17

SEM WITH SMALL SAMPLES

Two-step modeling and factor score regression versus Bayesian estimation with informative priors

Sanne C. Smid

DEPARTMENT OF METHODOLOGY AND STATISTICS, UTRECHT UNIVERSITY, UTRECHT, THE NETHERLANDS

Yves Rosseel

DEPARTMENT OF DATA ANALYSIS, GHENT UNIVERSITY, GHENT, BELGIUM

Introduction

Bayesian estimation is regularly suggested as a beneficial method when sample sizes are small, as pointed out by systematic literature reviews in many fields, such as: organizational science (Kruschke, 2010), psychometrics (Rupp, Dey & Zumbo, 2004), health technology (Spiegelhalter, Myles, Jones & Abrams, 2000), epidemiology (Rietbergen, Debray, Klugkist, Janssen & Moons, 2017), education (König & Van de Schoot, 2017), medicine (Ashby, 2006) and psychology (Van de Schoot, Winter, Ryan, Zondervan-Zwijnenburg & Depaoli, 2017). Similarly, many simulation studies have shown the advantages of applying Bayesian estimation to address small sample size issues for structural equation models (SEMs), instead of using frequentist methods (see, for example, Depaoli, 2013; B. O. Muthén & Asparouhov, 2012; Stegmueller, 2013; Van de Schoot, Broere, Perryck, Zondervan-Zwijnenburg & Van Loey, 2015; Van Erp, Mulder & Oberski, 2018). However, as discussed in McNeish (2016) and echoed in the systematic literature review of Smid, McNeish, Miočević and Van de Schoot (2019), the use of Bayesian estimation with only diffuse default priors can cause extremely biased estimates when samples are small. The specification of informative priors is therefore required when Bayesian estimation is used with small samples.

Besides using Bayesian estimation with informative priors, there are also options for analyzing SEMs with small samples within the frequentist framework. Many studies have shown that the use of maximum likelihood (ML) estimation with small

samples can result in convergence problems, inadmissible parameter solutions and biased estimates (see, for example, Boomsma, 1985; Nevitt & Hancock, 2004). Two newly introduced and promising frequentist methods to analyze SEMs with small samples are two-step modeling (two-step) and factor score regression (FSR). A recent development is the implementation of two-step and FSR in the accessible software `lavaan` (Rosseel, 2012), as discussed in Chapter 16 (Rosseel). In two-step, the measurement models for the latent variables are estimated separately as a first step. As a second step, the remaining parameters are estimated while the parameters of the measurement models are kept fixed to their estimated values. Two-step originates from work of Burt (1976) and Anderson and Gerbing (1988), and more recent work can be found in the latent class literature (e.g., Bakk, Oberski, & Vermunt, 2014). In FSR, each latent variable in the model is replaced by factor scores and subsequently path analysis or regression analysis is run using those factor scores. Recent developments in FSR can be found in studies of Croon (2002); Devlieger, Mayer and Rosseel (2016); Devlieger and Rosseel (2017), Hoshino and Bentler (2013), and Takane and Hwang (2018).

No simulation studies were found in which two-step and FSR are compared to Bayesian estimation. Therefore, the goal of this chapter is to examine the performance of the following estimation methods under varying sample sizes: two-step, FSR, ML estimation and Bayesian estimation with three variations in the specification of prior distributions. The remainder of the chapter is organized as follows: next, the statistical model will be discussed, as well as software details, the simulation conditions, and evaluation criteria. Then, results of the simulation study will be described. We end the chapter with a summary of the results, and recommendations on when to use which estimation method in practice.

Simulation design

Statistical model

The model of interest in this simulation study is an SEM in which latent variable X is predicting latent variable Y; see Figure 17.1. Both latent variables are measured by three continuous indicators. The model and population values are similar to the model discussed in Rosseel and Devlieger (2018). The parameter of interest in the current chapter is the regression coefficient β. The standardized regression coefficient, β^Z, is 0.243, which can be considered a small effect according to Cohen (1988).

Software details

Data sets were generated and analyzed in `R` version 3.4.4. (R Core Team, 2013), using packages `lavaan` version 0.6–1 (Rosseel, 2012) for the analyses of two-step, FSR and ML; and `blavaan` version 0.3–2 (Merkle & Rosseel, 2018) for the analyses of the Bayesian conditions. Example code of the analyses using the six

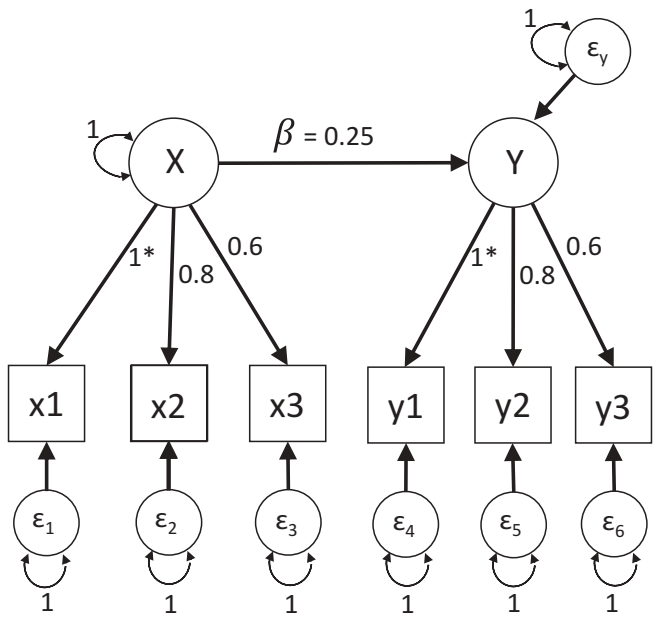

FIGURE 17.1 The model and unstandardized population values used in the simulation study. For scaling, the first factor loading for each factor is fixed to 1 (denoted by 1* in the figure), and the means of the latent variables are fixed to zero (not shown in the figure)

estimation methods can be found in supplemental file S1. All simulation code and supplemental files are available online (osf.io/bam2v/).

Six levels of sample size were examined, and for each sample size 1,000 data sets were generated according to the model and population values shown in Figure 17.1. Each generated data set was analyzed using six estimation methods. Accordingly, a total of 6 (sample size) * 6 (estimation methods) = 36 cells were investigated in the simulation design.

Simulation conditions

Six levels of sample size are studied: 10, 20, 50, 100, 250 and 500 to investigate how sample size influences the performance of the varying estimation methods. For the current model, sample sizes of 10 and 20 are extremely small. A sample size of 50 is considered small, and sample sizes of 100 and 250 are considered medium. The sample size of 500 is considered large and included as a benchmark.

Six estimation methods are considered in the current study: three frequentist estimation methods – two-step, FSR and ML – and Bayesian estimation with three types of prior specifications. For the three frequentist methods, all default settings of the `lavaan` package were used. For the default settings, see the help page for `lavOptions()` in the `lavaan` package. For the Bayesian methods, we used four chains instead of the two default chains. In terms of convergence, we used the Potential

Scale Reduction (PSR) factor, set it to a stricter criterion of 1.01, and used the following minimum number of iterations: a fixed burn-in period of 10,000 iterations (specified in `blavaan` by `adapt = 2,000, burnin = 8,000`), and for the sampling period 20,000 iterations (specified in `blavaan` by `sample = 20.000`)[1]. As an additional check, we visually assess convergence for two randomly selected data sets for each of the sample sizes and the Bayesian conditions (2 data sets * 6 sample sizes * 3 Bayesian conditions = 36 cases), by inspecting the traceplots for all parameters.

Three variants of prior specifications were examined, and all priors were specified for unstandardized parameters: BayesDefault, BayesInfoI, and BayesInfoII; see Table 17.1. The BayesDefault condition refers to a naïve use of Bayesian estimation, where only `blavaan` default priors are used. The BayesInfoI and BayesInfoII conditions refer to research situations where weakly prior information is available. In BayesInfoI, weakly informative priors are specified for the factor loadings, and `blavaan` default priors are specified for the remaining parameters. In BayesInfoII, weakly informative priors are used for both the factor loadings *and* regression coefficient β, in combination with `blavaan` default priors for the remaining parameters. Weakly informative priors were specified as follows: we set the mean hyperparameter of the normal distribution equal to the population value, and the precision hyperparameter equal to 1.

Evaluation criteria

For each of the estimation methods and sample sizes, the occurrence of convergence problems and warnings will be assessed. For the parameter of interest, regression coefficient β, the following evaluation criteria will be used to evaluate the performance under the varying estimation methods and sample sizes: relative mean bias, relative median bias, mean squared error (MSE), coverage and power. All evaluation criteria will be computed across completed replications[2].

Relative mean bias shows the difference between the average estimate across completed replications and the population value, relative to the population value. Relative median bias shows the relative difference between the median

TABLE 17.1 Specified prior distributions for the three Bayesian conditions

Parameter	*BayesDefault*	*BayesInfoI*	*BayesInfoII*
Factor loadings	$\mathcal{N}(0, 0.01)$	$\mathcal{N}(pop, 1)$	$\mathcal{N}(pop, 1)$
Regression coefficient β	$\mathcal{N}(0, 0.01)$	$\mathcal{N}(0, 0.01)$	$\mathcal{N}(pop, 1)$
Variances latent variables*	$G(1, 0.5)$	$G(1, 0.5)$	$G(1, 0.5)$
Intercepts observed variables	$\mathcal{N}(0, 0.01)$	$\mathcal{N}(0, 0.01)$	$\mathcal{N}(0, 0.01)$
Residual variances observed variables*	$G(1, 0.5)$	$G(1, 0.5)$	$G(1, 0.5)$

Note. The column BayesDefault shows the blavaan default priors (Merkle & Rosseel, 2018).

* Note that in blavaan the default priors are placed on precisions, which is the inverse of the variances. Abbreviations: $\mathcal{N}$ = Normal distribution with mean μ and precision τ; G = Gamma distribution with shape α and rate β parameters on the precision (which equals an Inverse Gamma prior with shape α and rate β parameters on the variance); *pop* = population value used in data generation.

across completed replications and the population value. The relative mean and median bias are computed by:

$$\textit{Relative mean bias} = \left[\left(\bar{\theta} - \theta\right)/\theta\right] \times 100,$$

$$\textit{Relative median bias} = \left[\left(\tilde{\theta} - \theta\right)/\theta\right] \times 100,$$

where $\bar{\theta}$ denotes the mean across completed replications, θ is the population value used for data generation, and $\tilde{\theta}$ denotes the median across completed replications. Values of relative mean and median bias below -10% or above +10% represent problematic levels of bias (Hoogland & Boomsma, 1998).

MSE is a combination of variability and bias across completed replications, where lower values indicate more stable and less biased estimates across replications. The MSE is computed by: $MSE = (\sigma)^2 + \left(\bar{\theta} - \theta\right)^2$, where σ is the standard deviation across completed replications, $\bar{\theta}$ denotes the average estimate across completed replications and θ is the population value (Casella & Berger, 2002). A narrower distribution of estimates across replications (i.e., less-variable estimates) leads to a smaller standard deviation across completed replications. Besides, the closer the estimated values are to the population value across completed replications, the smaller the amount of bias. MSE will be lower (and thus preferable) when the standard deviation and amount of bias across completed replications are small.

Coverage shows the proportion of completed replications for which the symmetric 95% confidence (for frequentist methods) or credibility (for Bayesian methods) interval contains the specified population value. Coverage values can range between 0 and 100, and values within the [92.5; 97.5] interval are considered to represent good parameter coverage (Bradley, 1978).

Finally, statistical power is expressed as the proportion of estimates for which the 95% confidence (for frequentist methods) or credibility (for Bayesian methods) interval did not contain zero, across completed replications. Power values can range from 0 to 100, where values above 80 are preferred (Casella & Berger, 2002).

Results

Convergence

With small samples, we encountered severe convergence problems when frequentist methods were used; see Table 17.2. Differences between the three frequentist methods were especially visible when $n < 100$. With $n < 100$, two-step resulted in most non-converged cases, followed by ML, and finally followed by FSR.

The three Bayesian conditions produced results in all 1,000 requested replications under all sample sizes[3]. However, when visually examining trace plots (for 2 randomly selected data sets $*$ 6 sample sizes $*$ 3 Bayesian conditions = 36 cases), severe convergence problems were detected for the smaller sample sizes, such as mode-switching; see Figure 17.2A. Mode-switching is defined as a chain that moves back

TABLE 17.2 Number of completed replications, number of warnings about negative variance estimates, and number of completed replications without negative variance estimates for two-step, FSR and ML under varying sample sizes.

	Completed replications out of 1,000 requested replications			*Number (%) of warnings of the completed replications*			*Number of completed replications without negative variance estimates*		
n	*Two-step*	*FSR*	*ML*	*Two-step*	*FSR*	*ML*	*Two-step*	*FSR*	*ML*
10	475	641	533	259 (54.5%)	432 (67.4%)	446 (83.7%)	216	209	87
20	605	797	744	167 (27.6%)	360 (45.2%)	419 (56.3%)	438	437	325
50	809	970	955	41 (5.1%)	202 (20.8%)	217 (22.7%)	768	768	738
100	950	999	997	9 (0.9%)	58 (5.8%)	52 (5.2%)	941	941	945
250	1000	1000	1000	0	0	1 (0.1%)	1000	1000	999
500	999	1000	1000	0	1 (0.1%)	0	999	999	1000

Note. *n* = sample size, Two-step = two-step modeling; FSR = factor score regression, ML = maximum likelihood estimation.

and forth between different modes (Erosheva & Curtis, 2011; Loken, 2005), such as the chains in Figure 17.2A which move back and forth between values 5 and -5.

To further examine the extent of Bayesian convergence problems, we assessed trace plots for another 25 randomly selected data sets (resulting in 25 data sets * 6 sample sizes * 3 Bayesian conditions = 450 cases). In the assessment of these 25 selected data sets, mode-switching only occurred when BayesDefault was used when $n = 10$ or 20. Mode-switching disappeared when weakly informative priors were specified; see Figures 17.2B and 17.2C. Besides mode-switching, mild spikes were also detected when $n < 100$; see Figure 17.2D. Spikes are extreme values that are sampled during Markov Chain Monte Carlo iterations, and could be seen as severe outliers. The appearance of spikes was reduced by the specification of weakly informative priors; see Figures 17.2E and 17.2F. From $n = 100$ onward, no convergence problems were detected when default priors were used. For more details on the convergence checks and more examples of trace plots, see supplemental file S2 (osf.io/bam2v/).

Warnings

For all small sample sizes, the three frequentist methods lead to a high percentage of warnings within the number of completed replications; see Table 17.2. All warnings were about negative variance parameters[4]. Differences between the three methods were especially present when $n < 100$. For these sample sizes, ML led to the highest percentage of warnings, followed by FSR, and followed by two-step. As can be seen in Table 17.2, the number of warnings decreased when sample size increased. The number of completed replications without warnings about negative variance estimates is higher for two-step and FSR compared to ML, especially when $n < 100$.

For BayesDefault, three warnings about a small effective sample size occurred for $n = 10$, and two for $n = 20$[5]. No warnings occurred in the BayesInfoI and BayesInfoII conditions.

Results for regression coefficient β

In Figure 17.3, the relative mean bias (top) and relative median bias (bottom) are presented for the varying sample sizes and estimation methods. Because of the large discrepancy between the mean relative bias and median relative bias for sample sizes below 100, we plotted the complete distribution of parameter estimates for β across replications; see Figure 17.4. For all estimation methods, an increase in sample size led to: a decrease in the number of outliers; a narrower distribution of estimates (i.e., estimates are more stable across replications); and estimates closer to the population value. With samples as small as 10 and 20, the distributions of estimates are wider and a lot of outliers are present, which are signs of unstable estimates across replications. ML produced the most extreme outliers (up to 37.57 when $n = 10$). FSR and two-step show the narrowest distribution of estimates, indicating relatively stable behavior across replications.

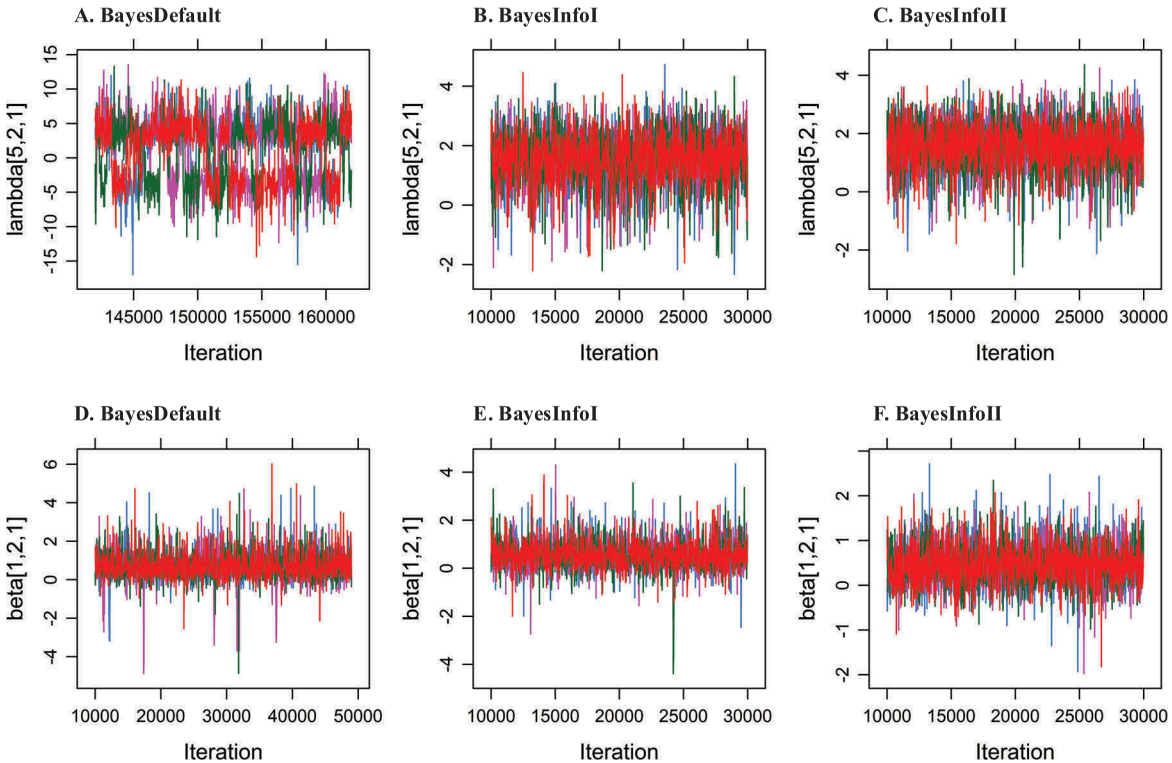

FIGURE 17.2 Trace plots for factor loading 5 (A–C), and regression coefficient β (D–F) after the analysis of BayesDefault, BayesInfoI and BayesInfoII. Trace plots A–C correspond to the analysis of replicated data set 802 (within the simulation study) with a sample size of 10. Trace plots D–F correspond to the analysis of replicated data set 260 (within the simulation study) with a sample size of 20

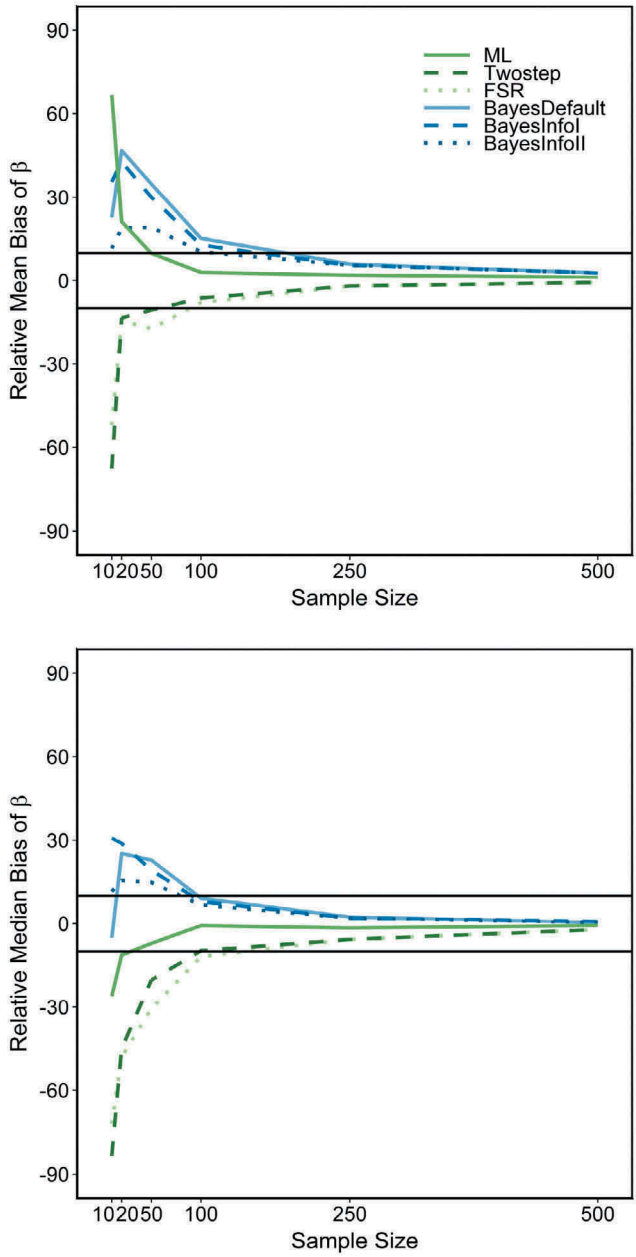

FIGURE 17.3 Relative Mean Bias (top) and Relative Median Bias (bottom) for parameter β, under varying sample sizes and estimation methods. The static black horizontal lines represent the desired $\pm 10\%$ interval

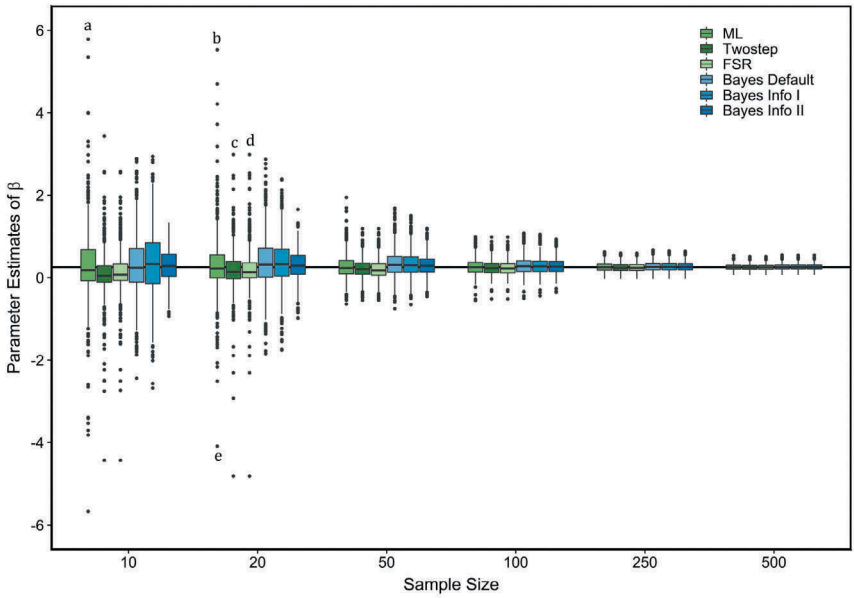

FIGURE 17.4 Distribution of the estimates for parameter β across completed replications, per estimation method and sample size. The static black horizontal line denotes the true population value of 0.25 for β. Outliers are displayed as black circles, and outliers outside the interval [-6; 6] are denoted as follows: [a] denotes 11.39, 11.46, 14.87, 37.57 for ML when $n = 10$; [b] denotes 6.49, 8.89, 9.12 for ML when $n = 20$; [c] denotes 6.86, 6.89 for two-step when $n = 20$; [d] denotes 6.86, 6.89 for FSR when $n = 20$; and [e] denotes -17.76 for ML when $n = 20$

Overall, BayesInfoII offers the best compromise between bias and stability: a narrow distribution of estimates, a mean and median close to the population value, and the smallest number of outliers. When $n = 100$, the differences between estimation methods become smaller; and the estimates become more stable across replications. For sample sizes of 250 and 500, differences between estimation methods are negligible and all estimation methods led to unbiased relative means and medians.

MSE for the regression coefficient β can be found in Figure 17.5A. Results are comparable to those shown in Figures 17.3 and 17.4. Differences between methods are especially visible when sample sizes are below 100. From $n = 100$ onward, MSE values are all close to zero. ML shows the highest MSE values for$n = 10$ and 20. BayesInfoI shows higher MSE than BayesDefault for $n = 10$, which was also visible in Figure 17.4 from the wider distribution of BayesInfoI relative to the distribution of BayesDefault for $n = 10$. The lowest MSE values are reported for BayesInfoII, followed by FSR, two-step, BayesDefault and BayesInfoI at $n = 10$. MSE values for FSR, two-step, BayesDefault and BayesInfoI are similar at $n = 20$, while BayesInfoII keeps the lowest MSE value. When

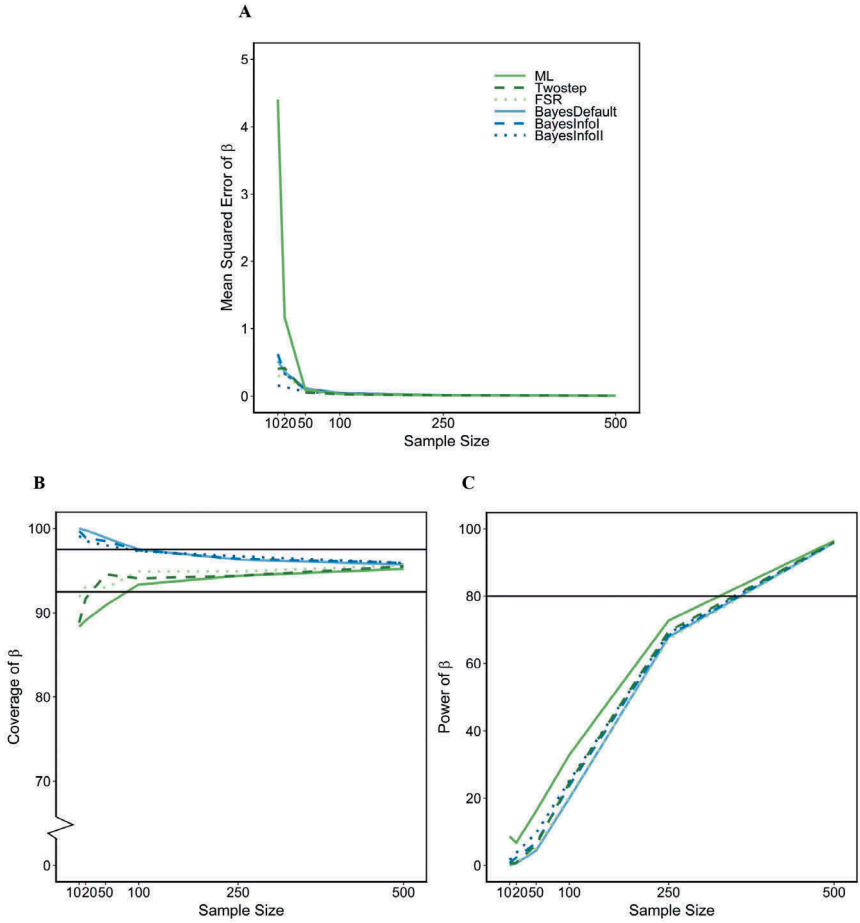

FIGURE 17.5 Mean Squared Error (A), Coverage (B), and Power (C) for parameter β, under varying sample sizes and estimation methods. The static black horizontal lines in subfigure B represent the [92.5; 97.5] coverage interval, and the black horizontal line in subfigure C represents the desired 80% power level

$n = 50$ MSE values are comparable between methods, and from $n = 100$ onward the differences in MSE between methods are negligible.

Coverage results for regression coefficient β can be found in Figure 17.5B. All estimation methods show adequate coverage levels from $n = 100$ onward. For $n < 100$, the three Bayesian conditions show excessive coverage (> 97.50), although this slightly improved under BayesInfoI and BayesInfoII. Within the three frequentist methods, two-step and FSR resulted in higher coverage levels than ML. When $n < 100$, ML shows undercoverage (< 92.50), while FSR only shows slight undercoverage when $n = 10$, and two-step when $n = 10$ and 20.

Results in terms of power can be found in Figure 17.5C. For all estimation methods, power is extremely low when the sample size is small, and only reached the desirable power level when $n = 500$. Across all sample sizes, the highest power levels are found for ML, followed by BayesInfoII, BayesInfoI, and two-step. The lowest power levels are found for FSR and BayesDefault.

Results for remaining parameters

Besides regression coefficient β, 12 remaining parameters are estimated in the model: two variances for latent variables, four factor loadings and six residual variances[6]. In supplemental file S3 (osf.io/bam2v/), the distributions of parameter estimates across replications are displayed for the remaining parameters.

Estimates for these 12 parameters seem similar across estimation methods and have good statistical properties when $n = 250$ and 500. However, with sample sizes of 100 and below, frequentist methods show many (extreme) outliers and wide distributions, indicating unstable results across replications. Bayesian methods show notably fewer outliers and in general narrower distributions than the frequentist methods, especially under BayesInfoI and BayesInfoII conditions, although the medians of the distributions still deviate from the population values when $n < 100$.

Conclusion

In this chapter, we assessed – under varying sample sizes – the performance of three frequentist methods: two-step, FSR and ML estimation; and Bayesian estimation with three variations in prior specification. With sample sizes of 250 and 500, differences between estimation methods are negligible, and all methods led to stable and unbiased estimates. Consistent with existing simulation literature (e.g., Depaoli & Clifton, 2015; Hox & Maas, 2001; Van de Schoot et al., 2015) we found that ML led to severe convergence problems and a large amount of negative variance parameters when sample sizes are small. Compared to ML, both two-step and FSR led to better convergence rates without negative variances. Also, with small samples, two-step and FSR resulted in more stable results across replications and less extreme parameter estimates than ML. When Bayesian estimation was used with default priors, problematic mode-switching behavior of the chains did occur under small samples ($n = 10,\ 20$), even though the PSR values indicated that the overall model had converged. The presence of mode-switching can be a sign that the model is too complex for the data (Erosheva & Curtis, 2011).

Power is low for all estimation methods and only with a sample size of 500 was the desired level of 80 reached. The use of *weakly informative* priors (i.e., BayesInfoI and BayesInfoII conditions), as well as the specification of `blavaan` default priors for the remaining parameters, could explain why ML led to slightly higher power levels than Bayesian estimation in the current chapter (as opposed to previous studies; for example, Miočević, MacKinnon & Levy, 2017; Van de Schoot et al., 2015).

Also, the differences in power between default and informative prior conditions were smaller in the current chapter than expected. In previous studies (e.g., Van de Schoot et al., 2015; Zondervan-Zwijnenburg, Depaoli, Peeters & Van de Schoot, 2019), priors with varying precision hyperparameters (e.g., 10 and 1) were compared to `Mplus` default priors with a precision hyperparameter of 10^{-10} (L. K. Muthén & Muthén, 1998–2017). In the current chapter, the difference in precision hyperparameters between the informative (precision = 1) and default (precision = 0.01) conditions is noticeably smaller. This could explain why the increase in power with informative priors is lower in the current chapter than expected based on previous studies. Note that the level of informativeness of a prior distribution can only be interpreted relative to the observed data characteristics, and is therefore not generalizable to other studies (i.e., a weakly informative prior in one study can act as a highly informative prior in another study that uses different measurement instruments).

In summary, with extremely small sample sizes, all frequentist estimation methods showed signs of breaking down (in terms of non-convergence, negative variances, and extreme parameter estimates), as well as the Bayesian condition with default priors (in terms of mode-switching behavior). When increasing the sample size is not an option, we recommend using Bayesian estimation with informative priors. However, note that the influence of the prior on the posterior is extremely large with relatively small samples. Even with thoughtful choices of prior distributions, results should be interpreted with caution (see also Chapter 4 by Veen and Egberts) and a sensitivity analysis should be performed; see Depaoli and Van de Schoot (2017) and Van Erp et al. (2018) on how to perform a sensitivity analysis. When no prior information is available or researchers prefer not to use Bayesian methods, two-step and FSR are a safer choice than ML, although they can still result in non-convergence, negative variances, and biased estimates.

However, note that by adjusting the implementation of two-step and FSR, non-convergence problems could be circumvented by using an alternative non-iterative estimation method (instead of ML) to estimate the measurement and structural models (see Takane & Hwang, 2018); and as discussed in Chapter 16. In addition, negative variances could be avoided by restricting the parameter space to only allow positive values for variance parameters. Therefore, the preferred approach to implement two-step and FSR in small sample contexts should be further examined. We hope the current chapter is a starting point for future research in those directions.

Acknowledgement

The first author was supported by a grant from the Netherlands Organisation for Scientific Research: NWO-VIDI-452-14-006.

Notes

1 When the PSR criterion is not reached after the specified minimum number of iterations, the number of iterations is automatically increased until the PSR criterion is met. We adjusted the blavaan default for the maximum time that the software uses to increase the amount of iterations to "24 hours" instead of the default "5 minutes".
2 We defined completed replications as replications for which (1) the model did converge according to the optimizer and (2) for which for all parameters standard errors could be computed. If the model did not converge or standard errors were not computed for one or more parameters, we defined the replication as incomplete and excluded the replication from the aggregation of the results. All simulation code can be found in supplemental file S4 (osf.io/bam2v/).
3 Note that the number of iterations in the Bayesian analyses was automatically increased until the PSR criterion of 1.01 was reached.
4 The warning message that occurred for two-step, FSR and ML was: "some estimated ov [observed variables] variances are negative". For two-step and ML, a second message also occurred: "some estimated lv [latent variables] variances are negative".
5 The warning message for BayesDefault: "Small effective sample sizes (< 100) for some parameters". The effective sample size expresses the amount of information in a chain while taking autocorrelation into account; see Chapter 4.
6 Note that when FSR is used, only three parameters are estimated: regression coefficient β, the variance of latent variable X and the variance of latent variable Y.

References

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin*, *103*(3), 411.

Ashby, D. (2006). Bayesian statistics in medicine: A 25 year review. *Statistics in Medicine*, *25* (21), 3589–3631. doi:doi.org/10.1002/sim.2672.

Bakk, Z., Oberski, D. L., & Vermunt, J. K. (2014). Relating latent class assignments to external variables: Standard errors for correct inference. *Political Analysis*, *22*(4), 520–540. doi:10.1093/pan/mpu003.

Boomsma, A. (1985). Nonconvergence, improper solutions, and starting values in LISREL maximum likelihood estimation. *Psychometrika*, *50*(2), 229–242. doi:10.1007/BF02294248.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical Statistical Psychology*, *31*(2), 144–152.

Burt, R. S. (1976). Interpretational confounding of unobserved variables in structural equation models. *Sociological Methods & Research*, *5*(1), 3–52. doi:10.1177/004912417600500101.

Casella, G., & Berger, R. L. (2002). *Statistical inference* (2nd ed.). Pacific Grove, CA: Duxbury.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Croon, M. (2002). Using predicted latent scores in general latent structure models. In G. A. Marcoulides & I. Moustaki (Eds.), *Latent variable and latent structure models* (pp. 195–223). Mahwah, NJ: Lawrence Erlbaum.

Depaoli, S. (2013). Mixture class recovery in GMM under varying degrees of class separation: Frequentist versus Bayesian estimation. *Psychological Methods*, *18*(2), 186–219. doi:10.1037/a0031609.

Depaoli, S., & Clifton, J. P. (2015). A Bayesian approach to multilevel structural equation modeling with continuous and dichotomous outcomes. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*(3), 327–351. doi:10.1080/10705511.2014.937849.

Depaoli, S., & Van de Schoot, R. (2017). Improving transparency and replication in Bayesian statistics: The WAMBS-Checklist. *Psychological Methods*, *22*(2), 240–261. doi:10.1037/met0000065.

Devlieger, I., Mayer, A., & Rosseel, Y. (2016). Hypothesis testing using factor score regression: A comparison of four methods. *Educational and Psychological Measurement*, *76*(5), 741–770. doi:10.1177/0013164415607618.

Devlieger, I., & Rosseel, Y. (2017). Factor score path analysis: An alternative for SEM. *Methodology*, *13*, 31–38. doi:10.1027/1614-2241/a000130.

Erosheva, E. A., & Curtis, S. M. (2011). *Dealing with rotational invariance in Bayesian confirmatory factor analysis*. Technical Report. Seattle, WA: Department of Statistics, University of Washington, 35.

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods Research in Higher Education*, *26*(3), 329–367.

Hoshino, T., & Bentler, P. M. (2013). Bias in factor score regression and a simple solution. In A. De Leon & K. Chough (Eds.), *Analysis of mixed data* (pp. 43–61). Boca Raton, FL: Chapman & Hall/CRC Press.

Hox, J. J., & Maas, C. J. M. (2001). The accuracy of multilevel structural equation modeling with pseudobalanced groups and small samples. *Structural Equation Modeling: A Multidisciplinary Journal*, *8*(2), 157–174. doi:10.1207/S15328007SEM0802_1.

König, C., & Van de Schoot, R. (2017). Bayesian statistics in educational research: A look at the current state of affairs. *Educational Review*, 1–24. doi:10.1080/00131911.2017.1350636.

Kruschke, J. K. (2010). Bayesian data analysis. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(5), 658–676. doi:10.1002/wcs.72.

Loken, E. (2005). Identification constraints and inference in factor models. *Structural Equation Modeling: A Multidisciplinary Journal*, *12*(2), 232–244. doi:10.1207/s15328007sem1202_3.

McNeish, D. (2016). On using Bayesian methods to address small sample problems. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(5), 750–773. doi:10.1080/10705511.2016.1186549.

Merkle, E. C., & Rosseel, Y. (2018). Blavaan: Bayesian structural equation models via parameter expansion. *Journal of Statistical Software*, *85*(4), 1–30. doi:10.18637/jss.v085.i04.

Miočević, M., MacKinnon, D. P., & Levy, R. (2017). Power in Bayesian mediation analysis for small sample research. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(5), 666–683. doi:10.1080/10705511.2017.1312407.

Muthén, B. O., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods*, *17*(3), 313–335. doi:10.1037/a0026802.

Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus user's guide*. Los Angeles, CA: Muthén & Muthén.

Nevitt, J., & Hancock, G. R. (2004). Evaluating small sample approaches for model test statistics in structural equation modeling. *Multivariate Behavioral Research*, *39*(3), 439–478. doi:10.1207/S15327906MBR3903_3.

R Core Team. (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rietbergen, C., Debray, T. P. A., Klugkist, I., Janssen, K. J. M., & Moons, K. G. M. (2017). Reporting of Bayesian analysis in epidemiologic research should become more transparent. *Journal of Clinical Epidemiology*, *86*, 51–58. e52 10.1016/j.jclinepi.2017.04.008.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36.

Rosseel, Y., & Devlieger, I. (2018). *Why we may not need SEM after all*. Amsterdam: Meeting of the SEM Working Group.

Rupp, A. A., Dey, D. K., & Zumbo, B. D. (2004). To Bayes or not to Bayes, from whether to when: Applications of Bayesian methodology to modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *11*(3), 424–451. doi:10.1207/s15328007sem1103_7.

Smid, S. C., McNeish, D., Miočević, M., & Van de Schoot, R. (2019). Bayesian versus frequentist estimation for structural equation models in small sample contexts: A systematic review. *Structural Equation Modeling: A Multidisciplinary Journal*. doi:10.1080/10705511.2019.1577140.

Spiegelhalter, D. J., Myles, J. P., Jones, D. R., & Abrams, K. R. (2000). Bayesian methods in health technology assessment: A review. *Health Technology Assessment*, *4*(38), 1–130.

Stegmueller, D. (2013). How many countries for multilevel modeling? A comparison of frequentist and Bayesian approaches. *American Journal of Political Science*, *57*(3), 748–761. doi:10.1111/ajps.12001.

Takane, Y., & Hwang, H. (2018). Comparisons among several consistent estimators of structural equation models. *Behaviormetrika*, *45*(1), 157–188. doi:10.1007/s41237-017-0045-5.

Van de Schoot, R., Broere, J. J., Perryck, K. H., Zondervan-Zwijnenburg, M., & Van Loey, N. E. (2015). Analyzing small data sets using Bayesian estimation: The case of posttraumatic stress symptoms following mechanical ventilation in burn survivors. *European Journal of Psychotraumatology*, *6*(1), 25216. doi:10.3402/ejpt.v6.25216.

Van de Schoot, R., Winter, S. D., Ryan, O., Zondervan-Zwijnenburg, M., & Depaoli, S. (2017). A systematic review of Bayesian articles in psychology: The last 25 years. *Psychological Methods*, *22*(2), 217–239. doi:10.1037/met0000100.

Van Erp, S., Mulder, J., & Oberski, D. L. (2018). Prior sensitivity analysis in default Bayesian structural equation modeling. *Psychological Methods*, *23*(2), 363–388. doi:10.1037/met0000162.

Zondervan-Zwijnenburg, M. A. J., Depaoli, S., Peeters, M., & Van de Schoot, R. (2019). Pushing the limits: The performance of ML and Bayesian estimation with small and unbalanced samples in a latent growth model. *Methodology*. doi:https://doi.org/10.1027/1614-2241/a000162.