Estimation of Bayes Factors in a Class of Hierarchical Random Effects Models using a Geometrically Ergodic MCMC Algorithm

Hani Doss and James P. Hobert
Department of Statistics
University of Florida

Abstract

We consider a Bayesian random effects model that is commonly used in meta-analysis, in which the random effects have a t distribution, with degrees of freedom parameter to be estimated. We develop a Markov chain Monte Carlo algorithm for estimating the posterior distribution in this model, and establish geometric convergence of the algorithm. The geometric convergence rate has important theoretical and practical ramifications. Indeed, it implies that, under standard second moment conditions, the ergodic averages used to estimate posterior quantities of interest satisfy central limit theorems. Moreover, it guarantees the consistency of a batch means estimate of the asymptotic variance in the CLT, which in turn allows for the construction of asymptotically valid standard errors. We show how our Markov chain can be used, in conjunction with an importance sampling method, to carry out an empirical Bayes approach for estimating the degrees of freedom parameter. To illustrate our methodology we consider a meta-analysis of studies that link intake of non-steroidal anti-inflammatory drugs to a reduction in colon cancer risk, in which some of the studies are outliers. To model the distribution of the study effects we consider the family of t distributions, as well as a family of mixtures of Dirichlet process priors centered at the t distributions, and show how our methodology can be used to make a choice of model.

Key words and phrases: Dirichlet process, Empirical Bayes, importance sampling, meta-analysis

1 Introduction

Bayesian hierarchical models are often used to deal with random effects, and a commonly used model is the following:

$$\text{conditional on } \theta_i, \qquad Y_i \overset{\text{ind}}{\sim} \mathcal{N}(\theta_i, \sigma_i^2), \quad i = 1, \ldots, K, \tag{1.1a}$$

$$\text{conditional on } \mu, \tau, \qquad \theta_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \tau^2), \quad i = 1, \ldots, K, \tag{1.1b}$$

$$(\mu, \tau) \sim \nu_c. \tag{1.1c}$$

In (1.1a), Y_i is a single summary statistic from experiment i, based on a sample of size m_i, and θ_i is a latent parameter particular to that experiment. (Usually the variance is unknown, but as is commonly done in meta-analysis, we assume that experiment i gives an estimate $\hat{\sigma}_i^2$ that is accurate enough so that assuming $\hat{\sigma}_i^2$ equals the true value does not cause any problem.) The prior ν_c in (1.1c), indexed by the hyperparameter $c = (c_1, c_2, c_3, c_4)$, is the prior in which $\mu \sim \mathcal{N}(c_1, c_2)$, $1/\tau^2 \sim \text{Gam}(c_3, c_4)$, and μ and τ are independent. Typically one takes something like $c = (0, 1000, .1, .1)$ or anything giving a fairly diffuse prior.

Whereas the normality assumption in line (1.1a) is typically supported by some theoretical result, such as the asymptotic normality of maximum likelihood estimates, the normality assumption in line (1.1b) generally doesn't have any justification and is made solely for the sake of convenience. In certain situations, for example when study effects are highly disparate and one wishes to accommodate outliers, a good alternative to line (1.1b) is $\theta_i \overset{\text{iid}}{\sim} t_d(\mu, \tau^2)$, where $t_d(\mu, \tau^2)$ is the t distribution with d degrees of freedom, location μ, and scale τ. We will then want to select d, with the choice $d = \infty$ signifying the choice of the normal distribution. In all of the many previous uses of this model, the degrees of freedom parameter has been chosen in an ad-hoc manner [see, e.g., Smith et al. (1995) and Gelman et al. (2004, sect. 17.4), among many others]—with a decision on whether to use a t rather than a normal often based on whether or not inferences are different—and subsequent inference carried out without careful consideration of the validity of the inference.

Here we seek a principled approach for specifying the degrees of freedom parameter. Denote $Y = (Y_1, \dots, Y_K)^T$ and $\theta = (\theta_1, \dots, \theta_K)^T$. Let ρ_d be the prior density of (θ, μ, τ) specified by model (1.1) when we use a t distribution with d degrees of freedom instead of a normal in (1.1b), and let $\rho_{d,y}$ be the corresponding posterior given $Y = y$. The marginal density of Y is given by

$$m_d(y) = \iiint \ell_y(\theta)\rho_d(\theta, \mu, \tau)\, d\theta\, d\mu\, d\tau,$$

where $\ell_y(\theta)$ is the likelihood function specified in (1.1a). For convenience of notation, ρ_∞, $\rho_{\infty,y}$ and $m_\infty(y)$ will denote the corresponding quantities when a normal distribution is used in (1.1b). An empirical Bayes approach selects the value of d that maximizes m_d (we now suppress the dependence of $m_d(y)$ on y, since y is fixed throughout). In principle, the maximizing value of d can be obtained via the EM algorithm, by treating θ as missing. Implementation of the EM algorithm appears to present some nontrivial issues; but a more significant problem is that it gives only the maximizing value, whereas in the present situation it is also of interest to know the entire marginal likelihood function m_d. For example, in one of the two illustrations in Section 4, the likelihood at ∞ is nearly the same as the likelihood at the maximum (see Figure 2(A)), and this is useful to know, since we would not use a t model if we knew that a normal model is adequate.

To discuss estimation of m_d as d varies, define the Bayes factor for the model based on the t distribution with d degrees of freedom vs. the model based on the normal distribution by

$$B(d, \infty) = \frac{m_d}{m_\infty}. \tag{1.2}$$

Clearly the information about d in $B(d, \infty)$ is the same as the information in m_d, but as will be seen shortly, it is much easier to estimate $B(d, \infty)$ than it is to estimate m_d. Typical Markov chain Monte Carlo (MCMC) algorithms for dealing with model (1.1) give as output a sequence $(\theta^{(i)}, \mu^{(i)}, \tau^{(i)})$, $i = 1, \dots, n$ for which the marginal distribution is approximately the posterior distribution. If an ergodic chain is run under model (1.1)—with a normal distribution in (1.1b)—then $B(d,\infty)$ may be estimated very conveniently by

$$\hat{B}(d,\infty) = \frac{1}{n}\sum_{i=1}^{n} \frac{\rho_d(\theta^{(i)},\mu^{(i)},\tau^{(i)})}{\rho_\infty(\theta^{(i)},\mu^{(i)},\tau^{(i)})}. \tag{1.3}$$

Indeed, we have

$$\begin{aligned}
\hat{B}(d,\infty) &= \frac{m_d}{m_\infty}\frac{1}{n}\sum_{i=1}^{n} \frac{\ell_y(\theta^{(i)})\rho_d(\theta^{(i)},\mu^{(i)},\tau^{(i)})/m_d}{\ell_y(\theta^{(i)})\rho_\infty(\theta^{(i)},\mu^{(i)},\tau^{(i)})/m_\infty} \\
&= \frac{m_d}{m_\infty}\frac{1}{n}\sum_{i=1}^{n} \frac{\rho_{d,y}(\theta^{(i)},\mu^{(i)},\tau^{(i)})}{\rho_{\infty,y}(\theta^{(i)},\mu^{(i)},\tau^{(i)})} \\
&\xrightarrow{\text{a.s.}} \frac{m_d}{m_\infty}\iiint \frac{\rho_{d,y}(\theta,\mu,\tau)}{\rho_{\infty,y}(\theta,\mu,\tau)}\rho_{\infty,y}(\theta,\mu,\tau)\,d\theta\,d\mu\,d\tau \;=\; \frac{m_d}{m_\infty}.
\end{aligned} \tag{1.4}$$

When we present an estimate such as (1.3), it is important to also provide error margins for the estimate and unfortunately, this is rarely done with estimates produced by MCMC (Flegal et al. 2008); we return to this point in Section 5 of the present paper. Now, whereas the almost sure convergence in (1.4) results from simple ergodicity of the chain, a central limit theorem, which is required in order to produce asymptotically valid standard errors, requires further regularity conditions. To be specific, suppose that we want to estimate the posterior expectation of some function $g(\theta,\mu,\tau)$. The ergodic theorem implies that $(1/n)\sum_{i=1}^{n} g(\theta^{(i)},\mu^{(i)},\tau^{(i)})$ is a strongly consistent estimator. However, a central limit theorem for this estimator may not exist unless (i) the Markov chain mixes fast enough and (ii) the random variable $g(\theta,\mu,\tau)$ has enough moments (with respect to $\rho_{\infty,y}$). Various sets of conditions exist (Chan and Geyer 1994) but here we mention only that a typical condition is that the Markov chain is geometrically ergodic and that $g(\theta,\mu,\tau)$ has a finite moment of order $2+\epsilon$, for some $\epsilon > 0$.

Geometric ergodicity of the standard Markov chains for dealing with model (1.1), a fixed scan Gibbs and a block Gibbs sampler, was established by Hobert and Geyer (1998). Unfortunately, the moment condition is problematic: the random variable $\rho_d(\theta,\mu,\tau)/\rho_\infty(\theta,\mu,\tau)$ does not even have a finite second moment when $(\theta,\mu,\tau) \sim \rho_{\infty,y}$, because the tails of the

t distribution are much heavier than those of the normal. Thus, the standard chains yield a Bayes factor estimate which does not satisfy a CLT, and since there is no CLT, the target parameter that the usual methods of estimating standard errors (such as those based on batch means and regeneration) are designed to estimate is not even defined.

Suppose that instead we run a chain based on model (1.1), but with a t distribution with d_1 degrees of freedom ($d_1 < \infty$) in (1.1b). It is not hard to see that $\rho_d(\theta, \mu, \tau)/\rho_{d_1}(\theta, \mu, \tau)$ has moments of all orders when $(\theta, \mu, \tau) \sim \rho_{d_1,y}$, and so the estimate

$$\hat{B}(d, d_1) = \frac{1}{n} \sum_{i=1}^{n} \frac{\rho_d(\theta^{(i)}, \mu^{(i)}, \tau^{(i)})}{\rho_{d_1}(\theta^{(i)}, \mu^{(i)}, \tau^{(i)})} \tag{1.5}$$

will satisfy a central limit theorem if we can establish geometric ergodicity of the chain.

In this paper we present an efficient Markov chain algorithm for estimating the posterior distribution for model (1.1) with a t distribution in line (1.1b), and establish geometric ergodicity of the chain. The benefits of these results are that they enable us to deal with the preliminary model choice issue of selecting the degrees of freedom parameter; and once we have decided on the model and run a Markov chain for that model, the combination of a central limit theorem and a method for estimating the variance such as batching will allow us to get valid error margins for estimates we obtain in subsequent inference. Our development puts the use of this model on a firm footing, which is useful since this is one of the most commonly used models in applied Bayesian work.

There is a close correspondence between our random effects model and the one considered by Hobert and Geyer (1998). Consequently, the associated block Gibbs samplers are also similar, although there are some significant differences; e.g., our algorithm has three steps per iteration while theirs has only two. Due to the similarity between the two algorithms, we are able to exploit a few of the simple (exact) expectation calculations from Hobert and Geyer (1998) in the early stages of our proof of geometric ergodicity. However, the latter part of the proof contains new analysis.

This paper is organized as follows. In Section 2 we describe a model that is equivalent to the version of model (1.1) in which the distribution of the random effects is a t, and describe the Gibbs sampling algorithm. In Section 3 we state and prove a theorem that asserts that our Markov chain is geometrically ergodic, and discuss estimation of variability. Section 4 gives an illustration of the use of Bayes factors to select the model in a meta-analysis example. Section 5 contains closing remarks that include a discussion of results on geometric ergodicity in concrete models of statistical interest.

2 Models and a Gibbs Sampling Algorithm

2.1 A Hierarchical Model With t-Distributed Random Effects

Consider the following hierarchical model:

$$\text{conditional on } \theta, \mu, \lambda_\theta, \quad Y_i \overset{\text{ind}}{\sim} \mathcal{N}(\theta_i, \gamma_i^{-1}), \quad i = 1, \ldots, K, \tag{2.1a}$$

$$\text{conditional on } \mu, \lambda_\theta, \qquad \theta_i \overset{\text{iid}}{\sim} t_d(\mu, \lambda_\theta^{-1}), \quad i = 1, \ldots, K, \tag{2.1b}$$

$$\mu \sim \mathcal{N}(\mu_0, \lambda_0^{-1}) \qquad\qquad \lambda_\theta \sim \text{Gam}(\alpha, \beta), \tag{2.1c}$$

where at the bottom level μ and λ_θ are independent. In (2.1a) and (2.1b), $\gamma_i = 1/\sigma_i^2$ and $\lambda_\theta = 1/\tau^2$; it is much more convenient to work with the γ_i's and λ_θ. Note that (2.1) is the same as (1.1) except that we have a t distribution instead of a normal in the middle of the hierarchy. As mentioned earlier, we assume that experiment i gives an estimate $\hat{\gamma}_i$ that is for practical purposes equal to γ_i. This is commonly done and does not present a problem unless the individual studies involve very small samples (DuMouchel 1990), in which situation one would want to put prior distributions on the γ_i's. For definiteness, we note that by $X \sim \text{Gam}(\alpha, \beta)$ we mean X is a random variable supported on the positive half-line with density proportional to $x^{\alpha-1}e^{-x\beta}$, and by $X \sim t_d(\mu, \lambda_\theta^{-1})$ we mean that X is a random variable whose density is proportional to $\left[d + \lambda_\theta(x-\mu)^2\right]^{-(d+1)/2}$. Let π denote the prior density of $(\theta, \mu, \lambda_\theta)$ under this model, and let π_y denote the posterior density given $Y = y$.

Now consider a more complex hierarchical model given by

$$\text{conditional on } \theta, \lambda, \mu, \lambda_\theta, \quad Y_i \overset{\text{ind}}{\sim} \mathcal{N}(\theta_i, \gamma_i^{-1}), \quad i = 1, \ldots, K, \tag{2.2a}$$

$$\text{conditional on } \lambda, \mu, \lambda_\theta, \qquad \theta_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \lambda_\theta^{-1}\lambda_i^{-1}), \quad i = 1, \ldots, K, \tag{2.2b}$$

$$\mu \sim \mathcal{N}(\mu_0, \lambda_0^{-1}) \qquad \lambda_\theta \sim \text{Gam}(\alpha, \beta) \qquad \lambda_i \overset{\text{iid}}{\sim} \text{Gam}(d/2, d/2),\ i = 1, \ldots, K, \tag{2.2c}$$

where $\lambda = (\lambda_1, \ldots, \lambda_K)^T$, and at the bottom level everything is independent. For this model, let π^* and π^*_y denote the prior and posterior densities of $(\theta, \lambda, \mu, \lambda_\theta)$, respectively, and let $m^*(y)$ denote the marginal density of Y.

We will write $\pi_y(\theta, \mu, \lambda_\theta)$ and $\pi(\theta, \mu, \lambda_\theta \,|\, y)$ interchangeably, and we will slightly abuse notation and use π generically to denote various distributions under model (2.1), e.g. $\pi(\theta)$, $\pi(\mu)$, and $\pi(\lambda_\theta)$ will denote the marginal distributions of θ, μ, and λ_θ, respectively. A similar remark applies to π^*. Also, even though π, π_y, π^*, π^*_y and m^* all depend on d, this dependence is suppressed, since d is now fixed.

Clearly,

$$\pi^*(\theta, \lambda, \mu, \lambda_\theta \,|\, y) = \frac{f(y \,|\, \theta, \lambda, \mu, \lambda_\theta)\pi^*(\theta \,|\, \lambda, \mu, \lambda_\theta)\pi^*(\lambda)\pi^*(\mu)\pi^*(\lambda_\theta)}{m^*(y)}, \tag{2.3}$$

where $f(y \,|\, \theta, \lambda, \mu, \lambda_\theta)$ is the likelihood function given in (2.2a). Note that $f(y \,|\, \theta, \lambda, \mu, \lambda_\theta)$ does not depend on λ, μ, and λ_θ, and this is the function that was denoted $\ell_y(\theta)$ in Section 1. The posterior marginal density of $(\theta, \mu, \lambda_\theta)$ is $\pi^*(\theta, \mu, \lambda_\theta \,|\, y) = \int \pi^*(\theta, \lambda, \mu, \lambda_\theta \,|\, y)\, d\lambda$, and a calculation shows that $\int \pi^*(\theta, \lambda, \mu, \lambda_\theta \,|\, y)\, d\lambda = \pi(\theta, \mu, \lambda_\theta \,|\, y)$. (This is the standard calculation that shows that a t distribution may be viewed as a gamma mixture of scaled normals.) We conclude that $\pi^*(\theta, \mu, \lambda_\theta \,|\, y) = \pi(\theta, \mu, \lambda_\theta \,|\, y)$. Hence, if we can develop an MCMC algorithm with stationary density $\pi^*(\theta, \lambda, \mu, \lambda_\theta \,|\, y)$, then we can use this MCMC algorithm to study $\pi(\theta, \mu, \lambda_\theta \,|\, y)$—we just throw away the λ's.

We note that the prior in (2.1c) specifies that μ and λ_θ are independent. This is different from another "normal / inverse gamma" prior in which $\lambda_\theta \sim \text{Gam}(\alpha, \beta)$, and given λ_θ, $\mu \sim \mathcal{N}(\mu_0, \lambda_0^{-1}\lambda_\theta^{-1})$, which is often used because it is conjugate to the normal distribution with two parameters unknown (see e.g. Berger 1985, p. 288). In our situation, there is no reason to prefer this normal / inverse gamma prior.

2.2 Conditional Distributions and a Gibbs Sampler

Let $\zeta = (\theta_1, \ldots, \theta_K, \mu)^T$. We will consider a block Gibbs sampler whose components are λ_θ, λ and ζ. (In what is below, the data vector Y is fixed, and all distributions are conditional on $Y = y$, although this conditioning is not stated explicitly.) It is easy to show that

$$\text{given } \lambda, \zeta, \quad \lambda_\theta \sim \text{Gam}\Big(\frac{K}{2} + \alpha, \frac{1}{2}\sum_i \lambda_i(\theta_i - \mu)^2 + \beta\Big).$$

It is also easy to show that

$$\text{given } \lambda_\theta, \zeta, \quad \lambda_i \overset{\text{ind}}{\sim} \text{Gam}\Big(\frac{d+1}{2}, \frac{\lambda_\theta}{2}(\theta_i - \mu)^2 + \frac{d}{2}\Big).$$

Now, it follows from (2.3) that the conditional density of ζ given λ and λ_θ is proportional to

$$\exp\bigg\{-\frac{1}{2}\bigg[-2\sum_{i=1}^K \theta_i(\gamma_i y_i + \mu\lambda_\theta\lambda_i) + \sum_{i=1}^K \theta_i^2(\lambda_\theta\lambda_i + \gamma_i) + \mu^2\Big(\lambda_0 + \lambda_\theta\sum_{i=1}^K \lambda_i\Big) - 2\mu\lambda_0\mu_0\bigg]\bigg\}.$$

From this we deduce that given λ and λ_θ, ζ has a multivariate normal distribution. All we have to do is identify the mean vector and the covariance matrix. This unnormalized multivariate normal density is nearly identical to the corresponding density in Hobert and Geyer's (1998) block Gibbs sampler, and the calculations we now perform to identify the mean and covariance matrix follow theirs closely. First, the covariance matrix, V, satisfies

$$V^{-1} = \begin{pmatrix} D^2 & -\lambda_\theta\lambda \\ -\lambda_\theta\lambda^T & \lambda_0 + \lambda_\theta\sum_{i=1}^K \lambda_i \end{pmatrix},$$

in which D is a $K \times K$ diagonal matrix whose i-th diagonal element is $d_{ii} = \sqrt{\lambda_\theta \lambda_i + \gamma_i}$. The mean, ζ_0, is the solution to

$$V^{-1}\zeta_0 = (\gamma_1 y_1, \gamma_2 y_2, \ldots, \gamma_K y_K, \lambda_0 \mu_0)^T.$$

Let

$$t = \sum_{i=1}^{K} \frac{\gamma_i \lambda_\theta \lambda_i}{\lambda_\theta \lambda_i + \gamma_i} = \lambda_\theta \sum_{i=1}^{K} \lambda_i - \lambda_\theta^2 \sum_{i=1}^{K} \lambda_i^2 (\lambda_\theta \lambda_i + \gamma_i)^{-1}.$$

Write the Cholesky factorization of the precision matrix as $V^{-1} = LL^T$, where L is a lower-triangular matrix. The elements of L can be calculated "by hand" and it is straightforward to show that

$$L^{-1} = \begin{pmatrix} D^{-1} & 0 \\ \frac{c^T D^{-1}}{\sqrt{\lambda_0 + t}} & \frac{1}{\sqrt{\lambda_0 + t}} \end{pmatrix},$$

where c^T is a $1 \times K$ row vector whose i-th element is $\lambda_\theta \lambda_i / d_{ii}$. We can now easily compute V and ζ_0. Indeed, $V = (L^{-1})^T L^{-1}$ and

$$\zeta_0 = (L^{-1})^T L^{-1} (\gamma_1 y_1, \gamma_2 y_2, \ldots, \gamma_K y_K, \lambda_0 \mu_0)^T$$

To simulate from the conditional distribution of ζ given λ, λ_θ, draw a $(K+1)$-variate standard normal, Z, and take $(L^{-1})^T Z + \zeta_0$. We now have all the conditional distributions needed to implement a Gibbs sampler on $(\lambda_\theta, \lambda, \zeta)$.

Next we give the variances and covariances (that is, the elements of V), along with upper bounds that will be used later. We have

$$\begin{aligned}
\mathrm{Var}(\theta_i \,|\, \lambda_\theta, \lambda) &= \frac{1}{\lambda_\theta \lambda_i + \gamma_i} \left[1 + \frac{\lambda_\theta^2 \lambda_i^2}{(\lambda_\theta \lambda_i + \gamma_i)(\lambda_0 + t)} \right] \le \frac{1}{\gamma_i} + \frac{1}{\lambda_0}, \\
\mathrm{Cov}(\theta_i, \theta_j \,|\, \lambda_\theta, \lambda) &= \frac{\lambda_\theta^2 \lambda_i \lambda_j}{(\lambda_\theta \lambda_i + \gamma_i)(\lambda_\theta \lambda_j + \gamma_j)(\lambda_0 + t)} \le \frac{1}{\lambda_0}, \\
\mathrm{Cov}(\theta_i, \mu \,|\, \lambda_\theta, \lambda) &= \frac{\lambda_\theta \lambda_i}{(\lambda_\theta \lambda_i + \gamma_i)(\lambda_0 + t)} \le \frac{1}{\lambda_0}, \\
\mathrm{Var}(\mu \,|\, \lambda_\theta, \lambda) &= \frac{1}{\lambda_0 + t} \le \frac{1}{\lambda_0}.
\end{aligned}$$

Moreover,

$$
\begin{aligned}
E(\mu \,|\, \lambda_\theta, \lambda) &= \sum_{j=1}^{K} \gamma_j y_j \operatorname{Cov}(\theta_j, \mu \,|\, \lambda_\theta, \lambda) + \lambda_0 \mu_0 \operatorname{Var}(\mu \,|\, \lambda_\theta, \lambda) \\
&= \frac{1}{\lambda_0 + t} \left[\sum_{j=1}^{K} \frac{\gamma_j \lambda_\theta \lambda_j y_j}{\lambda_\theta \lambda_i + \gamma_j} + \mu_0 \lambda_0 \right].
\end{aligned}
$$

Note that $E(\mu \,|\, \lambda_\theta, \lambda)$ is a convex combination of the y_j's and μ_0. Thus, it is uniformly bounded by a constant. Also,

$$
\begin{aligned}
E(\theta_i \,|\, \lambda_\theta, \lambda) &= \sum_{j=1}^{K} \gamma_j y_j \operatorname{Cov}(\theta_i, \theta_j \,|\, \lambda_\theta, \lambda) + \lambda_0 \mu_0 \operatorname{Cov}(\theta_i, \mu \,|\, \lambda_\theta, \lambda) \\
&= \frac{\lambda_\theta \lambda_i}{\lambda_\theta \lambda_i + \gamma_i} \left[\frac{1}{\lambda_0 + t} \left[\sum_{j=1}^{K} \frac{\gamma_j \lambda_\theta \lambda_j y_j}{\lambda_\theta \lambda_j + \gamma_j} + \mu_0 \lambda_0 \right] \right] + \frac{\gamma_i y_i}{\lambda_\theta \lambda_i + \gamma_i}.
\end{aligned}
$$

This shows that $E(\theta_i \,|\, \lambda_\theta, \lambda)$ is a convex combination of $E(\mu \,|\, \lambda_\theta, \lambda)$ and y_i, so is also uniformly bounded by a constant.

3 Geometric Ergodicity and Valid Estimates of Variability

3.1 Geometric Ergodicity

Consider a block-Gibbs sampler on the state space

$$\mathcal{X} = \mathbb{R}_+^K \times \mathbb{R}_+ \times \mathbb{R}^{K+1}$$

that updates λ, then λ_θ, then ζ; that is, if we write the current state as $(\lambda', \lambda'_\theta, \zeta')$ and the next state as $(\lambda, \lambda_\theta, \zeta)$, then the Markov transition density is given by

$$k(\lambda, \lambda_\theta, \zeta \,|\, \lambda', \lambda'_\theta, \zeta') = \pi_y^*(\lambda \,|\, \lambda'_\theta, \zeta')\pi_y^*(\lambda_\theta \,|\, \lambda, \zeta')\pi_y^*(\zeta \,|\, \lambda, \lambda_\theta). \tag{3.1}$$

Let $K^m(x, \cdot)$ denote the m-step Markov transition distribution corresponding to (3.1), and let Π_y^* denote the distribution corresponding to π_y^*.

Harris Ergodicity of the chain governed by (3.1) is the condition that $\|K^m(x, \cdot) - \Pi_y^*(\cdot)\| \to 0$ for all $x \in \mathcal{X}$, where $\|\cdot\|$ denotes supremum over all Borel subsets of $\mathcal{X}$. This condition is guaranteed by the so-called "usual regularity conditions," namely that the chain has an invariant probability measure, is irreducible, aperiodic, and Harris recurrent; see Theorem 13.0.1 of Meyn and Tweedie (1993). These usual regularity conditions are typically easy to check; they are implied for example if the Markov transition function has a density which is everywhere positive, which is the case for (3.1). Geometric ergodicity is the much stronger condition that there exist a constant $c \in [0, 1)$ and a function $M : \mathcal{X} \to [0, \infty)$ such that for any $m \in \mathbb{N}$,

$$\|K^m(x, \cdot) - \Pi_y^*(\cdot)\| \leq M(x)c^m \quad \text{for all } x.$$

Theorem 1 *The chain driven by* (3.1) *is geometrically ergodic.*

Proof As in Hobert and Geyer (1998), we will prove that the Gibbs sampler is geometrically ergodic by finding a "drift function" $w : \mathbb{R}_+^K \times \mathbb{R}_+ \times \mathbb{R}^{K+1} \to \mathbb{R}_+$ that is *unbounded off compact sets* [see (3.2) below] and satisfies

$$E\big(w(\lambda, \lambda_\theta, \zeta) \,\big|\, \lambda', \lambda'_\theta, \zeta'\big) \leq \rho w(\lambda', \lambda'_\theta, \zeta') + L,$$

where $\rho \in (0, 1)$, $L \in \mathbb{R}$, and where the expectation is taken with respect to the transition density in (3.1).

We will need to calculate some conditional expectations with respect to the transition density in (3.1). We use "last" as a shorthand for the variables of the last iteration, i.e. $(\lambda', \lambda'_\theta, \zeta')$. Conditional expectations given "last" are computed iteratively as follows:

$$E[w(\lambda, \lambda_\theta, \zeta) \,|\, \text{last}] = E\Big\{E\big\{E[w(\lambda, \lambda_\theta, \zeta) \,|\, \lambda, \lambda_\theta] \,\big|\, \lambda, \text{last}\big\} \,\Big|\, \text{last}\Big\}.$$

Define the following functions:

$$w_1 = \frac{1}{\lambda_\theta}, \qquad w_2 = \sum_i \frac{1}{\sqrt{\lambda_i}}, \qquad w_3 = \sum_i (\theta_i - \mu)^2, \qquad w_4 = \sum_i \gamma_i (y_i - \theta_i)^2,$$

$$w_5 = e^{c\lambda_\theta}, \qquad w_6 = \sum_i e^{c\lambda_i}, \qquad w_7 = \sum_i \big[\lambda_\theta(\theta_i - \mu)^2 + d\big]^{\frac{1}{2}},$$

where c is a positive constant. Our drift function will take the form $w = \sum_{i=1}^7 A_i w_i$ where the A_i's are positive constants to be determined. By definition, the function w is unbounded off compact sets if

$$\text{the level set } \{(\lambda, \lambda_\theta, \zeta) : w(\lambda, \lambda_\theta, \zeta) \le T\} \text{ is compact for every } T > 0 \tag{3.2}$$

(see Meyn and Tweedie 1993, p. 191). However since w is continuous, to show (3.2) it is enough to show that $|\theta_i|$ is bounded for each i, $|\mu|$ is bounded, λ_θ is bounded away from both 0 and ∞, and the same is true for the λ_i's. Since $w_5 \to \infty$ as $\lambda_\theta \to \infty$ and $w_1 \to \infty$ as $\lambda_\theta \to 0$, we know that λ_θ is contained as specified. A similar argument involving w_6 and w_2 shows that the λ_i's are also contained. Since $w_4 \to \infty$ as $|\theta_i| \to \infty$, we have θ_i contained, and given that θ_i is contained, $w_3 \to \infty$ as $|\mu| \to \infty$ so μ is contained as well. We conclude that w is unbounded off compact sets.

We now start computing the required expectations. The terms w_5 and w_6 are easy to bound. Let $0 < c < \min\{\beta, d/2\}$ and note that

$$E\big(e^{c\lambda_\theta} \,\big|\, \lambda, \text{last}\big) = \left(\frac{\beta + \frac{1}{2}\sum_i \lambda_i(\theta_i' - \mu')^2}{\beta + \frac{1}{2}\sum_i \lambda_i(\theta_i' - \mu')^2 - c}\right)^{\alpha + K/2} \le \left(\frac{\beta}{\beta - c}\right)^{\alpha + K/2} = \text{const},$$

where "const" is a quantity that is independent of any variables (it does of course depend on K and the hyperparameters). Similarly,

$$E\big(e^{c\lambda_i} \,\big|\, \text{last}\big) = \left(\frac{d/2 + \frac{\lambda_\theta'}{2}(\theta_i' - \mu')^2}{d/2 + \frac{\lambda_\theta'}{2}(\theta_i' - \mu')^2 - c}\right)^{(d+1)/2} \le \left(\frac{d/2}{d/2 - c}\right)^{(d+1)/2} = \text{const}.$$

Hence, $E\big(e^{c\lambda_\theta} \,\big|\, \text{last}\big) \le \text{const}$ and $E\big(e^{c\lambda_i} \,\big|\, \text{last}\big) \le \text{const}$.

Using the inequalities stated at the end of Section 2.2, we have

$$E(w_3(\zeta) \,|\, \lambda_\theta, \lambda) = \sum_i \operatorname{Var}\big[(\theta_i - \mu) \,\big|\, \lambda_\theta, \lambda\big] + \sum_i \Big(E\big[(\theta_i - \mu) \,\big|\, \lambda_\theta, \lambda\big]\Big)^2 \le \text{const}.$$

Thus, $E(w_3(\zeta)\,|\,\lambda_\theta) \le$ const. Similarly, $E(w_4(\zeta)\,|\,\text{last}) \le$ const.

Now, as long as $K \ge 2$, $\frac{K}{2} + \alpha > 1$, and we have

$$E(w_1(\lambda_\theta)\,|\,\lambda, \text{last}) = \frac{2\beta + \sum_i \lambda_i(\theta_i' - \mu')^2}{K + 2\alpha - 2}.$$

Hence,

$$\begin{aligned}
E(w_1(\lambda_\theta)\,|\,\text{last}) &= \frac{2\beta + \sum_i (\theta_i' - \mu')^2 E(\lambda_i\,|\,\text{last})}{K + 2\alpha - 2} \\
&= \frac{2\beta + \sum_i \left[\frac{(d+1)(\theta_i'-\mu')^2}{\lambda_\theta'(\theta_i'-\mu')^2 + d}\right]}{K + 2\alpha - 2} \\
&\le \frac{2\beta}{K + 2\alpha - 2} + \frac{(d+1)}{d(K + 2\alpha - 2)} \sum_i (\theta_i' - \mu')^2 \\
&= \text{const} + \frac{(d+1)}{d(K + 2\alpha - 2)} w_3(\zeta').
\end{aligned}$$

Since $(d+1)/2 > 1/2$, we have

$$E\big(\lambda_i^{-1/2}\,\big|\,\text{last}\big) = \frac{\Gamma\big(\frac{d}{2}\big)}{\sqrt{2}\,\Gamma\big(\frac{d+1}{2}\big)} \big[\lambda_\theta'(\theta_i' - \mu')^2 + d\big]^{\frac{1}{2}},$$

so,

$$E(w_2(\lambda)\,|\,\text{last}) = \frac{\Gamma\big(\frac{d}{2}\big)}{\sqrt{2}\,\Gamma\big(\frac{d+1}{2}\big)} \sum_i \big[\lambda_\theta'(\theta_i' - \mu')^2 + d\big]^{\frac{1}{2}} = \frac{\Gamma\big(\frac{d}{2}\big)}{\sqrt{2}\,\Gamma\big(\frac{d+1}{2}\big)}\, w_7(\lambda_\theta', \zeta').$$

Jensen's inequality implies that

$$E\Big\{ \big[\lambda_\theta(\theta_i - \mu)^2 + d\big]^{\frac{1}{2}} \,\Big|\, \lambda, \lambda_\theta \Big\} \le \Big\{ \lambda_\theta E\big[(\theta_i - \mu)^2\,\big|\,\lambda, \lambda_\theta\big] + d \Big\}^{\frac{1}{2}} \le \big[a\lambda_\theta + d\big]^{\frac{1}{2}},$$

where a is a positive constant. A second application of Jensen's inequality gives

$$E\big(\sqrt{a\lambda_\theta + d}\,\big|\,\lambda, \text{last}\big) \le \big[aE(\lambda_\theta\,|\,\lambda, \text{last}) + d\big]^{\frac{1}{2}} = \left[\frac{a(K + 2\alpha)}{\sum_i \lambda_i(\theta_i' - \mu')^2 + 2\beta} + d\right]^{\frac{1}{2}} \le \text{const}.$$

Hence, $E(w_7(\lambda_\theta, \zeta)\,|\,\text{last}) \le$ const.

Now, choose any $\rho \in (0, 1)$ and define

$$A_1 = \rho \frac{d(K + 2\alpha - 2)}{(d+1)} \quad \text{and} \quad A_2 = \rho \frac{\sqrt{2}\,\Gamma\big(\frac{d+1}{2}\big)}{\Gamma\big(\frac{d}{2}\big)}.$$

Let $A_3 = A_4 = \cdots = A_7 = 1$. Putting everything together, we have

$$E\big[w(\lambda, \lambda_\theta, \zeta) \,\big|\, \lambda', \lambda'_\theta, \zeta'\big] \le \text{const} + \rho w_3(\zeta') + \rho w_7(\lambda'_\theta, \zeta') \le \text{const} + \rho w(\lambda', \lambda'_\theta, \zeta').$$

The chain is therefore geometrically ergodic. $\square$

Let $x_0, x_1, \ldots$ be a Markov chain driven by (3.1), let l be a real-valued function of x, and suppose we wish to form confidence intervals for the posterior expectation of $l(x)$. Suppose the chain is geometrically ergodic and there exists $\epsilon > 0$ such that $E\big(|l(x)|^{2+\epsilon}\big) < \infty$. Theorem 18.5.3 of Ibragimov and Linnik (1971) implies that, with $\mathrm{Var}(l(x_0))$ and $\mathrm{Cov}(l(x_0), l(x_j))$ calculated under the assumption that x_0 has the stationary distribution, the series

$$\kappa^2 = \mathrm{Var}(l(x_0)) + 2\sum_{j=1}^{\infty} \mathrm{Cov}(l(x_0), l(x_j)) \tag{3.3}$$

converges absolutely, and if $\kappa^2 > 0$, then with x_0 having an arbitrary distribution, the estimate $\bar{l}_n = (1/n)\sum_{i=0}^{n-1} l(x_i)$ satisfies

$$n^{1/2}\big(\bar{l}_n - E[l(x) \,|\, y]\big) \xrightarrow{\text{d}} \mathcal{N}(0, \kappa^2) \qquad \text{as } n \to \infty.$$

Construction of Valid Standard Error Estimates The existence of the CLT for $\bar{l}_n$ is important from a practical standpoint. Indeed, in conjunction with a consistent estimator of the asymptotic variance, it allows for the construction of a valid asymptotic standard error for $\bar{l}_n$. There are many ways to estimate the asymptotic variance. The standard methods are batch means (Jones et al. 2006), spectral methods (Geyer 1992), and regenerative simulation (Mykland et al. 1995). Each of these estimators is consistent under regularity conditions that include geometric ergodicity of the Markov chain. Batch means, which is the simplest of the three methods to implement, is the one we use in our examples.

4 Illustration: Meta-Analysis of Studies on Non-Steroidal Anti-Inflammatory Drugs and Risk of Colon Cancer

An important application of random effects models is in the area of meta-analysis. Here we study a body of literature that considers the effect of non-steroidal anti-inflammatory drugs (NSAIDs) on the risk of colon cancer, a subject currently of considerable interest and controversy in the medical literature (see, e.g. Iwama 2009).

Over the last 15 years, a large number of studies have investigated the relationship between use of NSAIDs and development of colon cancer, either at the epidemiological or at the cellular and molecular level, and several have strongly suggested that long-term use of NSAIDs significantly decreases the risk of colon cancer. But the studies have been inconsistent, with some suggesting a weak beneficial effect and one even suggesting a negative effect. Harris et al. (2005) gives a review of this work and discusses the epidemiological studies that have appeared in the medical literature. Each study reports a risk ratio for NSAIDs use vs. no NSAIDs use. This risk ratio is either simply an odds ratio obtained from a case-control study or an odds ratio based on a multiple logistic regression analysis that takes into account important risk factors for colon cancer.

It is not surprising that the studies give inconsistent results, since there is heterogeneity in the subject pools (characteristics such as age, ethnicity, and health status vary across the studies), and in the way the data were obtained (covariates to collect, statistical method to use, etc.). It is certainly of interest to carry out a meta-analysis of these studies, and because of the heterogeneity, it seems clear that the meta-analysis should be based on a random effects model. There have been some meta-analyses in the medical literature, but these were very informal: none have used a random effects model (all used fixed effects) and they have dealt with the conflicting conclusions in ad hoc ways, for example by simply throwing out studies with outlying results.

Table 1 of Harris et al. (2005) gives summary information on 21 studies that relate NSAIDs intake and risk of colon cancer. For each study, the following information is given:

- Observed risk ratio for NSAIDs use vs. no NSAIDs use.
- Confidence interval for true risk ratio (this is equivalent to giving the standard error for the estimate).
- Type of NSAID used.
- Dose of NSAID. This information is available for some, but not all, of the studies.

We will carry out two analyses of this data set, in order to illustrate how apparent non-normality and outliers affect the choice of model to be used, and the effect that model choice has on inference.

We follow convention and work on the log scale, because the normal approximation to the distribution of a study-specific observed odds ratio is better on that scale. The vertical lines in the left panel of Figure 1 give a visual description of the data. The locations on the x-axis are the observed log odds ratios for the 21 studies, and the heights of the lines are proportional to the reciprocals of the reported standard errors. Also given by the figure is an estimate of the distribution of the study-specific log odds ratios, using a kernel density estimate that is based on the observed log odds ratios, with weights that reflect the estimated standard errors. (This density estimate should be viewed with caution, since it is based on the estimated log odds ratios, and not the log odds ratios themselves.) We note that there are two left outliers and one right outlier, although whether or not these are significant enough to warrant using a t distribution remains to be seen.

The dose variable is available for all 15 studies for which the NSAID is aspirin, but is not available for any of the other studies. This variable turns out to be quite important, and none of the reviews in the medical literature have considered it. For each study j, let L_j,

ω_j, and x_j denote the observed log risk ratio, standard error, and dose, respectively, for that study. Let ξ_j denote the true log risk ratio, i.e. ξ_j is the log risk ratio that would be obtained if the sample sizes for study j were infinite. We consider a linear relationship between the latent variable ξ_j and the dose x_j, i.e. we write $\xi_j = \alpha_j + \theta_j x_j$. It is easy to see that $\alpha_j = 0$, since at dose 0 we must have $\xi_j = 0$. So if we start with the model $L_j \sim \mathcal{N}(\xi_j, \omega_j^2)$ (justified by standard asymptotic theory), we may write the model equivalently as $L_j/x_j \sim \mathcal{N}\big(\theta_j, (\omega_j/x_j)^2\big)$. Therefore, if we use a t distribution to model the distribution of the θ_j's, we are led to precisely (2.1), with $Y_j = L_j/x_j$ having standard deviation $\sigma_j = \omega_j/x_j$, and where θ_j plays the role of log risk ratio for study j if the dose for that study was equal to 1. The right panel of Figure 1 gives a visual display of the data for the 15 aspirin studies, with dose taken into account. The deviation from normality now appears stronger.

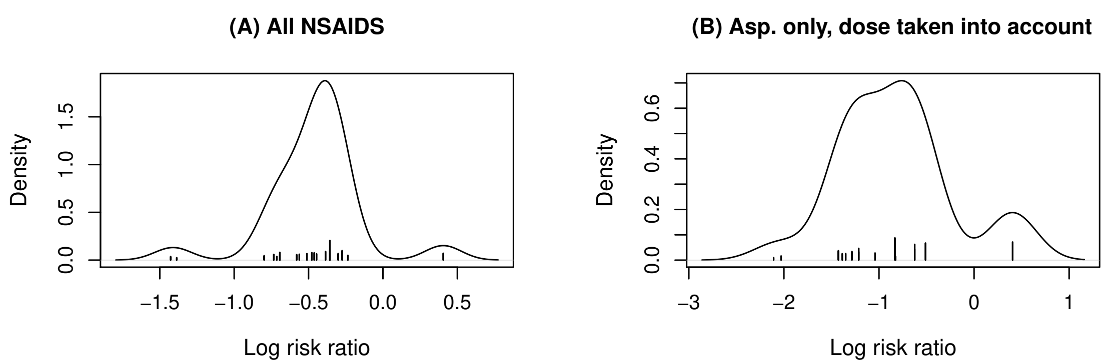

Figure 1: Estimate of distribution of the study-specific effect for two versions of the colon cancer data set. Left panel pertains to all 21 studies. Study results are represented by vertical lines, whose locations are the log risk ratios and whose heights are proportional to the reciprocals of the standard errors. Right panel pertains to the 14 aspirin studies, all of which include information on dose. The locations of the vertical lines are the log dose-adjusted risk ratios and the heights are proportional to the reciprocals of the standard errors.

We would like to determine whether for either version of the data set using a t distribution is warranted, and if so determine the degrees of freedom parameter. To this end, for each version of the data set, we ran a Markov chain of length $20{,}000$ based on model (2.1) with degrees of freedom parameter $d_1 = 4$ and a normal / inverse gamma prior at the bottom of the hierarchy. From this, we calculated $\hat{B}(d, d_1)$ for $d \in [.01, 20]$ and $d = \infty$, using formula (1.5). Figure 2 shows our results. The left panel is for all 21 studies (dose not taken into account), and the right panel is for the aspirin studies, with dose taken into account. In each case, the center line is actually a plot of $\hat{B}(d, d_1)/\hat{B}(\infty, d_1)$, i.e. an estimate of $B(d, \infty)$ (it's more natural to estimate $B(d, \infty)$ rather than $\hat{B}(d, d_1)$, i.e. take the normal distribution as the reference point). Also plotted are 95% confidence bands, obtained by the method of batching, using 20 batches (this choice is consistent with recommendations made by Flegal et al. (2008, sect. 3.1)). The confidence bands are valid pointwise. Although the left panel in Figure 1 suggests that for the group of all studies, dose not taken into account, outliers are present, the plot in the left panel of Figure 2 does not indicate that a t distribution is needed. The situation is different for the aspirin studies, where dose is taken into account. The right panel in Figure 2 suggests that one should use a t distribution, with about 3 degrees of freedom.

Before settling on the model based on the t_3 distribution for the aspirin studies, we consider a model in which the non-normality of the random effects is handled through a Dirichlet process prior. Specifically, we replace line (2.1b) with

$$\text{conditional on } F, \qquad \theta_i \overset{\text{iid}}{\sim} F, \quad i = 1, \dots, K, \tag{4.1a}$$

$$\text{conditional on } (\mu, \lambda_\theta), \quad F \sim \mathcal{D}_{M\, t_d(\mu, \lambda_\theta^{-1})}, \tag{4.1b}$$

and keep lines (2.1a) and (2.1c) the same. In (4.1a), $\mathcal{D}_{M\, t_d(d, \mu, \lambda_\theta^{-1})}$ denotes the Dirichlet process prior centered at the $t_d(\mu, \lambda_\theta^{-1})$ distribution, and with precision parameter M. For fixed values of M and d, the prior on F specified by lines (4.1b) and (2.1c) is centered at the two parameter family of t_d distributions, and the support of this prior is the set of all probability distributions on the real line. If M is very large, the model essentially reduces

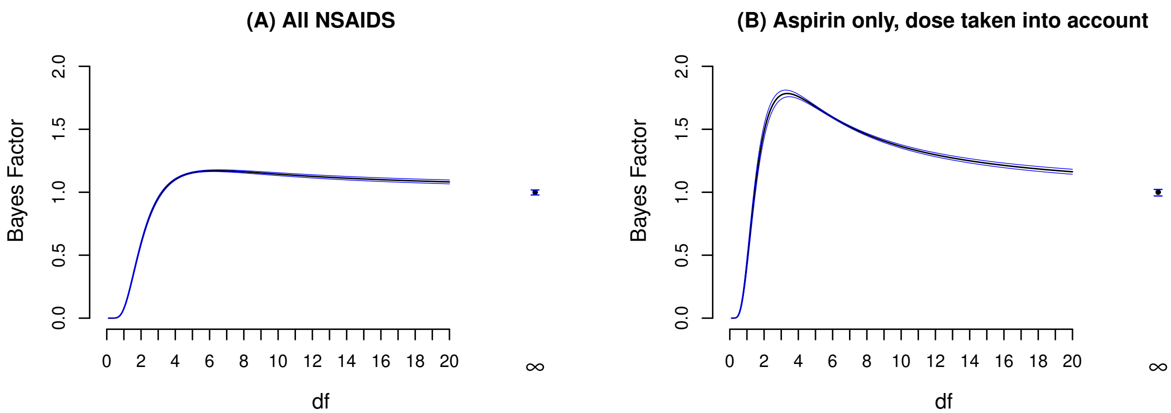

Figure 2: Estimates of Bayes factors, together with confidence bands, for the t vs. the normal distributions, for two versions of the colon cancer data. Left panel is for the group of all studies, dose not taken into account. The plot does not suggest that a t distribution is needed. Right panel is for the aspirin studies, with dose taken into account. The plot suggests a t distribution with about 3 degrees of freedom.

to model (2.1). See Antoniak (1974) for a discussion of mixtures of Dirichlet processes.

For this more general model, we need to determine both the precision parameter M and the degrees of freedom parameter d. Let $h = (M, d)$, and let m_h denote the marginal likelihood of the data under hyperparameter value h. To estimate h, we may proceed as we did before, i.e. fix a value $h_1 = (M_1, d_1)$ and define $B(h, h_1) = m_h/m_{h_1}$. Now let ρ_h be the distribution of $(\theta, \mu, \lambda_\theta)$ under the model specified by the mixture of Dirichlet processes. This distribution is not absolutely continuous with respect to Lebesgue measure on $\mathbb{R}^K \times \mathbb{R} \times \mathbb{R}_+$, because there is positive probability that the θ_i's are not all distinct, but for $h \neq h_1$, ρ_h is absolutely continuous with respect to ρ_{h_1}, and Doss (2009) obtains a formula for the Radon-Nikodym derivative $[d\rho_h/d\rho_{h_1}]$. Let $\theta_{(1)}, \dots, \theta_{(k)}$ denote the distinct values of $\theta_1, \dots, \theta_K$. When specialized to the present context, his formula is

$$\left[\frac{d\rho_h}{d\rho_{h_1}}\right](\theta, \mu, \lambda_\theta) = \left\{\prod_{r=1}^{k} \frac{t_d(\mu, \lambda_\theta^{-1})(\theta_{(r)})}{t_{d_1}(\mu, \lambda_\theta^{-1})(\theta_{(r)})}\right\} \left(\frac{M}{M_1}\right)^k \left\{\frac{\Gamma(M)\Gamma\big(M_1 + K\big)}{\Gamma(M_1)\Gamma(M + K)}\right\}. \qquad (4.2)$$

In (4.2), $t_d(\mu, \lambda_\theta^{-1})(\theta_{(r)})$ denotes the density of the $t_d(\mu, \lambda_\theta^{-1})$ distribution, evaluated at $\theta_{(r)}$, and Γ is the gamma function. The interpretation of this formula is as follows. Suppose $M > M_1$. If k, the number of distinct values in the vector $(\theta_1, \dots, \theta_K)$, is large, then the term $(M/M_1)^k$ is large. This is to be expected: The model for which the Dirichlet precision parameter is M_1 expected to see more ties, but didn't see them. So the model with Dirichlet precision parameter M better explains the data. The term in the first set of braces on the right side of (4.2) is the usual likelihood ratio of the t_d vs. the t_{d_1} distributions, except that it is based on only the distinct values of $\theta_1, \dots, \theta_K$. The term in the second set of braces on the right side of (4.2) does not involve $(\theta, \mu, \lambda_\theta)$, so is a constant that can be ignored.

To estimate $B(h, h_1)$ as $h = (M, d)$ varies, we proceed as we did before. Suppose that $(\theta^{(i)}, \mu^{(i)}, \lambda_\theta^{(i)})$, $i = 1, \dots, n$ is an ergodic Markov chain whose stationary distribution is the posterior distribution of $(\theta, \mu, \lambda_\theta)$ under the mixture of Dirichlet processes model. We form the estimate (1.5), except that the ratio of densities is replaced by the Radon-Nikodym derivative $[d\rho_h/d\rho_{h_1}]$. The estimate is valid, because the integral in (1.4) is still 1 if the ratio of densities is replaced by the Radon-Nikodym derivative, and the integration is with respect to the probability measure ρ_{h_1}. There are many algorithms for generating Markov chains for this model. We will not discuss these algorithms and instead refer to Jain and Neal (2007) for a review and recent developments. Using our own implementation, we ran a Markov chain of length $100{,}000$ under the model indexed by $M_1 = 3$ and $d_1 = 3$. Figure 3 gives plots of the Bayes factors as M varies, for four values of d, including $d = \infty$, i.e. the normal distribution. (The lines are actually plots of $\hat{B}(h, h_1)/\hat{B}((\infty, \infty), h_1)$, i.e. they are scaled so that we get 1 for the parametric model based on the normal distribution.) The analysis based on Figure 3 (and on plots for other values of d, not shown in Figure 3) suggests that there is no need for a model based on mixtures of Dirichlet processes, and that in fact the outliers are adequately accommodated by simply using a t_3 distribution.

We therefore ran Markov chains to estimate the posterior distributions under model (2.1) with a t_3 distribution in (2.1b) and also with a normal distribution in (2.1b), for the aspirin

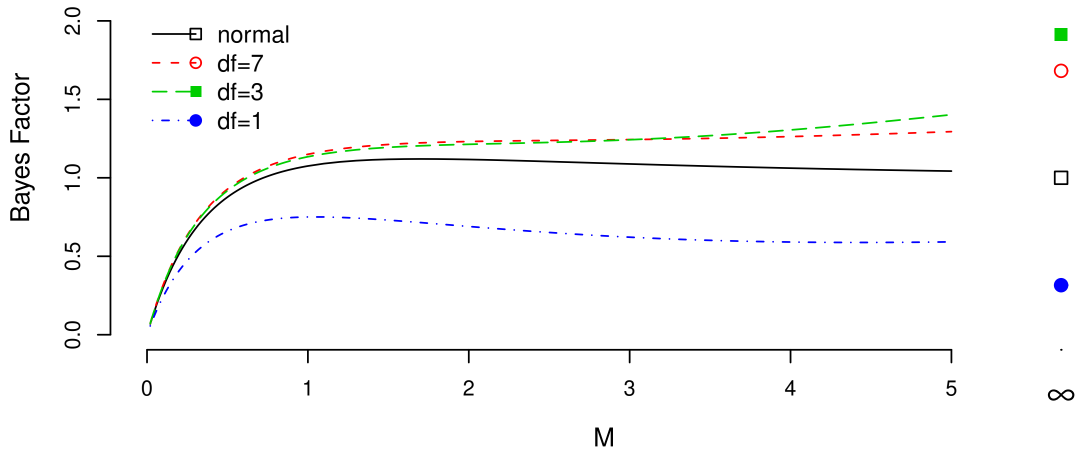

Figure 3: Model assessment for the aspirin and colon cancer data, when dose is taken into account. Shown are plots of Bayes factors for Dirichlet models centered at the location/scale families of normal and t distributions with 1, 3 and 7 degrees of freedom, as M varies. The Bayes factor is highest for the t_3 distribution, with $M = \infty$, which corresponds simply to a parametric model based on the t_3 distribution.

studies (dose taken into account). From the output, estimates of various quantities of interest can be obtained. In particular, we considered the predictive distribution of the log risk ratio for a future study (this quantity is of special interest in a meta-analysis that uses a random effects model, because in this case the unit is the study and not the individual in a study). Figure 4 shows the predictive distribution of the log risk ratio for a future study for the normal and t_3 models. As can be seen from the figure, the t_3 model gives stronger evidence of the effectiveness of aspirin: the distribution of θ_{new} is shifted to the left and also has significantly less spread. This is because the study for which the log risk ratio is $.4$ (see Figure 1 (B)) has less influence in the t_3 model. (When the influential study is removed, the two distributions are virtually identical. Also, for the full data set of 21 studies, where dose is not taken into account, the normal and t_d distributions for $d \geq 3$ all give essentially the same inferences, and this is consistent with the plot in Figure 2 (A)). With the Markov chain lengths of 10^5 that we used and the method of batching (using 40 batches), standard errors for all four quantities in the figure legend are less than $.004$.

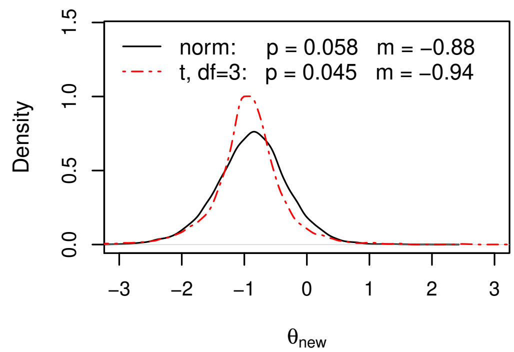

Figure 4: Distribution of θ_{new}, the log risk ratio for a future aspirin study, under the normal and t_3 models. Here, $p = P(\theta_{\text{new}} > 0)$ and m stands for mean.

5 Discussion

Before the MCMC revolution, when classical Monte Carlo methods based on iid samples were used to estimate intractable integrals, it would have been deemed unacceptable to report a Monte Carlo estimate without an accompanying asymptotic standard error (based on the CLT). Unfortunately, this seems to have changed with the advent of MCMC. In fact, it is actually uncommon to see an MCMC estimate accompanied by a standard error. Indeed, Flegal et al. (2008) examined all the articles published in the 2006 volumes of *Journal of the American Statistical Association*, *Biometrika* and *Journal of the Royal Statistical Society*, Series B. They found that MCMC methods had been used in 39 of the articles, and in only 3 of the 39 cases had the authors "directly addressed the Monte Carlo error in the reported estimates." This is due at least in part to the fact that it is harder, both theoretically and methodologically, to deal with the standard error problem in the MCMC context. Indeed, it is far more difficult to prove that MCMC-based estimates obey CLTs and, even when such a CLT is known to exist, finding consistent estimates of the variance in the CLT is not straightforward (see, e.g., Geyer 1992 and Jones et al. 2006). The cleanest way of establishing CLTs for MCMC-based estimators is to prove that the underlying Markov chain converges at a geometric rate (Chan and Geyer 1994; Roberts and Rosenthal 1997, 1998). Unfortunately, this is generally very difficult in the setting of continuous state spaces, and it is therefore not surprising that very few of the MCMC algorithms that are currently used in statistical applications are known to be geometrically ergodic. The exceptions include the MCMC algorithms studied by Mengersen and Tweedie (1996), Roberts and Tweedie (1996), Hobert and Geyer (1998), Roy and Hobert (2007), Marchev and Hobert (2004), Roberts and Rosenthal (1999), Jarner and Hansen (2000), and Papaspiliopoulos and Roberts (2008). (It should be noted that there is not a single result in the literature that gives geometric ergodicity for any of the Markov chains used to estimate the posterior distribution in mixture of Dirichlet process models such as the one we considered in Section 4. A comparison of the various chains through a theoretical evaluation is an interesting problem, but one we believe is difficult.)

Acknowledgements We thank two referees and an Associate Editor for helpful comments and suggestions. This work was supported by NSF Grant DMS-08-05860.

References

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics* **2** 1152–1174.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis (Second Edition)*. Springer-Verlag, New York.

Chan, K. S. and Geyer, C. J. (1994). Comment on "Markov chains for exploring posterior distributions". *The Annals of Statistics* **22** 1747–1758.

Doss, H. (2009). Hyperparameter and model selection for nonparametric Bayes problems via Radon-Nikodym derivatives. Tech. rep., Department of Statistics, University of Florida.

DuMouchel, W. (1990). Bayesian meta-analysis. In *Statistical Methodology in the Pharmaceutical Sciences*. Marcel Dekker (New York).

Flegal, J. M., Haran, M. and Jones, G. L. (2008). Markov chain Monte Carlo: Can we trust the third significant figure? *Statistical Science* **23** 250–260.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2004). *Bayesian Data Analysis (Second Edition)*. Chapman and Hall.

Geyer, C. J. (1992). Practical Markov chain Monte Carlo (with discussion). *Statistical Science* **7** 473–511.

Harris, R., Beebe-Donk, J., Doss, H. and Burr, D. (2005). Aspirin, Ibuprofen and other non-steroidal anti-inflammatory drugs in cancer prevention: A critical review of non-selective COX-2 blockade. *Oncology Reports* **13** 559–584.

Hobert, J. P. and Geyer, C. J. (1998). Geometric ergodicity of Gibbs and block Gibbs samplers for a hierarchical random effects model. *Journal of Multivariate Analysis* **67** 414–430.

Ibragimov, I. A. and Linnik, Y. V. (1971). *Independent and Stationary Sequences of Random Variables*. Wolters-Noordhoff, Groningen.

Iwama (2009). NSAIDs and colorectal cancer prevention. *Journal of Gastroenterology* **44** Suppl. 19 72–76.

Jain, S. and Neal, R. M. (2007). Splitting and merging components of a nonconjugate Dirichlet process mixture model. *Bayesian Analysis* **2** 445–472.

Jarner, S. F. and Hansen, E. (2000). Geometric ergodicity of Metropolis algorithms. *Stochastic Processes and their Applications* **85** 341–361.

Jones, G. L., Haran, M., Caffo, B. S. and Neath, R. (2006). Fixed-width output analysis for Markov chain Monte Carlo. *Journal of the American Statistical Association* **101** 1537–1547.

Marchev, D. and Hobert, J. P. (2004). Geometric ergodicity of van Dyk and Meng's algorithm for the multivariate Student's t model. *Journal of the American Statistical Association* **99** 228–238.

Mengersen, K. L. and Tweedie, R. L. (1996). Rates of convergence of the Hastings and Metropolis algorithms. *The Annals of Statistics* **24** 101–121.

Meyn, S. P. and Tweedie, R. L. (1993). *Markov Chains and Stochastic Stability*. Springer-Verlag, New York, London.

Mykland, P., Tierney, L. and Yu, B. (1995). Regeneration in Markov chain samplers. *Journal of the American Statistical Association* **90** 233–241.

Papaspiliopoulos, O. and Roberts, G. (2008). Stability of the Gibbs sampler for Bayesian hierarchical models. *The Annals of Statistics* **36** 95–117.

Roberts, G. O. and Rosenthal, J. S. (1997). Geometric ergodicity and hybrid Markov chains. *Electronic Communications in Probability* **2** 13–25.

Roberts, G. O. and Rosenthal, J. S. (1998). Markov chain Monte Carlo: Some practical implications of theoretical results (with discussion). *The Canadian Journal of Statistics* **26** 5–31.

Roberts, G. O. and Rosenthal, J. S. (1999). Convergence of slice sampler Markov chains. *Journal of the Royal Statistical Society,* Series B **61** 643–660.

Roberts, G. O. and Tweedie, R. L. (1996). Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika* **83** 95–110.

Roy, V. and Hobert, J. P. (2007). Convergence rates and asymptotic standard errors for Markov chain Monte Carlo algorithms for Bayesian probit regression. *Journal of the Royal Statistical Society,* Series B **69** 607–623.

Smith, T. C., Spiegelhalter, D. J. and Thomas, A. (1995). Bayesian approaches to random-effects meta-analysis: A comparative study. *Statistics in Medicine* **14** 2685–2699.