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

SAMUEL SELLAM
ALAIN FORCIOLI

Introduction de la notion d'écart entre sous-espaces vectoriels en analyse de données

Revue française d'automatique, d'informatique et de recherche opérationnelle. Recherche opérationnelle, tome 14, n° 3 (1980), p. 283-301.

<http://www.numdam.org/item?id=RO_1980__14_3_283_0>

R.A.I.R.O. Recherche opérationnelle/Operations Research
(vol. 14, n° 3, août 1980, p. 283 à 301)

INTRODUCTION DE LA NOTION D'ÉCART ENTRE SOUS-ESPACES VECTORIELS EN ANALYSE DE DONNÉES (*)

par Samuel SELLAM et Alain FORCIOLI [1]

Résumé. — *L'analyse de données se caractérise par l'étude de tableaux de valeurs. Mais l'analyse diffère suivant l'angle sous lequel on examine ces tableaux. Le présent article expose une vision « géométrique » dans laquelle on considère les données comme un ensemble de vecteurs d'un espace vectoriel : vouloir ainsi étudier de tels tableaux implique alors de pouvoir comparer entre eux des espaces ou des sous-espaces vectoriels.*

Pour ce faire, nous définissons un indice d'écart entre sous-espaces vectoriels d'un espace vectoriel euclidien de dimension finie et nous présentons des propriétés importantes de cet indice d'écart. Nous abordons ensuite la mise en œuvre de cet indice dans le cadre d'une analyse en composantes principales (ACP). Nous en donnerons ensuite une formulation générale dans l'espace des variables vectorielles avant de conclure sur un exemple numérique.

Abstract. — *If the data analysis is regarded in a geometrical point of view in which the datas are vectors in a vectorial space, a way to compare two or more data matrix may be to measure the distance between two vectorial subspaces.*

In this paper a "gap index" between vectorial subspaces of a normed vectorial space is defined together with important properties. Then a particular application to the principal component analysis is examined. A general formulation of this concept in the space of vectorial variables is finally produced. A numerical example follows as a conclusion.

I. INDICE D'ÉCART ENTRE DEUX SOUS-ESPACES VECTORIELS F ET G DE L'ESPACE E

En présence d'un nombre important, soit n, de variables mesurées par un certain nombre « d'objets », on peut chercher à réduire le nombre de variables en ne conservant que les variables les plus significatives, c'est-à-dire en gardant le maximum d'information. Nombre de travaux [JOLL, ESCO, BRAU] ont traité le sujet par des méthodes différentes. Mais la combinatoire importante

(*) Reçu janvier 1979.
([1]) Laboratoire Informatique Lyon-I, Villeurbanne.

de telles méthodes oblige à passer par des procédures de types hiérarchiques ne permettant pas en général d'atteindre l'optimum par rapport à la mesure choisie. C'est pourquoi il nous a paru intéressant de définir un modèle qui permette d'atteindre l'optimum réel.

L'analyse en composantes principales permet, en particulier, d'obtenir les « facteurs » dont on sait d'une part que chacun d'eux forme une combinaison linéaire des variables, d'autre part qu'ils sont ordonnés et que les « k premiers facteurs » emmagasinent le maximum d'information que l'on puisse avoir dans un espace de dimension k. Une méthode de sélection d'au plus q ($q < n$) variables peut alors consister à déterminer une mesure de la proximité entre le sous-espace vectoriel F engendré par les k premiers facteurs par exemple et les sous-espaces vectoriels $G\,p$ engendrés par p variables (p n'ayant pas une valeur fixe mais tel que $p \leqq q$). Il conviendra alors de retenir le sous-espace $G\,p$ qui minimise cette proximité tout en restant attentif à la combinatoire développée pour un tel problème afin de ne pas recourir à des procédures heuristiques ou hiérarchiques ne donnant pas en général l'optimum.

1. Indice d'écart. Définition

Soit $\mathscr{T}$ le treillis des sous-espaces vectoriels d'un espace vectoriel euclidien E de dimension n. Soient F et G deux éléments quelconques de ce treillis, F et G non nécessairement de même dimension. On définit

$$S_F = \{x \in F;\ \|x\| = 1\} \qquad \text{(de même } S_G\text{)},$$

soit alors

$$\theta(F, G) = \operatorname{Inf}\left[\operatorname*{Max}_{x \in S_F} d(x, G),\ \operatorname*{Max}_{y \in S_G} d(y, F)\right],$$

avec

$$d(x, G) = \|x - \operatorname{proj}_G x\|.$$

On a sur θ les deux propriétés évidentes suivantes :

- $\forall F \in \mathscr{T},\ \theta(F, F) = 0$;
- $\forall F \in \mathscr{T},\ \forall G \in \mathscr{T},\ \theta(F, G) = \theta(G, F)$.

Aussi dirons-nous que θ définit sur $\mathscr{T}$ un « indice d'écart » au sens de Bourbaki [BOUR].

Remarque 1 : On trouvera quelquefois dans la littérature [CAIL] la dénomination d'indice de dissimilarité au lieu d'indice d'écart.

Remarque 2 : Bien des auteurs [GLAZ], [KATO] ont déjà introduit une notion proche de l'indice d'écart pour mesurer la proximité entre éléments de $\mathscr{T}$. Ils définissent ainsi

$$\mu(F, G) = \operatorname{Sup}\left[\operatorname*{Max}_{x \in S_F} d(x, G), \operatorname*{Max}_{y \in S_G} d(y, F)\right].$$

Mais si μ définit sur $\mathscr{T}$ une *distance*, on démontre que (théorème de Bela sz Nagy) :

$$\dim F \neq \dim G \quad \Rightarrow \quad \mu(F, G) = 1.$$

Ainsi, si nous avions utilisé μ plutôt que θ, nous n'aurions pu comparer entre eux que des sous-espaces vectoriels de même dimension. Ce qui, dans certains cas, aurait soit empêcher de résoudre les questions posées, soit développer une combinatoire prohibitive. C'est pourquoi il nous a été nécessaire d'appauvrir dans un premier temps la notion de distance en notion d'indice d'écart. Mais nous verrons qu'en fait θ définit lui aussi une distance sur l'ensemble des sous-espaces vectoriels de même dimension et qu'il se révèle donc plus riche que μ pour le problème posé.

2. Propriété de l'indice d'écart

Nous énonçons dans ce paragraphe les théorèmes et propriétés fondamentaux sur θ. Pour ne pas alourdir notre propos, nous ne traiterons pas des démonstrations qui sont abondamment décrites dans [SELL 1].

THÉORÈME FONDAMENTAL :

- *si* $\dim F \leqq \dim G$ *alors* $\theta(F, G) = \operatorname*{Max}_{x \in S_F} d(x, G)$;
- *si* $\dim G \leqq \dim F$ *alors* $\theta(F, G) = \operatorname*{Max}_{y \in S_G} d(y, F)$.

COROLLAIRE : *Si* $\dim F = \dim G$ *alors* $\operatorname*{Max}_{x \in S_F} d(x, G) = \operatorname*{Max}_{y \in S_G} d(y, F)$.

PROPRIÉTÉ I : $\forall$ F et G sous-espaces vectoriels de E :

$$0 \leqq \theta(F, G) \leqq 1.$$

PROPRIÉTÉ II :

$$\left.\begin{array}{c}\theta(F, G) = 0 \\ \text{et } \dim F \leqq \dim G\end{array}\right\} \Leftrightarrow F \subseteq G.$$

THÉORÈME II : $\forall$ *F et G sous espaces vectoriels de E :*

$$\theta(F, G) = \theta(F^{\perp}, G^{\perp}).$$

PROPRIÉTÉ III : Si dim $G \leqq$ dim F et si G' est un sous-espace vectoriel de G, alors

$$\theta(F, G') \leqq \theta(F, G),$$

PROPRIÉTÉ IV (duale de III) : Si dim $F \leqq$ dim G et si G est un sous-espace vectoriel de G', alors : $\theta(F, G') \leqq \theta(F, G)$.

PROPRIÉTÉ V : Lorsque F et G sont deux sous-espaces vectoriels quelconques les deux propriétés suivantes ne sont pas vérifiées :

- $\theta(F, G) = 0 \Rightarrow F = G$;
- $\theta(F, G) \leqq \theta(F, H) + \theta(H, G)$ (inégalité triangulaire).

THÉORÈME III : *Si l'on désigne par $\mathscr{T}_l$ le sous-ensemble des sous-espaces vectoriels de dimension l, alors* θ *définit sur $\mathscr{T}_l$ une distance.*

La démonstration de ce dernier théorème est immédiate si l'on considère le corollaire énoncé et qu'on rapproche θ de μ qui définit sur $\mathscr{T}$, donc sur $\mathscr{T}_l$, une distance.

II. MISE EN ŒUVRE DE L'INDICE D'ÉCART DANS LE CADRE D'UNE ACP

1. Position relative des variables et des facteurs dans une ACP

On dispose d'un certain nombre d'observations mesurées sur n variables. Après avoir centré les variables, on estime la matrice V des variances-covariances.

La diagonalisation de la matrice V donne la matrice C des vecteurs propres. Cette matrice des vecteurs propres peut être interprétée comme la matrice de passage de la base initiale $B = (e_1, ..., e_n)$ (base des variables) à la base $B^* = (e_1^*, ..., e_n^*)$ (base des facteurs).

On peut considérer toute observation comme un vecteur d'un espace vectoriel euclidien E de dimension n. Soit alors X un vecteur de E. Les coordonnées

$$X = \begin{vmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{vmatrix} \quad \text{et} \quad Z = \begin{vmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{vmatrix},$$

de ce vecteur **X** respectivement dans les bases B^* et B sont liés par les relations :

$$X = C^{-1} Z, \tag{1}$$

$$Z = C.X. \tag{2}$$

Par ailleurs, puisque V est symétrique, la matrice C des vecteurs propres est une matrice orthogonale [CAIL]. On a donc la relation (si C^τ désigne la matrice transposée de la matrice C) :

$$C^\tau = C^{-1} \Leftrightarrow C.C^\tau = C^\tau.C = I_n, \tag{3}$$

soit alors G le sous-espace vectoriel de E, engendré par les l premiers vecteurs de la base B et soit $B_G = (e_1, ..., e_l)$ une base de G.

Soit F le sous-espace vectoriel de E, engendré par les k premiers vecteurs de la base B^* et soit $B_F = (e_1^*, ..., e_k^*)$ une base de F.

On supposera dans ce qui suit que $k \leqq l$.

Soit X_F un vecteur quelconque de F. Les coordonnées de ce vecteur sont données par

$$X_F = \begin{vmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{vmatrix} \text{ dans la base } B_F \quad \text{et} \quad X = \begin{vmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \\ 0 \\ \vdots \\ 0 \end{vmatrix} \text{ dans la base } B^*.$$

Les coordonnées de ce vecteur dans la base B sont, d'après (2) :

$$Z = C.X = \begin{vmatrix} z_1 \\ \vdots \\ z_n \end{vmatrix}.$$

Or, $x_{k+1} = \ldots = x_n = 0$. On peut donc écrire :

$$Z = M_0 X_F, \tag{4}$$

où M_0 est la matrice constituée par les k premières colonnes de la matrice C.

Par ailleurs

$$||\mathbf{X}_F||^2 = Z^\tau . Z = X_F^\tau M_0^\tau M_0 X_F. \tag{5}$$

Projetons alors le vecteur $\mathbf{X}_F$ sur le sous-espace G. Soit $\mathbf{Z}_G$ cette projection, Les coordonnées de $\mathbf{Z}_G$ dans la base B_G sont les l premières coordonnées du vecteur $\mathbf{X}_F$ dans la base B. Soit

$$Z_G = \begin{vmatrix} z_1 \\ \vdots \\ z_l \end{vmatrix}.$$

Or ces coordonnées sont obtenues dans (4) en ne retenant que les l premières lignes de la matrice Z. Soit alors A_0 la matrice constituée par les l premières lignes de la matrice M_0 :

$$Z_G = A_0 . X_F \tag{6}$$

et

$$||\mathbf{Z}_G||^2 = Z_G^\tau . Z_G = X_F^\tau . A_0^\tau . A_0 . X_F. \tag{7}$$

Enfin, puisque le vecteur $\mathbf{Z}_G$ est la projection orthogonale du vecteur $\mathbf{X}_F$ sur le sous-espace G, nous pouvons appliquer le théorème de Pythagore :

$$||\mathbf{X}_F||^2 = ||\mathbf{Z}_G||^2 + ||\mathbf{X}_F - \mathbf{Z}_G||^2 \Leftrightarrow ||\mathbf{X}_F - \mathbf{Z}_G||^2 = ||\mathbf{X}_F||^2 - ||\mathbf{Z}_G||^2$$
$$\Leftrightarrow ||\mathbf{X}_F - Z_G||^2 = X_F^\tau M_0^\tau M_0 X_F - X_F^\tau A_0^\tau A_0 X_F,$$

soit

$$||\mathbf{X}_F - \mathbf{Z}_G||^2 = X_F (M_0^\tau M_0 - A_0^\tau A_0) X_F \tag{8}$$

Si l'on veut donner une interprétation de la position relative de l variables par rapport à k facteurs, on peut alors utiliser toute fonctionnelle qui permette de mesurer l'écart entre les sous-espaces vectoriels F et G. Nous avons montré dans le paragraphe I que l'indice décart θ entre sous-espaces vectoriels remplit cette fonction.

On a considéré $k \leqq l$ soit dim $F \leqq$ dim G. Dans ces conditions l'indice d'écart entre F et G est donné par les formules (*cf.* § I) :

$$\theta(F, G) = \operatorname*{Max}_{X_F \in S_F} d(X_F, G) = \operatorname*{Max}_{X_F \in S_F} ||X_F - Z_G||,$$

si l'on désigne par S_F la boule unité du sous-espace F et par Z_G la projection du vecteur X_F sur le sous-espace G.

D'après la formule (8), maximiser $\| \mathbf{X}_F - \mathbf{Z}_G \|$ revient à maximiser $X_F^\tau (M_0^\tau M_0 - A_0^\tau A_0) X_F$ et le faire pour les vecteurs de S_F revient à imposer sur les vecteurs $\mathbf{X}_F$ la contrainte $\| \mathbf{X}_F \| = 1$, soit $X_F^\tau M_0^\tau M_0 X_F = 1$.

Rechercher la valeur de $\theta(F, G)$ revient donc à résoudre le problème suivant :

$$\left.\begin{array}{rl} \text{Maximiser} & X_F^\tau (M_0^\tau M_0 - A_0^\tau A_0) X_F \\ \text{avec} & X_F^\tau M_0^\tau M_0 X_F = 1 \end{array}\right\} \tag{9}$$

Or la matrice C étant orthogonale (soit $C^\tau . C = I_n$), on a

$$M_0^\tau . M_0 = I_k \tag{10}$$

En effet, la matrice C est la matrice constituée des n vecteurs colonnes $(Y_1, \ldots, Y_n)$ tels que

$$\forall i, \forall j \quad < yi, yj > = 0 \qquad \text{si} \qquad i \neq j$$

et

$$< yi, yj > = 1 \qquad \text{si} \qquad i = j.$$

La matrice M étant la sous-matrice de C constituée des k premiers vecteurs colonnes de C, l'égalité (10) est bien vérifiée.

La formulation (9) du problème est donc équivalente à la formulation suivante :

$$\left.\begin{array}{rl} \text{Maximiser} & X_F^\tau (I_k - A_0^\tau A_0) X_F \\ \text{avec} & X_F^\tau . X_F = 1. \end{array}\right\} \tag{11}$$

III. RÉSOLUTION DU PROBLÈME

La méthode des multiplicateurs de Lagrange [LEFE] donne :

$$\left\{\begin{array}{l} \dfrac{\partial}{\partial X_F} \left[X_F^\tau (I_k - A_0^\tau A_0) X_F - \mu (X_F^\tau . X_F - 1) \right] = 0 \\ \dfrac{\partial}{\partial \mu} \left[X_F^\tau (I_k - A_0^\tau A_0) X_F - \mu (X_F^\tau . X_F - 1) \right] = 0 \end{array}\right.$$

$$\Leftrightarrow \left\{\begin{array}{l} 2(I_k - A_0^\tau . A_0) X_F - 2\mu X_F = 0 \\ X_F^\tau . X_F = 1 \end{array}\right.$$

soit encore

$$\left.\begin{aligned}(I_k - A_0^\tau . A_0)\, X_F &= \mu\, X_F \\ X_F^\tau . X_F &= 1\end{aligned}\right\} \quad \text{et} \qquad (12)$$

On en conclut donc que le vecteur X_F de norme égale à 1 qui maximise $d^2(\mathbf{X}_F, G)$ est le vecteur propre $\mathbf{X}_{F_1}$ correspondant à la plus grande valeur propre μ de la matrice $I_k - A_0^\tau . A_0$.

Remarquons que la matrice $I_k - A_0^\tau . A_0$ étant définie non négative toutes ses valeurs propres sont positives ou nulles.

Dans ces conditions

$$d^2(\mathbf{X}_{F_1}, G) = X_{F_1}^\tau (I_k - A_0^\tau . A_0)\, X_{F_1} = \mu\, X_{F_1}^\tau . X_{F_1}$$

et, puisque : $X_{F_1}^\tau . X_{F_1} = 1$, on a

$$d^2(\mathbf{X}_{F_1}, G) = \mu.$$

Soit

$$\theta(F, G) = d(\mathbf{X}_{F_1}, G) = \sqrt{\mu}. \qquad (13)$$

Ainsi, pour calculer la valeur de l'indice d'écart entre les sous-espaces vectoriels F et G, il suffit de calculer la plus grande valeur propre de la matrice $I_k - A_0^\tau . A_0$.

IV. PROPRIÉTÉS DES MATRICES DU TYPE $A_0^\tau . A_0$

a) *Nouvelle formulation du problème*

D'une manière générale, si l'on désigne par M une matrice carrée quelconque, les matrices M et $I - M$ ont mêmes vecteurs propres et le vecteur propre X correspondant à la valeur propre μ de la matrice M correspondant à la valeur propre $\lambda = 1 - \mu$ de la matrice $I - M$.

Il est alors évident que si l'on range les valeurs propres de M dans l'ordre décroissant $[\mu_1 \geqq \mu_2 \ldots \geqq \mu_k]$, les valeurs propres correspondantes de $I - M$ sont rangées en ordre croissant $[\lambda_1 \leqq \lambda_2 \leqq \ldots \leqq \lambda_k]$.

Si l'on revient alors à notre problème, pour calculer la valeur de $\theta(F, G)$, on pourra calculer indifféremment la plus grande valeur propre de la matrice $I_k - A_0^\tau . A_0$ ou la plus petite valeur propre de la matrice $A_0^\tau . A_0$.

En d'autres termes, une formulation équivalente à la formulation (11) est la suivante :

$$\left.\begin{array}{c}\text{Minimiser} \quad X_F^{\tau} A_0^{\tau} . A_0 X_F \\ \text{avec} \quad X_F^{\tau} . X_F = 1.\end{array}\right\} \tag{14}$$

Le vecteur X_F cherché est alors le vecteur propre correspondant à la plus petite valeur propre λ de la matrice $A_0^{\tau} A_0$:

$$\theta(F, G) = \sqrt{1-\lambda}.$$

Remarque : Le raisonnement qui précède montre que toutes les valeurs propres de la matrice $I_k - A_0^{\tau} . A_0$ (ou de la matrice $A_0^{\tau} . A_0$) appartiennent à l'intervalle fermé [0, 1].

b) *Considérations sur les sous-matrices d'une matrice orthogonale*

Soit C une matrice orthogonale de dimension n et soit A_0, A_1, A_2, A_3 des sous-matrices de C représentées sur la figure 1.

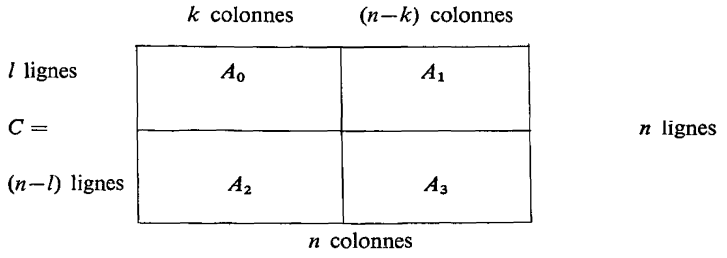


Figure 1.

Nous supposons dans tout ce qui suit que $k \leqq l$.

Nous allons dans ce qui suit exhiber un certain nombre de propriété sur les matrices A_0, A_1, A_2 et A_3. Nous ne mentionnerons pas les démonstrations qui sont abondamment décrites dans [SELL2].

PROPRIÉTÉ I : Les valeurs propres de la matrices $A_0^{\tau} . A_0$ sont valeurs propres de la matrice $A_0 . A_0^{\tau}$ et les valeurs propres de la matrice $A_0 . A_0^{\tau}$ qui ne sont pas valeurs propres de $A_0^{\tau} . A_0$ sont nulles.

LEMME : *Les matrices A_0 et A_1 vérifient l'égalité suivante :*

$$A_0 . A_0^{\tau} + A_1 A_1^{\tau} = I_l.$$

PROPRIÉTÉ II : Si l'on désigne par μi ($i = 1$ à k) les valeurs propres rangées dans l'ordre croissant de la matrice $I_k - A_0^\tau . A_0$ et par δi ($i = 1$ à l) les valeurs propres rangées dans l'ordre croissant de la matrice $A_1 . A_1^\tau$, on a les égalités :

(*a*) $\mu i = \delta i$ pour $i = 1$ à k;

(*b*) $\delta i = 1$ pour $i = k+1$ à l.

PROPRIÉTÉ III : Les matrices $I_k - A_0^\tau . A_0$ et $I_{n-l} - A_3 . A_3^\tau$ ont les mêmes plus grandes valeurs propres. Toutes les valeurs propres de l'une, qui ne sont pas valeurs propres de l'autre (du fait des dimensions respectives de ces matrices) sont nulles.

Remarque : On a vu que le calcul de la plus grande valeur propre de la matrice $I_k - A_0 . A_0^\tau$ fournit la valeur de $\theta(F, G)$. De même le calcul de la plus grande valeur propre de $I_{n-l} - A_3 . A_3^\tau$ fournit la valeur de $\theta(F, G)$. La propriété III démontre que $\theta(F, G) = \theta(F^\perp, G^\perp)$ propriété déjà démontrée.

Des calculs similaires sur les matrices A_1 et A_2 permettent, en prenant en compte les dimensions de ces matrices, de calculer $\theta(F, G^\perp)$ et $\theta(F^\perp, G)$.

V. NOUVELLE PROPRIÉTÉ DE L'INDICE D'ÉCART

THÉORÈME : *Soit F et G deux sous-espaces vectoriels d'un espace euclidien E, F étant engendré par un sous-ensemble de facteurs et G par un sous-ensemble de variables d'une analyse en composantes principales. Si la dimension de G est supérieure ou égale à la dimension de F, il existe alors un sous-espace vectoriel G' du sous-espace G tel que*

$$\dim G' = \dim F \quad \textit{et} \quad \theta(F, G') = \theta(F, G).$$

Posons $\dim F = k$ et $\dim G = l$ (avec $k \leqq l$). Désignons par $\lambda_1 \ldots \lambda_k$ les valeurs propres rangées en ordre décroissant de la matrice $A_0^\tau . A_0$ [cf. *fig.* 1], On sait alors (prop. I) que $\lambda_1 \ldots \lambda_k$ sont valeurs propres de $A_0 . A_0^\tau$ et que les $l-k$ autres valeurs propres de cette matrice sont nulles

$$\lambda_{k+1} = \ldots = \lambda_l = 0.$$

Désignons par $\mathbf{X}_1, \ldots, \mathbf{X}_l$ les vecteurs propres associés à la matrice $A_0 . A_0^\tau$ ils sont aussi vecteurs propres de la matrice $I_l - A_0 . A_0^\tau$ et sont associés aux valeurs propres $\mu_1 \ldots \mu_l$ telles que

$$\forall i \in [1, l], \quad \mu_i = 1 - \lambda_i.$$

Les valeurs propres de la matrice $I_k - A_0^\tau . A_0$ sont alors $\mu_1, \ldots, \mu_k$, et ces valeurs sont rangées en ordre croissant. On a alors

$$\theta(F, G) = \sqrt{\mu_k}.$$

Soit alors G' le sous-espace vectoriel engendré par les vecteurs $\mathbf{X}_1 \ldots \mathbf{X}_k$, On a alors dim G' = dim $F = k$.

Par ailleurs, puisque $(\mathbf{X}_1, \ldots, \mathbf{X}_l)$ forment une base de G, G' est un sous-espace vectoriel de G. Calculons alors l'écart entre le sous-espace G' et le sous-espace F. Soit $\theta\,(F, G')$.

Puisque dim G' = dim F, on a

$$\theta(F, G) = \underset{\mathbf{Y} \in S_{G'}}{\text{Max}}\, d(\mathbf{Y}, F) = \underset{\mathbf{Y} \in S_{G'}}{\text{Max}} \left|\left| \mathbf{Y} - \text{proj}_F \mathbf{Y} \right|\right|.$$

Nous ne pouvons reproduire directement sur le sous-espace G' les raisonnements effectués au paragraphe I.1. Mais, étant donné que G' est un sous-espace de G, nous pouvons faire un raisonnement identique sur G en imposant aux vecteurs étudiés d'appartenir à G'.

Ainsi, rechercher $\theta\,(F, G')$ c'est résoudre le problème :

$$\text{Maximiser} \quad Y^\tau (I_l - A_0 A_0^\tau) Y$$
$$\text{avec} \quad Y^\tau Y = 1 \quad \text{et} \quad Y \in G'$$

Le vecteur $\mathbf{Y}$ cherché est alors le vecteur propre de la matrice $I_l - A_0 A_0^\tau$ parmi les vecteurs propres de cette matrice qui sont dans G', associé à la plus grande valeur propre. On a donc $\mathbf{Y} = \mathbf{X}_k$ et

$$\theta(F, G') = \left|\left| \mathbf{X}_k - \text{proj}_F \mathbf{X}_k \right|\right| = \theta(F, G).$$

C.Q.F.D.

VI. UN ALGORITHME POUR LE CHOIX DE VARIABLES REPRÉSENTATIVES

VI.1. Principales phases de l'algorithme

L'algorithme se décompose en trois phases principales :

a) Une analyse en composantes principales sur un tableau de n variables préalablement centrées. Nous ne détaillerons pas cette phase, le problème étant abondamment traité dans un certain nombre d'ouvrages [LEBF, LEMT, BERT].

b) Une phase qui traite le problème du choix des k facteurs et détermine ainsi le sous-espace F. Nous n'aborderons pas dans cet article ce problème des critères de choix car ils sont inhérents au problème traité et aux objectifs poursuivis. Tout au plus, peut-on signaler que cette phase devra être rendue interactive pour faciliter le dialogue entre le décideur et le système.

c) Une phase, que nous allons traiter plus en détail, de détermination du sous-espace G des variables qui minimise l'indice d'écart entre F et les sous-espaces de variables possibles.

VI.2. Détermination du sous-espace des variables

a) *Règle de simplification*

Il est bien évident que la combinatoire développée par la troisième phase de l'algorithme risque d'être prohibitive eu égard aux temps d'exécution. C'est pourquoi il paraît nécessaire de déterminer des règles de simplification afin d'alléger cette combinatoire.

Si l'on désigne par G le sous-espace engendré par les variables $(e_1, \ldots, e_\rho)$ avec ρ inférieur ou égal à k, et par G_1 le sous-espace engendré par $(e_1, \ldots, e_{j-1}, e_{j+1}, e_\rho)$, on a

$$\theta(G_1, F) \leqq \theta(G, F) \qquad \text{(propriété III, § I).}$$

En vertu de cette propriété, lorsqu'on aura déterminé en cours d'algorithme, la plus petite valeur de $\theta(F, G)$ jusqu'alors obtenue, on pourra éliminer de la combinatoire sur les variables toutes les variables e_i pour lesquelles $\theta(e_i, F)$ est supérieur à cette plus petite valeur, puisque pour tout espace G_1, engendré par e_i en particulier, on aura

$$\theta(F, G_1) \geqq \theta(e_i, F) > \theta(F, G).$$

Il convient de noter, avant d'aller plus avant, qu'une telle règle de simplification ne peut s'appliquer que si la dimension l du sous-espace G considéré est inférieur ou égal à k. Si $l > k$, on calculera alors, pour se retrouver dans les conditions d'application de cette règle $\theta(F^\perp, G^\perp)$ dont on sait qu'il est égal à $\theta(F, G)$.

b) *L'algorithme en phase* 3

Nous considérons que la dimension du sous-espace G cherché est comprise entre deux bornes m et M. Nous savons par ailleurs que, pour tous sous-espaces F et G, $\theta(F, G) \leqq 1$. Cette propriété nous permet d'initialiser la valeur de θ cherchée, soit S_θ, à 1.

Un organigramme général de l'algorithme en phase 3 est alors donné en figure 2 (nous y désignons par $e_1, \ldots, e_n$ les variables et par l la dimension du sous-espace G).

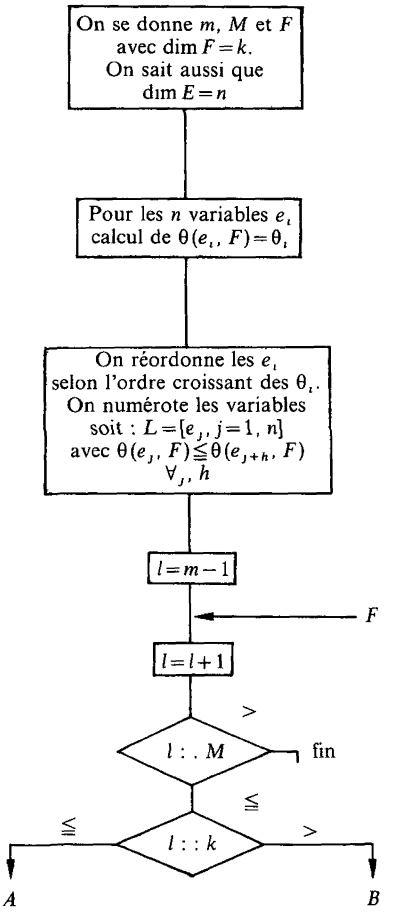


Figure 2. — L'algorithme en phase 3.

L'organigramme suivant permet, pour une plage de dimensions fixée ($l \in [m, M]$) et pour un sous-espace F constant de facteurs, de donner pour chaque valeur de l le « meilleur » sous-espace G_l (dim $G_l = l$) de variables au sens de $\theta(G_l, F)$ (G_l et F sont sous-espaces de E).

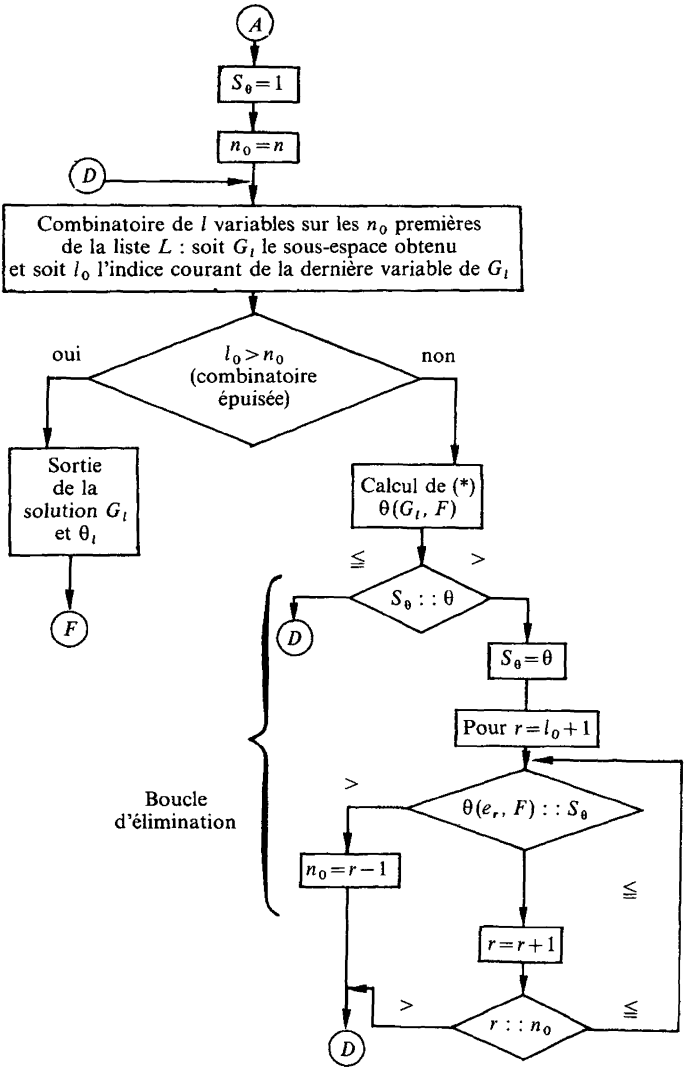


Figure 2. — L'algorithme en phase 3 (suite).

La partie B se traite de manière analogue à la partie A mais sur $F^\perp$ et $G^\perp$: au « meilleur » $G^\perp_{n-l}$ par rapport à $F^\perp_{n-k}$ correspondra la solution G_l par rapport à F_k (on a $G^\perp_{n-l} + G_l = E$ et $F^\perp_{n-k} + F_k = E$).

(*) Ce calcul peut faire l'objet d'une optimisation par la prise en compte des propriétés de sous-matrices citées précédemment.

Il est à noter qu'il faut dans B considérer non $\theta(e_j, F)$ mais $\theta(e_j, F^{\perp})$ en se souvenant que $\theta(e_j, F^{\perp}) + \theta(e_j, F) = 1$.

VI.3. Application de θ à l'analyse des données sur variables vectorielles

Ce paragraphe n'a pour but que de montrer la portée très générale de la notion d'indice d'écart établie au début de cet article. Seuls quelques résultats sont exposés et l'on trouvera un développement plus complet dans [SELL3].

Considérons deux variables vectorielles V et W avec

$$V = V_1, V_2, \ldots, V_k,$$

$$W = W_1, W_2, \ldots, W_l \qquad \text{avec} \quad k \leqq l.$$

Supposons que ces variables soient mesurées sur le même échantillon $\mathscr{E}$ de N individus, on peut alors toujours dire que $V_1, V_2, \ldots, V_k$ $W_1, W_2, \ldots, W_l$ sont des vecteurs de l'espace $\mathscr{E} = R^N$. Par conséquent V et W peuvent être considérés comme deux sous-espaces vectoriels du même espace R^N. Le problème posé revient à trouver une formulation de l'indice d'écart θ entre V et W.

Si on note F et G respectivement les matrices $(k \times N)$ et $(l \times N)$ associés aux variables vectorielles V et W mesurées sur $\mathscr{E}$ que l'on munit de la métrique M et convenant d'autre part de noter $T(B)$ la plus grande valeur propre d'une matrice B, on a alors

$$\theta(V, W) = \sqrt{T(B)},$$

avec

$$B = I_k - (F^{\tau} MF)^{-1} F^{\tau} MG (G^{\tau} MG)^{-1} G^{\tau} MF.$$

Avec ce résultat on dispose d'une mesure entre variables vectorielles et l'on peut imaginer d'appliquer la plupart des techniques classiques d'Analyse des Données à ce type très général de variables.

On peut à ce sujet faire deux remarques :

Remarque 1 : θ définit sur l'espace des variables vectorielles de même dimension une distance. On aura donc dans cet espace :

$$\theta(V_1, V_2) = 0 \quad \Leftrightarrow \quad V_1 = V_2,$$

$$\theta(V_1, V_2) = \theta(V_2, V_1,),$$

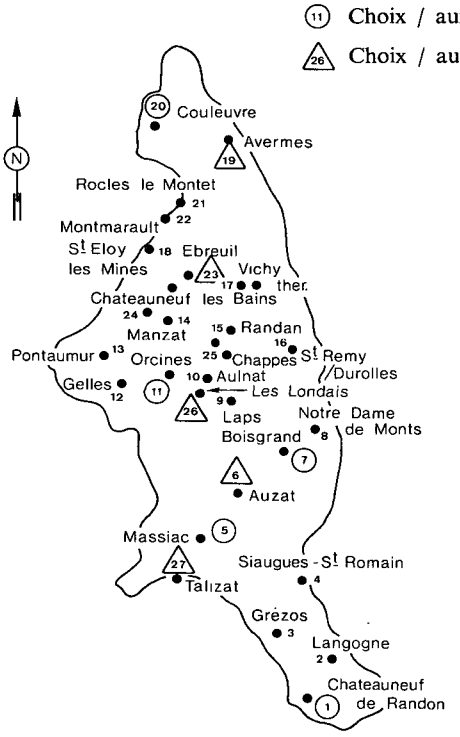
$$\theta(V_1, V_2) \leqq \theta(V_1, V_3) + \theta(V_3, V_2).$$

Remarque 2 : Appliqué à des variables vectorielles unidimensionnelles on peut exprimer θ en fonction du coefficient de corrélation classique. En effet on pourra vérifier que l'on a la relation

$$\theta^2(V_1, V_2) = 1 - \rho^2(V_1, V_2).$$

VI.4. Une illustration du problème de sélection : le choix de stations pluviométriques

Si l'on reprend les données utilisées par Escoufier dans [ESCO] concernant les hauteurs de chutes de pluie de 27 stations situées dans le bassin versant de l'Allier, on peut vouloir pour des raisons économiques par exemple choisir certaines d'entre elles.



Bassin versant de l'Allier.

Deux stratégies sont possibles :

– choisir des variables (ici les stations) qui présentent le minimum de redondance et le maximum d'informations (forte variance);

	Première stratégie					Deuxième stratégie					Ensemble de variables				
Variables sélectionnées.....	1	5	7	11	20	9	10	19	21	23	Toutes				
Matrice de corrélation.....		1	5	7	11	20		9	10	19	21	23			
	1						9								
	5	0.379					10	0.843							
	7	0.490	0.270				19	0.710	0.658						
	11	0.429	0.275	0.648			21	0.657	0.529	0.935					
	20	0.335	0.167	0.572	0.655		23	0.690	0.684	0.923	0.878				
Moyenne des corrélations..	0,422					0,751					0,605				
Variance des variables.....	1580	843	1640	1990	1340	607	333	636	985	625					
Valeurs propres :											5 premières valeurs propres				
— brutes.....	4402.5	1191	667	605	526	2633	368.5	110	44	31	17185	3539	1177	1004	921
— %.....	59.54	16.12	9.02	8.19	7.12	82.63	11.56	3.45	1.38	0.97	63.14	13.0	4.32	3.69	3.38
Conditionnement (*)..	0.346					0.1084					0				

(*) Il s'agit du conditionnement de la matrice X des observations mesurées sur les variables centrées. On a cond $(X) = \sqrt{\rho(V)/\sigma(V)}$ avec $V = X^{\tau} X$ la matrice des variances-covariances, $\rho(V)$ = plus petite valeur propre de V, $\sigma(V)$ = plus grande valeur propre de V. Ce conditionnement donne le degré de singularité de la matrice V (*voir* [ROB, FORS], ...).

– choisir des variables qui présentent une bonne interdépendance et qui soient relativement stables dans leur comportement (faible variance).

Il ne sera pas présenté ici de comparaisons avec les résultats obtenus par Escoufier car la signification des différences n'est pas encore complètement établie. On peut éventuellement les envisager en fonction des deux stratégies dont nous présentons ici les résultats comparatifs pour le choix de 5 stations parmi 27.

La première stratégie consiste logiquement à rechercher la combinaison la plus proche d'un sous-espace vectoriel formé de premiers facteurs alors que la seconde correspond à un calcul par rapport à un sous-espace de derniers facteurs.

On a choisi ici de prendre d'un côté les 7 premiers facteurs ([1]) soit environ 90 % de l'inertie totale, et de l'autre les 20 restants.

L'analyse des résultats montre que la première stratégie correspond à des variables bien réparties géographiquement et qui présentent une interdépendance faible, alors que la deuxième conduit à un choix de stations présentant une situation géographique limitée à la moitié nord de la zone de référence et une interdépendance relativement élevée.

BIBLIOGRAPHIE

[BERT] P. BERTIER et J. M. BOUROCHE, *Analyse des données Multidimensionnelles*, P.U.F., 1975.

[BOUR] N. BOURBAKI, *Topologie Générale*, chap. 9, § 1, TG 9 1.

[BRAU] J. M. BRAUN, *Étude des séries chronologiques multiples par l'Analyse des données*, Rapport CEA-R 4561, 1974.

[CAIL] F. CAILLIEZ et P. PAGES, *Introduction à l'Analyse des données*, S.M.A.S.H., 1976.

[ESCO] Y. ESCOUFIER, *Échantillonnage dans une population de variables aléatoires réelles*, Thèse de doctorat d'État, Université de Montpellier, 1970.

[FORS] G. FORSYTHE et C. B. MOLER, *Computer Solution of Linear Algebraic Systems*, Prentice-Hall, Inc., 1967.

[GANT] F. R. GANTMACHER, *Théorie des matrices*, Dunod, Paris, 1966.

[GLAZ] GLAZMAN et LIUBITCH, *Analyse linéaire dans les espaces de dimension finie*, Éditions de Moscou, 1972.

[JOLL] I. T. JOLLIFFE, *Discording Variables in Principal Component Analysis I : Artificial Data*, Appl. Statist., vol. 21, p. 160-173.

[KATO] J. KATO, *Perturbation Theory for Linear Operations*, Springer-Verlag, 1976.

([1]) On note généralement une certaine stabilité des solutions en fonction des facteurs choisis. Ici par exemple le résultat est identique pour 5, 6 et 7 facteurs.

[LEBF] L. LEBART et J. P. FENELON, *Statistique et Informatique Appliquées*, Dunod, Paris, 1971.

[LEFE] J. LEFEBVRE, *Introduction aux Analyses Statistiques Multidimensionnelles*, Masson, Paris, 1976.

[LEMT] L. LEBART, A. MORINEAU et N. TABARD, *Techniques de la description statistiques*, Dunod, Paris, 1977.

[ROB] F. ROBERT, *Méthodes Itératives pour des Systèmes d'équations linéaires*, Polycopie Analyse numérique, Université Lyon-I.

[SELL1] S. SELLAM et A. FORCIOLI, *Notion d'écart entre sous espaces vectoriels*, Rapport interne, Laboratoire Informatique de Lyon-I, 1979.

[SELL2] S. SELLAM et A. FORCIOLI, *Mise en œuvre de l'indice d'écart dans une A.C.P.* Rapport interne, Laboratoire Informatique Lyon-I, 1979.

[SELL3] S. SELLAM, A. FORCIOLI et J. AZENCOT, *Une distance entre variables vectorielles de même dimension*, Journées de statistiques Malakoff, 28-31 mai 1979.