CHAPTER 38

Foundations and Extensions of Bayesian Structural Equation Modeling

Sarah Depaoli
David Kaplan
Sonja D. Winter

In the years since our chapter on Bayesian SEM (BSEM) appeared in the first edition of this volume (Kaplan & Depaoli, 2012), Bayesian inference has become more popular as an alternative to frequentist statistical methodology in the social and behavioral sciences (van de Schoot, Winter, Zondervan-Zwijnenburg, Ryan, & Depaoli, 2017). In addition, considerable advances have been made within Bayesian computation, which is now becoming more widely available in open-source computing environments. As such, Bayesian methods, and BSEM in particular, are increasingly being used to address substantive problems.

The purpose of this chapter is to provide an update to Kaplan and Depaoli (2012). We limit our discussion of introductory material and focus our attention on new developments in BSEM. In our view, there are four major developments in Bayesian inference directly relevant to BSEM. The first development is computational. In this chapter, we focus on the advent of Hamiltonian Monte Carlo (HMC) estimation as implemented in the open-source software program `Stan` (Stan Development Team, 2021) and available in the open-source SEM program `blavaan` (Merkle & Rosseel, 2018). The second development has been in model selection. For this chapter, we review standard methods of model selection but discuss newer methods based directly on evaluating predictive accuracy using *leave-one-out cross-validation* (LOOCV). The third development concerns treating a structural model as one of a large number of models that could have generated the data. Greater predictive accuracy can be obtained by not selecting a single model but by averaging over the set of models—so-called “Bayesian model averaging.” Finally, we also focus on a flexible treatment of restricting parameters (e.g., restricting cross-loadings to being near zero) through the use of priors.

The organization of this chapter is as follows. To begin, other chapters in this volume provide a full account of basic and advanced concepts in SEM, and we assume that the reader is familiar with these topics. Given that assumption, the next section provides a brief introduction to Bayesian ideas, including Bayes theorem, the nature of prior distributions, describing the posterior distribution, and Bayesian model building. More detail can be found in standard Bayesian statistical texts; introductory texts include Gill (2002) and Kruschke (2015), whereas Gelman and colleagues (2014) and Kaplan (2023) offer a more advanced treatment of Bayesian statistics. Following that, we provide a brief overview of Markov chain Monte Carlo (MCMC) sampling, and particularly HMC, that we use for the empirical examples in this chapter. Next, we introduce the general form of the BSEM. This is followed by a discussion of Bayesian model evaluation and selection, including the use of model averaging techniques. Next, we introduce the use of approximate-zero priors for releasing restrictions in SEMs, then highlight the importance of conducting a prior sensitivity analysis to examine the impact of prior settings on statistical and substantive results. We then present Bayesian regularization methods implemented in SEM. We conclude this chapter with final thoughts about the future of BSEM in the applied and methodological literature, as well as a general discussion of how the Bayesian approach to SEM can lead to a pragmatic and evolutionary development of knowledge in the social and behavioral sciences.

Companion Website

As supplementary material to the contents of this chapter, we include several examples of BSEM implemented in `Stan` and `blavaan` at the companion website.

Text boxes are placed throughout the chapter to highlight relevant demonstrations and features that are presented at the companion website. An extensive write-up, R code (with annotation), data, and results for all examples are presented at the companion website.

PRELIMINARIES ON BAYES THEOREM AND PRIOR DISTRIBUTIONS

In this section, we set the notation and concepts of Bayesian statistics that will be necessary for later developments. Much more thorough treatments of Bayesian statistics can be found in Gelman and colleagues (2014) and Kaplan (2014).

Following closely the discussion in Kaplan (2014), the goal of statistical inference is to obtain estimates of the unknown parameters (denoted as θ) given the data (denoted as *y*). For our purposes, the model parameters are those that come from from SEMs, including structural regression parameters and measurement model parameters. The key difference between Bayesian statistical inference and frequentist statistical inference concerns the nature of θ. In the frequentist tradition, the assumption is that θ is unknown but has a fixed value that we wish to estimate. In Bayesian statistical inference, θ is also considered unknown; however, similar to the data, θ is viewed as a random variable possessing a *prior probability distribution* that encodes our uncertainty about the true value of θ before having seen the data. Because both the observed data *y* and the parameters θ are assumed to be random variables, the probability calculus allows us to express the joint probability of the parameters and the data as a function of the conditional distribution of the data given the parameters, and the prior distribution, namely

$$p(\theta, y) = p(y|\theta)\,p(\theta) \tag{38.1}$$

where $p(\theta, y)$ is the joint distribution of the parameters and the data, $p(y \mid \theta)$ is the distribution of the data conditional on the parameters—that is, the model—and $p(\theta)$ is the prior distribution. Bayes theorem is then defined as

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(y|\theta)\,p(\theta)}{p(y)} \tag{38.2}$$

where $p(\theta \mid y)$ is referred to as the *posterior distribution* of the parameters θ given the observed data *y* representing our updated knowledge about the parameters of interest after having encountered the data and is equal to the data distribution $p(y \mid \theta)$ times the prior distribution of the parameters $p(\theta)$ normalized by $p(y)$ so that the posterior distribution sums (or integrates) to one.

Prior Distributions

The general approach to considering the choice of a prior distribution on the parameters θ is based on how much information we believe we have *prior* to data collection and how accurate we believe that information to be. The strength of Bayesian inference lies precisely in its ability to incorporate our uncertainty about θ directly into our statistical models; prior distributions are of three general types: (1) noninformative, (2) weakly informative, and (3) informative priors (Gelman et al., 2014).

Noninformative Priors

In some cases, we may not be in possession of enough previous information to aid in drawing posterior inferences. Or, from a policy perspective, it may be prudent to refrain from providing subjective probabilities of effects of interest, letting the data speak for themselves. Regardless, from a Bayesian perspective, this lack of information is still important to consider and incorporate into our statistical models. In other words, it is equally as important to quantify our ignorance as it is to quantify our cumulative understanding of a problem at hand.

The standard approach to quantifying our ignorance about θ is to use noninformative prior distributions. In a case with no prior knowledge to draw from, perhaps the most common noninformative prior distribution to use is the uniform distribution $U(\alpha, \beta)$ over some sensible range of values from α to β. The uniform distribution essentially signals that we believe that the value of our parameter of interest lies in the range $\beta - \alpha$ and that all values in between have equal probability.

Weakly Informative Priors

Situated between noninformative and informative priors are *weakly informative* priors. Weakly informative priors are distributions that provide a method for incorporating less information than one actually has in a particular situation. Specifying weakly informative priors can be useful for many reasons. First, it is doubtful that one has complete ignorance of a problem for which a noninformative prior such as the uniform distribution is appropriate. Rather, it is likely that one can consider a range of parameter values that are reasonable. Second, as discussed in Gelman and colleagues (2014), weakly informative priors can be useful in theory testing, where it may appear unfair to specify strong priors in the direction of one's theory. Rather, specifying weakly informative priors in the opposite direction of a theory would then require the theory to pass a higher standard of evidence. Third, weakly informative priors are particularly useful in small-sample-size situations in which the prior distribution can overwhelm the likelihood (Gelman et al., 2014).

Informative Priors

Finally, it may be the case that previous research, expert opinion, or both, can be brought to bear on a problem and systematically incorporated into the prior distribution. Such priors are referred to as *informative* and require that the analyst commit to the shape of the distribution. For example, if a parameter of interest, such as a regression coefficient, is assumed to have a normal prior distribution, then the analyst must commit to specifying the average value and the precision around that value. Given that informative priors are inherently subjective in nature, they can be inconsistent with the population parameter. Fortunately, Bayesian theory provides numerous methods for assessing the sensitivity of results to the choice of prior distributions.

Bayesian versus Frequentist Comparisons

It is beyond the scope of this chapter to outline all of the differences between Bayesian and frequentist methods. A more complete treatment of the issue can be found in Wagenmakers (2007); we note several relevant distinctions:

1. Bayesian inference is the only paradigm of statistics that allows for the quantification of uncertainty. Uncertainty is present not only in our knowledge of parameters of interest, but also in the very models used to estimate those parameters. Central to Bayesian theory and practice is that the intervals around parameter estimates (so-called *credible intervals*) are more honest; models demonstrate better long-run predictive accuracy if uncertainty is directly addressed rather than ignored (Kaplan, 2014). In this chapter, we address both parameter and model uncertainty in SEM.

2. Bayesian methods provide a paradigm for dealing with the well-documented problems with the p-value. In particular, one criticism of the p-value is that it violates the *likelihood principle*, which states that in making inference or decisions about a parameter after data are observed, all relevant observational information is contained in the likelihood function for the observed data (Jeffreys, 1961; Wagenmakers, 2007). The p-value violates this principle insofar as reference to the distribution of the chosen test statistic refers to values that equal or exceed the observed value (i.e., the p-value includes values that were not observed in the data at hand). In contrast, Bayesian inference draws inferences only on data that were observed and summarized in the likelihood.

3. In large samples, Bayesian approaches and frequentist approaches will converge, though with differing interpretations. As noted earlier, frequentist parameters are treated as fixed; only uncertainty due to sampling variability can be estimated through reference to the estimate's standard error. Bayesian estimates, by contrast, are interpreted probabilistically. This provides a richer interpretation than the simple decision of whether a parameter estimate is statistically significant.

BASIC IDEAS OF MCMC SAMPLING

Two of the most common approaches to parameter estimation in SEM are *maximum likelihood* and *weighted least squares*. The focus of frequentist parameter estimation is the derivation of point estimates and standard errors of model parameters that, under standard assumptions, have desirable asymptotic properties such as consistency, asymptotic normality, and efficiency (see, e.g., Silvey, 1975).

In contrast to maximum likelihood estimation and other estimation methods within the frequentist paradigm, Bayesian inference focuses on calculating expectations of the posterior distribution of the model parameters. For very simple problems, this can be handled analytically. However, for complex, high-dimensional problems involving multiple integrals, the task of calculating expectations can be virtually impossible. So rather than attempting to analytically solve these high-dimensional problems, we can instead use well-established mathematical computation methods to draw samples from a *target distribution* of interest (in our case, the posterior distribution) and summarize the distribution formed by those samples. This is referred to as *Monte Carlo integration*.

Monte Carlo integration is based on first drawing T samples of the parameters of interest $\{\theta_t, t = 1, \ldots, T\}$ from the posterior distribution $p(\theta | y)$ and approximating the expectation by

$$E\left[p\left(\theta|y\right)\right] \approx \frac{1}{T}\sum_{t=1}^{T} p\left(\theta_t|y\right) \qquad (38.3)$$

Assuming the samples are independent of one another, the law of large numbers ensures that the approximation in Equation 38.3 will be increasingly accurate as T increases. However, an important feature of Monte Carlo integration of particular relevance to BSEM is that the samples do not have to be drawn independently. All that is required is that the sequence $\{\theta_t, t = 1, \ldots, T\}$ yields samples that have explored the support of the distribution (Gilks, Richardson, & Spiegelhalter, 1996).[1]

One approach to sampling throughout the support of a distribution while also relaxing the assumption of independent sampling is through the use of a Markov chain. Formally, a *Markov chain* is a sequence of dependent samples of random variables $\{\theta^t\}$

$$\theta^0, \theta^1, \ldots, \theta^t, \ldots \qquad (38.4)$$

such that the conditional probability of θ^t given all of the past variables depends only on θ^{t-1}—that is, only on the immediate past variable.

The Markov chain has a number of important properties, not the least of which is that over a long sequence, the chain will *forget* its initial state θ^0 and converge to its stationary distribution $p(\theta | y)$, which does not depend either on the number of samples T or on the initial state θ^0. The number of iterations prior to the stability of the distribution is referred to as *burn-in* or *warm-up* samples. Letting m represent the initial number of burn-in samples, we can obtain an *ergodic average* of the posterior distribution $p(\theta | y)$ as

$$\bar{p}\left(\theta|y\right) = \frac{1}{T-m}\sum_{t=m+1}^{T} p\left(\theta^t|y\right) \qquad (38.5)$$

The idea of conducting Monte Carlo sampling through the construction of Markov chains defines MCMC. The key issue now is how to specifically draw values from the posterior distribution. Three popular algorithms have been developed for this purpose: the *random walk Metropolis–Hastings* (MH) algorithm, the *Gibbs sampler*, and the HMC. The MH algorithm and the Gibbs sampler were presented in the first edition of the *Handbook*. We review them here briefly but spend more time discussing HMC.

Random Walk MH Algorithm

One of the earliest, yet still common, methods for constructing a Markov chain is referred to as the *random walk MH* algorithm (Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953). The basic steps of the MH algorithm are as follows. First, a starting value θ^0 is obtained and a sequence $\theta^0, . . \theta^{t-1}$ is desired. The next element in the chain, θ^t, is obtained by first obtaining a proposal value θ^* from a so-called *proposal* distribution (also referred to as a *jumping distribution*, which we will denote as $q(\theta^* | \theta^{t-1})$.[2] This proposal distribution could be, for example, a standard normal distribution with mean zero and some variance. Second, assuming symmetric proposal distributions (i.e., $q(\theta^* | \theta^{t-1}) = q(\theta^{t-1} | \theta^*)$) the algorithm *accepts* the candidate value with an *acceptance probability*

$$p\left(\theta^*|\theta^{t-1}\right) = \min\left\{1, \frac{p\left(\theta^*\right)}{p\left(\theta^{t-1}\right)}\right\} \qquad (38.6)$$

Notice that the numerator of Equation 38.6 is the probability of the candidate value, and the denominator is the probability of the current value. Thus, if the odds ratio $p(\theta^*)/p(\theta^{t-1}) > 1.0$, then the probability of acceptance of the candidate value is 1.0—that is, the algorithm accepts the candidate value with certainty. However, if the odds ratio is less than 1.0, then the algorithm can move to the next value or stay at the current value, which is determined by a random draw from a U(0,1) distribution.

The Gibbs Sampler

Another popular algorithm for MCMC is the *Gibbs sampler*, which is a special case of the MH algorithm. Consider that the goal is to obtain the joint posterior distribution of two model parameters, say, θ_1 and θ_2, given some data y, written as $p(\theta_1, \theta_2 | y)$. Dropping the conditioning on y for notational simplicity, what is required is to sample from $p(\theta_1 | \theta_2)$ and $p(\theta_2 | \theta_1)$. In the first step, an arbitrary value for θ_2 is chosen, say θ_2^0. We next obtain a sample from $p(\theta_1 | \theta_2^0)$. Denote this value as θ_1^1. With this new value, we then obtain a sample θ_2^1 from $p(\theta_2 | \theta_1^1)$. The Gibbs algorithm continues to draw samples using previously obtained values until two long chains of values for both θ_1 and θ_2 are formed. After discarding the burn-in samples, the remaining samples are then considered to be samples drawn from the marginal distributions of $p(\theta_1)$ and $p(\theta_2)$.

Hamiltonian Monte Carlo

The development of the MH and Gibbs sampling algorithms and their implementation in readily accessible software programs has made it possible to bring Bayesian statistics into mainstream practice. However, these two algorithms suffer from a severe practical limitation—namely, as the dimensionality of the parameter space increases, the number of directions that the algorithm can search increases exponentially while, at the same time, the MH acceptance probability in Equation 38.6 decreases. Thus, the MH and Gibbs algorithms can take an unacceptably long time to converge to the posterior distribution, resulting in a highly inefficient use of computer resources (Hoffman & Gelman, 2014). Given that the dimensionality of the parameter space of SEMs is typically large, the inefficiency of MH and Gibbs is problematic. An approach to addressing this problem has come through the development of *HMC*. HMC underlies the `Stan` programming environment. Excellent intuitive introductions to HMC have been provided by Betancourt (2018, 2019).

The problem associated with the inefficient use of computer resources when implementing MH or Gibbs algorithms stems from the geometry of probability distributions when the number of parameters increase. In particular, while the density of a distribution is largest in the neighborhood near the mode, the volume of that neighborhood decreases as the number of parameters increase and thus has inconsequential impact on the calculation of expectations. At the same time, as the number of parameters increase, the region far away from the mode has greater volume but much smaller density, and thus also contributes negligibly to the calculation of expectations. The neighborhood between these extremes is called the *typical set*, which is a subspace of the support of the distribution. This "Goldilocks zone" represents a region where the volume and density are just right, and where the mass is sufficient to produce reasonable expectations. Outside of the typical set, the contribution to the calculation of expectations is inconsequential, wasting computing resources (Betancourt, 2018).

The inefficiency of the MH and Gibbs algorithms is due to the random walk nature of these algorithms. In the best case scenario, for a small number of parameters, the MH algorithm will be biased toward the tails of the distribution where the volume is high, rejecting proposal values if the density is small. This will result in a relatively efficient exploration of the typical set. However, as the number of parameters increases, the volume outside the typical set will dominate the volume inside the typical set. Thus, the Markov chain will mostly end up outside the typical set, yielding proposals with low probabilities and hence more rejections by the algorithm. This results in the Markov chain getting stuck outside the typical set and moving slowly, as is often observed when employing MH in practice.

The solution to the problem of the Markov chain getting stuck outside the typical set is to come up with an approach that is capable of making large jumps across regions of the typical set, such that the typical set is fully explored. This is the goal of HMC. Specifically, HMC exploits the geometry of the typical set and constructs transitions that "glide across the typical set towards new, unexplored neighborhoods" (Betancourt, 2018, p. 18). The key to gliding across the typical set is to carefully choose an auxiliary momentum parameter to the probabilistic system. This momentum parameter is essentially a first-order gradient calculated from the log-posterior distribution.

No-U-Turn Sampler

HMC yields a much more efficient exploration of the posterior distribution compared to random walk MH and Gibbs. However, HMC does require user-specified parameters that can still result in a degree of computational inefficiency.

These parameters are referred to as the step size ϵ and the number of so-called *leapfrog* steps L. If ϵ is too large, then the acceptance rates will be too low. On the other hand, if ϵ is too small, then computation time is wasted because the algorithm is taking small steps. Regarding the leapfrog steps, if L is too small, then the draws will be too close to each other, resulting in random walk behavior and slow mixing of the chains. When mixing is slower, there is typically higher dependency on the starting values for the chains, requiring a longer chain to determine convergence (see next section). If L is too large, then computational resources will be wasted because the algorithm will loop back and repeat its steps (Hoffman & Gelman, 2014). This is a point worth considering because mixing time can be long for complex SEMs, and wasted computational power can unnecessarily increase the run time (which may already be longer as compared to less complex models). Although ϵ can be adjusted "on the fly" through the use of adaptive MCMC, deciding on the appropriate value of L is more difficult, and a poor choice of either parameter can lead to serious computational inefficiency. To solve these problems, the *No-U-Turn Sampler* (NUTS) algorithm developed by Hoffman and Gelman (2014) is designed to mimic the dynamics of HMC, while not requiring the user to specify ϵ or L. The NUTS algorithm is implemented in `Stan` (Stan Development Team, 2021).

CONVERGENCE DIAGNOSTICS

Given the computational intensity of MCMC, it is essential for Bayesian inference that we assess the convergence of the MCMC algorithm. The importance of assessing convergence stems from the very nature of MCMC, in that the MCMC algorithm is designed to converge in distribution rather than to a point estimate. Because there is not a single adequate assessment of convergence, it is important to inspect several different diagnostics that examine varying aspects of convergence. Three primary methods of assessing convergence are available in `Stan` and `blavaan`—all of which are demonstrated in the accompanying online supplementary materials.

Trace Plots

Perhaps the most common diagnostic for assessing MCMC convergence is to examine the so-called called *trace* or *history* plots, which plot the posterior parameter estimate obtained for each iteration of the chain. Typically, a parameter will appear to converge if the sample estimates form a tight horizontal band across the history of the iterations forming the chain. However, using this method as an assessment for convergence is rather crude since merely viewing a tight trace plot does not indicate that convergence was actually obtained. As a result, this method is more likely to be an indicator of nonconvergence (Mengersen, Robery, & Guihenneuc-Jouyax, 1999). For example, if two chains for the same parameter are sampled from different areas of the target distribution and the estimates over the history of the chain stay separated, then that would be evidence of nonconvergence. Likewise, if a plot shows substantial fluctuation or jumps in the chain, it is likely that the chain linked to that parameter has not reached convergence.

Posterior Density Plots

Another useful tool for diagnosing issues with the convergence of the Markov chain is the *posterior density plot*. This is a plot of the draws for each parameter in the model and for each chain. This plot is important to inspect insofar as the summary statistics for the parameters of the model are calculated from these posterior draws. If we consider, for example, a path coefficient in an SEM, with an associated conjugate normal prior, then we would expect the posterior density of that plot to be normally distributed. Any strong deviations from normality and, in particular, any serious bimodality in the plot would suggest issues with convergence of that parameter, which could possibly be resolved through more iterations, better choice of priors, or both.

Autocorrelation Plots and Effective Sample Size

One should also examine the speed at which the draws from the posterior distribution achieve independence. As noted earlier, draws from the posterior distribution using a Markov chain are not, initially, independent of one another. However, the chain should eventually "forget" its initial state and converge to a set of independent and stationary draws from the posterior distribution. We can determine how quickly the chain has forgotten its initial state by inspecting the autocorrelation function (ACF) plot, defined as follows: Let θ^t ($t = 1, \ldots, T$) be the t^{th} component (parameter draw) of a stationary Markov chain. Then the *lag-k* autocorrelation can be written as

$$\rho^k = \text{corr}\left(\theta^t, \theta^{t+k}\right) \tag{38.7}$$

In general, the lag-1 autocorrelation should be close to 1.0. However, we also expect that the components of the Markov chain will become independent as k increases. Thus, we prefer that the autocorrelation decrease quickly over the number of iterations. If this is not the case, it is evidence that the chain is "stuck" and thus not providing a full exploration over the support of the target distribution. In general, positive autocorrelation will be observed, but in some cases negative autocorrelation is possible, indicating fast convergence of the estimated value to the equilibrium value.

Related to the autocorrelation diagnostic is the *effective sample size* (ESS), which is an estimate of the number of independent draws from the posterior distribution. The ESS is calculated as

$$\text{ESS} = \frac{N}{1 + 2\sum_{t=1}^{\infty} \rho^t} \tag{38.8}$$

Because samples from the posterior distribution are not independent, we expect from Equation 38.8 that the ESS will be smaller than the total number of draws. If the ratio of the ESS to the total number of draws is close to 1.0, this is evidence that the algorithm as achieved mostly independent draws. Values less than 0.1 are a cause for concern; however, we note that this ratio is highly dependent on the choice of MCMC algorithm, number of burn-in iterations, and number of post–burn-in iterations.

One approach to addressing the problem of autocorrelation and the associated problem of lower ESS ratio is to use *thinning*. Suppose we request that the algorithm draws every 10th iteration from the posterior distribution until 3,000 draws are completed. The algorithm has 30,000 iterations, but we thin the sample by keeping only every 10th draw, resulting in 3,000 saved draws. Although thinning reduces memory burden, the autocorrelation is typically also reduced, resulting in a higher ESS.

Potential Scale Reduction Factor

When implementing an MCMC algorithm, one of the most common diagnostics is the *potential scale reduction factor* (see, e.g., Gelman, 1996; Gelman & Rubin, 1992a, 1992b), often denoted as $\hat{R}$. This diagnostic is based on analysis of variance and is intended to assess convergence among several parallel chains with varying starting values. Specifically, Gelman and Rubin (1992a) proposed a method where an overestimate and an underestimate of the variance of the target distribution are formed. The overestimate of the variance of the target distribution is measured by the between-chain variance; the underestimate is measured by the within-chain variance (Gelman, 1996). The theory is that if the ratio of these two sources of variance is equal to one, then this is evidence that the chains have converged. If the $\hat{R} > 1.01$, then this is typically a cause for concern. Brooks and Gelman (1998) added an adjustment for sampling variability in the variance estimates and also proposed a multivariate extension of the potential scale reduction factor, which does not include the sampling variability correction.

The $\hat{R}$ diagnostic is calculated for all chains over all iterations. A problem with $\hat{R}$ originally noted by Gelman and colleagues (2014) and further discussed in Vehtari, Gelman, Simpson, Carpenter, and Bürkner (2021) is that it sometimes does not detect nonstationarity. A relatively new version of the potential scale reduction factor is available in `Stan`. This version, referred to as the *Split* $\hat{R}$, is designed to address the problem that the conventional $\hat{R}$ fails to detect. The *Split* $\hat{R}$ quantifies the variation of a set of Markov chains initialized in SEM from location points in parameter space. This is accomplished by splitting the chain in two, then calculating the *Split* $\hat{R}$ on twice as many chains. So, if one is using four chains with 500 iterations per chain, the *Split* $\hat{R}$ is based on eight chains with 250 iterations per chains.

SPECIFICATION OF BAYESIAN STRUCTURAL EQUATION MODELING

Following general notation, denote the measurement model as

$$\mathbf{y} = \boldsymbol{\alpha} + \boldsymbol{\Lambda\eta} + \mathbf{Kx} + \boldsymbol{\epsilon} \tag{38.9}$$

where $\mathbf{y}$ is a vector of manifest variables, $\boldsymbol{\alpha}$ is a vector of measurement intercepts, $\boldsymbol{\Lambda}$ is a factor loading matrix, $\boldsymbol{\eta}$ is a vector of latent variables, $\mathbf{K}$ is a matrix of regression coefficients relating the manifest variables $\mathbf{y}$ to observed variables $\mathbf{x}$, and $\boldsymbol{\epsilon}$ is a vector of uniquenesses with covariance matrix $\boldsymbol{\Xi}$, assumed to be diagonal. The structural model relating common factors to each other and possibly to a vector of manifest variables $\mathbf{x}$ is written as

$$\boldsymbol{\eta} = \boldsymbol{\nu} + \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\Gamma}\mathbf{x} + \boldsymbol{\zeta} \tag{38.10}$$

where $\boldsymbol{\nu}$ is a vector of structural intercepts, $\mathbf{B}$ and $\boldsymbol{\Gamma}$ are matrices of structural coefficients, and $\boldsymbol{\zeta}$ is a vector of structural disturbances with covariance matrix $\boldsymbol{\Psi}$, which is assumed to be diagonal.

Conjugate Priors for SEM Parameters

To specify the prior distributions, it is notationally convenient to arrange the model parameters as sets of common conjugate distributions. Parameters with the subscript *n* follow a normal distribution, while those with the subscript *IW* follow a inverse-Wishart distribution. Let $\boldsymbol{\theta}_{norm} = \{\boldsymbol{\alpha}, \boldsymbol{\nu}, \boldsymbol{\Lambda}, \mathbf{B}, \boldsymbol{\Gamma}, \mathbf{K}\}$ be the vector of free model parameters that are assumed to follow a normal distribution, and let $\boldsymbol{\theta}_{IW} = \{\boldsymbol{\Xi}, \boldsymbol{\Psi}\}$ be the vector of free model parameters that are assumed to follow the inverse-Wishart distribution. Formally, we write

$$\boldsymbol{\theta}_{norm} \sim N(\boldsymbol{\mu}, \boldsymbol{\Omega}) \tag{38.11}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Omega}$ are the mean and variance hyperparameters (i.e., the parameters that govern the shape and scale of the prior distribution), respectively, of the normal prior. For blocks of variances and covariances in $\boldsymbol{\Xi}$ and $\boldsymbol{\Psi}$, we assume that the prior distribution is inverse-Wishart:[3]

$$\boldsymbol{\theta}_{IW} \sim IW(\mathbf{R}, \delta) \tag{38.12}$$

where $\mathbf{R}$ is a positive definite matrix, and $\delta > q - 1$, where q is the number of distinct variables in the covariance matrix. Different choices for $\mathbf{R}$ and δ will yield different degrees of "informativeness" for the inverse-Wishart distribution.

In addition to the conventional SEM parameters and their priors, an additional model parameter is required for the mixture SEMs. Specifically, it is required that we estimate the mixture proportions, which we will denote as $\boldsymbol{\pi}$. In this specification, the class labels assigning an individual to a particular latent class follows a multinomial distribution with parameters *n*, the sample size, and $\boldsymbol{\pi}$ is a vector of latent class proportions. The conjugate prior for trajectory class proportions is the Dirichlet(τ) distribution with hyperparameters $\tau = (\tau_1, \ldots, \tau_T)$, where T is the number of trajectory classes and $\sum_{\tau=1}^{T} = 1$.

Examples of BSEM

The companion website presents examples of several model forms implemented using Bayesian estimation, which include the following: confirmatory factor analysis (CFA), SEM, and latent growth curve modeling. Prior specification is described for each example provided.

BAYESIAN MODEL EVALUATION

SEM, by its very nature, involves the specification, estimation, and testing of models that purport to represent the underlying structure of data. In this case, SEM is not only a noun describing a broad class of methodologies, but it is also a verb—an activity on the part of a researcher to describe and analyze a phenomenon of interest. The chapters in this handbook have described the nuances of SEM from the frequentist domain—many attending to issues of specification, power, and model modification. In this section, we consider model evaluation and comparison from the Bayesian perspective, highlighting some of the more popular techniques that can be used for evaluating BSEMs.

Posterior Predictive Checking

The general idea behind posterior predictive checking is that there should be little, if any, discrepancy between data generated by the model and the actual data. In essence, posterior predictive checking is a method for assessing the specification quality of the model from the viewpoint of predictive accuracy. Any deviation between the model-generated data and the actual data suggests possible model misspecification.

Posterior predictive checking utilizes the posterior predictive distribution of replicated data. Following

Gelman and colleagues (2014), let $\tilde{y}$ be data replicated from our current model; that is,

$$\begin{aligned} p(\tilde{y}|y) &= \int p(\tilde{y}|\theta)\,p(\theta|y)\,d\theta \\ &= \int p(\tilde{y}|\theta)\,p(y|\theta)p(\theta)\,d\theta \end{aligned} \tag{38.13}$$

Notice that the second term, $p(\theta\,|\,y)$, on the right-hand side of Equation 38.13 is simply the posterior distribution of the model parameters. In words, Equation 38.13 states that the distribution of future observations given the present data, $p(\tilde{y}\,|\,y)$, is equal to the probability distribution of the future observations given the parameters, $p(\tilde{y}\,|\,\theta)$, weighted by the posterior distribution of the model parameters. Thus, posterior predictive checking accounts for both the uncertainty in the model parameters and the uncertainty in the data.

As a means of assessing the fit of the model, posterior predictive checking implies that the replicated data should match the observed data quite closely if we are to conclude that the model fits the data. One approach to quantifying model fit in the context of posterior predictive checking incorporates the notion of Bayesian p-values. Denote by $T(y)$ a model test statistic based on the data, and let $T(\tilde{y})$ be the same test statistic but defined for the replicated data. Then, the Bayesian p-value is defined to be

$$p\text{-value} = pr\left(T(\tilde{y}) \geq T(y)\middle|y\right) \tag{38.14}$$

Equation 38.14 measures the proportion of observations in the replicated data that exceeds that of the actual data.

Examples of Bayesian Model Evaluation

As a means for assessing model fit, the posterior predictive checking procedure is demonstrated at the companion website in the context of CFA, SEM, and latent growth curve modeling.

LOOCV

Recently, model evaluation has turned to the question of cross-validation. Specifically, the question concerns how well a model predicts the data. A popular method to assess the cross-validation quality of a model is referred to as *LOOCV*, which is a special case of k-fold cross-validation (k-fold CV) when $k = n$. In k-fold CV, a sample is split into k groups (folds), and each fold is taken to be the validation set, with the remaining $k - 1$ folds serving as the training set. For LOOCV, each observation serves as the validation set with the remaining $n - 1$ observations serving as the training set. LOOCV is available in the R software program `loo` (Vehtari, Gabry, Yao, & Gelman, 2019) and implemented in `Stan` and `blavaan`.[4]

Following Vehtari, Gelman, and Gabry (2017), let y_i ($i = 1, \ldots, n$) be an n-dimensional vector of data following a distribution conditional on model parameters θ—namely, $p(y\,|\,\theta) = \prod_{i=1}^{n} p(y_i\,|\,\theta)$. Given a prior distribution on the parameters, $p(\theta)$, we can obtain the posterior distribution, $p(\theta\,|\,y)$ as well as a posterior predictive distribution of predicted values $\tilde{y}$ written as $p(\tilde{y}\,|\,y) = \int p(\tilde{y}_i\,|\,\theta)\,p(\theta\,|\,y)\,d\theta$. The Bayesian LOOCV rests on the derivation of the *expected log pointwise predictive density* (elpd) for new data, defined as

$$\text{elpd} = \sum_{i=1}^{n} \int p_t(\tilde{y}_i) \log p(\tilde{y}_i|y)\,d\tilde{y}_i \tag{38.15}$$

where $p_t(\tilde{y}_i)$ represents the distribution of the true but unknown data-generating process for the predicted values $\tilde{y}_i$ and where Equation 38.15 is approximated by cross-validation procedures. The elpd provides a measure of predictive accuracy for the n data points taken one at a time (Vehtari et al., 2017). From here, the Bayesian leave-one-out (LOO) estimate can be written as

$$\text{elpd}_{loo} = \sum_{i=1}^{n} \log p(y_i|y_{-i}) \tag{38.16}$$

where

$$p(y_i|y_{-i}) = \int p(y_i|\theta)\,p(\theta|y_{-i})\,d\theta \tag{38.17}$$

which is the LOO predictive distribution using the log predictive score to assess predictive accuracy.

It can be time-consuming to calculate exact LOOCV. To remedy this, Vehtari and colleagues (2017) developed a fast and stable approach to obtaining LOOCV referred to as *Pareto-smoothed importance sampling* (PSIS-LOO). The PSIS approach is implemented in the loo program (Vehtari et al., 2019).

Example of Bayesian Cross-Validation

The LOOCV approach is demonstrated at the companion website in the context of an SEM based on political democracy data.

BAYESIAN SEM MODEL SELECTION

The Bayesian framework does not adopt the frequentist orientation to null hypothesis significance testing. Instead, as with posterior predictive checking, a key component of Bayesian statistical modeling is a framework for model choice, with the idea that the model will be used for prediction. For this chapter, we focus on Bayes factors, the Bayesian information criterion (BIC), the deviance information criterion (DIC), and the leave-one-out information criterion (LOOIC) as methods for choosing among a set of competing models.

Bayes Factors

A very simple and intuitive approach to model building and model selection uses so-called *Bayes factors* (Kass & Raftery, 1995). An excellent discussion of Bayes factors and the problem of hypothesis testing from the Bayesian perspective can be found in Raftery (1995). In essence, the Bayes factor provides a way to quantify the odds that the data favor one hypothesis over another. A key benefit of Bayes factors is that models do not have to be nested.

To begin, consider two competing models, denoted as M_1 and M_2, that could be nested within a larger space of alternative models (i.e., > 2 models can also be compared using this method). For example, these could be two regression models with a different number of variables, or two SEMs specifying very different directions of mediating effects. Furthermore, let θ_1 and θ_2 be two parameter vectors. From Bayes' theorem, the posterior probability that, say, M_1, is the correct model can be written as

$$p(M_1|y)=\frac{p(y|M_1)p(M_1)}{p(y|M_1)p(M_1)+p(y|M_2)p(M_2)} \quad (38.18)$$

Notice that $p(y \mid M_1)$ does not contain model parameters θ_1. To obtain $p(y \mid M_1)$ requires integrating over θ_1; that is,

$$p(y|M_1)=\int p(y|\theta_1,M_1)p(\theta_1|M_1)d\theta_1 \quad (38.19)$$

where the terms inside the integral are the likelihood and the prior, respectively. The quantity $p(y \mid M_1)$ has been referred to as the *integrated likelihood* for model M_1 (Raftery, 1995). Perhaps a more useful term is the *predictive probability of the data* given M_1. A similar expression can be written for M_2.

With these expressions, we can move to comparing our two models, M_1 and M_2. The goal is to develop a quantity expressing the extent to which the data support M_1 over M_2. One quantity could be the posterior odds of M_1 over M_2, expressed as

$$\frac{p(M_1|y)}{p(M_2|y)}=\frac{p(y|M_1)}{p(y|M_2)}\times\left[\frac{p(M_1)}{p(M_2)}\right] \quad (38.20)$$

Notice that the first term on the right-hand side of Equation 38.20 is the ratio of two integrated likelihoods. This ratio is referred to as the *Bayes factor* for M_1 over M_2, denoted here as B_{12}. In line with Kass and Raftery (1995, p. 776), our prior opinion regarding the odds of M_1 over M_2, given by $p(M_1)/p(M_2)$ is weighted by our consideration of the data, given by $p(y \mid M_1)/p(y \mid M_2)$. This weighting gives rise to our updated view of evidence provided by the data for either hypothesis, denoted as $p(M_1 \mid y)/p(M_2 \mid y)$. An inspection of Equation 38.20 also suggests that the Bayes factor is the ratio of the posterior odds to the prior odds.

In practice, there may be no prior preference for one model over the other. In this case, the prior odds are neutral and $p(M_1) = p(M_2) = 1/2$. When the prior odds ratio equals 1, then the posterior odds is equal to the Bayes factor.

The Bayesian Information Criterion

A popular measure for model selection used in both frequentist and Bayesian applications is based on an approximation of the Bayes factor and is referred to as the *BIC*, also referred to as the Schwarz criterion (Schwarz, 1978). Raftery (1995) examines generalizations of the BIC to a broad class of models, but it is important to note there are many current criticisms of the BIC; please see the companion website for these.

Consider again two models, M_1 and M_2, with M_2 nested in M_1. Under conditions where there is little prior information, Raftery (1995) has shown that an approximation of the Bayes factor can be written as

$$2\log B_{12}=\chi^2_{12}-df_{12}\log n \quad (38.21)$$

where χ^2_{12} is the conventional likelihood ratio chi-square obtained from testing M_1 against M_2; df_{12} is the difference in the degrees of freedom associated with each test.

The BIC can be written as

$$BIC=-2\log\left(\hat{\theta}\middle|y\right)+q\log(n) \quad (38.22)$$

where $-2\log\left(\hat{\theta} \mid y\right)$ describes model fit, $q\log(n)$ is a penalty for model complexity, where q represents the number of variables in the model, and n is the sample size.

As with Bayes factors, the BIC is often used for model comparison. Specifically, the difference between two BIC measures comparing, say M_1 to M_2, can be written as

$$\begin{aligned}\Delta\left(BIC_{12}\right) &= BIC_{(M_1)} - BIC_{(M_2)} \\ &= \log\left(\hat{\theta}_1 \middle| y\right) - \log\left(\hat{\theta}_2 \middle| y\right) \\ &-\frac{1}{2}\left(q_1 - q_2\right)\log\left(n\right)\end{aligned} \tag{38.23}$$

Rules of thumb have been developed to assess the quality of the evidence favoring one model over another using Bayes factors and the comparison of BIC values from two competing models. Following Kass and Raftery (1995, p. 777) and using M_1 as the reference model:

BIC difference	Bayes factor	Evidence against M_2
0 to 1	1 to 3	Weak
2 to 6	3 to 20	Positive
6 to 10	20 to 150	Strong
> 10	> 150	Very strong

The Deviance Information Criterion

Although the BIC is derived from a fundamentally Bayesian perspective, it is often productively used for model comparison in the frequentist domain. Another model comparison method derived in the Bayesian framework was developed by Spiegelhalter, Best, Carlin, and van der Linde (2002) based on the notion of *Bayesian deviance.*

Consider a particular probability model for a set of data, defined as $p(y \mid \theta)$. Then, *Bayesian deviance* can be defined as

$$D\left(\theta\right) = -2\log\left[p\left(y\middle|\theta\right)\right] + 2\log\left[h\left(y\right)\right] \tag{38.24}$$

where, according to Spielgelhalter and colleagues (2002), the term $h(y)$ is a standardizing factor that does not involve model parameters and thus is not involved in model selection. Note that although Equation 38.24 is similar to the BIC, it is not, as currently defined, an explicit Bayesian measure capturing model comparison. To accomplish this, we use Equation 38.24 to obtain a posterior mean over θ by defining

$$DIC = E_\theta\left\{-2\log\left[p\left(y\middle|\theta\right)\middle|y\right] + 2\log\left[h\left(y\right)\right]\right\} \tag{38.25}$$

where E_θ is the expected value of θ. Similar to the BIC, the model with the smallest DIC among a set of competing models is preferred.

A criticism of the DIC is that it rests on the assumption that the posterior distribution is normal. Moreover, some have argued that the DIC is not fully Bayesian insofar as it rests on point estimates of the posterior distribution and does not exploit the entire posterior distribution. In addition, because the data are being used to both provide the point estimate and evaluate the model, it tends select overfitted models.

LOOIC

It is useful to note that an information criterion based on LOO (*LOOIC*) can be easily derived as

$$\text{LOOIC} = -2\widehat{\text{elpd}}_{loo} \tag{38.26}$$

which places the LOOIC on the "deviance scale" (see Vehtari et al., 2017, for more details on implementing the LOOIC in `loo`). Among a set of competing models, the one with the smallest LOOIC is considered best from an out-of-sample pointwise predictive point of view. The LOOIC is available in `blavaan` (Merkle & Rosseel, 2018).

Examples of Bayesian Model Selection

Various Bayesian model selection indices are demonstrated at the companion website for selecting across different forms of a Bayesian CFA model examining positive and negative experiences with reading. These indices are used to select across models implementing different patterns of cross-loadings and different prior distributions.

BAYESIAN MODEL AVERAGING IN SEM

The various approaches to SEM model selection ultimately lead to a single model to be used for explanation and/or prediction. The problem with choosing a single model, regardless of the method of model selection, is that the uncertainty manifested in the selection process is not taken into account. Nevertheless, this uncertainty is propagated into inferences and decisions based on the chosen model. In other words, upon selecting a final model, one is behaving as though the model's posterior probability given the data is 1.0, and this can lead to "over-confident inferences and decisions that are more risky than one thinks they are" (Hoeting, Madigan, Raftery, & Volinsky, 1999, p. 382).

A solution to the problem of capturing model uncertainty in the context of SEM model selection is to simply not select a final model, but rather average over a (large) space of models that could have generated the data. This is the method of *Bayesian model averaging* (BMA; e.g., Clyde, 2003; Draper, 1995; Hoeting et al., 1999; Madigan & Raftery, 1994; Raftery, Madigan, & Hoeting, 1997; Raftery & Zheng, 2003; Steel, 2020).

Statistical Specification of BMA

Following Madigan and Raftery (1994), consider a quantity of interest such as a future observation, denoting this quantity as $\tilde{y}$. Our goal is to obtain an optimal prediction of $\tilde{y}$, in the sense that the utility of predicting $\tilde{y}$ is maximized. Next, consider a set of competing SEMs (substantively selected by the researcher), $\mathcal{M} = \{M_k\}_{k=1}^{K}$, that are not necessarily nested. The posterior distribution of $\tilde{y}$ given data D can be written as a mixture distribution

$$p(\tilde{y}|D) = \sum_{k=1}^{K} p(\tilde{y}|M_k)\, p(M_k|D) \tag{38.27}$$

where $p(M_k|D)$ is the posterior probability of model M_k written as

$$p(M_k|D) = \frac{p(D|M_k)\, p(M_k)}{\sum_{l=1}^{K} p(D|M_l)\, p(M_l)} \tag{38.28}$$

where the first term in the numerator on the right-hand side of Equation 38.28 is the probability of the data given model k, also referred to as the *integrated likelihood* written as

$$p(y|M_k) = \int p(y|\theta_k, M_k)\, p(\theta_k|M_k)\, d\theta_k \tag{38.29}$$

where $p(\theta_k|M_k)$ is the prior distribution of the parameters θ_k under model M_k (Raftery et al., 1997). The posterior model probabilities can be considered mixing weights associated with the mixture distribution given in Equation 38.27 (Clyde & Iversen, 2013). The second term $p(M_k)$ on the right-hand side of Equation 38.28 is the prior model probability for model k, allowing each model to have a different prior probability based on past performance of that model or a belief regarding which of the models might be the true model.[5] The denominator of Equation 38.28 ensures that $p(M_k|y)$ integrates to 1.0, as long as the true model is in the set of models under consideration. The case where the true model is not in the set of competing models is discussed further in Bernardo and Smith (2000), Clyde and Iversen (2013), and Kaplan (2021).

Following Kaplan and Lee (2016) closely, the approach to BMA for SEMs draws on the fact that path diagrams within the SEM tradition can be seen as special cases of so-called *directed acyclic graphs* (DAGs), the latter having been developed by Pearl (2009). BMA over DAGs has also been discussed in Madigan and Raftery (1994), but only recently connected to SEM.

The general steps of Kaplan and Lee's (2016) algorithm are as follows: (1) specify an initial model of interest recognizing that this may not be the model that generated the data; (2) starting with the initial model represented as a DAG, implement a search over the DAG to reduce the space of models to a reasonable size while maintaining the distinction between exogenous, mediating, and endogenous variables; (3) obtain the posterior model probabilities for each model; and (4) obtain the weighted average of structural parameters over each model, weighted by the posterior model probabilities.

In their paper, Kaplan and Lee (2016) compared the predictive performance of the BMA SEM to the initially specified BSEM by computing the reduced form of the models and calculating the log score and the predictive coverage. The results revealed that when the true model is known, BMA does not yield necessarily better predictive performance compared to nonaveraged models, particularly for large sample sizes. However, their case study using data from an international, large-scale assessment reveals that the BMA provides modestly better posterior predictive performance compared to the initially specified model.

Example of BMA within SEM

The companion website includes an example of implementing BMA for an SEM using Programme for International Student Assessment data predicting a reading assessment. An inclusion probability is provided for all model parameters, highlighting some parameters with high inclusion probabilities and others with lower probabilities of inclusion in the predictive model.

FLEXIBILITY OF BAYESIAN SEM VIA PRIORS

The Bayesian estimation framework has the ability to enhance the richness of results obtained within SEM. Much of this is due to the use of prior distributions, which allow for probabilistic interpretations of final model results. Priors can be used to incorporate substantive theory within the estimation process. In addition, the specification of priors can provide added flexibility to research questions examined through SEMs. In this section, we highlight some methods for using priors as a flexible approach for model specification. We also discuss the important issue of conducting a prior sensitivity analysis when implementing BSEM.

Implementing Near-Zero Priors for Added Flexibility

An attractive feature of Bayesian estimation for SEMs is the added flexibility that it provides. Construction of SEMs is based on patterns of fixed and free parameters. At least in some cases, parameters that are fixed to zero may not be *exactly* zero in the population. Treating such a parameter as fixed to zero can embed specification errors into the model. For example, the CFA model is a highly restricted model that is traditionally estimated with many cross-loadings fixed to zero. Within the frequentist framework, freeing all cross-loadings can result in a nonidentified model. However, Bayesian methods can be used to mitigate this identification problem through use of implementing *near-zero* or *approximate-zero* priors (Muthén & Asparouhov, 2012). This Bayesian approach allows for restrictions of fixed cross-loadings to be relaxed.

Near-zero priors (typically) follow a normal distribution with a mean hyperparameter of zero and a narrowed variance hyperparameter. The small-variance hyperparameter ensures that the prior is highly informative surrounding zero. Models such as the CFA can incorporate priors reflecting this near-zero status for secondary or negligible cross-loadings (see, e.g., Moore, Reise, Depaoli, & Haviland, 2015). These cross-loadings are no longer fixed to zero in the model, which implies that restrictions implemented in frequentist CFA have been relaxed. Through use of these priors, a less restricted version of the CFA can be estimated, which may in turn improve interpretations or model assessment.

Ultimately, the researcher determines how precise the near-zero prior should be. The variance hyperparameter is associated with the strength of the researcher's beliefs. Smaller (nonzero) variances indicate a relatively stronger belief that the loadings are zero. As the magnitude of the variance hyperparameter increases, the belief of the parameter being equal to zero relaxes. Larger variance hyperparameters allow the data to determine the loading strength to a greater extent.

Near-zero priors placed on cross-loadings allow for some "wiggle" room surrounding the cross-loadings without changing the substantive meaning of the factors. Successful application of near-zero priors requires a careful assessment of the variance hyperparameter settings. These priors should allow for meaningful cross-loadings to be estimated, while keeping the negligible cross-loadings close to zero. One recommended strategy is to start with a very small variance hyperparameter (e.g., 0.001), then incrementally increase the size until model fit improvements diminish or the model results lose substantive interpretability. For this approach, indices such as the DIC and posterior predictive *p*-value have been recommended for selecting the optimal variance hyperparameter setting for near-zero priors (cf. Asparouhov, Muthén, & Morin, 2015; Pokropek, Schmidt, & Davidov, 2020).[6]

Near-zero priors are not only useful when considering nonzero cross-loadings in a CFA. In fact, there are many other applications within SEM where near-zero priors allow for a more flexible framing of the model. Notably, near-zero priors can be extended into the case of assessing for measurement invariance (MI). Traditional MI approaches typically hold parameters to be exactly equal across groups (or time, in the case of longitudinal assessments of MI) during the different stages of invariance testing.[7] Near-zero priors provide an alternative approach, where *approximate equivalence* is examined as opposed to *exact equivalence* across groups. The Bayesian approach allows for small differences in parameter estimates across groups rather than holding the parameters equal. For example, the traditional MI testing approach is to assess whether the difference for a specific parameter between Group 1 and Group 2 is equal to zero. The Bayesian approach with near-zero priors does not require the parameter to be exactly equal across groups for invariance to hold. Instead, the difference between the Group 1 and Group 2 parameter estimates would be *approximately* zero. This approximation adds flexibility to what is considered to be invariant.

The near-zero prior is placed on a difference parameter, which captures the difference of a single parameter between two groups. For example, assume we are interested in examining group differences for a factor loading for Item 2 on Factor 1. The near-zero prior would capture the *difference* between the loading for Group 1 (G1) and the loading for Group 2 (G2). In traditional (frequentist) MI approaches, this difference would be set to zero, representing exact equivalence between the two groups. In contrast to this traditional approach, Bayesian approximate MI allows the factor loading difference to vary (even if only slightly) from zero through the use of a near-zero prior. An example of this prior looks as follows:

$$\lambda_{21}^{(G1)} - \lambda_{21}^{(G2)} \sim N[0,\ 0.001] \tag{38.30}$$

where the loading for Factor 1, Item 2 (λ_{21}) is compared across groups (G1 and G2, in this case) by setting up a difference parameter. The difference between the groups is assumed to be normally distributed ($\mathcal{N}$), with a mean hyperparameter of zero and a variance hyperparameter set to some predetermined value specified by the researcher (e.g., 0.001 in this example).

Overall, there are many benefits to using the Bayesian approximate MI approach. More accurate parameter estimates can be obtained, small (nonzero) cross-loadings can be included in the measurement model, and the Bayesian approximate MI approach performs better (e.g., through fit measures) than partial MI when parameter differences are small (Muthén & Asparouhov, 2013; Pokropek, Davidov, & Schmidt, 2019; van de Schoot et al., 2013). In addition, the use of near-zero priors allows for approximate invariance to be examined even when exact invariance does not hold for model parameters (cf. Winter & Depaoli, 2019). However, in order for the Bayesian approximate MI approach to yield correct interpretations, differences between parameters across groups must be small and nonsystematic; for more on this topic, see the *alignment issue* in Muthén and Asparouhov (2013).

Example of Near-Zero Priors

The companion website includes an example of implementing near-zero priors in the context of CFA. Rather than fixing cross-loadings to zero, they are treated as being *approximately* zero through a near-zero prior, highlighting a more flexible version of the conventional CFA specification.

Prior Specification and Sensitivity Analysis

There are many differences between Bayesian methods and frequentist estimation, but the most notable is the presence of the prior distribution. Prior distributions play an important role in any Bayesian model, but they have the ability to be especially impactful in the context of SEMs.

Scaling and parameter transformations are common within SEM (see e.g., Bollen, 1989; Grimm & Liu, 2016; Kline, 2016; Muthén & Asparouhov, 2002; Ulitzsch, Holtmann, Schultze, & Eid, 2017). Within the frequentist framework, the choice to specify SEM parameters a certain way is largely to improve estimation efficiency or the interpretation of model results, but the model form or likelihood is seldom affected. However, modifying the specification of a parameter within the Bayesian framework typically requires a new set of priors. For example, if a covariance parameter is respecified as a correlation, then a new prior distribution is likely to be specified as well. The prior for the covariance would need to accommodate a potentially large range of values because a covariance can be any number, but a prior for a correlation may be bounded between ± 1 to prevent impossible values.

Not only can individual priors be altered in a model (e.g., altering a prior for a parameter specified as a covariance vs. a correlation), but the prior distribution *strategy* can also be changed. Covariance matrices are commonly embedded within SEMs (e.g., latent factor covariances in a multifactor CFA), and there are two different strategies that can be used when specifying priors for the matrix. The first method is to treat the covariance matrix as a whole and specify a multivariate prior distribution for the entire matrix (e.g., specifying an inverse Wishart distribution for a covariance matrix). A second method is to use a *separation strategy* of prior implementation, where univariate priors are placed on each element of the matrix. For example, a covariance matrix can be decomposed into standard deviation (diagonal) and correlation (off-diagonal) elements (Barnard, McCulloch, & Meng, 2000). In this case, the standard deviations may be specified with half-Cauchy priors (for example), and the correlations can be specified with bounded uniform priors.[8] Research has indicated that, with all else held equal, multivariate priors produce final model results that are quite different from separation strategy priors within SEM (Depaoli, Liu, & Marvin, 2021; Liu, Zhang, & Grimm, 2016). In some cases, there is an advantage to using separation strategy priors for SEMs, in that results obtained are more accurate in some simulation conditions as compared to multivariate priors.

Capturing the influence that priors have on estimated posteriors is one of the most important aspects of model interpretation, especially given the various ways that priors can be specified. One method that can aid in better understanding a prior's influence is to conduct a thorough sensitivity analysis of the prior. A sensitivity analysis can be used to examine the impact that priors have on final model results.[9] It allows the researcher to examine the impact of prior settings on final results in a methodical manner.

The researcher will often specify original priors based on desired previous knowledge (or even through use of software default settings). After posteriors are estimated and inferences are described, the researcher can then examine the robustness of results to deviations in the priors specified in the original model. For example, if a prior for a factor loading in the original analysis is $\mathcal{N}(0.6, 0.1)$, then the mean and variance hyperparameters can be systematically altered in a prior sensitivity analysis. The prior, $\mathcal{N}(\mu, 0.1)$, can be altered in the following way:

- Original setting: $\mu = 0.6$.
- Examine settings lower than 0.6, where $\mu = -0.2$, 0, 0.2, and 0.4.
- Examine settings greater than 0.6, where $\mu = 0.8$ and 1.0.

Next, the variance hyperparameter can be altered, while keeping the mean hyperparameter at 0.6. The prior, $\mathcal{N}(0.6, \sigma^2)$, can be altered in the following way.

- Original setting: $\sigma^2 = 0.1$.
- Examine settings lower than 0.1, where $\sigma^2 =0.05$ and 0.01.
- Examine settings greater than 0.1, where $\sigma^2 =0.5$, 1, 10, and 100.

Then the settings for the mean and variance hyperparameters would be fully crossed to form a thorough sensitivity analysis. In other words, each of the mean hyperparameter settings listed would be examined under all of the variance hyperparameter settings.

Many Bayesian researchers (see, e.g., Depaoli & van de Schoot, 2017; Kruschke, 2015; Muthén & Asparouhov, 2012) recommend that model results be presented alongside a sensitivity analysis. This information will help to promote a more thorough understanding of the robustness of the findings. It is important to note here that there is no *right* or *wrong* result within a prior sensitivity analysis. If findings indicate that a posterior is heavily influenced by the prior setting, then that is not necessarily a problem because much can be learned from the discrepancy between the data and the prior (e.g., it could be that theory, via particular settings of the prior, influences the posterior to a great extent). This type of finding may highlight the need for careful theory building. At the very least, it can address the need for reporting full sensitivity analysis results to highlight varying inferences based on different priors or models. The point here is to be transparent about the specification and impact of the priors, and a sensitivity analysis can be a valuable tool for understanding the influence of prior settings in BSEM.

Example of a Sensitivity Analysis

The companion website includes an example of a prior sensitivity analysis using the latent growth curve model. Priors were systematically altered for the latent intercept mean to highlight the robustness of results under different priors.

Bayesian Regularization

One potential advantage of BSEM is that the incorporation of prior information can supplement the information provided by the data. This aspect of Bayesian methodology is popular, in part, because it can allow for complex models (e.g., models with many predictors) to be estimated using a sample of data that is relatively small in size. In this case, the information provided by the relatively small sample size would be compensated for by including additional information via the priors (e.g., through implementing [weakly] informative priors).

However, even with the use of more informative priors, there are some instances where overfitting the model becomes a concern. Overfitting can occur when the number of predictors in the model is relatively large in comparison to the number of observations in the sample. For example, when the number of predictors increases to a level approaching or larger than the number of subjects in the sample, then the model is considered to be overfit according to the sample. Overfitting the model in this manner can result in a lack of generalizability to other samples.

There are several penalization methods that can be implemented to help mitigate this issue, whether the researcher is working in the frequentist or Bayesian framework. Regardless of the method implemented, penalization techniques are used to shrink small coefficients toward zero and allow large coefficients to remain large. This process can increase estimate bias. However, it also avoids overfitting the model to that particular data set. By avoiding overfitting, there is a greater degree of generalizability to other samples of data despite the increased bias levels.

In some cases, Bayesian penalization results are comparable to frequentist penalized estimates (Park & Casella, 2008). However, this should not imply that the choice between Bayesian and frequentist penalization methods is an arbitrary one. van Erp (2020) details three main benefits for implementing penalization using Bayesian methods.

The first benefit is that there is a natural implementation of penalization via priors within Bayesian methods. When using Bayesian estimation, penalization techniques are incorporated through *shrinkage priors*. Shrinkage priors can be strategically specified to shrink small coefficients toward zero. Under Bayesian penalization, the penalty term is captured by a hyperparameter within the shrinkage prior. This hyperparameter is defined through its own *hyperprior* (a prior distribution placed on the hyperparameter). The hyperprior can be manipulated to increase or reduce the amount of shrinkage in estimated effects. A larger penalty term (via the shrinkage prior) will increase the amount of shrinkage, whereas a penalty term of zero will not produce shrinkage of effects. The second benefit is that the penalty term is estimated in the same step as the other model parameters. In other words, the penalty term is built into the model estimation process, since it is incorporated directly into the model via a prior. In turn, that prior can be specified in a flexible manner through different settings, controlling for the degree of shrinkage as the researcher sees fit. The third benefit of estimating Bayesian penalty terms is that many different forms of penalties can be implemented. There are classic penalty techniques, such as the ridge and lasso methods, which can be incorporated easily with Bayesian counterparts. In addition, there are methods that are exclusively Bayesian in nature, including elastic net (a hybrid of the ridge and lasso approaches, although not yet implemented for SEMs), the spike-and-slab prior, horseshoe prior, and so forth; for more information on different forms of shrinkage priors, see van de Schoot and colleagues (2021). For simplicity, we confine our discussion to the ridge (Hoerl & Kennard, 1970; Hsiang, 1975) and lasso (Park & Casella, 2008; Tibshirani, 1996) penalty methods.

Ridge Regression

Ridge regression is a penalization method that can be used to avoid the issues resulting from situations with a large number of predictors and a relatively small sample size. This approach uses a constraint on the model parameters such that $\hat{\beta}$ is selected to minimize the penalized sum of squares written as follows:

$$\hat{\beta}^{ridge} = \arg\min \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} x_{ij}\beta_j \right)^2 + \nu \sum_{j=1}^{k} \beta_j^2 \right\} \quad (38.31)$$

where there are $i = 1, \ldots, n$ subjects and $j = 1, \ldots, k$ parameters, x is the predictor, y is the outcome, β is a regression coefficient (e.g., a path coefficient in SEM), and ν is a penalty term that controls the amount of shrinkage, with larger values of ν magnifying the amount of shrinkage of β_j parameters toward zero. The ridge regression process is sometimes also referred to as L2-norm regularization because the penalty added $\left(\nu \sum_{j=1}^{k} \beta_j^2\right)$ is equivalent to squaring the magnitude of the regression coefficients.

In Bayesian ridge regression, the penalty term (ν) is captured through normally distributed priors placed on the regression slope parameters. The normal distribution mean hyperparameter is fixed at zero within this approach in order to control shrinkage toward zero. The variance hyperparameter is typically rescaled to be in standard deviation form. That hyperparameter defines the degree of spread that the distribution exhibits. In addition, putting the scale hyperparameter in terms of a standard deviation allows the half-Cauchy hyperprior to be placed on it.

Lasso Regression

The lasso penalization procedure produces estimates as follows:

$$\hat{\beta}^{lasso} = \arg\min\left\{\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{k} x_{ij}\beta_j\right)^2 + \nu\sum_{j=1}^{k}\left|\beta_j\right|\right\} \quad (38.32)$$

where the L1-norm regularization is used, which adds a penalty $\left(\nu\sum_{j=1}^{k}\left|\beta_j\right|\right)$ that is equivalent to taking the absolute value of the magnitude of the regression coefficients.

Bayesian lasso penalization uses a different shrinkage prior as compared to the Bayesian ridge approach. Specifically, the lasso method can be implemented in the Bayesian framework by specifying a double exponential prior on regression slopes. The double exponential distribution is also referred to as the "Laplace distribution." This distribution is ideal because it peaks at zero, which shrinks small coefficients toward zero. However, the double exponential can be set to have thick tails (in both directions), allowing the large coefficients to remain large. Given that the distribution is centered at zero to control shrinkage toward zero, the mean hyperparameter setting is always fixed to zero. The scale, or dispersion, of the double exponential distribution is the hyperparameter that researchers can alter when implementing the penalization. That hyperparameter defines the amount of spread and the thickness of the tails, which controls the degree of shrinkage in coefficients. A half-Cauchy hyperprior can be specified on the standard deviation hyperparameter, akin to the Bayesian ridge approach.

Although the ridge and lasso approaches are similarly implemented in the Bayesian framework, these techniques can produce different penalized coefficient results. Depending on the hyperparameter settings, the lasso approach can result in more shrinkage for the small estimates but less shrinkage for the large estimates. This result is a function of the double exponential distribution implemented in the lasso approach. The double exponential distribution is more peaked around zero, and it has heavier tails compared to the normal distribution used in the ridge approach. Regardless of the approach implemented, Bayesian penalization can be a useful tool when attempting to avoid overfitting a complex model to small samples. In addition, these approaches further highlight the modeling flexibility that Bayesian methods provide through the flexible implementation of priors.

Two Examples of Bayesian Penalized SEM

The companion website includes two examples of Bayesian penalization. Both examples fit a multiple indicator, multiple cause (MIMIC) model with several model predictors to Holzinger and Swineford (1939) data. The first example uses the Bayesian ridge approach, with a normally distributed prior used for the penalty term. The second example implements the Bayesian lasso method, with a double exponential distribution specified for the penalty term. A comparison between approaches is also provided.

DISCUSSION

We have sought in this chapter to present an accessible introduction to BSEM. An overview of Bayesian concepts as well as a brief introduction of Bayesian computation has also been provided. A general framework of Bayesian computation within the BSEM framework was also presented, along with several examples covering first- and second-generation SEM (presented at the companion website). With the advent of open source software for Bayesian computation, such as packages found in R (R Development Core Team, 2008) and programs using the BUGS language (Lunn, Thomas, Best, & Spiegelhalter, 2000), as well as the available MCMC estimator in *Mplus* (Muthén & Muthén, 1998–2017), researchers can now implement Bayesian methods for a wide range of research problems.

The relative ease of Bayesian computation in the SEM framework raises the important question of why one would choose to use this method—particularly when it can often provide results that are very close to that of frequentist approaches such as maximum likelihood. In our judgment, the answer lies in the major distinction between the Bayesian approach and the frequentist approach; that is, in the elicitation, specification, and incorporation of prior distributions on the model parameters. In addition, the Bayesian estimation framework increases modeling flexibility, allowing for a wider range of models to be estimated.

As Skrondal and Rabe-Hesketh (2004, p. 206) point out, there are four reasons why one would adopt the use of prior distributions, one of which is "truly" Bayesian, while the others represent a more "pragmatic" approach to Bayesian inference. The truly Bayesian approach would specify prior distributions that reflect elicited prior knowledge. For example, in the context of SEM applied to educational problems, one might specify a normal prior distribution on the regression coefficient relating socioeconomic status (SES) to achievement, where the hyperparameter on the mean of the regression coefficient is obtained from previous research. Given that an inspection of the literature suggests roughly the same values for the regression coefficient, a researcher might specify a small value for the variance of the regression coefficient, reflecting a high degree of precision. Pragmatic approaches, on the other hand, might specify prior distributions for the purposes of achieving model identification, constraining parameters so they do not drift beyond their boundary space (e.g., Heywood cases), or simply because the application of MCMC can sometimes make problems tractable that would otherwise be very difficult in more conventional frequentist settings.

Although we concur with the general point that Skrondal and Rabe-Hesketh (2004) are making, we do not believe that the distinction between "true" Bayesians and "pragmatic" Bayesians is necessarily the correct one to be made. If there is a distinction to be made, we argue that it is between Bayesians and pseudo-Bayesians, where the latter implement MCMC as "just another estimator." Rather, we adopt the pragmatic perspective that the usefulness of a model lies in whether it provides good predictions. The specification of priors based on subjective knowledge can be subjected to quite pragmatic procedures in order to sort out the best predictive model, such as the use of posterior predictive checking.

In addition, this chapter has highlighted methods that can potentially improve the predictive nature of a model, including through BMA and the use of near-zero priors. Bayesian methodology offers a range of tools that can be used to enhance the way in which substantive questions in SEM are explored, but much of this added flexibility is tied to the specification of the priors.

What Bayesian theory forces us to recognize is that it is possible to bring in prior information on the distribution of model parameters, but that this requires a deeper understanding of the "elicitation problem" (see Abbas, Budescu, & Gu, 2010; Abbas, Budescu, Yu, & Haggerty, 2008; O'Hagan et al., 2006). The general idea is that through a careful review of prior research on a problem, and/or the careful elicitation of prior knowledge from experts and/or key stakeholders, relatively precise values for hyperparameters can be obtained and incorporated into a Bayesian specification. Alternative elicitations can be directly compared via Bayesian model selection measures as described earlier. It is through (1) the careful and rigorous elicitation of prior knowledge, (2) the incorporation of that knowledge into our statistical models, and (3) a rigorous approach to the selection among competing models that a pragmatic *and* evolutionary development of knowledge can be realized. This is precisely the advantage that Bayesian statistics, and BSEM in particular, has over its frequentist counterparts. Now that the theoretical and computational foundations have been established, the benefits of BSEM will be realized in terms of how it provides insights into important substantive problems.

NOTES

1. The *support* of a distribution is the smallest closed interval (or set in the multivariate case). The elements of the interval/set are members of the distribution. Outside the support of the distribution, the probability of the element is zero.

2. For ease of notation, we are suppressing conditioning on the data *y*.

3. Note that in the case where there is only one element in the block, the prior distribution is assumed to be inverse-gamma (i.e., $\boldsymbol{\theta}_{IW} \sim IG(a, b)$).

4. The *widely applicable information criterion* (WAIC) has also been advocated for model selection. Although the WAIC and LOOCV are asymptotically equivalent (Watanabe, 2010), the implementation of LOOCV in the loo package is more robust in finite samples with weak priors or influential observations (Vehtari et al., 2017)

5. The existence of a true model is an important topic in BMA and is beyond the scope of this chapter. See Bernardo and Smith (2000) and Kaplan (2021).

6. Asparouhov et al. (2015) provide a helpful list of steps for implementing near-zero priors. However, we note that even taking these criteria into account, the decision of what variance hyperparameter to use is highly subjective.

7. Although we describe the near-zero priors in the context of assessing MI across groups, the same concepts can be applied to longitudinal studies of MI akin to Liu and West (2018).

8. With larger factor structures (e.g., three or more), independent priors on the factor standard deviations and correlations will not always lead to positive definite matrices. Researchers implementing this strategy with more than two

factors should be mindful to check the positive definite status of the resulting matrices.

9. The same process can be used for conducting a sensitivity analysis on the likelihood (via the statistical model).

REFERENCES

Abbas, A. E., Budescu, D. V., & Gu, Y. (2010). Assessing joint distributions with isoprobability countours. *Management Science, 56*, 997–1011.

Abbas, A. E., Budescu, D. V., Yu, H.-T., & Haggerty, R. (2008). A comparison of two probability encoding methods: Fixed probability vs. fixed variable values. *Decision Analysis, 5*, 190–202.

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

Barnard, J., McCulloch, R., & Meng, X.-L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica, 10*, 1281–1311.

Bernardo, J. M., & Smith, A. F. M. (2000). *Bayesian theory*. New York: Wiley.

Betancourt, M. (2018). *A conceptual introduction to Hamiltonian Monte Carlo*. Retrieved from https://arxiv.org/pdf/1701.02434.pdf.

Betancourt, M. (2019). *Probalistic computation*. https://betanalpha.github.io/assets/case_studies/probabilistic_computation.html.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics, 7*, 434–455.

Clyde, M. A. (2003). Model averaging. In S. J. Press (Ed.), *Subjective and objective Bayesian statistics* (pp. 320–335). Hoboken, NJ: Wiley-Interscience.

Clyde, M. A., & Iversen, E. S. (2013). Bayesian model averaging in the M-open framework. In P. Damien, P. Dellaportas, N. G. Polson, & D. A. Stephens (Eds.), *Bayesian theory and applications* (pp. 483–498). Oxford, UK: Oxford University Press.

Depaoli, S., Liu, H., & Marvin, L. (2021). Parameter specification in Bayesian CFA: An exploration of multivariate and separation strategy priors. *Structural Equation Modeling: A Multidisciplinary Journal, 28*, 699–715.

Depaoli, S., & van de Schoot, R. (2017). Improving transparency and replication in Bayesian statistics: The WAMBS-checklist. *Psychological Methods, 22*, 240–261.

Draper, D. (1995). Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society B, 57*, 55–98.

Gelman, A. (1996). Inference and monitoring convergence. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (Eds.), *Markov Chain Monte Carlo in practice* (pp. 131–143). New York: Chapman & Hall.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D., Vehatari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (3rd ed.). London: Chapman & Hall.

Gelman, A., & Rubin, D. B. (1992a). Inference from iterative simulation using multiple sequences. *Statistical Science, 7*, 457–511.

Gelman, A., & Rubin, D. B. (1992b). A single series from the Gibbs sampler provides a false sense of security. In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian statistics 4* (pp. 625–631). Oxford, UK: Oxford University Press.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). Introducing Markov Chain Monte Carlo. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (Eds.), *Markov Chain Monte Carlo in Practice* (pp. 1–19). London: Chapman and Hall.

Gill, J. (2002). *Bayesian methods: A social and behavioral sciences approach*. London: Chapman & Hall/CRC.

Grimm, K., & Liu, Y. (2016). Residual structures in growth models with ordinal outcomes. *Structural Equation Modeling: A Multidisciplinary Journal, 23*, 466–475.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12* (1), 55–67.

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science, 14*, 382–417.

Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn Sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research, 15*, 1593–1623.

Holzinger, K., & Swineford, F. (1939). *A study in factor analysis: The stability of a bifactor solution* (Supplementary educational monograph, No. 48). Chicago: University of Chicago Press.

Hsiang, T. C. (1975). A Bayesian view on ridge regression. *The Statistician, 24*, 267–268.

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford, UK: Oxford University Press.

Kaplan, D. (2023). *Bayesian statistics for the social sciences* (2nd ed.). New York: Guilford Press.

Kaplan, D. (2021). On the quantification of model uncertainty: A Bayesian perspective. *Psychometrika, 86*(1), 215–238.

Kaplan, D., & Depaoli, S. (2012). Bayesian structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 650–673). New York: Guilford Press.

Kaplan, D., & Lee, C. (2016). Bayesian model averaging over directed acyclic graphs with implications for the predictive performance of structural equation models. *Structural Equation Modeling, 23*, 343–353.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*, 773–795.

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). New York: Guilford Press.

Kruschke, J. K. (2015). *Doing Bayesian analysis: A tutorial with R, Jags, and STAN*. San Diego, CA: Elsevier.

Liu, H., Zhang, Z., & Grimm, K. J. (2016). Comparison of inverse Wishart and separation-strategy priors for Bayesian estimation of covariance parameter matrix in growth curve analysis. *Structural Equation Modeling: A Multidisciplinary Journal, 23*, 354–367.

Liu, Y., & West, S. G. (2018). Longitudinal measurement non-invariance with ordered-categorical indicators: How are the parameters in second-order latent linear growth models affected? *Structural Equation Modeling: A Multidisciplinary Journal, 25*, 762–777.

Lunn, D., Thomas, A., Best, N., & Spiegelhalter, D. (2000). Winbugs—A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing, 10*, 325–337.

Madigan, D., & Raftery, A. E. (1994). Model selection and accounting for model uncertainly in graphical models using Occam's window. *Journal of the American Statistical Association, 89*, 1535–1546.

Mengersen, K. L., Robery, C. P., & Guihenneuc-Jouyax, C. (1999). MCMC convergence diagnostics: A review. *Bayesian Statistics, 6*, 415–440.

Merkle, E. C., & Rosseel, Y. (2018). blavaan: Bayesian structural equation models via parameter expansion. *Journal of Statistical Software, 85*(4), 1–30.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics, 21*, 1087–1091.

Moore, T. M., Reise, S. P., Depaoli, S., & Haviland, M. G. (2015). Iteration of partially specified target matrices in exploratory and Bayesian confirmatory factor analysis. *Multivariate Behavioral Research, 50*, 149–161.

Muthén, B. O., & Asparouhov, T. (2002). *Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus*. Unpublished manuscript. http://www.statmodel.com/download/webnotes/CatMGLong.pdf.

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

Muthén, B. O., & Asparouhov, T. (2013). *BSEM measurement invariance analysis. Mplus web note: No. 17.* Unpublished manuscript. https://www.statmodel.com/examples/webnotes/webnote17.pdf.

Muthén, L. K., & Muthén, B. (1998–2017). *Mplus user's guide* (7th ed.). Los Angeles: Authors.

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., et al. (2006). *Uncertain judgements: Eliciting experts' probabilities*. West Sussex, UK: Wiley.

Park, T., & Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association, 103*, 681–686.

Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). Cambridge, UK: Cambridge University Press.

Pokropek, A., Davidov, E., & Schmidt, P. (2019). A Monte Carlo simulation study to assess the appropriateness of traditional and newer approaches to test for measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 26*, 724–744.

Pokropek, A., Schmidt, P., & Davidov, E. (2020). Choosing priors in Bayesian measurement invariance modeling: A Monte Carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal, 27*, 750–764.

R Development Core Team. (2008). *R: A language and environment for statistical computing* [Computer software manual]. www.R-project.org.

Raftery, A. E. (1995). Bayesian model selection in social research (with discussion). In P. V. Marsden (Ed.), *Sociological methodology* (Vol. 25, pp. 111–196). New York: Blackwell.

Raftery, A. E., Madigan, D., & Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association, 92*, 179–191.

Raftery, A. E., & Zheng, Y. (2003). Discussion: Performance of Bayesian model averaging. *Journal of the American Statistical Association, 98*, 931–938.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461–464.

Silvey, S. D. (1975). *Statistical inference*. Boca Raton, FL: CRC Press.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Boca Raton, FL: Chapman & Hall/CRC.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, B (Statistical Methodology), 64*, 583–639.

Stan Development Team. (2021). *Stan modeling language users guide and reference manual,version 2.26* [Computer software manual]. https://mc-stan.org.

Steel, M. F. J. (2020). Model averaging and its use in economics. *Journal of Economic Literature, 58*, 644–719.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, B (Methodological), 58*, 267–288.

Ulitzsch, E., Holtmann, J., Schultze, M., & Eid, M. (2017). Comparing multilevel and classical confirmatory factor analysis parameterizations of multirater data: A Monte Carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal, 24*, 80–103.

van de Schoot, R., Depaoli, S., King, R., Kramer, B., Martens, K., Tadesse, M., et al. (2021). Bayesian statistical modelling. *Nature Reviews Methods Primers, 1*, 1–26.

van de Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J., & Muthén, B. (2013). Facing off with Scylla and

Charybdis: A comparison of scalar, partial, and the novel possibility of approximate measurement invariance. *Frontiers in Psychology: Quantitative Psychology and Measurement, 4*, 1–15.

van de Schoot, R., Winter, S., Zondervan-Zwijnenburg, M., Ryan, O., & Depaoli, S. (2017). A systematic review of Bayesian applications in psychology: The last 25 years. *Psychological Methods, 22*, 217–239.

van Erp, S. (2020). A tutorial on Bayesian penalized regression with shrinkage priors for small sample sizes. In R. van de Schoot & M. Miočević (Eds.), *Small sample size solutions* (pp. 71–84). New York: Taylor & Francis.

Vehtari, A., Gabry, J., Yao, Y., & Gelman, A. (2019). *loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models* (R package version 2.1.0). https://CRAN.R-project.org/package=loo.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing, 27*, 1413–1432.

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved $\hat{R}$ for assessing convergence of MCMC. *Bayesian Analysis, 16*, 667–718.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin and Review, 14*, 779–804.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research, 11*, 3571–3594.

Winter, S. D., & Depaoli, S. (2019). An illustration of Bayesian approximate measurement invariance with longitudinal data and a small sample size. *International Journal of Behavioral Development, 44*, 371–382.