Hypothesis Testing and Model Comparison in Two-level Structural Equation Models

Sik-Yum Lee and Xin-Yuan Song
Department of Statistics
The Chinese University of Hong Kong

One basic and important problem in two-level structural equation modeling is to find a good model for the observed sample data. This article demonstrates the use of the well-known Bayes factor in the Bayesian literature for hypothesis testing and model comparison in general two-level structural equation models. It is shown that the proposed methodology is flexible, and can be applied to situations with a wide variety of nonnested models. Moreover, some problems encountered in using existing methods for goodness-of-fit assessment of the proposed model can be alleviated. An illustrative example with some real data from an AIDS care study is presented.

Introduction

Traditionally, many statistical methods assume identically and independently distributed observations. However, in practice, it is often to encounter hierarchically structured data which are collected from individuals that are nested within larger clusters (see, e.g., Aitkin & Longford, 1986; Goldstein, 1986; Goldstein & McDonald, 1988; among others). Examples might well be drawing of random samples of students from within random samples of classes or schools; or individuals from within random samples of families, etcetera. For this kind of multilevel data, the assumption of independence is not realistic, because individuals within a cluster (or a group) are expected to share certain common influential factors and hence produce correlated observations. Analyzing multilevel data as a single random sample can lead to erroneous results. Hence, a lot of attention has been devoted to develop models and methods that take account of the hierarchical
structure (see, e.g., the above listed citations, de Leeuw & Kreft, 1986; Raudenbush, 1995; Raudenbush & Bryk, 1988; Thum, 1997; among others).

Structural equation modeling is a widely used multivariate method in assessing causations and correlations among observed and latent variables. The need for more general models that take into account the correlated structure of multilevel data is well recognized; see for example, McDonald and Goldstein (1989), Lee (1990), and McDonald (1993) for some theoretical developments; Muthén (1991, 1994), Lee and Poon (1992), Longford and Muthén (1992), Raudenbush (1995), and Lee and Tsang (1999) for developments of various computational algorithms.

One important statistical inference beyond estimation is related to testing of various hypotheses about the model; for example, the goodness-of-fit of the posited model to the sample data. In the field of structural equation modeling, a common 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. As pointed out in the literature (see, e.g., Berger & Delampady, 1987; Berger & Sellke, 1987; Kass & Raftery, 1995) there are difficulties with such an approach. Some of those related to structural equation modeling are as follows.

1. Tests on the basis of $P$-values tend to reject the null hypothesis frequently with large sample sizes; see Raftery (1986) for a concrete example. Various descriptive fit indexes, such as the well-known normed or non-normed fit indexes (Bentler & Bonett, 1980) and the comparative fit index (Bentler, 1992) have been proposed as complementary measures for the goodness-of-fit of the model. Very often, the values of the fit indexes are over 0.95, but the $P$-values of the $\chi^2$-test are less that 0.01. Under these situations, conclusions drawn from these two testing methods seem contradictory.

2. The $P$-value of a significance test in hypothesis testing is a measure of evidence against the null model, not a means of supporting/proving the model.

3. The significance tests as well as the descriptive fit indexes mentioned above cannot be applied to test nonnested hypotheses or to compare nonnested models.

In addition to the above problems relating to standard structural equation model (SEM) with independently and identically distributed (i.i.d.) observations, we encounter one further difficulty in hypothesis testing with the more complex multilevel models. The goodness-of-fit test as well as the fit indexes are all based on some test statistics which are claimed to have a central or a non-central $\chi^2$ distribution with certain degrees of freedom (see, Bentler, 1992). The fundamental background of these test statistics is based
on the theory of likelihood ratio test in conjunction with maximum likelihood (ML) estimation with i.i.d. observations. For multilevel models with unbalanced data, in order to get independent observations, individual observations (of dimensions $p$) within a group, say group $g$ with sample size $N_g$, have to be pooled together to form a single long vector of dimension $pN_g$. The resulting likelihood function is the product of $G$ independent but non-identical random vectors with different dimensions $pN_g$. Hence, the basic i.i.d. assumption in the ML estimation is not satisfied. How serious this violation of assumption to hypothesis testing with the likelihood ratio chi-square test is unknown.

Therefore, it is important to find a testing procedure to overcome the above practical and theoretical problems. In this article, an approach using the Bayes factor (e.g. Berger, 1985) is proposed. The development of Bayes factor is quite general/universal and it has been applied to a wide variety of models, see Kass and Raftery (1995). The main aim of this article is to illustrate how it can be used for hypothesis testing and model comparison in the context of two-level SEMs. As a by-product, this approach can be applied to test or compare functional constraints among the parameters in the model.

**Model**

Consider a data set of $p$-variate observations $x_{gi}$, $i = 1, \ldots, N_g$, within groups $g = 1, \ldots, G$. The group-level sample sizes $N_g$ may differ from group to group. Assuming that, conditional on the group mean $v_g$, observations within each group are i.i.d. as $N(v_g, \Sigma_{gw})$; that is

$$
(x_{gi} | v_g) \overset{D}{=} N(v_g, \Sigma_{gw}),
$$

where $\Sigma_{gw}$ is the within-group covariance matrix. Moreover, we assume that the mean vectors $\{v_g, g = 1, \ldots, G\}$ are i.i.d. and

$$
v_g \overset{D}{=} N(\mu, \Sigma_B),
$$

where $\Sigma_B$ is the between-group covariance matrix. The covariance matrices $\Sigma_{gw}$ and $\Sigma_B$ can have any general form such as the LISREL model (Jöreskog & Sörbom, 1996) or the Bentler and Weeks (1980) model in EQS (Bentler, 1992). Hence, the model considered is a general two-level model with unbalanced designs. To achieve valid statistical inference on this general model, certain assumptions on the sample sizes have to be satisfied. For example, $G$ has to be sufficiently large in analyzing the between group
structures $\Sigma_g$ and $N_g$ has to be sufficiently large in estimating the specific parameters involved in group $g$. In many applications of the two-level models, it is assumed that $\Sigma_g = \Sigma_w$. For simplicity, we assume that $\Sigma_g = \Sigma_w$ and $\mu = 0$; but keep in mind that the proposed procedure can be extended to the general case with minor modifications.

Let $\theta$ be the $q \times 1$ vector of parameters, so that $\Sigma_g = \Sigma_B(\theta)$ and $\Sigma_w = \Sigma_w(\theta)$ are corresponding matrix-valued functions of $\theta$. For example, a two-level factor analysis model is given by

\begin{equation}
\mathbf{x}_{gi} = \lambda_B \xi_g + \epsilon_g + \lambda_w \eta_{gi} + \delta_{gi}, \quad g = 1, \ldots, G, \quad i = 1, \ldots, N_g,
\end{equation}

where $\xi_g$ is the common factor at the group level with distribution $N(0, \Phi_B)$, $\eta_{gi}$ is the common factors at the individual level with distribution $N(0, \Phi_w)$, $\epsilon_g$ and $\delta_{gi}$ are the error measurements at the group level and individual level with distributions $N(0, \Psi_B)$ and $N(0, \Psi_w)$ respectively, where $\Psi_B$ and $\Psi_w$ are diagonal covariance matrices. In this model,

\begin{equation}
\Sigma_B = \lambda_B \Phi_B \lambda_B^T + \Psi_B, \quad \text{and} \quad \Sigma_w = \lambda_w \Phi_w \lambda_w^T + \Psi_w.
\end{equation}

The parameter vector $\theta$ contains unknown parameters in $\lambda_B$, $\lambda_w$, $\Phi_B$, $\Phi_w$, $\Psi_B$ and $\Psi_w$.

Suppose we are interesting in a class of specific models under the framework of the general model defined by Equations 1 and 2, say $M_1, \ldots, M_k$. We call the model that is in common with every $M_k$ the basic model. Thus, each particular model can be represented by a basic model, which may be the saturated model where $\Sigma_B$ and $\Sigma_w$ are general positive definite matrices without any structures, or a special model of the general model given in Equations 1 and 2 together with an appropriate set of constraints in $\theta$. More specifically, a model $M_k$ can be defined by a basic model, and the following constraints among its parameters in $\theta$:

\begin{equation}
C_k(\theta) = [C_1^{(k)}(\theta), \ldots, C_r^{(k)}(\theta)] = 0,
\end{equation}

where $C_1^{(k)}$, ..., $C_r^{(k)}$ are any linear or nonlinear functions of $\theta$. Let $\theta_k$ be the vector of parameters in $M_k$ that satisfies $C_k(\theta) = 0$. For example, a two-level factor analysis model with equal cross-level loading matrices can be defined by a basic model as specified in Equation 3 together with the constraint: $\lambda_B - \lambda_w = 0$, or $\lambda_B = \lambda_w = \lambda$. In this special case,

\begin{equation}
\Sigma_B = \lambda \Phi_B \lambda^T + \Psi_B, \quad \Sigma_w = \lambda \Phi_w \lambda^T + \Psi_w,
\end{equation}
and the parameter vector contains the unknown parameters in $\Lambda$, $\Phi_B$, $\Phi_W$, $\Psi_B$ and $\Psi_W$. Another special case can be defined with the constraint: $\Phi_B = \Phi_W = \Phi$. In this case,

\[
\Sigma_B = \Lambda_B \Phi_B \Lambda_B^T + \Psi_B, \quad \Sigma_W = \Lambda_W \Phi_W \Lambda_W^T + \Psi_W.
\]  

Note that models given in Equations 6 and 7 are not nested within each other.

**Bayes Factor for Hypothesis Testing or Model Comparison**

We will use the Bayes factor to compare two competing models $M_1$ and $M_2$ for a given data set. Bayes factor is a well-known method (see, e.g. Berger, 1985), and it is still the focus of attention in recent Bayesian literature for hypothesis testing and model selection, see for example Aitkin (1991), Raftery (1995), Newton and Raftery (1994), Gelfand and Dey (1994), Berger and Pericchi (1996), Draper (1995), and O’Hagan (1995), among others. Application of this approach to model selection in single level SEMs has been proposed by Raftery (1993). Here, we first describe briefly its general development and then demonstrate its application to two-level SEMs.

Suppose the given data $D$ with sample size $n$ have arisen under one of the two competing models $M_1$ and $M_2$ according to a probability density $p(D|M_1)$ or $p(D|M_2)$. For model $M_k$, $k = 1, 2$, let $p(M_k)$ be the prior probability and $p(M_k|D)$ be the posterior probability. From the Bayes theorem, we obtain

\[
p(M_k|D) = \frac{p(D|M_k)p(M_k)}{p(D|M_1)p(M_1) + p(D|M_2)p(M_2)}, \quad k = 1, 2.
\]

Hence,

\[
\frac{p(M_1|D)}{p(M_2|D)} = \frac{p(D|M_1)p(M_1)}{p(D|M_2)p(M_2)}.
\]

The Bayes factor is defined as

\[
B_{12} = \frac{p(D|M_1)}{p(D|M_2)}.
\]
which is a summary of the evidence provided by the data in favor of \( M_1 \), as opposed to \( M_2 \). When the prior probabilities of the two hypotheses are equal, the Bayes factor is the posterior odds. Hence, the Bayes factor can be used as a quantity for model comparison. Unlike the significance test approach that is based on \( P \)-values, this comparison does not depend on the assumption that either model is ‘true’, and can be applied to nonnested models.

When the two competing models \( M_1 \) and \( M_2 \) only involve single distributions with no unknown parameters, the Bayes factor \( B_{12} \) is the likelihood ratio. In the general case where there are unknown parameters in either or both \( M_1 \) and \( M_2 \), \( B_{12} \) still has the form of a likelihood ratio. However, this test statistic is different from the usual large-sample likelihood ratio test in which the null hypothesis must be nested within the alternative hypothesis. Moreover, the densities \( p(D|M_k) \) in Equation 9, \( k = 1, 2 \), are obtained by integrating over the parameter space, that is

\[
p(D|M_k) = \int p(D|\theta_k, M_k)\pi(\theta_k|M_k)d\theta_k,
\]

where \( \theta_k \) is the \( d_k \times 1 \) parameter vector in \( M_k \), \( \pi(\theta_k|M_k) \) is its prior density, and \( p(D|\theta_k, M_k) \) is the probability density of \( D \) given \( \theta_k \), which may also be interpreted as the likelihood function of \( \theta \) evaluated at \( \theta_k \) under \( M_k \). The dimension of the above complicated integral is equal to the dimension of \( \theta_k \). Hence, very often, it is very difficult to obtain \( B_{12} \) analytically, and various analytic and numerical approximations have been proposed in the literature (Kass & Raftery, 1995). For example, see the Laplace method of approximation (Tierney & Kadane, 1986) and its alternative forms (Kass & Vaidyanathan, 1992; Tierney, Kass & Kadane, 1989), as well as Chib’s (1995) procedure with the Gibbs sampler algorithm. The above methodologies are rather involved and depend on the prior density \( \pi(\theta_k|M_k) \).

In this article, the following simple approximation that is not depending on the prior density is used to approximate \( 2\log B_{12} \):

\[
2\log B_{12} \approx 2S = 2\left[ \log p(D|\hat{\theta}_1, M_1) - \log p(D|\hat{\theta}_2, M_2) \right] - (d_1 - d_2)\log n,
\]

where \( S \) is often called the Schwarz criterion (Schwarz, 1978), \( \hat{\theta}_1 \) and \( \hat{\theta}_2 \) are the maximum likelihood estimates of \( \theta_1 \) and \( \theta_2 \) under \( M_1 \) and \( M_2 \), respectively. According to Kass and Raftery (1995), \( 2\log B_{12} \), with \( B_{12} \) defined in Equation 9, can be interpreted via the following criterion:
This commonly used criterion provides a descriptive scale about standards of evidence. Since it uses the asymmetric natural logarithm transformation of $B_{12}$, the cut-points are not symmetric. It gives more detailed guidelines relating to $M_1$. We can evaluate $2\log B_{12}$ if more detailed descriptive statements about evidence of $M_2$ are required.

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

$$
\frac{S - \log B_{12}}{\log B_{12}} \to 0,
$$

thus $S$ may be viewed as an approximation to $\log B_{12}$. This approximation is of order 0(1), hence it does not give the exact $\log B_{12}$ even for large samples. However, since the interpretation is on the natural logarithm scale, it provides a reasonable indication of evidence. As pointed out by Kass and Raftery (1995), it can be used for scientific reporting as long as the number of degrees of freedom ($d_1 - d_2$) involved in the comparison is small relative to the sample size $n$.

Suppose the competing models $M_1$ and $M_2$ are defined by an identical basic model (e.g. the saturated model) together with appropriate sets of constraints $C_1(\theta) = 0$ and $C_2(\theta) = 0$, respectively. Comparing $M_1$ and $M_2$ is then equivalent to comparing the evidence for $C_1(\theta) = 0$ and $C_2(\theta) = 0$. The Bayes factor $B_{12}$ is a summary of the evidence provided by the data in favor of the constraint $C_1(\theta) = 0$ as opposed to constraint $C_2(\theta) = 0$. In the computation of $2\log B_{12}$ via Equation 10, $\hat{\theta}_k$ is the corresponding constrained ML estimate and $d_k$ is replaced by $d_k - r_k$. In the special case where $M_2$ is the saturated model that involves no constraints, $r_2 = 0$ and $d_2$ is the number of unknown parameters in the saturated model. Substituting $r_2 = 0$ and $d_2$ in Equation 10, we can compute $2\log B_{12}$ which gives the summary of the evidence in favor of $C_1(\theta) = 0$ as opposed to the saturated model. As a result, arbitrary constraints (linear or nonlinear) on the parameter vectors can be compared or tested. Since statistical theory on the constrained ML estimation of two-level SEMs is not yet established, this kind of inference cannot be done rigorously via the existing classical likelihood ratio test.
Let $z^T_g = (x^T_{g1}, ..., x^T_{gN_g})$ be the $1 \times N_g$ vector that contains all observations in group $g$. The distribution of $z_g$ is $N(0, (J_g \otimes \Sigma_g) + (I_g \otimes \Sigma_w))$, where $J_g$ is an $N_g \times N_g$ square matrix of unit elements and $I_g$ is the identity matrix of order $N_g$. Since observations from different groups are independent, $z_h$ and $z_g$ are independent for $h \neq g$. It can be shown that the log-likelihood function (ignoring a constant which remains the same for different models) based on the data $D = \{z_1, ..., z_G\}$ is given by

$$\log p(D|\theta, M) = -\frac{1}{2} \sum_{g=1}^{G} N_g \left[ \log |\Sigma_g + \text{tr}(\Sigma_g^T S_g)\right] + \left(N_g - 1\right) \left[ \log |\Sigma_w + \text{tr}(\Sigma_w^T S_w)\right].$$

where $\Sigma_g = \Sigma_w + N_g \Sigma_g$, and

$$S_g = N_g^{-2} \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} x_{gi} x_{gj}^T, \quad S_w = \left[N_g \left(N_g - 1\right)\right]^{-1} \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} x_{gi} (x_{gi} - x_{gj})^T.$$
On the basis of the criterion given in Equation 11, if $2\log B_{23} < 0$, $M_3$ is the better model; while $M_2$ is the better model if $2\log B_{23} > 2$. No conclusion can be drawn if $0 \leq 2\log B_{23} < 2$. For model selection problem that involves a large number of possible models, the search strategy given in Raftery (1995) can be used. To save space, this strategy is not presented here.

For most applications of the two-level SEMs, constraints considered in the literature so far are very simple. They are either in form: $\theta_i^{(k)} = \text{constant}$, say 0 or 1; or $\theta_i^{(k)} - \theta_j^{(k)} = 0$. Longford and Muthén (1992) developed a scoring type algorithm for obtaining ML estimates of the parameters that subject to the above simple constraints in a two-level factor analysis model. Lee and Poon (1998) developed an efficient EM type algorithm that can be applied to general SEMs. Moreover, they showed that if $\Sigma_b$ and $\Sigma_w$ are the LISREL model or the Bentler and Weeks (1980) model, the well-known software LISREL (Jöreskog & Sörbom, 1996) or EQS (Bentler, 1992) can be used to obtain the ML estimate. Once we get the ML estimate, the Bayes factor can be calculated. Constrained ML estimates that are subjected to more general linear or nonlinear constraints can be obtained via the EM type algorithm with more computational effort, see Lee and Tsang (1999). However, the M-step of their proposed EM algorithm can be completed via LISREL (Jöreskog & Sörbom, 1996) or EQS (Bentler, 1992) with simple auxiliary programs.

**Illustrative Example**

A portion of the real data set in a study of the relationship between AIDS and the use of condom among Philippines sex workers (Morisky et al., 1998) is used to illustrate the proposed Bayes factor approach. The whole data set was collected from female commercial sex workers in 97 establishments (bars, right clubs, Karaoke TV, and massage parlors) in cities of Philippines. The questionnaire contains items on socio-demographic background, knowledge, attitudes, belief, behaviours, self efficacy for condom use, alcohol and drug use and social desirability. As pointed out by Morisky et al., influence of the establishment policies is substantial. Hence, a two-level model should be used to take into account the influential factor of the establishments. As an illustration of the proposed method, only six variables are used. The questions corresponding to the first three variables are: How great is the risk of getting AIDS or the AIDS virus from (a) kissing a person with the AIDS virus on the cheek; (b) deep kissing with someone who has the AIDS virus; and (c) sexual intercourse with someone who has the AIDS virus using a condom; while the questions corresponding to the last three variables are: (d) how much of a threat do you think AIDS is to the health
of people; (e) what are the chances that you yourself might get AIDS; and (f) how worried are you about getting AIDS. These are polytomous variables measured in a five-point scale. Since our main purpose is to illustrate the proposed methodology, these variables are treated as continuous variables with a multivariate normal distribution. The final results may be affected by the polytomous nature of the variable and hence real conclusion for the theory from analysis of this AIDS data set should be drawn with caution. Also, for simplicity, we deleted those observations with missing entries in the analysis. The within-establishment sample sizes are summarized in Table 1. For example, there are 6 establishments with sample sizes equal to 1, 11 establishments with sample size equals to 2, and so on. The total sample size is equal to 758. To unify the scales of the variables, the raw data were standardized. Motivated by Morisky et al. to take into account the second level effect on the influence of the establishment policies, the resulting data set was analyzed via a two-level confirmatory factor analysis model. On the basis of some exploratory analysis (see also Lee & Tsang, 1999), we begin our analysis with the following specifications:

$$\Lambda_B = \begin{bmatrix}
\Lambda_B(1,1) & \Lambda_B(2,1) & \Lambda_B(3,1) & 0^* & \Lambda_B(5,1) & 0^*
\end{bmatrix}^T,$$
$$\Phi_B = 1.0^*, \quad \Psi_B = \text{diag} [\Psi_B(1,1),...,\Psi_B(6,6)],$$

$$\Lambda_w = \begin{bmatrix}
\Lambda_w(1,1) & \Lambda_w(2,1) & \Lambda_w(3,1) & 0^* & 0^* & 0^* \\
0^* & 0^* & 0^* & \Lambda_w(4,2) & \Lambda_w(5,2) & \Lambda_w(6,2)
\end{bmatrix}^T,$$
$$\Phi_w = \begin{bmatrix}
1.0^* & \Phi_w(2,1) \\
\Phi_w(2,1) & 1.0^*
\end{bmatrix}, \quad \Psi_w = \text{diag} [\Psi_w(1,1),...,\Psi_w(6,6)].$$

Table 1
The Distribution of $N_g$ in the AIDS Data

<table>
<thead>
<tr>
<th>$N_g$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>Subtotal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>6</td>
<td>11</td>
<td>13</td>
<td>6</td>
<td>5</td>
<td>6</td>
<td>11</td>
<td>7</td>
<td>7</td>
<td>2</td>
<td>74</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$N_g$</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>19</th>
<th>28</th>
<th>59</th>
<th>Subtotal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>3</td>
<td>6</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>23</td>
</tr>
</tbody>
</table>

Total | 97 |
where elements with an asterisk are fixed at the preassigned values. The total number of unknown parameters is 23. The between-group covariance structure is a single factor model with the group factor representing the influence of the establishments on the sex workers attitude to the risk of getting AIDS and their worry of AIDS. The within-group covariance structure is a factor model with two non-overlapping factors, where the first three manifest variables are loaded on the first factor only and the last three manifest variables are loaded on the second factor only. The first one can be roughly interpreted as one about the risk of getting AIDS, while the other as one about the worry of AIDS.

To illustrate the use of Bayes factor in testing the goodness-of-fit of a model, we consider a two-level model $M_1$ that is defined by the model with specifications Equation 15 together with the following constraints $C_1(\theta) = 0$ given in Lee and Tsang (1999):

$$
\begin{align*}
(i) & \quad \Lambda_B^2(j,1) + \Psi_B(j,j) + \Lambda_w^2(j,1) + \Psi_w(j,j) = 1.0, \quad \text{for } j = 1,2,3, \\
(ii) & \quad \Lambda_B^2(j,1) + \Psi_B(j,j) + \Lambda_w^2(j,2) + \Psi_w(j,j) = 1.0, \quad \text{for } j = 4,5,6, \\
(iii) & \quad \Lambda_w(4,2) = \Lambda_w(5,2) = \Lambda_w(6,2) \\
(iv) & \quad \Lambda_w(3,1) = \frac{1}{3} \Lambda_w(1,1) = \frac{1}{4} \Lambda_w(2,1).
\end{align*}
$$

The first six nonlinear constraints were used to fix the diagonal elements of the estimated covariance matrix of the manifest variables to 1.0; while the last four linear constraints specify some relationships among the parameters in the loading matrix of the within-group structure. For completeness, the constrained ML estimates are presented in Table 2. At these ML estimates, $\log p(D|\theta_1, M_1)$ is equal to -1947.1, while the corresponding value for the saturated model is equal to -1928.9. Also, $n = 758, d_1 = 23, r_1 = 10, d_s = 42$, hence from Equation 13, $2\log B_{1s} = 155.87$. On the basis of the criterion given in Equation 11, it can be concluded that the data provide decisively evidence in favor of $M_1$. Hence, a two-level confirmatory factor analysis model has been established for the given data. Note that the application of the approach with significance test via $P$-values to assess the goodness-of-fit of this posited model with nonlinear constraints is questionable. The nonzero estimates of the between-group level factor loadings indicate substantial heterogeneity across establishments. From the constrained estimates of the within-group factor loadings, the influences of the non-overlapping factors to the observed variables satisfied the prescribed linear relationships. For
example, the factor about the worry of AIDS has equal effects on the observed variables.

To reaffirm the heterogeneity across establishments using the Bayes factor, we consider the following model $M_0$ with basic specifications in Equation 15 and constraints:

\begin{equation}
\begin{aligned}
C_0(\Theta) = 0: & \Lambda_B = 0, \Psi_B(j, j) = 0, j = 1, \ldots, 6.
\end{aligned}
\end{equation}

This model is equivalent to a single-level confirmatory factor analysis model that does not take into account the influential factor of the establishments. It has 13 unknown parameters in $\Lambda_w, \Phi_w$ and $\Psi_w$. We observed that $\log p(D | \hat{\Theta}_0, M_0)$ is equal to -2059.0. To compare with the saturated model, we found that $2\log B_{01} = -67.73$. Hence, it can be concluded from Equation 11 that this single-level model should be rejected.

To illustrate the proposed approach further, the following models are compared using the Bayes factor.

$M_2$. A model with specifications given in Equation 15 and the following constraints: (i), (ii) and (iii) as given in Equation 16, and $\Lambda_w(1, 1) = \Lambda_w(2, 1) = \Lambda_w(3, 1)$. 

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Parameter</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Lambda_B (1, 1)$</td>
<td>0.265</td>
<td>$\Lambda_w (1, 1)$</td>
<td>0.590</td>
</tr>
<tr>
<td>$\Lambda_B (2, 1)$</td>
<td>0.316</td>
<td>$\Lambda_w (2, 1)$</td>
<td>0.784</td>
</tr>
<tr>
<td>$\Lambda_B (3, 1)$</td>
<td>0.196</td>
<td>$\Lambda_w (3, 1)$</td>
<td>0.196</td>
</tr>
<tr>
<td>$\Lambda_B (4, 1)$</td>
<td>0.(fixed)</td>
<td>$\Lambda_w (4, 2)$</td>
<td>0.370</td>
</tr>
<tr>
<td>$\Lambda_B (5, 1)$</td>
<td>-.329</td>
<td>$\Lambda_w (5, 2)$</td>
<td>0.370</td>
</tr>
<tr>
<td>$\Lambda_B (6, 1)$</td>
<td>0.(fixed)</td>
<td>$\Lambda_w (6, 2)$</td>
<td>0.370</td>
</tr>
<tr>
<td>$\Psi_B (1, 1)$</td>
<td>0.018</td>
<td>$\Psi_w (1, 1)$</td>
<td>0.560</td>
</tr>
<tr>
<td>$\Psi_B (2, 2)$</td>
<td>0.025</td>
<td>$\Psi_w (2, 2)$</td>
<td>0.258</td>
</tr>
<tr>
<td>$\Psi_B (3, 3)$</td>
<td>0.044</td>
<td>$\Psi_w (3, 3)$</td>
<td>0.876</td>
</tr>
<tr>
<td>$\Psi_B (4, 4)$</td>
<td>0.090</td>
<td>$\Psi_w (4, 4)$</td>
<td>0.770</td>
</tr>
<tr>
<td>$\Psi_B (5, 5)$</td>
<td>0.135</td>
<td>$\Psi_w (5, 5)$</td>
<td>0.619</td>
</tr>
<tr>
<td>$\Psi_B (6, 6)$</td>
<td>0.165</td>
<td>$\Psi_w (6, 6)$</td>
<td>0.697</td>
</tr>
<tr>
<td>$\Phi_w (2, 1)$</td>
<td>0.250</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
\(M_3\). A model with two factors in \(\Sigma_B\) such that

\[
\Lambda_B = \begin{bmatrix}
\Lambda_B^{(1,1)} & \Lambda_B^{(2,1)} & \Lambda_B^{(3,1)} & 0^* & 0^* & 0^*
\end{bmatrix}^T,
\]

\[
\Phi_B = \begin{bmatrix}
10^* & \Phi_B^{(2,1)}
\end{bmatrix}^T,
\]

\[
\Psi_B = diag[\Psi_B^{(1,1)}, \ldots, \Psi_B^{(6,6)}].
\]

and same specifications of \(\Lambda_w, \Phi_w,\) and \(\Psi_w\) as given in Equation 15, together with same constraints as given in Equation 16.

\(M_4\). Same specifications in the parameter matrices as given in \(M_3\) for \(\Sigma_B\) and \(\Sigma_w\), together with constraints (i) and (ii) given in Equation 16, and for \(j = 1, \ldots, 6,\)

\[
\Lambda_B(j, 1) = \Lambda_w(j, 1), \text{ and } \Lambda_B(j, 2) = \Lambda_w(j, 2).
\]

Note that \(M_1, M_2, M_3\) and \(M_4\) are nonnested models. The motivation to consider \(M_3\) with two factors in \(\Sigma_B\) is to study whether the influence of the establishments can be better accounted by two factors instead of a single factor as before. The motivation for \(M_4\) is to assess cross-level relationship of the factor loadings. To save space, constrained ML estimates of parameters in these models are not presented. The logp(D | \(\Theta_k, M_k\)) values for \(M_2, M_3\) and \(M_4\) are equal to -1993.7, -1956.3 and -1975.5; respectively. The computed \(2\log B_{hs}\) and \(2\log B_{hk}\), with \(h, k = 1, 2, 3, 4\) are presented in Table 3. For example, \(2\log B_{4s} = 91.76, 2\log B_{21} = -93.20,\) and \(2\log B_{43} = 25.14.\) From \(2\log B_{hs}\), it seems that all the models considered are more preferable than the saturated model. Based on more comparisons among themselves, we can conclude that \(M_1\) is the best among these four models.

Table 3
2\(\log B_{hs}\) and 2\(\log B_{hk}\) for Models Comparison

<table>
<thead>
<tr>
<th>Model (_h)</th>
<th>(M_1)</th>
<th>(M_2)</th>
<th>(M_3)</th>
<th>(M_4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(s)</td>
<td>155.87</td>
<td>62.69</td>
<td>117.58</td>
<td>91.76</td>
</tr>
<tr>
<td>(M_1)</td>
<td>-93.20</td>
<td>-38.29</td>
<td>-63.43</td>
<td>-63.43</td>
</tr>
<tr>
<td>(M_2)</td>
<td>54.91</td>
<td></td>
<td>29.77</td>
<td></td>
</tr>
<tr>
<td>(M_3)</td>
<td></td>
<td></td>
<td>25.14</td>
<td></td>
</tr>
</tbody>
</table>
To conclude, we have shown that a single-level confirmatory factor analysis model is inadequate to fit the AIDS data set. Several nonnested plausible two-level models have been found. Using the Bayes factor approach, we are able to identify the best model among these plausible models. The above conclusion on model comparison cannot be achieved by existing theory in two-level structural equation modeling.

From the above illustration with various models and their associated constraints, it can be seen that the Bayes factor is rather flexible and useful in model testing and comparison for two-level SEMs. The stepwise strategies proposed by Raftery (1993) can be used for situations with a large number of competitive models.

**Discussion**

Bayes factor (see Berger, 1985; Kass & Raftery, 1995) is the most important statistic in Bayesian model selection. It has been applied extensively to a lot of statistical problems, such as nonnested regression models (Carlin & Chib, 1995), density estimation (Escobar & West, 1995; Roeder & Wasserman, 1997), multiple change-point problems (Green, 1995), mixture models (Chib, 1995; Richardson & Green, 1997), multi-sample factor analysis model and multivariate linear model with polychotomous variables (Song & Lee, 2001, in press), among many others. In this article, a Bayes factor approach is proposed for hypothesis testing and model comparison in general two-level SEMs with unbalanced designs and parameters subject to nonlinear constraints. For two-level SEMs, some software programs give some chi-square goodness-of-fit statistics on the basis of the classical likelihood ratio test that is associated with the ML estimation. However, the basic i.i.d. assumption required by ML estimation is violated. Hence, we feel very uneasy to use the classical asymptotic result to claim directly that the likelihood ratio statistic has a chi-square distribution. The Bayes factor approach does not have the above theoretical problem. Moreover, this approach can alleviate other difficulties and deficiencies in using fitting indexes or significance tests. For example, it can be applied to nonnested models and does not tend to reject the null hypothesis frequently with large sample sizes. Although the approach is described in the context of two-level models with $\mu = 0$ and invariant within-group covariance structures, it can be applied to more general models. For example, if $\mu \neq 0$ in $M_1$, then the parameter vector $\theta_1$ will involve $\mu$ and the Bayes factor can be similarly computed via Equation 10 with a modified $d_1$.

In this article, the natural logarithm of the Bayes factor is approximated via the Schwarz criterion. Minus twice the Schwarz criterion is commonly
called the Bayesian information criterion (BIC); sometimes an arbitrary constant is added. As pointed out by Kass and Raftery (1995), the Schwarz criterion is easy to use, does not require evaluation of prior distribution and provides a reasonable indication of the evidence. The Akaike (1973) information criterion in comparing models \(M_1\) and \(M_2\) is given by

\[
AIC = -2[\log p(D|\hat{\Theta}_1, M_1) - \log p(D|\hat{\Theta}_2, M_2)] + (d_1 - d_2),
\]

which does not involve the sample size \(n\). Comparing Equations 10 and 18, we see that our approach tends to favor simpler models more than those selected by AIC.

The computation of the Bayes factor or its various approximations has received a great deal of attention in the literature of statistics and it is still an area of active research, see the survey papers by Kass and Raftery (1995), and DiCiccio, Kass, Raftery, and Wasserman (1997) and references therein. For two-level SEMs, the popularity of the Bayes factor approach depends on how easy the general users can obtain the ML estimates. As pointed out by Lee and Tsang (1999), only minor modification of the existing software such as LISREL (Jöreskog & Sörbom, 1996) and EQS (Bentler, 1992) is required in obtaining ML estimates that subject to simple linear constraints. Hopefully, we expect to see such options to be included in these common software in the near future. In the absence of the ML estimates, a further approximation of the Bayes factor can be computed in terms of the less optimal but more familiar estimates, such as those given in Muthén (1994). The evaluation of the performance of such an approximation via simulation studies would be interesting. The performance of the exact Bayes factor should also be evaluated, as well as its approximation by \(S\).

References


Accepted October, 2000.