si $\underline{c}$ est le vecteur propre de B_z associé à sa plus grande valeur propre, la combinaison linéaire discriminante des facteurs $\underline{z}_j$ est donc $\sum_{j=1}^{m-p} c_j\ \underline{z}_j$.

Il ne reste plus qu'à identifier cette combinaison avec $\sum_{i=1}^{p} X_i\ \underline{u}_i$ pour obtenir les codages discriminants du premier ordre :

Si les $\underline{b}_i^j$ désignent les codages descriptifs de toutes les variables qualitatives des divers ordres on a donc :

$$\underline{z}_j = \sum_{i=1}^{p} X_i\ \underline{b}_i^j$$

d'où : $$\sum_{i=1}^{p} X_i\ \underline{u}_i = \sum_{j=1}^{m-p} c_j \sum_{i=1}^{p} X_i\ \underline{b}_i^j$$

ce qui donne :

$$\boxed{\underline{u}_i = \sum_{j=1}^{m-p} c_j\ \underline{b}_i^j}$$

ou globalement :

$$\underline{u} = \sum_{j=1}^{m-p} c_j\ \underline{b}^j$$

Le résultat est identique à celui obtenu pour la régression : les codages discriminants s'obtiennent par combinaison linéaire des codages descriptifs avec pour coefficients ceux de la fonction discriminante des codages descriptifs.

On obtient les codages discriminants d'ordre supérieur en considérant les

autres vecteurs propres de la matrice B_z .

Avec les deux premiers codages discriminants on peut faire une représentation graphique plane des individus selon le meilleur plan discriminant et représenter les proximités entre les modalités des variables explicatives et celles de la variable à expliquer au moyen du cercle des corrélations avec $\underline{z}_1$ et $\underline{z}_2$.

Cas particulier de la discrimination entre deux groupes.

On sait qu'il n'y a alors qu'une seule forme linéaire discriminante qui est fournie par la fonction de FISHER $H^{-1}\ (\underline{\mu}_1 - \underline{\mu}_2)$.

Or ici, puisque nous effectuons une discrimination sur les facteurs $\underline{z}$ la matrice $H = Z'D_p Z$ est égale à la matrice unité.

La forme linéaire discriminante est donc simplement donnée par :

$$\boxed{\underline{c} = \underline{\mu}_1 - \underline{\mu}_2}$$

c'est-à-dire la différence des moyennes des codages descriptifs de chacun des groupes. On voit donc l'intérêt de remplacer le tableau de données X par celui des facteurs $\underline{z}$.

Le codage discriminant $\underline{u}$ vaut alors :

$$\boxed{\underline{u}_i = \sum_{j=1}^{m-p} (\mu_1^j - \mu_2^j)\ \underline{b}_i^j}$$

où μ_1^j et μ_2^j représentent respectivement les coordonnées des centres de gravité des deux groupes sur le $j^{ème}$ axe factoriel descriptif.

3. Le cas de variables explicatives indépendantes et la méthode des "moindres moyens carrés".

D'importantes simplifications surviennent si les variables explicatives sont indépendantes ; étudions tout d'abord le cas de l'analyse discriminante ordinaire lorsqu'il y a non corrélation par blocs.

a) L'analyse discriminante ordinaire lorsque les variables explicatives

se partitionnent en p blocs orthogonaux deux à deux.

Supposons que le tableau de données puisse se mettre sous la forme :

$$(A \mid X) = (A \parallel X_1 \mid X_2 \mid \ldots \mid X_p)$$

avec $V_{ij} = 0$ si $i \neq j$. Les variables explicatives sont supposées centrées comme toujours.

Dans ce cas le projecteur P sur $\bigoplus_{i=1}^{p} W_i$ n'est autre que la somme des projecteurs sur chacun des W_i et :

$$P = X(X'D_p X)^{-1} X' = X V^{-1} X' = \sum_{i=1}^{p} P_i \quad \text{car} \quad X'D_p X = V$$

si $\underline{u}$ et $\underline{a}$ désignent comme au paragraphe 1) les vecteurs réalisant le minimum de $\| A\,\underline{a} - X\,\underline{u}\|^2_{D_p}$ sous une condition de normalisation pour $\underline{a}$ on a donc :

$$(A'D_p A)^{-1} A'D_p X V^{-1} X' D_p A\,\underline{a} = \lambda\,\underline{a}$$

en multipliant par A les deux membres on voit donc que $A\,\underline{a}$ est vecteur propre de $P_A \left(\sum_{i=1}^{p} P_i \right)$:

$$\sum_{i=1}^{p} P_A P_i A\,\underline{a} = \lambda A\,\underline{a}$$

Quant à $\underline{u}$ on a de même :

$$V^{-1} X'D_p A (A'D_p A)^{-1} A'D_p X\,\underline{u} = \lambda\,\underline{u}$$

d'où : $\sum_{i=1}^{p} P_i P_A X\,\underline{u} = \lambda\, X\,\underline{u}$

de plus si on écrit $X\,\underline{u} = X_1\,\underline{u}_1 + X_2\,\underline{u}_2 + \ldots + X_p\,\underline{u}_p$ on trouve :

$$X_i\,\underline{u}_i = P_i A\,\underline{a}$$

b) Discrimination sur variables qualitatives indépendantes.

Si nous supposons que $\phi^2_{ij} = 0$ pour $i \neq j$ nous sommes alors dans un cas analogue au précédent à ceci près que les variables ne sont pas centrées car ce sont des indicatrices de modalités. Cependant, le fait que l'analyse centrée et l'analyse non centrée conduisent aux mêmes résultats au facteur trivial $\underline{1}$ près,

nous autorise à remplacer $X'D_p X$ par V, (Nous supposerons dans ce paragraphe que $D_p = \frac{1}{n} I_n$ pour simplifier) alors que $X_i' D_p X_j$ ne peut être nul.

En fait pour des variables de ϕ^2 nul on trouve pour $X_i' D_p X_j$ le tableau des produits des probabilités marginales.

L'équation $(A'D_p A)^{-1} A'D_p X V^{-1} X'D_p A \underline{a} = \lambda \underline{a}$ s'écrit alors :

$$(A'A)^{-1} A'X V^{-1} X' A \underline{a} = \lambda \underline{a}$$

Où V^{-1} représente la matrice diagonale des inverses des effectifs de toutes les modalités des p variables. $(A'A)^{-1}$ la matrice des inverses des effectifs des classes à discriminer.

Suivant en cela M. MASSON [17] on reconnaît l'équation d'une analyse canonique formelle sur le tableau $A'X$ (voir chapitre I).

Or le tableau $A'X$ n'est autre que la juxtaposition des p tableaux de contingences que l'on peut former en croisant la variable à expliquer avec les p variables explicatives :

$$A'X = (A'X_1 \mid A'X_2 \mid \ldots \mid A'X_p)$$

$A'X =$

	1 2	m_1	1 2	m_2	
1					
2	n_{ij}^1		n_{ij}^2		. . . . .
⋮					
k					

L'analyse discriminante revient donc à faire une analyse des correspondances formelles sur cette juxtaposition de tableaux de contingence dans le cas de variables explicatives indépendantes.

Dans le cas très particulier d'une discrimination entre deux groupes comme $\underline{a}$ est déterminé par les contraintes de centrage et de réduction tout se passe comme si les codages des variables explicatives s'obtenaient par analyse canonique séparées, il suffit alors d'appliquer la formule du chapitre I § I. 3. pour chaque variable explicative : les codages descriptifs sont alors égaux aux codages discriminants.

c) La méthode des "moindres moyens carrés".

Si on applique les formules précédentes lorsque les variables explicatives ne sont pas indépendantes, on obtient un résultat qui n'est évidemment pas celui fourni par l'analyse discriminante directe mais qui possède certaines propriétés intéressantes. De plus cette technique est moins onéreuse du point de vue calcul.

Proposition.

Le vecteur $\underline{a}$ tel que $\sum_{i=1}^{p} P_A P_i A\underline{a} = \lambda A\underline{a}$ réalise le minimum de :

$$\sum_{i=1}^{p} \| A\underline{a} - X_i \underline{u}_i \|^2 \text{ sous la contrainte } \| A\underline{a} \|^2 = 1 .$$

En effet rechercher à minimiser cette quantité est un problème analogue à celui de la méthode du chapitre II où on recherchait un élément de $\bigoplus_{i=1}^{p} W_i$ le plus proche possible de chaque W_i au sens de la somme des carrés des corrélations multiples ; ici au lieu de chercher cet élément dans $\bigoplus_{i=1}^{p} W_i$ on le cherche dans W_a. Ceci revient au même que de maximiser :

$$\sum_{i=1}^{p} \underline{z}' P_i \underline{z} = \underline{z}' \left(\sum_{i=1}^{p} P_i \right) \underline{z}$$

sous les deux contraintes $\| \underline{z} \|^2 = 1$ et $\underline{z} \in W_a$.

Ecrire $\underline{z} \in W_a$ est équivalent à écrire $P_A \underline{z} = \underline{z}$. Notre problème est alors de maximiser $\underline{z}' P_A \left(\Sigma P_i \right) \underline{z}$ avec $\| \underline{z} \| = 1$, ce qui est réalisé si $\underline{z}$ est vecteur propre de $P_A \left(\Sigma P_i \right)$.

On voit donc que $A\underline{a}$ au lieu de minimiser $\| A\underline{a} - X\underline{u} \|^2$ minimise $\Sigma \| A\underline{a} - X_i \underline{u}_i \|^2$ d'où le nom de méthode des moindres moyens carrés que nous donnons à cette technique.

Le fait que les codages discriminants $\underline{u}_i$ s'obtiennent alors en écrivant $X_i \underline{u}_i = P_i A\underline{a}$ a pour conséquence que si une variable explicative est indépendante ($\phi^2 = 0$) de la variable à expliquer son codage non trivial est identiquement nul : en effet W_a et W_i sont alors orthogonaux au vecteur $\underline{1}$ près.

Cette propriété peut être intéressante dans certains cas car elle peut

permettre d'éliminer des variables parasites et elle n'est pas vérifiée par l'analyse canonique ordinaire qui peut fournir un codage non nul à une variable indépendante du critère à expliquer.

4. Retour sur la méthode séquentielle de M. MASSON [17].

Afin de limiter le nombre de variables explicatives aux seules réellement discriminantes, MASSON a proposé la méthode suivante de pas à pas par adjonction progressive des variables :

On cherche à la première étape, la variable qualitative parmi les p dont le pouvoir discriminant d'ordre 1 est le plus fort : ceci revient à effectuer les p analyses canoniques de A avec X_i et à retenir le tableau X_i tel que la première valeur propre non triviale soit la plus grande possible.

On remplace alors cette variable qualitative par son premier codage. Notons par convention X_1 le tableau retenu et $\underline{\xi}_1 = X_1 \underline{u}_1$ le codage résultant des individus, normé à 1.

A la deuxième étape on cherche parmi les $p-1$ variables qualitatives restantes celle qui associée à $\underline{\xi}_1$ considérée comme une variable numérique, donne la meilleure discrimination selon le même sens que précédemment : ceci revient à effectuer les $p-1$ analyses canoniques de A avec $Y_i = (\underline{\xi}_1 \mid X_i)$ et à retenir le tableau X_i qui fournit la meilleure première valeur propre non triviale. Convenons de noter X_2 ce tableau.

Afin de déterminer le codage discriminant $\underline{u}_2$ associé à cette variable, MASSON propose la contrainte de normalisation suivante : $\underline{u}_2$ est choisi tel que $\begin{pmatrix} 1 \\ \underline{u}_2 \end{pmatrix}$ est le vecteur des codages relatif au tableau Y_i ; en d'autres termes on exige que la combinaison linéaire canonique relative à Y_i ait une première composante égale à 1, ceci dans le but de ne pas modifier le codage obtenu à la première étape. Le codage des individus résultant de la prise en compte des deux variables est alors la somme des codages des individus issus de chaque variable :

$$\underline{\xi}_2 = \underline{\xi}_1 + X_2 \underline{u}_2$$

On continue en cherchant à la troisième étape parmi les $(p-2)$ variables restantes celle qui donne la meilleure discrimination en effectuant les $(p-2)$

analyses canoniques de A contre $Y_i = (\xi_2 \mid X_i)$ et ainsi de suite.

On arrête l'introduction de nouvelles variables lorsque la discrimination obtenue est jugée satisfaisante.

Cette méthode élégante a l'avantage de ne traiter à chaque fois que des tableaux de tailles modestes de nombres de colonnes égaux ou plus à $(\sup m_i) + 1$, au lieu de $\Sigma m_i - p$, ce qui peut être un avantage décisif dans le traitement de fichiers importants.

Cependant deux critiques nous semblent devoir être formulées :

1) Choisir la variable qualitative assurant la plus grande première valeur propre n'est pas forcément synonyme de meilleure discrimination possible si les variables explicatives n'ont pas toutes même nombre de modalités, car les valeurs propres extraites ne sont pas directement comparables.

Ceci est manifeste si il s'agit d'une discrimination entre deux groupes : on sait en effet que l'unique valeur propre est alors égale au ϕ^2 de contingence du croisement de la variable à expliquer avec la variable explicative (voir chapitre I § I.3.) et on ne peut comparer des ϕ^2 de tableaux de dimensions différentes, une grande valeur de ϕ^2 peut ne pas être statistiquement significative si la dimension du tableau est très grande. Il faut alors comparer la valeur de $n\phi^2$ avec la distribution d'un χ^2 dont le nombre de degrés de liberté serait ici $m_i - 1$, pour la variable X_i, car on sait que dans l'hypothèse d'indépendance en probabilité $n\phi^2$ dans un tableau de contingence (p,q) est distribué selon une loi de $\chi^2_{(p-1)(q-1)}$.

Nous proposons de prendre comme mesure du pouvoir discriminant la quantité $P(\chi^2_{m_i-1} < n\phi^2)$ qui permet alors, en se ramenant à une probabilité, de comparer des valeurs différentes du ϕ^2. On choisira alors comme variable entrant au premier pas celle qui maximise la quantité précédente.

Cette modification n'est cependant valable qu'au premier pas et pour une discrimination entre deux groupes ; dans les autres cas la distribution du pouvoir discriminant d'une valeur propre n'étant pas connue dans l'hypothèse d'indépendance.

2) Le codage obtenu par la méthode progressive en introduisant k variables n'est pas celui qu'on aurait obtenu en effectuant l'analyse discriminante directe avec ces k variables qualitatives. En effet, les codages des

variables introduites à l'étape i ne sont pas remis en cause à l'étape i + 1 qui ne fait que coder une nouvelle variable : il n'y a aucune raison par exemple que le codage de la première variable obtenue par l'analyse canonique de A et X_1 seul, soit le même que celui obtenu dans l'analyse complète de A avec $(X_1 \mid X_2 \mid \ldots \mid X_p)$ (sauf si les variables explicatives sont indépendantes statistiquement dans le cas d'une discrimination entre deux groupes).

5. Application pratique.

Nous avons appliqué ces méthodes précédentes à un problème réel de discrimination entre deux catégories de clients d'une société financière (variable C à deux modalités) décrits par 5 variables explicatives V_i possèdant respectivement 8, 4, 2, 6 et 3 modalités, soit au toal 23 variables indicatrices. L'échantillon limité que nous avons étudié comportait 50 individus de chacune des deux catégories à discriminer et se ventilait selon les tableaux de croisement annexés à la fin de ce paragraphe.

a) Discrimination sur facteurs $\underline{z}$.

L'application de la méthode du § 2 conduit au vecteur $\underline{u}$ suivant, des codages discriminants :

$\underline{u}_1$ = - 0,209 ; 1,239 ; 1,757 ; 1,077 ; - 0,483 ; - 0,417 ; - 0,1 ; - 0,697

$\underline{u}_2$ = - 1,221 ; - 1,665 ; - 0,013 ; - 0,629

$\underline{u}_3$ = - 0,262 ; 1,92

$\underline{u}_4$ = 1,499 ; 1,166 ; 0,481 ; 0,692 ; - 1,066 ; - 1,485

$\underline{u}_5$ = - 2,194 ; - 0,887 ; 0,629 .

Les conventions de signes adoptées font que les individus de la catégorie 1 doivent avoir des codages plutôt positifs et ceux de la catégorie 2 des codages plutôt négatifs.

La règle d'affectation qui s'en déduit alors (affecter à la catégorie 1 les individus codés positifs ...) donne alors le tableau de classement suivant :

affectation \ origine	1	2
1	40	9
2	10	41

soit 81 bien classés sur 100 .

Si on convient de limiter à $\alpha = 0,1$ (c'est-à-dire 5 individus) le taux de mal-classés de la catégorie 1, la discrimination s'interprète comme une procédure de test et le taux de mal-classés de la deuxième catégorie est alors $\beta = 0,2$ soit une puissance empirique de $1 - \beta = 0,8$.

b) Méthode des moindres carrés moyens.

Chaque variable est codée par son codage optimal obtenu par croisement séparé avec la variable à expliquer selon la formule du chapitre I § I.3. :

$$\sqrt{\phi^2}\ b_j = \frac{p_{ij}}{p_{.j}}\ a_1 + \frac{p_{2j}}{p_{.j}}\ a_2 \quad \text{avec ici} \quad a_1 = 1 \quad a_2 = -1$$

puisque $p_{1.} = \frac{1}{2}$.

Une analyse discriminante ordinaire effectuée sur les 5 variables ainsi codées donne les résultats suivants (il s'agit en fait ici d'une régression multiple où la variable explicative prend les valeurs 1 ou - 1 selon la catégorie des individus).

$\underline{u}_1 = - 0,046 ; 0,257 ; 0,428 ; 0,514 ; - 0,14 ; - 0,073 ; 0 ; - 0,285$

$\underline{u}_2 = - 0,108 ; - 0,325 ; 0,096 ; 0,075$

$\underline{u}_3 = - 0,017 ; 0,128$

$\underline{u}_4 = 0,817 ; 0,51 ; 0,327 ; 0,382 ; - 0,635 ; - 0,761$

$\underline{u}_5 = - 0,189 ; - 0,063 ; 0,05$

Le tableau de classement selon le critère positif - négatif est :

affectation \ origine	1	2
1	40	7
2	10	43

83 bien classés sur 100

Pour un risque $\alpha = 0,1$ on a trouvé $1 - \beta = 0,64$.

c) La méthode séquentielle de M. MASSON.

Les ϕ^2 de contingence de C avec les V_i valent :

V i	V1	V2	V3	V4	V5
100 ϕ_i^2	15,56	12,11	9,47	50,97	18,1

La première variable à introduire est donc V4 dont le codage est donné par :

$\underline{u}_4$ = 1,401 ; 0,875 ; 0,56 ; 0,654 ; - 1,089 ; - 1,304

Si on tient compte des degrés de liberté selon notre remarque du § 4 on ne constate pas de changement car dans cet exemple au plus fort ϕ^2 correspond aussi la plus forte probabilité pour la variable de χ^2 associée.

Avec V4 seule le tableau de classement selon le critère positif - négatif est :

affectation \ origine	1	2
1	48	14
2	2	36

soit 84 bien classés sur 100

Les effectifs des diverses catégories ne permettent pas de mesurer $1 - \beta$ à $\alpha = 0,1$.

A la deuxième étape la variable à introduire est V2 car les 4 analyses canoniques possibles de C avec V4 codée et V_i donnent les résultats suivants :

Variable à introduire	V1	V2	V3	V5
Pouvoir discriminant total λ	0,517	0,573	0,563	0,543

Le codage obtenu pour V2 est :

$\underline{u}_2$ = - 0,0173 ; 0,46 ; 0,483 ; 1,793 ; - 0,29 ; - 0,896 ; 0,325 ; - 0,489

Le tableau de classement pour la discrimination avec V2 et V4 selon

le critère positif - négatif est :

affectation \ origine	1	2
1	42	12
2	8	38

80 bien classés sur 100

A $\alpha = 0,1$ on trouve $1 - \beta = 0,72$.

Nous n'avons pas poursuivi l'algorithme car notre propos était de comparer ces résultats à ceux du codage optimal que l'on aurait obtenu par analyse directe de C avec V2 et V4 .

L'analyse directe donne les codages :

$\underline{u}_4 = 0,883 ; 0,999 ; 0,394 ; 0,668 ; - 1,015 ; - 1,17$

$\underline{u}_2 = 0,143 ; 0,93 ; 1,809 ; 0,536 ; - 0,463 ; - 0,3 ; - 0,464 ; - 0,795$

Le tableau de classement selon le critère positif - négatif est :

affectation \ origine	1	2
1	46	12
2	4	38

84 bien classés sur 100

A $\alpha = 0,1$ on a trouvé $1 - \beta = 0,76$.

On peut remarquer que si les codages obtenus pour la première variable diffèrent peu, ceux de la deuxième sont très différents (et pas seulement en raison de la contrainte de normalisation différente) ; la qualité de la discrimination obtenue est sensiblement meilleure par l'analyse directe.

Il peut donc être intéressant de combiner analyse directe et méthode séquentielle de la manière suivante :

La méthode séquentielle assez économique en temps calcul est utiliséepour sélectionner les "meilleures" variables discriminantes, ensuite on effectue une analyse directe sur les k variables retenues pour améliorer le codage.

On aura cependant remarqué sur cet exemple simple que les diverses techniques de codages envisagées conduisent à des discriminations assez voisines

bien qu'avec des codages numériques souvent différents.

TABLEAU DE BURT DES VARIABLES

EXPLICATIVES (OU MATRICE X'X)

	V1								V2				V3		V4						V5		
	46	0	0	0	0	0	0	0	3	10	14	19	40	6	3	6	7	12	4	14	7	11	28
	0	4	0	0	0	0	0	0	0	0	0	4	3	1	0	1	0	2	1	0	0	0	4
	0	0	12	0	0	0	0	0	1	0	1	10	0	3	1	4	1	5	0	1	0	1	11
V1	0	0	0	2	0	0	0	0	0	0	0	2	2	0	0	0	0	1	0	1	0	0	2
	0	0	0	0	11	0	0	0	1	2	0	8	10	1	1	0	0	5	3	2	2	2	7
	0	0	0	0	0	14	0	0	2	3	2	7	14	0	0	4	2	2	1	5	1	4	9
	0	0	0	0	0	0	2	0	1	0	1	0	2	0	0	0	0	1	0	1	0	0	2
	0	0	0	0	0	0	0	0	0	1	0	8	8	1	1	1	0	2	0	4	1	3	5
	3	0	1	0	1	2	1	0	8	0	0	0	8	0	0	0	2	1	0	5	3	3	2
V2	10	0	0	0	2	3	0	1	0	16	0	0	16	0	0	2	1	4	2	7	5	6	5
	14	0	1	0	0	2	1	0	0	0	18	0	18	0	1	2	0	7	3	5	1	6	11
	19	4	10	2	8	7	0	8	0	0	0	58	46	12	5	12	7	18	4	12	2	6	50
	40	3	9	2	10	14	2	8	8	16	18	46	88	0	4	11	8	27	9	29	11	10	58
V3	6	1	3	0	1	0	0	1	0	0	0	12	0	12	2	5	2	3	0	0	0	2	10
	3	0	1	0	1	0	0	1	0	0	1	5	4	2	5	0	0	0	0	0	0	1	5
	6	1	4	0	0	4	0	1	0	2	2	12	11	5	0	16	0	0	0	0	0	3	13
	7	0	1	0	0	2	0	0	2	1	0	7	8	2	0	0	10	0	0	0	1	2	7
V4	12	2	5	1	5	2	1	2	1	4	7	18	27	3	0	0	0	30	0	0	2	3	25
	4	1	0	0	3	1	0	0	0	2	3	4	9	0	0	0	0	0	0	0	0	5	4
	14	0	1	1	2	5	1	4	5	7	5	12	29	0	0	0	0	0	0	29	8	7	14
	7	0	0	0	2	1	0	1	3	5	1	2	11	0	0	0	1	2	0	8	11	0	0
V5	11	0	1	0	2	4	0	3	3	6	6	6	19	2	1	3	2	3	5	7	0	21	0
	28	4	11	2	7	9	2	5	2	5	11	50	58	10	5	13	7	25	4	14	0	0	58

CROISEMENTS DE LA VARIABLE DICHOTOMIQUE A EXPLIQUER C AVEC LES VARIABLES EXPLICATIVES (OU MATRICE A'X)

	V1								V2				V3	
C	21	3	11	2	4	6	1	2	3	2	11	34	39	11
	25	1	1	0	7	8	1	7	5	14	7	24	49	1

	V4						V5		
C	6	13	7	22	1	1	0	7	43
	0	3	3	8	8	28	11	14	25

CHAPITRE IV

L'ETUDE DIRECTE DES PROXIMITES ENTRE ENSEMBLES D'OBSERVATIONS ET ENTRE VARIABLES QUALITATIVES AU MOYEN D'OPERATEURS LINEAIRES

Les méthodes factorielles précédentes permettent de représenter les proximités entre les variables de différents ensembles ou entre les modalités de variables qualitatives, mais ne permettent pas de représenter directement les proximités entre ensembles ou entre variables qualitatives.

Reprenant les travaux d'Y. ESCOUFIER [8] et J.P. PAGES [19], qui associent à un ensemble de variables un être mathématique unique qui est un opérateur linéaire lié aux composantes principales de cet ensemble, nous examinerons les choix possibles de métrique sur l'ensemble des individus en fonction du but recherché et à partir d'une distance adéquate entre opérateurs nous exposerons la méthode de [19] pour représenter les proximités entre ensembles, que nous relierons à celle du chapitre II.

L'application, nouvelle, au cas des variables qualitatives permettra de dégager des indices de dépendance dont l'un, pour un choix particulier d'opérateurs ne sera autre que le coefficient de TSCHUPROW.

CHAPITRE IV

L'ETUDE DIRECTE DES PROXIMITES ENTRE ENSEMBLES D'OBSERVATIONS ET ENTRE VARIABLES QUALITATIVES AU MOYEN D'OPERATEURS LINEAIRES

Les méthodes factorielles précédentes permettent de représenter les proximités entre les variables de différents ensembles ou entre les modalités de variables qualitatives, mais ne permettent pas de représenter directement les proximités entre ensembles ou entre variables qualitatives.

Reprenant les travaux d'Y. ESCOUFIER [8] et J.P. PAGES [19], qui associent à un ensemble de variables un être mathématique unique qui est un opérateur linéaire lié aux composantes principales de cet ensemble, nous examinerons les choix possibles de métrique sur l'ensemble des individus en fonction du but recherché et à partir d'une distance adéquate entre opérateurs nous exposerons la méthode de [19] pour représenter les proximités entre ensembles, que nous relierons à celle du chapitre II.

L'application, nouvelle, au cas des variables qualitatives permettra de dégager des indices de dépendance dont l'un, pour un choix particulier d'opérateurs ne sera autre que le coefficient de TSCHUPROW.

I. OPERATEURS ET NORMES D'OPERATEURS ASSOCIES A DES TABLEAUX STATISTIQUES.

1. Généralités sur les normes.

$\mathbb{R}^p$ et $\mathbb{R}^n$ étant munis d'une structure euclidienne définie par des formes quadratiques P et N respectivement, nous conviendrons de représenter les applications linéaires de $\mathbb{R}^p$ dans $\mathbb{R}^n$ par des matrices en ayant fait choix de base respectivement P et N - orthonormées. P et N seront aussi représentées par des matrices.

L'ensemble des matrices rectangles associées aux applications linéaires de $\mathbb{R}^p$ dans $\mathbb{R}^n$ peut être normé de plusieurs façons différentes ; toutes les normes sont équivalentes au sens mathématique puisque les espaces sont de dimension finie mais elles ne conduisent pas toutes aux mêmes résultats pratiques, en particulier certaines d'entre elles sont associées à des produits scalaires mais pas d'autre.

Citons :

a) La norme déduite de la norme des opérateurs :

$$\|A\|^2 = \sup \frac{\|A\underline{x}\|_N^2}{\|\underline{x}\|_P^2} = \sup \frac{\underline{x}'A'NA\underline{x}}{\underline{x}'P\underline{x}}$$

$\|A\|^2 = \lambda_1$ où λ_1 est la plus grande valeur propre de $P^{-1}A'NA$.

En particulier si $N = I_n$ $P = I_p$ la norme de A est la racine carrée de la plus grande valeur propre de A'A. Cette norme n'est pas associée à un produit scalaire.

b) La norme euclidienne ordinaire :

Si A est une matrice de termes a_{ij} $\|A\|^2 = \sum_{i=1}^{n} \sum_{j=1}^{p} a_{ij}^2$

Cette norme est associée au produit scalaire $<A, B> = \text{Trace } (A'B)$ comme on peut le vérifier aisément. (Trace $A'B = \Sigma\Sigma\ a_{ij}\ b_{ij}$).

La norme euclidienne est celle de la somme des valeurs propres μ_i de A'A :

$$\|A\|^2 = \sum_{i=1}^{p} \mu_i$$

c) La norme utilisée pour les opérateurs d'ESCOUFIER

Pour des raisons qui apparaîtront plus loin nous utiliserons pour les matrices carrées la forme bilinéaire symétrique positive suivante :

$$< A, B > = \text{Trace } (A\,B)$$

qui permet de définir la semi-norme $\|A\|^2 = \text{Trace } A^2 = \sum_{i=1}^{n} \lambda_i^2$

où les λ_i sont les valeurs propres de A.

Cette semi-norme est une norme si on se restreint aux matrices diagonalisables car alors la nullité de toutes les valeurs propres est équivalente à celle de la matrice.

Dans le cas particulier de matrices symétriques cette norme se confond avec la norme euclidienne ordinaire, sinon elle munit l'ensemble des matrices diagonalisables d'une structure euclidienne différente.

2. Le cas des tableaux de données statistiques.

Si tout tableau de données X_i peut être considéré comme une matrice rectangle (n, m_i) on ne peut néanmoins comparer directement ces tableaux à l'aide des normes définies précédemment car ils peuvent être de dimensions différentes.

Il faut donc substituer à chaque tableau X_i un opérateur de $\mathbb{R}^n$ dans lui-même (car n est la dimension commune des X_i) c'est-à-dire une matrice carrée à laquelle nous pourrons appliquer la norme définie par la trace du produit.

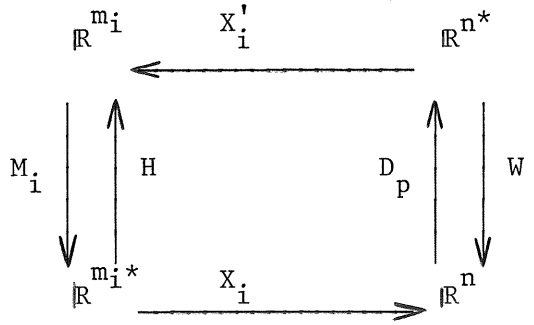
Nous exigerons que cette matrice soit diagonalisable afin d'avoir une véritable norme et non une semi-norme. Il semble naturel d'exiger de ces opérateurs qu'ils aient même espace image que les tableaux X_i de départ, c'est-à-dire W_i. Il est souhaitable aussi que la D_p-orthogonalité de deux espaces W_i et W_j se traduisent par l'orthogonalité des opérateurs associés pour le produit scalaire de la trace.

Il est facile de voir que les opérateurs $O_i = X_i M X'_i D_p$ répondent à toutes ces conditions avec M_i matrice symétrique (m_i, m_i) définie positive ;

si on considère M_i comme définissant une métrique sur l'espace des individus décrit par l'ensemble de variables de tableau X_i, $X_i M X_i'$ caractérise les distances deux à deux des n points représentatifs des individus dans $\mathbb{R}^{m_i}$.

L'opérateur O_i est équivalent au tableau de données X_i en ce sens que ses vecteurs propres sont les composantes principales du tableau X_i lorsque les individus sont munis de la métrique M_i. Le sens de la norme choisie pour les opérateurs s'éclaire alors : c'est la somme des carrés des valeurs propres de l'A.C.P. de métrique M_i du tableau X_i.

Ces opérateurs ont été introduits par Y. ESCOUFIER, qui, dans un cadre probabiliste, cherchait à résumer en un seul opérateur un ensemble de variables aléatoires et voulait que cet opérateur ait pour vecteurs propres les composantes principales de l'ensemble. Notons que dans ce cadre probabiliste, la possibilité d'utiliser des métriques différentes sur l'espace des individus comme en analyse des données, n'était pas exploitée. Par la suite, ces opérateurs ont été appliqués en analyse des données par BRAUN et PAGES notamment qui ont reconnu en eux l'application WD_p de l'analyse en composantes principales selon le schéma de dualité :



3. Normes, équivalences et indices de proximité.

Deux tableaux de données seront considérés comme équivalents si la distance entre leurs opérateurs associés est nulle :

$$X_i \sim X_j \quad \text{si} \quad \|O_i - O_j\|^2 = 0$$

Ceci veut dire que X_i et X_j ont même système de composantes principales. La nature de l'équivalence dépend en fait du choix de la métrique M_i

pour chaque tableau X_i . Considérons les principales métriques utilisées en pratique :

a) $M_i = I_{m_i} \quad \forall i$.

Deux tableaux de données ayant même composantes principales ordinaires (matrice des variances-covariances) sont équivalents. Ceci veut dire que les deux nuages des n points représentatifs des individus selon les deux tableaux de données se déduisent l'un de l'autre par une isométrie.

b) $M_i = D_{1/\sigma^2} \quad \forall i$.

Deux tableaux sont équivalents s'ils ont même composantes principales normées (matrice des corrélations). L'interprétation géométrique est la même qu'en a) à ceci près que l'on a opéré préalablement une réduction des données. Les deux nuages de points se déduisent l'un de l'autre par similitude.

c) $M_i = (X_i' D_p X_i)^{-1} = V_{ii}^{-1} \quad \forall i$.

C'est la métrique de MAHALANOBIS ; les opérateurs O_i ne sont autres que les projecteurs D_p-orthogonaux associés aux espaces W_i .

Deux tableaux de données sont alors équivalents s'ils engendrent le même espace W : c'est l'équivalence considérée en analyse canonique. X_i et X_j ont alors mêmes composantes principales car dans les métriques V_{ii}^{-1} et V_{jj}^{-1} tout vecteur de W_i (resp. W_j) est composante principale de X_i (resp. X_j).

Le choix de la métrique M_i dépend donc essentiellement du but recherché.

Le produit scalaire $< O_i, O_j > = \text{Trace } (O_i, O_j)$ munissant l'ensemble des opérateurs de la forme $O_i = X_i M X_i' D_p$ d'une structure euclidienne, il est alors possible de mesurer la proximité entre tableaux de données au moyen de l'indice qui est le cosinus déduit du produit scalaire:

$$\cos (O_i, O_j) = \frac{\text{Trace } (O_i O_j)}{\sqrt{\text{Trace } O_i^2 \quad \text{Trace } O_j^2}}$$

Cet indice traduit globalement la proximité entre deux tableaux de données X_i et X_j .

Examinons-les pour les 3 choix de métriques envisagées plus haut.

a) $M_i = I_{m_i}$ $\quad O_i = X_i X_i' D_p$

$$\cos(O_i, O_j) = \frac{\text{Trace}\,(X_i X_i' D_p X_j X_j' D_p)}{\sqrt{\text{Trace}\,(X_i X_i' D_p)^2 \,\text{Trace}\,(X_j X_j' D_p)^2}}$$

Or $\text{Trace}\,(X_i X_i' D_p X_j X_j' D_p) = \text{Trace}\,(X_i' D_p X_j X_j' D_p X_i) = \text{Trace}\,(V_{ij} V_{ji})$ par commutation sous l'opérateur Trace de X_i avec le reste.

$\text{Trace}\,(V_{ij} V_{ji})$ n'est autre que le carré de norme euclidienne ordinaire de la matrice V_{ij} des intercovariances.

D'où : $$\cos(O_i, O_j) = \frac{\text{Trace}\, V_{ij} V_{ji}}{\sqrt{\text{Trace}\,(V_{ii}^2)\,\text{Trace}\,(V_{jj})^2}}$$

L'indice angulaire est nul si $V_{ij} = 0$ donc si $W_i \underset{D_p}{\perp} W_j$ et seulement dans ce cas.

$\text{Trace}\,(V_{ii}^2)$ est égale à la somme des carrés des valeurs propres de la matrice de covariance de X_i.

Cet indice est une généralisation multidimensionnelle du carré du coefficient de corrélation puisque si X_i et X_j sont unicolonnes, $\underline{x}_i$ et $\underline{x}_j$, alors $V_{ii} = \sigma_i^2$ $\quad V_{jj} = \sigma_j^2$ $\quad V_{ij} = \text{COV}\,(\underline{x}_i, \underline{x}_j)$ d'où :

$$\text{Trace}\,(O_i, O_j) = (\text{COV}\,(\underline{x}_i, \underline{x}_j))^2$$

$$\cos(O_i, O_j) = r^2.$$

b) $M_i = D_{1/\sigma^2}$

Les résultats sont semblables aux précédents à ceci près qu'il faut remplacer les matrices de variance par les matrices de corrélation. L'orthogonalité entre opérateurs est encore équivalente à la D_p-orthogonalité des espaces engendrés et l'indice angulaire vaut :

$$\cos(O_i, O_j) = \frac{\text{Trace}\,(R_{ij} R_{ji})}{\sqrt{\text{Trace}\, R_{ii}^2 \,\text{Trace}\, R_{jj}^2}}$$

Lorsque X_i et X_j sont unicolonnes, même conclusion que précédemment.

c) $M_i = (X_i' D_p X_i)^{-1}$

Les opérateurs O_i étant les projecteurs D_p-orthogonaux P_i sur W_i on a : $\text{Trace } (P_i P_j) = \sum_{k=1}^{m_i} (r_{ij}^{(k)})^2$ si $m_i < m_j$.

Le produit scalaire est donc égal à la somme des carrés des coefficients de corrélation canonique.

$$\text{Trace } P_i^2 = \text{Trace } P_i = m_i \quad \text{si} \quad \text{rang } X_i = m_i .$$

L'indice angulaire vaut donc :

$$\boxed{\text{COS } (P_i, P_j) = \frac{\Sigma (r_{ij}^{(k)})^2}{\sqrt{m_i m_j}}}$$

L'indice n'est nul que si tous les coefficients de corrélation canonique sont nuls, c'est-à-dire si W_i et W_j sont D_p-orthogonaux.

L'indice ne peut être égal à 1 que si $m_i = m_j$ et $r_{ij}^{(k)} = 1 \quad \forall k$ c'est-à-dire dans le cas où $W_i = W_j$.

Lorsque $m_i \neq m_j$ on a :

$$\text{COS } (P_i, P_j) < \sqrt{\frac{\inf (m_i, m_j)}{\sup (m_i, m_j)}}$$

L'égalité ne pouvant être atteinte que si $W_i \subset W_j$ ou $W_j \subset W_i$.

II. APPLICATION A L'ANALYSE DES GROUPES DE DONNEES.

1. Représentation euclidienne des proximités entre opérateurs.

Au produit scalaire de la trace pour les opérateurs est associée une distance euclidienne entre opérateurs donc entre tableaux de données par :

$$d^2(X_i, X_j) = \text{Trace } (O_i)^2 + \text{Trace } (O_j)^2 - 2 \text{ Trace } (O_i\ O_j)$$

Ces distances étant euclidiennes nous savons qu'il existe un espace euclidien de dimension ou plus égale à $p-1$, si p est le nombre de tableaux étudiés, dans lequel nous pouvons représenter les opérateurs comme des points en conservant leurs distances mutuelles.

La méthode classique de l'analyse factorielle sur tableaux de distance permet d'obtenir les coordonnées de ces points sur le repère des axes principaux d'inertie, ainsi que des représentations approchées dans des espaces de dimension inférieure en ne retenant que certains espaces principaux.

Rappelons brièvement cette méthode [6] :

On constitue à partir du tableau des distances $d_{ij} = d(X_i, X_j)$ la matrice W de terme général :

$$W_{ij} = p\ (d^2_{i.} + d^2_{.j} - d^2_{ij} - d^2_{..})$$

où :
$$d^2_{i.} = \frac{1}{p} \sum_j d^2_{ij} \qquad d^2_{..} = \frac{1}{p} \sum_i d^2_{i.}$$

Alors la matrice W admet pour vecteurs propres les composantes principales du nuage des p points. La coordonnée sur le $k^{ième}$ axe du point représentatif du tableau X_i sera donc la $i^{ème}$ composante du $k^{ième}$ vecteur propre de W si on l'a normé de telle sorte que la somme de carrés de ses composantes soit égale à μ_k valeur propre correspondante.

On obtient une représentation plane approchée du nuage des opérateurs en considérant les deux premiers vecteurs propres de W, la part d'inertie expliquée étant donnée par :

$$\frac{\mu_1 + \mu_2}{\text{Trace } W}$$

On a ici $\text{Trace } W = p^2\ d^2_{..} = \sum_i \sum_j d^2(X_i, X_j)$

2. Représentation des proximités entre espaces W_i ou entre projecteurs P_i.

Nous pouvons à présent complèter la méthode factorielle du chapitre II qui consistait à diagonaliser $\sum_{i=1}^{p} P_i$, en étudiant directement les distances entre tableaux de données, la métrique choisie étant celle de MAHALANOBIS V_{ii}^{-1}, ce qui revient à étudier les distances entre les P_i.

On a alors $d^2(P_i, P_j) = m_i + m_j - 2 \sum_k (r_{ij}^{(k)})^2$, ce qui entraîne d'ailleurs que :

$$| m_j - m_i | \leqslant d^2(P_i, P_j) \leqslant m_i + m_j$$

L'analyse factorielle du tableau des distances permet alors une représentation plane des proximités entre espaces W_i plus intéressante que celle du chapitre II § IV. 3. : si la part d'inertie portée par le plan principal est suffisante la proximité de deux points P_i et P_j reflètera celle des espaces W_i et W_j.

Il est possible de relier par une formule les valeurs propres λ_i de l'analyse factorielle du chapitre I et la valeur propre μ_i de l'analyse du tableau des distances entre projecteurs.

En effet, la trace de W qui vaut $\sum_i \sum_j d^2(P_i, P_j)$ peut se décomposer en :

$$\text{Trace } W = \sum_{i=1}^{p} \sum_{j=1}^{p} (m_i + m_j - 2 \text{ Trace } P_i P_j)$$

soit : $$\text{Trace } W = 2pm - 2 \sum_{i=1}^{p} \sum_{j=1}^{p} \text{Trace } (P_i P_j) \quad \text{si} \quad m = \sum_{i=1}^{p} m_i$$

Or, nous avons montré au chapitre II que la somme des carrés des coefficients de corrélation canoniques de toutes les analyses possibles, avec répétition était égale à $\sum_{i=1}^{m} \lambda_i^2$.

d'où : $$\boxed{\sum_{i=1}^{p} \mu_i = 2 \left[pm - \sum_{i=1}^{m} \lambda_i^2 \right]}$$

On vérifie bien que si $W_1 = W_2 = \ldots = W_p$, $\sum_{i=1}^{p} \mu_i = 0$, (car tous les

projecteurs sont confondus) entraîne que $\sum_{i=1}^{m} \lambda_i^2 = pm$, en effet si tous les W_i sont confondus alors $\lambda_i = p$ est la seule valeur propre non nulle de $V^{-1}H$ avec un ordre de multiplicité égal à la dimension commune des W_i, mettons m_1, d'où $\Sigma \lambda_i^2 = m_1 p^2$ et comme alors $m = m_1 p$ on a bien le résultat annoncé.

On peut noter au passage que $\mu_p = 0$ dans tous les cas car le nuage des p points représentatifs des projecteurs se trouve dans un espace de dimension maximale $p-1$.

Remarque.

Il est possible de calculer les distances entre projecteurs P_i et P_j sans être obligés d'effectuer l'analyse canonique de X_i avec X_j si les variables de chaque ensemble sont séparément D_p-orthogonales : c'est le cas en particulier si on a remplacé les variables de X_i et X_j par les composantes principales. Dans ce cas, la somme des carrés des coefficients de corrélation canonique est égale à la somme des carrés des termes de R_{ij} matrice d'intercorrélation de X_i avec X_j : en effet la somme des carrés des termes d'une matrice est invariante pour toute transformation orthogonale sur chacun des espaces : il suffit alors de faire dans chaque espace la transformation orthogonale qui transforme les vecteurs D_p-orthogonaux par hypothèses, en les variables canoniques. La matrice R_{ij} s'écrit alors sur les bases des variables canoniques :

$$R_{ij} = \left[\begin{array}{c|c} \begin{matrix} r_1 & & \\ & r_2 & \\ & & \ddots \\ & & & r_{m_i} \end{matrix} & \\ \hline & \end{array}\right] \quad \text{si } m_i < m_j .$$

où les r_i sont les coefficients de corrélation canonique. (Il faut compléter dans W_j la base des variables canoniques par $m_j - m_i$ vecteurs formant une base quelconque de la partie de W_j D_p-orthogonale à W_i) .

Le résultat précédent n'est qu'une généralisation de la propriété bien connue du coefficient de corrélation multiple d'une variable avec p - variables non corrélées.

III. REPRESENTATION DES PROXIMITES ENTRE VARIABLES QUALITATIVES.

1. La distance associée au ϕ^2 et le T de TSCHUPROW.

Si nous appliquons les résultats du paragraphe précédent à des tableaux logiques représentant des variables qualitatives une précaution doit être prise pour le choix des projecteurs : en effet, si on utilise la formule habituelle $P_i = X_i (X_i' X_i)^{-1} X_i'$ (ici $D_p = \frac{1}{n} I_n$), l'orthogonalité de deux projecteurs est impossible car tous les espaces ont en commun le vecteur $\underline{1}$.

Il faut donc considérer les projecteurs P_{io} sur W_{io} sous-espace de W_i orthogonal à $\underline{1}$ des variables numériques centrées compatibles avec la variable qualitative n° i .

Dans ces conditions alors que $\text{Trace}\ (P_i\ P_j) = 1 + \phi^2_{ij}$ on aura donc $\text{Trace}\ (P_{io}\ P_{jo}) = \phi^2_{ij}$ et $\text{Trace}\ (P^2_{io}) = m_i - 1$.

L'orthogonalité de deux projecteurs pour le produit scalaire défini par la trace sera alors équivalente à la nullité du ϕ^2, c'est-à-dire à l'indépendance statistique des deux variables qualitatives.

L'indice angulaire de dépendance entre deux variables qualitatives i et j sera donc :

$$\frac{\text{Trace}\ (P_{io}\ P_{jo})}{\sqrt{\text{Trace}\ P^2_{io}\ \text{Trace}\ P^2_{jo}}}$$

soit :

$$\boxed{T^2 = \frac{\phi^2_{ij}}{\sqrt{(m_i - 1)\ (m_j - 1)}}}$$

Cet indice n'est autre que le carré du coefficient T de TSCHUPROW qui avait été introduit par cet auteur afin de ramener le ϕ^2 d'un tableau de contingence à une mesure comprise entre 0 et 1, car la valeur maximale du ϕ^2 dépend des dimensions du tableau de contingence.

L'interprétation nouvelle du coefficient de TSCHUPROW, que nous donnons, comme cosinus de l'angle entre les deux projecteurs centrés associés à deux variables qualitatives, peut redonner ainsi un intérêt à une mesure de dépendance

quelque peu tombée en désuetude.

L'égalité à 1 de T ne peut être atteinte que si $m_i = m_j$ lorsque le tableau de contingence, après avoir réordonné convenablement lignes et colonnes, se met sous forme diagonale. Plus généralement la valeur maximale de T est atteinte dans ce cas qui est celui de l'association complète :

$$X_i' X_j = \left[\begin{array}{c|c} n_1 & \\ & n_2 \\ & \ddots \\ & n_{m_i} \end{array}\right]$$

Si $m_i < m_j$ $\qquad T_{max} = \sqrt{\dfrac{m_i - 1}{m_j - 1}}$ et $\phi^2_{ij} = m_i - 1$

car $\phi^2_{ij} = \left[\sum_i \sum_j \left(\dfrac{(n_{ij})^2}{n_{i.}\, n_{.j}} \right) - 1 \right]$ vaut alors $\sum_{i=1}^{m_i} (1) - 1$

puisque $n_{ii} = n_{i.} = n_{.j}$ et $n_{ij} = 0$ si $i \neq j$.

Il est clair que si T atteint sa valeur maximale on peut en pratique supprimer les modalités qui ne sont jamais prises et se ramener au cas $m_i = m_j$ du tableau de contingence carrée, à condition toutefois de disposer d'un échantillon assez grand.

On peut utiliser aussi le coefficient de CRAMER :

$$C = \frac{\phi^2_{ij}}{\inf \{(m_i - 1)(m_j - 1)\}}$$ qui peut atteindre la valeur 1 pour des tableaux rectangulaires mais qui n'a pas la même interprétation angulaire que celui de TSCHUPROW.

On peut définir une distance euclidienne entre variables qualitatives par :

$$\boxed{d^2_{i,j} = m_i + m_j - 2 - 2\,\phi^2_{ij} = m_i + m_j - 2\,(1 + \phi^2_{ij})}$$

et de même que pour les tableaux numériques on aura :

$$| m_j - m_i | < d^2_{ij} < m_i + m_j - 2$$

ce qui ne fait que traduire l'inégalité $0 \leqslant \phi^2_{ij} \leqslant \inf \{ (m_i - 1) (m_j - 1) \}$.

On pourra donc en effectuant une analyse factorielle sur le tableau des distances entre variables qualitatives obtenir une représentation euclidienne des proximités entre ces variables, mais alors il n'est plus possible de représenter en même temps les proximités entre les modalités de ces variables.

Notons qu'il ne faut pas se laisser abuser par des proximités qui peuvent être trompeuses comme nous en avons fait la remarque au chapitre précédent : de petites distances d^2_{ij} donc de grandes valeurs de ϕ^2 ne doivent pas être interprétées trop vite comme un indice absolument certain de dépendance à cause des dimensions des tableaux (voir le problème des degrés de liberté des χ^2 associés).

2. Une application possible à l'analyse non linéaire des données.

La méthode de représentation précédente des proximités entre variables qualitatives peut, à notre avis, être un outil intéressant pour analyser les dépendances non linéaires entre variables numériques.

En effet, lorsque l'on dispose d'un tableau de données numériques l'étude de la matrice des corrélations et l'analyse en composantes principales normée qui s'en déduit ne met en évidence que des dépendances linéaires entre variables statistiques.

Si on procède à un découpage en classes de ces variables pour les rendre qualitatives, on sait qu'on s'affranchit ainsi du caractère linéaire pour prendre en compte des liaisons d'autres nature. En étudiant alors la dépendance entre les variables qualitatives ainsi obtenues on mettra en évidence la liaison fonctionnelle entre ces variables : deux variables qui se trouvaient éloignées sur le cercle des corrélations de l'A.C.P. pourront alors se retrouver proches dans la représentation plane déduite de l'analyse du tableau des distances entre variables qualitatives s'il existe une dépendance non linéaire entre elles.

On réalisera donc en quelque sorte une analyse en composantes principales non linéaire si on peut s'exprimer ainsi.

3. Une nouvelle mesure de dépendance entre variables qualitatives.

A titre d'exercice nous avons examiné ce que donnait pour des variables qualitatives l'indice angulaire entre opérateurs si on choisissait comme métrique sur l'espace des individus non pas $M_i = V_{ii}^{-1}$ mais $M_i = I_{m_i}$.

L'indice angulaire vaut d'après le § I. 3. :

$$\frac{\text{Trace } (V_{12}\, V_{21})}{\sqrt{\text{Trace } V_{11}^2 \;\; \text{Trace } V_{22}^2}} \; .$$

Il faut ici centrer les variables indicatrices pour pouvoir parler de matrices de covariance et nous trouvons alors au facteur n près, que :

V_{12} est la matrice de terme général $p_{ij} - p_{i.}\, p_{.j}$

V_{21} est sa transposée

V_{11} est la matrice de termes diagonaux $p_{i.}(1-p_{i.})$ et non diagonaux $-p_{i.}p_{i'.}$.

V_{22} est la matrice analogue à V_{11} avec $p_{.j}$

L'indice angulaire associé est donc:

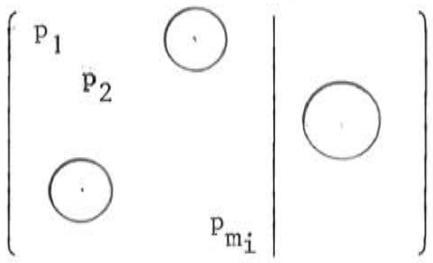
$$I = \frac{\sum_i \sum_j (p_{ij} - p_{i.}\, p_{.j})^2}{\sqrt{\left[\sum_i (p_{i.}(1-p_{i.}))^2 + \sum\sum (p_{i.}\, p_{i'.})^2\right]\left[\sum_j (p_{.j}(1-p_{.j}))^2 + \sum\sum (p_{.j}\, p_{.j'})^2\right]}}$$

Cet indice qui, à notre connaissance, n'a jamais été proposé, jouit des propriétés suivantes :

$I = 0$ est équivalent à l'indépendance car $p_{ij} = p_{i.}\, p_{.j} \quad \forall i,j$.

$I = 1$ est équivalent à l'association complète entre les variables qualitatives :

On vérifie en effet que s'il y a au plus un terme non nul par ligne et par colonne du tableau de contingence, c'est-à-dire si celui-ci est diagonal au sens large :

$\frac{1}{n} X_i' X_j =$  après une permutation des modalités.

alors $p_{ij} = 0$ si $i \neq j$ $\quad p_{ii} = p_{i.} = p_{.i}$ le numérateur vaut alors

$\Sigma (p_{i.}(1-p_{i.}))^2 + \Sigma\Sigma (p_{i.} p_{i'.})^2$ et le dénominateur aussi.

d'où le résultat annoncé. La réciproque est alors immédiate.

4. Application numérique.

Nous avons appliqué les résultats du § 1 à l'exemple numérique du chapitre III § III.5. afin d'étudier les proximités entre les 6 variables qualitatives C, V1, V2, V3, V4, V5.

Le tableau des ϕ^2 de contingence était le suivant, entre parenthèses figure le nombre de degrés de liberté.

	V1	V2	V3	V4	V5
C	0,1556 (7)	0,1211 (3)	0,0947 (1)	0,5097 (5)	0,181 (2)
V1		0,3009 (21)	0,05157 (7)	0,2999 (35)	0,1031 (14)
V2			0,09875 (3)	0,4634 (15)	0,3042 (6)
V3				0,141 (5)	0,021 (2)
V4					0,2232 (10)

L'analyse factorielle du tableau des distances définies par :

$d^2_{ij} = m_i + m_j - 2 - 2\,\phi^2_{ij}$ a donné les résultats ci-dessous :

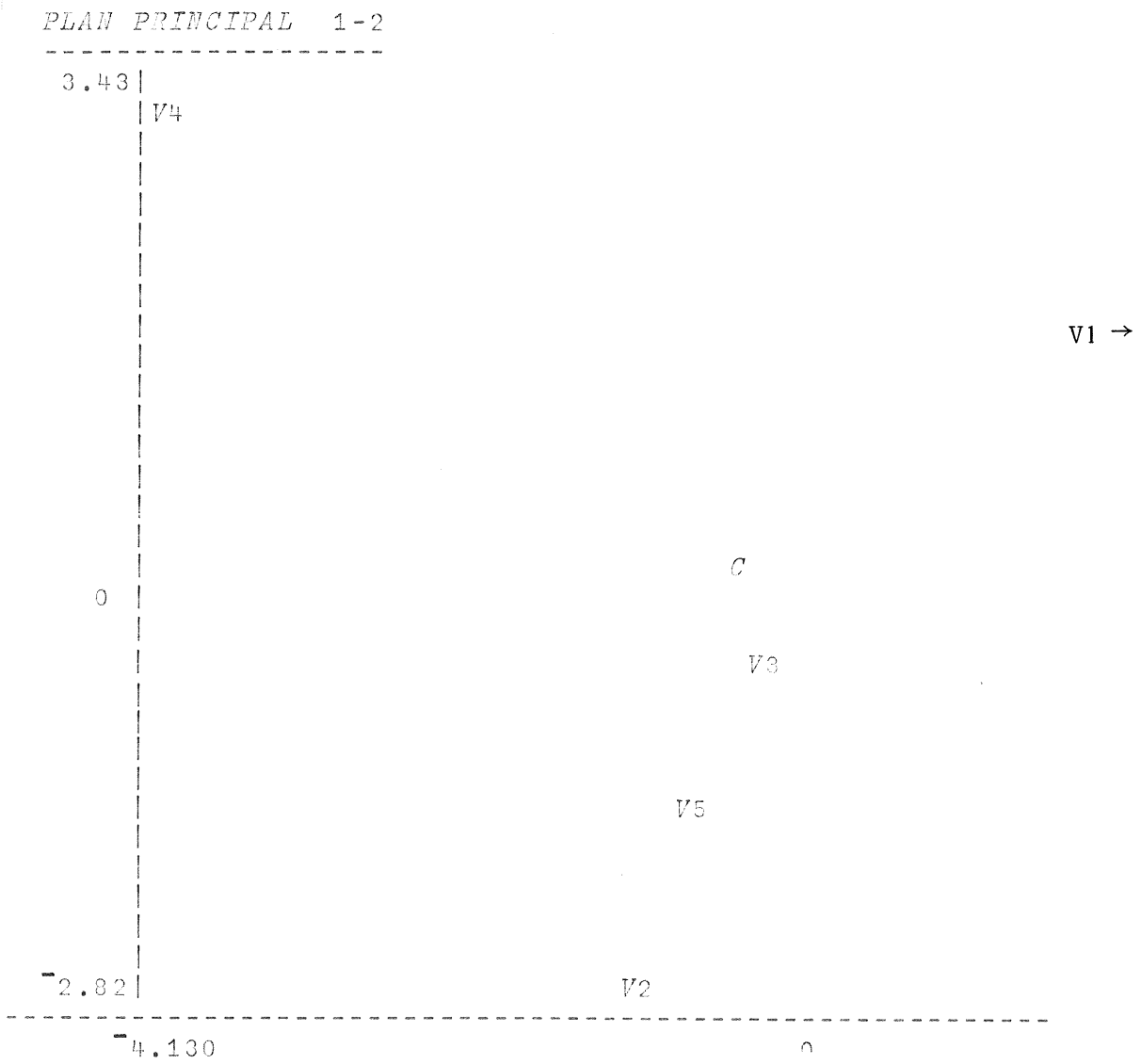
TABLEAU DES DISTANCES AU CARRE d^2_{ij} :

	C	V1	V2	V3	V4	V5
C	0.000E-01	7.688E00	3.758E00	1.811E00	4.981E00	2.638E00
V1	7.688E00	0.000E-01	3.938E00	7.897E00	1.140E01	8.794E00
V2	3.758E00	3.938E00	0.000E-01	3.802E00	7.073E00	4.332E00
V3	1.811E00	7.897E00	3.802E00	0.000E-01	5.718E00	2.958E00
V4	4.981E00	1.140E01	7.073E00	5.718E00	0.000E-01	6.554E00
V5	2.638E00	8.794E00	4.332E00	2.958E00	6.554E00	0.000E-01

	VALEUR PROPRE	POURCENTAGE D'INERTIE	CUMUL
1	11.603	.658	.658
2	4.030	.229	.887
3	1.543	.088	.975
4	.436	.025	.999
5	.011	.001	1.000

VECTEURS PROPRES NORMES A LA RACINE CARREE DE LA VALEUR PROPRE

-4.357E-01	1.159E-01
7.072E00	1.323E00
-1.135E00	-2.821E00
-4.113E-01	-4.977E-01
-4.130E00	3.434E00
-8.339E-01	-1.554E00



La représentation obtenue n'est cependant pas très satisfaisante car elle dépend trop fortement des nombres de modalités des diverses variables, c'est-à-dire des degrés de liberté. Or, ceux-ci sont très différents (de 1 jusqu'à 35).

Afin de nous affranchir des degrés de liberté remarquons tout d'abord que puisque les distances sont euclidiennes il est en fait inutile de calculer le tableau de distances et d'effectuer ensuite une A.F.T.D. car le tableau des ϕ^2 est alors un tableau de produit scalaire et il suffit de le diagonaliser directement, effectuant ainsi une analyse en composantes principales ordinaires si on considère les ϕ^2 comme des covariances généralisées.

Dans ces conditions on obtiendra une analyse indépendante des normes des variables, donc des degrés de liberté, si on considère le tableau des cosinus

d'angles associés aux produits scalaires, c'est-à-dire le tableau des T^2 de TSCHUPROW. La diagonalisation du tableau des T^2 est alors l'analogue d'une A.C.P. sur matrices de corrélations.

La matrice des T^2 de TSCHUPROW est alors définie par :

	V1	V2	V3	V4	V5
C	0,0588	0,0547	0,051	0,2279	0,128
V1		0,0657	0,0195	0,0507	0,0276
V2			0,057	0,1196	0,1242
V3				0,0631	0,0148
V4					0,0706

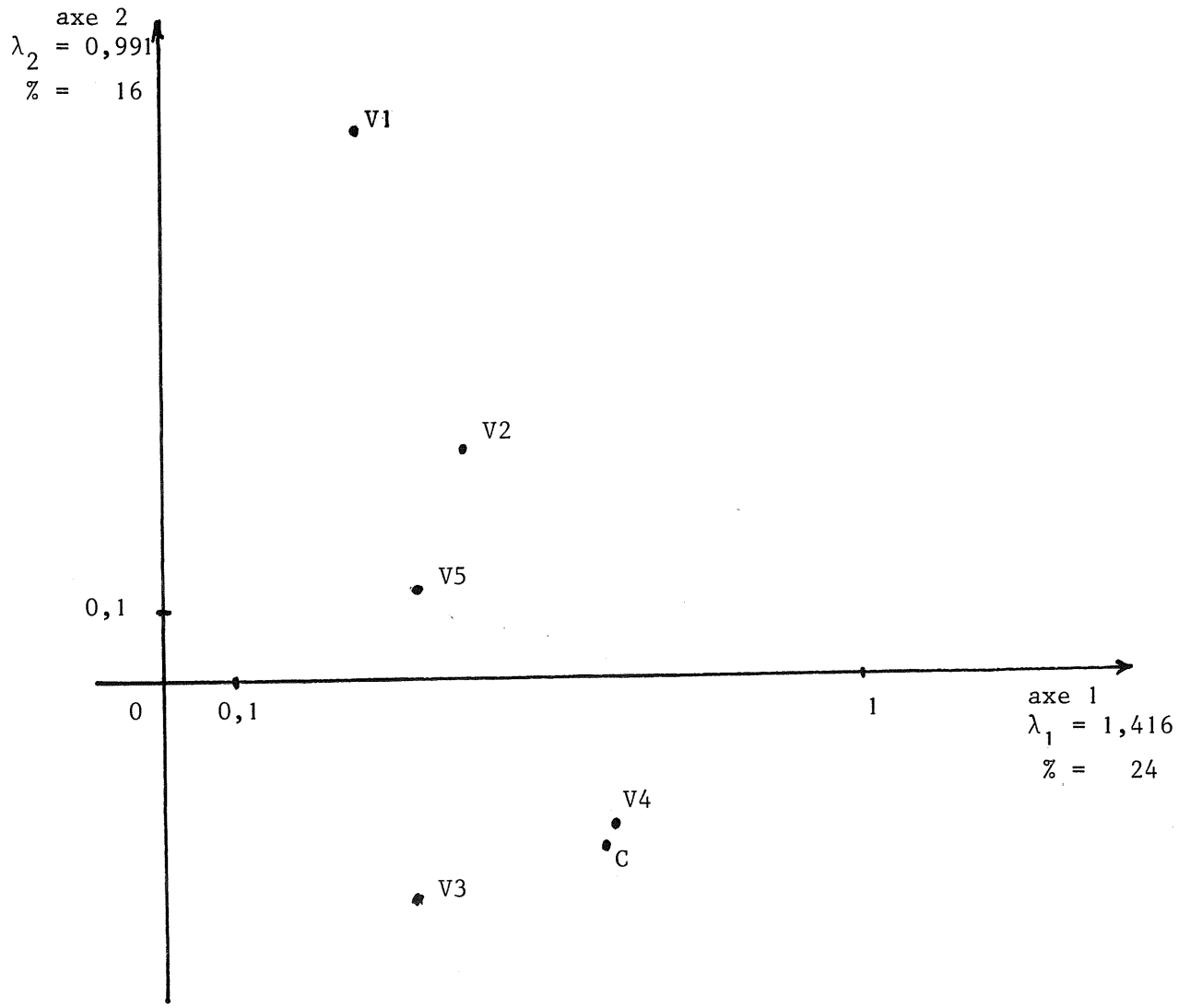
Sa diagonalisation donne les résultats suivants :

	Valeurs propres	pourcentage d'inertie	cumul
λ_1	1,416	0,236	0,236
λ_2	0,991	0,165	0,401
λ_3	0,982	0,164	0,565
λ_4	0,964	0,160	0,725
λ_5	0,899	0,150	0,875
λ_6	0,748	0,125	1

Vecteurs propres normés à la racine carrée de la valeur propre :

	axe 1	axe 2
C	0,63	- 0,24
V1	0,28	0,79
V2	0,43	0,33
V3	0,36	- 0,32
V4	0,64	- 0,21
V5	0,36	0,13

La représentation graphique est alors bien plus satisfaisante car les proximités correspondent alors aux liaisons fortes entre variables. On retrouve graphiquement le fait, déjà vu au chapitre III, que la variable la plus liée à C est V4 .



Représentation des "corrélations" entre variables qualitatives par diagonalisation de la matrice T^2 de TSCHUPROW.

BIBLIOGRAPHIE

[1] ANDERSON T.W. -
Introduction to multivariate statistical analysis.
WILEY (1958).

[2] BENZECRI J.P. -
L'analyse des données.
Tome II, DUNOD (1973).

[3] BRAUN J.M. -
Thèse de troisième cycle.
Université de Paris VI (1973).

[4] BURT C. -
The factorial analysis of qualitative data.
British Journal Statist. Psychol. (3) 166 - 185 (1950).

[5] CARROLL J.D. -
A generalization of canonical correlation analysis to three or more sets of variables.
Proc. 76th Conv. Amer. Psych. Assoc. p. 227 - 228 (1968).

[6] C E E E -
Analyse des données multidimensionnelles.
Centre d'Etudes Economiques d'entreprise (1971).

[7] DAUXOIS J. et POUSSE A. -
Analyse canonique de deux tribus.
Pub. n° 1 - 74, Laboratoire de Stat. Univ. P. Sabatier Toulouse (1974).

[8] ESCOUFIER Y. -
Echantillonnage dans une population de variables aléatoires réelles. *Thèse de Doctorat d'Etat - Univ. Montpellier (1970).*

[9] HORST P. -
Obtaining a composite measure from a number of different measures of the same attribute.
Psychometrika 1 p. 183 - 187 (1936).

[10] HORST P.
Relations among m sets of variables.
Psychometrika 26 p. 129 - 149 (1961).

[11] JAN DE LEEUW -
Canonical analysis of categorical data.
Université Leyde (non daté).

[12] KENDALL M.G. et STUART A. -
The advanced Theory of statistics.
Tome II, GRIFFIN (1967).

[13] KETTENRING R.J. -
Canonical analysis of several sets of variables.
Biometrika 58 p. 433 - 451 (1971).

[14] LANCASTER H.O. -
The structure of bivariate distributions.
Annals - Math. Stat. 29 p. 719 - 736 (1958).

[15] LANCASTER H.O. -
The Chi - Square Distribution.
WILEY (1969).

[16] MASSON M. -
C.R. Acad. Sc.
t. 278 Série A p. 803 - 806 (1974).

[17] MASSON M. -
Processus linéaires et analyse de données non linéaires.
Thèse de Doctorat d'Etat, Univ. Paris VI (1974).

[18] NAOURI J.C. -
Analyse factorielle des correspondances continues.
Thèse de Doctorat d'Etat, Univ. Paris VI (1971).

[19] PAGES J.P. -

A propos des opérateurs d'Y. Escouffier.
Note CEA Fontenay (1975).

[20] RAO C.R. -

Linear Statistical Inference and its Applications.
WILEY (1965).

[21] ROMEDER J.M. -

Méthodes et programmes d'analyse discriminante.
DUNOD (1973).

[22] SAPORTA G. -

Dépendance et codage de deux variables aléatoires.
A paraître Rev. Stat. Appliquée 23 n° 1 (manuscrit remis le 6-2-74) (1975).

[23] SEARLE S.R. -

Linear Models.
WILEY (1971).

[24] STEEL R.G.D. -

Minimum generalized variance for a set of linear functions
Annals - Math. Stat. 22 p. 456 - 460 (1951).

[25] WILLIAMS E.J. -

Use of scores for the analysis of association in contingency tables.
Biometrika 39 p. 274 - 289 (1952).

[26] WOLD S. -

QQ Regression : a NIPALS procedure for regression with qualitative and quantitative variables.
Univ. Umea. Suède (non daté).