Contributions to Bayesian Structural Equation Modeling

Séverine Demeyer[1,2], Nicolas Fischer[1], and 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

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.

Keywords: structural equation modeling, Bayesian statistics, Gibbs sampling, latent variables, identifiability

1 Structural equation models

1.1 Context

Structural equation models (SEMs) are multivariate latent variable models used to represent latent structures of causality in data. The observed (manifest) variables are associated with latent variables in the outer (measurement) model and causality links are assumed between latent variables in the inner (structural) model. This situation typically arises with satisfaction surveys as illustrated in section 3 where the observed variables are the questions and the latent variables are loyalty, satisfaction and image as in figure 2.

1.2 Model

Denoting Y_i the row vector of observed values for individual i on the p manifest variables and Z_i the row vector of scores of individual i on the q latent variables, the measurement model is expressed as

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

where E_i is the measurement error term distributed $E_i \sim \mathcal{N}(0, \Sigma_\varepsilon)$ with Σ_ε diagonal and θ is the $q \times p$ matrix of regression coefficients.

If Z_i were known, the measurement model (1) would reduce to a linear regression model.

Denoting H_i the endogenous latent variables and Ξ_i the exogeneous latent variables, the structural equations are equivalently given by the following expressions

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

where Π is the $q_1 \times q_1$ matrix of regression coefficients between endogeneous latent variables , Γ is the $q_2 \times q_1$ matrix of regression coefficients between endogeneous and exogeneous latent variables. Δ_i is the error term distributed $\Delta_i \sim \mathcal{N}\left(0, \Sigma_\delta\right)$, independent with Ξ_i and Ξ_i is distributed $\mathcal{N}\left(0, \Phi\right)$.

1.3   The role of latent variables and identifiability constraints

Basically LV are unidimensional concepts, measured on manifest variables (usually on different scales). They provide practitioners with useful unobserved information on individuals. Since latent variables are not observed they are unscaled. Unlike Palomo et al. (2007) who advocate a free mean and variance model for latent variables, this paper aims to prove that latent variables should be standardized in the run of the analysis, for identifiability concerns, see section 2.6 for a brief overview. This standardization step is showed to be an heuristics leading to the improved Gibbs sampling algorithm presented in section 2.4.

2   Bayesian estimation of SEM

2.1   Bayesian estimation

In the latent variable model defined by equations (1) and (2), well-known techniques of data augmentation and imputation, see Tanner and Wong (1987), are implemented in a Gibbs algorithm (see section 2.4) under normality and conjugacy assumptions. See Box and Tiao (1973) for calculations in multivariate Normal models and Gelman et al.(2004) for Gibbs sampling.

2.2   Conditional posterior distribution of latent variables

Let $\Theta = \{\theta, \Sigma_\varepsilon, \Pi_0, \Gamma, \Sigma_\delta, \Phi\}$. The conditional posterior distribution of the latent variables is expressed as

$$
\begin{aligned}
\left[Z_i | Y_i, \Theta\right] &\propto \left[Y_i | Z_i, \Theta\right] \left[Z_i | \Theta\right] \\
&\propto \left[Y_i | Z_i, \theta, \Sigma_\varepsilon\right] \left[Z_i | \Pi_0, \Gamma, \Sigma_\delta, \Phi\right] \\
&\propto \left[Y_i | Z_i, \theta, \Sigma_\varepsilon\right] \left[H_i | \Xi_i, \Pi_0, \Gamma, \Sigma_\delta\right] \left[\Xi_i | \Phi\right]
\end{aligned}
\tag{3}
$$

where $[Y_i|Z_i, \theta, \Sigma_\varepsilon]$ is the likelihood of individual i computed from the measurement model (1) and $[Z_i|\Pi_0, \Gamma, \Sigma_\delta]$ is the joint prior distribution of latent variables deduced from the structural equations (2)

$$Z_i|\Pi_0, \Gamma, \Sigma_\delta, \Phi \sim \mathcal{N}(0, \Sigma_Z) \tag{4}$$

$$\Sigma_Z = \begin{pmatrix} \left(\Pi_0^t\right)^{-1} \left(\Gamma^t \Phi \Gamma + \Sigma_\delta\right) \Pi_0^{-1} & \left(\Pi_0^t\right)^{-1} \Gamma^t \Phi \\ \Phi \Gamma \Pi_0^{-1} & \Phi \end{pmatrix} \tag{5}$$

Immediate computation gives

$$Z_i|Y_i, \theta, \Sigma_\varepsilon, \Lambda, \Sigma_\delta, \Phi \sim \mathcal{N}\left(D\theta \Sigma_\varepsilon^{-1} Y_i, D\right) \tag{6}$$

where $D^{-1} = \theta \Sigma_\varepsilon^{-1} \theta^t + \Sigma_Z^{-1}$.

2.3   Conditional posterior distributions of parameters

The conditional posterior distribution of parameters is expressed as

$$\begin{aligned} [\Theta|Y, Z] &\propto [Y, Z|\Theta] [\Theta] \\ &\propto [Y|Z, \theta, \Sigma_\varepsilon] [\theta, \Sigma_\varepsilon] [H|\Xi, \Lambda, \Sigma_\delta] [\Lambda, \Sigma_\delta] [\Xi|\Phi] [\Phi] \end{aligned} \tag{7}$$

where prior independance between inner and outer parameters is assumed.

The last expression implies that the posterior distributions of parameters can be computed separately from the following expressions

$$\begin{aligned} [\theta, \Sigma_\varepsilon|Y, Z] &= [Y|Z, \theta, \Sigma_\varepsilon] [\theta, \Sigma_\varepsilon] \\ [\Lambda, \Sigma_\delta|Y, Z] &= [H|\Xi, \Lambda, \Sigma_\delta] [\Lambda, \Sigma_\delta] \\ [\Phi|Z] &= [\Xi|\Phi] [\Phi] \end{aligned} \tag{8}$$

Let $\theta_k = (\theta_{k1} \dots \theta_{kn_k})$, the vector of regression coefficients of block k where n_k is the number of manifest variables in block k with $\theta_{k1} = 1$ for identifiability (see section 2.6), Λ_k the kth column of Λ and $\Sigma_{\varepsilon k}$ and $\Sigma_{\delta k}$ the associated error terms. Conjugate prior distributions are

$$\begin{aligned} \theta_{kj}|\Sigma_{\varepsilon kj} &\sim \mathcal{N}(\theta_{0k}, \Sigma_{\varepsilon kj}\Sigma_{\varepsilon 0k}), \ \Sigma_{\varepsilon kj}^{-1} \sim Gamma\left(\alpha_{0\varepsilon k}, \beta_{0\varepsilon k}\right) \\ \Lambda_k|\Sigma_{\delta k} &\sim \mathcal{N}(\Lambda_{0k}, \Sigma_{\delta k}\Sigma_{\delta 0k}), \ \Sigma_{\delta k}^{-1} \sim Gamma\left(\alpha_{0\delta k}, \beta_{0\delta k}\right) \\ &\qquad\qquad \Phi^{-1} \sim InvWishart\left(R_0, d_0\right) \end{aligned} \tag{9}$$

Let Y_{kj} the jth manifest variable of block k, and Z_k the associated latent variable. Combining (8) and (9) gives the posterior distributions

$$\theta_{kj}|Y, Z, \Sigma_{\varepsilon kj} \sim \mathcal{N}\left(D_{kj}A_{kj}, \Sigma_{\varepsilon kj}D_{kj}\right)$$

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

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

$$D_{kj} = \left(Z_k^t Z_k + \Sigma_{\varepsilon 0k}^{-1}\right)^{-1}, \ A_{kj} = \Sigma_{\varepsilon 0k}^{-1}\theta_{0k} + Z_k^t Y_{kj}$$

$$\Lambda_k|Y, Z, \Sigma_{\delta k} \sim \mathcal{N}\left(\tilde{D}_k\tilde{A}_k, \Sigma_{\varepsilon k}\tilde{D}_k\right)$$

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

$$\tilde{D}_k = \left(Z_k^t Z_k + \Sigma_{\delta k}^{-1}\right)^{-1}, \ \tilde{A}_k = \Sigma_{\delta k}^{-1}\Lambda_{0k} + Z^t H_k$$

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

2.4   The Gibbs sampler

Gibbs algorithm (see figure 1) alternates sampling in the conditionnal posterior distribution of parameters given data and latent variables (step 1), and sampling in the conditional posterior distributions of latent variables given data and parameters (step 3 to 7). Step 2 is the heuristics whose role is to ensure identifiability of the model by scaling the latent variables.

Initialisation : $\theta^0, \Sigma_{\varepsilon}^0, \Lambda^0, \Sigma_{\delta}^0, \Phi^0$
At iteration t :

a. sampling in the conditional posterior distribution of latent variables:
   $Z^t \sim Z|Y, \theta^{t-1}, \Sigma_{\varepsilon}^{t-1}, \Lambda^{t-1}, \Sigma_{\delta}^{t-1}, \Phi^{t-1}$
b. **standardization of latent variables:** define Z^{*t} the standardized LV
c. $\Sigma_{\varepsilon}^t \sim \Sigma_{\varepsilon}|Y, Z^{*t}, \theta^{t-1}, \Lambda^{t-1}, \Sigma_{\delta}^{t-1}, \Phi^{t-1}$
d. $\theta^t \sim \theta|Y, Z^{*t}, \Sigma_{\varepsilon}^t, \Lambda^{t-1}, \Sigma_{\delta}^{t-1}, \Phi^{t-1}$
e. $\Sigma_{\delta}^t \sim \Sigma_{\delta}^|Y, Z^{*t}, \Lambda^{t-1}, \theta^t, \Sigma_{\varepsilon}^t, \Phi^{t-1}$
f. $\Lambda^t \sim \Lambda|Y, Z^{*t}, \Sigma_{\delta}^t, \theta^t, \Sigma_{\varepsilon}^t, \Phi^{t-1}$
g. $\Phi^t \sim \Phi|Y, Z^{*t}, \theta^t, \Sigma_{\varepsilon}^t, \Lambda^t, \Sigma_{\delta}^t$

Fig. 1. Steps of Gibbs algorithm

After enough runs of Gibbs algorithm, conditional posterior simulations are supposed to be drawn from the marginal distributions of parameters.

2.5 Validation

Validation is based on Posterior Predictive p-values as developed in Gelman and al. (1996). PP p-values are derived from posterior predictive distributions, integrated out both parameters and latent variables. Let y_{rep} be a simulated dataset under the same model that generated the observed dataset y, say H_0, and the same parameters Θ and latent variables Z. The posterior predictive distribution of y_{rep} is then defined as:

$$
\begin{aligned}
\mathbb{P}^{H_0}\left(y_{rep}|y\right) &= \int \mathbb{P}^{H_0}\left(y_{rep},\Theta,Z|y\right)\,\mathrm{d}\Theta\mathrm{d}Z \\
&= \int \mathbb{P}^{H_0}\left(y_{rep}|\Theta,Z\right)\left[\Theta,Z|y\right]\mathrm{d}\Theta\mathrm{d}Z
\end{aligned}
\tag{10}
$$

The PP p-value is defined as the tail probability of a given discrepancy function D (analogous to the use of statistics to compute classical p-values) under the posterior predictive distribution :

$$
\begin{aligned}
PPp\left(y\right) &= \mathbb{P}^{H_0}\left(D\left(y_{rep},\Theta,Z\right) \geq D\left(y,\Theta,Z\right)|y\right) \\
&= \int \mathbb{P}^{H_0}\left(D\left(y_{rep},\Theta,Z\right) \geq D\left(y,\Theta,Z\right)\right)\left[\Theta,Z|y\right]\mathrm{d}\Theta\mathrm{d}Z
\end{aligned}
\tag{11}
$$

Lee (2007) selected the following discrepancy function to test SEMs:

$$
D\left(y_{rep},\Theta,Z\right) = \sum_{i=1}^{n}\left(y_{rep_i} - Z^i\theta^i\right)^t \Sigma_\varepsilon^i \left(y_{rep_i} - Z^i\theta^i\right)
\tag{12}
$$

where θ^i and Σ_ε^i are the current values in the run of Gibbs algorithm.

The PP p-value is thus computed as the proportion of runs for which $D\left(y_{rep},\Theta^i,Z^i\right)$ is higher than $D\left(y,\Theta^i,Z^i\right)$. H_0 is not rejected if the PP p-value is near 0.5.

2.6 Identifiability issues

Identifiability of structural equation models is the injectivity of the likelihood function integrated out the latent variables namely

$$
\forall Y_i,\ [Y_i|\Theta] = \left[Y_i|\tilde{\Theta}\right] \implies \Theta = \tilde{\Theta}
\tag{13}
$$

where Y_i is marginally distributed as $\mathcal{N}\left(\mathbf{0},\Sigma_Y\right)$ and $\Sigma_Y = \theta^t\Sigma_Z\theta + \Sigma_\varepsilon$.

With the notations of section 2.3 and denoting $\Sigma_Z = \{\rho_{ij},\ 1 \leq i,j \leq K\}$, Σ_Y is the block matrix

$$
\Sigma_Y = \begin{pmatrix}
\rho_{11}\theta_1\theta_1^t + \Sigma_{\varepsilon 1} & \rho_{12}\theta_1\theta_2^t & \dots & \rho_{1K}\theta_1\theta_K^t \\
\rho_{12}\theta_1\theta_2^t & \rho_{22}\theta_2\theta_2^t + \Sigma_{\varepsilon 2} \dots & & \vdots \\
\vdots & & & \\
\rho_{1K}\theta_1\theta_K^t & \dots & & \dots\,\rho_{KK}\theta_K\theta_K^t + \Sigma_{\varepsilon K}
\end{pmatrix}
\tag{14}
$$

Identifiability constraints are directly derived from definition (13) applied to the Normal likelihood, giving $\Sigma_Y = \tilde{\Sigma_Y}$.

The identifiability equations arising from this equality are

$$\rho_{kk}\theta_{ki}^2 + \sigma_{ki}^2 = \tilde{\rho}_{kk}\tilde{\theta}_{ki}^2\tilde{\sigma}_{ki}^2, \ i = 1\ldots n_k, \ k = 1\ldots K \tag{15}$$

$$\rho_{kk}\theta_{ki}\theta_{kj} = \tilde{\rho}_{kk}\tilde{\theta}_{ki}\tilde{\theta}_{kj}, \ 1 \leq i < j \leq n_k, \ k = 1\ldots K \tag{16}$$

$$\rho_{kk'}\theta_{ki}\theta_{k'j} = \tilde{\rho}_{kk'}\tilde{\theta}_{ki}\tilde{\theta}_{k'j}, \ 1 \leq i \leq n_k, \ 1 \leq j \leq n_{k'}, \ k = 1\ldots K \tag{17}$$

Equations (15) and (16) are derived from the block diagonal elements of Σ_Y and equation (17) is derived from the extra block diagonal elements of Σ_Y.

If $\theta_{k1} = \tilde{\theta}_{k1}$ and $\rho_{kk} = \rho_{kk'}$ for a fixed k then equation (16) gives $\theta_{kj} = \tilde{\theta}_{kj}$ for all j. Reporting in equation 15 gives $\sigma_{ki}^2 = \tilde{\sigma}_{ki}^2$ for all k, i. Reporting in equation (17) gives $\rho_{kk'} = \tilde{\rho}_{kk'}$ for all k, k'. Consequently, a sufficient set of conditions for identifiability is $\theta_{k1} = 1$ and $\rho_{kk} = 1$ for all k.

The latter constraint is actually expressed in terms of the inner parameters, see the application, obtained by equating to 1 the diagonal elements of Σ_Z given in expression 5. However, posterior sampling of parameters given these constraints is complicated. The heuristics, consisting in standardizing latent variable after they have been drawn in their posterior distribution, should overcome this difficulty, as shown in the application.

3   Application

Consider a part of ECSI model considering only relationships between loyalty, satisfaction and image (see figure 2), on a subset of $n = 202$ individuals with no missing data. The full dataset is the demonstration dataset of XLStat software (http://www.xlstat.com/) with ordinal variables treated as continuous variables. The algorithm is implemented with R software.

Let θ_0 and λ_0 denote the common prior values of parameters. Priors on parameters are chosen to reflect confidence in the causality links: $\theta_0 = 0.5$, $\Lambda_0 = 0.5$, $\Sigma_{\varepsilon 0} = 1$, $\Sigma_{\delta 0} = 1$ and $\Phi_0 = 1$. Early convergence of Gibbs algorithm for all the parameters and law autocorrelation in posterior samples is observed. Formula (5) applied to this model with $\Pi_0 = \begin{pmatrix} 1 & 0 \\ -\pi_{12} & 1 \end{pmatrix}$ and $\Gamma = (\lambda_1 \lambda_2)$ gives the following expression of Σ_Z

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

Identifiability constraints are given by equating to 1 the diagonal elements (see section 2.6)

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

Posterior samples of these constraints are computed from the posterior samples of parameters. These distributions are centred in 1 with low dispersion as showed in figure 2, which tends to support the heuristics.
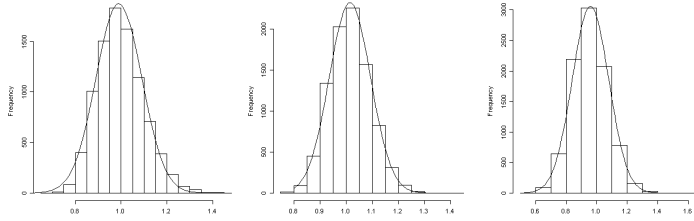
Fig. 2. Posterior distributions of the constraints a), b) and c) from left to right

Parameters of interest in SEMs are the correlations between manifest variables and latent variables and between latent variables. In table 1 θ_{12}, θ_{22}, θ_{23}, θ_{32}, θ_{33}, θ_{34}, θ_{35} and λ_2 are correlation coefficients whereas π_{12} and λ_1 are coefficients of a multiple regression. For identifiability $\theta_{11} = 1$, $\theta_{21} = 1$ and $\theta_{31} = 1$ (see section 2.6). From table 1, *satisfaction* and *Image* are highly correlated (0.796), meaning that *Image* has a great influence on *Satisfaction*. All the correlations are represented in the summary graph of figure 2.

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

Table 1. Regression coefficients: posterior mean and standard deviation (sd)

The PPp-value $0.37 < 0.5$ is due to the poor adjustment of data with the Normal distribution. This example however shows interesting features of Bayesian analysis, like hypothesis testing with PPp-values and the possibility to look at the variability of parameters and to the variability of functions of parameters.

4   Conclusion and perspectives

Posterior distributions of all the parameters of SEMs are derived under Normality and conjugacy assumptions. They are useful material to investigate aspects of the model like the variability of parameters and functions of parameters and hypothesis testing. The Gibbs algorithm augmented by the heuristics presented in this paper, converges rapidly, with low autocorrelations in posterior samples, thus reducing the number of iterations needed.
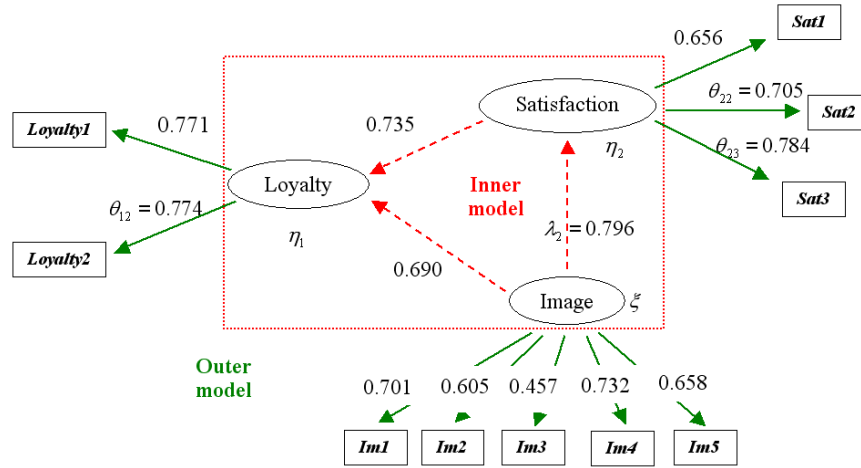
Fig. 3. Graph of correlations

More generally, this paper advocates systematic computation of identifiability constraints. Future work will concern structural equation modelling of mixed continuous and categorical data.

Acknowledgements

References

BOX, G. E. P. and TIAO G.C. (1973) : *Bayesian Inference in Statistical Analysis (Wiley Classics Library)*. Wiley.

GELMAN, A., MENG, X. L. and STERN, H. (1996) : Posterior Predictive Assessment of Model Fitness via Realized Discrepancies. *Statistica Sinica 6, 733-807*.

GELMAN, A., CARLIN, J.B., STERN, H.S., RUBIN, D.B. (2004) : *Bayesian Data Analysis (Texts in Statistical Science)*. Chapman & Hall/CRC.

LEE, S. Y. (2007) : *Structural Equation Modelling: A Bayesian Approach (Wiley Series in Probability and Statistics)*. Wiley.

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.

TANNER, M.A., WONG, W.H. (1987) : The Calculation of Posterior Distributions by Data Augmentation. *Journal of the American Statistical Association 82, 528–540*.