The Effect of Improper Priors on Gibbs Sampling in Hierarchical Linear Mixed Models

James P. Hobert*
Department of Statistics
University of Florida
Gainesville, FL 32611

George Casella
Biometrics Unit
Cornell University
Ithaca, NY 14853

*James P. Hobert is Assistant Professor, Department of Statistics, University of Florida, Gainesville, FL 32611. George Casella is Liberty Hyde Bailey Professor of Biological Statistics, Biometrics Unit, Cornell University, Ithaca, NY 14853. This research was supported by National Institute of Health Sciences Training Grant EHS-5-T32-ES07261-03 and National Science Foundation Grants DMS-9305547 and INT-9216784. The authors are grateful to Chuck McCulloch, Brett Presnell, Christian Robert, Marty Wells, the associate editor, and three referees for helpful comments and suggestions.

Abstract

Often, either from a lack of prior information or simply for convenience, variance components are modeled with improper priors in hierarchical linear mixed models. Although the posterior distributions for these models are rarely available in closed form, the usual conjugate structure of the prior specification allows for painless calculation of the Gibbs conditionals. Thus, the Gibbs sampler may be employed to explore the posterior distribution without ever having established propriety of the posterior. An example is given which shows that the output from a Gibbs chain corresponding to an improper posterior may appear perfectly reasonable. Thus, one can not expect the Gibbs output to provide a "red flag," informing the user that the posterior is improper. The user must demonstrate propriety before a Markov chain Monte Carlo technique is used. A theorem is given which classifies improper priors according to the propriety of the resulting posteriors. Applications concerning Bayesian analysis of animal breeding data and the location of maxima of unwieldy (restricted) likelihood functions are discussed. Gibbs sampling with improper posteriors is then considered in more generality. The concept of functional compatibility of conditional densities is introduced and is used to construct an invariant measure for a class of Markov chains. These results are used to show that Gibbs chains corresponding to improper posteriors are, in theory, quite ill behaved.

KEYWORDS: Animal Breeding; Compatibility; Functional Compatibility; Improper Posterior; Markov Chain; Monte Carlo; Variance Components.

1. INTRODUCTION

Posterior distributions corresponding to hierarchical linear mixed models are usually unavailable in closed form, even when conjugate priors are used. The Gibbs sampler (Geman and Geman 1984; Gelfand and Smith 1990) is an easily implemented simulation technique which allows one to draw random variables from the posterior distribution. Realizations of these random variables can be used to form Monte Carlo approximations of many features of the (joint) posterior, including marginal posterior densities of the variance components.

The hyperparameters in these models, because of either convenience or a lack of prior information, are often modeled with improper priors. Unfortunately, the same mathematical intractability that necessitates use of the Gibbs sampler, also makes demonstrating propriety of the posterior distribution a difficult task. This difficulty can lead to the use of the Gibbs sampler when the posterior distribution is improper. There are many examples of this in the (statistical and other) literature (see Sections 3 and 4).

To solidify these ideas, consider the simple one-way random effects model

$$y_{ij} = \beta + u_i + \epsilon_{ij} \qquad i = 1, 2, \ldots, k \qquad j = 1, 2, \ldots, J \tag{1.1}$$

where it is assumed that the u_i's (the random effects) are iid $N(0, \sigma^2)$ and the ϵ_{ij}'s (white noise) are iid $N(0, \sigma_\epsilon^2)$. 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 model fits nicely into a (Bayesian) conditionally independent hierarchical model (Kass and Steffey 1989) by writing (1.1) as a two stage hierarchy and specifying priors on the three unknown parameters

$$\begin{gathered} y_{ij}|\beta, \mathbf{u}, \sigma_\epsilon^2 \sim N(\beta + u_i, \sigma_\epsilon^2) \\ \beta \sim \pi(\beta) \qquad \mathbf{u} \sim N_k\left(\mathbf{0}, \mathbf{I}\sigma^2\right) \qquad \sigma_\epsilon^2 \sim \pi\left(\sigma_\epsilon^2\right) \\ \sigma^2 \sim \pi\left(\sigma^2\right) \end{gathered} \tag{1.2}$$

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.

The usual technique for calculating the conditional densities required for Gibbs sampling (Gibbs conditionals) is to use the conditional independence in model (1.2) to write the posterior density as

$$\pi\left(\sigma^2, \sigma_\epsilon^2, \mathbf{u}, \beta|\mathbf{y}\right) \propto f\left(\mathbf{y}|\beta, \mathbf{u}, \sigma_\epsilon^2\right) f\left(\mathbf{u}|\sigma^2\right) \pi(\beta) \pi\left(\sigma_\epsilon^2\right) \pi\left(\sigma^2\right) \tag{1.3}$$

and then pick the functional forms of each of the necessary conditionals off the right-hand side of (1.3). For the benefit of those unfamiliar with the tricks of Markov chain Monte Carlo (MCMC), we describe this in a bit more detail. Technically, when improper priors are used, the posterior density is not a conditional probability density. (See Berger (1985, p. 132) for an explanation and Hartigan (1983, Chap. 3) for a different point of view.) However, in a Bayesian analysis the function $\pi\left(\sigma^2, \sigma_\epsilon^2, \mathbf{u}, \beta|\mathbf{y}\right)$ is considered the conditional density of the parameters given the data, and is used as a joint density in the parameters which happens to involve the fixed, known quantity $\mathbf{y}$. Thus, for example, $f\left(\sigma^2|\sigma_\epsilon^2, \mathbf{u}, \beta, \mathbf{y}\right)$ is given by $\pi\left(\sigma^2, \sigma_\epsilon^2, \mathbf{u}, \beta|\mathbf{y}\right) / \int \pi\left(\sigma^2, \sigma_\epsilon^2, \mathbf{u}, \beta|\mathbf{y}\right) d\sigma^2$ and since the denominator is constant with respect to σ^2, as a function of σ^2, $f\left(\sigma^2|\sigma_\epsilon^2, \mathbf{u}, \beta, \mathbf{y}\right)$ is proportional to the right-hand side of (1.3). Each of the Gibbs conditionals can be computed in this manner.

A specific example of model (1.2) 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 $\pi\left(\sigma^2\right) \propto 1/\sigma^2$. Suppose that we have data for which this model is appropriate and we wish to use the Gibbs sampler to make inferences about the posterior distribution. Proportionality (1.3) is used to calculate the Gibbs conditionals, as described above, and we find that $(\sigma^2|\text{all others})$ and $(\sigma_\epsilon^2|\text{all others})$ have inverted gamma forms and $(\mathbf{u}|\text{all others})$ and $(\beta|\text{all others})$ have normal forms. One may conclude from this that the Gibbs sampler could be used to construct a Markov chain whose stationary distribution is the posterior distribution. Hill (1965), however, shows that the posterior for this model is improper (and suggests several alternatives including Jeffreys's prior). Thus, although we used it to calculate the Gibbs conditionals, (1.3) is meaningless because the right-hand side is not integrable. Impropriety implies that there does not exist a joint density to which the Gibbs conditionals correspond, and they are therefore called incompatible conditional densities (Arnold and Press 1989).

This example suggests the two important questions that are addressed in this paper: 1. Which improper priors yield proper posteriors in the general hierarchical linear mixed model? and 2. In general, what can be said about Gibbs Markov chains that correspond to improper posteriors?

A hierarchical linear mixed model with parametric improper priors is discussed in Section 2 where a theorem is given which classifies the improper priors according to the propriety of the resulting posterior distributions. This theorem is similar, in spirit, to those given in Ibrahim and Laud (1991), who consider the use of Jeffreys's prior in generalized linear models (GLM's), Dey, Gelfand and Peng (1994), who discuss the use of improper priors in overdispersed GLM's, and Natarajan and McCulloch (1995), which deals with mixed models for binomial responses. Another related paper is Zeger and Karim (1991) who discuss the use of improper priors and Gibbs sampling in GLM's. Section 3 of the present paper considers two applications of the theorem, one involving the Bayesian analysis of animal breeding data, and the other, finding maximum likelihood and restricted maximum likelihood estimates when closed form solutions are unavailable.

A general discussion of Gibbs sampling when the posterior is improper is given in Section 4. An invariant measure is constructed for the class of Markov chains generated using the Gibbs algorithm in conjunction with a set of functionally compatible conditional densities. This result shows that Gibbs Markov chains are null (null recurrent or transient) when the posterior is improper and therefore cannot converge (in the usual sense). An insidious feature of this problem is that a null Gibbs chain may be undetectable to the practitioner, that is, the resulting Monte Carlo approximations appear completely reasonable. Such an occurrence has already appeared in a number of published works (Gelfand, Hills, Racine-Poon, and Smith 1990; Geyer 1992; Wang, Rutledge and Gianola 1993, 1994). This is a dangerous situation because the Gibbs sampler will lead to seemingly reasonable inferences about a nonexistent posterior distribution. An example using model (1.2) is given at the end of Section 4. Section 5 contains concluding remarks.

2. HIERARCHICAL LINEAR MIXED MODELS

2.1 The Models

The hierarchical linear mixed models introduced in this section have the standard (noninformative) flat prior on the fixed effects and a parametric "power" improper prior on the variance components. Special cases of this parametric set-up include standard forms, such as discussed in Hill (1965) and Tiao and Tan (1965), as well as some nonstandard forms, such as flat priors. The main results of this section concern the identification of priors which lead to proper posteriors.

Our parametric prior specification leads to a manageable set of Gibbs conditionals, identical in form to those given in Gelfand and Smith (1990) who describe the Gibbs sampler for the one-way random effects model with proper priors. However, it is not only convenience that motivates our use of these priors. The main propriety result (Theorem 1) has applications in the analysis of animal breeding data and in likelihood theory. The results of this section also motivate the general discussion of Gibbs sampling with improper posterior distributions in Section 4.

The model equation is defined as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon} \tag{2.1}$$

where $\mathbf{y}$ is an $n \times 1$ vector of data, $\boldsymbol{\beta}$ is a $p \times 1$ vector of fixed effects (parameters), $\mathbf{u}$ is a $q \times 1$ vector of random effects (random variables), $\mathbf{X}$ and $\mathbf{Z}$ are known design matrices whose dimensions are $n \times p$ and $n \times q$, respectively, and $\boldsymbol{\epsilon}$ is an $n \times 1$ vector of residual errors.

One may object to the use of the term "fixed effect" in a Bayesian model since, from a Bayesian standpoint, all "effects" are random. A frequentist's decision to regard an effect as fixed or random is complicated one (McCulloch 1994). If, however, a frequentist decides that an effect is, indeed, random and provides a (prior) distribution, a Bayesian might be willing to use that distribution as a prior. In our Bayesian mixed models, the terms "fixed effect" and "random effect" are used to distinguish the effects with priors that tend to arise from frequentist considerations from the others.

The typical Bayesian hierarchy for mixed models begins with the assumptions

$$\begin{aligned}&(i)\ \mathbf{u}|\sigma_1^2,\ldots,\sigma_r^2 \sim N_q(\mathbf{0},\mathbf{D})\\&(ii)\ \boldsymbol{\epsilon}|\sigma_\epsilon^2 \sim N_n(\mathbf{0},\mathbf{I}\sigma_\epsilon^2)\end{aligned} \tag{2.2}$$

where $\mathbf{u} = (\mathbf{u}_1' \ \mathbf{u}_2' \ldots \mathbf{u}_r')'$, $\mathbf{u}_i$ is $q_i \times 1$, $\mathbf{D} = \oplus_{i=1}^r \mathbf{I}_{q_i}\sigma_i^2$, and $\sum_{i=1}^r q_i = q$. The r subvectors of $\mathbf{u}$ correspond to the r different random factors in the experiment. We assume that $\mathbf{X}$ is full column rank so that $\mathbf{X}'\mathbf{X}$ is invertible.

Clearly, (ii) implies that $\mathbf{y}|\mathbf{u},\sigma_\epsilon^2,\boldsymbol{\beta} \sim N_n(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u},\mathbf{I}\sigma_\epsilon^2)$ and with the priors mentioned above, our hierarchical model may be written as

$$\mathbf{y}|\mathbf{u},\sigma_\epsilon^2,\boldsymbol{\beta} \sim N_n(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u},\mathbf{I}\sigma_\epsilon^2)$$

$$\pi(\boldsymbol{\beta}) \propto 1 \qquad \mathbf{u}|\sigma_1^2,\ldots,\sigma_r^2 \sim N_q(\mathbf{0},\mathbf{D}) \qquad \pi_\epsilon(\sigma_\epsilon^2|b) \propto \left(\sigma_\epsilon^2\right)^{-(b+1)} \tag{2.3}$$

$$\pi_i(\sigma_i^2|a_i) \propto \left(\sigma_i^2\right)^{-(a_i+1)}$$

where the a_i's and b are known and the following conditional independence assumptions are in force: (1) given $\mathbf{u}$, $\mathbf{y}$ is conditionally independent of $\sigma_1^2,\ldots,\sigma_r^2$, (2) given $\sigma_1^2,\ldots,\sigma_r^2$, $\mathbf{u}$ is conditionally independent of $\boldsymbol{\beta}$ and σ_ϵ^2, and (3) $\boldsymbol{\beta}$, σ_ϵ^2, and $\sigma_1^2,\ldots,\sigma_r^2$ are a priori independent.

This hierarchical model is important for at least two reasons. First, similar models are used to analyze data in many fields including animal breeding (Wang et al. 1993, 1994, Datta 1992) and small area estimation (Ghosh 1994, Datta and Ghosh 1991). The animal breeders assume that the random variables within each subvector $\mathbf{u}_i'$ are not necessarily independent, as they are in our model ($\mathbf{D}$ is diagonal). However, our assumption is merely for convenience, and we show in the next section that a simple reparameterization allows our results concerning the propriety of the posterior for model (2.3) to apply to the animal breeders' model.

The second reason stems from the fact that there is a well-known, important connection between the likelihood function for the linear mixed model and the posterior distribution for a special case of model (2.3). The frequentist version of model (2.3) is, of course

$$\mathbf{y}|\mathbf{u},\sigma_\epsilon^2,\boldsymbol{\beta} \sim N_n(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u},\mathbf{I}\sigma_\epsilon^2) \tag{2.4}$$

$$\mathbf{u}|\sigma_1^2,\ldots,\sigma_r^2 \sim N_q(\mathbf{0},\mathbf{D})$$

where $\boldsymbol{\beta}$, σ_ϵ^2, and $\sigma_1^2,\ldots,\sigma_r^2$ are viewed as fixed, unknown parameters.

The likelihood function is defined by integrating over all possible values of the unobservable random effects (Hill 1965, Searle, Casella and McCulloch 1992, p. 322).

$$\begin{aligned} L\left(\sigma_1^2,\ldots,\sigma_r^2,\sigma_\epsilon^2,\boldsymbol{\beta}|\mathbf{y}\right) &\overset{\text{def}}{=} f(\mathbf{y}|\sigma_1^2,\ldots,\sigma_r^2,\sigma_\epsilon^2,\boldsymbol{\beta}) \\ &= \int_{\Re^q} f(\mathbf{y}|\mathbf{u},\sigma_\epsilon^2,\boldsymbol{\beta})f(\mathbf{u}|\sigma_1^2,\ldots,\sigma_r^2)\,d\mathbf{u}. \end{aligned} \tag{2.5}$$

Consider the Bayesian model obtained by taking $a_1 = \cdots = a_r = b = -1$ in (2.3), i.e., placing flat priors on all of the variance components. Denote the corresponding posterior density by π_l. If the random effects, $\mathbf{u}$, are integrated out of π_l, the resulting marginal posterior is proportional to the likelihood function in (2.5). That is,

$$L\left(\sigma_1^2,\ldots,\sigma_r^2,\sigma_\epsilon^2,\boldsymbol{\beta}|\mathbf{y}\right) \propto \pi_l(\sigma_1^2,\ldots,\sigma_r^2,\sigma_\epsilon^2,\boldsymbol{\beta}|\mathbf{y}) = \int_{\Re^q} \pi_l(\sigma_1^2,\ldots,\sigma_r^2,\sigma_\epsilon^2,\mathbf{u},\boldsymbol{\beta}|\mathbf{y})\,d\mathbf{u}. \tag{2.6}$$

The *restricted* likelihood function is used when (frequentist) inference centers on the variance components and the fixed effects are considered nuisance parameters. The restricted likelihood is the density function of a linear transformation of the data, $\mathbf{K}'\mathbf{y}$, given the variance components, viewed as a function of the variance components. The matrix $\mathbf{K}$ can be any $n \times (n-p)$ matrix of rank $n-p$ such that $\mathbf{K}'\mathbf{X} = 0$. This transformation is justified in a number of ways (Searle et al. 1992, p. 249) and leads to data whose density is not a function of the fixed effects, $\boldsymbol{\beta}$. The values of the variance components which maximize this likelihood are called the restricted maximum likelihood (REML) estimates. It is well known that the restricted likelihood function can be calculated simply by integrating $\boldsymbol{\beta}$ out of the full likelihood in (2.5). The connection between the restricted likelihood function and the posterior distribution corresponding to the flat prior hierarchy is therefore

$$\begin{aligned} L_r\left(\sigma_1^2,\ldots,\sigma_r^2,\sigma_\epsilon^2|\mathbf{y}\right) &= \int_{\Re^p} L\left(\sigma_1^2,\ldots,\sigma_r^2,\sigma_\epsilon^2,\boldsymbol{\beta}|\mathbf{y}\right)d\boldsymbol{\beta} \\ &\propto \int_{\Re^p} \pi_l(\sigma_1^2,\ldots,\sigma_r^2,\sigma_\epsilon^2,\boldsymbol{\beta}|\mathbf{y})d\boldsymbol{\beta} \\ &= \pi_l(\sigma_1^2,\ldots,\sigma_r^2,\sigma_\epsilon^2|\mathbf{y}). \end{aligned} \tag{2.7}$$

The ability to classify the posterior distributions corresponding to the flat prior hierarchy, π_l, as proper or not is useful when it is desired to use the Gibbs sampler to explore likelihood functions via (2.6) and (2.7). An example is given in the next section.

2.2 Propriety Results and Gibbs Sampling

Before we state the theorem indicating which values of $a_1,\ldots,a_r$ and b yield proper posteriors, we consider how model (2.3) lends itself to Gibbs sampling. Assume that

$2a_i > -q_i \; \forall i$ and $2b > -n$. Use the conditional independence assumptions to write the posterior as

$$\pi(\sigma_1^2,\ldots,\sigma_r^2,\sigma_\epsilon^2,\mathbf{u},\boldsymbol{\beta}|\mathbf{y}) \propto f(\mathbf{y}|\mathbf{u},\sigma_\epsilon^2,\boldsymbol{\beta}) f(\mathbf{u}|\sigma_1^2,\ldots,\sigma_r^2)\pi(\boldsymbol{\beta})\pi_\epsilon(\sigma_\epsilon^2|b)\prod_{i=1}^{r}\pi_i(\sigma_i^2|a_i) \tag{2.8}$$

where f is used to represent a generic density.

The functional forms of the Gibbs conditionals can be picked off of the right-hand side of (2.8) as discussed in Section 1. The results are as follows.

$$f\left(\sigma_i^2|\sigma_1^2,\ldots,\sigma_{i-1}^2,\sigma_{i+1}^2,\ldots,\sigma_r^2,\mathbf{y},\mathbf{u},\sigma_\epsilon^2,\boldsymbol{\beta}\right) = IG\left(a_i+\frac{q_i}{2},\,\frac{2}{\mathbf{u}_i'\mathbf{u}_i}\right)$$

$$f\left(\sigma_\epsilon^2|\sigma_1^2,\ldots,\sigma_r^2,\mathbf{y},\mathbf{u},\boldsymbol{\beta}\right) = IG\left(b+\frac{n}{2},\,2\left\{(\mathbf{y}-(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u}))'\,(\mathbf{y}-(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u}))\right\}^{-1}\right)$$

$$f\left(\mathbf{u}|\sigma_1^2,\ldots,\sigma_r^2,\mathbf{y},\sigma_\epsilon^2,\boldsymbol{\beta}\right) = N_q\left(\left(\mathbf{Z}'\mathbf{Z}+\sigma_\epsilon^2\mathbf{D}^{-1}\right)^{-1}\mathbf{Z}'\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right),\,\sigma_\epsilon^2\left(\mathbf{Z}'\mathbf{Z}+\sigma_\epsilon^2\mathbf{D}^{-1}\right)^{-1}\right)$$

$$f\left(\boldsymbol{\beta}|\sigma_1^2,\ldots,\sigma_r^2,\mathbf{y},\sigma_\epsilon^2,\mathbf{u}\right) = N_p\left(\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\left(\mathbf{y}-\mathbf{Z}\mathbf{u}\right),\,\sigma_\epsilon^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right) \tag{2.9}$$

where IG stands for inverted gamma and we say that $X \sim IG(r,s)$ if $f_X(t) \propto t^{-r-1}\exp(-1/st)$ for positive t.

If $2a_i \le -q_i$ for some i or $2b \le -n$, then at least one of the conditionals is improper, since the inverted gamma density is defined only when both parameters are positive (Berger 1985, p. 561). Clearly, one improper conditional implies an improper posterior.

Although it may be tempting to assume that propriety of the conditionals in (2.9) implies propriety of the posterior distribution, the example in Section 1 shows that this is false. Indeed, there are many values of the vector $(a_1, a_2, \ldots a_r, b)$ which simultaneously yield proper conditionals ($2a_i > -q_i \; \forall\, i$ and $2b > -n$) and an improper posterior. Thus, in general, if one incorrectly assumes propriety of a posterior and writes down a (false) proportionality statement like (2.8), it may happen that the Gibbs conditionals are all proper densities. Such a situation is very dangerous because, if the output from the Gibbs sampler fails to warn the user that the posterior is improper, the result could be inferences about a nonexistent posterior distribution. We now state the theorem.

Theorem 1. Let $t = \text{rank}\,(\mathbf{P}_X \mathbf{Z}) = \text{rank}\,(\mathbf{Z}'\mathbf{P}_X\mathbf{Z}) \leq q$ where we define $\mathbf{P}_X = \left(\mathbf{I} - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right)$. There are two cases:

1. If $t = q$ or if $r = 1$ then conditions (i), (ii), and (iii) below are necessary and sufficient for the propriety of the posterior distribution of model (2.3).
2. If $t < q$ and $r > 1$ then conditions (i), (ii), and (iii) below are sufficient for the propriety of the posterior distribution of model (2.3) while necessary conditions result when (ii) is replaced with (ii') $q_i > -2a_i$.

$$(i)\ a_i < 0$$

$$(ii)\ q_i > q - t - 2a_i$$

$$(iii)\ n + 2\textstyle\sum a_i + 2b - p > 0.$$

3. APPLICATIONS

3.1 An Animal Breeding Model

The following hierarchical model is employed in Wang et al. (1993, 1994) for the Bayesian analysis of animal breeding data

$$\mathbf{y}|\mathbf{u}_1, \ldots, \mathbf{u}_r, \sigma_\epsilon^2, \boldsymbol{\beta} \sim N_n(\mathbf{X}\boldsymbol{\beta} + \sum_{i=1}^{r} \mathbf{Z}_i\mathbf{u}_i, \mathbf{I}\sigma_\epsilon^2)$$

$$\pi(\boldsymbol{\beta}) \propto 1 \qquad \mathbf{u}_i|\sigma_i^2 \sim N_{q_i}(\mathbf{0}, \mathbf{G}_i\sigma_i^2) \qquad \sigma_\epsilon^2|\nu_\epsilon, s_\epsilon^2 \sim IG\left(\frac{\nu_\epsilon}{2}, \frac{2}{\nu_\epsilon s_\epsilon^2}\right) \tag{3.1}$$

$$\sigma_i^2|\nu_i, s_i^2 \sim IG\left(\frac{\nu_i}{2}, \frac{2}{\nu_i s_i^2}\right)$$

where the $\mathbf{u}_i$'s are independent and the other model assumptions are the same as those of model (2.3) except, as mentioned above, the random effects in each subvector $\mathbf{u}_i$ have a correlation structure described by the correlation matrix $\mathbf{G}_i$. The matrices $\mathbf{G}_1, \ldots, \mathbf{G}_r$ are known positive definite matrices which "... contain functions of known coefficients of coancestry" (Wang et al. 1993). They suggest that the hyperparameters $s_\epsilon^2, s_1^2, \ldots, s_r^2$ be assigned the prior value of the corresponding variance component and refer to $\nu_\epsilon, \nu_1, \ldots, \nu_r$ as "degree of belief" parameters. The values of these parameters are subjectively chosen to reflect the faith that the experimenter/statistician have in $s_\epsilon^2, s_1^2, \ldots, s_r^2$ as prior estimates of the variance components. This parameterization is considered to be intuitively pleasing in that

$$\mathrm{E}\left(\sigma_i^2\right) = \frac{\nu_i s_i^2}{\nu_i - 2} \qquad \text{and} \qquad \mathrm{Var}\left(\sigma_i^2\right) = \frac{2\nu_i^2 s_i^4}{\left(\nu_i - 2\right)^2 \left(\nu_i - 4\right)} \tag{3.2}$$

when these moments exist. Therefore, if it is believed a priori that $\sigma_i^2 = s_i^2$, but the degree of belief, ν_i, is small, then the prior on σ_i^2 is conservative; having a relatively large variance and a mean larger than s_i^2. The variance and $\mathrm{E}\left(\sigma_i^2\right) - s_i^2$ both go to zero as the degree of belief goes to infinity. Wang et al. (1993) use model (3.1) with the degree of belief parameters all set to zero, which is supposed to reflect prior ignorance about the variance components. Theorem 1 is now used to show that this prior specification leads to an improper posterior distribution.

In model (3.1), if we reparameterize from

$$\left(\mathbf{u}_1, \ldots, \mathbf{u}_r, \sigma_1^2, \ldots, \sigma_r^2, \sigma_\epsilon^2, \boldsymbol{\beta}\right) \quad \text{to} \quad \left(\mathbf{u}_1^*, \ldots, \mathbf{u}_r^*, \sigma_1^2, \ldots, \sigma_r^2, \sigma_\epsilon^2, \boldsymbol{\beta}\right)$$

where $\mathbf{u}_i^* = \mathbf{G}_i^{-\frac{1}{2}} \mathbf{u}_i$, then model (3.1) with degree of belief parameters set to 0, can be written in the form (2.3) with $\mathbf{u} = \left(\mathbf{u}_1^{*\prime} \cdots \mathbf{u}_r^{*\prime}\right)'$, $\mathbf{Z} = \left[\mathbf{Z}_1 \mathbf{G}_1^{\frac{1}{2}}\ \mathbf{Z}_2 \mathbf{G}_2^{\frac{1}{2}} \cdots \mathbf{Z}_r \mathbf{G}_r^{\frac{1}{2}}\right]$, a partitioned matrix, and the a_i's and b all set to 0. Theorem 1 may now be applied and implies that the posterior distribution $\pi\left(\sigma_\epsilon^2, \sigma_1^2, \ldots, \sigma_r^2, \mathbf{u}_1^*, \ldots, \mathbf{u}_r^*, \boldsymbol{\beta} | \mathbf{y}\right)$ is improper since condition (i) is violated. This clearly implies the impropriety of $\pi\left(\sigma_\epsilon^2, \sigma_1^2, \ldots, \sigma_r^2, \mathbf{u}_1, \ldots, \mathbf{u}_r, \boldsymbol{\beta} | \mathbf{y}\right)$ as well, since the difference is only a linear transformation. Wang et al. (1993) actually point out that the posterior is improper, but employ the Gibbs sampler and find no suggestion from the output that the posterior is improper. The futility of this type of application of the Gibbs sampler, and the fictitious answers it can give, are discussed in Section 4.

Wang et al. (1994) suggest that zero degree of belief parameters should not be used because the resulting posterior is improper. They suggest that, instead, ignorance should be modeled by placing flat priors on all of the variance components because such a prior specification leads to a proper posterior, but they provide no proof of this. (In order to get flat priors out of the inverted gamma priors in model (3.1), the degree of belief parameters must be set to -2 and the a priori variance estimates to 0.) In fact, flat priors do not always yield a proper posterior. If flat priors are used in a balanced one-way random effects model with three classes, for example, condition (ii) of Theorem 1 is violated and the posterior is improper.

Theorem 1 can be quite useful when the data follow a hierarchical linear mixed model and simple improper priors are used on the variance components. Its use requires only a simple (if computer intensive) rank calculation and no integration at all.

3.2 Likelihood Estimation

Closed form solutions for maximum likelihood (ML) and restricted maximum likelihood (REML) estimators of the parameters in model (2.4) are often unavailable. In such cases, numerical optimization methods like the Newton-Raphson and EM algorithms can be used to calculate these estimates (Searle et al. 1992, Chapter 8). These algorithms are not guaranteed to locate the global maximum of a likelihood function from an arbitrary starting point, but are much more likely to do so when the starting value is close to the ML (or REML) estimate. Standard starting values are easily computed, unbiased estimates of the variance components, which need not be near the ML (or REML) estimate. When a hierarchical version of the mixed model with flat priors yields a proper posterior distribution, the Gibbs sampler can be used to simulate from a density that is proportional to the likelihood function. The output from such a Gibbs sampler can be used to find better starting values.

Consider a case in which a numerical method is to be used to locate the REML estimate. As in Section 2, let π_l represent the posterior density when flat priors are used in model (2.3). Suppose an application of Theorem 1 shows that π_l is a proper posterior (as will usually be the case in practice). Write the Gibbs chain as $\left(\sigma_\epsilon^{2(j)}, \sigma_1^{2(j)}, \ldots, \sigma_r^{2(j)}, \mathbf{u}'^{(j)}, \boldsymbol{\beta}'^{(j)}\right)$, $j \geq 1$, and let $\boldsymbol{\sigma}^{2(j)} = \left(\sigma_\epsilon^{2(j)}, \sigma_1^{2(j)}, \ldots \sigma_r^{2(j)}\right)'$ The propriety of π_l guarantees that, as $j \to \infty$, these random vectors converge in distribution to a random vector with density π_l. Therefore, equation (2.7) shows that as $j \to \infty$, $\boldsymbol{\sigma}^{2(j)}$ converges in distribution to a random vector whose density is proportional to the restricted likelihood function. Now, insofar as the mode of this density corresponds to an area of high probability, we expect some of the values of $\boldsymbol{\sigma}^{2(j)}$ to be near the REML estimate, once the chain is burned-in. Therefore, reasonable starting values for the numerical optimization methods would be the $\boldsymbol{\sigma}^{2(j)}$'s yielding the largest values of the restricted likelihood function. (Incidentally, the modes of the mixed model likelihoods and restricted likelihoods that we've considered all corresponded to areas of high probability.)

This method is attractive because the Gibbs sampler is so straightforward in this situation. Simple matrix calculations and the ability to simulate normal and inverted gamma random variables are all that is required. Other uses of Markov chain Monte Carlo techniques in likelihood theory are found in Geyer (1991), Geyer and Thompson (1992), and Casella and Berger (1994).

4. GIBBS SAMPLING WITH IMPROPER POSTERIORS

If a complicated hierarchical model with improper priors is postulated, it will often be the case that demonstration of propriety of the posterior will be mathematically tedious, if not impossible. On the other hand, many such models have some type of conjugate structure that makes calculation of the Gibbs conditionals a simple exercise in the recognition of common functional forms. If propriety has not been demonstrated, however, calculating these densities via recognition requires assuming (possibly incorrectly) that a proportionality like (2.8) holds. If a set of proper densities results, it is tempting to assume that the posterior distribution is proper, but this may not be true. The end result is that the Gibbs sampler may be employed in conjunction with a set of conditionals corresponding to an improper posterior distribution. Wang et al. (1993) and Gelfand et al. (1990) both analyze data using the Gibbs sampler in conjunction with one-way random effects models with improper posteriors. (The model labeled I in Section 4 of Gelfand et al. (1990) is slightly different from model (2.3) in that a proper, normal prior is placed on the fixed effects, but the techniques in the proof of Theorem 1 can be used to show that the resulting posterior is improper.) Both articles show plots of approximate marginal posterior densities and give other results which seem completely reasonable. It is not at all obvious in these examples that the posterior distribution is improper and, in fact, that all inferences are to nonexistent posterior distributions.

In this section, we show that Gibbs Markov chains constructed using conditionals from an improper posterior are null, i.e., null recurrent or transient, and therefore do not enjoy the convergence properties associated with Gibbs chains corresponding to proper posteriors. Thus, although Monte Carlo approximations based on the output from a null Gibbs chain may appear reasonable, as in Wang et al. (1993) and Gelfand et al. (1990), their limiting behavior is often quite unreasonable (Hobert and Casella 1995). Section 4.2 gives a simple example using the one-way random model from Section 1.

4.1 Improper Posteriors and Null Markov Chains

The members of a set of conditional density functions are called *compatible* if there exists a joint density function which generates them (Arnold and Press 1989). We call the members *functionally compatible* if there exists a (possibly nonintegrable) function that acts as a joint density with respect to generating the conditions. This section contains results concerning the behavior of Markov chains constructed using the Gibbs algorithm in conjunction with functionally compatible conditional densities. The behavior of Gibbs chains corresponding to improper posteriors can be gleaned from these results since Gibbs conditionals are functionally compatible.

Suppose that N_x and N_y are subsets of $\Re$ and that $f_{X|Y}(x|y)$ and $f_{Y|X}(y|x)$ are real-valued functions with domain $\Re^2$ such that for any $y \in N_y$, $f_{X|Y}(x|y)$ is a probability density function (pdf) in x with support N_x, and similarly for any $x \in N_x$, $f_{Y|X}(y|x)$ is a pdf in y with support N_y. (In this section, X and Y are generic random variables which do not necessarily correspond to parameters or data from a Bayesian model.) An obvious question is: Under what conditions will there exist a joint pdf, $f_{X,Y}(x, y)$, with support $N = N_x \times N_y$, whose conditional pdf's are $f_{X|Y}(x|y)$ and $f_{Y|X}(y|x)$. Following Arnold and Press (1989), $f_{X|Y}(x|y)$ and $f_{Y|X}(y|x)$ will be called compatible conditional densities when such a joint density exists. In general, they will be called candidate conditional densities to reflect the fact that there may not exist a joint density to which they correspond. We begin with something weaker than compatibility.

Definition 1. If there exists a real-valued function $g\,(x,y)$ with domain N such that

$$f_{X|Y}(x|y) = \frac{g(x,y)}{\int_{N_x} g(x,y)dx} \qquad \text{and} \qquad f_{Y|X}(y|x) = \frac{g(x,y)}{\int_{N_y} g(x,y)dy} \tag{4.1}$$

then $f_{X|Y}(x|y)$ and $f_{Y|X}(y|x)$ are *functionally compatible.*

Note that $g\,(x,y)$ need not be integrable. Thus, functional compatibility is necessary, but not sufficient, for compatibility. For example, consider the two exponential conditional densities from Example 2 of Casella and George (1992): $f_{X|Y}(x|y) = y\exp(-xy)$ and $f_{Y|X}(y|x) = x\exp(-yx)$ ($N_x = N_y = \Re_+$). These candidate conditionals are functionally compatible since the conditions of Definition 1 are satisfied using $g\,(x,y) = \exp(-xy)$. This does not imply compatibility, however, since $\exp(-xy)$ is not integrable.

In order to make our point without introducing unwieldy notation, we will stick to the simple example concerning two candidate conditional densities. It is a simple matter, however, to generalize Definition 1 and all of the results of this section to the case of arbitrarily many candidate conditionals (Hobert and Casella 1995). We will therefore take the liberty of referring to the general results.

Clearly, Gibbs conditionals that are calculated via proportionalities like (2.8) and (1.3), are functionally compatible, with the (possibly improper) posterior density serving as g. Thus, everything in this section concerning functionally compatible conditional densities is directly relevant to such Gibbs conditionals.

We now develop a result which allows one to check for functional compatibility and to construct g when it exists. Fix x_0 and y_0 in N_x and N_y, respectively, and define two functions

$$g_1\,(x,y) = \frac{f_{X|Y}(x|y)f_{Y|X}(y|x_0)}{f_{X|Y}(x_0|y)} \qquad \text{and} \qquad g_2\,(x,y) = \frac{f_{Y|X}(y|x)f_{X|Y}(x|y_0)}{f_{Y|X}(y_0|x)}. \tag{4.2}$$

If $f_{X|Y}(x|y)$ and $f_{Y|X}(y|x)$ are compatible, then the joint density is unique and is proportional to both $g_1\,(x,y)$ and $g_2\,(x,y)$ (Besag 1974; Gelman and Speed 1993). Thus, compatibility requires that the ratio of g_1 to g_2 be constant. This condition is actually necessary and sufficient for functional compatibility. (See Brook (1965) for general results concerning factorizations like those in (4.2).)

Theorem 2. The candidate conditionals $f_{X|Y}(x|y)$ and $f_{Y|X}(y|x)$ are functionally compatible if and only if the ratio of $g_1\,(x,y)$ to $g_2\,(x,y)$ is constant. If they are functionally compatible, then both g_1 and g_2 can serve as g, which is unique up to constant multiples.

If the compatibility of a pair of candidate conditional densities is in question, one should first establish whether or not they are functionally compatible. If they are not, then they are not compatible either. If they are, then compatibility follows if and only if g is integrable. This result is stated formally in the following theorem. (Arnold and Press (1989) prove a similar result without the concept of functional compatibility.)

Theorem 3. If $f_{X|Y}(x|y)$ and $f_{Y|X}(y|x)$ are functionally compatible, then they are compatible if and only if $\int\int g(x,y)\,dx\,dy < \infty$. Moreover, if they are compatible then the normalized version of g is the joint density.

Given some starting value for Y, say y^*, consider constructing a Gibbs chain using $f_{X|Y}(x|y)$ and $f_{Y|X}(y|x)$ assuming only that they are functionally compatible. Write the Gibbs chain as $(X_1, Y_1), (X_2, Y_2), \ldots$ and the algorithm as $X_1 \sim f_{X|Y}(\cdot|y^*)$ followed by

$$\begin{gathered} Y_i \sim f_{Y|X}(\cdot|x_i) \\ X_{i+1} \sim f_{X|Y}(\cdot|y_i) \end{gathered}$$

for $i = 1, 2, \ldots$. The results above will be used to establish that this Markov chain is positive recurrent if and only if the conditionals are compatible.

Let $P((x,y),A)$ denote the probability that the chain will be in the set $A \subseteq N$ after the next iteration, given that it is currently at the point (x,y). (Assume A is a two-dimensional Borel set throughout.) $P((x,y),A)$ is called the Markov transition function (Meyn and Tweedie 1992, p. 65) and is given by

$$P((x,y),A) = \int_A f_{X|Y}(s|y) f_{Y|X}(t|s) d(s,t) \tag{4.3}$$

(Robert 1993; Schervish and Carlin 1990). A measure ν on the set N is an invariant measure for this Markov chain if for any $A \subseteq N$,

$$\nu(A) = \int_N \nu(d(x,y)) P((x,y),A). \tag{4.4}$$

The word *invariant* refers to the fact that if ν is a probability measure and the starting value of the chain is generated according to ν, then the distribution of (X_i, Y_i) is ν $\forall$ i. Define a measure ν_G on the set N as follows.

$$\nu_G(A) = \int_A g(x,y)\, d(x,y)\,. \tag{4.5}$$

Theorem 4. The measure ν_G is an invariant measure for the Gibbs chain with Markov transition function $P(\cdot,\cdot)$ in (4.3).

Theorems 3 and 4 can be used to deduce the behavior of the Gibbs Markov chain. Although our main interest is in the chains associated with incompatible conditionals, the well-known compatible case is mentioned for completeness. Suppose that $f_{X|Y}(x|y)$ and $f_{Y|X}(y|x)$ are compatible. Theorem 3 implies that ν_G is (up to a constant multiple) the probability measure associated with the joint density, $f_{X,Y}(x,y)$. Therefore, in the compatible case, Theorem 4 simply states the well known result (Tierney 1995) that the probability measure corresponding to the joint density is the invariant measure for the Gibbs Markov chain. When the Markov chain possesses an invariant probability measure, it is called positive recurrent and under (weak) regularity conditions (Meyn and Tweedie 1992, p. 310), $(X_j, Y_j) \Rightarrow (X, Y) \sim f_{X,Y}$ where the double arrow denotes convergence in distribution. This is basically why the Gibbs sampler "works" when the posterior distribution is proper.

Gibbs conditionals from an improper posterior are a special case of functionally compatible conditional densities which are not compatible. The general versions of Theorems 3 and 4 can be used to show that when a posterior (of arbitrary dimension) is improper, the Gibbs Markov chain possesses an invariant measure with infinite mass and is therefore null, i.e., not positive recurrent (Meyn and Tweedie 1992, p. 231). It follows that if A is any compact set in the parameter space that contains the starting value, the probability of the chain being in the set A after n, say, iterations converges to zero as $n \to \infty$ (see Meyn and Tweedie 1992, p. 454 for a more precise statement). Therefore, when the posterior is improper, the random vectors of a Gibbs Markov chain cannot converge in distribution to any random vector whose probability distribution puts positive mass on compact sets containing the starting value.

These results can be used to show that many Monte Carlo approximations have undesirable limiting behavior when the posterior is improper (see Section 4.2). However, as a reviewer has pointed out, it may be possible to use null Gibbs chains to make inferences about lower dimensional functions of the parameters that have proper posteriors.

4.2 Practical Considerations and an Example

The behavior of a null Gibbs chain can often be anticipated by studying the improper posterior. For example, suppose $(X, Y) \in \Re^2_+$ is the parameter in a Bayesian model which yields the improper posterior $\pi(x, y) \propto y^{d-1} \exp(-xy)$ where $d > 0$. The Gibbs conditionals are easy to compute: $X|Y = y \sim$ exponential$(1/y)$ and $Y|X = x \sim$ gamma$(d, 1/x)$. The posterior is improper because $\int\int \pi(x, y)\, dx\, dy = \int y^{d-2} dy = \Gamma(d) \int x^{-d} dx = \infty$ no matter what the value of d. If $d < 1$, then the "marginal" of Y has an infinite amount of mass near the origin, while the marginal of X has an infinite amount of mass in the limit towards infinity. As we might expect, when the Gibbs chain with $d = 1/2$ is run, the y component gets "absorbed" at zero and the x component "escapes" to infinity. Similarly when $d > 1$, the marginal of X has an infinite amount of mass near the origin, while the marginal of Y has an infinite amount of mass in the limit towards infinity. When the Gibbs chain with $d = 3/2$ is run, the y component escapes to infinity and the x component gets absorbed at zero. When $d = 1$, the Gibbs conditionals are the exponential conditionals from Section 4.1, and both of the marginals have infinite mass near the origin and in the limit towards infinity. Surprisingly, this Gibbs chain is relatively well behaved, periodically returning to the origin in between long charges towards one of the "absorbing states" associated with $d \neq 1$.

Sometimes the behavior of a null Gibbs chain is a function of the starting values.

The following example was suggested by a referee. Consider the hierarchical model

$$\begin{aligned} \mathbf{X} &\sim N_2(\boldsymbol{\mu}, \mathbf{I}) \\ \boldsymbol{\mu} &\sim N_2(\mathbf{1}\theta, \mathbf{I}\tau^2) \\ \pi(\theta, \tau^2) &\propto \tau^{-2}. \end{aligned} \tag{4.6}$$

The resulting posterior is improper due to an infinite amount of mass near $\tau^2 = 0$. Suppose that $x_1 = -x_2 = 10$. This posterior density possesses a well defined peak whose mode is near the point $(\mu_1, \mu_2, \theta, \tau^2) = (9.8, -9.8, 0, 48)$. This peak is well defined because it is separated from the infinite mass near $\tau^2 = 0$ by a region of extremely small mass. Starting the Gibbs chain with τ^2 very close to zero, say 10^{-30}, causes the τ^2 component to be absorbed at zero. On the other hand, if τ^2 is started at a more typical value, say one, then absorption does not occur. The null Gibbs chain apparently "gets stuck" in a "reasonable" part of the parameter space due to the very small probability of a transition to the "bad" part of the space, where absorption would occur. See Geyer (1992, p. 481) for a similar example.

Since this posterior is only an approximation based on an inexact prior and model, one might be willing to simply restrict τ^2 to be larger than some positive ϵ, in order to make the posterior proper. In fact, when the Gibbs sampler is started with a relatively large τ^2 component, the Gibbs output appears to be providing realizations from this "restricted" posterior, but it is not. No matter what the starting values, an arbitrarily small τ^2 component is possible at every iteration. Thus, not only is it impossible for the random vectors of the Gibbs Markov chain to converge in distribution to a random vector from the restricted posterior, but absorption could occur at any time.

The τ^2 step of this Gibbs algorithm entails simulating an $IG(1, 2/((\mu_1 - \theta)^2 + (\mu_2 - \theta)^2))$ random variable. Consider modifying this step by simulating from the inverted gamma density restricted to (ϵ, ∞). This modified Gibbs sampler can be used to simulate from the restricted posterior. Note that unless $(\mu_1 - \theta)^2 + (\mu_2 - \theta)^2$ is very small, the inverted gamma density has negligible mass in the region $(0, \epsilon)$. Thus, if the starting value of τ^2 is relatively large, there will probably be little difference in the outputs from these two Gibbs samplers (over a finite number of iterations) because, at each τ^2 step, there will be little difference between the inverted gamma density and its restriction to (ϵ, ∞). We conclude that, although the theory implies that the unmodified Gibbs chain cannot converge in distribution to the restricted posterior, the output from such a chain may provide a reasonable approximation.

We have just shown that null Gibbs chains can get stuck in reasonable parts of the parameter space for long periods of time. In these cases, output from the Gibbs sampler can produce nice looking pictures of the supposed marginal posterior densities, particularly when the posterior density is computed as an average of conditional densities. However, as the results of this section show, there can be no actual distribution to which the Gibbs picture corresponds. Thus, when we previously referred to the Gibbs-based conclusions of Wang et al. (1993) and Gelfand et al. (1990) as "fictitious," this was based on the observation that, in the problems they analyzed, there can be no conclusions about a posterior distribution because such a distribution does not exist.

In order to demonstrate just how reasonable some of these null Gibbs chains can appear, we give an example. Consider the one-way random effects model from Section 1 with $k = 7$ and $J = 5$. In order to simulate some data we set $\sigma^2 = 5$, $\sigma_\epsilon^2 = 2$, and $\beta = 10$. The vector $(u_1, \ldots, u_7)$ was simulated by generating seven iid $N(0, 5)$ random variables and the vector $(\epsilon_{11}, \ldots, \epsilon_{75})$ was simulated by generating 35 iid $N(0, 2)$ random variables. These numbers were combined according to (1.1). We use the hierarchical model (1.2) with the priors $\pi(\beta) \propto 1$, $\pi\left(\sigma_\epsilon^2\right) \propto 1/\sigma_\epsilon^2$ and $\pi\left(\sigma^2\right) \propto 1/\sigma^2$ which yield an improper posterior (Hill 1965 or take $a = b = 0$ in (2.3) and use Theorem 1). A Gibbs chain was constructed using the conditionals given in (2.9). We denote the chain by $\left(\sigma^{2(j)}, \sigma_\epsilon^{2(j)}, \mathbf{u}'^{(j)}, \beta^{(j)}\right)$, $j \geq 1$. At the start, all parameters were set to one, except for the overall mean, β, which was set to eight. The chain was first allowed to run for 15,000 iterations; keep in mind that the word "burn-in" is not appropriate for these initial iterations because the chain is null and is therefore not converging (in the usual sense). The sole purpose of these initial iterations was to provide the chain with ample opportunity to misbehave and alert us that something may be wrong; it never did. We chose 15,000 because a typical burn-in would probably be in the hundreds (see Gelfand et al. (1990) and Wang et al. (1993)) so that if our chain did not misbehave during the burn-in stage, neither would that of an unknowing experimenter.

After the initial 15,000 iterations, the output from the 15,001st through the 16,000th was collected. Figure 1 is a histogram of the 1,000 effect variances from the null Gibbs chain, that is, $\sigma^{2(j+15,000)}$, $j = 1, 2, \ldots, 1000$, with a Monte Carlo approximation of the supposed marginal posterior density superimposed. Figure 2 is the analogue of Figure 1 for the error variance component. The density approximations in Figures 1 and 2 were calculated using the usual "average of conditional densities" approximation (see formula 2.9 in Casella and George 1992). All of these plots appear perfectly reasonable even though the posterior distribution is improper. (In fact, it can be shown (Hobert and Casella 1995) that the Monte Carlo density approximations have almost sure pointwise limits of zero or no limit at all.) Clearly, if one were unaware of the impropriety, plots like these could lead to seriously misleading conclusions.

This particular posterior is improper due to an infinite amount of mass near $\sigma^2 = 0$. Given the above discussion, one might expect that if the starting value of σ^2 were near zero, the σ^2 component of the Gibbs chain would be absorbed at 0. This is not the case, however. In fact, the σ^2 component and the random effects components move towards zero, but eventually they all return to a reasonable part of the space. For example, we started the chain with $\sigma^2 = 10^{-50}$ and after 20,000 iterations the σ^2 component was approximately 10^{-122} and the largest magnitude of any of the random effects components was about 10^{-60}. The chain was allowed to run for a total of one million iterations, after which all of the components were back in a reasonable part of the parameter space. This Gibbs chain behaves somewhat like the one constructed with the exponential conditionals in that it leaves the "center" of the space for long periods of time, but eventually returns. Such behavior is consistent with null recurrence.

5. DISCUSSION

The fact that it is possible to implement the Gibbs sampler without checking that the posterior is proper is dangerous. What magnifies the problem is that a null Gibbs chain may not provide a "red flag," indicating that something is wrong. This entire problem, however, is based on the initial assumption of propriety necessary to write down the proportionality (like (1.3) in Section 1) which is used to identify the Gibbs conditionals. Clearly, care must be taken to show that such a proportionality is valid, when improper priors are used, before a Markov chain Monte Carlo technique is employed. The examples above and those in the references demonstrate that one should not charge blindly ahead and expect to be informed of an improper posterior by the Markov chain itself.

One way to avoid improper posteriors is to use proper priors. In mixed models, ignorance can be modeled by using a normal prior with very large variance for the fixed effects and inverted gamma priors with very small parameter values for the variance components. The Gibbs conditionals for such a model are easily derived and have the same form as those in (2.9).

The authors have experimented with some diagnostics for null Markov chains (Hobert 1994), but have not met with much success. Typically, the diagnostics work well only in cases where they are not really needed, that is, when the Markov chain is clearly misbehaved. Note that common diagnostics for monitoring "convergence" of the Markov chain are really not appropriate for the cases considered here, since these diagnostics are working under the assumption that the Markov chain is positive recurrent. Thus, they are not diagnosing *if* the chain will converge but rather *when* it will converge. It seems that, for now, the only foolproof way of avoiding the problem is to use proper priors or results like Theorem 1 and those in Ibrahim and Laud (1991) and Dey et al. (1994) which give sufficient conditions for propriety of posteriors for classes of improper priors.

REFERENCES

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

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" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 36, 192–236.

Brook, D. (1964), "On the Distinction Between the Conditional Probability and the Joint Probability Approaches in the Specification of Nearest-Neighbour Systems," *Biometrika,* 51, 481–483.

Casella, G., and Berger, R. L. (1994), "Estimation with Selected Binomial Information or, Do You Really Believe that Dave Winfield is Batting .471?," *Journal of the American Statistical Association,* 89, 1080–1090.

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

Datta, G. S. (1992), "A Unified Bayesian Prediction Theory for Mixed Linear Models with Application," *Statistics and Decisions,* 10, 337–365.

Datta, G. S., and Ghosh, M. (1991), "Bayesian Prediction in Linear Models: Applications to Small Area Estimation," *The Annals of Statistics,* 19, 1748–1770.

Dey, D. K., Gelfand, A. E., and Peng, F (1994), "Overdispersed Generalized Linear Models," Technical Report, University of Connecticut, Dept. of Statistics.

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

Geman, S., and Geman, D. (1984), "Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 6, 721–741.

Geyer, C. J. (1991), "Markov Chain Monte Carlo Maximum Likelihood," in *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface,* ed. E. M. Keramides, Fairfax, VA: Interface Foundation.

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

Geyer, C. J., and Thompson, E. A. (1992), "Constrained Monte Carlo Maximum Likelihood for Dependent Data" (with discussion), *Journal of the Royal Statistical Society,* Ser. B, 54, 657–699.

Ghosh, M., and Rao, J. N. K. (1994), "Small Area Estimation: An Appraisal" (with discussion), *Statistical Science,* 9, 55–93.

Hartigan, J. A. (1983), *Bayes Theory,* New York: Springer-Verlag.

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

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

Hobert, J. P., and Casella, G. (1995), "Functional Compatibility, Markov Chains and Gibbs Sampling with Improper Posteriors," Technical Report No. 481, University of Florida, Department of Statistics.

Ibrahim, J. G., and Laud, P. W. (1991), "On Bayesian Analysis of Generalized Linear Models Using Jeffreys's Prior," *Journal of the American Statistical Association,* 86, 981–986.

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.

Marshall, A. W., and Olkin, I. (1979), *Inequalities: Theory of Majorization and its Applications*, New York: Academic Press.

McCulloch, C. E. (1994), "Fixed and Random Effects and Best Prediction," Technical Report BU-1241-M, Cornell University, Biometrics Unit.

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

Natarajan, R., and McCulloch, C. E. (1995), "A Note on the Existence of the Posterior Distribution for a Class of Mixed Models for Binomial Responses," *Biometrika,* 82, 639–643.

Robert, C. P. (1993), "Convergence Assessments for Markov Chain Monte Carlo Methods," To Appear: *Statistical Science.*

Schervish, M. J., and Carlin, B. P. (1992), "On the Convergence of Successive Substitution Sampling," *Journal of Computational and Graphical Statistics,* 1, 111–127.

Searle, S. R. (1982), *Matrix Algebra Useful for Statistics*, New York: John Wiley and Sons.

Searle, S. R., Casella, G., and McCulloch, C. E. (1992), *Variance Components*, New York: John Wiley and Sons.

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" (with discussion), *The Annals of Statistics,* 22, 1701–1762.

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.

Wang, C. S., Rutledge, J. J., and Gianola, D. (1994), "Bayesian Analysis of Mixed Linear Models via Gibbs Sampling with an Application to Litter Size of Iberian Pigs," *Genetique, Selection, Evolution,* 26, 1–25.

Zeger, S. L., and Karim, M. R. (1991), "Generalized Linear Models With Random Effects; A Gibbs Sampling Approach," *Journal of the American Statistical Association,* 86, 79–86.

APPENDIX: THEOREM PROOFS

A.1 Proof of Theorem 1

Before developing conditions under which the posterior is proper, we define some notation and state two lemmas which are required in the sequel. Let the (real-valued) eigenvalues of a $v \times v$ symmetric matrix $\mathbf{S}$ be written as

$$\lambda_{max}(\mathbf{S}) = \lambda_1(\mathbf{S}) \geq \lambda_2(\mathbf{S}) \geq \cdots \geq \lambda_{v-1}(\mathbf{S}) \geq \lambda_v(\mathbf{S}) = \lambda_{min}(\mathbf{S}). \tag{A.1}$$

Further, let $\lambda_{sp}(\mathbf{S})$ denote the smallest non-zero eigenvalue. Then we have the following results

Lemma 1. If c is a scalar and $\mathbf{S}$ is non-negative definite (n.n.d.) then

$$\lim_{c\to\infty} \mathbf{S}\left[\mathbf{S} + \frac{\mathbf{I}}{c}\right]^{-1} \mathbf{S} = \mathbf{S}, \tag{A.2}$$

that is, in the limit, $(\mathbf{S} + \mathbf{I}/c)^{-1}$ is a generalized inverse of $\mathbf{S}$.

Proof. Since $\mathbf{S}$ is symmetric, it can be factored as $\mathbf{S} = \mathbf{H}'\mathbf{\Lambda}\mathbf{H}$ where $\mathbf{H}$ is orthogonal and $\mathbf{\Lambda}$ is a diagonal matrix of the eigenvalues of $\mathbf{S}$. Now

$$\lim_{c\to\infty} \mathbf{S}\left[\mathbf{S} + \frac{\mathbf{I}}{c}\right]^{-1} \mathbf{S} = \lim_{c\to\infty} \mathbf{H}'\mathbf{\Lambda}\mathbf{H}\mathbf{H}'\left[\mathbf{\Lambda} + \frac{\mathbf{I}}{c}\right]^{-1}\mathbf{H}\mathbf{H}'\mathbf{\Lambda}\mathbf{H} = \lim_{c\to\infty} \mathbf{H}'\mathbf{\Lambda}\left[\mathbf{\Lambda} + \frac{\mathbf{I}}{c}\right]^{-1}\mathbf{\Lambda}\mathbf{H}.$$

Assume (w.l.o.g) that $\mathbf{\Lambda} = \text{diag}\left(\lambda_1(\mathbf{S}), \lambda_2(\mathbf{S}), \ldots, \lambda_t(\mathbf{S}), 0, \ldots, 0\right)$ where $t \leq v$ is the rank of $\mathbf{S}$. Then

$$\mathbf{\Lambda}\left[\mathbf{\Lambda} + \frac{\mathbf{I}}{c}\right]^{-1}\mathbf{\Lambda} = \text{diag}\left(\frac{c\lambda_1^2}{c\lambda_1+1}, \frac{c\lambda_2^2}{c\lambda_2+1}, \ldots, \frac{c\lambda_t^2}{c\lambda_t+1}, 0, \ldots, 0\right).$$

and the result follows since

$$\lim_{c\to\infty} \frac{c\lambda_i^2}{c\lambda_i+1} = \lambda_i.$$

□

Lemma 2. (Marshall and Olkin 1979). If two symmetric matrices $\mathbf{S}_1$ and $\mathbf{S}_2$ are both n.n.d. then

$$\prod_{i=1}^{v}\left[\lambda_i(\mathbf{S}_1) + \lambda_i(\mathbf{S}_2)\right] \leq |\mathbf{S}_1 + \mathbf{S}_2| \leq \prod_{i=1}^{v}\left[\lambda_i(\mathbf{S}_1) + \lambda_{v-i+1}(\mathbf{S}_2)\right].$$

We first prove case (1) of Theorem 1. From equation (2.8), we have

$$\pi(\sigma_1^2,\ldots,\sigma_r^2,\sigma_\epsilon^2,\mathbf{u},\boldsymbol{\beta}|\mathbf{y}) = \frac{f(\mathbf{y}|\mathbf{u},\sigma_\epsilon^2,\boldsymbol{\beta})f(\mathbf{u}|\sigma_1^2,\ldots,\sigma_r^2)\pi(\boldsymbol{\beta})\pi_\epsilon(\sigma_\epsilon^2|b)\prod_{i=1}^r \pi_i(\sigma_i^2|a_i)}{m(\mathbf{y})} \tag{A.3}$$

where $m(\mathbf{y})$, the marginal density of the data, is given by

$$m(\mathbf{y}) = \int f(\mathbf{y}|\mathbf{u},\sigma_\epsilon^2,\boldsymbol{\beta})f(\mathbf{u}|\sigma_1^2,\ldots,\sigma_r^2)\pi(\boldsymbol{\beta})\pi_\epsilon(\sigma_\epsilon^2|b)\prod_{i=1}^r \pi_i(\sigma_i^2|a_i)\,d\mathbf{u}\,d\boldsymbol{\beta}\,d\sigma_\epsilon^2\prod_{i=1}^r d\sigma_i^2. \tag{A.4}$$

It is straightforward to show that

$$\int f(\mathbf{y}|\mathbf{u},\sigma_\epsilon^2,\boldsymbol{\beta})f(\mathbf{u}|\sigma_1^2,\ldots,\sigma_r^2)\pi(\boldsymbol{\beta})\,d\mathbf{u}\,d\boldsymbol{\beta} = \frac{\exp\left\{\frac{1}{2}\mathbf{y}'\left(\mathbf{M}_1\mathbf{X}\mathbf{M}_2\mathbf{X}'\mathbf{M}_1 - \mathbf{M}_1\right)\mathbf{y}\right\}}{(2\pi)^{\frac{N-p}{2}}\left(\sigma_\epsilon^2\right)^{\frac{N-q-p}{2}}|\mathbf{D}|^{\frac{1}{2}}\,|\mathbf{X}'\mathbf{X}|^{\frac{1}{2}}\left|\sigma_\epsilon^2\mathbf{D}^{-1}+\mathbf{Z}'\mathbf{P}_X\mathbf{Z}\right|^{\frac{1}{2}}} \tag{A.5}$$

where $\mathbf{P}_X$ is defined in the statement of Theorem 1 and

$$\mathbf{M}_1 = \left(\mathbf{Z}\mathbf{D}\mathbf{Z}' + \mathbf{I}\sigma_\epsilon^2\right)^{-1} \quad \text{and} \quad \mathbf{M}_2 = \left(\mathbf{X}'\left(\mathbf{Z}\mathbf{D}\mathbf{Z}' + \mathbf{I}\sigma_\epsilon^2\right)^{-1}\mathbf{X}\right)^{-1}. \tag{A.6}$$

Note that (A.5) is the restricted likelihood function in (2.7). We may now write

$$\int f(\mathbf{y}|\mathbf{u},\sigma_\epsilon^2,\boldsymbol{\beta})f(\mathbf{u}|\sigma_1^2,\ldots,\sigma_r^2)\pi(\boldsymbol{\beta})\pi_\epsilon(\sigma_\epsilon^2|b)\prod_{i=1}^r \pi_i(\sigma_i^2|a_i)\,d\mathbf{u}\,d\boldsymbol{\beta}\prod_{i=1}^r d\sigma_i^2 = \frac{\pi_\epsilon(\sigma_\epsilon^2|b)}{(2\pi)^{\frac{N-p}{2}}\left(\sigma_\epsilon^2\right)^{\frac{N-q-p}{2}}|\mathbf{X}'\mathbf{X}|^{\frac{1}{2}}}\cdot \int \frac{\exp\left\{\frac{1}{2}\mathbf{y}'\left(\mathbf{M}_1\mathbf{X}\mathbf{M}_2\mathbf{X}'\mathbf{M}_1 - \mathbf{M}_1\right)\mathbf{y}\right\}\prod_{i=1}^r \pi_i(\sigma_i^2|a_i)}{|\mathbf{D}|^{\frac{1}{2}}\left|\sigma_\epsilon^2\mathbf{D}^{-1}+\mathbf{Z}'\mathbf{P}_X\mathbf{Z}\right|^{\frac{1}{2}}}\prod_{i=1}^r d\sigma_i^2. \tag{A.7}$$

Consider the exponential in (A.8). Write it as

$$f(\mathbf{t}) = \exp\left\{\frac{1}{2}\mathbf{y}'\left(\mathbf{M}_1\mathbf{X}\mathbf{M}_2\mathbf{X}'\mathbf{M}_1 - \mathbf{M}_1\right)\mathbf{y}\right\} \tag{A.8}$$

where $\mathbf{t} = (\sigma_1^2,\ldots,\sigma_r^2)$. A lengthy differentiation argument will show that $f(\mathbf{t})$ is non-decreasing in each of its arguments. Upper and lower bounds are now developed for $f(\mathbf{t})$. The lower bound is simple

$$f(\mathbf{0}) = \exp\left\{-\frac{1}{2\sigma_\epsilon^2}\mathbf{y}'\mathbf{P}_X\mathbf{y}\right\}. \tag{A.9}$$

The next step is to find an upper bound. We let the variance components go to ∞ at the same rate, so we will calculate $\lim_{c\to\infty} f(\mathbf{1}c)$, where $\mathbf{1}$ is a one-vector. Applying the Schur complement formula (Searle 1982 p.261) we have

$$\left(\mathbf{Z}\mathbf{D}\mathbf{Z}' + \mathbf{I}\sigma_\epsilon^2\right)^{-1} = \frac{\mathbf{I}}{\sigma_\epsilon^2} - \frac{1}{\sigma_\epsilon^2}\mathbf{Z}\left(\sigma_\epsilon^2\mathbf{D}^{-1} + \mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'. \tag{A.10}$$

Lemma 1 and (A.10) together give

$$\lim_{c\to\infty}\left(\mathbf{Z}\mathbf{I}c\mathbf{Z}' + \mathbf{I}\sigma_\epsilon^2\right)^{-1} = \frac{1}{\sigma_\epsilon^2}\mathbf{P}_Z \tag{A.11}$$

where $\mathbf{P}_Z = \mathbf{I} - \mathbf{Z}\left(\mathbf{Z}'\mathbf{Z}\right)^{-}\mathbf{Z}'$ with $-$ denoting a generalized inverse. A slightly more complicated calculation involving the Schur complement and another application of Lemma 1 shows that

$$\lim_{c\to\infty}\mathbf{X}'\mathbf{P}_X\mathbf{X}\left[\mathbf{X}'\left(\mathbf{I} - \mathbf{Z}\left(\mathbf{Z}'\mathbf{Z} + \frac{\sigma_\epsilon^2\mathbf{I}}{c}\right)^{-1}\mathbf{Z}'\right)\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{P}_X\mathbf{X} = \mathbf{X}'\mathbf{P}_X\mathbf{X}. \tag{A.12}$$

Equations (A.11) and (A.12) yield

$$\lim_{c\to\infty} f(\mathbf{1}c) = \exp\left\{-\frac{1}{2\sigma_\epsilon^2}\mathbf{y}'\left(\mathbf{P}_Z - \mathbf{P}_Z\mathbf{X}\left(\mathbf{X}'\mathbf{P}_Z\mathbf{X}\right)^{-}\mathbf{X}'\mathbf{P}_Z\right)\mathbf{y}\right\} \tag{A.13}$$

where writing $\left(\mathbf{X}'\mathbf{P}_Z\mathbf{X}\right)^{-}$ as $\left(\mathbf{T}'\mathbf{T}\right)^{-}$ for $\mathbf{T} = \mathbf{P}_Z\mathbf{X}$ shows that $\mathbf{P}_Z\mathbf{X}\left(\mathbf{X}'\mathbf{P}_Z\mathbf{X}\right)^{-}\mathbf{X}'\mathbf{P}_Z$ is invariant to the generalized inverse. Also, $\mathbf{P}_Z\mathbf{X}\left(\mathbf{X}'\mathbf{P}_Z\mathbf{X}\right)^{-}\mathbf{X}'\mathbf{P}_Z\mathbf{X} = \mathbf{P}_Z\mathbf{X}$ which means that $\mathbf{P}_Z - \mathbf{P}_Z\mathbf{X}\left(\mathbf{X}'\mathbf{P}_Z\mathbf{X}\right)^{-}\mathbf{X}'\mathbf{P}_Z$ is idempotent. Combining the results above gives

$$\exp\left\{-\frac{1}{2\sigma_\epsilon^2}\mathbf{y}'\mathbf{P}_X\mathbf{y}\right\} \le f(\mathbf{t}) \le \exp\left\{-\frac{1}{2\sigma_\epsilon^2}\mathbf{y}'\left(\mathbf{P}_Z - \mathbf{P}_Z\mathbf{X}\left(\mathbf{X}'\mathbf{P}_Z\mathbf{X}\right)^{-}\mathbf{X}'\mathbf{P}_Z\right)\mathbf{y}\right\}. \tag{A.14}$$

Conditions (i), (ii), and (iii) are first shown to be sufficient for integrability. Using the upper bound

$$\int f(\mathbf{y}|\mathbf{u},\sigma_\epsilon^2,\boldsymbol{\beta})f(\mathbf{u}|\sigma_1^2,\ldots,\sigma_r^2)\pi(\boldsymbol{\beta})\pi_\epsilon(\sigma_\epsilon^2|b)\prod_{i=1}^{r}\pi_i(\sigma_i^2|a_i)\,d\mathbf{u}\,d\boldsymbol{\beta}\prod_{i=1}^{r}d\sigma_i^2 \le$$

$$\frac{\pi_\epsilon(\sigma_\epsilon^2|b)\exp\left\{-\frac{1}{2\sigma_\epsilon^2}\mathbf{y}'\left(\mathbf{P}_Z - \mathbf{P}_Z\mathbf{X}\left(\mathbf{X}'\mathbf{P}_Z\mathbf{X}\right)^{-}\mathbf{X}'\mathbf{P}_Z\right)\mathbf{y}\right\}}{(2\pi)^{\frac{N-p}{2}}\left(\sigma_\epsilon^2\right)^{\frac{N-q-p}{2}}|\mathbf{X}'\mathbf{X}|^{\frac{1}{2}}}.$$

$$\int \frac{\prod_{i=1}^{r} \pi_i(\sigma_i^2|a_i)}{|\mathbf{D}|^{\frac{1}{2}} \left|\sigma_\epsilon^2 \mathbf{D}^{-1} + \mathbf{Z}'\mathbf{P}_X\mathbf{Z}\right|^{\frac{1}{2}}} \prod_{i=1}^{r} d\sigma_i^2. \qquad (A.15)$$

Focusing on the integrand, Lemma 2 gives

$$\prod_{i=1}^{q} \left[\lambda_i\left(\sigma_\epsilon^2 \mathbf{D}^{-1}\right) + \lambda_i\left(\mathbf{Z}'\mathbf{P}_X\mathbf{Z}\right)\right] \leq \left|\sigma_\epsilon^2 \mathbf{D}^{-1} + \mathbf{Z}'\mathbf{P}_X\mathbf{Z}\right|$$

$$\leq \prod_{i=1}^{q} \left[\lambda_i\left(\sigma_\epsilon^2 \mathbf{D}^{-1}\right) + \lambda_{q-i+1}\left(\mathbf{Z}'\mathbf{P}_X\mathbf{Z}\right)\right] \qquad (A.16)$$

Assume that $t = q$. (A slightly different argument is necessary when $t < q$ and $r = 1$ which is explained below.) When $t = q$, equation (A.16) yields

$$\prod_{i=1}^{r} \left[\frac{\sigma_\epsilon^2}{\sigma_i^2} + \lambda_{min}\left(\mathbf{Z}'\mathbf{P}_X\mathbf{Z}\right)\right]^{q_i} \leq \left|\sigma_\epsilon^2 \mathbf{D}^{-1} + \mathbf{Z}'\mathbf{P}_X\mathbf{Z}\right| \leq \prod_{i=1}^{r} \left[\frac{\sigma_\epsilon^2}{\sigma_i^2} + \lambda_{max}\left(\mathbf{Z}'\mathbf{P}_X\mathbf{Z}\right)\right]^{q_i} \qquad (A.17)$$

and employing the lower bound in (A.17) yields an upper bound for the integral in (A.15)

$$\int \frac{\prod_{i=1}^{r} \pi_i(\sigma_i^2|a_i)}{|\mathbf{D}|^{\frac{1}{2}} \left|\sigma_\epsilon^2 \mathbf{D}^{-1} + \mathbf{Z}'\mathbf{P}_X\mathbf{Z}\right|^{\frac{1}{2}}} \prod_{i=1}^{r} d\sigma_i^2 \leq \prod_{i=1}^{r} \int \frac{\left(\sigma_i^2\right)^{-(a_i+1)}}{\left(\lambda_{min}\left(\mathbf{Z}'\mathbf{P}_X\mathbf{Z}\right)\sigma_i^2 + \sigma_\epsilon^2\right)^{\frac{q_i}{2}}} d\sigma_i^2. \qquad (A.18)$$

The generic form of the integrals on the right-hand side in (A.18) can be written as

$$l^{\frac{q_i}{2}} \int \frac{t^{-(a_i+1)}}{\left(l\sigma_\epsilon^2 + t\right)^{\frac{q_i}{2}}} dt \qquad (A.19)$$

where $l = \left(\lambda_{min}\left(\mathbf{Z}'\mathbf{P}_X\mathbf{Z}\right)\right)^{-1}$. This integral will be finite if and only if $a_i > 0$ and $q_i > -2a_i$ and when these conditions hold, the integral in (A.19) equals $c_i\left(\sigma_\epsilon^2\right)^{-(a_i+q_i/2)}$, where c_i is constant in σ_ϵ^2. Therefore, since we've assumed that these two conditions hold for $i = 1, 2, \ldots, r$, we may write

$$\int f(\mathbf{y}|\mathbf{u}, \sigma_\epsilon^2, \boldsymbol{\beta}) f(\mathbf{u}|\sigma_1^2, \ldots, \sigma_r^2) \pi(\boldsymbol{\beta}) \pi_\epsilon(\sigma_\epsilon^2|b) \prod_{i=1}^{r} \pi_i(\sigma_i^2|a_i)\, d\mathbf{u}\, d\boldsymbol{\beta} \prod_{i=1}^{r} d\sigma_i^2 \leq$$

$$\frac{\pi_\epsilon(\sigma_\epsilon^2|b) \exp\left\{-\frac{1}{2\sigma_\epsilon^2}\mathbf{y}'\left(\mathbf{P}_Z - \mathbf{P}_Z\mathbf{X}\left(\mathbf{X}'\mathbf{P}_Z\mathbf{X}\right)^{-}\mathbf{X}'\mathbf{P}_Z\right)\mathbf{y}\right\}}{(2\pi)^{\frac{N-p}{2}} \left(\sigma_\epsilon^2\right)^{\frac{N-q-p}{2}} |\mathbf{X}'\mathbf{X}|^{\frac{1}{2}}} \cdot c \cdot \left(\sigma_\epsilon^2\right)^{-\left(\frac{q}{2} + \sum a_i\right)}. \qquad (A.20)$$

Sufficiency will follow if it can be shown that (A.20) is integrable with respect to σ_ϵ^2. This follows from condition (*iii*) and the fact that, aside from a constant, (A.20) is an inverted gamma density in σ_ϵ^2.

Necessity is simple given all of the bounds above. Using the lower bound on the exponential in (A.14) and the upper bound in (A.17) yields

$$\int f(\mathbf{y}|\mathbf{u},\sigma_\epsilon^2,\boldsymbol{\beta})f(\mathbf{u}|\sigma_1^2,\ldots,\sigma_r^2)\pi(\boldsymbol{\beta})\pi_\epsilon(\sigma_\epsilon^2|b)\prod_{i=1}^r \pi_i(\sigma_i^2|a_i)\,d\mathbf{u}\,d\boldsymbol{\beta}\prod_{i=1}^r d\sigma_i^2 \geq$$

$$\frac{\pi_\epsilon(\sigma_\epsilon^2|b)\exp\left\{-\frac{1}{2\sigma_\epsilon^2}\mathbf{y}'\mathbf{P}_X\mathbf{y}\right\}}{(2\pi)^{\frac{N-p}{2}}(\sigma_\epsilon^2)^{\frac{N-q-p}{2}}|\mathbf{X}'\mathbf{X}|^{\frac{1}{2}}}\cdot\prod_{i=1}^r\int\frac{(\sigma_i^2)^{-(a_i+1)}}{\left(\lambda_{max}\left(\mathbf{Z}'\mathbf{P}_X\mathbf{Z}\right)\sigma_i^2+\sigma_\epsilon^2\right)^{\frac{q_i}{2}}}d\sigma_i^2. \tag{A.21}$$

This inequality demonstrates the necessity of conditions (*i*) and (*ii*) since the right-hand integral will diverge if either (or both) fails to hold. If both hold, an argument similar to one above shows

$$\int f(\mathbf{y}|\mathbf{u},\sigma_\epsilon^2,\boldsymbol{\beta})f(\mathbf{u}|\sigma_1^2,\ldots,\sigma_r^2)\pi(\boldsymbol{\beta})\pi_\epsilon(\sigma_\epsilon^2|b)\prod_{i=1}^r \pi_i(\sigma_i^2|a_i)\,d\mathbf{u}\,d\boldsymbol{\beta}\prod_{i=1}^r d\sigma_i^2 \geq$$

$$\frac{\pi_\epsilon(\sigma_\epsilon^2|b)\exp\left\{-\frac{1}{2\sigma_\epsilon^2}\mathbf{y}'\mathbf{P}_X\mathbf{y}\right\}}{(2\pi)^{\frac{N-p}{2}}(\sigma_\epsilon^2)^{\frac{N-q-p}{2}}|\mathbf{X}'\mathbf{X}|^{\frac{1}{2}}}\cdot c'\cdot\left(\sigma_\epsilon^2\right)^{-\left(\frac{q}{2}+\sum a_i\right)} \tag{A.22}$$

where c' is constant in σ_ϵ^2. As a function of σ_ϵ^2, the right-hand side of (A.22) is, again, an inverted gamma density and this makes clear the necessity of condition (*iii*).
For the case when $t < q$ and $r = 1$,

$$\begin{aligned}\left|\sigma_\epsilon^2\mathbf{D}^{-1}+\mathbf{Z}'\mathbf{P}_X\mathbf{Z}\right|^{\frac{1}{2}} &= \left|\mathbf{I}\frac{\sigma_\epsilon^2}{\sigma_1^2}+\mathbf{Z}'\mathbf{P}_X\mathbf{Z}\right|^{\frac{1}{2}}\\ &= \left|\mathbf{I}\frac{\sigma_\epsilon^2}{\sigma_1^2}+\mathbf{H}'\boldsymbol{\Lambda}\mathbf{H}\right|^{\frac{1}{2}}\end{aligned}$$

where $\mathbf{H}$ is orthogonal and $\boldsymbol{\Lambda}$ is a diagonal matrix of the eigenvalues of $\mathbf{Z}'\mathbf{P}_X\mathbf{Z}$. Since $\mathbf{Z}'\mathbf{P}_X\mathbf{Z}$ is a n.n.d. matrix with rank $t < q$, it has t positive eigenvalues and $q-t$ zero eigenvalues. Therefore

$$\left[\frac{\sigma_\epsilon^2}{\sigma_1^2}+\lambda_{sp}\right]^{\frac{t}{2}}\left(\frac{\sigma_\epsilon^2}{\sigma_1^2}\right)^{\frac{q-t}{2}} \leq \left|\sigma_\epsilon^2\mathbf{D}^{-1}+\mathbf{Z}'\mathbf{P}_X\mathbf{Z}\right|^{\frac{1}{2}} \leq \left[\frac{\sigma_\epsilon^2}{\sigma_1^2}+\lambda_{max}\right]^{\frac{t}{2}}\left(\frac{\sigma_\epsilon^2}{\sigma_1^2}\right)^{\frac{q-t}{2}} \tag{A.23}$$

where $\lambda_{sp} = \lambda_{sp}\left(\mathbf{Z}'\mathbf{P}_X\mathbf{Z}\right)$ and $\lambda_{max} = \lambda_{max}\left(\mathbf{Z}'\mathbf{P}_X\mathbf{Z}\right)$. The proof above can then be used with (A.23) in place of (A.17).
That is the end of the proof of Case 1 of Theorem 1. The only thing that changes going from case 1 to case 2 is that in case 2, the matrix $\mathbf{Z}'\mathbf{P}_X\mathbf{Z}$ has some zero eigenvalues which result in a slightly different lower bound for $\left|\sigma_\epsilon^2\mathbf{D}^{-1}+\mathbf{Z}'\mathbf{P}_X\mathbf{Z}\right|$. In particular, Lemma 1 can be used to show that

$$\prod_{i=1}^{t}\left[\lambda_i\left(\sigma_\epsilon^2\mathbf{D}^{-1}\right)+\lambda_{sp}\left(\mathbf{Z}'\mathbf{P}_X\mathbf{Z}\right)\right]\prod_{i=t+1}^{q}\left[\lambda_i\left(\sigma_\epsilon^2\mathbf{D}^{-1}\right)\right]\leq\left|\sigma_\epsilon^2\mathbf{D}^{-1}+\mathbf{Z}'\mathbf{P}_X\mathbf{Z}\right|$$

$$\leq\prod_{i=1}^{r}\left[\frac{\sigma_\epsilon^2}{\sigma_i^2}+\lambda_{max}\left(\mathbf{Z}'\mathbf{P}_X\mathbf{Z}\right)\right]^{q_i}. \tag{A.24}$$

The fact that the upper bound in (A.24) is the same as it was in case 1 (see (A.17)) means that the proof of necessity for case 1 can be used again for case 2. Basically, the proof of sufficiency for case 1 also works again except that some extra work is needed to integrate over $\Re_+^r$ in (A.15). Using the lower bound in (A.24) directly in the integral over $\Re_+^r$ is impossible because it requires knowledge of the smallest eigenvalue of $\mathbf{D}^{-1}$ which changes depending on your location in $\Re_+^r$. We avoid this problem by introducing the following mutually exclusive sets

$$\left\{\left(\sigma_{i_1}^2,\sigma_{i_2}^2,\ldots,\sigma_{i_r}^2\right)\in\Re_+^r \text{ s.t. } \sigma_{i_1}^2<\sigma_{i_2}^2<\cdots<\sigma_{i_r}^2\right\} \tag{A.25}$$

where $(i_1, i_2, \ldots, i_r)$ is one of the $r!$ permutations of $(1, 2, \ldots, r)$. On each of these sets the eigenvalues of $\mathbf{D}^{-1}$ (the inverses of the variance components) have a constant ordering. We can therefore use the lower bound to integrate over each of these $r!$ sets and then add the results to get the full integral over $\Re_+^r$. For instance, let S denote the set in (A.25) with $(i_1, i_2, \ldots, \imath_r) = (1, 2, \ldots, r)$. On the set S, the lower bound in (A.24) becomes

$$\prod_{i=1}^{t}\left[\lambda_i\left(\sigma_\epsilon^2\mathbf{D}^{-1}\right)+\lambda_{sp}\left(\mathbf{Z}'\mathbf{P}_X\mathbf{Z}\right)\right]\prod_{i=t+1}^{q}\left[\lambda_i\left(\sigma_\epsilon^2\mathbf{D}^{-1}\right)\right]=$$

$$\left(\frac{\sigma_\epsilon^2}{\sigma_1^2}+\lambda_{sp}\right)^{q_1}\left(\frac{\sigma_\epsilon^2}{\sigma_2^2}+\lambda_{sp}\right)^{q_2}\cdots\left(\frac{\sigma_\epsilon^2}{\sigma_r^2}+\lambda_{sp}\right)^{t-\sum_{i=1}^{r-1}q_i}\left(\frac{\sigma_\epsilon^2}{\sigma_r^2}\right)^{q-t}.$$

where again $\lambda_{sp}=\lambda_{sp}\left(\mathbf{Z}'\mathbf{P}_X\mathbf{Z}\right)$ and from this follows

$$\int_S\frac{\prod_{i=1}^r\pi_i(\sigma_i^2|a_i)}{|\mathbf{D}|^{\frac{1}{2}}\left|\sigma_\epsilon^2\mathbf{D}^{-1}+\mathbf{Z}'\mathbf{P}_X\mathbf{Z}\right|^{\frac{1}{2}}}\prod_{i=1}^{r}d\sigma_i^2$$

$$\leq\int_S\frac{(\sigma_\epsilon^2)^{\frac{t-q}{2}}\prod_{i=1}^r\pi_i(\sigma_i^2|a_i)}{\left(\sigma_\epsilon^2+\sigma_r^2\lambda_{sp}\left(\mathbf{Z}'\mathbf{P}_X\mathbf{Z}\right)\right)^{\frac{t}{2}-\sum_{i=1}^{r-1}\frac{q_i}{2}}\prod_{i=1}^{r-1}\left(\sigma_\epsilon^2+\sigma_i^2\lambda_{sp}\left(\mathbf{Z}'\mathbf{P}_X\mathbf{Z}\right)\right)^{\frac{q_i}{2}}}\prod_{i=1}^{r}d\sigma_i^2$$

$$\leq\int_{\Re_+^r}\frac{(\sigma_\epsilon^2)^{\frac{t-q}{2}}\prod_{i=1}^r\pi_i(\sigma_i^2|a_i)}{\left(\sigma_\epsilon^2+\sigma_r^2\lambda_{sp}\left(\mathbf{Z}'\mathbf{P}_X\mathbf{Z}\right)\right)^{\frac{t}{2}-\sum_{i=1}^{r-1}\frac{q_i}{2}}\prod_{i=1}^{r-1}\left(\sigma_\epsilon^2+\sigma_i^2\lambda_{sp}\left(\mathbf{Z}'\mathbf{P}_X\mathbf{Z}\right)\right)^{\frac{q_i}{2}}}\prod_{i=1}^{r}d\sigma_i^2$$

since the integrand is positive. The rest of the proof of sufficiency for case 2 follows closely that of case 1. □

A.2 Proof of Theorem 2

Assume that $f_{X|Y}$ and $f_{Y|X}$ are functionally compatible. From the definition we have that $g_i\,(x,y) \propto g\,(x,y)$, $i = 1, 2$, which implies that the ratio of g_1 to g_2 is constant. Now suppose that the ratio is constant. Clearly, $g_1\,(x,y) \,/ \int g_1\,(x,y)\,dx = f_{X|Y}$. Furthermore, the constant ratio implies that $g_1\,(x,y) \,/ \int g_1\,(x,y)\,dy = f_{Y|X}$. Therefore, Definition 1 is satisfied with g_1 serving the role of g in (4.1). □

A.3 Proof of Theorem 3

If $f_{X|Y}$ and $f_{Y|X}$ are compatible, then g must be a proportional to the joint density. Conversely, if the integral is finite, then g is normalizable and compatibility follows. □

A.4 Proof of Theorem 4

$$\begin{aligned}
&\int_N \nu\,(d(x,y))\,P((x,y),A) \\
&\qquad = \int_{N_x}\int_{N_y}\left[\int_A f_{X|Y}(s|y)f_{Y|X}(t|s)d(s,t)\right] g(x,y)dx\,dy \\
&\qquad = \int_A\left[\int_{N_y}\int_{N_x} g(x,y)f_{X|Y}(s|y)f_{Y|X}(t|s)dx\,dy\right] d(s,t) \\
&\qquad = \int_A\left[\int_{N_y} g(s,y)f_{Y|X}(t|s)dy\right] d(s,t) \\
&\qquad = \int_A g(s,t)d(s,t) \\
&\qquad = \nu\,(A)
\end{aligned}$$

where the third and fourth equalities follow from functional compatibility. □

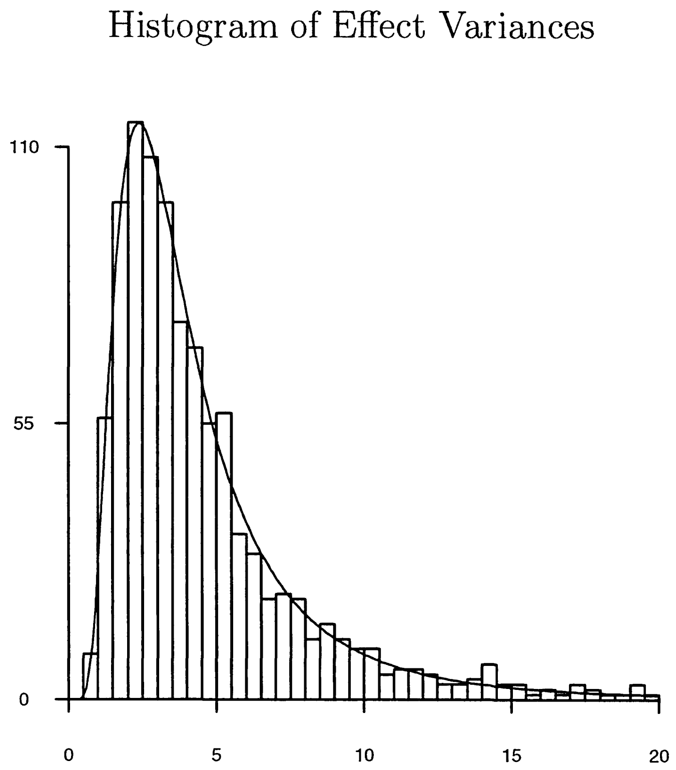

Figure 1. Histogram of the 1000 values of the effect variance from the null Gibbs chain, that is, a histogram of $\sigma^{2(j+15,000)}$ for $j = 1, 2, \ldots, 1000$. Superimposed is the approximate (supposed) marginal posterior density of σ^2. An appropriately scaled version of $\hat{\pi}_{\sigma^2|\mathbf{y}}(t|\mathbf{y})$ is on the ordinate with t on the abscissa. (Actually, 15 of the 1,000 values of the effect variance, ranging from 21.0 to 45.1, were not included in the histogram.)

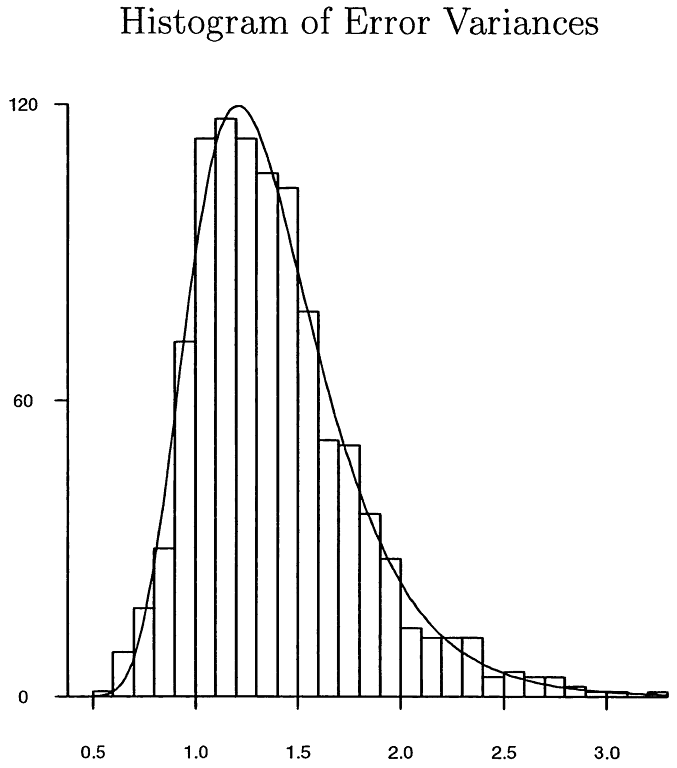

Figure 2. Histogram of the 1000 values of the error variance from the null Gibbs chain, that is, a histogram of $\sigma_\epsilon^{2(j+15,000)}$ *for* $j = 1, 2, \ldots, 1000$. *Superimposed is the approximate (supposed) marginal posterior density of* σ_ϵ^2. *An appropriately scaled version of* $\hat{\pi}_{\sigma_\epsilon^2|\mathbf{y}}(t|\mathbf{y})$ *is on the ordinate with* t *on the abscissa.*