REVUE DE STATISTIQUE APPLIQUÉE

J. PONTIER
MYRIAM NORMAND

À propos de généralisation de l'analyse canonique

Revue de statistique appliquée, tome 40, n° 1 (1992), p. 57-75
<http://www.numdam.org/item?id=RSA_1992__40_1_57_0>

Rev. Statistique Appliquée, 1992, **XXXIX** (1), 57-75

A PROPOS DE GÉNÉRALISATION DE L'ANALYSE CANONIQUE

J. PONTIER[1] et Myriam NORMAND[2]

[1] *Laboratoire d'Analyse de Données et Biométrie, U.F.R.A.P.S.*
Université Claude Bernard Lyon, 27-29 boulevard du 11 novembre 1918
69622 VILLEURBANNE CEDEX France
[2] *C.R.E.U.S.E.T., Université de Saint Etienne, 33 rue du 11 novembre 1918*
42023 SAINT ETIENNE CEDEX France

RÉSUMÉ

La généralisation de l'analyse canonique proposée par Carroll, développée par Saporta, représente, par rapport au point de vue initial de Hotelling, un changement d'objectif. A l'analyse du comportement de chaque paquet de variables relativement à l'autre, cette généralisation substitue une vue globale sur l'ensemble de tous les paquets de variables. Nous proposons dans cet article une autre approche de cette recherche de généralisation de l'analyse canonique, respectant le point de vue initial (analyse de chaque paquet relativement aux autres). Les deux approches sont appliquées à un même exemple numérique, ce qui nous permet de mettre en évidence les potentialités de chacune, et leur complémentarité.

Mots-clés : *analyse canonique, analyse canonique généralisée, stratégie* Π, *stratégie* Σ

ABSTRACT

Generalized canonical analysis (GCA) as proposed by Carroll and expanded by Saporta, shows some change of objective in regard to the initial Hotelling point of view on canonical correlation analysis. GCA allows a global panorama upon several sets of variables, and does not analyze relations between these sets of variables. Herein we propose another way to generalize canonical correlation analysis, closest to the initial point of view. An application of the two approaches to the same numerical example allows us to show possibilities of each one, and their complementarity.

Key-words : *canonical analysis, generalized canonical analysis,* Π*-strategy,* Σ*-strategy*

1. Introduction

L'histoire de l'analyse canonique est l'histoire de la recherche de relations (linéaires) entre deux ensembles de variables quantitatives. En 1935, Hotelling [6] détermine quelle est la fonction linéaire des variables constituant l'un des deux ensembles, qui est la plus proche d'une fonction linéaire des variables constituant

l'autre ensemble (proximité mesurée par la corrélation multiple). C'est l'année suivante que cet auteur lance véritablement l'analyse canonique, en étendant et formalisant ce premier résultat, ce qu'il fait dans les termes suivants [7] :

"The statistic (1.1) is invariant under internal linear transformations of either set, as will be proved in Section 4. Another example of such a statistic is provided by the maximum multiple correlation with either set of a linear function of the other set, which has been the subject of a brief study§ . This problem of finding, not only a best predictor among the linear functions of one set, but at the same time the function of the other set which it predicts most accurately, will be solved in Section 3 in a more symmetrical manner. When the influence of these two linear functions is eliminated by partial correlation, the process may be repeated with the residuals. In this way we may obtain a sequence of pairs of variates, and of correlations between them, which in the aggregate will fully characterize the invariant relations between the sets, in so far as these can be represented by correlation coefficients. They will be called *canonical variates and canonical correlations*. Every invariant under general linear internal transformations, such for example z, will be seen to be a function of the canonical correlations.

...

... § Harold Hotelling. "The Most Predictable Criterion" in *Journal of Educational Psychology*, Vol. XXVI. pp. 139-142, February, 1935."

Généralement, on désigne par *analyse canonique* la détermination de ces corrélations canoniques entre deux ensembles de variables, et des variables canoniques correspondantes qui, ainsi associées par couples, constituent ce qu'on appelle les couples canoniques. Cependant, dans ce contexte, l'analyse peut être poussée plus loin, permettant notamment de déterminer les fonctions linéaires des variables de l'un des deux ensembles, qui sont sans corrélation avec toute fonction linéaire de l'autre. Evoqué d'abord dans l'article de Hotelling [7], puis par certains auteurs présentant la théorie de l'analyse canonique (Benzecri [1], Cailliez et Pagès [2]), cet aspect de l'analyse est semble-t-il resté longtemps sinon ignoré, du moins négligé, considéré comme sans utilité pratique. En 1987, Pontier, Jolicœur et Pernin [11] sont revenus sur ce sujet, montrant le parti que l'on pouvait tirer de la détermination de ces fonctions linéaires d'un ensemble de variables, sans corrélation avec l'autre ensemble. Ces auteurs ont proposé d'appeler *analyse canonique complète*, l'analyse canonique ainsi poussée jusqu'au bout. Pontier et Pernin [13] ont montré les relations existant entre cette analyse canonique complète et la «méthode LONGI» (Pernin [10], Pontier et Pernin [12]). Malgré un intérêt théorique considérable certain, l'analyse canonique trouve jusqu'à présent l'essentiel de ses applications concrètes dans les cas où l'un au moins des deux ensembles de variables est constitué par les indicatrices des modalités d'un caractère qualitatif : elle est alors appelée analyse discriminante dans le cas d'un caractère qualitatif, analyse des correspondances (simple) dans le cas de deux caractères qualitatifs. Tout ceci est développé de manière extensive dans la plupart des manuels d'analyse des données, par exemple dans le récent ouvrage de Pontier, Dufour et Normand [14].

La généralisation de l'analyse canonique à plus de deux ensembles de variables est un problème auquel ont été apportées des solutions diverses. L'objectif du présent article est de présenter notre contribution à la discussion et à la résolution de ce problème.

2. Vous avez dit généralisation ?

Les origines de l'analyse canonique, rappelées ci-dessus, montrent qu'au départ il s'agissait de chercher des relations «angulaires» entre deux ensembles de variables, relations exprimées en termes de maximisation de cosinus (ou de coefficient de corrélation linéaire). Le problème initial est donc, comme la notion d'angle, spécifique du nombre deux. L'analyse canonique entre deux ensembles de variables possède plusieurs propriétés intéressantes. Certaines sont spécifiques de ce nombre deux. Les autres sont éventuellement généralisables à plus de deux ensembles de variables, mais elles ne semblent pas toutes généralisables en même temps.

En 1971, Kettenring [8] passe en revue les généralisations proposées jusqu'alors, et qui possèdent les deux caractéristiques suivantes, essentielles à ses yeux :

– coïncider avec l'analyse canonique ordinaire, dans le cas où le nombre d'ensembles de variables est égal à deux ;

– mettre en évidence, dans chacun des ensembles de variables, une variable canonique (fonction linéaire des variables constituant cet ensemble), ces variables canoniques possédant la propriété d'optimiser telle ou telle fonction de leurs intercorrélations.

En 1975, Saporta [15] reprend ce problème, en accordant une attention privilégiée à la solution proposée par Carroll en 1968 [3]. En résumé, au lieu de rechercher directement une fonction optimale dans chacun des ensembles de variables, on passe par un intermédiaire : c'est dans l'ensemble de toutes les variables qu'on recherche une fonction optimale (par exemple, la plus proche, en un certain sens, de chacun des sous-ensembles de variables). Cette fonction optimale générale est ensuite projetée sur chacun des sous-ensembles, et ce sont ces projections qui constituent les variables canoniques recherchées. Saporta approfondit l'idée de Carroll, souligne qu'effectivement l'analyse proposée par cet auteur coïncide bien avec l'analyse canonique ordinaire dans le cas de deux ensembles de variables, et développe l'aspect calculatoire (matriciel) de la mise en pratique de cette méthode.

A l'heure actuelle, c'est la généralisation proposée par Carroll, formalisée par Saporta, qui est considérée comme l'*analyse canonique généralisée* (ACG). Sa variante la plus utilisée semble être l'*analyse des correspondances multiples* (ACM), qui est l'ACG appliquée au cas où les variables sont les indicatrices des modalités de caractères qualitatifs : à chacun des caractères qualitatifs correspond donc un ensemble de variables, au sens de l'analyse canonique généralisée. Les relations entre cette ACG et d'autres méthodes d'analyse des données ont été décrites par exemple par Casin et Turlot [4], par Tenenhaus [18].

Il est à noter que, dans l'analyse canonique ordinaire, on obtient directement les variables canoniques, en diagonalisant un *produit* de matrices («stratégie Π»). Dans l'analyse canonique généralisée, on obtient les variables canoniques intermédiaires (celles qu'on projettera ensuite sur les divers ensembles de variables), en diagonalisant une *somme* de matrices («stratégie Σ»). Dans le cas de *deux* ensembles de variables, les deux stratégies aboutissent en fin de compte au même

résultat : c'est cet état de fait qui justifie que l'ACG soit considérée comme une généralisation de l'analyse canonique. Par ailleurs, l'ACG est très directement une généralisation de l'analyse en composantes principales normée (ACPN). Partant de k variables, cette dernière analyse construit des fonctions linéaires de ces k variables, dont la somme des carrés des corrélations avec ces variables est maximum. L'ACG, elle, fait exactement la même chose à partir non pas de k variables, mais à partir de k *ensembles* de variables : elle construit des fonctions linéaires des variables, dont la somme des carrés des corrélations *multiples* avec ces k ensembles de variables est maximum. Cette propriété n'avait pas échappé à Saporta [15], qui conclut (p. II.14) «La méthode factorielle proposée pour généraliser l'analyse canonique est identique à une analyse en composantes principales dans laquelle chaque espace d'observation E_i relatif à un tableau X_i est muni de sa métrique de Mahalanobis V_{ii}^{-1}». Il en fut de même pour Tenenhaus qui, en 1977, voulant analyser un ensemble de tableaux disjonctifs de caractères qualitatifs, redécrivit la méthode pour réaliser une analyse en composantes principales d'un ensemble de variables nominales [17] (version ACM de l'ACG). Dans son récent ouvrage sur l'analyse des données, Saporta [16] dit explicitement (p. 194) «L'analyse canonique généralisée est donc une ACP sur des groupes de variables», et plus loin, parlant de l'ACM (p. 217) «cette méthode possède des propriétés qui ... en font l'équivalent de l'analyse en composantes principales pour des variables qualitatives».

En définitive, l'ACG apparaît comme une généralisation directe de l'ACPN (stratégie Σ), comme une généralisation indirecte de l'analyse canonique (stratégie Π). C'est la raison qui pousse Pagès, Cailliez et Escoufier [9], en 1979, à déclarer préférer appeler cette méthode *analyse en composantes principales réduite généralisée*. Nous partageons pleinement ce point de vue. Chacune des deux stratégies a des avantages et des inconvénients. S'agissant d'explorer les relations entre plusieurs ensembles de variables, par le moyen d'une généralisation de l'analyse canonique, nous trouvons regrettable le fait d'avoir privilégié la seule stratégie Σ. Etant une généralisation de l'ACPN, l'analyse canonique généralisée se prête très bien à une formalisation du type classique «triplet / schéma de dualité». Nous pensons que ce n'est là ni une raison nécessaire, ni une raison suffisante, pour tourner le dos à la stratégie Π, dans ce problème de généralisation de l'analyse canonique. Nous consacrerons le paragraphe suivant au développement d'une proposition à ce sujet.

3. Une généralisation selon la stratégie Π

L'objectif majeur des analyses factorielles est d'obtenir des «facteurs», fonctions linéaires de variables, possédant diverses propriétés optimales, et deux à deux orthogonaux. C'est cette orthogonalité des facteurs qui autorise la décomposition de chaque variable elle-même, en fragments deux à deux orthogonaux, d'où s'ensuivent les diverses interprétations relationnelles de graphiques tels que les cercles de corrélations et les cartes factorielles.

Pour entrer dans le détail de notre propos, nous introduisons quelques notations, qui sont celles utilisées dans [11], [12], [13], [14].

• $\Omega = \{\omega_1, \omega_2, ..., \omega_n\}$ est un ensemble de n unités d'observation (individus, unités statistiques, ...) ;

• $\{\pi_1, \pi_2, ..., \pi_n\}$ est un ensemble de «poids» des unités d'observation, $0 < \pi_j < 1$, $\sum_{j=1}^{n} \pi_j = 1$;

• $[\Omega]$ est l'ensemble des applications $\Omega \to \mathbb{R}$ (les éléments de $[\Omega]$ sont habituellement appelés les «variables», ou «variables quantitatives») ; parmi les éléments de $[\Omega]$ figurent les indicatrices des unités d'observation, soit :

Ω_j = l'indicatrice de ω_j, définie par $\Omega_j(w) = 1$ si $\omega = \omega_j$, $\Omega_j(\omega) = 0$ sinon $(j = 1, 2, ..., n)$

L'ensemble $[\Omega]$ a une structure d'espace euclidien de dimension n, dans lequel le produit scalaire est défini par l'expression : $\varphi(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{n} \pi_j \mathbf{x}(\omega_j)\mathbf{y}(\omega_j)$ (métrique des poids des individus). L'ensemble $\{\Omega_1, \Omega_2, ..., \Omega_n\}$ des indicatrices est une base φ-orthogonale de $[\Omega]$ (la base «indicatrice»).

• Tout ensemble $X = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_p\}$ d'éléments de $[\Omega]$ engendre un sous-espace de $[\Omega]$, noté $[X]$, de dimension au plus égale à p. Φ_{XX} note la matrice des produits scalaires $\varphi(\mathbf{x}_s, \mathbf{x}_t)$, $s = 1$ à p, $t = 1$ à p. En particulier $\Phi_{\Omega\Omega}$, matrice des produits scalaires des indicatrices, est la matrice diagonale d'ordre n dont les termes diagonaux sont les poids π_j.

• Si $X = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_p\}$ et $Y = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_q\}$ sont deux ensembles d'éléments de $[\Omega]$, Φ_{XY} note la matrice des produits scalaires croisés $\varphi(\mathbf{x}_s, \mathbf{y}_t)$, $s = 1$ à p, $t = 1$ à q.

Pour simplifier l'exposé, on supposera ci-après que les $\mathbf{x}_i$ (resp. $\mathbf{y}_j$) sont linéairement indépendants, cas auquel il est toujours possible de se ramener quelle que soit la situation initiale. Ceci assure que la dimension de $[X]$ (resp. $[Y]$) est exactement égale à p (resp. q), et que la matrice Φ_{XX} (resp. Φ_{YY}) est inversible.

Si $\mathbf{z}$ est un vecteur de $[X]$, il est aussi un vecteur de $[\Omega]$. Pour ce vecteur $\mathbf{z}$, nous serons donc éventuellement amenés à distinguer sa matrice par rapport à la base indicatrice de $[\Omega]$, matrice notée $\mathbf{Z}_\Omega$, et sa matrice par rapport à la base de $[X]$, matrice notée $\mathbf{Z}_X$. Ces deux matrices de $\mathbf{z}$ sont liées par les relations $\mathbf{Z}_\Omega = \mathbf{X}_\Omega \mathbf{Z}_X$ et $\mathbf{Z}_X = \Phi_{XX}^{-1}\, {}^t\mathbf{X}_\Omega \Phi_{\Omega\Omega} \mathbf{Z}_\Omega$, où $\mathbf{X}_\Omega$ note le tableau (n lignes, p colonnes) contenant les valeurs individuelles x_{ij} ($i = 1$ à p, $j = 1$ à n) des variables $\mathbf{x}_i$: ce tableau est ici interprété comme une matrice, constituée par la concaténation horizontale des matrices $\mathbf{X}_{i_\Omega}$ (matrices des $\mathbf{x}_i$ par rapport à la base indicatrice de $[\Omega]$.

Si $\mathbf{z}$ est un vecteur de$[\Omega]$, nous noterons proj$(\mathbf{z}|[X])$ sa projection φ-orthogonale sur le sous-espace $[X]$.

Enfin, nous serons amenés à considérer la somme directe de deux sous-espaces $[X]$ et $[Y]$ de $[\Omega]$, dans la situation particulière où ces deux sous-espaces sont φ-orthogonaux (tout vecteur non nul de l'un est φ-orthogonal à tout vecteur non nul de l'autre). Dans ce cas, nous parlerons de somme directe φ-orthogonale des deux sous-espaces, somme que nous noterons : $[X] \oplus^{\perp} [Y]$.

Rappelons maintenant les résultats fondamentaux de l'analyse canonique complète des deux ensembles X et Y, qui est aussi l'analyse canonique complète des deux sous-espaces $[X]$ et $[Y]$ de $[\Omega]$ (*cf.* [11], [14]).

Proposition 1 (analyse canonique complète de $[X]$ et $[Y]$). *Soient $[X]$ et $[Y]$ deux sous-espaces de $[\Omega]$, de dimensions respectives p et q, engendrés respectivement par $X = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_p\}$ et $Y = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_q\}$. L'analyse canonique de $[X]$ et $[Y]$ repose sur les éléments propres des deux matrices :*

$$\mathbf{A}_{XX} = \Phi_{XX}^{-1}\Phi_{XY}\Phi_{YY}^{-1}\Phi_{YX} \quad \textit{et} \quad \mathbf{B}_{YY} = \Phi_{YY}^{-1}\Phi_{YX}\Phi_{XX}^{-1}\Phi_{XY}$$

Ces éléments propres sont ainsi caractérisés :

a) Les deux matrices ont en commun μ valeurs propres strictement positives dont

- *ν valeurs propres égales à 1 :*

$$\lambda_1 = \lambda_2 = \ldots = \lambda_\nu = 1$$

- *$\mu - \nu$ valeurs propres strictement comprises entre 0 et 1 :*

$$1 > \lambda_{\nu+1} \geq \lambda_{\nu+2} \geq \ldots \geq \lambda_\mu > 0$$

Les $p - \mu$ (resp. $q - \mu$) autres valeurs propres sont nulles :

$$\lambda_{\mu+1} = \lambda_{\mu+2} = \ldots = \lambda_p = 0 \; ; \quad \lambda'_{\mu+1} = \lambda'_{\mu+2} = \ldots = \lambda'_q = 0$$

b) Dans $[X]$, les p vecteurs propres $\mathbf{u}_1$, $\mathbf{u}_2$, ..., $\mathbf{u}_p$ de $\mathbf{A}_{XX}$, φ-normés, appelés les fonctions canoniques *ou* vecteurs canoniques *de $[X]$, constituent une base φ-orthonormée de $[X]$ appelée* la base de $[X]$ canonique par rapport à $[Y]$*; dans $[Y]$, les q vecteurs propres $\mathbf{v}_1$, $\mathbf{v}_2$, ..., $\mathbf{v}_q$ de $\mathbf{B}_{YY}$, φ-normés, appelés les* fonctions canoniques *ou* vecteurs canoniques *de $[Y]$, constituent une base φ-orthonormée de $[Y]$ appelée* la base de $[Y]$ canonique par rapport à $[X]$.

c) Les fonctions canoniques $\mathbf{u}_r$ et $\mathbf{v}_r$, associées à la même valeur propre λ_r strictement positive, constituent le r-ème couple canonique. *Les deux éléments d'un même couple canonique sont liés entre eux par les relations exprimées en d et e ci-après. Chaque élément d'un couple canonique est φ-orthogonal aux éléments des autres couples canoniques.*

d) Si 1 est valeur propre (commune à $\mathbf{A}_{XX}$ et à $\mathbf{B}_{YY}$), d'ordre de multiplicité ν, alors les vecteurs propres $\mathbf{u}_1$, $\mathbf{u}_2$, ..., $\mathbf{u}_\nu$ de $\mathbf{A}_{XX}$ associés à cette valeur propre 1, et les vecteurs propres $\mathbf{v}_1$, $\mathbf{v}_2$, ..., $\mathbf{v}_\nu$ de $\mathbf{B}_{YY}$ associés à cette même valeur propre, engendrent le même sous-espace de $[\Omega]$, soit l'intersection $[X] \cap [Y]$, de dimension ν, dont ils constituent une base :

$$[\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n] = [X]_1 = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_\nu] = [Y]_1 = [X] \cap [Y]$$

e) $\mathbf{u}_{\nu+1}, \ldots, \mathbf{u}_\mu$, *vecteurs propres de* $\mathbf{A}_{XX}$ *associés aux* $\mu - \nu$ *valeurs propres strictement comprises entre* 0 *et* 1, *et* $\mathbf{v}_{\nu+1}, \ldots, \mathbf{v}_\mu$, *vecteurs propres de* $\mathbf{B}_{YY}$ *associés respectivement à ces mêmes valeurs propres, sont tels que :*

$$\begin{cases} \mathbf{v}_r = \dfrac{1}{\sqrt{\lambda_r}}\widehat{\mathbf{u}}_r = \dfrac{1}{\sqrt{\lambda_r}} \operatorname{proj}(\mathbf{u}_r \mid [Y]); \ \mathbf{V}_{rY} = \dfrac{1}{\sqrt{\lambda_r}}\Phi_{YY}^{-1}\Phi_{YX}\mathbf{U}_{rX} \\ \mathbf{u}_r = \dfrac{1}{\sqrt{\lambda_r}}\widehat{\mathbf{v}}_r = \dfrac{1}{\sqrt{\lambda_r}} \operatorname{proj}(\mathbf{v}_r \mid [X]); \ \mathbf{U}_{rX} = \dfrac{1}{\sqrt{\lambda_r}}\Phi_{XX}^{-1}\Phi_{XY}\mathbf{U}_{rY} \\ \cos(\mathbf{u}_r, \mathbf{v}_r) = \sqrt{\lambda_r} \qquad (r = \nu+1, \ldots, \mu) \end{cases}$$

$\mathbf{u}_{\nu+1}, \ldots, \mathbf{u}_\mu$ *engendrent le sous-espace* $[X]_2$, *de dimension* $\mu - \nu$, *ensemble des vecteurs de* $[X]$ *obliques par rapport à* $[Y]$ *(c'est-à-dire : formant avec leur projection* φ*-orthogonale sur* $[Y]$ *un angle dont le carré du cosinus n'est ni nul, ni égal à 1).*

$\mathbf{v}_{\nu+1}, \ldots, \mathbf{v}_\mu$ *engendrent le sous-espace* $[Y]_2$, *de dimension* $\mu - \nu$, *ensemble des vecteurs de* $[Y]$ *obliques par rapport à* $[X]$.

f) Tous les vecteurs propres $\mathbf{u}_{\mu+1}, \ldots, \mathbf{u}_p$ *sont* φ*-orthogonaux à* $[Y]$*; ils engendrent* $[X]_3 = [X] \cap [Y]^\perp$, *de dimension* $p - \mu$.

Tous les vecteurs propres $\mathbf{v}_{\mu+1}, \ldots, \mathbf{v}_q$ *sont* φ*-orthogonaux à* $[X]$*; ils engendrent* $[Y]_3 = [Y] \cap [X]^\perp$, *de dimension* $q - \mu$.

g) On a les relations :

$$[X] = [X]_1 \oplus^\perp [X]_2 \oplus^\perp [X]_3; \quad [Y] = [Y]_1 \oplus^\perp [Y]_2 \oplus^\perp [Y]_3$$

Notons $[X] + [Y]$ le sous-espace de $[\Omega]$ engendré par $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_p, \mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_q\}$. $[X] + [Y]$ est de dimension au plus égale à $p + q$. L'analyse canonique de $[X]$ et $[Y]$ induit une décomposition «canonique» de $[X] + [Y]$:

Proposition 2 (décomposition canonique complète de $[X] + [Y]$**).** *Les décompositions canoniques complètes des deux sous-espaces* $[X]$ *et* $[Y]$ *l'un par rapport à l'autre (prop. 1 ci-dessus) induisent une décomposition canonique complète du sous-espace* $[X] + [Y]$, *sous la forme :*

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

dans laquelle :

- $[X] \cap [Y] = [X]_1 = [Y]_1$ *a pour base* φ*-orthonormée* $(\mathbf{w}_1, ..., \mathbf{w}_\nu)$

  où $\mathbf{w}_i = \mathbf{u}_i = \mathbf{v}_i = \frac{1}{2}(\mathbf{u}_i + \mathbf{v}_i)$*; on a* $\dim([X] \cap [Y]) = \nu$*;*

  $[X] \cap [Y]$ *est* la composante de $[X] + [Y]$ commune à $[X]$ et à $[Y]$*;*

- $[X]_2 \oplus [Y]_2$ *a pour base* φ*-orthonormée* $(\mathbf{w}_{\nu+1}, ..., \mathbf{w}_{2\mu-\nu})$

  où $\mathbf{w}_{2i-\nu-1} = \dfrac{1}{\sqrt{2+2\sqrt{\lambda_i}}}(\mathbf{u}_i + \mathbf{v}_i)$ *et* $\mathbf{w}_{2i-\nu} = \dfrac{1}{\sqrt{2-2\sqrt{\lambda_i}}}(\mathbf{u}_i - \mathbf{v}_i)(i = \nu+1$ *à* $\mu)$*;*

on a $\dim([X]_2 \oplus [Y]_2) = 2(\mu - \nu)$;

$[X]_2 \oplus [Y]_2$ *est* la composante de $[X] + [Y]$ interdépendante de $[X]$ et $[Y]$;

- $[X] \cap [Y]^\perp$ *a pour base* φ*-orthonormée* $(\mathbf{w}_{\mu-\nu+\mu+1}, ..., \mathbf{w}_{\mu-\nu+p})$

  où $\mathbf{w}_{\mu-\nu+i} = \mathbf{u}_i (i = \mu + 1$ *à* $p)$; $\dim([X] \cap [Y]^\perp) = p - \mu$;

  $[X] \cap [Y]^\perp$ *est* la composante de $[X] + [Y]$ spécifique de $[X]$;

- $[Y] \cap [X]^\perp$ *a pour base* φ*-orthonormée* $(\mathbf{w}_{p-\nu+\mu+1}, ..., \mathbf{w}_{p-\nu+q})$

  où $\mathbf{w}_{p-\nu+i} = \mathbf{v}_i (i = \mu + 1$ *à* $q)$; $\dim([Y] \cap [X]^\perp) = q - \mu$;

  $[Y] \cap [X]^\perp$ *est* la composante de $[X] + [Y]$ spécifique de $[Y]$.

On a $\dim([X] + [Y]) = p + q - \nu$. *L'ensemble des vecteurs* $\mathbf{w}_i (i = 1$ *à* $p + q - \nu)$ *est une base* φ*-orthonormée de* $[X] + [Y]$.

Nous retiendrons de tout ceci que l'analyse canonique complète des deux sous-espaces $[X]$ et $[Y]$ induit une factorisation complète φ-orthogonale de chacun des trois sous-espaces $[X]$, $[Y]$, et $[X] + [Y]$: dans chacun d'eux, cette analyse permet d'obtenir une base φ-orthonormée, dont les éléments ont des relations angulaires particulières avec les autres espaces.

Considérons maintenant k ensembles de variables X_1, X_2, ..., X_k, avec $X_i = \{\mathbf{x}_1^i, \mathbf{x}_2^i, \ldots, \mathbf{x}_{p_i}^i\}$. X_i engendre le sous-espace $[X_i]$. Notons $[X] = [X_1] + [X_2] + \ldots + [X_k]$ le sous-espace de $[\Omega]$ engendré par l'ensemble des $p_1 + p_2 + \ldots + p_k$ variables $\mathbf{x}_s^i$, $s = 1$ à p_i, $i = 1$ à k.

L'analyse canonique généralisée, au sens de Carroll, permet d'obtenir une base φ-orthogonale de $[X]$, dont les éléments possèdent des propriétés, rappelées plus haut, de proximité relativement aux sous-espaces $[X_i]$, $i = 1$ à k. La projection des vecteurs de cette base sur chacun des sous-espaces $[X_i]$, définit dans ce sous-espace des « variables canoniques ». Celles-ci ne sont cependant pas deux à deux orthogonales, *sauf dans le cas particulier où* $k = 2$ (*cf.* Saporta [15], [16]). De ce fait, dans la pratique, elles ne sont qu'exceptionnellement utilisées. *A contrario*, ce sont les projections des variables sur $[X]$ rapporté à sa base φ-orthogonale, qui captent toute l'attention de l'analyste des données. Tout ceci, d'une part met bien en relief le rôle « analyse en composantes principales » joué par l'analyse canonique généralisée, et d'autre part fait jouer à l'espace $[X]$ (et ses sous-espaces définis par les composantes obtenues par l'ACG, stratégie Σ) le rôle du *compromis* au sens décrit par Escoufier [5].

La fonction première de l'analyse canonique, au sens de Hotelling, entre les deux sous-espaces $[X]$ et $[Y]$, est la caractérisation de chacun de ces deux sous-espaces d'après le comportement angulaire de ses éléments relativement à l'autre sous-espace ; d'où une factorisation complète φ-orthogonale de $[X]$ (resp. $[Y]$) répartissant les éléments de $[X]$ (resp. $[Y]$) en trois catégories : ceux qui sont dans $[Y]$ (resp. $[X]$), ceux qui sont « obliques » par rapport à $[Y]$ (resp. $[X]$), ceux qui sont φ-orthogonaux à $[Y]$ (resp. $[X]$). Cette factorisation complète φ-orthogonale de $[X]$ comme de $[Y]$, est obtenue directement par l'analyse canonique complète (stratégie Π). Elle est obtenue indirectement par l'ACG (stratégie Σ), en passant par l'intermédiaire d'une factorisation complète φ-orthogonale de $[X] + [Y]$, car il se trouve que, *dans ce cas particulier de* deux *sous-espaces* $[X]$ et $[Y]$, la base de

$[X]$ canonique par rapport à $[Y]$, et la base de $[Y]$ canonique par rapport à $[X]$, sont les projections φ-orthogonales respectivement sur $[X]$ et sur $[Y]$, de la base de $[X]+[Y]$ déterminée par l'ACG. Mais cette propriété n'est plus vraie dès lors que l'on considère plus de deux sous-espaces. Ainsi, l'ACG gomme complètement la fonction première de l'analyse canonique, en n'aboutissant pas – sauf dans le cas particulier de deux sous-espaces – à une factorisation complète φ-orthogonale de chacun des sous-espaces, factorisation caractérisant la comportement angulaire des vecteurs de ce sous-espace relativement aux vecteurs des autres sous-espaces. Le passage de l'analyse canonique ordinaire (deux sous-espaces) à l'ACG (un nombre quelconque de sous-espaces) se fait donc par la généralisation d'une propriété somme toute secondaire de l'analyse canonique ordinaire, généralisation au prix d'une perte de l'objectif premier (Hotelling) de l'analyse canonique.

Les remarques que nous venons de formuler ci-dessus ne diminuent en rien les mérites de l'ACG, en tant que méthode factorielle permettant d'obtenir une bonne vue globale d'un ensemble de plusieurs paquets de variables. Ces remarques avaient seulement pour but de mettre en évidence ce qui, à partir de la situation particulière de deux paquets de variables, est effectivement généralisé, et ce qui ne l'est pas. Or, nous pensons que ce qui n'est pas généralisé par l'ACG mérite quand même quelque considération. C'est pourquoi nous proposons une autre approche, plus proche du point de vue de Hotelling, ayant pour objectif d'obtenir, dans chacun des sous-espaces $[X_i]$, une base φ-orthogonale ayant des relations optimales de proximité avec l'ensemble des variables n'appartenant pas à ce sous-espace.

Plus précisément, notons $[X_i]'$ le sous-espace de $[\Omega]$ (et aussi de $[X]$) engendré par l'ensemble de toutes les variables actives sauf celles définissant $[X_i]$:

$$[X_i]' = [X_1] + \ldots + [X_{i-1}] + [X_{i+1}] + \ldots + [X_k]$$

Cette définition ne fait pas obstacle à l'éventualité que certaines des variables $\mathbf{x}_s^i$ définissant $[X_i]$ (et donc supprimées de la liste des variables engendrant $[X_i]'$) puissent se trouver aussi dans d'autres sous-espaces, ou être fonctions linéaires de variables appartenant à d'autres sous-espaces. A ce titre-là, elles figurent aussi dans $[X_i]'$.

L'analyse canonique complète des deux sous-espaces $[X_i]$ et $[X_i]'$ produit les résultats rappelés ci-avant à la proposition 1. En particulier, elle permet d'obtenir une base φ-orthonormée de $[X_i]$, permettant de décomposer ce sous-espace en ses trois sous-espaces canoniques relativement à $[X_i]'$, c'est-à-dire :

• $[X_i]_1$, ensemble des vecteurs de $[X_i]$ communs à $[X_i]$ et à $[X_i]'$. Tout vecteur de $[X_i]_1$ est fonction linéaire de certains vecteurs de $[X_i]'$, et pas forcément de vecteurs appartenant à l'un seulement des sous-espaces $[X_s](s = 1, ..., i-1, i+1, ..., k)$; $[X_i]_1$ est donc la composante de $[X_i]$ exactement reconstituée à l'aide de fonctions linéaires de variables appartenant aux autres sous-espaces. C'est la composante de $[X_i]$ *commune* avec l'ensemble des autres sous-espaces.

• $[X_i]_2$, ensemble des vecteurs de $[X_i]$ obliques par rapport à $[X_i]'$. Tout vecteur de $[X_i]_2$ forme un angle ni nul, ni droit, avec tout vecteur de $[X_i]'$; $[X_i]_2$ est donc la composante de $[X_i]$ approximativement reconstituée à l'aide de fonctions

linéaires de variables appartenant aux autres sous-espaces. C'est la composante de $[X_i]$ *oblique* par rapport à l'ensemble des autres sous-espaces.

• $[X_i]_3$, ensemble des vecteurs de $[X_i]$ φ-orthogonaux à $[X_i]'$. Tout vecteur de $[X_i]_3$ est φ-orthogonal à tout vecteur de $[X_i]'$; $[X_i]_3$ est donc la composante de $[X_i]$ non reconstituable à l'aide de fonctions linéaires de variables appartenant aux autres sous-espaces, c'est la composante *spécifique* de $[X_i]$.

La réitération de cette analyse canonique complète, pour chacun des sous-espaces $[X_i]$, $i = 1$ à k, munit donc chacun de ces sous-espaces d'une base φ-orthonormée, ayant des relations angulaires «intéressantes» avec l'ensemble des autres sous-espaces. Cette opération est *une* généralisation de l'analyse canonique (complète) de deux sous-espaces, en ce sens qu'elle coïncide avec cette analyse lorsque $k = 2$. Dans le paragraphe suivant, nous allons mettre en pratique, sur un même exemple numérique, l'ACG au sens de Carroll et Saporta, et la méthode que nous proposons ici. Pour la commodité de l'exposé, nous désignerons ces deux généralisations de l'analyse canonique, respectivement par les sigles ACGΣ et ACGΠ.

4. Exemple

Les données que nous analyserons ici sont publiés dans Pontier, Dufour, Normand [14] (tableau A.3 p. 392-394). Il s'agit d'observations faites sur 64 handballeurs, soit huit caractères morphologiques quantitatifs, et deux caractères qualitatifs : le poste de jeu (cinq modalités), et le niveau de jeu de l'équipe à laquelle ils appartiennent (quatre modalités). Dans cette analyse, les 17 variables qui résultent de ces observations ont été réparties en quatre groupes, définis ci-dessous.

Le premier groupe est constitué par les mesures de 4 caractères morphologiques, liés à l'individu dans le sens longitudinal :

1. Taille debout (TD)
2. Taille assis (TA) ou segment supérieur (du siège au sommet du crâne)
3. Hauteur utilisable (HU) (hauteur totale, bras levés)
4. Longueur des jambes (LJ)

Le deuxième groupe est constitué par les mesures de 4 caractères morphologiques liés à la «largeur», constitutive ou fonctionnelle, de l'individu :

1. Diamètre biacromial (DB) (largeur d'épaule)
2. Envergure (EN)
3. Longueur des bras (LB)
4. Empan de la main porteuse de la balle (EM)

Le troisième groupe est constitué par les indicatrices des modalités du caractère qualitatif poste de jeu :

1. Ailier (AI)
2. Arrière Centre (AC)

3. Pivot (PI)

4. Arrière Latéral (AL)

5. Gardien (GA)

Le quatrième groupe est constitué par les indicatrices des modalités du caractère qualitatif niveau de jeu :

1. Equipe de niveau national 1A (NA)

2. Bataillon de Joinville (BJ)

3. Equipe de niveau national 1B (NB)

4. Equipe de niveau national 2 (N2).

Notons $[\Omega]$ l'espace (de dimension 64) engendré par les indicatrices des handballeurs, $[X_1]$ et $[X_2]$ les sous-espaces engendrés respectivement par les variables centrées constituant les premier et deuxième groupes, $[X_3]$ et $[X_4]$ les sous-espaces engendrés respectivement par les indicatrices constituant les troisième et quatrième groupes.

Sur ces données, nous avons réalisé les deux analyses canoniques généralisées : ACGΣ, ACGII. Notre objectif n'est pas d'analyser en profondeur la situation concrète d'où sont issues ces données numériques : il s'agit pour nous d'illustrer une méthode d'analyse. Aussi, nous prions le lecteur d'excuser la brièveté de nos commentaires, et l'absence d'un certain nombre de graphiques, classiques ou non, qu'autorisent les méthodes utilisées.

ACG stratégie Σ

La stratégie Σ permet d'obtenir une base φ-orthogonale particulière du sous-espace $[X] = [X_1] + [X_2] + [X_3] + [X_4]$ engendré par l'ensemble des 17 variables. Avec 15 valeurs propres non nulles, ce sous-espace se révèle être de dimension 15. Les 15 composantes canoniques, éléments de cette base φ-orthogonale, définissent des sous-espaces factoriels, en particulier des plans factoriels, permettant d'obtenir des représentations synthétiques, globales, de l'ensemble des 17 variables, ou de l'ensemble des 64 individus.

Le tableau 1 donne, pour chacune des 15 composantes canoniques issues de l'ACGΣ, la somme des carrés des quatre coefficients de corrélation multiple entre cette composante canonique et, respectivement, chacun des quatre groupes de variables (aucune de ces sommes ne peut donc être supérieure à 4, maximum théorique). Parmi les multiples représentations graphiques possibles, nous présentons ici seulement les deux les plus classiques : les corrélations des variables initales avec les deux premières composantes canoniques (figure 1 : cercle de corrélation), et les coordonnées individuelles des 64 handballeurs selon ces deux composantes (figure 2 : carte factorielle).

On notera, à partir de la figure 1, que les variables les plus étroitement (et positivement) corrélées avec la première composante canonique caractérisent les dimensions générales du corps, tant en longueur (TD, LB, LJ) qu'en largeur (DB, EN) : cette composante est un «facteur de taille». L'interprétation de la

TABLEAU 1
Sommes des carrés des coefficients de corrélation multiple
des quatre groupes de variables avec chaque composante

Numéro de la composante canonique	Somme des carrés des coefficients de corrélation multiple
1	2.55
2	1.87
3	1.55
4	1.42
5	1.30
6	1.67
7	0.99
8	0.96
9	0.84
10	0.67
11	0.55
12	0.46
13	0.29
14	0.27
15	0.07

deuxième composante canonique est moins nette : elle rassemble les joueurs de niveau national A (NA), ceux qui ont un empan important (EM), les pivots (PI)... La carte factorielle de la figure 2 confirme les tendances exprimées dans le cercle de corrélation : par exemple, les pivots (individus numéros 4, 8, 9, 11, 22, 24, 45, 48, 50, 56, 59) sont tous dans la partie supérieure gauche du graphique.

ACG stratégie II

La stratégie II permet de compléter l'analyse précédente, notamment dans l'étude des relations entre les groupes de caractères. A titre d'exemple, nous avons représenté les relations du premier groupe de caractères (4 variables quantitatives) avec les trois autres groupes (figure 3), et celles du troisième groupe (5 indicatrices des modalités d'une variable qualitative) avec les trois autres groupes (figure 4).

Précisons le principe des représentations graphiques utilisées dans ces figures 3 et 4. L'analyse canonique complète de $[X_i]$ et de $[X_i]' = [X_1] + ... + [X_{i-1}] + [X_{i+1}] + ... + [X_k]$ permet d'obtenir une base φ-orthonormée de $[X_i]$, canonique par rapport à $[X_i]'$. La dimension de $[X_i]$ étant p_i, sa base canonique est constituée de p_i vecteurs, notés $\mathbf{u}_{i1}$, $\mathbf{u}_{i2}$, ..., $\mathbf{u}_{ip_i}$, de norme unité, deux à deux φ-orthogonaux. Soit $\mathbf{z}$ un vecteur non nul, élément de $[\Omega]$, et $\cos(\mathbf{z}, \mathbf{u}_{is})$ le cosinus du φ-angle

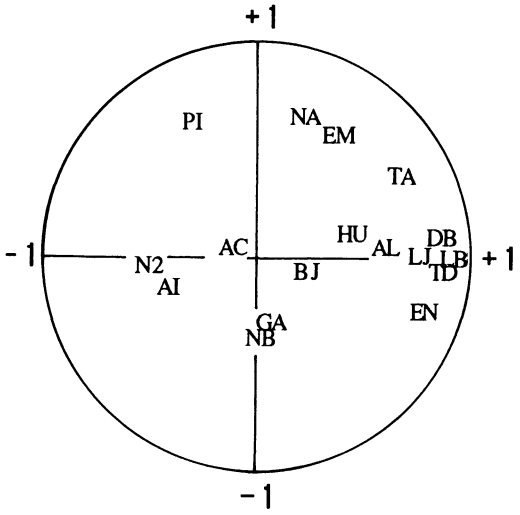

FIGURE 1
Corrélations des 17 caractères avec les deux premières composantes canoniques de l'ACGΣ

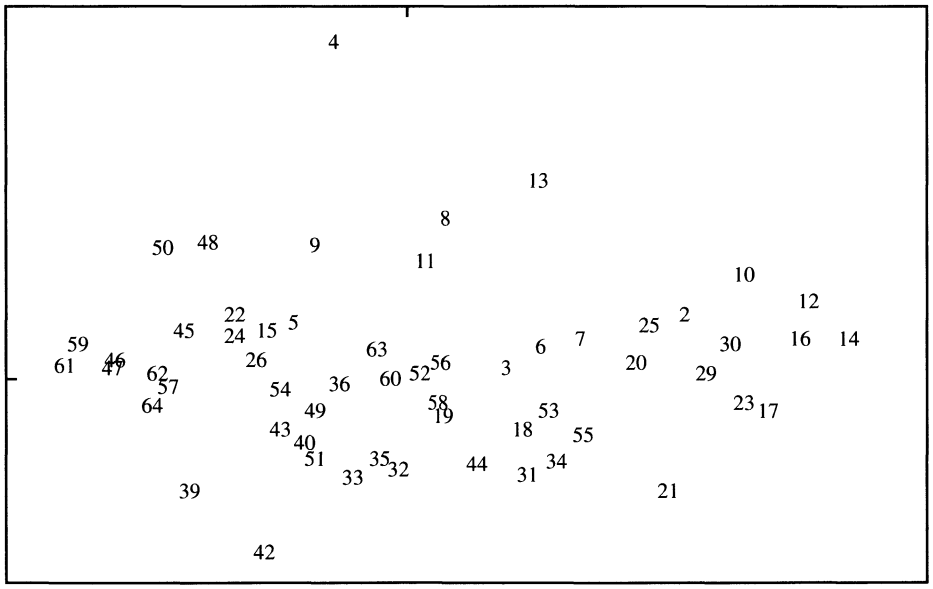

FIGURE 2
Représentation des 64 individus dans le plan factoriel défini par les deux premières composantes canoniques (ACGΣ) de $[X]$

de $\mathbf{z}$ et de $\mathbf{u}_{is}(s = 1, 2, ..., p_i)$. La somme $\sum_{s=1}^{k} \cos^2(\mathbf{z}, \mathbf{u}_{is})$ est toujours comprise entre 0 et 1, avec les propriétés extrêmales suivantes : elle est égale à 1 si et seulement si $\mathbf{z}$ appartient au sous-espace $[X_i]$, elle est égale à 0 si et seulement si $\mathbf{z}$ est φ-orthogonal à tout vecteur non nul de $[X_i]$. Considérons maintenant la représentation cartésienne suivante des vecteurs de $[\Omega]$: dans $\mathbb{R}^{p_i}$, rapporté à un système d'axes orthonormés, associons à $\mathbf{z} \in [\Omega]$ le point M_z dont les coordonnées sont les valeurs des $\cos(\mathbf{z}, \mathbf{u}_{is})$, $s = 1$ à p_i. Alors M_z est dans la sphère de rayon 1, dont le centre est l'origine O des axes de référence, avec les propriétés extrêmales suivantes : M_z est sur la surface de la sphère si et seulement si $\mathbf{z}$ appartient à $[X_i]$, M_z coïncide avec l'origine O si et seulement si $\mathbf{z}$ est φ-orthogonal à $[X_i]$. La projection orthogonale de M_z sur le plan défini par deux quelconques des axes de référence, est un point M'_z situé dans le cercle de centre O, de rayon 1 (cercle de cosinus). Compte tenu de ce qui précède, l'interprétation de la position de ce point M'_z, relativement au centre et à la circonférence du cercle, ne sera pas la même selon que le vecteur $\mathbf{z}$ appartient ou n'appartient pas à $[X_i]$. Dans les figures 3 et 4, nous avons ainsi projeté sur le plan associé aux deux premières composantes canoniques d'un certain groupe de variables X_i ($i = 1$ dans le cas de la figure 3, $i = 3$ dans le cas de la figure 4) la totalité des 17 variables actives. Pour faciliter la lecture, nous avons éclaté cette représentation en autant de diagrammes qu'il y a de groupes de variables, soit 4, et nous avons encadré par un carré ceux des diagrammes correspondant aux groupes autres que X_i, pour attirer l'attention sur une situation différente, appelant éventuellement des règles d'interprétation différentes.

Le cercle de corrélation de la figure 3 montre que la première composante canonique, dans $[X_1]$, peut être interprétée comme un facteur de taille longitudinale, puisque les quatre caractères morphologiques du premier groupe sont très positivement corrélés avec cette composante (les individus à silhouette longiligne auront tendance à avoir une forte valeur de cette composante). La deuxième composante, opposant la taille assis (TA) et la longueur des jambes (LJ), traduit la façon dont la longueur totale est répartie entre le haut et le bas du corps (les individus ayant des jambes courtes par rapport au tronc, auront tendance à avoir une forte valeur de cette composante).

Dans cette même figure 3, les trois autres cercles de corrélation situent les variables des trois autres groupes, par rapport à ces deux premières composantes canoniques du groupe 1. Ainsi, dans le deuxième groupe, l'envergure (EN) et la longueur des bras (LB) sont très proches de la composante longitudinale, ce qui est morphologiquement normal. Ces deux caractères ont été classés *a priori* par nous dans le deuxième groupe, en raison de leur utilisation «horizontale» sur le terrain (largeurs «fonctionnelles»). L'empan (EM) et le diamètre biacromial (DB), qui au contraire sont des largeurs «constitutives» (morphologiques), sont moins liées à cette composante longitudinale. Par ailleurs, aucune de ces quatre variables du deuxième groupe n'est significativement liée à la deuxième composante canonique, ce qui n'a rien d'étonnant. Tout ceci nous suggère que la relation qu'a le deuxième groupe de variables avec le premier est essentiellement une relation de taille, «plus un individu est grand en hauteur, plus il est grand en largeur».

Les cinq postes de jeu, constituant le troisième groupe de variables, sont essentiellement corrélés à la première composante canonique du premier groupe. Un gradient va des postes de pivot (PI) et d'ailier (AI) occupés par des sportifs de hauteur relativement faible, au poste d'arrière latéral (AL) occupé par des sportifs plus grands.

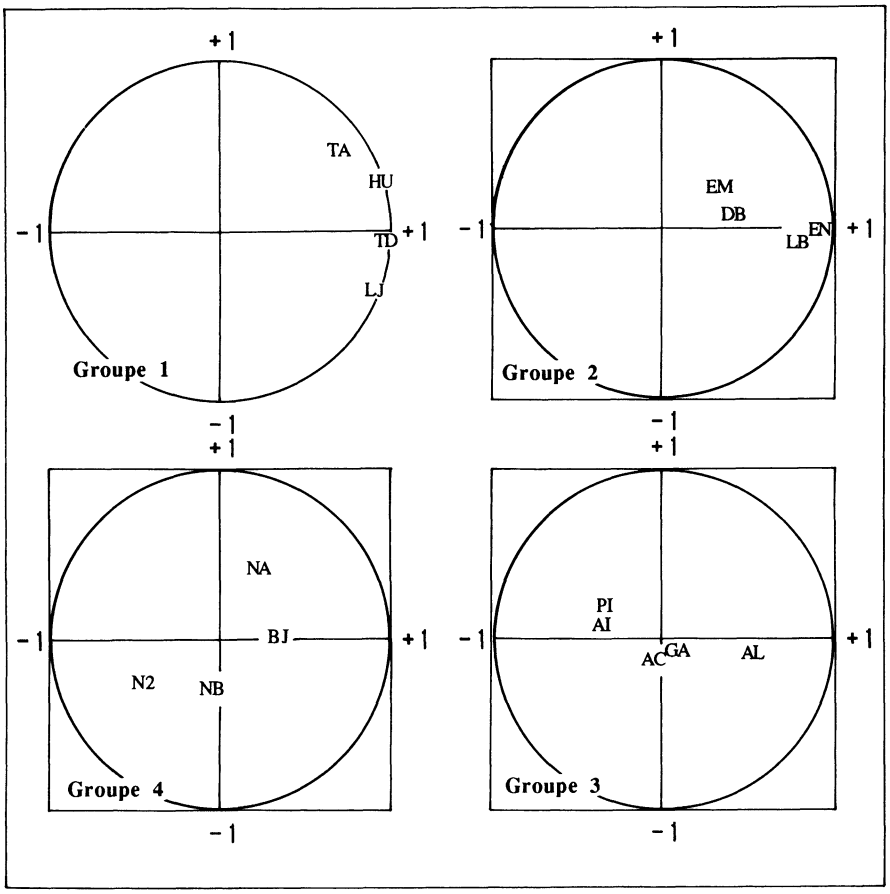

FIGURE 3
Représentation des corrélations des variables de chaque groupe avec les deux premières composantes canoniques de $[X_1]$.
Commentaire dans le texte

Enfin les quatre niveaux de jeu, constituant le quatrième groupe de variables, sont relativement peu corrélés tant avec la première composante canonique qu'avec la deuxième : la morphologie ne saurait à elle seule expliquer le niveau de jeu. On note cependant un gradient, orienté *grosso modo* selon la première bissectrice,

plaçant les niveaux les plus faibles en bas à gauche (tendance : taille moins élevée, jambes longues par rapport au tronc).

La figure 4 situe les 17 variables, groupe par groupe, par leurs corrélations avec les deux premières composantes canoniques du troisième groupe (celui-ci est constitué par les indicatrices des 5 postes de jeu). Le cercle des corrélations du groupe 3 fait ressortir, dans le sens de l'axe horizontal (première composante canonique du groupe), un gradient allant de l'arrière latéral (AL) à l'ailier (AI). La deuxième composante canonique (axe vertical) de ce troisième groupe dissocie les pivots (PI) des gardiens (GA), pratiquement confondus selon la première composante canonique. On va tenter d'interpréter la disposition particulière des cinq postes de jeu, dans ce cercle de corrélation, en examinant les cercles de corrélation correspondant aux trois autres groupes de variables.

Une disposition en gradient des cinq postes de jeu a déjà été remarquée plus haut, lors de l'étude de la première composante canonique du premier groupe de variables (voir fig. 3, groupe 3) : les projections des cinq postes de jeu sur le premier axe canonique du premier groupe sont, au sens près (sans signification, le signe des vecteurs propres étant arbitraire), disposées dans le même ordre que les projections de ces mêmes cinq postes de jeu sur le premier axe canonique du troisième groupe. Cette quasi homothétie des deux gradients suggère l'existence d'une assez forte corrélation (négative) entre ces deux composantes canoniques (le coefficient de corrélation entre les deux est égal à $-0{,}558$). Dans le cercle de corrélation correspondant au groupe 1 (fig. 4), les quatre variables constituant ce premier groupe sont très regroupées, proches du premier axe canonique, aux alentours de l'abscisse $-0{,}5$. Ceci confirme la remarque qui précède : le gradient des cinq postes de jeu selon la première composante canonique du troisième groupe semble pouvoir être expliqué par l'ensemble des quatre caractères morphologiques constituant le premier groupe.

Toujours dans la figure 4, l'examen du cercle de corrélation correspondant au groupe 2 suggère que l'on pourrait rechercher l'explication de la disjonction pivots / gardiens dans la dissociation des quatre caractères morphologiques du deuxième groupe, en deux sous-groupes : les largeurs constitutionnelles d'une part (empan et diamètre biacromial), les largeurs fonctionnelles d'autre part (envergure et longueur des bras). Ainsi les pivots auraient tendance à être plus larges d'épaules et à avoir de plus grandes mains que les gardiens ; ces derniers auraient en revanche une envergure plus importante, des bras plus longs.

Quant aux quatre niveaux de jeu, représentés dans le cercle de corrélation associé au groupe 4, ils restent très regroupés à proximité de l'origine, traduisant l'indépendance du niveau relativement au poste (quel que soit son niveau, une équipe comporte forcément les cinq postes de jeu !). Soulignons, à propos de cet exemple, la différence de règle d'interprétation de la position d'un point proche du centre du cercle. Dans le cas du groupe 3, la variable AC (par exemple), proche du centre, est simplement « mal représentée » dans le plan associé aux deux premières composantes canoniques de ce groupe 3 ; mais nous sommes assurés qu'elle sera mieux représentée dans d'autres plans, parce que dans $\mathbb{R}^5$, le point représentant la variable AC est sur la surface de la sphère. Au contraire, dans le cercle associé au groupe 4, les variables NA, N2, NB, BJ, toutes les quatre proches du centre, peuvent très bien rester proches du centre dans tous les plans de représentation : ceci

ne ferait que traduire la quasi φ-orthogonalité de chacune de ces quatre variables, par rapport au sous-espace $[X_i]$ (ce qui, dans le cas présent, est cohérent avec la nature des deux variables qualitatives poste et niveau de jeu : ces deux variables sont indépendantes par définition, car une équipe est constituée des mêmes postes, quel que soit son niveau ; ce sont les aléas des effectifs observés qui créent une petite déviation par rapport à la φ-orthogonalité).

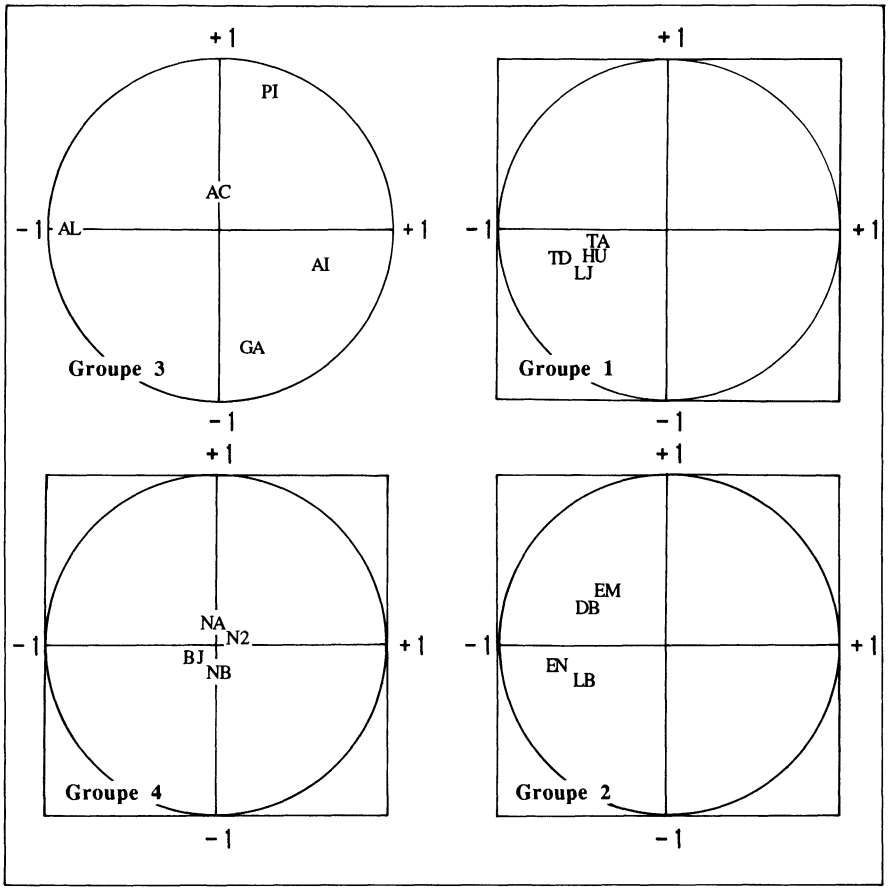

FIGURE 4
Représentation des corrélations des variables de chaque groupe avec les deux premières composantes canoniques de $[X_3]$.
Commentaire dans le texte

5. Conclusion

La généralisation de l'analyse canonique proposée par Carroll permet, en se plaçant dans $[X] = \sum_{i=1}^{k} [X_i]$, d'obtenir une vue globale de l'ensemble des $[X_i], i = 1$ à k. La généralisation que nous proposons permet, en se plaçant successivement dans chacun des sous-espaces $[X_i]$, d'avoir une vue globale de l'ensemble des autres sous-espaces. La première adopte la stratégie Σ, la seconde la stratégie Π, c'est pourquoi nous suggérons de les désigner respectivement par ACGΣ et ACGΠ. Ces deux généralisations de l'analyse canonique ont des objectifs différents, et d'une certaine façon se complètent : reprenant la terminologie d'Escoufier [5], nous dirons :

– que l'ACGΣ, en projetant toutes les variables sur un sous-espace (optimal, tenant compte de la répartition des variables en groupes), construit un *compromis* ;

– que l'ACGΠ, en analysant chaque groupe de variables au travers de l'image que donne de lui l'ensemble des autres variables, est un instrument d'*analyse fine*.

L'exemple numérique que nous avons (partiellement) traité, permet de dégager quelques grandes lignes de l'utilisation pratique de l'ACGΠ. Par cette méthode, chaque groupe de variables est factorisé en fonction des relations qu'il a avec l'ensemble des variables des autres groupes. Les facteurs (les composantes canoniques du groupe) étant destinés à servir de référence, par exemple dans des représentations graphiques des individus (cartes factorielles), on cherche à les interpréter grâce aux représentations graphiques des corrélations entre toutes les variables d'une part, et les composantes canoniques de ce groupe particulier de variables d'autre part.

Par exemple, pour k groupes de variables, on peut associer à chacun des groupes :

– le cercle de corrélation, où sont représentées les variables de ce groupe, relativement à deux composantes canonique du groupe ;

– les $k - 1$ cercles de corrélation, dans chacun desquels sont représentées les variables de l'un des autres groupes, relativement aux deux mêmes composantes canoniques que précédemment.

L'examen simultané de cet ensemble de k représentations associées à un même groupe de variables, peut permettre à la fois l'interprétation des composantes canoniques de ce groupe, et l'interprétation de la disposition des points-variables par rapport aux axes canoniques.

RÉFÉRENCES

[1] BENZECRI J. P. (1973). *L'analyse des données : 2. L'analyse des correspondances*. Paris, Dunod éd., 619 p.

[2] CAILLIEZ F. et PAGES J.-P. (1976). *Introduction à l'analyse des données*. Paris, S.M.A.S.H. éd., 616 p.

[3] CARROLL J. D. (1968). A generalization of canonical correlation analysis to three or more sets of variables. *Proceedings of the 76th annual Convention of the American Psychological Association*, p. 227-228

[4] CASIN Ph. et TURLOT J.-C. (1986). Une présentation de l'analyse canonique généralisée dans l'espace des individus. *Revue de Statistique Appliquée*, **XXXV**, 3, 65-75

[5] ESCOUFIER Y. (1985). Objectifs et procédures de l'analyse conjointe de plusieurs tableaux de données. *Statistique et Analyse des Données*, **10**, 1, 1-10

[6] HOTELLING H. (1935). The most predictable criterion. *The Journal of Educational Psychology*, **XXVI**, 139-142

[7] HOTELLING H. (1936). Relations between two sets of variables. *Biometrika*, **28**, 321-377

[8] KETTENRING J. R. (1971). Canonical analysis of several sets of variables. *Biometrika*, **58**, 3, 433-451

[9] PAGES J.-P., CAILLIEZ F., ESCOUFIER Y. (1979). Analyse factorielle : un peu d'histoire et de géométrie. *Revue de Statistique Appliquée*, **XXVII**, 1, 5-28

[10] PERNIN M.-O. (1986). *Contribution à la méthodologie d'analyse de données longitudinales. Exemple de la croissance chez l'être humain (Auxologie)*. Lyon, Thèse Dipl. Doct. Université Claude Bernard, 264 p.

[11] PONTIER J., JOLICŒUR P., PERNIN M.-O. (1987). Analyse canonique complète. *Statistique et Analyse des Données*, **12**, 1-2, 124-148

[12] PONTIER J., PERNIN M.-O. (1987). 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., p. 49-65

[13] PONTIER J., PERNIN M.-O. (1989). Relations entre analyse canonique complète et méthode LONGI. *Revue de Statistique Appliquée*, **XXXVII**, 4, 67-82

[14] PONTIER J., DUFOUR A.-B., NORMAND M. (1990). *Le modèle euclidien en analyse des données*. Bruxelles, Editions de l'Université de Bruxelles, 428 p.

[15] SAPORTA G. (1975). *Liaisons entre plusieurs ensembles de variables et codage de données qualitatives*. Paris, Thèse Doct. 3ème Cycle, Université Pierre et Marie Curie

[16] SAPORTA G. (1990). *Probabilités, Analyse des Données et Statistique*. Paris, Editions Technip, 493 p.

[17] TENENHAUS M. (1977). Analyse en composantes principales d'un ensemble de variables nominales ou numériques. *Revue de Statistique Appliquée*, **XXV**, 2, 39-56

[18] TENENHAUS M. (1986). Generalized canonical analysis, canonical analysis and applications. *Communication à l'Ecole de Printemps d'Analyse des Données, Chania, Crête, 31 mars - 4 avril 1986.*