Prior-Posterior Predictive P-values

Tihomir Asparouhov and Bengt Muthén

Mplus Web Notes: No. 22

Version 2

April 27, 2017

1 Introduction

The Bayesian SEM introduced in Muthén and Asparouhov (2012) uses small-variance priors to expand standard SEM models into a more flexible and more realistic set of models. When a hypothesized SEM model is rejected by the data, a BSEM model can relax the rigid framework of the SEM model by adding small-variance priors to parameters that are fixed to zero in the SEM model. That way the conceptual framework of the SEM model is preserved, and by allowing parameters to be approximately zero instead of being fixed to zero the discrepancy between the SEM model and the data can be resolved. In this process the BSEM model can also parse out meaningful model misspecifications from small model misspecifications that can also be the cause of model rejections when such small misspecifications are in great number or when the sample size is so large that even small misspecifications are enough to reject the model.

The BSEM model estimation typically requires multiple model estimations, see Asparouhov, Muthén and Morin (2015), varying the size of the variance of the small priors and using the posterior predictive p-value (PPP) to monitor the distance between the data and the model. In this process no particular prior variance is preferred, rather, the prior variance is adjusted gradually to maintain identifiability of the model while resolving model fit and separating parameters that have minor deviations from zero from substantively important misspecifications. The choice of the variance of the small prior is further complicated by the fact that they are sample size dependent, i.e., a prior variance that works for one data set might not be appropriate for similar data set of much larger sample size. This is because in the Bayesian methodology as the sample size increases to infinity the prior influence is lost. To resolve that problem the BSEM estimation requires the prior to be set with much smaller variance to overcome the effect of the data. Typically the BSEM estimation starts with a very small prior variance that is guaranteed to produce a well identified model close in meaning and parameter estimates to the original SEM model and gradually increase the prior variance until the estimation no longer converges or the PPP value no longer improves. The PPP values used with the BSEM model estimation is based on the chi-square discrepancy function, see Scheines, Hoijtink and Boomsma (1999) and Asparouhov and Muthén (2010).

Recently Hoijtink and van de Schoot (2017) point out that there is a need for a different type of hypothesis testing than the one that the PPP

method provides which attaches a practical meaning to the prior variance value. Consider a model M that has a set of major parameters θ_1 and a set of minor parameters θ_2. If the θ_2 parameters are fixed to zero the data rejects the model and thus we are interested in estimating a BSEM model where the θ_2 parameters are not-exactly zero but approximately zero, i.e., within small deviations of zero. More specifically we want to test the specific hypothesis that the θ_2 parameters are within a small range around 0, provided by the normal distribution $N(0, v)$ where v is a small variance, not any small variance but a particular value of v that has some practical meaning. The PPP testing in the BSEM estimation does not provide an answer for that hypothesis as it generally does not attach a specific meaning to the small-variance prior and the small-variance prior is not sample size independent. Hoijtink and van de Schoot (2017) construct a method called the prior posterior predictive p-value (PPPP) that accomplishes exactly that in the simple case of a linear regression with one dependent variable and one predictor. In this note we show how a slight modification of Hoijtink and van de Schoot (2017) PPPP approach allows us to generalize the PPPP method for general SEM models. In what follows we discuss the details of that modification, the Mplus Version 8 implementation of the PPPP, and illustrate the performance of the method with several simulation studies and empirical examples.

2 The prior posterior predictive p-value

Let's again consider a model M with a set of major parameters θ_1 and a set of minor parameters θ_2 where we want to test the hypothesis that the minor parameters can be considered approximately zero within the range of a $N(0, v)$ distribution where v is a particular small variance. We use the following notation $M(\theta_1, \theta_2)$ to represent the model M with parameter values θ_1 and θ_2. We call the pure model the model where the minor parameters are set to 0, i.e., $M(\theta_1, \theta_2 = 0)$. Let θ denote the vector of all model parameters $\theta = (\theta_1, \theta_2)$ and $G(Y, \theta_1, \theta_2)$ denote the chi-quare for the model $M(\theta_1, \theta_2)$ and the data Y.

First let's note that the PPP in a general Bayesian estimation can be computed using any discrepancy function, not just the standard structural equation model chi-square discrepancy function. A general discrepancy function is defined as a function of the data and the model parameters. Let $f(Y, \theta)$ be a general discrepancy function. The PPP for this discrepancy function is

defined as the

$$PPP = P(f(Y, \theta) < f(\tilde{Y}, \theta)) \tag{1}$$

where the probability is computed over the posterior distribution of θ given the prior of θ and observed data and $\tilde{Y}$ is generated data assuming the model and the θ parameters. For a general discussion on posterior predictive checking see Gelman et al. (2004). The discrepancy function generally is a function that measures the distance between the data and the model. It can be chosen for example to be the squared difference between the sample and the model-implied means.

In what follows we define the discrepancy function to be the distance between the data and the pure model, i.e., we define

$$f(Y, \theta_1, \theta_2) = G(Y, \theta_1, 0) \tag{2}$$

Let $F(\theta_2)$ denote the PPP value using the above discrepancy function for the Bayesian model estimation of model M where the minor parameters are fixed to the values θ_2 while the parameters θ_1 are set free. The function $F(\theta_2)$ essentially compares the distance between the observed data and the pure model and the distance between the pure model and data that are θ_2 contaminated, i.e., data that originate from the $M(\theta_1, \theta_2)$ model, rather than the pure $M(\theta_1, 0)$ model.

We define the prior posterior predictive p-values as

$$PPPP = E(F(\theta_2)) \tag{3}$$

where the expectation is taken over the $N(0, v)$ distribution.

Hoijtink and van de Schoot (2017) illustrate this concept through a simple example rather than define the PPPP as a general concept. However, there is a slight difference in their definition and the definition given above. If we replace equation (2) with the following equation

$$f(Y, \theta_1, \theta_2) = G(Y, \hat{\theta_1}, 0) \tag{4}$$

where $\hat{\theta_1}$ denotes the maximum-likelihood estimate for the θ_1 parameter obtained from the data Y for the pure model $M(\theta_1, 0)$ then we arrive at the definition used by Hoijtink and van de Schoot (2017) for the simple regression example. This difference amounts to how the distance is measured between the data and the pure model. Our definition uses the current value of θ_1 while

the Hoijtink and van de Schoot (2017) definition uses the maximum likelihood value of θ_1. Since it is impractical to compute the maximum-likelihood value at each MCMC iteration we resort to the simplification in equation (2) and the simulation studies shown below demonstrate that the performance of this method is similar to the performance of the Hoijtink and van de Schoot (2017) method.

3 The Mplus implementation of the PPPP

The PPPP in Mplus version 8 is computed for every model for which the PPP is computed. To trigger this computation a small-variance normal prior has to be specified for a parameter which is a slope, a loading or an intercept parameter. A normal prior with variance smaller than 1 is considered a small-variance prior. The computation of the PPPP is performed before the general Bayes estimation, i.e., two separate MCMC sequences are conducted. The first one results in the computation of the PPPP value while the second MCMC chain results in the computation of the PPP value and model parameter estimates. Thus the PPPP methodology can be used simultaneously with the BSEM model estimation.

Following the Hoijtink and van de Schoot (2017) algorithmic descriptions, the computation of the PPPP amounts to two modifications of the PPP computation. The first one is that the θ_2 parameters are generated at each iteration from their prior distribution rather than the posterior distribution. The second modification is the switch in the discrepancy function, where the chi-square function is replaced by the function in equation (2) measuring the distance between the data and the pure model (where the standard PPP would use the distance between the data and the contaminated model).

To compute $F(\theta_2)$ in equation (3) for a particular draw of θ_2, a complete MCMC sequence has to be run to convergence for that particular value of θ_2. Hoijtink and van de Schoot (2017) instead suggest that just one MCMC iteration is conducted for each θ_2 draw and point out that results remain unchanged when compared to the case of much longer MCMC sequence for each value of θ_2. Ultimately this amounts to updating θ_2 at each iteration or updating it every k-th iteration assuming the MCMC sequence for a particular value of θ_2 converges in k iterations. The Mplus implementation of this is as follows. Using the thin command a specification of thin=k will cause the θ_2 parameters to be updated every k-th iteration. Note that the Mplus

default for the thin option is 1.

It is important to note that any prior that is not considered a small-variance prior for the purpose of the computation of the PPPP is not included in the θ_2 vector. Only those parameters that are given small-variance priors in the model prior statement are included in the θ_2 vector. Note that the diff priors are NOT included in the θ_2 vector either, but that may change in the future.

Note that the PSR convergence criterion is used for both MCMC runs, the PPP run and the PPPP run. The MCMC chain for the PPPP computation is technically not a proper Bayes MCMC chain but rather a random Bayesian averaging of MCMC chains. Nevertheless the PSR can still be used to monitor convergence. Many of the BSEM models where small-variance priors are used are actually unidentified models that are used for exploratory purposes. In many cases the BSEM estimation for such models will not converge due to the prior not being strict enough. If the BSEM model does not converge the PPPP will not be reported.

In other cases the PPPP run will not converge but the PPP run will. The PPPP is still computed in these cases from the MCMC chain sequence. Generally speaking the PPPP can be computed for unidentified models as it is only based on model distance and doesn't require the Bayesian mixed MCMC chains to be converging and the model to be identified. However, it should be fairly unusual that the more relaxed PPP run converges while the PPPP run does not. We have found that using the thin command can improve the convergence rate for the PPPP run. It is possible to prefix the number of MCMC iterations run for both the PPP and the PPPP MCMC chains, using the FBITER option, and deal with convergence separately.

4 Simulation Study: Regression Analysis

In this section we will replicate the simulation study results given in Table 5 in Hoijtink and van de Schoot (2017). The model is a simple regression example

$$Y = \alpha + \beta X + \varepsilon. \tag{5}$$

We are interested in testing the hypothesis that β is approximately zero. We specify a small-variance prior $N(0, 0.01)$ for the β parameter. We generate data according to the above model and various values of β to evaluate the performance of the PPPP. The covariate X is generated from a standard

Table 1: Simulation results for linear regression example: average PPPP (percent rejection)

β	0	0.1	0.2	0.3	0.707
N=20	.49(.01)	.46(.04)	.39(.09)	.32(.15)	.02(.90)
N=50	.53(.03)	.51(.04)	.34(.13)	.19(.35)	.00(1.00)
N=100	.58(.00)	.47(.07)	.25(.22)	.08(.55)	.00(1.00)

normal distribution. In this simulation study we generate the data using the following parameters $\alpha = 0$ and $Var(\varepsilon) = 1 - \beta^2$, so that $Var(Y) = 1$. Data is generated for several values of $\beta = 0, 0.1, 0.2, 0.3, 0.707$. We generate 100 samples of sizes 20, 50, and 100 for each value of β. Table 1 shows the results of this simulation study. We report the average PPPP value across the 100 replications as well as the percentage rejection of the hypothesis that β is approximately zero at the 5% nominal level. Table 5 in Hoijtink and van de Schoot (2017) is based on a single data set rather than the average across 100 replications so it is not possible to directly compare the results, however, the results are fairly close and certainly the pattern is the same.

The most importantly result from Table 1 is the rejection rate for $\beta = 0.3$. The PPPP rejection rate is .15, .35 and .55, i.e., they increase. The value $\beta = 0.3$ is outside the range of the normal prior $N(0, 0.01)$ and therefore we want the test to reject the hypothesis. The PPPP rejection rates increase as the sample size increases just as a classic p-value would do. In that same case PPP value rejection rates are .04, .06 and .00, i.e., as the sample size increases the rejection rates decrease. This is counter intuitive as we would expect that with an increase of the sample size and the information in the data, the hypothesis testing would reach the right conclusion. To summarize again the PPP value can not be used to test the hypothesis that a parameter is approximately zero and it should be used within the BSEM framework as it is outlined in Asparouhov, Muthén and Morin (2015). On the other hand the PPPP value can be used to test that hypothesis and appears to behave similar to the classic p-value when it comes to sample size. The advantage of the PPPP value over the classic p-value based tests is that it allows the test of an approximate hypothesis where minor deviations from the hypothesis are not a reason to reject it.

5 Simulation Study: Factor Analysis

In this section we consider the factor analysis simulation study discussed in Hoijtink and van de Schoot (2017) and we will compute the PPPP values for the two hypotheses considered in that article. The factor analysis model is a 2 factor model with 6 indicators, for $p = 1, ..., 6$

$$Y_p = \nu_p + \lambda_{1p}\eta_1 + \lambda_{2p}\eta_2 + \varepsilon_p. \tag{6}$$

We generate the data using the following parameter values $\nu_p = 0$, $\theta_p = Var(\varepsilon_p) = .35$ for $p = 1, 3, 4$, and 6, and $\theta_p = .51$ for $p = 2$, and 5. The loading matrix Λ is given by

$$\begin{bmatrix} 0.7 & -0.4 \\ 0.7 & 0 \\ 0.7 & 0.4 \\ -0.4 & 0.7 \\ 0 & 0.7 \\ 0.4 & 0.7 \end{bmatrix}. \tag{7}$$

The two hypotheses that were considered in Hoijtink and van de Schoot (2017) are

$$H1 : \lambda_{15}, \lambda_{22} \sim N(0, 0.01)$$

and

$$H2 : \lambda_{14}, \lambda_{15}, \lambda_{16}, \lambda_{21}, \lambda_{21}, \lambda_{23} \sim N(0, 0.01).$$

We expect hypothesis H1 to not be rejected and hypothesis H2 to be rejected as the value 0.4 is outside of the range of $N(0, 0.01)$. The Hoijtink and van de Schoot (2017) critique of the PPP method is that the PPP values do not reject either hypothesis as sample size increases, which is what one would expect as a prefixed prior has no effect on the model estimation, including test of fit, for sufficiently large sample size. The BSEM methodology relies on reducing the variance of the prior as sample size increases, but here we are holding it fixed as the above hypotheses require.

We also include the following two hypotheses for further illustration purposes using a larger prior variance of 0.1

$$H3 : \lambda_{14}, \lambda_{15}, \lambda_{16}, \lambda_{21}, \lambda_{21}, \lambda_{23} \sim N(0, 0.1).$$

and

$$H4 : \lambda_{14} \sim N(0, 0.1), \lambda_{15}, \lambda_{21} = 0.$$

Table 2: Simulation results for factor analysis example: average PPPP (percent rejection)

N	H1	H2	H3	H4	H5
50	.42(.02)	.01(.93)	.11(.22)	.32(.04)	.45(.02)
100	.51(.00)	.00(1.00)	.06(.52)	.32(.01)	.45(.02)
500	.73(.00)	.00(1.00)	.03(.83)	.26(.01)	.37(.01)
1000	.80(.00)	.00(1.00)	.03(.92)	.27(.00)	.33(.00)
5000	.88(.00)	.00(1.00)	.02(.99)	.26(.00)	.34(.00)

Note that hypothesis H4 can be implemented in Mplus in two ways both of which are equivalent. The parameters λ_{15}, λ_{21} can be treated as parameters fixed to 0, or they can be treated as parameters having $N(0, 0)$ prior. The second approach will essentially include them in the PPPP testing, however, the two approaches are equivalent and yield identical results.

Finally, we include the hypothesis H5, which is identical to H2 but we generate the data using this loading matrix

$$\begin{bmatrix} 0.7 & -0.1 \\ 0.7 & 0 \\ 0.7 & 0.1 \\ -0.1 & 0.7 \\ 0 & 0.7 \\ 0.1 & 0.7 \end{bmatrix}. \tag{8}$$

Because the 0.1 cross-loadings are within the $N(0, 0.01)$ we expect hypothesis H5 to not be rejected.

The results of this simulation are given in Table 2 for various sample sizes and are computed over 100 replications for each sample size. Note that Table 2 corresponds to Table 4 in Hoijtink and van de Schoot (2017), where only the PPP values are computed, as they did not include PPPP value computation beyond the linear regression example.

Clearly the PPPP resolves the issue with the PPP and the PPPP behavior is similar to the classic p-value behavior when it comes to sample size. The PPPP value decreases as the sample size increases. Table 2 results show that hypothesis H1 is not rejected while hypothesis H2 is rejected confirming our expectations. Next we compare hypotheses H3 and H4. Hypothesis H4 is not

rejected which is in line with our expectations that the true value of -0.4 is within the range of the prior $N(0, 0.1)$. In addition, the PPPP value appears to be close to the probability that a value as large as 0.4 by absolute value will be drawn from the small-variance prior $N(0, 0.1)$, which is $\Phi(0.4/\sqrt{.1}) \approx .2$, where Φ is the standard normal function. On the other hand the hypothesis H3 is rejected. The interpretation of that is as follows. The 6 cross-loading values that we tested are -0.4, -0.4, 0, 0, 0.4, 0.4. While one 0.4 value, by absolute value, can occur as the draw from the $N(0, 0.1)$ distribution, 4 out of 6 is not likely. The probability that 4 or more will be greater or equal than 0.4 by absolute value is approximately $.2^4.8^2\binom{6}{4} + .2^5.8\binom{6}{5} + .2^6 \approx 0.017$ which is exactly what the PPPP value converges to as sample size increases. This means that at the 5% nominal level we should expect H3 to be rejected and indeed the PPPP results confirm that.

Finally we consider hypothesis H5. In this case the PPPP correctly does not reject the hypothesis and it concludes that the cross-loading values -0.1, -0.1, 0, 0, 0.1, 0.1 are approximately zero and can be assumed to be white noise parameters coming from the small-variance distribution $N(0, 0.01)$. The fact that hypothesis H5 is not rejected is the real strength of the PPPP method and it is where the real advantage of this method can be seen very clearly.

6 The difference between PPPP and PPP

The PPP is a test of model fit. It tests the fit to the data of the model with the given priors and is based on comparing the model with the unrestricted mean and variance covariance model.

The PPPP is not a test of model fit. It is a test for the minor parameters θ_2 in the model. If the test does not reject, this should NOT be interpreted as a test of model fit result, i.e., it should NOT be interpreted as evidence that the model fits the data. The PPPP is more similar, but not equivalent, to the Wald test implemented in Mplus with the MODEL TEST command, that is, a test for specific parameters.

The proper interpretation of the PPPP is as follows. If the test does not reject, the minor parameters can be assumed to come from $N(0, v)$ distribution. More broadly speaking, if the PPPP does not reject, that means that there is no evidence in the data for the minor parameters in model $M(\theta_1, \theta_2)$ to be outside the $N(0, v)$ distribution. The PPPP does not consider the fact that another set of parameters θ_3 (minor or major) that are missing from

Table 3: Simulation results for misspecified factor analysis example: average value (percent rejection)

N	PPP	PPPP
50	.28(.13)	.33(.07)
100	.28(.08)	.39(.00)
500	.02(.89)	.28(.01)
1000	.00(1.00)	.26(.00)
5000	.00(1.00)	.24(.00)

the model $M(\theta_1, \theta_2)$ may not be zero and a reason for the model to be inadequate for this data. If such a parameter θ_3 indeed exists then the PPP will reject the model correctly while the PPPP will not reject. Note that it is not correct to say that the PPPP incorrectly does not reject the model, rather, the proper interpretation is that it does not reject the hypothesis that the θ_2 parameters are approximately zero. The PPPP does not make inference about other model misspecifications that can not be resolved by the θ_2 parameters being tested.

We will illustrate this point with the following simulation study. Consider again the factor analysis example and data generation for hypothesis H1. We modify the data generation by including a residual covariance between Y_1 and Y_2 and we set that to 0.3. Using multiple sample sizes N and 100 replicated data sets for each sample size we obtain the average values and rejection rates for PPP and PPPP. The results are reported in Table 3. The PPP correctly rejects the model for sample size $N = 500, 1000$ and 5000. For smaller sample sizes there is not enough power to reject the model and establish significance of the misfit. On the other hand the PPPP never rejects. That is because it is not a test of fit for the model. It is only a test for the λ_{15} and λ_{22} parameters. These loadings are unrelated to the residual correlation between Y_1 and Y_2 where the misfit occurs, i.e., there is nothing in the data that will cause these loadings to be estimated as larger non-zero cross-loadings and thus the PPPP does not reject.

In the above simulation study we illustrated the fact that the PPPP does not replace the PPP. The PPPP is a targeted test for a set of minor parameters θ_2 and a large p-value does not guarantee model fit. The PPP is a test of model fit. The PPPP is a test for the hypothesis that the minor

parameters θ_2 are approximately zero, while the PPP is a test for the hypothesis that the structural model $M(\theta_1, \theta_2)$, where the θ_2 parameters are approximately zero, fits as well as the unrestricted mean and variance covariance model. The two hypotheses that the PPP and the PPPP address could not be more different. However, the hypotheses become equivalent in some special circumstance such as, for example, the case where the model $M(\theta_1, \theta_2)$ is the same as or equivalent to the unrestricted mean and variance covariance model. For example this is the case of the regression example, but it is not the case for the factor analysis example. If the minor parameters θ_2 include also all the residual covariance parameters in the factor analysis example this would also be the case of equivalent hypotheses. Note that with the current Mplus implementation only the intercept, slopes, and loadings parameters can be included in the PPPP test.

The above simulation study is an example where the PPP rejects while the PPPP does not not reject. The simulation study for hypothesis H2 in the previous section is an example of the opposite where the PPP does not reject while the PPPP rejects. We did not include the PPP values for that simulation study but these can be found in Hoijtink and van de Schoot (2017). The PPP does not reject in that case because for sufficiently large sample size the fixed prior $N(0, v)$ is not strict enough to have an effect on the estimation, which is dominates by the data, and any size cross-loadings are allowed. As a result the estimated BSEM model becomes identical to the data generating model and thus is not rejected. When the PPP and the PPPP yield opposite conclusions proper interpretation is critical. Clearly there is substantial room for misuse if these tests are interpreted incorrectly.

A proper utilization of the PPP and the PPPP would still involve first fitting a model via the BSEM approach of Muthén and Asparouhov (2012) that relies primarily on PPP for model evaluation. Once the model is fitted and the PPP value does not reject the model we can separate the parameters into major parameters θ_1 and minor parameters θ_2 and test for the size of the minor parameters using the PPPP test. Note that this may require a separate run. The PPP prior variance value does not need to correspond to the PPPP prior variance value. The PPP prior variance is guided by the BSEM fitting process described in Asparouhov, Muthén and Morin (2015) while the PPPP prior variance is guided by what is considered substantively different from zero. In the next section we illustrate the proper use of the PPP and the PPPP using empirical examples.

7 Using the PPP and the PPPP with empirical examples

In this section we revisit the Holzinger-Swineford mental abilities example discussed in Muthén and Asparouhov (2012). The example consists of a four-factor model where the factors are measured by 19 indicators. The sample consists of two groups obtained from two separate schools: Grant-White school (N = 145) and the Pasteur school (N = 156). A simple structure for the factor loadings is rejected by the data in both groups and thus we consider the model which includes a total of $19 \cdot 3 = 57$ cross-loading with small-variance prior of $N(0, 0.01)$.

The PPP and the PPPP of these analyses are presented in Table 4 for the two groups. No rejections are observed in either of the groups. The proper interpretation of these results is as follows. The fact that the PPP does not reject the model means that a four-factor model with some minor cross-loadings fits the data well. Detailed analysis of the cross-loading results, obtained from the posterior distribution in the BSEM estimation, can reveal if the cross-loadings should be considered minor or major. The PPP doesn't carry such information. It only asserts that generally the cross loadings are a sufficient model modification to resolve the model misfit of the simple four-factor analysis structure and there is no need to add further modifications such as adding an additional factor or adding additional residual covariances between for the indicator variables. The fact that the PPPP test does not reject means that the cross loadings can be considered approximately zero and coming from a normal distribution $N(0, 0.01)$. Clearly the PPPP and the PPP implications for the model are completely different. Note also that the PPP values are smaller in both groups. This is because the PPP tests a much stricter hypothesis than the PPPP. The PPP tests the hypothesis that the cross-loadings are small and the residual covariances are zero, while the PPPP tests only that the cross-loadings are small, although the meaning of small is much stricter than it is for the PPP.

Now we repeat the above analysis on a sample that is created by doubling the original sample, i.e., every observation is entered twice. This is an artificially created data set and the analysis has no practical implications. We only use it here to illustrate the proper use of the PPP and the PPPP in empirical examples. The results of this analysis is presented in Table 5 when using the small-variance prior $N(0, 0.01)$ and in Table 6 when using

Table 4: Holzinger-Swineford four-factor model

school	PPP	PPPP
Grant-White	.36	.52
Pasteur	.16	.25

Table 5: Holzinger-Swineford four-factor model with doubled sample and prior N(0,0.01)

school	PPP	PPPP
Grant-White	.00	.02
Pasteur	.00	.00

the small-variance prior $N(0, 0.1)$. The PPP results in both tables indicate that in these larger samples there is enough evidence to reject the four-factor model and additional modifications are needed. The potential modifications that can be explored are adding an additional factor or adding additional indicator residual covariances. In addition the PPP confidence interval values for the difference of the chi-square values for the observed and the replicated data changed only slightly between the two prior variances, which indicates that much of the model fit improvement is already gained at these prior levels and no further gains can be expected.

The PPPP test rejects the hypothesis of $N(0, 0.01)$ cross-loadings but does not reject the hypothesis of $N(0, 0.1)$ cross-loadings. Overall the conclusion is that allowing the cross-loadings to be a bit bigger will get us closer to fitting the data but it will not be enough. This analysis again illustrates the complementary nature of the two methods if interpreted correctly.

8 Conclusion

Hoijtink and van de Schoot (2017) proposed a new method for testing a hypothesis that a set of parameters are approximately zero within the BSEM framework. They defined the PPPP method only for simple regression models, however. In this note we generalized the method to the SEM modeling

Table 6: Holzinger-Swineford four-factor model with doubled sample and prior N(0,0.1)

school	PPP	PPPP
Grant-White	.00	1.00
Pasteur	.00	1.00

framework and illustrated its performance with simulation studies and empirical examples.

We also clarified the difference between the PPPP and the PPP methods. The PPP method is a test of fit, while the PPPP method is not a test of fit. The PPPP method is a test for the hypothesis that a set of parameters are approximately zero. Proper interpretation is essential for both the PPPP and the PPP values. The two methods are not a contradiction of each other, even when the two values are not the same. The PPP value should be strictly used as it is outlined in Muthén and Asparouhov (2012) and Asparouhov, Muthén and Morin (2015). That is, the prior variance is adjusted gradually to maintain identifiability of the model while resolving model fit and separating parameters that have minor deviations from zero from substantively important misspecifications. Most importantly, the PPP value should not be confused with the PPPP value or with being the result of hypothesis testing such as the hypotheses H1-H5 illustrated in the factor analysis simulation example. On the other hand the PPPP values should not be seen as being a test of model fit.

References

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

[2] Asparouhov, T., Muthén, B. & Morin, A. J. S. (2015). Bayesian structural equation modeling with cross-loadings and residual covariances: Comments on Stromeyer et al. Journal of Management, 41, 1561-1577.

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

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

[5] Scheines, R.; Hoijtink, H. and Boomsma, A. (1999) Bayesian estimation and testing of structural equation models. Psychometrika, 64,37-52.

[6] Hoijtink, H. and van de Schoot, R. (2017) Testing Small Variance Priors Using Prior-Posterior Predictive p Values. Psychological Methods. http://dx.doi.org/10.1037/met0000131