Statistique et analyse des données

Jacques Benasseni

Stabilité du pouvoir discriminant des facteurs par rapport à des perturbations des données en analyse linéaire discriminante

Statistique et analyse des données, tome 10, n° 2 (1985), p. 1-28

<http://www.numdam.org/item?id=SAD_1985__10_2_1_0>

Statistiques et Analyse de Données
Septembre 1985 Vol. 10 n° 2 pp. 1-28

STABILITE DU POUVOIR DISCRIMINANT DES FACTEURS PAR RAPPORT A DES PERTURBATIONS DES DONNEES EN ANALYSE LINEAIRE DISCRIMINANTE

Jacques BENASSENI

Unité de Biométrie
INRA - ENSA - USTL
34060 MONTPELLIER CEDEX.

Résumé : *On étudie successivement les perturbations engendrées par la permutation des classes d'appartenance de deux unités statistiques, la suppression ou le transfert d'une classe à l'autre d'une unité statistique en analyse linéaire discriminante. Dans chacun des trois cas on s'intéresse au comportement du pouvoir discriminant des facteurs en proposant des bornes aux variations maximales que l'on peut observer pour les valeurs propres.*

Abstract : *This paper is devoted to the consequences of some perturbations of the data in discriminant analysis. Bounds to the variations of the eigenvalues, which characterize discriminant power of the factors, are suggested when two observations are exchanged from one class to an other and when an observation is deleted or put in another class.*

Mots clés : *Analyse linéaire discriminante, Stabilité des valeurs propres.*

Indices de classification STMA : *06-060 ; 00-050.*

Manuscrit reçu le 7 mai 1985
révisé le 16 janvier 1986

0 - INTRODUCTION

On considère un tableau X regroupant les mesures sur p variables d'un échantillon de base de n unités statistiques (u.s.) x_i considérées comme colonnes du tableau pour i = 1, ..., n. Un poids p_i étant attribué à chaque u.s. ($p_i > 0$, $\sum_{i=1}^{n} p_i = 1$), on associe aux variables du tableau la matrice de variance V pxp définie par $V = \sum_{i=1}^{n} p_i(x_i-g)(x_i-g)'$ avec $g = \sum_{i=1}^{n} p_i x_i$. On suppose que l'on dispose également d'une variable qualitative y à q modalités qui permet de répartir les u.s. en q classes C_j, j = 1, ..., q. Chaque classe C_j est caractérisée par son poids π_j, somme des poids des u.s. qui la composent, et par son centre de gravité $g_j = \frac{1}{\pi_j} \sum_{x_i \in C_j} p_i x_i$. On désigne par G la matrice pxq formée par les q centres de gravité g_j, j = 1, ..., q et l'on pose $\Delta = \text{diag } \pi_j$. Alors, si V est supposée inversible, on sait que l'analyse linéaire discriminante du tableau X par rapport à la variable qualitative y peut être envisagée comme l'analyse en composantes principales (A.C.P.) du triplet statistique (G, V^{-1}, Δ). Si l'on désigne par $B = \sum_{j=1}^{q} \pi_j (g_j-g)(g_j-g)'$ la matrice de variance interclasses, les valeurs propres de $B V^{-1}$ caractérisent le pouvoir discriminant des variables que fournit l'analyse.

Dans un contexte probabiliste un certain nombre d'auteurs se sont intéressés aux effets d'erreurs de classement dans l'échantillon de base sur l'efficacité de l'analyse discriminante. Ces études ont été en général traitées avec l'hypothèse d'une loi de distribution parente connue et dans le cas de deux classes seulement. Sous l'hypothèse de normalité de l'échantillon de base J. Mc LACHLAN [5] s'est ainsi intéressé au problème lorsqu'on suppose que les u.s. ont initialement la même probabilité d'être mal classées. Considérant ce modèle irréaliste P.A. LACHENBRUCH [4] a repris l'étude avec des modèles où les observations qui sont les plus proches de la moyenne de la mauvaise classe ont une probabilité plus grande que les autres d'être mal classées.

Plus récemment T.J. O'NEILL [6] s'est intéressé à des problèmes similaires par le biais du maximum de vraisemblance pour des lois de distributions générales de type exponentielle. La liste de ces travaux n'est bien évidemment pas exhaustive.

Pour notre part nous abordons les problèmes qui viennent d'être évoqués dans un contexte non probabiliste qui nous dispense d'avoir recours à des distributions théoriques sous jacentes aux données. Nous nous proposons simplement d'étudier d'un point de vue algébrique les fluctuations des valeurs propres pour des perturbations types de l'échantillon de base à savoir :

- le transfert d'une u.s. d'une classe à une autre.
- la permutation des classes d'appartenance de deux u.s.
- la suppression d'une u.s. de l'échantillon de base.

Notre objectif est de prolonger les résultats obtenus en [1], [2] et [3] en ce qui concerne la stabilité des valeurs propres dans l'A.C.P. classique ou l'A.F.C. et de montrer qu'une procédure mathématique simple est dans chacun des cas évoqués suffisante pour obtenir des bornes à la variation des valeurs propres.

Nous espérons ainsi fournir à l'utilisateur un outil lui permettant, sans avoir besoin de réaliser une nouvelle analyse, de savoir si certains types d'erreurs dans les données de base sont à même ou non d'influencer significativement les valeurs propres.

Nous commençons par présenter dans le paragraphe qui suit les résultats mathématiques élémentaires sur lesquels nous fonderons notre approche puis nous appliquons ces résultats à chacun des trois cas de perturbation qui viennent d'être évoqués. Une illustration pratique est proposée dans le dernier paragraphe à partir de l'exemple classique des Iris de Fisher.

1 - PRELIMINAIRES MATHEMATIQUES

Les valeurs propres d'une matrice carrée A d'ordre m seront notées $\lambda_i(A)$, i=1,...,m avec $\lambda_1(A) \geqslant \lambda_2(A) \geqslant \ldots \geqslant \lambda_m(A)$. Le produit scalaire au sens d'une métrique M de $\mathbb{R}^m$ sera noté $<,>_M$ et la norme associée $||\ ||_M$. On utilisera la notation plus simple <,> et $||\ ||$ pour la métrique identité. On a alors les résultats suivants :

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

$$(1)\quad \lambda_i(A+B) \leqslant \lambda_j(A) + \lambda_k(B)$$

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

$$(3)\quad \lambda_m(B) + \lambda_i(A) \leqslant \lambda_i(A+B) \leqslant \lambda_i(A) + \lambda_1(B)$$

On trouvera ce résultat en [7] par exemple. On notera que la relation (3) n'est qu'un cas particulier des relations (1) et (2).

Proposition 2 : Soient u et v deux vecteurs colonnes de $\mathbb{R}^m$ linéairement indépendants et r, s, t trois réels vérifiant $rt - s^2 \neq 0$. Alors la matrice $A = r\,u\,u' + s(uv' + vu') + t\,v\,v'$ est de rang deux et ses valeurs propres non nulles s'expriment sous la forme $\frac{1}{2}[\alpha \pm \sqrt{\alpha^2 - 4(rt-s^2)\beta}]$ avec $\alpha = r\,||u||^2 + 2\,s\,<u,v> + t\,||v||^2$ et $\beta = ||u||^2\,||v||^2 - <u,v>^2$.

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

Preuve : Tout vecteur colonne w orthogonal à u et v vérifie Aw = 0. Ainsi A est au plus de rang deux. Par conséquent tout vecteur propre associé à une valeur propre non nulle $\lambda(A)$ de A est de la forme $a\,u + b\,v$, $(a,b) \in \mathbb{R}^2$, et l'on a :

$$A(a\,u + b\,v) = (\lambda(A))\,(a\,u + b\,v)$$

En développant cette expression il est immédiat de voir que $\lambda(A)$ peut être considérée comme valeur propre de la matrice 2x2 suivante :

$$\begin{bmatrix} r\,||u||^2 + s\,\langle u,v\rangle & s\,||v||^2 + r\,\langle u,v\rangle \\ s\,||u||^2 + t\,\langle u,v\rangle & t\,||v||^2 + s\,\langle u,v\rangle \end{bmatrix}$$

car les vecteurs u et v ont été supposés linéairement indépendants. Cette dernière matrice est de rang deux car $rt - s^2 \neq 0$ et les valeurs propres s'obtiennent sans difficulté.

Corollaire 3 : Si $A = uu' - vv'$ on a :

$$\lambda_1(A) = \frac{1}{2}\left[||u||^2 - ||v||^2 + ||u+v||\ ||u-v||\right]$$

$$\lambda_m(A) = \frac{1}{2}\left[||u||^2 - ||v||^2 - ||u+v||\ ||u-v||\right]$$

2 - TRANSFERT D'UNE UNITE STATISTIQUE D'UNE CLASSE DANS UNE AUTRE

Dans ce paragraphe on étudie le cas où l'u.s. k appartenant initialement à la classe C_i est transférée dans la classe C_j. Cette modification des données se traduit par la transformation du triplet statistique initial (G, V^{-1}, Δ) en un nouveau triplet $(\tilde{G}, V^{-1}, \tilde{\Delta})$. Pour étudier les variations des valeurs propres nous commençons par expliciter $\tilde{G}$ et $\tilde{\Delta}$ par rapport à leurs homologues initiaux G et Δ.

Le nouveau tableau pxq $\tilde{G}$ des centres de gravité des classes se définit par $\tilde{G} = (\tilde{g}_1, \ldots, \tilde{g}_q)$ avec :

$$\tilde{g}_\ell = g_\ell \text{ si } \ell \neq i \text{ et } \ell \neq j$$

$$\tilde{g}_i = \frac{1}{\pi_i - p_k}\left[\left(\sum_{x_\ell \in C_i} p_\ell\, x_\ell\right) - p_k x_k\right] = \frac{1}{\pi_i - p_k}\left[\pi_i g_i - p_k x_k\right]$$

$$\tilde{g}_j = \frac{1}{\pi_j + p_k}\left[\left(\sum_{x_\ell \in C_j} p_\ell\, x_\ell\right) + p_k x_k\right] = \frac{1}{\pi_j + p_k}\left[\pi_j g_j + p_k x_k\right]$$

La nouvelle matrice $\tilde{\Delta} = \text{diag } \pi_j$ des poids attribués aux centres de gravité $\tilde{g}_\ell$, $\ell = 1, \ldots, q$, vérifie :

$$\tilde{\pi}_\ell = \pi_\ell \quad \text{si } \ell \neq i \quad \text{et} \quad \ell \neq j \quad , \quad \tilde{\pi}_i = \pi_i - p_k \quad , \quad \tilde{\pi}_j = \pi_j + p_k$$

Il est immédiat de constater que g qui est le centre de gravité des g_ℓ pour les poids π_ℓ reste le centre de gravité des $\tilde{g}_\ell$ pour les poids $\tilde{\pi}_\ell$ $\ell = 1, \ldots, q$. On a alors :

$$\tilde{B} = \sum_{\ell=1}^{q} \tilde{\pi}_\ell \, (\tilde{g}_\ell - g) \, (\tilde{g}_\ell - g)'$$

$$= \sum_{\substack{\ell=1 \\ \ell \neq i,j}}^{q} \pi_\ell (g_\ell - g)(g_\ell - g)' + \tilde{\pi}_i (\tilde{g}_i - g)(\tilde{g}_i - g)' + \tilde{\pi}_j (\tilde{g}_j - g)(\tilde{g}_j - g)'$$

A partir des relations qui viennent d'être explicitées comme liant $\tilde{\pi}_i$, $\tilde{\pi}_j$, $\tilde{g}_i$, $\tilde{g}_j$ à leurs homologues initiaux on obtient par un calcul simple l'expression :

$$\tilde{B} = B + [p_k/(\pi_i - p_k)] \, M_i + [p_k/(\pi_j + p_k)] \, M_j \qquad (2.1)$$

avec

$$M_i = \pi_i (g_i - g)(g_i - g)' - \pi_i [(g_i - g)(x_k - g)' + (x_k - g)(g_i - g)'] + p_k (x_k - g)(x_k - g)'$$

$$M_j = -\pi_j (g_j - g)(g_j - g)' + \pi_j [(g_j - g)(x_k - g)' + (x_k - g)(g_j - g)'] + p_k (x_k - g)(x_k - g)'$$

On voit donc en décomposant V^{-1} sous la forme $V^{-1} = S \, S'$ que la matrice $\tilde{B} \, V^{-1}$ a les mêmes valeurs propres que la matrice symétrique :

$$S' \, \tilde{B} \, S = S' \, B \, S + [p_k/(\pi_i - p_k)] \; S' \, M_i \, S + [p_k/(\pi_j + p_k)] \, S' \, M_j \, S \qquad (2.2)$$

Comme on a $\pi_i\ p_k - \pi_i^2 < 0$ (car $p_k < \pi_i$) et $-\pi_j p_k - \pi_j^2 < 0$ la proposition 2 permet de voir que les matrices $S'M_i\ S$ et $S'\ M_j\ S$ sont chacune de rang deux avec leurs deux valeurs propres non nulles de signes opposés. D'après la proposition l'expression des valeurs propres fait intervenir des termes de la forme $||S'u||^2$, $||S'v||^2$ et $\langle S'u, S'v\rangle$ (avec $u = S'(g_\ell - g)$ pour $\ell = i$ ou j et $v = S'(x_k - g)$) qui ne nécessitent cependant pas la connaissance de S puisqu'on a :

$$||S'u||^2 = u'\ S\ S'\ u = u'\ V^{-1}\ u = ||u||^2_V{}^{-1}$$

$$||S'v||^2 = v'\ S\ S'\ v = v'\ V^{-1}\ v = ||v||^2_V{}^{-1}$$

$$\langle S'u, S'v\rangle = u\ S\ S'\ v = u'\ V^{-1}\ v = \langle u, V^{-1}\ v\rangle = \langle u, v\rangle_V{}^{-1}$$

En appliquant la proposition 1 à l'expression (2.2) on obtient une famille d'encadrements des valeurs propres de $\tilde{B}\ V^{-1}$ avec pour $r+s+t \leqslant m+2$ et $m = 1, \ldots, p$

$$\lambda_m(\tilde{B}\ V^{-1}) \leqslant \lambda_r(B\ V^{-1}) + \frac{p_k}{\pi_i - p_k}\lambda_s(S'M_i\ S) + \frac{p_k}{\pi_j + p_k}\lambda_t(S'M_j\ S)$$

$$\lambda_{p-m+1}(\tilde{B}\ V^{-1}) \geqslant \lambda_{m-r+1}(B\ V^{-1}) + \frac{p_k}{\pi_i - p_k}\lambda_{p-s+1}(S'M_iS) + \frac{p_k}{\pi_j + p_k}\lambda_{p-t+1}(S'M_jS)$$

On notera qu'en pratique les inégalités les plus intéressantes sont pour $m=1, \ldots, p$ les suivantes :

$$\lambda_m(\tilde{B}\ V^{-1}) \leqslant \lambda_m(B\ V^{-1}) + \frac{p_k}{\pi_i - p_k}\lambda_1(S'M_iS) + \frac{p_k}{\pi_i + p_k}\lambda_1(S'M_jS)$$

$$\lambda_m(\tilde{B}\ V^{-1}) \geqslant \lambda_m(B\ V^{-1}) + \frac{p_k}{\pi_i - p_k}\lambda_p(S'M_iS) + \frac{p_k}{\pi_i + p_k}\lambda_p(S'M_jS)$$

On notera encore : si $p \geq 3$ pour $m = 2, \ldots, p$

$$\lambda_m(\tilde{B}\, V^{-1}) \leq \lambda_{m-1}(B\, V^{-1}) + \min\left[\frac{p_k}{\pi_i - p_k}\lambda_1\,(S'M_iS),\ \frac{p_k}{\pi_i + p_k}\lambda_1(S'M_jS)\right]$$

si $p \geq 3$ pour $m = 1, \ldots, p-1$

$$\lambda_m(\tilde{B}\, V^{-1}) \geq \lambda_{m+1}(B\, V^{-1}) + \max\left[\frac{p_k}{\pi_i - p_k}\lambda_p(S'M_iS),\ \frac{p_k}{\pi_i + p_k}\lambda_p(S'M_jS)\right]$$

et enfin : pour $m = 3, \ldots, p$, si $p \geq 3$ $\quad \lambda_m(\tilde{B}\, V^{-1}) \leq \lambda_{m-2}(B\, V^{-1})$

pour $m = 1, \ldots, p-2$, si $p \geq 3$ $\quad \lambda_m(\tilde{B}\, V^{-1}) \geq \lambda_{m+2}(B\, V^{-1})$

Remarques : 1. On notera l'importance du poids p_k de x_k par rapport au poids total de sa classe d'origine C_i puisque si p_k est très proche de π_i, le terme $p_k/(\pi_i - p_k)$ peut devenir très grand. Les encadrements autorisent alors des variations importantes des valeurs propres ce qui paraît intuitivement tout à fait logique.

2. Il est possible de donner une autre formulation au problème en écrivant :

$$S'\tilde{B}\, S = S'B\, S + S'N_i\, S + S'N_jS \qquad (2.3)$$

avec :

$$N_i = \tilde{\pi}_i(\tilde{g}_i - g)(\tilde{g}_i - g)' - \pi_i(g_i - g)(g_i - g)'$$

$$N_j = \tilde{\pi}_j(\tilde{g}_j - g)(\tilde{g}_j - g)' - \pi_j(g_j - g)(g_j - g)'$$

Alors en posant pour $\ell = i, j$

$$a_\ell = \tilde{\pi}_\ell\, ||\tilde{g}_\ell - g||^2_{V^{-1}} - \pi_\ell\, ||g_\ell - g||^2_{V^{-1}}$$

$$b_\ell = ||\sqrt{\tilde{\pi}_\ell}\,(\tilde{g}_\ell - g) + \sqrt{\pi_\ell}\,(g_\ell - g)||_{V^{-1}}\ ||\sqrt{\tilde{\pi}_\ell}\,(\tilde{g}_\ell - g) - \sqrt{\pi_\ell}\,(g_\ell - g)||_{V^{-1}}$$

Le corollaire 3 permet de voir que :

$$\lambda_1(S'N_\ell S) = \frac{1}{2}(a_\ell+b_\ell)\ ,\ \lambda_p(S'N_\ell S) = \frac{1}{2}(a_\ell-b_\ell)$$

et $\quad \lambda_r(S'N_\ell S) = 0 \qquad r = 2,\ldots,p-1$

En appliquant la proposition 1 à la relation (2.3) on obtient alors à nouveau une famille d'encadrements. Ces derniers, à l'inverse des précédents qui ne faisaient intervenir que des paramètres du triplet initial (G, V^{-1}, Δ) nécessitent le calcul de $\tilde{g}_i$, $\tilde{g}_j$, $\tilde{\pi}_i$ et $\tilde{\pi}_j$.

3 - PERMUTATION DES CLASSES D'APPARTENANCE DE DEUX UNITES STATISTIQUES

Soient x_ℓ une u.s. de la classe C_i et x_k une u.s. de la classe C_j avec $i \neq j$. On se propose ici d'étudier le comportement des valeurs propres lorsqu'on transfère l'u.s. x_ℓ dans la classe C_j et l'u.s. x_k dans la classe C_i. La démarche adoptée est analogue à celle du paragraphe précédent. La matrice de variance V n'étant pas affectée par la modification envisagée, le triplet statistique (G, V^{-1}, Δ) se trouve transformé en un triplet $(\tilde{G}, V^{-1}, \tilde{\Delta})$ dont nous allons exprimer les éléments $\tilde{G}$ et $\tilde{\Delta}$ en fonction des paramètres initiaux.

La transformation de $\Delta = \text{diag}\ \pi_r$ en $\tilde{\Delta} = \text{diag}\ \tilde{\pi}_r$ se caractérise pour $r = 1, \ldots, q$ de la manière suivante :

$$\tilde{\pi}_r = \pi_r \quad \text{si} \quad r \neq i \text{ et } r \neq j\ ,\ \tilde{\pi}_i = \pi_i + (p_k-p_\ell),\ \tilde{\pi}_j = \pi_j-(p_k-p_\ell)$$

Le nouveau tableau pxq $\tilde{G}$ des centres de gravité des classes se définit par $\tilde{G} = (\tilde{g}_1, \ldots, \tilde{g}_q)$ avec pour $r = 1, \ldots, q$:

$$\tilde{g}_r = g_r \quad \text{si } r \neq i \text{ et } r \neq j$$

$$\tilde{g}_i = \frac{1}{\tilde{\pi}_i}\left[\left(\sum_{x_s \in C_i} p_s\, x_s\right) - p_\ell\, x_\ell + p_k\, x_k\right]$$

$$\tilde{g}_i = \frac{1}{\pi_i+(p_k-p_\ell)}[\pi_i g_i - p_\ell x_\ell + p_k x_k]$$

$$\tilde{g}_j = \frac{1}{\tilde{\pi}_j}[(\sum_{x_s \in G C_j} p_s x_s) - p_k x_k + p_\ell x_\ell]$$

$$= \frac{1}{\pi_j-(p_k-p_\ell)}[\pi_j g_j - p_k x_k + p_\ell x_\ell]$$

La nouvelle matrice $\tilde{B}$ de variance interclasses s'exprime comme il suit :

$$\tilde{B} = \sum_{r=1}^{q} \tilde{\pi}_r(\tilde{g}_r-g)(\tilde{g}_r-g)'$$

$$= B + \tilde{\pi}_i(\tilde{g}_i-g)(\tilde{g}_i-g)' - \pi_i(g_i-g)(g_i-g)' + \tilde{\pi}_j(\tilde{g}_j-g)(\tilde{g}_j-g)' - \pi_j(g_j-g)(g_j-g)'$$

En posant $y = p_k(x_k-g) - p_\ell(x_\ell-g)$ et en utilisant les expressions de $\tilde{\pi}_i$, $\tilde{g}_i$ et $\tilde{\pi}_j$, $\tilde{g}_j$ qui viennent d'être données on obtient facilement l'expression suivante pour $\tilde{B}$:

$$\tilde{B} = B + \frac{1}{\pi_i+(p_k-p_\ell)} M_i + \frac{1}{\pi_j-(p_k-p_\ell)} M_j$$

avec :

$$M_i = yy' + \pi_i[(g_i-g)y' + y(g_i-g)'] - \pi_i(p_k-p_\ell)(g_i-g)(g_i-g)'$$

$$M_j = yy' - \pi_j[(g_j-g)y' + y(g_j-g)'] + \pi_j(p_k-p_\ell)(g_j-g)(g_j-g)'$$

Afin de pouvoir ensuite utiliser la proposition 1 on raisonne sur la matrice symétrique $S'\tilde{B}S$ (avec $V^{-1} = SS'$) qui a les mêmes valeurs propres que $\tilde{B}\,V^{-1}$ et l'on a :

$$S'\tilde{B}S = S'BS + \frac{1}{\pi_i+(p_k-p_\ell)} S'M_iS + \frac{1}{\pi_j-(p_k-p_\ell)} S'M_jS \qquad (3.1)$$

La proposition 2 donne tous les renseignements nécessaires à la connaissance des valeurs propres des matrices $S'M_rS$ pour $r \neq i,j$ lorsqu'on l'applique avec $u = S'y$ et $v = S'(g_r-g)$. En ce qui concerne $S'M_iS$ on constate d'après la proposition 2 que les deux valeurs propres non nulles sont de signe opposé puisque :

$$-\pi_i(p_k-p_\ell) - \pi_i^2 = -\pi_i\,[(\pi_i-p_\ell) + p_k] < 0 \qquad ((\pi_i-p_\ell) > 0)$$

Il en est de même pour $S'M_jS$ puisque :

$$\pi_j(p_k-p_\ell) - \pi_j^2 = -\pi_j\,[(\pi_j-p_k) + p_\ell] < 0 \qquad ((\pi_j-p_k) > 0)$$

D'autre part comme dans le paragraphe précédent on peut remarquer que les expressions des valeurs propres de ces deux matrices, telles qu'elles sont données par la proposition 2, ne font pas intervenir S mais seulement V^{-1} au travers du produit scalaire $<,>_{V^{-1}}$ et de la norme $||\ ||_{V^{-1}}$ qui lui est associée.

Par application de la proposition 1, on obtient une famille d'encadrements analogues dans leur forme à ceux du paragraphe précédent et que nous ne détaillons donc pas.

On notera qu'en pratique les poids des u.s. sont généralement pris tous égaux à 1/n. On a alors $p_k = p_\ell$, Δ est alors invariant ($\tilde{\Delta} = \Delta$) et il en résulte des simplifications dans les expressions notamment pour M_i et M_j dont le troisième terme disparaît.

4 - SUPPRESSION D'UNE UNITE STATISTIQUE

On étudie ici les conséquences de la suppression d'une u.s. x_ℓ qui appartient initialement à la classe C_i. D'un point de vue mathématique ce cas est plus complexe à étudier que les deux précédents puisque la matrice de variance V est affectée par la perturbation, si bien que le triplet (G, V^{-1}, Δ) se trouve transformé en un triplet $(\tilde{G}, \tilde{V}^{-1}, \tilde{\Delta})$. Comme dans les deux paragraphes qui précèdent nous commençons par exprimer les éléments du nouveau triplet en fonction de ceux du triplet initial.

On note tout d'abord que les poids des u.s. différentes de x_ℓ doivent être normalisés par le coefficient $(1 - p_\ell)^{-1}$ de manière à ce que leur somme reste toujours égale à l'unité. Après suppression de l'u.s. x_ℓ le poids initial p_k d'une u.s. x_k avec $k \neq \ell$ se trouve donc transformé en $\tilde{p}_k = p_k/(1-p_\ell)$ et l'on a bien

$$\sum_{\substack{k=1 \\ k\neq\ell}}^{n} \tilde{p}_k = 1$$

Il est alors facile de caractériser la transformation de $\Delta = \text{diag } \pi_k$ en $\tilde{\Delta} = \text{diag } \tilde{\pi}_k$. On a pour $k=1,\ldots,q \quad k\neq i$

$$\tilde{\pi}_k = \sum_{x_j \in C_k} \tilde{p}_j = (1-p_\ell)^{-1} \sum_{x_j \in C_k} p_j = \pi_k/(1-p_\ell)$$

et

$$\tilde{\pi}_i = \sum_{\substack{x_j \in C_i \\ j \neq \ell}} \tilde{p}_j = (1 - p_\ell)^{-1} \sum_{\substack{x_j \in C_j \\ j \neq \ell}} p_j = \frac{\pi_i - p_\ell}{1 - p_\ell}$$

Le tableau G des moyennes se trouve transformé en un tableau $\tilde{G} = (\tilde{g}_1, \ldots, \tilde{g}_q)$ qui se définit de la manière suivante :

pour $k = 1, \ldots, q \quad k \neq i$

$$\tilde{g}_k = \frac{1}{\tilde{\pi}_k} \sum_{x_j \in C_k} \tilde{p}_j x_j = \frac{1-p_\ell}{\pi_k} \sum_{x_j \in C_k} \frac{p_j}{1-p_\ell} x_j = \frac{1}{\pi_k} \sum_{x_j \in C_k} p_j x_j = g_k$$

et

$$\tilde{g}_i = \frac{1}{\tilde{\pi}_i} \sum_{\substack{x_j \in C_i \\ j \neq \ell}} \tilde{p}_j x_j = \frac{1-p_\ell}{\pi_i - p_\ell} \sum_{\substack{x_j \in C_i \\ j \neq \ell}} \frac{p_j}{1-p_\ell} x_j = \frac{1}{\pi_i - p_\ell} \left[\sum_{x_j \in C_i} p_j x_j - p_\ell x_\ell \right]$$

$$= \frac{1}{\pi_i - p_\ell} [\pi_i g_i - p_\ell x_\ell]$$

La matrice de variance se trouve transformée pour sa part en $\tilde{V}$ définie par :

$$\tilde{V} = \sum_{\substack{k=1 \\ k\neq\ell}}^{n} \tilde{p}_k (x_k-\tilde{g})(x_k-\tilde{g})' \quad \text{avec} \quad \tilde{g} = \sum_{\substack{k=1 \\ k\neq\ell}}^{n} \tilde{p}_k x_k = \sum_{k=1}^{q} \tilde{\pi}_k \tilde{g}_k$$

et l'on peut écrire en utilisant le théorème d'Huygens :

$$\tilde{V} = \frac{1}{1-p_\ell} V - \frac{p_\ell}{1-p_\ell} (x_\ell-g)(x_\ell-g)' - (g-\tilde{g})(g-\tilde{g})'$$

En remarquant que $g-\tilde{g} = \frac{p_\ell}{1-p_\ell} (x_\ell-g)$ on obtient alors facilement :

$$\tilde{V} = \frac{1}{1-p_\ell} [V - \frac{p_\ell}{1-p_\ell} (x_\ell-g)(x_\ell-g)']$$

On suppose qu'après suppression de l'u.s. x_ℓ la matrice de variance reste inversible ce qui sera vérifié si $p_\ell(x_\ell-g)' V^{-1}(x_\ell-g) \neq 1-p_\ell$ puisqu'on peut alors écrire :

$$\tilde{V}^{-1} = (1-p_\ell)[V^{-1} + c_\ell V^{-1}(x_\ell-g)(x_\ell-g)' V^{-1}] \qquad (4.1)$$

où c_ℓ désigne le coefficient défini par :

$$c_\ell = p_\ell / [1-p_\ell-p_\ell(x_\ell-g)' V^{-1}(x_\ell-g)]$$

Après avoir exprimé les éléments $\tilde{G}$, $\tilde{V}^{-1}$, $\tilde{\Delta}$ du nouveau triplet en fonction des paramètres initiaux nous allons étudier le comportement de la nouvelle matrice de variance interclasses

$$\tilde{B} = \sum_{j=1}^{q} \tilde{\pi}_j (\tilde{g}_j-\tilde{g})(\tilde{g}_j-\tilde{g})'$$

Par une nouvelle application du théorème d'Huygens on obtient :

$$\tilde{B} = \sum_{j=1}^{q} \tilde{\pi}_j (\tilde{g}_j-g)(\tilde{g}_j-g)' - (g-\tilde{g})(g-\tilde{g})'$$

soit compte tenu de l'expression des $\tilde{\pi}_j$ et des $\tilde{g}_j$ $\quad j=1,\ldots,q$

$$\tilde{B} = \frac{1}{1-p_\ell} B + \frac{\pi_i-p_\ell}{1-p_\ell} (\tilde{g}_i-g)(\tilde{g}_i-g)' - \frac{\pi_i}{1-p_\ell} (g_i-g)(g_i-g)'-(g-\tilde{g})(g-\tilde{g})' \qquad (4.2)$$

Considérons $A = \frac{\pi_i-p_\ell}{1-p_\ell} (\tilde{g}_i-g)(\tilde{g}_i-g)' - (g-\tilde{g})(g-\tilde{g})'$

On obtient sans difficultés $(g-\tilde{g})(g-\tilde{g})' = (\frac{p_\ell}{1-p_\ell})^2 (x_\ell-g)(x_\ell-g)'$ et en utilisant l'égalité $\tilde{g}_i = \frac{1}{\pi_i-p_\ell}[\pi_i g_i - p_\ell x_\ell]$ on peut écrire $(\tilde{g}_i-g)(\tilde{g}_i-g)'$ sous la forme :

$$\frac{\pi_i^2}{(\pi_i-p_\ell)^2} (g_i-g)(g_i-g)' - \frac{\pi_i p_\ell}{(\pi_i-p_\ell)^2} [(g_i-g)(x_\ell-g)' + (x_\ell-g)(g_i-g)']$$

$$+ \frac{p_\ell^2}{(\pi_i-p_\ell)^2} (x_\ell-g)(x_\ell-g)'$$

On obtient alors pour A l'expression

$$\frac{\pi_i^2}{(\pi_i-p_\ell)(1-p_\ell)} (g_i-g)(g_i-g)' - \frac{\pi_i p_\ell}{(\pi_i-p_\ell)(1-p_\ell)} [(g_i-g)(x_\ell-g)'+(x_\ell-g)(g_i-g)']$$

$$+ \frac{p_\ell^2(1-\pi_i)}{(\pi_i-p_\ell)(1-p_\ell)^2} (x_\ell-g)(x_\ell-g)'$$

En utilisant la décomposition $\frac{\pi_i^2}{(\pi_i-p_\ell)(1-p_\ell)^2} = \frac{\pi_i}{1-p_\ell} + \frac{\pi_i p_\ell}{(\pi_i-p_\ell)(1-p_\ell)}$ et en revenant à l'expression (4.2) on obtient en définitive :

$$\tilde{B} = \frac{1}{1-p_\ell} [B + \frac{1}{\pi_i-p_\ell} D] \qquad (4.3)$$

avec :

$$D = \pi_i p_\ell\ (g_i-g)(g_i-g)' - \pi_i p_\ell\ [(g_i-g)(x_\ell-g)'+(x_\ell-g)(g_i-g)']$$

$$+ \frac{p_\ell^2\ (1-\pi_i)}{1 - p_\ell}\ (x_\ell-g)(x_\ell-g)'$$

En décomposant $\tilde{V}^{-1}$ sous la forme $\tilde{V}^{-1} = \tilde{R}\ \tilde{R}'$ on voit que $\tilde{B}\ \tilde{V}^{-1}$ a les mêmes valeurs propres que $\tilde{R}'\ \tilde{B}\ \tilde{R}$ qui est symétrique et s'écrit d'après (4.3) sous la forme :

$$\tilde{R}'\ \tilde{B}\ \tilde{R} = \frac{1}{1-p_\ell}\ [\tilde{R}'\ B\ \tilde{R} + \frac{1}{\pi_i - p_\ell}\ \tilde{R}'D\ \tilde{R}\ ] \qquad (4.4)$$

L'application de la proposition 1 à la matrice symétrique $\tilde{R}'\tilde{B}\ \tilde{R}$ nécessite la connaissance des valeurs propres de $\tilde{R}'\ B\ \tilde{R}$ et $\tilde{R}'\ D\ \tilde{R}$ que nous allons successivement étudier.

Etudions tout d'abord les valeurs propres de $\tilde{R}'\ B\ \tilde{R}$ qui sont identiques à celles de $B\tilde{R}\tilde{R}' = B\tilde{V}^{-1} = (1-p_\ell)\ [\ BV^{-1} + c_\ell\ BV^{-1}(x_\ell-g)(x_\ell-g)'V^{-1}]$

Décomposons à son tour B sous la forme $B = SS'$. On voit que $B\ \tilde{V}^{-1}$ a les mêmes valeurs propres que la matrice symétrique :

$$S'\ \tilde{V}^{-1}\ S = (1-\ p_\ell)\ [S'\ V^{-1}S + c_\ell\ S'\ V^{-1}\ (x_\ell-g)(x_\ell-g)'\ V^{-1}S]$$

à laquelle on peut donc appliquer la proposition 1 ce qui donne pour r=1,...,p et t et s vérifiant $t+s \leqslant r+1$:

$$\lambda_r(\tilde{R}'B\tilde{R}) \leqslant (1-p_\ell)\ [\lambda_t(S'V^{-1}S)+c_\ell\ \lambda_s(S'V^{-1}(x_\ell-g)(x_\ell-g)'\ V^{-1}S)\ ]$$

et

$$\lambda_{p-r+1}(\tilde{R}'B\tilde{R}) \geqslant (1-p_\ell)\ [\ \lambda_{p-t+1}(S'V^{-1}S) + c_\ell\ \lambda_{p-s+1}(S'V^{-1}(x_\ell-g)(x_\ell-g)'V^{-1}S)]$$

compte tenu du fait que $\lambda_r(\tilde{R}'B\tilde{R}) = \lambda_r(B\tilde{V}^{-1}) = \lambda_r(S'\tilde{V}^{-1}S)$

Il est immédiat de constater que $S'V^{-1}(x_\ell-g)(x_\ell-g)'V^{-1}S$ est de rang un et que sa valeur propre non nulle est égale à $(x_\ell-g)'V^{-1}BV^{-1}(x_\ell-g)$.

En pratique les inégalités précédentes se ramènent alors aux expressions suivantes :

$$\lambda_r(BV^{-1}) \leqslant \frac{1}{1-p_\ell}\ \lambda_r(B\tilde{V}^{-1}) \leqslant \lambda_r(BV^{-1}) + c_\ell\ (x_\ell-g)'V^{-1}BV^{-1}\ (x_\ell-g) \quad (4.5)$$

et

$$\lambda_r(B\tilde{V}^{-1}) \leqslant (1-p_\ell)\ \lambda_{r-1}\ (BV^{-1}) \quad \text{pour} \quad r \geqslant 2$$

Examinons à présent les valeurs propres de $\tilde{R}'D\tilde{R}$. Elles se déduisent de la proposition 2 avec :

$$r = \pi_i p_\ell \ , \ s = -\pi_i p_\ell \ , \ t = \frac{p_\ell^2\ (1-\pi_i)}{1-p_\ell} \ , \ u = \tilde{R}'(g_i-g) \ , \ v = \tilde{R}'(x_\ell-g)$$

Comme $rt-s^2 = \frac{\pi_i p_\ell^2\ (p_\ell-\pi_i)}{1-p_\ell} < 0$ on en déduit que $\lambda_1(\tilde{R}'DR) > 0$ et $\lambda_p(\tilde{R}'D\tilde{R}) < 0$, les autres valeurs propres étant nulles. On remarquera qu'il est possible d'exprimer ces deux valeurs propres uniquement à partir des éléments du triplet initial (G, V^{-1}, Δ) et que les expressions ne nécessitent par la connaissance de $\tilde{R}$ mais font intervenir le produit scalaire $<,>_{\tilde{V}^{-1}}$ et la norme $||.||_{\tilde{V}^{-1}}$. On utilise alors la relation (4.1) entre $\tilde{V}^{-1}$ et V^{-1} pour exprimer $<,>_{\tilde{V}^{-1}}$ et $||.||_{\tilde{V}^{-1}}$.

Enfin l'application de la proposition 1 à la relation (4.4) permet d'obtenir pour $r+k \leqslant m+1$ $m = 1, \ldots, p$:

$$\lambda_m(\tilde{B}\ \tilde{V}^{-1}) \leqslant \frac{1}{1-p_\ell}[\lambda_r(B\ \tilde{V}^{-1}) + \frac{1}{\pi_i-p_\ell}\ \lambda_k(\tilde{R}'\ D\ \tilde{R})]$$

$$\lambda_{p-m+1}(\tilde{B}\ \tilde{V}^{-1}) \geqslant \frac{1}{1-p_\ell}[\lambda_{p-r+1}(B\ \tilde{V}^{-1}) + \frac{1}{\pi_i-p_\ell}\ \lambda_{p-k+1}(\tilde{R}'D\ \tilde{R})]$$

En pratique compte tenu de (4.5) on remarquera principalement les encadrements suivants : pour $m = 1, \ldots, p$

$$\lambda_m(\tilde{B}\ \tilde{V}^{-1}) \leqslant \lambda_m(B\ V^{-1}) + c_\ell\ (x_\ell-g)'\ V^{-1}B\ V^{-1}(x_\ell-g) + \frac{1}{(1-p_\ell)(\pi_i-p_\ell)}\ \lambda_1(\tilde{R}'D\tilde{R})$$

$$\lambda_m(\tilde{B}\ \tilde{V}^{-1}) \geqslant \lambda_m(B\ V^{-1}) + \frac{1}{(1-p_\ell)(\pi_i-p_\ell)}\ \lambda_p(\tilde{R}'D\tilde{R}) \quad (4.6)$$

obtenus pour $r = m$ et $k = 1$.

On notera aussi si $p \geq 3$

$$\text{pour } m \geq 2 \quad \lambda_m(\tilde{B}\tilde{V}^{-1}) \leq \lambda_{m-1}(BV^{-1}) + c_\ell\ (x_\ell - g)'\ V^{-1}\ B\ V^{-1}(x_\ell - g)$$

$$\text{pour } m \geq 3 \quad \lambda_m(\tilde{B}\tilde{V}^{-1}) \leq \lambda_{m-2}(BV^{-1}) \qquad (4.7)$$

$$\text{pour } m \leq p-1 \quad \lambda_m(\tilde{B}\tilde{V}^{-1}) \geq \lambda_{m+1}(BV^{-1})$$

Les inégalités (4.6) ont l'intérêt de donner un encadrement de la m-ième valeur propre de $\tilde{B}\ \tilde{V}^{-1}$ à partir de la m-ième valeur propre de $B\ V^{-1}$. Elles peuvent cependant être assez lourdes à manipuler à cause des expressions de $\lambda_1(\tilde{R}'D\tilde{R})$ et de $\lambda_p(\tilde{R}'D\tilde{R})$. Aussi lorsque $\lambda_m(BV^{-1})$ sera assez proche de $\lambda_{m-1}(BV^{-1})$ ou de $\lambda_{m+1}(BV^{-1})$ il pourra être plus intéressant d'utiliser les inégalités (4.7).

5 - ILLUSTRATION DES RESULTATS

Nous présentons, à partir de l'exemple classique des iris de Fisher, une illustration des encadrements des valeurs propres en ce qui concerne la permutation des classes d'appartenance de deux u.s. et la suppression d'une u.s.

A) Présentation des données

Les données sont constituées par un ensemble de 150 iris répartis en 3 groupes et pour lesquels on a mesuré quatre variables : longueur et largeur des sépales, longueur et largeur des pétales. Les trois groupes sont d'effectifs égaux (50 iris) et chacun correspond à une variété particulière : variété setosa pour le groupe 1, versicolor pour le groupe 2, virginica pour le groupe 3. Les iris de chacun des trois groupes sont repérés par un numéro compris entre 1 et 50. L'analyse linéaire discriminante a été réalisée en utilisant le programme MAHAL 3 de la bibliothèque ADDAD, elle a conduit aux deux valeurs propres non nulles : $\lambda_1(BV^{-1}) = 0.970$ et $\lambda_2(BV^{-1}) = 0.222$ respectivement associées aux vecteurs propres :

$$u_1 = (0.039,\ -0.027,\ 0.049,\ -0.0145)$$

et

$$u_2 = (0.083, -0.193, -0.253, -0.002)$$

Le tableau 1 donne les coordonnées de chaque iris sur les facteurs discriminants.

B) Permutation des classes d'appartenance de deux unités statistiques

Nous nous proposons ici d'illustrer les résultats du paragraphe 3. D'un point de vue pratique le poids de chaque iris est pris égal à 1/n (n=150) ce qui entraîne certaines simplifications dans la formulation des encadrements. Pour deux u.s. k et ℓ quelconques on a en effet : $p_k - p_\ell = 0$ et la relation (3.1) prend alors la forme :

$$S'\tilde{B}\,S = S'B\,S + \frac{1}{\pi_i}\,S'M_iS + \frac{1}{\pi_j}\,S'M_jS$$

avec : $y = \frac{1}{n}(x_k - x_\ell)$

$$M_i = yy' + \pi_i[(g_i-g)y' + y(g_i-g)']$$

$$M_j = yy' - \pi_j[(g_j-g)y' + y(g_j-g)']$$

On obtient donc en définitive :

$$S'\tilde{B}\,S = S'B\,S + (\frac{1}{\pi_i} + \frac{1}{\pi_j})\,S'yy'S + S'(g_i-g_j)y'S + S'y(g_i-g_j)'S$$

Il est commode de désigner par A la matrice

$$(\frac{1}{\pi_i} + \frac{1}{\pi_j})\,S'yy'S + S'(g_i-g_j)y'S + S'y(g_i-g_j)'S$$

On obtient alors les encadrements :

$$\lambda_i(BV^{-1}) + \lambda_p(A) \leq \lambda_i(\tilde{B}V^{-1}) \leq \lambda_i(BV^{-1}) + \lambda_1(A) \qquad (4.1)$$

TABLEAU 1

IRIS	GROUPE SETOSA		GROUPE VERSICOLOR		GROUPE VIRGINICA	
	FACTEUR 1	FACTEUR 2	FACTEUR 1	FACTEUR 2	FACTEUR 1	FACTEUR 2
1	- 0.976	- 5.815	0.795	- 5.074	1.536	- 7.543
2	- 1.153	- 6.717	0.798	- 6.644	1.264	- 7.710
3	- 0.816	- 5.259	0.625	- 5.960	1.240	- 6.666
4	- 1.091	- 6.138	0.592	- 4.853	1.335	- 5.903
5	- 0.900	- 5.692	0.420	- 4.332	1.189	- 5.490
6	- 1.048	- 6.615	0.505	- 4.457	1.277	- 5.611
7	- 0.818	- 5.070	0.823	- 5.779	1.135	- 5.827
8	- 0.834	- 6.872	0.408	- 4.523	1.512	- 6.199
9	- 0.881	- 5.234	0.712	- 5.622	1.335	- 5.903
10	- 1.057	- 6.393	0.527	- 5.128	1.245	- 6.102
11	- 0.924	- 6.439	0.695	- 6.219	1.204	- 4.014
12	- 1.291	- 7.561	0.982	- 4.433	1.469	- 7.404
13	- 0.781	- 5.030	0.771	- 5.638	1.309	- 6.511
14	- 1.141	- 6.294	0.560	- 4.486	1.479	- 4.702
15	- 0.847	- 5.306	0.684	- 6.366	1.561	- 6.452
16	- 0.828	- 5.449	0.883	- 5.959	1.038	- 5.095
17	- 1.026	- 5.788	0.580	- 6.010	1.561	- 7.236
18	- 0.959	- 5.217	0.315	- 4.760	1.401	- 8.024
19	- 1.044	- 6.202	0.652	- 5.516	1.570	- 8.099
20	- 0.827	- 6.347	0.721	- 5.128	1.190	- 6.231
21	- 1.231	- 7.026	0.738	- 6.996	1.387	- 7.726
22	- 0.869	- 5.344	0.578	- 3.588	1.372	- 7.087
23	- 0.981	- 7.237	0.618	- 5.444	1.295	- 5.897
24	- 0.944	- 5.698	0.572	- 3.579	1.084	- 6.401
25	- 0.979	- 5.927	0.816	- 5.514	1.561	- 6.704
26	- 0.918	- 5.091	1.021	- 6.865	1.060	- 6.482
27	- 0.895	- 6.252	1.158	- 5.148	1.257	- 5.911
28	- 0.727	- 6.326	0.544	- 5.500	1.528	- 7.495
29	- 1.032	- 6.012	0.563	- 4.374	1.670	- 5.780
30	- 0.755	- 7.046	0.790	- 4.514	1.359	- 7.429
31	- 0.817	- 5.485	0.854	- 5.964	1.284	- 7.243
32	- 0.781	- 5.588	0.576	- 5.456	1.698	- 5.061
33	- 1.358	- 7.348	0.676	- 5.203	1.265	- 4.037
34	- 0.891	- 5.613	0.475	- 4.204	1.085	- 6.100
35	- 1.237	- 8.374	0.745	- 6.014	1.513	- 8.113
36	- 0.964	- 5.924	0.804	- 4.206	1.173	- 5.171
37	- 0.825	- 5.337	1.041	- 4.682	1.472	- 6.351
38	- 1.019	- 6.536	0.286	- 5.222	1.352	- 5.631
39	- 0.916	- 4.979	0.435	- 6.012	1.977	- 5.268
40	- 1.047	- 6.798	0.621	- 5.815	1.461	- 5.300
41	- 0.961	- 7.011	0.986	- 6.078	1.477	- 5.071
42	- 1.103	- 6.511	0.765	- 4.664	1.281	- 6.955
43	- 1.136	- 7.569	0.641	- 5.406	1.571	- 8.356
44	- 0.624	- 4.211	0.793	- 5.852	1.361	- 6.677
45	- 1.037	- 6.951	0.712	- 4.906	1.148	- 7.135
46	- 1.020	- 6.121	0.800	- 5.101	1.744	- 7.841
47	- 0.995	- 6.455	0.601	- 5.790	1.570	- 6.674
48	- 0.897	- 5.837	0.589	- 5.448	1.325	- 5.750
49	- 0.945	- 5.767	0.836	- 5.362	1.414	- 5.851
50	- 1.089	- 6.509	0.766	- 4.521	1.554	- 7.361

avec : $\lambda_1(A) = \frac{1}{2}(\alpha + \sqrt{\alpha^2 + 4\beta})$

$\lambda_p(A) = \frac{1}{2}(\alpha - \sqrt{\alpha^2 + 4\beta})$

et

$$\alpha = (\frac{1}{\pi_i} + \frac{1}{\pi_j}) \; ||y||^2_{V^{-1}} + 2 \langle g_i - g_j, y \rangle_{V^{-1}}$$

$$\beta = ||y||^2_{V^{-1}} \; ||g_i - g_j||^2_{V^{-1}} - \langle g_i - g_j, y \rangle^2_{V^{-1}}$$

Pour tester la précision de ces encadrements nous avons procédé de la manière suivante : une procédure de tirage au hasard nous a permis de choisir deux groupes parmi les trois en présence puis une u.s. dans chacun de ces groupes. Nous avons alors étudié le comportement des inégalités (4.1) pour la permutation des deux u.s. retenues. Après avoir répété 200 fois la procédure il est apparu que les encadrements restent en général d'amplitude relativement faible ce qui traduit une bonne stabilité de l'analyse par rapport au type de perturbation mis en oeuvre sur les données. Le tableau 2 qui suit propose en illustration les résultats correspondant aux 25 premiers tirages au hasard. Pour la permutation de chaque couple d'u.s. il donne les bornes inférieures m_i et supérieure M_i proposées par les inégalités (4.1) en ce qui concerne la variation de la valeur propre i (i = 1,2).

Pour certains des 200 couples d'u.s. retenus par la procédure les encadrements apparaissent comme étant un peu moins précis. Il nous a semblé intéressant de vérifier si cette performance moins bonne dans certains cas était en rapport ou non avec une variation effective des valeurs propres. A cet effet nous avons calculé ces dernières de manière exacte pour certaines situations correspondant à des encadrements apparemment moins précis que l'ensemble. Pour la permutation de chaque couple d'u.s. le tableau 3 donne les nouvelles valeurs propres $\lambda_i(\tilde{B}\, V^{-1})$ (i=1,2) ainsi que les bornes inférieure m_i et supérieure M_i proposées par les inégalités (4.1).

TABLEAU 2

IDENTIFICATION DES U.S. PERMUTEES				m_1	M_1	m_2	M_2
u.s.	Groupe	u.s.	Groupe				
2	1	43	2	0.916	0.971	0.168	0.223
40	2	48	3	0.954	0.981	0.206	0.233
6	1	40	3	0.877	0.992	0.129	0.244
12	2	39	3	0.938	0.981	0.191	0.234
20	2	33	3	0.952	1.001	0.204	0.253
1	2	7	3	0.940	0.987	0.192	0.239
15	1	19	3	0.870	0.988	0.122	0.240
17	1	36	2	0.911	0.973	0.163	0.225
28	2	49	3	0.943	0.984	0.196	0.237
3	2	30	3	0.927	0.987	0.179	0.239
6	1	12	2	0.882	0.994	0.134	0.246
42	1	47	2	0.916	0.980	0.168	0.232
10	1	41	3	0.879	0.989	0.131	0.241
18	1	19	3	0.866	0.988	0.118	0.240
11	2	14	3	0.959	0.997	0.211	0.249
26	1	48	2	0.932	0.979	0.185	0.231
21	2	38	3	0.963	0.991	0.215	0.244
14	1	38	2	0.912	0.986	0.164	0.238
9	1	29	3	0.875	0.987	0.127	0.239
19	2	40	3	0.944	0.993	0.196	0.246
5	1	31	2	0.928	0.977	0.180	0.229
1	2	22	3	0.934	0.972	0.186	0.224
16	1	39	3	0.867	0.990	0.119	0.242
39	1	48	2	0.933	0.980	0.185	0.232
2	2	24	3	0.961	0.979	0.214	0.231

TABLEAU 3

Permutation n°	IDENTIFICATION DES U.S. PERMUTEES				$\lambda_1(\tilde{B}V^{-1})$	m_1	M_1	$\lambda_2(\tilde{B}V^{-1})$	m_2	M_2
	u.s.	Groupe	u.s.	Groupe						
1	13	1	43	3	0.902	0.863	0.996	0.209	0.115	0.248
2	12	1	5	2	0.936	0.881	0.990	0.187	0.133	0.243
3	26	2	33	3	0.968	0.958	1.025	0.267	0.210	0.277
4	35	1	36	2	0.930	0.867	0.988	0.176	0.119	0.240
5	33	1	32	3	0.879	0.865	0.992	0.230	0.117	0.244
6	35	1	29	3	0.883	0.869	0.993	0.231	0.121	0.245
7	2	1	43	3	0.888	0.874	0.978	0.216	0.126	0.230
8	11	2	33	3	0.965	0.950	1.020	0.258	0.203	0.272
9	14	1	50	3	0.891	0.858	0.999	0.218	0.110	0.251
10	2	1	31	3	0.899	0.863	1.005	0.220	0.115	0.257
11	2	2	11	3	0.966	0.963	1.014	0.263	0.215	0.266
12	33	1	38	3	0.890	0.864	1.003	0.228	0.116	0.255
13	49	1	50	3	0.898	0.859	1.003	0.216	0.111	0.255
14	33	1	46	3	0.879	0.848	0.995	0.220	0.100	0.252

La lecture de ce dernier tableau amène un certain nombre d'enseignements dont les principaux sont les suivants :

Il apparaît tout d'abord que la variation des valeurs propres consécutive à la permutation des classes d'appartenance de deux u.s. peut être plus importante qu'on ne pourrait l'imaginer compte tenu de l'effectif assez grand des u.s. Ainsi en est-il de la première valeur propre pour les permutations 5, 6, 7, 9, 12, 13 et de la deuxième valeur propre pour les permutations 2, 3, 4, 8 et 11. L'amplitude assez "importante" de certains encadrements se trouve donc justifiée par une variation en rapport de l'une ou l'autre des valeurs propres.

Les encadrements donnent une idée assez exacte des variations maximales qui peuvent être observées. Ainsi, sur les 200 répétitions de la procédure, la borne la plus faible a été obtenue avec la valeur 0.848 ce qui est relativement proche de la diminution maximale effectivement enregistrée pour la première valeur propre avec 0.879. De même la valeur extrême de M_2 = 0.277, que l'on note pour la permutation n° 3, donne une bonne idée de l'augmentation maximale de la deuxième valeur propre qui est enregistrée avec 0.267.

L'amplitude des encadrements semble également un bon indicateur pour orienter l'utilisateur vers des couples d'u.s. pour lesquelles l'appartenance aux classes initiales présente un caractère déterminant sur les valeurs propres. Dans cet esprit on note par exemple à partir des permutations n° 5 et 8 que le passage de l'u.s. 33 du groupe 3 au groupe 2 a tendance à diminuer le pouvoir discriminant du premier facteur tout en augmentant celui du deuxième. Une remarque analogue peut être formulée en ce qui concerne l'u.s. 2 du groupe 1 et la diminution de la première valeur propre.

A côté de ces aspects positifs on notera cependant un inconvénient lié à la structure théorique des encadrements qui est la même pour les deux valeurs propres. Il en résulte que pour une permutation donnée entraînant une variation sensible de l'une seulement des valeurs propres, ce sont les encadrements correspondant aux deux valeurs propres qui réagissent simultanément ; l'un seulement correspond alors à la réalité.

Pour expliciter ce point de vue considérons par exemple la permutation 11. D'après la relation (4.1), $\lambda_1(A)$ se doit d'être suffisamment grande pour répondre à l'augmentation de la deuxième valeur propre (et l'on obtient une borne M_2 précise), mais il en découle une mauvaise précision pour la borne M_1. La même remarque peut s'appliquer à la permutation 5 : la valeur propre $\lambda_p(A)$ répond à la diminution de la première valeur propre avec une borne m_1 précise mais elle induit une borne m_2 qui n'est absolument pas représentative de la deuxième valeur propre. Cet aspect a déjà été souligné en [2] dans le contexte de l'A.C.P.

Pour conclure d'une manière plus ponctuelle ces commentaires on notera que compte tenu de la valeur initiale de la première valeur propre la borne M_1 dépasse parfois l'unité. On peut bien évidemment la ramener dans ces cas à 1 puisqu'on sait qu'en analyse linéaire discriminante les valeurs propres ne peuvent dépasser cette valeur.

C) Suppression d'une unité statistique

Nous avons ici calculé les encadrements qui correspondent à la suppression de chacune des u.s. et dont la formulation est donnée à la relation (4.6). Comme il serait fastidieux d'en fournir la liste exhaustive le tableau 4 donne seulement ceux correspondant à la suppression des u.s. n° 1 à 15 de chacun des trois groupes, ce qui est suffisant pour avoir une idée générale de leur précision. Les bornes inférieure et supérieure à la variation de la i-ième valeur propre sont comme précédemment désignées par m_i et M_i (i=1,2).

L'examen du tableau 4 (qui constitue un extrait assez représentatif de ce que l'on observe sur l'ensemble des 150 u.s.) met en évidence que les encadrements (4.6) sont en moyenne d'amplitude inférieure à ce que l'on observe dans le cadre de la permutation des classes d'appartenance de deux u.s. Cette précision des encadrements semble plus spécialement marquée pour les iris du groupe 2 ce qui peut s'expliquer par la position centrale de ces derniers sur les facteurs principaux initiaux.

TABLEAU 4

N° DE L'U.S. SUPPRIMEE	GROUPE 1				GROUPE 2				GROUPE 3			
	m_1	M_1	m_2	M_2	m_1	M_1	m_2	M_2	m_1	M_1	m_2	M_2
1	0.957	0.983	0.209	0.235	0.959	0.977	0.211	0.230	0.945	0.990	0.197	0.242
2	0.948	0.990	0.200	0.242	0.967	0.982	0.219	0.234	0.936	0.999	0.188	0.251
3	0.957	0.985	0.209	0.237	0.961	0.982	0.213	0.235	0.959	0.979	0.211	0.231
4	0.950	0.990	0.202	0.242	0.962	0.973	0.214	0.225	0.956	0.985	0.208	0.237
5	0.951	0.989	0.203	0.241	0.955	0.978	0.207	0.230	0.945	0.999	0.198	0.252
6	0.946	0.992	0.198	0.244	0.957	0.976	0.209	0.228	0.954	0.989	0.206	0.241
7	0.955	0.988	0.207	0.240	0.968	0.973	0.220	0.226	0.960	0.982	0.212	0.234
8	0.947	0.991	0.199	0.243	0.956	0.978	0.208	0.230	0.959	0.980	0.211	0.233
9	0.953	0.989	0.205	0.241	0.961	0.979	0.213	0.231	0.956	0.985	0.208	0.237
10	0.954	0.985	0.206	0.237	0.964	0.972	0.217	0.224	0.955	0.985	0.207	0.237
11	0.947	0.992	0.199	0.244	0.968	0.977	0.220	0.229	0.958	1.001	0.210	0.254
12	0.937	1.003	0.189	0.255	0.951	0.982	0.203	0.234	0.950	0.985	0.202	0.238
13	0.952	0.991	0.204	0.243	0.964	0.976	0.217	0.228	0.950	0.988	0.202	0.241
14	0.944	0.995	0.196	0.247	0.959	0.974	0.211	0.226	0.957	0.994	0.209	0.246
15	0.952	0.990	0.204	0.242	0.968	0.979	0.220	0.231	0.955	0.983	0.207	0.235

TABLEAU 5

U.S. SUPPRIMEE n°	Groupe	$c_\ell \|\|x_\ell - g\|\|^2$	$\lambda_p(\tilde{R}'D\,\tilde{R})$	$\lambda_1(\tilde{R}'D\,\tilde{R})$	$\lambda_1(\tilde{B}\,\tilde{V}^{-1})$	m_1	M_1	$\lambda_2(\tilde{B}\,\tilde{V}^{-1})$	m_2	M_2
35	3	0.016	- 0.030	0.009	0.970	0.939	0.995	0.217	0.192	0.247
35	1	0.027	- 0.038	0.014	0.970	0.931	1.011	0.225	0.184	0.263
46	3	0.019	- 0.032	0.009	0.970	0.938	0.998	0.217	0.190	0.250
33	1	0.024	- 0.036	0.012	0.970	0.934	1.006	0.222	0.186	0.258
28	3	0.013	- 0.036	0.017	0.970	0.934	1.001	0.217	0.186	0.253
21	1	0.019	- 0.034	0.014	0.970	0.936	1.003	0.221	0.188	0.255
24	2	0.002	- 0.007	0.005	0.970	0.962	0.977	0.212	0.200	0.234
19	2	0.001	- 0.005	0.004	0.970	0.964	0.975	0.221	0.216	0.227

Si les bornes obtenues révèlent une bonne stabilité de l'analyse par rapport à la perturbation étudiée, il est toutefois intéressant comme au paragraphe B) de les confronter aux variations réelles des valeurs propres. Pour ce faire, nous avons calculé les nouvelles valeurs propres correspondant à la suppression de certaines u.s. pour lesquelles les encadrements semblent moins précis. Les notations étant celles du paragraphe 4, le tableau 5 donne pour chaque u.s. x_ℓ supprimée les valeurs de $c_\ell \, ||x_\ell - g||^2$, de $\lambda_1(\tilde{R}'D\,\tilde{R})$ et $\lambda_p(\tilde{R}'D\,\tilde{R})$ (qui interviennent dans l'expression de (4.6)), les valeurs propres $\lambda_i(\tilde{B}\,\tilde{V}^{-1})$ (i=1,2) ainsi que les bornes m_i et M_i correspondantes.

L'enseignement principal qu'apporte ce dernier tableau réside dans la stabilité des valeurs propres par rapport à la suppression d'une u.s. puisque toutes les variations sont d'amplitude inférieure à 0.01. Les encadrements traduisent bien dans l'ensemble cette précision, ils semblent toutefois moins précis pour les u.s. des groupes 1 et 3 puisqu'ils autorisent une variation moyenne des valeurs propres de 0.03. En ce sens il sont peut être un peu moins pertinents que ceux du paragraphe B). La précision est toutefois toujours suffisante pour signaler à l'utilisateur que la suppression d'aucune u.s. n'est à même d'entraîner des fluctuations sensibles des valeurs propres.

6 - CONCLUSION

Il nous semble possible parmi d'autres interprétations de considérer les perturbations étudiées comme le résultat d'erreurs dans la saisie des données : mauvais codage de la classe d'appartenance d'une u.s., inversion des codages de deux u.s., non prise en compte involontaire d'une u.s. D'un point de vue pratique les résultats obtenus permettent alors de chiffrer la variation maximale qui peut en résulter sur les valeurs propres. L'exemple traité montre que la précision des encadrements est assez bonne pour pouvoir effectivement intéresser l'utilisateur et le guider vers des u.s. jouant un rôle déterminant sur les valeurs propres.

En ce sens les encadrements proposés nous semblent pouvoir prendre place dans la liste des aides à l'interprétation qui permettent au statisticien d'analyser ses résultats. Il est enfin intéressant de remarquer que dans les trois cas étudiés la perturbation se traduit en terme de matrice de rang 2 d'une forme bien particulière ce qui unifie l'approche et la replace dans le contexte général des travaux effectués en A.C.P. sur le même sujet.

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

[1] BENASSENI, J., "Une contribution à l'étude de la stabilité en analyse factorielle", Thèse de Doctorat de 3ème cycle, U.S.T.L., Montpellier, 1984.

[2] BENASSENI, J., "Influence des poids des unités statistiques sur les valeurs propres en analyse en composantes principales", Revue de Statistique Appliquée, à paraître.

[3] ESCOFIER, B., "Stabilité et approximation en analyse factorielle", Thèse de Doctorat d'Etat, Université de Paris VI, 1979.

[4] LACHENBRUCH, P.A., "Discriminant analysis when the initial samples are misclassified II : Non random misclassification models", Technometrics, 1974, Vol. 16, n° 3, p. 419-424.

[5] Mc LACHLAN, G.J., "Asymptotic results for discriminant analysis when the initial samples are misclassified", Technometrics, 1972, Vol. 14, n° 2, p. 415-422.

[6] O'NEIL, T.J., "The general distribution of the error rate of a classification procedure with application to logistic regression discrimination", JASA, 1980, Vol. 75, p. 154-160.

[7] WILKINSON, J.H., The algebraic eigenvalue problem, Clarendon Press, Oxford, 1969.