

# Hypothesis Testing and Model Selection Via Posterior Simulation

Adrian E. Raftery  
University of Washington \*

## 1 Introduction

To motivate the methods described in this chapter, consider the following inference problem in astronomy (Soubiran, 1993). Until fairly recently, it has been believed that the Galaxy consists of two stellar populations, the disk and the halo. More recently, it has been hypothesized that there are in fact three stellar populations, the old (or thin) disk, the thick disk, and the halo, distinguished by their spatial distributions, their velocities, and their metallicities. These hypotheses have different implications for theories of the formation of the Galaxy. Some of the evidence for deciding whether there are two or three populations is shown in Figure 1, which shows radial and rotational velocities for  $n = 2,370$  stars.

A natural model for this situation is a mixture model with  $J$  components, namely

$$y_i = \sum_{j=1}^J \rho_j f(y_i|\theta_j) \quad (i = 1, \dots, n), \quad (1)$$

where  $y = (y_1, \dots, y_n)$  are the observations,  $\rho = (\rho_1, \dots, \rho_J)$  is a probability vector,  $f(\cdot|\theta)$  is a probability density for each  $\theta$ , and  $\theta_j$  is the parameter vector for the  $j$ th component. The inference problem can then be cast as one of choosing between the two-component and the three-component models, i.e. between  $J = 2$  and  $J = 3$  in equation (1).

Standard frequentist hypothesis testing theory does not apply to this problem because the regularity conditions that it requires do not hold, and because the two- and three-component models are not nested (Titterton, Smith and Makov, 1985). Various *ad-hoc* adjustments have been proposed, however. The standard Bayesian solution consists of calculating the Bayes factor for the two-component model against the three-component model, but, as far as I know, this has not been worked out analytically, and it is hard. A

---

\*This research was supported by ONR contract no. N00014-91-J-1074. I am grateful to Gilles Celeux for helpful discussions, to Caroline Soubiran for sharing her data, to Julian Besag for pointing out equation (15), to Jon Wellner for pointing out a useful reference, and to Wally Gilks, Andrew Gelman, Steven Lewis and David Madigan for helpful comments.

## Velocities of 2370 stars in the Galaxy

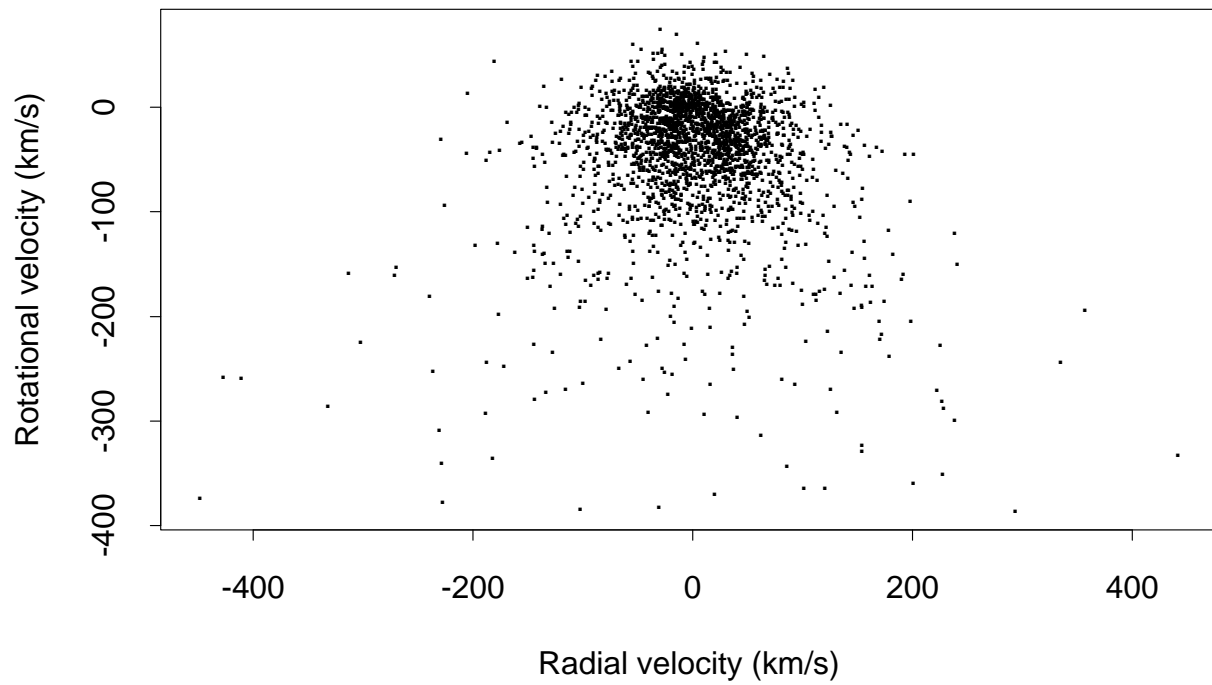


Figure 1: Radial and rotational velocities for 2,370 stars in the Galaxy. Source: Soubiran (1993).

heuristic approximation based on cluster analysis has been proposed (Banfield and Raftery, 1993). Posterior simulation via MCMC is a good way to *estimate* the parameters in equation (1). In this chapter, I will describe ways of calculating Bayes factors from posterior simulation output that also allow MCMC to be used for *testing*.

In Section 2, the Bayesian approach to hypothesis testing, model selection and accounting for model uncertainty is reviewed. These are all based on Bayes factors. The Bayes factor is the ratio of the marginal likelihoods under the two competing models, i.e. the probabilities of the data with all the model parameters integrated out. Thus the marginal likelihood is the key quantity needed for this approach. In Sections 3 and 4, I review ways of estimating the marginal likelihood of a model from posterior simulation output. In Section 3, the harmonic mean estimator and other importance sampling and related estimators are reviewed. In Section 4, two new estimators, the Laplace-Metropolis estimator and the Candidate's estimator, are introduced. These both consisting of adjusting the maximized likelihood, or another likelihood value. In Section 5, I return to the mixture problem and apply the methods described to the motivating problem from astronomy. In Section 6, several other solutions to the problem are briefly reviewed, and outstanding issues are discussed.

## 2 Hypothesis Testing, Model Selection and Accounting for Model Uncertainty Using Bayes Factors

The standard Bayesian solution to the hypothesis testing problem is to represent both the null and alternative hypotheses as parametric probability models, and to compute the Bayes factor for one against the other. The Bayes factor,  $B_{10}$  for a model  $M_1$  against another model  $M_0$  given data  $D$  is the ratio of posterior to prior odds, namely

$$B_{10} = \text{pr}(D|M_1)/\text{pr}(D|M_0), \quad (2)$$

the ratio of the marginal likelihoods. In equation (2),

$$\text{pr}(D|M_k) = \int \text{pr}(D|\theta_k, M_k)\text{pr}(\theta_k|M_k)d\theta_k, \quad (3)$$

where  $\theta_k$  is the vector of parameters of  $M_k$ , and  $\text{pr}(\theta_k|M_k)$  is its prior density ( $k = 0, 1$ ).

One important use of the Bayes factor is as a summary of the evidence for  $M_1$  against  $M_0$  provided by the data. It can be useful to consider twice the logarithm of the Bayes factor, which is on the same scale as the familiar deviance and likelihood ratio test statistics. I use the following rounded scale for interpreting  $B_{10}$ , which is based on that of Jeffreys (1961), but is more granular and slightly more conservative than his.

$B_{10}$	$2 \log B_{10}$	Evidence for $M_1$
$< 1$	$< 0$	Negative (supports $M_0$ )
1 to 3	0 to 2	Not worth more than a bare mention
3 to 12	2 to 5	Positive
12 to 150	5 to 10	Strong
$> 150$	$> 10$	Very strong

The model selection problem arises when one initially considers several models, perhaps a very large number, and wishes to select one of them. A Bayesian solution to this problem is to choose the model with the highest posterior probability. However, the purpose of choosing a model needs to be kept in mind, and if the principal decision-making problem is really one of choosing a model, then utilities should be introduced (Kadane and Dickey, 1980).

The marginal likelihoods yield posterior probabilities of all the models, as follows. Suppose that  $K$  models,  $M_1, \dots, M_K$ , are being considered. Then, by Bayes' theorem, the posterior probability of  $M_k$  is

$$\text{pr}(M_k|D) = \text{pr}(D|M_k) \text{pr}(M_k) / \sum_{r=1}^K \text{pr}(D|M_r) \text{pr}(M_r). \quad (4)$$

In the examples, I will take all the models to have equal prior probabilities, corresponding to prior information that is “objective” or “neutral” between competing models (e.g. Berger, 1985), but other prior information about the relative plausibility of competing models can easily be taken into account.

The ultimate goal of an investigation is often estimation, prediction or decision-making, rather than model selection *per se*. In that case, selecting a single model ignores model uncertainty and so will underestimate uncertainty about quantities of interest, thus, for example, biasing policy choices in favor of policies that are riskier in fact than the analysis appears to indicate (Hodges, 1987). It is better to take account explicitly of model uncertainty.

The posterior model probabilities given by equation (4) lead directly to solutions of the prediction, decision-making and inference problems that take account of model uncertainty. The posterior distribution of a quantity of interest  $\Delta$ , such as a structural parameter to be estimated, a future observation to be predicted, or the utility of a course of action, is

$$\text{pr}(\Delta|D) = \sum_{k=0}^K \text{pr}(\Delta|D, M_k) \text{pr}(M_k|D), \quad (5)$$

where  $\text{pr}(\Delta|D, M_k) = \int \text{pr}(\Delta|D, \theta_k, M_k) \text{pr}(\theta_k|D, M_k) d\theta_k$ . The combined posterior mean and standard deviation are

$$E[\Delta|D] = \sum_{k=0}^K \hat{\Delta}_k \text{pr}(M_k|D), \quad (6)$$

$$\text{Var}[\Delta|D] = \sum_{k=0}^K \left( \text{Var}[\Delta|D, M_k] + \hat{\Delta}_k^2 \right) \text{pr}(M_k|D) - E[\Delta|D]^2, \quad (7)$$

where  $\hat{\Delta}_k = E[\Delta|D, M_k]$  (Raftery, 1993a).

The number of models considered can be very large, and then direct evaluation of equation (5) will often be impractical. For example, in variable selection for regression with  $p$  candidate independent variables, the number of possible models is  $K = 2^p$ , so that if  $p = 20$ ,  $K$  is about one million. In such cases, two strategies for the evaluation of equation (5) have been proposed. One of these, Occam’s Window, consists of selecting and averaging over a much smaller set of models (Madigan and Raftery, 1994). The other, MCMC model composition (MC<sup>3</sup>), is a MCMC algorithm that moves through model space and generates a sample from the posterior distribution of the “true” model, which is used to approximate (5) (Madigan and York, 1993). Bayes factors and posterior model probabilities calculated from posterior simulation of the model parameters can be used as part of either the Occam’s Window or MC<sup>3</sup> algorithms.

For reviews of Bayes factors and their use in accounting for model uncertainty, see Kass and Raftery (1994) and Draper (1994).

### 3 Importance Sampling Estimators of the Marginal Likelihood

The marginal likelihood,  $\text{pr}(D|M_k)$ , is the key quantity needed for Bayesian hypothesis testing, model selection and accounting for model uncertainty. In this section and the next, I deal only with estimating the marginal likelihood for a single model, and so I will drop the notational dependence on  $M_k$ . In this section, I outline several importance and related estimators of the marginal likelihood based on a sample  $\{\theta^{(t)} : t = 1, \dots, T\}$  from the posterior distribution of the parameter  $\theta$  of the model. Such a sample can be generated by direct analytic simulation, MCMC, the weighted likelihood bootstrap (Newton and Raftery, 1994a), or other methods.

Unlike other posterior simulation methods, MCMC algorithms yield samples that are not independent and are only approximately drawn from the posterior distribution. Here I will assume that enough initial “burn-in” values have been discarded for the approximation to be good; see the chapters by Gelman (1994) and Raftery & Lewis (1994) for ways of ensuring this. I will also ignore the dependency between samples for the following reason. Since  $\{\theta^{(t)}\}$  defines a Markov chain, it is typically  $\phi$ -mixing at a geometric rate, in the sense of Billingsley (1968). Thus estimators that are simulation-consistent (as  $T \rightarrow \infty$ ) for independent samples from the posterior will usually also be so for MCMC samples, by the laws of large numbers for  $\phi$ -mixing processes in Billingsley (1968). MCMC algorithms can be made to yield approximately independent posterior samples by subsampling; see the chapter by Raftery & Lewis (1994).

Let  $L(\theta) = \text{pr}(D|\theta)$  be the likelihood and  $\pi(\theta) = \text{pr}(\theta)$  be the prior. Let  $\|X\|_h = T^{-1} \sum_{t=1}^T X(\theta^{(t)})$ , where  $X(\cdot)$  is a function and  $\{\theta^{(t)}\}$  is a sample of size  $T$  from the probability density  $h(\theta) / \int h(\phi)d\phi$ ,  $h$  being a positive function. Then the marginal likelihood is  $\text{pr}(D) = \int L\pi d\theta$ .

Importance sampling can be used to evaluate  $\text{pr}(D)$ , as follows. Suppose that we can sample from a density proportional to the positive function  $g(\theta)$ , say the density  $cg(\theta)$ , where  $c^{-1} = \int g(\theta)d\theta$ . Then

$$\begin{aligned}\text{pr}(D) &= \int L(\theta)\pi(\theta)d\theta \\ &= \int L(\theta) \left[ \frac{\pi(\theta)}{cg(\theta)} \right] (cg(\theta)) d\theta.\end{aligned}\tag{8}$$

Given a sample  $\{\theta^{(t)} : t = 1, \dots, T\}$  from the density  $cg(\theta)$ , then, as suggested by equation (8), a simulation-consistent estimator of  $\text{pr}(D)$  is

$$\begin{aligned}\hat{\text{pr}}(D) &= T^{-1} \sum_{t=1}^T \frac{L(\theta^{(t)}) \pi(\theta^{(t)})}{cg(\theta^{(t)})} \\ &= \|L\pi/cg\|_g.\end{aligned}$$

If  $c$  cannot be found analytically, it remains only to estimate it from the MCMC output. A simulation-consistent estimator of  $c$  is  $\hat{c} = \|\pi/g\|_g$ . This yields the general importance sampling estimator of  $\text{pr}(D)$  with importance sampling function  $g(\cdot)$ , namely

$$\hat{\text{pr}}_{IS} = \frac{\|L\pi/g\|_g}{\|\pi/g\|_g},\tag{9}$$

(Newton and Raftery, 1994b). Here,  $g$  is a positive function; if it is normalized to be a probability density, then equation (9) becomes  $\hat{\text{pr}}_{IS} = \|L\pi/g\|_g$  (Neal, 1994).

The simplest such estimator results from taking the prior as importance sampling function, so that  $g = \pi$ . This is

$$\hat{\text{pr}}_1(D) = \|L\|_\pi,$$

a simple average of the likelihoods of a sample from the prior. This is just the simple Monte-Carlo estimator of the integral  $\text{pr}(D) = \int L\pi d\theta$  (Hammersley and Handscomb, 1964). This was mentioned by Raftery and Banfield (1991), and was investigated in particular cases by McCulloch and Rossi (1991). A difficulty with  $\hat{\text{pr}}_1(D)$  is that most of the  $\theta^{(t)}$  have small likelihood values if the posterior is much more concentrated than the prior, so that the simulation process will be quite inefficient. Thus the estimate is dominated by a few large values of the likelihood, and so the variance of  $\hat{\text{pr}}_1(D)$  is large and its convergence to a Gaussian distribution is slow. These problems were apparent in the examples studied in detail by McCulloch and Rossi (1991). The variance of  $\hat{\text{pr}}_1(D)$  is roughly  $O(T^{-1}n^{d/2}|W|^{1/2})$  where  $n$  is the sample size,  $d$  is the number of parameters, and  $W$  is the prior variance matrix. To obtain reasonable precision with this method thus rapidly becomes expensive as the sample size increases, as the dimension increases, or as the prior becomes more diffuse.

Posterior simulation produces a sample from the posterior distribution, and so it is natural to take the posterior distribution as the importance sampling function. This yields the importance sampling estimator with importance sampling function  $g = L\pi$ , namely

$$\hat{\text{pr}}_2(D) = \|1/L\|_{\text{post}}^{-1},\tag{10}$$

where “post” denotes the posterior distribution; this is the harmonic mean of the likelihood values (Newton and Raftery, 1994a). It converges almost surely to the correct value,  $\text{pr}(D)$ , as  $T \rightarrow \infty$ , but it does not, in general, satisfy a Gaussian central limit theorem, and the variance of  $\hat{\text{pr}}_2(D)^{-1}$  is usually infinite. This manifests itself by the occasional occurrence of a value of  $\theta^{(t)}$  with small likelihood and hence large effect, so that the estimator  $\hat{\text{pr}}_2(D)$  can be somewhat unstable.

To get around this problem, Newton and Raftery (1994a) suggested using as importance sampling function in equation (9) a mixture of the prior and posterior densities, namely  $g(\theta) = \delta\pi(\theta) + (1 - \delta)\text{pr}(\theta|D)$ , where  $0 < \delta < 1$ . The resulting estimator,  $\hat{\text{pr}}_3(D)$ , has the efficiency of  $\hat{\text{pr}}_2(D)$  due to being based on many values of  $\theta$  with high likelihood, but avoids its instability and does satisfy a Gaussian central limit theorem. However, it has the irksome aspect that one must simulate from the prior as well as the posterior. This may be avoided by simulating all  $T$  values from the posterior distribution and imagining that a further  $\delta T/(1 - \delta)$  values of  $\theta$  are drawn from the prior, all of them with likelihoods  $L(\theta^{(t)})$  equal to their expected value,  $\text{pr}(D)$ . The resulting estimator,  $\hat{\text{pr}}_4(D)$ , may be evaluated using a simple iterative scheme, defined as the solution  $x$  of the equation

$$x = \frac{\delta T/(1 - \delta) + \sum[L_t/\{\delta x + (1 - \delta)L_t\}]}{\delta T/\{(1 - \delta)x\} + \sum\{\delta x + (1 - \delta)L_t\}^{-1}},$$

where  $L_t = L(\theta^{(t)})$  and the summations are over  $t = 1, \dots, T$ .

Another modification of the harmonic mean estimator,  $\hat{\text{pr}}_1(D)$ , is

$$\hat{\text{pr}}_5(D) = \left\| \frac{f}{L\pi} \right\|_{\text{post}}^{-1}, \quad (11)$$

where  $f$  is a function of  $\theta$  and is any probability density; this was mentioned by Gelfand and Dey (1994). It is unbiased and simulation-consistent, and satisfies a Gaussian central limit theorem if the tails of  $f$  are thin enough, specifically if

$$\int \{f^2/L\pi\} < \infty. \quad (12)$$

If  $\theta$  is one-dimensional, if the posterior distribution is normal, and if  $f$  is normal with mean equal to the posterior mean and variance equal to  $\kappa$  times the posterior variance, then the mean squared error of  $\hat{\text{pr}}_5(x)$  is minimized when  $\kappa = 1$ . This suggests that high efficiency is most likely to result if  $f$  is roughly proportional to the posterior density.

Meng and Wong (1993) proposed the alternative  $\hat{\text{pr}}_6(D) = \|L\pi g\|_{\pi}/\|\pi g\|_{\text{post}}$  where  $g$  is a positive function. Like  $\hat{\text{pr}}_3(D)$ , this has the disadvantage of needing simulation from the prior as well as the posterior. They considered an optimal choice of  $g$  and showed how it can be computed from an initial guess. This appears promising but has yet to be extensively tested.

Rosenkranz (1992) evaluated the estimators of the marginal likelihood,  $\hat{\text{pr}}_1(D)$ ,  $\hat{\text{pr}}_2(D)$ ,  $\hat{\text{pr}}_3(D)$  and  $\hat{\text{pr}}_4(D)$ , in the contexts of normal models, hierarchical Poisson-gamma models for counts with covariates, unobserved heterogeneity and outliers, and a multinomial model

with latent variables. She found that analytic approximations via the Laplace method gave greater accuracy for much less computation *and* human time than the posterior simulation estimators; the problem is that the Laplace method is not always applicable. Among the posterior simulation estimators, she found  $\hat{p}_3(D)$  with a large value of  $\delta$  (close to 1) to have the best performance. The harmonic mean estimator,  $\hat{p}_2(D)$  is easy to compute and her experience, as well as that of Carlin and Chib (1993), is that with substantial numbers of iterations (at least 5,000), it gave results accurate enough for the granular scale of interpretation in Section 2; see also Rosenkranz and Raftery (1994).

There has been less systematic evaluation of  $\hat{p}_5(D)$  or  $\hat{p}_6(D)$ . In a very simple one-dimensional example (testing  $\mu = 0$  in a  $N(\mu, 1)$  model) with a well chosen  $f$  function,  $\hat{p}_5(D)$  performed very well in a small numerical experiment that I carried out. In more complex and high-dimensional examples, such as the mixture model described in Section 1, experience to date is less encouraging. The estimator  $\hat{p}_5(D)$  appears to be sensitive to the choice of  $f$  function, and it seems hard in practice to ensure that the condition (12) holds. If  $f$  is not well chosen,  $\hat{p}_5(D)$  can give highly inaccurate answers, as illustrated in Section 5. More research is required on this issue.

## 4 Estimating the Marginal Likelihood by Adjusting the Maximized Likelihood

### 4.1 The Laplace-Metropolis Estimator

Rosenkranz (1992) found that the Laplace method produces much more accurate estimates of the marginal likelihood than posterior simulation for several very different models and for large amounts of simulation. However, the Laplace method is often not applicable because the derivatives that it requires are not easily available. This is particularly true for complex models of the kind for which posterior simulation, especially MCMC, is often used.

The idea of the Laplace-Metropolis estimator is to get around the limitations of the Laplace method by using posterior simulation to *estimate* the quantities it needs. The Laplace method for integrals (e.g. de Bruijn, 1970, Section 4.4) is based on a Taylor series expansion of the real-valued function  $f(u)$  of the  $d$ -dimensional vector  $u$ , and yields the approximation

$$\int e^{f(u)} du \approx (2\pi)^{d/2} |A|^{1/2} \exp\{f(u^*)\}, \quad (13)$$

where  $u^*$  is the value of  $u$  at which  $f$  attains its maximum, and  $A$  is minus the inverse Hessian of  $f$  evaluated at  $u^*$ . When applied to equation (3) it yields

$$p(D) \approx (2\pi)^{d/2} |\Psi|^{1/2} \text{pr}(D|\tilde{\theta})\text{pr}(\tilde{\theta}), \quad (14)$$

where  $d$  is the dimension of  $\theta$ ,  $\tilde{\theta}$  is the posterior mode of  $\theta$ , and  $\Psi$  is minus the inverse Hessian of  $h(\theta) = \log\{\text{pr}(D|\theta)\text{pr}(\theta)\}$ , evaluated at  $\theta = \tilde{\theta}$ . Arguments similar to those in the Appendix of Tierney and Kadane (1986) show that in regular statistical models the relative

error in equation (14), and hence in the resulting approximation to  $B_{10}$ , is  $O(n^{-1})$ , where  $n$  is sample size.

Thus the Laplace method requires the posterior mode,  $\tilde{\theta}$ , and  $|\Psi|$ . The simplest way to estimate  $\tilde{\theta}$  from posterior simulation output, and probably the most accurate, is to compute  $h(\theta^{(t)})$  for each  $t = 1, \dots, T$  and take the value for which it is largest. If the likelihood is hard to calculate, however, this may take too much computer time. A simple alternative is to use the multivariate median, or  $L_1$  center, defined as the value of  $\theta^{(t)}$  that minimizes  $\sum_s |\theta^{(s)} - \theta^{(t)}|$ , where  $|\cdot|$  denotes  $L_1$  distance; see, e.g., Small (1990). This is suboptimal but often yields values of  $h(\theta^{(t)})$  close to those at the posterior mode, and can be much cheaper computationally. Even cheaper is the componentwise posterior median, computed as the estimated posterior median of each parameter individually. This performs well in the majority of cases, but can sometimes give poor results. A fourth possibility is to estimate the posterior mode directly from the posterior sample using nonparametric density estimation.

The matrix  $\Psi$  is asymptotically equal to the posterior variance matrix, as sample size tends to infinity, and so it would seem natural to approximate  $\Psi$  by the estimated posterior variance matrix from the posterior simulation output. The main problem with this is that it is sensitive to the occasional distant excursions to which MCMC trajectories can be prone, and so I prefer to estimate the posterior variance matrix using a robust but consistent estimator variance matrix with a high breakdown point. This means that the estimator continues to perform well even if the proportion of outliers is high. I use the weighted variance matrix estimate with weights based on the minimum volume ellipsoid estimate of Rousseeuw and van Zomeren (1990); this is implemented in the S-PLUS function `cov.mve` using the genetic algorithm of Burns (1992).

The resulting estimator of the marginal likelihood typically involves less computation than the estimators in Section 3. It remains to be systematically evaluated, but initial indications are promising. In a small numerical experiment with a  $N(\mu, 1)$  model, it yielded values of the marginal likelihood that were accurate to within about 5% based on 600 iterations. It seems to be more stable than the estimators in Section 3. I conjecture that under quite mild conditions (finite posterior fourth moments are probably enough), its relative error is  $(O(n^{-\frac{1}{2}}) + O(T^{-\frac{1}{2}}))$  or less. This has not been proved, but see Kass and Vaidyanathan (1992) for relevant results and references. If true, this would usually be accurate enough for interpretation on the granular scale of Section 2. A short generic S-PLUS function to calculate it is shown in Figure 2.

## 4.2 The Candidate's Estimator

Bayes' theorem says that

$$\text{pr}(\theta|D) = \frac{\text{pr}(D|\theta)\text{pr}(\theta)}{\text{pr}(D)},$$

from which it follows that

$$\text{pr}(D) = \text{pr}(D|\theta) \cdot \left( \frac{\text{pr}(\theta)}{\text{pr}(\theta|D)} \right). \quad (15)$$

```

mcmcbf <- function (theta, data=NULL, method="likelihood") {
# Inputs: theta: A sequence of samples from the posterior distribution.
#       This is a (niter x p) matrix, where the parameter is of dimension p.
# data: Input data for calculating likelihoods
# method: if "likelihood", the empirical posterior mode is used
# if "L1center", the multivariate median is used
# if "median", the componentwise posterior median is used.
# Value: The Laplace-Metropolis estimator of the marginal likelihood.
# NOTE: The user must supply the functions llik and lprior to calculate the
# log-likelihood and log-prior for input values of the parameter.

theta <- as.matrix (theta); niter <- length (theta[,1]); p <- length (theta[1,])
if (method == "likelihood") { h <- NULL
for (t in (1:niter)) h <- c(h, llik(theta[t,],data) + lprior (theta[t,]) )
hmax <- max(h) }

if (method == "L1center") { L1sum <- NULL; oneniter<- as.matrix (rep(1,niter))
onep <- as.matrix (rep(1,p))
for (t in (1:niter)) { thetat <- theta[t,]; thetatmat <- oneniter %*% thetat
L1sum <- c(L1sum, sum(abs((theta-oneniter%*%thetat)%*%onep)) ) }
argL1center <- min ((1:niter)[L1sum==min(L1sum)])
thetaL1center <- theta[argL1center,]
hmax <- llik(thetaL1center,data) + lprior(thetaL1center) }

if (method=="median") { thetamed <- apply (theta,2,median)
hmax <- llik(thetamed,data) + lprior(thetamed) }
if (p==1) logdetV <- 2*log(mad(theta[,1])) else
logdetV <- sum(log(eigen(cov.mve(theta,print=F)$cov)$values ))
hmax + 0.5 * p * log(2*3.14159) + 0.5 * logdetV }

```

Figure 2: S-PLUS function to calculate the Laplace-Metropolis estimator of the marginal likelihood.

This was pointed out to me by Julian Besag, who noted that it is also related to his “Candidate’s formula” for Bayesian prediction (Besag, 1989), whence the name of the estimator I will describe.

Typically,  $\text{pr}(D|\theta)$  and  $\text{pr}(\theta)$  can be calculated directly. If  $\text{pr}(\theta|D)$  is also available in closed form, then equation (15) allows  $\text{pr}(D)$  to be calculated analytically without integration (Julian Besag, personal communication). This will often not be the case, but one can still use equation (15) if one has a sample from the posterior. One can then simply *estimate*  $\text{pr}(\theta|D)$  by nonparametric density estimation from the posterior sample. This is inexpensive computationally because the density is required at only one point. Practical multivariate density estimators have been proposed by Terrell (1990) and many others.

What value of  $\theta$  should be used? Equation (15) holds for all values of  $\theta$ , so in principle one could use any value. However, the most precise estimate of  $\text{pr}(\theta|D)$  would usually result from using the posterior mode or a value very close to it. Any value from the central part of the posterior sample should do almost as well.

### 4.3 The Data Augmentation Case

#### 4.3.1 Latent Data

MCMC methods often involve introducing *latent data*  $z$ , that is such that when  $z$  is known, the “complete data likelihood”,  $\text{pr}(D, z|\theta)$  has a simple form. This is the idea underlying the EM algorithm (Dempster, Laird and Rubin, 1977), its stochastic generalizations (see the chapter by Diebolt and Ip, 1994), and its Bayesian analogue, the IP algorithm (Tanner and Wong, 1987). In MCMC methods, values of both  $z$  and  $\theta$  are generated from their joint posterior distribution. The latent data can consist, for example, of individual random effects in a random effects or hierarchical model, of group memberships in a mixture model (see Section 5), or of missing data.

Typically  $z$  includes quantities that cannot be consistently estimated, in which case the conditions for the Laplace approximation to be good may not hold. Further, the latent data are often very numerous so that the matrix  $\Psi$  in equation (3) is of very high dimension, in which case there may also be numerical problems with the Laplace approximation. Rosenkranz (1992) has shown that for calculating marginal likelihoods there can be considerable advantages to integrating over the latent data directly, especially when there are conditional independence properties that can be exploited to reduce the dimensions of the integrals involved. Here I outline some strategies for doing this.

#### 4.3.2 The Conditional Independence Case

Suppose that the data can be partitioned as  $D = (D_1, \dots, D_n)$  and the latent data as  $z = (z_1, \dots, z_n)$ , and that

$$(z_i \perp z_j \mid \theta), \tag{16}$$

$$(D_i \perp D_j \mid z_i, z_j, \theta) \quad (i \neq j), \tag{17}$$

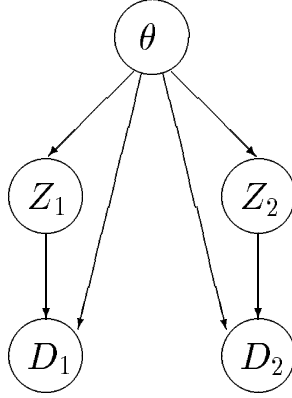


Figure 3: Graphical model representation of the conditional independence model of equations (16) and (17), when  $n = 2$ .

where  $\perp$  denotes independence. This covers many situations, including most random effects models. A graphical model representation of such a model is shown in Figure 3.

Now suppose that the MCMC algorithm has been run, yielding a sample from the joint posterior distribution of  $\theta$  and  $z$ ,  $\text{pr}(\theta, z|D)$ . We discard the values of  $z$ , and are left with a sample from the (marginal) posterior distribution of  $\theta$ ,  $\text{pr}(\theta|D)$ . We can now apply the Laplace-Metropolis or the Candidate’s estimator as before. It remains only to calculate  $\log \text{pr}(D|\tilde{\theta})$ , and we can use the fact that

$$\log \text{pr}(D|\tilde{\theta}) = \sum_{i=1}^n \log L_i(\tilde{\theta}),$$

where

$$L_i(\tilde{\theta}) = \int \text{pr}(D_i|z_i, \tilde{\theta})\text{pr}(z_i|\tilde{\theta})dz_i. \quad (18)$$

$L_i(\tilde{\theta})$  is typically a low-dimensional integral (often of one dimension), and so it is readily amenable to various integration strategies, including the following:

1. *Analytic evaluation:* If the distribution of  $z_i$  is conjugate to that of  $D_i$ , the integral (18) can often be evaluated analytically. This was illustrated by Rosenkranz (1992) in the case of hierarchical Poisson-gamma models.

2. *Summation:* If the  $z_i$  are discrete (as in Section 5), then the integral (18) can be directly and simply evaluated as

$$L_i(\tilde{\theta}) = \sum_{j=1}^J \text{pr}(D_i|z_i = j, \tilde{\theta}) \text{Pr}[z_i = j|\tilde{\theta}]. \quad (19)$$

3. *Laplace method:* The Laplace method can itself be applied to each  $L_i(\tilde{\theta})$  individually. If this idea is used with the Laplace-Metropolis estimator, the result is a kind of “iterated

Laplace method”, in which the Laplace method is applied at each of two stages. Analytic expressions for the maximum of the integrand in equation (18) and/or the required second derivatives may be available, making things easier. Otherwise, numerical approximations can be used, especially if the dimension of  $z_i$  is low. This is not guaranteed to perform well but the Laplace method has been found to provide remarkable accuracy even for very small “sample sizes”; see Grunwald, Raftery and Guttorp (1993) for one example of this.

4. *Quadrature*: If the dimension of  $z_i$  is low, then numerical quadrature may well work well for evaluating  $L_i(\tilde{\theta})$ .

### 4.3.3 Latent Data Without Conditional Independence

If conditions such as (16) and (17) do not hold, it is harder to evaluate  $L(\tilde{\theta}) = \text{pr}(D|\tilde{\theta}) = \int \text{pr}(D|z, \tilde{\theta})\text{pr}(z|\tilde{\theta})dz$ . One way of doing it is via an importance sampling estimator using the posterior simulation output for both  $z$  and  $\theta$ . We have

$$L(\tilde{\theta}) \approx \frac{\sum_{t=1}^T w(z^{(t)}) \text{pr}(D|z^{(t)}, \tilde{\theta})}{\sum_{t=1}^T w(z^{(t)})}, \quad (20)$$

where

$$w(z^{(t)}) = \text{pr}(z^{(t)}|\tilde{\theta})/\text{pr}(z^{(t)}|D). \quad (21)$$

In equation (21),

$$\begin{aligned} \text{pr}(z^{(t)}|D) &= \int \text{pr}(z^{(t)}|\theta, D)\text{pr}(\theta|D)d\theta \\ &\approx T^{-1} \sum_{s=1}^T \text{pr}(z^{(t)}|\theta^{(s)}, D). \end{aligned} \quad (22)$$

Then  $L(\tilde{\theta})$  is evaluated by substituting (22) into (21) and then (21) into (20). This estimator has the advantage that it involves only prior and posterior ordinates, and no integrals. However, it does require the availability of a way of calculating  $\text{pr}(z^{(t)}|\theta^{(s)}, D)$ .

If  $\theta$  is absent from the model, or is assumed known, then (20) reduces to the harmonic mean estimator of the marginal likelihood. However, preliminary inspection suggests that if  $\theta$  is present, the estimator (20) will tend not to suffer from the instability of the harmonic mean estimator. This is because the smoothing effect of averaging over the values of  $\theta$  may help to avoid the overwhelming importance of individual samples that can plague the harmonic mean estimator.

The estimator (20) is related to, but not the same as, the estimator of Geyer and Thompson (1992). In their formulation, the  $z^{(t)}$  were simulated from  $\text{pr}(z|\theta_0)$  for a fixed  $\theta_0$ , rather than from the posterior distribution,  $\text{pr}(z|D)$ . More recently, Geyer (1993) has proposed an estimator in which the  $z^{(t)}$  are simulated from a mixture,  $\sum_{j=1}^J \gamma_j \text{pr}(z|\theta_j)$ .

## 5 Application: Determining the Number of Components in a Mixture

I now return to the motivating application of Section 1. I first briefly recall the basic ideas of Gibbs sampling for Gaussian mixtures, I then give a simulated example, and finally I return to the problem of the number of disks in the Galaxy.

### 5.1 Gibbs Sampling for Gaussian Mixtures

I consider the one-dimensional Gaussian mixture of equation (1), where  $\theta_j = (\mu_j, v_j)$  and  $f(\cdot|\theta_j)$  is a normal density with mean  $\mu_j$  and variance  $v_j$ .

I use the prior densities

$$\begin{cases} v_j \sim IG(\omega_j/2, \lambda_j/2) \\ (\mu_j|v_j) \sim N(\xi_j, v_j/\tau_j) \\ \rho = (\rho_1, \dots, \rho_J) \sim \text{Dirichlet}(\alpha_1, \dots, \alpha_J), \end{cases} \quad (23)$$

where  $IG(\alpha, \beta)$  denotes the inverse gamma density defined by  $f(x) = \beta^\alpha \Gamma(\alpha)^{-1} x^{-(\alpha+1)} \exp(-\beta/x)$ , for  $x > 0$ . The prior parameters are chosen so that the prior distribution is relatively flat over the range of values that could be expected given the range of the data. The values used were  $\omega_j = 2.56$ ,  $\lambda_j = 0.72 \hat{\text{Var}}(y)$ ,  $\xi = \bar{y}$ ,  $\tau_j = (2.6/(y_{\max} - y_{\min})^2)$ , and  $\alpha_j = 1$  ( $j = 1, \dots, J$ ). For derivations and discussion of such “reference proper priors” for Bayes factors, see Raftery (1993c) and Raftery, Madigan and Hoeting (1993).

In order to use the Gibbs sampler, I introduce the latent data  $z = (z_1, \dots, z_n)$ , where  $z_i = j$  if  $y_i \sim N(\mu_j, v_j)$ , i.e. if  $y_i$  belongs to the  $j$ th component of the mixture. The required conditional posterior distributions are then given by Diebolt and Robert (1994); see also the chapter by Robert (1994). The Gibbs sampler is initialized by dividing the data into  $J$  equal-sized chunks of contiguous data points, using the resulting means and variances for  $\mu$  and  $v$ , and setting  $\rho_j = 1/J$  ( $j = 1, \dots, J$ ). It proceeds by drawing first  $z$ , and then  $\rho$ ,  $v$  and  $\mu$  in turn from their conditional posterior distributions, and iterating. There is thus no need to initialize  $z$ .

### 5.2 A Simulated Example

I consider a sample of size  $n = 100$  data points generated from the model described in Section 5.1 with  $J = 2$  components,  $\mu = (0, 6)$ ,  $v = (1, 4)$ , and  $\rho = (\frac{1}{2}, \frac{1}{2})$ . The Gibbs sampler was run with 600 iterations, of which the first 20 were discarded. The data, together with the true and estimated densities, are shown in Figure 4. The estimated density is the mixture of two normal densities corresponding to the estimated posterior means of the parameters. The sequences of  $\mu$  values are shown in Figure 5 together with the corresponding likelihoods.

Corresponding quantities for the 3-component model are shown in Figure 6. The Gibbs sampler makes some very distant excursions, which occur when one of the groups becomes

## True and Estimated Densities

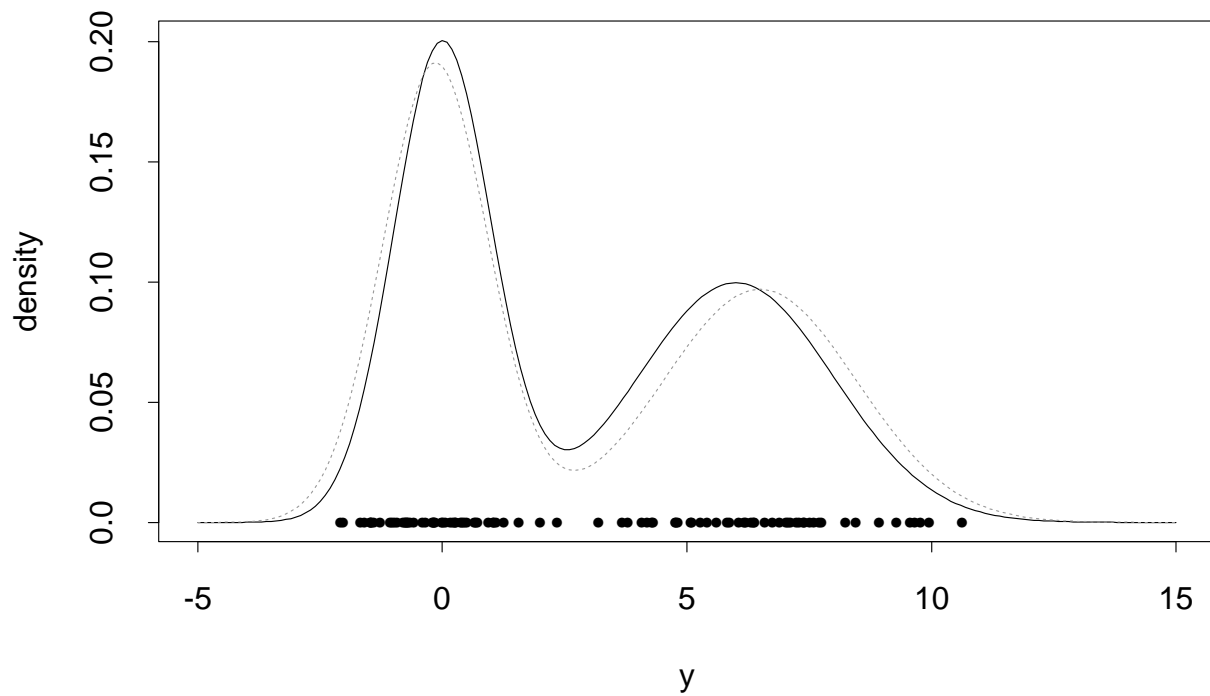


Figure 4: 100 points simulated from the  $\frac{1}{2}N(0, 1) + \frac{1}{2}N(6, 2^2)$  density, with the true (solid) and estimated (dotted) densities overlaid.

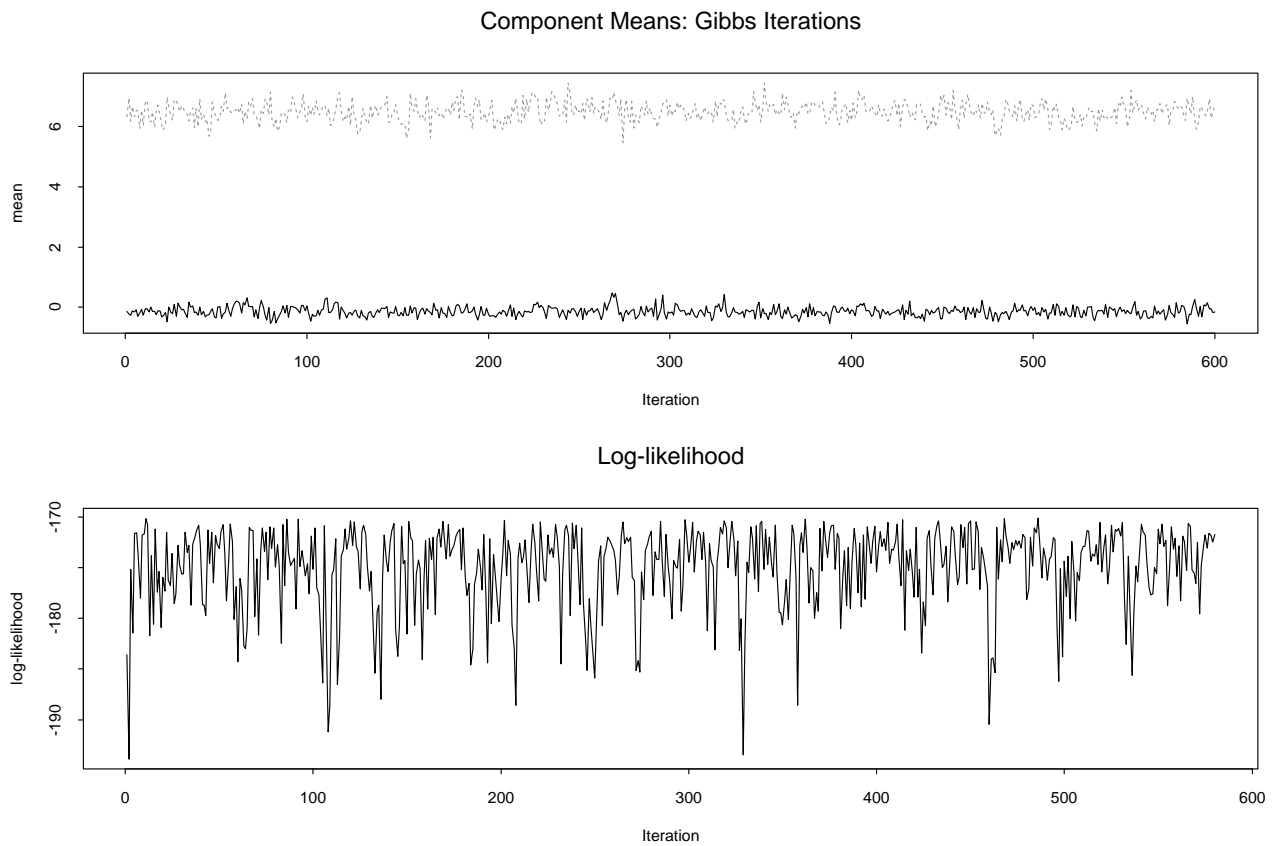


Figure 5: Results of the Gibbs sampler for a two-component Gaussian mixture run on the data of Figure 4 with 600 iterations: (a) Component means and (b) log-likelihood for each iteration.

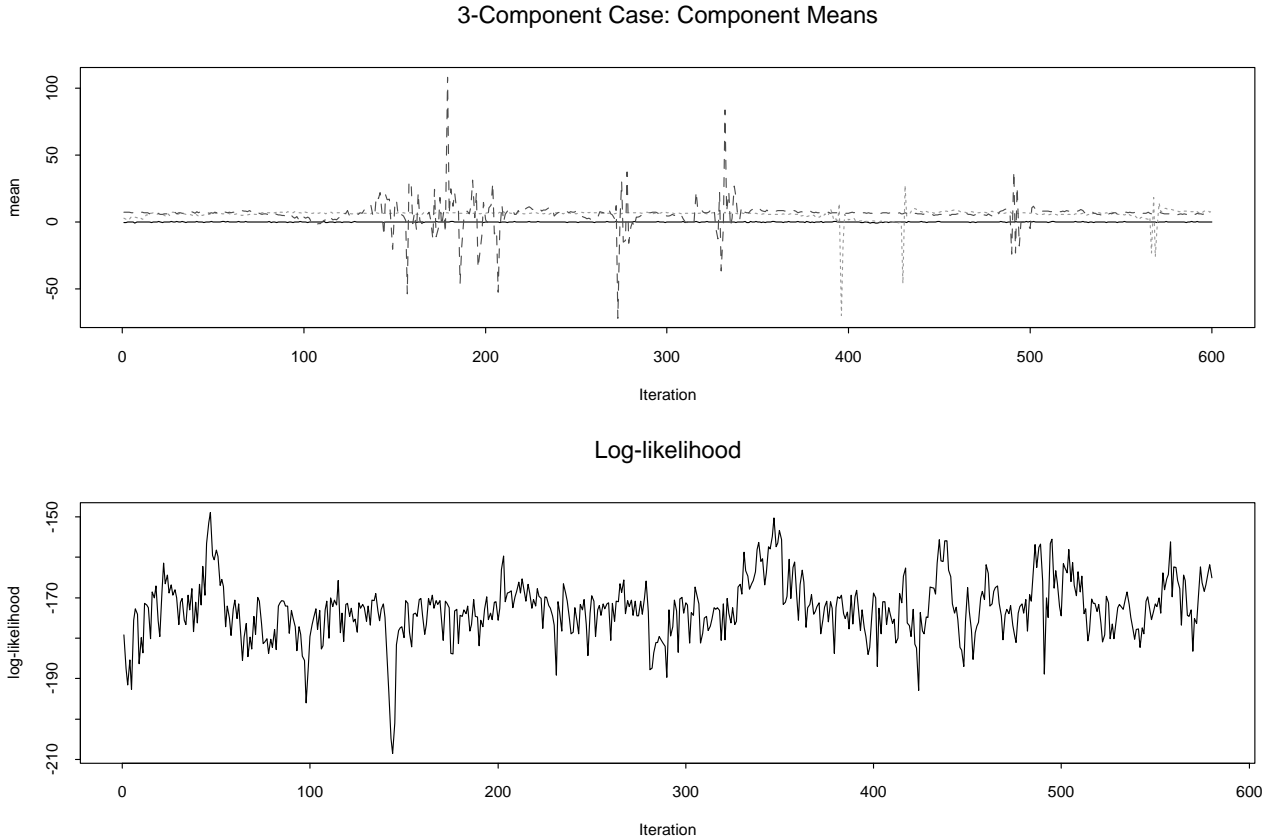


Figure 6: Results of the Gibbs sampler based on a three-component Gaussian mixture model run on the data of Figure 4 with 600 iterations: (a) Component means and (b) log-likelihood for each iteration.

empty. The algorithm has no problem returning after these excursions, and the corresponding likelihoods are not very low, but these episodes nevertheless require careful handling when estimating Bayes factors.

The various estimators of the log-marginal likelihood are shown in Table 1. The Laplace-Metropolis estimator was computed using equations (14), (18) and (19). The Gelfand-Dey estimator (11),  $\hat{p}_{r_5}(D)$ , is computed using for  $f(\cdot)$  a product of independent densities. For the  $\mu$ ,  $v$  and  $\rho$  parameters these are normal with mean and variance equal to those of the simulated values, while for each  $z_i$  they are multinomial with probabilities again estimated from the Gibbs sampler output. Maximized log-likelihoods are shown, with maximization over both  $\theta$  and  $(\theta, z)$ . The Schwarz approximation to the log-marginal likelihood is given as a general indication of a ballpark value, even though it is not known to be valid in this case. This is defined by  $\log \hat{p}_{\text{Schwarz}}(D) = \max_t \log \text{pr}(D|\theta^{(t)}) - \frac{1}{2}d \log(rn)$ , where  $d$  is the number of parameters and  $r$  is the dimension of  $y_i$  (here  $r = 1$ ). Finally, the crude AWE (Approximate Weight of Evidence) approximation of Banfield and Raftery (1993) is also shown, defined as  $\log \hat{p}_{\text{AWE}}(D) = \max_t \log \text{pr}(D|\theta^{(t)}, z^{(t)}) - (d + \frac{3}{2}) \log(rn)$ .

Table 1: Log marginal likelihood estimates for the simulated data based on 600 Gibbs sampler iterations. The log marginal likelihood for the one-component model is  $-271.8$  (calculated analytically).

Estimator	Number of Components		
	2	3	4
Laplace-Metropolis	-249.8	-248.6	-251.7
Harmonic Mean	-188.1	-202.1	-203.7
$\log \hat{p}r_5(D)$	37.6	127.0	181.6
Schwarz	-249.7	-256.7	-263.6
AWE	-200.6	-197.7	-211.7
$\max_t \ell(\theta^{(t)})$	-238.2	-238.3	-238.3
$\max_t \ell(\theta^{(t)}, z^{(t)})$	-170.1	-148.9	-144.5
Number of parameters	5	8	11

The Laplace-Metropolis estimator is the only one that seems to be in the right ballpark, and this indicates that there are probably either two or three groups but that the data do not distinguish clearly between these two possibilities. Even though the data were generated from a two-component model, Figure 4 shows that there is a second gap in the data around  $y = 8$ , so that someone analyzing the data by eye might well come to the same somewhat uncertain conclusion. For this example, the three ways of calculating the Laplace-Metropolis estimator gave answers that were very close.

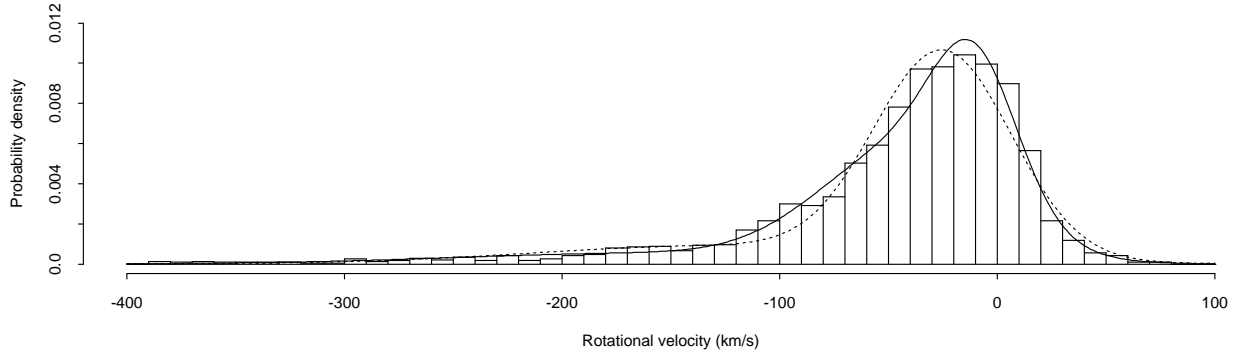
The Gelfand-Dey estimator implies that there is decisive evidence for four components against one, two or three. Even though the true values of the marginal likelihoods are not known, this seems somewhat implausible, and suggests the estimator to be rather inaccurate. A more carefully crafted  $f$  function seems to be needed. The harmonic mean estimator favors the two-component model, but it seems somewhat inaccurate with this fairly small number of iterations. The AWE estimator, used with the qualitative guidelines of Banfield and Raftery (1993), would, like the Laplace-Metropolis estimator, lead one to consider the two- and three-component models, and so it is as accurate here as its creators claimed! Except for  $\hat{p}r_5(D)$ , these less accurate estimators give better estimates of the Bayes factor between the two most likely models than of the marginal likelihoods themselves.

### 5.3 Determining the Number of Disks in the Galaxy

The Gibbs sampler was run using the rotational velocities for the 2,370 stars shown in Figure 1 for the two- and three-component models. Only 100 iterations were used because of computational constraints, and the first 50 of these were discarded. The trajectories were quite stable in this case, perhaps because of the considerable amount of data.

The log-marginal likelihoods, computed using the Laplace-Metropolis estimator, were:

### 2- and 3-component mixture models for stellar velocities



### Components of the 3-component mixture model

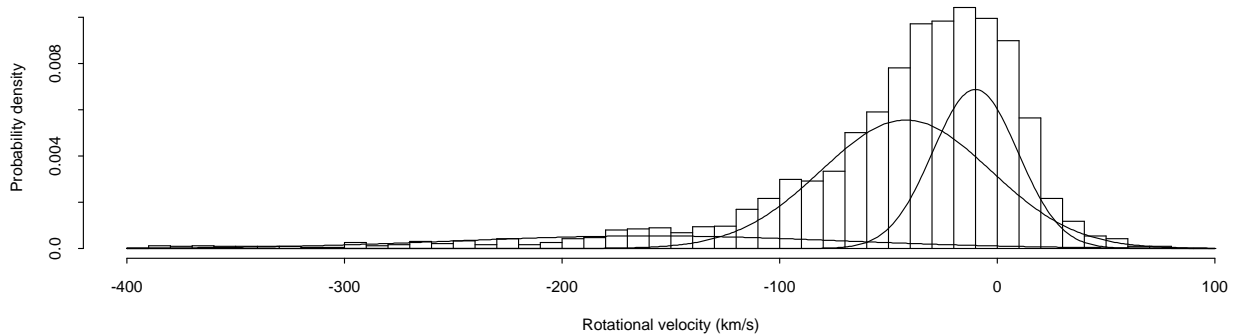


Figure 7: Gibbs sampler results for the astronomy data of Figure 1, with histogram of the data: (a) Estimated densities for the 2-component (dashed) and 3-component (solid) models; (b) Estimated components of the 3-component mixture model.

one component:  $-13141.0$ ; two components:  $-12615.5$ ; three components:  $-12609.6$ . Thus the rotational velocities alone appear to provide strong evidence for a three-component model over a two-component model. This is in agreement with recent astronomical opinion (Soubran, 1993, and references therein).

Figure 7(a) shows the fitted marginal densities of the data under the two competing models. The differences are subtle, but close inspection does support the better fit of the three-component model. The estimated densities under both models are strongly unimodal, so the ability of the method to distinguish between them is impressive. Both models fit quite well, indicating that the choice of a normal distribution for the individual components is satisfactory. The estimated components of the mixture are shown in Figure 7(b).

In another chapter, Robert (1994) analyzed a related but much smaller data set different approach (Kullback-Leibler divergence) to choose between the one- and two-component models.

## 6 Discussion

I have described various methods for computing marginal likelihoods, and hence doing Bayesian hypothesis testing, model selection and accounting for model uncertainty using posterior simulation. Research on this topic is at an early stage and much remains to be done. Preliminary experience with the Laplace-Metropolis estimator introduced here is encouraging. The importance sampling estimators described in Section 3 seem to have drawbacks, although future research may overcome these. One way of doing this would be to develop a good automatic way of choosing the  $f$  function for the modified harmonic mean estimator,  $\hat{\text{pr}}_5(D)$  of equation (11), suggested by Gelfand and Dey (1994). The ideas of Meng and Wong (1993) seem promising here.

Some general methods of calculating the marginal likelihood have been considered in the statistical physics literature under the name “free energy estimation”. The approaches are not automatic and require analytical effort to tailor them to statistical applications. However, they may be of use in certain problems. See Neal (1993) for references.

In this chapter I have concentrated on methods for estimating the marginal likelihood. In some situations there are methods for directly estimating Bayes factors with posterior simulation. For example, suppose one wishes to compare  $M_0$  with  $M_1$ , where the parameter of  $M_1$  is  $\theta_1 = (\omega, \psi)$ , and  $\theta_0$  is specified by setting  $\omega = \omega_0$ . Then Verdinelli and Wasserman (1993) have shown that if  $\text{pr}(\omega|\psi, D, M_1)$  is available in closed form, the Savage density ratio can be used to provide an estimator of the Bayes factor. This is in the form of an unweighted average, and so it may well have good performance.

Unfortunately, this result is often not directly applicable. For example, the comparison of mixture models in Section 5 cannot readily be cast in this form. Also, if  $\text{pr}(\omega|\psi, D, M_1)$  is not in closed form, Verdinelli and Wasserman (1993) resort to an importance sampling estimator of Chen (1992). This is closely related to the estimator  $\hat{\text{pr}}_5(D)$  whose performance to date is variable, and so its quality needs to be further evaluated.

Carlin and Chib (1993) suggested including a model indicator variable in the MCMC scheme and defining “pseudo-priors” for  $(\theta_1|H_2)$  and  $(\theta_2|H_1)$ . This involves designing and running a special MCMC algorithm to calculate Bayes factors. Similar suggestions have been made by Carlin and Polson (1991) and George and McCulloch (1993).

In the data-augmentation case, an alternative approach to estimating a Bayes factor is available when the latent data  $z$  is present in both models and has the same meaning in each one. Then the Bayes factor can be simulation-consistently estimated as the average of the “complete-data” Bayes factors, namely

$$\begin{aligned} B_{10} &\approx T^{-1} \sum_{t=1}^T B_{10}(z^{(t)}) \\ &= T^{-1} \sum_{t=1}^T \frac{\text{pr}(D, z^{(t)}|M_1)}{\text{pr}(D, z^{(t)}|M_0)}, \end{aligned}$$

where the  $z^{(t)}$  are simulated from their posterior distribution under  $M_0$ . The  $B_{10}(z^{(i)})$  are then often easy to calculate, or at least to approximate fairly well, for example using the

Laplace method. When  $z$  is present in  $M_1$  but not in  $M_0$ , we again recover the harmonic mean estimator of  $\text{pr}(D|M_1)$  (Raftery, 1993b). This is related to previous work of Thompson and Guo (1991) on the calculation of likelihood ratios.

## References

- Banfield, J.D. and Raftery, A.E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49, 803–821.
- Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*. New York: Springer-Verlag.
- Besag, J.E. (1989). A candidate's formula: A curious result in Bayesian prediction. *Biometrika*, 76, 183.
- Billingsley, P. (1968). *Convergence of Probability Measures*. New York: Wiley.
- Burns, P.J. (1992). A genetic algorithm for robust regression estimation. Statistical Sciences, Inc., Seattle, Wash.
- Carlin, B.P. and Chib, S. (1993). Bayesian model choice via Markov chain Monte Carlo. Research Report 93-006, Division of Biostatistics, University of Minnesota.
- Carlin, B.P. and Polson, N.G. (1991). Inference for nonconjugate Bayesian models using the Gibbs sampler. *Canadian J. Statist.*, 19, 399-405.
- Chen, M-H. (1992). Importance weighted marginal Bayesian posterior density estimation. Technical Report, Department of Statistics, Purdue University.
- de Bruijn, N.G. (1970). *Asymptotic Methods in Analysis*. Amsterdam: North-Holland.
- Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with Discussion). *J. R. Statist. Soc., ser. B*, 39, 1–37.
- Diebolt, J. and Ip, E. (1994). TITLE. In *Practical Markov Chain Monte Carlo* (W.R. Gilks, D.J. Spiegelhalter and S. Richardson, eds.), to appear.
- Diebolt, J. and Robert, C. (1994). Estimation of finite mixture distributions through Bayesian sampling. *J. R. Statist. Soc., ser. B*, 56, 363–376.
- Draper, D. (1994). Assessment and propagation of model uncertainty (with Discussion). *J. R. Statist. Soc. B*, 56, to appear.
- Gelfand, A.E. and Dey, D.K. (1994). Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society, series B*, 56, to appear.
- Gelman, A. (1994). Model validation. In *Practical Markov Chain Monte Carlo* (W.R. Gilks, D.J. Spiegelhalter and S. Richardson, eds.), to appear.
- George, E.I. and McCulloch, R.E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88, 881–889.
- Geyer, C.J. (1993). Estimating normalizing constants and reweighting mixtures in Markov chain Monte Carlo. Technical Report no. 568, School of Statistics, University of Minnesota.
- Geyer, C.J. and Thompson, E.A. (1992). Constrained Monte Carlo maximum likelihood for dependent data (with Discussion). *J. Roy. Statist. Soc., ser. B*, 54, 657–699.
- Grunwald, G.K., Raftery, A.E. and Guttorp, P. (1993). Time series of continuous proportions. *J. Roy. Statist. Soc., ser. B*, 55, 103–116.

- Hammersley, J.M. and Handscomb, D.C. (1964). *Monte Carlo Methods*. London: Chapman and Hall.
- Hodges, J.S. (1987). Uncertainty, policy analysis and statistics. *Statistical Science*, 2, 259–291.
- Jeffreys, H. (1961). *Theory of Probability*, Third Edition. Oxford University Press.
- Kadane, J.B. and Dickey, J.M. (1980). Bayesian Decision Theory and the Simplification of Models. In *Evaluation of Econometric Models*, (eds. J. Kmenta and J. Ramsey), Academic Press, 245–268.
- Kass, R.E. and Raftery, A.E. (1994). Bayes factors. *Journal of the American Statistical Association*, to appear.
- Kass, R.E. and Vaidyanathan, S. (1992). Approximate Bayes factors and orthogonal parameters, with application to testing equality of two Binomial proportions. *J. Royal Statist. Soc. B.*, 54, 129–144.
- Madigan, D. and Raftery, A.E. (1994) Model selection and accounting for model uncertainty in graphical models using Occam’s window. *Journal of the American Statistical Association*, to appear.
- Madigan, D. and York, J. (1993) Bayesian graphical models for discrete data. Technical Report no. 259, Department of Statistics, University of Washington.
- McCulloch, R.E. and Rossi, P.E. (1991). A Bayesian approach to testing the arbitrage pricing theory. *Journal of Econometrics*, 49, 141–168.
- Meng, X.L. and Wong, W.H. (1993). Simulating ratios of normalizing constants via a simple identity. Technical Report no. 365, Department of Statistics, University of Chicago.
- Neal, R.M. (1993). Probabilistic inference using Markov chain Monte Carlo methods based on Markov chains. Technical Report CRG-TR-93-1, Department of Computer Science, University of Toronto.
- Neal, R.M. (1994). Contribution to the discussion of “Approximate Bayesian inference by the weighted likelihood bootstrap.” *Journal of the Royal Statistical Society, series B*, 56, 41–42.
- Newton, M.A. and Raftery, A.E. (1994a). Approximate Bayesian inference by the weighted likelihood bootstrap (with Discussion). *Journal of the Royal Statistical Society, series B*, 56, 3–48.
- Newton, M.A. and Raftery, A.E. (1994b). Reply to the discussion of “Approximate Bayesian inference by the weighted likelihood bootstrap.” *Journal of the Royal Statistical Society, series B*, 56, 43–48.
- Raftery, A.E. (1993a). Bayesian model selection in structural equation models. In *Testing Structural Equation Models* (eds. K.A. Bollen and J.S. Long), Beverly Hills: Sage, pp. 163–180.
- Raftery, A.E. (1993b). Discussion of three papers on the Gibbs sampler and other Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, series B*, 53, 85.
- Raftery, A.E. (1993c). Approximate Bayes factors and accounting for model uncertainty in generalized linear models. Technical Report no. 255, Department of Statistics, University of Washington.
- Raftery, A.E. and Banfield, J.D. (1991). Stopping the Gibbs sampler, the use of morphology and other issues in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43,

32–43.

- Raftery, A.E. and Lewis, S.M. (1994). The number of iterations, convergence diagnostics, and generic Metropolis algorithms. In *Practical Markov Chain Monte Carlo* (W.R. Gilks, D.J. Spiegelhalter and S. Richardson, eds.), to appear.
- Raftery, A.E., Madigan, D.M. and Hoeting, J. (1993). Accounting for model uncertainty in linear regression models. Technical Report no. 262, Department of Statistics, University of Washington.
- Robert, C. (1994). Mixtures of distributions: Inference and estimation. In *Practical Markov Chain Monte Carlo* (W.R. Gilks, D.J. Spiegelhalter and S. Richardson, eds.), to appear.
- Rosenkranz, S. (1992). The Bayes factor for model evaluation in a hierarchical Poisson model for area counts, Ph.D. dissertation, Department of Biostatistics, University of Washington, 1992.
- Rosenkranz, S. and Raftery, A.E. (1994). Covariate selection in hierarchical models of hospital admission counts: A Bayes factor approach. Technical Report no. 268, Department of Statistics, University of Washington.
- Rousseeuw, P.J. and van Zomeren, B.C. (1990). Unmasking multivariate outliers and leverage points (with discussion). *J. Amer. Statist. Ass.*, 85, 633–651.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461–464.
- Small, C.G. (1990). A survey of multivariate medians. *Int. Statist. Rev.*, 58, 263–277.
- Soubiran, C. (1993). Kinematics of the Galaxy’s stellar populations from a proper motion survey. *Astronomy and Astrophysics*, to appear.
- Tanner, M. and Wong, W. (1987). The calculation of posterior distributions by data augmentation (with discussion). *J. Amer. Statist. Assoc.*, 82, 528–550.
- Terrell, G.R. (1990). The maximal smoothing principle in density estimation. *J. Amer. Statist. Ass.*, 85, 470–477.
- Thompson, E.A. and Guo, S.W. (1991). Evaluation of likelihood ratios for complex genetic models. *IMA J. Math. Appl. Med. Biol.*, 8, 149–169.
- Tierney, L., and Kadane, J.B. (1986) Accurate approximations for posterior moments and marginal densities. *J. Amer. Statist. Ass.*, 81, 82–86.
- Titterton, D.M., Smith, A.F.M. and Makov, U.D. (1985). *Statistical Analysis of Finite Mixture Distributions*. New York: Wiley.
- Verdinelli, I. and Wasserman, L. (1993a). A note on generalizing the Savage-Dickey density ratio. Technical Report no. 573, Department of Statistics, Carnegie-Mellon University.