J. R. Statist. Soc. B (1995)
57, *No.* 3, *pp.* 473–484

Bayesian Model Choice via Markov Chain Monte Carlo Methods

By BRADLEY P. CARLIN† and SIDDHARTHA CHIB

University of Minnesota, Minneapolis, USA *Washington University, St Louis, USA*

[Received July 1993. Final revision August 1994]

SUMMARY

Markov chain Monte Carlo (MCMC) integration methods enable the fitting of models of virtually unlimited complexity, and as such have revolutionized the practice of Bayesian data analysis. However, comparison across models may not proceed in a completely analogous fashion, owing to violations of the conditions sufficient to ensure convergence of the Markov chain. In this paper we present a framework for Bayesian model choice, along with an MCMC algorithm that does not suffer from convergence difficulties. Our algorithm applies equally well to problems where only one model is contemplated but its proper size is not known at the outset, such as problems involving integer-valued parameters, multiple changepoints or finite mixture distributions. We illustrate our approach with two published examples.

Keywords: BAYES FACTOR; FINITE MIXTURE MODEL; GIBBS SAMPLER; INTEGER-VALUED PARAMETERS; MODELS OF VARYING SIZE; MULTIPLE CHANGEPOINT MODEL; NON-NESTED MODELS

1. INTRODUCTION

Practitioners are increasingly turning to Bayesian methods for the analysis of complicated statistical models. This move seems due in large part to the advent of inexpensive high speed computers and the simultaneous rapid development of stochastic integration methodology, especially Markov chain Monte Carlo (MCMC) approaches such as the Gibbs sampler (Gelfand and Smith, 1990). As the diversity of the recent applied Bayesian references attests, the MCMC approach is so generally applicable and easy to use that the class of candidate models for a given data set now appears limited only by the user's imagination. However, with this generality has come the temptation to fit models so large that their parameters are unidentified, or nearly so. In extremely complicated hierarchical and random effects settings, this lack of identifiability may be subtle and requires insightful reparameterization of the model (see for example Vines *et al.* (1994) and Gelfand *et al.* (1995)). In even less standard modelling scenarios, the problem may only become apparent through multichain diagnostic methods (Gelman and Rubin, 1992).

Models for which the Markov chain convergence conditions are not satisfied present another problem raised by the apparent broadness of the class of candidate Bayesian models. For example, models of varying size (i.e. having dimension that is not fixed at the outset by the analyst) fall into this category, since at the gth MCMC iteration a parameter may be forced out of the model, creating an absorbing state in the algorithm and thus violating a condition necessary for convergence

†*Address for correspondence*: Division of Biostatistics, School of Public Health, University of Minnesota, Box 303 Mayo Memorial Building, Minneapolis, MN 55455-0392, USA.
E-mail: brad@muskie.biostat.umn.edu

© 1995 Royal Statistical Society 0035–9246/95/57473

(Tierney, 1994). MCMC methods have been used in several recent Bayesian analyses of models of varying size, involving for example integer-valued parameters (West, 1993), multiple changepoints (McCulloch and Tsay, 1994) and finite mixtures (Diebolt and Robert, 1994; Escobar and West, 1995). Although standard MCMC theory does not apply when the dimension of the parameter space is not fixed, some researchers have developed specialized MCMC algorithms for this. For example, Green (1994) offers a 'reversible jump' Metropolis–Hastings algorithm for handling an unknown number of changepoints in a Poisson process, and Grenander and Miller (1994) used a similar approach in comparing competing hypotheses in pattern theory.

In fact, this same convergence issue arises in applying the MCMC technology to *any* problem involving a choice between K competing Bayesian model specifications —even if the candidates are simple and/or nested. To see this, note that we could let M be an integer-valued parameter which indexes the model collection. Since the issue of model selection arises in virtually every Bayesian data analysis, we adopt this framework in what follows and show how Gibbs sampling methodology may be suitably modified to handle choice (or averaging) across a finite collection of models without destroying convergence. Section 2 presents our algorithm and gives advice on its proper tuning to speed convergence. Section 3 illustrates our approach with two data examples, the first involving a choice between two non-nested regression models, and the second fitting a finite mixture model in a case where the proper number of components is unknown.

2. BAYES FACTORS AND SAMPLING-BASED METHODS

Consider the problem of choosing between K models for an observed data vector $\mathbf{y}$. The models need not be nested, and hence we assume that corresponding to each is a distinct parameter vector $\boldsymbol{\theta}_j$, $j = 1, \ldots, K$. As suggested in Section 1, let M be an integer-valued parameter that indexes the model. Our interest lies in $p(M = j|\mathbf{y})$, $j = 1, \ldots, K$, the posterior probabilities of each of the K models, and perhaps also in $p(\boldsymbol{\theta}_j|\mathbf{y})$, $j = 1, \ldots, K$, the posterior distributions of the parameter vectors under the respective models.

Many researchers have suggested the use of MCMC methods to simplify the tasks of assessing model adequacy and selecting the model that is best supported by the data. Gelfand *et al.* (1992) created diagnostics based on the cross-validation predictive distribution $p(y_r|\mathbf{y}_{(r)})$, where $\mathbf{y}_{(r)}$ is the data vector with the rth point deleted. This approach is consistent with the predictive purposes to which the chosen model is often put and has the advantage of remaining feasible where the posterior distributions $p(\boldsymbol{\theta}_j|\mathbf{y})$ are proper but the prior distributions $p(\boldsymbol{\theta}_j)$ are not. Still, a choice between two Bayesian models, say $M = 1$ and $M = 2$, is most commonly based on the *Bayes factor*, which is the ratio of posterior to prior odds in favour of model 2. By Bayes's theorem, this is computable as

$$B_{21} = p(\mathbf{y}|M = 2)/p(\mathbf{y}|M = 1),$$

the ratio of the observed marginal densities for the two models. Newton and Raftery (1994) showed how MCMC samples from the posterior distribution may be used to obtain estimates of these two marginal densities. Although the directness of this solution is attractive, in our limited experience we have found that these estimates

can be somewhat unstable (Albert and Chib, 1993). For a comprehensive review of Bayes factors, their computation and their usage in Bayesian hypothesis testing, see Kass and Raftery (1995).

Alternatively, Carlin and Polson (1991) added the model indicator M into the sampling scheme, so that at convergence the resulting Gibbs iterates $\{M^{(g)}, g = 1, \ldots, G\}$ form a sample from the marginal posterior distribution for M, $p(M|\mathbf{y})$. This solution is in keeping with the 'data augmentation' spirit of many MCMC algorithms but suffers from the aforementioned violations of the convergence condition unless each model can use the same parameterization (e.g. in the case of choosing between competing distributional specifications). George and McCulloch (1993) used a similar approach for selecting an appropriate set of explanatory variables in a regression setting, but again the need to ensure convergence leads to a method that never actually eliminates a regressor from the full model, but only forces it to be close to 0 with high probability.

2.1. *Technical Development*

We circumvent these difficulties by viewing the prior distributions $p(\boldsymbol{\theta}_j)$, $j = 1, \ldots, K$, as part of the Bayesian model specification, and also allowing these distributions to depend on the model indicator M. Suppose that corresponding to model j we have a likelihood $f(\mathbf{y}|\boldsymbol{\theta}_j, M = j)$ and a prior $p(\boldsymbol{\theta}_j|M = j)$. Since we are assuming that M merely provides an indicator about which particular $\boldsymbol{\theta}_j$ is relevant to $\mathbf{y}$, we have that $\mathbf{y}$ is independent of $\{\boldsymbol{\theta}_{i\neq j}\}$ given that $M = j$. In addition, since our primary goal is the computation of Bayes factors, we assume that each prior $p(\boldsymbol{\theta}_j|M = j)$ is proper (though possibly quite vague). For simplicity we assume complete independence among the various $\boldsymbol{\theta}_j$ given the model indicator M, and thus we may complete the Bayesian model specification by choosing proper 'pseudopriors' $p(\boldsymbol{\theta}_j|M \neq j)$. From our conditional independence assumptions,

$$p(\mathbf{y}|M = j) = \int f(\mathbf{y}|\boldsymbol{\theta}, M = j)\, p(\boldsymbol{\theta}|M = j)\, d\boldsymbol{\theta} = \int f(\mathbf{y}|\boldsymbol{\theta}_j, M = j)\, p(\boldsymbol{\theta}_j|M = j)\, d\boldsymbol{\theta}_j,$$

and so the form given to $p(\boldsymbol{\theta}_j|M \neq j)$ is irrelevant. Thus, as the name suggests, a pseudoprior is not really a prior but only a conveniently chosen linking density, required to define completely the joint model specification. (We defer the specifics of pseudoprior selection until Section 2.2.) Then given prior model probabilities $\pi_j \equiv P(M = j)$ such that $\Sigma_{j=1}^{K} \pi_j = 1$, and writing $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K\}$ the joint distribution of $\mathbf{y}$ and $\boldsymbol{\theta}$ when $M = j$ is

$$p(\mathbf{y}, \boldsymbol{\theta}, M = j) = f(\mathbf{y}|\boldsymbol{\theta}_j, M = j)\left\{\prod_{i=1}^{K} p(\boldsymbol{\theta}_i|M = j)\right\}\pi_j.$$

Now, to implement the Gibbs sampler, we need the full conditional distributions of each $\boldsymbol{\theta}_j$ and M. The former is given by

$$p(\boldsymbol{\theta}_j|\boldsymbol{\theta}_{i\neq j}, M, \mathbf{y}) \propto \begin{cases} f(\mathbf{y}|\boldsymbol{\theta}_j, M = j)\, p(\boldsymbol{\theta}_j|M = j), & M = j, \\ p(\boldsymbol{\theta}_j|M \neq j), & M \neq j, \end{cases} \tag{1}$$

i.e. when $M = j$ we generate from the usual model j full conditional; when $M \neq j$ we generate from the linking density (we shall use the terms 'pseudoprior' and

'linking density' interchangeably). In cases where $p(\boldsymbol{\theta}_j|M=j)$ is taken to be conjugate with its likelihood, both of these generations are straightforward.

For M we have

$$p(M=j|\boldsymbol{\theta},\mathbf{y})=\frac{f(\mathbf{y}|\boldsymbol{\theta}_j,M=j)\left\{\prod_{i=1}^{K}p(\boldsymbol{\theta}_i|M=j)\right\}\pi_j}{\sum_{k=1}^{K}f(\mathbf{y}|\boldsymbol{\theta}_k,M=k)\left\{\prod_{i=1}^{K}p(\boldsymbol{\theta}_i|M=k)\right\}\pi_k}. \tag{2}$$

Since M is a discrete finite parameter, its generation is routine as well. Hence all the required full conditional distributions are well defined, and under the usual regularity conditions (Roberts and Smith, 1993) the algorithm will produce samples from the correct joint posterior distribution. In particular, the ratio

$$\hat{p}(M=j|\mathbf{y})=\frac{\text{number of } M^{(g)}=j}{\text{total number of } M^{(g)}}, \qquad j=1,\ldots,K, \tag{3}$$

provides simple estimates that may be used to compute the Bayes factor between any two of the models. Standard errors for these estimates are easy to obtain even if the $M^{(g)}$ output stream exhibits autocorrelation, through the use of batching or perhaps more sophisticated spectral decomposition techniques (Ripley (1987), chapter 6). We illustrate such computations in the examples of Section 3.

2.2. *Implementational Notes*

Notice that, in contrast with equation (3), summarization of the collection of $\boldsymbol{\theta}_j^{(g)}$ samples is not useful. This is because what is of interest in this case is not the marginal posterior densities $p(\boldsymbol{\theta}_j|\mathbf{y})$, but rather the *conditional* posterior densities $p(\boldsymbol{\theta}_j|M=j,\mathbf{y})$. However, suppose that in addition to $\boldsymbol{\theta}$ we have a vector of nuisance parameters, say η, common to all models. Then the fully marginal posterior density of η, $p(\eta|\mathbf{y})$, may be of some interest. We would not need to create a pseudoprior for η, since the data are informative about η regardless of the value of M. But in this case great caution must be taken to ensure that η has the same interpretation in both models. For example, suppose that we wish to choose between the two nested regression models

$$M=1: \quad y_i=\alpha+\epsilon_i, \qquad \epsilon_i \overset{\text{IID}}{\sim} N(0,\sigma^2), \quad i=1,\ldots,n,$$

and

$$M=2: \quad y_i=\alpha+\beta x_i+\epsilon_i, \qquad \epsilon_i \overset{\text{IID}}{\sim} N(0,\tau^2), \quad i=1,\ldots,n,$$

so that $\boldsymbol{\theta}_1=\sigma$, $\boldsymbol{\theta}_2=(\beta,\tau)$ and $\eta=\alpha$. But α is playing two different roles here: 'grand mean' in model 1, and 'intercept' in model 2. The corresponding posteriors could be quite different if, for example, the observed y_i-values were centred near 0, while the observed x_i-values were centred far from 0. The resulting bimodal shape for $p(\alpha|\mathbf{y})$ could wreak great havoc with convergence of the Gibbs algorithm, making jumps between $M=1$ and $M=2$ extremely unlikely.

Poor choices of the linking densities $p(\boldsymbol{\theta}_j|M\neq j)$ can have a similar deleterious effect on convergence. Good choices will produce $\boldsymbol{\theta}_j^{(g)}$-values that are consistent with the data, so that $p(M=j|\boldsymbol{\theta},\mathbf{y})$ will still be reasonably large at the next M

update step. Failure to generate competitive pseudoprior values will again result in intolerably high autocorrelations in the $M^{(g)}$-chain: hence slow convergence. To avoid this, we recommend obtaining preliminary estimates of the model-specific posterior distributions $p(\boldsymbol{\theta}_j | M = j, \mathbf{y})$, perhaps by using first-order (normal) approximations, or other parametric forms designed to mimic the output from K individual MCMC runs. Matching the linking densities as nearly as possible to the true model-specific posteriors should produce a reasonably well mixing final algorithm. Note that we are *not* using the data to help to select the prior, but only the pseudoprior. More specific guidance on pseudoprior selection is provided in the context of our Section 3 examples.

Finally, if for a particular data set one of the $p(M = j | \mathbf{y})$ is extremely large, the realized chain will exhibit slow convergence due to the resulting nearly absorbing state in the algorithm. In this case, the π_j may be adjusted to correct the imbalance; the final value of B_{ji} will still reflect the true odds in favour of $M = j$ suggested by the data. This adjustment may be done adaptively during the early stages of the algorithm, before samples are retained for Bayes factor estimation.

It is tempting to skip the generation of actual pseudoprior values and instead simply to keep $\boldsymbol{\theta}_j^{(g)}$ at its current value when $M^{(g)} \neq j$. But, although seemingly reasonable, such an algorithm is clearly not a Gibbs sampler in the strict sense, since the nodes visited are determined by the current values in the realized Markov chain. We might instead attempt to portray this non-visitation as a Metropolis–Hastings rejection step, but then we arrive at a chain with a transition kernel that depends on the particular $M^{(g)}$-values generated. Gelfand and Sahu (1994) gave a simple example of such a chain that does *not* converge to the proper stationary distribution. Hence the convergence properties of this simplified algorithm are unclear. In any case, we cannot dispense with the linking densities entirely, since they are required for the M update step in equation (2).

3. DATA EXAMPLES

3.1. *Non-nested Regression Models*

Efron (1984) considered fitting two plausible straight line models to the data set of Williams (1959), displayed in Table 1. For $n = 42$ specimens of radiata pine, the maximum compressive strength parallel to the grain y_i was measured, along with the specimen's density, x_i, and its density adjusted for resin content, z_i (resin contributes much to the density but little to the strength of the wood). It is desired to compare the two models $M = 1$ and $M = 2$ where

$$M = 1: \quad y_i = \alpha + \beta x_i + \epsilon_i, \qquad \epsilon_i \overset{\text{IID}}{\sim} N(0, \sigma^2), \quad i = 1, \ldots, n,$$

and

$$M = 2: \quad y_i = \gamma + \delta z_i + \epsilon_i, \qquad \epsilon_i \overset{\text{IID}}{\sim} N(0, \tau^2), \quad i = 1, \ldots, n.$$

Hence $\boldsymbol{\theta}_1 = (\alpha, \beta, \sigma)$ and $\boldsymbol{\theta}_2 = (\gamma, \delta, \tau)$. After centring the x_i and z_i at their means, we place $N((3000, 185)^{\mathrm{T}}, \mathrm{diag}(10^6, 10^4))$ priors on $(\alpha, \beta)^{\mathrm{T}}$ and $(\gamma, \delta)^{\mathrm{T}}$, and inverse gamma priors on σ^2 and τ^2, both having mean and standard deviation equal to 300^2. Each of these priors is roughly centred on the appropriate least squares parameter estimate but is extremely vague (though still proper). The inverse gamma

TABLE 1
Radiata pine compressive strength data

Case (i)	y_i	x_i	z_i	*Case (i)*	y_i	x_i	z_i
1	3040	29.2	25.4	22	3840	30.7	30.7
2	2470	24.7	22.2	23	3800	32.7	32.6
3	3610	32.3	32.2	24	4600	32.6	32.5
4	3480	31.3	31.0	25	1900	22.1	20.8
5	3810	31.5	30.9	26	2530	25.3	23.1
6	2330	24.5	23.9	27	2920	30.8	29.8
7	1800	19.9	19.2	28	4990	38.9	38.1
8	3110	27.3	27.2	29	1670	22.1	21.3
9	3160	27.1	26.3	30	3310	29.2	28.5
10	2310	24.0	23.9	31	3450	30.1	29.2
11	4360	33.8	33.2	32	3600	31.4	31.4
12	1880	21.5	21.0	33	2850	26.7	25.9
13	3670	32.2	29.0	34	1590	22.1	21.4
14	1740	22.5	22.0	35	3770	30.3	29.8
15	2250	27.5	23.8	36	3850	32.0	30.6
16	2650	25.6	25.3	37	2480	23.2	22.6
17	4970	34.5	34.2	38	3570	30.3	30.3
18	2620	26.2	25.7	39	2620	29.9	23.8
19	2900	26.7	26.4	40	1890	20.8	18.4
20	1670	21.1	20.0	41	3030	33.2	29.4
21	2540	24.1	23.9	42	3030	28.2	28.2

priors for σ^2 and τ^2 also double as the linking densities (pseudopriors) for these parameters, whereas for the remaining components we use independent univariate normal linking densities that roughly equal the corresponding first-order approximation to the posterior. More specifically, we let $\alpha|(M = 2) \sim N(3000, 52^2)$, $\beta|(M = 2) \sim N(185, 12^2)$, $\gamma|(M = 1) \sim N(3000, 43^2)$ and $\delta|(M = 1) \sim N(185, 9^2)$. (An alternative would be to transform σ^2 and τ^2 to the log-scale and to use trivariate normal approximations to the posterior densities for $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ as our pseudopriors. Although this more-involved approach would account for the dependence within $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, our simple method appears adequate in this example, as we show below.)

Using $\pi_1 = \pi_2 = 0.5$ for an initial run of five parallel Gibbs chains for 5000 iterations each, we observed only eight instances where the generated $M^{(g)}$ equalled 1. To correct this imbalance, we settled on $\pi_1 = 0.9995$ and $\pi_2 = 0.0005$, and performed our production run of five parallel Gibbs chains for 50000 iterations each. We started the five chains at disparate points in the sample space and, after considering plots of the realized chains, sample autocorrelations and the monitoring statistic of Gelman and Rubin (1992), were satisfied with the convergence of our algorithm. The resulting point estimates from equation (3) are $\hat{p}(M = 1|\mathbf{y}) = 0.3114$ and $\hat{p}(M = 2|\mathbf{y}) = 0.6886$, so that an estimated standard deviation for these estimates (assuming independent samples) is given by $\sqrt{(0.3114 \times 0.6886/250000)} = 0.00093$. However, the realized $M^{(g)}$-chains exhibited significant positive autocorrelation through lag 4, so we batched the output into 2500 groups of length 100 to obtain the somewhat larger (and presumably more accurate) estimate $\widehat{\text{sd}}\{\hat{p}(M = 2|\mathbf{y})\} = 0.00166$. Hence an approximate 95% confidence interval for

the posterior probability that $M = 2$ is given by (0.6853, 0.6918). Converting to Bayes factors, we have a point estimate of 4420 for B_{21}, with (4353, 4487) as the corresponding 95% confidence interval—overwhelming evidence in favour of the adjusted density model. This is consistent with (though apparently more precise than) the frequentist result reported by Efron (1984), namely a two-sided significance level of less than 0.10.

Our prescription of matching the pseudopriors to the model-specific posteriors provides a systematic approach to producing a well mixing algorithm. To investigate the degradation in performance resulting from a poorly matched pseudoprior, we doubled the standard deviations in our normal pseudopriors for α, β, γ and δ. We found that, whereas the estimated Bayes factor remained roughly the same, the $M^{(g)}$-chains now exhibited significant positive autocorrelation through lag 10, and our batched standard deviation estimate $\widehat{\mathrm{sd}}\{\hat{p}(M = 2|\mathbf{y})\}$ increased from 0.00166 to 0.00231. Similarly, quadrupling these pseudoprior standard deviations led to significant positive autocorrelation through lag 40, and a further increase in $\widehat{\mathrm{sd}}\{\hat{p}(M = 2|\mathbf{y})\}$ to 0.00413.

3.2. *Mixture Model with Unknown Number of Components*

Another important illustration of our ideas is possible in the context of finite mixture models. Evans *et al.* (1992) used Monte Carlo Bayesian methods to analyse a two-component normal mixture model under a non-informative prior. In our case, we wish to compare a D_1-component normal mixture with a mixture having D_2 components. To illustrate, consider the data set in Table 2 on velocities of 82 galaxies from six well-separated conic sections of the corona borealis region, originally presented by Postman *et al.* (1986). A histogram of these data is shown in Fig. 1(a). An important issue with these data is whether they arise from a multimodal distribution, as would be implied by astronomical theories concerning the clustering of galaxies. A nonparametric kernel density estimate could be used to estimate the number of modes (i.e. the number of galactic clusters), but the answer to this important question would then be quite sensitive to the choice of

TABLE 2
Velocities (kilometres per second) for galaxies in the corona borealis region

9172	9350	9483	9558	9775	10227
10406	16084	16170	18419	18552	18600
18927	19052	19070	19330	19343	19349
19440	19473	19529	19541	19547	19663
19846	19856	19863	19914	19918	19973
19989	20166	20175	20179	20196	20215
20221	20415	20629	20795	20821	20846
20875	20986	21137	21492	21701	21814
21921	21960	22185	22209	22242	22249
22314	22374	22495	22746	22747	22888
22914	23206	23241	23263	23484	23538
23542	23666	23706	23711	24129	24285
24289	24366	24717	24990	25633	26960
26995	32065	32789	34279		

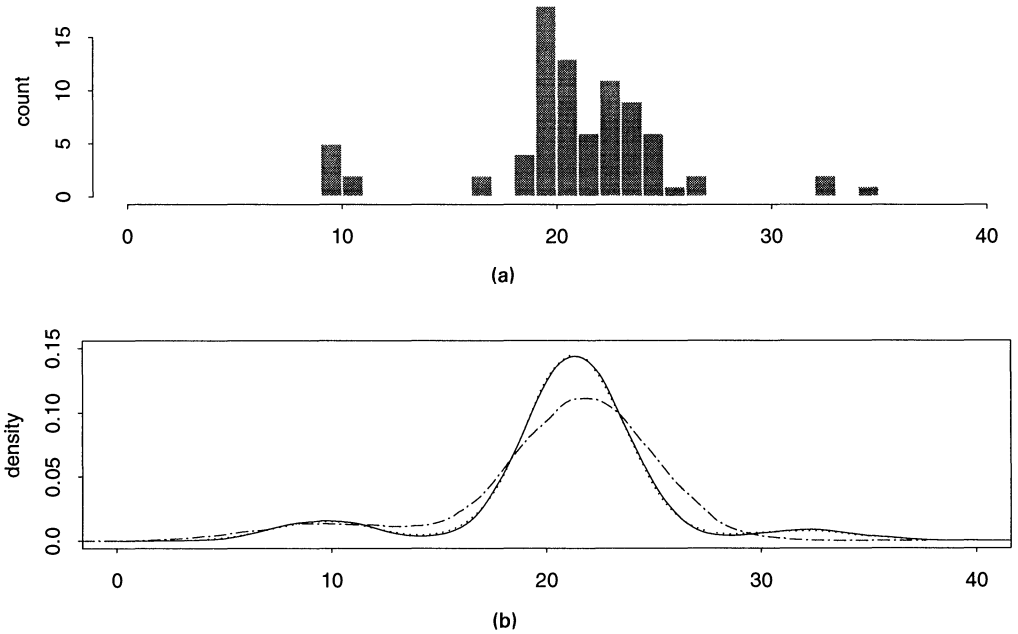

Fig. 1. (a) Histogram of velocities ($\times 10^3$ km s^{-1}) and (b) estimated predictive densities, normal mixture models (———, four component; ·····, three component; — · — ·, two component) for the galaxy data

smoothing window width. Roeder (1990) developed a confidence set of plausible densities for these data that featured between three and seven modes. On the basis of this information and the appearance of the histogram, we compare a three-component normal mixture model with a mixture having four components by using our parametric Monte Carlo Bayesian approach.

Suppose that the density function of the datum y_i in the kth component of the jth model is given by $\phi(y_i|\mu_{jk}, \sigma_j^2)$, where $\phi(\ |\ ,\ )$ denotes the normal density function, and let q_{jk} denote the corresponding mixing probabilities for each component in the mixture. Then, given a vector of independent samples $\mathbf{y} = (y_1, \ldots, y_n)$ the two models under consideration are given by

$$M = 1: \quad f(y_i|\mu_1, \sigma_1^2, q_1) = \sum_{k=1}^{3} q_{1k}\phi(y_i|\mu_{1k}, \sigma_1^2),$$

and

$$M = 2: \quad f(y_i|\mu_2, \sigma_2^2, q_2) = \sum_{k=1}^{4} q_{2k}\phi(y_i|\mu_{2k}, \sigma_2^2),$$

where in obvious notation $\mu_1 = (\mu_{11}, \mu_{12}, \mu_{13})$, $q_1 = (q_{11}, q_{12}, q_{13})$ and similarly for μ_2 and q_2.

As discussed in West (1992) and Diebolt and Robert (1994), Monte Carlo Bayesian inference for either model is facilitated by supplementing the parameters with latent observation-specific index variables: $s_{i1} \in \{1, 2, 3\}$ for $M = 1$; $s_{i2} \in \{1, 2, 3, 4\}$ for $M = 2$. These s_{ij} are distributed (independently over i) according to q_j

and signify the component population from which each observation arose. Hence, for a given model, the complete conditional distributions required to implement the Gibbs sampler arise as follows.

(a) Conditioned on $s_j = \{s_{ij}\}$, the observations are classified into three different populations for $j = 1$, and four different populations for $j = 2$. The complete conditional distributions of μ_{jk} and σ_j^2 are obtained easily provided that each component mean follows an independent normal prior and the σ_j^2 follow inverse gamma priors.

(b) Conditioned on μ_j, σ_j^2 and s_j, a Dirichlet prior on q_j leads to a Dirichlet posterior as well, with parameters of the prior revised according to the number of observations assigned to each population.

(c) Conditioned on μ_j, σ_j^2 and q_j, the posterior probability mass function of s_{ij} is given by the expression $\Pr(s_{ij} = k | \mathbf{y}, \mu_2, \sigma_j^2) \propto q_{jk}\phi(y_i | \mu_{jk}, \sigma_j^2)$, where k runs from 1 to 3 if $j = 1$ and from 1 to 4 if $j = 2$.

On the basis of a run of 30000 iterations for each model, we obtain the results summarized in Tables 3 and 4. The prior means and standard deviations listed are from the rather vague normal, Dirichlet and inverse gamma distributions chosen for μ_{jk}, q_j and σ_j^2 respectively. Also, 'Num SE' denotes the numerical standard error of the posterior mean, while 'Corr' gives the sample first-order autocorrelation of the simulated values. Table 4 reports high autocorrelations for μ_{22}, μ_{23}, q_{22} and q_{23}, and posterior standard deviations for q_{22} and q_{23} that are slightly *larger* than their prior values. These facts suggest that there is not enough information in the likelihood to estimate all four components and their mixing probabilities accurately in the larger model. A similar message is conveyed by Fig. 1(b), which for each model plots the estimated predictive distribution for a future observation y_f, $\hat{f}(y_f | \mathbf{y})$. (The predictive density for the poorly fitting two-component mixture model is also shown for comparison.) The substantial overlap of the second and third modes in the four-component model suggests almost no improvement in fit over the three-component model.

For a more formal comparison of these two models, we use the approach of Section 2 in conjunction with the algorithm for sampling the parameters for a given

TABLE 3
Summary of the three-component normal mixture model for the galaxy data

Parameter	*Prior*		*Posterior*			
	Mean	*Standard deviation*	*Mean*	*Num SE*	*Standard deviation*	*Corr*
μ_{11}	9.000	5.000	9.674	0.005	0.823	0.070
μ_{12}	18.000	5.000	21.337	0.002	0.273	0.040
μ_{13}	30.000	5.000	31.922	0.009	1.258	0.237
σ_1^2	20.000	20.000	5.224	0.005	0.832	0.071
q_{11}	0.333	0.236	0.095	0.000	0.032	0.012
q_{12}	0.333	0.236	0.854	0.000	0.039	0.047
q_{13}	0.333	0.236	0.051	0.000	0.025	0.087

TABLE 4
Summary of the four-component normal mixture model for the galaxy data

Parameter	*Prior*		*Posterior*			
	Mean	*Standard deviation*	*Mean*	*Num SE*	*Standard deviation*	*Corr*
μ_{21}	9.000	5.000	9.669	0.004	0.764	0.054
μ_{22}	18.000	5.000	20.838	0.019	0.872	0.811
μ_{23}	22.000	5.000	21.926	0.021	0.909	0.828
μ_{24}	30.000	5.000	32.110	0.008	1.178	0.233
σ_2^2	15.000	15.000	4.422	0.014	0.933	0.455
q_{21}	0.136	0.100	0.092	0.000	0.030	0.008
q_{22}	0.364	0.139	0.430	0.003	0.144	0.884
q_{23}	0.364	0.139	0.427	0.003	0.144	0.883
q_{24}	0.136	0.100	0.051	0.000	0.023	0.060

model described in (a)–(c) above, i.e. we sample over M, $\boldsymbol{\theta}_1 = (\mu_1, \sigma_1^2, q_1, \{s_{i1}\})$ and $\boldsymbol{\theta}_2 = (\mu_2, \sigma_2^2, q_2, \{s_{i2}\})$. Given the large dimension of the $\boldsymbol{\theta}_j$ it is important that the linking densities (pseudopriors) be carefully specified if frequent moves between models are to be realized. As such, we matched normal, Dirichlet and inverse gamma linking densities for the μ_j, q_j and σ_j^2 respectively with the appropriate posterior estimates in Tables 3 and 4. Note that the functional form of $f(\mathbf{y}|\mu_j, q_j, \sigma_j^2, M = j)$ is available for use in equation (2) as a product of the terms in equation (4). Therefore, if samples for (μ_j, q_j, σ_j^2) given $M = j$ are obtained by resampling from the output of the model-specific preliminary runs, we may implement our algorithm in Section 2 without including the latent data in the sampling order. Alternatively, an acceptably accurate pseudoprior mass function for each s_{ij} can be determined by the observed $\{s_{ij}^{(g)}\}$ relative frequencies in the model-specific preliminary runs.

Using the latter approach, the same model-specific priors $p(\boldsymbol{\theta}_j|M = j)$ as before and letting $\pi_1 = 0.35$ and $\pi_2 = 0.65$, the combined sampler was run for $G = 30000$ iterations. For model choice, it suffices to focus on the chain corresponding to M. Regardless of the initial starting model, the sampler is observed to move reasonably well between the two models. The point estimate (3) of $p(M = 2|\mathbf{y})$ is found to be 0.5153, with an estimated standard error of 0.0146 (the latter based on the means of 300 batches of length 100). This translates into a point estimate for B_{21} of 0.572, and a 95% confidence interval of (0.511, 0.642). (By comparison, the estimated Bayes factor in favour of the three-component mixture over the two-component mixture is roughly 196000.) Hence the Bayes factor agrees with our earlier assessment of indifference between the three- and four-component models. However, it is important to note that, because of the aforementioned inability of the data to identify the four-component model fully, the Bayes factor can be made to prefer this larger model by altering the model-specific priors. For example, if we replace the Dirichlet(1, 1, 1) and Dirichlet(1.5, 4, 4, 1.5) priors on q_1 and q_2 with the more informative Dirichlet(3, 8, 3) and Dirichlet(3, 8, 3, 3) distributions respectively, the Bayes factor estimate increases to 284. This highlights the well-

known sensitivity of Bayes factors to the prior inputs when the likelihood for one or both of the models does not convey much information.

ACKNOWLEDGEMENTS

The work of the first-named author was supported in part by National Institute of Allergy and Infectious Diseases First Independent Research Support and Transition award 1-R29-AI33466. The authors thank Alan Gelfand, Andrew Gelman, John Geweke, Charles Geyer, Jay Kadane, Peter Müller and Adrian Raftery for helpful discussions during the course of our investigation, and two referees whose comments led to substantial improvements in the presentation.

REFERENCES

Albert, J. and Chib, S. (1993) Bayesian model checking for binary and categorical response data. *Technical Report*. Department of Mathematics and Statistics, Bowling Green State University, Bowling Green.

Carlin, B. P. and Polson, N. G. (1991) Inference for nonconjugate Bayesian models using the Gibbs sampler. *Can. J. Statist.*, **19**, 399–405.

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

Efron, B. (1984) Comparing non-nested linear models. *J. Am. Statist. Ass.*, **79**, 791–804.

Escobar, M. and West, M. (1995) Bayesian density estimation and inference using mixtures. *J. Am. Statist. Ass.*, to be published.

Evans, M., Guttman, I. and Olkin, I. (1992) Numerical aspects in estimating the parameters of a mixture of normal distributions. *J. Comput. Graph. Statist.*, **1**, 351–365.

Gelfand, A. E., Dey, D. K. and Chang, H. (1992) Model determination using predictive distributions with implementation via sampling-based methods (with discussion). In *Bayesian Statistics 4* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 147–167. Oxford: Oxford University Press.

Gelfand, A. E. and Sahu, S. K. (1994) On Markov chain Monte Carlo acceleration. *J. Comput. Graph. Statist.*, **3**, 261–276.

Gelfand, A. E., Sahu, S. K. and Carlin, B. P. (1995) Efficient parameterizations for generalized linear mixed models (with discussion). In *Bayesian Statistics 5* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith). Oxford: Oxford University Press. To be published.

Gelfand, A. E. and Smith, A. F. M. (1990) Sampling-based approaches to calculating marginal densities. *J. Am. Statist. Ass.*, **85**, 398–409.

Gelman, A. and Rubin, D. B. (1992) Inference from iterative simulation using multiple sequences (with discussion). *Statist. Sci.*, **7**, 457–511.

George, E. I. and McCulloch, R. E. (1993) Variable selection via Gibbs sampling. *J. Am. Statist. Ass.*, **88**, 881–889.

Green, P. J. (1994) A Bayesian analysis of the multiple change-point problem for a Poisson process. *Technical Report*. Department of Mathematics, University of Bristol, Bristol.

Grenander, U. and Miller, M. I. (1994) Representations of knowledge in complex systems (with discussion). *J. R. Statist. Soc.* B, **56**, 549–603.

Kass, R. E. and Raftery, A. E. (1995) Bayes factors and model uncertainty. *J. Am. Statist. Ass.*, to be published.

McCulloch, R. E. and Tsay, R. S. (1994) Bayesian analysis of threshold autoregressive processes with a random number of regimes. In *Computing Science and Statistics: Proc. 25th Symp. Interface* (eds M. E. Tarter and M. D. Lock), pp. 253–262. Fairfax Station: Interface Foundation of North America.

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

Postman, M., Huchra, J. P. and Geller, M. J. (1986) Probes of large-scale structures in the Corona Borealis region. *Astron. J.*, **92**, 1238–1247.

Ripley, B. D. (1987) *Stochastic Simulation.* New York: Wiley.

Roberts, G. O. and Smith, A. F. M. (1993) Simple conditions for the convergence of the Gibbs sampler and Metropolis–Hastings algorithms. *Stoch. Processes Appl.*, **49**, 207–216.

Roeder, K. (1990) Density estimation with confidence sets exemplified by superclusters and voids in galaxies. *J. Am. Statist. Ass.*, **85**, 617–624.

Tierney, L. (1994) Markov chains for exploring posterior distributions (with discussion). *Ann. Statist.*, to be published.

Vines, S. K., Gilks, W. R. and Wild, P. (1994) Fitting Bayesian multiple random effects models. Submitted to *Statist. Comput.*

West, M. (1992) Modelling with mixtures (with discussion). In *Bayesian Statistics 4* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 503–524. Oxford: Oxford University Press.

——(1993) Predictions for finite populations under biased sampling. *Discussion Paper 93-A11.* Institute for Statistics and Decision Sciences, Duke University, Durham.

Williams, E. (1959) *Regression Analysis.* New York: Wiley.