A Bayesian model selection method with applications

Xin-Yuan Song, Sik-Yum Lee*

Department of Statistics, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong

Received 1 May 2001; received in revised form 1 January 2002

Abstract

In this paper, we consider Bayesian model selection using the well-known Bayes factor. A method on the basis of path sampling for computing the ratio of two normalizing constants involved in the Bayes factor is proposed. The key idea is to construct a continuous path to link up the competing models, then the Bayes factor can be estimated efficiently by means of grids in [0,1] and observations simulated from the posterior distribution of the parameters. This method is applied to non-nested regression models, mixture models with an unknown number of components, and a general latent variable model with mixed continuous and polytomous variables. Analyses of some real data sets are presented to illustrate the efficiency and flexibility of the method. © 2002 Elsevier Science B.V. All rights reserved.

Keywords: Bayes factor; Path sampling; Posterior simulation; Gibbs sampler; Latent variable models; Mixture models

1. Introduction

Model selection is a major issue in many statistical problems such as non-nested regression models, density estimation, multiple change-points problem, mixture models with unknown number of components, see Carlin and Chib (1995), Chib (1995), Gelfand and Dey (1994), George and McCulloch (1993), and Green (1995); among others. Bayes factors (see Berger, 1985) are perhaps the most important tool for Bayesian model selection. In general, since the required marginal densities involve intractable high-dimensional integrals, computation of Bayes factor is challenging and has received much attention in the literature. Proposed methods include the Schwarz...
criterion (Schwarz, 1978), estimations of the marginal density via posterior simulations (see, e.g., Chib, 1995; DiCiccio et al., 1997), and jump Markov chain methods (Green, 1995), among others. One objective of this article is to apply the path sampling (Gelman and Meng, 1998) for computing Bayes factors. The key idea of path sampling is to construct a continuous path to link two densities, then the ratio of the normalizing constants and hence the Bayes factor can be approximated by grids in [0,1] and sample draws from the appropriate posterior distribution via some Markov chain Monte Carlo (MCMC) methods. It will be demonstrated that this approach is flexible, computationally efficient and simple to implement, please refer to more discussions at the end of Section 2.2.1.

The method will be applied to test the number of components in a mixture model. Mixture models arise in many contexts, including behavioral science, medicine, economics and environmental science. They have been used in modeling heterogeneity, handling outliers (Pettit and Smith, 1985), and density estimation (Roeder and Wasserman, 1997), among others. Since there has been strong interest in finite mixture distributions (see, e.g., McLachlan and Basford, 1988; Titterington et al., 1985, among others), this is an interesting problem in its own right. For estimation with a fixed number of components, say $K$, a variety of methods have been proposed. Examples are the method of moments (Lindsay and Basak, 1993), Bayesian treatment with MCMC techniques (Diebolt and Robert, 1994; Robert, 1996); and the maximum likelihood method (Hathaway, 1985). For mixture models with $K$ treated as random, Richardson and Green (1997) developed a full Bayesian analysis on the basis of the jump MCMC method introduced by Green (1995). For the challenging problem of testing the number of components, the classical likelihood-based inference encountered serious difficulties (see Aitkin et al., 1981). As pointed out by Richardson and Green (1997), the Bayes paradigm is particularly suited to the problem with an unknown $K$.

The method for model selection will also be applied to a general latent variable model with mixed continuous and polytomous variables. Latent variable models (e.g., Bentler, 1992; Jöreskog and Sörbom, 1996) with continuous variables have been applied extensively to problems in behavioral, medical and social sciences. For single- and two-level models with continuous data, the issue of model selection has been addressed, respectively, by Raftery (1993), and Lee and Song (2001) on the basis of the Bayesian information criterion (BIC, Schwarz, 1978) which is a simple approximation of the Bayes factor. Recently, a lot of attention has been devoted to analyze models with mixed continuous and polytomous variables. For instance, Muthén (1984), and Lee et al. (1995) developed multi-stage methods on the basis of polyserial and polychoric correlations; Roboussin and Liang (1998) proposed a quadratic estimation equation approach; Sammel et al. (1997) linked the conditional distribution of the manifest variables given the latent variables with an exponential family model and then analyzed the data with a generalized linear model; Shi and Lee (2000) investigated the ML estimation and implemented a Monte-Carlo EM algorithm (Dempster et al., 1977; Wei and Tanner, 1990) to get the solution, and Lee and Zhu (2000) developed a Bayesian approach for nonlinear models. The work cited above on models with polytomous variables are mainly emphasized on estimation. Limited results have been achieved for model selection, except for Song et al. (2001) which was based on BIC, and Song
and Lee (2001) which considered the factor analysis model and computed the Bayes factor via Chib’s (1995) method.

The paper is organized as follows. In Section 2, we consider a method to compute Bayes factor via path sampling. Several features that are different from the existing methods are mentioned and an illustrative example in non-nested regression models is given. The topic on selecting the number of components in mixture models is discussed in Section 3. Illustrative examples with some published univariate and multivariate data are presented. The application to a latent variable model with mixed continuous and polytomous variables is explored and illustrated with a real example in Section 4. A discussion is given in Section 5.

2. Model selection via Bayes factor

2.1. Computing Bayes factor using path sampling

Suppose the observed data \( y \) have arisen under one of the \( K + 1 \) competing models \( \{M_0, \ldots, M_K\} \) which may be nested or non-nested. Let \( p(y|M_k) \) be the probability density of \( y \) given \( M_k \). The choice between any two different models, say \( M_0 \) and \( M_1 \), is commonly based on the Bayes factor that is defined by

\[
B_{10} = \frac{p(y|M_1)}{p(y|M_0)}.
\]

See Kass and Raftery (1995) for a comprehensive review of Bayes factors, their application to Bayesian hypothesis testing and the following criterion for interpretation:

<table>
<thead>
<tr>
<th>( \log B_{10} )</th>
<th>(&lt; 0)</th>
<th>(0–1)</th>
<th>(1–3)</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>

In general, the marginal density \( p(y|M_k) \) is obtained by integrating over the corresponding parameter space, that is

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

where \( \theta_k \) is the parameter vector in \( M_k \), \( p(\theta_k|M_k) \) is the prior density and \( p(y|\theta_k, M_k) \) is the likelihood. Since this marginal density usually cannot be evaluated in an analytic form, various numerical methods have been proposed to estimate it. In fact, besides the rough approximation of the Bayes factor via the Schwarz criterion (or BIC), most methods are concentrated on estimating \( p(y|M_k) \) via posterior simulation. These include the Bartlett adjustment of the Laplace approximation (Tierney and Kadane, 1986; DiCiccio and Stern, 1993), importance sampling and bridge sampling (see, DiCiccio et al., 1997), reciprocal importance sampling (Gelfand and Dey, 1994), Chib’s (1995) method using Gibbs outputs; among others. See DiCiccio et al. (1997) for a comprehensive comparative study. Recently, developed MCMC methods for simulating observations from posterior distributions greatly enhance their applicability. Generally, all
these methods directly involve the prior density of $\theta$. Most of them also require some estimates of the location and scale of the posterior, which may require extra effort to compute.

In this article, the powerful path sampling (Gelman and Meng, 1998) is utilized to evaluate the logarithm of Bayes factor. Instead of estimating the marginal density, this approach estimates the Bayes factor directly as the ratios of the appropriate normalizing constants. The idea is from the equality $p(\theta | y) = p(y, \theta) / p(y)$, where the marginal density $p(y)$ can be regarded as a normalizing constant of $p(y | \theta)$. Following Gelman and Meng (1998), we consider the following class of densities on the same space with a continuous path parameter $t \in [0, 1]$: 

$$p(\theta | y, t) = \frac{1}{z(t)} p(y, \theta | t),$$

where $z(t) = p(y | t)$. To compute the Bayes factor, we require to construct a continuous path to link $M_1$ and $M_0$ such that $z(1) = p(y | 1) = p(y | M_1)$ and $z(0) = p(y | 0) = p(y | M_0)$. Taking logarithm, differentiating (1) with respect to $t$, and assuming the legitimacy of interchange of integration with differentiation, we get

$$\frac{d \log z(t)}{dt} = \int \frac{1}{z(t)} \frac{d}{dt} p(y, \theta | t) d\theta = E_\theta \left[ \frac{d}{dt} \log p(y, \theta | t) \right],$$

where the expectation $E_\theta$ is taken with respect to the distribution $p(\theta | y, t)$. Suppose the prior distribution of $\theta$ is independent of $t$, then $d \log z(t)/dt = d \log p(y, \theta | t)/dt$. Let

$$U(y, \theta, t) = \frac{d}{dt} \log p(y | \theta, t),$$

which is not dependent on the prior density of $\theta$, then

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

Let $p(t)$ be a prior density of $t$, it follows that

$$\log B_{10} = E_{\theta, t} \left[ \frac{U(y, \theta, t)}{p(t)} \right],$$

where the expectation is taken with respect to the joint density of $p(\theta | y, t)p(t)$. Note that since the prior of $\theta$ is independent of $t$, this approach does not require to compute $p(\theta)$ directly. To approximate $\log B_{10}$, we follow the idea of Gelman and Meng (1998) to numerically evaluate the integral over $t$ via the trapezoidal rule as in Ogata (1989). Specifically, we first order the values of different $t_i$ from fixed grids $\{t_{(s)}\}_{s=0}^S$ such that $t_{(0)} = 0 < t_{(1)} < t_{(2)} < \cdots < t_{(S)} < t_{(S+1)} = 1$, and estimate $\log B_{10}$ by

$$\log \hat{B}_{10} = \frac{1}{2} \sum_{s=0}^S (t_{(s+1)} - t_{(s)})(\hat{U}_{(s+1)} + \hat{U}_{(s)}),$$

(5)
where
\[ \bar{U}(\theta) = J^{-1} \sum_{j=1}^{J} U(\theta^{(j)}, t(s)) \] (6)
in which \( \{\theta^{(j)}, j = 1, \ldots, J\} \) are observations simulated from \( p(\theta | y, t(s)) \).

The introduced method on the basis of path sampling (Gelman and Meng, 1998) has several nice features. Its implementation is simple, the only main task is in simulating \( \{\theta^{(j)}, j = 1, \ldots, J\} \) from \( p(\theta | y, t(s)) \) for every \( s \), which can be done efficiently with some appropriate MCMC methods. In general, as pointed out by Gelman and Meng (1998), we can always construct a continuous path to link two competing models with the same support. Hence the method can be applied to a wide variety of problems. Unlike some methods in estimating the marginal likelihood via posterior simulation, it does not require to estimate the location and/or scale parameters in the posterior. Instead, Bayesian estimates of these parameters under \( M_0 \) and \( M_1 \) can be obtained easily via the sample mean and sample variance (or covariance matrix) of the simulated observations at \( t = 0 \) and 1. Distinct from most existing approaches, the prior density is not directly involved in the evaluation. Finally, the logarithm scale of Bayes factor is computed, which is generally more stable than the ratio scale.

2.2. An example from non-nested regression models

As an illustration of the proposed method, we analyze a data set (Williams, 1959) which was considered by Efron (1984) and Carlin and Chib (1995) in fitting two plausible straight line models. The data set (see Carlin and Chib, 1995, Table 1) contains \( n = 42 \) specimens of radiate pines, the strength \( y_i \) of a piece of wood was measured along with its density \( x_i \) and its adjusted density \( z_i \). It is desired to compare the following two non-nested regression models:

\[ M_0: y_i = \alpha + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, n, \]
\[ M_1: y_i = \alpha + \beta_2 z_i + \epsilon_i, \quad i = 1, \ldots, n, \]

where \( \epsilon_i \) are i.i.d. \( N[0, \sigma^2] \). To apply the proposed method, the competing models are linked up by \( t \in [0, 1] \) as follows:

\[ M_t: y_i = \alpha + (1 - t)\beta_1 x_i + t\beta_2 z_i + \epsilon_i. \]

The log-likelihood function for \( \theta = (\alpha, \beta_1, \beta_2, \sigma) \) is given by
\[
\log p(y|\theta, t) = -\frac{n}{2} \left\{ \log(2\pi) + n \log \sigma^2 \right. \\
\left. + \sum_{i=1}^{n} \left[ y_i - \alpha - (1 - t)\beta_1 x_i - t\beta_2 z_i \right]^2 / \sigma^2 \right\}. 
\] (7)

It follows from (3) that
\[
U(y, \theta, t) = \sum_{i=1}^{n} \left[ y_i - \alpha - (1 - t)\beta_1 x_i - t\beta_2 z_i \right](-\beta_1 x_i + \beta_2 z_i) / \sigma^2. 
\] (8)
To compute the estimate of $\log B_{10}$ via (5) and (6), $t(\alpha)$ are obtained from fixed grids in $[0,1]$ and random observations of $\theta$ are simulated via the Gibbs sampler (Geman and Geman, 1984) from the posterior distribution $p(\theta|y,t)$. According to Carlin and Chib (1995), the following independent prior distributions on parameters of $\theta$ were taken: $\alpha \sim N[3000,10^6]$, $\beta_1 \sim N[185,10^4]$, $\beta_2 \sim N[185,10^4]$, and $\sigma^2$ follows an inverted Gamma ($IG$) distribution with the mean and standard deviation equal to $300^2$. These priors are roughly based on the appropriate least-squares estimates but are extremely vague. We monitored the convergence of parallel runs of the Gibbs sampler via the method of Gelman and Rubin (1992). Since $p(\theta|y,t)$ is very simple, it required very little computing time to simulate the observations and the Gibbs sampler converged very quickly in about 50 iterations. After convergence, a total of $J = 550$ observations were collected to estimate $\hat{U}(y,0,t)$. The number of grids was taken to be 20. Here the total number of Gibbs iterations required to achieve the result is 12,000. The computed $\log B_{10}$ is equal to 8.647. Hence, there is a strong evidence in favor of $M_1$. This conclusion is the same as that given in Carlin and Chib (1995).

Bayesian estimate of $\theta = (\alpha, \beta, \sigma)$, which is obtained from an approximation of the posterior mean using observations simulated under the selected model $M_1$, and its standard error estimate are equal to $\{2992.18, 182.43, 277.44\}$ and $\{42.64, 9.14, 29.25\}$, respectively.

The data set has been reanalyzed with the following independent prior distributions that are significantly different from above: $\alpha \sim N[0,10^6]$, $\beta_1 \sim N[0,10^4]$, $\beta_2 \sim N[0,10^4]$, and $\sigma^2$ follows an inverted Gamma distribution with mean and standard deviation equal to 300. $\log B_{10}$ is equal to 8.728, hence $M_1$ is selected again. Bayesian estimate of $\theta$ and standard error estimate are, respectively, equal to

<table>
<thead>
<tr>
<th>Prior distribution</th>
<th>Estimated logarithm of Bayes factors with different hyper-parameters values</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\pi \propto D(1,...,1)$</td>
<td>$(v_0, \rho_0) = (3, 20)$ $\log B_{21}$ $5.440$ $7.207$ $7.918$</td>
</tr>
<tr>
<td>$\mu_k \propto N[20,10^2]$</td>
<td>$\log B_{12}$ $7.194$ $7.527$ $6.277$</td>
</tr>
<tr>
<td>$\sigma_k^2 \propto IG[v_0, \rho_0]$</td>
<td>$\log B_{21}$ $5.440$ $5.725$ $5.594$</td>
</tr>
<tr>
<td>$\pi \propto D(1,...,1)$</td>
<td>$\sigma_0^2 = 10^2$ $\sigma_0^2 = 15^2$ $\sigma_0^2 = 20^2$</td>
</tr>
<tr>
<td>$\sigma_k^2 \propto IG[3,20]$</td>
<td>$\log B_{21}$ $5.440$ $5.216$ $6.671$</td>
</tr>
<tr>
<td>$\mu_k \propto N[20,\sigma_k^2]$</td>
<td>$\log B_{12}$ $7.194$ $6.122$ $5.555$</td>
</tr>
<tr>
<td>$\mu_k \propto N[\mu_0, 10^2]$</td>
<td>$\pi \propto D(1,1)$ $\pi \propto D(1,3)$ $\pi \propto D(3,1)$</td>
</tr>
<tr>
<td>$\sigma_k^2 \propto IG[3,20]$</td>
<td>$\log B_{21}$ $5.440$ $6.025$ $4.572$</td>
</tr>
<tr>
<td>$\log B_{32}$ $7.194$ $6.654$ $5.737$</td>
<td></td>
</tr>
</tbody>
</table>
{2987.17, 181.03, 262.73} and {40.30, 8.68, 27.93}. These results are very close to the results of previous analysis.

3. Testing the number of components in mixture models

3.1. Formulation

We consider the following $K$-component mixture model for independent vector observations $y_i$:

$$f(y_i) = \sum_{k=1}^{K} \pi_k f_k(y_i | \theta_k), \quad i = 1, \ldots, n,$$

where $f_k(\cdot | \theta_k)$ is the density for the $k$th group with parameter vector $\theta_k$, $\pi_k$ are component probabilities which are nonnegative and sum to 1, and $K$ is the unknown number of components. We denote the parameters of this model by $\theta = (\theta_1, \ldots, \theta_K)$ and $\pi = (\pi_1, \ldots, \pi_K)$. The preliminary objective is the inference about the unknown number of components; while Bayesian estimates of $\theta, \pi$ and their related statistics can be obtained as by-products.

Without loss of generality, for the observed data matrix $Y = (y_1, \ldots, y_n)$, let

$$M_1: Y|\theta, \pi \sim \sum_{k=1}^{K} \pi_k f_k(Y | \theta_k)$$

corresponding to a model of $K$ components, and

$$M_0: Y|\theta, \pi^* \sim \sum_{k=1}^{h} \pi_k^* f_k(Y | \theta_k), \quad h < K$$

corresponding to a model of $h$ components with component probabilities $\pi_k^*$, where $1 \leq h < K$. To apply the proposed method in computing $\log B_{10}$, these competing models are linked up by a path with $t \in [0, 1]$ as follows:

$$M_t: Y|\theta, \pi, t \sim \sum_{k=1}^{h} \pi_k^* f_k(Y | \theta_k) + t \sum_{k=h+1}^{K} \pi_k f_k(Y | \theta_k), \quad h < K$$

where $a_1, \ldots, a_h$ are given nonnegative weights such that $a_1 + \cdots + a_h = 1$. Clearly, when $t = 1$, $M_t$ reduces to $M_1$; when $t = 0$, $M_t$ reduces to $M_0$ with $a_k^* = \pi_k + a_k (\pi_{h+1} + \cdots + \pi_K)$, $k=1, \ldots, h$. The weights $a_k$ represent the increase of the corresponding component probabilities from a $K$-component model to a $h$-component model. A natural and simple suggestion for practical applications is to take $a_k = h^{-1}$. The log-likelihood function
can be written as
\[
\log p(Y|\theta, \pi, t) = \sum_{i=1}^{n} \log \left\{ \sum_{k=1}^{K} \left[ \pi_k + (1-t) a_k \sum_{m=h+1}^{K} \pi_m \right] f_k(y_i|\theta_k) \right. \\
\left. + \sum_{k=h+1}^{K} t \pi_k f_k(y_i|\theta_k) \right\}.
\]

By differentiation with respect to \( t \), we have
\[
U(Y, \theta, \pi, t) = \sum_{i=1}^{n} \frac{-\sum_{m=h+1}^{K} \pi_m \sum_{k=1}^{h} a_k f_k(y_i|\theta_k) + \sum_{k=h+1}^{K} \pi_k f_k(y_i|\theta_k)}{\sum_{k=1}^{K} \left[ \pi_k + (1-t) a_k \sum_{m=h+1}^{K} \pi_m \right] f_k(y_i|\theta_k) + \sum_{k=h+1}^{K} t \pi_k f_k(y_i|\theta_k)}.
\]

Estimate of \( \log B_{10} \) can be obtained via (5) and (6) with a sample of observations \{\{\theta^{(j)}, \pi^{(j)}\} : j = 1, \ldots, J\} that are simulated from \( p(\theta, \pi|Y, t_{(s)}) \) via the Gibbs sampler.

3.2. Implementation of the Gibbs sampler

The implementation plan of the Gibbs sampler in drawing observations from the posterior distribution \( p(\theta, \pi|Y, t_{(s)}) \) is conventional and similar to that given in Diebolt and Robert (1994). Based on the nature of mixture models, it is helpful to utilize the idea of data augmentation by introducing a group variable \( w_i \) for \( y_i \) as a latent allocation variable. The observed data \( Y \) will be augmented with the matrix of latent variables \( W = (w_1, \ldots, w_n) \) in the Gibbs sampler. We follow the common assumption that \( w_i \) is independently drawn from the following distribution:
\[
p(w_i = k) = \pi_k, \quad k = 1, \ldots, K
\]

and given \( w_i \), observations are drawn independently from the respective subpopulation. The Gibbs sampler is implemented as follows: At the \((j+1)\)th iteration with current \((\pi^{(j)}, \theta^{(j)})\): (i) generate \( W^{(j+1)} \) from \( p(W|Y, \pi^{(j)}, \theta^{(j)}, t) \); then (ii) generate \((\pi^{(j)}, \theta^{(j)})\) from \( p(\theta, \pi|Y, W^{(j+1)}, t) \). To keep our discussion concrete, our detailed exposition will be limited to the case of normal mixture so that \( \theta_k = (\mu_k, \Sigma_k) \), where \( \mu_k \) is the mean vector and \( \Sigma_k \) is the covariance matrix. However, the methodology is generic and can be applied to more general situations. Let \( \mu_k \) denotes the first component of \( \mu_k \), we assume that \( \mu_1(1) < \cdots < \mu_K(1) \) to avoid identification problems. Formulae for \( p(W|Y, \pi, \theta) \) and \( p(\theta, \pi|Y, W) \) are presented below; while \( p(W|Y, \pi, \theta, t) \) and \( p(\theta, \pi|Y, W, t) \) can be easily obtained by incorporating these formulae with \( t \).

Conditional distributions will be presented in the context of a \( K \)-component mixture model with unknown parameters \( \pi = (\pi_1, \ldots, \pi_K) \), \( \theta_k = (\mu_k, \Sigma_k) : k = 1, \ldots, K \}. We assume prior distributions of \( \pi, \mu_k, \Sigma_k \), \( k = 1, \ldots, K \) are mutually independent. Prior distribution of \( \pi \) is taken to be the symmetric Dirichlet distribution, \( \pi \propto D(\alpha, \ldots, \alpha) \). The following conjugate type prior distributions of \( \mu_k \) and \( \Sigma_k \) are used:
\[
\mu_k \propto N(\mu_{0k}, \Omega_{0k}), \quad \Sigma_k \propto IW[R_{0k}, \rho_{0k}],
\]
where $\mu_k$, $\Omega_0k$, $R_0k$, $\rho_0k$ are hyper-parameters, and $\text{IW}[\cdot,\cdot]$ denotes an inverted Wishart distribution which reduces to an inverted Gamma $\text{IG}(v_0k, \rho_0k)$ in univariate case. Let $\Omega_k = (\Omega_0k^{-1} + n_k \Sigma_k^{-1})^{-1}$, $R_k = (Y_k Y_k^T + R_0k^{-1})^{-1}$, $\bar{B}_k = n_k^{-1} \sum_{i:w_i = k} (y_i - \mu_k)$, where $n_k$ is the total number of $i$ such that $w_i = k$, $Y_k$ is sub-matrix of $Y$ such that all the $i$th columns with $w_i \neq k$ are deleted. It can be shown that conditional distributions of $\pi, \mu_k$, and $\Sigma_k$ are

\[
p(\pi|Y, W) = p(\pi|W) \propto D(x + n_1, \ldots, x + n_K),
\]
\[
p(\mu_k|Y, \Sigma_k) \propto N[\Omega_k(n_k \Sigma_k^{-1} \bar{B}_k + \Omega_0k^{-1} \mu_0k), \Omega_k],
\]
\[
p(\Sigma_k|Y, \mu_k) \propto \text{IW}[R_k, \rho_0k + n_k].
\]

Consider the conditional distribution $p(W|Y, \theta, \pi)$. Based on the definition of the model and assumptions

\[
p(W|Y, \theta, \pi) = \prod_{i=1}^n p(w_i|y_i, \theta, \pi) \propto \prod_{i=1}^n p(w_i|\pi) p(y_i|w_i, \theta).
\]

It follows that

\[
p(w_i = r|y_i, \theta, \pi) = \frac{\pi_r f_r(y_i|\mu_r, \Sigma_r)}{\sum_{k=1}^K \pi_k f_k(y_i|\mu_k, \Sigma_k)}
\]

where $f_k(y_i|\mu_k, \Sigma_k)$ is the density function of $N[\mu_k, \Sigma_k]$. Hence, the required conditional distributions are some familiar distributions. Since simulating observations from these distributions is straightforward and requires little computational effort, the proposed method is efficient and simple to implement.

3.3. Illustrative examples

The first illustrative example is on the basis of well-known data set about velocities of 82 galaxies from well-separated conic sections of the Corona Borealis region, originally presented by Postman et al. (1986). See the histogram in Richardson and Green (1997). This data set has been analyzed by Carlin and Chib (1995), Chib (1995), Escobar and West (1995), Richardson and Green (1997), and Roeder and Wasserman (1997). Results of Carlin and Chib (1995) indicate that the Gibbs output from models with four or five components display nonvanishing serial correlations for extremely high lags. Hence, it is expected to encounter difficulties with convergence and nonidentifiability of parameters. Similar to Chib (1995), a model will be selected among competing models with less than or equal to three components.

Let $M_K$ be the following $K$-component model:

\[
M_K: f(y) = \sum_{k=1}^K \pi_k \phi(y_i|\mu_k, \sigma_k^2),
\]

where $\phi(\cdot|\cdot)$ denote the normal density function. Prior distributions of the parameters corresponding to this univariate case are taken as

\[
\pi \propto D(1, \ldots, 1), \quad \mu_k \propto N[20, 10^2], \quad \sigma_k^2 \propto \text{IG}(3, 20),
\]
which are the same as those given in Chib (1995). These priors reflect weak prior information about the parameters.

In applying our method to compute the logarithm Bayes factor $\log B_{kh}$, for comparing different models $M_k$ and $M_h$ with this data set, we take $a_k = h^{-1}$ and $S = 20$ grids over $[0,1]$. The convergence of the Gibbs sampler in simulating observations for computing $\hat{U}(s)$ in (6) is monitored by the ‘estimated potential scale reduction (EPSR)’ (Gelman and Rubin, 1992) values of the parameters and several parallel sequences of observations generated with significantly different starting values. To give some ideas about the convergence behaviors associated with the estimation of $\log B_{32}$, the EPSR values at $t = 1.0$ are presented in Fig. 1. It can be seen that the algorithm converged quickly within 200 iterations. Hence, we settled for a burn-in phase of 200 Gibbs iterations and further collected $J = 1000$ observations to get $\hat{U}(s)$. The total number of Gibbs iterations for a comparison of two models is about 24,000. The computed logarithm Bayes factors are: $\log B_{23} = 5.440$, and $\log B_{32} = 7.194$. According to the criterion given by Kass and Raftery (1995) for the logarithm scale of Bayes factors, a three-component model is selected. Basically, this conclusion agrees with those reported in the literature on the analysis of normal mixture model via posterior simulation, such as Carlin and Chib (1995), Chib (1995), and Roeder and Wasserman (1997). For completeness, Fig. 2 displays smooth estimates of predictive densities of $y_i$ obtained from 200 MCMC samples after spline smoothing.

Fig. 1. EPSR values against the number of iterations.
To study the sensitivity of the proposed method to prior inputs, the Bayes factors were estimated with differently perturbed hyper-parameters values. Results obtained from some perturbations corresponding to prior inputs of $\pi$, $\mu_k$, and $\sigma_k^2$ are given in Table 1. These results indicate that estimates of logarithm of Bayes factors under different prior inputs are reasonably stable and they clearly give the same conclusion that a three-component mixture model is recommended. As expected, Bayesian estimates under different priors are close; they are not reported to save space.

The second example is for illustrating the feasibility of the proposed method for mixture models with a larger number of parameters. A subset of the Fisher’s Iris data (Andrews and Herzberg, 1985) that had been analyzed by Lindsay and Basak (1993) was used. The data consist of 50 observations each of four-dimensional measurements on sepal length, sepal width, petal length and petal width on two species of iris (Setosa and Versicolor). In our analysis, the species label was ignored. Prior distributions of the parameters were taken as

$$\pi \propto D(1, 1, 1), \quad \mu_k \propto N(\mu_0, \Phi_0), \quad \Sigma_k \propto IW(0.25I, 20),$$

(17)

where $I$ is an identity matrix, $\mu_0$ and $\Phi_0$ are the sample mean and half of the sample covariance matrix obtained from the data. This choice of prior inputs is for the purpose of illustration only, and we are not necessary recommending these priors for substantive analyses of practical problems. Again, we took $a_k = h^{-1}$, and 20 grids over $[0, 1]$. We observed that the Gibbs sampler again converged within 200 iteration. The similar convergence summary is not presented to save space. After a burn-in phase of 200 iterations and a total of $J = 2000$ additional observations were collected to get $\widetilde{U}_i$. The estimated logarithm of Bayes factors are equal to $\log \widetilde{B}_{21} = 12.157$ and $\log \widetilde{B}_{32} = -0.544$. 

![Fig. 2. Smooth estimates of predictive densities of $y_i$ obtained from 200 MCMC samples.](image)
Table 2
Bayesian estimates of parameters in a 2-component model

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Parameter</th>
<th>Estimate</th>
<th>Parameter</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \pi_1 )</td>
<td>0.499</td>
<td>( \pi_1 )</td>
<td>0.145</td>
<td>( \pi_1 )</td>
<td>0.254</td>
</tr>
<tr>
<td>( \mu_1(1) )</td>
<td>5.003</td>
<td>( \mu_1(2) )</td>
<td>3.424</td>
<td>( \mu_1(3) )</td>
<td>0.247</td>
</tr>
<tr>
<td>( \mu_1(4) )</td>
<td>1.463</td>
<td>( \mu_2(1) )</td>
<td>5.935</td>
<td>( \mu_2(2) )</td>
<td>2.770</td>
</tr>
<tr>
<td>( \mu_2(3) )</td>
<td>4.257</td>
<td>( \mu_2(4) )</td>
<td>1.326</td>
<td>( \sigma_1(1,1) )</td>
<td>0.078</td>
</tr>
<tr>
<td>( \sigma_1(2,1) )</td>
<td>0.013</td>
<td>( \sigma_1(2,2) )</td>
<td>0.071</td>
<td>( \sigma_1(2,3) )</td>
<td>0.008</td>
</tr>
<tr>
<td>( \sigma_1(3,1) )</td>
<td>0.199</td>
<td>( \sigma_1(3,2) )</td>
<td>0.010</td>
<td>( \sigma_1(3,3) )</td>
<td>0.005</td>
</tr>
<tr>
<td>( \sigma_1(4,1) )</td>
<td>0.071</td>
<td>( \sigma_1(4,2) )</td>
<td>0.013</td>
<td>( \sigma_1(4,3) )</td>
<td>0.013</td>
</tr>
<tr>
<td>( \sigma_1(4,4) )</td>
<td>0.032</td>
<td>( \sigma_1(4,3) )</td>
<td>0.005</td>
<td>( \sigma_1(4,3) )</td>
<td>0.056</td>
</tr>
<tr>
<td>( \sigma_1(4,4) )</td>
<td>0.077</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

These results suggest a definite conclusion in selecting a two-component mixture model. For completeness, Bayesian estimates corresponding to this model are presented in Table 2. To give some rough idea about the sensitivity to prior inputs, the data set has been reanalyzed with some different hyper-parameters values. For example, by halving and doubling \( \phi_0 \) in (17), the estimated logarithm Bayes factors \( \{ \log B_{21}, \log B_{32} \} \) are equal to \( \{ -1.517, -0.837 \} \) and \( \{ 8.643, -0.866 \} \), respectively. The conclusion of selecting a two-component model is not changed with these different prior inputs.

4. Latent variable models with mixed continuous and polytomous variables

4.1. Computation of the Bayes factor for model selection

Consider a latent variable model with \( p \times 1 \) manifest random vector \( y_i \) that satisfies the following measurement model:

\[
y_i = \mu + A\zeta_i + \epsilon_i, \quad i = 1, \ldots, n,
\]

where \( \mu \) is a \( p \times 1 \) mean vector, \( A \) is a \( p \times q \) matrix, \( \zeta_i \) is a \( q \times 1 \) latent vector, \( \epsilon_i \) is a \( p \times 1 \) random vector of error measurements with distribution \( N(0, \Sigma_\epsilon) \) which is independent of \( \zeta_i \), and \( \Sigma_\epsilon \) is a diagonal matrix. Suppose the latent vector \( \zeta_i \) is partitioned into \((\eta_i^T, \xi_i^T)^T\), where \( \eta_i \) is a \( q_1 \times 1 \) vector, \( \xi_i \) is a \( q_2 \times 1 \) vector and \( q_1 + q_2 = q \). It is assumed that the partitioned latent vectors satisfy the following structural equation:

\[
\eta_i = \Pi \eta_i + \Gamma \xi_i + \delta_i,
\]

where \( \Pi \) and \( \Gamma \) are \( q_1 \times q_1 \) and \( q_1 \times q_2 \) matrices of unknown parameters such that \( I_{q_1} - \Pi \) is nonsingular, \( \xi_i \) and \( \delta_i \) are independently distributed as \( N(0, \Phi) \) and \( N(0, \Psi_\delta) \), respectively; and \( \Psi_\delta \) is a diagonal matrix. It can be shown that by selecting an appropriate partition of \( \Lambda \), the above model is equivalent to a LISREL (Jöreskog and Sörbom, 1996) model or the Bentler–Weeks’ model in EQS (Bentler, 1992). Assuming that \( y_i \) is composed with a \( r \times 1 \) vector \( x_i \) of observed measurements and a \( s \times 1 \) vector \( v_i \) of unobserved measurements, where \( r + s = p \). Without loss of generality,
let \( y_i = (x_i^T, v_i^T)^T \), and the information of \( v = (v_1, \ldots, v_s)^T \) is given by an observed polytomous vector \( z \) such that

\[
z = \begin{pmatrix}
  z(1) \\
  \vdots \\
  z(s)
\end{pmatrix}
\]

\[\begin{array}{c}
  \alpha_{1, z(1)} < v(1) \leq \alpha_{1, z(1)+1}, \\
  \vdots \\
  \alpha_{s, z(s)} < v(s) \leq \alpha_{s, z(s)+1},
\end{array}\]  

(20)

where \( z(k) \) is an integral value that belongs to the set \( \{0, 1, \ldots, b_k\} \) for \( k = 1, \ldots, s \); and \( \alpha_{k,0} = -\infty \), \( \alpha_{k,b_k+1} = \infty \). Hence, for the \( k \)th variable, there are \( b_k + 1 \) categories and \( \alpha_{k,j} \) are unknown threshold parameters that define the categories.

It has been pointed out by Lee et al. (1995) that models with polytomous variables are not identified without imposing identification conditions. To solve this problem, we use the method given in Shi and Lee (2000) to fix \( \alpha_{k,1} \) and \( \alpha_{k,b_k}, k = 1, \ldots, s \) at preassigned values. Moreover, we follow the common practice in structural equation modeling to identify the covariance structure of \( y_i \) by fixing appropriate elements in \( \Sigma, \Pi, \Psi_e, \Psi_\delta \) and \( \Phi \), and \( z \) that contains all unknown thresholds in \( \alpha_{k,j}, k = 1, \ldots, s \).

Let \( M_0 \) and \( M_1 \) be two identified competing models of interest. To apply the proposed method for estimating the Bayes factor, we need to define a path to link \( M_0 \) and \( M_1 \), and then simulate observations from the appropriate posterior distribution. To give a more specific illustration, consider the following models with different measurement and structural equations:

\[
M_0: y = \mu + A_0 \zeta + e, \quad \eta = \Pi \eta + \Gamma_0 \zeta + \delta,
\]

\[
M_1: y = \mu + A_1 \zeta + e, \quad \eta = \Pi \eta + \Gamma_1 \zeta + \delta.
\]

These two models can be linked up by \( t \in [0, 1] \) as follows:

\[
M_t: y = \mu + [(1 - t)A_0 + tA_1] \zeta + e, \quad \eta = \Pi \eta + [(1 - t)\Gamma_0 + t\Gamma_1] \zeta + \delta.
\]

For this complicated latent variable model \( M_t \), direct analysis of the posterior distribution on the basis of the observed data \( X = (x_1, \ldots, x_n) \) and \( Z = (z_1, \ldots, z_n) \) is very complicated. Inspired by the idea of data augmentation (Tanner and Wong, 1987), the observed data are augmented with the latent variables \( \Xi = (\zeta_1, \ldots, \zeta_n) \) and latent continuous measurements \( V = (v_1, \ldots, v_n) \) in the analysis. Then the complete-data log-likelihood is equal to

\[
\log p(X, Z, \Xi, V | \theta, \alpha, t) = \log p(X, Z, V | \Xi, \eta, \zeta, \alpha, t) + \log p(\Xi | \theta, \alpha, t)
\]

\[
= - \frac{1}{2} \left\{ (p + q) \log(2\pi) + n \log|\Psi_e| + n \log|\Psi_\delta| - 2n \log|I_{q1} - \Pi| + n \log|\Phi| \right\}
\]

\[
+ \sum_{i=1}^{n} (y_i - \mu - A_1 \zeta_i)^T \Psi_e^{-1} (y_i - \mu - A_1 \zeta_i)
\]
\[ \begin{aligned} &+ \sum_{i=1}^{n} (\eta_i - \Pi\eta_i - \Gamma_i\epsilon_i)\Phi^{-1}(\eta_i - \Pi\eta_i - \Gamma_i\epsilon_i) \\ &+ \sum_{i=1}^{n} \epsilon_i\Phi^{-1}\epsilon_i - \sum_{i=1}^{n} \log I(v_i \in A_i) \end{aligned} \]

where \( A_i = (1 - t)A_0 + tA_1, \Gamma_i = (1 - t)\Gamma_0 + t\Gamma_1, I(v \in A) \) is an indicator function which takes the value 1 if \( v \in A \) zero otherwise, and

\[ A_i = (x_{i,1}, x_{i,2}) \times \cdots \times (x_{i,2}, x_{i,2} + 1). \]

By differentiation with respect to \( t \), we have

\[ U(X, Z, \Xi, V, 0, x, t) = \sum_{i=1}^{n} \left[ (y_i - \mu - A_i\epsilon_i)\Phi^{-1}A_i\epsilon_i - (\eta_i - \Pi\eta_i - \Gamma_i\epsilon_i)\Phi^{-1}\Gamma_i\epsilon_i \right], \]

where \( A_0 = A_1 - A_0 \) and \( \Gamma_0 = \Gamma_1 - \Gamma_0 \). Observations from the posterior distribution \( p[\mu, A, \Pi, \Gamma, \Phi, \Psi_\epsilon, \Psi_\delta, x, \Xi, V | X, Z, t] \) are then simulated by a hybrid algorithm that combines the Gibbs sampler (Geman and Geman, 1984) and the Metropolis–Hastings (MH) algorithm (Metropolis et al., 1953; Hastings, 1970) for computing the Bayes factor via (5) and (6). The derivation of the required conditional distributions, which are the familiar Gamma, normal, univariate truncated normal and inverted Wishart distributions, and the implementation plan of the algorithm are standard and similar to that given in Shi and Lee (1998). To save space, the tedious details are not presented.

4.2. An illustrative real example

A small portion of the ICPSR data set collected in the project WORLD VALUES SURVEY 1981–1984 AND 1990–1993 (1994) (World Values Study Group, ICPSR Version) is analyzed for illustration. The whole data set was collected in 45 societies around the world on broad topics such as work, religious belief, meaning and purpose of life, family life, contemporary social issues, etc. The data obtained from Canada are used to illustrate our proposed method. Eight variables in the original data set (with corresponding question numbers 180, 96, 116, 117, 62, 179, 252, and 254) that are related with respondent’s job, religious belief, and homelife were taken as manifest variables in \((y_{1,1}, \ldots, y_{1,8})^{T}\). After deleting cases with missing entries, the sample size is 470. Among them, \((y_{1,1}, y_{1,2})\) are related to life, \((y_{1,3}, y_{1,4})\) are related to job satisfaction, \((y_{1,5}, y_{1,6})\) are related to religious belief, and \((y_{1,7}, y_{1,8})\) are related to job attitude. Measurements associated with \(y_{1,5}\) and \(y_{1,6}\) were based on a five-point scale, while all others were measured by a 10-point scale. Thus, variables \(y_{1,5}\) and \(y_{1,6}\) were treated as polytomous and the others were treated as continuous.
Table 3
Bayesian estimates of parameters in latent variable model $M_2$

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Est</th>
<th>Parameter</th>
<th>Est</th>
<th>Parameter</th>
<th>Est</th>
<th>Parameter</th>
<th>Est</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda_{21}$</td>
<td>1.006</td>
<td>$\beta_{21}$</td>
<td>-0.112</td>
<td>$\psi_1$</td>
<td>1.112</td>
<td>$\mu_1$</td>
<td>8.414</td>
</tr>
<tr>
<td>$\lambda_{42}$</td>
<td>1.015</td>
<td>$\gamma_{11}$</td>
<td>0.136</td>
<td>$\psi_2$</td>
<td>0.853</td>
<td>$\mu_2$</td>
<td>7.999</td>
</tr>
<tr>
<td>$\lambda_{63}$</td>
<td>-1.004</td>
<td>$\gamma_{22}$</td>
<td>-0.539</td>
<td>$\psi_3$</td>
<td>4.087</td>
<td>$\mu_3$</td>
<td>3.871</td>
</tr>
<tr>
<td>$\lambda_{84}$</td>
<td>1.584</td>
<td>$\phi_{11}$</td>
<td>0.453</td>
<td>$\psi_4$</td>
<td>2.714</td>
<td>$\mu_4$</td>
<td>3.204</td>
</tr>
<tr>
<td>$\alpha_{12}$</td>
<td>-0.166</td>
<td>$\phi_{12}$</td>
<td>0.086</td>
<td>$\psi_5$</td>
<td>0.547</td>
<td>$\mu_5$</td>
<td>0.007</td>
</tr>
<tr>
<td>$\alpha_{13}$</td>
<td>0.332</td>
<td>$\phi_{22}$</td>
<td>0.890</td>
<td>$\psi_6$</td>
<td>0.558</td>
<td>$\mu_6$</td>
<td>-0.014</td>
</tr>
<tr>
<td>$\alpha_{22}$</td>
<td>0.479</td>
<td>$\psi_1$</td>
<td>1.309</td>
<td>$\psi_7$</td>
<td>1.669</td>
<td>$\mu_7$</td>
<td>8.049</td>
</tr>
<tr>
<td>$\alpha_{23}$</td>
<td>0.733</td>
<td>$\psi_2$</td>
<td>1.925</td>
<td>$\psi_8$</td>
<td>2.289</td>
<td>$\mu_8$</td>
<td>7.446</td>
</tr>
</tbody>
</table>

As an illustration, consider the following latent variable model:

$$y = \mu + A^T \zeta + \epsilon,$$

where $1$'s and $0$'s are fixed parameters, and $\zeta = (\eta_1, \eta_2, \xi_1, \xi_2)^T$. In this specification, the latent variables in $\zeta$ can be interpreted as factors about life, job satisfaction, religious belief and job attitude. Suppose we are interested in the following nonnested models whose measurement equations are the same as given in (21) but with different structural equations:

- $M_0$: $\left( \begin{array}{c} \eta_1 \\ \eta_2 \end{array} \right) = \left( \begin{array}{c} \gamma_{11} \\ \gamma_{21} \end{array} \right) \left( \begin{array}{c} \xi_1 \\ \xi_2 \end{array} \right) + \left( \begin{array}{c} \delta_1 \\ \delta_2 \end{array} \right)$,
- $M_1$: $\left( \begin{array}{c} \eta_1 \\ \eta_2 \end{array} \right) = \left( \begin{array}{c} 0 \\ \beta_{21} \end{array} \right) \left( \begin{array}{c} \eta_1 \\ \eta_2 \end{array} \right) + \left( \begin{array}{c} \gamma_{11} \\ 0 \end{array} \right) \left( \begin{array}{c} \xi_1 \\ \xi_2 \end{array} \right) + \left( \begin{array}{c} \delta_1 \\ \delta_2 \end{array} \right)$,
- $M_2$: $\left( \begin{array}{c} \eta_1 \\ \eta_2 \end{array} \right) = \left( \begin{array}{c} 0 \\ \beta_{21} \end{array} \right) \left( \begin{array}{c} \eta_1 \\ \eta_2 \end{array} \right) + \left( \begin{array}{c} 0 \\ \gamma_{12} \end{array} \right) \left( \begin{array}{c} \xi_1 \\ \xi_2 \end{array} \right) + \left( \begin{array}{c} \delta_1 \\ \delta_2 \end{array} \right)$.

The proposed method for computing Bayes factors is applied to select one of these competing models. The number of grids in $[0,1]$ is taken to be $20$. In the implementation of the hybrid algorithm that combines the Gibbs sampler and MH algorithm, we used conjugate prior distributions with hyper-parameters values given by estimates obtained from an initial run with non-informative priors. We observed that the algorithm converged quickly within 500 iterations. To save space, the similar convergence behaviors are not presented. For each $t(s)$, an additional 2000 observations were collected after convergence for computing $\hat{B}_{01}$ and $\hat{B}_{12}$. We found that $\log \hat{B}_{01} = -0.206$ and $\log \hat{B}_{12} = 6.588$. These results suggest that $M_2$ is better than $M_1$ and is significantly better than $M_3$. Hence, $M_2$ is selected. For completeness, the Bayesian estimates of the structural parameters associated with $M_2$ are presented in Table 3. Other statistics, such as the highest probability density (HPD) intervals (Chen and Shao, 1999), can...
be obtained by standard methods using the simulated observations. The data have been reanalyzed using different prior inputs. We obtain close estimates of the logarithm of Bayes factors and the same conclusion as above.

## 5. Discussion

We have introduced a method for computing the logarithm of Bayes factor. The main idea is utilization of the powerful path sampling for estimating the ratio of the normalizing constants. The procedure is simple to implement, although it requires a little insight to construct a path that links up the competing models. Its main computational task is to draw a simulated sample of the parameters from the posterior distribution at a number of grids in [0,1]. In the illustrative examples, the number of grids is taken to be 20. We have reanalyzed all the examples with only 10 grids in [0,1] and the results obtained are close to those obtained with 20 grids. For example, in the example with latent variable models, the estimated logarithm of Bayes factors with 10 grids are equal to $\log B_{01} = -0.211$ and $\log B_{12} = 6.655$. Clearly, half of the computational effort can be saved by using 10 grids in [0,1]. Using a SUN Enterprise 4500 Server, the computation times for calculating one logarithm Bayes factor in each of the numerical examples are given in Table 4.

The important findings in a comparative study of DiCiccio et al. (1997) are that methods for computing Bayes factors via posterior simulations are reasonably accurate and the bridge sampling provides a substantial improvement. Gelman and Meng (1998) showed that the limit using infinitely many bridges in the bridge sampling leads to the path sampling. Hence, path sampling can be regarded as an extension of bridge sampling, and we naturally expect that it has potential for even more improvement. Another method for estimating Bayes factor via posterior simulation is the procedure proposed by Chib (1995). It has been applied to a multiple-sample factor analysis model with polytomous variables (Song and Lee, 2001). The following logarithm of the marginal density:

$$\log p(y) = \log p(y|\theta^*) + \log p(\theta^*) - \log p(\theta^*|y)$$

under one of the competing models is computed, where $\theta^*$ is usually taken to be the underlying ML estimate of $\theta$. Its key feature is to estimate the posterior density $p(\theta^*|y)$ by a sequence of simulations followed by a sequence of lower-dimensional density estimates. In addition, the likelihood function $p(y|\theta^*)$ is needed to be evaluated. This is an efficient procedure for models with a small number of parameters.
However, when applying to complicated models that involve a large number of parameters and latent quantities introduced to simplify simulation, the computational burden is likely to be heavy. Take the latent variable model as described in Section 4.1 as an example, the corresponding posterior density $p[\mu, \Lambda, \Gamma, \Phi, \Psi, \Psi', z, \Xi, \Sigma, V | X, Z]$ requires a considerable amount of computational effort to evaluate. Moreover, it also requires substantial effort to compute the ML estimate and the likelihood function that involves intractable multiple integrals. In contrast, the proposed procedure is more simple to implement.

Meng and Schilling (1998) pointed out that given two models having unnormalized densities with the same support, say $p_0(\theta)$ and $p_1(\theta)$, we can always construct a continuous path to link them. For example, the density of the linked model using a geometric path with a scalar parameter $t$ in $[0,1]$ is $p_0(\theta)^{1-t}p_1(\theta)^t$. However, for some complex situations, it is difficult to sample from the posterior distributions relating to this kind of linked models. Hence, it is difficult to apply the path sampling under situations where a simple path is not available. This may be a limitation of the path sampling. We conjecture that for a sufficient large $S$, solutions obtained from different paths are close to each other, because they are good approximations of the same integral given in (4). Of course, further simulation studies are required to draw a definite conclusion.

Acknowledgements

This research is fully supported by a grant from the Research Grant Council of the Hong Kong Special Administrative Region (Project No. CUHK 4346/01H). The authors are thankful to anonymous reviewers for valuable comments, and grateful to ICPSR and the relevant funding agency for allowing use of the data in the example.

References


