Biostatistics (2001), **2**, 4, *pp.* 485–500
Printed in Great Britain

Empirical Bayes Gibbs sampling

GEORGE CASELLA
Department of Statistics, University of Florida, Gainesville, FL 32611, USA
casella@stat.ufl.edu

SUMMARY

The wide applicability of Gibbs sampling has increased the use of more complex and multi-level hierarchical models. To use these models entails dealing with hyperparameters in the deeper levels of a hierarchy. There are three typical methods for dealing with these hyperparameters: specify them, estimate them, or use a 'flat' prior. Each of these strategies has its own associated problems. In this paper, using an empirical Bayes approach, we show how the hyperparameters can be estimated in a way that is both computationally feasible and statistically valid.

Keywords: Bayesian computation; Consistency; Empirical Bayes; Hierarchical models; Likelihood.

1. INTRODUCTION

Computation using the Gibbs sampler (Geman and Geman, 1984; Gelfand and Smith, 1990) has made Bayesian estimation in complex hierarchical models not only feasible, but almost routine. A consequence of this development is that, to do a Bayesian analysis, an experimenter may be asked to specify values for hyperparameters that are in deep levels of a hierarchy. A difficulty arises here in that such values may be difficult to specify, defying not only intuition, but any obvious connection to the problem at hand.

One consequence of this difficulty with hyperparameters is that they tend to be ignored. There is some basis for this, as there is a certain 'robustness' to specification of parameters that lie deeper in a hierarchy (Goel and DeGroot, 1981). However, there is not a universal robustness and, especially if the hyperparameter is estimated, it could be important to assess the actual sensitivity of the inference to the specification of the hyperparameter.

For example, we look at data analysed by Efron (1996), a meta-analysis of randomized trials of a new surgical treatment for stomach ulcers. The data, given in Appendix A.4, are paired binomial experiments in 39 medical centers. (Efron (1996) presents data for 41 medical centers, but drops the last two in his analysis. For consistency we also eliminate those cases (but see Morris, 1996, for another opinion). We also have flipped Efron's definition of 'success' so that here a 'success' is a desirable outcome.) In medical center (experiment) j, $j = 1, 2, \ldots, 39$ we observe

$$Y_{1j} \sim \text{binomial}(n_{1j}, p_{1j}) \text{ and } Y_{2j} \sim \text{binomial}(n_{2j}, p_{2j}),$$

where Y_{1j} are the successes from the treatment, and Y_{2j} are the successes from the control. As we might expect there to be a relationship between p_{1j} and p_{2j} for each j, we could model this in a hierarchical model using a bivariate distribution with a correlation between p_{1j} and p_{2j}. The standard bivariate Dirichlet distribution is an obvious candidate, but it has the property that the variables must be negatively correlated. In terms of this example, this would force a negative correlation between success rates within a medical center, which does not seem reasonable. With a slight redefinition, however, we can have a form

of the bivariate Dirichlet in which correlation between p_{1j} and p_{2j} is non-negative. (Appendix A.2 gives more details on this and alternative priors.) This leads to the hierarchical model

$$\begin{aligned} Y_{1j} \sim \text{binomial}(n_{1j}, p_{1j}) \quad \text{and} \quad Y_{2j} \sim \text{binomial}(n_{2j}, p_{2j}), \quad j = 1, 2, \ldots, 39 \\ \pi(p_{1j}, p_{2j}) \quad \propto \quad (1 - p_{1j})^{a-1} p_{2j}^{b-1} (p_{1j} - p_{2j})^{c-1} \end{aligned} \tag{1}$$

where a, b, and c are hyperparameters.

To carry out an analysis of this model, it is necessary to specify values for the hyperparameters a, b, and c. Although these parameters can be given meaning in terms of means and variances of the prior distribution (see Appendix A.2), deciding on specific values could be difficult. Moreover, if the values were specified, it would be wise to then somehow assess the sensitivity of the answer. We propose to estimate these parameters and then do the Bayesian analysis, and demonstrate that this methodology is a valid one.

Estimation of hyperparameters seems common, especially in Gibbs sampling implementations, but is not often accompanied with an assessment of the effect of the estimation. For the ulcer data, Figure 1 shows the results of the analysis that we propose. The posterior distribution is calculated using estimated values of the hyperparameters. Along with this posterior density, we also display an envelope of densities corresponding to the values of the hyperparameters that fall within one standard deviation of the estimated values. This range of hyperparameter values is constructed using likelihood methods, and we can actually construct valid (asymptotically) confidence sets with respect to the marginal distribution of the data. Moreover, a display similar to this can be constructed for various ranges of the hyperparameters, which could give a good idea of the effect of the value of the hyperparameters on the resulting posterior density, and thus identify the importance of the hyperparameter in the overall inference. Details are given in Section 4.

The actual model that we are fitting is, in fact, an empirical Bayes model in the spirit of Morris (1983) and Efron (1996). Moreover, the EM/Gibbs algorithm that we describe in Section 3 can be adapted to the set-up of Efron (1996). To overcome some of the difficulty in calculating marginal MLEs, Efron considered models based on the exponential family. Using the EM/Gibbs algorithm, the applicability of Efron's models can be expanded. Moreover, in contrast to the comments of Gelfand (1996), with our new algorithm it may be the case that the empirical Bayes model is now easier to fit than a hierarchical model. We discuss this point further in Section 5.

A common implementation of the hierarchy (1) is to estimate some subset of a, b, and c (and specify the remaining ones), and calculate $\pi(p_i|\mathbf{y}, \hat{a}, \hat{b}, \hat{c})$ from the Gibbs sampler. Strictly speaking, this is not a Bayesian analysis, as the posterior distribution will depend on estimated hyperparameters. In fact, this is a 'classical empirical Bayes' case (see, for example, Morris, 1983) where we need to account for the variation due to the estimated hyperparameters. We proceed as follows. Conditional on the values of the hyperparameters, we are content to use a Bayesian posterior distribution. However, at the hyperparameter level, we will use maximum likelihood theory to estimate the parameters and assess the errors, and then use these estimates to understand how precise our estimated posterior distribution is. We are thus using a frequentist error calculation to assess the accuracy of a Bayesian inference, which is in the spirit of a robust Bayes analysis. Our main concern is with the properties of such an estimated procedure.

Bayesian inference based on data-dependent priors is not new, and can be traced back to, at least, Berger (1984), although one could argue that the nonparametric empirical Bayes formulation of Robbins (1964, 1983) is also a case of this. (However, Robbins was not directly concerned with a Bayesian inference, rather with a minimax property.) In addition to the previously mentioned parametric empirical Bayes formulation of Morris (1983) (see also Carlin and Louis, 1996), other recent uses of data-dependent priors include O'Hagan (1995) and Berger and Perrichi (1996). Shively *et al.* (1999) use data-dependent priors in a way similar to ours, but do not go into detail on assessment of the effect of hyperparameter

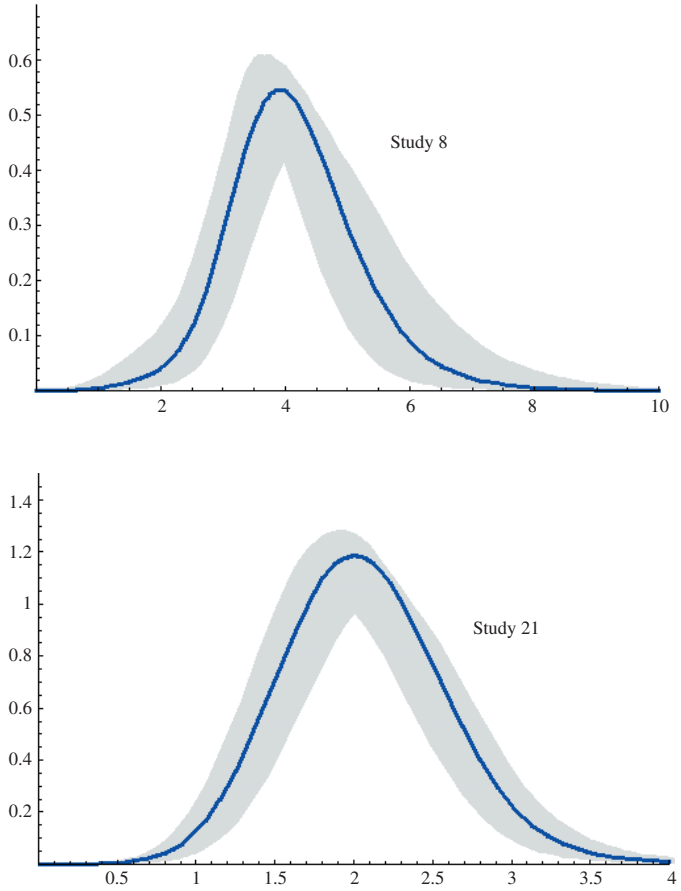

Fig. 1. Posterior density of the logodds of treatment versus control. The top panel is study 8 and the bottom panel is study 21. The envelope around the posterior consists of posterior densities with hyperparameter values within one standard deviation of the estimated hyperparameters. Note that the envelope is tighter for study 21, which had more observations than study 8.

error. Wasserman (2000), working in mixture models, shows that data-dependent priors can yield second-order-correct frequentist confidence intervals.

The remainder of the paper is organized as follows. In Section 2 we develop the empirical Bayes Gibbs sampler, look at some theoretical properties of the estimated posterior distribution, and give a convergence result (with the proof in Appendix A.1). Section 3 goes into detail on the implementation of the algorithm. Here we show that with very little overhead, the entire empirical Bayes Gibbs sampler can be implemented using little more than the calculations of the original Gibbs sampler. In Section 4 we apply the empirical Bayes Gibbs sampler to the data described in the introduction. Section 5 contains a discussion.

2. The empirical Bayes Gibbs sampler

The Gibbs sampler has found many of its applications in hierarchical models. We begin by defining a 'generic hierarchy', from which we develop the Gibbs sampler and its empirical Bayes version. Let X

have sampling distribution $f(x|\theta, \psi)$, where θ is a parameter of interest and ψ is a nuisance parameter. We observe p independent copies of X, each with its own parameter θ. We then model the θ_i with a common prior distribution, whose parameters may, in turn, have a prior distribution. This all results in what has come to be known as a *conditionally independent hierarchical model* (Kass and Steffey, 1989):

$$\begin{aligned} X_i &\sim f(x|\theta_i, \psi) \quad i = 1, 2, \ldots, p \\ \theta_i &\sim \pi(\theta|\lambda, \psi) \\ \lambda &\sim g(\lambda|\psi). \end{aligned} \tag{2}$$

Although we model λ as a common parameter (the simple empirical Bayes case), neither λ nor ψ need be scalars.

In a typical application of the Gibbs sampler we specify ψ and, based on observations $\mathbf{x} = (x_1, \ldots, x_p)$, we would set up iterations between the full conditional posterior distributions

$$\begin{aligned} \theta^{(j+1)} &\sim \pi(\theta|\mathbf{x}, \psi, \lambda^{(j)}) \\ \lambda^{(j+1)} &\sim g(\lambda|\mathbf{x}, \psi, \theta^{(j+1)}) \end{aligned} \tag{3}$$

for $j = 1, \ldots, M$, to produce estimates of the marginal posteriors

$$\pi(\theta|\mathbf{x}, \psi) \qquad \text{and} \qquad g(\lambda|\mathbf{x}, \psi).$$

If the experimenter is unable, or unwilling, to specify ψ, an alternative is to first estimate ψ with $\hat{\psi}$ and run the *empirical Bayes Gibbs sampler* iterations

$$\begin{aligned} \theta &\sim \pi(\theta|\mathbf{x}, \hat{\psi}, \lambda) \\ \lambda &\sim g(\lambda|\mathbf{x}, \hat{\psi}, \theta). \end{aligned}$$

The Gibbs sampler works as usual—that is, as in (3)—and, for example, produces the estimated posterior distribution

$$\hat{\pi}(\theta|\mathbf{x}, \hat{\psi}) = \frac{1}{M} \sum_{j=1}^{M} \pi(\theta|\mathbf{x}, \hat{\psi}, \lambda^{(j)}). \tag{4}$$

Our fundamental concern is to understand in what sense we can consider $\hat{\pi}(\theta|\mathbf{x}, \hat{\psi})$ to be an estimate of $\pi(\theta|\mathbf{x}, \psi)$.

The classic results on consistency of Bayes estimators date back to Doob (1949), and are given a rigorous treatment by Schwartz (1965) and Diaconis and Freedman (1986); see also Schervisch (1995, Section 7.4.1). The basic theme of these results is that as the amount of data increases without bound, the posterior distribution tends to a point mass at the true value of θ. In the situation that we are considering, these results are not exactly what we want, as our situation is closer to an empirical Bayes model. The results of Datta (1991) are more in the spirit of the present model, but again are not exactly applicable. However, the structure that we assume, especially the underlying likelihood estimation, allows a fairly straightforward development of the needed theory.

Starting from the generic hierarchy (2), the posterior distributions of interest are $\pi(\theta_i|\mathbf{x}, \psi)$, for $i = 1, \ldots, p$. The Gibbs sampler estimates these densities with the average of the conditional densities as long as these latter densities are known in closed form. With an estimated hyperparameter $\hat{\psi}$, the MLE of ψ under the marginal distribution $m(\mathbf{x}|\psi)$, the development in Appendix A.1 gives conditions under

which, for any measurable set A, we have for each $i = 1, 2, \ldots, p$,

$$\int_A \left| \frac{1}{M} \sum_{j=1}^{M} \pi(\theta_i | \mathbf{x}, \hat{\psi}, \lambda^{(j)}) - \pi(\theta_i | \mathbf{x}, \psi) \right| \mathrm{d}\theta_i \to 0, \tag{5}$$

as $M, p \to \infty$.

3. Implementation

The results of the previous section give us some assurance that the inferences we draw from the estimated posterior will be reasonable. Since we are using maximum likelihood methods to estimate ψ, and we are basing our inferences on the marginal distribution of $X_1, \ldots, X_p$, we have available the 'machinery' of likelihood to help us make these inferences. However, before getting to this point, we first address a point that, in practice, could be a problem.

To use a marginal MLE to estimate the hyperparameter ψ requires computation of the marginal likelihood function, which could require a high-dimensional integration. (For example, a reasonable hierarchical model for the famous salamander data of McCullagh and Nelder (1989) can result in the need to evaluate six intractable 20-dimensional integrals to evaluate the likelihood function: see Karim and Zeger (1992) and Hobert (2000).) Moreover, we note that one of the strengths of the Gibbs sampler is that it avoids having to use high-dimensional integration to compute marginals, so it is counterproductive to re-introduce such a calculation. Fortunately, for the structure induced by the Gibbs sampler, and for calculation of a marginal MLE, there is an EM algorithm that is virtually automatic to implement.

For the generic hierarchy (2), notice that the marginal likelihood for ψ can be written as

$$L(\psi | \mathbf{x}) = \frac{L(\psi | \mathbf{x}, \boldsymbol{\theta}, \lambda)}{\pi(\boldsymbol{\theta}, \lambda | \mathbf{x}, \psi)} \tag{6}$$

where $L(\psi | \mathbf{x}, \boldsymbol{\theta}, \lambda)$ is the conditional likelihood of ψ and $\pi(\boldsymbol{\theta}, \lambda | \mathbf{x}, \psi)$ is the posterior distribution of $(\boldsymbol{\theta}, \lambda)$ given ψ. By taking logs, this expression for $L(\psi | \mathbf{x})$ leads to the identity

$$E[\log L(\psi | \mathbf{x}) | \psi_0] = E[\log L(\psi | \mathbf{x}, \boldsymbol{\theta}, \lambda) | \psi_0] - E[\log \pi(\boldsymbol{\theta}, \lambda | \mathbf{x}, \psi) | \psi_0] \tag{7}$$

where the expectation is taken with respect to $\pi(\boldsymbol{\theta}, \lambda | \mathbf{x}, \psi_0)$. Equation (7) is the basic identity on which the EM algorithm is built, and the sequence

$$\psi^{(k+1)} = \mathrm{argmax}_{\psi} E[\log L(\psi | \mathbf{x}, \boldsymbol{\theta}, \lambda) | \psi^{(k)}]$$

converges to the MLE of ψ. Rather than actually computing this expectation, we will instead use the Monte Carlo version of the EM algorithm, which has iterations

$$\psi^{(k+1)} = \mathrm{argmax}_{\psi} \frac{1}{M} \sum_{j=1}^{M} \log L(\psi | \mathbf{x}, \boldsymbol{\theta}^{(j)}, \lambda^{(j)}). \tag{8}$$

As noted, calculation of the conditional likelihood on the right side is straightforward. However, what makes this EM algorithm automatic is that the expectation is taken with respect to the distribution $\pi(\boldsymbol{\theta}, \lambda | \mathbf{x}, \psi^{(k)})$, which is exactly the output from the *original* Gibbs sampler. That is, if we specify that $\psi = \psi^{(k)}$, the original Gibbs sampler produces a sample from the distribution that we want. So we have the following algorithm to produce the empirical Bayes Gibbs sampler posterior estimate.

Algorithm For the generic hierarchy (2)

1. Set $k = 0$ and initialize $\psi^{(0)}$.
2. Generate a sample $(\boldsymbol{\theta}^{(j)}, \lambda^{(j)})$, $j = 1, \ldots, M$ from the Gibbs sampler which iterates between $\pi(\boldsymbol{\theta}|\mathbf{x}, \lambda, \psi^{(k)})$ and $\pi(\lambda|\mathbf{x}, \boldsymbol{\theta}, \psi^{(k)})$.
3. Update $\psi^{(k)}$ with the Monte Carlo EM iteration (8) and return to 2.
4. At convergence of $\psi^{(k)}$ to the marginal MLE $\hat{\psi}$, produce a final Gibbs sample from $\pi(\boldsymbol{\theta}|\mathbf{x}, \lambda, \hat{\psi})$ and $\pi(\lambda|\mathbf{x}, \boldsymbol{\theta}, \hat{\psi})$.

The final Gibbs sample can then be used to construct posterior estimates such as (4). We can, therefore, produce the empirical Bayes Gibbs sample merely by looping on the original Gibbs sampler, and no integrations are required for calculation of the marginal likelihood estimates. The only additional computation is the maximization in the EM algorithm.

4. PRACTICE

To illustrate the use of the algorithm of Section 3, we look at the ulcer data of Section 1. Based on the hierarchical model (1), we implement the Gibbs sampler using the full conditional posteriors

$$\begin{aligned} \pi(p_{1j}|p_{2j}) &\propto p_{1j}^{y_{1j}}(1-p_{1j})^{n_{1j}-y_{1j}+a-1} p_{2j}^{b-1}(p_{1j}-p_{2j})^{c-1} \\ \pi(p_{2j}|p_{1j}) &\propto p_{2j}^{y_{1j}+b-1}(1-p_{2j})^{n_{2j}-y_{2j}}(p_{1j}-p_{2j})^{c-1} \\ & j = 1, 2, \ldots, 39. \end{aligned} \tag{9}$$

To simplify the model for the hyperparameters, we assume that $a = b$, so we only fit two hyperparameters. With this parametrization, we find that increasing the hyperparameter a will increase our estimate of the underlying success probabilities, while increasing c will decrease the underlying correlation. Looking at the moment formulae in Appendix A.2 will make this clear.

The output from the Gibbs sampler and the EM algorithm are samples of pairs (p_{1j}, p_{2j}), $j = 1, 2, \ldots, 39$ from the joint posterior distribution. For this hierarchy we can also write down the form of the joint posterior. Thus, to display the results of the analysis we have a number of choices—we can display joint or marginal posterior distributions.

Another possibility is to display a transformation of the posterior distribution. Since we are interested in comparing p_{1j} to p_{2j}, we have chosen to transform the bivariate posterior to display the logodds. This entails a bivariate transformation and a marginalization. Although the bivariate transformation is tedious, it can be done analytically. (We have done it using Mathematica (Wolfram, 1998).) However, the marginalization is not a tractable integral, and it must be accomplished using a Monte Carlo marginalization. Thus (suppressing j), for $t = \log(\frac{p1}{1-p1}/\frac{p2}{1-p2})$ the final posterior that we display is

$$\begin{aligned} \pi(t|\mathbf{y}, \hat{a}, \hat{c}) &= \int \pi(t, p_2|\mathbf{y}, \hat{a}, \hat{c})\, \mathrm{d}p_2 \\ &\approx \frac{1}{M}\sum_{i=1}^{M} \frac{\pi(t, p_2^{(i)}|\mathbf{y}, \hat{a}, \hat{c})}{\pi(t^{(i)}, p_2^{(i)}|\mathbf{y}, \hat{a}, \hat{c})} w(t^{(i)}) \end{aligned}$$

where $(t^{(i)}, p_2^{(i)})$, $i = 1, 2, \ldots, M$ are the output from the Gibbs sampler and $w(\cdot)$ is an arbitrary density on $(0, \infty)$. (See Appendix A.2 for details.)

Figure 1 shows the posterior distributions of the logodds from two different studies along with an envelope of densities corresponding to the values of the hyperparameters that fall within one standard deviation of the estimated values. The envelope gives some idea of the sensitivity of the posterior to the estimated values of the hyperparameters, and, in particular, gives an idea of how sensitive are the endpoints

of a confidence interval. Study 8 has fewer observations than Study 21, which results in a tighter envelope for the latter study.

To calculate the standard errors, we have implemented an identity due to Oakes (1999). The identity as presented by Oakes is more suitable to regular EM rather than Monte Carlo EM, but with a little modification it becomes easily implementable for Monte Carlo EM. See Appendix A.3 for details.

It should be mentioned that the model we are considering is rather simplistic, as we would expect to use covariates to further model the success probabilities. For example, the success probabilities should certainly be a function of both patient and medical center covariates. The methodology that we describe here can be adapted to these more complicated models.

5. Discussion

We have described a combination of Bayesian and frequentist methodologies, together with some Monte Carlo computing algorithms. Naturally, there are concerns, with perhaps the most important ones relating to the validity of the statistical inference. In this final section we address a variety of issues.

5.1 *Inference*

The methodology discussed here, which is empirical Bayes in nature, leads naturally to an inference that is a combination of Bayesian and frequentist inference. For example, Figure 1 can be seen as a Bayesian posterior density with a frequentist confidence set around it. Although some might find this unsettling, it merely reflects the inference that can be done. That is, for the parameters that can be modeled with a prior distribution, a posterior distribution is calculated. For those parameters without a prior distribution, frequentist inference (through likelihood) becomes the only option. One resulting inference is a range of Bayes posterior distributions, where the range reflects the frequentist uncertainty in the hyperparameter.

Such an inference is in the spirit of a robust Bayes analysis (Berger, 1990, 1994; Wasserman, 1990), but is really fundamentally different. One goal of robust Bayes analysis is to find a class of priors for which the resulting inference is 'robust' in that it does not quantitatively change as the prior varies through the class. This is not our goal. Rather, we use the data to estimate the prior, within a class, that provides the best fit (in the sense of maximum likelihood). Then, rather than assessing the robustness of the inference to the class of priors, we present a diagnostic for assessing the robustness of the inference over a range of likely values of the estimated parameter.

With a generic hierarchy such as (2), the deepest point in the hierarchy is the point where the inference shifts from Bayesian to frequentist. That is, as we go down the hierarchy, every parameter has a prior until we get to ψ. So the inference on ψ is frequentist. The method and algorithm of Section 3 details the EM/Gibbs sampler that results in consistent estimates of posterior distributions and asymptotically valid confidence sets.

From a graph such as Figure 1, one can also attach a Bayesian inference to the envelope of densities. For example, for a fixed value of ψ, we can construct a credible region for θ, C_ψ. If we choose C_ψ so that $P(\theta \in C_\psi|\psi, \mathbf{x}) = 1 - \gamma_1$, and ψ varies in a set S, a Bayesian inference from this setup is

$$P(\theta \in C_\psi, \psi \in S|\mathbf{x}) = \int_S P(\theta \in C_\psi|\psi, \mathbf{x})\pi(\psi|\mathbf{x})\,\mathrm{d}\psi = (1-\gamma_1)P(\psi \in S|\mathbf{x}),$$

where $\pi(\psi|\mathbf{x})$ is the posterior distribution of ψ. Moreover, we could assume that there is some 'probability matching' prior for ψ. That is, there is a prior distribution for which the Bayesian and frequentist coverage probabilities agree to some high order (see, for example, Datta and Ghosh (1995) or the review paper of Kass and Wasserman (1996)). If we construct an approximate $1 - \gamma_2$ confidence set for ψ using this

methodology, we then have the approximation

$$P(\theta \in C_\psi, \psi \in S|\mathbf{x}) \approx (1-\gamma_1)(1-\gamma_2).$$

Alternatively, one can use Laplace approximations (Tierney and Kadane, 1986) to obtain approximations to these probabilities.

5.2 *Computation*

The actual implementation of the methodology, and the inference, is almost automatic, and uses only the standard tools of the Gibbs sampler. Let us first note that although we have used examples in which the Gibbs sampler was a bivariate sampler, nothing precludes the method from being using on a more complex Gibbs structure.

There are a number of ways in which the computations described here can be improved, some of which are rather easy. For example, to implement the Monte Carlo EM algorithm of Section 3, there is no need to re-run the Gibbs sampler at each update of the EM sequence. This is because the Monte Carlo expected log likelihood can be recalculated using importance sampling, a point that is noted by Booth and Hobert (1999). To see this, recall the Monte Carlo EM iteration (8). In the calculation of the average log likelihood at the kth step of the sequence, we generated a sample $(\boldsymbol{\theta}_k^{(j)}, \lambda_k^{(j)})$, $j = 1, \ldots, M$ from $\pi(\boldsymbol{\theta}, \lambda|\mathbf{x}, \psi^{(k)})$. (Here we use a subscript k for the EM iteration.) If instead of having a sample for this distribution, suppose that we continually reuse the first generated sample $(\boldsymbol{\theta}_0^{(j)}, \lambda_0^{(j)})$, $j = 1, \ldots, M$ from $\pi(\boldsymbol{\theta}, \lambda|\mathbf{x}, \psi^{(0)})$. We then have the importance sampling approximation

$$\frac{1}{M}\sum_{j=1}^{M} \log L(\psi|\mathbf{x}, \boldsymbol{\theta}_k^{(j)}, \lambda_k^{(j)}) \approx \frac{1}{M}\sum_{j=1}^{M} \log L(\psi|\mathbf{x}, \boldsymbol{\theta}_0^{(j)}, \lambda_0^{(j)}) \frac{\pi(\boldsymbol{\theta}_0^{(j)}, \lambda_0^{(j)}|\mathbf{x}, \psi^{(k)})}{\pi(\boldsymbol{\theta}_0^{(j)}, \lambda_0^{(j)}|\mathbf{x}, \psi^{(0)})}. \tag{10}$$

One difficulty remains in using (10), in that in the Gibbs sampler we do not know the form of $\pi(\boldsymbol{\theta}, \lambda|\mathbf{x}, \psi)$. However, from (6), we see that $\pi(\boldsymbol{\theta}, \lambda|\mathbf{x}, \psi) = L(\psi|\mathbf{x}, \boldsymbol{\theta}, \lambda)/L(\psi|\mathbf{x})$, where $L(\psi|\mathbf{x}, \boldsymbol{\theta}, \lambda)$ is merely the product of the original densities in the hierarchy. But most importantly, $L(\psi|\mathbf{x})$ plays no role in the maximization of (10) as it only depends of the parameters $\psi^{(0)}$ and $\psi^{(k)}$. Hence in the maximization of the EM sequence we use

$$\operatorname{argmax}_\psi \frac{1}{M}\sum_{j=1}^{M} \log L(\psi|\mathbf{x}, \boldsymbol{\theta}_k^{(j)}, \lambda_k^{(j)}) \approx \operatorname{argmax}_\psi \frac{1}{M}\sum_{j=1}^{M} \log L(\psi|\mathbf{x}, \boldsymbol{\theta}_0^{(j)}, \lambda_0^{(j)}) \frac{L(\psi^{(k)}|\mathbf{x}, \boldsymbol{\theta}_0^{(j)}, \lambda_0^{(j)})}{L(\psi^{(0)}|\mathbf{x}, \boldsymbol{\theta}_0^{(j)}, \lambda_0^{(j)})}. \tag{11}$$

This strategy can be refined to use periodic updates of the $(\boldsymbol{\theta}, \lambda)$ sample, which should improve the approximation in (11).

A further computing refinement is possible using methodology developed in Levine and Casella (2001), which builds on the work of Booth and Hobert (1999). Booth and Hobert describe an algorithm to help choose the size of the Monte Carlo sample used at each step of the EM algorithm, but need independent samples for their algorithm. Using regeneration ideas developed by Robert *et al.* (1998), Levine and Casella extended the algorithm to dependent samples, making it implementable for MCMC samplers.

5.3 *Alternatives*

There are different approaches to modeling and analysing data such as the ulcer data, and some may not agree with the approach taken here. One reviewer felt that the ulcer data is better analysed by

parametrizing it directly in terms of the logodds, and either fitting a fixed trial effect and a random treatment effect, or a full bivariate distribution on control and treatment response logodds. These are worthy alternatives, and in either case the EM/Gibbs algorithm can be used to fit the models.

Another alternative to the model considered here is the full Bayesian hierarchical model. In cases where all priors can be specified, this is the preferred model to use. Thus, the methodology presented here is in no way a 'competitor' to the full hierarchical set-up, but rather an option when the full hierarchy cannot be specified. In such cases, rather than using arbitrary 'flat' priors, one may consider estimation of the hyperparameters and the empirical Bayes solution.

When there are 'flat' or other 'noninformative' priors on the hyperparameter, the hierarchical Bayes approach can also, in many instances, provide a good alternative. However, some care must be taken in these situations as there are times when the implementation of a 'flat' prior can be problematic, and can lead (possibly without detection) to models with improper posterior distributions (Natarajan and McCulloch, 1995; Hobert and Casella, 1996). For example, in the generic hierarchy (2), we can adopt the approach of George *et al.* (1993, 1994). Starting from the conditional likelihood $L(\psi|\mathbf{x}, \boldsymbol{\theta}, \lambda) = f(\mathbf{x}|\boldsymbol{\theta}, \psi)\pi(\boldsymbol{\theta}|\lambda, \psi)g(\lambda|\psi)$, normalize L as $L^* = \int L \, \mathrm{d}\psi$. Then use the Gibbs sampler

$$\begin{aligned} \boldsymbol{\theta}^{(j+1)} &\sim \pi(\boldsymbol{\theta}|\mathbf{x}, \psi^{(j)}, \lambda^{(j)}) \\ \lambda^{(j+1)} &\sim g(\lambda|\mathbf{x}, \psi^{(j)}, \boldsymbol{\theta}^{(j+1)}) \\ \psi^{(j+1)} &\sim L^*(\psi|\mathbf{x}, \boldsymbol{\theta}^{(j+1)}, \lambda^{(j+1)}) \end{aligned} \tag{12}$$

to produce the marginal posterior densities $\pi(\boldsymbol{\theta}|\mathbf{x})$ and $g(\lambda|\mathbf{x})$.

There are a number of complications with this approach that make it less appealing than the EM/Gibbs approach. Firstly, $L^* = \int L \, \mathrm{d}\psi$ may not be finite. (This can, of course, be fixed by instead calculating $\int Lh(\psi) \, \mathrm{d}\psi$ for some prior h, and then we are back in the fully specified proper hierarchical model.) Secondly, generation of samples in (12) can be quite difficult, even if a prior h is used. Thirdly, the consistency results of Section 2 may not apply here, so the frequentist interpretation is less clear. (Choosing h to be a 'probability matching' prior, as Efron (1996) suggests, could alleviate this, but the computational difficulties still remain.)

Various other refinements remain to be explored for this methodology. Other than applying these methods to other MCMC schemes, two interesting paths are (i) exploring the effect of other estimates of the hyperparameters, especially robust estimates, and (ii) exploring the effect of optimizing (or in some way varying) the shape of the hyperparameter confidence set.

ACKNOWLEDGEMENTS

Thanks to Andy Rosalsky for helpful discussions, to the Editor for encouragement, and to two very careful referees for valuable comments. This research was supported by NSF grant DMS 9971586.

APPENDIX A

A.1 *Convergence of the posterior distribution*

We will work with the hierarchy

$$\begin{aligned} X_i &\sim f(x|\theta_i, \psi) \quad i = 1, 2, \ldots, p \\ \theta_i &\sim \pi(\theta|\lambda, \psi) \\ \lambda &\sim g(\lambda|\psi), \end{aligned} \tag{13}$$

where $m(x|\psi) = \int\int f(x|\theta, \psi)\pi(\theta|\lambda, \psi)g(\lambda|\psi)\,d\theta\,d\lambda$ is the marginal distribution.

For unknown ψ, the Gibbs sampling estimate of the posterior density of the kth component of θ, θ_k, is given by $\hat{\pi}(\theta_k|\mathbf{x}, \hat{\psi}) = \frac{1}{M}\sum_{j=1}^{M}\pi(\theta_k|\mathbf{x}, \hat{\psi}, \lambda^{(j)})$. For fixed k, we are concerned with the limiting behavior of this estimate as M and $p \to \infty$. We define

$$h_\psi(\psi') = \int \pi(\theta_k|\mathbf{x}, \lambda, \psi')g(\lambda|\mathbf{x}, \psi)\,d\lambda$$

$$\hat{h}_\psi(\psi') = \frac{1}{M}\sum_{i=1}^{M}\pi(\theta_k|\mathbf{x}, \lambda^{(i)}, \psi'), \text{ where } \lambda^{(i)} \sim g(\lambda|\mathbf{x}, \theta^{(i)}, \psi),$$

so the subscript refers to the parameter of the density of integration (or generation) and the argument refers to the parameter of the integrand (or summand). We want to show that if $\hat{\psi}_p \to \psi$ (as $p \to \infty$) almost everywhere (with respect to the density m), then $\hat{h}_{\hat{\psi}_p}(\hat{\psi}_p) \to h_\psi(\psi)$ (as $p \to \infty$) in the sense that for all measurable A,

$$\int_A |\hat{\pi}(\theta_k|\mathbf{x}, \hat{\psi}_p) - \pi(\theta_k|\mathbf{x}, \psi)|\,d\theta_k \to 0$$

almost everywhere as M and $p \to \infty$.

To ease notation a bit we have suppressed putting a subscript on the function $\hat{h}$ to show that it depends on M. The estimator $\hat{\psi}_p$ depends on p but not M. In the following lemma we will need $M \to \infty$ with p, and we define a sequence M_p such that $M_p \to \infty$ as $p \to \infty$. The details of this are contained in the proof of the following lemma.

LEMMA A.1 For the hierarchy (13), suppose that

(i) $\hat{\psi}_p \to \psi$ (as $p \to \infty$) almost everywhere (with respect to m);
(ii) $h_\psi(\psi')$ is continuous in both ψ and ψ';
(iii) $\hat{h}_\psi(\psi')$ is continuous in ψ' and stochastically equicontinuous in ψ: that is, given $\varepsilon > 0, \exists\delta > 0$ with $|\hat{h}_{\psi_1}(\psi') - \hat{h}_{\psi_2}(\psi')| < \varepsilon$ for all $|\psi_1 - \psi_2| < \delta$ except on a set with g-measure 0;
(iv) the Gibbs sampler produces an ergodic Markov chain.

Then there exists a sequence M_p, with $\lim_{p\to\infty} M_p = \infty$, for which

$$|\hat{h}_{\hat{\psi}_p}(\hat{\psi}_p) - h_\psi(\psi)| \to 0 \text{ as } p \to \infty \tag{14}$$

almost everywhere (with respect to the densities g and m).

Proof. Suppose that ψ_0 is the true value of ψ, and let $\varepsilon > 0$ be given. Use the triangle inequality to write

$$|\hat{h}_{\hat{\psi}_p}(\hat{\psi}_p) - h_{\psi_0}(\psi_0)| \leqslant |\hat{h}_{\hat{\psi}_p}(\hat{\psi}_p) - \hat{h}_{\psi_0}(\psi_0)| + |\hat{h}_{\psi_0}(\psi_0) - h_{\psi_0}(\psi_0)|, \tag{15}$$

where we have added $\pm\hat{h}_{\psi_0}(\psi_0)$. The second term on the right does not involve $\hat{\psi}_p$, and converges to 0 by the Ergodic theorem. Specifically, for each p, choose $M_p^{(1)}$ so that this term is less than $\varepsilon/3$.

To deal with the first term in (15) we write

$$|\hat{h}_{\hat{\psi}_p}(\hat{\psi}_p) - \hat{h}_{\psi_0}(\psi_0)| \leqslant |\hat{h}_{\hat{\psi}_p}(\hat{\psi}_p) - \hat{h}_{\hat{\psi}_p}(\psi_0)| + |\hat{h}_{\hat{\psi}_p}(\psi_0) - \hat{h}_{\psi_0}(\psi_0)|, \tag{16}$$

where we have added $\pm\hat{h}_{\hat{\psi}_p}(\psi_0)$. As $p \to \infty$, the first term on the right in (16) goes to zero by assumption (iii). Now, by Egorov's theorem, we can choose p large enough so that $|\hat{\psi}_p - \psi_0| < \delta$ except on a set with probability less than $\varepsilon/2$. Also using assumption (iii), we then find $M_p^{(2)}$ so that

$$|\hat{h}_{\psi'}(\psi_0) - \hat{h}_{\psi_0}(\psi_0)| \leqslant \varepsilon/3 \quad \text{for all } |\psi' - \psi_0| < \delta,$$

except on a set with g-measure less than $\varepsilon/2$, and it then follows that the second term is bounded outside this set by

$$|\hat{h}_{\hat{\psi}_p}(\psi_0) - \hat{h}_{\psi_0}(\psi_0)| \leqslant \sup_{\psi':|\psi'-\psi_0|<\delta} |\hat{h}_{\psi'}(\psi_0) - \hat{h}_{\psi_0}(\psi_0)| \leqslant \varepsilon/3.$$

Thus, for arbitrary $\varepsilon > 0$, we choose p large enough, and $M_p = \max\{M_p^{(1)}, M_p^{(2)}\}$. Then the left side of (15) is bounded by three terms, each of which can be made less than $\varepsilon/3$ except on a set with probability less than ε, so (14) is established.

THEOREM A.1 Under the conditions of Lemma A.1, for each measurable A, as M and $p \to \infty$,

$$\int_A \left| \frac{1}{M} \sum_{j=1}^{M} \pi(\theta_k|\mathbf{x}, \hat{\psi}_p, \lambda^{(j)}) - \pi(\theta_k|\mathbf{x}, \psi) \right| \mathrm{d}\theta_k \to 0, \tag{17}$$

almost everywhere (with respect to the densities g and m).

Proof. The proof is a direct consequence of Lemma A.1 and Scheffé's Lemma (Resnick, 1999, Lemma 8.2.1), which we state for completeness.

Scheffé's Lemma. Let $\{F, F_n, n \geqslant 1\}$ be probability distributions with densities $\{f, f_n, n \geqslant 1\}$. Then,

1. $\sup_{B\in\mathcal{B}(\Re)} |F_n(B) - F(B)| = \frac{1}{2}\int |f_n(x) - f(x)|\,\mathrm{d}x$.
2. If $f_n(x) \to f$ almost everywhere, then $\int |f_n(x) - f(x)|\,\mathrm{d}x \to 0$, and thus $F_n(x) \to F$ in total variation.

So, by Scheffé's Lemma, the convergence (14) implies (17).

REMARK In the proof of Lemma A.1, p and M_p depend on ψ_0 so this is an existence proof. The equicontinuity assumption it is not that restrictive in practice, as $\hat{h}$ is a function of the densities π and g of (13). The variables λ and ψ are typically continuous variables, with the density π being continuous in λ, and g being continuous in ψ.

A.2 *The bivariate Dirichlet*

Random variables X_1 and X_2 have the Dirichlet distribution with parameters (a, b, c) if they have density function

$$f(x_1, x_2) = \frac{\Gamma(a+b+c)}{\Gamma(a)\Gamma(b)\Gamma(c)} x_1^{a-1} x_2^{b-1} (1 - x_1 - x_2)^{c-1},$$
$$0 \leqslant x_1, x_2 \leqslant 1, \quad x_1 + x_2 \leqslant 1.$$

This distribution has the property that X_1 and X_2 are negatively correlated. For the ulcer data introduced in Section 1, a negative correlation between p_1 and p_2 does not make sense—we would expect the correlation within a medical center to be non-negative.

If we define $p_1 = 1 - x_1$ and $p_2 = x_2$, then

$$\pi(p_1, p_2) = \frac{\Gamma(a+b+c)}{\Gamma(a)\Gamma(b)\Gamma(c)} (1 - p_1)^{a-1} p_2^{b-1} (p_1 - p_2)^{c-1},$$
$$0 \leqslant p_1, p_2 \leqslant 1, \quad p_1 \geqslant p_2, \tag{18}$$

and now p_1 and p_2 have non-negative correlation. In fact

$$\mathrm{E}p_1 = \frac{b+c}{a+b+c} \qquad \mathrm{E}p_2 = \frac{b}{a+b+c}$$

$$\mathrm{Var}\, p_1 = \frac{a(b+c)}{(a+b+c)^2(a+b+c+1)} \qquad \mathrm{Var}\, p_2 = \frac{b(a+c)}{(a+b+c)^2(a+b+c+1)}$$

$$\mathrm{Cov}(p_1, p_2) = \frac{ab}{(a+b+c)^2(a+b+c+1)}.$$

It also happens that the standard restriction that the sum of the variables is less than 1 becomes the restriction that $p_{1j} \geqslant p_{2j}$: that is, that the treatment has a greater success probability than the control.

Although the restriction $p_1 \geqslant p_2$ seems very sensible, as one would believe that the new procedure is no worse than the old, one of the reviewers pointed out that the restriction that the new procedure is no worse than the old might be frowned on by regulatory agencies (among others). There are alternative priors that do not impose this restriction. Here is one that preserves the conjugate hierarchical structure. Referring to (18), we let our prior be $\pi(p_1, p_2)$ with probability α, $0 \leqslant \alpha \leqslant 1$, and let it be $\pi(p_2, p_1)$, with probability $1-\alpha$. This prior allows $p_1 \geqslant p_2$ and $p_2 \geqslant p_1$, while preserving the correlation structure. Moreover, $Z \sim$ Bernoulli(α), and conditional on Z, we have the same posterior distributions as in (9). Then we can either estimate α, or put a prior on it to compute the EM/Gibbs sampler.

For the Gibbs sampler used in (9), the full conditionals are of the form

$$\pi(p_1|p_2) \propto p_1^{y_1}(1-p_1)^{n_1-y_1+a-1}(p_1-p_2)^{c-1}, \quad p_1 \geqslant p_2$$

$$\pi(p_2|p_1) \propto p_2^{y2+b-1}(1-p_2)^{n_2-y_2}(p_1-p_2)^{c-1}, \quad p_2 \leqslant p_1.$$

An easy way to generate from these densities is to use an accept–reject algorithm with candidate random variables from a truncated beta density. In particular, for generating p_1:

1. Generate $W \sim$ beta$(y_1 + c, n_1 - y_1 + a)$. If $W > p_2$ go to step 2, otherwise generate another W.
2. Generate $U \sim$ Uniform$(0, 1)$. If

$$U < \left(\frac{W - p_2}{W * (1 - p_2)}\right)^{c-1}$$

accept $p_1 = W$, otherwise go to step 1.

Recall that in the accept–reject algorithm, to get an observation from the target density f by generating an observation from the candidate density g, the observation is accepted if $U < \frac{1}{M}\frac{f}{g}$, where $M = \sup \frac{f}{g}$. It is easy to show that

$$\frac{w^{y_1}(1-w)^{n_1-y_1+a-1}(w-p_2)^{c-1}}{w^{y_1+c-1}(1-w)^{n_1-y_1+a-1}} \leqslant (1-p_2)^{c-1}$$

for $w \geqslant p_2$, which yields the bound in the algorithm.

A similar algorithm is used for generating p_2, using a beta$(y_2 + b, n_2 - y_2 + c)$.

Calculating the joint distribution of the logodds can be done analytically, but the following Mathematica code can also be used.

```
f[x1_, x2_] := x1^y1 * (1 - x1)^n1-y1+a-1 x2^y2+b-1 * (1 - x2)^n2-y2 * (x1 - x2)^c-1
so := Solve[{v1 == Log[x1/(1 - x1) / x2/(1 - x2)], v2 == x2}, {x1, x2}]
g[v1_, v2_] = f[x1/.so[[1]], x2/.so[[2]]]
              *Abs[Det[Outer[D, First[{x1, x2}/.so], {v1, v2}]]]
```

A.3 *Standard errors*

To obtain error estimates for the marginal likelihood estimators, we adapt a formula derived by Oakes (1999). Using our notation, starting from (7), Oakes shows that

$$\frac{\partial^2}{\partial\psi^2}\log L(\psi|\mathbf{x}) = \left\{\frac{\partial^2}{\partial\psi'^2}E[\log L(\psi'|\mathbf{x},\boldsymbol{\theta},\lambda)|\psi] + \frac{\partial^2}{\partial\psi'\partial\psi}E[\log L(\psi'|\mathbf{x},\boldsymbol{\theta},\lambda)|\psi]\right\}\Bigg|_{\psi'=\psi}$$

where the expectation is taken with respect to $\pi(\boldsymbol{\theta},\lambda|\mathbf{x},\psi)$. If we now take the derivative inside the expectation, we can rewrite Oakes' identity as

$$\begin{aligned}\frac{\partial^2}{\partial\psi^2}\log L(\psi|\mathbf{x}) &= E\left(\frac{\partial^2}{\partial\psi^2}\log L(\psi|\mathbf{x},\boldsymbol{\theta},\lambda)\Bigg|\psi\right)\\ &\quad + E\left[\left(\frac{\partial}{\partial\psi}\log L(\psi|\mathbf{x},\boldsymbol{\theta},\lambda)\right)\left(\frac{\partial}{\partial\psi}\log\pi(\boldsymbol{\theta},\lambda|\mathbf{x},\psi)\right)\Bigg|\psi\right],\end{aligned} \tag{19}$$

which allows the Monte Carlo evaluation

$$\begin{aligned}\frac{\partial^2}{\partial\psi^2}\log L(\psi|\mathbf{x}) &= \frac{1}{M}\sum_{j=1}^{M}\frac{\partial^2}{\partial\psi^2}\log L(\psi|\mathbf{x},\boldsymbol{\theta}^{(j)},\lambda^{(j)})\\ &\quad + \frac{1}{M}\sum_{j=1}^{M}\left(\frac{\partial}{\partial\psi}\log L(\psi|\mathbf{x},\boldsymbol{\theta}^{(j)},\lambda^{(j)})\right)\left(\frac{\partial}{\partial\psi}\log\pi(\boldsymbol{\theta}^{(j)},\lambda^{(j)}|\mathbf{x},\psi)\right),\end{aligned}$$

where $(\boldsymbol{\theta}^{(j)},\lambda^{(j)})$, $j=1,\ldots,M$ are from the Gibbs sampler which iterates between $\pi(\boldsymbol{\theta}|\mathbf{x},\lambda,\psi)$ and $\pi(\lambda|\mathbf{x},\boldsymbol{\theta},\psi)$.

If the density $\pi(\boldsymbol{\theta},\lambda|\mathbf{x},\psi)$ is not available, we can further modify (19) by recalling that

$$\pi(\boldsymbol{\theta},\lambda|\mathbf{x},\psi) = \frac{L(\psi|\mathbf{x},\boldsymbol{\theta},\lambda)}{L(\psi|\mathbf{x})}.$$

If we take logs, differentiate, and use the fact (Oakes, 1999, equation 5) that

$$\frac{\partial}{\partial\psi}\log L(\psi|\mathbf{x}) = E\left(\frac{\partial}{\partial\psi}\log L(\psi|\mathbf{x},\boldsymbol{\theta},\lambda)\Bigg|\psi\right)$$

we can rewrite (19) as

$$\begin{aligned}\frac{\partial^2}{\partial\psi^2}\log L(\psi|\mathbf{x}) &= E\left(\frac{\partial^2}{\partial\psi^2}\log L(\psi|\mathbf{x},\boldsymbol{\theta},\lambda)\Bigg|\psi\right)\\ &\quad + E\left[\left(\frac{\partial}{\partial\psi}\log L(\psi|\mathbf{x},\boldsymbol{\theta},\lambda)\right)^2\Bigg|\psi\right] - \left[E\left(\frac{\partial}{\partial\psi}\log L(\psi|\mathbf{x},\boldsymbol{\theta},\lambda)\Bigg|\psi\right)\right]^2,\end{aligned}$$

which only uses the complete-data likelihood, and is readily evaluated with a Monte Carlo sum. We note that this last equation can be expressed in the rather pleasing form

$$\frac{\partial^2}{\partial\psi^2}\log L(\psi|\mathbf{x}) = E\left(\frac{\partial^2}{\partial\psi^2}\log L(\psi|\mathbf{x},\boldsymbol{\theta},\lambda)\Bigg|\psi\right) + \mathrm{var}\left(\frac{\partial}{\partial\psi}\log L(\psi|\mathbf{x},\boldsymbol{\theta},\lambda)\Bigg|\psi\right).$$

A.4 *The ulcer data*

Data for the 39 experiments (medical centers). Each set of observations is {successes, number of trials} for treatment and control, respectively.

Exp.	Trt	Control	Exp.	Trt	Control	Exp.	Trt	Control
1	{8, 15}	{2, 13}	14	{14, 21}	{20, 25}	27	{12, 17}	{10, 15}
2	{11, 19}	{8, 16}	15	{22, 25}	{21, 32}	28	{10, 10}	{2, 14}
3	{29, 34}	{35, 39}	16	{7, 11}	{4, 10}	29	{22, 22}	{16, 24}
4	{29, 36}	{27, 31}	17	{8, 10}	{2, 10}	30	{16, 18}	{11, 21}
5	{9, 12}	{12, 12}	18	{30, 31}	{23, 27}	31	{14, 15}	{6, 13}
6	{3, 7}	{0, 4}	19	{24, 28}	{16, 31}	32	{16, 24}	{12, 27}
7	{13, 17}	{11, 24}	20	{36, 43}	{27, 43}	33	{6, 12}	{2, 9}
8	{15, 16}	{3, 16}	21	{34, 40}	{8, 21}	34	{20, 20}	{18, 23}
9	{11, 14}	{15, 22}	22	{14, 18}	{34, 39}	35	{13, 17}	{14, 16}
10	{36, 38}	{20, 32}	23	{54, 68}	{61, 74}	36	{30, 40}	{8, 20}
11	{6, 12}	{0, 8}	24	{15, 21}	{13, 21}	37	{13, 16}	{14, 16}
12	{5, 7}	{2, 9}	25	{6, 6}	{0, 6}	38	{30, 34}	{14, 19}
13	{12, 21}	{17, 24}	26	{9, 10}	{10, 15}	39	{31, 38}	{22, 37}

REFERENCES

BERGER, J. O. (1984). The robust Bayesian viewpoint. In Kadane, J. (ed.), *Robustness of Bayesian Analysis*, Amsterdam: North-Holland.

BERGER, J. O. (1990). Robust Bayesian analysis: sensitivity to the prior. *Journal of Statistical Planning and Inference* **25**, 303–328.

BERGER, J. O. (1994). An overview of robust Bayesian analysis (with discussion). *Test* **3**, 5–124.

BERGER, J. O. AND PERRICHI, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association* **91**, 109–122.

BILLINGSLEY, P. (1995). *Probability and Measure*, 3rd edn. New York: Wiley.

BOOTH, J. G. AND HOBERT, J. P. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society,* Series B **61**, 265–285.

CARLIN, B. P. AND LOUIS, T. A. (1996). *Bayes and Empirical Bayes Methods for Data Analysis*. London: Chapman and Hall.

DATTA, S. (1991). On the consistency of posterior mixtures and its applications. *Annals of Statistics* **19**, 338–353.

DATTA, S. AND GHOSH, M. (1995). Some remarks on noninformative priors. *Journal of the American Statistical Association* **90**, 1357–1363.

DIACONIS, P. AND FREEDMAN, D. (1986). On the consistency of Bayes estimates. *Annals of Statistics* **14**, 1–26.

DOOB, J. L. (1949). Application of the theory of martingales. *Le Calcal des Probabilités et ses Applications, Colloques Internationaux du Centre National de la Recherche Scientifique* **13**, 23–27.

EFRON, B. (1996). Empirical Bayes methods for combining likelihood (with discussion). *Journal of the American Statistical Association* **91**, 538–565.

GELFAND, A. E. (1996). Comment on Efron's paper. *Journal of the American Statistical Association* **91**, 551–552.

GELFAND, A. E. AND SMITH, A. F. M. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**, 398–409.

GEMAN, S. AND GEMAN, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721–741.

GEORGE, E. I., MAKOV, U. AND SMITH, A. F. M. (1993). Conjugate likelihood distributions. *Scandinavian Journal of Statistics* **20**, 147–156.

GEORGE, E. I., MAKOV, U. AND SMITH, A. F. M. (1994). Fully Bayesian hierarchical analysis for exponential families via Monte Carlo computation. In Freedman, P. R. and Smith, A. F. M. (eds), *Aspects of Uncertainty*, New York: Wiley, pp. 181–198.

GOEL, P. AND DEGROOT, M. (1981). Information about hyperparameters in hierarchical models. *Journal of the American Statistical Association* **76**, 140–147.

HOBERT, J. P. (2000). Hierarchical models: a current computational perspective. *Journal of the American Statistical Association* **95**, 1312–1315.

HOBERT, J. P. AND CASELLA, G. (1996). The effect of improper priors on Gibbs sampling in hierarchical linear models. *Journal of the American Statistical Association* **91**, 1461–1473.

KARIM, M. R. AND ZEGER, S. L. (1992). Generalized linear models with random effects; salamander mating revisited. *Biometrics* **48**, 631–644.

KASS, R. E. AND STEFFEY, D. (1989). Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). *Journal of the American Statistical Association* **84**, 717–726.

KASS, R. E. AND WASSERMAN, L. (1996). The selection of prior distributions by formal rules (corr: 1998 V93 p. 412). *Journal of the American Statistical Association* **91**, 1343–1370.

LEVINE, R. A. AND CASELLA, G. (2001). Implementations of the Monte Carlo EM algorithm. *Journal of Computational and Graphical Statistics* **10**.

MCCULLAGH, P. AND NELDER, J. A. (1989). *Generalized Linear Models*, 2nd edn. London: Chapman and Hall.

MORRIS, C. N. (1983). Parametric empirical Bayes inference: theory and applications (with discussion). *Journal of the American Statistical Association* **78**, 47–65.

MORRIS, C. N. (1996). Discussion of Efron's paper. *Journal of the American Statistical Association* **91**, 555–558.

NATARAJAN, R. AND MCCULLOCH, C. E. (1995). A note on the existence of the posterior distribution for a class of mixed models for binomial responses. *Biometrika* **82**, 639–643.

OAKES, D. (1999). Direct calculation of the information matrix via the EM algorithm. *Journal of the Royal Statistical Society,* Series B **61**, 479–482.

O'HAGAN, A. (1995). Fractional Bayes factors for model comparison. *Journal of the Royal Statistical Society,* Series B **57**, 99–118.

RESNICK, S. I. (1999). *A Probability Path*. Boston: Birkhäuser.

ROBBINS, H. (1964). The empirical Bayes approach to statistical decision problems. *Annals of Mathematical Statistics* **35**, 1–20.

ROBBINS, H. (1983). Some thoughts on empirical Bayes estimation. *Annals of Statistics* **11**, 713–723.

ROBERT, C. P. AND CASELLA, G. (1999). *Monte Carlo Statistical Methods*. New York: Springer.

ROBERT, C. P., RYDÉN, T. AND TITTERINGTON, D. M. (1998). Convergence controls for MCMC algorithms, with applications to hidden Markov chains. Technical Report from MCMC preprint service `http://www.stats.bris.ac.uk/MCMC`.

SCHERVISCH, M. J. (1995). *Theory of Statistics*. New York: Springer.

SHIVELY, T. S., KOHN, R. AND WOOD, S. (1999). Variable selection and function estimation in additive nonparametric regression using a data-based prior (with discussion). *Journal of the American Statistical Association* **94**, 777–806.

SCHWARTZ, L. (1965). On Bayes procedures. *ZeitWahr* **4**, 10–26.

TIERNEY, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Annals of Statistics* **22**, 1701–1786.

TIERNEY, L. AND KADANE, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association* **81**, 82–86.

WASSERMAN, L. (1990). Recent methodological advances in robust Bayesian inference (with discussion). In Bernardo, J. M., Berger, J. O., Dawid, A. P. and Smith, A. F. M. (eds), *Bayesian Statistics* 4, Oxford University Press, pp. 483–502.

WASSERMAN, L. (2000). Asymptotic inference for mixture models using data-dependent priors. *Journal of the Royal Statistical Society,* Series B **62**, 159–180.

WOLFRAM, S. (1998). *Mathematica* 3.0. Champagne, Ill: Wolfram Media.

[*Received 30 July, 1999*; *first revision 5 January, 2000*; *second revision 20 July, 2000; accepted for publication 30 May, 2001*]