REVUE DE STATISTIQUE APPLIQUÉE

M. TENENHAUS

L'approche PLS

Revue de statistique appliquée, tome 47, n° 2 (1999), p. 5-40

<http://www.numdam.org/item?id=RSA_1999__47_2_5_0>

Rev. Statistique Appliquée, 1999, **XLVII** (2), 5-40

L'APPROCHE PLS

M. Tenenhaus

Groupe HEC (Jouy-en-Josas)

RÉSUMÉ

Nous allons étudier dans cet article l'approche PLS proposée par Herman Wold. Cette méthode permet d'analyser un ensemble de J blocs de variables. On suppose que chaque bloc est résumé par une variable latente et qu'il existe des relations structurelles entre les variables latentes. L'approche PLS permet d'estimer les variables latentes et les relations structurelles. L'approche PLS est à l'algorithme LISREL ce que l'analyse en composantes principales est à l'analyse factorielle en facteurs communs et spécifiques. L'approche PLS est une méthode très synthétique puisqu'elle contient comme cas particuliers l'analyse en composantes principales, l'analyse canonique, l'analyse des redondances, la régression PLS, l'analyse canonique généralisée aux sens de Horst et de Carroll, au niveau de la première composante. De plus l'approche PLS permet l'analyse de tableaux avec données manquantes en utilisant l'algorithme NIPALS et la régression PLS. Nous décrivons dans cet article les grandes lignes de l'algorithme NIPALS et de la régression PLS avec leurs principales propriétés. Nous présentons ensuite plus en détail l'approche PLS, ses liens avec la régression PLS et montrons qu'elle contient tous les cas particuliers mentionnés. Un exemple entièrement traité termine l'exposé de l'approche PLS.

Mots-clés : *Approche PLS, Modélisation douce, NIPALS, Régression PLS, Relations structurelles sur variables latentes, Analyse canonique généralisée.*

ABSTRACT

This paper is devoted to a presentation of the Partial Least Squares approach (Soft Modeling) proposed by Herman Wold and to its links with PLS regression. This method allows to analyse a set of J blocks of variables. We suppose that each block can be summarized by one latent variable and that it exists structural relationships between the latent variables. The PLS approach allows to estimate the latent variables and the structural relationships. The PLS approach can be compared to the LISREL algorithm as principal component analysis to factor analysis. The PLS approach is a very synthetic method since it contains as particular cases principal component analysis, canonical correlation analysis, redundancy analysis, PLS regression, and generalized canonical analysis according to Horst and Carroll, at the first component level. Furthermore, the PLS approach allows to analyse tables with missing data by using the NIPALS algorithm and PLS regression. In this paper, we give a general description of the NIPALS algorithm and of PLS regression with their main properties. Next, we present the PLS approach in detail, its links with PLS regression and we show that it contains all mentioned particular cases. A completely analyzed example ends the presentation of the PLS approach.

Keywords : *PLS approach, Soft modeling, NIPALS, PLS regression, Structural-equation model with latent variables, generalized canonical analysis.*

1. Introduction

Dans un cadre très général appelé *Partial Least Squares* (PLS), Herman et Svante Wold ont proposé des méthodes d'analyse des données permettant d'étudier J blocs de variables observées sur les mêmes individus.

La méthode NIPALS (*Nonlinear estimation by iterative Partial Least Squares*), proposée par Wold (1966), permet d'étudier un seul bloc de variables ($J = 1$). Elle conduit à l'analyse en composantes principales lorsque les données sont complètes, mais fonctionne également lorsqu'il y a des données manquantes.

La régression PLS permet de relier un bloc de variables à expliquer à un bloc de variables explicatives ($J = 2$). Elle a été proposée par Wold, Martens & Wold (1983). On obtient les composantes PLS par applications successives de l'analyse factorielle inter-batteries de Tucker (1958). L'utilisation des principes de l'algorithme NIPALS permet le traitement des données manquantes. Il peut y avoir beaucoup plus de variables que d'observations. La régression PLS est sans doute actuellement la meilleure réponse au problème de la multicolinéarité en régression multiple.

Le cas de J blocs a été étudié dans le cadre de la modélisation de relations structurelles sur variables latentes (*Path models with latent variables*). L'estimation de ces modèles peut être abordée de deux manières très différentes : l'approche maximum de vraisemblance ou l'approche PLS.

L'approche maximum de vraisemblance a été développée par Jöreskog (1970) à travers le logiciel LISREL (Jöreskog et Sörbom (1979,1984) et Hayduk (1987)). Cette approche est disponible dans le logiciel SAS (Proc CALIS) et dans le logiciel AMOS (Arbuckle, 1997) diffusé par SPSS.

L'approche PLS proposée par Wold (1975, 1982, 1985) est aussi décrite dans Lohmöller (1989) et Fornell & Cha (1994). L'approche PLS a été particulièrement développée en France par Valette-Florence (1988a,b, 1990) pour des applications en Marketing. L'approche PLS est disponible dans le programme LVPLS 1.8 de Lohmöller (1987). Ce programme et sa documentation sont diffusés gratuitement par J.J. McArdle (Department of Psychology, University of Virginia, Charlottesville, VA 22903, USA). Ils sont aussi accessibles par internet depuis le site de l'Université de Virginie.

Les approches LISREL et PLS ont été comparées dans Jöreskog et Wold (1982). Les différences entre l'analyse factorielle et l'analyse en composantes principales se retrouvent entre ces deux approches.

L'approche maximum de vraisemblance repose sur des hypothèses de multinormalité et permet une modélisation de la matrice des covariances entre les variables observées. Il peut y avoir des problèmes d'identification et non convergence de l'algorithme. Les variables latentes ne sont pas estimées au niveau des individus.

Par contraste, l'approche PLS est d'une grande simplicité. Il y a peu d'hypothèses probabilistes. On modélise directement les données à l'aide d'une succession de régressions simples ou multiples. Il n'y a aucun problème d'identification et les variables latentes sont estimées au niveau des individus. L'approche NIPALS et la régression PLS permettent le traitement des données manquantes. L'approche PLS, appelée aussi modélisation douce (*soft modeling*) par Herman Wold, corres-

pond tout à fait à l'esprit de l'analyse des données. Elle contient d'ailleurs comme cas particuliers l'analyse en composantes principales, l'analyse canonique, l'analyse des redondances, la régression PLS, l'analyse canonique généralisée de Horst et l'analyse canonique généralisée de Carroll, au niveau de la première composante. Malgré toutes ces qualités l'approche PLS semble avoir atteint un niveau de diffusion voisin de celui de l'analyse des correspondances dans les années soixante-dix.

Nous décrivons dans cet article les grandes lignes de l'algorithme NIPALS et de la régression PLS avec leurs principales propriétés. Nous présentons ensuite plus en détail l'approche PLS et ses liens avec la régression PLS. Nous montrons qu'elle contient tous les cas particuliers mentionnés. Un exemple entièrement traité termine enfin cet exposé de l'approche PLS.

2. La méthode NIPALS

L'algorithme NIPALS (Wold, 1966) permet de réaliser une analyse en composantes principales d'un tableau individus × variables X avec données manquantes, sans avoir à supprimer les individus à données manquantes ni à estimer les données manquantes.

Décrivons les principes de cet algorithme.

La formule de décomposition de l'analyse en composantes principales d'un tableau X formé de variables centrées s'écrit

$$X = \sum_{h=1}^{a} t_h p_h' \tag{1}$$

où a est le rang de la matrice X, t_h la h-ième composante principale et p_h le h-ième axe factoriel. Inversement on peut considérer la formule (1) comme un modèle et chercher à «estimer» les vecteurs t_h et p_h. On retrouve les composantes principales et les axes factoriels en imposant au modèle (1) des contraintes d'orthogonalité sur les vecteurs t_h et d'orthonormalité sur les vecteurs p_h.

Nous notons dans cette section $\mathbf{x}_j$ la j-ième colonne du tableau X et x_i le vecteur-colonne obtenu en transposant la i-ième ligne du tableau X. Chaque valeur t_{hi} de la composante principale t_h pour l'individu i représente également le coefficient de régression de la régression simple sans constante de x_i sur p_h. De même, chaque coordonnée p_{hj} du vecteur p_h représente le coefficient de régression de la régression simple sans constante de $\mathbf{x}_j$ sur t_h.

Décrivons l'étape courante de l'algorithme NIPALS. On part d'un vecteur arbitraire $t_h^{(1)}$ orthogonal aux vecteurs précédents $t_1, ..., t_{h-1}$ en choisissant une colonne de la matrice $X - \sum_{i=1}^{h-1} t_i p_i'$ représentant les résidus de la régression de X sur $t_1, ..., t_{h-1}$. On obtient ensuite un vecteur $p_h^{(1)}$ en normant le vecteur formé des coefficients de régression de la régression de X sur $t_h^{(1)}$. On obtient un nouveau vecteur

$t_h^{(2)}$ en régressant chaque x_i sur $p_h^{(1)}$. Cette procédure est itérée jusqu'à convergence des vecteurs $t_h^{(k)}$ vers la solution t_h.

Lorsqu'il n'y a pas de données manquantes, cet algorithme conduit à un vecteur t_h égal à la h-ième composante principale du tableau X. Mais lorsqu'il y a des données manquantes, il reste possible d'appliquer les différentes phases de l'algorithme présenté sur les données disponibles. On obtient alors des vecteurs t_h et p_h représentant des «estimations» de la h-ième composante principale et du h-ième axe factoriel du tableau X sans données manquantes. De plus la formule de reconstitution (1) permet au final d'estimer les valeurs de ces données manquantes. On trouvera une description plus détaillée de cet algorithme et un exemple d'application dans Tenenhaus (1998). Par ailleurs l'algorithme NIPALS est disponible dans les logiciels SIMCA-P (Umetri, 1996) et The Unscrambler (Camo, 1996).

3. La régression PLS

La régression PLS (Wold, Martens & Wold, 1983) permet de modéliser la liaison entre un bloc de variables Y et un bloc de variables X. Cette méthode consiste à rechercher dans un premier temps des composantes orthogonales t_h, combinaisons linéaires des variables X, expliquant au mieux à la fois les X et les Y. Les équations de régression PLS sont ensuite obtenues en régressant chaque variable Y sur les composantes t_h, puis en exprimant ces régressions en fonction des variables X d'origine. L'algorithme d'origine est présenté sous la forme d'un algorithme permettant d'intégrer les données manquantes selon les principes de NIPALS. Nous décrivons cet algorithme en détail dans Tenenhaus (1998). Lorsqu'il n'y a pas de données manquantes, les composantes PLS t_h peuvent être obtenues comme solutions successives de problèmes d'optimisation. Martens & Næs (1989) distinguent la régression PLS1 (une variable Y à expliquer) de la régression PLS2 (plusieurs variables Y à expliquer). Nous allons décrire dans cette section les critères permettant d'obtenir les composantes PLS dans ces deux situations, lorsqu'il n'y a pas de données manquantes, et exposer les principales propriétés mathématiques de ces méthodes.

3.1 La régression PLS1

Le tableau Y est formé d'une seule variable notée y. Le tableau X est formé de p colonnes notées dans cette section x_j. Toutes ces variables sont supposées centrées-réduites. Les composantes PLS orthogonales t_h peuvent s'écrire en fonction de X ($t_h = Xw_h^*$), mais sont en fait calculées en les écrivant en fonction du résidu X_{h-1} de la régression de X sur $t_1, ..., t_{h-1}$ ($t_h = X_{h-1}w_h$), où l'on a posé $X_0 = X$.

On recherche à chaque étape h le vecteur normé w_h maximisant le critère

$$\text{cov}(y, X_{h-1}w_h) \tag{2}$$

Le vecteur normé w_h maximisant le critère (2) est donné par la formule

$$w_h = \frac{X'_{h-1}y}{\|X'_{h-1}y\|}$$

On déduit ensuite la matrice $W_h^* = [w_1^*, ..., w_h^*]$ des matrices $W_h = [w_1, ..., w_h]$ et $P_h = [p_1, ..., p_h]$, où $p_h = X't_h/t'_ht_h$, à l'aide de la formule

$$W_h^* = W_h(P'_hW_h)^{-1} \tag{3}$$

On trouvera la justification de cette formule dans Tenenhaus (1998).

L'équation de régression PLS de y sur X, obtenue en utilisant les h premières composantes PLS $t_1, ..., t_h$, est calculée en régressant y sur les composantes $t_1, ..., t_h$, puis en exprimant cette régression en fonction de X. L'équation de régression de y sur les composantes $t_1, ..., t_h$ s'écrit

$$y \approx c_1t_1 + ... + c_ht_h \tag{4}$$

où $c_l = y't_l/t'_lt_l$. Posons $C_h = [c_1, ..., c_h]$. On obtient l'expression de l'équation de régression PLS (4) en fonction de X :

$$y \approx X[w_1^*c_1 + ... + w_h^*c_h] \tag{5}$$

$$\approx XW_h^*C'_h \tag{6}$$

Notons $\mathbf{b}_h = W_h^*C'_h$ le vecteur des coefficients de régression PLS b_j. On déduit de (6) la formule de régression PLS reliant y aux variables $x_1, ..., x_p$

$$y \approx \sum_{j=1}^{p} b_jx_j \tag{7}$$

$$\approx \sum_{j=1}^{p}(\sum_{l=1}^{h} w_{lj}^*c_l)x_j \tag{8}$$

Le coefficient de régression b_j représente le produit scalaire entre la j-ième ligne de la matrice W_h^* et le vecteur-ligne C_h. Ce résultat permet de justifier les cartes des variables habituellement construites en régression PLS à l'aide des vecteurs w_l^* et c_l. Dans le cas où deux composantes PLS suffisent à expliquer y à l'aide des X ($h = 2$), les variables x_j et y sont représentées dans un plan par les points (w_{1j}^*, w_{2j}^*) et (c_1, c_2). Le produit scalaire de ces deux points représente le coefficient de régression b_j. Par conséquent les coefficients de régression b_j sont positifs, nuls ou négatifs selon que les vecteurs du plan représentant les variables x_j et y forment un angle aigu, droit ou obtus.

Une présentation plus directe de la régression PLS1 a été proposée par de Jong (1993a) dans le cadre de son algorithme SIMPLS décrit plus loin dans le paragraphe 3.2. Il montre que les composantes PLS $t_h = Xw_h^*$ sont obtenues, à une normalisation près, en cherchant, pour des valeurs successives de l'indice h, à maximiser le critère

$$\text{cov}(y, Xw_h^*) \tag{9}$$

sous les contraintes

(i) le vecteur w_h^* est normé,

(ii) la composante $t_h = Xw_h^*$ est orthogonale aux composantes $t_1, ..., t_{h-1}$.

La logique de la régression PLS1 devient très claire lorsqu'on réécrit le critère (9) sous la forme

$$\text{cov}(y, Xw_h^*) = \text{cor}(y, Xw_h^*) \times \sqrt{\text{var}(y)} \times \sqrt{\text{var}(Xw_h^*)} \tag{10}$$

Pour $h = 1$ la régression PLS1 apparaît comme un compromis entre la régression multiple de y sur X ($\text{cor}(y, Xw_1^*)$ maximum) et la recherche de la première composante principale de X ($\text{var}(Xw_1^*)$ maximum). Pour l'indice général h on recherche une nouvelle composante $t_h = Xw_h^*$ orthogonale aux composantes précédentes $t_1, ..., t_{h-1}$ et aussi explicative que possible de y ($\text{cor}(y, Xw_h^*)$ maximum) et de X ($\text{var}(Xw_h^*)$ maximum).

Suivant de Jong (1993a), on obtient très simplement les vecteurs w_h^* :

Pour $h = 1$:

$$w_1^* = \frac{X'y}{\|X'y\|} \tag{11}$$

À une normalisation près, les coordonnées du vecteur w_1^* représentent donc les corrélations entre les variables x_j et y

Pour $h > 1$:

On calcule tout d'abord le résidu $\widetilde{X}'_{h-1}$ de la régression de X' sur les vecteurs $X't_1, ..., X't_{h-1}$. Puis on obtient

$$w_h^* = \frac{\widetilde{X}'_{h-1}y}{\left\|\widetilde{X}'_{h-1}y\right\|} \tag{12}$$

Indiquons deux propriétés intéressantes de la régression PLS1 découvertes par de Jong (1993b, 1995).

Dans l'article intitulé «*PLS fits closer than PCR*», il montre que la corrélation multiple entre y et les h premières composantes PLS est nécessairement supérieure ou égale à la corrélation multiple entre y et les h premières composantes principales

du tableau X. Ce résultat est évident pour $h = 1$, mais ne l'est plus pour $h > 1$ et de Jong en donne alors la démonstration.

Dans l'article «*PLS shrinks*» de Jong étudie l'évolution du vecteur $\mathbf{b}_h$ des coefficients des variables X dans l'équation de régression PLS construite à partir des h premières composantes PLS. Il montre que la norme du vecteur $\mathbf{b}_h$ croît avec l'indice h et que, pour $h = a = \text{rang}(X)$,

$$\mathbf{b}_a = (X'X)^+ X'y \tag{13}$$

où $(X'X)^+$ est l'inverse généralisé de Moore-Penrose de la matrice $X'X$. Autrement dit, lorsqu'on utilise toutes les composantes PLS disponibles, la régression PLS fournit la solution des équations normales de norme minimum.

3.2 La régression PLS2

Le tableau Y est maintenant formé de q colonnes notées y_k et supposées centrées-réduites. Le tableau X est formé de p colonnes x_j supposées centrées-réduites. Comme en régression PLS1, les composantes PLS orthogonales t_h peuvent s'écrire en fonction de X ($t_h = Xw_h^*$), mais sont en fait calculées en les écrivant en fonction du résidu X_{h-1} de la régression de X sur $t_1, ..., t_{h-1}$ ($t_h = X_{h-1}w_h$).

On recherche à chaque étape h le vecteur normé w_h maximisant le critère

$$\sum_{k=1}^{q} \text{cov}^2(y_k, X_{h-1}w_h) \tag{14}$$

Le vecteur w_h est obtenu par analyse factorielle inter-batteries (Tucker, 1958) des tableaux X_{h-1} et Y : c'est le vecteur propre de la matrice $X'_{h-1}YY'X_{h-1}$ associé à la plus grande valeur propre.

On peut ensuite calculer les vecteurs w_h^* à l'aide de la formule (3).

L'équation de régression PLS de Y sur X, obtenue en utilisant les h premières composantes PLS $t_1, ..., t_h$, est calculée en régressant Y sur les composantes $t_1, ..., t_h$, puis en exprimant cette régression en fonction de X. L'équation de régression de Y sur les composantes $t_1, ..., t_h$ s'écrit maintenant

$$Y \approx t_1c'_1 + ... + t_hc'_h \tag{15}$$

où $c_l = Y't_l/t'_lt_l$. Posons $C_h = [c_1, ..., c_h]$. On obtient l'expression de l'équation de régression PLS (15) en fonction de X :

$$Y \approx X[w_1^*c'_1 + ... + w_h^*c'_h] \tag{16}$$

$$\approx XW_h^*C'_h \tag{17}$$

Notons $B = W_h^* C_h'$ la matrice des coefficients de régression PLS b_{jk}. On déduit de (17) la formule de régression PLS reliant y_k aux variables $x_1, ..., x_p$

$$y_k \approx \sum_{j=1}^{p} b_{jk} x_j \tag{18}$$

$$\approx \sum_{j=1}^{p} (\sum_{l=1}^{h} w_{lj}^* c_{lk}) x_j \tag{19}$$

Le coefficient de régression b_{jk} représente le produit scalaire entre la j-ième ligne de la matrice W_h^* et la k-ième ligne de la matrice C_h. Ce résultat justifie, comme en régression PLS1, les cartes des variables construites à l'aide des vecteurs w_l^* et c_l. Dans le cas où deux composantes PLS suffisent à expliquer Y à l'aide des X ($h = 2$), les variables x_j et y_k sont représentées dans un plan par les points (w_{1j}^*, w_{2j}^*) et (c_{1k}, c_{2k}). Le produit scalaire de ces deux points représente le coefficient de régression b_{jk}. Les coefficients de régression b_{jk} sont positifs, nuls ou négatifs selon que les vecteurs représentant les variables x_j et y_k forment un angle aigu, droit ou obtus.

Un algorithme plus direct, SIMPLS (*Straightforward Implementation of a statistically inspired Modification of the* ***PLS*** *method*) a été proposé par de Jong (1993a). On recherche successivement des composantes $t_h = Xa_h$ maximisant le critère

$$\sum_{k=1}^{q} \text{cov}^2(y_k, Xa_h) \tag{20}$$

sous les contraintes

(i) le vecteur a_h est normé,

(ii) la composante $t_h = Xa_h$ est orthogonale aux composantes $t_1, ..., t_{h-1}$.

La décomposition de (20) en

$$\sum_{k=1}^{q} \text{cor}^2(y_k, Xa_h) \times \text{var}(Xa_h) \tag{21}$$

montre que l'algorithme SIMPLS réalise un compromis entre une analyse des redondances de Y par rapport à X ($\sum_{k=1}^{q} \text{cor}^2(y_k, Xa_h)$ maximum) et une analyse en composantes principales de X (var(Xa_h) maximum).

On obtient les vecteurs a_h comme suit :

Pour $h = 1$:

On retrouve que le vecteur a_1 est vecteur propre de la matrice $X'YY'X$ associé à la plus grande valeur propre.

Pour $h > 1$:

On calcule tout d'abord le résidu $\widetilde{X}'_{h-1}$ de la régression de X' sur les vecteurs $X't_1, ..., X't_{h-1}$. Puis on obtient a_h comme vecteur propre de la matrice $\widetilde{X}'_{h-1}YY'\widetilde{X}_{h-1}$ associé à la plus grande valeur propre.

L'algorithme SIMPLS donne systématiquement des résultats très voisins de la régression PLS2. Cet algorithme est disponible dans la Proc PLS du logiciel SAS (version 6.12). D'un point de vue pratique, la régression PLS reste cependant compétitive car elle permet l'analyse de tableaux *avec données manquantes*. Elle est disponible dans les logiciels SIMCA-P et The Unscrambler.

4. L'approche PLS

L'objectif de cet article est de montrer les nombreux intérêts théoriques et pratiques de l'approche PLS. Nous allons présenter l'algorithme PLS de base (Wold, 1982), en exploitant systématiquement le point de vue géométrique de Bookstein (1982). Nous indiquerons les extensions de l'algorithme proposées par Lohmöller (1989). Enfin nous comparerons sur un exemple l'utilisation des logiciels SIMCA-P et LVPLS 1.8.

4.1 Les données, les hypothèses et le modèle

Les données sont formées de J blocs de variables $X_j = \{x_{j1}, ..., x_{jk_j}\}$ observées sur n individus. Les variables x_{jh} sont appelées «variables manifestes» et supposées centrées-réduites. On conservera la notation X_j pour la matrice des données observées du j-ième groupe.

Relation entre les variables manifestes et les variables latentes

Chaque groupe de variables constitue l'expression observable d'une «variable latente» ξ_j centrée-réduite.

On distingue deux manières de relier les variables manifestes x_{jh} d'un bloc à leur variable latente ξ_j.

Un bloc est formé de variables manifestes x_{jh} *réflectives* (en anglais : *reflective (outward) model*) si l'on considère qu'elles reflètent la variable latente ξ_j, qu'elles en sont une conséquence, qu'elles sont créées par celle-ci. Elles sont reliées à ξ_j par l'équation linéaire

$$x_{jh} = \lambda_{jh}\xi_j + \varepsilon_{jh} \tag{22}$$

où ε_{jh} est un terme aléatoire de moyenne nulle et non corrélé à la variable latente ξ_j.

Un bloc est formé de variables manifestes x_{jh} *formatives* (en anglais : *formative (inward) model*) si c'est la variable latente ξ_j qui est créée par l'ensemble des variables du bloc, qu'elle en est une conséquence. La variable latente ξ_j vérifie l'équation

linéaire

$$\xi_j = \sum_h \pi_{jh} x_{jh} + \delta_j \tag{23}$$

où δ_j est un terme aléatoire de moyenne nulle et non corrélé aux variables manifestes x_{jh}.

Il est utile (mais non indispensable) de préciser les signes d_{jh} des coefficients λ_{jh} ou π_{jh}.

Relation entre les variables latentes

Le phénomène étudié est décrit par des relations structurelles entre les variables latentes. On distingue les variables latentes endogènes (elles sont expliquées par d'autres variables latentes) des variables latentes exogènes (elles sont toujours explicatives). Il est d'usage de noter η_j les variables latentes endogènes et ξ_i les variables latentes exogènes. Par soucis de simplicité au niveau des notations, nous préférons noter ξ_i toutes les variables latentes. Les relations structurelles entre les variables latentes sont de la forme

$$\xi_j = \sum_i \beta_{ji} \xi_i + \zeta_j \tag{24}$$

où ζ_j est un terme aléatoire de moyenne nulle et non corrélé aux variables latentes explicatives ξ_i apparaissant dans le second membre de (24). Certains coefficients β_{ji} sont structurellement nuls et la variable correspondante ξ_i n'apparaît donc pas dans l'équation (24). On a en particulier $\beta_{jj} = 0$ et la variable ξ_j n'apparaît pas à droite du signe égal dans (24).

L'utilisateur définit les équations (24). Il peut préciser le sens de la liaison entre la variable latente dépendante ξ_j et les autres variables latentes indépendantes ξ_i apparaissant dans l'équation (24) :

$$\begin{aligned} c_{ij} = c_{ji} &= +1, \quad \text{si } \operatorname{cor}(\xi_i, \xi_j) > 0 \\ &= -1, \quad \text{si } \operatorname{cor}(\xi_i, \xi_j) < 0 \\ &= 0 \quad \text{si } \operatorname{cor}(\xi_i, \xi_j) = 0; \quad \text{soit } \beta_{ij} = \beta_{ji} = 0 \end{aligned}$$

Les coefficients c_{ij} non nuls peuvent aussi être estimés à partir des données.

On représente les données et le modèle sous la forme d'un schéma fléché. Les variables manifestes sont représentées par des rectangles et les variables latentes par des ellipses. Les relations de causalité décrites par les relations (22), (23) et (24) sont symbolisées par des flèches. L'origine de la flèche est la variable «cause» (variable explicative) et la pointe de la flèche est la variable «effet» (variable à expliquer). Pour les blocs réflectifs les flèches partent des variables latentes et sont pointées vers les variables manifestes (*outwards direction*). La situation est inversée pour les blocs formatifs : les flèches vont des variables manifestes vers les variables latentes (*inwards direction*). Dans une optique de prévision, il est plutôt naturel que les blocs

endogènes soient réflectifs et les blocs exogènes formatifs. On représente également sur le schéma les signes c_{ij} des corrélations entre les variables latentes reliées entre elles et les signes d_{jh} des coefficients λ_{jh} ou π_{jh}.

L'exemple décrit dans la figure 1 suffit à présenter les aspects les plus importants de la méthode. Il y a trois blocs de variables. Les variables latentes ξ_1 et ξ_2 sont exogènes et la variable latente ξ_3 endogène. Les deux premiers blocs sont formatifs et le troisième bloc est réflectif.

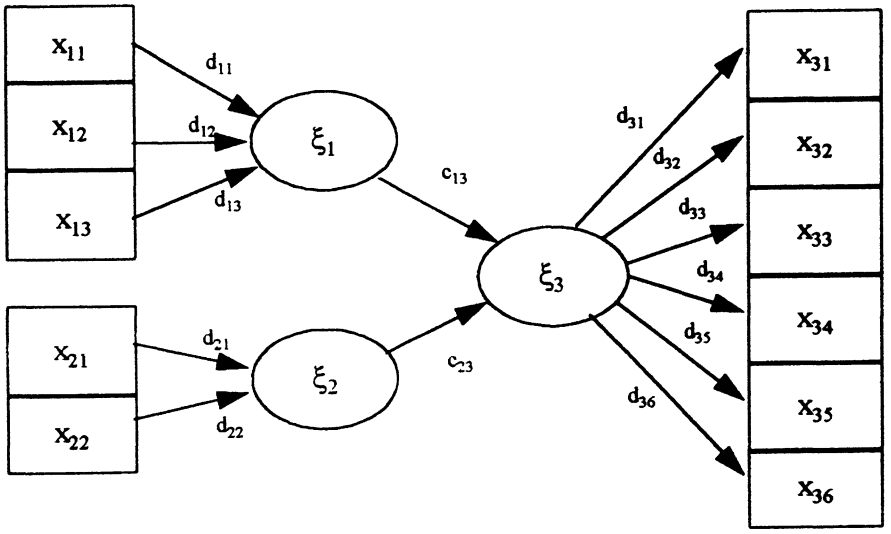


FIGURE 1
Réseau de causalité pour trois groupes de variables

4.2 Estimation des variables latentes

Les variables latentes ξ_j sont estimées de deux manières différentes : 1) l'estimation externe Y_j à partir des variables manifestes x_{jh}, et 2) l'estimation interne Z_j à partir des estimations externes Y_i des variables latentes ξ_i liées à ξ_j.

Estimation externe Y_j de la variable latente ξ_j

L'estimation externe Y_j de la variable latente ξ_j est construite comme une combinaison linéaire des variables manifestes x_{jh} :

$$Y_j = \sum_h w_{jh} x_{jh} = X_j w_j \tag{25}$$

où w_j est le vecteur-colonne des coefficients w_{jh}. On impose à la variable Y_j d'être centrée-réduite.

Estimation interne Z_j de la variable latente ξ_j

On définit une autre approximation Z_j de ξ_j appelée estimation interne de ξ_j. Elle est construite à l'aide des estimations externes Y_i des variables latentes ξ_i liées à ξ_j :

$$Z_j \propto \sum_{\{i|i\neq j, c_{ij}\neq 0\}} e_{ji} Y_i \tag{26}$$

où le signe $\propto$ signifie que la variable située à gauche de ce signe est obtenue par réduction de la variable située à droite.

Lohmöller (1989) décrit trois manières de choisir les coefficients e_{ji}.

(1) Le schéma centroïde

Le schéma centroïde (*Centroid weighting scheme)* consiste à poser $e_{ji} = c_{ji}$, où les coefficients non nuls c_{ji} (égaux à 1 ou – 1) sont fournis par l'utilisateur ou bien estimés par les signes des corrélations r_{ji} entre les variables latentes estimées Y_j et Y_i. C'est le schéma proposé par Wold dans l'algorithme PLS de base.

(2) Le schéma factoriel

Dans le schéma factoriel (*factor weighting scheme)* on pose $e_{ji} = r_{ji} = \text{cor}(Y_j, Y_i)$. L'approximation Z_j est alors construite comme la première composante PLS réduite dans la régression PLS de Y_j sur l'ensemble des variables Y_i liées à Y_j ($c_{ji} \neq 0$).

(3) Le schéma structurel

Dans la troisième approche, Lohmöller distingue deux groupes de variables ξ_i liées à ξ_j. D'une part les variables ξ_i explicatives de la variable ξ_j. Elles apparaissent à droite du signe égal dans l'équation (24). D'autre part les variables ξ_i expliquées par la variable ξ_j. Elles apparaissent à gauche du signe égal dans une des équations structurelles (24). On note b_{ji} les coefficients de régression de la régression multiple de Y_j sur les variables Y_i estimations des variables ξ_i explicatives de la variable ξ_j. Le schéma structurel (*path weighting scheme*) consiste à poser $e_{ji} = b_{ji}$ si la variable ξ_i est explicative de la variable ξ_j, et $e_{ji} = r_{ji}$ si la variable ξ_i est expliquée par la variable ξ_j.

En reliant les estimations externes Y_j et internes Z_j de chaque variable latente ξ_j, Wold obtient des conditions de stationnarité qui permettent de déterminer les variables Y_j. On résoud les équations de stationnarité par un processus itératif. La convergence du processus est prouvée dans le cas de deux groupes et constatée dans la pratique pour les situations plus générales.

Liaison entre les estimations externes Y_j et internes Z_j des variables latentes ξ_j

Wold propose deux modes de relation entre les deux approximations Y_j et Z_j de la variable latente ξ_j. Nous proposons une troisième solution plus générale consistant à utiliser la régression PLS.

Le Mode A (outwards direction)

Dans le mode A la variable Y_j est reliée à la variable Z_j par la formule

$$Y_j \propto \sum_h \text{cor}(x_{jh}, Z_j) x_{jh} \tag{27}$$

L'équation (27) peut aussi s'écrire

$$Y_j \propto X_j X_j' Z_j \tag{28}$$

D'où la condition de stationnarité pour une variable Y_j dans le mode A :

$$Y_j \propto X_j X_j' \sum_{\{i|i\neq j, c_{ij}\neq 0\}} e_{ji} Y_i \tag{29}$$

Le Mode B (inwards direction)

Dans le mode B la variable Y_j est obtenue par régression multiple de Z_j sur les colonnes x_{jh} de la matrice X_j, puis réduction :

$$Y_j \propto X_j (X_j' X_j)^{-1} X_j' Z_j \tag{30}$$

D'où la condition de stationnarité pour une variable Y_j dans le mode B :

$$Y_j \propto X_j (X_j' X_j)^{-1} X_j' \sum_{\{i|i\neq j, c_{ij}\neq 0\}} e_{ji} Y_i \tag{31}$$

On trouve dans la documentation du programme LVPLS 1.8 de Lohmöller (1987) des indications sur le choix du mode en fonction de la nature des blocs :

1) Le mode A est plutôt adapté à des blocs réflectifs et le mode B à des modes formatifs.

2) Le mode A est plutôt adapté à des blocs endogènes et le mode B à des blocs exogènes.

3) Lorsque la matrice X_j n'est pas de rang plein, le mode A s'impose.

Utilisation de la régression PLS

On peut remarquer que le mode A correspond à une utilisation de la régression PLS : $Y_j \propto X_j X_j' Z_j$ représente la prévision de Z_j réduite dans la régression PLS de Z_j sur X_j en utilisant *une seule* composante PLS (ou bien, ce qui revient au même, Y_j représente la première composante PLS réduite). Dans le mode B, $Y_j \propto X_j (X_j' X_j)^{-1} X_j' Z_j$ représente aussi la prévision de Z_j réduite dans la régression PLS de Z_j sur X_j en utilisant *toutes* les composantes PLS disponibles.

On peut donc penser à une solution intermédiaire consistant à construire Y_j comme la prévision de Z_j réduite obtenue par régression PLS de Z_j sur X_j, où le nombre de composantes est choisi par validation croisée (procédure *Autofit* de SIMCA-P). Cette approche présente un double avantage : 1) il y a une solution de continuité entre les modes A et B, et le choix est fait en fonction des données, et 2) la régression PLS fonctionne même avec des données manquantes.

Calcul des estimations externes Y_j des variables latentes ξ_j

L'algorithme itératif proposé par Wold est maintenant naturel. À l'étape initiale on part de valeurs initiales des estimations externes Y_i fixées selon des règles décrites plus loin. On obtient à l'aide des équations (29) et/ou (31) de nouvelles valeurs de ces variables. Les estimations Y_j des variables latentes ξ_j sont obtenues en itérant ce processus jusqu'à convergence. Les vecteurs w_j se déduisent ensuite des équations (29) et (31).

Il semble que le choix du schéma de construction des estimations internes Z_j ait peu d'influence sur le calcul des estimations externes Y_j. Noonan et Wold (1982) ont comparé ces différents schémas sur un modèle à 16 variables latentes décrites par 49 variables manifestes évaluées sur un échantillon d'environ 3 200 individus. Ils ont constaté que les variables Y_j variaient très peu d'un schéma à l'autre. Dans la conclusion de leur article, ils conseillent donc l'utilisation du schéma centroïde. Les autres schémas ont cependant un intérêt théorique car c'est leur utilisation qui permet de montrer que l'approche PLS contient comme cas particuliers l'analyse en composantes principales et l'analyse canonique généralisée aux sens de Horst et de Carroll au niveau de la première composante (*cf.* les sections 4.5 et 4.6).

De même, le choix des estimations externes initiales Y_i influe peu sur la convergence de l'algorithme. S'il n'y a pas de données manquantes, on peut choisir de démarrer avec une variable manifeste représentant bien le concept décrit par le bloc. S'il y a des données manquantes, une bonne solution consiste à choisir la première composante principale du bloc estimée par la procédure NIPALS.

4.3 Estimation des relations structurelles et mesure de la qualité du modèle par le test de validation croisée de Stone-Geisser

On estime les paramètres des modèles définis par les équations (22), (23) et (24) par régression en remplaçant les variables latentes ξ_j par leur estimation Y_j. Si les coefficients c_{ij} sont fixés *a priori*, la vérification des similitudes des signes entre les corrélations entre Y_i et Y_j et les coefficients c_{ij} est un premier test de cohérence.

On étudie la capacité du modèle à prédire les variables manifestes x_{jh} des blocs endogènes X_j. On relie tout d'abord la variable manifeste x_{jh} à sa variable latente estimée Y_j par régression simple :

$$x_{jh} \approx p_{jh} Y_j \tag{32}$$

Les équations (24) sont estimées par régression multiple de Y_j sur les variables Y_i qui lui sont structurellement liées ($\beta_{ji} \neq 0$) :

$$Y_j \approx \sum_{i:\beta_{ji}\neq 0} b_{ji} Y_i \tag{33}$$

On peut mesurer la qualité du modèle défini par l'équation (24) et tester des hypothèses structurelles (nullité de certains coefficients β_{ji}) par validation croisée. Pour cela Wold utilise l'approche de Stone (1974) et Geisser (1974) qui, pour reprendre son expression, «***fits soft modeling like hand in glove***». Tout d'abord on peut relier le bloc endogène X_j aux blocs X_i correspondant à des coefficients $c_{ij} \neq 0$ en remplaçant dans les équations (32) et (33) Y_j par $X_j w_j$ et on obtient l'équation de prédiction :

$$pred\ X_j = (\sum_{i:c_{ij}\neq 0} b_{ji} X_i w_i) p'_j \tag{34}$$

où $p'_j = (p_{j1}, ..., p_{jk_j})$. On découpe ensuite les données du tableau X_j en G classes. Wold (1982, p. 31) précise la construction de ces classes. On lit le tableau X_j colonne par colonne et on numérote les données x_{jhi} de 1 à $n \times k_j$. La première classe est formée des données numérotées $1, G+1, 2G+1, ...$, la deuxième classe des données numérotées $2, G+2, 2G+2, ...$, et ainsi de suite jusqu'à la G-ième classe. On élimine à tour de rôle chaque classe d'individus. On estime à chaque fois le modèle sur les $G-1$ classes restantes et on utilise la formule de prévision (34) pour prédire les valeurs de la classe exclue. On note $\widehat{pred}\ X_j$ la prévision de X_j obtenue de cette manière. D'où une somme des carrés résiduelle

$$SCR_j = \left\| X_j - \widehat{pred}\ X_j \right\|^2$$

Par ailleurs on calcule une prévision «naïve» de la valeur x_{jhi} de la variable x_{jh} pour l'individu i en calculant la moyenne $\overline{x}_{jh(-i)}$ de la variable x_{jh} en excluant l'individu i. Wold propose de mesurer la qualité de reconstitution du tableau X_j à l'aide du critère de Stone-Geisser Q_j^2 défini par

$$Q_j^2 = 1 - \frac{SCR_j}{\sum_h \sum_i (x_{jhi} - \overline{x}_{jh(-i)})^2}$$

L'expérience montre que le critère Q_j^2 est relativement robuste au choix du nombre de classes G. Wold conseille de choisir entre 5 et 10 classes. Il faut de plus bien répartir dans le tableau X_j la localisation des données supprimées à tour de rôle. Il est préférable de choisir pour G un nombre entier premier par rapport à k_j et n. Dans le logiciel LVPLS 1.8 le nombre de classes G doit simplement être différent du nombre n d'individus. Le critère Q_j^2 est analogue à un coefficient de détermination R^2, mais il peut être négatif. Dans ce dernier cas le modèle étudié n'est pas acceptable.

Considérons maintenant une hypothèse H_0 consistant le plus souvent à annuler des coefficients β_{ji} dans l'équations (24). On peut calculer les sommes de carrés résiduelles $[SCR_j]_0$ et $[SCR_j]_1$ correspondant respectivement à l'estimation du modèle sous l'hypothèse H_0 et sans hypothèse restrictive particulière. On en déduit des valeurs associées Q^2_{j0} et Q^2_{j1} du critère de Stone-Geisser. On rejette l'hypothèse H_0 lorsque Q^2_{j0} est nettement inférieur à Q^2_{j1}.

4.4 Étude du cas de deux blocs

Nous montrons dans cette section comment l'utilisation de l'algorithme PLS de base permet de retrouver les principales méthodes utilisées pour relier deux groupes de variables au niveau de la première étape. On précise dans le tableau 1 les méthodes correspondant aux différents choix possibles des modes A ou B de calcul pour Y_1 et Y_2.

TABLEAU 1
Équivalence entre l'algorithme PLS appliqué à deux blocs de variables X_1 et X_2 et la première étape de différentes méthodes

	Analyse canonique	Analyse factorielle inter-batteries	Analyse des redondances de X_2 par rapport à X_1
Mode de calcul pour Y_1	B	A	B
Mode de calcul pour Y_2	B	A	A

Pour vérifier ces résultats il suffit d'écrire les conditions de stationnarité (28) et (30) pour ces différentes situations. On suppose pour simplifier, et sans enlever de généralité, que les variables latentes ξ_1 et ξ_2 sont corrélées positivement. Par conséquent $Z_1 = Y_2$ et $Z_2 = Y_1$.

Analyse canonique

Les conditions de stationnarité s'écrivent en utilisant le mode B pour Y_1 et Y_2 :

$$Y_1 \propto X_1(X_1'X_1)^{-1}X_1'Y_2$$

$$Y_2 \propto X_2(X_2'X_2)^{-1}X_2'Y_1$$

L'algorithme PLS converge donc bien vers les premières composantes canoniques de l'analyse canonique de X_1 et X_2.

Analyse factorielle inter-batteries

Les conditions de stationnarité s'écrivent en utilisant le mode A pour Y_1 et Y_2 :

$$Y_1 \propto X_1X_1'Y_2$$

$$Y_2 \propto X_2X_2'Y_1$$

L'algorithme PLS converge donc bien vers les premières composantes de l'analyse factorielle inter-batteries des deux tableaux X_1 et X_2.

On peut remarquer que le cas particulier où $X_1 = X_2$ permet d'obtenir la première composante principale réduite Y_1 de X_1 puisque Y_1, vecteur propre de X_1X_1' associé à la plus grande valeur propre, est aussi vecteur propre de $(X_1X_1')^2$ associé à la plus grande valeur propre.

Analyse des redondances de X_2 par rapport à X_1

Les conditions de stationnarité s'écrivent en utilisant le mode B pour Y_1 et le mode A pour Y_2 :

$$Y_1 \propto X_1(X_1'X_1)^{-1}X_1'Y_2$$

$$Y_2 \propto X_2X_2'Y_1$$

L'algorithme PLS converge vers la première composante Y_1 de l'analyse des redondances de X_2 par rapport à X_1 : $Y_1 = X_1w_1$ est vecteur propre de la matrice $X_1(X_1'X_1)^{-1}X_1'X_2X_2'$ associé à la plus grande valeur propre et on retrouve donc bien que w_1 est vecteur propre de la matrice $(X_1'X_1)^{-1}X_1'X_2X_2'X_1$ associé à la plus grande valeur propre.

Différentes possibilités d'étendre l'algorithme PLS vers la recherche d'autres composantes $Y_j^{(h)}$ ont été proposées par Apel et Wold (1982), Wold (1984) et Lohmöller (1989). Nous avons étudié dans Tenenhaus (1998, chapitre 11) celles qui conduisaient aux méthodes présentées dans le tableau 1 de cet article.

4.5 Approche PLS et Analyse Canonique Généralisée de J blocs de variables

Nous allons montrer dans cette section que les différentes options de l'approche PLS (Mode A ou Mode B, schéma de construction des estimations internes de type centroïde, factoriel ou structurel) permettent de retrouver l'analyse canonique généralisée aux sens de Horst(1961) et de Carroll (1968).

On considère J blocs de variables $X_1, ..., X_j, ..., X_J$. Chaque bloc de variables X_j représente la version observable d'une variable latente ξ_j. Pour estimer ces variables latentes ξ_j, Wold (1982) propose d'introduire un nouveau bloc de variables X formé de la juxtaposition des J blocs X_j et résumé par une variable latente ξ. On considère ensuite un modèle structurel où chaque variable ξ_j est prédictive de la variable ξ. Le schéma fléché correspondant à cette situation est visualisé dans la figure 2.

Lohmöller (1989) a montré que les équations stationnaires de l'analyse canonique généralisée de Horst (critère SUMCOR) pouvaient être retrouvées en estimant les variables latentes du modèle structurel décrit dans la figure 2 en choisissant le Mode B et le schéma centroïde. Nous avons aussi constaté que les équations stationnaires obtenues en choisissant le Mode B et le schéma factoriel permettaient de retrouver l'analyse canonique généralisée de Carroll.

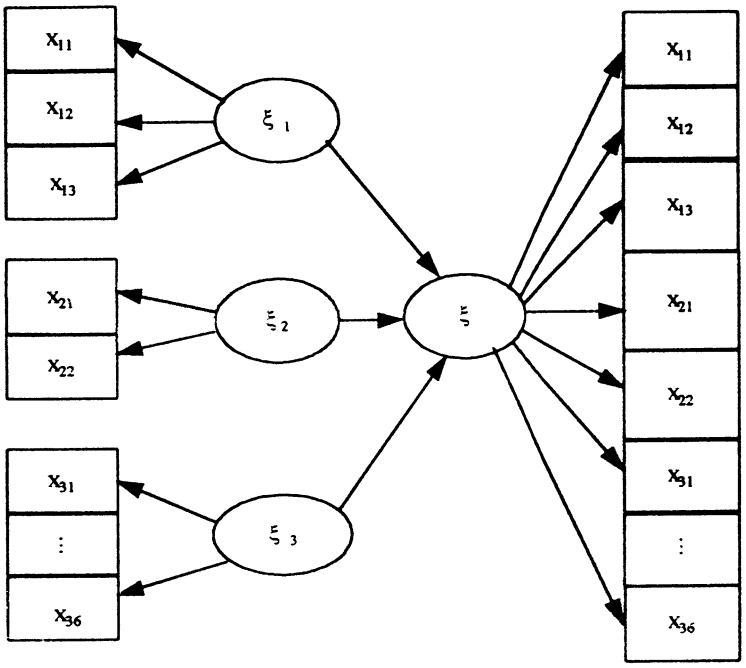


FIGURE 2
Schéma fléché pour l'analyse de J blocs.

4.5.1 Approche PLS et analyse canonique généralisée de Horst

Rappelons tout d'abord l'approche de Horst en suivant Saporta (1975).
On recherche des variables centrées-réduites $V_j = X_j a_j$ maximisant le critère

$$\sum_j \sum_k \text{cor}(X_j a_j, X_k a_k) \tag{35}$$

On cherche donc le maximum du critère

$$\sum_j \sum_k a_j' (\frac{1}{n} X_j' X_k) a_k \tag{36}$$

sous les contraintes

$$a_j' (\frac{1}{n} X_j' X_j) a_j = 1 \tag{37}$$

L'utilisation des multiplicateurs de Lagrange conduit aux équations de stationnarité suivantes :

$$\sum_k X_j' X_k a_k - \lambda_j X_j' X_j a_j = 0 \tag{38}$$

En posant $V = \sum_k X_k a_k$ et $P_j = X_j(X_j X_j')^{-1} X_j$ les équations (38) s'écrivent aussi

$$P_j V = \lambda_j V_j \tag{39}$$

soit

$$P_j(\sum_k X_k a_k) = \lambda_j X_j a_j \tag{40}$$

Les vecteurs $a_1, ..., a_J$ maximisant le critère (36) sous les contraintes (37) doivent donc vérifier les équations stationnaires (40).

Par ailleurs, on déduit des équations (38) et des contraintes (37) l'égalité

$$\sum_j \sum_k a_j'(\frac{1}{n} X_j' X_k) a_k = \sum_j \lambda_j \tag{41}$$

Il faut donc choisir la solution des équations stationnaires (40) maximisant $\sum_j \lambda_j$.

On remarque que la variance de V est aussi égale à $\sum_j \lambda_j$:

$$\begin{aligned} \frac{1}{n} V'V &= \frac{1}{n}(\sum_j X_j a_j)'(\sum_j X_j a_j) \\ &= \sum_j \sum_k a_j'(\frac{1}{n} X_j' X_k) a_k \\ &= \sum_j \lambda_j \end{aligned}$$

On déduit de (37) et (39)

$$\begin{aligned} \lambda_j &= \frac{1}{n} V_j' P_j V \\ &= \frac{1}{n} V_j' V \\ &= \text{cov}(V_j, V) \end{aligned}$$

Et par conséquent

$$\text{cor}(V_j, V) = \frac{\lambda_j}{\sqrt{\sum_j \lambda_j}}$$

La méthode de Horst consiste aussi à rechercher le maximum du critère

$$\sum_j \text{cor}(V_j, V) \tag{42}$$

sous les contraintes :

(1) les variables $V_j = X_j a_j$ sont centrées-réduite

(2) $V = \sum_j V_j$

À l'optimum, la valeur de (42) est donc égale à $\sqrt{\sum_j \lambda_j}$.

Montrons maintenant que l'utilisation du Mode B et du schéma centroïde conduit également aux équations (40) lorsqu'on suppose que les corrélations entre les variables latentes ξ_j et ξ sont positives.

Les estimations externes centrées-réduites Y_j et Y des variables latentes ξ_j et ξ s'écrivent

$$Y_j = X_j w_j$$

et

$$Y = Xw$$

Les estimations internes Z_j et Z sont calculées en utilisant l'hypothèse de positivité des corrélations entre les variables latentes ξ_j et ξ. On obtient donc

$$Z_j = Y$$

et

$$Z \propto \sum_{j=1}^{J} Y_j$$

Les équations de stationnarité de l'approche PLS sous le Mode B s'écrivent maintenant

$$Y_j \propto P_j Y \tag{43}$$

et

$$Y \propto X(X'X)^{-1}X\left(\sum_{j=1}^{J} Y_j\right) \tag{44}$$

Comme $\sum_{j=1}^{J} Y_j$ appartient à l'espace engendré par les colonnes de X, on déduit

$$Y \propto \sum_{j=1}^{J} Y_j \tag{45}$$

et par conséquent, en reprenant (43), on obtient les équations

$$X_j w_j \propto P_j(\sum_{j=1}^{J} X_j w_j) \tag{46}$$

analogues aux équations de stationnarité (40). D'où l'équivalence annoncée.

Lohmöller a montré qu'on pouvait aussi retrouver la méthode de Horst sans faire intervenir de variable auxiliaire ξ. Il considère le schéma fléché de la figure 3. Les variables latentes de chaque bloc sont maintenant toutes liées entre elles.

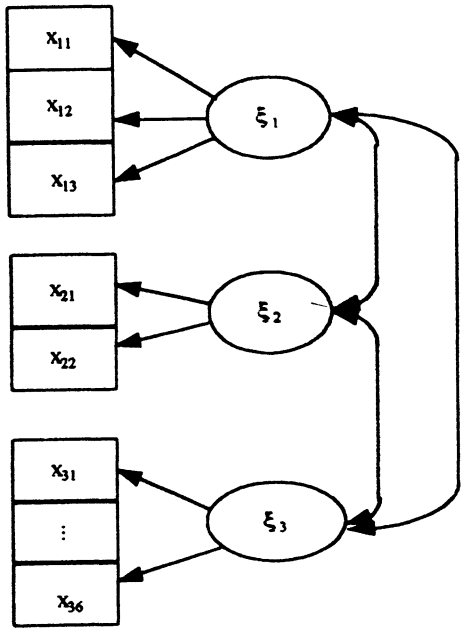


FIGURE 3
Schéma fléché pour l'analyse canonique généralisée de Horst.

Montrons que l'utilisation du Mode B et du schéma centroïde sur le modèle structurel de la figure 3 conduit encore aux équations (40).

Les estimations externes centrées-réduites Y_j des variables latentes ξ_j s'écrivent

$$Y_j = X_j w_j$$

Les estimations internes Z_j sont calculées en supposant toutes les corrélations entre les ξ_j positives. On obtient donc

$$Z_j \propto \sum_{k \neq j}^{J} Y_k$$

Les équations de stationnarité de l'approche PLS sous le Mode B s'écrivent maintenant

$$Y_j \propto P_j(\sum_{k \neq j}^{J} Y_k)$$

soit aussi, puisque $P_j(Y_j) = Y_j$,

$$Y_j \propto P_j(\sum_{k=1}^{J} Y_k)$$

et par conséquent les équations

$$X_j w_j \propto P_j(\sum_{j=1}^{J} X_j w_j)$$

analogues aux équations de stationnarité (40). D'où le résultat.

4.5.2 Approche PLS et analyse canonique généralisée de Carroll

Rappelons l'analyse canonique généralisée de Carroll. On recherche une variable auxiliaire centrée-réduite $V = Xa$ maximisant le critère

$$\sum_j R^2(V, X_j) \tag{47}$$

Le critère (47) s'écrivant

$$\frac{1}{n} \sum_j V' P_j V \tag{48}$$

on en déduit que la variable V est vecteur propre de la matrice $\sum_j P_j$ associé à la plus grande valeur propre qui est alors la valeur du maximum de (47).

Montrons maintenant que l'utilisation du Mode B et du schéma factoriel sur le modèle de la figure 2 permet de retrouver la solution de l'analyse canonique généralisée de Carroll.

Les estimations externes centrées-réduites Y_j et Y des variables latentes ξ_j et ξ s'écrivent

$$Y_j = X_j w_j$$

et

$$Y = Xw$$

Les estimations internes Z_j et Z sont calculées en supposant toutes les corrélations entre les ξ_j et ξ positives. On obtient donc

$$Z_j = Y$$

et

$$Z \propto \sum_{j=1}^{J} r_j Y_j$$

où r_j représente la corrélation entre les variables Y et Y_j.

Les équations de stationnarité de l'approche PLS sous le Mode B s'écrivent maintenant

$$Y_j \propto P_j Y$$

et

$$Y \propto X(X'X)^{-1}X(\sum_{j=1}^{J} r_j Y_j)$$

La variable Y étant centrée-réduite, l'écart-type de la variable $P_j Y$ est égal à la corrélation r_j et par conséquent

$$Y_j = \frac{1}{r_j} P_j Y \tag{49}$$

Comme $\sum_{j=1}^{J} r_j Y_j$ appartient à l'espace engendré par les colonnes de X, on déduit

$$Y \propto \sum_{j=1}^{J} r_j Y_j \tag{50}$$

Par conséquent, en reprenant (49) et (50), on obtient l'équation

$$Y \propto (\sum_j P_j)Y \tag{51}$$

montrant que Y est vecteur propre de $\sum_j P_j$.

Dans l'approche PLS, on part d'une variable Y égale par exemple à x_{11}, et on utilise de manière itérative les équations (51), (49) et (50). Cette méthode converge bien vers le vecteur propre de la matrice $\sum_j P_j$ associé à la plus grande valeur propre. D'où l'équivalence annoncée.

4.6 Approche PLS et analyse en composantes principales de X

Lohmöller (1989) a étudié l'utilisation du Mode A et du schéma structurel pour estimer les variables latentes du modèle structurel de la figure 2. Il a montré que les équations stationnaires définies par ce modèle avaient pour solution une variable Y égale à la première composante principale réduite de X et des variables Y_j représentant les fragments de la première composante principale construits sur les blocs X_j puis réduits. Nous reprenons ici sa démarche.

Les estimations externes centrées-réduites Y_j et Y des variables latentes ξ_j et ξ s'écrivent

$$Y_j = X_j w_j$$

et

$$Y = Xw$$

Les estimations internes Z_j et Z sont calculées en supposant toutes les corrélations entre les ξ_j et ξ positives. On obtient donc

$$Z_j = Y$$

et

$$Z \propto \sum_{j=1}^{J} b_j Y_j \tag{52}$$

où b_j représente le coefficient de régression de Y_j dans la régression multiple de Y sur $Y_1, ..., Y_J$.

Les équations de stationnarité de l'approche PLS sous le Mode A s'écrivent maintenant

$$Y_j \propto X_j X_j' Y \tag{53}$$

et

$$Y \propto XX'(\sum_{j=1}^{J} b_j Y_j) \tag{54}$$

Soit $Y = Xa = \sum_j X_j a_j$, la première composante principale réduite du bloc X. Les variables $X_j a_j$ représentent les fragments de Y définis sur les blocs X_j. Notons Y_j la variable $X_j a_j$ réduite. Les coefficients de régression b_j de Y_j dans la régression multiple de Y sur $Y_1, ..., Y_J$ représentent donc tout simplement les écarts-types des fragments $X_j a_j$ et par conséquent

$$Y = \sum_{j=1}^{J} b_j Y_j \tag{55}$$

Les quantités Y, Y_j, et b_j ainsi définies vérifient bien les équations (53) et (54) puisqu'en analyse en composantes principales le vecteur a est colinéaire au vecteur des covariances entre les variables x_{jh} et la composante principale Y.

Nous avons donc montré que la première composante principale réduite de X et ses fragments sur chaque bloc X_j vérifient bien les équations de stationnarité (53) et (54). Mais il faudrait encore vérifier que l'approche PLS converge vers cette solution. Lohmöller a constaté ce résultat dans la pratique, mais n'en fournit pas une démonstration mathématique.

4.7 Traitement d'un exemple

Les données de cet exemple proviennent d'un article de Russet (1964) et sont reproduites et analysées dans Gifi (1990) et dans Tenenhaus (1998). Russet cherche à montrer que l'inégalité économique entraine l'instabilité politique.

Pour mesurer l'inégalité économique Russet utilise des variables décrivant la répartition des terres agricoles, le produit national brut par tête et le pourcentage de personnes actives travaillant dans l'agriculture :

– La variable GINI représente l'écart entre la courbe de Lorenz de la répartition des terres et la droite d'égalité mesuré à l'aide de l'indice de concentration de Gini.

– La variable FARM correspond au pourcentage de fermiers possèdant la moitié des terres, en commençant par les plus petites surfaces. Si FARM vaut 90%, alors la moitié des terres est possédée par 10% des fermiers.

– La variable RENT mesure le pourcentage de fermiers locataires de leurs terres.

– La variable GNPR (gross national product per capita) est le produit national brut par tête en dollars U.S. en 1955.

– La variable LABO est égale au pourcentage de personnes actives travaillant dans l'agriculture.

Il y a quatre mesures de l'instabilité politique :

– La variable INST est une fonction du nombre de responsables du pouvoir exécutif et du nombre d'années pendant lesquelles le pays a été indépendant entre 1945 et 1961. Cet indice varie entre 0 (très stable) et 17 (très instable).

– La variable ECKS est l'indice d'Eckstein calculé sur la période 1946-1961. Il mesure le nombre de conflits violents entre communautés sur cette période.

– La variable DEAT est le nombre de personnes tuées lors de manifestations violentes sur la période 1950-1962.

– La variable DEMO classe les pays en trois groupes : démocratie stable, démocratie instable et dictature.

Les données d'origine sont reproduites dans le tableau 2.

Gifi décrit une utilisation de l'algorithme CANALS (canonical correlation with alternating least squares) permettant de construire des transformations des variables optimisant l'analyse canonique entre les variables économiques transformées et les variables politiques transformées. Les transformations présentées dans Gifi (op. cité, p. 230) suggèrent d'utiliser les variables suivantes : GINI, FARM, Ln(RENT+1), Ln(GNPR), Ln(LABO), Exp(INST-16.3), Ln(ECKS+1), Ln(DEAT+1). La variable DEMO étant qualitative, on la remplace par les trois variables indicatrices des états politiques : DEMOSTAB, DEMOINST et DICTATUR.

Il s'agit de relier les variables mesurant l'instabilité politique aux variables décrivant la situation économique. Les variables GINI, FARM, LRENT = Ln(RENT+1), LGNPR = Ln(GNPR) et LLABO = Ln(LABO) constituent les prédicteurs X. Les réponses Y sont les variables EINST = Exp(INST-16.3), LECKS = Ln(ECKS+1), LDEAT = Ln(DEAT+1), DEMOSTAB, DEMOINST et DICTATUR.

L'approche PLS permet de modéliser la relation entre l'inégalité économique et l'instabilité politique en considérant trois groupes de variables. Le premier groupe X_1 est formé des variables x_{11} = GINI, x_{12} = FARM et x_{13} = Ln(RENT+1) décrivant l'inégalité dans la répartition des terres. Le deuxième groupe X_2, constitué des variables x_{21} = Ln(GNPR) et x_{22} = Ln(LABO), décrit le développement industriel. Les variables politiques transformées x_{31} = Exp(INST-16.3), x_{32} = Ln(ECKS+1), x_{33} = Ln(DEAT+1), x_{34} = DEMOSTAB, x_{35} = DEMOINST et x_{36} = DICTATUR forment le troisième groupe X_3 et correspondent à l'instabilité politique. Pour faciliter les calculs toutes les variables sont préalablement centrées-réduites. Les notations sont néanmoins conservées pour plus de simplicité. On suppose que chaque groupe X_j est résumé par une variable latente ξ_j : ξ_1 représente l'inégalité agricole, ξ_2 le niveau de développement industriel et ξ_3 l'instabilité politique. On cherche à relier l'instabilité politique ξ_3 à l'inégalité économique ξ_1 et au développement industriel ξ_2. Nous supposons que les trois blocs sont réflectifs car ils ne sont que le reflet de situations évidemment plus complexes.

Les équations du modèle étudié s'écrivent :

$$x_{1h} = \pi_{1h}\xi_1 + \varepsilon_{1h},\ h = 1, ..., 3 \tag{56.1}$$

$$x_{2h} = \pi_{2h}\xi_2 + \varepsilon_{2h},\ h = 1, 2 \tag{56.2}$$

$$x_{3h} = \pi_{3h}\xi_3 + \varepsilon_{3h},\ h = 1, ..., 6 \tag{56.3}$$

TABLEAU 2
Les données de Russett

Pays	gini	farm	rent	gnpr	labo	inst	ecks	deat	demo
Argentine	86.3	98.2	32.9	374	25	13.6	57	217	2
Australie	92.9	99.6	–	1215	14	11.3	0	0	1
Autriche	74.0	97.4	10.7	532	32	12.8	4	0	2
Belgique	58.7	85.8	62.3	1015	10	15.5	8	1	1
Bolivie	93.8	97.7	20.0	66	72	15.3	53	663	3
Brésil	83.7	98.5	9.1	262	61	15.5	49	1	2
Canada	49.7	82.9	7.2	1667	12	11.3	22	0	1
Chili	93.8	99.7	13.4	180	30	14.2	21	2	2
Colombie	84.9	98.1	12.1	330	55	14.6	47	316	2
Costa Rica	88.1	99.1	5.4	307	55	14.6	19	24	2
Cuba	79.2	97.8	53.8	361	42	13.6	100	2900	3
Danemark	45.8	79.3	3.5	913	23	14.6	0	0	1
Rép. Dominic.	79.5	98.5	20.8	205	56	11.3	6	31	3
Equateur	86.4	99.3	14.6	204	53	15.1	41	18	3
Egypte	74.0	98.1	11.6	133	64	15.8	45	2	3
Espagne	78.0	99.5	43.7	254	50	0.0	22	1	3
Etats-Unis	70.5	95.4	20.4	2343	10	12.8	22	0	1
Finlande	59.9	86.3	2.4	941	46	15.6	4	0	2
France	58.3	86.1	26.0	1046	26	16.3	46	1	2
Guatemala	86.0	99.7	17.0	179	68	14.9	45	57	3
Grèce	74.7	99.4	17.7	239	48	15.8	9	2	2
Honduras	75.5	97.4	16.7	137	66	13.6	45	111	3
Inde	52.2	86.9	53.0	72	71	3.0	83	14	1
Irak	88.1	99.3	75.0	195	81	16.2	24	344	3
Irlande	59.8	85.9	2.5	509	40	14.2	9	0	1
Italie	80.3	98.0	23.8	442	29	15.5	51	1	2
Japon	47.0	81.5	2.9	240	40	15.7	22	1	2
Libye	70.0	93.0	8.5	90	75	15.8	8	0	3
Luxembourg	63.8	87.7	18.8	1194	23	12.8	0	0	1
Nicaragua	75.7	96.4	–	254	68	12.8	16	16	3
Norvège	66.9	87.5	7.5	969	26	12.8	1	0	1
Nlle Zélande	77.3	95.5	22.3	1259	16	12.8	0	0	1
Panama	73.7	95.0	12.3	350	54	15.6	29	25	3
Pays-Bas	60.5	86.2	53.3	708	11	13.6	2	0	1
Pérou	87.5	96.9	–	140	60	14.6	23	26	3
Philippines	56.4	88.2	37.3	201	59	14.0	15	292	3
Pologne	45.0	77.7	0.0	468	57	8.5	19	5	3
RFA	67.4	93.0	5.7	762	14	3.0	4	0	2
Royaume-uni	71.0	93.4	44.5	998	5	13.6	12	0	1
Salvador	82.8	98.8	15.1	244	63	15.1	9	2	3
Sud Vietnam	67.1	94.6	20.0	133	65	10.0	50	1000	3
Suède	57.7	87.2	18.9	1165	13	8.5	0	0	1
Suisse	49.8	81.5	18.9	1229	10	8.5	0	0	1
Taiwan	65.2	94.1	40.0	132	50	0.0	3	0	3
Uruguay	81.7	96.6	34.7	569	37	14.6	1	1	1
Venezuela	90.0	99.3	20.6	762	42	14.9	36	111	3
Yougoslavie	43.7	79.8	0.0	297	67	0.0	9	0	3

et

$$\xi_3 = \beta_{31}\xi_1 + \beta_{32}\xi_2 + \zeta_3 \tag{57}$$

On peut faire des hypothèses naturelles sur les signes des paramètres des équations (56) :

$$\pi_{1h} \geqslant 0,\ h = 1, ..., 3 \tag{58.1}$$

$$\pi_{21} \geqslant 0 \quad \text{et} \quad \pi_{22} \leqslant 0 \tag{58.2}$$

$$\pi_{3h} \geqslant 0,\ h = 1, ..., 3,\ \pi_{34} \leqslant 0,\ \pi_{35} \geqslant 0,\ \pi_{36} \geqslant 0 \tag{58.3}$$

Les variables x_{11} = GINI, x_{12} = FARM et x_{13} = Ln(RENT+1) sont corrélées positivement à l'inégalité dans la répartition des terres. La variable x_{21} = Ln(GNPR) est corrélée positivement au développement industriel et x_{22} = Ln(LABO) négativement. Les variables x_{31} = Exp(INST-16.3), x_{32} = Ln(ECKS+1), x_{33} = Ln(DEAT+1), x_{35} = DEMOINST et x_{36} = DICTATUR sont corrélées positivement à l'instabilité politique et la variable x_{34} = DEMOSTAB négativement. Ces conditions sont utilisées dans la phase d'estimation du modèle pour choisir les valeurs initiales des Y_j et ensuite pour le valider.

L'approche PLS permet d'étudier de manière précise la relation entre l'inégalité économique ξ_1, le développement industriel ξ_2 et l'instabilité politique ξ_3 à travers l'équation (57). On peut faire l'hypothèse que $\beta_{31} > 0$ et $\beta_{32} < 0$. On pose donc $c_{13} = c_{31} = +1$, $c_{23} = c_{32} = -1$ et tous les autres coefficients c_{ij} sont nuls. Suivant les recommandations de Lohmöller, nous choisissons le mode A pour estimer les variables latentes. Pour calculer les estimations internes nous avions choisi le schéma centroïde dans Tenenhaus (1998). Nous allons retenir dans cet article le schéma factoriel. Il paraît en effet plus logique de pondérer les variables du problème par les corrélations adéquates. On déduit de la formule (29) et du schéma factoriel les équations de stationnarité suivantes :

$$Y_1 \propto X_1 X_1' Y_3 \tag{59}$$

$$Y_2 \propto X_2 X_2' Y_3 \tag{60}$$

$$Y_3 \propto X_3 X_3' (r_{31} Y_1 + r_{32} Y_2) \tag{61}$$

où $r_{ij} = \text{cor}(Y_i, Y_j)$.

L'algorithme PLS consiste à partir d'une solution initiale, ici $Y_1 = x_{11}$, $Y_2 = x_{21}$, $Y_3 = x_{31}$ puisque les coefficients $\pi_{11}, \pi_{21}, \pi_{31}$ sont positifs, à utiliser les équations (59) à (61) pour obtenir une nouvelle solution, puis à itérer cette procédure jusqu'à convergence. Les équations (59) à (61) montrent qu'à chaque étape :

– la nouvelle variable Y_1 est la première composante PLS réduite de la régression PLS de la variable Y_3, calculée à l'étape précédente, sur X_1,

– la nouvelle variable Y_2 est la première composante PLS réduite de la régression PLS de la variable Y_3 sur X_2,

– la nouvelle variable Y_3 est la première composante PLS réduite de la régression PLS de la variable $\widehat{Y}_3 = r_{31} Y_1 + r_{32} Y_2$, calculée à l'étape précédente, sur le bloc X_3.

On peut aussi noter que la variable $\widehat{Y}_3$ est la première composante PLS de la régression PLS de la variable Y_3, calculée à l'étape à l'étape précédente, sur les variables Y_1 et Y_2.

Nous avons utilisé le programme de régression PLS de SIMCA-P pour calculer les variables Y_j aux différentes étapes. Ceci nous a permis de gérer sans difficulté les quelques données manquantes du tableau de données.

Les résultats de l'algorithme PLS

L'algorithme PLS a convergé en trois itérations. Les variables latentes estimées Y_j sont données dans le tableau 3. Elles sont combinaisons linéaires des variables de leur groupe. On peut les écrire en revenant aux données d'origine :

$$Y_1 = -9.65 + 0.032 \times GINI + 0.077 \times FARM + 0.10 \times LRENT$$

$$Y_2 = -0.69 + 0.57 \times LGNPR - 0.77 \times LLABO$$

$$Y_3 = -1.58 + 0.42 \times EINST + 0.20 \times LECKS + 0.12 \times LDEAT$$
$$+ 0 \times DEMOSTAB + 0.77 \times DEMOINST + 1.32 \times DICTATUR$$

Tous les coefficients des variables sont cohérents au niveau des signes. On vérifie de même la cohérence des signes des corrélations entre les x_{jh} et Y_j :

	gini	farm	lrent
cor(x_{1h}, Y_1)	0.975	0.986	0.517

	lgnpr	llabo
cor(x_{2h}, Y_2)	0.950	– 0.955

	einst	lecks	ldeat	demostab	demoinst	dictatur
cor(x_{3h}, Y_3)	0.345	0.810	0.790	– 0.863	0.074	0.749

Nous avons comparé les estimations Y_j obtenues en utilisant le schéma centroïde (Tenenhaus, 1998) et le schéma factoriel utilisé ici. Les corrélations entre les estimations issues de ces deux schémas sont respectivement de 1, 1 et 0.9997. Nous retrouvons donc ici les résultats de Noonan et Wold mentionnés plus haut.

Nous avons également utilisé sur ces données le programme LVPLS 1.8 de Lohmöller. Nous ne savons pas exactement comment ce programme prend en compte les données manquantes. Pour permettre la comparaison nous avons donc estimé les valeurs manquantes de LRENT par régression de LRENT sur Y_1. Les données

manquantes de l'Australie, du Nicaragua et du Pérou sont respectivement estimés par 3.40, 3.00 et 3.21. Les estimations des variables latentes obtenues en utilisant LVPLS 1.8 et l'approche présentée dans cet article sont très proches.

On peut écrire l'équation de régression reliant l'instabilité politique Y_3 à l'inégalité économique Y_1 et au niveau de développement industriel Y_2 :

$$\underset{}{Y_3} \approx \underset{(2.177)}{0.211\, Y_1} - \underset{(-7.231)}{0.701\, Y_2}$$

Le R^2 vaut 0.627 et les t, qui apparaissent entre parenthèses, sont tous les deux «significatifs». Nous avons par ailleurs utilisé le programme LVPLS 1.8 sur les données complétées pour réaliser la validation croisée. Nous avons choisi $G = 30$. On obtient un coefficient $Q_3^2 = 0.231$: il est naturellement plus difficile de prédire les différentes variables manifestes x_{3h} que la variable latente Y_3. Les coefficients de régression b_{31} et b_{32} du modèle reliant Y_3 à Y_1 et Y_2 ont respectivement pour moyennes 0.2095 et – 0.6963 et pour écarts-types 0.0141 et 0.0137.

Les résultats sont visualisés dans la figure 4 sous la forme d'un schéma fléché. On indique sur les flèches reliant les variables manifestes x_{jh} à leur variable latente ξ_j les corrélations $\text{cor}(x_{jh}, Y_j)$ et sur les flèches reliant les variables latentes explicatives ξ_1 et ξ_2 à la variable latente ξ_3 les coefficients de régression b_{31} et b_{32}. On représente dans la figure 5 les pays dans le plan des variables Y_1, Y_2, en indiquant leur régime politique.

On déduit de ces résultats que l'instabilité politique dépend nettement plus du sous-développement industriel que de l'inégalité dans la répartition des terres. À l'exception de l'Inde toutes les démocraties stables ont un niveau industriel supérieur à la moyenne et le Venezuela est la seule dictature ayant un niveau industriel supérieur à la moyenne. Le cumul d'une grande inégalité dans la répartition des terres et d'un sous-développement industriel va de pair avec un régime dictatorial. Faute d'une vision plus précise sur le monde en 1961, nous n'allons pas plus loin dans les commentaires.

Remerciements

Cette recherche a été financée par la Fondation HEC. Je remercie Pierre Cazes, Carole Donada, Jacques Obadia et Pierre Valette-Florence pour leurs commentaires sur différents points de la première version du manuscrit. Ils m'ont aidé à mieux comprendre l'approche PLS et permis d'améliorer significativement la lisibilité de cet article aux niveaux théorique et pratique.

TABLEAU 3
Estimation des variables latentes

Pays	Y_1	Y_2	Y_3
Argentine	0.95	0.24	0.69
Australie	1.22	1.36	– 1.58
Autriche	0.39	0.25	– 0.48
Belgique	– 0.81	1.51	– 0.87
Bolivie	1.10	– 1.57	1.49
Brésil	0.77	– 0.65	0.23
Canada	– 1.52	1.66	– 0.96
Chili	1.22	– 0.32	– 0.01
Colombie	0.80	– 0.44	0.75
Costa-Rica	0.91	– 0.48	0.26
Cuba	0.74	– 0.18	1.67
Danemark	– 1.98	0.81	– 1.50
Rép. Dominicaine	0.71	– 0.72	0.55
Equateur	0.96	– 0.68	0.97
Egypte	0.45	– 1.07	0.88
Espagne	0.82	– 0.51	0.44
Etats-unis	0.19	1.99	– 0.95
Finlande	– 1.03	0.30	– 0.29
France	– 0.88	0.80	0.45
Guatémala	0.99	– 0.95	1.10
Grèce	0.62	– 0.52	0.03
Honduras	0.49	– 1.08	1.11
Inde	– 0.95	– 1.50	– 0.37
Irak	1.17	– 1.03	1.48
Irlande	– 1.06	0.06	– 1.08
Italie	0.71	0.22	0.24
Japon	– 1.79	– 0.38	0.12
Libye	– 0.09	– 1.42	0.26
Luxembourg	– 0.62	0.97	– 1.57
Nicaragua	0.41	– 0.75	0.66
Norvège	– 0.62	0.75	– 1.43
Nouvelle Zélande	0.42	1.28	– 1.57
Panama	0.21	– 0.39	1.02
Pays-Bas	– 0.74	1.23	– 1.34
Pérou	0.83	– 0.99	0.85
Philippines	– 0.75	– 0.77	1.03
Pologne	– 2.28	– 0.26	0.55
RFA	– 0.21	1.09	– 0.50
Royaume-Uni	0.13	2.03	– 1.05
Salvador	0.81	– 0.71	0.45
Sud Vietnam	0.02	– 1.09	1.37
Suède	– 0.85	1.39	– 1.58
Suisse	– 1.54	1.62	– 1.58
Taiwan	– 0.01	– 0.89	0.01
Uruguay	0.69	0.18	– 1.28
Venezuela	1.14	0.25	1.14
Yougoslavie	– 2.16	– 0.65	0.19

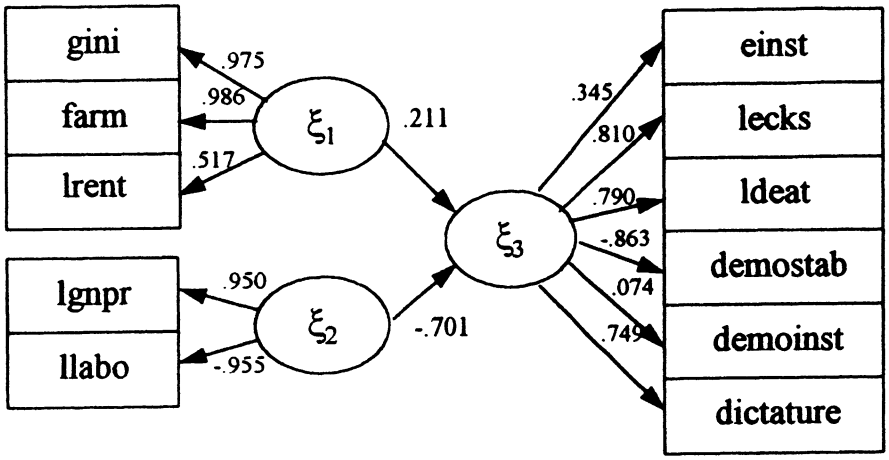
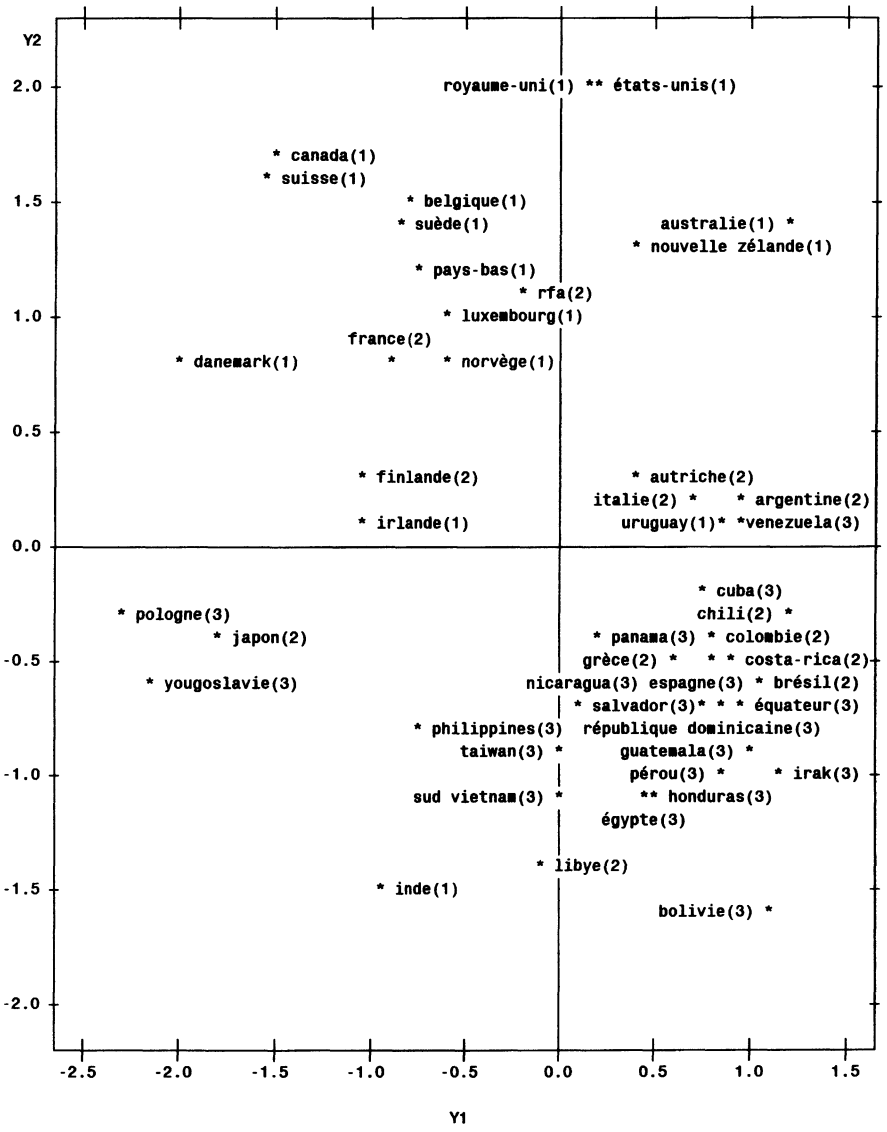


FIGURE 4
Réseau de causalité entre l'instabilité politique et l'inégalité économique



Y_1 = Inégalité dans la répartition des terres,

Y_2 = Développement industriel

Régime politique entre parenthèses : 1 = démocratie stable, 2 = démocratie instable, 3 = dictature

FIGURE 5

Carte des pays représentés dans le plan Y_1, Y_2

Références

[1] APEL H., WOLD H. (1982), «Soft modeling with latent variables in two or more dimensions : PLS estimation and testing for predictive relevance», in *System under indirect observation, vol. 2*, K.G. Jöreskog & H. Wold (Eds), North-Holland, Amsterdam, pp. 209–247.

[2] ARBUCKLE J.L. (1997), *Amos Users' Guide, Version 3.6*, SPSS Inc., Chicago.

[3] BOOKSTEIN F.L. (1982), «The geometric meaning of soft modeling, with some generalizations», in *System under indirect observation, vol. 2*, K.G. Jöreskog & H. Wold (Eds), North-Holland, Amsterdam, pp. 55–74.

[4] CAMO AS (1996), *The Unscrambler, The Unscrambler 6 users's guide*, Camo AS, Olav Tryggvasons gt. 24, N-7019 Trondheim, Norway.

[5] CARROLL J.D. (1968), «A generalization of canonical correlation analysis to three or more sets of variables», *Proc. 76^{th} Conv. Amer. Psych. Assoc.* pp. 227–228.

[6] FORNELL C., CHA J. (1994), «Partial Least Squares», in *Advanced methods in marketing reseach*, R.P. Bagozzi (Ed.), Blackwell, Cambridge, USA.

[7] GEISSER S. (1974), «A predictive approach to the random effect model», *Biometrika*, vol. 61, pp. 65–89.

[8] GIFI A. (1990), *Nonlinear Multivariate Analysis*, John Wiley & Sons, Chichester.

[9] HAYDUCK L.A. (1987), *Structural equation modeling with LISREL*. The John Hopkins University Press, Baltimore.

[10] HORST P. (1961), «Relations among m sets of variables», *Psychometrika*, vol. 26, pp. 129–149.

[11] DE JONG S. (1993a), «SIMPLS : An alternative approach to partial least squares regression», *Chemometrics and Intelligent Laboratory Systems*, vol. 18, pp. 251–263.

[12] DE JONG S. (1993b), «PLS fits closer than PCR», *Journal of Chemometrics*, vol. 7, pp. 551–557.

[13] DE JONG S. (1995), «PLS shrinks», *Journal of Chemometrics*, vol. 9, pp. 323–326.

[14] JÖRESKOG K.G. (1970), «A general method for analysis of covariance structure», *Biometrika*, vol. 57, pp. 239–251.

[15] JÖRESKOG K.G., SÖRBOM D. (1979), *Advances in factor analysis and structural equation models*. Abt Books, Cambridge.

[16] JÖRESKOG K.G., SÖRBOM D. (1984), *LISREL VI : Analysis of linear structural relationships by maximum likelihood, instrumental variables, and least squares methods*. Scientific Software, Mooresville.

[17] JÖRESKOG K.G., WOLD H. (1982), «The ML and PLS techniques for modeling with latent variables : historical and comparative aspects», in *System under indirect observation, vol. 1*, K.G. JÖRESKOG & H. Wold (Eds), North-Holland, Amsterdam, pp. 263–270.

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

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

[20] MARTENS H., NæS T. (1989), *Multivariate Calibration*. John Wiley & Sons, New York.

[21] NOONAN R. & WOLD H. (1982), «PLS path modeling with indirectly observed variables : a comparison of alternative estimates for the latent variable», in *System under indirect observation, vol. 2*, K.G. Jöreskog & H. Wold (Eds), North-Holland, Amsterdam, pp. 75–94.

[22] RUSSET B.M. (1964), «Inequality and instability», *World Politics*, vol. 21, pp. 442–454.

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

[24] SAS Institute Inc. (1996), *SAS/STAT Software, version 6.12, Proc PLS*, Cary, NC : SAS Institute Inc.

[25] STONE M. (1974), «Cross-validatory choice and assessment of statistical predictions», *Journal of the Royal Statistical Society, Series B*, vol. 36, pp. 111–133.

[26] TENENHAUS M. (1998), *La Régression PLS : Théorie et Pratique*. Technip, Paris.

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

[28] UMETRI AB (1996), *SIMCA-P for Windows, Graphical Software for Multivariate Process Modeling*, Umetri AB, Box 7960, S-90719 Umeå, Sweden.

[29] 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.

[30] 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.

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

[32] WOLD H. (1966), «Estimation of principal components and related models by iterative least squares», in *Multivariate Analysis*, Krishnaiah, P.R. (Ed.), Academic Press, New York, pp. 391–420.

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

[34] 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.

[35] 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.

[36] WOLD S. (1984), «Three PLS algorithms according to SW», in *Report from the symposium MULTDAST (multivariate data analysis in science and technology)*, Wold, S. (Ed.) Umeå, Sweden.

[37] WOLD S., MARTENS H., WOLD H. (1983), «The multivariate calibration problem in chemistry solved by the PLS method», in *Proc. Conf. Matrix Pencils*, Ruhe, A. & Kågstrøm, B. (Eds), March 1982, Lecture Notes in Mathematics, Springer Verlag, Heidelberg, pp. 286–293.