L'analyse exploratoire multidimensionnelle d'un modèle structurel fondée sur une classe de critères de covariance généralisée

Xavier Bry, Patrick Redont, Thomas Verron

L'analyse exploratoire multidimensionnelle d'un modèle structurel fondée sur une classe de critères de covariance généralisée

X. Bry [1], P. Redont [1], Th. Verron [2]

[1] *I3M, Université Montpellier 2, Place Eugène Bataillon, 34095 Montpellier*
[2] *SEITA-ITG, Centre de recherche SCR, 4 rue André Dessaux, 45404 Fleury les Aubrais*

Résumé

Notre but est d'explorer un modèle structurel: plusieurs groupes de variables décrivant les mêmes unités sont supposées structurés autour de dimensions latentes liées entre elles par un modèle linéaire pouvant comporter plusieurs équations. Ce type de modèle est couramment traité par des méthodes ne considérant qu'une dimension latente par groupe. Cependant, les modèles conceptuels relient couramment des concepts structurellement multidimensionnels, sans que l'on sache a priori combien de dimensions interviennent, ni lesquelles. Nous proposons une classe de critères pouvant mesurer la qualité d'un modèle structurel. Cette classe contient le critère de covariance fondant la régression PLS, ainsi que la covariance multiple fondant la méthode SEER. Elle contient également des critères liés à la rotation quartimax. Tous les critères de cette classe sont appelés à être maximisés sous contrainte de norme unité des vecteurs en argument. Nous donnons un programme d'optimisation libre équivalent, ainsi qu'un algorithme pour le résoudre. Cette optimisation est utilisée à l'intérieur d'un algorithme plus général (nommé THEME, pour: Thematic Equation Model Explorer) permettant la recherche dans chaque groupe de toutes les dimensions utiles au modèle. THEME extrait des composantes thématiques localement hiérarchisées.
Mots-clés : Approche PLS, Modèles à équations structurelles, PLS, SEER, THEME.

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 that may have several equations. 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, and how many and which dimensions should be involved is not known. We first propose a general class of criteria suitable to measure the quality of a Structural Equation Model. This class contains the covariance criteria used in PLS Regression methods and the Multiple Covariance criterion of the SEER method. It also contains quartimax-related criteria. All criteria in the class must be maximized under a unit norm constraint. We give an equivalent unconstrained maximization program, and algorithms to solve it. This maximization is used within a general algorithm named THEME (Thematic Equation Model Explorer),

which allows to search the structures of groups for all dimensions useful to the model. THEME extracts locally nested structural component models.
Keywords : Path Modeling, PLS, SEER,Structural Equation Models, THEME.

1. Introduction

Le contexte est celui des modèles à équations structurelles (MES): R groupes de variables, $X_1, \ldots, X_R$ décrivant les mêmes n unités sont supposés structurés autour d'un nombre restreint de dimensions reliées entre elles par un modèle linéaire. Les MES sont couramment traités en utilisant des variables latentes (VL): les variables de chaque groupe X_r sont supposées refléter une seule et même VL, qu'il s'agit d'estimer. Deux démarches sont couramment adoptées. La première, PLS Path Modeling [Chin et Newsted (1999); Tenenhaus (1998); Lohmöller (1989)], n'étant fondée sur l'optimisation d'aucun critère global, reste purement empirique. La seconde consiste à optimiser un critère global interprétable. Selon le critère choisi, on obtient différentes méthodes [Hwang et Takane (2004); Smilde *et al.* (2000); Jöreskog et Wold (1982)]. Cette approche plus rigoureuse se paye souvent de difficultés dans le traitement des petits échantillons. Jusqu'ici, ces méthodes ont supposé que chaque groupe était structuré autour d'une dimension sous-jacente unique impliquée dans la modélisation linéaire. On peut objecter que si tel est le cas, les variables du groupes étant fortement corrélées entre elles, cette dimension peut être estimée très simplement par leur première composante principale (CP). Le modèle linéaire relie *a posteriori* ces CP. Le problème d'identification des dimensions utiles au modèle se pose véritablement lorsque les groupes de variables illustrent des concepts à structure réellement multidimensionnelle, ce à quoi les modélisateurs sont le plus souvent confrontés au départ, sans savoir combien de dimensions interviennent ni, *a fortiori*, lesquelles. Il est alors primordial de pouvoir explorer la structure de ces groupes, en liaison avec le modèle linéaire, de sorte à extraire des groupes les dimensions utiles à ce dernier. Les méthodes telles que Multiblock PLS [Wangen et Kowalski (1988); Westerhuis *et al.* (1998)] tentent de le faire, mais faute d'un critère reflétant correctement les liaisons *partielles* de chaque groupe explicatif à son groupe dépendant, elles doivent incorporer certaines étapes correctives de déflation empiriques et assez arbitraires. Afin d'étendre la régression PLS au cas où un groupe dépendant dépend de plusieurs groupes prédicteurs, nous avons proposé dans Bry *et al.* (2009) de maximiser un critère global que nous avons appelé *covariance multiple*. Cette maximisation est la base d'un algorithme d'exploration d'un modèle à un seul groupe dépendant: Structural Equation Exploratory Regression (SEER). SEER extrait autant de composantes par groupe qu'on en veut (jusqu'à épuiser la dimension du groupe), ordonnées de façon clairement interprétables selon un principe d'emboîtement local. L'utilisation d'un critère global rend possible la sélection arrière des composantes. Nous étendons ici ce critère de covariance multiple de deux façons: 1) le critère est généralisé de sorte à pouvoir prendre en compte la force structurelle des composantes de façon plus flexible, incluant des mesures orientées vers le dépistage de faisceaux de variables; 2) il est aussi étendu de sorte à pouvoir traiter un modèle compor-

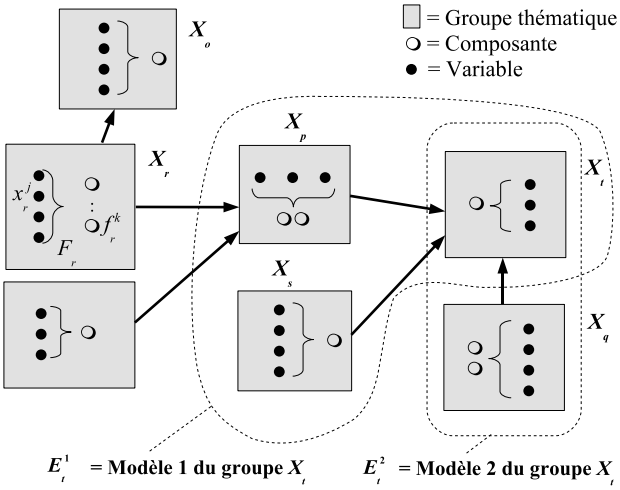
tant un nombre quelconque de groupes dépendants ou prédicteurs. Nous proposons enfin des algorithmes de maximisation du critère.

2. Modèle et problème:

2.1. Modèle thématique

Conceptuellement, tout groupe de variables $X_r = \left(x_r^1, \ldots, x_r^{J_r}\right)$ peut dépendre de, ou contribuer à modéliser d'autres groupes. Le réseau de dépendances est le *modèle thématique* (cf. fig. 1). Un groupe peut faire l'objet de plusieurs modélisations; notons E_r^h le modèle h du groupe dépendant X_r. Les liaisons sont plus précisément supposées concerner des dimensions sous-jacentes aux groupes. Nous voulons faire en sorte que les dimensions de X_r puissent être révélées à travers K_r composantes $f_r^1, \ldots, f_r^{K_r}$, où $\forall r, k : f_r^k = X_r v_r^k$.

Figure 1: Modèle thématique



2.2. Problème

On attend des composantes qu'elles donnent à leur modèle un bon ajustement, mais aussi qu'elles aient quelque force structurelle, i.e. restituent une part non résiduelle de la variance de leur groupe, de sorte que le modèle final soit robuste et aisé à interpréter. À chaque groupe X_r, on associe une métrique euclidienne M_r. La force structurelle de la composante $f_r^k = X_r v_r^k$ est mesurée par une fonction de la forme: $S(v_r^k) = \sum_{t=1}^{T} (v_r^{k\prime} A_{rt} v_r^k)^a$ où les matrices A_{rt} sont symétriques positives (s.p.) et $a > 0$. Généralement, v_r^k est contraint par: $v_r^{k\prime} D_r v_r^k = 1$, où D_r est une matrice définie s.p. appropriée. Nous considérons les cas particuliers suivants:

- Variance de la composante: $S(v_r^k) = V(f_r^k) = v_r^{k\prime} X_r{}' P X_r v_r^k$; $a = 1$; $T = 1$; $A_r = X_r{}' P X_r$; $D_r = M_r^{-1}$.

- Variance du groupe restituée par la composante. Toutes les variables étant réduites, on prend: $S(v_r^k) = \sum_{x_r^j \in X_r} \rho^2(f_r^k, x_r^j) = \sum_{x_r^j \in X_r} \langle f_r^k | x_r^j \rangle_P^2 = \sum_{x_r^j \in X_r} f_r^{k\prime} P x_r^j x_r^{j\prime} P f_r^k = v_r^{k\prime} (X_r{}' P X_r)^2 v_r^k$, où $D_r = X_r{}' P X_r$, de sorte que $\|f_r^k\|_P^2 = 1$; $a = 1$, $T = 1$, $A_r = (X_r{}' P X_r)^2$

- Corrélation variable-composante avec exposant pair. Toutes les variables étant réduites, on prend:

$$S(v_r^k) = \sum_{x_r^j \in X_r} a_j \rho^{2a}(f_r^k, x_r^j) = \sum_{x_r^j \in X_r} a_j \langle X_r v_r^k | x_r^j \rangle_P^{2a} = \sum_{j=1}^{J_r} a_j (v_r^{k\prime} X_r{}' P x_r^j x_r^{j\prime} P X_r v_r^k)^a,$$

avec $D_r = X_r{}'PX_r$. Alors: $T = J_r$, $A_{rj} = a_j^{1/a} w_j w_j{}'$ où $w_j = X_r{}'Px_r^j$. La motivation d'un tel critère est la même que quartimax des CP: rapprocher les composantes d'éventuels faisceaux de variables. Le paramètre a permet de régler l'attraction de la composante vers les faisceaux locaux (a plus élevé pour une attraction plus forte). Les pondérations a_j permettent l'encodage de thèmes flous partiellement superposés.

Le group X_r dépendant d'autres groupes $X_s, \ldots, X_t$ dans un certain modèle E_r^h, chaque composante f_r^k de X_r est modélisée linéairement en fonction des composantes de ces groupes $\{f_s^l, s{\in}P_r^h, l \leq K_s\}$, où P_r^h est l'ensemble d'indices des groupes prédicteurs dans E_r^h. La qualité d'ajustement du modèle de f_r^k est simplement mesurée par son coefficient R^2, noté $R^2(f_r^k|\{f_s^l, s \in P_r^h, l \leq K_s\})$, ou plus simplement $R^2(E_r^h)$.

3. Exploration du modèle thématique

3.1. Covariance Multiple Généralisée

Covariance multiple

La covariance multiple proposée dans Bry *et al.* (2009) est l'une des extensions possibles de la covariance binaire classique (en valeur absolue).
Définition: Soit y une variable modélisée comme combinaison linéaire des variables $x^1, \ldots, x^S$, la covariance multiple de y sur $x^1, \ldots, x^S$ est définie comme:

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

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

Covariance Multiple Généralisée

Étant donnée une mesure de force structurelle $S(v)$ d'une composante $f = Xv$, et le modèle linéaire entre composantes: $f_d = \sum_{p\in P(d)} b_p f_p + \varepsilon$, $P(d)$ étant l'ensemble d'indices des composantes prédictrices de f_d, nous appellerons *covariance multiple généralisée* (CMG) de f_d sur $\{f_p, p{\in}P(d)\}$:

$$CMG(f_d|\{f_p, p \in P(d)\}) \;=\; \left[S(v_d)\left(\prod_{p\in P(d)} S(v_p)\right) R^2(f_d|\{f_p, p \in P(d)\})\right]^{1/2}$$

3.2. Le critère de THEME pour les composantes de rang 1

Soit E_r l'ensemble des modèles impliquant f_r, et e_r leur nombre. Le premier critère dont nous proposons la maximisation par les composantes de rang 1 est:

$$C_1 \;=\; \prod_{r=1}^{R} (S(v_r))^{e_r} \prod_{d,h} R^2(f_d|\{f_s, s \in P_d^h\})$$

C_1 amalgame la force structurelle des composantes avec la qualité d'ajustement du modèle de régression reliant les composantes.

3.3. Au delà d'une composante par groupe

Lorsque l'on veut plus généralement extraire K_r composantes du groupe X_r, on doit tenir compte de ce que les composantes prédictrices du modèle $E_d^h : \{f_p^k; p{\in}P_d^h, \forall p : k = 1, K_p\}$ doivent prédire non une mais K_d composantes $\{f_d^k;\ k = 1,\ K_d\}$ dans le groupe dépendant X_d. Donc, la recherche de composantes explicatives doit être fondée sur l'optimisation d'un critère de CMG agrégé. Notons que:

$$\sum_{k=1}^{K_d} CMG^2(f_d^k|\{f_p^l|p \in P_d^h, l \leq K_p\}) \;=\; \left(\prod_{p\in P_d^h; l\leq K_p} S(v_p^l)\right) \sum_{k=1}^{K_d} S(v_d^k) R^2(f_d^k|\{f_p^l|p \in P_d^h, l \leq K_p\})$$

Un principe d'emboîtement doit par ailleurs être défini de sorte à rendre l'ordre des composantes interprétable.

Le principe d'emboîtement local (PEL)

Notre PEL stipule que:

a) La composante de rang k du groupe X_r : f_r^k, doit servir à prédire au mieux - au regard de la CMG - *toutes* les composantes d'un groupe X_d qui en dépend selon un modèle E_d^h, lorsqu'elle est associée à *toutes* les composantes prédisant ce groupe dans E_d^h à l'exception des composantes de rang plus élevé de X_r's, considérées comme non encore disponibles, i.e.:

$$\{f_r^l;\ l < k\} \cup \{f_t^l; r{\in}P_d^h, t{\in}P_d^h, t \neq r, l \leq K_t\}$$

b) La composante de rang k du groupe X_r: f_r^k, doit être prédite au mieux - au regard du critère - par toutes les composantes des groupes explicatifs de X_r selon un certain modèle, sous la contrainte d'orthogonalité:

$$\forall l < k : f_r^{k\prime} P f_r^l = 0$$

Ce principe séquentiel n'est pas a priori compatible avec l'existence d'un critère global que l'ensemble de toutes les composantes maximiserait. Mais il est aisé de définir, pour chaque composante f_r^k, le critère que selon le PEL, elle devrait maximiser:

$$\left(S(v_r^k)\right)^{e_r} \left(\prod_{h|P_r^h\neq\emptyset} R^2(f_r^k|\{f_p^l, p \in P_r^h, l \leq K_p\})\right) \prod_{(s,h)|r\in P_s^h} \left(\sum_{l=1}^{K_s} S(v_s^l) R^2(f_s^l|\{f_q^m, q \in P_s^h, m \leq K_q\})\right)$$

Maximisation du critère par la composante courante f_r^k

Nous montrons que le critère (1) peut être mis sous la forme générique:

$$\left(\sum_{t=1}^{T}(v_r^{k\prime}A_{rkt}v_r^k)^a\right)^{e_r} \quad \prod_{l=1}^{e_r}\frac{f_r^{k\prime}C_{rkl}f_r^k}{f_r^{k\prime}D_{rkl}f_r^k}$$

Par suite, nous montrons que sa maximisation contrainte équivaut à la minimisation libre d'une certaine fonction φ. La minimisation de φ peut être obtenue à l'aide d'outils généralistes, mais nous proposons également un algorithme spécifique aux bonnes performances, et facile à programmer.

4. Application

Diverses simulations seront présentées, ainsi qu'une application sur données chimiométriques: à partir des caractéristiques chimiques du tabac et physiques du design de la cigarette, on produira un modèle explicatif et prédictif des composés de la fumée.

Bibliographie

[1] Bry X., Verron T., Cazes P. (2009) : *Exploring a physico-chemical multi-array explanatory model with a new multiple covariance-based technique: Structural equation exploratory regression*, Anal. Chim. Acta, 642 45–58.

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

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

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

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

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

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

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

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