Rejoinder: Mastering a new method

Bengt Muthén & Tihomir Asparouhov

April 21, 2012

Abstract

This rejoinder discusses the general comments on how to use BSEM wisely and how to get more people better trained in using Bayesian methods. Responses to specific comments cover how to handle sign switching, non-convergence and non-identification, and prior choices in latent variable models. Two new applications are included. The first one revisits the Kaplan science model by considering priors on primary parameters. The second one applies BSEM to the bi-factor model that was hypothesized in the original Holzinger-Swineford study.

Keywords: Statistics training, sign switching, non-identification, bi-factor model.

Introduction

We are pleased about the generally positive reactions in the commentaries. There is agreement about the value of applying Bayesian priors to achieve more flexible modeling. There is also agreement about the potential of model modification using BSEM results and that studies are needed to compare this to ML modification indices. Posterior predictive checking will also benefit from further study as will modeling with small-variance priors for residual correlations. Following are responses to comments about possible misuses of BSEM and the need for training, as well as responses to some particular technical points made in the commentaries. The rejoinder ends with two new applications as further motivation to learn the method.

Potential misuse and analysis choices

Bayes and BSEM, as all statistical methods, can be misused in many ways and it is our hope that our article will stimulate thoughtful rather than careless applications, including those that merely jump on an opportunity to get published by using a new method. New methods often suffer from such faddish problems as seen again and again in the history of methodology; see e.g. Muthén (1989) for an early discussion of poor applications of SEM. To help temper the tendency to misuse, following are two examples of what not to do. Consider the context of informative priors for non-zero parameters in a mediation model that is just-identified and cannot be rejected by data. There is a risk that a researcher favors a prior that provides a mediated effect that is significant (credibility interval not covering zero), arguing that this prior reflects his/her well-reasoned theories about the mediation

mechanism. With a strong enough prior, the influence from the data is overwhelmed and the posterior estimate is closer to the prior than the likelihood- (data-) based ML estimate. This suggests that the argument for using a strong prior must be based on relevant previous empirical evidence in a series of studies, not merely subjective notions. A second example refers to the replication of studies. Suppose a researcher conducts a small-sample study viewed as a continuation of a line of several similar previous studies, but perhaps with a slightly better design. The evidence from the previous studies can be given as a strong prior if the parameter in question obtains an estimate with little variation across the studies, which may happen if the sample sizes are large and the studies designed and carefully carried out in the same way. In this case, the data from the new study are given almost no weight because the prior has a stronger influence than the new data. The risk is that when the posterior estimate from the Bayes analysis of the new study is very different from the likelihood-based ML point estimate, the researcher carelessly chooses the Bayes estimate. With a strong prior, it is crucial to consider whether the new study is sufficiently close in design and execution to previous studies that generate the prior. In fact, collecting data from a small sample may be meaningless if new data are not given a chance to contradict previous studies.

While the above examples pertain to Bayes in general, the MacCallum et al. commentary raises a question specific to BSEM concerning the use of informative priors for almost zero parameters in the eight analyses of the Kaplan model:

> "might it be appropriate in general to always run a model of this nature, where all fixed zeros are converted to parameters

assigned small-variance priors?"

To clarify, the zeros for the conventional SEM pertain to cross-loadings and residual correlations in the measurement models, as well as direct effects. The answer to this question depends on the stage of evidence-building that the particular research area has reached. In early research stages where the focus is on exploration, priors should be non-informative or only mildly informative, that is have large variances, letting the data be the essential influence on the estimation. Here, ML and Bayes estimates are approximately the same. With respect to the measurement part of the model, imposing as little structure as possible can be done via EFA, or ESEM in SEM settings, where ESEM can be generalized to BESEM (Bayesian ESEM) to not leave the Bayes framework. As empirical evidence is accumulated via new studies, more informative priors in line with the zero-mean, small-variance priors of BSEM can be introduced. At the far end of this spectrum, the prior variances move toward exactly zero, resulting in classic CFA and SEM. It is this Bayesian flexibility that makes it attractive for the theory investigation that SEM focuses on. The strength of using zero-mean, small-variance priors is that they let data suggest modifications, where parameters can be freed if that is indicated.

In this connection, a warning is warranted. The fact that the BSEM model may fit well when such parameter estimates substantially deviate from zero may invite misuse of the method. These deviations may mask important model misspecifications, such as the need for a different number of factors or different structural relations. It should be stressed that if significant and substantively large deviations from zero are observed, it is important to investigate the reasons for this.

Related to this, MacCallum et al. state that

> "there is a tradeoff inherent in BSEM relative to conventional likelihood-based SEM in that implementation of BSEM in practice may be subject to a variety of difficulties and uncertainties. The Bayesian approach to SEM introduces complexities in model specification and estimation that may not be familiar to users of likelihood-based SEM;"

We think the potential difficulties are worth the effort and the effort will become less and less as both more Bayes training gets underway and software implementation improves by benefitting from more and more practical experience in the SEM area. Similar difficulties arose when moving researchers to SEM, with seemingly huge technical obstacles when coming from the tradition of regression-based path analysis with EFA factor scores. Proper training is needed but feasible. Note that the inherent tradeoff does not imply that BSEM is more subjective than classic, frequentist SEM as a reader might infer from the commentaries. On the contrary, the use of informative priors in Bayes makes explicit and quantifies the subjectivity that can remain hidden in classic SEM. Formal guidelines should and no doubt will be developed for how to use Bayes wisely with SEM. Some of them will be built into software to make applications easier.

Training

Given the above comments, it is clear that we agree with the reviewers about the importance of providing sufficient training to understand how to use Bayes and BSEM wisely. To master any new method takes training

and getting many learners up to speed takes time. How to best teach Bayes methodology is a good question. The topic was recently debated on the internet in the Epidemiology section of the American Statistical Association (for excerpts, please contact the first author). Epidemiology is akin to Psychology in the need to learn about Bayes without being a statistics expert and several arguments were made for better inclusion of Bayes in basic training. Our view is that low-level statistics training in Psychology graduate programs should have at least 10% devoted to Bayes, medium-level statistics courses should have at least 25% devoted to Bayes, and for departments with statistics emphases there should be upper-level courses devoted to Bayes. Muthén (1989) stresses the importance of methods-specializing "bridgers", students who do not specialize in the substantive area of Psychology and are not aiming to be Ph.D. statisticians, but are trained to be able to reach into both realms. Bridgers are important for making Bayes become effectively used in behavioral sciences. They are needed to help with the analyses of those who instead choose to specialize in substantive areas and therefore will not have time to go as deep in their methods studies. In regard to methods training, it becomes clear that behavioral science departments who do not admit student who want to specialize in methodology are making a mistake.

It should be added that university courses are not the only avenue for methods training. The abundance of short courses and workshops also provide ways to more quickly move the field forward. In addition, there are currently several applied articles, overview chapters, and books in the pipe line for publication, giving practical advice for how to benefit from Bayes and BSEM, including how to use them in Mplus. Not long ago, rather

few members of Psychology departments in the U.S. could be described as competent Bayes users. Our feeling is that the rate of increase is now very high.

Response to MacCallum, Edwards, and Cai

We cover three major technical issues in the MacCallum et al. commentary: Sign switching of loadings, non-convergence/non-identification, and dependence on priors. It is shown that all of these potential difficulties can be addressed. Our general aim is to raise awareness of some of the important issues and to suggest practical ways to handle them. A non-technical approach is taken, where our points are mostly made using graphs.

Sign switching in latent variable models

MacCallum et al. points out that when MCMC is applied to latent variable models, the signs of factor loadings for a given factor may switch over the MCMC iterations. With maximum-likelihood estimation it is known that sign switching corresponds to an ignorable reversal of the direction of the factor. With MCMC, however, it can lead to non-convergence or falsely assuming convergence while obtaining meaningless results that average positive and negative loadings. It is unclear how common this phenomenon is for latent variable models. In our experience, it is not common for models with large loadings and small standard errors. We have seen this for complex models with small loadings and large standard errors like exploratory factor analysis and bi-factor models. It may also occur for

multi-trait, multi-method models. Although sign switching can occur, it can be detected in the graphs produced in Mplus Bayesian analysis. To avoid sign switching in factor models, we introduce a relabeling algorithm which, for each MCMC iteration and for each factor, computes the sum of all loadings and changes the signs of all loadings if this sum is negative (see Asparouhov and Muthén, 2012). This induces a reversal of the direction of the factor.

An example that illustrates the problem and its solution uses a bi-factor model for the Holzinger-Swineford Grant-White data described in our paper. In this example, the set of 19 variables is augmented by adding five tests that are intended to measure the general factor (deduction, numerical puzzles, problem reasoning, series completion, arithmetic problems). All 24 variables are specified to measure a general factor. As before, the first 19 variables also measure the four specific factors of spatial, verbal, speed, and memory in line with the analyses of our paper, whereas the last five variables measure only the general factor.

Figure 1 shows the trace for the loading of the variable visual on the spatial factor using 50,000 MCMC iterations and two chains corresponding to the blue and red colors. The iterations to the right of the red vertical line represent the posterior distribution of the loading estimate. It is seen that each chain's iterations oscillate between positive and negative loading values of similar absolute value. The posterior distribution is bi-modal as shown in Figure 2 with modes at ± 0.25. The two modes are almost equally high and the average value of the posterior distribution is therefore close to zero, which is a misleading result. The 95% credibility interval cannot be used because it accepts the posterior as bimodal, whereas only one of

the modes is of interest. All four factor loadings for the spatial factor show these trace and posterior problems. The problem does not occur for other loadings in the model presumably because those loadings are larger and have smaller standard errors. The differences in factor loading size and variability are confirmed by maximum-likelihood analysis.

[Figure 1 about here.]

[Figure 2 about here.]

Now consider the results using the relabeling algorithm. For the same factor loading as above, Figure 3 shows the trace and Figure 4 the posterior distribution when avoiding the sign switching. The trace is now showing a desirable convergence picture and the posterior is unimodal.

[Figure 3 about here.]

[Figure 4 about here.]

The sign switching relabeling approach also makes it possible to carry out exploratory factor analysis using Bayes. In exploratory factor analysis, modeling with a high number of factors may result in relatively small factor loadings with large standard errors for some factors. This necessitates relabeling when using Bayesian analysis to obtain the unrotated loadings. Bayesian exploratory factor analysis is studied in Asparouhov and Muthén (2012).

Non-convergence and non-identification in latent variable models

McCallum et al. points out that the determination of convergence of MCMC is more involved than in, for example, maximum-likelihood estimation. Mac Callum et al. also points to the fact that model identification is more difficult to determine with MCMC than with maximum-likelihood estimation. Both of these issues are potential problems with MCMC. However, we believe that there are practical solutions some of which we describe below. Our paper discusses the Gelman-Rubin Potential Scale Reduction (PSR) approach and notes in the Appendix that this needs to be augmented by a long run because the PSR criterion is sometimes fulfilled prematurely. A further check of convergence is discussed here, which also addresses the non-identification problem.

Consider as an example a one-factor model that is not identified in maximum-likelihood terms due to attempting to estimate factor indicator intercepts as well as a factor mean. The indeterminacy of the non-identification can be shown as follows for indicator y_j.

$$E(y_j) = \nu_j + \lambda_j\ \alpha = \nu_j^* + \lambda_j\ \alpha^*, \tag{1}$$

$$\alpha^* = \alpha + c, \tag{2}$$

$$\nu_j^* = \nu_j - \lambda_j\ c, \tag{3}$$

for any value of c. This means that there is an infinite number of parameter values that create the same $E(y_j)$, so that the intercepts ν_j and the factor mean α are not identified. The factor loadings and the residual variances of the factor indicators are, however, identified.

As a real-data example, consider attempting to estimate the factor mean in a one-factor model for the four spatial factor items of the Holzinger-Swineford Grant-White data. Figure 5 shows the trace for 50,000 MCMC iterations and two chains, clearly indicating a convergence problem in that the two chains do not overlap. Using too few iterations, however, results in a premature decision of convergence using PSR. This occurs around iteration 1500 and again at iteration 16,000.

[Figure 5 about here.]

A Monte Carlo simulation study in Mplus is used to study the performance of PSR in this situation. Four variables measure one factor with a sample size of $n = 500$. Using 10 replications, none of the replications converges by PSR at 100,000 MCMC iterations using two chains. In all replications, however, the PSR convergence criterion is fulfilled at several earlier points during the iterations, illustrating the need for a longer run.

To avoid the premature convergence determination by PSR, the PSR criterion can be augmented with a Kolmogorov-Smirnov test of equality of the posterior distributions generated by the different chains. In the above Monte Carlo replications, in all cases that PSR prematurely signals convergence, the Kolmogorov-Smirnov test rejects equality of the two posterior distributions and therefore signals that more iterations are needed. The rejections appear for the 10 intercepts and the factor mean, thereby suggesting an indeterminacy among these parameters in line with (2) - (3).

In other models that are not identified in maximum-likelihood terms, the non-identification can be seen by exploding parameter values involved in the indeterminacies. This in turn leads to convergence failure due to a

non-positive definite posterior covariance matrix. The convergence failure protects against using a non-identified model.

It should be emphasized that the above discussion of non-identification refers to parameters for which improper priors and improper posteriors are considered. For example, mean and intercept parameters in the above example have normal priors with mean zero and infinite variance. In this case, the non-identified parameters have unlimited range because neither the prior or the data contribute information to the parameters. It is for this case that identification problems can be detected by Kolmogorov-Smirnov tests and/or non-positive definite posterior covariance matrix failure. With proper priors and posteriors, however, the Bayesian analysis always results in an identified model in that the priors contribute the necessary information when the data do not. This is discussed in Garrett and Zeger (2000), which gives further references. Garrett and Zeger (2000) consider comparisons of posterior with priors to judge how much information the data provide about parameters. This might be an avenue for detecting model non-identification, although our investigation of this has not been fruitful, perhaps due to indeterminacies involving complex functions of several parameters.

Variance priors and induced priors

MacCallum et al. discusses induced priors in connection with their Figure 1. Their word induced may be given different meanings and we want to discuss their comments to avoid misunderstandings by readers. Their discussion may give the impression that the total estimated variance in a Bayesian analysis is a fixed value that matches the observed sample variance.

Instead the total variance is an estimated parameter that has a posterior distribution, and where the mean, median, or mode may not agree with the sample variance. Likewise, their discussion might give the impression that priors on the factor loadings induce possibly inappropriate priors on the residual variances, which is not the case. The prior for the unique variance is not induced when the word is interpreted as deterministically related, but is separately specified. Information in the data, in this case the total sample variance, does not influence the prior. Only the posterior distribution can be influenced by the data.

The example and prior presented in Figure 1 of MacCallum et al. is not relevant to the typical BSEM application because each factor indicator would have at least one loading which is not presumed approximately zero. Instead, consider a two-factor model for the Holzinger-Swineford tests for the spatial and verbal factors. We specify a non-restrictive prior for the loading of the visual test on the spatial factor using a normal prior with mean zero and variance 10. A cross-loading is allowed for the visual test on the verbal factor using a normal prior with mean zero and variance 0.01. For the residual variance, we specify the inverse-Gamma prior IG(1,1). We consider the prior of the corresponding standardized residual variance, which is shown in Figure 6. There is nothing undesirable about this prior because all values on the [0,1] interval are plausible. In the example presented in MacCallum et al. there is only one loading which is assumed to be zero or near zero according to the prior. In that case one should also expect a standardized unique variance to be near 1 and that is what the induced prior says in their Figure 1.

It should also be noted that we have found that different choices of priors

for residual variances do not make a notable difference in the estimation of key parameters such as factor loadings. We agree, however, that the choice of priors for residual correlations is a topic that warrants further study.

[Figure 6 about here.]

Response to Rindskopf

Rindskopf suggests several ways to extend BSEM analysis using priors. We give examples of two of these suggestions: Using priors for primary parameters and doing bi-factor analysis with small-variance priors for cross-loadings in the Holzinger-Swineford model discussed earlier. We hope that these novel analyses will further stimulate readers to learn about these new methods.

Priors on primary parameters: The Kaplan science model

Rindskopf suggested using informative priors not only for the minor loadings referred to as cross-loadings, but also for major loadings. The Kaplan science model in Figure 3 of our paper offers an interesting example where this idea is useful. Table 16 of our paper shows eight models and here we add one more that fits well in terms of PP p-value. The 12 direct effects from the indicators of the factors involve and challenge on the scigra10 and sciach outcomes are given small-variance priors, adding to the informative small-variance priors for structural coefficients, direct effects of covariates onto the factor indicators, and cross-loadings. Correlated residuals are, however, not included.

Figure 7 shows the posterior distribution of the regression coefficient for the scigra10 outcome regressed on the challenge factor from a BSEM analysis with 10,000 iterations and recording every 50th iteration to reduce auto-correlation. This posterior, as well as those for the three direct effects from the factor indicators of the challenge factor, is rather rectangular with a wide span from negative to positive values with a middle point about zero. This means that the posteriors are not informative about the parameter values. This may be a reasonable outcome in a case where maximum-likelihood estimation finds that two of these three indicators have significant direct effects. Reducing the prior variance from 0.01 to 0.005 for the direct effects does not change this picture. In line with Rindskopf's suggestion, it can be argued that the Kaplan model implies an a priori hypothesis of a major coefficient for the scigra10 outcome regressed on the challenge factor. Corresponding to this, the prior N(1,0.25) is used for the effect, implying that a near-zero effect is unlikely given that a zero effect is two standard deviations below the prior mean. This prior changes the results considerably in that convergence is faster and the results show a posterior with much less variation and a significant positive effect of the challenge factor on the scigra10 outcome, as well as a large direct effect from one indicator. The PP p-value of 0.455 indicates an acceptable model.

[Figure 7 about here.]

BSEM applied to the bi-factor model

Rindskopf suggested that BSEM might be useful for bi-factor analysis. Returning to the Holzinger-Swineford example we started with to illustrate sign switching, this is yet another illustration of BSEM strength.

When applying the bi-factor model to the 24-variable Grant-White Holzinger-Swineford data, using the sign relabeling algorithm, it is found that the model does not fit well as judged by the PP p-value of 0.001. The BSEM approach of small-variance priors for cross-loadings is, however, applicable also in this case. Using prior variances of 0.01 results in a well-fitting model with a PP p-value of = 0.477. The solution is shown in Table 1. Only two cross-loadings are significant, where the connection of the arithmetic test with the speed factor and its addition item is well known.

[Table 1 about here.]

It is interesting to note that the BSEM analysis is the only approach known to us that both fits well for these classic data and corresponds to the Holzinger-Swineford hypothesized model structure. A regular bi-factor confirmatory factor model does not fit well by ML or Bayes. ML bi-factor ESEM (Asparouhov & Muthén, 2009) gives a well-fitting model but does not fully capture the hypothesized model structure. The same is true for the recently introduced ML bi-factor exploratory factor analysis approach of Jennrich and Bentler (2011). The Mplus scripts for all these analyses are available from the authors.

References

[1] Asparouhov, T. & Muthén (2012). Bayesian exploratory factor analysis. Paper in preparation.

[2] Garrett, E. & Zeger, S.L. (2000). Latent class model diagnosis. *Biometrics*, 56, 1055-1067.

[3] Jennrich, R.I. & Bentler, P.M. (2011). Exploratory bi-factor analysis. *Psychometrika*, 76, 537-549.

[4] Muthén, B. (1989). Teaching students of Educational Psychology new sophisticated statistical techniques. In M.C. Wittrock & F. Farley (eds.), *The future of Educational Psychology* (pp. 181-189). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

List of Figures

Figure 1: Trace for the loading of the variable visual on the Spatial factor in a bi-factor model for Holzinger-Swineford's Grant-White data ($n = 145$)

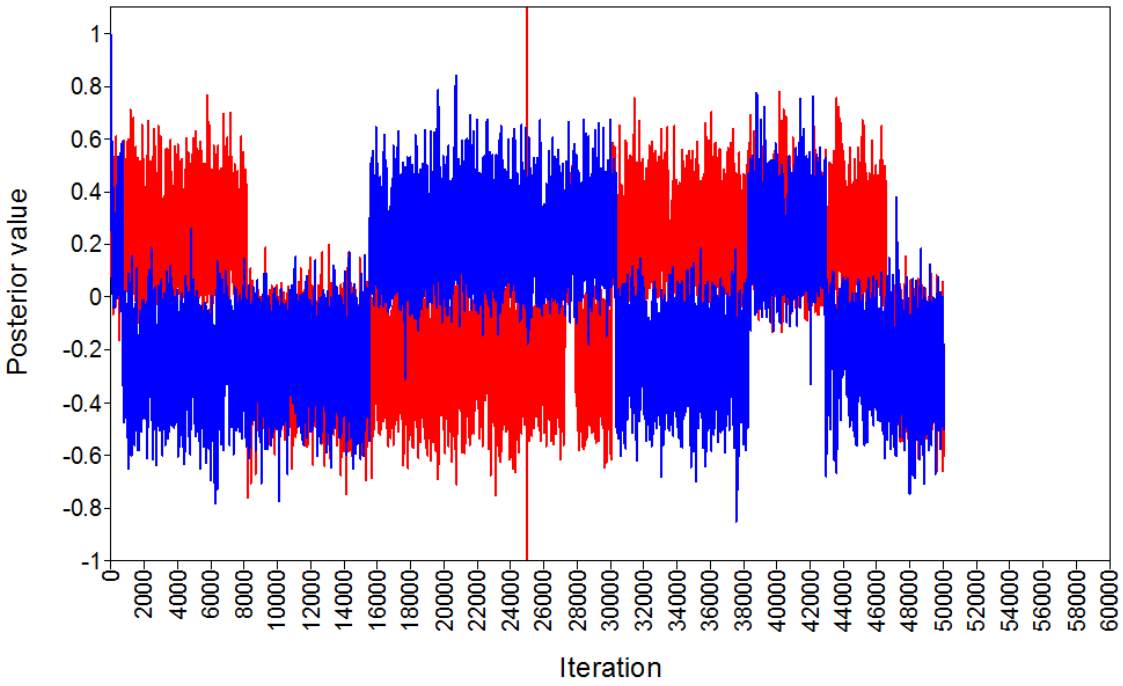


Figure 2: Posterior distribution for the loading of the variable visual on the Spatial factor in a bi-factor model for Holzinger-Swineford's Grant-White data ($n = 145$)

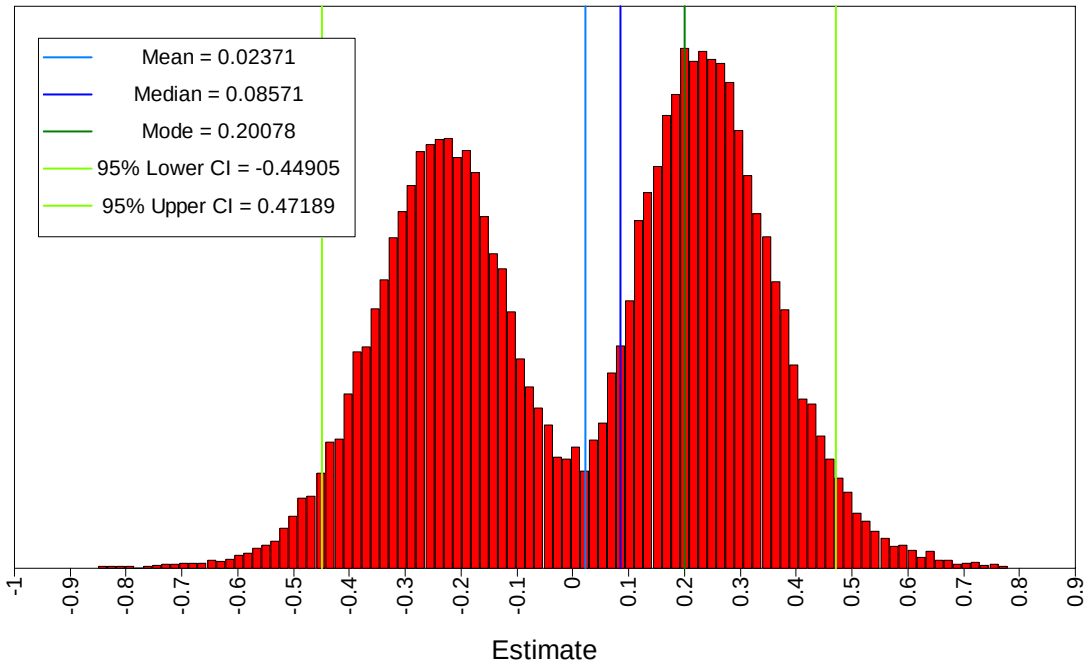


Figure 3: Trace for the loading of the variable visual on the specific factor Spatial in a bi-factor model for Holzinger-Swineford's Grant-White data ($n = 145$) after relabeling

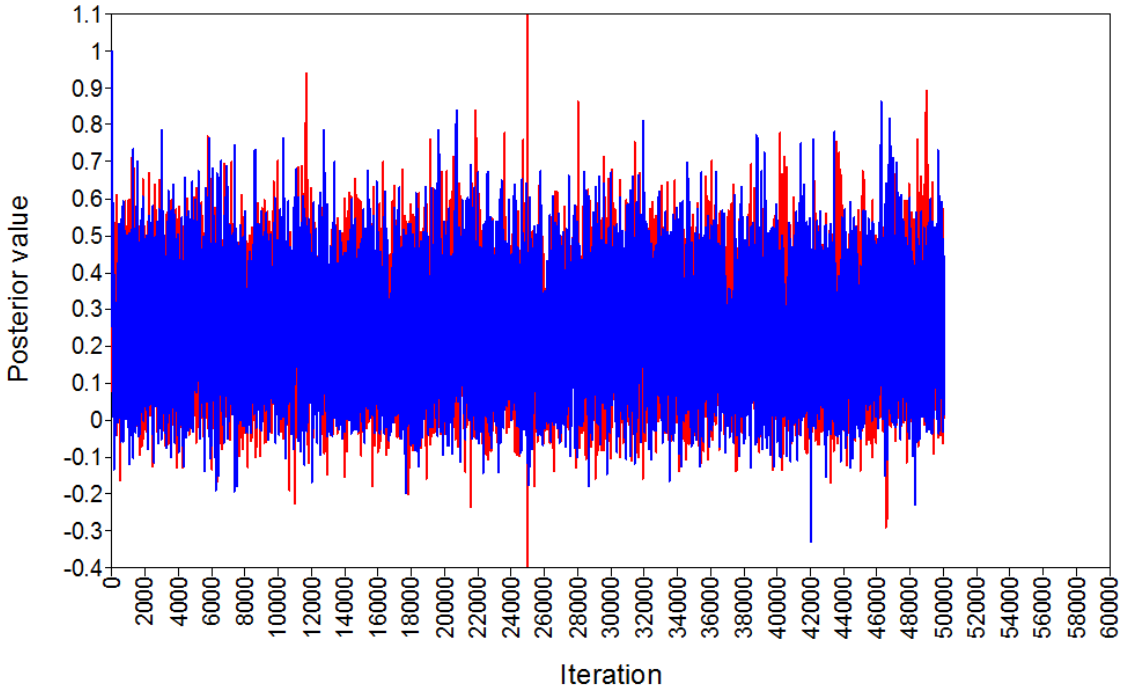


Figure 4: Posterior distribution for the loading of the variable visual on the Spatial factor in a bi-factor model for Holzinger-Swineford's Grant-White data ($n = 145$) after relabeling

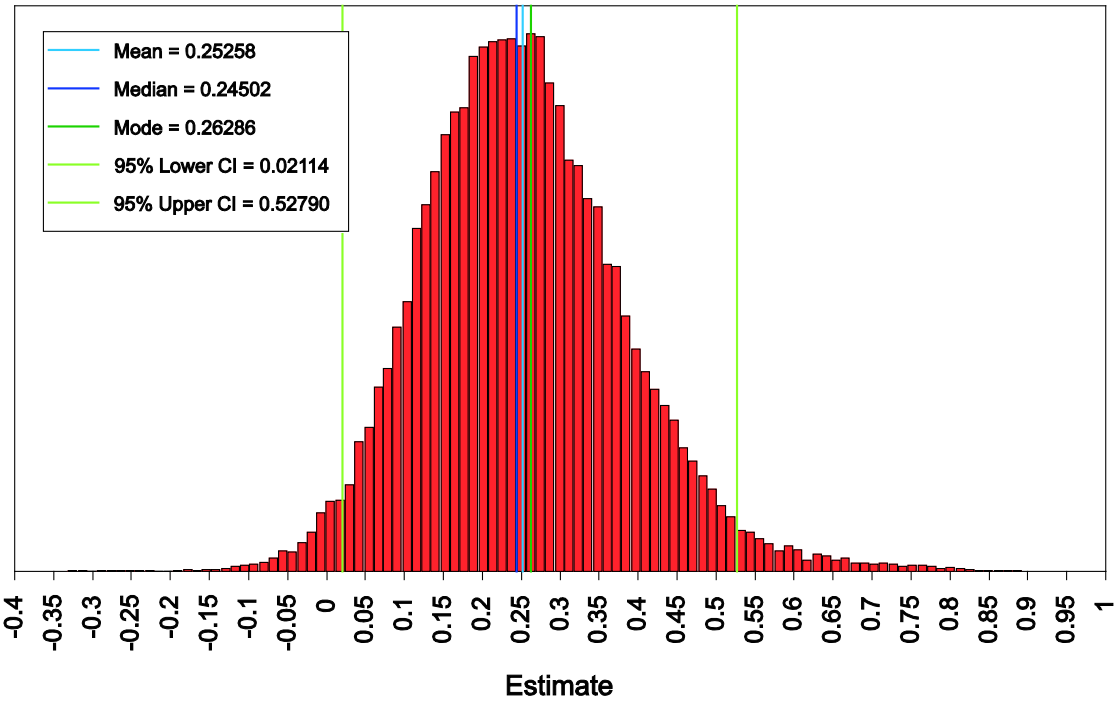


Figure 5: Trace for the non-identified factor mean of the Spatial factor in a factor model for the spatial items of the Holzinger-Swineford's Grant-White data ($n = 145$)

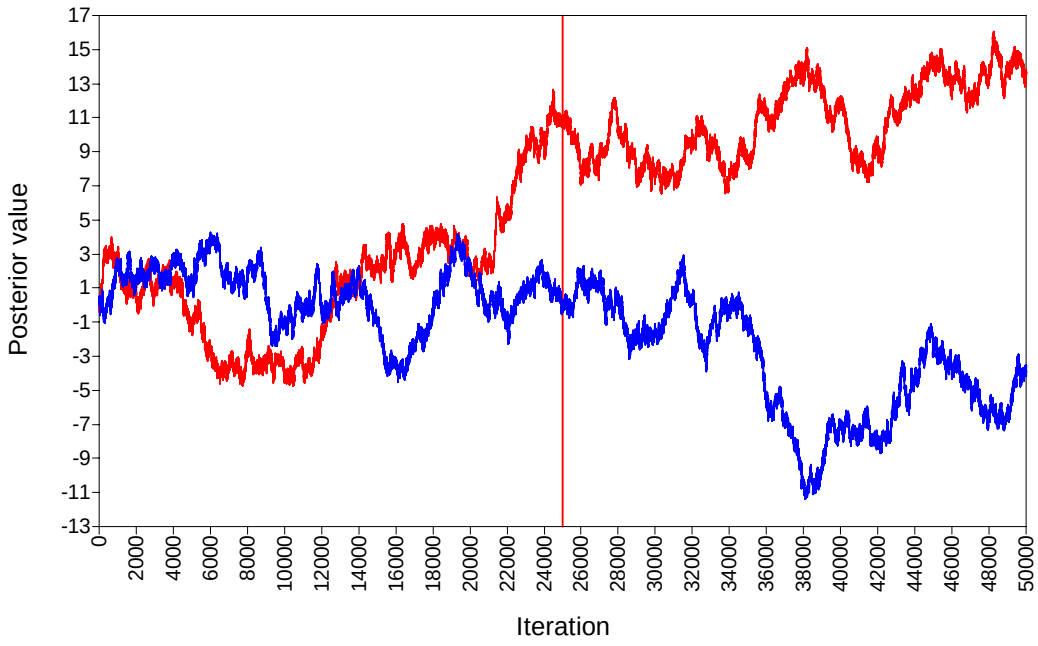


Figure 6: Prior distribution for the standardized residual variance of the visual test in a factor model for the spatial and verbal items of the Holzinger-Swineford's Grant-White data ($n = 145$)

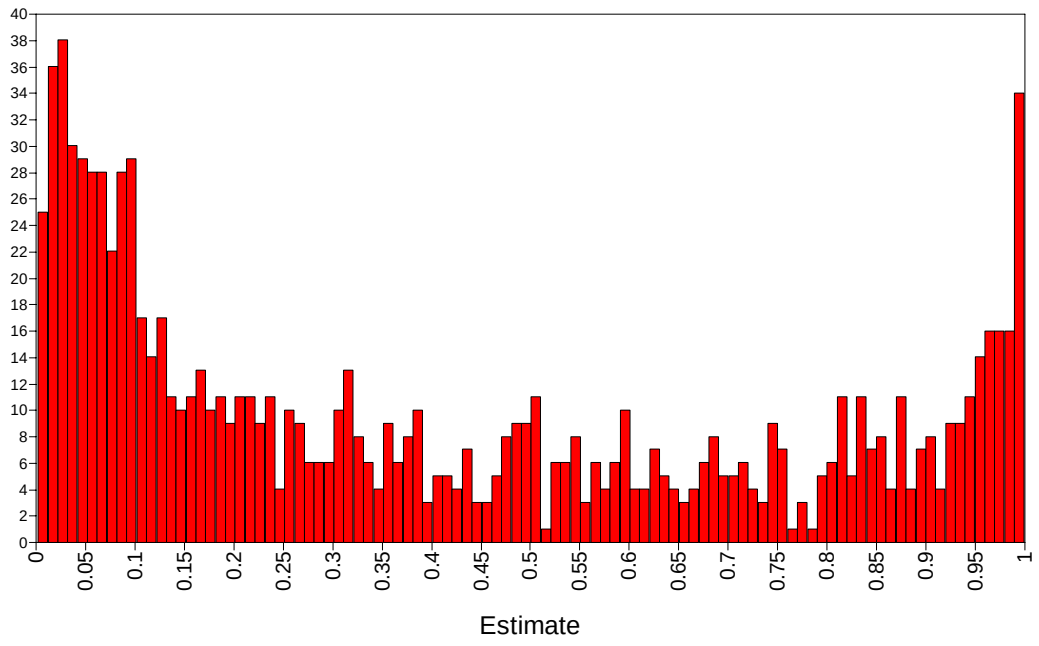
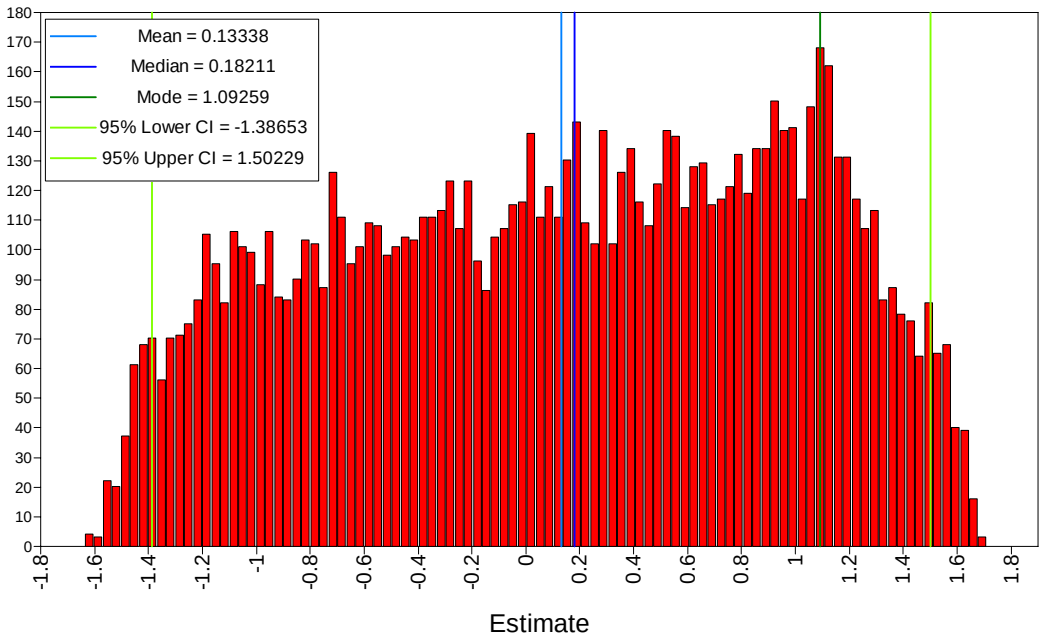


Figure 7: Posterior distribution for the regression of the scigra10 outcome on the challenge factor in the Kaplan model with direct effects from factor indicators to scigra10 and sciach



List of Tables

Table 1: BSEM solution for bi-factor modeling of Holzinger-Swineford's 24-variable Grant-White data using small-variance priors for cross-loadings

Factor Loadings

	General	Spatial	Verbal	Speed	Memory
visual	0.625	**0.356***	-0.027	-0.009	-0.003
cubes	0.440	**0.239**	-0.037	-0.045	-0.046
paper	0.418	**0.368***	0.052	0.047	0.044
flags	0.589	**0.343***	0.008	-0.058	-0.036
general	0.609	0.024	**0.532***	0.041	-0.040
paragrap	0.578	0.016	**0.596***	-0.033	0.050
sentence	0.582	-0.021	**0.607***	0.017	-0.041
wordc	0.622	0.029	**0.360***	0.053	-0.017
wordm	0.618	-0.023	**0.590***	-0.091	0.010
addition	0.446	-0.163	0.018	**0.647***	0.021
code	0.514	0.020	0.040	**0.400***	0.135
counting	0.501	0.009	-0.107	**0.553***	-0.039
straight	0.593	0.183*	0.017	**0.406***	-0.039
wordr	0.379	-0.013	0.051	0.000	**0.425***
numberr	0.370	0.025	-0.001	-0.016	**0.397***
figurer	0.531	0.102	-0.062	-0.058	**0.341***
object	0.442	-0.102	0.022	0.069	**0.507***
numberf	0.551	0.043	-0.102	0.088	**0.289***
figurew	0.454	-0.031	-0.002	0.001	**0.215**
deduct	0.666	-0.017	0.070	-0.097	0.035
numeric	0.668	0.041	-0.073	0.105	-0.019
problemr	0.671	-0.012	0.058	-0.093	0.027
series	0.743	0.027	0.036	-0.039	-0.023
arithmet	0.664	-0.150	0.076	0.172*	0.046

Factor Correlations

	General	Spatial	Verbal	Speed	Memory
General	1.000				
Spatial	0.000	1.000			
Verbal	0.000	0.000	1.000		
Speed	0.000	0.000	0.000	1.000	
Memory	0.000	0.000	0.000	0.000	1.000