Statistique et analyse des données

Jacques Pontier
Pierre Jolicoeur
Marie-Odile Pernin

Analyse canonique complète

Statistique et analyse des données, tome 12, n° 1-2 (1987), p. 124-148
<http://www.numdam.org/item?id=SAD_1987__12_1-2_124_0>

Statistique et Analyse des Données
1987 - Vol.12 n° 1 et 2 - pp. 124-148

ANALYSE CANONIQUE COMPLETE

Jacques PONTIER *, Pierre JOLICOEUR **, Marie-Odile PERNIN *

* Laboratoire d'Analyse de Données et Biométrie,
Université Claude Bernard, 43 boulevard du 11 novembre 1918
69622 VILLEURBANNE CEDEX
FRANCE

** Département de Sciences Biologiques, Université de Montréal
C.P. 6128 MONTREAL Qué H3C 3J7
CANADA

Résumé : *L'analyse canonique, opération classique en analyse des données, possède des ressources incomplètement exploitées. Cet article explicite certaines propriétés liées aux sous-espaces propres associés aux valeurs propres nulles des matrices diagonalisées de l'analyse canonique, et indique des possibilités d'utilisation de ces propriétés.*

Abstract : *Though canonical correlation analysis is a classical method in data analysis, its properties are generally underexploited. In this paper we show several properties of subspaces associated with null eigenvalues. We suggest several possible uses of these properties.*

Mots clés : *Analyse canonique, valeurs propres nulles, projections orthogonales, sous-espaces orthogonaux.*

Indices de classification STMA : *06-040, 07-040*

1 - INTRODUCTION

L'analyse canonique est une opération bien connue en analyse des données, et qui s'interprète de diverses manières selon le contexte choisi (voir par exemple la formulation par la recherche de fonctions linéaires à corrélation maximum, ou par la recherche de sous-espaces unidimensionnels invariants par double projection ; voir aussi les formulations liées à des applications particulières : analyse discriminante, analyse des correspondances).

Manuscrit reçu le 18.6.87, révisé le 4.12.87

Cette opération n'est cependant pas utilisée pleinement, en ce sens qu'une part importante de ses propriétés est habituellement négligée, bien que cette part soit prometteuse de retombées intéressantes dans de nombreux domaines de l'analyse des données, tant sur le plan théorique que sur le plan des applications. Il s'agit des propriétés des sous-espaces propres associés aux valeurs propres nulles des matrices diagonalisées. L'objet de cet article est d'expliciter un certain nombre de ces propriétés, et de donner des exemples de possibilités d'utilisation.

Nous consacrons le §2 à un rappel des propriétés de l'analyse canonique ordinaire, c'est-à-dire de l'analyse canonique telle qu'on la pratique habituellement entre deux "blocs de variables". Ces propriétés étant très classiques, nous nous excusons par avance auprès du lecteur qui trouverait ce rappel superflu : la justification de sa présence est qu'il nous permet de mettre en place notre système de notations, de présenter le contexte dans lequel nous interprétons l'analyse canonique, et d'introduire comme un prolongement naturel les propriétés constituant l'objet principal de notre propos.

Le §3 montre comment, moyennant quelques calculs supplémentaires, les résultats habituellement obtenus par l'analyse canonique ordinaire se trouvent complétés par d'autres résultats dont il est permis d'attendre des applications théoriques et concrètes intéressantes. L'objet du §4 est d'esquisser l'étude théorique de quelques-unes de ces applications.

Ce texte développe les idées exposées par l'un des auteurs (J. P.) lors des XIXes Journées de Statistique à Lausanne (mai 1987).

2 - RAPPELS SUR L'ANALYSE CANONIQUE ORDINAIRE

Nous nous placerons ici dans le cas de l'analyse canonique "simple", c'est-à-dire celle qui concerne deux "blocs de variables", observées sur un nombre fini d'"individus". Nous préciserons au préalable les concepts et notations que nous allons utiliser.

La situation à analyser se présente concrètement sous la forme d'un tableau de données du type suivant, dans lequel les ω_j (j = 1 à n) désignent les unités statistiques observées (les "individus") ; les x_r (r = 1 à p) et les y_s (s = 1 à q) symbolisent des "variables" observées sur les ω_j ; x_{rj} et y_{sj} notent les valeurs numériques respectives de la variable x_r et de la variable y_s, telles qu'observées sur l'unité ω_j. Enfin, les π_j sont des coefficients de pondération affectés aux unités observées, coefficients strictement positifs

dont la somme est égale à 1 par convention ; souvent ces coefficients sont égaux (chacun étant donc égal à 1/n), mais dans ce qui suit nous ne supposerons pas être dans ce cas particulier.

obs.	poids	x_1	x_2 ...	x_p	y_1	y_2 ...	y_q
ω_1	π_1	x_{11}	x_{21} ...	x_{p1}	y_{11}	y_{21} ...	y_{q1}
ω_2	π_2	x_{12}	x_{22} ...	x_{p2}	y_{12}	y_{22} ...	y_{q2}
:	:	:	:	:	:	:	:
ω_n	π_n	x_{1n}	x_{2n} ...	x_{pn}	y_{1n}	y_{2n} ...	y_{qn}

Nous interprèterons cette situation selon le modèle euclidien suivant. Soit Ω l'ensemble $\{\omega_1, \omega_2, \ldots, \omega_n\}$ des n unités statistiques, dont chacune est affectée du "poids" positif π_j. L'ensemble, noté $[\Omega]$, des fonctions $\Omega \rightarrow \mathbb{R}$, est un espace vectoriel de dimension n. Par exemple, les x_r et les y_s sont des vecteurs de $[\Omega]$. Pour chaque j = 1 à n, soit Ω_j la fonction indicatrice de ω_j : $\Omega_j(\omega) = 1$ si $\omega = \omega_j$, $= 0$ sinon. L'ensemble des n fonctions indicatrices constitue une base de l'espace vectoriel $[\Omega]$, base que nous appelons pour cette raison la base indicatrice de $[\Omega]$ (appellation préférée à "base canonique" pour éviter toute ambiguité, dans le contexte de l'analyse canonique).

La pondération des unités statistiques permet de définir sur $[\Omega]$ un produit scalaire noté φ, soit :

$$x, y \in [\Omega] \quad \rightarrow \varphi(x,y) = \sum_j \pi_j \, x(\omega_j) \, y(\omega_j)$$

La métrique induite dans $[\Omega]$ par ce produit scalaire est la métrique diagonale des poids.

Dans $[\Omega]$ considérons les deux systèmes de vecteurs, $X = \{x_1, x_2, ..., x_p\}$ et $Y = \{y_1, y_2, ..., y_q\}$. Ces deux systèmes engendrent respectivement les sous-espaces [X] et [Y] de $[\Omega]$. Quitte à substituer à chacun des systèmes X, Y, un système libre linéairement équivalent (engendrant le même sous-espace), nous pouvons toujours supposer que chacun de ces systèmes est libre, et de ce fait constitue une base du sous-espace qu'il engendre.

Si u est un vecteur de [X] (donc aussi vecteur de $[\Omega]$), U_X note sa matrice par rapport à la base X, alors que U_Ω note sa matrice par rapport à la base indicatrice de $[\Omega]$.

Nous noterons Φ_{XX} le tableau des produits scalaires $\varphi(x_r,x_s)$. Ce tableau est interprétable comme la matrice (par rapport à la base X) de la restriction du produit scalaire φ au sous-espace [X]. De façon analogue, nous noterons Φ_{YY} le tableau des produits scalaires $\varphi(y_t,y_u)$, Φ_{XY} le tableau des produits scalaires $\varphi(x_r,y_t)$, enfin Φ_{YX} le tableau transposé de Φ_{XY}. Avec ce système de notation, le tableau de données X, de dimensions n÷p, est interprétable comme la matrice X_Ω obtenue par concaténation des matrices (par rapport à la base indicatrice de [Ω]) des vecteurs $x_1, x_2, \ldots, x_p$.

La réflexion φ-orthogonale de [X] sur [Y] est l'application linéaire de [X] dans [X] définie comme la composée des projections φ-orthogonales de [X] dans [Y] puis de [Y] dans [X]. Cette opération est schématisée à la figure 1.

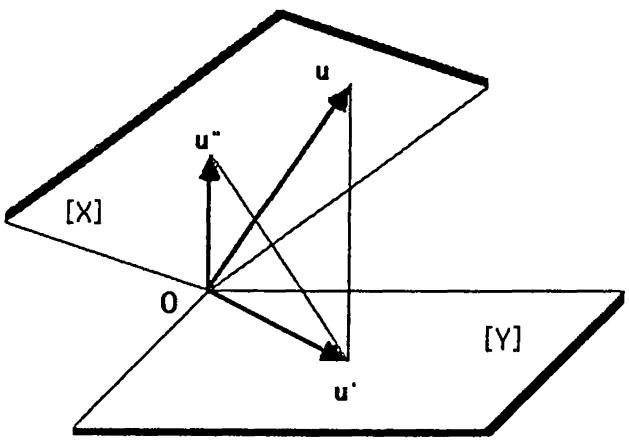
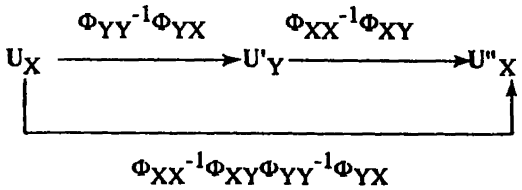


Fig 1. Représentation schématique de l'opération de "réflexion" ou "double projection" de [X] sur [Y] : un vecteur **u** de [X] est projeté en **u'** sur [Y] ; cet **u'** est à son tour projeté en **u"** sur [X] (il s'agit dans les deux cas de projections φ-orthogonales).

L'expression matricielle (par rapport à X) de cette opération est :

$$M_{XX} = \Phi_{XX}^{-1}\Phi_{XY}\Phi_{YY}^{-1}\Phi_{YX}$$



A partir de là nous pouvons obtenir $U''_\Omega = X_\Omega U''_X = X_\Omega M_{XX} U_X$, c'est-à-dire le tableau des valeurs numériques prises par les variables u sur les individus.

On définira de façon analogue la réflexion φ-orthogonale de [Y] sur [X], application linéaire de [Y] dans lui-même, composée des deux projections φ-orthogonales de [Y] dans [X] puis de [X] dans [Y]. La matrice (par rapport à Y) de cette application est donc : $N_{YY} = \Phi_{YY}^{-1}\Phi_{YX}\Phi_{XX}^{-1}\Phi_{XY}$.

L'analyse canonique ordinaire entre [X] et [Y] est la recherche :
- d'une part, dans [X], des directions (sous-espaces unidimensionnels) invariantes par réflexion φ-orthogonale sur [Y] : ce sont les directions propres de la matrice M_{XX} ;
- d'autre part, dans [Y], des directions invariantes par réflexion φ-orthogonale sur [X] : ce sont les directions propres de la matrice N_{YY}.

Les propriétés couramment utilisées sont les suivantes :

a) M_{XX} et N_{YY} ont le même rang, noté ici μ ($\mu \leq \min(p,q)$).

b) Elles ont les mêmes μ valeurs propres non nulles, notées $\lambda_1, \lambda_2, \ldots, \lambda_\mu$, et l'on a $1 \geq \lambda_i > 0$ quel que soit $i = 1,\ldots, \mu$.

c) Soit λ_i l'une de ces valeurs propres non nulles. u_i et v_i sont des vecteurs propres associés à λ_i, respectivement pour les matrices M_{XX} et N_{YY} ; U_{iX} et V_{iY} sont les matrices (respectivement par rapport à X et par rapport à Y) de ces vecteurs propres. Alors :

$\lambda_i = \cos^2(u_i, v_i)$ [il s'agit du cosinus du φ-angle, défini par le produit scalaire φ] ; si u_i et v_i sont chacun φ-normés, alors :

$$U_{iX} = (1/\sqrt{\lambda_i})\,\Phi_{XX}^{-1}\Phi_{XY}\,V_{iY} \quad \text{et} \quad V_{iY} = (1/\sqrt{\lambda_i})\,\Phi_{YY}^{-1}\Phi_{YX}\,U_{iX}$$

d) Les éventuelles valeurs propres autres que les λ_i, $i = 1$ à μ, sont toutes nulles.

Ce sont ces valeurs propres nulles, ou plus exactement les sous-espaces propres associés, qui vont faire l'objet du "complément" à apporter à l'analyse canonique. Les vecteurs propres associés aux valeurs propres nulles issues de la diagonalisation des matrices M_{XX} et N_{YY}, ne semblent pas couramment utilisés. ANDERSON [1] n'évoque même pas leur existence éventuelle. CAILLIEZ et PAGES ([2], chap XI) sont suffisamment explicites quant à l'existence des sous-espaces propres associés aux valeurs propres nulles ; ils n'en exploitent cependant aucune propriété. Plus près de nous, TAKEUCHI, YANAI & MUKHERJEE [6], bien que développant l'interprétation "projectionniste" de l'analyse des données, n'évoquent que très sommairement les sous-espaces en question.

3 - ANALYSE CANONIQUE COMPLETE

La considération de tous les sous-espaces propres associés à la matrice M_{XX} va nous permettre de définir une décomposition de [X] en somme directe de trois sous-espaces φ-orthogonaux deux à deux, chacun d'eux possédant des propriétés particulières relativement au sous-espace [Y]. Ceci nous conduira à appeler cette décomposition la décomposition complète de [X] canonique par rapport à [Y].

De la même façon, nous serons amenés à considérer la décomposition complète de [Y] canonique par rapport à [X]. Ensuite, à partir des éléments de ces deux décompositions, celle de [X] et celle de [Y], nous verrons comment obtenir une "décomposition canonique complète" de la somme [X]+[Y], c'est-à-dire du sous-espace de [Ω] engendré par XUY.

3.1- Décomposition complète de [X] canonique par rapport à [Y]

En analyse canonique, chacune des valeurs propres des deux matrices à diagonaliser (M_{XX} et N_{YY}) s'interprète comme le carré du cosinus d'un angle. Cette propriété, que nous avons rappelée au §2 pour ce qui concerne les valeurs propres non nulles, est valable également pour les valeurs propres nulles : une valeur propre nulle caractérise alors une relation de φ-orthogonalité que nous allons expliciter. Nous examinons ici le cas de la diagonalisation de la matrice M_{XX}.

Cette matrice, d'ordre p, de rang $\mu \leq p$, possède μ valeurs propres strictement positives, parmi lesquelles ν sont égales à 1 (éventuellement $\nu = 0$). Les $p-\mu$ autres valeurs propres sont nulles.

A l'ensemble des ν valeurs propres égales à 1, correspond un sous-espace de [X], noté $[X]_1$ engendré par ν vecteurs propres $u_1, \ldots, u_\nu$ associés respectivement à ces ν valeurs propres. Ces vecteurs propres sont deux à deux φ-orthogonaux, et constituent donc une base φ-orthogonale de $[X]_1$, lequel est donc de dimension ν. Ce sous-espace $[X]_1$ est l'ensemble des vecteurs de [X] qui coïncident avec leur propre réflexion φ-orthogonale sur [Y] : c'est donc l'intersection $[X] \cap [Y]$.

Aux $\mu - \nu$ valeurs propres strictement comprises entre 0 et 1, sont associés des vecteurs propres $u_{\nu+1}, \ldots, u_\mu$, engendrant un sous-espace $[X]_2$ dans [X]. Ces vecteurs propres étant φ-orthogonaux deux à deux, ils constituent une base φ-orthogonale de $[X]_2$, qui de ce fait est de dimension $\mu-\nu$. $[X]_2$ est l'ensemble de tous les vecteurs de [X] qui ne sont ni dans [Y], ni φ-orthogonaux à [Y] ; nous dirons pour cette raison que ce sont les vecteurs de [X] "obliques" par rapport à [Y].

Enfin, aux $p-\mu$ valeurs propres nulles, correspond un sous-espace propre $[X]_3$ de [X], de dimension $p-\mu$, dont tout élément est φ-orthogonal non seulement à tout vecteur de $[X]_1$ ou de $[X]_2$, mais encore à tout vecteur de [Y]. $[X]_3$ est l'ensemble des vecteurs de [X] φ-orthogonaux à [Y], c'est-à-dire $[X]_3 = [X] \cap [Y]^{\perp}$, où $[Y]^{\perp}$ note le supplément φ-orthogonal de [Y] dans $[\Omega]$.

L'espace [X] se trouve donc décomposé en la somme directe des trois sous-espaces $[X]_1$, $[X]_2$ et $[X]_3$; c'est cette décomposition particulière de [X], issue de l'analyse canonique entre [X] et [Y], que nous appelons la décomposition complète de [X] canonique par rapport à [Y].

Notons $u_{\mu+1}, \ldots, u_p$ une base φ-orthogonale de $[X]_3$. Par convention, pour $i = \mu+1$ à p, le vecteur u_i sera considéré comme un vecteur propre associé à la valeur propre nulle λ_i. Cette convention, non contradictoire avec les propriétés du sous-espace propre associé à la valeur propre 0 d'ordre $p-\mu$, nous permet d'homogénéiser la présentation des propriétés liant les p valeurs propres de M_{XX}, les p vecteurs propres qui leurs sont associés, et les trois sous-espaces $[X]_1$, $[X]_2$, $[X]_3$ définis ci-dessus. Ces propriétés sont récapitulées dans le tableau suivant (le symbole $\oplus^{\perp}$ entre deux sous-espaces désigne la "somme directe orthogonale" de ces deux sous-espaces, signifiant par là la somme de deux sous-espaces dont l'intersection est réduite au vecteur nul, et qui sont φ-orthogonaux entre eux) :

$$\left.\begin{array}{lcl} \lambda_1 = 1 & \Leftrightarrow & u_1 \\ \vdots & & \vdots \\ \lambda_\nu = 1 & \Leftrightarrow & u_\nu \end{array}\right\} \Rightarrow [X]_1 = [u_1] \oplus^\perp \ldots \oplus^\perp [u_\nu] \quad \dim [X]_1 = \nu$$

$$\left.\begin{array}{lcl} 1 > \lambda_{\nu+1} > 0 & \Leftrightarrow & u_{\nu+1} \\ \vdots & & \vdots \\ 1 > \lambda_\mu > 0 & \Leftrightarrow & u_\mu \end{array}\right\} \Rightarrow [X]_2 = [u_{\nu+1}] \oplus^\perp \ldots \oplus^\perp [u_\mu] \quad \dim [X]_2 = \mu - \nu$$

$$\left.\begin{array}{lcl} \lambda_{\mu+1} = 0 & \Leftrightarrow & u_{\mu+1} \\ \vdots & & \vdots \\ \lambda_p = 0 & \Leftrightarrow & u_p \end{array}\right\} \Rightarrow [X]_3 = [u_{\mu+1}] \oplus^\perp \ldots \oplus^\perp [u_p] \quad \dim [X]_3 = p - \mu$$

$[X]_1$ = $[X] \cap [Y]$ = ensemble des $u \in [X]$ appartenant aussi à $[Y]$
$[X]_2$ = ensemble des $u \in [X]$ "obliques" par rapport à $[Y]$
$[X]_3$ = $[X] \cap [Y]^\perp$ = ensemble des $u \in [X]$ φ-orthogonaux à $[Y]$
$[X]$ = $[X]_1 \oplus^\perp [X]_2 \oplus^\perp [X]_3$

3.2 - Décomposition complète de [Y] canonique par rapport à [X]

Nous définissons d'une manière analogue la décomposition complète de [Y] canonique par rapport à [X]. Cette décomposition résulte de la diagonalisation "complète" de la matrice N_{YY}. Cette matrice, d'ordre q, de rang $\mu \leq q$, possède μ valeurs propres strictement positives, identiques aux valeurs propres strictement positives de la matrice M_{XX}. Les $q-\mu$ autres valeurs propres sont nulles. L'espace [Y] se trouve ainsi décomposé en somme directe de trois sous-espaces, deux à deux φ-orthogonaux, notés $[Y]_1$, $[Y]_2$, $[Y]_3$, sous-espaces propres de N_{YY} associés respectivement à l'ensemble des valeurs propres égales à 1, à l'ensemble des valeurs propres strictement comprises entre 0 et 1, à l'ensemble des valeurs propres égales à 0 :

$[Y]_1$ = $[Y] \cap [X]$ = ensemble des $v \in [Y]$ appartenant aussi à $[X]$
$[Y]_2$ = ensemble des $v \in [Y]$ "obliques" par rapport à $[X]$
$[Y]_3$ = $[Y] \cap [X]^\perp$ = ensemble des $v \in [Y]$ φ-orthogonaux à $[X]$
$[Y]$ = $[Y]_1 \oplus^\perp [Y]_2 \oplus^\perp [Y]_3$
$\dim [Y]_1 = \nu$; $\dim [Y]_2 = \mu - \nu$; $\dim [Y]_3 = q - \mu$; $\dim [Y] = q$

3.3 - Décomposition canonique complète de [X] + [Y]

[X]+[Y] note le sous-espace de [Ω] engendré par $\{x_1, \ldots, x_p, y_1, \ldots, y_q\}$. Il est de dimension au plus égale à p+q. La décomposition complète de [X] canonique par rapport à [Y], et la décomposition complète de [Y] canonique par rapport à [X], que nous venons de définir ci-dessus, permettent d'obtenir une décomposition de [X]+[Y] sous la forme d'une somme directe de quatre sous-espaces deux à deux φ-orthogonaux, et dont chacun a un lien particulier avec [X] et/ou avec [Y] :

$$[X] + [Y] = ([X] \cap [Y]) \oplus^{\perp} ([X]_2 \oplus [Y]_2) \oplus^{\perp} [X]_3 \oplus^{\perp} [Y]_3$$

* $[X] \cap [Y]$ est de dimension ν. Il est identique à $[X]_1$, identique aussi à $[Y]_1$: c'est le sous-espace de [X]+[Y] commun à [X] et à [Y]. Il est engendré aussi bien par des vecteurs de [X] (par exemple : $u_1, \ldots, u_\nu$) que par des vecteurs de [Y] (par exemple : $v_1, \ldots, v_\nu$). Nous pourrons prendre comme base φ-orthonormée les $w_i = (u_i + v_i)/2$ (i = 1 à ν).

* $[X]_2 \oplus [Y]_2$ est de dimension $2(\mu-\nu)$. C'est le sous-espace de [X] + [Y] qui dépend à la fois de [X] et de [Y], étant engendré par la base $\{u_{\nu+1}, \ldots, u_\mu, v_{\nu+1}, \ldots, v_\mu\}$. Pour cette raison nous l'appelons le sous-espace de [X] + [Y] interdépendant de [X] et de [Y]. Notons que la base indiquée ci-dessus n'est pas φ-orthogonale ; rappelons en effet que $\varphi(u_i,v_i) = \lambda_i > 0$ dans ce cas. On obtiendra à partir de là une base φ-orthonormée constituée par les $w_{2i-\nu-1} = (u_i+v_i)/\sqrt{2+2\sqrt{\lambda_i}}$ et les $w_{2i-\nu} = (u_i-v_i)/\sqrt{2-2\sqrt{\lambda_i}}$, pour i = ν+1 à μ.

* $[X]_3$ est de dimension p-μ. C'est le sous-espace de [X] + [Y] ne dépendant que de [X] ; il est φ-orthogonal à [Y], et engendré par $\{u_{\mu+1}, \ldots, u_p\}$. Nous l'appelons le sous-espace de [X] + [Y] spécifique de [X]. Nous prendrons $w_{i+\mu-\nu} = u_i$, pour i = μ+1 à p.

* $[Y]_3$ est de dimension q-μ. C'est le sous-espace de [X] + [Y] ne dépendant que de [Y] ; il est φ-orthogonal à [X], et engendré par $\{v_{\mu+1}, \ldots, v_q\}$. Nous l'appelons le sous-espace de [X] + [Y] spécifique de [Y]. Nous prendrons $w_{i+p-\nu} = v_i$, pour i = μ+1 à q.

3.4 - Problèmes pratiques posés par l'analyse canonique complète

La réalisation des calculs nécessités par l'analyse canonique complète ne pose pas de problème grave. Les vecteurs propres associés aux valeurs propres strictement positives sont obtenus selon les méthodes habituelles aux diverses versions de l'analyse canonique,

puisqu'il s'agit là de la pratique ordinaire.

L'obtention d'une base φ-orthonormée de $[X]_3$, et d'une base φ-orthonormée de $[Y]_3$ peut poser un problème numérique, les vecteurs propres associés aux valeurs propres nulles ayant la réputation de ne pas être fiables. Aussi, nous dissocions ce problème de celui de la détermination de vecteurs propres associés aux valeurs propres nulles de M_{XX} (resp. N_{YY}). Nous procédons de la manière suivante, par exemple pour obtenir une base φ-orthogonale de $[X]_3$.

a) Dans [X], nous disposons déjà d'une base du sous-espace $[X]_1 \oplus^{\perp} [X]_2$, base constituée par les vecteurs $u_1, \ldots, u_\nu, \ldots, u_\mu$, vecteurs propres associés aux valeurs propres strictement positives de la matrice M_{XX} (ces vecteurs propres, obtenus par les méthodes habituelles, définissent les composantes canoniques de l'analyse canonique ordinaire). Nous construisons alors les x_i', projections φ-orthogonales sur $[X]_1 \oplus^{\perp} [X]_2$ des vecteurs x_i (i = 1 à p), base de [X]. Les vecteurs $z_i = x_i - x_i'$ sont tous dans $[X]_3$, dont ils constituent un système de générateurs.

b) Nous calculons la matrice Φ_{ZZ} des produits scalaires de ces p vecteurs z_i. Cette matrice, d'ordre p, est de rang p - μ. A partir de Φ_{ZZ}, nous disposons d'au moins deux méthodes permettant d'obtenir une base φ-orthonormée de $[X]_3$:
- les vecteurs propres associés aux valeurs propres strictement positives de Φ_{ZZ} constituent une base φ-orthogonale de $[X]_3$: c'est la "base principale", constituée par les composantes principales de la matrice Φ_{ZZ}, à partir de laquelle nous obtenons facilement une base φ-orthonormée ; cette façon de procéder est donc en définitive l'ACP de l'ensemble des résidus des projections φ-orthogonales des x_i sur $[X]_1 \oplus^{\perp} [X]_2$.
- une procédure du type Gram-Schmidt améliorée, nous permet de construire une base φ-orthonormée.

Nous avons essayé ces deux méthodes ; la seconde nous paraît plus rapide que la première.

Nous obtiendrons de la même manière une base φ-orthogonale de $[Y]_3$. Ayant ainsi obtenu tous les vecteurs de base nécessaires, tant dans [X] que dans [Y], il sera intéressant de les φ-normer, en vue de leur utilisation. La base φ-orthonormée de l'espace inter-dépendant $[X]_2+[Y]_2$ sera alors constituée par l'ensemble des $(u_i+v_i)/\sqrt{2+2\sqrt{\lambda_i}}$ et des $(u_i-v_i)/\sqrt{2-2\sqrt{\lambda_i}}$, pour i = ν+1 à μ.

3.5 - Exemple

L'étude qui suit se place dans un contexte de recherche de moyens d'améliorer la technique de sportifs de haut niveau. L'analyse des relations existant entre la performance du sportif, et la manière dont il effectue certains mouvements, est un outil de cette recherche. L'exemple présenté concerne la natation (crawl). La performance du nageur est évidemment liée, entre autres, à son mouvement de bras. Ce mouvement, cyclique, est étudié en détail par l'examen image par image d'un enregistrement vidéo de la partie immergée du nageur en action.

Les éléments d'appréciation de la performance réalisée par le nageur pendant un cycle sont :

x_1 = l'*avancée totale du cycle* = la distance, en mètres, parcourue pendant le cycle étudié (distance mesurée par le déplacement du bassin) ;

x_2 = la *vitesse du cycle* = la vitesse moyenne, en mètres par seconde, réalisée pendant le cycle étudié.

Le cycle étudié est sélectionné parmi la dizaine de cycles nécessités par la traversée de la piscine. L'étude du mouvement pendant ce cycle est basée sur l'observation de la variation de l'angle du bras par rapport à l'horizontale ; le mouvement de rotation complète du bras est subdivisé en cinq phases successives : de 0° à 45°, de 45° à 90°, de 90° à 135°, de 135° à 180°, de 180° à 360° (au cours de cette dernière phase le bras, constamment hors de l'eau, n'est pas filmé par la caméra) (voir figure 2). A chacune des cinq phases, on associe sa durée relative (en %) par rapport à la durée totale du cycle : ces cinq durées relatives constituent les cinq variables $y_1, ..., y_5$ dont nous disposons pour caractériser le mouvement du bras. Remarquons dès à présent que ces cinq dernières variables ne sont pas linéairement indépendantes, leur somme étant contrainte à être égale à 100. On connaît d'autre part l'indice de performance (i.p.) de chaque nageur, c'est-à-dire le rapport (en %) de la vitesse record atteinte par ce nageur au 100 m crawl, sur la vitesse du record mondial (1.86 m/s pour les filles, 2.04 m/s pour les garçons). Cet indice de performance n'est directement fonction ni de la performance du nageur pendant le cycle observé, ni des mouvements du nageur pendant ce cycle ; nous considèrerons ici qu'il s'agit d'une variable "supplémentaire".

En vue de l'analyse des relations entre performance et mouvement, les 8 variables décrites ci-dessus ont été notées sur 25 nageurs (hommes) de haut niveau. On a choisi une pondération uniforme de ces 25 nageurs.

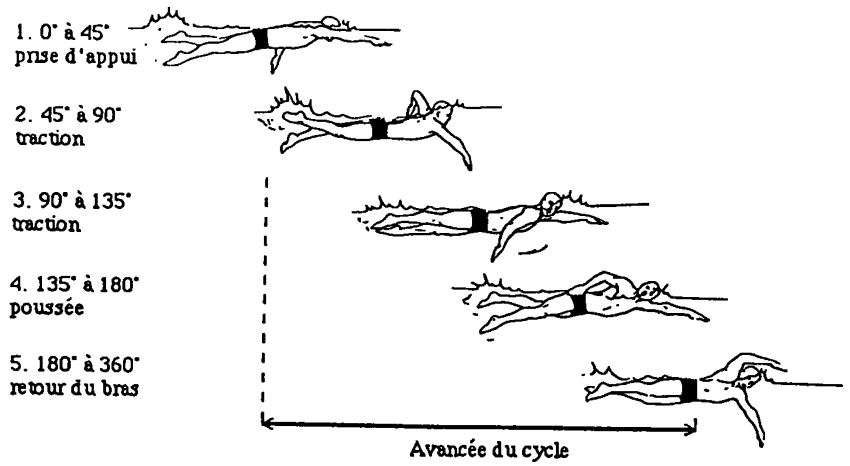


Fig 2. Représentation schématique des cinq phases du mouvement d'un bras du nageur, d'après ROUARD [5].

La partie "ordinaire" de l'analyse canonique entre les deux ensembles de variables, doit nous permettre de lier la performance au mouvement du bras. La partie "complémentaire" de cette même analyse, est destinée à mettre en évidence, le cas échéant, ce qui dans la performance est sans corrélation avec le mouvement, et ce qui dans le mouvement est sans corrélation avec la performance. Dans l'espace [Ω], de dimension 25, nous réalisons l'analyse canonique complète entre les deux sous-espaces [X], engendré par {1, x_1, x_2,}, et [Y], engendré par {1, y_1, y_2, y_3, y_4} ; nous avons écarté la variable y_5, combinaison linéaire des 4 autres variables y_s. Le vecteur noté 1 est la "variable" constamment égale à 1 ; sa présence parmi les générateurs du sous-espace [X] et parmi ceux du sous-espace [Y], signifie que l'on recherchera des fonctions linéaires des variables, avec terme constant, et que les matrices de produits scalaires intervenant dans la diagonalisation seront les matrices de covariances ou de corrélations (matrices notées S), des variables x_r (r=1, 2) et y_s (s=1 à 4). Les valeurs numériques de ces variables sont contenues dans le tableau 1 ; elles nous ont été aimablement communiquées par Mme A. ROUARD, professeur d'Education Physique et Sportive à l'Université Claude Bernard Lyon I, que nous remercions ici. Dans le présent article, ces données ne sont utilisées qu'à titre d'illustration de la méthode d'analyse de données que nous présentons ; le lecteur désireux d'en savoir plus sur l'étude de natation qui en est à l'origine, pourra se reporter à ROUARD [5]

i.p.	x_1	x_2	y_1	y_2	y_3	y_4	y_5
94.35	2.50	1.79	31.43	5.71	11.43	14.29	37.14
98.15	2.83	1.91	35.14	10.81	10.81	10.81	32.43
85.43	1.98	1.60	32.26	9.68	9.68	19.35	29.03
80.90	2.21	1.53	30.56	8.33	11.11	16.67	33.33
82.82	2.42	1.59	36.84	7.89	5.26	13.16	36.84
77.55	2.26	1.45	33.33	10.26	10.26	12.82	33.33
79.97	2.35	1.55	34.21	7.89	10.53	13.16	34.21
87.89	2.21	1.67	36.36	6.06	12.12	12.12	33.33
90.33	2.53	1.76	30.56	13.89	11.11	13.89	30.56
89.34	2.40	1.71	31.43	8.57	11.43	17.14	31.43
90.33	2.40	2.00	33.33	6.67	10.00	13.33	36.67
92.56	2.50	1.79	31.43	8.57	11.43	14.29	34.29
91.35	2.41	1.77	35.29	5.88	5.88	14.71	38.24
92.56	2.50	1.74	33.33	11.11	11.11	13.89	30.56
92.04	2.33	1.71	26.47	8.82	8.82	20.59	35.29
89.84	2.62	1.72	39.47	7.89	10.53	10.53	31.58
89.02	2.31	1.80	34.38	9.38	9.38	12.50	34.38
90.84	2.43	1.69	36.11	11.11	8.33	13.89	30.56
95.09	2.42	1.78	32.35	8.82	11.76	17.65	29.41
90.00	2.50	1.84	32.35	8.82	8.82	14.71	35.29
87.11	2.28	1.73	36.36	9.09	9.09	15.15	30.30
83.53	2.45	1.61	31.58	7.89	10.53	15.79	34.21
83.53	2.10	1.59	30.30	9.09	9.09	15.15	36.36
89.67	2.34	1.72	32.35	8.82	11.76	11.76	35.29
90.84	2.35	1.78	36.36	9.09	9.09	15.15	30.30

Tab 1. Valeurs des variables x_1, x_2 (caractérisant la performance) et des variables y_1, y_2, y_3, y_4, y_5 (caractérisant le mouvement du bras) mesurées sur 25 nageurs (hommes) de haut niveau. La colonne i.p. contient l'indice de performance, variable supplémentaire sans lien calculatoire avec les variables précédentes.

Le sous-espace [X] est de dimension 3, le sous-espace [Y] est de dimension 5. Ces deux sous-espaces n'ont en commun que le sous-espace trivial unidimensionnel engendré par le vecteur $u_1 = v_1 = 1$ (sous-espace des constantes) : $[X]_1 = [Y]_1 = [1]$ (correspondant à la valeur propre triviale $\lambda_1 = 1$). La diagonalisation de la matrice $S_{XX}^{-1}S_{XY}S_{YY}^{-1}S_{YX}$, d'ordre 2, nous donne les deux autres valeurs propres strictement positives λ_2 et λ_3, auxquelles correspondent dans [X] (resp. [Y]) les deux composantes canoniques obliques u_2, u_3 (resp. v_2, v_3), base φ-orthogonale de $[X]_2$ (resp. $[Y]_2$). Le sous-espace [X] ne possède pas de composante φ-orthogonale à [Y] ($[X]_3$ est réduit au vecteur 0) ; par contre le sous-espace [Y] possède une composante φ-orthogonale à [X], soit $[Y]_3$ engendré par les vecteurs v_4 et v_5. Les valeurs numériques des valeurs propres et des composantes des vecteurs propres sont récapitulées ci-dessous :

$$\lambda_1 = 1 \quad \begin{cases} u_1 = \mathbf{1} \\ v_1 = \mathbf{1} \end{cases}$$

$$\lambda_2 = 0.27181 \quad \begin{cases} u_2 = -11.6354\,\mathbf{1} + 6.9608\,x_1 - 2.8994\,x_2 \\ v_2 = 6.5816\,\mathbf{1} - 0.0613\,y_1 + 0.1998\,y_2 - 0.0020\,y_3 - 0.4329\,y_4 \end{cases}$$

$$\lambda_3 = 0.02028 \quad \begin{cases} u_3 = -11.4182\,\mathbf{1} - 2.1466\,x_1 + 9.6527\,x_2 \\ v_3 = 7.5985\,\mathbf{1} - 0.0709\,y_1 - 0.5250\,y_2 + 0.1900\,y_3 - 0.1729\,y_4 \end{cases}$$

$$\lambda_4 = 0 \qquad v_4 = -9.8716\,\mathbf{1} + 0.03182\,y_1 + 0.1307\,y_2 + 0.5754\,y_3 + 0.1325\,y_4$$

$$\lambda_5 = 0 \qquad v_5 = -25.6804\,\mathbf{1} + 0.5311\,y_1 + 0.0378\,y_2 + 0.2692\,y_3 + 0.3416\,y_4$$

Les expressions des x_r et de y_s en fonction respectivement des u_k et des v_t sont les suivantes

$$\text{dans [X] :} \quad \begin{cases} x_1 = 2.3852\,\mathbf{1} + 0.1583\,u_2 + 0.0476\,u_3 \\ x_2 = 1.7133\,\mathbf{1} + 0.0352\,u_2 + 0.1142\,u_3 \end{cases}$$

$$\text{dans [Y] :} \quad \begin{cases} y_1 = 33.3440\,\mathbf{1} + 1.2101\,v_2 + 0.3391\,v_3 - 1.1063\,v_4 + 2.1341\,v_5 \\ y_2 = 8.8071\,\mathbf{1} + 0.6576\,v_2 - 1.5217\,v_3 + 0.5807\,v_4 - 0.1620\,v_5 \\ y_3 = 9.9747\,\mathbf{1} + 0.2854\,v_2 + 0.5001\,v_3 + 1.5712\,v_4 + 0.0055\,v_5 \\ y_4 = 14.4994\,\mathbf{1} - 2.1790\,v_2 - 0.7528\,v_3 + 0.4173\,v_4 - 0.3770\,v_5 \\ y_5 = 33.3749\,\mathbf{1} - 0.0260\,v_2 + 1.4353\,v_3 - 1.4629\,v_4 - 1.6007\,v_5 \end{cases}$$

La décomposition complète de chaque variable x_r, de chaque variable y_s, comme fonction linéaire d'un certain nombre de variables artificielles (ici les composantes canoniques) φ-orthonormées, autorise l'utilisation des règles habituelles d'interprétation de ces variables artificielles, reposant sur les corrélations entre variables et "facteurs" (cf les cercles de corrélations). Nous ne procèderons pas ici à une interprétation détaillée de tous les résultats de l'analyse ci-dessus ; nous donnerons seulement, sous forme de graphiques

commentés, quelques exemples de possibilités offertes par les composantes spécifiques, et par la décomposition canonique globale de [X] + [Y]. Dans notre exemple, les composantes spécifiques ne concernent que le sous-espace [Y] : ce sont v_4 et v_5. Chacune d'elles est, par construction, incorrélée tant avec les variables x_r qu'avec les autres composantes canoniques u_k et v_t. Leurs corrélations avec les variables y_s sont consignées dans le tableau suivant, dans lequel on remarque notamment d'une part la forte corrélation entre v_4 et y_3, d'autre part les corrélations relativement élevées (en valeur absolue) qu'a v_5 avec y_1 et y_5.

	y_1	y_2	y_3	y_4	y_5
v_4	-0.4078	0.3292	0.9389	0.1758	-0.5625
v_5	0.7868	-0.0918	0.0033	-0.1589	-0.6155

L'ensemble des corrélations entre les 8 variables du tableau 1, et les deux composantes spécifiques v_4 et v_5, fait l'objet des représentations graphiques de la figure 3.

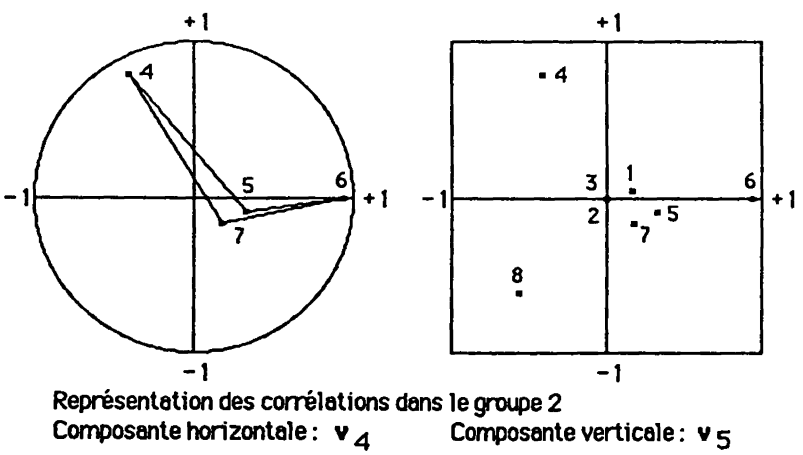


Fig 3. Représentation des corrélations des diverses variables, avec les deux composantes spécifiques du mouvement, soit v_4 et v_5. Les variables sont numérotées dans l'ordre des colonnes du tableau 1 : 1=i.p., 2=x_1, 3=x_2, 4=y_1, 5=y_2, 6=y_3, 7=y_4, 8=y_5 (voir commentaire dans le texte).

La partie gauche de cette figure 3 est le "cercle de corrélation" contenant les projections, sur le plan [v_4, v_5], des quatre variables engendrant l'espace [Y]. On note la quasi-identité entre y_3 et v_4, ainsi que la bonne proximité entre y_4 et v_5. La "trajectoire" reliant les quatre points rappelle seulement la relation d'ordre chronologique cyclique entre ces quatre variables.

A droite, dans un carré, sont représentées les projections sur ce même plan des huit variables considérées au tableau 1. Par rapport au système d'axes, les variables 4 à 7 occupent les mêmes positions que dans le graphique de gauche ; les variables 2 et 3, appartenant à l'espace [X], sont évidemment représentées à l'origine des coordonnées, puisqu'elles sont incorrélées à v_4 et v_5 ; les variables supplémentaires 1 (indice de performance) et 8 (phase 5 du mouvement) sont aussi représentées.

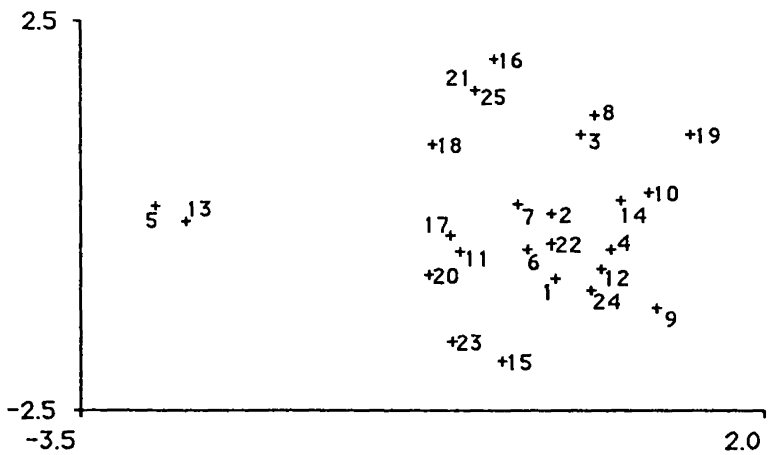


Fig 4. Représentation des 25 nageurs sur le plan défini par les deux composantes canoniques spécifiques du mouvement : v_4 (axe horizontal) et v_5 (axe vertical). On peut se rendre compte de la position particulière occupée par les sujets numéros 5 et 13.

La représentation graphique des 25 individus dans le plan [v_4,v_5], met en évidence la position très excentrique des numéros 5 et 13, caractérisés par une valeur de v_4 très inférieure aux valeurs de v_4 chez les autres nageurs (fig. 4). Or v_4 est fortement liée à y_3, et de fait les deux nageurs en question sont les deux dont la phase 3 du mouvement est

proportionnellement la plus rapide (tab. 1). Par contre l'ensemble des 23 autres nageurs est représenté par un nuage de points approximativement discoïdal, suggérant une bonne homogénéité de cet ensemble, et une éventuelle reprise de l'analyse après avoir écarté les sujets numéros 5 et 13.

L'espace [X] + [Y], engendré par l'ensemble des 7 variables $1, x_1, x_2, y_1, y_2, y_3, y_4$, se trouve muni, par la décomposition canonique complète ci-dessus, d'une base φ-orthonormée constituée à partir des bases canoniques respectives de [X] et de [Y], comme expliqué au § 3.3. L'espace [X] + [Y] est de dimension 7 ; la composition des vecteurs w_t (t = 1 à 7) de cette base canonique est indiquée dans le tableau 2. $w_1 = 1$ engendre l'intersection [X] $\cap$ [Y], réduite ici aux constantes. w_6 et w_7 s'identifient respectivement à v_4 et v_5, engendrant le sous-espace spécifique $[Y]_3$.

	1	x_1	x_2	y_1	y_2	y_3	y_4
w_1	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
w_2	-2.8973	3.9905	-1.6622	-0.0351	0.1146	-0.0012	-0.2482
w_3	-18.6187	7.1143	-2.9633	0.0627	-0.2042	0.0021	0.4425
w_4	-2.5270	-1.4201	6.3859	-0.0469	-0.3473	0.1257	-0.1144
w_5	-14.5204	-1.6391	7.3704	0.0541	0.4008	-0.1451	0.1320
w_6	10.0513	0.0000	0.0000	-1.1063	0.5807	1.5712	0.4173
w_7	-64.3213	0.0000	0.0000	2.1341	-0.1620	0.0055	-0.3770

Tab 2. Valeurs numériques des coefficients des variables $1, x_r, y_s$, dans l'expression de chaque composante canonique w_t. On note en particulier que w_1 s'identifie à 1, que w_2 à w_5 sont des composantes "interdépendantes" de [X] et de [Y], enfin que w_6 et w_7 sont spécifiques de [Y].

4 - EXTENSIONS DE L'ANALYSE CANONIQUE COMPLETE

4.1 - Régression linéaire

Nous considérons la situation suivante. Sur les unités statistiques constituant l'ensemble Ω, nous observons trois catégories de variables : un bloc de variables $X = \{x_1, \ldots, x_p\}$, un bloc de variables $Y = \{y_1, \ldots, y_q\}$, et une variable z. Nous souhaitons exprimer z au moyen d'une fonction linéaire des x_r, faisant ressortir les éventuelles liaisons entre les x_r et les y_s. Ce problème relève donc d'une régression linéaire multiple, dont les termes seront analysés pour tenir compte des y_s. L'interprétation euclidienne de la situation est la suivante. Dans l'espace $[\Omega]$, considérons les deux sous-espaces [X] et [Y], et le vecteur z. Soit z' la projection φ-orthogonale de z sur [X] : z' est la fonction linéaire des x_r la plus proche de z (au sens des moindres carrés). La décomposition complète de [X] canonique par rapport à [Y], induit une décomposition de z' en trois composantes deux à deux φ-orthogonales :

z'_1, projection de z' sur $[X]_1$, est la partie de z' susceptible d'être exprimée comme fonction linéaire des y_s uniquement ;

z'_2, projection de z' sur $[X]_2$, est la partie de z' oblique par rapport à l'ensemble des y_r (nous entendons par là que son produit scalaire avec une quelconque fonction linéaire des y_r n'est ni nul, ni égal à 1) ;

z'_3, projection de z' sur $[X]_3$, est la partie de z' φ-orthogonale à toute fonction linéaire des y_r.

Ainsi, z s'exprime sous la forme d'une somme de quatre termes φ-orthogonaux :

$$z = z'_1 + z'_2 + z'_3 + (z-z')$$

dont les trois premiers sont chacun une fonction linéaire des x_r, chacun des trois ayant avec les y_s les relations décrites plus haut (voir fig. 5). Quant au terme z-z', c'est le "résidu", φ-orthogonal à [X] ; en cas de besoin, ce résidu peut à son tour être fractionné (par projection sur [Y]) en une part φ-orthogonale à [Y] et une part appartenant à $[Y]_3$. Les quatre termes du second membre étant deux à deux φ-orthogonaux, le carré de la norme de z se trouve décomposé en la somme des carrés des normes de ces quatre termes. Par exemple, si les variables considérées sont centrées, ceci traduit simplement une décomposabilité de la variance de z en somme de quatre variances interprétables en fonction des relations entre les x_r et les y_s.

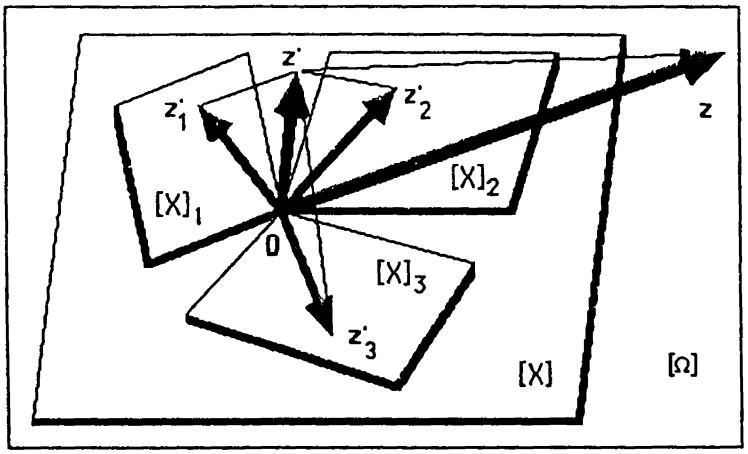


Fig 5. Le sous-espace [X] a été décomposé, dans l'analyse canonique complète avec un autre sous-espace [Y], en somme directe de trois sous-espaces φ-orthogonaux deux à deux $[X]_1$, $[X]_2$, $[X]_3$. De ce fait, la projection de z sur [X], soit z' (régression linéaire) se trouve elle-même décomposée complètement en somme de trois composantes z'_1, z'_2, z'_3 deux à deux φ-orthogonales.

Si nous projetons la variable z sur l'espace [X] + [Y], la décomposition canonique complète de ce dernier, telle que décrite au § 3.3, induit une décomposition de z en somme de cinq composantes deux à deux φ-orthogonales : les projections φ-orthogonales de z respectivement sur $[X] \cap [Y]$, sur $[X]_2 \oplus [Y]_2$, sur $[X]_3$ et sur $[Y]_3$, et le résidu, fraction de z φ-orthogonale à l'espace [X] + [Y]. La variance totale de z est alors décomposée en cinq parts additives, dont chacune est indicative de l'importance prise par telle ou telle des cinq composantes dans la variabilité totale de z ; on réalise ainsi une sorte d'analyse de la variance, susceptible à l'avenir de donner lieu à des tests statistiques de signification, adaptés à cette décomposition. A chacune des trois décompositions possibles de z, correspond une expression de z comme fonction linéaire des composantes canoniques de l'espace de projection, fonction à laquelle s'ajoute le résidu. Ainsi, pour la variable z = i.p. projetée par exemple sur l'espace [X] + [Y], nous obtenons l'expression :

$$\text{i.p.} = 88.6015\,1 + 1.2948\,w_2 + 1.7132\,w_3 + 2.4319\,w_4 + 2.7603\,w_5 + 0.8079\,w_6 + 0.2686\,w_7 + z'$$

(z' note le résidu de cette décomposition)

La figure 6 permet une visualisation, sous la forme de trois diagrammes, des parts relatives prises par chacune des composantes canoniques (autres que 1) dans la variabilité de i.p., respectivement dans [X] (premier groupe), dans [Y] (deuxième groupe), et dans [X] + [Y] (ensemble des deux groupes). Cette figure permet, mieux que le tableau des valeurs numériques de ces parts relatives de variance (tableau non donné ici), de localiser les détails de chacun des trois partages de la variance totale de i.p. Le tableau 3 donne une récapitulation des résultats numériques globaux de ces trois décompositions.

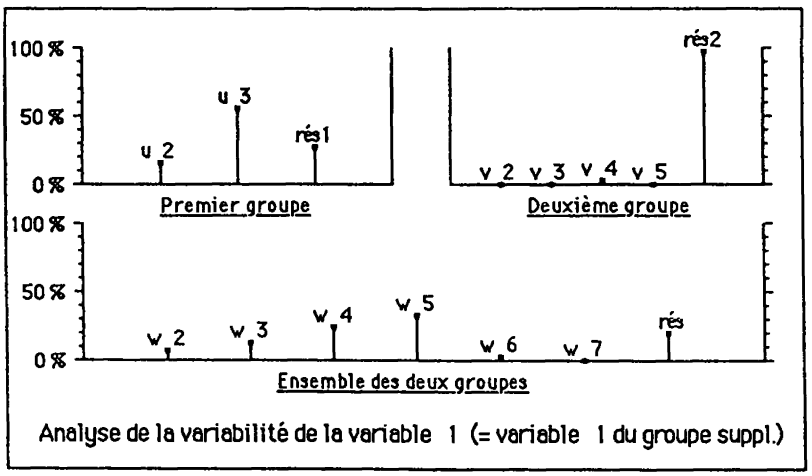


Fig 6. La variable supplémentaire i.p. étant projetée respectivement sur le sous-espace [X] (premier groupe de variables actives), sur [Y] (deuxième groupe), et sur [X] + [Y] (ensemble), sa variance se trouve décomposée en parts additives, contributions respectives des composantes canoniques du sous-espace concerné. Dans chacun des cas s'ajoute une part résiduelle, due au fait que i.p. n'appartient à aucun des trois sous-espaces.

	1er GROUPE	2ème GROUPE	ENSEMBLE
Commun	0.00	0.00	0.00
Interdép.	72.45	0.36	76.61
Spécif. 1	0.00	---	0.00
Spécif. 2	---	3.06	3.06
Résidu	27.55	96.58	20.33
TOTAUX	100.00 %	100.00 %	100.00 %
Analyse de la variabilité de la variable 1 (= variable 1 du groupe supplémentaire)			

Tab 3. La variance de la variable n°1 (i.p.) est égale à 23.6851. La projection φ-orthogonale de cette variable "supplémentaire" sur chacun des sous-espaces [X], [Y], [X] + [Y], engendre une décomposition de cette variance en parts additives, exprimées dans ce tableau en pourcentages de la variance totale. On note dans tous les cas une part résiduelle non négligeable.

4.2 - Analyse des correspondances complète

L'analyse des correspondances est l'analyse canonique entre l'ensemble $A = \{a_1, \ldots, a_p\}$ des fonctions indicatrices des p modalités d'une variable qualitative A, et l'ensemble $B = \{b_1, \ldots, b_q\}$ des fonctions indicatrices des q modalités d'une variable qualitative B. Le tableau des données se présente sous la forme suivante (dans laquelle, par convention, 01 note la présence soit de 0, soit de 1) :

obs.	poids	a_1	a_2	...	a_p	b_1	b_2	...	b_q
ω_1	π_1	01	01	...	01	01	01	...	01
ω_2	π_2	01	01	...	01	01	01	...	01
:	:	:	:		:	:	:		:
ω_n	π_n	01	01	...	01	01	01	...	01

Dans [Ω], les deux sous-espaces [A] et [B] engendrés par les deux ensembles de fonctions indicatrices, ont respectivement pour dimension p et q. Ils ont en commun au moins le sous-espace unidimensionnel [1] (ensemble des constantes) : nous sommes assurés que $[A] \cap [B]$ n'est pas réduit à **0**. L'analyse des correspondances ordinaire fournit les couples canoniques associés aux μ-1 valeurs propres strictement positives (on ne tient pas compte de la valeur propre triviale 1 associée au vecteur propre trivial **1** ; mais il peut exister d'autres valeurs propres égales à 1, non triviales, qui, elles, sont à considérer). Ces couples canoniques définissent :

* dans [A], les μ-1 codages canoniques des modalités de la variable A, codages qui sont aussi une base φ-orthogonale de $([1]^{\perp} \cap [A]_1) \oplus^{\perp} [A]_2$;

* dans [B], les μ-1 codages canoniques correspondants, des modalités de la variable B, codages constituant une base φ-orthogonale de $([1]^{\perp} \cap [B]_1) \oplus^{\perp} [B]_2$.

L'analyse canonique complète entre [A] et [B] fournit, en plus des μ-1 premiers couples canoniques, les p-μ codages de base engendrant $[A]_3 = [A] \cap [B]^{\perp}$, et les q-μ codages de base engendrant $[B]_3 = [B] \cap [A]^{\perp}$.

La décomposition complète de [A+B] peut être intéressante par exemple dans une situation d'analyse de la variance à deux facteurs contrôlés. **x** étant une "variable" (caractère quantitatif observé), A et B deux variables qualitatives (facteurs contrôlés) à respectivement p et q modalités, nous sommes amenés à considérer dans l'espace [Ω], le vecteur **x**, et les deux sous-espaces [A] et [B]. La projection φ-orthogonale de **x** sur [A+B] est notée $\mathbf{x}'_{A+B}$, et l'on a $\mathbf{x} = \mathbf{x}'_{A+B} + (\mathbf{x} - \mathbf{x}'_{A+B})$; le dernier terme de ce second membre est le "résidu", orthogonal à [A+B] ; si **x** est centré (c'est-à-dire orthogonal à [1]), les variances des différents termes de cette relation s'identifient aux carrés des normes, et la variance de **x** est ainsi décomposée en la variance de $\mathbf{x}'_{A+B}$ plus la variance résiduelle. La variance de $\mathbf{x}'_{A+B}$ est elle-même décomposée en quatre parts additives, par projections sur les quatre sous-espaces canoniques de [A+B]. De son côté, l'analyse de la variance décompose la variance de **x** en une part attribuée à A, une part attribuée à B, éventuellement une part attribuée à l'"interaction" entre A et B, et une part résiduelle. Ces deux décompositions, celle associée à l'analyse de la variance, et celle associée à l'analyse canonique complète, ne coïncident pas ; elles ont cependant entre elles des relations que nous étudions par ailleurs.

4.3 - Cas de la méthode LONGI

Nous avons eu l'occasion d'utiliser les ressources de l'analyse des correspondances complète, en construisant la méthode LONGI, que nous avons présentée aux XVIIes Journées de Statistique à Pau (1985), et au 3rd International Symposium on Data Analysis à Bruxelles (1985). Cette méthode a été créée pour analyser la situation suivante : p variables x_i (caractères morphologiques) ont été observées sur n individus i_j (enfants, en cours de croissance), chacun de ces individus ayant été observé à q dates successives d_k. L'éventualité de données manquantes doit être prise en considération, en raison de l'absence possible de l'enfant à la date prévue pour les mesures. Dans cette situation, nous nous proposions d'obtenir des indices multivariés (combinaisons linéaires des variables x_i) qui caractérisent du mieux possible le phénomène évolutif de croissance, indépendamment de l'individu. Nous nous proposions aussi d'obtenir des indices multivariés qui situent au mieux chacun des individus dans l'ensemble des individus, indépendamment de la date d'observation. Nous avons obtenu de tels indices, en procédant de la manière suivante. Considérant les deux variables qualitatives I (individu), et D (date d'observation), l'analyse canonique complète entre [I] et [D] nous a permis notamment d'obtenir $[I]_3 = [I] \cap [D]^{\perp}$, ensemble des codages des individus φ-orthogonaux à tout codage des dates, et $[D]_3 = [D] \cap [I]^{\perp}$, ensemble des codages des dates φ-orthogonaux à tout codage des individus. Ensuite, l'analyse canonique (ordinaire) entre [X] et $[D]_3$ nous a fourni les fonctions linéaires des x_i caractérisant les dates d'observation "indépendamment" des individus observés (caractérisation du phénomène de croissance) ; l'analyse canonique entre [X] et $[I]_3$ nous a fourni les fonctions linéaires des x_i caractérisant les individus "indépendamment" des dates d'observation (situation de chaque individu parmi les autres). Les données manquantes ne font nullement obstacle à cette démarche : s'il y en a, les sous-espaces $[I]_2$ et $[D]_2$ ne sont pas réduits au vecteur 0 (ce qui est le cas s'il n'y a pas de données manquantes). Nous n'avons ci-dessus que donné une esquisse de la méthode LONGI ; celle-ci est développée beaucoup plus complètement dans PERNIN [3] et dans PONTIER et PERNIN [4].

4.4 - Analyse discriminante complète

L'analyse discriminante est l'analyse canonique entre d'une part un ensemble de p variables x_r, r = 1 à p, auxquelles on adjoint implicitement la "variable", notée 1, constamment égale à 1 (c'est elle qui va fournir le coefficient de centrage), et d'autre part l'ensemble des q fonctions indicatrices c_s, s = 1 à q, des modalités d'une variable qualitative C. Le tableau des données est du type suivant :

obs.	poids	1	x_1	x_2	... x_p	c_1	c_2	... c_q
ω_1	π_1	1	x_{11}	x_{21}	... x_{p1}	01	01	... 01
ω_2	π_2	1	x_{12}	x_{22}	... x_{p2}	01	01	... 01
:	:	:	:	:	:	:	:	:
ω_n	π_n	1	x_{1n}	x_{2n}	... x_{pn}	01	01	... 01

Dans $[\Omega]$, les deux sous-espaces [X] et [C] engendrés par les deux ensembles de variables, ont respectivement pour dimension p+1 et q. Ils ont en commun au moins le sous-espace unidimensionnel [1] (ensemble des constantes) : nous sommes assurés que $[X] \cap [Y]$ n'est pas réduit à **0**. L'analyse discriminante ordinaire fournit les couples canoniques associés aux μ-1 valeurs propres strictement positives (on ne tient pas compte de la valeur propre triviale 1 associée au vecteur propre trivial **1** ; mais il peut exister d'autres valeurs propres égales à 1, non triviales, qui, elles, sont à considérer). Ces couples canoniques définissent :

* dans [X], les μ-1 fonctions discriminantes, qui sont aussi une base φ-orthogonale du sous-espace $([1]^{\perp} \cap [X]_1) \oplus^{\perp} [X]_2$;

* dans [C], les μ-1 codages correspondants, base φ-orthogonale de $([1]^{\perp} \cap [C]_1) \oplus^{\perp} [C]_2$.

La décomposition complète de [X] canonique par rapport à [C] fournit, en plus des μ-1 fonctions discriminantes, les p-μ fonctions de base engendrant $[X]_3$. Ce dernier espace, φ-orthogonal à [C], est l'ensemble de toutes les fonctions linéaires des x_r qui sont φ-orthogonales à tout codage des modalités de la variable qualitative C : en ce sens on peut dire que la détermination de $[X]_3$ est une analyse "antidiscriminante", fournissant les fonctions linéaires des x_r qui sont "indépendantes" de tout codage de C. La décomposition complète de [C] canonique par rapport à [X] fournit, en plus des μ-1 codages associés aux fonctions discriminantes, les q-μ codages de base engendrant $[C]_3$. Ce dernier espace, φ-orthogonal à [X], est l'ensemble de tous les codages des modalités de la variable qualitative C qui sont φ-orthogonaux à toute fonction linéaire des variables x_r (en particulier aux variables x_r elles-mêmes). La recherche de $[C]_3$ fournit donc les codages des modalités de C qui ne peuvent être mis en évidence par aucune combinaison linéaire des variables x_r ; peut-être est-il alors possible de parler d'analyse "super-discriminante" ?

REFERENCES

[1] ANDERSON T. W. *An introduction to multivariate statistical analysis.* John Wiley, New York, 1958.

[2] CAILLIEZ F. & PAGES J.-P. *Introduction à l'analyse des données.* S.M.A.S.H., Paris, 1976.

[3] PERNIN M.-O. *Contribution à la méthodologie d'analyse de données longitudinales. Exemple de la croissance chez l'être humain (Auxologie).* Thèse Dipl. Doct., Université Claude Bernard, Lyon, 1986.

[4] PONTIER J., PERNIN M.-O. *Multivariate and longitudinal data on growing children : solution using LONGI.* Proceedings of the Third Symposium on Data Analysis : the ins and outs of solving real problems, held june 10-12, 1985, in Brussels, Belgium.. London, Plenum ed., 1987, p. 49-65

[5] ROUARD A. *Etude biomécanique du crawl. Evolution des paramètres cinématiques et électromyographiques avec la vitesse.* Thèse Dipl. Doct., Université Claude Bernard, Lyon, 1987.

[6] TAKEUCHI K., YANAI H. & MUKHERJEE B. N.*The foundations of multivariate analysis. A unified approach by means of projection onto linear subspaces.* Wiley Eastern Ltd, New Delhi, 1984.