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

Adrian E. Raftery
University of Washington

Michael A. Newton
University of Wisconsin

Jaya M. Satagopan
Memorial Sloan-Kettering Cancer Center

Pavel N. Krivitsky
University of Washington

Working Paper no. 60
Center for Statistics and the Social Sciences
University of Washington
Seattle, Wash., USA

April 13, 2006

Abstract

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. 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 illustrate the proposed methods through several examples. One of these is the selection of the dimension for the latent space social network model of Hoff, Raftery and Handcock (2002). When applied to the well-known monks' social network data of Sampson (1968), our methods yield a surprising result: we find that the monks' social network can be well represented by a latent space model with just one dimension.

Contents

List of Tables

List of Figures

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,

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

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 $\mathrm{Var}\{[\pi(y|\theta)]^{-1}|y\}$) 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 log-likelihood 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 using only the likelihoods from the posterior simulation, 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 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 $\mathrm{Var}\{[\pi(y|\theta)]^{-1}|y\}$. This variance, in turn, is determined by the second noncentral moment $\mathrm{E}\{[\pi(y|\theta)]^{-2}|y\}$ 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.

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 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 prior predictive density. 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 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 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.

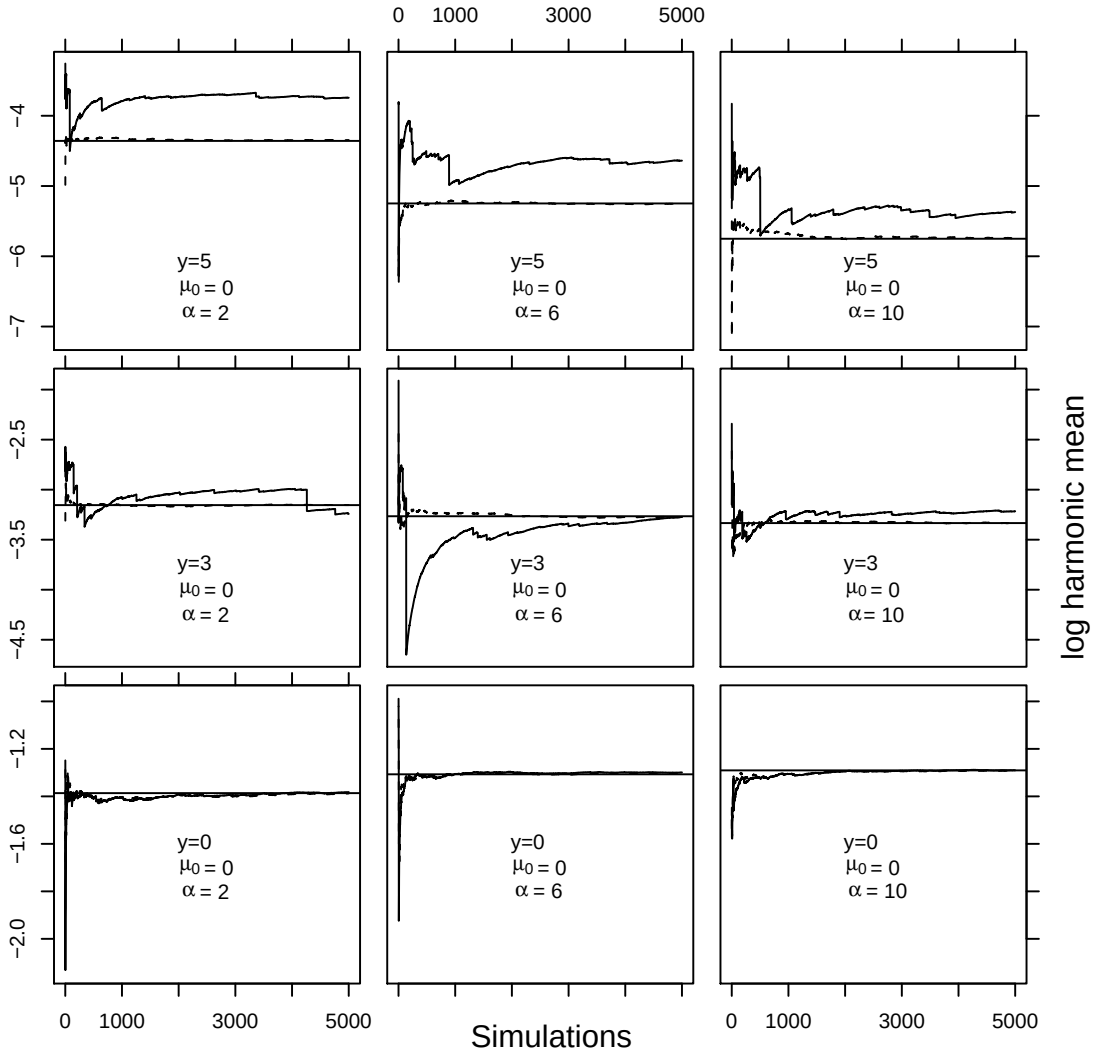

Figure 1: Normal (bold line) and stabilized t-based (dotted line) harmonic mean estimates of the log integrated likelihood compared with the true value (dashed line), when the data y follow a univariate normal distribution as described in Section 2. The estimate based on the first B values simulated 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.

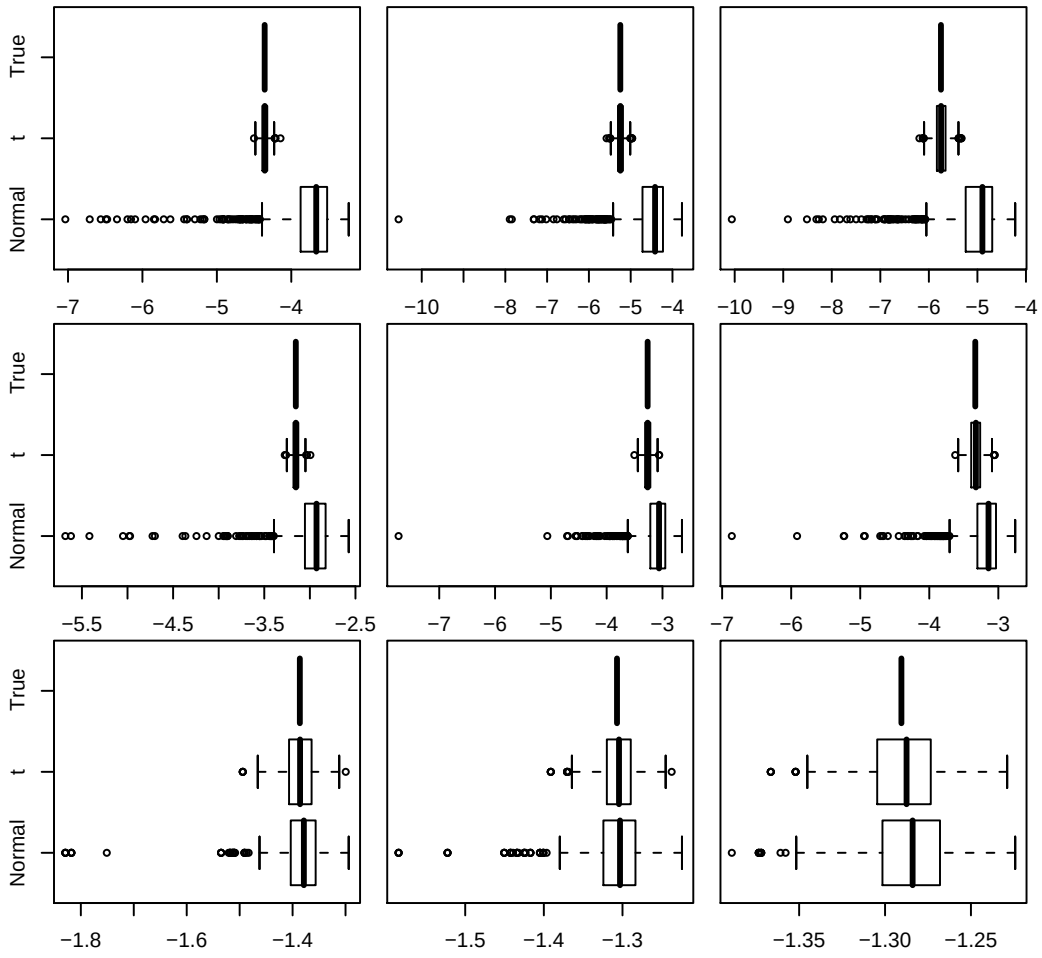

Figure 2: Boxplots to assess the variability of the estimated integrated likelihood. Shown are the true integrated likelihood, and the normal and stabilized t-based harmonic mean estimators, both on the logarithmic scale. The estimates are obtained from 1000 Monte Carlo samples of size 1000. These estimates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

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. 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	0.79	0.90	0.94
		(5.46)	(10.38)	(13.32)	(15.88)
(5, 0, 6)	190.19	0.50	0.81	0.90	0.95
		(23.87)	(45.36)	(58.22)	(69.37)
(5, 0, 10)	314.38	0.53	0.78	0.88	0.93
		(62.44)	(118.64)	(152.27)	(181.44)
(3, 0, 2)	23.44	0.49	0.82	0.90	0.95
		(1.29)	(2.44)	(3.14)	(3.74)
(3, 0, 6)	26.20	0.49	0.78	0.89	0.93
		(2.41)	(4.57)	(5.87)	(6.99)
(3, 0, 10)	28.05	0.48	0.79	0.88	0.93
		(3.57)	(6.78)	(8.71)	(10.37)
(0, 0, 2)	4.00	0.47	0.79	0.90	0.93
		(0.17)	(0.32)	(0.41)	(0.49)
(0, 0, 6)	3.70	0.48	0.77	0.87	0.93
		(0.12)	(0.22)	(0.28)	(0.34)
(0, 0, 10)	3.63	0.47	0.81	0.86	0.93
		(0.12)	(0.22)	(0.28)	(0.34)

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

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

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), for example, is a useful review from a statistical perspective. We reconsider the particular model

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

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. (Note that we can condition on marker information m because its marginal distribution $\pi(m)$ is not dependent 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:

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

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

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

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

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

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, $\mathrm{St}_n(y|\mu+\alpha^{'}g,I,\zeta)$.

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. 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 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, these (p_i) 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

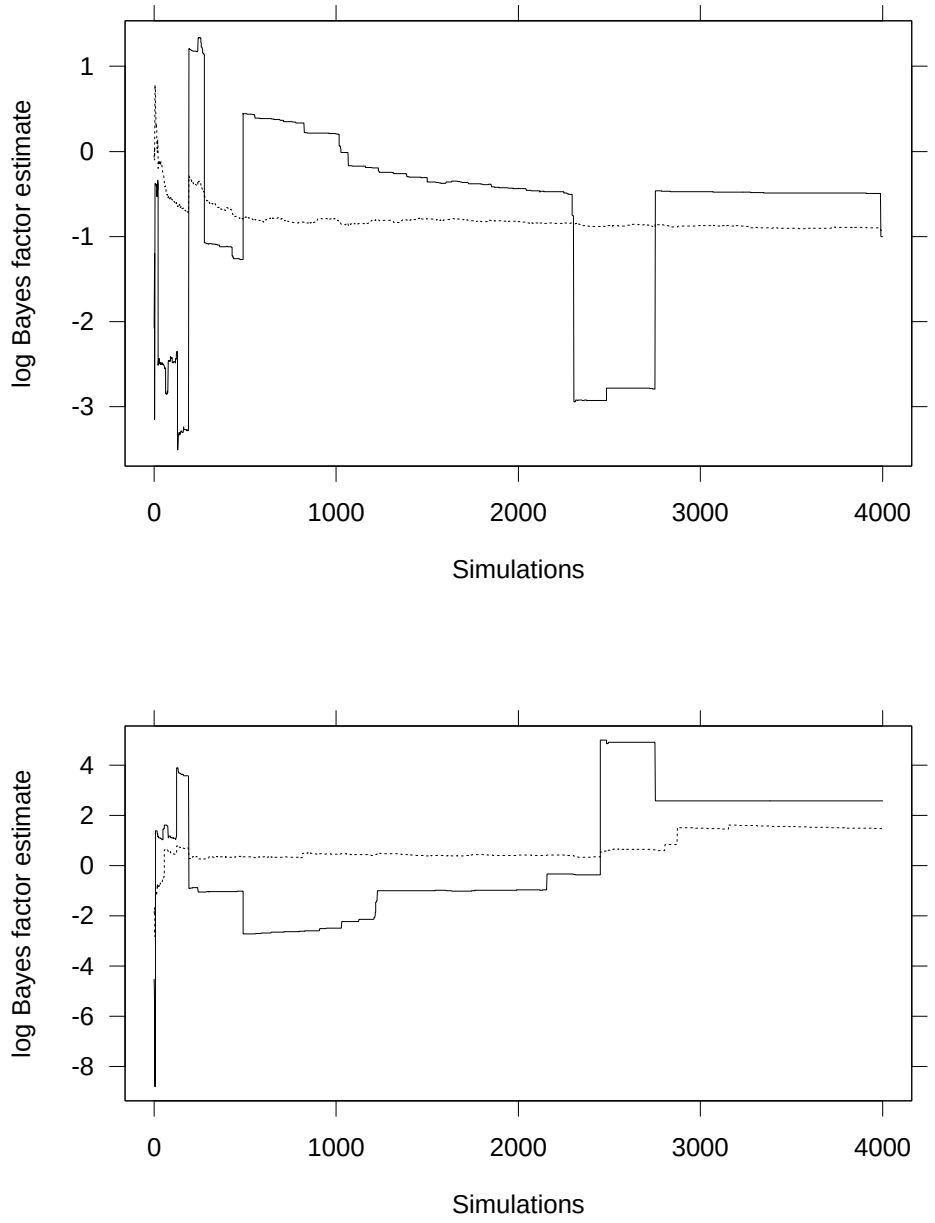

Figure 3: Log Bayes factor Estimates for the Flowering Time Data, based on MCMC. The log Bayes factor based on the first B saved scans of the MCMC run is plotted against B. The comparison between one locus and two loci models is shown on the top. The bottom figure corresponds to the comparison between the two 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.

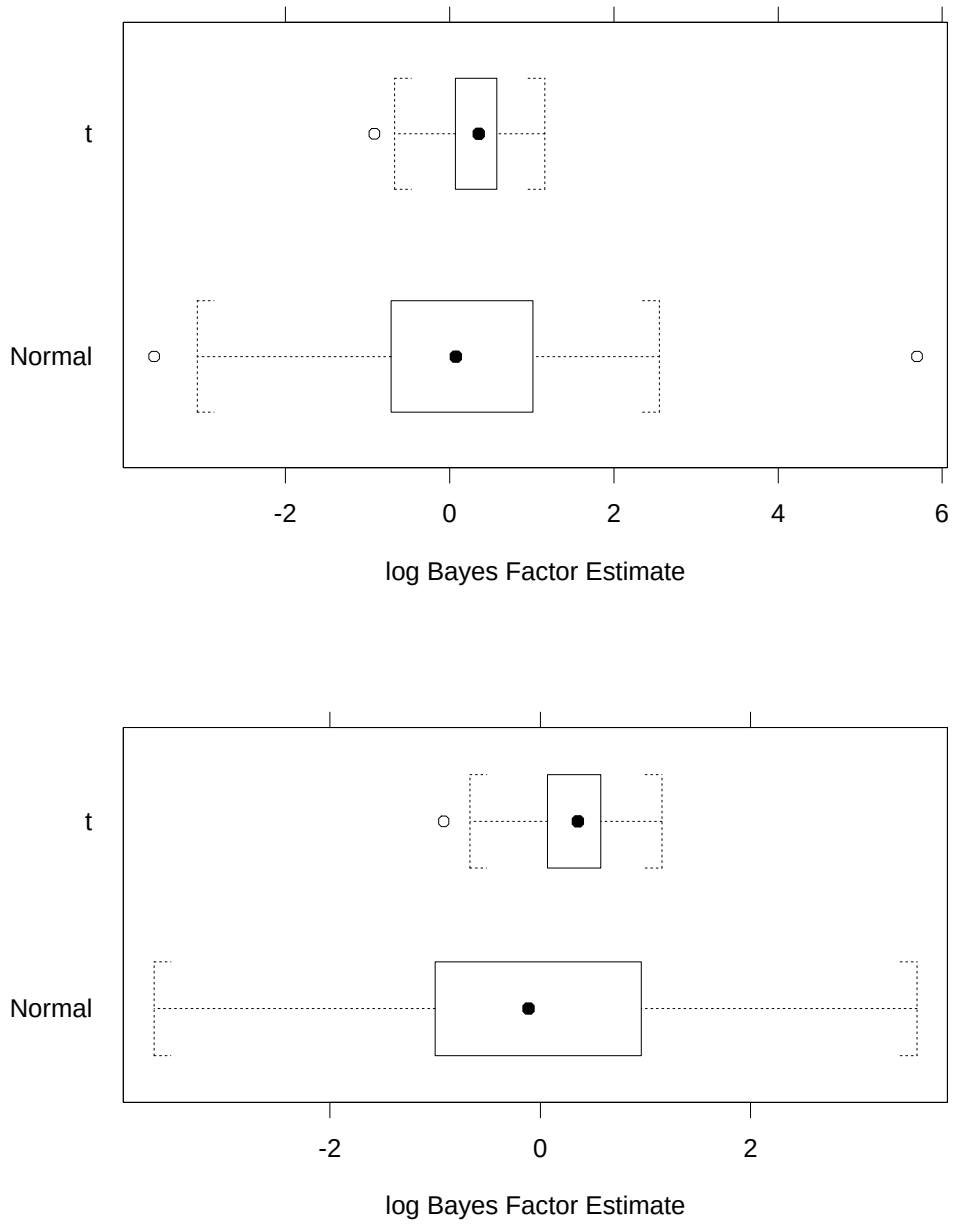

Figure 4: Assessing the Variability of the Log Bayes Factor Estimates for the Flowering Time Data, using 75 replications of the MCMC run. The top panel shows the comparison between the one-locus and two-loci models, and the bottom panel shows the comparison between the two-loci and three-loci models. 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.

out both (p_i) and the hyperparameters a and b. It is routine to sample the full parameter set $\theta = (p, a, b)$ from its posterior distribution (Gelman, Carlin, Stern, and Rubin 1996). 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 $a > \max(y_i)$ 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

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

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

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

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, to be 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

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

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

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

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 $\hat{\alpha} = s^2_\ell$ and $\hat{\ell}_{\max} = \bar{\ell} + s^2_\ell$, where $\bar{\ell}$ and s^2_ℓ 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 so 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 $\max_t \ell_t > \hat{\ell}_{\max}$ 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

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

In practice, λ will be close to 1, but slightly 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 very close to 1, and 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

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

and by analogy we define

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

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

$$\begin{aligned} \log \hat{\pi}_{\mathrm{BICM}}(y) &= \hat{\ell}_{\max} - \frac{\hat{d}}{2}\log(n) && (17) \\ &= \bar{\ell} - s_\ell^2\left(\log(n) - 1\right). && (18) \end{aligned}$$

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) have shown in another context that when different choices are possible, they each give valid approximations to the integrated likelihood, corresponding to different unit information

priors, that differ 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) give an example of one way of determining this, and another example in a hierarchical model is given in Section 4.3. Pauler (1998) in her equation (11) has proposed a modified definition for hierarchical models, called S_M, and showed its validity in her Theorem 2. In this approach each parameter has potentially has a different "n" associated with it, corresponding to the number of data points involved in estimating it.

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

$$\text{AIC} = 2\ell_{\max} - 2d, \tag{19}$$

which we can estimate by

$$\begin{aligned} \text{AICM} &= 2\hat{\ell}_{\max} - 2\hat{d} && (20)\\ &= 2\hat{\ell}_{\max} - 4s^2_\ell && (21)\\ &= 2(\bar{\ell} - s^2_\ell). && (22) \end{aligned}$$

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.

As we have noted, the moment estimator of α implies that $\hat{d} = 2s^2_\ell$ 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, $p_D = 2(\log \pi(\bar{\theta}|y) - \bar{\ell})$. 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) gives 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.

When our estimator $\hat{d}$ is replaced by p_D, $\hat{\ell}_{\max}$ is equivalent to the estimator given in equation (19) of Spiegelhalter et al. (2002), and AICM becomes equivalent to DIC as defined by Spiegelhalter et al. (2002), although the derivations are different.

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 $r = \frac{1}{2}$, $a_n = O(n^{-\frac{1}{2}})$, and $S = \{\theta : \theta^T \Sigma \theta\}$ where Σ is the Fisher information matrix, so that S is an ellipse. When this is not the case they say that the distribution is "fan-shaped." They show that in general under these conditions

$$\ell_{\max} - \ell_t \sim \text{Gamma}(rd, 1). \tag{23}$$

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

They give several simple examples where this is not the case. 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 1 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 loglikelihoods (12) becomes approximately normal. The posterior distribution of the reciprocal of the likelihood is then approximately lognormal, leading to the estimator

$$\log \hat{\pi}_{\text{LN}}(y) = \bar{\ell} - \frac{1}{2} s_\ell^2. \tag{24}$$

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

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

so that $\log \hat{\pi}_{\mathrm{LN}}(y)$ 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 our estimates $\hat{d}$, $\hat{\ell}_{\max}$ and $\pi_{\mathrm{BICM}}(y)$, 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 $\mathrm{MVN}_d(\mu, I)$ random vectors, and the prior for μ is $\mu \sim \mathrm{MVN}_d(0, I)$. The sufficient statistic is then just the d-dimensional $\bar{y} \sim \mathrm{MVN}_d(\mu, I/n)$. We simulated values of μ from its posterior distribution $\mu|y \sim \mathrm{MVN}_d(n\bar{y}/(n+1), I/(n+1))$. 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.

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 extremely 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 almost perfect. Finally, the lower right panel shows the approximated and true log integrated likelihoods; again the estimation was extremely 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 (17) for $\hat{\pi}_{\text{BICM}}(y)$. However, this would be a model-specific adjustment and take us beyond the generic estimates that we are aiming for here.

4.3 Example: Latent Space Models for Social Networks

Social network data consists 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 the latent positions, with

$$\log\left(\frac{\Pr(y_{ij}=1)}{\Pr(y_{ij}=0)}\right) = \beta - |z_i - z_j|, \tag{25}$$

$$z_i \overset{\text{iid}}{\sim} \text{MVN}_q(0, \sigma^2 I). \tag{26}$$

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 $\sigma^2 \sim \sqrt{10}$ Inverse χ^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.

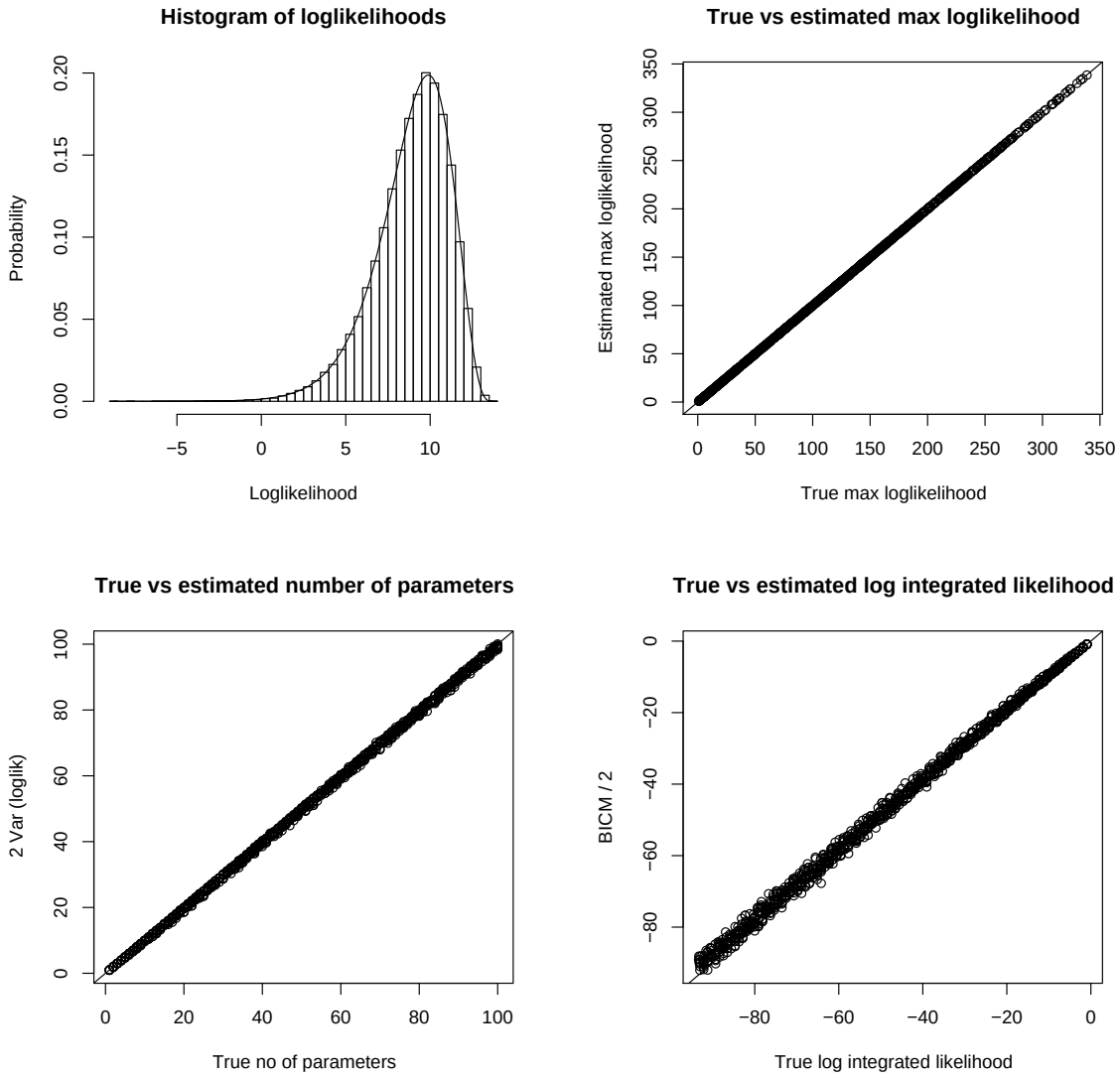

Figure 5: Multivariate normal simulation study of the shifted gamma estimates. Upper left: Histogram of the loglikelihoods for one dataset with $d = 10$ parameters and $n = 100$ data points, with the fitted gamma density superimposed. Upper right: The estimated maximum achievable loglikelihood, $\hat{\ell}_{\max}$, 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.

Table 2: Comparing Dimensions in the Latent Space Model Using the Integrated Likelihood. 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.

q	$\hat{\ell}_{\max}$	$\hat{d}$	$\log \hat{\pi}_{\mathrm{BICM}}(y)$	$\ell_{\max}$	# par
1	–128.6	20.4	–158.1	–129.1	20
2	–109.6	38.0	–164.6	–110.3	38
3	–87.8	66.1	–183.4	–89.9	56
4	–79.3	78.6	–192.9	–73.3	74

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 are traditionally 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 forming $\hat{\pi}_{\mathrm{BICM}}(y)$, n is taken to be the number of actors, 18, rather than the number of links (88), or the number of possible links (306). This corresponds to a unit information prior, where a unit of information is thought of as the information that would be gleaned by observing all the links of one actor. See Volinsky and Raftery (2000) for analogous reasoning in a different setting. In fact, the same model would be chosen no matter which of these values of n was chosen.

The results are shown in Table 2. In addition to our estimates, 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 exactly the same as the number involved in the MCMC in this kind of hierarchical latent variable model, but in this case they do seem to line up rather well.

According to the $\hat{\pi}_{\mathrm{BICM}}(y)$ estimate of the integrated likelihood, the preferred latent space model for these data is a one-dimensional one. This is somewhat surprising at first sight, as these data have usually been represented in two dimensions for visualization purposes. But Figure 6 helps to show why this might be a reasonable choice 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

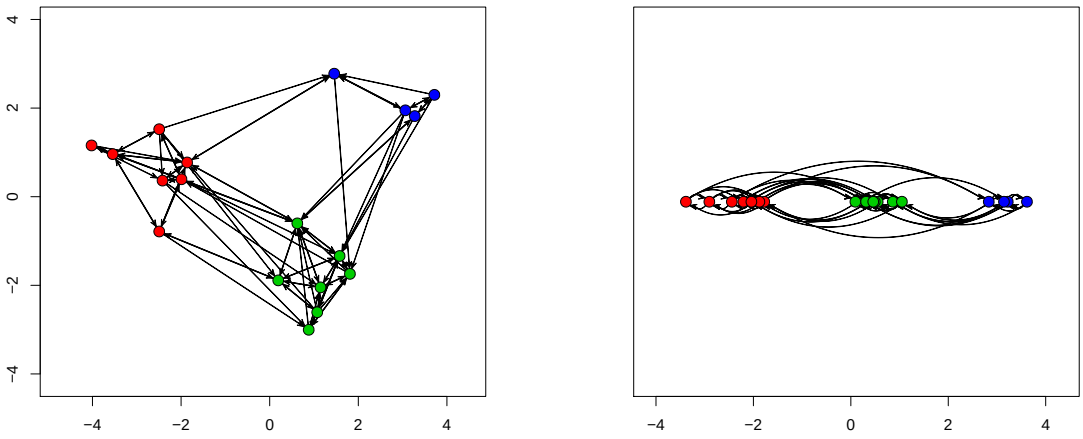

Figure 6: Estimated Latent Positions of Monks in Social Network Example. Left panel: Two-dimensional latent positions with links also shown. Right panel: One-dimensional latent positions. In both plots, the known groupings of the monks are shown: Red = Loyal opposition; Green = Young Turks; Blue = Outcasts.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

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.

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 the main features of the data, as suggested by the integrated likelihoods. This one-dimensional representation of the data has not been noted before as far as we know, and in a personal communication Mark Handcock pointed out to us that it is in line with the detailed ethnographic description of the monks' dynamics in Sampson (1968).

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. In this article 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 likelihood by a 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.

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) suggests that acceptable choices may be possible by considering what a reasonable choice of a unit of information for a unit information prior would be. 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 did not satisfy 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 et al (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 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.

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 transdimensional 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 research of Raftery and Krivitsky was supported by NIH grant 8R01EB 002137-02. The authors are grateful to Mark Handcock and Marijtje van Duijn for extensive and helpful discussions. They are also grateful to Peter Hoff and Matthew Stephens for useful comments.

Appendix I: Student's t

Student t

Copying Bernardo and Smith (1994, page 122),

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

where

$$c = \frac{\Gamma((\alpha+1)/2)}{\Gamma(\alpha/2)\ \Gamma(1/2)}\left(\frac{\lambda}{\alpha}\right)^{1/2}.$$

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

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

Set

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

It can be easily shown that the maxima 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

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

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 $1 \le a(\mu) \le a(\mu^*)$, 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 $a \le b$. 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 $a(\alpha) \le b(\alpha)$ 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 $a(\alpha) \le b(\alpha)$ if

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

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

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csáski (Eds.), *Proceedings of the Second International Symposium on Information Theory*, pp. 267–281. Budapest: Akadémiai Kiadó.

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

Andrews, D. F. and C. L. Mallows (1974). Scale mixtures of normality. *Journal of the Royal Statistical Society, Series B 36*, 99–102.

Bartolucci, F., L. Scaccia, and A. Mira (2006). Efficient Bayes factor estimation from the reversible jump output. *Biometrika 93*, 41–52.

Basu, S. and S. Chib (2003). Marginal likelihood and Bayes factors for Dirichlet process mixture models. *Journal of the American Statistical Association 98*, 224–235.

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

Bos, C. S. (2002). A comparison of marginal likelihood computation methods. Discussion Paper 2002-084/4, Tinbergen Institute, Faculty of Economics and Business Administration, Vrije Unversiteit Amsterdam.

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

Celeux, G., M. Hurn, and C. Robert (2000). Computational and inferential difficulties with mixture posterior distribution. *Journal of the American Statistical Association 95*, 957–970.

Chen, M.-H. (1994). Importance-weighted marginal Bayesian posterior density estimation. *Journal of the American Statistical Association 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. *Journal of the American Statistical Association 90*, 1313–1321.

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

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

Chipman, H., E. I. George, and R. E. McCulloch (2001). The practical implementation of Bayesian model selection. In *Model Selection*, Volume 38 of *IMS Lecture Notes – Monograph Series*, pp. 65–134. Institute of Mathematical Statistics.

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

Dawid, A. P. (1991). Fisherian inference in likelihood and prequential frames of reference (with Discussion). *Journal of the Royal Statistical Society, Series B 53*, 79–109.

DiCiccio, T. J., R. E. Kass, A. E. Raftery, and L. Wasserman (1997). Computing Bayes factors by combining simulation and asymptotic approximation. *Journal of the American Statistical Association 92*, 903–915.

DiCiccio, T. J., R. E. Kass, A. E. Raftery, and L. Wasserman (1997). Computing Bayes factors by combining simulation and asymptotic approximations. *Journal of the American Statistical Association 92*, 903–915.

Diebolt, J. and C. P. Robert (1994). Bayesian estimation of finite mixture distributions. *Journal of the Royal Statistical Society, Series B 56*, 363–375.

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

Fan, J., H.-N. Hung, and W.-H. Wong (2000). Geometric understanding of likelihood ratio statistics. *Journal of the American Statistical Association 95*, 836–841.

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

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (1996). *Bayesian Data Analysis*. Chapman and Hall.

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

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

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

Handcock, M. S., A. E. Raftery, and J. Tantrum (2005). Model-based clustering for social networks. Working Paper 46, Center for Statistics and the Social Sciences, University of Washington.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky (1999). Bayesian model averaging: A tutorial (with discussion). *Statistical Science 15*, 193–195. Correction: vol. 15, pp. 193-195. The corrected version is available at http://www.stat.washington.edu/www/research/online/hoeting1999.pdf.

Hoff, P., A. E. Raftery, and M. S. Handcock (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association 97*, 1090–1098.

Johnson, V. E. (1999). Posterior distributions on normalizing constants. Working Paper 98–26, Institute for Statistics and Decision Sciences, Duke University.

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

Kass, R. E. and L. Wasserman (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association 90*, 928–934.

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

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

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

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

Nandram, B. and H. Kim (2002). Marginal likelihood for a class of Bayesian generalized linear models. *Journal of Statistical 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 A. E. Raftery (1994). Approximate Bayesian inference by the weighted likelihood bootstrap (with discussion). *Journal of the Royal Statistical Society, Series B 56*, 3–48.

Oh, M.-S. (1999). Estimation of posterior density functions from a posterior sample. *Computational Statistics and Data Analysis 29*, 411–427.

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

Pritchard, J. K., M. Stephens, and P. Donnelly (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. *Annals of Statistics 31*, 1695–1731.

Richardson, S. (2002). Discussion of Spiegelhalter et al. *Journal of the Royal Statistical Society, Series B 64*, 626–627.

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

Sampson, S. F. (1968). A novitiate in a period of change: An experimental and case study of relationships. Unpublished Ph. D. dissertation, Department of Sociology, Cornell University.

Satagopan, J. M., B. S. Yandell, M. A. Newton, and T. C. Osborn (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. *Annals of Statistics 6*, 497–511.

Sinharay, S. and H. S. Stern (2005). An empirical comparison of methods for computing Bayes factors in generalized linear mixed models. *Journal of Computational and Graphical Statistics 14*, 415–435.

Sisson, S. A. (2005). Transdimensional Markov chains: A decade of progress and future perspectives. *Journal of the American Statistical Association 100*, 1077–1089.

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. van der Linde (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B 64*, 583–639.

Steele, R., A. E. Raftery, and M. Emond (2006). Computing normalizing constants for finite mixture models via incremental mixture importance sampling (IMIS). *Journal of Computational and Graphical Statistics*, to appear.

Stephens, M. (2000). Dealing with label-switching in mixture models. *Journal of the Royal Statistical Society, Series B, Methodological 62*, 795–809.

Verdinelli, I. and L. Wasserman (1995). Computing Bayes factors using a generalization of the Savage-Dickey density ratio. *Journal of the American Statistical Association 90*, 614–618.

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

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

Neerlandica 59, 107–118.