RAIRO. RECHERCHE OPÉRATIONNELLE

J. BENASSENI

Perturbation des poids des unités statistiques et approximation en analyse en composantes principales

RAIRO. Recherche opérationnelle, tome 21, n° 2 (1987), p. 175-198

<http://www.numdam.org/item?id=RO_1987__21_2_175_0>

R.A.I.R.O. Recherche opérationnelle/Operations Research
(vol. 21, n° 2, mai 1987, p. 175 à 198)

PERTURBATION DES POIDS DES UNITÉS STATISTIQUES ET APPROXIMATION EN ANALYSE EN COMPOSANTES PRINCIPALES (*)

par J. BENASSENI (1)

Résumé. – *En analyse en composantes principales, après que les poids de certaines unités statistiques aient été modifiés, on étudie la précision d'approximations des valeurs propres et des vecteurs propres récemment proposées. Les résultats sont illustrés à partir d'un exemple.*

Mots clés : Approximations; valeur propre; vecteur propre; influence; perturbation; analyse en composantes principales.

Abstract. – *When the weights of some of the observations are modified in principal component analysis, the accuracy of recent approximations for the eigenvalues and eigenvectors is discussed. An illustration of the results is given with an example.*

Keywords : Approximations; Eigenvalue; Eigenvector; Influence; Perturbation; Principal Component Analysis.

I. INTRODUCTION

Dans son interprétation des résultats d'une analyse en composantes principales (A.C.P.) le statisticien souhaite souvent disposer d'éléments lui permettant d'évaluer dans quelle mesure ces résultats peuvent être tributaires de telle ou telle unité statistique (u. s.) ou petit groupe d'u. s. Une approche traditionnelle en ce domaine consiste à étudier les variations des résultats découlant de la suppression de ces u. s. ou d'une variation des poids qui leur sont attribués.

Les problèmes liés à ce type de préoccupation sont initialement abordés par Escofier et Leroux [1976 (*a*) et (*b*)] et Escofier (1979) qui fournissent des bornes à la variation des valeurs propres et vecteurs propres lorsqu'on

(*) Reçu septembre 1986.
(1) Unité de Biométrie, 9, place Pierre-Viala, 34060 Montpellier Cedex.

R.A.I.R.O. Recherche opérationnelle/Operations Research, 0399-0559/87/02 175 24/$ 4.40

envisage certaines modifications des données en analyse factorielle des correspondances (suppression d'un élément, regroupement en classes) ou en A.C.P. (ajout ou suppression d'un groupe d'éléments, influence d'un descripteur). Dans le prolongement de ces derniers travaux et dans le cadre précis de l'influence des poids des u. s. qui nous intéresse ici des bornes à la variation des valeurs propres sont ensuite proposées par Benasseni [(1985) et (1986]. Parallèlement lorsque certaines u. s. sont supprimées Critchley (1985) utilise une approche différente qui par le biais de développements limités fournit des approximations des nouvelles caractéristiques spectrales avec l'aspect très intéressant de pouvoir ainsi cerner le mécanisme de variation de ces dernières. D'un point de vue pratique il reste cependant fondamental de pouvoir apprécier la qualité de ces approximations et en déduire une localisation précise des nouvelles caractéristiques spectrales sans avoir besoin de recalculer l'analyse. On montre ici comment l'utilisation de résultats classiques sur les quotients de Rayleigh permet de fournir des éléments de réponse à ce problème et d'améliorer en particulier très sensiblement les résultats antérieurs de Benasseni [(1985) et (1986] concernant la localisation des valeurs propres. L'accent est mis sur une formulation explicite du problème pour les approximations d'ordre zéro et un. Notre approche est faite dans le cadre d'une modification progressive des poids des u. s. ce qui permet d'englober le cas de leur suppression comme cas particulier et limite où leur poids est réduit à zéro ce qui généralise très légèrement l'approche de Critchley.

Pour fixer les notations on considère l'A.C.P. d'un ensemble de n u. s. caractérisées par les vecteurs colonnes x_l $l=1, \ldots, n$ de leur p coordonnées correspondant aux mesures faites pour chacune d'elles sur les p variables étudiées. A chaque u. s. l est attribué un poids initial $w_l \geqq 0$ avec $\sum_{l=1}^{n} w_l = 1$. On étudie tout d'abord la modification du poids d'une seule u. s. puis celle des poids d'un petit groupe d'u. s. Dans un but de concision on limite pour l'essentiel la présentation au cadre de l'A.C.P. sur matrice de variance. On notera cependant que l'intégralité des résultats se transpose au cadre de l'A.C.P. sur matrice de corrélation sans particularité notoire si ce n'est une complexité accrue des formulations analytiques. Tout au long de l'exposé les valeurs propres sont supposées distinctes et classées par ordre décroissant.

II. MODIFICATION DU POIDS D'UNE SEULE UNITÉ STATISTIQUE EN A.C.P. SUR MATRICE DE VARIANCE

On suppose ici que le poids initial w_j de l'u. s. j est transformé en $\tilde{w}_j$, les poids w_l, $l \neq j$ étant seulement normalisés en $\tilde{w}_l = \alpha_j w_l$ par le coefficient $\alpha_j = (1-\tilde{w}_j)/(1-w_j)$ de manière à avoir $\sum_{l=1}^{n} \tilde{w}_l = 1$.

1. Approximations des valeurs propres et des vecteurs propres

Posons $g=\sum_{l=1}^{n} w_l x_l$ et $\beta_j=1-\alpha_j$; il est facile de voir [Benasseni (1985)] que la transformation de la matrice de variance initiale V en $\tilde{V}$ peut être formulée comme suit :

$$\tilde{V}=\alpha_j[V+\beta_j(x_j-g)(x_j-g)^t]$$
$$=V+\beta_j[(x_j-g)(x_j-g)^t-V]-\beta_j^2(x_j-g)(x_j-g)^t. \quad \text{(II.1)}$$

La modification de V en $\tilde{V}$ transforme les valeurs propres initiales λ_i en $\tilde{\lambda}_i$ et les vecteurs propres u_i correspondants en $\tilde{u}_i$. En pratique $\beta_j=(\tilde{w}_j-w_j)/(1-w_j)$ reste petit. Les valeurs propres ayant été supposées distinctes, supposons, en suivant Wilkinson (1965), que les nouvelles valeurs propres et vecteurs propres s'expriment sous la forme :

$$\tilde{\lambda}_i=\lambda_i+\sum_{k=1}^{m}\beta_j^k\mu_k+O(\beta_j^{m+1})$$

et

$$\tilde{u}_i=u_i+\sum_{k=1}^{m}\beta_j^k v_k+O(\beta_j^{m+1})$$

m étant un entier arbitraire. Supposons u_i et $\tilde{u}_i$ normalisés, alors d'un point de vue théorique, en écrivant $\tilde{V}\tilde{u}_i=\tilde{\lambda}_i\tilde{u}_i$ et en identifiant dans cette égalité les termes correspondant aux différentes puissances de β_j il est possible, compte tenu des relations $\tilde{u}_i^t\tilde{u}_i=u_i^t u_i=1$, de déterminer les paramètres μ_k et v_k, $k=1,\ldots,m$. En pratique cependant la complexité des expressions augmente très rapidement avec l'ordre m que l'on considère pour le développement.

Il est commode de noter pour $m\geqq 1$:

$$\tilde{u}_i^{(m)}=u_i+\sum_{k=1}^{m}\beta_j^k v_k$$

et

$$\tilde{\lambda}_i^{(m)}=\lambda_i+\sum_{k=1}^{m}\beta_j^k\mu_k$$

les approximations d'ordre m de $\tilde{u}_i$ et $\tilde{\lambda}_i$ et de compléter la notation par $\tilde{u}_i^{(0)} = u_i$ et $\tilde{\lambda}_i^{(0)} = \lambda_i$. Dans la partie pratique de notre exposé nous discuterons principalement la précision des approximations d'ordre zéro et un. Si l'on décompose $x_j - g$ dans la base des vecteurs propres u_k de V avec $x_j - g = \sum_k z_{jk} u_k$ on obtient alors en suivant Critchley (1985) dans son utilisation du lemme de Sibson (1979) :

$$\tilde{\lambda}_i^{(1)} = \lambda_i + \beta_j (z_{ji}^2 - \lambda_i)$$

$$\tilde{u}_i^{(1)} = u_i - \beta_j z_{ji} \sum_{k \neq i} z_{jk} (\lambda_k - \lambda_i)^{-1} u_k.$$

Remarques : (1) Il peut arriver (notamment lorsque initialement λ_i est proche de λ_{i-1} ou λ_{i+1}) que le développement $\tilde{\lambda}_i = \tilde{\lambda}_i^{(m)} + O(\beta_j^m)$ de la i-ième valeur propre dans l'ordre décroissant ne corresponde pas à la i-ième valeur propre de $\tilde{V}$ (mais par exemple à celle de rang $i-1$ ou $i+1$ ce que l'on constate si $\tilde{\lambda}_i^{(m)} > \tilde{\lambda}_{i-1}^{(m)}$ ou $\tilde{\lambda}_{i+1}^{(m)} > \tilde{\lambda}_i^{(m)}$). En pratique ce phénomène ne pose pas de problème, il suffit de réordonner les valeurs propres perturbées $\tilde{\lambda}_i = \tilde{\lambda}_i^{(m)} + O(\beta_j^{m+1})$ et de renuméroter les vecteurs propres associés $\tilde{u}_i = \tilde{u}_i^{(m)} + O(\beta_j^{m+1})$. Dans la suite de l'exposé l'écriture $\tilde{\lambda}_i^{(m)}$ et $\tilde{u}_i^{(m)}$ désignera donc toujours les approximations de la i-ième valeur propre de $\tilde{V}$ et du vecteur propre normalisé qui lui est associé.

(2) On notera que lorsqu'on dispose d'une approximation normalisée $\tilde{u}_i^{(m)}$ de $\tilde{u}_i$, une approximation naturelle de $\tilde{\lambda}_i$ est donnée par le quotient de Rayleigh $q_i^{(m)} = (\tilde{u}_i^{(m)})^t \tilde{V} \tilde{u}_i^{(m)}$. Un des intérêts de cette nouvelle approximation réside principalement en ce qu'il est facile d'en apprécier la précision par les résultats simples qui sont présentés dans le paragraphe qui suit.

2. Précision des approximations et localisation des valeurs et vecteurs propres

Une méthodologie classique pour appréhender les erreurs d'approximation repose sur l'utilisation du quotient de Rayleigh $q_i^{(m)}$ défini dans la remarque précédente à partir du vecteur normalisé $\tilde{u}_i^{(m)}$. Suivons encore une fois Wilkinson (1965) (p. 172-173) en adaptant sa présentation dans le cas complexe pour une matrice normale, au cas de la matrice symétrique réelle $\tilde{V}$ qui nous intéresse ici.

Posons $\varepsilon_i^{(m)} = \| \tilde{V} \tilde{u}_i^{(m)} - q_i^{(m)} \tilde{u}_i^{(m)} \|$ et supposons que la décomposition de $\tilde{u}_i^{(m)}$ dans la base orthonormée des vecteurs propres $\tilde{u}_k$ $k = 1, \ldots, p$ de $\tilde{V}$ soit telle que $\tilde{u}_i^{(m)} = \sum_k c_k \tilde{u}_k$ avec $c_i > 0$. Soit encore a un réel positif tel que les $(p-1)$

valeurs propres $\tilde{\lambda}_k$, $k=1,\ldots,p$, $k\neq i$ de $\tilde{V}$ vérifient $|\tilde{\lambda}_k - q_i^{(m)}| > a$. Alors :

$$\|\tilde{u}_i - \tilde{u}_i^{(m)}\|^2 \leqq \{\varepsilon_i^{(m)}/a\}^2 \{1+(\varepsilon_i^{(m)}/a)^2\} \qquad \text{(II.2)}$$

et si $(\varepsilon_i^{(m)}/a)<1$:

$$|\tilde{\lambda}_i - q_i^{(m)}| \leqq \{(\varepsilon_i^{(m)})^2/a\} / \{1-(\varepsilon_i^{(m)}/a)^2\}. \qquad \text{(II.3)}$$

Remarques : (1) Les relations (II.2) et (II.3) donnent une localisation de $\tilde{u}_i$ et $\tilde{\lambda}_i$ dont la précision dépend de la qualité d'approximation de $\tilde{u}_i^{(m)}$ et $q_i^{(m)}$. En particulier l'inégalité (II.3) correspond à l'encadrement suivant :

$$d_i^{(m)} = q_i^{(m)} - \left\{\frac{(\varepsilon_i^{(m)})^2}{a}\right\} \Big/ \left\{1-\left(\frac{\varepsilon_i^{(m)}}{a}\right)^2\right\} \leqq \tilde{\lambda}_i$$

$$\leqq q_i^{(m)} + \left\{\frac{(\varepsilon_i^{(m)})^2}{a}\right\} \Big/ \left\{1-\left(\frac{\varepsilon_i^{(m)}}{a}\right)^2\right\} = D_i^{(m)} \qquad \text{(II.4)}$$

(2) Puisque l'on a supposé $\|\tilde{u}_i\| = \|\tilde{u}_i^{(m)}\| = 1$, il est facile de formuler la relation (II.2) en terme de cosinus avec :

$$1 - \{(\varepsilon_i^{(m)})^2/2a^2\} \{1+(\varepsilon_i^{(m)}/a)^2\} \leqq \cos(\tilde{u}_i^{(m)}, \tilde{u}_i) \leqq 1. \qquad \text{(II.5)}$$

(3) L'utilisation pratique des inégalités (II.2) et (II.3) nécessite la détermination d'une valeur à donner au coefficient a. Cette détermination suppose que l'on puisse disposer d'une première localisation, même un peu grossière, des valeurs propres $\tilde{\lambda}_k$ afin de pouvoir les situer par rapport au quotient de Rayleigh $q_i^{(m)}$. Cette localisation fait l'objet du paragraphe qui suit.

3. Une première localisation des valeurs propres $\tilde{\lambda}_k$ de $\tilde{V}$

3.1. *Approche utilisant les valeurs propres initiales λ_k de V*

Benasseni (1985 et 1986) fournit, en utilisant les inégalités de Weyl, des encadrements des nouvelles valeurs propres $\tilde{\lambda}_k$ à partir des valeurs propres initiales λ_k. Rappelons brièvement la formulation des principaux encadrements dans le cas $\beta_j<0$ par exemple.

Pour $k=1,\ldots,p$ les bornes supérieures des encadrements des $\tilde{\lambda}_k$ sont données par l'inégalité :

$$\tilde{\lambda}_k \leqq \alpha_j \lambda_k. \qquad \text{(II.6)}$$

En ce qui concerne les bornes inférieures le choix doit être fait parmi les inégalités suivantes :

– pour $k=1,\ldots,p$:

$$\alpha_j(\lambda_k+\beta_j\|x_j-g\|^2)\leqq\tilde{\lambda}_k \tag{II.7}$$

– pour $k=1,\ldots,p-1$:

$$\alpha_j\lambda_{k+1}\leqq\tilde{\lambda}_k \tag{II.8}$$

– pour $k=1,\ldots,p$ si V inversible :

$$\alpha_j(1+\beta_j\|x_j-g\|^2_{V^{-1}})\lambda_k\leqq\tilde{\lambda}_k \tag{II.9}$$

avec

$$\|x_j-g\|^2_{V^{-1}}=(x_j-g)^t V^{-1}(x_j-g)$$

– pour $k=1,\ldots,p$:

$$\alpha_j[\lambda_k+\{\beta_j/2\}\{z_{jk}^2+\sqrt{z_{jk}^4+4z_{jk}^2(\|x_j-g\|^2-z_{jk}^2)}\}]\leqq\tilde{\lambda}_k \tag{II.10}$$

si lorsque $k\neq 1$ la condition : $\lambda_{k-1}-\lambda_k>-\beta_j\{\|x_j-g\|^2-z_{jk}^2\}$ est vérifiée.

Il n'est pas possible d'un point de vue théorique de déterminer la plus précise des inégalités (II.7), (II.8) ou (II.9), par contre lorsque sa condition d'utilisation est vérifiée la relation (II.10) est toujours plus précise que (II.7). En fait cette condition n'apparaît restrictive dans la pratique que pour les plus petites valeurs propres qui sont souvent proches les unes des autres.

3.2. *Une deuxième approche*

L'approche précédente qui localise $\tilde{\lambda}_k$ à partir de λ_k, valeur propre de même rang, peut ne pas être toujours pleinement satisfaisante notamment dans des situations d'inversion des valeurs propres. On peut alors procéder comme suit. Un calcul de quelques lignes permet, en écrivant $V=\sum\lambda_k u_k u_k^t$ et toujours avec $x_j-g=\sum z_{jk}u_k$, de réexprimer la relation (II.1)

$$\tilde{V}=\alpha_j\{V+\beta_j(x_j-g)(x_j-g)^t\}$$

sous la forme

$$\tilde{V}=\alpha_j\{\tilde{\tilde{V}}+\beta_j(vw^t+wv^t+ww^t)\} \tag{II.11}$$

avec :

$$\tilde{\tilde{V}}=\sum_{k\neq r,\,s}\lambda_k u_k u_k^t+(\lambda_r+\beta_j z_{jr}^2)u_r u_r^t+\beta_j z_{jr}z_{js}(u_r u_s^t+u_s u_r^t)+(\lambda_s+\beta_j z_{js}^2)u_s u_s^t$$

$$v = z_{jr} u_r + z_{js} u_s \qquad \text{et} \qquad w = \sum_{k \neq r,\, s} z_{jk} u_k = (x_j - g) - v$$

où r et s sont deux entiers distincts et arbitraires compris entre 1 et p dont le choix sera discuté un peu plus loin.

Posons

$$\gamma_r = \lambda_r + \beta_j z_{jr}^2, \qquad \gamma_s = \lambda_s + \beta_j z_{js}^2, \qquad \gamma_{rs} = \beta_j z_{jr} z_{js}.$$

En utilisant les résultats algébriques élémentaires présentés par Benasseni (1986), on remarque que les valeurs propres de $\tilde{\tilde{V}}$ sont constituées par les $(p-2)$ valeurs propres λ_k, $k \neq r$, s et par les deux valeurs propres

$$(1/2)\{\gamma_r + \gamma_s \pm \sqrt{(\gamma_r + \gamma_s)^2 - 4(\gamma_r \gamma_s - \gamma_{rs}^2)}\}.$$

Ces valeurs propres classées par ordre décroissant seront notées $\mu_1 \geqq \mu_2 \geqq \ldots \geqq \mu_p$. Désignons par $\nu_1 \geqq \nu_2 \geqq \ldots \geqq \nu_p$ les valeurs propres de $A = vw^t + w^t v + ww^t$, il est alors facile de voir en se référant une nouvelle fois aux techniques utilisées dans Benasseni (1986) que :

$$\nu_1 = (1/2)[\|x_j - g\|^2 - (z_{jr}^2 + z_{js}^2) + \sqrt{\{\|x_j - g\|^2 - (z_{jr}^2 + z_{js}^2)\}^2 + 4\{\|x_j - g\|^2 - (z_{jr}^2 + z_{js}^2)\}\{z_{jr}^2 + z_{js}^2\}}] \geqq 0$$

$$\nu_i = 0, \qquad i = 2, \ldots, p-1$$

$$\nu_p = (1/2)[\|x_j - g\|^2 - (z_{jr}^2 + z_{js}^2) - \sqrt{\{\|x_j - g\|^2 - (z_{jr}^2 + z_{js}^2)\}^2 + 4\{\|x_j - g\|^2 - (z_{jr}^2 + z_{js}^2)\}\{z_{jr}^2 + z_{js}^2\}}] \leqq 0.$$

En appliquant les inégalités de Weyl à la relation (II.11) on obtient en définitive les encadrements suivants pour $k = 1, \ldots, p$:

si $\beta_j > 0$

$$\alpha_j(\mu_k + \beta_j \nu_p) \leqq \tilde{\lambda}_k \leqq \alpha_j(\mu_k + \beta_j \nu_1) \qquad \text{(II.12)}$$

si $\beta_j < 0$

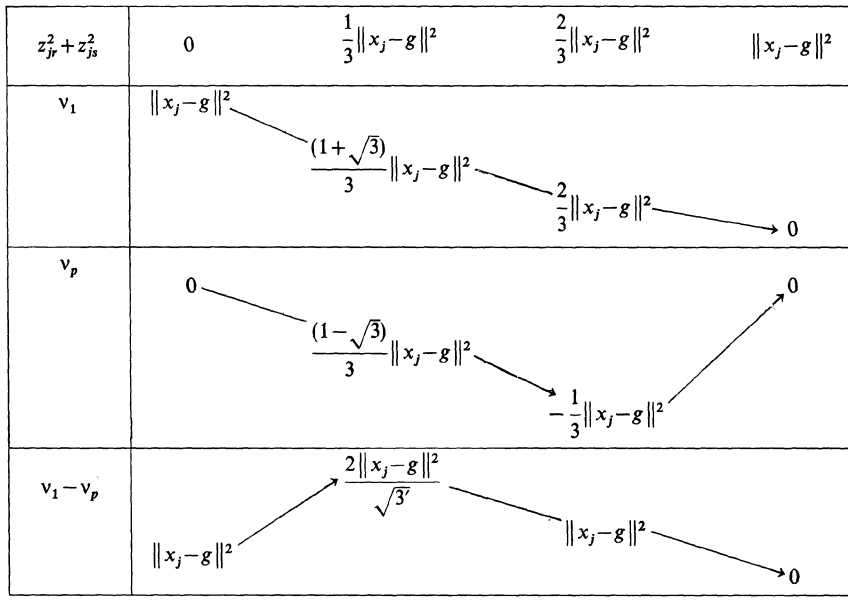
$$\alpha_j(\mu_k + \beta_j \nu_1) \leqq \tilde{\lambda}_k \leqq \alpha_j(\mu_k + \beta_j \nu_p) \qquad \text{(II.13)}$$

et dans les deux cas si $p \geqq 3$:

$$\alpha_j \mu_{k+1} \leqq \tilde{\lambda}_k \qquad \text{si} \quad k = 1, \ldots, p-1$$

$$\alpha_j \mu_{k-1} \geqq \tilde{\lambda}_k \qquad \text{si} \quad k = 2, \ldots, p.$$

La précision de ces nouvelles bornes par rapport au choix de z_{jr} et z_{js} est liée aux variations (résumées dans le tableau qui suit) de ν_1, ν_p et $\nu_1 - \nu_p$ en fonction de z_{jr}^2 et z_{js}^2.



On constate que la précision des encadrements devient parfaite lorsque $z_{jr}^2+z_{js}^2=\|x_j-g\|^2$ i.e. lorsque l'u. s. j se trouve dans le plan factoriel engendré par u_r et u_s ($\tilde{V}=\alpha_j \tilde{\tilde{V}}$). En pratique les encadrements apparaissent intéressants lorsqu'on se rapproche de cette situation extrême i. e. lorsque $z_{jr}^2+z_{js}^2>(2/3)\|x_j-g\|^2$ et l'on est donc souvent conduit à choisir r et s tels que $z_{jr}^2+z_{js}^2$ soit le plus grand possible. Cependant, on est parfois conduit à d'autres choix lorsque l'on s'intéresse plus particulièrement soit à la borne inférieure, soit à la borne supérieure à donner à une valeur propre.

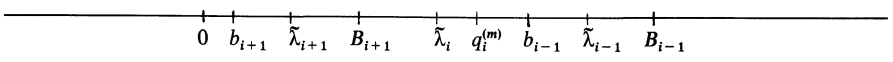
4. Localisation du quotient de Rayleigh par rapport aux valeurs propres de $\tilde{V}$

D'une manière générale les résultats du paragraphe 3 conduisent à des encadrements exacts des valeurs propres $\tilde{\lambda}_k$ de $\tilde{V}$. Désignons par b_k et B_k les bornes correspondant à l'encadrement le plus précis que fournit l'ensemble

des inégalités pour la valeur propre $\tilde{\lambda}_k$ $k=1, \ldots, p$ avec $b_k \leqq \tilde{\lambda}_k \leqq B_k$. Deux situations peuvent se présenter dans la localisation du quotient de Rayleigh $q_i^{(m)}$.

4.1. La situation la plus courante est caractérisée par l'inégalité $B_{i+1} < b_{i-1}$. Deux possibilités sont alors à envisager :

$$(a)\ B_{i+1} < q_i^{(m)} < b_{i-1}$$



On a alors $|\tilde{\lambda}_k - q_i^{(m)}| > a$ $k=1, \ldots, p$ $k \neq i$ en donnant au coefficient a introduit dans les relations (II.2) et (II.3) la valeur $a = \min(b_{i-1} - q_i^{(m)},\ q_i^{(m)} - B_{i+1})$.

Les relations (II.3) et (II.4) sont utilisables sous réserve que la condition $\varepsilon_i^{(m)}/a < 1$ soit vérifiée. On obtient en particulier par (II.4) un encadrement $d_i^{(m)} \leqq \tilde{\lambda}_i \leqq D_i^{(m)}$ pour $\tilde{\lambda}_i$. On peut généralement obtenir par la même approche des encadrements $d_k^{(m)} \leqq \tilde{\lambda}_k \leqq D_k^{(m)}$ $k=i-1,\ i+1$ pour $\tilde{\lambda}_{i-1}$ et $\tilde{\lambda}_{i+1}$. On a presque toujours en pratique $d_{i-1}^{(m)} > b_{i-1}$ et $B_{i+1} > D_{i+1}^{(m)}$ on peut alors améliorer la précision de $d_i^{(m)}$ et $D_i^{(m)}$ en prenant $b_{i-1} = d_{i-1}^{(m)}$ et $B_{i+1} = D_{i+1}^{(m)}$ dans la définition de a. On peut d'une manière analogue améliorer la précision de $d_{i-1}^{(m)}$ et $D_{i+1}^{(m)}$ ce qui entraîne une nouvelle amélioration pour $d_i^{(m)}$ et $D_i^{(m)}$ et en continuant par itération, l'amélioration des bornes liées à une valeur propre entraîne une amélioration de proche en proche des bornes liées aux valeurs propres voisines. On notera cependant que l'amélioration devient généralement insignifiante d'un point de vue pratique au bout d'une ou deux itérations.

Enfin on notera que lorsque $i=1$, il suffit d'avoir $q_1^{(m)} > B_2$ pour définir alors a par $a = q_1^{(m)} - B_2$. La remarque se transpose au cas $i=p$.

$$(b)\ q_i^{(m)} \leqq B_{i+1} \qquad \text{ou} \qquad q_i^{(m)} \geqq b_{i-1}$$

Ce cas correspond généralement à une situation où $\tilde{u}_i^{(m)}$ est une mauvaise approximation de $\tilde{u}_i$. Il en est alors de même de $q_i^{(m)}$ pour $\tilde{\lambda}_i$. Étant assez éloigné de $\tilde{\lambda}_i$, $q_i^{(m)}$ peut alors vérifier $q_i^{(m)} \leqq B_{i+1}$ ou $q_i^{(m)} \geqq b_{i-1}$. Il convient simplement d'affiner la précision de $\tilde{u}_i^{(m)}$ en utilisant un développement limité d'ordre supérieur de manière à se retrouver dans la situation précédente.

4.2. Il arrive que l'on ait $B_{i+1} \geqq b_{i-1}$ dans quelques cas où certaines valeurs propres sont initialement très rapprochées les unes des autres. Il n'est plus possible alors de déterminer a ni donc d'appliquer les résultats du paragraphe II.2.

Cette situation est en général indicatrice en elle-même d'une grande instabilité de ces valeurs propres et des vecteurs propres associés [*cf.* les termes en $(\lambda_k-\lambda_i)^{-1}$ dans l'expression de $\tilde{u}_i^{(1)}$] et doit généralement inciter le statisticien à recalculer l'analyse.

5. Approximations d'ordre zéro et un

L'approximation d'ordre zéro est caractérisée par $\tilde{u}_i^{(0)}=u_i$ et $\tilde{\lambda}_i^{(0)}=\lambda_i$. On suppose ici que l'on est dans la situation $B_{i+1}<q_i^{(0)}<b_{i-1}$ qui permet de déterminer simplement le coefficient a intervenant dans les expressions (II.2) et (II.3) en posant d'après le paragraphe 4,

$$a=\min(b_{i-1}-q_i^{(0)},\ q_i^{(0)}-B_{i+1}) \qquad \text{si} \quad i\neq 1 \quad \text{et} \quad i\neq n,$$

$$a=q_1^{(0)}-B_2 \quad \text{si} \quad i=1 \qquad \text{ou} \qquad a=b_{p-1}-q_p^{(0)} \quad \text{si} \quad i=p.$$

Un calcul immédiat montre que $q_i^{(0)}$ et $\varepsilon_i^{(0)}$ s'expriment sous la forme

$$q_i^{(0)}=\alpha_j(\lambda_i+\beta_j z_{ji}^2)=\lambda_i+\beta_j(z_{ji}^2-\lambda_i)-\beta_j^2 z_{ji}^2=\tilde{\lambda}_i^{(1)}-\beta_j^2 z_{ji}^2$$

$$(\varepsilon_i^{(0)})^2=\alpha_j^2\beta_j^2 z_{ji}^2(\|x_j-g\|^2-z_{ji}^2).$$

L'inégalité (II.3) est subordonnée dans son utilisation à la condition $\varepsilon_i^{(0)}/a<1$.

Posons $\omega=a/\alpha_j\beta_j$. Il est facile de voir que cette condition correspond à la région hachurée comprise entre la courbe d'équation :

$$\|x_j-g\|^2=(\omega/z_{ji}^2)+z_{ji}^2$$

et la première bissectrice d'équation

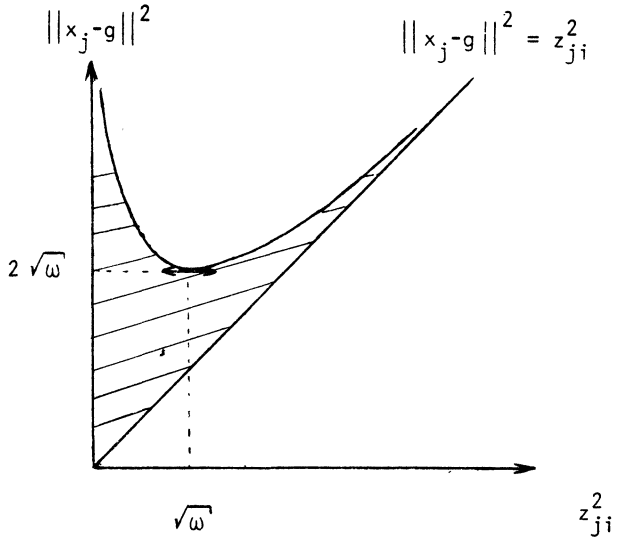
$$\|x_j-g\|^2=z_{ji}^2.$$

On constate ainsi

– que la région d'application des inégalités est d'autant plus grande que a l'est c'est-à-dire que $q_i^{(0)}$ est éloigné de $\tilde{\lambda}_{i-1}$ et $\tilde{\lambda}_{i+1}$;

– que lorsque z_{ji} devient grand la condition n'est plus vérifiée que si x_j-g est pratiquement colinéaire à u_i. Lorsque z_{ji} devient au contraire petit la condition est pratiquement toujours vérifiée;

– que dans les cas extrêmes où (x_j-g) est colinéaire ou orthogonal à u_i ($x_j-g=z_{ji}u_i$ ou $z_{ji}=0$) la condition est toujours vérifiée puisque $\varepsilon_i^{(0)}=0$. Dans ces cas on a $\tilde{u}_i=u_i$ et $\tilde{\lambda}_i=\alpha_j(\lambda_i+\beta_j z_{ji}^2)$ si $x_j-g=z_{ji}u_i$ ou $\tilde{\lambda}_i=\alpha_j\lambda_i$ si $z_{ji}=0$.



Les inégalités (II.2) :

$$\|\tilde{u}_i - u_i\|^2 \leqq (\varepsilon_i^{(0)}/a)^2 \{1 + (\varepsilon_i^{(0)}/a)^2\}$$

et (II.3) [lorsque $(\varepsilon_i^{(0)}/a) < 1$] :

$$|\tilde{\lambda}_i - q_i^{(0)}| \leqq \{(\varepsilon_i^0)^2/a\} / \{1 - (\varepsilon_i^{(0)}/a)^2\}$$

(avec $\tilde{u}_i^{(0)} = u_i$) permettent d'apprécier la qualité des approximations de $\tilde{u}_i$ par u_i et de $\tilde{\lambda}_i$ par $q_i^{(0)}$. Comme on a :

$$q_i^{(0)} = \alpha_j(\lambda_i + \beta_j z_{ji}^2) = \lambda_i + \beta_j(z_{ji}^2 - \lambda_i) - \beta_j^2 z_{ji}^2 = \tilde{\lambda}_i^{(1)} - \beta_j^2 z_{ji}^2$$

on peut déduire de (II.3) de manière indirecte des renseignements sur la qualité d'approximation de $\tilde{\lambda}_i$ par $\tilde{\lambda}_i^{(0)} = \lambda_i$ ou par $\lambda_i^{(1)}$. Ainsi on a par exemple :

$$-\beta_j^2 z_{ji}^2 - [\{(\varepsilon_i^{(0)})^2/a\} / \{1 - (\varepsilon_i^{(0)}/a)^2\}] \leqq \tilde{\lambda}_i - \tilde{\lambda}_i^{(1)} \leqq [\{(\varepsilon_i^{(0)})^2/a\} / \{1 - (\varepsilon_i^{(0)}/a)^2\}] - \beta_j^2 z_{ji}^2$$

On remarquera en particulier que le quotient de Rayleigh

$$q_i^{(0)} = (\tilde{u}_i^{(0)})^t \tilde{V} \tilde{u}_i^{(0)} = u_i^t \tilde{V} u_i$$

(associé donc à l'approximation d'ordre zéro $\tilde{u}_i^{(0)} = u_i$ pour les vecteurs propres) ne diffère de l'approximation d'ordre 1 $\tilde{\lambda}_i^{(1)}$ que par un terme en β_j^2.

Un calcul simple montre par ailleurs que le quotient de Rayleigh d'ordre un $q_i^{(1)} = (\tilde{u}_i^{(1)})^t \tilde{V} (\tilde{u}_i^{(1)})$ (avec $\tilde{u}_i^{(1)}$ normalisé) peut se formuler comme suit :

$$q_i^{(1)} = [1 + \beta_j^2 z_{ji}^2 \sum_{k \neq i} z_{jk}^2 (\lambda_k - \lambda_i)^{-2}]^{-1}$$

$$\times [q_i^{(0)} + \beta_j^2 z_{ji}^2 \sum_{k \neq i} z_{jk}^2 (\lambda_k - 2)(\lambda_k - \lambda_i)^{-1} + \beta_j^3 z_{ji}^2 r + \beta_j^4 z_{ji}^2 s]$$

r et s étant deux réels dont les expressions découlent des calculs.

On voit donc que $q_i^{(1)}$ ne diffère de $q_i^{(0)}$ que par des termes d'ordre supérieur ou égal à β_j^2 mais que ces termes peuvent ne pas être négligeables si l'une des valeurs propres λ_k, $k \neq i$, se trouve suffisamment proche de λ_i.

6. Stabilité de la représentation des unités statistiques dans les plans factoriels

D'un point de vue pratique le statisticien utilise la représentation des u. s. dans les plans factoriels pour disposer de manière globale d'une vision approximative des positions relatives des u. s. Il est intéressant pour lui de pouvoir apprécier les distorsions maximales dont pourront être affectées ces représentations compte tenu de la modification de poids envisagée. Posons

$$\tilde{g} = \sum_{i=1}^{n} \tilde{w}_i x_i \qquad \text{et} \qquad x_k - \tilde{g} = \sum_{i=1}^{p} \tilde{z}_{ki} \tilde{u}_i,$$

$$x_k - g = \sum_{i=1}^{p} z_{ki} u_i \qquad (k = 1, \ldots, n).$$

Dans le plan factoriel engendré par les axes r et s la position relative de deux u. s. k et l dépend respectivement avant et après perturbation des quantités $|z_{ki} - z_{li}|$ et $|\tilde{z}_{ki} - \tilde{z}_{li}|\, i = r, s$. Aussi est-il intéressant de pouvoir les comparer. Pour ce faire on peut écrire :

$$|\tilde{z}_{ki} - \tilde{z}_{li}| = |\tilde{u}_i^t (x_k - x_l)|$$

$$\leqq |\{\tilde{u}_i^t - (\tilde{u}_i^{(m)})^t\}\{x_k - x_l\}| + |(\tilde{u}_i^{(m)})^t (x_k - x_l)|$$

$$\leqq \|\tilde{u}_i - \tilde{u}_i^{(m)}\| \, \|x_k - x_l\| + |\tilde{z}_{ki}^{(m)} - \tilde{z}_{li}^{(m)}|$$

avec $\tilde{z}_{ki}^{(m)} = (\tilde{u}_i^{(m)})^t (x_k - \tilde{g})$, $\tilde{z}_{li}^{(m)} = (\tilde{u}_i^{(m)})^t (x_l - \tilde{g})$ et en utilisant l'inégalité de Cauchy-Schwarz.

D'une manière symétrique on obtient

$$|\tilde{z}_{ki}^{(m)} - \tilde{z}_{li}^{(m)}| \leqq \|\tilde{u}_i - \tilde{u}_i^{(m)}\| \, \|x_k - x_l\| + |\tilde{z}_{ki} - \tilde{z}_{li}|$$

d'où l'on déduit l'encadrement

$$\left|\tilde{z}_{ki}^{(m)} - \tilde{z}_{li}^{(m)}\right| - \left\|\tilde{u}_i - \tilde{u}_i^{(m)}\right\| \left\|x_k - x_l\right\| \leqq \left|\tilde{z}_{ki} - \tilde{z}_{li}\right| \leqq \left|\tilde{z}_{ki}^{(m)} - \tilde{z}_{li}^{(m)}\right| + \left\|\tilde{u}_i - \tilde{u}_i^{(m)}\right\| \left\|x_k - x_l\right\|$$

dans lequel on utilise la majoration de $\left\|\tilde{u}_i - \tilde{u}_i^{(m)}\right\|$ fournie par (II.2).

III. EXTENSION DES RÉSULTATS

1. Modification des poids de plusieurs unités statistiques dans l'ACP sur matrice de variance

Supposons sans perte de généralité que les poids w_j des r u. s. x_j $j=1, \ldots, r$ (r petit devant n) soient modifiés en $\tilde{w}_j$ $\left(0 \leqq \tilde{w}_j < 1,\ 0 \leqq \sum_{j=1}^{r} \tilde{w}_j < 1\right)$, les poids w_k $k=r+1, \ldots, n$ des $(n-r)$ u. s. restantes étant seulement normalisés par le coefficient α en $\tilde{w}_k = \alpha\, w_k$ avec

$$\alpha = \left(1 - \sum_{j=1}^{r} \tilde{w}_j\right) \Big/ \left(1 - \sum_{j=1}^{r} w_j\right)$$

de manière à avoir $\sum_{k=1}^{n} \tilde{w}_k = 1$. La matrice de variance initiale V se trouve transformée en $\tilde{V}$ qu'un calcul simple permet d'exprimer sous la forme :

$$\tilde{V} = \alpha V + \sum_{j=1}^{r} (\tilde{w}_j - \alpha w_j)(x_j - g)(x_j - g)^t - \sum_{j=1}^{r} \sum_{k=1}^{r} (\tilde{w}_j - \alpha w_j)(\tilde{w}_k - \alpha w_k)(x_j - g)(x_k - g)^t \quad \text{(III.1)}$$

En pratique il arrive souvent que les u. s. x_j $j=1, \ldots, r$ aient initialement une influence homogène ($w_j = w$ $j=1, \ldots, r$) et que l'on désire que la modification des poids conserve cette homogénéité ($\tilde{w}_j = \tilde{w}$, $j=1, \ldots, r$). Alors $\alpha = (1 - r\tilde{w})/(1 - rw)$ et avec $\beta = (1-\alpha)/r$ on a $\tilde{w}_k - \alpha w_k = \beta$ $k=1, \ldots, r$. On peut ainsi écrire :

$$\tilde{V} = \alpha V + \beta \sum_{j=1}^{r} (x_j - g)(x_j - g)^t - \beta^2 \sum_{j=1}^{r} \sum_{k=1}^{r} (x_j - g)(x_k - g)^t$$

$$\tilde{V}=V+\beta\left\{\sum_{j=1}^{r}(x_j-g)(x_j-g)^t-rV\right\}-\beta^2\sum_{j=1}^{r}\sum_{k=1}^{r}(x_j-g)(x_k-g)^t \quad \text{(III.2)}$$

$\beta=(\tilde{w}-w)/(1-rw)$ étant en général petit on peut exprimer comme précédemment $\tilde{u}_i$ et $\tilde{\lambda}_i$ sous la forme :

$$\tilde{u}_i=\tilde{u}_i^{(m)}+O(\beta^{m+1}), \qquad \tilde{\lambda}_i=\tilde{\lambda}_i^{(m)}+O(\beta^{m+1})$$

avec

$$\tilde{u}_i^{(0)}=u_i, \qquad \tilde{\lambda}_i^{(0)}=\lambda_i \qquad \text{et} \qquad \tilde{u}_i^{(m)}=u_i+\sum_{k=1}^{m}\beta^k v_k,$$

$$\tilde{\lambda}_i^{(m)}=\lambda_i+\sum_{k=1}^{m}\beta^k\mu_k \qquad (m\geqq 1).$$

A l'ordre 1 les approximations se formulent simplement comme suit :

$$\tilde{\lambda}_i^{(1)}=\lambda_i+\beta\left\{\sum_{j=1}^{r}z_{ji}^2-r\lambda_i\right\} \qquad \text{(III.3)}$$

$$\tilde{u}_i^{(1)}=u_i-\beta\left\{\sum_{j=1}^{r}z_{ji}\sum_{k\neq i}(\lambda_k-\lambda_i)^{-1}z_{jk}u_k\right\}. \qquad \text{(III.4)}$$

Au coefficient β près ces relations mettent en relief le caractère additif de la formulation des approximations d'ordre 1, dans le cas de la modification des poids de plusieurs u. s., par rapport au cas d'une modification du poids d'une seule u. s.

En introduisant le quotient de Rayleigh

$$q_i^{(m)}=(\tilde{u}_i^{(m)})^t\,\tilde{V}\,\tilde{u}_i^{(m)}$$

on peut utiliser comme précédemment les relations (II.2) et (II.3) sans problème particulier si ce n'est celui de la détermination du coefficient a que l'on peut résoudre en considérant la perturbation des r poids comme le résultat de r perturbations consécutives du poids d'une seule u. s. à chaque fois. Pour éviter une formulation trop lourde explicitons simplement cette détermination dans le cas le plus courant (qui sera illustré à partir d'un exemple au paragraphe suivant) où un poids $w_i=1/n$ étant initialement attribué à chaque u. s. on supprime les r u. s. $x_1, \ldots, x_r$. La situation est caractérisée par $w_i=1/n$ $i=1, \ldots, n$; $\tilde{w}_i=0$ $i=1, \ldots, r$, $\tilde{w}_i=1/(n-r)$ $i=r+1, \ldots, n$.

On peut considérer que l'on supprime les u. s. les unes après les autres en r étapes successives. Soit $V_{(l)}(1 \leqq l \leqslant r)$ la matrice de variance obtenue à l'étape l après suppression des u. s. $x_1, \ldots, x_l$. Supposons que l'on ait obtenu des encadrements $b_{k\,(l)} \leqq \lambda_{k\,(l)} \leqq B_{k\,(l)}$ pour ses valeurs propres $\lambda_{k\,(l)} k = 1, \ldots, p$. Nous allons voir que l'on peut en déduire des encadrements $b_{k\,(l+1)} \leqq \lambda_{k\,(l+1)} \leqq B_{k\,(l+1)}$ pour les valeurs propres $\lambda_{k\,(l+1)}$ de la matrice de variance $V_{(l+1)}$. Alors il est clair qu'une procédure en r étapes permettra à partir de $b_{k\,(0)} = \mathrm{B}_{k\,(0)} = \lambda_k$ (valeurs propres initiales de V) d'arriver à des bornes inférieures $b_{k\,(r)} = b_k$ et supérieures $B_{k\,(r)} = B_k$ qui permettront une première localisation des valeurs propres $\lambda_{k\,(r)} = \tilde{\lambda}_k$ de $V_{(r)} = \tilde{V}$. Cette localisation permet en général de définir le coefficient a en suivant la démarche du paragraphe II. 4.

A partir de (II. 1) on peut exprimer $V_{(l+1)}$ en fonction de $V_{(l)}$. On a alors :

$$V_{(l+1)} = \{(n-l)/(n-l-1)\}\{V_{(l)} - (1/(n-l-1))(x_{l+1} - g_l)(x_{l+1} - g_l)^t\} \quad \text{(III. 5)}$$

avec

$$g_l = \{1/(n-l)\} \sum_{i=l+1}^{n} x_i = g + \{1/(n-l)\}\left\{lg - \sum_{i=1}^{l} x_i\right\}.$$

Les résultats du paragraphe II. 3. 1 permettent d'obtenir à partir de (III. 5) une première localisation des valeurs propres $\lambda_{k\,(l+1)}$ de $V_{(l+1)}$ à partir des valeurs propres $\lambda_{k\,(l)}$ de V_l. Ainsi en utilisant par exemple (II. 6) et (II. 7) pour fixer les idées on obtient :

$$\{(n-l)/(n-l-1)\}\{\lambda_{k\,(l)} - [1/(n-l-1)]\,\|x_{l+1} - g_l\|^2\} \leqq \lambda_{k\,(l+1)}$$

$$\lambda_{k\,(l+1)} \leqq \{(n-l)/(n-l-1)\}\,\lambda_{k\,(l)}$$

d'où l'on déduit puisque l'on a supposé :

$$b_{k\,(l)} \leqq \lambda_{k\,(l)} \leqq \mathrm{B}_{k\,(l)}$$

$$b'_{k\,(l+1)} = \{(n-l)/(n-l-1)\}\{b_{k\,(l)} - [1/(n-l-1)]\,\|x_{l+1} - g_l\|^2\} \leqq \lambda_{k\,(l+1)}$$

$$\lambda_{k\,(l+1)} \leqq \{(n-l)/(n-l-1)\}\,B_{kl} = B'_{k\,(l+1)}$$

Les bornes $b'_{k\,(l+1)}$ et $B'_{k\,(l+1)}$ $k = 1, \ldots, p$ peuvent généralement être affinées sensiblement par l'utilisation de l'inégalité (II. 4). Intéressons nous à la i-ième valeur propre de $V_{(l+1)}$ par exemple. A partir de l'expression de $V_{(l+1)}$ en fonction de V :

$$V_{(l+1)} = V - \{1/(n-l-1)\}\left\{\sum_{j=1}^{l+1} (x_j - g)(x_j - g)^t - (l+1)\,V\right\}$$

$$-\{1/(n-l-1)^2\}\sum_{j=1}^{l+1}\sum_{k=1}^{l+1}(x_j-g)(x_k-g)^t$$

telle qu'elle résulte de (III.2), on obtient à l'ordre m l'approximation $u_{i\,(l+1)}^{(m)}$ du i-ième vecteur propre de $V_{l+1)}$ (à l'ordre 1

$$u_{i\,(l+1)}^{(1)}=u_i+\{1/(n-l-1)\}\left\{\sum_{j=1}^{l+1}z_{zi}\sum_{k\neq i}(\lambda_k-\lambda_i)^{-1}z_{jk}u_k\right\}\Bigg).$$

Soient

$$q_{i\,(l+1)}^{(m)}=(u_{i\,(l+1)}^{(m)})^t V_{(l+1)}(u_{i\,(l+1)}^{(m)})$$

et

$$\varepsilon_{i\,(l+1)}^{(m)}=\| V_{(l+1)}u_{i\,(l+1)}^{(m)}-q_{i\,(l+1)}^{(m)}u_{i\,(l+1)}^{(m)}\|.$$

Alors si

$$a_{l+1}=\min\,(q_{i\,(l+1)}^{(m)}-B'_{i+1\,(l+1)},\qquad b'_{i-1\,(l+1)}-q_{i\,(l+1)}^{(m)})>0$$

et si $\varepsilon_{i\,(l+1)}^{(m)}/a_{l+1}<1$ on peut utiliser les relations (II.3) et (II.4) pour localiser $\lambda_{i\,(l+1)}$ par rapport à $q_{i\,(l+1)}^{(m)}$. En particulier l'encadrement (II.4) fournit les bornes $b_{i\,(l+1)}$ et $B_{i\,(l+1)}$ pour $\lambda_{i\,(l+1)}$. Si $a_{l+1}\leqq 0$ ou $\varepsilon_{i\,(l+1)}^{(m)}/a_{l+1}>1$ on pose $b_{i\,(l+1)}=b'_{i\,(l+1)}$ et $B_{i\,(l+1)}=B'_{i\,(l+1)}$. En r étapes on obtient les bornes $b_{k\,(r)}=b_k$ et $B_{k\,(r)}$ qui permettent la détermination du paramètre a.

2. Modification de poids en ACP sur matrice de corrélation

Lorsqu'on modifie le poids w_j de l'u. s. j en $\tilde{w}_j$, les autres poids étant seulement normalisés (*cf.* situation du paragraphe II), la matrice de corrélation initiale R se trouve transformée en $\tilde{R}$ que l'on peut exprimer sous la forme $\tilde{R}=T(R+\beta_j y_j y_j^t)\,T$ [Benasseni (1985)] avec $y_j^t=(y_{j_1},\ldots,y_{jp})$ le vecteur initial des coordonnées centrées réduites de l'u. s. j et $T=\mathrm{diag}(1+\beta_j y_{jk}^2)^{-1/2}$. Posons $\Delta=\mathrm{diag}\,y_{jk}^2$, alors T peut être exprimée par un développement limité à l'ordre m de la forme :

$$T=I+\sum_{k=1}^{m}\beta_j^k(\omega_k\Delta^k)+O(\beta_j^{m+1})\qquad(\omega_k\text{ réel})$$

et il en est de même pour $\tilde{R}$ avec

$$\tilde{R} = R + \sum_{k=1}^{m} \beta_j^k A_k + O(\beta_j^{m+1})$$

les A_k désignant des matrices symétriques.

On déduit de ce dernier développement des approximations d'ordre m, $\tilde{\lambda}_i^{(m)}$ et $\tilde{u}_i^{(m)}$, pour les valeurs propres et les vecteurs propres de $\tilde{R}$, auxquels on applique les résultats précédents fondés sur les quotients de Rayleigh.

La détermination du coefficient a se fait toujours à partir d'une première localisation des valeurs propres qui se calque sur la présentation faite au paragraphe II.3 [mais on pourra à nouveau pour plus de détails consulter Benasseni (1985) ou (1986)].

A titre d'exemple on notera les expressions suivantes à l'ordre 1.

$$\tilde{R} = R + \beta_j \{ y_j y_j^t - (1/2)(R\Delta + \Delta R) \} + O(\beta_j^2).$$

Avec $u_k^t = (u_{k_1}, \ldots, u_{k_p})$ le k-ième vecteur propre de R et z_{jk} la coordonnée de y_j sur l'axe engendré par u_k on a :

$$\tilde{\lambda}_i^{(1)} = \lambda_i + \beta_j \left\{ z_{ji}^2 - \lambda_i \sum_{k=1}^{p} u_{ik}^2 y_{jk}^2 \right\}$$

$$\tilde{u}_i^{(1)} = u_i - \beta_j \left\{ z_{ji} \sum_{k \neq i} (\lambda_k - \lambda_i)^{-1} z_{jk} u_k \right.$$

$$\left. - (1/2) \sum_{k \neq i} (\lambda_k + \lambda_i)(\lambda_k - \lambda_i)^{-1} \sum_{l=1}^{p} u_{kl} u_{il} y_{jl}^2 \right\}.$$

Pour des raisons de concision nous ne détaillons pas le cas de la modification de plusieurs poids en A.C.P. sur matrice de corrélation. Les développements reprennent avec une formulation plus complexe la méthodologie exposée au paragraphe III.1.

IV. ILLUSTRATION PRATIQUE

L'illustration proposée utilise un tableau de données présenté par Diday *et al.* (1982). Ce tableau donne 12 notes moyennes attribuées selon différents critères par un groupe d'utilisateurs à 76 ordinateurs (fabriqués par 10 constructeurs). On obtient donc un tableau de 76 u. s. et 12 variables. Les ordinateurs sont répartis en fonction des types de modèles et des constructeurs

TABLE

CONSTRUCTEURS ET MODELES	EFFECTIF r	i	$\tilde{\lambda}_i$	$q_i^{(0)}$	$q_i^{(1)}$	$\tilde{\lambda}_i^{(1)}$
AMDALH	1	1	74.0561	74.0388	74.0561	74.0513
		2	39.5761	39.5756	39.5761	39.5767
BURROUGHS	8	1	72.7537	71.6207	72.6764	71.8951
		2	26.5296	22.1492	23.3284	24.3462
CONTROL DATA	4	1	75.5748	75.4222	75.5746	75.4684
		2	39.0619	39.0960	39.0616	39.1694
DIGITAL EQUIP^T	2	1	75.4258	75.3701	75.4253	75.3829
		2	37.9536	37.7240	37.9537	37.7877
HONEYWELL	10	1	61.9363	61.4113	61.8978	62.8993
		2	41.1366	40.9762	41.1319	40.9788
IBM 360	9	1	81.7839	81.7228	81.7824	81.7379
		2	42.0592	41.9049	42.0572	42.0314
IBM 370	11	1	82.0626	81.4918	82.0615	82.1246
		2	41.3325	41.5842	41.3253	42.1033
IBM 3	6	1	70.4784	70.3742	70.4763	71.0341
		2	42.3636	42.3628	42.3636	42.3631
IBM	4	1	74.9367	74.7547	74.9329	74.7549
		2	39.4829	39.2717	39.4664	39.3493
ITEL	1	1	73.0626	73.0206	73.0626	73.0466
		2	39.5697	39.5712	39.5697	39.5723
NCR	8	1	74.2818	73.9919	74.2805	74.1208
		2	42.3041	42.3578	42.3031	42.3619
UNIVAC	11	1	70.7144	70.2004	70.7084	70.3734
		2	40.7952	40.6184	40.7859	40.6264
XEROX	1	1	74.9842	74.9840	74.9842	74.9841
		2	38.7698	38.7580	38.7698	38.7698
HO7 - HO8 - UNB	3	1	59.3040	58.1881	59.1319	58.2110
		2	38.9566	39.8030	39.1068	39.8050

AU 1

$d_i^{(0)} \ll \tilde{\lambda}_i \ll D_i^{(0)}$	$d_i^{(1)} \ll \tilde{\lambda}_i \ll D_i^{(1)}$	$\lvert\tilde{\lambda}_i - q_i^{(0)}\rvert$	$\lvert\tilde{\lambda}_i - q_i^{(1)}\rvert$	$\lvert\tilde{\lambda}_i - \tilde{\lambda}_i^{(1)}\rvert$
74.0071 ; 74.0705 39.5638 ; 39.5873	74.0561 ; 74.0561 39.5761 ; 39.5761	0.0173 0.0006	0.0000 0.0000	0.0048 0.0005
69.7508 ; 73.4907 / /	72.5757 ; 72.7772 / /	1.1330 4.3804	0.0773 3.2011	0.8587 2.1833
75.2246 ; 75.6198 38.6103 ; 39.5818	75.5743 ; 75.5750 39.0600 ; 39.0633	0.1527 0.0342	0.0002 0.0002	0.1065 0.1075
75.3008 ; 75.4393 37.2622 ; 38.1859	75.4248 ; 75.4259 37.9500 ; 37.9575	0.0557 0.2296	0.0005 0.0001	0.0429 0.1659
60.0982 ; 62.7244 40.6169 ; 41.3356	61.8165 ; 61.9791 41.1210 ; 41.1427	0.5250 0.1604	0.0385 0.0048	0.9630 0.1578
81.6384 ; 81.8072 41.4117 ; 42.3980	81.7800 ; 81.7847 42.0492 ; 42.0652	0.0611 0.1544	0.0016 0.0020	0.0460 0.0279
80.7559 ; 82.2277 39.1966 ; 43.9717	82.0600 ; 82.0631 41.3028 ; 41.3477	0.5708 0.2517	0.0010 0.0072	0.0621 0.7709
70.1470 ; 70.6013 42.3598 ; 42.3659	70.4721 ; 70.4805 42.3636 ; 42.3637	0.1042 0.0008	0.0021 0.0000	0.5557 0.0005
74.4761 ; 75.0333 38.8094 ; 39.7341	74.9268 ; 74.9390 39.4367 ; 39.4961	0.1820 0.2112	0.0038 0.0165	0.1819 0.1336
72.9396 ; 73.1016 39.5512 ; 39.5911	73.0625 ; 73.0626 39.5697 ; 39.5697	0.0419 0.0015	0.0000 0.0000	0.0160 0.0027
73.5414 ; 74.4423 41.9613 ; 42.7544	74.2788 ; 74.2823 42.3009 ; 42.3054	0.2899 0.0537	0.0013 0.0010	0.1610 0.0578
69.2251 ; 71.1758 39.8329 ; 41.4040	70.7000 ; 70.7168 40.7546 ; 40.8172	0.5139 0.1768	0.0060 0.0093	0.3409 0.1688
74.9837 ; 74.9843 38.7290 ; 38.7869	74.9842 ; 74.9842 38.7698 ; 38.7698	0.0002 0.0118	0.0000 0.0000	0.0001 0.0000
56.3008 ; 60.0755 38.4068 ; 41.1993	58.9385 ; 59.3253 38.8478 ; 39.3659	1.1158 0.8465	0.1720 0.1503	1.0930 0.8484

TABL

CONSTRUCTEURS DE MODELES	EFFECTIF r	i	$\|\|\tilde{u}_i - u_i\|\| =$	$\|\|\tilde{u}_i - u_i\|\| \leqslant$	$\|\|\tilde{u}_i - \tilde{u}_i^{(1)}\|\| =$
AMDALH	1	1 2	0.0170 0.0124	0.0303 0.0285	0.0001 0.0001
BURROUGHS	8	1 2	0.1430 0.9197	0.2288 /	0.0379 0.7240
CONTROL DATA	4	1 2	0.0599 0.0767	0.0742 0.1932	0.0022 0.0060
DIGITAL EQUIP^T	2	1 2	0.0355 0.1360	0.0431 0.1934	0.0032 0.0090
HONEYWELL	10	1 2	0.1078 0.0785	0.2553 0.1580	0.0305 0.0146
IBM 360	9	1 2	0.0352 0.0855	0.0463 0.1753	0.0053 0.0120
IBM 370	11	1 2	0.1125 0.1400	0.1399 0.4278	0.0042 0.0206
IBM 3	6	1 2	0.0420 0.0090	0.0900 0.0141	0.0062 0.0010
IBM	4	1 2	0.0597 0.1139	0.0892 0.1649	0.0084 0.0323
ITEL	1	1 2	0.0262 0.0159	0.0492 0.0372	0.0004 0.0003
NCR	8	1 2	0.0821 0.0905	0.1201 0.1531	0.0056 0.0071
UNIVAC	11	1 2	0.1034 0.1225	0.1839 0.2220	0.0125 0.0298
XEROX	1	1 2	0.0024 0.0201	0.0027 0.0462	0.0001 0.0005
HO7 - HO8 - UNB	3	1 2	0.2181 0.2112	0.3258 0.3115	0.0904 0.0899

EAU 2

$\|\|\tilde{u}_i - \tilde{u}_i^{(1)}\|\| \leqslant$	$\cos(u_i, \tilde{u}_i) =$	$\cos(u_i, \tilde{u}_i) \geqslant$	$\cos(\tilde{u}_i^{(1)}, \tilde{u}_i) =$	$\cos(\tilde{u}_i^{(1)}, \tilde{u}_i) \geqslant$
0.0001	0.9999	0.9848	1.0000	0.9999
0.0002	0.9999	0.9857	1.0000	1.0000
0.0490	0.9900	0.8856	0.9993	0.9755
/	0.5770	/	0.7379	/
0.0031	0.9982	0.9629	1.0000	0.9984
0.0114	0.9971	0.9034	1.0000	0.9943
0.0039	0.9994	0.9784	1.0000	0.9981
0.0173	0.9907	0.9033	1.0000	0.9913
0.0626	0.9942	0.8724	0.9995	0.9687
0.0266	0.9969	0.9210	0.9999	0.9867
0.0077	0.9994	0.9768	1.0000	0.9961
0.0220	0.9963	0.9123	0.9999	0.9890
0.0062	0.9937	0.9300	1.0000	0.9969
0.0423	0.9902	0.7861	0.9998	0.9788
0.0122	0.9991	0.9550	1.0000	0.9939
0.0018	1.0000	0.9929	1.0000	0.9991
0.0131	0.9982	0.9554	1.0000	0.9934
0.0395	0.9935	0.9176	0.9994	0.9802
0.0005	0.9996	0.9754	1.0000	0.9997
0.0006	0.9999	0.9814	1.0000	0.9997
0.0074	0.9966	0.9400	1.0000	0.9963
0.0115	0.9959	0.9235	1.0000	0.9943
0.0168	0.9947	0.9080	0.9999	0.9916
0.0431	0.9925	0.8890	0.9996	0.9784
0.0001	1.0000	0.9986	1.0000	1.0000
0.0008	0.9998	0.9769	1.0000	0.9996
0.0989	0.9762	0.8371	0.9959	0.9505
0.1376	0.9777	0.8442	0.9960	0.9312

en 13 groupes d'effectifs variés. Un quatorzième groupe, dont l'intérêt sera discuté ultérieurement, a été introduit de manière artificielle en rassemblant deux ordinateurs Honeywell (H07, H08) avec l'un du groupe Univac (UNB).

L'A.C.P. est effectuée sur matrice de variance puisque les variables sont de nature homogène et nous présentons ici les résultats correspondant aux variations des valeurs propres et vecteurs propres occasionnées par la suppression de chacun des groupes. Pour des raisons de concision nous limitons notre présentation aux deux plus grandes valeurs propres (initialement $\lambda_1 = 74{,}004\,6$ et $\lambda_2 = 39{,}137\,3$) et aux vecteurs propres associés. Le tableau I rassemble les résultats relatifs aux valeurs propres. Y figurent pour la suppression de chaque groupe et pour $i = 1, 2$, la nouvelle valeur propre $\tilde{\lambda}_i$, ses approximations $q_i^{(0)}$, $q_i^{(1)}$ et $\tilde{\lambda}_i^{(1)}$ ainsi que les écarts avec ces dernières et les encadrements que l'on obtient pour $\tilde{\lambda}_i$ à partir de la relation (II.4) avec $m = 0$ ($d_i^{(0)} \leqq \tilde{\lambda}_i \leqq D_i^{(0)}$) et $m = 1$ ($d_i^{(1)} \leqq \tilde{\lambda}_i \leqq D_i^{(1)}$). Dans le tableau II on trouve les résultats relatifs aux vecteurs propres avec pour chaque groupe supprimé et pour $i = 1, 2$, la différence en norme et le cosinus entre le nouveau vecteur propre $\tilde{u}_i$ et ses approximations d'ordre zéro et un, ainsi que la majoration de cette différence obtenue par application de la relation II.2 et la minoration du cosinus donnée par (II.5). Tous les résultats numériques sont donnés avec une précision de 10^{-4} qui est largement suffisante d'un point de vue pratique.

Les principaux commentaires que nous suggère la lecture des tableaux I et II sont les suivants :

Il apparaît tout d'abord que la qualité des approximations peut varier assez sensiblement selon le groupe d'ordinateurs supprimé ce qui souligne donc bien, a posteriori, la nécessité d'une méthodologie permettant d'apprécier cette qualité. On note ainsi par exemple que l'approximation de $\tilde{\lambda}_i$ par $q_i^{(1)}$ et $\lambda_i^{(1)}$ est très satisfaisante pour la suppression du groupe IBM 360, un peu moins précise pour la suppression des trois ordinateurs H07-H08-UNB, plutôt approximative en ce qui concerne $\tilde{\lambda}_2$ pour la suppression du groupe Burroughs.

On constate que dans l'exemple étudié les relations (II.2), (II.3), (II.4), (II.5) ont pu être utilisées (par le biais de la procédure itérative présentée au paragraphe III.1) dans tous les cas sauf pour la deuxième valeur propre et le vecteur propre associé, dans l'étude de la suppression du groupe Burroughs. Dans ce dernier cas la variation du deuxième axe est très importante ($\| u_2 - \tilde{u}_2 \| = 0{,}92$ et $\cos(u_2, \tilde{u}_2) = 0{,}58$) et l'approximation d'ordre 1 imprécise ($\| u_2 - \tilde{u}_2^{(1)} \| = 0{,}72$, $\cos(u_2, \tilde{u}_2^{(1)}) = 0{,}74$). Dans la mise en œuvre de la méthode itérative présentée en III.1, on s'est alors trouvé lors d'une étape dans la situation du paragraphe II.4.1.*b* qui ne permet plus de définir le paramètre

a, $q_2^{(0)}$ et $q_2^{(1)}$ étant trop éloignés de $\tilde{\lambda}_2$. En fait l'échec de la méthode lié à la mauvaise précision de $\tilde{u}_2^{(1)}$ et de $\tilde{q}_2^{(1)}$ par voie de conséquence n'est pas vraiment surprenant si l'on examine les coordonnées z_{j_2} des ordinateurs Burroughs sur le deuxième axe. Ces dernières sont toutes du même signe et grandes en valeur absolue. On peut alors penser que dans le développement

$$\tilde{u}_2 = u_2 + \sum_{k=1}^{m} \beta^k v_k + O(\beta_j^{m+1}),$$

$\beta^2 v_2$ devient prépondérant par rapport à βv_1 puisque v_2 [dont on trouvera l'expression dans Critchley (1985)] fait intervenir des termes en $z_{j_2}^2$et surtout $z_{j_2}^3$.

Les variations des valeurs propres et vecteurs propres ne sont pas directement liées à l'effectif du groupe supprimé. A cet égard les formulations (III. 3) et (III. 4) sont explicites et le dernier groupe (H07-H08-UNB) a été formé pour montrer que la suppression d'un petit nombre d'u. s. (ici $r=3$) particulières peut avoir des conséquences plus importantes que celle d'un groupe d'effectif plus élevé comme Honeywell ou IBM 360.

Pour les groupes supprimés ici (à l'exception du groupe Burroughs) les approximations d'ordre $m=1$, $\tilde{u}_i^{(1)}$ et $q_i^{(1)}$ apparaissent en général d'une qualité suffisante pour dispenser de recalculer l'analyse ou de s'intéresser à des approximations d'ordre supérieur. Les encadrements $d_i^{(1)} \leqq \tilde{\lambda}_i \leqq D_i^{(1)}$ traduisent d'une manière très précise cette qualité d'approximation de $\tilde{\lambda}_i$ par $q_i^{(1)}$ et améliorent de façon sensible les encadrements antérieurs [Benasseni (1985), (1986)] qui restent cependant indispensables dans la détermination du paramètre a (paragraphes II. 3, II. 4). On note en outre que $q_i^{(1)}$ est pratiquement toujours plus précis que $\tilde{\lambda}_i^{(1)}$. D'une manière analogue la majoration de $\| \tilde{u}_i - \tilde{u}_i^{(1)} \|$ (relation (II. 2) et la minoration du cosinus entre u_i et $\tilde{u}_i^{(1)}$ [relation (II. 5)] sont un assez bon reflet de l'écart réel entre $\tilde{u}_i$ et $\tilde{u}_i^{(1)}$. Les résultats proposés nous semblent donc pouvoir être considérés comme des outils utiles dans l'étude de la stabilité des résultats d'une A.C.P.

REMERCIEMENTS

L'auteur remercie le Professeur B. Lemaire pour ces commentaires sur le manuscrit original.

BIBLIOGRAPHIE

J. BENASSENI, *Influence des poids des unités statistiques sur les valeurs propres en*

analyse en composantes principales, Revue de Statistique Appliquée, vol. 33, n° 4, 1985, p. 41-55.

J. BENASSENI, *Une amélioration d'un résultat concernant l'influence d'une unité statistique sur les valeurs propres en analyse en composantes principales*, Statistique et Analyse des Données, vol. 11, n° 1, 1986, p. 42-63.

F. CRITCHLEY, *Influence in Principal Component Analysis*, Biometrika, vol. 72, n° 3, 1985, p. 627-636.

E. DIDAY, J. LEMAIRE, J. POUGET et F. TESTU, *Eléments d'analyse de données* Dunod, 1982.

B. ESCOFIER et B. LEROUX, (*a*), *Étude de trois problèmes de stabilité en analyse factorielle*, Publications de l'I.S.U.P., vol. 21, 1976, p. 1-48.

B. ESCOFIER et B. LEROUX, (*b*) *Influence d'un élément sur les facteurs en analyse des correspondances*, Cahiers de l'Analyse des Données, vol. 1, n° 3, 1976, p. 297-318.

B. ESCOFIER, B. LEROUX, *Mesure de l'influence d'un descripteur sur une analyse en composantes principales*, Publications de l'I.S.U.P., vol. 22, 1977, p. 25-44.

B. ESCOFIER, *Stabilité et approximation en analyse factorielle*, Thèse de Doctorat, Université Paris-VI, 1979.

J. H. WILKINSON, *The Algebraic Eigenvalue Problem*, Clarendon Press, 1965.