Journal of Mathematical Psychology 56 (2012) 135–148

Contents lists available at SciVerse ScienceDirect

Journal of Mathematical Psychology

journal homepage: www.elsevier.com/locate/jmp

A tutorial on the Bayesian approach for analyzing structural equation models

Xin-Yuan Song, Sik-Yum Lee *

Department of Statistics, The Chinese University of Hong Kong, Shatin, NT, Hong Kong

ARTICLE INFO

Article history:
Received 31 January 2011
Received in revised form
31 January 2012
Available online 31 March 2012

Keywords:
Structural equation models
Bayesian analysis
Markov chain Monte Carlo (MCMC) methods

ABSTRACT

In this paper, we provide a tutorial exposition on the Bayesian approach in analyzing structural equation models (SEMs). SEMs, which can be regarded as regression models with observed and latent variables, have been widely applied to substantive research. However, the classical methods and most commercial software in this area are based on the covariance structure approach, which would encounter serious difficulties when dealing with complicated models and/or data structures. In contrast, the Bayesian approach has much more flexibility in handling complex situations. We give a brief introduction to SEMs and a detailed description of how to apply the Bayesian approach to this kind of model. Advantages of the Bayesian approach are discussed, and results obtained from a simulation study are provided for illustration. The intended audience is statisticians/methodologists who either know about SEMs or simple Bayesian statistics, and Ph.D. students in statistics, psychometrics, or mathematical psychology.

1. Introduction

In social and psychological research, it is common to encounter latent constructs that cannot be directly measured by a single observed variable. To assess the nature of a latent construct, a combination of observed variables is needed. For example, a number of related items should be combined to evaluate intelligence. In statistical inference, a latent construct is analyzed through a latent variable, which is appropriately defined by a combination of several observed variables.

Structural equation modeling is a powerful multivariate tool for studying interrelationships among observed and latent variables. This statistical model is very popular in behavioral, educational, and social research. The basic structural equation model (SEM), for example the widely used LISREL model (Jöreskong & Sörbom, 1996), consists of two components. The first component is a confirmatory factor analysis (CFA) model, which groups the highly correlated observed variables into latent variables and takes the measurement errors into account. This component can be regarded as a regression model with latent variables, in which the observed variables are regressed on a smaller number of latent variables. As the covariance matrix of the latent variables is allowed to be nondiagonal, the correlations/covariances of the latent variables can be evaluated. As various effects of explanatory latent variables on key outcome latent variables of interest cannot be assessed by the CFA model of the first component, a second component is needed. This component is again a regression-type model with latent variables, in which the dependent latent variables of main interest are regressed on other latent variables; see a concrete example in Section 2. As a result, the SEM is conceptually formulated by two familiar regression models. However, as the latent variables in the model are random, the standard technique in regression analysis cannot be applied to analyze SEMs.

In substantive research, it is often important to develop an appropriate model to evaluate the impacts of some explanatory observed and latent variables on the key outcome variables. Based on its particular formulation, the SEM is very useful for achieving the above objective. Furthermore, it is easy to appreciate the key ideas of the SEM. In fact, to apply the model to substantive research, one only needs to understand the basic concepts of latent variables and to be familiar with the regression model. As a result, this model has been extensively applied to behavioral, educational, psychological, and social research.

The covariance structure approach in analyzing SEMs lays emphasis on fitting the covariance structure of the proposed model to the sample covariance matrix computed from the observed data. For simple SEMs and when the underlying distribution of the observed data is normal, this classical approach works fine for datasets with reasonably large sample sizes. However, it will encounter serious difficulties for many complicated situations: for example, when deriving the covariance structure is difficult, or the data structures are complex.

Recently, the Bayesian approach has been applied to analyze many advanced SEMs that are useful for practical medical and socio-psychological research. The Bayesian approach, whose statistical development is based on raw observations rather

* Corresponding author.
E-mail address: sylee@sta.cuhk.edu.hk (S.-Y. Lee).

0022-2496/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmp.2012.02.001

than the sample covariance matrix, has several advantages. For example, the Bayesian approach allows the use of genuine prior information in addition to the information that is available in observed data for producing results; it provides better statistics for goodness-of-fit and model comparison, and also other useful statistics such as the mean and percentiles of the posterior distribution; and it can give more reliable results for small samples (see Dunson (2000) and Lee and Song (2004)). Hence, the Bayesian approach is rather popular in analyzing SEMs. For instance, this approach has been applied to analyze (i) SEMs with mixed continuous and discrete variables (Dunson & Herring, 2005; Song, Lee et al., 2009; van Onna, 2002), (ii) nonlinear SEMs (Arminger & Muthen, 1998; Lee & Song, 2003), (iii) multilevel SEMs (Ansari & Jedidi, 2000; Song & Lee, 2004), (iv) semiparametric SEMs (Song, Xia, & Lee, 2009; Yang & Dunson, 2010), (v) SEMs with missing data (Song & Lee, 2002), and (vi) longitudinal SEMs (Dunson, 2003; Song, Lu, Hser, & Lee, 2011), among others. Moreover, the Bayesian SEM has actually been applied to substantive real research in various disciplines: see, for example, Jiang and Mahadevan (2009) (engineering); Ansari, Jedidi, and Jagpal (2000) (marketing research); Wu, Heringstad, and Gianola (2010) (genetics), Arhonditsis et al. (2006) (ecology), and Song, Lee et al. (2009) (diabetic studies).

The main objective of this article is to provide a brief introduction to some basic SEMs and to the Bayesian approach in analyzing these models. Bayesian estimation and model comparison, together with the freely available software WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000), will be discussed.

2. Basic concepts of SEMs

Let $\mathbf{y} = (y_1, \ldots, y_p)^T$ be a p by 1 vector of observed variables that have been selected for the analysis, and let $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_q)^T$ be a q by 1 vector of latent variables that are related to the observed variables in $\mathbf{y}$. The link between the observed variables and all the latent variables in $\boldsymbol{\omega}$ is defined by the following *measurement equation*:

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

where $\boldsymbol{\Lambda}$ is a p by q matrix of unknown factor loadings, and $\boldsymbol{\epsilon}$ is a p by 1 random vector of measurement (residual) errors. Here, $\boldsymbol{\epsilon}$ is distributed as $N[0, \boldsymbol{\Psi}_\epsilon]$, where $\boldsymbol{\Psi}_\epsilon$ is a diagonal matrix. Let $\boldsymbol{\omega} = (\boldsymbol{\eta}^T, \boldsymbol{\xi}^T)^T$, where $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$ are q_1 by 1 and $q_2 (= q - q_1)$ by 1 random vectors which respectively contain the outcome and explanatory latent variables in $\boldsymbol{\omega}$. The effects of $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_{q_2})^T$ on $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_{q_1})^T$ are assessed by the following *structural equation*:

$$\boldsymbol{\eta} = \boldsymbol{\Pi}\boldsymbol{\eta} + \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\delta}, \tag{2}$$

where $\boldsymbol{\Pi}$ and $\boldsymbol{\Gamma}$ are matrices of unknown coefficients, and $\boldsymbol{\delta}$ is a residual vector which is distributed as $N[0, \boldsymbol{\Psi}_\delta]$. It is assumed that $\boldsymbol{\xi}$ is distributed as $N[0, \boldsymbol{\Phi}]$, and that $\boldsymbol{\epsilon}$ and $\boldsymbol{\delta}$ are independent. Moreover, it is assumed that $|\mathbf{I} - \boldsymbol{\Pi}|$ is a positive constant which does not involve elements in $\boldsymbol{\Pi}$. Eqs. (1) and (2) define the most basic linear SEM. See Figs. 1 and 3 for path diagrams associated with some SEMs. Moreover, a nonlinear SEM is presented at the end of Section 3.3, while other generalizations (such as an SEM with structured mean, a mixture SEM, and a multilevel SEM) can be found in Lee (2007).

Like applications of other statistical models, applications of SEMs rely on substantive knowledge to build a model. It is a confirmatory tool rather than an exploratory tool. Practically, we usually have a clear objective of the study, and some basic background about the key structure of the model that is obtained either from subject knowledge or from preliminary data analysis. This basic background will be used in both the specification and the interpretation of the model.

Eqs. (1) and (2) of a specified SEM decide the key numbers q_1 and q_2. In real applications, the basic background of the p manifest variables in $\mathbf{y}$ usually give good choices of q_1 and q_2. To give a simple illustrative example in medical research, suppose that we are interested in studying the effects of blood pressure and obesity on kidney failure, which is assessed by the urinary albumin creatinine ratio (ACR) and plasma creatinine (PCr). In addition to ACR and PCr, suppose that the other observed variables are systolic blood pressure (SBP), diastolic blood pressure (DBP), body mass index (BMI), hip index (HIP), and waist index (WST). From medical knowledge of these variables, the following are clear. (i) The observed variables ACR and PCr provide key information about kidney disease severity; hence, they are grouped together to form the independent latent variable that can be interpreted as 'kidney disease severity, η'. (ii) {SBP, DBP} are taken as the observed variables to form the latent variable that can be interpreted as 'blood pressure, ξ_1'. (iii) Similarly, the observed variables {BMI, HIP, WST} are grouped together to form the latent variable that can be interpreted as 'obesity, ξ_2'. Hence, it is natural to group the seven observed variables into three latent variables, and take $q_1 = 1$, and $q_2 = 2$. Let

$$\boldsymbol{\Lambda}^T = \begin{bmatrix} 1.0^* & \lambda_{21} & 0^* & 0^* & 0^* & 0^* & 0^* \\ 0^* & 0^* & 1.0^* & \lambda_{42} & 0^* & 0^* & 0^* \\ 0^* & 0^* & 0^* & 0^* & 1.0^* & \lambda_{63} & \lambda_{73} \end{bmatrix}, \tag{3}$$

where parameters with asterisks are fixed at the preassigned values. According to common practice in factor analysis, the fixed value of 1.0 is used to specify the 'scale' of the unknown parameters (factor loadings) in that column. The structure of the factor loading matrix $\boldsymbol{\Lambda}$ reflects the fact that SBP and DBP are clear indicators of 'blood pressure, ξ_1', whilst BMI, HIP, and WST are not; and BMI, HIP, and WST are clear indicators of 'obesity, ξ_2', whilst SBP and DBP are not. Obviously, ACR and PCr are the observed variables related to η. Hence the measurement equation is defined by $\mathbf{y} = \boldsymbol{\Lambda}\boldsymbol{\omega} + \boldsymbol{\epsilon}$, where $\boldsymbol{\Lambda}$ is given by (3), $\boldsymbol{\omega} = (\eta_1, \xi_1, \xi_2)^T$, and $\boldsymbol{\epsilon}$ is a 7 by 1 random vector of measurement errors. The latent variables are related through the following structural equation:

$$\eta = \gamma_1\xi_1 + \gamma_2\xi_2 + \delta; \tag{4}$$

here, $\boldsymbol{\Pi} = \mathbf{0}$, and $\boldsymbol{\Gamma} = (\gamma_1, \gamma_2)$. The covariance matrices of $(\xi_1, \xi_2)^T$ and $(\epsilon_{11}, \epsilon_{12}, \epsilon_{13}, \epsilon_{14}, \epsilon_{15}, \epsilon_{16}, \epsilon_{17})^T$ are respectively equal to

$$\boldsymbol{\Phi} = \begin{bmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{22} \end{bmatrix}, \quad \text{and}$$

$$\boldsymbol{\Psi}_\epsilon = \text{diag}(\psi_{\epsilon 1}, \psi_{\epsilon 2}, \psi_{\epsilon 3}, \psi_{\epsilon 4}, \psi_{\epsilon 5}, \psi_{\epsilon 6}, \psi_{\epsilon 7}).$$

As SEMs can be regarded as regression models with latent variables, the interpretation of the unknown parameters in $\boldsymbol{\Lambda}$, γ_1, and γ_2, is the same as the interpretation of the regression coefficients in a regression model. In the above example, the effects of blood pressure (ξ_1) and obesity (ξ_2) on kidney disease severity (η) are assessed through γ_1 and γ_2. The path diagram associated with the above SEM is presented in Fig. 1.

3. A Bayesian approach in estimating SEMs

The basic objective of this section is to introduce a Bayesian approach for analyzing SEMs. In contrast to the existing covariance structure analysis approach, we focus on the use of raw observations rather than the sample covariance matrix. The following general strategy will be used to solve the difficulties that are induced by the complexity of the model and the data. First, we treat the latent variables in the model as missing data, and then we analyze

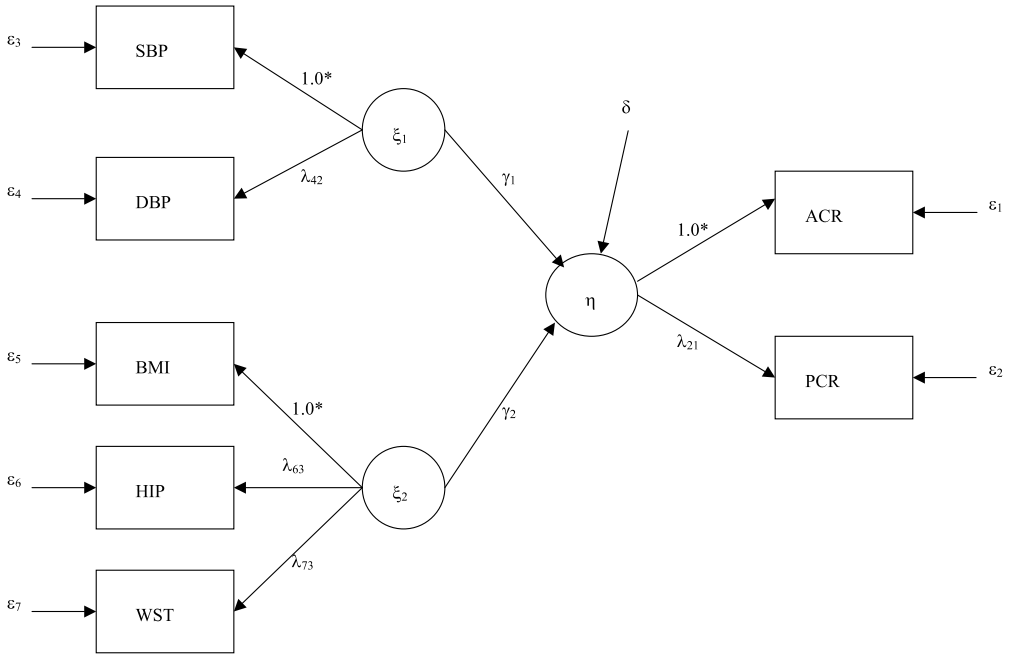

Fig. 1. Path diagram associated with the SEM. In this figure and Fig. 3, observed variables are represented by rectangles, and latent variables are represented by circles (or ellipses).

the model on the basis of the complete dataset that contains the observed data and the missing data by applying some Markov chain Monte Carlo (MCMC) tools (see Gelman, Carlin, Stern, and Rubin (1995)) in statistical computing. As the complete dataset is much easier to handle than the observed dataset, the difficulties that are induced by the complexity of the latent variables are alleviated.

The basic nice feature of a Bayesian approach is its flexibility to utilize useful prior information in conducting statistical inference. In many practical problems, statisticians may have good prior information from some sources, such as the knowledge of experts, and analyses of similar data and/or past data. For example, in an organization and management research activity that involves latent variables about job performance and job satisfaction, we may have some prior information about the correlation of these latent variables, say a relatively large value that exceeds 0.4; we may also have some prior information on the values of factor loadings, say a relatively large loading that corresponds to the observed variable 'salary' and the latent factor 'job satisfaction'. As pointed out by many important articles in Bayesian analyses of SEMs (Dunson, 2000; Lee & Song, 2004), another nice feature of sampling-based Bayesian methods is that they depend less on asymptotic theory, and hence have the potential to produce reliable results even with small samples.

3.1. Basic concepts in Bayesian estimation

The Bayesian approach is well recognized in the statistics literature as an attractive approach to analyze a wide variety of models (Berger, 1985; Box & Tiao, 1973; Congdon, 2003; Zellner, 1971). To introduce this approach in general, we let M be an arbitrary statistical model with a vector of unknown parameters $\boldsymbol{\theta}$, and let $\mathbf{Y}$ be the observed dataset of raw observations with a sample size n. In a Bayesian approach, $\boldsymbol{\theta}$ is considered to be random with a distribution (called the prior distribution) and an associated (prior) density function, say $p(\boldsymbol{\theta})$. Let $p(\mathbf{Y}, \boldsymbol{\theta}|M)$ be the probability density function of the joint distribution of $\mathbf{Y}$ and $\boldsymbol{\theta}$ under M. The behavior of $\boldsymbol{\theta}$ under the given data $\mathbf{Y}$ is fully described by the posterior distribution of $\boldsymbol{\theta}$. Let $p(\boldsymbol{\theta}|\mathbf{Y}, M)$ be the density function of the posterior distribution, which is called the posterior density function. The posterior distribution of $\boldsymbol{\theta}$ or its density plays the most important role in the Bayesian analysis of the model. Based on a well-known identity in probability, we have $p(\mathbf{Y}, \boldsymbol{\theta}|M) = p(\mathbf{Y}|\boldsymbol{\theta}, M)p(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\mathbf{Y}, M)p(\mathbf{Y}|M)$. As $p(\mathbf{Y}|M)$ does not depend on $\boldsymbol{\theta}$, and can be regarded as a constant with fixed $\mathbf{Y}$, we have

$$\log p(\boldsymbol{\theta}|\mathbf{Y}, M) \propto \log p(\mathbf{Y}|\boldsymbol{\theta}, M) + \log p(\boldsymbol{\theta}). \tag{5}$$

Note that $p(\mathbf{Y}|\boldsymbol{\theta}, M)$ is the likelihood function. Hence, the posterior density function incorporates the sample information and the prior information through the likelihood function $p(\mathbf{Y}|\boldsymbol{\theta}, M)$ and the prior density function $p(\boldsymbol{\theta})$. Note also that $p(\mathbf{Y}|\boldsymbol{\theta}, M)$ depends on the sample size, whereas $p(\boldsymbol{\theta})$ does not. In large samples, the prior distribution of $\boldsymbol{\theta}$ plays a less important role, and the posterior density function $\log p(\boldsymbol{\theta}|\mathbf{Y}, M)$ is close to the log-likelihood function $\log p(\mathbf{Y}|\boldsymbol{\theta}, M)$. Hence, Bayesian and maximum likelihood (ML) approaches are asymptotically equivalent, and the Baysian estimates have the same optimal properties as the ML estimates. When the sample sizes are small or moderate, the prior distribution of $\boldsymbol{\theta}$ plays a significant role in the Bayesian approach. For substantive research problems in which the sample sizes are small or moderate, prior information about the parameter vector $\boldsymbol{\theta}$ can be incorporated into the Bayesian analysis through the prior distribution of $\boldsymbol{\theta}$ in order to achieve better results (see below for the utilization of useful prior information in the analysis). For many practical problems, researchers may have good prior information from experts, from analyses of similar or past data, or from some other sources. More accurate results can be achieved by incorporating the appropriate prior information in the analysis through the prior distribution of $\boldsymbol{\theta}$. When applying the Bayesian approach for analyzing SEMs, M will represent an arbitrary SEM. In the following, the symbol M will be suppressed.

The prior distribution of $\boldsymbol{\theta}$ represents the distribution of possible parameter values, from which the parameter $\boldsymbol{\theta}$ has been drawn. Basically, there are two kinds of prior distributions, namely noninformative prior distributions and informative prior distributions. Noninformative prior distributions are used when we have little prior information, and hence the prior distributions play a minimal role in the posterior distribution. The associated

prior density is regarded as vague, diffuse, flat or noninformative, for example a density that is proportional to a constant or has an extremely large variance. It can be seen from (5) that the Bayesian approach and the ML approach are basically equivalent if the priors are noninformative. This flexibility can be regarded as an advantage of the Bayesian approach. For an informative prior distribution, we may have good prior knowledge about this distribution, either from closely related data or from subjective knowledge of experts. Usually, an informative prior distribution has its own parameters, which are called hyperparameters. Theoretically, Bayesian methods can be developed with general priors. A commonly used informative prior distribution in the general Bayesian approach to statistical problems is the conjugate prior distribution, in which the posterior distribution follows the same parametric form as the prior distribution (see Gelman et al. (1995)). To give a simple example, we consider the following binomial model with parameter θ. The likelihood function of an observation y is given by

$$p(y|\theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}.$$

If the prior density of θ is given by $p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$, which is a beta distribution with hyperparameters α and β, then the posterior distribution of θ is given by

$$\begin{aligned} p(\theta|y) &\propto p(y|\theta)p(\theta) \\ &\propto \theta^y(1-\theta)^{n-y}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &= \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1}. \end{aligned}$$

Thus, $p(\theta|y)$ and $p(\theta)$ are of the same form, and $p(\theta)$ is a conjugate prior of θ.

The Bayesian estimate of $\boldsymbol{\theta}$ is usually defined as the mean of $p(\boldsymbol{\theta}|\mathbf{Y})$. Theoretically, it could be obtained via integration. For most situations, the integration does not have a closed form. However, if we can simulate a sufficiently large number of draws from $p(\boldsymbol{\theta}|\mathbf{Y})$ by means of some efficient algorithm in statistical computing (see Gelman et al. (1995)), we can approximate the mean, and/or other useful statistics, through the simulated observations. Hence, it is important to develop efficient methods for drawing observations from the posterior distribution (or the posterior density).

3.2. Bayesian estimation of an SEM

In applying the Bayesian approach to analyze SEMs, we consider the model M be an arbitrary SEM with latent variables. The posterior distributions of parameters and latent variables can be estimated by using a sufficiently large number of observations drawn from the posterior distribution of unknown parameters through efficient MCMC methods. Means as well as quantiles of this posterior distribution can be estimated from the simulated observations. These quantities are useful in making statistical inferences. For example, the Bayesian estimates of unknown parameters and latent variables can be obtained from the corresponding sample means of the posterior distribution. However, for most SEMs, the posterior distribution $p(\boldsymbol{\theta}|\mathbf{Y})$ is complicated. It is difficult to derive this distribution and simulate observations from it. A major breakthrough for posterior simulation is the idea of data augmentation that is proposed by Tanner and Wong (1987). The strategy is to treat latent quantities as hypothetical missing data and augment the observed data with them so that the posterior distribution (or the posterior density) based on the complete dataset is relatively easy to handle. This is particularly useful for SEMs that involve latent variables. The feature that makes SEMs different from common regression models is the existence of latent variables, and this is what causes difficulties in analyzing the model. However, if the latent variables are observed, the SEM will become the familiar regression model or simultaneous equation model that can be handled without much difficulty. More specifically, instead of working on the intractable posterior density $p(\boldsymbol{\theta}|\mathbf{Y})$, we will work on $p(\boldsymbol{\theta}, \boldsymbol{\Omega}|\mathbf{Y})$, where $\boldsymbol{\Omega}$ is the set of latent variables in the model. For most cases, $p(\boldsymbol{\theta}, \boldsymbol{\Omega}|\mathbf{Y})$ is still not in closed form, and it is difficult to deal with directly. However, based on the complete dataset $(\boldsymbol{\Omega}, \mathbf{Y})$, the conditional distribution $[\boldsymbol{\theta}|\boldsymbol{\Omega}, \mathbf{Y}]$ is usually standard; moreover, the conditional distribution $[\boldsymbol{\Omega}|\boldsymbol{\theta}, \mathbf{Y}]$ can be derived from the definition of the model without much difficulty. Consequently, we can apply some MCMC methods to simulate observations from $p(\boldsymbol{\theta}, \boldsymbol{\Omega}|\mathbf{Y})$ by drawing observations iteratively from their full conditional densities $p(\boldsymbol{\theta}|\boldsymbol{\Omega}, \mathbf{Y})$ and $p(\boldsymbol{\Omega}|\boldsymbol{\theta}, \mathbf{Y})$. A useful algorithm to do this is the following Gibbs sampler (Geman & Geman, 1984).

In the model M, suppose that the parameter vector $\boldsymbol{\theta}$ and the latent matrix $\boldsymbol{\Omega}$ are respectively decomposed into the following components or subvectors: $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_a)$ and $\boldsymbol{\Omega} = (\boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_b)$. The Gibbs sampler is a Markov chain algorithm which performs an alternating conditional sampling at each of its iteration. It cycles through the components of $\boldsymbol{\theta}$ and $\boldsymbol{\Omega}$, drawing each component conditional on the values of all the other components. More specifically, at the j-th iteration with current values $\boldsymbol{\theta}^{(j)} = (\boldsymbol{\theta}_1^{(j)}, \ldots, \boldsymbol{\theta}_a^{(j)})$ and $\boldsymbol{\Omega}^{(j)} = (\boldsymbol{\Omega}_1^{(j)}, \ldots, \boldsymbol{\Omega}_b^{(j)})$, it simulates, in turn,

$$\begin{aligned} &\boldsymbol{\theta}_1^{(j+1)} \quad \text{from } p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2^{(j)}, \ldots, \boldsymbol{\theta}_a^{(j)}, \boldsymbol{\Omega}^{(j)}, \mathbf{Y}), \\ &\boldsymbol{\theta}_2^{(j+1)} \quad \text{from } p(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1^{(j+1)}, \ldots, \boldsymbol{\theta}_a^{(j)}, \boldsymbol{\Omega}^{(j)}, \mathbf{Y}), \\ &\vdots \\ &\boldsymbol{\theta}_a^{(j+1)} \quad \text{from } p(\boldsymbol{\theta}_a|\boldsymbol{\theta}_1^{(j+1)}, \ldots, \boldsymbol{\theta}_{a-1}^{(j+1)}, \boldsymbol{\Omega}^{(j)}, \mathbf{Y}), \\ &\boldsymbol{\Omega}_1^{(j+1)} \quad \text{from } p(\boldsymbol{\Omega}_1|\boldsymbol{\theta}^{(j+1)}, \boldsymbol{\Omega}_2^{(j)}, \ldots, \boldsymbol{\Omega}_b^{(j)}, \mathbf{Y}), \\ &\boldsymbol{\Omega}_2^{(j+1)} \quad \text{from } p(\boldsymbol{\Omega}_2|\boldsymbol{\theta}^{(j+1)}, \boldsymbol{\Omega}_1^{(j+1)}, \ldots, \boldsymbol{\Omega}_b^{(j)}, \mathbf{Y}), \\ &\vdots \\ &\boldsymbol{\Omega}_b^{(j+1)} \quad \text{from } p(\boldsymbol{\Omega}_b|\boldsymbol{\theta}^{(j+1)}, \boldsymbol{\Omega}_1^{(j+1)}, \ldots, \boldsymbol{\Omega}_{b-1}^{(j+1)}, \mathbf{Y}). \end{aligned} \tag{6}$$

There are $a + b$ steps in the j-th iteration of the Gibbs sampler. At each step, each component in $\boldsymbol{\theta}$ and $\boldsymbol{\Omega}$ is updated conditional on the latest values of the other components. One may simulate the components in $\boldsymbol{\Omega}$ first, then the components in $\boldsymbol{\theta}$; or vice versa. Most of the full conditional distributions in (6) are the standard normal, gamma, or Wishart distributions. Simulating observations from them is rather straightforward. For nonstandard conditional distributions, the Metropolis–Hastings (MH) algorithm (Hastings, 1970; Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953) may have to be used for efficient simulation.

It has been shown (Geman & Geman, 1984) that, under mild regularity conditions, the joint distribution of $(\boldsymbol{\theta}^{(j)}, \boldsymbol{\Omega}^{(j)})$ converges to the desired posterior distribution $[\boldsymbol{\theta}, \boldsymbol{\Omega}|\mathbf{Y}]$, after a sufficiently large number of iterations (usually called burn-in phase), say J. Observations obtained at the early iterations should be discarded, because they still do not belong to the target distribution. The required number of iterations for achieving convergence of the Gibbs sampler can be determined by plots of the simulated sequences of the individual parameters. At convergence, parallel sequences generated with different starting values should mix well together. An example of sequences that have no reached convergence is presented in Fig. 2, while some examples of sequences for which convergence looks reasonable are given in Fig. 4 in relation to the analysis of a simulated dataset presented in Section 4. An alternative method for monitoring convergence is based on the 'estimated potential scale reduction (EPSR, $\hat{R}$) values'. The computation of these values is presented in Appendix A (see Lee Appendix 4.2). A minor problem with iterative simulation

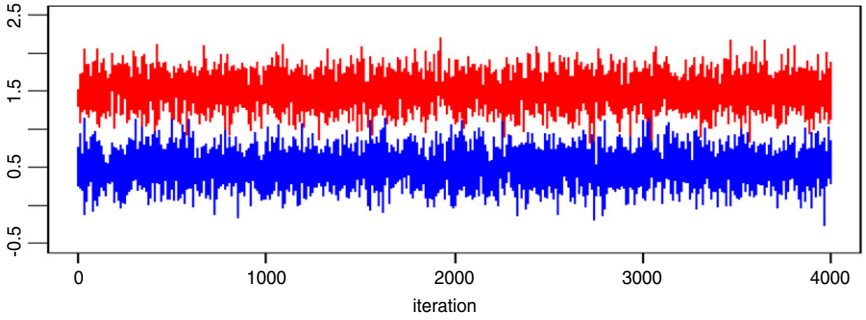

Fig. 2. MCMC chains which have not reached convergence.

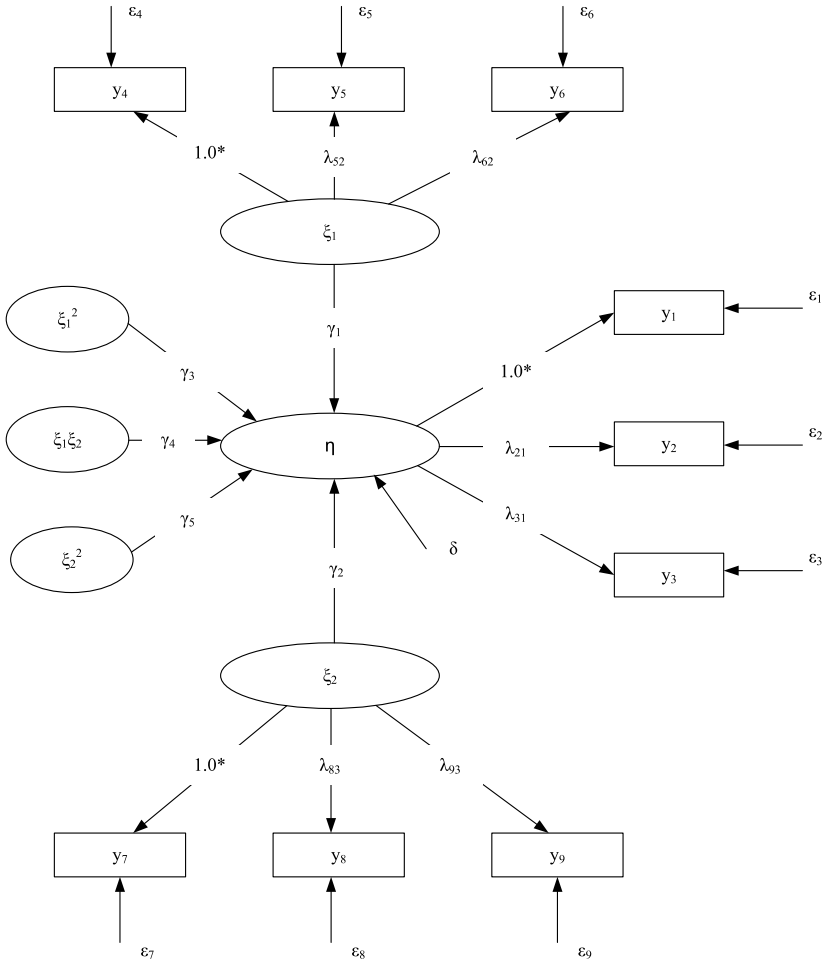

Fig. 3. Path diagram associated with the nonlinear SEM in the simulation study.

draws is their within-sequence correlation. To reduce the serial correlation between consecutive observations, samples may be collected in cycles with indices $J + c, J + 2c, \ldots, J + Tc$ for some spacing c (see Gelfand and Smith (1990)). However, in most practical applications, a small c will suffice for many statistical analyses such as obtaining estimates of unknown parameters and their standard error estimates (see Albert and Chib (1993) and Zeger and Karim (1991)). In the illustrative example presented in Section 4, we use $c = 1$.

Statistical inference of the model can then be conducted on the basis of simulated observations from $p(\boldsymbol{\theta}, \boldsymbol{\Omega}|\mathbf{Y})$, namely, $\{(\boldsymbol{\theta}^{(k)}, \boldsymbol{\Omega}^{(k)}): k = 1, \ldots, K\}$. The Bayesian estimate of $\boldsymbol{\theta}$ as well as the standard error estimate can be obtained from

$$\hat{\boldsymbol{\theta}} = K^{-1} \sum_{k=1}^{K} \boldsymbol{\theta}^{(k)}, \tag{7}$$

$$\widehat{\mathrm{Var}}(\boldsymbol{\theta}|\mathbf{Y}) = (K-1)^{-1} \sum_{k=1}^{K} (\boldsymbol{\theta}^{(k)} - \hat{\boldsymbol{\theta}})(\boldsymbol{\theta}^{(k)} - \hat{\boldsymbol{\theta}})^T. \tag{8}$$

More statistical inference on $\boldsymbol{\theta}$ can be carried out based on the simulated sample, $\{\boldsymbol{\theta}^{(k)}: k = 1, \ldots, K\}$. For instance, the 2.5%, 50%, and 97.5% quantiles of the sampled distribution of an individual parameter give a 95% posterior interval and convey skewness in its marginal posterior density. The total number

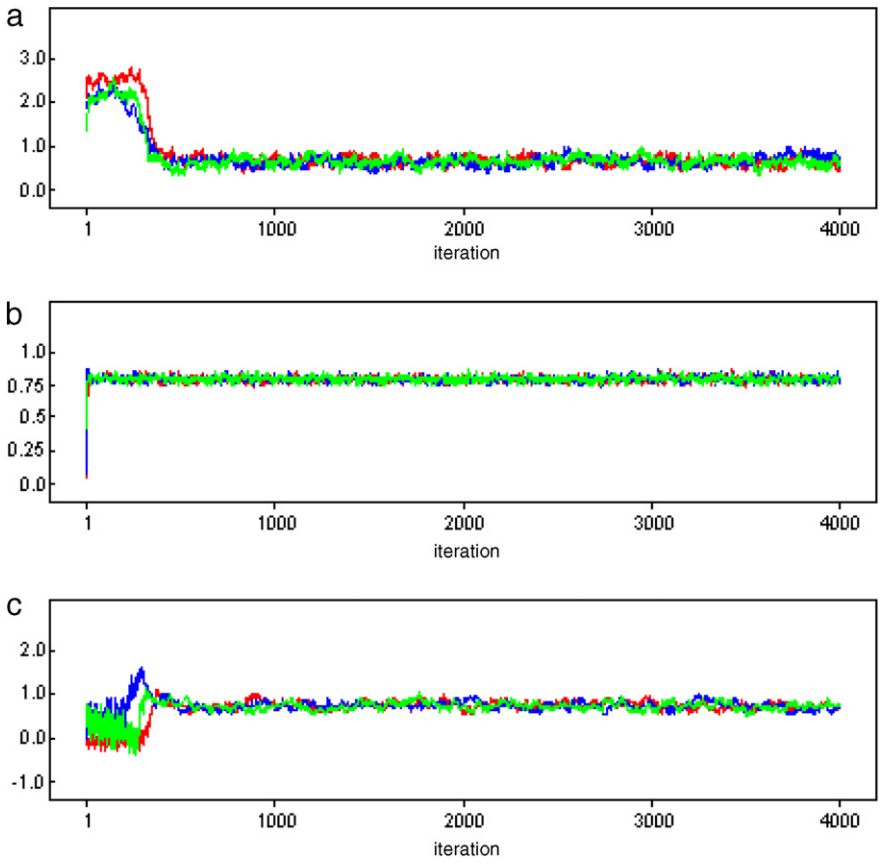

Fig. 4. (a)–(c) are plots of parallel sequences corresponding to different starting values of μ_1, λ_{21}, and γ_3.

of draws, K, that is required for statistical analysis depends on the form of the posterior distribution. For most SEMs, 1000 draws after convergence are sufficient. Clearly, different choices of K would produce similar estimates, but they are not exactly equal.

For any individual $\mathbf{y}_i$, let $\boldsymbol{\omega}_i$ be the vector of latent variables, and let $E(\boldsymbol{\omega}_i|\mathbf{y}_i)$ and $\mathrm{Var}(\boldsymbol{\omega}_i|\mathbf{y}_i)$ be the posterior mean and the posterior covariance matrix. A Bayesian estimate $\hat{\boldsymbol{\omega}}_i$ can be obtained through $\{\boldsymbol{\Omega}^{(k)}, k = 1, \ldots, K\}$ as follows:

$$\hat{\boldsymbol{\omega}}_i = K^{-1} \sum_{k=1}^{K} \boldsymbol{\omega}_i^{(k)} = \widehat{E}(\boldsymbol{\omega}_i \mid \mathbf{y}_i), \tag{9}$$

where $\boldsymbol{\omega}_i^{(k)}$ is the i-th column of $\boldsymbol{\Omega}_i^{(k)}$. This gives a direct Bayesian estimate which is not expressed in terms of the structural parameter estimates. A consistent estimate of $\mathrm{Var}(\boldsymbol{\omega}_i|\mathbf{y}_i)$ can be similarly obtained as in (8).

To apply the Bayesian approach in estimating the unknown parameter vector $\boldsymbol{\theta}$ and the latent variable in $\boldsymbol{\Omega}$ in a specified SEM, we need to derive the components in the full conditional distributions $p(\boldsymbol{\theta}|\boldsymbol{\Omega}, \mathbf{Y})$ and $p(\boldsymbol{\Omega}|\boldsymbol{\theta}, \mathbf{Y})$, and then construct a computer program based on some MCMC methods such as the Gibbs sampler, to simulate observations for statistical inference. Those who do not wish to go through the above tasks can use the software WinBUGS (Lunn et al., 2000), which is useful for producing reliable Bayesian statistics for a very wide range of statistical models. The algorithm used in WinBUGS is mainly developed using MCMC techniques, such as the Gibbs sampler (Geman & Geman, 1984), and the MH algorithm (Hastings, 1970; Metropolis et al., 1953). It has been shown that, under broad conditions, this software can provide simulated samples from the joint posterior distribution of the unknown quantities, such as parameters and latent variables in the model. Bayesian estimates of the unknown parameters and their standard error estimates in the model can be obtained from these samples for conducting statistical inferences. The advanced version of the program is WinBUGS 1.4 running under Windows, which is developed by the Medical Research Council (MRC) Biostatistics Unit (Cambridge, UK) and the Department of Epidemiology and Public Health of the Imperial College School of Medicine at St. Mary's Hospital (London).

3.3. Bayesian estimation of the confirmatory factor analysis (CFA) model

To illustrate the Bayesian estimation and the associative MCMC method described above, we present a detailed application in the context of the CFA model, which is essentially the measurement equation of an SEM. In the next section, we show how the Bayesian approach can be extended naturally to an SEM which also involves a structural equation. Consider the following CFA model: for $i = 1, \ldots, n$,

$$\mathbf{y}_i = \boldsymbol{\Lambda}\boldsymbol{\omega}_i + \boldsymbol{\epsilon}_i, \tag{10}$$

where $\mathbf{y}_i$, $\boldsymbol{\Lambda}$, and $\boldsymbol{\epsilon}_i$ are defined as in (1). Let $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$ be the observed data matrix, $\boldsymbol{\Omega} = (\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_n)$ be the matrix of latent factor scores, and $\boldsymbol{\theta}$ be the parameter vector that contains the unknown elements of $\boldsymbol{\Lambda}$, $\boldsymbol{\Phi}$, and $\boldsymbol{\Psi}_\epsilon$ in the model. In the Bayesian analysis, we will treat the latent factor scores in $\boldsymbol{\Omega}$ as hypothetical missing data, and augment the observed dataset $\mathbf{Y}$ with $\boldsymbol{\Omega}$ in the posterior analysis. A sufficiently large sample of $(\boldsymbol{\theta}, \boldsymbol{\Omega})$ from the joint posterior distribution $[\boldsymbol{\theta}, \boldsymbol{\Omega}|\mathbf{Y}]$ is generated by the following Gibbs sampler algorithm. At the $(j+1)$-th iteration with current

values of $\boldsymbol{\Omega}^{(j)}$, $\boldsymbol{\Psi}_\epsilon^{(j)}$, $\boldsymbol{\Lambda}^{(j)}$, and $\boldsymbol{\Phi}^{(j)}$:

$$\begin{aligned}&\text{(i) Generate } \boldsymbol{\Omega}^{(j+1)} \quad \text{from } p(\boldsymbol{\Omega}|\boldsymbol{\Psi}_\epsilon^{(j)}, \boldsymbol{\Lambda}^{(j)}, \boldsymbol{\Phi}^{(j)}, \mathbf{Y}).\\ &\text{(ii) Generate } \boldsymbol{\Psi}_\epsilon^{(j+1)} \quad \text{from } p(\boldsymbol{\Psi}_\epsilon|\boldsymbol{\Omega}^{(j+1)}, \boldsymbol{\Lambda}^{(j)}, \boldsymbol{\Phi}^{(j)}, \mathbf{Y}).\\ &\text{(iii) Generate } \boldsymbol{\Lambda}^{(j+1)} \quad \text{from } p(\boldsymbol{\Lambda}|\boldsymbol{\Omega}^{(j+1)}, \boldsymbol{\Psi}_\epsilon^{(j+1)}, \boldsymbol{\Phi}^{(j)}, \mathbf{Y}).\\ &\text{(iv) Generate } \boldsymbol{\Phi}^{(j+1)} \quad \text{from } p(\boldsymbol{\Phi}|\boldsymbol{\Omega}^{(j+1)}, \boldsymbol{\Psi}_\epsilon^{(j+1)}, \boldsymbol{\Lambda}^{(j+1)}, \mathbf{Y}).\end{aligned} \tag{11}$$

The conditional distributions in (11) are required in the implementation of the Gibbs sampler.

The derivation of $p(\boldsymbol{\Omega}|\boldsymbol{\Psi}_\epsilon, \boldsymbol{\Lambda}, \boldsymbol{\Phi}, \mathbf{Y}) = p(\boldsymbol{\Omega}|\boldsymbol{\theta}, \mathbf{Y})$ is based on the fact that, for $i = 1, \ldots, n$, $\boldsymbol{\omega}_i$ are mutually independent, and $\mathbf{y}_i$ are also mutually independent, given $(\boldsymbol{\omega}_i, \boldsymbol{\theta})$. Hence, we have

$$p(\boldsymbol{\Omega}|\mathbf{Y}, \boldsymbol{\theta}) = \prod_{i=1}^{n} p(\boldsymbol{\omega}_i|\mathbf{y}_i, \boldsymbol{\theta}) \propto \prod_{i=1}^{n} p(\boldsymbol{\omega}_i|\boldsymbol{\theta})\, p(\mathbf{y}_i|\boldsymbol{\omega}_i, \boldsymbol{\theta}). \tag{12}$$

Moreover, since the conditional distributions of $\boldsymbol{\omega}_i$, given $\boldsymbol{\theta}$, and $\mathbf{y}_i$, given $(\boldsymbol{\omega}_i, \boldsymbol{\theta})$, are $N(\mathbf{0}, \boldsymbol{\Phi})$ and $N(\boldsymbol{\Lambda}\boldsymbol{\omega}_i, \boldsymbol{\Psi}_\epsilon)$, respectively, it can be shown (see Lindley and Smith (1972, pp. 4–5)) that the conditional distribution of $\boldsymbol{\omega}_i$, given $(\mathbf{y}_i, \boldsymbol{\theta})$, is

$$\begin{aligned}[\boldsymbol{\omega}_i|\mathbf{y}_i, \boldsymbol{\theta}] \overset{D}{=} N\big[&(\boldsymbol{\Phi}^{-1} + \boldsymbol{\Lambda}^T\boldsymbol{\Psi}_\epsilon^{-1}\boldsymbol{\Lambda})^{-1}\boldsymbol{\Lambda}^T\\ &\boldsymbol{\Psi}_\epsilon^{-1}\mathbf{y}_i, (\boldsymbol{\Phi}^{-1} + \boldsymbol{\Lambda}^T\boldsymbol{\Psi}_\epsilon^{-1}\boldsymbol{\Lambda})^{-1}\big].\end{aligned} \tag{13}$$

Hence, the conditional distribution of $\boldsymbol{\Omega}$, given $(\mathbf{Y}, \boldsymbol{\theta})$, can be obtained from (12) and (13).

The conditional distribution of $\boldsymbol{\theta}$, given $(\mathbf{Y}, \boldsymbol{\Omega})$, is proportional to $p(\boldsymbol{\theta})p(\mathbf{Y}, \boldsymbol{\Omega}|\boldsymbol{\theta})$. Hence, it is necessary to select the prior density function $p(\boldsymbol{\theta})$ that represents the prior information of $\boldsymbol{\theta}$. Based on the property of $(\boldsymbol{\Lambda}, \boldsymbol{\Psi}_\epsilon)$ and $\boldsymbol{\Phi}$ in a factor analysis model, we specify the prior distribution of $\boldsymbol{\theta}$ as follows: $p(\boldsymbol{\theta}) = p(\boldsymbol{\Lambda}, \boldsymbol{\Phi}, \boldsymbol{\Psi}_\epsilon) = p(\boldsymbol{\Lambda}, \boldsymbol{\Psi}_\epsilon)\, p(\boldsymbol{\Phi})$. Moreover, the conditional distribution of $\mathbf{Y}$, given $\boldsymbol{\Omega}$, only depends on $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}_\epsilon$, and the distribution of $\boldsymbol{\Omega}$ only involves $\boldsymbol{\Phi}$. Hence,

$$\begin{aligned}p(\boldsymbol{\Lambda}, \boldsymbol{\Psi}_\epsilon, \boldsymbol{\Phi}|\mathbf{Y}, \boldsymbol{\Omega}) &= p(\boldsymbol{\theta}|\mathbf{Y}, \boldsymbol{\Omega}) \propto p(\mathbf{Y}, \boldsymbol{\Omega}|\boldsymbol{\theta})p(\boldsymbol{\theta})\\ &= p(\mathbf{Y}|\boldsymbol{\theta}, \boldsymbol{\Omega})p(\boldsymbol{\Omega}|\boldsymbol{\theta})p(\boldsymbol{\theta})\\ &= p(\mathbf{Y}|\boldsymbol{\theta}, \boldsymbol{\Omega})p(\boldsymbol{\Omega}|\boldsymbol{\theta})p(\boldsymbol{\Lambda}, \boldsymbol{\Psi}_\epsilon)p(\boldsymbol{\Phi})\\ &= [p(\boldsymbol{\Lambda}, \boldsymbol{\Psi}_\epsilon)p(\mathbf{Y}|\boldsymbol{\Lambda}, \boldsymbol{\Psi}_\epsilon, \boldsymbol{\Omega})]\\ &\quad \cdot [p(\boldsymbol{\Omega}|\boldsymbol{\Phi})p(\boldsymbol{\Phi})].\end{aligned} \tag{14}$$

Since the first term of the product on the right-hand side of (14) depends only on $(\boldsymbol{\Lambda}, \boldsymbol{\Psi}_\epsilon)$ while the second term depends only on $\boldsymbol{\Phi}$, the marginal conditional densities $p(\boldsymbol{\Lambda}, \boldsymbol{\Psi}_\epsilon|\mathbf{Y}, \boldsymbol{\Omega})$ and $p(\boldsymbol{\Phi}|\mathbf{Y}, \boldsymbol{\Omega})$ are proportional to $p(\boldsymbol{\Lambda}, \boldsymbol{\Psi}_\epsilon)p(\mathbf{Y}|\boldsymbol{\Lambda}, \boldsymbol{\Psi}_\epsilon, \boldsymbol{\Omega})$ and $p(\boldsymbol{\Omega}|\boldsymbol{\Phi})p(\boldsymbol{\Phi})$, respectively. As a result, these densities can be treated separately. The following conjugate prior distributions are considered for $(\boldsymbol{\Lambda}, \boldsymbol{\Psi}_\epsilon)$ and $\boldsymbol{\Phi}$. Let $\psi_{\epsilon k}$ and $\boldsymbol{\Lambda}_k^T$ be the k-th diagonal elements of $\boldsymbol{\Psi}_\epsilon$ and the k-th row of $\boldsymbol{\Lambda}$, respectively. For convenience, we assume that, for any $k \neq h$, the prior distribution of $\psi_{\epsilon k}$ is independent of $\psi_{\epsilon h}$, and $\boldsymbol{\Lambda}_k$ is independent of $\boldsymbol{\Lambda}_h$. The prior distributions are given by

$$\begin{aligned}&\psi_{\epsilon k}^{-1} \overset{D}{=} \text{Gamma}[\alpha_{0\epsilon k}, \beta_{0\epsilon k}],\\ &[\boldsymbol{\Lambda}_k|\psi_{\epsilon k}] \overset{D}{=} N[\boldsymbol{\Lambda}_{0k}, \psi_{\epsilon k}\mathbf{H}_{0yk}], \quad \text{and}\\ &\boldsymbol{\Phi} \overset{D}{=} W_r[\mathbf{R}_0^{-1}, \rho_0],\end{aligned} \tag{15}$$

where $\text{Gamma}(\alpha, \beta)$ represents the gamma distribution with shape parameter $\alpha > 0$ and inverse scale parameter $\beta > 0$, $IW_r[\cdot, \cdot]$ denotes an r-dimensional inverted Wishart distribution, and $\alpha_{0\epsilon k}$, $\beta_{0\epsilon k}$, $\boldsymbol{\Lambda}_{0k}$, ρ_0, and the positive definite matrices $\mathbf{H}_{0yk}$ and $\mathbf{R}_0$ are hyperparameters whose values are assumed to be given by prior information of previous studies or other sources. Let $\mathbf{Y}_k^T$ be the k-th row of $\mathbf{Y}$, $\mathbf{A}_k = (\mathbf{H}_{0yk}^{-1} + \boldsymbol{\Omega}\boldsymbol{\Omega}^T)^{-1}$, $\mathbf{a}_k = \mathbf{A}_k(\mathbf{H}_{0yk}^{-1}\boldsymbol{\Lambda}_{0k} + \boldsymbol{\Omega}\mathbf{Y}_k)$, and $\beta_{\epsilon k} = \beta_{0\epsilon k} + 2^{-1}(\mathbf{Y}_k^T\mathbf{Y}_k - \mathbf{a}_k^T\mathbf{A}_k^{-1}\mathbf{a}_k + \boldsymbol{\Lambda}_{0k}^T\mathbf{H}_{0yk}^{-1}\boldsymbol{\Lambda}_{0k})$; then it can be shown as in Appendix B that, for $k = 1, \ldots, p$, the conditional distribution of $(\boldsymbol{\Lambda}_k, \psi_{\epsilon k}^{-1})$, given $\mathbf{Y}$ and $\boldsymbol{\Omega}$, is independently distributed as the following Normal–Gamma distribution (Broemeling, 1985):

$$\begin{aligned}&[\psi_{\epsilon k}^{-1}|\mathbf{Y}, \boldsymbol{\Omega}] \overset{D}{=} \text{Gamma}[n/2 + \alpha_{0\epsilon k}, \beta_{\epsilon k}],\\ &\text{and} \quad [\boldsymbol{\Lambda}_k|\mathbf{Y}, \boldsymbol{\Omega}, \psi_{\epsilon k}^{-1}] \overset{D}{=} N[\mathbf{a}_k, \psi_{\epsilon k}\mathbf{A}_k].\end{aligned} \tag{16}$$

Since $p(\boldsymbol{\Lambda}_k, \psi_{\epsilon k}^{-1}|\mathbf{Y}, \boldsymbol{\Omega}) = p(\psi_{\epsilon k}^{-1}|\mathbf{Y}, \boldsymbol{\Omega})p(\boldsymbol{\Lambda}_k|\mathbf{Y}, \boldsymbol{\Omega}, \psi_{\epsilon k}^{-1})$, the conditional distribution of $(\boldsymbol{\Lambda}_k, \psi_{\epsilon k}^{-1})$, given $(\mathbf{Y}, \boldsymbol{\Omega})$, can be obtained via (16).

To derive $p(\boldsymbol{\Phi}|\mathbf{Y}, \boldsymbol{\Omega})$, we first note from (14) that it is proportional to $p(\boldsymbol{\Phi})p(\boldsymbol{\Omega}|\boldsymbol{\Phi})$. As $\boldsymbol{\omega}_i$ are independent, we have

$$p(\boldsymbol{\Phi}|\mathbf{Y}, \boldsymbol{\Omega}) \propto p(\boldsymbol{\Phi}) \prod_{i=1}^{n} p(\boldsymbol{\omega}_i|\boldsymbol{\Phi}).$$

From the prior distribution of $\boldsymbol{\Phi}$ given in (15), and the fact that the distribution of $\boldsymbol{\omega}_i$, given $\boldsymbol{\Phi}$, is $N(\mathbf{0}, \boldsymbol{\Phi})$, we have

$$\begin{aligned}p(\boldsymbol{\Phi}|\mathbf{Y}, \boldsymbol{\Omega}) &\propto \left[|\boldsymbol{\Phi}|^{-(\rho_0+r+1)/2} \exp\left\{-\frac{1}{2}\text{tr}(\mathbf{R}_0^{-1}\boldsymbol{\Phi}^{-1})\right\}\right]\\ &\quad \times \left[|\boldsymbol{\Phi}|^{-n/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\boldsymbol{\omega}_i^T\boldsymbol{\Phi}^{-1}\boldsymbol{\omega}_i\right\}\right]\\ &= |\boldsymbol{\Phi}|^{-(n+\rho_0+r+1)/2} \exp\left[-\frac{1}{2}\text{tr}\left\{\boldsymbol{\Phi}^{-1}(\boldsymbol{\Omega}\boldsymbol{\Omega}^T + \mathbf{R}_0^{-1})\right\}\right].\end{aligned}$$

It follows from Zellner (1971) that

$$[\boldsymbol{\Phi}|\mathbf{Y}, \boldsymbol{\Omega}] \overset{D}{=} IW_r[\boldsymbol{\Omega}\boldsymbol{\Omega}^T + \mathbf{R}_0^{-1},\ n + \rho_0]. \tag{17}$$

3.4. Bayesian estimation of the standard linear SEM

In this section, we further illustrate the Bayesian estimation by considering a standard linear SEM that is defined by (1) and (2). Let $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$ and $\boldsymbol{\Omega} = (\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_n)$, and let $\boldsymbol{\theta}$ be the vector of unknown parameters in $\boldsymbol{\Lambda}$, $\boldsymbol{\Psi}_\epsilon$, $\boldsymbol{\Pi}$, $\boldsymbol{\Gamma}$, $\boldsymbol{\Phi}$, and $\boldsymbol{\Psi}_\delta$. In the posterior analysis, we augment the observed data $\mathbf{Y}$ with the matrix of latent variable $\boldsymbol{\Omega}$, and consider the joint posterior distribution $[\boldsymbol{\theta}, \boldsymbol{\Omega}|\mathbf{Y}]$. Again, a sufficiently large number of observations are generated from this posterior distribution through the Gibbs sampler, which is implemented as follows. At the $(j+1)$-th iteration with current values of $\boldsymbol{\Omega}^{(j)}$ and $\boldsymbol{\theta}^{(j)}$:

$$\begin{aligned}&\text{(i) Generate } \boldsymbol{\Omega}^{(j+1)} \quad \text{from } p(\boldsymbol{\Omega}|\boldsymbol{\theta}^{(j)}, \mathbf{Y}),\\ &\text{(ii) Generate } \boldsymbol{\theta}^{(j+1)} \quad \text{from } p(\boldsymbol{\theta}|\boldsymbol{\Omega}^{(j+1)}, \mathbf{Y}).\end{aligned} \tag{18}$$

Note that $\boldsymbol{\theta}$ involves components that correspond to $\boldsymbol{\Lambda}$, $\boldsymbol{\Psi}_\epsilon$, $\boldsymbol{\Pi}$, $\boldsymbol{\Gamma}$, $\boldsymbol{\Phi}$, and $\boldsymbol{\Psi}_\delta$. This Gibbs sampler is similar to the one for the CFA model, except that more components corresponding to the additional parameters are involved.

The SEM is a straightforward generalization of the CFA model through an additional structural equation (2). The conditional distributions involved in the Gibbs sampler that correspond to the measurement equation are very similar to those that are associated with the CFA model. This is also true for the structural equation, because it is essentially a regression model or a factor analysis model with latent variables. Hence, the generalization of the conditional distributions in the Gibbs sampler from a CFA model to an SEM does not involve too much difficulty.

Under a similar definition and assumption, $p(\boldsymbol{\Omega}|\mathbf{Y}, \boldsymbol{\theta})$ can be expressed as in (12), with the conditional distribution of $\boldsymbol{\omega}_i$, given $(\mathbf{y}_i, \boldsymbol{\theta})$, being similarly given as in (13), expect that $\boldsymbol{\Phi}$ is replaced by

the following covariance matrix of $\boldsymbol{\omega}$, $\boldsymbol{\Sigma}_\omega$. Based on the SEM defined in (1) and (2), we have

$$\boldsymbol{\Sigma}_\omega = \begin{bmatrix} \boldsymbol{\Pi}_0^{-1}(\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}^T + \boldsymbol{\Psi}_\delta)\boldsymbol{\Pi}_0^{-T} & \boldsymbol{\Pi}_0^{-1}\boldsymbol{\Gamma}\boldsymbol{\Phi} \\ \boldsymbol{\Phi}\boldsymbol{\Gamma}^T\boldsymbol{\Pi}_0^{-T} & \boldsymbol{\Phi} \end{bmatrix}.$$

The conditional distribution of $\boldsymbol{\theta}$, given $(\mathbf{Y}, \boldsymbol{\Omega})$, is proportional to $p(\boldsymbol{\theta})p(\mathbf{Y}, \boldsymbol{\Omega}|\boldsymbol{\theta})$. Let $\boldsymbol{\theta}_y$ be the unknown parameters in $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}_\epsilon$ associated with the measurement equation; and let $\boldsymbol{\theta}_\omega$ be the unknown parameters in $\boldsymbol{\Gamma}$, $\boldsymbol{\Phi}$, and $\boldsymbol{\Psi}_\delta$ associated with the structural model with latent variables. It is natural to assume that the prior distribution of $\boldsymbol{\theta}_y$ is independent of the prior distribution of $\boldsymbol{\theta}_\omega$; that is, $p(\boldsymbol{\theta}) = p(\boldsymbol{\theta}_y)p(\boldsymbol{\theta}_\omega)$. Moreover, $p(\mathbf{Y}|\boldsymbol{\Omega}, \boldsymbol{\theta}) = p(\mathbf{Y}|\boldsymbol{\Omega}, \boldsymbol{\theta}_y)$ and $p(\boldsymbol{\Omega}|\boldsymbol{\theta}) = p(\boldsymbol{\Omega}|\boldsymbol{\theta}_\omega)$. Hence,

$$p(\boldsymbol{\theta}_y, \boldsymbol{\theta}_\omega|\mathbf{Y}, \boldsymbol{\Omega}) \propto [p(\mathbf{Y}|\boldsymbol{\Omega}, \boldsymbol{\theta}_y)p(\boldsymbol{\theta}_y)][p(\boldsymbol{\Omega}|\boldsymbol{\theta}_\omega)p(\boldsymbol{\theta}_\omega)].$$

Since the first term of the product on the right-hand side depends only on $\boldsymbol{\theta}_y$, whereas the second term depends only on $\boldsymbol{\theta}_\omega$, the marginal conditional densities $\boldsymbol{\theta}_y$ and $\boldsymbol{\theta}_\omega$ are proportional to $p(\mathbf{Y}|\boldsymbol{\Omega}, \boldsymbol{\theta}_y)p(\boldsymbol{\theta}_y)$ and $p(\boldsymbol{\Omega}|\boldsymbol{\theta}_\omega)p(\boldsymbol{\theta}_\omega)$, respectively. Consequently, these conditional densities can be treated separately. The marginal conditional distribution of $\boldsymbol{\theta}_y$ is $p(\boldsymbol{\Lambda}, \boldsymbol{\Psi}|\mathbf{Y}, \boldsymbol{\Omega})$. Under the conjugate prior distributions given in (15), this conditional distribution can be obtained from (16).

Now, consider the conditional distribution of $\boldsymbol{\theta}_\omega$ that is proportional to $p(\boldsymbol{\Omega}|\boldsymbol{\theta}_\omega)p(\boldsymbol{\theta}_\omega)$. Let $\boldsymbol{\Omega}_1 = (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n)$ and $\boldsymbol{\Omega}_2 = (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n)$. Since the distribution of $\boldsymbol{\xi}_i$ only involves $\boldsymbol{\Phi}$, $p(\boldsymbol{\Omega}_2|\boldsymbol{\theta}_\omega) = p(\boldsymbol{\Omega}_2|\boldsymbol{\Phi})$. Moreover, it is natural to assume that the prior distribution of $\boldsymbol{\Phi}$ is independent of the prior distributions of $\boldsymbol{\Pi}$, $\boldsymbol{\Gamma}$, and $\boldsymbol{\Phi}_\delta$. Hence,

$$\begin{aligned} p(\boldsymbol{\Omega}|\boldsymbol{\theta}_\omega)p(\boldsymbol{\theta}_\omega) = {} & [p(\boldsymbol{\Omega}_1|\boldsymbol{\Omega}_2, \boldsymbol{\Pi}, \boldsymbol{\Gamma}, \boldsymbol{\Psi}_\delta)p(\boldsymbol{\Pi}, \boldsymbol{\Gamma}, \boldsymbol{\Psi}_\delta)] \\ & \times [p(\boldsymbol{\Omega}_2|\boldsymbol{\Phi})p(\boldsymbol{\Phi})], \end{aligned}$$

and the marginal conditional densities of $(\boldsymbol{\Pi}, \boldsymbol{\Gamma}, \boldsymbol{\Psi}_\delta)$ and $\boldsymbol{\Phi}$ can be treated separately. Consider a conjugate-type prior distribution of $\boldsymbol{\Phi}$ as $\boldsymbol{\Phi}^{-1} \overset{D}{=} W_{r_2}[\mathbf{R}_0, \rho_0]$, a Wishart distribution with hyperparameters ρ_0, and a positive definite matrix $\mathbf{R}_0$. It can be shown by similar reasoning as in Section 3.2 that the conditional distribution of $\boldsymbol{\Phi}$, given $\boldsymbol{\Omega}_2$, is given by

$$[\boldsymbol{\Phi}|\boldsymbol{\Omega}_2] \overset{D}{=} IW_{r_2}[\boldsymbol{\Omega}_2\boldsymbol{\Omega}_2^T + \mathbf{R}_0^{-1}, n + \rho_0]. \tag{19}$$

Note that $\boldsymbol{\eta}_i = \boldsymbol{\Lambda}_\omega\boldsymbol{\omega}_i + \boldsymbol{\delta}_i$, where $\boldsymbol{\Lambda}_\omega = (\boldsymbol{\Pi}, \boldsymbol{\Gamma})$. This is very similar to a factor analysis model; and, when $\boldsymbol{\Omega}$ is given, this is a regression model. Let $\psi_{\delta k}$ be the k-th diagonal element of $\boldsymbol{\Psi}_\delta$, and let $\boldsymbol{\Lambda}_{\omega k}^T$ be the row vector that contains the unknown parameters in the k-th row of $\boldsymbol{\Lambda}_\omega$. The prior distributions of $\boldsymbol{\Lambda}_{\omega k}$ and $\psi_{\delta k}^{-1}$ are similarly specified as the following conjugate-type prior distributions:

$$\begin{aligned} & \psi_{\delta k}^{-1} \overset{D}{=} \text{Gamma}[\alpha_{0\omega k}, \beta_{0\omega k}], \quad \text{and} \\ & [\boldsymbol{\Lambda}_{\omega k}|\psi_{\delta k}] \overset{D}{=} N[\boldsymbol{\Lambda}_{0\omega k}, \psi_{\delta k}\mathbf{H}_{o\omega k}], \quad k = 1, 2, \ldots, k_1, \end{aligned} \tag{20}$$

where $\alpha_{0\delta k}$, $\beta_{0\delta k}$, and $\mathbf{H}_{0\omega k}$ are given hyperparameters. Moreover, it is assumed that, for $h \neq k$, $(\psi_{\delta k}, \boldsymbol{\Lambda}_{\omega k})$, and $(\psi_{\delta h}, \boldsymbol{\Lambda}_{\omega h})$ are independent. Then, following the same reasoning as before, it can be shown that

$$\begin{aligned} & [\psi_{\delta k}^{-1}|\boldsymbol{\Omega}] \overset{D}{=} \text{Gamma}[n/2 + \alpha_{0\delta k}, \beta_{0\delta k}], \quad \text{and} \\ & [\boldsymbol{\Lambda}_{\omega k}|\boldsymbol{\Omega}, \psi_{\delta k}^{-1}] \overset{D}{=} N[\mathbf{a}_{\omega k}, \psi_{\delta k}\mathbf{A}_{\omega k}], \end{aligned} \tag{21}$$

where $\mathbf{A}_{\omega k} = (\mathbf{H}_{0\omega k}^{-1} + \boldsymbol{\Omega}\boldsymbol{\Omega}^T)^{-1}$, $\mathbf{a}_{\omega k} = \mathbf{A}_{\omega k}(\mathbf{H}_{0\omega k}^{-1}\boldsymbol{\Lambda}_{0\omega k} + \boldsymbol{\Omega}\boldsymbol{\Omega}_{1k})$, and $\beta_{\delta k} = \beta_{0\delta k} + \frac{1}{2}(\boldsymbol{\Omega}_{1k}^T\boldsymbol{\Omega}_{1k} - \mathbf{a}_{\omega k}^T\mathbf{A}_{\omega k}^{-1}\mathbf{a}_{\omega k} + \boldsymbol{\Lambda}_{0\omega k}^T\mathbf{H}_{0\omega k}^{-1}\boldsymbol{\Lambda}_{0\omega k})$, in which $\boldsymbol{\Omega}_{1k}^T$ is the k-th row of $\boldsymbol{\Omega}_1$.

The above conditional distributions involved in the Gibbs sampler are standard; hence, simulating observations from these conditional distributions is fast and straightforward. The key idea of data augmentation and the Gibbs sampler can be extended to handle more complex SEMs. For instance, consider a nonlinear SEM with the following measurement equation and nonlinear structural equation: for $i = 1, \ldots, n$,

$$\begin{aligned} & \mathbf{y}_i = \boldsymbol{\mu} + \boldsymbol{\Lambda}\boldsymbol{\omega}_i + \boldsymbol{\epsilon}_i; \\ & \boldsymbol{\eta}_i = \boldsymbol{\Pi}\boldsymbol{\eta}_i + \boldsymbol{\Gamma}\mathbf{H}(\boldsymbol{\xi}_i) + \boldsymbol{\delta}_i, \end{aligned} \tag{22}$$

where $\boldsymbol{\mu}$ is an intercept, $\boldsymbol{\Lambda}$, $\boldsymbol{\Pi}$, $\boldsymbol{\Gamma}$, $\boldsymbol{\eta}_i$, $\boldsymbol{\xi}_i$, $\boldsymbol{\epsilon}_i$, and $\boldsymbol{\delta}_i$ are similarly defined as before, and $\mathbf{H}(\boldsymbol{\xi}) = (h_1(\boldsymbol{\xi}), \ldots, h_t(\boldsymbol{\xi}))^T$ is a nonzero vector-valued function with nonzero, known, and linearly independent differentiable functions $h_1, \ldots, h_t$, and $t \geq q_2$. In the Bayesian estimation, the observed data $\mathbf{Y}$ are again augmented with $\boldsymbol{\Omega}$, and the Gibbs sampler is also implemented; see (18). In Step (i), note that

$$\begin{aligned} p(\boldsymbol{\Omega}|\boldsymbol{\theta}, \mathbf{Y}) = {} & \prod_{i=1}^n p(\boldsymbol{\omega}_i|\mathbf{y}_i, \boldsymbol{\theta}) \\ \propto {} & \prod_{i=1}^n p(\mathbf{y}_i|\boldsymbol{\omega}_i, \boldsymbol{\theta})p(\boldsymbol{\eta}_i|\boldsymbol{\xi}_i, \boldsymbol{\theta})p(\boldsymbol{\xi}_i|\boldsymbol{\theta}), \end{aligned}$$

where $p(\boldsymbol{\omega}_i|\mathbf{y}_i, \boldsymbol{\theta})$ is proportional to

$$\begin{aligned} \exp\Big\{ & -\frac{1}{2}\boldsymbol{\xi}_i^T\boldsymbol{\Phi}^{-1}\boldsymbol{\xi}_i - \frac{1}{2}(\mathbf{y}_i - \boldsymbol{\mu} - \boldsymbol{\Lambda}\boldsymbol{\omega}_i)^T\boldsymbol{\Psi}_\epsilon^{-1}(\mathbf{y}_i - \boldsymbol{\mu} - \boldsymbol{\Lambda}\boldsymbol{\omega}_i) \\ & - \frac{1}{2}[\boldsymbol{\eta}_i - \boldsymbol{\Pi}\boldsymbol{\eta}_i - \boldsymbol{\Gamma}\mathbf{H}(\boldsymbol{\xi}_i)]^T\boldsymbol{\Psi}_\delta^{-1}[\boldsymbol{\eta}_i - \boldsymbol{\Pi}\boldsymbol{\eta}_i - \boldsymbol{\Gamma}\mathbf{H}(\boldsymbol{\xi}_i)]\Big\}. \end{aligned}$$

The MH algorithm will be used to generate observations from this conditional distribution. The deviations and the expressions of the conditional distributions involved in Step (ii) of the Gibbs sampler are similar to those presented as before.

4. A simulation study

In order to illustrate the empirical performance of the Bayesian approach for analyzing SEMs, in this section, we report results obtained from a simulation study. Let $\mathbf{y} = (y_1, \ldots, y_9)^T$ be a random vector that follows a nonlinear SEM as defined in (22). The parameters in $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$ of the measurement equation were specified as follows:

$$\boldsymbol{\mu} = (\mu_1, \ldots, \mu_9)^T, \quad \text{and}$$

$$\boldsymbol{\Lambda}^T = \begin{bmatrix} 1 & \lambda_{21} & \lambda_{31} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & \lambda_{52} & \lambda_{62} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & \lambda_{83} & \lambda_{93} \end{bmatrix},$$

where 1s and 0s in $\boldsymbol{\Lambda}$ were treated as fixed parameters to identify the model. Let $\boldsymbol{\omega} = (\eta, \xi_1, \xi_2)^T$. The nonlinear structural equation is defined as follows:

$$\eta = \gamma_1\xi_1 + \gamma_2\xi_2 + \gamma_3\xi_1^2 + \gamma_4\xi_1\xi_2 + \gamma_5\xi_2^2 + \delta. \tag{23}$$

The path diagram of this model is given in Fig. 3. The true values of the unknown parameters are as follows: $\mu_1 = \cdots = \mu_9 = 0.5$, $\lambda_{21} = \cdots = \lambda_{93} = 0.8$, $\psi_{\epsilon 1} = \cdots = \psi_{\epsilon 9} = 0.36$, $\gamma_1 = \gamma_2 = 0.3$, $\gamma_3 = \gamma_4 = \gamma_5 = 0.8$, $\psi_\delta = 0.16$, $\phi_{11} = \phi_{22} = 1$, and $\phi_{12} = 0.5$. On the basis of the above setting, data with a sample size of 300 were generated for analysis in each replication. To reveal the sensitivity of Bayesian analysis to the prior inputs, the following two different prior inputs were considered.

Prior I: The elements in $\boldsymbol{\mu}_0$, $\boldsymbol{\Lambda}_{0k}$, $\boldsymbol{\Lambda}_{0\omega k}$ are taken to be true values, $\mathbf{H}_{0yk}$ and $\mathbf{H}_{0\omega k}$ are identity matrices with appropriate dimensions, $\alpha_{0\epsilon k} = \alpha_{0\delta} = 9$, $\beta_{0\epsilon k} = \beta_{0\delta} = 4$, $\rho_0 = 7$, and $\mathbf{R}_0 = 4\boldsymbol{\Phi}_0$, where the elements in $\boldsymbol{\Phi}_0$ are the true values.

Table 1
Bayesian estimates of the parameters in the simulation study under different prior inputs via our C program.

Parameter	Prior I		Prior II		Parameter	Prior I		Prior II	
	AB	RMS	AB	RMS		AB	RMS	AB	RMS
μ_1	0.035	0.097	0.054	0.124	$\psi_{\epsilon 1}$	0.000	0.039	0.008	0.108
μ_2	0.027	0.078	0.041	0.088	$\psi_{\epsilon 2}$	0.000	0.038	0.000	0.038
μ_3	0.030	0.072	0.044	0.087	$\psi_{\epsilon 3}$	0.003	0.030	0.003	0.030
μ_4	0.006	0.067	0.002	0.075	$\psi_{\epsilon 4}$	0.009	0.039	0.011	0.040
μ_5	0.006	0.055	0.001	0.064	$\psi_{\epsilon 5}$	0.001	0.028	0.002	0.029
μ_6	0.006	0.052	0.000	0.058	$\psi_{\epsilon 6}$	0.010	0.034	0.010	0.033
μ_7	0.006	0.067	0.001	0.073	$\psi_{\epsilon 7}$	0.014	0.041	0.015	0.042
μ_8	0.013	0.055	0.008	0.057	$\psi_{\epsilon 8}$	0.001	0.029	0.002	0.030
μ_9	0.002	0.056	0.003	0.059	$\psi_{\epsilon 9}$	0.004	0.030	0.004	0.030
λ_{21}	0.001	0.020	0.003	0.030	γ_1	0.022	0.138	0.005	0.150
λ_{31}	0.002	0.019	0.001	0.029	γ_2	0.008	0.118	0.005	0.141
λ_{52}	0.013	0.053	0.014	0.055	γ_3	0.004	0.010	0.002	0.108
λ_{62}	0.006	0.046	0.008	0.049	γ_4	0.016	0.128	0.021	0.137
λ_{83}	0.012	0.051	0.011	0.052	γ_5	0.033	0.112	0.027	0.114
λ_{93}	0.017	0.053	0.017	0.056	ϕ_{11}	0.028	0.105	0.030	0.112
ψ_δ	0.090	0.094	0.092	0.096	ϕ_{12}	0.010	0.079	0.012	0.083
					ϕ_{12}	0.042	0.113	0.041	0.119

Prior II: the elements in $\boldsymbol{\mu}_0$, $\boldsymbol{\Lambda}_{0k}$, $\boldsymbol{\Lambda}_{0\omega k}$ are zero, $\mathbf{H}_{0yk}$ and $\mathbf{H}_{0\omega k}$ are ten times identity matrices with appropriate dimensions, $\alpha_{0\epsilon k} = \alpha_{0\delta} = 9$, $\beta_{0\epsilon k} = \beta_{0\delta} = 4$, $\rho_0 = 4$, and $\mathbf{R}_0$ is the identity matrix.

Prior I can be regarded as informative prior inputs while Prior II can be regarded as rough ad hoc prior inputs. We first conducted the analysis using our own C program. Based on some pilot runs, we found that the algorithm converged in about 2000 iterations. We took a burn-in phase of 2000 iterations, and collected $J = 2000$ observations after convergence for obtaining the Bayesian estimates and their standard errors estimates. The simulation results are obtained from 100 replications. The absolute bias (AB) and the root mean square (RMS) errors are reported in Table 1. It can be seen that the absolute bias and the root mean square errors are both small, which indicates that the Bayesian approach provides accurate estimates of the unknown parameters. We also found that the estimates were similar under the two different prior inputs. It seems that the Bayesian estimates are not sensitive to the prior inputs with the related sample size $n = 300$.

The datasets used in the simulation study have been reanalyzed by WinBUGS, based on the same prior inputs. We observe that the WinBUGS program converged in 4000 iterations. To reveal convergence, plots of some parameter values against the MCMC iterations in a randomly selected replication are presented in Fig. 4. It can be seen from these plots that the sequences of parameter values generated by different starting points mixed well after 4000 iterations. As before, we collected 2000 observations after convergence to obtain the estimates and their standard errors. The results are reported in Table 2. It can be seen that the estimates are very close to those obtained from the C program. Moreover, plots of empirical posterior distributions for some parameters are presented in Fig. 5. The WinBUGS codes for analyzing this example based on Prior I inputs are given in Appendix D. The datasets in this simulation study are provided at http://www.sta.cuhk.edu.hk/xysong/JMP.

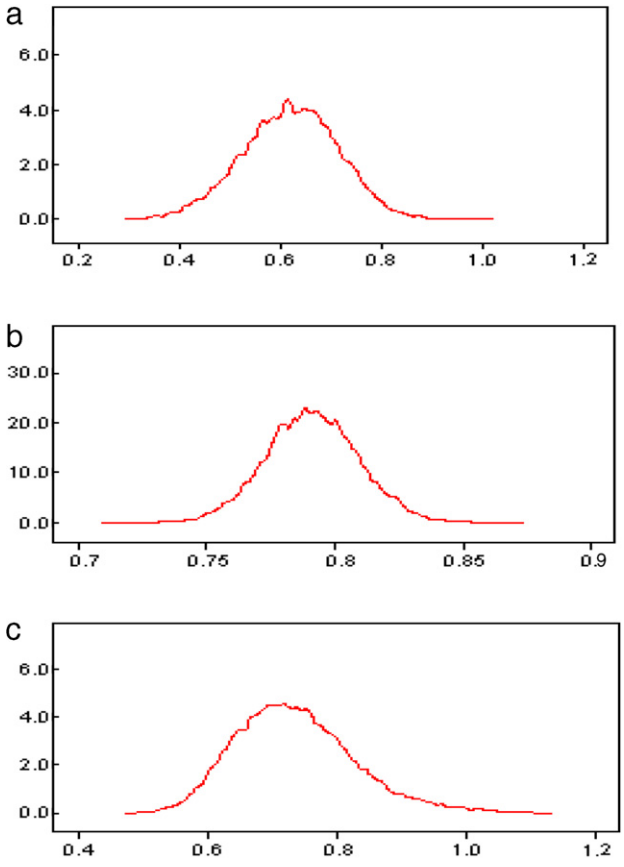

Fig. 5. Posterior densities of μ_1, λ_{21}, and γ_3.

5. Bayesian model comparison

One important statistical inference in SEMs is related to model comparison or testing various hypotheses about the model. In the field of structural equation modeling, a classical approach in 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. However, as pointed out in the statistics literature (see Berger and Sellke (1987); Berger and Delampady (1987)), the p-value of a significance test in hypothesis testing, which is related only to the type-I error, is a measure of evidence against the null model, but not a means of supporting the null model. For complex SEMs, the asymptotic distributions of these test statistics are usually hard to derive. Moreover, significance tests cannot be applied to test nonnested hypotheses or to compare nonnested models. The main objective of this section is to introduce two well-known model comparison statistics, namely the Bayes factor (see Kass and Raftery (1995), Lodewyckx et al. (2011) and Morey, Rouder, Pratte, and Speckman (2011)) and the Deviance Information Criterion (DIC; Spiegelhalter, Best, Carlin, and van der Linde (2002)), which do not have the above-mentioned problems.

The Bayes factor is a well-known statistic for model comparison. Suppose that the observed data $\mathbf{Y}$ with a sample size n have

Table 2
Bayesian estimates of the parameters in the simulation study under different prior inputs via WinBUGS.

Parameter	Prior I		Prior II		Parameter	Prior I		Prior II	
	AB	RMS	AB	RMS		AB	RMS	AB	RMS
μ_1	0.038	0.086	0.036	0.088	$\psi_{\epsilon 1}$	0.002	0.038	0.002	0.038
μ_2	0.030	0.074	0.028	0.075	$\psi_{\epsilon 2}$	0.001	0.037	0.001	0.037
μ_3	0.031	0.069	0.030	0.070	$\psi_{\epsilon 3}$	0.003	0.030	0.004	0.030
μ_4	0.016	0.068	0.024	0.075	$\psi_{\epsilon 4}$	0.013	0.040	0.012	0.040
μ_5	0.014	0.058	0.021	0.063	$\psi_{\epsilon 5}$	0.003	0.028	0.003	0.028
μ_6	0.014	0.055	0.021	0.059	$\psi_{\epsilon 6}$	0.011	0.034	0.011	0.034
μ_7	0.018	0.065	0.028	0.072	$\psi_{\epsilon 7}$	0.015	0.041	0.012	0.041
μ_8	0.022	0.056	0.030	0.062	$\psi_{\epsilon 8}$	0.000	0.029	0.001	0.030
μ_9	0.012	0.056	0.021	0.062	$\psi_{\epsilon 9}$	0.004	0.031	0.006	0.031
λ_{21}	0.001	0.019	0.002	0.019	γ_1	0.046	0.138	0.072	0.159
λ_{31}	0.002	0.018	0.002	0.018	γ_2	0.036	0.118	0.064	0.140
λ_{52}	0.020	0.054	0.018	0.054	γ_3	0.026	0.100	0.023	0.105
λ_{62}	0.014	0.050	0.012	0.050	γ_4	0.032	0.132	0.027	0.145
λ_{83}	0.014	0.051	0.009	0.052	γ_5	0.025	0.102	0.013	0.106
λ_{93}	0.020	0.054	0.015	0.053	ϕ_{11}	0.045	0.108	0.041	0.109
ψ_δ	0.086	0.089	0.088	0.091	ϕ_{12}	0.015	0.078	0.012	0.080
					ϕ_{12}	0.046	0.111	0.034	0.114

arisen under one of the two competing models M_1 and M_0 according to probability densities $p(\mathbf{Y}|M_1)$ and $p(\mathbf{Y}|M_0)$, respectively. For $k = 0, 1$, let $p(M_k)$ and $p(M_k|\mathbf{Y})$ be the prior and posterior probability densities, respectively. From the Bayes theorem, we obtain

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

The Bayes factor for evaluating M_1 against M_0 is defined as

$$B_{10} = \frac{p(\mathbf{Y}|M_1)}{p(\mathbf{Y}|M_0)}. \tag{24}$$

Hence, the Bayes factor is a summary of the evidence provided by the data in favor of M_1 as opposed to M_0, and it measures how well M_1 predicts the data relative to M_0. As suggested by Kass and Raftery (1995), the resulting statistic can be roughly interpreted according to the following table:

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

(25)

See Lee (2007, Chapter 5) for more discussion about using this table, and the 'Discussion' in Gelman, Meng, and Stern (1996) about various issues on model comparison.

The prior distributions of the parameters are involved in the Bayes factor; see Eq. (26) below. As pointed out by Kass and Raftery (1995), Bayes factors are sensitive to prior inputs, and using noninformative priors on the parameters under M_1 will force the Bayes factor to favor the competitive model M_0. In most Bayesian analyses of SEMs, the proper conjugate-type prior distributions that involve prior inputs of the hyperparameters have been used. To study the stability of the results, a common method (see Kass and Raftery (1995) and Lee (2007)) is to perturb the prior inputs and then evaluate the sensitivity of the recomputed Bayes factor values. In general, the marginal densities $p(\mathbf{Y}|M_k)$, $k = 0, 1$, are obtained by integrating over the parameter space; that is,

$$p(\mathbf{Y}|M_k) = \int p(\mathbf{Y}|\boldsymbol{\theta}_k, M_k)p(\boldsymbol{\theta}_k|M_k)d\boldsymbol{\theta}_k, \tag{26}$$

where $\boldsymbol{\theta}_k$ is the parameter vector in M_k, $p(\boldsymbol{\theta}_k|M_k)$ is its prior density, and $p(\mathbf{Y}|\boldsymbol{\theta}_k, M_k)$ is the probability density of $\mathbf{Y}$, given $\boldsymbol{\theta}_k$. The dimension of the above integral is equal to the dimension of $\boldsymbol{\theta}_k$. Usually, it is very difficult to obtain B_{10} analytically, and various analytic and numerical approximations have been proposed in the literature (see Dicicio, Kass, Raftery, and Wasserman (1997)). Here, the procedure based on idea of the path sampling (Gelman & Meng, 1998) is proposed to compute the Bayes factor. Path sampling, which is a generalization of importance sampling and bridge sampling (Meng & Wong, 1996), has several nice features. The application of path sampling in computing the logarithm of the Bayes factor is described in Appendix C.

To illustrate the path sampling procedure in computing the logarithm of the Bayes factor for model comparison, we let M_0 be the model from the simulation study (i.e., the model with the nonlinear structural equation; see Eq. (23)). The competing models M_1 and M_2 are defined by the same measurement equation and the following different structural equations. Hence (and we then start by giving M_0 from Eq. (23) as well, so the reader can easily spot the difference between the models)

M_1: $\eta = \gamma_1\xi_1 + \gamma_2\xi_2 + \gamma_3\xi_1^2 + \delta$,

M_2: $\eta = \gamma_1\xi_1 + \gamma_2\xi_2 + \delta$.

The corresponding linked models are

M_{t01}: $\eta = \gamma_1\xi_1 + \gamma_2\xi_2 + \gamma_3\xi_1^2 + (1-t)\{\gamma_4\xi_1\xi_2 + \gamma_5\xi_2^2\} + \delta$,

M_{t02}: $\eta = \gamma_1\xi_1 + \gamma_2\xi_2 + (1-t)\{\gamma_3\xi_1^2 + \gamma_4\xi_1\xi_2 + \gamma_5\xi_2^2\} + \delta$,

M_{t12}: $\eta = \gamma_1\xi_1 + \gamma_2\xi_2 + (1-t)\gamma_3\xi_1^2 + \delta$.

Clearly, when $t = 0$, $M_{t01} = M_0$, $M_{t02} = M_0$, and $M_{t12} = M_1$; when $t = 1$, $M_{t01} = M_1$, $M_{t02} = M_2$, and $M_{t12} = M_2$. The log-likelihood functions corresponding to the linked models, and their derivatives with respect to t, are given in Lee (2007).

In the path sampling procedure, we take $S = 20$ in computing the Bayes factor; see (C.3). For each $t_{(s)}$, we take a burn-in phase of 2000 iterations in the MCMC algorithm, and 2000 observations are collected after convergence for computing the logarithm Bayes factor. To give some idea about the sensitivity of the procedure with respect to prior inputs, the two different prior inputs mentioned in the simulation study are considered. The summaries of the estimated logarithm Bayes factors based on 100 replications are presented in Table 3. Based on (25), the Bayes factor always chooses true model M_0 in all replications. Moreover, it can be seen that M_1 is better than the linear model M_2. Finally, the summaries corresponding to DIC values obtained from WinBUGS for M_0, M_1, and M_2 with different prior inputs are also given in Table 3. Based on these results, we can conclude that both the Bayes factor and the DIC values choose the true model correctly.

Another goodness-of-fit or model comparison statistic is the DIC; see Spiegelhalter et al. (2002). For a competing model M_k with

Table 3
Summary of estimated log Bayes factors and DIC values under different prior inputs.

	Prior I		Prior II	
	Mean	SD	Mean	SD
$\log B_{10}$	−167.910	31.039	−115.197	16.184
$\log B_{20}$	−442.418	30.559	−176.502	17.282
$\log B_{21}$	−368.981	48.279	−189.242	25.924
DIC_0	5508.752	73.062	5509.038	73.098
DIC_1	5648.058	76.129	5739.811	72.628
DIC_2	5712.052	75.167	5710.573	75.093

a vector of unknown parameter $\boldsymbol{\theta}_k$, the DIC is defined as

$$\text{DIC}_k = \overline{D(\boldsymbol{\theta}_k)} + d_k, \tag{27}$$

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

$$\overline{D(\boldsymbol{\theta}_k)} = E_{\boldsymbol{\theta}_k}\{-2\log p(\mathbf{Y}|\boldsymbol{\theta}_k, M_k)|\mathbf{Y}\}. \tag{28}$$

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

$$d_k = E_{\boldsymbol{\theta}_k}\{-2\log p(\mathbf{Y}|\boldsymbol{\theta}_k, M_k)|\mathbf{Y}\} + 2\log p(\mathbf{Y}|\hat{\boldsymbol{\theta}}_k), \tag{29}$$

in which $\hat{\boldsymbol{\theta}}_k$ is the Bayesian estimate of $\boldsymbol{\theta}_k$. Let $\{\boldsymbol{\theta}_k^{(j)}: j = 1, \ldots, J\}$ be a sample of observations simulated from the posterior distribution. The expectation in (28) and (29) can be estimated as follows:

$$E_{\boldsymbol{\theta}_k}\{-2\log p(\mathbf{Y}|\boldsymbol{\theta}_k, M_k)|\mathbf{Y}\} = -\frac{2}{J}\sum_{j=1}^{J}\log p(\mathbf{Y}|\boldsymbol{\theta}_k^{(j)}, M_k). \tag{30}$$

In model comparison, the model with the smaller DIC value is selected. In analyzing a hypothesized model, WinBUGS produces a DIC value which can be used for model comparison. As pointed out in the WinBUGS User Manual (Spiegelhalter, Thomas, Best, & Lunn, 2003), in practical applications of the DIC, it is important to note the following. (i) 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. (ii) The DIC can be applied to nonnested models. Similar to the Bayes factor, the DIC gives clear conclusion to support the null hypothesis or the alternative hypothesis. (iii) The DIC assumes the posterior mean to be a good estimate of the parameter.

6. Final remarks

This article introduces a Bayesian approach in analyzing structure equation models. This approach can utilize prior information for achieving better statistical results and give more reliable inference for small samples. In contrast to the covariance structure approach used by most SEM software, the statistical results developed by this approach are based on properties of the raw observations rather than the properties of the sample covariance matrix. This Bayesian approach has several advantages. First, the first-moment properties of raw observations are simpler than the second-moment properties of the sample covariance matrix, and hence the Bayesian approach can deal with more complex situations. Second, it gives a direct estimate of latent variables. Third, this approach can directly model raw observations and latent variables through the familiar regression model; hence it gives more direct interpretation.

The key idea in the development of the Bayesian approach is data augmentation coupled with efficient MCMC methods. The strategy is to treat all the latent quantities (such as latent variables, latent measurements, missing data, etc.) as hypothetical missing data and augment the observed data with them, and then use MCMC methods to draw observations from the posterior distribution based on the complete data (see Lee (2007), for a more comprehensive treatment). This strategy has been applied to develop Bayesian methodologies for analyzing important generalizations of the basic SEMs, such as nonlinear SEMs, multilevel SEMs, and longitudinal SEMs. These extensions illustrate that the Bayesian approach to the analysis of SEMs has wide applicability and broad appeal.

Acknowledgments

This paper is fully supported by grant GRF404711 from the Research Grant Council of the Hong Kong Special Administration Region. We are deeply thankful to the Editor and Dr. Dora Matzke, Dr. C.V. (Conor) Dolan, and Dr. Rens van de Schoot for their valuable comments, which substantially improved the paper, and to Dr. J.H. Cai for his assistance in providing the numerical results.

Appendix A. EPSR value

Assessing the convergence of a Monte Carlo simulation procedure should be on the basis of several simulation sequences generated independently from different starting values. The following approach (see Gelman (1996)) involves monitoring each scalar estimate (e.g. parameter) of interest separately. Let n be the length of the each sequence, after discarding the first part of the simulations. For each scalar estimate, say ψ, let ψ_{jk} ($j = 1, \ldots, n$; $k = 1, \ldots, K$) be the draws from K parallel sequences of length n. The between-sequence and within-sequence variances are computed as

$$B = \frac{n}{K-1}\sum_{k=1}^{K}(\psi_{\cdot k} - \psi_{\cdot\cdot})^2, \quad \text{where}$$

$$\psi_{\cdot k} = n^{-1}\sum_{j=1}^{n}\psi_{jk},\ \psi_{\cdot\cdot} = K^{-1}\sum_{k=1}^{K}\psi_k$$

$$W = \frac{1}{K}\sum_{k=1}^{K}s_k^2, \quad \text{where } s_k^2 = (n-1)^{-1}\sum_{j=1}^{n}(\psi_{jk} - \psi_{\cdot k})^2.$$

The estimate of $\text{Var}(\psi|\mathbf{Y})$, the marginal posterior variance of the estimate, is then obtained by a weighted average of B and W as follows:

$$\widehat{\text{Var}}(\psi) = \frac{n-1}{n}W + \frac{1}{n}B.$$

The 'estimated potential scale reduction (EPSR)' is defined as

$$\hat{R}^{1/2} = [\widehat{\text{Var}}(\psi)/W]^{1/2}.$$

As the simulation converges, $\hat{R}^{1/2}$ should be close to 1.0. In monitoring converges, all EPSR values for all scalar estimates are computed. The whole simulation procedure is said to be converged if all the EPSR values are less than 1.2.

Appendix B. Derivations of conditional distributions

To simplify notation in the derivation of $p(\boldsymbol{\Lambda}_k, \psi_{\epsilon_k}|\mathbf{Y}, \boldsymbol{\Omega})$, we let $\nu_k = \psi_{\epsilon k}^{-1}$. From (15), the conjugate prior density of ν_k, and the conjugate prior density of $\boldsymbol{\Lambda}_k$, given ν_k, are proportional to $\nu_k^{\alpha_{0k}-1}\exp(-\beta_{0k}\nu_k)$ and $\nu_k^{r/2}\exp\{-\frac{1}{2}(\boldsymbol{\Lambda}_k-\boldsymbol{\Lambda}_{0k})^T\mathbf{H}_{0k}^{-1}(\boldsymbol{\Lambda}_k-\boldsymbol{\Lambda}_{0k})\nu_k\}$, respectively. It can be seen that the likelihood of $\mathbf{Y}$ is given by

$$p(\mathbf{Y}|\boldsymbol{\Lambda}, \boldsymbol{\Psi}_\epsilon, \boldsymbol{\Omega}) \propto |\boldsymbol{\Psi}_\epsilon|^{-n/2}\ \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{y}_i - \boldsymbol{\Lambda}\boldsymbol{\omega}_i)^T\boldsymbol{\Psi}_\epsilon^{-1}(\mathbf{y}_i - \boldsymbol{\Lambda}\boldsymbol{\omega}_i)\right\}.$$

Let $\mathbf{Y}_k^T$ be the k-th row of $\mathbf{Y}$, y_{ki} be the i-th component of $\mathbf{Y}_k^T$, $\mathbf{A}_k^* = (\boldsymbol{\Omega}\boldsymbol{\Omega}^T)^{-1}\boldsymbol{\Omega}\mathbf{Y}_k$, and $a = \mathbf{Y}_k^T\mathbf{Y}_k - \mathbf{Y}_k^T\boldsymbol{\Omega}(\boldsymbol{\Omega}\boldsymbol{\Omega}^T)^{-1}\boldsymbol{\Omega}\mathbf{Y}_k = \mathbf{Y}_k^T\mathbf{Y}_k -$

$\mathbf{A}_k^{*T}(\boldsymbol{\Omega}\boldsymbol{\Omega}^T)\mathbf{A}_k^*$; then the exponential term in $p(\mathbf{Y}|\boldsymbol{\Lambda}, \boldsymbol{\Psi}_\epsilon, \boldsymbol{\Omega})$ can be expressed as

$$
\begin{aligned}
&-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{y}_i-\boldsymbol{\Lambda}\boldsymbol{\omega}_i)^T\boldsymbol{\Psi}_\epsilon^{-1}(\mathbf{y}_i-\boldsymbol{\Lambda}\boldsymbol{\omega}_i)\\
&=-\frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{p}\psi_{\epsilon k}^{-1}(y_{ki}-\boldsymbol{\Lambda}_k^T\boldsymbol{\omega}_i)^2\\
&=-\frac{1}{2}\sum_{k=1}^{p}\left\{\nu_k\left[\sum_{i=1}^{n}y_{ki}^2-2\boldsymbol{\Lambda}_k^T\sum_{i=1}^{n}y_{ki}\boldsymbol{\omega}_i\right.\right.\\
&\quad\left.\left.+\ \mathrm{tr}\left(\boldsymbol{\Lambda}_k\boldsymbol{\Lambda}_k^T\sum_{i=1}^{n}\boldsymbol{\omega}_i\boldsymbol{\omega}_i^T\right)\right]\right\}\\
&=-\frac{1}{2}\sum_{k=1}^{p}\left\{\nu_k\left[\mathbf{Y}_k^T\mathbf{Y}_k-2\boldsymbol{\Lambda}_k^T\boldsymbol{\Omega}\mathbf{Y}_k+\boldsymbol{\Lambda}_k^T(\boldsymbol{\Omega}\boldsymbol{\Omega}^T)\boldsymbol{\Lambda}_k\right]\right\}\\
&=-\frac{1}{2}\sum_{k=1}^{p}\left\{\nu_k\left[\mathbf{Y}_k^T\mathbf{Y}_k-\mathbf{Y}_k^T\boldsymbol{\Omega}^T(\boldsymbol{\Omega}\boldsymbol{\Omega}^T)^{-1}\boldsymbol{\Omega}\mathbf{Y}_k\right]\right.\\
&\quad+\nu_k\left[\boldsymbol{\Lambda}_k-(\boldsymbol{\Omega}\boldsymbol{\Omega}^T)^{-1}\boldsymbol{\Omega}\mathbf{Y}_k\right]^T(\boldsymbol{\Omega}\boldsymbol{\Omega}^T)\\
&\quad\left.\times\left[\boldsymbol{\Lambda}_k-(\boldsymbol{\Omega}\boldsymbol{\Omega}^T)^{-1}\boldsymbol{\Omega}\mathbf{Y}_k\right]\right\}\\
&=-\frac{1}{2}\sum_{k=1}^{p}\{\nu_k[a+(\boldsymbol{\Lambda}_k-\mathbf{A}_k^*)^T(\boldsymbol{\Omega}\boldsymbol{\Omega}^T)(\boldsymbol{\Lambda}_k-\mathbf{A}_k^*)]\}.
\end{aligned}
$$

Therefore, it follows from the likelihood of $\mathbf{Y}$ and the conjugate densities of $\boldsymbol{\Lambda}_k$ and ν_k that

$$
\begin{aligned}
p(\boldsymbol{\Lambda}_k, \nu_1, \ldots, \nu_p|\mathbf{Y}, \boldsymbol{\Omega}) &\propto \prod_{k=1}^{p}\left[\nu_k^{n/2+r/2+\alpha_{0\epsilon k}-1}\right.\\
&\quad\times\exp\left\{-\frac{1}{2}\nu_k[(\boldsymbol{\Lambda}_k-\mathbf{A}_k^*)^T(\boldsymbol{\Omega}\boldsymbol{\Omega}^T)(\boldsymbol{\Lambda}_k-\mathbf{A}_k^*)\right.\\
&\quad\left.\left.+\ (\boldsymbol{\Lambda}_k-\boldsymbol{\Lambda}_{0k})^T\mathbf{H}_{0yk}^{-1}(\boldsymbol{\Lambda}_k-\boldsymbol{\Lambda}_{0k})]-\nu_k(\beta_{0\epsilon k}+a/2)\right\}\right]\\
&=\prod_{k=1}^{p}p(\boldsymbol{\Lambda}_k, \nu_k|\mathbf{Y}, \boldsymbol{\Omega}).
\end{aligned}
$$

From the above equation, it can be seen that the conditional distributions of $(\boldsymbol{\Lambda}_k, \nu_k)$, given $(\mathbf{Y}, \boldsymbol{\Omega})$, are mutually independent for $k = 1, \ldots, p$. Hence, it suffices to derive $p(\boldsymbol{\Lambda}_k, \nu_k|\mathbf{Y}, \mathbf{Z})$.

Let $\mathbf{A}_k = (\mathbf{H}_{0yk}^{-1}+\boldsymbol{\Omega}\boldsymbol{\Omega}^T)^{-1}$ and $\mathbf{a}_k = \mathbf{A}_k(\mathbf{H}_{0yk}^{-1}\boldsymbol{\Lambda}_{0k}+\boldsymbol{\Omega}\mathbf{Y}_k)$. It follows that

$$
\begin{aligned}
&(\boldsymbol{\Lambda}_k-\mathbf{A}_k^*)^T(\boldsymbol{\Omega}\boldsymbol{\Omega}^T)(\boldsymbol{\Lambda}_k-\mathbf{A}_k^*)+(\boldsymbol{\Lambda}_k-\boldsymbol{\Lambda}_{0k})^T\mathbf{H}_{0yk}^{-1}(\boldsymbol{\Lambda}_k-\boldsymbol{\Lambda}_{0k})\\
&\quad=(\boldsymbol{\Lambda}_k-\mathbf{a}_k)^T\mathbf{A}_k^{-1}(\boldsymbol{\Lambda}_k-\mathbf{a}_k)-\mathbf{a}_k^T\mathbf{A}_k^{-1}\mathbf{a}_k\\
&\qquad+\mathbf{A}_k^{*T}\boldsymbol{\Omega}\boldsymbol{\Omega}^T\mathbf{A}_k^*+\boldsymbol{\Lambda}_{0k}^T\mathbf{H}_{0yk}^{-1}\boldsymbol{\Lambda}_{0k}.
\end{aligned}
$$

Hence,

$$
\begin{aligned}
p(\boldsymbol{\Lambda}_k, \nu_k|\mathbf{Y}, \boldsymbol{\Omega}) &= p(\nu_k|\mathbf{Y}, \boldsymbol{\Omega})\,p(\boldsymbol{\Lambda}_k|\mathbf{Y}, \boldsymbol{\Omega}, \nu_k)\\
&\propto\left[\nu_k^{n/2+\alpha_{0\epsilon k}-1}\exp\{-\beta_{\epsilon k}\nu_k\}\right]\\
&\quad\cdot\left[\nu_k^{r/2}\exp\left\{-\frac{1}{2}(\boldsymbol{\Lambda}_k-\mathbf{a}_k)^T\mathbf{A}_k^{-1}(\boldsymbol{\Lambda}_k-\mathbf{a}_k)\nu_k\right\}\right],
\end{aligned}
$$

where $\beta_{\epsilon k}=\beta_{0\epsilon k}+2^{-1}(\mathbf{Y}_k^T\mathbf{Y}_k-\mathbf{a}_k^T\mathbf{A}_k^{-1}\mathbf{a}_k+\boldsymbol{\Lambda}_{0k}^T\mathbf{H}_{0yk}^{-1}\boldsymbol{\Lambda}_{0k})$. Thus, the posterior distribution of $(\boldsymbol{\Lambda}_k, \nu_k)$, given $\mathbf{Y}$ and $\mathbf{Z}$, is the following Normal–Gamma distribution (Broemeling, 1985):

$$
[\nu_k|\mathbf{Y}, \boldsymbol{\Omega}]\overset{D}{=}\mathrm{Gamma}[n/2+\alpha_{0\epsilon k}, \beta_{\epsilon k}],\quad\text{and}
$$

$$
[\boldsymbol{\Lambda}_k|\mathbf{Y}, \boldsymbol{\Omega}, \nu_k]\overset{D}{=}N[\mathbf{a}_k, \nu_k^{-1}\mathbf{A}_k].
$$

Appendix C. Path sampling

In general, let $\mathbf{Y}$ be the matrix of observed data, and let $\boldsymbol{\Omega}$ be the matrix of latent variables in the model. 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 $p(\mathbf{Y}, \boldsymbol{\Omega}, \boldsymbol{\theta})$ taken as the unnormalized density. Now, for a continuous parameter t in [0, 1], let

$$
\begin{aligned}
z(t) &= p(\mathbf{Y}|t)=\int p(\mathbf{Y}, \boldsymbol{\Omega}, \boldsymbol{\theta}|t)d\boldsymbol{\Omega}d\boldsymbol{\theta}\\
&=\int p(\mathbf{Y}, \boldsymbol{\Omega}|\boldsymbol{\theta}, t)p(\boldsymbol{\theta})d\boldsymbol{\Omega}d\boldsymbol{\theta},
\end{aligned}
\tag{C.1}
$$

with $p(\boldsymbol{\theta})$ being 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 $B_{10} = z(1)/z(0)$. Taking the logarithm and then differentiating (C.1) with respect to t, and assuming the legitimacy of interchange of integration with differentiation, it can be shown that

$$
\frac{d\log z(t)}{dt}=E_{\Omega,\theta}\left[\frac{d}{dt}\log p(\mathbf{Y}, \boldsymbol{\Omega}, \boldsymbol{\theta}|t)\right],
$$

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

$$
\begin{aligned}
U(\mathbf{Y}, \boldsymbol{\Omega}, \boldsymbol{\theta}, t) &= \frac{d}{dt}\log p(\mathbf{Y}, \boldsymbol{\Omega}, \boldsymbol{\theta}|t)\\
&=\frac{d}{dt}\log p(\mathbf{Y}, \boldsymbol{\Omega}|\boldsymbol{\theta}, t),
\end{aligned}
\tag{C.2}
$$

we have

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

To numerically evaluate the integral over t, we first order the unique values of S fixed grids $\{t_{(s)}\}_{s=0}^{S}$ such that $t_{(0)} = 0 < t_{(1)} < t_{(2)} < \cdots < t_{(S)} < t_{(S+1)} = 1$, and then estimate $\log B_{10}$ by

$$
\widehat{\log B_{10}}=\frac{1}{2}\sum_{s=0}^{S}(t_{(s+1)}-t_{(s)})(\bar{U}_{(s+1)}+\bar{U}_{(s)}),
\tag{C.3}
$$

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

$$
\bar{U}_{(s)}=J^{-1}\sum_{j=1}^{J}U(\mathbf{Y}, \boldsymbol{\Omega}^{(j)}, \boldsymbol{\theta}^{(j)}, t_{(s)}),
\tag{C.4}
$$

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

In a comprehensive comparative study of various computing methods in computing the Bayes factor, Diciccio et al. (1997) pointed out bridge sampling (Meng & Wong, 1996) is an attractive method. Gelman and Meng (1998) showed that path sampling is a generalization of bridge sampling and importance sampling. Hence, it is expected that path sampling can give a more accurate result than bridge sampling in computing the Bayes factor. Moreover, we can always construct a continuous path to link two competing models with the same support. Hence, the method can be applied to a wide variety of problems. Unlike some methods in estimating the marginal likelihood via posterior simulation, it does not require one to estimate the location and/or scale parameters in the posterior. Distinct from most existing approaches, the prior density is not directly involved in the evaluation. Finally, the logarithm scale of Bayes factor is computed, which is generally more stable than the ratio scale.

```
model {
    for (i in 1:N) {
        #measurement equation
        for (j in 1:9) { y[i,j]~dnorm(mu[i,j], psi[j]) }

        mu[i,1]<-u[1]+eta[i]  mu[i,2]<-u[2]+lam[1]*eta[i]
        mu[i,3]<-u[3]+lam[2]*eta[i]
        mu[i,4]<-u[4]+xi[i,1] mu[i,5]<-u[5]+lam[3]*xi[i,1]
        mu[i,6]<-u[6]+lam[4]*xi[i,1]
        mu[i,7]<-u[7]+xi[i,2] mu[i,8]<-u[8]+lam[5]*xi[i,2]
        mu[i,9]<-u[9]+lam[6]*xi[i,2]

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

        nu[i]<-gam[1]*xi[i,1]+gam[2]*xi[i,2]+gam[3]*xi[i,1]*xi[i,1]
                +gam[4]*xi[i,1]*xi[i,2]+gam[5]*xi[i,2]*xi[i,2]

        xi[i,1:2]~dmnorm(ux[1:2], phi[1:2,1:2])
    } #model definition

    #prior inputs
    for (i in 1:9) { u[i]~dnorm(0.5,1) }

    lam[1]~dnorm(0.8, psi[2])  lam[2]~dnorm(0.8, psi[3])
    lam[3]~dnorm(0.8, psi[5])  lam[4]~dnorm(0.8, psi[6])
    lam[5]~dnorm(0.8, psi[8])  lam[6]~dnorm(0.8, psi[9])

    for (i in 1:9) { psi[i]~dgamma(9,4)  sgm[i]<-1/psi[i] }

    gam[1]~dnorm(0.3, psd)  gam[2]~dnorm(0.3, psd)
    gam[3]~dnorm(0.8, psd)  gam[4]~dnorm(0.8, psd)  gam[5]~dnorm(0.8, psd)

    psd~dgamma(9,4)  sgd<-1/psd

    phi[1:2,1:2]~dwish(R[1:2,1:2], 7) phx[1:2,1:2]<-inverse(phi[1:2,1:2])
}
```

Box I.

Appendix D. WinBUGS code

The WinBUGS code is given in Box I.

References

Ansari, A., & Jedidi, K. (2000). Bayesian factor analysis for multilevel binary observations. *Psychometrika*, *65*, 475–498.

Ansari, A., Jedidi, K., & Jagpal, S. (2000). A hierarchical Bayesian methodology for treating heterogeneity in structural equation models. *Marketing Science*, *19*, 328–347.

Albert, J. H., & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, *88*, 669–679.

Arhonditsis, G. B., Stow, C. A., Steinberg, L. J., Kenney, M. A., Lathrop, R. C., McBride, S. J., et al. (2006). Exploring ecological patterns with structural equation modeling and Bayesian analysis. *Ecological Modelling*, *192*, 385–409.

Arminger, G., & Muthen, B. (1998). A Bayesian approach to nonlinear latent variable models using the Gibbs sampler and the Metropolis–Hastings algorithm. *Psychometrika*, *63*, 271–300.

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis*. New York: Springer-Verlag.

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

Berger, J. O., & Delampady, M. (1987). Testing a point null hypotheses. *Statistical Science*, *3*, 317–335.

Broemeling, L. D. (1985). *Bayesian analysis of linear models*. New York: Marcel Dekker Inc.

Box, G. E. P., & Tiao, G. C. (1973). *Bayesian inference in statistical analysis*. Reading, MA: Addison-Wesley.

Congdon, P. (2003). *Applied Bayesian modeling*. Hoboken, New York: John Wiley.

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

Dunson, D. B. (2000). Bayesian latent variable models for clustered mixed outcomes. *Journal of the Royal Statistical Society, Series B*, *62*, 355–366.

Dunson, D. B. (2003). Dynamic latent trait models for multidimensional longitudinal data. *Journal of American Statistical Association*, *98*, 555–563.

Dunson, D. B., & Herring, A. (2005). Bayesian latent variable models for mixed discrete outcomes. *Biostatistics*, *6*, 11–25.

Gelfand, A. E., & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, *85*, 398–409.

Gelman, A. (1996). Inference and monitoring convergence. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice* (pp. 131–144). London: Chapman and Hall.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. London: Chapman & Hall Ltd.

Gelman, A., & Meng, L. (1998). Simulating normalizing constant: from importance sampling to bridge sampling to path sampling. *Statistical Science*, *6*, 733–759.

Gelman, A., Meng, S. L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, *6*, 733–759.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*, 721–741.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their application. *Biometrika*, *57*, 97–109.

Jiang, X., & Mahadevan, S. (2009). Bayesian structural equation modeling method for hierarchical model validation. *Reliability Engineering and System Safety*, *94*, 796–809.

Jöreskong, K. G., & Sörbom, D. (1996). *LISREL 8: structural equation modeling with the SIMPLIS command language*. Hove and London: Scientific Software International.

Kass, R. E., & 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.

Lee, S. Y., & Song, X. Y. (2003). Model comparison of nonlinear structural equation models with fixed covariates. *Psychometrika*, *68*, 27–47.

Lee, S. Y., & Song, X. Y. (2004). Evaluation of the Bayesian and maximum likelihood approaches in analyzing structural equation models with small sample sizes. *Multivariate Behavioral Research, 39*, 653–686.

Lindley, D. V., & Smith, A. F. M. (1972). Bayes estimates for the linear model (with discussion). *Journal of the Royal Statistical Society, Series B, 34*, 1–42.

Lodewyckx, T., Kim, W., Lee, M. D., Tuerlincks, F., Kuppens, P., & Wagenmakers, E. (2011). A tutorial on Bayes factor estimation with the product space method. *Journal of Mathematical Psychology*, *55*, 331–347.

Lunn, D. J., Thomas, A., Best, N. G., & Spiegelhalter, D. J. (2000). WinBUGS–a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, *10*, 325–337.

Meng, X. L., & Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica*, *6*, 831–860.

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

Morey, R. D., Rouder, J. N., Pratte, M. S., & Speckman, P. L. (2011). Using MCMC chain outputs to efficiently estimate Bayes factors. *Journal of Mathematical Psychology*, *55*, 368–378.

Song, X. Y., & Lee, S. Y. (2002). Analysis of structural equation model with ignorable missing continuous and polytomous data. *Psychometrika*, *67*, 261–288.

Song, X. Y., & Lee, S. Y. (2004). Bayesian analysis of two-level nonlinear structural equation models with continuous and polytomous data. *British Journal of Mathematical and Statistical Psychology*, *57*, 29–52.

Song, X. Y., Lee, S. Y., Ma, R. C. W., So, W. Y., Cai, J. H., Ying, W., et al. (2009). Phenotype–genotype interactions on renal function in type 2 diabetes—an analysis using structural equation modeling. *Diabetologia*, *52*, 1543–1553.

Song, X. Y., Xia, Y. M., & Lee, S. Y. (2009). Bayesian semiparametric analysis of structural equation models with mixed continuous and unordered categorical variables. *Statistics in Medicine*, *28*, 2253–2276.

Song, X. Y., Lu, Z. H., Hser, Y. I., & Lee, S. Y. (2011). A Bayesian approach for analyzing longitudinal structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *18*, 183–194.

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

Spiegelhalter, D. J., Thomas, A., Best, N. G., Lunn, D., (2003). WinBUGS user manual. Version 1.4. Cambridge, England: MRC Biostatistics Unit.

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

van Onna, M. J. H. (2002). Bayesian estimation and model selection in ordered latent class models for polytomous items. *Psychometrika*, *67*, 519–538.

Wu, X. L., Heringstad, B., & Gianola, D. (2010). Bayesian structural equation models for inferring relationships between phenotypes: a review of methodology, identifiability, and applications. *Journal of Animal Breeding and Genetics*, *127*, 3–15.

Yang, M. G., & Dunson, D. B. (2010). Bayesian semiparametric structural equation models with latent variables. *Psychometrika*, *75*, 675–693.

Zeger, S. L., & Karim, M. R. (1991). Generalized linear models with random effects: a Gibbs sampling approach. *Journal of the American Statistical Association*, *86*, 79–86.

Zellner, A. (1971). *An introduction to Bayesian inference in econometrics*. New York: John Wiley & Sons, Inc.