REVUE DE STATISTIQUE APPLIQUÉE

X. BRY

Une autre approche de l'analyse factorielle : l'analyse en résultantes covariantes

Revue de statistique appliquée, tome 49, n° 3 (2001), p. 5-38

<http://www.numdam.org/item?id=RSA_2001__49_3_5_0>

Rev. Statistique Appliquée, 2001, **XLIX** (3), 5-38

UNE AUTRE APPROCHE DE L'ANALYSE FACTORIELLE : L'ANALYSE EN RÉSULTANTES COVARIANTES

X. Bry

Professeur au Département de Statistique de l'E.N.E.A., BP 5084, Dakar Fann.
Email : dsd-enea@refer.sn

RÉSUMÉ

Nous proposons ici une mesure de liaison entre une variable F et un groupe de variables X : la résultante de F sur X, qui est à la fois quantitative et directionnelle, comme la régression linéaire de F sur X. Mais à la différence de la régression, la résultante considère les corrélations internes à X comme une information structurelle dont on peut tenir compte à un degré variable.

Nous montrons que les problèmes qui fondent les méthodes factorielles classiques peuvent être reformulés à partir de résultantes. On le fait d'abord pour l'Analyse en Composantes Principales et l'Analyse Canonique Généralisée. Puis, on reprend diverses méthodes d'exploration des liaisons entre deux groupes de variables (régression PLS, Analyse des Redondances Maximales, Analyse Canonique). La réécriture de leur formulaire en termes de résultantes fournit à certaines de ces techniques une voie simple et directe d'extension, en particulier au traitement des variables qualitatives. L'inconvénient général de ces méthodes reste toutefois de ne pas systématiquement fournir des facteurs décorrélés dans chacun des groupes.

Reprenant la problématique de ces méthodes, on leur construit enfin une alternative générale fournissant toujours des facteurs linéairement non redondants : l'Analyse en Résultantes Covariantes (ARC). Deux exemples d'application permettant la comparaison pratique de la régression PLS et de l'ARC sont donnés en fin d'article.

Mots-clés : *Analyse factorielle, Analyse en Composantes principales, Régression PLS, Analyse canonique, Analyse des redondances maximales.*

ABSTRACT

In this work, we propose a measure of association between a variable F and a group of variables X. This measure, called resultant of F on X, is both quantitative and directional, as is multiple regression of F on X. But unlike regression, the resultant considers internal correlations of X as relevant structural information that can be taken into account to a chosen degree.

We show that questions grounding classical factoring methods can be reformulated using resultants. We first do it for Principal Components Analysis and Generalized Canonical Analysis. We then consider various methods investigating correlations between two groups of variables (PLS regression, Maximal Redundancy Analysis, Canonical Analysis). Rewriting their formulas using resultants allows these methods a simple and direct extension, particularly to categorical data processing. These techniques still have a general drawback : they do not systematically yield uncorrelated components in both groups.

We finally build up a general alternate method that always gives such uncorrelated components : Covarying Resultants Analysis (CRA). Two application examples are given, that allow practical comparison between PLS regression and CRA.

Keywords : *Factor Analysis, Principal Component Analysis, PLS regression, Canonical Analysis, Maximal Redundancy Analysis.*

Notations

I_n désigne couramment la matrice identité de taille n.

Les minuscules u, v, w désignent des vecteurs de coefficients.

X, Y, Z désignent des tableaux matriciels décrivant I individus (en ligne) à l'aide de variables (en colonne).

Ces lettres sont indifféremment utilisées pour désigner les groupes de variables en question.

Les minuscules x, y, z désignent indifféremment les vecteurs-colonnes de ces matrices et les variables des groupes correspondants.

Les minuscules grecques désignent des scalaires.

Le sous-espace vectoriel de $\mathbf{R}^I$ engendré par un groupe de variable X sera noté $< X >$.

M et N sont des matrices symétriques définies positives de tailles $(J \times J)$ et $(K \times K)$ pondérant les variables des groupes respectifs Y (variables y^j, $j = 1$ à J) et Z (variables z^k, $k = 1$ à K).

Lorsqu'un groupe de variables X est décomposé en L sous-groupes, ceux-ci seront notés $X_1, \ldots X_L$.

F et G désignent des facteurs construits respectivement par combinaison linéaire des variables de Y et des variables de Z.

Le projecteur orthogonal sur un sous-espace E sera noté Π_E.

La matrice diagonale ayant pour éléments diagonaux $a, b...$ sera notée $diag(a, b...)$.

La matrice bloc-diagonale ayant pour blocs diagonaux les sous-matrices $A, B...$ sera notée $diag(A, B...)$.

Le produit scalaire entre deux vecteurs x et y sera noté $\langle x|y \rangle$.

L'inertie d'un nuage $\{x^j\}_{j=1 \text{ à } J}$ projeté sur un axe $< F >$ sera notée

$Inertie^{<F>}(\{x^j\}_{j=1 \text{ à } J})$.

Le signe $\propto$ dénotera la proportionnalité de deux vecteurs, ex : $x \propto y$.

1. Préliminaire : résultante d'un facteur sur un groupe de variables

Les méthodes classiques d'Analyse Factorielle sont fondées sur une mesure de dispersion particulière : la variance. La liaison entre variables est alors naturellement

mesurée par leur covariance. On étend classiquement cette mesure à la liaison d'une variable F à un groupe X de deux façons : la part de variance de F explicable par une combinaison linéaire de X (R^2 de la régression multiple de F sur X), et la variance totale de X expliquée par F (somme des carrés des corrélations de F avec les variables de X). Ces indicateurs sont scalaires. On remarque cependant que dans la régression de F sur X, outre le R^2 mesurant l'intensité de la liaison, la partie expliquée de la variable donne la direction de liaison dans le groupe. La régression n'envisage le groupe X qu'en tant que sous-espace de prédicteurs, et ne considère pas les corrélations internes à X comme une information structurelle intéressante. Alternativement, nous proposons ici une mesure vectorielle plus générale de la liaison entre une variable F et un groupe X : la résultante de F sur X. Il s'agit d'une variable indiquant dans quelle direction et avec quelle intensité la dispersion selon le groupe X concorde globalement avec la dispersion selon F. Contrairement à la régression, toutefois, la résultante peut tenir compte, à divers degrés, des corrélations internes au groupe.

1.1. Résultante simple

• On considère un groupe de variables numériques $X = (x^1, ..., x^J)$ et une variable F, toutes centrées réduites, décrivant I individus. Pour simplifier les écritures, on munit ces individus de poids unitaires. Les variables sont donc traduites par des vecteurs de $\mathbf{R}^I$, cet espace étant muni de la métrique diagonale des poids individuels $P = I_I$ (matrice identité de taille I) [1]. Les vecteurs $x^1, ..., x^J$ et F sont donc normés.

• On définit la ***résultante simple*** de F sur le groupe X (*cf.* fig. 1), notée $R_X(F)$, comme la somme des projections orthogonales de F sur les x^j : $R_X(F) = \sum_{j=1}^{J} \Pi_j F$

avec $\Pi_j = x^j x^{j'}$

Soit, matriciellement : $R_X F = XX'F$

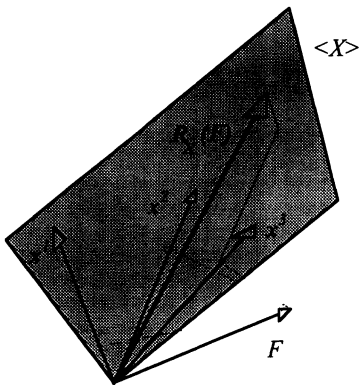


FIGURE 1

[1] Dans le cas d'une matrice P de poids quelconques, on peut ramener l'espace à la métrique identité au prix de la transformation $x \mapsto x^* = P^{1/2}x$ des vecteurs qui s'y trouvent.

1.2. Interprétation

Nous montrons ici que *la variance de la résultante simple contient l'inertie des x^j dans la direction $< F >$, ainsi que des termes traduisant le fait que les x^j varient ensemble comme F.*

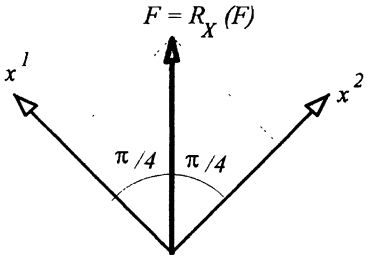
Calculons la variance de la résultante $R_X(F)$:

$$\|R_X(F)\|^2 = \left\| \sum_j \cos(F, x^j) x^j \right\|^2$$

$$= \sum_j \cos^2(F, x^j) + \sum_{j \neq m} \cos(F, x^j)\cos(F, x^m)\cos(x^m, x^j)$$

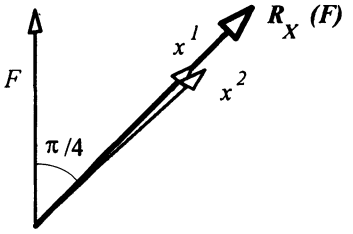
Le terme $\sum_j \cos^2(F, x^j) = \sum_j \langle x^j \mid F \rangle^2$ correspond à l'inertie de la projection du nuage des x^j sur $< F >$. On note d'autre part qu'on peut toujours, en changeant éventuellement x^j en son opposé, faire en sorte que les $\cos(F, x^j)$ soient positifs. Dès lors, le signe de chaque produit $\cos(F, x^j)\cos(F, x^m)\cos(x^m, x^j)$ est celui de $\cos(x^m, x^j)$. La valeur absolue de ce terme sera d'autant plus élevée que x^j et x^m sont corrélées à F, mais aussi que x^j et x^m sont liées entre elles en valeur absolue.

Pour le problème de mesure globale des liaisons des variables dans la direction $< F >$, la variance de la résultante de F peut représenter une alternative à la variance des x^j expliquée par F (leur inertie dans la direction de F). La différence de comportement entre ces deux mesures de liaison apparaît clairement à l'examen des deux situations présentées dans la figure 2, dans lesquelles les x^j ont la même inertie sur $< F >$.



situation 1 :

les x^j forment un angle de 45° avec F, mais sont décorrélées.



situation 2 :

les x^j forment un angle de 45° avec F, mais sont parfaitement corrélées. La variance de la résultante double.

FIGURE 2

En somme, la résultante fournit à la fois une information quantitative et directionnelle sur la liaison variable-groupe, alors que l'inertie ne donne que l'élément quantitatif. De plus, quand l'inertie ne fait que mesurer la corrélation des variables

avec F, la variance de la résultante accorde une «prime» au fait que les variables soient corrélées entre elles.

1.3. Résultante généralisée

• On considère à présent des variables quelconques $x^1, \ldots x^J$ (non nécessairement réduites), munies de poids $m_1, \ldots m_J$. En notant $M = diag(m_j)$, on définit la résultante généralisée de F sur le groupe X comme : $R_X^M F = XMX'F$. Plus généralement, on pourra utiliser dans l'écriture précédente une matrice M définie positive, ce qui revient à pondérer des combinaisons linéaires particulières des variables.

• L'opérateur de résultante est symétrique [2] .

• En prenant $M = (X'X)^{-1}$, on retrouve à titre de cas particulier la régression multiple de F sur X.

• Considérons le cas où le groupe X est structuré en sous-groupes X_j. On suppose les $X_j'X_j$ inversibles.

Si l'on désire tenir compte des liaisons entre sous-groupes, mais pas des liaisons à l'intérieur d'un sous-groupe, on prendra pour matrice de pondération la matrice bloc-diagonale suivante :

$$M = \begin{bmatrix} (X_1'X_1)^{-1} & \ldots & 0 \\ \vdots & & \vdots \\ 0 & \ldots & (X_J'X_J)^{-1} \end{bmatrix}$$

La résultante généralisée est alors : $R_X^M(F) = \sum_{j=1}^{J} X_j(X_j'X_j)^{-1}X_j'F = \sum_{j=1}^{J} \Pi_{<X_j>}F$. Nous qualifierons cette résultante de *multiple*. Elle est particulièrement intéressante dans le cas de variables qualitatives. En effet, une telle variable Y est représentée par l'ensemble des indicatrices de ses H modalités. D'après l'analyse de variance à un facteur, la partie expliquée par Y d'une variable quantitative F est sa projection sur le sous-espace $< Y >$ de dimension H engendré par les indicatrices. Cette projection est la somme de deux composantes :

– la première est la projection de F sur le vecteur e dont toutes les composantes valent 1, et est proportionnelle à la moyenne de F;

– la seconde est la projection de F sur l'orthogonal $e^{\perp}$ de e dans $< Y >$. Cet orthogonal est le sous-espace de dimension $H - 1$ engendré par les H indicatrices projetées sur $e^{\perp}$, c'est-à-dire centrées.

Seule la seconde composante traduit la liaison entre les variables F et Y. En particulier, le carré de la norme de cette seconde composante est égal à la variance de F entre les classes de Y.

---

[2] XMX' est la matrice des M-produits scalaires entre lignes de X correspondant à l'opérateur d'Escoufier.

Il est donc naturel, en présence de variables qualitatives dans un groupe X, de représenter chacune d'elles par le sous-espace engendré par ses indicatrices centrées. On calculera alors la résultante multiple de F sur X en considérant comme sous-groupe l'ensemble des indicatrices centrées de chaque facteur [3] .

• On peut également vouloir tenir compte des liaisons internes à chaque sous-groupe, mais en harmonisant la contribution des différents sous-groupes. Cette démarche est au fondement de l'Analyse Factorielle Multiple (AFM, *cf.* [5]). On cherchera alors à pondérer chaque sous-groupe X_j par un scalaire ϖ_j adéquat. La matrice M s'écrira donc $M = diag(diag(\varpi_j))_j$:

$$M = \begin{bmatrix} \begin{bmatrix} \varpi_1 & & \\ & \ldots & \\ & & \varpi_1 \end{bmatrix} & \ldots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \ldots & \begin{bmatrix} \varpi_J & & \\ & \ldots & \\ & & \varpi_J \end{bmatrix} \end{bmatrix}$$

• Changement de base :

Supposons que l'on rapporte l'espace $\mathbf{R}^J$, muni de la métrique M, à une quelconque base $\{u_\alpha\}_{\alpha=1 \text{ à } J}$ M-orthonormée. Les vecteurs définis par $u^*_\alpha = M^{1/2}u_\alpha$ forment un système I_J-orthonormé. Les coordonnées des individus dans la base des u_α sont données par : $h_\alpha = XMu_\alpha$. Soient U, U^* et H les matrices dont les colonnes sont respectivement les u_α, u^*_α et les h_α. On a : $H = XMU$. L'opérateur de résultante simple sur H est : $R_H = HH' = XMUU'\,MX' = XM^{1/2}U^*U'^*M^{1/2}X'$. Or : $U^*U'^* = I_J$; par conséquent : $R_H = HH' = XMX' = R^M_X$.

Cette invariance de l'opérateur de résultante est particulièrement utile lorsqu'on ne dispose pas des variables originelles mais seulement des facteurs qui les représentent (issus d'une procédure de multidimensional scaling, par exemple).

2. L'Analyse Factorielle d'un groupe de variables à partir des résultantes

2.1. Maximisation de la variance d'une résultante généralisée

Soient un groupe de variables numériques X, et une matrice M de pondération de celles-ci. On recherche une base factorielle permettant une représentation synthétique des corrélations internes à X. Soit une variable F normée. On résout alors le programme : $\mathbf{P}^*$: $\underset{\|F\|=1}{Max}\ \text{var}(R^M_X F)$

$$\text{var}(R^M_X F) = \langle R^M_X F \mid R^M_X F \rangle = F'(XMX')(XMX')F$$

[3] On montre en fait aisément qu'il est indifférent de centrer ou non ces indicatrices.

Les F solutions sont donc les vecteurs propres de $(XMX')^2$, qui sont aussi ceux de XMX', associés aux carrés des valeurs propres λ de XMX'. Ces dernières étant positives, le classement des vecteurs propres par ordre décroissant de valeur propre coïncide pour les deux matrices. Or, la diagonalisation de XMX' répond au problème de maximisation suivant :

$$\underset{\|F\|=1}{Max}\ F'(XMX')F \Longleftrightarrow \underset{\|F\|=1}{Max}\ \textit{Inertie}^{<F>}(\{x^j\}_{j=1\ \grave{a}\ J})$$

Et dans ce cas, chaque valeur propre de XMX' est l'inertie des x^j dans la direction du vecteur propre correspondant. On reconnaît là l'ACP du tableau X avec la métrique M.

On remarquera que les extréma F sont colinéaires à leur résultante : $R_X(F) = XMX'F = \lambda F$. En particulier, ce sont des combinaisons linéaires des variables x^j.

• L'ACP de X avec la métrique M procède donc indifféremment des programmes :

$$\mathbf{P}^* : \underset{\|F\|=1}{Max} \langle R_X^M F \mid R_X^M F\rangle \qquad \text{et} \qquad \mathbf{Q}^* : \underset{u'M\,u=1}{Max} \langle XMu \mid XMu\rangle$$

2.2. Cas particuliers

• Le groupe X étant structuré en sous-groupes X_j, l'usage de la résultante multiple fournit immédiatement l'Analyse Canonique Généralisée de X (ACG).

• X étant toujours structuré en sous-groupes X_j, et si l'on désigne par μ_j la plus grande valeur propre d'ACP de X_j, l'usage de la métrique $M = \textit{diag}\left(\textit{diag}\left(\frac{1}{\mu_j}\right)\right)_j$ conduit à l'AFM de X.

3. Analyses factorielles classiques des liaisons entre deux groupes de variables

3.1. Cadre général

3.1.1. Le problème

On considère deux groupes Y et Z à respectivement J et K variables, et pondérés respectivement par M et N, matrices définies positives. On voudrait photographier la structure de chaque groupe dans une base factorielle de son espace, de sorte à faire apparaître les liaisons entre les deux groupes. Les trois techniques classiques que nous allons examiner dans ce cadre (PLS, Analyse des Redondances Maximales, Analyse Canonique) peuvent être dérivées d'un même programme

étendant $\mathbf{Q}^*$, qui consiste à maximiser sous contraintes la covariance des facteurs des deux groupes, soit [4] :

$$\boxed{\mathbf{Q} : \underset{\substack{u'Mu=1 \\ v'Nv=1}}{Max} \langle YMu \mid ZNv \rangle}$$

3.1.2 Résolution du programme $\mathbf{Q}$

Le lagrangien de $\mathbf{Q}$ est :

$$L = u'MY'ZNv - \lambda(u'Mu - 1) - \mu(v'Nv - 1)$$

$$\frac{\partial L}{\partial u} = 0 \Leftrightarrow MY'ZNv = 2\lambda Mu \ (1) \ ; \ \frac{\partial L}{\partial v} = 0 \Leftrightarrow NZ'YMu = 2\mu Nv \ (1')$$

(1) et (1') $\Rightarrow$ (2) $NZ'YMY'ZNv = \eta Nv$

et (2') $MY'ZNZ'YMu = \eta Mu$ avec $\eta = 4\lambda\mu$

On a par ailleurs :

$$u'(1) \Leftrightarrow u'MY'ZNv = 2\lambda \ ; \ v'(1') \Leftrightarrow v'NZ'YMy = 2\mu$$

Ce qui implique $\lambda = \mu = u'MY'ZNv$ à maximiser. Donc $\eta = 4\lambda^2$ est la plus grande valeur propre de $MY'ZNZ'YM$ et $NZ'YMY'ZN$.

Enfin, en notant $F = YMu$ et $G = ZNv$, on obtient :

$$Y(1) \Leftrightarrow R_Y^M G = \sqrt{\eta} F \ (3) \ ; \ Z(1') \Leftrightarrow R_N^N F = \sqrt{\eta} G \ (3')$$

et par suite :

$$(4) \ R_Z^N R_Y^M G = \eta G \ ; \ (4') \ R_Y^M R_Z^N F = \eta F$$

Les opérateurs $R_Z^N R_Y^M$ et $R_Y^M R_Z^N$ seront appelés *opérateurs de résultante croisée.*

• *Lorsque Z est réduit à une variable z*

Le programme $\mathbf{Q1}$: $Max_{u'Mu=1} \langle YMu|z \rangle$ a pour solution : $u \propto Y'z$, ce qui donne :

$$F = YMu \propto YMY'z = R_Y^M z$$

[4] Une autre formulation de ce programme est donnée dans [8].

3.2. Cas particuliers

Nous allons à présent considérer le cas de deux groupes de variables quantitatives. Dans un deuxième temps, nous verrons comment traiter des données qualitatives ou mixtes.

3.2.1. La régression PLS

• *La méthode*

Y et Z sont ici des groupes de variables quantitatives. La première étape de la régression PLS consiste à trouver les facteurs Yu et Zv de covariance maximale sous la double contrainte $\|u\| = 1$, $u \in \mathbf{R}^K$ et $\|v\| = 1$, $v \in \mathbf{R}^L$, avec l'argumentation heuristique suivante :

$$\mathrm{cov}(Yu, Zv) = \sqrt{\mathrm{var}(Yu)}.\sqrt{\mathrm{var}(Zv)}.\rho(Yu, Zv)$$

avec : $\mathrm{var}(Yu) = \|Yu\|^2 = u'Y'Yu =$ inertie des individus le long de $< u >$ dans $\mathbf{R}^J$;

$\mathrm{var}(Zv) = \|ZV\|^2 = v'Z'Zv =$ inertie des individus le long de $< v >$ dans $\mathbf{R}^K$;

$\rho(Yu, Zv)$ mesurant la corrélation entre les facteurs des deux bases.

Dans le programme **Q**, on pose donc ici $M = I_J$, $N = I_K$. Les équations (2) et (2') donnant les vecteurs u et v deviennent alors :

$$Z'YY'Zv = \eta v \; ; \; Y'ZZ'Yu = \eta u$$

Et les équations (4) et (4') deviennent : $R_Z R_Y G = \eta G$; $R_Y R_Z F = \eta F$

Les facteurs F (respectivement G), vecteurs propres de matrices non symétriques, ne sont pas a priori décorrélés deux à deux.

La seconde étape de PLS, recherchant des facteurs explicatifs non linéairement redondants, doit donc projeter le problème sur l'orthogonal du premier facteur $F_1 = Yu_1$. On réitère alors le procédé à partir des résidus de régression des variables de Y et Z sur ce facteur.

• *Caractéristiques de la méthode*

Le programme de la première étape de PLS est conçu pour mettre en rapport deux structures de variables qu'elle traite symétriquement (il n'est donc pas encore nécessairement question d'expliquer l'une à partir de l'autre). La résolution de ce programme souffre à ce stade de quelques points faibles :

– Elle ne fournit pas d'emblée des facteurs décorrélés au sein de chaque groupe. L'un des deux groupes (Y) étant pris comme groupe explicatif, PLS doit, pour obtenir une suite de facteurs explicatifs non redondants, résoudre à chaque étape un programme nouveau.

– Ce faisant, PLS ne fournit toujours pas de facteurs du groupe expliqué orthogonaux deux à deux. Ceci pose un problème de représentation en base factorielle des variables de ce groupe. En outre, cette non-orthogonalité des facteurs G représente une perte d'efficacité dans la synthèse des variables de Z, perte d'autant plus grande que les G sont linéairement redondants.

– PLS ne permet donc pas une exploration *symétrique* de la communauté de structure des groupes Y et Z. S'il n'y a pas d'orientation causale claire entre les deux groupes, la méthode paraît peu adaptée à la recherche de liens entre eux.

Si cette orientation causale existe, PLS possède en revanche des atouts :

– La prise en compte de la structure de corrélations internes aux groupes dans la construction des facteurs, ce qui en augmente la robustesse et facilite leur interprétation.

– La possibilité de traiter des groupes de variables sujettes à multicolinéarité.

– La possibilité, en calculant un nombre suffisant de facteurs explicatifs, de retrouver l'entier pouvoir explicatif des régressions ordinaires de Z sur Y.

3.2.2. Analyse des Redondances Maximales (ARM) ou ACP sur Variables Instrumentales

• *La méthode :*

Y et Z sont ici encore des groupes quantitatifs. On cherche à réaliser une ACP du groupe Z orientée vers l'explication par Y, sans tenir compte de l'intra-structure de Y. L'inertie des individus dans l'espace des z^l doit ici être prise en compte, mais pas leur inertie dans l'espace des y^k.

On cherche ainsi à résoudre le programme : $\underset{\substack{w'Y'Yw=1 \\ t't=1}}{Max} \langle Yw \mid Zt \rangle$.

Pour retrouver la formulation du programme **Q**, on pose donc :

$$w = Mu; t = Nv; M = (Y'Y)^{-1}; N = I_K$$

Les équations (4) et (4′) deviennent alors :

$$\Pi_{<Y>} ZZ'F = \eta F \; ; \; ZZ'\Pi_{<Y>} G = \eta G \quad \text{avec} \quad \Pi_{<Y>} = Y(Y'Y)^{-1}Y'$$

On note que l'appartenance de F au sous-espace $< Y >$ entraîne $\Pi_{<Y>} F = F$, de sorte que la première équation s'écrit aussi : $\Pi_{<Y>} ZZ'\Pi_{<Y>} F = \eta F$. Les facteurs F sont ainsi vecteurs propres d'une matrice symétrique, et par conséquent orthogonaux deux à deux. Ce n'est a priori pas le cas des facteurs G.

Une fois obtenus les facteurs F, on les utilisera comme facteurs explicatifs en régressant sur eux les variables de Z.

• *Caractéristiques de la méthode*

– L'ARM ne traite pas symétriquement les deux groupes. L'orientation explicative doit donc être claire au départ : Y est explicatif, et Z à expliquer.

– Contrairement à PLS, l'ARM fournit d'emblée des facteurs F décorrélés. Cette non-redondance est importante dans la mesure où ces facteurs sont explicatifs. Malheureusement, l'intra-structure de Y n'ayant pas été prise en compte, ces facteurs peuvent éventuellement comporter une part importante de bruit. Leur interprétation en est fragilisée.

– L'ARM ne peut *directement* traiter le cas d'un groupe explicatif Y comportant des multicolinéarités ($Y'Y$ n'est dans ce cas pas inversible).

– Les facteurs expliqués G sont linéairement redondants.

– Comme pour PLS, en calculant un nombre suffisant de facteurs explicatifs, on retrouve l'entier pouvoir explicatif des régressions ordinaires de Z sur Y.

3.2.3. Analyse Canonique (AC)

Cette fois, on ne tient compte de l'intra-structure d'aucun des deux groupes. On cherche donc à résoudre :

$$\underset{\|Yw\|=1;\|Zt\|=1}{Max} \langle Yw \mid Zt \rangle.$$

On retrouve le programme **Q** en posant : $w = Mu$; $t = Nv$; $M = (Y'Y)^{-1}$; $N = (Z'Z)^{-1}$.

Les équations (4) et (4′) deviennent alors :

$$\Pi_{<Y>}\Pi_{<Z>}F = \eta F \; ; \; \Pi_{<Z>}\Pi_{<Y>}G = \eta G$$

Comme : $F \in< Y >\Rightarrow F = \Pi_{<Y>}F$ (et de même $G \in< Z >\Rightarrow G = \Pi_{<Z>}G$), ces équations équivalent respectivement à :

$$\Pi_{<Y>}\Pi_{<Z>}\Pi_{<Y>}F = \eta F \; ; \; \Pi_{<Z>}\Pi_{<Y>}\Pi_{<Z>}G = \eta G$$

Les facteurs F (resp. G) sont ainsi vecteurs propres de matrices symétriques, et donc deux à deux orthogonaux.

• *Caractéristiques de la méthode*

– L'AC traite symétriquement les deux groupes. Elle ne nécessite donc pas d'orientation explicative particulière, et a plutôt une vocation exploratoire.

– L'AC fournit d'emblée des facteurs F (respectivement G) décorrélés. Mais les intra-structures de Y et Z n'ayant pas été prises en compte, ces facteurs peuvent ne pas représenter des faits structurels forts de leurs groupes respectifs.

– L'AC ne peut *directement* traiter le cas de groupes Y et Z comportant des multicolinéarités.

3.2.4. Méthodes d'analyse de groupes qualitatifs ou mixtes

Les équations (3) et (3′) permettent d'étendre simplement la méthodologie aux variables qualitatives. Si l'un ou l'autre des deux groupes contient des variables qualitatives, chacune y est représentée par les indicatrices de ses modalités et est considérée comme un sous-groupe. On utilise alors la résultante multiple du facteur sur le groupe dans les équations (3) et (3′).

Sur les plans factoriels directs, une modalité pourra être représentée par un point ayant pour coordonnées factorielles les moyennes des facteurs sur l'ensemble des individus possédant la modalité.

• Dans le cas de deux groupes qualitatifs, il est intéressant de noter (*cf.* [3]) que cette extension correspond à une technique classique : l'Analyse des Correspondances du tableau de contingence croisant les deux groupes.

• De la même manière, si le groupe Y est qualitatif et Z quantitatif, la méthode conduit à l'ACP du tableau des moyennes partielles des z^k par modalités des y^j., i.e. à l'analyse inter-classes de Z. En particulier, si l'on prend $N = (Z'Z)^{-1}$, on obtient l'Analyse Factorielle Discriminante.

3.3. Conclusion

Les méthodes classiques présentées ici ont une unité formelle plaisante. Elles représentent des variantes d'une même méthode de base, qui s'avère fondée sur la diagonalisation des deux opérateurs de résultante croisée $R_Y R_Z$ et $R_Z R_Y$. La souplesse de cette méthode permet de faire face à de nombreuses situations. Cependant, dans plusieurs de ses variantes, la méthode ne fournit pas spontanément des facteurs deux à deux décorrélés. Nous avons donc cherché à construire une méthode alternative remplissant les conditions suivantes :

– Synthétiser chaque groupe en une base factorielle propre, de sorte à faire également apparaître les liaisons entre les groupes.

– Etre aussi souple que la méthode générale exposée ci-dessus, c'est à dire permettre de régler à volonté l'importance relative des intra-structures et inter-structures.

– Fournir systématiquement des facteurs linéairement non redondants.

4. Une alternative générale : l'Analyse en Résultantes Covariantes

4.1. Principe de la méthode

On a vu que l'ACP d'un groupe X de variables pondérées par M procède indifféremment de deux programmes : $\mathbf{P}^*$ et $\mathbf{Q}^*$. Soient deux groupes Y et Z respectivement pondérés par M et N. On a vu précédemment comment le programme $\mathbf{Q}$, généralisant $\mathbf{Q}^*$, étendait l'ACP aux deux groupes. De même, le programme $\mathbf{P}^*$

suggère l'extension **P** suivante :

$$\mathbf{P} : \underset{\|F\|=1;\|G\|=1}{Max} \langle R_Y^M F \mid R_Z^N G \rangle$$

Il s'agit donc de maximiser la covariance des résultantes des facteurs sur leurs groupes respectifs.

4.2. Résolution du programme P

4.2.1. Résolution générale

• Note : la maximisation doit être effectuée sous les contraintes de sous-espace : $F \in < Y >$ et $G \in < Z >$. En fait, il apparaît que les solutions de la maximisation libérée de ces contraintes les vérifient.

• Le lagrangien de **P** est : $L = F' R_Y^M R_Z^N G - \lambda F' F - \mu G' G$

$$\frac{\partial L}{\partial F} = 0 \Leftrightarrow R_Y^M R_Z^N G = 2\lambda F \ (5); \ \frac{\partial L}{\partial G} = 0 \Leftrightarrow R_Z^N R_Y^M F = 2\mu G \ (5')$$

En notant $\eta = 4\lambda\mu$, on a : (5) et (5') $\Rightarrow R_Y^M R_Z^N R_Z^N R_Y^M F = \eta F$ (6) et $R_Z^N R_Y^M R_Y^M R_Z^N G = \eta G$ (6')

En outre, $F'(5) = G'(5')$ implique $\lambda = \mu$, à maximiser. On prendra donc η maximale.

La diagonalisation des produits des opérateurs de résultante croisée fournit ainsi des couples de facteurs (F, G) associés à une même valeur propre. On classe ces couples par ordre de valeurs propres décroissantes. La solution du programme initial est le couple de rang 1 : (F_1, G_1).

• Les équations (6) et (6') montrent, d'une part, que $F \in < Y >$ et $G \in < Z >$, et d'autre part, que les facteurs F (respectivement G), vecteurs propres d'une matrice symétrique, sont deux à deux orthogonaux. On reprend ensuite le programme **P**, en y ajoutant les contraintes $F \perp F_1$, $G \perp G_1$. On obtient alors le couple de rang 2 du classement précédent, etc.

• Si l'on considère Y comme un groupe explicatif et Z comme un groupe à expliquer, on pourra régresser le groupe Z et ses facteurs G sur les facteurs explicatifs F.

• On notera que les facteurs ne sont pas a priori colinéaires à leur résultante. Les résultantes des F (respectivement G) ne sont d'ailleurs pas orthogonales entre elles.

• On n'oubliera pas non plus que la corrélation prise en compte dans le critère de l'ARC est celle des résultantes et non celle des facteurs eux-mêmes. La première tient davantage compte que la seconde des liaisons dans les groupes. Mais les deux types de corrélations fournissent des indications utiles sur la liaison entre les groupes.

• *Caractéristiques de la méthode*

– Dans une optique exploratoire, la méthode a pour avantage de permettre un traitement symétrique des deux groupes. L'ARC, initialement, ne cherche pas explicitement à expliquer Z à partir de Y, mais seulement des structures homologues dans les deux groupes.

– L'ARC fournit systématiquement des facteurs non redondants.

– En jouant sur les pondérations des variables, on peut régler de manière fine le rôle des intra-structures et inter-structures.

– Dans l'optique de l'explication du groupe Z par le groupe Y, la décorrélation des facteurs F est un avantage. Dans la régression des variables de Z (ou des facteurs G) sur les facteurs F, la part de variance expliquée par chaque facteur F n'est autre que le cosinus carré de l'angle entre la variable régressée et F. Ces régressions fournissent des équations du type : $z^k = \sum_\alpha \beta_{k\alpha} F_\alpha + \varepsilon$. En y reportant l'expression des facteurs F en fonction des variables explicatives y^j, on peut reconstruire un modèle «consolidé» des variables à expliquer en fonction des explicatives : $z^k = \sum_j b_{kj} y^j + \varepsilon$.

– Toutefois, l'ARC risque de ne pas fournir assez de facteurs pour accéder au plein pouvoir explicatif de Y. En effet, les opérateurs $R_Z^N R_Y^M R_Y^M R_Z^N$ et $R_Y^M R_Z^N R_Z^N R_Y^M$ ont un rang égal au plus petit des rangs de Y et de Z. Le nombre des facteurs associés à une valeur propre non nulle que fournit leur diagonalisation est égal à ce rang. Donc, si le groupe Z contient moins de variables que Y, les facteurs F ne pourront a priori engendrer $< Y >$ entier.

On peut alors adopter deux attitudes, selon les résultats fournis par la première étape de l'ARC :

Si les facteurs F fournis ont un pouvoir explicatif de Z jugé suffisant, on s'en contentera pour les régressions (la réduction dimensionnelle des groupes est l'un des objectifs pratiques de la méthode).

Si tel n'est pas le cas, on pourra adopter la même démarche que dans PLS pour la recherche de facteurs explicatifs supplémentaires : régresser toutes les variables de Y et Z sur un certain nombre des premiers facteurs F obtenus, et relancer l'ARC sur les résidus de régression [5] . On obtient alors de nouveaux facteurs explicatifs, orthogonaux aux premiers, ainsi que de nouveaux facteurs G, mais qui, eux, ne sont pas orthogonaux aux précédents.

• *Calcul pratique de l'ARC*

On diagonalise par exemple la matrice $S = R_Y^M R_Z^N R_Z^N R_Y^M$ pour obtenir les facteurs F (équation (6)). A partir de ceux-ci, on calcule les facteurs G à l'aide de $(5') \Leftrightarrow G = \frac{1}{\sqrt{\eta}} R_Z^N R_Y^M F$.

Note : si le nombre I d'individus est beaucoup plus grand que ceux J et K des variables, ce qui sera fréquent en pratique, la matrice S est inutilement encombrante :

[5] Cette ARC devant prendre en compte l'inégalité de norme des résidus, ceux-ci ne doivent pas être préalablement réduits. On leur donne par ailleurs la même structure de pondération N que les variables z^l dont ils sont issus.

elle est de dimension $I \times I$, mais de rang au plus égal à inf(J, K). On note alors que :

$$Y'(6) \Leftrightarrow Y'YMY'ZNZ'ZNZ'YMw = \eta w \ (7) \quad \text{avec} \quad w = Y'F$$

La matrice à diagonaliser est alors : $H = Y'YMY'ZNZ'ZNZ'YM$, de dimension $J \times J$.

A partir de ses couples propres (w, η), on retrouve les facteurs F de l'ARC en utilisant (6) :

$$(6) \Leftrightarrow YMY'ZNZ'ZNZ'YMw = \eta F \Leftrightarrow F = \frac{1}{\eta} YMY'ZNZ'ZNZ'YMw$$

4.2.2. Cas où Z est réduit à une variable z (ARC1)

Dans ce cas le facteur G, comme sa résultante sur z, est colinéaire à z elle-même. Le programme $\mathbf{P}$ équivaut donc ici à : $\mathbf{P1}$: $\underset{\|F\|=1}{Max} \langle R_Y^M F \mid z \rangle$. Sa résolution donne : $F \propto R_Y^M z$. On remarque qu'il s'agit aussi de la solution de $\mathbf{Q1}$. Le facteur F est donc la résultante (normée) de z sur le groupe Y. Le cas particulier $M = (Y'Y)^{-1}$ redonne la régression multiple de z sur Y. Dans le cas le plus général, M permet de moduler l'importance donnée à l'intra-structure du groupe Y dans la recherche d'une explication de z.

4.3. Rapport entre les facteurs de rang 1 de P et de Q

Il est aisé de montrer que lorsque les facteurs F_1 et G_1 de $\mathbf{Q}$ sont très proches de composantes principales des ACP respectives de Y et de Z (pondérés par M et N), ce qui risque notamment d'être le cas dès que les groupes sont fortement liés (ils ont une première composante principale voisine), alors l'ARC donnera également F_1 et G_1 pour facteurs de rang 1.

Preuve : reprenons les équations reliant les premiers facteurs de $\mathbf{Q}$:

$$(3) \quad YMY'G = \sqrt{\eta} F \ : \ (3') \quad ZNZ'F = \sqrt{\eta} G$$

Or, si F (resp. G) est très proche d'un facteur d'ACP de Y (resp. Z), on a :

$$(8) \quad YMY'F \propto F \ ; \ (8') \quad ZNZ'G \propto G$$

Alors :

$$ZNZ'(8) \Leftrightarrow ZNZ'YMY'F \propto ZNZ'F \, ; \, YMY'(8') \Leftrightarrow YMY'ZNZ'G \propto YMY'G$$

En tenant compte de (3) et (3′), on a donc :

$$(8) \quad ZNZ'YMY'F \propto G \ ; \ (8') \quad YMY'ZNZ'G \propto F$$

Ce qui n'est autre que la caractérisation des facteurs de l'ARC.

On ne sera donc pas étonné, en pratique, par l'éventuelle grande similitude des facteurs de rang 1 des deux analyses.

4.4. Présentation et interprétation des résultats

• On note d'abord la force des liaisons dépistées à l'aide des éléments suivants :

– La variance de chaque résultante $R_Y F$ (resp. $R_Z G$) témoigne des liaisons intra-groupe dans la direction du facteur F (resp. G). Alternativement, on pourra calculer la variance de ce facteur (égale à l'inertie du groupe sur le facteur).

– La corrélation de ces résultantes mesure la liaison inter-groupe correspondante.

– La corrélation de chaque couple (F, G) mesure également une liaison partielle entre groupes, mais en tenant moins compte des liaisons internes aux groupes.

• On interprète ensuite les facteurs et les résultantes à l'aide des variables des groupes, à partir des graphiques suivants :

– La représentation des variables y^k, z^l et des G dans la base des F, leur coordonnée sur un facteur étant égale à leur corrélation avec ce dernier. On pourra également y représenter la résultante normée de chaque facteur.

– La représentation homologue de ces variables dans la base des G_α.

• Les liaisons dépistées entre variables sont ensuite examinées du point de vue des individus, ce qui permet de trouver ceux qui contribuent positivement à ces liaisons et ceux qui s'y opposent. On utilise les graphiques représentant le nuage des individus repérés par leurs valeurs pour les F (respectivement G).

Enfin, on pourra fournir les indicateurs habituels de qualité de représentation des individus sur les axes et les plans : CO2, QLT.

4.5. L'ARC alternative à PLS

4.5.1. La méthode

• PLS est obtenue en portant dans $\mathbf{Q}$: $M = I_J$, $N = I_K$. On reprend ce choix de pondérations dans le programme $\mathbf{P}$ de l'ARC. La méthode obtenue sera qualifiée d'ARC *simple*.

L'argumentation heuristique est la suivante :

$$\mathrm{cov}(R_Y(F), R_Z(G)) = \sqrt{\mathrm{var}(R_Y(F))}.\sqrt{\mathrm{var}(R_Z(G))}.\rho(R_Y(F), R_Z(G))$$

La maximisation de la variance de chacune des résultantes tend à produire un facteur résumant bien les corrélations entre variables dans le groupe correspondant, tandis que leur corrélation exerce une contrainte orientant la recherche vers les structures communes.

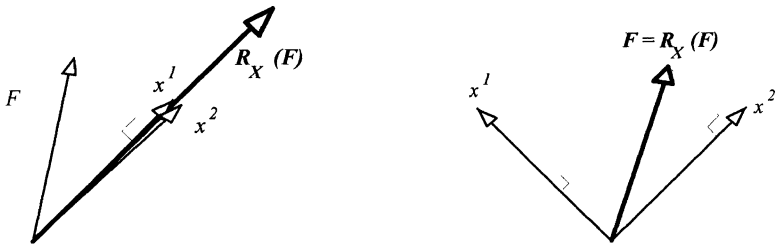
4.5.2. Comportement de l'ARC simple : quelques cas extrêmes

• Concernant un groupe de variables, deux situations extrêmes peuvent se présenter :

– situation 1 : ces variables présentent une structure de liaison maximale : elles sont de corrélation 1 en valeur absolue. Elles sont donc représentées par des vecteurs colinéaires.

– situation 2 : ces variables ne présentent aucune structure de liaison : elles sont toutes deux à deux décorrélées. Les vecteurs qui les représentent sont deux à deux orthogonaux.

Dans les deux cas, il est particulièrement aisé de voir ce que devient la résultante d'un facteur engendré par le groupe (figure 3).



situation 1 : situation 2 :

FIGURE 3

Dans la situation 1, la résultante est colinéaire aux x^j et sa norme est maximale lorsque l'on prend F colinéaire aux x^j aussi. Dans la situation 2, la base des x^j est orthogonale, F s'y décompose comme la somme de ses projections orthogonales sur les x^j. F est donc égal à sa résultante et sa norme est constante.

On peut, à partir de ces cas limites, croiser les situations dans le problème à deux groupes et chercher à quoi revient la maximisation de $\langle \underline{R}_Y F \mid \underline{R}_Z G \rangle$.

– *Situation 1-1 :* les deux groupes sont dans la situation 1 (les variables de chaque groupe sont colinéaires). Dans ce cas, l'angle des deux résultantes reste constant : c'est celui des deux groupes. Maximiser la covariance revient donc (si l'angle entre les groupes n'est pas nul) à maximiser les normes des résultantes, ce qui peut être fait séparément dans chaque groupe. Le facteur optimal de chaque groupe est celui qui a la direction des variables du groupe, et le sens qui donne un cosinus positif avec celui de l'autre groupe.

– *Situation 2-2 :* les deux groupes sont dans la situation 2 (les variables de chaque groupe sont orthogonales). Dans ce cas, la norme des deux résultantes reste constante égale à 1. Maximiser la covariance revient donc à maximiser le cosinus des résultantes, ce qui est exactement l'Analyse Canonique (AC). L'AC ne prend pas en compte les structures de liaison internes des groupes. Etant donné qu'ici, il n'y en a aucune, il est satisfaisant de retrouver cette technique.

– *Situation 1-2 (mixte) :* l'un des groupes (*e.g.* Z) est constitué de variables parfaitement corrélées (en valeur absolue) et l'autre (Y) de variables orthogonales.

Dans ce dernier, la résultante est égale au facteur F et garde une norme constante égale à 1. Dans le premier groupe, la résultante du facteur G reste colinéaire aux variables et a une norme maximale lorsque le facteur l'est aussi. Reste donc à maximiser le cosinus des deux résultantes, ce qui est atteint en projetant G sur $< Y >$. On obtient la régression multiple de Z sur Y.

Là encore, l'intra-structure de Z est parfaitement respectée, et son explication optimale par l'autre groupe est trouvée. Ce dernier n'ayant pas d'intra-structure, il ne compte qu'en tant que sous-espace.

• Considérons enfin la situation des deux groupes l'un par rapport à l'autre.

Supposons que Y et Z sont orthogonaux entre eux, à l'exception de deux variables – une dans chaque groupe : y^1 et z^1 – qui sont corrélées.

$$\operatorname{cov}(R_Y F, R_Z G) = \sum_k \sum_l \operatorname{cov}(\Pi_{y^k} F, \Pi_{z^l} G)$$

$$= \operatorname{cov}(\Pi_{y^1} F, \Pi_{z^1} G) = \|\Pi_{y^1} F\| \, \|\Pi_{z^1} G\| \, \langle y^1 \mid z^1 \rangle$$

Cette covariance est maximale lorsque $\|\Pi_{y^1} F\|$ et $\|\Pi_{z^1} G\|$ le sont, i.e. pour $F = y^1$ et $G = z^1$. L'ARC a donc dépisté les dimensions liées des deux groupes.

4.6. L'ARC alternative à l'ARM

L'ARM est obtenue en portant dans $\mathbf{Q}$: $M = (Y'Y)^{-1}$ et $N = I_K$. On reprend ce choix de pondérations dans le programme $\mathbf{P}$ de l'ARC. On gomme ainsi l'effet de l'intra-structure du groupe explicatif Y, tandis qu'on tient compte de celle du groupe à expliquer Z.

4.7. L'Analyse Canonique est une ARC

Porter dans $\mathbf{Q}$: $M = (Y'Y)^{-1}$ et $N = (Z'Z)^{-1}$ conduit les équations (6) et (6′) à :

$$\Pi_{<Y>}\Pi_{<Z>}\Pi_{<Z>}\Pi_{<Y>}F = \eta F \quad \text{et} \quad \Pi_{<Z>}\Pi_{<Y>}\Pi_{<Y>}\Pi_{<Z>}G = \eta G$$

Ce qui, compte tenu de $\Pi^2_{<Z>} = \Pi_{<Z>}$, $\Pi^2_{<Y>} = \Pi_{<Y>}$, $\Pi_{<Y>}F = F$ et $\Pi_{<Z>}G = G$, redonne les équations de l'Analyse Canonique.

4.8. L'ARC sur données structurées en sous-groupes

Pour généraliser les situations précédentes, on suppose ici que les groupes Y et Z sont divisés en respectivement J et K sous-groupes : $Y = (Y_1, \ldots Y_J)$ et $Z = (Z_1, \ldots Z_K)$ matérialisant chacun un point de vue sur les données.

• Si la redondance interne à un point de vue doit compter autant que celle entre les points de vue, alors on utilisera l'ARC sans tenir compte des sous-groupes (ARC simple).

• Si la redondance interne à un point de vue ne doit pas avoir d'impact, alors qu'il faut tenir compte de la convergence entre les points de vue, l'usage de la résultante multiple (*cf.* § 1.3.) permettra de ne considérer le sous-groupe qu'en tant que sous-espace.

La résultante multiple est également utilisée pour appliquer l'ARC aux données qualitatives ou mixtes. Chaque variable qualitative d'un groupe est représentée par l'ensemble des indicatrices de ses modalités, considéré comme un sous-groupe.

• Enfin, s'il s'agit d'harmoniser les contributions des intra-structures des sous-groupes, on utilisera pour matrice de pondération du groupe la matrice $M = diag\left(diag\left(\frac{1}{\mu_j}\right)\right)_j$ où μ_j est la première valeur propre d'ACP du sous-groupe j (cf. § 2.2.). En effet, la résultante d'un facteur sur le groupe entier peut s'écrire comme la somme de ses sous-résultantes sur les différents sous-groupes. Si on ne pondère aucune variable, la norme carrée maximale de la sous-résultante de F sur X_j est μ_j^2 (*cf.* § 2.1.). Ce sont ces normes que la pondération ci-dessus permet d'égaliser.

5. Deux exemples d'application comparée de PLS et de l'ARC simple

5.1. Données de Linnerud (cf. annexe)

On a appliqué les deux méthodes aux données de Linnerud (*cf.* [6]). Vingt usagers d'un club de gymnastique ont été soumis à trois exercices physiques (tractions à la barre fixe, flexions, sauts), et on a par ailleurs mesuré sur ces usagers trois paramètres physiques (poids, tour de taille et pouls). Il s'agit d'expliquer la structure des résultats aux exercices à l'aide de celle des paramètres. On a donc :

$Y = \{y_1 = \text{poids},\ y_2 = \text{tour de taille},\ y_3 = \text{pouls}\}$;

$Z = \{z_1 = \text{tractions},\ z_2 = \text{flexions},\ z_3 = \text{sauts}\}$

a) Facteurs

Les corrélations entre facteurs F et G sont données dans le tableau 1.

TABLEAU 1

PLS			ARC				
			corrélations	RF1	RF2	G1	G2
			RG1	0,50		0,99	
corrélations	G1	G2	RG2		0,12		0,93
F1	**0,56**	0,00	F1	1,00		**0,55**	
F2	0,20	**0,30**	F2		0,69		**0,33**

La corrélation entre les résultantes de rang 2 est faible. Les deux groupes sont donc globalement peu liés dans cette direction. Les facteurs correspondants ont une corrélation plus forte, mais représentent moins bien leur groupe (corr(F2,RF2) = 0,69). Les corrélations de type (F_α, G_α) fournies par les deux méthodes sont très proches. Pour les facteurs de rang 1, cela vient du fait qu'ils sont pratiquement identiques dans les deux cas, comme on le voit sur le tableau suivant.

b) Corrélations entre facteurs des deux analyses (tableau 2) :

TABLEAU 2

corrélations	F1PLS	F2PLS	G1PLS	G2PLS
F1ARC	**1,00**	0,00		
F2ARC	0,00	0,24		
G1ARC			**1,00**	0,81
G2ARC			0,00	0,05

Les facteurs de rang 2 fournis par les deux méthodes n'ont pas grand chose à voir. On voit plus loin que la structure de Y est presque unidimensionnelle, ce qui fait que la plus grande partie des variables explicatives y^k et de ce qu'elles peuvent expliquer des z^l est captée par le facteur F_1. Le facteur F_2 est donc assez faiblement déterminé, et sa variabilité d'une méthode à l'autre est peu surprenante.

c) Représentation des variables

La représentation des variables sur le plan (F1,F2) de PLS est donnée par la figure 4. Compte tenu de la non-orthogonalité de G_1 et G_2 dans PLS, il est impossible de représenter les variables dans leur base comme on le fait dans celle des F. Les représentations des variables dans les plans (F1,F2) et (G1,G2) de l'ARC sont données par les figures 5 et 6.

On remarque sur la figure 5 que les variables de Y sont moins bien représentées sur F_2 ARC qu'elles ne le sont sur F_2 PLS. Mais cette qualité n'est pas seule en jeu : la liaison de ce facteur avec un facteur représentatif du groupe Z doit également être considérée.

La figure 6 montre que le facteur G_2 de l'ARC oppose partiellement tractions et sauts. Il est donc susceptible de nuancer l'information de l'axe 1, qui confond les trois variables à expliquer. Le facteur G_2 de PLS ne remplit pas ce rôle, dans la mesure où ses corrélations avec les variables à expliquer sont :

corrélations	tractions	flexions	sauts
G2 PLS	0,76	0,72	0,60

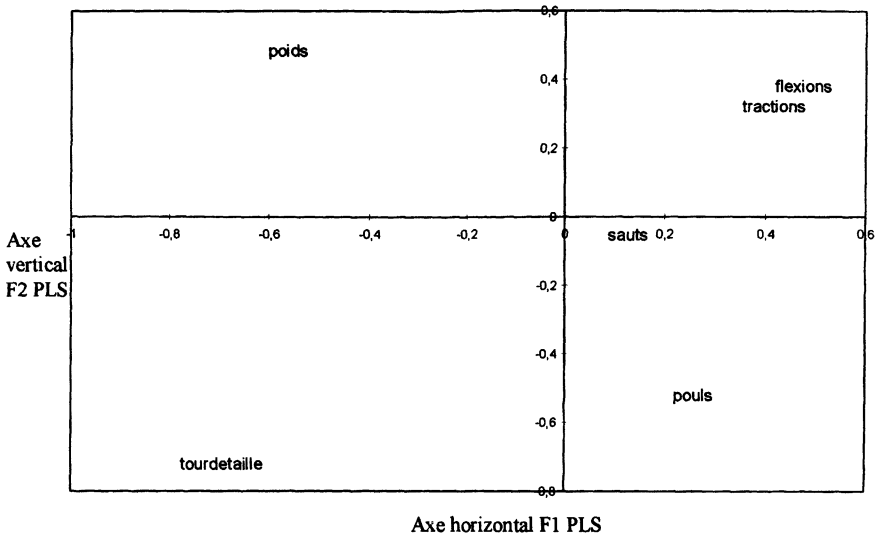


FIGURE 4
Représentation des variables dans le plan (F1,F2) de PLS

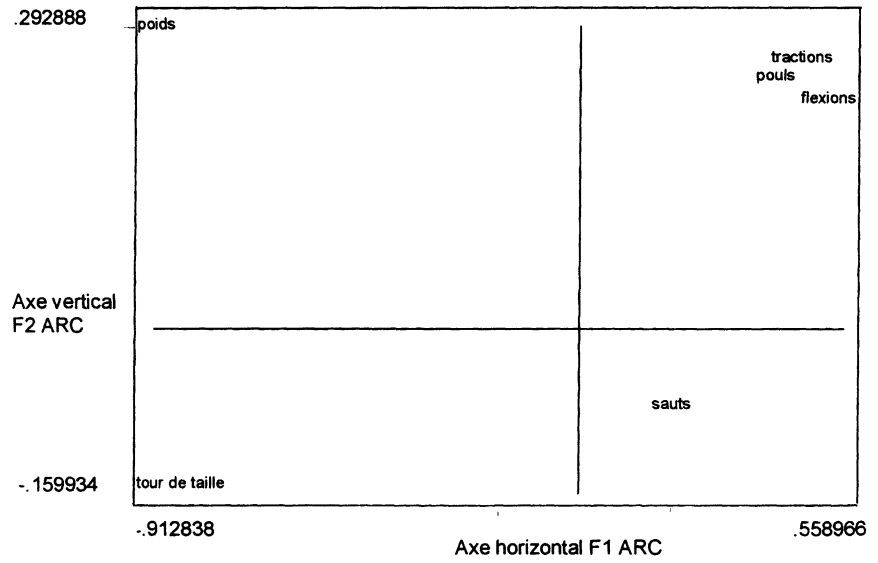


FIGURE 5
Représentation des variables dans le plan (F1,F2) de l'ARC

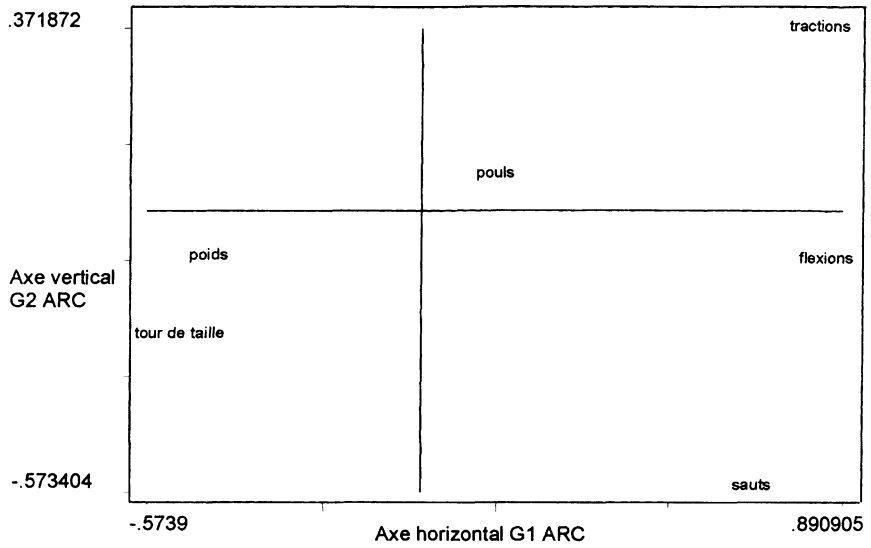


FIGURE 6
Représentation des variables dans le plan (G1,G2) de l'ARC

d) Représentation des individus

La représentation des individus dans le plan (F1,F2) de PLS est donnée par la figure 7. La forte corrélation des facteurs G1 et G2 de PLS rend peu pertinente la représentation des individus sur ce couple de facteurs. Les représentations des individus dans les plans (F1,F2) et (G1,G2) de l'ARC sont données par les figures 8 et 9.

Ces représentations permettent dans les deux cas de dépister les individus contribuant positivement à la liaison dépistée par un axe α : les projections d'un tel point sur F_α et G_α sont voisines (c'est le cas de l'individu n pour les deux facteurs), et ceux y contribuant négativement : leurs projections sur F_α et G_α sont opposées (l'individu t apparaît ainsi ne pas suivre la liaison dépistée par le premier facteur : ses résultats aux exercices sont médiocres malgré son très faible poids).

Sur les facteurs F_1 et G_1, les projections du nuage sont évidemment presque identiques dans PLS et l'ARC (les facteurs sont quasiment les mêmes). Ce n'est pas du tout le cas sur les facteurs de rang 2. L'axe G2 de l'ARC permet notamment de faire ressortir l'individu k, original en ce qu'il a obtenu un très bon score aux tractions en dépit de résultats médiocres aux autres exercices, et de caractéristiques physiques moyennes.

e) Parts de variance expliquées par les deux premiers facteurs

• Les parts de variance expliquées par les facteurs F de PLS sont données dans le tableau 3. La part de variance totale expliquée par les facteurs $G1$ et $G2$ de PLS est de 0,74.

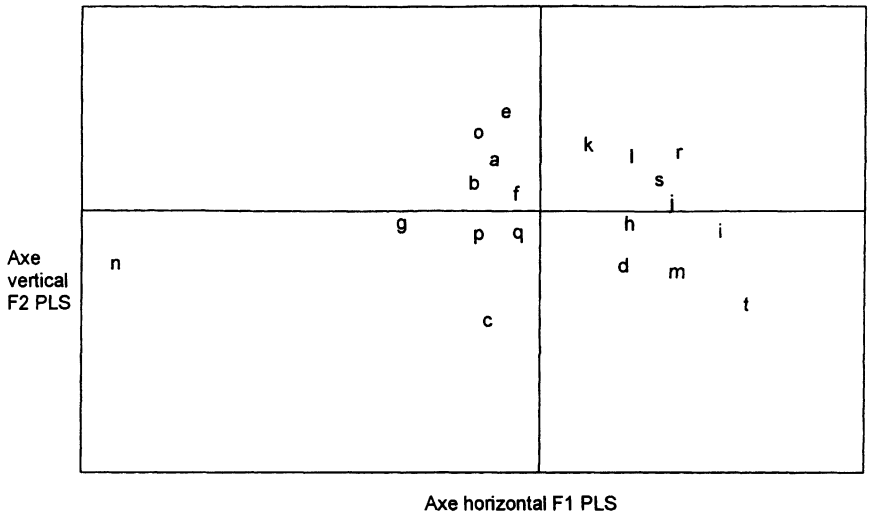


FIGURE 7
Représentation des individus dans le plan (F1,F2) de PLS

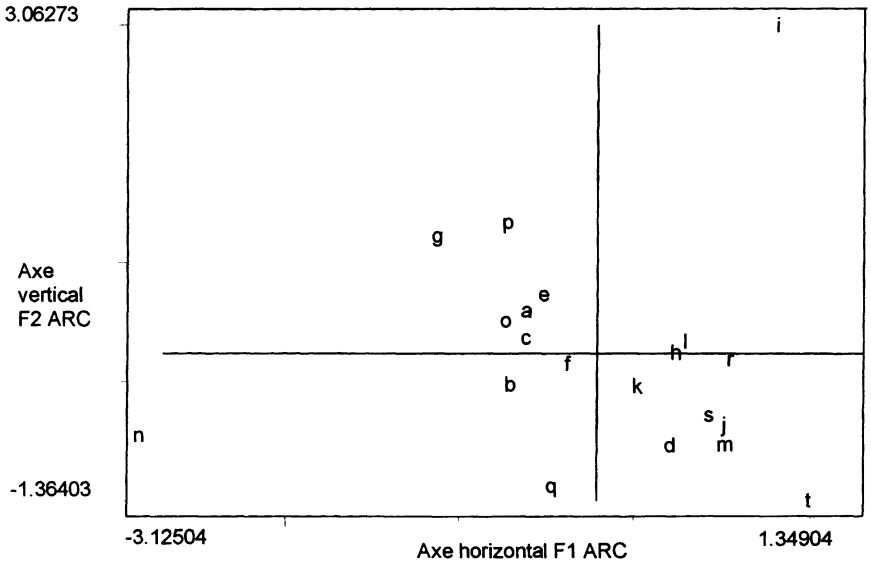


FIGURE 8
Représentation des individus dans le plan (F1,F2) de l'ARC

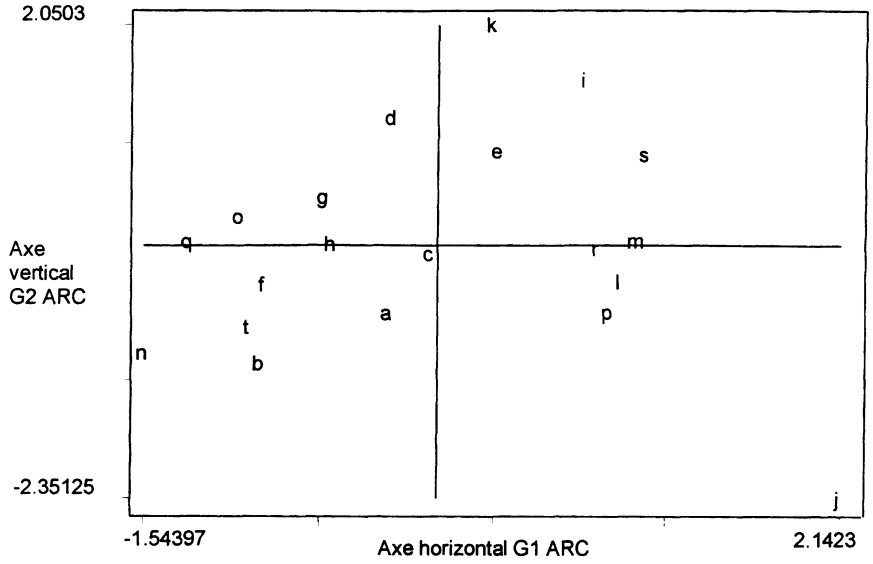


FIGURE 9
Représentation des individus dans le plan (G1,G2) de l'ARC

TABLEAU 3
Parts de variance expliquées par les facteurs F de PLS

	variables explicatives				variables expliquées			
	poids	**tour de taille**	**pouls**	total	**tractions**	**flexions**	**sauts**	total
F1PLS	0,90	0,93	0,19	**0,67**	0,25	0,35	0,05	**0,21**
F2PLS	0,00	0,06	0,61	**0,22**	0,04	0,05	0,00	**0,03**
F1+F2PLS	**0,90**	**0,99**	**0,80**	**0,90**	**0,28**	**0,40**	**0,05**	**0,24**

• Les parts de variance expliquées par les facteurs F et G de l'ARC sont données respectivement dans les tableaux 4 et 5.

Notes : PLS fait un certain cas du pouls, variable très peu liée à l'autre groupe, alors que l'ARC ne reprend quasiment pas cette variable. C'est ce qui fait baisser la part de variance des y^k expliquée par le plan (F_1, F_2) de l'ARC par rapport à celui de PLS.

La part de variance expliquée par le facteur F_2 de l'ARC (5 %) est très légèrement supérieure à celle du facteur F_2 de PLS (3 %), ce qui donne au

TABLEAU 4
Parts de variance expliquées par les facteurs F de l'ARC

	variables explicatives				variables expliquées			
	poids	**tour de taille**	**pouls**	total	**tractions**	**flexions**	**sauts**	total
F1ARC	0,90	0,92	0,19	**0,67**	0,25	0,35	0,05	**0,22**
F2ARC	0,10	0,03	0,06	**0,06**	0,08	0,06	0,01	**0,05**
F1+F2ARC	**1,00**	**0,95**	**0,25**	**0,73**	**0,33**	**0,41**	**0,06**	**0,27**

TABLEAU 5
Parts de variance expliquées par les facteurs G de l'ARC

	variables expliquées			
	tractions	**flexions**	**sauts**	total
G1ARC	0,78	0,88	0,56	**0,74**
G2ARC	0,15	0,01	0,37	**0,18**
G1+G2ARC	**0,93**	**0,89**	**0,92**	**0,92**

plan (F_1, F_2) de l'ARC un pouvoir prédictif très légèrement meilleur que le plan homologue de PLS.

Les facteurs G_1 et G_2 de l'ARC permettent ici de mieux représenter l'ensemble des variables z^l (92 % de variance captée par le plan) que ceux de PLS (74 % de variance captée).

Conclusion : Dans cet exemple, les deux méthodes concordent sur l'essentiel : le facteur 1. La représentation des variables explicatives y est moins bonne dans l'ARC que dans PLS (l'ARC faisant mal apparaître la variable peu liée à l'autre groupe, ce qu'on pourra considérer comme une qualité), mais la représentation des variables de Z est meilleure : la visualisation de leurs corrélations avec les facteurs G est beaucoup plus facile, et le plan (G_1, G_2) capte une plus grande part de leur structure.

Ces constatations ne sont guère surprenantes : dans l'ARC, le peu de liaison du pouls avec l'autre groupe empêche sa prise en compte par le facteur F_2, ce qui en retour «libère» le facteur G_2, qui peut alors répondre à l'exigence de représentativité du groupe Z, dans la mesure où ses variables sont liées au groupe Y, ce qui est le cas des trois variables à expliquer. Dans PLS, les choses se passent différemment : le facteur F_2 est construit pour bien représenter les variables explicatives non prises en compte par F_1 et les variables de Z mal expliquées par F_1, dans la mesure où les deux lots sont liés. Ici, les trois variables de Z étant relativement bien expliquées par F_1, il ne reste presque rien du groupe Z à prendre en compte, alors qu'il reste toujours une

variable à représenter dans le groupe Y (le pouls). C'est elle que le facteur F_2 PLS vient illustrer.

Il nous semble ici que du point de vue descriptif, l'ARC s'est mieux comportée que PLS : il paraît en effet plus important de bien représenter, sur les facteurs, les variables de chaque groupe qui sont liées à l'autre groupe, plutôt que les variables d'un groupe ou l'autre en particulier.

f) Synthèse des modèles finaux de Z en fonction de Y

Il a été noté (*cf.* [4]), dans l'application de PLS à ces données, que la synthèse du modèle explicatif à partir du seul premier facteur F1 devait être préférée à celle fondée sur les deux premiers facteurs F1 et F2, dans la mesure où F2 n'est lié qu'à une variable explicative (pouls), très peu liée aux variables à expliquer, donc susceptible de n'être qu'un bruit. On peut bien sûr tirer la même conclusion concernant les résultats fournis par l'ARC, dont le facteur F2 n'apparaît lié à aucune variable explicative. L'ARC fournissant ici quasiment le même premier facteur que PLS, la synthèse à partir du premier facteur fournit pratiquement les mêmes modèles dans les deux analyses (tableau 6).

TABLEAU 6
Coefficients des modèles reconstitués à partir du premier facteur

	cste		**poids**		**tour de taille**		**pouls**	
	PLS	ARC	PLS	ARC	PLS	ARC	PLS	ARC
Tractions	29,20	30,24	−0,04	−0,05	−0,44	−0,44	0,06	0,05
Flexions	430,25	437,96	−0,62	−0,64	−6,27	−6,25	0,86	0,76
Sauts	150,48	157,04	−0,18	−0,19	−1,76	−1,85	0,24	0,23

5.2. Scores régionaux à l'élection présidentielle française de 1995

On dispose, pour les 26 régions françaises, des scores obtenus aux deux tours par les candidats à l'élection présidentielle de 1995 [6]. Ces scores sont calculés comme le rapport des suffrages aux électeurs inscrits.

On a appliqué PLS et l'ARC à ces données en prenant les scores du premier tour comme variables explicatives et ceux du second tour comme variables à expliquer. On cherche ainsi à obtenir une vision macroscopique du report de voix entre les deux tours.

Les résultats obtenus par les deux méthodes sont très bons, car les résultats régionaux des deux tours sont fortement liés. En conséquence, ces méthodes étant

[6] Le candidat Jacques Cheminade a été ôté des données pour cause de scores insignifiants.

faites pour dépister en priorité les structures fortes communes, elles donnent ici des résultats proches.

Le groupe Z est constitué de trois variables : les scores au second tour de MM. Jospin et Chirac, ainsi que le taux d'abstention. La somme de ces trois variables fait invariablement 100 %. Par conséquent, la version centrée de ces variables n'engendre qu'un sous-espace de dimension 2. L'ARC ne fournira donc au départ que deux paires de facteurs (F, G). Comparons, dans un premier temps, les résultats des deux méthodes sur leur premier plan factoriel.

On note d'abord l'identité des facteurs F_1 (resp. G_1) de l'ARC et PLS, et la très forte corrélation (*cf.* tableau 7) des facteurs F_2 (resp. G_2). Les représentations sur le plan (F_1, F_2) des deux méthodes seront donc pratiquement isomorphes.

TABLEAU 7
Corrélations entre facteurs de PLS et de l'ARC

corrélations	F1PLS	F2PLS	corrélations	G1PLS	G2PLS
F1ARC	1,00		G1ARC	1,00	
F2ARC		−0,82	G2ARC		−0,81

Ici, on n'observe pas de divergence entre les facteurs de rang 2 de l'ARC et PLS (au contraire de ce qui a été noté sur les données de Linnerud). Cela vient du fait que la liaison entre les deux groupes, pleinement bidimensionnelle, détermine fortement les *deux* facteurs.

Les corrélations des paires de facteurs (F, G) dans chaque méthode sont données dans le tableau 8.

TABLEAU 8

PLS :

corrélations	G1PLS	G2PLS
F1PLS	**0,95**	
F2PLS		**0,69**

ARC :

corrélations	RF1	RF2	G1ARC	G2ARC
RG1	0,93		1,00	
RG2		0,41		1,00
F1ARC	1,00		**0,95**	
F2ARC		0,83		**0,75**

La liaison partielle de rang 2 entre les groupes mesurée par la corrélation (F2PLS,G2PLS) est moins forte que celle que mesure la corrélation (F2ARC,G2ARC). Les facteurs G1 et G2 de PLS sont partiellement redondants (leur corrélation égale

0,2), contrairement à ceux de l'ARC. Dans cette dernière, la corrélation (F2,G2) un peu plus élevée permettra un recollement des nuages légèrement meilleur.

Représentations sur les plans factoriels (figures 10 à 13) :

On se contentera de donner celles de l'ARC, puisqu'elle permet la double représentation (F_1, F_2) et (G_1, G_2), et que PLS donne sur (F_1, F_2) une image très similaire à celle de l'ARC.

On constate sur les figures 11 et 13 un très bon recollement des images, ce qui dénote ici (l'ensemble des variables étant bien représenté) un fort mécanisme macroscopique de report de voix. Globalement, les régions ayant eu les taux d'abstention les plus forts sont les mêmes pour les deux tours, celles qui ont au premier tour donné leurs meilleurs scores aux candidats de gauche sont celles qui votent le plus pour Jospin au second tour, et celles dans lesquelles les candidats de droite ont eu leurs meilleurs résultats votent davantage pour Chirac au second tour. Les phénomènes se situant en rupture par rapport à cette structure générale, a priori plus intéressants, réclameront une analyse fine. On note par exemple la corrélation relativement faible des scores de Chirac au premier et au second tour. On examinera également dans le détail le cas des régions qui échappent au mécanisme général, i.e. les régions dont les projections sur les deux graphiques ne sont pas voisines. C'est le cas par exemple de la Corse, plus chiraquienne au second tour que le mécanisme d'ensemble ne le laissait prévoir, de la Réunion, qui s'est moins abstenue et a davantage voté pour Jospin qu'attendu, et original du Limousin, fortement chiraquien aux deux tours.

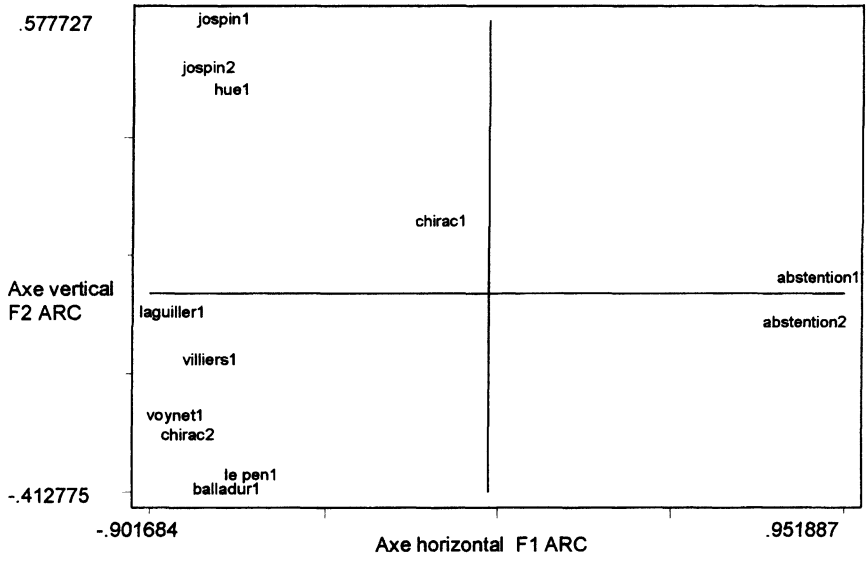


FIGURE 10
Représentation des variables dans le plan (F1,F2) de l'ARC

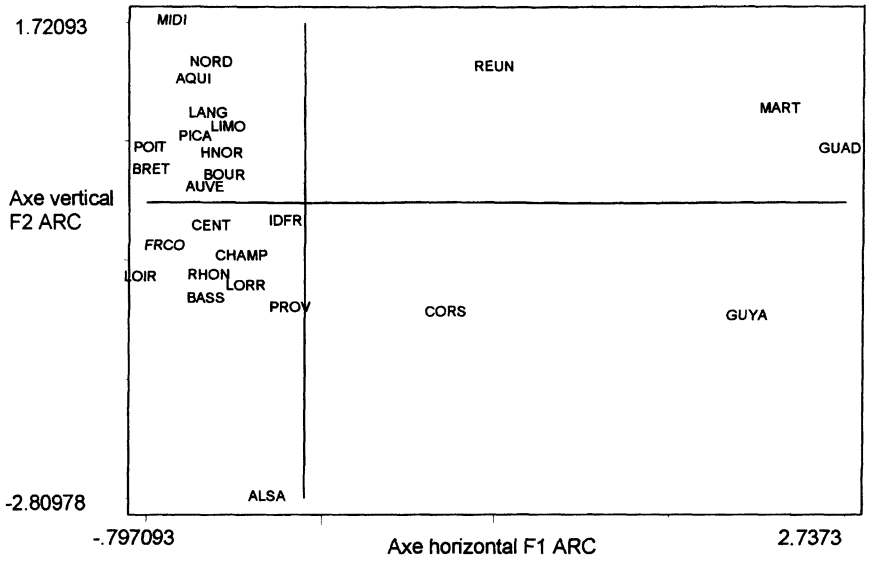


FIGURE 11
Représentation des individus dans le plan (F1,F2) de l'ARC

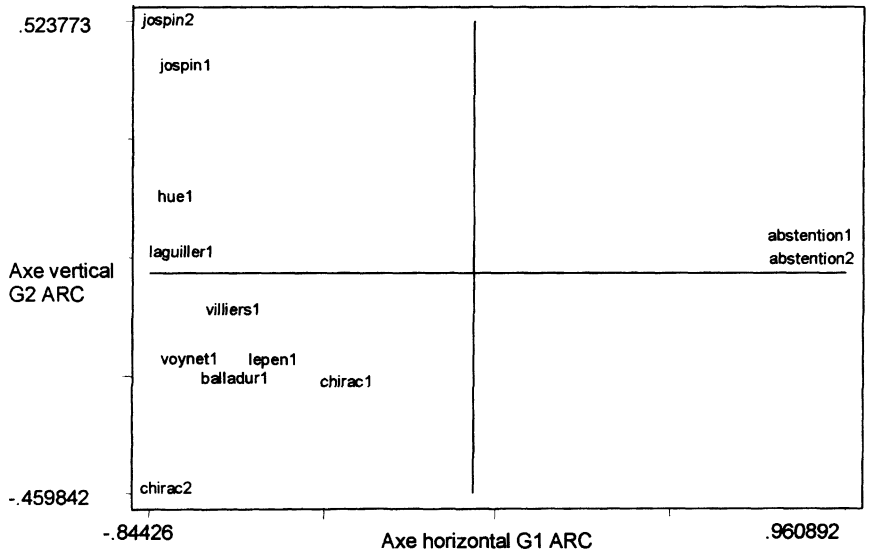


FIGURE 12
Représentation des variables dans le plan (G1,G2) de l'ARC

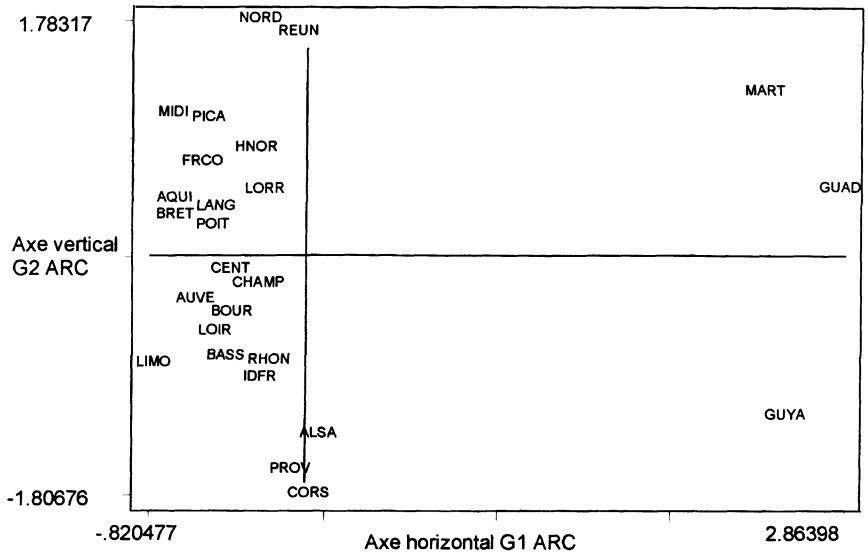


FIGURE 13
Représentation des individus dans le plan (G1,G2) de l'ARC

Considérons à présent les parts de variance des scores du second tour expliquée par les deux premiers facteurs F dans chaque analyse (tableau 9).

TABLEAU 9
Parts de variance du 2^{e} tour expliquée par les deux premiers facteurs du 1er tour

	PLS		ARC
	F1+F2		**F1+F2**
jospin2	0,76	<	0,86
chirac2	0,71	<	0,79
abstentions2	0,95	>	0,91
Total	**0,81**	<	**0,85**

La prévision à partir des deux premiers facteurs explicatifs apparaît globalement meilleure dans l'ARC que dans PLS [7] . Dans cet exemple, la première étape de l'ARC

[7] A titre de comparaison, la régression des scores du second tour sur les deux premières composantes principales du premier tour donne une part de variance expliquée globale de 0,79. Ceci atteste simplement que les structures des deux groupes analysés sont en rapport très étroit. Dans le cas de liens plus marginaux, PLS et l'ARC présenteraient un avantage comparatif plus important.

a donc mieux fonctionné que PLS : le recollement d'image est un peu meilleur (grâce à une plus forte corrélation entre les facteurs F_2 et G_2), et la prévision également.

On cherche ensuite à améliorer la prédiction du second tour en mettant en œuvre la seconde phase de l'ARC : on régresse ici les variables des deux groupes sur les deux premiers facteurs, et on procède une nouvelle fois à l'ARC sur les résidus obtenus. Ces résidus n'ont bien sûr pas été renormés, afin de prendre en compte les phénomènes non encore expliqués selon leur ampleur, dans la recherche d'un nouveau facteur explicatif. Les résultats sont donnés dans le tableau 10.

TABLEAU 10
Parts de variance des scores du 2^e tour expliquée par les trois premiers facteurs du 1^{er} tour

	PLS		ARC
	F1+F2+F3		**F1+F2+F3**
jospin2	0,87	>	0,86
chirac2	0,93	<	0,94
abstentions2	0,96	=	0,96
Total	**0,92**	=	**0,92**

6. Conclusion

L'usage des résultantes dans la formulation des problèmes factoriels nous a paru souple et général. Il permet de formuler les problèmes fondant les méthodes classiques, mais également parfois de les traiter de manière alternative, à l'aide de l'ARC. Il rend en outre immédiate l'extension des méthodes pour variables quantitatives aux données qualitatives et mixtes.

Dans les deux micro-applications présentées ici, l'ARC simple nous a semblé remplir ses objectifs de manière satisfaisante. Sa phase exploratoire est plus simple à programmer que PLS et ses représentations graphiques sont plus naturelles. Il est également possible d'en faire un usage explicatif à l'aide d'une version itérative à l'image de PLS. L'ARC explicative, cependant, est justiciable du même type de reproche que celui que l'on peut faire à PLS, à savoir la fourniture de pseudo-modèles linéaires purement empiriques, qui ne sont qu'une éventuelle première étape de modélisation, et doivent évidemment être dépassés par de véritables modèles économétriques. Il subsiste notamment un problème de fond à laisser intervenir dans un modèle des variables explicatives fortement corrélées, ou pire : liées par une multicolinéarité stricte. Ces liaisons empêchent en effet toute interprétation des coefficients du modèle en termes d'effets «propres», i.e. de variations relatives de la variable endogène par rapport à l'exogène *toutes choses égales par ailleurs dans le*

modèle. Dans le cas d'une multicolinéarité, les coefficients des variables exogènes sont indéterminés si on ne rajoute une contrainte. Cette indétermination est fondamentale, car elle traduit l'impossibilité d'interpréter des coefficients en termes d'effets propres. PLS et l'ARC, à travers leur programme d'optimisation, rajoutent une contrainte de manière sous-jacente (il s'agit d'une contrainte de minimisation de norme du vecteur des coefficients (*cf.* [2]). Le fait d'obtenir un jeu de coefficients bien déterminés ne doit, pour autant, pas occulter le fait qu'aucun de ceux-ci n'est interprétable comme un effet propre. Il nous semble donc sage de se contenter d'un usage purement descriptif de ces méthodes, et de limiter l'utilisation des pseudo-modèles qu'elles fournissent à une prévision «aveugle». Ce qui n'empêche nullement d'utiliser les résultats de l'analyse exploratoire pour construire un véritable modèle des phénomènes étudiés sur une base débruitée.

Remerciements

Je tiens à remercier chaleureusement Odile Wolber et Pierre Cazes, ainsi que Michel Grun-Rehomme, Dominique Guyot et Dominique Desbois, pour leurs encouragements, suggestions et appui documentaire.

Bibliographie

[1] BRY X. (1994), *Analyses Factorielles Simples*, Economica Poche.

[2] BRY X. (1995), *Analyses Factorielles Multiples*, Economica Poche.

[3] CAZES P., BAUMERDER A., BONNEFOUS S., PAGÈS J.P. (1977), *Codage et analyse des tableaux logiques, introduction à la pratique des variables qualitatives*, Cahiers du BURO n° 27.

[4] DESBOIS D. (1999), *Régression PLS*, Cahier des techniques de l'INRA n° 41, INRA.

[5] ESCOFIER B., PAGÈS J. (1990), *Analyses factorielles simples et multiples*, Dunod.

[6] JACKSON J.E. (1991), *A User's Guide to Principal Components*, Wiley, New-York.

[7] LEBART L., MORINEAU A., PIRON M. (1995), *Statistique Exploratoire Multidimensionnelle*, Dunod.

[8] TENENHAUS M. (1998), *La régression PLS, théorie et pratique*, Technip.

Annexe : données

Données de Linnerud

identifiant	poids	tour de taille	pouls	tractions	flexions	sauts
a	191	36	50	5	162	60
b	189	37	52	2	110	60
c	193	38	68	12	101	101
d	162	35	62	12	105	37
e	189	35	46	13	155	58
f	182	36	56	4	101	42
g	211	38	56	8	101	38
h	167	34	60	6	125	40
i	176	31	74	15	200	40
j	154	33	56	17	251	250
k	169	34	50	17	120	38
l	166	33	52	13	210	115
m	154	34	64	14	215	105
n	247	46	50	1	50	50
o	193	36	46	6	70	31
p	202	37	62	12	210	120
q	176	37	54	4	60	25
r	157	32	52	11	230	80
s	156	33	54	15	225	73
t	138	33	68	2	110	43

Scores électoraux des régions

	1er tour (%)									2ème tour (%)		
régions	chirac	jospin	balladur	le pen	laguiller	voynet	hue	villiers	absten-tions	jospin	chirac	absten-tions
ALSACE	12,90	13,06	18,98	19,66	3,80	2,96	2,44	3,35	22,61	30,12	41,40	28,48
AQUITAINE	16,87	21,18	13,93	8,86	4,18	2,39	7,42	3,69	21,26	39,14	39,51	21,35
AUVERGNE	19,42	18,11	13,28	8,90	4,49	2,41	8,12	3,45	21,58	36,27	41,66	22,08
BASSE NORMANDIE	17,22	16,89	17,25	9,71	4,70	2,74	5,01	4,12	22,13	34,43	42,47	23,10
BOURGOGNE	15,48	18,83	14,22	11,21	3,76	2,55	6,63	4,02	23,09	35,08	40,82	24,10
BRETAGNE	16,55	20,46	17,47	8,10	4,74	3,19	6,28	3,24	19,76	39,17	40,11	20,73
CENTRE	15,02	17,33	15,35	11,58	4,00	2,48	6,88	4,73	22,40	35,85	39,74	24,41
CHAMPAGNE ARDENNE	15,30	16,37	14,16	14,03	4,14	2,38	6,05	4,02	23,32	35,43	39,33	25,24
CORSE	20,32	13,46	13,22	6,98	2,00	1,66	6,35	1,47	34,36	29,38	43,08	27,54
FRANCHE COMTE	15,36	19,21	14,58	12,73	4,26	3,84	5,17	3,80	20,85	39,34	38,47	22,19
GUADELOUPE	12,59	11,57	4,79	1,01	0,74	0,45	1,18	0,30	67,07	23,36	19,03	57,61
GUYANE	16,27	9,87	6,90	3,30	1,54	1,07	0,77	0,76	59,16	19,41	26,20	54,40
HAUTE NORMANDIE	13,79	17,70	14,34	12,90	4,85	2,38	8,31	3,32	22,20	38,32	36,34	25,34
ILE DE FRANCE	18,49	17,14	11,95	10,36	3,85	2,53	6,41	2,61	26,50	32,68	41,75	25,57
LANGUEDOC ROUSSILLON	14,10	19,00	12,65	14,34	3,97	2,36	8,72	3,32	21,35	37,79	38,71	23,50
LIMOUSIN	29,79	19,58	6,96	5,18	3,47	2,18	10,87	2,33	19,45	35,85	44,73	19,42
LORRAINE	13,68	16,34	14,88	16,17	4,74	2,56	5,00	3,14	23,26	36,60	37,25	26,14
MARTINIQUE	10,75	13,42	8,70	0,61	0,98	0,47	1,30	0,33	63,12	27,16	18,96	53,87
MIDI PYREENEES	16,14	23,36	13,39	9,08	4,25	2,74	6,87	3,39	20,54	40,88	37,78	21,33
NORD PAS DE CALAIS	13,16	18,70	13,21	13,30	4,55	1,79	9,89	3,22	21,98	41,03	33,47	25,51
PAYS DE LA LOIRE	15,26	17,98	17,67	7,58	4,40	3,01	5,04	7,87	20,96	34,99	41,56	23,45
PICARDIE	15,11	17,90	12,90	14,33	4,69	2,10	8,53	3,82	20,40	39,98	37,01	23,01
POITOU CHARENTES	16,29	19,93	14,47	7,75	4,06	2,73	6,19	5,69	22,64	38,04	38,84	23,12
PROVENCE ALPES COTE D'AZUR	14,08	14,22	14,42	16,19	3,38	2,19	7,07	3,33	24,94	30,19	42,69	27,12
REUNION	21,46	18,52	8,26	1,76	1,48	1,16	6,43	1,36	38,98	40,44	31,67	27,89
RHONE ALPES	13,94	17,02	15,55	14,16	4,09	3,13	5,78	3,72	22,38	33,11	41,29	25,59