Estimator choices with categorical outcomes

Muthén, B., Muthén., L. & Asparouhov, T.

March 24, 2015

Table 1: Comparisons of estimators for categorical factor analysis (+ implies an advantage and - implies a disadvantage)

Criteria	Weighted least squares	Maximum likelihood	Bayes
Large number of factors	+	−	+
Large number of variables	−	+	+
Large number of subjects	+	−	−
Small number of subjects	−	+	+
Statistical efficiency	−	+	+
Missing data handling	−	+	+
Test of LRV structure	+	−	+
Ordered polytomous variables	+	−	−
Heywood cases	−	−	+
Zero cells	−	+	+
Residual correlations	+	−	±

It is instructive to compare the three estimation procedures of weighted least squares, maximum likelihood, and Bayes from the point of view of their relative strengths in practice. A summary of the comparisons is given in Table 1. Weighted least squares estimation is the default in Mplus because of several important advantages, most notably its computational speed. This does not imply that it is always the best choice, but it can be a good starting point. Following is a discussion of the relative merits of the three estimators for the criterion of each row in Table 1.

The number of factors has little influence on the computations for weighted least squares. In contrast, maximum-likelihood is at a disadvantage with a large number of factors because this leads to a large number of dimensions of numerical integration, which is both time and storage space (memory) consuming. In this context, a large number of factors means four or more factors. With four factors and the default of 15 integration points per dimension, a total of $15^4 = 50,625$ points are used, leading to slow computations. With exploratory factor analysis, the Mplus default is to use only 7 integration point per dimension with the idea that a sufficiently good approximation to the factor pattern is obtained. With many factors, it may be advantageous to use Monte Carlo integration for maximum likelihood estimation because in this case the computational work does not depend on the number of factors. A total of 5000 points typically gives good results, but in some cases 500 points give a sufficiently good approximation.

Maximum-likelihood and Bayes have the advantage of more easily handling a large number of variables than weighted least-squares. This is because the weight matrix of weighted least-squares grows by the power of four with the number of variables. In this context, a large number of variables means say 50 or more variables. In contrast, the computational work of maximum-likelihood and Bayes increases only linearly as a function of the number of variables. Weighted least squares can be speeded up by using options to not compute χ^2 or standard errors, which are more computationally demanding than getting the point estimates.

The number of subjects has ignorable influence on weighted least squares estimation. Maximum likelihood estimation is at a potential disadvantage with a large number of subjects because its computational time is a function of the product of the number of integration points and the sample size. Maximum likelihood using Monte Carlo integration is advantageous in such cases. The number of factors interacts with the sample size in determining the relative computational burden of maximum likelihood and Bayes. With small to medium-size samples, the number of factors has much less influence on the Bayes computations than on the maximum likelihood computations. With large samples, say several thousand, the Bayes advantage can be reversed.

A small number of subjects, say less than 200, has a negative impact on the quality of the weighted least squares tests of model fit and parameter standard errors. Parameter estimates are less negatively affected. The unweighted least squares estimator ULSMV can be advantageous to WLSMV in small samples (Forero et al. 2009). The full-information estimation of maximum likelihood may be preferable in such cases (Forero & Maydeu-Olivares, 2009). While less explored, the full-information approach of Bayes may also be preferable to weighted least squares.

Maximum-likelihood estimation uses all available information in the data and is optimal from a statistical point of view because it is an efficient estimator, that is, its parameter estimates have as small standard errors as possible. Bayesian estimation shares this quality because Bayes and maximum-likelihood are asymptotically equivalent when non-informative priors are used for Bayes. In contrast, weighted least-squares uses limited information from only the second-order moments, that is, bivariate distributions from pairs of items, and ignores information from higher-order moments. The loss of information may not be great, however, as indicated for example by Christofferson (1975).

Missing data is handled better by maximum-likelihood and Bayesian estimation than by weighted least squares. The most optimal approach of estimation under the assumption of missing at random (MAR) is allowed with maximum-likelihood and Bayesian estimation, but not with weighted least-squares. MAR allows missingness for a subject to be influenced by the variables that are observed for the subject. Estimation draws on all available data. In contrast, by using only limited information from pairs of variables, weighted least-squares requires a pairwise present approach (Asparouhov & Muthén, 2010d). When estimating a LRV correlation, information is not used for subjects who has missing data on either of the two variables. Weighted least-squares allows missingness to be a function of only observed covariates, not observed outcomes. In the factor models discussed in this chapter, there are no covariates, and weighted least-squares is correct only under the assumption of missing completely at random (MCAR). MCAR is a much stronger assumption than MAR. In this way, the simplicity advantage of using bivariate information is a disadvantage from a missing data point of view.

It is of interest to test the structure on the LRV correlations imposed by the factor model. The LRV correlations are a relevant target when the normality assumption for the factors is combined with probit to create a multivariate normal distribution for the latent response variables. Weighted least squares gives a χ^2 test of model fit to the sample LRV correlations. Although this does not test fit against the data, this is nevertheless a useful way to study the factor structure. Maximum likelihood does not offer such a test. Bayes offers this type of test using the approach of posterior predictive checking, where a discrepancy function in the form of a χ^2 test is obtained by u^* data generated from the estimated model, and where correlations based on these generated data are compared to those estimated from the model.

With polytomous variables, weighted least squares estimation is as fast as with binary variables. With maximum likelihood estimation, polytomous variables lead to slower computations. This is partly due to optimizing with respect to many more parameters due to having more threshold parameters. A more critical aspect, however, is due to data collapsing. With categorical variables, a time-saving device is to summarize the raw data in terms of response pattern frequencies. With binary variables, this collapsed data information is much less extensive than with polytomous variables. Bayes estimation is considerably slowed down when moving from binary to polytomous variables. This is for algorithmic reasons due to the technical necessity of switching from the regular Markov Chain Monte Carlo approach of Gibbs sampling to Metropolis-Hastings.

Heywood cases, that is, solutions corresponding to negative residual variances, may occur with binary items even though residual variances are not free parameters to be estimated. Often, Heywood cases lead to non-convergence, although milder cases (smaller negative variances) may converge. With weighted least squares and the Delta parameterization the latent response variable variances are standardized to one. The latent response variable variance explained by the factors (the communality) must not exceed one in order for the residual variance to be obtained as a positive remainder. When it exceeds one, a Heywood case occurs and a negative residual variance is printed. With the Theta parameterization, the residual variance is fixed at one and a Heywood case materialize as an exploding factor loading.

Maximum likelihood estimation likewise uses a fixed residual variance and Heywood cases materialize as exploding loadings. Bayesian analysis avoids Heywood cases by considering latent response variables with residual variance fixed at one.

It is common to observe zero cells in the frequency table with binary variables that have strong skewness or are observed in small samples, or where both those conditions hold. In such cases the estimation of tetrachoric correlations is problematic in that not enough information is available (see Brown & Benedetti, 1977; Savalei, 2011). With a zero cell in a 2×2 table, the tetrachoric correlation is theoretically either $+1$ or -1 depending on the position of the zero cell. This means that sample tetrachorics are not a good summary of the data structure and therefore fitting the model to these correlations as is done with weighted least-squares can give distorted parameter estimates. In case of a zero cell, a common approach is to add $0.5/n$ to the zero cell, but a better approach is perhaps to avoid using the correlation, either by deleting a variable, creating one three-category variable from the non-zero cells of the two binary variables, or switching to another estimator. Maximum-likelihood estimation and Bayesian analysis may be more successful in such settings in that the intermediate step of estimating tetrachorics is avoided and the model fitted directly to the likelihood.

Frequently, there is a need to add residual correlations to the factor model. The estimators differ with respect to the ease with which this can be done. Under probit, the latent response variables have a multivariate normal distribution which has the advantage of making available a correlation parameter for every pair of latent response variables. This implies that a factor model can be easily expanded to allow for correlated residuals. This is a strength of weighted least-squares and Bayesian estimation, where the latent response variables play a part in the computations. With weighted least squares using EFA, the correlated residuals can be added to the model using the exploratory structural equation modeling (ESEM) approach. With Bayes, ESEM is not yet available and correlated residuals can be added only with confirmatory factor analysis models. In contrast, with maximum-likelihood estimation, the latent response variables do not play a part in the computations. Given the conditional probability curve formulation of maximum-likelihood, any residual correlation violates the conditional independence. Under logit, normality of the factors and logistic conditional probability curves do not lead to a known form for the multivariate distribution of the latent response variables. In fact, this distribution has to be expressed by integrating over the factors. Also, a multivariate logistic distribution with free correlation coefficients does not exist. In this way, with maximum-likelihood estimation any residual correlation requires that the numerical integration approach is altered and expanded, for example by including additional minor factors that represent the residual correlations.