REVUE DE STATISTIQUE APPLIQUÉE

J.-P. PAGES
F. CAILLIEZ
Y. ESCOUFIER

Analyse factorielle : un peu d'histoire et de géométrie

Revue de statistique appliquée, tome 27, n° 1 (1979), p. 5-28
<http://www.numdam.org/item?id=RSA_1979__27_1_5_0>

ANALYSE FACTORIELLE : UN PEU D'HISTOIRE ET DE GÉOMÉTRIE

J.-P. PAGES*, F. CAILLIEZ**, Y. ESCOUFIER***

Cet article reprend en partie un travail mené avec P. CAZES et coll. (Bibl. [14]) et complété avec R. TOMASSONE (Communication au colloque CNRS sur l'Elaboration et la Justification des Modèles en Biologie, 1978).

SOMMAIRE

(*) L.S.E.E.S. Département de Protection, Centre d'Etudes Nucléaires de Fontenay aux Roses.

(**) Centre Technique Forestier Tropical, Nogent sur Marne.

(***) Université des Sciences et Techniques du Languedoc, Montpellier.

Revue de Statistique Appliquée, 1979, vol. XXVII, n° 1

Mots-clés : **Analyses factorielles, schéma de dualité, opérateurs, analyse des correspondances, tableaux de contingence, tableaux disjonctifs complets.**

0. – INTRODUCTION

On considère en France que l'analyse des données recouvre principalement deux ensembles de techniques : les premières, qui relèvent de la géométrie euclidienne et conduisent à l'extraction de valeurs et de vecteurs propres, sont appelées "analyses factorielles" ; les secondes, dites de "classification automatique" sont caractérisées par le choix d'un indice de proximité et d'un algorithme d'agrégation ou de désagrégation qui permettent d'obtenir une partition ou un arbre de classification.

On s'efforcera ici de décrire les grands traits de la nouvelle école française d'analyse factorielle ; cette école, où existent des tendances [8] [11] [30], mais qui est dominée incontestablement par la personnalité de J.P. BENZECRI [6] [7] est caractérisée :

- par une grande méfiance vis-à-vis de la statistique inférentielle classique : remise en cause du modèle probabiliste a priori et en particulier du modèle gaussien ; priorité aux données et particulièrement aux données multidimensionnelles.

- par un retour à la géométrie : abandon du langage de la statistique mathématique, position critique à l'égard du langage matriciel ; exploitation systématique de la dualité.

- par la diversité des problèmes abordés : hier instrument privilégié du chercheur en sciences sociales, l'analyse factorielle est devenue un outil d'investigation ordinaire qui est utilisé sans difficulté, grâce aux programmes existants, par le médecin, l'ingénieur, le vendeur, le gestionnaire, le journaliste, etc.

- par la place privilégiée occupée par certaines techniques d'analyse factorielle : l'analyse factorielle des correspondances très prisée en France n'est pratiquement pas utilisée dans les autres pays et cela particulièrement dans les pays anglo-saxons (malgré un article de M.O. HILL [21]) ; on ne fait guère plus appel à l'analyse factorielle au sens de SPEARMAN.

- par certaines pratiques particulières : recours systématiques à la "pratique des points supplémentaires" qui permet en particulier d'effectuer des régressions graphiques au moindre coût ; utilisation d'indices permettant de juger de la stabilité des analyses ou de l'influence de certains éléments, etc.

Après un récit bref sur l'évolution des idées et des techniques en analyse factorielle qui permettra de bien différencier l'analyse factorielle au sens de SPEARMAN (école psychométrique américaine) des techniques relevant de l'analyse en composantes principales ou de l'analyse canonique, on propose, en s'appuyant sur les types de tableaux qu'il est courant de leur soumettre, une classification sommaire de ces dernières. On introduit alors le langage géométrique du schéma de dualité, langage qui permet de présenter de façon élégante et concise les principales stratégies et les techniques qui s'en déduisent, utilisées à l'heure actuelle en analyse factorielle.

1. – FRAGMENTS D'HISTOIRE

Etant donnée une matrice de corrélation (p x p) R, existe-t-il de façon unique deux matrices (p x p) V_1 et V_2, où V_1 est symétrique semi-définie positive de rang k (inférieur à p) et V_2 est diagonale à coefficients positifs, telles que :

$$R = V_1 + V_2 \ (^1)$$

C'est sur ce problème de décomposition d'une forme quadratique définie positive qu'est tombé C. SPEARMAN en 1904 [40] quand il a cherché à exprimer de façon mathématique sa conception unidimensionnelle de l'intelligence. Pour SPEARMAN le "degré de dépendance" [5] entre les variables qu'il considérait (les réponses à différents tests), qui n'est autre que le rang de V_1, devait être égal à 1 ; aussi cette décomposition de R n'était-elle possible que si tous les déterminants internes, sans termes provenant de la diagonale de R, d'ordre 2 étaient nuls (différences tétrades). Ainsi s'est trouvée fondée la première école psychométrique américaine *d'analyse factorielle.*

Le problème que se posait SPEARMAN et que se sont posé après lui ses nombreux disciples [1] [19] [25] [28] [42] entre dans la catégorie des problèmes que l'on peut se poser quand on désire résumer à l'aide d'un petit nombre de dimensions une information multidimensionnelle ; la réduction recherchée "réduire le rang de R" se faisait dans une certaine optique : "reconstituer au mieux les corrélations entre les tests" compte tenu d'idées a priori bien affirmées sur le phénomène étudié.

Karl PEARSON en 1901 [34] puis Harold HOTELLING en 1933 [22] ont abordé de façon différente la recherche de ce petit nombre de dimensions permettant de reconstituer l'information initiale ; le premier article de K. PEARSON s'intitulait "On lines and planes of closest fit to systems of points in space" : les individus-colonnes du tableau à analyser étant considérés comme des vecteurs d'un espace à p dimensions, on proposait de réduire la dimension de l'espace en projetant le nuage des points individus sur le sous-espace de dimension k (k petit fixé) permettant d'ajuster au mieux le nuage. Ainsi était inventée l'*"analyse en composantes principales"*.

La recherche géométrique que proposait K. PEARSON, qui ne faisait aucune hypothèse, au contraire de SPEARMAN, sur la nature de la distribution associée aux variables considérées (le modèle gaussien que postulent SPEARMAN et ceux qui l'ont suivi leur permet à la fois, raisonnant sur des échantillons, d'inférer et de traduire covariances et corrélations en termes de dépendance ou d'indépendance statistique), et qui conduit, si les variables sont réduites, à extraire les valeurs et vecteurs propres de la matrice des corrélations, ne conduit pas en général à la solution, k étant fixé, au problème de SPEARMAN.

---

(1) On consultera avec intérêt l'article de A. KOBILINSKY publié dans ce même numéro.

On pourra s'en convaincre aisément en traitant du modèle à corrélations égales (distribution cylindrique des points individus dans l'espace à p dimensions). Si tous les coefficients de corrélation sont égaux à r, la matrice (p x p) R s'écrit :

$$R = r\,\underline{j}\,\underline{j}' + (1 - r)\,I_p$$

où $\underline{j}$ est le vecteur (colonne) dont toutes les coordonnées sont égales à 1, I_p désignant la matrice identité de rang p.

Si r est positif, la matrice $V_2 = (1 - r)\,I_p$ étant bien diagonale, le degré de dépendance k, qui n'est autre que le rang de la matrice $V_1 = r\,j\,j'$, est égal à 1 ; la seule décomposition du type $R = V_1 + V_2$ respectant les conditions de SPEARMAN est, si p est au moins égal à 3, la décomposition précédente. Il existe donc bien un "facteur général" : une seule dimension suffit pour expliquer les corrélations constatées entre les variables considérées.

Le sous-espace propre associé à la valeur propre $\lambda_1 = 1 + (p - 1)\,r$ de R n'est autre que la droite engendrée par $\underline{j}$; la valeur propre $\lambda_2 = 1 - r$, dont l'ordre de multiplicité est égal à $p - 1$, admet pour sous-espace propre l'hyperplan orthogonal à $\underline{j}$. Aussi l'analyse en composantes principales fournit-elle la décomposition $R = V_1 + V_2$, où les matrices :

$$V_1 = \frac{1 + (p - 1)\,r}{p}\,\underline{j}\,\underline{j}' \quad \text{et} \quad V_2 = \frac{r - 1}{p}\,\underline{j}\,\underline{j}' + (1 - r)\,I_p$$

représentent les formes quadratiques d'inertie des nuages obtenus en projetant les points individus sur l'axe engendré par $\underline{j}$ et l'hyperplan orthogonal à $\underline{j}$ respectivement ; V_2 est ici de rang $p - 1$.

Pour PEARSON l'axe engendré par $\underline{j}$ n'est le premier axe principal que si r est positif ($\lambda_1 > \lambda_2$) ; même si, r étant positif, on se contente d'une représentation unidimensionnelle, le meilleur ajustement du nuage conduit à une décomposition de R qui n'est pas celle que propose l'analyse de SPEARMAN.

Même si les problèmes respectivement posés par SPEARMAN et PEARSON sont différents, l'analyse en composantes principales fournit en général une excellente approximation de ce que l'on recherche en *"analyse factorielle au sens de SPEARMAN"*, et cela quoique la solution de PEARSON, où la somme des rangs de V_1 et V_2 est égale à p, ne puisse théoriquement pratiquement jamais (il faudrait que V_2 soit de rang $p - k$) coïncider avec une solution au problème de SPEARMAN. La similitude des objectifs (réduire), la proximité entre les résultats obtenus ont conduit à bien des controverses. Le succès que connait particulièrement à l'heure actuelle la solution de PEARSON résulte de sa simplicité ; elle ne pose pas de problème d'unicité [1] et est obtenue sans avoir recours à d'autres itérations que celles exigées par l'extraction des valeurs et vecteurs propres.

Si on dispose non pas de variables quantitatives mais de variables dichotomiques (variables qualitatives à deux modalités) on peut aborder avec les *"structures latentes"* introduites par P. LAZARSFELD [29] un problème tout à fait analogue à celui qu'avait posé SPEARMAN : existe-t-il une variable qualitative (ses modalités définiront les classes latentes) telle que, conditionnellement à cette variable, les variables dichotomiques considérées soient indépendantes ? Opérant ici directement sur les probabilités, l'existence et l'unicité de la solution créent encore bien des difficultés.

K. PEARSON et H. HOTELLING n'ont vu l'analyse en composantes principales que sous son aspect le plus restrictif : les points individus étaient tous munis des mêmes poids ; la distance euclidienne classique était considérée. Le jour où l'on a pris conscience que l'on pouvait jouer à la fois sur les poids et sur la métrique – en France l'influence de J.P. BENZECRI a été à ce niveau déterminante [6] – on s'est aperçu qu'en fait l'analyse en composantes principales recouvrait tout un ensemble de techniques, que nous dirons d'*analyse factorielle*, répondant à des objectifs apparemment différents [11]. C'est à ces analyses factorielles que désormais nous nous intéresserons.

L'étude géométrique des positions relatives de deux sous-espaces d'un espace euclidien que nous appelons *analyse canonique* [11] a été introduite par H. HOTELLING [23] en 1936 ; l'objectif initial était d'exprimer au mieux à l'aide d'un petit nombre de couples de variables la liaison entre deux ensembles de caractères quantitatifs. Compte tenu des moyens de calcul dont on disposait et des tendances dominantes de l'époque dans le domaine de la statistique (on était à l'apogée de la statistique inférentielle anglo-saxonne) l'accent a été mis fort longtemps en analyse canonique, non pas sur les procédés graphiques qu'elle offrait à la statistique descriptive (et l'on sait combien ces procédés sont utilisés maintenant en analyse des données), mais sur les cosinus (les coefficients de corrélation canonique, généralisations des coefficients de corrélation et de corrélation multiple) qu'elle permettait de calculer et qui fournissaient un moyen, connaissant la distribution du plus grand d'entre eux [35] [20], de juger de la significativité des liaisons entre les deux paquets de variables considérées sous hypothèse de normalité des distributions.

C'est à ce même problème géométrique que fait référence l'*analyse factorielle discriminante* [18] [36] [37] qui permet de décrire la liaison entre une variable qualitative et un ensemble de variables quantitatives ; ici les cosinus sont traduits en termes de rapport de corrélation.

L'analyse factorielle des correspondances sur tableau de contingence qui a été introduite en 1962 par J.P. BENZECRI [6] [7] peut être encore considérée comme une variante de l'analyse canonique ; ici les sous-espaces dont il s'agit de préciser les positions relatives sont caractéristiques des deux variables qualitatives dont on se propose d'analyser les liaisons. L'originalité en 1962 des propositions de J.P. BENZECRI ne résidait que très partiellement dans la technique qui était proposée ; il s'agissait beaucoup plus d'une remise en cause radicale des schémas de pensée et des pratiques ayant cours à cette époque en France. En effet le livre de chevet du statisticien qui analysait plus de deux variables étant alors le fameux "Introduction to multivariate analysis" de T.W. ANDERSON [2], l'analyse des données multidimensionnelles ne se concevait pratiquement alors que dans le cadre de la statistique inférentielle en univers gaussien. Ce fut à proprement parler, et les aveugles de l'époque (nous en fûmes) l'ont ressenti plus que tout autre, une révolution dans le petit monde de la statistique française : le modèle gaussien et au-delà le modèle probabiliste devinrent les symboles d'une pratique où l'on déclare vert ce que l'on sait pertinemment ne pas l'être, où l'on s'interroge sur des problèmes annexes d'existence (x est-il lié à y ?) sans jamais regarder finement ce qui est (comment x est-il lié à y ?) où l'on discute doctement sur des préalables (échantillonnage) en hésitant à prendre le taureau par les cornes (manipuler les données existantes)... etc. La position positiviste de la nouvelle vague a parfois choqué : affirmait-on sa naïveté ou voulait-on la polémique ? Elle a eu le mérite de provoquer un salutaire retour aux sources qui s'imposait d'ailleurs compte tenu des possibilités nouvelles qu'offraient les moyens de stockage et de calcul puissants.

L'extension du problème de HOTELLING à l'étude des positions relatives de k sous-espaces vectoriels [12] [26] a conduit à des pratiques particulières dont l'*analyse factorielle des correspondances sur tableau disjonctif* est, dans le cas des variables qualitatives, une illustration. Ce fait que l'analyse factorielle des correspondances permette de traiter, de façon tout à fait justifiée, aussi bien les tableaux de contingence que les tableaux logiques (ils sont considérés dans l'analyse comme s'ils étaient des tableaux de contingence) et de faire face encore à d'autres situations [14], a pu faire dire à certains qu'elle était "universelle".

C'est en 1969 [16] qu'Y. ESCOUFIER proposa de manipuler les tableaux de données comme on manipule les lignes et les colonnes de ces tableaux en leur associant des *opérateurs* considérés comme vecteurs d'un espace euclidien ; il est intéressant de noter qu'ESCOUFIER ne s'est aperçu que bien après de la généralité de ses propositions simplement par le fait que, raisonnant dans l'espace des variables aléatoires du second ordre (espace L^2), il utilisait à l'époque le langage du calcul des probabilités. Le jour où il a adopté le "schéma de dualité" [11] comme instrument clef d'expression, c'est-à-dire le jour où il est revenu à la géométrie en se donnant la possibilité de jouer sur les métriques dans les deux espaces en dualité où "individus" et "caractères" sont respectivement représentés comme des points, il s'est rendu compte que ses opérateurs, qui permettaient de comparer différents types de tableaux de données dans une grande diversité d'optiques, constituaient une sorte de méta langage pour discuter fort simplement de toutes les techniques d'analyse factorielle (on ne parle pas ici de l'analyse au sens de SPEARMAN) et que le produit scalaire qu'il avait introduit pour mesurer les angles entre opérateurs généralisait les différents indices que l'on trouve en statistique pour mesurer des angles, (coefficients et rapports de corrélations ; chi-deux) [33].

Signalons que la psychométrie américaine a connu, plus tôt, une évolution comparable par certains côtés (remise en cause implicite du modèle probabiliste a priori) à celle de la statistique multidimensionnelle française. Dans ses écrits W.S. TORGERSON [42] indique comment procéder pour effectuer une analyse factorielle directement sur un tableau de distances ou plus généralement de dissimilarités ; nous appelons cette technique, qui répond à un problème posé depuis longtemps chez les psychométriciens [46], *"analyse factorielle sur tableau de distances"* [11]. On met ici l'accent, exploitant la dualité entre espace des observations (individus) et espace des variables [45], sur le fait qu'il est possible d'extraire, de façon hiérarchique, des descripteurs quantitatifs de la seule information constituée par un tableau décrivant des proximités entre objets ; on mesure tout l'intérêt de la méthode proposée si ces objets sont "caractéristiques" du phénomène étudié, c'est-à-dire si l'explication du phénomène peut être recherchée dans les différences entre objets. Un problème de même nature a été abordé par J.D. CARROLL [13] : les tableaux de dissimilarités D_1, D_2, ..., D_k de taille (n x n), caractérisant les perceptions que manifestent k juges à propos de n objets, peuvent-ils être considérés comme résultant de métriques différentes appliquées sur une même batterie de descripteurs (variables) quantitatifs ? Ce problème, qu'il est facile de traiter en termes d'opérateurs, fait référence à une équivalence entre tableaux de dissimilarités identique à celle qui est considérée en analyse canonique. La manière d'aborder l'analyse des proximités a fait un pas en avant en 1962 avec R.N. SHEPARD [38] [39] qui a montré que l'information contenue dans un tableau de dissimilarités résidait pour l'essentiel dans l'ordre dans lequel se rangent ces dissimilarités ; l'équivalence qui en résulte joue un rôle fondamental en classification automatique [24]. Empruntant aux idées de SHEPARD, J.B. KRUSKAL [27] a alors proposé une technique nouvelle pour

construire une image euclidienne à partir d'un tableau de dissimilarités ne prenant en compte que l'ordre sur les dissimilarités ; cette technique connaît à l'heure actuelle aux Etats-Unis, particulièrement dans le domaine des études de marché, un succès considérable. Il faut encore citer parmi ceux qui ont fait beaucoup pour l'évolution des pratiques statistiques aux Etats-Unis J.W. TUKEY [43] qui est l'auteur d'un ensemble de propositions originales et simples concernant la description des données ; le titre de l'un de ses derniers ouvrages "Exploratory data analysis" [44] indique clairement quel est le rôle dans l'esprit de TUKEY des techniques descriptives d'analyse de données (et aussi des techniques rapides d'estimation) dont font partie les analyses factorielles : si elles permettent de s'interroger sur les phénomènes étudiés, et cela dans un certain ordre, les faits mis en évidence, quand on ne travaille que sur des fragments (échantillons) et non sur des totalités (populations), ne permettent en toute rigueur que d'émettre des hypothèses qu'il reste encore à valider ; valider les hypothèses est le rôle dévolu aux techniques dites "confirmatoires" (confirmatory data analysis) qui, si elles consistent souvent à refaire les calculs sur des sous-échantillons ou sur des échantillons simulés (on peut citer par exemple les épreuves de validation proposées par L. LEBART en analyse factorielle des correspondances [6] [31]), laissent malgré tout une place beaucoup plus grande aux enseignements de la statistique mathématique.

2. – GRANDS TYPES DE TABLEAUX DE DONNEES ET CLASSIFICATION SOMMAIRE DES TECHNIQUES D'ANALYSE FACTORIELLE (RAPPEL)

Nous nous restreignons ici aux techniques d'analyse factorielle relevant de l'analyse en composantes principales ou de l'analyse canonique et conduisant à des solutions en termes de valeurs et de vecteurs propres.

2.1. – Types principaux de tableaux se prêtant directement à une analyse factorielle

On peut distinguer trois types principaux :

- *Les tableaux "individus x caractères"*

Ces tableaux à p lignes (caractères) et n colonnes (individus) seront analysés différemment suivant que les caractères sont :

- quantitatifs,
- dichotomiques (caractères qualitatifs à deux modalités),
- qualitatifs.

Dans le dernier cas le tableau "individus x caractères" est construit en empilant les unes sur les autres les variables indicatrices associées aux modalités (variables valant 1 ou 0 suivant que l'individu prend ou ne prend pas la modalité) ; ce tableau "disjonctif" est dit "complet" si toutes les modalités sont considérées.

- *Les mesures définies sur un produit cartésien*

Ici les lignes ou les colonnes, qui jouent des rôles symétriques, peuvent être réunies par sommation, les nombres positifs constituant ces tableaux pouvant être interprétés comme des masses. Les tableaux de contingence, qui résultent du croisement de deux variables qualitatives, appartiennent à cette famille.

• *Les tableaux de dissimilarités ou de distances*

Ces tableaux carrés symétriques décrivent les proximités entre n objets ; si on a affaire à des dissimilarités, l'indice d considéré ne vérifie, pour tout i et pour tout couple (i , i'), que les deux propriétés :

$$d_{ii} = 0 \; ; \; d_{ii'} = d_{i'i} > 0.$$

2.2 Classification sommaire des techniques

Avec les techniques d'analyse factorielle on dispose de moyens de contrôle, et de réduction efficaces des données analogues aux histogrammes en univers unidimensionnel : une formule simple permet de reconstituer de façon approximative le tableau initial à partir des dimensions retenues, dimensions qui donnent souvent un éclairage nouveau sur le phénomène étudié.

Contrôler, réduire, extraire les dimensions principales sous-jacentes au phénomène peuvent donc être considérés comme des objectifs communs aux analyses factorielles.

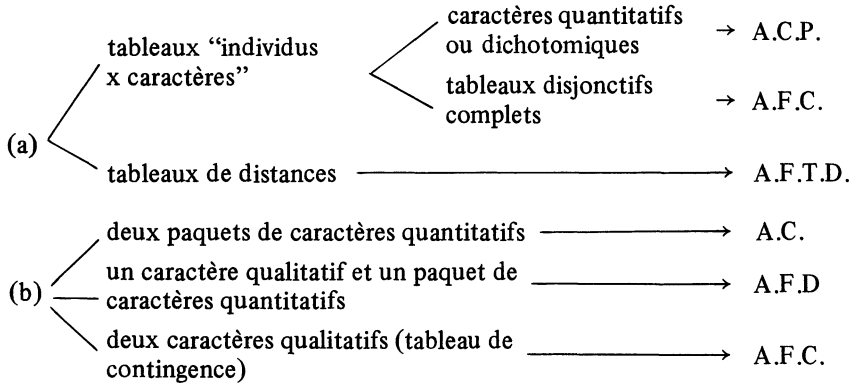
Les graphiques issus des analyses factorielles offrent des descriptions simples des lignes ou des colonnes des tableaux considérés ; en se référant aux notions d'"individu" (ou objet, ou observation...) et de "caractère" (ou variable, ou dimension...), qui induisent des équivalences de nature différente, on peut donc distinguer les deux objectifs suivants :

(a) : décrire au mieux les proximités entre individus
(b) : décrire au mieux les liaisons entre caractères.

Si on adopte les sigles :

A.C.P. pour analyse en composantes principales
A.C. pour analyse canonique
A.F.D. pour analyse factorielle discriminante
A.F.C. pour analyse factorielle des correspondances
A.F.T.D. pour analyse factorielle sur tableau de distances

la classification suivante des techniques usuelles, qui ne reflète que de loin les pratiques, peut être alors proposée :



Bien des transformations peuvent être effectuées sur les données, on passera très souvent par exemple, pour explorer des liaisons plus générales que les liaisons linéaires ou fonctionnelles, du quantitatif au qualitatif.

3. GEOMETRIE EN ANALYSE FACTORIELLE

Cette partie reprend très largement un exposé effectué à Uppsala [17] : il s'agit de bien mettre en évidence en utilisant le langage de la géométrie les rapports entre les différentes analyses. Une attention plus particulière est portée sur l'analyse factorielle des correspondances.

3.1 Le langage du schéma de dualité

3.1.1 Le triplet (X, M, D_p)

La donnée d'un tableau individus x caractères X va toujours de pair en analyse factorielle avec la donnée de deux métriques euclidiennes :

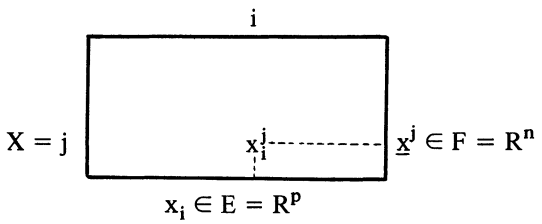
– la première, représentée par la matrice définie positive M (p x p) permet de mesurer les proximités entre les n individus colonnes ;

– ayant muni les individus des poids $p_i (p_i > 0 ; \Sigma\, p_i = 1)$, la seconde est représentée par la matrice diagonale définie positive $D_p (n \times n)$ dont les éléments diagonaux sont les p_i. Les produits scalaires et les cosinus des angles entre les caractères centrés (de moyenne nulle) ne sont autres alors que les covariances et les corrélations.

Nous parlerons le plus souvent non pas du tableau X, mais du triplet (X, M, D_p).

3.1.2 Un bien joli schéma : le schéma de dualité

Voici un tableau "individus x caractères" X(p x n) :



Notons $\{\underline{e}_j / j = 1, .., p\}$ et $\{\underline{f}_i / i = 1, .., n\}$ les bases canoniques des espaces euclidiens E (métrique M) et F (métrique D_p) de dimensions respectives p et n.

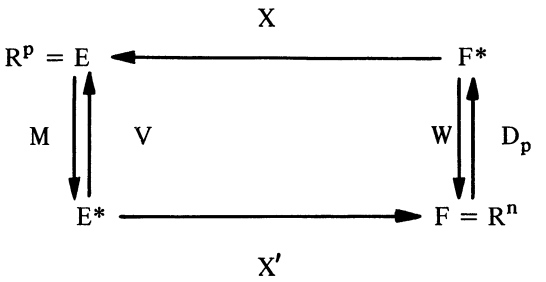
Les caractères peuvent être représentés, si on note $\{\underline{e}_j^* / j = 1, .., p\}$ la base duale du dual E* de E :

– dans F par les vecteurs $\underline{x}^j$ dont les coordonnées sont les données x_i^j ;
– dans E par les vecteurs de base $\underline{e}_j$;
– dans E* par les vecteurs de base $\underline{e}_j^*$.

De même les individus peuvent être représentés, si on note $\{\underline{f}_i^* / i = 1, .., n\}$ la base duale du dual F* de F :

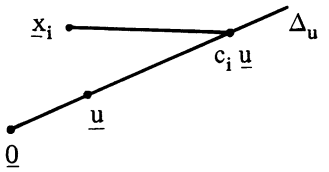
- dans E par les vecteurs $\underline{x}^i$ dont les coordonnées sont les x_i^j ;
- dans F par les vecteurs de base $\underline{f}_i$;
- dans F* par les vecteurs de base $\underline{f}_i^*$.

Au triplet (X, M, D_p) on peut donc associer le schéma suivant dit "de dualité" [11] :



Le tableau de données X est ici considéré comme la matrice d'une application linéaire qui associe les deux représentations $\underline{f}_i^*$ et $\underline{x}_i$ d'un même individu, la transposée X' associant les représentations $\underline{e}_j^*$ et $\underline{x}^j$ d'un même caractère.

La métrique M apparaît dans le schéma sous la forme d'un isomorphisme : si $\underline{u}$ est un vecteur normé de E, la forme linéaire $M\underline{u}$ fournit les coordonnées par rapport à $\underline{u}$ des projections sur la droite Δ_u engendrée par $\underline{u}$.



$$c_i = < M\underline{u}, \underline{x}_i >$$

A l'application $W = X' M X$ est associé un "écart" euclidien (la matrice W n'est en général que semi définie positive) qui permet de mesurer dans F* des angles et des distances entre vecteurs $\underline{f}_i^*$ identiques à ceux que l'on mesure dans E entre vecteurs $\underline{x}_i$ avec la métrique M.

De façon symétrique la métrique D_p apparaît sous la forme d'un isomorphisme, elle induit sur E* l'écart euclidien représenté dans le schéma par l'application $V = X D_p X'$ qui permet de calculer dans E* des produits scalaires et des distances identiques à celles que l'on calcule dans F. Si les caractères sont centrés, la matrice V n'est autre que la matrice des variances-covariances.

3.1.3 Quelques acrobaties

3.1.3.1 Opérateurs de projection (projecteurs)

Le projecteur A sur le sous-espace X'(E*) des combinaisons linéaires des caractères $\underline{x}^j$, s'écrit, s'il est de dimension p :

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

Si on prend pour métrique M la métrique de "Mahalanobis" $M = V^{-1}$, l'opérateur $W D_p$ n'est autre que A.

3.1.3.2 Aperçu sur l'analyse en composantes principales

Les *axes principaux* Δ_{uj}, axes les plus proches du nuage des individus $\underline{x}_i$ au sens de la métrique M, sont définis par les équations :

$$V M \underline{u}_j = \lambda_j \underline{u}_j$$

$$||\underline{u}_j||_M = 1$$

Ces axes définissent une base M orthonormée et V^{-1} orthogonale (si V est inversible) dans l'espace E ; la valeur propre λ_j est égale au moment d'inertie par rapport à l'hyperplan orthogonal à Δ_{uj}.

Les *facteurs principaux* $\underline{v}_j = M \underline{u}_j$, qui permettent de calculer les coordonnées dans le système des axes principaux, vérifient les équations :

$$M V \underline{v}_j = \lambda_j \underline{v}_j$$

$$||\underline{v}_j||_{M^{-1}} = 1 \; ; \; ||\underline{v}_j||_V = \sqrt{\lambda_j}$$

Ces facteurs définissent dans E* une base M^{-1} orthonormée et V-orthogonale.

Les *composantes principales* $\underline{c}^j = X' \underline{v}_j = X' M \underline{u}_j$, dont les coordonnées ne sont autres que celles des individus $\underline{x}_i$ dans la base des axes principaux, vérifient les équations :

$$W D_p \underline{c}^j = \lambda_j \underline{c}^j$$

$$||\underline{c}^j||_{D_p} = \sqrt{\lambda_j}$$

Les composantes principales définissent une base D_p-othogonale dans le sous-espace X'(E*) de F.

Remarque : vers l'analyse factorielle sur tableau de distances

La matrice W peut être calculée directement à partir des distances $d_{ii'} = ||\underline{x}_i - \underline{x}_{i'}||$ et des poids p_i à l'aide de la formule [42] :

$$w_{ii'} = \frac{1}{2} (d^2_{i.} + d^2_{i'.} - d^2_{..} - d^2_{ii'})$$

avec :

$$d^2_{i.} = \Sigma \, p_{i'} \, d^2_{ii'} \; ; \; d^2_{..} = \Sigma \, p_i p_{i'} d^2_{ii'}$$

Si on note D le tableau (n x n) des distances $d_{ii'}$ cette formule montre que pour effectuer l'analyse en composantes principales la donnée du triplet (X, M, D_p) est équivalente à la donnée du couple (D, D_p).

3.1.3.3 Jeux de métriques

L'analyse en composantes principales sur variables réduites, qui conduit à la diagonalisation de la matrice des corrélations R, correspond au choix pour M, X étant centré, de la matrice D_{1/σ^2} des inverses des carrés des écarts-types.

En analyse factorielle au sens de SPEARMAN, la variante de JORESKOG [25], choisissant pour M la matrice diagonale diag (V^{-1}) dont la diagonale n'est autre que celle de V^{-1}, revient à une analyse en composantes principales.

L'espace E étant muni de la métrique de Mahalanobis $M = V^{-1}$, l'analyse factorielle discriminante n'est autre qu'une analyse en composantes principales effectuée sur un nuage de centres de gravité muni des poids associés aux groupes d'individus considérés [11] etc.

3.1.4 Comparaison de tableaux et opérateurs

L'opérateur $U = W D_p$, caractéristique du triplet (X, M, D_p) et du couple (cf. 3.1.3.2) (D, D_p), a été introduit par Y. ESCOUFIER [16] pour comparer des triplets (X, M, D_p) ou des couples (D, D_p). Le sous-espace G des applications linéaires D_p-symétriques de F dans F étant muni du produit scalaire P défini à partir de la trace (Tr) :

$$U \in G \quad V \in G \quad \Rightarrow \quad P(U, V) = Tr(UV)$$

On peut alors manipuler les opérateurs, donc les triplets (X, M, D_p) ou les couples (D, D_p), comme on manipule les lignes ou les colonnes des tableaux, à l'aide de l'analyse factorielle [11] [33].

3.2 Stratégies Π et Σ

3.2.1 Positions relatives de deux sous-espaces vectoriels

L'étude des positions relatives de deux ou de k sous-espaces vectoriels est un problème géométrique central en analyse factorielle.

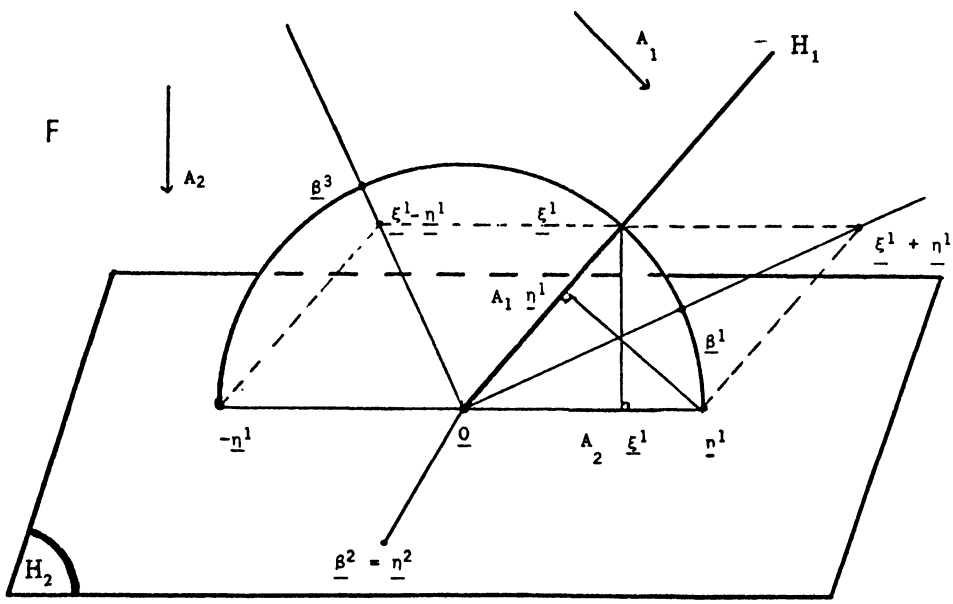


FIGURE 1

Pour caractériser les positions relatives de deux sous-espaces vectoriels H_1 et H_2 d'un espace euclidien F (les sous-espaces qui leur sont orthogonaux sont respectivement notés $H_1^{\perp}$ et $H_2^{\perp}$) on dispose de deux stratégies :

STRATEGIE Π : chercher respectivement dans H_1 et H_2 les vecteurs $\underline{\xi}^1$ et $\underline{\eta}^1$ normés faisant un angle minimum ; puis chercher, toujours dans ces sous-espaces, les vecteurs normés $\underline{\xi}^2$ et $\underline{\eta}^2$, orthogonaux aux précédents, faisant un angle minimum, etc.

STRATEGIE Σ : chercher dans F, le vecteur normé $\underline{\beta}^1$ équidistant de H_1 et H_2 faisant un angle minimum avec ces deux sous-espaces ; puis chercher $\underline{\beta}^2$ dans F, normé et orthogonal à $\underline{\beta}^1$, équidistant de H_1 et H_2, faisant un angle minimum avec ces deux sous-espaces, etc.

Si les projecteurs (cf. 3.1.3.1) associés aux sous-espaces H_1 et H_2 sont notés A_1 et A_2, la stratégie Π conduit aux équations :

$$A_1 A_2 \underline{\xi}^j = \lambda_j \underline{\xi}^j$$

$$\underline{\eta}^j = \frac{1}{\sqrt{\lambda_j}} A_2 \underline{\xi}^j$$

$$||\underline{\xi}^j|| = ||\underline{\eta}^j|| = 1$$

Les valeurs propres λ_j ne sont autres que les carrés des cosinus des angles entre les $\underline{\xi}^j$ et les $\underline{\eta}^j$.

La stratégie Σ conduit aux équations :

$$(A_1 + A_2)\underline{\beta}^j = \mu_j \underline{\beta}^j$$

$$||\underline{\beta}^j|| = 1.$$

En tirant les vecteurs propres $\underline{\beta}$ de l'opérateur $A_1 + A_2$ on obtient en réalité (figure 1) un système orthonormé de vecteurs situés :

– soit dans $H_1 \cap H_2$; la valeur propre μ est alors égale à 2 ;

– soit à égale distance de H_1 et H_2, l'angle formé avec ces sous-espaces étant inférieur à 45 degrés ; homothétiques aux vecteurs $(\underline{\xi} + \underline{\eta})$, il leur est associé une valeur propre μ comprise entre 1 et 2 ($\mu = 1 + \sqrt{\lambda}$) ;

– soit dans $H_1^{\perp} \cap H_2$ ou $H_1 \cap H_2^{\perp}$, la valeur propre associée est égale à 1 ;

– soit à égale distance de H_1 et H_2, l'angle formé avec ces deux sous-espaces étant supérieur à 45 degrés ; homothétiques aux vecteurs $(\underline{\xi} - \underline{\eta})$, il leur est associé une valeur propre μ comprise entre 0 et 1 ($\mu = 1 - \sqrt{\lambda}$) ;

– soit dans $H_1^{\perp} \cap H_2^{\perp}$; la valeur propre μ associée est alors nulle.

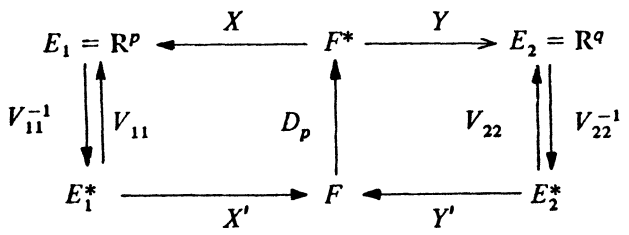
En analyse factorielle, les stratégies Π et Σ correspondent à deux optiques différentes ; l'optique "analyse canonique" pour la première, l'optique "analyse en composantes principales" en ce qui concerne la seconde.

La stratégie Σ se généralise immédiatement. Pourquoi ne pas caractériser les positions relatives de k sous-espaces vectoriels H_i par les vecteurs normés $\underline{\beta}$ obtenus en diagonalisant la somme des projecteurs ΣA_i ? Nous appellons, ne suivant en cela ni CARROLL ni KETTENRING, *analyse en composantes principales* (et non analyse canonique) *réduite généralisée* la technique conduisant à diagonaliser ΣA_i [12] [26].

Pratiquement toutes les techniques d'analyse factorielle conduisant à des solutions en terme de vecteurs et de valeurs propres peuvent être considérées comme découlant des deux stratégies précédentes.

3.2.2 *Stratégie* Π *et variables quantitatives*

C'est la stratégie Π que l'on utilise de façon classique en analyse canonique [11] quand on cherche à analyser les liaisons entre deux paquets de caractères quantitatifs (tableaux X, (p x n) et Y, (q x n)). Le double schéma de dualité suivant est alors considéré :



Si les caractères sont centrés les matrices $V_{11} = X\, D_p\, X'$ et $V_{22} = Y\, D_p\, Y'$ sont les matrices des variances et covariances associées respectivement aux $\underline{x}^i$ et aux $\underline{y}^j$. La matrice $V_{12} = X\, D_p\, Y' = V'_{21}$ désigne alors la matrice des covariances entre les $\underline{x}^i$ et les $\underline{y}^j$. Les projecteurs A_1 et A_2 sur les sous-espaces H_1 et H_2 de F engendrés par les caractères sont les opérateurs caractéristiques des triplets (X, V_{11}^{-1}, D_p) et (Y, V_{22}^{-1}, D_p) (cf. 3.1.4) :

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

Aux caractères canoniques $\underline{\xi}^j$ et $\underline{\eta}^j$ que l'on peut obtenir directement par les équations données au paragraphe 3.2.1, on fait correspondre les facteurs canoniques $\underline{a}_j \in E_1^*$ et $\underline{b}_j \in E_2^*$:

$$\underline{\xi}^j = X'\, \underline{a}_j \; ; \; \underline{\eta}^j = Y'\, \underline{b}_j.$$

Ces facteurs canoniques sont obtenus directement par les équations :

$$B_1\, B_2\, \underline{a}_j = \lambda_j\, \underline{a}_j$$

$$\underline{b}_j = \frac{1}{\sqrt{\lambda_j}} B_2\, \underline{a}_j$$

$$\|\underline{a}_j\|_{V_{11}} = \|\underline{b}_j\|_{V_{22}} = 1$$

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

3.2.3 *Stratégie* Σ *et variables quantitatives*

Rechercher les vecteurs propres de $A_1 + A_2$ revient à effectuer l'analyse en composantes principales du triplet (Z, M, D_p) avec :

$$Z = \begin{pmatrix} X \\ Y \end{pmatrix} \quad \text{et} \quad M = \begin{pmatrix} V_{11}^{-1} & 0 \\ 0 & V_{22}^{-1} \end{pmatrix}$$

En effet, les composantes principales $\underline{c}^j$ associées au triplet (Z, M, D_p) qui vérifient :

$$(A_1 + A_2)\,\underline{c}^j = \mu_j\,\underline{c}^j$$

$$\|\underline{c}^j\|_{D_p} = \sqrt{\mu_j}$$

sont les homothétiques des vecteurs normés $\underline{\beta}^j$ définis au paragraphe 3.2.1 :

$$\underline{c}^j = \sqrt{\mu_j}\,\underline{\beta}^j.$$

De façon générale, quand on cherche à analyser k tableaux de données X_i en diagonalisant la somme des projecteurs $\Sigma\, A_i$ on effectue en fait l'analyse en composantes principales du triplet (Z, M, D_p) avec :

$$Z = \begin{bmatrix} X_1 \\ X_2 \\ \cdot \\ \cdot \\ \cdot \\ X_k \end{bmatrix} \quad \text{et} \quad M = \begin{bmatrix} V_{11}^{-1} & & & 0 \\ & V_{22}^{-1} & & \\ & & \cdot & \\ & & & \cdot \\ & & & \cdot \\ 0 & & & V_{kk}^{-1} \end{bmatrix}$$

L'analyse en composantes principales sur variables réduites (métrique D_{1/σ^2}) apparaît donc bien comme un cas particulier de cette anlyse en composantes principales réduite généralisée.

3.3 Sur le caractère universel de l'analyse factorielle des correspondances

3.3.1 Représentation des variables qualitatives : codage et dualité

Les variables qualitatives sont considérées ici comme permettant de décrire un ensemble Ω de n individus munis de poids p_ω :

$$\Omega = \{\omega/\omega = 1, \ldots, n\}$$

$$p_\omega > 0, \Sigma\, p_\omega = 1$$

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

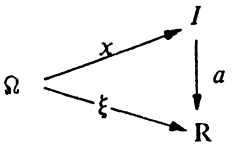
où : a est l'application codage : $I \xrightarrow{a} R$

x^i est la variable indicatrice associée à la $i^{ième}$ modalité de x :

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

$$x^i(\omega) = 0 \text{ sinon}$$

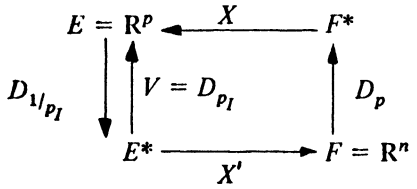
$a^i = a(i)$ est le codage numérique de la $i^{ième}$ modalité de x.



Si l'application a est injective, les variables x et ξ qui induisent sur Ω la même partition sont équivalentes ; les codages sont alors tous différents. A l'ensemble des variables indicatrices est associé le *tableau disjonctif* (tableau logique) :

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

L'espace F étant muni de la distance "en moyenne quadratique" D_p, on a le schéma de dualité suivant :



Si $\mathcal{P}(I)$ désigne l'ensemble des parties de I :

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

- p_I désigne la loi de probabilité sur $(I, \mathcal{P}(I))$ définie par les probabilités p_i de prendre les différentes modalités i de x.

- Le dual E* de E est considéré comme l'espace des variables définies sur I.

- La métrique induite sur E* par D_p, représentée dans le schéma par l'application $V = XD_pX'$, admet pour matrice D_{p_I}, matrice diagonale dont les éléments diagonaux sont les p_i.

- Restreinte au simplexe des lois de probabilités la métrique D_1/p_I induite sur E par D_{p_I} (métrique V^{-1} de Mahalanobis) est ce que l'on appelle *"la distance du chi-deux de centre* p_I'' [6] [11].

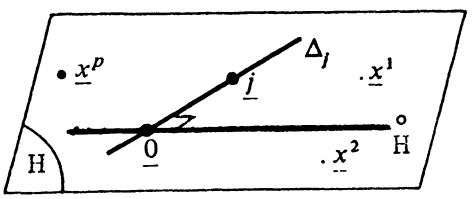
- W D_p est le projecteur A sur $H = X'(E^*)$

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

$$H = X'(E^*) = \{\underline{\xi} / \underline{\xi} = X'(\underline{a}) \; ; \underline{a} \in E^*\}$$

Dans le langage du calcul des probabilités, H est la représentation dans F de l'ensemble des fonctions mesurables sur $(I, \mathcal{P}(I))$.

Remarque : Codage centré : On peut se placer dans F orthogonalement à la "droite des constantes" Δ_j, c'est-à-dire, ne considérant que les codages centrés, caractériser la variable x, non pas par le sous-espace H et le projecteur A, mais par le sous-espace $\overset{\circ}{H}$ et le projecteur associé $\mathring{A}$.



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

On supposera dans la suite de ce paragraphe 3.3 que les poids p_ω associés aux individus sont tous égaux à $\frac{1}{n}$.

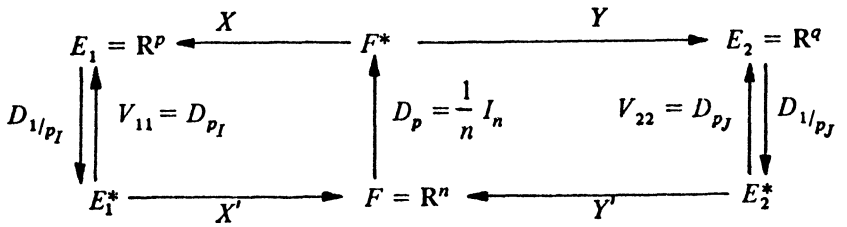
3.3.2 Stratégie Π – *Analyse factorielle des correspondances sur tableau de contingence.*

Le tableau de contingence N, dont les éléments sont notés n_{ij}, décrit l'imbrication des partitions associées aux deux caractères qualitatifs x et y, dont les ensembles de modalités sont notés respectivement I et J. On note :

$$p_{ij} = \frac{n_{ij}}{n} \qquad p_{i.} = \sum_j p_{ij} \qquad p_{.j} = \sum_i p_{ij}$$

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

Considérons le double schéma de dualité :



Analyser les liaisons entre les caractères x et y revient à préciser les positions relatives des sous-espaces $H_1 = X'(E_1^*)$ et $H_2 = Y'(E_2^*)$ représentatifs de ces caractères dans F ; l'"analyse factorielle des correspondances" est la technique consistant à analyser ces positions à l'aide de la stratégie Π : effectuer une analyse factorielle des correspondances revient donc à effectuer l'analyse canonique des indicatrices $\underline{x}^i$ et $\underline{y}^j$. Les facteurs canoniques (Cf. 3.2.2) par exemple sont donnés par les équations [11] :

$$B_1 B_2 \underline{a}_j = \lambda_j \underline{a}_j$$

$$\underline{b}_j = \frac{1}{\sqrt{\lambda_j}} B_2 \underline{a}_j$$

$$\|\underline{a}_j\|_{D_{p_I}} = \|\underline{b}_j\|_{D_{p_I}} = 1$$

avec ici :

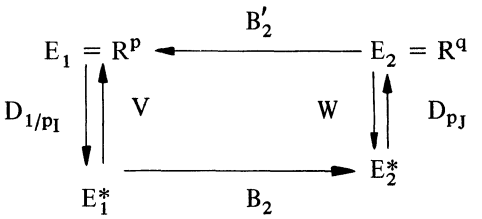
$$B_1 = V_{11}^{-1} V_{12} = D_{1/p_I} P \; ; \; B_2 = V_{22}^{-1} V_{21} = D_{1/p_J} P'$$

On notera la forme particulière des matrices B_1' et B_2' dont les colonnes sont les lois conditionnelles "sachant i" et "sachant j", respectivement définies par les probabilités $p_j^i = \frac{p_{ij}}{p_{i.}}$ et $p_i^j = \frac{p_{ij}}{p_{.j}}$. Ces lois conditionnelles sont représentées par les vecteurs $\underline{p}_J^i$ et $\underline{p}_I^j$ dans les espaces E_2 et E_1.

Ayant éliminé le couple de facteurs trivial $(\underline{a}_0, \underline{b}_0)$ correspondant au vecteur $\underline{j} = \Sigma \underline{x}^i = \Sigma \underline{y}^j$, dont toutes les coordonnées sont égales à 1, situé dans l'intersection des sous-espaces H_1 et H_2, le couple de facteurs canoniques $(\underline{a}_j, \underline{b}_j)$ fournit le $j^{ième}$ codage simultané des modalités de x et y rendant maximum la corrélation entre les caractères $\underline{\xi} = X'(\underline{a})$ et $\underline{\eta} = Y'(\underline{b})$ (j varie de 1 à p-1 si $p < q$).

Effectuer l'analyse canonique précédente revient à effectuer l'analyse en composantes principales du nuage des lois conditionnelles $\underline{p}_I^j$ "sachant j", ces lois étant munies des poids $p_{.j}$, et le simplexe des lois de probabilités définies sur I étant muni de la métrique du chi-deux de centre p_I (cf. 3.3.1) : c'est-à-dire à effectuer l'analyse en composantes principales du triplet $(B_2', D_{1/p_I}, D_{p_j})$.

Le schéma de dualité suivant est alors considéré :



La coordonnée de la loi $\underline{p}_I^j$ par rapport au $\ell^{ième}$ axe principal n'est autre que $\sum_i p_i^j a_\ell^i = \sqrt{\lambda_\ell}\, b_\ell^j$. Donc, au facteur $\sqrt{\lambda_\ell}$ près, b_ℓ^j (coordonnée de la modalité j de y) est au barycentre des a_ℓ^i (coordonnées des modalités i de x) affectés des poids p_i^j. En général, les valeurs propres λ_ℓ étant petites (et pour des raisons de symétrie), on préfère représenter simultanément les a_ℓ^i et b_ℓ^j. On dit que la représentation est non barycentrique.

Remarque :

$$T_r(A_1 A_2) = T_r(B_1 B_2) = \sum_{j=1}^{p-1} \lambda_j + 1 = \phi^2 + 1.$$

Le produit scalaire (cf. 3.1.4) entre les projecteurs associés aux sous-espaces H_1 et H_2 est ici égal à 1 près au phi-deux [11] associé au tableau des probabilités $P\left(\phi^2 = \frac{\chi^2}{n}\right)$.

Si on se place dans F orthogonalement à la droite des constantes, c'est-à-dire si on se restreint aux codages centrés (cf. remarque du paragraphe 3.3.1), on trouve :

$$\mathrm{Tr}\,(\mathring{A}_1 \mathring{A}_2) = \sum_{j=1}^{p-1} \lambda_j = \phi^2 = \chi^2/n$$

Alors le cosinus de l'angle entre les opérateurs $\mathring{A}_1$ et $\mathring{A}_2$ est le T^2 introduit par TSCHUPROV :

$$\cos(\mathring{A}_1, \mathring{A}_2) = \frac{\mathrm{Tr}\,(\mathring{A}_1 \mathring{A}_2)}{\sqrt{\mathrm{Tr}(\mathring{A}_1^2)\,\mathrm{Tr}(\mathring{A}_2^2)}} = \frac{\phi^2}{\sqrt{(p-1)(q-1)}} = T^2$$

3.3.3 *Stratégie Σ : vers l'analyse factorielle des correspondances sur tableau disjonctif complet*

Analyser les positions relatives des sous-espaces H_1 et H_2 à l'aide de la stratégie Σ conduit à considérer le tableau disjonctif "complet" $Z = \begin{vmatrix} X \\ Y \end{vmatrix}$ et le schéma de dualité (cf. 3.2.3) :

$$\begin{array}{ccc} & Z & \\ E = R^p & \longleftarrow & F^* \\ M \downarrow \uparrow V & & W \downarrow \uparrow D_p \\ E^* & \longrightarrow & F = R^n \\ & Z' & \end{array}$$

Ici :

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

$$V = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix} = \begin{bmatrix} D_{p_I} & P \\ P' & Dp_J \end{bmatrix}$$

et :

$$W D_p = A_1 + A_2.$$

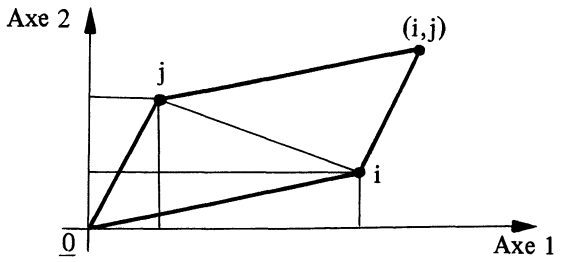
Les vecteurs $\underline{\beta}^j$, vecteurs propres normés de $A_1 + A_2$, associés aux valeurs propres μ_j sont alors homothétiques aux composantes principales $\underline{c}^j$ du triplet (Z, M, D_p) :

$$\underline{\beta}^j = X' \underline{a}_j + Y' \underline{b}_j$$

Si μ_j est différent de 1, $\underline{\beta}^j$ est équidistant des sous-espaces H_1 et H_2 : $||\underline{a}_j||_{D_{p_I}} = ||\underline{b}_j||_{D_{p_J}} = \frac{1}{\sqrt{2}}$.

On ne retient ici, comme précédemment, que les vecteurs $\underline{\beta}^j$ centrés ; ils sont orthogonaux au vecteur $\underline{j}$, vecteur propre trivial de $A_1 + A_2$ de valeur propre 2. L'analyse en composantes principales que l'on effectue ici sur des tableaux X et Y *non centrés* est donc équivalente à celle que l'on pourrait effectuer sur les tableaux *centrés.*

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



Au facteur $\frac{1}{\sqrt{2}}$ près, si les deux plus grandes valeurs propres μ_j sont supérieures à 1, on retrouve pour les modalités i et j la représentation simultanée non barycentrique décrite en 3.3.2.

Cette analyse en composantes principales effectuée sur le tableau disjonctif complet Z revient à l'analyse que proposait C. BURT [10] en 1950.

3.3.4 L'universalité

3.3.4.1 Deux variables qualitatives

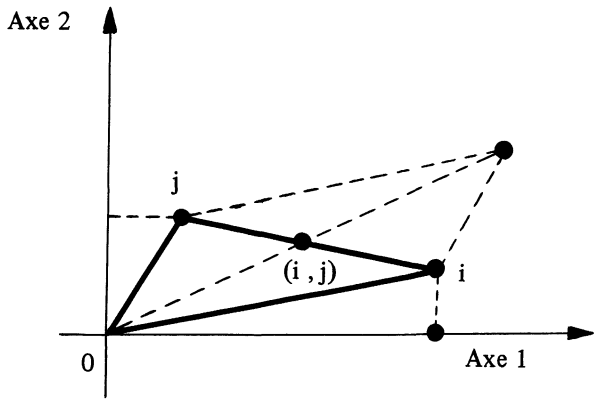
Analyser les positions relatives de H_1 et H_2 à l'aide de la stratégie Σ revient, et c'est ici qu'apparaît l'universalité de l'analyse factorielle des correspondances, *à effectuer l'analyse factorielle des correspondances sur le tableau disjonctif complet Z comme s'il était un tableau de contingence* (stratégie Π).

Posons en effet $B_2' = \frac{1}{2} Z$ (B_2 est le tableau des lois conditionnelles associées aux individus) et considérons le schéma de dualité :

$$\begin{array}{ccc} & \xleftarrow{\frac{1}{2} Z} & \\ E = R^{p+q} & & F^* \\ 2M \uparrow \downarrow \frac{1}{4} V & & \frac{1}{2} W \uparrow \downarrow D_p = \frac{1}{n} I_n \\ E^* & \xrightarrow[\frac{1}{2} Z']{} & F = R^n \end{array}$$

où V, M et W désignent les mêmes matrices que dans le schéma de dualité précédent. Les composantes principales sont ici au facteur $\frac{1}{\sqrt{2}}$ près identiques à celles obtenues à l'aide de la stratégie Σ. Les facteurs principaux $\underline{a}$ et $\underline{b}$ sont de même identiques aux précédents au facteur $\sqrt{2}$ près : compte tenu que des résultats identiques à une homothétie près fournissent des représentations graphiques équivalentes, on peut donc considérer que la représentation des deux caractères obtenue ici est identique à celle obtenue dans la Σ- analyse du tableau Z.

L'individu (i,j) est représenté dans cette analyse au milieu du segment reliant les modalités i et j c'est-à-dire par un vecteur identique à celui de la Σ-analyse, au facteur $\frac{1}{2}$ près.



Remarque : On montre sans difficulté [14], si on effectue l'analyse factorielle des correspondances sur le tableau :

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

considéré comme s'il était un tableau de contingence, que l'on retrouve encore, à peu de choses près (facteurs homothétiques, valeurs propres égales aux μ_j^2 au facteur $\frac{1}{4}$ près), les résultats fournis par la Σ – analyse du tableau Z.

3.3.4.2 k variables qualitatives

La pratique usuelle consistant à décrire simultanément l'ensemble des n individus et les ensembles $I_1, I_2 \ldots, I_k$ des modalités de k variables qualitatives en effectuant l'analyse factorielle des correspondances du tableau disjonctif complet Z obtenu en "empilant" toutes les variables indicatrices s'explique compte tenu des résultats de 3.3.4.1. Procéder ainsi revient, à des différences mineures près qui n'influent pas sur l'interprétation des résultats, à utiliser la stratégie Σ pour caractériser les positions relatives des sous-espaces $H_1, H_2, \ldots H_k$ (analyse en composantes principales réduite généralisée).

Pour éviter d'introduire le tableau disjonctif complet Z qui peut être très grand, on peut s'appuyer sur la remarque du paragraphe 3.3.4.1 et substituer à Z la généralisation à k variables qualitatives du tableau V précédent, couramment appelé "tableau de BURT".

3.3.4.3 Autres utilisations de l'analyse factorielle des correspondances : combinaison des stratégies Π *et* Σ

Quand on cherche à analyser la liaison entre une variable qualitative "à expliquer" x (sous-espace H et projecteur A) et p variables qualitatives "explicatives" $y^1, \ldots y^p$ (sous-espaces H_j et projecteurs B_j), différentes stratégies peuvent être utilisées [14] :

• l'une, classique en analyse de la variance, consiste à analyser à l'aide de la stratégie Π les positions relatives des sous-espaces H et ΣH_j (ΣH_j est engendré par l'ensemble de toutes les variables indicatrices associées aux y^j) ; pour tenir compte des interactions on "croise" entre elles les variables qualitatives intervenant dans ces interactions.

• Une autre consiste à extraire les vecteurs et valeurs propres du produit $A \,.\, \Sigma\, B_j$; cette stratégie revient à la précédente quand les variables y^j considérées sont indépendantes.

On montre sans difficulté qu'appliquer cette deuxième stratégie revient à analyser, par l'analyse factorielle des correspondances, le tableau obtenu en empilant les tableaux de contingence croisant la variable x avec les différentes variables y^j, ce tableau étant considéré comme s'il était un tableau de contingence.

Cette pratique se généralise immédiatement au cas où on a à analyser la liaison entre deux paquets de variables qualitatives (projecteurs A_i et projecteurs B_j) ; la pratique consistant à diagonaliser le produit $\Sigma\, A_i .\, \Sigma\, B_j$ revient à analyser, à l'aide de l'analyse factorielle des correspondances, le tableau obtenu en empilant et en juxtaposant les différents tableaux de contingence croisant les variables x^i avec les variables y^j, ce tableau étant considéré comme s'il était un tableau de contingence. L'analyse factorielle des correspondances sur tableau de BURT (cf. 3.3.4.2) apparaît alors comme un cas particulier de cette pratique.

3.3.5 Conclusion sur l'analyse factorielle des correspondances

Le succès que connaît à l'heure actuelle en France l'analyse factorielle des correspondances résulte en partie des propriétés remarquables rappelées dans les paragraphes précédents : cette technique permet de faire face à des situations très différentes (objectifs et tableaux de données différents). Des programmes performants ont été mis au point, certains étant particulièrement adaptés à l'analyse des tableaux disjonctifs complets [32].

BIBLIOGRAPHIE

[1] ANDERSON T.W. (1956). – *Statistical inference in factor analysis.* Proceedings of the third Berkeley symposium in mathematical statistics and probability.

[2] ANDERSON T.W. (1958). – *Introduction to multivariate analysis.* Wiley, New-York.

[3] ANSCOMBE F.J. (1963). – Tests of goodness of fit. *J.R. S.S.*, série B, vol. 25, part 1, p. 81-94.

[4] ANSCOMBE F.J. (1967). – Topics in the investigation of linear relations fitted by the method of least squares (avec discussion). *J.R. S.S.*, série B, vol. 29, part 1, p. 1-52.

[5] BARGMANN R. (1957). – A study of independance and dependance in multivariate normal analysis. University of North California, *Mimeograph series*, n° 186.

[6] BENZECRI J.P. (1973). – *L'analyse des données.* Dunod, Paris, (2 tomes).

[7] BENZECRI J.P. (1976). – Histoire et préhistoire de l'analyse des données. *Les Cahiers de l'analyse des données*, n° 1, 2, 3, 4 ..., Dunod.

[8] BERTIER P., BOUROCHE J.M. (1975). – *Analyse des données multidimensionnelles.* Presses Universitaires de France.

[9] G.E.P. Box (1976). – Science and statistics. *J.A.S.A.*, vol. 71, p. 791-799.

[10] BURT C. (1950). – The factorial analysis of qualitative data. *British Journal of Psychology*, Stat. sec. III.

[11] CAILLIEZ F., PAGES J.P. (1976). – *Introduction à l'analyse des données.* S.M.A.S.H., Paris (9, rue Duban, 75016 Paris).

[12] CARROLL J.D. (1968). – A generalization of canonical correlation analysis to three or more sets of variables. *Proceedings of the 76th annual convention of the american psychological association*, p. 227-228.

[13] CARROLL J.D., CHANG J.J. (1970). – Analysis of individual differences in multidimensional scaling via an n way generalization of "Eckart-Young" decomposition. *Psychometrika*, vol. 35, n° 3, p. 283-319.

[14] CAZES P., BAUMERDER A. et coll. (1977). – Codage et analyse des tableaux logiques. Introduction à la pratique des variables qualitatives. *Cahiers du Bureau Universitaire de recherche opérationnelle*, n° 27, Université Paris VI.

[15] DRAPER N.R., SMITH H. (1966). – *Applied regression analysis.* Wiley, New-York.

[16] ESCOUFIER Y. (1970). – *Echantillonnage dans une population de variables aléatoires réelles.* Publications de l'Institut de Statistique des Universités de Paris, Université Paris VI.

[17] ESCOUFIER Y., CAILLIEZ F., PAGES J.P. (1978). – Géométrie et techniques particulières en analyse factorielle. Communication présentée à : European meeting of psychometrics and mathematical psychology, University of Uppsala (Suède), 15 au 18 juin.

[18] FISHER R.A. (1936). – The use of multiple measurements in taxonomic problems. *Annals of eugenics*, vol. 7.

[19] HARMAN H.H. (1960). – *Modern factor analysis.* Chicago University Press.

[20] HECK D.L. (1960). – Charts of some upper percentage points of the distribution of the largest characteristic root. *Annals of mathematical statistics*, vol. 31.

[21] HILL M.O. (1974). – Correspondence analysis : a neglected multivariate method. *Applied statistics*, vol. 23, p. 340-354.

[22] HOTELLING H. (1933). – Analysis of a complex of statistical variables into principal components. *The Journal of educational psychology*, vol. 24, p. 417-441 et 498-520.

[23] HOTELLING H. (1936). – Relations between two sets of variables. *Biometrika*, vol. 28, p. 321-377.

[24] JARDINE N., SIBSON R. (1971). – Mathematical taxonomy. Wiley, Londres.

[25] JORESKOG K.G. (1963). – Statistical estimation in factor analysis : a new technique and its foundations. Almqvist-Wuksell, Stockholm.

[26] KETTENRING J.R. (1971). – Canonical analysis of several sets of variables. *Biometrika*, vol. 58, n° 3, p. 433-451.

[27] KRUSKAL J.B. (1964). – Multidimensional scaling by optimizing goodness of fit to a non metric hypothesis – Non metric multidimensional scaling : a numerical method. *Psychometrika*, vol. 29, n° 1 et 2.

[28] LAWLEY D.N., MAXWELL A.E. (1963). – *Factor analysis as a statistical method.* Butterworths, Londres.

[29] LAZARSFELD P. (1968). – *Latent structure analysis.* Houghton Mifflin Company.

[30] LEBART L., FENELON J.P. (1971). – *Statistique et informatique appliquées.* Dunod, Paris.

[31] LEBART L. (1977). – La validité des résultats en analyse des données. *Consommation*, n° 1.

[32] LEBART L., MORINEAU A., TABARD N. (1977). – *Techniques de la description statistique : méthodes et logiciels pour l'analyse des grands tableaux.* Dunod, Paris.

[33] PAGES J.P., ESCOUFIER Y., CAZES P. (1976). – Opérateurs et analyse des tableaux à plus de deux dimensions. *Cahiers du Bureau Universitaire de Recherche Opérationnelle,* n° 25, Université Paris VI.

[34] PEARSON K. (1901). – On lines and planes of closest fit to systems of points in space. *Phil. Mag,* n° 2 (6ème série), p. 559-572.

[35] PILLAI K.C.S. (1956). – On the distribution of the largest or the smallest root of a matrix in multivariate analysis. *Biometrika,* vol. 43, p. 122-127.

[36] ROMEDER J.M. (1973). – *Méthodes et programmes d'analyse discriminante.* Dunod, Paris.

[37] SEBESTYEN G.S. (1962). – *Decision making processes in pattern recognition.* The Mac Millan Company.

[38] SHEPARD R.N. (1962). – The analysis of proximities : multidimensional scaling with an unknown distance function. *Psychometrika,* vol. 27, n° 2 et 3, p. 125-140 et 219-246.

[39] SHEPARD R.N. (1974). – Representation of structure in similarity data : problems and prospects. *Psychometrika,* vol. 39, n° 4.

[40] SPEARMAN C. (1904). – General intelligence objectively determined and measured. *American Journal of Psychology,* vol. 15, p. 201-292.

[41] THURSTONE L.L. (1947). – *Multiple factor analysis.* Chicago University Press.

[42] TORGERSON W.S. (1958). – *Theory and methods of scaling.* Wiley, New-York.

[43] TUKEY J.W. (1962). – The future of data analysis. *Annals of mathematical statistics,* vol. 33, p. 1-67.

[44] TUKEY J.W. (1977). – *Exploratory data analysis.* Addison Wesley Publishing Company.

[45] WHITTLE P. (1948). – On principal components and least squares methods of factor analysis. *Skandinavisk aktuarietids-krift,* vol. 19.

[46] YOUNG G., HOUSEHOLDER A.S. (1938). – Discussion of a set of points in terms of their mutual distances. *Psychometrika,* vol. 3.