BAYESIAN STATISTICS 8, pp. 1–45.
J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid,
D. Heckerman, A. F. M. Smith and M. West (Eds.)
© Oxford University Press, 2007

Estimating the Integrated Likelihood via Posterior Simulation Using the Harmonic Mean Identity

Adrian E. Raftery
University of Washington, USA
raftery@u.washington.edu

Michael A. Newton
University of Wisconsin, USA
newton@stat.wisc.edu

Jaya M. Satagopan
Sloan-Kettering Cancer Center, USA
satagopj@mskcc.org

Pavel N. Krivitsky
University of Washington, USA
pavel@stat.washington.edu

Summary

The integrated likelihood (also called the marginal likelihood or the normalizing constant) is a central quantity in Bayesian model selection and model averaging. It is defined as the integral over the parameter space of the likelihood times the prior density. The Bayes factor for model comparison and Bayesian testing is a ratio of integrated likelihoods, and the model weights in Bayesian model averaging are proportional to the integrated likelihoods. We consider the estimation of the integrated likelihood from posterior simulation output, aiming at a generic method that uses only the likelihoods from the posterior simulation iterations. The key is the harmonic mean identity, which says that the reciprocal of the integrated likelihood is equal to the posterior harmonic mean of the likelihood. The simplest estimator based on the identity is thus the harmonic mean of the likelihoods. While this is an unbiased and simulation-consistent estimator, its reciprocal can have infinite variance and so it is unstable in general.

We describe two methods for stabilizing the harmonic mean estimator. In the first one, the parameter space is reduced in such a way that the modified estimator involves a harmonic mean of heavier-tailed densities, thus resulting in a finite variance estimator. The resulting estimator is stable. It is also self-monitoring, since it obeys the central limit theorem, and so confidence intervals are available. We discuss general conditions under which this reduction is applicable.

Adrian E. Raftery is Blumstein-Jordan Professor of Statistics and Sociology at the University of Washington in Seattle, Michael A. Newton is a Professor of Statistics and of Biostatistics and Medical Informatics at the University of Wisconsin in Madison, Jaya M. Satagopan is Associate Attending Biostatistician at the Memorial Sloan-Kettering Cancer Center in New York City, and Pavel N. Krivitsky is Graduate Research Assistant at the University of Washington Department of Statistics in Seattle. The research of Raftery and Krivitsky was supported by NIH grant 8R01EB 002137-02.

The second method is based on the fact that the posterior distribution of the log-likelihood is approximately a gamma distribution. This leads to an estimator of the maximum achievable likelihood, and also an estimator of the effective number of parameters that is extremely simple to compute from the loglikelihoods, independent of the model parametrization, and always positive. This yields estimates of the log integrated likelihood, and posterior simulation-based analogues of the BIC and AIC model selection criteria, called BICM and AICM. We provide standard errors for these criteria. We illustrate the proposed methods through several examples.

Keywords and Phrases: AIC; AICM; Basketball; Bayes Factor; Bayesian Model Averaging; Beta-Binomial; BIC; BICM; DIC; Effective Number of Parameters; Fan-Shaped Distribution; Gamma Distribution; Genetics; Hierarchical Model; Latent Space; Marginal Likelihood; Markov Chain Monte Carlo; Maximum Likelihood; Model Comparison Model Selection; Normalizing Constant; Poisson-Gamma; Random Effects Model; Robust Linear Model; Social Networks; Unit Information Prior

1. INTRODUCTION

The integrated likelihood, also called the marginal likelihood or the normalizing constant, is an important quantity in Bayesian model comparison and testing: it is the key component of the Bayes factor (Kass and Raftery 1995; Chipman, George, and McCulloch 2001). The Bayes factor is the ratio of the integrated likelihoods for the two models being compared. When taking account of model uncertainty using Bayesian model averaging, the posterior model probability of a model is proportional to its prior probability times the integrated likelihood (Hoeting, Madigan, Raftery, and Volinsky 1999).

Consider data y, a likelihood function $\pi(y|\theta)$ from a model for y indexed by a parameter θ, in which both y and θ may be vector-valued, and a prior distribution $\pi(\theta)$. The integrated likelihood of y is then defined as

$$\pi(y) = \int \pi(y|\theta)\pi(\theta)\,d\theta.$$

The integrated likelihood is the normalizing constant for the product of the likelihood and the prior in forming the posterior density $\pi(\theta|y)$. Furthermore, as a function of y prior to data collection, $\pi(y)$ is the prior predictive density.

Evaluating the integrated likelihood can present a difficult computational problem. Newton and Raftery (1994) showed that $\pi(y)$ can be expressed as an expectation with respect to the posterior distribution of the parameter, thus motivating an estimate based on a Monte Carlo sample from the posterior. By Bayes's theorem,

$$\frac{1}{\pi(y)} = \int \frac{\pi(\theta|y)}{\pi(y|\theta)}\,d\theta = E\left\{ \left. \frac{1}{\pi(y|\theta)} \right| y \right\}. \qquad (1)$$

Equation (1) says that the integrated likelihood is the posterior harmonic mean of the likelihood, and so we call it the *harmonic mean identity.* This suggests that the integrated likelihood $\pi(y)$ can be approximated by the sample harmonic mean of the likelihoods,

$$\hat{\pi}_{\mathrm{HM}}(y) = \left[\frac{1}{B} \sum_{t=1}^{B} \frac{1}{\pi(y|\theta^t)} \right]^{-1}, \tag{2}$$

based on B draws $\theta^1, \theta^2, \ldots, \theta^B$ from the posterior distribution $\pi(\theta|y)$. This sample might come out of a standard Markov chain Monte Carlo implementation, for example. Though $\hat{\pi}_{\mathrm{HM}}(y)$ is consistent as the simulation size B increases, its precision is not guaranteed.

The simplicity of the harmonic mean estimator (2) is its main advantage over other more specialized techniques (Chib 1995; Green 1995; Meng and Wong 1996; Raftery 1996; Lewis and Raftery 1997; DiCiccio, Kass, Raftery, and Wasserman 1997; Chib and Jeliazkov 2001). It uses only within-model posterior samples and likelihood evaluations which are often available anyway as part of posterior sampling. A major drawback of the harmonic mean estimator is its computational instability. The estimator is consistent but may have infinite variance (measured by $\mathrm{Var}\big\{[\pi(y|\theta)]^{-1}|y\big\})$ across simulations, even in simple models. When this is the case, one consequence is that when the cumulative estimate of the harmonic mean estimate (2) based on the first B draws from the posterior is plotted against B, the plot has occasional very large jumps, and looks unstable.

In this article we describe two approaches to stabilizing the harmonic mean estimator. In the first method, the parameter space is reduced such that the modified estimator involves a harmonic mean of heavier-tailed densities, thus resulting in a finite variance estimator. We develop general conditions under which this method works. The resulting estimator obeys the central limit theorem, yielding confidence intervals for the integrated likelihood. In this way it is self-monitoring.

The second approach is based on the fact that the posterior distribution of the loglikelihood is approximately a shifted gamma distribution. This leads to an estimator of the maximum achievable likelihood, and also an estimator of the effective number of parameters that is very simple to compute, uses only the likelihoods from the posterior simulation, is independent of the model parametrization, and is always positive. This yields estimates of the log integrated likelihood, and posterior simulation-based analogues of the BIC and AIC model selection criteria, called BICM and AICM. Standard errors of these criteria are also provided. We illustrate the proposed methods through several examples.

In Section 2 we describe the parameter reduction method and in Section 3 we give several examples. In Section 4 we describe the shifted gamma approach and we report a small simulation study and an example. In Section 5 we discuss limitations and possible improvements of the methods described here, and we mention some of the other methods proposed in the literature.

2. STABILIZING THE HARMONIC MEAN ESTIMATOR BY PARAMETER REDUCTION

An overly simple but helpful example to illustrate our first method is the model in which $\theta = (\mu, \psi)$ records the mean and precision of a single normally distributed data point y. A conjugate prior is given by

$$\begin{aligned} \psi &\sim \mathrm{Gamma}(\alpha/2, \alpha/2) \\ (\mu|\psi) &\sim \mathrm{Normal}(\mu_0, n_0\psi), \end{aligned}$$

where α, n_0, and μ_0 are hyperparameters (e.g., Bernardo and Smith, 1994, page 268 or Appendix I). The integrated likelihood, $\pi(y)$, is readily determined to be the ordinate of a t density, $\mathrm{St}(y|\mu_0, n_0/(n_0{+}1), \alpha)$ in the notation of Bernardo and Smith (1994, page 122 or Appendix I). Were we to approximate $\pi(y)$ using equation (2), instead of taking the analytically determined value, we could measure the stability of the estimator with the variance $\mathrm{Var}\big\{[\pi(y|\theta)]^{-1}|y\big\}$. This variance, in turn, is determined by the second noncentral moment $\mathrm{E}\big\{[\pi(y|\theta)]^{-2}|y\big\}$, which is proportional to

$$\int\int \psi^{\alpha/2} \exp\left\{\frac{\psi}{2}[(y-\mu)^2 - n_0(\mu-\mu_0)^2 - \alpha]\right\} d\psi d\mu,$$

and which is infinite in this example owing to the divergence of the integral in μ for each ψ. The reciprocal of the light-tailed normal density forms too large an integrand to yield a finite posterior variance, and hence the harmonic mean estimator is unstable.

An alternative estimator, supported equally by the basic equation (1), is

$$\hat{\pi}_{\mathrm{SHM}}(y) = \left[\frac{1}{B}\sum_{t=1}^{B}\frac{1}{\pi(y|\mu^t)}\right]^{-1}, \tag{3}$$

which we call a stabilized harmonic mean. In (3), μ^t is the mean component of $\theta^t = (\mu^t, \psi^t)$, and thus is a draw from the marginal posterior distribution $\pi(\mu|y)$. The stabilized harmonic mean is formed not from standard likelihood values, but rather from marginal likelihoods obtained by integrating out the precision parameter ψ. It is straightforward to show that this integrated likelihood has the form of a t ordinate,

$$\pi(y|\mu) = \mathrm{St}\left\{y|\mu, (\alpha+1)/[\alpha + n_0(\mu-\mu_0)^2], \alpha+1\right\}.$$

The intuition motivating (3) is that since $\pi(y|\mu)$ has a heavier tail than $\pi(y|\theta)$, averages of reciprocal ordinates become averages of less variable quantities than in (2). Measuring stability as above, we observe that

$$\mathrm{E}\left\{[\pi(y|\mu)]^{-2}\big|\, y\right\} \propto \int \frac{\{1+[(y-\mu)^2 + n_0(\mu-\mu_0)^2]/\alpha\}^{\alpha/2+1}}{\{1+n_0(\mu-\mu_0)^2/\alpha\}^{\alpha+1}} d\mu \tag{4}$$

is finite when $\alpha > 1$ and $n_0 > 0$. This result is proved in Appendix II.

Figure 1 compares the harmonic mean $\hat{\pi}_{\mathrm{HM}}(y)$ to the stabilized harmonic mean $\hat{\pi}_{\mathrm{SHM}}(y)$ for various parameter settings of this simple normal example. For each case, both estimates use a common sample of $B = 5,000$ independent and identically distributed posterior draws for the mean μ and precision ψ. Shown for each sample is the value of both estimators using ever larger amounts of the sample. Figure 1 shows clearly how the infinite variance of the harmonic mean estimator manifests itself in practice. Every so often a parameter value with a very small likelihood is generated from the posterior, and this yields a very large value of the reciprocal of the likelihood, which in turn greatly reduces $\hat{\pi}_{\mathrm{HM}}(y)$. Subsequently, $\hat{\pi}_{\mathrm{HM}}(y)$ increases gradually, until another very small likelihood is encountered. Improved performance of the stabilized harmonic mean is evident in Figure 1. The t-based estimator $\hat{\pi}_{\mathrm{SHM}}(y)$ converges much more rapidly than the standard estimator, and does not exhibit the same pattern of occasional massive changes. To further validate this observation, we recomputed both final estimators on 1000 independent posterior samples of size $B = 1000$ (Figure 2). Relative stability of the $\hat{\pi}_{\mathrm{SHM}}(y)$ is clearly indicated.

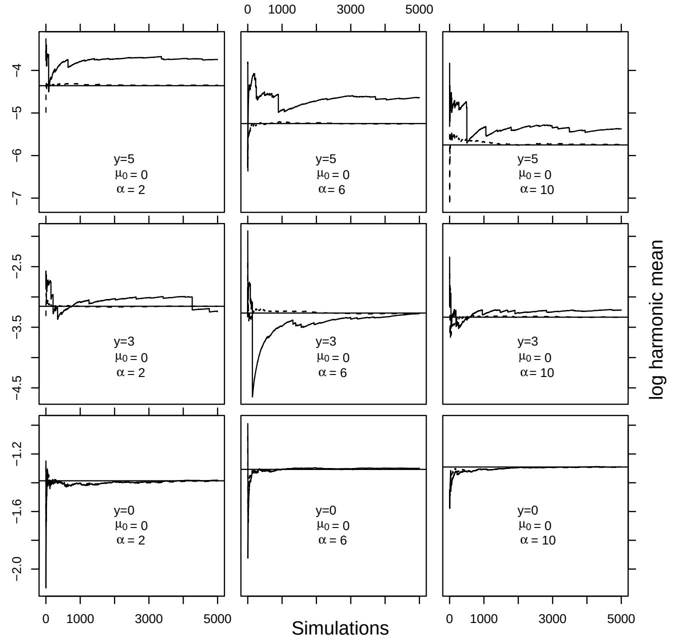

Figure 1: *Normal (bold line) and stabilized t-based (dotted line) harmonic mean estimates of the log integrated likelihood compared with the true value (dashed line), when the data y follow a univariate normal distribution as described in Section 2. The estimate based on the first B values simulated from the posterior distribution is plotted against B for one set of 5,000 values simulated from the posterior in each situation. The top row of the figure displays the harmonic mean estimates when* $y = 5$ *and* $\mu_0 = 0$. *The second row corresponds to* $y = 3$ *and* $\mu_0 = 0$. *The bottom row gives the figures for* $y = 0$ *and* $\mu_0 = 0$. *The three columns correspond to* α *values of 2, 6 and 10. The value of* n_0 *is 1. The plot shows that the normal estimate is unstable but the stabilized estimate is much more stable and converges rapidly to the correct value.*

The reciprocal estimator $\{\hat{\pi}_{\mathrm{SHM}}(y)\}^{-1}$ is a sum of quantities that have finite variance, and so it has a limiting normal distribution by the central limit theorem. This fact can be used to obtain a confidence interval for the integrated likelihood. Table 1 gives the coverage probabilities and the average length of the confidence intervals for the parameter values in Figures 1 and 2, using 1000 independent Monte Carlo samples each of size $B = 1000$. The empirical coverage probabilities are close

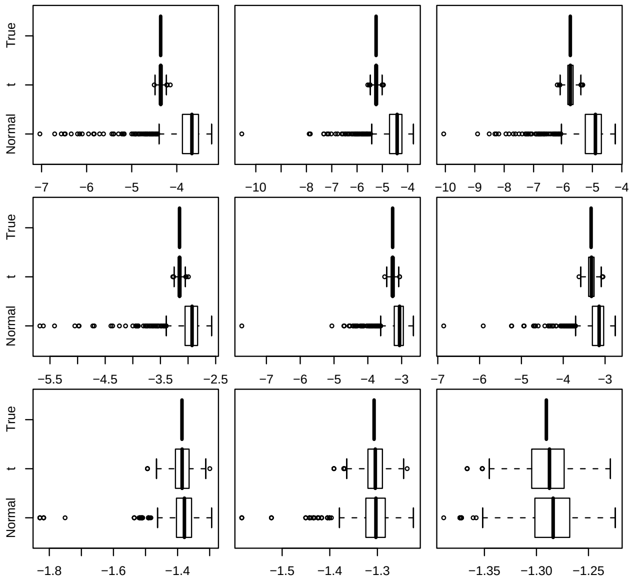

Figure 2: *Boxplots to assess the variability of the estimated integrated likelihood. Shown are the true integrated likelihood, and the normal and stabilized t-based harmonic mean estimators, both on the logarithmic scale. The estimates are obtained from 1000 Monte Carlo samples of size 1000. These estimates are shown for the same configurations of parameters as in Figure 1.*

to their nominal levels. This makes the method a self-monitoring one, in that even if the estimate it provides is imprecise, this will be made clear to the user.

The multivariate normal model is a direct extension of the univariate normal example discussed above. The standard estimator, obtained using equation (2), is a harmonic mean of multivariate normal densities. This can be easily shown to be an unstable estimator of the integrated likelihood. Integrating the precision parameter leads to a heavier tailed multivariate t density, which can be used to obtain a stable estimator analogous to equation (3).

The stabilized harmonic mean estimator was first reported in a statistical genetics application in which numerical stability of a $t-$based harmonic mean was observed (Satagopan, Yandell, Newton, and Osborn 1996). Section 3.1 presents a detailed study of this case. Although the genetical model used by these authors was rather specialized, the method to obtain a more stable estimate is quite general: approximate $\pi(y)$ by a harmonic mean of values $\pi[y|h(\theta^t)]$, where $\theta^1, \theta^2, \ldots, \theta^B$ form a sample from the posterior distribution $\pi(\theta|y)$. The function $h(\theta)$ must reduce the parameter space as much as possible, while not making the calculation of the marginal likelihood $\pi[y|h(\theta)]$ too difficult. In the examples we work out, $h(\theta)$ is of lower dimension than θ, typically obtained by integrating out one or several of the components. Taking $h(\theta)$ to be constant is an extreme case; $\pi[y|h(\theta)]$ then becomes

Table 1: *Coverage Probabilities for 50%, 80%, 90%, and 95% Confidence Intervals for the Stablilized Harmonic Mean Estimator, for the situations shown in Figures 1 and 2, for 1000 Monte Carlo samples each of size 1000. The average lengths of the confidence intervals for the reciprocal of the likelihood are shown in parentheses. Column 1 shows the parameters used in the simulation, column 2 shows the true value of* $\{\pi(y)\}^{-1}$*, and columns 3, 4, 5, and 6 give the coverage probabilities.*

(y, μ_0, α)	True $\{\pi(y)\}^{-1}$	50%	80%	90%	95%
(5, 0, 2)	78.09	0.49 (5.46)	0.79 (10.38)	0.90 (13.32)	0.94 (15.88)
(5, 0, 6)	190.19	0.50 (23.87)	0.81 (45.36)	0.90 (58.22)	0.95 (69.37)
(5, 0, 10)	314.38	0.53 (62.44)	0.78 (118.64)	0.88 (152.27)	0.93 (181.44)
(3, 0, 2)	23.44	0.49 (1.29)	0.82 (2.44)	0.90 (3.14)	0.95 (3.74)
(3, 0, 6)	26.20	0.49 (2.41)	0.78 (4.57)	0.89 (5.87)	0.93 (6.99)
(3, 0, 10)	28.05	0.48 (3.57)	0.79 (6.78)	0.88 (8.71)	0.93 (10.37)
(0, 0, 2)	4.00	0.47 (0.17)	0.79 (0.32)	0.90 (0.41)	0.93 (0.49)
(0, 0, 6)	3.70	0.48 (0.12)	0.77 (0.22)	0.87 (0.28)	0.93 (0.34)
(0, 0, 10)	3.63	0.47 (0.12)	0.81 (0.22)	0.86 (0.28)	0.93 (0.34)

the integrated likelihood $\pi(y)$. Of course, if this were computable there would be no need to calculate an approximation, and in any case, the harmonic mean estimator would have zero variance. To form harmonic means from reduced distributions is a general variance reduction technique.

Theorem 1 *If h is a measurable function of θ then*

$$\text{Var}\left\{ \frac{1}{\pi[y|h(\theta)]} \middle| y \right\} \le \text{Var}\left\{ \frac{1}{\pi[y|\theta]} \middle| y \right\}.$$

Either variance may be infinite. If the left hand side is infinite, then the right hand side is infinite also.

To avoid measure-theoretic considerations, we prove Theorem 1 only under the additional condition that $h(\theta)$ is a dimension-reducing transformation: i.e. $\theta = (\alpha, \beta)$, $h(\theta) = \alpha$, and both α and β range freely so that the prior density $\pi(\theta) = \pi(\alpha)\pi(\beta|\alpha)$ is well-defined. See Appendix III for a proof. In certain hierarchical models, where analytical integration is possible on one or two levels, it may be possible to identify useful reductions $h(\theta)$ to facilitate stable harmonic mean calculations.

Gelfand and Dey (1994) noted an extension of the basic identity (1) which justifies estimating the integrated likelihood by the harmonic mean of $\pi(y|\theta^t)\pi(\theta^t)/f(\theta^t)$ where, as before, the θ^t's are sampled from the posterior, but now $\pi(\theta)$ is the prior density and $f(\theta)$ is any (normalized) density on the parameter space. The idea is to choose f carefully so as to minimize Monte Carlo error. We show in Section 3.3 that our proposed stabilization can be combined with this technique for improved performance. Indeed there is some synergy in this combination because the proposed stabilization reduces the dimension of θ, thus making it simpler to identify a useful f function.

3. STABILIZED HARMONIC MEAN ESTIMATOR: EXAMPLES

3.1. *Statistical Genetics Example*

Linear models are used frequently in quantitative genetics to relate variation in a measured trait (phenotype) to variation in underlying genes affecting the trait (genotype). Doerge, Zeng, and Weir (1997) provide a useful review from a statistical perspective. We reconsider the particular model

$$y_i \quad = \quad \mu + \sum_{j=1}^{s} \alpha_j g_{i,j} + \epsilon_i, \qquad i = 1, \cdots, n, \tag{5}$$

used by Satagopan *et al.* (1996) to infer the genetic causes of variation in the time-to-flowering phenotype in the plant species *Brassica napus*. In (5), the i indicates different plants in a sample of size $n = 105$, the phenotypes $y = (y_i)$ are the logarithms of the times to flowering, and the decomposition on the right-hand-side characterizes the expected phenotype conditional on the genotype $g_i = (g_{i,j})$ at a set of s different genetic loci. Here ϵ_i is modeled as a mean zero normally distributed disturbance with variance σ^2 independent of genetic factors, μ is the marginal expected phenotype and α_j is the genetic effect of the jth quantitative trait locus (QTL). From the particular experimental design, each genotype $g_{i,j}$ takes one of two possible values, coded as $\{-1, 1\}$, with equal marginal probability.

The model (5) would be rather standard except that the genotypes $g = (g_i)$ are unobserved; in fact, for each i they represent the values of a random process defined over the whole genome and evaluated at s distinct positions $\lambda = (\lambda_1, \ldots, \lambda_s)$, the s putative QTLs. The number of QTLs, s, is unknown, as are their positions λ and their effects $\alpha = (\alpha_1, \ldots, \alpha_s)$. Indirect information about the QTL genotypes comes through genotype data $m = (m_i)$, obtained in this example from a panel of 10 molecular markers in the chromosomal region of interest. The statistical problem is to infer the unknown parameters $\theta = (\mu, \alpha, \lambda, \sigma^2)$ from marker and phenotype data (m, y), and considering missing genotypes g.

Satagopan *et al.* (1996) presented a Bayesian solution in which Markov chain Monte Carlo (MCMC) was used to sample the posterior distribution of all the unknowns conditional on s, the number of QTLs, separately for a range of values of s. To infer s, the integrated likelihood $\pi(y|m, s)$ was approximated for each s via a harmonic mean, and this enabled calculation of Bayes factors

$$\mathrm{BF}(s_1, s_2) = \pi(y|m, s_1)/\pi(y|m, s_2). \tag{6}$$

We reconsider this calculation in further detail. We can condition on marker information m because its marginal distribution $\pi(m)$ does not depend on any of the unknown parameters.

The prior for θ factorizes into a uniform prior over ordered loci $\lambda = (\lambda_1, \ldots, \lambda_s)$ within the chromosomal region under consideration and a conjugate prior for μ, $\alpha = (\alpha_j)$, and σ^2:

$$\begin{aligned} \pi(\mu|\sigma^2) &= \text{Normal}(\mu_0, \sigma^2/n_0), \\ \pi(\alpha_j|\sigma^2) &= \text{Normal}(\alpha_{0,j}, \sigma^2/n_{0,j}), \qquad j = 1, \cdots, s \\ \pi(\sigma^2) &= \text{Inverse Gamma}(\zeta/2, \zeta/2), \end{aligned}$$

where $\mu_0 = 5$, $n_0 = 1$, $\alpha_{0,j} = 5$, $n_{0,j} = 1$, for each j and $\zeta = 8$. Fixing the number of loci s, one complete scan of the MCMC sampler updates each element of θ and all the missing genotypes in g. See Satagopan *et al.* (1996) for further details on the component updates. A total of 3 chains, corresponding to $s = 1, 2$, and 3, were obtained. For a fixed s (= 1, 2, or 3), we report results below based on a chain of length 400,000 complete scans, subsampled every 100 scans, with the first 100 saved states removed as burn-in; diagnostics indicated that the resulting subsampled scans were close to being independent. Thus this corresponds to an effective independent sample size of about 3,900 for estimating the genetic effect parameters.

Unknowns (θ^t, g^t) are sampled from their posterior distribution conditional on observed phenotypes y, marker genotypes m and the model dimension parameter s. Invoking the standard harmonic mean argument, as in (2), we approximate $\pi(y|m, s)$ by

$$\hat{\pi}_{\text{HM}}(y|m, s) = \left[\frac{1}{B} \sum_{t=1}^{B} \frac{1}{\pi(y|m, \theta^t, g^t, s)} \right]^{-1}. \tag{7}$$

As in the simple normal example of Section 2, a problem arises with (7) because we are averaging reciprocals of normal ordinates. To stabilize the estimator, we integrate out the variance parameter σ^2 and obtain

$$\hat{\pi}_{\text{SHM}}(y|m, s) = \left[\frac{1}{B} \sum_{t=1}^{B} \frac{1}{\pi(y|m, h(\theta^t), g^t, s)} \right]^{-1}, \tag{8}$$

where $h()$ returns all components of θ except the variance parameter. In (8), $\pi(y|m, h(\theta^t), g^t, s)$ is a scaled t density, $\text{St}_n(y|\mu + \alpha^{'} g, I, \zeta)$.

Figure 3 shows the cumulative Bayes factor estimates obtained from three chains, ($s = 1, 2$, and 3), based on integrated likelihood estimates in either (7) and (8). Evidently the stabilization has worked in this more complicated example: there are fewer massive changes in the estimate. Numerically, we obtain $BF(1, 2) = 0.368$ using (7), and $BF(1, 2) = 0.395$ using the stabilized estimator (8). The estimates of $BF(2, 3)$ are rather more disparate: 13.15 and and 4.39, respectively. In any case we would conclude that the two-locus model is most likely *a posteriori*.

Figure 4 indicates the Monte Carlo sampling variability of the two estimators. The above computations were replicated 75 times. To reduce the computational burden of the simulation, we used a value of B equal to half of the earlier value.

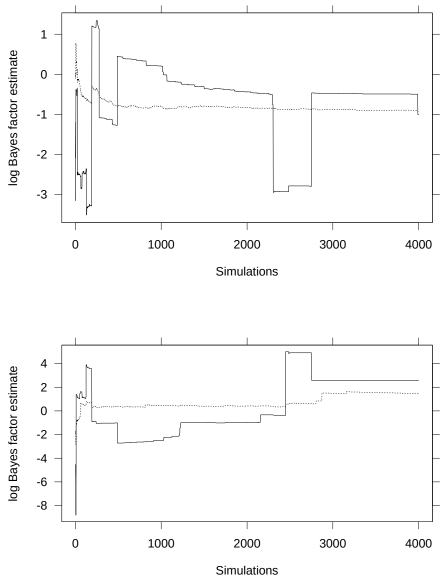

Figure 3: *Log Bayes Factor Estimates for the Flowering Time Data, based on MCMC. The log Bayes factor based on the first B saved scans of the MCMC run is plotted against B. The comparison between the one-locus and two-loci models is shown on the top. The bottom figure corresponds to the comparison between the two-loci and three-loci models. The bold line is the standard harmonic mean estimate of the log Bayes factor, and the dotted line is the stabilized t-based estimate. The plot shows that the stabilized estimate is much more stable than the standard one.*

The side-by-side boxplots further confirm the success of the stabilization in the present example.

We note that other dimension-reducing transformations $h(\cdot)$ could be used in this example. For example, we could sum out the genotype values g and thus average reciprocals of finite mixtures of normals (or t's). It may also be possible to integrate out the genetic effects α. Neither of these has been attempted here.

3.2. *Beta–Binomial Example*

A naturally occurring hierarchical model has observable counts $y = (y_i)$, $i = 1, \ldots, m$, arising as conditionally independent binomial random variables with numbers of trials (n_i) and success probabilities $p = (p_i)$. In turn, the p_i's are modeled as conditionally independent beta variables with canonical hyperparameters, a and b say, upon which some further prior distribution $\pi(a, b)$ is placed. To obtain the probability of y in this model, we must integrate out both the p_i's and the hyperparameters a and b. It is routine to sample the full parameter set $\theta = (p, a, b)$ from its

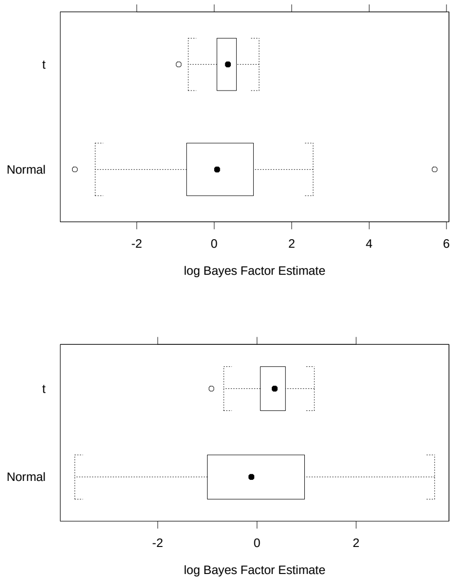

Figure 4: *Assessing the Variability of the Log Bayes Factor Estimates for the Flowering Time Data, using 75 replications of the MCMC run. The top panel shows the comparison between the one-locus and two-loci models, and the bottom panel shows the comparison between the two-loci and three-loci models. In each panel, the variability among the stabilized t-based estimates is shown on top, and that among the standard normal estimates is shown below.*

posterior distribution (Gelman, Carlin, Stern and Rubin, 2003). For example, an MCMC simulation might update each p_i from its Beta full-conditional distribution, and then resort, perhaps, to a random-walk proposal to update a and b.

The basic harmonic mean combines reciprocals of binomial likelihoods from the posterior sample, and, it turns out, can be quite unstable. As before, stability is determined by the second noncentral moment

$$E\left\{[\pi(y|\theta)]^{-2}|y\right\} \propto \int\int\prod_i \left\{\int p^{a-1-y_i}(1-p)^{b-1-n_i+y_i}\,dp\right\}\pi(a,b)\,da\,db.$$

Unless we take an extreme prior $\pi(a,b)$ which ensures that $a > \max(y_i)$ and $b > \max(n_i - y_i)$, this integral can diverge. Typically, a prior extreme enough to avoid this divergence would be unrealistically peaked. This is unsatisfactory, ruling out the standard (unstabilized) harmonic mean estimator as a practical tool for the beta-binomial model.

It is straightforward to stabilize the harmonic mean by reducing the dimension of θ as in previous examples. One possibility is to take $h(\theta) = (a,b)$; i.e. to integrate out all the binomial success probabilities. In this conjugate structure, we have a closed form beta-binomial expression for $\pi\{y|h(\theta)\}$, namely

$$\pi\{y|h(\theta)\} = \prod_i \frac{\Gamma(n_i+1)}{\Gamma(n_i-y_i+1)\Gamma(y_i+1)} \frac{\Gamma(a+b)}{\Gamma(a+b+n_i)} \frac{\Gamma(a+y_i)}{\Gamma(a)} \frac{\Gamma(b+n_i-y_i)}{\Gamma(b)}. \quad (9)$$

The harmonic mean of these beta-binomial probabilities, calculated from the (a, b)'s sampled from their posterior, is consistent for the integrated likelihood. We may expect this to be more stable since the beta-binomial distribution is more diffuse than the binomial, and so the reciprocals of the probabilities may not be as extreme. The stability of this estimator is determined by the second noncentral moment, which satisfies

$$E\left\{[\pi(y|a,b)]^{-2}|y\right\} \quad \leq \quad \int (a+b+n_{\max}-1)^m \pi(a,b)\ da\ db,$$

where $n_{\max} = \max n_i$. Stability is ensured when prior moments of a and b exist.

Data on free-throw percentages from the National Basketball Association (NBA) provide an interesting demonstration of the harmonic mean calculations. On March 9, 1999, there were 414 active NBA players of whom 374 had attempted at least one free throw by that point in the season. Among these 374 players, the numbers of attempts (n_i) ranged from 1 to 205, with a mean of about 35. We model y_i, the number of made free throws by player i, as Binomial with n_i trials and unknown success probability p_i. The average free throw percentage y_i/n_i is about 70% in the data reported at `www.yahoo.com` (and available from the authors).

We consider the problem of evaluating the integrated likelihood $\pi(y)$ under the hierarchical beta–binomial model given above. This would be useful when comparing this model with other hypothesized models for these data. We place independent standard exponential priors on $a - \epsilon$ and $b - \epsilon$ where $\epsilon = 1$ is a lower truncation point of the prior. MCMC was used to simulate the posterior. The following numerical results are based on a single chain of length 2.5 million complete scans, subsampled every 50 scans, and with the first 100 saved states removed as burn-in. Significant trends were not detected in the output and standard MCMC diagnostics indicated that little dependence remained in the saved states. Computations were done separately on a second run and we saw no appreciable differences.

We calculated natural logarithms of the product binomial likelihood and the product beta-binomial likelihood (9). From these values we obtained the standard harmonic mean estimate and the stabilized one. The log estimates were –817.0 and –942.9 respectively; these are quite different. The standard estimate is known to be unstable. Indeed the variance of the sampled loglikelihood values was 146.3 while that of the sampled log beta-binomial values was only 4.1. Variance on the log scale does not tell the whole story because we are averaging on the anti-log scale; it is outliers (having very low likelihood) that are particularly influential, but still variance gives some indication.

Suspecting that some additional improvements could be made, we combined the stabilization technique with the method of Gelfand and Dey (1994) discussed at the end of Section 2, using a Gaussian approximation to the posterior $\pi(a, b|y)$ as the density f. The estimate becomes a harmonic mean of the values $\pi(y|a, b)\pi(a, b)/f(a, b)$, with (a, b)'s sampled from their posterior. The main advantage of this adjustment is that now the influence of individual sample points is greatly diminished.

The estimated log integrated likelihood is -951.4, which matches a brute force grid-based numerical integration of $\pi(y|a,b)\pi(a,b)$ almost exactly. Thus we see that the initial stabilization method worked fairly well and was easily improved.

3.3. *Other Reductions: A Simple Poisson-Gamma Model*

Sometimes useful reductions are hard to find, and the natural approach we have considered of integrating out a parameter does not work. A simple example is when y has a Poisson distribution with mean $\gamma\lambda$, and γ is exponentially distributed with mean 1 and independent of λ *a priori.* The standard harmonic mean estimator of $\pi(y)$ uses samples $\theta^i = (\lambda^i, \gamma^i)$ from $\pi(\theta|y)$, and averages the reciprocals of Poisson probabilities. Stability depends on the second noncentral moment

$$\mathrm{E}\left\{ [\pi(y|\theta)]^{-2} \middle| \, y\right\} \quad \propto \quad \int\int \frac{1}{(\gamma\lambda)^y} \exp\{-\gamma(1-\lambda)\}\pi(\lambda)\, d\gamma d\lambda.$$

Note that the inner integral diverges for any $\lambda > 1$, so that the standard harmonic mean is unstable. The natural reduction would be to take $h(\theta) = \lambda$. Thus the marginal likelihood $\pi[y|h(\theta)] = \pi(y|\lambda)$ is a geometric distribution $\lambda^y/(1+\lambda)^{(y+1)}$. Stability here hinges upon

$$\mathrm{E}\left\{ [\pi(y|\lambda)]^{-2} \middle| \, y\right\} \quad \propto \quad \int \left(\frac{1+\lambda}{\lambda}\right)^y (1+\lambda)\, \pi(\lambda)\, d\lambda.$$

For small λ, the dominant term of the integrand is $\pi(\lambda)/\lambda^y$, and so stability of the modified harmonic mean depends on the prior, though for a standard Gamma prior, for example, this integral can diverge. In other words, both variances in Theorem 1 equal infinity. Thus integrating out γ does not produce a stabilized harmonic mean estimator in this case.

Another, further reduction does work, however. Consider the case where λ, like γ, has a prior exponential distribution with mean 1. Suppose that $h(\theta) = 0$ if $\lambda \leq \epsilon$, and $h(\theta) = \lambda$ if $\lambda > \epsilon$, where ϵ is a small predetermined constant. Then $\pi[y|h(\theta) = 0] \approx \epsilon^{y+1}/(y+1)$ (better approximations are readily available if necessary), and it is easily shown that $E\{\pi[y|h(\theta)]^{-2}|y\} < \infty$. Thus, with this refinement, the modified harmonic mean estimator is stable.

4. SHIFTED GAMMA ESTIMATOR OF THE INTEGRATED LIKELIHOOD

4.1. *Shifted Gamma Estimator*

We now consider a different approach to stabilizing the harmonic mean estimate. If MCMC is used to simulate from the posterior, we suppose that the the output has been thinned in such as way that we have an approximately independent sequence of loglikelihoods $\{\ell_t : t = 1, \ldots, B\}$.

We use the fact that asymptotically (as the amount of data underlying the likelihoods increases to infinity, not the number of samples from the posterior), the posterior distribution of the loglikelihoods is given by

$$\ell_{\max} - \ell_t \sim \mathrm{Gamma}(\alpha, 1), \tag{10}$$

where $\ell_{\max}$ is the maximum achievable loglikelihood, and $\alpha = d/2$ where d is the dimension of the parameter θ, i.e. the number of parameters in the underlying model

(Bickel and Ghosh 1990; Dawid 1991). In (10), a Gamma(α, λ^{-1}) distribution with shape parameter α and scale parameter λ has the density

$$f_X(x) = \frac{x^{\alpha-1}\exp(-x/\lambda)}{\Gamma(\alpha)\lambda^{\alpha}}. \tag{11}$$

With this definition, $E(X) = \alpha\lambda$, and $\mathrm{Var}(X) = \alpha\lambda^2$. This can also be viewed as a scaled χ^2 distribution with $d = 2\alpha$ degrees of freedom. Fan, Hung, and Wong (2000) showed that (10) holds under more general conditions than the usual Wald-type conditions required for the likelihood ratio test statistic to be asymptotically χ^2.

In principle, we could use the asymptotic approximation (10) directly to approximate the posterior harmonic mean and hence the integrated likelihood. There are three main difficulties with this, however. First, in general we will not know $\ell_{\max}$ from a posterior sample, because the maximum likelihood will typically not be reached. In practice, the difference between $\ell_{\max}$ and the maximum observed loglikelihood in the MCMC sample can be quite large when the number of parameters is big. Second, in general, we will not know the effective number of parameters, d, especially in hierarchical and other random effects models of the kind often estimated using MCMC. Third, with the posterior distribution (10) of the loglikelihoods, the posterior harmonic mean, and hence the integrated likelihood, is infinite.

The first two difficulties can be resolved by noting that simple moment estimators of $\ell_{\max}$ and α are available. Under the assumption (10), $E[\ell_{\max} - \ell_t] = \alpha$ and $\mathrm{Var}(\ell_t) = \alpha$. Replacing the expectation and variance of ℓ_t by their sample equivalents and solving, we thus get the moment estimators $\hat{\alpha} = s_\ell^2$ and $\hat{\ell}_{\max} = \bar{\ell} + s_\ell^2$, where $\bar{\ell}$ and s_ℓ^2 are the sample mean and variance of the ℓ_t's.

It is clear that $\ell_{\max}$ is at least as big as the largest observed loglikelihood, $\max_t \ell_t$. Thus we could refine the moment estimator of $\ell_{\max}$ to take account of this, as $\hat{\ell}^*_{\max} = \max\{\hat{\ell}_{\max}, \max_t \ell_t\}$, or $\hat{\ell}^{**}_{\max} = \max\{\hat{\ell}_{\max}, \max_t \ell_t + \delta\}$, where δ is some small positive number that is small on the typical scale of loglikelihoods, such as 0.01. We have found, however, that it rarely happens that $\hat{\ell}_{\max}$ is smaller than $\max_t \ell_t$, and that even when it does, the difference is very small. Thus we have not found this refinement of much use in practice.

The third difficulty implies that the approximation (10) is not accurate enough for any actual data that would be encountered. One possibility is to modify it by allowing a scale parameter that is not exactly equal to 1, so that the approximate posterior distribution becomes

$$\ell_{\max} - \ell_t \sim \mathrm{Gamma}(\alpha, \lambda^{-1}), \tag{12}$$

where $\lambda < 1$. In practice, λ will be less than 1, but not much less than 1.

Given the approximation (12), we can find the integrated likelihood using the fact that if $X \sim \mathrm{Gamma}(\alpha, \lambda^{-1})$, then the moment generating function of X is

$$m_X(t) = E[e^{tX}] = (1 - \lambda t)^{-\alpha}. \tag{13}$$

Combining the harmonic mean identity (1) with equations (12) and (13), we see that the integrated likelihood is given by

$$\log \pi(y) = \log E[e^{-\ell_t}|y] = \ell_{\max} + \alpha \log(1 - \lambda). \tag{14}$$

This has an interesting similarity to the BIC approximation to the log integrated likelihood,

$$\log \hat{\pi}_{\mathrm{BIC}}(y) = \ell(\hat{\theta}) - \frac{d}{2}\log(n), \tag{15}$$

where $\hat{\theta}$ is the maximum likelihood estimator, so that $\ell(\hat{\theta}) = \ell_{\max}$, the maximum achievable loglikelihood. In general, under regularity conditions,

$$\log \pi(y) = \log \hat{\pi}_{\mathrm{BIC}}(y) + O_P(1), \tag{16}$$

(Schwarz 1978), so that the relative error in $\log \hat{\pi}_{\mathrm{BIC}}(y)$ tends to zero asymptotically. If the prior $\pi(\theta)$ is a normal unit information prior, then the approximation is more accurate and the $O_P(1)$ term in (16) is replaced by $O_P(n^{-1/2})$ (Kass and Wasserman 1995, Raftery 1995). We have that $\alpha = d/2$, and so $-\log(1-\lambda)$ in (14) corresponds to $\log(n)$ in (15).

We already have estimates of $\ell_{\max}$ and α in (14), and so to obtain an estimate of the integrated likelihood it remains only to estimate λ. Unfortunately this is difficult, because λ is typically close to 1, and so the value of $\pi(y)$ is sensitive to its precise value. On the other hand, the loglikelihoods $\{\ell_t\}$ typically do not allow us to distinguish well between values of λ close to 1. We have experimented with Bayesian and other estimators of λ, but so far the estimates we have tried have not been very accurate. This is a topic of ongoing research.

In the meantime we suggest a posterior simulation-based version of BIC. BIC is defined by

$$\mathrm{BIC} = 2\ell(\hat{\theta}) - d\log(n), \tag{17}$$

and by analogy we define

$$\mathrm{BICM} = 2\hat{\ell}_{\max} - \hat{d}\log(n), \tag{18}$$

where BICM stands for BIC–Monte (Carlo). This yields the following approximation to the log integrated likelihood:

$$\begin{aligned} \log \hat{\pi}_{\mathrm{BICM}}(y) &= \hat{\ell}_{\max} - \frac{\hat{d}}{2}\log(n) & (19)\\ &= \bar{\ell} - s_\ell^2\left(\log(n) - 1\right). & (20) \end{aligned}$$

One difficulty with this criterion is that the sample size n is not always well-defined, particularly in the kind of models commonly estimated by MCMC. Volinsky and Raftery (2000) showed that in another context, when different choices are possible, they each give valid approximations to the integrated likelihood, corresponding to different unit information priors, that differed in the definition of a "unit". Thus a reasonable choice may follow by considering what a reasonable definition of a "unit" is. Volinsky and Raftery (2000) gave an example of one way of determining this.

The definition (17) of BIC, on which BICM as defined by (18) is based, is adequate for fixed effects models, but not for hierarchical or random effects models, as shown for example by Pauler (1998) and Berger, Ghosh, and Mukhopadhyay (2003). Pauler (1998) in her equation (11) proposed a modified definition of BIC for hierarchical models, called S_M, and showed its validity in her Theorem 2. In this approach each parameter potentially has a different "n" associated with it, corresponding to the effective sample size, or order of information involved in estimating it, and the definition of BIC becomes

$$\mathrm{BIC} = 2\ell(\hat{\theta}) - \sum_{k=1}^{K} \log(n_k), \tag{21}$$

where n_k is the effective sample size involved in the estimation of the k-th parameter, θ_k. We consider a slight modification of this, namely

$$\mathrm{BIC} = 2\ell(\hat{\theta}) - \sum_{k=1}^{K} \log(n_k + 1). \tag{22}$$

This is asymptotically equivalent to (21), but unlike (21) it assigns a nonzero, although small penalty even when $n_k = 1$. It also remains defined even when $n_k = 0$, i.e. when there are no data relevant to that parameter, assigning no penalty in that case, which seems appropriate. In general, determining n_k involves assessing the Fisher or observed information for θ_k (Pauler 1998), but we will take as a rough approximation the number of data points that participate in the estimation of θ_k.

This leads to a modified definition of BICM. Parameters are divided into classes according to the number of data points that participate in the estimation of each one, and are ordered according to the value of n_k. The random effects will be last, and will be assigned an effective number of parameters equal to $\hat{d} - K'$, where K' is the number of parameters already accounted for. Thus we have

$$\mathrm{BICM} = 2\hat{\ell}_{\max} - \sum_{k=1}^{K'} \log(n_k + 1) - (\hat{d} - K')\log(n_{K'+1}), \tag{23}$$

An example of the use of equation (23) is given in Section 4.3.

In a similar way, we can write down a posterior simulation-based version of AIC (Akaike 1973). AIC can be defined as

$$\mathrm{AIC} = 2\ell_{\max} - 2d, \tag{24}$$

which we can estimate by

$$\begin{aligned} \mathrm{AICM} &= 2\hat{\ell}_{\max} - 2\hat{d} && (25)\\ &= 2\hat{\ell}_{\max} - 4s_\ell^2 && (26)\\ &= 2(\bar{\ell} - s_\ell^2). && (27) \end{aligned}$$

Thus AICM is seen to be a very simply computed penalized version of the posterior mean of the loglikelihoods, using only the loglikelihoods from the posterior simulation. There is a substantial literature on the relative merits of AIC and BIC, and many of the same arguments could probably be made about AICM and BICM. Our derivation of BICM is as an approximation to the log integrated likelihood, but AICM does not have such an interpretation.

We can obtain standard errors of BICM and AICM using the facts that $\mathrm{Var}(\bar{\ell}) \approx d/(2B)$ and

$$\mathrm{Var}(s_\ell^2) \approx d(11d/4 + 12)/B,$$

together with the approximate posterior independence of $\bar{\ell}$ and s_ℓ^2. The criteria BICM (in both the forms we have given) and AICM are both of the form $a\bar{\ell} - bs_\ell^2$. The standard error of a criterion of this kind is thus

$$\sqrt{a^2\hat{d}/(2B) + b^2\hat{d}(11\hat{d}/4 + 12)/B}.$$

Note that this standard error takes account only of the Monte Carlo variation, i.e. it is an estimate of the standard deviation of the criterion over repeated posterior simulation runs of the same length. It does not take account of error in the approximation to the log integrated likelihood or of sampling variation in the data themselves. It is a standard error of BICM, when BICM is viewed as an estimator of the BICM value that would be obtained asymptotically if the number of draws from the posterior grew without bound; similarly for AICM. Note also that it depends crucially on the assumption that the posterior simulation draws are approximately independent, so that if MCMC is used, the posterior sample would have to be thinned enough for this to be the case.

As we have noted, the moment estimator of α implies that $\hat{d} = 2s_\ell^2$ can be viewed as an estimator of the effective number of parameters. Spiegelhalter, Best, Carlin, and van der Linde (2002) proposed a different estimator of the effective number of parameters from posterior simulation, $p_D = 2(\log \pi(\bar{\theta}|y) - \bar{\ell})$, where $\bar{\theta}$ is the mean of the values of θ simulated from the posterior. In our limited experience, we have found that p_D and $\hat{d}$ are similar and that both work well in situations where the number of parameters is known.

However, Spiegelhalter *et al.* (2002) have pointed out that p_D is not invariant to the model's parameterisation because it involves the posterior mean of the parameters, $\bar{\theta}$, and that this noninvariance can be consequential. They also pointed out that p_D can be negative. In addition, p_D may not be well defined in situations where the meaning of $\bar{\theta}$ is not clear, such as multinomial parameters, or finite mixture models where the unobserved group memberships are included in the MCMC scheme (Diebolt and Robert 1994). A similar problem arises when there is near posterior nonidentifiability such as label-switching in mixture models or random effects without identifying constraints (Celeux, Hurn, and Robert 2000; Stephens 2000). One way around this is to use a posterior mode of θ instead of $\bar{\theta}$, but Richardson (2002) gave several examples of mixture models where p_D with this definition inadequately penalizes model complexity. The estimator $\hat{d}$ is defined simply and unambiguously in all those cases.

The estimator $\hat{d}$ was also derived by Gelman *et al.* (2003, Section 6.7), and used by them instead of p_D in their alternative definition of DIC, the measure of fit originally defined by Spiegelhalter *et al.* (2002). AICM is equivalent to Gelman *et al.* (2003)'s definition of DIC. Gelman *et al.* (2003) used this alternative definition because the deviance function was not available to their MCMC program and so their program could not compute p_D routinely, whereas it could compute $\hat{d}$; see also Sturtz, Ligges and Gelman (2005).

An interesting observation follows from the results of Fan *et al.* (2000). They consider the situation where, roughly speaking, the level-w contour of the likelihood function has the form $\hat{\theta} + a_n w^r S$, where $\hat{\theta}$ is the maximum likelihood estimator, $r > 0$ is a constant, $a_n \to 0$ is a sequence, and S is a surface in R^d. The standard situation where the likelihood contours are elliptical has $r = 1/2$, $a_n = O(n^{-1/2})$, and $S = \{\theta : \theta^T \Sigma \theta\}$ where Σ is the Fisher information matrix, so that S is an ellipse.

When the contours are not elliptical, they say that the distribution is "fan-shaped." They show that in general under these conditions

$$\ell_{\max} - \ell_t \sim \text{Gamma}(rd, 1). \tag{28}$$

In the standard, elliptical situation with $r = \frac{1}{2}$, this reduces to (10) as before.

They give several simple examples within their class where the likelihood contours are not elliptical. One is inference about the minimum of a shifted exponential distribution whose scale parameter is known. In that case they show that $r = 1$. Thus the "effective number of parameters" in that case is 2, even though there is only one actual parameter. This illustrates the fact that the term "effective number of parameters" is really just a figure of speech. It suggests that what is important for estimating the integrated likelihood is the shape parameter of the approximating gamma distribution, not a literal count of the parameters in the model. The arguments above suggest that the former may continue to be well approximated by $2s_\ell^2$ even when this does not coincide with a simple count of the number of parameters.

Finally, we note that when the number of parameters (not necessarily data points) becomes large, the shifted gamma approximation to the posterior distribution of the loglikelioods (12) becomes approximately normal. The posterior distribution of the reciprocal of the likelihood is then approximately lognormal, leading to the estimator

$$\log \hat{\pi}_{\text{LN}}(y) = \bar{\ell} - \tfrac{1}{2}s_\ell^2. \tag{29}$$

This was proposed by Pritchard, Stephens, and Donnelly (2000), who also noted that a better approximation might be available by using a gamma distribution for the loglikelihoods, thus prefiguring the present work, although they did not develop their observation further. Pritchard *et al.* (2000) proposed and used $\log \hat{\pi}_{\text{LN}}(y)$ as a model choice criterion rather than an estimator of the log integrated likelihood. It is interesting to note that

$$\log \hat{\pi}_{\text{LN}}(y) = \hat{\ell}_{\max} - \tfrac{3}{4}\hat{d},$$

so that $\log \hat{\pi}_{\text{LN}}(y)$ is a penalized version of the estimated maximum loglikelihood, with a penalty similar to but smaller than that of AICM, equal to $\frac{3}{4}\hat{d}$ rather than $\hat{d}$ as for AICM.

4.2. *Multivariate Normal Simulation Experiment*

In order to assess the estimators $\hat{d}$, $\hat{\ell}_{\max}$ and $\pi_{\text{BICM}}(y)$, we first carried out a small simulation study using a canonical multivariate normal situation. The data $y_1, \ldots, y_n$ are independent and identically distributed $\text{MVN}_d(\mu, I)$ random vectors, and the prior for μ is $\mu \sim \text{MVN}_d(0, I)$. The sufficient statistic is then just the d-dimensional $\bar{y} \sim \text{MVN}_d(\mu, I/n)$. We simulated values of μ from its posterior distribution $\mu|y \sim \text{MVN}_d(n\bar{y}/(n+1), I/(n+1))$. The loglikelihoods are then given by

$$\ell_t = \log p(\bar{y}|\mu^t) = \frac{d}{2}\log(n/2\pi) - \frac{n}{2}\sum_{j=1}^{d}(\bar{y}_j - \mu_j)^2.$$

The true maximum likelihood is $\frac{d}{2}\log(n/2\pi)$ and the true log integrated likelihood is

$$\pi(y) = \frac{d}{2}\log\left(\frac{n}{(n+1)2\pi}\right) - \frac{1}{2}\frac{n}{n+1}\sum_{j=1}^{d}\bar{y}_j^2.$$

Our goal was to see how the method worked under a wide range of values of d and n, so we fixed μ at $(0.15, \ldots, 0.15)$. We simulated values of the number of parameters d from a discrete uniform distribution on the integers from 1 to 100, and we simulated values of the sample size n from a discretized log-uniform distribution with $\log(n) \sim U[3, 9]$, so that approximately, n ranged from 20 to 8,000, with a median of 400, subject to the constraint that $d < n$. Thus the simulation encompassed standard situations with a small number of parameters and a large sample size, and also situations where there were almost as many parameters as data points, ranging up to moderately large numbers of parameters (100). For each pair of values of d and n sampled, a dataset consisting of $\bar{y}$ was drawn, and then the posterior distribution was simulated. Altogether, 1000 datasets were simulated, and for each dataset a sample of size 100,000 was drawn from the posterior. This is a standard fixed effects model and so for BICM we used the definition (18).

The results are shown in Figure 5. The upper left panel shows the histogram of loglikelihoods for one dataset with $d = 10$ and $n = 100$, together with the fitted gamma distribution superimposed. The fit is extremely good, and this was the case for all the datasets that we examined. The upper right panel shows the estimated maximum achievable loglikelihood plotted against the true maximum likelihood for the 1000 simulated datasets. The estimation was good, even in cases with larger number of parameters, where the largest loglikelihood among those sampled, $\max_t \ell_t$, was much smaller than the true maximum loglikelihood. The lower left panel shows the estimated number of parameters plotted against the true number; again the estimation was very good. Finally, the lower right panel shows the approximated and true log integrated likelihoods; again the estimation was good.

In the simulated situation, the prior used was a unit information prior, so it is of interest to see what happens if a different prior is used. We experimented with situations where the prior was $\mu \sim \text{MVN}_d(0, \sigma^2)$ where $\sigma^2 \neq 1$. Note that the unit information prior corresponds to $\sigma^2 = 1$. The good results for $\hat{d}$ and $\hat{\ell}_{\max}$ remained unchanged. As long as σ^2 was larger than about 0.2, i.e. as long as the prior was not highly informative, the value of $\log \hat{\pi}_{\text{BICM}}(y)$ remained very highly correlated with the true value of $\log \pi(y)$. The slope of the line in the lower right panel of Figure 5 was no longer unity, but the fact that the correlation remained very high means that model comparisons based on the estimated log integrated likelihoods would remain accurate. A more accurate approximation to the absolute value of $\pi(y)$ could be obtained by replacing $\log(n)$ by $\log(\sigma^2 n)$ in the expression (19) for $\hat{\pi}_{\text{BICM}}(y)$. However, this would be a model-specific adjustment and would take us beyond the generic estimates that we are aiming for here.

4.3. *Example: Latent Space Models for Social Networks*

Social network data consist of observations on relations between actors, for example whether one individual says she likes another. Often such data are binary, in which a directed or undirected relation between actor i and actor j either exists or does not. In this case, the data consist of values of y_{ij} for $i, j = 1, \ldots, n$, where i and j index the n actors, and $y_{ij} = 1$ if the relation from i to j exists and $y_{ij} = 0$ if it does not.

Hoff, Raftery, and Handcock (2002) introduced the latent position model for data such as these. In this model, each actor i is assumed to be associated with an observed or latent position in an unobserved q-dimensional Euclidean "social space", denoted by z_i. Then the model says that the y_{ij} are conditionally independent given

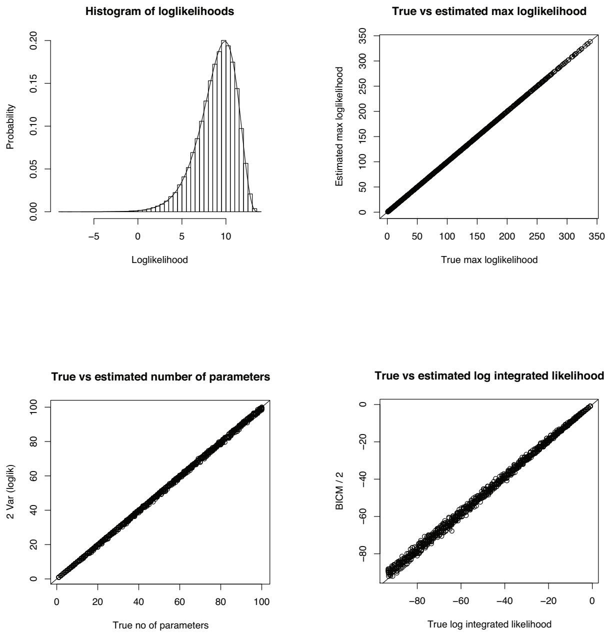

Figure 5: *Multivariate Normal Simulation Study of the Shifted Gamma Estimator. Upper left: Histogram of the loglikelihoods for one dataset with* $d = 10$ *parameters and* $n = 100$ *data points, with the fitted gamma density superimposed. Upper right: The estimated maximum achievable loglikelihood,* $\hat{\ell}_{\max}$*, plotted against the true maximum loglikelihood for the 1000 simulated datasets. Lower left: The estimated number of parameters,* $\hat{d}$*, plotted against the true number of parameters for the 1000 datasets. Lower right: The estimated log integrated likelihood,* $\log \hat{\pi}_{\mathrm{BICM}}(y)$*, plotted against the true log integrated likelihood for the 1000 simulated datasets. In the last three plots, the solid line is the* $y = x$ *or identity line.*

the latent positions, with

$$\log\left(\frac{\Pr(y_{ij}=1)}{\Pr(y_{ij}=0)}\right) = \beta - |z_i - z_j|, \tag{30}$$

$$z_i \overset{\mathrm{iid}}{\sim} \mathrm{MVN}_q(0, \sigma^2 I). \tag{31}$$

There are just two parameters for which priors are needed, β and σ^2, and we use the priors $\beta \sim N(0, 10^2)$ and $\sigma^2 \sim \sqrt{10}\ \mathrm{Inverse}\,\chi^2_3$. These priors are proper but reasonably spread out. Estimation is carried out by MCMC on β, σ^2 and the z_i's.

Here we consider a well-known dataset on the relations among 18 monks collected by Sampson (1968). Each monk was asked with which other monks he had positive relations. Based on extensive analyses of these and much other data, the 18 monks have traditionally been classified into three groups: the Loyal Opposition, the Young Turks, and the Outcasts. Hoff *et al.* (2002) analyzed a subset of these data, and the fuller dataset we analyze here was previously analyzed by Handcock, Raftery, and Tantrum (2005).

Interest focuses here on the choice of dimension, and MCMC estimation is carried out for each dimension $q = 1, 2, 3, 4$. In computing BICM, the issue of how to define the penalty term arises. This is a random effects model, and so we use the form (23). There are three "groups" of parameters: β (one parameter) with associated effective sample size n_1, σ^2 (one parameter) with associated effective sample size n_2, and the latent positions z_i, which are random effects (all remaining parameters), with associated effective sample size n_3.

The number of actors in the data is 18 so that the number of potential links is $\binom{18}{2} = 306$, and the number of actual links is 88. The parameter β is estimated from a logistic regression with 306 cases and 88 "successes", and we take the associated effective sample size to be $n_1 = 88$, following arguments analogous to those of Volinsky and Raftery (2000). The parameter σ^2 is estimated from data on the 18 actors, and so we take the associated effective sample size to be $n_2 = 18$. Finally, most of the information about an actor's latent position z_i comes from the links to and from that actor. There are an average of $88/18 = 4.9$ links per actor, and so we take the effective sample size associated with the random effects to be $n_3 = 4.9$. BICM is then defined as

$$\text{BICM} = \hat{\ell}_{\max} - \log(n_1 + 1) - \log(n_2 + 1) - (\hat{d} - 2)\log(n_3 + 1). \tag{32}$$

Note that the values of n_1 and n_2 chosen do not affect model selection, because the corresponding terms cancel in computing differences between BICM values for different models, which are what matter for model comparisons.

The results are shown in Table 2. In addition to our estimates of the maximized likelihood, estimates of the maximized loglikelihood by numerical optimization are shown. These agree reasonably closely with our estimates. Also, the number of parameters involved in the MCMC simulation is shown, and this corresponds fairly well with $\hat{d}$, the estimated number of parameters. There is no reason to expect the effective number of parameters to be the same as the number of parameters over which the MCMC algorithm iterates in this kind of hierarchical latent variable model, but in this case they do line up rather well.

Table 2: *Comparing Dimensions in the Latent Space Social Network Model. q is the dimension of the latent space, $\ell_{\max}$ is the maximized loglikelihood from a numerical optimisation routine, and # par is the total number of parameters estimated, including the latent position coordinates. The best values of BICM and AICM are shown in bold.*

q	$\hat{\ell}_{\max}$	$\ell_{\max}$	$\hat{d}$	# par	$\log \hat{\pi}_{\text{BICM}}(y)$	SE	$\frac{1}{2}$AICM	SE
1	–128.6	–129.1	20.4	20	–148.6	0.3	–149.0	0.4
2	–109.6	–110.3	38.0	38	**–145.3**	0.9	**–147.6**	0.7
3	–87.8	–89.9	66.1	56	–148.4	1.0	–154.0	1.1
4	–79.3	–73.3	78.6	74	–151.0	1.2	–157.9	1.3

According to the $\hat{\pi}_{\text{BICM}}(y)$ estimate of the integrated likelihood, the preferred latent space model for these data is a two-dimensional one. These data have usually been visualized in two dimensions, so this agrees with previous practice, although we are not aware of any previous efforts to choose the dimension of the latent space in a formal way. AICM makes the same choice, although by a small margin over the one-dimensional model.

Figure 6 shows the estimated latent positions for these data. The left panel shows the estimated two-dimensional positions. The three well-known groups are clearly delineated. It is clear that the density of links is highest within each group. However, the Young Turks have some links to both of the other groups, while the Loyal Opposition and the Outcasts are joined by very few links. This suggests that a one-dimensional arrangement with the Young Turks in the middle might represent the main features of the data adequately.

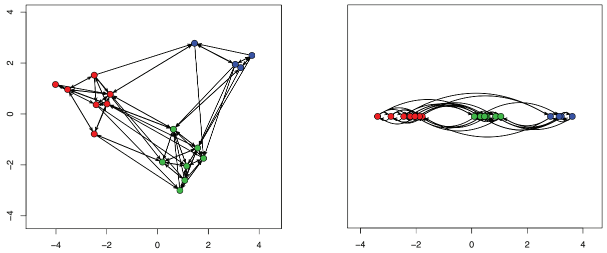

Figure 6: *Estimated Latent Positions of Monks in Social Network Example. Left panel: Two-dimensional latent positions with links also shown. Right panel: One-dimensional latent positions. In both plots, the known groupings of the monks are shown: Red = Loyal opposition; Green = Young Turks; Blue = Outcasts. Both the one-dimensional and the two-dimensional latent position models give results that are consistent with the known groupings.*

The right panel of Figure 6 shows the one-dimensional estimated latent positions. The three main groups are as well identified by the one-dimensional model as by the two-dimensional model. Again it seems reasonable that the Young Turks have a more central position, suggesting that a one-dimensional latent space captures most of the main features of the data, as suggested by the relatively small differences in BICM and AICM between the one- and two-dimensional models.

Our method provides standard errors for BICM and AICM, and these are also shown in Table 2. These increase rapidly, and roughly proportionally with the number of parameters. They can be used to calculate standard errors of the difference between the BICM values for two different models using the fact that the values for different models are independent. We use the standard formula

$$\text{SE}\ (\text{BICM}_1 - \text{BICM}_2) = \sqrt{\text{SE}(\text{BICM}_1)^2 + \text{SE}(\text{BICM}_2)^2}.$$

A similar formula holds for differences between AICM values. On the key model comparison, between the one- and two-dimensional latent space models, the standard error of the difference between the values of $\log \hat{\pi}_{\text{BICM}}(y)$ for the two models is about 0.95, suggesting that the observed difference of 3.3 would be unlikely to change sign if more MCMC runs were done. The standard error of the difference between the two values of $\frac{1}{2}$AICM is about 0.8, casting some doubt on whether the observed difference of 1.3 would persist in a longer MCMC run. If one wanted to select one model on the basis of AICM, this suggests that a longer MCMC run should be used.

5. DISCUSSION

Our final goal is a generic method that estimates the integrated likelihood using only the likelihoods given a set of draws from the posterior. We have investigated approaches to this based on the harmonic mean identity, which says that the integrated likelihood is the posterior harmonic mean of the likelihood. The most obvious esimator from this, the sample posterior harmonic mean of the likelihoods, is unbiased and simulation-consistent, but does not have finite variance in general and so is often unstable (Newton and Raftery 1994).

We have investigated two approaches to more stable estimation of the integrated likelihood using the harmonic mean identity. The first is to reduce the parameter space and then use the sample posterior harmonic mean; by judiciously choosing the likelihood to be used this can yield stable and finite variance estimators. The second approach involves modeling the posterior distribution of the loglikelihood by a shifted gamma distribution. This leads to estimates of the effective number of parameters and the true maximum likelihood that seem to work well, and hence to posterior-simulation-based analogues of the well-known BIC and AIC criteria, called BICM and AICM.

Our first approach takes advantage of dimension-reducing transformations on the parameter space. The proposed variance stabilizing method extends a very simple tool into a range of widely used hierarchical statistical models. As illustrated in Section 3, dimension reduction is straightforward in certain hierarchical models. Sometimes the natural approach of integrating out a nuisance parameter does not yield a stabilized estimator, however, and one must search farther. We have given one example in Section 3.3, a simple Poisson-Gamma model, where the natural approach does not work directly, but a slight refinement of the $h(\cdot)$ function does yield a stabilized estimator. The trick used there to find this refined h function was based on the fact that the estimator is stable if and only if $E\{\pi[y|h(\theta)]^{-2}|y\} < \infty$. We wrote this expectation as an integral, identified the part of the range of integration responsible for the integral being infinite, and effectively carried out the integration over that small part of the space via analytic approximation, thus defining a new h function. Dimension reduction for variance stabilization may not be an effective method to compute normalizing constants in certain very hard problems. In the cases we have studied, we have shown that it is possible to stabilize the harmonic mean estimator and obtain estimates that are much more accurate, but still easy to calculate.

Another application of our first stabilization approach includes robust linear models (Andrews and Mallows 1974; Carlin and Louis 1996). The robust linear model has an error term distributed as $Z/\sqrt{U}$, where Z and U are independent, Z has a centered normal distribution, and U has a χ^2 distribution. The standard harmonic mean estimator can have infinite variance. A stabilized harmonic mean estimator can then be obtained by integrating out the denominator U.

Hierarchical models that involve standard distributions may be good candidates for our first approach. For one thing, MCMC is well understood for within-model posterior simulation. Furthermore, the integrations required for dimension reduction may be solved analytically. The simplicity of the resulting stabilized harmonic mean is its main advantage.

Our second approach involves modeling the posterior distribution of the loglikelihoods by a shifted gamma distribution. This fits the observed distribution of loglikelihoods well in some applications, and leads to very simple estimates of the effective number of parameters and the true maximum likelihood that seem of good quality. This in turn yields posterior-simulation-based analogues of the BIC and AIC criteria, BICM and AICM. It also provides simple standard errors for these criteria, which can be useful both for assessing the results and for deciding whether enough samples have been drawn from the posterior for model comparison purposes.

The BICM criterion we have defined requires the specification of sample size, and this may be problemmatical in some applications. The analogies with the results of Volinsky and Raftery (2000) suggest that in fixed effects models acceptable choices may be possible by considering what a reasonable choice of a unit of information for a unit information prior would be. Analogies with the results of Pauler (1998) suggest a corresponding approach for random effects or hierarchical models. In our examples, these approaches have worked fairly well.

It would be desirable, however, to have a fully automated solution where this parameter could be estimated from the posterior simulation output. We have investigated various possible solutions to this, mostly Bayesian estimates of the gamma distribution parameters that exploit the prior information that the scale parameter is less than 1, but not much less than 1. The results so far have not satisfied us fully, however, and so we did not present them here.

The general idea explored here, of estimating the posterior harmonic mean of the likelihood by modeling the loglikelihoods, may yield progress by using models other than the shifted gamma distribution. For example, it may be possible to make progress by recognizing that in regular models the posterior distribution of the loglikelihood can be approximated asymptotically by a shifted and scaled noncentral chi-squared distribution with a small noncentrality parameter, perhaps better than by the (central) shifted and scaled gamma distribution we have been using so far. The estimation of the scale and noncentrality parameters is delicate, however.

Another approach might take advantage of the work that has been done on approximating the posterior distribution of the loglikelihood using Edgeworth expansions. Bickel and Ghosh (1990) proposed such an expansion where the leading term is of the form (10). This expansion would not in itself be useful for the present purpose because the leading term still yields an infinite log integrated likelihood, but the basic idea may be fruitful in a modified form. Other expansions that have been proposed might also be useful; many of these are reviewed by Reid (2003).

A range of other methods for computing integrated likelihoods from posterior simulation have been proposed. Most of these methods are not generic algorithms that use only the output of the posterior simulation; in most cases they require additional simulations or model-specific calculations. Other methods have been proposed for estimating Bayes factors or posterior model probabilities, but not the underlying integrated likelihoods themselves. Subsets of the different methods have been reviewed and compared by DiCiccio, Kass, Raftery, and Wasserman (1997), Han and Carlin (2001), Bos (2002), Clyde and George (2004), Sinharay and Stern (2005), and Rossi, Allenby and McCulloch (2005, chapter 6).

Newton and Raftery (1994) proposed modifications of the harmonic mean estimator using real or imaginary draws from the prior, and these have been applied, for example by Zijlstra, van Duijn, and Snijders (2005), with some success, but they are still somewhat unstable. As we discussed in Section 2, Gelfand and Dey (1994) proposed a method that can be viewed as a generalization of the harmonic mean estimator. It requires the careful choice of a function of the entire parameter vector, tailored for each application, and so is not as generic as the methods we have been discussing, although with a good choice of function it can perform well. As we have shown in Section 3.2, it can be combined with our first approach to achieve further improvements.

The method of Chib (1995) was developed for the specific case where posterior simulation is done by Gibbs sampling. It is based on the conditional probability formula for the normalizing constant, and requires running specially designed auxiliary conditional MCMC samplers. Chib and Jeliazkov (2001) extended this to the case of the Metropolis-Hastings algorithm, in which case it requires a different auxiliary simulation algorithm additional to the main MCMC algorithm. These methods have been successfully applied to specific models, for example by Albert and Chib (2001), Chib, Nardard, and Shephard (2002), and Basu and Chib (2003). However, Neal (1999) showed that Chib (1995)'s application of the idea to mixture models was incorrect, and Rossi et al (2005, Section 6.9) showed the instability of the method due to large outliers in the posterior simulation.

The method of Chib (1995) was developed for the specific case where posterior simulation is done by Gibbs sampling. It is based on the conditional probability formula for the normalizing constant, and requires running specially designed auxiliary conditional MCMC samplers. Chib and Jeliazkov (2001) extended this to the case of the Metropolis-Hastings algorithm, in which case it requires a different auxiliary simulation algorithm additional to the main MCMC algorithm.

Oh (1999) proposed a method based on an identity that requires knowledge of full conditional posterior densities. Lockwood and Schervish (2005) proposed two methods, one a brute force method, and the other a sequential approach that is related to the method of Oh (1999). Chen (2005), building on Chen (1994), proposed a method that uses another identity. It involves the use of latent variables and the proposed optimal version of the method requires knowledge of the full conditional posterior distribution of the parameters given the latent variables, including all normalizing constants.

A version of the Laplace method in which the required posterior modes and Hessian matrices are estimated from posterior simulation output, called the Laplace-Metropolis method, was proposed by Raftery (1996) and Lewis and Raftery (1997). This is a generic method but can depend on the model's parameterization, and may not work well for very high-dimensional models. Importance sampling based methods have also been proposed (Nandram and Kim 2002; Steele, Raftery, and Emond 2006), but these can also require model-specific computations.

Several methods have been proposed for estimating Bayes factors, or ratios of integrated likelihoods, but not the integrated likelihoods themselves. These include the Savage-Dickey ratio and a generalization of it (Verdinelli and Wasserman 1995), and bridge sampling (Meng and Wong 1996; Mira and Nicholls 2004). Johnson (1999) has proposed a method for estimating the integrated likelihood that involves simulating from a second density as well as the posterior; it seems that for its performance to be good the second density needs to be carefully chosen taking account of the situation at hand.

A general approach to estimating posterior model probabilities is to use trans-dimensional MCMC, pioneered by Green (1995) with his introduction of reversible jump MCMC; a review of this area is given by Sisson (2005). These methods can be used to estimate Bayes factors, but not the underlying integrated likelihoods. Bayes factors can be read off the output of transdimensional MCMC directly, and more efficient approaches to estimating Bayes factors from transdimensional MCMC have been discussed by Bartolucci, Scaccia, and Mira (2006). Godsill (2001) has pointed out that integrating out parameters analytically can improve the efficiency of transdimensional MCMC; this is analogous to our proposal here to stablize the harmonic mean estimator by parameter reduction.

ACKNOWLEDGEMENTS

The authors are grateful to Mark Handcock and Marijtje van Duijn for extensive and helpful discussions. They also thank Jim Berger, David Dunson, Peter Hoff, Donna Pauler, Matthew Stephens and Aki Vehtari for useful comments.

REFERENCES

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *Proc. 2nd Internat. Symp. Information Theory* (B. N. Petrov and F. Csáski, eds.), 267–281. Budapest: Akadémiai Kiadó.

Albert, J. H. and Chib, S. (2001). Sequential ordinal modeling with applications to survival data. *Biometrics* **57**, 829–836.

Andrews, D. F. and Mallows, C. L. (1974). Scale mixtures of normality. *J. Roy. Statist. Soc. B* **36**, 99–102.

Bartolucci, F., Scaccia, L. and Mira, A. (2006). Efficient Bayes factor estimation from the reversible jump output. *Biometrika* **3**, 41–52.

Basu, S. and Chib, S. (2003). Marginal likelihood and Bayes factors for Dirichlet process mixture models. *J. Amer. Statist. Assoc.* **98**, 224–235.

Berger, J. O., Ghosh, J. K. and Mukhopadhyay, H. (2003). Approximations and consistency of Bayes factors as model dimension grows. *J. Statist. Planning and Inference* **112**, 241–258.

Bickel, P. J. and Ghosh, J. K. (1990). Decomposition of the likelihood ratio statistic and the Bartlett correction – A Bayesian argument. *Ann. Statist.* **18**, 1070–1090.

Bos, C. S. (2002). A comparison of marginal likelihood computation methods. *Tech. Rep.*, Vrije Unversiteit Amsterdam, The Netherlands..

Carlin, B. P. and Louis, T. A. (1996). *Bayes and Empirical Bayes Methods for Data Analysis*, Chapter 6, pp. 209–211. London: Chapman and Hall

Celeux, G., Hurn, M. and Robert ,C. (2000). Computational and inferential difficulties with mixture posterior distribution. *J. Amer. Statist. Assoc.* **95**, 957–970.

Chen, M.-H. (1994). Importance-weighted marginal Bayesian posterior density estimation. *J. Amer. Statist. Assoc.* **89**, 818–824.

Chen, M.-H. (2005). Computing marginal likelihoods from a single MCMC output. *Statistica Neerlandica* **59**, 16–29.

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

Chib, S. and Jeliazkov, I. (2001). Marginal likelihood from the Metropolis-Hastings output. *J. Amer. Statist. Assoc.* **96**, 270–281.

Chib, S., Nardard, F. and Shephard, N. (2002). Markov chain Monte Carlo methods for stochastic volatility models. *J. Econometrics* **108**, 281–316.

Chipman, H., George, E. I. and McCulloch, R. E. (2001). The practical implementation of Bayesian model selection. *Model Selection* (IMS Lecture Notes) **38**) Hayward, CA: IMS, 65–134.

Clyde, M. and George, E. I. (2004). Model uncertainty. *Statist. Science* **19**, 81–94.

Dawid, A. P. (1991). Fisherian inference in likelihood and prequential frames of reference. *J. Roy. Statist. Soc. B* **53**, 79–109, (with discussion).

DiCiccio, T. J., Kass, R. E., Raftery, A. E. and Wasserman, L. (1997). Computing Bayes factors by combining simulation and asymptotic approximations. *J. Amer. Statist. Assoc.* **92**, 903–915.

Diebolt, J. and Robert, C. P. (1994). Bayesian estimation of finite mixture distributions. *J. Roy. Statist. Soc. B* **56**, 363–375.

Doerge, R. W., Zeng, Z.-B. and Weir, B. S. (1997). Statistical issues in the search for genes affecting quantitative traits in experimental populations. *Statist. Science* **12**, 195–219.

Fan, J., Hung, H.-N. and Wong, W.-H. (2000). Geometric understanding of likelihood ratio statistics. *J. Amer. Statist. Assoc.* **95**, 836–841.

Gelfand, A. E. and Dey, D. K. (1994). Bayesian model choice: asynptotics and exact calculations. *J. Roy. Statist. Soc. B* **56**, 501–514.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2004). *Bayesian Data Analysis*, 2nd ed. Boca Raton, FL: Chapman and Hall/CRC Press.

Godsill, S. J. (2001). On the relationship between Markov chain Monte Carlo methods for model uncertainty. *J. Comp. Graphical Statist.* **10**, 230–248.

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

Han, C. and Carlin, B. P. (2001). Markov chain Monte Carlo methods for computing Bayes factors: A comparative review. *J. Amer. Statist. Assoc.* **96**, 1122–1132.

Handcock, M. S., Raftery, A. E. and Tantrum, J. (2005). Model-based clustering for social networks. *Tech. Rep.*, University of Washington, USA.

Hoeting, J. A., Madigan,D., Raftery, A. E. and Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statist. Science* **15**, 193–195 (with discussion).

Hoff, P., Raftery, A. E. and Handcock, M. S. (2002). Latent space approaches to social network analysis. *J. Amer. Statist. Assoc.* **97**, 1090–1098.

Johnson, V. E. (1999). Posterior distributions on normalizing constants. *Tech. Rep.*, ISDS, Duke University, USA.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *J. Amer. Statist. Assoc.* **90**, 773–795.

Kass, R. E. and Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *J. Amer. Statist. Assoc.* **90**, 928–934.

Lewis, S. M. and Raftery, A. E. (1997). Estimating Bayes factors via posterior simulation with the Laplace-Metropolis estimator. *J. Amer. Statist. Assoc.* **92**, 648–655.

Lockwood, J. R. and Schervish, M. J. (2005). MCMC strategies for computing Bayesian predictive densities for censored multivariate data. *J. Comp. Graphical Statist.* **14**, 395–414.

Meng, X.-L. and Wong, W.-H. (1996). Simulating ratios of normalizing constants: a theoretical exploration. *Statistica Sinica* **6**, 831–860.

Mira, A. and Nicholls, G. (2004). Bridge estimation of the probability density at a point. *Statistica Sinica* **4**, 603–612.

Nandram, B. and Kim, H. (2002). Marginal likelihood for a class of Bayesian generalized linear models. *J. Statist. Computation and Simulation* **72**, 319–340.

Neal, R. M. (1999). Erroneous results in 'Marginal likelihood from Gibbs output'. http://www.cs.utoronto.ca/˜radford.

Newton, M. A. and Raftery, A. E. (1994). Approximate Bayesian inference by the weighted likelihood bootstrap. *J. Roy. Statist. Soc. B* **56**, 3–48, (with discussion).

Oh, M.-S. (1999). Estimation of posterior density functions from a posterior sample. *Comput. Statist. Data Anal.* **29**, 411–427.

Pauler, D. K. (1998). The Schwarz criterion and related methods for normal linear models. *Biometrika* **85**, 13–27.

Pritchard, J. K., Stephens, M. and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959.

Raftery, A. E. (1995). Bayesian model selection in social research (with discussion). *Sociological Methodology* **25**, 111–196.

Raftery, A. E. (1996). Hypothesis testing and model selection. In W. R. Gilks, D. J. Spiegelhalter, and S. Richardson (Eds.), *Markov Chain Monte Carlo in Practice*, pp. 163–188. London: Chapman and Hall.

Reid, N. (2003). Asymptotics and the theory of inference. *Ann. Statist.* **31**, 1695–1731.

Richardson, S. (2002). Discussion of Spiegelhalter *et al. J. Roy. Statist. Soc. B* **64**, 626–627.

Rossi, P. E., Allenby, G. M. and McCulloch, R. (2005). *Bayesian Statistics and Marketing.* Chichester: Wiley

Sampson, S. F. (1968). *A Novitiate in a Period of Change: An Experimental and Case Study of Relationships.* Ph.D. Thesis, Cornell University, USA.

Satagopan, J. M., Yandell, B. S., Newton, M. A. and Osborn, T. C. (1996). A Bayesian approach to detect quantitative trait loci using Markov chain Monte Carlo. *Genetics* **144**, 805–816.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 497–511.

Sinharay, S. and Stern, H. S. (2005). An empirical comparison of methods for computing Bayes factors in generalized linear mixed models. *J. Comp. Graph. Statist.* **14**, 415–435.

Sisson, S. A. (2005). Transdimensional Markov chains: A decade of progress and future perspectives. *J. Amer. Statist. Assoc.* **100**, 1077–1089.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *J. Roy. Statist. Soc. B* **64**, 583–639 (with discussion).

Steele, R., Raftery, A. E. and Emond, M. (2006). Computing normalizing constants for finite mixture models via incremental mixture importance sampling (IMIS). *J. Comp. Graphical Statist.* **15**, 712–734.

Stephens, M. (2000). Dealing with label-switching in mixture models. *J. Roy. Statist. Soc. B* **62**, 795–809.

Sturtz, S., Ligges, U. and Gelman, A. (2005). R2WinBUGS: A package for running WinBUGS from R. *J. Statist. Software* **12**, 1–17.

Verdinelli, I. and Wasserman, L. (1995). Computing Bayes factors using a generalization of the Savage-Dickey density ratio. *J. Amer. Statist. Assoc.* **90**, 614–618.

Volinsky, C. T. and Raftery, A. E. (2000). Bayesian information criterion for censored survival models. *Biometrics* **56**, 256–262.

Zijlstra, B. J. H., van Duijn, M. A. J. and Snijders, T. A. B. (2005). Model selection in random effects models for directed graphs using approximated Bayes factors. *Statistica Neerlandica* **59**, 107–118.

APPENDIX I: STUDENT'S t

Student t

Copying Bernardo and Smith (1994, page 122),

$$\mathrm{St}(x|\mu,\lambda,\alpha) = c\left[1+\frac{\lambda}{\alpha}(x-\mu)^2\right]^{-(\alpha+1)/2}, \qquad c = \frac{\Gamma((\alpha+1)/2)}{\Gamma(\alpha/2)\;\Gamma(1/2)}\left(\frac{\lambda}{\alpha}\right)^{1/2}.$$

Multivariate Student t

Using the notation of Bernardo and Smith (1994, page 139),

$$\mathrm{St}_n(x|\mu,\lambda,\alpha) = c\left[1+\frac{1}{\alpha}(x-\mu)^T\lambda(x-\mu)\right]^{-(\alpha+n)/2},$$

where

$$c = \frac{\Gamma((\alpha+n)/2)}{\Gamma(\alpha/2)\ (\alpha\pi)^{n/2}}\det(\lambda)^{1/2}.$$

x and μ are of dimension n. λ is a symmetric, positive-definite $n \times n$ matrix, and $\alpha > 0$.

APPENDIX II: PROOF OF EQUATION (4)

Define

$$f(\mu) = \tfrac{n_0}{\alpha}(\mu-\mu_0)^2 \qquad \text{and} \qquad g(\mu) = \tfrac{1}{\alpha}(y-\mu)^2 \ .$$

Set

$$a(\mu) = 1 + \frac{g(\mu)}{1+f(\mu)} \ .$$

It can be easily shown that the maximum of the continuous function $a(\mu)$ occurs at $\mu^* = \mu_0 - \alpha/[n_0(y-\mu_0)]$, and the maximum value of the function is

$$a(\mu^*) = 1 + \frac{1}{n_0} + g(\mu_0) \ .$$

Further $a(\mu) \to 1+1/n_0$, as $\mu \to \pm\infty$. The expected value of interest can be written as

$$E\left\{\frac{1}{[\pi(y|\mu)]^2}|y\right\} \quad \propto \quad \int [a(\mu)]^{\alpha/2+1}[1+f(\mu)]^{-\alpha/2}d\mu \ ,$$

where $[1+f(\mu)]^{-\alpha/2}$ is proportional to a t-density of the form

$$\mathrm{St}(\mu|\mu_0, n_0(\alpha-1)/\alpha, \alpha-1) \ .$$

Since $1 \leq a(\mu) \leq a(\mu^*)$, the integral on the right hand side is finite by dominated convergence theorem when $\alpha > 1$ and $n_0 > 0$.

APPENDIX III: PROOF OF THEOREM 1

Define $\alpha = h(\theta)$, write $\theta = (\alpha, \beta)$, and set

$$a = \mathrm{E}\left\{\frac{1}{[\pi(y|\alpha)]^2}\middle| y\right\} \qquad \text{and} \qquad b = \mathrm{E}\left\{\frac{1}{[\pi(y|\theta)]^2}\middle| y\right\}.$$

Since both $1/\pi(y|\alpha)$ and $1/\pi(y|\theta)$ have common expectation $1/\pi(y)$, it suffices to show that $a \leq b$. Expanding b, we have

$$\begin{aligned} b &= \int\int \frac{1}{[\pi(y|\alpha,\beta)]^2}\,\pi(\alpha,\beta|y)\,d\beta\,d\alpha \\ &= \int\int \frac{1}{[\pi(y|\alpha,\beta)]^2}\,\pi(\beta|\alpha,y)\,p(\alpha|y)\,d\beta\,d\alpha \\ &= \int b(\alpha)\,\pi(\alpha|y)\,d\alpha \end{aligned}$$

where

$$b(\alpha) = \int \frac{1}{[\pi(y|\alpha,\beta)]^2}\,\pi(\beta|\alpha,y)\,d\beta.$$

By contrast,

$$a = \int a(\alpha)\pi(\alpha|y)\,d\alpha$$

where

$$a(\alpha) = \frac{1}{[\pi(y|\alpha)]^2}.$$

Therefore, it is sufficient to prove that $a(\alpha) \leq b(\alpha)$ for all α. Simplifying $b(\alpha)$, we have

$$\begin{aligned} b(\alpha) &= \int \frac{1}{[\pi(y|\alpha,\beta)]^2}\pi(\beta|\alpha,y)\,d\beta \\ &= \int \frac{1}{[\pi(y|\alpha,\beta)]^2}\,\frac{\pi(y|\alpha,\beta)\,\pi(\beta|\alpha)\,\pi(\alpha)}{\pi(y|\alpha)\,\pi(\alpha)}\,d\beta \\ &= \frac{1}{\pi(y|\alpha)}\int \frac{\pi(\beta|\alpha)}{\pi(y|\alpha,\beta)}\,d\beta. \end{aligned}$$

Cancelling one factor $1/\pi(y|\alpha)$, we have $a(\alpha) \leq b(\alpha)$ if

$$\frac{1}{\pi(y|\alpha)} \leq \int \frac{\pi(\beta|\alpha)}{\pi(y|\alpha,\beta)}\,d\beta.$$

This follows by Jensen's inequality using the distribution $\pi(\beta|\alpha)$. In the event that one or another of the integrals diverges, $a(\alpha) \leq b(\alpha)$ must continue to hold.

DISCUSSION

NICHOLAS G. POLSON (*University of Chicago, USA*)

The authors are to be congratulated on their extension of the popular harmonic mean (HM) estimator for marginal likelihoods. They propose a stabilised harmonic mean (SHM) estimator for stabilising the possible infinite variance of the original harmonic mean estimator. A number of examples illustrating their approach are given. They also discuss an application to model comparison and provide a posterior simulation-based alternative to BIC which they term BICM. This measure requires an estimate of the maximum achievable log integrated likelihood $l_{\max}$ and they show how to use the MCMC draws to achieve this. A common theme throughout the paper is that there is extra information, particularly in the tails, from the MCMC output for harmonic means. This information can be thoughtfully used to provide better estimates than the usual ergodic averaging approach. In this discussion I will focus on three issues: 1) The Monte Carlo convergence rate for harmonic mean estimators based on a method described in Wolpert (2002), 2) Alternative approaches to marginal likelihoods based on extensions of the Savage density ratio, see Verdinelli and Wasserman (1995) and Jacquier and Polson (2002) and 3) an MCMC approach for computing the maximum achievable log-integrated likelihood, see Jacquier, Johannes and Polson (2006).

First, the basic problem that the harmonic mean estimator tackles is the estimation of the marginal likelihood defined by $\pi(y) = \int \pi(y|\theta)p(\theta)d\theta$. This is a central problem in Bayesian inference and forms the basis for model selection and comparison. The harmonic mean estimator has become popular due to its simplicity. It simply takes the MCMC draws $\{\theta^{(t)}\}_{t=1}^{B}$ and computes

$$\hat{\pi}_{HM}(y) = \frac{1}{\frac{1}{B}\sum_{t=1}^{B}\frac{1}{\pi(y|\theta^{(t)})}}$$

as an estimator. A caveat noted by the authors is that this estimator can have infinite variance. The proposal described here is a stabilised harmonic mean estimator of the form

$$\hat{\pi}_{SHM}(y) = \frac{1}{\frac{1}{B}\sum_{t=1}^{B}\frac{1}{\pi(y|\mu^{(t)})}}$$

where $\mu = h(\theta)$ is a dimensionality reduction. The intuition is that this marginalisation will lead to a heavier-tailed distribution $\pi(y|\mu)$ which in turn will lead to an estimator with finite variance.

Another caveat with these types of estimators is that they can have slow convergence properties in B. From a practical perspective this implies that a few large outliers can dominate the estimator which leads to large Monte Carlo errors. Wolpert (2002) discusses this issue and proposes a method to accelerate the convergence. Specifically, the issue is as follows: standard ergodic averaging yields

$$\frac{S_B}{B} \rightarrow \frac{1}{\hat{\pi}(y)}$$

where $S_B = \frac{1}{B}\sum_{t=1}^{B}\frac{1}{\pi(y|\theta^{(t)})}$. However, this occurs at a *very slow* rate given by

$$\frac{S_B}{B} \approx \frac{1}{\hat{\pi}(y)} + ZB^{\frac{1}{\alpha}-1}$$

where $\alpha \approx 1$. The asymptotic distribution Z can be characterised as a fully-skewed stable distribution $S_\alpha(\delta_B, \gamma_B)$. Unfortunately, the convergence rate can be poor. In a simple normal location problem for example, the convergence rate in B is

$$\frac{1}{\alpha} - 1 = \left(1 + \frac{n\tau^2}{\sigma^2}\right)^{-1} \approx 0.$$

Hence, the Monte Carlo error can be large. One solution is to use the information in the whole distribution of S_B/B rather than simple ergodic averaging. Specifically, we can use quantile information to estimate $(\alpha, \delta_B, \gamma_B)$ and hence estimate $1/\pi(y) = E(S_B/B)$ using the moment identity

$$E\left(\frac{S_B}{B}\right) = \hat{\delta} - \hat{\gamma} \tan \frac{\pi\hat{\alpha}}{2}$$

Again the intuition is that there is extra information in the whole distribution of the MCMC draws.

It should also be noted that slow convergence can occur in other approaches for estimating marginal likelihoods. For example, the marginal likelihood approach in Chib (1995) where the estimator

$$\hat{\pi}_C(y) = \frac{\pi(y|\theta^\star)\pi(\theta^\star)}{\frac{1}{B}\sum_{t=1}^{B} \pi(\theta^\star|\phi^{(t)}, y)}$$

is used for any $\theta^\star$. From a purely Monte Carlo perspective it is always wise to avoid estimators than are ratios of averages

One class of problems where we can avoid the use of ratios of MC averages is the class of models where extensions of the Savage-density ratio approach to calculating Bayes factors ($\mathcal{BF}$) applies. There is a long history discussing the relationship between marginal likelihoods, Bayes factors and their use in model selection. A common approach for nested models is to estimate

$$\mathcal{BF} = \frac{1}{B}\sum_{t=1}^{B} \frac{p(\theta_1^0|\theta_2^{(t)}, y)}{p(\theta_1^0)}$$

With the use of data augmentation many of the restrictive conditions for the original approach can be relaxed. For example, Verdinelli and Wasserman (1995) generalises this estimator to

$$\mathcal{BF} = \hat{C}\frac{p(\theta_1|y)}{p(\theta_1)} \quad \text{where} \quad \hat{C} = \frac{1}{B}\sum_{t=1}^{B} \frac{p_1(\theta_2^{(t)})}{p(\theta_1, \theta_2^{(t)})}$$

Jacquier and Polson (2002) provides an extension that relaxes the assumption the $p(\theta_2|\theta_1^0) = p(\theta_2|\mathcal{M}_1)$. Here we obtain as estimator of the form

$$\mathcal{BF} = \frac{1}{B}\sum_{t=1}^{B} \frac{p(y|\theta_1^1, \theta_2^{(t)})}{p(y|\theta_1^{(t)}, \theta_2^{(t)})} \frac{p_1(\theta_2^{(t)})}{p(\theta_2^{(t)}|\theta_1^{(t)})}$$

When these approaches apply, the key is that they avoid Monte Carlo estimates that are reciprocals. Standard central limit theorem type convergence results in B also hold for these procedures.

Finally, there is a useful MCMC alternative for estimating the maximum achievable log integrated likelihood for model comparison, see Jacquier, Johannes and Polson (2006). Specifically, suppose that you are interested in an integrated likelihood of the form $\mathcal{L}(\theta_1) = \int_{\theta_2} f(y|\theta_1, \theta_2)p(\theta_2|\theta_1)d\theta_2$ for a given model specification.

Now one can consider a MCMC posterior analysis where we copy the variable θ_2 to be integrated over, J independent times. This leads to a joint posterior of the form

$$p_J\left(\theta_2^J, \theta_1|y\right) \propto \prod_{j=1}^{J} f(y|\theta_1, \theta_2^j)p(\theta_2^j, \theta_1)$$

This joint density has the property that on marginalising out the θ_2 parameter leads to a marginal $\pi_J(\theta_1|y) \propto \exp\left(J \ln \mathcal{L}(\theta_1)\right)$. This marginal collapses on $\hat{\theta}_1$ as $J \to \infty$ and so $l_{\max} = \ln \mathcal{L}(\hat{\theta}_1)$ is easily estimated.

Using MCMC to simulate from $p_J\left(\theta_2^J, \theta_1|y\right)$ is straightforward as we can iteratively simulate the conditionals

$$\begin{array}{rc} \theta_2^j|\theta_1, y \sim & p(\theta_2^j|\theta_1, y) \\ \theta_1|\theta_2^J, y \sim & \prod_{j=1}^{J} p(\theta_1|\theta_2^j, y) \end{array}$$

In many situation this approach gives reasonable answers for small values of J.

In summary, this paper provides main insights and suggestions into tackling the hard problem of computing marginal likelihoods. One area where the literature is currently silent is in describing why Markov Chain MC sampling has been so successful for computing marginalisation constants as compared to more standard importance sampling techniques which have similar Monte carlo averaging properties. Maybe it is due to the authors' intuition that there's a lot of extra information in the whole distribution, particularly the tails, of harmonic mean estimators?

BRADLEY P. CARLIN (*University of Minnesota, USA*) and
DAVID J. SPIEGELHALTER (*MRC Biostatistics Unit Cambridge, UK*)

Congratulations to the authors for a fine extension and updating of their earlier ideas in Bayesian model choice using the harmonic mean estimator, AIC, and BIC. The first author is a long-time supporter and developer of BIC-related tools for Bayesian data analysis, and this paper makes a significant and welcome addition to this literature.

We wish to discuss the authors' estimate of the effective number of model parameters, $\hat{d} = 2s_\ell^2$, or twice the sample variance of the log-likelihood samples ℓ_t. The authors obtain this as the result of a simple moment estimate of α in their shifted gamma model (Section 4.1). In fact this same estimate appeared in an early version of the original DIC paper by Spiegelhalter *et al.* (2002); see for example equation (18) in the August 2001 version of the paper, available online at `ftp://muskie.biostat.umn.edu/pub/2001/rr2001-013.pdf.gz`. In that paper, the reasoning underlying the estimate was based on the approximation

$$D(\theta) \approx D(\hat{\theta}) + \chi_p^2 \, , \tag{33}$$

where $D(\theta) = -2\ell(\theta)$, the *deviance*, which when prior information is weak is essentially identical to the shifted gamma model in (10). Approximation (33) holds provided the posterior distribution $p(\theta \mid y)$ can be reasonably well approximated by a multivariate normal distribution. Taking posterior expectations of both sides produces

$$E[D(\theta) \mid y] \approx D(\hat{\theta}) + p \ ,$$

motivating the basic formula for our measure of effective model size, $p_D = \overline{D} - D(\hat{\theta})$, where $\overline{D}$ is the sample mean of the MCMC deviance samples, $D_t \equiv -2\ell_t$. If we however take the posterior *variance* of both sides of (33), we obtain

$$Var[D(\theta) \mid y] \approx 2p \ .$$

This suggests the alternative empirical estimate of model size $p_V = s_D^2/2$, where s_D^2 is the sample variance of the deviance samples. Clearly the relationship between D and ℓ means that $p_V = \hat{d}$. Incidentally, the use of half the variance of the deviance as the effective number of parameters is also recommended in Gelman *et al.* (2004, pp.181–182), and is the method used to compute DIC in the `R2WinBUGS` package for `R`; see `http://cran.r-project.org/src/contrib/Descriptions/R2WinBUGS.html.`

However, the main point here is not to claim credit for this variance-based estimate of effective model size, but instead to clarify why we deleted it from the final version of the DIC paper! We tried for a long time to convince ourselves that this attractive and parameterisation-invariant quantity was an appropriate measure, but failed. The crucial test for us was whether, in the normal hierarchical model, the effective number of parameters was given as $\mathrm{tr}(H)$, where H is the "hat" matrix that projects the observations onto their fitted values. As described in Spiegelhalter *et al.* (2002), this measure has been derived from numerous perspectives, in particular in a recent extension of AIC to normal mixed models (Vaida and Blanchard, 2005). (To be honest, our proposal for p_D was motivated primarily by working backwards from this desired result.)

Consider the simulation exercise carried out by the authors in Section 4.2 for the general case with prior variance σ^2. Denote $n_0 = 1/\sigma^2$, the effective sample size in the prior, and let $\rho = n/(n + n_0)$ so that $\mu_i | \overline{y}_i \sim \mathrm{N}(\rho\overline{y}_i, \rho n_0)$. Hence ρ can be interpreted as the "shrinkage," or the proportion of the posterior precision arising from the likelihood: in this situation $p_D = \mathrm{tr}(H) = d\rho$.

The posterior distribution of the deviance can be derived in closed form in this situation, and is a sum of shifted non-central chi-square distributions (Spiegelhalter et al., 2002) with exact variance

$$V = 2\rho^2[d + 2(1-\rho)n_0 \sum_{j=1}^{d} \overline{y}_j^2] \ .$$

Since $\overline{y}_j \sim \mathrm{N}(0, (\rho n_0)^{-1})$, V has a sampling expectation

$$E[V] = 2\rho^2[d + 2(1-\rho)d/\rho] = 2d\rho(2-\rho) \ .$$

Hence $p_V = V/2 \approx d\rho(2-\rho)$, compared to $p_D = d\rho$. When ρ is near 1 the two methods will closely agree: we note that ρ is 0.99 in Figure 5 and generally very close to 1 in other situations explored by the authors. However in any interesting application of hierarchical models there will be non-negligible shrinkage and values of ρ closer to 0.5 may be more typical. In such cases we might expect p_V to substantially *overestimate* the effective number of parameters.

This is born out in the following analysis of the well-known Scottish lip cancer data, introduced by Clayton and Kaldor (1987) and most famously analyzed by Besag *et al.* (1991). Table 3 presents a side-by-side comparison of p_D and $\hat{d}$ for a variety of models, all using a non-Gaussian (Poisson) likelihood. The models we consider are easily fit in `WinBUGS 1.4` (`http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml`) or `OpenBUGS` (`http://mathstat.helsinki.fi/openbugs/`) since this dataset is an example in the `GeoBUGS` User Manual: click on `Map` and pull down to `Manual`, then click on `Examples`. The full model for the log-relative risk of lip cancer in county i is given by

$$log(RR)_i = \beta_0 + \beta_1 x_i + \phi_i + \theta_i \; , \tag{34}$$

where β_0 is an intercept, β_1 is the effect of a single covariate (AFF, the percentage of county i's population engaged in agriculture, fishing, or forestry; essentially a surrogate for sunlight exposure), the ϕ_i are spatial clustering random effects assigned a conditionally autoregressive (CAR) prior, and the θ_i are pure heterogeneity random effects assigned an i.i.d. normal prior. Several authors have investigated whether these data support inclusion of either or both of the two sets of random effects, with the general consensus being that the clustering random effects are helpful, but the heterogeneity terms add little of additional value. The saturated model simply fits independent effects for each area and hence has no covariate structure.

Table 3: *Comparison of p_D and $\hat{d}$ for the Scottish lip cancer data.*

model	p_D	ΔDIC	$\hat{d}$	ΔAICM
full	33.1	—	42.9	—
clustering only	29.2	–0.5	45.0	5.5
heterogeneity only	40.0	11.6	59.4	21.3
fixed effects only	1.97	151.9	1.92	142.2
saturated model	52.8	18.1	57.8	13.4

Table 3 compares the effective model sizes p_D and $\hat{d}$, as well as the corresponding overall model choice statistics, DIC and AICM. Both of these latter two quantities are expressed in terms of change relative to the full model (34), with smaller values indicating preferred models. All of our computations were done in `WinBUGS`, using a single chain of 10,000 samples retained following a 1000-iteration burn-in period. We see that the performance of DIC and AICM are comparable, both obtaining rough equivalence between the full and clustering only models, which are slightly preferred over the heterogeneity only model and strongly preferred over the fixed effects only model. The two effective model size statistics also behave similarly in the simple fixed effects case, both obtaining an answer very close to the correct value of 2.0. However, the $\hat{d}$ values do emerge as significantly larger than p_D for all of the random effects models. The $\hat{d}$ value for the heterogeneity only model seems especially doubtful, since the upper bound here (obtained by simply counting parameters) is only 59: there are 56 county-level random effects, 2 fixed regression effects, and a single variance parameter (one could also argue that this upper bound should be 56, the number of data points). Also, $\hat{d}$ for the heterogeneity only model exceeds the value of $\hat{d}$ for the saturated model; it seems anomalous that a heterogeneity-only model can somehow be more complex than the saturated model.

Much larger MCMC sample sizes did not eliminate these problems with $\hat{d}$. More broadly, while the model choice issue here is fairly clear-cut, in other hierarchical modeling examples AICM may tend to select somewhat simpler models than DIC, due to its significantly higher effective model size penalty.

Regarding the use of $\hat{d}$ in BIC: in our reply to the discussion of Spiegelhalter *et al.* (2002) we said we could find no strong reason to use p_D in BIC, attractive though it would be. The authors define BICM "by analogy"; we wonder if there is any stronger theoretical justification for this assumption?

Naturally, far more extensive investigations of DIC, AICM, BICM, and related methods are warranted. Perhaps p_D's most embarrassing feature is that it can sometimes be negative, and clearly $\hat{d}$ does avoid this problem. Other recent work in this area includes that of Lu *et al.* (2004), who extended the approach of Hodges and Sargent (2001) to generalized linear mixed model settings, obtaining effective model size estimates that respect both the lower (0) and upper (raw parameter count) boundaries. Also noteworthy is the forthcoming *Bayesian Analysis* discussion paper by Celeux *et al.* (2006), who offered a variety of "repaired" versions of p_D for missing data (especially mixture model) settings.

DAVID DRAPER (*University of California at Santa Cruz, USA*)

I would like to add to the discussion of this interesting and stimulating paper by posing two questions.

As noted, for example, by Draper (1995), two Laplace approximations for computing log Bayes factors—when comparing parametric models M_j indexed by parameter vectors θ_j of dimension k_j, on the basis of a data set y consisting of n conditionally exchangeable observations—are

$$\ln p(y|M_j) = \frac{1}{2}k_j \ln(2\pi) - \frac{1}{2}\ln|\hat{I}_j| + \ln p(y|\hat{\theta}_j, M_j) + \ln p(\hat{\theta}_j|M_j) + O(n^{-1}), \quad (35)$$

where $\hat{\theta}_j$ is either the mode of the posterior distribution $p(\theta_j|y, M_j)$ or the MLE and $\hat{I}_j$ is the observed information matrix evaluated at $\hat{\theta}_j$, and

$$\ln p(y|M_j) = -\frac{1}{2}k_j \ln n + \ln p(y|\hat{\theta}_j, M_j) + O(1), \quad (36)$$

the latter a large-n approximation to the former that is recognizable as the basis of BIC (cf. Equation (15) in the paper under discussion here). Raftery (1995) noted that "it is possible to improve on [Equation (35) in this discussion contribution] in its MLE form by taking a single Newton step toward the posterior mode, starting at the MLE." Equations (35) and (36) above (and the refinement suggested by Raftery 1995) have the advantage of relative computational simplicity and perhaps speed (if reliable maximization software is handy) when compared with some of the ideas explored in the paper under discussion here. Do the authors have any experiences they can share that would shed light on the speed-versus-accuracy tradeoffs inherent in a comparison of their methods with Laplace-based approaches, when both are appropriate to consider?

As is well known, Bayes factors involve comparing quantities of the form

$$\begin{aligned} p(y|M_j) &= \int \left[\prod_{i=1}^{n} p(y_i|\theta_j, M_j)\right] p(\theta_j|M_j)\, d\theta_j \\ &= E_{(\theta_j|M_j)} L(\theta_j|y, M_j), \end{aligned} \quad (37)$$

i.e., Bayes factors are based on comparisons of expectations of likelihoods with respect to the *priors* in the models under comparison, and this is why they behave so unstably as model selection criteria with diffuse priors, as a function of how the diffuseness is specified. Many ad hoc methods for attempting to cope with this instability have by now been suggested, including {partial, intrinsic, fractional} Bayes factors, well calibrated priors, conventional priors, intrinsic priors, expected posterior priors, and so on (e.g., Pericchi 2005); the list seems as endless as its ad-hockery is disheartening. It is arguable (e.g., Draper and Krnjajić 2006) that this is a good reason for shifting attention in Bayesian model specification away from Bayes factors and toward model selection criteria, such as the predictive log score, that do not suffer from such instabilities when diffuse prior information is all that is available; but (a) the dependence of the mixing weights in Bayesian model averaging (BMA) on ratios of Bayes factors and (b) the central role that BMA plays in appropriately propagating model uncertainty combine to leave the impression that evaluating Bayes factors with diffuse prior information cannot, regrettably, be entirely avoided. Do the authors see any possibility that any of the ideas they have explored for stabilizing MCMC-based *estimates* of Bayes factors might be used to help stabilize *the Bayes factors themselves* in the presence of diffuse prior information?

CHRIS SHERLOCK and PAUL FEARNHEAD (*Lancaster University, UK*)

We would like to congratulate the author on a stimulating paper. The goal of a simple and automatic procedure for estimating the integrated likelihood is an important one, and we are pleased to see some further work extending the harmonic mean estimator.

We decided to test out the BICM and AICM procedures described within the paper for the problem of model selection for Markov-modulated Poisson Processes. These are models for the occurence of events through time, and assume an underlying continuous-time Markov process $X(t)$ which has state space $\{1, \dots, K\}$. The data consists of the times of events from a Poisson process of time-varying intensity $\lambda(t)$ which depends on the underyling $X(t)$ process. So conditional on $X(t) = k$ we have $\lambda(t) = \lambda_k$. The parameters of the model are the different intensities $\{\lambda_1, \dots, \lambda_K\}$, and the entries in the rate matrix of $X(t)$.

We applied this model to analyse the occurence of Chi-sites along the lagging strand of the genome of Ecoli (see Fearnhead and Sherlock 2006 for details of this application). For $K > 1$, this is a non-trivial example, for example the likelihood is symmetric across re-labelling of the states of the $X(t)$ process whereas we have used a non-symetric informative prior so that we have a complex multi-modal posterior. However, Fearnhead and Sherlock 2006 describe how to implement a Gibbs sampler for this model, and it is thus possible to use the idea of Chib (1995) to get accurate estimates of the integrated likelihood for $K = 2$ and 3 (the $K = 1$ case can be calculated analytically). Thus we have a “correct” answer to compare the results of BICM and AICM to.

As the paper suggests, calculating BICM or AICM values from MCMC output is simple and quick. The results we obtained, together with twice the integrated likelihood as calculated from the method of Chib (1995) are shown in Table 4. (We give twice the integrated likelihood values as it is this that BICM and AICM estimate.) As suggested by the theory, the plots of the log-likelihood values output from the MCMC run show that these do closely follow a shifted gamma distribution (results not shown).

For the $K = 1$ case BICM appears to give a better estimate of twice the integrated likelihood than AICM, but it gives substantial underestimates for the $K = 2$ and $K = 3$ case. As a result BICM incorrectly shows strong evidence for $K = 1$. By comparison AICM performs well: while it over-estimates twice the integrated likelihood for all three models, the relative estimates are very close to the truth. Based on AICM you would correctly choose $K = 2$, and have appropriate estimates of the strength of evidence for this model over $K = 1$ and $K = 3$ respectively.

For comparison we also tried the harmonic mean estimator, and this estimator also performed very well. Twice the harmonic mean estimates are -935.5, -927.2 and -929.9 for $K = 1$, 2 and 3 respectively. While these again are over-estimates of the true iterated likelihoods, the relative estimates are close to being correct.

Table 4: *BICM, AICM and (twice) integrated likelihood values (*$2\times$ *IL) for the Chi-site data.*

	$K = 1$	$K = 2$	$K = 3$
2× IL	-937.9	-927.7	-931.0
BICM	-937.6	-940.0	-946.3
AICM	-934.9	-925.3	-928.8

Finally we wonder if you could say something about the conditions required for the posterior distribution of the likelihood values (10) to hold - do these require the standard regularity conditions for the likelihood ratio statistic to asymptotically have a χ^2_d distribution? Also is it possible to get a higher-order result for this limiting distribution, which could then be integrated to give a (finite) estimate of the integrated likelihood? Such a result may give theoretical justification for choosing one of AICM or BICM over the other in different scenarios.

I. CLAIRE GORMLEY and T. BRENDAN MURPHY
(*Trinity College Dublin, Ireland*)

We would like to congratulate the authors on their excellent paper on computing integrated likelihoods which is a topic of great importance in Bayesian model comparison.

We would like to discuss our experiences in applying AICM, BICM and alternatives when selecting the dimensionality for a latent space model for rank data.

Background

Irish elections employ a voting system called proportional representation by means of a single transferable vote (PR-STV). In the PR-STV system, voters rank some or all of the candidates in order of preference. Votes are counted and subsequently transferred between candidates, using the voter preferences, in a series of counts to determine who gets elected. More details on the electoral system and the counting of votes is given in Gormley and Murphy (2005).

Hence, Irish elections generate rank data recording the preferences of the voters for candidates. We have recently developed mixture models for modeling rank data in the context of analyzing Irish election data (Gormley and Murphy, 2005) and Irish college application data (Gormley and Murphy, 2006a).

Gormley and Murphy (2006b) develops a latent space model for rank data where voters and candidates are located in a latent space $\mathbf{Z} \subseteq \mathbb{R}^D$. Each voter is located at position z_i $(i = 1, 2, \ldots, M)$ and each candidate at location ζ_j $(j = 1, 2, \ldots, N)$. Let $d(z_i, \zeta_j)$ be the squared Euclidean distance from voter i to candidate j. We would expect that voter i will give high preferences to the candidates that are closest and low preferences to those that are distant.

Let

$$p_{ij} = \frac{\exp\{-d(z_i, \zeta_j)\}}{\sum_{j'=1}^{N} \exp\{-d(z_i, \zeta_j')\}}$$

be the probability of voter i selecting candidate j in first position. These probabilities can be used to give the probability of a vote (i.e. candidate ordering) using the Placket-Luce model (Plackett, 1975)

$$\mathbf{P}\{\mathbf{x}_i|\mathbf{p}\} = \prod_{t=1}^{n_i} \frac{p_{ic(i,t)}}{\sum_{s=t}^{N} p_{ic(i,s)}},$$

where $c(i, 1), c(i, 2), \ldots, c(i, n_i)$ is the ordered list of the candidates selected by voter i and $c(i, n_i + 1), c(i, n_i + 2), \ldots, c(i, N)$ is an arbitrary ordering of the candidates not selected by voter i.

These models are fitted in a Bayesian framework and samples from the posterior distribution are generated using a random walk Metropolis-Hastings algorithm.

Choice of dimensionality

An important issue when fitting these models is the choice of D, the dimensionality of the latent space.

In this discussion, we will concentrate on data from an exit poll taken at the 1997 Irish Presidential Election. We fitted latent space models for D=1, 2 and 3 and computed the AICM, BICM (using equation (23) from the paper), DIC (Spiegelhalter et al, 2002) and the Pritchard et al (2000) criterion for each model; the values obtained are given in Table 5.

Table 5: *The values of AICM, BICM, DIC and Pritchard et al's criterion computed for the 1997 Presidential Election exit poll data for dimensionality D equal to 1, 2 and 3. The model selected is highlighted in bold.*

	AICM	BICM	DIC	Pritchard
$D = 1$	**-20171**	**-19244**	18270	**17855**
$D = 2$	-23217	-21221	17966	18246
$D = 3$	-26621	-23671	**17923**	19270

There is good consistency across the results for the AICM, BICM and the Pritchard et al criterion with each selecting $D = 1$ whereas DIC selects $D = 3$. A simple principal components analysis of the estimated candidate locations suggests that most of the variation in candidate locations is explained by a single dimension which captures the race between McAleese (winner of the election) and the other candidates (Scallon, Banotti, Roche and Nally).

Hence, AICM and BICM select a dimensionality that is consistent with the estimated candidate configurations. This contrasts with DIC which doesn't appear to have a strong enough penalty on dimensionality.

CHRISTIAN P. ROBERT (*Université Paris Dauphine, France*) and
NICOLAS CHOPIN (*University of Bristol, UK*)

Comparison

The issue of approximating marginal densities obviously remains an up-to-date concern for Bayesian Statistics, since two invited talks at this conference are centred around it, namely the present paper and the alternative proposal of Skilling *(this volume)*. While we are not completely convinced of the advantages of nested sampling (see our discussion in this volume), we would welcome the authors' opinion of the respective worths of both approaches. In particular, Skilling's *(this volume)* perspective is completely in line with the second approach of the present paper, that is, based on a (prior or posterior) distribution of the likelihood function, since Skilling's marginal is expressed as $\pi(y) = \mathbb{E}_\pi[L]$, while Raftery et al.'s marginal is $\pi(y) = \mathbb{E}[L^{-1}|y]$.

Potential dangers

Let us first state that we find the representation of Newton and Raftery (1994) quite interesting in that it allows for an approximation of the marginal density based on the output of the MCMC simulation of the posterior $\pi(\theta|y)$ (rather than from the prior as in nested sampling). Its major drawback is however the disastrous feature of a potential infinite variance, against which the Rao-Blackwellised solution proposed in this version does not always work. We can take for instance the case of the normal variance, $y|\sigma \sim \mathcal{N}(0, \sigma^2)$ and $\pi(\sigma^2|y) \sim \mathcal{IG}(3/2, 1 + y^2/2)$ where nested sampling provides an approximation of $\pi(y)$ in agreement with a decrease of the error in $\sqrt{n}$ (see Fig. 7 (left), while the harmonic mean approximation has no variance and obviously varies much more for the same computational effort. A further difficulty is that, in complex settings, the infinite variance of the harmonic estimator may remain undetected. For instance, a run up to $5,000,000$ iterations produces an apparent decrease of the error in $\sqrt{n}$ in Fig. 7 (right).

Bayes factors

A feature also common to both Skilling's and Raftery *et al.* 's approaches is that they do not easily adapt to multiple models environments, as those encountered in model choice and the computation of Bayes factors, for which generic approximations methods like path sampling (Gelman and Meng, 1998) are readily available, being based on simulations from the (alternative) posterior distributions.

Asymptotic approximations

To go back to the Gamma approximation, we are a bit concerned with the $\log(1-\alpha)$ in the expression of $\log \pi(y)$ in Eqn. (14). Would the Gamma approximation in Eqn. (10) be *exact* (by chance, or because the model has some specific structure, *e.g.*, Gaussian), then $\log \pi(y)$ would be infinite. This does not seem intuitive, and it also casts some doubt on the practicality of this approach. Beyond this specific point, one must always be wary of approximations methods that do not provide a way of evaluating, even roughly, the approximation error (like BIC). In that respect,

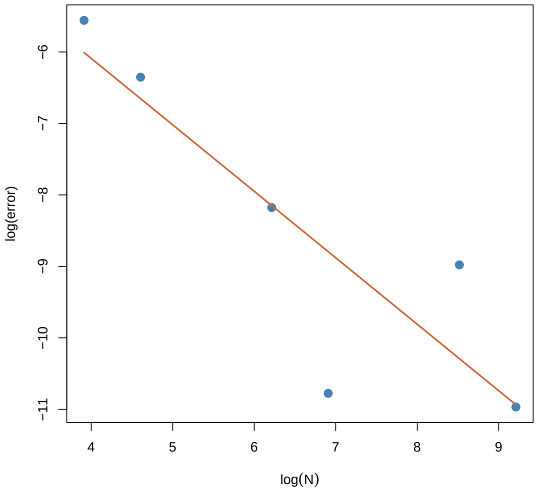

Figure 7: *Comparison of the error evolution for nested sampling and harmonic approximation:* (left) *Evolution of the error in Skilling's nested sampling approximation of the marginal density when* $y = 5$ *in a* $y|\sigma \sim \mathcal{N}(0, \sigma^2)$ *and* $\pi(\sigma^2|y) \sim \mathcal{IG}(3/2, 1 + y^2/2)$ *model, using* N *initial simulations from the exponential prior and* $j = N/2$ *replications.* (right) *Evolution of the error of the harmonic approximation for the same problem, using* 1000 *times more simulations from the posterior than in nested sampling.*

a sequence of increasingly accurate (and possibly increasingly expensive to compute) approximations, would be preferable, as the authors suggest briefly in the conclusion.

Remark

A final bibliographical remark is about the study of DIC in missing data models: Celeux *et al.* (2006) give a detailed analysis of the multiple possible interpolations of p_D and of DIC in such setups, agreeing with the authors about its instability.

REPLY TO THE DISCUSSION

Basic idea

We are very grateful to all the discussants for their stimulating discussions.

The basic idea of our paper is that the integrated likelihood can be estimated from the posterior distribution of the loglikelihood, using the harmonic mean identity. This reduces the problem of estimating the integrated likelihood to a one-dimensional one. The simplest way to do is via the harmonic mean estimator of Newton and Raftery (1994), but while this is simulation-consistent, it often has infinite variance.

We suggested two ways of getting integrated likelihood estimators with better properties. The first is to reduce the parameter space. The second, on which most of the discussants focused, is to model the posterior distribution of the log-likelihood parametrically, estimate the parameters of the resulting model from MCMC or other posterior simulation output, and then apply the harmonic mean identity to the resulting estimated model. We proposed a shifted scaled gamma distribution as a possible approximating model, and showed that it gave reasonable results in some examples.

However, this is not the only approximating model that could be used. One could instead use a shifted scaled noncentral χ^2 distribution; this is exact for Gaussian fixed effects models. One could also use a sum of shifted scaled noncentral χ^2 distributions; this is exact for an ANOVA-type Gaussian random effects model, as shown by Spiegelhalter *et al.* (2002). Our experience so far, including with random effects models, is that the shifted gamma distribution does provide a good approximation in a wide range of situations; this conforms to the experience reported by Sherlock and Fearnhead, for example.

Sherlock and Fearnhead asked whether one could obtain a higher-order expansion for the asymptotic gamma approximation, and then integrate this. Bickel and Ghosh (1990) proposed an expansion of this type. Their expansion couldn't be used directly for this purpose, because the terms in it lead to infinite estimates of the integrated likelihood. However, it does seem possible that an expansion of this kind could be obtained that would be useful in the present context.

A related possibility, not mentioned in our paper or by any of the discussants, would be to approximate the posterior distribution of the loglikelihood by a mixture of normals. Roeder and Wasserman (1997) showed that this can approximate a wide range of distributions, including gamma-like distributions, and proposed ways of estimating this model. This might have the advantage of leading to more stable estimates, as the tail declines quickly. The estimator of the integrated likelihood proposed by Pritchard *et al.* (2000) could be viewed as a special case of such a normal mixture model, with just one term.

Polson discussed some very interesting unpublished work of Wolpert (2002), who indicated that the posterior distribution of the harmonic mean estimator is asymptotically a stable law, with a stable index that is typically just large enough to ensure that the mean exists. This suggests the possibility of using an approximating stable distribution of the reciprocal likelihood (rather than the loglikelihod) as the basis for a stabilized harmonic mean estimator. Wolpert's paper doesn't report numerical experience, but it is worth investigating further. Nevertheless, it seems likely that it would be as hard to estimate the stable index of the stable distribution as it is to estimate the scale parameter of the gamma distribution that we suggest; this may only shift the difficulty to another arena.

The effective number of parameters

Carlin and Spiegelhalter suggest that $\hat{d}$ may be an overestimate of the effective number of parameters, referring to the normal random effects example. However, we introduce $\hat{d}$ primarily as an estimate of (twice) the scale parameter in the approximating gamma distribution rather than of the actual number of parameters; in regular fixed effects models these coincide asymptotically. However, even in fixed effects models they don't always coincide: for example in the exponential distribution example discussed by Fan *et al.* (2000), the number of degrees of freedom in the χ^2 distribution that approximates the posterior distribution of the loglikelihoods is twice the number of parameters.

In order to investigate this further, we simulated some data and posterior distributions from the normal random effects model discussed by Carlin and Spiegelhalter, as described by Spiegelhalter *et al.* (2002). For this model, p_D does not depend on the data, whereas $\hat{d}$ does. Carlin and Spiegelhalter suggested that $\hat{d}$ will tend to overestimate the effective number of parameters relative to p_D, but we found that $\hat{d}$ was often actually slightly smaller than p_D.

In this model, the true $\ell_{\max}$ is known, and this makes it easier to estimate both α and λ in the approximating gamma distribution, using a moment estimator, for example. We found that in this situation, moment estimators of the gamma distribution parameters were very similar to maximum likelihood or other estimators. This provides an alternative to both p_D and $\hat{d}$.

We found that (i) the shifted gamma distribution continued to fit well for a wide variety of situations including ones with large shrinkage; (ii) our moment estimates of α tended to be *larger* than either $p_D/2$ or $\hat{d}/2$; and (iii) the gamma distribution tended to fit the observed distribution better with the moment estimates that with either $\hat{d}$ or p_D. When n_i, the sample size in the i-th group, was the same for all groups, we found that $-\log(1-\hat{\lambda})$ was well approximated by $\log(n_i+1)$, lending some support to the random effects BICM approximation (22).

Theory

Carlin and Spiegelhalter asked what the theoretical rationale for BICM is, and, in a related question, Sherlock and Fearnhead asked what regularity conditions are needed for the asymptotic posterior distribution of the loglikelihood (10) to hold. The regularity conditions are given by Fan *et al.* (2000); these are weaker than the conditions required for the asymptotic χ^2 distribution of the likelihood ratio test statistic to hold. In particular, the contours of the loglikelihood do not need to be asymptotically elliptical.

The theoretical rationale for BICM is outlined in Section 4.1 of our paper. In brief, Raftery (1995, Section 4.1), drawing heavily on Kass and Wasserman (1995), showed that with a unit information prior, the integrated likelihood for a fixed effects model can be approximated using BIC, with error of order $O(n^{-1/2})$. BIC involves the maximized loglikelihood, $\ell_{\max}$, and the number of parameters, d. We assume that neither of these is available explicitly, and that they must be estimated from posterior simulation. Under the weak regularity conditions we have mentioned, the maximized loglikelihood (suitably normalized) can be estimated consistently from posterior simulation output by $\hat{\ell}_{\max}$, and the number of parameters can be consistently estimated by $\hat{d}$. Thus BICM provides a consistent estimator of BIC, and hence of twice the integrated likelihood.

This theoretical rationale applies to fixed effects models. However, Pauler (1998) has provided a theoretical derivation of an extension of BIC to random effects models, and it seems that this could be the basis for a theoretical rationale for BICM also for random effects models. More work on this is required, however.

Chopin and Robert asked what happens if the gamma approximation is exact. However, this will be the case only with an improper prior, and in that case the integrated likelihood is undefined in any event. Chopin and Robert also asked about standard errors for BICM and AICM. These are now in the paper (they were not in the version of the paper presented at the Valencia meeting).

The harmonic mean estimator

Our work was motivated by the observation that the harmonic mean estimator of the (reciprocal) integrated likelihood has infinite variance, a fact pointed out in the paper that introduced it (Newton and Raftery 1994). In spite of this apparently undesirable behavior, the original harmonic mean estimator has been widely used, and several researchers have reported satisfactory results with it, including Sherlock and Fearnhead in their discussion here.

The results of Wolpert (2002), reported by Polson, shed some light on this. Wolpert showed that the harmonic mean estimator does have a distribution. In spite of having infinite variance, most of the mass of this distribution may be fairly concentrated. If one is comparing two models, and the distributions of the harmonic mean estimators of their integrated likelihoods have little overlap, then the model comparison is clear. This is often the case. A simple way of assessing these distributions is by estimating the harmonic mean estimator from replicated posterior simulation runs, as done for example by Zijlstra *et al.* (2005).

The Wolpert results suggest that the distribution of the harmonic mean estimator from a subset of a posterior simulation run will be similar to that from the run as whole. Thus a reasonable approach could be to divide a posterior simulation run (suitably thinned) into, say, 9 or 19 batches, and use the resulting distribution as a rough sampling distribution of the harmonic mean estimator itself. This should be good enough at least for assessing whether the result of a comparison between two models is decisive. Thus this provides some theoretical support for continued use of the harmonic mean estimator in practice, at least in some situations.

Experience with BICM and AICM

Gormley and Murphy reported positive experience with BICM and AICM, and we were glad to see their results.

Sherlock and Fearnhead gave results for a modulated Poisson process in which accurate values of the integrated likelihood were available. AICM and the harmonic mean estimator worked well, but BICM did not. In particular, BICM supported a one-state model over a two-state model, unlike the true integrated likelihood. In the version of the paper on which they commented and that was presented at Valencia, only the fixed effects version (18) of BICM was given explicitly. If instead we use the version (22), with $n_k = n/K$, which seems more appropriate, then the second line of their table (for BICM) becomes −937.6, −936.6, −939.6. BICM is now more accurate than before and in particular supports the two-state model. It is still not fully satisfactory for this example, however.

Comparisons with other methods

Several discussants asked about comparisons with other methods. In our paper we gave a fairly extensive literature review of relevant methods. However, our goal here is specific: a generic method for estimating the integrated likelihood of a model from posterior simulation output that uses only the loglikelihoods of the simulated parameter values, and in particular does not involve model-specific algebra or computation. Several of the methods mentioned do not fall into this category, and so we have not viewed them as directly comparable for the present purposes.

Draper asked about a comparison with the Laplace method. The Laplace method can be efficient and highly accurate, but it is not generic, and indeed applying it to a specific model can require a great deal of work. Polson mentioned his very interesting work on estimation of the maximized likelihood and extensions of the Savage density ratio method. These methods are generally for computing ratios of integrated likelihoods rather than individual ones, and Polson's work is for specific models, and so has a somewhat different goal from ours.

Chopin and Robert asked about a comparison between our approach and that of Skilling in the present volume. Skilling's approach is also generic, and so comparable to ours. We agree with Chopin and Robert's comments. Skilling's method involves integrating over the prior, and in the low-dimensional examples he gives, this works well. However, in the much higher-dimensional models commonly tackled using MCMC, this may well not work so well. We would be interested to see further experience with this proposal.

Finally, Chopin and Robert suggested that our approach was not well suited to situations with multiple models. However, it yields an estimate of the integrated likelihood for each model estimated, and so provides a ready way to compare all the models estimated in a round of data analysis. Thus it does seem well adapted to situations where multiple models are being fitted.

ADDITIONAL REFERENCES IN THE DISCUSSION

Besag, J., York, J. C. and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Ann. Inst. Statist. Math.* **43**, 1–59, (with discussion).

Celeux, G., Forbes, F., Robert, C. P. and Titterington, D. M. (2006). Deviance information criteria for missing data models. *Bayesian Analysis*, (with discussion) (to appear).

Clayton, D. G. and Kaldor, J. M. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics* **43**, 671–681.

Draper, D. (1995). Assessment and propagation of model uncertainty. *J. Roy. Statist. Soc. B* **57**, 45–97 (with discussion).

Draper, D. and Krnjajić, M. (2006). Bayesian model specification. Submitted.

Fearnhead, P. and Sherlock, C. (2006). An exact Gibbs sampler for the Markov modulated Poisson process. *J. Roy. Statist. Soc. B* **68**, 767–784.

Gormley, I. C. and Murphy, T. B. (2005). Exploring Heterogeneity in Irish Voting Data: A Mixture Modeling Approach. *Tech. Rep.*, 05/09, Department of Statistics, Trinity College Dublin, Ireland.

Gormley, I. C. and Murphy, T. B. (2006a). Analysis of Irish Third-Level College Applications Data. *J. Roy. Statist. Soc. A* **169**, 361–380.

Gormley, I. C. and Murphy, T. B. (2006b). A Latent Space Model for Rank Data. *Tech. Rep.*, 06/02, Department of Statistics, Trinity College Dublin, Ireland.

Hodges, J. S. and Sargent, D. J. (2001). Counting degrees of freedom in hierarchical and other richly-parameterised models. *Biometrika* **88**, 367–379.

Jacquier, E. and Polson, N. G. (2002) Odds Ratios for Stochastic Volatility Models. *Tech. Rep.*, University of Chicago, USA.

Jacquier, E., Johannes, M. and Polson, N. G. (2006) MCMC Maximum Likelihood for Latent State Models. *J. Econometrics* (to appear).

Lu, H., Hodges, J. and Carlin, B. P. (2004). Measuring the complexity of generalized linear hierarchical models. *Tech. Rep.*, 2004–002, Division of Biostatistics, University of Minnesota. Under revision for *Can. J. Statist.*

Pericchi, L. (2005). Unifying concepts for methods of Bayesian model choice. Submitted.

Plackett, R. L. (1975). The analysis of permutations. *Appl. Statist.* **24**, 193–202.

Raftery, A. E. (1995). Discussion of "Assessment and propagation of model uncertainty," by Draper D, *J. Roy. Statist. Soc. B* **57**, 78–79.

Vaida, F. and Blanchard, S. (2005). Conditional Akaike information for mixed-effects models. *Biometrika* **92**, 351–370.

Wolpert, R. (2002). Stable Limit laws for Marginal Probabilities from MCMC streams: Acceleration of Convergence. *Tech. Rep.*, ISDS, Duke University, USA.