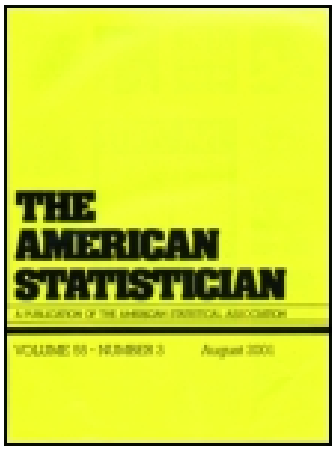

The American Statistician
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/utas20

Bayesian Multimodel Inference by RJMCMC: A Gibbs Sampling Approach

Richard J. Barker [a] & William A. Link [b]
[a] Department of Mathematics and Statistics , University of Otago , P.O. Box 56 Dunedin, New Zealand
[b] USGS Patuxent Wildlife Research Center , Laurel , MD , 20708
Published online: 11 Sep 2013.

To cite this article: Richard J. Barker & William A. Link (2013) Bayesian Multimodel Inference by RJMCMC: A Gibbs Sampling Approach, The American Statistician, 67:3, 150-156, DOI: 10.1080/00031305.2013.791644

To link to this article: http://dx.doi.org/10.1080/00031305.2013.791644

Bayesian Multimodel Inference by RJMCMC: A Gibbs Sampling Approach

Richard J. BARKER and William A. LINK

Bayesian multimodel inference treats a set of candidate models as the sample space of a latent categorical random variable, sampled once; the data at hand are modeled as having been generated according to the sampled model. Model selection and model averaging are based on the posterior probabilities for the model set. Reversible-jump Markov chain Monte Carlo (RJMCMC) extends ordinary MCMC methods to this metamodel. We describe a version of RJMCMC that intuitively represents the process as Gibbs sampling with alternating updates of a categorical variable M (for *Model*) and a "palette" of parameters $\boldsymbol{\psi}$, from which any of the model-specific parameters can be calculated. Our representation makes plain how model-specific Monte Carlo outputs (analytical or numerical) can be post-processed to compute model weights or Bayes factors. We illustrate the procedure with several examples.

KEY WORDS: Bayes factors; Posterior model probabilities; Reversible jump Markov chain Monte Carlo.

1. INTRODUCTION

Model selection is a fundamental statistical issue with naturally Bayesian solutions (Robert and Marin 2008) based on posterior model probabilities, or the Bayes factors from which they can be calculated. Suppose models M_1 and M_2 are to be compared on the basis of data $\boldsymbol{y}$. Each model is described by a family of densities $f_i(\boldsymbol{y};\boldsymbol{\theta}_i)$, with parameters $\boldsymbol{\theta}_i$ having prior distributions $\pi_i(\theta_i)$ over parameter spaces Θ_i $(i = 1, 2)$. The Bayes factor in favor of M_1 is

$$B_{12} = \frac{\int f_1(\boldsymbol{y};\boldsymbol{\theta}_1)\pi_1(\boldsymbol{\theta}_1)d\boldsymbol{\theta}_1}{\int f_2(\boldsymbol{y};\boldsymbol{\theta}_2)\pi_2(\boldsymbol{\theta}_2)d\boldsymbol{\theta}_2}. \quad (1)$$

Posterior and prior odds in favor of M_1 are related by the formula

$$\frac{\Pr(M_1|\boldsymbol{y})}{\Pr(M_2|\boldsymbol{y})} = B_{12} \times \frac{\Pr(M_1)}{\Pr(M_2)}.$$

Richard J. Barker is Professor, Department of Mathematics and Statistics, University of Otago, P.O. Box 56 Dunedin, New Zealand (E-mail: *rbarker@maths.otago.ac.nz*). William A Link, USGS Patuxent Wildlife Research Center, Laurel, MD 20708 (E-mail: *wlink@usgs.gov*).

Two challenges for implementing Bayesian multimodel inference relate to the Bayes factor. The first arises in attempting objective Bayesian analysis, in which posterior inference is dominated by the data rather than the priors. The choice among various minimally informative priors on parameters may be of little consequence for inference about the parameters themselves but nevertheless have a profound effect on the Bayes factor. Indeed, if the model-specific priors $\pi_i(\boldsymbol{\theta})$ are improper, the Bayes factor may not be well defined (Han and Carlin 2001). Second, calculation of the marginal likelihoods, as in the numerator and denominator of Equation (1), can be difficult.

This article focuses on the computational challenges, presenting a simple and intuitive version of reversible jump Markov chain Monte Carlo (RJMCMC) (Green 1995). For cases where the number of models under consideration is moderate, our approach allows models to be fit one at a time using ordinary MCMC methods, then post-processed to compute Bayes factors comparing the models. We also illustrate the problems associated with prior choice for objective Bayesian analysis, suggesting as a solution the use of "nonpreferential priors" (Link and Barker 2008).

1.1 Calculation of Bayes Factors

In some cases (e.g., exponential families and conjugate priors, Kass and Raftery 1995), analytic expressions are available for calculating Bayes factors. For the remaining cases some authors have suggested computing approximations to Bayes factors from model-specific MCMC output (see Bartolucci, Scaccia, and Mira 2006, for a brief review). Proposals have been based on estimation of marginal likelihoods (e.g., Newton and Raftery 1994; Chib 1995) or ratios of marginal likelihoods (e.g., Meng and Wong 1996). A drawback of these approaches is that they can be complicated to implement and require a lot of "bookkeeping" (Bartolucci, Scaccia, and Mira 2006). Also, methods based on the harmonic mean approach by Newton and Raftery (1994) have been shown to be unstable (Chib 1995). Worse, related estimates of marginal likelihood such as those by Congdon (2006) have been shown to be biased (Robert and Marin 2008).

An alternative to estimating Bayes factors using approximate marginal likelihoods is to build an MCMC sampler on an enlarged state space that includes the model as an unknown, described as a categorical variable M with sample space $\mathcal{M}$. Bayes factors can then be computed as ratios of posterior model

odds (relative frequencies in the MCMC sampler) to specified prior odds. Carlin and Chib (1995) proposed an algorithm that searches over the composite space of model $M \in \mathcal{M}$ and model parameters $\boldsymbol{\theta}_j \in \Theta_j$ given by $\mathcal{M} \times \prod_{j\in\mathcal{M}} \Theta_j$, where the products are Cartesian products. Godsill (2001) gave a general composite space representation of the multimodel inference problem that leads to the algorithm by Carlin and Chib (1995) as a special case.

A disadvantage of the Carlin and Chib (1995) approach is that the user needs to specify a potentially large number of "pseudo-priors" that must be drawn from at every iteration of the Markov chain. Also, their method does not exploit any interrelationships among parameters. The RJMCMC method by Green (1995) searches over $\mathcal{M} \times \cup_{j\in\mathcal{M}}\Theta_j$, which is of lesser dimension than the full product space explored by the algorithm by Carlin and Chib (1995). Godsill (2001) showed that the RJMCMC algorithm by Green (1995) could also be derived as a particular form of Metropolis-Hastings updating from his product-space formulation thereby linking RJMCMC and the method by Carlin and Chib (1995).

When the parameters of different models are interrelated, RJMCMC is an effective way of building a Metropolis-Hastings sampler that exploits these relationships when proposing moves from one model to another. However, as with methods for estimating marginal likelihoods, the RJMCMC approach can be a challenge to understand and difficult to implement.

Godsill (2001) suggested that a hybrid approach that adopts the best aspects of RJMCMC and the algorithm by Carlin and Chib (1995) would be desirable. Here we elaborate one such hybrid solution based on a mild restriction to RJMCMC as outlined by Link and Barker (2010). We portray RJMCMC as simple Gibbs sampling, alternate sampling of full conditional distributions for M and a universal parameter $\boldsymbol{\psi}$ called the "palette" from which each model's parameters can be constructed. Our representation is simple and intuitive, and offers an advantage not considered by Godsill (2001)—the ability to post-process MCMC output from independently fitted models while exploiting interrelationships among parameters of the models.

Details on RJMCMC as Gibbs sampling of model and palette are given in Section 2, using a simple example for illustration. In Section 3, we give two further examples illustrating the challenges of implementing Bayesian multimodel inference. We conclude with a brief discussion in Section 4.

2. THE RJMCMC PALETTE

As outlined by Green (1995), a key step in RJMCMC is the specification of bijections describing relationships between the parameters of various models. These bijections allow reduction of the search space from the full product space considered by Chib (1995) to the union space $\mathcal{M} \times \cup_{j\in\mathcal{M}}\Theta_j$. There are $\binom{K}{2}$ such bijections, where K is the number of models in the model set $\mathcal{M}$; these are used to translate the parameters of one model into the parameters of another. When the dimensions of the two model differ, supplemental variables are introduced; these in conjunction with the bijections ensure detailed balance and convergence of the Markov chain to the target distribution Robert and (Casella 2010).

To demystify RJMCMC for an audience familiar only with ordinary MCMC, Link and Barker (2010) outlined an alternative, slightly restricted reformulation of RJMCMC. They introduced the idea of a universal parameter $\boldsymbol{\psi}$ called the "palette" from which all model-specific parameters can be calculated.

Instead of the welter of model-specific parameters, all of varying dimensions, there is a single quantity, $\boldsymbol{\psi}$, summarizing them all and of fixed dimension. More specifically, $\boldsymbol{\psi}$ is a vector of dimension $d \geq \max\{d_k\}$ where $d_k = \dim(\boldsymbol{\theta}_k)$, $k \in \mathcal{M}$. Parameter vector $\boldsymbol{\theta}_k$ can be recovered from the palette $\boldsymbol{\psi}$ by means of a known (invertible) mapping $g_k(\boldsymbol{\psi}) = \boldsymbol{\xi}_k = (\boldsymbol{\theta}_k', \boldsymbol{u}_k')'$; there are K such mappings. Vector $\boldsymbol{u}_k$ is irrelevant to model M_k, serving only to match the dimension of $\boldsymbol{\xi}_k$ and $\boldsymbol{\psi}$, so that $g_k(\cdot)$ can be defined as a bijection. For example, if model M_2 has parameter space of dimension 7, and $d = 10$, vector $\boldsymbol{u}_2$ will have dimension 3.

The main distinction between this presentation of RJMCMC and that by Green (1995) is that it is based on a set of K bijections, rather than the $\binom{K}{2}$ bijections Green describes. A full set of $\binom{K}{2}$ bijections consistent with Green's formulation can be obtained by compositions from the smaller set, viz. $g_{ij}(\cdot) = g_j \circ g_i^{-1}(\cdot)$.

With the palette in mind, RJMCMC is seen to be nothing more than Gibbs sampling, with alternating updates of M and $\boldsymbol{\psi}$. Full conditional distributions $[\boldsymbol{\psi}|\cdot]$ and $[M|\cdot]$ are determined by the joint model

$$[\boldsymbol{y}, \boldsymbol{\psi}, M] = [\boldsymbol{y}|\boldsymbol{\psi}, M][\boldsymbol{\psi}|M][M].$$

We describe the priors $[\boldsymbol{\psi}|M]$ in Section 2.1 and the full conditionals in Section 2.2.

2.1 Priors on Parameters

For each model M_k, we have a prior distribution for parameter $\boldsymbol{\theta}_k$, which by a slight abuse of notation we write as $[\boldsymbol{\theta}_k|M_k]$ (rather than $[\boldsymbol{\theta}_k|M = M_k]$). For multimodel inference, we require priors for the palette $[\boldsymbol{\psi}|M_k]$. Because $\boldsymbol{\psi} = g_k^{-1}(\boldsymbol{\xi}_k)$, the prior $[\boldsymbol{\psi}|M_k]$ is obtained by applying the change of variables theorem to the prior distribution $[\boldsymbol{\xi}_k|M_k]$. Since

$$[\boldsymbol{\xi}_k|M_k] = [\boldsymbol{\theta}_k, \boldsymbol{u}_k|M_k] = [\boldsymbol{\theta}_k|M_k]\,[\boldsymbol{u}_k|\boldsymbol{\theta}_k, M_k], \tag{2}$$

all that is needed is a specification of $[\boldsymbol{u}_k|\boldsymbol{\theta}_k, M_k]$. We make the convenient assumption that $\boldsymbol{u}_k$ is conditionally independent of $\boldsymbol{\theta}_k$, so that $[\boldsymbol{u}_k|\boldsymbol{\theta}_k, M_k] = [\boldsymbol{u}_k|M_k]$. The specific choice of $[\boldsymbol{u}_k|M_k]$ has no bearing on inference, but a good choice improves performance of the RJMCMC algorithm. Letting $f_k(\boldsymbol{\xi}_k) = [\boldsymbol{\xi}_k|M_k]$, the change of variables theorem yields

$$[\boldsymbol{\psi}|M_k] = f_k\left(g_k(\boldsymbol{\psi})\right)\left|\frac{\partial g_k(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}}\right|. \tag{3}$$

Example 1: Binomial success rates. Suppose $y_i \sim B(n_i, p_i)$ and we have observations $y_1 = 8$, $n_1 = 20$, $y_2 = 16$, and $n_2 = 30$. To evaluate the evidence for $p_1 \neq p_2$ versus $p_1 = p_2$, we consider two models:

M_1: p_i unrelated, with independent Be(α_i, β_i) priors, $i = 1, 2$.
M_2: $p_1 = p_2 \equiv \pi$ with prior Be(α_π, β_π) on π.

Here, we use $\text{Bin}(n, p)$ and $\text{Be}(a, b)$ to describe the binomial and beta distributions; the density function for $Y \sim \text{Bin}(n, p)$ will be denoted by $\text{Bin}(y; n, p)$ and that of $X \sim \text{Be}(a, b)$ by $\text{Be}(x; a, b)$.

A full product space MCMC sampler for comparing M_1 and M_2 requires search over a three-dimensional space associated with p_1, p_2, and π. Instead, we construct a two-dimensional palette, $\boldsymbol{\psi} = (\psi_1, \psi_2)'$. In M_1, we associate p_i with ψ_i; in M_2, we associate π with the weighted average $\bar{\psi} = (n_1\psi_1 + n_2\psi_2)/(n_1 + n_2)$ and a supplemental variable u with ψ_2. Thus, we define bijections $g_1(\boldsymbol{\psi}) = \boldsymbol{\psi} = (p_1, p_2)'$ and $g_2(\boldsymbol{\psi}) = \mathbf{A}\boldsymbol{\psi} = (\pi, u)'$ where

$$\mathbf{A} = \begin{pmatrix} \dfrac{n_1}{n_1 + n_2} & \dfrac{n_2}{n_1 + n_2} \\ 0 & 1 \end{pmatrix}.$$

For M_1, the bijection $g_1(\boldsymbol{\psi})$ is the identity, with Jacobian determinant $= 1$, so calculation of the prior is straightforward using Equations (2) and (3):

$$[\boldsymbol{\psi}|M_1] = \prod_{i=1}^{2} \text{Be}(\psi_i; \alpha_i, \beta_i) \times 1.$$

Under M_2, we have specified a beta prior for π; to calculate $[\boldsymbol{\psi}|M_2]$, we require a prior $[u|M_2]$. Because the choice has no bearing on inference, we can choose any prior we wish: an appealing choice is $[u|M_2] = \text{Be}(\alpha_u, \beta_u)$, with $\alpha_u = \alpha_2 + y_2$ and $\beta_u = \beta_2 + n_2 - y_2$. This is the posterior distribution for p_2 under M_1, ensuring that candidate values for p_2 are reasonable when Markov chain transitions from M_2 to M_1 are considered.

With these choices, the prior for $\boldsymbol{\psi}$ under M_2 is

$$[\boldsymbol{\psi}|M_2] = \text{Be}(\bar{\psi}; \alpha_\pi, \beta_\pi) \times \text{Be}(\psi_2; \alpha_u, \beta_u) \times \frac{n_1}{n_1 + n_2},$$

where $n_1/(n_1 + n_2)$ is the Jacobian determinant of the inverse transformation $\mathbf{A}^{-1}(\pi, u)$.

2.2 Gibbs Sampling

Our implementation of RJMCMC consists of alternating draws from $[\boldsymbol{\psi}|M, \mathbf{y}]$ and $[M|\boldsymbol{\psi}, \mathbf{y}]$.

2.2.1 *Drawing From* $[\boldsymbol{\psi}|M_k, \mathbf{y}]$

Sampling $\boldsymbol{\psi}$ from its full conditional distribution can be accomplished by drawing $\boldsymbol{\theta}_k$ from $[\boldsymbol{\theta}_k|M_k, \mathbf{y}]$, then sampling $\mathbf{u}_k$ from its prior $[\mathbf{u}_k|\boldsymbol{\theta}_k, M_k] = [\mathbf{u}_k|M_k]$ to form $\boldsymbol{\xi}_k = (\boldsymbol{\theta}_k', \mathbf{u}_k')'$, and then computing $\boldsymbol{\psi} = g_k^{-1}(\boldsymbol{\xi}_k)$.

The draw from $[\boldsymbol{\theta}_k|M_k, \mathbf{y}]$ can either be made directly, if the distribution is of convenient form, or by simulation. An alternative is to take a random draw from the stored MCMC output of an earlier analysis of model M_k.

2.2.2 *The Full-Conditional for* $\boldsymbol{M}$

The full-conditional $[M|\boldsymbol{\psi}, \mathbf{y}]$ is categorical with sample space

$\mathcal{M} = \{M_1, M_2, \ldots, M_K\}$ and

$$\Pr(M_k \mid \cdot) = \frac{[\mathbf{y} \mid \boldsymbol{\psi}, M_k]\,[\boldsymbol{\psi} \mid M_k]\,[M_k]}{\sum_j [\mathbf{y} \mid \boldsymbol{\psi}, M_j]\,[\boldsymbol{\psi} \mid M_j]\,[M_j]}, \tag{4}$$

for $k = 1, \ldots, K$. The likelihood $[\mathbf{y} \mid \boldsymbol{\psi}, M_k]$ is simply $[\mathbf{y} \mid \boldsymbol{\theta}_k, M_k]$; the prior $[\boldsymbol{\psi} \mid M_k]$ is calculated as described in Section 2.1.

If we are willing to calculate all of these probabilities, we can update M by a direct draw from this full-conditional distribution. Chain frequencies for $M = M_k$ converge to posterior probabilities $\Pr(M_k|\mathbf{y})$. Somewhat greater efficiency is available by monitoring $\Pr(M_k \mid \cdot)$, chain means of which also converge to the posterior model probabilities (Rao-Blackwellization, Casella and Robert 1996).

As an alternative to sampling the full conditional distribution for M, we can update the model by a Metropolis-Hastings step. Suppose that a move from M_j to M_k is proposed with probability $J(M_k|M_j)$. The move is made with probability

$$r = \min\left\{\frac{\Pr(M_k|\cdot)J(M_j|M_k)}{\Pr(M_j|\cdot)J(M_k|M_j)}, 1\right\};$$

otherwise, the chain remains in its current state. The advantage of this approach is that we need not compute all of the values $\Pr(M_k|\cdot)$; the cost is typically greater autocorrelation in the chain of values M.

Example 1: Binomial success rates (cont.). Our RJMCMC sampler for the binomial rates example proceeds as follows:

1. Given the current value of M, sample $\boldsymbol{\psi}$:
   - Under M_1, we sample p_i from $[p_i|M_1, \mathbf{y}] = \text{Be}(y_i + \alpha_i, n_i - y_i + \beta_i)$, for $i = 1, 2$. We then compute $\boldsymbol{\psi} = (p_1, p_2)'$.
   - Under M_2, we sample π from $[\pi|M_2, \mathbf{y}] = \text{Be}(y_1 + y_2 + \alpha_\pi, n_1 + n_2 - y_1 - y_2 + \beta_\pi)$ and u from $[u|M_2] = \text{Be}(y_2 + \alpha_2, n_2 - y_2 + \beta_2)$. We then compute $\boldsymbol{\psi} = \mathbf{A}^{-1}(\pi, u)'$.
2. Given the current value of $\boldsymbol{\psi}$, sample M. We use Equation (4) to compute full conditional model probabilities. The first step is to calculate model-specific parameters $(\boldsymbol{\theta}_k)$ from $\boldsymbol{\psi}$. Note that if the last sampled value of M was M_2, the value of $\boldsymbol{\psi}$ may correspond to an inadmissible value of p_1 under M_1. In this case, $\Pr(M_1|\cdot) = 0$ and the Markov chain for M will remain at M_2. We enforce this constraint by including the indicator function $\mathbf{I}(\psi_1 \in (0, 1))$ in the next equation.
   In general, $\Pr(M_1|\cdot)/\Pr(M_2|\cdot) =$

$$\left\{\prod_{i=1}^{2} \text{Bin}(y_i; n_i, \psi_i) \times \text{Be}(\psi_i; \alpha_i, \beta_i) \times \Pr(M_1) \times \mathbf{I}(\psi_1 \in (0, 1))\right\} \Big/ \left\{\prod_{i=1}^{2} \text{Bin}(y_i; n_i, \bar{\psi}) \times \text{Be}(\bar{\psi}; \alpha_\pi, \beta_\pi) \times \text{Be}(\psi_2; \alpha_u, \beta_u) \times \Pr(M_2) \times \frac{n_1}{n_1 + n_2}\right\}.$$

   Writing $P_1 = \Pr(M_1|\cdot)$, the above equation is $P_1/(1 - P_1)$; we solve for P_1 and sample M_1 with probability P_1, and M_2 otherwise.

To illustrate, we set $\alpha_1 = \beta_1 = \alpha_2 = \beta_2 = \alpha_\pi = 1$. Under M_1, we generate $p_1 \sim \text{Be}(9, 13)$ and $p_2 \sim \text{Be}(17, 15)$ and set $\boldsymbol{\psi} = (p_1, p_2)'$. Under M_2, we generate $\pi \sim \text{Be}(25, 27)$ and $u \sim \text{Be}(17, 15)$ and set $\boldsymbol{\psi} = \mathbf{A}^{-1}(\pi, u)' = ((5\pi - 3u)/2, u)'$. Assuming equal prior model weights, we sample M_1 with probability P_1, where

$$\frac{P_1}{1 - P_1} = \frac{\psi_1^8(1 - \psi_1)^{12}}{\bar{\psi}^{24}(1 - \bar{\psi})^{26}} \times \frac{16!\,14!}{31!} \times \frac{50}{20} \times \mathbf{I}(\psi_1 \in (0, 1)),$$

with $\bar{\psi} = (20\psi_1 + 30\psi_2)/50$; otherwise we sample M_2.

A chain of length 10^6 had $M = M_1$ 343,068 times, from which we estimate $\Pr(M_1|\boldsymbol{y}) = 0.343$; the posterior mean for P_1 was 0.342, which agrees to three decimal places with the true value, based on beta-binomial marginal distributions under the two models. We note that the chain for M mixed well, starting at M_2, making 184,109 transitions from M_2 to M_1 and 184,110 from M_1 to M_2. The autocorrelation for M was 0.14 at lag 1, and 0.035 at lag 2. The Bayes factor in favor of M_2 is the ratio of posterior odds to prior odds, estimated by (656,932/343,068) $\div$ 1 = 1.92, which agrees with the exact value to three significant digits.

2.3 Marginalizing the Gibbs Sampler for M

Consider the chain of sampled values $M^{(b)}$, with $b = 1, 2, \ldots, B$ representing sequence numbers for draws of M. The transition probabilities characterize the Markov chain for M; knowledge of them allows calculation of the stationary distribution of the chain, that is, of the posterior model probabilities $\Pr(M|\boldsymbol{y})$. Specifically, let $\phi_{ij} = \Pr(M^{(b+1)} = M_j | M^{(b)} = M_i)$, and let $\boldsymbol{\Phi} = (\{\phi_{ij}\})$ denote the transition matrix. The posterior model probabilities are the limiting distribution obtainable by normalizing the left eigenvector of the transition matrix associated with the eigenvalue 1.0 (Seber 2008). For example, with $K = 2$ the posterior odds $\Pr(M_2|\boldsymbol{y})/\Pr(M_1|\boldsymbol{y}) = \phi_{12}/\phi_{21}$. This observation suggests an alternative means of estimating posterior model probabilities, without actually producing a Markov chain of draws for M.

For a fixed value i in $1, \ldots, K$ we sample $[\boldsymbol{\psi}|M_i, \boldsymbol{y}]$ as before, that is, combining values sampled from $[\boldsymbol{\theta}_i|M_i, \boldsymbol{y}]$ and $[\boldsymbol{u}_i|M_i]$ to produce $\boldsymbol{\xi}_i$, then calculating $\boldsymbol{\psi} = g_i^{-1}(\boldsymbol{\xi}_i)$. Using this value for $\boldsymbol{\psi}$ we calculate $\Pr(M_j|\cdot)$, also as before. Instead of now sampling M, we draw another sample $[\boldsymbol{\psi}|M_i, \boldsymbol{y}]$ and again calculate $\Pr(M_j|\cdot)$, $j = 1, 2, \ldots, K$. The average value of vector $(\Pr(M_1|\cdot), \Pr(M_2|\cdot), \ldots, \Pr(M_K|\cdot))'$ across draws from $[\boldsymbol{\psi}|M_i, \boldsymbol{y}]$ is a Rao-Blackwellized estimate of the ith row of $\boldsymbol{\Phi}$. Repeating the process for $i = 1, 2, \ldots, K$, we obtain an estimate of $\boldsymbol{\Phi}$ from which we can estimate posterior model probabilities $\Pr(M_j|\boldsymbol{y})$.

This approach allows us to fix the number of samples associated with each model, rather than sampling models with frequencies proportional to their posterior probabilities.

3. EXAMPLES

3.1 Geometric Versus Poisson

Consider the dataset $\boldsymbol{y} = (0, 1, 2, 3, 8)'$. The sample variance is nearly 3.5 times larger than the sample mean, favoring a geometric distribution, $f_p(y) = p(1 - p)^y$, over a Poisson distribution, $f_\mu(y) = \exp(-\mu)\mu^y/y!$; we seek an objective Bayesian assessment of the relative support for the two models.

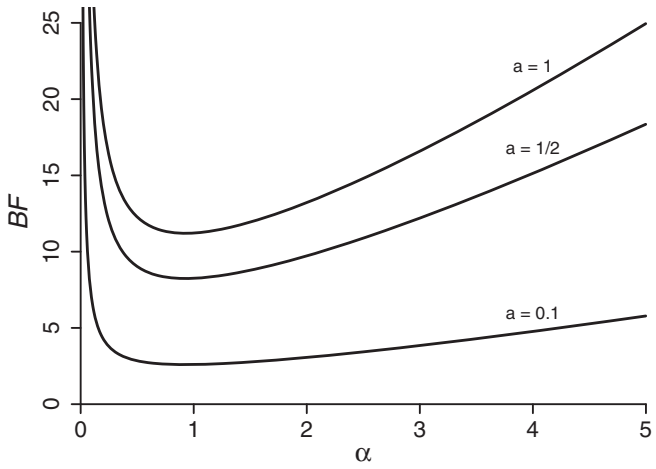

Figure 1. Bayes factors (BF) in favor of the geometric versus the Poisson model resulting from gamma priors with $\alpha = \beta$ and beta priors with $a = b$; these are plotted for fixed values a as functions of α.

This example provides clear illustration of the sensitivity of Bayes factors to choices of model-specific parameters. A gamma prior $Ga(\alpha, \beta)$ (shape = α, rate = β) for μ is conjugate under the Poisson assumption, and a beta prior $\text{Be}(a, b)$ is conjugate for p under the geometric assumption; with these priors, marginal distributions for $\boldsymbol{y}$ are easily calculated in closed form, so the Bayes factor is as well. Posterior inference for p or for μ is fairly insensitive to choices of $0 < a, b \leq 1$, and $0 < \alpha, \beta \leq 1$, but the Bayes factor is not. The extreme sensitivity to priors is evident in Figure 1, which displays Bayes factors in favor of the geometric distribution resulting from gamma priors with $\alpha = \beta$ and beta priors with $a = b$; these are plotted for fixed values a as functions of α.

It is possible to choose priors for the parameters of the two models such that the marginal distributions of individual y_i are identical; the models are distinguished by the joint distributions of $\boldsymbol{y}$. This seems an appealing feature for an objective comparison of families of distributions, avoiding favoring one family over another due to choice of priors on unknown parameters within the families. We have previously described prior specifications as "nonpreferential" when models are distinguished by their marginal joint distributions rather than the marginal distributions of individual observations (Link and Barker 2008).

The first step to obtaining nonpreferential priors is to note that geometric random variables are exponential mixtures of Poisson random variables, so the two models can be expressed as specifying $y_i|\lambda_i \sim \text{Pois}(\lambda_i)$, the difference being in variation among values λ_i.

M_1: $y_i|\lambda_i \sim \text{Pois}(\lambda_i)$, $\lambda_i \equiv \mu$, so $y_i|\mu$ are exchangeable $\text{Pois}(\mu)$ random variables; μ has prior $g(\mu)$.

M_2: $y_i|\lambda_i \sim \text{Pois}(\lambda_i)$, λ_i are iid $\text{Exp}(\alpha)$ random variables, so $y_i|\alpha$ are exchangeable geometric random variables: $\Pr(y_i = k|\alpha) = p(1 - p)^k$ where $p = \alpha/(1 + \alpha)$; α has prior $h(\alpha)$.

Suppose that under M_1 we sample μ from an $\text{Exp}(\alpha)$ distribution, with α having the same prior as under M_2; that is, we

set

$$g(\mu) = \int \alpha \exp(-\alpha\mu) h(\alpha) d\alpha.$$

The result is that the priors on individual values λ_i are identical under the two models; the difference is that the values of λ_i are constant under the Poisson model M_1, but exchangeable under the geometric model, M_2. Thus, the models have common marginal distributions for individual y_i, but distinct joint distributions for $\boldsymbol{y}$.

The proposed form of the priors complicates analytical calculation of the Bayes factor, but RJMCMC can be implemented.

3.1.1 RJMCMC for Poisson Versus Geometric

To conduct RJMCMC, we must define a palette $\boldsymbol{\psi} = (\psi_1, \psi_2, \ldots, \psi_6)'$ and bijections $g_1(\boldsymbol{\psi}) = (u_1, \ldots, u_4, \mu, \alpha)' = (\boldsymbol{u}_1', \boldsymbol{\theta}_1')'$ and $g_2(\boldsymbol{\psi}) = (\lambda_1, \ldots, \lambda_5, \alpha)'$. We must assign model-specific priors on $\boldsymbol{\psi}$, conforming to the desired priors for the original models.

We begin with M_2, setting $g_2(\boldsymbol{\psi}) = \boldsymbol{\psi}$. Thus, under M_2, $\lambda_i = \psi_i$ for $i = 1, 2, \ldots, 5$, and $\alpha = \psi_6$. The prior $[\boldsymbol{\psi}|M_2]$ is defined by supposing $\psi_6 \sim h(\cdot)$, and that given ψ_6, variables $\psi_1, \psi_2, \ldots, \psi_5$ have independent Exp(ψ_6) priors. Thus,

$$[\boldsymbol{\psi}|M_2] = h(\psi_6) \times \prod_{i=1}^{5} \text{Ga}(\psi_i; 1, \psi_6).$$

The identity bijection and this choice of priors trivially satisfy the specifications for M_2.

Given that under M_2, $\psi_i = \lambda_i$ for $i = 1, 2, \ldots, 5$, and that under M_1, $\lambda_i \equiv \mu$, it seems sensible that under M_1 we should set $\mu = \bar{\psi}_{1:5} = \frac{1}{5}\sum_{i=1}^{5} \psi_i$. We also retain the association of ψ_6 with α. For convenience (as will become clear) we define the bijection

$$g_1(\boldsymbol{\psi}) = \left(\frac{\psi_1}{5\bar{\psi}_{1:5}}, \frac{\psi_2}{5\bar{\psi}_{1:5}}, \frac{\psi_3}{5\bar{\psi}_{1:5}}, \frac{\psi_4}{5\bar{\psi}_{1:5}}, \bar{\psi}_{1:5}, \psi_6 \right)'.$$

As before, our prior $[\boldsymbol{\psi}|M_1]$ will have $\psi_6 \sim h(\cdot)$, and the distributions of other ψ_i's conditionally independent given ψ_6. This time, however, we specify that ψ_i, $i = 1, 2, \ldots 5$ are exchangeable Ga(1/5, ψ_6/5) random variables. Thus,

$$[\psi|M_1] = h(\psi_6) \times \prod_{i=1}^{5} \text{Ga}(\psi_i; 1/5, \psi_6/5).$$

It follows from elementary properties of the gamma distribution that, conditional on ψ_6, the mean $\bar{\psi}_{1:5} = \frac{1}{5}\sum_{j=1}^{5} \psi_j$ has an Exp(ψ_6) distribution and that, independently, $\frac{1}{5\bar{\psi}_{1:5}}(\psi_1, \psi_2, \psi_3, \psi_4, \psi_5)'$ has a Dirichlet distribution with parameter vector $(1/5, 1/5, \ldots, 1/5)'$. Thus, $[\mu|\alpha, M_1] = \text{Exp}(\alpha)$ and $[\alpha|M_1] = h(\alpha)$; furthermore, $[\boldsymbol{u}, \boldsymbol{\theta}_1|M_1] = [\boldsymbol{u}|M_1][\boldsymbol{\theta}_1|M_1]$.

3.1.2 Implementation

Choosing an improper gamma prior $h(\alpha) = \text{Ga}(\alpha; 0, 0)$ is equivalent to choosing an improper Ga(0, 0) prior for μ under M_1, and an improper Be(0, 0) prior for p under M_2. Thus, model-specific posteriors are $[\mu|\boldsymbol{y}, M_1] = \text{Ga}(14, 5)$ and $[p|\boldsymbol{y}, M_1] = \text{Be}(5, 14)$. We use these to produce a Gibbs sampler for M and $\boldsymbol{\psi}$, as follows.

First consider draws of $[\boldsymbol{\psi}|M, \boldsymbol{y}]$. Under M_1, we sample $\mu \sim \text{Ga}(14, 5)$, and set $\bar{\psi}_{1:5} = \mu$. We independently sample $\boldsymbol{\pi}$ from a Dirichlet distribution with parameter vector $(1/5, 1/5, \ldots, 1/5)'$, and set $(\psi_1, \psi_2, \psi_3, \psi_4, \psi_5)' = 5\bar{\psi}_{1:5}\boldsymbol{\pi}$. We complete the draw of $[\boldsymbol{\psi}|M_1, \boldsymbol{y}]$ by drawing from the full conditional for ψ_6 under M_1, which is an exponential distribution with parameter $\bar{\psi}_{1:5}$. Under M_2, we begin by converting a draw $p \sim \text{Be}(5, 14)$ to $\psi_6 = \alpha = p/(1-p)$. Noting that

$$[\boldsymbol{\psi}|M_1, \boldsymbol{y}, \psi_6] \propto \prod_{i=1}^{5} \left\{ \frac{\psi_i^{y_i} e^{-\psi_i}}{y_i!} \, e^{-\psi_6 \psi_i} \right\},$$

we perform independent draws ψ_i from Ga($y_i + 1, \alpha + 1$) for $i = 1, 2, \ldots, 5$, to complete the draw of $[\boldsymbol{\psi}|M_2, \boldsymbol{y}]$.

Now consider draws of $[M|\boldsymbol{\psi}, \boldsymbol{y}]$. Given $\boldsymbol{\psi}$, we need only compute

$$P_1 = \frac{[\boldsymbol{y}|M_1, \boldsymbol{\psi}]\,[\boldsymbol{\psi}|M_1]\,[M_1]}{[\boldsymbol{y}|M_1, \boldsymbol{\psi}]\,[\boldsymbol{\psi}|M_1]\,[M_1] + [\boldsymbol{y}|M_2, \boldsymbol{\psi}]\,[\boldsymbol{\psi}|M_2]\,[M_2]},$$

and average P_1 across draws for $\boldsymbol{\psi}$.

We analyzed the dataset $\boldsymbol{y} = (0, 1, 2, 3, 8)'$, producing a Markov chain of length 10,000. Using equal prior model weights, convergence to the exact value for $\Pr(M = 1|\boldsymbol{y})$ of 0.917 ($\text{BF}_{12} = 12.17$) is rapid occurring within about 10,000 iterations (Figure 2).

3.2 Regression Models

Selection among regression models is often conducted using indicator variables in describing regression functions (O'Hara and Sillanpää 2009). For example, Link and Barker (2006) reported an analysis based on fitting logistic regression models to the return rates for brown trout expressed in terms of sex S_i and length L_i effects. Modeling the return indicator

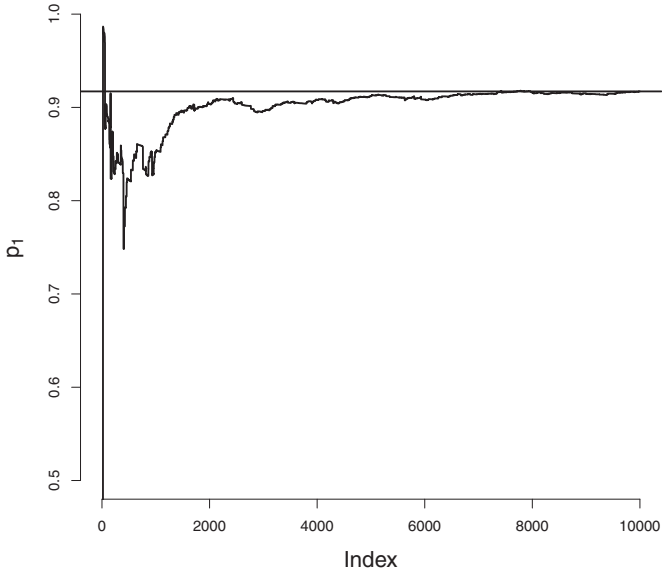

Figure 2. Index plot of the cumulative estimates of the posterior probability $\Pr(M = 1|\boldsymbol{y})$ for the Poisson versus geometric example. The horizontal black line represents the exact value of 0.917.

$y_i \sim \text{Bern}(p_i)$ they considered five models for $\eta_i = \text{logit}(p_i)$, namely M_1: β_0 (constant); M_2: $\beta_0 + \beta_1 S_i$; M_3: $\beta_0 + \beta_2 L_i$; M_4: $\beta_0 + \beta_1 S_i + \beta_2 L_i$, and M_5: $\beta_0 + \beta_1 S_i + \beta_2 L_i + \beta_3 S_i L_i$.

The five models might be expressed simultaneously as

$$\eta_i = \beta_0 + \mathbf{I}_1\beta_1 S_i + \mathbf{I}_2\beta_2 L_i + \mathbf{I}_1\mathbf{I}_2\mathbf{I}_3\beta_3 S_i L_i,$$

where $\mathbf{I}_1$ and $\mathbf{I}_2$ are indicator variables for inclusion of sex and length effects, and $\mathbf{I}_3$ is an indicator for inclusion of the interaction term. Modeling the indicators as independent Bernoulli observations with success parameters γ_i, one can calculate prior model probabilities $\pi_k = \Pr(M = M_k)$. Note, however, that with three γ_i's, we can show that one cannot calculate the full range of admissible five-dimensional priors $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_5)'$; one cannot, for instance, assign equal prior weights to all models. The approach leads to a restriction on prior model probabilities that is also reflected in posterior model probabilities.

It is possible to represent the full range of priors on model with four independent indicator variables $\mathbf{J}_j$:

$$\begin{aligned}\eta_i = \beta_0 &+ [(1-\mathbf{J}_4)(1-\mathbf{J}_3)\mathbf{J}_2(1-\mathbf{J}_1) + (1-\mathbf{J}_4)\mathbf{J}_3 + \mathbf{J}_4]\,\beta_1 S_i \\ &+ [(1-\mathbf{J}_4)(1-\mathbf{J}_3)\mathbf{J}_2\mathbf{J}_1 + (1-\mathbf{J}_4)\mathbf{J}_3 + \mathbf{J}_4]\,\beta_2 L_i \\ &+ \mathbf{J}_4\beta_3 S_i L_i,\end{aligned}$$

but the interpretation of the indicators becomes more complex. Furthermore, the result is equivalent to the much simpler

$$\begin{aligned}\eta_i = \beta_0 &+ \mathbf{I}(M \in \{M_2, M_4, M_5\})\beta_1 S_i \\ &+ \mathbf{I}(M \in \{M_3, M_4, M_5\})\beta_2 L_i + \mathbf{I}(M = M_5)\beta_3 S_i L_i,\end{aligned}$$

obtained by describing the models by the categorical variable M, as in RJMCMC.

Another advantage of RJMCMC for variable selection is the ease with which we may assign model-specific priors on parameters: we may wish, for instance, that the prior on β_1 be different under M_2 and M_5. Link and Barker (2006) assigned independent mean-zero normal priors to β_j's, with variances $(d_k V)^{-1}$ depending on the number d_k of coefficients in the linear predictor under M_k. The choice equalizes the prior mean and variance of the linear predictor η_i across models, thus approximating the nonpreferentiality criterion discussed in Section 3.1. Having standardized the regressors S, L, and SL, the prior variances of η_i average V^{-1}.

To implement RJMCMC for the trout return data, we define a five-dimensional parameter $\boldsymbol{\psi}$ and bijections $g_k(\boldsymbol{\psi}) = \boldsymbol{\psi}$ for $k = 1, 2, \ldots, 5$, with

$$\begin{aligned}\boldsymbol{\psi} &= (\beta_0, u_1, u_2, u_3, V)', \quad \text{if } M = M_1, \\ \boldsymbol{\psi} &= (\beta_0, \beta_1, u_2, u_3, V)', \quad \text{if } M = M_2, \\ \boldsymbol{\psi} &= (\beta_0, u_1, \beta_2, u_3, V)', \quad \text{if } M = M_3, \\ \boldsymbol{\psi} &= (\beta_0, \beta_1, \beta_2, u_3, V)', \quad \text{if } M = M_4, \\ \boldsymbol{\psi} &= (\beta_0, \beta_1, \beta_2, \beta_3, V)', \quad \text{if } M = M_5.\end{aligned}$$

For all models, we assigned $\psi_5 (\equiv V)$ a Ga(3.29, 7.80) prior. This choice was motivated by the observation that $\text{logit}(p) \sim \mathcal{N}(0, V^{-1})$ and $V \sim \text{Ga}(3.29, 7.80)$ implies an approximately $U(0, 1)$ prior for p. The prior specifications $[\boldsymbol{\psi}|M_k]$, $k = 1, 2, \ldots, 5$ were completed by specifying that conditional on ψ_5, $[\psi_i|\psi_5, M_k] = N(\psi_i; 0, (n_k V)^{-1})$ are independent normal distributions for $i = 1, 2, 3, 4$.

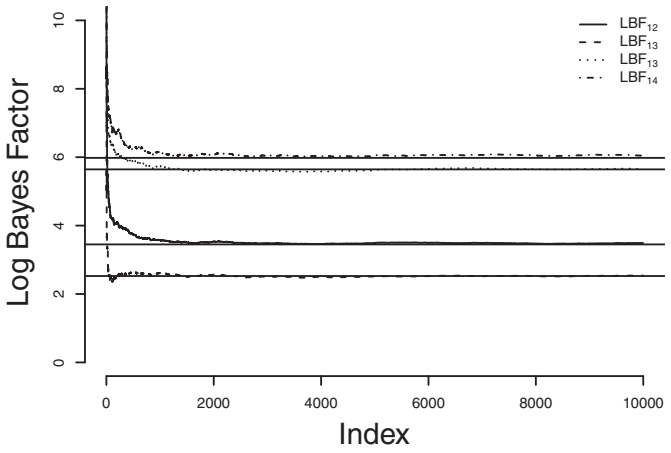

Figure 3. Index plot of the cumulative estimates of the log-scale Bayes factor for the trout logistic regression example. The horizontal black lines represent the values obtained after 500,000 iterations and LBF_{1j} is used to denote the log-scale Bayes factor in favor of model 1 relative to model j.

We estimated the Bayes factors by estimating the transition matrix for constant prior model weights. Our estimates converged rapidly to the values reported by Link and Barker (2006) (Figure 3).

4. DISCUSSION

Bayesian inference offers an appealing framework for multimodel inference, but the difficulties of computing Bayes factors or posterior model probabilities can be a barrier to implementation. The palette version of RJMCMC presented here has the appeal of simplicity: the process is nothing more than Gibbs sampling of model M and a universal parameter $\boldsymbol{\psi}$, from which parameters for each candidate model can be computed.

RJMCMC is typically conceived of as a simultaneous investigation of parameters and models. However, if model-specific posterior distributions for parameters $[\boldsymbol{\theta}_k|M_k, \boldsymbol{y}]$ are known, the focus of RJMCMC can be shifted to the relative support for models alone. The form of $[\boldsymbol{\theta}_k|M_k, \boldsymbol{y}]$ may be known analytically (as in cases of conjugacy) or through preliminary model-specific MCMC applications. In the latter case, the palette representation of RJMCMC can be considered as a post-processing of model-specific analyses. Given $[\boldsymbol{\theta}_k|M_k, \boldsymbol{y}]$, sampling $[\boldsymbol{\psi}|M_k, \boldsymbol{y}]$ is straightforward.

We may thus break the problem of multimodel inference into distinct model-fitting and model-comparison steps. The model-fitting steps involve model-specific evaluation of posterior distributions, that is, $[\boldsymbol{\theta}_k|M_k, \boldsymbol{y}]$, from which we obtain $[\boldsymbol{\psi}|M_k, \boldsymbol{y}]$. The model-comparison step consists of constructing a Markov chain sampler for a categorical variable on the model set. As noted, we need not sample this Markov chain, but only to compute its transition probabilities, which determine posterior model probabilities. Transition probability ϕ_{kj} (the probability of a move from M_k to M_j) is obtained as the average value of $\Pr(M = M_j|\boldsymbol{\psi}, \boldsymbol{y})$ over values $\boldsymbol{\psi}$ sampled from $[\boldsymbol{\psi}|M_k, \boldsymbol{y}]$.

Viewed thus, model-specific posterior distributions $[\boldsymbol{\theta}_k|M_k, \boldsymbol{y}]$ are inputs to model comparisons; an associate editor noted that one might regard model-specific MCMC outputs as computational "objects" for further analysis.

This two-stage approach to model fitting and comparison seems natural. For example, an alternative model comparison approach in Bayesian modeling is to first fit a set of models under consideration and then to compare them using the deviance information criterion (DIC) by Spiegelhalter et al. (2002). The approach we have outlined has the advantage of allowing for coherent comparison of models by use of posterior model probabilities.

Suppose that a dataset has been analyzed under K models, and that Bayes factors BF_{kj}, $1 \leq k, j \leq K$ have been computed. A new model M_{K+1} can be readily added to the model set by comparing it to a single model (say, M_1) from the original set. Constructing a Markov chain sampler on the set $\{M_1, M_{K+1}\}$, the posterior odds in favor of M_{K+1} are given by ratio of transition probabilities $\phi_{1,K+1}/\phi_{K+1,1}$ for this new chain; from these we calculate the new Bayes factor $\mathrm{BF}_{K+1,1}$. Bayes factors for comparing the new model to previous models can then be calculated as $\mathrm{BF}_{K+1,j} = \mathrm{BF}_{K+1,1} \times \mathrm{BF}_{1,j}$.

In considering use of RJMCMC for estimating Bayes factors or posterior model probabilities, there are technical and philosophical issues to be faced. The choice of bijections relating $\boldsymbol{\psi}$ to the parameters of various models requires some thought, as does the choice of priors on dimension-matching parameters $\boldsymbol{u}_k$; these choices affect the mixing of the Markov chain for the model. One potentially serious issue with BMI is inferential sensitivity to the choice of priors on parameters; it is well-known that Bayes factors are sensitive to this choice, especially with vague priors.

Our view is that if we want to choose among families of models, and the families are described by parameters about which we have little or no prior knowledge, then the choice of priors for those parameters should be made with the object of minimizing its influence on our choice among families of models. If, for instance, we wish to penalize models for complexity, that penalty should be made explicitly by our choice of prior weights on models, rather than accidentally, due to priors on parameters having different dimensions. In this spirit, our choice of priors in the Poisson versus geometric example equated the marginal sampling distributions $[y_i|M_k]$ so that differences among models reflect differences in the marginal joint distributions of data $[\boldsymbol{y}|M_k]$ rather than the effect of priors.

[Received December 2012. Revised March 2013.]

REFERENCES

Bartolucci, F., Scaccia, L., and Mira, A. (2006), "Efficient Bayes Factor Estimation From the Reversible Jump Output," *Biometrika*, 93, 41–52. [150]

Carlin, B. P., and Chib, S. (1995), "Bayesian Model Choice via Markov Chain Monte Carlo Methods," *Journal of the Royal Statistical Society,* Series B, 57, 473–484. [151]

Casella, G., and Robert, C. (1996), "Rao-Blackwellisation of Sampling Schemes," *Biometrika*, 83, 81–94. [152]

Chib, S. (1995), "Marginal Likelihood From the Gibbs Output," *Journal of the American Statistical Association*, 90, 1313–1321. [150,151]

Congdon, P. (2006), "Bayesian Model Choice Based on Monte Carlo Estimates of Posterior Model Probabilities," *Computational Statistics and Data Analysis*, 50, 346–357. [150]

Godsill, S. J. (2001), "On the Relationship Between Markov Chain Monte Carlo Computation and Bayesian Model Determination," *Journal of Computational and Graphical Statistics*, 10, 230–248. [151]

Green, P. J. (1995), "Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination," *Biometrika*, 82, 711–732. [150,151]

Han, C., and Carlin, B. (2001), "Markov Chain Monte Carlo Methods for Computing Bayes Factors: A Comparative Review," *Journal of the American Statistical Association*, 96, 1122–1133. [150]

Kass, R., and Raftery, A. (1995), "Bayes Factors," *Journal of the American Statistical Association*, 90, 115–114. [150]

Link, W. A., and Barker, R. J. (2006), "Model Weights and the Foundations of Multi-Model Inference," *Ecology*, 87, 2626–2635. [154,155]

——— (2008), "Bayes Factors and Multimodel Inference," *Environmental and Ecological Statistics*, 3, 597–618. [150,153]

——— (2010), *Bayesian Inference With Ecological Applications*, London: Academic Press. [151]

Meng, X. L., and Wong, W. H. (1996), "Simulating Ratio of Normalizing Constant via a Simple Identity: A Theoretical Exploration," *Statistica Sinica*, 6, 831–860. [150]

Newton, M. A., and Raftery, A. E. (1994), "Approximate Bayesian Inference With the Weighted Likelihood Bootstrap," *Journal of the Royal Statistical Society,* Series B, 56, 3–48. [150]

O'Hara, R. B., and Sillanpää, M. J. (2009), "Review of Bayesian Variable Selection Methods: What, How and Which," *Bayesian Analysis*, 4, 85–118. [154]

Robert, C. P., and Casella, G. (2010), *Monte Carlo Statistical Methods* (2nd ed.), New York: Springer. [151]

Robert, C. P., and Marin, J.-M. (2008), "On Some Difficulties With a Posterior Probability Approximation Technique," *Bayesian Analysis*, 3, 427–442. [150]

Seber, G. A. F. (2008), *A Matrix Handbook for Statisticians*, Hoboken, NJ: John Wiley. [153]

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002), "Bayesian Measures of Model Complexity and Fit," *Journal of the Royal Statistical Society,* Series B, 64, 583–639. [156]