Utilisation des modèles à équations structurelles en analyse sensorielle

Michel Tenenhaus
HEC School of Management (GRECHEC),
1 rue de la Libération, Jouy-en-Josas, France
tenenhaus@hec.fr

Résumé

Deux écoles concurrentes se sont imposées dans le domaine de la modélisation des équations structurelles. La première approche, appelée « Covariance-based SEM » (*SEM = Structural Equation Modeling*), s'est développée autour de Karl Jöreskog. Cette approche a pour objectif de modéliser la matrice de covariance des variables observées. Elle peut être considérée comme une généralisation de l'analyse factorielle en facteurs communs et spécifiques au cas de plusieurs tableaux de données reliés entre eux par des liens de causalité. La seconde approche s'est développée autour de Herman Wold sous le nom de « PLS » (*Partial Least Squares*), puis de « Component-based SEM ». Elle a pour objectif la construction de scores résumant au mieux chaque bloc de variables tout en tenant compte du réseau de causalité. C'est une généralisation de l'analyse en composantes principales au cas de plusieurs tableaux de données reliés entre eux par des liens de causalité. Plus récemment Hwang et Takane (2004) ont proposé une méthode « Component-based SEM » : la méthode GSCA (*Generalized Structural Component Analysis*). Elle permet une recherche de scores optimisant un critère global. Nous allons discuter de l'utilisation de ces trois approches en donnant des exemples d'utilisation en analyse sensorielle.

Mots-clés : PLS, ULS, SEM, GSCA.

1. Utilisation de la méthode d'estimation ULS en modélisation d'équations structurelles

Pour une présentation détaillée des modèles à équations structurelles, on peut consulter Bollen (1989). Un modèle à équations structurelles est formé de deux modèles : le modèle structurel et le modèle de mesure.

Le modèle structurel

On considère un ensemble de variables latentes centrées (variables inobservables) reliées entre elles par des liens de causalité. Notons $\boldsymbol{\eta}$ un vecteur colonne formé de m variables latentes endogènes (c'est à dire expliquées par d'autres variables latentes) et $\boldsymbol{\xi}$ un vecteur colonne formé de k variables latentes exogènes (c'est à dire purement explicatives). Le modèle structurel reliant le vecteur $\boldsymbol{\eta}$ aux vecteurs $\boldsymbol{\eta}$ et $\boldsymbol{\xi}$ s'écrit

$$\boldsymbol{\eta} = \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta} \quad (1)$$

où $\mathbf{B}$ est une matrice $m \times m$ de coefficients de régression à diagonale nulle, $\boldsymbol{\Gamma}$ une matrice $m \times k$ de coefficients de régression et $\boldsymbol{\zeta}$ un vecteur aléatoire de dimension m. On suppose que la matrice $\mathbf{I-B}$ est inversible.

Le modèle de mesure

Chaque variable latente (inobservable) est décrite par un ensemble de variables manifestes (observables) supposées centrées. Le vecteur $\mathbf{y}_j = (y_{j1},...,y_{jp_j})'$ des variables manifestes reliées à la variable latente endogène η_j peut s'écrire en fonction de η_j : $\mathbf{y}_j = \boldsymbol{\lambda}_j^y \eta_j + \boldsymbol{\varepsilon}_j$. Le vecteur $\mathbf{y} = (\mathbf{y}_1',...,\mathbf{y}_m')'$ s'écrit

$$\mathbf{y} = \boldsymbol{\Lambda}_y \boldsymbol{\eta} + \boldsymbol{\varepsilon} \quad (2)$$

où $\mathbf{\Lambda_y}$ est une matrice diagonale par bloc, chaque bloc étant formé du vecteur $\boldsymbol{\lambda}_j^y$, et $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_1', ..., \boldsymbol{\varepsilon}_m')'$.

De même le vecteur **x** des variables manifestes reliées aux variables latentes exogènes s'écrit en fonction de $\boldsymbol{\xi}$:

(3) $$\mathbf{x} = \mathbf{\Lambda_x}\boldsymbol{\xi} + \boldsymbol{\delta}$$

Décomposition de la matrice des covariances des variables manifestes

Posons $\mathbf{\Phi} = Cov(\boldsymbol{\xi}) = E(\boldsymbol{\xi}\boldsymbol{\xi}')$, $\boldsymbol{\psi} = Cov(\boldsymbol{\zeta}) = E(\boldsymbol{\zeta}\boldsymbol{\zeta}')$, $\mathbf{\Theta_\varepsilon} = Cov(\boldsymbol{\varepsilon}) = E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')$, et $\mathbf{\Theta_\delta} = Cov(\boldsymbol{\delta}) = E(\boldsymbol{\delta}\boldsymbol{\delta}')$. On suppose que les vecteurs aléatoires $\boldsymbol{\xi}$, $\boldsymbol{\zeta}$, $\boldsymbol{\varepsilon}$, et $\boldsymbol{\delta}$ sont indépendants entre eux et que les matrices de covariances $\mathbf{\Psi}$, $\mathbf{\Theta_\varepsilon}$, $\mathbf{\Theta_\delta}$ des termes d'erreur sont diagonales. On obtient alors la décomposition suivante de la matrice des covariances entre les variables manifestes :

(4) $$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_{xx} & \mathbf{\Sigma}_{xy} \\ \mathbf{\Sigma}_{yx} & \mathbf{\Sigma}_{yy} \end{bmatrix} = \begin{bmatrix} \mathbf{\Lambda_x}\mathbf{\Phi}\mathbf{\Lambda_x}' + \mathbf{\Theta_\delta} & \mathbf{\Lambda_x}\mathbf{\Phi}\mathbf{\Gamma}'\left[(\mathbf{I}-\mathbf{B})'\right]^{-1}\mathbf{\Lambda_y}' \\ \mathbf{\Lambda_y}\left[(\mathbf{I}-\mathbf{B})\right]^{-1}\mathbf{\Gamma}\mathbf{\Phi}\mathbf{\Lambda_x}' & \mathbf{\Lambda_y}[(\mathbf{I}-\mathbf{B})^{-1}(\mathbf{\Gamma}\mathbf{\Phi}\mathbf{\Gamma}'+\mathbf{\Psi})][(\mathbf{I}-\mathbf{B})']^{-1}\mathbf{\Lambda_y}' + \mathbf{\Theta_\varepsilon} \end{bmatrix}$$

Notons $\Omega = \left\{\mathbf{\Lambda_x}, \mathbf{\Lambda_y}, \mathbf{B}, \mathbf{\Gamma}, \mathbf{\Phi}, \mathbf{\Psi}, \mathbf{\Theta_\varepsilon}, \mathbf{\Theta_\delta}\right\}$ l'ensemble des paramètres du modèle et $\mathbf{\Sigma}(\Omega)$ la matrice (4). La méthode proposée par McDonald (1996) consiste à utiliser la méthode d'estimation *ULS* (*Unweighted Least Squares*) en *SEM* en mettant à 0 les estimations $\hat{\mathbf{\Theta}}_\varepsilon$ et $\hat{\mathbf{\Theta}}_\delta$ des matrices de covariances $\mathbf{\Theta_\varepsilon}$ et $\mathbf{\Theta_\delta}$. Il s'agit de rechercher les paramètres $\hat{\mathbf{\Lambda}}_x, \hat{\mathbf{\Lambda}}_y, \hat{\mathbf{B}}, \hat{\mathbf{\Gamma}}, \hat{\mathbf{\Phi}}, \hat{\mathbf{\Psi}}$ minimisant le critère

(5) $$\left\| \mathbf{S} - \mathbf{\Sigma}(\hat{\mathbf{\Lambda}}_\mathbf{x}, \hat{\mathbf{\Lambda}}_\mathbf{y}, \hat{\mathbf{B}}, \hat{\mathbf{\Gamma}}, \hat{\mathbf{\Phi}}, \hat{\mathbf{\Psi}}, \mathbf{0}, \mathbf{0}) \right\|^2$$

Il s'agit donc de rechercher une décomposition de la matrice des covariances empirique en fonction des paramètres du modèle structurel et du modèle de mesure. Les estimations des matrices de covariances $\mathbf{\Theta_\varepsilon} = Cov(\boldsymbol{\varepsilon})$, $\mathbf{\Theta_\delta} = Cov(\boldsymbol{\delta})$ des termes résiduels du modèle de mesure sont mises à 0. Autrement dit les estimations des variances des termes résiduels du modèle de mesure sont intégrées dans les termes de la diagonale de la matrice des erreurs de reconstitution $\mathbf{S} - \mathbf{\Sigma}(\hat{\mathbf{\Lambda}}_\mathbf{x}, \hat{\mathbf{\Lambda}}_\mathbf{y}, \hat{\mathbf{B}}, \hat{\mathbf{\Gamma}}, \hat{\mathbf{\Phi}}, \hat{\mathbf{\Psi}}, \mathbf{0}, \mathbf{0})$. Dans la mesure où l'approche de McDonald est une généralisation de l'analyse en composantes principales, elle peut être utilisée avec un petit nombre d'observations et une matrice de covariance observée **S** de rang non plein. Cette approche est donc parfaitement utilisable en analyse sensorielle.

2. L'approche de McDonald pour estimer les variables latentes

Comme il est d'usage dans l'approche PLS (voir section 3), nous désignons maintenant une variable latente par la lettre ξ, que son type soit endogène ou exogène. McDonald (1996) propose d'évaluer la variable latente ξ_j à l'aide de la formule

(6) $$\hat{\xi}_j \propto \sum_k \tilde{w}_{jk} x_{jk}$$

où $\tilde{w}_{jk} = \widetilde{Cov}(x_{jk}, \xi_j)$ et où $\propto$ signifie que le terme de gauche représente le terme de droite centré-réduit. Le coefficient de régression λ_{jk} de la variable latente ξ_j dans la régression de la variable manifeste x_{jk} sur la variable latente ξ_j est estimé par

(7) $$\hat{\lambda}_{jk} = \widetilde{Cov}(x_{jk}, \xi_j) / \widetilde{Var}(\xi_j)$$

Nous en déduisons que la formule (6) peut aussi s'écrire

(8) $$\hat{\xi}_j \propto \sum_k \hat{\lambda}_{jk} x_{jk}$$

Notons $\mathbf{X}_j$ le tableau des variables manifestes x_{jk} relatives à la variable latente ξ_j. La variable latente estimée centrée-réduite peut finalement s'écrire

(9) $$\hat{\boldsymbol{\xi}}_j = \mathbf{X}_j \mathbf{w}_j$$

Ainsi l'approche de McDonald revient à estimer la variable latente ξ_j à l'aide de la première composante PLS centrée-réduite estimée dans la régression PLS de la variable latente ξ_j sur les variables manifestes $x_{j1},...,x_{jp_j}$. Nous avons utilisé cette approche sur l'exemple de dégustation de vins rouge de la Loire de Pagès, Asselin, Morlat & Robichet (1987). Cet exemple portait sur 21 vins et 27 variables. Nous avons également montré l'efficacité de cette approche dans une étude portant sur 6 jus d'orange. Les résultats de ces études sont disponibles dans Tenenhaus (2007a).

3. L'approche PLS

Les développement les plus récents de l'approche PLS sont décrits en détail dans Tenenhaus, Vinzi Esposito, Chatelin & Lauro (2005). Nous allons résumer ici l'algorithme.

Le modèle structurel (1) est plutôt appelé modèle interne dans l'approche PLS. Le modèle de mesure reliant les variables manifestes à leurs variables latentes est lui appelé modèle externe. Lorsque ce modèle correspond aux équations (2) ou (3), on est dans un mode réflectif. Il peut aussi arriver que le lien entre les variables manifestes et la variable latente soit inversé. C'est la variable latente ξ_j qui est fonction de ses variables manifestes à travers l'équation de régression suivante :

(10) $$\xi_j = \sum_h \tau_{jh} x_{jh} + \delta_j$$

Il s'agit du mode formatif.

On appelle estimation externe de la variable latente ξ_j la variable centrée-réduite $\hat{\boldsymbol{\xi}}_j = \mathbf{X}_j \mathbf{w}_j$. On définit également une estimation interne $\mathbf{z}_j$ de la variable latente ξ_j à l'aide de la formule

(11) $$\mathbf{z}_j = \sum_{\xi_k \text{ connectée à } \xi_j} e_{jk} \hat{\boldsymbol{\xi}}_k$$

Trois schémas de définition sont utilisés pour construire les coefficients e_{jk}. Dans le schéma centroïde, $e_{jk} = signe(cor(\hat{\xi}_j, \hat{\xi}_k))$. Dans le schéma factoriel, $e_{jk} = cor(\hat{\xi}_j, \hat{\xi}_k)$. Enfin, dans le schéma structurel, la valeur de e_{jk} dépend du statut de la variable latente ξ_j par rapport à ξ_k. Si la VL ξ_j est expliquée par la VL ξ_k, alors e_{jk} est égal au coefficient de régression de $\hat{\xi}_k$ dans la régression de $\hat{\xi}_j$ sur l'ensemble de ses VL explicatives. Si par contre la VL ξ_j explique la VL ξ_k, alors $e_{jk} = cor(\hat{\xi}_j, \hat{\xi}_k)$.

Deux modes de calcul sont utilisés pour calculer les poids $\mathbf{w}_j$:

Dans le mode A :

(12) $$w_{jh} \propto Cov(\mathrm{z}_j, \mathrm{x}_{jh})$$

Dans le mode B :

(13) $$\mathrm{w}_j \propto (\mathrm{X}_j' \mathrm{X}_j)^{-1} \mathrm{X}_j' \mathrm{z}_j$$

où le signe $\propto$ indique maintenant que le vecteur $\mathbf{w}_j$ est normalisé de manière à obtenir une VL estimée centrée-réduite. Le mode A est plutôt adapté au mode réflectif et le mode B au mode formatif.

L'algorithme PLS est itératif. A l'étape initiale, on choisit des vecteurs de poids $\mathbf{w}_j$ arbitraires. Le plus courant est de choisir le vecteur $\mathbf{w}_j$ avec toutes ses coordonnées égales. A l'aide de l'équation (11) on

calcule les estimations internes $\mathbf{z}_j$, puis à l'aide des équations (12) ou (13) de nouveaux poids $\mathbf{w}_j$. Cet algorithme converge pratiquement toujours dans la pratique.

Nous avons utilisé cette approche PLS sur un exemple de dégustation de jus d'orange. Cet exemple portait sur 6 jus d'orange et 53 variables. Les résultats de cette étude ont été publiés dans Tenenhaus, J. Pagès, L. Ambroisine, C. Guinot, (2005).

L'approche de McDonald décrite dans la section 2 peut entrer dans le cadre PLS. Dans l'approche PLS habituelle sous le mode A, on obtient aussi les poids w_{jk} par régression simple de chaque variable x_{jk} sur l'estimation interne $\mathbf{z}_j$ de la variable latente ξ_j. Il est nécessaire de calculer explicitement l'estimation interne $\mathbf{z}_j$ pour obtenir ces poids. Trois procédures sont proposées dans les logiciels PLS : les schémas centroïde, factoriel et structurel décrits plus haut. Par contre les logiciels SEM fournissent directement les poids (saturations ou *loadings*) représentant pour chaque x_{jk} une estimation du coefficient de régression de la variable latente « théorique » ξ_j dans la régression de x_{jk} sur ξ_j. On peut donc utiliser, à la place du coefficient de régression de l'estimation interne $\mathbf{z}_j$, le coefficient de régression estimé de la variable latente « théorique » ξ_j. Nous avons proposé cette procédure de calcul des poids tout simplement basée sur les sorties des logiciels SEM dans Tenenhaus, Vinzi Esposito, Chatelin & Lauro (2005). Nous l'avions nommée « le schéma LISREL » et retrouvé alors sans le savoir le choix proposé par McDonald.

4. La méthode GSCA de Hwang et Takane

En 2004, Hwang et Takane ont proposé la méthode GSCA (*Generalized Structured Component Analysis*). C'est une méthode concurrente de l'approche PLS. Son principal avantage est d'être basée sur un critère global à optimiser. Présentons ce critère. Les variables latentes sont toutes notées ξ_j. On distingue cependant les variables latentes endogènes et exogènes, ainsi que les réflectives et formatives. On note $\hat{\boldsymbol{\xi}}_j = \mathbf{X}_j \mathbf{w}_j$ une estimation centrée-réduite de la variable latente ξ_j. On recherche les poids $\mathbf{w}_j$ minimisant le critère

$$(14) \qquad \sum_{\xi_j \text{ réflectif}} \left\| \mathbf{X}_j - \hat{\boldsymbol{\xi}}_j \mathbf{c}_j' \right\|^2 + \sum_{\substack{\xi_j \text{ endogène,} \\ \xi_k \text{ expliquant } \xi_j}} \left\| \hat{\boldsymbol{\xi}}_j - \sum_k b_{jk} \hat{\boldsymbol{\xi}}_k \right\|^2$$

où $\mathbf{c}_j$ est un vecteur de poids associés aux variables manifestes x_{jh}.

Une variable latente ξ_j peut aussi être une variable latente du second ordre. C'est une VL non directement reliée à des variables manifestes, mais reliée à d'autres variables latentes dans un sens de causalité. Par exemple, si ξ_1 et ξ_2 sont des variables latentes reliées à des variables manifestes, une variable latente ξ_3 est du second ordre relative à ξ_1 et ξ_2 si ξ_3 est cause de ξ_1 et ξ_2. Dans ce cas, on estime la variable latente ξ_3 à l'aide de la formule

$$(15) \qquad \hat{\boldsymbol{\xi}}_3 = w_{31} \hat{\boldsymbol{\xi}}_1 + w_{32} \hat{\boldsymbol{\xi}}_2 .$$

Nous avons aussi exploré l'utilisation de la méthode GSCA en analyse des tableaux multiples, appelée aussi analyse factorielle confirmatoire dans l'univers de la modélisation des équations structurelles. Si l'on suppose chaque bloc réflectif, la recherche des premières composantes centrées-réduites $\hat{\boldsymbol{\xi}}_j = \mathbf{X}\mathbf{w}_j$ de chaque bloc $\mathbf{X}_j$ et d'une première composante globale centrée-réduite $\hat{\boldsymbol{\xi}} = \sum_j w_j \hat{\boldsymbol{\xi}}_j$ est menée en minimisant le critère

$$(16) \qquad \sum_j \left\| \mathbf{X}_j - \hat{\boldsymbol{\xi}}_j \mathbf{c}_j' \right\|^2 + \sum_j \left\| \hat{\boldsymbol{\xi}}_j - \hat{\boldsymbol{\xi}} \right\|^2$$

Si l'on suppose chaque bloc formatif, la recherche des premières composantes centrés-réduites $\hat{\xi}_j = \mathbf{X}\mathbf{w}_j$ de chaque bloc $\mathbf{X}_j$ et d'une première composante globale centrée-réduite $\hat{\xi} = \sum_j w_j \hat{\xi}_j$ est alors menée en minimisant le critère

(17) $$\sum_j \left\| \hat{\xi}_j - \hat{\xi} \right\|^2 = \sum_j \left\{ \left\| \hat{\xi}_j \right\|^2 + \left\| \hat{\xi} \right\|^2 - 2\hat{\xi}_j' \hat{\xi} \right\}$$

Les variables latentes $\hat{\xi}_j = \mathbf{X}_j \mathbf{w}_j$ et $\hat{\xi} = \sum_j w_j \hat{\xi}_j$ étant centrées réduites, on retrouve le critère

(18) $$\text{Maximiser} \sum_j Cor(\hat{\xi}_j, \hat{\xi})$$

Nous présentons l'application de cette approche sur des données de satisfaction de consommateurs dans Tenenhaus (2008b).

Références

Bollen, K. A. (1989) : *Structural equations with latent variables.* John Wiley & Sons.

Hwang, H. & Takane Y. (2004) : Generalized structured component analysis, *Psychometrika*, 69, 1, 81-99.

McDonald, R.P. (1996) : Path analysis with composite variables, *Multivariate Behavioral Research*, 31 (2), 239-270.

Pagès J., Asselin C., Morlat R., Robichet J. (1987): Analyse factorielle multiple dans le traitement de données sensorielles : Application à des vins rouges de la vallée de la Loire, *Sciences des aliments*, 7, 549-571.

Tenenhaus M. (2008a): Structural Equation Modelling for small samples. *Working paper* n° 885. HEC Paris, Jouy-en-Josas.

Tenenhaus M. (2008b): Component-based Structural Equation Modelling for small samples. *Total Quality Management & Business Excellence.* Submitted.

Tenenhaus M., Esposito Vinzi V., Chatelin Y.-M., Lauro C. (2005) : PLS path modeling. *Computational Statistics & Data Analysis,* 48, 159-205.

Tenenhaus, M. & Hanafi M. (2008) : A bridge between PLS path modelling and multi-block data analysis », in *Handbook of Partial Least Squares (PLS): Concepts, Methods and Applications* (V. Esposito Vinzi, W. Chin, J. Henseler, H. Wang, Eds), Volume II in the series of the Handbooks of Computational Statistics, Springer, à paraître.

Tenenhaus M., Pagès J., Ambroisine L., Guinot C. (2005) : PLS methodology to study relationships between hedonic judgments and product characteristics, *Food Quality and Preference*, 16, 315-325.