REVUE DE STATISTIQUE APPLIQUÉE

Y. ESCOUFIER

Les liaisons entre groupes d'aléas

Revue de statistique appliquée, tome 19, n° 2 (1971), p. 5-17

<http://www.numdam.org/item?id=RSA_1971__19_2_5_0>

LES LIAISONS ENTRE GROUPES D'ALÉAS

Y. ESCOUFIER

L'auteur précise les liens étroits qui existent entre l'analyse canonique telle qu'elle est présentée dans les traités classiques d'analyse statistique multidimensionnelle et l'analyse factorielle des correspondances développée par M. Benzecri et ses collaborateurs. L'explicitation du contexte mathématique sous-jacent à ces méthodes permet une généralisation à des aléas vectoriels, chose qui n'a pas été abordée jusqu'à ce jour. L'exposé comprend des exemples détaillés pour chacune des trois méthodes.

I - L'ANALYSE CANONIQUE

I_1. Soit X un vecteur aléatoire partitionné en deux sous-vecteurs X_1 et X_2 $\left(X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}\right)$ possédant respectivement p et q composantes. Nous supposerons dans la suite de l'exposé que le vecteur X est centré.

L'analyse canonique suppose que les composantes X_1^i (i = 1, ..., p) de X_1 et X_2^j (j = 1, ..., q) de X_2 sont des fonctions de carrés sommables sur un espace probabilisé $(\Omega, \mathcal{F}, P)$.

Nous verrons plus loin l'intérêt qu'il peut y avoir à ce que le vecteur X soit un vecteur gaussien sur $(\Omega, \mathcal{F}, P)$; mais cette hypothèse plus forte, nécessaire pour l'inférence statistique, est superflue dans un premier temps si l'on se limite à la présentation algébrique du problème.

Soient L_X^2 l'espace de Hilbert des combinaisons linéaires des composantes de X muni du produit scalaire $\langle U, V \rangle = E(UV)$, et $L_{X_1}^2$ et $L_{X_2}^2$ les sous-espaces de L_X^2 engendrés respectivement par les composantes de X_1 et celles de X_2.

Le probleme que veut résoudre l'analyse canonique est de trouver les couples (U, V), $U \in L_{X_1}^2$, $V \in L_{X_2}^2$, $||U|| = ||V|| = 1$, pour lesquels $E(UV)$ est maximum ou, ce qui revient au même, tels que l'angle entre U et V soit minimum.

I_2. On peut remarquer ([4]) que pour U donné le vecteur V qui fait avec U un angle minimum est colinéaire à la projection $P_1(U)$ de $U \in L_{X_1}^2$ dans $L_{X_2}^2$. Donc, pour U donné :

$$V = \frac{P_1(U)}{||P_1(U)||}$$

De la même manière, pour V donné, U est colinéaire à la projection $P_2(V)$ de $V \in L^2_{X_2}$ dans $L^2_{X_1}$ si bien que pour V donné :

$$U = \frac{P_2(V)}{||P_2(V)||}$$

Il s'ensuit que les couples (U, V) solutions du problème posé sont tels que :

$$V = \frac{P_1 \circ P_2(V)}{||P_1(U)|| \, ||P_2(V)||} \qquad \text{et} \qquad U = \frac{P_2 \circ P_1(U)}{||P_1(U)|| \, ||P_2(V)||}$$

c'est dire que :

Les couples (U, V) solution du problème posé sont tels que U soit vecteur propre de l'opérateur $P_1 \circ P_2$ associé à la valeur propre $||P_1(U)|| \, ||P_2(V)||$ et V le vecteur propre de $P_2 \circ P_1$ associé à la même valeur propre.

Insertion n° 1

Remarque $I_{2.a}$: La transformation $P_1 \circ P_2$ est une transformation symétrique de L_{X_2} dans lui-même. La transformation $P_2 \circ P_1$ est une transformation symétrique de $L^2_{X_1}$ dans lui-même.

On a en effet dans L^2_X, pour $U \in L^2_{X_1}$ et $V \in L^2_{X_2}$:

$$< U, V > = < P_1(U), V > = < U, P_2(V) >$$

D'où :

$$\begin{aligned} < U, P_2 \circ P_1(U') > &= < U, P_1(U') > \\ &= < P_1(U), P_1(U) > \\ &= < P_1(U), U' > \\ &= < P_2 \circ P_1(U), U' > \end{aligned}$$

Il en découle que si $L^2_{X_1}$ et $L^2_{X_2}$ sont finis, l'opérateur $P_1 \circ P_2$ a des valeurs propres positives et des vecteurs propres deux à deux orthogonaux.

Remarque $I_{2.b}$: P_1 et P_2 sont des opérateurs de projection si bien que :

$$\{||U|| = ||V|| = 1\} \Longrightarrow \{0 \leqslant ||P(U)|| \; ||P(V)|| \leqslant 1\}$$

Les valeurs propres de $P_1 \circ P_2$ et de $P_2 \circ P_1$ sont comprises entre 0 et 1.

Remarque $I_{2.c}$: Les opérateurs $P_2 \circ P_1$ et $P_1 \circ P_2$ ne peuvent avoir des valeurs propres égales à l'unité que si :

$$L^2_{X_1} \cap L^2_{X_2} \neq 0$$

Le nombre de valeur propre égale à l'unité sera égal à la dimension de $L^2_{X_1} \cap L^2_{X_2}$.

Remarque $I_{2.d}$. Si U est le vecteur propre de $P_2 \circ P_1$ associé à la valeur propre λ, le vecteur propre V correspondant à la même valeur propre est égal à $P_1(U)$. En effet :

$$\{P_2 \circ P_1(U) = \lambda U\} \Longrightarrow \{P_1 \circ P_2(P_1(U)) = \lambda P_1(U)\} .$$

Insertion n° 2

Le premier couple $< U_1, V_1 >$ solution du problème est donc fourni par les vecteurs propres de $P_2 \circ P_1$ et $P_1 \circ P_2$ associés à la plus grande valeur propre différente de l'unité. Si l'on cherche un second couple $< U_2, V_2 >$ tel que U_2 soit orthogonal à U_1 et V_2 à V_1, on trouvera le couple des vecteurs propres associés à la valeur propre immédiatement inférieure et ainsi de suite.

I_3. Le calcul explicite des valeurs propres et des vecteurs propres des opérateurs $P_1 \circ P_2$ et $P_2 \circ P_1$, conduit à chercher les matrices associées à ces opérateurs dans les bases X_1 et X_2.

Pour cela, nous devons écrire que pour tout élément X_1^i de la base X_1, le vecteur $X_1^i - P_1(X_1^i)$ est orthogonal à tout élément X_2^j de la base X_2. Soit :

$$\forall i = 1, \ldots, p \; ; \; \forall j = 1, \ldots, q : < X_1^i - P_1(X_1^i), \; X_2^j > = 0$$

$P_1(X_1^i)$ est un élément de $L^2_{X_2}$ que l'on peut écrire : $\sum_{k=1}^{q} M_k^i \, X_2^k$.

D'où, les $p \times q$ égalités précédentes deviennent :

$$\forall i = 1, \ldots, p \; ; \; \forall j = 1, \ldots, q : < X_1^i, X_2^j > - \sum_{k=1}^{q} M_k^i < X_2^k, X_2^j > = 0 .$$

Si on appelle Σ_{12} la matrice d'élément $(\Sigma_{12})_{ij} = < X_1^i, X_2^j >$; Σ_{22} la matrice d'élément $(\Sigma_{22})_{k,j} = < X_2^k, X_2^j >$ et M la matrice dont la $i^{ième}$ ligne est formée des M_k^i; les $p \times q$ égalités précédentes se résument en l'égalité matricielle :

$$\Sigma_{12} - M \; \Sigma_{22} = 0$$

D'où :

$$M = \Sigma_{12} \; \Sigma_{22}^{-1} .$$

Convenons d'appeler Σ_{21} la transposée de Σ_{12} , M est la matrice des coefficients des $P_1(X_1^i)$, les coordonnéees de $P_1(X_1^i)$ étant rangées dans la ligne i. On en déduit que la matrice associée à P_1 est :

$$^t(\Sigma_{12} \; \Sigma_{22}^{-1}) = \Sigma_{22}^{-1} \; \Sigma_{21} .$$

De même, en appelant Σ_{11} la matrice d'élément $(\Sigma_{11})_{ij} = < X_1^i, X_1^j >$ on établirait que la matrice associée à P_2 est $\Sigma_{11}^{-1} \; \Sigma_{12}$.

Il en découle que la matrice associée à $P_1 \circ P_2$ est $\Sigma_{22}^{-1} \; \Sigma_{21} \; \Sigma_{11}^{-1} \; \Sigma_{12}$ et la matrice associée à $P_2 \circ P_1$: $\Sigma_{11}^{-1} \; \Sigma_{12} \; \Sigma_{22}^{-1} \; \Sigma_{21}$.

Remarque $I_{3.a}$: Dans l'utilisation traditionnelle de l'analyse canonique, les statisticiens se contentent de donner les valeurs propres de $P_1 \circ P_2$ appelées corrélations canoniques et les vecteurs propres appelés variables canoniques. Il nous paraît intéressant de suggérer de représenter les X_2^j dans l'espace $L^2_{X_2}$ rapportés aux vecteurs propres de $P_1 \circ P_2$ ainsi que les $P_1(X_1^i)$. Si on appelle Σ_{12}^i la $i^{ième}$ ligne de Σ_{12}, d'après ce qui précède, on a :

$$P_1(X_1^i) = \Sigma_{12}^i \; \Sigma_{22}^{-1} X_2$$

si bien que la proximité des points représentants X_1^i et X_2^j est liée à l'importance de X_2^j dans la régression de X_1^i par rapport à X_2.

De plus :

$$\begin{aligned} ||P_1(X_1^i) - P_1(X_1^k)||^2 &= < P_1(X_1^i) - P_1(X_1^k),\ P_1(X_1^i) - P_1(X_1^k) > \\ &= < (\Sigma_{12}^i - \Sigma_{12}^k)\Sigma_{22}^{-1} X_2,\ {}^t X_2 \Sigma_{22}^{-1}\, {}^t(\Sigma_{12}^i - \Sigma_{12}^k) > \\ &= (\Sigma_{12}^i - \Sigma_{12}^k)\, \Sigma_{22}^{-1}\, {}^t(\Sigma_{12}^i - \Sigma_{12}^k) \end{aligned}$$

Là encore, il apparaît que les points représentant $P_1(X_1^i)$ et $P_2(X_1^k)$ seront d'autant plus proches que les variables X_1^i et X_1^k ont des covariances avec X_2^j voisines.

I_4. Nous avons supposé jusqu'à maintenant que les matrices Σ_{12}, Σ_{21}, Σ_{11} et Σ_{22} étaient connues. Ce n'est pas toujours le cas et le plus souvent, elles doivent être estimées. L'hypothèse d'un vecteur X gaussien est alors interessante puisqu'elle nous permet d'obtenir le résultat suivant : ([1] p. 298).

Appelons Σ la matrice $\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$ et S son estimation du maximum de vraisemblance. Alors dans le cas où toutes les valeurs propres de $\Sigma_{11}^{-1}\, \Sigma_{12}\, \Sigma_{22}^{-1}\, \Sigma_{21}$ sont différentes, les vecteurs propres normes sont définis de manière unique si l'on s'astreint à leur imposer une première composante positive. Ceci permet d'assurer que les estimations du maximum de vraisemblance des vecteurs propres et des valeurs propres de $\Sigma_{11}^{-1}\, \Sigma_{12}\, \Sigma_{22}^{-1}\, \Sigma_{21}$ sont ceux de $S = S_{11}^{-1}\, S_{12}\, S_{22}^{-1}\, S_{21}$.

I_5. Exemple.

Pour illustrer ce que nous venons de dire, nous prendrons les donnée présentées par Anderson ([1] p. 303). Le lecteur pourra ainsi comparer l'intérêt de notre approche par rapport à la présentation habituelle du problème. L'étude porte sur les deux premiers fils d'un échantillon de 25 familles.

Le vecteur X_1 a deux composantes ; X_1^1 est la longueur de la tête du fils aîné ; X_1^2 sa largeur. X_2 a également deux composantes ; X_2^1 est la longueur de la tête du fils cadet ; X_2^2 sa largeur. Les calculs sont faits sur les variables centrées réduites. La matrice S est la suivante :

$$S = \left(\begin{array}{c|c} S_{11} & S_{12} \\ \hline S_{21} & S_{22} \end{array}\right) = \left(\begin{array}{cc|cc} 1,0000 & 0,7346 & 0,7108 & 0,7040 \\ 0,7346 & 1,0000 & 0,6932 & 0,7086 \\ \hline 0,7108 & 0,6932 & 1,0000 & 0,8392 \\ 0,7040 & 0,7086 & 0,8392 & 1,0000 \end{array}\right)$$

D'où :

$$S_{22}^{-1}\, S_{21} = \begin{pmatrix} 0,4058 & 0,3332 \\ 0,3635 & 0,4290 \end{pmatrix}$$

$$S_{11}^{-1} S_{12} = \begin{pmatrix} 0,4379 & 0,3986 \\ 0,3714 & 0,4157 \end{pmatrix}$$

et

$$S_{22}^{-1} S_{21} S_{11}^{-1} S_{12} = \begin{pmatrix} 0,3014 & 0,3001 \\ 0,3184 & 0,3231 \end{pmatrix}$$

La première valeur propre de cette matrice est $\lambda_1 = 0,6215$. Le vecteur propre Y_1 qui lui est associé tel que $E(Y_1^2) = 1$ est :

$$Y_1 = 0,5045\ X_2^1 + 0,5383\ X_2^2 .$$

De la même manière le vecteur propre Y_2 tel que $E(Y_2^2) = 1$ associé à la seconde valeur propre $\lambda_2 = 0,0029$ est :

$$Y_2 = 1,7680\ X_2^1 - 1,7586\ X_2^2 .$$

En inversant la matrice $\begin{pmatrix} 0,5045 & 0,5383 \\ 1,7680 & -1,7586 \end{pmatrix}$, on obtient :

$$X_2^1 = 0,9564\ Y_1 + 0,2927\ Y_2$$

$$X_2^2 = 0,9614\ Y_1 - 0,2743\ Y_2$$

et en utilisant ${}^t(S_{22}^{-1} S_{21})$:

$$P_1(X_1^1) = 0,8141\ Y_1 + 0,0429\ Y_2$$

$$P_1(X_1) = 0,7311\ Y_1 - 0,0201\ Y_2$$

Dans ce cas très simple, l'interprétation n'est pas particulièrement riche ; on pourra noter cependant que les points représentant X_2^1 et X_2^2 sont placés à une distance de l'origine des axes égale à leur variance (ici 1,0000). Les points représentant $P_1(X_1)$ et $P_2(X_1^2)$ sont placés de part et d'autres de l'axe Y_1 et très près de celui-ci. Cette dernière proximité s'explique par l'analogie qu'il y a entre les troix équations :

$$Y_1 = 0,5045\ X_2^1 + 0,5383\ X_2^2$$

$$P_1(X_1^1) = 0,4858\ X_2^1 + 0,3635\ X_2^2$$

$$P_2(X_1^2) = 0,3332\ X_1^2 + 0,4290\ X_2^2$$

$P_1(X_1^1)$ est placé du même côté de X_2^1 car X_2^1 intervient plus dans la régres- de $P_1(X_1^1)$ que X_2^2. Il en va de même pour $P_2(X_1^2)$ et X_2.

On notera enfin que la distance de $P_1(X_1^1)$ à l'origine des axes représente la part de variation expliquée par X_1^2 et X_2^2 dans la variation totale de X_1^1. La même remarque vaut pour $P_1(X_1^2)$.

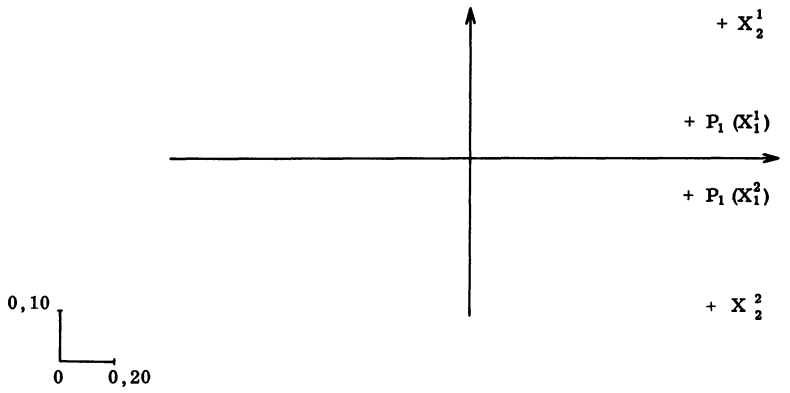
Remarque : En multipliant a gauche le vecteur $\begin{pmatrix} 0,5045 \\ 0,5383 \end{pmatrix}$ qui définit le premier vecteur de $S_{22}^{-1} S_{21} S_{11}^{-1} S_{12}$ par $S_{11}^{-1} S_{12}$ on obtient le premier vecteur propre de $S_{11}^{-1} S_{12} S_{22}^{-1} S_{21}$, soit :

$$\tilde{Z}_1 = 0,4355\ X_1^1 + 0,4111\ X_1^2 .$$

En posant alors $Z_1 = \frac{\tilde{Z}_1}{11\tilde{Z}.11}$, on a vérifié que :

$$E(Y_1\ Z_1) = \sqrt{\lambda_1}.$$

Analyse canonique



II - L'ANALYSE FACTORIELLE DES CORRESPONDANCES

II 1. L'analyse factorielle des correspondances peut être présentée comme un cas particulier de l'analyse canonique dans lequel Ω est un ensemble produit de deux ensembles finis $I \times J$; I à p éléments ; J à q éléments ; $\Omega = I \times J$ est probabilisé par la donnée des probabilités p(i ,j) des couples (i ,j) pour $i = 1, \ldots, p$; $j = 1, \ldots, q$.

Le vecteur X_1 est le vecteur des fonctions δ_i définies sur I par :

$$\begin{cases} \delta_i(k) = 0 & \text{si} \quad k \neq i \\ \delta_i(k) = 1 & \text{si} \quad k = i \end{cases}$$

De même le vecteur X_2 est le vecteur des fonctions η_j définies sur J par :

$$\begin{cases} \eta_j(k) = 0 & \text{si} \quad k \neq j \\ \eta_j(k) = 0 & \text{si} \quad k = j \end{cases}$$

Explicitons dans ce cas particulier les éléments des différentes matrices.

$$(\Sigma_{11})_{i,k} = < \delta_i, \delta_k > = \sum_{l=1}^{p} \sum_{h=1}^{q} \delta_i(l)\ \delta_k(h)\ p(l\ ,h)$$

d'où :

$$(\Sigma_{11})_{i,k} = p(i) \quad \text{si} \quad i = k$$

$$(\Sigma_{11})_{i,k} = 0 \quad \text{si} \quad i \neq k$$

$$(\Sigma_{12})_{i,j} = < \delta_i , \eta_j > = \sum_{k=1}^{p} \sum_{l=1}^{q} \delta_i(k)\ \eta_j(l)\ p(k,l) = p(i,j) \ .$$

On en déduit :

$$(\Sigma_{11}^{-1}\ \Sigma_{12})_{i,j} = \frac{p(i,j)}{p(i)}$$

et

$$(\Sigma_{11}^{-1}\ \Sigma_{12})_{i,j} = \frac{p(i,j)}{p(j)}$$

D'où :

$$(\Sigma_{22}^{-1}\ \Sigma_{21}\ \Sigma_{11}^{-1}\ \Sigma_{12})_{j_1,j_2} = \sum_{i=1}^{p} \frac{p(i,j_1)\ p(i,j_2)}{p(i)\ p(j_1)} \ .$$

II_2. En particulier $P_1(\delta_1) = \Sigma_{12}^{i}\ \Sigma_{22}^{-1}\ X_2 = \sum_{j=1}^{q} \frac{p(i,j)}{p(j)}\ \eta_j$.

Ainsi, le point représentant $P_1(\delta_1)$ sera le centre de gravité des points représentant les η_j affectés des masses :

$$\frac{p(i,j)}{p(j)} \ .$$

D'autre part :

$$||P_1(\delta_i) - P_1(\delta_k)||^2 = (\Sigma_{12}^{i} - \Sigma_{12}^{k})\,\Sigma_{22}^{-1}\ {}^{t}(\Sigma_{12}^{i} - \Sigma_{12}^{k})$$

$$= \sum_{j=1}^{q} \frac{(p(i,j) - p(k,j))^2}{p(j)}$$

$$= \sum_{j=1}^{q} p(j)\ \left(\frac{p(i,j)}{p(j)} - \frac{p(k,j)}{p(j)}\right)^2 \ .$$

Remarque $II_{2.a}$: Si au lieu de projetter les éléments δ_i et δ_k on effectuait la projection de $\frac{\delta_i}{p(i)}$ et $\frac{\delta_i}{p(k)}$ on obtiendrait :

$$||P_1\left(\frac{\delta_i}{p(i)}\right) - P\ \left(\frac{\delta}{p(k)}\right)||^2 = \sum_{j=1}^{q} p(j)\ \left(\frac{p(i,j)}{p(i)\ p(j)} - \frac{p(k,j)}{p(k)\ p(j)}\right)^2$$

Ce résultat a été le but des travaux de M. Benzécri et de ses collaborateurs. Il les a conduit à proposer l'analyse factorielle des correspondances, méthode

dont il faut souligner l'intérêt puisqu'elle permet de prendre en compte des ensembles I et J d'éléments qui n'étaient pas abordables par les méthodes statistiques classiques.

Remarque $II_{2.b}$: Dans le cas particulier que nous étudions, l'ensemble $L^2_{X_1} \cap L^2_{X_2}$ n'est pas vide : il contient les fonctions constantes sur $I \times J$. Nous aurons donc une valeur propre égale à 1 dont nous ne tiendrons pas compte.

II_3. Exemple : Nous traiterons un exemple très simple qui nous permettra de mettre en évidence les différents calculs à effectuer. Soit J un ensemble à 2 éléments ; I un ensemble à trois éléments ; sur une population de 50 individus, les éléments de I et de J sont apparus selon la table de correspondances suivantes :

J \ I	i_1	i_2	i_3
j_1	1	2	17
j_2	14	13	3

On établit sans difficultés que :

$$\Sigma_{22}^{-1}\,\Sigma_{21} = \begin{pmatrix} 1/20 & 2/20 & 17/20 \\ 14/30 & 13/30 & 3/30 \end{pmatrix}$$

$$\Sigma_{11}^{-1}\,\Sigma_{12} = \begin{pmatrix} 1/15 & 14/15 \\ 2/15 & 13/15 \\ 17/20 & 3/20 \end{pmatrix}$$

D'où :

$$\Sigma_{22}^{-1}\,\Sigma_{21}\;\Sigma_{11}^{-1}\,\Sigma_{12} = \begin{pmatrix} 0,7392 & 0,2608 \\ 0,1739 & 0,8261 \end{pmatrix}$$

Le premier vecteur propre associée à la valeur propre 1 est sans intérêt (Remarque $II_{2.b}$).

La seconde valeur propre est égale à 0,5652. Il lui est associé le vecteur propre $Y = 1,2243\ \eta_1 - 0,8168\ \eta_2$. On vérifie aisément que $||Y|| = 1$.

La base des η_j est une base orthogonale ; donc la coordonnée de η_1 sur Y est 1,2243 et la coordonnée de η_2 est − 0,8168. On en déduit les coordonnées de $P_1(\delta_1)$ sur Y. En effet,

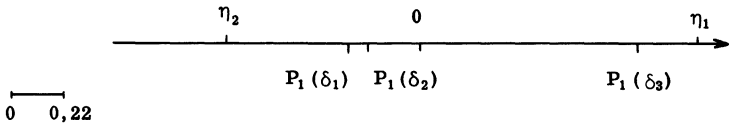
$$P_1(\delta_1) = \Sigma_{12}\;\Sigma_{22}\,X_2 = \frac{1}{20}\,\eta_1 + \frac{14}{30}\,\eta_2$$

et la coordonnée sur Y est donc :

$$\frac{1,2243}{20} - \frac{14 \times 0,8168}{30} = -0,3200$$

On établit de la même manière que la coordonnée de $P_1(\delta_2)$ sur Y est $-0,2314$ et celle de $P_1(\delta_3)$: 1,0747. Les points se placent donc sur l'axe Y dans un ordre conforme aux relations qui existent entre les éléments des ensembles I et J.

Analyse des correspondances



III - GENERALISATION A DES ALEAS VECTORIELS

III$_1$. Nous nous proposons dans ce paragraphe de généraliser l'analyse canonique à des aléas vectoriels. Au lieu de considérer comme en I des vecteurs X_1 et X_2 dont les composantes X_1^i et X_2^j sont des variables aléatoires centrées sur $(\Omega, \mathcal{F}, P)$, nous supposerons maintenant que X_1 et X_2 sont des familles de vecteurs aléatoires centrés sur $(\Omega, \mathcal{F}, P)$; c'est dire que les X_1^i et X_2^j sont des vecteurs aléatoires. Les composantes des vecteurs seront supposées de variances finies ($X^{i,k} \in L^2(\Omega, \mathcal{F}, P)$).

Soit $X^{i,k}$, la $k^{ième}$ composante du $i^{ième}$ vecteur de la famille α ; ($\alpha = 1, 2$; $i = 1, \ldots, n_\alpha$; $k = 1, \ldots, n_\alpha^i$). Dans un travail précédent [5], nous avons associé au vecteur X_α^i, l'opérateur V_α^i de $L^2(\Omega, \mathcal{F}, P)$ en lui-même, defini par :

$$V_\alpha^i(Y) = \frac{1}{n_\alpha^i} \sum_{k=1}^{n_\alpha^i} E(X_\alpha^{i,k} Y) X_\alpha^{i,k}.$$

L'intérêt de cet opérateur réside, pour une part, dans les propriétés suivantes, dont on trouvera les démonstrations dans le travail cité.

Propriété III$_{1.a}$: L'opérateur V_α^i a les mêmes vecteurs propres que l'opérateur Λ_α^i défini par la matrice des variances et covariances du vecteur X_α^i; ses valeurs propres sont égales à celles de Λ_α^i divisées par n_α^i.

Dans la mesure donc, où les éléments propres de Λ_α^i sont caractéristiques de X_α^i, V_α^i est caractéristique de ce dernier vecteur.

Propriété III$_{1.b}$: L'opérateur V_α^i appartient à la classe des opérateurs de Hilbert-Schmidt (la somme des carrés des valeurs propres est finie) et, sur cette classe, on peut définir un produit scalaire qui la munit d'une structure d'espace de Hilbert.

En particulier, on en déduit une distance entre opérateur qui est nulle si et seulement si les deux opérateurs en cause ont les mêmes éléments propres.

Propriété $III_{1.c}$: Pour deux opérateurs V_α^i et V_β^j, ce produit scalaire peut s'écrire :

$$< V_\alpha^i , V_\beta^j > = \frac{\sum_{k=1}^{n_\alpha^i} \sum_{l=1}^{n_\beta^j} [E (X_\alpha^{i,k} X_\beta^{j,l})]^2}{n_\alpha^i n_\beta^j}$$

Appelons H la classe des opérateurs de Hilbert-Schmidt. En identifiant les X_α^i aux opérateurs V_α^i qui leur sont associés, nous pouvons énoncer le problème de l'analyse canonique pour des aléas vectoriels de la manière suivante :

Trouver U de norme unité dans le sous-espace de H engendré par les V_1^i ($i = 1, \ldots, n_1$) et V de norme unité dans le sous-espace de H engendré par les V_2^j ($j = 1, \ldots, n_2$) tel que $< U , V >$ soit maximum.

Ce problème qui est exactement le même que celui que nous avons posé en I nous conduira aux mêmes solutions à savoir les éléments propres de la matrice $\Sigma_{22}^{-1} \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12}$ dans laquelle :

$$(\Sigma_{11})_{kl} = < V_1^k , V_1^l >$$

$$(\Sigma_{12})_{kl} = (\Sigma_{21})_{lk} = < V_1^k , V_2^l >$$

et :

$$(\Sigma_{22})_{kl} = < V_2^k , V_2^l >$$

III_2. Exemple. Pour illustrer ce paragraphe, nous avons retenu une famille X_1 de trois vecteurs X_1^1, X_1^2 et X_1^3 ayant chacun trois composantes ; X_2 est une famille de trois vecteurs à deux composantes chacuns. X_α^i ($\alpha = 1, 2$; $i = 1, 2, 3$) est le vecteur des notes obtenues par un étudiant pour la matière i pendant l'année α.

Nous nous proposons d'étudier globalement les liaisons entre les ensembles de notes X_2^1, X_2^2, X_2^3 et leur influence sur les ensembles X_1^1, X_1^2, X_1^3 sur la base d'une population de 110 étudiants.

La matrice Σ est donnée par la table $III_{2.a}$:

Table $III_{2.a}$

	X_1^1	X_1^2	X_1^3	X_2^1	X_2^2	X_2^3
X_1^1	0,3918					
X_1^2	0,0678	0,4051				
X_1^3	0,0746	0,0667	0,4227			
X_2^1	0,1603	0,0782	0,0645	0,6235		
X_2^2	0,0508	0,0198	0,0190	0,0285	0,5040	
X_2^3	0,1276	0,1349	0,1471	0,0558	0,0189	0,5964

On en déduit successivement :

$$\Sigma_{22}^{-1}\ \Sigma_{21} = \begin{pmatrix} 0,2366 & 0,1050 & 0,0811 \\ 0,0803 & 0,0253 & 0,0242 \\ 0,1893 & 0,2156 & 0,2383 \end{pmatrix}$$

$$\Sigma_{11}^{-1}\ \Sigma_{12} = \begin{pmatrix} 0,3767 & 0,1219 & 0,2355 \\ 0,1199 & 0,0256 & 0,2535 \\ 0,0614 & 0,0177 & 0,2434 \end{pmatrix}$$

$$\Sigma_{22}^{-1}\ \Sigma_{21}\ \Sigma_{11}^{-1}\ \Sigma_{12} = \begin{pmatrix} 0,1067 & 0,0330 & 0,1020 \\ 0,0348 & 0,0109 & 0,0312 \\ 0,1118 & 0,0328 & 0,1572 \end{pmatrix}$$

La table $III_{2.b}$ donne les vecteurs propres normés au sens du produit scalaire que nous avons défini :

V_1	V_2	V_3
0,7285	0,9346	0,4953
0,2314	0,3838	- 1,3316
0,9571	- 0,8839	- 0,0875

Il faut comprendre : $V_1 = 0,7285\ X_2^1 + 0,2314\ X_2^2 + 0,9571\ X_2^3$.

Notons que les valeurs propres sont respectivement 0,2510 ; 0,0238 et 0. On peut en déduire déjà que la liaison entre les deux familles de vecteurs est assez faible.

En inversant la matrice des coefficients des V_i sur les X_2^j, on obtient les xoefficients des X_2^j sur les V_i. La table $III_{2.c}$ les fournit :

Table $III_{2.c}$

X_2^1	X_2^2	X_2^3
0,5212	0,1435	0,6134
0,5343	0,2362	- 0,4638
0,2445	- 0,6578	- 0,0271

Il faut comprendre : $X_2^1 = 0,5212\ V_1 + 0,5343\ V_2 + 0,2445\ V_3$.

On peut également calculer les coordonnées de $P_1(X_1^i)_{i=1,2,3}$ sur les axes V_1, V_2, V_3.

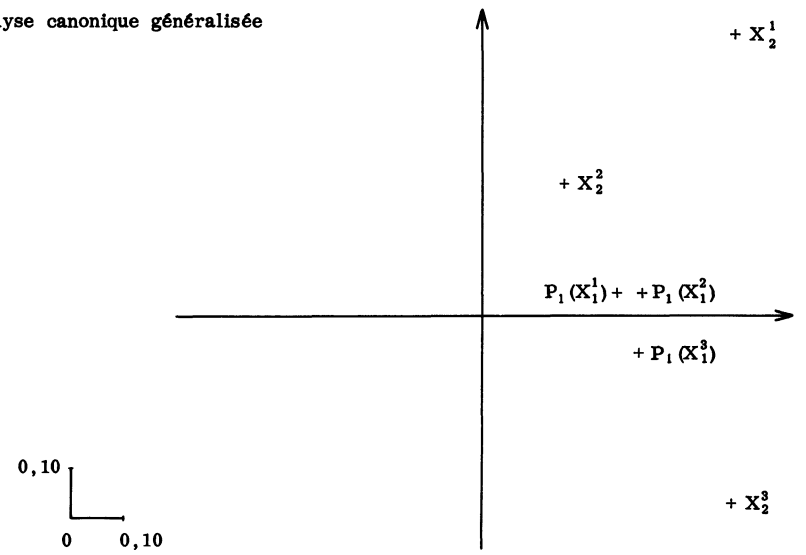
Pour les deux premiers axes, les résultats sont consignés dans la table $III_{2.d}$:

Table $III_{2.d}$

	$P_1(X_1^1)$	$P_1(X_1^2)$	$P_1(X_1^3)$
V_1	0,2510	0,1906	0,1919
V_2	0,0576	0,0379	- 0,0615

Que peut-on déduire d'une représentation des points correspondant aux X_2^j et au $P_1(X_1)$ dans un plan rapporté à V_1 et V_2 ?

On peut voir tout d'abord que les vecteurs X_2^1, X_2^2 et X_2^3 sont peu liés entre eux puisque des liaisons étroites se traduiraient par une grande proximité de ces points. On constate également que les points représentant $P_1(X_1^1)$, $P_1(X_1^2)$ et $P_1(X_1^3)$ restent assez éloignés des points représentant X_2^1, X_2^2, X_2^3: c'est dire que ces derniers expliquent peu les précédents. On perçoit toutefois l'influence prépondérante de X_2^3 dans $P_1(X_1^3)$ et celle de X_2^1 dans $P_1(X_1^1)$ par leur position par rapport à l'axe V_1.



IV - CONCLUSION

A une époque où le nombre des données disponibles augmente sans cesse, la méthode proposée dans la troisième partie de ce travail peut permettre des traitements qui ne supposent ni choix effectués sans critères objectifs parmi les variables, ni réduction de l'information disponible par pas-passage à des moyennes ou autres caractéristiques.

De plus, il paraît certain qu'un raisonnement analogue permettra une généralisation de toutes les méthodes d'analyse multidimensionnelle à des familles de vecteurs. C'est dans ce sens que nous travaillons.

Je tiens à remercier M. Le Professeur B. CHARLES de l'aide morale et technique qu'il m'a sans cesse apporté ; sans lui ce travail n'aurait pas été possible.

BIBLIOGRAPHIE

[1] T.W. ANDERSON - An Introduction to multivariate statistical analysis . John Xiley and Sons, 1958.

[2] J.P. BENZECRI - Distance distributionnelle et métrique du χ^2 en Analyse factorielle des correspondances. Polycopié, 1970.

[3] B. CORDIER - L'analyse factorielle des correspondances - Thèse de Doctorat de 3ème cycle - Université de Rennes, 1965.

[4] B. ESCOFFIER (CORDIER) - Sur l'Analyse des correspondances. Note polycopiée, 1970.

[5] Y. ESCOUFIER - Echantillonnage dans une population de variables aléatoires réelles - Thèse de Doctorat d'Etat - Université de Montpellier. 1970.