Finite Mixture EFA in Mplus

November 16, 2007

In this document we describe the Mixture EFA model estimated in Mplus. Four types of dependent variables are possible in this model: normally distributed, ordered categorical with logit or probit link, Poisson distributed with the exponential link function, and censored variables. Inflation is not available for the Censored and Poisson variables.

Suppose that we estimate a K class model with M factors and P dependent variables. Denote the variables by $Y_1, ..., Y_P$ and the normally distributed factors by $\eta_1, ..., \eta_M$. Let η be the vector of all latent factors $\eta = (\eta_1, ..., \eta_M)$. The Mixture model is based on a single categorical latent class variable C.

For a normally distributed variable Y_p we estimate the following model in class k

$$Y_p = \nu_{kp} + \lambda_{kp}\eta + \varepsilon_p$$

where ν_{kp} is the intercept parameter, λ_{kp} is a vector of loadings of dimension M, and ε_p is a zero mean normally distributed residual with variance θ_{kp}.

For an ordered categorical variable Y_p we estimate the following model in class k

$$P(Y_p = j) = F(\tau_{kpj} - \lambda_{kp}\eta) - F(\tau_{kpj-1} - \lambda_{kp}\eta)$$

for $j = 1, ..., r_p$ where r_p is the number of categories that the variable Y_p takes. The parameters τ_{kpj} are monotonically increasing for j and for identification purposes $\tau_{kpr_p} = \infty$ and $\tau_{kp0} = -\infty$. The function F is either the standard normal distribution function, for probit link, or the logit distribution function $F(x) = 1/(1 + Exp(-x))$, for logit link. Alternatively we can specify the model as follows

$$Y_p = j \Leftarrow \tau_{kpj-1} \leq Y_p^* < \tau_{kpj}$$

where

$$Y_p^* = \lambda_{kp}\eta + \varepsilon_p$$

where ε_p is a residual with distribution F.

For Poisson distributed variables we estimate the following model in class k

$$P(Y_p = j) = e^{-Y_p^*}\frac{(Y_p^*)^j}{j!}$$

where

$$Y_p^* = \nu_{kp} + \lambda_{kp}\eta$$

and the parameters to be estimated are again the intercept ν_{kp} and the loading vector λ_{kp}.

For censored variables Y_p we estimate the following model in class k

$$Y_p = \begin{cases} Y_p^* & \text{if } Y_p^* > c_p \\ c_p & \text{if } Y_p^* \leq c_p \end{cases}$$

where c_p is the censoring limit and Y_p^* is latent normally distributed variable

$$Y_p^* = \nu_{kp} + \lambda_{kp}\eta + \varepsilon_p$$

where ν_{kp}, λ_{kp}, and the variance θ_{kp} of the zero mean residual ε_p are to be estimated. The above model is for censored variables with a lower end bound. Similar model is available for censored variables with an upper end bound.

We also estimate an unrestricted correlation matrix Ψ_k for the factors η in class k when we estimate the model with oblique rotation. If we estimate the model with orthogonal rotation the correlation matrix is fixed to the identity matrix, i.e., the factors are assumed standard normal and orthogonal in all classes. Finally we estimate an unrestricted distribution for the latent class variable C, i.e., we estimate the parameters $p_k = P(C = k)$.

The above model is not identified in principle. To be identified the model has to include an additional $M(M-1)$ restrictions for oblique rotations or $M(M-1)/2$ restrictions for orthogonal rotations. Before we proceed with a loading rotation algorithm however we standardize the loadings with respect to the $Var(Y_p^*)$. For normally distributed Y_p we assume that $Y_p^* = Y_p$. We construct the standardized loadings λ_{kp}^* as follows

$$\lambda_{kp}^* = \lambda_{kp}/\sqrt{(Var(Y_p^*))}$$

where

$$Var(Y_p^*) = \lambda_{kp}\Psi_p\lambda_{kp}^T + \theta_{kp}$$

where for censored and normal variables θ_{kp} is as specified in the model, for categorical probit link variable it is $\theta_{kp} = 1$, for categorical logit link variable $\theta_{kp} = \pi^2/3$ and for Poisson variables $\theta_{kp} = 0$. Similarly we standardize the θ_{kp} parameter

$$\theta_{kp}^* = \theta_{kp}/Var(Y_p^*)$$

Note also that as constructed the standardized loadings are on the correlation scale, that is, if Λ_k^* is the matrix of all standardized loadings and Θ_k^* is the diagonal matrix with all θ_{kp}^* on the diagonal, the estimated correlation matrix of $Y^* = (Y_1^*, ..., Y_P^*)$ is

$$\Lambda_k^*\Psi\Lambda_k^{*T} + \Theta_k^*.$$

We now define the rotation criteria that will identify the loadings and the factor correlation Ψ. All oblique factor rotations are defined by a square matrix H of dimension M such that HH^T has ones on the diagonal. All orthogonal rotations are defined by orthogonal square matrices of dimension M, i.e., $HH^T = I$, where I is the identity matrix. All such factor rotations lead to equivalent factor models with M factors. We estimate the rotation that minimizes the simplicity function, i.e., the rotation criteria

$$Q(\Lambda^* H)$$

across all rotation matrices H, where the rotation criteria can be any rotation criteria such as cf-varimax, quartimin, geomin etc, supported by Mplus. With this additional constraint the loadings and factor correlation are uniquely defined.

We now focus on the output reported by Mplus. For each class the rotation is performed independently, since all loadings and residual covariances are class specific. In the Mplus output we report the rotated standardized loadings $\Lambda^* H$, where H is the optimal rotation. Standard errors for the rotated standardized loadings are also reported. In addition the class specific intercepts ν_{kp} are reported, as well as the threshold parameters τ_{kpj}. These parameters are reported in their original metric, however the threshold parameters τ_{kpj} are also reported in the standardized correlation metric. Denote these by τ_{kpj}^*. Consequently the estimated probabilities for each category is computed as follows

$$P(Y_p = j|C = k) = \Phi^{-1}(\tau_{kpj}^*) - \Phi^{-1}(\tau_{kpj-1}^*).$$

This computation is exact for the probit link function, however it is only approximate for the logit link function.

The Mixture EFA model estimation can be challenging in some instances. When all dependent variables are normally distributed there is no numerical integration involved in the estimation and the computation is fairly quick, however sufficient number of random starts should be used to ensure that the global log-likelihood maximum is reached. When some of the variables are not normally distributed, i.e., Poisson, censored, and ordered categorical variables, numerical integration is used for all factors and thus the computation will be significantly slower. With Poisson, censored, and ordered categorical variables the Mixture EFA model is possible but because of the numerical integration and the random starts perturbation the computational time might be substantial. Mixture EFA with binary variables is a particularly difficult model to estimate because of the flexibility of the model and fairly little information provided by binary variables - in particular it is fairly easy to exceed or approach the maximum degrees of freedom when only a few binary variables are used. In addition for Mixture EFA models with categorical variables, the best log-likelihood value found in multiple starting value perturbations, can be difficult to replicate, again due to the flexibility of the model.

Additional information on mixture factor analysis can be found in McLachlan and Peel (2000) and McLachlan et al. (2004). Mixture factor analysis with categorical variables is discussed in Muthen and Asparouhov (2006). Mixture EFA analysis is illustrated in Example 4.4, Mplus User's Guide (Muthen and Muthen, 1998-2007).

References

McLachlan, G. & Peel, D. (2000). Finite mixture models. New York: John Wiley & Sons.

McLachlan, G.J., Do, K.A., & Ambroise, C. (2004). Analyzing microarray gene expression data. New York: John Wiley & Sons.

Muthen, B. & Asparouhov, T. (2006). Item response mixture modeling: Application to tobacco dependence criteria. Addictive Behaviors, 31, 1050-1066.

Muthen, L.K. & Muthen, B.O. (1998-2007). Mplus User's Guide. Fifth Edition. Los Angeles, CA: Muthen & Muthen