Estimation par algorithme EM pour un modèle à facteurs et à équation structurelle

Tami Myriam, Xavier Bry, Christian Lavergne
IMAG, Universités de Montpellier
Place Eugène Bataillon, Montpellier, France
myriam.tami@umontpellier.fr, xavier.bry@umontpellier.fr,
christian.lavergne@univ-montp3.fr

Résumé - *Nous proposons une nouvelle méthode d'estimation d'un modèle à facteurs et à équation structurelle. Notre méthode est fondée sur la maximisation de la fonction de vraisemblance par algorithme EM. Nous montrerons que cette approche permet les estimations des facteurs en plus de celles des coefficients du modèle. À travers une étude sur données simulées et d'une analyse de sensibilité nous avons évalué les performances de cette méthode. Nous présenterons une application de cette approche sur des données réelles environnementales.*

Mots clés - Algorithme EM, modèles à équations structurelles, Modèles à facteurs, variables latentes, estimation.

1 Introduction

Nous nous plaçons dans le contexte des modèles à équations structurelles (*Structural Equation Models, SEM*). On suppose avoir observé plusieurs groupes de variables Y, X^1,..., X^p sur les mêmes unités. De chacun de ces groupes est extrait une variable non observable dite latente. Ces variables latentes sont alors liées par des relations de causalité. Le modèle peut aussi être enrichi de variables explicatives également observées. Dans la littérature, deux familles de méthodes existent. L'une, connue sous le nom de PLS (*Partial Least Squares*) et développée par Herman Wold (1975) et ses successeurs, repose sur l'usage de composantes comme variables latentes. La seconde, issue des travaux de Karl Jöreskog (1970), est fondée sur le maximum de vraisemblance et fait appel à des facteurs comme variables latentes. La méthode engendrée porte le nom de LISREL (*LInear Structural RELations*). Ces deux approches ont été comparées dans plusieurs travaux [SS06]. Dans un objectif d'interprétation a priori du modèle, ces derniers ont montré qu'il est préférable d'utiliser un modèle à facteurs plutôt qu'un modèle à composantes, lequel est computationnellement plus efficace mais trop contraint. Nous nous plaçons dans le paradigme de l'estimation par maximum de vraisemblance et nous proposons un algorithme EM ; algorithme bien adapté dans le cas de variables latentes. Contrairement aux méthodes citées plus haut, cette approche a l'avantage d'estimer les facteurs en plus des paramètres du modèle, tout en restant efficace en terme de temps de calcul. Nous présenterons une application de cette méthode sur des données réelles.

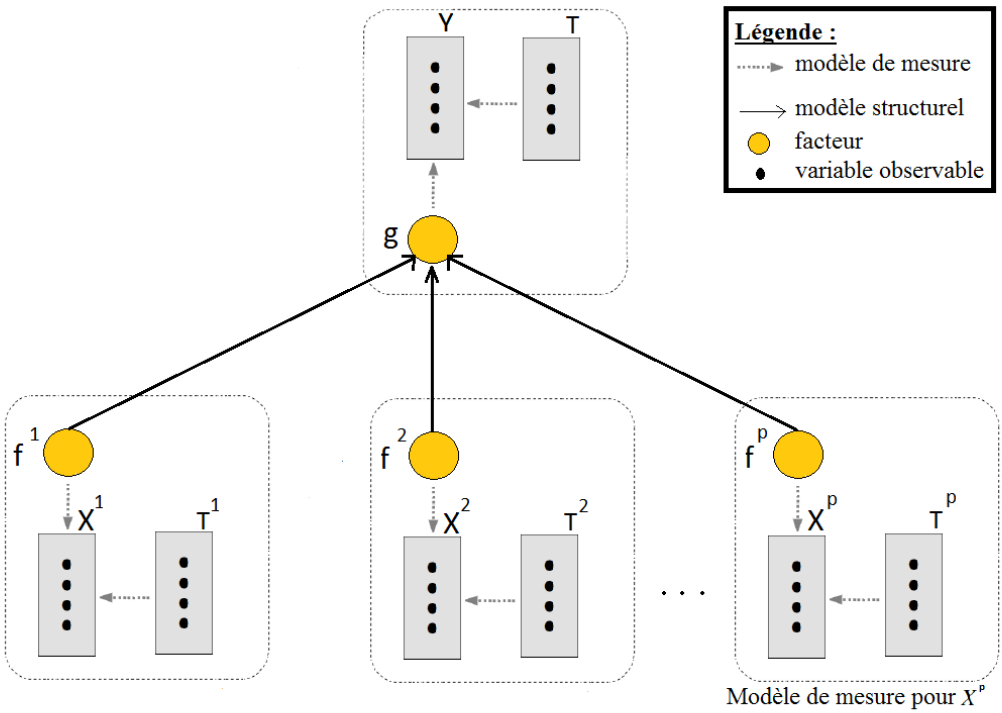

FIGURE 1 – Modèle structurel à p groupes explicatifs et un groupe dépendant

2 Structure générale du modèle

2.1 Notations du 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 $p \in [\![1, m]\!]$ 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 (dit aussi structurel), la variable latente g est dépendante de l'ensemble des variables latentes f^1, ..., f^p (cf. figure 1). Quant aux modèles externes (dit aussi de mesure), chacun forme un groupe de variables $X^m = (X_1^m, ...X_{q_m}^m)$ (resp. $Y = (Y_1, ...Y_{q_Y})$) comme dépendant du facteur f^m (resp. g). Dépendances pouvant être chacune enrichie par une dépendance supplémentaire aux covariables T^m (resp. T).
Notons que ce modèle peut se généraliser à plusieurs facteurs dans chaque groupe sans difficultés.

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

Introduisons D (resp. D^m) une matrice $r_T \times q_Y$ (resp. $r_m \times q_m$) de coefficients pondérateurs, b (resp. a^m) un vecteur $1 \times q_Y$ (resp. $1 \times q_m$) de coefficients pondérateurs et ε^Y (resp. ε^m) une matrice des erreurs $n \times q_Y$ (resp. $n \times q_m$), associées au groupe de variable Y (resp. X^m). On notera également ε^g la matrice des erreurs associées à g. Le modèle peut alors être formulé

ainsi :

$$\left\{\begin{array}{ccc} Y & = & TD + gb' + \varepsilon^Y \\ \forall m \in [\![1,p]\!], X^m & = & T^m D^m + f^m a^{m\prime} + \varepsilon^m \\ g & = & f^1 c^1 + \cdots + f^p c^p + \varepsilon^g \end{array}\right. \quad (1)$$

Nous imposons que les éléments de la première colonne des matrices de covariables T et T^m soient fixés à 1. Ainsi, la première ligne de D et de chaque matrice D^m correspondront aux paramètres de moyenne. On y adjoint, sous contraintes d'identifiabilité, les hypothèses suivantes :
$\forall m \in [\![1,p]\!]$, $f^m \sim \mathcal{N}(0,1)$; $\varepsilon_i^m \sim \mathcal{N}(0,\, \psi_m)$, $\varepsilon_i^Y \sim \mathcal{N}(0,\, \psi_Y)$
où $\psi_m = diag(\sigma^2_{m,j})_{j\in[\![1,q_m]\!]}$ de dimension $q_m \times q_m$ et $\psi_Y = diag(\sigma^2_{Y,j})_{j\in[\![1,q_Y]\!]}$ de dimension $q_Y \times q_Y$; $\varepsilon_i^g \sim \mathcal{N}(0,\, 1)$; $g \sim \mathcal{N}(0,\, (c^1)^2 + ... + (c^p)^2 + 1)$ et enfin, ε_i^g, mutuellement indépendants des f^m pour tout observation $i \in [\![1,n]\!]$ et $\forall m \in [\![1,p]\!]$ ε^Y et ε^m sont indépendant. .

3 Estimation par algorithme EM

L'algorithme EM de Dempster, Laird et Rubin (1977) [DLR77] 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.

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. 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 $j \in [\![1, q_m]\!]$ ou $j \in [\![1, q_Y]\!]$ on écrira $D_{,j}$ et b_j (resp. $D^1_{,j}$, $D^2_{,j}$, $a^1_{,j}$, a^2_j) les coefficients pondérateurs associés à Y_j (resp. X_j^1, X_j^2) et ε^Y_{ij} (resp. ε^1_{ij}, ε^2_{ij}) 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 :

$$\left\{\begin{array}{ccc} y_i' & = & t_i{}'D + g_i b' + \varepsilon_i{}^{y\prime} \\ x_i^{1\prime} & = & t_i^{1\prime} D^1 + f_i^1 a^{1\prime} + \varepsilon_i{}^{1\prime} \\ x_i^{2\prime} & = & t_i^{2\prime} D^2 + f_i^2 a^{2\prime} + \varepsilon_i{}^{2\prime} \\ g_i & = & f_i^1 c^1 + f_i^2 c^2 + \varepsilon_i{}^g \end{array}\right. \quad (2)$$

Où nous faisons les simplifications suivantes $\psi_Y = \sigma_Y^2$, $\psi_1 = \sigma_1^2$,$\psi_2 = \sigma_2^2$. On note
$\theta = \{D, D^1, D^2, b, a^1, a^2, c^1, c^2, \sigma_Y^2, \sigma_1^2, \sigma_2^2\}$ l'ensemble des paramètres du modèle. Ainsi, la dimension de θ est :

$$K = 5 + q_Y(r_T + 1) + \sum_{m=1}^{2} q_m(r_m + 1)$$

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 :

$$\begin{aligned}\mathcal{L}(\theta; z, h) = -\frac{1}{2}\sum_{i=1}^{n}\{&ln|\psi_Y| + ln|\psi_1| + ln|\psi_2| \\ &+ (y_i - D't_i - g_i b)'\psi_Y^{-1}(y_i - D't_i - g_i b) \\ &+ (x_i^1 - D^{1'}t_i^1 - f_i^1 a^1)'\psi_1^{-1}(x_i^1 - D^{1'}t_i^1 - f_i^1 a^1) \\ &+ (x_i^2 - D^{2'}t_i^2 - f_i^2 a^2)'\psi_2^{-1}(x_i^2 - D^{2'}t_i^2 - f_i^2 a^2) \\ &+ (g_i - c^1\ f_i^1 - c^2\ f_i^2)^2 + (f_i^1)^2 + (f_i^2)^2\} + \lambda\end{aligned}$$

Où θ est l'ensemble des paramètres du modèle de dimension K et λ une constante. Cependant, à cause de la simplification du modèle (1), dans notre cas (2), $\theta = \{D, D^1, D^2, b, a^1, a^2, c^1, c^2, \sigma_Y^2, \sigma_1^2, \sigma_2^2\}$.
En effet, $\psi_Y = \sigma_Y^2 Id_{q_Y}$, $\psi_1 = \sigma_1^2 Id_{q_1}$ et $\psi_2 = \sigma_2^2 Id_{q_2}$.
Pour maximiser la fonction log-vraisemblance par algorithme EM nous résolvons :

$$\mathbb{E}_z^h[\frac{\partial}{\partial\theta}\mathcal{L}(\theta; z, h)] = 0 \tag{3}$$

(cf. [Fou02]).
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 :

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

Nous obtenons des formules solutions explicites de (3) que nous ne présentons pas ici pour ne pas alourdir le document. Cependant vous pouvez y avoir accès sur l'article en preprint [TBL15].

3.2 L'algorithme

Pour estimer les paramètres de θ et les facteurs g, f^m, nous proposons l'algorithme qui suit où nous notons $[t]$ la t^{ieth}-itération de l'algorithme.

1. Initialisation = choix des valeurs initiales des paramètres $\theta^{[0]}$.

2. À l'itération courante $t \geq 1$, jusque satisfaction du critère d'arrêt on procède comme suit :
   (a) **E-step** : Avec $\theta^{[t-1]}$,
      i. On calcule explicitement la distribution $h_i|z_i$ pour tout $i \in [\![1, n]\!]$.
      ii. On estime les valeurs des facteurs $\widetilde{g}^{[t]}$, $\widetilde{f^m}^{[t]}$, $m \in \{1, 2\}$.

iii. On calcule $\widetilde{\gamma}^{[t]}$ and $\widetilde{\phi^m}^{[t]}$, $m \in \{1, 2\}$.

(b) **M-step** :

i. On actualise θ à $\theta^{[t]}$ en introduisant $\widetilde{g}^{[t]}$, $\widetilde{\gamma}^{[t]}$ et $\widetilde{f^m}^{[t]}$, $\widetilde{\phi^m}^{[t]}$, $m \in \{1, 2\}$ dans les formules solutions de (3) (cf. [TBL15]).

3. Nous utilisons le critère d'arrêt suivant avec ϵ le plus petit possible :

$$\sum_{k=1}^{K} \frac{|\theta^{*[t+1]}[k] - \theta^{*[t]}[k]|}{\theta^{*[t+1]}[k]} < \epsilon$$

où θ^* est le vecteur de dimension K contenant tous les paramètres scalaires de l'ensemble des paramètres de θ.

4 Performances de l'approche et application

Suite à une analyse de sensibilité, nous montrerons les performances de la méthodes. 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.

Références

[Bac87] F. Bacher. LES MODÈLES STRUCTURAUX EN PSYCHOLOGIE PRÉSENTATION D'UN MODÈLE : LISREL Première partie. *Le Travail Humain*, 50(4) :347–370, January 1987.

[Bac88] F. Bacher. LES MODÈLES STRUCTURAUX EN PSYCHOLOGIE PRÉSENTATION D'UN MODÈLE : LISREL. *Le Travail Humain*, 51(4) :273–288, January 1988.

[Bol14] Kenneth A. Bollen. *Structural Equations with Latent Variables*. John Wiley & Sons, August 2014.

[BRVC12] Xavier Bry, Patrick Redont, Thomas Verron, and Pierre Cazes. THEME-SEER : a multidimensional exploratory technique to analyze a structural model using an extended covariance criterion. *J. Chemometrics*, 26(5) :158–169, May 2012.

[BTVM13] X. Bry, C. Trottier, T. Verron, and F. Mortier. Supervised component generalized linear regression using a PLS-extension of the Fisher scoring algorithm. *Journal of Multivariate Analysis*, 119 :47–60, August 2013.

[BV15] Xavier Bry and Thomas Verron. THEME : Thematic model exploration through multiple costructure maximization. *J. Chemometrics. Accepted but unpublished*, August 2015.

[DLR77] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1) :1–38, January 1977.

[EVT14] V Esposito Vinzi and L Trinchera. Modèles à équations structurelles, approches basées sur les composantes, 2014.

[Fou02] Jean-Louis Foulley. Algorithme EM : Théorie et application au modèle mixte. *Journal de la Société française de statistique*, 143(3-4) :57–109, 2002.

[HT04] Heungsun Hwang and Yoshio Takane. Generalized structured component analysis. *Psychometrika*, 69(1) :81–99, March 2004.

[Jö70] K. G. Jöreskog. A general method for analysis of covariance structures. *Biometrika*, 57(2) :239–251, January 1970.

[Jak07] Emmanuel Jakobowicz. *Contributions aux modèles d'équations structurelles à variables latentes*. phdthesis, Conservatoire national des arts et metiers - CNAM, October 2007.

[Jor03] Michael I. Jordan. *An introduction to probabilistic graphical models*. preparation, 2003.

[JS82] Karl G. Jöreskog and Dag Sörbom. Recent Developments in Structural Equation Modeling. *Journal of Marketing Research*, 19(4) :404–416, November 1982.

[Loh13] Jan-Bernd Lohmöller. *Latent Variable Path Modeling with Partial Least Squares*. Springer Science & Business Media, November 2013.

[MM14] George A. Marcoulides and Irini Moustaki. *Latent Variable and Latent Structure Models*. Psychology Press, April 2014.

[Sai06] Mohamed Saidane. *Modèles à Facteurs Conditionnellement Hétéroscédastiques et à Structure Markovienne Cachée pour les Séries Financières*. phdthesis, Université Montpellier II - Sciences et Techniques du Languedoc, July 2006.

[SS06] Valentina Stan and Gilbert Saporta. Une comparaison expérimentale entre les approches PLS et LISREL. In *38 èmes Journées de Statistique, Clamart, France*, X, France, January 2006.

[TBL14] Myriam Tami, Xavier Bry, and Christian Lavergne. Estimation of structural equation models with factors by EM algorithm. In *JDS 2014 Rennes*, Rennes, France, June 2014.

[TBL15] Myriam Tami, Xavier Bry, and Christian Lavergne. Em estimation of a structural equation model (preprint). Montpellier, France, September 2015.

[TT11] Arthur Tenenhaus and Michel Tenenhaus. Regularized Generalized Canonical Correlation Analysis. *Psychometrika*, 76(2) :257–284, March 2011.

[TVCL05] Michel Tenenhaus, Vincenzo Esposito Vinzi, Yves-Marie Chatelin, and Carlo Lauro. PLS path modeling. *Computational Statistics & Data Analysis*, 48(1) :159–205, January 2005.

[VCHW10] Vincenzo Esposito Vinzi, Wynne W. Chin, Jörg Henseler, and Huiwen Wang. *Handbook of Partial Least Squares : Concepts, Methods and Applications*. Springer Science & Business Media, March 2010.

[WK89] L. E. Wangen and B. R. Kowalski. A multiblock partial least squares algorithm for investigating complex chemical systems. *J. Chemometrics*, 3(1) :3–20, January 1989.