Insper

A Tutorial on the Computation of Bayes Factors

Hedibert Freitas Lopes

Insper Working Paper
WPE: 341/2014

A Tutorial on the Computation of Bayes Factors

Hedibert Freitas Lopes

INSPER Institute of Education and Research
Rua Quatá 300, Vila Olímpia
São Paulo/SP - Brazil - 04546-042
E-mail: hedibertFL@insper.edu.br

Abstract

In this review paper we revisit several of the existing schemes that approximate predictive densities and, consequently, Bayes factors. We also present the reversible jump MCMC scheme, which can be thought of as an MCMC scheme over the space of models. These approaches are applied to select the number of common factors in the basic normal linear factor model, which is a high profile example within the psychometrics community.

1 Introduction

Bayesian model comparison is commonly (but not exclusively) performed by computing posterior model probabilities. Suppose that the competing models can be enumerated and are represented by the set $\mathcal{M} = \{\mathcal{M}_1, \mathcal{M}_2, \ldots\}$. Under model $\mathcal{M}_j$ with corresponding parameter vector θ_j, the posterior distribution of θ_j is obtained in the usual manner

$$p(\theta_j|\mathcal{M}_j, y) = \frac{p(y|\theta_j, \mathcal{M}_j)p(\theta_j|\mathcal{M}_j)}{p(y|\mathcal{M}_j)}, \tag{1}$$

where $p(y|\theta_j, \mathcal{M}_j)$ and $p(\theta_j|\mathcal{M}_j)$ respectively represent the likelihood and the prior distribution of θ_j under model $\mathcal{M}_j$. Predictive densities (aka normalizing constants or marginal likelihoods)

$$p(y|\mathcal{M}_j) = \int p(y|\theta_j, \mathcal{M}_j)p(\theta_j|\mathcal{M}_j)d\theta_j, \tag{2}$$

play an important role in Bayesian model selection and comparison. The posterior odds of model $\mathcal{M}_j$ relative to model $\mathcal{M}_k$ is defined as $Pr(\mathcal{M}_j|y)/Pr(\mathcal{M}_k|y)$, which is the product of the prior odds $Pr(\mathcal{M}_j)/Pr(\mathcal{M}_k)$ of model $\mathcal{M}_j$ relative to model $\mathcal{M}_k$ by the *Bayes factor*,

$$B_{jk} = \frac{p(y|\mathcal{M}_j)}{p(y|\mathcal{M}_k)} \ . \tag{3}$$

The Bayes factor can be viewed as the weighted likelihood ratio of $\mathcal{M}_j$ to $\mathcal{M}_k$. Hence, the posterior model probability for model j is

$$Pr(\mathcal{M}_j|y) = \left\{ \sum_{k=1}^{\infty} B_{kj} \frac{Pr(\mathcal{M}_k)}{Pr(\mathcal{M}_j)} \right\}^{-1} , \tag{4}$$

which depends on all prior odds ratios and all Bayes factors involving model j. When the prior model probabilities are uniformly distributed, the posterior model probabilities equal the Bayes factor. Jeffreys (1961) recommends the use of the following rule of thumb to decide between models j and k: when B_{jk} is above 100, between 10 and 100 and between 3 and 10, there is decisive, strong or substantial evidence against k, respectively.

Markov Chain Monte Carlo methods freed the Bayesian community by accurately approximating posterior distributions, $p(\theta_j|\mathcal{M}_j, y)$, for virtually all sorts of model structures. However, one can argue that the hardest computational task for an applied Bayesian is the computation of the normalizing constant $p(y|\mathcal{M}_j)$, which involves a multidimensional integral in θ_j.

In this review paper we revisit several of the existing schemes that approximate predictive densities, $p(y|\mathcal{M})$, and, consequently, Bayes factors. We also present the reversible jump MCMC scheme, which can be thought of as an MCMC scheme over the space of models. These two approaches are presented in Sections 3 and 4, respectively, following Section 2 where the basic normal linear factor model is introduced as a motivational high profile example within the psychometrics community. The factor model appears again in Section 5 through a computationally intensive simulated exercise. We conclude in Section 6.

2 Factor analysis

We will use the basic linear Gaussian factor model framework as an illustration of the computation of Bayes factors. In the factor analysis case, competing models have distinct number of common factors. The origin of factor analysis can be tracked back to Spearman's (1904) seminal paper on general intelligence. At the time, psychologists were trying to define intelligence by a single, all-encompassing unobservable entity, the g factor. Spearman studied the influence of the g factor on examinees test scores on several domains: pitch, light, weight, classics, french, english and mathematics. At the end of the day, the g factor would provide a mechanism to detect common correlations among such imperfect measurements.

Spearman's (1904) one-factor model based on p test domains (measurements) and n examinees (individuals) can be written as

$$y_{ij} = \mu_j + \beta_j g_i + \varepsilon_{ij}, \tag{5}$$

for $i = 1, \ldots, n$, $j = 1, \ldots, p$, where y_{ij} is the score of examinee i on test domain j, μ_j is the mean of test domain j, g_i is the value of the intelligence factor for person i, β_j is the loading of test domain j onto the intelligence factor g and ϵ_{ij} is the random error term for person i and test domain j. For subsequent developments, mainly in psychology studies, see Burt (1940), Holzinger and Harman (1941) and Thomson (1953), amongst others, where the factors had *a priori* known structure.

The extension to multiple factors as well as its formal statistical framework come many decades later. Multiple factor analysis were first introduced by Thurstone (1935,1947) and Lawley (1940,1953), along with estimation via centroid method and maximum likelihood, respectively. Hotelling (1955) proposed a more robust method of estimation, the method of principal components, while Anderson and Rubin (1956) formalized and elevated factor analysis to the realm of statistically and probabilistically sound modeling schemes. Maximum likelihood estimation became practical in the late 1960s through the work of Joreskog(1967,1969). A further improvement was achieved in the early 1980s through the EM algorithms of Rubin and Thayer (1982,1983); see also Bentler and Tanaka (1983). In the late 1980s, Anderson and Amemiya (1988) and Amemiya and Anderson (1990), studied the asymptotic behavior of estimation and hypothesis testing for a large class of factor analysis under

general conditions, while Akaike (1987) proposed an information criterion to selecting the proper number of common factors. To celebrate the centennial of Spearman (1904), The L. L. Thurstone Psychometric Laboratory, University of North Carolina at Chapel Hill, hosted in May 2004 a workshop entitled *Factor Analysis at 100: Historical Developments and Future Directions*. The papers presented at the meeting appeared in Cudeck and MacCallum (2007).

Bayesian normal linear factor analysis. Let $y_i = (y_{i1}, \ldots, y_{ip})'$, for $i = 1, \ldots, n$, be a p-dimensional vector with the measurements on p related variables (Spearman's tests, attributes, macroeconomic or financial time-series, census sectors, monitoring stations, to name a few examples). The basic normal linear factor model assumes that ys are independent and identically distributed $N(0, \Omega)$, ie. a zero-mean multivariate normal with a $p \times p$ non-singular variance matrix Ω. Loosely speaking, a factor model usually rewrites Ω, which depends of $q = p(p+1)/2$ variance and covariance components, as a function of d parameters, where d is potentially many orders of magnitude smaller than q.

More specifically, for any positive integer $k \leq p$, a standard normal linear k-factor model for y_i is written as

$$y_i | f_i, \beta, \Sigma, k \sim N(\beta f_i, \Sigma) \tag{6}$$

$$f_i | H, k \sim N(0, H) \tag{7}$$

where f_i is the k-dimensional vector of common factors, β is the $p \times k$ matrix of factor loadings, $\Sigma = \text{diag}(\sigma_1^2, \cdots, \sigma_p^2)$ is the covariance of the specific factors and $H = \text{diag}(h_1, \ldots, h_k)$ is the covariance matrix of the common factors. The uniquenesses σ_i^2s, also known as idiosyncratic or specific variances, measure the residual variability in each of the data variables once that contributed by the factors is accounted for. Conditionally on the common factors, f_k, the measurements in y_i are independent. In other words, the common factors explain all the dependence structure among the p variables and, based on equations (6) and (7), the unconditional, constrained covariance matrix of y_i (a function of k) becomes

$$\Omega = \beta H \beta' + \Sigma. \tag{8}$$

Parsimony. The matrix Ω depends on $d = (p+1)(k+1) - 1$, the number of elements of β, H and Σ, a number usually considerably smaller than $q = p(p+1)/2$, the number of elements of the unconstrained Ω. In practical problems, especially with larger values of p, the number of factors k will often be small relative to p, so most of the variance-covariance structure is explained by a few number of common factors. For example, when $p = 100$ and $k = 10$, a configuration commonly found in modern applications of factor analysis, $q = 5050$ and $d = 1110$, or roughly $q = 5d$. Similarly, when $p = 1000$ and $k = 50$, if follows that $q = 500500$ and $d = 51050$, or roughly $q = 10d$. Such drastic reduction in the number of unrestricted parameters renders factor modeling inherently a parsimonious inducing technique. See Lopes and West (2004) for additional details on identifiability, invariance, reduced-rank and rotation issues.

Prior specification. The unconstrained components of β are independent and identically distribution (i.i.d.)$N(\mathcal{M}_0, C_0)$, the diagonal components β_{ii}s are i.i.d. truncated normal from below at zero, denoted here by $N_{(0,\infty)}(\mathcal{M}_0, C_0)$, and the idiosyncratic variances, σ_i^2 are i.i.d. $IG(\nu/2, \nu s^2/2)$. The hyperparameters $\mathcal{M}_0$, C_0, ν and s^2 are known. It is worth mentioning that the above prior specification has been extended and modified many times over to accommodate specific characteristics of the scientific modeling under consideration. Lopes et al. (2008), for example, utilize spatial proximity to parameterize the columns of β when modeling pollutants across Eastern US monitoring stations.

Posterior inference. Early MCMC-based posterior inference in standard factor analysis appears in, among others, Geweke and Zhou (1996) and Lopes and West (2004). Conditionally on the number of factors, kk, they basically propose and implement a standard Gibbs sampler that cycles through the full conditional distributions of $p(\beta|f, \Sigma, y, k)$, $p(\Sigma|f, \beta, y, k)$ and $p(f|\beta, \Sigma, y, k)$, which following well known distributions when conditionally conjugate priors are used. Here we assume that $H = I_k$ and that β is block lower triangular, for identification structure, with strictly positive diagonal elements. See Lopes and West (2004) for additional details. Lopes (2014) presents an extensive overview of modern Bayesian factor modeling, including prior

and posterior robustness, mixture of factor analyzers, factor analysis in time series and macro-econometric modeling and sparse factor structures in micro-econometrics and genomics. He also lists some of the recent contributions to the literature on non-Bayesian (large dimensional and/or dynamic) factor analysis.

3 Computing Bayes factor

In what follows, we present several approximations to the predictive density, most of them based on Monte Carlo draws from the posterior distribution and/or from auxiliary distributions. To simplify the notation, we rewrite equation (2) as

$$p(y) = \int p(y|\theta)p(\theta)d\theta, \tag{9}$$

omitting any explicit dependence on model $\mathcal{M}_i$. Recall that if $\mathcal{M}_k$ is a k-factor model (Section 2), θ_k represents the parameters of the model and $\theta_k = (\beta_k, \Sigma_k)$ corresponds to the factor loadings matrix and the idiosyncratic covariance matrix.

3.1 Normal approximation

The normal approximation to the posterior leads to $p(y|\hat{\theta})p(\hat{\theta})(2\pi)^{d/2}|V|^{1/2}$ as an approximation to $p(y)$, which is based on the evaluation of the values of $\hat{\theta}$, the posterior mode, and V, an asymptotic approximation for the posterior variance matrix. Sampling-based approximations for $\hat{\theta}$ and V can be constructed if a sample $\theta^{(1)}, \ldots, \theta^{(N)}$ from the posterior is available. The mode $\hat{\theta}$ can be estimated as the sample value $\tilde{\theta}$ for which $p(\theta|y)$ is largest, i.e., $p(\tilde{\theta}|y) = \max_j\{p(\theta^{(j)}|y)\}$. Similarly, estimates for the posterior variance matrix may be given in the case of an independent sample by $\tilde{V} = \frac{1}{N}\sum_{j=1}^{N}(\theta^{(j)} - \bar{\theta})(\theta^{(j)} - \bar{\theta})'$, where $\bar{\theta} = \frac{1}{N}\sum_{j=1}^{N}\theta^{(j)}$. Therefore,

$$\hat{p}_0 = p(\tilde{\theta})p(y|\tilde{\theta})(2\pi)^{d/2}|\tilde{V}|^{1/2}$$

is the normal approximation to $p(y)$. Lewis and Raftery (1997) named this estimator the *Laplace-Metropolis* estimator. Kass and Raftery (1995), Raftery (1996) and DiCiccio et al. (1997), among others, discussed alternative calculations of the value of $\hat{\theta}$ when computation of $p(\theta|y)$ is expensive and of the value of $\hat{V}$ with the use of robust estimators.

3.2 Monte Carlo approximations

The (simple) Monte Carlo estimate derived from the identity of Equation (9) is

$$\hat{p}_1 = \frac{1}{N}\sum_{j=1}^{N} p(y|\theta^{(j)})$$

where $\theta^{(j)}, \ldots, \theta^{(N)}$ is a sample from the prior distribution $p(\theta)$. Raftery (1996) argued that this estimator does not work well in cases of disagreement between prior and likelihood. It averages likelihood values that are chosen according to the prior. In general, the likelihood is more concentrated than the prior and the majority of θ_i will be placed in low likelihood regions. Even for large values of n, this estimate will be influenced by a few sampled values, making it very unstable.

An alternative is to perform importance sampling with the aim of boosting sampled values in regions where the integrand is large. This approach is based on sampling from the importance density $g(\theta) = cg^*(\theta)$ where g^* is the unnormalized form of the density and c is a normalizing constant. It is easy to see that

$$p(y) = E_g\left\{\frac{p(y|\theta)p(\theta)}{g(\theta)}\right\}, \tag{10}$$

where E_g denotes an expectation with respect to the importance distribution $g(\theta)$. This form motivates new estimates

$$\hat{p}_2 = \frac{1}{N}\sum_{j=1}^{N}\frac{p(y|\theta^{(j)})p(\theta^{(j)})}{g(\theta^{(j)})} \quad \text{and} \quad \hat{p}_3 = \frac{\sum_{j=1}^{N} p(y|\theta^{(j)})p(\theta^{(j)})/g^*(\theta^{(j)})}{\sum_{j=1}^{N} p(\theta^{(j)})/g^*(\theta^{(j)})},$$

for the cases where k is, respectively, known and unknown, and $\theta^{(1)}, \ldots, \theta^{(N)}$ is a sample from the importance density $g(\theta)$.

Newton and Raftery (1994) propose two well-known special cases are $g(\theta) = \pi(\theta)$ or $g(\theta) = \delta p(\theta) + (1-\delta)\pi(\theta)$. The first case leads to the *harmonic mean estimator*:

$$\hat{p}_4 = \left(\frac{1}{N}\sum_{j=1}^{N}\frac{1}{p(y|\theta^{(j)})}\right)^{-1}. \tag{11}$$

Its simplicity makes it a very appealing estimator and its use is recommended provided N is large enough. Despite its consistency, this estimator is strongly affected

by small likelihood values. Raftery (1996) relates this weakness to the occasional divergence of the variance of the terms in (11). See the recent discussion by Wolpert and Schmidler (2012).

The second case proposes an estimator that is a compromise between $\hat{p}_1$, derived from prior draws, and $\hat{p}_4$, derived from posterior draws:

$$\hat{p}_5^{(i)} = \frac{\sum_{j=1}^{N} p(y|\theta^{(j)})\{\delta\hat{p}_5^{(i-1)} + (1-\delta)p(y|\theta^{(j)})\}^{-1}}{\sum_{j=1}^{N}\{\delta\hat{p}_5^{(i-1)} + (1-\delta)p(y|\theta^{(j)})\}^{-1}} \tag{12}$$

for $i = 1, 2, \ldots$ and, say, $\hat{p}_5^{(0)} = \hat{p}_4$. A small number of iterations is usually enough for convergence. The estimator avoids the instability of $\hat{p}_4$ with the additional cost of also simulating from the prior.

Another generalization of the harmonic mean estimator was obtained by Gelfand and Dey (1994) based on the identity

$$\int g(\theta)\frac{p(y)p(\theta|y)}{p(y|\theta)p(\theta)}d\theta = 1. \tag{13}$$

Sampling $\theta^{(1)}, \ldots, \theta^{(N)}$ from the posterior leads to the estimate

$$\hat{p}_6 = \left(\frac{1}{n}\sum_{j=1}^{n}\frac{g(\theta_j)}{f(y|\theta_j)p(\theta_j)}\right)^{-1}. \tag{14}$$

Even though the method is specified for any density g, appropriate choices are very important for a good practical implementation. Gelfand and Dey (1994) suggested using g as an importance density for the posterior and to take a normal or t distribution that approximates π with moments based on the sample of θ. Raftery (1996) presented a simple example where g was taken in product forms for each parameter component. The estimates obtained are highly inaccurate, showing that some skill is required in choosing g.

3.3 Bridge sampler

Meng and Wong (1996) introduced the *bridge sampling* to estimate ratios of normalizing constants by noticing that

$$p(y) = \frac{E_g\{\alpha(\theta)p(\theta)p(y|\theta)\}}{E_{p(\theta|y)}\{\alpha(\theta)g(\theta)\}} \tag{15}$$

for any arbitrary *bridge* function $\alpha(\theta)$ with support encompassing both supports of the posterior density π and the proposal density g. If $\alpha(\theta) = 1/g(\theta)$ then the bridge estimator reduces to the simple Monte Carlo estimator $\hat{p}_1$. Similarly, if $\alpha(\theta) = \{p(\theta)p(y|\theta)g(\theta)\}^{-1}$ then the bridge estimator is a variation of the harmonic mean estimator. They showed that the optimal mean square error α function is $\alpha(\theta) = \{g(\theta) + (M/N)\pi(\theta)\}^{-1}$, which depends on $f(y)$ itself. By letting $\omega^{(j)} = (\theta^{(i)})p(\theta^{(i)})/g(\theta^{(i)})$, for $j = 1, \ldots, N$ and $\tilde{\omega}^{(i)} = p(y|\tilde{\theta}^{(i)})p(\tilde{\theta}^{(i)})/g(\tilde{\theta}^{(i)})$, for $j = 1, \ldots, M$, they devised the iterative scheme to estimate $p(y)$:

$$\hat{p}_7^{(i)} = \frac{\frac{1}{M}\sum_{j=1}^{M} \tilde{\omega}^{(j)}[s_1\tilde{\omega}^{(j)} + s_2\hat{p}_7^{(i-1)}]^{-1}}{\frac{1}{N}\sum_{j=1}^{N}[s_1\omega^{(j)} + s_2\hat{p}_7^{(i-1)}]^{-1}} ,$$

for $i = 1, 2, \ldots$, $s_1 = N/(M+N)$, $s_2 = M/(M+N)$ and, say, $\hat{p}_7^{(0)} = \hat{p}_4$. A small number of iterations is usually enough for convergence. See also Meng and Schilling (1996), Gelman and Meng (1997) and Meng and Schilling (2002). Gelman and Meng (1998) generalized the bridge sampling by replacing one (possibly long) bridge by infinitely many shorter bridges or, as they call it, a *path*.

3.4 Candidate's estimators

A very simple estimate, usually called the *candidate's estimator* (Besag, 1989), can be derived from the fact that $p(y) = p(y|\theta)p(\theta)/p(\theta|y)$ for any value of θ. Typically, $p(y|\theta)$ and $p(\theta)$ are easy to calculate but $p(\theta|y)$ is not. However, if a sample of π is available, some form of histogram smoothing can be applied to get an estimate of $p(\theta|y)$. Chib (1995) introduced an alternative estimate of $p(\theta|y)$ when full conditional densities are available in closed form, as in Gibbs sampling. For simplicity, we will show here the case where $\theta = (\theta_1, \theta_2)$, so that $p(\theta_1, \theta_2|y) = p(\theta_2|\theta_1, y)p(\theta_1|y)$. The conditional $p(\theta_2|\theta_1, y)$ can be evaluated exactly for any pair (θ_1, θ_2), while $p(\theta_1|y)$ can be approximated by

$$\hat{p}(\theta_1|y) = \frac{1}{N}\sum_{j=1}^{N} p(\theta_1|\theta_2^{(j)}, y)$$

where $\theta_2^{(1)}, \ldots, \theta_2^{(N)}$ are draws from $p(\theta_2|y)$, obtained by Gibbs sampler. Therefore,

$$\hat{p}_8 = \frac{p(y|\theta)p(\theta)}{\hat{p}(\theta_1|y)p(\theta_2|\theta_1, y)}, \tag{16}$$

is, for any value $\theta = (\theta_1, \theta_2)$, the candidate's estimator of $p(y)$. θ should be chosen so that $\hat{\pi}$ has the smallest possible estimation error. This narrows the choice of θ to the central region of the posterior where π is likely to be estimated more accurately. Simple choices are the mode and the mean but any value in that region should be adequate. Chib and Jeliazkov (2001,2005) extended the above idea for cases where some (or none) of the full conditional densities are of unknown form and difficult to sample from and Metropolis-Hastings output is available. Mira and Nicholls (2004) showed that Chib and Jeliazkov's estimator is a special case of the bridge sampler. DiCiccio et al. (1997), Han and Carlin (2001) and Lopes and West (2004), among others, compared several of estimators introduced in this section.

4 Computing posterior model probabilities

We present the Reversible Jump algorithm as introduced in Green (1995). Among many others, Barbieri and O'Hagan (1996), Richardson and Green (1997), Dellaportas et al. (2002), Huerta and West (1999), Denison et al. (1998), Huerta and Lopes (2000), Lopes et al. (2008) utilized the RJMCMC algorithm to a series of models. We also explore its relationship to Carlin and Chib's (1995) pseudo-prior method. Particular attention is given to the *Metropolized Carlin-Chib* algorithm simultaneously introduced by Dellaportas et al. (2002) and Godsill (2001). The results presented here are mainly based on the developments from Dellaportas et al. (2002) and Godsill (2001). Additional overview and/or further extensions can be found in (Chen et al., 2000, Section 9.5), and (Gamerman and Lopes, 2006, Chapter 7).

Suppose that the competing models can be enumerable and are represented by the set $\mathcal{M} = \{\mathcal{M}_1, \mathcal{M}_2, \ldots\}$. Under model $\mathcal{M}_k$, the posterior distribution is

$$p(\theta_k|y, k) \propto p(y|\theta_k, k)p(\theta_k|k) \tag{17}$$

where $p(y|\theta_k, k)$ and $p(\theta_k|k)$ represent the probability model and the prior distribution of the parameters of model $\mathcal{M}_k$, respectively. Then,

$$p(\theta_k, k|y) \propto p(k)p(\theta_k|k, y) \tag{18}$$

4.1 Reversible Jump MCMC

The RJMCMC methods involve Metropolis-Hastings type algorithms that move a simulation analysis between models defined by (k, θ_k) to $(k', \theta_{k'})$ with different defining dimensions k and k'. The resulting Markov chain simulations jump between such distinct models and form samples from the joint distribution $p(\theta_k, k)$. The algorithm are designed to be reversible so as to maintain detailed balance of a irreducible and aperiodic chain that converges to the correct target measure. Further details of the general methodology and ideas can be found in Green (1995).

Here we present the algorithm in a schematic form. If the current state of the Markov chain is (k, θ_k), then one possible version of the RJMCMC algorithm is as follows:

Step 1. Propose a visit to model $\mathcal{M}_{k'}$ with probability $J(k \to k')$.

Step 2. Sample u from a proposal density $q(u|\theta_k, k, k')$.

Step 3. Set $(\theta_{k'}, u') = g_{k,k'}(\theta_k, u)$, where $g_{k,k'}(\cdot)$ is a bijection between (θ_k, u) and $(\theta_{k'}, u')$, where u and u' play the role of matching the dimensions of both vectors.

Step 4. The acceptance probability of the new model, $(\theta_{k'}, k')$ can be calculated as the minimum between one and

$$\underbrace{\frac{p(y|\theta_{k'}, k')p(\theta_{k'})p(k')}{p(y|\theta_k, k)p(\theta_k)p(k)}}_{\text{model ratio}} \underbrace{\frac{J(k' \to k)q(u'|\theta_{k'}, k', k)}{J(k \to k')q(u|\theta_k, k, k')} \left|\frac{\partial g_{k,k'}(\theta_k, u)}{\partial(\theta_k, u)}\right|}_{\text{proposal ratio}} \tag{19}$$

Looping through steps 1-4 generates a sample $\{k_l, l = 1, \ldots, L\}$ for the model indicators and $Pr(k|y)$ can be estimated by

$$\widehat{Pr}(k|y) = \frac{1}{L}\sum_{l=1}^{L} 1_k(k_l) \tag{20}$$

where $1_k(k_l) = 1$ if $k = k_l$ and zero otherwise. The choice of the model proposal probabilities, $J(k \to k')$, and the proposal densities, $q(u|k, \theta_k, k')$, must be cautiously made, especially in highly parameterized problems.

Independent sampler: If all parameters of the proposed model are generated from the proposal distribution, then $(\theta_{k'}, u') = (u, \theta_k)$ and the Jacobian in (19) is one.

Standard Metropolis-Hastings: When the proposed model k' equals the current model k, the loop through steps 1-4 corresponds to the traditional Metropolis-Hastings algorithm Metropolis et al. (1995); Hastings (1970); Peskun (1973); Chib and Greenberg (1995).

Posterior densities as proposal densities: If $p(\theta_k|y,k)$ is available in close form for each model $\mathcal{M}_k$, then $q(u'|\theta_{k'}, k', k) = p(\theta_k|y,k)$ and the acceptance probability (equation 19) reduces to the minimum between one and

$$\frac{p(k')p(y|k')}{p(k)p(y|k)} \frac{J(k' \to k)}{J(k \to k')} \tag{21}$$

using the fact that $p(y|\theta_k, k)p(\theta_k)p(k) = p(\theta_k, k|y)p(y|k)$. Again, the Jacobian equals one. The predictive density or normalizing constant, $p(y|k)$, is also available in close form. Moreover, if $J(k' \to k) = J(k \to k')$, the acceptance probability is the minimum between one and the posterior odds ratio from model $\mathcal{M}_{k'}$ to model $\mathcal{M}_k$, that is the move is automatically accepted when model $\mathcal{M}_{k'}$ has higher posterior probability than model $\mathcal{M}_k$; otherwise the posterior odds ratio determines how likely is to move to a lower posterior probability model.

4.2 Metropolized Carlin and Chib's algorithm

Let $\Theta = (\theta_k, \theta_{-k})$ be the vector containing the parameters of all competing models. Then the joint posterior of (Θ, k) is

$$p(\Theta, k|y) \propto p(k)p(y|\theta_k, k)p(\theta_k|k)p(\theta_{-k}|\theta_k, k) \tag{22}$$

where $p(\theta_{-k}|\theta_k, k)$ are *pseudo-prior* densities Carlin and Chib (1995). Carlin and Chib propose a Gibbs sampler where the full posterior conditional distributions are

$$p(\theta_k|y, k, \theta_{-k}) \propto \begin{cases} p(y|\theta_k, k)p(\theta_k|k) & \text{if } k = k' \\ p(\theta_k|k') & \text{if } k = k' \end{cases} \tag{23}$$

and

$$p(k|\Theta, y) \propto p(y|\theta_k, k)p(k) \prod_{m \in \mathcal{M}} p(\theta_m|k) \tag{24}$$

Notice that the pseudo-prior densities and the RJMCMC's proposal densities have similar functions. As a matter of fact, Carlin and Chib suggest using pseudo-prior distributions that are close to the posterior distributions within each competing model.

The main problem with Carlin and Chib's Gibbs sampler is the need of evaluating and drawing from the pseudo-prior distributions at each iteration of the MCMC scheme. This problem can be overwhelmingly exacerbated in large situations where the number of competing models is relatively large (See Clyde, 1999, for applications and discussions in variable selection in regression models).

To overcome this last problem Dellaportas et al. and Godsill (2001) proposes "Metropolizing" Carlin and Chib's Gibbs sampler. If the current state of the Markov chain is at (θ_k, k), then they suggest proposing and accepting/rejecting a move to a new model in the following way:

Step 1. Propose a new model $\mathcal{M}_{k'}$ with probability $J(k \to k')$.

Step 2. Generate $\theta_{k'}$ from the pseudo-prior $p(\theta_{k'}|k)$.

Step 3. The acceptance probability of the new model, k' can be calculated as the minimum between one and

$$\frac{p(y|\theta_{k'}, k')p(k')J(k' \to k) \prod_{m \in \mathcal{M}} p(\theta_m|k')}{p(y|\theta_k, k)p(k)J(k \to k') \prod_{m \in \mathcal{M}} p(\theta_m|k)}$$

which can be simplified to

$$\frac{p(y|\theta_{k'}, k')p(k')J(k' \to k)p(\theta_{k'}|k')p(\theta_k|k')}{p(y|\theta_k, k)p(k)J(k \to k')p(\theta_k|k)p(\theta_{k'}|k)} \tag{25}$$

since the other pseudo-prior densities cancel out.

Once again, if $p(\theta_k|y, k)$ is available in close form for each model $\mathcal{M}_k$, and $p(\theta_k|k') = p(\theta_k|y, k)$, then the acceptance probability in (25) reduces to (21). As we have mentioned earlier the pseudo-prior densities and the RJMCMC's proposal densities have similar functions and the closer their are to the competing models' posterior probabilities the better the sampler mixing.

5 Factor analysis revisited

We consider a one-factor model for a seven-dimensional problem generating one hundred observations. In each of 1,000 simulations, one-hundred observations were drawn from a one-factor models defined by parameters

$$\begin{aligned} \beta' &= (0.995, 0.975, 0.949, 0.922, 0.894, 0.866, 0.837), \\ \text{diag}(\Sigma) &= (0.01, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30). \end{aligned}$$

Each such simulated data set was analysed using the MCMC and reversible jump methodologies, and also subject to study using the range of model selected criteria and methods described above. This study explored $k-$factor models for each data set, with up to three possible factors in each case.

MCMC analyses utilised the prior distributions based on the following hyper-parameter values: $m_0 = 0$ and $C_0 = 1$ define the prior distribution for β, while σ_k^2, $\nu_{0i} = 2.2$ and $\nu_{0i}s_{0i}^2 = 0.1$ define the prior distribution for each σ_k^2 such that $E(\sigma_k^2) = 0.5$. The MCMC and reversible jump samplers were based on $M_0 = 10,000$ iterations as burn-in, followed by a further 10,000 iterates that were sampled every ten steps to produce a final MCMC sample of size 1,000. In generating proposals in the RJMCMC methods, we adopted $a = 18$, $b = 2$ and

$$J = \begin{pmatrix} 0.0 & 1.0 & 0.0 \\ 0.5 & 0.0 & 0.5 \\ 0.0 & 1.0 & 0.0 \end{pmatrix}.$$

Among the candidate methods for model selected, the "Newton and Raftery" technique requires the specification of a control parameter, δ; this was set at $\delta = 0.05$, and the number of iterations at 1,000.

Table 1 displays results from this simulation analysis. We repeated the model fitting exercises for 1,000 different data sets generated independently from the one-factor model. The table provides simple counts of the number of times that each k-factor model achieved the highest posterior probability. For example, the harmonic estimator method, $\hat{p}_4$, selected the one-factor model 428 times out of 1,000, and the three-factor model 314 times out of 1,000. Evidently, most of the approximation

methods are very reliable in favoring the one-factor model, as is the RJMCMC (the "gold standard") approach. Bridge sampling, $\hat{p}_7$, agrees with the RJMCMC approach. Relatively poor results are achieved by the harmonic mean method ($\hat{p}_4$), Newton-Raftery estimator ($\hat{p}_5$), AIC, and at some extent the candidate's estimator ($\hat{p}_8$), which all tend to prefer higher numbers of factors a significant proportion of the time. In terms of model selection *per se*, as opposed to exploring model uncertainty more formally, the BIC methods are relatively accurate and, of course, rather easier to compute.

	k		
Method	$k=1$	$k=2$	$k=3$
$\hat{p}_0$	1000	0	0
$\hat{p}_4$	428	258	314
$\hat{p}_5$	467	234	299
$\hat{p}_6$	1000	0	0
$\hat{p}_7$	1000	0	0
$\hat{p}_8$	954	46	0
RJMCMC	1000	0	0
BIC	1000	0	0
AIC	854	135	11

Table 1: *Model comparison:* AIC= $l_k + 2d_k$ (Akaike, 1987) and BIC= $l_k + \log(n)d_k$ (Schwarz, 1978), where $l_k = -2\log p(y|\hat{\beta}_k, \hat{\Sigma}_k)$, n is the number of observations, d_k is the number of parameters in the k-factor model and $\hat{\beta}_k$ and $\hat{\Sigma}$ are the maximum likelihood estimates of β and Σ, respectively.

6 Conclusion

We review the standard computational methods to approximate the normalizing constant, a key ingredient to computing Bayes factors. We have presented RJMCMC schemes that by-passes the computation of normalizing constants altogether and

directly approximate posterior model probabilities. We have used the standard normal linear factor model as a motivational example to illustrate the implementation of such methods. For additional discussion on Bayes factors and their approximations for model comparison, see Kass and Raftery (1995), DiCiccio et al. (1997), Han and Carlin (2001), Kadane and Lazar (2004), Lopes and West (2004) and Chapter 7 of Gamerman and Lopes (2006), amongst others.

References

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

Amemiya, Y. and T. W. Anderson (1990). Asymptotic chi-square tests for a large class of factor analysis models. *The Annals of Statistics 18*, 1453–1463.

Anderson, T. W. and Y. Amemiya (1988). The asymptotic normal distribution of estimators in factor analysis under general conditions. *The Annals of Statistics 16*, 759–771.

Anderson, T. W. and H. Rubin (1956). Statistical inference in factor analysis. In J. Neyman (Ed.), *Proceedings of the Third Berkeley Symposium of Mathematical Statistics and Probability*, Volume 5, pp. 111–150. University of California Press, Berkeley.

Barbieri, M. and A. O'Hagan (1996). A reversible jump MCMC sampler for Bayesian analysis of ARMA time series. Technical report, Dipartimento di Statistica,Probabilità e Statistiche Applicate, Università La Sapienza,Roma.

Bentler, P. M. and J. S. Tanaka (1983). Problems with EM algorithms for ML factor analysis. *Psychometrika 48*, 247–251.

Besag, J. (1989). A candidate's formula: A curious result in Bayesian prediction. *Biometrika 76*, 183–183.

Burt, C. (1940). *The Factors of the Mind: An Introduction to Factor Analysis in Psychology*. University of London Press, London.

Carlin, B. and S. Chib (1995). Bayesian model choice via Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, B 57*, 473–484.

Chen, M.-H., Q.-M. Shao, and J. Ibrahim (2000). *Monte Carlo methods in Bayesian computation*. New York: Springer-Verlag.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association 90*, 1313–1321.

Chib, S. and E. Greenberg (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician 49*, 327–335.

Chib, S. and I. Jeliazkov (2001). Marginal likelihood from the Metropolis-Hastings output. *Journal of the American Statistical Association 96*, 270–281.

Chib, S. and I. Jeliazkov (2005). Accept-reject Metropolis-Hastings sampling and marginal likelihood estimation. *Statistica Neerlandica 59*, 30–44.

Clyde, M. (1999). Bayesian model averaging and model search strategies (with discussion). In J. Bernardo, J. Berger, A. Dawid, and A. Smith (Eds.), *Bayesian Statistics 6*, pp. 157–185. John Wiley.

Cudeck, R. and R. C. MacCallum (2007). *Factor Analysis at 100: Historical Developments and Future Directions*. Routledge.

Dellaportas, P., J. Forster, and I. Ntzoufras (2002). On Bayesian model and variable selection using MCMC. *Statistics and Computing 12*, 27–36.

Denison, D. G. T., B. K. Mallick, and A. F. M. Smith (1998). Automatic Bayesian curve fitting. *Journal of the Royal Statistical Society. Series B 60*, 333–350.

DiCiccio, T., R. Kass, A. Raftery, and L. Wasserman (1997). Computing Bayes' factors by combining simulation and asymptotic approximations. *Journal of the American Statistical Association 92*, 903–915.

Gamerman, D. and H. F. Lopes (2006). *Markov Chain Monte Carlo - Stochastic Simulation for Bayesian Inference* (2nd ed.). Chapman&Hall/CRC.

Gelfand, A. and D. Dey (1994). Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society, Ser. B 56*, 501–514.

Gelman, A. and X. L. Meng (1997). On monte carlo methods for estimating ratios of normalizing constants. *Annals of Statistics 25*, 1563–1594.

Gelman, A. and X. L. Meng (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science 13*, 163–185.

Geweke, J. and G. Zhou (1996). Measuring the pricing error of the arbitrage pricing theory. *The Review of Financial Studies 9*, 557–587.

Godsill, S. J. (2001). On the relationship between Markov chain Monte Carlo methods for model uncertainty. *Journal of Computational and Graphical Statistics 10*(2), 230–248.

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

Han, C. and B. P. Carlin (2001). Markov chain monte carlo methods for computing bayes factors: A comparative review. *Journal of the American Statistical Association 96*, 1122–1132.

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

Holzinger, K. J. and H. H. Harman (1941). *Factor analysis: A Synthesis of Factorial Methods*. University of Chicago Press, Chicago.

Hotelling, H. (1955). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology 24*, 417–441.

Huerta, G. and H. F. Lopes (2000). Bayesian forecasting and inference in latent structure for the Brazilian industrial production index. *Brazilian Review of Econometrics 20*, 1–26.

Huerta, G. and M. West (1999). Priors and component structures in autoregressive time series models. *Journal of the Royal Statistical Society-Series B 61*, 881–899.

Jeffreys, H. (1961). *Theory of Probability (3rd edition)*. Oxford University Press.

Joreskog, K. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika 32*, 443–482.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika 34*, 183–220.

Kadane, J. B. and N. A. Lazar (2004). Methods and criteria for model selection. *Journal of the American Statistical Association 99*, 279–290.

Kass, R. and A. Raftery (1995). Bayes' factor. *Journal of the American Statistical Association 90*, 773–795.

Lawley (1940). The estimation of factor loadings by the method of maximum likelihood. *Proceedings of the Royal Society of Edinburgh 60*, 64–82.

Lawley (1953). Further investigations in factor estimation. *Proceedings of the Royal Society of Edinburgh 60*, 64–82.

Lewis, S. and A. Raftery (1997). Estimating Bayes' factors via posterior simulation with the Laplace-Metropolis estimator. *Journal of the American Statistical Association 92*, 648–655.

Lopes, H. F. (2014). Modern bayesian factor analysis. In I. Jeliazkov and X.-S. Yang (Eds.), *Bayesian inference in the Social Sciences*. Wiley.

Lopes, H. F., E. Salazar, and D. Gamerman (2008). Spatial dynamic factor models. *Bayesian Analysis 5*, 1–30.

Lopes, H. F. and M. West (2004). Bayesian model assessment in factor analysis. *Statistica Sinica 14*, 41–67.

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

Meng, X. L. and S. Schilling (1996). Fitting full-information factor models and an empirical investigation of bridge sampling. *Journal of the American Statistical Association 91*, 1254–1267.

Meng, X.-L. and S. Schilling (2002). Warp bridge sampling. *Journal of Computational and Graphical Statistics 11*, pp. 552–586.

Metropolis, N., A. Rosenbluth, M. Rosenbluth, and A. Teller (1995). Equations of state calculations by fast computing machines. *Journal of Chemical Physics 21*, 1087–1092.

Mira, A. and G. K. Nicholls (2004). Bridge estimation of the probability density at a point. *Statistica Sinica 14*, 603–612.

Newton, M. and A. Raftery (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society, Ser. B 56*, 3–48.

Peskun, P. (1973). Optimum Monte-Carlo sampling using Markov chains. *Biometrika 60*, 607–612.

Raftery, A. (1996). Hypothesis testing and model selection. In W. Gilks, S. Richardson, and D. Spiegelhalter (Eds.), *Markov Chain Monte Carlo in Practice*. Chapman and Hall.

Richardson, S. and P. Green (1997). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination (with discussion). *Journal of the Royal Statistical Society - Series B 59*, 731–758.

Rubin, D. B. and D. T. Thayer (1982). EM algorithms for ML factor analysis. *Psychometrika 47*, 69–76.

Rubin, D. B. and D. T. Thayer (1983). More on the EM for factor analysis. *Psychometrika 48*, 253–257.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics 6*, 461–464.

Spearman (1904). General intelligence objectively determined and measured. *American Journal of Psychology 15*, 201–292.

Thomson, G. H. (1953). *The Factor Analysis of Human Ability*. University of London Press, London.

Thurstone, L. (1947). *Multiple Factor Analysis*. University of Chicago Press.

Thurstone, L. L. (1935). *Vectors of the Mind*. University of Chicago Press, Chicago.

Wolpert, R. L. and S. C. Schmidler (2012). α-stable limit laws for harmonic mean estimators of marginal likelihoods. *Statistica Sinica 22*, 1233–1251.