Approche bayésienne des modèles à équations structurelles utilisant l'expansion paramétrique

Séverine Demeyer [1,3] & Jean-Louis Foulley [2] & Nicolas Fischer [1] & Gilbert Saporta [3]

[1] *Laboratoire National de Métrologie et d'Essais, LNE, 1 rue Gaston Boissier, Paris*
(E-mail: sev.demey@yahoo.fr, nicolas.fischer@lne.fr*)*
[2] *INRA-GABI-PSGen, UMR 1313, 78350 Jouy-en-Josas, France*
[3] *Chaire de Statistique Appliquée & CEDRIC, CNAM, 292 rue Saint Martin, Paris*

Résumé

Les modèles à équations structurelles à variables latentes (SEM) sont utilisés pour représenter des relations de causalité dans les données, de telle façon que la structure de corrélation des variables observées est résumée dans la structure de corrélation de variables latentes construites à cet effet. Ce papier propose une analyse bayésienne des modèles SEM reposant sur l'analyse de la matrice de covariance des variables latentes utilisant l'expansion paramétrique pour surmonter les problèmes d'identifiabilité. Ce papier est appliqué à l'estimation d'un modèle structurel schématisant un processus de mesure de polluants de l'eau.

Abstract

Structural Equation Models with latent variables (SEM) are multivariate models used to model causality relationships in data (observed variables), such that the correlation structure of the observed variables is transferred into the correlation structure of the latent variables. A Bayesian approach of SEM is proposed based on the analysis of the covariance matrix of latent variables using parameter expansion to overcome identifiability issues. This paper is applied to the estimation of a SEM modelling a measurement process of water pollutants.

Mots-clés: Modèles à équations structurelles, variables latentes, identifiabilité, analyse bayésienne, augmentation des données, expansion paramétrique, algorithme de Gibbs, métrologie, comparaisons interlaboratoires, connaissances d'experts

1 Spécification du modèle SEM pour des variables observées mixtes

Les modèles SEM sont des modèles multivariés permettant de représenter des structures latentes de causalité dans les données. Les variables observées sont associées à des variables latentes dans le modèle externe dit aussi modèle de mesure et les relations

de causalité supposées entre variables latentes sont représentées dans le modèle interne dit aussi modèle structurel, voir figure 1.

1.1 Modèle de mesure

Soit Y_i le vecteur ligne des réalisations des variables observées mixtes continues, binaires et catégorielles ordonnées pour l'individu i sur les p variables observées, réparties en q blocs disjoints. On suppose que chaque bloc reflète un concept unidimensionnel, modélisé par une unique variable latente continue. Soit Z_i le vecteur ligne des q variables latentes continues pour l'individu i.

Les modèles SEM à variables observées mixtes sont définis dans le cadre des modèles linéaires généralisés où les variables observées binaires et catégorielles ordonnées sont quantifiées en réponses latentes suivant Albert et Chib (1993), en utilisant des liens probit.

Soit $Y_i^* = \left\{Y_{ikj}^*,\, k = 1...q,\, j = 1...n_k\right\}$ le vecteur ligne des réponses latentes où n_k est nombre de variables observées dans le bloc k.

Le modèle interne relie chaque réponse latente à sa variable latente associée de manière réflexive (car chaque variable observée reflète sa variable latente). On suppose que les variables observées sont conditionnellement indépendantes sachant les variables latentes.

L'écriture matricielle du modèle externe pour l'individu i est

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

où E_i est le terme d'erreur distribué $E_i \sim \mathcal{N}(0, \Sigma_\varepsilon)$ avec Σ_ε diagonale et θ est la matrice $q \times p$ des coefficients de régression.

1.2 Modèle interne: modélisation alternative

Soit H_i les variables latentes endogènes et Ξ_i les variables latentes exogènes. Les équations structurelles sont les équations simultanées définies par

$$H_i = H_i\Pi + \Xi_i\Gamma + \Delta_i \tag{2}$$

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 endogènes et les variables latentes exogènes. Δ_i est le terme d'erreur distribué $\Delta_i \sim \mathcal{N}(0, \Sigma_\delta)$ avec Σ_δ diagonale, indépendante de Ξ_i et Ξ_i est distribué $\mathcal{N}(0, \Phi)$.

Puisqu'une approche bayésienne permet de travailler avec la distribution jointe des variables latentes, il est équivalent de travailler avec la matrice de corrélation des variables latentes (sous l'hypothèse de multinormalité des distributions conditionnelles des Z_i) de telle sorte que le modèle interne considéré dans ce papier est donné par

$$Z_i | R_Z \sim N(0, R_Z) \tag{3}$$

où R_Z est une matrice de corrélation..

De plus R_Z^{-1} contient les coefficients de régression des régressions possibles entre variables latentes.

1.3 Identifiabilité des modèles SEM et expansion paramétrique

Dans ce papier, l'expansion paramétrique est utilisée pour surmonter les problèmes d'identifiabilité liés à la nature non observée des variables latentes. On surmonte ces problèmes en fixant la moyenne et la variance des variables latentes. Il est aisé de contraindre la moyenne à 0 lors de l'imputation des variables latentes. En revanche fixer une échelle est plus astucieux et est l'objet de cette section.

L'expansion paramétrique consiste dans ce cas à simuler la matrice de covariance des variables latentes puis à la transformer en matrice de corrélation (voir la section 2.1) de telle sorte que cette matrice soit la matrice de covariance des mêmes variables sous forme réduite.

2 Estimation bayésienne des modèles SEMs à variables observées mixtes

2.1 Implémentation de l'expansion parametrique

L'implémentation de l'expansion paramétrique réalisée dans ce papier imite l'implémentation de l'expansion paramétrique dans le cas des algorithmes PX-EM comme exposée dans Liu et al. (1998) et diffère de l'implémentation habituelle pour les algorithmes MCMC décrite dans Liu et Wu (1999).

L'expansion paramétrique (Liu et al. (1998)) consiste à travailler avec des paramètres non identifiés dans le modèle des données complètes $f(Y, Z|\theta)$ et à indicer des modèles de données augmentées $p(Y, W|\theta, \alpha)$ ainsi que des variables latentes augmentées W correspondant à une valeur d'un paramètre d'expansion noté α, de telle sorte que la vraisemblance observée soit conservée, c'est-à-dire vérifie

$$f(Y|\theta) = \int f(Y, Z|\theta)\,\mathrm{d}Z = \int p(Y, W|\theta, \alpha)\,\mathrm{d}W \tag{4}$$

La transformation indicée par le paramètre d'expansion est un C^1 difféomorphisme.

Dans ce papier les variances des variables latentes structurelles sont les paramètres d'expansion.

Dans le modèle des données complètes $Z \sim N(0, R_Z)$ où R_Z est une matrice de corrélation.

On introduit les paramètres non identifiés de variance des variables latentes $\alpha_1, ..., \alpha_q$ et on définit la matrice diagonale $\alpha = diag(\alpha_1, ..., \alpha_q)$. Il s'ensuit que $W = \alpha^{\frac{1}{2}} Z$ sont les

nouvelles variables latentes dans le modèle étendu indicé par α où $W \sim N(0, \Sigma_W)$ avec $\Sigma_W = \alpha^{\frac{1}{2}} R_Z \alpha^{\frac{1}{2}}$ une matrice de covariance.

La matrice de corrélation des variables latentes est finalement obtenue par $R_Z = \alpha^{-\frac{1}{2}} \Sigma_W \alpha^{-\frac{1}{2}}$.

On applique également l'expansion paramétrique aux variances résiduelles des réponses latentes où le paramètre d'expansion est la variance résiduelle, voir Meza et al. (2009) .

2.2 L'algorithme PX-Gibbs

L'algorithme PX-Gibbs d'estimation des modèles SEM sur données observées mixtes repose sur le calcul de toutes les distributions conditionnelles *a posteriori* des grandeurs suivant une implémentation de Gibbs classique (voir Lee (2007)) augmenté du calcul des distributions conditionnelles *a posteriori* des paramètres d'expansion dans le modèle des données augmentées.

Précisément l'algorithme PX-Gibbs nécessite deux schémas PX menant à un algorithme en trois grandes étapes décrites ci-après, homologues des étapes implémentées dans l'algorithme PX-EM. Pour une description complète de ces trois étapes on se reportera à Demeyer et al. (2011).

- **Etape 1:** implémentation PX dans les modèles probit pour générer des réponses latentes vérifiant la contrainte d'identifiabilité de variance résiduelle unité conditionnellement aux variables latentes structurelles et aux valeurs courantes des paramètres dans le modèle des données complètes.

- **Etape 2:** implémentation PX dans le modèle structurel pour générer des imputations des variables latentes structurelles vérifiant la contrainte de la matrice de covariance sous la forme d'une matrice de corrélation, conditionnellement aux réponses latentes simulées à l'étape 1 et aux valeurs courantes des paramètres dans le modèles des données complètes.

- **Etape 3:** tirage des paramètres externes dans leur distribution conditionnelle *a posteriori* dans le modèle des données complètes.

3 Application

La méthodologie a été appliquée en métrologie pour étudier les relations entre des sources de biais de mesure lors de comparaisons interlaboratoires.

La schématisation du processus de mesure par des experts a fait apparaître trois concepts latents: la préparation de la mesure, la mesure et le contrôle qualité, ayant un effet direct sur la qualité des mesures des laboratoires et donc sur les biais de mesure des laboratoires. Ces concepts ont des relations entre eux et sont associés aux variables observées (les sources de biais) selon le modèle SEM de la figure 1.

Le modèle a été estimé à partir des réponses de 18 laboratoires à un questionnaire relatif à la mesure des micropolluants organiques dans l'eau (pesticides). Les résultats souffrent du petit nombre de laboratoires participant qui se ressent en terme de précision des estimations mais permettent toutefois d'identifier des tendances. Ainsi d'après les données le contrôle qualité a plus d'influence sur la préparation de la mesure que sur la mesure et on retrouve bien que la préparation de la mesure a une influence sur la mesure.

Ce schéma mène à l'interprétation suivante: pour améliorer la qualité des mesures, on doit agir sur les variables du contrôle qualité et parmi elles, une action est à mettre en oeuvre sur les sources de biais les plus corrélées au contrôle qualité c'est-à-dire la correction par rapport à l'étalon et la nature de l'étalonnage.

Il apparaît clairement que les concepts latents résument la structure des sources de biais observées et peuvent ainsi être utilisés dans une analyse des résultats de mesure des laboratoires. Pour plus de détails le lecteur intéressé se reportera aux travaux de thèse de Demeyer (2011).

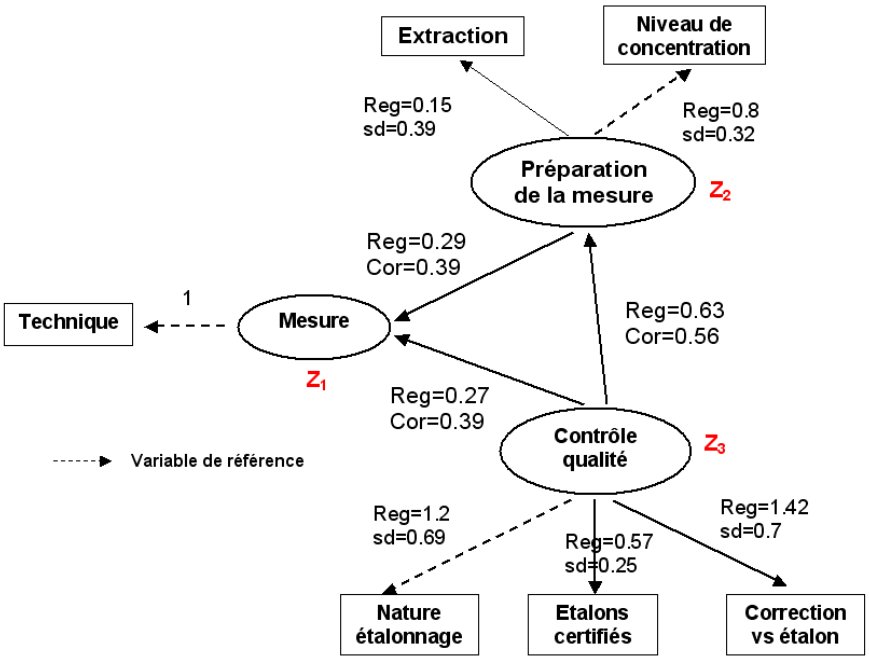

Figure 1: Estimation du modèle structurel. *Reg* est l'estimation du coefficient de régression, *sd* est l'écart-type associé et *cor* est l'estimation du coefficient de corrélation.

4 Conclusion

En conclusion le cadre bayésien d'estimation des modèles SEM présenté permet l'intégration immédiate de données observées continues, binaires et catégorielles ordonnées fréquemment rencontrées simultanément en pratique et offre un cadre rigoureux dans la perspective d'imputation de données manquantes ou censurées, de comparaison et de validation de modèles.

Plus généralement, le cadre bayésien permet d'intégrer ou de combiner un modèle SEM à d'autres modèles en actualisant les distributions conditionnelles *a posteriori* des différents modèles. Cette pratique permet par exemple d'intégrer une expertise à un modèle de mesure existant.

Bibliographie

[1] Albert, J.H., Chib S., (1993) *Bayesian analysis of binary and polychotomous response data*, Journal of the American Statistical Association, 88(422):669–679.

[2] Demeyer, S. et Foulley, J.-L. et Fischer, N. et Saporta, G. (2011) *Bayesian analysis of structural equation modelling using parameter expansion*, in Statistical Learning and Data Science, Chapman & Hall /CRC, à paraître.

[3] Demeyer, S. (2011) *Approche bayésienne de l'évaluation de l'incertitude de mesure: application aux comparaisons interlaboratoires*, Thèse du Conservatoire National des Arts et Métiers, Paris.

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

[5] Liu, C. et Rubin, D. B. et Wu, Y. N. (1998) *Parameter expansion to accelerate EM: The PX-EM algorithm*, Biometrika, 85:755–770.

[6] Liu, J. S. et Wu, Y. N. (1999), *Parameter expansion for data augmentation*, Journal of the American Statistical Association, 94(448):1264–1274.

[7] Meza, C. et Jaffrézic, F. et Foulley, J.-L. (2009), *Estimation in the probit normal model for binary outcomes using the SAEM algorithm*, Computational Statistics and Data Analysis, 53(4), 1350–1360.