MCMC Methods for Computing Bayes Factors: A Comparative Review

By Cong Han and Bradley P. Carlin

Division of Biostatistics, School of Public Health, University of Minnesota, Mayo Mail Code 303, Minneapolis, Minnesota 55455–0392, U.S.A.

Correspondence author: Bradley P. Carlin
telephone: (612) 624-6646
fax: (612) 626-0660
email: brad@muskie.biostat.umn.edu

April 3, 2001

MCMC Methods for Computing Bayes Factors: A Comparative Review

Summary

The problem of calculating posterior probabilities for a collection of competing models and associated Bayes factors continues to be a formidable challenge for applied Bayesian statisticians. Current approaches that take advantage of modern Markov chain Monte Carlo (MCMC) computing methods include those that attempt to sample over some form of the joint space created by the model indicators and the parameters for each model, others that sample over the model space alone, and still others that attempt to estimate the marginal likelihood of each model directly (since the collection of these is equivalent to the collection of model probabilities themselves). In this paper we review several of these methods, and subsequently compare them in the context of three examples, the first a simple regression example, the second a more challenging hierarchical longitudinal model, and the third a binary data latent variable model. We find that the joint model-parameter space search methods perform adequately but can be difficult to program and tune, while the marginal likelihood methods are often less troublesome and require less in the way of additional coding. Our results suggest that the latter methods may be most appropriate for practitioners working in many standard model choice settings, while the former remain important for comparing models of varying dimension (e.g. multiple changepoint models), or models whose parameters cannot be easily updated in relatively few blocks. We caution however that *all* of the methods we compare require significant human and computer effort, suggesting that less formal Bayesian model choice methods may offer a more realistic alternative in many cases.

Key words: Bayesian model choice; Gibbs sampler; marginal likelihood; Metropolis-Hastings algorithm; reversible jump sampler.

1 Introduction

The traditional approach to Bayesian model selection is concerned with the following situation. Suppose the observed data $\mathbf{y}$ are generated by a model $j \in \mathcal{M}$, where $\mathcal{M}$ is the finite set of competing models. Corresponding to model j, there is a distinct unknown parameter vector $\boldsymbol{\theta}_j$ of dimension n_j, and a prior model probability $\pi_j \equiv P(M = j)$, where $\sum_{j \in \mathcal{M}} \pi_j = 1$. Let Θ_j be the set of all possible values for $\boldsymbol{\theta}_j$, so that $\boldsymbol{\theta}_j \in \Theta_j \subset \mathbb{R}^{n_j}$, and let $\boldsymbol{\theta}$ be the collection of all model-specific $\boldsymbol{\theta}_j$'s. Interest lies in obtaining the posterior probabilities for the various models, $P(M = j|\mathbf{y})$, either to arrive at a single "best" model, or to determine the posterior distribution of some quantity of interest ψ which is common to all models via model averaging (c.f. Carlin and Louis, 2000, p.49). Due to possible differences among the π_j, a choice between two models (say, models 1 and 2) is often based not on the posterior odds, but on the *Bayes factor*,

$$B_{21} = \frac{P(M = 2 \mid \mathbf{y})/P(M = 1 \mid \mathbf{y})}{P(M = 2)/P(M = 1)} = \frac{p(\mathbf{y} \mid M = 2)}{p(\mathbf{y} \mid M = 1)}\,, \tag{1}$$

i.e., the ratio of posterior to prior odds in favor of model 2. The Bayes factor is often thought of as the weight of evidence in favor of model 2 provided "by the data," though Lavine and Schervish (1999) show that a more accurate interpretation is that B_{21} captures the *change* in the odds in favor of model 2 as we move from prior to posterior. Since the prior probabilities π_j are known in advance, equation (1) shows that the collection of *marginal likelihoods* $p(\mathbf{y}|M = j)$ is equivalent to the model probabilities themselves, and hence could also be considered the quantities of key interest. Kass and Raftery (1995) provide a comprehensive review of Bayes factors, including their interpretation, computation or approximation, robustness to the model-specific prior distributions $p(\boldsymbol{\theta}_j|M = j)$ selected, and exemplification in a variety of scientific applications. Recent textbooks reviewing modern computational approaches for Bayes factors include Carlin and Louis (2000) and

Chen, Shao, and Ibrahim (2000).

As a side comment, note that if the model-specific prior $p(\boldsymbol{\theta}_j|M=j)$ is improper (as is often the case in objective Bayesian data analyses using noninformative priors), then the marginal likelihood $p(\mathbf{y}|M=j) = \int f(\mathbf{y}|\boldsymbol{\theta}_j, M=j)p(\boldsymbol{\theta}_j|M=j)d\boldsymbol{\theta}_j$ necessarily is as well, and so the Bayes factor (1) is not well-defined. Numerous solutions have been proposed to this problem, ranging from the use of various "pseudo-Bayes factor" approaches, such as the *intrinsic Bayes factor* (Berger and Pericchi, 1996) and the *fractional Bayes factor* (O'Hagan, 1995), to the use of asymptotic methods and the *Bayesian Information Criterion*, or BIC (Schwarz, 1978). In this paper, however, the methods we seek to review and compare have all been developed with proper Bayes factors in mind. As such, we do not discuss these objective Bayesian alternatives further, instead referring the reader to the excellent recent review and critical comparison of these techniques by Berger and Pericchi (2001).

Several methods (c.f. Carlin and Louis, 2000, Sec. 6.3.1) seek to estimate the marginal likelihood $\hat{p}(\mathbf{y}|M=j)$ directly for each model using ordinary Monte Carlo sampling, but these approaches are difficult to implement for high-dimensional models. In such cases, Bayesian analyses are now typically carried out using Markov chain Monte Carlo (MCMC) computing methods such as the Gibbs sampler (Gelfand and Smith, 1990) and the Metropolis-Hastings (M-H) algorithm (Hastings, 1970). While these algorithms enable direct estimation of posterior and predictive quantities of interest, they do not readily lend themselves to estimating aspects of the marginal distribution. Chib (1995) does however give an indirect method for estimating marginal likelihoods from Gibbs sampling output, an idea that has recently been extended to output from the Metropolis-Hastings algorithm by Chib and Jeliazkov (2001).

A slightly more direct (and more common) approach to estimating posterior model probabilities using MCMC has been to include the model indicator M as a parameter in the sampling order. Once the sampler has converged, the proportion of times the sampler visits model j is then a

simulation-consistent estimate of $P(M = j|\mathbf{y})$. In fact, it is sometimes possible to integrate the parameters $\boldsymbol{\theta}$ out of the joint specification $(\boldsymbol{\theta}, M)$ in closed form, yielding a sampler that operates (quite efficiently) over the model space alone. Unfortunately, such an approach is only possible for fairly special model classes, such as graphical models for discrete data (Madigan and York, 1995) and multiple regression models with conjugate priors (Raftery, Madigan, and Hoeting, 1997). As a result, most model settings require that the MCMC search be over the model and parameter space jointly. That is, the joint sampling space is $\mathcal{M} \times \prod_{j \in \mathcal{M}} \Theta_j \subset \mathcal{M} \times \prod_{j \in \mathcal{M}} \mathbb{R}^{n_j}$. Besides the marginal posterior model probabilities $P(M = j|\mathbf{y})$, this joint search also permits posterior estimation of the parameters under each model, $p(\boldsymbol{\theta}_j|M = j, \mathbf{y})$, simply by conditioning on the samples produced when the chain is currently in state $M = j$.

In this paper, we seek to compare and contrast some of methods currently available for MCMC-based Bayes factor computation, both in terms of ease of use and understanding, and in the accuracy of the results they obtain. In Section 2 we review several such methods, including the product space search of Carlin and Chib (1995), the reversible jump sampler of Green (1995), the "Metropolized Carlin and Chib" (MCC) method of Dellaportas, Forster, and Ntzoufras (2001) and the related composite model space approach of Godsill (2001), and the marginal likelihood approaches of Chib (1995) and Chib and Jeliazkov (2001). In Section 3 we compare all of these methods in the context of a simple, non-hierarchical regression example. Here we find all the methods perform reasonably well, albeit after differing amounts and kinds of tuning. Section 4 then considers the more difficult case of choosing between a linear and a piecewise-linear longitudinal model with random effects in a dataset drawn from an large AIDS clinical trial. Here the much larger parameter space and extended hierarchy renders some of the methods infeasible and suggests a general strategy for choosing amongst models of this type. Section 5 compares the most promising methods in a binary data setting where an auxiliary set of latent variables may or may not be used to assist with the

computation. Finally, in Section 6 we discuss our findings and offer recommendations on how to proceed in various Bayesian model choice settings.

2 Description of Methods

2.1 Product Space Search

The method of Carlin and Chib (1995), which we abbreviate "CC", assumes that corresponding to model j, the likelihood is $f(\mathbf{y}|\boldsymbol{\theta}_j, M = j)$ and the prior is $p(\boldsymbol{\theta}_j|M = j)$. It is assumed that M is merely an indicator of which $\boldsymbol{\theta}_j$ is relevant to $\mathbf{y}$, and therefore $\mathbf{y}$ is independent of $\{\boldsymbol{\theta}_{j'\neq j}\}$ given the model indicator M. Thus as mentioned above, the sampler operates over the *product space*, $\mathcal{M} \times \prod_{j\in\mathcal{M}} \Theta_j$. As with all the methods we discuss, proper priors are required for the sensible computation of proper (not pseudo) Bayes factors; prior independence among $\boldsymbol{\theta}_j$'s given M is also assumed for simplicity. Note that under the conditional independence assumption,

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

where $\boldsymbol{\theta}$ is the collection $\{\boldsymbol{\theta}_{j'}, j' \in \mathcal{M}\}$. Hence the distributions $p(\boldsymbol{\theta}_j|M \neq j)$ become irrelevant in this calculation, and we may choose these *pseudo-priors* in any way we like. A Gibbs sampler is then defined over the product space by the full conditional distributions

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

and

$$P(M = j|\boldsymbol{\theta}, \mathbf{y}) = \frac{f(\mathbf{y}|\boldsymbol{\theta}_j, M = j)[\prod_{j'\in\mathcal{M}} p(\boldsymbol{\theta}_{j'}|M = j)]\pi_j}{\sum_{k\in\mathcal{M}}\{f(\mathbf{y}|\boldsymbol{\theta}_k, M = k)[\prod_{j'\in\mathcal{M}} p(\boldsymbol{\theta}_{j'}|M = k)]\pi_k\}} \ , \tag{2}$$

where of course M is discrete taking finitely many possible values.

Under the usual regularity conditions, this Gibbs sampling algorithm will produce samples from the correct joint posterior distribution. Provided that the sampling chain for the model indicator mixes sufficiently well, the posterior probability of model j can be estimated by

$$\hat{P}(M = j|\mathbf{y}) = \frac{1}{G}\sum_{g=1}^{G} I(M^{(g)} = j), \tag{3}$$

which can in turn be used to estimate a Bayes factor as

$$\hat{B}_{jj'} = \frac{\hat{P}(M = j|\mathbf{y})/\hat{P}(M = j'|\mathbf{y})}{P(M = j)/P(M = j')} .$$

Note that we are free to choose the prior model probabilities $\pi_j = P(M = j)$ arbitrarily, since the Bayes factor $\hat{B}_{jj'}$ will "divide out" their effect anyway. As such, we will typically choose the π_j so that the algorithm visits each model in roughly equal proportion. This will in turn allow more precise estimation of the posterior odds ratio, hence the Bayes factor. Of course, since the true $P(M = j|\mathbf{y})$ are unknown, some experimentation with preliminary runs will typically be required to select computationally efficient π_j values.

In addition, the performance of this method is optimized when the pseudo-priors match the corresponding model-specific posteriors as nearly as possible (Carlin and Chib, 1995; Godsill, 2001). Note that an operational drawback of this method is that, the observations of Green and O'Hagan (1998) notwithstanding, draws must be made from each pseudo-prior at every iteration in order for equation (3) to produce acceptably accurate results. This limits the method's practicality if the cardinality of $\mathcal{M}$ is at all large.

2.2 "Metropolized" product space search

Dellaportas, Forster, and Ntzoufras (2001) propose a hybrid Gibbs-Metropolis strategy. In their strategy, the model selection step is based on a proposal for a move from model j to j', followed by acceptance or rejection of this proposal. That is, the method is a "Metropolized Carlin and Chib" (MCC) approach, which proceeds as follows:

1. Let the current state be $(j, \boldsymbol{\theta}_j)$, where $\boldsymbol{\theta}_j$ is of dimension n_j.

2. Propose a new model j' with probability $h(j, j')$.

3. Generate $\boldsymbol{\theta}_{j'}$ from a pseudo-prior $p(\boldsymbol{\theta}_{j'}|M \neq j')$ as in Carlin and Chib's method.

4. Accept the proposed move (from j to j') with probability

$$\alpha_{j \to j'} = \min \left\{ 1, \ \frac{f(\mathbf{y}|\boldsymbol{\theta}_{j'}, M = j')p(\boldsymbol{\theta}_{j'}|M = j')p(\boldsymbol{\theta}_j|M = j')\pi_{j'}h(j', j)}{f(\mathbf{y}|\boldsymbol{\theta}_j, M = j)p(\boldsymbol{\theta}_j|M = j)p(\boldsymbol{\theta}_{j'}|M = j)\pi_j h(j, j')} \right\}. \tag{4}$$

Thus by "Metropolizing" the model selection step, the MCC method needs to sample only from the pseudo-prior for the proposed model j'. Here the move is a Gibbs step or a sequence of Gibbs steps when $j' = j$. Posterior model probabilities and Bayes factors can be estimated as before.

2.3 Reversible Jump MCMC

This method, originally due to Green (1995), is another strategy that samples over the model and parameter space, but which avoids the full product space search of the Carlin and Chib (1995) method (and the associated pseudo-prior specification and sampling), at the cost of a less straightforward algorithm operating on the *union* space, $\mathcal{M} \times \bigcup_{j \in \mathcal{M}} \Theta_j$. It generates a Markov chain that can "jump" between models with parameter spaces of different dimensions, while retaining the aperiodicity, irreducibility, and detailed balance conditions necessary for MCMC convergence.

A typical reversible jump (RJ) algorithm proceeds as follows.

1. Let the current state of the Markov chain be $(j, \boldsymbol{\theta}_j)$, where $\boldsymbol{\theta}_j$ is of dimension n_j.

2. Propose a new model j' with probability $h(j, j')$.

3. Generate $\mathbf{u}$ from a proposal density $q(\mathbf{u}|\boldsymbol{\theta}_j, j, j')$.

4. Set $(\boldsymbol{\theta}'_{j'}, \mathbf{u}') = \mathbf{g}_{j,j'}(\boldsymbol{\theta}_j, \mathbf{u})$, where $\mathbf{g}_{j,j'}$ is a deterministic function that is 1-1 and onto. This is a "dimension-matching" function, specified so that $n_j + dim(\mathbf{u}) = n'_j + dim(\mathbf{u}')$.

5. Accept the proposed move (from j to j') with probability

$$\alpha_{j \to j'} = \min \left\{ 1, \ \frac{f(\mathbf{y}|\boldsymbol{\theta}'_{j'}, M = j') p(\boldsymbol{\theta}'_{j'}|M = j') \pi_{j'} h(j', j) q(\mathbf{u}'|\boldsymbol{\theta}_{j'}, j', j)}{f(\mathbf{y}|\boldsymbol{\theta}_j, M = j) p(\boldsymbol{\theta}_j|M = j) \pi_j h(j, j') q(\mathbf{u}|\boldsymbol{\theta}_j, j, j')} \left| \frac{\partial \mathbf{g}(\boldsymbol{\theta}_j, \mathbf{u})}{\partial (\boldsymbol{\theta}_j, \mathbf{u})} \right| \right\}. \quad (5)$$

When $j' = j$, the move can be either a standard Metropolis-Hastings or Gibbs step. Posterior model probabilities and Bayes factors may be estimated as before.

Several variations or simplifications of reversible jump MCMC have been proposed for various model classes; see e.g. Richardson and Green (1997) in the context of mixture modeling, and Knorr-Held and Rasser (2000) for a spatial disease mapping application. The "jump diffusion" approach of Phillips and Smith (1996) can also be thought of as a variant on the reversible jump idea. Most recently, Stephens (2000) describes a "birth-death" alternative to reversible jump methods applicable in any context where the parameters of interest may be viewed as a point process having an explicit likelihood; this approach appears applicable in a wide variety of model choice setups.

2.4 Using Partial Analytic Structure

Godsill (2001) proposes use of a "composite model space," which is essentially the setting of Carlin and Chib (1995) except that parameters are allowed to be "shared" between different

models. While this parameter-sharing idea is somewhat controversial, a standard Gibbs sampler applied to this composite model produces the CC method, while a more sophisticated Metropolis-Hastings approach produces a version of the reversible jump algorithm that avoids the "dimension matching" step present in its original formulation (see step 4 in Subsection 2.3 above). This step is often helpful for challenging problems (such as when moving to a higher-dimensional model containing parameters whose values would not plausibly equal any of those in the current model; see e.g. the changepoint example in Section 4 of Green, 1995), but may be unnecessary for simpler problems.

Along these lines, Godsill (2001) outlines a reversible jump method that takes advantage of *partial analytic structure* (PAS) in the Bayesian model. This procedure is applicable when there exists a subvector $(\boldsymbol{\theta}_{j'})_{\mathcal{U}}$ of the parameter vector $\boldsymbol{\theta}_{j'}$ for model j' such that $p((\boldsymbol{\theta}_{j'})_{\mathcal{U}}|(\boldsymbol{\theta}_{j'})_{-\mathcal{U}}, M = j', \mathbf{y})$ is available in closed form, and in the current model j, there exists an equivalent subvector $(\boldsymbol{\theta}_j)_{-\mathcal{U}}$ (the elements of $\boldsymbol{\theta}_j$ *not* in subvector $\mathcal{U}$) of the same dimension as $(\boldsymbol{\theta}_{j'})_{-\mathcal{U}}$. Operationally:

1. Let the current state be $(j, \boldsymbol{\theta}_j)$, where $\boldsymbol{\theta}_j$ is of dimension n_j.

2. Propose a new model j' with probability $h(j, j')$.

3. Set $(\boldsymbol{\theta}_{j'})_{-\mathcal{U}} = (\boldsymbol{\theta}_j)_{-\mathcal{U}}$.

4. Accept the proposed move with probability

$$\alpha_{j \to j'} = \min \left\{ 1, \ \frac{p(j'|(\boldsymbol{\theta}_{j'})_{-\mathcal{U}}, \mathbf{y}) h(j', j)}{p(j|(\boldsymbol{\theta}_j)_{-\mathcal{U}}, \mathbf{y}) h(j, j')} \right\}, \tag{6}$$

   where $p(j|(\boldsymbol{\theta}_j)_{-\mathcal{U}}, \mathbf{y}) = \int p(j, (\boldsymbol{\theta}_j)_{\mathcal{U}}|(\boldsymbol{\theta}_j)_{-\mathcal{U}}, \mathbf{y}) \, d(\boldsymbol{\theta}_k)_{\mathcal{U}}$.

5. If the model move is accepted, update the parameters of the new model $(\boldsymbol{\theta}_{j'})_{\mathcal{U}}$ and $(\boldsymbol{\theta}_{j'})_{-\mathcal{U}}$ using standard Gibbs or Metropolis-Hastings steps; otherwise, update the parameters of the

old model $(\boldsymbol{\theta}_j)_{\mathcal{U}}$ and $(\boldsymbol{\theta}_j)_{-\mathcal{U}}$ using standard Gibbs or Metropolis-Hastings steps.

Note that model move proposals of the form $j \rightarrow j$ always have acceptance probability 1, and therefore when the current model is proposed, this algorithm simplifies to standard Gibbs or Metropolis-Hastings steps. Note that multiple proposal densities may be needed for $(\boldsymbol{\theta}_j)_{\mathcal{U}}$ across models since, while this parameter is common to all of them, its interpretation and posterior support may differ. Troughton and Godsill (1999) discuss and expand on this idea, showing how the update step for $(\boldsymbol{\theta}_j)_{\mathcal{U}}$ may be skipped when a proposed model move is rejected.

2.5 Marginal Likelihood Estimation

Chib (1995) provides a method for computing the marginal density of the data (i.e., the marginal likelihood) $p(\mathbf{y}|M = j)$ when full conditionals for the parameters (or more feasibly, blocks of parameters) are available in closed form. As alluded to above, this method differs from those in the previous three subsections in that it does not include the model indicator M in the sampling order, but instead belongs to the class of methods that seek a marginal density estimate $\hat{p}(\mathbf{y}|M = j)$ directly for each model. This algorithm thus works with the models one at a time, and so we drop the model indicator for now to simplify the notation.

For a model where $\mathbf{y}$ denotes the observed data and $\boldsymbol{\theta}$ denotes the unknown parameter vector, Chib (1995) notes that $\forall\, \boldsymbol{\theta} \in \Theta$, $p(\mathbf{y}) = f(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})/p(\boldsymbol{\theta}|\mathbf{y})$ by Bayes Rule. Therefore an estimate of $\log p(\mathbf{y})$ is

$$\log \hat{p}(\mathbf{y}) = \log f(\mathbf{y}|\tilde{\boldsymbol{\theta}}) + \log p(\tilde{\boldsymbol{\theta}}) - \log \hat{p}(\tilde{\boldsymbol{\theta}}|\mathbf{y}) \,, \tag{7}$$

where $\hat{p}(\boldsymbol{\theta}|\mathbf{y})$ is an estimate of the posterior distribution, and we choose $\tilde{\boldsymbol{\theta}} \in \Theta$ to be a point of high posterior density so as to maximize the accuracy of this approximation.

Chib deals with the case where $\boldsymbol{\theta}$ can be partitioned into several blocks such that the full

conditional for each block is available in closed form. For simplicity, we illustrate in the case of two blocks, i.e., $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ where $p(\boldsymbol{\theta}_1|\mathbf{y},\ \boldsymbol{\theta}_2)$ and $p(\boldsymbol{\theta}_2|\mathbf{y},\ \boldsymbol{\theta}_1)$ are available in closed form. Note that

$$p(\tilde{\boldsymbol{\theta}}|\mathbf{y}) = p(\tilde{\boldsymbol{\theta}}_1, \tilde{\boldsymbol{\theta}}_2|\mathbf{y}) = p(\tilde{\boldsymbol{\theta}}_2|\tilde{\boldsymbol{\theta}}_1, \mathbf{y})p(\tilde{\boldsymbol{\theta}}_1|\mathbf{y}) \tag{8}$$

where $p(\tilde{\boldsymbol{\theta}}_1|\mathbf{y})$ can be estimated by

$$\hat{p}(\tilde{\boldsymbol{\theta}}_1|\mathbf{y}) = \frac{1}{G}\sum_{g=1}^{G} p(\tilde{\boldsymbol{\theta}}_1|\boldsymbol{\theta}_2^{(g)}, \mathbf{y}) \tag{9}$$

with g indexing (post-convergence) iterations. Thus, we have

$$\log \hat{p}(\mathbf{y}) = \log f(\mathbf{y}|\tilde{\boldsymbol{\theta}}_1, \tilde{\boldsymbol{\theta}}_2) + \log p(\tilde{\boldsymbol{\theta}}_1, \tilde{\boldsymbol{\theta}}_2) - \log p(\tilde{\boldsymbol{\theta}}_2|\tilde{\boldsymbol{\theta}}_1, \mathbf{y}) - \log \hat{p}(\tilde{\boldsymbol{\theta}}_1|\mathbf{y}) \tag{10}$$

with the first three terms on the right hand side being available in closed form. Exponentiating and then repeating for each model will give all the quantities needed for calculation of Bayes factors.

The extension from two to B parameter blocks replaces equation (8) with a factoring of the joint posterior into B components. This in turn requires a total of $(B-1)$ Gibbs sampling runs of G samples each to estimate the various factors. For instance, in the case of $B = 3$ parameter blocks, the (now reduced, not full) conditional term $\hat{p}(\tilde{\boldsymbol{\theta}}_2|\tilde{\boldsymbol{\theta}}_1, \mathbf{y})$ could be estimated in a manner similar to that in equation (9), but where the samples are drawn from a reduced Gibbs algorithm which uses only two full conditional distributions, namely $p(\boldsymbol{\theta}_2|\mathbf{y}, \tilde{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_3)$ and $p(\boldsymbol{\theta}_3|\mathbf{y}, \tilde{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_2)$. Clever partitioning of the parameter vector into only a few blocks (each still having a closed form full conditional) is thus desirable for increasing computational accuracy and reducing programming and sampling time.

Of course, equation (9) requires us to know the normalizing constant for the full conditional

distribution $p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, \mathbf{y})$, thus precluding its use with full conditionals updated using Metropolis-Hastings (rather than Gibbs) steps. To remedy this, Chib and Jeliazkov (2001) extend the approach, which takes a particularly simple form in the case where the parameter vector $\boldsymbol{\theta}$ can be updated in a single block. Let $\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y}) = \min\{1, [p(\boldsymbol{\theta}'|\mathbf{y})q(\boldsymbol{\theta}', \boldsymbol{\theta}|\mathbf{y})]/[p(\boldsymbol{\theta}|\mathbf{y})q(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y})]\}$, the probability of accepting a M-H proposal $\boldsymbol{\theta}'$ generated from a candidate density $q(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y})$ (note that this density is allowed to depend on the data $\mathbf{y}$). Chib and Jeliazkov (2001) then show

$$p(\tilde{\boldsymbol{\theta}}|\mathbf{y}) = \frac{E_1\left\{\alpha(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}|\mathbf{y})q(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}|\mathbf{y})\right\}}{E_2\left\{\alpha(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}|\mathbf{y})\right\}}, \tag{11}$$

where E_1 is the expectation with respect to the posterior $p(\boldsymbol{\theta}|\mathbf{y})$ and E_2 is the expectation with respect to the candidate density $q(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}|\mathbf{y})$. The numerator is then estimated by averaging the product in braces with respect to draws from the posterior, while the denominator is estimated by averaging the acceptance probability with respect to draws from $q(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}|\mathbf{y})$, given the fixed value $\tilde{\boldsymbol{\theta}}$. Note that this calculation does not require knowledge of the normalizing constant for $p(\boldsymbol{\theta}|\mathbf{y})$. Plugging this estimate of (11) into (7) completes the estimation of the marginal likelihood. When there are two or more blocks, Chib and Jeliazkov (2001) illustrate an extended version of this algorithm using multiple MCMC runs, similar to the Chib (1995) approach for the Gibbs sampler outlined above.

3 Numerical Illustration: Non-nested Linear Regression Models

3.1 Data and Models

The data in Table 1 are taken from Williams (1959), and give the maximum compressive strength parallel to the grain y_i, the density x_i, and the resin-adjusted density z_i for $n = 42$ specimens of

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

Table 1: Radiata pine compressive strength data

radiata pine. Carlin and Chib (1995) use these data to compare the two linear regression models

$$M = 1: \quad y_i = \alpha + \beta(x_i - \bar{x}) + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2), \quad i = 1, \cdots, n;$$

$$M = 2: \quad y_i = \gamma + \delta(z_i - \bar{z}) + \eta_i, \quad \eta_i \overset{iid}{\sim} N(0, \tau^2), \quad i = 1, \cdots, n.$$

That is, $\mathcal{M}=\{1,\ 2\}$, $\boldsymbol{\theta}_1 = (\alpha, \beta, \sigma^2)^\top$, and $\boldsymbol{\theta}_2 = (\gamma, \delta, \tau^2)^\top$. Notice we could have set $\alpha = \gamma$ and $\sigma^2 = \tau^2$, (i.e., allow the models to share a common intercept and error variance), but we maintain a full composite model space for the time being.

Following the analysis of Carlin and Chib (1995) we adopt $N\left((3000, 185)^\top,\ Diag(10^6, 10^4)\right)$ priors on $(\alpha, \beta)^\top$ and $(\gamma, \delta)^\top$, and $IG\left(3\ ,\ (2 \cdot 300^2)^{-1}\right)$ priors on σ^2 and τ^2, where $IG(a, b)$ represents the inverse gamma distribution with density function $f(v) = \exp(-\frac{1}{bv})/[\Gamma(a) b^a v^{a+1}],\ v > 0,\ a, b >$

0 . That is, σ^2 and τ^2 both have prior mean and standard deviation equal to 300^2. These priors are roughly centered on the corresponding least squares solutions, but are rather vague. We assume prior independence among all the parameters given the corresponding model indicators.

To fix the pseudo-priors, we suppose that $(\alpha, \beta)^\top | M = 2 \sim N\left((3000, 185)^\top,\ Diag(52^2, 12^2)\right)$, $(\gamma, \delta)^\top | M = 1 \sim N\left((3000, 185)^\top,\ Diag(43^2, 9^2)\right)$, $\sigma^2 | M = 2 \sim IG\left(3\ ,\ (2 \cdot 300^2)^{-1}\right)$, and $\tau^2 | M = 1 \sim IG\left(3\ ,\ (2 \cdot 300^2)^{-1}\right)$. Independence among the components of the pseudo-priors is also assumed; that is, α, β, and σ^2 are assumed independent given $M = 2$, and similarly for γ, δ, and τ^2 given $M = 1$. The pseudo-priors for α, β, γ, and δ roughly equal the corresponding posterior distributions, which were obtained through preliminary runs of the individual models; for the variance components, the pseudo-priors simply equal the priors.

The full conditional distributions of the model specific parameters are also bivariate normal and inverse gamma, respectively, and that of M is routinely available using equation (2). We choose π_1 and π_2 to be 0.9995 and 0.0005, respectively, to ensure that the two models are visited in roughly equal proportion by the Markov chain.

For the MCC method, the pseudo-priors that serve as proposal densities are the same as those given above. We adopt the simple model-switching probabilities $h(1,1) = h(1,2) = h(2,1) = h(2,2) = 0.5$, and again use $\pi_1 = 0.9995$ and $\pi_2 = 0.0005$. When the current model is $M = 1$ and the proposed model is $M = 2$, the acceptance probability is

$$\alpha_{1\to 2} = \min\left\{1,\ \frac{f(\mathbf{y}|\gamma, \delta, \tau^2, M = 2)p(\gamma, \delta, \tau^2 | M = 2)p(\alpha, \beta, \sigma^2 | M = 2)\pi_2}{f(\mathbf{y}|\alpha, \beta, \sigma^2, M = 1)p(\alpha, \beta, \sigma^2 | M = 1)p(\gamma, \delta, \tau^2 | M = 1)\pi_1}\right\}\ ;$$

when the current model is $M = 2$ and the proposed model is $M = 1$, the acceptance probability is

$$\alpha_{2\to 1} = \min\left\{1,\ \frac{f(\mathbf{y}|\alpha, \beta, \sigma^2, M = 1)p(\alpha, \beta, \sigma^2 | M = 1)p(\gamma, \delta, \tau^2 | M = 1)\pi_1}{f(\mathbf{y}|\gamma, \delta, \tau^2, M = 2)p(\gamma, \delta, \tau^2 | M = 2)p(\alpha, \beta, \sigma^2 | M = 2)\pi_2}\right\}\ .$$

For the reversible jump method, we use the same priors, but no pseudo-priors are involved. In addition, we use log transforms of the error variances, i.e., $\lambda = \log \sigma^2$, $\omega = \log \tau^2$, to simplify the choice of proposal density. Again, the probabilities of proposing new models are $h(1,1) = h(1,2) = h(2,1) = h(2,2) = 0.5$. Note that the dimension matching requirement is automatically satisfied without generating an additional random vector; moreover, the Jacobian term in the acceptance probability (5) is equal to 1.

Since the two regression models are similar in interpretation, when a move between models is proposed, we simply set $(\alpha, \beta, \lambda)^\top = (\gamma, \delta, \omega)^\top$ or $(\gamma, \delta, \omega)^\top = (\alpha, \beta, \lambda)^\top$. The acceptance probabilities are then given by $\alpha_{1\to 2} = \min\left\{1,\ \frac{f(\mathbf{y}|\gamma,\delta,\omega,M=2)\pi_2}{f(\mathbf{y}|\alpha,\beta,\lambda,M=1)\pi_1}\right\}$, and $\alpha_{2\to 1} = \min\left\{1,\ \frac{f(\mathbf{y}|\alpha,\beta,\lambda,M=1)\pi_1}{f(\mathbf{y}|\gamma,\delta,\omega,M=2)\pi_2}\right\}$.

When the proposed model is the same as the current model, we update using the standard Metropolis step by choosing the proposal density to be trivariate normal distribution centered at the current value. Thus for model 1 we draw $(\alpha^*, \beta^*, \lambda^*)^\top \sim N\left((\alpha^{(k)}, \beta^{(k)}, \lambda^{(k)})^\top\ ,\ Diag(5000, 250, 1)\right)$ and set $(\alpha^{(k+1)}, \beta^{(k+1)}, \lambda^{(k+1)})^\top = (\alpha^*, \beta^*, \lambda^*)^\top$ with probability $r = \min\left\{1, \frac{p(\alpha^*,\beta^*,\lambda^*)}{p(\alpha^{(k)},\beta^{(k)},\lambda^{(k)})}\right\}$. Similar results hold for model 2. Alternatively, we may perform the "within model" updates using a standard Gibbs step, thus avoiding the log transform and the trivariate normal proposal density, and instead simply using the corresponding full conditional distributions.

We may also use the reversible jump method on a somewhat reduced model, i.e., we analytically integrate the slopes and intercepts (α, β, γ, and δ) out of the model and use proposal densities for σ^2 and τ^2 that are identical to the corresponding priors. The acceptance probabilities are $\alpha_{1\to 2} = \min\left\{1,\ \frac{\pi_2}{\pi_1}\frac{f(\mathbf{y}|\tau^2,M=2)}{f(\mathbf{y}|\sigma^2,M=1)}\right\}$, and $\alpha_{2\to 1} = \min\left\{1,\ \frac{\pi_1}{\pi_2}\frac{f(\mathbf{y}|\sigma^2,M=1)}{f(\mathbf{y}|\tau^2,M=2)}\right\}$.

To implement the partial analytic structure algorithm, we treat σ^2 and τ^2 as a single parameter (as required to set $(\boldsymbol{\theta}_j)_{-\mathcal{U}} = (\boldsymbol{\theta}_{j'})_{-\mathcal{U}}$ in step 3 of the algorithm) and denote it by σ^2; we also use $h(1,1) = h(1,2) = h(2,1) = h(2,2) = 0.5$. The acceptance probabilities simplify to $\alpha_{1\to 2} = \min\left\{1,\ \frac{P(M=2|\sigma^2,\mathbf{y})}{P(M=1|\sigma^2,\mathbf{y})}\right\}$, and $\alpha_{2\to 1} = \min\left\{1,\ \frac{P(M=1|\sigma^2,\mathbf{y})}{P(M=2|\sigma^2,\mathbf{y})}\right\}$.

Finally for the Chib (1995) marginal likelihood method, we also run a two-block Gibbs sampler that treats σ^2 or τ^2 as θ_1, and $(\alpha, \beta)^\top$ or $(\gamma, \delta)^\top$ as $\boldsymbol{\theta}_2$. For each of our methods, five independent chains are sampled using different starting values. Each chain is run for 60,000 iterations, of which the first 10,000 are treated as pre-convergence "burn-in" and discarded. In this and all subsequent examples, we checked the adequacy of the burn-in period two ways: graphically, by plotting the sample traces from the five chains versus iteration on a single set of axes; and numerically, by computing the value of the lag 1 sample autocorrelation for a representative subset of the parameters.

3.2 Results

Using traditional (non-Monte Carlo) numerical integration, Green and O'Hagan (1998) find that the "exact" posterior probability for model 1 under the present specification is 0.29135, which corresponds to a 0.70865 posterior probability for model 2 and a Bayes factor of about 4862 in favor of model 2. Table 2 summarizes our results, reporting the estimated posterior probability for model 2, a batched standard deviation estimate for this probability (using 2500 batches of 100 consecutive iterations), an approximate 95% confidence interval (CI) for the posterior probability for model 2, the Bayes factor in favor of model 2, $\hat{B}_{21}$, the lag 1 sample autocorrelation for the model indicator, $\hat{\rho}(1)$, the percentage of moves between the 2 models, $\hat{P}(M^{(g)} \neq M^{(g-1)}|\mathbf{y})$, and the execution time for the `FORTRAN` program in seconds. Model 2 is clearly preferred, as indicated by the huge Bayes factor. In addition, Table 2 indicates that the original CC method has resulted in a smaller batched standard deviation and a lower sample lag 1 autocorrelation than its Metropolized version; still, it is significantly slower, presumably due to the required generation from both pseudo-priors at every iteration. The CC and MCC methods also have smaller batched standard deviations than those obtained using reversible jump, whether Metropolis or Gibbs updates are used in the latter. One

method	$\hat{P}(M=2\|\mathbf{y})$	SD	95% CI for $P(M=2\|\mathbf{y})$	$\hat{B}_{21}$	$\hat{\rho}(1)$	% move	time
CC	.70806	.001721	(.70469, .71144)	4848.4	.567	.179	22.8"
MCC	.71195	.002061	(.70791, .71599)	4940.7	.673	.134	12.2"
RJ-M	.70861	.004058	(.70066, .71657)	4861.3	.589	.170	18.7"
RJ-G	.70906	.002394	(.70437, .71376)	4871.9	.593	.168	7.9"
RJ-R	.70750	.002004	(.70357, .71142)	4835.1	.660	.141	6.7"
PAS	.71035	.001800	(.70682, .71388)	4902.4	.591	.168	7.8"
Chib-1				4860.7			13.6"
Chib-2				4860.3			14.0"
target	.70865			4862			

Table 2: Comparison of different methods, for the simple linear regression example. Here, CC=Carlin and Chib's method; MCC=Metropolized Carlin and Chib method; RJ-M=reversible jump using Metropolis steps if the current model is proposed; RJ-G=reversible jump using Gibbs steps if the current model is proposed; PAS=Godsill's partial analytic structure method, where σ^2 and τ^2 are treated as the same parameter; RJ-R=reversible jump on the reduced model (i.e., with the regression coefficients integrated out); Chib-1=Chib's method evaluated at posterior means; and Chib-2=Chib's method evaluated at frequentist LS solutions.

apparent drawback of the MCC method is that it has the highest lag 1 sample autocorrelation and the lowest move percentage among the six methods for which these two indices are available.

The reversible jump algorithm operating on the reduced model (with the regression coefficients integrated out) appears to be slightly more accurate and faster than the corresponding algorithms operating on the full model. Of course, some extra effort is required to do the integration before programming, and posterior samples are obviously not produced for any parameters no longer appearing in the sampling order. Still, this latter criticism appears minor since our primary interest here lies in posterior model probabilities and the Bayes factor, rather than model-specific estimates.

The Gibbs reversible jump (RJ-G) method appears to offer the best combination of acceptable accuracy, runtime, and pre-programming analytic burden. The reduced reversible jump (RJ-R) and PAS methods are slightly better, but require more pre-run analytic work. We caution that none of these algorithms are any better than their proposal densities; in particular, many of these methods could likely be improved using model-specific proposal densities for σ^2 and τ^2 (or the common σ^2

in the PAS method), rather than the (common) prior distributions for each. We also note that rerunning the PAS method using a single long chain of 1,000,000,000 iterations (after a 10,000-iteration burn-in period) reproduces the Green and O'Hagan (1998) target model probability to four decimal places. More specifically, we obtain an estimated model 2 probability of 0.70861 and a batched (10,000,000 batches of 100 each) standard deviation of 0.0000285, resulting in a 95% confidence interval of $(0.70856, 0.70867)$, covering the true value 0.70865. This offers validation for the PAS approach of using a common σ^2 (instead of separate σ^2 and τ^2 for the two models).

Calculating the sample standard error for Chib's marginal likelihood method is a bit more complicated (the author's suggested approach involves a spectral density estimate and the delta method), so for now we merely observe that its point estimates are closer to the true values than those of five of the other six methods. Note that using the frequentist least squares solutions as $\tilde{\boldsymbol{\theta}}$ ("Chib-2" in the table) greatly simplifies the programming and at no apparent detriment to the method's accuracy. Runtimes are also competitive with the reversible jump and MCC methods.

As indicated by Carlin and Chib (1995), good choices of the pseudo-priors are needed to implement the CC method and its Metropolized version. This usually requires preliminary runs of each of the competing models, thus adding overhead to programming and computing. Hence, it is interesting to investigate the performance of these two methods when these distributions are chosen suboptimally. To do this, we first specify pseudo-priors that are overdispersed relative to their true posteriors by taking a $N\left((3000, 185)^\top,\ Diag(95^2, 21^2)\right)$ distribution for both $(\alpha, \beta)^\top | M = 2$ and $(\gamma, \delta)^\top | M = 1$. We also investigate an underdispersed pseudo-prior choice for $(\alpha, \beta)^\top | M = 2$ and $(\gamma, \delta)^\top | M = 1$, namely a $N\left((3000, 185)^\top,\ Diag(24^2, 5^2)\right)$.

Tables 3 and 4 summarize the results for the two suboptimal pseudo-priors. In these two cases, we can see that Carlin and Chib's method still results in a lower lag 1 sample autocorrelation and a smaller batched standard deviation. All four of the confidence intervals in Tables 3 and 4 cover the

method	$\hat{P}(M=2\|\mathbf{y})$	SD	95% CI for $P(M=2\|\mathbf{y})$	$\hat{B}_{21}$	$\hat{\rho}(1)$	% move	time
CC	.71012	.002416	(.70538, .71485)	4896.8	.765	.097	22.8"
MCC	.70962	.002996	(.70375, .71549)	4885.1	.834	.068	12.0"
target	.70865			4862			

Table 3: Comparison of CC and MCC methods using overdispersed pseudo-priors

method	$\hat{P}(M=2\|\mathbf{y})$	SD	95% CI for $P(M=2\|\mathbf{y})$	$\hat{B}_{21}$	$\hat{\rho}(1)$	% move	time
CC	.70902	.002410	(.70429, .71375)	4870.9	.759	.099	22.9"
MCC	.70761	.003324	(.70109, .71412)	4837.7	.831	.070	12.1"
target	.70865			4862			

Table 4: Comparison of CC and MCC methods using underdispersed pseudo-priors

corresponding true value, but precision (as measured both by the SDs and the width of the interval estimates) is noticeably lower than in Table 2. Overall, estimation of the Bayes factor in favor of model 2 is degraded for both methods, as are lag 1 sample autocorrelations and move probabilities. Thus, MCC appears to exhibit no more robustness to suboptimally specified pseudo-priors than does the CC method, though it does retain its nearly two-to-one advantage over CC in terms of runtimes.

4 Numerical Illustration: Hierarchical Longitudinal Models

4.1 Data and Models

In an AIDS clinical trial originally described by Abrams et al. (1994), 467 patients were randomized to two treatment groups, didanosine (ddI) and zalcitabine (ddC). CD4 lymphocyte counts were measured for each available subject at study entry and at the 2, 6, 12, and 18 month follow-up visits, but many observations are missing due either to death or loss to followup.

Due to the right skew in the CD4 measurements (which are positive by definition), we transform to the square root scale, letting Y_{ij} be the square root CD4 count for the the i^{th} subject at the j^{th}

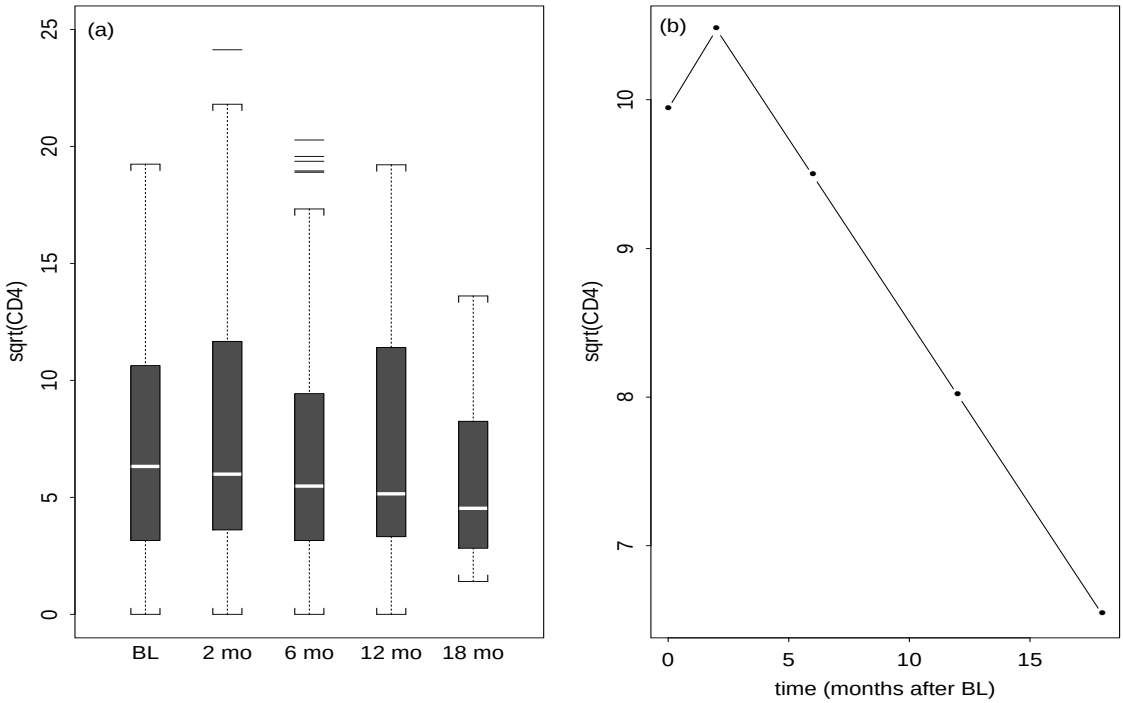

Figure 1: Raw data and fitted changepoint model, ddI/ddC clinical trial: (a) boxplots of square root CD4 count over time, ddI group; (b) fitted changepoint model for AIDS-negative patients, ddI group.

monitoring point, for $j = 1, \ldots, s_i$ and $i = 1, \ldots, n$. (See Figure 1(a) for boxplots of the raw data in the ddI group.) Let $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{is_i})^\top$ be the response vector for the i^{th} subject. We then seek to compare the two mixed effects models:

$$M = 1: \quad \mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\alpha} + \mathbf{W}_i\boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\beta}_i \overset{iid}{\sim} N(\mathbf{0}, \mathbf{V}), \quad \boldsymbol{\epsilon}_i \overset{ind}{\sim} N(0, \sigma^2\mathbf{I}_{s_i}), \text{ and}$$

$$M = 2: \quad \mathbf{Y}_i = \mathbf{P}_i\boldsymbol{\gamma} + \mathbf{Q}_i\boldsymbol{\delta}_i + \boldsymbol{\eta}_i, \quad \boldsymbol{\delta}_i \overset{iid}{\sim} N(\mathbf{0}, \mathbf{U}), \quad \boldsymbol{\eta}_i \overset{ind}{\sim} N(0, \tau^2\mathbf{I}_{s_i}),$$

where the $s_i \times 3$ design matrix $\mathbf{W}_i$ has j^{th} row $(1, t_{ij}, \max\{0, t_{ij} - 2\})$ and the $s_i \times 2$ design matrix $\mathbf{Q}_i$ has j^{th} row $(1, t_{ij})$ with $t_{ij} \in \{0, 2, 6, 12, 18\}$. In addition, $\mathbf{X}_i = (\mathbf{W}_i \mid z_{1i}\mathbf{W}_i \mid z_{2i}\mathbf{W}_i)$ and $\mathbf{P}_i = (\mathbf{Q}_i \mid z_{1i}\mathbf{Q}_i \mid z_{2i}\mathbf{Q}_i)$, where z_{1i} is an indicator variable for being in the ddI treatment group, and z_{2i} an indicator variable for having a positive AIDS diagnosis at baseline ($t = 0$). To obtain a

conjugate model specification, we use multivariate normal priors for $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$, inverse gamma priors for σ^2 and τ^2, and Wishart priors for $\mathbf{V}^{-1}$ and $\mathbf{U}^{-1}$. Our notation for the Wishart distribution is such that

$$\mathbf{x}_{k\times k} \sim W(\boldsymbol{\Omega}_{k\times k}\,,\,\nu) \Leftrightarrow f(\mathbf{x}) = \frac{|\mathbf{x}|^{(\nu-k-1)/2}\exp\left[-\frac{1}{2}\mathrm{tr}(\boldsymbol{\Omega}^{-1}\mathbf{x})\right]}{|\boldsymbol{\Omega}|^{\nu/2}2^{\nu k/2}\pi^{k(k-1)/4}\prod_{j=1}^{k}\gamma(\frac{\nu+1-j}{2})},\ \mathbf{x}_{k\times k} \text{ positive definite.}$$

The parameter ν is sometimes referred to as the *degrees of freedom* for the Wishart distribution; larger ν values correspond to a more informative prior.

Note that $\boldsymbol{\beta}_i$ and $\boldsymbol{\delta}_i$ are subject-specific random effects, the inclusion of which reflect the fact that measurements obtained from the same individual are more similar than measurements from different individuals. The formulation of Model 1 allows for a possible changepoint in y_{ij} at month 2; such a "boost" in CD4 count is the desired clinical consequence of both drugs (Goldman et al., 1996). Figure 1(b) shows the fitted changepoint model (Model 1) for AIDS-negative patients in the ddI group, to provide a better idea of the difference between this and the straight-line linear decline model (Model 2). We seek to determine whether a changepoint is in fact necessary to adequately model our data using the CC, RJ, MCC, and marginal likelihood methods.

Following the justification in Carlin and Louis (1996, p.279), we choose priors for the first model as follows: $\boldsymbol{\alpha}|M=1 \sim N(\mathbf{c},\ \mathbf{D})$, $\mathbf{V}^{-1}|M=1 \sim W\left((\rho\mathbf{R})^{-1},\ \rho\right)$, and $\sigma^2|M=1 \sim IG(a,\ b)$, where a=3, b=0.005, $\mathbf{c} = (10,0,0,0,0,0,-3,0,0)^\top$, $\mathbf{D} = Diag(4,1,1,0.01,1,1,1,1,1)$, ρ=24, and $\mathbf{R} = Diag(4,\frac{1}{16},\frac{1}{16})$. Next, since the second model is just a simplification of the first, we set its priors to be the corresponding simplifications as well; namely, $\boldsymbol{\gamma}|M=2 \sim N(\mathbf{c}^*,\ \mathbf{D}^*)$, $\mathbf{U}^{-1}|M=2 \sim W\left((\rho^*\mathbf{R}^*)^{-1},\ \rho^*\right)$, and $\tau^2|M=2 \sim IG(a^*,\ b^*)$, where a^*=3, b^*=0.005, $\mathbf{c}^* = (10,0,0,0,-3,0)^\top$, $\mathbf{D}^* = Diag(4,1,0.01,1,1,1)$, ρ^*=24, and $\mathbf{R}^* = Diag(4,\frac{1}{16})$.

For priors on the model probabilities we use $\pi_1 = \pi_2 = 0.5$, and for probabilities of propos-

ing models we use $h(1,1) = h(1,2) = h(2,1) = h(2,2) = 0.5$. After preliminary model-specific runs (5 chains per model, 2000 iterations per chain), we choose pseudo-priors from the obvious distributional families (multivariate normal for $\boldsymbol{\alpha}$, $\boldsymbol{\gamma}$, $\boldsymbol{\beta}_i$, and $\boldsymbol{\delta}_i$; inverse gamma for σ^2 and τ^2; and Wishart with $\nu = 24$ for $\mathbf{V}^{-1}$ and $\mathbf{U}^{-1}$) with parameters chosen to roughly match the corresponding posterior estimates based on the post-convergence iterations of the preliminary runs.

For the marginal likelihood method we begin with a trick recommended by Chib and Carlin (1999). For Model 1, we write $p(\{\boldsymbol{\beta}_i\}, \boldsymbol{\alpha}|\sigma^2, \mathbf{V}^{-1}, \mathbf{y}) = p(\{\boldsymbol{\beta}_i\}|\boldsymbol{\alpha}, \sigma^2, \mathbf{V}^{-1}, \mathbf{y})p(\boldsymbol{\alpha}|\sigma^2, \mathbf{V}^{-1}, \mathbf{y})$, and recognize that the reduced conditional $p(\boldsymbol{\alpha}|\sigma^2, \mathbf{V}^{-1}, \mathbf{y})$ is available in closed form. That is, we marginalize the likelihood over the random effects, obtaining

$$\mathbf{Y}_i|\boldsymbol{\alpha}, \sigma^2, \mathbf{V}, M = 1 \sim N\left(\mathbf{X}_i\boldsymbol{\alpha},\ \sigma^2\mathbf{I}_{s_i} + \mathbf{W}_i\mathbf{V}\mathbf{W}_i^\top\right) , \tag{12}$$

from which we derive the reduced conditional for $\boldsymbol{\alpha}$. This trick reduces autocorrelation in the chains somewhat, and also allows us to work with the appropriate three-block extension of equation (10),

$$\begin{aligned}\log f(\mathbf{Y}|M = 1) = {} & \log f(\mathbf{Y}|\tilde{\boldsymbol{\alpha}}, \tilde{\sigma}^2, \tilde{\mathbf{V}}^{-1}, M = 1) + \log p(\tilde{\boldsymbol{\alpha}}, \tilde{\sigma}^2, \tilde{\mathbf{V}}^{-1}|M = 1) \\ & - \log p(\tilde{\boldsymbol{\alpha}}|\mathbf{Y}, \tilde{\sigma}^2, \tilde{\mathbf{V}}^{-1}, M = 1) - \log \hat{p}(\tilde{\sigma}^2|\mathbf{Y}, \tilde{\mathbf{V}}^{-1}, M = 1) - \log \hat{p}(\tilde{\mathbf{V}}^{-1}|\mathbf{Y}, M = 1)\end{aligned}$$

for Model 1, and similarly for Model 2.

For the RJ method, we first reduce the dimension even further by using the normality in equation (12) to also integrate out the fixed effects $\boldsymbol{\alpha}$. This leads to rather complicated expressions for the reduced likelihoods $f(\mathbf{y}|\sigma^2, \mathbf{V}^{-1})$ and $f(\mathbf{y}|\tau^2, \mathbf{U}^{-1})$, but reduced total model dimensions of only 7 and 4 for Models 1 and 2, respectively. Our RJ method then used septivariate and quadrivariate normal independence chain proposals for suitably transformed versions of the two reduced parameter vectors, namely $\boldsymbol{\psi}^\top = (\log\sigma^2, \log[(V^{-1})_{11}], (V^{-1})_{21}, (V^{-1})_{31}, \log[(V^{-1})_{22}], (V^{-1})_{32}, \log[(V^{-1})_{33}])$ and

$\boldsymbol{\phi}^\top = (\log\tau^2, \log[(U^{-1})_{11}], (U^{-1})_{21}, \log[(U^{-1})_{22}])$. This is a simplified version of the "tailored independence chain" approach of Chib and Greenberg (1998, p.351); values of the mean and covariance parameters for these two proposal densities were obtained from preliminary Gibbs sampler runs for each model. A "quick reject" step handled those rare occasions on which proposals corresponding to non-positive definite $\mathbf{V}$ or $\mathbf{U}$ were generated. We further used $h(1,2) = h(2,1) = 1$ and $h(1,1) = h(2,2) = 0$; i.e., the RJ algorithm proposed a model switch at every iteration. The acceptance probability when the current model is 1 is then

$$\alpha_{1\to 2} = \min\left\{1,\ \frac{f(\mathbf{y}|\tau^2, \mathbf{U}^{-1}, M=2)\, p(\tau^2, \mathbf{U}^{-1}|M=2)\, q_7(\sigma^2, \mathbf{V}^{-1})\, \pi_2}{f(\mathbf{y}|\sigma^2, \mathbf{V}^{-1}, M=1)\, p(\sigma^2, \mathbf{V}^{-1}|M=1)\, q_4(\tau^2, \mathbf{U}^{-1})\, \pi_1}\right\}, \tag{13}$$

where q_7 and q_4 denote the two proposal densities; the expression for $\alpha_{2\to 1}$ follows similarly. Finally, we chose $\pi_1 = 70/71$ and $\pi_2 = 1/71$ in an effort to better balance the proportion of time spent by the algorithm in each model. We used 3 chains of 60,000 iterations, discarding the first 10,000 from each chain as burn-in.

4.2 Results

Using the Chib (1995) marginal likelihood method, our 5 chains of 1000 post-burn-in iterations each estimated the log marginal likelihood as –3581.795 for Model 1 and –3577.578 for Model 2. These values produce an estimated Bayes factor of $\hat{B}_{21} = \exp[-3577.578 - (-3581.795)] = 67.83$, moderate to strong evidence in favor of Model 2 (the simpler model without a change point). This is consistent with previous findings (Carlin and Louis, 1996, p.283–285) using less formal model choice tools (cross-validation residuals and conditional predictive ordinates) which suggested that the additional complexity offered by the changepoint model was not required for adequate explanation of these data. To get an idea of the variability in $\hat{B}_{21}$, we replicated our calculations

three more times, obtaining Model 1 log marginal likelihoods of $(-3581.891, -3581.828, -3582.169)$, Model 2 log marginal likelihoods of $(-3577.620, -3577.549, -3577.575)$, and hence Bayes factors of $(71.59, 72.19, 98.88)$. This serves as a reminder that small changes among the former can lead to fairly pronounced changes in the latter – though even with our modest MCMC sample sizes, not enough to change our overall conclusion of "moderate to strong evidence in favor of Model 2."

For the CC and MCC methods, we failed to obtain convergence even after running 3 chains of 1,500,000 iterations each: results vary dramatically among different chains and with different random number seeds. The main problem we encountered was the rarity of acceptance of a proposed jump between models, even after significant tinkering with our pseudo-prior densities. It does not seem feasible to specify psuedopriors separately for each of the 467 $\boldsymbol{\beta}_i$'s, yet using a common pseudo-prior for all of them seems to produce poor candidates overall, hence poor switching between the two models. Marginalizing the $\boldsymbol{\beta}_i$ out of the model as in (12) would significantly ease the pseudo-prior specification burden, but also eliminate the closed form for the σ^2 and $\mathbf{V}^{-1}$ full conditionals (though one could imagine "Metropolis-within-Gibbs" updates for those two parameters).

For the RJ method, we were able to obtain convergence using 3 chains of 50,000 post-burn-in iterations each: the algorithm accepted a proposed model jump 60.8% of the time, producing a lag 1 sample autocorrelation for M of –0.21849. The model 2 probability estimate is $\hat{P}(M = 2|y) = 0.51961$ with batched standard deviation 0.00173, for a 95% CI of (0.51622, 0.52300). The corresponding Bayes factor is $\hat{B}_{21} = 75.71$, which is quite close to 76.74, the result obtained by averaging our four Chib (1995) loglikelihood estimates for each model and then computing the Bayes factor. To make an even fairer comparison between these two methods, we also applied the marginal likelihood method to the reduced (seven- and four-dimensional) models used by RJ. Since no closed forms exist for the parameters in this case, the Chib and Jeliazkov (2001) extension was required. Using 5 chains of 1000 post-burn-in iterations each, this in turn produced $\hat{B}_{21} = 76.50$,

also consistent with previous results.

Overall, our experience with RJ and the other joint model-parameter space search methods was a rather frustrating one for this problem. Without the marginalization over both the fixed and random mean effects, none of these algorithms could be implemented satisfactorily. But this marginalization creates problems of its own: it is quite technical, and leads to a very complicated form for the M-H acceptance ratio (13), thus increasing the chance of an algebraic or computational error. Even with the marginalization, such methods remain difficult to tune. For example, when one model is decisively worse than the other the sampler may favor the better model, but be unable to estimate the Bayes factor with any degree of accuracy. In our case we were able to set π_1 and π_2 to adjust for this to some extent – but only because we already had a good estimate of B_{21} from the marginal likelihood method! This need for preliminary information is of course in addition to the preliminary runs required to specify proposal (or pseudo-prior) densities for each model, which are typically needed to successfully implement a joint model-parameter space search method.

By contrast, the marginal likelihood methods were relatively easy to program and tune, and seemed to perform adequately using fewer MCMC samples (20,000 versus 150,000). Estimating standard errors is more difficult, but simply replicating the entire procedure a few times with different random number seeds provided an acceptable idea of the procedure's order of accuracy. These methods also did not require preliminary runs (only a point of high posterior density, $\tilde{\boldsymbol{\theta}}$), and in the case of the Gibbs sampler, only a rearrangement of existing computer code.

5 Numerical Illustration: Binary Data Latent Variable Models

5.1 Data and Models

Chib (1995) and Chib and Jeliazkov (2001) illustrate their techniques using a dataset originally presented by Brown (1980) in which the response variable is the presence ($y_i = 1$) or absence

($y_i = 0$) of prostatic nodal involvement for $n = 53$ prostate cancer patients. Among the potential predictor variables are the result of an X-ray exam, negative ($x_i = 0$) or positive ($x_i = 1$), and the size of the tumor, small ($z_i = 0$) or large ($z_i = 1$). Here we compare two of the probit regression models considered by Chib (1995) and Chib and Jeliazkov (2001), namely

$$M = 1: \quad \Pr(y_i = 1) = \Phi(\alpha + \beta x_i), \quad i = 1, \cdots, n, \text{ and}$$
$$M = 2: \quad \Pr(y_i = 1) = \Phi(\gamma + \delta z_i), \quad i = 1, \cdots, n,$$

where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal random variable. To ensure closed-form full conditional distributions, Chib's (1995) method requires the augmentation of the parameter space with latent variables ω_i, $i = 1, \ldots, n$. Specifically, for Model 1 we have

$$\omega_i \sim N(\alpha + \beta x_i \, , \, 1) \, .$$

Then if $y_i = I(\omega_i > 0)$, it is easily shown (Carlin and Polson, 1992; Albert and Chib, 1993) that this formulation is equivalent to the probit model, and that the full conditional distributions for $(\alpha, \beta)^\top$ and ω_i are normal and truncated normal, respectively. We seek to implement the reversible jump algorithm in this setting, and compare it to these marginal likelihood methods in terms of performance and ease of use.

We use the same prior distributions for the parameters as in Chib (1995) and Chib and Jeliazkov (2001), namely, a $N\left((0.75, 0.75)^\top, Diag(5^2, 5^2)\right)$ for both $(\alpha, \beta)^\top$ and $(\gamma, \delta)^\top$. Our RJ algorithm always proposes a jump and no within-model update occurs; that is, $h(1,2) = h(2,1) = 1$ and $h(1,1) = h(2,2) = 0$. The acceptance probabilities are thus given by $\alpha_{1\to 2} = \min\left\{1 \, , \, \frac{\pi_2 f(\mathbf{y}|\gamma,\delta,M=2)p(\gamma,\delta|M=2)q_1(\alpha,\beta)}{\pi_1 f(\mathbf{y}|\alpha,\beta,M=1)p(\alpha,\beta|M=1)q_2(\gamma,\delta)}\right\}$ and $\alpha_{2\to 1} = \min\left\{1 \, , \, \frac{\pi_1 f(\mathbf{y}|\alpha,\beta,M=1)p(\alpha,\beta|M=1)q_2(\gamma,\delta)}{\pi_2 f(\mathbf{y}|\gamma,\delta,M=2)p(\gamma,\delta|M=2)q_1(\alpha,\beta)}\right\}$.

Convenient prior model probabilities for this problem (determined from a preliminary run, or

perhaps by considering the maximum likelihood for the fitted model in each case) are $\pi_1 = 0.125$ and $\pi_2 = 0.875$. For the proposal densities, we use bivariate normal densities with diagonal covariance matrices, with the mean vectors and variance elements close to the maximum-likelihood estimates and their squared standard errors, respectively. Specifically, a $N\left((-0.7, 1.3)^\top,\ Diag(0.25^2, 0.45^2)\right)$ is used for q_1, while a $N\left((-0.9, 1.0)^\top,\ Diag(0.30^2, 0.40^2)\right)$ is chosen as q_2.

5.2 Results

Chib (1995) reports that the marginal log-likelihood of Model 1 is –35.323 while that of Model 2 is –37.234, indicating a Bayes factor of 6.76 in favor of Model 1; Chib and Jeliazkov (2001) report the two log-likelihoods to be -35.33 and -37.23, corresponding to a Bayes factor of 6.69. Using five independent reversible jump chains of 50,000 iterations each (after deleting the first 10,000 iterations from each chain), we obtained a Bayes factor of 6.67, with corresponding 95% confidence interval $[6.61, 6.73]$ (based on a batched standard deviation estimate using 2500 batches of 100 iterations each). The lag 1 autocorrelation of the model indicator is an insignificant –0.143, and the overall acceptance rate is 57%.

In contrast to the results of the previous section, here we tend to favor the RJ algorithm over the marginal likelihood methods. The former method runs well and quickly in its regular form, producing essentially uncorrelated draws of the model indicator while requiring no "flash of insight" to augment the parameter space with the appropriate latent variables. (The Chib and Jeliazkov extension of course avoids this problem, but at the cost of a more involved algorithm using equation (11).) Also, estimated standard errors for the Bayes factor are straightforwardly obtained under RJ, without resort to the delta method as required by the marginal likelihood methods.

6 Summary, Discussion, and Recommendations

For the non-hierarchical linear regression model, we were able to obtain acceptable results for all of the methods we investigated, though the amount and difficulty of the programming effort and pre-programming analytic work varied rather widely. The CC and MCC methods continued to perform reasonably when their pseudo-priors were under- or over-dispersed. The RJ and PAS methods ran more quickly than the other model space search methods, but the marginal likelihood methods seemed to produce the highest degree of accuracy for roughly comparable runtimes. This is in keeping with the intuition that some gain in precision should accrue to MCMC methods that avoid a model space search.

For our hierarchical longitudinal model, only the RJ and marginal likelihood methods were able to produce estimates of the Bayes factor. The marginal likelihood methods required a fair amount of "bookkeeping," but seemed easier to tune, and required little in the way of additional coding beyond what was already required for the individual models themselves. As in the linear regression example, the RJ method produced a given number of samples in about half the time taken by the marginal likelihood methods, but the latter did not require nearly as large a total sample size to produce an acceptably accurate result.

Our binary data model provides a setting where both the marginal likelihood and RJ methods can demonstrate their respective strengths. With the addition of latent variables, the former approach becomes straightforward, and retains its characteristic implementational and computational simplicity. Without the latent variables, however, the latter approach seems preferable, running quickly and producing direct estimates of the model probabilities and their standard errors.

Overall, then, it appears difficult to make a general recommendation between marginal likelihood and reversible jump methods that will be appropriate in all model settings. We are inclined to

conclude that the marginal likelihood methods appear to offer a better and safer approach to recommend to practitioners seeking to choose amongst a collection of standard (e.g., hierarchical linear) models. They also appear better suited to implementation in standard software packages, as they reuse existing code to a large extent; by contrast, the joint model-parameter space search methods require extra, typically problem-specific coding (say, to specify move types), as well as additional algorithm tuning.

We hasten to add however that the blocking required by the marginal likelihood methods may preclude their use in some settings, such as spatial models using Markov random field priors (which involve large numbers of random effect parameters that cannot be analytically integrated out of the likelihood nor readily updated in blocks). Models of varying dimension (e.g. multiple changepoint models) provide another setting where reversible jump may offer the only feasible alternative for estimating a Bayes factor. The method can of course be difficult to implement (due for example to the required Jacobian calculations, or the algorithm tuning problems we faced in Section 4), but if one sticks to simple moves rather than elaborate constructions (as we have done in Section 5), the effort can remain at a reasonable level. Still, in settings where latent variables cannot be integrated out of the likelihood in closed form (e.g, the multivariate probit model in Chib and Jeliazkov, 2001), RJ may well have the same difficulties we found in Section 4 before doing such integration.

The marginal likelihood methods would also seem impractical if the number of candidate models were very large (e.g., in variable selection problems having 2^p possible models, corresponding to each of p predictors being either included or excluded). However, here we caution that the ability of joint model and parameter space search methods to sample effectively over such large spaces is very much in doubt; see for example Clyde et al. (1996). Moreover it may not be feasible to write code that specifies moves between all pairs of models; actual implementation of RJ may require some ordering of the models (as is done in models having an unknown number of changepoints

or mixture components). Still, more efficient joint search algorithms (such as the PAS method) continue to emerge, and thus we urge ongoing investigation into their applicability in comparing high-dimensional models.

Finally, we remark that *all* of the methods we have discussed require substantial time and effort (both human and computer) for a rather modest payoff, namely a collection of posterior model probability estimates, possibly augmented with associated standard error estimates. Besides being only single number summaries of relative model worth, Bayes factors are not interpretable under improper priors on any components of any of the candidate parameters, and have also been criticized on theoretical grounds (see e.g. Draper, 1995, and the associated discussion). As a result, one might conclude that none of the methods considered herein is appropriate for everyday, "rough and ready" model comparison, and instead search for more computationally realistic alternatives. Recent such suggestions include cross-validatory residual analyses (Gelfand et al., 1992) and various fairly informal conditional predictive schemes (Laud and Ibrahim, 1995; Waller et al., 1997), as well as more formal decision-theoretic methods for minimizing posterior predictive loss (Gelfand and Ghosh, 1998). Also, Spiegelhalter et al. (1998) suggest a generalization of the Akaike information criterion (AIC) that is based on the posterior distribution of the deviance statistic. The reader is referred to Carlin and Louis (2000, Sec. 6.5) for a brief overview of some of these alternative Bayesian model selection methods.

Acknowledgements

The research of both authors was supported by National Institute of Allergy and Infectious Diseases Grant R01-AI41966. The authors thank Profs. Sid Chib, Mike Daniels, Simon Godsill, and Peter Green for helpful discussions and invaluable technical assistance in the preparation of

this manuscript. The authors also thank two anonymous referees for comments and suggestions that significantly improved the manuscript.

References

[1] Abrams, D.I., Goldman, A.I., Launer, C., Korvick, J.A., Neaton, J.D., Crane, L.R., Grodesky, M., Wakefield, S., Muth, K., Kornegay, S., Cohn, D.L., Harris, A., Luskin-Hawk, R., Markowitz, N., Sampson, J. H., Thompson, M., Deyton, L., and the Terry Beirn Community Programs for Clinical Research on AIDS (1994). Comparative trial of didanosine and zalcitabine in patients with human immunodeficiency virus who are intolerant of or have failed zidovudine therapy. *New England Journal of Medicine*, **330**, 657-662.

[2] Albert, J.H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *J. Amer. Statist. Assoc.*, **88**, 669–679.

[3] Berger, J.O. and Pericchi, L.R. (1996). The intrinsic Bayes factor for linear models. In *Bayesian Statistics 5*, J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, eds., Oxford: Oxford University Press, pp. 25–44.

[4] Berger, J.O. and Pericchi, L.R. (2001). Objective Bayesian methods for model selection: introduction and comparison. To appear *J. Statist. Plan. Inf.*

[5] Brown, B.W. (1980). Prediction analysis for binary data. In *Biostatistics Casebook*, (R.G. Miller, Jr., B. Efron, B.W. Brown, Jr. and L.E. Moses, eds.), pp. 3-18, New York: Wiley.

[6] Carlin, B.P., and Chib, S. (1995). Bayesian model choice via Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Ser. B*, **57**, 473-484.

[7] Carlin, B.P., and Louis, T.A. (2000). *Bayes and Empirical Bayes Methods for Data Analysis, 2nd edition*. Boca Raton, FL: Chapman and Hall/CRC Press.

[8] Carlin, B.P. and Polson, N.G. (1992). Monte Carlo Bayesian methods for discrete regression models and categorical time series. In *Bayesian Statistics 4*, J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, eds., Oxford: Oxford University Press, pp. 577–586.

[9] Chen, M.-H., Shao, Q.-M., and Ibrahim, J.G. (2000). *Monte Carlo Methods in Bayesian Computation*. New York: Springer-Verlag.

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

[11] Chib, S., and Carlin, B.P. (1999). On MCMC sampling in hierarchical longitudinal models. *Statistics and Computing*, **9**, 17–26.

[12] Chib, S., and Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika*, **85**, 347–361.

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

[14] Clyde, M., DeSimone, H., and Parmigiani, G. (1996). Prediction via orthogonalized model mixing. *J. Amer. Statist. Assoc.*, **91**, 1197–1208.

[15] Dellaportas, P., Forster, J.J., and Ntzoufras, I. (2001). On Bayesian model and variable selection using MCMC. To appear *Statistics and Computing*.

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

[17] 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*, J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, eds., Oxford: Oxford University Press, pp. 147–167.

[18] Gelfand, A.E. and Ghosh, S.K. (1998). Model choice: a minimum posterior predictive loss approach. *Biometrika*, **85**, 1–11.

[19] Gelfand, A.E. and Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.*, **85**, 398–409.

[20] Godsill, S. J. (2001). On the relationship between MCMC model uncertainty methods. To appear *J. Comput. Graph. Statist.*

[21] Goldman, A.I., Carlin, B.P., Crane, L.R., Launer, C., Korvick, J.A., Deyton, L., and Abrams, D. I. (1996). Response of $CD4^+$ and clinical consequences to treatment using ddI or ddC in patients with advanced HIV infection. *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology*, **11**, 161–169.

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

[23] Green, P.J. and O'Hagan, A. (1998). Carlin and Chib do not need to sample from pseudo-priors. Research Report 98-1, Department of Statistics, University of Nottingham.

[24] Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.

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

[26] Knorr-Held, L., and Rasser, G. (2000). Bayesian detection of clusters and discontinuities in disease maps. *Biometrics*, **56**, 13–21.

[27] Laud, P. and Ibrahim, J. (1995). Predictive model selection. *J. Roy. Statist. Soc., Ser. B*, **57**, 247–262.

[28] Lavine, M. and Schervish, M.J. (1999). Bayes factors: What they are and what they are not. *The American Statistician*, **53**, 119–122.

[29] Madigan, D. and York, J. (1995). Bayesian graphical models for discrete data. *International Statistical Review*, **63**, 215–232

[30] O'Hagan, A. (1995). Fractional Bayes factors for model comparison (with discussion). *J. Roy. Statist. Soc., Ser. B*, **57**, 99–138.

[31] Phillips, D.B. and Smith, A.F.M. (1996). Bayesian model comparison via jump diffusions. In *Markov Chain Monte Carlo in Practice*, eds. W.R. Gilks, S. Richardson, and D.J. Spiegelhalter, London: Chapman and Hall, pp. 215–239.

[32] Raftery, A.E., Madigan, D., and Hoeting, J.A. (1997). Bayesian model averaging for linear regression models. *J. Amer. Statist. Assoc.*, **92**, 179–191

[33] Richardson, S. and Green, P.J. (1997). Bayesian analysis of mixtures with an unknown number of components. *J. Roy. Statist. Soc., Ser. B* **59**, 731–758.

[34] Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.*, **6**, 461–464.

[35] Spiegelhalter, D.J., Best, N., and Carlin, B.P. (1998). Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models. Research Report 98–009, Division of Biostatistics, University of Minnesota. Submitted to *J. Roy. Statist. Soc., Ser. B.*

[36] Stephens, M. (2000). Bayesian analysis of mixture models with an unknown number of components – an alternative to reversible jump methods. *Annals of Statistics*, **28**, 40–74.

[37] Troughton, P.T. and Godsill, S.J. (1999). MCMC methods for restoration of nonlinearly distorted autoregressive signals. *Signal Processing*, **81**, 83–97.

[38] Waller, L.A., Carlin, B.P., Xia, H. and Gelfand, A.E. (1997). Hierarchical spatio-temporal mapping of disease rates. *Journal of the American Statistical Association*, **92**, 607–617.

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