UNIVERSITÉ PARIS XI

U.E.R. MATHÉMATIQUE

91405 ORSAY FRANCE

N° 200 - 76.58

OPTIMALITE DE L'ANALYSE MULTICANONIQUE PAR LA METHODE DE CAROLL

A. KOBILINSKY.

Publication mathématique d'Orsay.

OPTIMALITE DE L'ANALYSE MULTICANONIQUE PAR LA METHODE DE CAROLL.

A. KOBILINSKY

I. INTRODUCTION.

Par analyse multicanonique, nous entendons toute généralisation à k groupes de variables ($k \geq 2$) de l'analyse canonique introduite par HOTELLING.

Plusieurs critères d'analyse multicanonique ont été proposés : HORST, STEEL, CAROLL, KETTENRING. La méthode que nous proposons ici est équivalente à celle de CAROLL qui cherche une variable - facteur canonique - rendant maximale la somme des carrés des coefficients de corrélation multiple entre elle et chacun des groupes, et itère sous contrainte d'orthogonalité.

KETTENRING présente l'analyse multicanonique comme la recherche de k variables canoniques (une par groupe) les plus correlées possible. Cette conception est difficile à exploiter sur le plan statistique : le choix des contraintes pour la recherche d'un second k-uplet de variables canoniques étant rendu délicat par le fait que la seule contrainte naturelle, non corrélation entre premier et second k-uplet, est beaucoup trop forte quand k dépasse 2.

Du point de vue algorithmique, les facteurs canoniques de la méthode de CAROLL s'obtiennent par simple diagonalisation d'une matrice symétrique. Dans les autres méthodes, le nombre de maximums locaux des critères utilisés est inconnu, et même si l'on suppose que la suite des solutions obtenues dans les algorithmes utilisés converge vers un maximum local (*), on ne possède aucun moyen de vérifier que ce maximum est absolu.

La méthode de CAROLL, qui est donc plus simple à interpréter et mieux connue sur le plan numérique, peut en outre se présenter (**) comme la comparaison par analyse en composantes principales des formes quadratiques g et f définies sur l'espace E des combinaisons linéaires formelles $\alpha' X = \alpha_1 x_1 + \ldots \alpha_p x_p$ des variables étudiées par :

$$g(\alpha' X) = \alpha' \Sigma \alpha \qquad \text{et} \qquad f(\alpha' X) = \alpha' \Sigma_o \alpha$$

Σ étant la matrice de covariance des variables $x_1, \ldots, x_p$, et Σ_o la matrice obtenue en annulant dans Σ les covariances entre variables de groupes différents. Les facteurs canoniques sont les combinaisons $v' X$ obtenues à partir des vecteurs propres de $\Sigma_o^{-1} \Sigma$: cette méthode peut donc être considérée comme la recherche des directions dans lesquelles la variance s'écarte le plus de ce qu'elle serait sous l'hypothèse d'indépendance entre les groupes de variables.

(*) Dans le cas du critère de HORST (somme des corrélations), nous avons donné (KOBILINSKY - 1975) une démonstration de la convergence vers un maximum local pour un algorithme légèrement différent de celui de KETTENRING.

Les démonstrations de convergence données dans DAUXOIS et POUSSE (1975) pour les méthodes de HORST (p. 48), STEEL (p. 60), KETTENRING (p. 55) sont erronées : si pour tout i, u_i est une valeur d'adhérence de la suite $u_i(n)$, le k-uplet $(u_1, \ldots, u_k)$ n'est pas nécessairement une valeur d'adhérence de la suite de k-uplet $(u_1(n), \ldots, u_k(n))$. L'égalité $\phi(u_1, \ldots, u_k) = \lim \phi_i(n)$ n'est donc pas prouvée.

(**) Elle est présentée comme telle par POUSSE et DAUXOIS. La présentation donnée ici, qui n'utilise pas le schéma de dualité, diffère cependant notablement de celle de POUSSE et DAUXOIS.

Le but de ce texte est d'une part de montrer que plusieurs critères apparemment différents conduisent en fait à la méthode de CAROLL, d'autre part d'appliquer les propriétés de l'analyse en composantes principales à cette même méthode.

Dans un premier temps, nous donnons donc plusieurs définitions équivalentes de cette méthode (parmi lesquelles celle de CAROLL).

Puis nous exprimons les propriétés d'optimalité en variation et perte d'information de l'analyse en composante principale énoncées par OKAMOTO (1969) dans le cas où Σ_0 est l'identité, sous une forme géométrique qui s'applique directement à l'analyse de CAROLL.

Enfin, après avoir donné un nouveau critère d'optimalité qui généralise le critère d'optimalité en corrélation (OKAMOTO - 1969, Théorème 4.7), nous en déduisons une propriété remarquable de l'espace défini par les r premiers facteurs canoniques.

2. NOTATIONS ET RAPPELS.

2.1 / Dans ce qui suit, on désigne par :

E un espace vectoriel réel (e.v.r.) de dimension finie p

F un sous-espace de E de dimension r

f une forme bilinéaire symétrique définie positive sur E

g une forme bilinéaire symétrique sur E

U l'endomorphisme f-symétrique tel que $g(x, y) = f(x, Uy)$

P un projecteur g-orthogonal (i.e. tel que $\forall x, \forall y \quad g(x - Px, Py) = 0$) de rang r

$M_1, \ldots, M_k$ des sous-espaces de E de somme directe E

$P_1, \ldots, P_k$ des projecteurs g-orthogonaux sur $M_1, \ldots, M_k$

2.2 /

On notera de la même façon une forme bilinéaire et la forme quadratique qui lui est associée : soit $g(x)$ pour $g(x, x)$.

La restriction de g à F sera notée $g_{|F}$.

Si $\varphi : G \to E$ est un homomorphisme d'e.v.r., on notera $g_o\varphi$ la forme quadratique (ainsi que la forme bilinéaire associée) définie sur G par : $g_o\varphi\,(y) = g(\varphi(y))$.

2.3 /

L'analyse en composantes principales de $\frac{g}{f}$ est la recherche d'une base f-orthonormée et g-orthogonale. Si $e_1, \ldots, e_p$ est une telle base, supposée ordonnée de telle façon que $g(e_1) \geq \ldots \geq g(e_p)$, on appellera :

e_i le $i^{\text{ème}}$ vecteur propre ou la $i^{\text{ème}}$ composante principale de $\frac{g}{f}$

$g(e_i)$ la $i^{\text{ème}}$ valeur propre de $\frac{g}{f}$

$g(e_1) + \ldots + g(e_p)$ la trace de $\frac{g}{f}$

$g(e_1) \ldots g(e_p)$ le déterminant de $\frac{g}{f}$

2.4 /

Les vecteurs propres, valeurs propres, la trace et le déterminant de $\frac{g}{f}$ sont les vecteurs propres, valeurs propres, la trace et le déterminant de U. Si $x_1, \ldots, x_p$ est une base de E, A et B les matrices respectives de g et f dans cette base, $A^{-1}B$ est la matrice de U. Si $x_1, \ldots, x_p$ est une base f-orthonormée, A est l'identité et la trace de $\frac{g}{f}$ est égale à $g(x_1) + \ldots + g(x_p)$.

2.5 / On notera :

$e_1(F), \ldots, e_r(F)$ les vecteurs propres de $\frac{g_{|F}}{f_{|F}}$

$\lambda_1(F), \ldots, \lambda_r(F)$ les valeurs propres correspondantes

$e_1(P), \ldots, e_r(P)$ les r premiers vecteurs propres, choisis dans le f-orthogonal du noyau de P, de $\frac{g_o P}{f}$

$\lambda_1(P), \ldots, \lambda_r(P)$ les valeurs propres correspondantes

$e_1, \ldots, e_p$ les vecteurs propres de $\frac{g}{f}$

$\lambda_1, \ldots, \lambda_p$ les valeurs propres correspondantes

E_r l'espace engendré par $e_1, \ldots, e_r$

G_r l'espace engendré par $e_{p-r+1}, \ldots, e_p$

Q_r le projecteur f-orthogonal sur E_r

R_r le projecteur f-orthogonal sur G_r .

3. DEFINITION DE L'ANALYSE MULTICANONIQUE PAR LA METHODE DE CAROLL.

Dans cette section, g est supposée positive et définie sur chaque M_i, f est la forme bilinéaire égale à g sur les M_i qui rend les M_i orthogonaux deux à deux.

Définition : L'analyse multicanonique par la méthode de CAROLL des espaces $M_1, \ldots, M_k$ et de g est l'analyse en composantes principales de $\frac{g}{f}$. Les facteurs canoniques de l'analyse sont les vecteurs propres $e_1, e_2, \ldots$ de $\frac{g}{f}$.

Dans les applications, $M_1, \ldots, M_k$ seront des sous-espaces d'un espace $L^2(\Omega, \mathcal{A}, u)$, E leur somme directe, g la forme quadratique canoniquement induite sur E par la métrique de covariance.

3.1 / Proposition 3.1

U est égal à $P_1 + \ldots + P_k$.

Démonstration :

Soient $y = y_1 + \ldots + y_k$, $y_i \in M_i$ et x des vecteurs de E.

$$g(x, y) = g(x, y_1) + \ldots + g(x, y_k) = g(P_1 x, y_1) + \ldots + g(P_k x, y_k) =$$

$$f(P_1 x, y_1) + \ldots + f(P_k x, y_k) = f(P_1 x, y) + \ldots + f(P_k x, y) = f(U x, y)$$

3.2 / Une infinité dénombrable de définitions de l'analyse multicanonique !!

Pour $z \in Z$, soit f_z la forme bilinéaire symétrique définie sur l'orthogonal du noyau de g par $f_z(x) = f(x, U^z x)$. On vérifie aisément que f est toujours définie positive.

Proposition 3.2

z_1 et z_2 appartenant à Z, $z_1 \neq z_2$, on peut définir l'analyse multicanonique de g et $M_1, \ldots, M_k$ comme l'analyse en composantes principales de $\frac{f_{z_1}}{f_{z_2}}$. Les facteurs canoniques n'appartenant pas à Ker g sont les vecteurs propres de cette analyse, tandis que ses valeurs propres sont celles de $\frac{g}{f}$ élevées à la puissance $z_1 - z_2$.

Démonstration :

L'endomorphisme f_{z_2}-symétrique associé à f_{z_1} est précisément $U^{z_1 - z_2}$, i.e $f_{z_1}(x, y) = f_{z_2}(U^{z_1 - z_2} x, y)$. Or $U^{z_1 - z_2}$ a les mêmes vecteurs propres que U et ses valeurs propres sont celles de U élevées à la puissance $z_1 - z_2$.

3.3 / Quelques cas particuliers - la définition de CAROLL.

On peut appliquer 3.2 avec les formes $f_0(x) = f(x)$, $f_1(x) = g(x)$, $f_2(x) = g(x, U x) = g(x, P_1 x) + \ldots + g(x, P_k x) = g(P_1 x) + \ldots + g(P_k x)$, $f_3(x) = g(U x)$.

Ainsi $$\frac{f_2(x)}{f_1(x)} = \frac{g(x, P_1 x)}{g(x)} + \ldots + \frac{g(x, P_k x)}{g(x)} = \cos^2(x, M_1) + \ldots + \cos^2(x, M_k)$$

Cette égalité montre l'équivalence entre l'analyse définie ici et celle définie par CAROLL (1968).

4. OPTIMALITE DE L'ANALYSE EN COMPOSANTES PRINCIPALES.

4.1 / Théorème 4.1

Pour tout F, on a les inégalités : $\lambda_1 = \lambda_1(E_r) \geq \lambda_1(F) \geq \lambda_1(G_r) = \lambda_{p-r+1}$

$$\lambda_2 = \lambda_2(E_r) \geq \lambda_2(F) \geq \lambda_2(G_r) = \lambda_{p-r+2}$$

$$\dots$$

$$\lambda_r = \lambda_r(E_r) \geq \lambda_r(F) \geq \lambda_r(G_r) = \lambda_p$$

Remarque : Ce théorème entraîne que toute fonction des $\lambda_i(F)$ croissante en chaque λ_i est maximum quand $F = E_r$. On notera que la trace et le déterminant de $\frac{g_{|F}}{f_{|F}}$ sont de telles fonctions.

Démonstration du théorème 4.1 :

L'espace engendré par $e_1(F), \dots, e_i(F)$, de dimension i, et l'espace engendré par $e_i, \dots, e_p$, de dimension p-i+1, ont un vecteur commun e. En écrivant e successivement comme combinaison linéaire de $e_1(F), \dots, e_i(F)$, puis de $e_i, \dots, e_p$, on voit que : $\lambda_i \geq \frac{g(e)}{f(e)} \geq \lambda_i(F)$.

On peut formuler le théorème 4.1 de la façon suivante :

Il existe pour tout F de dimension r une isométrie pour la métrique f $\varphi : F \to E_r$ qui incrémente les normes pour g :

$$f(x) = f(\varphi(x)) \qquad g(x) \leq g(\varphi(x))$$

On déduit immédiatement de ce théorème le théorème 4.3 d'OKAMOTO (1969).

4.2. / Alors que la propriété étudiée en 4.1 est presque symétrique en f et g, les propriétés énoncées en 4.2 dissymétrisent les rôles de f et g : on ne cherche plus l'espace où f et g diffèrent le plus, mais le projecteur P qui rend $g_o P$ "le plus grand possible" ou au contraire "le plus petit possible". f permet de définir une pondération : $g_o P(x)$ est pondéré par $\frac{1}{f(x)}$.

Théorème 4.2

Pour tout P, on a les inégalités : $\lambda_1 = \lambda_1(Q_r) \geq \lambda_1(P) \geq \lambda_1(R_r) = \lambda_{p-r+1}$

$$\lambda_2 = \lambda_2(Q_r) \geq \lambda_2(P) \geq \lambda_2(R_r) = \lambda_{p-r+2}$$

$$\ldots$$

$$\lambda_r = \lambda_r(Q_r) \geq \lambda_r(P) \geq \lambda_r(R_r) = \lambda_p$$

Démonstration :

a. Inégalités relatives à Q_r.

On utilise la g-orthogonalité de P et un argument analogue à celui utilisé dans la démonstration de 4.1 : l'espace engendré par $e_i, \ldots, e_p$ et l'espace engendré par $e_1(P), \ldots, e_i(P)$ ont un vecteur commun e ; on a alors les inégalités suivantes :

$$\lambda_i \geq \frac{g(e)}{f(e)} \geq \frac{g_o P(e)}{f(e)} \geq \lambda_i(P).$$

b. Inégalités relatives à R_r

La g-orthogonalité de P n'est pas utilisée ici. L'espace engendré par $e_1, \ldots, e_{p-r+i}$ et l'espace engendré par $P(e_i(P)), \ldots, P(e_r(P))$ ont un vecteur commun Pe . e étant f-orthogonal au noyau de P, on a $f(Pe) = f(Pe - e) + f(e) \geq f(e)$ et donc : $\lambda_{p-r+i} \leq \frac{g(Pe)}{f(Pe)} \leq \frac{g(Pe)}{f(e)} \leq \lambda_i(P)$.

Remarque : On peut faire des remarques analogues à celles qui ont été faites en 4.1 : tout fonction des $\lambda_i(P)$ croissante en chaque λ_i est maximum quand $P = Q_r$. Il en est donc ainsi de la trace et du déterminant de $\frac{g_o P}{f}$.

Il existe une f-isométrie $\varphi : E \to E$ telle que : $g_o P(x) \leq g_o Q_r(\varphi(x))$.

On peut déduire le théorème 4.4 d'OKAMOTO des inégalités relatives à R_r en remarquant que si V est un endomorphisme de rang $p - r$, P le projecteur g-orthogonal sur le g-orthonal de Im V, $g(Px) \leq g(x - Vx)$.

Les inégalités relatives à Q_r généralisent le théorème 4.7 d'OKAMOTO (1969) (Optimalité en corrélation). La quantité qu'il rend maximum est la trace de $\frac{g_o P}{f}$, g étant définie par la matrice de covariance, f par la matrice identité dans la base des x_i, P étant le projecteur sur l'espace engendré par les coordonnées du vecteur aléatoire y. On déduit de ces inégalités le théorème suivant :

4.3 / Nous reprenons ici les notations de la section 3.

Théorème 4.3

Si l'on prend comme indice de proximité entre deux espaces M et L la somme $d^2(M, L)$ des carrés des coefficients de corrélation canonique entre M et L, l'espace engendré par les r premiers facteurs canoniques est l'espace F de dimension r qui rend maximum la somme : $d^2(M_1, F) + \ldots + d^2(M_k, F)$.

Démonstration :

Soit $y_1, \ldots, y_p$ la base f-orthonormée de E obtenue en réunissant les vecteurs canoniques dans M_1 de l'analyse canonique de M_1 et F, ..., les vecteurs canoniques dans M_k de l'analyse canonique de M_k et F. Soit P le projecteur g-orthogonal sur F. La trace de $\frac{g_o P}{f}$ est égale à $g_o P(y_1) + \ldots + g_o P(y_p)$ (cf. 2.4). Comme $g(y_i) = f(y_i) = 1$, $g_o P(y_i)$ est le cosinus carré de l'angle entre y_i et F. Le critère de la proposition n'est donc autre que la trace de $\frac{g_o P}{f}$, dont on sait qu'elle est maximum quand $F = E_r$.

BIBLIOGRAPHIE :

1. CAROLL , J.D. (1968). A generalization of canonical correlation analysis to three or more sets of variables. Proc. 76th Conv. Amer. Psych. Assoc. p. 227-228.

2. DAUXOIS, J., et POUSSE, A. (1975). K-canoniques analyses. Publication du laboratoire de Statistique de l'Université Paul Sabatier. Toulouse.

3. HORST, P. (1961). Relations among m sets of variables. Biometrika, vol. 58, p. 433-451

4. KETTENRING, R.J. (1971). Canonical analysis of several sets of variables. Biometrika, vol. 58, p. 433-451.

5. KOBILINSKY, A. (1975). Sur les relations entre plusieurs groupes de variables. Document interne du laboratoire de Biométrie de l'INRA. Versailles.

6. OKAMOTO, M. (1969). Optimality of principal components. In Multivariate Analysis, 2 (P.R. Krishnaiah, éd.), p. 673-685. Academic Press, New York.

7. PAGES, J.P., NAKACHE J.P., MAILLES, J.P., CA ILLIEZ , F (1971) Analyse des données multidimensionnelles, C3E, Paris, Tome II, Chapitre 9

8. LEFORT, G. Communications personnelles.