REVUE FRANÇAISE D'AUTOMATIQUE, D'INFORMATIQUE ET DE RECHERCHE OPÉRATIONNELLE. RECHERCHE OPÉRATIONNELLE

SAMUEL SELLAM
ALAIN DUSSAUCHOY

Une mesure pour l'interprétation d'un ensemble de facteurs dans une analyse en composantes principales

Revue française d'automatique, d'informatique et de recherche opérationnelle. Recherche opérationnelle, tome 14, n° 1 (1980), p. 43-51.

<http://www.numdam.org/item?id=RO_1980__14_1_43_0>

R.A.I.R.O. Recherche opérationnelle/Operations Research
(vol. 14, n° 1, février 1980, p. 43 à 51)

UNE MESURE POUR L'INTERPRÉTATION D'UN ENSEMBLE DE FACTEURS DANS UNE ANALYSE EN COMPOSANTES PRINCIPALES (*)

par Samuel SELLAM (1) et Alain DUSSAUCHOY (1)

Résumé. — *On présente une généralisation de la notion de « contribution d'un facteur à la variance d'une variable » [1] au cas d'un ensemble de facteurs par rapport à un ensemble de variables et l'on démontre ensuite que cette contribution peut être exprimée en fonction du « Produit scalaire entre opérateurs » [4, 5, 3] d'Escoufier.*

Abstract. — *Here, the notion of "factor contribution to the variance of a variable" is generalized to the case of several factors with respect to several variables. It is shown that this contribution can be expressed in terms of the "scalar product between operators" (Escoufier [4, 5, 3]).*

I. RAPPEL DE LA NOTION DE « CONTRIBUTION D'UN FACTEUR A LA VARIANCE D'UNE VARIABLE »

Soit $\mathscr{X}$ un vecteur aléatoire à n dimensions constitué des n variables aléatoires centrées à analyser.

Soit V la matrice de variance-covariance du vecteur aléatoire $\mathscr{X}$.

Soit D_λ la matrice diagonale des valeurs propres de V et C sa matrice des vecteurs propres normalisés.

Soit $\mathscr{Y}$ le vecteur aléatoire des facteurs défini à partir du vecteur aléatoire $\mathscr{X}$ par l'équation

$$\mathscr{Y} = C^t \mathscr{X}. \tag{1}$$

La propriété d'orthogonalité (2) de la matrice C permet d'écrire (1) sous la forme

$$\mathscr{X} = C \mathscr{Y}. \tag{2}$$

(*) Reçu janvier 1979.

(1) Laboratoire des Méthodes informatiques appliquées à la Gestion, Université Claude-Bernard, Lyon-I.

(2) $C^t C = CC^t = I_n$ matrice unité.

R.A.I.R.O. Recherche opérationnelle/Operations Research, 0399-0842/1980/ 43 /$ 5.00

Soit X_p la p-ième variable aléatoire de $\mathscr{X}$, si on note $c_{.p}$ la p-ième ligne de C, on peut déduire de (2) l'expression de X_p en fonction de $\mathscr{Y}$:

$$X_p = c_{.p}\mathscr{Y}. \tag{3}$$

Considérons maintenant la variance de la variable aléatoire X_p :

$$\mathrm{Var}(X_p) = E[X_p X_p^t].$$

Du fait de (3) et de la linéarité de l'espérance mathématique on en déduit

$$\mathrm{Var}(X_p) = E[c_{.p}\mathscr{Y}\mathscr{Y}^t c_{.p}^t] = c_{.p} E[\mathscr{Y}\mathscr{Y}^t] c_{.p}^t. \tag{4}$$

Or, les variables aléatoires Y_j du vecteur $\mathscr{Y}$ sont non corrélées par construction donc

$$E[\mathscr{Y}\mathscr{Y}^t] = D_\lambda. \tag{5}$$

[Le *j-ième élément de* D_λ étant la variance de Y_j (Var (Y_j))].

On peut donc écrire (4) sous la forme

$$\mathrm{Var}(X_p) = \sum_{j=1}^{n} c_{jp}^2 \mathrm{Var}(Y_j). \tag{6}$$

L'équation (6) exprime la linéarité de la variance des variables en fonction des variances des facteurs. On peut donc en déduire que la contribution brute du facteur Y_l à la variance de X_p est la quantité $c_{lp}^2 \mathrm{var}(Y_l)$ et sa contribution relative notée $\mathscr{C}_{Y_l/X_p}$ sera donnée par

$$\mathscr{C}_{Y_l/X_p} = \frac{c_{lp}^2 \mathrm{Var}(Y_l)}{\sum_{j=1}^{n} c_{jp}^2 \mathrm{Var}(Y_j)} = \frac{c_{lp}^2 \mathrm{Var}(Y_l)}{\mathrm{Var}(X_p)}$$

II. « CONTRIBUTION D'UN ENSEMBLE DE FACTEURS A LA VARIANCE D'UN ENSEMBLE DE VARIABLES »

Soit $\mathscr{X}_l$ un vecteur aléatoire formé de l variables aléatoires du vecteur $\mathscr{X}$ et, soit $\mathscr{Y}_k$ un vecteur aléatoire formé de k facteurs du vecteur $\mathscr{Y}$.

Si on définit la quantité $\mathscr{I}_{\mathscr{X}_l}$ comme :

$$\mathscr{I}_{\mathscr{X}_l} = \sum_{j=1}^{l} \mathrm{Var}(X_j)$$

($\mathscr{I}$ *n'est autre qu'une généralisation de la variance* [3]).

Compte tenu de l'équation (3), on peut écrire :

$$\mathcal{I}_{\mathcal{X}_l} = \sum_{j=1}^{l} E[X_j X_j^t] = \sum_{j=1}^{l} E[c_{.j} \mathcal{Y}\mathcal{Y}^t c_{.j}^t].$$

Comme pour (4) la linéarité de l'espérance mathématique permet d'écrire :

$$\mathcal{I}_{\mathcal{X}_l} = \sum_{j=1}^{l} c_{.j} E[\mathcal{Y}\mathcal{Y}^t] c_{.j}^t.$$

Soit, compte tenu de (5) :

$$\mathcal{I}_{\mathcal{X}_l} = \sum_{j=1}^{l} \sum_{i=1}^{n} c_{ij}^2 \operatorname{var}(Y_i).$$

Si on regroupe les k facteurs de $\mathcal{Y}_k$ on a

$$\mathcal{I}_{\mathcal{X}_l} = \sum_{i=1}^{k} \sum_{j=1}^{l} c_{ij}^2 \operatorname{Var}(Y_i) + \sum_{i=k+1}^{n} \sum_{j=1}^{l} c_{ij}^2 \operatorname{Var}(Y_i).$$

Ainsi $\mathcal{I}_{\mathcal{X}_l}$ est réparti entre une contribution de $\mathcal{Y}_k$ et une contribution du complémentaire dans $\mathcal{Y}$ de $\mathcal{Y}_k$ (noté dans la suite $\mathcal{Y}_k^{\perp}$).

On peut donc en déduire l'expression de la « contribution de k facteurs à la variance de l variables » que l'on notera $\mathcal{C}_{\mathcal{Y}_k/\mathcal{X}_l}$:

$$\mathcal{C}_{\mathcal{Y}_k/\mathcal{X}_l} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{l} c_{ij}^2 \operatorname{Var}(Y_i)}{\sum_{j=1}^{l} \sum_{i=1}^{n} c_{ij}^2 \operatorname{Var}(Y_i)} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{l} c_{ij}^2 \operatorname{Var}(Y_i)}{\sum_{j=1}^{l} \operatorname{Var}(X_j)}.$$

Une autre expression de $\mathcal{C}_{\mathcal{Y}_k/\mathcal{X}_l}$ plus intéressante pour la suite peut être établie : pour cela il faut introduire les éléments suivants :

Soit B la sous-matrice $(n \times l)$ composée des l colonnes de la matrice C^t correspondant aux l variables de $\mathcal{X}_l$ et A la sous-matrice $(k \times l)$ composée des k lignes de B correspondant aux k facteurs de $\mathcal{Y}_k$.

Et si l'on convient que $D_{\lambda k}$ est la matrice $(k \times k)$ diagonale des valeurs propres de $\mathcal{Y}_k$, alors on pourra vérifier que :

$$\mathcal{C}_{\mathcal{Y}_k/\mathcal{X}_l} = \frac{\operatorname{Trace}(A^t D_{\lambda k} A)}{\operatorname{Trace}(B^t D_{\lambda} B)}.$$

III. PROPRIÉTÉS DE LA « CONTRIBUTION » $\mathcal{C}$

Le lecteur pourra aisément vérifier les trois propriétés suivantes :

III$_1$ = Si $k = n$ alors $\mathcal{C}_{\mathcal{Y}_k/\mathcal{X}_l} = 1$ *pour tout* $\mathcal{X}_l$;

III$_2$ = Si $l = n$ alors $\mathcal{C}_{\mathcal{Y}_k/\mathcal{X}_l} = \sum_{j=1}^{k} \lambda_j / \sum_{j=1}^{n} \lambda_j$;

Avec λ_j la j-ième valeur propre :

Nous reconnaissons en III_2 la formule de la part de variance expliquée par k facteurs :

$III_3 = \mathscr{C}_{\mathscr{G}_k / \mathscr{X}_i} + \mathscr{C}_{\mathscr{G}_k^{\perp} / \mathscr{X}_i} = 1.$

On dispose ainsi d'une mesure pour interpréter n'importe quel sous-ensemble de facteurs déduits d'une analyse en composantes principales et nous allons voir que cette mesure, si elle ne répond pas aux axiomes de la distance, peut toutefois s'exprimer en fonction du produit scalaire entre opérateurs d'Escoufier.

IV. RAPPEL DE LA MISE EN ŒUVRE DU PRODUIT SCALAIRE ENTRE OPÉRATEURS [3, 5]

Soit le tableau X des p caractères notés x^i mesurés sur le nuage I des n individus muni de la métrique M_1 et, soit le tableau Y des q caractères notés y^j mesurés sur le même nuage I muni de la métrique M_2.

On a le double schéma de dualité suivant : (pour plus de détails sur le formalisme employé *voir* [3] ou [5]) :

$$\begin{array}{ccccc} E_1 = R^p & \xleftarrow{X} & F^* & \xrightarrow{Y} & E_2 = R^q \\ \downarrow M_1 \uparrow V_{11} & & \downarrow W_1 \uparrow D_p \downarrow W_2 & & \uparrow V_{22} \downarrow M_2 \\ E_1^* & \xrightarrow[X']{} & F = R^n & \xleftarrow[Y']{} & E_2^* \end{array}$$

Avec F^*, E_1^*, E_2^* les espaces duaux de F (espace des caractères), E_1 (espace des individus repérés par les p caractères x^i), E_2 (espace des individus repérés par les q caractères Y^j) respectivement et :

- D_p la métrique diagonale des poids des n individus dans l'espace des caractères;
- $V_{11} = X D_p X^t$, $V_{12} = X D_p Y^t = V_{21}^t$, $V_{22} = Y D_p Y^t$;
- $W_1 = X^t M_1 X$, $U_1 = W_1 D_p$, $W_2 = Y^t M_2 Y$, $U_2 = W_2 D_p$.

La comparaison des triplets (X, M_1, D_p) et (Y, M_2, D_p) revient à la comparaison des opérateurs U_1 et U_2.

Nous rappelons que le produit scalaire entre opérateurs d'Escoufier s'exprime par

$$P(U_1, U_2) = \text{Trace}\,(U_1 U_2).$$

Si U_1 et U_2 sont les matrices associées aux opérateurs U_1 et U_2.

Nous rappelons également sans le démontrer un résultat qui nous sera utile

$$P(U_1, U_2) = \text{trace}\,(U_1 U_2) = \text{trace}\,(V_{21} M_1 V_{12} M_2).$$

V. LE SCHÉMA DE DUALITÉ D'UNE ANALYSE EN COMPOSANTES PRINCIPALES

Soit le tableau X de p caractères x^i mesurés sur le nuage I des n individus muni de la métrique M_1.

On rappelle qu'une analyse en composantes principales revient à calculer entre autres choses la matrice C des vecteurs propres d'une certaine matrice notée V_{11} des variances covariances des caractères.

Tenant compte de ce qui précède :

Soit le tableau $Y = C^t X$ ([3]) des p facteurs y^j « mesurés » sur le nuage I muni de la métrique M_2.

Nous aurons le schéma de dualité suivant :

$$\begin{array}{ccccc} E_1 = R^p & \xleftarrow{X} & F^* & \xrightarrow{Y = C^t X} & E_2 = R^p \\ M_1 \downarrow \uparrow V_{11} & & W_1 \downarrow \uparrow D_p \quad W_2 \downarrow \uparrow V_{22} & & M_2 \downarrow \uparrow \\ E_1^* & \xrightarrow[X^t]{} & F & \xleftarrow[Y^t]{} & E_2^* \end{array}$$

Notons que dans le cas d'une analyse en composantes principales V_{22} n'est autre que D_λ la matrice diagonale des valeurs propres donc, nous pourrons écrire en écriture matricielle

$$V_{11} = X D_p X^t = C D_\lambda C^t,$$

$$V_{22} = Y D_p Y^t = C^t X D_p X^t C = C^t V_{11} C = D_\lambda,$$

$$V_{12} = X D_p Y^t = X D_p X^t C = V_{11} C = C^t D_\lambda,$$

$$V_{21} = C^t V_{11} = D_\lambda C^t,$$

VI. PRODUIT SCALAIRE ENTRE OPÉRATEURS DÉFINIS SUR DES SOUS-ENSEMBLES DE CARACTÈRES

Restant dans le cadre d'une analyse en composantes principales, nous allons procéder à une projection de l'espace $E_1 = R^p$ des p variables sur le sous-espace $E_{10} = R^l$ correspondant à un sous-ensemble de l variables ($l \leqq p$) et de l'espace

([3]) Si les colonnes de C sont les vecteurs propres de V_{11} on a

$$X = C Y,$$

comme $C C^t = C^t C = I$, il en résulte que

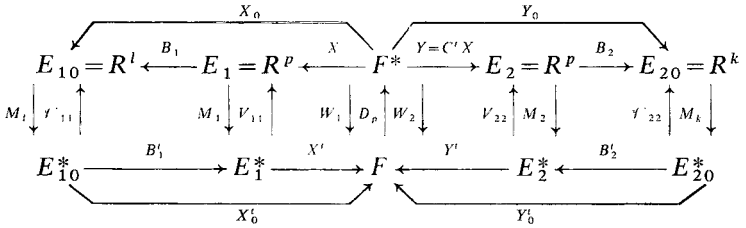
$$Y = C^t X.$$

$E_2 = R^p$ des p facteurs sur un sous-espace $E_{20} = R^k$ correspondant à un sous-ensemble de k facteurs ($k \leq p$). Pour cela nous disposons de :

B_1 : Matrice de projection de R^p sur R^l (B_1 sera une matrice diagonale dont une sous-matrice de rang l est une matrice unité et dont les autres éléments sont nuls).

B_2 : Matrice de projection de R^p sur R^k (B_2 est construit selon le même principe que B_1).

A l'aide de B_1 et B_2 (4), nous pouvons compléter le schéma de dualité du paragraphe V :



Dans ce schéma remarquons que :

$$\mathscr{V}_{11} = X_0 D_p X_0^t = B_1 X D_p X^t B_1^t = B_1 V_{11} B_1^t,$$

$$\mathscr{V}_{22} = Y_0 D_p Y_0^t = B_2 Y D_p Y^t B_2^t = B_2 C^t V_{11} C B_2^t = B_2 D_\lambda B_2^t,$$

$$\mathscr{V}_{12} = X_0 D_p Y_0^t = B_1 X D_p X^t B_2^t = B_1 V_{11} C B_2^t,$$

$$\mathscr{V}_{21} = B_2 C^t V_{11} B_1^t.$$

Nous pouvons de même qu'au paragraphe IV définir $\mathscr{U}_1$ et $\mathscr{U}_2$ comme suit

$$\text{si } \begin{cases} \mathscr{W}_1 = X_0^t M_l X_0 \\ \mathscr{W}_2 = Y_0^t M_k Y_0 \end{cases} \quad \text{alors } \begin{cases} \mathscr{U}_1 = \mathscr{W}_1 D_p, \\ \mathscr{U}_2 = \mathscr{W}_2 D_p \end{cases}$$

et poser le produit scalaire

$$P(\mathscr{U}_1, \mathscr{U}_2) = \text{Trace } (\mathscr{V}_{21} M_l \mathscr{V}_{12} M_k).$$

D'après les égalités précédentes, on en déduit que

$$P(\mathscr{U}_1, \mathscr{U}_2) = \text{Trace } (B_2 C^t V_{11} B_1^t M_l B_1 V_{11} C B_2^t M_k).$$

Or $V_{11} C = C D_\lambda$ de même $C^t V_{11} = D_\lambda C^t$.

Donc

$$P(\mathscr{U}_1, \mathscr{U}_2) = \text{Trace}(B_2 D_\lambda C^t B_1^t M_l B_1 C D_\lambda B_2^t M_k).$$

(4) Les notations sont les mêmes entre une application et la matrice qui lui est associée.

Si on pose $D_{\lambda k}$ la matrice déduite de D_λ dans laquelle on a annulé les $p-k$ éléments ne correspondant pas aux facteurs choisis et A la matrice déduite de la matrice C^t dans laquelle on a annulé les $p-k$ lignes ne correspondant pas aux facteurs choisis et les $p-l$ colonnes ne correspondant pas aux variables choisies, on peut vérifier que

$$B_2 D_\lambda C^t B_1^t = D_{\lambda k} A$$

et que

$$B_1 C D_\lambda B_2^t = A^t D_{\lambda k}.$$

On en déduit la formulation suivante du produit scalaire

$$P(\mathcal{U}_1, \mathcal{U}_2) = \text{Trace } (D_{\lambda k} A M_l A^t D_{\lambda k} M_k).$$

VII. EXPRESSION DE LA « CONTRIBUTION » EN FONCTION DU PRODUIT SCALAIRE ENTRE OPÉRATEURS

Avant de tenter de faire le lien entre la « contribution » $\mathcal{C}_{\mathcal{Y}_k/\mathcal{X}_l}$ et le produit scalaire $P(\mathcal{U}_1, \mathcal{U}_2)$, il faut faire les remarques suivantes :

1° $\mathcal{Y}_k$ et $\mathcal{X}_l$ étant des vecteurs aléatoires, nous devons passer à des réalisations de ces vecteurs : soit Y_k et X_l;

2° si on a une matrice quelconque T_1 à k lignes et k colonnes qui est une sous-matrice d'une matrice T_2 dans laquelle tous les autres éléments autres que ceux de T_1 sont nuls alors la trace de ces deux matrices reste inchangée : Trace (T_1) = trace (T_2). Nous pouvons donc confondre ces deux types de matrice au niveau de leur trace.

Ceci étant posé, si nous prenons pour métriques M_k et M_l les métriques suivantes :

$M_k = D_{\lambda k}^{-1}$ (soit la matrice des inverses des valeurs propres) et $M_l = I$ (matrice unité), alors :

$$P(\mathcal{U}_1, \mathcal{U}_2) = \text{Trace } (D_{\lambda k} A A^t),$$

soit :

$$P(\mathcal{U}_1, \mathcal{U}_2) = \text{Trace } (A^t D_{\lambda k} A).$$

D'autre part, exprimons $\mathcal{U}_1$:

$$\mathcal{U}_1 = \mathcal{W}_1 D_p = X_0^t M_l X_0 D_p = X^t B_1^t M_l B_1 X D_p.$$

Calculons la trace de $\mathcal{U}_1$ avec l'hypothèse $M_l = I$:

$$\text{Trace } (\mathcal{U}_1) = \text{Trace } (X^t B_1^t B_1 X D_p) = \text{Trace } (B_1 X D_p X^t B_1^t)$$
$$= \text{Trace } (B_1 V_{11} B_1^t) = \text{Trace } (B_1 C D_\lambda C^t B_1^t).$$

Si nous remarquons que le produit $C^t B_1^t$ n'est autre qu'une matrice qui contient la matrice B du paragraphe II et des éléments nuls partout ailleurs, nous pouvons écrire :

$$\text{Trace } (\mathcal{U}_1) = \text{Trace } (B^t D_\lambda B).$$

A ce point, nous disposons de tous les éléments pour exprimer $\mathcal{C}_{Y_k/X_l}$ en fonction des opérateurs $\mathcal{U}_1$ et $\mathcal{U}_2$:

$$\mathcal{C}_{Y_k/X_l} = \frac{\text{Trace } (\mathcal{U}_1\, \mathcal{U}_2)}{\text{Trace } (\mathcal{U}_1)} \qquad \text{soit} \quad \mathcal{C}_{Y_k/X_l} = \frac{P(\mathcal{U}_1, \mathcal{U}_2)}{\text{Trace } (\mathcal{U}_1)}.$$

VIII. COMMENTAIRES

On pourrait penser pouvoir établir la mesure symétrique de $\mathcal{C}_{\mathcal{Y}_k/\mathcal{X}_l}$ soit $\mathcal{C}_{\mathcal{X}_l/\mathcal{Y}_k}$. Nous ne détaillerons pas les calculs, mais si nous suivons la démarche du paragraphe II, nous obtenons l'expression de $\mathcal{C}_{\mathcal{X}_l/\mathcal{Y}_k}$ sous la forme

$$\mathcal{C}_{\mathcal{X}_l/\mathcal{Y}_k} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{l} c_{ij}^2 \text{ var}(X_j)}{\text{Trace } D_{\lambda k}} + \frac{TR}{\text{Trace } D_{\lambda k}}.$$

Dans cette expression TR désigne les termes rectangles dus au fait que $E(\mathcal{X}\,\mathcal{X}^t)$ n'est pas une matrice diagonale.

Si nous nous limitons comme mesure de la contribution cherchée à la quantité

$$\mathcal{C}_{\mathcal{X}_l/\mathcal{Y}_k} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{l} c_{ij}^2 \text{ var}(X_j)}{\text{Trace } D_{\lambda k}},$$

en négligeant les termes rectangles (pensant avoir là une sorte de contribution « pure » des variances des variables), on se heurte à des paradoxes : en effet, sous l'hypothèse de variables réduites, si l'on pose $k=1$ et $l=n$ on obtient la quantité

$$\mathcal{C}_{\mathcal{X}_l/\mathcal{Y}_k} = \frac{\sum_{j=1}^{n} c_{ij}^2}{\lambda_k} = \frac{1}{\lambda_k},$$

qui peut, s'il s'agit des premiers facteurs pour lesquels $\lambda_k \geqq 1$, être supérieur à l'unité; ce qui voudrait dire que la contribution de $\mathcal{X}_l$ à la variance de $\mathcal{Y}_k$ est supérieure à 100 % ! L'explication du phénomène est simple : TR peut être négatif.

Nous devons donc nous contenter de la mesure $\mathscr{C}_{y_k/x_i}$ qui, si elle présente des limites dues à la remarque précédente (qui montre que ce n'est pas une distance), a toutefois le gros intérêt d'être très suggestive. On peut penser à deux types d'applications :

– interprétation des résultats d'une analyse en composantes principales;

– validation de procédures telles que « l'échantillonnage dans une population de variables aléatoires réelles » [4] ou telles que la recherche de « séries chronologiques multiples résumées » [2].

On peut toutefois se ramener à un concept de distance au prix d'une certaine complexification (*voir* [6]).

BIBLIOGRAPHIE

1. A. Bonnafous, *La logique de l'investigation économétrique*, Dunod, Paris, 1973.
2. J.-M. Braun, *Étude des séries chronologiques multiples par l'Analyse des Données*, Rapport C.E.A.-R-4561, 1974.
3. F. Caillez et P. Pages, *Introduction à l'Analyse des Données*, S.M.A.S.H., Paris, 1976.
4. Y. Escoufier, *Échantillonnage dans une population de variables aléatoires réelles*, Thèse de doctorat d'État, Université de Montpellier, 1970.
5. J. P. Pages, Y. Escoufier et P. Cazes, *Opérateurs et analyse des tableaux à plus de 2 dimensions*, Cahier n° 25 du Bureau universitaire de recherche opérationnelle, Institut de Statistiques, Paris, 1976.
6. S. Sellam, *Un concept de proximité entre sous-espaces vectoriels et son application à l'Analyse des Données*. Thèse 3e Cycle, Université Lyon-I, 1979.