Comparison of computational methods for high dimensional item factor analysis

Tihomir Asparouhov and Bengt Muthén

November 9, 2012

Abstract

In this article we conduct a simulation study to compare several methods for estimating confirmatory and exploratory item factor analysis using the software programs Mplus and IRTPRO. When the number of factors is bigger than three or four the standard numerical integration methodology used for computing the maximum-likelihood estimates is intractable due to the exponentially large number of integration points needed to compute the likelihood. Several methods have been developed recently to overcome these computational problems however they have not been directly compared previously. In this paper we present a simulation study to compare maximum likelihood estimation based on Montecarlo integration, maximum likelihood estimation based on Metropolis-Hastings Robbins-Monro algorithm, maximum likelihood estimation based on two-tier integration, Bayesian estimation and the weighted least square estimation.

1 Introduction

Full information maximum likelihood estimation for factor analysis with categorical variables is a useful estimation method particularly in the presence of missing data. The most commonly used estimation method for exploratory and confirmatory factor analysis with categorical data is based on the weighted least squares methodology developed in Muthen (1984) and implemented in the Mplus package, Muthén and Muthén (1998-2010), among others. However this method is asymptotically consistent only when the missing data is missing completely at random (MCAR) but not if the missing data is missing at random (MAR), see Little and Rubin (1987) and Section 3 in Asparouhov and Muthén (2010b). On the other hand full information estimates are consistent even if the missing data is missing at random (MAR). Other reasons for preferring full information estimation over weighted least squares have also been cited in the literature such as efficiency gains as well as obtaining information criteria which are useful for example in comparing non-nested model. In addition the weighted least squares estimation is based on a multivariate polychoric/thetrachoric correlation matrix estimated from pairwise estimation which often is not a valid positive definite correlation matrix, i.e., the weighted least square estimation may actually be misrepresenting the data even when the model fits well the polychoric/thetrachoric correlation matrix because an unrestricted probit model may not fit the data well. In all of the simulations described in this article however we include the weighted least square (WLSMV estimator) estimation as implemented in Mplus.

The standard approach for obtaining the maximum-likelihood estimates involves the evaluation of multidimensional integrals using numerical integration, see Muthén and Asparouhov (2009), such as the Gauss - Hermite quadrature. With the current computing power this approach is generally limited to a maximum of there or four dimensions of integration which essentially corresponds to having at most three or four latent variables in the model. In many practical applications however where the multivariate data contains a large number of observed categorical variables it is necessary to include more than three of four factors.

A number of different approaches have been proposed in the literature and implemented in statistical packages to deal with this multi-dimensionality problem. In Muthén and Asparouhov (2009) the Montecarlo integration has been suggested and that method is implemented in Mplus for confirmatory factor analysis. In Asparouhov and Muthén (2009) the method has also been applied to exploratory factor analysis and that methodology is implemented in Mplus as well.

Another method for obtaining approximate full-information maximum likelihood estimates is the Bayesian estimation. Asymptotically the Bayesian posterior distribution is the same as the asymptotic distribution inferred by the ML method. When the sample size is sufficiently large the priors of the model parameters have no effect on the estimation, particularly when the priors are chosen to be some non-informative priors or weakly informative priors. Thus, one can use the mode/median or the mean of the posterior distribution obtained from the Bayesian estimation as an approximate ML estimates. Here the fact that these estimates are only approximate is not essential. As the likelihood can not be evaluated precisely by any method due to the fact that it has no explicit form, any estimation method should be considered only approximate. Since the Bayesian estimates are asymptotically the same as the ML estimates and they will have the same asymptotic advantages as the ML estimates. The Bayesian estimation is also implemented in Mplus. The estimation of confirmatory factor analysis with categorical data is described in Asparouhov and Muthén (2010a) and Asparouhov and Muthén (2010b). The Bayesian estimation of exploratory factor analysis implemented in Mplus is described in Section 3 below.

For some special confirmatory factor analysis models such as the bifactor model it is possible to reformulate the model in a special way so that a model with many factors requiring large multidimensional integration can actually be reformulated so that it requires only 2 or 3 dimensional integration. This method was pioneered in Gibbons and Hedeker (1992), and more recently has been generalized in Cai (2010a) and Cai et al. (2011a). The method is implemented in the software package IRTPRO, see Cai et alt. (2011b) and we will use that package to evaluate the performance of this method. The two-tier integration method can also be estimated in Mplus as a two-level multiple group model where the general factors are between level factors, while the specific factors are within level factors and the multiple groups represent the different blocks of variables that are correlated beyond the general factors. We will also use the two-tier integration method as implemented in Mplus for comparison. Prior to Mplus version 7 the two-tier estimation has to be setup as a two-level mixture model with observed class variable however in Version 7 this is no longer necessary and the two-tier estimation will be used even if the model is setup as a regular item factor analysis model, meaning that the program will automatically determine if the model is a bifactor-like model that allows for more optimal two-tier integration.

Finally in our simulation we will use the Metropolis-Hastings Robbins-Monro algorithm for confirmatory and exploratory factor analysis with categorical data as described in Cai (2010b) and Cai (2010c) and implemented in the IRTPRO package.

In Section 2 we describe the confirmatory and exploratory item factor analysis model as well as the bifactor model. In Section 3 we describe the Bayesian estimation of exploratory factor analysis model.

2 The item factor analysis model

In this section we describe the general item factor analysis model as well as the two special models we will use in the simulation studies: the bifactor model and the exploratory factor analysis model. Let U_{ij} be the $j-$th observed categorical variable for individual i, which takes values $0, ..., k_j$ for $j = 1, ..., P$ and $i = 1, ..., N$. Suppose that η is a vector of m latent variables with mean 0 and variance covariance matrix Ψ. The model is described by the following equation

$$P(U_{ij} = k) = \Phi(\tau_{kj} - \lambda_j \eta^T) - \Phi(\tau_{k-1,j} - \lambda_j \eta^T) \qquad (1)$$

where Φ is either the logit or probit distribution function, τ_{kj} are the threshold parameters that are estimated when $k > 0$ and $k < K_j$. When $k = 0$, $\tau_{kj} = -\infty$. When $k = K_j$, $\tau_{kj} = \infty$. The loading parameter vector λ_j is the j-th row of the loading matrix Λ.

In the above model there are m^2 unidentified parameters. The general EFA model is identified by fixing the parameters in Λ above the main diagonal to 0, i.e, $\lambda_{ij} = 0$ if $j > i$. This removes $m(m-1)/2$ parameters from the model and the remaining $m(m+1)/2$ parameter identifications come from fixing the Ψ matrix to the identity matrix. The above constraints specify the unrotated EFA model. The rotated EFA model is identified via a rotation criterion function which minimizes

$$f(\Lambda)$$

where

$$\Lambda = \Lambda_0 (H^T)^{-1}$$

over all orthogonal or oblique rotation matrices H and Λ_0 is the unrotated EFA loading matrix. Orthogonal rotations are those for which the matrix H is orthogonal, i.e., $HH^T = I$ where I is the identity matrix, while for oblique rotations HH^T has diagonal values of 1. The rotation criterion function f is generally designed to reduce the number of non-zero entries in the loading matrix, i.e., to simplify the patterns of the loadings. Various different rotation functions are used in practice, see Browne (2001).

The second special factor analysis model we are interested in is the bifactor confirmatory factor analysis model which in its simplest form is defined as having one general factor for which all loadings are estimated and all variables load on the general factor and at most one other factor. The variance covariance matrix for the factors is fixed to the identity matrix for identification purposes. For variations of the bifactor model see Cai et alt. (2011a).

3 Bayesian Estimation of Exploratory Factor Analysis

In this section we describe the Bayesian estimation of the EFA model as implemented in Mplus. Suppose that we are estimating an exploratory factor analysis model

$$Y = \Lambda\eta + \varepsilon$$

with p dependant variables and m factors. The first step in the Bayesian estimation is the estimation of the unrotated model as a CFA model using the MCMC method described in Asparouhov and Muthén (2010a) for general CFA models. The model is estimated until convergence. The second half of the generated MCMC sequence is used to form the posterior distribution of the unrotated EFA parameters. To obtain the posterior distribution of the rotated parameters we simply rotate the generated unrotated parameters in every MCMC iteration, using oblique or orthogonal rotation. Thus, in each MCMC iteration the rotation criteria is minimized to convert the unrotated values into the rotated. The rotated values from all MCMC iterations are then used to estimate the rotated posterior distribution as well as point estimates and standard errors for the rotated parameters.

The Baysian estimation of general latent variable models with categorical variables is based on the Gibbs sampler where underlying latent response variables are generated for every categorical variable, see Asparouhov and Muthén (2010a). After the latent response variables are generated the Gibbs sampler steps are the same as if the observed variables were continuous. Thus the above description of the Bayesian estimation of the EFA model applies also the EFA models with categorical variables.

This MCMC estimation is complicated by identification issues that are similar to label switching in the Bayesian estimation of Mixture models. There are two types of identification issues in the Bayes EFA estimation. The first type is identification issues related to the unrotated parameters. These issues are the same as the identification issues in a general Bayesian estimation of a CFA model. The second type of identification issues are those related to the rotated parameters.

3.1 Identification of the unrotated parameters

In a general CFA the signs of all loadings are generally not identified. A model with all factor loadings reversed has the same meaning and interpretation and fits the data equally well. Thus the posterior distribution of a loading parameter is always symmetric around 0. The multivariate posterior distribution of all loading parameters has 2^m symmetric modes because each of the signs of the m factors is unidentified. As the MCMC sequence generates values from this posterior distribution those values may or may not be in any of the 2^m modes. This depends on the sample size and the complexity of the model. The complex the model and the smaller the sample size is the more likely the more likely the MCMC sequence will jump from one of these modes to the other. The large the sample size is and the simpler the model is the sharper and more narrow the modes are and the more disconnected they are. This produces a probability of nearly 0 that a jump from one mode to another will occur. Nevertheless, as we estimate the loadings posterior distribution we need to make inference about the mean of the posterior, the median of the posterior and the variance of the posterior using only one of these 2^m. If we don't, then the mean and the median of every loading will be 0 and of the variance will be inflated. The only statistic that is not essentially distorted by the multiple modes is the mode of the posterior. Any one of the true and equal modes is equally good. As the MCMC sequence is generated we reflect all generated values into one of the symmetric 2^m modes and thus the generated values will form only one mode. To reflect the generated values however we need lines of symmetry. Multivariate Distributions are difficult to visualize and they may have many different lines of symmetry. Thus the actual definition is what really consists of $1/2^m-$th of the true posterior distribution is not necessarily defined in a unique way. Different lines of symmetry can be constructed. While these different lines of symmetries would generally not affect the mean and the median (the mode is again not affected at all) by much, posterior inferences that are affected by tail probabilities can be affected more substantially. In the Mplus implementation of general CFA model estimation we use the lines of symmetry defined by the following parameters constraints. For each factor j, $j = 1, ..., m$, for sign identification purposes we constraint the model parameters so that

$$\sum_{i=1}^{p} \lambda_{ij} > 0$$

where λ_{ij} is the loading value for Y_i on factor η_j. The above inequalities divide the full posterior distribution into symmetric and equivalent modes. If the MCMC sequence generates loading values that do not satisfy the above constraint the signs of all loadings are reversed, the constraints are thus satisfied, and the MCMC sequence continues as usual. This way the MCMC generated parameters populate only one of the 2^m symmetric modes. Any MCMC path visiting several of the modes of the posterior distribution is essentially mirrored into one of its symmetric components.

In MPlus convergence of the MCMC sequence is evaluated by comparing the generated posterior distribution across different MCMC chains. If the above sign identification is not performed then the different MCMC chains may converge towards different modes and the MCMC sequence may result in non-convergence or extremely slow convergence.

In the MCMC sequence the signs of the factors η are not changed or monitored. After the loadings are generated new factor estimates are generated and thus sign reversal for η is not needed.

The sign identification as described above is needed not just for the unrotated solution of the EFA model but also for a general CFA model or any general latent variable model.

3.2 Identification of the rotated parameters

The rotated loading parameters also have the sign unidentification problem that the unrotated parameters have. In addition to the sign unidentification the rotated parameters have unidentification due to the fact that factors can be permuted in any order to obtain an equivalent model. Typically the rotation criterion is symmetric with respect to the factors and thus the order of the factors or their signs is not identified by the rotation criterion. If such an order is not established the generated values from the MCMC sequence could be accumulated incorrectly and could be all mixed up. If the rotation criterion is symmetric with respect to all factors the posterior distribution of the rotated solution will have $2^m m!$ symmetric modes. Again we are only interested in obtaining one of the modes of this posterior distribution.

To resolve the sign and permutation unidentification we minimize and the following function

$$\sum_{i,j}(s_j \lambda_{i\sigma(j)} - L_{ij})^2 \tag{2}$$

over all possible factor permutations σ and all possible sign allocations $s_j =$

± 1, where λ_{ij} in the above formula is the rotated solution and L_{ij} is a target loading matrix which is obtained by averaging the rotated estimates over the proceeding MCMC iterations. When multiple MCMC chains are computed the same target matrix L is used for all chains.

In Mplus the optimization of (2) is conducted twice for all iterations. In stage one the target matrix L is being updated after each iteration so that it is the average of all rotated, permuted and sign adjusted solutions in the previous iterations. After all iterations are adjusted for permutation and sign allocations the final target matrix is constructed as the average across all iterations. In stage two using that final target matrix L as a constant, the optimization of (2) is performed again for all iterations. Using a constant target matrix is important to ensure that exactly one of the symmetric modes is constructed from the MCMC sequence. During the second stage typically only a few iterations will need adjustments because the final target matrix is generally close to the target matrices used in stage one.

4 Simulation study for the bifactor model

In this section we study the performance of the various methods for the bifactor model. We consider a model with 35 observed categorical variables each with 7 categories. A similar example was considered in Cai (2010a) where the observed variables represent a quality of life instrument. We generate 100 samples of size N=500 according to a bifactor model with 8 factors where all variables load on the first factor while each of the remaining specific factors have 5 variables each with non-zero loadings. Variables 1,..,5 load on the second factor, variables 6,...,10 load on the third factor etc. All loadings are set to 1 and we set the 6 thresholds values for all variables to be: $\pm 2.5, \pm 1.25, \pm 0.5$. We use the logit distribution function in the data generation.

Six methods are compared in this simulation and the results are presented in Tables 1 and 2. For all methods the default settings of convergence criteria and algorithmic options were used. The first method is the Mplus two tier method which uses two tier integration based on two dimensional numerical integration. The second method is the Mplus Monte 500 method which uses the Montecarlo integration method based on 500 integration points for the 8 dimensional integral evaluated in the log-likelihood. The third method is the same as the second but it uses 5000 integration points. The fourth method is

Table 1: Absolute bias, coverage and log-likelihood for the bifactor model.

Method	λ_{11}	λ_{12}	$\tau_{1,1}$	Log-Likelihood
Mplus Two Tier	.02(.95)	.02(.97)	.02(.97)	-31664.1
Mplus Monte 500	.03(.92)	.07(.92)	.02(.94)	-31760.8
Mplus Monte 5000	.03(.94)	.01(.95)	.01(.96)	-31678.5
Mplus Bayes	.04(.94)	.01(.95)	.00(.97)	-
Mplus WLSMV	.00(.96)	.02(.96)	.03(.95)	-
IRTPRO Two Tier	.02(.98)	.00(.96)	.01(.98)	-31680.7

the Mplus estimation of this model with the Bayesian methodology using all default uninformative or weakly informative priors and the posterior median as the point estimates. The Bayesian methodology in Mplus uses the probit link function thus we had to rescale / multiply all parameters by 1.7. The fifth method is the weighted least squares method implemented in Mplus using the WLSMV estimator. The sixth method is the two tier implementation in IRTPRO.

We present the results only for 3 parameters. The model is symmetric and thus all loadings can be permuted into the first two loading by relabeling the variables and the factors. Thus presenting the results only for those parameters is sufficient. Table 1 shows that all methods have negligible bias and all methods yield good coverage. The likelihood values estimated by the different methods are also nearly identical with the exception of the Montecarlo integration method with 500 integration points. This finding suggest that if a precise estimate of the log-likelihood is needed the number of integration points should be increased to 5000 or more. Table 2 shows that all standard error estimates are the same across the different methods and are approximately equal to the standard deviation of the parameter estimates across replications and therefore are correct. There is only one exception. The standard error obtained in IRTPRO for the general factor loadings are overestimated by 47%.

Table 2: Average standard error, ratio between average standard error and standard deviation for the bifactor model.

Method	λ_{11}	λ_{12}	$\tau_{1,1}$
Mplus Two Tier	0.12(0.97)	0.16(1.00)	0.19(1.10)
Mplus Monte 500	0.12(0.91)	0.15(0.99)	0.18(1.05)
Mplus Monte 5000	0.12(0.89)	0.15(0.94)	0.19(1.09)
Mplus Bayes	0.13(0.97)	0.16(1.01)	0.19(1.09)
Mplus WLSMV	0.13(0.97)	0.15(0.98)	0.18(1.08)
IRTPRO Two Tier	0.17(1.47)	0.16(1.01)	0.18(1.02)

5 Simulation study for the EFA model

In this section we study the performance of several estimation methods for the EFA model with categorical items. First we consider a model with 35 observed variables each with 7 categories. We generate 100 samples of size N=500 according to a model with 7 factors where 5 different variables load on each factor. Variable 1,..,5 load on the first factor, variables 6,...,10 load on the second factor, etc. The estimation of the parameters is typically done on a standardized scale. To define the standardized scale for the EFA model we provide an alternative parameterization for the item factor analysis model. Instead of defining the model through equation (1) we can define it using the underlying variable U^*_{ij} which is defined through the following equation

$$U_{ij} = k \Leftrightarrow \tau_{k-1,j} < U^*_{ij} < \tau_{kj}. \tag{3}$$

The item factor analysis model is then defined as a standard factor analysis model for the U^*_{ij} variables

$$U^*_{ij} = \lambda_j \eta^T + \varepsilon_{ij} \tag{4}$$

where ε_{ij} has a standard normal or logit distribution depending on the distribution function used in equation (1). The standardized parameters are defined as the standardized parameters of equation (4), i.e., the standardized loading parameters are $\lambda_j/\sqrt{Var(U^*_{ij})}$. Similarly the standardized threshold parameters are $\tau_{kj}/\sqrt{Var(U^*_{ij})}$. Using a standardized metric is helpful when

Table 3: Absolute bias, coverage and log-likelihood for EFA model with ordered polytomous variables.

Method	λ_{11}	λ_{12}	Log-Likelihood
Mplus Monte 500	.01(0.97)	.00(0.83)	-28580.3
Mplus Monte 5000	.01(0.96)	.00(0.87)	-28578.4
Mplus Bayes	.01(.90)	.00(.96)	-
Mplus WLSMV	.00(.94)	.00(.89)	-
IRTPRO MHRM	.00(.54)	.00(.65)	-28665.2

models and parameters based on different link functions are compared. For the remaining part of this section all parameters are on standardized scale.

To generate the data we set all non-zero loadings in the EFA model to 0.7 and the 6 thresholds values for all variables are set to $\pm 1.5, \pm 1, \pm 0.4$. We use the probit link function to generate the data.

Five methods are compared in this simulation and the results are presented in Tables 3 and 4. The first method is the Mplus Monte 500 method which uses the Montecarlo integration method based on 500 integration points for the 7 dimensional integral evaluated in the log-likelihood. The second method is the same as the first but it uses 5000 integration points. The third method is the Bayesian estimation implemented in Mplus for the EFA model. The fourth method is the weighted least squares method implemented in Mplus using the WLSMV estimator. The fifth method is the Metropolis-Hastings Robbins-Monro algorithm implemented in IRTPRO. Again we report the results just for the first two loadings. Due to model symmetry any other loading parameter is equivalent to one of these two parameters. All the methods perform well in terms of bias, however the Metropolis-Hastings Robbins-Monro method underestimates the standard errors and the coverage drops down to 54%. The Montecarlo integration method using 500 integration points also has a slight underestimation of the standard error for the second loading. This suggest that increasing the number of integration points from 500 to 5000 can lead to significant improvement in the precision of the standard errors.

Next we describe a simulation study for the EFA model with binary items. The simulation setup is the same as for the EFA model with ordered polytomous variables. To generate binary variables one threshold value is used in

Table 4: Average standard error, ratio between average standard error and standard deviation for the EFA model with ordered polytomous variables.

Method	λ_{11}	λ_{12}
Mplus Monte 500	0.033(1.00)	0.031(0.72)
Mplus Monte 5000	0.033(0.99)	0.035(0.81)
Mplus Bayes	0.030(0.97)	0.032(0.98)
Mplus WLSMV	0.030(0.97)	0.038(0.85)
IRTPRO MHRM	0.012(0.42)	0.026(0.65)

Table 5: Absolute bias, coverage and log-likelihood for EFA model with binary variables.

Method	λ_{11}	λ_{12}	Log-Likelihood
Mplus Monte 500	.02(0.97)	.00(0.82)	-10759.4
Mplus Monte 5000	.02(0.97)	.00(0.89)	-10753.8
Mplus Bayes	.00(.96)	.00(.97)	-
Mplus WLSMV	.00(.95)	.00(.92)	-
IRTPRO MHRM	.01(.42)	.01(.72)	-10763.5

equation (3). The threshold value is set to 0 for all variables. The results for this simulation study are presented in Tables 5 and 6 and the findings are the same as for the ordered polytomous case. All the methods perform well in terms of bias and coverage with the exception of the Metropolis-Hastings Robbins-Monro method which underestimates the standard errors and the coverage drops down to 42%.

Table 6: Average standard error, ratio between average standard error and standard deviation for the EFA model with binary variables.

Method	λ_{11}	λ_{12}
Mplus Monte 500	0.053(1.05)	0.045(0.75)
Mplus Monte 5000	0.053(1.04)	0.051(0.85)
Mplus Bayes	0.051(1.06)	0.044(1.13)
Mplus WLSMV	0.048(1.04)	0.054(0.88)
IRTPRO MHRM	0.014(0.33)	0.035(0.53)

6 Simulation study for the EFA bifactor model

The bifactor model described in Section 4 can be analyzed also as an EFA model, i.e., the bifactor model can be estimated even when the loading pattern matrix is unknown. Jennrich and Bentler (2011) proposed a rotation criterion which is designed to uncover bifactor loading matrices. Here we conduct a simulation study to evaluate the performance of this method for categorical data. We consider a model with 35 observed binary variables and 8 factors. All variables load on the first factor while each of the remaining specific factors have 5 different variables with non-zero loadings. Variables 1,..,5 load on the second factor, variables 6,...,10 load on the third factor, etc. All of the general factor loadings are set to 2 and the specific factor loadings are set to 1. The residual variance is set to 1. The threshold values for all variables is set to 0. The probit link function is used to generate 100 samples of size N=1000. The model is estimated using the WLSMV method, and the full information Monetcarlo integration method with 500 and 5000 integration points as well as the Bayesian estimation method. The estimation is based on the bi-factor Quartimax rotation criterion

$$f(\Lambda) = \sum_{i=1}^{p} \sum_{j=2}^{m} \sum_{l \neq j, l > 1}^{m} \lambda_{ij}^2 \lambda_{il}^2. \tag{5}$$

The bi-factor rotation functions are generally susceptible to multiple local optima. Thus it is important to use randomized starting values during the rotation optimization. In this simulation 100 random starting values are used.

Table 7: Absolute bias and coverage for bifactor EFA model with binary variables.

Method	λ_{11}	λ_{12}	λ_{13}
Mplus WLSMV	.01(.86)	.02(.88)	01(.84)
Mplus Monte 500	.02(.59)	.02(.45)	01(.58)
Mplus Monte 5000	.00(.90)	.01(.88)	01(.77)
Mplus Bayes	.01(.96)	.01(.89)	00(.99)

Table 7 shows the absolute bias and coverage for three of the loadings. The first loading λ_{11} is the loading for the general factor. The second loading λ_{12} is the loading for the specific factor. The third loading λ_{13} is loading for a different specific factor, i.e., its true value is 0. All other loadings are equivalent to those by symmetry. The results are presented in standardized scale as EFA models are typically presented. Thus the true loading value for the general loadings is 0.816 and the true value for the specific loading is 0.408. The results suggest that all estimators are unbiased. The standard errors are approximated well with the exception of the Montecarlo integration method with 500 integration points. Increasing the number of integration points to 5000 appears to improve substantially the accuracy of the standard errors.

7 Model selection testing

In this section we evaluate several different methods for model selection and testing for high dimensional structural equation models with categorical data. In this simulation study we consider a two-group CFA model with 25 observed binary variables and five factors in each group. Each factor is measured by 5 different variables and there are no cross loadings. We generate 100 samples of size N=500 with group membership being equally likely for each observation. All loadings are set to 1 and we set the threshold values to 0. We use the logit distribution function in the data generation. The factor covariance matrices are estimated as free parameters and for data generation purposes all factor covariances are set to 0. In the first group the factor variances are set to 1 and in the second group the factor variances are set to 1.2. We estimated two models. The unrestricted H1 model estimates a model where all threshold values are free and unequal across the two groups and all loadings are unequal and free across the two groups. The factor variances in each group are fixed to 1 and the covariances are estimated as free parameters. The unrestricted model has 25 loading parameters in each group and 25 threshold parameters as well as 10 covariance parameters. In total the unrestricted model has 120 parameters. One typical test for multiple group analysis in structural equation models is the test of measurement invariance. To test measurement invariance we want to know if the loadings are equal across the two groups and any differences between the two groups are due to factor distributions differences. It is of interest to know if the actual measurement model is identical between the two groups and the factors can be interpreted the same way. There are multiple ways such a test can be conducted. One way is to estimate the restricted model H0 which has all loadings equal across the two groups, factor variances fixed to 1 in the first group and free in the second group. Factor covariances free and unequal across the two groups as well as the threshold parameters. This restricted model has 25 parameters and 10 covariances in each group, 25 loading parameters and 5 variance parameters in the second group for a total of 100 parameters. The data is generated so that both the unrestricted H1 and the restricted model H0 are both true. The two models are also nested in each other. Thus if we have a way of computing the log-likelihood value L at the maximum-likelihood estimates for the two models we can simply conduct a likelihood ratio test (LRT) using the statistic

$$T = 2(L_{H_1} - L_{H_1}).$$

When the null hypothesis of measurement invariance is true the statistic T is distributed as a chi-square distribution with 20 degrees of freedom. $\chi(20)$ and if the T statistic is higher than the 95% percentile of this distribution the null hypothesis is rejected. Since the log-likelihood cannot be computed explicitly and all likelihood based methods provide only an approximate estimate of the log-likelihood it is unclear if such an approximation is sufficiently accurate for the purpose of conducting the LRT. The first three methods that we evaluate in our simulation study are the LRT test based on the Montecarlo integration method as implemented in Mplus using 500 and 5000 integration points as well as the MHRM method as implemented in IRTPRO. Since both models are correct the null hypothesis should not be rejected and across the 100 replication we expected the LRT rejection rate near the nominal level of

5%. The average test statistic value should be near the degrees of freedom value which is 20 in this case.

The Bayesian information criteria (BIC) can also be used for model selection. This criteria also depends on the log-likelihood value L

$$BIC = -2L + k\log(n)$$

where n is the sample size and k is the number of model parameters. Using this criterion the model with the smallest BIC is selected. Asymptotically BIC will select the simplest correct model. In our simulation this is the H0 model since it has fewer number of parameters. Using BIC we expect to have 0% rejection rate of the H0 model if the log-likelihood estimate for L is sufficiently accurate. We evaluate the performance of BIC obtained by the Montecarlo method with 500 and 5000 integration points as well as the MHRM method.

Another method for testing measurement invariance is the Wald test. This method does not rely on the log-likelihood value but uses the asymptotic variance covariance of the parameter estimates to evaluate the statistical significance of parameter constraints. The measurement invariance test in our example can be evaluated using the Wald test for 20 equations evaluated for the parameters in the H1 model. These equations imply the proportionality of the loadings in the H1 model. If λ_{ijg} is the loading estimate of factor j on variable i in group g in the H1 model the first four equations can be written as follows

$$0 = \lambda_{111} - \lambda_{112}\frac{\lambda_{211}}{\lambda_{212}} \tag{6}$$

$$0 = \lambda_{111} - \lambda_{112}\frac{\lambda_{311}}{\lambda_{312}} \tag{7}$$

$$0 = \lambda_{111} - \lambda_{112}\frac{\lambda_{411}}{\lambda_{412}} \tag{8}$$

$$0 = \lambda_{111} - \lambda_{112}\frac{\lambda_{511}}{\lambda_{512}} \tag{9}$$

The remaining 16 equations are written similarly and imply proportionality of the loadings for each factor. We evaluate the performance of the Wald test using the Montecarlo estimation method using 500 and 5000 integration points as well as the WLSMV estimator in Mplus. Again the rejection rate for the Wald test should be near the nominal level of 5%. The average test statistic value should be near the degrees of freedom value which is again 20.

Table 8: Rejection rates for testing measurement invariance

Method	Rejection Rate	Average Test Statistic
LRT Mplus Monte 500	8%	20.7
LRT Mplus Monte 5000	8%	21.3
LRT IRTPRO MHRM	28%	28.2
BIC Mplus Monte 500	0%	-
BIC Mplus Monte 5000	0%	-
BIC IRTPRO MHRM	10%	-
Wald Mplus Monte 500	2%	16.6
Wald Mplus Monte 5000	2%	17.2
Wald Mplus WLSMV	6%	19.9
Difftest Mplus WLSMV	8%	22.3

Finally, the last method we evaluate in this simulation study is the difference test associated with the WLSMV estimator which uses the difference between the fit function values for the H1 and H0 model. The difference is adjusted so that test statistic approximates a chi-square distribution with 20 degrees of freedom, see Asparouhov and Muthén (2006) and Asparouhov and Muthén (2010c). Again the rejection rate for the difference test should be near the nominal level of 5%. The average test statistic value should be near 20.

The results of the simulation study are presented in Table 7. All the methods perform well with the exception of the methods based on the MHRM log-likelihood value. The LRT method based on MHRM leads to an inflated test statistic and 28% rejection rate. Using the the BIC value with MHRM leads to 10% rejection rate. Both of these indicate that the log-likelihood estimate is not as reliable with this method. On the other hand both the LRT and the BIC method imply that the Montecarlo integration method works well and there is no gain in precision by increasing the number of integration points. The Wald test with all estimators and the difference test work well in this example as well.

The Wald test as presented here depends on the asymptotic normality of the equations that are being evaluated. These equations are non-linear and even if all parameter estimates are normally distributed for finite sample size the equations that are being tested may not be normally distributed. This may lead to substantial deviation from the expected chi-square distribution and consequently inflated or deflated rejection rates. Deflated rejection rates are not a problem if the null hypothesis is correct but when the hypothesis is not-correct the deflation in the rejection rate will result in a loss of power. Thus it is important to specify the Wald test equations to as close to linear as possible. There are many equivalent ways to specify the Wald test equations and in general it is hard to decide which set of equations is the closest to linear equation. Different specifications for the Wald test will lead to slightly different results. Asymptotically, i.e., if the sample size is sufficiently large the Wald test would be independent of the actual specification. However for finite sample size that will not be true and linearity in the specification is important. To test the measurement invariance in our simulation study the specification provided above is optimal. To illustrate the dependence of the Wald test in finite sample size on the test equation specification consider the following set of equations

$$0 = \lambda_{111} * \lambda_{212} - \lambda_{112}\lambda_{211} \tag{10}$$

$$0 = \lambda_{111} * \lambda_{312} - \lambda_{112}\lambda_{311} \tag{11}$$

$$0 = \lambda_{111} * \lambda_{412} - \lambda_{112}\lambda_{411} \tag{12}$$

$$0 = \lambda_{111} * \lambda_{512} - \lambda_{112}\lambda_{511} \tag{13}$$

Equations (10-13) are equivalent to equations (6-9). These equations directly involve product of parameters and can be expected to provide worse chi-square fit. We conduct a simulation study to illustrate this issue using the WLSMV estimator and data sets of varying sample size. The results are presented in Table 8. For the same sample size of 500, the Wald test based on equations (10-13) produced an average test statistic of 16.7 and a 0% rejection rate, while the Wald test based on equations (6-9) produced an average test statistic of 19.9 and a 6% rejection rate. Thus the results confirm that in finite sample size the specification (6-9) performs better, while the alternative specification (10-13) shows deflated rejection rates for smaller sample sizes. For sample size of 2000 or more the alternative parameterization also works well.

Table 9: Rejection rates for alternative Wald test using WLSMV estimator

Sample size	Rejection Rate	Average Test Statistic
500	0%	16.7
1000	0%	19.0
2000	2%	19.9
5000	4%	19.7

8 Comparing Computational Time

Computational times are important when comparing different algorithms. However, such comparison should not proceed the comparisons of the quality of the estimators. None of the methods described in this paper are computationally challenging. Comparison between algorithms that are dramatically different is much easier than comparison that are very similar. For example, the computational time for obtaining the ML estimates based on numerical integration using quadrature grows exponentially with the increase of the number of factors while the Montecarlo integration does not depend on the number of factors at all. The numerical error in the Monetecarlo integration depends on the number of integration points Q and is proportional to $1/\sqrt{Q}$ but is independent of the number of factors. Thus it is easy to illustrate and understand the fundamental principle behind the fact that Montecarlo integration is better for high dimensional model estimation than regular quadrature integration. However when methods that are quite similar such as the Montecarlo integration, RMRM method, and the full Baysian estimation method any timing comparison should be taken lightly because there is no fundamental principle that yield a clear advantage of one method over another. These methods are similar because they all use stochastic approximations to obtain the ML estimates.

The algorithms in all cases are quite complex. It is not easy to evaluate the error in the ML parameter estimates that is due to stochastic approximation. Most stochastic methods have a very tight relationship between the error in the stochastic approximation and the amount of computing involved. The larger the number of integration point in the Montecarlo integration method the bigger the computational time and the smaller the error. Comparing the computational times is not informative unless we are able to ensure that the

Table 10: Computational times in seconds

Method	Example 1	Example 2	Example 3	Example 4
Mplus Monte 500	146	62	17	37
Mplus Monte 5000	220	456	122	235
Mplus WLSMV	1	6	2	1
Mplus Bayes	1409	119	632	-
Mplus Two-Tier	-	95	-	-
IRTPRO MHRM	1601	-	21	15
IRTPRO Two-Tier	-	350	-	-

methods are producing a comparable stochastic error and that is not possible due to the complexities of the algorithms. Thus any time comparison on these methods have no tangible implications. Time comparison is also complicated by the fact that they also depend on the convergence criterions of the maximization algorithms and the various technical options that are specific for each method. Nevertheless we present some computational times as an illustration. These results may not replicate for other models or data. In Cai (2010b), computational times are presented for similar models and algorithms, and the results presented there are quite different from the computational times we observed. This again points out that general comparison between algorithms and software implementations is generally unreliable and may not replicate for different models and data.

The timing results are presented in Table 9 for the various estimation methods and four examples. Example 1 is a real data example that estimates a 4 factor orthogonal EFA model for 17 antisocial behavior binary items from the National Longitudinal Survey of Youth. Example 2 is the simulated bifactor example. Example 3 is the simulated EFA model with binary variables. Example 4 is the simulated 2 group example estimated under the measurement invariance. The WLSMV estimator is always the fastest estimator by a large margin. Among the full information estimators for different examples show different results. None of the estimators use a prohibitive amount of time and are feasible is these settings.

References

[1] Asparouhov, T. and Muthén, B. (2006). Robust Chi Square Difference Testing with Mean and Variance Adjusted Test Statistics. Mplus Web Notes: No. 10. May 26, 2006. http://statmodel.com/download/webnotes/webnote10.pdf

[2] Asparouhov, T. and Muthén, B. (2009). Exploratory structural equation modeling. Structural Equation Modeling, 16, 397-438.

[3] Asparouhov, T. and Muthén, B. (2010a). Bayesian analysis using Mplus: Technical implementation. Technical Report. http://statmodel.com/download/Bayes3.pdf

[4] Asparouhov T. and Muthén B. (2010b) Bayesian Analysis of Latent Variable Models using Mplus. Technical Report. http://statmodel.com/download/BayesAdvantages18.pdf

[5] Asparouhov, T. and Muthén, B. (2010c). Simple Second Order Chi-Square Correction. Mplus technical report. http://statmodel.com/download/WLSMV_new_chi21.pdf

[6] Browne, M.W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research* **36**, 111-150.

[7] Cai, L. (2010a). A Two-Tier Full-Information Item Factor Analysis Model with Applications. Psychometrika 75, 581612.

[8] Cai, L. (2010b). High-dimensional exploratory item factor analysis by a Metropolis- Hastings Robbins-Monro algorithm. Psychometrika, 75, 33-57.

[9] Cai, L. (2010c). Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. Journal of Educational and Behavioral Statistics, 35, 307335.

[10] Cai, L., Yang, J. S., and Hansen, M. (2011a) Generalized Full-Information Item Bifactor Analysis. Psychological Methods, 16, 221248.

[11] Cai, L., du Toit, S. H. C., & Thissen, D. (2011b). IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling [Computer software]. Chicago: SSI International.

[12] Gibbons, R.D., and Hedeker, D. (1992). Full-information item bifactor analysis. Psychometrika, 57, 423436.

[13] Jennrich R.I. and Bentler P.M. (2011) Exploratory bi-factor analysis. Psychometrika, 76, 1-13.

[14] Little, R.J.A. and Rubin, D.B. (1987) Statistical Analysis with Missing Data. J. Wiley & Sons, New York.

[15] Muthén, B. & Asparouhov, T. (2009). Growth mixture modeling: Analysis with non-Gaussian random effects. In Fitzmaurice, G., Davidian, M., Verbeke, G. & Molenberghs, G. (eds.), Longitudinal Data Analysis, pp. 143-165. Boca Raton: Chapman & Hall/CRC Press.

[16] Muthén, L.K. and Muthén, B.O. (1998-2010). Mplus Users Guide. Sixth Edition. Los Angeles, CA: Muthén & Muthén

[17] Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. Psychometrika, 49, 115-132.