MATHÉMATIQUES ET SCIENCES HUMAINES

B. ESCOFIER
B. LEROUX

Étude des questionnaires par l'analyse des correspondances. Modification du codage des questions ou de leur nombre et stabilité de l'analyse

Mathématiques et sciences humaines, tome 49 (1975), p. 5-27
<http://www.numdam.org/item?id=MSH_1975__49__5_0>

© Centre d'analyse et de mathématiques sociales de l'EHESS, 1975, tous droits réservés.

L'accès aux archives de la revue « Mathématiques et sciences humaines » (http://msh.revues.org/) implique l'accord avec les conditions générales d'utilisation (http://www.numdam.org/conditions). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM
Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
http://www.numdam.org/

Math. Sci. hum. (13e année, n°49, 1975, p. 5-27)

ETUDE DES QUESTIONNAIRES PAR L'ANALYSE DES CORRESPONDANCES MODIFICATION DU CODAGE DES QUESTIONS OU DE LEUR NOMBRE ET STABILITE DE L'ANALYSE

par

B. ESCOFIER* et B. LEROUX**

L'analyse des correspondances est souvent utilisée dans le traitement des résultats d'un questionnaire et cette méthode s'est révélée particulièrement efficace à condition de prendre certaines précautions en codant les données. Le plus souvent il est préférable d'associer à chaque question un ensemble de modalités de réponses s'excluant mutuellement, chaque individu devant choisir une des modalités et une seule. On construit alors un tableau de correspondance entre l'ensemble I des individus interrogés et l'ensemble J des modalités de réponses à toutes les questions, en codant $k_{ij} = 1$ si l'individu i a choisi la modalité j et zéro sinon. Les données sont alors sous forme "disjonctive complète.

La plupart des questionnaires se prêtent à ce type de codage : par exemple, dans le cas de questions n'admettant que les réponses oui et non, on code deux modalités de réponses pour chaque question, l'une associée au oui, et l'autre au non. Quand une question correspond à une variable continue, (par exemple l'âge), on la partitionne en quelques classes ; on appelle alors modalité de réponse, chacune de ces classes. Ce codage très souple permet de mêler variables logiques et quantitatives.

L'analyse de données de ce type a des propriétés particulières intéressantes qu'il est utile de connaître pour le traitement et l'interprétation des résultats. Nous rappelons brièvement ces propriétés dans le premier paragraphe.

L'étude de variables quantitatives par ce codage pose immédiatement le problème du choix du nombre de classes et son influence sur les

* Département de Mathématiques. I.N.S.A., Rennes.
** U.E.R. de Mathématiques. Université René Descartes, Paris.

facteurs. La multiplication du nombre de classes permet de décrire plus finement la variable mais n'est pas sans inconvénient : la taille du tableau augmente, l'effectif dans chaque classe diminue et les résultats deviennent souvent difficilement interprétables.

Dans cet article, nous avons étudié ce problème. Nous montrons que lorsque le nombre de questions est grand, et que la valeur propre associée au facteur est bien séparée des autres, alors le regroupement de modalités d'une question perturbe peu le facteur. Mais le plus souvent, les facteurs peuvent changer et même disparaître. Le codage est un problème délicat mais très important : deux codages apparemment proches peuvent donner des résultats d'analyse différents.

Un autre problème se pose souvent au niveau de l'interprétation : il peut apparaître, dans une analyse, une question peu homogène aux autres ou bien d'inertie prédominante pour un facteur. On fait souvent une nouvelle analyse en la supprimant afin de mesurer son influence sur la détermination des facteurs. Les résultats présentés ici permettent dans certains cas d'assurer la stabilité des facteurs et donc d'éviter une autre analyse. Nous traitons plus particulièrement des questions à deux modalités de réponses, la connaissance de leur position par rapport aux axes factoriels, permet en effet d'affiner considérablement les résultats.

L'influence sur les facteurs de l'adjonction d'une ou plusieurs questions se pose souvent. Dans certains cas, les réponses à ces questions n'étaient pas disponibles au moment de l'analyse, ou bien elles n'ont pas été traitées pour diminuer la taille du tableau à analyser ou parce qu'elles paraissaient peu homogènes aux autres.

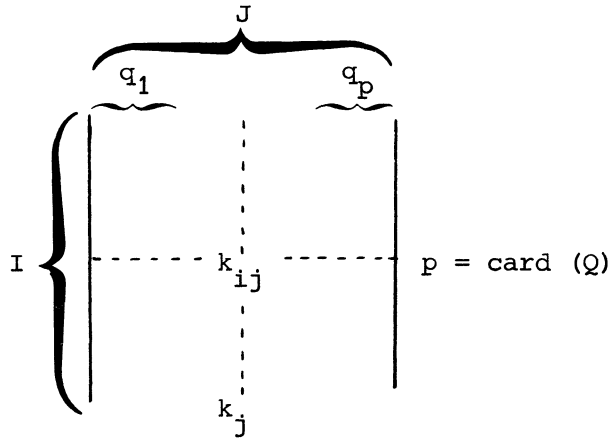
Nous étudions aussi ce problème et nous pouvons dans certains cas, assurer la stabilité des facteurs ; les questions à deux modalités de réponses mises en éléments supplémentaires font l'objet d'une étude particulière dont les résultats semblent très intéressants.

Enfin, dans un dernier paragraphe, nous rappelons des résultats relatifs aux éléments propres de deux endomorphismes symétriques de $\mathbb{R}^n$, A et C = A + B, résultats que nous utilisons dans les paragraphes précédents (démontrés en [5] et [6]).

1 - PROPRIETES DE L'ANALYSE DES QUESTIONNAIRES MIS SOUS FORME DISJONCTIVE COMPLETE [1] et [3]

1-0 : Notations

R_I désigne l'espace vectoriel des mesures sur l'ensemble fini I, R^I celui des fonctions sur I.



On note k_{IJ} la correspondance définie sur I × J, J est divisé en une suite Q de sous-ensembles q (ou questions). Les nombres k_{ij} sont égaux à zéro ou un. On note $k_j = \sum_{i \in I} k_{ij}$

Les données étant sous forme disjonctive complète :

$$\forall i \in I, \forall q \in Q \qquad \sum_{j \in q} k_{ij} = 1$$

$$\sum_{j \in J} k_{ij} = \text{card}(Q) = p$$

$$\forall q \in Q \qquad \sum_{j \in q} k_j = \text{card}(I) = n$$

et

$$\sum_{i \in I} \sum_{j \in J} k_{ij} = np$$

Désignons par f_I^j le profil de l'élément $j \in J$: $f_I^j = \{k_{ij}/k_j \mid i \in I\}$

En analyse des correspondances on définit le nuage

$N(J) = \{(f_I^j, f_j = k_j/np) \mid j \in J\}$ dans l'espace vectoriel $\mathbb{R}_I$ muni de la métrique M^{II} ($m^{ii'} = \delta^i_{i'}\, n$)

1-1 : Distance du χ^2 entre l'élément de J

Soit X l'ensemble des individus ayant choisi la modalité de réponse j

et Y l'ensemble des individus ayant choisi la modalité de réponse j', X Δ Y la différence symétrique entre X et Y, alors la distance du chi-deux entre j et j' est :

$$d^2(j,j') = n \times \frac{\text{card}(X \Delta Y)}{\text{card}(X) \times \text{card}(Y)}$$

Dans le cas particulier où j et j' s'excluent mutuellement, par exemple j et j' sont deux modalités de réponses d'une même question, alors :

$$d^2(j,j') = n \left(\frac{1}{k_j} + \frac{1}{k_{j'}}\right)$$

Si g désigne le centre de gravité du nuage $d^2(j,g) = \frac{n}{k_j} - 1$

la distance augmente avec la rareté de la réponse.

1-2 : Distance du χ^2 entre éléments de I

$$d^2(i,i') = \frac{n}{p} \sum \left\{ \frac{1}{k_j} \mid j \in \{\text{modalités de réponses choisies par i et non par i'}\}\right\}$$

1-3 : Barycentre des réponses à une même question

$$\forall q \in Q \qquad \sum_{j \in q} \frac{k_j}{n} \frac{k_{ij}}{k_j} = \frac{1}{n} = \sum_{j \in J} \frac{k_j}{np} \frac{k_{ij}}{k_j}$$

Autrement dit, les modalités de réponses d'une même question ont le même barycentre que N(J). Cette propriété se conservant en projection, pour chaque espace factoriel, les modalités de réponses d'une même question ont donc leur barycentre à l'origine. En particulier si q a deux modalités de réponses q^+ et q^-, elles sont alignées avec le centre de gravité du nuage.

1-4: Inertie des éléments de J (questions)

$$\forall j \in J \text{ on a :} \qquad In(j) = \frac{n - k_j}{np}$$

où In(j) désigne l'inertie de j, cette inertie est d'autant plus

faible que la modalité de réponse est souvent choisie.

Pour une question, on a :

$$- In(q) = \sum_{j \in q} In(j) = \frac{[card(q) - 1]}{p}$$

Si toutes les questions ont le même nombre de modalités de réponses, alors elles ont toutes même inertie. Dans le cas où les p questions q n'ont que deux modalités de réponses, $In(q) = \frac{1}{p}$

- l'inertie du nuage $In(J) = \frac{card\ J}{p} - 1$, elle vaut $(r-1)$ si toutes les questions ont r modalités de réponses, et 1 si elles n'en ont que deux.

1-5 : Equivalence entre l'analyse I × J et celles des correspondances I × I et J × J déduite de la correspondance I × J

De la correspondance I × J, on déduit une correspondance s_{JJ} symétrique sur J × J telle que :

$$\forall j \in J \quad \forall j' \in J \quad s_{jj'} = \sum_{i \in I} k_{ij}\, k_{ij'}$$

Alors $s_{jj'}$ est le nombre d'individus qui ont choisi simultanément les modalités de réponses j et j'.

Tout facteur sur J de la correspondance I × J associé à la valeur propre σ est facteur de la correspondance J × J associé à la valeur propre σ^2 et réciproquement. L'analyse d'un tableau mis sous forme disjonctive complète est donc équivalente à celle d'un tableau de fréquences.

De même, l'analyse de la correspondance symétrique t_{II} définie sur I × I par

$$\forall i \in I \quad \forall i' \in I \quad t_{ii'} = \sum_{j \in J} \frac{k_{ij}\, k_{i'j}}{k_j}$$

est identique à celle de la correspondance I × J. Ce dernier résultat est d'ailleurs valable pour toute correspondance.

1-6 : Analyse de la sous-correspondance I × q associée à une question

Etudions la correspondance k_{Iq} entre l'ensemble des individus et l'ensemble des modalités de réponses de la question q. La correspondance symétrique q × q déduite (cf. § 1-5) est diagonale puisque les modalités de réponses d'une même question s'excluent deux à deux. Les éléments diagonaux sont égaux à k_j, nombre d'individus ayant choisi la modalité de réponse j.

Si la question q a r modalités de réponses, alors la correspondance I × q a (r-1) valeurs propres égales à un et les autres sont nulles.

2 - AJOUT (ET SUPPRESSION) D'UNE QUESTION : CAS GENERAL

Dans ce paragraphe, on compare les résultats de l'analyse I × J à ceux de l'analyse I × (JUq_a), q_a désigne la question ajoutée et r le nombre de ses modalités de réponses.

Dans le cas des questionnaires mis sous forme disjonctive complète, la distribution marginale sur I est uniforme, les nuages N(J) et $N(JUq_a)$ et $N(q_a)$ ont donc même centre de gravité et induisent sur R_I la même métrique M^{II}.

2- 1 : Formes quadratiques d'inertie

Les trois nuages N(J), $N(JUq_a)$ et $N(q_a)$ ayant même centre de gravité, on peut comparer les facteurs non triviaux de ces nuages par l'étude de leur forme quadratique d'inertie autour de l'origine, notées respectivement S_{II}, T_{II} et R_{II}.

On rappelle que $S_{ii'} = \sum_{j \in J} k_{ij}\, k_{i'j}/(np \times k_j)$

On a : $T_{II} = \frac{p}{(p+1)}\, S_{II} + \frac{1}{(p+1)}\, R_{II}$

Les facteurs (et leur inertie) des nuages N(J), $N(JUq_a)$ et $N(q_a)$ sont respectivement les éléments propres de $M^{II} \circ S_{II}$, $M^{II} \circ T_{II}$ et $M^{II} \circ R_{II}$, applications M^{-1}-symétriques.

On peut donc appliquer les résultats relatifs à la somme de deux endomorphismes symétriques (cf. § 5-2). On désigne respectivement par σ, τ et ρ les valeurs propres de MoS, MoT et MoR, rangées par ordre décroissant.

2-2 : Comparaison des valeurs propres des correspondances $I \times J$ et $I \times (JUq_a)$

La question q_a ayant r modalités de réponses, le nuage $N(q_a)$ a (r-1) valeurs propres égales à un, les autres étant nulles, on a donc $\rho_1 = 1$ et $\rho_n = 0$. On obtient donc :

Pour les valeurs propres σ de $I \times J$ et τ de $I \times (JUq_a)$, p étant le nombre de questions, q_a exclue, on a :

$$\forall s \in]n] \quad \frac{p}{p+1}\,\sigma_s \leqslant \tau_s \leqslant \frac{p}{p+1}\,\sigma_s + \frac{1}{p+1}$$

Pour les taux d'inertie notés taux (J) et taux (JUq_a) extraits par les s premiers facteurs des nuages N(J) et $N(JUq_a)$:

$$\frac{\text{card } J-p}{\text{card}J+r-(p+1)}\ \text{taux (J)} \leqslant \text{taux } (JUq_a) \leqslant \frac{\text{card } J-p}{\text{card}J+r-(p+1)}\ \text{taux (J)} +$$

$$+ \begin{cases} \dfrac{s}{\text{card}J+r-(p+1)} & \text{si } s \leqslant r-1 \\ \dfrac{r-1}{\text{card}J+r-(p+1)} & \text{si } s > r-1 \end{cases}$$

Si les questions ont toutes le même nombre r de modalités de réponses on a :

$$\forall s \in]n] \quad \frac{p}{p+1}\ \text{taux(J)} \leqslant \text{taux}(JUq_a) \leqslant \frac{p}{p+1}\ \text{taux (J)} + \begin{cases} \dfrac{s}{(r-1)(p+1)} & \text{si } s \leqslant r-1 \\ \dfrac{1}{(p+1)} & \text{si } s > r-1 \end{cases}$$

Remarque

On ne peut conclure sur le sens de variation des valeurs propres et des taux d'inertie quand on ajoute une question. Naturellement, si le nombre p de questions est grand les valeurs propres et les taux d'inertie seront peu modifiés

II-3 : Comparaison des facteurs sur I des correspondances I × J et I × (JUq_a)

Pour comparer deux à deux les facteurs des correspondances I × J et I × (JUq_a), on étudie leur angle dans R^I muni de la métrique M^{-1}. Le cosinus de cet angle n'est autre que la corrélation entre les facteurs, plus l'angle est petit, plus la corrélation est proche de un. Mais il n'est pas suffisant de comparer les facteurs deux à deux. En effet, si deux valeurs propres sont voisines, les facteurs associés sont peu stables, ils peuvent s'échanger, mais souvent le plan qu'ils engendrent ne varie pas. Dans certains cas, on ne peut assurer la stabilité d'un facteur, il est alors intéressant d'examiner celle d'un sous-espace qui le contient. Pour cela nous étudions l'écart entre deux sous-espaces de R^I engendrés par des facteurs de même rang. On mesure cet écart par l'angle maximum entre un vecteur de l'un et sa projection orthogonale sur l'autre [§ 5-1]. Le cosinus de cet angle minore la corrélation multiple entre un facteur de l'un de ces deux sous-espaces et ceux engendrant l'autre. Plus l'angle est petit, plus les deux sous-espaces factoriels sont proches. Pour mesurer la stabilité d'un sous-espace on cherche la borne supérieure de cet angle, pour cela on applique les résultats du § 5-2.

- l'angle θ entre les sous-espaces engendrés par les s premiers facteurs sur I des nuages N(J) et N(JUq_a) est tel que :

$$\text{Si } \sigma_s - \sigma_{s+1} > \frac{1}{p} \text{ alors } \theta < \frac{\pi}{4} \text{ et } \sin 2\theta < \frac{1}{p(\sigma_s - \sigma_{s+1})}$$

- l'angle θ entre les sous-espaces engendrés par lesfacteurs de rang s, s+1, ... , s + ℓ est tel que

Si $\delta p \geqslant 1$ avec $\delta = \inf\{(\sigma_{s-1} - \sigma_s), (\sigma_{s+\ell} - \sigma_{s+\ell+1})$

Alors $\sin\theta \leqslant \dfrac{1}{2p\delta - 1}$

Pour comparer les facteurs deux à deux il suffit de poser $\ell = 0$ dans la formule précédente.

Ces majorations sont d'autant plus petites que les valeurs propres de l'espace factoriel considéré sont bien séparées des autres et que le nombre de questions est grand.

2-4 : Ajout de plusieurs questions

On généralise immédiatement les résultats du paragraphe précédent au cas de l'ajout d'un nombre quelconque p_a de questions. Pour les valeurs propres, on a :

$$\forall s \in]n] \quad \frac{p_a}{p + p_a}\ \sigma_s \leqslant \tau_s \leqslant \frac{p}{p + p_a}\ \sigma_s + \frac{p_a}{p + p_a}$$

L'angle θ entre les sous-espaces invariants engendrés par les s premiers facteurs est tel que :

$$\text{Si } \sigma_s - \sigma_{s+1} \geqslant \frac{p_a}{p} \quad \text{alors} \quad \theta < \frac{\pi}{4} \quad \text{et} \quad \sin 2\theta \leqslant \frac{p_a}{p(\sigma_s - \sigma_{s+1})}$$

L'angle θ entre les sous-espaces engendrés par les facteurs de rang s, ... , s+ℓ est tel que :

$$\text{si } \delta p \geqslant p_a \quad \text{avec} \quad \delta = \inf((\sigma_{s-1} - \sigma_s), (\sigma_{s+\ell} - \sigma_{s+\ell+1})) \quad \text{Alors } \sin\theta \leqslant \frac{p_a}{2p\delta - 1}$$

2-5 : Suppression d'une question

Etudions maintenant les perturbations des résultats de l'analyse provoqués par la suppression d'une question q_a.

De manière analogue au paragraphe précédent, on compare les valeurs propres (σ) de l'analyse où la question q_a a été supprimée à celles (τ) de l'analyse des p+1 questions :

$$\forall s \in]n] \quad \frac{p+1}{p}\tau_s - \frac{1}{p} \leqslant \sigma_s \leqslant \frac{p+1}{p}\tau_s$$

Soit θ l'angle entre les sous-espaces engendrés par les s premiers facteurs des correspondances $I \times J$ et $I \times (JUq_a)$:

$$\text{Si } \tau_s - \tau_{s+1} > \frac{1}{p+1} \quad \text{alors } \theta < \frac{\pi}{4} \text{ est sin } 2\theta < \frac{1}{(p+1)(\tau_s - \tau_{s+1})}$$

Remarquons que ces résultats se déduisent de ceux du paragraphe précédent en remplaçant σ par τ et p par p+1.

2-6 : Commentaires

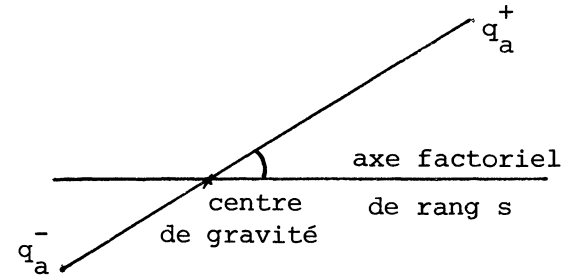
Les bornes obtenues ci-dessus pour l'angle entre les sous-espaces engendrés par les premiers facteurs sont optima ; elles peuvent être atteintes. Ceci montre bien que, lorsque les valeurs propres sont peu séparées ou que le nombre de questions ajoutées ou retirées n'est pas très inférieur à celui des questions analysées, les perturbations sur les facteurs peuvent devenir très importantes. Ceci a souvent été constaté en pratique ; en particulier, la suppression de questions peut entraîner la disparition complète d'un facteur. Les tableaux de données mis sous forme disjonctive complète sont sensibles aux modifications du nombre de questions étudiées.

Sans informations supplémentaires sur les questions ajoutées ou retirées, on ne peut obtenir de meilleures majorations. Cependant, lorsque ces questions sont à deux modalités de réponses, on connait, (ou on peut calculer facilement) leurs positions par rapport aux axes factoriels ; Ce problème fait l'objet du paragraphe suivant : on précisera la variation des facteurs et on améliorera considérablement les bornes.

3 - AJOUT (ET RETRAIT) D'UNE QUESTION A DEUX MODALITES DE REPONSES q_a^+ et q_a^-

Les résultats précédents s'appliquent évidemment à ce cas particulier, mais on peut obtenir des majorations beaucoup plus fines en utilisant les carrés des corrélations de la question avec chaque axe factoriel. Ils sont calculés dans les programmes d'analyse des correspondances. Pour les éléments intervenant dans l'analyse et pour ceux mis en éléments supplémentaires, ils sont égaux aux contributions relatives de la question au facteur. Lorsqu'on retire une question, on connait donc toujours ces coefficients, par contre si on l'ajoute, il faut qu'elle ait été étudiée en élément de poids nul.

Nous avons vu au § 1, que les deux modalités de réponses sont alignées avec le centre de gravité du nuage et que la correspondance $I \times \{q_a^-, q_a^+\}$ n'a qu'une valeur propre non nulle.



La forme quadratique d'inertie du sous-nuage $\{q_a^+, q_a^-\}$, par rapport au centre de gravité, est de rang 1, l'angle de son vecteur propre avec chaque facteur est donné par la contribution relative de q_a^+ ou q_a^- à ce facteur. Nous pouvons donc appliquer les résultats du § 5-3 à la somme (p+1) MoT = p MoS + MoR puisque MoR est de rang un.

Soit Φ_s (resp. ψ_s) l'angle entre la question q_a et le sous-espace factoriel correspondant aux s plus grandes valeurs propres de la correspondance $I \times J$ (resp. $I \times (JUq_a)$), alors $\cos^2\Phi_s$ (resp. $\cos^2\psi_s$) est égal à la somme des contributions relatives de q_a aux s premiers facteurs de $I \times J$ (resp. $I \times (JUq_a)$).

3 -1 : Comparaison des valeurs propres de $I \times J$ et $I \times (JUq_a)$

Nous étudions les perturbations provoquées sur les valeurs propres de $I \times J$ par l'adjonction de la question q_a

Entre les valeurs propres σ de I × J et τ de I × (JUq_a), p étant le nombre de questions (q_a exclue), on a les relations :

$$\forall s \in]n], \; s \neq 1 \; : \quad \frac{p}{p+1}\,\sigma_s \leqslant \tau_s \leqslant \frac{p}{p+1}\,\sigma_s + \frac{\sin^2\Phi_{s-1}}{p+1}$$

Pour la somme des s premières valeurs propres, on a :

$$\forall s \in]n] \quad \frac{p}{p+1}\sum_{i=1}^{s}\sigma_i + \frac{\cos^2\Phi_s}{p+1} \leqslant \sum_{i=1}^{s}\tau_i \leqslant \frac{p}{p+1}\sum_{i=1}^{s}\sigma_i + \frac{1}{p+1}$$

(Cette inégalité permet en particulier de comparer τ_1 et σ_1).

Pour les taux d'inertie notés taux (J) et taux (JUq_a) extraits par les s premiers facteurs, on a :

$$\frac{\text{card J - p}}{\text{cardJ-p+1}}\ \text{taux (J)} + \frac{\cos^2\Phi_s}{\text{cardJ-p+1}} \leqslant \text{taux } (JUq_a) \leqslant \frac{\text{card J-p}}{\text{cardJ-p+1}}\ \text{taux (J)} + \frac{1}{\text{cardJ-p+1}}$$

Si chaque question a deux modalités de réponses, la formule se simplifie :

$$\frac{p}{p+1}\ \text{taux (J)} + \frac{\cos^2\Phi_s}{p+1} \leqslant \text{taux } (JUq_a) \leqslant \frac{p}{p+1}\ \text{taux (J)} + \frac{1}{p+1}$$

De manière analogue, on étudie les perturbations dûes au <u>retrait de la question q_a</u>, on a par exemple :

$$\forall s \in]n] \qquad \frac{p+1}{p}\,\tau_s - \frac{\cos^2\psi_s}{p} \leqslant \sigma_s \leqslant \frac{p+1}{p}\,\tau_s$$

<u>Commentaires</u>

Quand on étudie l'<u>ajout d'une nouvelle question</u> à l'analyse, la connaissance de ses contributions relatives aux facteurs, permet d'améliorer les résultats du § 2-2, d'autant plus que Φ_{s-1} est plus petit. Les valeurs propres diminuent dès que $\sin^2\Phi_{s-1} < \sigma_s$, en particulier si la question q_a appartient au sous-espace factoriel correspondant aux s premiers facteurs, les va-

leurs propres de rang supérieur sont multipliées par p/(p+1). En général, on ne peut déterminer le sens de variation des taux d'inertie, cependant si Φ_s est petit, plus précisément si $\cos^2\Phi_s >$ taux(J), le taux d'inertie extrait par les s premiers facteurs augmente.

De même, quand on supprime une question de l'analyse, les résultats du § 2 -5 sont améliorés. Les valeurs propres augmentent dès que $\cos^2\psi_s$ est inférieur à τ_s ; si la question ajoutée est orthogonale au sous-espace associé aux s premiers facteurs, les valeurs propres de rang supérieur sont multipliées par le coefficient (p+1)/p.

3.2 : Comparaison des facteurs sur I des correspondances I × J et I ×(JUq$_a$)

Pour l'ajout d'une question les résultats du § 2-3 sont améliorés. L'angle θ entre les sous-espaces engendrés par les s premiers facteurs sur I des nuages N(J) et N(JUq$_a$), p étant le nombre de questions (q_a exclue), est tel que :

$$\text{si} \quad \sigma_s - \sigma_{s+1} > \frac{1}{p} \quad \text{alors} \quad \theta < \frac{\pi}{4} \quad \text{et} \quad \operatorname{tg} 2\theta \leq \frac{\sin 2\Phi_s}{p(\sigma_s - \sigma_{s+1}) + \cos 2\Phi_s}$$

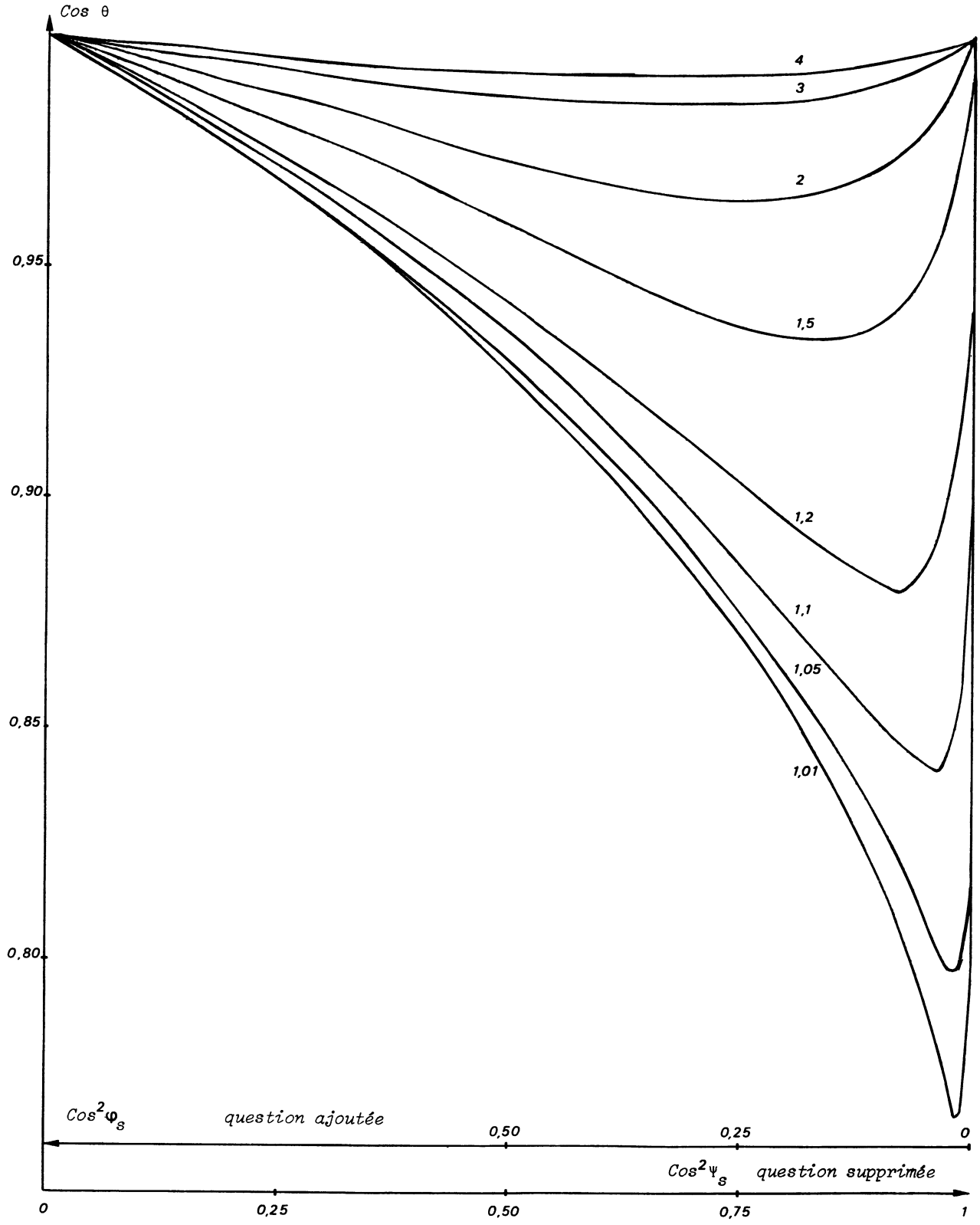
L'angle θ entre les sous-espaces engendrés par les facteurs de rang s,..., s+ℓ de I × J et I ×(JUq$_a$), Φ désignant l'angle entre q_a et le sous-espace factoriel associé à I × J, est tel que

$$\text{si} \quad \delta = \inf \{(\sigma_{s-1} - \sigma_s), (\sigma_{s+\ell} - \sigma_{s+\ell+1})\} > \frac{1}{p} \sin^2 \Phi_{s-1}$$

$$\text{alors} \qquad \sin\theta < \frac{\sin\Phi \cos\Phi}{p\delta - \sin^2 \Phi_{s-1}}$$

De même, dans l'étude du retrait d'une question, on a pour les sous-espaces correspondant aux s premiers facteurs :

$$\text{si} \quad \tau_s - \tau_{s+1} > \frac{1}{p+1} \quad \text{alors} \quad \theta < \frac{\pi}{4} \quad \text{et} \quad \operatorname{tg} 2\theta \leq \frac{\sin 2\psi_s}{(p+1)(\tau_s - \tau_{s+1}) - \cos 2\psi_s}$$


Cos θ
4
3
2
1,5
1,2
1,1
1,05
1,01
0,95
0,90
0,85
0,80
Cos²φs
question ajoutée
0,50
0,25
0
Cos²ψs
question supprimée
0
0,25
0,50
0,75
1

Pour les sous-espaces intermédiaires, on obtient,

$$\text{si} \quad \delta = \inf \{ (\tau_{s-1} - \tau_s), (\tau_{s+\ell} - \tau_{s+\ell+1}) \} > \frac{1}{p+1} \cos^2 \psi_{s+\ell}$$

$$\text{alors} \quad \sin \theta \leq \frac{\sin \psi \cos \psi}{(p+1)\delta - \cos^2 \psi_{s+\ell}}$$

Commentaires

Avec les résultats du paragraphe précédent, il était rare de pouvoir assurer la stabilité des facteurs, alors qu'en utilisant les corrélations de la question ajoutée avec les axes factoriels, on peut très fréquemment le faire et donc éviter de nouvelles analyses.

Les courbes ci-jointes donnent en fonction des contributions relatives de la question ajoutée ($\cos^2 \Phi_s$) ou retirée ($\cos^2 \Psi_s$) la borne du cosinus de l'angle entre les sous-espaces associés aux s plus grandes valeurs propres des deux correspondances pour quelques valeurs d'un paramètre [$p(\sigma_s - \sigma_{s+1})$ pour l'ajout et $(p+1)(\tau_s - \tau_{s+1})$ pour le retrait] .
Les minimas de ces courbes correspondent aux majorations du paragraphe 2. Ces courbes présentent un pic, d'autant plus aigu que la valeur du paramètre est plus petite. Sauf à l'intérieur d'un très petit intervalle, les corrélations permettent d'améliorer notablement les résultats précédents et donc d'assurer la stabilité des facteurs.

Les faisceaux de courbes correspondant à l'ajout et au retrait se déduisent l'un de l'autre par symétrie par rapport à la droite $\cos^2 \Phi_s = \cos^2 \psi_s = \frac{1}{2}$. On les a donc figurés sur le même graphique en traçant deux axes horizontaux, l'un associé à $\cos^2 \Phi_s$ et l'autre à $\cos^2 \Psi_s$.

On constate que, dans le cas de l'ajout d'une question q_a, le sous-espace factoriel étudié ne varie pas du tout si q_a en est proche (si $\cos^2 \Phi_s > 1/2$, $\cos \theta$ est encore supérieur à 0,925 quand le paramètre vaut 1,01 , alors que son minimum est 0,767). Si la question q_a est dans

l'orthogonal du sous-espace factoriel étudié, celui-ci ne varie pas non plus, mais cette stabilité est détruite dès que l'on s'en éloigne un peu.

Dans le cas de la suppression d'une question, la situation est inversée, la stabilité est plus durable quand la question q_a est au voisinage de l'orthogonal du sous-espace étudié que lorsqu'elle en est proche.

Pour les facteurs de même rang, (ou, plus généralement, pour les sous-espaces factoriels intermédiaires), nos majorations montrent que si la question ajoutée ou retirée est proche de l'axe factoriel (sin Φ ou sin Ψ proche de zéro) ou lui est presque orthogonale (cos Φ ou cos Ψ proche de zéro), le facteur est stable.

Les majorations données pour les sous-espaces intermédiaires peuvent évidemment s'appliquer aux sous-espaces associés aux plus grandes valeurs propres. Les bornes obtenues sont moins bonnes mais les conditions d'application étant moins strictes, on est parfois amené à les utiliser . Elles permettent en particulier d'affirmer que le sous-espace est toujours fixe quand il contient la question q_a ou lui est orthogonal.

4 - REGROUPEMENT DE PLUSIEURS MODALITES DE REPONSES D'UNE QUESTION

On se propose ici d'étudier les modifications des résultats de l'analyse des correspondances dues au regroupement de plusieurs modalités de réponses J_c d'une même question en une modalité notée c. On dira qu'un individu a choisi cette modalité c s'il a choisi l'une des modalités regroupées. De la correspondance sur I × J on déduit ainsi une correspondance entre I et C l'ensemble des nouvelles modalités de réponses. Dans le tableau de correspondance, cela revient à remplacer les colonnes représentant les modailités à regrouper par une seule colonne égale à leur somme.

La correspondance I × C est du même type que la correspondance I × J : le tableau associé ne comprend que des 0 et des 1 et est sous forme disjonctive complète.

Pour simplifier, nous étudions le cas où l'on regroupe les modalités d'une seule question, la généralisation à plusieurs questions est immédiate.

Ceci est un cas particulier du problème du regroupement en classes pour une correspondance quelconque, que nous avons déjà traité [5]. D'après le théorème de Huyghens, la forme quadratique d'inertie T_{II} du nuage N(J) est la somme de la forme quadratique d'inertie S_{II} du nuage N(C) et de celle du sous nuage $N(J_c)$ par rapport à son centre de gravité. L'analyse de N(J) et de N(C) se fait dans le même espace euclidien $\mathbb{R}_I$ muni de la métrique M^{II}, les facteurs et leur inertie sont les éléments propres de $M^{II} o T_{II}$ et de $M^{II} o S_{II}$ respectivement. On pourra appliquer les résultats relatifs à la somme de deux endomorphismes symétriques [cf. § 5-2] à la somme :

$$M^{II} \circ T_{II} = M^{II} \circ S_{II} + M^{II} \circ R_{II}$$

Montrons que les valeurs propres de MoR sont nulles ou égales à 1/p.

$$\text{Soit} \quad I_c = \{i \mid i \in I, \sum_{j \in J_c} k_{ij} = 1\}$$

Le support du sous-nuage $N(J_c)$ est contenu dans $R_{I_c} \subset R_I$, MoR est donc nul en dehors de ce support. De plus, l'analyse de la correspondance $I_c \times J_c$ est celle d'une question à card (J_c) modalités de réponses exclusives, elle admet donc (cf. § 1-6) card J_c - 1 valeurs propres égales à un, les autres sont nulles. Or la forme quadratique des moments d'inertie du nuage $N(J_c)$, associé à la correspondance $I_c \times J_c$ est $\frac{np}{\text{card } I_c} \times R$, le centre de gravité du nuage est $\{x_i = \frac{1}{\text{card } I_c} \mid i \in I_c\}$, la métrique associée est donc égale à $\frac{\text{card } I_c}{n} M$. Les valeurs propres de $\left(\frac{\text{card } I_c}{n} M\right) \circ \left(\frac{np}{\text{card } I_c} R\right)$ sont égales à un (ou nulles), par conséquent celles de MoR sont égales à $\frac{1}{p}$ ou nulles.

4.1 : Comparaison des valeurs propres des correspondances I × J et I × C

Pour les valeurs propres de I × J et I × C, p étant le nombre de questions, on a :

$$\forall\, s \in]\, n\,] \quad \sigma_s \leqslant \tau_s \leqslant \sigma_s + \frac{1}{p}$$

Pour les taux d'inertie, notés taux(J) et taux(C) extraits par les s premiers facteurs des nuages N(J) et N(C) :

$$\text{taux(C)} \times \left(1 - \frac{\text{card } J_c - 1}{\text{card } J - p}\right) \leqslant \text{taux(J)} \leqslant \text{taux(C)} \times \left(1 - \frac{\text{card } J_c - 1}{\text{card } J - p}\right) + \begin{cases} \dfrac{1}{\text{card } J-p} & \text{si } s < \text{card } J_c - 1 \\ \dfrac{\text{card } J_c - 1}{\text{card } J-p} & \text{si } s \geqslant \text{card } J_c - 1 \end{cases}$$

Remarques

Le regroupement de plusieurs modalités de réponses, comme tout regroupement en classes, diminue les valeurs propres. Le nombre de modalités regroupées n'intervient pas. Si le nombre p de questions est grand, le regroupement de modalités d'une question diminue peu les valeurs propres. Mais si on a effectué des regroupements dans les p questions, entre les valeurs propres τ de I × J et σ de I × C, on a :

$$\sigma_s \leqslant \tau_s \leqslant \sigma_s + 1$$

Et on ne peut rien conclure de plus ; il en est de même pour le sens de variation des taux d'inertie extraits par les s premiers facteurs.

4.2 : Comparaison des facteurs sur I des correspondances I × J et I × C

- L'angle θ entre les sous-espaces engendrés par les s premiers facteurs sur I des nuages N(J) et N(C) est tel que :

$$\text{Si} \quad \sigma_s - \sigma_{s+1} > \frac{1}{p} \quad \text{alors} \quad \theta < \frac{\pi}{4} \quad \text{et} \quad \sin 2\theta \leqslant \frac{1}{p(\sigma_s - \sigma_{s+1})}$$

- L'angle θ entre les sous-espaces engendrés par les facteurs de rang s, s+1,....., s+ℓ est tel que

$$\text{si} \quad \delta p \geq 1 \quad \text{avec} \quad \delta = \inf \{(\sigma_{s-1} - \sigma_s), (\sigma_{s+\ell} - \sigma_{s+\ell+1})\}$$

$$\text{alors} \quad \sin\theta \leq \frac{1}{2p\delta - 1}$$

4.3 : Regroupements de modalités pour plusieurs questions

On généralise immédiatement les résultats précédents dans le cas où l'on effectue des regroupements dans un nombre quelconque p' de questions.

Pour les valeurs propres, on a :

$$\forall s \in]\, n \,] \qquad \sigma_s < \tau_s < \sigma_s + \frac{p'}{p}$$

L'angle θ entre les sous-espaces engendrés par les s premiers facteurs est tel que :

$$\text{Si} \quad \sigma_s - \sigma_{s+1} \geq \frac{p'}{p} \quad \text{alors} \quad \theta < \frac{\pi}{4} \quad \text{et} \quad \sin 2\theta \leq \frac{p'}{p(\sigma_s - \sigma_{s+1})}$$

Commentaires

Les majorations obtenues sont d'autant plus petites que les valeurs propres de l'espace factoriel considéré sont bien séparées des autres et que le nombre de questions modifiées est petit par rapport au nombre total de questions, Remarquons que, le nombre de modalités regroupées dans une question n'intervient pas dans les majorations, et que les dernières sont identiques à celles obtenues dans l'étude de l'ajout ou du retrait d'une question. Elles sont optimales, on peut donc en conclure

que le regroupement de modalités d'une question peut entraîner une perturbation de l'analyse aussi importante que sa suppression. Ces bornes permettent très rarement de conclure à la stabilité des facteurs ; ce phénomène n'est pas surprenant puisque nous avons déjà montré [5] que l'analyse des correspondances est peu perturbée, seulement si les éléments regroupés sont proches. Or des modalités s'excluant mutuellement ne sont jamais proches (cf. § 1.1). Les tableaux de données mis sous forme disjonctive complète sont donc sensibles aux regroupements de modalités de réponses. Cependant, il ne faut pas en conclure qu'il est nécessaire de multiplier le nombre de modalités de réponses, cette pratique risquant de faire apparaître des facteurs parasites. On remarque de plus que si, des regroupements sont effectués dans toutes les questions, tous les facteurs peuvent changer, dans certains cas ils peuvent même disparaître (voir par exemple [2] page 21).

5 - RAPPELS : COMPARAISON DES ELEMENTS PROPRES DE DEUX ENDOMORPHISMES SYMETRIQUES

5.1 : Définition : Soit $\mathcal{E}$ un espace euclidien de dimension n. Soient E et F deux sous-espaces de $\mathcal{E}$ ayant même dimension, on appelle angle entre E et F l'angle maximum entre un vecteur de l'un et sa projection orthogonale sur l'autre.

(Pour définir la position entre deux sous-espaces, voir [4]).

5.2 : Comparaison des éléments propres des endomorphismes symétriques A et C = A+B [5]

Soient A, B, C trois endomorphismes symétriques de $\mathcal{E}$ tels que C = A+B, dont les valeurs propres respectives sont notées α_s, β_s et γ_s et rangées par ordre décroissant. Alors :

5.2.1 : Comparaison des valeurs propres

$$- \forall s \in]n] \quad \sigma_s + \beta_n \leq \gamma_s \leq \alpha_s + \beta_1$$

$$- \quad \forall\, s \in]\,n]\qquad \sum_{i=1}^{s} \alpha_i + \sum_{i=n-s+1}^{n} \beta_i \lesssim \sum_{i=1}^{s} \gamma_i \lesssim \sum_{i=1}^{s} \alpha_i + \sum_{i=1}^{s} \beta_i$$

5.2.2 : Comparaison des sous-espaces invariants

- Soit θ l'angle entre les deux sous-espaces engendrés respectivement par les r premiers vecteurs propres de A et C.

$$\text{Si}\quad \beta_1 - \beta_n < \alpha_r - \alpha_{r+1}\quad ,\quad \text{alors}\quad \theta < \frac{\pi}{4}\quad \text{et}\quad \sin 2\theta < \frac{\beta_1 - \beta_n}{\alpha_r - \alpha_{r+1}}$$

cette borne est optimale.

- Soit θ l'angle entre les deux sous-espaces engendrés respectivement par les vecteurs propres de rang k,..., k+ℓ de A et C :

$$\text{Si}\quad \beta_1 - \beta_n < \delta = \inf\{(\alpha_{k-1} - \alpha_k), (\alpha_{k+r} - \alpha_{k+r+1})\}$$

$$\text{on a}\quad \sin\theta \lesssim \frac{\beta_1 - \beta_n}{2\delta - (\beta_1 - \beta_n)}$$

5.3 : Comparaison des éléments propres des endomorphismes symétriques A et C = A+B, avec B de rang un [6]

Soient A, B, C trois endomorphismes symétriques de $\mathcal{E}$ avec C = A+B et B de rang un et non négative.

Notons [Z] la direction propre de B associée à la valeur propre non nulle, Ψ_s (resp. Φ_s) l'angle entre [Z] et le sous-espace engendré par les s premiers vecteurs propres de C (resp. A). On a :

5.3.1 : Comparaison des valeurs propres

$$-\ (\forall\, s \in]\,n])\qquad \gamma_s - \beta \cos^2 \Psi_s \lesssim \alpha_s \lesssim \gamma_s$$

$$(\forall\, s \in]\,n])\qquad \sum_{i=1}^{s} \gamma_i - \beta \cos^2 \Psi_s \lesssim \sum_{i=1}^{s} \alpha_i$$

$$- (\forall_s : 1 < s \leq n) \quad \alpha_s \leq \gamma_s \leq \alpha_s + \beta \sin^2 \Phi_{s-1}$$

$$- (\forall_s \in]n]) \quad \sum_{i=1}^{s} \gamma_i \leq \sum_{i=1}^{s} \alpha_i + \beta \cos^2 \Phi_s$$

5.3.2 : Comparaison des sous-espaces invariants

- Soit θ l'angle entre les deux sous-espaces engendrés respectivement par les r premiers vecteurs propres de A et C.

$$\text{si} \quad \gamma_s - \gamma_{s+1} > \beta \quad \text{alors} \quad \theta < \frac{\pi}{4} \quad \text{et} \quad \operatorname{tg} 2\theta \leq \frac{\beta \sin 2\psi_s}{\gamma_s - \gamma_{s+1} - \beta \cos 2\psi_s}$$

$$\text{si} \quad \alpha_s - \alpha_{s+1} > \beta \quad \text{alors} \quad \theta < \frac{\pi}{4} \quad \text{et} \quad \operatorname{tg} 2\theta \leq \frac{\beta \sin 2\phi_s}{\alpha_s - \alpha_{s+1} + \beta \cos 2\phi_s}$$

Ces deux bornes sont optimales.

- Soit θ l'angle entre les sous-espaces invariants de C et A associés aux valeurs propres de rang $s,\ldots,s+\ell$. Soit Ψ (resp. Φ) l'angle entre $[z]$ et ce sous-espace invariant de C (resp. A), on a :

$$- \text{Si} \quad \delta = \inf \{(\gamma_{s-1} - \gamma_s), (\gamma_{s+\ell} - \gamma_{s+\ell+1})\} > \beta \cos^2 \Psi_{s+\ell}$$

$$\sin \theta \leq \frac{\beta \sin \psi \cos \psi}{\delta - \beta \cos^2 \psi_{s+\ell}}$$

$$- \text{Si} \quad \delta = \inf \{(\alpha_{s-1} - \alpha_s), (\alpha_{s+\ell} - \alpha_{s+\ell+1})\} > \beta \sin^2 \phi_{s-1}$$

$$\sin \theta \leq \frac{\beta \sin \Phi \cos \Phi}{\delta - \beta \sin^2 \Phi_{s-1}}$$

B I B L I O G R A P H I E

[1] BENZECRI, J.P. et collaborateurs, *L'Analyse des données, Tome 2,*
PARIS, Dunod, 1973.

[2] BENZECRI, J.P., *Les scrutins en 1967 à l'Assemblée des Nations Unies,*
ISUP, PARIS, 1974, Ronéo.

[3] BENZECRI, J.P., *Sur l'analyse des tableaux binaires associés à une correspondance multiple,*
ISUP, PARIS, 1970, Ronéo.

[4] DEMPSTER, A.P., *Elements of continuous multivariate analysis,*
NEW YORK, Addison Wesley, 1969.

[5] ESCOFIER, B. et LE ROUX, B., *Etude de trois problèmes de stabilité en analyse factorielle,*
PARIS, Publications de l'Institut de Statistiques, à paraître 1974.

[6] ESCOFIER, B. et LE ROUX, B., *Mesure de l'influence d'un descripteur sur les résultats d'une analyse en composantes principales,*
1974, Ronéo.