Functional Compatibility, Markov Chains and Gibbs Sampling with Improper Posteriors

James P. Hobert*
Department of Statistics
University of Florida

George Casella†
Biometrics Unit
Cornell University

April 20, 1995

*Research supported by National Institute of Health Sciences Training Grant EHS-5-T32-ES07261-03 and National Science Foundation Grants DMS-9305547 and INT-9216784. This is technical report #481 in the Statistics Department, University of Florida, Gainesville, FL 32611.

†Research supported by National Science Foundation Grants DMS-9305547 and INT-9216784. This is technical report BU-1280-M in the Biometrics Unit, Cornell University, Ithaca, NY 14853.

AMS 1991 *subject classifications.* Primary 62F15; secondary 60J27

Key words and phrases. Bayesian Hierarchical Models, Compatible Conditional Densities, Improper Priors, Markov Transition Functions, Null Markov Chains

Abbreviated title. Gibbs Sampling and Improper Posteriors

Abstract

The members of a set of conditional probability density functions are called *compatible* if there exists a joint probability density function which generates them. We generalize this concept by calling the conditionals *functionally compatible* if there exists a (possibly non-integrable) function that behaves like a joint density as far as generating the conditionals according to the probability calculus. A necessary and sufficient condition for functional compatibility is given, which provides a method of calculating this function, if it exists. A Markov transition function is then constructed using a set of functionally compatible conditional densities and it is shown, using the compatibility results, that the associated Markov chain is positive recurrent if and only if the conditionals are compatible. A Gibbs Markov chain, constructed via "Gibbs conditionals" from a hierarchical model with an improper posterior, is a special case. Monte Carlo approximations based on Gibbs chains are shown to have undesirable limiting behavior when the posterior is improper. The results are applied to a Bayesian hierarchical one-way random effects model with an improper posterior distribution. The model is simple, but also quite similar to some models with improper posteriors which have been used in conjunction with the Gibbs sampler in the literature.

1. Introduction

Consider two real valued functions $f_1(x_1, x_2)$ and $f_2(x_1, x_2)$ with domain $\Re^2$. Suppose that there exist two sets, A_1 and A_2, in $\Re$ such that for any $x_2 \in A_2$, f_1 is a probability density in x_1 whose support is A_1 and similarly, for any $x_1 \in A_1$, f_2 is a probability density in x_2 with support A_2. The functions f_1 and f_2 may be thought of as conditional probability densities and will hereafter be written as $f_1(x_1|x_2)$ and $f_2(x_2|x_1)$. Arnold and Press (1989) give necessary and sufficient conditions for the existence of a joint density function $f(x_1, x_2)$ whose conditionals are given by f_1 and f_2. When such an f exists, f_1 and f_2 are called *compatible* conditional densities. Arnold and Press allow A_1 and A_2 to depend on x_2 and x_1, respectively. If f_1 and f_2 are compatible and the support sets are fixed, that is $A_1(x_2) = A_1$ and $A_2(x_1) = A_2$, then results of Besag (1974) show that the joint density is unique (and satisfies the *positivity* condition). The following simple example from Gourieroux and Monfort (1979) shows that uniqueness does not necessarily hold when the support sets are not fixed.

Example 1. Define f_1 and f_2 by

$$f_1(x_1|x_2) = \begin{cases} I(x_1 \in [1,2]) & \text{if } x_2 \in [1,2] \\ I(x_1 \in [2,3]) & \text{if } x_2 \in [2,3] \\ 0 & \text{otherwise} \end{cases}$$

$$f_2(x_2|x_1) = \begin{cases} I(x_2 \in [1,2]) & \text{if } x_1 \in [1,2] \\ I(x_2 \in [2,3]) & \text{if } x_1 \in [2,3] \\ 0 & \text{otherwise} \end{cases}$$

where $I(\cdot)$ is the indicator function. The support sets of f_1 and f_2 clearly depend on x_2 and x_1, respectively. These conditionals are compatible, but any joint density of the form

$$f(x_1, x_2) = \alpha I(x_1 \in [1,2]) I(x_2 \in [1,2]) + (1-\alpha) I(x_1 \in [2,3]) I(x_2 \in [2,3])$$

with $\alpha \in (0,1)$ will produce them.

In Section 2, we consider the compatibility of the set of conditional densities, $f_1(x_1|x_2, \ldots, x_m), \ldots, f_m(x_m|x_1, \ldots, x_{m-1})$, under the assumption that the support sets are fixed. Our approach is to first introduce a necessary (but not sufficient) condition for compatibility, which we call *functional compatibility*. Conditional densities are functionally compatible if there exists a function g (possibly non-integrable) which, if treated as a joint density, generates the conditionals. For example, f_1 and f_2 are functionally compatible if there exists a function $g(x_1, x_2)$ such that $g/\int g\, dx_1 = f_1$ and $g/\int g\, dx_2 = f_2$. Clearly, if no such g exists, the conditionals cannot be compatible. On the other hand, the existence of g does not guarantee compatibility since g may not be normalizable. For instance, consider the exponential conditionals of Casella and George (1992, Example 2): $f_1(x_1|x_2) = x_2 \exp(-x_1 x_2)$ and

$f_2(x_2|x_1) = x_1 \exp(-x_1 x_2)$. The non-integrable function $g(x_1, x_2) = \exp(-x_1 x_2)$, if treated as a joint density, does yield f_1 and f_2 as its conditionals, thus f_1 and f_2 are functionally compatible, but they are not compatible (see Theorem 2).

A necessary and sufficient condition for functional compatibility is given (Theorem 1) which allows one to check for functional compatibility and construct g if it exists. Compatibility of the conditionals follows if and only if g is integrable. Thus, if the compatibility of a set of conditionals is in question, one may first check whether or not they are functionally compatible. If they are not, then they are not compatible either, and if they are, the integral of g must be checked.

The necessary and sufficient condition for functional compatibility is based on the following argument. Assume that the support sets of f_1 and f_2 are fixed. If f_1 and f_2 are compatible, and we let $f(x_1, x_2)$ denote the unique joint density, then for any particular $x_1' \in A_1$ and $x_2' \in A_2$, we have (Besag 1974, Gelman and Speed 1993)

$$f(x_1, x_2) \propto \frac{f_1(x_1|x_2)\, f_2(x_2|x_1')}{f_1(x_1'|x_2)} \quad \text{and} \quad f(x_1, x_2) \propto \frac{f_2(x_2|x_1)\, f_1(x_1|x_2')}{f_2(x_2'|x_1)}. \tag{1.1}$$

Therefore, if we are given f_1 and f_2, and compatibility is in question, a necessary condition for compatibility is that the ratio of the two right-hand sides be constant for any point (x_1', x_2'). This condition is actually necessary and sufficient for functional compatibility and when it is satisfied, either of the right-hand sides will serve as g.

In Section 3, we consider a Markov transition function constructed using a set of functionally compatible conditional densities. It is shown that a σ-finite measure, π, constructed using g, is an invariant measure for the associated Markov chain (see Theorem 3). Results from Section 2 imply that π is a finite measure (normalizable) if and only if the conditional densities (used to construct the transition function) are compatible. It follows that the chain is positive recurrent if and only if the conditional densities are compatible. Section 3 ends with a general result for a class of null chains which describes the limiting behavior of averages.

The results of Sections 2 and 3 are relevant in situations where the Gibbs sampler (Gelfand and Smith 1990, Tierney 1995) is applied in an attempt to explore an *improper* posterior distribution. The remainder of this section is a discussion of this particular application and Section 4 gives an example concerning a Bayesian hierarchical random effects model (with an improper posterior) which is similar to models with improper posteriors which have been employed in the literature.

Often, either from a lack of prior information or simply for convenience, improper priors are assigned to the hyperparameters of Bayesian hierarchical models. When improper priors are used in any stage of a hierarchical model, the resulting posterior distribution must be checked for propriety. The integration necessary to check propriety and calculate posterior quantities of interest can be daunting, however. When the posterior is proper, the Gibbs sampler can often be used to simulate from the posterior distribution. The simulation results can then be used to calculate Monte Carlo approximations of the posterior quantities of interest, thus avoiding difficult in-

tegration. Unfortunately, if one mistakenly assumes propriety, it may still be possible to apply the Gibbs sampler. Consider the following example.

Example 2. Let Y_1, Y_2, Y_3 be iid $\mathrm{N}(\mu, \sigma^2)$ and suppose that the improper prior on the parameters is $\pi(\mu, \sigma^2) = I_{\Re_+}(\sigma^2)$ where $I_S(\cdot)$ is the indicator of the set S. It is not difficult to show that the posterior is improper, that is

$$\int_{\Re_+}\int_{\Re} (\sigma^2)^{-\frac{3}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_i (y_i-\mu)^2\right) d\mu\, d\sigma^2 = \infty. \tag{1.2}$$

Given this knowledge, consider what would have happened had we assumed that the posterior distribution was proper and applied the Gibbs sampler. If we had assumed propriety, we would have written the posterior as

$$\pi(\mu, \sigma^2|y_1, y_2, y_3) \propto (\sigma^2)^{-\frac{3}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_i (y_i-\mu)^2\right) I_{\Re_+}(\sigma^2). \tag{1.3}$$

Use of the Gibbs sampler in this situation would require $f_1(\mu|\sigma^2, \mathbf{y})$, the conditional density of μ given σ^2 and the data, and $f_2(\sigma^2|\mu, \mathbf{y})$, the conditional density of σ^2 given μ and the data. These densities can be calculated by recognition based on (1.3) and it follows that f_1 is $\mathrm{N}(\overline{y}, \sigma^2/3)$ and f_2 is $\mathrm{IG}\left(1/2, 2\left(\sum (y_i-\mu)^2\right)^{-1}\right)$ [IG is the inverted gamma distribution (see Berger 1985 p.561)]. Note that f_1 and f_2 are functionally compatible, with π serving as g, but not compatible since π is not integrable.

Given a starting value for μ, say $\mu^{(0)} = \overline{y}$, a Gibbs chain,

$$\sigma^{2(1)}, \mu^{(1)}, \sigma^{2(2)}, \mu^{(2)}, \sigma^{2(3)}, \mu^{(3)}, \sigma^{2(4)}, \ldots \tag{1.4}$$

could be constructed in the usual manner. (The symbol $\sigma^{2(i)}$ represents the ith value of σ^2 in the Gibbs chain.) We would then be under the impression that $\left(\mu^{(n)}, \sigma^{2(n)}\right)$ converges in distribution to a random variable whose distribution is the "posterior distribution."

Figures 1 and 2 show the first 1,000 values of $\ln|\mu^{(i)}|$ and $\ln\left(\sigma^{2(i)}\right)$, respectively, for one realization of this Gibbs chain. (The data, y_1, y_2, and y_3, were simulated from a standard normal distribution.) The chain is apparently out of control. At the 1,000th iteration, the magnitude of the μ component is up to about 10^{37} and the σ^2 component is up to about 10^{65}.

Thus, the Gibbs chain in Example 2 provides a "red flag" warning us that there may be a problem. If an experimenter had mistakenly assumed propriety of the posterior in Example 2, collected three data points whose mean and standard deviation were near 0 and 1, respectively, and then simulated a Gibbs chain like the one shown in Figures 1 and 2, he would probably question his assumption regarding propriety and discover his mistake before any damage was done.

If Gibbs chains corresponding to improper posteriors always "misbehaved," there would be no reason to worry about demonstrating propriety before applying the Gibbs

sampler, since we would discover an improper posterior through the Gibbs output. This is not the case, however. Sometimes the output from Gibbs chains corresponding to improper posteriors appears perfectly reasonable, that is, the Gibbs chains do not provide a "red flag." These situations are very dangerous because one ends up making inferences about a nonexistent posterior distribution. Such instances can be found in the literature (see Section 4), thus the properties of such chains, and the associated Monte Carlo approximations, are of practical interest.

In general, "Gibbs conditionals" calculated via a proportionality, as are those in Example 2, are functionally compatible. Therefore the results from Sections 2 and 3 may be applied and show that, under some mild regularity conditions, a Gibbs sampler is positive recurrent if and only if the posterior distribution is proper. (Note that this fact precludes the use of standard "convergence diagnostics" (Cowles and Carlin 1994) for detection of improper posteriors through Gibbs output, since the diagnostics are based on the assumption that the Gibbs chain is positive recurrent.) It follows from the results of Section 3 that, although the output from Gibbs chains corresponding to improper posteriors may appear reasonable and can even lead to nice looking pictures of (nonexistent) marginal posterior densities, the limiting behavior of the Monte Carlo approximations is quite undesirable.

2. Compatibility of Conditional Densities

2.1. The Problem

Consider m measure spaces $(A_i, \mathcal{B}_i, \mu_i)$, $i = 1, \ldots, m$, where each $A_i \subseteq \Re^{n_i}$, $\mathcal{B}_i$ is the corresponding Borel σ-algebra, and μ_i is Lebesgue measure when A_i is uncountable and counting measure otherwise. Put $A = A_1 \times \cdots \times A_m$ and $A_{-i} = A_1 \times \cdots \times A_{i-1} \times A_{i+1} \times \cdots \times A_m$. Let x_i denote an element of A_i so that $\mathbf{x} = (x_1, \ldots, x_m)$ represents an element of A. Also, elements of A_{-i} will be written $\mathbf{x}_{-i} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_m)$.

Suppose that there are functions $f_i(x_i|x_{-i}) : A \to [0, \infty)$, $i = 1, \ldots, m$, such that for every $x_{-i} \in A_{-i}$, $f_i(\cdot|x_{-i})$ is a probability density function with respect to μ_i whose support set is A_i. The sets A_i are assumed fixed in that they may not depend on x_{-i}.

Example 3. Take $A_i = \Re$ and let $f_1, \ldots, f_m$ have the Gaussian forms

$$f_i(x_i|x_{-i}) = \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{1}{2}\left(x_i - p_i \sum_{j \neq i} x_j \right)^2 \right\}$$

where the p_i's are constants.

The question of interest is as follows: *Does there exists a joint probability density whose conditional densities are the f_i's?* We refer to $f_1, \ldots, f_m$ as candidate conditional densities. They are called *compatible* if there exists a function, $f(x_1, \ldots, x_m)$:

$A \to [0, \infty)$, which is a probability density with respect to the product measure $\mu = \mu_1 \times \cdots \times \mu_m$, having support set A, such that

$$\frac{f(x_1, \ldots, x_m)}{\int_{A_i} f(x_1, \ldots, x_m)\, \mu(dx_i)} = f_i(x_i | x_{-i}) \tag{2.1}$$

for $i = 1, \ldots, m$.

Arnold and Press (1989) give necessary and sufficient conditions for compatibility when $m = 2$ in the more general setting where the support sets of the candidate conditionals are not assumed fixed. In the remainder of this section, we consider the compatibility of $f_1, \ldots, f_m$.

2.2. Compatibility versus Functional Compatibility

We begin by defining *functional compatibility.*

Definition 1. Let $f_1, \ldots, f_m$ be the set of candidate conditional densities described above. If there exists a function $g(x_1, \ldots, x_m) : A \to [0, \infty)$ such that

$$\frac{g(x_1, \ldots, x_m)}{\int_{A_i} g(x_1, \ldots, x_m)\, \mu(dx_i)} = f_i(x_i | x_{-i}) \tag{2.2}$$

for $i = 1, \ldots, m$, then $f_1, \ldots, f_m$ are *functionally compatible.*

Functional compatibility is necessary, but not sufficient, for compatibility since g may not be a probability density. For instance, the function $g(x_1, x_2) = \exp(-x_1 x_2)$ generates the exponential conditionals discussed in the Introduction, but it is clearly not a probability density since its integral over the positive quadrant diverges. Note that "Gibbs conditionals" calculated via a proportionality, like (1.3), are functionally compatible. A necessary and sufficient condition for functional compatibility is now developed. The condition is constructive in that it gives the form of g in terms of $f_1, \ldots, f_m$.

Suppose, for a moment, that our candidate conditionals are compatible. Write the joint density as $f(x_1, \ldots, x_m)$. Besag (1974) shows that if $(x'_1, \ldots, x'_m)$ is any fixed point in A and $(l_1, \ldots, l_m)$ is any one the $m!$ permutations of $(1, 2, \ldots, m)$, then

$$f(x_1, \ldots, x_m) \propto \frac{\prod_{j=1}^{m} f_{l_j}\left(x_{l_j} | x_{l_1}, \ldots, x_{l_{j-1}}, x'_{l_{j+1}}, \ldots, x'_{l_m}\right)}{\prod_{j=2}^{m} f_{l_j}\left(x'_{l_j} | x_{l_1}, \ldots, x_{l_{j-1}}, x'_{l_{j+1}}, \ldots, x'_{l_m}\right)} \tag{2.3}$$

on A. (Note that $f_i(x_i | x_{-i}) > 0$ whenever $(x_1, \ldots, x_m) \in A$ so the denominator is never zero.) Thus, $f(x_1, \ldots, x_m)$ is unique when the candidate conditionals are compatible.

If, on the other hand, the compatibility of $f_1, \ldots, f_m$ is in question, the $m!$ versions of (2.3) can be constructed and compatibility ruled out if the ratio of any two is not

constant. For $i = 1, \ldots, m!$ let $l^i = (l_1^i, l_2^i, \ldots, l_m^i)$ represent the permutations of $(1, 2, \ldots, m)$. For fixed $(x_1', \ldots, x_m') \in A$, define

$$g_i(x_1, \ldots, x_m) = \frac{\prod_{j=1}^m f_{l_j^i}\left(x_{l_j^i} | x_{l_1^i}, \ldots, x_{l_{j-1}^i}, x'_{l_{j+1}^i}, \ldots, x'_{l_m^i}\right)}{\prod_{j=2}^m f_{l_j^i}\left(x'_{l_j^i} | x_{l_1^i}, \ldots, x_{l_{j-1}^i}, x'_{l_{j+1}^i}, \ldots, x'_{l_m^i}\right)}. \tag{2.4}$$

Theorem 1. *The candidate conditional densities $f_1, \ldots, f_m$ are functionally compatible if and only if for each fixed $(x_1', \ldots, x_m') \in A$, the ratio*

$$\frac{g_i(x_1, \ldots, x_m)}{g_j(x_1, \ldots, x_m)} \tag{2.5}$$

is constant for all $i \neq j$. Moreover, if they are functionally compatible, then any g_i will serve as g which is unique up to constant multiples.

Proof. First assume that $f_1, \ldots, f_m$ are functionally compatible. Consider some permutation l^i. Define the function $g^*\left(x_{l_1^i}, x_{l_2^i}, \ldots, x_{l_m^i}\right) = g(x_1, x_2, \ldots, x_m)$. Clearly

$$g_i(x_1, \ldots, x_m) \propto \frac{g^*\left(x_{l_1^i}, x'_{l_2^i}, \ldots, x'_{l_m^i}\right) g^*\left(x_{l_1^i}, x_{l_2^i}, x'_{l_3^i}, \ldots, x'_{l_m^i}\right) \cdots g^*\left(x_{l_1^i}, \ldots, x_{l_m^i}\right)}{g^*\left(x_{l_1^i}, x'_{l_2^i}, \ldots, x'_{l_m^i}\right) g^*\left(x_{l_1^i}, x_{l_2^i}, x'_{l_3^i}, \ldots, x'_{l_m^i}\right) \cdots g^*\left(x_{l_1^i}, \ldots, x_{l_{m-1}^i}, x'_{l_m^i}\right)}.$$

Thus, for every i, $g_i(x_1, \ldots, x_m) \propto g(x_1, \ldots, x_m)$ and the condition is satisfied.

Now assume that the condition is satisfied. Take any l^i and any fixed point $(x_1', \ldots, x_m') \in A$. It will be shown that g_i generates $f_1, \ldots, f_m$ as in (2.2). It is clear that

$$\frac{g_i(x_1, \ldots, x_m)}{\int_{A_{l_m^i}} g_i(x_1, \ldots, x_m)\ \mu_{l_m^i}(dx_{l_m^i})} = f_{l_m^i}\left(x_{l_m^i} | x_{-l_m^i}\right)$$

since $x_{l_m^i}$ appears only once in g_i. Let $u \in \{1, 2, \ldots, m-1\}$. Employing the condition of the theorem, we have

$$\int_{A_{l_u^i}} g_i(x_1, \ldots, x_m)\ \mu_{l_u^i}(dx_{l_u^i}) = c(x_1', \ldots, x_m') \int_{A_{l_u^i}} g_j(x_1, \ldots, x_m)\ \mu_{l_u^i}(dx_{l_u^i})$$

where $c(x_1', \ldots, x_m')$ is a constant and l^j is such that $l_m^j = l_u^i$. Now use the fact that

$$\int_{A_{l_u^i}} g_j(x_1, \ldots, x_m)\ \mu_{l_u^i}(dx_{l_u^i}) = \frac{g_i(x_1, \ldots, x_m)}{c(x_1', \ldots, x_m')\, f_{l_u^i}\left(x_{l_u^i} | x_{-l_u^i}\right)}$$

and the result follows. □

In terms of the Hammersley-Clifford Theorem (Besag 1974), functional compatibility is equivalent to having constructed the candidate conditional densities using appropriate "G-functions" without regard for the integrability condition.

Example 3 cont. Consider the case $m = 3$. Some simple calculations show that

$$g_i(x_1, x_2, x_3) \propto \exp\Big\{-\frac{1}{2}\Big[\sum_{j=1}^{3} x_{l_j^i}^2 - 2\Big(x_{l_1^i} x'_{l_2^i}\left(p_{l_1^i} - p_{l_2^i}\right) + x_{l_1^i} x'_{l_3^i}\left(p_{l_1^i} - p_{l_3^i}\right) + x_{l_2^i} x'_{l_3^i}\left(p_{l_2^i} - p_{l_3^i}\right) + p_{l_3^i} x_{l_3^i}\left(x_{l_1^i} + x_{l_2^i}\right) + p_{l_2^i} x_{l_1^i} x_{l_2^i}\Big)\Big]\Big\}.$$

Thus, for instance, if $l^1 = (1, 2, 3)$ and $l^2 = (1, 3, 2)$, we have

$$g_1/g_2 = \exp\left\{(p_2 - p_3)\left(x_2 x'_3 + x_3 x'_2 - x_2 x_3\right)\right\} \tag{2.6}$$

which is constant only if $p_2 = p_3$. Similar considerations lead to the conclusion that these three candidate conditionals are functionally compatible only when $p_i = p$. Analogously, $f_1, \ldots, f_m$ are functionally compatible only if $p_i = p$, $i = 1, \ldots, m$, and in that case

$$g(x_1, \ldots, x_m) \propto \exp\left\{-\frac{1}{2}\mathbf{x}'\mathbf{M}_m\mathbf{x}\right\} \tag{2.7}$$

where $\mathbf{M}_m = -p\mathbf{J}_m + (1+p)\mathbf{I}_m$ where $\mathbf{J}_m$ is an m-dimensional square matrix of 1's and $\mathbf{I}_m$ is an m-dimensional identity matrix.

If the compatibility of a general set $f_1, \ldots, f_m$ is in question, the first step is to check that they are functionally compatible using the condition in Theorem 1. If they are not functionally compatible, then they are not compatible. If they are functionally compatible, then they are compatible if and only if g is integrable. More formally, we have

Theorem 2. *The functionally compatible conditional densities $f_1, \ldots, f_m$ are compatible if and only if*

$$\int_{A_1} \cdots \int_{A_m} g(x_1, \ldots, x_m)\ \mu_m(dx_m) \cdots \mu_1(dx_1) < \infty.$$

Proof. If they are compatible then g must be proportional to the joint density. Conversely, if the integral is finite, then g is normalizable and compatibility follows. □

Example 3 cont. Assume that $p_i = p$ so that $f_1, \ldots, f_m$ are functionally compatible. According to Theorem 2, they are compatible if and only if (2.7) is integrable, which will be the case only if $\mathbf{M}_m$ is positive definite. Since the eigenvalues of $\mathbf{M}_m$ are $(1+p)$ and $1 - p(m-1)$, (2.7) will be integrable only when $p \in \left(-1, \frac{1}{m-1}\right)$ and in that case the joint density corresponding to $f_1, \ldots, f_m$ is an m-dimensional normal with mean $\mathbf{0}$ and covariance matrix $\mathbf{M}_m^{-1}$.

3. A Markov Chain

In this section, a Markov transition function is constructed using the functionally compatible conditional densities $f_1, \ldots, f_m$. The measure $\pi(S) = \int_S g$ is shown to be invariant for the associated Markov chain which, in light of Theorem 2, implies that the chain is positive recurrent if and only if $f_1, \ldots, f_m$ are compatible. Gibbs Markov chains corresponding to improper posterior distributions are a special case and are therefore null (not positive recurrent). We conclude with a general result for null chains which can be used to describe the limiting behavior of some standard Monte Carlo approximations.

3.1. Construction

Let $f_1, \ldots, f_m$ be a set of *continuous* functionally compatible conditional densities and let $\mathcal{B}$ represent the product σ-algebra corresponding to A. Consider the function $P : A \times \mathcal{B} \to [0, 1]$ given by

$$P(\mathbf{x}, S) = \tag{3.1}$$
$$\int_S f_1(t_1|x_2, \ldots, x_m) f_2(t_2|t_1, x_3, \ldots, x_m) \cdots f_m(t_m|t_1, \ldots, t_{m-1}) \mu(d(t_1, \ldots, t_m)).$$

For any $\mathbf{x} \in A$, $P(\mathbf{x}, \cdot)$ is a probability measure on $\mathcal{B}$. Also, for any $S \in \mathcal{B}$, $P(\cdot, S)$ is a lower semi-continuous function (see the Appendix), which implies that it's measurable (Billingsley 1986 p.188). Therefore, P is a *Markov transition function* (Meyn and Tweedie 1992, Chapter 3) which defines a discrete time, time homogeneous Markov chain $\Phi = \{\phi_0, \Phi_1, \Phi_2, \ldots\}$ on the product space A^∞. The initial state of the chain is $\Phi_0 = \phi_0$ and the transition probabilities are now briefly described. For any $i = 0, 1, 2, \ldots$, the conditional distribution of Φ_{i+1} given that $\Phi_i = \phi_i$ is $P(\phi_i, \cdot)$. For $n \geq 2$, define the *n-step Markov transition functions* inductively as

$$P^n(\mathbf{x}, S) = \int_A P(\mathbf{x}, d\mathbf{y}) P^{n-1}(\mathbf{y}, S).$$

For any $i = 0, 1, 2, \ldots$ and any $n = 2, 3, \ldots$, the conditional distribution of Φ_{i+n} given that $\Phi_i = \phi_i$ is $P^n(\phi_i, \cdot)$. Thus, for example, $P^n(\phi_0, S)$ is the probability that the chain is in the set S after the first n steps. The Markov chain Φ is μ-irreducible and aperiodic since the f_i's are strictly positive on A.

3.2. Positive Recurrence and Compatibility

Define a measure, $\pi(\cdot)$, on the measurable space $(A, \mathcal{B})$ using g of (2.2) as follows

$$\pi(S) = \int_S g(x_1, \ldots, x_m) \mu(d(x_1, \cdots, x_m)). \tag{3.2}$$

It is assumed throughout that $\pi(\cdot)$ is σ-finite.

Theorem 3. *The measure $\pi(\cdot)$ defined in (3.2) is a σ-finite invariant measure for Φ, that is, for any $S \in \mathcal{B}$*

$$\pi(S) = \int_A \pi(d\mathbf{x})\, P(\mathbf{x}, S). \tag{3.3}$$

Proof. We give a proof for m=2. Extension to the general case is straightforward.

$$\begin{aligned}
&\int_A \pi(d(x_1, x_2))\, P((x_1, x_2), S) \\
&= \int_{A_1}\int_{A_2}\left[\int_S f_1(t_1|x_2) f_2(t_2|t_1)\mu(d(t_1, t_2))\right] g(x_1, x_2)\, \mu(dx_2)\, \mu(dx_1) \\
&= \int_S \left[\int_{A_2}\int_{A_1} g(x_1, x_2)\, f_1(t_1|x_2) f_2(t_2|t_1)\mu(dx_1)\, \mu(dx_2)\right] \mu(d(t_1, t_2)) \\
&= \int_S \left[\int_{A_2} g(t_1, x_2)\, f_2(t_2|t_1)\mu(dx_2)\right] \mu(d(t_1, t_2)) \\
&= \int_S g(t_1, t_2)\, \mu(d(t_1, t_2)) \\
&= \pi(S)
\end{aligned} \tag{3.4}$$

where the third and fourth equalities follow from functional compatibility, that is, from (2.2). □

If π is finite, it is the unique (up to constant multiples) invariant measure and Φ is positive recurrent, otherwise Φ is null (Meyn and Tweedie 1993, p230). This fact, together with Theorems 2 and 3 give us the following result.

Theorem 4. *The Markov chain Φ is positive recurrent if and only if $f_1, \ldots, f_m$ are compatible.*

Although our main interest is in the chains resulting from incompatible f_i's, the well-known compatible case is discussed briefly for completeness. Tierney (1991) shows that if Φ is positive recurrent, and the probability measure $P(\mathbf{x}, \cdot)$ is absolutely continuous w.r.t. π for all $\mathbf{x} \in A$, then Φ is positive Harris recurrent. (Harris recurrence is stronger than recurrence: for any set $V \in \mathcal{B}$ such that $\mu(V) > 0$ and any starting point $\phi_0 \in A$, a Harris recurrent chain visits V an infinite number of times with probability one, while a recurrent chain has only an infinite expected number of visits to V.) Since the f_i's are all strictly positive on A, $\pi(S) = 0$ implies that $\mu(S) = 0$ for any $S \in \mathcal{B}$, which clearly implies that $P(\mathbf{x}, S) = 0$, no matter what the value of $\mathbf{x}$. Thus, if π is finite, Φ is positive Harris recurrent.

Assuming that π is finite, let $\pi'(\cdot) = \pi(\cdot)/\pi(A)$. Successful use of the Gibbs sampler relies on two facts about Φ which follow from positive Harris recurrence (Meyn and Tweedie 1992, Theorems 13.0.1 and 17.0.1). First, for any starting value

$\phi_0 \in A$, the probability measures given by $P^n(\phi_0, \cdot)$ converge in total variation to the probability measure π' as $n \to \infty$. This implies that the Φ_n converge in distribution to a random variable with distribution π'. Second, the law of large numbers holds, that is, if t is a real-valued function with domain A such that $\int |t(\mathbf{x})| \pi'(d\mathbf{x})$ is finite, then $\frac{1}{n}\sum_{i=1}^{n} t(\Phi_i) \to \int t(\mathbf{x})\, \pi'(d\mathbf{x})$ with probability one.

3.3. A General Result for Null Chains

Let $\Gamma = (\gamma_0, \Gamma_1, \Gamma_2, \ldots)$ be a Markov chain on a product space, A^∞, where A is a Euclidean space of the type described at the beginning of Section 2.1. Let R and $\mathbf{P}_{\gamma_0}$ denote the Markov transition function and the probability law for the entire chain, respectively. (We use Γ and R here to avoid confusion with Φ and P.) One definition is required before the result is stated. The chain, Γ, is called a *Feller chain* if $R(\cdot, S)$ is a lower semi-continuous function for every $S \in \mathcal{B}$.

Theorem 5. *Suppose that Γ is an aperiodic, μ-irreducible, null, Feller Markov chain where the support of μ has non-empty interior. If $t : A \to \Re_+$ is a bounded measurable function for which, given $\epsilon > 0$, there exists a compact set $C \in A$ such that $t(y) \leq \epsilon$ $\forall\, y \in C^c$, then*

$$\liminf_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} t(\Gamma_i) = 0 \quad a.s. \tag{3.5}$$

Proof. Choose $\epsilon \in (0,1)$ and let $C_1 \subset C_2 \subset \cdots$ be a sequence of compact sets in A such that $\gamma_0 \in C_1$ and such that $t(y) \leq \epsilon^j$ when $y \in C_j^c$. The conditions of the theorem imply that if $C \in A$ is a compact set containing γ_0, then $\lim_{n\to\infty} R^n(\gamma_0, C) = 0$ (Meyn and Tweedie 1993 pp. 127, 454). Furthermore, by the consistency of Cesaro summation (Billingsley 1986 p.572) we have $\lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} R^i(\gamma_0, C) = 0$. Thus, we may choose a subsequence, $\{n_j\}$, of the positive integers such that

$$\sum_{j=1}^{\infty} \left(\frac{1}{n_j} \sum_{i=1}^{n_j} R^i(\gamma_0, C_j) \right) < \infty.$$

It will be shown that

$$\lim_{j\to\infty} \frac{1}{n_j} \sum_{i=1}^{n_j} t(\Gamma_i) = 0 \quad \text{a.s.}$$

According to the first Borel-Cantelli Lemma, it is enough to show that for any $\delta > 0$,

$$\sum_{j=1}^{\infty} \mathbf{P}_{\gamma_0} \left(\frac{1}{n_j} \sum_{i=1}^{n_j} t(\Gamma_i) > \delta \right) < \infty.$$

Let M be an upper bound for t. Since $t \le MI_{C_j} + \epsilon^j I_{C_j^c}$ for any j (I is an indicator), we have

$$\begin{aligned}
\sum_{j=1}^{\infty} \mathbf{P}_{\gamma_0}\left(\frac{1}{n_j}\sum_{i=1}^{n_j} t(\Gamma_i) > \delta\right) &\le \sum_{j=1}^{\infty} \mathbf{P}_{\gamma_0}\left(\frac{1}{n_j}\sum_{i=1}^{n_j}\left(MI_{C_j}(\Gamma_i) + \epsilon^j I_{C_j^c}(\Gamma_i)\right) > \delta\right) \\
&\le \frac{1}{\delta}\sum_{j=1}^{\infty}\frac{1}{n_j}\sum_{i=1}^{n_j} E\left(MI_{C_j}(\Gamma_i) + \epsilon^j I_{C_j^c}(\Gamma_i)\right) \\
&\le \frac{M}{\delta}\sum_{j=1}^{\infty}\frac{1}{n_j}\sum_{i=1}^{n_j} R^i(\gamma_0, C_j) + \frac{1}{\delta}\sum_{j=1}^{\infty}\frac{1}{n_j}\sum_{i=1}^{n_j}\epsilon^j \\
&< \infty
\end{aligned}$$

where the second step follows from Markov's inequality. □

This result can be used to demonstrate that many standard Monte Carlo approximations used in Gibbs sampling have undesirable limiting behavior when the posterior is improper. The example developed in the next section shows that these undesirable properties are not always apparent from the Gibbs output.

4. A Gibbs Sampling Application

In this section we discuss a Bayesian hierarchical version of the one-way random effects model which has an improper posterior. This model is similar to the hierarchical model in Example 2 in that if one assumes that the posterior is proper, the Gibbs conditionals and hence a Gibbs chain may be constructed. Unlike the Gibbs chain in Example 2, however, this chain is (seemingly) well-behaved and provides no warning that the posterior is improper. The results from the previous sections are used to demonstrate that, although they may seem well-behaved, the Monte Carlo approximations constructed using this Gibbs chain have undesirable limiting behavior.

Although the model discussed in this section is quite simplistic, it is a special case of a hierarchical linear mixed model (Hobert and Casella 1994) which may possess an improper posterior depending on which improper priors are placed on the variance components. Models of this type possessing improper posteriors have been employed in the literature (see the references below) and this example is therefore of practical as well as theoretical interest.

Consider the simple one-way random effects model

$$y_{ij} = \beta + u_i + \epsilon_{ij} \tag{4.1}$$

where $i = 1, 2, \ldots, k$, $j = 1, 2, \ldots, J$. It is assumed that the u_i's (the random effects) are iid $\mathrm{N}(0, \sigma^2)$ and the ϵ_{ij}'s (white noise) are iid $\mathrm{N}(0, \sigma_\epsilon^2)$. The u_i's and ϵ_{ij}'s are assumed independent. The overall mean, β, and the variance components, σ^2 and σ_ϵ^2, are considered unknown parameters.

This frequentist model fits nicely into a Bayesian conditionally independent hierarchical model (Kass and Steffey 1989) by writing (4.1) as a two stage hierarchy and specifying priors on the three unknown parameters

$$\begin{gathered} y_{ij}|\beta, \mathbf{u}, \sigma_\epsilon^2 \sim \mathrm{N}(\beta + u_i, \sigma_\epsilon^2) \\ \beta \sim \pi(\beta) \quad \mathbf{u}|\sigma^2 \sim \mathrm{N}_k(\mathbf{0}, \mathbf{I}\sigma^2) \quad \sigma_\epsilon^2 \sim \pi\left(\sigma_\epsilon^2\right) \\ \sigma^2 \sim \pi\left(\sigma^2\right) \end{gathered} \tag{4.2}$$

where $\mathbf{u}' = (u_1, \ldots, u_k)$ and the priors $\pi(\beta)$, $\pi\left(\sigma_\epsilon^2\right)$ and $\pi\left(\sigma^2\right)$ must be elicited. (It is often assumed that the variance components are not independent *a priori*, that is, $\pi\left(\sigma_\epsilon^2\right)$ is often allowed to depend on σ^2. The Gibbs sampler is more difficult to implement in these situations, however, because simulating from the "Gibbs conditionals" is not easy. Lehmann (1983, p.248) and Chaloner (1987) both discuss such models and give further references.)

A specific example of model (4.2) discussed by Hill (1965) and Tiao and Tan (1965) has $\pi(\beta) \propto 1$, $\pi\left(\sigma_\epsilon^2\right) \propto 1/\sigma_\epsilon^2$ and $\pi\left(\sigma^2\right) \propto 1/\sigma^2$ where the last two are restricted to $\Re_+$. Hill (1965) shows that the posterior distribution corresponding to this model is improper. If, however, propriety of the posterior were incorrectly assumed, as was done for similar models in Gelfand *et al.* (1990, Model I, Section 4) and Wang *et al.* (1993, p.44), then the "Gibbs conditionals" could be computed (see Example 2) and the result is

$$\begin{gathered} f_i\left(u_i|\mathbf{u}_{-i}, \beta, \sigma_\epsilon^2, \sigma^2, \mathbf{y}\right) = \mathrm{N}\left(\tfrac{\sigma^2}{J\sigma^2+\sigma_\epsilon^2}\left(y_{i\cdot} - J\beta\right), \tfrac{\sigma_\epsilon^2\sigma^2}{J\sigma^2+\sigma_\epsilon^2}\right) \quad i = 1, \ldots, k \\ f_{k+1}\left(\beta|\mathbf{u}, \sigma_\epsilon^2, \sigma^2, \mathbf{y}\right) = \mathrm{N}\left(\overline{y}_{\cdot\cdot} - \overline{u}_\cdot, \tfrac{\sigma_\epsilon^2}{Jk}\right) \\ f_{k+2}\left(\sigma_\epsilon^2|\mathbf{u}, \beta, \sigma^2, \mathbf{y}\right) = \mathrm{IG}\left(Jk/2, 2\left(\textstyle\sum_i \sum_j \left(y_{ij} - \beta - u_i\right)^2\right)^{-1}\right) \\ f_{k+3}\left(\sigma^2|\mathbf{u}, \beta, \sigma_\epsilon^2, \mathbf{y}\right) = \mathrm{IG}\left(k/2, 2\left(\mathbf{u}'\mathbf{u}\right)^{-1}\right) \end{gathered} \tag{4.3}$$

where $\overline{y}_{\cdot\cdot} = \sum_{i,j} y_{ij}/Jk$, $y_{i\cdot} = \sum_j y_{ij}$, $\overline{u}_\cdot = \sum_i u_i/k$, IG stands for inverted gamma, and $\mathbf{y}$ represents the data. (We say $X \sim \mathrm{IG}\left(a, b\right)$ if it has support $\Re_+$ and $f_X\left(x\right) \propto \left[x^{a+1}\exp\left(1/xb\right)\right]^{-1}$.) Thus, $f_1, \ldots, f_{k+3}$ are a set of continuous functionally compatible conditional densities. They are not compatible, however, since the posterior is improper (see Theorem 2). Theorem 4 tells us that the Markov chain, Φ, constructed using $f_1, \ldots, f_{k+3}$ is null, that is, the Gibbs chain is null.

As mentioned above, this is an example of a situation in which the Gibbs output does not provide a "red flag" informing us that the posterior is improper. Suppose that we are under the impression that the posterior corresponding to the model (4.2) is proper and that we have data for which this model is appropriate. It is desired to simulate from the posterior (using the Gibbs algorithm) and construct Monte Carlo estimates of (1) $f_{\sigma^2|\mathbf{y}}\left(\cdot|\mathbf{y}\right)$, the marginal posterior density of σ^2, and (2) $E^\pi I_{[1,2]}\left(\beta\right)$, the posterior probability that β is in the interval $[1, 2]$. Write the Gibbs chain as

$$\Phi_i = \left[\mathbf{u}^{(i)}, \beta^{(i)}, \sigma_\epsilon^{2(i)}, \sigma^{2(i)}\right] \quad i = 0, 1, 2, \ldots \tag{4.4}$$

where the zeros indicate starting values. We might approximate $f_{\sigma^2|\mathbf{y}}(\cdot|\mathbf{y})$ at the point a using

$$\hat{f}_{\sigma^2|\mathbf{y}}(a|\mathbf{y}) = \frac{1}{n}\sum_{i=b}^{b+n} f_{k+3}\left(a|\mathbf{u}^{(i)}, \beta^{(i)}, \sigma_\epsilon^{2(i)}, \mathbf{y}\right) \tag{4.5}$$

and $E^\pi I_{[1,2]}(\beta)$ using (Liu, Wong and Kong 1994)

$$\frac{1}{n}\sum_{i=b}^{b+n}\int_{[1,2]} \frac{1}{\sqrt{2\pi\sigma_\epsilon^{2(i)}/Jk}} \exp\left(-\frac{Jk}{2\sigma_\epsilon^{2(i)}}\left(t - \overline{y}_{\cdot\cdot} - \overline{u}_{\cdot}^{(i)}\right)^2\right) dt \tag{4.6}$$

where b is the "burn-in" and n is "large."

Before considering the limiting behavior of these approximations, we give an example of how well-behaved they appear. Figure 3 shows the pointwise estimate $\hat{f}_{\sigma^2|\mathbf{y}}(a|\mathbf{y})$ (b=15,000, n=1,000) from a realization of (4.4) based on data simulated using $i = 7$, $j = 5$, $\beta = 10$, $\sigma^2 = 5$, and $\sigma_\epsilon^2 = 2$. A histogram of $\sigma^{2(j+15{,}000)}$, $j = 1, \ldots, 1000$, is shown in the same figure. Note that the density approximation and histogram appear reasonable and in no way warn the user of an improper posterior. One might believe that the chain would eventually misbehave if it were allowed to run for a long time, but this is not the case. Some of these chains were run for millions of iterations and never misbehaved.

Theorem 5 shows that for any point a,

$$\liminf_{n\to\infty} \frac{1}{n}\sum_{i=b}^{b+n} f_{k+3}\left(a|\mathbf{u}^{(i)}, \beta^{(i)}, \sigma_\epsilon^{2(i)}, \mathbf{y}\right) = 0 \text{ a.s.}$$

Thus, at each point, the Monte Carlo approximation has an almost sure limit of zero or none at all. Similarly

$$\liminf_{n\to\infty} \frac{1}{n}\sum_{i=b}^{b+n}\int_{[1,2]} \frac{1}{\sqrt{2\pi\sigma_\epsilon^{2(i)}/Jk}} \exp\left(-\frac{Jk}{2\sigma_\epsilon^{2(i)}}\left(t - \overline{y}_{\cdot\cdot} - \overline{u}_{\cdot}^{(i)}\right)^2\right) dt = 0 \text{ a.s.}$$

There are many approximations to which Theorem 5 does not apply. For example, a more intuitive approximation of $E^\pi I_{[1,2]}(\beta)$, which is sometimes more variable than (4.6) (Liu, Wong and Kong 1994), is

$$\frac{1}{n}\sum_{i=b}^{b+n} I_{\Re^k\times[1,2]\times\Re_+\times\Re_+}(\Phi_i). \tag{4.7}$$

Theorem 5 cannot be applied to this approximation because the indicator function does not satisfy the necessary conditions. On the other hand, if the indicator in (4.7) were replaced with $I_{[-M,M]^k\times[1,2]\times[M^{-1},M]\times[M^{-1},M]}(\Phi_i)$ where M is some large, positive number, the approximation would be practically the same, and Theorem 5 could be applied.

5. Concluding Remarks

When Bayesian hierarchical models with improper priors are employed, the high-dimensional integration required to calculate posterior quantities of interest is often extremely difficult. The ability to use the Gibbs sampler in these situations is usually a blessing, but may be a curse. Sometimes a perfectly good set of "Gibbs conditionals" may be calculated from a hierarchical model with an improper posterior distribution. Since demonstrating propriety of the posterior is not a necessary step in calculating the "Gibbs conditionals" (and usually involves the same complicated integration that one is avoiding by using the Gibbs sampler), the experimenter might simply assume propriety and use the Gibbs sampler to calculate the "posterior quantities of interest." The problem is that the resulting Gibbs output may appear perfectly reasonable (see Section 4) which could lead to inferences about a nonexistent posterior distribution.

This paper contains some general theory which can be used to characterize the behavior of "improper Gibbs" chains, that is, Gibbs Markov chains constructed using "Gibbs conditionals" associated with an improper posterior. We have generalized Arnold and Press's (1989) notion of *compatibility* by calling conditional densities *functionally compatible* if there exists a positive function, g, which behaves as the joint density function in every way except that it need not be integrable. Theorem 1 gives a necessary and sufficient condition for functional compatibility as well as the form of g (when it exists). "Gibbs conditionals" corresponding to improper posteriors are *functionally compatible* due to the manner in which they are constructed. This implies that "improper" Gibbs chains are special cases of the chain defined by the Markov transition function in Section 3 and are thus <u>not</u> positive recurrent, i.e., they are null (either transient or null recurrent).

Sometimes when an "improper" Gibbs chain is simulated, the output appears "out of control" (see Example 2) and therefore warns the user that there is a problem. The danger occurs when "improper" Gibbs chains produce nice looking output either because they are "almost" positive recurrent (like a chain constructed with the normal conditionals in Example 3 with $p = 1/(m-1)$) or because they "get stuck" in a "nice" part of the space (Geyer 1992). Our results show that although some "improper" Monte Carlo approximations may appear reasonable, they either have an almost certain limit of zero or none at all.

Ideally, a hierarchical model (with improper priors) should always be shown to possess a proper posterior distribution before being used as a model for data. However, for many hierarchical models, demonstrating propriety is extremely difficult, while employing the Gibbs sampler is almost trivial. Thus, the ability to use the Gibbs output to diagnose positive recurrence (propriety) would be useful. One such diagnostic, described in Hobert (1994), is based on the fact that an infinite mean return time (to a compact set) implies that the Gibbs chain is null, i.e., that the posterior is improper. Independent Gibbs chains are used to collect a random sample of return times (to some arbitrary compact set) and the technique suggested by Hill

(1975) is used to decide if the return time distribution has an infinite mean or not. Unfortunately, this technique seems to be effective in detecting improper posteriors only in cases where the chain is clearly out of control.

There is an important distinction between the diagnostics for positive recurrence and the so-called "convergence diagnostics" proposed in the MCMC literature (see, for example, Robert 1993, Tanner 1993, p.114, Gelman and Rubin 1992, Roberts 1992, and Raftery and Banfield 1991). The latter *assume* that the chain is positive recurrent and use the output to provide information about when Monte Carlo approximations are "close enough" to the true values. They are not designed to detect *if* the Gibbs chain converges (positive recurrence), nor even *when* the Gibbs chain has converged; it never does. Thus, one should not count on "convergence diagnostics" to detect an improper posterior.

There are many Monte Carlo approximations which are Cesaro averages of functions which do not satisfy the conditions of Theorem 5. Although our intuition tells us that many of these approximations should also have undesirable limiting behavior, our results do not apply. Results describing the limiting behavior of averages of functions which do not satisfy the "arbitrarily small off of compact sets" condition of Theorem 5, (like the indicator function in (4.7)) would clearly be useful.

References

Arnold, B. C., and Press, S. J. (1989), "Compatible Conditional Distributions," *Journal of the American Statistical Association,* **84**, 152–156.

Bartle, R. G. (1976), *The Elements of Real Analysis* (2nd ed.), New York: Wiley.

Berger, J. O. (1985), *Statistical Decision Theory and Bayesian Analysis* (2nd ed.), New York: Springer-Verlag.

Besag, J. (1974), "Spatial Interaction and the Statistical Analysis of Lattice Systems," *Journal of the Royal Statistical Society*: Ser. B, **36**, 192–236.

Billingsley, P. (1986), *Probability and Measure*, New York: Wiley.

Casella, G., and George, E. I. (1992), "Explaining the Gibbs Sampler," *The American Statistician,* **46**, 167–174.

Chaloner, K. (1987), "A Bayesian Approach to the Estimation of Variance Components for the Unbalanced One-Way Random Model," *Technometrics,* **29**, 323–337.

Cowles, M. K., and Carlin, B. P. (1994), "Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review," Technical Report # 94-008, Division of Biostatistics, University of Minnesota, Minneapolis, MN, 55455, USA.

Gelfand, A. E., and Smith, A. F. M. (1990), "Sampling Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association,* **85**, 398–409.

Gelfand, A. E., Hills, S. E., Racine-Poon, A., and Smith, A. F. M. (1990), "Illustration of Bayesian Inference in Normal Data Models Using Gibbs Sampling," *Journal of the American Statistical Association,* **85**, 972–985.

Gelman, A., and Rubin, D. B. (1992), "Inference from Iterative Simulation Using Multiple Sequences," *Statistical Science,* **7**, 457–511.

Gelman, A., and Speed, T. P. (1993), "Characterizing a Joint Probability Distribution by Conditionals," *Journal of the Royal Statistical Society*: Ser. B, **55**, 185–188.

Geyer, C. J. (1992), "Practical Markov Chain Monte Carlo," *Statistical Science,* **7**, 473-511.

Gourieroux, C., and Monfort, A. (1979), "On the Characterization of a Joint Probability Distribution by Conditional Distributions," *Journal of Econometrics* **10**, 115–118.

Hill, B. M. (1965), "Inference about Variance Components in the One-Way Model," *Journal of the American Statistical Association,* **60**, 806–825.

Hill, B. M. (1975), "A Simple General Approach to Inference about the Tail of a Distribution," *The Annals of Statistics,* **3**, 1163–1174.

Hobert, J. P. (1994), "Occurrences and Consequences of Nonpositive Markov Chains in Gibbs Sampling," Ph.D. Thesis, Cornell University, Ithaca.

Hobert, J. P., and Casella, G. (1994), "The Effect of Improper Priors on Gibbs Sampling in Hierarchical Linear Mixed Models," Technical Report #469, Statistics Department, University of Florida, Gainesville, Florida, 32611, USA.

Kass, R. E., and Steffey, D. (1989), "Approximate Bayesian Inference in Conditionally Independent Hierarchical Models (Parametric Empirical Bayes Models)," *Journal of the American Statistical Association* **84**, 717–726.

Lehmann, E. L. (1983), *Theory of Point Estimation,* Pacific Grove, CA: Wadsworth.

Liu, J., Wong, W. H., and Kong, A. (1994), "Covariance Structure of the Gibbs Sampler with applications to the comparisons of Estimators and Augmentation Schemes," *Biometrika,* **81**, 27–40.

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

Raftery, A. E., and Banfield, J. D. (1991), "Stopping the Gibbs Sampler, the Use of Morphology, and Other Issues in Spatial Statistics," *Annals of the Institute of Statistical Mathematics,* **43**, 32–43.

Robert, C. P. (1993), "Convergence Assessments for Markov Chain Monte Carlo Methods," Technical Report, Dept. de Math., Univ. de Rouen, France.

Roberts, G. O. (1992), "Convergence Diagnostics of the Gibbs Sampler," In *Bayesian Statistics 4* (eds. J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), Oxford: Oxford University Press, 775–782.

Tanner, M. A. (1993), *Tools for Statistical Inference,* New York: Springer-Verlag.

Tiao, G. C., and Tan, W. Y. (1965), "Bayesian Analysis of Random-Effect Models in the Analysis of Variance. I. Posterior Distribution of Variance Components," *Biometrika,* **52**, 37–53.

Tierney, L. (1995), "Markov Chains for Exploring Posterior Distributions," To Appear: *The Annals of Statistics.*

Wang, C. S., Rutledge, J. J., and Gianola, D. (1993), "Marginal Inferences about Variance Components in a Mixed Linear Model using Gibbs Sampling," *Genetique, Selection, Evolution,* **25**, 41–62.

6. Appendix

Recall (Bartle 1976 p.180) that $P(\cdot, S)$ is lower semi-continuous at the point $\mathbf{x}^* \in A$ if

$$\liminf_{\mathbf{x}\to\mathbf{x}^*} P(\mathbf{x}, S) \geq P(\mathbf{x}^*, S)$$

where

$$\liminf_{\mathbf{x}\to\mathbf{x}^*} P(\mathbf{x}, S) = \lim_{r\to 0} \inf \{P(\mathbf{x}, S) : 0 < \|\mathbf{x} - \mathbf{x}^*\| < r, \mathbf{x} \in A\}.$$

Lemma 1. *For any $S \in \mathcal{B}$ and any sequence $\mathbf{x}_n \in A$ such that $\mathbf{x}_n \to \mathbf{x}^*$*

$$\liminf_{n\to\infty} P(\mathbf{x}_n, S) \geq P(\mathbf{x}^*, S).$$

Proof. Write the integrand in (3.1) as $k(\mathbf{t}, \mathbf{x})$. Define $f_n(\mathbf{t}) = k(\mathbf{t}, \mathbf{x}_n)$. By the continuity of the conditional densities, we have $f_n(\mathbf{t}) \to k(\mathbf{t}, \mathbf{x}^*)$ for all $\mathbf{t}$ and the result follows by Fatou's Lemma. □

Theorem 6. *For $S \in \mathcal{B}$, $P(\cdot, S)$, is lower semi-continuous.*

Proof. Define the following notation

$$\phi(r) = \inf \{P(\mathbf{x}, S) : 0 < \|\mathbf{x} - \mathbf{x}^*\| < r, \mathbf{x} \in A\}.$$

First, if $\mathbf{x}^*$ is not a limit point of A, then the theorem is trivial, so assume $\mathbf{x}^*$ is a limit point and that the theorem is false, that is,

$$\liminf_{\mathbf{x}\to\mathbf{x}^*} P(\mathbf{x}, S) = l < P(\mathbf{x}^*, S).$$

Let $r_n \to 0$. Then $\lim_{n\to\infty} \phi(r_n) = l$ and for each $k = 1, 2, \ldots$, there exists an N_k s.t. $|\phi(r_n) - l| < 2^{-k}$ whenever $n \geq N_k$. We may clearly assume that $N_{k+1} > N_k$. Let $\mathbf{x}_k \in \{\mathbf{x} : 0 < \|\mathbf{x} - \mathbf{x}^*\| < r_{N_k}, x \in A\}$ be such that $|P(\mathbf{x}_k, S) - \phi(r_{N_k})| < 2^{-k}$. Then $\mathbf{x}_k \to \mathbf{x}^*$ and $|P(\mathbf{x}_k, S) - l| < 2^{1-k}$, but this contradicts Lemma 1. □

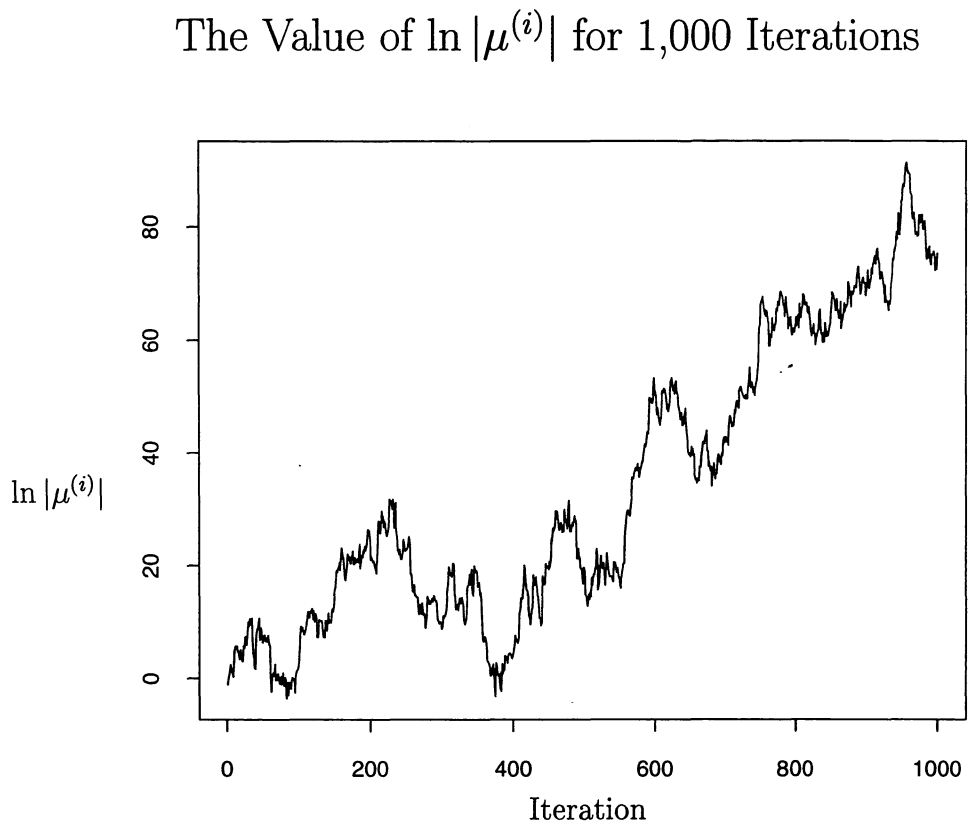

Figure 1: The natural logarithm of the absolute value of the $\mu^{(i)}$'s versus i for the first 1,000 iterations of a Gibbs chain. The data, (y_1, y_2, y_3), were realizations of independent standard normals. The densities used to build the chain were $\mu|\sigma^2 \sim \mathrm{N}(\overline{y}, \sigma^2/3)$ and $\sigma^2|\mu \sim \mathrm{IG}\left(1/2, 2\left(\sum (y_i - \mu)^2\right)^{-1}\right)$.

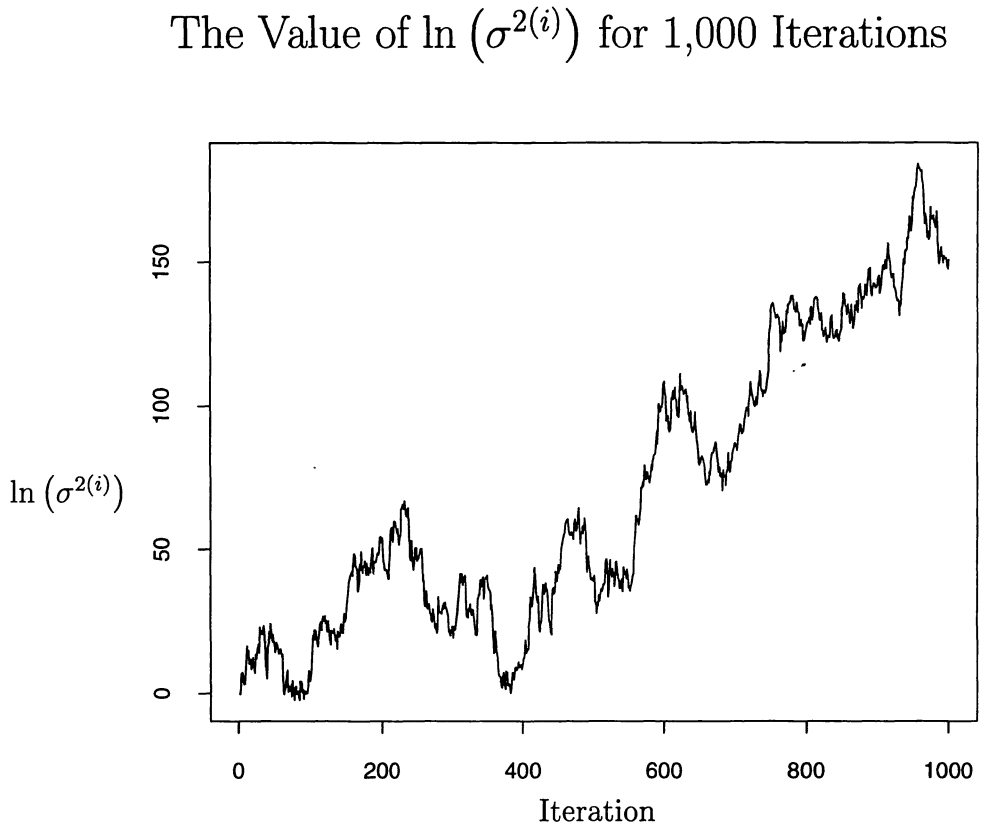

Figure 2: The natural logarithm of the $\sigma^{2(i)}$'s versus i for the first 1,000 iterations of a Gibbs chain. The data, (y_1, y_2, y_3), were realizations of independent standard normals. The densities used to build the chain were $\mu|\sigma^2 \sim \mathrm{N}\left(\overline{y}, \sigma^2/3\right)$ and $\sigma^2|\mu \sim \mathrm{IG}\left(1/2, 2\left(\sum (y_i - \mu)^2\right)^{-1}\right)$.

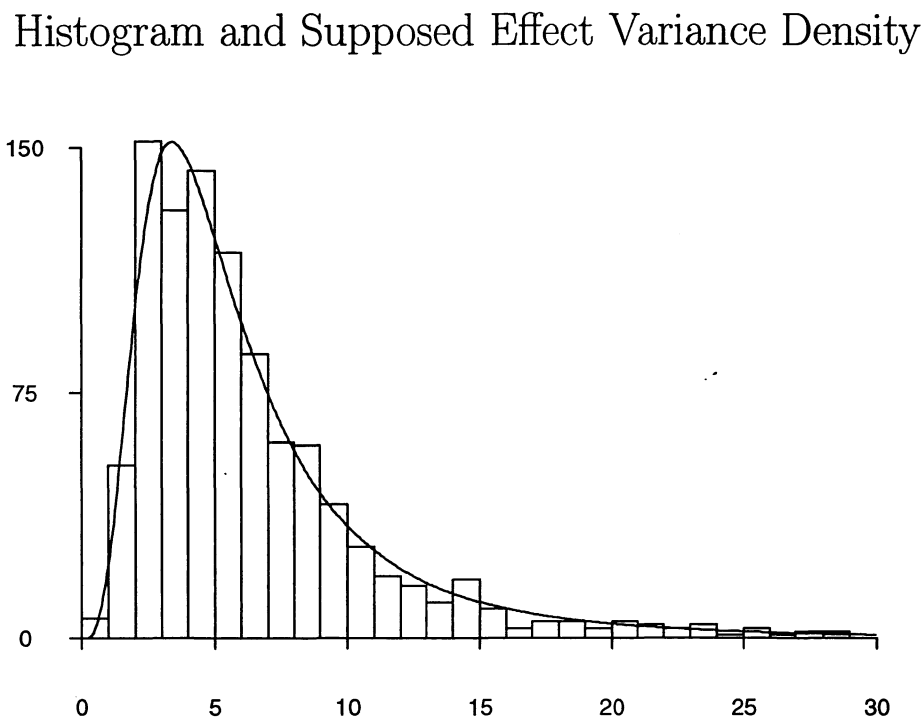

Figure 3: Histogram of the 1000 values of the effect variance from the null Gibbs chain, that is, a histogram of $\sigma^{2(j+15,000)}$ for $j = 1, 2, \ldots, 1000$. Superimposed is the approximate (supposed) marginal posterior density of σ^2. An appropriately scaled version of $\hat{f}_{\sigma^2|\mathbf{y}}(a|\mathbf{y})$ is on the ordinate with a on the abscissa. (Actually, eight of the 1,000 values of the effect variance, ranging from 38.1 to 169.7, were not included in the histogram.)