

REVUE DE STATISTIQUE APPLIQUÉE

X. BRY

Une méthode d'estimation empirique d'un modèle à variables latentes: l'analyse en composantes thématiques

Revue de statistique appliquée, tome 51, n° 2 (2003), p. 5-45

http://www.numdam.org/item?id=RSA_2003__51_2_5_0

© Société française de statistique, 2003, tous droits réservés.

L'accès aux archives de la revue « Revue de statistique appliquée » (<http://www.sfds.asso.fr/publicat/rsa.htm>) implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
<http://www.numdam.org/>

UNE MÉTHODE D'ESTIMATION EMPIRIQUE D'UN MODÈLE À VARIABLES LATENTES : L'ANALYSE EN COMPOSANTES THÉMATIQUES

X. BRY

Consultant

BP 5420, Dakar Fann, Sénégal. Email : bryxavier@yahoo.fr

RÉSUMÉ

On considère un groupe Z de variables à expliquer à partir de R groupes de prédicteurs $Y_1, \dots, Y_R$. On suppose que chaque groupe est produit par quelques variables latentes, et que les variables latentes de Z sont modélisables linéairement en fonction de celles des Y_r . Nous proposons ici une méthode d'estimation de ces variables latentes et du modèle qu'elles satisfont. Un lien est établi entre cette méthode et l'approche PLS.

Mots-clés : *Analyse Canonique, Analyse en Composantes Thématiques, Analyse en Résultats Covariantes, Approche PLS, Moindres Carrés Partiels, Régression PLS, Variables Latentes.*

ABSTRACT

Given a group of variables Z to be predicted from R groups of predictors $Y_1, \dots, Y_R$, we assume that each group is the result of fewer latent variables, and that those of Z can be modelled linearly using those of groups Y_r . We then propose a technique for estimating all latent variables together with the model they satisfy. A link is established between this technique and PLS Path Modelling.

Keywords : *Canonical Analysis, Covarying Resultant Analysis, Latent Variables, PLS Path Modelling, PLS Regression, Thematic Components Analysis.*

Notations

I_n désigne la matrice identité de taille n . Lorsque la dimension de la matrice est sans ambiguïté, elle sera simplement notée I .

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

Les minuscules grecques désignent des scalaires.

X, Y, Z désignent des tableaux matriciels décrivant H individus (en ligne) à l'aide de variables (en colonne). Ces lettres sont indifféremment utilisées pour désigner les groupes de variables en question.

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

Y et Z contiennent respectivement J et K variables. On note ainsi : $Y = (y^j)_{j=1 \text{ à } J}$ et $Z = (z^k)_{k=1 \text{ à } K}$.

M et N sont des matrices symétriques définies positives de tailles $(J \times J)$ et $(K \times K)$ pondérant les variables des groupes respectifs Y et Z .

Le sous-espace vectoriel de $\mathbb{R}^H$ engendré par un groupe de variables X sera noté $\langle X \rangle$.

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

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

Le projecteur sur un sous-espace E parallèlement à un supplémentaire F sera noté : Π_E^F

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

L'orthogonalité s'entend, sauf mention explicite contraire, par rapport à la métrique canonique I .

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

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

Quelques acronymes :

AC : Analyse Canonique

ACP : Analyse en Composantes Principales

ACT : Analyse en Composantes Thématiques

ARC : Analyse en Résultantes Covariantes

ARM : Analyse des Redondances Maximales

PLS : Partial Least Squares

RCP : Régression sur Composantes Principales

N.B. Les variables sont systématiquement supposées centrées.

Introduction

Modélisation d'un groupe de variables en fonction d'un autre - l'optique des Moindres Carrés Partiels :

Disposant d'un groupe Z de variables à expliquer (éventuellement réduit à une seule variable z) et d'un groupe Y de variables explicatives, on cherche une prédiction linéaire des variables de Z à partir de Y .

Lorsqu'on n'impose qu'une condition d'optimalité de l'ajustement telle que les moindres carrés ordinaires, on obtient la régression linéaire multiple. On peut toutefois demander à la prédiction d'utiliser des structures fortes des variables explicatives. On parie alors que la force de ces structures peut assurer à la prédiction une certaine robustesse, le bruit contenu dans les variables explicatives étant, dans une certaine mesure, exclu du mécanisme prédictif. On postule ainsi l'existence, à l'intérieur de Y , de variables latentes matérialisant les structures fortes les plus à même de prédire

Z . Ces variables latentes sont estimées par des facteurs, et la prédiction de Z est faite à partir de ces seuls facteurs, ce qui motive l'appellation de *moindres carrés «partiels»*. Exprimer ces facteurs en fonction des variables explicatives originelles, on peut reformuler le modèle comme combinaison linéaire de ces dernières. On obtient ainsi une régression *régularisée*¹ de Z sur Y .

Il existe plusieurs techniques d'extraction des facteurs explicatifs.

- On peut chercher les facteurs en fonction de leur seule force structurelle² dans Y . On utilise alors les premiers facteurs d'ACP de Y comme nouvelles variables explicatives. Cette approche conduit à la Régression sur Composantes Principales (RCP, cf. [5]).
- On peut vouloir orienter davantage les facteurs vers l'explication de Z . On demande alors aux facteurs de réaliser un compromis entre l'ACP de Y et l'ajustement au groupe Z . Selon le critère choisi, on obtient la régression PLS (cf. [7],[3]) ou l'Analyse en Résultats Covariantes (cf. [3]).

Modélisation faisant intervenir R groupes explicatifs :

On suppose ici disposer de R groupes explicatifs $Y_1, \dots, Y_R$, chacun correspondant à un thème, *i.e.* ayant une unité conceptuelle claire et un rôle propre dans l'explication de Z . Pour plus de simplicité, on considère momentanément que Z est réduit à une seule variable z . On cherche les dimensions des groupes explicatifs qui jouent un rôle dans la prédiction de z .

- Si l'on choisit de se passer de variables latentes, on procède, lorsqu'elle est techniquement possible, à la régression multiple de z sur Y . On note que la prédiction $\hat{z}$ obtenue peut être écrite comme somme de ses composantes $\hat{z}_r$ sur les différents groupes Y_r . Ces composantes peuvent être comparées entre elles, en norme et en direction. Le nombre et la redondance des variables d'un groupe Y_r n'ont ici d'effet qu'à travers la dimension du sous-espace engendré. Si celle-ci est assez élevée, la composante $\hat{z}_r$ pourra contribuer grandement à la prédiction de z . Si, dans le même temps, les structures fortes sous-jacentes de Y_r sont peu nombreuses, $\hat{z}_r$ sera fondée sur une grande part de bruit, donc sans robustesse.

- On peut *a contrario* chercher à fonder la prédiction sur des variables latentes représentant des structures fortes des groupes. On pourra alors utiliser les redondances (corrélations) internes aux groupes comme indications structurelles importantes, et tenir compte de la complémentarité entre eux pour la détermination des facteurs les plus utiles. Le bruit interne à chaque groupe sera ainsi quelque peu écarté de la modélisation et l'on cherchera, dans les différents groupes, des facteurs qui se combinent le mieux possible pour prédire z . Une telle approche est bien adaptée à la situation, fréquente dans l'approche exploratoire de la modélisation statistique, où l'analyste a construit un schéma conceptuel global indiquant la dépendance d'un thème Z relativement à plusieurs autres, mais où il ne dispose pas, pour formaliser cette dépendance, de relations fonctionnelles précises entre variables observées. Il

¹ On entend par là : assise sur les dimensions structurelles fortes du groupe prédictif plutôt que sur l'ensemble des dimensions, bruit compris.

² *i.e.* leur corrélation globale avec les variables de Y .

lui faut alors simultanément extraire l'information pertinente de chaque thème et le modèle reliant ces différentes informations. Une telle recherche impose de ne pas mélanger les thèmes explicatifs. On écartera donc les méthodes à deux groupes (PLS, ARM, AC...), qui, mélangeant les thèmes, fournissent des facteurs *a priori* hybrides.

L'approche PLS proposée par H. Wold (cf. [7], [9]) permet d'estimer une variable latente par groupe explicatif, en faisant certaines hypothèses sur le sens des liaisons entre ces variables et z . Nous proposerons une méthode qui s'affranchit de telles hypothèses et fournit une base de facteurs dits *thématiques* dans chaque groupe.

1. Quelques modèles simples fondés sur des variables latentes

Dans la situation générale que nous traiterons, R variables latentes agissent sur une variable latente ou observée. On parlera de modèle *multivarié* lorsque R est au moins égal à 2 (cf. figure 2), et de modèle *bivarié* lorsqu'il est réduit à 1 (cf. figure 1). Ces appellations font donc référence au nombre de variables *latentes* du modèle.

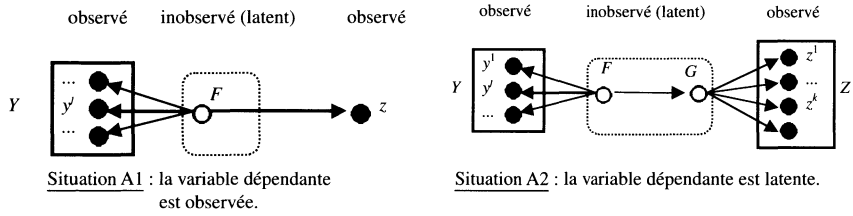


FIGURE 1
Modèles bivariés

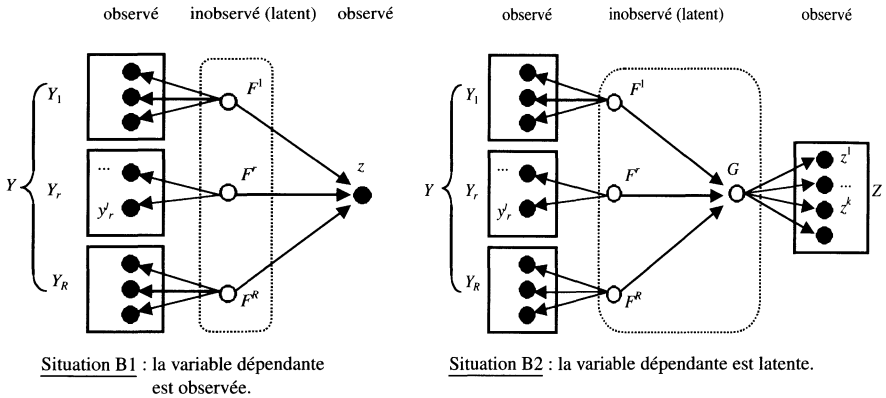


FIGURE 2
Modèles multivariés

Note : Un arc orienté d'une variable A vers une variable B signifie que A détermine B (au moins partiellement).

La modélisation multivariée cherchant à séparer les effets, sur la variable dépendante, de différents déterminants F^r , l'estimation de chaque variable explicative latente devra tenir compte de l'existence des autres. La technique d'estimation y sera donc moins simple que dans le cas bivarié.

- Dans les modèles illustrés ci-dessus, on n'a fait apparaître qu'une variable latente par groupe de variables. Cela peut éventuellement correspondre à l'hypothèse que la structure fondamentale de chaque groupe est unidimensionnelle. Cependant, dans une optique exploratoire et dans beaucoup de domaines d'application, cette hypothèse n'est pas naturelle. C'est le cas par exemple en sciences humaines, où les groupes thématiques sont le plus souvent clairement multidimensionnels. Il est alors essentiel d'identifier toutes les dimensions « utiles » de chaque groupe. Les procédures d'estimation doivent, dans ce cas, permettre une exploration multidimensionnelle de chaque groupe.

2. Traitement des modèles bivariés

On considère 2 groupes de variables Y et Z pondérés respectivement par les matrices M et N . Ces matrices servent à moduler la prise en compte de la structure de corrélations interne à chaque groupe dans la recherche des liaisons entre eux. Par exemple, pour tenir compte de la structure des corrélations internes à Y , on prendra $M = I$, tandis que pour gommer leur intervention, on prendra $M = (Y'Y)^{-1}$. De même pour Z et N .

2.1. Le groupe Z est réduit à une seule variable z (situation A1)

2.1.1. Programme commun à PLS et à la régression multiple

- Le programme à résoudre est :

$$\mathbf{Q1} : \underset{u'Mu=1}{\text{Max}} \langle YMu|z \rangle$$

- On note $F = YMu$. La solution de **Q1** est caractérisée par (cf. [3]) : $u = \eta Y'z$, tel que $u'Mu = 1$. Le facteur correspondant est alors : $F = \eta YMY'z$ (1).

- La variable $YMY'z$ a été appelée *résultante* de la variable z sur le groupe Y pondéré par M , et notée $R_Y^M z$. Outre que la direction de la résultante est solution de **Q1**, sa norme, lorsque toutes les variables de Y et z sont réduites, est une mesure de l'intensité de la liaison entre z et Y (cf. [3]). L'opérateur de résultante YMY' , qui est aussi l'opérateur d'Escoufier, est noté R_Y^M .

Lorsque $M = I$, on obtient la résultante *simple* de z sur Y : $R_Y z = YY'z$.

Lorsque l'on choisit $M = (Y'Y)^{-1}$ pour neutraliser les corrélations internes à Y , on obtient la régression de z sur Y : $\Pi_{(Y)} z = Y(Y'Y)^{-1}Y'z$.

2.1.2. Le programme de l'Analyse en Résultantes Covariantes

Les propriétés de direction et de norme de la résultante la rendent utilisable dans les programmes de recherche de liaisons entre groupes de variables. On a ainsi proposé, pour le présent problème, le programme alternatif suivant (cf. [3]) :

$$\mathbf{P1} : \text{Max}_{\|F\|=1} \langle R_Y^M F | z \rangle$$

La solution de **P1** est caractérisée par : $F \propto R_Y^M z$ avec $\|F\| = 1$ (cf. [3]). On retrouve, normée, la solution de **Q1**.

2.2. Cas général : Z contient plusieurs variables (Situation A2)

2.2.1. Programme commun à PLS, l'ARM et l'AC

• Les programmes fournissant les facteurs de rang 1 de PLS, de l'ARM et de l'AC sont des cas particuliers du programme **Q** suivant (cf. [3] [7]) :

$$\mathbf{Q}(Y, M; Z, N) : \text{Max}_{\substack{u'Mu=1 \\ v'Nv=1}} \langle YMu | ZNv \rangle$$

Le programme de PLS correspond au choix $M = I, N = I$; il utilise à la fois les corrélations internes de Y et celles de Z . Celui de l'ARM correspond à : $M = (Y'Y)^{-1}; N = I$; il n'utilise que les corrélations internes de Z , considérant Y comme sous-espace. Enfin, le programme de l'AC correspond à : $M = (Y'Y)^{-1}; N = (Z'Z)^{-1}$; il ne s'intéresse aux groupes Y et Z qu'à travers les sous-espaces $\langle Y \rangle$ et $\langle Z \rangle$ qu'ils engendrent.

Notation. – Le programme $\mathbf{Q}(Y, (Y'Y)^{-1}; Z, N)$ sera noté $\mathbf{Q}(\langle Y \rangle; Z, N)$ pour indiquer qu'il ne s'intéresse qu'au sous-espace engendré par Y .

• Solution du programme **Q** :

En posant $F = YMu$ et $G = ZNv$, la résolution conduit aux caractérisations équivalentes suivantes des solutions de rang 1 (cf. [3]) :

$$\text{Caractérisation (a)} : (2) R_Y^M G = \sqrt{\eta} F; \quad (2') R_Z^N F = \sqrt{\eta} G$$

$$\text{Caractérisation (b)} : (3) R_Y^M R_Z^N F = \eta F; \quad (3') R_Z^N R_Y^M G = \eta G$$

où η est la plus grande valeur propre des opérateurs $R_Z^N R_Y^M$ et $R_Y^M R_Z^N$.

- Deux propriétés s'avèreront utiles par la suite :

PROPRIÉTÉ P1. – Si $W \perp Y$ ($W'Y = 0$), alors le programme $Q(Y, M; (Z + W), N)$ a la même solution que $Q(Y, M; Z, N)$.

Preuve. $\langle YMu|(Z + W)Nv \rangle = \langle YMu|ZNv \rangle + \langle YMu|WNv \rangle$. Or $W'Y = 0$, donc $\langle YMu|WNv \rangle = v'NW'YMu = 0$.

PROPRIÉTÉ P2. – Soit $W \perp Y$ et soit N la matrice de pondération de (Z, W) . On écrira : $N = \begin{bmatrix} N_Z & N_{ZW} \\ N_{WZ} & N_W \end{bmatrix}$, où $\dim(N_Z) = (K, K)$. Si N est bloc-diagonale, de sorte que le bloc N_{ZW} correspondant au croisement de Z et W soit nul (c'est en particulier le cas pour $N = I$), alors le programme $Q(Y, M; (Z, W), N)$ a la même solution que $Q(Y, M; Z, N_Z)$.

Preuve. En effet, dans ce cas, $\langle YMu|(Z, W)Nv \rangle = \langle YMu|ZN_Zv_Z + WN_Wv_W \rangle = \langle YMu|ZN_Zv_Z \rangle$. La contrainte $\|v\|_N^2 = 1$ s'écrit : $\|v_Z\|_{N_Z}^2 + \|v_W\|_{N_W}^2 = 1$. On a donc à chercher $\max_{u'Mu=1; \|v_Z\|_{N_Z}^2 \leq 1} \langle YMu|ZN_Zv_Z \rangle$, qui est trivialement atteint pour $\|v_Z\|_{N_Z}^2 = 1$. Le programme équivaut donc à : $\max_{u'Mu=1; \|v_Z\|_{N_Z}^2=1} \langle YMu|ZN_Zv_Z \rangle$

- Facteurs de rang 2, etc. :

Une fois extrait le facteur de rang 1 de chaque groupe, la seconde étape consiste *grosso modo* à reprendre la recherche de facteurs sous certaines contraintes d'orthogonalité aux premiers facteurs. Les choses se passent différemment selon la méthode choisie, puisque suivant le choix de M et N , les solutions des équations fournies par les conditions du premier ordre de Q présentent ou non certaines orthogonalités.

Les conditions du premier ordre de la première étape de PLS, par exemple, fournissent un lot de facteurs explicatifs non orthogonaux. Comme on recherche des facteurs explicatifs non linéairement redondants, on choisit, pour calculer le deuxième facteur, de projeter le problème sur l'orthogonal du premier $F_1 = Yu_1$. Soient respectivement Y^* et Z^* les résidus de régression de Y et Z sur F_1 , on résout donc le programme $Q(Y^*, I; Z^*, I)$, qu'on notera plus simplement $Q(Y^*; Z^*)$. D'après la propriété P1, $Q(Y^*; Z^*) \iff Q(Y^*; Z)$. En effet, $Z = \hat{Z} + Z^*$ avec $\hat{Z} \propto F_1$, donc $\hat{Z} \perp Y^*$ et $Q(Y^*; Z^*) \iff Q(Y^*; \hat{Z} + Z^*)$.

Note. Il n'est pas équivalent (cf. annexe 1, propriété 1') de :

- i) résoudre le programme de PLS sur Y^* et Z (ou Z^*)
- ii) résoudre le programme de PLS sur Y et Z (ou Z^*) avec la contrainte $F \perp F_1$.

2.2.2. Programme de l'ARC

- Le programme proposé ici est :

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

- Solution du programme $\mathbf{P}$:

On parvient aisément aux deux caractérisations équivalentes suivantes des solutions (cf. [3]) :

$$\text{Caractérisation (a) : (4) } R_Y^M R_Z^N G = \sqrt{\eta} F = F; (4') R_Z^N R_Y^M F = \sqrt{\eta} G$$

$$\text{Caractérisation (b) : (5) } R_Y^M R_Z^N R_Z^N R_Y^M F = \eta F; (5') R_Z^N R_Y^M R_Y^M R_Z^N G = \eta G$$

avec η maximale.

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

Les équations (5) et (5') montrent, d'une part, que $F \in \langle Y \rangle$ et $G \in \langle Z \rangle$, et d'autre part, que les facteurs F (respectivement G), vecteurs propres d'une matrice symétrique, sont deux à deux orthogonaux.

Pour la recherche des facteurs de rang 2, on reprend le programme $\mathbf{P}$, en y ajoutant les contraintes $F \perp F_1, G \perp G_1$. On obtient alors le couple de rang 2 du classement précédent, etc.

• ARC itérative

On peut concevoir, lors de l'étape 2, la variante suivante : comme dans PLS, une fois trouvé le premier facteur F_1 , on projette le problème initial sur son orthogonal afin de déterminer le deuxième facteur explicatif F_2 , etc.

Note. Y^* et Z^* étant les tableaux associés respectivement à Y et Z après projection orthogonale sur l'hyperplan orthogonal à F_1 , il est équivalent (cf. annexe 1, propriété 2') de :

- i) résoudre le programme de l'ARC sur Y^* et Z^*
- ii) résoudre le programme de l'ARC sur Y et Z^* avec la contrainte $F \perp F_1$.

2.3. Modèles et pseudo-modèles

Chacune des méthodes ci-dessus exposées peut être utilisée pour construire une prédiction régularisée du groupe Z à partir du groupe Y . Une fois calculés les facteurs, les variables dépendantes sont régressées dessus (éventuellement sur les q premiers seulement), ce qui donne un modèle linéaire de chaque variable en fonction des facteurs : $z^k = \sum_{\alpha} a_{\alpha k} F_{\alpha} + e^k$, où e^k est un résidu orthogonal aux F_{α} .

Ensuite, chacun des facteurs F_α étant exprimé comme combinaison linéaire des variables y^j , l'équation ci-dessus conduit à une expression de z^k en fonction des y^j : $z^k = \sum_j b_{jk} y^j + e^k$.

On obtient une telle expression même dans le cas où les y^j présentent des multicollinéarités. On ne peut donc pas *a priori* interpréter le coefficient b_{jk} comme la variation de z^k consécutive à une variation unitaire de y^j les autres explicatives restant fixées, mais simplement comme une pondération de y^j dans un prédicteur de z^k . Aussi préférons-nous employer le terme de *pseudo-modèle* pour les combinaisons linéaires obtenues.

3. Estimation d'un modèle multivarié par l'approche PLS

L'approche PLS proposée par H. Wold (cf. [7] [9]) étend la méthode de recherche de variables latentes correspondant au programme **Q** au cas de L groupes de variables (dont les situations B1 et B2 font partie). Les solutions fournies par l'approche PLS coïncident avec celles de **Q1** et **Q** dans le cas particulier des situations bivariées A1 et A2.

On considère L groupes de variables $X_1, \dots, X_l, \dots, X_L$. On fait les hypothèses suivantes (cf. [7]) :

H1 : chaque groupe X_l est essentiellement unidimensionnel, c'est-à-dire est engendré par une variable latente unique F_l . Chaque variable du groupe peut ainsi s'écrire $x_l^k = a_l^k F_l + \varepsilon_l^k$, où ε_l^k est un bruit de moyenne nulle et non corrélé à F_l .

H2 : Les variables latentes sont liées par des relations structurelles de la forme : $F_l = \sum_{m \neq l} \beta_{ml} F_m + \omega_l$, où ω_l est non corrélée aux F_m du membre de droite. Certains des β_{ml} sont *a priori* supposés nuls³.

H3 : Le sens de la corrélation entre chaque variable latente dépendante F_l et chaque variable latente F_m figurant parmi ses déterminants est supposé connu. On notera c_{ml} le signe (± 1) de cette corrélation.

3.1. L'algorithme

Etape 0 (initialisation) :

Une valeur initiale $F_l(0)$ est déterminée pour chaque variable latente F_l , en l'égalant par exemple à une des variables de son groupe avec laquelle elle est positivement corrélée.

³ L'hypothèse H2 correspond au graphe des relations causales entre les variables latentes. Un coefficient β_{ml} non nul correspond à un arc causal orienté de F_m vers F_l .

Etape n :

Phase 1 (estimation externe de chaque variable) :

On calcule la résultante de chaque variable latente issue de l'étape $n - 1$ sur son groupe. Cette résultante, une fois normée, remplace l'ancienne valeur. On a donc remplacé $F_l(n - 1)$ par $\Phi_l = X_l X_l' F_l(n - 1)$ réduite.

Phase 2 (estimation interne de chaque variable) :

On pose : $F_l(n) = \sum_{m \neq l} c_{ml} \Phi_m$ normée.

Fin : on s'arrête lorsque les variables latentes obtenues sont assez stables.

3.2. Bilan

- Une interprétation heuristique : la phase 1 (estimation externe) de l'étape courante «cale» chaque variable latente sur une structure forte de son groupe. La phase 2 (estimation interne) recale chaque variable latente F_l sur les relations qu'elle présente avec les autres, en calculant une pseudo-résultante de F_l sur le groupe $\{F_m, m \neq l\}$ utilisant les signes des covariances entre variables plutôt que les covariances. Au point de convergence, l'estimation externe et l'estimation interne s'accordent en un sens; elles restent néanmoins distinctes, car l'estimation externe d'une variable latente F_l appartient au sous-espace engendré par le groupe X_l , ce qui n'est pas le cas *a priori* de son estimation interne.

- Si l'hypothèse unidimensionnelle H1 paraît contraignante⁴, elle n'est toutefois pas explicitement utilisée par l'algorithme et peut ainsi être relâchée. Après l'extraction d'une première variable latente dans un groupe, on pourrait y poursuivre la recherche d'autres variables latentes, sous contrainte d'orthogonalité à la première par exemple.

- H3 nous paraît être une hypothèse plus problématique dans la mesure où la corrélation entre F_m et F_l est une corrélation simple, et non une corrélation partielle. La première est symétrique et ne tient pas compte de l'intervention d'autres variables dans le modèle. La seconde, par contre, est orientée, et traduit explicitement l'effet d'une variation unitaire de F_m sur F_l , *les autres déterminants de F_l restant inchangés*. Les corrélations simple et partielle peuvent fort bien être de signes opposés. Une confusion entre elles est une confusion entre deux problématiques : la bivariée, qui ne s'intéresse qu'à la liaison statistique globale entre deux variables (tous autres déterminants de l'une ou de l'autre étant ignorés, ce qui en fait un outil peu adapté à l'exploration des causalités) et la problématique multivariée, qui cherche explicitement à séparer les effets des différents déterminants sur la variable dépendante.

⁴ Elle semble en outre rendre presque superflue la stratégie d'estimation proposée, dans la mesure où si l'unidimensionnalité d'un groupe est nette, sa première composante principale en estimera très bien l'unique variable latente.

Ajoutons à cela que dans la plupart des situations, on n'a pas d'idée *a priori* sur le sens de la liaison entre une variable dépendante et l'un de ses déterminants, ce sens dépendant des autres déterminants introduits dans le modèle.

Reprenant la formule d'estimation interne $F_l(n) = \sum_{m \neq l} c_{ml} \Phi_m$, Lohmöller (cf. [6], [8]) a proposé un choix des coefficients c_{ml} qui permet de se passer de toute hypothèse *a priori* sur le sens des liaisons :

- 1) lorsque la variable latente F_l est expliquée par les variables F_m , les c_{ml} seront pris égaux aux coefficients de régression de Φ_l sur les Φ_m ;
- 2) lorsque F_l est une variable explicative de F_m , c_{ml} sera pris égal à la corrélation $\rho(\Phi_l, \Phi_m)$.

Supprimant toute contrainte *a priori* sur le sens des liaisons, cette stratégie assouplit notablement l'estimation des variables latentes.

En ce qui concerne le respect du schéma multivarié reliant les variables latentes, la stratégie proposée nous paraît discutable : le choix 1 semble judicieux car il correspond à une logique multivariée; mais le choix 2 est toujours axé sur une liaison bivariée, la corrélation simple. Or, il nous semble que les phases d'estimation externe et interne que l'on tente de faire converger ont chacune leur logique propre :

- l'estimation externe, cherchant à caler les variables latentes sur des structures de corrélation fortes des groupes, utilise naturellement la redondance bivariée entre les variables du groupe;
- l'estimation interne, au contraire, cherche à caler les variables latentes sur un modèle linéaire explicatif, c'est-à-dire à séparer les effets des groupes explicatifs sur le groupe à expliquer; elle ressortit donc à une logique multivariée.

Plus loin, nous proposons une procédure d'estimation interne des variables latentes qui s'inscrit dans une problématique exclusivement multivariée (cf. § 6.1.).

• Il a été noté dans [8] que le mode d'estimation interne semblait avoir peu d'influence sur l'estimation externe fournissant les facteurs.

Le fait nous semble tout à fait prévisible lorsque l'hypothèse unidimensionnelle est bien vérifiée, parce que l'estimation de la variable latente a alors très peu de latitude.

En présence de groupes vraiment multidimensionnels, par contre, la proximité des résultats n'est nullement assurée. En effet, lorsque les variables explicatives d'un modèle ne sont pas orthogonales deux à deux, on sait que les schémas bivariés et multivariés conduisent à des estimations divergentes. Il s'ensuit que la recherche des dimensions « utiles » pour l'un des schémas ne donnera pas *a priori* les mêmes que pour l'autre. Une illustration de cette divergence est donnée au § 6.1.

4. Estimation du modèle multivarié par l'Analyse en Composantes Thématiques

Nous allons ici construire une autre extension des programmes **Q** et **P** aux situations B1 et B2 qui soit davantage fondée sur la problématique multivariée de séparation des effets que l'approche PLS.

Nous allons également relâcher totalement les contraintes H1 et H3. La méthode proposée ne supposera aucune connaissance *a priori* sur les données en dehors du schéma causal.

Le principe de la méthode est simple. Considérons la situation B1. On projette la variable dépendante z sur le sous-espace engendré par l'ensemble Y des variables explicatives. La prédiction obtenue $\hat{z}$ peut être exprimée comme somme de composantes $\hat{z}_r$ appartenant aux sous-espaces $\langle Y_r \rangle$. Chacune de ces composantes $\hat{z}_r$ est la projection de z sur $\langle Y_r \rangle$ parallèlement aux autres composantes. La méthode que nous proposons utilise ce fait, en alternance avec un calcul de résultante, pour rapprocher chaque composante des structures fortes de son groupe. Ce rapprochement n'est pas fait isolément, mais en fonction des valeurs courantes des autres composantes, de sorte qu'au point de convergence, les composantes obtenues dans les différents groupes sont co-déterminées.

4.1. Préliminaires

4.1.1. Projecteur sur un sous-espace $\langle W_1 \rangle$ parallèlement à un sous-espace supplémentaire $\langle W_2 \rangle$

L'espace E étant décomposé comme $\langle W_1 \rangle \oplus \langle W_2 \rangle$, on cherche l'expression de la projection, notée Π_1^2 , d'un vecteur x sur $\langle W_1 \rangle$ parallèlement à $\langle W_2 \rangle$.

Si l'on note Π_2 le projecteur orthogonal sur $\langle W_2 \rangle^\perp$, on a $\Pi_2 = (I - W_2(W_2'W_2)^{-1}W_2')$, et Π_1^2x est défini par :

$$\Pi_1^2x \in \langle W_1 \rangle \iff \Pi_1^2x = W_1b \quad \text{et} \quad (x - \Pi_1^2x) \in \langle W_2 \rangle \iff \Pi_2(x - \Pi_1^2x) = 0.$$

Des équations précédentes on tire :

$$\Pi_2(x - W_1b) = 0 \iff \Pi_2x = \Pi_2W_1b$$

Π_2W_1 est de plein rang en colonne, dès que $W_1 \cap W_2 = \{0\}$ (en effet, $\langle \Pi_2W_1 \rangle$ est de dimension inférieure à celle de $\langle W_1 \rangle$ si et seulement si $\langle W_1 \rangle$ contient une dimension également présente dans $\langle W_2 \rangle$).

Dès lors, on peut écrire :

$$\begin{aligned} \Pi_2W_1b = \Pi_2x &\implies W_1'\Pi_2W_1b = W_1'\Pi_2x \iff b = (W_1'\Pi_2W_1)^{-1}W_1'\Pi_2x \\ &\implies \Pi_1^2x = W_1b = W_1(W_1'\Pi_2W_1)^{-1}W_1'\Pi_2x \end{aligned}$$

Soit :

$$\Pi_1^2 = W_1(W_1'\Pi_2W_1)^{-1}W_1'\Pi_2$$

Remarque. – Ainsi, ce projecteur peut être vu comme un projecteur orthogonal au sens de la métrique Π_2 .

4.1.2. Régression multiple

Considérons la situation B1. En l'absence de multicollinéarité entre les sous-groupes Y_r , la régression de z sur Y permet d'écrire de façon unique : $z = \hat{z} + \hat{u}$, avec : $\hat{u} \perp \langle Y \rangle$; $\hat{z} = \sum_{r=1}^R \hat{z}_r$, où $\forall r, \hat{z}_r \in \langle Y_r \rangle$.

Le vecteur $\hat{z}_r$ est la projection de $\hat{z}$ sur $\langle Y_r \rangle$ parallèlement au sous-espace engendré par les autres composantes : $D_s = \sum_{s \neq r} \langle \hat{z}_s \rangle$. Or, d'après le § 4.1.1, on peut exprimer ce projecteur comme un projecteur Π_{E_r} -orthogonal, où Π_{E_r} est le projecteur orthogonal sur $E_r = (D_r)^\perp \cap \langle Y \rangle$. On a ainsi :

$$\hat{z}_r = \Pi_{\langle Y_r \rangle}^{D_r} z = Y_r(Y_r'\Pi_{E_r}Y_r)^{-1}Y_r'\Pi_{E_r}z.$$

On remarque que $\hat{z}_r$ est aussi la projection orthogonale de z elle-même sur $\langle Y_r \rangle$ au sens de Π_{E_r} :

$$\begin{aligned} E_r \subset \langle Y \rangle &\implies \Pi_{E_r}\hat{u} = 0 \\ &\implies Y_r(Y_r'\Pi_{E_r}Y_r)^{-1}Y_r'\Pi_{E_r}(\hat{z} + \hat{u}) = Y_r(Y_r'\Pi_{E_r}Y_r)^{-1}Y_r'\Pi_{E_r}\hat{z} \end{aligned}$$

4.2. Situation B1 : la variable dépendante est observée

On cherche, comme dans la régression multiple, à décomposer partiellement z en composantes complémentaires appartenant aux sous-espaces $\langle Y_r \rangle$, chaque composante devant par ailleurs s'appuyer sur des structures fortes de $\langle Y_r \rangle$. Comme nous n'avons pas fait d'hypothèse d'unidimensionnalité des groupes Y_r , chaque composante pourra être construite à partir de plusieurs facteurs. Chaque facteur devrait être un compromis entre :

- l'exigence de complémentarité avec ceux des autres groupes pour la prédiction de z . Il faut donc que les facteurs du groupe Y_r permettent d'approcher la composante $\hat{z}_r$ de $\hat{z}$ sur $\langle Y_r \rangle$;
- l'exigence de « force structurelle », *i.e.* de représentativité du facteur par rapport à la structure interne de son groupe.

4.2.1. Une première extension de Q1 : l'ACT1

• Au groupe Y_r est associée une matrice de pondération des variables M_r . Sous la contrainte $u'_r M_r u_r = 1$, la norme carrée usuelle (variance) du facteur $F^r = Y_r M_r u_r$ mesurera la force structurelle du facteur dans le groupe Y_r .

• Nous allons chercher à résoudre un programme du type :

$\text{Max}_{u'_r M_r u_r = 1} \langle Y_r M_r u_r | z \rangle_C$ où C est une matrice convenablement choisie pour que soit vérifiée la propriété suivante, assurant la continuité avec la régression multiple :

PROPRIÉTÉ P3. – Si le groupe Y_r n'a pas de structure (ou si l'on ne tient pas compte de sa structure), la résolution du programme doit fournir pour facteur F^r la composante partielle sur Y_r de la régression multiple de z sur $\langle Y_r, F^{-r} \rangle$, où F^{-r} désigne l'ensemble des facteurs obtenus dans les autres groupes.

Ne pas tenir compte de la structure de Y_r revient à prendre $M_r = (Y'_r Y_r)^{-1}$. D'autre part, comme $\langle Y_r M_r u_r | z \rangle_C = \langle Y_r M_r u_r | Cz \rangle$, le programme ci-dessus est un cas du programme **Q1**, et sa solution s'en déduit :

$$F^r = Y_r M_r u_r \propto Y_r M_r Y'_r C z = Y_r (Y'_r Y_r)^{-1} Y'_r C z \iff F^r = \Pi_{\langle Y_r \rangle} C z \quad (a)$$

$$\text{On doit par ailleurs avoir : } F^r = \Pi_{\langle Y_r \rangle}^{\langle F^{-r} \rangle} z \quad (b)$$

De (a) et (b), on déduit alors :

$$\Pi_{\langle Y_r \rangle} C z = \Pi_{\langle Y_r \rangle}^{\langle F^{-r} \rangle} z \implies C z = \Pi_{\langle Y_r \rangle}^{\langle F^{-r} \rangle} z + t, \text{ où } t \perp \langle Y_r \rangle.$$

On remarque enfin que le vecteur t n'a en réalité aucune importance, puisque :

$$\begin{aligned} \langle Y_r M_r u_r | t \rangle &= 0 \implies \langle Y_r M_r u_r | C z \rangle = \langle Y_r M_r u_r | \Pi_{\langle Y_r \rangle}^{\langle F^{-r} \rangle} z + t \rangle \\ &= \langle Y_r M_r u_r | \Pi_{\langle Y_r \rangle}^{\langle F^{-r} \rangle} z \rangle \end{aligned}$$

On peut donc par exemple prendre $t = 0$, ce qui donne $C = \Pi_{\langle Y_r \rangle}^{\langle F^{-r} \rangle}$. On aboutit ainsi au programme :

$$\text{Max}_{u'_r M_r u_r = 1} \langle Y_r M_r u_r | \Pi_{\langle Y_r \rangle}^{\langle F^{-r} \rangle} z \rangle = \mathbf{Q1}(Y_r, M_r; \Pi_{\langle Y_r \rangle}^{\langle F^{-r} \rangle} z)$$

Résolution du programme :

La solution se déduit de celle de **Q1** : $F^r \propto R_{Y_r}^{M_r} \Pi_{\langle Y_r \rangle}^{\langle F^{-r} \rangle} z$

Le facteur cherché est la résultante sur Y_r (pondéré par M_r) de la projection de z sur $\langle Y_r \rangle$ parallèlement aux autres facteurs.

Nous exploitons ci-dessous cette propriété pour déterminer itérativement les facteurs thématiques de rang 1 les uns par rapport aux autres. Schématiquement, on détermine les valeurs initiales des facteurs en résolvant le programme **Q1** sur z et chacun des Y_r pris isolément. Puis, à l'étape courante, on projette z sur chaque $\langle Y_r \rangle$ parallèlement aux facteurs des autres groupes, et on calcule la résultante sur Y_r de cette projection, que l'on norme⁵, ce qui fournit la nouvelle valeur du facteur.

Pour les facteurs de rang deux et plus, on procède de même, mais en remplaçant chaque Y_r par ses résidus de régression sur les facteurs de rang antérieur du groupe, et en incluant systématiquement dans F^{-r} tous les facteurs de rang antérieur de tous les groupes.

On obtient l'algorithme suivant :

Algorithme de l'ACT1 (algorithme A)

Etape 1 (facteurs de rang 1) :

Itération 0 (initialisation) :

– On détermine une valeur initiale $F_1^r(0)$ du premier facteur de chaque groupe Y_r , e.g. en effectuant **Q1** entre z et Y_r , ce qui donne : $F_1^r(0) = R_{Y_r}^{M_r} z$ normée.

Itération $k + 1$, $k \geq 0$:

– Pour chaque groupe Y_r la direction du facteur $F_1^r(k + 1)$ est déterminée en résolvant : **Q1**($Y_r, M_r; C_1^r(k)z$), avec $C_1^r(k) = \Pi_{\langle Y_r \rangle}^{(F_1^{-r}(k))}$, ce qui donne : $F_1^r(k + 1) = R_{Y_r}^{M_r} \Pi_{\langle Y_r \rangle}^{(F_1^{-r}(k))} z$ normé.

Pratiquement, on régresse z sur Y_r et $F_1^{-r}(k)$, on prend la composante sur $\langle Y_r \rangle$ de la partie prédite et on calcule sa résultante sur le groupe, que l'on norme.

– Si les facteurs trouvés à l'itération $k + 1$ sont suffisamment proches de ceux fournis par l'itération k , on s'arrête. On note $F_1^1, \dots, F_1^R$ les facteurs de rang 1 finaux.

Etape $n \geq 2$ (facteurs de rang n) :

Cette étape a pour but de trouver, dans chaque groupe où subsistent des dimensions non prises en compte, un facteur orthogonal à ceux déjà trouvés dans les étapes précédentes. Pour alléger les notations, on fera ici comme si l'on cherchait des dimensions supplémentaires dans tous les groupes. On procède comme suit :

– On projette chaque groupe Y_r dans lequel on cherche un nouveau facteur sur l'orthogonal des facteurs qu'on y a trouvé jusque là. Le nouveau facteur est en effet voulu non redondant avec les précédents. On note Y_r^* les résidus de régression de Y_r sur $\{F_1^1, \dots, F_{n-1}^1\}$. On note par ailleurs F_{n-1} l'ensemble des facteurs obtenus dans tous les groupes lors des étapes précédentes, soit :

$$F_{n-1} = \{F_1^1, \dots, F_1^R, F_2^1, \dots, F_2^R, \dots, F_{n-1}^1, \dots, F_{n-1}^R\}.$$

⁵ Le fait de normer les facteurs n'est utile qu'à la comparaison des valeurs successives, pour déterminer le point d'arrêt.

Itération 0 (initialisation) :

– On détermine une valeur initiale $F_n^r(0)$ du facteur de rang n de chaque groupe Y_r^* retenu, par exemple en résolvant le programme **Q1**($Y_r^*, M_r; C_n^r(0)z$) où : $C_n^r(0) = \Pi_{\langle Y_r^* \rangle}^{\langle F_{n-1}^r \rangle}$, ce qui donne : $F_n^r(0) = Y_r^* M_r Y_r^{*'} \Pi_{\langle Y_r^* \rangle}^{\langle F_{n-1}^r \rangle}$ normé. Pratiquement, on régresse z sur Y_r^* et F_{n-1} , on prend la composante sur $\langle Y_r^* \rangle$ de la partie prédite et on calcule sa résultante sur le groupe, que l'on norme.

Itération $k + 1$, $k \geq 0$:

$F_n^{-r}(k)$ désigne l'ensemble des facteurs obtenus à l'itération k , à l'exception du r -ième, soit : $\{F_n^1(k), \dots, F_n^{r-1}(k), F_n^{r+1}(k), \dots, F_n^R(k)\}$.

– Pour chaque groupe Y_r , le facteur $F_n^r(k+1)$ est calculé en résolvant : **Q1**($Y_r^*, M_r; C_n^r(k)z$), où $C_n^r(k) = \Pi_{\langle Y_r^* \rangle}^{\langle F_{n-1}^r, F_n^{-r}(k) \rangle}$, ce qui donne : $F_n^r(k+1) = R_{Y_r^*}^{M_r} \Pi_{\langle Y_r^* \rangle}^{\langle F_{n-1}^r, F_n^{-r}(k) \rangle} z$ normé. Pratiquement, on régresse z sur Y_r^* , F_{n-1} et $F_n^{-r}(k)$, on prend la composante sur $\langle Y_r^* \rangle$ de la partie prédite et on calcule sa résultante sur le groupe, que l'on norme.

– Si les facteurs trouvés à l'itération $k+1$ sont suffisamment proches de ceux fournis par l'itération k , on s'arrête. On note $F_n^1, \dots, F_n^R$ les facteurs finaux.

Notes.

- A l'étape courante, il n'est pas nécessaire d'utiliser le résidu de régression de z sur les facteurs trouvés dans les étapes précédentes, puisque toutes les projections de z effectuées à l'étape courante sont faites parallèlement à ces facteurs.
- On pouvait aussi bien utiliser **P1** à la place de **Q1** dans l'étape courante, ces programmes ayant des solutions colinéaires.
- L'algorithme **A** sera utilisé par d'autres méthodes, moyennant un changement de matrice C – voire le remplacement de **Q1** par un autre programme – dans l'étape courante.

• Comportement de la méthode dans diverses situations

a) Dimension commune à deux groupes (multicolinéarité) :

Plusieurs cas doivent être distingués, selon l'importance structurelle de la dimension commune dans chacun des groupes.

a1) La dimension commune est résiduelle dans les deux groupes (on parlera de *multicolinéarité résiduelle*)

Cette dimension ne constitue jamais un facteur au cours des étapes de l'algorithme, et ne pose donc pas de problème.

a2) La dimension commune est résiduelle dans l'un des groupes, et constitue une structure forte de l'autre (on parlera de *multicolinéarité faible*).

Considérons la figure 3. Les flèches a et c (respectivement b et d) représentent des sous-ensembles de variables du groupe Y_1 (resp. Y_2). Les flèches épaisses (b et

c) représentent les structures fortes, et les minces (a et d) les structures résiduelles de chacun des groupes. La direction de c , notée δ , est identique à celle de d , et donc commune aux deux sous-espaces $\langle Y_1 \rangle$ et $\langle Y_2 \rangle$. Les directions a , b et c sont supposées orthogonales entre elles. Enfin, la variable z est très bien prédite par le sous-espace $\langle Y_2 \rangle$.

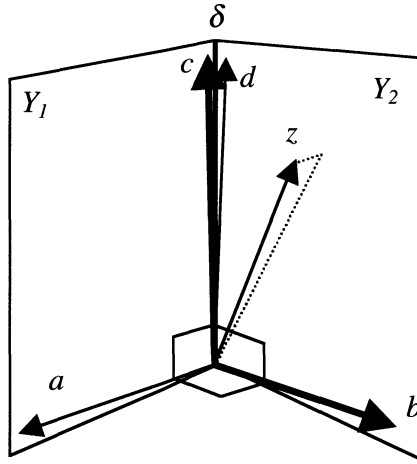


FIGURE 3
Multicolinéarité faible

A l'initialisation (Q1 de z sur les deux groupes séparément), on obtient pour Y_1 et Y_2 respectivement les facteurs c et b . On constate que l'itération suivante, concernant Y_2 , commence par la projection de z sur $\langle Y_2 \rangle$ parallèlement à c . Comme $c \in \langle Y_2 \rangle$, cette projection n'existe pas.

Nous devons ici faire face à un problème théorique⁶. Dans la mesure où la dimension commune est résiduelle dans le groupe Y_2 , on aurait aimé que la méthode fût insensible à son existence. Nous pouvons provisoirement contourner l'obstacle en débruitant au préalable chaque groupe. Nous proposons plus loin une autre adaptation de l'ACT non équivalente à celle-ci, mais insensible aux multicolinéarités faibles.

a3) La dimension commune est une structure forte dans chacun des groupes (multicolinéarité forte, cf. fig. 4).

Le même problème se pose évidemment que dans le cas d'une multicolinéarité faible, mais ne peut être contourné par débruitage préalable des groupes. La difficulté est toutefois ici légitime : une confusion entre structures de deux groupes doit entraîner l'impossibilité de séparer leurs effets sur z . Le contraire eût été suspect.

⁶ Sa portée pratique est toutefois limitée, puisque si le nombre des variables libres d'un groupe n'excède pas $(H - 1)$, dimension de l'espace engendré par toutes les variables centrées, il y a très peu de chances que le sous-espace engendré par ce groupe contienne *exactement* le facteur d'un autre groupe.

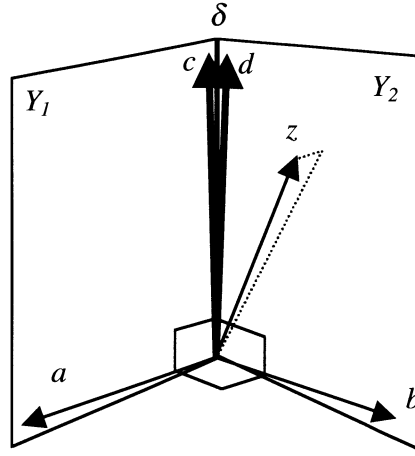


FIGURE 4
Multicolinéarité forte

b) Si les groupes explicatifs sont 2 à 2 orthogonaux, l'ACTI équivaut à QI séparément sur chaque groupe.

En effet, $\forall s \neq r, F_n^s \perp \langle Y_r \rangle$. La projection de z sur $\langle Y_r \rangle$ parallèlement aux facteurs des autres groupes est donc la projection orthogonale sur $\langle Y_r \rangle$.

Or :

$$\langle Y_r M_r u | \Pi_{\langle Y_r \rangle} z \rangle = \langle Y_r M_r u | \Pi_{\langle Y_r \rangle} z + \Pi_{\langle Y_r \rangle^\perp} z \rangle = \langle Y_r M_r u | z \rangle$$

On est donc à chaque étape ramené au programme **Q1** séparément sur chaque groupe.

c) Si chaque groupe est réduit à une variable, on retrouve la régression ordinaire sur Y .

En effet, dès la première itération, le facteur correspondant à chaque groupe n'est autre que la seule variable s'y trouvant. La seconde itération fournit la composante partielle sur Y_r de la régression de z sur Y .

d) Dans les groupes, les variables qui sont orthogonales au vrai modèle de z sont éliminées par la méthode.

Examinons le cas simple et immédiatement généralisable de deux groupes (cf. figure 5). A, B et C désignent des sous-espaces. On suppose $A \subset \langle Y_1 \rangle, B \subset \langle Y_2 \rangle, C = (A + B)^\perp \cap (\langle Y_1 \rangle \cup \langle Y_2 \rangle)$ et $z = a + b$, où $a \in A$ et $b \in B$. On a alors $z \perp C$.

L'étape 0 (**Q1** séparément de z sur Y_1 et sur Y_2) élimine C d'après la propriété P2 (cf. § 2.2.1.). Les déterminations initiales des facteurs de rang 1 se trouvent dans A et B respectivement.

Les projections de z sur le sous-espace d'un groupe se font parallèlement à des directions de $A + B$, donc orthogonalement à C . Par conséquent, les déterminations

ultérieures des facteurs de rang 1 demeurent dans A et B . Il en va de même pour les facteurs de rang supérieur.

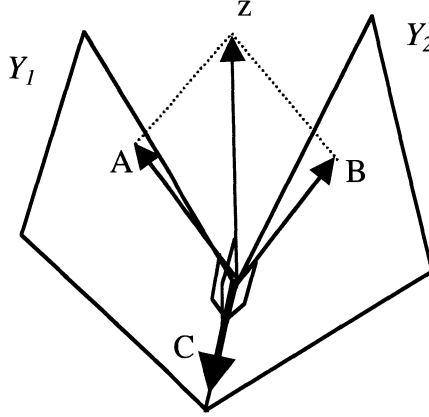


FIGURE 5

Variables orthogonales au modèle

4.2.2. Une adaptation de Q1 qui résiste aux multicollinéarités faibles : PACT1 (Pseudo-ACT1)

On a montré, dans le § 4.1.1., que la projection de z sur $\langle W_1 \rangle$ parallèlement à un sous-espace supplémentaire $\langle W_2 \rangle$ est sa projection orthogonale sur $\langle W_1 \rangle$ au sens de la métrique $\Pi_{\langle W_2 \rangle^\perp}$. Elle maximise donc $\cos(z, W_1)$ au sens de la métrique $\Pi_{\langle W_2 \rangle^\perp}$. Ainsi, d'après le § 4.1.2., dans la régression de z sur Y , chaque composante $\hat{z}_r$ maximise le cosinus de l'angle fait par z et $\langle Y_r \rangle$ au sens de Π_{E_r} .

Pour tenir compte, à présent, de la structure interne de Y_r , le programme $\text{Max}_{F \in \langle Y_r \rangle} \cos_{\Pi_{E_r}}(F, z)$ peut être remplacé par :

$$\text{Max}_{u' M_r u = 1} \langle Y_r M_r u | z \rangle_{\Pi_{E_r}} \iff \text{Max}_{u' M_r u = 1} \left(\|Y_r M_r u\|_{\Pi_{E_r}} \cos_{\Pi_{E_r}}(z, Y_r M_r u) \right)$$

La norme $\|Y_r M_r u\|_{\Pi_{E_r}}$, sous la contrainte $u' M_r u = 1$, mesure ici la force structurelle du facteur F dans Y_r , mais elle le fait dans un sens particulier. En effet, $\|Y_r M_r u\|_{\Pi_{E_r}}^2 = \|\Pi_{E_r} Y_r M_r u\|^2$. Or, $\Pi_{E_r} Y_r$ représente les résidus de régression du groupe Y_r sur le sous-espace $D_r = \sum_{s \neq r} \langle \hat{z}_s \rangle$. Il s'agit de la partie du groupe Y_r ne

possédant aucune redondance linéaire avec les composantes trouvées dans les autres groupes. Sous la contrainte $\|u\|_{M_r} = 1$, $\|\Pi_{E_r} Y_r M_r u\|^2$ représente donc l'inertie, sur l'axe $\langle u \rangle$, des individus décrits par $\Pi_{E_r} Y_r$.

• L'algorithme PACT1

On utilise l'algorithme A de l'ACT1, en y prenant :

$$C_1^r(k) = \Pi_{\langle F_1^{-r}(k) \rangle^\perp}; \quad C_n^r(0) = \Pi_{\langle F_{n-1} \rangle^\perp}; \quad C_n^r(k) = \Pi_{\langle F_{n-1}, F_n^{-r}(k) \rangle^\perp}$$

• **Comportement de la méthode dans des situations particulières :**

a) *Dimension commune à deux groupes (multicolinéarité).*

a1) La multicolinéarité résiduelle ne pose pas plus de problème que dans **ACT1**.

a2) Multicolinéarité faible :

Considérons à nouveau la figure 3. A l'initialisation (**Q1** de z sur les deux groupes séparément), on obtient pour Y_1 et Y_2 respectivement les facteurs c et b . L'étape suivante consiste à appliquer d'une part **Q1** à Y_1 et $\Pi_{\langle b \rangle^\perp} z$, ce qui redonne c , et d'autre part **Q1** à Y_2 et $\Pi_{\langle c \rangle^\perp} z$, ce qui redonne b . On retrouve l'étape 0. Les facteurs sont donc stables. On modélise alors z en fonction de c et b .

a3) Multicolinéarité forte :

Considérons la figure 4. L'initialisation donne c ($= d$) pour facteur de chaque groupe. La seconde itération donne respectivement a et b (chacun des groupes est projeté sur $\langle c \rangle^\perp$). La troisième itération redonne c . On se trouve donc dans un cycle, et il n'y a pas convergence des solutions. Ce phénomène est bien compréhensible : l'«hésitation» de la méthode provient de l'inséparabilité des structures fortes des deux groupes.

b) *Si les groupes explicatifs sont 2 à 2 orthogonaux, PACT1 équivaut à Q1 séparément sur chaque groupe*

En effet,

$$\begin{aligned} (\forall s \neq r, F_n^s \perp \langle Y_r \rangle) &\implies \langle F_n^{-r}(k) \rangle^\perp \supset \langle Y_r \rangle \supset \langle Y_r^* \rangle \implies \Pi_{\langle F_n^{-r}(k) \rangle^\perp} Y_r^* = Y_r^* \\ &\implies \langle Y_r^* M_r u | \Pi_{\langle F_n^{-r}(k) \rangle^\perp} z \rangle = \langle \Pi_{\langle F_n^{-r}(k) \rangle^\perp} Y_r^* M_r u | z \rangle = \langle Y_r^* M_r u | z \rangle \end{aligned}$$

c) *Si chaque groupe est réduit à une variable, on retrouve la régression ordinaire sur Y*

La situation est identique à celle de l'ACT1.

d) *Dans les groupes, les variables qui sont orthogonales au vrai modèle de z sont éliminées par la méthode.*

Reprenons la figure 5. L'étape 0 est la même que celle de l'ACT1. On obtient donc : $F_1^1(0) = a \subset A$ et $F_1^2(0) = b \subset B$. Par suite : $\langle F_1^{-1}(0) \rangle^\perp = \langle F_1^2(0) \rangle^\perp = \langle b \rangle^\perp$, donc $\Pi_{\langle F_1^{-1}(0) \rangle^\perp} z = \Pi_{\langle b \rangle^\perp} (a + b) = \Pi_{\langle b \rangle^\perp} a$. Or, la projection orthogonale sur $\langle b \rangle^\perp$ se fait parallèlement à $\langle b \rangle$, donc $\Pi_{\langle b \rangle^\perp} a$ reste orthogonal à C . De même, on obtient l'orthogonalité de $\Pi_{\langle F_1^{-2}(0) \rangle^\perp} z$ à $\hat{C}$. Les programmes **Q1** déterminant les valeurs des facteurs à l'itération 1 se trouvent ainsi cantonnés à $C^\perp$. Ce schéma se reproduit exactement aux itérations suivantes.

• **Bilan**

Un prix assez fort doit être payé pour la résistance aux multicolinéarités faibles, puisque PACT1 perd la propriété P3 assurant la continuité avec la régression multiple.

En effet, le programme $\text{Max}_{u_r' M_r u_r = 1} \langle Y_r M_r u_r | \Pi_{(F-r)^\perp} z \rangle$, lorsque $M_r = (Y_r' Y_r)^{-1}$, a pour solution $Y_r M_r u_r = \Pi_{(Y_r)} \Pi_{(F-r)^\perp} z$, qui est en général différent de $\Pi_{(Y_r)}^{(F-r)} z$.

Les propriétés autres que P3 sont conservées.

4.2.3. Exemple d'application : détermination du prix de 21 périodiques

On a relevé les caractéristiques de base de 21 publications périodiques (données en annexe 2). Ces caractéristiques sont ainsi organisées : outre le prix du numéro, on a un groupe de caractéristiques physiques (poids, nombre de pages, surface, type de papier, mode d'impression couleur, nombre de numéros par an) et un groupe de caractéristiques de contenu (domaine et pourcentages de surface occupés par différents types de rubriques). Les groupes des caractéristiques physiques et de contenu sont considérés comme potentiellement explicatifs du prix. On désire extraire des composantes physiques (suffixées par *a*) et des composantes de contenu (suffixées par *b*) susceptibles d'expliquer une part suffisante des variations de prix.

Les caractéristiques nominales ont été codées par l'ensemble de leurs indicatrices. On a ainsi obtenu 9 variables physiques et 16 variables de contenu, soit 25 variables explicatives. Le centrage des variables les place toutes dans un sous-espace de dimension 20 de $\mathbb{R}^{21}$. Les multicolinéarités sont nombreuses. En effet, il y a d'une part une multicolinéarité par caractéristique nominale. On remarquera d'autre part que si l'on enlève une modalité pour chacune des 4 caractéristiques nominales, il reste 21 variables, soit une de plus que nécessaire pour engendrer un sous-espace de dimension 20.

Dès le départ, le prix est parfaitement exprimable (et d'une infinité de manières) comme combinaison linéaire des caractéristiques retenues, sans que cela révèle *a priori* le moindre rapport entre les deux. La régression multiple est impossible à réaliser directement, et de ce fait la séparation des rôles entre caractéristiques physiques et de contenu.

Nous avons appliqué à ces données deux méthodes de régression du prix sur facteurs thématiques : la RCP et l'ACT. La RCP utilise les facteurs obtenus par ACP séparées des thèmes, calculés indépendamment du prix. L'ACT fournit des facteurs orientés vers l'explication du prix. On comparera les résultats donnés par chacune des méthodes pour un même nombre de facteurs thématiques.

a) Régression sur Composantes Principales des groupes

• ACP des groupes⁷

Facteur physiques :

Inertie des facteurs physiques : Fa1 : 60.1% Fa2 : 15.3%

Le premier facteur physique de l'ACP, captant une part importante de l'inertie du groupe, oppose les périodiques de faible poids et de forme peu sophistiquée (papier journal, imprimés en noir et blanc ou bichromie) aux autres (poids et nombre de pages importants, papier glacé, quadrichromie). Le second facteur, captant une part d'inertie

⁷ On a omis les graphiques pour des raisons d'encombrement. Seule la description sommaire des facteurs est donnée.

beaucoup plus faible et médiocrement illustré, distingue l'impression monochrome des autres modes, et reprend l'opposition entre papier journal et papier glacé.

Facteurs de contenu :

Inertie des facteurs de contenu : Fb1 : 24.9% Fb2 : 18.3%

L'inertie captée par les deux premiers facteurs de contenu est assez faible. Le premier distingue les numéros Hors-Série et/ou les publications de vulgarisation scientifique et/ou les périodiques dominés par les articles de fond. Le second facteur oppose les caractéristiques de contenu du domaine : *vie courante* (informations pratiques et commerciales) à celles des domaines : *arts & littérature et information & opinion* (informations brèves, interviews). La configuration du nuage est triangulaire, et l'interprétation globale du plan est préférable à l'interprétation séparée des facteurs. Logiquement, dans la modélisation du prix, les facteurs Fb1 et Fb2 devraient intervenir conjointement.

• *Régression du prix sur les facteurs d'ACP :*

On utilise d'abord le premier facteur de chaque groupe.

PSEUDO-MODÈLE 1.

Exogènes :	Fa1,	Fb1	→	$R^2 = 0.6397$
Coefficients :	0.64	-0.41		
P.critique :	0.00	0.01		

La part de variance expliquée par Fa1 et Fa2 est assez bonne (64%). Chacun de ces facteurs joue son rôle; nous avons toutefois vu que l'interprétation du premier facteur de contenu est trop partielle pour être très intéressante.

Introduisons à présent le second facteur de chaque groupe.

PSEUDO-MODÈLE 2.

Exogènes :	Fa1,	Fa2,	Fb1,	Fb2	→	$R^2 = 0.6676$
Coefficients :	0.68	0.14	-0.41	0.09		
P.critique :	0.00	0.34	0.01	0.58		

Le gain de variance expliquée est très faible. Les probabilités critiques associées aux coefficients des exogènes indiquent que les facteurs de rang 2 introduits n'ont pas de rôle important à jouer dans la modélisation.

b) Analyse en Composantes Thématiques

• *ACTI*

On a pris ici M_1 et M_2 égales à l'identité. La convergence a été jugée suffisante (variation inférieure à 0.1%) après 4 itérations.

Facteurs physiques :

Inertie des facteurs physiques : Fa1 : 60% Fa2 : 12%

La corrélation entre premiers facteurs physiques (Fa1) de l'ACP et l'ACT est de 0.99. Le facteur Fa1 est le même dans les deux analyses. Ce n'est pas le cas des autres facteurs.

La figure 7 montre que le second facteur physique dégagé par l'ACT est, contrairement à celui de l'ACP, davantage illustré par le poids et la périodicité que par les modes d'impression couleur.

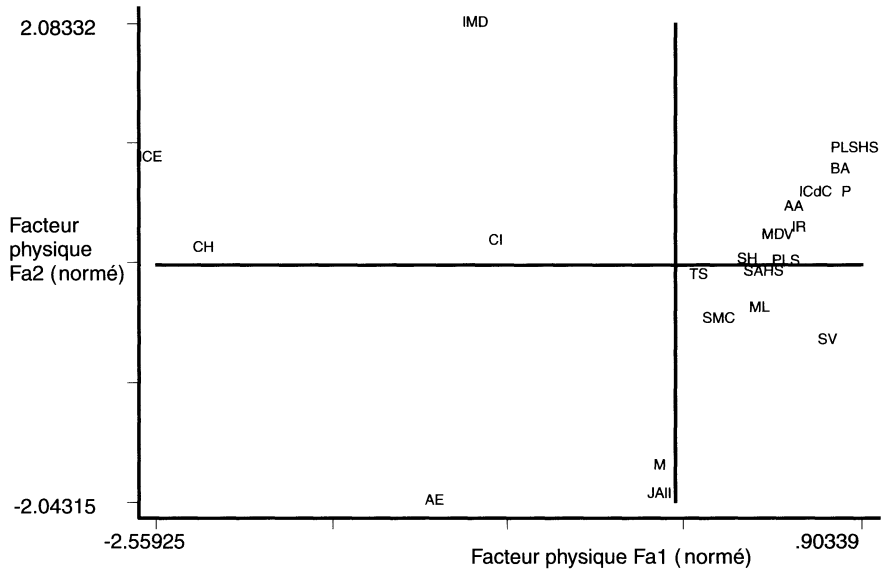


FIGURE 6

Plan (1,2) de l'ACT; individus selon les facteurs physiques du prix

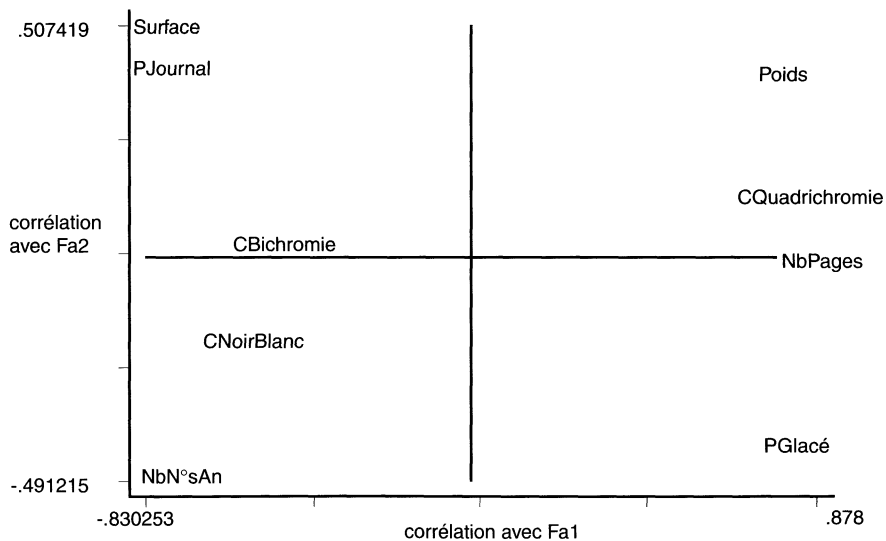


FIGURE 7

Plan (1,2) de l'ACT; corrélations entre variables physiques et facteurs physiques du prix

Facteurs de contenu :

Inertie des facteurs de contenu : Fb1 : 18% Fb2 : 17%

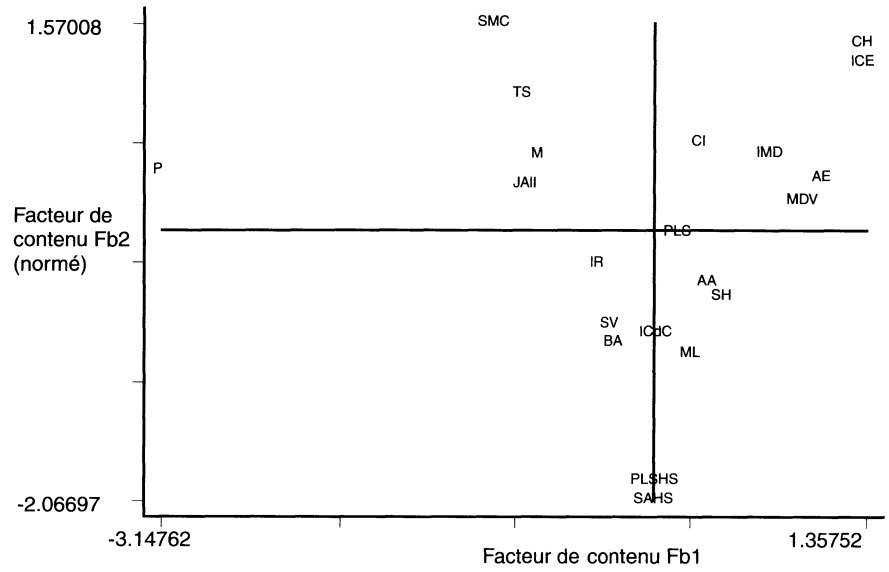


FIGURE 8

Plan (1,2) de l'ACT; individus selon les facteurs de contenu du prix

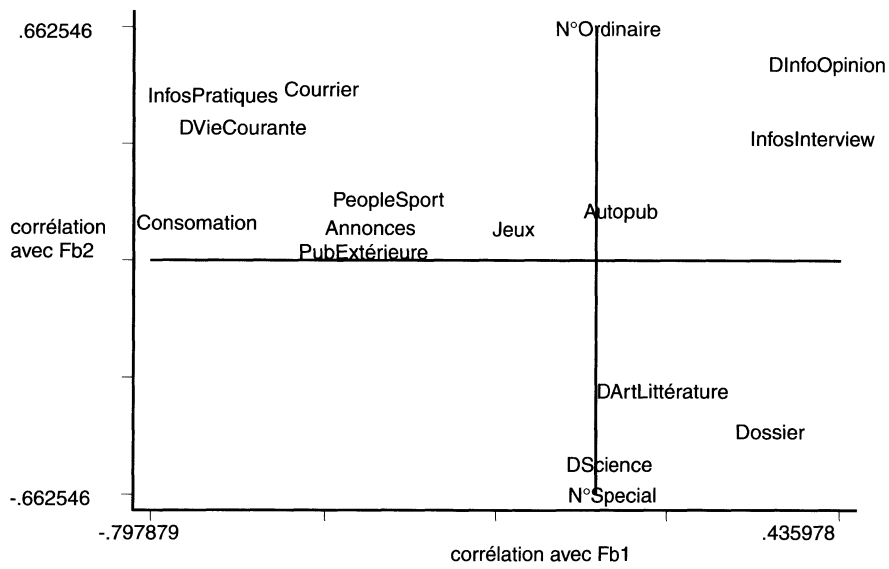


FIGURE 9

Plan (1,2) de l'ACT; corrélations entre variables de contenu et facteurs de contenu du prix

Les facteurs de contenu captent, comme en ACP, une inertie modeste. La figure 9 permet d'interpréter Fb1 globalement comme un facteur consumériste (les valeurs négatives traduisant une plus forte orientation vers la consommation). Il est particulièrement bien illustré par la part des rubriques de consommation, mais également négativement corrélé à la part de la publicité extérieure et des petites annonces. Le plan (Fb1,Fb2) discrimine assez bien les différents domaines de publication.

• *Régression du prix :*

PSEUDO-MODÈLE 1.

Exogènes :	Fa1,	Fb1	→	$R^2 = 0.77$
Coefficients :	1.04	0.62		
P.critique * :	0.00	0.00		

La part de variance du prix expliquée par les facteurs Fa1 et Fb1 (77%) est meilleure que dans la RCP (64%). Les rôles de ces facteurs sont nets. On peut utiliser les renseignements fournis par cette exploration pour construire un premier modèle du prix fondé sur deux exogènes. On prend la variable exogène la plus illustrative de chacun des facteurs : le nombre de pages pour Fa1, et la part de l'information de consommation pour Fb1.

Modèle 1.

Exogènes :	Nb pages,	% Consommation	→	$R^2 = 0.57$
Coefficients :	0.73	-0.58		
P.critique :	0.00	0.00		

La part de variance expliquée par ces deux seules variables (57%) s'avère relativement satisfaisante.

On introduit ensuite les facteurs de rang 2 dans la modélisation. La régression du prix sur Fa1, Fa2, Fb1, Fb2 permet d'expliquer 85% de la variance du prix. Cependant, le rôle de Fb2 ne paraît pas si important (probabilité critique* = 0.08). Comme son interprétation n'est pas d'une grande force, il est loisible de l'abandonner.

PSEUDO-MODÈLE 2.

Exogènes :	Fa1,	Fa2,	Fb1	→	$R^2 = 0.82$
Coefficients :	1.01	0.22	0.58		
P.critique * :	0.00	0.05	0.00		

La part de variance expliquée par ce pseudo-modèle (82%) est notablement meilleure que celle du pseudo-modèle à 4 facteurs de la RCP (67%).

Pour tirer un modèle du prix utilisant un petit nombre des variables originelles, on prendra deux variables fortement illustratives du plan (Fa1, Fa2) et peu corrélées

* Le calcul des facteurs de l'ACT impliquant la variable endogène, la probabilité critique n'est en toute rigueur pas utilisable pour l'inférence dans ce type de pseudo-modèle. On ne la fait figurer ici qu'à titre descriptif, comme indicateur de l'importance de la liaison marginale entre le facteur et l'endogène.

entre elles : le poids et le papier journal, ainsi que la variable la plus corrélée avec Fb1 : la part des rubriques de consommation.

Modèle 2.

Exogènes :	Poids,	Papier journal,	% Consommation	→	$R^2 = 0.80$
Coefficients :	0.65	-0.34	-0.57		
P.critique :	0.00	0.01	0.00		

L'ACT nous a conduit à voir le prix comme étant principalement déterminé par deux facteurs physiques et un facteur de contenu, nous menant très rapidement à un modèle parcimonieux (3 variables), efficace (80% de variance expliquée) et d'interprétation très claire.

4.3. Situation B2 : la variable dépendante est latente

4.3.1. Adaptation utilisant le programme Q (méthode Q-ACT)

- Outre la recherche des facteurs explicatifs latents, la méthode Q-ACT recherchée devra cette fois opérer une réduction dimensionnelle de Z , en y cherchant des facteurs expliqués G structurellement représentatifs. On remarque que le simple remplacement, dans Q-ACT1, des programmes de types $Q1(.; Cz)$ par $Q(.; CZ, N)$ ne fournit pas de facteurs de Z . En effet, les programmes Q n'y utilisent que des projections partielles de Z sur les $\langle Y_r \rangle$, et non Z lui-même.

- On remarque par ailleurs que la propriété P3 (cf. § 4.2.1.) n'a plus guère de signification pour le programme Q , le groupe Z contenant plusieurs variables. Il devient nécessaire d'étendre cette propriété. On demandera ainsi à l'étape courante de Q-ACT, si le groupe thématique Y_r dont on veut extraire un facteur est privé de structure, de conduire à la même solution que si l'on travaillait globalement avec le sous-espace engendré par Y_r et les autres facteurs thématiques (F^{-r}). Soit, plus formellement :

PROPRIÉTÉ P3'. – Si le groupe Y_r n'a pas de structure (ou si l'on ne tient pas compte de celle-ci), Q-ACT doit donner à chaque étape, dans $\langle Y_r \rangle$, la composante sur Y_r de la solution F du programme $Q(\langle Y_r, F^{-r} \rangle; Z, N)$.

Remarque 1. – $Q(\langle Y_r, F^{-r} \rangle; Z, N)$ considérant le groupe (Y_r, F^{-r}) comme sous-espace, le facteur F solution doit maximiser $\cos(F, G)$; par conséquent, on a $F \propto \Pi_{\langle Y_r, F^{-r} \rangle} G$, et la composante de F sur $\langle Y_r \rangle$ est, en direction, la même que celle de G .

Remarque 2. – Lorsque $N = (Z'Z)^{-1}$, et que les groupes thématiques sont sans structure interne, la propriété P3' assure que Q-ACT revient à faire l'Analyse Canonique entre Y et Z . Dans le cas plus particulier encore où Z code une variable qualitative à l'aide de ses indicatrices, on a le même résultat concernant l'Analyse Factorielle Discriminante.

- On construit ci-dessous une extension de Q vérifiant P3', en s'appuyant sur deux propriétés :

- L'extension Q-ACT1 de $Q1$ vérifie P3.

– En itérant correctement **Q1**, on obtient la solution de **Q**. En effet, les relations (2) et (2') vérifiées par le couple de facteurs (F, G) solution de **Q** $(Y, M; Z, N)$ montrent que F est solution de **Q1** $(Y, M; G)$ et que G est solution de **Q1** $(Z, N; F)$.

Algorithme de Q-ACT (algorithme B)

Etape 1 (facteurs de rang 1) :

Itération 0 (initialisation) :

– On détermine une valeur initiale $F_1^r(0)$ du premier facteur de chaque groupe Y_r , e.g. en normant la solution F de **Q** $(Y_r, M_r; Z, N)$. Le groupe des facteurs 1 obtenus est noté $F_1(0)$.

– On détermine un facteur $G_1(0)$ dans le groupe Z en effectuant **Q** $(\langle F_1(0) \rangle; Z, N)$.

Itération $k + 1$, $k \geq 0$:

– Pour chaque groupe Y_r la valeur courante $F_1^r(k+1)$ du 1^{er} facteur est calculée en normant la solution F de : **Q1** $(Y_r, M_r; C_1^r(k)G_1(k))$, où : $C_1^r(k) = \Pi_{\langle Y_r \rangle}^{\langle F_1^{r-k}(k) \rangle}$.

On a donc $F_1^r(k+1) = R_{Y_r}^{M_r} \Pi_{\langle Y_r \rangle}^{\langle F_1^{r-k}(k) \rangle} G_1(k)$ normé. Pratiquement, on régresse $G_1(k)$ sur Y_r et $F_1^{r-k}(k)$, on prend la composante sur $\langle Y_r \rangle$ de la partie prédite, on en calcule la résultante sur Y_r , et on la norme.

– On détermine le facteur $G_1(k+1)$ dans le groupe Z en normant la solution G de : **Q** $(\langle F_1(k) \rangle; Z, N)$, où $F_1(k)$ désigne l'ensemble des facteurs $F_1^r(k)$, $1 \leq r \leq R$.

– Si les facteurs trouvés à l'itération $k + 1$ sont suffisamment proches de ceux fournis par l'itération k , on s'arrête. On note $F_1^1, \dots, F_1^R$ et G_1 les facteurs finaux.

Etape $n \geq 2$ (facteurs de rang n) :

– On régresse Z sur $\{G_1, \dots, G_{n-1}\}$ ainsi que chaque groupe Y_r sur $\{F_1^r, \dots, F_{n-1}^r\}$. On note : $Z^* = \Pi_{\langle G_1, \dots, G_{n-1} \rangle}^\perp Z$ et $Y_r^* = \Pi_{\langle F_1^r, \dots, F_{n-1}^r \rangle}^\perp Y_r$ les résidus de ces régressions⁸. On note par ailleurs F_{n-1} l'ensemble des facteurs obtenus dans tous les groupes lors des étapes précédentes, soit :

$$F_{n-1} = \{F_1^1, \dots, F_1^R, F_2^1, \dots, F_2^R, \dots, F_{n-1}^1, \dots, F_{n-1}^R\}.$$

Itération 0 (initialisation) :

– On détermine un facteur $F_n^r(0)$ dans chaque groupe Y_r^* retenu, e.g. en normant la solution F de **Q** $(Y_r^*, M_r; C_n^r(0)Z^*)$ où : $C_n^r(0) = \Pi_{\langle Y_r^* \rangle}^{\langle F_{n-1}^r \rangle}$. Le groupe formé par les facteurs obtenus est noté $F_n(0)$. On pose $\bar{F}_n(0) = F_n(0) \cup F_{n-1}$.

⁸ Variantes :

1) Si l'on cherche à pousser jusqu'au bout la modélisation d'une *unique* variable latente de Z , on continuera à travailler avec Z à la place de Z^* dans l'étape courante.

2) On peut aussi concevoir ici, dans une optique purement explicative, de régresser Z sur l'ensemble des facteurs F précédemment trouvés. On perd alors l'orthogonalité des facteurs G , et de ce fait la possibilité de représenter facilement les z^k dans la base des G .

– On détermine un facteur $G_n(0)$ dans le groupe Z^* en normant la solution G de $\mathbf{Q}(\langle \overline{F}_n(0) \rangle; Z^*, N)$.

Itération $k + 1$, $k \geq 0$:

$F_n^{-r}(k)$ désigne l'ensemble des facteurs obtenus à l'itération k , à l'exception du r -ième, soit : $\{F_n^1(k), \dots, F_n^{r-1}(k), F_n^{r+1}(k), \dots, F_n^R(k)\}$.

– Pour chaque groupe Y_r , le facteur $F_n^r(k + 1)$ est calculé en normant la solution F de : $\mathbf{Q1}(Y_r^*, M_r; C_n^r(k)G_n(k))$, où $C_n^r(k) = \Pi_{\langle Y_r^* \rangle}^{\langle F_{n-1}^{-r}, F_n^{-r}(k) \rangle}$. On a donc $F_n^r(k + 1) = R_{Y_r^*}^{M_r} \Pi_{\langle Y_r^* \rangle}^{\langle F_{n-1}^{-r}, F_n^{-r}(k) \rangle} G_n(k)$ normé. Pratiquement, on régresse $G_n(k)$ sur Y_r^* , F_{n-1} et $F_n^{-r}(k)$, on prend la composante sur $\langle Y_r^* \rangle$ de la partie prédite, on en calcule la résultante sur Y_r^* , et on la norme.

– On détermine le facteur $G_n(k + 1)$ dans le groupe Z en normant la solution G de $\mathbf{Q}(\langle \overline{F}_n(k) \rangle; Z, N)$, où $\overline{F}_n(k) = F_n(k) \cup F_{n-1}$.

– Si les facteurs trouvés à l'itération $k + 1$ sont suffisamment proches de ceux fournis par l'itération k , on s'arrête. On note $F_n^1, \dots, F_n^R$ et G_n les facteurs finaux.

• **Q-ACT vérifie la propriété P3' :**

Preuve. A l'itération $k + 1$ de l'étape 1, le facteur $F_1^r(k + 1)$ du groupe Y_r est calculé en résolvant : $\mathbf{Q1}(Y_r, M_r; C_1^r(k)G_1(k))$ où : $C_1^r(k) = \Pi_{\langle Y_r \rangle}^{\langle F_1^{r-r}(k) \rangle}$ et $M_r = (Y_r' Y_r)^{-1}$. P3 étant alors vérifiée, $F_1^r(k + 1)$ est la composante sur Y_r de $\Pi_{\langle Y_r, F^{-r}(k) \rangle} G_1(k)$. Au point de convergence, F_1^r sera donc la composante sur Y_r de $\Pi_{\langle Y_r, F^{-r} \rangle} G_1$. D'autre part, G_1 sera la solution de $\mathbf{Q}(\langle F_1 \rangle; Z, N)$, par conséquent sera caractérisé par : $R_Z^N \Pi_{\langle F_1 \rangle} G_1 = \eta G_1$.

Or, $\Pi_{\langle F_1 \rangle} G_1 = \Pi_{\langle F_1^r, F_1^{-r} \rangle} G_1$, et comme F_1^r est la composante sur Y_r de $\Pi_{\langle Y_r, F^{-r} \rangle} G_1$, soit : $F_1^r \propto \Pi_{\langle Y_r \rangle}^{F^{-r}} \Pi_{\langle Y_r, F^{-r} \rangle} G_1$ (que l'on notera plus simplement Γ_1^r), on a : $\Pi_{\langle F_1^r, F_1^{-r} \rangle} G_1 = \Pi_{\langle \Gamma_1^r, F_1^{-r} \rangle} G_1 = \Pi_{\langle Y_r, F_1^{-r} \rangle} G_1$. Donc, la caractérisation ci-dessus équivaut à : $R_Z^N \Pi_{\langle Y_r, F^{-r} \rangle} G_1 = \eta G_1$ qui est celle du facteur G solution de $\mathbf{Q}(\langle Y_r, F^{-r} \rangle; Z, N)$.

Il en va de même à l'itération $k + 1$ de l'étape n .

Remarque. – Une version Q-PACT résistant aux multicolinéarités faibles peut être construite en procédant dans B aux mêmes substitutions de matrices C qu'entre Q-ACT1 et Q-PACT1.

4.3.2. Adaptation utilisant le programme P (méthode P-ACT)

• Raisonnant comme pour $\mathbf{Q}$, on note que les solutions (F, G) de $\mathbf{P}(Y, M; Z, N)$ peuvent être obtenues par itération des équations (4) et (4'). Par conséquent, F sera solution de $\mathbf{P1} = \mathbf{Q1}(Y, M; R_Z^N G)$. On adapte donc l'algorithme de Q-ACT de la façon suivante : dans l'algorithme B , on remplacera les programmes de type $\mathbf{Q}$ par leur homologue de type $\mathbf{P}$ et les programmes de type $\mathbf{Q1}(V, A; CG)$ par $\mathbf{P1} = \mathbf{Q1}(V, A; CR_Z^N G)$.

• La propriété $P3'$ n'est plus respectée, mais une autre extension de $P3$, notée $P3''$, l'est :

PROPRIÉTÉ $P3''$. Si le groupe Y_r n'a pas de structure (ou si l'on ne tient pas compte de celle-ci), P -ACT doit donner à chaque étape, dans $\langle Y_r \rangle$, la composante partielle sur Y_r de $R_Z^N G$ parallèlement à F^{-r} , où G est la solution du programme : $P(\langle Y_r, F^{-r} \rangle; Z, N)$.

La preuve est identique à celle de la propriété $P3'$ pour Q -ACT, à condition d'y remplacer G_1 par $R_Z^N G_1$.

Remarque 1. – Il est important de noter que dans le cas où $N = (Z'Z)^{-1}$, $R_Z^N G$ étant égale à G , la propriété $P3''$ équivaut strictement à $P3'$.

Remarque 2. – Une version P -PACT résistant aux multicollinéarités faibles peut être construite en procédant dans B aux mêmes substitutions de matrices C qu'entre Q -ACT1 et Q -PACT1.

4.3.3 Exemple d'application : données de Russet sur l'instabilité politique

• Les données utilisées ici sont présentées dans [7] et [8], où elles sont traitées par l'approche PLS. Brièvement, elles concernent des pays observés entre 1945 et 1962. Il s'agissait pour Russet d'expliquer l'instabilité politique à partir de l'inégalité économique et du niveau de développement.

Le groupe à expliquer concerne l'instabilité politique et contient les variables suivantes :

- *exinst* est un indice d'instabilité fonction du nombre de responsables du pouvoir exécutif et du nombre d'années pendant lesquelles le pays a été indépendant entre 1945 et 1961.
- *Lecks* = $\text{Ln}(ecks+1)$ où *ecks* (indice d'Eckstein) est le nombre de conflits violents entre communautés sur la période 1946-1961.
- *Ldeat* = $\text{Ln}(deat+1)$ où *deat* est le nombre de personnes tuées lors de manifestations violentes sur la période 1950-1962.
- *demosta*, *demoins* et *dictat*, qui sont les indicatrices respectives des régimes suivants : démocratie stable, démocratie instable et dictature.

Le premier groupe explicatif (inégalité économique) est ainsi composé :

- *gini* = indice de concentration de Gini
- *farm* = pourcentage de fermiers possédant la moitié des terres, en commençant par les petites surfaces.
- *Lrent* = $\text{Ln}(rent+1)$ où *rent* = pourcentage de fermiers locataires de leurs terres.

Le second groupe explicatif (développement) comprend :

- *Lgnpr* = $\text{Ln}(gnpr)$, où *gnpr* = PNB par tête en \$ en 1955.
- *Llabo* = $\text{Ln}(labo)$, où *labo* = pourcentage de personnes actives travaillant dans l'agriculture.

• Le traitement par Q-ACT (utilisant des matrices identité pour M_1 , M_2 et N) a fourni les résultats suivants :

- Trois itérations ont été suffisantes.
- Les facteurs de rang 1 de l'ACT (notés respectivement Fa1, Fb1 et G1) sont très proches des variables latentes estimées par l'approche PLS (notées respectivement fa1, fb1 et g1), comme en témoignent les corrélations suivantes :

$$\text{corr}(\text{Fa1}, \text{fa1}) = 0.9903 \quad \text{corr}(\text{Fb1}, \text{fb1}) = -1.0000 \quad \text{corr}(\text{G1}, \text{g1}) = 0.9491$$

Rappelons que les facteurs de l'ACT ont été obtenus sans utiliser d'hypothèse *a priori* sur le sens des liaisons entre les variables latentes. L'interprétation des facteurs de l'ACT se fait *a posteriori*, en fonction de leurs corrélations avec les variables du groupe. Ainsi, Fa1 est un facteur d'inégalité économique, Fa2 un facteur anticorrélé au développement économique, et G1 un facteur d'instabilité politique.

- Le modèle estimé entre facteurs de rang 1 est :

$$\begin{array}{lcl} \text{Modèle :} & G1 & = \quad 0.20\text{Fa1} \quad + \quad 0.72\text{Fb1} \quad \implies R^2 = 0.634 \\ & (\text{Probabilités critiques}) & \quad (0.04) \quad \quad (0.00) \end{array}$$

La qualité d'ajustement du modèle fourni par l'ACT (0.634) est pratiquement identique à celle de l'approche PLS (0.623).

• L'ACT fournit des facteurs de rang 2 : Fa2 et Fb2 dans les deux groupes explicatifs, et G2 dans le groupe à expliquer. Les facteurs Fa2 et Fb2 s'avèrent n'avoir pas de rôle important dans la prévision de G1. Le facteur G2 discrimine les démocraties stables des instables, et il est lié au seul facteur Fa2 (positivement corrélé avec *Lrent*, mais mal illustré par elle).

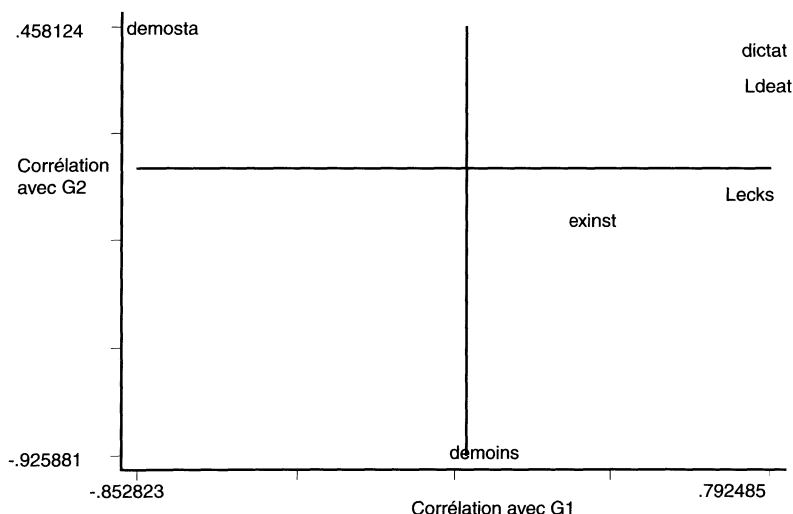


FIGURE 10

Caractéristiques d'instabilité politique dans le plan (G1, G2)

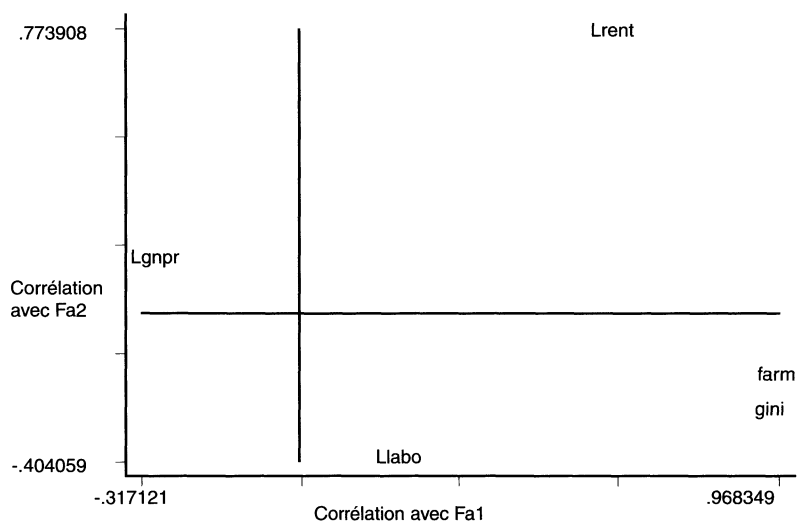


FIGURE 11

*Caractéristiques d'inégalité économique dans le plan (Fa1, Fa2)
(variables de développement en supplémentaire)*

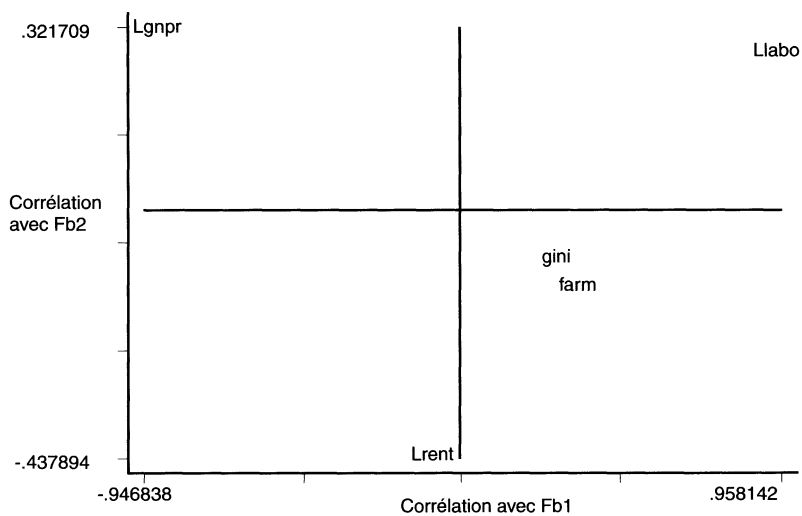


FIGURE 12

*Caractéristiques de développement économique dans le plan (Fb1, Fb2)
(variables d'inégalité en supplémentaire)*

5. Remarques pratiques

- Pour ne pas perdre de vue les rapports que les facteurs d'un thème entretiennent avec les autres thèmes, nous conseillons de projeter, dans chaque plan thématique, les variables des autres thèmes en supplémentaire.

- Les facteurs thématiques n'étant pas *a priori* orthogonaux *entre thèmes*, on pourra trouver commode, après les avoir interprétés, de présenter leur structure de corrélation à l'aide d'une ACP.

6. Liens entre l'ACT et l'approche PLS

6.1. Une estimation interne alternative pour l'approche PLS

La construction de l'ACT nous permet de proposer une procédure d'estimation interne pour la phase 2 de l'approche PLS, qui respecte la logique multivariée de cette phase.

- On suppose tout d'abord que dans le modèle reliant les variables latentes, aucune variable n'est simultanément explicative et dépendante. L'estimation interne est alors conduite comme suit :

- Si F_l est dépendante d'autres variables latentes, on l'estimera par sa régression (normée) sur ces dernières.
- Si F_l fait partie des variables explicatives $\{F_m\}$ d'une autre variable latente F_k , alors, on régressera Φ_k sur $\{X_l, \{\Phi_m\}\}$ et on prendra comme nouvelle estimation interne de F_l la composante sur $\langle X_l \rangle$ de cette régression, que l'on normera. Autrement dit, on projette Φ_k sur $\langle X_l \rangle$ parallèlement à $\langle \{\Phi_m\} \rangle$.

- Ce schéma multivarié d'estimation se comporte différemment des schémas bivariés ou hybrides décrits au § 3. Considérons la situation représentée sur la figure 13. Les directions a , b et c sont supposées être des structures fortes dans leur groupe. Le groupe explicatif Y_1 contient a et b , tandis que son homologue Y_2 est réduit à c . La variable à expliquer z appartient ici au plan (a, c) . La dimension de Y_1 la plus utile au modèle multivarié est donc a . Or, un schéma d'estimation interne de F_1 fondé sur une liaison bivariée entre F_1 et z privilégiera toujours la dimension b , bien plus corrélée à z que a .

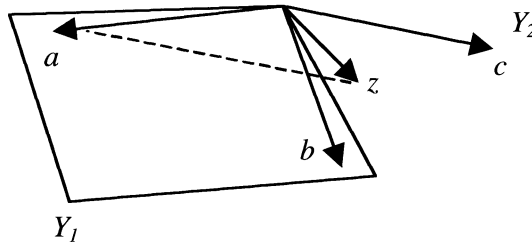


FIGURE 13

Par contre, la procédure que nous proposons, projetant z sur $\langle Y_1 \rangle$ parallèlement à c , trouve immédiatement la direction a .

• Dans le cas d'un modèle à plusieurs équations où une même variable latente intervient à la fois comme variable explicative et comme variable dépendante, un compromis doit être trouvé entre les deux modes d'estimation interne correspondants. Si la variable intervient dans s équations du modèle, elle peut faire l'objet d'autant d'estimations internes, tantôt comme variable explicative, tantôt comme variable dépendante. L'hypothèse d'unicité de la variable latente nous conduit naturellement à l'estimer par le premier facteur d'ACP de ces estimations internes.

Une forte divergence entre ces dernières indiquerait l'inadéquation du modèle n'envisageant qu'une variable latente dans le groupe.

6.2. Liens avec l'ACT

La procédure d'estimation interne proposée précédemment rend l'approche PLS très voisine de l'étape 1 de Q-ACT (extraction des facteurs de rang 1) lorsque $N = I$. Notons respectivement F_l et G les variables latentes explicatives et expliquées, et $F_l(n)$ et $G(n)$ leurs estimations à l'itération n . On constate que :

- L'enchaînement estimation interne / estimation externe de chaque variable latente explicative F_l est faite de la même façon dans les deux méthodes, *i.e.* par calcul de la résultante sur le groupe X_l de la projection de la variable à expliquer estimée G sur $\langle X_l \rangle$ parallèlement aux autres variables explicatives latentes.
- L'enchaînement estimation interne / estimation externe de la variable à expliquer de l'approche PLS consiste à régresser G sur $\{F_l\}$, puis à calculer la résultante sur Z de la partie expliquée. Or, cette partie expliquée n'est autre que la résultante de G sur $\{F_l\}$ lorsqu'on utilise pour métrique la matrice de variance inverse des F_l , *i.e.* $V(n-1) = (F(n-1)'F(n-1))^{-1}$ à l'itération courante. Par conséquent, on a : $G(n) \propto R_Z^I R_{F(n-1)}^{V(n-1)} G(n-1)$. Dans l'ACT, $G(n)$ est la solution de $Q(F(n-1), V(n-1); Z, I)$. C'est donc le vecteur propre de $R_Z^I R_{F(n-1)}^{V(n-1)}$ associé à la plus grande valeur propre. Le calcul de $G(n)$ dans l'approche PLS peut ainsi être considéré comme la première itération du calcul de $G(n)$ dans l'ACT.

7. Conclusion

L'ACT est une technique d'extraction de variables latentes, mais elle peut aussi être utilisée pour la sélection de prédicteurs efficaces, comme nous l'avons vu dans l'exemple des magazines.

Concernant l'extraction des variables latentes, l'ACT s'affranchit de toute hypothèse tant sur la dimensionnalité des groupes que sur le sens des liaisons, ce qui rend son usage souple et en permet une utilisation vraiment exploratoire. En particulier, l'ACT fournit, pour chaque thème, des représentations en plans factoriels

des variables et des individus, ce qui permet une analyse thématique fine de ces derniers en relation avec le groupe à prédire. L'extraction de plusieurs facteurs par groupe, typique de l'analyse des données, impose de restreindre la complexité du modèle latent à une seule équation, contrairement à l'approche PLS, qui gère plusieurs équations sous contrainte d'unicité de la variable latente dans chaque groupe.

Concernant la sélection de prédicteurs, l'ACT se trouve en concurrence avec les méthodes de sélection automatique. Celles-ci peuvent être très efficaces sur le plan de l'ajustement, mais ne tiennent aucun compte des structures fortes révélées par les corrélations des variables, ni de la structuration thématique de celles-ci, et ne laissent pas l'initiative à l'analyste. Une variable pertinente peut être supplantée par une variable qui lui est très corrélée mais qui a un sens beaucoup moins fort et qui ressortit éventuellement à un thème différent. L'interprétation du modèle et sa robustesse risquent d'être sacrifiées à l'ajustement. Nous pensons qu'il est très préférable de donner à l'analyste une image globale des structures thématiques utiles et de l'y laisser choisir les variables qu'il trouve les plus pertinentes. Les résultats de la sélection manuelle pourront toujours être comparés à ceux de la sélection automatique, à titre indicatif.

Remerciements

Je tiens à remercier vivement Odile Wolber, Pierre Cazes et Michel Tenenhaus pour leurs observations, suggestions et appui documentaire. Merci également à Ludovic Lebart pour ses encouragements.

BIBLIOGRAPHIE

- [1] Bry X. (1994) *Analyses Factorielles Simples*, Economica Poche.
- [2] Bry X. (1995) *Analyses Factorielles Multiples*, Economica Poche.
- [3] Bry X. (2001) *Une autre approche de l'Analyse Factorielle : l'Analyse en Résultantes Covariantes*, RSA vol. 49, n° 3, pp. 5-38.
- [4] Escofier B., Pagès J. (1990) *Analyses factorielles simples et multiples*, Dunod.
- [5] Lebart L., Morineau A., Piron M. (1995) *Statistique exploratoire multidimensionnelle*, Dunod.
- [6] Lohmöller J.-B. (1989) *Latent Variables Path Modelling with Partial Least Squares*, Physica-Verlag, Heidelberg.
- [7] Tenenhaus M. (1998) *La régression PLS, théorie et pratique*, Technip.
- [8] Tenenhaus M. (1999) *L'approche PLS*, RSA vol. 47, n° 2, pp. 5-40.
- [9] Wold H. (1985) *Partial Least Squares*, Encyclopedia of Statistical Sciences, John Wiley & Sons, pp. 581-591.

Annexe 1 : propriétés des programmes P et Q

Y , Z et z désignant respectivement deux groupes de variables et une variable, Φ étant un facteur de $\langle Y \rangle$, et Y^* , Z^* et z^* les résidus respectifs de Y , Z et z régressés sur Φ , on a les propriétés suivantes :

PROPRIÉTÉ 1.

Il n'y a pas, en général, équivalence entre les programmes :

- (a) $\mathbf{Q1}(Y, M; z)$ sous la contrainte supplémentaire : $YMu \perp \Phi$
- (b) $\mathbf{Q1}(Y^*, M; z)$

Ceci vaut en particulier si z est remplacée par z^* . Toutefois, on a l'exception suivante :

Lorsque $M = (Y'Y)^{-1}$, $\mathbf{Q1}(Y, M; z^*)$ sous la contrainte $YMu \perp \Phi \iff \mathbf{Q1}(Y^*, M; z^*)$

PROPRIÉTÉ 1'.

Il n'y a pas, en général, équivalence entre les programmes :

- (a) $\mathbf{Q}(Y, M; Z, N)$ sous la contrainte supplémentaire : $YMu \perp \Phi$
- (b) $\mathbf{Q}(Y^*, M; Z, N)$

Ceci vaut en particulier si Z est remplacée par Z^* . Toutefois, on a l'exception suivante :

Lorsque $M = (Y'Y)^{-1}$, $\mathbf{Q1}(Y, M; Z^*, N)$ sous la contrainte $YMu \perp \Phi \iff \mathbf{Q1}(Y^*, M; Z^*, N)$

PROPRIÉTÉ 2.

Il n'y a pas, en général, équivalence entre les programmes :

- (a) $\mathbf{P1}(Y, M; z)$ sous la contrainte supplémentaire : $YMu \perp \Phi$
- (b) $\mathbf{P1}(Y^*, M; z)$

L'équivalence est toutefois vraie si z est remplacé par z^* .

PROPRIÉTÉ 2'.

Il n'y a pas, en général, équivalence entre les programmes :

- (a) $\mathbf{P}(Y, M; Z, N)$ sous la contrainte supplémentaire : $YMu \perp \Phi$
- (b) $\mathbf{P}(Y^*, M; Z, N)$

L'équivalence est toutefois vraie si Z est remplacé par Z^* .

Preuve de 1.

- $\mathbf{Q1}(Y^*, M; z)$ a pour solution :

$$F \propto Y^* M Y'^* z = \Pi_{\langle \Phi \rangle^\perp} Y M Y' \Pi_{\langle \Phi \rangle^\perp} z \quad (1)$$

• Considérons $\mathbf{Q1}(Y, M; z)$ sous la contrainte supplémentaire : $YMu \perp \Phi$,
soit : $\text{Max}_{\substack{u'Mu=1 \\ \Phi'YMu=0}} \langle YMu | z \rangle$

Le lagrangien de ce programme s'écrit : $L = z'YMu - \lambda u'Mu - \mu \Phi'YMu$

Les conditions du 1^{er} ordre donnent :

$$\frac{\partial L}{\partial u} = 0 \iff MY'z - 2\lambda Mu - \mu MY'\Phi = 0 \iff 2\lambda u = Y'z - \mu Y'\Phi \quad (2)$$

$$u'M(2) \iff 2\lambda = u'MY'z, \text{ car } u'MY'\Phi = 0$$

Notons $\alpha = \Phi'YMY'\Phi$.

$$\text{Alors, } \Phi'YM(2) \iff 0 = \Phi'YM'z - \mu\alpha \iff \mu = \frac{\Phi'YMY'z}{\Phi'YMY'\Phi}.$$

$$\text{Finalement, } (2) \iff (u'MY'z)u = Y'z - \frac{Y'\Phi(\Phi'YMY'z)}{\Phi'YMY'\Phi}$$

$$\iff (u'MY'z)u = \left(I - \frac{Y'\Phi\Phi'YM}{\Phi'YMY'\Phi} \right) Y'z \quad (3)$$

$$\text{Et } YM(3) \iff F = YMu \propto \left(I - \frac{YMY'\Phi\Phi'}{\Phi'YMY'\Phi} \right) YMY'z \quad (4)$$

Lorsqu'on remplace z par $z^* = \Pi_{\langle \Phi \rangle^\perp} z$, (1) reste inchangée :

$$F \propto \Pi_{\langle \Phi \rangle^\perp} YMY'z^*, \text{ tandis que (4) devient : } F \propto \left(I - \frac{YMY'\Phi\Phi'}{\Phi'YMY'\Phi} \right) YMY'z^*.$$

Or, $\Pi_{\langle \Phi \rangle^\perp} = I - \frac{\Phi\Phi'}{\Phi'\Phi} \neq I - \frac{YMY'\Phi\Phi'}{\Phi'YMY'\Phi}$ en général. On remarquera toutefois qu'on a l'égalité pour $M = (Y'Y)^{-1}$, car $\Phi \in \langle Y \rangle \implies Y(Y'Y)^{-1}Y'\Phi = \Phi$.

Preuve de 1'.

- $\mathbf{Q}(Y^*, M; Z, N)$ a des solutions caractérisées par :

$$(1) \quad F \propto R_{Y^*}^M G \iff F \propto \Pi_{\langle \Phi \rangle^\perp} YMY'\Pi_{\langle \Phi \rangle^\perp} G$$

$$(1') \quad G \propto R_Z^N F \iff G \propto ZN Z' F$$

• Considérons $\mathbf{Q}(Y, M; Z, N)$ sous la contrainte supplémentaire : $YMu \perp \Phi$,
soit : $\text{Max}_{\substack{u'Mu=1, v'Nv=1 \\ \Phi'YMu=0}} \langle YMu | ZNv \rangle$

Le lagrangien de ce programme s'écrit :

$$L = v'NZ'YMu - \lambda(u'Mu - 1) - \mu(v'Nv - 1) - \kappa\Phi'YMu$$

Les conditions du 1^{er} ordre donnent :

$$(2) \quad \frac{\partial L}{\partial u} = 0 \iff MY'ZNv - 2\lambda Mu - \kappa MY'\Phi = 0;$$

$$(2') \quad \frac{\partial L}{\partial v} = 0 \iff NZ'YMu - 2\mu Nv = 0$$

$u'(2)$ et $v'(2') \implies \lambda = \mu$. On pose $2\lambda = 2\mu = \alpha = u'MY'ZNv$.

D'autre part,

$$\Phi'Y(2) \iff \Phi'YMY'ZNv = \kappa \Phi'YMY'\Phi \iff \kappa = \frac{\Phi'YMY'ZNv}{\Phi'YMY'\Phi}.$$

Dès lors :

$$(2) \iff MY'ZNv - (u'MY'ZNv)Mu - \frac{MY'\Phi(\Phi'YMY'ZNv)}{\Phi'YMY'\Phi} = 0 \quad (3)$$

$$\begin{aligned} \text{Et : } Y(3) &\iff YMY'G - (u'MY'ZNv)F - \frac{YMY'\Phi\Phi'YMY'}{\Phi'YMY'\Phi} G = 0 \\ &\iff \left(I - \frac{YMY'\Phi\Phi'}{\Phi'YMY'\Phi}\right) YMY'G - (u'MY'ZNv)F = 0 \\ &\implies F \propto \left(I - \frac{YMY'\Phi\Phi'}{\Phi'YMY'\Phi}\right) YMY'G \end{aligned} \quad (4)$$

$$\text{Par ailleurs : } Z(2') \iff ZNZ'F - 2\mu G = 0 \quad (3') \implies G \propto ZNZ'F \quad (4')$$

On constate que $(4') \iff (1')$, alors qu'en général, $(4) \not\iff (1)$.

Lorsqu'on remplace en particulier Z par $Z^* = \Pi_{\langle\Phi\rangle^\perp} Z$, on a :

$$(1) \iff F \propto \Pi_{\langle\Phi\rangle^\perp} YMY'\Pi_{\langle\Phi\rangle^\perp} Z^*Nv = \Pi_{\langle\Phi\rangle^\perp} YMY'\Pi_{\langle\Phi\rangle^\perp} ZNv.$$

Donc, (1) ne change pas.

$$(4) \iff F \propto \left(I - \frac{YMY'\Phi\Phi'}{\Phi'YMY'\Phi}\right) YMY'\Pi_{\langle\Phi\rangle^\perp} ZNv.$$

Or, en général, $I - \frac{YMY'\Phi\Phi'}{\Phi'YMY'\Phi} \neq \Pi_{\langle\Phi\rangle^\perp}$ (si $M = (Y'Y)^{-1}$, toutefois, on a l'égalité).

Preuve de 2.

• $\mathbf{P1}(Y^*, M; z)$ a une solution colinéaire à celle de $\mathbf{Q1}(Y^*, M; z)$, soit :

$$F \propto Y^*MY^{*'}z = \Pi_{\langle\Phi\rangle^\perp} YMY'\Pi_{\langle\Phi\rangle^\perp} z \quad (1)$$

• Considérons $\mathbf{P1}(Y, M; z)$ sous la contrainte supplémentaire : $YMu \perp \Phi$, soit : $\text{Max}_{\substack{F'F=1 \\ \Phi'F=0}} \langle R_Y^M F | z \rangle$

Le lagrangien de ce programme s'écrit : $L = z'R_Y^M F - \lambda(F'F - 1) - \mu\Phi'F$

Les conditions du 1^{er} ordre donnent :

$$\frac{\partial L}{\partial F} = 0 \iff R_Y^M z - 2\lambda F - \mu\Phi = 0 \implies F \propto R_Y^M z - \mu\Phi \quad (2)$$

$\Phi'(2)$ et $\Phi'F = 0 \implies \mu = \frac{\Phi'R_Y^M Z}{\Phi'\Phi}$. Il s'ensuit :

$$(2) \iff F \propto R_Y^M z - \Phi \frac{\Phi'R_Y^M Z}{\Phi'\Phi} = \left(I - \frac{\Phi\Phi'}{\Phi'\Phi} \right) R_Y^M Z$$

Soit : $F \propto \Pi_{\langle\Phi\rangle^\perp} YMY'z$. Par conséquent, si z est remplacé par $z^* = \Pi_{\langle\Phi\rangle^\perp} z$, on retrouve la solution de $\mathbf{P1}(Y^*, M; z)$. Donc, $\mathbf{P1}(Y, M; z^*)$ sous la contrainte : $YMu \perp \Phi$ équivaut à $\mathbf{P1}(Y^*, M; z)$ et à $\mathbf{P1}(Y^*, M; z^*)$.

Preuve de 2'.

• $\mathbf{P}(Y^*, M; Z, N)$ a des solutions caractérisées par :

$$(1) \quad R_{Y^*}^M R_Z^N G = \alpha F \iff \Pi_{\langle\Phi\rangle^\perp} YMY' \Pi_{\langle\Phi\rangle^\perp} ZN'Z'G = \alpha F$$

$$(1') \quad R_Z^N R_{Y^*}^M F = \alpha G \iff ZN'Z' \Pi_{\langle\Phi\rangle^\perp} YMY' \Pi_{\langle\Phi\rangle^\perp} F = \alpha G$$

• Considérons $\mathbf{P}(Y, M; Z, N)$ sous la contrainte supplémentaire : $F \perp \Phi$, soit :

$$\text{Max}_{F'F=1; G'G=1} \langle R_Y^M F | R_Z^N G \rangle$$

$$F'\Phi=0$$

On prendra ici Φ normé, comme F et G .

Le lagrangien du programme est :

$$L = F'R_Y^M R_Z^N G - \lambda(F'F - 1) - \mu(G'G - 1) - \kappa F'\Phi$$

Les conditions du premier ordre sont :

$$(2) \quad \frac{\partial L}{\partial F} = 0 \iff R_Y^M R_Z^N G - 2\lambda F - \kappa\Phi = 0;$$

$$(2') \quad \frac{\partial L}{\partial G} = 0 \iff R_Z^N R_Y^M F - 2\mu G = 0$$

$F'(2)$ et $G'(2') \iff \lambda = \mu$. On notera $\alpha = 2\lambda = 2\mu$.

En outre, $\Phi'(2) \iff \kappa = \Phi'R_Y^M R_Z^N G$. Dès lors :

$$(2) \iff R_Y^M R_Z^N G - \Phi\Phi'R_Y^M R_Z^N G = \alpha F \iff (I - \Phi\Phi')R_Y^M R_Z^N G = \alpha F$$

$$\iff \Pi_{\langle\Phi\rangle^\perp} R_Y^M R_Z^N G = \alpha F \quad (3)$$

$$\text{On obtient aussi : } (2') \iff R_Z^N R_Y^M F = \alpha G \quad (3')$$

Généralement, (3) $\nleftrightarrow$ (1) et (3') $\nleftrightarrow$ (1'). Toutefois, si l'on remplace Z par Z^* , il vient :

$$(3) \iff \Pi_{\langle \Phi \rangle^\perp} Y M Y' Z^* N Z^{*'} G = \alpha F \iff \Pi_{\langle \Phi \rangle^\perp} Y M Y' \Pi_{\langle \Phi \rangle^\perp} Z^* N Z^{*'} G = \alpha F \iff (1) \text{ où } Z \text{ est remplacé par } Z^*.$$

D'autre part :

$$(3') \iff Z^* N Z^{*'} Y M Y' F = \alpha G, \text{ et comme } \Pi_{\langle \Phi \rangle^\perp} F = F, (3') \iff Z^* N Z^{*'} \Pi_{\langle \Phi \rangle^\perp} Y M Y' \Pi_{\langle \Phi \rangle^\perp} F = \alpha G \iff (1') \text{ où } Z \text{ est remplacé par } Z^*.$$

Donc, $\mathbf{P}(Y, M; Z^*, N)$ sous la contrainte : $YMu \perp \Phi$ équivaut à $\mathbf{P}(Y^*, M; Z^*, N)$.

Annexe 2 : données

Nom	Identifiant	Prix	Poids (g)	Nb Pages	Nb N°s/An	Surface (cm ²)	Papier 1 = journal 2 = glacé	Couleur 1 = N&B 2 = Bichromie 3 = Quadrichromie
PourLaScience	PLS	39	250	116	12	588	2	3
PourLaScienceHS	PLSHS	45	380	132	4	588	2	3
SciencesHumaines	SH	33	220	80	12	624	2	3
ManièreDeVoir	MDV	45	230	100	4	624	2	3
MagazineLittéraire	IML	32	185	108	12	567	2	3
Sciences&AvenirHS	SAHS	25	185	100	4	598	2	3
JalIntelligent	JALI	18	120	100	52	526	2	3
Parents	P	14	340	136	12	675	2	3
ArtActuel	AA	28	280	100	6	616	2	3
LaRecherche	IR	39	285	124	12	670	2	3
BeauxArts	BA	39	380	132	12	627	2	3
CanardEnchaîné	ICE	8	40	8	52	2016	1	1
CharlieHebdo	CH	10	60	16	52	1282	1	2
MondeDiplomatique	IMD	25	140	32	12	1504	1	3
Marianne	M	15	142	84	52	553	2	3
AlternativesEconomiques	AE	18	82	34	52	567	2	1
Science&Vie	SV	23	220	170	12	408	2	3
60MConsommateurs	SMC	26	142	76	12	591	2	3
TribuneSanté	TS	30	160	48	12	696	2	3
CourrierInternational	CI	18	185	76	52	913	1	3
CahiersDuCinéma	ICdC	35	340	132	12	664	2	3

Nom	Identifiant	Domaine 1 = Sciences 2 = Arts & Littérature 3 = Info & Opinion 4 = Vie courante	Type de N° 0 = ordinaire 1 = spécial (Hors Série)	% Auto-publicité	% Publicité Externe	% Infos brèves Interviews Critique Littéraire	% Dossiers & Analyses	% Jeux	% Annonces	% Courrier	% People, Faits divers, Sport	% Consommation (info non critique)	% Infos Pratiques
PourLaScience	PLS	1	0	6,3	9,8	16,2	60,0	6,0	0,0	0,9	0,0	0,0	0,0
PourLaScienceHS	PLSHS	1	1	6,8	3,8	0,0	87,1	0,0	0,0	0,0	0,0	0,0	0,0
SciencesHumaines	SH	1	0	9,4	5,5	22,8	66,3	0,0	0,0	0,0	0,0	0,0	0,0
ManièreDeVoir	MDV	3	0	1,0	3,0	0,0	92,0	0,0	0,0	0,0	0,0	0,0	0,0
MagazineLittéraire	IML	2	0	1,9	7,8	26,4	62,0	0,0	0,0	0,0	0,0	0,0	0,0
Sciences&AvenirHS	SAHS	1	1	0,0	2,0	8,0	88,0	0,0	0,0	0,0	0,0	0,0	0,0
JalIntelligent	JALI	3	0	1,0	7,5	23,5	45,0	0,0	6,0	1,0	8,2	3,0	0,7
Parents	P	4	0	1,8	31,6	3,7	25,0	1,8	1,8	2,2	2,2	12,9	10,3
ArtActuel	AA	2	0	3,5	5,5	51,5	28,0	0,0	0,0	0,0	0,0	0,0	0,0
LaRecherche	IR	1	0	2,8	9,4	25,8	52,0	2,8	2,4	2,7	0,0	1,3	0,0
BeauxArts	BA	2	0	4,0	15,2	37,6	30,3	0,0	0,0	0,0	0,0	3,8	0,0
CanardEnchaîné	ICE	3	0	0,6	0,0	66,5	31,3	1,6	0,0	0,0	0,0	0,0	0,0
CharlieHebdo	CH	3	0	5,8	0,7	72,7	22,9	0,0	0,0	1,0	0,0	0,0	0,0
MondeDiplomatique	IMD	3	0	5,4	12,7	10,4	70,3	0,0	0,0	1,2	0,0	0,0	0,0
Marianne	M	3	0	0,3	8,3	33,3	34,5	1,2	0,0	4,8	4,8	8,3	0,0
AlternativesEconomiques	AE	3	0	5,3	0,0	0,0	88,2	0,0	0,0	0,0	0,0	0,0	0,0
Science&Vie	SV	1	0	4,0	9,0	25,0	55,9	0,0	0,0	1,5	0,0	4,1	0,0
60MConsommateurs	SMC	4	0	13,5	1,3	13,2	50,0	0,0	0,0	6,6	0,0	0,0	10,5
TribuneSanté	TS	4	0	3,1	6,8	34,4	39,6	0,0	0,0	2,1	0,0	2,1	8,3
CourrierInternational	CI	3	0	1,5	37,8	29,0	26,3	0,0	0,0	0,2	0,0	0,0	0,0
CahiersDuCinéma	ICdC	2	0	4,3	17,0	29,2	48,1	0,0	0,0	0,6	0,0	0,0	0,0

Notes sur les intitulés des variables de contenu :

Auto-publicité : publicité pour abonnements, numéros spéciaux, produits dérivés (CDROM, sites internet...)

Publicité Extérieure : Publicité commerciale traditionnelle ou publi-reportages, hors auto-publicité. ...

Infos Interviews : Informations brèves (peu ou pas d'analyse), interviews, critique littéraire.

Dossiers & Analyses : Articles de fond, dossiers, enquêtes.

Jeux : Jeux, tests, concours.

Annonces : Petites annonces

Courrier : Courrier des lecteurs, forums, contributions.

People, Sport : Faits divers, informations sportives

Consommation : Information commerciale *non-critique* (les bancs d'essais sont exclus).

Infos Pratiques : Agendas, guides non-commerciaux (droit, santé, administration, etc.)