MATHÉMATIQUES ET SCIENCES HUMAINES

I. C. LERMAN

Analyse du phénomène de la « sériation » à partir d'un tableau d'incidence

Mathématiques et sciences humaines, tome 38 (1972), p. 39-57

<http://www.numdam.org/item?id=MSH_1972__38__39_0>

Math. Sci. hum. (11e année, n° 38, 1972, p. 39-57)

ANALYSE DU PHÉNOMÈNE DE LA «SÉRIATION» A PARTIR D'UN TABLEAU D'INCIDENCE

par

I. C. LERMAN [1]

I. INTRODUCTION

Le problème que pose l'archéologue pour la recherche des « sériations chronologiques» est en fait un problème fréquent dans les Sciences de la Nature et de l'Homme. Un des buts du spécialiste est en effet de cerner à partir d'un tableau de données un phénomène qui évolue et la variable qui conditionne cette évolution.

Commençons par formuler de façon intuitive le problème, en nous appuyant sur un exemple tiré de l'Archéologie. On suppose établi pour la description d'un ensemble E de p tombes, un ensemble fini A de n types d'objets. La description est matérialisée par un tableau d'incidence (ε_{lj}); $l = 1, 2, \ldots, n$ et $j = 1, 2, \ldots, p$, dont l'ensemble des lignes est indexé par A et celui des colonnes par E. $\varepsilon_{lj} = 1$ si le type d'objet l est présent dans la tombe j et 0 sinon.

Au cours du temps des divers types d'objets se succèdent; une tombe correspondra à un âge d'autant plus reculé qu'elle contiendra les types d'objets les plus vieux. Il s'agit de découvrir sur A l'ordre chronologique. L'hypothèse H du spécialiste qui rend possible la solution du problème est la suivante:

« Un type d'objet donné existe pendant une période *continue* de temps; de plus, si a_{t_1}, a_{t_2} et a_{t_3} sont trois types apparus aux dates t_1, t_2 et t_3 avec $t_1 < t_2 < t_3$, le type a_{t_2} est plus 'proche' de a_{t_1} que ne l'est a_{t_3}, respectivement, le type a_{t_2} est plus 'proche' de a_{t_3} que ne l'est a_{t_1}.» La notion de proximité entre deux types sera établie à partir du nombre de tombes possédant simultanément les deux types par rapport à ceux qui possèdent soit l'un soit l'autre.

L'hypothèse H n'est, de l'avis même de l'archéologue qu'une approximation de la réalité. Il se peut en effet qu'au cours des âges, certains types réapparaissent. Il faut toutefois espérer que H est vraie à des fluctuations assez négligeables près pour qu'on puisse appréhender le problème par la statistique. En fait, il nous sera possible de juger de la validité d'une telle hypothèse au moyen d'un test qui mesurera le degré d'invraisemblance de l'hypothèse d'absence de structure au profit de celle définie par H; mais il nous faut une longue expérience pour définir le seuil à adopter pour ce test.

On ne restreint en rien la généralité du problème si on suppose la condition C_0 suivante.

C_0: il n'existe pas deux types d'objets distincts qui soient simultanément présents ou absents de chacune des tombes.

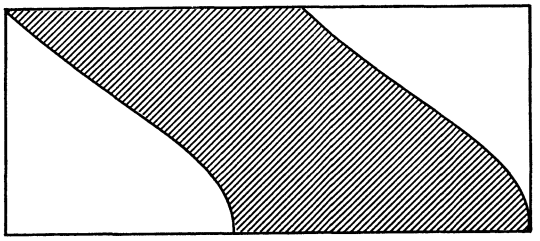
1. Centre de Mathématiques Appliquées et de Calcul de la Maison des Sciences de l'Homme.

Cette condition exprime qu'il n'existe pas deux lignes identiques de la matrice d'incidence. Si tel n'est pas le cas, on s'y ramène aisément en décrivant la matrice ligne par ligne et en supprimant tout vecteur ligne déjà rencontré. Il résultera en effet de l'analyse mathématique que la répétition d'une ligne ne fournit aucune information supplémentaire pour la solution du problème de la sériation.

La condition suivante C_1 peut paraître très restrictive; C_1: le nombre de tombes où un type d'objet donné apparaît est le même quel que soit le type. Cette condition signifie que le nombre de composantes égales à 1 dans un même vecteur ligne de la matrice d'incidence est constant.

Des considérations statistiques nous permettront de ramener le problème général à celui où cette condition C_1 se trouve satisfaite.

Si l'hypothèse H de l'Archéologue est exacte, il est aisé de voir qu'une permutation des lignes et des colonnes du tableau d'incidence pourra l'amener à la *forme* σ suivante qui tient compte des conditions C_0 et C_1 ci-dessus.



La partie hachurée est celle chargée de 1.

Si k désigne le nombre commun de composantes égales à 1 dans une même ligne du tableau l'expression mathématique de cette forme σ peut être une application qui à une ligne donnée de rang l, associe l'entier $c\,(l)$ tel que:

$$\varepsilon_{lj} = \begin{cases} 0 & \text{si} \quad j < c\,(l) \quad \text{ou} \quad j \geqslant c\,(l) + k \\ 1 & \text{si} \quad c\,(l) \leqslant j < c\,(l) + k. \end{cases}$$

La fonction $l \to c\,(l)$ étant strictement croissante, $c\,(1) = 1$ et $c\,(n) + k = p$.

Avant de nous engager plus avant, expliquons pourquoi ce travail; en d'autres termes, par quoi se distingue notre point de vue des autres méthodes d'approche du problème.

Signalons d'abord la technique de J. Bertin (*cf.* [3]). Cette dernière est essentiellement graphique; noircissant les cases du tableau d'incidence T où un 1 apparaît, l'expérimentateur opère, au moyen de réglettes, directement par permutation des lignes et des colonnes de façon à reconnaître visuellement, au mieux, la forme σ sous-jacente à l'hypothèse du spécialiste. Une telle procédure est très liée à l'intuition et à l'expérience de l'opérateur, elle n'est pas automatique et ne peut par conséquent traiter de très grands tableaux.

D'autre part, certaines méthodes connues attaquent le problème de la sériation à partir de techniques complexes élaborées pour d'autres problèmes et rendent compte des résultats obtenus pour des formes particulières du tableau d'incidence. Compte tenu de la complexité de ces techniques, il est difficile d'analyser pourquoi le phénomène de la sériation se manifeste d'une façon plutôt que d'une autre.

Nous venons en fait de présenter notre principale critique relative à la méthode de D. G. Kendall (*cf.* [5]), que nous allons rapidement esquisser.

Le point de départ de cette méthode est un algorithme de J. B. Kruskal (*cf.* [6]), MDSCAL, qui tente de réaliser une idée de R. N. Shepard (cf. [8]) : « Étant donné un ensemble de p points dans un espace de dimension n, déterminer une représentation euclidienne de cet ensemble dans un espace de faible dimension de façon à respecter au mieux le système des inégalités entre les distances des points. » Bien que consacré par l'expérience, l'algorithme de J. B. Kruskal ne semble pas avoir un fondement mathématique clair. L'ensemble des p points considéré par D. G. Kendall est celui des vecteurs colonnes de la matrice d'incidence; chaque vecteur colonne qui définit la description d'une tombe est un point dans un espace de dimension n. Les distances entre les points sont données par la métrique euclidienne où par conséquent le produit scalaire $\sum_l \varepsilon_{lj}\, \varepsilon_{lk}$ définit la proximité entre les deux points représentant les tombes j et k; c'est le nombre de types simultanément présents dans les deux tombes qui définit ainsi leur mesure de ressemblance. On applique MDSCAL pour obtenir une configuration des points dans un espace à 3 dimensions. D. G. Kendall établit la matrice d'inertie 3×3 du nuage des points obtenu qu'il projette sur le plan des deux premiers vecteurs propres. On constate alors, que pour une forme σ du tableau d'incidence, les points s'organisent sur ce plan en une courbe rappelant la forme du « fer à cheval »; l'ordre des points sur la courbe définit l'ordre des colonnes ou son inverse pour σ. Cet ordre permet de retrouver celui chronologique sur l'ensemble des tombes.

Nous envisageons quant à nous le problème plus direct de la sériation des types; dans ce cas d'ailleurs, la justification statistique de la mesure de proximité que nous adopterons apparaît plus clairement. Cette mesure de proximité permet de ramener le problème à celui où la condition C_1 ci-dessus est satisfaite. Nous étudierons dans ce cas la question de l'unicité de la solution (*cf.* § III). Pour une classe assez générale de tableaux d'incidence admettant la forme σ, il existe une représentation, des vecteurs lignes d'un tableau par des points d'un vecteur orienté de longueur 1, telle que les distances entre les points respectent exactement les « écarts » entre les vecteurs lignes du tableau, calculés conformément à la mesure de proximité établie (*cf.* § IV). Le problème se trouve ainsi réduit à la détermination d'une distribution d'un ensemble de point, dont on connaît le système des distances, sur un segment de droite orienté. Le lemme 1 du paragraphe IV permet de retrouver cette distribution à partir de l'un de ses deux points extrêmes. Pour une forme σ très générale qui peut même correspondre à plusieurs « dimensions » relativement « indépendantes » les unes des autres; c'est une analyse simultanée de la moyenne et de la variance des proximités qui permettra une représentation géométrique du tableau des données. Cette représentation, qui ne nécessite pas la diagonalisation d'une matrice, sera commentée par rapport à celle que fournit l'analyse factorielle présentée du point de vue de J. P. Benzécri (*cf.* [2], (b)).

II. MESURE DE PROXIMITÉ SUR L'ENSEMBLE A DES ÉLÉMENTS DESCRIPTIFS

Reprenons un instant le tableau d'incidence pour en donner diverses formulations. Ce tableau (ε_{lj}), $1 \leqslant l \leqslant n$ et $1 \leqslant j \leqslant p$, permet en général le croisement de deux ensembles finis, le premier A (**card** $A = n$), formé d'attributs descriptifs et le second E (**card** $E = p$), formé des objets ou sujets d'une population à laquelle le spécialiste s'intéresse. A chaque attribut a se trouve associé une ligne du tableau qui est un vecteur logique $\vec{a} = (a_1, a_2, \ldots, a_j, \ldots, a_p)$ où a_j est égal à 1 si l'attribut a est présent chez l'objet codé j et 0 sinon; $\vec{a}$ est un point du cube $\{ 0, 1 \}^p$. De façon équivalente, l'attribut a est représenté par le sous-ensemble E_a ($E_a \subset E$) des objets qui le possèdent. La représentation est ainsi définie au moyen d'une application A dans l'ensemble P (E) des parties de E. La matrice d'incidence nous transmet donc A comme un échantillon dans $\{ 0, 1 \}^p$ ou dans P (E).

A la représentation de A correspond celle duale de E; un vecteur colonne de la matrice d'incidence définit la description d'un élément de E. On peut regarder ce tableau comme définissant une distribution de masses égales à 1 ou 0 sur le rectangle :

$[1, 2, \ldots, n] \times [1, 2, \ldots, p]$ de N^2 où N est l'ensemble des entiers.

Par rapport à un même objet j, deux attributs a et b sont dits avoir une association positive (resp. négative) si a et b sont simultanément présents (resp. absents) chez j.

La notion de proximité entre attributs aura une importance cruciale; elle sera définie à partir de la représentation dans P (E) ou dans $\{0, 1\}^p$. On suppose pour cela que l'ensemble A des attributs de description est établi de telle sorte que seule une association positive entre deux attributs données contribue à leur similarité; cette circonstance correspond à la situation la plus fréquente.

Relativement à deux parties L et K de E représentant en l'occurrence deux éléments l et k de A, introduisons $\mu_l = \textbf{card}\ L/p$, $\mu_k = \textbf{card}\ K/p$ et $s = \textbf{card}\ L \cap K$ qui est la base de la mesure de proximité envisagée. Selon une suggestion de P. Achard, nous allons centrer et réduire la statistique s en nous référant à l'hypothèse N, que nous avions par ailleurs introduite (*cf.* [7], (b), ch. IV), où L (resp. K) serait tiré au niveau du simplexe P (E) défini par **card** L (resp. **card** K), ce dernier étant muni d'une mesure uniforme. Dans le cadre de l'hypothèse N on peut admettre que la statistique s suit une loi de Poisson de paramètre $p\ \mu_l\ \mu_k$ (*cf.* [4] ch. VI § 5). D'où la mesure de proximité:

$$S\ (l, k) = \frac{s - p\ \mu_l\ \mu_k}{\sqrt{p\ \mu_l\ \mu_k}} \tag{1}$$

qui suit dans l'hypothèse N une loi de probabilité très voisine de la loi normale centrée et réduite.

On peut se rendre compte que si on effectue, dans le tableau d'incidence des données, le changement de mesure:

$$\varepsilon_{lj} \to \frac{\varepsilon_{lj} - \mu_l}{\sqrt{\mu_l}\ \sqrt{p}} = \varepsilon'_{lj},$$

la statistique

$$s = \sum_{j=1}^{p} \varepsilon_{lj}\ \varepsilon_{kj}$$

devient:

$$S = \sum_{j=1}^{p} \varepsilon'_{lj}\ \varepsilon'_{kj}$$

en vertu de la relation:

$$\sum_{j=1}^{p} (\varepsilon_{lj} - \mu_l)\ (\varepsilon_{kj} - \mu_k) = s - p\ \mu_l\ \mu_k.$$

S correspond ainsi au produit scalaire euclidien sur le tableau transformé. Si μ_l est constant pour tout l dans A, s et S sont équivalents du point de vue de la préordonnance associée (*cf.* [7], (a), ch. 1).
Nous établirons les différents résultats dans ce dernier cas. Puisque notre analyse est basée sur l'étude des proximités et que S corrige s lorsque le nombre d'objets où un même attribut est présent n'est pas constant; nous étendrons au cas général l'algorithme obtenu.

Dans les paragraphes suivants III et IV on a $\mu_l = \mu$ pour toute ligne l du tableau d'incidence.

III. PROBLÈME DE L'UNICITÉ DE LA SOLUTION

Nous allons poser quelques définitions qui nous permettront de préciser notre vocabulaire.

Relativement à une forme σ (*cf.* § I), posons pour tout i, $c_i = c\ (i)/p$, la fonction $i \to c\ (i)$ a été

définie au paragraphe précédent. La forme σ sera dite réduite à la *forme parallélogrammique* π si pour tout couple de lignes d'indices l et k ($l < k$) on a $\frac{c_k - c_l}{(k - l)} =$ constante. Le tableau sera dit *horizontalement enchaîné* si pour tout couple de lignes (l, k) est *strictement positif* l'entier

$$s(l, k) = \sum_{j=1}^{p} \varepsilon_{lj}\, \varepsilon_{lk}.$$

Compte tenu de la signification des lignes, la condition exprime que pour tout couple d'attributs, il existe au moins un objet qui les possède simultanément.

Le tableau sera dit faiblement *horizontalement enchaîné* si pour tout couple de lignes (l, k), il existe une suite d'indices $(i_1, i_2, \ldots, i_r)$ telle que

$$\min \{ s(l, i_1), \quad s(i_1, i_2), \quad \ldots, \quad s(i_r, k) \} > 0.$$

Un tableau enchaîné l'est faiblement; si un tableau n'est pas faiblement enchaîné nous dirons qu'il est *disconnexe.*

Dans ce cas une *composante connexe* est définie comme la restriction du tableau à un ensemble maximal de lignes tel que le tableau restreint soit faiblement horizontalement enchaîné.

1. Proposition

La condition nécessaire et suffisante pour qu'un tableau d'incidence de zéros et de uns puisse être ramené à la forme σ est qu'il existe une permutation des colonnes telle que la partie chargée de toute ligne définisse un intervalle de l'ensemble des colonnes.

La condition est évidemment nécessaire; elle est aussi suffisante, l'ordre des lignes étant déterminé par la suite croissante des valeurs de $c(i)$; $i = 1, 2, \ldots, n$.

A toute solution σ définie pour une permutation donnée des colonnes, il correspond bijectivement une solution σ' définie pour la permutation inverse qui détermine sur les lignes l'ordre inverse de celui correspondant à σ. Dans la pratique, une information extérieure permettra au spécialiste de choisir entre deux solutions σ et σ'.

Relativement à une permutation des colonnes qui définit une forme σ du tableau, l'ordre des lignes est défini de façon unique. Chaque ligne définit sur l'ensemble des colonnes un préordre total à trois classes dont la classe médiane correspond à l'intervalle de la ligne chargé de uns. *L'intersection des différents préordres définit sur l'ensemble des colonnes un préordre total π à n classes* (n étant le nombre de lignes). En disant qu'une colonne est présente dans une ligne si à leur intersection se trouve un 1; une même classe du préordre total π est formé de colonnes simultanément présentes ou absentes de toute ligne du tableau.

Une permutation P des colonnes définit un ordre total O sur l'ensemble des colonnes; P sera dite *compatible* avec le préordre total π s'il en est ainsi de O; c'est-à-dire ($x < y$ pour l'ordre quotient défini par π) $\Rightarrow$ ($x < y$ pour O). Désignons toujours par π un préordre total sur l'ensemble des colonnes associé à une forme σ du tableau et soit π' le préordre total inverse où l'ordre quotient sur les classes définies par π est inversé. Il est évident que toute permutation des colonnes compatibles avec π ou π' donne au tableau une forme σ; de plus deux permutations compatibles avec le même préordre total π se déduisent l'une de l'autre par un produit de permutations dont chacune opère sur une même classe de π.

2. Proposition

Si le tableau est faiblement horizontalement enchaîné, les seules permutations des colonnes qui laissent invariante une forme σ sont celles qui sont compatibles avec π ou π'.

Nous allons montrer que pour une forme σ, les classes de π ainsi que l'ordre quotient, à son inverse près, sont déterminés de façon unique. En effet une classe extrême du préordre est définie par un ensemble maximal de colonnes présentes dans exactement une ligne. Cette classe extrême va définir la première ligne du tableau à partir de laquelle seront déterminées en même temps que l'ordre total sur les lignes, les autres classes du préordre selon π ou π'. La $k$ième ligne est celle, l, parmi les lignes non encore rangées, pour laquelle est maximum le nombre de colonnes simultanément présentes dans les lignes $(k-1)$ et l. L'ordre total sur les lignes permet de définir le préordre total π ou π' selon que la classe extrême initialement considérée est la première ou la dernière de π.

En conclusion, si une forme σ du tableau existe, l'ordre total sur les lignes permettant de l'obtenir est déterminé à l'ordre inverse près; d'autre part, la proposition précédente nous permet, à partir d'une forme σ obtenue, d'énumérer toutes les permutations de colonnes définissant une forme σ.

IV. Représentation sur un segment de droite orienté

On ne restreint en rien la généralité de ce qui va suivre en supposant le segment de droite de longueur 1.

1. Lemme

Soit $\{ c_1, c_2, \ldots, c_n \}$ l'ensemble des abscisses de n points répartis de façon quelconque sur un segment de droite orienté $\overrightarrow{AB}$ de longueur 1; $c_1 < c_2 < \ldots < c_n$. On a (α): celui des n points dont la moyenne des distances à l'ensemble des points est la plus grande est nécessairement un point extrême. (β): celui des n points dont la variance des distances à l'ensemble des points est la plus grande est aussi nécessairement un point extrême.

(α) *Moyenne des distances d'un point*

Désignons chacun des n points par son rang en allant de A vers B, le point i étant ainsi d'abscisse c_i.

Nous allons comparer la moyenne des distances de 1 à celle de l pour $l \leqslant n/2$. La suite des distances de 1 est:

$$0, c_2 - c_1, c_3 - c_1, \ldots, c_l - c_1, c_{l+1} - c_1, \ldots, c_n - c_1,$$

d'où la moyenne des distances de 1:

$$M(1) = \frac{1}{n}\left(\sum_{i=1}^{n} c_i - nc_1\right).$$

La suite des distances de l est:

$$c_l - c_1, c_l - c_2, \ldots, c_l - c_{(l-1)}, 0, c_{(l+1)} - c_l, c_{l+2} - c_l, \ldots, c_n - c_l.$$

d'où la moyenne des distances de l:

$$M(l) = \frac{1}{n}\left((2l - n)\, c_l - \sum_{i=1}^{l} c_i + \sum_{i=l+1}^{l} c_i\right).$$

La différence:

$$M(1) - M(l) = \frac{1}{n}\left((n - 2l)\, c_l + 2 \sum_{i=1}^{l} c_i - nc_1\right)$$

$$= \frac{1}{n}\left((n - 2l)\,(c_l - c_1) + 2 \sum_{i=1}^{l} (c_i - c_1)\right)$$

or:

$$n - 2l \geqslant 0,\ c_l - c_1 > 0 \qquad \text{et} \qquad c_i - c_1 \geqslant 0 \qquad \text{pour} \qquad 1 \leqslant i \leqslant l,$$

donc:

$$M(1) - M(l) > 0 \qquad \text{pour tout } l \text{ tel que } 2l \leqslant n.$$

L'inégalité est encore vraie si n est impair et si $2l = n + 1$; en effet, on a alors:

$$M(1) - M(l) = \frac{1}{n}\left((c_l - c_1) + 2 \sum_{i=1}^{(l-1)} (c_i - c_1)\right)$$

finalement:

$$M(1) = \max_{1 \leqslant l \leqslant [(n+1)/2]} M(l)$$

où $[(n + 1)/2]$ désigne la partie entière de $(n + 1)/2$.

Symétriquement:

$$M(n) = \max_{l > [n/2]} M(l).$$

La partie (α) du lemme se trouve démontrée. Par conséquent, si la fonction distance est donnée, l'ordre des points, ou son inverse, sur le segment orienté, peut être déterminé à partir du point dont la moyenne des distances est la plus grande.

(β) *Variance des distances d'un point.*

Comparons la variance des distances de 1 à celle de l pour $l \leqslant n/2$. La variance des distances de 1 est:

$$V(1) = \frac{1}{n} \sum_{i=1}^{n} (c_i - c_1)^2 - \frac{1}{n^2}\left(\sum_{i=1}^{n} c_i - nc_1\right)^2.$$

Prenons le point 1 pour origine et posons $d_i = c_i - c_1$, on a:

$$V(1) = \frac{1}{n} \sum_{i=1}^{n} d^2 - \left(\frac{1}{n} \sum_{i=1}^{n} d_i\right)^2.$$

La variance des distances de l est:

$$V(l) = \frac{1}{n} \sum_{i=1}^{n} (c_l - c_i)^2 - \frac{1}{n^2}\left[\sum_{i=1}^{l} (c_l - c_i) + \sum_{i=l+1}^{n} (c_i - c_l)\right]^2$$

soit:

$$V(l) = \frac{1}{n} \sum_{i=1}^{n} (d_l - d_i)^2 - \frac{1}{n^2}\left[\sum_{i=1}^{l} (d_l - d_i) + \sum_{i=l+1}^{n} (d_i - d_l)\right]^2.$$

Calculons la différence V (1) — V (*l*),

$$V(1) - V(l) = \frac{d_l}{n} \sum_{i=1}^{n} (2d_i - d_l) - \frac{1}{n^2}\left[2 \sum_{i=1}^{l} d_i + (n - 2l)\, d_l\right] \times \left[2 \sum_{i=l+1}^{n} d_i - (n - 2l)\, d_l\right].$$

Un calcul, dont nous ne donnons pas le détail ici, montre que:

$$V(1) - V(l) = \frac{4}{n^2} d^2 \left(\sum_{l+1}^{n} e_j - l\right)\left((n - l) - \sum_{1}^{l} e_i\right)$$

où $e_j = d_j/d_l$.

Or $e_j = d_j/d_l > 1$ pour $j > l$, de sorte que

$$\sum_{l+1}^{n} e_j - l > n - 2l \geqslant 0.$$

D'autre part $e_i = d_i/d_l \leqslant 1$ pour $i \leqslant l$, de sorte que:

$$(n - l) - \sum_{1}^{l} e_i \geqslant n - 2l \geqslant 0.$$

Donc: V (1) $\geqslant$ V (*l*) pour $l \leqslant n/2$.

La partie (β) du lemme se trouve ainsi démontrée.

Par conséquent, le point pour lequel V est maximum nous permettra de retrouver l'ordre, ou son inverse, des points. En effet, à partir du point extrême on utilisera l'algorithme des « enchaînements successifs» où à chaque pas on détermine le plus voisin du dernier retenu.

2. Expression de la mesure de proximité pour une forme σ

Remettons-nous en mémoire le tableau d'incidence et reprenons le problème de la « sériation» où on cherchera à découvrir l'ordre total sur les lignes du tableau. Avec les notations définies précédemment (*cf.* § III), l'expression de la mesure de proximité entre deux lignes est:

$$S(l, k) = \sum_{j=1}^{p} \frac{\varepsilon_{lj} - \mu}{\sqrt{\mu\sqrt{p}}} \frac{\varepsilon_{kj} - \mu}{\sqrt{\mu\sqrt{p}}} = \frac{s - p\,\mu^2}{\mu\sqrt{p}}$$

puisqu'on suppose $\mu_l = \mu$ pour toute ligne l du tableau.

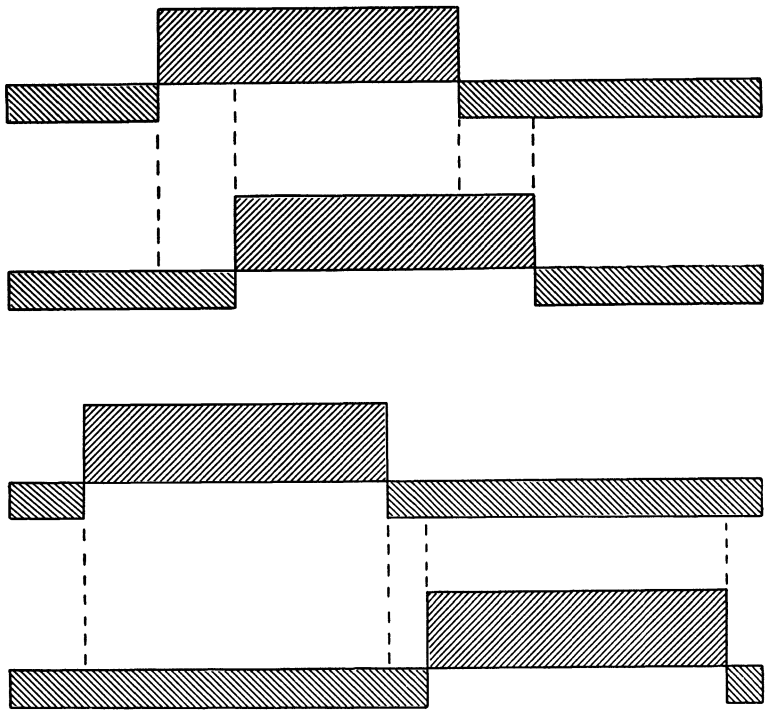
Rappelons le changement de mesure:

$$\varepsilon_{lj} \to \varepsilon'_{lj} = \frac{\varepsilon_{lj} - \mu}{\sqrt{\mu\sqrt{p}}}$$

pour lequel:

$$S(l, k) = \sum_{j} \varepsilon'_j\, \varepsilon'_{kj}$$

Supposons que le tableau d'incidence admette la forme σ et désignons par l et k ($l < k$), les rangs respectifs de deux lignes pour σ. Les deux lignes peuvent se présenter, après le changement de mesure $\varepsilon_{lj} \to \varepsilon'_{lj}$, sous l'une des deux formes suivantes, où les hachures ////// expriment une charge positive égale à $(1 - \mu)/\sqrt{\mu\sqrt{p}}$ et où celles \\\\\\, une charge négative, égale à $-\sqrt{\mu}/\sqrt{p}$. Le cas *a*) est caractérisé par $c_k \leqslant c_l + \mu$, le cas *b*) par $c_k > c_l + \mu$.



Les diverses valeurs de $\varepsilon'_j\ \varepsilon'_{kj}$ sont:

$$\begin{aligned}
\mu / \sqrt{p} \quad &\text{si} \quad \varepsilon_{lj} = \varepsilon_{kj} = 0 \\
-(1-\mu) / \sqrt{p} \quad &\text{si} \quad \varepsilon_{lj} = 0,\ \varepsilon_{kj} = 1 \\
-(1-\mu) / \sqrt{p} \quad &\text{si} \quad \varepsilon_{lj} = 1,\ \varepsilon_{kj} = 0 \\
(1-\mu)^2 / \mu \sqrt{p} \quad &\text{si} \quad \varepsilon_{lj} = \varepsilon_{kj} = 1.
\end{aligned}$$

De sorte que l'expression de la mesure de proximité $S\ (l,\ k)$ est:

a') *dans le cas* a) *où* $c_k \leqslant c_l + \mu$:

$$S\ (l,\ k) = \sum_{1 \leqslant j \leqslant pc_l} \mu / \sqrt{p} - \sum_{pc_l + 1 \leqslant j \leqslant pc_k} (1-\mu) / \sqrt{p} + \sum_{pc_k + 1 \leqslant j \leqslant p\,(c_l + \mu)} (1-\mu)^2 / \mu \sqrt{p}$$

$$- \sum_{p\,(c_l + \mu) + 1 \leqslant j \leqslant p\,(c_k + \mu)} (1-\mu) / \sqrt{p} + \sum_{p\,(c_k + \mu) + 1 \leqslant j \leqslant p} \mu / \sqrt{p}$$

$$= \frac{1}{\sqrt{p}} \Big\{ [pc_l + p\ (1 - c_k - \mu)]\ \mu + [p\ (c_l + \mu - c_k)]\ (1-\mu)^2 / \mu - 2p\ (c_k - c_l)\ (1-\mu) \Big\}$$

$$= \frac{\sqrt{p}}{\mu} \Big\{ [(c_l - c_k) + (1-\mu)]\ \mu^2 + [(c_l - c_k) + \mu]\ (1-\mu)^2 + 2p\ (c_l - c_k)\ (1 - \mu) \Big\}$$

d'où en regroupant on obtient:

$$S\ (l,\ k) = \frac{\sqrt{p}}{\mu}\ [\mu\ (1-\mu) - (c_k - c_l)]$$

La valeur de $S\ (l,\ k)$ pour $c_k = c_l + \mu$ est $-\sqrt{p}\ \mu$; d'autre part, $S\ (l,\ k) \geqslant -\sqrt{p}\ \mu$ pour $c_k \leqslant c_l + \mu$.

b') *dans le cas* b) *où* $c_k > c_l + \mu$

$$S(l, k) = \frac{1}{\sqrt{p}} \left\{ p c_l \mu - p \mu (1 - \mu) + p (c_k - c_l - \mu) \mu - p \mu (1 - \mu) + p (1 - c_k - \mu) \mu \right\}$$
$$= - \sqrt{p} \mu$$

finalement:

$$S(l, k) = \begin{cases} \frac{\sqrt{p}}{\mu} [\mu (1 - \mu) - (c_k - c_l)] & \text{si} \quad c_k \leqslant c_l + \mu \\ - \sqrt{p} \mu & \text{si} \quad c_k \geqslant c_l + \mu \end{cases}$$

Introduisons l'« écart» $D(l, k)$:

$$D(l, k) = - \frac{\mu}{\sqrt{p}} S(l, k) + \mu (1 - \mu);$$

on a:

$$D(l, k) = \begin{cases} c_k - c_l & \text{si} \quad c_k \leqslant c_l + \mu \\ \mu & \text{si} \quad c_k \geqslant c_l + \mu. \end{cases}$$

2.1 *Théorème*

Si un tableau d'incidence remplit les conditions

(*i*) le nombre de 1 dans toute ligne est le même (*i.e.*, $\mu_l = \mu$ pour tout l),
(*ii*) le tableau est horizontalement enchaîné,
(*iii*) le tableau admet la forme σ,

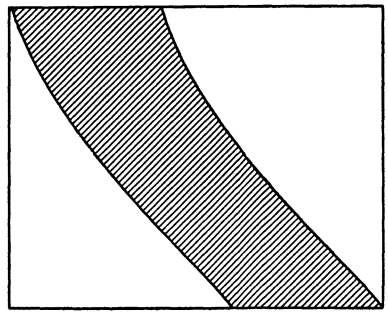
alors, il existe une représentation, sur un segment de droite orienté de longueur 1, des lignes du tableau par des points du segment dont l'ordre est celui relatif à σ et dont les distances sont les écarts $S(l, k)$.

En effet, la condition (*ii*) est équivalente à $c_n < c_1 + \mu$. Le lemme ci-dessus nous permet de conclure.

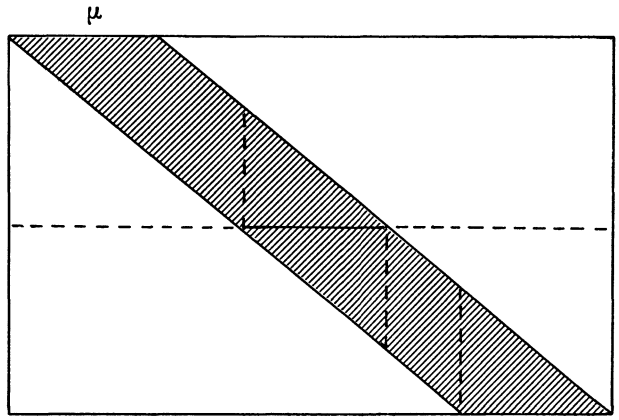
2.2. *Théorème*

Si un tableau d'incidence remplit les conditions (*i*), (*ii*) et (*iii*) du théorème précédent, on a (α) celle des n lignes dont la moyenne des proximités S à l'ensemble des lignes est la plus faible est nécessairement une ligne extrême pour la forme σ du tableau; de même (β) celle des n lignes dont la variance des proximités à l'ensemble des lignes est la plus grande est nécessairement une ligne extrême pour la forme σ.

Dans le cas d'un tableau faiblement enchaîné comme l'indique la figure suivante, où les hachures définissent la partie chargée du tableau; on peut montrer aisément que la ligne dont la moyenne des écarts D est la plus grande est nécessairement une ligne extrême.



Mais la propriété (β) du théorème précédent n'est plus vérifiée en général dans ce cas. Pour le montrer, nous allons examiner le cas d'un tableau faiblement enchaîné ayant une forme parallélogrammique π suffisamment inclinée pour que l'intervalle chargé de la ligne médiane ait avec chacun des intervalles chargés des deux lignes extrêmes une intersection vide.



Dans les calculs qui vont suivre, 1 désignera la première ligne du tableau et l, la ligne médiane. Nous allons comparer d'une part les moyennes; d'autre part, les variances des écarts de 1 et de l à l'ensemble des lignes du tableau.

Rappelons que pour le cas d'une forme π:

$$D(l, k) = \begin{cases} c_k - c_l = \alpha (k - l) & \text{si} \quad c_k \leqslant c_l + \mu \\ \mu & \text{si} \quad c_k \geqslant c_l + \mu \end{cases}$$

où $k \geqslant l$ et où α est une constante.

La suite des valeurs des écarts de 1 est donc:

$$\alpha, 2\alpha, \ldots, \left(\frac{\mu}{\alpha} - 1\right)\alpha, \mu, \mu, \ldots, \mu;$$

la suite se termine par $\left(n - \frac{\mu}{\alpha}\right)$ termes tous égaux à μ.

La suite des valeurs des écarts de l est:

$$\alpha, 2\alpha, \ldots, \left(\frac{\mu}{\alpha} - 1\right)\alpha, \alpha, 2\alpha, \ldots, \left(\frac{\mu}{\alpha} - 1\right)\alpha, \mu, \mu, \ldots, \mu;$$

la suite comprend deux fois la séquence $\alpha, 2\alpha, \ldots, \left(\frac{\mu}{\alpha} - 1\right)\alpha$ et se termine par $\left(n - 2\frac{\mu}{\alpha} + 1\right)$ termes tous égaux à μ.

Moyenne des écarts de 1, $\mathcal{M}$ (1)

Posons $q = \mu / \alpha$; on a:

$$\begin{aligned} \mathcal{M}(1) &= \frac{1}{n}\left[\alpha q (q - 1) / 2 + (n - q)\mu\right] \\ &= \frac{1}{n}\left[\mu\left(n - \frac{q + 1}{2n}\right)\right] \\ &= \mu\left(1 - \frac{q + 1}{2n}\right). \end{aligned}$$

Moyenne des écarts de l, $\mathscr{M}$ (l)

$$\mathscr{M}(l) = \frac{1}{n}[\alpha\, q\,(q-1) + (n-2q+1)\,\mu]$$

$$= \mu\left(1 - \frac{q}{n}\right)$$

D'où: $\mathscr{M}(1) > \mathscr{M}(l)$.

Variance des écarts de 1, $\mathscr{V}$ (1) *et de l,* $\mathscr{V}$ (l)

$$\mathscr{V}(1) = \frac{1}{n}\left[\sum_{1 \leqslant k \leqslant (q-1)} \alpha^2 k^2 + (n-q)\,\mu^2\right] - \mu^2\,[1 - (q+1)/2n]^2$$

$$\mathscr{V}(l) = \frac{1}{n}\left[2\sum_{1 \leqslant k \leqslant (q-1)} \alpha^2 k^2 + (n-2q+1)\,\mu^2\right] - \mu^2\left(1 - \frac{q}{n}\right)^2.$$

Étudions le signe de la différence $\mathscr{V}(l) - \mathscr{V}(1)$.

$$\mathscr{V}(l) - \mathscr{V}(1) = \frac{1}{n}[\alpha^2\,(q-1)\,q\,(2q-1)/6 - (q-1)\,\mu^2] - \mu^2\left\{\left((1-\frac{q}{n}\right)^2 - [1-(q+1)/2n]^2\right\}.$$

Un calcul élémentaire nous montre que:

$$\mathscr{V}(l) - \mathscr{V}(1) \simeq \frac{\mu^2 q}{3n} - \frac{\mu^2 q}{n} - \mu^2\left[\left(1 - \frac{q}{n}\right)^2 - \left(1 - \frac{q}{2n}\right)^2\right]$$

l'écart entre le second et le premier nombre étant de l'ordre de $\mu^2 / 2n$. Le membre de droite peut se mettre sous la forme:

$$\frac{\mu^2 q}{n}\left[-\frac{2}{3} + \left(1 - \frac{3}{4}\frac{q}{n}\right)\right]$$

$$= \frac{\mu^2 q}{n}\left(\frac{1}{3} - \frac{3}{4}\frac{q}{n}\right)$$

Donc: $\mathscr{V}(l) > \mathscr{V}(1)$ si $\frac{q}{n} > \frac{4}{9}$ c'est-à-dire $\frac{\mu}{\alpha} < \frac{4}{9}n$.

2.3. *Proposition*

Si un tableau d'incidence admet une forme π pour laquelle $\mu/\alpha < 4n/9$, pour cette forme, la variance des écarts de la ligne médiane est supérieure à celle d'une ligne extrême.

Pour un tableau faiblement horizontalement enchaîné et admettant une forme σ comme l'indique la figure 2.1 ci-dessus, l'ordre des lignes, ou son inverse, peut être déterminé en utilisant l'algorithme des «enchaînements successifs» à partir d'une ligne extrême l_1, définie comme étant celle dont la moyenne des écarts est la plus grande. Cet algorithme consiste à retenir au $k$ième pas la ligne l_k la plus proche de celle l_{k-1}. Le théorème de représentation 2.1 n'est plus vrai si un tableau remplit les conditions (*i*), (*iii*), mais n'est que faiblement horizontalement enchaîné.

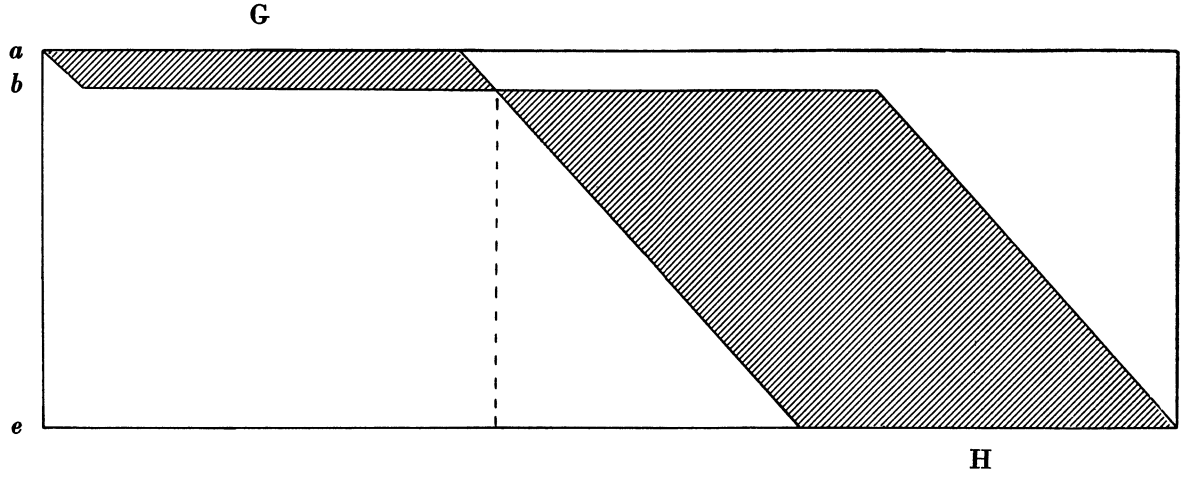
V. ANALYSE SIMULTANÉE DE LA MOYENNE ET DE LA VARIANCE DES PROXIMITÉS

Nous allons examiner quelques exemples où la forme σ fait apparaître des blocs parallélogrammiques; deux blocs successifs étant tels que la dernière ligne du premier et la première ligne du second aient en

commun pour tout au plus un petit intervalle de l'ensemble des colonnes, une charge simultanée égale à 1. Ces exemples, pour lesquels $\mu_l = \mu$ pour tout l, nous permettrons de préciser un algorithme de représentation géométrique des tableaux d'incidence des données qu'on appliquera dans le cas général.

1. Exemple 1

Soit le tableau d'incidence *ramené à la forme* σ comme il est indiqué dans la figure.



La partie hachurée est celle qui est chargée de uns; elle est constituée de deux blocs G et H ayant la forme de parallélogramme. m est l'indice de la ligne qui termine le bloc G; on a $m = n / 10.\ (m + 1)$ est l'indice de la ligne qui commence le bloc H. On supposera dans les calculs qui vont suivre n grand; en tout cas assez grand pour que, dans ces calculs on puisse confondre $(m - 1)$ ou $(m + 1)$ avec m, sans effet sensible sur le résultat.

Soit comme il est indiqué dans la figure, $\mu = 1 / 3$.

Lorsque les lignes l et k $(l < k)$, appartiennent à un même bloc, on a:

$$\frac{c_k - c_l}{(k - l)} = \text{constante } \alpha \text{ où } \boxed{\alpha = 1 / 3n}$$

a est le vecteur représenté par la première ligne du tableau; b, celui représenté par la $(m + 1)$ème ligne et e, celui représenté par la dernière ligne.

Un calcul analogue à celui qui a abouti à la proposition 2.3 du paragraphe précédent, permet d'établir le tableau des valeurs:

$$\mathcal{M}(a) = 0{,}302 \qquad \mathcal{M}(b) = 0{,}1684$$
$$\mathcal{V}(a) = 0{,}00899 \qquad \mathcal{V}(b) = 0{,}00989.$$

$\mathcal{M}(a)$ (resp. $\mathcal{M}(b)$) est la moyenne des écarts de a (resp. de b);
$\mathcal{V}(a)$ (resp. $\mathcal{V}(b)$) est la variance des écarts de a (resp. de b).

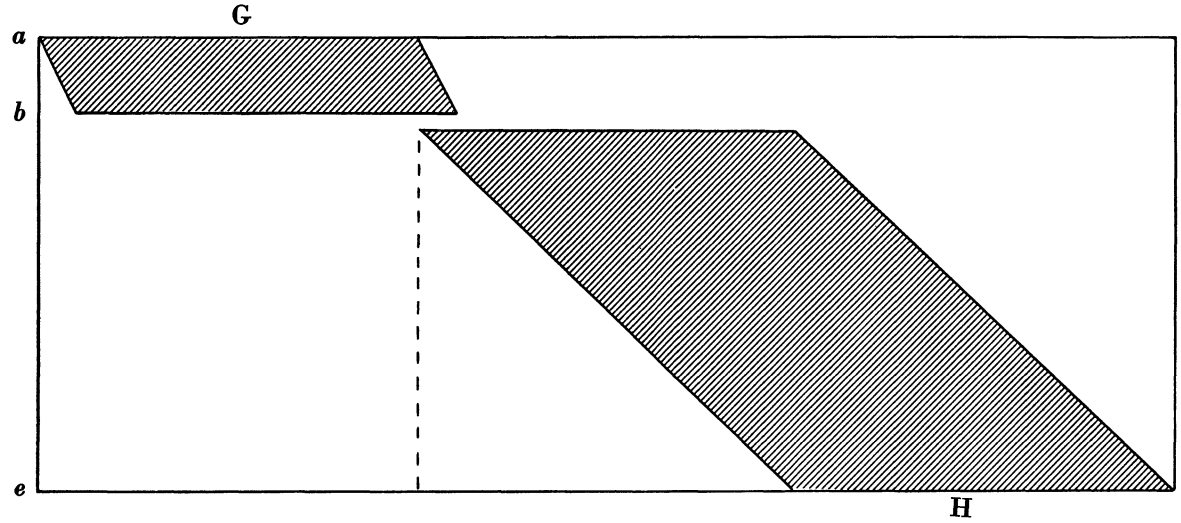
Ainsi $\mathcal{M}(a)$ est sensiblement plus grand que $\mathcal{M}(b)$ alors que $\mathcal{V}(b) > \mathcal{V}(a)$. Dans le cas de notre tableau, il est surtout intéressant de découvrir la présence des deux blocs, quitte par la suite à ranger les lignes de chacun d'entre eux. Certes, une classification automatique permet de détecter les deux blocs; mais ici, nous l'effectuerons plus simplement par une analyse simultanée de la variance et de la moyenne des proximités. En effet, à partir de b dont la variance des proximités est maximum, a peut être obtenu parmi les vecteurs lignes différents de b qui rendent maximum le rapport homogène:

$$\mathscr{V}(x) \,/\, (S(x, b))^2,$$

où $S(x, b)$ est la mesure de proximité entre a et b. Dans notre exemple, ce sont les deux lignes extrêmes du bloc G qui réalisent le maximum du rapport ci-dessus. En appelant, dans l'exemple considéré, I. (resp. J), l'ensemble totalement ordonné des colonnes associées à G (resp. H); on remarque que le tableau conserve la forme σ si on place J avant I (*cf.* § IV). Ce ne sera plus le cas pour le second exemple que nous allons envisager et qui nous permettra d'améliorer notre intuition du problème.

2. Exemple 2

Le tableau d'incidence se présente comme suit lorsqu'il est ramené à la forme σ.



$$(1 - 2\mu) = = \mu\ 1 \,/\, 3.$$

L'indice m de la ligne qui termine le bloc G est ici égal à $n \,/\, 6$ où n est le nombre total de lignes. Les lignes m et $(m + 1)$ ont en commun pour une tranche de l'ensemble des colonnes une charge simultanée égale à 1. n est supposé assez grand et on a $\mu = 1 \,/\, 3$.

Lorsque les lignes l et k appartiennent au même bloc G, on a:

$$\frac{c_k - c_l}{(k - l)} = \text{constante } \alpha, \text{ où } \alpha = 1 \,/\, 5n.$$

D'autre part, lorsque les lignes l et k appartiennent au même bloc H, on a:

$$\frac{c_k - c_l}{(k - l)} = \text{constante } \beta, \text{ où } \beta = 2 \,/\, 5n.$$

a est toujours le point représenté par la première ligne du tableau;

b celui représenté par la $(m + 1)$ème ligne. Les résultats du calcul, présentés avec les notations adoptées ci-dessus sont:

$$\mathscr{M}(a) = 0{,}280 \qquad \mathscr{M}(b) = 0{,}192$$
$$\mathscr{V}(a) = 0{,}014 \qquad \mathscr{V}(b) = 0{,}011.$$

On a ici $\mathscr{M}(a) > \mathscr{M}(b)$ et $\mathscr{V}(a) > \mathscr{V}(b)$; toutefois le rapport $\mathscr{M}(a) \,/\, \mathscr{M}(b)$ est sensiblement plus élevé que celui $\mathscr{V}(a) \,/\, \mathscr{V}(b)$.

L'algorithme des enchaînements successifs aurait permis ici de reconstituer l'ordre des lignes en partant de a dont la moyenne des écarts $\mathcal{M}$ est maximum. Mais cet algorithme ne met pas en évidence la présence des deux blocs G et H; il faut pour cela recourir à une analyse simultanée de la variance et de la moyenne des proximités comme il a été fait allusion dans l'exemple ci-dessus. En effet, pour le point b il se produit un saut positif brutal de:

$$\mathscr{V}(x) \;/\; (S(x, a))^2$$

lorsque x parcourt l'ensemble des vecteurs lignes du tableau différents de a et rangés par proximité décroissante à a. Le vecteur ligne pour lequel le rapport précédent est maximum est b ou e (il faut faire le calcul pour e); de toute façon l'un quelconque de ces deux points permet de définir la dimension sous-jacente à H; c'est-à-dire, l'ordre ou son inverse des attributs de description représentés par les lignes de H ramené à la forme σ.

3. Formule d'analyse de la variance des proximités

Considérons le tableau carré $n \times n$ des proximités S_{lk}:

$$S(l, k) = \frac{s - p\,\mu_l\,\mu_k}{\sqrt{p\,\mu_l\,\mu_k}}$$

Nous allons commencer par effectuer une analyse de la variance globale des proximités selon les lignes du tableau carré auquel on aura ôté le contenu de la diagonale (cette analyse est d'ailleurs identique à celle selon les colonnes du tableau carré).

Posons:

$$\overline{S}_l = \frac{1}{(n-1)} \sum_{\{k \,/\, k \neq l\}} S_{lk} = \text{ moyenne des proximités de } l.$$

$$\overline{S} = \frac{1}{n} \sum_{1 \leqslant l \leqslant n} \overline{S}_l = \text{ moyenne globale des proximités.}$$

Décomposons la différence $(S_{lk} - \overline{S})$ comme suit:

$$(S_{lk} - \overline{S}) = (S_{lk} - \overline{S}_l) + (\overline{S}_l - \overline{S})$$
$$(S_{lk} - \overline{S})^2 = (S_{lk} - \overline{S}_l)^2 + (\overline{S}_l - \overline{S}^2) + 2\,(S_{lk} - \overline{S}_l)\,(\overline{S}_l - \overline{S})$$

en sommant par rapport à k et pour $k \neq l$, on obtient:

$$\sum_{\{k \,/\, k \neq k\}} (S_{lk} - \overline{S})^2 = \sum_{\{k \,/\, k \neq l\}} (S_{lk} - \overline{S}_l)^2 + (n-1)\,(\overline{S}_l - \overline{S})^2 + 0.$$

en sommant par rapport à l et en divisant par $n\,(n-1)$ les deux membres on a :

$$\frac{1}{n\,(n-1)} \sum_{\{(l,k) \,/\, l \neq k\}} (S_{lk} - \overline{S})^2 = \frac{1}{n} \sum_{1 \leqslant l \leqslant n} \frac{1}{(n-1)} \sum_{\{k \,/\, k \neq l\}} (S_{lk} - \overline{S}_l)^2 + \frac{1}{n} \sum_{1 \leqslant l \leqslant n} (S_l - \overline{S})^2.$$

Dans cette formule le premier membre définit la dispersion du nuage des points représentant les attributs dans le simplexe P (E) (*cf.* § II). Cette dispersion se décompose d'une part, en la moyenne des variances des proximités de chacun des éléments descriptifs avec les autres (variances intra-lignes)

et d'autre part, la variance inter-lignes; cette dernière est une mesure de la distorsion par rapport à un état sphérique des données défini par $\bar{S}_l$ constant pour tout l, $S(l, k)$ est un produit scalaire pour le tableau transformé $\varepsilon_{lk} \to \varepsilon'_{lk}$ (*cf.* § II).

3.1 *Tests*

En se référant à l'hypothèse N (*cf.* § II), on peut aisément effectuer des tests.

1) Test d'absence de structure vis-à-vis de l'hypothèse de la « sériation» où le tableau peut être, au contenu d'un petit nombre de cases près, être ramené à la forme σ.

2) Test d'absence de structure par rapport à celle définie par une disposition sphérique des données pour la métrique qui nous intéresse.

Le premier test sera basé sur la plus grande valeur observée de $\sum_{\{k/k\neq l\}}(S_{lk} - \bar{S}_l)^2$ et le second sur la valeur de $\sum_{1\leqslant l\leqslant n}(S_l - \bar{S})^2$.

Si le tableau des données n'est pas de grande dimension, on peut effectuer les tests à partir de simulations du tableau d'incidence dans l'hypothèse N, un nombre suffisant de fois; nous disposons d'un programme qui permet de le faire. On comparera les valeurs observées des statistiques avec leurs distributions empiriques.

Autrement, on peut constater que pour p grand, dans l'hypothèse N, la suite des valeurs:

$$S(l, 1), S(l, 2), \ldots, S(l, l-1), S(l, l+1), \ldots, S(l, n) \qquad (*)$$

des proximités d'une ligne l fixée avec les autres lignes du tableau, peut être considérée comme une suite de $(n-1)$ réalisations indépendantes d'une variable aléatoire normale centrée réduite; de sorte que $\sum_{\{k/k\neq l\}}(S_{lk} - \bar{S})^2$ est pour l fixé, la réalisation d'un χ^2 à $(n-2)$ degrés de liberté. Les n valeurs de cette statistique obtenues pour l variant de 1 à n, ne sont pas rigoureusement indépendantes dans l'hypothèse N; en effet, pour les n suites telles que (*), chaque $S(l, k)$ pour l différent de k, se retrouve dans exactement deux suites différentes. Toutefois ce degré de dépendance est faible et l'est d'autant que n est grand. Par conséquent, on se référera à la distribution de la plus grande valeur de n statistiques indépendantes du χ^2 à $(n-2)$ degrés de liberté pour juger de l'importance relative de la plus grande valeur observée de $\sum_{\{k/k\neq l\}}(S_{lk} - \bar{S}_l)^2$. De même on se référera à la loi du χ^2 à $(n-1)$ degrés de libertés pour juger de la relative petitesse de $\sum_{1\leqslant l\leqslant n}(S_l - \bar{S})^2$.

4. Détermination d'un plan de représentation géométrique

Algorithme [1]

L'indice l_1 pour lequel est maximum la variance $\frac{1}{(n-1)}\sum_{\{k/k\neq l\}}(S_{lk} - \bar{S}_l)^2$ définira le premier axe $Ax1$ du plan de représentation; Ax1 sera pris horizontal et orienté de gauche à droite.

L'élément l_2 qui définira le second axe Ax2 doit satisfaire deux conditions:

1. Nous tenons à remercier très vivement M. J. M. Dernsteïn, de l'Institut de Programmation, qui a écrit le programme relatif à cet algorithme.

a) Une valeur de $S(l_1, l_2)$ voisine de 0 ; l_2 devant être assez indépendant de l_1. $S(l_1, l_2)$ est le produit scalaire des deux vecteurs lignes l_1 et l_2 de la matrice (ε'_{ij}) transformée de celle d'incidence (ε_{ij}).

b) Une valeur élevée de $\mathscr{V}(l_2)$ qui nous assurera du caractère discriminant du second axe.

Par conséquent, nous prendrons pour l'indice l_2, qui détermine Ax2, celui qui rend maximum le rapport homogène $\mathscr{V}(k) / (S(l_1, k)^2$.

Le point d'intersection des deux axes aura comme abscisse commune S (l_1, l_2) sur chacun des deux axes.

Si t est la valeur de $S(l_1, l_2)$, l'angle α des deux axes sera défini par $\pi(t)$ π

$$\text{où} \qquad \pi(t) = \Pr^N \{ S(l, k) < t \} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-x^2/2}\, dx.$$

Ainsi, l'angle des deux axes définit le degré d'indépendance des deux attributs de description qu'ils représentent respectivement.

Sur le plan des deux axes, un objet k sera représenté par le point de coordonnée $(S(l_1, k), S(l_2, k))$

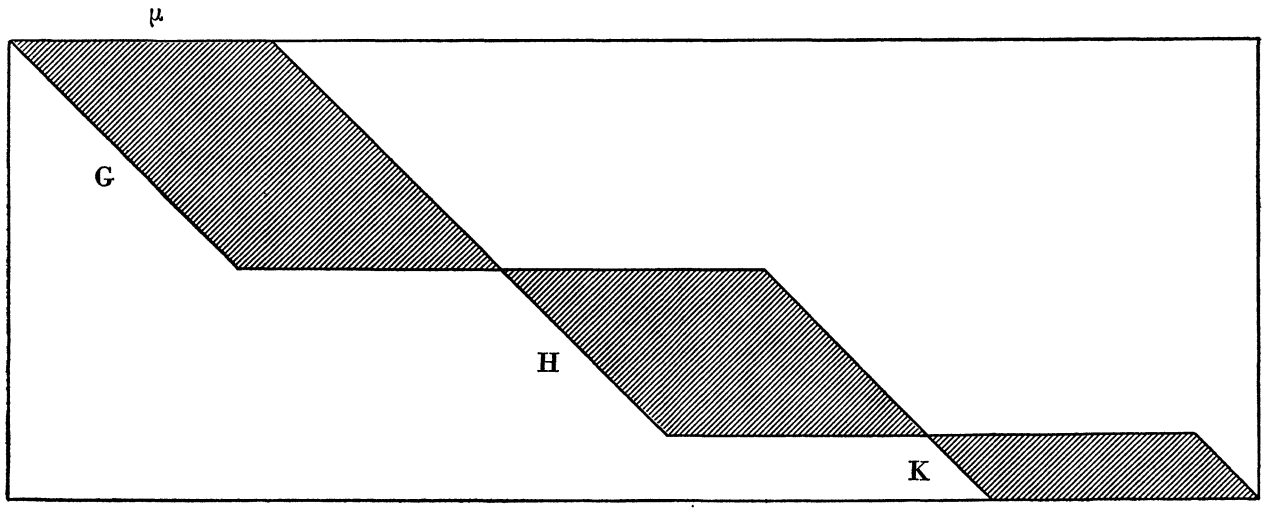
On peut de proche en proche déterminer de nouveaux axes de référence définissant, lorsqu'elles existent, de nouvelles dimensions ; ainsi le troisième axe Ax3 sera défini à partir de l'indice l_3 qui satisfait :

$$max \left\{ min \left(\frac{\mathscr{V}(k)}{(S(l_1, k))^2}, \frac{\mathscr{V}(k)}{(S(l_2, k))^2} \right) \right\}.$$

Examinons dans le plan des deux premiers axes Ax1 et Ax2, les exemples 1) et 2) du paragraphe précédent. Pour l'exemple 1), les points représentant les lignes du bloc H (resp. G) se trouvent tous situés dans l'ordre défini par la forme σ du tableau d'incidence, sur Ax1 (resp. Ax2).

Pour l'exemple 2), les points représentant les lignes du bloc G (resp. H) se trouvent rangés selon Ax1 (resp. Ax2) dans l'ordre défini par la forme σ. La plupart des points relatifs au bloc G (resp. H) se trouvent *sur* Ax1 (resp. Ax2), ceux qui n'y sont pas sont au voisinage de l'origine.

Considérons un troisième exemple où le tableau d'incidence peut prendre la forme ci-dessous.



Les points relatifs aux lignes de G (resp. H) sont situés, dans le plan des deux premiers axes, sur Ax1 (resp. Ax2) selon l'ordre défini par la forme σ. Les diverses lignes du bloc K sont représentés en un même point d'égales coordonnées $(-\sqrt{p}\ \mu, -\sqrt{p}\ \mu)$. Un troisième axe Ax3 définira la dimension sous-jacente à K.

Nous nous proposons, au paragraphe suivant, d'expliciter la méthode par rapport à l'analyse factorielle présentée du point de vue de J. P. Benzécri (*cf.* [2]).

VI. COMPARAISON AVEC L'ANALYSE FACTORIELLE

Discutons la recherche du premier axe puisque les autres s'en déduisent de proche en proche.

En ce qui nous concerne, le premier axe de discrimination Ax1 est défini à partir d'un vecteur ligne du tableau d'incidence transformé. Si $\vec{\eta}_k$ désigne un vecteur ligne courant

$$\vec{\eta}_k = (((\varepsilon_{kj} - \mu_k) / \sqrt{\mu_k \sqrt{p}}) / j = 1, 2, \ldots, p)$$

l'incidence l_1 qui définit le premier axe est, rappelons le, celui qui rend maximum par rapport à l.

$$\sum_{k=1}^{n} \left\{ \vec{\eta}_l \left(\vec{\eta}_k - \vec{\eta}_l \left(\frac{1}{n} \sum_{k=1}^{n} \vec{\eta}_k \right) \right) \right\}^2.$$

Si les lignes représentent par exemple des attributs de description, c'est un attribut effectivement présent qui sera le point extrême droite du premier axe qu'il caractérise. La suite des projections, au sens de notre métrique, des autres attributs sur cet axe définira le premier facteur.

Pour l'analyse factorielle la plus proche de la méthode, le premier axe est défini par un vecteur unitaire $\vec{v}$, celui qui rend maximum

$$\sum_{k=1}^{n} (\vec{\eta}_k \,.\, \vec{v})^2$$

$\vec{\eta}_k \,.\, \vec{v}$ étant le produit scalaire euclidien.

Par conséquent, dans notre méthode il s'agit, intuitivement parlant, d'un jugement des données de l'intérieur. La technique ne nécessite pas la diagonalisation d'une matrice.

Signalons qu'un des points de départ de ce travail a été une remarque pratique: cherchant à découvrir une classification sur des données pour lesquelles une analyse factorielle des correspondances avait été appliquée, nous avons commencé par déterminer au moyen d'une technique analysée dans ([7], (c)) et dont ce texte présente une généralisation systématique, les éléments les plus neutres et ceux, les plus discriminants. Nous avons observé, dans le plan des deux premiers axes factoriels, les éléments les plus neutres se grouper autour de l'origine et ceux les plus discriminants aux extrémités du premier axe dont l'importance était d'ailleurs sensiblement supérieure au second.

En fait, c'est au niveau d'une classe bien cohérente, résultant d'une classification automatique, que nous envisageons d'appliquer notre méthode pour étudier la position relative des divers éléments de la classe.

Ce traitement s'applique à n'importe quel tableau de données $A \times E$ pourvu que les variables descriptives de A définissent toutes le même type de structure algébrique sur E. L'analyse n'est en effet basée que sur les proximités et nous avons étendu, en le précisant, le principe de définition de la mesure

de proximité entre deux variables indicatrices de parties sur E (*cf.* § II), au cas d'un couple de variables définissant, soit un couple de partitions, soit un couple de préordres totaux, soit un couple d'ordres totaux, soit enfin, un couple de mesures positives sur E (*cf.* [7] (d)).

Il arrive souvent dans des études pour le développement économique et social que A soit formé d'échelles (chaque variable détermine sur E un préordre total). Si une classification automatique sur A permet de dégager les principales « dimensions » du développement, on peut espérer que l'analyse d'une même classe de variables par cette méthode, suivie d'une recherche plus précise étudiée dans ([7]), (a) établira pour la « dimension » étudiée, un enchaînement entre les différents états du développement.

BIBLIOGRAPHIE

[1] ACHARD, P., "Biais statistique sur les indices de similarité", *note interne*, CMAC (Maison des Sciences de l'Homme), Paris, sept. 1970.

[2] BENZÉCRI, J. P., *a*) "Ordre latéral entre lois de probabilités sur un ensemble ordonné" ; *b*) "Représentations euclidiennes" ; *c*) "Distance distributionnelle et métrique du χ^2 en analyse factorielle des correspondances", *Publications du Laboratoire de Statistique Mathématique*, Paris, ISUP.

[3] BERTIN, J., "Traitement graphique de l'information", *Annales*, 24e année, Paris, Armand Colin, janvier-février 1969.

[4] FELLER, W., *An introduction to probability theory and its applications*, vol. I, 2nd ed., New York-London, John Wiley, 1964.

[5] KENDALL, D. G., "Seriation from abundances matrices", *Proc. Conference on mathematical methods in the archaeological and historical sciences*, Mamaïa-Roumanie, sept. 1970.

[6] KRUSKAL, J. B., "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis", *Psychometrika*, vol. 29, 1964.

[7] LERMAN, I. C., *a*) "Essai sur l'analyse hiérarchique", *Math. Sci. hum.*, no 17, 1966, pp. 37-46 ; *b*) *Les bases de la classification automatique*, coll. Programmation, Paris, Gauthier-Villars, 1970 ; *c*) "Sur l'analyse des données préalable à une classification automatique", *Math. Sci. hum.*, no 32, 1970 ; *d*) "Mesure de proximité entre structures algébriques de même type ; application à la classification automatique", *rapport interne*, CMAC (Maison des Sciences de l'Homme), Paris, juin 1971.

[8] SHEPARD, R. N., "The analysis of proximities : multidimensional scaling with an unkown distance function", *Psychometrika*, vol. 27, 1962.

[9] DE LA VEGA, W. F., "Sur deux techniques de sériation", *note*, Centre d'Analyse Documentaire pour l'Archéologie (CNRS), Marseille, 1971.