HAL open science

Un critère de covariance multiple permettant l'extension de la régression PLS à plusieurs groupes prédicteurs

Xavier Bry, Thomas Verron, Pierre Cazes

To cite this version:

Xavier Bry, Thomas Verron, Pierre Cazes. Un critère de covariance multiple permettant l'extension de la régression PLS à plusieurs groupes prédicteurs. 41èmes Journées de Statistique, SFdS, Bordeaux, 2009, Bordeaux, France, France. inria-00386701

HAL Id: inria-00386701

https://inria.hal.science/inria-00386701v1

Submitted on 22 May 2009

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Un critère global de covariance multiple permettant l'extension de la régression PLS à plusieurs groupes prédicteurs

X. Bry [1] , T. Verron [2], P. Cazes [3]

[1] *I3M, Université Montpellier 2, Place Eugène Bataillon, 34090 Montpellier*
[2] *ALTADIS - SCR, 4 rue André Dessaux, 45404 Fleury les Aubrais*
[3] *CEREMADE, Univ. Paris Dauphine, Pl. Lattre de Tassigny 75016 Paris*

Résumé : Nous cherchons à explorer un modèle structurel : plusieurs groupes de variables décrivant les mêmes unités sont supposés structurés autour de dimensions latentes liées entre elles par un modèle linéaire. Ce type de modèle est classiquement traité par des méthodes supposant unique la dimension structurant chaque groupe. Mais souvent, les modèles conceptuels articulent entre eux des concepts multidimensionnels. Nous proposons une méthode : SEER (Structural Equation Exploratory Regression) qui permet d'explorer la structure des groupes à la recherche de toutes les dimensions utiles au modèle. Fondée sur la maximisation d'un critère de covariance multiple, SEER étend à la fois la Régression PLS et PLS Path Modeling. Une application comparée à des données physico-chimiques a montré un net avantage de SEER tant sur le plan explicatif que prédictif.
Mots-clefs : *PLS Path Modeling, PLS, SEER, Structural Equation Models.*

Abstract : Our aim is to explore a structural model: several variable groups describing the same units are assumed to be structured around latent dimensions that are linked together through a linear model. This type of model is commonly dealt with by methods assuming that the latent dimension in each group is unique. However, conceptual models generally link concepts which are multidimensional. We propose a method: SEER (Structural Equation Exploratory Regression), which allows to search the structures of groups for all dimensions useful to the model. Based on the maximization of a multiple covariance criterion, SEER extends both PLS Regression and PLS Path Modeling. A compared application on physicochemical data has shown a clear advantage of SEER, as far as both explanation and prediction are concerned.
Keywords: *PLS Path Modeling, PLS, SEER, Structural Equation Models.*

1 Introduction

Le contexte est celui des modèles à équations structurelles (SEM) : plusieurs groupes de variables décrivant les mêmes unités sont supposés structurés autour de dimensions liées entre elles par un modèle linéaire. Les SEM sont classiquement traités selon le paradigme des variables latentes : chaque groupe de variables est supposé produit par une variable latente qu'il s'agit d'estimer. Deux approches sont couramment utilisées. La première, très empirique, PLS Path Modeling [1,4], n'est fondée sur aucun critère global. La seconde consiste à maximiser un critère global. Selon le critère choisi, on obtient diverses méthodes telles que [3,5,2]. La plus grande rigueur de cette approche se paye de difficulté à traiter les échantillons au nombre d'observations petit devant celui des variables. Ces méthodes supposent chaque groupe structuré autour d'une dimension unique. Mais, les modélisateurs disposent souvent d'un modèle conceptuel articulant entre eux des concepts *a priori* multidimensionnels. Il est alors important d'explorer la structure des groupes afin d'en extraire la part utile à la modélisation. C'est ce que fait la Régression PLS [6] dans le cas particulier où un groupe dépendant est modélisé à partir d'un seul groupe prédicteur. RPLS est fondée sur la maximisation de la covariance des composantes. Des extensions de RPLS [7,8], ont été proposées pour traiter le cas de plusieurs groupes de prédicteurs ; mais elles n'optimisent plus de critère global. Nous proposons ici une méthode, SEER (Structural Equation Exploratory Regression), fondée sur la maximisation d'une covariance multiple, qui étend ainsi RPLS. SEER extrait un nombre quelconque de composantes clairement hiérarchisées par groupe.

2 Données et problème :

2.1 Données

Soient n unités décrites par plusieurs groupes de variables dont l'un (Y) est supposé linéairement dépendant des autres ($X_1, \ldots, X_R$). Dans l'exemple traité, $n = 28$ cigarettes sont décrites par 58 variables partitionnées en 4 groupes : $Y(minSC)$ contient les concentrations de 14 composés mineurs de la fumée, $X_1(TChem)$ 29 caractéristiques chimiques du tabac, X_2 ($CPhys$) 10 caractéristiques physiques de la cigarette, et X_3 ($MajSC$) 5 concentrations de composés majeurs de la fumée. Le groupe Y (resp. X_r) possède K (resp.

J_r) variables et on lui associe une matrice métrique N (resp. M_r) de taille K (resp. J_r).

2.2 Objectif & modèle

L'objectif est d'expliquer et prédire Y à partir de $X_1, \ldots, X_R$. Les groupes sont articulés selon un modèle conceptuel *a priori*, ou *modèle thématique* (cf. fig. 2). Ce modèle, destiné à évoluer selon les résultats des estimations, peut rester rudimentaire dans le premier temps de l'exploration.

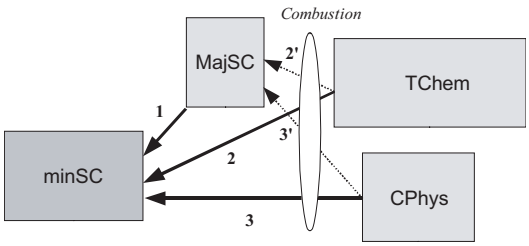

Fig1 : Premier modèle conceptuel

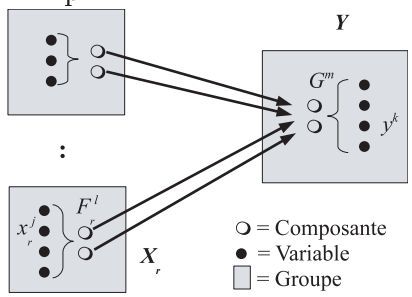

Fig2 : Modèle à composantes

Le premier modèle conceptuel de nos données est motivé par deux considérations :
- Les composants de la fumée dépendent de la composition du tabac et des caractéristiques physiques de la cigarette (flèches 2, 3, 2' et 3') ;
- Une part a priori résiduelle de *minSC* pourrait provenir de la seule machine à fumer. L'existence d'un tel artefact pourrait alors être détectée en utilisant *MajSC* en 3me groupe prédicteur comme proxy pour certaines caractéristiques non observées de la machine.

2.3 Composantes

Les structures intéressantes de X_r (resp. Y) sont supposées pouvoir être captées *via* un nombre J'_r (resp. K) de composantes $F^1_r, \ldots, F^{J'_r}_r$ (resp. $G^1, \ldots, G'^K$). On impose *a priori* : $\forall (j,r)$: $F^j_r = X_r M_r u^j_r$, $\forall k$: $G^k = YNv^k$, $\forall (j,r)$: $u^{j\prime}_r M_r u^j_r = 1$ et $\forall k$: $v^{k\prime} N v^k = 1$, de sorte que la variance de F^j_r (resp. G^k) représente l'inertie sur l'axe $< u^j_r >$ (resp. $< v^k >$) des observations selon (X_r, M_r) (resp. Y muni de N).
N.B. Quand $Y = \{y\}$, on obtient un modèle univarié, et $G = y$.

3 Structural Equation Exploratory Regression

3.1 Critères de covariance multiple

Covariance Multiple

Notre covariance multiple est une extension possible de la covariance binaire, que RPLS maximise au rang 1.
Définition : y étant modélisée linéairement en fonction de $x^1, \ldots,\ x^S$, nous appelons *Covariance Multiple de y sur $x^1, \ldots,\ x^S$* :

$$CM(y|x^1, \ldots, x^S) = \left[\left(V(y) \prod_{s=1}^{S} V(x^s) \right) R^2(y|x^1, \ldots, x^S) \right]^{1/2}$$

où $R^2(y \mid x^1,\ \ldots,\ x^S)$ est le R^2 de la régression de y sur $\{x^1,\ \ldots,\ x^S\}$.

Critère pour le modèle univarié

- Considérons le modèle correspondant à la fig.2 quand $Y = \{y\}$. On cherchera à maximiser :

$$C_1(y; F_1, \ldots, F_R) = CM^2(y|F_1, \ldots, F_R) = \left(V(y) \prod_{r=1}^{R} V(F_r) \right) R^2(y|F_1, \ldots, F_R) \qquad (1)$$

C_1 amalgame la force structurelle des composantes F_r, mesurée par leur variance, au R^2 de régression de y sur celles-ci.
- Pour tout $1 \leq r \leq R$, soit $F_{-r} = \{F_s, s \neq r\}$ et Π_E le projecteur orthogonal sur un sous-espace E de $\mathbb{R}^n$. Nous montrons que C_1 peut être formulé ainsi :

$$C_1 = F_r{}' F_r \frac{F_r{}' A_r(y) F_r}{F_r{}' B_r F_r}, \text{ où :} B_r = \Pi_{F_{-r}^{\perp}};\ A_r(y) = (y' \Pi_{F_{-r}} y) B_r + B_r{}' y y' B_r \qquad (2)$$

Critère pour le modèle multivarié

- Si $N = diag(n_k)_{k=1 \text{ à } K}$, on peut directement tirer de C_1 le critère suivant :

$$C_2(Y, N; F_1, \ldots, F_R) = \sum_{k=1}^{K} n_k C_1(y^k; F_1, \ldots, F_R) \qquad (3)$$

- De (2) et (3) s'ensuit que pour tout $1 \leq r \leq R$, C_2 peut être écrit :

$$C_2 = F_r{}' F_r \frac{F_r{}' A_r F_r}{F_r{}' B_r F_r} \text{ où :} B_r = \Pi_{F_{-r}^{\perp}};\ A_r = \Pi_{F_{-r}}^{\perp} tr(YNY' \Pi_{F_{-r}}) + \Pi_{F_{-r}}^{\perp} (YNY') \Pi_{F_{-r}}^{\perp}$$

3.2 Composantes F de rang 1

Prenant $C = C_1$ pour le modèle univarié et $C = C_2$ pour le multivarié, il s'agit de résoudre le programme $Q : \underset{\forall r:\ u_r{}'M_r u_r=1}{max} C$.
Nous proposons de maximiser C sur chaque F_r à tour de rôle, F_{-r} étant considéré comme fixe. Nous montrons que le programme de la maximisation contrainte $Q_r : \underset{u_r{}'M_r u_r=1}{max} C$ équivaut à la maximisation libre $S_r : \underset{u}{max}$ $(\ln C - u_r{}'M_r u_r)$, ce qui permet d'utiliser des méthodes générales de maximisation. Nous donnons aussi un algorithme alternatif ayant la solution de Q_r pour point fixe.

3.3 Calcul de la composante de rang k de X_r : F_r^k

En supposant connu le nombre J'_r de composantes désiré dans chaque X_r, on remplace X_r par ses résidus de régression sur $F_r^1, \ldots, F_r^{k-1}$ pour calculer F_r^k. Cependant, afin que la hiérarchie des composantes ne pose pas problème, elle doit rester locale au groupe : $F_r^1, \ldots, F_r^{k-1}$ sont calculées successivement mais en considérant toutes les composantes des autres groupes comme fixes. Ainsi, chaque composante d'un groupe vient compléter les précédentes, à structures données dans les autres groupes. Comme en pratique les nombres J'_r sont inconnus *ab initio*, nous proposons une stratégie de sélection arrière.

3.4 Composantes G du modèle multivarié

Une fois calculé l'ensemble $F = \{F_r^k\}_{r,k}$, on cherche les composantes G de Y les mieux prédites par F, *via* l'ACPVI de Y sur F.

4 Application aux données cigarettes

Nous avons comparé les résultats de SEER sur nos données à ceux de RPLS et PLSPM. Les mécanismes prédicteurs fournis par les modèles estimés ont été évalués par validation croisée. La division thématique des prédicteurs a permis d'identifier les effets *directs*, ce qui est essentiel au plan explicatif. Elle a notamment montré que la machine à fumer n'introduit pas d'artefact. Le modèle auquel SEER nous a conduit possède non seulement la meilleure capacité prédictive des modèles comparés, mais aussi le plus grand nombre de composantes interprétables, donc le plus grand pouvoir explicatif.

Conclusion :

Tout modèle explicatif est bâti sur un modèle conceptuel des données, qui implique un partitionnement thématique des prédicteurs. La dimension des concepts n'étant pas toujours connue, il est essentiel de disposer d'une méthode d'exploration multidimensionnelle des groupes de variables qui les mesurent. La recherche de dimensions optimales impose de maximiser un critère global amalgamant la force structurelle de ces dimensions à l'ajustement du modèle explicatif. Il est enfin souhaitable, dans le contexte multidimensionnel, de borner le nombre des composantes *via* un mécanisme de sélection lié au critère optimisé. SEER tente de satisfaire ces exigences. La covariance multiple sur laquelle elle est fondée en fait une extension de la Régression PLS au cas de prédicteurs partitionnés. La représentation multidimensionnelle qu'elle offre des groupes permet le raffinement progressif du modèle conceptuel, ce qui constitue un avantage sur les méthodes d'estimation des SEM. Son application aux données a conduit à un modèle plus riche sur le plan explicatif et meilleur sur le plan prédictif que le meilleur des modèles fournis par RPLS et PLSPM.

Références bibliographiques

[1] Chin, W.W., Newsted, P.R., (1999) : *Structural equation modeling analysis with small samples using PLS.* In : Statistical Strategies for Small Sample Research. Sage, 307–341.

[2] Hwang, H., and Takane, Y. (2004). *Generalized structured component analysis.* Psychometrika, 69, 81-99.

[3] Jöreskog, K. G. and Wold, H. (1982) *The ML and PLS techniques for modeling with latent variables : historical and competitive aspect*, Systems under indirect observation, Part 1, 263 – 270

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

[5] Smilde, A.K., Westerhuis, J.A., Boquée, R., (2000). *Multiway multiblock component and covariates regression models.* J. Chem. 14, 301–331.

[6] Tenenhaus M. (1998) : *La régression PLS* - Technip.

[7] Wangen L., Kowalski B. (1988) : *A multiblock partial least squares algorithm for investigating complex chemical systems.* J. Chem.; 3 : 3–20.

[8] Westerhuis, J.A., Kourti, K., Macgregor, J.F., (1998) : *Analysis of multiblock and hierarchical PCA and PLS models.* J. Chem. 12, 301–321.