Erroneous Results in "Marginal Likelihood from the Gibbs Output"

Radford M. Neal
Department of Statistics and Department of Computer Science
University of Toronto, Toronto, Ontario, Canada
http://www.cs.utoronto.ca/∼radford/
radford@stat.utoronto.ca

8 September 1998 (italic comments added 16 March 1999)

The following letter was submitted for publication in JASA. The editor declined to publish it, apparently because he and the referees believed that any problem is with the Gibbs sampling method for mixtures, rather than with Chib's use of this algorithm. This is not true. The Gibbs sampler for the dataset in question mixes well within one of the $k!$ symmetrical modes. That it does not mix well between these modes is and was well known to all concerned. The point at issue is whether or not this lack of mixing between symmetrical modes invalidates the results of applying Chib's method. Chib, the referees, and the editor apparently believe it does not. They are wrong. Anyone wishing to confirm this can first verify the correctness of the short S program included, and then run it themselves. This will establish that Chib's results are incorrect. The fact that the discrepancies match what would be expected according to the arguments presented in the letter indicates that these incorrect results are due to the general problem that I describe, rather than being due to a failure of Gibbs sampling for this particular dataset to mix well within one of the symmetrical modes.

To the best of my knowledge, none of the JASA referees considered this evidence. The referee reports consist simply of unsupported assertions that Chib's method must be correct, and therefore any discrepancy must be the fault of the Gibbs sampling algorithm.

This problem with the application of Chib's method will affect its use in calculating marginal likelihoods for mixture models, hidden markov models, neural networks, and other models with similar symmetries. As noted in the letter, it is possible to fix this problem, though the fixes involve substantial additional computation, lessening the attractiveness of Chib's general method.

Due to the potentially serious consequences if Chib's method is used for such models in real applications, I feel that I have a responsibility to inform the scientific community of this error, and am therefore distributing this letter through my web page.

A paper by S. Chib published in the *Journal of the American Statistical Association* (vol. 90, pp. 1313-1321, 1995), presents a method for finding the marginal likelihood of a Bayesian model using information obtained from Gibbs sampling. This method is valid in general, but its application to mixture models in Section 4.2.1 of the paper is incorrect, as I show here by computing the marginal likelihoods by a different method, obtaining different answers than Chib. The problem appears to be an incorrect treatment of the

Model fitted	*Chib's estimate*	*C program*	*S program*
2 components, equal variances	-240.464 (.006)	-239.764 (.005)	-239.798 (.050)
3 components, equal variances	-228.620 (.008)	-226.803 (.040)	—
3 components, unequal variances	-224.138 (.086)	-226.791 (.089)	—

Table 1: Estimates for the log of the marginal likelihood of various models for the galaxy data. The standard errors are given in parentheses.

non-identifiability present in mixture models. If this problem is corrected, Chib's method should be able to produce correct answers, but the computation time for his method will either grow as the factorial of the number of mixture components, or will require the possibly expensive computation of constrained conditional densities.

In Chib's mixture model example, the velocities of 82 galaxies were modeled as coming from a mixture of normals. Chib considered a model with two components having equal variances, a model with three components having equal variances, and a model with three components whose variances were not constrained to be equal. In all cases, the means of these components had independent normal priors with mean 20 and variance 100. The variance parameter (or parameters) were given inverse gamma priors with shape parameter 3 and scale parameter 20. The prior for the mixing proportions was uniform over the simplex of valid probabilities.

To compare with Chib's results, I estimated the marginal likelihood, which is the expected probability density of the data under the prior, by the simple procedure of averaging the probability density of the data for a large sample of points drawn from the prior distribution. Though quite tedious, this is feasible for these simple models using today's computers. Estimates based on 10^8 points drawn from the prior were computed using a program written in C, and are shown in Table 1, along with Chib's estimates (both are reported in terms of the log of the marginal likelihood). As a further check, I also show an estimate for the model with two components that was obtained using a program written in S, which uses a different random number generator. This estimate was based on only 10^6 points, because S is much slower than C. The S program used is shown in Figure 1 in order to make clear exactly how the computation was done.

As seen in the table, the results I found by sampling from the prior are not compatible with Chib's results. Since the method described in Chib's paper is clearly correct at an abstract level, the error must lie in how it was applied to these models. I believe the problem arises because mixture models are not identifiable, since relabelling the mixture components does not change the probability density for the data. This non-identifiability causes no real problems when estimating posterior expectations using Gibbs sampling — one simply looks at quantities that are invariant under relabeling, in which case it makes no difference how the components are labeled. However, it is essential that the different labelings of mixture components be treated in a consistent manner when calculating the marginal likelihood.

Naive application of Chib's method to mixture models will give the correct answer provided the Gibbs sampling chain visits all labelings of the components. This will usually occur in theory, but in practice the time required for a Gibbs sampling chain to sample all labelings may be very long, since these labelings correspond to modes in the posterior distribution that will often be isolated from each other. This lack of mixing can be solved by introducing special relabeling transitions into the Markov chain. When the number of mixture components, k, is not too large, Chib's method should then produce acceptable results. When k is large, however, estimation of the posterior probability of a chosen parameter vector, θ^*, which is central to Chib's method, will likely be very inefficient, taking about $k!$ longer than one might naively expect, since the chain will spend most of its time in the vicinity of the $k!-1$ relabelings of θ^*, rather than in the vicinity of θ^* itself.

I pointed out this potential difficulty to Chib after reading a draft of his paper. From the subsequent discussion, it appears that in fact Chib relabels the mixture components after each Gibbs sampling step so that their means are in increasing order, though this is not mentioned in the published paper. This reordering of course guarantees that other labelings are not visited. Estimation of the posterior probability of θ^* will then be based on values for the latent variables that are always appropriate for this labeling. When θ^* has negligible probability conditional on latent variables appropriate for another labeling, as is typical when the parameters are well-estimated, the posterior probability of θ^* will then be overestimated by a factor of $k!$, leading to an underestimate of the marginal likelihood by this same factor.

This appears to explain the incorrect results Chib obtained for the two and three component models with equal variances, since the differences from the correct results for these models are indeed approximately $\log(2!) = 0.693$ and $\log(3!) = 1.792$. For the three component model with unequal variances, however, Chib's estimate is actually higher than the true value, and hence cannot be explained in this way. For this model, the parameters are not so well estimated, and it may be that there is a non-negligible probability of moving from one labelling of the components to another in a single Gibbs sampling scan. However, it is difficult to see how this effect can do any more than cancel the tendency to underestimate the log marginal likelihood by $\log(3!)$. The higher estimate obtained by Chib may therefore be due to some other factor.

According to an anonymous JASA referee, the figure of -224.138 *for the log of the marginal likelihood for the three component model with unequal variances that was given in Chib's paper is a "typo", with the correct figure being* -228.608. *With this correction, the difference between my results and Chib's results are approximately* $\log(3!)$, *explaining this anomaly. One should note, however, that in general the difference will not always be* $\log(k!)$ *— this will be true only when the symmetrical modes are isolated with respect to the Gibbs sampling updates.*

It would be possible to use Chib's method to compute the marginal likelihood for a model in which the labeling of the components is constrained to produce some ordering. (This marginal likelihood is in fact the same as the marginal likelihood for the unconstrained model.) To do this correctly, one would need to use a prior that incorporates the constraint, under which the prior density for parameters obeying the constraint would be a factor of $k!$ larger than under the prior for the unconstrained model. One would also need to use Gibbs sampling updates (for the marginal likelihood calculation, though not necessarily for the actual sampling) in which the constraint is enforced when computing the conditional densities. This would avoid the factor of $k!$ slowdown from introducing relabeling steps into the Markov chain, but computation of the constrained conditional densities might be expensive.

Finally, one should note that similar problems may arise with the Markov mixture model example in Section 4.2.2 of the paper.

```
N <- 1000000          # Size of sample from prior

# Generate N sets of parameters from the prior.

p  <- runif(N)        # Mixing proportion
v  <- rgamma(N,3)/20  # Reciprocal of variance
u1 <- rnorm(N,20,10)  # Mean for component no. 1
u2 <- rnorm(N,20,10)  # Mean for component no. 2

# Find the likelihoods for these sets of parameters.

galaxies <- scan("galaxy.data")

l <- rep(1,N)         # Vector of likelihoods

for (x in galaxies)
{ l <- l * ( p * sqrt(v/(2*pi)) * exp(-v*(x-u1)^2/2)
       + (1-p) * sqrt(v/(2*pi)) * exp(-v*(x-u2)^2/2) );
}

# Find mean likelihood, and associated std. error.

m <- mean(l)          # Estimate for marg. likelihood
e <- sqrt(var(l)/N)   # Standard error for estimate

# Print estimate & std. error, in direct and log form.

cat (m, e, log(m), e/m, "\n")
```

Figure 1: S program for calculating the marginal likelihood of the model with two components having equal variances.