Chapter 4 Bayesian Model Comparison and Model Checking

As one of the main goals of SEMs is the evaluation of some simultaneous hypotheses about the interrelationships among the observed variables, latent variables, and fixed covariates, testing of various hypotheses about the model is certainly an important topic of interest.

In the field of structural equation modeling, the classical approach for hypothesis testing is to use the significance tests on the basis of p-values that are determined by some asymptotic distributions of the test statistics. In general, there may be serious problems associated with such an approach. See Lee (2007, Chapter 5) for a discussion of these problems in relation to SEMs.

$H_0: \theta = \theta_0$ vs $H_1: \theta \neq \theta_0$

$L = -2 \dfrac{\ell(\theta_0)}{\ell(\theta)}$

$\Rightarrow$ Model Comparison

The main objectives of this chapter are:

(I) to introduce various Bayesian statistics for hypothesis testing and model comparison,

(II) to provide some statistical methods for assessment of the goodness-of-fit of the posited model and for diagnostic of the model.

Organization of this chapter:

- An introduction of the Bayes factor, including its computation and application. ~ AIC & BIC
- Some other methods for model comparison.
- An illustrative example.
- Methods for model checking and goodness-of-fit.

In the Bayesian approach, testing the null hypothesis H_0 against its alternative hypothesis H_1 can be regarded as comparing two models corresponding to H_0 and H_1. For instance, we consider the SEM as follows:

$$\mathbf{y}_i = \boldsymbol{\mu} + \boldsymbol{\Lambda}\boldsymbol{\omega}_i + \boldsymbol{\epsilon}_i \tag{1}$$

$$\eta_i = b_1 d_i + \gamma_1 \xi_{i1} + \gamma_2 \xi_{i2} + \gamma_3 \xi_{i1}\xi_{i2} + \gamma_4 \xi_{i1}^2 + \gamma_5 \xi_{i2}^2 + \delta_i. \tag{2}$$

Test if the nonlinear term is necessary.

Suppose we want to test

$$H_0 : \ \gamma_3 = \gamma_4 = \gamma_5 = 0, \quad \text{vs} \quad H_1 : \ \gamma_3 \neq 0, \ \gamma_4 \neq 0, \ \gamma_5 \neq 0$$

We can define an SEM, M_0, with a measurement equation defined by (1) and a structural equation defined by

$$\eta_i = b_1 d_i + \gamma_1 \xi_{i1} + \gamma_2 \xi_{i2} + \delta_i.$$

This gives a model corresponding to H_0. The model M_1 that corresponds to the alternative hypothesis H_1 is defined by Equations (1) and (2).

Bayesian statistics for model comparison in this book:

- Bayes factor — a ratio of marginal likelihoods. It's computationally challenging, and path sampling is used for computation.
- Deviance Information Criterion (DIC) — an analog of the AIC. It aims to seek an appropriate model by compromising the goodness-of-fit and model complexity under a Bayesian framework. WinBUGS directly provides the DIC values for many complex SEMs.
- L-measure — a criterion-based method. It measures the performance of a model by a combination of how close its predictions are to the observed data and the variability of the predictions.

In this section, we introduce an important Bayesian statistic, the Bayes factor (Berger, 1985; Kass and Raftery, 1995) for model comparison. Suppose that the given data set $\mathbf{Y}$ with a sample size n has arisen under one of the two competing models M_1 and M_0 according to probability densities $p(\mathbf{Y}|M_1)$ or $p(\mathbf{Y}|M_0)$. Let $p(M_0)$ be the prior probability of M_0 and $p(M_1) = 1 - p(M_0)$, and let $p(M_k|\mathbf{Y})$ be the posterior probability for $k = 0,\ 1$. From the Bayes theorem, we have

$\int_\theta p(Y|M_k,\theta)\, p(\theta|M_k)\, d\theta$. a marginal

$$p(M_k|\mathbf{Y}) = \frac{p(\mathbf{Y}|M_k)p(M_k)}{p(\mathbf{Y}|M_1)p(M_1) + p(\mathbf{Y}|M_0)p(M_0)}, \qquad k = 0,\ 1.$$

Hence,

posterior odds Bayes factor prior odds.

$$\frac{p(M_1|\mathbf{Y})}{p(M_0|\mathbf{Y})} = \frac{p(\mathbf{Y}|M_1)}{p(\mathbf{Y}|M_0)}\frac{p(M_1)}{p(M_0)}. \tag{3}$$

The Bayes factor for comparing M_1 and M_0 is defined as

$$B_{10} = \frac{p(\mathbf{Y}|M_1)}{p(\mathbf{Y}|M_0)}. \tag{4}$$

So, posterior odds = Bayes factor $\times$ prior odds. In the special case where M_1 and M_0 are equally probable a priori so that $p(M_1) = p(M_0) = 0.5$, the Bayes factor is equal to the posterior odds in favor of M_1. We note the following aspects of the Bayes factor:

1. It may reject a null hypothesis associated with M_0, or may equally provide evidence in favor of the null hypothesis or the alternative hypothesis associated with M_1.
2. Unlike the significance test approach that is based on the likelihood ratio criterion and its asymptotic test statistic, the comparison based on the Bayes factor does not depend on the assumption that either model is 'true'. From Bayesian perspective, no model is true, or all models are true.
3. It can be seen from (4) that the same data set is used in the comparison; hence, it does not favor the alternative hypothesis (or M_1) in extremely large samples.
4. It can be applied to compare nonnested models M_0 and M_1.

The criterion (see Kass and Raftery, 1995) that is used for interpreting B_{10} and $2 \log B_{10}$ is given in Table 4.1:

Why is the threshold not symmetric ?

B_{10}	$2 \log B_{10}$	Evidence against $H_0(M_0)$
< 1	< 0	Negative (supports $H_0(M_0)$)
1 to 3	0 to 2	Not worth more than a bare mention
3 to 20	2 to 6	Positive (supports $H_1(M_1)$)
20 to 150	6 to 10	Strong
> 150	> 10	Decisive

BIC

The interpretation of evidence provided by Table 4.1 depends on the specific context. $M_1 \cap M_0 = \varnothing$

A. For two nonnested models, say M_1 and M_0, we select M_0 if $2 \log B_{10}$ is negative. If $2 \log B_{10}$ is in $(0, 2)$, we may interpret that M_1 is slightly better than M_0 and hence it may be better to select M_1. The choice of M_1 is more definite if $2 \log B_{10}$ is larger than 6.

B. For two nested models, say M_0 is nested in M_1, $2 \log B_{10}$ is most likely larger than zero. If $2 \log B_{10}$ is larger than 6. The above criterion will suggest a decisive conclusion to select M_1. However, if $2 \log B_{10}$ is in $(0, 2)$, then the difference between M_0 and M_1 is not significant. Under this situation, great caution should be taken in drawing conclusions. According to the 'parsimony' guideline in practical applications, it may be desirable to select the simpler model M_0.

C. For marginal cases, it is always helpful to conduct other analysis, for example residual analysis, to cross-validate the results.

The choice of prior inputs is an important issue when applying Bayes factor to the comparison of M_0 and M_1. As pointed out by Kass and Raftery (1995), using a prior with a very large spread on the parameters under M_1 as to make it "noninformative" will force the Bayes factor to favor the competing model M_0. This is known as the "Bartlett's paradox".

Usually, we take the following considerations on the prior inputs:

1. priors on parameters under the model comparison are generally taken to be proper and not having a too big spread.
2. The conjugate families with reasonable spreads are appropriate choices.
3. To cope with situations without prior information, a simple method suggested by Kass and Raftery (1995) is to set aside part of the data to use as a training sample coped with a noninformative prior distribution to produce an informative prior inputs. The Bayes factor is then computed from the remainder of the data.
4. A sensitivity analysis should be conducted to check the effects of prior inputs to model comparison results.

Let $\boldsymbol{\theta}_k$ be the random parameter vector associated with M_k. From

$$p(\boldsymbol{\theta}_k, \mathbf{Y}|M_k) = p(\mathbf{Y}|\boldsymbol{\theta}_k, M_k)p(\boldsymbol{\theta}_k|M_k),$$

we have

Computational Challenge.

$$p(\mathbf{Y}|M_k) = \int p(\mathbf{Y}|\boldsymbol{\theta}_k, M_k)p(\boldsymbol{\theta}_k|M_k)d\boldsymbol{\theta}_k, \tag{5}$$

where $p(\boldsymbol{\theta}_k|M_k)$ is the prior density of $\boldsymbol{\theta}_k$ and $p(\mathbf{Y}|\boldsymbol{\theta}_k, M_k)$ is the probability density of $\mathbf{Y}$ given $\boldsymbol{\theta}_k$. The dimension of this integral is equal to the dimension of $\boldsymbol{\theta}_k$.

Obtaining B_{10} analytically is difficult. Various analytic and numerical approximations have been proposed in the literature. See Chib (1995), DiCiccio *et al.* (1997), Gelman and Meng (1998), and Chib and Jeliazkov (2001). We will discuss the application of path sampling to compute the Bayes factor for model comparison.

Let

- $\mathbf{Y}$ — the matrix of observed data,
- $\boldsymbol{\Omega}$ — the matrix of latent variables.

For SEMs with latent variables, direct application of path sampling to compute the Bayes factor is difficult. We utilize the idea of data augmentation to solve the problem. From the equality

$$p(\boldsymbol{\Omega}, \boldsymbol{\theta}|\mathbf{Y}) = p(\mathbf{Y}, \boldsymbol{\Omega}, \boldsymbol{\theta})/p(\mathbf{Y}),$$

where $p(\mathbf{Y})$ can be treated as the normalizing constant of $p(\boldsymbol{\Omega}, \boldsymbol{\theta}|\mathbf{Y})$, with the complete-data probability density $p(\mathbf{Y}, \boldsymbol{\Omega}, \boldsymbol{\theta})$ taking as the unnormalized density. Now, consider the following class of densities which are denoted by a continuous parameter t in $[0, 1]$:

$$p(\boldsymbol{\Omega}, \boldsymbol{\theta}|\mathbf{Y}, t) = \frac{1}{z(t)} p(\mathbf{Y}, \boldsymbol{\Omega}, \boldsymbol{\theta}|t), \quad \frac{p(Y, \Omega, \theta|t)}{P(Y|t)}. \tag{6}$$

where

$$z(t) = p(\mathbf{Y}|t) = \int p(\mathbf{Y}, \boldsymbol{\Omega}, \boldsymbol{\theta}|t) d\boldsymbol{\Omega} d\boldsymbol{\theta} = \int p(\mathbf{Y}, \boldsymbol{\Omega}, |\boldsymbol{\theta}, t) p(\boldsymbol{\theta}) d\boldsymbol{\Omega} d\boldsymbol{\theta}. \tag{7}$$

In computing the Bayes factor, we construct a path using the parameter t in $[0,1]$ to link two competing models M_1 and M_0 together, so that

Why are both side equal?

$$z(1) = p(\mathbf{Y}|1) = p(\mathbf{Y}|M_1), \quad z(0) = p(\mathbf{Y}|0) = p(\mathbf{Y}|M_0),$$

You should find such link function satisfying such relationship.

and $B_{10} = z(1)/z(0)$. Taking logarithm and then differentiating (7) with respect to t, and assuming the legitimacy of interchange of integration with differentiation, we have

$$\begin{aligned}
\frac{d \log z(t)}{dt} &= \int \frac{1}{z(t)} \frac{d}{dt} p(\mathbf{Y}, \Omega, \theta|t) d\Omega d\theta \\
&= \int \frac{d}{dt} \log p(\mathbf{Y}, \Omega, \theta|t) \cdot p(\Omega, \theta|\mathbf{Y}, t) d\Omega d\theta \qquad (8) \\
&= E_{\Omega,\theta} \left[\frac{d}{dt} \log p(\mathbf{Y}, \Omega, \theta|t) \right],
\end{aligned}$$

where $E_{\Omega,\theta}$ denotes the expectation with respect to $p(\Omega, \theta|\mathbf{Y}, t)$.

$U(Y, \Omega, \theta, t)$

$$\int_0^1 \frac{d \log z(t)}{dt} dt = \log z(1) - \log z(0) = \log \frac{z(1)}{z(0)} = \int E_{\Omega,\theta} \left[\frac{d}{dt} \log p(Y, \Omega, \theta|t) \right] dt$$

Let

$$U(\mathbf{Y}, \boldsymbol{\Omega}, \boldsymbol{\theta}, t) = \frac{d}{dt} \log p(\mathbf{Y}, \boldsymbol{\Omega}, \boldsymbol{\theta}|t) = \frac{d}{dt} \log p(\mathbf{Y}, \boldsymbol{\Omega}|\boldsymbol{\theta}, t), \tag{9}$$

which does not involve the prior density $p(\boldsymbol{\theta})$, we have

$$\log B_{10} = \log \frac{z(1)}{z(0)} = \int_0^1 E_{\Omega,\theta}[U(\mathbf{Y}, \boldsymbol{\Omega}, \boldsymbol{\theta}, t)]dt.$$

The method given in Ogata (1989) is used to numerically evaluate the integral over t. Specifically, we first order the unique values of fixed grids $\{t_{(s)}\}_{s=1}^{S}$ between [0,1] such that $0 = t_{(0)} < t_{(1)} < \cdots < t_{(S)} < t_{(S+1)} = 1$, and estimate $\log B_{10}$ by

$$\widehat{\log B_{10}} = \frac{1}{2} \sum_{s=0}^{S} (t_{(s+1)} - t_{(s)})(\bar{U}_{(s+1)} + \bar{U}_{(s)}), \tag{10}$$

where

$$\bar{U}_{(s)} = J^{-1} \sum_{j=1}^{J} U(\mathbf{Y}, \boldsymbol{\Omega}^{(j)}, \boldsymbol{\theta}^{(j)}, t_{(s)}), \tag{11}$$

in which $\{(\boldsymbol{\Omega}^{(j)}, \boldsymbol{\theta}^{(j)}),\ j = 1, \cdots, J\}$ are drawn from $p(\boldsymbol{\Omega}, \boldsymbol{\theta}|\mathbf{Y}, t_{(s)})$.

Where does it come from?

Steps in implementing the path sampling procedure:

1. Define a link model M_t to link M_0 and M_1, such that when $t = 0$, $M_t = M_0$; and when $t = 1$, $M_t = M_1$.
2. Obtain $U(\mathbf{Y}, \boldsymbol{\Omega}, \boldsymbol{\theta}, t)$ by differentiating the logarithm of the complete-data likelihood function under M_t with respect to t
3. Estimate $\log B_{10}$ via (9) and (10). For most SEMs, $S = 20$ and $J = 1,000$ provide results that are accurate enough for many practical applications. Experiences indicate that $S = 10$ is also acceptable for simple SEMs.

Nice features of the path sampling:

1. Its implementation is simple. The main programming task is simulating observations from $p(\boldsymbol{\Omega}, \boldsymbol{\theta}|\mathbf{Y}, t_{(s)})$.
2. It's easy to construct a continuous path to link most competing models. Thus, the method can be applied to the comparison of a wide variety of models.
3. Its computation does not directly include the prior density.
4. Its logarithm scale is generally more stable than the ratio scale.
5. It is a generalization of bridge sampling, so it has potential to produce more accurate results than other sampling methods.

In applying path sampling, it is required to find a path t in $[0, 1]$ to link the competing models M_0 and M_1. For most cases, finding such a path is fairly straightforward. However, for some complex situations that involve very different M_1 and M_0, it is difficult to find a path that directly links the competing models.

When they come from the same family.

A possible solution is using appropriate auxiliary models, M_a, $M_b, \cdots$, to link M_1 and M_0. e.g., suppose that M_a and M_b are appropriate auxiliary models such that M_a can be linked with M_1 and M_b; and M_b can be linked with M_0. Then

$$\frac{p(\mathbf{Y}|M_1)}{p(\mathbf{Y}|M_0)} = \frac{p(\mathbf{Y}|M_1)/p(\mathbf{Y}|M_a)}{p(\mathbf{Y}|M_0)/p(\mathbf{Y}|M_a)}, \quad \text{and} \quad \frac{p(\mathbf{Y}|M_0)}{p(\mathbf{Y}|M_a)} = \frac{p(\mathbf{Y}|M_0)/p(\mathbf{Y}|M_b)}{p(\mathbf{Y}|M_a)/p(\mathbf{Y}|M_b)}.$$

Hence, $\log B_{10} = \log B_{1a} + \log B_{ab} - \log B_{0b}$. Each logarithm Bayes factor can be computed through path sampling.

The objectives of this simulation study are

1. to reveal the performance of path sampling in computing Bayes factor,
2. to evaluate the sensitivity of the results to prior inputs.

Consider a nonlinear SEM with fixed covariates defined by (2.21) and (2.22). The model includes:

$y = Ac + \Lambda w + \varepsilon.$

- 8 observed variables;
- 2 fixed covariates $\{c_{i1}, c_{i2}\}$ in the measurement equation, where c_{i1} is sampled from a multinomial distribution which takes values 1.0, 2.0, and 3.0 with probabilities $\Phi^*(-0.5)$, $\Phi^*(0.5) - \Phi^*(-0.5)$, and $1.0 - \Phi^*(0.5)$, and c_{i2} is sampled from $N[0, 1]$;
- 3 latent variables $\{\eta_i, \xi_{i1}, \xi_{i2}\}$;
- 1 fixed covariate d_i in the structural equation, where d_i is sampled from a Bernoulli distribution that takes 1.0 with probability 0.7 and 0.0 with probability 0.3.

The true population values in matrices $\mathbf{A}$, $\mathbf{\Lambda}$, and $\mathbf{\Psi}_\epsilon$ are:

$$\mathbf{A}^T = \begin{bmatrix} 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 \\ 0.7 & 0.7 & 0.7 & 0.7 & 0.7 & 0.7 & 0.7 & 0.7 \end{bmatrix},$$

$$\mathbf{\Lambda}^T = \begin{bmatrix} 1 & 1.5 & 1.5 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1.5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1.5 & 1.5 \end{bmatrix}, \quad \mathbf{\Psi}_\epsilon = \mathbf{I}_8,$$

where 1's and 0's in $\mathbf{\Lambda}$ are fixed to identify the model, and $\mathbf{I}_8$ is an 8×8 identity matrix. $\phi_{11} = \phi_{22} = 1.0$, and $\phi_{21} = 0.15$. The structural equation is:

$$\eta_i = 1.0 d_i + 0.5\xi_{i1} + 0.5\xi_{i2} + 1.0\xi_{i2}^2 + \delta_i,$$

where $\psi_\delta = 1.0$.

Random samples $\{\mathbf{y}_i,\ i = 1, \cdots, n\}$ with $n = 300$ were generated for the simulation study. A total of 100 replications were taken.

We are interested in comparing models with different structural equations. Hence, models with the same measurement equation and the following structural equations are considered in the model comparison:

$$
\begin{aligned}
M_0: &\quad \eta_i = bd_i + \gamma_1\xi_{i1} + \gamma_2\xi_{i2} + \gamma_{22}\xi_{i2}^2 + \delta_i, \\
M_1: &\quad \eta_i = bd_i + \gamma_1\xi_{i1} + \gamma_2\xi_{i2} + \delta_i, \\
M_2: &\quad \eta_i = bd_i + \gamma_1\xi_{i1} + \gamma_2\xi_{i2} + \gamma_{12}\xi_{i1}\xi_{i2} + \delta_i, \\
M_3: &\quad \eta_i = bd_i + \gamma_1\xi_{i1} + \gamma_2\xi_{i2} + \gamma_{11}\xi_{i1}^2 + \delta_i, \\
M_4: &\quad \eta_i = bd_i + \gamma_1\xi_{i1} + \gamma_2\xi_{i2} + \gamma_{12}\xi_{i1}\xi_{i2} + \gamma_{11}\xi_{i1}^2 + \delta_i, \\
M_5: &\quad \eta_i = \gamma_1\xi_{i1} + \gamma_2\xi_{i2} + \gamma_{22}\xi_{i2}^2 + \delta_i, \\
M_6: &\quad \eta_i = bd_i + \gamma_1\xi_{i1} + \gamma_2\xi_{i2} + \gamma_{12}\xi_{i1}\xi_{i2} + \gamma_{11}\xi_{i1}^2 + \gamma_{22}\xi_{i2}^2 + \delta_i.
\end{aligned}
$$

$M_1 \subset M_0$

$M_6 \supset M_0$

Here, M_0 is the true model, M_1 is a linear model, M_2, M_3, and M_4 are nonnested in M_0, M_5 is nested in M_0, and M_0 is nested in the most general model M_6.

To provide an illustration for applying path sampling procedure to model comparison, the implementation in estimating $\log B_{02}$ for comparing M_0 and M_2 is given here.

Let $\boldsymbol{\theta} = (\tilde{\boldsymbol{\theta}}, \boldsymbol{\Gamma}_\omega)$, and $\boldsymbol{\theta}_t = (\tilde{\boldsymbol{\theta}}, \boldsymbol{\Gamma}_{t\omega})$, where $\boldsymbol{\Gamma}_\omega = (b, \gamma_1, \gamma_2, \gamma_{12}, \gamma_{22})$, $\boldsymbol{\Gamma}_{t\omega} = (b, \gamma_1, \gamma_2, (1-t)\gamma_{12}, t\gamma_{22})$, and $\tilde{\boldsymbol{\theta}}$ includes other unknown common parameters in M_0 and M_2. The procedure consists of the following steps:

Step 1: Select a link model M_t to link M_0 and M_2. Here, M_t is defined with the same measurement model as in M_0 and M_2, but with the following structural equation:

$$M_t: \quad \eta_i = bd_i + \gamma_1 \xi_{i1} + \gamma_2 \xi_{i2} + (1-t)\gamma_{12}\xi_{i1}\xi_{i2} + t\gamma_{22}\xi_{i2}^2 + \delta_i.$$

When $t = 1, M_t = M_0$; when $t = 0, M_t = M_2$.

Step 2: At the fixed grid $t = t_{(s)}$, generate observations $\{(\boldsymbol{\Omega}^{(j)}, \boldsymbol{\theta}^{(j)}),\ j = 1, \cdots, J\}$ from $p(\boldsymbol{\Omega}, \boldsymbol{\theta}|\mathbf{Y}, t_{(s)})$ by using some MCMC methods, such as the Gibbs sampler and the MH algorithm, as in the Bayesian estimation.

Step 3: Calculate $U(\mathbf{Y}, \boldsymbol{\Omega}^{(j)}, \boldsymbol{\theta}^{(j)}, t_{(s)})$ by substituting $\{(\boldsymbol{\Omega}^{(j)}, \boldsymbol{\theta}^{(j)}),\ j = 1, \cdots, J\}$ to the following equation:

$$U(\mathbf{Y}, \boldsymbol{\Omega}, \boldsymbol{\theta}, t_{(s)}) = d \log p(\mathbf{Y}, \boldsymbol{\Omega}, \boldsymbol{\theta}, t)/dt\big|_{t=t_{(s)}}$$

$$= -\sum_{i=1}^{n}\{\eta_i - bd_i - \gamma_1\xi_{i1} - \gamma_2\xi_{i2} - (1 - t_{(s)})\gamma_{12}\xi_{i1}\xi_{i2} - t_{(s)}\gamma_{22}\xi_{i2}^2\}$$

$$(\gamma_{12}\xi_{i1}\xi_{i2} - \gamma_{22}\xi_{i2}^2)/\psi_\delta.$$

Step 4: Calculate $\bar{U}_{(s)}$; see (11). MC Expectation of U

Step 5: Repeat *Step 2* to *Step 4* until all $\bar{U}_{(s)},\ s = 0, \cdots, S+1$ are calculated. Then, $\widehat{\log B_{02}}$ is estimated via (10).

In the sensitivity analysis, the prior inputs are perturbed as follows. Under prior inputs $\alpha_{0\epsilon k} = \alpha_{0\delta k} = 8$, $\beta_{0\epsilon k} = \beta_{0\delta k} = 10$, and $\rho_0 = 20$, we consider the following three types of prior inputs for $\mathbf{A}_{0k}, \mathbf{\Lambda}_{0k}, \mathbf{\Lambda}_{0\omega k}$, and $\mathbf{R}_0^{-1}$:

(I) $\mathbf{A}_{0k}$, $\mathbf{\Lambda}_{0k}$, and $\mathbf{\Lambda}_{0\omega k}$ are selected to be the true parameter matrices, and $\mathbf{R}_0^{-1} = (\rho_0 - q_2 - 1)\mathbf{\Phi}_0$, where elements in $\mathbf{\Phi}_0$ are the true parameters values.

(II) The hyperparameters specified in (I) are equal to half of those given in (I).

(III) The hyperparameters specified in (I) are equal to twice of those given in (I).

Moreover, under Type (I) prior inputs as given above, we consider the following prior inputs for $\alpha_{0\epsilon k}$, $\alpha_{0\delta k}$, $\beta_{0\epsilon k}$, $\beta_{0\delta k}$, and ρ_0:

(IV) $\alpha_{0\epsilon k} = \alpha_{0\delta k} = 3$, $\beta_{0\epsilon k} = \beta_{0\delta k} = 5$, and $\rho_0 = 12$.

(V) $\alpha_{0\epsilon k} = \alpha_{0\delta k} = 12$, $\beta_{0\epsilon k} = \beta_{0\delta k} = 15$, and $\rho_0 = 30$.

For every case, $\mathbf{\Sigma}_0$, $\mathbf{H}_{0yk}$, and $\mathbf{H}_{0\omega k}$ were taken as the identity matrices.

We took 20 grids in $[0, 1]$, and collected $J = 1,000$ iterations after discarding 500 burn-in iterations at each grid in the computation of the logarithm Bayes factor via path sampling. Estimates of $\log B_{0k}$, $k = 1, \cdots, 6$ under the different prior inputs were computed. The mean and standard deviation of $\widehat{\log B_{0k}}$ were also computed on the basis of 100 replications. Results corresponding to $\widehat{\log B_{0k}}$, $k = 1, \cdots, 5$ and $\widehat{\log B_{60}}$ are reported in the following Table 4.2.

the B_{0k} means Bayes ratio of the models that are presented in the previous pages.

	Mean (Std)				
	prior I	prior II	prior III	prior IV	prior V
$\log B_{01}$	106.28 (25.06)	107.58 (25.15)	102.96 (24.81)	103.87 (22.71)	104.61 (23.92)
$\log B_{02}$	102.16 (24.91)	103.45 (25.02)	99.17 (24.54)	99.98 (22.67)	100.49 (23.47)
$\log B_{03}$	109.51 (25.63)	111.23 (25.74)	105.96 (25.19)	107.20 (23.81)	108.24 (24.59)
$\log B_{04}$	105.23 (25.31)	106.61 (25.47)	101.83 (24.90)	103.16 (23.78)	103.69 (24.12)
$\log B_{05}$	17.50 (5.44)	18.02 (5.56)	16.65 (5.21)	18.02 (5.34)	17.85 (5.30)
$\log B_{60}$	0.71 (0.54)	0.71 (0.51)	0.69 (0.55)	0.78 (0.67)	0.75 (0.65)

the results of distinct priors are quite robust.

convention:

larger model is put on the upper place while the smaller one is put on the lower place.

Moreover, for each $k = 1, \cdots, 6$, we evaluate

$$D(\mathrm{I} - \mathrm{II}) = \max\{|\widehat{\log B_{0k}}(\mathrm{I}) - \widehat{\log B_{0k}}(\mathrm{II})|\}$$

disturbance difference between different prior input.

as well as $D(\mathrm{I}-\mathrm{III})$ and $D(\mathrm{IV}-\mathrm{V})$ similarly, where $\widehat{\log B_{0k}}(\mathrm{I})$ is the estimate of $\log B_{0k}$ under prior (I) and so on, and 'max' is the maximum taken over the 100 replications. The results are presented in the Table 4.3:

	$\log B_{01}$	$\log B_{02}$	$\log B_{03}$	$\log B_{04}$	$\log B_{05}$	$\log B_{60}$
D(I-II)	6.55	5.47	8.22	5.24	2.18	0.27
D(I-III)	7.84	9.33	10.23	10.17	3.07	0.31
D(IV-V)	14.03	17.86	13.65	4.87	1.91	0.25

M_0 : $\eta_i = bd_i + \gamma_1\xi_{i1} + \gamma_2\xi_{i2} + \gamma_{22}\xi_{i2}^2 + \delta_i$, $M_1 \subset M_0$

M_1 : $\eta_i = bd_i + \gamma_1\xi_{i1} + \gamma_2\xi_{i2} + \delta_i$,

M_2 : $\eta_i = bd_i + \gamma_1\xi_{i1} + \gamma_2\xi_{i2} + \gamma_{12}\xi_{i1}\xi_{i2} + \delta_i$,

M_3 : $\eta_i = bd_i + \gamma_1\xi_{i1} + \gamma_2\xi_{i2} + \gamma_{11}\xi_{i1}^2 + \delta_i$,

M_4 : $\eta_i = bd_i + \gamma_1\xi_{i1} + \gamma_2\xi_{i2} + \gamma_{12}\xi_{i1}\xi_{i2} + \gamma_{11}\xi_{i1}^2 + \delta_i$,

M_5 : $\eta_i = \gamma_1\xi_{i1} + \gamma_2\xi_{i2} + \gamma_{22}\xi_{i2}^2 + \delta_i$,

M_6 : $\eta_i = bd_i + \gamma_1\xi_{i1} + \gamma_2\xi_{i2} + \gamma_{12}\xi_{i1}\xi_{i2} + \gamma_{11}\xi_{i1}^2 + \gamma_{22}\xi_{i2}^2 + \delta_i$.

$M_6 \supset M_0$

Interpretations and findings from Tables 4.2 and 4.3:

1. M_0 is much better than the linear model M_1, the nonnested models M_2, M_3, and M_4, and the nested model M_5. Thus, the correct model is selected.
2. In comparison with the encompassing model M_6, we found that out of 100 replications under prior (I), 75 of $\widehat{\log B_{60}}$ were in the interval $(0.0, 1.0)$, 23 of them were in $(1.0, 2.0)$, and only 2 of them were in $(2.0, 3.0)$. Since M_0 is simpler than M_6, it should be selected if $\widehat{\log B_{60}}$ is in $(0.0, 2.0)$. Thus, the true model is selected in 98 out of the 100 replications.
3. From Table 4.2, the means and standard deviations of $\widehat{\log B_{0k}}$ obtained under different prior inputs are close to each other, indicating that $\widehat{\log B_{0k}}$ is not very sensitive to the given prior inputs and sample size.
4. From Table 4.3, even for the worst situation with the maximum absolute deviation, the estimated logarithm of Bayes factors under different prior inputs give the same conclusion.

WinBUGS does not have an option to compute the Bayes factor. However, as mentioned in Section 3.5, WinBUGS can be run in batch mode using scripts, and the R2WinBUGS (Sturtz, Ligges and Gelman, 2005) package makes use of this feature and provides tools to call WinBUGS directly after data manipulation in R. Hence, Bayes factor can be computed via WinBUGS and R2WinBUGS. The WinBUGS code for comparing M_0 and M_2 is given in Appendix 4.1.

This WinBUGS code must be stored in a separate file (say 'model.txt') within an appropriate directory (say C:\Bayes Factor\) when computing the logarithm Bayes factor via path sampling, and $\bar{U}_{(s)}$ at each grid is computed from WinBUGS via the bugs($\cdot$) function in R2WinBUGS; the logarithm Bayes factor is then computed using the $\bar{U}_{(s)}$, $s = 0, \cdots, S+1$. The related R code for computing the $\log B_{02}$, including data generation, is given in Appendix 4.2.

An approximation of $2 \log B_{10}$ that does not depend on the prior density is the following Schwarz criterion S^* (Schwarz, 1978):

What is Schwarz criterion S^*.

What is this?

$$2 \log B_{10} \cong 2S^* = 2\{\log p(\mathbf{Y}|\tilde{\boldsymbol{\theta}}_1, M_1) - \log p(\mathbf{Y}|\tilde{\boldsymbol{\theta}}_0, M_0)\} - (d_1 - d_0) \log n, \tag{12}$$

where $\tilde{\boldsymbol{\theta}}_1$ and $\tilde{\boldsymbol{\theta}}_0$ are the maximum likelihood (ML) estimates of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_0$ under M_1 and M_0, respectively; d_1 and d_0 are the dimensions of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_0$, and n is the sample size. Minus $2S^*$ is the following well-known Bayesian Information Criterion (BIC) for comparing M_1 and M_0:

$$\text{BIC}_{10} = -2S^* \cong -2 \log B_{10} = 2 \log B_{01}. \tag{13}$$

The interpretation of BIC_{10} can be based on Table 4.1.

What is the d_k ? It also appears in AIC and DIC!

Alternatively, for each M_k, $k = 0,\ 1$, we can define

$$\text{BIC}_k = -2 \log p(\mathbf{Y}|\tilde{\theta}_k, M_k) + d_k \log n. \tag{14}$$

Hence $2 \log B_{10} \cong \text{BIC}_0 - \text{BIC}_1$. Based on Table 4.1, the model M_k with the smaller BIC_k value is selected.

As n tends to infinity, it has been shown (Schwarz, 1978) that

$$\frac{S^* - \log B_{10}}{\log B_{10}} \to 0,$$

thus S^* may be viewed as an approximation to $\log B_{10}$. This approximation is of order $O(1)$, thus S^* does not give the exact $\log B_{10}$ even for large samples. However, as pointed out by Kass and Raftery (1995), it can be used for scientific reporting as long as the number of degrees of freedom $(d_1 - d_0)$ involved in the comparison is small relative to the sample size n.

Remarks:

1. BIC is simple and can be applied even when the priors $p(\boldsymbol{\theta}_k|M_k)$ $(k = 1, 0)$ are hard to specify precisely.
2. The ML estimates of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_0$ are involved in the computation of BIC. In practice, since the Bayesian estimates and the ML estimates are close to each other, they can be used to compute the BIC. The order of approximation is not changed.
3. For complex SEMs, the observed-data log-likelihoods are usually intractable multiple integrals. Under such situations, path sampling can be applied to evaluate $p(\mathbf{Y}|\tilde{\boldsymbol{\theta}}_k, M_k)$, by fixing $\boldsymbol{\theta}_k$ at its estimate $\tilde{\boldsymbol{\theta}}_k$ rather than treating it as random (see Song and Lee, 2006).

The Akaike Information Criterion (AIC; Akaike, 1973) associated with a competing model M_k is given by

unknown constant

2 d_k Log of BIC$_k$.

$$\text{AIC}_k = -2 \log p(\mathbf{Y}|\tilde{\boldsymbol{\theta}}_k, M_k) + 2d_k, \tag{15}$$

which does not involve the sample size n. The interpretation of AIC_k is similar to BIC_k. Hence, M_k is selected if its AIC_k is smaller.

Comparing AIC and BIC, we see that BIC tends to favor simpler models.

Another model comparison statistic that compromises the goodness-of-fit and model complexity is the Deviance Information Criterion (DIC). This statistic is intended as a generalization of AIC. Under a competing model M_k with $\boldsymbol{\theta}_k$, the DIC is defined as

$$\text{DIC}_k = \overline{D(\boldsymbol{\theta}_k)} + d_k, \tag{16}$$

where $\overline{D(\boldsymbol{\theta}_k)}$ measures the goodness-of-fit of the model, and is defined as

$$\overline{D(\boldsymbol{\theta}_k)} = E_{\boldsymbol{\theta}_k}\{-2\log p(\mathbf{Y}|\boldsymbol{\theta}_k, M_k)|\mathbf{Y}\}. \tag{17}$$

Here, d_k is the effective number of parameters in M_k, and is defined as

$$d_k = E_{\boldsymbol{\theta}_k}\{-2\log p(\mathbf{Y}|\boldsymbol{\theta}_k, M_k)|\mathbf{Y}\} + 2\log p(\mathbf{Y}|\tilde{\boldsymbol{\theta}}_k), \tag{18}$$

in which $\tilde{\boldsymbol{\theta}}_k$ is the Bayesian estimate of $\boldsymbol{\theta}_k$.

Let $\{\boldsymbol{\theta}_k^{(j)},\ j = 1, \cdots, J\}$ be a sample of observations simulated from the posterior distribution. The expectations in (17) and (18) can be estimated as follows:

empirical expectation → mathematical expectation.

$$E_{\boldsymbol{\theta}_k}\{-2\log p(\mathbf{Y}|\boldsymbol{\theta}_k, M_k)|\mathbf{Y}\} = -\frac{2}{J}\sum_{j=1}^{J}\log p(\mathbf{Y}|\boldsymbol{\theta}_k^{(j)}, M_k). \quad (19)$$

In practical applications, the model with the smaller DIC value is selected.

The computational burden of DIC is on simulating $\{\boldsymbol{\theta}_k^{(j)},\ j = 1, \cdots, J\}$ from the posterior distribution; and thus is lighter than that of the Bayes factor.

WinBUGS produces the DIC value for model comparison. In applying DIC, we should note the following:

(I) There are circumstances, such as mixture models, in which WinBUGS does not give the DIC values.

log Bio ∈ (0,1)

(II) If the difference in DIC is small, say less than 5, and the models make very different inferences, then just reporting the model with the lowest DIC could be misleading.

You have to use other Criterion to test your models, or just regard the two models are the same and choose the simpler one.

(III) DIC can be applied to nonnested models. Moreover, similar to the Bayes factor, BIC, and AIC, DIC gives clear conclusion to support the null hypothesis or the alternative hypothesis.

The L_ν-measure is developed from the predictive distribution of the data with a sum of two components. One component involves the means of the posterior predictive distribution, and the other is related to the variances. Hence, it measures the performance of a model by a combination of how close its predictions are to the observed data and the variability of the predictions.

Let

- $\mathbf{Y}$ — the observed data,
- $p(\mathbf{Y}, \boldsymbol{\theta})$ — the joint density that corresponds to a model M with a parameter vector $\boldsymbol{\theta}$.
- $\mathbf{Y}^{\text{rep}} = (\mathbf{y}_1^{\text{rep}}, \cdots, \mathbf{y}_n^{\text{rep}})$ — future values of $\mathbf{Y}$, which have the same sampling density as $p(\mathbf{Y}|\boldsymbol{\theta})$.

The basic idea is that good models should give predictions close to what have been observed. Several criteria, such as the Euclidean distance between $\mathbf{Y}$ and $\mathbf{Y}^{\text{rep}}$, can be considered. In this book, we first consider the following statistic: For some $\delta > 0$, let

$$L_1(\mathbf{Y}, \mathbf{B}, \delta) = E[\text{tr}(\mathbf{Y}^{\text{rep}} - \mathbf{B})^T(\mathbf{Y}^{\text{rep}} - \mathbf{B})] + \delta \text{ tr}(\mathbf{Y} - \mathbf{B})^T(\mathbf{Y} - \mathbf{B}), \quad (20)$$

$E[\text{tr}(Y^{rep} - B)^T(Y^{rep} - B) | Y]$.

where the expectation is taken with respect to the posterior predictive distribution of $[\mathbf{Y}^{\text{rep}}|\mathbf{Y}]$. Note that this statistic reduces to the Euclidean distance by setting $\mathbf{B} = \mathbf{Y}$. By setting $\mathbf{B}$ as the minimizer of (20), and substituting it to (20), it can be shown that (Ibrahim, Chen and Sinha, 2001)

$$L_\nu(\mathbf{Y}) = \sum_{i=1}^{n} \text{tr}\{\text{Cov}(\mathbf{y}_i^{\text{rep}}|\mathbf{Y})\} + \nu \sum_{i=1}^{n} \text{tr}[\{E(\mathbf{y}_i^{\text{rep}}|\mathbf{Y}) - \mathbf{y}_i\}\{E(\mathbf{y}_i^{\text{rep}}|\mathbf{Y}) - \mathbf{y}_i\}^T], \quad (21)$$

Why $L_\nu(Y)$ is defined like this, a weird formula?

where $\nu = \delta/(\delta + 1)$. This statistic is called the L_ν-measure.

Note that this L_ν-measure is a sum of two components. The first component relates to the variability of the predictions, and the second component measures how close its predictions to the observed data. Clearly, a small value of the L_ν-measure indicates that the corresponding model gives a prediction close to the observed value, and the variability of the prediction is also low. Hence, the model with the smallest L_ν-measure is selected from a collection of competing models.

Obviously, $0 \leq \nu \leq 1$, where $\nu = 0$ if $\delta = 0$, and ν tends to one as δ tends to infinity. This quantity can be interpreted as a weight term in the second component of $L_\nu(\mathbf{Y})$. Using $\nu = 1$ gives equal weight to the squared bias and the variance component. However, allowing ν to vary provides more flexibility in the trade-off between bias and variance. In the context of a linear model, Ibrahim, Chen and Sinha (2001) provided some theoretical results and argued that $\nu = 0.5$ is a desirable and justifiable choice for model selection.

bias term is less important!

In applying the L_ν-measure for model assessment and model selection for SEMs, we have to evaluate $\text{Cov}(\mathbf{y}_i^{\text{rep}}|\mathbf{Y})$ and $E(\mathbf{y}_i^{\text{rep}}|\mathbf{Y})$, which involve intractable multiple integrals. Based on the identities:

$$E(\mathbf{y}_i^{\text{rep}}|\mathbf{Y}) = E\{E(\mathbf{y}_i^{\text{rep}}|\boldsymbol{\Omega}, \boldsymbol{\theta})|\mathbf{Y}\}$$
$$E\{\mathbf{y}_i^{\text{rep}}(\mathbf{y}_i^{\text{rep}})^T|\mathbf{Y}\} = E[E\{\mathbf{y}_i^{\text{rep}}(\mathbf{y}_i^{\text{rep}})^T|\boldsymbol{\Omega}, \boldsymbol{\theta}\}|\mathbf{Y}],$$

the consistent estimates of $E(\mathbf{y}_i^{\text{rep}}|\mathbf{Y})$ and $\text{Cov}(\mathbf{y}_i^{\text{rep}}|\mathbf{Y})$ can be obtained from the MCMC sample simulated from the full conditional distributions via the Gibbs sampler and/or the MH algorithm.

In this section, we present an example to illustrate the application of the L_ν-measure. The model is defined by

$$\mathbf{y}_i = \boldsymbol{\mu} + \boldsymbol{\Lambda}\boldsymbol{\omega}_i + \boldsymbol{\epsilon}_i, \quad \text{and} \tag{22}$$

$$\boldsymbol{\eta}_i = \boldsymbol{\Pi}\boldsymbol{\eta}_i + \boldsymbol{\Gamma}\mathbf{F}(\boldsymbol{\xi}_i) + \boldsymbol{\delta}_i, \tag{23}$$

where the definitions of $\boldsymbol{\mu}$, $\boldsymbol{\Lambda}$, $\boldsymbol{\omega}_i$, $\cdots$ are the same as before.

plugin.

To compute the L_ν-measure, let $\boldsymbol{\Lambda}_\eta$ and $\boldsymbol{\Lambda}_\xi$ be the submatrix of $\boldsymbol{\Lambda}$ corresponding to $\boldsymbol{\eta}_i$ and $\boldsymbol{\xi}_i$, respectively, it follows that

$$\mathbf{y}_i = \boldsymbol{\mu} + \boldsymbol{\Lambda}_\eta \boldsymbol{\Pi}_0^{-1}\{\boldsymbol{\Gamma}\mathbf{F}(\boldsymbol{\xi}_i) + \boldsymbol{\delta}_i\} + \boldsymbol{\Lambda}_\xi \boldsymbol{\xi}_i + \boldsymbol{\epsilon}_i, \tag{24}$$

where $\boldsymbol{\Pi}_0 = \mathbf{I} - \boldsymbol{\Pi}$. As $\mathbf{Y}^{\text{rep}} = (\mathbf{y}_1^{\text{rep}}, \cdots, \mathbf{y}_n^{\text{rep}})$ has the same density as $p(\mathbf{Y}|\boldsymbol{\theta})$, we have

$$E(\mathbf{y}_i^{\text{rep}}|\boldsymbol{\Omega}, \boldsymbol{\theta}) = \boldsymbol{\mu} + \boldsymbol{\Lambda}_\eta \boldsymbol{\Pi}_0^{-1}\boldsymbol{\Gamma}\mathbf{F}(\boldsymbol{\xi}_i) + \boldsymbol{\Lambda}_\xi \boldsymbol{\xi}_i, \tag{25}$$

$$\text{Cov}(\mathbf{y}_i^{\text{rep}}|\boldsymbol{\Omega}, \boldsymbol{\theta}) = \boldsymbol{\Lambda}_\eta \boldsymbol{\Pi}_0^{-1}\boldsymbol{\Psi}_\delta(\boldsymbol{\Lambda}_\eta \boldsymbol{\Pi}_0^{-1})^T + \boldsymbol{\Psi}_\epsilon. \tag{26}$$

To compute the L_ν-measure given in (21), we use the following identities to utilize the simulated observations already available in the estimation:

$$E(\mathbf{y}_i^{\text{rep}}|\mathbf{Y}) = E\{E(\mathbf{y}_i^{\text{rep}}|\boldsymbol{\Omega},\boldsymbol{\theta})|\mathbf{Y}\}, \quad \text{and}$$
$$\text{Cov}(\mathbf{y}_i^{\text{rep}}|\mathbf{Y}) = E\{\text{Cov}(\mathbf{y}_i^{\text{rep}}|\boldsymbol{\Omega},\boldsymbol{\theta})|\mathbf{Y}\} + \text{Cov}\{E(\mathbf{y}_i^{\text{rep}}|\boldsymbol{\Omega},\boldsymbol{\theta})|\mathbf{Y}\}.$$

Let $\{(\boldsymbol{\theta}^{(j)}, \boldsymbol{\Omega}^{(j)}),\ j = 1, \cdots, J\}$ be simulated observations from $p(\boldsymbol{\theta}, \boldsymbol{\Omega}|\mathbf{Y})$, it follows from (20), (21), and the above identities that:

$$\widehat{E}(\mathbf{y}_i^{\text{rep}}|\mathbf{Y}) = \frac{1}{J}\sum_{j=1}^{J} \mathbf{m}_i^{(j)},$$

$$\widehat{\text{Cov}}(\mathbf{y}_i^{\text{rep}}|\mathbf{Y}) = \frac{1}{J}\sum_{j=1}^{J}[\boldsymbol{\Lambda}_\eta^{(j)}(\boldsymbol{\Pi}_0^{(j)})^{-1}\boldsymbol{\Psi}_\delta^{(j)}(\boldsymbol{\Lambda}_\eta^{(j)}(\boldsymbol{\Pi}_0^{(j)})^{-1})^T + \boldsymbol{\Psi}_\epsilon^{(j)}] +$$
$$\frac{1}{J}\sum_{j=1}^{J}\mathbf{m}_i^{(j)}\mathbf{m}_i^{(j)T} - \left(\frac{1}{J}\sum_{j=1}^{J}\mathbf{m}_i^{(j)}\right)\left(\frac{1}{J}\sum_{j=1}^{J}\mathbf{m}_i^{(j)}\right)^T,$$

where $\mathbf{m}_i^{(j)} = \boldsymbol{\mu}^{(j)} + \boldsymbol{\Lambda}_\eta^{(j)}(\boldsymbol{\Pi}_0^{(j)})^{-1}\{\boldsymbol{\Gamma}^{(j)}\mathbf{F}(\boldsymbol{\xi}_i^{(j)})\} + \boldsymbol{\Lambda}_\xi^{(j)}\boldsymbol{\xi}_i^{(j)}$. Hence, an estimate of the L_ν measure defined by (21) can be obtained.

We use the following data set to illustrate model comparison via various statistics. A small portion of the Inter-university Consortium for Political and Social Research (ICPSR) data set collected in project WORLD VALUES SURVEY 1981-1984 and 1990-1993 (World Values Study Group, ICPSR Version) is considered.

Six variables in the original data set obtained from United Kingdom (variables 180, 96, 62, 176, 116 and 117; see Appendix 1.1) that related to respondents' job, religious belief, and home life were taken as observed variables in $\mathbf{y} = (y_1, \cdots, y_6)^T$. After deleting missing data, the sample size was 197. Among them,

- (y_1, y_2) — measure 'life'
- (y_3, y_4) — measure 'religious belief'
- (y_5, y_6) — measure 'job satisfaction'

Variable y_3 was measured in a five-point scale, while all others were measured in a ten-point scale, and they were treated as continuous for brevity.

The competing models are defined with a measurement equation with three latent variables $\{\eta, \xi_1, \xi_2\}$ and the following loading matrix:

$$\mathbf{\Lambda}^T = \begin{bmatrix} 1 & \lambda_{21} & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & \lambda_{42} & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & \lambda_{63} \end{bmatrix}.$$

The structural equations of the competing models are given as follows:

$$\begin{aligned}
M_1: &\quad \eta_i = \gamma_1 \xi_{i1} + \gamma_2 \xi_{i2} + \delta_i, \\
M_2: &\quad \eta_i = \gamma_1 \xi_{i1} + \gamma_2 \xi_{i2} + \gamma_3 \xi_{i1}^2 + \delta_i, \\
M_3: &\quad \eta_i = \gamma_1 \xi_{i1} + \gamma_2 \xi_{i2} + \gamma_3 \xi_{i2}^2 + \delta_i, \\
M_4: &\quad \eta_i = \gamma_1 \xi_{i1} + \gamma_2 \xi_{i2} + \gamma_3 \xi_{i1} \xi_{i2} + \delta_i, \\
M_5: &\quad \eta_i = \gamma_1 \xi_{i1} + \gamma_2 \xi_{i2} + \gamma_3 \xi_{i1}^2 + \gamma_4 \xi_{i2}^2 + \gamma_5 \xi_{i1} \xi_{i2} + \delta_i.
\end{aligned}$$

The following hyperparameters were selected in the analysis:

- $\alpha_{0\epsilon k} = \alpha_{0\delta} = 10,\ \beta_{0\epsilon k} = \beta_{0\delta} = 8$;
- $\mathbf{H}_{0yk}$ and $\mathbf{H}_{0\omega k}$ are diagonal matrices with diagonal element 0.25;
- $\rho_0 = 20,\ \mathbf{\Sigma}_0 = \mathbf{I}_6,\ \mathbf{R}_0^{-1} = 2\tilde{\mathbf{\Phi}}$, where
- $\mathbf{\Lambda}_{0k} = \tilde{\mathbf{\Lambda}}_{0k}$, and $\mathbf{\Gamma}_{0k} = \tilde{\mathbf{\Gamma}}_{0k}$;

where $\tilde{\mathbf{\Lambda}}_{0k}, \tilde{\mathbf{\Gamma}}_{0k}$, and $\tilde{\mathbf{\Phi}}$ were the Bayesian estimates obtained on the basis of M_1 and noninformative prior distributions.

When checking convergence, we found that the MCMC algorithm converged within 2,000 iterations. Hence, the burn-in iterations were taken as 2,000. The following results were obtained through additional 2,000 observations collected after convergence.

The following values of the $L_{0.5}$ measure were obtained:

1. $L_{(1)} = 3657.8, L_{(2)} = 3652.67, L_{(3)} = 3702.8, L_{(4)} = 3568.4,$ and $L_{(5)} = 3853.5$, where $L_{(k)}$ is the $L_{0.5}$ measure corresponding to M_k. Based on these results, M_4 is selected.
2. The DIC values obtained from WinBUGS are equal to: $\text{DIC}_{(1)} = 4093.0$, $\text{DIC}_{(2)} = 4090.5$, $\text{DIC}_{(3)} = 4093.9$, $\text{DIC}_{(4)} = 4081.6$, and $\text{DIC}_{(5)} = 4087.6$. Based on the DIC values, M_4 is selected again.
3. $2 \log B_{14} = -5.336$, $2 \log B_{24} = -8.626$, $2 \log B_{34} = -5.748$, and $2 \log B_{54} = -0.246$. Again, M_4 is selected.

Hence, we draw the same conclusion that a nonlinear SEM with an interaction is selected for fitting the data set.

The model comparison statistics discussed in previous sections can be used to assess the goodness-of-fit of the hypothesized model by taking M_0 or M_1 to be the saturated model. $\mathcal{N}(\mu, \Sigma)$ μ, Σ are free, i.e. without any structure to be constrained

However, for some complex SEMs, it is rather difficult to define a saturated model. For example, in the analysis of nonlinear SEMs, the distribution of the observed random vector associated with the hypothesized model is not normal. Thus, the model assuming a normal distribution with a general unstructured covariance matrix cannot be regarded as a saturated model? Under these situations, the model comparison statistics, such as the Bayes factor, BIC, AIC, and DIC cannot be applied to access goodness-of-fit of the hypothesized model.

A simple and more convenient alternative without involving basic saturated model is the posterior predictive p-values (PP p-values) introduced by Meng (1994) on the basis of the posterior assessment in Rubin (1984).

What's the Ω?

Let $D(\mathbf{Y}|\boldsymbol{\theta}, \boldsymbol{\Omega})$ be a discrepancy measure that is used to capture the discrepancy between the hypothesized model M_0 and the data, and let $\mathbf{Y}^{\text{rep}}$ be the generated hypothetical replicate data. The PP p-value is defined by

$$p_B(\mathbf{Y}) = Pr\{D(\mathbf{Y}^{\text{rep}}|\boldsymbol{\theta}, \boldsymbol{\Omega}) \geq D(\mathbf{Y}|\boldsymbol{\theta}, \boldsymbol{\Omega})|\mathbf{Y}, M_0\}, \tag{27}$$

which is the upper-tail probability of the discrepancy measure under its posterior predictive distribution. See Appendix 4.3 for computation of $p_B(\mathbf{Y})$. The PP p-values not far from 0.5 indicate that the realized discrepancies are near the center of the posterior predictive distribution of the discrepancy measure. Hence, a hypothesized model may be considered as plausible when its PP p-value is reasonably close to 0.5.

otherwise, they are quite different.

Many common model checking methods in data analysis, such as residual analysis, can be incorporated in the Bayesian analysis. An advantage of the sampling-based Bayesian approach for SEMs is that we can obtain the estimates of the latent variables through the posterior simulation so that reliable estimates of the residuals in the measurement equation and the structural equation can be obtained. The graphical interpretation of these residuals is similar to those in other statistical models, for example, regression.

As an illustration of the basic idea, consider the SEMs with fixed covariates. Estimates of the residuals in the measurement equation can be obtained from (2.13) as:

$$\hat{\epsilon}_i = \mathbf{y}_i - \hat{\mathbf{A}}\mathbf{c}_i - \hat{\Lambda}\hat{\omega}_i, \quad i = 1, \cdots, n, \tag{28}$$

where $\hat{\mathbf{A}}$, $\hat{\Lambda}$, and $\hat{\omega}_i$ are Bayesian estimates that are obtained from the corresponding simulated observations through the MCMC methods. Plots of $\hat{\epsilon}_i$ versus $\hat{\omega}_i$ give useful information for the fit of the measurement equation. For a reasonably good fit, the plots should lie within two parallel horizontal lines that are not widely separated apart and centered at zero.

Estimates of residuals in the structural equation can be obtained from (2.15) as:

$$\hat{\boldsymbol{\delta}}_i = (\mathbf{I} - \hat{\boldsymbol{\Pi}})\hat{\boldsymbol{\eta}}_i - \hat{\mathbf{B}}\mathbf{d}_i - \hat{\boldsymbol{\Gamma}}\hat{\boldsymbol{\xi}}_i, \quad i = 1, \cdots, n, \tag{29}$$

where $\hat{\boldsymbol{\Pi}}$, $\hat{\mathbf{B}}$, $\hat{\boldsymbol{\Gamma}}$, $\hat{\boldsymbol{\eta}}_i$, and $\hat{\boldsymbol{\xi}}_i$ are Bayesian estimates. The interpretation and the use of plots of $\hat{\boldsymbol{\delta}}_i$ and $\hat{\boldsymbol{\epsilon}}_i$ are similar.

The residual estimates $\hat{\boldsymbol{\epsilon}}_i$ can also be used for outliers analysis. A particular observation $\mathbf{y}_i$ whose residual is far from zero may be informally regarded as an outlier. Moreover, the QQ plots of $\hat{\epsilon}_{ij}$, $j = 1, \cdots, p$, and $\hat{\delta}_{ik}$, $k = 1, \cdots, q_1$, can be used to check the assumption of normality.