Estimation par algorithme EM pour modèles à facteurs et à équations structurelles

Myriam Tami & Xavier Bry & Christian Lavergne

I3M, Universités de Montpellier, myriam.tami@univ-montp2.fr, xavier.bry@univ-montp2.fr, christian.lavergne@univ-montp2.fr

Résumé. Introduits dans les années 1970 par Jöreskog, les modèles d'équations structurelles à facteurs permettent de mettre en relation des variables non observables, dites latentes. À l'origine, ces modèles avaient, pour seule méthode d'estimation des paramètres, l'analyse de la structure de covariance connue sous le nom de LISREL (LInear Structural RELations). Cette dernière peut être fondée sur un critère de maximum de vraisemblance ou d'autres critères. Depuis, d'autres méthodes plus rapides, car fondées sur des composantes, ont été proposées. Il y a eu tout d'abord, PLS-PM (Partial Least Squares Path Modeling) de Wold (1985), suivie par GSCA (Generalized Structured Component Analysis) de Hwang et Takane (2004) puis RGCCA (Regularized Generalized Canonical Correlation Analysis) de Tenenhaus (2011) et THEME (THematic Equation Model Explorer) de Bry (2012). Nous proposons ici, dans le cadre des modèles d'équations structurelles à facteurs, une méthode d'estimation fondée sur l'algorithme EM. Nous présenterons une application de cette méthode sur des données réelles.

Mots-clés. Modèles à équations structurelles (MES), Modèles à facteurs, variables latentes, estimation, algorithme EM, LISREL.

Abstract. Introduced in the 1970s by Jöreskog, structural equation models with factors allow to connect unobservable variables, called latent. Originally, the only parameter estimation method used on that models was analysis of covariance structure known as LISREL (Linear Structural RELations). This one may be based on a criterion of maximum likelihood or other criteria. Since then, faster methods based on components have been proposed. The first one was PLS -PM (Partial Least Squares Path Modeling) by Wold (1985), followed by GSCA (Generalized Structured Component Analysis) by Hwang and Takane (2004), RGCCA (Regularized Generalized Canonical Correlation Analysis) by Tenenhaus (2011) and THEME (THematic Equation Model Explorer) by Bry (2012). We propose in the context of structural equation models with factors, an estimation method based on the EM algorithm. We present an application of this method on real data.

Keywords. Structural equation modeling (SEM), factor models, latent variables, estimation method, EM algorithm, LISREL.

1 Introduction

Nous nous plaçons dans le contexte des modèles à équations structurelles (MES) où plusieurs groupes de variables Y, X^1,..., X^P décrivent les mêmes n unités. Chacun des groupes de variables est supposé refléter une seule variable latente, laquelle est aussi en relation avec d'autres variables latentes. Les variables latentes sont liées par des relations de causalité qui constituent le modèle interne dit aussi structurel. En outre, la relation entre chaque variable latente et les variables observées qui la reflète est spécifiée par le modèle externe (cf. fig. 1). Pour expliciter l'intensité des relations entre les différents constituants du modèle complet (interne + externe), nous cherchons à estimer les paramètres du modèle. Dans la littérature, deux courants de pensée s'affrontent. L'un développé par Herman Wold (1975) et ses successeurs utilisant comme variables latentes des composantes. Ce courant est à l'origine par exemple de la méthode des moindres carrés partiels (PLS) et d'autres citées plus haut. Le second issu des travaux de Karl Jöreskog (1970) est fondé sur le maximum de vraisemblance et utilise comme variables latentes des facteurs. La méthode engendrée porte le nom de LISREL. Ces deux approches ont été comparées dans plusieurs travaux, celui de Stan et Saporta (2006) en est un exemple. Dans un objectif d'interprétation a priori du modèle, il est préférable d'utiliser un modèle à facteurs à un modèle à composantes lequel est computationnellement plus efficace mais trop contraint. Ainsi, nous proposons ici une méthode d'estimation d'un modèle structurel à facteurs dont l'estimation par le maximum de vraisemblance est réalisé via l'algorithme EM. En effet, celui ci constitue une méthodologie générale d'estimation d'un modèle en présence de variables latentes.

2 Structure générale du modèle

2.1 Modèle à P groupes explicatifs et un groupe dépendant

Nous nous plaçons dans le cadre d'un modèle d'équations structurelles à variables latentes où pour $p \in [\![1, P]\!]$ chacun des groupes de variables Y, X^1,..., X^P dépend d'une variable latente (respectivement de G, F^1, ..., F^P). Au niveau du modèle interne, la variable latente G est dépendante de l'ensemble des variables latentes F^1, ..., F^P (cf. fig. 1). Quant aux modèles externes, chacun forme un groupe de variables $X^p = (X_1^p, ...X_{q_p}^p)$ (resp. $Y = (Y_1, ...Y_{q_Y})$) comme dépendant du facteur F^p (resp. G).
Notons que ce modèle peut se généraliser à plusieurs facteurs dans chaque groupe sans difficultés.

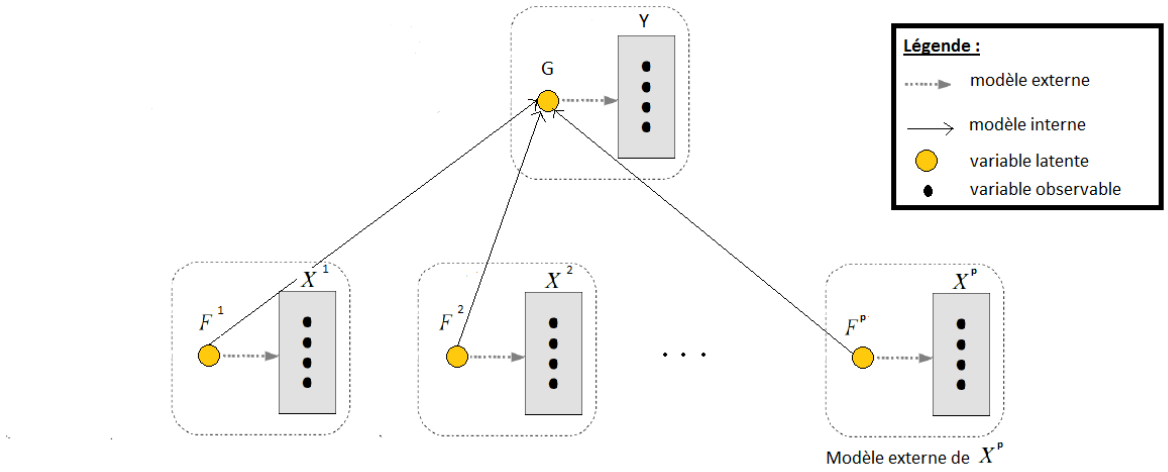

Figure 1: Modèle structurel à P groupes explicatifs et un groupe dépendant

2.2 Formulation du modèle à P groupes explicatifs et un groupe dépendant

Introduisons μ^Y (resp. μ^{X^p}) les vecteurs moyennes, B (resp. A) matrices des coefficients pondérateurs et ε^Y (resp. ε^{X^p}) les matrices des erreurs, associées aux groupes de variable Y (resp. X^p). On notera également ε^G la matrice des erreurs associées à G. Le modèle peut alors être formulé ainsi :

$$\begin{cases} Y = \mathbb{1}_n \mu^{Y\prime} + GB + \varepsilon^Y \\ X^1 = \mathbb{1}_n \mu^{X\prime} + F^1 A^1 + \varepsilon^{X^1} \\ \vdots \\ X^P = \mathbb{1}_n \mu^{X^{P\prime}} + F^P A^P + \varepsilon^{X^P} \\ G = F^1 C^1 + \cdots + F^P C^P + \varepsilon^G \end{cases}$$

On y adjoint, sous contraintes d'identifiabilité, les hypothèses suivantes :
$\forall p \in [\![1, P]\!]$, $F^p \sim \mathcal{N}(0, \mathrm{I})$; $\varepsilon_i^{X^p} \sim \mathcal{N}(0, \psi_{X^p})$, $\varepsilon_i^Y \sim \mathcal{N}(0, \psi_Y)$
où $\psi_{X^p} = \begin{pmatrix} \sigma^2_{p,1} & & 0 \\ & \ddots & \\ 0 & & \sigma^2_{p,q_p} \end{pmatrix}$ de dimension $q_p \times q_p$ et $\psi_Y = \begin{pmatrix} \sigma^2_{Y,1} & & 0 \\ & \ddots & \\ 0 & & \sigma^2_{Y,q_Y} \end{pmatrix}$ de dimension $q_Y \times q_Y$; $\varepsilon_i^G \sim \mathcal{N}(0, I)$; $G \sim \mathcal{N}(0, C^{1\prime}C^1 + ... + C^{P\prime}C^P + I)$ et enfin, ε_i^G, mutuellement indépendants des F^p pour tout observation $i \in [\![1, n]\!]$.

3 Estimation par algorithme EM

L'algorithme EM de Dempster, Laird et Rubin (1977) est une procédure générale pour maximiser la vraisemblance. Il est souvent utilisé dans le cas de problèmes à données manquantes. Dans le cadre des modèles d'équations structurelles à variables latentes, les données manquantes correspondent aux facteurs. Afin d'estimer les paramètres du modèle, cet algorithme procède en deux étapes E (pour "Expectation") et M (pour "Maximization").
Nous le présentons avec une restriction à $P = 2$ groupes explicatifs dans les sections suivantes où nous numéroterons les étapes E et M par 1 et 2.

3.1 L'algorithme EM dans le cadre d'un modèle du type 2 groupes explicatifs et un groupe dépendant

Nous sommes ici en présence de deux groupes explicatifs et un groupe dépendant. On notera en minuscules dans la suite les facteurs communs g (resp. f^1 et f^2) en fonction desquels le modèle exprime les variables Y_1,..., Y_{q_Y} (resp. X_1^1,..., $X_{q_1}^1$ et X_1^2,..., $X_{q_2}^2$). Le facteur g est dépendant des facteurs f^1 et f^2. Soit q_Y (resp. q_1, q_2) le nombre de variables étudiées Y_j (resp.X_j^1, X_j^2) et n le nombre d'observations dont on dispose pour chacune des variables. Pour $j \in [\![1, q_p]\!]$ ou $j \in [\![1, q_Y]\!]$ on écrira μ_j^Y (resp. $\mu_j^{X^1}$, $\mu_j^{X^2}$) les espérances de Y_j (resp. X_j^1, X_j^2), b_j (resp. $a_{,j}^1$, a_j^2) les coefficients pondérateurs associés à Y_j (resp. X_j^1, X_j^2) et ε_{ij}^Y (resp. $\varepsilon_{ij}^{X^1}$, $\varepsilon_{ij}^{X^2}$) les erreurs, associées à la variable Y_j (resp. X_j^1, X_j^2). On notera également ε_i^g les erreurs associées à g. Pour i une observation, le modèle peut être formulé selon le système d'équations suivant :

$$\begin{cases} y_i' = \mu^{Y'} + g_i b' + {\varepsilon_i}^{Y'} & \text{(2a)} \\ x_i^{1'} = \mu^{X^{1'}} + f_i^1 a^{1'} + {\varepsilon_i}^{X^{1'}} & \text{(2b)} \\ x_i^{2'} = \mu^{X^{2'}} + f_i^2 a^{2'} + {\varepsilon_i}^{X^{2'}} & \text{(2c)} \\ g_i = f_i^1 c^1 + f_i^2 c^2 + {\varepsilon_i}^g & \text{(2d)} \end{cases}$$

Où $b' = (b_1, ..., b_{q_Y})$, $a^{1'} = (a_1^1, ..., a_{q_1}^1)$, $a^{2'} = (a_1^2, ..., a_{q_2}^2)$ et c^1, c^2 réels.
Les hypothèses de ce modèle sont du même type que celles du modèle à p groupes explicatifs et un groupe dépendant avec $P = 2$.
Pour $z = (y, x^1, x^2)$ et $h = (g, f^1, f^2)$ la log vraisemblance s'écrit :

$$\begin{aligned} \mathcal{L}(\theta; z, h) = -\frac{1}{2} \sum_{i=1}^{n} \{ & ln|\psi_Y| + ln|\psi_{X^1}| + ln|\psi_{X^2}| + (y_i - \mu^Y - g_i b)' \psi_Y^{-1} (y_i - \mu^Y - g_i b) \\ & + (x_i^1 - \mu^{X^1} - f_i^1 a^1)' \psi_{X^1}^{-1} (x_i^1 - \mu^{X^1} - f_i^1 a^1) + (x_i^2 - \mu^{X^2} - f_i^2 a^2)' \psi_{X^2}^{-1} (x_i^2 - \mu^{X^2} - f_i^2 a^2) \\ & + (g_i - c^1 \ f_i^1 - c^2 \ f_i^2)^2 + (f_i^1)^2 + (f_i^2)^2 \} + cte \end{aligned}$$

où l'ensemble des paramètres du modèle est $\theta = \{\mu^Y, \mu^{X^1}, \mu^{X^2}, b, a^1, a^2, c^1, c^2, \psi_Y, \psi_{X^1}, \psi_{X^2}\}$.
Pour obtenir les estimateurs des paramètres nous résolvons :

$$\mathbb{E}_y^f[\frac{\partial}{\partial\theta} ln[p(y, f; \theta)]] = 0 \tag{3}$$

Pour ce faire, nous utilisons pour chaque i une observation :

$$h_i|z_i \sim \mathcal{N}(m_i = \begin{pmatrix} m_{1i} \\ m_{2i} \\ m_{3i} \end{pmatrix}, \Sigma_i = \begin{pmatrix} \sigma_{11_i} & \sigma_{12_i} & \sigma_{13_i} \\ \sigma_{21_i} & \sigma_{22_i} & \sigma_{23_i} \\ \sigma_{31_i} & \sigma_{32_i} & \sigma_{33_i} \end{pmatrix})$$

Et nous notons :

$$\begin{array}{ll} \widetilde{\gamma}_i = \mathbb{E}_{z_i}^{h_i}[g_i^2] = (\mathbb{E}_{z_i}^{h_i}[g_i])^2 + \mathbb{V}_{z_i}^{h_i}[g_i] = {m_1}_i^2 + \sigma_{11_i}, & \widetilde{g}_i = \mathbb{E}_{z_i}^{h_i}[g_i] = m_{1i}, \\ \widetilde{\phi_i^1} = \mathbb{E}_{z_i}^{h_i}[(f_i^1)^2] = (\mathbb{E}_{z_i}^{h_i}[f_i^1])^2 + \mathbb{V}_{z_i}^{h_i}[f_i^1] = {m_2}_i^2 + \sigma_{22_i}, & \widetilde{f_i^1} = \mathbb{E}_{z_i}^{h_i}[f_i^1] = m_{2i}, \\ \widetilde{\phi_i^2} = \mathbb{E}_{z_i}^{h_i}[(f_i^2)^2] = (\mathbb{E}_{z_i}^{h_i}[f_i^2])^2 + \mathbb{V}_{z_i}^{h_i}[f_i^2] = {m_3}_i^2 + \sigma_{33_i}, & \widetilde{f_i^2} = \mathbb{E}_{z_i}^{h_i}[f_i^2] = m_{3i}. \end{array}$$

3.2 L'algorithme

L'algorithme procède en deux étapes à chaque itération [t].

1. Avec la valeur courante $\theta^{[t]} = \{\mu^{Y^{[t]}}, \mu^{X^{1}{}^{[t]}}, \mu^{X^{2}{}^{[t]}}, b^{[t]}, a^{1^{[t]}}, c^{1^{[t]}}, a^{2^{[t]}}, c^{2^{[t]}}, \psi_Y^{[t]}, \psi_{X^1}^{[t]}, \psi_{X^2}^{[t]}\}$ on calcule pour chaque observation $i \in \{1, ..., n\}$:
$\widetilde{\gamma}_i = {m_1}_i^{[t]^2} + \sigma_{11_i}^{[t]}$; $\widetilde{\phi_i^1} = {m_2}_i^{[t]^2} + \sigma_{22_i}^{[t]}$; $\widetilde{\phi_i^2} = {m_3}_i^{[t]^2} + \sigma_{33_i}^{[t]}$; $\widetilde{g}_i = {m_1}_i^{[t]}$; $\widetilde{f_i^1} = {m_2}_i^{[t]}$; $\widetilde{f_i^2} = {m_3}_i^{[t]}$.

2. On actualise la valeur courante $\theta^{[t+1]}$ en utilisant les formules solutions de (3). Nous ne les présentons pas toutes afin de ne pas alourdir le document mais en donnons quelques exemples :

$$\mu^{X^{1}{}^{[t+1]}} = \overline{x^1} - a^{1^{[t]}}\overline{\widetilde{f^1}}, \quad a^{1^{[t+1]}} = \frac{\overline{\widetilde{f^1}x^1} - \overline{x^1}\,\overline{\widetilde{f^1}}}{\overline{\widetilde{\phi^1}} - \overline{\widetilde{f^1}}^2},$$
$$\sigma_{1,j}^{2}{}^{[t+1]} = \frac{1}{n}\sum_{i=1}^{n}\{(x_i^1 - \widehat{\mu^{X^1}}^{[t]})^2 + (\widehat{a^1}^{[t]})^2\widetilde{\phi_i^1} - 2(x_i^1 - \widehat{\mu^{X^1}}^{[t]})\widehat{a^1}^{[t]}\widetilde{f_i^1}\}$$

La nouvelle valeur :
$\theta^{[t+1]} = \{\mu^{Y^{[t+1]}}, \mu^{X^{1}{}^{[t+1]}}, \mu^{X^{2}{}^{[t+1]}}, b^{[t+1]}, a^{1^{[t+1]}}, c^{1^{[t+1]}}, a^{2^{[t+1]}}, c^{2^{[t+1]}}, \psi_Y^{[t+1]}, \psi_{X^1}^{[t+1]}, \psi_{X^2}^{[t+1]}\}$ va permettre de mettre à jour $\widetilde{\gamma}_i$, $\widetilde{\phi_i^1}$, $\widetilde{\phi_i^2}$, $\widetilde{g}_i$, $\widetilde{f_i^1}$ et $\widetilde{f_i^2}$ dans l'étape 1. On repasse alors à l'étape 2, et ainsi de suite jusqu'à convergence de l'algorithme.

4 Application

Diverses simulations seront présentées ainsi qu'une application sur des données réelles de notre méthode d'estimation. Application qui nous permettra de proposer un modèle explicatif.

Bibliographie

[1] Bry, X., Redont, P., Verron, T. (2012), *THEME-SEER : a multidimensional exploratory technique to analyze a structural model using an extended covariance criterion*, *Journal of chemometrics*, Montpellier.
[2] Esposito Vinzi, V. et Trinchera, L. (2008), *Modèles à équations structurelles, approches basée sur les composantes*, URL : `http://www.academia.edu/390381/Modeles_a_equations_structurelles_approches_basees_sur_les_composantes`, Naples.
[3] Foulley, J-L. (2002), *Algorithme EM : Theorie et application au modèle mixte*, *Journal de la Société Française de Statistique*, Jouyen-Josas.
[4] Fox, J. (2002), *Structural Equation Models*, `http://cran.r-project.org/doc/contrib/Fox-Companion/appendix-sems.pdf`.
[5] Hwang, H. et Takane, Y. (2004), *Generalized structured component analysis*, *Psychometrika*, 81–99.
[6] Jakobowicz, E. (2007), *Contributions aux modèles d'équations structurelles à variables latentes*, *Thèse*, Paris, 81–99.
[7] Jöreskog, K. (1970), *A general method for analysis of covariance structure*, *Biometrika*.
[8] Rivera, P. et Satorra, A. (2002), *Latent Variable and Latent Structure Models*, *Marcoulides, G. et Moustaki, I.*, New Jersey, 85–102.
[9] Saidane, M. (2006), *Modèles à facteurs conditionnellement hétéroscédastiques et à structure markoviene cachée pour les séries financières, Thèse*, Montpellier.
[10] Stan, V. et Saporta, G. (2006), *Une comparaison expérimentale entre les approches PLS et LISREL*, Paris.
[11] Tenenhaus, A. et Tenenhaus, M. (2011), *Regularized generalized canonical correlation analysis*, *Psychometrika*, Gif-sur-Yvette et Jouy-en-Josas.
[13] Wold, H. (1975), *Modelling in complex situations with soft information, in hird World Congress of Econometric Societ*, Canada.
[14] Wold, H. (1985), *Partial Least Squares, Encyclopedia of Statistical Sciences*, Vol. 6, Wiley, New York, 581–591.