EM estimation of a Structural Equation Model

Myriam Tami*[1], Xavier Bry[1] and Christian Lavergne[1]

[1]Universities of Montpellier, IMAG, France
*Email: myriam.tami@umontpellier.fr

Abstract: We propose an estimation method of a Structural Equation Model (SEM). It consists in viewing the Latent Variables (LV's) as missing data and using the EM algorithm to maximize the whole model's likelihood, which simultaneously provides estimates not only of the model's coefficients, but also of the values of LV's. Through a simulation study, we investigate how fast and accurate the method is, and eventually apply it to real data.

Introduction

The proposed approach is an estimation method of a SEM linking latent factors. It provides estimates of the coefficients of the model and its factors at the same time. This method departs from more classical methods such as LISREL. In fact, LISREL mainly focuses on the covariance structure and the LV scores estimation is based on a least-squares technique performed on mere measurement equations. Contrary to PLS-like methods, we do not constrain factors to belong to the spaces spanned by the Observed Variables (OV's), but only to be normally distributed.

1 The model and data notations

The data consists in blocks of OV's describing the same n independent units. $Y = \{y_i^j\}$ (resp. $X^m = \{x_i^{j,m}\}$); $i \in [\![1, n]\!]$, $j \in [\![1, q_Y]\!]$ (resp. $j \in [\![1, q_m]\!]$, $m \in [\![1, p]\!]$) is the $n \times q_Y$ (resp. $n \times q_m$) matrix coding the dependent block of OV's (resp. m^{ieth}-explanatory block of OV's), identified with its column-vectors. T (resp. T^m) refers to a $n \times r_T$ (resp. $n \times r_m$) matrix of covariates. For the sake of simplicity, the SEM we handle here is a restricted one. It contains only one structural equation, relating a dependent latent factor g, underlying block Y, to p explanatory latent factors f^m respectively underlying blocks X^m. The SEM consists of $p + 1$ measurement equations and one structural equation :

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

where, $\varepsilon^g \in \mathbb{R}^n$ (resp. ε^Y, ε^m) is a disturbance vector (resp. are disturbance matrices) and $\forall m \in [\![1,p]\!]$, $\theta = \{D, D^m, b, a^m, c^1, c^2, \psi_Y, \psi_m\}$ is the set of parameters. The main assumptions of this model are the following: f^m are standard normal; g is normal with zero-mean, and its expectation conditional on all f^m is a linear combination of them; $\varepsilon^g \sim \mathcal{N}(0,1)$; ε^g is independent of ε^Y and ε^m, $\forall m \in [\![1,p]\!]$.

2 Estimation using the EM algorithm

We propose to carry out likelihood maximization through an iterative Expectation-Maximization algorithm (Dempster et al. (1977)). If we consider factors as missing data, EM algorithm enables us to estimate the factors. Let $Z = (Y, X^1, \ldots, X^p)$ be the OV's, $h = (g, f^1, \ldots, f^p)$ the LV's. To maximize the log-likelihood associated with the complete data $\mathcal{L}(\theta; Z, h)$, in the EM framework, we must solve: $\mathbb{E}_z^h \left[\frac{\partial}{\partial\theta}\mathcal{L}(\theta; Z, h)\right] = 0$. Thanks to the explicit solutions of this system and the distribution of h conditional on Z we design an algorithm. It is a rapidly converging iterative procedure starting from a good initialization. The iteration equations have been given in detail in Bry et al. (2016) and will be presented.

3 Results and application

A sensitivity analysis has been performed to investigate how the quality of estimations could be affected by the number n of observations and the number q of OV's in each block. The results were that the sample size n proved to have more impact on the quality of parameter estimation and factor reconstruction than the number of OV's. We advise to use a minimal sample size of $n = 100$.

Conclusion

This method can estimate quickly and precisely factors of the SEM (in addition to estimating its loadings) by maximization of the whole model's likelihood. Various simulations and an application on real data will be presented.

References

Bry, X., Lavergne, X., Tami, M. (2016). EM estimation of a Structural Equation Model *in review.*

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society.*