Approche bayésienne des modèles à équations structurelles

Séverine Demeyer [1,2] & Nicolas Fischer [1] & Gilbert Saporta [2]

[1] *LNE, Laboratoire National de Métrologie et d'Essais*
29 avenue Roger Hennequin, 78197 Trappes, France, severine.demeyer@lne.fr
[2] *Chaire de statistique appliquée & CEDRIC, CNAM*
292 rue Saint Martin, Paris, France

Résumé

Les modèles à équations structurelles (SEMs) sont des modèles multivariés à variables latentes utilisés pour modéliser les structures de causalité dans les données. Une approche bayésienne d'estimation et de validation des modèles SEMs est proposée et l'identifiabilité des paramètres est étudiée. Cette étude montre qu'une étape de réduction des variables latentes au sein de l'algorithme de Gibbs permet de garantir l'identifiabilité des paramètres. Cette heuristique permet en fait d'introduire les contraintes d'identifiabilité dans l'analyse. Pour illustrer ce point, les contraintes d'identifiabilité sont calculées dans une application en marketing, dans laquelle les distributions des contraintes sont obtenues par combinaisons des tirages *a posteriori* des paramètres.

Abstract

Structural equation models (SEMs) are multivariate latent variable models used to model causality structures in data. A Bayesian estimation and validation of SEMs is proposed and identifiability of parameters is studied. The latter study shows that latent variables should be standardized in the analysis to ensure identifiability. This heuristics is in fact introduced to deal with complex identifiability constraints. To illustrate the point, identifiability constraints are calculated in a marketing application, in which posterior draws of the constraints are derived from the posterior conditional distributions of parameters.

Mots clés : modèles à équations structurelles, variables latentes, algorithme de Gibbs, identifiabilité, méthodes bayésiennes, marketing

1 Modèles à équations structurelles

1.1 Contexte

Les variables observées (manifestes) sont groupées puis associées aux variables latentes dans le modèle externe (modèle de mesure) et les relations de causalité entre les variables

latentes sont représentées dans le modèle interne (structurel) Cette situation est typique des études de satisfaction en marketing, comme illustré dans l'application, où les variables observées sont des questions regroupées selon des thématiques et les variables latentes sont ces thématiques, à savoir la satisfaction, la fidélité et l'image (voir figure 3).

1.2 Modèle

Le vecteur ligne Y_i des valeurs observées pour l'individu i sur les p variables manifestes est exprimé en fonction du vecteur ligne Z_i de ses scores sur les q variables latentes, par le modèle de régression suivant (appelé le modèle de mesure) :

$$Y_i = Z_i\theta + \varepsilon_i,\ 1 \leq i \leq n \tag{1}$$

avec $\varepsilon_i \sim \mathcal{N}(0, \Sigma_\varepsilon)$ et où θ est la matrice $p \times q$ des coefficients de régression.

Si Z_i était connu, le modèle de mesure (1) serait un modèle de régression linéaire classique. En notant H_i les variables latentes endogènes et Ξ_i les variables latentes exogènes, les équations structurelles sont données de manière équivalente par les trois expressions :

$$\begin{aligned} H_i &= H_i\Pi + \Xi_i\Gamma + \delta_i \\ H_i &= Z_i\Lambda + \delta_i \qquad \Lambda^t = \left(\Pi^t\,\Gamma^t\right) \\ \Pi_0^t H_i &= \Gamma^t\Xi_i + \delta_i \qquad \Pi_0 = Id - \Pi \end{aligned} \tag{2}$$

où Π est la matrice $q_1 \times q_1$ des coefficients de régression entre les variables latentes endogènes , Γ est la matrice $q_2 \times q_1$ des coefficients de régression entre les variables latentes endogènes et exogènes, δ_i est indépendant de Ξ_i, $\delta_i \sim \mathcal{N}(0, \Sigma_\delta)$, et $\Xi_i \sim \mathcal{N}(0, \Phi)$.

1.3 Etude de l'identifiabilité des paramètres

Garantir l'identifiabilité des modèles à équations structurelles est équivalent à vérifier l'injectivité de la fonction de vraisemblance intégrée sur les variables latentes. En notant $\Theta = \{\theta, \Sigma_\varepsilon, \Pi_0, \Gamma, \Sigma_\delta, \Phi\}$ l'ensemble des paramètres du modèles, l'identifiabilité s'écrit :

$$\forall Y_i,\ [Y_i|\Theta] = \left[Y_i|\tilde{\Theta}\right] \Longrightarrow \Theta = \tilde{\Theta} \tag{3}$$

où Y_i est marginalement distribué $\mathcal{N}(\mathbf{0}, \Sigma_Y)$ et $\Sigma_Y = \theta^t\Sigma_Z\theta + \Sigma_\varepsilon$ où Σ_Z est la matrice de covariance des variables latentes.

En notant de plus $\theta_k = (\theta_{k1} \dots \theta_{kn_k})$, le vecteur des coefficients de régression du bloc k et $\Sigma_Z = \{\rho_{ij},\ 1 \leq i, j \leq K\}$, Σ_Y est la matrice bloc :

$$\Sigma_Y = \begin{pmatrix} \rho_{11}\theta_1\theta_1^t + \Sigma_{\varepsilon 1} & \rho_{12}\theta_1\theta_2^t & \dots & \rho_{1K}\theta_1\theta_K^t \\ \rho_{12}\theta_1\theta_2^t & \rho_{22}\theta_2\theta_2^t + \Sigma_{\varepsilon 2} & \dots & \vdots \\ \vdots & & & \\ \rho_{1K}\theta_1\theta_K^t & \dots & \dots & \rho_{KK}\theta_K\theta_K^t + \Sigma_{\varepsilon K} \end{pmatrix} \tag{4}$$

La définition de l'identifiabilité (3) appliquée à une vraisemblance gaussienne donne $\Sigma_Y = \tilde{\Sigma_Y}$. Les équations d'identifiabilité qui découlent de cette égalité sont :

$$\rho_{kk}\theta_{ki}^2 + \sigma_{ki}^2 = \tilde{\rho}_{kk}\tilde{\theta}_{ki}^2\tilde{\sigma}_{ki}^2,\ i = 1 \dots n_k,\ k = 1 \dots K \tag{5}$$

$$\rho_{kk}\theta_{ki}\theta_{kj} = \tilde{\rho}_{kk}\tilde{\theta}_{ki}\tilde{\theta}_{kj},\ 1 \leq i < j \leq n_k,\ k = 1 \dots K \tag{6}$$

$$\rho_{kk'}\theta_{ki}\theta_{k'j} = \tilde{\rho}_{kk'}\tilde{\theta}_{ki}\tilde{\theta}_{k'j},\ 1 \leq i \leq n_k,\ 1 \leq j \leq n_{k'},\ k = 1 \dots K \tag{7}$$

Les équations (5) et (6) sont obtenues en égalant les éléments bloc diagonaux de Σ_Y et $\tilde{\Sigma_Y}$ et l'équation (7) vient de l'égalisation des éléments hors des blocs diagonaux.

Si $\theta_{k1} = \tilde{\theta}_{k1}$ et $\rho_{kk} = \rho_{kk'}$ pour un k fixé, alors l'équation (6) donne $\theta_{kj} = \tilde{\theta}_{kj}$ pour tous les j. Reporter ce résultat dans l'équation 5 donne $\sigma_{ki}^2 = \tilde{\sigma}_{ki}^2$ pour tout k, i. Reporter ce dernier résultat dans l'équation (7) donne $\rho_{kk'} = \tilde{\rho}_{kk'}$ pour tout k, k'. En conséquence, $\theta_{k1} = 1$ and $\rho_{kk} = 1$ pour tout k constitue un ensemble suffisant de conditions d'identifiabilité.

La contrainte $\rho_{kk} = 1$ s'exprime en fait en fonction des paramètres intérieurs, obtenue en égalant à 1 les éléments diagonaux de Σ_Z. Cependant, la simulation *a posteriori* des paramètres conditionnellement à ces contraintes est compliquée. L'heuristique, qui consiste à réduire les variables latentes après qu'elles ont été tirées dans leur distribution conditionnelle *a posteriori*, permet de contourner cette difficulté (voir l'application).

2 Estimation bayésienne des modèles SEM

Dans ce modèle à variables latentes, les techniques d'augmentation des données et d'imputation, voir Tanner and Wong (1987), sont implémentées dans un algorithme de Gibbs sous des hypothèses de normalité et de conjugaison. On se reportera à Box and Tiao (1973) pour les calculs dans les modèles multivariés gaussiens et à Gelman et al. (2004) pour des détails sur l'échantillonneur de Gibbs. On reporte ci-dessous les expressions finales des distributions *a posteriori* invoquées par l'algorithme de Gibbs (voir figure 1) :

$$Z_i|Y_i, \theta, \Sigma_\varepsilon, \Lambda, \Sigma_\delta, \Phi \sim \mathcal{N}\left(D\theta\Sigma_\varepsilon^{-1}Y_i, D\right),\ D^{-1} = \theta\Sigma_\varepsilon^{-1}\theta^t + \Sigma_Z^{-1}$$

$$\theta_{kj}|Y, Z, \Sigma_{\varepsilon kj} \sim \mathcal{N}\left(D_{kj}A_{kj}, \Sigma_{\varepsilon kj}D_{kj}\right),\ D_{kj} = \left(Z_k^tZ_k + \Sigma_{\varepsilon 0k}^{-1}\right)^{-1},\ A_{kj} = \Sigma_{\varepsilon 0k}^{-1}\theta_{0k} + Z_k^tY_{kj}$$

$$\Sigma_{\varepsilon k1}^{-1} \sim \mathcal{G}\left(\frac{n}{2} + \alpha_{0\varepsilon kj}, \beta_{0\varepsilon kj} + \frac{1}{2}\left(Y_{kj} - Z_k\right)^t\left(Y_{kj} - Z_k\right)\right)$$

$$\Sigma_{\varepsilon kj}^{-1} \sim \mathcal{G}\left(\frac{n}{2} + \alpha_{0\varepsilon kj}, \beta_{0\varepsilon kj} + \frac{1}{2}\left[Y_{kj}^tY_{kj} - \left(D_{kj}A_{kj}\right)^tD_{kj}^{-1}D_{kj}A_{kj} + \frac{\theta_{0k}^2}{\Sigma_{\varepsilon 0k}}\right]\right)$$

Initialisation : $\theta^0, \Sigma_\varepsilon^0, \Lambda^0, \Sigma_\delta^0, \Phi^0$. A l'itération t :

1. $Z^t \sim Z|Y, \theta^{t-1}, \Sigma_\varepsilon^{t-1}, \Lambda^{t-1}, \Sigma_\delta^{t-1}, \Phi^{t-1}$
2. **réduction des variables latentes**: soit Z^{*t} la VL réduite
3. $\Sigma_\varepsilon^t \sim \Sigma_\varepsilon|Y, Z^{*t}, \theta^{t-1}, \Lambda^{t-1}, \Sigma_\delta^{t-1}, \Phi^{t-1}$
4. $\theta^t \sim \theta|Y, Z^{*t}, \Sigma_\varepsilon^t, \Lambda^{t-1}, \Sigma_\delta^{t-1}, \Phi^{t-1}$
5. $\Sigma_\delta^t \sim \Sigma_\delta^|Y, Z^{*t}, \Lambda^{t-1}, \theta^t, \Sigma_\varepsilon^t, \Phi^{t-1}$
6. $\Lambda^t \sim \Lambda|Y, Z^{*t}, \Sigma_\delta^t, \theta^t, \Sigma_\varepsilon^t, \Phi^{t-1}$
7. $\Phi^t \sim \Phi|Y, Z^{*t}, \theta^t, \Sigma_\varepsilon^t, \Lambda^t, \Sigma_\delta^t$

Figure 1: Etapes de l'algorithme de Gibbs

$$\Lambda_k|Y, Z, \Sigma_{\delta k} \sim \mathcal{N}\left(\tilde{D}_k\tilde{A}_k, \Sigma_{\varepsilon k}\tilde{D}_k\right),\ \tilde{D}_k = \left(Z_k^t Z_k + \Sigma_{\delta k}^{-1}\right)^{-1},\ \tilde{A}_k = \Sigma_{\delta k}^{-1}\Lambda_{0k} + Z^t H_k$$

$$\Sigma_{\delta k}^{-1} \sim \mathcal{G}\left(\frac{n}{2} + \alpha_{0\delta k}, \beta_{0\delta k} + \frac{1}{2}\left[Y_k^t Y_k - \left(\tilde{D}_k\tilde{A}_k\right)^t \tilde{D}_k{}^{-1}\tilde{D}_k\tilde{A}_k + \Lambda_{0k}^t\Sigma_{\delta k}^{-1}\Lambda_{0k}\right]\right)$$

$$\Phi|Z \sim InvWishart\left(\Xi^t\Xi + R_0^{-1}, n + d_0\right)$$

où les paramètres indicés par 0 sont les paramètres à priori des distributions conjuguées correspondantes.

L'algorithme de Gibbs (voir figure 1) alterne le tirage dans les distributions conditionnelles *a posteriori* des paramètres sachant les données (Etape 1) et le tirage dans les distributions conditionnelles *a posteriori* des variables latentes sachant les paramètres (Etapes 3 à 7). L'étape 2 est l'heuristique qui permet d'assurer l'identifiabilité des paramètres du modèle en réduisant les variables latentes (VL).

La validation du modèle peut-être réalisée en calculant les "Posterior Predictive p-values" (voir Gelman and al. (1996)). Les PP p-values sont calculées à partir des distribution prédictive *a posteriori* intégrées sur les paramètres et les variables latentes. Le modèle n'est pas rejeté si la PP p-value est proche de 0.5.

3 Application

Les relations entre la fidélité, la satisfaction et l'image sont étudiées dans le cadre des modèles ECSI sur un sous-ensemble de $n = 202$ individus sans données manquantes provenant du jeu de données de démonstration du logiciel XLStat. Les variables catégorielles sont centrées, réduites et traitées comme continues. L'algorithme est implémenté sous R.

On note θ_0 et λ_0 les valeurs *a priori* des paramètres. Les valeurs *a priori* choisies reflètent la confiance dans les liens de causalité : $\theta_0 = 0.5$, $\Lambda_0 = 0.5$, $\Sigma_{\varepsilon 0} = 1$, $\Sigma_{\delta 0} = 1$

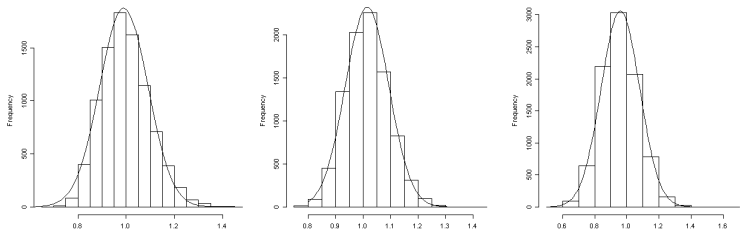

Figure 2: Distributions *a posteriori* des contraintes b) de gauche à droite

	θ_{12}	θ_{22}	θ_{23}	θ_{32}	θ_{33}	θ_{34}	θ_{35}	π_{12}	λ_1	λ_2
moyenne	0.774	0.705	0.784	0.605	0.457	0.732	0.658	0.475	0.307	0.796
écart-type	0.060	0.051	0.053	0.063	0.067	0.059	0.059	0.127	0.130	0.047

Table 1: Coefficients de régression: moyenne et écart-type *a posteriori*

and $\Phi_0 = 1$. On observe une convergence rapide de l'algorithme de Gibbs pour tous les paramètres et une faible autocorrélation dans les échantillons *a posteriori*. En explicitant $\Pi_0 = \begin{pmatrix} 1 & 1 \\ -\pi_{12} & 1 \end{pmatrix}$ et $\Gamma = (\lambda_1 \lambda_2)$ on obtient l'expression suivante de Σ_Z :

$$\Sigma_Z = \begin{pmatrix} \lambda_1^2 + \Sigma_{\delta 1} + \pi_{12}\lambda_1\lambda_2 + \pi_{12}^2\left(\lambda_2^2 + \Sigma_{\delta 2}\right) & \lambda_1\lambda_2 + \pi_{12}^2\left(\lambda_2^2 + \Sigma_{\delta 2}\right) & \Phi\left(\lambda_1 + \lambda_1\lambda_2\right) \\ \lambda_1\lambda_2 + \pi_{12}^2\left(\lambda_2^2 + \Sigma_{\delta 2}\right) & \lambda_2^2 + \Sigma_{\delta 2} & \Phi\lambda_2 \\ \Phi\left(\lambda_1 + \lambda_1\lambda_2\right) & \Phi\lambda_2 & \Phi \end{pmatrix}$$

D'après la section 1.3, les contraintes d'identifiabilité sont (où les contraintes b) sont obtenues en égalant à 1 les éléments diagonaux de Σ_Z) :

$$a)\theta_{11} = 1,\ \theta_{21} = 1, \theta_{31} = 1$$
$$b)\Phi = 1,\ \lambda_2^2 + \Sigma_{\delta 2} = 1,\ \lambda_1^2 + \Sigma_{\delta 1} + \pi_{12}\lambda_1\lambda_2 + \pi_{12}^2\left(\lambda_2^2 + \Sigma_{\delta 2}\right) = 1$$

Des échantillons *a posteriori* de ces contraintes sont calculés à partir des échantillons *a posteriori* des paramètres. Ces distributions sont centrées en 1 avec une faible dispersion comme le montre la figure 2, ce qui est en faveur de l'heuristique.

Les valeurs d'intérêt pour cette application sont les corrélations entre les variables manifestes et les variables latentes et entre les variables latentes. Dans la table 1, θ_{12}, θ_{22}, θ_{23}, θ_{32}, θ_{33},θ_{34}, θ_{35} et λ_2 sont des coefficients de corrélation alors que π_{12} et λ_1 sont les coefficients d'une régression multiple. *Satisfaction* et *Image* sont fortement corrélées (0.796, voir table 1), ce qui signifie que l'*Image* a une influence importante sur la *Satisfaction*. Toutes les corrélations sont représentées dans le graphique de la figure 3.

La PP p-value de $0.37 < 0.5$ vient de ce que les données, catégorielles, ne suivent pas des lois gaussiennes. Il demeure que cet exemple illustre des caractéristiques intéressantes

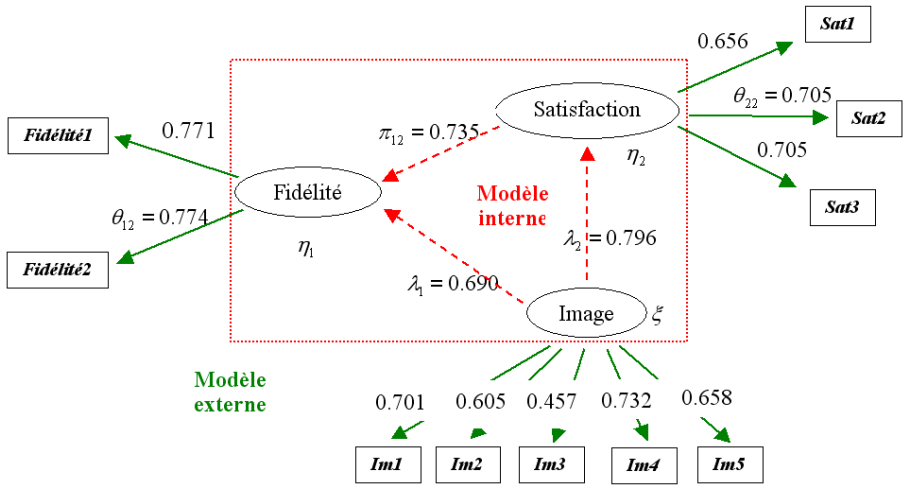

Figure 3: Graphique des corrélations

des approches bayésiennes, comme les tests d'hypothèses avec les PPp-values et la possibilité d'observer la variabilité des paramètres et la variabilité de fonctions des paramètres.

4 Conclusion et perspectives

Les distributions *a posteriori* des paramètres des modèles SEMs sont calculées sous les hypothèses de normalité et de conjugaison. Ces distributions permettent d'étudier différents aspects du modèle comme la variabilité des paramètres et de fonctions de paramètres et aussi de réaliser des tests d'hypothèses. L'algorithme de Gibbs proposé pour prendre en compte les contraintes d'identifiablité converge rapidement avec de faibles autocorrélations des simulations *a posteriori*, réduisant ainsi le nombre de simulations nécessaires. La suite de ce travail consiste en l'étude des modèles SEMs sur données manifestes catégorielles.

Bibliographie

[1] Box, G. E. P. et Tiao G.C. (1973) *Bayesian Inference in Statistical Analysis*, Wiley.
[2] Gelman, A., Meng, X. L. et Stern, H. (1996) Posterior Predictive Assessment of Model Fitness via Realized Discrepancies. *Statistica Sinica*, 6, 733-807.
[3] Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B. (2004) *Bayesian Data Analysis*, Chapman & Hall/CRC.
[4] Lee, S. Y. (2007) *Structural Equation Modelling: A Bayesian Approach*, Wiley.
[5] Palomo, J., Dunson, D. B. and Bollen, K. (2007) Bayesian Structural Equation Modeling. In: S. Y. Lee (Ed): *Handbook of latent variable and related models*, Elsevier, 163–188.
[6] Tanner, M.A., Wong, W.H. (1987) The Calculation of Posterior Distributions by Data Augmentation, *Journal of the American Statistical Association*, 82, 528–540.