4 Bayesian Model Comparison and Model Checking

4.1 Introduction

In Chapter 3, we introduced the Bayesian approach for estimating parameters in SEMs. We showed that this approach when coupled with MCMC methods provides an efficient and flexible tool for fitting SEMs. 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, as pointed out in the statistics literature (see e.g., Berger and Delampady, 1987; Berger and Sellke 1987; Kass and Raftery, 1995), there are serious problems associated with such an approach. See Lee (2007, Chapter 5) for a discussion of these problems in relation to SEMs.

The main objectives of this chapter are: (i) to introduce various Bayesian statistics for hypothesis testing and model comparison, and (ii) to provide some statistical methods for assessment of the goodness-of-fit of the posited model and for diagnostic of the model. In our Bayesian approach, we will consider the issue of hypothesis testing as model comparison, mainly because a hypothesis can be represented via a specific model. Hence, 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. We use the artificial example presented in Section 3.5 as an illustrative example of the above idea. Suppose that we are interested in testing $H_0: \ \gamma_3 = \gamma_4 = \gamma_5 = 0$, against $H_1: \ \gamma_3 \neq 0, \ \gamma_4 \neq 0$, and $\gamma_5 \neq 0$;

see Equation (3.19). We can define an SEM, M_0, with a measurement equation defined by (3.18) and a structural equation defined by $\nu_i = b_1 d_i + \gamma_1 \xi_{i1} + \gamma_2 \xi_{i2}$. This gives a model corresponding to H_0. The model M_1 that corresponds to the alternative hypothesis H_1 is defined by Equations (3.18) and (3.19). Similarly, other null and alternative hypotheses can be assessed as a model comparison problem. Hence, in this book, we will use the general term 'model comparison' to represent hypothesis testing and model selection.

A common Bayesian statistic for model comparison in the field of SEMs is the Bayes factor (see Lee, 2007). It has been shown that this statistic has many nice statistical properties (see Kass and Raftery, 1995). Computationally, the evaluation of Bayes factor can be difficult. Recently, various algorithms for computing the Bayes factor have been developed on the basis of posterior simulation via MCMC methods. Based on a comparative study on a variety of algorithms, DiCiccio *et al.* (1997) concluded that bridge sampling is an attractive method. However, Gelman and Meng (1998) showed that path sampling is a direct extension of bridge sampling and can give even better results. In addition to the Bayes factor, we will introduce several other Bayesian statistics for model comparison, namely Bayesian Information Criterion (BIC), Akaike Information Criterion (AIC), Deviance Information Criterion (DIC), and the L_ν-measure, a criterion-based statistic. The Bayes factor and/or the abovementioned statistics will be applied to cope with the model comparison problem in the context of various complex SEMs and data structures, see subsequent chapters in this book.

We will give sufficient technical details for researchers to implement their own program in computing the aforementioned Bayesian statistics. For applied researchers who don't want to write their own program, WinBUGS directly provides the DIC values for many complex SEMs and data structures. Moreover, by utilizing the program R2WinBUGS,

results provided by WinBUGS can be conveniently used to compute some other model comparison statistics.

An introduction of the Bayes factor will be presented in Section 4.2. Here, discussions related to path sampling and WinBUGS for computing this statistic will be included; and an application of the methodology to SEMs with fixed covariates will be provided. Some other methods for model comparison are given in Section 4.3. An illustrative example is given in Section 4.4. Methods for model checking and goodness-of-fit are discussed in Section 4.5.

4.2 Bayes Factor

In this section, we introduce an important Bayesian statistic, the Bayes factor (Berger, 1985) for model comparison. This statistic has a solid logical foundation that offers great flexibility. It has been extensively applied to a lot of statistical models (Kass and Raftery, 1995) and SEMs (Lee, 2007).

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

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

Hence,

$$\frac{p(M_1|\mathbf{Y})}{p(M_0|\mathbf{Y})} = \frac{p(\mathbf{Y}|M_1)p(M_1)}{p(\mathbf{Y}|M_0)p(M_0)}. \tag{4.1}$$

The Bayes factor for comparing M_1 and M_0 is defined as

$$B_{10} = \frac{p(\mathbf{Y}|M_1)}{p(\mathbf{Y}|M_0)}. \tag{4.2}$$

From (4.1), we see that 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. In general, it is a summary of evidence provided by the data in favor of M_1 as oppose to M_0, or in favor of M_0 as oppose to M_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. 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'. Moreover, it can be seen from (4.2) 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. Finally, 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. Kass and Raftery (1995) pointed out that these categories furnish appropriate guidelines for practical applications of the Bayes factor. Depending on the competing models M_0 and M_1 in fitting a given data set, if the Bayes factor (or 2log Bayes factor) rejects the null hypothesis H_0 that is associated with M_0, we can conclude that the data give evidence to support the alternative hypothesis H_1 that is associated with M_1. Similarly, if the Bayes factor rejects H_1, a definite conclusion of supporting H_0 can be attained.

Table 4.1 here

The interpretation of evidence provided by Table 4.1 depends on the specific context. For two nonnested competing models, say M_1 and M_0, we should 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. For two nested competing models, say M_0 is nested in M_1, $2\log B_{10}$ is most likely larger than zero. If M_1 is significantly better than M_0, $2\log B_{10}$ can be much larger than 6. Then 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 worth more than a bare mention'. 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. The criterion given in Table 4.1 is a suggestion, it is not necessary to regard it as a strict rule. Similar to other data analyses, for conclusions drawn from the marginal cases, it is always helpful to conduct other analysis, for example residual analysis, to cross-validate the results. Generally speaking, model selection should be approached on a problem-by-problem basis. In certain circumstances, the opinions from experts may also be taken into account.

The prior distribution of $\boldsymbol{\theta}$, $p(\boldsymbol{\theta})$, has to be specified in computing a Bayes factor. Compared to Bayesian estimates, the values of Bayes factor are more sensitive to prior inputs. Hence, 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". To avoid this difficulty, priors on parameters under the model comparison are generally taken to be proper and not having a too big spread. The conjugate families with reasonable spreads are appropriate choices. Prior inputs for the hyperparameters in the conjugate prior distributions may come from analyses of past or similar data, or from the subjective knowledge of experts. 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 which is combined with a noninformative prior distribution to produce an informative prior distribution. The Bayes factor is then computed from the remainder of the data. More advanced methods have been suggested; see for example O'Hagan (1995), and Berger and Pericchi (1996), among others. To study the sensitivity of the Bayes factor to the choice of prior inputs in terms of the hyperparameter values, a common method (see Kass and Raftery, 1995; Lee and Song, 2003; among others) is to perturb the prior inputs. For example, if the prior distribution is $N[\mu_0, \sigma_0^2]$ in which the given hyperparameters are μ_0 and σ_0^2, the hyperparameters may be perturbed by changing μ_0 to $\mu_0 \pm c$ and halving or doubling σ_0^2, and the Bayes factor is recomputed accordingly.

4.2.1 Path Sampling

From its definition, we observe that Bayes factor involves the density $p(\mathbf{Y}|M_k)$. Let $\boldsymbol{\theta}_k$ be the random parameter vector associated with M_k. From the fact that $p(\boldsymbol{\theta}_k, \mathbf{Y}|M_k) = p(\mathbf{Y}|\boldsymbol{\theta}_k, M_k)p(\boldsymbol{\theta}_k|M_k)$, we have

$$p(\mathbf{Y}|M_k) = \int p(\mathbf{Y}|\boldsymbol{\theta}_k, M_k)p(\boldsymbol{\theta}_k|M_k)d\boldsymbol{\theta}_k, \tag{4.3}$$

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 $\boldsymbol{\theta}_k$. This quantity can be interpreted as the marginal likelihood of the data, obtained by integrating the joint density of $(\mathbf{Y}, \boldsymbol{\theta}_k)$ over $\boldsymbol{\theta}_k$. It can also be interpreted as the predictive probability of the data; that is, the probability of observing the data that actually were observed, calculated before any data became available. Sometimes, it is also called an integrated likelihood. Note that, as in the computation of the likelihood ratio statistic but unlike in some other applications of the likelihood, all constants appearing in the definition of the

likelihood $p(\mathbf{Y}|\boldsymbol{\theta}_k, M_k)$ must be retained when computing B_{10}. Very often, it is difficult to obtain B_{10} analytically, and various analytic and numerical approximations have been proposed in the literature. For example, Chib (1995) and Chib and Jeliazkov (2001) developed efficient algorithms for computing the marginal likelihood through MCMC chains produced by the Gibbs sampler and by the MH algorithm, respectively. Based on the results of DiCiccio *et al.* (1997), and the recommendation of Gelman and Meng (1998), we will discuss the application of path sampling to compute the Bayes factor for model comparison. To simplify notation, 'M_k' will be suppressed; hence $p(\mathbf{Y}) = p(\mathbf{Y}|M_k)$, etc.

In general, let $\mathbf{Y}$ be the matrix of observed data, and $\boldsymbol{\Omega}$ be the matrix of latent variables in the model. For SEMs which involve latent variables, direct application of path sampling (Gelman and Meng, 1998) in computing the Bayes factor is difficult. Similar to Bayesian estimation, we utilize the idea of data augmentation (Tanner and Wong, 1987) to solve the problem. Below we use the similar reasoning as in Gelman and Meng (1998) to briefly show that path sampling can be applied to compute the logarithm of the Bayes factor by augmenting $\mathbf{Y}$ with $\boldsymbol{\Omega}$. The main result is given by Equations (4.8) and (4.9), with the definition of $U(\mathbf{Y}, \boldsymbol{\Omega}, \boldsymbol{\theta}, t)$ given by (4.7). Readers who are not interested in the technical derivation may jump to these equations. From the equality $p(\boldsymbol{\Omega}, \boldsymbol{\theta}|\mathbf{Y}) = p(\mathbf{Y}, \boldsymbol{\Omega}, \boldsymbol{\theta})/p(\mathbf{Y})$, the marginal density $p(\mathbf{Y})$ 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 $[0, 1]$:

$$p(\boldsymbol{\Omega}, \boldsymbol{\theta}|\mathbf{Y}, t) = \frac{1}{z(t)} p(\mathbf{Y}, \boldsymbol{\Omega}, \boldsymbol{\theta}|t), \tag{4.4}$$

where

$$z(t) = p(\mathbf{Y}|t) = \int p(\mathbf{Y}, \mathbf{\Omega}, \boldsymbol{\theta}|t) d\mathbf{\Omega} d\boldsymbol{\theta} = \int p(\mathbf{Y}, \mathbf{\Omega}, |\boldsymbol{\theta}, t) p(\boldsymbol{\theta}) d\mathbf{\Omega} d\boldsymbol{\theta}, \tag{4.5}$$

with $p(\boldsymbol{\theta})$ be the prior density of $\boldsymbol{\theta}$ which is assumed to be independent of t.

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 $z(1) = p(\mathbf{Y}|1) = p(\mathbf{Y}|M_1)$, $z(0) = p(\mathbf{Y}|0) = p(\mathbf{Y}|M_0)$, and $B_{10} = z(1)/z(0)$. Taking logarithm and then differentiating (4.5) with respect to t, and assuming the legitimacy of interchange of integration with differentiation, we have

$$\begin{aligned} \frac{d \log z(t)}{dt} =& \int \frac{1}{z(t)} \frac{d}{dt} p(\mathbf{Y}, \mathbf{\Omega}, \boldsymbol{\theta}|t) d\mathbf{\Omega} d\boldsymbol{\theta} \\ =& \int \frac{d}{dt} \log p(\mathbf{Y}, \mathbf{\Omega}, \boldsymbol{\theta}|t) \cdot p(\mathbf{\Omega}, \boldsymbol{\theta}|\mathbf{Y}, t) d\mathbf{\Omega} d\boldsymbol{\theta} \\ =& E_{\Omega,\theta} \left[\frac{d}{dt} \log p(\mathbf{Y}, \mathbf{\Omega}, \boldsymbol{\theta}|t) \right], \end{aligned} \tag{4.6}$$

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

$$U(\mathbf{Y}, \mathbf{\Omega}, \boldsymbol{\theta}, t) = \frac{d}{dt} \log p(\mathbf{Y}, \mathbf{\Omega}, \boldsymbol{\theta}|t) = \frac{d}{dt} \log p(\mathbf{Y}, \mathbf{\Omega}|\boldsymbol{\theta}, t), \tag{4.7}$$

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

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

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

$$\widehat{\log B_{10}} = \frac{1}{2} \sum_{s=0}^{S} (t_{(s+1)} - t_{(s)})(\bar{U}_{(s+1)} + \bar{U}_{(s)}), \tag{4.8}$$

where $\bar{U}_{(s)}$ is the following average of the values of $U(\mathbf{Y}, \mathbf{\Omega}, \boldsymbol{\theta}, t)$ based on simulation draws at $t = t_{(s)}$,

$$\bar{U}_{(s)} = J^{-1} \sum_{j=1}^{J} U(\mathbf{Y}, \mathbf{\Omega}^{(j)}, \boldsymbol{\theta}^{(j)}, t_{(s)}), \tag{4.9}$$

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

To apply the path sampling procedure, we need to define a link model M_t to link M_0 and M_1, such that when $t = 0$, $M_t = M_0$; and when $t = 1$, $M_t = M_1$. Then, we 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, and finally estimate $\log B_{10}$ via (4.7) and (4.8). Note that the form and the derivative of the complete-data likelihood are not difficult. The main computation is on simulating the sample of observations $\{(\boldsymbol{\Omega}^{(j)}, \boldsymbol{\theta}^{(j)}),\ j = 1, \cdots, J\}$ from $p(\boldsymbol{\Omega}, \boldsymbol{\theta}|\mathbf{Y}, t_{(s)})$, for $s = 0, \cdots, S + 1$. This task can be done via some efficient MCMC methods, such as the Gibbs sampler and the MH algorithm as described in the last chapter; see illustrative examples given in other chapters in this book. 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. However, a smaller J is not recommended.

The path sampling approach has several nice features. Its implementation is simple, the main programming task is simulating observations from $p(\boldsymbol{\Omega}, \boldsymbol{\theta}|\mathbf{Y}, t_{(s)})$. As pointed out by Gelman and Meng (1998), we can always construct a continuous path to link two competing models. Thus, the method can be applied to the comparison of a wide variety of models. Bayesian estimates of the unknown parameters and latent variables under M_0 and M_1 can be obtained easily via the simulated observations at $t = 0$ and $t = 1$. In contrast to most existing methods in computing the Bayes factor, the path sampling procedure does not directly include the prior density in the computation. Furthermore, the logarithm scale of Bayes factor is computed, which is generally more stable than the ratio scale. Finally, as path sampling is a generalization of bridge sampling, it has potential to produce more accurate results.

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. Most of the time, this difficulty can be solved by using appropriate auxiliary models, $M_a,\ M_b, \cdots$, in between M_1 and M_0. For example, 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 of the Bayes factor can be computed through path sampling. See an illustrative example in Section 5.2.4.

4.2.2 A Simulation Study

The objectives of this simulation study are to reveal the performance of path sampling in computing Bayes factor, and to evaluate the sensitivity of the results to prior inputs. Random observations were generated from a nonlinear SEM with fixed covariates defined by (2.21) and (2.22). The specific model involves eight observed variables which are related to two fixed covariates $\{c_{i1}, c_{i2}\}$ in the measurement equation, and three latent variables $\{\eta_i, \xi_{i1}, \xi_{i2}\}$ and one fixed covariate d_i in the structural equation. The first fixed covariate 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 $1.0 - \Phi^*(0.5)$, respectively; where $\Phi^*(\cdot)$ is the distribution function of $N[0,1]$; while the second covariate c_{i2} is sampled from $N[0,1]$. The true population values in matrices $\mathbf{A}$, $\mathbf{\Lambda}$, and $\mathbf{\Psi}_\epsilon$ are given as follows:

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

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

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. The true variances and covariance of ξ_{i1} and ξ_{i2} are $\phi_{11} = \phi_{22} = 1.0$, and $\phi_{21} = 0.15$. These two explanatory latent variables are related to the outcome latent variable η_i by

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

where d_i is sampled from a Bernoulli distribution that takes 1.0 with probability 0.7 and 0.0 with probability 0.3; and $\psi_\delta = 1.0$. Based on these specifications, 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 for each case.

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:

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

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

a more detailed illustration for the application of path sampling procedure to model comparison of nonlinear SEMs, the implementation of path sampling 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:

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

Clearly, 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:

$$\begin{aligned} &U(\mathbf{Y}, \boldsymbol{\Omega}, \boldsymbol{\theta}, t_{(s)}) = d \log p(\mathbf{Y}, \boldsymbol{\Omega}|\boldsymbol{\theta}, t)/dt\big|_{t=t_{(s)}} \\ &= -\sum_{i=1}^{n} (\eta_i - bd_i - \gamma_1\xi_{i1} - \gamma_2\xi_{i2} - (1-t_{(s)})\gamma_{12}\xi_{i1}\xi_{i2} - t_{(s)}\gamma_{22}\xi_{i2}^2)(\gamma_{12}\xi_{i1}\xi_{i2} - \gamma_{22}\xi_{i2}^2)/\psi_\delta. \end{aligned}$$

Step 4: Calculate $\bar{U}_{(s)}$; see (4.9).

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 (4.8).

Conjugate prior distributions (see for example, Equations (3.6) and (3.8)) are used in the Bayesian analysis. In the sensitivity analysis concerning about the prior inputs, we

followed the suggestion of Kass and Raftery (1995) to perturb them 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, the covariance matrices $\mathbf{\Sigma}_0$, $\mathbf{H}_{0yk}$, and $\mathbf{H}_{0\omega k}$ were taken as the identity matrices with appropriate dimensions. Moreover, 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 of the 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 Table 4.2. Moreover, for each $k = 1, \cdots, 6$, we evaluate

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

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 Table 4.3; for example, the maximum difference of the estimates of $\log B_{01}$ obtained via priors (I) and (II) is 6.55. From the rows of Table 4.2, we observe that the means and standard deviations of $\widehat{\log B_{0k}}$ obtained under different prior inputs are close to each other. This indicates that the estimate of $\log B_{0k}$ is not very sensitive to these prior inputs under a sample size of 300. We also see from Table 4.3 that even for the worst situation with the maximum absolute deviation, the estimated logarithm of Bayes factors under different prior inputs give the same conclusion for selecting the model based on the criterion given in Table 4.1.

Tables 4.2 and 4.3 here

It is clear from Table 4.2 that 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. In comparison with the encompassing model M_6, we found that out of 100 replications under prior (I), 75 of the $\widehat{\log B_{60}}$ 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. Owing to randomness, in only 2 of 100 replications the $\widehat{\log B_{60}}$ mildly support the encompassing model. Although the encompassing model is not the true model, it should not be regarded as an incorrect model for fitting the data.

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 of the 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 of the 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.

4.3 Other Model Comparison Statistics

4.3.1 Bayesian Information Criterion and Akaike Information Criterion

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

$$2\log B_{10} \cong 2S^* = 2\{\log p(\mathbf{Y}|\tilde{\boldsymbol{\theta}}_1, M_1) - \log p(\mathbf{Y}|\tilde{\boldsymbol{\theta}}_0, M_0)\} - (d_1 - d_0)\log n, \tag{4.10}$$

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:

$$\text{BIC}_{10} = -2S^* \cong -2\log B_{10} = 2\log B_{01}. \tag{4.11}$$

The interpretation of BIC_{10} can be based on Table 4.1. Alternatively, for each M_k, $k = 0,\ 1$, we can define

$$\text{BIC}_k = -2\log p(\mathbf{Y}|\tilde{\boldsymbol{\theta}}_k, M_k) + d_k \log n. \tag{4.12}$$

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. The BIC is appealing in that it is relatively simple and can be applied even when the priors $p(\boldsymbol{\theta}_k|M_k)$ $(k = 1, 0)$ are hard to specify precisely. 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 and the BIC obtained can be interpreted using the criterion given in Table 4.1. See Raftery (1993) for an application of BIC to the standard LISREL model that is based on the normal assumption and a linear structural equation. Under this simple case, the computation of the observed-data log-likelihood $\log p(\mathbf{Y}|\tilde{\boldsymbol{\theta}}_k, M_k)$ is in closed form and its computation is straightforward. 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) for an application of path sampling to evaluate the observed-data log-likelihood function.

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

$$\text{AIC}_k = -2 \log p(\mathbf{Y}|\tilde{\boldsymbol{\theta}}_k, M_k) + 2d_k, \tag{4.13}$$

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 (4.12) with (4.13), we see that BIC tends to favor simpler models.

4.3.2 Deviance Information Criterion

Another model comparison statistic that compromises the goodness-of-fit and model complexity is the Deviance Information Criterion (DIC), see Spiegelhalter *et al.* (2002). This statistic is intended as a generalization of AIC. Under a competing model M_k with a vector of unknown parameter $\boldsymbol{\theta}_k$, the DIC is defined as

$$\text{DIC}_k = \overline{D(\boldsymbol{\theta}_k)} + d_k, \tag{4.14}$$

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

$$\overline{D(\boldsymbol{\theta}_k)} = E_{\boldsymbol{\theta}_k}\{-2\log p(\mathbf{Y}|\boldsymbol{\theta}_k, M_k)|\mathbf{Y}\}. \tag{4.15}$$

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

$$d_k = E_{\boldsymbol{\theta}_k}\{-2\log p(\mathbf{Y}|\boldsymbol{\theta}_k, M_k)|\mathbf{Y}\} + 2\log p(\mathbf{Y}|\tilde{\boldsymbol{\theta}}_k), \tag{4.16}$$

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 (4.15) and (4.16) can be estimated as follows:

$$E_{\boldsymbol{\theta}_k}\{-2\log p(\mathbf{Y}|\boldsymbol{\theta}_k, M_k)|\mathbf{Y}\} = -\frac{2}{J}\sum_{j=1}^{J}\log p(\mathbf{Y}|\boldsymbol{\theta}_k^{(j)}, M_k). \tag{4.17}$$

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. In the analysis of a hypothesized model, WinBUGS (Spiegelhalter, *et al.*, 2003) produces a DIC value which can be used for model comparison. Thus, it is very convenient to apply DIC in practice.

As pointed out in the WinBUGS manual (Spiegelhalter *et al.*, 2003), it is important to note the following in practical application of DIC: (i) DIC assumes the posterior mean to be a good estimate of the parameter. There are circumstances, such as mixture models, in which WinBUGS does not give the DIC values. (ii) If the difference in DIC is small, for example less than 5, and the models make very different inferences, then just reporting the model with the lowest DIC could be misleading. (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. Detailed discussions of DIC can be found in Spiegelhalter *et al.* (2002), and Celeux *et al.* (2006).

4.3.3 L_ν-Measure

The L_ν-measure can be viewed as a criterion-based method for Bayesian model assessment which is developed on the basis of the predictive approach with future values of a replicate experiment. More specifically, this statistic 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, whereas 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 $\mathbf{Y}$ be the observed data, and let $p(\mathbf{Y}, \boldsymbol{\theta})$ be the joint density that corresponds to a model M with a parameter vector $\boldsymbol{\theta}$. Considering the predictive approach of using a future response vector for model comparison, we propose a Bayesian model selection statistic through future responses $\mathbf{Y}^{\text{rep}} = (\mathbf{y}_1^{\text{rep}}, \cdots, \mathbf{y}_n^{\text{rep}})$, 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

$$L_1(\mathbf{Y}, \mathbf{B}, \delta) = E[\text{tr}(\mathbf{Y}^{\text{rep}} - \mathbf{B})^T(\mathbf{Y}^{\text{rep}} - \mathbf{B})] + \delta \text{ tr}(\mathbf{Y} - \mathbf{B})^T(\mathbf{Y} - \mathbf{B}), \tag{4.18}$$

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 (4.18), and substituting it to (4.18), it can be shown that (Ibrahim, Chen and Sinha, 2001)

$$L_\nu(\mathbf{Y}) = \sum_{i=1}^{n} \text{tr}\{\text{Cov}(\mathbf{y}_i^{\text{rep}}|\mathbf{Y})\} + \nu \sum_{i=1}^{n} \text{tr}[\{E(\mathbf{y}_i^{\text{rep}}|\mathbf{Y}) - \mathbf{y}_i\}\{E(\mathbf{y}_i^{\text{rep}}|\mathbf{Y}) - \mathbf{y}_i\}^T], \tag{4.19}$$

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.

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:

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

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.

4.4 An Illustration

In this section, we present an example to illustrate the application of the above discussed statistics to model comparison related to nonlinear SEMs. As discussed in Section 2.4.1 in Chapter 2, the model is defined by

$$\mathbf{y}_i = \boldsymbol{\mu} + \boldsymbol{\Lambda}\boldsymbol{\omega}_i + \boldsymbol{\epsilon}_i, \quad \text{and} \tag{4.20}$$

$$\boldsymbol{\eta}_i = \boldsymbol{\Pi}\boldsymbol{\eta}_i + \boldsymbol{\Gamma}\mathbf{F}(\boldsymbol{\xi}_i) + \boldsymbol{\delta}_i, \tag{4.21}$$

where the definitions of $\boldsymbol{\mu},\ \boldsymbol{\Lambda},\ \boldsymbol{\omega}_i,\ \cdots$ are the same as described in Section 2.4.1.

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

$$\mathbf{y}_i = \boldsymbol{\mu} + \boldsymbol{\Lambda}_\eta\boldsymbol{\Pi}_0^{-1}\{\boldsymbol{\Gamma}\mathbf{F}(\boldsymbol{\xi}_i) + \boldsymbol{\delta}_i\} + \boldsymbol{\Lambda}_\xi\boldsymbol{\xi}_i + \boldsymbol{\epsilon}_i, \tag{4.22}$$

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 $p(\mathbf{Y}|\boldsymbol{\Omega},\boldsymbol{\theta})$, we have

$$E(\mathbf{y}_i^{\text{rep}}|\boldsymbol{\Omega},\boldsymbol{\theta}) = \boldsymbol{\mu} + \boldsymbol{\Lambda}_\eta\boldsymbol{\Pi}_0^{-1}\boldsymbol{\Gamma}\mathbf{F}(\boldsymbol{\xi}_i) + \boldsymbol{\Lambda}_\xi\boldsymbol{\xi}_i, \tag{4.23}$$

$$\text{Cov}(\mathbf{y}_i^{\text{rep}}|\boldsymbol{\Omega},\boldsymbol{\theta}) = \boldsymbol{\Lambda}_\eta\boldsymbol{\Pi}_0^{-1}\boldsymbol{\Psi}_\delta(\boldsymbol{\Lambda}_\eta\boldsymbol{\Pi}_0^{-1})^T + \boldsymbol{\Psi}_\epsilon. \tag{4.24}$$

To compute the L_ν-measure given in (4.19), we use the following identities to utilize the simulated observations already available in the estimation:

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

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

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 (4.20), (4.21), and the above identities that:

$$\begin{aligned}
\widehat{E}(\mathbf{y}_i^{\mathrm{rep}}|\mathbf{Y}) =& \frac{1}{J}\sum_{j=1}^{J}\mathbf{m}_i^{(j)},\\
\widehat{\mathrm{Cov}}(\mathbf{y}_i^{\mathrm{rep}}|\mathbf{Y}) =& \frac{1}{J}\sum_{j=1}^{J}[\boldsymbol{\Lambda}_\eta^{(j)}(\boldsymbol{\Pi}_0^{(j)})^{-1}\boldsymbol{\Psi}_\delta^{(j)}(\boldsymbol{\Lambda}_\eta^{(j)}(\boldsymbol{\Pi}_0^{(j)})^{-1})^T + \boldsymbol{\Psi}_\epsilon^{(j)}]+\\
& \frac{1}{J}\sum_{j=1}^{J}\mathbf{m}_i^{(j)}\mathbf{m}_i^{(j)T} - \big(\frac{1}{J}\sum_{j=1}^{J}\mathbf{m}_i^{(j)}\big)\big(\frac{1}{J}\sum_{j=1}^{J}\mathbf{m}_i^{(j)}\big)^T,
\end{aligned}$$

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 (4.19) 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) were related to life, (y_3, y_4) were related to religious belief, and (y_5, y_6) were related to job satisfaction. Variable y_3 was measured in a five-point scale, while all others were measured in a ten-point scale. As the purpose of this example is for illustration, they were all 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:

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

Hence, η, ξ_1, and ξ_2 can be roughly interpreted as 'life', 'religious belief', and 'job satisfaction', respectively. The structural equations of the competing models are given as follows: For $i = 1, \cdots, n$,

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

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, $\rho_0 = 20$, $\mathbf{\Sigma}_0 = \mathbf{I}_6$, $\mathbf{R}_0^{-1} = 2\tilde{\mathbf{\Phi}}$, $\mathbf{\Lambda}_{0k} = \tilde{\mathbf{\Lambda}}_{0k}$, 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. We found that the MCMC algorithm converged within 2,000 iterations. Results were obtained through 2,000 observations collected after convergence. The following values of the $L_{0.5}$ measure were obtained: $L_{(1)} = 3657.8, L_{(2)} = 3652.67, L_{(3)} = 3702.8, L_{(4)} = 3568.4$, 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. 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. We have used the Bayes factor to compare

M_4 with others, and obtained the following results: $2\log B_{14} = -5.336$, $2\log B_{24} = -8.626$, $2\log B_{34} = -5.748$, 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.

4.5 Goodness-of-fit and Model Checking Methods

4.5.1 Posterior Predictive p-value

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. 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). 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

$$p_B(\mathbf{Y}) = Pr\{D(\mathbf{Y}^{\text{rep}}|\boldsymbol{\theta}, \boldsymbol{\Omega}) \geq D(\mathbf{Y}|\boldsymbol{\theta}, \boldsymbol{\Omega})|\mathbf{Y}, M_0\}, \tag{4.25}$$

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.

4.5.2 Residual Analysis

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 as described in Section 2.3 of Chapter 2. Estimates of the residuals in the measurement equation can be obtained from (2.13) as:

$$\hat{\boldsymbol{\epsilon}}_i = \mathbf{y}_i - \hat{\mathbf{A}}\mathbf{c}_i - \hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\omega}}_i, \quad i = 1, \cdots, n, \tag{4.26}$$

where $\hat{\mathbf{A}}$, $\hat{\boldsymbol{\Lambda}}$, and $\hat{\boldsymbol{\omega}}_i$ are Bayesian estimates that are obtained from the corresponding simulated observations through the MCMC methods. Plots of $\hat{\boldsymbol{\epsilon}}_i$ versus $\hat{\boldsymbol{\omega}}_i$ 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:

$$\hat{\boldsymbol{\delta}}_i = (\mathbf{I} - \hat{\boldsymbol{\Pi}})\hat{\boldsymbol{\eta}}_i - \hat{\mathbf{B}}\mathbf{d}_i - \hat{\boldsymbol{\Gamma}}\hat{\boldsymbol{\xi}}_i, \quad i = 1, \cdots, n, \tag{4.27}$$

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. More concrete examples of residual analysis in the context

of real data sets will be presented in subsequent chapters.

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.

Appendix 4.1 WinBUGS Code

```
model {
    for (i in 1:N) {
        for (j in 1:8) { y[i,j]~dnorm(mu[i,j], psi[j]) }
        mu[i,1]<-a[1,1]*x[i,1]+a[1,2]*x[i,2]+eta[i]
        mu[i,2]<-a[2,1]*x[i,1]+a[2,2]*x[i,2]+lam[1]*eta[i]
        mu[i,3]<-a[3,1]*x[i,1]+a[3,2]*x[i,2]+lam[2]*eta[i]
        mu[i,4]<-a[4,1]*x[i,1]+a[4,2]*x[i,2]+xi[i,1]
        mu[i,5]<-a[5,1]*x[i,1]+a[5,2]*x[i,2]+lam[3]*xi[i,1]
        mu[i,6]<-a[6,1]*x[i,1]+a[6,2]*x[i,2]+xi[i,2]
        mu[i,7]<-a[7,1]*x[i,1]+a[7,2]*x[i,2]+lam[4]*xi[i,2]
        mu[i,8]<-a[8,1]*x[i,1]+a[8,2]*x[i,2]+lam[5]*xi[i,2]

        #structural equation
        eta[i]~dnorm(nu[i], psd)

        nu[i]<-b*z[i]+gam[1]*xi[i,1]+gam[2]*xi[i,2]
              +t*gam[3]*xi[i,1]*xi[i,2]+(1-t)*gam[4]*xi[i,2]*xi[i,2]

        u[i]<-(eta[i]-nu[i])*psd*(gam[3]*xi[i,1]*xi[i,2]
             -gam[4]*xi[i,2]*xi[i,2])

        xi[i,1:2]~dmnorm(zero[1:2], phi[1:2,1:2])
    }   #end of i

    ubar<-sum(u[])

    #prior distribution
    lam[1]~dnorm(1.5,psi[2])    lam[2]~dnorm(1.5,psi[3])
    lam[3]~dnorm(1.5,psi[5])    lam[4]~dnorm(1.5,psi[7])
    lam[5]~dnorm(1.5,psi[8])

    b~dnorm(1, psd)             gam[1]~dnorm(0.5,psd)
    gam[2]~dnorm(0.5,psd)       gam[3]~dnorm(1.0,psd)
    gam[4]~dnorm(0.0,psd)
```

```
    for (j in 1:8) {
        psi[j]~dgamma(8,10)
        a[j,1]~dnorm(1.0,1)     a[j,2]~dnorm(0.7,1)
    }

    psd~dgamma(8,10)    phi[1:2,1:2]~dwish(R[1:2,1:2], 20)
} #end of model
```

Appendix 4.2 R code in Bayes Factor Example

```
library(mvtnorm)   #Load mvtnorm package
library(R2WinBUGS) #Load R2WinBUGS package

N=300                          #Sample size
AZ=matrix(NA, nrow=N, ncol=2) #Fixed covariates in measurement equation
BZ=numeric(N)                  #Fixed covariate in structural equation
XI=matrix(NA, nrow=N, ncol=2) #Explanatory latent variables
Eta=numeric(N)                 #Outcome latent variables
Y=matrix(NA, nrow=N, ncol=8)  #Observed variables

#The covariance matrix of xi
phi=matrix(c(1, 0.15, 0.15, 1), nrow=2)

p=numeric(3); p[1]=pnorm(-0.5); p[2]=pnorm(0.5)-p[1]; p[3]=1-pnorm(0.5)

#Generate the data
for (i in 1:N) {
    AZ[i,1]=sample(1:3, 1, prob=p); AZ[i,2]=rnorm(1,0,1)
    BZ[i]=rbinom(1,1,0.7)

    XI[i,]=rmvnorm(1, c(0,0), phi)

    delta=rnorm(1,0,1)
    Eta[i]=BZ[i]+0.5*XI[i,1]+0.5*XI[i,2]+XI[i,2]*XI[i,2]+delta

    eps=rnorm(8,0,1)
    Y[i,1]=Eta[i]+eps[1]
```

```
    Y[i,2]=1.5*Eta[i]+eps[2]
    Y[i,3]=1.5*Eta[i]+eps[3]
    Y[i,4]=XI[i,1]+eps[4]
    Y[i,5]=1.5*XI[i,1]+eps[5]
    Y[i,6]=XI[i,2]+eps[6]
    Y[i,7]=1.5*XI[i,2]+eps[7]
    Y[i,8]=1.5*XI[i,2]+eps[8]

    for (j in 1:8) { Y[i,j]=Y[i,j]+AZ[i,1]+0.7*AZ[i,2] }
}

R=matrix(c(17.0 2.55, 2.55, 17.0), nrow=2)

data=list(N=300, zero=c(0,0), x=AZ, z=BZ, R=R, y=Y, t=NA) #Data

init1=list(lam=rep(0,5), a=matrix(rep(0,16), nrow=8, byrow=T),
           gam=rep(0,4), b=0,    psi=rep(1,8),    psd=1,
           phi=matrix(c(1, 0, 0, 1), nrow=2))

init2=list(lam=rep(1,5), a=matrix(rep(1,16), nrow=8, byrow=T),
           gam=rep(1,4), b=1,    psi=rep(2,8),    psd=2,
           phi=matrix(c(2, 0, 0, 2), nrow=2))

inits=list(init1, init2) #Initial values

parameters=c("ubar")

#Path sampling
for (i in 1:21) {
    data$t<-(i-1)*0.05
    model<-bugs(data,inits,parameters,
                model.file="C:/Bayes Factor/model.txt",
                n.chains=2,n.iter=1500,n.burnin=500,n.thin=1,
                bugs.directory="C:/Program Files/WinBUGS14/",
                working.directory="C:/Bayes Factor/")
    u[i]<-model$mean$ubar
```

```
}

#Caluate log Bayes factor
logBF=0
for (i in 1:20) { logBF=logBF+(u[i+1]+u[i])*0.05/2 }
print(logBF)
```

Appendix 4.3 PP p-value for Model Assessment

Based on the posterior predictive assessment as discussed in Rubin (1984), Gelman, Meng and Stern (1996) introduced a Bayesian counterpart of the classical p-value by defining a posterior predictive (PP) p-value for model checking. To apply the approach for establishing a goodness-of-fit assessment of a hypothesized model M_0 with parameter vector $\boldsymbol{\theta}$, observed data $\mathbf{Y}$ and latent data $\boldsymbol{\Omega}$, we consider a discrepancy variable $D(\mathbf{Y}|\boldsymbol{\theta}, \boldsymbol{\Omega})$ for measuring the discrepancy between $\mathbf{Y}$ and the generated hypothetical replicate data $\mathbf{Y}^{\text{rep}}$. Then, the PP p-value is defined as

$$\begin{aligned} p_B(\mathbf{Y}) &= Pr\{D(\mathbf{Y}^{\text{rep}}|\boldsymbol{\theta}, \boldsymbol{\Omega}) \geq D(\mathbf{Y}|\boldsymbol{\theta}, \boldsymbol{\Omega})|\mathbf{Y}, M_0\}, \\ &= \int I\{D(\mathbf{Y}^{\text{rep}}|\boldsymbol{\theta}, \boldsymbol{\Omega}) \geq D(\mathbf{Y}|\boldsymbol{\theta}, \boldsymbol{\Omega})\} p(\mathbf{Y}^{\text{rep}}, \boldsymbol{\theta}, \boldsymbol{\Omega}|\mathbf{Y}, M_0) d\mathbf{Y}^{\text{rep}} d\boldsymbol{\theta} d\boldsymbol{\Omega}. \end{aligned}$$

where $I(\cdot)$ is an indicator function. The probability is taken over the following joint posterior distribution of $(\mathbf{Y}^{\text{rep}}, \boldsymbol{\theta}, \boldsymbol{\Omega})$ given $\mathbf{Y}$ and M_0:

$$p(\mathbf{Y}^{\text{rep}}, \boldsymbol{\theta}, \boldsymbol{\Omega}|\mathbf{Y}, M_0) = p(\mathbf{Y}^{\text{rep}}|\boldsymbol{\theta}, \boldsymbol{\Omega}) p(\boldsymbol{\theta}, \boldsymbol{\Omega}|\mathbf{Y}).$$

In almost all our applications to SEMs considered in this book, we take the chi-square discrepancy variable such that $D(\mathbf{Y}^{\text{rep}}|\boldsymbol{\theta}, \boldsymbol{\Omega})$ has a chi-squared distribution with d^* degrees of freedom. Thus, the PP p-value is equal to

$$\int p\{\chi^2(d^*) \geq D(\mathbf{Y}|\boldsymbol{\theta}, \boldsymbol{\Omega})\} p(\boldsymbol{\theta}, \boldsymbol{\Omega}|\mathbf{Y}) d\boldsymbol{\theta} d\boldsymbol{\Omega}.$$

A Rao-Blackwellized type estimate of this PP p-value is:

$$\hat{p}_B(\mathbf{Y}) = J^{-1} \sum_{j=1}^{J} Pr\{\chi^2(d^*) \geq D(\mathbf{Y}|\boldsymbol{\theta}^{(j)}, \boldsymbol{\Omega}^{(j)})\},$$

where $\{(\boldsymbol{\theta}^{(j)}, \boldsymbol{\Omega}^{(j)}),\ j = 1, \cdots, J\}$ are observations simulated during the estimation. The computational burden for obtaining this $\hat{p}_B(\mathbf{Y})$ is light.

References

Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. In B. N. Petrox and F. Caski (eds), *Second International Symposium on Information Theory*, p. 267. Budapest, Hungary: Akademiai Kiado.

Berger, J. O. (1985) *Statistical Decision Theory and Bayesian Analysis*. New York: Springer-Verlag.

Berger, J. O. and Delampady, M. (1987) Testing precise hypotheses. *Statistical Science*, **3**, 317-352.

Berger, J. O. and Pericchi, L. R. (1996) The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, **91**, 109-122.

Berger, J. O. and Sellke, T. (1987) Testing a point null hypothesis: The irreconcilability of P values and evidence. *Journal of the American Statistical Association*, **82**, 112-122.

Celeux, G., Forbes, F., Robert, C. P. and Titterington, D. M. (2006) Deviance information criteria for missing data models. *Bayesian Analysis*, **1**, 651-674.

Chib, S. (1995) Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, **90**, 1313-1321.

Chib, S. and Jeliazkov, I. (2001) Marginal likelihood from the Metropolis-Hastings output. *Journal of the American Statistical Association*, **96**, 270-281.

DiCiccio, T. J., Kass, R. E., Raftery, A. and Wasserman, L. (1997) Computing Bayes factors by combining simulation and asymptotic approximations. *Journal of the American Statistical Association*, **92**, 903-915.

Gelman, A. and Meng, X. L. (1998) Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science*, **13**, 163-185.

Gelman, A., Meng, X. L. and Stern, H. (1996) Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, **6**, 733-760.

Ibrahim, J. G., Chen, M. H. and Sinha, D. (2001) Criterion-based methods for Bayesian model assessment. *Statistica Sinica*, **11**, 419-443.

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

Lee, S. Y. (2007) *Structural Equation Modeling: A Bayesian Approach*. UK: John Wiley & Sons, Ltd.

Lee, S. Y. and Song, X. Y. (2003) Bayesian model selection for mixtures of structural equation models with an unknown number of components. *British Journal of Mathematical and Statistical Psychology*, **56**, 145-165.

Meng, X. L. (1994) Posterior predictive p-values. *The Annals of Statistics*, **22**, 1142-1160.

Ogata, Y. (1989) A Monte Carlo method for high dimensional integration. *Numerische Mathematik*, **55**, 137-157.

O'Hagan, A. (1995) Fractional Bayes Factor for Model Comparison. *Journal of the Royal Statistical Society, Series B*, **57**, 99-138.

Raftery, A. E. (1993) Bayesian model selection in structural equation models. In K. A. Bollen and J. S. Long (eds), *Testing Structural Equation Models*, pp. 163-180. Beverly Hills, CA: Sage.

Rubin, D. B. (1984) Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, **12**, 1151-1172.

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

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

Spiegelhalter, D. J., Thomas, A., Best, N. G. and Lunn, D. (2003) *WinBUGS User Manual. Version 1.4*. Cambridge, UK: MRC Biostatistics Unit.

Song, X. Y. and Lee, S. Y. (2006) Model comparison of generalized linear mixed models. *Statistics in Medicine*, **25**, 1685-1698.

Sturtz, S., Ligges, U. and Gelman, A. (2005) R2WinBUGS: A package for running WinBUGS from R. *Journal of Statistical Software*, **12**, 1-16.

Tanner, M. A. and Wong, W. H. (1987) The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American statistical Association*, **82**, 528-550.

World Values Study Group (1994) World Values Survey, 1981-1984 and 1990-1993. ICPSR version. Ann Arbor, MI: Institute for Social Research (producer). Ann Arbor, MI: Inter-university Consortium for Political and Social Research (distributor).

Table 4.1: Interpretation of Bayes factor.

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

Table 4.2: Means and standard deviations of the estimated $\log B_{0k}$ in the simulation study.

	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)

Table 4.3: Maximum absolute differences of $\log B_{0k}$ under several different prior settings.

	$\log B_{01}$	$\log B_{02}$	$\log B_{03}$	$\log B_{04}$	$\log B_{05}$	$\log B_{60}$
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