# THEME : Une méthode et un logiciel pour l'analyse exploratoire multidimensionnelle d'un modèle structurel.

**X. Bry**[a], **T. Verron**[b] and **P. Redont**[a]

[a]Institut de Mathématiques et de Modélisation de Montpellier
Université Montpellier II
Place Eugène Bataillon CC 051 - 34095, Montpellier, France
xavier.bry@univ-montp2.fr
patrick.redont@univ-montp2.fr

[b]ITG-SEITA
Centre de recherche SCR
4 rue André Dessaux, 45000 Fleury les Aubrais, France
thomas.verron@fr.imptob.com

**Mots clefs** : Approche PLS, Équations Structurelles, PLS, SEER, THEME.

Le contexte est celui des modèles à équations structurelles (SEM) : $R$ groupes de variables observées, $X_1, ..., X_R$ décrivant les mêmes $n$ unités sont supposés structurés autour d'un nombre restreint de variables latentes (non observées) $f$ reliées entre elles par un modèle linéaire (ML). Considérons par exemple 5 groupes : $X_1$, structuré autour de deux variables latentes $\{f_1^1, f_1^2\} := F_1$, $X_2$ autour d'une seule, $\{f_2\} := F_2$, $X_3$ autour de $\{f_3\} := F_3$, $X_4$ autour de $\{f_4^1, f_4^2, f_4^3\} := F_4$ et $X_5$ autour de $\{f_5^1, f_5^2\} := F_5$. On suppose savoir que sur le plan causal, $X_3$ dépend de $X_1$ et $X_2$, tandis que $X_5$ dépend de $X_3$ et $X_4$. Les équations structurelles à estimer sont alors :

$$F_3 = ML(F_1, F_2)\,;\; F_5 = ML(F_3, F_4)$$

Les SEM sont couramment traités de façon restrictive, en supposant que les variables observées de chaque groupe $X_r$ reflètent une seule et même variable latente, qu'il s'agit d'estimer. Deux démarches sont alors couramment adoptées. La première, PLS Path Modeling (PLSPM) [2,5,7], n'étant fondée sur l'optimisation d'aucun critère global, reste purement empirique. Sa fondation théorique est insuffisante, notamment parce que les liaisons partielles, qui sont l'essence des modèles de régression multiple, n'y sont pas correctement prises en compte. La seconde consiste à optimiser un critère global interprétable issu d'une modélisation statistique véritable. Ces modèles permettent une prise en compte correcte des liaisons partielles entre variables. Selon le critère choisi, on obtient différentes méthodes ([4,6,3]). Cette approche plus rigoureuse se paye souvent de problèmes de convergence, notamment dans le traitement des petits échantillons. Récemment, [10] ont proposé une méthode, RGCCA, inspirée de PLSPM mais contrairement à celle-ci, fondée sur l'optimisation d'un critère global souple, ce qui permet d'en assurer la convergence. Elle souffre cependant du même défaut de prise en compte des liaisons partielles que PLSPM.

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 [8,9] 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 arbitraires.
La méthode générale que nous proposons, THEME [11], consiste à rechercher dans chaque groupe de variables un petit nombre de composantes (combinaisons linéaires des variables) qui 1) possèdent une certaine force structurelle (i.e. s'approchent autant que possible des dimensions de mesure que sont les variables du groupe, s'éloignant au contraire des dimensions de bruit), et 2) satisfont le système d'équations structurelles qui relie les groupes. THEME utilise la maximisation d'un critère qui prend en compte les liaisons partielles entre composantes, afin d'extraire 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 par validation croisée. Par ailleurs, le critère que nous proposons permet de définir de façon très flexible ce que nous entendons par "force structurelle". Par exemple, on peut demander à une composante de capter une part importante de la variance totale du groupe dont elle est issue (optique ACP), ou bien de mettre en évidence un faisceau de variables (optique quartimax) de taille variable, entre autres possibilités.
THEME est implémentée sous forme d'un logiciel développé en R. Une application en est faite à des données chimiométriques, avec variation des divers paramètres et comparaison des résultats.

**Références**


[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.*
[10] Tenenhaus A, Tenenhaus M. Psychometrika 2011 ; 76 : 257-284.
[11] Bry X., Verron T., Redont P., Cazes P. (2012) : *THEME-SEER : a multidimensional exploratory technique to analyze a structural model using an extended covariance criterion* - Journal of Chemometrics, 26, pp 158-169.