Bayesian model comparison of nonlinear structural equation models with missing continuous and ordinal categorical data

Sik-Yum Lee and Xin-Yuan Song

1Department of Statistics, The Chinese University of Hong Kong
2Department of Statistics, Zhongshan University, China

Missing data are very common in behavioural and psychological research. In this paper, we develop a Bayesian approach in the context of a general nonlinear structural equation model with missing continuous and ordinal categorical data. In the development, the missing data are treated as latent quantities, and provision for the incompleteness of the data is made by a hybrid algorithm that combines the Gibbs sampler and the Metropolis–Hastings algorithm. We show by means of a simulation study that the Bayesian estimates are accurate. A Bayesian model comparison procedure based on the Bayes factor and path sampling is proposed. The required observations from the posterior distribution for computing the Bayes factor are simulated by the hybrid algorithm in Bayesian estimation. Our simulation results indicate that the correct model is selected more frequently when the incomplete records are used in the analysis than when they are ignored. The methodology is further illustrated with a real data set from a study concerned with an AIDS preventative intervention for Filipina sex workers.

1. Introduction

Structural equation models (SEMs) are useful for analysing correlations and causation among observed and latent variables (see Bentler, 1992; Bollen, 1989; Jöreskog & Sörbom, 1996; Legler & Ryan, 1997; Sammel & Ryan, 1996; and Sammel, Ryan, & Legler, 1997). There are now more than a dozen user-friendly packages to cope with the strong demand in various fields. These packages and their associated theories are developed for latent variables related by linear functions. However, it is recognized that nonlinear relations among the variables are important for developing more correct and meaningful models (see Kenny & Judd, 1984; Bagozzi, Baumgartner, & Yi, 1992; Schumacker and Marcoulides, 1998; and references therein). Recently, the nonlinear structural equation model (NSEM) has received a lot of attention. Simple estimation methods that apply the
LISREL program (Jöreskog & Sörbom, 1996) have been proposed (see, for example, Jaccard & Wan, 1995; Ping, 1996; Jöreskog & Yang, 1996). A two-stage estimation method has been developed by Bollen and Paxton (1998). More statistically sound approaches have also been developed for fitting some NSEMs (see, for example, Arminger & Mutheén, 1998; Zhu & Lee, 1998; Wall & Amemiya, 2000). The papers cited above deal with continuous data. Motivated by the fact that a lot of practical data are ordinal categorical (polytomous), Lee and Zhu (2000) developed a Bayesian approach for estimating NSEMs with mixed continuous and ordinal categorical variables.

Missing data are very common in applications. It is well recognized that the contribution of the incomplete data should be taken into account because of the threat of informative missingness and because information contained in the incomplete record might otherwise be discarded. In general, the analysis of data sets that have incomplete records has a long history and is still an active research area in statistics; see, for example, the multiple imputation method (Rubin, 1987) and the other techniques in Little and Rubin (1987). However, contributions to SEMs with missing data are comparatively limited. Existing methods (see Lee, 1986; Allison, 1987; Jamshidian & Bentler, 1999) are inefficient for handling missing data sets that involve a large number of missing patterns, and they only deal with linear SEMs based on continuous data. Nothing has yet been published on NSEMs with missing ordinal categorical data.

In this paper, we are concerned with the analysis of NSEMs with missing continuous and ordinal categorical data. We will assess the effect of data missing at random (MAR) with an ignorable mechanism. More importantly, we wish to investigate the contribution of the incomplete data to model comparison (or equivalently, model selection and hypothesis testing). In particular, we wish to study whether the simple approach of ignoring the incomplete data leads to the selection of an inferior model. To achieve our main goals, we develop an approach for comparison of NSEMs on the basis of the Bayes factor (see Berger, 1985; Kass & Raftery, 1995) which is well recognized as an important tool in Bayesian model comparison. The computation of a Bayes factor is generally difficult (see, for example, Kass & Raftery, 1995; DiCiccio, Kass, Raftery, & Wasserman, 1997), especially for complicated NSEMs with a large number of unknown parameters and latent variables, as well as many latent and missing quantities. In the implementation of the procedure, observations from the appropriate posterior distribution have to be sampled. This is done by a hybrid algorithm that combines the Gibbs sampler (Geman & Geman, 1984) and the Metropolis–Hastings (MH) algorithm (Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953; Hastings, 1970). It will be shown that the Bayesian estimates of the unknown parameters in the selected model can be obtained as by-products from the simulated observations.

The present work is motivated by a study (Morisky et al., 1998) of the effects of establishment policies, knowledge and attitudes to condom use among Filipina sex workers. Since the nature of sex work promotes the spread of AIDS and other sexually transmitted diseases, promotion of safer sexual practice among sex workers is an important issue. The primary concern is with inferences from an AIDS preventative intervention for Filipina sex workers. The data set was collected from female sex workers in cities of the Philippines. The entire questionnaire consisted of 134 items, covering attitudes, beliefs, behaviours, self-efficacy for condom use, and social desirability. Latent psychological determinants such as sex workers’ risk behaviours, knowledge and attitudes associated with AIDS and condom use are important issues to be assessed and will be used to modify the planned intervention. A part of this data set has been used by Shi and Lee (2000) to illustrate the maximum likelihood estimation of
the linear factor analysis model. For simplicity, they analysed only the fully observed data. As some embarrassing questions were involved, we find a large amount of missing data. Hence it is desirable to make use of the incomplete records. Moreover, it is also necessary to extend the simple linear factor analysis model to a more general NSEM for finding a better model to fit the data.

Section 2 provides the description of an NSEM with missing continuous and ordinal categorical data. Bayesian posterior simulation in the context of the proposed model is discussed in Section 3. A procedure for computing the Bayes factor for model comparison is developed in Section 4. Results of a simulation study to examine the empirical performance of the proposed procedure are reported in Section 5. A real example is presented in Section 6, and a discussion is given in Section 7.

2. NSEM with missing continuous and ordinal categorical data

Consider a $p \times 1$ random vector $\mathbf{u}_i$ that satisfies the measurement equation

$$\mathbf{u}_i = \mathbf{\mu} + \Lambda \mathbf{\omega}_i + \mathbf{e}_i, \quad i = 1, \ldots, n,$$

where $\mathbf{\mu}$ is a mean vector, $\Lambda$ is an unknown parameter matrix, $\mathbf{\omega}_i$ is a $q \times 1$ vector of latent variables, $\mathbf{e}_i$ is a vector of error measurements with distribution $N(0, \varPsi_\epsilon), \varPsi_\epsilon = \text{diag}(\psi_{e1}, \ldots, \psi_{eP})$, and $\mathbf{\omega}_i$ and $\mathbf{e}_i$ are independent. It is assumed that $\{\mathbf{u}_i, i = 1, \ldots, n\}$ are mutually independent. To handle more complex situations, the latent vector $\mathbf{\omega}_i$ is partitioned into subvectors $(\mathbf{\eta}_i^T, \mathbf{\xi}_i^T)^T$ which satisfy the nonlinear structural equation

$$\mathbf{\eta}_i = \mathbf{B}\mathbf{\eta}_i + \mathbf{\Gamma}\mathbf{F}(\mathbf{\xi}_i) + \mathbf{\delta}_i,$$

where $\mathbf{\eta}_i$ and $\mathbf{\xi}_i$ are $q_1 \times 1$ and $q_2 \times 1$ latent random vectors, respectively; $\mathbf{F}(\mathbf{\xi}_i) = (f_1(\mathbf{\xi}_i), \ldots, f_m(\mathbf{\xi}_i))^T$ is a vector-valued function with differentiable functions $f_1, \ldots, f_m$ and $m \geq q_2$; and $\mathbf{B}$ and $\mathbf{\Gamma}$ are unknown parameter matrices. Moreover, it is assumed that $|\mathbf{I}_{q_1} - \mathbf{B}|$ is positive and independent of $\mathbf{B}, \mathbf{\xi}$ and $\mathbf{\delta}$ are independently distributed as $N(0, \varPsi)$ and $N(0, \varPsi_\delta)$, respectively; and $\varPsi_\delta = \text{diag}(\psi_{\delta1}, \ldots, \psi_{\delta q_1})$. Let $\varPi = (\mathbf{B}, \mathbf{\Gamma}), \ \varTheta(\mathbf{\omega}_i) = (\mathbf{\eta}_i^T, \mathbf{\xi}_i^T)^T$; the structural equation (2) can be written as:

$$\mathbf{\eta}_i = \varPi \varTheta(\mathbf{\omega}_i) + \mathbf{\delta}_i.$$

This model, with nonlinear causal effects among latent variables, is a generalization of the factor analysis model considered in Shi and Lee (2000).

Without loss of generality, suppose $\mathbf{u}_i = (\mathbf{x}_i^T, \mathbf{y}_i^T)^T$, where $\mathbf{x}_i$ is an $r \times 1$ vector of observed measurements and $\mathbf{y}_i$ is an $s \times 1$ vector of unobserved measurements. The information of $\mathbf{y} = (y_1, \ldots, y_s)^T$ is given by an observable ordinal categorical vector $\mathbf{z}$ such that

$$\mathbf{z} = \begin{pmatrix}
\begin{array}{c}
z(1) \\
\vdots \\
z(s)
\end{array}
\end{pmatrix}
\begin{array}{ll}
\text{if } \alpha_{1, z(1)} < y(1) \leq \alpha_{1, z(1) + 1}, \\
\vdots \\
\text{if } \alpha_{s, z(s)} < y(s) \leq \alpha_{s, z(s) + 1},
\end{array}$$

where $z(k)$ is an integer value in the set $\{0, 1, \ldots, b_k\}$ for $k = 1, \ldots, s, \ \alpha_{k, 0} = -\infty$ and $\alpha_{k, b_k + 1} = \infty$. Hence, for the $k$th variable, there are $b_k + 1$ categories which are defined by unknown threshold parameters $\alpha_{k,j}$. 


Without imposing any identification conditions, models with ordinal categorical variables are overparameterized. To solve this problem, we fix $\alpha_{k,1}$ and $\alpha_{k,b_k}$, $k = 1, \ldots, s$, at preassigned values. To identify the covariance structure of $u_i$, we follow the common practice of fixing appropriate elements in $\mu$, $\Lambda$ and $\Pi$ at preassigned values. Moreover, the choice of $F(\xi)$ is not arbitrary. We assume that there are no dependent columns in the matrix $\partial F(\xi)/\partial \xi$.

To deal with the missing-data problem, let $x_i = \{x_{i,obs}, x_{i,mis}\}$ and $z_i = \{z_{i,obs}, z_{i,mis}\}$, where $x_{i,obs}$ and $z_{i,obs}$ represent the observed data, while $x_{i,mis}$ and $z_{i,mis}$ represent the missing data. We assume that missing data are MAR with an ignorable mechanism (Little & Rubin, 1987). For a fully observed $\{x_i, z_i\}$ data point, $x_{i,mis}$ and $z_{i,mis}$ are empty. Let $y_i = \{y_{i,obs}, y_{i,mis}\}$ represent the latent continuous measurements, where $y_{i,obs}$ and $y_{i,mis}$ correspond to $z_{i,obs}$ and $z_{i,mis}$, respectively. Let $u_{i,obs} = \{x_{i,obs}, y_{i,obs}\}$ and $u_{i,mis} = \{x_{i,mis}, y_{i,mis}\}$, so $u_i = \{u_{i,obs}, u_{i,mis}\}$. Bayesian analysis of the proposed NSEM will be studied on the basis of the observed data set $\{(x_{i,obs}, z_{i,obs}); i = 1, \ldots n\}$.

3. Simulation of the posterior distribution

An integral part of the proposed procedure for model comparison is the simulation of observations from the appropriate posterior distribution. By way of preparation, posterior simulation of the current NSEM is discussed here. Since it is closely related to estimation, we approach this issue as an estimation problem.

Let $\theta$ be the vector of structural parameters that involves unknown parameters in $\mu$, $\Lambda$, $\Psi$, $\Pi$ and $\Psi_b$, and $\alpha$ be the vector of unknown thresholds in an identified NSEM. In a Bayesian approach, $\theta$ and $\alpha$ are regarded as random. Let $p(\theta, \alpha)$ be the prior density of $\theta$ and $\alpha$, and let $X_{obs} = \{x_{i,obs}; i = 1, \ldots, n\}$ and $Z_{obs} = \{z_{i,obs}; i = 1, \ldots, n\}$ be observed continuous and ordinal categorical data, respectively. A common way of obtaining the Bayesian estimates of $\theta$ and $\alpha$ is from the expectation of their joint posterior distribution with given $X_{obs}$ and $Z_{obs}$. The joint posterior density of $\alpha$ and $\theta$ given $X_{obs}$ and $Z_{obs}$ is

$$p(\theta, \alpha|X_{obs}, Z_{obs}) \propto p(\theta, \alpha)p(X_{obs}, Z_{obs}|\theta, \alpha).$$

Owing to the complexity induced by the nonlinear structural equation and the existence of the missing and ordinal categorical data, the likelihood $p(X_{obs}, Z_{obs}|\theta, \alpha)$ is very complicated. Finding a Bayesian solution via direct evaluation of the expectation of this posterior density is difficult. To solve the problem, the posterior simulation will be conducted using the technique of data augmentation (Tanner & Wong, 1987). Let $X_{mis} = \{x_{i,mis}; i = 1, \ldots, n\}$, $Y_{obs} = \{y_{i,obs}; i = 1, \ldots, n\}$ and $Y_{mis} = \{y_{i,mis}; i = 1, \ldots, n\}$ be the latent data, and $\Omega = \{\omega_i; i = 1, \ldots, n\}$ be the collection of latent variables. Further, let $U_{obs} = \{u_{i,obs}; i = 1, \ldots, n\}$, $U_{mis} = \{u_{i,mis}; i = 1, \ldots, n\}$, $X = \{X_{obs}, X_{mis}\}$, $Y = \{Y_{obs}, Y_{mis}\}$ and $U = \{U_{obs}, U_{mis}\}$. In addition to the real missing data, we treat $Y_{mis}$ and $\Omega$ as missing data. The observed data $\{X_{obs}, Z_{obs}\}$ are augmented with the missing quantities $\{\Omega, X_{mis}, Y_{mis}, Y_{obs}\} = \{\Omega, U_{mis}, Y_{obs}\}$ in the posterior analysis. A sufficiently large number of random observations will be simulated from the joint posterior distribution $[\theta, \Omega, U_{mis}, \alpha, Y_{obs}|X_{obs}, Z_{obs}]$ so that it can be approximated adequately from the empirical distribution of the generated observations. The following Gibbs sampler (Geman & Geman, 1984) will be used. At the $j$th iteration, with current values $\theta^{(j)}, \Omega^{(j)}, U_{mis}^{(j)}, \alpha^{(j)}$, and $Y_{obs}^{(j)}$, we generate $\theta^{(j+1)}$ from $p(\theta|X_{obs}, Z_{obs}, \Omega^{(j)}, U_{mis}^{(j)}, \alpha^{(j)}, Y_{obs}^{(j)})$; $\Omega^{(j+1)}$, $U_{mis}^{(j+1)}$, $\alpha^{(j+1)}, Y_{obs}^{(j+1)}$, and $\Omega^{(j+1)}$ from $p(\Omega|X_{obs}, Z_{obs}, \theta^{(j+1)}, U_{mis}^{(j+1)}, \alpha^{(j+1)}, Y_{obs}^{(j+1)}), U_{mis}^{(j+1)}$ from
that is, $\frac{1}{2}$ are the elements in values with probability 1.0. For brevity, we only present the results for the case where all identifying the model can be regarded as distributions concentrating on the preassigned investigated by Lee and Zhu (2000). According to their suggestion, the following can be shown.

The conditional distributions required in the implementation of the Gibbs sampler can be obtained from (7) and (8).
it follows from (1) that
\[
p(U_{\text{mis}}|X_{\text{obs}}, Z_{\text{obs}}, \Omega, \theta, \alpha, Y_{\text{obs}}) = \prod_{i=1}^{n} p(u_{i,\text{mis}}|\theta, \omega_i),
\]
where \(u_{i,\text{mis}}\) is a \(p_i \times 1\) subvector of \(\mu\) with elements corresponding to observed components deleted, \(\Lambda_{i,\text{mis}}\) is a \(p_i \times q\) submatrix of \(\Lambda\) with rows corresponding to observed components deleted, and \(\Psi_{i,\text{mis}}\) is a \(p_i \times p_i\) submatrix of \(\Psi\) with the appropriate rows and columns deleted. Hence, even though the form of \(U_{\text{mis}}\) is complicated with many distinct missing patterns, its conditional distribution only involves a product of very simple univariate normal distributions. The computational burden of simulating \(U_{\text{mis}}\) is light.

Finally, we consider the joint conditional distribution of \((\alpha, Y_{\text{obs}})\) given \(X_{\text{obs}}, Z_{\text{obs}}, \Omega, U_{\text{mis}}\) and \(\theta\). In the common situation with little or no information about the thresholds, the following non-informative prior distribution is used:
\[
p(\alpha_k) = p(\alpha_{k,2}, \ldots, \alpha_{k,b_k-1}) \propto \text{constant}, \quad \text{for } \alpha_{k,2} < \ldots < \alpha_{k,b_k-1}, \quad k = 1, \ldots, s.
\]
Since \(\Psi\) is diagonal, \((\alpha, Y_{\text{obs}})\) is conditionally independent of \(X_{\text{obs}}\) and \(U_{\text{mis}}\) with \(\Omega\) and \(\theta\) given. Let \(Z_{k,\text{obs}}\) be the \(k\)th row of \(Z_{\text{obs}}\), \(Y_{k,\text{obs}}\) be the \(k\)th row of \(Y_{\text{obs}}\) corresponding to \(Z_{k,\text{obs}}\), and \(\mu_k\) be the \(k\)th row of \(\mu\). Given \(\Omega\) and \(\theta\), since \(\Psi\) is diagonal, \((\alpha_k, Y_{k,\text{obs}})\) is also conditionally independent of \((\alpha_0, Y_{0,\text{obs}})\) for \(k \neq b\). Hence, we only need to derive marginal conditional distributions of \((\alpha_k, Y_{k,\text{obs}})\) for \(k = 1, \ldots, p\). Since the observations are independent, \(p(\alpha_k, Y_{k,\text{obs}}|Z_{k,\text{obs}}, \Omega, \theta)\) is equal to
\[
p(\alpha_k|Z_{k,\text{obs}}, \Omega, \theta)p(Y_{k,\text{obs}}|\alpha_k, Z_{k,\text{obs}}, \Omega, \theta) = p(\alpha_k|Z_{k,\text{obs}}, \Omega, \theta) \prod_{i=1}^{n} p(y_{k,\text{obs},i}|\alpha_k, z_{k,\text{obs},i}, \omega_i, \theta),
\]
where
\[
p(\alpha_k|Z_{k,\text{obs}}, \Omega, \theta) \propto \prod_{i=1}^{n_{z,k}} \Phi^+(\psi_{ek}^{-1/2}[\alpha_{k,z_{k,obs},i+1} - \mu_k - \Lambda_k^T \omega_i])
\]
\[
- \Phi^+(\psi_{ek}^{-1/2}[\alpha_{k,z_{k,obs},i} - \mu_k - \Lambda_k^T \omega_i]),
\]
and
\[
p(y_{k,\text{obs},i}|\alpha_k, z_{k,\text{obs},i}, \omega_i, \theta) \overset{D}{=} N(\mu_k + \Lambda_k^T \omega_i, \psi_{ek})I_{\Delta, \theta}(y_{k,\text{obs},i}),
\]
where \(n_{z,k}\) is the number of \(z_{k,\text{obs}}\) in \(Z_{k,\text{obs}}, I_{\Delta, \theta}(y)\) is an index function which takes 1 if \(y \in A\) and 0 otherwise, and \(\Phi^+(\cdot)\) is the cumulative distribution function of \(N[0, 1]\).

Simulating observations from the familiar normal, gamma and inverted Wishart distributions that are involved in \(p(\theta|X_{\text{obs}}, Z_{\text{obs}}, \Omega, U_{\text{mis}}, \alpha, Y_{\text{obs}})\) and \(p(U_{\text{mis}}|X_{\text{obs}}, Z_{\text{obs}}, \Omega, U_{\text{mis}}, \alpha, Y_{\text{obs}})\) is straightforward. However, conditional densities involved in \(p(\Omega|X_{\text{obs}}, Z_{\text{obs}}, \theta, U_{\text{mis}}, \alpha, Y_{\text{obs}})\) and \(p(\alpha, Y_{\text{obs}}|Z_{\text{obs}}, \theta, \Omega)\) are non-standard and complex. The well-known Metropolis–Hastings algorithm (Metropolis et al., 1953; Hastings, 1970) is used to simulate observations from these conditional distributions. Hence, the proposed algorithm for posterior simulation is a hybrid algorithm that combines the Gibbs sampler and the MH algorithm. Technical details concerning implementation of the MH algorithm are presented in Appendix B.

Let \((\theta^{(j)}, \Omega^{(j)}, U_{\text{mis}}^{(j)}, \alpha^{(j)}, Y_{\text{obs}}^{(j)}, j = 1, \ldots, J)\) be random observations collected after the convergence of the hybrid algorithm. Joint Bayesian estimates of \(\theta, \alpha\) and \(\Omega\) as well
as their covariance matrices can be obtained easily as the corresponding sample means and sample covariance matrices. For example,

\[
\hat{\theta} = J^{-1} \sum_{j=1}^{J} \theta^{(j)}, \quad \text{Cov}(\theta) = (J - 1)^{-1} \sum_{j=1}^{J} (\theta^{(j)} - \hat{\theta})(\theta^{(j)} - \hat{\theta})^T.
\]

To save space, other related statistics are not discussed.

4. Bayesian model comparison

In analysing NSEMs, a useful method for comparing models uses the Bayes factor (Berger, 1985). Let \( M_0 \) and \( M_1 \) be two competing NSEMs, and \( D \) be the observed data. The Bayes factor is defined as

\[
B_{10} = \frac{p(D|M_1)}{p(D|M_0)};
\]

it is the summary of evidence provided by the data in favour of \( M_1 \) as opposed to \( M_0 \). As pointed out by Kass and Raftery (1995), this approach has several key advantages over the commonly used classical significance tests with \( p \)-values: first, it does not tend to reject the null hypothesis frequently with large sample sizes; second, it provides not only a measure of evidence against the null model but also a measure of support for the alternative model; and third, it can be used to compare non-nested models. Kass and Raftery (1995) also suggested the following logarithmic scale for interpreting the Bayes factor:

<table>
<thead>
<tr>
<th>( \log B_{10} )</th>
<th>(&lt; 0)</th>
<th>0 to 1</th>
<th>1 to 3</th>
<th>( &gt; 3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Support ( M_0 )</td>
<td>No conclusion</td>
<td>Support ( M_1 )</td>
<td>Strongly support ( M_1 )</td>
<td></td>
</tr>
</tbody>
</table>

The use of these criteria is straightforward for comparing non-nested models. For nested models, where \( M_0 \) is nested in \( M_1 \), we usually have \( P(D|M_0) \leq p(D|M_1) \). This implies that \( \log B_{10} \) is usually non-negative. The inequality \( \log B_{10} < 1 \) may indicate that model \( M_1 \) is not significantly better than the simpler model \( M_0 \), hence \( M_0 \) may be selected.

Very often, the marginal density, \( p(D|M_1) \) or \( p(D|M_0) \), is an intractable multi-dimensional integral and \( B_{10} \) does not have a closed form. Various analytic and numerical approximations have been proposed in the literature; for example, the Laplace method of approximation (Tierney & Kadane, 1986) and its alternative forms (Kass & Vaidyanathan, 1992), reciprocal importance sampling (Gelfand & Dey, 1994), importance sampling and bridge sampling (see DiCiccio et al., 1997), and the procedures using Gibbs outputs (Chib, 1995) and the MH algorithm outputs (Chib & Jeliazkov, 2001). According to the comparative study in DiCiccio et al. (1997), bridge sampling is attractive. Gelman and Meng (1998) developed a method, called path sampling, for computing ratios of normalizing constants of probability models. They showed that path sampling is more general than bridge sampling and hence may be superior for computing the Bayes factor. Inspired by Gelman and Meng (1998), we develop our procedure for computing the Bayes factor on the basis of path sampling with \( D = (X_{\text{obs}}, Z_{\text{obs}}) \).
Direct application of path sampling to the current NSEM with missing continuous and ordered categorical data is difficult. The difficulty is alleviated by augmenting the observed data with the latent quantities $(\Omega, U_{\text{mis}}, Y_{\text{obs}})$. Based on the equality

$$p(\theta, \Omega, U_{\text{mis}}, \alpha, Y_{\text{obs}} | X_{\text{obs}}, Z_{\text{obs}}) = \frac{p(\theta, \Omega, U_{\text{mis}}, \alpha, Y_{\text{obs}}, X_{\text{obs}}, Z_{\text{obs}})}{p(X_{\text{obs}}, Z_{\text{obs}})},$$

the marginal density $p(X_{\text{obs}}, Z_{\text{obs}})$ is treated as the normalizing constant of $p(\theta, \Omega, U_{\text{mis}}, \alpha, Y_{\text{obs}} | X_{\text{obs}}, Z_{\text{obs}})$. Consider the following class of densities defined by a continuous parameter $t$ in $[0, 1]$:

$$p(\theta, \Omega, U_{\text{mis}}, \alpha, Y_{\text{obs}} | X_{\text{obs}}, Z_{\text{obs}}, t) = p(\theta, \Omega, U_{\text{mis}}, \alpha, Y_{\text{obs}}, X_{\text{obs}}, Z_{\text{obs}} | t) / z(t),$$

where $z(t) = p(X_{\text{obs}}, Z_{\text{obs}} | t)$. Let $t$ be a parameter linking the competing models $M_0$ and $M_1$ such that $z(1) = p(X_{\text{obs}}, Z_{\text{obs}} | t = 1) = p(X_{\text{obs}}, Z_{\text{obs}} | M_1)$ and $z(0) = p(X_{\text{obs}}, Z_{\text{obs}} | t = 0) = p(X_{\text{obs}}, Z_{\text{obs}} | M_0)$; then $B_{10} = z(1)/z(0)$. Based on the reasoning given in Gelman and Meng (1998), it can be shown (see Appendix C) that the logarithm of the Bayes factor is estimated as follows:

$$\log \hat{B}_{10} = \frac{1}{2} \sum_{s=1}^S (t(s+1) - t(s))(\tilde{V}(s+1) + \tilde{V}(s)),$$

where $t(0) = 0 < t(1) < \ldots < t(s) < t(s+1) = 1$, which are fixed grids in $[0, 1]$, and

$$\tilde{V}(s) = J^{-1} \sum_{j=1}^J V(\theta^{(j)}, \Omega^{(j)}, U_{\text{mis}}^{(j)}, \alpha^{(j)}, Y_{\text{obs}}^{(j)}, X_{\text{obs}}^{(j)}, Z_{\text{obs}}^{(j)}, t^{(s)}),$$

in which $\{ (\theta^{(j)}, \Omega^{(j)}, U_{\text{mis}}^{(j)}, \alpha^{(j)}, Y_{\text{obs}}^{(j)}, X_{\text{obs}}^{(j)}, Z_{\text{obs}}^{(j)}, t^{(s)}) \}$ is a sample of observations simulated from $p(\theta, \Omega, U_{\text{mis}}, \alpha, Y_{\text{obs}} | X_{\text{obs}}, Z_{\text{obs}}, t^{(s)})$ via the hybrid algorithm, and

$$V(\theta, \Omega, U_{\text{mis}}, \alpha, Y_{\text{obs}}, X_{\text{obs}}, Z_{\text{obs}}, t) = \frac{d}{dt} \log p(\Omega, U_{\text{mis}}, Y_{\text{obs}}, X_{\text{obs}}, Z_{\text{obs}} | \theta, \alpha, t).$$

This procedure has some nice features. First, the implementation is simple even for the current complicated NSEMs with many unknown parameters, latent and missing quantities. Since the observations required in the computation can be simulated by the hybrid algorithm developed in the estimation, little extra programming is necessary. The prior density of $\theta$ is not directly involved in the computation, so it may be less sensitive to prior inputs in some applications. Moreover, $\log B_{10}$ is computed, which is generally more stable than $B_{10}$ itself. Finally, the Bayesian estimates and their standard errors estimates under $M_0$ and $M_1$ can be obtained easily as by-products from the simulated observations at $t = 0$ and $t = 1$; see (12).

As pointed out by Gelman and Meng (1998), one can always construct a path to link two unnormlized densities with the same support. This makes the proposed procedure flexible in handling various NSEMs. As an illustration, consider its application to the following competing NSEMs:

$$M_0: u_i = \mu_0 + \Lambda_0 \omega_i + \epsilon_i, \quad \eta_i = \Pi_0 G_0(\omega_i) + \delta_i, \quad i = 1, \ldots, n;$$

$$M_1: u_i = \mu_1 + \Lambda_1 \omega_i + \epsilon_i, \quad \eta_i = \Pi_1 G_1(\omega_i) + \delta_i, \quad i = 1, \ldots, n,$$

where $\{\mu_0, \Lambda_0, \Pi_0, G_0(\omega)\}$ and $\{\mu_1, \Lambda_1, \Pi_1, G_1(\omega)\}$ are two sets of parameters and functions of $\omega$ that associate with $M_0$ and $M_1$, respectively. In general, some components
in one set may be equal to or different from the corresponding components in the other set. These models can be linked up by $t \in [0, 1]$ as follows:

$$
M_t : \mathbf{u}_i = (1 - t)[\mathbf{\mu}_0 + \Lambda_0 \omega_i] + t[\mathbf{\mu}_1 + \Lambda_1 \omega_i] + \mathbf{\epsilon}_i,
$$

$$
\eta_i = (1 - t)\Pi_0 \mathbf{G}_0(\omega_i) + t\Pi_1 \mathbf{G}_1(\omega_i) + \delta_i, \quad i = 1, \ldots, n.
$$

Clearly, when $t = 0$, $M_t = M_0$; and when $t = 1$, $M_t = M_1$. It follows from (1) and (3) that $\log p(\Omega, \mathbf{U}_{\text{mis}}, \mathbf{Y}_{\text{obs}}, \mathbf{X}_{\text{obs}}, \mathbf{Z}_{\text{obs}} | \theta, \alpha, t)$ is equal to

$$
\sum_{i=1}^n \left[ C - \frac{1}{2} \{ \mathbf{\epsilon}_i(t)^T \mathbf{\Psi}_\epsilon^{-1} \mathbf{\epsilon}_i(t) + \delta_i(t)^T \mathbf{\Psi}_\delta^{-1} \delta_i(t) \} \right], \quad (16)
$$

where $C$ is a constant independent of $t$, $\mathbf{\epsilon}_i(t) = \mathbf{u}_i - (1 - t)(\mathbf{\mu}_0 + \Lambda_0 \omega_i) - t(\mathbf{\mu}_1 + \Lambda_1 \omega_i)$, and $\delta_i(t) = \eta_i - (1 - t)\Pi_0 \mathbf{G}_0(\omega_i) - t\Pi_1 \mathbf{G}_1(\omega_i)$. In (16), $\theta$ is the parameter vector in the linked model. It contains all the common and distinct parameters in $\mathbf{\mu}_0, \mathbf{\mu}_1, \Lambda_0, \Lambda_1, \Pi_0, \Pi_1, \Phi, \mathbf{\Psi}_\epsilon$ and $\mathbf{\Psi}_\delta$. Differentiating the above likelihood function with respect to $t$, we have

$$
V(\theta, \Omega, \mathbf{U}_{\text{mis}}, \alpha, \mathbf{Y}_{\text{obs}}, \mathbf{X}_{\text{obs}}, \mathbf{Z}_{\text{obs}}, t) = \mathbf{\epsilon}_i(t)^T \mathbf{\Psi}_\epsilon^{-1} [\mathbf{\mu}_1 + \Lambda_1 \omega_i - \mathbf{\mu}_0 - \Lambda_0 \omega_i]
$$

$$
+ \delta_i(t)^T \mathbf{\Psi}_\delta^{-1} [\Pi_1 \mathbf{G}_1(\omega_i) - \Pi_0 \mathbf{G}_0(\omega_i)]. \quad (17)
$$

The Bayes factor can be computed from (14), (15), and (17) with a sample \{$(\theta^{(j)}, \Omega^{(j)}, \mathbf{U}_{\text{mis}}^{(j)}, \alpha^{(j)}, \mathbf{Y}_{\text{obs}}^{(j)})$, $j = 1, \ldots, f$\} simulated by the hybrid algorithm.

It is well known that Bayesian estimation is not sensitive to prior inputs, especially when the sample sizes are large (see Kass & Raftery, 1995). However, the Bayes factor may be more sensitive to priors in model comparison. Moreover, it may be problematic if the prior distributions of the parameters involved in the comparison are non-informative; see Spiegelhalter and Smith (1982) in the context of linear and log-linear models. In our analysis with ordinal categorical variables, we assign non-informative prior distributions to the nuisance threshold parameters that are not directly included in the competing models for comparison. As pointed out by Kass and Raftery (1995), the choice of non-informative prior for the nuisance parameters does not greatly affect the results. We expect that the above point is also valid in our approach based on path sampling because, in addition to the arguments in Kass and Raftery (1995), the thresholds and their prior densities are not directly involved in the computation formulæ.

5. A simulation study

The main purpose is to study the effect of missing data in model comparison and estimation. Random observations were generated from an NSEM defined by (1) and (2) with six manifest variables that are related with latent variables $\eta$, $\xi_1$ and $\xi_2$. The specifications in (1) are

$$
\mathbf{\mu} = (0, \ldots, 0)^T, \quad \text{and} \quad \Lambda^T = \begin{bmatrix}
1 & \lambda_{21} & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & \lambda_{42} & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & \lambda_{63}
\end{bmatrix}, \quad (18)
$$

where the 1s and 0s are fixed, and $\lambda_{21}, \lambda_{42}$ and $\lambda_{63}$ are unknown parameters with true values 0.8, 0.7 and 0.8, respectively. The structural equation is given by
\[ \eta = \gamma_1 \xi_1 + \gamma_2 \xi_2 + \gamma_{11} \xi_1^2 + \delta, \]

with the true values of \( \gamma_1, \gamma_2 \) and \( \gamma_{11} \) equal to 0.6, 0.6 and 0.3, respectively. The true variances of \( \varepsilon_1 \) and \( \delta \) are all equal to 0.5, and true values of elements in \( \phi_{jk} \) are \( \phi_{11} = \phi_{22} = 1.0 \), and \( \phi_{12} = 0.2 \). Continuous measurements corresponding to the last two variables were transformed to ordinal categorical data via true thresholds \( \alpha_1 = \alpha_2 = (-1.0, -0.6, 0.6, 1.0) \), where \(-1.0 \) and 1.0 were held fixed.

In each replication, a complete data set with 500 random observations was generated. Then the MAR missing data were created as follows. First, 100 fully observed data points were randomly selected, and the sample means of the first four variables, \( \bar{x}(1), \bar{x}(2), \bar{x}(3) \) and \( \bar{x}(4) \) were computed. Then, for each element \( x(1), x(2), x(3) \) and \( x(4) \), and in each and every of the remaining 400 observations, we randomly generated an observation \( v \) from \( U[0, 1] \) to decide whether the element was missing or not. More specifically, we randomly generated four independent observations \( v(1), v(2), v(3) \) and \( v(4) \) from \( U[0, 1] \), then \( x(1) \) was deleted if \( x(1) + \bar{x}(1) + \bar{x}(2) > v(1) - 1.0 \), otherwise \( x(2) \) was deleted if \( x(2) + \bar{x}(1) + \bar{x}(2) > v(2) - 1.0 \), also \( x(3) \) was deleted if \( x(3) + \bar{x}(1) > v(3) \), otherwise \( x(4) \) was deleted if \( x(4) + \bar{x}(1) > v(4) \). In the missing data sets created, all entries relating to \( y(1) \) and \( y(2) \) were retained, and about two-thirds of the observations contained one or more missing entries in \( x(1), \ldots, x(4) \) (see Table 1).

Table 1. Estimated log \( B_{10} \) in ten replications

<table>
<thead>
<tr>
<th>Replication no.</th>
<th>No. of FOD</th>
<th>ALL</th>
<th>FOD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>I</td>
<td>II</td>
</tr>
<tr>
<td>1</td>
<td>179</td>
<td>6.805</td>
<td>8.201</td>
</tr>
<tr>
<td>2</td>
<td>156</td>
<td>5.523</td>
<td>5.777</td>
</tr>
<tr>
<td>3</td>
<td>172</td>
<td>6.350</td>
<td>7.427</td>
</tr>
<tr>
<td>4</td>
<td>144</td>
<td>2.868</td>
<td>3.005</td>
</tr>
<tr>
<td>5</td>
<td>165</td>
<td>4.694</td>
<td>5.280</td>
</tr>
<tr>
<td>6</td>
<td>161</td>
<td>4.121</td>
<td>5.035</td>
</tr>
<tr>
<td>7</td>
<td>175</td>
<td>3.543</td>
<td>3.880</td>
</tr>
<tr>
<td>8</td>
<td>193</td>
<td>5.530</td>
<td>6.730</td>
</tr>
<tr>
<td>9</td>
<td>184</td>
<td>7.241</td>
<td>8.550</td>
</tr>
<tr>
<td>10</td>
<td>150</td>
<td>7.365</td>
<td>8.280</td>
</tr>
</tbody>
</table>

‘FOD’ stands for fully observed data, and ‘ALL’ for all data.

We will compare the above NSEM (\( M_1 \)) with a linear SEM (\( M_0 \)) based on the same measurement equation as specified and the linear structural equation 
\[ \eta = \gamma_1 \xi_1 + \gamma_2 \xi_2 + \delta. \]
To give some idea of the sensitivity of the results to prior inputs, three types of hyperparameter values in the conjugate distributions were considered. First, we fixed \( H_{0ek} = I \) and \( H_{0sk} = I \) for all types. Then in type I, we took 
\[ \rho_0 = 10, \quad R_0^{-1} = 4I, \quad \alpha_{0ek} = \alpha_{0sk} = 8, \quad \beta_{0ek} = \beta_{0sk} = 10, \]
for all \( k \); the entry in \( \Pi_{0k} \) that corresponds to \( \gamma_{11} \) is 0, and other values in \( \Lambda_{0ek} \) and \( \Pi_{0k} \) are equal to the true parameters values. Hyperparameter values in type II and type III were respectively equal to half and twice those given in type I. On the basis of the fully observed data, and the fully observed and incomplete data, estimates of the log Bayes factors were obtained via the proposed procedure with \( S = 20 \) and \( J = 1000 \). Results obtained from 10 replications are reported in Table 1. Based on Kass and Raftery’s (1995) criterion for interpreting the Bayes factor, in most cases, results produced from all the
fully observed and incomplete data clearly suggest the correct NSEM, but results obtained from the fully observed data cannot provide a definite suggestion as to the correct NSEM. Hence, based on the Bayes factor, the correct model is selected more frequently when the incomplete records are used. It seems that the conclusions are not sensitive to the selected prior inputs.

To give some idea of the effect of missing data on the accuracy of estimation, Bayesian estimates of the parameters under $M_1$ were obtained in 100 replications, using both types of data sets. The means, the standard deviation (Std), and the root mean squares (RMS) of the estimates and the corresponding true values are reported in Tables 2 and 3. Estimates obtained from fully observed and incomplete data are reasonably accurate, given the fact that a lot of data are missing. The mean values given in Table 2 are closer to the corresponding true population values than those given in Table 3. Moreover, the standard deviation and RMS values reported in Table 2 are much smaller than those given in Table 3. These facts indicate that results obtained with the incomplete data are significantly better.

### Table 2. Bayesian estimation based on all the fully observed and incomplete data

<table>
<thead>
<tr>
<th>True parameter values</th>
<th>Type I</th>
<th>Type II</th>
<th>Type III</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Std</td>
<td>RMS</td>
</tr>
<tr>
<td>$\lambda_{21} = 0.8$</td>
<td>0.744</td>
<td>0.049</td>
<td>0.073</td>
</tr>
<tr>
<td>$\lambda_{42} = 0.7$</td>
<td>0.768</td>
<td>0.061</td>
<td>0.084</td>
</tr>
<tr>
<td>$\lambda_{63} = 0.8$</td>
<td>0.784</td>
<td>0.073</td>
<td>0.066</td>
</tr>
<tr>
<td>$\gamma_1 = 0.6$</td>
<td>0.763</td>
<td>0.099</td>
<td>0.190</td>
</tr>
<tr>
<td>$\gamma_2 = 0.6$</td>
<td>0.685</td>
<td>0.087</td>
<td>0.117</td>
</tr>
<tr>
<td>$\gamma_{11} = 0.3$</td>
<td>0.355</td>
<td>0.064</td>
<td>0.082</td>
</tr>
<tr>
<td>$\phi_{11} = 1.0$</td>
<td>0.867</td>
<td>0.100</td>
<td>0.168</td>
</tr>
<tr>
<td>$\phi_{12} = 0.2$</td>
<td>0.181</td>
<td>0.062</td>
<td>0.064</td>
</tr>
<tr>
<td>$\phi_{22} = 1.0$</td>
<td>1.008</td>
<td>0.130</td>
<td>0.141</td>
</tr>
<tr>
<td>$\psi_{e1} = 0.5$</td>
<td>0.616</td>
<td>0.102</td>
<td>0.139</td>
</tr>
<tr>
<td>$\psi_{e2} = 0.5$</td>
<td>0.540</td>
<td>0.067</td>
<td>0.063</td>
</tr>
<tr>
<td>$\psi_{e3} = 0.5$</td>
<td>0.505</td>
<td>0.064</td>
<td>0.038</td>
</tr>
<tr>
<td>$\psi_{e4} = 0.5$</td>
<td>0.491</td>
<td>0.049</td>
<td>0.039</td>
</tr>
<tr>
<td>$\psi_{e5} = 0.5$</td>
<td>0.594</td>
<td>0.083</td>
<td>0.112</td>
</tr>
<tr>
<td>$\psi_{e6} = 0.5$</td>
<td>0.587</td>
<td>0.065</td>
<td>0.102</td>
</tr>
<tr>
<td>$\psi_{2} = 0.5$</td>
<td>0.755</td>
<td>0.108</td>
<td>0.268</td>
</tr>
<tr>
<td>$\alpha_{12} = -0.6$</td>
<td>-0.602</td>
<td>0.044</td>
<td>0.045</td>
</tr>
<tr>
<td>$\alpha_{13} = 0.6$</td>
<td>0.597</td>
<td>0.044</td>
<td>0.038</td>
</tr>
<tr>
<td>$\alpha_{22} = -0.6$</td>
<td>-0.607</td>
<td>0.041</td>
<td>0.043</td>
</tr>
<tr>
<td>$\alpha_{23} = 0.6$</td>
<td>0.589</td>
<td>0.042</td>
<td>0.040</td>
</tr>
</tbody>
</table>

### 6. An illustrative example with the AIDS data set

By way of illustration, a portion of the real data set in the AIDS study (Morisky et al., 1998) as described in Section 1 is analysed. Six manifest variables ($u_{(1)}, \ldots, u_{(6)}$) were selected. Variables $u_{(1)}$ and $u_{(2)}$ relate to ‘worry about getting AIDS’, $u_{(3)}$ and $u_{(4)}$ to the ‘number of times of vaginal sex in the last seven days’ and the ‘average
weekly money (in pesos) earned as an entertainer’, while \( u_{(5)} \) and \( u_{(6)} \) are about the ‘attitudes to getting AIDS from sexual intercourse using a condom’. Variables \( u_{(3)} \) and \( u_{(4)} \) are continuous, while the others are ordinal categorical measurements on a five-point scale. After deleting obvious outliers, the data set contains 1 080 observations, only 754 of them fully observed. The missing-data patterns are presented in Table 4.

In this illustrative example, we assume that the values are MAR and the missingness mechanism is ignorable. To unify the scales of the continuous variables, the corresponding raw continuous data were standardized. The sample means and standard deviations of the continuous variables are \{1.58, 1.203.74\} and \{1.84, 1.096.32\}, respectively. The cell frequencies of each individual ordinal categorical variable ranged from 21 to 709. See Morisky et al. (1998) for other descriptive statistics.

Following Shi and Lee (2000) in identifying parameters associated with the ordinal categorical variables, \( a_{11}, a_{14}, a_{21}, a_{24}, a_{31}, a_{34}, a_{41} \) and \( a_{44} \) were fixed at \( 2.0.478, 1.034, -1.420, 0.525, -0.868, 0.559, -2.130 \) and \(-0.547\), respectively. These values were selected via

\[
\alpha_{kb} = \Phi^{-1}(f_k),
\]

where the \( f_k \) are observed cumulative marginal proportions of the categories with \( z_k < b \). Based on the meanings of the questions corresponding to the selected manifest variables, the data set was analysed by a model with three latent variables, \( \eta, \xi_1 \) and \( \xi_2 \), and the measurement equation as specified in (18).
Competing models associated with the same measurement equation but the following different structural equations are considered by way of illustration:

\[ M_1 : h = g_{1j1} + g_{2j2} + \delta, \]
\[ M_2 : h = g_{1j1} + g_{2j2} + g_{11j21} + \delta, \]
\[ M_3 : h = g_{1j1} + g_{2j2} + g_{12j1}j2 + \delta, \]
\[ M_4 : h = g_{1j1} + g_{2j2} + g_{22j22} + \delta. \]

Note that \( M_1 \) is nested in \( M_2, M_3 \) and \( M_4 \), while \( M_2, M_3 \) and \( M_4 \) are not nested. Estimated log Bayes factors are obtained by the proposed procedure with \( S = 20 \) and \( J = 1000 \). Assuming that we have no prior information from other sources, we conducted an initial Bayesian estimation based on \( M_1 \) with non-informative priors in order to get prior inputs of some hyperparameters. Three types of prior inputs are considered. Here, prior inputs in type I are the same as those in type I in the simulation study, except that the true values are replaced by the Bayesian estimates obtained from the initial estimation. Prior inputs in type II and type III are obtained in the same way as before.

We are interested in comparing the linear model with the nonlinear models. It is easy to construct a path to link the competing models. For example, the linked model for \( M_1 \) and \( M_2 \) is \( M_t : h = g_{1j1} + g_{2j2} + t\gamma_{11j1}j1 + \delta. \) Hence, when \( t = 0, M_t = M_1; \) and when \( t = 1, M_t = M_2. \) Results obtained on the basis of the 754 fully observed data and the total of 1080 fully observed and incomplete data are presented in the top half of Table 5. The estimated log Bayes factors are not very sensitive to the selected prior inputs. In comparing \( M_2 \) and \( M_1, \) log \( \hat{B}_{21} \) obtained from all the data clearly recommends the nonlinear model \( M_2 \) while log \( \hat{B}_{21} \) obtained from the fully observed data provides no suggestion. From log \( \hat{B}_{31} \) and log \( \hat{B}_{41}, \) the other nonlinear models are not significantly better than \( M_1. \)

To compare \( M_2 \) with more complex models, we consider the same models with the same measurement equation and the following structural equations:

\[ M_5 : h = g_{1j1} + g_{2j2} + \gamma_{11j1}j1^2 + \gamma_{12j1}j2 + \delta, \]
\[ M_6 : h = g_{1j1} + g_{2j2} + \gamma_{11j1}j1^2 + \gamma_{22j22} + \delta. \]
Here, $M_2$ is nested in $M_5$ and $M_6$. The estimated log Bayes factors are presented in the bottom half of Table 5. We observe that the results are not very sensitive to the selected prior inputs. Moreover, it seems that the more complex models are not significantly better than $M_2$, hence the simpler model $M_2$ is selected. We found that the posterior predictive $p$-value (Gelman, Meng, & Stern, 1996) corresponding to $M_2$ is 0.572. This indicates that $M_2$ fits the data well. In almost all cases, the EPSR values in monitoring the convergence of the hybrid algorithm for drawing observations from the posterior distribution are less than 1.2 after about 500 iterations. For completeness, Bayesian estimates of unknown parameters in $M_2$ obtained on the basis of different prior inputs and all the fully observed and incomplete data are reported in Table 6. The estimates of $\lambda_{63}$ and $\phi_{12}$ under the different prior inputs are not close. This indicates that the prior inputs may have an impact on some parameter estimates.

Based on the results obtained, an NSEM was chosen. Its specification of $\Lambda$ in the measurement equation suggests that there are three non-overlapping latent factors, $\eta$, $\xi_1$

Table 5. Estimated log Bayes factors under different prior inputs

<table>
<thead>
<tr>
<th>Priors</th>
<th>FOD</th>
<th>ALL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>I</td>
<td>II</td>
</tr>
<tr>
<td>$\log B_{21}$</td>
<td>0.551</td>
<td>0.864</td>
</tr>
<tr>
<td>$\log B_{31}$</td>
<td>0.296</td>
<td>0.276</td>
</tr>
<tr>
<td>$\log B_{41}$</td>
<td>0.811</td>
<td>0.871</td>
</tr>
<tr>
<td>$\log B_{52}$</td>
<td>0.419</td>
<td>0.496</td>
</tr>
<tr>
<td>$\log B_{62}$</td>
<td>0.364</td>
<td>0.344</td>
</tr>
</tbody>
</table>

‘FOD’ stands for fully observed data, and ‘ALL’ for all data.

Table 6. Bayesian estimates of parameters in $M_2$, under different prior inputs

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Bayesian estimate</th>
<th>Parameter</th>
<th>Bayesian estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda_{31}$</td>
<td>0.228 0.226 0.242</td>
<td>$\psi_{e1}$</td>
<td>0.593 0.510 0.660</td>
</tr>
<tr>
<td>$\lambda_{42}$</td>
<td>0.353 0.318 0.402</td>
<td>$\psi_{e2}$</td>
<td>0.972 0.973 0.975</td>
</tr>
<tr>
<td>$\lambda_{63}$</td>
<td>0.358 0.276 0.405</td>
<td>$\psi_{e3}$</td>
<td>0.519 0.433 0.586</td>
</tr>
<tr>
<td>$\gamma_1$</td>
<td>0.544 0.505 0.538</td>
<td>$\psi_{e4}$</td>
<td>0.943 0.943 0.940</td>
</tr>
<tr>
<td>$\gamma_2$</td>
<td>0.033 0.051 0.076</td>
<td>$\psi_{e5}$</td>
<td>0.616 0.545 0.716</td>
</tr>
<tr>
<td>$\gamma_{11}$</td>
<td>0.226 0.153 0.223</td>
<td>$\psi_{e6}$</td>
<td>1.056 1.086 1.048</td>
</tr>
<tr>
<td>$\phi_{11}$</td>
<td>0.508 0.592 0.442</td>
<td>$\alpha_{12}$</td>
<td>-0.030 -0.031 -0.015</td>
</tr>
<tr>
<td>$\phi_{12}$</td>
<td>0.029 0.028 0.460</td>
<td>$\alpha_{13}$</td>
<td>0.340 0.345 0.359</td>
</tr>
<tr>
<td>$\phi_{22}$</td>
<td>0.394 0.479 0.305</td>
<td>$\alpha_{22}$</td>
<td>-0.961 -0.964 -0.966</td>
</tr>
<tr>
<td>$\psi_{6}$</td>
<td>0.663 0.668 0.659</td>
<td>$\alpha_{23}$</td>
<td>-0.620 -0.623 -0.627</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\alpha_{32}$</td>
<td>-0.394 -0.399 -0.399</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\alpha_{33}$</td>
<td>0.257 0.255 0.256</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\alpha_{42}$</td>
<td>-1.604 -1.613 -1.604</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\alpha_{43}$</td>
<td>-0.734 -0.736 -0.738</td>
</tr>
</tbody>
</table>
and $\xi_2$, which can be roughly interpreted as ‘worry about AIDS’, ‘aggressiveness’ of sex workers, and ‘attitude to the risk of getting AIDS’. These latent factors are related by the nonlinear structural equations $\eta = 0.544 \xi_1 - 0.033 \xi_2 - 0.226 \xi_1^2 + \delta$, estimated with type I prior inputs. Thus, the ‘aggressiveness’ of the sexworkers has both linear and quadratic effects on ‘worrying about AIDS’. Plotting the quadratic curve of $\eta$ against $\xi_1$, we find that the maximum of $\eta$ is roughly at $\xi_1 = 1.2$, and $\eta$ decreases as $\xi_1$ moves away from 1.2 in both directions. Hence, the ‘more aggressive’ sex workers are not afraid of getting AIDS, while the ‘less aggressive’ ones are not worried about having AIDS. Conclusions obtained on the basis of other prior inputs are similar. From the model comparison results, the model with the quadratic term for ‘attitude to the risk of getting AIDS’ or the corresponding interaction term with ‘aggressiveness’ is not as good. Hence, these nonlinear relationships are not important, and it is not necessary to consider the more complicated model that involves both the interaction and quadratic terms. To save space, other less interesting interpretations are not discussed. Other models—for example, those with high-order terms in $\xi_1$ and/or different measurement equations—can be similarly analysed using the proposed method. To save space, these analyses are not included.

7. Discussion

Analysis with missing data has received a great deal of attention in statistics. In analysing NSEMs, taking into account the information carried by missing data is important for the following reasons: first, the data often have a substantial number of missing entries; second, the efficiency of the estimates can be improved substantially; and finally, a better model is likely to be selected. As we have seen from our numerical studies, ignoring the missing data may lead to inferior results and conclusions.

In this paper, a procedure for model comparison is developed on the basis of path sampling, in which a hybrid algorithm that combines the Gibbs sampler and the MH algorithm is used for simulating observations from the posterior distribution for computing the log Bayes factor. Bayesian estimates and other related statistics can be obtained as by-products. In the numerical studies, the number of grids $S$ taken in $[0,1]$ is 20, and the number of simulated observations $J$ for computing $\hat{V}_s$ is 1 000. We have obtained very similar results with $S = 10$ and 50, and $J = 2 000$ and 4 000. On the basis of a limited sensitivity analysis, we get the impression that the results are not very sensitive to the prior distribution in our analyses. However, a more detailed simulation study is required to draw more definite conclusions.

The implementation of the path sampling procedure is rather simple. Its main computational burden is the posterior simulation. Using $S = 20$ and $J = 1 000$, the computer time required to obtain an estimated log Bayes factor in the AIDS example is about 60 minutes, with a SUN Enterprise 4500 Server. The computer time can be reduced with a smaller $S$, say 10. We do not recommend the reduction of $J$ to less than 1 000.

The efficiency of the procedure and the validity of the statistical analysis are not affected by the missing patterns and the sample sizes within them. The procedure can handle various types of missing data with an ignorable missing mechanism. Clearly, development of statistical methods for handling the non-ignorable missing mechanisms is necessary.
Acknowledgements

The work described in this paper was fully supported by grants (CUHK 4243/02H) from the Research Grants Council of the Hong Kong Special Administration Region. We are grateful to the Editor and anonymous reviewers for valuable comments which improved the paper substantially, and to D. E. Morisky and J. A. Stein for providing their AIDS data set.

References


Appendix A: Conditional distribution of $\theta$

Let $\theta_u$ be the vector of unknown parameters in $\mu$, $\Lambda$ and $\Psi$, associated with the measurement equation; and $\theta_w$ be the vector of unknown parameters in $\Pi$, $\Phi$, and $\Psi_b$ associated with the structural equation. It is natural to assume that prior distributions of $\theta_u$ and $\theta_w$ are independent. Hence, $p(U|\Omega, \theta) = p(U|\Omega, \theta_u) \cdot p(\Omega|\theta) = p(\Omega|\theta_w) \cdot p(\Omega|\theta_u)$ and

$$p(\theta_u, \theta_w|U, \Omega) \propto [p(U|\Omega, \theta_u)p(\theta_u)]p(\Omega|\theta_u)p(\theta_w).$$

Since the first term on the right-hand side of the above expression depends only on $\theta_u$ while the second term depends only on $\theta_w$, the marginal conditional densities $p(\theta_u|\Omega)$ and $p(\theta_w|\Omega)$ are proportional to $p(U|\Omega, \theta_u)p(\theta_u)$ and $p(U|\Omega, \theta_w)p(\theta_w)$, respectively. As a result, these conditional densities can be treated separately.

Consider the marginal conditional distribution of $\theta_u$. For $k \neq b$, it is assumed that $(\psi_{ek}, A_{ek})$ and $(\psi_{eh}, A_{eh})$ are independent. Let $U_k$ be the $k$th row of $U$. It can be shown using reasoning similar to that of Shi and Lee (1998), that for $k = 1, \ldots, p$,

$$[\psi_{ek}^{-1}|U, \Omega] \overset{d}{=} \text{Gamma}[n/2 + \alpha_{0ek}, \beta_{ek}], \quad [A_{ek}|U, \Omega, \psi_{ek}] \overset{d}{=} N[A_{ek}^*, \psi_{ek}A_{ek}];$$

$$[\mu|U, \Omega, \Lambda, \Psi] \overset{d}{=} N\left[\Sigma^{-1} + n\Psi^{-1} - 1 \begin{pmatrix} n\Psi^{-1}B + \Sigma^{-1} & \Sigma^{-1} \\ \Sigma^{-1} & \Psi^{-1} \end{pmatrix}^{-1} \begin{pmatrix} \Sigma^{-1} \\ \Psi^{-1} \end{pmatrix}\right],$$

where $A_{ek} = (H_{ek}^{-1} + \Omega \Omega^T)^{-1}$, $A_{ek}^* = A_{ek}[H_{0ek}^{-1}A_{0ek} + \Omega_U^T]$, $\beta_{ek} = \beta_{0ek} + 2^{-1}(U_k^TU_k - A_{ek}^* A_{ek} + \Lambda_{0ek}^{-1}H_{0ek}^{-1}A_{0ek})$ and $B = \Sigma_{i=1}^n (u_i - \Lambda \omega_i)/n$.

Now consider the conditional distribution of $\theta_w$. Let $\Omega_{(1)} = (\eta_1, \ldots, \eta_n)$ and $\Omega_{(2)} = (\xi_1, \ldots, \xi_n)$. It is natural to assume that the prior distribution of $\Phi$ is independent of the prior distribution of $\Pi$ and $\Psi_b$. Since the distribution of $\xi_i$ only involves $\Phi$, it follows that

$$p(\Omega|\theta_w)p(\theta_w) \propto [p(\Omega_{(1)}|\Omega_{(2)}, \Pi, \Psi_b)p(\Pi, \Psi_b)]p(\Omega_{(2)}|\Phi)p(\Phi).$$

Hence, the marginal conditional densities of $(\Pi, \Psi_b)$ and $\Phi$ can again be treated separately. For $k \neq b$, it is assumed that $(\psi_{0ek}, \Pi_{ek})$ and $(\psi_{0ek}, \Pi_{eh})$ are independent. Let $IW_{q_2}[\cdot, \cdot]$ denote an inverted Wishart distribution with dimension $q_2$, $\Omega = (G(\omega_1), \ldots, G(\omega_n))$ and $\Omega_{(1)k}$ the $k$th row of $\Omega_{(1)}$. It can be shown that, for $k = 1, \ldots, q_1$,

$$[\psi_{0ek}^{-1}|\Omega] \overset{d}{=} \text{Gamma}[n/2 + \alpha_{0ek}, \beta_{0ek}], \quad [\Pi_{ek}|\Omega, \psi_{0ek}] \overset{d}{=} N[\Pi_{0ek}^*, \psi_{0ek}A_{0ek}],$$

$$[\Phi|\Omega] = [\Phi|\Omega_{(2)}] \overset{d}{=} IW_{q_2}[\Omega_{(2)}, \Theta_{(2)} + R_0^{-1}], n + \rho_0,$$

where $A_{0ek} = (H_{0ek}^{-1} + \tilde{\Omega}_{0ek} \tilde{\Omega}_{0ek}^T)^{-1}$, $\Pi_{0ek}^* = A_{0ek}[H_{0ek}^{-1}\Pi_{0ek} + \tilde{\Omega}_{0ek}^T\Omega_{(1)k}]$, $\beta_{0ek} = \beta_{0ek} + 2^{-1}(\Omega_{(1)k}^T\Omega_{(1)k} - \Pi_{0ek}^T A_{0ek}^{-1} \Pi_{0ek}^* + \Pi_{0ek}^T H_{0ek}^{-1} \Pi_{0ek}).$
Appendix B: Implementation of the Metropolis–Hastings algorithm

With given $X_{\text{obs}}, Y_{\text{obs}}$ and $U_{\text{mis}}$, the conditional distribution $[\Omega | X_{\text{obs}}, Z_{\text{obs}}, U_{\text{mis}}, \alpha, Y_{\text{obs}}]$ reduces to $[\Omega | U, \theta]$. Details on the implementation of the MH algorithm for drawing observations from this conditional distribution can be found in Lee and Zhu (2000). Hence, we only consider the MH algorithm for drawing observations from the target density $p(\alpha_k, Y_{k,\text{obs}} | Z_{\text{obs}}, \theta, \Omega)$ as given in (10) and (11).

The MH algorithm is a Markov chain Monte Carlo method for simulating observations from a target density with the help of a proposal distribution from which it is easy to sample. Inspired by Cowles (1996), a joint proposal density for $\alpha_k$ and $Y_{k,\text{obs}}$ is constructed as follows:

$$ p(\alpha_k, Y_{k,\text{obs}} | Z_{k,\text{obs}}, \Omega, \theta) = p(\alpha_k | Y_{k,\text{obs}}, \alpha_k, Z_{k,\text{obs}}, \Omega, \theta). $$ (B.1)

The MH algorithm for simulating new values $(\alpha_k^{(j+1)}, Y_{k,\text{obs}}^{(j+1)})$ at the $j$th iteration of the Gibbs sampler is implemented as follows. A candidate vector of thresholds $(\alpha_{k,2}, \ldots, \alpha_{k,b_k-1})$ is first simulated from the following univariate truncated normal distribution:

$$ \alpha_{k,z} \sim N(\alpha_{k,z}^{(j)}, \sigma_{\alpha_k}^2)I(\alpha_{k,z-1}, \alpha_{k,z+1}^{(j)})(\alpha_{k,z}), \quad \text{for } z = 2, \ldots, b_k - 1, $$ (B.2)

where $\alpha_{k,z}^{(j)}$ is the current value of $\alpha_{k,z}$, and $\sigma_{\alpha_k}^2$ is preassigned variance which is chosen to give an approximate acceptance rate of 0.44 (see Cowles, 1996). With this candidate-generating density, the acceptance probability for $(\alpha_k, Y_{k,\text{obs}})$ as a new observation $(\alpha_k^{(j+1)}, Y_{k,\text{obs}}^{(j+1)})$ is min$\{1, R_k\}$, where

$$ R_k = \frac{p(\alpha_{k}^{(j)}, Y_{k,\text{obs}}^{(j)}, Z_{k,\text{obs}}, \Omega, \theta)p(\alpha_{k}^{(j)}, Y_{k,\text{obs}}^{(j)} | \alpha_{k}, Y_{k,\text{obs}}^{(j)} , Z_{k,\text{obs}}, \Omega, \theta)}{p(\alpha_{k}^{(j)}, Y_{k,\text{obs}}^{(j)}, Z_{k,\text{obs}}, \Omega, \theta)p(\alpha_{k}, Y_{k,\text{obs}} | \alpha_{k}^{(j)}, Y_{k,\text{obs}}^{(j)}, Z_{k,\text{obs}}, \Omega, \theta)}. $$

It can be shown from (10), (11), (B.1) and (B.2) that

$$ R_k = \prod_{z=2}^{b_k-1} \frac{\Phi([\alpha_{k,z+1}^{(j)} - \alpha_{k,z}^{(j)}]/\sigma_{\alpha_k}) - \Phi([\alpha_{k,z-1}^{(j)} - \alpha_{k,z}^{(j)}]/\sigma_{\alpha_k})}{\Phi([\alpha_{k,z+1}^{(j)} - \alpha_{k,z}^{(j)}]/\sigma_{\alpha_k}) - \Phi([\alpha_{k,z-1}^{(j)} - \alpha_{k,z}^{(j)}]/\sigma_{\alpha_k})} \times \prod_{i=1}^{n_2} \Phi([\psi_{ek}^{-1/2} | \alpha_{k,z_{2,obs,i}+1}^{(j)} - \mu_k - \Lambda^T_k \omega_i] - \Phi([\psi_{ek}^{-1/2} | \alpha_{k,z_{2,obs,i}+1}^{(j)} - \mu_k - \Lambda^T_k \omega_i]) \times \Phi([\psi_{ek}^{-1/2} | \alpha_{k,z_{2,obs,i}}^{(j)} - \mu_k - \Lambda^T_k \omega_i]) - \Phi([\psi_{ek}^{-1/2} | \alpha_{k,z_{2,obs,i}}^{(j)} - \mu_k - \Lambda^T_k \omega_i]).$$

Since $R_k$ depends only on the old and new values of $\alpha_k$ and not on $Y_{k,\text{obs}}$, there is no requirement to generate a new $Y_{k,\text{obs}}$ in any MH iteration in which the new value of $\alpha_k$ is not accepted. For an accepted $\alpha_k$, a new $Y_{k,\text{obs}}$ is simulated from the univariate truncated normal distribution in (11) using the algorithm proposed by Roberts (1995).

Appendix C: Computation of the Bayes factor via path sampling

Let $D = (X_{\text{obs}}, Z_{\text{obs}})$, and $p(\theta, \alpha)$ be the prior density of $\theta$ and $\alpha$ which is independent of $t$. It follows from (13) that

$$ z(t) = p(D | t) = \int p(\theta, \Omega, U_{\text{mis}}, \alpha, Y_{\text{obs}}, D | t) d\theta d\Omega dU_{\text{mis}} d\alpha dY_{\text{obs}} $$

$$ = \int p(\Omega, U_{\text{mis}}, Y_{\text{obs}}, D | \theta, \alpha, t) p(\theta, \alpha) d\theta d\Omega dU_{\text{mis}} d\alpha dY_{\text{obs}.} $$

Bayesian model comparison of nonlinear structural equation models 149
Taking the logarithm and differentiating with respect to $t$, it can be shown by exactly the same arguments as in Gelman and Meng (1998) that

$$\frac{d \log z(t)}{dt} = E^* \frac{d \log p(\theta, \Omega, U_{\text{mis}}, \alpha, Y_{\text{obs}}, D|t)}{dt}, \quad \text{(C.1)}$$

where $E^*$ denotes the expectation with respect to the distribution of $p(\theta, \Omega, U_{\text{mis}}, \alpha, Y_{\text{obs}}|D, t)$. Since $p(\theta, \alpha)$ is independent of $t$,

$$V(\theta, \Omega, U_{\text{mis}}, \alpha, Y_{\text{obs}}, D, t) = \frac{d}{dt} \log p(\theta, \Omega, U_{\text{mis}}, \alpha, Y_{\text{obs}}, D|t)$$

$$= \frac{d}{dt} \log p(\Omega, U_{\text{mis}}, Y_{\text{obs}}, D|\theta, \alpha, t).$$

It follows from (C.1) that

$$\log B_{10} = \log \frac{z(1)}{z(0)} = \int_0^1 E^*[V(\theta, \Omega, U_{\text{mis}}, \alpha, Y_{\text{obs}}, D, t)] dt.$$

By numerically evaluating the integral using the trapezoidal rule, $\log B_{10}$ can be approximated as in (14).