Statistical Science
2013, Vol. 28, No. 1, 95–115
DOI: 10.1214/12-STS403

Bayes Model Selection with Path Sampling: Factor Models and Other Examples

Ritabrata Dutta and Jayanta K. Ghosh

Abstract. We prove a theorem justifying the regularity conditions which are needed for Path Sampling in Factor Models. We then show that the remaining ingredient, namely, MCMC for calculating the integrand at each point in the path, may be seriously flawed, leading to wrong estimates of Bayes factors. We provide a new method of Path Sampling (with Small Change) that works much better than standard Path Sampling in the sense of estimating the Bayes factor better and choosing the correct model more often. When the more complex factor model is true, PS-SC is substantially more accurate. New MCMC diagnostics is provided for these problems in support of our conclusions and recommendations. Some of our ideas for diagnostics and improvement in computation through small changes should apply to other methods of computation of the Bayes factor for model selection.

Key words and phrases: Bayes model selection, covariance models, path sampling, Laplace approximation.

1. BAYES MODEL SELECTION

Advances in MCMC techniques to compute the posterior for many complex, hierarchical models have been a major reason for success in Bayes modeling and analysis of complex phenomena (Andrieu, Doucet and Robert, 2004). These techniques along with applications are surveyed in numerous papers, including Chen, Shao and Ibrahim (2000), Liu (2008) and Robert and Casella (2004). Moreover, many Bayesian books on applications or theory and methods provide a quick introduction to MCMC, such as Gelman et al. (2004), Ghosh, Delampady and Samanta (2006), Gamerman and Lopes (2006) and Lynch (2007).

Just as the posterior for the parameters of a given model are important for calculating Bayes estimates, posterior variance, credibility intervals and a general description of the uncertainty involved, one needs to calculate Bayes factors for selecting one of several models. Bayes factors are the ratio of marginals of given data under different models, when more than one model is involved and one wishes to choose one from among them, based on their relative or posterior probability. The ratio of marginals measures the relative posterior probability or credibility of one model with respect to the other if we make the usual objective choice of half as prior probability for each model.

Although there are many methods for calculating Bayes factors, their success in handling complex modern models is far more limited than seems to be generally recognized. Part of the reason for lack of awareness of this is that model selection has become important relatively recently. Also, one may think that, in principle, calculation of a BF can be reduced to the calculation of a posterior, and hence solvable by the same methods as those for calculating the posterior. Reversible Jump MCMC (RJMCMC) is an innovative methodology due to Green (1995), based on this simple fact. However, two models essentially lead to two different sets of states for any Markov chain that connects them. The state spaces for different models often differ widely in their dimension. This may prevent good mixing and may show up in the known difficulties of tuning RJMCMC. For a discussion of tuning difficulties see Robert and Casella (2004).

Ritabrata Dutta is Ph.D. Student and Jayanta K. Ghosh is Professor, Department of Statistics, Purdue University, Lafayette, Indiana 47907, USA (e-mail: rdutta@purdue.edu; statrita2004@gmail.com; ghosh@purdue.edu).

Another popular method for calculating BF is path sampling (PS), which is due to Gelman and Meng (1998) and recently re-examined by Lefebvre et al. (2009). Our major goal is to explore PS further in the context of nested, relatively high-dimensional covariance models, rather than nonnested low-dimensional mean models, as in the last reference. The new examples show both similarities and sharp changes from the sort of behavior documented in Lefebvre et al. (2009).

We consider three paths, namely, the geometric mean path, the arithmetic mean path and the parametric arithmetic mean path, which appear in Gelman and Meng (1998), Lefebvre et al. (2009), Ghosh and Dunson (2008), Ghosh and Dunson (2009), Lee and Song (2002) and Song and Lee (2006). Other applications of path sampling and bridge sampling (with some modifications) appear in Lartillot and Philippe (2006), Friel and Pettitt (2008), Xie et al. (2011) and Fan et al. (2011). Our priors are usually the diffuse Cauchy priors, first suggested by Jeffreys (1961) and since then recommended by many others, including Berger (personal communication), Liang et al. (2008), Gelman (2006) and Ghosh and Dunson (2009). But we also examine other less diffuse priors too, going all the way to normal priors. Since Lefebvre et al. (2009) have studied applications of PS to mean like parameters, we focus on covariance models. We restrict ourselves generally to factor models for covariance, which have become quite popular in recent applications, for example, Lopes and West (2004), Ghosh and Dunson (2008), Ghosh and Dunson (2009) and Lee and Song (2002). The recent popularity of factor models is due to the relative ease with which they may be used to provide a sparse representation of the covariance matrix of multivariate normal data in many applied problems of finance, psychometry and epidemiology; see, for example, the last three references. Also, often it leads to interesting scientific insight; see Bartholomew et al. (2002).

In addition to prior, likelihood and path, there are other choices to be made before PS can be implemented, namely, a method of discretizing the path, for example, by equispaced points or adaptively (Lefebvre et al., 2009) and how to integrate the score function of Gelman and Meng (1998) at each point in the discrete path. A popular method is to use MCMC. These more technical choices are discussed later in the paper. Along with PS, we will consider other methods like Importance Sampling (IS) and its descendants like Annealed Importance Sampling (AIS), due to Neal (2001), and Bridge Sampling (BS), due to Meng and Wong (1996).

We now summarize our contribution in this paper.

In Section 2 we review what is known about path sampling and factor models. We introduce factor models, a suitable path and suitable diffuse t-priors. The path we use was first introduced in Gelman and Meng (1998) for mean models and by Lee and Song (2002) and Ghosh and Dunson (2009) for factor models.

In Section 2.4 we prove a theorem (Theorem 2.1) which essentially shows that except for the convergence of MCMC estimates for expected score function $E_t(U(\theta, t))$ at each grid point t in the path, all other needed conditions for PS will hold for our chosen path, prior and likelihood for factor models. In one of the remarks following the theorem we generalize this result to other paths. Remark 3 points to the need for some finite moments for the prior, not just for Theorem 2.1 to hold but for the posterior to behave well. Then in Remark 5 we provide a detailed, heuristic argument as to why the MCMC may fail dramatically by not mixing properly if the data has come from the bigger of the two models under consideration. If our heuristics is correct, and there is a small interval where $E_t(U(\theta, t))$ oscillates most, then a grid size that is a bit coarse will not only be a bit inaccurate, it will be very wrong. Even if the grid size is sufficiently small, one will need to do MCMC several times with different starting points just to realize PS will not work. Our new proposal avoids these problems but will require more time if many models are involved.

In Section 3 we give an argument as to why the above is unlikely to be true if the data has come from the smaller model. More importantly, in Section 3.3 we propose a modification of PS, which we call Path Sampling with Small Change (PS-SC) which is expected to do better.

Implementation of PS and PS-SC can be very time consuming due to the need of MCMC sampling for each grid point along the path. Time can be saved if we can implement PS and PS-SC by parallel computation, as noted by Gelman and Meng (1998).

In Section 3.4 we show MCMC output for the various cases discussed and validate our heuristics above. The diagnostics via projection into likelihood space should prove useful for other model selection problems. Our gold standard is PS-SC, based on an MCMC with the number of draws $m = 50{,}000$ and burn-in of 1000, if necessary. But actually in our examples $m = 6000$ and burn-in of 1000 suffices for PS-SC. For other model selection rules we also go up to $m = 50{,}000$ if

necessary. After Section 3.4, having shown our modified PS, namely, PS-SC, is superior to PS under both models, we do not consider PS in the rest of the paper.

In the last two sections we touch on the following related topics: effects of grid size, alternative path, alternative methods and performance of PS-SC and some other methods in very high-dimensional simulated and real examples. PS-SC seems to choose the true models in the simulated cases and relatively conservative models for real data. In Section 5 we explore various real life and high-dimensional factor models, with the object of combining PS-SC with two of the methods which do relatively well in Section 4 to reduce the time of PS-SC in problems with the number of factors rather high, say, 20 or 26, for which PS-SC can be quite slow. For these high-dimensional examples, we use Laplace approximation to marginals for preliminary screening of models. A few general comments on Laplace approximation in high-dimensional problems are in Section 5.

In Appendix A.1 we introduce briefly a few other methods like Annealed Importance Sampling (AIS) which we have compared with PS-SC. Finally, Appendix A.2 points to some striking differences between what we observe in factor models and what one might have expected from our familiar classical asymptotics for maximum likelihood estimates. Of course, as pointed out by Drton (2009), classical asymptotics does not apply here, but it surprised us that the differences would be so stark. It is interesting and worth pointing out that the Bayes methods like PS-SC can be validated partly theoretically and partly numerically in spite of a lack of suitable asymptotic theory.

2. PATH SAMPLING AND FACTOR MODELS

In the following subsections we review some basic facts about PS, including the definition of the three paths and the notion of an optimal path. More importantly, since our interest would be in model selection for covariance rather than mean, we introduce factor models and then PS for factor models in Sections 2.3 and 2.4.

Section 2.1 is mostly an introduction to PS and reviews previous work. After that we show the failure of PS-estimates in a toy problem related to the modeling of the covarince matrix in Section 2.2. In Section 2.3 we introduce factor models and our priors. Section 2.4 introduces paths that we consider for factor models and a theorem showing the regularity conditions needed for validity of PS under factor models. Then in a series of remarks we extend the theorem and also study and explain how the remaining ingredient of PS, namely, MCMC, can go wrong. We show a few MCMC outputs to support our arguments in Section 3.4. This particular theme is very important and will come up several times in later sections or subsections where related different aspects will be presented.

2.1 Path Sampling

Among the many different methods related to importance sampling, the most popular is Path Sampling (PS). However, PS is best understood as a limit of the simpler Bridge Sampling (BS) (Gelman and Meng, 1998). So we first begin with BS.

It is well known that unless the densities of the sampling and target distributions are close in relative importance sampling weights, Importance Sampling (IS) will have high variance as well as high bias. Due to the difficulty of finding a suitable sampling distribution for IS, one may try to reduce the difficulty by introducing a nonnormalized intermediate density $f_{1/2}$ that acts like a bridge between the nonnormalized sampling density f_1 and nonnormalized target density f_0 (Meng and Wong, 1996). One can then use the identity $Z_1/Z_0 = \frac{Z_{1/2}/Z_0}{Z_{1/2}/Z_1}$ and estimate both the numerator and denominator by IS. Extending this idea, Gelman and Meng (1998) constructed a whole path $f_t : t \in [0, 1]$ connecting f_0 and f_1. This is also like a bridge. Discretizing this, they get the identity $Z_1/Z_0 = \prod_{l=1}^{L} \frac{Z_{(l-1/2)}/Z_{(l-1)}}{Z_{(l-1/2)}/Z_{(l)}}$, which leads to a chain of IS estimates in the numerator and denominator. We call this estimate the Generalized Bridge Sampling (GBS) estimate.

More importantly, Gelman and Meng (1998) introduced PS, which is a new scheme, using the idea of a continuous version of GBS but using the log scale. The PS estimate is calculated by first constructing a path as in BS. Suppose the path is given by $p_t : t \in [0, 1]$ where for each t, p_t is a probability density. Then we have the following definition:

$$p_t(\theta) = \frac{1}{z_t} f_t(\theta), \tag{2.1}$$

where f_t is an unnormalized density and $z_t = \int f_t(\theta)\, d\theta$ is the normalizing constant. Taking the derivative of the logarithm on both sides, we obtain the following identity under the assumption of interchangeability of the order of integration and differentiation:

$$\begin{aligned} \frac{d}{dt}\log(z_t) &= \int \frac{1}{z_t}\frac{d}{dt} f_t(\theta)\mu(d\theta) \\ &= E_t\left[\frac{d}{dt}\log f_t(\theta)\right] = E_t\big[U(\theta, t)\big], \end{aligned} \tag{2.2}$$

where the expectation E_t is taken with respect to $p_t(\theta)$ and $U(\theta, t) = \frac{d}{dt} \log f_t(\theta)$. Now integrating (2.2) from 0 to 1 gives the log of the ratio of the normalizing constants, that is, log BF in the context of model selection:

$$\log\left[\frac{Z_1}{Z_0}\right] = \int_0^1 E_t\big[U(\theta, t)\big]\, dt. \tag{2.3}$$

To approximate the integral, we discretize the path with k points $t_{(0)} = 0 < t_{(1)} < \cdots < t_{(k)} = 1$ and draw m MCMC samples converging to $p_t(\theta)$ at each of these k points. Then estimate $E_t[U(\theta, t)]$ by $\frac{1}{m}\sum U(\theta^{(i)}, t)$ where $\theta^{(i)}$ is the MCMC output. To estimate the final log Bayes factor, commonly numerical integration schemes are used. It is clear that MCMC at different points "t" on the path can be done in parallel. We have used this both for PS and for our modification of it, namely, PS-SC introduced in Section 3.3.

Gelman and Meng (1998) showed there is an optimum path in the whole distribution space providing a lower bound for MCMC variance, namely,

$$\left[\arctan \frac{H(f_0, f_1)}{\sqrt{4 - H^2(f_0, f_1)}}\right]^2 \Big/ m,$$

where f_0 and f_1 are the densities corresponding to the two models compared and $H(f_0, f_1)$ is their Hellinger distance. Unfortunately in nested examples f_0 and f_1 are mutually orthogonal, so $H(f_0, f_1)$ takes the trivial value of two. Moreover, m is so large that the lower bound becomes trivial and unattainable. However, in a given problem, one path may be more suitable or convenient than another.

Following Gelman and Meng (1998) and Lefebvre et al. (2009), we consider three paths generally used for the implementation of PS. The Geometric Mean Path (GMP) and Arithmetic Mean Path (AMP) are defined by the mean [$f_t = f_0^{(1-t)} f_1^t$ and $f_t = tf_0 + (1-t)f_1$, resp.] of the densities of two competing models for each model $M_t : t \in (0, 1)$ along the path. Our notation for the Bayes factor is given later in equation (2.6).

One more common path is obtained by assuming a specific functional form f_θ for the density and then constructing the path in the parametric space ($\theta \in \Theta$) of the assumed density. If $\theta_t = t\theta_0 + (1-t)\theta_1$, then f_{t,θ_t} is the density of the model M_t, where $f_{0,\theta_0} = f_0$ and $f_{1,\theta_1} = f_1$. We denote this third path as the Parametric Arithmetic Mean Path (PAMP). The PAMP path was shown by Gelman and Meng (1998) to minimize the Rao distance in a path for model selection about normal means. More importantly, it is very convenient for use of MCMC, as shown for some factor models by Song and Lee (2006) and Ghosh and Dunson (2009), and for linear models by Lefebvre et al. (2009). Implementation of PS with the paths mentioned above is denoted as GMP-PS, AMP-PS and PAMP-PS. In view of the discussion in Lefebvre et al. (2009) regarding the degeneracy of the AMP-PS, we will only consider PAMP-PS and GMP-PS.

Unlike Lefebvre et al. (2009), who study models about means, our interest is in studying model selection for covariance models, specifically factor models with different number of factors. These are discussed in the Sections 2.3 and 2.4. Performance of PS for covariance models can be very different from the examples in Lefebvre et al. (2009). In the next subsection we give a toy example of covariance model selection where PS fails and our proposed modification PS-SC is also not applicable.

2.2 Covariance Model: Toy Example

To illustrate the difficulties in calculation of the BF that we discuss later, we begin by considering a problem where we can calculate the true value of the Bayes factor.

Assuming $Y_p \sim N(0, \Sigma)$, for some $m < p$ we wish to test whether $Y_{1,\ldots,m}$ and $Y_{m+1,\ldots,p}$ are independent or not. If $\Sigma = \left(\begin{smallmatrix} A_{11} & A_{12} \\ A_{12}' & A_{22} \end{smallmatrix}\right)$ where $Y_{1,\ldots,m} \sim N(0, A_{11})$ and $Y_{m+1,\ldots,p} \sim N(0, A_{22})$, then the competitive models for a fixed m will be $M_0 : A_{12} = 0$ vs $M_1 : A_{12} \neq 0$. Under M_1 we use an inverse-Wishart prior for the covariance matrix, as it helps us to calculate the true BF, using the conjugacy property of the prior. Under M_0 we take A_{11}, A_{22} to be independent, each with an inverse Wishart prior.

We illustrate the above problem with $p = 10$ and $m = 7$ for a positive definite matrix $\Sigma^0 = \left(\begin{smallmatrix} A_{11}^0 & A_{12}^0 \\ (A_{12}^0)' & A_{22}^0 \end{smallmatrix}\right)$ (given in Appendix A.3). We implement the path sampling for this problem connecting M_0 and M_1, using a Parametric Arithmetic Mean Path:

$$M_t : y_i \sim N\left(0, \Sigma = \begin{pmatrix} A_{11}^0 & tA_{12}^0 \\ t(A_{12}^0)' & A_{22}^0 \end{pmatrix}\right). \tag{2.4}$$

For every $0 \le t \le 1$, the Σ matrix is positive definite, being a convex combination of two positive definite matrices. For $t = 0$ and $t = 1$ we get the models M_0 and M_1.

We can estimate the Bayes factor by using the path sampling schemes as described earlier. We simulated two data sets, one each from M_0 and M_1, and report the true BF value with the PS estimate in Table 1. Here

TABLE 1
Performance of PS in toy example modeling covariance: Log Bayes factor (MCMC-standard deviation)

Method	Data 1	Data 2
True BF value	258.38	−132.87
PS estimate of BF	184.59 (0.012)	−20.11 (0.008)

the reported Bayes factor is defined as the ratio $\frac{m_1}{m_0}$, where m_1 and m_0 are the marginals under the models M_1 and M_0, respectively.

The values in the table show us that the estimated BF value is off by an order of magnitude when M_0 is true. The value is relatively stable as judged by the MCMC-standard deviation based on 10 runs and near to the true value for M_1.

2.3 Factor Models and Bayesian Specification of Prior

A factor model with k factors is defined as n i.i.d. observed r.v.'s

$$y_i = \Lambda\eta_i + \varepsilon_i, \quad \varepsilon_i \sim N_p(0, \Sigma),$$

where Λ is a $p \times k$ matrix of factor loadings,

$$\eta_i = (\eta_{i1}, \ldots, \eta_{ik})' \sim N_k(0, I_k)$$

is a vector of standard normal latent factors, and ε_i is the residual with diagonal covariance matrix $\Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2)$. Thus, we may write the marginal distribution of y_i as $N_p(0, \Omega)$, $\Omega = \Lambda\Lambda' + \Sigma$. This model implies that the sharing of common latent factors explains the dependence in the outcomes and the outcome variables are uncorrelated given the latent factors.

A factor model, without any other constraints, is nonidentifiable under orthogonal rotation. Post-multiplying Λ by an orthogonal matrix P, where P is such that $PP' = I_k$, we obtain exactly the same Ω as in the previous factor model. To avoid this, it is customary to assume that Λ has a full-rank lower triangular structure, restricting the number of free parameters in Λ and Σ to $q = p(k+1) - k(k-1)/2$, where k must be chosen so that $q \le p(p+1)/2$. The reciprocal of diagonal entries of Σ forms the precision vector here.

It is well known that maximum likelihood estimates for parameters in factor models may lie on boundaries and, hence, likelihood equations may not hold (Anderson, 1984). The Bayes estimate of Ω defined as average over MCMC outputs is well defined, easy to calculate and, being average of positive definite matrices, is easily seen to be positive definite. This fact is used to search for maximum likelihood estimates (mle) or maximum prior×likelihood estimates (mple) in a neighborhood of the Bayes estimate.

We also note for later use the following well-known simple fact, for example, Anderson (1984). If the likelihood is maximized over all positive definite matrices Ω, not just over factor models, then the global maximum for n independent observations exists and is given by

$$\hat{\Omega} = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})(y_i - \bar{y})'. \tag{2.5}$$

From the Bayes model selection perspective, a specification of the prior distribution for the free elements of Λ and Σ is needed. Truncated normal priors for the diagonal elements of Λ, normal priors for the lower triangular elements and inverse-gamma priors for $\sigma_1^2, \ldots, \sigma_p^2$ have been commonly used in practice due to conjugacy and the resulting simplification in posterior distribution. Prior elicitation is not common.

Ghosh and Dunson (2009) addressed the above problems by using the idea of Gelman (2006) to introduce a new class of default priors for the factor loadings that have good mixing properties. They used the Gibbs sampling scheme and showed there was good mixing and convergence. They used parameter expansion to induce a class of t or folded t-priors depending on sign constraints on the loadings. Suitable t-priors have been very popular. We use the same family of priors but consider a whole range of many degrees of freedom going all the way to the normal and use the same Gibbs sampler as in Ghosh and Dunson (2008). We have used a modified version of their code.

In the factor model framework, we stick to the convention of denoting the Bayes factor for two models with latent factors $h-1$ and h as

$$BF_{h,h-1} = \frac{m_h(x)}{m_{h-1}(x)}, \tag{2.6}$$

where $m_h(x)$ is the marginal under the model having h latent factors. *So the Bayes factor for the simpler model (defined as M_0) and complex model (defined as M_1) with $h-1$ and h latent factors will be defined as $BF_{h,h-1}$.* We choose the model with h and $h-1$ latent factors, respectively, depending on the value of the log Bayes factor being positive and negative. Alternatively, one may choose a model only when the value of $\log BF$ is decisively negative or positive, say, less than or greater than a chosen threshold.

2.4 Path Sampling for Factor Models

There are several variants of path sampling which have been explored in different setups, depending on choice of path, prior and other tuning parameters (grid size and MCMC sample size). In the factor model setup the parametric arithmetic mean path (PAMP) [used by Song and Lee (2006) and Ghosh and Dunson (2009)] seems to be the most popular one. We also consider Geometric Mean Path (GMP) along with the PAMP for the factor model.

By constructing a GM path from corresponding prior to the posterior, we can estimate the value of the log-marginal under both M_0 and M_1, which in turn leads us to an estimate of the log-BF. We will first describe the two paths and their corresponding score functions to be estimated along the path.

(i) *Parametric arithmetic mean path*: Lee and Song (2002) used this path in factor models, following an example in Gelman and Meng (1998). Ghosh and Dunson (2008) also used this path along with parameter expansion. Here we define M_0 and M_1 to be the two models corresponding to the factor model with factors $h-1$ and h, respectively, and then connect them by the path $M_t : y_i = \Lambda_t \eta_i + \varepsilon_i$, $\Lambda_t = (\lambda_1, \lambda_2, \ldots, \lambda_{h-1}, t\lambda_h)$, where λ_j is the jth column of the loading matrix. So for $t = 0$ and $t = 1$ we get the models M_0 and M_1. The likelihood function at grid point t is a MVN which is denoted as $f(Y|\Lambda, \Sigma, \eta, t)$. We have independent priors $\pi(\Lambda), \pi(\Sigma), \pi(\eta)$ and a score function,

$$U(\Lambda, \Sigma, \eta, Y, t) = \sum_{i=1}^{n} (y_i - \Lambda_t \eta_i)' \Sigma^{-1} \big(0^{p\times(h-1)}, \lambda_h\big)\eta_i. \tag{2.7}$$

For fixed and ordered grid points along the path $t_{(0)} = 0 < t_{(1)} < \cdots < t_{(S)} < t_{(S+1)} = 1$, our path sampling estimate for the log Bayes factor is

$$\log(\widehat{BF}_{h:h-1}) = \frac{1}{2}\sum_{s=0}^{S} (t_{s+1} - t_s)\big(\widehat{E}_{s+1}(U) + \widehat{E}_s(U)\big). \tag{2.8}$$

We simulate m samples of $(\Lambda_{t_s,i}, \Sigma_i, \eta_i : i = 1, \ldots, m)$ from the posterior distribution of $(\Lambda_{t_s}, \Sigma, \eta)$ at the point $0 \le t_s \le 1$ and use them to estimate $\widehat{E}_s(U) = \frac{1}{m}\sum U(\Lambda_{t_s,i}, \Sigma_i, \eta_i, y), \forall s = 1, \ldots, S+1$.

(ii) *Geometric mean path*: This path is constructed over the distributional space (Gelman and Meng, 1998), hence, we model the density for the model M_t at each point along the grid. We use the density $f_t(\Lambda, \Sigma, \eta|Y) = f(y|\Lambda, \Sigma, \eta)^t \pi(\Lambda, \Sigma, \eta)$ as the unnormalized density for the model M_t connecting the prior and the posterior, when $\pi(\Lambda, \Sigma, \eta)$ and $f(y|\Lambda, \Sigma, \eta)$ are the prior and the likelihood function, respectively. By using PS along this path we can find the log marginal for the models M_0 and M_1, as the normalizing constant for the prior is known. Hence, the $\log BF$ can be estimated by using those estimates of the log marginal for those models. The score function $U(\Lambda, \Sigma, \eta, Y, t)$ will be the log likelihood function $\log f(y|\Lambda, \Sigma, \eta)$.

The theorem below verifies the regularity conditions of path sampling for factor models. For PS to succeed we also need convergence of MCMC at each point in the path. That will be taken up after proving the theorem.

THEOREM 2.1. *Consider path sampling for factor models with parametric arithmetic mean path (PAMP) and likelihood as given above for factor models. Assume prior is proper and the corresponding score function is integrable w.r.t. the prior*:

(1) *The interchangeability of integration and differentiation in* (2.2) *is valid.*

(2) $E_t(U)$ *is finite as* $t \to 0$.

(3) *The path sampling integral for factor models, namely*, (2.3), *is finite.*

PROOF. Here, for notational convenience, we write $(\Lambda, \Sigma, \eta) = \theta$. When $f(Y|\theta)$ and $\pi(\theta)$ are the likelihood function of the data and the prior density function for the corresponding parameter, respectively, then the following is equivalent to showing equation (2.2):

$$\frac{d}{dt}\int_{-\infty}^{\infty} f(Y|\theta, t)\pi(\theta)\,d\theta = \int_{-\infty}^{\infty} \frac{d}{dt} f(Y|\theta, t)\pi(\theta)\,d\theta.$$

We can write the LHS as the following:

$$\begin{aligned}
&= \lim_{\delta\to 0}\int_{-\infty}^{\infty} \frac{f(Y|\theta, t+\delta) - f(Y|\theta, t)}{\delta}\pi(\theta)\,d\theta \\
&= \lim_{\delta\to 0}\int_{-\infty}^{\infty} f'\big(Y|\theta, t'\big)\pi(\theta)\,d\theta\ , t' \in [t, t+\delta] \\
&= \lim_{\delta\to 0}\int_{-\infty}^{\infty} U\big(Y|\theta, t'\big) f\big(Y|\theta, t'\big)\pi(\theta)\,d\theta,
\end{aligned}$$

where $t' \in [t, t+\delta]$. U is a quadratic function in θ and, hence, its absolute value is bounded above by a quadratic function in θ, free of t but depending on Y. $f(Y|\theta, t')$ is bounded by the global maximum of the MVN likelihood, say, M, achieved at $\widehat{\Omega}$ [equation (2.6)]. Now applying the moment assumptions for

$\pi(\theta)$, we can use the Dominated Convergence theorem (DCT) and take the limit within the integral sign. The rest of statements 2 and 3 follow similarly. □

In Remark 1 we extend the theorem to other paths. Then in a series of remarks we study various aspects like convergence and divergence of PS, that are closely related to the theorem. All the remarks are related to the theorem and insights gained from its proof. Remark 5 is the most important.

REMARK 1. For PS with GMP, the score function is the log likelihood function which can be bounded as before by using the RHS of equation (2.5). Also, $f(y|\Lambda, \Sigma, \eta)^t \le (1 \vee f(y|\hat{\Omega}))$ with $\hat{\Omega}$ as in equation (2.5). We believe a similar generalization holds for most paths modeling means of two models. Now the proof of Theorem 2.1 applies exactly as before (i.e., as for PAMP). We exhibit performance of PS for this path in Section 4.

REMARK 2. If we further assume the MCMC average at each point on the grid converges to the Expectation of the score function of MCMC, then the theorem implies the convergence of PS. We showed the integrand is continuous on [0, 1]. So by continuity it can be approximated by a finite sum. Now take the limit of the MCMC average at each of these finitely many grid points. However, even if the MCMC converges in theory, the rate of convergence may be very slow or there may be a problem with mixing even for $m = 50{,}000$, which we have taken as our gold standard for good MCMC. This problem will be apparent to some extent from high MCMC standard deviation.

REMARK 3. As $t \to 0$ the likelihood is practically independent of the extra parameters of the bigger model, so that a prior for those parameters (conditional on other parameters) will not learn much from data. In particular, the posterior for these parameters will remain close to the diffuse prior one normally starts with. If the prior fails to have the required finite moment in the theorem, the posterior will also be likely to have large values for moments, which may cause convergence problems for the MCMC. That's why we chose a prior making the score function integrable. In the proof of the theorem, we have assumed the first two moments of the prior to be finite. In most numerical work our prior is a t with 5 to 10 d.f.

REMARK 4. In the same vein, we suggest that even when the integral at t near zero converges, the convergence may be slow for the following reason. Consider a fixed $(\Lambda_t, \Sigma, \eta)$ with a large posterior or negative value of $U(\Lambda_t, \Sigma, \eta)L(\Lambda_t, \Sigma, \eta)$ at point t, the same large value will occur at $(\frac{1}{t}\Lambda_t, \Sigma, \eta)$ with prior weight $\pi(\frac{1}{t}\Lambda_t, \Sigma, \eta)$. For priors like t-distribution with low degrees of freedom, $\pi(\frac{1}{t}\Lambda_t, \Sigma, \eta)$ will not decay rapidly enough to substantially diminish the contribution of the large value of $U(\Lambda_t, \Sigma, \eta)L(\Lambda_t, \Sigma, \eta)$ at $(\Lambda_t, \Sigma, \eta)$.

REMARK 5. The structure of the likelihood and prior actually provides insight as to when the MCMC will not converge to the right distribution owing to bad mixing. To this end, we sketch a heuristic argument below, which will be supported in Section 3.4 by MCMC figures:

(1) The maximized likelihood remains the same along the whole path, because the path makes a one-to-one transformation of the parameter space as given below.

(2) If the MLE of λ_h at $t = 1$ is $\hat{\lambda}_h$, then the MLE at $t = t'$ is $\frac{\hat{\lambda}_h}{t'}$ (subject to variation due to MCMC at two different points at the path), which goes to infinity as t goes to zero. This happens as the $\hat{\lambda}_h$ remains the vector among λ_h' (where λ_h' is the MCMC sample from model M_t at t) having the highest maximum likelihood. Hence, as $t \to 0$, $\pi(\hat{\lambda}_h/t) \to 0$ at a rate determined by the tail of the prior. The conflict between prior and maximized likelihood may also be viewed as a conflict between the nested models, with the prior favoring the parsimonious smaller model. This inherent conflict in model selection seems to have the following implications for MCMC.

We expect to see a range (say, $[t_1, t_2]$) near zero showing a conflict between prior and maximized likelihood. Definitely the points t_1 and t_2 are not well specified, but we treat them as such so as to understand some underlying issues of mixing and convergence here. On the set of points $t > t_2$ the MCMC samples are expected to be around the points maximizing likelihood, whereas for $t < t_1$ they will be nearly zero due to the concentration around a value λ_h which is both the prior mode and the mle under M_0, namely, $\lambda_h = 0$. But for any point in the range $[t_1, t_2]$, they will span a huge part of the parameter space, ranging from points maximizing likelihood to ones having higher prior probability, showing a lot of fluctuations from MCMC to MCMC. The MCMC outputs in Section 3.4 show both clusters but having highly fluctuating values (Figure 1, Section 3.4) for the proportions of the clusters.

Equation (2.7) tells us that the score function is proportional to $\frac{\lambda_h'}{t}$ (where λ_h' is the MCMC sample from

model M_t at t). Hence, we will see $E_t(U)$ as an increasing function while $t \to t_2$ from the right-hand side [(2) in Remark 5]. This leads to a lot of variation of the estimate of $E_t(U)$ for different MCMC samples in the range $[t_1, t_2]$ as explained above. Also, as explained above, for $t < t_1$, the score function will concentrate near zero.

The width of the zone of conflict (here $t_2 - t_1$) will shrink, if we have a relatively strong decaying tail of the prior. On the other hand, for heavy-tailed priors we may see these above mentioned fluctuations for a longer range, causing a loss of mass from the final integration. These problems are aggravated by the high dimension of the problem and the diffuse spread of the prior on the high-dimensional space. This may mean the usual BF estimated by PS will be off by an order of magnitude. We will see the implications reflected in some figures and tables in the next section, when we study PAMP-PS for factor models in detail in Section 3.

REMARK 6. We have checked that adaptive choice of grid points by Lefebvre et al. (2009), which improves accuracy in their two examples with GMP, does not help in the case of the very large fluctuations described above. It seems to us that adaptive choice would work better when the two models tested are less dissimilar than the models in Remark 5, for example, when the smaller of two nested models is true (Section 3.1) or when our proposed modification of PS is used (Section 3.3). However, we have not verified this because even without adaptive choice, our new proposal worked quite well in our examples.

We note in passing that in both the examples of Lefebvre et al. (2009), the two models being tested have maximum likelihoods that differ by fifteen in the log scale, whereas for the models in Remark 5 they differ by much more, over a hundred.

3. WHAT DO ACTUAL COMPUTATIONS TELL US?

Following the discussion in the previous section, we would like to study the effects of the theoretical observations in the previous section for the implementation of path sampling. Here we only consider the PAMP for PS, and for notational convenience we will mention it as just PS. After studying estimated BF's in several simulated data sets (not reported here) from various factor models, we note a few salient features. Error in estimation of the BF or the discrepancy between different methods tends to be relatively large, if one of the following is true: the data has come from the complex model rather than the simpler model, the prior is relatively diffuse or the value of the precision parameters are relatively small. Different subsections study what happens if the complex or simpler model is true, the effect of the prior, the grid size and the MCMC size. These are done in Sections 3.1–3.3.

TABLE 2
Loading factors used for simulation

Factor 1	0.89	0	0.25	0	0.8	0	0.5
Factor 2	0	0.9	0.25	0.4	0	0.5	0

In Section 3.3 we introduce a new PS scheme, which operates through a chain of paths, each path involving two nested models with a small change between the contiguous pairs. The new scheme is denoted as Path Sampling with Small Changes (PS-SC). The effect of precision parameters will also be studied in this subsection for PS-SC. Then we study the MCMC samples and try to understand their behavior from the point of view of explaining the discrepancy between different methods for estimating Bayes factors and why PS-SC does better than PS in Section 3.4.

Our simulated data are similar to those of Ghosh and Dunson (2009) but have different parameters. *We use a* 2*-factor model and a* 1*-factor model as our complex model* M_1 *and simpler model* M_0*, respectively, to demonstrate the underlying issues.* The loading parameters and the diagonal entries of the Σ matrix are given in Tables 2 and 3. In simulation we take model M_0 or M_1 as true but Σ is not changed. Of course, if the one-factor model M_0 is true, then since it is nested in M_1, M_1 is also true.

3.1 Issues in Complex (2-Factor) Model

We will study the effect of grid size, prior and the behavior of MCMC, keeping in mind Theorem 2.1 and the remarks in Section 2. For path sampling with the PAM path, we now discuss the effect of the prior and the two tuning parameters, namely, the effect of the grid size and MCMC size, on the estimated value of the BF and their standard deviation. Following the discussion in Remarks 3 and 4, we know that $\lim_{t\to 0} E_t(U)$ is finite and path sampling converges under some finite

TABLE 3
Diagonal entries of Σ

0.2079	0.19	0.15	0.2	0.36	0.1875	0.1875

TABLE 4
PAM-PS: Dependance of $\log BF_{21}$ *over prior, 2-factor model true*

PS using grid size 0.01				
t_1	t_5	t_{10}	t_{90}	normal
2.62	14.42	22.45	70.20	70.25
3.67	11.90	21.39	68.70	68.72
3.00	13.43	21.31	47.06	47.21
4.29	13.17	18.49	48.03	48.13
4.20	13.11	18.48	47.70	47.74

moment assumption for the prior. The prior considered in PS by Ghosh and Dunson (2008) are Cauchy and half-Cauchy, which do not have any finite moments and so U is not integrable. We therefore choose a relatively diffuse prior, but with enough finite moments for U. For finite mean and variance one needs a t with at least four degrees of freedom. Our favorites are t-distributions with 5 to 10 degrees of freedom. We show results for 5 and 10 d.f. only. But we first explore the sensitivity of the estimate to changes in d.f. of the t-distribution as prior, over a range of 1 through 90. The BF values change considerably until we reach about 40 d.f. and then it stabilizes. In Table 4 we report the $\log BF$ values estimated for 5 data sets simulated from a 2-factor model using different priors. The behavior of the estimated $\log BF$ with the change of d.f. continuously from 1 to 100 is shown in Figure 1 for the 3rd data set.

We can see the estimate of the BF changing with the change in the pattern of the tail of the prior. The effect of the grid size and MCMC size on MCMC-

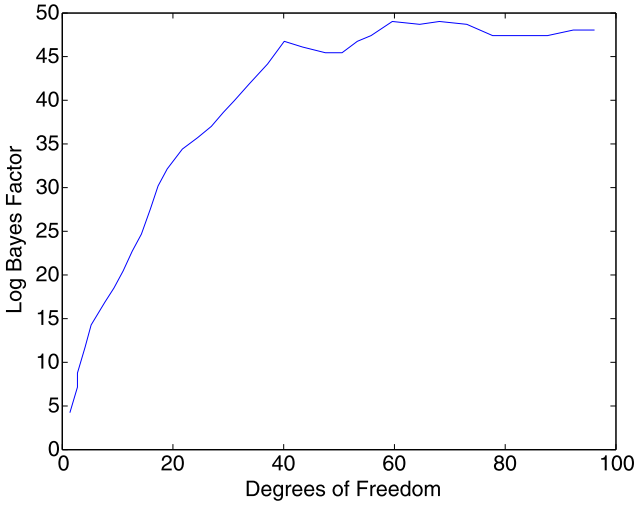

FIG. 1. *Dependance of* $\log BF_{21}$ *over prior for* 3*rd data set.*

TABLE 5
PAM-PS: Dependence of $\log BF_{21}$ *(MCMC-standard deviation) estimates over grid size and MCMC size, 2-factor model true*

Grid size		0.01	0.001
MCMC size	Prior	Data 2	Data 2
5000	t_{10}	21.26 (1.39)	21.26 (1.29)
	$N(0,1)$	66.89 (4.15)	67.21 (3.28)
50,000	t_{10}	23.71 (1.21)	23.57 (0.52)
	$N(0,1)$	68.21 (3.62)	68.23 (3.11)

standard deviation of the estimate are studied, using priors t_{10} and $N(0,1)$ and reported in Table 5. We report the mean of the estimates found from 25 different MCMC runs and the corresponding standard deviation as MCMC-standard deviation. The study has been done on the 2nd of the 5 data sets simulated from model 1 earlier.

As expected, Table 5 shows a major increase of MCMC size and finer grid size reduces the MCMC-standard deviation of the estimator. The difference between the mean values of BF estimated by t_{10} and $N(0,1)$ differ by an order of magnitude. We will study these issues as well as special patterns exhibiting MCMC in Section 3.4. Though the different variants of PS compared here differ in their estimated value of BF, they still choose the correct model 100% of the time.

3.2 Issues in Simpler (1-Factor) Model

Now we study the scenario when the 1-factor model is true focusing on the effect of prior, grid size and MCMC size on the estimated Bayes factor (Table 6). In this scenario the estimates do not change much with the change of prior, so we will report the estimates for prior t_{10} and $N(0,1)$ with different values of MCMC size and grid size.

This table shows us that the MCMC-standard deviation improves with the finer grid size and large MCMC

TABLE 6
PAM-PS: Dependence of $\log BF_{21}$ *(MCMC-standard deviation) estimates over grid size and MCMC size, while* 1*-factor model true*

Grid size		0.01	0.001
MCMC size	Prior	Data 1	Data 1
5000	t_{10}	−4.26 (0.054)	−4.27 (0.044)
	$N(0,1)$	−4.62 (0.052)	−4.60 (0.051)
50,000	t_{10}	−4.24 (0.012)	−4.24 (0.007)
	$N(0,1)$	−4.60 (0.006)	−4.62 (0.005)

size as expected, but the estimated values of BF_{21} remain mostly the same. As noted earlier, PS chooses the correct model 100% of the time when M_0 is true.

We explain tentatively why the calculation of BF is relatively stable when the lower dim model M_0 is true. Since M_0 is nested in M_1, M_1 is also true in this case, which in turn implies both max likelihoods (under M_0 and M_1) are similar and smaller than for data coming from M_1 true (but not M_0). This tends to reduce or at least is associated with the reduction of the conflict between the two models or prior and likelihood along the path mentioned in Remark 5.

Moreover, the score function for small t causes less problem since for data under M_0, λ_2' is relatively small compared with that for data generated under M_1.

So we see when two models are close in some sense, we expect their likelihood ratio will not fluctuate widely provided the parameters from the two parameter spaces are properly aligned, for example, if found by minimizing a K-L divergence between the corresponding densities or taking a simple projection from the bigger space to the smaller space. This is likely to make importance sampling more stable than if the two models were very different. It seems plausible that this stability or its lack in the calculation of BF will also show up in methods like PS that are derived from importance sampling in some way. Ingenious modifications of importance sampling seems to mitigate but not completely solve the problem. Following this idea of closer models in some sense, we modify PS in a similar manner below.

3.3 Path Sampling with Small Changes: Proposed Solution

In Remark 5, Section 3.1, a prior-likelihood conflict was identified as a cause of poor mixing. This will be re-examined in the next subsection. In the present subsection we propose a modification of PS which tries to solve or at least reduce the magnitude of this problem.

To solve this problem without having to give up our diffuse prior (we will be using t with 10 d.f. as our prior), we try to reduce the problem to a series of one-dimensional problems so that the competing models are close to each other. We calculate the Bayes factor by using the path sampling step for every single parameter that may be zero, keeping others fixed. It is easily seen that the original log Bayes factor is the sum of all the log Bayes factors estimated in these smaller steps. We denote this procedure as PS-SC (Path Sampling with Small Change) and implement with the parametric arithmetic mean path (PAMP). (As pointed out by a Referee, there is scope for exploring other paths, including a search for an optimal one, to reduce the MCMC-variance.) More formally, if we consider λ_2 as a p-dimensional vector, then M_0 and M_1 differ only in the last $p-1$ parameters, as λ_{21} is always zero due to the upper-triangular condition. We consider p models $M_i' : i = 1, \ldots, p$, where for model M_i' we have first i parameters of λ_2 being zero correspondingly. If we define $BF'_{i,i+1} = \frac{m_i(x)}{m_{i+1}(x)}$, when $m_i(x)$ is the marginal for the model M_i', then

$$\log BF_{21} = \sum_{i=1}^{p-1} \log BF'_{i,i+1}.$$

So we perform $p-1$ path sampling computations to estimate $\log BF'_{i,i+1}, \forall i = 1, \ldots, p-1$. And for each of the steps the score function will be of the following form:

$$\begin{aligned} &U_i'(\Lambda, \Sigma, \eta, Y, t) \\ &\quad = \sum_{j=1}^{n} (y_j - \Lambda_t \eta_i)' \\ &\qquad \cdot \Sigma^{-1}\big(0^{p\times(h-1)}, [0_i; \lambda_{h,i+1}; 0_{p-i-1}]\big)\eta_i, \end{aligned}$$

where $\Lambda_t = (\lambda_1, [0_i; t\lambda_{2,i+1}; \lambda_{2,(i+2,\ldots,p)}])$.

As in the case of the small model true, the max likelihoods under both models are close, and generally the two models are close, suggesting fluctuations are less likely and true BF is not very large. This seems generally to lead to stability of computation of BF.

Also, the parameter λ_2' is now one dimensional. So the score function is more likely to be small than when λ_2' is a vector as under PS. We also notice that in each step the score function is not anymore proportional to $\frac{\lambda_2'}{t}$ but rather to $\frac{\lambda_{2i}'}{t}$ which will be much smaller in value, hence reducing the fluctuation and loss of mass.

Computational implementation shows it to be stable for different MCMC size and grid size regarding MCMC-standard deviation and also produces a smooth curve of $E_t(U)$ for every single step. Here we use an MCMC size of 5000/50,000 and grid size of 0.01 for our study and report the corresponding estimated BF values for two data sets from 1-factor and 2-factor models, respectively. The MCMC-standard deviation of the estimates along with the mean of the estimated value over 25 MCMC runs are reported in Table 7. PS-SC has smaller standard deviation than PS under both M_0 and M_1. In Section 2 and Section 3.4, we argue that, at least under M_1, PS-SC provides a better estimate of BF.

TABLE 7
$\log BF_{21}$ (*MCMC-standard deviation*) *estimated by PAM-PS-SC and PAM-PS*

True model	MCMC size	PS-SC	PS (t_{10})	PS ($N(0, 1)$)
1-factor	5000	−8.09 (0.013)	−4.26 (0.054)	−4.62 (0.052)
1-factor	50,000	−8.08 (0.0067)	−4.24 (0.012)	−4.60 (0.0065)
2-factor	5000	80.14 (0.66)	21.26 (1.39)	66.89 (4.15)
2-factor	50,000	80.75 (0.54)	23.71 (1.21)	68.21 (3.62)

Now we see the effect of changing the precision parameters keeping the factor loadings as before. The diagonal entries of Σ are in Table 8. The precision of these 3 models lie in the ranges of [2.77, 6.55], [1.79, 2.44], [1.36, 1.66], respectively.

We study PS-SC for 6 data sets generated from the 3 models (2 data sets with $n = 100$ from each model: Data 1 from 1-factor and Data 2 from 2-factor model) and report the estimated Bayes factor value in Table 9.

The effect of precision parameters is seen on the estimated value of the Bayes Factor (BF), more prominently when the 2-factor model is true. Generally, the absolute value of the BF decreases with the decrease in the value of the precision parameters. For the smaller value of precision parameters, we expect the model selection to be less conclusive, explaining the pattern shown in the estimated BF values.

Under M_1, PS is often bad in estimating the Bayes Factor (BF_{21}), but since the true Bayes factor is large, it usually chooses the true model as often as PS-SC. When M_0 is true, PS is much better in estimating the Bayes factor, but since the Bayes factor is usually not that large, it does not choose M_0 all the time. The probability of choosing M_0 correctly depends on the data in addition to the true values of the parameters. PS-SC does better than PS in all these cases; it estimates BF_{21} better and chooses the correct model equally or more often. The sense in which PS-SC estimates BF_{21} better has been discussed in detail earlier in this section. Under M_0 PS-SC estimates BF_{21} better by having a smaller, that is, more negative, value than PS.

TABLE 8
Diagonal entries of Σ *in the* 3 *different models*: *the first one is modified from Ghosh and Dunson* (2008)

Model 1	0.2079	0.19	0.15	0.2	0.36	0.1875	0.1875
Model 2	0.553	0.52	0.48	0.54	0.409	0.55	0.54
Model 3	0.73	0.71	0.67	0.7	0.599	0.67	0.72

3.4 Issues Regarding MCMC Sampling

This subsection is best read along with the remarks in Section 2. We first study the graph of $E_t(U)$ and the likelihood values for the MCMC samples at t for both the t_{10} and $N(0, 1)$ prior (Figures 2 and 3). We will plot the likelihood as a scalar proxy since we can not show fluctuations of the vector of factor loadings in the MCMC output. The clusters of the latter can be inferred from the clusters of the former. *We will argue that there are two clusters at each grid point and the mixing proportion of the two clusters has a definite pattern.*

Under the true 2-factor model M_1, denote $\lambda' = [\lambda_1', \lambda_2']$, where λ_i' is the loading for the corresponding latent factor under M_t. Here λ_2' is a 7×1 vector and becomes zero, as it approaches M_0 from M_1 (as $t \to 0$). The posterior distribution at each M_t can be viewed roughly as a sort of mixture model with two components representing M_0 and M_1, the form of the likelihood as given in Theorem 2.1. In the diagram (Figure 4) of the log-likelihood of MCMC samples, we see two clear clusters around log-likelihood values −850 and −925, representing MCMC outputs with nonzero λ_2' and zero λ_2' values, respectively. We may think of them as coming from the component corresponding to M_1 (cluster 2) and the component corresponding to M_0 (cluster 1). Samples of both clusters are present in the range [0.03, 0.2], while samples appear to be predominantly from cluster 2 until $t = 0.1$. A good representation of samples from cluster 1 are only present in the range [0, 0.1]. In the range [0.03, 0.2], both clusters occur with proportions varying a lot. Moreover, here the magnitude of the score function is proportional to $\frac{\lambda_2'}{t}$. We see these fluctuations in Figure 4 in the region [0.03, 0.2]. This is also brought out by the MCMC standard deviation of $E_t(U)$ *which are of order of* 30–50 *in the log scale*.

We notice the absence of any samples from M_1 for $t < 0.03$, except some chaotic representation for a few

TABLE 9
$\log BF_{21}$ (*MCMC-standard deviation*) *estimation by PS-SC*: *effect of precision parameter*

Model	True model	Data	PS-SC	PS (t_{10})
Model 1	1-factor	Data 1	−8.09 (0.012)	−3.84 (0.055)
	2-factor	Data 2	71.59 (0.66)	19.81 (1.38)
Model 2	1-factor	Data 1	−11.01 (0.0066)	−3.09 (0.0277)
	2-factor	Data 2	51.41 (0.3658)	2.8 (1.9104)
Model 3	1-factor	Data 1	−5.13 (0.0153)	−2.6 (0.0419)
	2-factor	Data 2	3.975 (0.0130)	2.2 (0.3588)

random values of t (notice in the figure, a spike representing samples from M_1 at $t = 0.016$), clearly representing poor mixing of MCMC samples near the model M_0.

The new method PS-SC stabilizes the estimated Bayes factor value with a very small MCMC-standard deviation. Here we check through Figures 5 and 6 that it avoids prior-likelihood conflict and the problem about mixing for MCMC samples seen for the standard PS. We concentrate our study for the first step of PS-SC. In this step only the first component of λ_2', λ_{22}' converges to zero as $t \to 0$. So here we consider the spread of the MCMC sample of λ_{22}' for different values of t near $t = 0$, from both PS and PS-SC in Figures 5 and 6 by considering the histogram of MCMC sample of λ_{22}'. We can easily notice that the spread of the MCMC sample fluctuates in between the two modes in a chaotic manner showing poor or unstable mixing for PS, whereas PS-SC samples come from both the clusters and slowly shift toward the prior mode as $t \to 0$. We have also studied but do not report similar nice behavior regarding mixing of MCMC of PS-SC for the data simulated from the 1-factor model.

The poor mixing discussed above for MCMC outputs for PS will now be illustrated with plots of autocorrelation for λ_{22}' for different lags (Figure 7). For the sake of comparison, we do the same for PS-SC (Figure 8). Clearly, except very near $t = 0$, that is, in what we have called the chaotic zone, the autocorrelations for PS are much bigger than those for PS-SC. However, near $t = 0$, though plots in both Figures 7 and 8 are small, those for PS are slightly smaller. We have no simple explanation for this.

Poor mixing seems to lead to missing mass and random fluctuations for calculations for $E_t(U)$. This probably explains the discrepancy we have noticed in the estimation of BF by PS as compared with PS-SC. We now look at autocorrelations for a first factor loading

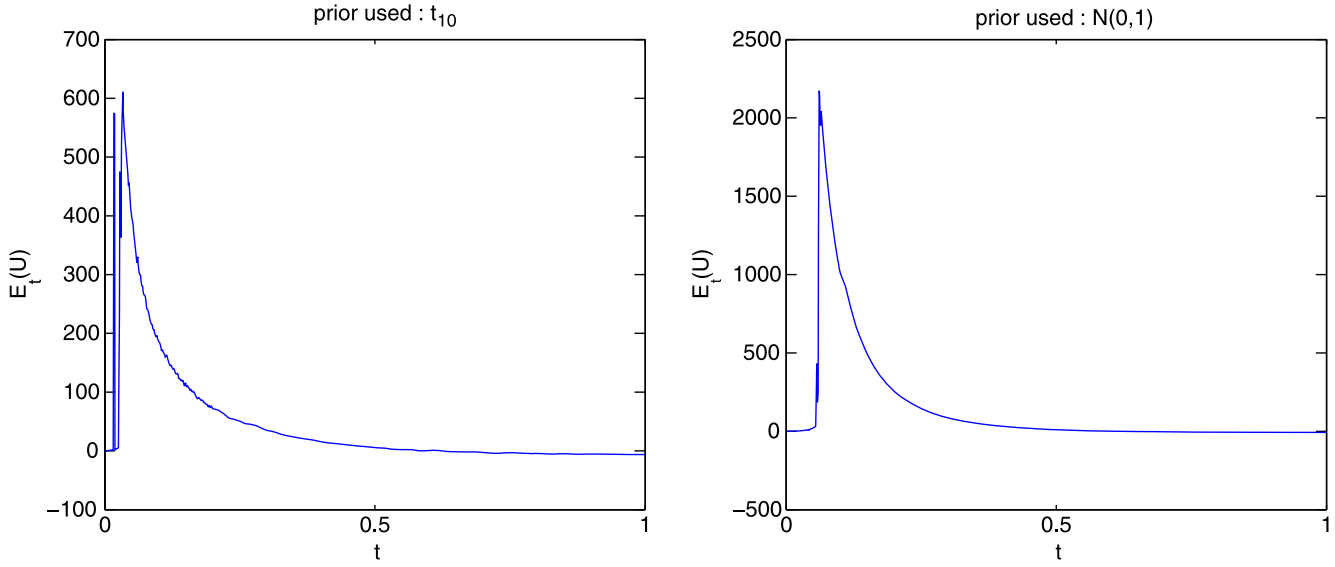

FIG. 2. $E_t(U)$ *for prior* t_{10} *and* $N(0, 1)$, 2*-factor model is true.*

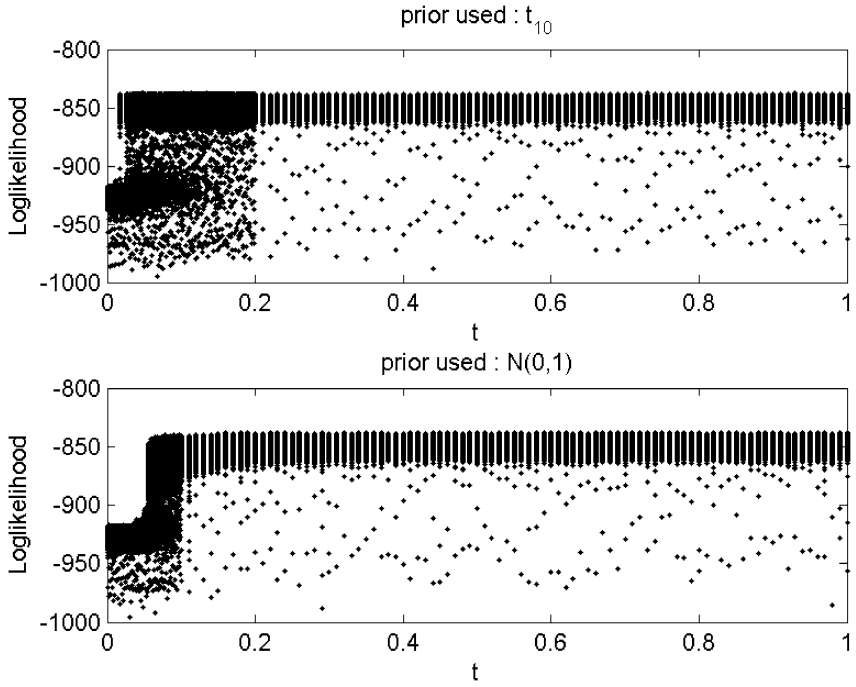

FIG. 3. *Log-likelihood for prior t_{10} and $N(0, 1)$, 2-factor model is true.*

in Figure 7 and second factor loading in Figure 8. The top rows in each of the two figures show zero autocorrelation, as they are very close to $t = 0$. On the other hand, high autocorrelations are shown in the next two rows. We believe they correspond to what we called a chaotic region. The bottom two rows of Figure 8 show small autocorrelation. They correspond to the second factor loading which comes only in model 2, and they also depict the zone dominated by model 2. The other figure is in the same zone as in the previous line, but the variable considered is a 1-factor loading. Here autocorrelation also eventually tends to 0, but its values

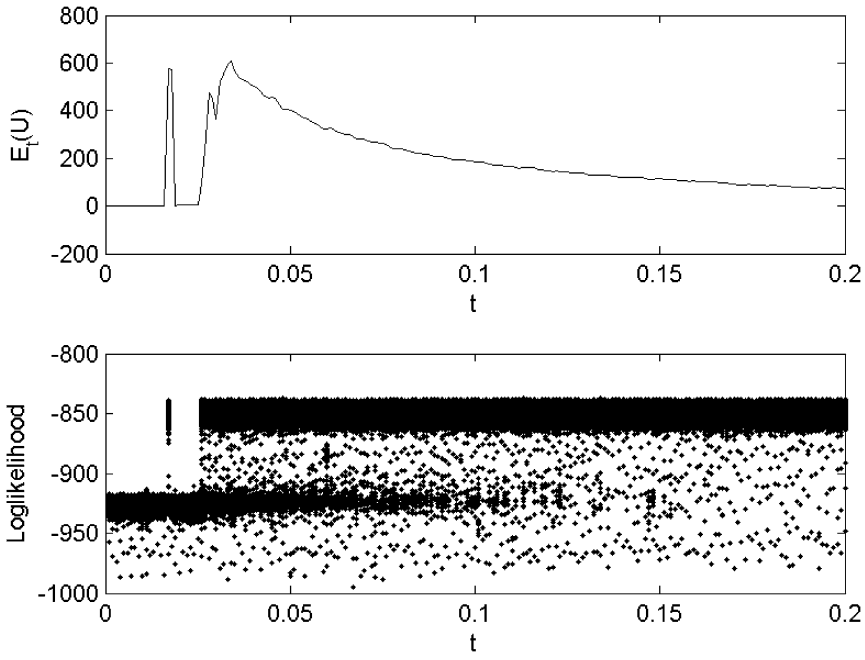

FIG. 4. *$E_t(U)$ and Log-likelihood for prior t_{10} in the range $t \in [0, 0.2]$, 2-factor model is true.*

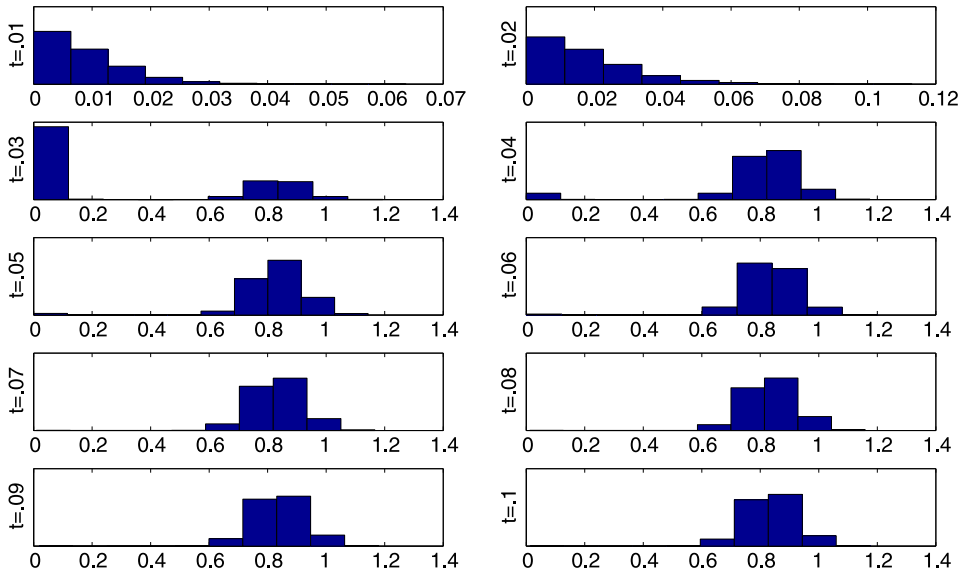

FIG. 5. *Histograms for λ'_{22} for different values of t near $t = 0$ (MCMC size used* 50,000*), using PS.*

are bigger than in Figure 8. We do not have any simple explanation for this higher autocorrelation.

The above discussion covers the case when the more complex model is true. If the simpler model (M_0) is true, as noted in Section 3.3 both PS and PS-SC perform well in estimating the Bayes factor as well as choosing the correct model. The Bayes factor based on PS-SC provides stronger support for the true model than the Bayes factor based on PS.

To check whether PSSC works well in other examples as in the factor model, we try to explore its impact on our earlier toy example. In this case, we were unable to implement path sampling with small changes, but rather used a pseudo-PSSC scheme. Going back to our example where we have taken $m = 7$ and $p = 10$, we define a sequence of models as the following:

$$M_i : y_i \sim N\left(0, \Sigma = \begin{pmatrix} A_{11} & 0 \\ 0 & A_{22} \end{pmatrix}\right)$$

when A_{11} is $(i \times i)$ matrix for $i = 7, 8, 9, 10$.

We can see our previously defined M_0 and M_1 are now M_7 and M_{10}, respectively. For our pseudo-PSSC, we

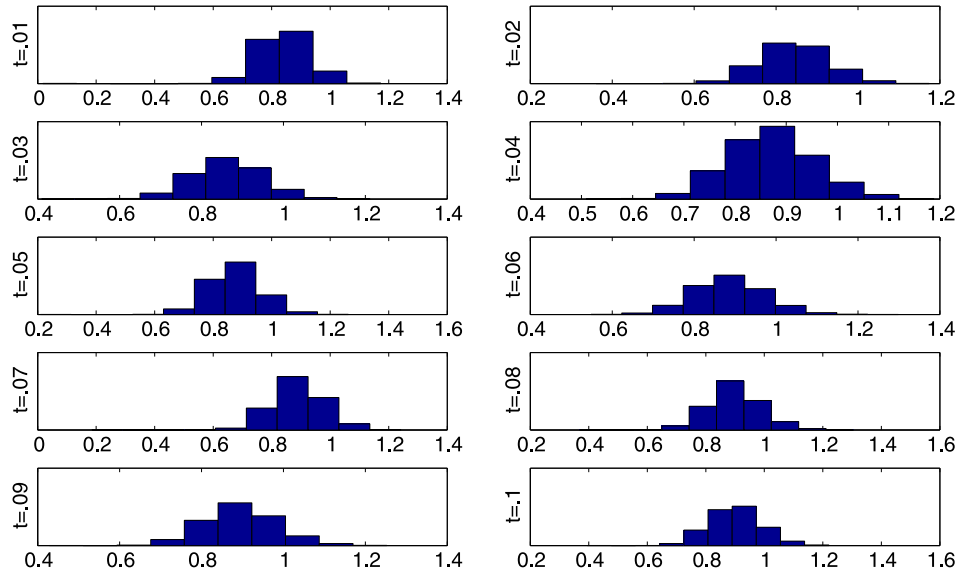

FIG. 6. *Histograms for λ'_{22} for different values of t near $t = 0$ (MCMC size used* 50,000*), using PS-SC.*

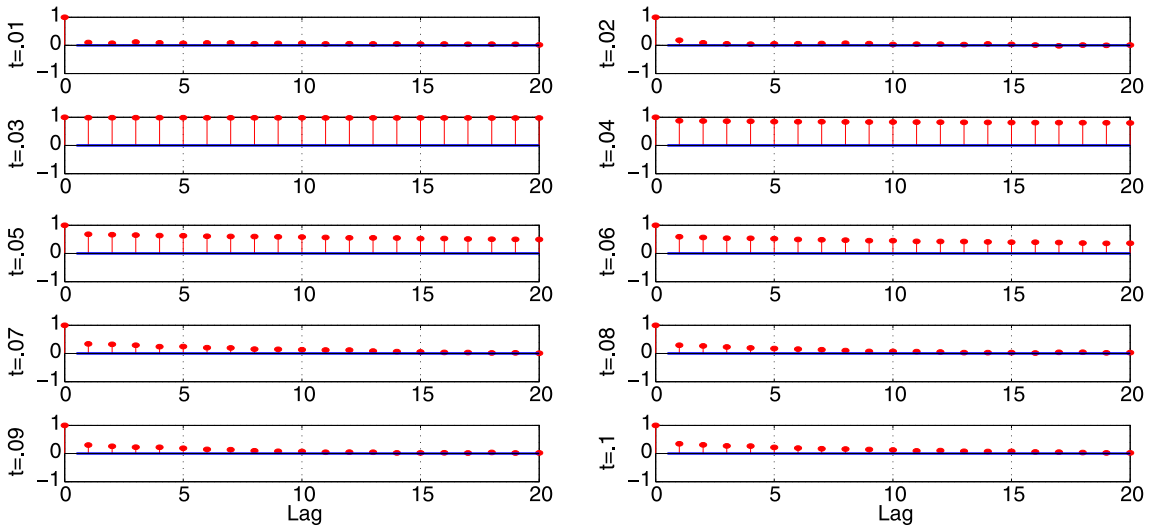

FIG. 7. *Autocorrelation for* λ'_{22} *for different values of* t *near* $t = 0$ (*MCMC size used* 50,000), *using PS*.

estimate $\log BF_{i,i+1}$ by $\log BF$ between the models M'_0 and M'_1, with $m = i$ and $p = i + 1$:

$$M'_t : y_i \sim N\left(0, \Sigma = \begin{pmatrix} A_{11} & tA_{12} \\ t(A_{12})' & A_{22} \end{pmatrix}\right).$$

Still being underestimates on each step, this method improves on the standard path sampling in terms of Bayes factor estimation, as we can see in the Table 10.

4. IMPLEMENTATION OF OTHER METHODS

We have explored several methods of estimating the ratio of normalizing constants, for example, the methods of Nielsen (2004), DiCiccio et al. (1997), Rue, Martino and Chopin (2009) and Chib (1995). The method of Rue, Martino and Chopin (2009) models a link function of means, but here we are concerned with models for the variance–covariance matrix. We could not use Chib's method here since for our parameter expanded prior the full conditionals of the original model parameters are not available. But we were able to implement the deterministic variational Bayes method of Nielsen (2004) and the Laplace approximation with a correction due to DiCiccio et al. (1997). Since the results were not satisfactory, we do not report them in this paper. In the variational Bayes approach, the method selected the correct model approx-

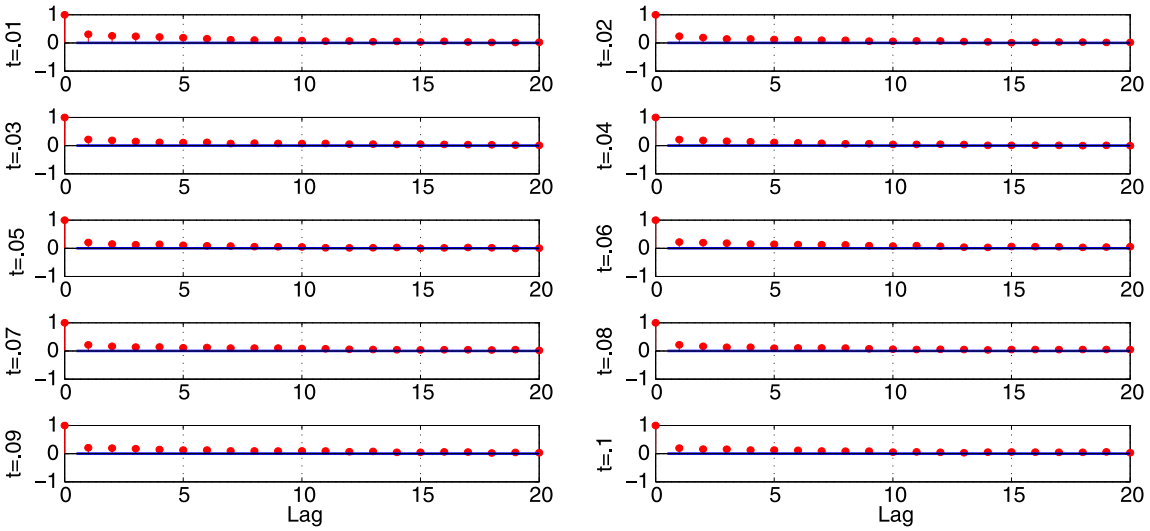

FIG. 8. *Autocorrelation for* λ'_{22} *for different values of* t *near* $t = 0$ (*MCMC size used* 50,000), *using PS-SC*.

TABLE 10
Performance of PS and pseudo-PS-SC in toy example modeling covariance: Log Bayes factor (MCMC standard deviation)

Method	Data 1	Data 2
True BF value	258.38	−132.87
PS estimate of BF	184.59 (0.012)	−20.11 (0.008)
pseudo-PSSC estimate of BF	195.35 (0.011)	−25.21 (0.007)

imately 80% of the time, but the estimated logBF values were considerably over (or under) estimated. The variational Bayes method is worth further study, possibly with suitable modifications. It appears to us it is still not understood when Belief Propagation provides a good approximation to a marginal or not, for example, Gamarnik, Shah and Wei (2010) commented: *Only recently we have witnessed an explosion of research for theoretical understanding of the performance of the BP algorithm in the context of various combinatorial optimization problems, both tractable and intractable (NP-hard) versions*.

Following the discussion in Section 2.4, we have implemented the GMP-PS. Here the marginal for both models is estimated by constructing a path between the prior distribution to the posterior distribution of the model. Due to very high-dimensionality of the model, the modes of prior and posterior distribution are far apart. So as discussed before, the MCMC sampling along the path fails to sample smoothly and fluctuates between the two modes in a chaotic way near the prior mode. Hence, the estimate of the marginal of both the models becomes very unstable. Due to the poor estimation of BF, this method also fails to choose the correct model very often. As in the case of GMP-PS, the AIS with the GM path also did not work well. Hence, we implemented the AIS with the PAM-path. Implementation of PAM-AIS is also very time intensive, so we have only implemented PAMP-AIS with MCMC sample size 5000. PAM-AIS not only shows very high MCMC-standard deviation, but it also fails to choose the correct model many a time, when the 2-factor model is correct. The last methods we looked at are the following:

(1) Importance Sampling (IS).
(2) Newton-Raftery approximation (BICM).
(3) Laplace/BIC type approximation (BICIM).

IS is the most easy to implement and shows moderately good results in choosing the correct model (Ghosh and Dunson, 2008)). We study the stability of Bayes factor values estimated by IS with the change of the MCMC size in Table 11.

TABLE 11
Study of IS, BICM and BICIM for different MCMC size: Estimated Bayes factor (MCMC standard deviation)

Method(MCMC-size)/ true model	2-factor model	1-factor model
IS (10,000)	109.78 (168.72)	0.0749 (0.1063)
IS (50,000)	97.12 (61.25)	−5.39 (84.35)
IS (100,000)	86.92 (110.35)	−3.07 (10.41)
IS (200,000)	83.66 (58.53)	−2.69 (2.96)
BICM (10,000)	68.66 (0.93)	−5.72 (0.62)
BICIM (10,000)	67.9 (0.11)	−5.3 (0.57)
PS-SC (5000)	80.75 (0.63)	−8.08 (0.0013)

Similarly, we also study the stability of the estimates of the Bayes factor by BICM and BICIM (explained in A.1.3 in the Appendix) using MCMC sample size 10,000, where both of these methods show significantly less amount of MCMC-standard deviation than other methods considered. Hence, we will only consider PS-SC, BICM and BICIM to explore model selection for a dimension much higher than previously considered.

5. EFFECT OF PRECISION PARAMETERS AND HIGH-DIMENSIONAL (SIMULATED AND REAL) DATA SET

Our goal is to explore if PS-SC may be made more efficient by combining with BICM and BICIM and also to explore the number of dimensions much higher than before and the real life examples.

In the examples in this section, p varies from 6 to 26. We have 2 examples of real life examples with $p = 6$ and 26 and a simulated example with $p = 20$. As expected, PS-SC still takes a long time, even with a parallel processing for high-dimensional examples. We explore whether PS-SC can be combined with BICM and BICIM to substantially reduce time, since their performance seems much faster than PS-SC.

We compare the behavior of these methods for a higher-dimensional model and for some real data sets taken from Ghosh and Dunson (2009) and Akaike (1987). We first consider one 3-factor model with $p = 20$ and $n = 100$ in Table 12.

We notice that all the methods are selecting correct models for all the 3 data sets, but based on our earlier discussion of PS-SC, we believe only this method provides a reliable estimate of BF. Now we will compare

TABLE 12
Simulated model ($p = 20$, $n = 100$) and ($k =$ the number of true factors): Comparison of log Bayes factor

Data	BF	PS-SC	BICM	BICIM
Data1 ($k = 1$)	BF_{21}	−25.91 (0.0233)	−32.68	−38.01
	BF_{32}	−24.84 (0.0594)	−21.18	−38.24
	BF_{43}	−22.79 (0.0483)	−19.81	−43.77
Data2 ($k = 2$)	BF_{21}	225.81 (4.2099)	248.09	219.87
	BF_{32}	−23.61 (0.0160)	−23.59	−46.17
	BF_{43}	−19.18 (0.0297)	−20.3	−47.98
Data3 ($k = 3$)	BF_{21}	152.07 (1.7422)	185.45	162.3
	BF_{32}	104.17 (2.5468)	198.1	168.54
	BF_{43}	−17.35 (0.0276)	−29.73	−48.24

the methods for some real data sets. We choose two data sets: "Rodent Organ Data" from Ghosh and Dunson (2009) and "26-variable Psychological Data" from Akaike (1987). These data sets have been normalized first before analyzing them further. We not only study the estimated Bayes factor but also the model chosen by them.

In the "Rodant Organ Data" the model chosen by PS-SC and other methods are, respectively, the 3-factor model and 2-factor model (Table 13). For the "26-variable Psychological Data," PS-SC and BICM/BICIM choose the model with 3 factors and 4 factors, respectively (Table 14). The models chosen by PS-SC and the other methods are close, but as expected differ a lot in their estimate of BFs.

There is still no rigorously proved Laplace approximation for relatively high-dimensional cases because of analytical difficulties. Problems of determining sample size in hierarchical modeling, pointed out by Clyde and George (2004), are avoided by both versions of our approximations (Appendix A.1.3). These two methods seem to be good as a preliminary searching method to narrow the field of plausible models before using PS-SC. This saves time relative to PS-SC for model search as seen in the previous examples.

TABLE 13
Rodant organ weight data ($p = 6$, $n = 60$): Comparison of log Bayes factor

Bayes factor	PS-SC	BICM	BICIM
$\log BF_{21}$	4.8	26.34	21.57
$\log BF_{32}$	10.52	−3.14	−10.01
$\log BF_{43}$	−3.28		

6. CONCLUSION

We have studied PS for factor models (and one other toy example) and have identified the component of PS that is most likely to go wrong and where. This is partly based on the fact that we have a relatively simple sufficient condition for factor models (Theorem 2.1). Typically, for the higher-dimensional model the MCMC output for finding the integral along grid points in the path may become quite unreliable at some parts of the path. Some insight about why it happens and how it can be rectified has been suggested. MCMC seems to be unreliable for PS when the higher-dimensional model is true. The problem is worse the more the two models differ, as when a very high-dimensional model is being compared to a low-dimensional model.

The suggestion for rectification was based on the intuition that PS, like importance sampling itself, seems more reliable when the two marginal densities in the Bayes factor are relatively similar, as is the case when the smaller of two nested models is true. Based on this intuition, we suggested PS-SC and justified PS-SC by comparing MCMC output and MCMC standard deviation of both PS-SC and PS.

TABLE 14
26-variable psychological data ($p = 26$, $n = 300$): Comparison of log Bayes factor

Bayes factor	PS-SC	BICM	BICIM
$\log BF_{21}$	122.82	205.27	188.19
$\log BF_{32}$	35.27	71.05	35.5
$\log BF_{43}$	−10.7	23.16	7.55
$\log BF_{54}$	−33.32	−4.63	−25.51
$\log BF_{65}$	−16.7	−17.32	−43.21

It is our belief that the above insights as to when things will tend to go wrong and when not, will also be valid for the other general strategy for selection from among nested models, namely, RJMCMC. Piyas Chakraborty in Purdue is working on a change point problem in several parameters where Shen and Ghosh (2011) have an accurate approximation to the Bayes factor, which may be used for validation. He will explore small changes as well as adaptive MCMC.

Our work has focused on model selection by Bayes factors, which seems very natural since it provides posterior probability for each model. However, model selection is a complicated business and one of its major purposes is also to find a model that fits the data well. Several model selecting statisticians feel this should also be done along with calculation of Bayes factors.

However, there has not been a good discussion on how one should put together the findings from the two different approaches. We hope to return to these issues in a future communication.

A natural future direction of our study of factor models is to add to the model an unknown mean vector with a regression setup. The problem now would be to simultaneously determine a parsimonious model for both the variance–covariance matrix and the mean vector. There are natural priors for these problems, but computation of the Bayes factor seems to be a challenging problem.

APPENDIX

A.1 Other Methods

A.1.1 *Importance sampling*. Suppose we have two densities proportional to two functions $f(x)$ and $g(x)$, which are feasible to evaluate at every x, but one of the distributions, say, the one induced by $f(x)$, is not easy to sample. Then the importance sampling (IS) estimate of the ratio of normalizing constants is based on m independent draws $x_1, \ldots, x_m$ generated from the distribution defined by $g(x)$. We first compute the importance weights $w_i = \frac{f(x_i)}{g(x_i)}$ and then define the IS estimate:

$$\frac{1}{m}\sum_{i=1}^{m} w_i. \tag{A.1}$$

Under the assumption that $g(x) \neq 0$ when $f(x) \neq 0$, $\frac{1}{m}\sum_{i=1}^{m} w_i$ converges as $m \to \infty$ to Z_f/Z_g, when $Z_f = \int f(x)\,dx$ and $Z_g = \int g(x)\,dx$ are the normalizing constants for $f(x)$ and $g(x)$. The variability of the IS estimate depends heavily on the variability of the weight functions. So to have a good IS estimate, we need to have $g(x)$ as a good approximation to $f(x)$, which is difficult to achieve in problems with high or moderately high-dimensional, possibly multimodal density.

Analysis of Bayesian factor models using IS has been introduced by Ghosh and Dunson (2008). The IS estimator of BF for factor models is based on m samples $\theta_i^{(h)}$ from the posterior distribution, under $M^{(h)}$

$$\widehat{BF}_{h-1,h} = \frac{1}{m}\sum_{i=1}^{m} \frac{p(y|\theta_i^{(h)}, k = h-1)}{p(y|\theta_i^{(h)}, k = h)}, \tag{A.2}$$

which in turn is based on the following identity:

$$\begin{aligned}&\int \frac{p(y|\theta^{(h)}, k = h-1)}{p(y|\theta^{(h)}, k = h)}\, p\big(\theta^{(h)}|y, k = h\big)\, d\theta^{(h)}\\ &= \int p\big(y|\theta^{(h)}, k = h-1\big)\frac{p(\theta^{(h)})}{p(y|k = h)}\, d\theta^{(h)} \qquad \text{(A.3)}\\ &= \frac{p(y|k = h-1)}{p(y|k = h)}.\end{aligned}$$

Ghosh and Dunson (2008) implemented IS with a parameter expanded prior. They also have noted that IS is fast and often (90%) chooses the correct model in simulation. In our simulation IS chooses a true bigger model correctly, but a 20% error rate was observed when the smaller model is true.

A.1.2 *Annealed importance sampling*. Following Neal (2001), we consider densities $p_t : t \in [0, 1]$ joining the densities p_0 and p_1. We choose densities by discretizing the path $p_{t_{(i)}}$ where $0 = t_{(1)} < \cdots < t_{(k)} = 1$ and then simulate a Markov chain designed to converge to $p_{t_{(k)}}$. Starting from the final states of the previous simulation, we simulate some number of iterations of a Markov chain designed to converge to $p_{t_{(k-1)}}$. Similarly, we simulate some iterations starting from the final steps of $p_{t_{(j)}}$ designed to converge to $p_{t_{(j-1)}}$ until we simulate some iterations converging to $p_{t_{(1)}}$. This sampling scheme produces a sample of points $x_1, \ldots, x_m$ and then we compute the weights $w_i = \frac{p_1(x_i)}{p_0(x_i)}$. Then the estimate of the ratio of normalizing constant becomes as follows:

$$\frac{1}{m}\sum_{i=1}^{m} w_i. \tag{A.4}$$

Notice that while both AIS and PS are based on MCMC runs along a path from one model to another, the MCMC'S are drawn at each point, but the details are very different. Due to the better spread of MCMC samples, the estimates in AIS seem to be better than

those calculated by IS when the smaller model is true, helping in correct model selection and also improving the estimation of Bayes factors. However, simulations show that AIS has the same problem as IS in estimating the Bayes factor when the bigger model is true.

A.1.3 *BIC type methods*: *Raftery–Newton and our method using information matrix*. In contrast to the methods previously discussed, we try to directly estimate the marginal under each model and then use these marginals to find the Bayes factor. We know that BIC is an approximation to the log-marginal based on a Laplace-type approximation of the log-marginal (Ghosh, Delampady and Samanta, 2006), under the assumption of i.i.d. observations. Thus,

$$\begin{aligned}\log\bigl(m(x)\bigr) &\approx \log\bigl(f(x|\hat{\theta})\pi(\hat{\theta})\bigr)\\ &\quad + (p/2)\log(2\pi) + (p/2)\log(n)\\ &\quad + \log\bigl(\bigl|H_{1,\hat{\theta}}^{-1}\bigr|^{1/2}\bigr),\end{aligned} \tag{A.5}$$

where $H_{1,\hat{\theta}}$ is the observed Fisher Information matrix evaluated at the maximum likelihood estimator using a single observation. For BIC we just use

$$\begin{aligned}\log\bigl(m(x)\bigr) &\approx \log\bigl(f(x|\hat{\theta})\pi(\hat{\theta})\bigr) + (p/2)\log(n)\\ &\approx \log\bigl(f(x|\hat{\theta})\bigr) + (p/2)\log(n),\end{aligned} \tag{A.6}$$

ignoring other terms as they are $O(1)$.

It is known BIC may be a poor approximation to the log-marginal in high-dimensions (Berger, Ghosh and Mukhopadhyay, 2003). To take care of this problem, Raftery et al. (2007) suggest the following. Simulate i.i.d. MCMC samples from the posterior distributions, evaluate independent sequence of log(prior × likelihood)s (log-p.l.) $\{l_t : t = 1, \ldots, m\}$, and then an estimate for the marginal is

$$\log\bigl(m(x)\bigr) \approx \bar{l} - s_l^2\bigl(\log(n) - 1\bigr), \tag{A.7}$$

where $\bar{l}$ and s_l^2 will be the sample mean and variance of l_t's. We call this method BICM, following the convention of Raftery et al. (2007).

In order to apply (A.5), we do not need to evaluate n since it cancels by combining the last two terms. This suggests the approximation (A.5) take care of the point raised by Clyde and George (2004). However, (A.7) does use n, but we do not know the impact on the approximation.

We have also used the Laplace approximation (A.5) without any change as likely to work better than the usual BIC. We compute the Information Matrix at the maximum prior × likelihood (mpl) value under the model and impute its value in the computation of the marginal. To find the mpl estimate, we use the MCMC sample from the posterior distribution and pick the maxima in that sample. Then we search for the mple in its neighborhood, using it as the starting point for the optimization algorithm. In our simulation study, it has been seen to give very good results similar to the computationally intensive numerical algorithms used to find the maximum of a function over the whole parameter space seen by taking repetition of MCMC runs and large MCMC samples. In the spirit of Raftery et al. (2007), we call this method BICIM, indicating the use of Information Matrix based Laplace Approximation. We also used several other modifications that did not give good results, so are not reported.

A.2 A Theoretical Remark on the Likelihood Function

It appears that the behavior of the likelihood, for example, its maximum, plays an important role in model selection, specifically in the kind of conflict we see between PS and the Laplace approximations (BICM, BICIM) when the bigger model is true (and the prior is a t with a relatively small d.f.). The behavior seems to be different from the asymptotic behavior of maximum likelihood under the following standard assumptions. Assume dimension of the parameter space is fixed and usual regularity conditions hold. Moreover, when the big model is true but the small model is assumed (so that it is a misspecified model), the Kullback–Liebler projection of the true parameter space to the parameter space of the small model exists (Bunke and Milhaud, 1998).

FACT. Assume the big model is true, and the small model is false. Then, as may be verified easily by the Taylor expansion,

(1) $\log L(\hat{\theta}_{\text{big}}) - \log L(\theta_{\text{true(big)}}) = O_P(1)$

(2) $\log L(\hat{\theta}_{\text{small}}) - \log L(\text{KL projection of } \theta_{\text{true(big)}} \text{ to } \Theta_{\text{small}}) = O_P(1)$

(3) $\log L(\theta_{\text{true(big)}}) - \log L(\text{KL projection of } \theta_{\text{true(big)}} \text{ to } \Theta_{\text{small}}) = O_P(n)$

and

(4)

$$\begin{aligned}&\log L(\hat{\theta}_{\text{big}}) - \log L(\hat{\theta}_{\text{small}})\\ &\quad = \log L(\theta_{\text{true(big)}})\\ &\quad\quad - \log L(\text{KL projection of } \theta_{\text{true(big)}} \text{ to } \Theta_{\text{small}})\\ &\quad\quad + O_P(1)\\ &\quad = O_P(n).\end{aligned} \tag{A.8}$$

The maximized likelihood for factor models substantially overestimates the true likelihood, unlike relation (1) above. Unfortunately, as pointed out in Drton (2009), the asymptotics of mle for factor models is still not fully worked out.

A.3 Matrix Used for the Toy Example

$$\Sigma^0 = \begin{pmatrix} 128.35 & 52.69 & -19.25 & -11.86 & 24.34 \\ 52.69 & 73.37 & -21.04 & -37.85 & 12.29 \\ -19.25 & -21.04 & 30.86 & 8.63 & -1.41 \\ -11.86 & -37.85 & 8.63 & 80.49 & 4.66 \\ 24.34 & 12.29 & -1.41 & 4.66 & 15.45 \\ 8.80 & 8.74 & -13.58 & 3.26 & 2.58 \\ 10.63 & 15.60 & -3.03 & -49.24 & 2.05 \\ 13.75 & 12.09 & -11.64 & -9.68 & 3.72 \\ -7.40 & -14.08 & 21.28 & 22.18 & -1.31 \\ -29.80 & -17.27 & 22.05 & 8.52 & -7.87 \end{pmatrix}$$

$$\begin{pmatrix} 8.80 & 10.63 & 13.75 & -7.40 & -29.80 \\ 8.74 & 15.60 & 12.09 & -14.08 & -17.27 \\ -13.58 & -3.03 & -11.64 & 21.28 & 22.05 \\ 3.26 & -49.24 & -9.68 & 22.18 & 8.52 \\ 2.58 & 2.05 & 3.72 & -1.31 & -7.87 \\ 31.37 & 11.62 & -4.85 & -16.89 & -20.10 \\ 11.62 & 58.09 & 7.00 & -19.58 & 5.16 \\ -4.85 & 7.00 & 26.59 & -3.04 & 11.17 \\ -16.89 & -19.58 & -3.04 & 31.81 & 22.86 \\ -20.10 & 5.16 & 11.17 & 22.86 & 64.68 \end{pmatrix}$$

A.4 Choice of Prior Under M_0

A referee has asked whether under M_0, the prior for the extra parameter can be chosen in a same optimal or philosophically compelling manner. This has been a long-standing problem, but the method followed for factor models is one of the standard procedures, apparently first suggested by Edwards, Lindman and Savage (1984).

This prior is mentioned by Edwards, Lindman and Savage (1984) and may be justified as follows. One tries to ensure the extra parameter has similar roles under both the models. If the joint prior of (θ_1, θ_2) under M_1 is $\pi(\theta_1, \theta_2)$, then the natural prior for $(\theta_2|\theta_1)$ is the usual conditional density of $\pi(\theta_2|\theta_1)$. In our case $\pi(\theta_1, \theta_2) = \pi(\theta_1)\pi(\theta_2)$. So $\pi(\theta_2|\theta_1)$ is as we have chosen. This is one of the standard default choices. Another default choice is due to Jeffreys (1961), but when θ_1, θ_2 are independent, both lead to the same choice. If we introduce a prior (e.g., minimizing MCMC-variance), it may not be acceptable to Bayesian philosophy.

ACKNOWLEDGMENTS

We thank Joyee Ghosh for helping us in discussions on factor models and sharing her code and Andrew Lewandowski for thought-provoking comments on an earlier draft.

REFERENCES

AKAIKE, H. (1987). Factor analysis and AIC. *Psychometrika* **52** 317–332. MR0914459

ANDERSON, T. W. (1984). *An Introduction to Multivariate Statistical Analysis*, 2nd ed. Wiley, New York. MR0771294

ANDRIEU, C., DOUCET, A. and ROBERT, C. P. (2004). Computational advances for and from Bayesian analysis. *Statist. Sci.* **19** 118–127. MR2082151

BARTHOLOMEW, D. J., STEELE, F., MOUSTAKI, I. and GABBRITH, J. I. (2002). *The Analysis and Interpretation of Multivariate Data for Social Scientists*. Chapman & Hall, Boca Raton, FL.

BERGER, J. O., GHOSH, J. K. and MUKHOPADHYAY, N. (2003). Approximations and consistency of Bayes factors as model dimension grows. *J. Statist. Plann. Inference* **112** 241–258. MR1961733

BUNKE, O. and MILHAUD, X. (1998). Asymptotic behavior of Bayes estimates under possibly incorrect models. *Ann. Statist.* **26** 617–644. MR1626075

CHEN, M.-H., SHAO, Q.-M. and IBRAHIM, J. G. (2000). *Monte Carlo Methods in Bayesian Computation*. Springer, New York. MR1742311

CHIB, S. (1995). Marginal likelihood from the Gibbs output. *J. Amer. Statist. Assoc.* **90** 1313–1321. MR1379473

CLYDE, M. and GEORGE, E. I. (2004). Model uncertainty. *Statist. Sci.* **19** 81–94. MR2082148

DICICCIO, T. J., KASS, R. E., RAFTERY, A. and WASSERMAN, L. (1997). Computing Bayes factors by combining simulation and asymptotic approximations. *J. Amer. Statist. Assoc.* **92** 903–915. MR1482122

DRTON, M. (2009). Likelihood ratio tests and singularities. *Ann. Statist.* **37** 979–1012. MR2502658

EDWARDS, W., LINDMAN, H. and SAVAGE, L. J. (1984). Bayesian statistical inference for psychological research. In *Robustness of Bayesian Analyses. Stud. Bayesian Econometrics* **4** 1–62. North-Holland, Amsterdam. MR0785366

FAN, Y., WU, R., CHEN, M.-H., KUO, L. and LEWIS, P. O. (2011). Choosing among partition models in Bayesian phylogenetics. *Mol. Biol. Evol.* **28** 523–532.

FRIEL, N. and PETTITT, A. N. (2008). Marginal likelihood estimation via power posteriors. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **70** 589–607. MR2420416

GAMARNIK, D., SHAH, D. and WEI, Y. (2010). Belief propagation for min-cost network flow: Convergence & correctness. In *Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms* 279–292. SIAM, Philadelphia, PA. MR2809676

GAMERMAN, D. and LOPES, H. F. (2006). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*, 2nd ed. Chapman & Hall/CRC, Boca Raton, FL. MR2260716

GELMAN, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Anal*. **1** 515–533 (electronic). MR2221284

GELMAN, A. and MENG, X.-L. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statist. Sci*. **13** 163–185. MR1647507

GELMAN, A., CARLIN, J. B., STERN, H. S. and RUBIN, D. B. (2004). *Bayesian Data Analysis*, 2nd ed. Chapman & Hall/CRC, Boca Raton, FL. MR2027492

GHOSH, J. K., DELAMPADY, M. and SAMANTA, T. (2006). *An Introduction to Bayesian Analysis*: *Theory and Methods*. Springer, New York. MR2247439

GHOSH, J. and DUNSON, D. B. (2008). *Random Effect and Latent Variable Model Selection. Lecture Notes in Statistics* **192**. Springer, New York. MR2761923

GHOSH, J. and DUNSON, D. B. (2009). Default prior distributions and efficient posterior computation in Bayesian factor analysis. *J. Comput. Graph. Statist*. **18** 306–320. MR2749834

GREEN, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82** 711–732. MR1380810

JEFFREYS, H. (1961). *Theory of Probability*, 3rd ed. Clarendon Press, Oxford. MR0187257

LARTILLOT, N. and PHILIPPE, H. (2006). Computing Bayes factors using thermodynamic integration. *Syst. Biol*. **55** 195–207.

LEE, S.-Y. and SONG, X.-Y. (2002). Bayesian selection on the number of factors in a factor analysis model. *Behaviormetrika* **29** 23–39. MR1894459

LEFEBVRE, G., STEELE, R., VANDAL, A. C., NARAYANAN, S. and ARNOLD, D. L. (2009). Path sampling to compute integrated likelihoods: An adaptive approach. *J. Comput. Graph. Statist*. **18** 415–437. MR2749839

LIANG, F., PAULO, R., MOLINA, G., CLYDE, M. A. and BERGER, J. O. (2008). Mixtures of g priors for Bayesian variable selection. *J. Amer. Statist. Assoc*. **103** 410–423. MR2420243

LIU, J. S. (2008). *Monte Carlo Strategies in Scientific Computing*. Springer, New York. MR2401592

LOPES, H. F. and WEST, M. (2004). Bayesian model assessment in factor analysis. *Statist. Sinica* **14** 41–67. MR2036762

LYNCH, S. M. (2007). *Introduction to Applied Bayesian Statistics and Estimation for Social Scientists*. Springer, New York.

MENG, X.-L. and WONG, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statist. Sinica* **6** 831–860. MR1422406

NEAL, R. M. (2001). Annealed importance sampling. *Stat. Comput*. **11** 125–139. MR1837132

NIELSEN, F. B. (2004). Variational approach to factor analysis and related models. Master's thesis, Institute of Informatics and Mathematical Modelling, Technical Univ. Denmark.

RAFTERY, A. E., NEWTON, M. A., SATAGOPAN, J. M. and KRIVITSKY, P. N. (2007). Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. In *Bayesian Statistics* 8 371–416. Oxford Univ. Press, Oxford. MR2433201

ROBERT, C. P. and CASELLA, G. (2004). *Monte Carlo Statistical Methods*, 2nd ed. Springer, New York. MR2080278

RUE, H., MARTINO, S. and CHOPIN, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Stat. Soc. Ser. B Stat. Methodol*. **71** 319–392. MR2649602

SHEN, G. and GHOSH, J. K. (2011). Developing a new BIC for detecting change-points. *J. Statist. Plann. Inference* **141** 1436–1447. MR2747912

SONG, X.-Y. and LEE, S.-Y. (2006). Model comparison of generalized linear mixed models. *Stat. Med*. **25** 1685–1698. MR2227347

XIE, W., LEWIS, P. O., FAN, Y., KUO, L. and CHEN, M.-H. (2011). Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Syst. Biol*. **60** 150–160.