Advances in Bayesian Model Fit Evaluation for Structural Equation Models

Tihomir Asparouhov and Bengt Muthén

November 7, 2019

Abstract

In this article we discuss the Posterior Predictive P-value (PPP) method in the presence of missing data, the Bayesian adaptation of the approximate fit indices RMSEA, CFI and TLI, as well as the Bayesian adaptation of the Wald test for nested models. Simulation studies are presented. We also illustrate how these new methods can be used to build BSEM models.

1 Introduction

This paper makes three contributions to Bayesian model fit evaluation for Structural Equation Models (SEM). First, we discuss improvements to the Posterior Predictive P-value (PPP) method in the presence of missing data. Second, Bayesian versions of the approximate fit indices RMSEA, CFI and TLI, constructed in Garnier-Villarreal and Jorgensen (2019), are studied in both the complete and the incomplete data cases. Third, the paper proposes a Bayesian adaptation of the classic Wald test of multiple hypothesis. These methods apply to standard SEM models estimated within the Bayesian framework as well as the Bayesian Structural Equation Models (BSEM) discussed in Muthén and Asparouhov (2012), where small variance priors are used to relax the SEM model to accommodate minor differences between the model and the observed data. We show how the Bayesian fit indices can be used instead of the PPP to build approximately well fitting BSEM models. Simulation studies are presented for illustration purposes. All of the above methods are implemented in Mplus 8.4.

2 PPP in the presence of missing data

The construction of the PPP value used in Mplus 8.3 and earlier versions is described in Asparouhov and Muthén (2010a). It follows standard posterior predictive checking methodology, see Gelman et al. (2004), based on a discrepancy function. The discrepancy function used in Mplus is the likelihood ratio test (LRT) function comparing the estimated model, the H0 model, and the unconstrained mean and variance covariance matrix model, typically referred to as the H1 model

$$D = D(m, S, \mu, \Sigma) = \frac{n}{2}\Big(log(|\Sigma|/|S|)+Tr(\Sigma^{-1}(S+(\mu-m)(\mu-m))-p\Big) \tag{1}$$

where S is the sample variance covariance, m is the sample mean, Σ is the model implied variance covariance, μ is the model implied mean, n is the sample size, and p is the number of variables in the model. Thus, the discrepancy function is the standard test of fit used with the ML estimation. In the posterior predictive checking, the discrepancy function is evaluated at the entirety of the posterior distribution of the model parameters, i.e., the discrepancy function is evaluated for every iteration in the MCMC estimation. Using the i-th iteration

H0 model parameter estimates, we compute the model implied mean and variance/covariance μ_i and Σ_i. We then compute the observed data discrepancy function as $D_i^{obs} = D(m, S, \mu_i, \Sigma_i)$. Using the i-th iteration H0 model, we generate a replicated data set of the same size as the original data set and compute the sample mean and variance of the replicated data m_i and S_i. The replicated data discrepancy function is computed as $D_i^{rep} = D(m_i, S_i, \mu_i, \Sigma_i)$. Finally the PPP value is computed as follows

$$PPP = P(D^{obs} < D^{rep}) \approx \frac{1}{L} \sum_{i=1}^{L} \delta_i \tag{2}$$

where L is the number of iterations in the MCMC estimation, and $\delta_i = 1$ if $D_i^{obs} < D_i^{rep}$ and 0 otherwise.

When there are no missing data the H1 model does not require an actual estimation. The sample mean and the sample variance covariances are the ML estimates for the H1 model parameters. These sample statistics are used for the construction of the discrepancy function. The discrepancy function, as defined and used in Mplus, does not use the posterior distribution of the H1 model parameter estimates. Instead it uses the actual H1 ML estimates. This is done for both the observed data as well as the replicated data.

In the presence of missing data, the H1 model ML estimates are not readily available. In Mplus 8.3 and earlier versions, the missing data are essentially ignored for the purposes of defining the discrepancy function. During the MCMC estimation the missing data are imputed and the discrepancy function uses these imputed values as if they are the actual observed values, i.e., the discrepancy function is defined as the test of fit for the observed and imputed data combined. This simplification is quite useful in terms of the speed of the computation. Unfortunately, it has the caveat that it weakens the information in the observed data. The imputed data are generated from the estimated model during the MCMC estimation and thus are an affirmation that the model is correct. Unlike the observed data, for which we do not know if there is a good fit for the model, the imputed data are a perfect fit for the model as they are generated from that model. As a result, the evidence of possible misfit between the observed data and the model is weakened by this method. Combining the observed data and the perfectly fitting imputed data weakens the power to detect misspecifications.

In Mplus 8.4, a new method for posterior predictive checking is implemented to remedy this situation. First, the discrepancy function is defined as the test of fit function for the observed data only, i.e., at each iteration in the MCMC estimation we compute the LRT function for the observed data only, using the current H0 model estimates as well as the H1 model estimates based on the observed data only. In addition, the replicated data are generated from the current H0 model, however, the replicated data are not a complete data set but an incomplete data set just like the observed data. For every missing observation in the observed data, we put a missing value in the replicated data in the exact same place so that the observed data and the replicated data have the same missing data patterns. Lastly, the discrepancy function for the replicated data is computed the exact same way as it is computed for the observed data. That is, the discrepancy function uses the LRT function for the H0 model using the current H0 model parameter estimates as well as the H1 model estimates obtained from the incomplete replicated data.

We can easily compute the H1 model parameter estimates for the observed data, however, these estimates are not easily available for the replicated data which changes at every iteration. To compute the discrepancy function for the replicated data, the EM algorithm must be performed for every replicated data set, which would become computationally very heavy. In fact, the software package blavaan, Merkle and Rosseel (2018), implements the EM algorithm at every iteration and as the authors point out, the PPP method becomes impractical altogether in the presence of missing data. To resolve this problem, instead of using the H1 ML estimates in the discrepancy function we use an H1 model parameter draw from the H1 model parameter posterior distribution. To be able to do that, however, we would still need to estimate the posterior distribution of the H1 model for every replicated data set which would still be computationally heavy. To resolve that problem, we use the following simplification. For every instance where the H1 model parameter estimates are needed, we perform a 10 iteration MCMC estimation of the H1 model and use the 10-th iteration as our draw. Running only 10 MCMC iterations for the H1 model is imperfect as a full convergence is neither checked nor enforced, however, the overall speed of the computation will not be compromised this way. Note that we use the 10-iteration approach not just for the replicated data but also for the observed data so that the comparison between the two is equitable.

There are two reasons why the 10-iteration approach works well. First, unlike structural models, estimating the H1 model is fast and simple. The MCMC autocorrelations in the H1 model estimation is small. Correlations between the H1 model parameters is also expected to be small. If the amount of missing data is moderate, the variation in the H1 parameter draws will be small as well, see Section 2.3.1 below. In many respects, the estimation of the H1 model amounts to estimating multiple bivariate normal distributions, i.e., increasing the number of variables does not increase the complexity of the model substantially. The second reason is as follows. In principle, we do not need to use in the definition of the discrepancy function the H1 model parameters distribution. We can use a different distribution. The discrepancy function can be defined in whatever way we want. Using the H1 model parameter distribution is useful because it directly addresses the model fit as compared to the fit provided by the H1 model, however, the discrepancy function can be defined differently. For example, if we have a way to approximate the H1 model parameter distribution, we can define the discrepancy function using that approximate distribution. The 10-iteration MCMC procedure provides exactly that, an approximation to the H1 model parameter distribution. Here is where the equitable approach becomes important. The way we compute the discrepancy function for the observed data must be exactly the same way we compute it for the replicated data. This is needed to preserve the validity of the posterior predictive checking. This second argument shows that even if the 10-iteration approach provides a mediocre approximation to the H1 model parameter distribution, the posterior predictive checking is still valid. One final issue to consider regarding the validity of the procedure is the starting values for the 10-iteration MCMC procedure. To obtain a comparable distribution for the H1 model parameters, we must use the same starting values for both the observed and the replicated data. These starting values are as follows. For the missing values, we use the current imputed missing values, i.e., those that are imputed from the H0 model in the current MCMC iteration. These are used to complement both the observed and the replicate data. The starting values for the means and the variance covariance parameters are the sample mean and variance covariance, computed when the missing values are added to the data set (observed or replicated).

We generate the replicated data sets so that they have the same amount of missing data as the real data set because we want the repli-

cated data to be comparable to the real data, under the null hypothesis that the H0 model is correct. Ideally, we would want the missing data in the replicated data set to have been obtained from the same missing data mechanism that generated the missing data for the observed data set. The missing data mechanism, however, is unknown and it is not estimated. Generating missing data completely at random in the replicated data could potentially lead to inappropriate PPP results for the simple reason that the replicated data are too different from the observed data due to a mismatch in the missing data mechanism. It turns out, however, that this is not the case. We use the estimated model to generate a complete replicated data set and then replace the generated values by missing values when the real data has a missing value. This missing data mechanism is not MCAR (missing completely at random) since the covariates in the replicated data are held equal to the covariates in the real data (only dependent variables are regenerated) and such covariates might have had influence on the probability of being missing in the real data (and therefore in the replicated data). Regardless, however, the missing data mechanism used to generate missing values for the replicated data is certainly MAR(missing at random). MAR is the model estimating assumption also for the real data set. Thus, both the real data and the replicated data have missing values which are MAR. It is well known that the likelihood of the observed data is independent of the missing data mechanism when the missing data are MAR, sea Little and Rubin (1987). This implies that the discrepancy function which is likelihood based is independent of the missing data mechanism. Thus, even though the missing data mechanism for the replicated and the real data sets are not identical, the discrepancy functions between these data sets are still comparable. Under the null hypothesis of correct H0 model, the discrepancy function of the replicated data and the real data should be similar.

Mplus computes the discrepancy function every 10-th MCMC iteration instead of every iteration to reduce the computational burden. In the presence of missing data, every time the discrepancy function is computed we also compute 10-iteration MCMC estimation for the H1 model. This averages out and thus the number of iterations used in the H1 model estimation is about the same as the number of iterations used in the H0 model estimation. To summarize, the new method of computing the PPP with missing data essentially amounts to estimating three models instead of one: the original H0 model using

the observed data, the H1 model for the observed data, as well as the H1 model for the replicated data. In most situations, however, the increase in the computational time will be negligible.

To formalize the new PPP computation in the presence with missing data, we define the discrepancy function D as follows

$$D(Y, \mu_1, \Sigma_1, \mu_0, \Sigma_0) = \mathscr{L}(Y|\mu_1, \Sigma_1) - \mathscr{L}(Y|\mu_0, \Sigma_0) \tag{3}$$

where Y represents the data and $\mathscr{L}(Y|\mu_j, \Sigma_j)$ represents the log-likelihood of Y based on the multivariate normal distribution with mean μ_j and variance/covariance Σ_j. Let Y^{obs} denote the observed data and let Y_i^{rep} denote the replicated data generated during the i-th MCMC iteration. The discrepancy function for the observed data is computed as

$$D_i^{obs} = D(Y^{obs}, \mu_{1i}(Y^{obs}), \Sigma_{1i}(Y^{obs}), \mu_{0i}, \Sigma_{0i}), \tag{4}$$

where $\mu_{1i}(Y^{obs})$ and $\Sigma_{1i}(Y^{obs})$ represent a random draw of the H1 model parameters estimates for Y^{obs}, and μ_{0i} and Σ_{0i} are the H0 model implied mean and variance/covariance obtained from the i-th iteration of the H0 model estimation. Similarly, the discrepancy function for the replicated data is computed as

$$D_i^{rep} = D(Y_i^{rep}, \mu_{1i}(Y_i^{rep}), \Sigma_{1i}(Y_i^{rep}), \mu_{0i}, \Sigma_{0i}), \tag{5}$$

where $\mu_{1i}(Y^{obs})$ and $\Sigma_{1i}(Y^{obs})$ represent a random draw of the H1 model parameters estimates for Y_i^{rep}. The PPP is then computed as in (2).

Next we illustrate the advantages of the new method using simulation studies.

2.1 Improving Power

In this section we compare the power of the new PPP method to the power of the PPP method implemented in Mplus 8.3. We generate data using a CFA model with two factors each measured by 3 indicators. The loadings for the first factor are all set to 1 while for the second factor they are set to 1, 0.8 and 0.8. The means of the indicators are set 0 and the residual variances are set to 1. The factor variances are set to 1 and the covariance between the two factor is set to 0.7. Missing data are generated completely at random at a constant rate for each indicator. The data are analyzed using a two factor

Table 1: Comparing power: average PPP (rejection rate)

Number of factors	Sample size	Missing rate	V8.3	V8.4
1	300	0.25	.16(.23)	.07(.67)
1	300	0.50	.40(.00)	.23(.16)
1	1000	0.25	.01(.97)	.00(1.0)
1	1000	0.50	.28(.00)	.03(.81)
2	300	0.25	.50(.00)	.51(.01)
2	300	0.50	.49(.00)	.49(.00)
2	1000	0.25	.48(.00)	.50(.00)
2	1000	0.50	.50(.00)	.51(.00)

CFA model (the true model) as well as a 1 factor CFA model. Mplus default non-informative priors are used for all of the estimations in this note.

Table 1 contains the results for various sample sizes and missing data rates based on 100 replications. The average PPP values, obtained with Mplus 8.3 and Mplus 8.4, are given as well as the rejection rates based on these PPP values. A model is rejected when the PPP value is below 0.05. We can clearly see that the new PPP method is much more powerful. It rejects the incorrect 1 factor model more often, while preserving the type I error rate, i.e., the false rejections when the model is the correct 2 factor model. Typical tests of fit will reach the nominal 5% rejection rate when the model is correct. Table 1 results show, however, that the PPP type I error is below the 5% rejection rate even for large sample size situations. This discrepancy occurs not just for the incomplete data case but also in the complete data case, see Asparouhov and Muthén (2010b). This discrepancy is due to the fact that the PPP value is not uniformly distributed as the P-value in the classical likelihood ratio tests, see Hjort et al. (2006).

2.2 Missing at random

In this section we illustrate the quality of the new PPP method when the missing data are missing at random (i.e. not completely at ran-

dom). Here we use the simple regression model

$$Y_1 = \alpha + \beta Y_2 + \varepsilon \tag{6}$$

where we set $\alpha = 0$, $\beta = 0.3$, $Var(\varepsilon) = Var(Y_2) = 1$. For this simulation we generate 100 data sets of size 500. We generate missing values according to the following model

$$P(Y_1 \text{ is missing}) = Exp(Y_2)/(1 + Exp(Y_2)). \tag{7}$$

This is a very informative missing data mechanism. Full likelihood estimation such as ML and Bayes would however yield unbiased estimates for the model parameters because the missing data mechanism depends only on observed values and in this example Y_2 has no missing values. Note here that the overall means in the population for both Y_1 and Y_2 are 0. When the Y_1 value is not missing, however, the mean of Y_1 is no longer 0, it is -0.12 while the mean of Y_2 is -0.4. When the Y_1 value is missing the mean of Y_2 is 0.4.

In this simulation example we treat the Y_2 variable as a dependent variable (i.e. not an actual covariate). While computing the PPP, replicated data are generated for both Y_1 and Y_2 according to the estimated model. In the model estimation, the missing data mechanism (7) is not know. The missing data generation for the replicated data (as described earlier) will amount to MCAR because there are no covariates in this model. Thus, in contrast to the analyzed data, in the replicated data the mean of Y_1 and Y_2 will be 0 regardless of weather or not Y_1 is missing. Nevertheless, this simulation study shows that these kinds of replicated data are sufficient to evaluate the model fit using the new PPP method.

First, we estimate the baseline model where the parameter β is fixed to zero. Using Mplus 8.4 we obtain an average PPP value of .01 and in 89% of the replications the baseline model is rejected. Because the information regarding the baseline model is contained solely in the full pattern observations when both Y_1 and Y_2 are observed we can expect this rejection rate to occur if we run the model only with the observed patterns. To obtain such a rejection rate we conduct a simulation study without any missing data and with sample size of 250 (given that approximately 50% of the Y_1 values are missing) and we obtain an average PPP value of .01 and a rejection rate of 95%. This result confirms that approximately the correct rejection rate has been obtained in the missing data simulation. If we conduct

the same experiment on the full regression model where β is estimated, the PPP value is .49 and no replications were rejected in the missing data case. In the no missing data case with a sample size of 250 we obtain an average PPP value of .50 and 0 rejection rate which again confirms that the missing data computation works correctly. The fact that the average PPP value is near 50% when the model is correct is very important. This can only occur if the replicated data are of the same quality as the observed data in terms of the discrepancy function. Thus, we confirm here that the missing data mechanism had no impact on the PPP value. Even though a different missing data mechanism was used for the replicated data, we obtain a discrepancy function that matches that of the observed data, i.e., it is equally likely to be higher or lower than that of the observed data.

The improvement in power described earlier can be found in this example as well. Estimating the baseline model with missing data with Mplus 8.3 yields an average PPP value of .14 with a rejection rate of 28% which is substantially lower than the corresponding value of 89% obtained using Mplus 8.4.

2.3 Additional points

2.3.1 Smooth transition between the missing data case and the complete data case

The missing data PPP provides a smooth transition between the missing and the non-missing case. Consider the situation when a single value is missing. It would be undesirable if the new method yields substantially different results from the complete data case. The likelihood of the observed data will change very little when a single observation is missing. In fact, the only difference between the missing data case and the complete data case would be the likelihood of the observation that is missing. The H1 model parameter estimates, however, will vary in the missing data case while in the complete case they will remain constant. This would be particularly the case when the overall sample size is moderate or small because of the wider posterior distribution. The H1 model estimates in the incomplete case would be selected from the entire posterior distribution while in the complete case the parameters will be fixed at their ML estimates. This additional variability in the incomplete case could potentially compromise the power of the PPP as the range of the discrepancy function val-

ues will be wider, even when just a single observation is missing. To eliminate this possibility, we implement a small modification of the methodology. In every step when the H1 model has to be estimated, we perform the 10-iteration MCMC estimation. At the end of that estimation, instead of taking the last H1 parameter estimates we simply use the sample mean and variance based on the current imputed values. The variability in these sample mean and variances will be much smaller than that of the H1 model parameter estimates as it will be primarily driven by the actual observed (non-missing) values. The smaller the number of missing value in the data, the smaller that variability will be. This way the impact on having a small number of missing observations will be negligible. In fact, if there are no missing observations in the data but we run the estimation as if there are missing values (by including the Mplus option missing=all(999); for example), the discrepancy function will be defined the exact same way as if we run the estimation as in the complete data case. Note, however, that the speed in the computation will still be lower if the analysis is run through the missing data algorithm. In addition, due to changes in the order of the random number generation, the PPP values will not be identical, even through the algorithms are identical. This difference is not in the quality of the PPP values, however, and the results should be similar.

This simple modification in the H1 model estimation is also useful when there are covariates in the model. Covariates, unlike the dependent variables, are never allowed to have missing values as no formal model is specified for the imputation of such missing values. Other dependent variables in the model may not have any missing values as well. For all such variables, the ML H1 model estimates are easy to obtain and due to this modification will be used in the definition of the discrepancy function.

2.3.2 The effect of varying H1 model estimates

A simple experiment can be constructed to evaluate the effect of the varying H1 model estimates. We use the regression model used in Section 2.2, where the data are generated using the non-zero regression coefficient but the estimated model has the coefficient fixed to zero. Missing data are generated completely at random for both variables at the rate of 50%. We use a sample size of 400 and obtain a rejection rate of 63% over 500 replications. The information regarding the

regression coefficient is contained entirely in the observations where both variables are observed, which on average is about 133 (given that 1/3 of the observations will have the first variable missing, 1/3 will have the second variable missing, and 1/3 will have no missing, Mplus excludes all observations that have both missing). We can therefore expect to get a similar rejection rate if we run the same simulation with sample size of 133 and no missing data. The result that we obtain from this simulation is a rejection rate of 66%. Notably here the missing data rejection rate is comparable to the corresponding rate of the non-missing case, which implies that if any loss of power exist it is very small. Some difference in the rejection rates is also likely due to the fact that in the missing data case the number of observations with full patterns vary while in the non-missing case it does not. Note also that this is a very simple and limited simulation study that provides a rough glimpse into the effect of varying H1 model estimates. More complicated examples may indeed manifest some more noticeable loss of power.

Note also that the results obtained in Section 2.2 also lead to the same conclusion. In the incomplete data case we obtained a rejection rate of 89% while the corresponding complete data case resulted in 95% rejection rate. The difference is sufficiently small to assume that the loss of power due to the variability in the H1 parameter estimates is minimal.

2.3.3 Negative values in the discrepancy function

The discrepancy function may yield negative values in the missing data situation which do not occur in the complete case. This occurs because the H1 model parameter estimates used for the definition of the discrepancy function are not the ML estimates. Therefore there is no guarantee (like in the complete data case) that the H0 and the H1 models at the current parameters values (obtained at a particular iteration of the MCMC estimation) are nested. This in principle is not a problem and in general the discrepancy function is not required to be positive. What is important is that the replicated data are treated the same way as the observed data in terms of the computation of the discrepancy function. This is why we defined the discrepancy function to use the H1 model parameter estimates the same way for both the real and the replicated data. We do not use a single H1 model estimation for the observed data (even though we could since the ob-

served data doesn't change across iterations). Instead, we always use a 10 iteration MCMC estimation. In some particulary difficult missing data situations with very informative missing data mechanisms it is possible that the 10 iterations are not enough for good model estimation. Nevertheless, because of the symmetry between the observed and replicated data we would expect good PPP performance even if the H1 model estimates have not converged. In our limited simulation studies, the 10-iteration MCMC estimation appeared to be sufficient for the purpose of obtaining a good quality PPP value.

2.3.4 Manipulating the PPP

In some respect the problem with the PPP version implemented in Mplus 8.3 goes beyond the loss of power. The method is technically speaking exposed to being manipulated. Consider a hypothetical example where a data set is analyzed using a model H0 and rejected by the PPP. If we attach now a large amount of missing data to the original data, that missing data will be imputed from the incorrect model. This imputed data will fit the incorrect model and therefore will weaken the evidence that the model does not fit. The more missing data we insert in the original data set the more the total data (imputed and observed) will appear to fit the model. As we keep on adding more and more missing data, eventually we can expect that the model will not be rejected by the PPP. This can potentially be done for any model of any degree of misfit. The new PPP method is safeguarded against this situation because it does not use the imputed data for model evaluation.

Mplus generally removes all observations that contain only missing values. Therefore the manipulative approach described above would not be as easy to implement. If we simply attach observations with all missing values Mplus will directly remove these observations from the analysis and the manipulation approach will fail. Mplus, however, will not remove an observation if just one variable is not missing. Thus, to implement the manipulation strategy, we would need to add a new variable, completely unrelated to the original model, which has no missing values at all. When this new variable is added, we can expand the original data with many missing observations which Mplus will not delete due to the new variable.

2.3.5 PPP power as compared to the power of the ML chi-square

Asparouhov and Muthén (2010b) point out that in the complete case situation the ML chi-square is more powerful than the PPP method. This discrepancy carries over in the incomplete case to a similar extent. If we run the example discussed in Section 2.3.2 using the ML estimation in the complete case we obtain a rejection rate of 93% while in the incomplete case we obtain a rejection rate of 90%. These values are much higher than the rejection rates of the PPP: 66% and 63% respectively.

The lower power of the PPP value can also be interpreted as an approximate fit index. For example, PPP values between 0.05 and 0.20 could be considered approximately fitting rather than completely fitting or completely rejected models. This approximately fitting interpretation, however, has the caveat that it does not carry over for large sample sizes. In large sample sizes the PPP behaves the same way as the ML chi-square, i.e., they are asymptotically equivalent. If we double the sample size in simulation study in Section 2.3.2, the PPP rejection rates for the complete and the incomplete case are 97% which nearly matches the 100% rejection rates obtained with the ML chi-square.

2.3.6 Improving PPP power by adjusting the cutoff value

At the heart of the lower power of the PPP, in the complete and the incomplete case, is the fact that the PPP value is not uniformly distributed between 0 and 1, as is the P-value in the classical likelihood ratio tests, see Hjort et al. (2006). The PPP rejection rates are obtained using the cutoff value of 0.05, i.e., we reject the model if the PPP value is smaller than 0.05. If the PPP value, however, is not uniformly distributed, the 0.05 value does not represent the 5-th percentile of the distribution. Typically, a larger value represents the 5-th percentile of the PPP distribution. As a result of that, the PPP-value will have lower rejection rate when the hypothesis is true, which results in lower type I error, as well as when the hypothesis is not true, which results in higher type II error and lower power. If we know the distribution of the PPP value, however, we can obtain the 5-th percentile of the distribution and use that value as the cutoff value. With this new cutoff value, we surely will obtain a type I error of 5% as well as lower type II error and improved power.

One complication in this approach is that we do not know the distribution of the PPP value. Furthermore, that distribution is not the same for all models. A simple way to see how much difference there is between the PPP value distribution and the uniform distribution is to look at the standard deviation of the PPP distribution. This standard deviation is computed in every Mplus Montecarlo output. The uniform distribution on the 0 to 1 interval has a standard deviation of 0.29. The further away the standard deviation of the PPP values is from this value, the bigger the discrepancy between the uniform distribution and the PPP value distribution and the bigger the need for the cutoff adjustment.

Consider for example the 2-factor CFA model discussed in Section 2.1. We generate and analyze the data with the correct 2-factor CFA model, using 25% missing data and sample size of 300, and we obtain the PPP standard deviation of 0.15. To estimate this quantity well, we use 3000 MCMC iterations, resulting in PPP computation based on the comparison of the original data and 300 replicated data sets. Next, consider the regression example discussed in Section 2.2. The PPP standard deviation is 0.03, which clearly indicates that this PPP distribution is much further away from being uniform than is the distribution of the PPP value of the 2-factor CFA. Notably, however, these PPP distributions appear to be largely independent of the sample size, missing data and the type of missing data. The 2-factor CFA with sample size of 1000 and 25% missing data yields a PPP standard deviation of 0.15. Without any missing data, we obtain a PPP standard deviation of 0.16 for both sample size of 300 and 1000. The regression model without the MAR missing data also yields a PPP standard deviation of 0.03. The fact that the sample size does not affect the PPP distribution has a simple explanation. The sample size essentially takes the role of a multiplicative factor in the discrepancy function and therefore it doesn't affect the comparison between the replicated and the observed data. A similar argument applies to the missing data effect.

The mean of the PPP distribution is typically near 0.5, given that the null hypothesis is correct. Note, however, that the standard deviation of the PPP is not a sufficient descriptor for the entire distribution. The shape of that distribution varies from one model to another. In the regression example, the skewness of the PPP distribution is 0.05 and the kurtosis is -.13, i.e., in this case the distribution is fairly close to a normal distribution, see Figure 1. In the two-factor CFA example,

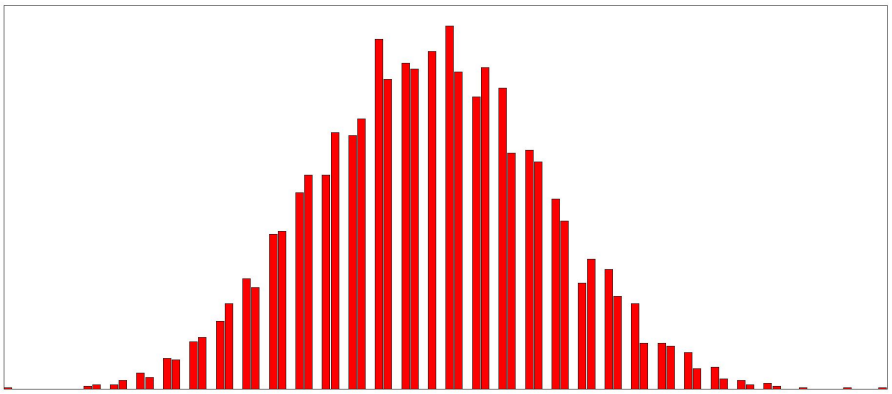

Figure 1: PPP distribution for the regression model

the skewness is -0.54 and the kurtosis is -0.27, i.e., the distribution is not close to a normal distribution but it has a heavy left tail, see Figure 2. For the uniform distribution the skewness is 0 and the kurtosis is -1.2. Estimating the 5-th percentile of the PPP distribution would need to be done by simply estimating the entire distribution.

Next, we illustrate how the adjusted cutoff value should be computed. Consider the example discussed in Section 2.1. The data is generated by a two-factor CFA model but is analyzed by a single factor model. In this illustration we use a sample size of 150 and no missing data. When we test the model fit of the single factor model using the PPP with the standard cutoff value of 0.05, we obtain a rejection rate of 62%. Using ML and the classic P-value, we obtain a rejection rate of 91%. Clearly the classic P-value outperforms the PPP value here. Next, we conduct a simulation study to obtain the distribution of the PPP. In a practical setting when a single model is estimated, the parameters for the simulation study should be chosen to be the same as the final results of the null model parameter estimates. In our simulation study, because we have multiple null model estimates, we use the average parameter estimates. Thus, we use the CFA 1-factor model average parameter estimates to construct a new simulation study where the data is generated and analyzed with a one

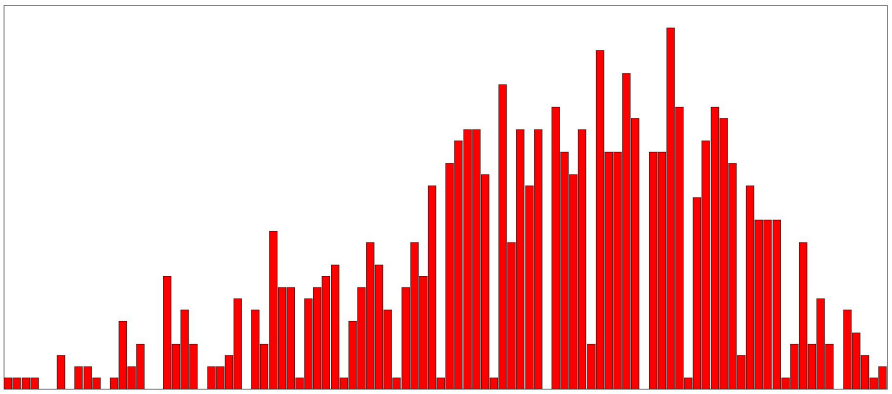

Figure 2: PPP distribution for the 2-factor CFA model

factor CFA model. We obtain the PPP distribution for that model and the 5-th percentile of that distribution is estimated to be 0.182. In this estimation we used 1000 replications and 3000 MCMC iterations for each replication. Using the new cutoff value of 0.182 we can now recompute the rejection rate for the PPP in the original simulation study and we obtain a rejection rate of 0.87. This nearly matches the ML rejection rate of 0.91. We conclude that the PPP power can be recovered simply by adjusting the cutoff value.

Note that in the above illustration, two simulation studies were described. The first one is where the data is generated as a 2-factor CFA but is analyzed as a 1-factor CFA. We use this simulation study to determine the rejection rate and the power of the testing procedure. The second simulation study is where the data is generated and analyzed with the null 1-factor CFA. The purpose of that simulation study is to determine the distribution of the PPP value and in particular the 5-th percentile, which will then be used as a cutoff value for the first simulation study. In practical settings where only a single data set is analyzed, only the second simulation study would be needed to determine the 5-th percentile of the PPP distribution. Because the PPP distribution and its 5-th percentile depend on the model, the second simulation is indeed required for the cutoff value

adjustment. Universal cutoff value adjustment that applies to all models can not be recommended. The cutoff adjustment method can be useful when the PPP value is between 0.05 and 0.25. For larger or smaller values, the cutoff adjustment method is unlikely to change the conclusion regarding the fit of the model.

Note, however, that the lower power of the PPP may not necessarily viewed as a problem, see Hoijtink et al.(2019). The lower power is always associated with a lower type I error, which in some situations can be viewed as an advantage of the Bayesian methodology.

2.3.7 PPP power as compared to the power of Wald type testing

PPP is useful when an entire structural model has to be evaluated. The test should not be used as a way to test nested models when we can test a hypothesis directly with the parameter estimates and their confidence limits. For individual parameters, using directly the confidence limits should be preferred. If multiple parameters need to be tested and a Wald type test (see below for the Bayesian adaptation of the Wald test) can be formulated, it should be used instead of PPP. This is because the asymptotics work much better for the standard errors of the parameters than for PPP. In fact in most cases (when there are no informative priors) the Bayesian parameters estimates and their standard errors match the ML results quite well. This means that the power of the Bayesian inference based on the standard errors will be identical to that of the ML method (and the ML chi-square). Consider again the simple regression model. The PPP test of the baseline model is equivalent to testing the significance of the β coefficient. Using a confidence interval directly yields 90% rejection of the baseline model in the complete and the incomplete case with the Bayesian estimation. Similarly, using the ML estimation we get 91% rejection rate for the complete case and the incomplete case. These also match the corresponding 93% and 90% rejection rates of the ML chi-square. In all of these cases, we get better power than the power of PPP using the 0.05 cutoff value which is 66% and 63% for the complete and incomplete case respectively.

2.3.8 New PPP method and its connection to the comparative fit indices CFI/TLI

Using the new PPP method is particularly important for the purpose of constructing comparative fit indices. Both CFI and TLI are based on evaluating the fit of the baseline model, which typically is a very poorly fitting model. In the baseline model estimation, the imputed data will be dramatically different from the observed data because the imputed data will have no correlations between the variables. Using a PPP method that combines the imputed data and the observed data, as the method implemented in Mplus 8.3 does, is bound to badly distort the PPP value, much more so for the baseline model than it would for other structural models that are not as poorly fitting. Using the new PPP method we can construct comparative fit indices that closely match their ML analogues. We use simulation studies below to illustrate this point.

2.3.9 Availability

Currently, the new PPP method is available only with continuous items, single level models (i.e. not for multilevel models), and is available for multiple group models. The above logic, however, indicates that the method would apply also for the case of missing categorical data as well as for multilevel models. Future releases of Mplus would upgrade the PPP in those situations as well.

3 Approximate Fit Indices

In this section we discuss how the approximate fit indices RMSEA, CFI and TLI used for the evaluation of SEM models are adopted to the Mplus Bayesian framework. Approximate fit indices are intended to circumvent a deficiency of rigorous testing procedures such as the chi-square test of fit which can reject a model even when the model misspecifications are minor, i.e., substantively insignificant. Structural models often involve a large number of variables and not all correlations between the variables can be fully accounted for with a simplified SEM model with a limited number of factors. If the sample size is sufficiently large, a small difference between the correlations observed in the data and the correlations implied by the model can become statistically significant and can be a reason to reject the model.

Augmenting the model by additional parameters such as residual correlations is one possibility to deal with model rejections. If the model is large, however, the process of adding parameters can become tedious and can compromise the power of the model based inference. Adding more parameters to the model makes the model less parsimonious and may increase the standard error of its parameters. This in turn will likely compromise the inference based on the model. In practical settings, it is often desirable to retain a substantively driven model as long as it fits the data approximately. Fit indices such as RMSEA, CFI and TLI are typically used to evaluate the approximate fit of a model. Hu and Bentler (1999) suggest the following fit index cut off values for reasonably well fitting models: RMSEA<0.06, TLI>0.95, CFI>0.95. Browne and Cudeck (1993) suggest an RMSEA cut off value of 0.05.

Hoofs et al. (2017) and Garnier-Villarreal and Jorgensen (2019) propose different methods for adapting fit indices to the Bayesian framework. The differences between these methods are fairly small in general but the Garnier-Villarreal and Jorgensen (2019) method appears to produce results closer to the ML results and thus we have adopted this approach in Mplus. To compute the RMSEA index, at each MCMC iteration we compute

$$RMSEA_i = \sqrt{max\left(0, \frac{D_i^{obs} - p^*}{(p^* - pD)N}\right)}\sqrt{G}. \tag{8}$$

Here i is the i-th iteration in the MCMC estimation. D_i^{obs} is the discrepancy function computed for the observed data at the i-th MCMC iteration. G is the number of groups in the model. p^* is the number of parameters in the H1 model. If p denotes the number of dependent variables in the model and q denotes the number of covariates in the model then $p^* = G(p(p+3)/2 + pq)$. pD is the estimated number of parameters for the H0 model which is typically produced in the Mplus output. pD is generally close to the number of parameters in the H0 model when there are no informative priors given for the model parameters. By default, Mplus computes the discrepancy function every 10-th iteration, for every chain in the model. As with the PPP estimation, Mplus uses only the second half of all iterations for inference. Thus if a model uses 1000 interactions to converge, using a 2 chain estimation, then RMSEA is computed in a total of 2(500/10)=100 iterations. The quantities $RMSEA_i$ are used to obtain the posterior

distribution of the RMSEA, which is then used to obtain the median, reported as the point estimate, as well as the confidence limits for the RMSEA.

The computation of the CFI and TLI are similar. For CFI we compute

$$CFI_i = 1 - \frac{D_i^{obs} - p^*}{D_{B,i}^{obs} - p^*}, \tag{9}$$

where $D_{B,i}^{obs}$ is the baseline model discrepancy function for the observed data computed at the $i-$th iteration of the baseline model MCMC estimation. CFI_i are then used to form the posterior distribution for the CFI and from there we obtain the point estimate and the confidence interval for the index. Similarly, for TLI we compute

$$TLI_i = \frac{(D_{B,i}^{obs} - pD_B)/(p^* - pD_B) - (D_i^{obs} - pD)/(p^* - pD)}{(D_{B,i}^{obs} - pD_B)/(p^* - pD_B) - 1}, \tag{10}$$

where $pD_B = 2pG$, i.e., we don't use the estimated number of parameters for the baseline model but the actual number of parameters. The baseline model is estimated with uninformative priors and thus the difference between the estimated and the actual number of parameters is expected to be small. Both TLI_i and CFI_i are truncated to the interval (0,1). The above definitions are derived from the ML definitions where the chi-square is replaced by $D_i^{obs} - pD$ and the degrees of freedom for the model is replaced by $p^* - pD$.

The approximate fit indices are intended to be used when the sample size is large, for example, when the sample size is more than 100 or even 200. When the sample size is small, if the PPP rejects the model, the differences between the data and the model must be quite large and the approximate fit should not be claimed regardless of what the fit indices are. If the PPP does not reject the model, then we can be confident that the model fits not only approximately well but also in the usual exact sense.

The Bayesian framework has an advantage over the ML framework as it provides confidence limits for all fit indices. The confidence limits can be used to determine when the sample size is too small to establish approximate fit. If the sample size is small, the confidence intervals will be large and most likely will contain the suggested cutoff values. When the confidence interval contains the cutoff value we can not be sure if the index is smaller or bigger than the cutoff value, i.e., the fit index is inconclusive. For CFI and TLI, if the 90% confidence interval

is above 0.95 we can claim that the model is approximately well fitting with certainty of 95% or more. If the confidence interval is below 0.95 we can claim with 95% certainty that the model is not fitting the data even approximately. Similar logic applies to the RMSEA. Using the entire RMSEA confidence interval, we reach one of the three possible conclusions: the fit index is inconclusive (the confidence interval contains 0.06), the fit index suggests that the model fits approximately well (the confidence interval is below 0.06), the fit index suggests that the model does not fit the data even approximately (the confidence limit is above 0.06).

Comparative fit indices are constructed by comparing the "distances" between the three models: the estimated model, the H1 model and the baseline model. If the baseline model is close to the H1 model, i.e., both models fit the data approximately well then there is roughly speaking not enough room for the estimates model to differentiate itself from the baseline model. Kenny (2015) recommends that RMSEA is computed for the baseline model and if that value is less than 0.158 (i.e. the baseline model provides a fairly good fitting model although not approximately well fitting model), the comparative fit indices CFI and TLI should not be used. RMSEA for the baseline model of less than 0.158 can be interpreted as an indicator that the distance between the baseline and the H1 model is too small (either due to small sample or due to very small correlations between the variables or both).

In the ML estimation, when the H0 is the same as the H1 model the CFI and TLI indices will both be 1. Similarly, if the H0 model is the same as the baseline model the two fit indices will be 0 in most cases. This will not be the case, however, for the Bayesian estimator due to the fact that there is random variation in the discrepancy function. Even if the H0 model is the same as the baseline model, $D^{obs}_{B,i}$ will not be the same as the D^{obs}_{i} because in the i-th MCMC iteration the estimates for the baseline and the H0 models will not be identical. The two models are estimated by two different MCMC sequences even though they are the same model in this instance.

In the next sections we illustrate the performance of the Bayes fit indices with several simulation studies.

3.1 Simulation example

In this section we use the example depicted in Figure 1 in Garnier-Villarreal and Jorgensen (2019). The model consist of 15 variables,

Y_1,...,Y_{15} measuring 3 factors f_1,...,f_3. The first factor is measured by the first five variables, the second is measured by the next five and the third is measured by the last five variables. The intercepts for all variables are set to 0, the factor loadings of the measurement model are set as follows $\lambda_{11} = \lambda_{21} = 0.7$, $\lambda_{31} = 0.75$, $\lambda_{41} = \lambda_{51} = 0.8$, $\lambda_{62} = \lambda_{72} = 0.7$, $\lambda_{82} = 0.75$, $\lambda_{92} = \lambda_{10,2} = 0.8$, $\lambda_{11,3} = \lambda_{12,3} = 0.7$, $\lambda_{13,3} = 0.75$, $\lambda_{14,3} = \lambda_{15,3} = 0.8$. There are two cross loadings in the model $\lambda_{42} = \lambda_{13} = 0.5$. The factor variances are set to 1 while the factor covariances are set to $\psi_{12} = 0.5$, $\psi_{13} = 0.3$, $\psi_{23} = 0.4$. The residual variances of the indicators are set as follows $\theta_1 = \theta_2 = \theta_6 = \theta_7 = \theta_{11} = \theta_{12} = .51$, $\theta_3 = \theta_8 = \theta_{13} = .4375$, $\theta_4 = \theta_5 = \theta_9 = \theta_{10} = \theta_{14} = \theta_{15} = .36$. We generate the data according to the above model and we estimate the model with 3 levels of misspecification. Level L_0 doesn't have any misspecifications. Level L_1 estimates the model without the λ_{42} cross loading. Level L_2 misspecification consist of omitting both cross loadings. We also include MCAR missing data in this simulation study and we vary the amount of missing data.

The average results for the Bayes and the ML estimators for the three different fit indices for various sample sizes and missing data levels are reported in Table 2. Each row in this table represents the average result over 100 replications. Both estimators produce nearly identical results when there are no missing data. We can also see that the fit index stabilizes as the sample size increases. This essentially makes the fit indices independent of the sample size and a valuable alternative to the chi-square test of fit and the PPP value. In the presence of missing data when the sample size is small, some differences between the fit indices are visible. The smaller the data and the bigger the amount of missing data, the bigger the difference. The Bayes fit indices tend to be more lenient which roughly corresponds to the lower power of the PPP as compared to the chi-square test of fit. As the sample size increases, however, even with large amount of missing data, the ML and the Bayes fit indices appear to converge to the same value.

Note again, however, that when the sample size is small, the fit indices are irrelevant. Consider the L2 model with sample size 300 and 50% missing data. With the Bayes estimator, 51% of the replications were rejected using PPP while all fit indices indicate approximately well fitting model in all 100 replications. If we compute the RMSEA for the baseline model we get an average value of 0.148, i.e., less than the cutoff value of 0.158 recommended by Kenny(2015). The sample

size in this case is too small to properly evaluate the approximate fit. Using PPP alone would be the preferred choice. For comparison, with sample size 300 and 25% missing data, the RMSEA for the baseline model is .214 on average. Note here that using confidence intervals on the fit indices is not helpful. The confidence limits for all three indices in this case are all of 0 length. The issue of not having enough distance between the baseline, the H0, and the H1 models can be detected only by computing the RMSEA of the baseline model. Note also that the ML fit indices CFI and TLI yield strange results as well in this situation. These values are .159 and .157 and they are very different from the values obtained for the same model with larger sample sizes.

We can also see in Table 2 how the chi-square and the PPP testing fails to recognize that the L1 and L2 models are approximately well fitting models. As the sample size increases the tests reject the models without being able to provide a nuanced conclusion regarding the level of misfit.

3.2 Using the confidence intervals for the fit indices

For our next illustration we use the example depicted in Figure 4 in Garnier-Villarreal and Jorgensen (2019). The model has 6 observed variables and 3 factors and is given by the following equations

$$y_1 = \mu_1 + \eta_1 + \varepsilon_1 \tag{11}$$

$$y_2 = \mu_2 + \lambda_2\eta_1 + \varepsilon_2 \tag{12}$$

$$y_3 = \mu_3 + \eta_2 + \varepsilon_3 \tag{13}$$

$$y_4 = \mu_4 + \lambda_4\eta_2 + \varepsilon_4 \tag{14}$$

$$y_5 = \mu_5 + \eta_3 + \varepsilon_5 \tag{15}$$

$$y_6 = \mu_6 + \lambda_6\eta_3 + \varepsilon_6 \tag{16}$$

$$\eta_1 = \beta_{12}\eta_2 + \beta_{13}\eta_3 + \xi_1 \tag{17}$$

$$\eta_2 = \beta_{23}\eta_3 + \xi_2 \tag{18}$$

The parameters used for the data generation are as follows: $\mu_i = 0$, $\lambda_i = 0.5$, $\beta_{12} = .6$, $\beta_{13} = -.435$, $\beta_{23} = -.6$. The residual variances for the six observed variables are set to 4.5, 3, 4.75, 2.5, 3, 2.5. The

Table 2: Comparing fit indices: Average values for Bayes/ML

Sample size	Missing	Misspecification	RMSEA	CFI	TLI	PPP/P-value rejections
100	0%	L0	.031/.029	.982/.985	.979/.981	.07/.18
300	0%	L0	.012/.011	.997/.997	.996/.996	.02/.08
1000	0%	L0	.005/.005	.999/.999	.999/.999	.02/.04
5000	0%	L0	.002/.002	1.00/1.00	1.00/1.00	.01/.05
100	0%	L1	.058/.056	.957/.960	.948/.951	.37/.65
300	0%	L1	.050/.051	.971/.970	.964/.963	.92/.98
1000	0%	L1	.051/.050	.970/.971	.964/.965	1.0/1.0
5000	0%	L1	.051/.051	.970/.971	.964/.964	1.0/1.0
100	0%	L2	.078/.075	.927/.931	.913/.917	.74/.90
300	0%	L2	.071/.071	.942/.941	.930/.929	1.0/1.0
1000	0%	L2	.071/.070	.941/.943	.929/.931	1.0/1.0
5000	0%	L2	.071/.070	.942/.943	.930/.930	1.0/1.0
300	10%	L0	.000/.012	1.00/.996	1.00/.996	.02/.10
1000	10%	L0	.000/.005	1.00/.999	1.00/.999	.02/.10
5000	10%	L0	.000/.002	1.00/1.00	1.00/1.00	.00/.02
300	10%	L1	.034/.047	.982/.970	.978/.964	.90/.96
1000	10%	L1	.044/.046	.974/.972	.969/.966	1.0/1.0
5000	10%	L1	.046/.047	.972/.971	.966/.965	1.0/1.0
300	10%	L2	.058/.066	.953/.941	.944/.928	1.0/1.0
1000	10%	L2	.064/.065	.945/.943	.934/.931	1.0/1.0
5000	10%	L2	.065/.066	.943/.942	.931/.930	1.0/1.0
300	25%	L0	.000/.015	1.00/.994	1.00/.993	.00/.07
1000	25%	L0	.000/.005	1.00/.999	1.00/.999	.01/.05
5000	25%	L0	.000/.002	1.00/1.00	1.00/1.00	.01/.04
300	25%	L1	.004/.041	.998/.969	.998/.962	.60/.83
1000	25%	L1	.031/.039	.983/.973	.980/.967	1.0/1.0
5000	25%	L1	.038/.040	.975/.972	.969/.966	1.0/1.0
300	25%	L2	.033/.059	.977/.939	.973/.926	.97/1.0
1000	25%	L2	.051/.057	.953/.944	.943/.932	1.0/1.0
5000	25%	L2	.056/.057	.946/.943	.935/.931	1.0/1.0
300	50%	L0	.000/.015	1.00/.168	1.00/.167	.00/.86
1000	50%	L0	.000/.006	1.00/.998	1.00/.997	.00/.07
5000	50%	L0	.000/.002	1.00/1.00	1.00/.999	.01/.08
300	50%	L1	.000/.041	1.00/.164	1.00/.163	.05/.93
1000	50%	L1	.000/.039	1.00/.976	1.00/.971	.86/.98
5000	50%	L1	.023/.028	.983/.975	.979/.970	1.0/1.0
300	50%	L2	.000/.059	1.00/.159	1.00/.157	.51/.98
1000	50%	L2	.022/.040	.982/.947	.978/.936	1.0/1.0
5000	50%	L2	.037/.041	.954/.946	.945/.935	1.0/1.0

residual variances for the factors are set to 4, 5, 7. Two residual correlations are used for the data generation $\theta_{13} = Cov(y_1, y_3) = 1.16$ and $\theta_{24} = Cov(y_2, y_4) = 1.1$.

We consider the L2 model misspecification where the residual correlations θ_{13} and θ_{24} are not included in the model and where the direct effect parameter β_{13} is not included as well. Table 3 shows the comparison between the Bayes and the ML fit indices. As in the previous section the fit indices are very close between the two estimators. Table 4 shows the rejection rates for each fit index using the cut off values 0.05 for RMSEA, 0.95 for CFI, and 0.95 for TLI.

In the first row of Table 4 we see the somewhat unusual situation where the model is rejected by the approximate fit indices but is not rejected by the PPP value and the chi-square P-value. When the sample size is 100, the rejection rates for RMSEA/CFI/TLI are higher than the PPP/P-value rejection rates. Such results are contradictory because if a model fits the data exactly it should also fit approximately. Generally we expect the fit indices rejection rates to be lower than that of the PPP/P-value. This problem disappears quickly as we increase the sample size. For sample size of 300 almost all rejection rates climb to near 100% with the exception of the CFI which is 85%. Regardless, the reversal of the rejection rates disappears as the sample size increases to 300 or more. The RMSEA for the baseline model with sample size of 100 is .353. Clearly this issue is not the same issue as the small sample size problem discussed in the previous section, i.e., the problem is not related to the H1, H0 and the baseline model being to close to each other.

One possible way to deal with this rejection rate reversal is to use the confidence intervals of the fit indices instead of their point estimates. For example, when using the CFI criterion we would reject the model if the entire 90% confidence interval falls below the .95 cutoff value. Table 5 contains the rejection rates using the 90% confidence intervals. The rejection rate reversal is now eliminated for CFI, and almost completely eliminated for TLI. For the RMSEA the rejection rate is still higher than that of the PPP but to a smaller extent. Using 0.06 as the cutoff value for the RMSEA improves the situation as well. The rejection rate in that case drops down to 63%. If in addition, we use the 95% confidence interval in the determination of the approximate fit we obtain a rejection rate of 56% which is close to the PPP rejection rate. Regardless of which approach is used, the PPP should take precedence in this situation. If an approximate fit

Table 3: Comparing fit indices: Average values for Bayes/ML

Sample size	RMSEA	CFI	TLI
100	.126/.139	.918/.924	.861/.838
300	.137/.143	.921/.922	.849/.834
1000	.143/.147	.921/.922	.841/.832

Table 4: Rejection rates based on fit indices point estimates: Bayes/ML

Sample size	RMSEA	CFI	TLI	PPP/P-value
100	.97/.98	.79/.69	.90/.91	.40/.80
300	1.0/1.0	.85/.88	.99/1.0	.99/1.0
1000	1.0/1.0	1.0/1.0	1.0/1.0	1.0/1.0

doesn't hold while exact fit appears to hold, we can rely more on the exact fit test. Approximate fit is a concept best suited for larger sample sizes. If a reversal in outcome occurs, the natural explanation is that the sample size is not sufficient to pursue the approximate fit concept. Note also that the rejection rates between Table 4 and 5 did not change substantially for sample size of 300 and 1000. We conclude that the fit indices confidence intervals approach is mostly useful for the small sample size situations. In an Mplus simulation study the confidence intervals can be obtained by specifying the results option in the montecarlo command.

Table 5: Rejection rates based on fit indices 90% confidence intervals: Bayes

Sample size	RMSEA	CFI	TLI	PPP
100	.68	.30	.54	.40
300	.99	.71	.95	.99
1000	1.0	1.0	1.0	1.0

4 Wald test in the Bayesian framework

Consider the situation when two nested models M1 and M2 are estimated and we want to test the hypothesis that the more restricted model, say M1, provides as good of a fit to the data as the less restricted model M2. Suppose that the M1 model is obtained from the M2 model by introducing parameter constraints on the M2 model parameters. With the maximum likelihood estimation this can be done using the LRT test or the Wald test, however, neither of these two tests are available in the Bayesian framework. The available options in the Bayesian framework are the DIC criterion, the BIC criterion and the PPP value. The BIC criterion can also be used to compute the Bayes factor

$$BF_{12} = Exp((BIC_2 - BIC_1)/2), \tag{19}$$

see Wagenmakers (2007). Neither of these methods are universally applicable. The DIC and BIC in many situations have too much variability to provide a reliable answer. In two-level models, for example, all random effects are treated as parameters in Mplus. Precise evaluation of DIC and BIC will require a very large number of MCMC iterations. In addition, DIC and BIC can not be computed for models with categorical variables. The Bayes factor can be computed without replying on BIC, see Hoijtink et al. (2019), but these method are not as simple as equation (19). The main limitation of the PPP method is that it doesn't test the hypothesis of M1 v.s. M2 directly. Instead, it tests the hypothesis of M1 v.s. the unrestricted model and M2 v.s. the unrestricted model. In practical settings, both M1 and M2 might be rejected, or both might not be rejected leaving us without a conclusion on the hypothesis M1 v.s. M2. It is possible to modify the PPP method so that the discrepancy function is the LRT of M1 v.s. M2. This version would have to mimic the algorithm described in Section 2 where both the H0 and the H1 models must be estimated. In this section, however, we describe a Wald test based alternative that is simpler to implement than the PPP, is faster to compute, applies to all modeling situations, and is more powerful.

If the M1 model is obtained from the M2 model by introducing just a single constraint in the parameters, it is possible to directly test the M1 model by estimating the M2 model and testing the validity of the parameter constraint using the M2 parameter estimates. This is typically done in Mplus by introducing a new parameter equal to the

parameter constraint in the Model Constraint command. The parameter constraint is computed in every MCMC iteration. The validity of the parameter constraint is then evaluated through the confidence interval of the constraint. If, however, the M1 model is obtained from the M2 model by introducing multiple parameter constraints, such a process becomes difficult. Each constraint must be evaluated separately and multiple testing issues must be addressed. In the ML framework, the Wald test resolves these issues seamlessly. Because the ML estimator is asymptotically equivalent to the Bayes estimator, we can formulate an equivalent Bayesian version of the Wald test. Here we describe this test and evaluate its performance through simulation studies.

Suppose that the M1 model is obtained from the M2 model by introducing the following L constraints

$$0 = W_1(\theta_2) \tag{20}$$

...

$$0 = W_L(\theta_2) \tag{21}$$

where θ_2 represents the M2 model parameters. To test the validity of the above constraints we estimate the M2 model. Let $\theta_{2,i}$ represent the $i-$th draw from the θ_2 posterior distribution. These typically are obtained from the second half draws of the MCMC estimation of M2. We then compute the vectors $W_i = (W_1(\theta_{2,i}), ..., W_L(\theta_{2,i}))$. Let $\overline{W}$ represent the sample mean of W_i and let S represent the sample variance covariance matrix. Under the null hypothesis that the model constraint equations (20-21) are correct,

$$T = \overline{W} S^{-1} \overline{W}^T \tag{22}$$

will asymptotically have a chi-square distribution with L degrees of freedom just as this is so in the ML framework. We can then use T as our test statistic and produce a P-value similar to how this is done in the ML framework. The Mplus implementation of this testing procedure is exactly the same as in the ML framework, i.e., we use the *Model Test* command to define the model constraints that need to be tested.

We illustrate the performance of the Bayes Wald test with the following two-level CFA model. The model has 4 observed variables

Table 6: Type I error for Wald test: rejection rates

C	Bayes	ML
100	.10	.11
200	.05	.04

Y_i, $i = 1, ..., 4$ measuring one factor on the within level and one factor on the between level and is given by the following equations

$$Y_i = Y_{w,i} + Y_{b,i} \tag{23}$$

$$Y_{w,i} = \lambda_{w,i}\eta_w + \varepsilon_{w,i} \tag{24}$$

$$Y_{b,i} = \mu_i + \lambda_{b,i}\eta_b + \varepsilon_{b,i}. \tag{25}$$

All loading parameters are set to 1, the within level residual variances are set to 1 on the within level and to 0.3 on the between level, all means are set to 0, the factor variance is set to 1 on the within level and to 0.4 on the between level. In this example, the model M2 is the same as the generating model. The model M1 is the model where the loadings on the within level are held equal to the loadings on the between level. This model is of interest because it implies that the between level factor can be interpreted as the random intercept of the total factor represented by the sum of the within and the between factor. If the loadings are not held equal, such an interpretation is problematic. Since the first loading is held fixed to 1 for identification purposes, the model constraint equations that need to be tested are as follows

$$\lambda_{w,2} = \lambda_{b,2} \tag{26}$$

$$\lambda_{w,3} = \lambda_{b,3} \tag{27}$$

$$\lambda_{w,4} = \lambda_{b,4}. \tag{28}$$

We generate 100 data sets with C clusters each of size 20, where $C = 100$ or $C = 200$. We analyze the data with the ML and the Bayes estimators using model M2 and we compute the Wald test for the above hypothesis. The results of this simulation study are presented in Table 6. We see that the type I errors are comparable for the two estimators. The model rejection rates are near the nominal level of 0.05.

Table 7: Power analysis for Wald test: rejection rates

C	Bayes	ML
100	.22	.22
200	.44	.48

Next we consider a simulation study to evaluate the power of the Wald test for the ML and the Bayes estimators. To do that we generate the data using unequal between and within loadings. We set $\lambda_{b,3} = \lambda_{b,4} = 1.2$ while the rest of the loadings are set at 1. The results of this simulation study are presented in Table 7. We can see that the power of the Wald test is comparable for the two estimators.

It is interesting to point out here that neither the PPP nor the DIC methods are able to provide an alternative. Consider the case when the loadings are equal. While the DIC is smaller for model M1 on average, when the DICs are compared within each replication, the DIC picks the correct M1 model only 56% of the time. This is due to the large variability of the DIC. Because of the large number of estimated parameters, which includes all random effects, it takes a very long MCMC sequence to reduce the variability and obtain more accurate DIC estimates. The PPP method is also unable to provide efficient testing. The PPP has 0% rejection rates in both cases: when the loadings are equal or when they are unequal. This is due to the lower power associated with the PPP method.

5 Building approximately well fitting BSEM models

In this section we illustrate how the approximate fit indices can be used to obtain approximately well fitting BSEM models. These kinds of models are discussed in detail in Muthén and Asparouhov (2012) as well as Asparouhov et al. (2015). The core of this modeling strategy is that when a model does not fit well we can add a multitude of additional parameters to the model. To preserve the original structural model we assign tiny priors to all the additional parameter, for example $N(0, 0.00001)$. If some of the additional parameters are needed to improve the model fit during the estimation, they will be able to

escape the tiny prior and indeed help with the model fit. The rest of the parameters that do not improve the model fit will remain near zero. We can then separate the parameters that escape the tiny priors and potentially add those to the original model or simply retain the BSEM model with the tiny priors.

The PPP plays a vital role in the BSEM modeling. It is used to evaluate the fit of the original model as well as the fit of the BSEM model based on the tiny priors. The PPP is used to determine the size of the tiny priors as well. To build approximately well fitting BSEM models, as an alternative to the perfectly fitting BSEM models, we simply replace the PPP role in this process with the approximate fit indices. The final outcome of this modeling strategy is that we can obtain a BSEM model that more closely resembles the substantively drawn model. This would be particularly useful when the sample size is large and many small deviations between the data and the model become significant. In such situations, using the PPP as a criterion may result in bigger priors and many more added parameters, both of which are undesirable.

We illustrate this process with the following simulated example. Consider a two-factor analysis model where each factor is measured by 7 indicators for a total of 14 dependent variables. The first seven variables measure the first factor while the last seven variables measure the second factor. We set all intercept parameters to 0, all loading parameters to 1, all residual variances to 1, all factor variances to 1 and the factor covariance to 0. In addition, we introduce 4 cross loadings $\lambda_{12} = \lambda_{22} = \lambda_{81} = \lambda_{91} = 0.3$ and 4 residual covariances between the indicators $\theta_{4,5} = \theta_{6,7} = \theta_{11,12} = \theta_{13,14} = 0.15$. Using this model we generate a data set of size $N = 5000$.

The first step in the estimation is to evaluate the model excluding any cross loadings or residual covariances. The outcome of that estimation is as follows: $PPP = 0$, $RMSEA = 0.063$, $CFI = 0.948$, and $TLI = 0.938$. Clearly the model fits fairly well but not well enough to claim exact fit or an approximate fit. Next we add to the model all cross-loadings, a total of 14, with a tiny prior $N(0, v)$. Without the tiny priors, such a model will be of course unidentified. When the tiny priors are introduced, however, all cross-loadings can be estimated. Table 8 contains the fit measure results for several values of v. We can see that the fit of the model improves as we increase v but the improvements are minimal after $v = 0.001$. We therefore select this as our approximately well fitting BSEM model. We pick

Table 8: Fit of BSEM model based on different tiny priors

v	PPP	RMSEA	CFI	TLI
0.00001	.000	.060	.951	.944
0.0001	.000	.050	.968	.961
0.001	.000	.036	.986	.980
0.01	.000	.034	.987	.982
0.1	.000	.034	.987	.982

the best fitting model with the smallest v. Choosing the smallest v is important as that will minimize the size of the cross loadings. All three approximate fit indices suggest that this is a well fitting model, while the PPP still rejects the model primarily because the large sample size and the fact that some of the covariances are not fitted well enough.

We can choose to retain the above BSEM model as our model of approximate fit or we can go further to analyze the cross loadings and retain only those that are needed. This way we convert the BSEM model to a standard SEM model that is not based on tiny priors but uses non-informative priors. The sizes of the estimated cross-loadings in the BSEM model in order of magnitude are as follows:. 19, .18, .18, .16, .10, ... We can include these large cross-loadings to the original SEM model in order of magnitude (or statistical significance) one at a time. First we estimate the model with the largest cross-loading, then with the largest 2, etc. Table 9 reports the fit of the model as we include the top L cross-loadings. The model fit improves as we add up to 4 loadings and it doesn't improve when we add the fifth. This clearly indicates that 4 cross-loadings is the right choice, confirming the model that generated the data. Note also that the SEM model with 4 cross-loadings matched the fit of the BSEM model with all of the cross-loadings. We conclude that we have extracted all of the possible fit improvement with just those 4 cross-loadings.

In this process we obtained an approximately well fitting BSEM and SEM models without having to pursue acceptable PPP value. We ignored the small residual correlations in the data, which are the reason for the PPP rejections. We retained the main structure of the CFA model and improved the model fit with the addition of the small cross-loadings. In conclusion, this illustration shows how the

Table 9: Fit of SEM model based on different number of cross-loadings

L	PPP	RMSEA	CFI	TLI
1	.000	.057	.958	.948
2	.000	.052	.966	.958
3	.000	.043	.977	.971
4	.000	.033	.987	.983
5	.000	.033	.987	.983

approximate fit indices can be used to extend the BSEM methodology to the framework of approximate fit.

6 Discussion

In this note we demonstrate the advantages of the new model fit methods implemented in Mplus 8.4. Approximate fit indices provide a valuable tool in the Bayesian framework. Some challenges remain, however. Using approximate fit indices in small sample size situations is not recommended. The exact fit methods should be preferred instead. The problem is that it is unclear at what level of sample size the switch between these two methodologies should occur. It is unclear how that level of sample size depends on the complexity of the model. We described two distinct situations where approximate fit indices fail for small sample sizes. The first one is when the distance between the baseline and the H1 model is too small. The second situation is when we have rejection rate reversals. Clearly further research is needed on this topic. Our simulations show that these problems are not related only to the Bayes estimator but also to the ML estimator. Maydeu-Olivares et al. (2018) focus their research on the confidence limits of the fit indices. In our simulations, such an approach resolved the rejection rate reversal problem to some extent. Mplus 8.4 simulation studies now include summary results also for the CFI and TLI indices for all estimators, as well as their confidence limits obtained with the Bayes estimator. This new feature could possibly facilitate further research or could enlighten real data applications.

References

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

[2] Asparouhov, T. & Muthén, B. (2010b). Bayesian analysis of latent variable models using Mplus http://www.statmodel.com/download/BayesAdvantages18.pdf

[3] Asparouhov, T., Muthén, B., & Morin, A. (2015). Bayesian Structural equation modeling with cross-loadings and residual covariances. Journal of Management, 41, 1561-1577.

[4] Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen, & J. S. Long, Testing structural equation models. Newbury Park, CA: Sage.

[5] Garnier-Villarreal, M., & Jorgensen, T. D. (2019). Adapting fit indices for Bayesian structural equation modeling: Comparison to maximum likelihood. Psychological Methods.

[6] Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B.(2004) Bayesian Data Analysis. London, Chapman & Hall.

[7] Hjort, N. L., Dahl, F. A. and Steinbakk, G. H. (2006). Post-processing posterior predictive p-values. J. Amer. Statist. Assoc. 101, 1157-1174.

[8] Hoijtink, H., Mulder, J., van Lissa, C., & Gu, X. (2019). A tutorial on testing hypotheses using the Bayes factor. Psychological Methods.

[9] Hoofs, H., van de Schoot, R., Jansen, N. W. H., & Kant, Ij. (2017). Evaluating Model Fit in Bayesian Confirmatory Factor Analysis With Large Samples: Simulation Study Introducing the BRMSEA. Educational and Psychological Measurement, 78 (4), 537-568.

[10] Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling, 6(1), 1-55.

[11] Kenny, D. (2015). Measuring Model Fit. http://www.davidakenny.net/cm/fit.htm

[12] Little. R. J. A,. and D. B. Rubin. (1987). Statistical analysis with missing data. Wiley. New York. New York. USA.

[13] Maydeu-Olivares, A., Shi, D., & Rosseel, Y. (2018). Assessing fit in structural equation models: A Monte-Carlo evaluation of RMSEA versus SRMR confidence intervals and tests of close fit. Structural Equation Modeling: A Multidisciplinary Journal, 25, 389-402.

[14] Merkle E.C., & Rosseel Y. (2018). blavaan: Bayesian Structural Equation Models via Parameter Expansion. Journal of Statistical Software, 85(4), 1–30.

[15] Muthén, B., & Asparouhov, T. (2012). Bayesian SEM: A more flexible representation of substantive theory. Psychological Methods, 17, 313-335.

[16] Wagenmakers, E.J. (2007). A practical solution to the pervasive problems ofp values. Psychonomic Bulletin & Review, 14(5), 779–804.