MATHÉMATIQUES ET SCIENCES HUMAINES

RÉMY VIREDAZ

Sur l'invariance de l'indice de redondance

Mathématiques et sciences humaines, tome 57 (1977), p. 59-75

<http://www.numdam.org/item?id=MSH_1977__57__59_0>

Math. Sci. hum. (15e année, n°57, 1977, p.59-75)

SUR L'INVARIANCE DE L'INDICE DE REDONDANCE [1]

Rémy VIREDAZ

1. La méthode des corrélations canoniques [2], ou analyse canonique, est une méthode de statistique multivariée destinée à mesurer et éventuellement à analyser la relation entre deux jeux de variables X et Y sur une même population. Il pourrait s'agir par exemple d'un jeu de variables décrivant des conditions socio-professionnelles d'une part, et d'un jeu de variables décrivant la symptomatologie psychiatrique d'autre part, observés sur l'ensemble des consultants d'une policlinique psychiatrique pendant une certaine période.

Le procédé est le suivant. On cherche d'abord une combinaison linéaire f_1 des variables du premier jeu et une combinaison linéaire g_1 des variables du second jeu telles que la corrélation de f_1 et g_1 soit maximale; celle-ci est appelée la *corrélation canonique* des deux jeux de variables X et Y . On cherche ensuite deux combinaisons linéaires f_2 et g_2 définies de la même façon mais sous la condition que toutes deux soient orthogonales (sans corrélation) aussi bien à f_1 qu'à g_1 ; et ainsi de suite par itération.

Pour mesurer la relation entre les deux jeux de variables X et Y , on s'est vite rendu compte que la corrélation canonique convient mal :

---

1. Cet article est issu d'un colloque tenu à l'Ecole des Sciences sociales et politiques de l'Université de Lausanne, chaire de Mathématiques pour les sciences humaines, à l'instigation du professeur Georges Leresche.

2. On pourra consulter, en français, Caillez et Pagès (1), et Lebart et Fénelon (2).

comme elle ne fait intervenir qu'une seule paire de facteurs, elle peut très bien être forte même si globalement les deux jeux ont peu de relation entre eux - car f_1 et g_1 peuvent être fortement corrélés entre eux tout en étant des facteurs relativement "marginaux" à l'intérieur de leur propre jeu. On a donc introduit une autre quantité, appelée la *redondance* du premier jeu de variables étant donné le second.

Cette quantité peut se calculer de deux façons : soit comme moyenne des carrés des corrélations multiples entre les variables du premier jeu et le second jeu (formule A ci-dessous § 4), soit comme somme des "indices de redondance" attachés à chaque facteur du premier jeu (formule B). - On calcule symétriquement la redondance du second jeu étant donné le premier.

Comme, par définition, la moyenne des carrés des corrélations multiples (A) est indépendante de la méthode d'orthogonalisation utilisée, on peut se demander si (ou : dans quelle mesure) la somme des indices de redondance (B) en est indépendante elle aussi.

D'après Miller et Farr (4) elle l'est, et dans tous les cas; mais leur démonstration est erronée (v. § 4) et leur exemple est particulier (facteurs principaux). Le problème reste donc posé.

2. Nous le traiterons géométriquement, en utilisant le fait que, sur l'espace vectoriel des variables de moyenne nulle définies sur une population donnée P, la *covariance* est un *produit scalaire*. Arrêtons-nous un moment sur ce point.

Soit P une population d'effectif N. Une variable (numérique) x sur P est une application de P dans le corps des nombres réels $\mathbb{R}$: à chaque individu a de P est associé un nombre, appelé score de a et noté $x(a)$. Définissons la somme de deux variables x et y comme la variable $x + y$ qui associe à a le score $x(a) + y(a)$, le produit d'une variable x par un nombre k comme la variable $k\,x$ qui associe à a le score $k\,x(a)$, et le produit de deux variables x et y comme la variable $x\,y$ qui associe à a le score $x(a)\,y(a)$. Les deux premières de ces opérations font de l'ensemble des variables sur P un *espace vectoriel*; et l'ensemble des variables de moyenne nulle en est un sous-espace.

Sur un tel espace vectoriel, on peut toujours introduire un *produit scalaire* et par là-même une métrique (longueurs, angles). Rappelons qu'un produit scalaire associe à tout couple de vecteurs (x; y) un nombre, noté $x \cdot y$, et ceci de manière bilinéaire (linéaire en x et linéaire en y), commutative ($x \cdot y = y \cdot x$), et définie-positive ($x \cdot x$ est positif, et nul seulement si x est le vecteur nul - en l'occurrence, si tous les individus de P ont le même score 0 pour la variable x). Pour la statistique - étant donné les formules de la moyenne, de la variance et de la covariance - il convient de choisir comme produit scalaire la moyenne de la variable-produit (il est facile de vérifier que c'est bien un produit scalaire) :

$$x \cdot y = \frac{1}{N} \sum_a x(a)\, y(a) \ .$$

En particulier, la covariance de deux variables x et y sera le produit scalaire des variables de moyenne nulle associées à x et à y. La covariance est comme le produit scalaire une forme bilinéaire commutative, mais elle n'est pas définie-positive ($\mathrm{cov}(x, x) = 0$ n'implique pas que x soit nul : il suffit que tous les individus de P aient le même score pour x). Mais sur le sous-espace des variables de moyenne nulle ces deux formes bilinéaires s'identifient. Comme nous ne considérerons que des variables de moyenne nulle, nous pouvons dire que la covariance de deux telles variables x et y est simplement le produit scalaire $x \cdot y$.

On a dès lors les équivalences suivantes entre concepts géométriques et statistiques :

produit scalaire, $x \cdot y$	covariance
carré de la longueur, $\|\|x\|\|^2$	variance
cosinus de l'angle :	
entre deux vecteurs	corrélation
entre un vecteur et un sous-espace	corrélation multiple
entre deux sous-espaces	corrélation canonique
vecteurs orthogonaux, $x \perp y$	variables ou facteurs non corrélés
carré de la longueur de	
la projection orthogonale	variance expliquée

etc.

En particulier, la méthode des corrélations canoniques peut être présentée comme suit. Soient X et Y deux faisceaux de vecteurs (dans un espace multidimensionnel). On cherche d'abord un vecteur f_1 situé dans le sous-espace engendré par X, et un vecteur g_1 situé dans le sous-espace engendré par Y, tels que l'angle entre f_1 et g_1 soit minimum. (Dans le cas particulier où X serait formé d'un seul vecteur, et Y de deux vecteurs, on obtiendrait ainsi l'angle entre la droite définie par X et le plan défini par Y.) On cherche ensuite deux vecteurs f_2 et g_2 définis de la même façon mais sous la condition que tous deux soient orthogonaux aussi bien à f_1 qu'à g_1 ; et ainsi de suite.

3. Nous utiliserons les notations suivantes.

Soient :

$X = (x_i)_{i = 1, p}$ et $Y = (y_j)_{j = 1, q}$ deux systèmes de vecteurs normés (c'est-à-dire de longueur unité);

(X) et (Y) les sous-espaces qu'ils engendrent;

$F = (f_h)_{h = 1, m}$ et $G = (g_k)_{k = 1, n}$ deux bases orthonormées quelconques de (X) et (Y).

(Si l'espace vectoriel considéré est celui des variables de moyenne nulle sur une population, X et Y sont deux jeux de variables préalablement standardisées (moyenne nulle, variance unité), et F et G deux jeux complets de facteurs orthogonaux.)

Posons x'_i = projection orthogonale de x_i sur (Y), et de même f'_h, et inversement y'_j, g'_k celles de y_j, g_k sur (X).

La h^e composante ("saturation") de x_i dans la base F est $x_{ih} = x_i \cdot f_h$; la k^e composante de x'_i dans la base G est $x'_{ik} = x'_i \cdot g_k = x_i \cdot g_k$. On définit de même y_{jk}, y'_{jh}, et $f'_{hk} = g'_{kh} = f_h \cdot g_k$.

On écrira $X \cdot Y$ la matrice des produits scalaires de X et Y, c'est-à-dire la matrice dont l'élément en ligne i et colonne j est $x_i \cdot y_j$. (Comme x_i et y_j sont de longueur 1, $x_i \cdot y_j$ est un cosinus, en statistique une corrélation : $X \cdot Y$ est en notation statistique R_{XY}.)

Par "variance" de X, on entendra la somme des variances des va-

riables, $\sum_i ||x_i||^2 = \text{trace}(X \cdot X)$ (= p , puisque les x_i sont normés). C'est la somme des carrés des longueurs des vecteurs du "faisceau" X .

On abrégera l'expression fréquente "proportion de variance de a expliquée par b " (où a aussi bien que b peuvent être soit une variable isolée soit un jeu de variables) en "pve(a, b)".

Pour faciliter la lecture, les indices i , j , h , k auront toujours le même rôle que ci-dessus (p.ex., il sera toujours entendu que i se rapporte aux vecteurs du système X et va de 1 à p). Quand les facteurs des deux jeux sont appariés, comme c'est le cas des facteurs canoniques, h et k seront souvent remplacés par ℓ .

4. Considérons alors les quantités

$$A = \frac{1}{p} \sum_i ||x'_i||^2 = \frac{1}{p} \sum_i \sum_k x'^2_{ik} ,$$

$$B = \sum_h \left[\left(\frac{1}{p} \sum_i x^2_{ih} \right) (||f'_h||^2) \right] = \frac{1}{p} \sum_i \sum_k \sum_h x^2_{ih} f'^2_{hk} .$$

A est la proportion de variance de X expliquée par Y , ou moyenne des carrés des corrélations multiples entre les x_i et (Y) .

B s'obtient en calculant, pour chaque facteur f_h , le produit $\text{pve}(X, f_h) \times \text{pve}(f_h, Y)$, et en additionnant ces produits.

Selon Miller et Farr (4), p. 317, il résulte d'un certain "théorème de multiplication" que ces $\text{pve}(X, f_h) \times \text{pve}(f_h, Y)$ sont les contributions de chaque f_h à pve(X, Y) , de sorte que B serait égal à A , et ceci, indépendamment de F et G (c'est-à-dire, quelle que soit la méthode d'orthogonalisation choisie). Ce théorème, s'il était vrai, serait extrêmement intéressant : en effet, puisque A , par définition, ne dépend pas des bases F et G , il aurait pour corollaire que B n'en dépend pas non plus. Il n'y aurait donc plus besoin, pour démontrer l'invariance de B , de la longue ("lengthy") démonstration de Miller (3) (ce dernier point semble avoir échappé à Miller et Farr). [3]

---

3. Miller (3) a-t-il réellement montré que n'importe quelles bases F et

Mais le raisonnement de Miller et Farr est incorrect. Un produit de proportions de variance expliquée n'a aucune raison d'être une proportion de variance expliquée. Le "théorème de multiplication" dont ils s'autorisent est un théorème (ou plutôt un axiome) qui s'applique aux probabilités d'événements indépendants, et n'a donc rien à faire ici! En fait, une variance "expliquée" n'est qu'une *longueur projetée* au carré; or si on projette les x_i sur f_h, puis qu'on projette le résultat sur (Y), on n'obtiendra en général pas le même résultat qu'en projetant les x_i sur (Y) directement. [4]

La question reste donc posée : est-ce que A = B indépendamment de F et de G ?

Comme A et B sont tous deux des sommes $\left(\frac{1}{p}\sum_i\sum_k \ldots\right)$, on peut même se demander si cette égalité a lieu terme à terme : est-ce que $x'^2_{ik} = \sum_h x^2_{ih} f'^2_{hk}$ pour tout i et pour tout k ? Ou, si ce n'est pas le cas, est-ce que du moins les sommes sur i, ou les sommes sur k, sont égales entre elles?

Enfin, au cas où ces égalités ne seraient pas vraies en général, dans quels cas particuliers le sont-elles?

5. Pour répondre à ces questions, le mieux est de commencer par étudier un exemple. Nous avons pris $p = m = 2$ pour X, $q = n = 3$ pour Y. Les bases orthonormées ont les propriétés suivantes : G est quelconque; P et Q sont les facteurs principaux de X et Y respectivement; U et V sont telles que $h \neq k \Rightarrow u_h \perp v_k$ (condition d'"orthogonalité mutuelle" entre les deux jeux de facteurs, qui est l'une des conditions imposées sur les facteurs canoniques).

G donnent pour B le même résultat, comme l'écrivent Miller et Farr (4), p. 318, ou seulement que les facteurs principaux donnent le même résultat que les facteurs canoniques? En tous cas, seule est vraie la seconde de ces propositions.

4. La variance d'un facteur f_h peut être choisie arbitrairement (nous l'avons prise égale à 1). La prendre égale à la variance expliquée par f_h peut induire en erreur, mais ne change rien au raisonnement, où la longueur de f_h n'intervient pas. D'ailleurs les projections des f_h ne sont pas orthogonales entre elles.

Données artificielles :

$$U = \begin{pmatrix} \frac{1}{\sqrt{5}} & 0 & \frac{2}{\sqrt{5}} & 0 & 0 \\ 0 & \frac{2}{\sqrt{5}} & 0 & \frac{1}{\sqrt{5}} & 0 \end{pmatrix} \qquad V = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

(par rapport à une certaine base orthonormée de l'espace somme);

$$U \cdot V = \begin{pmatrix} \frac{2}{\sqrt{5}} & 0 & 0 \\ 0 & \frac{1}{\sqrt{5}} & 0 \end{pmatrix};$$

$$F = \begin{pmatrix} \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{5}} \\ \frac{1}{\sqrt{5}} & -\frac{2}{\sqrt{5}} \end{pmatrix} U \qquad X = \begin{pmatrix} \frac{4}{5} & \frac{3}{5} \\ \frac{4}{5} & -\frac{3}{5} \end{pmatrix} F \qquad P = F$$

$$G = \begin{pmatrix} \frac{1}{3} & \frac{2}{3} & -\frac{2}{3} \\ 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{4}{3\sqrt{2}} & \frac{-1}{3\sqrt{2}} & \frac{1}{3\sqrt{2}} \end{pmatrix} V \qquad Y = \begin{pmatrix} \frac{1}{3} & \frac{2}{3} & \frac{2}{3} \\ \frac{2}{3} & \frac{1}{3} & \frac{2}{3} \\ \frac{1}{3} & -\frac{2}{3} & \frac{2}{3} \end{pmatrix} G$$

$$Y = \begin{pmatrix} .8963 & -.4207 & .1394 \\ .9855 & -.0648 & -.1566 \\ .5162 & .8545 & .0567 \end{pmatrix} Q \qquad \begin{aligned} \lambda_1 &= 2.0413 \\ \lambda_2 &= .9115 \\ \lambda_3 &= .0472 \end{aligned}$$

(les matrices de passage sont les matrices de corrélations $F \cdot U$, $X \cdot F$, $G \cdot V$, $Y \cdot G$ et $Y \cdot Q$);

$$X \cdot X = \begin{pmatrix} 1 & \frac{7}{25} \\ \frac{7}{25} & 1 \end{pmatrix} \qquad Y \cdot Y = \begin{pmatrix} 1 & \frac{8}{9} & \frac{4}{9} \\ \frac{8}{9} & 1 & \frac{1}{9} \\ \frac{4}{9} & \frac{1}{9} & 1 \end{pmatrix}$$

$$X \cdot Y = \begin{pmatrix} .6079 & .7068 & .6833 \\ .5104 & .5495 & .1333 \end{pmatrix}$$

(ces trois dernières matrices ne sont pas utiles pour notre propos, mais permettent de s'assurer que l'exemple est "suffisamment quelconque").

En calculant la matrice $X \cdot G$, puis en élevant chacun de ses termes au carré - ce que nous noterons $(X \cdot G)_2$ - on obtient la matrice des

x'^2_{ik} , dont on calcule ensuite les sommes des lignes et les sommes des colonnes; le total est égal à p A . Ce calcul sera désigné par "X vers G dir(ectement)" dans le tableau 1. On peut faire le même calcul en remplaçant G par V ou par Q .

Le produit matriciel des matrices $(X \cdot F)_2$ et $(F \cdot G)_2$ donne la matrice des $\sum_h x^2_{ih} f'^2_{hk}$, dont on calcule ensuite les sommes des lignes et les sommes des colonnes; le total est égal à p B . Ce calcul sera désigné par "X vers G via F". On peut faire le même calcul en remplaçant F (= P) par U , ou G par V ou par Q .

On compare ensuite chaque calcul "indirect" au calcul "direct" correspondant (tableau 1).

On procède ensuite de même pour Y , ce qui donne une nouvelle série d'exemples (tableau 2).

Les calculs ont été faits en valeurs exactes tant que c'était praticable, puis à 4 décimales par défaut; ici les résultats sont arrondis à deux décimales pour une meilleure lisibilité.

6. Outre le fait que A ne dépend ni de F ni de G , ni B de G (par définition), les tableaux 1 et 2 s'accordent à faire apparaître les résultats suivants :

THEOREME 1

Si les bases F et G sont mutuellement orthogonales ($h \neq k \Rightarrow f_h \perp g_k$), *alors :*
Pour tout i et pour tout k ,

$$x'^2_{ik} = \sum_h x^2_{ih} f'^2_{hk} . \qquad (1)$$

Démonstration. $f'_{hk} = f_h \cdot g_k = 0$ si $h \neq k$, de sorte que le second membre se réduit à $x^2_{ik} f'^2_{kk}$. D'autre part, l'équation $x_i = \sum_h x_{ih} f_h$ donne, par projection orthogonale sur g_k : $x'_{ik} = \sum_h x_{ih} f'_{hk}$; mais cette somme, pour la même raison, se réduit à $x'_{ik} = x_{ik} f'_{kk}$, d'où l'égalité à démontrer.

X

vers :

	V				Q				G			
dir.	.77	.01	0	.78	.61	.10	.07	.78	.06	.00	.72	.78
	.16	.16	0	.32	.27	.02	.02	.32	.16	.08	.08	.32
	.93	.17	0	1.10	.88	.12	.10	1.10	.22	.08	.80	1.10
via :												
U	.77	.01	0	.78	.66	.07	.05	.78	.09	.00	.69	.78
	.16	.16	0	.32	.16	.08	.08	.32	.09	.08	.15	.32
	.93	.17	0	1.10	.82	.15	.13	1.10	.18	.08	.84	1.10
P	.47	.08	0	.55	.44	.06	.05	.55	.11	.04	.40	.55
	.47	.08	0	.55	.44	.06	.05	.55	.11	.04	.40	.55
	.93	.17	0	1.10	.88	.12	.10	1.10	.22	.08	.80	1.10

Tableau 1

Y

vers :

	U			P		
dir.	.44	.06	.50	.49	.01	.50
	.58	.05	.63	.62	.02	.63
	.44	.03	.47	.26	.21	.47
	1.45	.14	1.60	1.37	.23	1.60
via :						
V	.44	.06	.50	.36	.13	.50
	.58	.05	.63	.47	.16	.63
	.44	.03	.47	.36	.11	.47
	1.45	.14	1.60	1.19	.41	1.60
Q	.56	.04	.60	.54	.07	.60
	.66	.03	.69	.65	.05	.69
	.23	.07	.30	.18	.12	.30
	1.45	.14	1.60	1.36	.23	1.60
G	.33	.06	.39	.25	.14	.39
	.36	.06	.41	.30	.12	.41
	.33	.06	.39	.25	.14	.39
	1.01	.17	1.18	.79	.39	1.18

Tableau 2

THEOREME 2

Si la base F est celle des facteurs principaux, alors :
Pour tout k , et quel que soit G ,

$$\sum_i x'^2_{ik} = \sum_i \sum_h x^2_{ih} f'^2_{hk} . \qquad (2)$$

Démonstration. Le second membre peut s'écrire $\sum_h \lambda_h f'^2_{hk}$, puisque $\sum_i x^2_{ih} = \lambda_h$ (h^e valeur propre de la matrice des corrélations X•X).
D'autre part, si A est la matrice des "saturations" factorielles (X = A F , avec ${}^tA\,A = \Lambda$, matrice diagonale des valeurs propres), on a X•G = A (F•G) , c'est-à-dire $x'_{ik} = x_i \cdot g_k = \sum_h a_{ih} f_h \cdot g_k$.
En prémultipliant X•G par sa transposée, il vient : (G•X)(X•G) = = (G•F) ${}^tA\,A$ (F•G) = (G•F) Λ (F•G) . En lisant les termes diagonaux de cette équation matricielle, on a, pour tout k , $\sum_i x'^2_{ik} = \sum_h \lambda_h f'^2_{hk}$, d'où l'égalité à démontrer.

COROLLAIRES

1.1 *Si F est telle qu'il existe G avec* $h \neq k \Rightarrow f_h \perp g_k$, *alors :*
Pour tout i (et quelle que soit la base G' utilisée pour calculer effectivement ces quantités),

$$||x'_i||^2 = \sum_h x^2_{ih} ||f'_h||^2 .$$

(Sommer sur k l'égalité (1).)

1.2 *Si F et G sont les bases des facteurs canoniques, alors A = B .*

(Sommer sur k et sur i l'égalité (1), et particulariser F et G .)

2.1 *Si F et G sont les bases des facteurs principaux, alors A = B .*

(Sommer sur k l'égalité (2), et particulariser G .)

2.2 *La formule B donne avec les facteurs principaux le même résultat qu'avec les facteurs canoniques.*

(Conséquence de 1.2 et 2.1, puisque A est invariant.)

On obtient ainsi deux théorèmes déjà connus (1.2 et 2.2), mais comme corollaires de deux théorèmes plus forts, et avec des démonstrations plus simples.

En revanche, quand F est quelconque, B peut ne pas être égal à A, et le tableau 2 en donne un exemple (où Y, X, G de l'exemple jouent les rôles des X, Y, F de la théorie). Le tableau 1 ne contient pas de tel exemple, parce que nous n'avons utilisé pour X que des bases particulières (facteurs canoniques, facteurs principaux).

7. La conséquence pratique des considérations théoriques qui précèdent est qu'il ne semble pas y avoir de sens à utiliser B, ni sa décomposition en somme des contributions des facteurs du jeu "expliqué", à savoir

$$B = \sum_h \left[\mathrm{pve}(X, f_h) \times \mathrm{pve}(f_h, Y)\right] ,$$

pour l'analyse de données bimultivariées.

Ce qui est important, c'est la proportion de variance de X expliquée par Y [ou inversement], c'est-à-dire, par définition (§3), la moyenne des carrés des corrélations multiples. Celle-ci peut se calculer par la formule A quelle que soit la méthode d'orthogonalisation utilisée, alors que la formule B ne s'applique qu'aux facteurs canoniques [5] ou principaux.

Si l'on veut décomposer cette proportion de variance expliquée, il paraît naturel de l'analyser en somme des contributions des facteurs du jeu "expliquant", non du jeu "expliqué" : en considérant Y comme expliquant X, on veut voir comment les facteurs de Y se partagent la tâche :

$$A = \mathrm{pve}(X, Y) = \sum_k \mathrm{pve}(X, g_k) .$$

Le fait important est que cette décomposition, elle aussi, est possible quelle que soit la méthode d'orthogonalisation, alors que la décomposition de B a peu d'intérêt puisqu'elle n'est possible que pour des

---

5. Ce que nous disons des facteurs canoniques s'applique en fait à tout double jeu de facteurs mutuellement orthogonaux, seule hypothèse utilisée dans le théorème 1.

facteurs qui ne sont guère interprétables, à savoir les facteurs canoniques ou les facteurs principaux, mais non pas des facteurs Varimax ou d'autres facteurs après rotation.

8. Notons d'ailleurs que dans le cas de l'analyse canonique la "mauvaise" décomposition se trouve coïncider avec la "bonne" : les contributions des facteurs du jeu expliqué, $\mathrm{pve}(X, f_\ell) \times \mathrm{pve}(f_\ell, Y)$, sont égales à celles du jeu expliquant, $\mathrm{pve}(X , g_\ell)$. On a en effet

$$\mathrm{pve}(x_i, f_\ell) \times \mathrm{pve}(f_\ell, Y) = x_{i\ell}^2 \, ||f'_\ell||^2 = x_{i\ell}^2 \, r_\ell^2 ,$$

où r_ℓ est la ℓ^e corrélation canonique, et d'autre part

$$\mathrm{pve}(x_i, g_\ell) = x'^2_{i\ell} = \sum_h x_{ih}^2 \, f'^2_{h\ell} = x_{i\ell}^2 \, r_\ell^2$$

(en utilisant le théorème 1 et le fait que $f'_{h\ell} = 0$ si $h \neq \ell$). En sommant ces égalités sur i , on obtient la proposition à démontrer.

Cela signifie que, dans le cas particulier des facteurs canoniques, $\frac{1}{p} \sum_i x_{i\ell}^2 \, r_\ell^2$ ne doit pas être considéré comme la contribution de f_ℓ à $\mathrm{pve}(X, Y)$, mais comme la contribution de g_ℓ , ou du couple $(f_\ell ; g_\ell)$. Pour les autres méthodes d'orthogonalisation où les facteurs des deux jeux ne sont pas appariés, l'usage de décomposer $\mathrm{pve}(X, Y)$ par rapport à F , au lieu de G , résulte d'une mauvaise généralisation.

Même dans le cas des facteurs principaux, où on pourrait attendre que les $\mathrm{pve}(X, f_h) \times \mathrm{pve}(f_h, Y)$ aient une signification, puisque leur somme donne un résultat intrinsèque, cette signification est trop compliquée pour être intéressante dans la pratique.

9. A dire vrai, on n'a jamais cherché dans $\mathrm{pve}(X, f_h) \times \mathrm{pve}(f_h, Y)$ quelque chose qui ait une signification, mais seulement une mesure pratique de l'importance d'un facteur par rapport à l'autre jeu. On a remarqué qu'en mesurant l'importance d'un facteur par rapport à l'autre jeu, ou importance "externe", par $\mathrm{pve}(f, Y)$, et en cherchant le facteur qui maximise cette quantité, on n'obtient pas une bonne mesure de l'interrelation, ou "chevauchement" (*overlap*), entre X et Y - car un facteur à forte $\mathrm{pve}(f, Y)$ a souvent peu d'importance "interne", c'est-à-dire de faibles corrélations avec son propre jeu. On a donc suggéré de pondérer

l'importance externe par l'importance interne, mesurée par $pve(X, f)$.

Mais cette procédure n'est qu'un emplâtre sur une jambe de bois, car $pve(f_h, Y) = ||f'_h||^2$ n'est même pas une bonne mesure de l'importance externe : elle peut être grande même si les y_j sont, dans leur ensemble, mal corrélés avec f_h, pourvu qu'il existe une combinaison linéaire des y_j qui soit proche de f_h. Une bonne mesure serait $pve(Y, f_h) = \frac{1}{q} \sum_j ||y'_j||^2$, la moyenne des carrés des corrélations entre f_h et les y_j. Notons que c'est précisément la contribution de f_h à la redondance totale $pve(Y, X) = \sum_h pve(Y, f_h)$ (cf. formule A §4).

Il n'est donc pas étonnant que le produit $pve(X, f_\ell) \times pve(f_\ell, Y)$ soit sans intérêt. Comme nous l'avons vu, il est simplement égal à $pve(X, g_\ell)$ pour les facteurs canoniques (ou autres facteurs mutuellement orthogonaux), de signification obscure pour les facteurs principaux, et sans aucune signification pour d'autres facteurs.

En général, la redondance de X, étant donné Y, est définie dans le cadre de la méthode des corrélations canoniques, et à l'aide de la formule B, et ensuite seulement on signale que le résultat est identique à celui de la formule plus simple A. La raison de ce détour est probablement historique : on s'est intéressé d'abord à la corrélation canonique (au carré) $pve(f_1, Y)$, qu'on a ensuite "corrigée" (§9) en un indice de redondance $pve(X, f_1) \times pve(f_1, Y)$, complété à son tour en $\sum_\ell [pve(X, f_\ell) \times pve(f_\ell, Y)] = B$, formule justifiée a posteriori par le théorème $B = A$. Mais il est en tout point préférable de définir directement A, indépendamment de la méthode des corrélations canoniques, et de reléguer B aux oubliettes.

10. En résumé, il semble recommandable, pour l'analyse bimultivariée, une fois extraits des facteurs orthogonaux, de ne pas calculer la décomposition de la formule B, mais de calculer les deux redondances totales comme

$$pve(Y, X) = \sum_h pve(Y, f_h) ,$$

$$pve(X, Y) = \sum_k pve(X, g_k) .$$

Ces calculs donnent comme résultat intermédiaire l'importance externe des facteurs,

$$\text{pve}(Y, f_h) = \sum_j y'^2_{jh} ,$$

$$\text{pve}(X, g_k) = \sum_i x'^2_{ik} ,$$

tandis que leur importance interne est donnée par

$$\text{pve}(X, f_h) = \sum_i x^2_{ih} ,$$

$$\text{pve}(Y, g_k) = \sum_j y^2_{jk} .$$

Il ne semble pas que cela ait un sens de multiplier l'importance externe par l'importance interne : il est plus clair de les garder séparées.

Cette procédure laisse libre de choisir la méthode d'orthogonalisation qui paraît le mieux convenir au problème considéré.

11. Effectuons ces calculs, pour notre exemple, dans le cas des facteurs principaux (F et G sont donc ici les P et Q du § 4).

Il faut calculer pour cela les sommes des colonnes des matrices $(X \cdot F)_2$, $(Y \cdot G)_2$, $(Y \cdot F)_2$, $(X \cdot G)_2$ (notation définie § 5), et diviser ces sommes par p (= 2) ou par q (= 3) selon le cas (tableau 3).

12. La remarque faite ci-dessus § 9 au sujet de pve(f, Y) nous amène à une autre remarque au sujet de l'analyse canonique en général. On a remarqué que cette méthode, maximisant pve(f, Y) et pve(g, X) , ne donne généralement pas des facteurs interprétables et utiles (4, p. 314). La raison n'en est pas qu'il s'agit d'une "maximisation purement mathématique" (ibid.) : une maximisation est toujours purement mathématique par nature, et pourtant elle peut donner des résultats interprétables ou utiles si le critère à maximiser est bien choisi. Les raisons sont plutôt celles-ci : 1^0 un facteur ne peut pas maximiser à la fois un critère d'importance interne et un facteur d'importance externe, de sorte qu'un facteur canonique n'a aucune raison d'être un facteur important dans son

Premier jeu X

Facteurs

f_1 f_2

Second jeu Y

Facteurs

g_1 g_2 g_3

Indices d'importance interne

.64	.36	
.64	.36	
1.28	.72	2
.64	.36	1

pve(X, f_h)

.80	.18	.02	
.97	.00	.02	
.27	.73	.00	
2.04	.91	.05	3
.68	.30	.02	1

pve(Y, g_k)

Indices d'importance externe

.49	.01	
.62	.02	
.26	.21	
1.37	.23	1.60
.46	.08	.53

pve(Y, f_h) pve(Y, X)

.61	.10	.07	
.27	.02	.02	
.88	.12	.10	1.10
.44	.06	.05	.55

pve(X, g_k) pve(X, Y)

Tableau 3

propre jeu; 2^0 pve(f, Y) est un mauvais critère d'importance externe, de sorte qu'un facteur canonique a peu de raisons d'être un facteur important par rapport à l'autre jeu.

C'est pourquoi l'objectif même de l'analyse canonique serait mieux servi par une méthode maximisant pve(Y, f) au lieu de pve(f, Y) [6]. C'est justement ce que propose van den Wollenberg (5, 6), et qu'il appelle "analyse de redondance". Cet auteur maximise la "redondance", non pas sous la forme $\text{pve}(Y, g_\ell) \times \text{pve}(g_\ell, X)$ - ce qui ne conduirait à rien puisque cette formule n'a de sens que si g_ℓ est un facteur canonique ou un facteur principal - mais sous la forme $\text{pve}(Y, f_\ell)$. Cette méthode ne donne pas des facteurs mutuellement orthogonaux (5); mais on peut, après n'avoir gardé que les facteurs dont la contribution est importante, effectuer une analyse canonique sur les dimensions retenues (6).

Il semble bien que le meilleur emploi de l'analyse canonique soit seulement comme *complément* d'une autre méthode. Le premier pas serait de choisir les dimensions qui expliquent un maximum de variance, soit dans leur propre jeu (analyse en composantes principales), soit dans l'autre jeu (analyse de redondance au sens de van den Wollenberg), selon qu'on veut des facteurs dont l'importance soit interne ou externe. Le second pas serait de transformer par rotation ces facteurs en d'autres mieux interprétables, ce qui peut se faire par des méthodes telles que les rotations Varimax, et peut-être aussi l'analyse canonique, cette dernière méthode ayant l'avantage (au moins pratique) de donner peu de corrélations non nulles entre les facteurs des deux jeux.

---

6. Cette remarque ne s'applique pas à l'analyse factorielle discriminante (considérée comme un cas particulier de l'analyse canonique). Dans cette méthode, en effet, on cherche une fonction de X qui soit aussi proche que possible d'une fonction constante sur les classes, ce qui correspond bien à maximiser la corrélation canonique.

BIBLIOGRAPHIE

(1) CAILLEZ F. et PAGÈS J.P., *Introduction à l'analyse des données*, Paris, SMASH, 1976.

(2) LEBART L. et FÉNELON J.P., *Statistique et informatique appliquées*, Paris, Dunod, 1971.

(3) MILLER J., *The development and application of bimultivariate correlation*, unpublished doctoral dissertation, State University of New York at Buffalo, 1969.

(4) MILLER J. and FARR S., "Bimultivariate redundancy : a comprehensive measure of interbattery relationship", *Multivariate Behavioral Research*, 6 (1971), 313-324.

(5) VAN DEN WOLLENBERG A., *Canonical redundancy analysis, an oblique alternative for canonical correlation analysis*, Report 75 MA 01, Université de Nimègues, 1975.

(6) VAN DEN WOLLENBERG A., *The way back to orthogonality : a rotational procedure for redundancy analysis*, Report 75 MA 08, Université de Nimègues, 1975.