BU-1413-M

Data-Driven Gibbs Sampling

First Draft

George Casella
Cornell University

July 6, 1998

Abstract

The wide applicability of Gibbs sampling has increased the use of more complex hierarchical models. In such situations, to perform a Bayesian analysis an experimenter may be faced with the task of specifying values for hyperparameters in the deeper levels of a hierarchy. Such specifications can be difficult, as intuition tends to break down. A typical remedy is to estimate these hyperparameters, and proceed as if they were exactly specified. We examine the impact of this, and investigate the properties, and resulting inference, from a Gibbs sampling algorithm used in this manner. We also detail a computational algorithm that, with little additional effort, can produce all of the required estimates.

Key words and phrases : Hierarchical models, Bayesian computation, Empirical Bayes, Consistency, Likelihood.

This research was supported by NSF Grant DMS 9625440, and this is paper BU-xxxx-M in the Department of Biometrics, Cornell University, Ithaca, NY 14853.
File draft1.tex

1 Introduction

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

As an example, a veterinarian was interested in modeling occurrences of clinical mastitis in dairy herds. If θ_i $i = 1, 2, \cdots, p$, is the mean rate of occurrence of clinical mastitis in herd i, and $X_i =$ is the observed number occurrences of mastitis in herd i, a hierarchical model is

$$\begin{aligned} X_i &\sim \text{Poisson } (\theta_i t_i) \\ \theta_i &\sim \text{Gamma } (\alpha, \beta) \\ \beta &\sim \text{Gamma } (a, b) \end{aligned} \tag{1}$$

where t_i is the known size of herd i. A typical goal of the analysis it to estimate the posterior distribution of θ_i. To do so requires specification of not only the prior parameter α, but also of the hyperparameters a and b. Although an experimenter may have some idea of a reasonable value for α (although specifying α can be a challenge), it is almost certain that specification of a and b will be guesswork.

The veterinarian was not willing to specify these values for β, and we were asked to estimate them. Such a procedure seems common, especially in Gibbs sampling implementations. For example, consider the approach of Gelfand and Smith (1990) in their seminal paper. In analyzing the oft analyzed "pump-failure data" (Gaver and O'Muircheartaigh 1987), which is modeled with the hierarchy (1) defining $X_i =$ number of failures, $t_i =$ known time to failure, Gelfand and Smith set $a = .1$, $b = 1$, and estimated α from the data. They did not address the effect of estimating α.

Figure 1: Posterior density for the mean rate of clinical mastitis infection in a herd, and envelope of density functions corresponding to a 90% (asymptotic) confidence cube on a, b, and α.

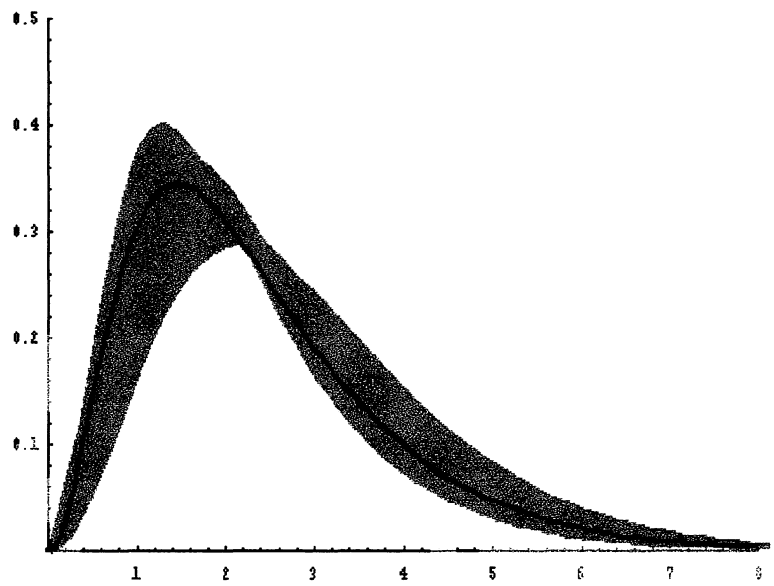


Figure 1 shows the results of the analysis that we propose. The posterior distribution is calculated using estimated values of the hyperparameters. Along with this posterior density, we also depict an envelope of densities corresponding to the values of the hyperparameters that lie in a 90% confidence cube. This confidence cube is constructed using likelihood methods, and is valid (asymptotically) with respect to the marginal distribution of the data. Moreover, a picture similar to this can be constructed for various ranges of the hyperparameters. This would give a good idea of the effect that the value of the hyperparameter had on the resulting posterior density, and thus identify the importance of the hyperparameter in the overall inference. Details are given in Section 5.

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 4 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 described here, 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 6.

A common implementation of the hierarchy (1) is to estimate some subset of α, a, and b (and specify the remaining ones), and calculate $\pi(\theta_i|\mathbf{x}, \hat{\alpha}, \hat{a}, \hat{b})$ 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 assess the errors, and then use these errors 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 much 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, to the robust Bayes formulation of 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 mention parametric empirical Bayes formulation of Morris (1983) (see also Carlin and Louis 1996), other recent uses of data-dependent priors include O'Hagan (1995), Berger and Perrichi (1996) and Shively, Kohn and Wood (1997). This last paper is similar in spirit to our approach, but does not go into as much detail on assessment of the effect of hyperparameter error.

The remainder of the paper is organized as follows. In Section 2 we develop the data-driven Gibbs sampler, and illustrate it with a "toy example" that we will repeat throughout. (The purpose of this example is to illustrate the working of the procedure in a simple situation.) Section 3 looks at some theoretical properties of the estimated posterior distribution, and gives a convergence result (with the proof in the Appendix). Section 4 goes into detail on the implementation in real problems. Here we show that with very

little overhead, the entire data-driven Gibbs sampler can be implemented using little more than the calculations of the original Gibbs sampler. In Section 5 we apply the data-driven Gibbs sampler to the data described in the introduction, and Section 6 contains a discussion.

2 The Data-Driven 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 data-driven 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)

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

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, where we assume that ψ is specified, based on the observations $\mathbf{x}_p = (x_1,\cdots,x_p)$ we would set up the iterations

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

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

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

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

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

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

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

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

Toy Example. To illustrate these points, we look at the following stylized situation, in which there is no Gibbs sampling. Suppose we observe $\mathbf{X}_p = \mathbf{x}_p = (x_1, \cdots, x_p)$ where

$$\begin{aligned} X_i &\sim \mathcal{N}(\theta_i, 1) \\ \theta_i &\sim \mathcal{N}(0, \tau^2) \end{aligned} \tag{5}$$

where τ^2 is unknown. Consider estimation of θ_1 (the other θ_is are similar). The desired posterior distribution is

$$\theta_1 \sim \mathcal{N}(bx_1, b) \text{ where } b = \frac{\tau^2}{1+\tau^2}$$

If we now try to estimate τ^2 with $\hat{\tau}^2$, an obvious candidate estimator is the MLE from the marginal likelihood

$$\begin{aligned} L(\tau^2|\mathbf{x}_p) &\propto \frac{e^{-\frac{1}{2}\sum \frac{(x_i-\mu)^2}{1+\tau^2}}}{(\tau^2)^{n/2}} \int \prod_i e^{-\frac{1}{2}\sum \frac{(\theta_i - bx_i)^2}{b}} d\theta_i \\ &\propto \frac{e^{-\frac{1}{2}\frac{\sum (x_i-\mu)^2}{1+\tau^2}}}{(1+\tau^2)^{n/2}}. \end{aligned} \tag{6}$$

We then compute $\hat{b} = \hat{\tau^2}/(1+\hat{\tau^2})$, and estimate the posterior distribution with $\mathcal{N}(\hat{b}x_1, \hat{b})$. How valid are the inferences that can be drawn from this estimated posterior distribution? ||

3 Convergence of the Posterior

The classic results on consistency of Bayes estimators date back to Doob (1948), 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 $\pi(\theta_r|\mathbf{x}_p, \psi)$, for $r = 1, \ldots, p$. The Gibbs sampler estimates this with the average of the conditional densities. With an estimated hyperparameter $\hat{\psi}_p$, the MLE of ψ under the marginal distribution $m(\mathbf{x}_p|\psi)$, the development in the Appendix gives conditions under which

$$\frac{1}{M}\sum_{j=1}^{M} \pi(\theta_r|\mathbf{x}_p, \hat{\psi}_p, \lambda^{(j)}) \stackrel{p\to\infty}{\to} \pi(\theta_r|x_r, \mathbf{x}_p, \psi) \tag{7}$$

in probability. More precisely, for a compact set A, we have that

$$\int_A \left| \frac{1}{M}\sum_{j=1}^{M} \pi(\theta_r|\mathbf{x}_p, \hat{\psi}_p, \lambda^{(j)}) - \pi(\theta_r|x_r, \mathbf{x}_p, \psi) \right| d\theta_r \to 0 \tag{8}$$

in probability.

We note that, in some sense, the Gibbs sampler is an irrelevant concern. Apply the triangle inequality to the integrand in (8) to see that

$$\begin{aligned} \left|\frac{1}{M}\sum_{j=1}^{M} \pi(\theta_i|\mathbf{x}_p, \hat{\psi}_p, \lambda^{(j)}) - \pi(\theta_i|\mathbf{x}_p, \psi)\right| \;\leq\; & \left|\frac{1}{M}\sum_{j=1}^{M} \pi(\theta_i|\mathbf{x}_p, \hat{\psi}_p, \lambda^{(j)}) - \pi(\theta_i|\mathbf{x}_p, \hat{\psi}_p)\right| \\ & + \left|\pi(\theta_i|\mathbf{x}_p, \hat{\psi}_p) - \pi(\theta_i|\mathbf{x}_p, \psi)\right| \end{aligned} \tag{9}$$

The first term on the right side of (9) only involves the estimated $\hat{\psi}_p$. Here we are running the Gibbs sampler with this estimated value, which has no

effect on the numerical properties of the Gibbs algorithm. So, as long as we are using an ergodic Markov chain, we can be sure that this term converges to zero. The second term is of more concern, as its convergence is governed by the probabilistic structure of the problem. Conditions for its convergence are detailed in the Appendix.

4 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. Referring back to (6), such a calculation could require a high-dimensional integration. Although we can do the calculation in the toy example, it is unlikely that such calculations can be done in practice. 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

$$L(\psi|\mathbf{x}) = \frac{L(\psi|\mathbf{x}, \boldsymbol{\theta}, \lambda)}{\pi(\boldsymbol{\theta}, \lambda|\mathbf{x}, \psi)}. \tag{10}$$

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. This expression for $L(\psi|\mathbf{x})$ leads to the EM identity

$$\begin{aligned} \log L(\psi|\mathbf{x}) &= E[\log L(\psi|\mathbf{x})|\psi_0] \\ &= E[\log L(\psi|\mathbf{x}, \boldsymbol{\theta}, \lambda)|\psi_0] - E[\log \pi(\boldsymbol{\theta}, \lambda|\mathbf{x}, \psi)|\psi_0] \end{aligned}$$

where the expectation is taken with respect to $\pi(\boldsymbol{\theta}, \lambda|\mathbf{x}, \psi_0)$.

Return to the Toy Example. In the toy example the conditional likelihood of τ^2 is

$$L(\tau^2|\mathbf{x},\boldsymbol{\theta}) \propto e^{-\frac{1}{2}\sum(x_i-\theta_i)^2} \frac{e^{-\frac{1}{2}\sum\frac{(\theta_i-\mu)^2}{\tau^2}}}{\tau^p}$$

We then get $\pi(\boldsymbol{\theta}|\mathbf{x},\hat{\tau}^2)$ from the EM iterations

$$\tau^{2(k+1)} = \text{argmax}_{\tau^2} E[\log L(\tau^2|\mathbf{x},\boldsymbol{\theta})|\tau^{2(k)}]$$

where the expectation is taken with respect to $\mathcal{N}(b^{(k)}\mathbf{x}, b^{(k)}I)$, and $b^{(k)} = \tau^{2(k)}/(1+\tau^{2(k)})$. Of course, the conditional likelihood function for τ^2 is merely the product of the densities in the hierarchy, making its calculation easy.

Rather than actually computing the above expectation, we will instead use the Monte Carlo version of the EM algorithm , which has iterations

$$\tau^{2(k+1)} = \text{argmax}_{\tau^2} \frac{1}{M}\sum_{j=1}^{M} \log L(\tau^2|\mathbf{x},\boldsymbol{\theta}^{(j)}),$$

where we generate a sample $\boldsymbol{\theta}^{(1)},\ldots,\boldsymbol{\theta}^{(M)}$ from $\mathcal{N}(b^{(k)}\mathbf{x}, b^{(k)}I)$. ||

In the generic hierarchy (2) we have the Monte Carlo EM iterations

$$\psi^{(k+1)} = \text{argmax}_{\psi} \frac{1}{M}\sum_{j=1}^{M} \log L(\psi|\mathbf{x},\boldsymbol{\theta}^{(j)},\lambda^{(j)}). \tag{11}$$

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 data-driven 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 on $\pi(\boldsymbol{\theta}|\mathbf{x},\lambda,\psi^{(k)})$ and $\pi(\lambda|\mathbf{x},\boldsymbol{\theta},\psi^{(k)})$

3. Update $\psi^{(k)}$ with the Monte Carlo EM iteration (11) and return to step 2 of the algorithm.

4. At convergence of $\psi^{(k)}$ to the marginal MLE $\hat{\psi}_p$, produce a final Gibbs sample from $\pi(\boldsymbol{\theta}|\mathbf{x}, \lambda, \hat{\psi}_p)$ and $\pi(\lambda|\mathbf{x}, \boldsymbol{\theta}, \hat{\psi}_p)$

The final Gibbs sample can then be used to construct posterior estimates such as (4). We can, therefore, produce the data-driven 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.

5 Practice

To illustrate the use of the algorithm of Section 4, we will first look at the model used by Gelfand and Smith (1990) to analyze the pump failure data. Starting from the hierarchy (1), they assumed that $a = .1$ and $b = 1$, and that α was unknown. This yields the Gibbs sampler

$$\begin{aligned} \theta_i|\mathbf{x}, \alpha, \beta &\sim \text{Gamma}(x_i + \alpha, [t_i + 1/\beta]^{-1}) \\ \beta|\mathbf{x}, \alpha, \boldsymbol{\theta} &\sim \text{IG}(p\alpha + a, [\textstyle\sum \theta_i + 1/b]^{-1}) \end{aligned} \tag{12}$$

where $\text{Gamma}(\alpha, \beta)$ is the gamma distribution and $\text{IG}(a, b)$ is the inverted gamma distribution. We then have the conditional likelihood function

$$L(\alpha|\mathbf{x}, \boldsymbol{\theta}, \beta) \propto \frac{e^{-\sum \theta_i[t_i+1/\beta]} \prod[\theta_i t_i]^{x_i+\alpha-1}}{[\Gamma(\alpha)\beta^\alpha]^p}$$

and marginal likelihood function $L(\alpha|\mathbf{x}) = L(\alpha|\mathbf{x}, \boldsymbol{\theta}, \beta)/\pi(\boldsymbol{\theta}, \beta|\mathbf{x}, \alpha)$.

The Algorithm of Section 4 is:

1. Set $k = 0$ and initialize $\hat{\alpha}^{(0)}$.

2. Update $\hat{\alpha}^{(k)} \rightarrow \hat{\alpha}^{(k+1)}$ by

$$\hat{\alpha}^{(k+1)} = \text{argmax}_\alpha \frac{1}{M} \sum_{j=1}^{M} \log L(\alpha|\mathbf{x}, \boldsymbol{\theta}^{(j)}, \beta^{(j)})$$

3. For $j = 1, \ldots, M$, generate from the Gibbs sampler (12)

$$\begin{aligned} \boldsymbol{\theta}^{(j)} &\sim \pi(\boldsymbol{\theta}|\mathbf{x}, \hat{\alpha}^{(k)}, \beta^{(j-1)}) \\ \beta^{(j)} &\sim \pi(\beta|\mathbf{x}, \hat{\alpha}^{(k)}, \boldsymbol{\theta}^{(j)}) \end{aligned}$$

and return to step 2.

4. On convergence of $\hat{\alpha}^{(k)} \rightarrow \hat{\alpha}$, generate

$$\begin{aligned} \boldsymbol{\theta}^{(j)} &\sim \pi(\boldsymbol{\theta}|\mathbf{x}, \hat{\alpha}, \beta^{(j-1)}) \\ \beta^{(j)} &\sim \pi(\beta|\mathbf{x}, \hat{\alpha}, \boldsymbol{\theta}^{(j)}) \end{aligned}$$

and estimate

$$\begin{aligned} \pi(\boldsymbol{\theta}|\mathbf{x}, \hat{\alpha}) &\approx \frac{1}{M} \sum_{j=1}^{M} \pi(\boldsymbol{\theta}|\mathbf{x}, \hat{\alpha}, \beta^{(j)}) \\ \pi(\beta|\mathbf{x}, \hat{\alpha}) &\approx \frac{1}{M} \sum_{j=1}^{M} \pi(\beta|\mathbf{x}, \hat{\alpha}, \boldsymbol{\theta}^{(j)}) \end{aligned}$$

which are consistent estimates of $\pi(\boldsymbol{\theta}|\mathbf{x}, \alpha)$ and $\pi(\beta|\mathbf{x}, \alpha)$.

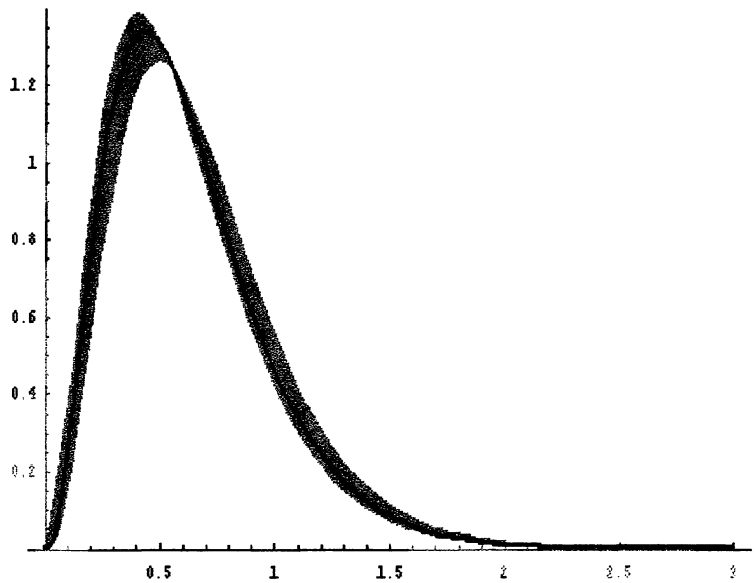
Note, once again, that with the exception of the maximization step, only the calculations of the original Gibbs sampler are needed.

In addition to calculating the values of the hyperparameters, the EM algorithm brings along methodology for calculating the standard errors of these estimates (see, for example, Tanner 1996, Section 4.4). The advantage to this is that along with the estimated posterior distribution, we can also display the envelope of posterior densities corresponding to a range of the hyperparameter. Figure 2 displays an estimated posterior along with an envelope of posterior densities corresponding to a 90% (asymptotic) confidence interval on the hyperparameter.

As an aside, we have also found that for smaller problems (like this one) some computer algebra programs can actually calculate

$$\begin{aligned} -\frac{\partial^2}{\partial \alpha^2} \log L(\alpha|\mathbf{x}) &= -\frac{1}{M} \sum_{j=1}^{M} \frac{\partial^2}{\partial \alpha^2} \log L(\alpha|\mathbf{x}, \boldsymbol{\theta}^{(j)}, \beta^{(j)}) \\ &\quad + \frac{1}{M} \sum_{j=1}^{M} \frac{\partial^2}{\partial \alpha^2} \log \pi(\boldsymbol{\theta}^{(j)}, \beta^{(j)}|\mathbf{x}, \alpha). \end{aligned}$$

Figure 2: Posterior density for mean time to failure of a nuclear pump, and envelope of density functions corresponding to a 90% (asymptotic) confidence interval on the hyperparameter α.



This again makes calculation quite simple, and the EM standard error approximations do not have to be used.

To further illustrate our methodology, and perhaps to make the pump failure analysis even more realistic, we now reanalyze the data with the hierarchy (1), but now assume that the three hyperparameters a, b, and α are all unknown. The methodology remains the same, and we still run the original Gibbs sampler (12), but now we use the conditional likelihood

$$L(\alpha|\mathbf{x},\boldsymbol{\theta},\boldsymbol{\beta}) \propto \prod_i \frac{e^{-\theta_i[t_i+1/\beta_i]}[\theta_i t_i]^{x_i+\alpha-1}}{\Gamma(\alpha)\beta_i^{\alpha+a+1}} \frac{e^{-1/b\beta_i}}{\Gamma(a)b^a}$$

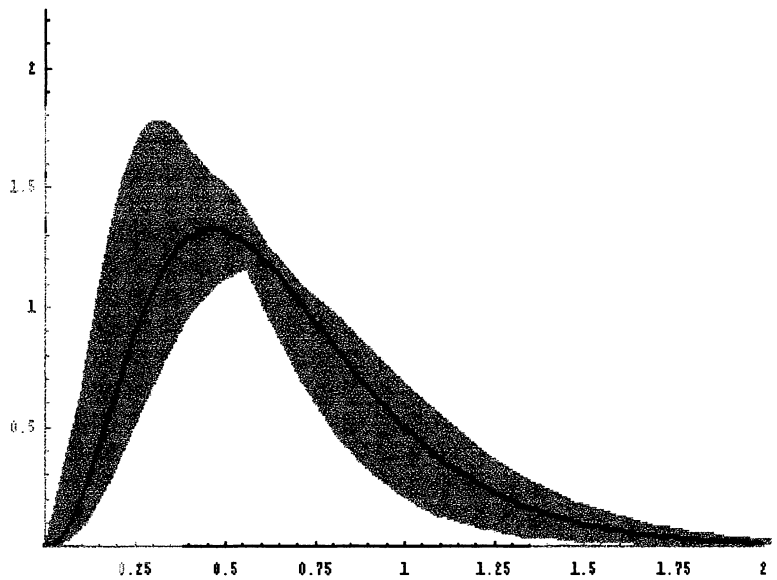
and marginal likelihood $L(\alpha,a,b|\mathbf{x}) = L(\alpha,a,b|\mathbf{x},\boldsymbol{\theta},\beta)/\pi(\boldsymbol{\theta},\beta|\mathbf{x},\alpha,a,b)$. The algorithm is as before, but we now update

$$(\hat{\alpha}^{(k)},\hat{a}^{(k)},\hat{b}^{(k)}) \rightarrow (\hat{\alpha}^{(k+1)},\hat{a}^{(k+1)},\hat{b}^{(k+1)}).$$

Figure 3 shows an estimated posterior density from this analysis, along with a 90% envelope of posterior density functions. Here we used a Bonferroni

inequality to construct a simultaneous 90% (asymptotic) confidence cube on a, b, and α.

Figure 3: Posterior density for mean time to failure of a nuclear pump, and envelope of density functions corresponding to a 90% (asymptotic) confidence cube on a, b, and α.



The analysis of the clinical mastitis data, shown in Figure 1, is done with virtually the identical Gibbs sampler as this one.

6 Discussion

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 3 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 re-

sulting 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)

The actual implementation of the methodology, and the inference, is almost automatic, and uses only the standard tools of the Gibbs sampler. 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, here based on likelihood. The method and algorithm of Section 4 details the EM/Gibbs sampler that results in consistent estimates of posterior distributions and asymptotically valid confidence sets. These confidence sets also provide a graphical means of assessing the effect of hyperparameter estimation, providing an easy way to assess sensitivity.

From a graph such as Figure 3, one can also attach a Bayesian inference to the envelope of densities. For example, for a fixed value of $\psi = (\alpha, a, b)$, 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

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

where $\pi(\psi|\mathbf{x})$ is the posterior distribution of ψ. Moreover, if we assume that there is some "probability matching" prior for ψ, and if we construct an approximate $1-\gamma_2$ confidence set for ψ using the likelihood 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 to obtain approximations to these probabilities.

One other point about implementation needs mentioning, a point that is also noted by Booth and Hobert (1998). To implement the Monte Carlo EM algorithm of Section 4, 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. To see this, recall the Monte Carlo EM iteration (11). In the calculation of the average log

likelihood at the k^{th} 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 clarity.) 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

$$\begin{aligned} & \frac{1}{M} \sum_{j=1}^{M} \log L(\psi | \mathbf{x}, \boldsymbol{\theta}_k^{(j)}, \lambda_k^{(j)}) \\ & \approx \frac{1}{M} \sum_{j=1}^{M} \log L(\psi | \mathbf{x}, \boldsymbol{\theta}_0^{(j)}, \lambda_0^{(j)}) \frac{\pi(\boldsymbol{\theta}_0^{(j)}, \lambda_0^{(j)} | \mathbf{x}, \psi^{(k)})}{\pi(\boldsymbol{\theta}_0^{(j)}, \lambda_0^{(j)} | \mathbf{x}, \psi^{(0)})}. \end{aligned} \tag{13}$$

One difficulty remains in using (13), in that in the Gibbs sampler we do not know the form of $\pi(\boldsymbol{\theta}, \lambda | \mathbf{x}, \psi)$. However, from (10), 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 (13) as it only enters through $\psi^{(0)}$ and $\psi^{(k)}$. Hence in the maximization of the EM sequence we use

$$\begin{aligned} & \text{argmax}_{\psi} \frac{1}{M} \sum_{j=1}^{M} \log L(\psi | \mathbf{x}, \boldsymbol{\theta}_k^{(j)}, \lambda_k^{(j)}) \\ & \approx \text{argmax}_{\psi} \frac{1}{M} \sum_{j=1}^{M} \log L(\psi | \mathbf{x}, \boldsymbol{\theta}_0^{(j)}, \lambda_0^{(j)}) \frac{L(\psi^{(k)} | \mathbf{x}, \boldsymbol{\theta}_0^{(j)}, \lambda_0^{(j)})}{L(\psi^{(0)} | \mathbf{x}, \boldsymbol{\theta}_0^{(j)}, \lambda_0^{(j)})}. \end{aligned} \tag{14}$$

This strategy can be refined to use periodic updates of the $(\boldsymbol{\theta}, \lambda)$ sample, which should improve the approximation in (14).

An obvious alternative to the model considered here is the full Bayesian hierarchical model with a "flat" or other "noninformative" prior on the hyperparameter. For example, in the generic hierarchy (2), we can adopt the approach of George, Makov, and Smith (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 $L^* = \int L \, d\psi$. Then use the Gibbs sampler

$$\begin{aligned} \boldsymbol{\theta}^{(j+1)} & \sim \pi(\boldsymbol{\theta} | \mathbf{x}, \psi^{(j)}, \lambda^{(j)}) \\ \lambda^{(j+1)} & \sim g(\lambda | \mathbf{x}, \psi, \boldsymbol{\theta}^{(j+1)}) \\ \psi^{(j+1)} & \sim L^*(\psi | \mathbf{x}, \boldsymbol{\theta}^{(j+1)}, \lambda^{(j+1)}) \end{aligned} \tag{15}$$

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, $L^* = \int L \, d\psi$ may not be finite. (This can, of course, be fixed by instead calculating $\int L \, h(\psi) d\psi$ for some prior h, but then the effect on the inference from such a prior need be assessed.) Secondly, generation of samples in (15) can be quite difficult, even if a prior h is used. Thirdly, the consistency results of Section 3 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 other way varying) the shape of the hyperparameter confidence set.

A Appendix

We will work with the generic hierarchy

$$
\begin{aligned}
X_i &\sim f(x|\theta_i, \psi) \quad i = 1, 2, \cdots, p \\
\theta_i &\sim \pi(\theta|\lambda, \psi) \\
\lambda &\sim g(\lambda|\psi).
\end{aligned} \tag{16}
$$

For unknown ψ, the Gibbs sampling estimate of the posterior $\pi(\theta_k|\mathbf{x}_p, \psi)$ is $\frac{1}{M}\sum_{j=1}^{M} \pi(\theta_k|\mathbf{x}_p, \hat{\psi}_p, \lambda^{(j)})$. We are concerned with the limiting behavior of this estimate as $p \to \infty$.

First, we need to be a bit more formal and, for fixed r, define $\mathbf{x}_p^{(-r)} = (x_1, \ldots, x_{r-1}, x_{r+1}, \ldots, x_p)$. This notation makes it clear that when we consider the posterior distribution for θ_r, it is conditional on x_r. The convergence of the Gibbs sampling estimate depends on the properties of the Markov chain produced, in particular we need the chain to be *ergodic* (see Tierney 1994, Section 3.2 or Robert and Casella 1999, Section 7.1.3). We then can now state the following theorem.

Theorem A.1 *Suppose that $\hat{\psi}_p$ is a consistent estimator of ψ with respect to the marginal distribution of $\mathbf{X}$, $m(\mathbf{x}|\psi)$, that $\pi(\theta_r|x_r, \mathbf{x}_p^{(-r)}, \psi)$ is a continuous*

function of ψ, and that the Gibbs sampler produces an ergodic Markov chain. Then, under the marginal distribution $m(\cdot|\psi)$, for each r and compact A, as $p \to \infty$,

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

in probability.

Proof First write, by the triangle inequality,

$$\begin{aligned} &\left| \frac{1}{M} \sum_{j=1}^{M} \pi(\theta_r | x_r, \mathbf{x}_p^{(-r)}, \hat{\psi}_p, \lambda^{(j)}) - \pi(\theta_r | x_r, \mathbf{x}_p^{(-r)}, \psi) \right| \\ &\quad \leq \left| \frac{1}{M} \sum_{j=1}^{M} \pi(\theta_r | x_r, \mathbf{x}_p^{(-r)}, \hat{\psi}_p, \lambda^{(j)}) - \pi(\theta_r | x_r, \mathbf{x}_p^{(-r)}, \hat{\psi}_p) \right| \\ &\quad + \left| \pi(\theta_r | x_r, \mathbf{x}_p^{(-r)}, \hat{\psi}_p) - \pi(\theta_r | x_r, \mathbf{x}_p^{(-r)}, \hat{\psi}_p) \right| \end{aligned} \tag{17}$$

Consider the first term on the right side of (17), which only concerns convergence of the Markov chain for fixed $\hat{\psi}_p$. As the integration is over a compact set A, ergodicity of the Markov Chain implies that given $\varepsilon > 0$, for each p we can choose the number of Monte Carlo iterations M_p to make the integral over A less than ε. Thus we only need concentrate on the second term.

Under the marginal distribution, since $\hat{\psi}_p$ is a consistent estimator of ψ, the continuity of π implies that $\pi(\theta_r | x_r, \mathbf{x}_p^{(-r)}, \hat{\psi}_p) \to \pi(\theta_r | x_r, \mathbf{x}_p^{(-r)}, \hat{\psi}_p)$ for each fixed θ_r, and the compactness of A completes the proof. □

There are two ways that the result of Theorem A.1 can be strengthened. First, the convergence can be strengthened to almost everywhere convergence. To do so requires additional requirements on the marginal distribution m, requirements that are typically satisfied in practice (see Schervisch 1995, Section 7.3.2). The other strengthening of the theorem is to remove the requirement of compactness of A. This requires additional requirements on both the marginal and conditional densities. To deal with the second term in (17), removal of compactness would necessitate a tail condition on $\pi(\theta_r | x_r, \mathbf{x}_p^{(-r)}, \psi)$, which should typically not be a problem in practice. Dealing with the first term seems a bit more problematic. Strengthening the

Markov chain requirement to *geometric ergodicity* still leaves a gap, as the bounding function must be integrable with respect to the marginal distribution m, and this is neither automatic nor easy to check. (Requiring *uniform ergodicity* would do, but Gibbs chains rarely satisfy this condition.) The other route is to require *tightness* (see Billingsley 1995, Section 25) of the family of distributions $\frac{1}{M}\sum_{j=1}^{M}\pi(\theta_r|x_r, \mathbf{x}_p^{(-r)}, \hat{\psi}_p, \lambda^{(j)})$. This seems the more promising route.

References

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

Berger, J. O. (1990). Robust Bayesian analysis: Sensitivity to the prior. *J. Statist. Plan. Inf.* **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 *J. Amer. Statist. Assoc.* **91** 109-122.

Billingsley, P. (1995). *Probability and Measure, Third Edition* New York: Wiley

Booth, J. G and Hobert, J. P. (1998). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. Technical Report, Department of Statistics, University of Florida.

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. *Ann. Statist.* **19** 338-353.

Diaconis, P. and Freedman, D. (1986). On the consistency of Bayes estimates. *Ann. Statist.* **14** 1-26.

Doob, J. L. (1948). Application of the theory of martingales. *Coll. Int. du C. N. R. S. Paris Ann. Probab.* **2** 183-201.

Efron, B. (1996). Empirical Bayes methods for combining likelihood (with discussion). *J. Amer. Statist. Assoc.* **91** 538-565.

Gaver, D. P. and O'Muircheartaigh, I. G. (1987). Robust empirical Bayes analysis of event rates. *Technometrics* **29** 1-15.

Gelfand, A. E. (1996). Comment on Efron's paper. *J. Amer. Statist. Assoc.* **91** 551-552.

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

Geman, S. and Geman, D. (1984) Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.* **6**, 721–741.

George, E.I., Makov, U. and Smith, A.F.M. (1993). Conjugate likelihood distributions. *Scandinavian J. Statist.* **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. *Aspects of Uncertainty*, P. R. Freedman and A. F. M. Smith, eds, 181-198. New York: Wiley

Goel, P. and DeGroot, M. (1981). Information about hyperparameters in hierarchical models. *J. Amer. Statist. Assoc.* **76**, 140-147.

Kass, R. E. and Steffey, D. (1989). Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). *J. Amer. Statist. Assoc.* **84**, 717-726.

Morris, C. N. (1983). Parametric empirical Bayes inference: Theory and applications (with discussion). *J. Amer. Statist. Assoc.* **78**, 47-65.

O'Hagan, A. (1995). Fractional Bayes factors for model comparison. *J. Roy. Statist. Soc. Ser. B* **57** 99-118.

Robbins, H. (1964). The empirical Bayes approach to statistical decision problems. *Ann. Math. Statist.* **35** 1-20.

Robbins, H. (1983). Some thoughts on empirical Bayes estimation. *Ann. Statist.* **11**, 713-723.

Robert, C. P. and Casella, G. (1999). *Monte Carlo Statistical Inference.* New York: Springer-Verlag

Schervisch, M. J. (1995). *Theory of Statistics.* New York: Springer-Verlag

Shively, T. S., Kohn, R. and Wood, S. (1997). Variable selection and function estimation in additive nonparametric regression using a data-based prior. Technical Report, Australian Graduate School of Management, University of New South Wales.

Schwartz, L. (1965). On Bayes procedures. *ZeitWahr* **4** 10-26.

Tanner, M. (1996) *Tools for Statistical Inference, 3rd edition.* New York: Springer-Verlag

Tierney, L. (1994) Markov chains for exploring posterior distributions (with discussion). *Ann. Statist.* . **22**, 1701–1786.

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