REVUE DE STATISTIQUE APPLIQUÉE

J. BENASSENI

Stabilité de l'analyse en composantes principales par rapport à une perturbation des données

Revue de statistique appliquée, tome 34, n° 3 (1986), p. 49-64

<http://www.numdam.org/item?id=RSA_1986__34_3_49_0>

STABILITÉ DE L'ANALYSE EN COMPOSANTES PRINCIPALES PAR RAPPORT A UNE PERTURBATION DES DONNÉES

J. BENASSENI

Unité de Biométrie, INRA-ENSA-USTL
9, place Viala, 34060 Montpellier

RÉSUMÉ

On étudie en analyse en composantes principales les conséquences sur les valeurs propres d'une modification de l'une des unités statistiques ou d'une variable. On montre en particulier qu'il est possible d'envisager une modification non triviale d'une u.s. et par extension de l'ensemble du tableau de données qui laisse invariante la matrice de variance et donc les valeurs propres et les axes principaux. Les résultats sont illustrés à partir de l'exemple bien connu des poissons d'Amiard.

ABSTRACT

This paper deals with the modification of an observation or a variate in principal component analysis and with its consequences upon the eigenvalues. A particular modification of the observations is pointed out such that the sample covariance matrix (and then the eigenvalues and principal axis) is kept unchanged. A practical illustration of the results, based on the well known example of Amiard fishes, is given.

Mots-clés : **Analyse en composantes principales, Modification d'une unité statistique, Modification d'une variable, Stabilité, Valeurs propres.**

1. INTRODUCTION

Une analyse en composantes principales (A.C.P.) se fonde sur la donnée d'un triplet statistique (X, Q, D). Le tableau X regroupe les mesures de n unités statistiques (u.s.) sur p variables. Chaque u.s. i, $i = 1, ..., n$ est caractérisée par le vecteur ligne $x_i = (x_{i1}, ..., x_{ip})$ de $\mathbf{R}^p$, chaque variable j, $j = 1, ..., p$ par le vecteur colonne v_j de $\mathbf{R}^n$ défini par $v_j' = (x_{1j}, ..., x_{nj})$. Un poids p_i étant attribué à chaque u.s. $(p_i \geqslant 0, \sum_{i=1}^{n} p_i = 1)$ la matrice $D = \text{diag } p_i$ joue le rôle de métrique sur l'espace $\mathbf{R}^n$ représentatif des variables tandis que la métrique Q sur $\mathbf{R}^p$ permet de calculer les distances entre u.s.

Lorsqu'on approfondit les propriétés de la méthode, il est intéressant de disposer de renseignements concernant la stabilité des inerties des différents axes factoriels par rapport à de petites fluctuations de l'un ou l'autre des trois paramètres du triplet (X, Q, D).

B. ESCOFIER en [4] et l'auteur en [2] ou [3] se sont intéressés au problème en ce qui concerne respectivement une perturbation de la métrique Q et une perturbation de la métrique D. Il reste néanmoins que pour le praticien, ce sont bien les problèmes de stabilité liés à d'éventuelles fluctuations du tableau X qui sont les plus importants et cela d'autant que la métrique Q elle-même est souvent liée à ce tableau.

Les conséquences sur les inerties d'une perturbation des variables ayant été étudiées en [4] par B. ESCOFIER lorsqu'on travaille sur les données centrées réduites, nous nous proposons ici de traiter principalement du problème d'une perturbation des u.s. Les inerties se définissent comme les valeurs propres des opérateurs $WD = (I - e'e\ D)\ X\ Q\ X'(I - D\ e'e)D$ ou $VQ = X'(I - D\ e'e\ D)\ D(I - e'e\ D)\ X\ Q$ avec $e = (1, ..., 1) \in \mathbf{R}^n$, W et V étant les matrices de produits scalaires respectivement entre u.s. et entre variables. Il n'y a aucune difficulté à constater que WD et VQ ont mêmes valeurs propres et, en ce qui concerne ces dernières, on peut raisonner sur les opérateurs symétriques $D^{1/2}\ W\ D^{1/2}$ et LVL' avec $D^{1/2} = \text{diag}\ \sqrt{p_i}$ et L défini par la décomposition de Choleski de Q selon $Q = L'L$.

Après avoir consacré le paragraphe 2 à quelques rappels algébriques nous présentons les principaux résultats dans le paragraphe 3 où nous soulignons en particulier l'existence d'une perturbation non triviale des u.s. n'affectant pas la matrice de variance. Nous montrons dans le paragraphe suivant comment la méthode utilisée pour traiter le problème d'une perturbation des u.s. se transpose au cadre d'une perturbation des variables et permet d'améliorer et de généraliser ce qui a été obtenu par B. ESCOFIER. Un dernier paragraphe illustre enfin les résultats obtenus en s'appuyant sur l'exemple des poissons d'Amiard dont on trouvera une présentation détaillée en [5].

2. PRÉLIMINAIRES ALGÉBRIQUES

Tout au long de l'étude, les valeurs propres d'une matrice carrée A d'ordre m seront notées $\lambda_i(A)$ $i = 1, ..., m$ avec $\lambda_1(A) \geqslant \lambda_2(A) \geqslant ... \geqslant \lambda_m(A)$.

L'ensemble de ce que nous présentons ici se fonde sur l'utilisation des deux propositions qui suivent. La première correspond à un résultat général classique que nous avons déjà utilisé pour traiter le problème d'une perturbation de D (cf. [3]). La seconde constitue un résultat technique simple plus spécifique du problème de perturbation de X que nous abordons ici.

Proposition 2.1

Soient A et B deux matrices symétriques d'ordre m. Alors pour i, j, k éléments de $\{1, ..., m\}$ vérifiant $j + k \leqslant i + 1$, les inégalités suivantes se trouvent vérifiées :

I — Inégalités de Weyl

$$\lambda_i(A + B) \leqslant \lambda_j(A) + \lambda_k(B) \tag{1}$$

$$\lambda_{m-i+1}(A + B) \geqslant \lambda_{m-j+1}(A) + \lambda_{m-k+1}(B) \tag{2}$$

$$\lambda_m(B) + \lambda_i(A) \leqslant \lambda_i(A + B) \leqslant \lambda_i(A) + \lambda_l(B) \tag{3}$$

II — Si A et B sont semi-définies positives

$$\lambda_i(AB) \leqslant \lambda_j(A)\,\lambda_k(B) \tag{4}$$

$$\lambda_{m-i+1}(AB) \geqslant \lambda_{m-j+1}(A)\,\lambda_{m-k+1}(B) \tag{5}$$

$$\lambda_m(B)\,\lambda_i(A) \leqslant \lambda_i(AB) \leqslant \lambda_i(A)\,\lambda_l(B) \tag{6}$$

Remarque

On trouvera une démonstration des inégalités de Weyl en [6]. Les inégalités II sont démontrées en [1] dans le cas où B est supposée définie positive. Il est immédiat d'étendre le résultat au cas où B est semi-définie positive seulement. Pour cela on pourra se reporter à [2].

Proposition 2.2

Soient u et v deux vecteurs colonnes de $\mathbf{R}^m$ linéairement indépendants, r, s, t trois réels. On désigne par $(.\,|\,.)$ le produit scalaire habituel sur $\mathbf{R}^m$ et par $\|.\|$ la norme associée. Alors si $rt - s^2 \neq 0$ la matrice $r\,u\,u' + s(uv' + vu') + t\,vv'$ est de rang deux et ses deux valeurs propres non nulles s'expriment sous la forme

$$\frac{1}{2}\left[\alpha \pm \sqrt{\alpha^2 - 4(rt - s^2)\beta}\right]$$

avec $\alpha = r\|u\|^2 + 2s(u\,|\,v) + t\|v\|^2$ et $\beta = \|u\|^2\,\|v\|^2 - (u\,|\,v)^2$.

De plus, ces deux valeurs propres sont de signes opposés si $rt - s^2 < 0$ et de même signe si $rt - s^2 > 0$.

La démonstration de ce résultat ne présente aucune difficulté, on pourra la trouver en [2].

Corollaire 2.3

Les deux valeurs propres non nulles de la matrice $uu' - vv'$ s'expriment sous la forme :

$$\frac{1}{2}\left[\|u\|^2 - \|v\|^2 \pm \|u + v\| \, \|u - v\|\right]$$

3. MODIFICATION DES UNITÉS STATISTIQUES DU TABLEAU

Nous nous intéressons ici à la modification des valeurs propres engendrée par la transformation d'une u.s. x_i en $\tilde{x}_i = x_i + e_i$ où $e_i = (e_{i1}, ..., e_{ip})$ est un vecteur quelconque de $\mathbf{R}^p$. La transformation de x_i en $\tilde{x}_i$ entraine une modification du centre de gravité initial g en $\tilde{g}$ avec :

$$g = \sum_{k=1}^{n} p_k x_k \quad \text{et} \quad \tilde{g} = \sum_{k=1}^{n} p_k x_k + p_i e_i = g + p_i e_i$$

et une modification de la matrice de variance initiale V en $\tilde{V}$ avec :

$$V = \sum_{k=1}^{n} p_k (x_k - g)' (x_k - g)$$

$$\tilde{V} = \sum_{\substack{k=1 \\ k \neq i}}^{n} p_k (x_k - \tilde{g})' (x_k - \tilde{g}) + p_i (\tilde{x}_i - \tilde{g})' (\tilde{x}_i - \tilde{g})$$

3.1. Relation entre V et $\tilde{V}$

Le théorème d'Huygens permet d'exprimer $\tilde{V}$ sous la forme :

$$\tilde{V} = \sum_{k=1}^{n} p_k (x_k - g)' (x_k - g) + p_i (\tilde{x}_i - g)' (\tilde{x}_i - g) - (g - \tilde{g})' (g - \tilde{g})$$

Compte tenu des relations $\tilde{x}_i = x_i + e_i$ et $\tilde{g} = g + p_i e_i$ on obtient alors :

$$\tilde{V} = V + p_i [e_i'(x_i - g) + (x_i - g)' e_i + e_i' e_i] - p_i^2 e_i' e_i$$

$$\tilde{V} = V + p_i [e_i'(x_i - g) + (x_i - g)' e_i + (1 - p_i) e_i' e_i] \qquad (3.1)$$

Cette relation amène principalement deux remarques :

1) On note tout d'abord que dans le cas d'une colinéarité de $x_i - g$ et de e_i avec $x_i - g = \alpha e_i$ ($\alpha \in \mathbf{R}$) la relation (3.1) devient

$$\tilde{V} = V + p_i (2\alpha + 1 - p_i) e_i' e_i \qquad (3.2)$$

On constate que lorsque $\alpha = (p_i - 1)/2$ on a $\tilde{V} = V$. La matrice de variance reste alors invariante et la matrice de corrélation qui peut s'en déduire également. Pour cette transformation les valeurs propres et les axes principaux de l'A.C.P. reste donc inchangés. Cette transformation se caractérise par $x_i - g = [(p_i - 1)/2] e_i$ ce que l'on peut encore traduire (au moyen d'un rapide calcul) par la relation

$$\tilde{x}_i - g = -\frac{1 + p_i}{1 - p_i} (x_i - g).$$

Lorsque p_i est petit, tout revient donc approximativement à prendre pour $\tilde{x}_i$ le symétrique de x_i par rapport à g. On notera que l'on peut appliquer de manière itérative cette transformation à d'autres u.s. et générer ainsi à partir du tableau initial des tableaux complètement différents mais auxquels correspondent toujours les mêmes matrices de variance et de corrélation. Cet aspect sera illustré dans le dernier paragraphe.

2) Dans le cas fréquent où l'on travaille avec la métrique $Q = \text{diag } 1/s_j^2$ (s_j^2 variance de la j-ième variable) la transformation de x_i en $\tilde{x}_i$ conduit à une nouvelle métrique $\tilde{Q} = \text{diag } 1/\tilde{s}_j^2$ ($\tilde{s}_j^2$ variance de la j-ième variable après transformation de x_i) et l'on a la relation $\tilde{Q} = Q(Q^{-1}\tilde{Q})$. Le passage de Q à $\tilde{Q}$ traduisant le changement d'échelle sur les variables se fait donc au travers de $\tilde{Q}Q^{-1} = \text{diag } s_j^2/\tilde{s}_j^2$. Comme la relation (3.1) montre que pour $j = 1, ..., p$ on a

$$\tilde{s}_j^2 = s_j^2 + p_i[2\, e_{ij}(x_{ij} - g_j) + (1 - p_i)\, e_{ij}^2]$$

avec $g = (g_1, ..., g_p)$, il est facile de voir pour quelles variables il y a contraction ou dilatation des longueurs selon que l'on a $s_j^2/\tilde{s}_j^2 < 1$ ou $s_j^2/\tilde{s}_j^2 > 1$. La situation se résume ainsi :

* si $x_{ij} - g_j > 0$:

$s_j^2/\tilde{s}_j^2 \geqslant 1$ pour $-2(x_{ij} - g_j)/(1 - p_i) \leqslant e_{ij} \leqslant 0$

$s_j^2/\tilde{s}_j^2 \leqslant 1$ pour $e_{ij} \leqslant -2(x_{ij} - g_j)/(1 - p_i)$ ou $e_{ij} \geqslant 0$

* si $x_{ij} - g_j = 0$:

on a $s_j^2/\tilde{s}_j^2 \leqslant 1$

* si $x_{ij} - g_j < 0$:

$s_j^2/\tilde{s}_j^2 \geqslant 1$ pour $0 \leqslant e_{ij} \leqslant -2(x_{ij} - g_j)/(1 - p_i)$

$s_j^2/\tilde{s}_j^2 \leqslant 1$ pour $e_{ij} \leqslant 0$ ou $e_{ij} \geqslant -2(x_{ij} - g_j)/(1 - p_i)$

3.2. Variation des valeurs propres

Considérons tout d'abord le cas où l'on effectue l'A.C.P. sur matrice de variance. La relation (3.1) s'écrit :

$$\tilde{V} = V + p_i B \tag{3.3}$$

si l'on pose $B = e_i'(x_i - g) + (x_i - g)' e_i + (1 - p_i)\, e_i' e_i$.

Dans le cas général où e_i est supposée non colinéaire à $x_i - g$, la proposition 2.2 appliquée avec $u = (x_i - g)'$, $v = e_i'$ et $r = 0$, $s = 1$, $t = 1 - p_i$, donne de manière précise les expressions des deux valeurs propres non nulles $\lambda_1(B)$ et $\lambda_p(B)$. A partir de (3.3) la proposition 2.1.I permet alors d'obtenir pour $k = 1, ..., p$:

$$p_i\lambda_p(B) + \lambda_k(V) \leqslant \lambda_k(\tilde{V}) \leqslant \lambda_k(V) + p_i\lambda_1(B)$$

La matrice B étant de rang deux, on note en outre que si $p \geqslant 3$ on a :

$$\lambda_k(\widetilde{V}) \leqslant \lambda_{k-1}(V) \quad \text{pour} \quad k = 2, ..., p$$
$$\lambda_k(\widetilde{V}) \geqslant \lambda_{k+1}(V) \quad \text{pour} \quad k = 1, ..., p-1$$

Remarque

Dans le cas particulier de colinéarité entre $x_i - g$ et e_i la relation (3.2) montre que $B = (2\alpha - 1 - p_i)\, e_i' e_i$ est de rang un.

On remarque alors qu'on peut déterminer le sens de variation des valeurs propres de la matrice de variance puisqu'on a pour $k = 1, ..., p$

si $\alpha > (p_i - 1)/2$ $\quad \lambda_k(V) \leqslant \lambda_k(\widetilde{V}) \leqslant \lambda_k(V) + p_i(2\alpha + 1 - p_i)\, \|e_i\|^2$

si $\alpha < (p_i - 1)/2$ $\quad \lambda_k(V) + p_i(2\alpha + 1 - p_i)\, \|e_i\|^2 \leqslant \lambda_k(\widetilde{V}) \leqslant \lambda_k(V)$

Lorsqu'on s'intéresse à l'A.C.P. sur matrice de corrélation, la remarque 2 du paragraphe 3.1 précèdent permet de cerner le mécanisme de transformation de la métrique $Q = \text{diag}\ 1/s_j^2$ en $\widetilde{Q} = \text{diag}\ 1/\widetilde{s}_j^2$. La matrice de corrélation $R = Q^{1/2} V Q^{1/2}$ (avec $Q^{1/2} = \text{diag}\ 1/s_j$) se trouve modifiée en $\widetilde{R} = \widetilde{Q}^{1/2} \widetilde{V} \widetilde{Q}^{1/2}$ (avec $\widetilde{Q}^{1/2} = \text{diag}\ 1/\widetilde{s}_j$). On obtient alors à partir de la relation (3.1) :

$$\widetilde{R} = \widetilde{Q}^{1/2} \widetilde{V} \widetilde{Q}^{1/2} =$$

$$\widetilde{Q}^{1/2} V \widetilde{Q}^{1/2} + p_i \widetilde{Q}^{1/2} [e_i' (x_i - g) + (x_i - g)' e_i + (1 - p_i)\, e_i' e_i]\, \widetilde{Q}^{1/2}$$

$$\widetilde{R} = \widetilde{Q}^{1/2} Q^{-1/2} R Q^{-1/2} \widetilde{Q}^{1/2} + p_i \widetilde{B} \qquad (3.4)$$

avec $\quad \widetilde{B} = \widetilde{Q}^{1/2} [e_i'(x_i - g) + (x_i - g)' e_i + (1 - p_i)\, e_i' e_i]\, \widetilde{Q}^{1/2}$

Comme précédemment, la proposition 2.2 appliquée avec

$$u = \widetilde{Q}^{1/2} (x_i - g)',$$

$v = \widetilde{Q}^{1/2} e_i'$ et $r = 0$, $s = 1$, $t = 1 - p_i$ donne les expressions des deux valeurs propres non nulles $\lambda_1(\widetilde{B})$ et $\lambda_p(\widetilde{B})$ de $\widetilde{B}$. A partir de (3.4) la partie I de la proposition 2.1 permet d'obtenir pour $k = 1, ..., p$:

$$\lambda_k(\widetilde{Q}^{1/2} Q^{-1/2} R Q^{-1/2} \widetilde{Q}^{1/2}) + p_i \lambda_p(\widetilde{B})$$

$$\leqslant \lambda_k(\widetilde{R}) \leqslant \lambda_k(\widetilde{Q}^{1/2} Q^{-1/2} R Q^{-1/2} \widetilde{Q}^{1/2}) + p_i \lambda_1(\widetilde{B} \qquad (3.5)$$

Comme $\widetilde{Q}^{1/2} Q^{1/2} R Q^{-1/2} \widetilde{Q}^{-1/2}$ a les mêmes valeurs propres que $R(Q^{-1}\widetilde{Q})$, la partie II de la même proposition conduit à l'encadrement :

$$\left[\min_j \frac{s_j^2}{\widetilde{s}_j^2}\right] \lambda_k(R) \leqslant \lambda_k(\widetilde{Q}^{1/2} Q^{-1/2} R Q^{-1/2} \widetilde{Q}^{1/2}) \leqslant \left[\max_j \frac{s_j^2}{\widetilde{s}_j^2}\right] \lambda_k(R)$$

On obtient alors en définitive à partir de la relation (3.5) pour $k = 1, ..., p$:

$$\left[\min_j \frac{s_j^2}{\widetilde{s}_j^2}\right] \lambda_k(R) + p_i \lambda_p(\widetilde{B}) \leqslant \lambda_k(\widetilde{R}) \leqslant \left[\max_j \frac{s_j^2}{\widetilde{s}_j^2}\right] \lambda_k(R) + p_i \lambda_1(\widetilde{B}) \quad (3.6)$$

Remarques

1) Il est possible d'obtenir simplement des encadrements des valeurs propres lorsqu'on modifie plusieurs u.s. du tableau. Il suffit de procéder par encadrements successifs en considérant que l'on modifie successivement une u.s. après l'autre.

2) On note naturellement que les résultats précédents se transposent de manière immédiate au cas d'une métrique $Q = \text{diag } q_j$ quelconque. Cette dernière se trouvant modifiée en $\widetilde{Q} = \text{diag } \widetilde{q}_j$, suite à la modification de l'u.s. i, les encadrements précédents s'écrivent alors :

$$\left[\min_j \frac{\widetilde{q}_j}{q_j}\right] \lambda_k(R) + p_i\lambda_p(\widetilde{B}) \leqslant \lambda_k(\widetilde{R}) \leqslant \left[\max_j \frac{\widetilde{q}_j}{q_j}\right] \lambda_k(R) + p_i\lambda_1(\widetilde{B})$$

4. MODIFICATION DES VARIABLES DU TABLEAU

On étudie ici les conséquences d'une modification de l'une des variables du tableau, les résultats se généralisant au cas de plusieurs variables sans difficulté.

On suppose que l'on s'intéresse à une modification de la variable j et que la métrique $Q = \text{diag } q_\ell$ se trouve alors transformée en $\widetilde{Q} = \text{diag } \widetilde{q}_\ell$ avec $\widetilde{q}_\ell = q_\ell$ pour $\ell = 1, ..., p$ et $\ell \neq j$. On suppose encore que toutes les variables v_ℓ $\ell = 1, ..., p$ sont centrées pour les poids de D et on envisage une modification de v_j qui conduit à une nouvelle variable centrée $\widetilde{v}_j$.

L'opérateur $D^{1/2} W D^{1/2} = \sum q_\ell (D^{1/2} v_\ell) (D^{1/2} v_\ell)'$ initial se trouve alors transformé en :

$$D^{1/2} \widetilde{W} D^{1/2} = \sum_{\ell}^{p} q_\ell (D^{1/2} v_\ell) (D^{1/2} v_\ell)' + \widetilde{q}_j (D^{1/2} \widetilde{v}_j) (D^{1/2} \widetilde{v}_j)'$$

$$D^{1/2} \widetilde{W} D^{1/2} = D^{1/2} W D^{1/2} + \widetilde{q}_j (D^{1/2} \widetilde{v}_j) (D^{1/2} \widetilde{v}_j)' - q_j (D^{1/2} v_j) (D^{1/2} v_j)' \tag{4.1}$$

Posons $A = \widetilde{q}_j (D^{1/2} \widetilde{v}_j) (D^{1/2} \widetilde{v}_j)' - q_j (D^{1/2} v_j) (D^{1/2} v_j)'$. Le corollaire 2.3 appliqué avec $u = \sqrt{\widetilde{q}_j}\, D^{1/2} \widetilde{v}_j$ et $v = \sqrt{q_j}\, D^{1/2} v_j$ donne les expressions des deux valeurs propres non nulles de A et l'on a donc par la proposition 2.1 pour $k = 1, ..., n$:

$$\lambda_n(A) + \lambda_k(D^{1/2} W D^{1/2}) \leqslant \lambda_k(D^{1/2} \widetilde{W} D^{1/2}) \leqslant \lambda_k(D^{1/2} W D^{1/2}) + \lambda_1(A) \tag{4.2}$$

La matrice A étant de rang deux on note également si $n \geqslant 3$:

$$\begin{aligned} &\text{pour } k \geqslant 2 \qquad \lambda_k(D^{1/2} \widetilde{W} D^{1/2}) \leqslant \lambda_{k-1}(D^{1/2} W D^{1/2}) \\ &\text{pour } k \leqslant n-1 \quad \lambda_k(D^{1/2} \widetilde{W} D^{1/2}) \geqslant \lambda_{k+1}(D^{1/2} W D^{1/2}) \end{aligned} \tag{4.3}$$

On constate que seuls les encadrements correspondant à $k \leqslant p$ sont intéressants puisque $D^{1/2}\,\widetilde{W}\,D^{1/2}$ est au plus de rang p par construction.

Remarques

1) Désignons par $\mathrm{var}(v_\ell)$ la variance d'une variable v_ℓ. On a alors $\|D^{1/2}\,v_\ell\|^2 = \mathrm{var}(v_\ell)$ puisque les variables ont été supposées centrées. A partir du corollaire 2.3, les valeurs propres $\lambda_1(A)$ et $\lambda_n(A)$ peuvent alors s'exprimer sous la forme suivante :

$$\lambda_1(A) = \frac{1}{2}\left[\mathrm{var}(\sqrt{\widetilde{q}_j}\,\widetilde{v}_j) - \mathrm{var}(\sqrt{q_j}\,v_j) + \sqrt{\mathrm{var}(\sqrt{\widetilde{q}_j}\,\widetilde{v}_j + \sqrt{q_j}\,v_j)\,\mathrm{var}(\sqrt{\widetilde{q}_j}\,\widetilde{v}_j - \sqrt{q_j}\,v_j)}\right]$$

$$\lambda_n(A) = \frac{1}{2}\left[\mathrm{var}(\sqrt{\widetilde{q}_j}\,\widetilde{v}_j) - \mathrm{var}(\sqrt{q_j}\,v_j) - \sqrt{\mathrm{var}(\sqrt{\widetilde{q}_j}\,\widetilde{v}_j + \sqrt{q_j}\,v_j)\,\mathrm{var}(\sqrt{\widetilde{q}_j}\,\widetilde{v}_j - \sqrt{q_j}\,v_j)}\right]$$

Dans le cas le plus courant où l'on travaille avec $Q = \mathrm{diag}\ 1/s_\ell^2$ et $\widetilde{Q} = \mathrm{diag}\ 1/\widetilde{s}_\ell^2$ ($s_\ell = \widetilde{s}_\ell$ pour $\ell \neq j$), $\sqrt{\widetilde{q}_j}\,\widetilde{v}_j$ et $\sqrt{q_j}\,v_j$ représentent la variable j centrée réduite après et avant modification respectivement. Les valeurs propres $\lambda_1(A)$ et $\lambda_n(A)$ prennent alors une forme particulièrement simple puisque d'une part $(\mathrm{var}(\sqrt{\widetilde{q}_j}\,\widetilde{v}_j) = \mathrm{var}(\sqrt{q_j}\,v_j) = 1$ et d'autre part la covariance entre $\sqrt{\widetilde{q}_j}\,\widetilde{v}_j$ et $\sqrt{q_j}\,v_j$ n'est autre que la corrélation $\mathrm{cor}(v_j, \widetilde{v}_j)$ entre la variable j avant perturbation et la variable j après perturbation. On obtient facilement :

$$\sqrt{\mathrm{var}(\sqrt{\widetilde{q}_j}\,\widetilde{v}_j + \sqrt{q_j}\,v_j)\,\mathrm{var}(\sqrt{\widetilde{q}_j}\,\widetilde{v}_j - \sqrt{q_j}\,v_j)} = 2\sqrt{1 - \mathrm{cor}^2(v_j, \widetilde{v}_j)}$$

et les encadrements (4.2) prennent alors la forme suivante :

$$\lambda_k(D^{1/2}\,W\,D^{1/2}) - \sqrt{1 - \mathrm{cor}^2(v_j, \widetilde{v}_j)} \leqslant \lambda_k(D^{1/2}\,\widetilde{W}\,D^{1/2}) \leqslant \lambda_k(D^{1/2}\,W\,D^{1/2}) + \sqrt{1 - \mathrm{cor}^2(v_j, \widetilde{v}_j)}$$

On retrouve ainsi une formulation obtenue par B. ESCOFIER [4] (p. 110-111) dans le cadre d'une approche moins générale.

2) Lorsqu'on s'intéresse à la modification de plusieurs variables, on procède par encadrements successifs comme dans le cas d'une modification des u.s.

5. ILLUSTRATION DES RÉSULTATS

Nous nous proposons ici d'illustrer les résultats obtenus en ce qui concerne une perturbation des u.s. en nous appuyant sur l'exemple des poissons d'Amiard. Cet exemple a l'avantage d'être classique et nous renvoyons à [5] pour une présentation détaillée. Nous nous bornerons ici à

rappeler que les données sont constituées par un ensemble de 24 poissons évoluant dans un milieu radioactif et pour lesquels on mesure 16 caractéristiques de taille et de radioactivité de différents organes. Le poisson 17 mort en cours d'expérimentation n'est pas pris en compte si bien qu'un poids de 1/23 est attribué à chacun des 23 poissons participant de manière effective à l'analyse statistique. L'ACP est effectuée sur matrice de corrélation car les 16 variables ont des variances sensiblement différentes.

Avant de nous intéresser à la perturbation laissant invariante la matrice de corrélation (remarque 1 du paragraphe 3.1), nous illustrons dans un premier temps les encadrements des valeurs propres fournis par la relation (3.6) lorsqu'on modifie une u.s. Nous avons choisi de modifier les coordonnées du poisson 20 qui a initialement la plus forte contribution au 1er axe factoriel. Le tableau qui suit donne les coordonnées initiales x_{20j} $j = 1, ..., 16$ de ce poisson puis ses coordonnées centrées réduites $(x_{20j} - g_j)/s_j$.

Variables	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
x_{20j} $j=1,...,16$	31	195	208	350	73	109	5	809	11	49	170	154	39	33	12	8
$(x_{20j}-g_j)/s_j$	2.120	1.782	1.369	2.127	2.843	-0.681	-1.310	2.474	2.949	-1.284	-1.170	-1.088	-0,805	-1.459	-0,631	-1.845

Initialement les deux premières valeurs propres ont respectivement pour valeurs $\lambda_1(R) = 7.607$ et $\lambda_2(R) = 3.763$.

Nous étudions les variations de ces deux valeurs propres pour différentes perturbations du poisson 20. Ces perturbations ont été choisies de manière à faire ressortir certains caractères essentiels des encadrements fournis par la relation (3.6). Cette dernière fait intervenir les deux valeurs propres $\lambda_1(\tilde{B})$ et $\lambda_p(\tilde{B})$ qui se définissent comme étant égales à $1/2\ [\alpha \pm \sqrt{\alpha^2 + 4\beta}]$ avec :

$$\alpha = 2\,(e_{20} \mid x_{20} - g)_{\tilde{Q}} + (1 - p_{20})\,\|e_{20}\|^2_{\tilde{Q}}$$

$$\beta = \|e_{20}\|^2_{\tilde{Q}}\,\|x_{20} - g\|^2_{\tilde{Q}} - (e_{20} \mid x_{20} - g)^2_{\tilde{Q}}$$

$(\cdot \mid \cdot)_{\tilde{Q}}$ désignant le produit scalaire sur $\mathbf{R}^p$ tel que pour deux vecteurs colonnes x et y on ait $(x \mid y)_{\tilde{Q}} = x'\tilde{Q}y$ et $\|\cdot\|_{\tilde{Q}}$ la norme associée.

Aussi nous résumons les caractéristiques de chaque perturbation étudiée dans un tableau où nous faisons figurer les 16 coordonnées du vecteur $\tilde{x}_{20} = x_{20} + e_{20}$ qui définit la perturbation, les rapports extrémaux $\max_{j=1,16} s_j^2/\tilde{s}_j^2$ et $\min_{j=1,16} s_j^2/\tilde{s}_j^2$ caractérisant le changement d'échelle, les quantités $\|x_{20} - g\|_{\tilde{Q}}$, $\|e_{20}\|_{\tilde{Q}}$, $(e_{20} \mid x_{20} - g)_{\tilde{Q}}$ qui interviennent dans les expressions de $\lambda_1(\tilde{B})$ et $\lambda_{16}(\tilde{B})$ elles-mêmes et enfin les deux premières valeurs propres $\lambda_i(\tilde{R})$ $i = 1,2$ que l'on obtient après perturbation ainsi que la borne inférieure m_i et la borne supérieure M_i qu'en donnent les encadrements de la relation (3.6).

1ère perturbation

Variables	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
$\tilde{x}_{20j}$ J=1,...,16	25	175	190	300	63	125	7	759	7	69	180	159	41	35	12	8
$s_j^2/\tilde{s}_j^2$ J=1,...,16	1.140	1.058	1.028	1.102	1.159	1.004	1.059	1.052	1.417	1.066	1.047	1.027	1.023	1.051	1.000	1.000

$\|\|x_{20}-g\|\|^2_{\tilde{Q}}$	$\|\|e_{20}\|\|^2_{\tilde{Q}}$	$(x_{20}-g\|e_{20})_{\tilde{Q}}$	$\lambda_1(\tilde{B})$	$\lambda_{16}(\tilde{B})$	$\lambda_1(\tilde{R})$	$[m_1, M_1]$	$\lambda_2(\tilde{R})$	$[m_2, M_2]$
56.911	7.087	-17.566	3.021	-31.374	7.220	[6.243 , 10.909]	3.947	[2.399 , 5.463]

2ème perturbation

Variables	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
$\tilde{x}_{20j}$ J=1,...,16	26	170	180	300	66	145	7	750	10	80	185	169	47	39	12	8
$s_j^2/\tilde{s}_j^2$ J=1,...,16	1.119	1.071	1.042	1.102	1.111	1.008	1.059	1.062	1.101	1.080	1.060	1.056	0.999	1.106	1.000	1.000

$\|\|x_{20}-g\|\|^2_{\tilde{Q}}$	$\|\|e_{20}\|\|^2_{\tilde{Q}}$	$(x_{20}-g\|e_{20})_{\tilde{Q}}$	$\lambda_1(\tilde{B})$	$\lambda_{16}(\tilde{B})$	$\lambda_1(\tilde{R})$	$[m_1, M_1]$	$\lambda_2(\tilde{R})$	$[m_2, M_2]$
53.978	10.502	-16.220	9.519	-31.914	6.919	[6.209 , 8.925]	4.130	[2.371 , 4.625]

3ème perturbation

Variables	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
$\tilde{x}_{20j}$ J=1,...,16	27	168	170	305	68	189	8	740	11	80	188	175	47	38	16	8
$s_j^2/\tilde{s}_j^2$ J=1,...,16	1.097	1.076	1.054	1.092	1.078	1.015	1.076	1.072	1.000	1.080	1.065	1.054	0.999	1.099	0.981	1.000

$\|\|x_{20}-g\|\|^2_{\tilde{Q}}$	$\|\|e_{20}\|\|^2_{\tilde{Q}}$	$(x_{20}-g\|e_{20})_{\tilde{Q}}$	$\lambda_1(\tilde{B})$	$\lambda_{16}(\tilde{B})$	$\lambda_1(\tilde{R})$	$[m_1, M_1]$	$\lambda_2(\tilde{R})$	$[m_2, M_2]$
52.816	14.054	-16.365	14.178	-33.465	6.778	[6.005 , 8.979]	4.306	[2.236 , 4.753]

4ème perturbation

Variables	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
$\tilde{x}_{20j}$ J=1,...,16	28	166	157	312	69	335	9	735	11	82	198	188	47	36	15	8
$s_j^2/\tilde{s}_j^2$ J=1,...,16	1.073	1.081	1.068	1.078	1.062	1.020	1.084	1.077	1.000	1.081	1.060	1.005	0.999	1.072	1.006	1.000

$\|\|x_{20}-g\|\|^2_{\tilde{Q}}$	$\|\|e_{20}\|\|^2_{\tilde{Q}}$	$(x_{20}-g\|e_{20})_{\tilde{Q}}$	$\lambda_1(\tilde{B})$	$\lambda_{16}(\tilde{B})$	$\lambda_1(\tilde{R})$	$[m_1, M_1]$	$\lambda_2(\tilde{R})$	$[m_2, M_2]$
52.498	18.350	-17.597	18.225	-35.867	6.687	[6.037 , 9.040]	4.394	[2.199 , 4.873]

D'un point de vue pratique le choix des perturbations proposées a été guidé par un double objectif :

— obtenir des variations des valeurs propres assez importantes pour pouvoir tester de manière valable le comportement des encadrements,

— obtenir une amplitude de perturbation (caractérisée par $\|e_{20}\|$) qui soit la plus grande possible tout en maintenant les rapports extrémaux des variances dans des limites raisonnables autour de l'unité. Notre but est ainsi que l'effet intrinsèque des perturbations envisagées ne soit pas occulté par le changement d'échelle de mesure sur les variables.

Nous présentons quatre perturbations. Le choix de la première perturbation a été guidé par notre intuition vis à vis des objectifs fixés. Chaque perturbation qui suit est construite à partir de la précédente en essayant de modifier les coordonnées de manière à se rapprocher toujours plus des objectifs proposés. Il convient à chaque fois de se reporter au tableau 1 des coordonnées initiales de x_{20} pour apprécier dans le détail la modification apportée à chaque coordonnée.

On constate que les perturbations proposées ont occasionné des variations sensibles des valeurs propres, diminution de la première valeur propre initialement égale à 7.607 jusqu'à 6.687, augmentation de la deuxième valeur propre initialement égale à 3.763 jusqu'à 4.394.

Pour la première valeur propre la borne inférieure m_1 apparaît toujours plus près de la vraie valeur $\lambda_1(\widetilde{R})$ que la borne supérieure M_1. A l'inverse, pour la deuxième valeur propre on note que la borne supérieur M_2 est toujours plus précise que la borne inférieure m_2. Ce phénomène est caractéristique des encadrements proposés qui présentent la même structure

$$\left[\min_j \frac{s_j^2}{\widetilde{s}_j^2}\right] \lambda_k(R) + p_i \lambda_p(\widetilde{B}) \leqslant \lambda_k(\widetilde{R}) \leqslant \left[\max_j \frac{s_j^2}{\widetilde{s}_j^2}\right] \lambda_k(R) + p_i \lambda_1(\widetilde{B})$$

quelle que soit la valeur propre $\lambda_k(\widetilde{R})$ considérée.

Ainsi $\lambda_p(\widetilde{B})$ se doit de suivre la diminution de la première valeur propre $\lambda_1(\widetilde{R})$ mais ce faisant conduit à une borne peu précise pour la deuxième valeur propre $\lambda_2(\widetilde{R})$ qui augmente au contraire. D'une manière analogue $\lambda_1(\widetilde{B})$ doit être suffisamment grande pour suivre l'augmentation de la deuxième valeur propre de $\lambda_2(\widetilde{R})$ mais ce faisant fournit une mauvaise borne supérieure pour la première valeur propre $\lambda_1(\widetilde{R})$.

En outre on note l'importance des rapports extrémaux des variances qui conditionnent souvent en grande partie la qualité des encadrements. Ainsi pour la première perturbation la mauvaise qualité de la borne M_1 est en partie imputable au rapport $\max_j (s_j^2/\widetilde{s}_j^2) = s_7^2/\widetilde{s}_7^2 = 1.417$. Il en résulte que l'encadrement $[m_1, M_1]$ est plus mauvais pour la première perturbation qui est d'amplitude relativement modeste ($\|e_{20}\|_{\widetilde{Q}}^2 = 7.087$ pour $\|x_{20} - g\|_{\widetilde{Q}}^2 = 56.911$) qe pour la quatrième qui est pourtant d'amplitude nettement supérieure

($\|e_{20}\|^2_{\tilde{Q}} = 18.350$ pour $\|x_{20} - g\|^2_{\tilde{Q}} = 52.498$) mais pour laquelle le rapport $\max_j (s_j^2/\tilde{s}_j^2) = s_7^2/\tilde{s}_7^2 = 1.084$ est bien plus proche de l'unité.

Cet exemple illustre bien la difficulté qu'il y a à prévoir les effets de la modification d'une u.s., le comportement déterminant des variances n'étant pas lié de manière simple à l'amplitude de la perturbation (cf. remarques 1 et 2 du paragraphe 3.1). A cet égard le comportement de la variance au cours des perturbations de la 12ième coordonnée (variable 12) est très représentatif. Alors qu'à partir de la valeur initiale de 154 on arrive progressivement au cours des quatre perturbations jusqu'à la valeur 188, on observe d'abord une augmentation du rapport des variances $s_{12}^2/\tilde{s}_{12}^2$ (perturbations 1 et 2) puis une diminution (perturbations 3 et 4).

En conclusion on voit donc que la modification d'une u.s. bien choisie dans un tableau de dimensions modestes, peut avoir des conséquences non négligeables et qu'il est souvent difficile de prévoir sur les inerties des axes.

Pour terminer nous illustrons la perturbation des u.s. qui a été soulignée dans la remarque 1 du paragraphe 3.1 comme laissant invariante la matrice des corrélations.

Etant donné un nuage d'u.s. dans $\mathbf{R}^p$ l'ACP définit le plan sur lequel la projection des u.s. conduira à la meilleure représentation de dimension 2 au sens du critère de l'inertie. La remarque permet de voir que ce plan reste le meilleur plan de projection pour une infinité de nuages très divers les uns des autres (mais auxquels est associée une même et unique matrice de corrélation) et qui se déduisent simplement du nuage initial. En effet la perturbation se définit par

$$\tilde{x}_i - g = -\frac{1 + p_i}{1 - p_i}(x_i - g)$$

On peut l'appliquer à une seule u.s. mais aussi de manière itérative à une partie ou à la totalité des u.s. du nuage. On remarque alors qu'à chaque itération le centre de gravité (qui intervient dans la définition de la perturbation) se trouve modifié si bien que le nouveau tableau que l'on obtient dépend de l'ordre dans lequel on modifie successivement les u.s.

Il est encore possible d'appliquer de manière itérative la perturbation à une partie ou à la totalité des u.s. du nouveau tableau, ce qui conduit à un troisième tableau et l'on peut répéter le processus autant de fois que l'on veut. On remarque que les tableaux obtenus sont toujours différents des précédents car le centre de gravité n'est jamais le même. Cette diversité de tableaux se traduit par une diversité des représentations des u.s. dans un même plan factoriel auquel correspond donc toujours la même inertie.

En illustration nous avons généré à partir de la perturbation étudiée différents tableaux pour lesquels nous donnons la représentation des u.s. dans le plan principal.

La figure 1 correspond à la représentation des u.s. du tableau initial. La figure 2 visualise la perturbation de l'u.s. 20 tandis que la figure 3 correspond

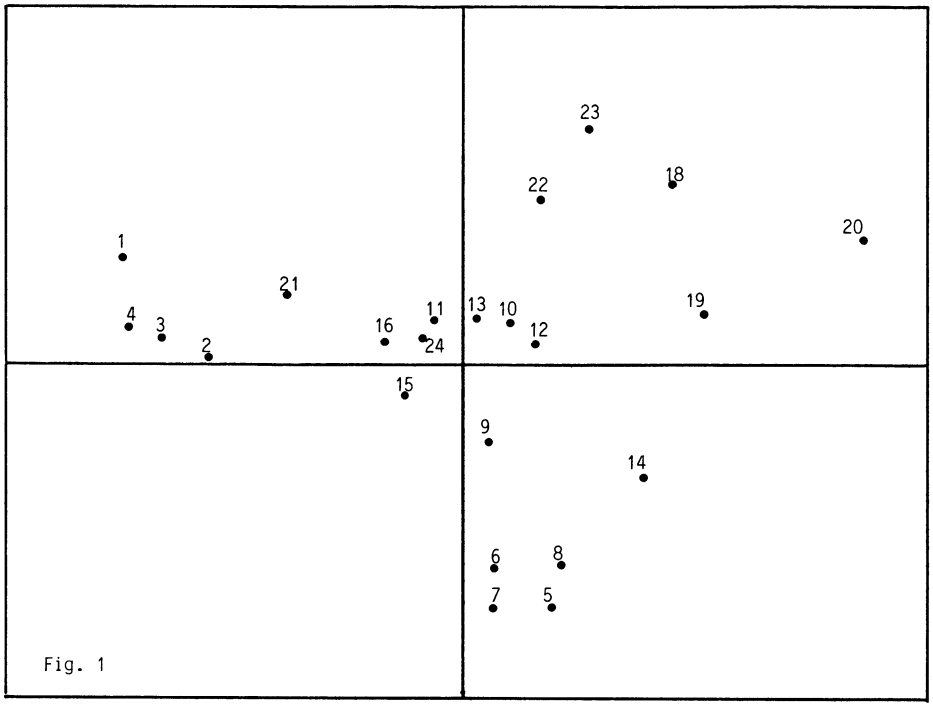
23
18
22
20
1
21
19
4
3
11
13
10
16
12
24
2
15
9
14
6
8
7
5

Fig. 1

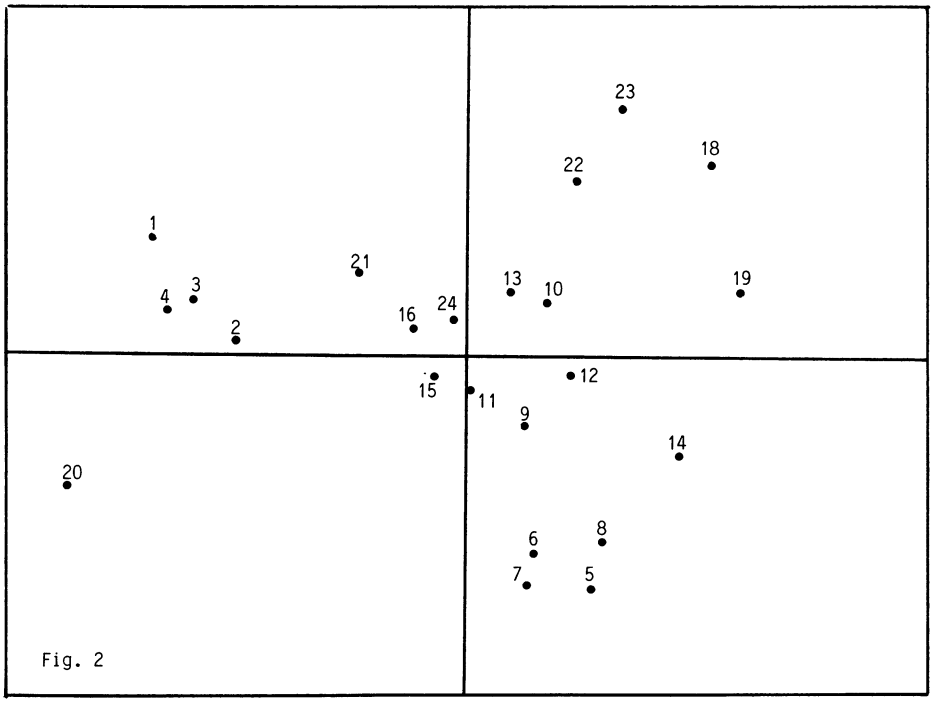
23
18
22
1
21
13
10
19
4
3
24
16
2
12
15
11
9
14
20
6
8
7
5

Fig. 2

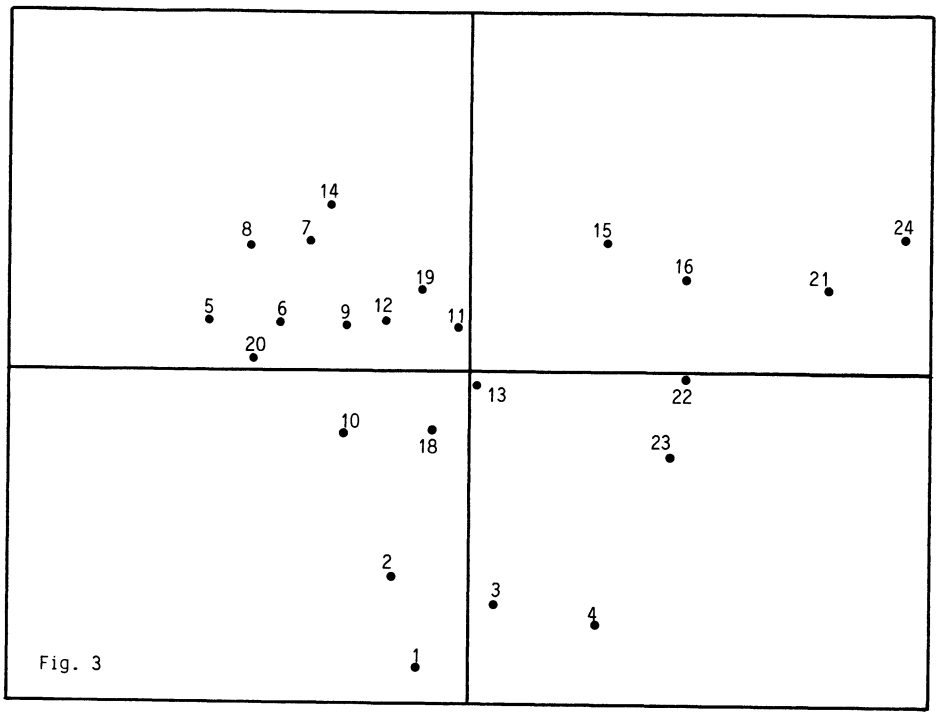
14
8
7
19
5
6
9
12
11
20
15
24
16
21
13
22
10
18
23
2
3
4
1

Fig. 3

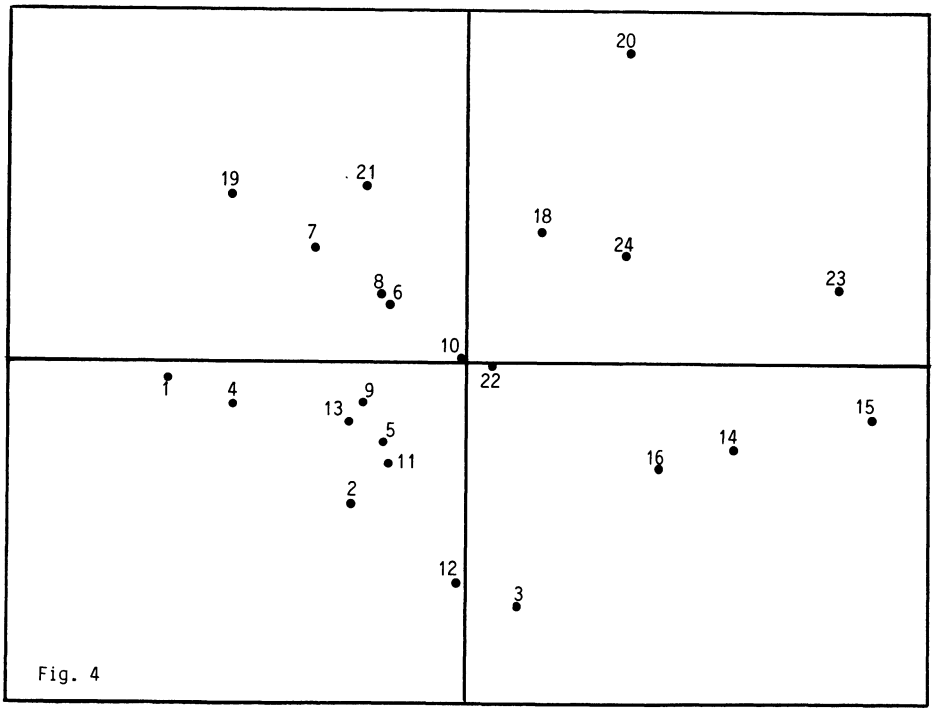
20
19
21
18
7
24
8
6
23
10
22
1
4
9
13
15
5
14
11
16
2
12
3

Fig. 4

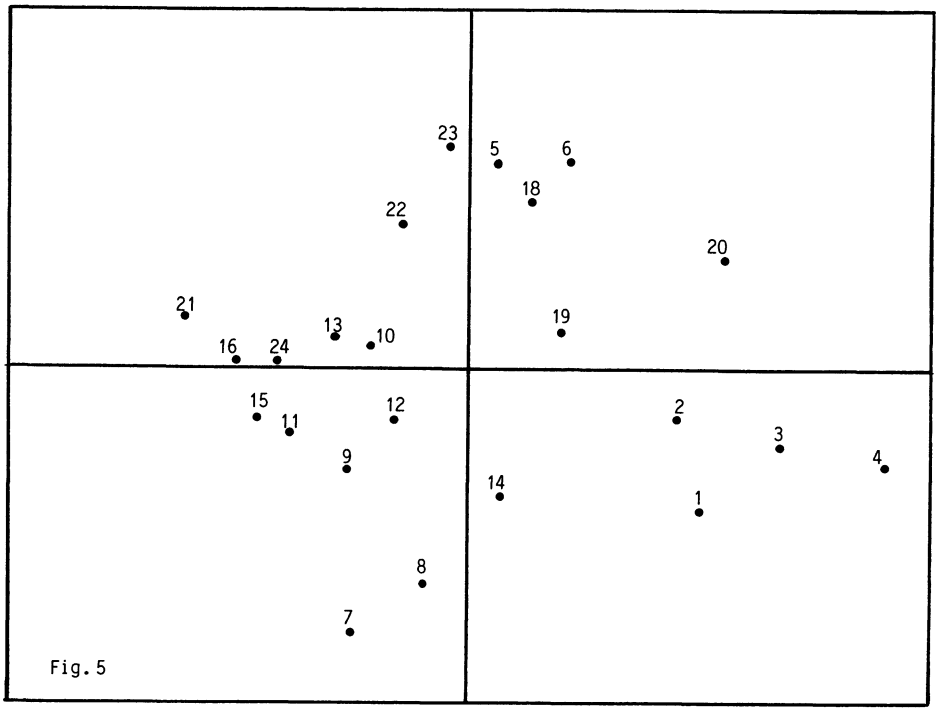
23
5
6
18
22
20
21
19
13
10
16
24
15
12
11
2
3
4
9
14
1
8
7

Fig. 5

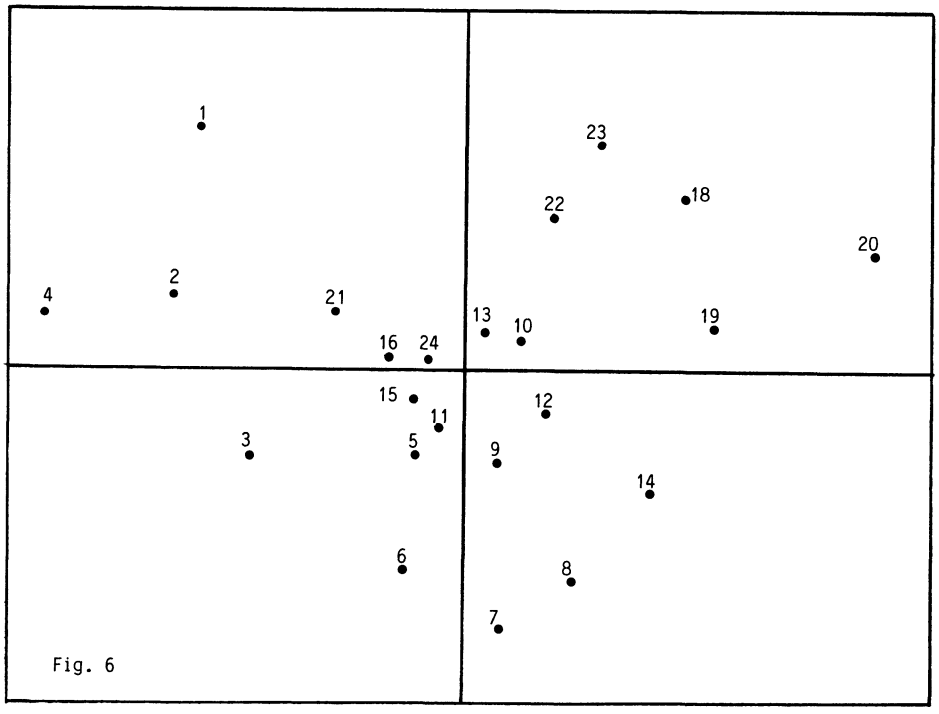
1
23
18
22
20
2
4
21
13
10
19
16
24
15
12
11
3
5
9
14
6
8
7

Fig. 6

à la perturbation itérative (dans l'ordre de 1 à 24) de la totalité des u.s. Si l'on applique la perturbation à la totalité des u.s. de ce nouveau tableau (toujours dans l'ordre de 1 à 24) et que l'on répète le processus un grand nombre de fois on obtient au terme de 18 itérations un tableau auquel correspond la représentation de la figure 4. La figure 5 correspond enfin à la perturbation dans l'ordre des six premières u.s. du tableau initial tandis que la figure 6 donne la représentation des u.s. lorsqu'on a répété 18 fois cette perturbation des six premières u.s. Les six figures donnent une idée de la diversité des tableaux auxquels correspondent toutefois la même matrice de corrélation et par conséquent les mêmes axes factoriels.

BIBLIOGRAPHIE

[1] T.W. ANDERSON, S. DASGUPTA. — "Some inequalities on characteristic roots of matrices", *Biometrika,* 1963, Vol. 50, p. 522-524.

[2] J. BENASSENI. — "Une contribution à l'étude de la stabilité en analyse factorielle", *Thèse de 3^{e} cycle,* U.S.T.L. Montpellier, 1984.

[3] J. BENASSENI. — "Influence des poids des unités statistiques sur les valeurs propres en analyse en composantes principales", *Revue de Statistique Appliquée,* Vol. XXXIII, n° 4, p. 41-55.

[4] B. ESCOFIER. — "Stabilité et approximation en analyse factorielle" *Thèse de Doctorat d'Etat,* Université P. et M. Curie, Paris VI, 1979.

[5] J.P. PAGES, F. CAILLIEZ. — Introduction à l'analyse des données, *SMASH,* 1976.

[6] J.H. WILKINSON. — *The algebraic eigenvalue problem,* Clarendon Press, Oxford, 1965.