REVUE DE STATISTIQUE APPLIQUÉE

JÉRÔME PAGÈS
MICHEL TENENHAUS

Analyse factorielle multiple et approche PLS

Revue de statistique appliquée, tome 50, n° 1 (2002), p. 5-33
<http://www.numdam.org/item?id=RSA_2002__50_1_5_0>

Rev. Statistique Appliquée, 2002, **L** (1), 5-33

ANALYSE FACTORIELLE MULTIPLE ET APPROCHE PLS

Jérôme PAGÈS[1], Michel TENENHAUS[2]

[1] *ENSA-INSFA, Rennes,* [2] *Groupe HEC, Jouy-en-Josas*

RÉSUMÉ

L'analyse factorielle multiple et l'approche PLS s'appliquent à des tableaux dans lesquels un même ensemble d'individus est décrit par plusieurs groupes de variables. Bien que de nature différente, tant par leur objet (l'une est purement descriptive, l'autre repose sur un modèle dont on estime les paramètres), que par les calculs qu'elles impliquent (l'une repose sur une diagonalisation, l'autre sur un algorithme itératif), elles possèdent un fort dénominateur commun : chaque groupe de variables est pris en compte non par le sous-espace qu'il engendre, mais par la répartition de l'inertie des variables dans ce sous-espace. L'article analyse les conséquences de ce dénominateur commun, mettant en évidence une remarquable convergence des résultats en pratique, ainsi que leurs spécificités. Il en résulte que, l'utilisation conjointe des deux méthodes constitue une méthodologie complète d'étude de tableaux multiples, comportant à la fois des aspects descriptifs et modélisateurs.

Ce texte a été présenté au syposium sur les méthodes PLS organisé par le CISIA et le groupe HEC les 5 et 6 octobre 1999 à Jouy-en-Josas.

Mots-clés : *Tableaux multiples, Analyse factorielle multiple, Approche PLS.*

ABSTRACT

Multiple Factor Analysis (MFA) and PLS approach (PLSA) deal with tables in which a same set of individuals is described by several groups of variables. Although they greatly differ, so much by their aim (MFA is a purely descriptive method, PLSA rests on a model whose parameters are estimated), than by calculations they imply (MFA performs the calculus of eigenvectors, PLSA uses iterative algorithm), they have a strong common denominator : each group of variables is taken into account not by the subspace which it generates, but by the distribution of the inertia of the variables in this subspace. The paper analyses the consequences of this common denominator, highlighting a remarkable convergence of the results in practice but also their specificities. It shows that the joint use of the two methods constitutes a complete methodology in the study of multiple tables, comprising both descriptive and modelling aspects.

This text was presented during symposium concerning PLS methods organised by CISIA and HEC group the 5th of October 1999 in Jouy-in-Josas (France).

Keywords : *Multiple tables, Multiple factor analysis, PLS approach.*

Introduction

L'analyse factorielle multiple a été proposée par Escofier & Pagès (1983, 1998) pour rechercher les structures communes à un ensemble de J tableaux de données observés sur les mêmes individus. Elle permet de visualiser ces structures communes au niveau des variables et des individus. Lorsque chaque tableau de données représente un ensemble de manifestations observables d'une variable latente et qu'il existe des relations de causalité explicites entre les variables latentes, il est intéressant d'utiliser l'approche PLS proposée par Wold (1975, 1982, 1985). Cette approche a été reprise et développée par Lohmöller (1987, 1989) sur les plans théorique et informatique. En France l'approche PLS a été particulièrement étudiée par Valette-Florence (1988a,b, 1990) et Tenenhaus (1999).

Lorsqu'il n'y a pas de relations de causalité entre les blocs, on peut encore adopter le point de vue de l'approche PLS. Pour cela Wold (1982) propose de constituer un bloc fictif constitué de la juxtaposition de tous les blocs et de relier chaque bloc individuel au bloc fictif. L'approche PLS permet alors de retrouver diverses méthodes comme l'analyse canonique généralisée de Horst (1961), celle de Carroll (1968), l'analyse en composantes principales et l'analyse factorielle multiple.

Cette article précise les liens entre l'approche PLS et l'analyse factorielle multiple. Les liens entre l'approche PLS et d'autres méthodes d'analyse de tableaux multiples sont présentés dans Guinot, Tenenhaus et Latreille (1999).

1. Données; notations

Les données auxquelles nous nous intéressons présentent une structure très classique : plusieurs groupes de variables ont été mesurées sur le même ensemble d'individus. Ces données peuvent être présentées sous la forme d'un tableau unique (*cf.* Figure 1) structuré en sous-tableaux. Nous notons : X le tableau complet; I l'ensemble des individus; K l'ensemble des variables (tous groupes confondus); J l'ensemble des sous-tableaux; K_j l'ensemble des variables du groupe j; $(K = U_j K_j)$; X_j le tableau associé au groupe j. Les symboles I, J, K ou K_j désignent à la fois l'ensemble et son cardinal. Une variable du groupe K_j est notée : $v_k (k \in K_j)$.

Pour simplifier l'exposé, sans perte de généralité, les variables sont supposées centrées-réduites et de même poids a priori. De même, les individus sont supposés de même poids $1/I$.

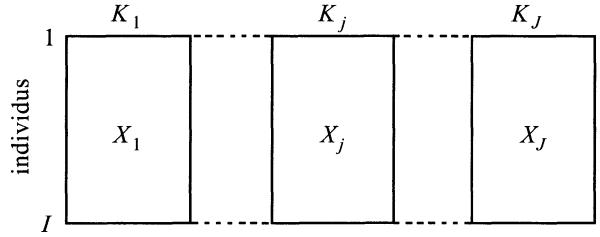


FIGURE 1
Tableau des données

2. Problématique

La problématique associée à ce type de tableau est très riche. Elle présente de nombreux aspects qui tous dérivent, plus au moins directement, de l'interrogation suivante : quelles sont les structures communes aux différents groupes ?

Pour cela, on construit des combinaisons linéaires des variables d'un seul groupe. Ces combinaisons linéaires sont appelées *variables canoniques* ou *composantes* ou encore *variables latentes*. Une structure commune est mise en évidence par un J-uplet de variables canoniques (une par groupe) corrélées entre elles.

La mise en évidence de plusieurs structures communes conduit à rechercher dans chaque groupe j une suite de combinaisons linéaires des variables de ce groupe $\{F_s^j;\ s = 1, S\}$ telles que, entre les groupes, les combinaisons ayant le même rang s $\{F_s^j;\ j = 1, J\}$ soient corrélées le plus possible.

On exprime que F_s^j est combinaison linéaire des variables v_k du groupe K_j en écrivant :

$$F_s^j = \sum_{k \in K_j} a_k^s v_k$$

Différentes méthodes s'inscrivent dans cette problématique. Nous en situons quelques-unes ci-après.

A) *Cas où les variables canoniques de même rang doivent seulement être corrélées entre elles.*

Dans le cas de deux groupes, ce problème est classique et est résolu par l'analyse canonique (Hotelling 1936). F_s^j est la variable canonique de rang s du groupe j. Les variables sont telles que :

$$r(F_s^1,\ F_s^2) \text{ maximum}$$

avec

$$r(F_s^j, F_t^j) = 0 \text{ pour } t < s \text{ et } \mathrm{Var}[F_s^j] = 1\ \forall t, s, j.$$

en notant $r(x, y)$ le coefficient de corrélation entre x et y.

Dans le cas de plus de deux groupes, plusieurs généralisations ont été proposées dont la plus féconde semble être celle de Carroll (1968) décrite plus loin (§ 3).

B) *Cas où les variables canoniques de même rang doivent être corrélées entre elles et aux variables du groupe qu'elles représentent.*

Dans le cas de deux groupes, une solution est donnée par l'analyse inter-batteries de Tucker (1958) bien décrite dans Tenenhaus (1998 p 23). Dans cette méthode, les variables canoniques, appelées ici composantes, sont telles que :

$$\mathrm{cov}(F_s^1, F_s^2) = r(F_s^1, F_s^2)\sqrt{\mathrm{Var}(F_s^1)\mathrm{Var}(F_s^2)} \text{ maximum}$$

Selon ce critère, F_s^1 et F_s^2 sont à la fois corrélées entre elles et «expliquent» une part élevée de la variance du groupe qu'elles représentent.

Les contraintes s'écrivent :

$$\sum_{k \in K_j} (a_k^s)^2 = 1 \quad \text{et} \quad \sum_{k \in K_j} a_k^s a_k^t = 0 \quad \forall s, t$$

Dans le cas général, une solution est donnée par l'AFM décrite plus loin (§ 4).

C) *Cas où tous les groupes ne jouent pas le même rôle.*

Dans le cas de deux groupes, l'un, noté Y, doit être expliqué par l'autre (noté X). La méthode de référence est la régression usuelle (lorsque Y est multidimensionnel, chaque variable de Y est «régressée» séparément) : comme l'analyse canonique usuelle dont elle peut être vue comme un cas particulier, la régression usuelle ne prend en compte les variables du tableau X qu'au travers du sous-espace qu'elles engendrent (et non de la répartition des variables dans ce sous-espace). Les régressions PLS (notée PLS1 si Y ne comporte qu'une variable, PLS2 si non; *cf.* Tenenhaus 1998) résolvent ce problème via des composantes analogues à celles de l'analyse de Tucker.

Dans le cas général, le rôle des groupes est spécifié par un modèle qui relie les variables, dites latentes, représentant les groupes. Ce modèle peut être figuré par un graphe dont les sommets sont les groupes et les arêtes, orientées, les relations de causalité entre groupes. L'approche PLS identifie ces variables latentes en imposant aux seuls groupes reliés par le modèle d'avoir des variables latentes corrélées entre elles.

3. Analyse canonique généralisée de CARROLL

On recherche des suites de variables canoniques, combinaisons linéaires des variables d'un groupe :

$$\left\{ F_s^j = \sum_{k \in K_j} a_k^s v_k ; s = 1, S ;\ j = 1, J \right\}$$

Ces variables ne sont pas obtenues directement. On recherche d'abord une suite de variables normées $\{z_s; s = 1, S\}$, dites générales ou auxiliaires, liées chacune le plus possible à l'ensemble des groupes.

Pour cela, il est nécessaire de définir d'abord une mesure de liaison entre une variable et un groupe de variable K_j. En ACG, cette mesure est le carré du coefficient de corrélation multiple. Soit

$$\text{Liaison } (z_s, K_j) = R^2(z_s, K_j)$$

Il est alors naturel de définir la liaison entre une variable z_s et l'ensemble des groupes de variables par :

$$\text{Liaison } (z_s, U_j K_j) = \Sigma_j R^2(z_s, K_j)$$

En ACG, la variable générale de rang s est définie par :

$$\Sigma_j R^2(z_s, K_j) \text{ maximum}$$
$$\text{avec } \mathrm{Var}[z_s] = 1 \text{ et } r(z_s, z_t) = 0 \; \forall t < s$$

Ces variables sont obtenues par la diagonalisation de $\sum_j X_j (X_j' X_j)^{-1} X_j'$

Une fois la variable générale z_s obtenue, la variable canonique F_s^j est définie par :

$$F_s^j = \sum_{k \in K_j} a_k^s v_k$$
$$\text{et } r(F_s^j, z_s) \text{ maximum}$$

Ce qui revient à réaliser la régression de z_s en fonction de X_j.

Dans cette démarche,

- z_s peut être considérée comme une structure commune aux différents groupes (par construction elle est liée autant que possible à tous les groupes);
- F_s^j la représentation de cette structure commune dans le groupe K_j (combinaison linéaire des variables de K_j liée à z_s).

L'intérêt théorique de l'ACG de Carroll est grand. Dans le cas de deux groupes, on retrouve l'analyse canonique usuelle. Dans le cas où chaque groupe est composé d'une seule variable quantitative, on retrouve l'ACP. Dans le cas où chaque groupe est composé des indicatrices d'une variable qualitative, on retrouve l'ACM. Tout ceci a été étudié par Saporta (1975).

4. Analyse Factorielle Multiple

L'AFM (Escofier & Pagès 1983 et 1998)) procède à la fois de la problématique de l'analyse factorielle (recherche de directions d'inertie maximum) et de l'analyse canonique (recherche de facteurs communs). C'est ce second point de vue que nous adoptons pour la présenter ici.

4.1. Pondération des variables

En AFM, chaque variable v_k du groupe K_j est affecté du poids : $m_k = 1/\lambda_1^j$ en appelant λ_1^j la première valeur propre de l'ACP séparée du groupe K_j. Ainsi, en faisant l'ACP séparée du groupe K_j avec les poids m_k, la première valeur propre

vaut 1. En d'autres termes, avec cette pondération, l'inertie axiale maximum de la configuration des individus, ou de celle des variables, vaut 1.

Notations : M_j est la matrice (égale à l'identité à un coefficient près) des poids des variables du groupe K_j et M la matrice diagonale des poids des variables tous groupes confondus.

4.2. Mesure de liaison entre une variable z et un groupe K_j

En AFM, cette mesure est définie de la façon suivante (les v_k étant supposées centrées-réduites) :

$$\mathcal{L}_{\mathfrak{g}}(z, K_j) = \sum_{k \in K_j} m_k r^2(z, v_k) = \frac{1}{\lambda_1^j} \sum_{k \in K_j} r^2(z, v_k)$$

Cette mesure est apparentée à la redondance (Van den Wollenberg 1977) qui, avec nos notations, s'écrit :

$$R_d(z, K_j) = \frac{1}{K_j} \sum_{k \in K_j} r^2(z, v_k)$$

Les deux mesures varient entre 0 et 1.

Elles vérifient la propriété :

$$\mathcal{L}_{\mathfrak{g}}(z, K_j) = 0 \Leftrightarrow R_d(z, K_j) = 0 \Leftrightarrow r(z, v_k) = 0 \; \forall k \in K_j$$

selon laquelle la liaison entre z et K_j est nulle lorsque z est non corrélée à chaque variable du groupe K_j.

Pour les deux mesures, la valeur de 1 correspond à une liaison maximum mais ce maximum n'a pas la même signification pour R_d et $\mathcal{L}_{\mathfrak{g}}$.

$$R_d(z, K_j) = 1 \Leftrightarrow r(z, v_k) = 1 \; \forall k \in K_j$$

Cette situation, dans laquelle toutes les variables du groupe K_j sont parfaitement corrélées entre elles, ne correspond pas, bien sûr, à une situation concrète.

$$\mathcal{L}_{\mathfrak{g}}(z, K_j) = 1 \Leftrightarrow z \text{ est la première composante principale de } K_j.$$

On retrouve ici l'idée de l'ACP selon laquelle les composantes principales sont des variables synthétiques, c'est-à-dire liées le plus possible aux variables initiales.

4.3. Recherche des variables générales

A l'instar de l'ACG, on cherche en AFM une suite de variables générales (ou auxilliaires) liées le plus possible avec l'ensemble des groupes.

Il est naturel de définir la liaison entre une variable z et l'ensemble des groupes K_j par :

$$\text{Liaison } (z, U_j K_j) = \Sigma_j \mathcal{L}_{\mathfrak{g}}(z, K_j)$$

En AFM la variable générale de rang s est donc définie par :

$$\Sigma_j \mathcal{L}_{\mathfrak{g}}(z_s, K_j) \text{ maximum}$$
$$\text{avec } \mathrm{Var}[z_s] = 1 \text{ et } r(z_s, z_t) = 0 \; \forall t < s$$

En remarquant que :

$$\mathcal{L}_{\mathfrak{g}}(z_s, K_j) = \frac{1}{I^2} z_s' X_j M_j X_j' z_s$$

La quantité à maximiser s'écrit :

$$\sum_j \mathcal{L}_{\mathfrak{g}}(z_s, K_j) = \frac{1}{I^2} \sum_j z_s' X_j M_j X_j' z_s = \frac{1}{I^2} z_s' X M X' z_s$$

Ce qui montre que z_s est la composante principale normée de rang s issue de l'ACP (les variables étant pondérées par M) du tableau complet X.

En notant λ_s la valeur propre de rang s de cette ACP et F_s la composante principale associée, on a : $F_s = \sqrt{\lambda_s} z_s$.

L'intérêt de cette démarche, par rapport à celle de Carroll, est double :

- elle conduit à des variables générales plus liées aux variables initiales tout en assurant, du fait des poids m_k, une forme d'équilibre entre les groupes de variables;
- étant des composantes principales, les variables générales bénéficient de toutes les propriétés de l'ACP.

4.4. Recherche des variables canoniques

Suivant toujours le principe de Carroll, on associe, à chaque variable générale z_s, une variable canonique F_s^j dans chaque groupe j. En AFM, cette variable est définie par :

$$F_s^j = \frac{1}{\sqrt{\lambda_s}} \frac{1}{I} X_j M_j X_j' z_s = \frac{1}{\lambda_s} \frac{1}{I} X_j M_j X_j' F_s = \frac{1}{\lambda_s \lambda_1^j} \frac{1}{I} X_j X_j' F_s$$

A un coefficient près, on retrouve la première composante PLS de la régression PLS de F_s en fonction de X_j.

L'intérêt de la démarche, par rapport à celle de CARROLL, est double :

- ici encore, elle conduit à des variables canoniques plus liées aux variables initiales;

- les F_s^j fournissent chacun une représentation des individus qui peut être superposée à celle de l'ACP (F_s) grâce aux deux propriétés décrites ci- après.

Propriété 1

$$\sum_j F_s^j = \sum_j \frac{1}{\lambda_s}\frac{1}{I} X_j M_j X_j' F_s = \frac{1}{\lambda_s}\frac{1}{I} X M X' F_s = F_s$$

Au coefficient $1/J$ près, la représentation de l'individu i par F_s est au barycentre des représentations de i par les F_s^j. En pratique, on dilate les F_s^j selon le coefficient J pour que $F_s(i)$ soit un barycentre exact. Soit :

$$F_s(i) = \frac{1}{J}\sum_j J F_s^j(i)$$

Dans la terminologie de l'AFM, on dit que l'image globale (*i.e.* du point de vue de l'ensemble des groupes) d'un individu est au centre de gravité de ses images partielles (*i.e.* du point de vue de l'un des groupes).

Propriété 2

Soit $G_s(k)$ la coordonnée de la variable v_k sur l'axe de rang s c'est-à-dire, ici où les variables sont centrées-réduites, le coefficient de corrélation entre v_k et F_s. Notons $v_k(i)$ la valeur de v_k pour l'individu i.

Appliquons à l'AFM la relation de transition usuelle de l'ACP :

$$F_s(i) = \frac{1}{\sqrt{\lambda_s}}\sum_j \frac{1}{\lambda_1^j}\sum_{k\in K_j} v_k(i) G_s(k)$$

Cette relation peut être restreinte au groupe j :

$$F_s^j(i) = \frac{1}{\sqrt{\lambda_s}}\frac{1}{\lambda_1^j}\sum_{k\in K_j} v_k(i) G_s(k)$$

Cette définition de F_s^j est équivalente à celle donnée au début de ce paragraphe, dans laquelle apparaît la quantité

$$\frac{1}{I} X_j' \frac{F_s}{\sqrt{\lambda_s}}$$

dont le k^{e} terme n'est autre que $G_s(k)$.

Cette relation restreinte au groupe j permet une interprétation des F_s^j d'un point de vue analytique (les coefficients $G_s(k)$ sont proportionnels à ceux de la combinaison linéaire exprimant F_s^j en fonction des v_k) et d'un point de vue géométrique (un

individu partiel i^j est du côté des variables pour lesquelles il a une forte valeur et à l'opposé de celles pour lesquelles il a une faible valeur).

Remarque

On peut exprimer matriciellement F_s et F_s^j en tant que combinaisons linéaires des variables initiales. Soit :

$$F_s = X\varphi_s \qquad F_s^j = X_j\varphi_s^j$$

On montre que φ_s^j est le fragment de φ_s correspondant aux variables du groupe j.

5. Régressions PLS et AFM : cas de deux groupes de variables

Avant d'aborder le cas général à J groupes, il est utile d'étudier le cas de deux groupes. Les propriétés respectives des variables canoniques de l'AFM et des composantes PLS apparaissent de façon plus flagrante.

Remarque. Les calculs concernant les méthodes PLS ont été réalisées à l'aide du logiciel LVPLS de Lohmöller (1989). Les options choisies sont les plus courantes : schéma factoriel pour les estimations internes, mode A pour les estimations externes.

5.1. Notations

C_s^j : composante PLS de rang s pour le groupe j. C_s^j est, pour les méthodes PLS, l'analogue de la variable canonique F_s^j pour l'AFM.

M_j : matrice diagonale des poids des variables; elle est en général confondue avec l'identité dans les présentations des méthodes PLS. Dans ce cas, en AFM, cette matrice est égale à la matrice identité mais au coefficient $1/\lambda_1^j$ près.

Dans le cas de la régression PLS, il n'y a que 2 groupes de variables qui, en outre, ne jouent pas le même rôle. Conformément à l'usage, nous réservons la lettre Y au tableau de données à expliquer (éventuellement réduit à une colonne) et X aux données explicatives. Les composantes PLS s'écrivent alors C_s^x et C_s^y; les variables canoniques de l'AFM : F_s^x et F_s^y.

5.2. Cas où l'un des deux groupes est réduit à une variable

Nous définissons la première composante de la régression PLS univariée (PLS1) en reprenant la relation donnée par Tenenhaus (1998) avec nos notations. Soit :

$$C_1^x \propto \frac{1}{I} X X' Y$$

où le terme à gauche du signe $\propto$ est égal au terme de droite centré-réduit.

En AFM (cf. § 4.4) :

$$F_1^x = \frac{1}{\lambda_1 \lambda_1^x} \frac{1}{I} XX4F_1$$

On met en évidence ici une analogie entre les deux approches : prendre en compte un ensemble de variables (ici les colonnes de X) non pas du point de vue du seul sous-espace qu'elles engendrent (comme en régression usuelle) mais en tenant compte de la distribution de ces variables dans le sous-espace. Ainsi, on montre que $(1/I)XX'Y$ est plus corrélé aux colonnes de X que ne l'est la prédiction par régression usuelle $X(X'X)^{-1}X'Y$. (Escofier & Pagès 1998 p. 163).

Les deux formules ne sont toutefois pas identiques : dans la régression PLS, l'opérateur $(1/I)XX'$ s'applique à la variable à prédire Y ; dans l'AFM, il s'applique à la variable générale F_1, qui exprime déjà un compromis entre Y et X. On peut donc s'attendre à :

- une corrélation entre C_1^x et Y plus forte qu'entre F_1^x et Y ;
- des corrélations entre C_1^x et les colonnes de X plus faibles qu'entre F_1^x et les colonnes de X.

En ce sens, C_1^x est intermédiaire entre F_1^x et l'estimation de Y par la régression usuelle (notée Y_{reg}); de son côté, F_1^x est intermédiaire entre C_1^x et la première composante principale de X (notée Cp_1X).

L'analogie ne va pas au-delà de la composante de rang 1. En effet, la régression PLS impose $r[C_s^x, C_t^x] = 0$ si $s \neq t$ ce qui n'est pas le cas en AFM.

Exemple des données de Cornell

Pour évaluer l'impact pratique de l'analogie entre PLS1 et l'AFM, ces deux méthodes ont été appliquées aux données de Cornell citées par Tenenhaus (1998 p. 78). Dans cet exemple, la première composante PLS et la première variable canonique (du groupe X) de l'AFM sont très proche : $r(C_1^x, F_1^x) = .9997$; plus généralement, le tableau 1 rassemble les corrélations entre les différentes prédictions (incluant Cp_1X, 1$^{\text{e}}$ composante principale de l'ACP de X).

TABLEAU 1

Données de Cornell : corrélation entre 4 prédictions. Régression usuelle (Y_{reg}) *et PLS* (C_1^x)*, AFM* (F_1^x) *et 1*^{ère} *composante principale de* X (Cp_1X)

	Y_{reg}	C_1^x	F_1^x	Cp_1X
Y_{reg}	1			
C_1^x (pls)	.9646	1		
F_1^x (afm)	.9589	.9997	1	
Cp_1X (acp)	.9506	.9981	.9993	1

Le tableau 2 donne quelques caractéristiques des différentes prédictions :

- $r(.,Y)$: corrélation entre la variable étudiée et la variable Y à prédire;
- $V_X = \Sigma_k r(.,v_k)^2$: variance de l'ensemble des variables de X expliquée par la variable étudiée;
- $\mathcal{L}_g(.,X)$: mesure de liaison entre X et la variable étudiée (cf. § 4.2).

TABLEAU 2

Données de Cornell : caractéristiques comparées de 4 prédictions : régression usuelle (Y_{reg}) et PLS (C_1^x), AFM (F_1^x) et 1ère composante principale de X (Cp_1X)

	$r(.,Y)$	V_X	$\mathcal{L}_g(.,X)$
Y_{reg}	.996	3.699	.919
C_1^x (pls)	.961	4.015	.997
F_1^x (afm)	.955	4.022	.999
Cp_1X (acp)	.947	4.026	1

Cet exemple illustre bien les résultats attendus. Sur les données «actives», en excluant la première composante principale qui sert seulement de référence :

- la régression usuelle fournit le meilleur ajustement et l'AFM le plus mauvais;
- la régression usuelle fournit la prédiction la moins liée aux variables explicatives et l'AFM la plus liée.

Dans ces données, Y_{reg} est très proche de la première composante principale de X. L'étroite corrélation entre C_1^x et F_1^x n'est donc pas très probante.

Exemple de données choisies

Un jeu de données choisies (tableau 3) met en évidence les différences entre les méthodes de façon flagrante.

TABLEAU 3

Données choisies pour la régression univariée.
A) Données. B) Régressions, AFM et ACP. C) Propriétés des colonnes de B)

A			B				C			
X1	X2	Y	Yreg	C_1^x	F_1^x	Cp_1X		r(.,Y)	V_X	$\mathcal{L}_g$
1,2	0,8	2,5	2	1,039	1,4292	1,39	Yreg	0,943	1,0000	.520
0,8	1,2	-0,5	0	0,959	1,3829	1,39	C_1^x (pls)	0,693	1,9201	.998
-0,8	-1,2	-0,5	0	-0,959	-1,3829	-1,39	F_1^x (afm)	0,678	1,9226	1.000
-1,2	-0,8	-1,5	-2	-1,039	-1,4292	-1,39	Cp_1X (acp)	0,667	1,9231	1

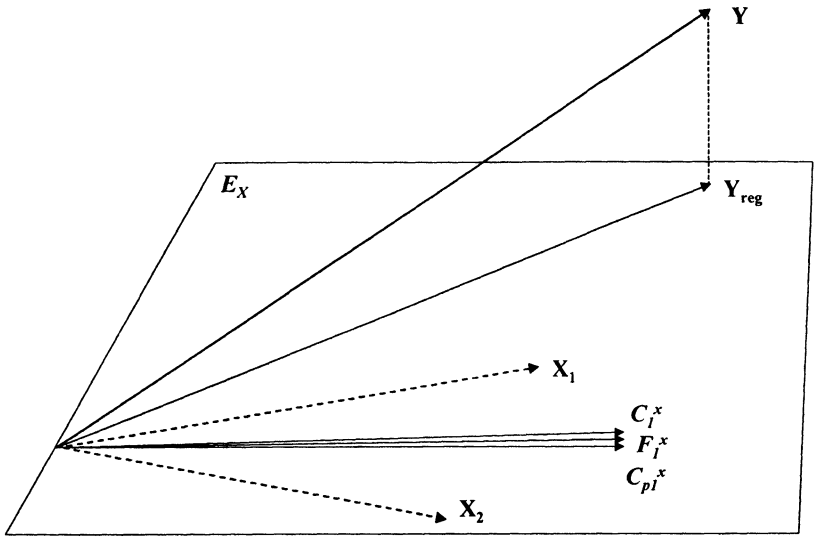


FIGURE 2

Données choisies pour la régression univariée (cf. tableau 3). Représentation des variables, respectant les angles et les longueurs (possible ici puisque l'espace des variables centrées est de dimension 3). E_X : sous-espace engendré par X_1 et X_2

L'ensemble des variables, initiales et calculées, est représentée figure 2.

On retrouve bien le résultat attendu, à savoir l'ordre Y_{reg}, C_1^x, F_1^x, Cp_1X décroissant du point de vue de la corrélation avec Y et croissant du point de vue de la variance expliquée de X.

L'étroite parenté entre la première composante PLS C_1^x et la première variable canonique de l'AFM F_1^x, est flagrante.

5.3. Cas général de deux groupes

Les deux premières composantes de la régression PLS multivariée (PLS2) vérifient :

$$C_1^x \propto (1/I)XX'C_1^y$$
$$C_1^y \propto (1/I)YY'C_1^x$$

où le terme à gauche du signe $\propto$ est égal au terme de droite centré-réduit. Ces équations assurent un compromis entre des composantes étroitement corrélées entre elles (cas des variables canoniques usuelles) et des composantes « représentant » bien leur groupe (cas des composantes principales de chacun des groupes).

L'AFM suit le même objectif dans le même esprit mais en procédant légèrement différemment :

$$F_1^x = \frac{1}{\lambda_1 \lambda_1^x} \frac{1}{I} X X' F_1$$
$$F_1^y = \frac{1}{\lambda_1 \lambda_1^y} \frac{1}{I} Y Y' F_1$$

L'analogie entre les deux méthodes est de même type que pour PLS1. Elle peut être illustrée en appliquant à la fois PLS2 et l'AFM aux données de Linnerud étudiées par Tenenhaus (1998, p. 142).

TABLEAU 4

Données de Linnerud : corrélations entre composantes PLS et variables canoniques de l'AFM

	C_1^x	C_1^y	F_1^x	F_1^y
C_1^x	1			
C_1^y	.554	1		
F_1^x	.998	.540	1	
F_1^y	.522	.996	.509	1

TABLEAU 5

Données de Linnerud : comparaison de prédictions
A : corrélations entre variables initiales (de Y) et composantes PLS ou variables canoniques de l'AFM.
B : propriétés des différentes prédictions. Pour Y_{reg}, il y a 3 prédictions puisque 3 régressions

	A			B					
	Y_1	Y_2	Y_3	V_x			$\mathcal{L}_g$		
Yreg	.583	.661	.232	1.412	1.628	1.653	.671	.774	.786
C_1^x	.486	.592	.203	2.084			.990		
F_1^x	.472	.579	.200	2.100			.998		
Cp,x	.460	.559	.195	2.104			1		

Les résultats obtenus illustrent les propriétés attendues :

- les composantes PLS sont très proches des variables canoniques de l'AFM;
- par rapport à la régression usuelle, les variables canoniques de l'AFM présentent les caractéristiques des composantes PLS de façon plus marquée : elles sont

moins corrélées aux variables à prédire (pour les données actives) et plus corrélées aux variables explicatives.

5.4. Conclusion sur les cas de deux groupes

Dans les cas étudiés, au niveau de la première composante, les résultats de l'AFM et de PLS sont très proches.

Plus dans le détail, la variables canonique F_1^x de l'AFM présente de façon amplifiée les caractéristiques de la première composante PLS. On peut se demander si, de ce fait, l'AFM ne peut pas être utilisée dans une optique de régression.

Ce n'est pas le cas. En pratique, la première composante PLS est souvent insuffisamment corrélée aux variables à prédire, ce à quoi on remédie en faisant intervenir une ou plusieurs des composantes PLS suivantes. Cette démarche est moins intéressante en AFM du fait de la non orthogonalité des variables canoniques entre elles.

6. Approche PLS et AFM : cas de J groupes de variables

6.1. L'approche PLS

Dans l'approche PLS, les données se composent, comme en AFM, de J groupes de variables définis sur les mêmes individus. Mais, en plus, un modèle relie ces groupes de la façon suivante :

- chaque groupe est représenté par une variable latente;
- les variables latentes sont reliées entre elles par un ensemble de relations linéaires.

Le cœur de l'algorithme recherche l'ensemble des variables latentes de façon telle que :

- chaque variable latente «représente» bien le groupe en ce sens qu'elle est corrélée aux variables de ce groupe; cette propriété est assurée, dans l'algorithme, par l'estimation externe de type Mode A;
- les variables latentes reliées par les équations du modèle sont corrélées entre elles; cette propriété est assurée, dans l'algorithme, par l'estimation interne.

Cette problématique est proche de celle de l'AFM, les variables canoniques jouant le rôle de variables latentes. La principale distinction entre les deux approches est que la notion de modèle n'existe pas en AFM : concrètement, en AFM, les variables canoniques (de même rang) sont corrélées entre elles autant que possible, sans privilégier certains couples d'entre elles.

6.2. Influence du modèle dans le calcul de variables latentes PLS

Avant de comparer les deux approches, il est utile d'examiner la dépendance des variables latentes PLS vis à vis du modèle adopté. A priori, on peut s'attendre à ce que les structures communes à tous les groupes apparaissent en tant que variables latentes quel que soit le modèle. Cette situation doit être en principe la plus fréquente : elle correspond au modèle postulé à l'origine (Wold H., 1975), dans lequel chaque groupe est unidimensionnel et bien représenté par une seule variable latente dont les variables brutes sont des expressions observables.

Ainsi, les données de Russett (Tenenhaus 1998 p 239) ont été soumises aux trois modèles dans lesquels un groupe est expliqué par les deux autres. Les corrélations entre les deux premières composantes PLS de chaque groupe pour les trois modèles sont données tableau 3. D'un modèle à l'autre, les composantes PLS homologues sont étroitement corrélées. Ainsi, le plus faible (en valeur absolue) coefficient de corrélation entre premières composantes homologues issues de modèles différents vaut .981 . Les deuxième composantes sont également stables d'un modèle à l'autre, à l'exception de celle du groupe 3 dans le modèle 2, sensiblement différente de son homologue dans les deux autres modèles. Cette instabilité est sans doute liée au caractère presque unidimensionnel du groupe 2 (91 % de son inertie est concentré dans une direction) qui joue un rôle central dans le modèle 2.

On remarquera au passage la non-corrélation entre composantes PLS d'un même groupe.

En revanche, s'il existe des structures fortes communes à certains groupes seulement, les composantes PLS peuvent varier considérablement selon le modèle comme l'illustre l'exemple de la figure 3.

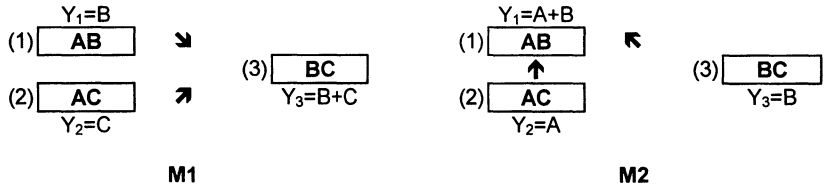


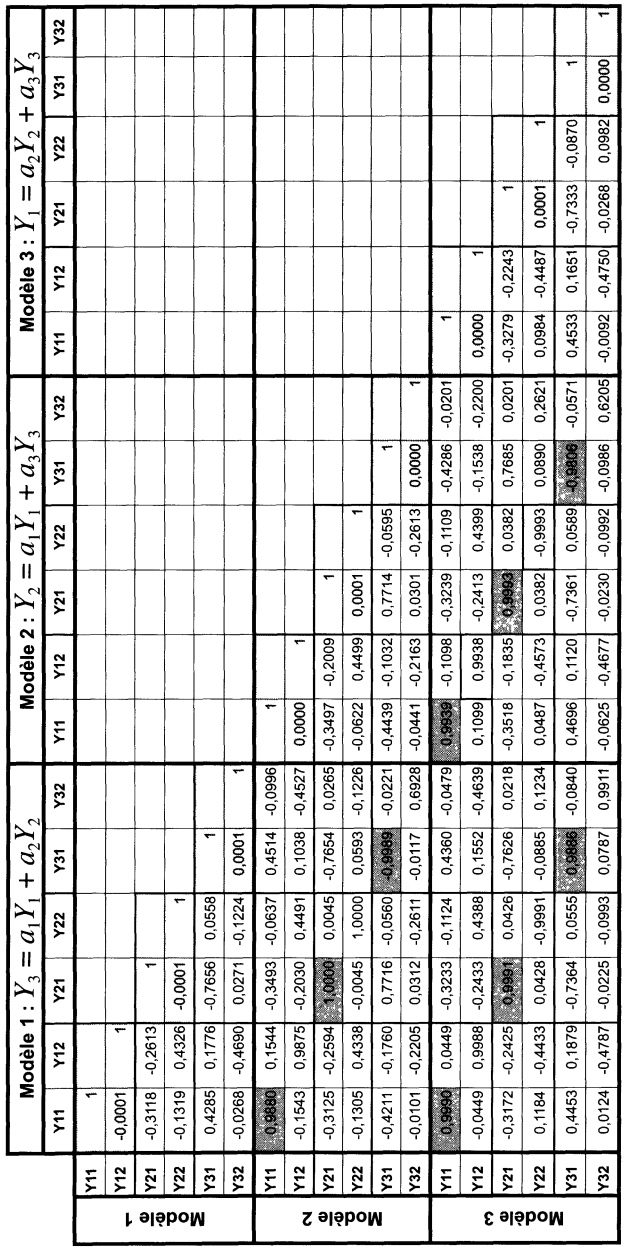
FIGURE 3

Deux modèles pour un même ensemble de données choisies. Chaque groupe est représenté par un rectangle contenant les variables du groupe (ex : le groupe 1 comprend les variables A et B); l'ensemble des groupes est bâti à partir de trois variables non corrélées A, B et C. Les liaisons de dépendance entre groupes sont matérialisées par des flèches; au-dessus de chaque rectangle figure la 1^e composante PLS du groupe : ainsi, pour le modèle 1, la composante PLS du groupe 1 est $Y_1 = B$

Dans cet exemple, à partir de 3 variables non corrélées A, B et C, on constitue les groupes (1), (2) et (3). Le schéma indique les premières composantes PLS (Y_1, Y_2, Y_3) obtenues à partir des modèles M1 (réduit à l'équation : $Y_3 = a_1 Y_1 + a_2 Y_2$) et M2 (réduit à l'équation : $Y_1 = a_2 Y_2 + a_3 Y_3$) matérialisés par des flèches.

TABLEAU 6

Données de Russett : corrélations entre composantes PLS obtenues par différents modèles sur les mêmes données. Yjs s^e composante PLS du groupe j. Définition des groupes § 6.4.1



		Modèle 1 : $Y_3 = a_1Y_1 + a_2Y_2$					Modèle 2 : $Y_2 = a_1Y_1 + a_3Y_3$					Modèle 3 : $Y_1 = a_2Y_2 + a_3Y_3$				
		Y11	Y12	Y21	Y22	Y31	Y32	Y11	Y12	Y21	Y22	Y31	Y32	Y11	Y12	Y21	Y22	Y31	Y32
Modèle 1	Y11	1																	
	Y12	-0,0001	1																
	Y21	-0,3118	-0,2613	1															
	Y22	-0,1319	0,4326	-0,0001	1														
	Y31	0,4285	0,1776	-0,7656	0,0558	1													
	Y32	-0,0268	-0,4690	0,0271	-0,1224	0,0001	1												
Modèle 2	Y11	0,9880	0,1544	-0,3493	-0,0637	0,4514	-0,0996	1											
	Y12	-0,1543	0,9875	-0,2030	0,4491	0,1038	-0,4527	0,0000	1										
	Y21	-0,3125	-0,2594	1,0000	0,0045	-0,7654	0,0265	-0,3497	-0,2009	1									
	Y22	-0,1305	0,4338	-0,0045	1,0000	0,0593	-0,1226	-0,0622	0,4499	0,0001	1								
	Y31	-0,4211	-0,1760	0,7716	-0,0560	0,9989	-0,0221	-0,4439	-0,1032	0,7714	-0,0595	1							
	Y32	-0,0101	-0,2205	0,0312	-0,2611	-0,0117	0,6928	-0,0441	-0,2163	0,0301	-0,2613	0,0000	1						
Modèle 3	Y11	0,9890	0,0449	-0,3233	-0,1124	0,4360	-0,0479	0,9939	-0,1098	-0,3239	-0,1109	-0,4286	-0,0201	1					
	Y12	-0,0449	0,9988	-0,2433	0,4388	0,1552	-0,4639	0,1099	0,9938	-0,2413	0,4399	-0,1538	-0,2200	0,0000	1				
	Y21	-0,3172	-0,2425	0,9981	0,0426	-0,7626	0,0218	-0,3518	-0,1835	0,9993	0,0382	0,7685	0,0201	-0,3279	-0,2243	1			
	Y22	0,1184	-0,4433	0,0428	-0,9991	-0,0885	0,1234	0,0487	-0,4573	0,0382	-0,9993	0,0890	0,2621	0,0984	-0,4487	0,0001	1		
	Y31	0,4453	0,1879	-0,7364	0,0555	0,9886	-0,0840	0,4696	0,1120	-0,7361	0,0589	0,9886	-0,0571	0,4533	0,1651	-0,7333	-0,0870	1	
	Y32	0,0124	-0,4787	-0,0225	-0,0993	0,0787	0,9911	-0,0625	-0,4677	-0,0230	-0,0992	-0,0986	0,6205	-0,0092	-0,4750	-0,0268	0,0982	0,0000	1

Sur cet exemple, selon le modèle, les variables latentes varient considérablement. Cette variation est en accord (prévisible) avec le modèle. Ainsi, le modèle 1 privilégie les liaisons entre les groupes 1 et 3 d'une part et 2 et 3 d'autre part : la variable A, commune aux seuls groupes 1 et 2 n'apparaît pas dans les variables latentes.

En revanche, le calcul des variables latentes ne dépend pas de l'orientation des liaisons dans le modèle (sauf lorsque l'estimation interne est réalisée selon le schéma structurel).

6.3. Comparaisons entre variables latentes PLS et variables canoniques de l'AFM

6.3.1. Cas où il existe une structure forte commune à tous les groupes

Dans ce cas, on peut s'attendre, indépendamment du modèle, à une grande similitude entre les variables latentes PLS et les variables canoniques de l'AFM. Ceci est illustré (*cf.* tableau 7) par le traitement des données de Russett en utilisant le modèle 1 du tableau 6.

TABLEAU 7

Données de Russett. Corrélations entre composantes PLS, variables canoniques de l'AFM et premières composantes principales. (Var : variance des variables du groupe qu'elle représente expliquée par la composante étudiée). Lg : liaison, cf. § 4.2, entre la composante et le groupe qu'elle représente

	PLS			AFM			ACP				
	C_1^1	C_1^2	C_1^3	F_1^1	F_1^2	F_1^3	Cp1	Cp2	Cp3	Var	$\mathcal{L}_g$
C_1^1	1									2,192	.973
C_1^2	-0,312	1								1,815	1.000
C_1^3	0,428	-0,766	1							2,712	.997
F_1^1	0,999	-0,299	0,420	1						2,216	.984
F_1^2	0,313	-1,000	0,765	0,301	1					1,815	1.000
F_1^3	0,430	-0,760	1,000	0,421	0,759	1				2,718	.999
Cp1	0,981	-0,254	0,385	0,989	0,256	0,387	1			2,252	1
Cp2	-0,313	1,000	-0,765	-0,301	-1,000	-0,760	-0,256	1		1,815	1
Cp3	0,432	-0,750	0,994	0,423	0,750	0,997	0,386	-0,750	1	2,721	1

Comme attendu :

- les composantes PLS et les variables canoniques de l'AFM sont très proches entre elles (mais aussi, proches des premières composantes principales);
- les composantes PLS extraient légèrement moins d'inertie que les variables canoniques de l'AFM.

- La composante PLS du groupe 3 est plus corrélée aux composantes PLS des groupes 1 et 2 que ne le sont les composantes obtenues par AFM ou ACP.

Les modèles finaux s'écrivent, les composantes PLS et les variables canoniques de l'AFM étant normées :

$$\begin{aligned} &\text{PLS} : C_1^3 = .2101C_1^1 - .7001C_1^2 && R^2 = .6261 \\ &\text{AFM} : F_1^3 = .2118F_1^1 - .6958F_1^2 && R^2 = .6177 \end{aligned}$$

Les deux méthodes conduisent ici pratiquement au même modèle.

6.3.2. Cas où il existe des structures fortes communes à certains groupes seulement

Dans ce cas, les résultats différent grandement d'une méthode à l'autre comme le montrent les traitements des données de la figure 3 (modèle M1) dont les résultats sont rassemblés tableau 8.

TABLEAU 8

Données choisies (cf. Figure 3, Modèle M1) : composantes PLS et variables canoniques issues de l'AFM.
Exemple : la $1^{ère}$ composante PLS du groupe 2 est confondue avec la variable C.
O : variable constante égale à 0

	PLS		AFM		
	C_1	C_2	F_1	F_2	F_3
Groupe 1	B	A	B	O	A
Groupe 2	C	A	O	C	A
Groupe 3	B+C	B-C	B	C	O

L'AFM fournit ici les trois facteurs communs aux couples de groupes et restitue bien la façon dont les données ont été construites. Remarquons que ici, les variables étant normées, l'ordre des facteurs est arbitraire.

L'approche PLS est, quant à elle, focalisée sur le groupe 3. Elle met en évidence une combinaison des variables du groupe 3 à la fois liée au groupe 1 et au groupe 2.

Ici, aucune méthode n'est meilleure que l'autre. Ces résultats illustrent les points de vue des méthodes qui sont, dans cet exemple, bien complémentaires.

Remarque : des résultats très proches ont été obtenus en bruitant légèrement ces données afin de se prémunir contre d'éventuels problèmes de convergence du programme LVPLS.

6.3.3. Cas de coïncidence entre les deux méthodes

Lohmöller (1989; démonstration reprise dans Tenenhaus (1999)) a montré que, en utilisant le mode A pour l'estimation externe, le schéma structurel pour l'estimation interne et un modèle dans lequel chaque groupe de variables explique un groupe

artificiel rassemblant toutes les variables, les fragments de la première composante principale du tableau complet vérifient les équations de stationnarité des variables latentes. La convergence de l'algorithme de l'approche PLS vers cette solution n'est pas démontrée mais a été constatée en pratique. Nous testons ici cette convergence dans le cas de l'AFM.

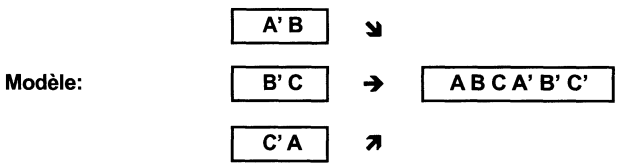
Le cœur de l'AFM pouvant être vu comme une ACP pondérée, et compte tenu de la remarque du § 4.4, l'approche PLS dans les conditions du § précédent doit conduire aux variables canoniques de rang 1 de l'AFM (à condition de transformer les données de façon telle que leur ACP, ici non normée, ait pour première valeur propre 1).

Nous avons constaté ce résultat dans le cas des données de Russett et l'avons testé sur les données du tableau 9.

TABLEAU 9

Données choisies pour tester le cas de coïncidence exacte entre l'approche PLS et l'AFM

	A	B	C	A'	B'	C'
	1	1	1	1,1	1,1	1,1
Données:	1	-1	-1	0,9	-1,1	-0,9
	-1	1	-1	-0,9	0,9	-1,1
	-1	-1	1	-1,1	-0,9	0,9



Les variables A, B et C sont centrées-réduites et non corrélées. Dans un premier temps, $A = A'$; $B = B'$; $C = C'$. Confronté à ces données particulières, l'algorithme du programme LVPLS ne converge pas. Dans un deuxième temps, $A' \approx A$; $B' \approx B$; $C' \approx C$. La coïncidence prévue par Lohmöller est à nouveau constatée (le coefficient de corrélation entre variables homologues entre les deux méthodes est toujours supérieur à .999999) sur ces données difficiles car sans facteur commun à tous les groupes.

6.4. Représentations associées aux composantes PLS et aux variables canoniques de l'AFM

6.4.1. Représentation des variables

Les composantes PLS successives d'un même groupe sont orthogonales. Il est donc possible de représenter l'ensemble des variables (tous groupes confondus) en projection sur un couple de composantes (généralement les deux premières) d'un même groupe de la façon usuelle :

- la coordonnée de la variable v_k sur la composante C_s^j vaut $r(v_k, C_s^j)$;

- l'ensemble des points représentant les variables s'interprète comme une projection du nuage des variables définis à partir de l'ensemble des coordonnées.

Pour les deux premières composantes, par exemple, il y a donc autant de représentations que de groupes, comme en analyse canonique usuelle; chaque représentation privilégie le point de vue d'un groupe.

Ce type de représentation n'est pas possible en AFM du fait de la non-orthogonalité des variables canoniques. En revanche, les variables générales F_s, qui bénéficient de la propriété $F_s = \Sigma_j F_s^j$, sont orthogonales et fournissent un cadre «moyen» pour représenter les variables.

Exemple des données de Russett

L'ensemble des variables est représenté d'une part sur deux premières composantes PLS du groupe 3 (à expliquer) et d'autre part sur les deux premiers axes de l'AFM (figure 4).

Comme attendu, la représentation sur les deux premières composantes PLS du groupe 3 privilégie... les variables du groupe 3. De son côté, en partie du fait de la pondération, la représentation de l'AFM assure des qualités de représentation plus homogènes d'un groupe à un autre (cf. tableau 10).

TABLEAU 10

Données de Russett. Qualités de représentation des groupes de variables fournies par l'approche PLS (deux premières composantes du groupe 3) et l'AFM (deux premiers axes). La qualité est mesurée par le rapport inertie projetée / inertie totale. Dans la colonne tous groupes, ces rapports ne sont pas exactement comparables entre eux puisque les variables ont des poids différents selon les méthodes

	Groupe 1	Groupe 2	Groupe 3	Tous groupes
1e composante PLS groupe 3	.08	.25	.33	0.25
2 composantes PLS groupe 3	0.18	0.54	0.67	0.52
1er axe AFM	0.27	0.71	0.39	0.43
2 axes AFM	0.78	0.83	0.41	0.61

Liste des variables, par groupe

Groupe 1 : inégalité dans la répartition des terres. *Gini* : indice de Gini; *Farm* : plus grand % de fermiers possédant la moitié des terres; *Rent* : % de fermiers locataires de leurs terres;

Groupe 2 : développement industriel. *Gnpr* : PNB par habitant; *Labo* : % d'actifs travaillant dans l'agriculture;

Groupe 3 : instabilité politique. *Inst* : indice de stabilité politique variant entre 0 (très stable) et 17 (très instable); *Ecks* : indice d'Eckstein mesurant le nombre de conflits violents sur la période 1946-1961; *Deat* : nombre de tués lors de manifestations

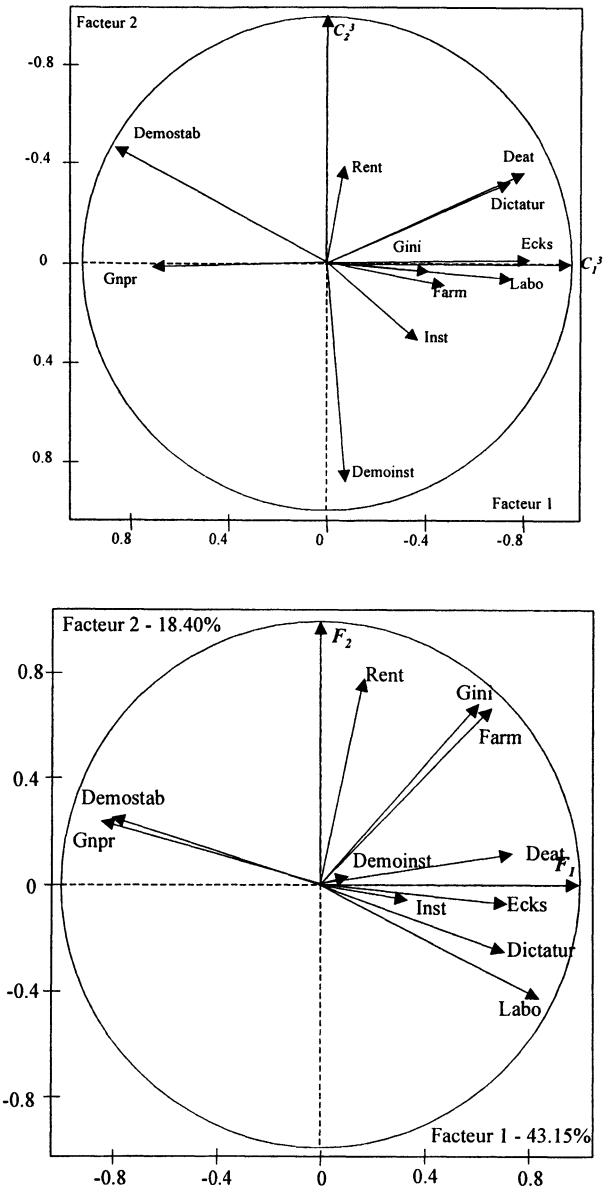


FIGURE 4

Données de Russett : représentation des variables.
En haut : sur les deux premières composantes PLS du groupe 3;
En bas : sur les deux premiers axes de l'AFM

violentes pendant la période 1950-1962 *Démo* : variable qualitative à 3 modalités : démocratie stable, démocratie instable, dictature. Cette variable est prise en compte au travers des indicatrices de ses modalités.

Ces deux représentations illustrent bien le point de vue de chacune des méthodes.

- Les composantes PLS du groupe 3 mettent bien en évidence les directions de ce groupe «explicables» par les variables des autres groupes. Ici, l'opposition *dictature* $\leftrightarrow$ *démocratie stable* est liée assez fortement aux variables de développement industriel, *i.e.* le produit national brut par habitant (*Gnpr*) et le % d'actifs travaillant dans l'agriculture (*Labo*), et très peu aux variables de répartition des terres. D'un autre côté, les démocraties instables ne sont pas caractérisées par les variables des autres groupes.
- Les axes de l'AFM mettent également en évidence le facteur commun entre les groupes 2 et 3 (de façon un peu plus symétrique, du point de vue de ces deux groupes, que PLS). La différence réside dans la bonne représentation du groupe 1 sur l'axe 2 (seuls *Gini* et *Farm* sont légèrement corrélées au facteur commun précité) impliquée par l'équilibre entre les groupes.

Ici encore, aucune représentation n'est en soi meilleure que l'autre : ce sont les points de vue qui différent.

Représentation complémentaire

Sur chacune de ces représentations de variables, on peut superposer une représentation des variables canoniques de l'AFM, des composantes PLS et des composantes principales des ACP séparées.

C'est ce qui est fait sur la figure 5 qui montre, par exemple pour le rang 1, l'étroite corrélation, pour chaque groupe, entre la variable canonique AFM, la composante PLS et la composante principale de l'ACP séparée.

6.4.2. Représentation des individus

Les représentations «naturelles» des individus dans le cadre de l'approche PLS consistent à croiser des variables latentes de même rang pour deux groupes différents. Ces variables ne sont pas orthogonales : elles sont corrélées par construction. Il est ainsi possible de mettre en évidence des individus présentant des valeurs «contradictoires» entre les variables des différents groupes.

Il n'y a pas, actuellement, d'autres représentations.

L'AFM, de son côté, fournit une représentation simultanée :

- des individus au travers de l'ensemble des données;
- des individus au travers des variables de chacun des groupes.

Cette représentation, qui comporte $J + 1$ points par individu (1 pour chaque groupe + 1 pour l'ensemble), est construite selon le principe suivant, qui comporte deux étapes.

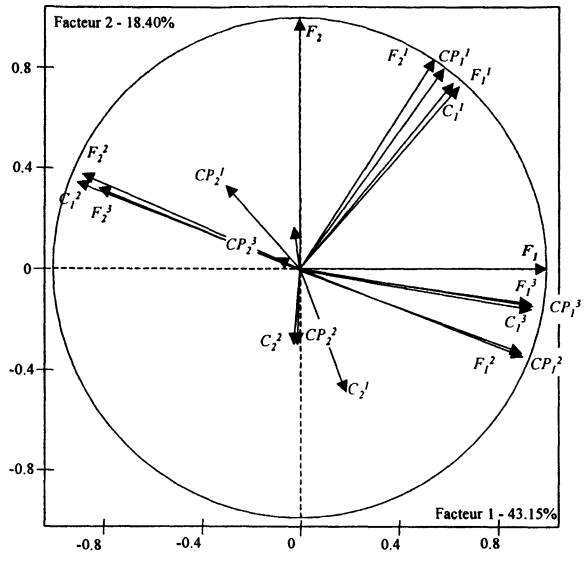


FIGURE 5

Données de Russett : représentation de variables calculées, sur les deux premiers axes de l'AFM. Pour le rang s et le groupe j : F_s^j variable canonique AFM, C_s^j composante PLS, CP_s^j composante principale de l'ACP séparée du groupe

1) Les variables générales F_s, étant des composantes principales, fournissent naturellement une représentation des individus. Celle-ci est reliée à la représentation des variables du § 6.4.1 par la relation de transition usuelle rappelée § 4.4.

Cette relation de l'analyse factorielle permet d'interpréter la représentation des individus en référence à celle des variables; elle justifie la représentation simultanée des individus et des variables utilisée quelquefois.

2) Les variables canoniques (multipliées par J) F_s^j peuvent être superposées à F_s. Cette superposition bénéficie des propriétés énoncées au § 4.4.

Exemple des données de Russett.

Exploitant les particularités de cet exemple, une bonne représentation des individus issue de l'approche PLS consiste à croiser les premières composantes des groupes 1 et 2 et à compléter les identificateurs des individus par un signe indiquant leur régime politique (Tenenhaus 1998 p. 244).

Ainsi, la figure 6 met-elle en évidence, par exemple :

- pour l'Australie et la Nouvelle Zélande une inégalité dans la répartition des terres grande compte tenu de leur fort développement industriel;
- pour la Pologne et la Yougoslavie, une faible inégalité dans la répartition des terres, compte tenu de leur plutôt faible développement industriel;
- pour l'Inde, un régime politique (démocratie stable) inhabituel compte tenu du très faible développement industriel.

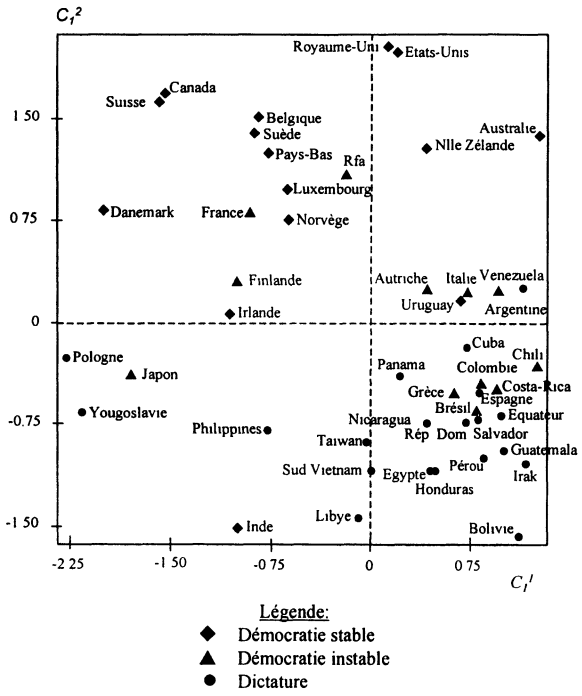


FIGURE 6

Données de Russett : représentation des individus par les premières composantes PLS des groupes 1 (inégalité dans la répartition des terres) et 2 (développement industriel)

Cette représentation est satisfaisante mais bénéficie du fait qu'il n'y a que 3 groupes et qu'une variable qualitative (le régime politique) résume clairement l'un d'entre eux.

De son côté, l'AFM fournit une première représentation des individus au travers des variables générales c'est à dire du point de vue des trois groupes de variables (*cf.* figure 7). Ainsi, les pays à droite du graphique constituent un ensemble relativement homogène du point de vue des 3 groupes. Les pays à gauche constituent en revanche un ensemble relativement homogène du point de vue des groupes 2 et 3 (1^{er} axe) et hétérogène du point de vue du groupe 1 (dont les variables sont liées à l'axe 2).

Ce faisant nous utilisons la dualité entre les figures 7 et 4.B. Par rapport à la figure 6, cette représentation présente donc l'avantage de bénéficier des propriétés de l'analyse factorielle.

A cette première représentation, l'AFM superpose une représentation des individus vus par un groupe seulement (individus dits partiels). Cette superposition est régie par les propriétés énoncées en § 4.4.

Sur la figure 8, nous limitons cette représentation à quelques individus remarquables. Le commentaire qui suit concerne uniquement les coordonnées sur le premier

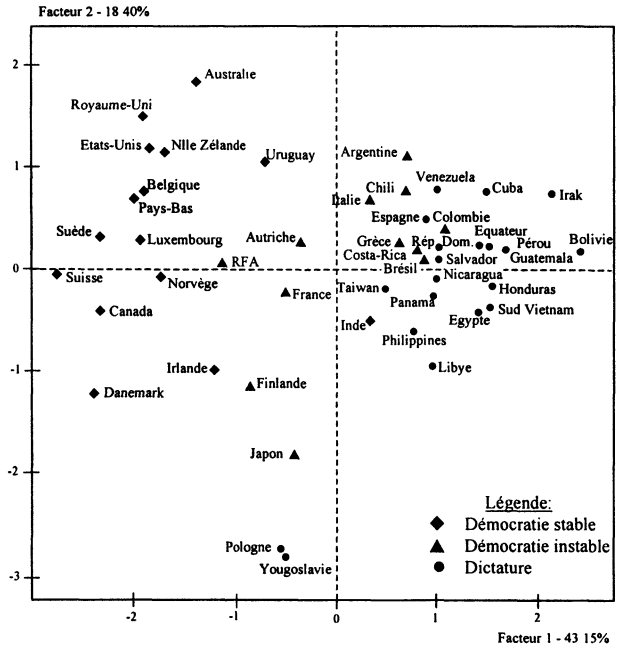


FIGURE 7

Données de Russett : représentation des individus sur le premier plan de l'AFM

facteur, seul à pouvoir être considéré comme commun aux 3 groupes. Il est étayé par le tableau 11 qui reprend quelques données.

Espagne. Ce pays est très homogène selon les 3 points de vue. Il est plutôt inégalitaire du point de vue de la répartition des terres (*Gini*, *Farm* et *Rent* élevés), un peu moins développé que la moyenne sur le plan industriel (*Gnpr* bas, *Labo* élevé) et sa qualification de dictature est tempérée par une relative stabilité des responsables exécutifs (*Inst*) et un faible nombre de tués (*Deat*). Ce type de pays illustre bien la notion de facteur commun.

Australie. Le régime de ce pays (démocratie stable) correspond bien à son développement industriel élevé. En revanche il est très inégalitaire dans la répartition des terres.

Pologne. Le régime de ce pays (dictature tempérée par une stabilité et un nombre de tués relativement bas) correspond à son développement industriel un peu moins élevé que la moyenne. En revanche, il est très égalitaire dans la répartition des terres.

Ces deux pays illustrent bien le lien étroit entre développement industriel et stabilité politique et le très faible lien entre ces deux groupes et l'inégalité dans la répartition des terres.

Inde. Malgré son très faible développement industriel (mais en accord avec une répartition des terres plutôt égalitaire), ce pays est plutôt du côté de la stabilité

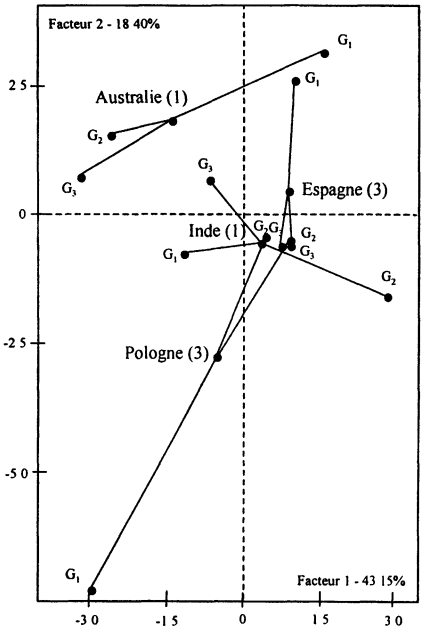


FIGURE 8

Données de Russett : représentation des individus partiels de trois pays sur le premier plan de l'AFM

politique. Sa qualification de démocratie stable est tempérée par un nombre élevé de conflits violents (*Ecks*) et plutôt élevé de tués (*Deat*). Variables

6.5. Conclusion sur les représentation graphiques

Les représentations de variables fournies par l'approche PLS et l'AFM sont très différentes et présentent chacune leur intérêt, lié à l'esprit de la méthode :

- celles de l'AFM représentent mieux l'ensemble des variables;
- celles de l'approche PLS, orientées par le groupe à expliquer, illustrent ce qui dans ce groupe est lié à des variables d'autres groupes.

Les représentations des individus différent également entre les deux méthodes. Le point de vue de l'AFM permet des représentations plus riches.

7. Conclusion

L'analyse simultanée de plusieurs groupes de variables définies sur les mêmes individus s'organise toujours autour de la notion de structures communes aux groupes de variables. Ces structures sont mises en évidence à l'aide de combinaisons linéaires de variables de chaque groupe, appelées variables canoniques au variables latentes.

TABLEAU 11

Données de Russett. Extrait des données, brutes (en haut) et transformées centrées réduites (en bas)

	Variables								
	Gini	*Farm*	*Rent*	*Gnpr*	*Labo*	*Inst*	*Ecks*	*Deat*	*Demo*
Espagne	78.0	99.5	43.7	254	50	0.0	22	1	3
Australie	92.9	99.6	28.9	1215	14	11.3	0	0	1
Pologne	45.0	77.7	0.0	468	57	8.5	19	5	3
Inde	52.2	86.9	53.0	72	71	3.0	83	14	1
moyenne	71.4	92.9	21.9	559.6	42.4	12.4	23.2	131.4	-

	Gini*	Farm*	Rent*	Gnpr*	Labo*	Inst*	Ecks*	Deat*	Demo*
Espagne	0.46	0.99	1.06	-0.47	0.51	-0.83	0.45	-0.53	-
Australie	1.51	1.00	0.64	1.29	-1.32	-0.79	-1.88	-0.83	-
Pologne	-1.85	-2.30	-2.93	0.22	0.70	-0.83	0.34	-0.06	-
Inde	-1.35	-0.91	1.26	-1.90	1.01	-0.83	1.40	0.34	-
moyenne	0	0	0	0	0	0	0	0	-

On peut chercher ces variables en se préoccupant uniquement de leurs corrélations entre elles; c'est le cas de l'analyse canonique. On peut aussi les chercher en se préoccupant en outre de la part de variance que chacune représente dans son groupe : c'est le cas de l'AFM et de l'approche PLS.

La différence profonde entre ces deux approches réside dans l'existence d'un modèle sous-jacent à l'approche PLS. Concrètement, ce modèle focalise la recherche de structures communes à certains couples de groupes seulement alors que l'AFM fait jouer le même rôle à tous les groupes. En pratique, cette distinction joue peu s'il existe des structures fortes communes à tous les groupes, mise en évidence naturellement par les deux méthodes, quel que soit le modèle choisi pour l'approche PLS. En revanche, s'il existe des structures fortes communes à certains groupes seulement, les deux méthodes sont susceptibles de donner des résultats très différents.

D'un point de vue plus technique, chacune des méthodes présente des propriétés spécifiques dont l'une en particulier concerne les utilisateurs :

- dans l'approche PLS, les variables canoniques d'un même groupe sont orthogonales entre elles;
- l'AFM propose une représentation superposée des individus vus par chacun des groupes de variables.

Par ailleurs, de même que l'AFM bénéficie de toutes les propriétés de la méthodologie factorielle (en particulier les relations de transition), l'approche PLS bénéficie des possibilités inhérentes à PLS (en particulier un traitement élégant des

données manquantes grâce à l'algorithme NIPALS, et des épreuves de validité – *cf.* Tenenhaus 1998 p. 61 et 138). Enfin, l'approche PLS repose sur un modèle dont on estime les paramètres, aspect totalement absent en AFM.

Il résulte, des convergences entre ces deux approches et de leurs spécificités, une véritable complémentarité qui peut être mise à profit pour définir des méthodologies d'analyse de tableaux multiples, conjuguant à la fois des aspects descriptifs (issus de l'AFM) et modélisant (issus de l'approche PLS).

Ainsi, par exemple, l'AFM peut être utilisée en amont de l'approche PLS, pour aider la construction de groupes de variables unidimensionnels (condition d'application originelle de l'approche PLS) et liés à un groupe à prédire. Pour cela, on définit les groupes de façon telle que, chacun contient des variables à la fois homogènes quant à leur signification et corrélées fortement à un même facteur de l'AFM. On construit alors un modèle PLS par facteur de l'AFM. Un exemple d'application se trouve dans Pagès et Tenenhaus (2001).

8. Remerciements

Il nous est agréable de remercier ici Cécile Lavanant, étudiante du DESS de l'Université de Rennes 2, qui, dans le cadre de son stage à l'ENSAR, a réalisé l'ensemble des traitements informatiques utilisés dans ce travail.

9. Bibliographie

CARROLL J.D. (1968), A generalization of canonical correlation analysis to three or more sets of variables, *Proc. 76th Conv. Amer. Psych. Assoc.*, pp. 227- 228.

ESCOFIER B. & PAGES J. (1983), Méthode pour l'analyse de plusieurs groupes de variables. Application à la caractérisation des vins rouges du Val de Loire. *Revue de Statistique Appliquée*, Vol. XXXI, n° 2, pp. 43-59.

ESCOFIER B. & PAGES J. (1994), Multiple Factor Analysis (AFMULT package). *Computational statistics & data analysis*, 18, pp. 121-140.

ESCOFIER B. & PAGES J. (1998), *Analyses factorielles simples et multiples; objectifs, méthodes et interprétation*. 3e édition révisée et enrichie, en particulier de deux nouveaux chapitres, 284 p., Dunod, Paris.

GUINOT C., TENENHAUS M., LATREILLE J. (1999), Approche PLS et Analyse de tableaux multiples. Application à l'étude des habitudes cosmétiques de 1012 femmes d'Ile de France. *Actes du Symposium PLS 99*, Cisia-Ceresta/Groupe HEC, Jouy-en-Josas.

HORST P. (1961), Relations among m sets of variables, *Psychometrika*, vol. 26, pp. 129-149.

HOTELLING H. (1936), Relations between two sets of variates, *Biometrika*, 28, pp. 321-377.

LOHMÖLLER J.-B. (1987), *LVPLS Program Manual, Version 1.8*, Zentralarchiv für Empirische Sozialforschung, Köln.

LOHMÖLLER J.-B. (1989), *Latent Variables Path Modeling with Partial Least Squares*, Physica-Verlag, Heidelberg.

PAGES J. & TENENHAUS M. (2001), Multiple Factor Analysis and PLS Approach. Application to the analysis of the relationship between physicochemical variables, sensory profiles and hédonic judgments. *Chemometrics and Intelligent Laboratory Systems*, 58, pp. 261-273.

SAPORTA G. (1975), *Liaisons entre plusieurs ensembles de variables et codage des données qualitatives*. Thèse de troisième cycle, Université Pierre et Marie Curie.

TENENHAUS M. (1998), *La régression PLS, Théorie et Pratique*, Technip, Paris.

TENENHAUS M. (1999), L'Approche PLS, *Revue de Statistique Appliquée*, vol. XLVII (2), pp. 5-40.

TENENHAUS M. & MORINEAU A. Eds (1999), Les méthodes PLS. Symposium international pls'99. Cisia. Paris

TUCKER L.R. (1958), An inter-battery method of factor analysis, *Psychometrika*, vol. 23, n° 2, pp. 111-136.

VALETTE-FLORENCE P. (1988A), Analyse structurelle comparative des composantes des systèmes de valeurs selon Kahle et Rokeach, *Recherche et Applications en Marketing*, vol. III, n° 1.

VALETTE-FLORENCE P. (1988B), Spécificité et apports des méthodes d'analyse multivariée de la deuxième génération, *Recherche et Applications en Marketing*, vol. III, n° 4.

VALETTE-FLORENCE P. (1990), Analyse structurelle et analyse typologique : illustration d'une démarche complémentaire, *Recherche et Applications en Marketing*, vol. V, n° 1.

VAN DEN WOLLENBERG A.L. (1977), Redundancy analysis : an alternative for canonical correlation, *Psychometrika*, vol. 42, pp. 207-219.

WOLD H. (1975), *Modeling in Complex Situations with Soft Information*, Third World Congress of Econometric Society, August 21-26, Toronto, Canada.

WOLD H. (1982), Soft Modeling : the basic design and some extensions, in *System under indirect observation*, vol. 2, K.G. Jöreskog & H. Wold (Eds), North-Holland, Amsterdam, pp. 1-54.

WOLD H. (1985), Partial Least Squares, in *Encyclopedia of Statistical Sciences*, vol.6, Kotz, S. & Johnson, N.L. (Eds), John Wiley & Sons, New York, pp. 581-591.