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

James P. Hobert; George Casella

Journal of the American Statistical Association, Vol. 91, No. 436. (Dec., 1996), pp. 1461-1473.

Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28199612%2991%3A436%3C1461%3ATEOIPO%3E2.0.CO%3B2-1

Journal of the American Statistical Association is currently published by American Statistical Association.

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at http://www.jstor.org/about/terms.html. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at http://www.jstor.org/journals/astata.html.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is an independent not-for-profit organization dedicated to creating and preserving a digital archive of scholarly journals. For more information regarding JSTOR, please contact support@jstor.org.

http://www.jstor.org/
Tue Aug 22 10:27:32 2006

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

James P. HOBERT and George CASELLA

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 used to explore the posterior distribution without ever having established propriety of the posterior. An example is given showing that the output from a Gibbs chain corresponding to an improper posterior may appear perfectly reasonable. Thus one cannot 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 that 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.

KEY WORDS: 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 (Gelfand and Smith 1990; Geman and Geman 1984) is an easily implemented simulation technique that 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 are often modeled with improper priors, because of either convenience or a lack of prior information. 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 Secs. 3 and 4).

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

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

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 $N(0, \sigma_\varepsilon^2)$. The u_i's and ε_{ij}'s are assumed to be independent. The overall mean, β, and the variance components, σ^2 and σ_ε^2, are considered to be unknown parameters.

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 Charles E. McCulloch, Brett Presnell, Christian Robert, Marty Wells, the associate editor, and three referees for helpful comments and suggestions.

This model fits nicely into a (Bayesian) conditionally independent hierarchical model (Kass and Steffey 1989) by writing (1) as a two-stage hierarchy and specifying priors on the three unknown parameters

$$y_{ij}|\beta, \mathbf{u}, \sigma_\varepsilon^2 \sim N(\beta + u_i, \sigma_\varepsilon^2),$$

$$\beta \sim \pi(\beta) \qquad \mathbf{u} \sim N_k(\mathbf{0}, \mathbf{I}\sigma^2) \qquad \sigma_\varepsilon^2 \sim \pi(\sigma_\varepsilon^2),$$

and

$$\sigma^2 \sim \pi(\sigma^2), \tag{2}$$

where $\mathbf{u} = (u_1, \ldots, u_k)'$ and the priors $\pi(\beta), \pi(\sigma_\varepsilon^2)$, and $\pi(\sigma^2)$ 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 (2) to write the posterior density as

$$\pi(\sigma^2, \sigma_\varepsilon^2, \mathbf{u}, \beta|\mathbf{y}) \propto f(\mathbf{y}|\beta, \mathbf{u}, \sigma_\varepsilon^2) f(\mathbf{u}|\sigma^2) \pi(\beta) \pi(\sigma_\varepsilon^2) \pi(\sigma^2) \tag{3}$$

and then pick the functional forms of each of the necessary conditionals off the right side of (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 $\pi(\sigma^2, \sigma_\varepsilon^2, \mathbf{u}, \beta|\mathbf{y})$ is considered to be the conditional density of the parameters given the data and is used as a joint density in the parameters that happen to involve the fixed, known quantity $\mathbf{y}$. Thus, for example, $f(\sigma^2|\sigma_\varepsilon^2, \mathbf{u}, \beta, \mathbf{y})$ is given by

© 1996 American Statistical Association
Journal of the American Statistical Association
December 1996, Vol. 91, No. 436, Theory and Methods

$\pi(\sigma^2, \sigma_\varepsilon^2, \mathbf{u}, \beta|\mathbf{y}) / \int \pi(\sigma^2, \sigma_\varepsilon^2, \mathbf{u}, \beta|\mathbf{y})\, d\sigma^2$, and because the denominator is constant with respect to σ^2, as a function of σ^2, $f(\sigma^2|\sigma_\varepsilon^2, \mathbf{u}, \beta, \mathbf{y})$ is proportional to the right side of (3). Each of the Gibbs conditionals can be computed in this manner.

A specific example of model (2) discussed by Hill (1965) and Tiao and Tan (1965) has $\pi(\beta) \propto 1, \pi(\sigma_\varepsilon^2) \propto 1/\sigma_\varepsilon^2$ and $\pi(\sigma^2) \propto 1/\sigma^2$. 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 (3) is used to calculate the Gibbs conditionals, as described earlier and we find that (σ^2|all others) and (σ_ε^2|all others) have inverted gamma forms and ($\mathbf{u}$|all others) and (β|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. However, Hill (1965) showed that the posterior for this model is improper (and suggested several alternatives, including Jeffreys's prior). Thus, although we used it to calculate the Gibbs conditionals, (3) is meaningless because the right side is not integrable. Impropriety implies that there does not exist a joint density to which the Gibbs conditionals correspond, and thus they are called incompatible conditional densities (Arnold and Press 1989).

This example suggests the two important questions addressed in this article:

1. Which improper priors yield proper posteriors in the general hierarchical linear mixed model?
2. In general, what can be said about Gibbs Markov chains that correspond to improper posteriors?

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

In Section 4 we give a general discussion of Gibbs sampling when the posterior is improper. 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 (2) is given at the end of Section 4. In Section 5 we present 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 setup include standard forms, such as those discussed by 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 that lead to proper posteriors.

Our parametric prior specification leads to a manageable set of Gibbs conditionals, identical in form to those given by Gelfand and Smith (1990), who described 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

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}, \tag{4}$$

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 $n \times q$, and $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of residual errors.

One may object to the use of the term "fixed effect" in a Bayesian model, because, 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

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

and

$$\text{(b)} \quad \boldsymbol{\varepsilon}|\sigma_\varepsilon^2 \sim N_n(\mathbf{0}, \mathbf{I}\sigma_\varepsilon^2), \tag{5}$$

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

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

$$\pi(\boldsymbol{\beta}) \propto 1 \qquad \mathbf{u}|\sigma_1^2, \ldots, \sigma_r^2 \sim N_q(\mathbf{0}, \mathbf{D}),$$

$$\pi_\varepsilon(\sigma_\varepsilon^2|b) \propto (\sigma_\varepsilon^2)^{-(b+1)},$$

and

$$\pi_i(\sigma_i^2|a_i) \propto (\sigma_i^2)^{-(a_i+1)}, \tag{6}$$

where the a_i's and b are known and the following conditional independence assumptions are in force: (a) given $\mathbf{u}, \mathbf{y}$ is conditionally independent of $\sigma_1^2, \ldots, \sigma_r^2$; (b) given $\sigma_1^2, \ldots, \sigma_r^2$, $\mathbf{u}$ is conditionally independent of $\boldsymbol{\beta}$ and σ_ε^2; and (c) $\boldsymbol{\beta}, \sigma_\varepsilon^2$, 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 (Datta 1992; Wang et al. 1993, 1994) and small area estimation (Datta and Ghosh 1991; Ghosh 1994). 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 (6) 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 (6). The frequentist version of model (6) is, of course,

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

and

$$\mathbf{u}|\sigma_1^2, \ldots, \sigma_r^2 \sim N_q(\mathbf{0}, \mathbf{D}), \tag{7}$$

where $\boldsymbol{\beta}, \sigma_\varepsilon^2$, 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).

$$\begin{aligned} &L(\sigma_1^2, \ldots, \sigma_r^2, \sigma_\varepsilon^2, \boldsymbol{\beta}|\mathbf{y}) \\ &\quad \stackrel{\text{def}}{=} f(\mathbf{y}|\sigma_1^2, \ldots, \sigma_r^2, \sigma_\varepsilon^2, \boldsymbol{\beta}) \\ &\quad = \int_{\Re^q} f(\mathbf{y}|\mathbf{u}, \sigma_\varepsilon^2, \boldsymbol{\beta}) f(\mathbf{u}|\sigma_1^2, \ldots, \sigma_r^2)\, d\mathbf{u}. \end{aligned} \tag{8}$$

Consider the Bayesian model obtained by taking $a_1 = \cdots = a_r = b = -1$ in (6); that is, 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, then the resulting marginal posterior is proportional to the likelihood function in (8); that is,

$$\begin{aligned} &L(\sigma_1^2, \ldots, \sigma_r^2, \sigma_\varepsilon^2, \boldsymbol{\beta}|\mathbf{y}) \\ &\quad \propto \pi_l(\sigma_1^2, \ldots, \sigma_r^2, \sigma_\varepsilon^2, \boldsymbol{\beta}|\mathbf{y}) \\ &\quad = \int_{\Re^q} \pi_l(\sigma_1^2, \ldots, \sigma_r^2, \sigma_\varepsilon^2, \mathbf{u}, \boldsymbol{\beta}|\mathbf{y})\, d\mathbf{u}. \end{aligned} \tag{9}$$

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 that maximize this likelihood are called the 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 (8). Thus the connection between the restricted likelihood function and the posterior distribution corresponding to the flat prior hierarchy is

$$\begin{aligned} &L_r(\sigma_1^2, \ldots, \sigma_r^2, \sigma_\varepsilon^2|\mathbf{y}) \\ &\quad = \int_{\Re^p} L(\sigma_1^2, \ldots, \sigma_r^2, \sigma_\varepsilon^2, \boldsymbol{\beta}|\mathbf{y})\, d\boldsymbol{\beta} \\ &\quad \propto \int_{\Re^p} \pi_l(\sigma_1^2, \ldots, \sigma_r^2, \sigma_\varepsilon^2, \boldsymbol{\beta}|\mathbf{y})\, d\boldsymbol{\beta} \\ &\quad = \pi_l(\sigma_1^2, \ldots, \sigma_r^2, \sigma_\varepsilon^2|\mathbf{y}). \end{aligned} \tag{10}$$

The ability to classify the posterior distributions corresponding to the flat prior hierarchy, π_l, as proper or not is useful when using the Gibbs sampler to explore likelihood functions via (9) and (10) is desired. We give an example 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 (6) lends itself to Gibbs sampling. Assume that $2a_i > -q_i$ for all i and $2b > -n$. Use the conditional independence assumptions to write the posterior as

$$\begin{aligned} &\pi(\sigma_1^2, \ldots, \sigma_r^2, \sigma_\varepsilon^2, \mathbf{u}, \boldsymbol{\beta}|\mathbf{y}) \\ &\quad \propto f(\mathbf{y}|\mathbf{u}, \sigma_\varepsilon^2, \boldsymbol{\beta}) f(\mathbf{u}|\sigma_1^2, \ldots, \sigma_r^2) \pi(\boldsymbol{\beta}) \pi_\varepsilon(\sigma_\varepsilon^2|b) \\ &\qquad \times \prod_{i=1}^{r} \pi_i(\sigma_i^2|a_i), \end{aligned} \tag{11}$$

where f is used to represent a generic density.

The functional forms of the Gibbs conditionals can be picked off of the right side of (11) as discussed in Section

1. The results are as follows:

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

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

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

and

$$f(\boldsymbol{\beta}|\sigma_1^2,\ldots,\sigma_r^2,\mathbf{y},\sigma_\varepsilon^2,\mathbf{u}) = N_p((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{y}-\mathbf{Z}\mathbf{u}),\sigma_\varepsilon^2(\mathbf{X}'\mathbf{X})^{-1}), \quad (12)$$

where IG stands for inverted gamma and we say that $X\sim \mathrm{IG}(r,s)$ 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, because 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 (12) implies propriety of the posterior distribution, the example in Section 1 shows that this is false. Indeed, there are many values of the vector $(a_1,a_2,\ldots,a_r,b)$ that simultaneously yield proper conditionals ($2a_i>-q_i$ for all 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 (11), then 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 $t=\mathrm{rank}(\mathbf{P}_X\mathbf{Z})=\mathrm{rank}(\mathbf{Z}'\mathbf{P}_X\mathbf{Z})\le q$ where we define $\mathbf{P}_X=(\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')$. There are two cases:

1. If $t=q$ or if $r=1$ then the following conditions (a), (b), and (c) are necessary and sufficient for the propriety of the posterior distribution of model (6):

(a) $a_i<0$,
(b) $q_i>q-t-2a_i$,
(c) $n+2\sum a_i+2b-p>0$.

2. If $t<q$ and $r>1$ then the foregoing conditions (a), (b), and (c) are sufficient for the propriety of the posterior distribution of model (6) while necessary conditions result when (b) is replaced with (b′) $q_i>-2a_i$.

3. APPLICATIONS

3.1 An Animal Breeding Model

The following hierarchical model was used by Wang et al. (1993, 1994) for the Bayesian analysis of animal breeding data:

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

$$\pi(\boldsymbol{\beta})\propto 1 \qquad \mathbf{u}_i|\sigma_i^2\sim N_{q_i}(\mathbf{0},\mathbf{G}_i\sigma_i^2),$$

$$\sigma_\varepsilon^2|\nu_\varepsilon,s_\varepsilon^2\sim \mathrm{IG}\left(\frac{\nu_\varepsilon}{2},\frac{2}{\nu_\varepsilon s_\varepsilon^2}\right),$$

and

$$\sigma_i^2|\nu_i,s_i^2\sim \mathrm{IG}\left(\frac{\nu_i}{2},\frac{2}{\nu_i s_i^2}\right), \quad (13)$$

where the $\mathbf{u}_i$'s are independent and the other model assumptions are the same as those of model (6) except, as mentioned earlier, 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 that "contain functions of known coefficients of coancestry" (Wang et al. 1993). These authors suggested that the hyperparameters $s_\varepsilon^2,s_1^2,\ldots,s_r^2$ be assigned the prior value of the corresponding variance component and referred to $\nu_\varepsilon,\nu_1,\ldots,\nu_r$ as "degree of belief" parameters. The values of these parameters are subjectively chosen to reflect the faith that the experimenter/statistician have in $s_\varepsilon^2,s_1^2,\ldots,s_r^2$ as prior estimates of the variance components. This parameterization is considered to be intuitively pleasing in that

$$E(\sigma_i^2)=\frac{\nu_i s_i^2}{\nu_i-2} \quad\text{and}\quad \mathrm{var}(\sigma_i^2)=\frac{2\nu_i^2 s_i^4}{(\nu_i-2)^2(\nu_i-4)} \quad (14)$$

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 $E(\sigma_i^2)-s_i^2$ both go to zero as the degree of belief goes to infinity. Wang et al. (1993) used model (13) with the degree of belief parameters all set to zero, which is supposed to reflect prior ignorance about the variance components. We now use Theorem 1 to show that this prior specification leads to an improper posterior distribution.

In model (13), if we reparameterize from

$$(\mathbf{u}_1,\ldots,\mathbf{u}_r,\sigma_1^2,\ldots,\sigma_r^2,\sigma_\varepsilon^2,\boldsymbol{\beta})$$

to

$$(\mathbf{u}_1^*,\ldots,\mathbf{u}_r^*,\sigma_1^2,\ldots,\sigma_r^2,\sigma_\varepsilon^2,\boldsymbol{\beta}),$$

where $\mathbf{u}_i^*=\mathbf{G}_i^{-1/2}\mathbf{u}_i$, then model (13) with degree of belief parameters set to zero can be written in the form (6) with $\mathbf{u}=(\mathbf{u}_1^{*\prime},\ldots,\mathbf{u}_r^{*\prime})'$, $\mathbf{Z}=[\mathbf{Z}_1\mathbf{G}_1^{1/2}\mathbf{Z}_2\mathbf{G}_2^{1/2},\ldots,\mathbf{Z}_r\mathbf{G}_r^{1/2}]$, a partitioned matrix, and the a_i's and b all set to zero. Theorem 1 may now be applied and implies that the posterior distribution $\pi(\sigma_\varepsilon^2,\sigma_1^2,\ldots,\sigma_r^2,\mathbf{u}_1^*,\ldots,\mathbf{u}_r^*,\boldsymbol{\beta}|\mathbf{y})$ is improper, because condition (a) is violated. This clearly implies the im-

propriety of $\pi(\sigma_\varepsilon^2, \sigma_1^2, \ldots, \sigma_r^2, \mathbf{u}_1, \ldots, \mathbf{u}_r, \boldsymbol{\beta}|\mathbf{y})$ as well, because the difference is only a linear transformation. Wang et al. (1993) actually pointed out that the posterior is improper, but used the Gibbs sampler and found no suggestion from the output that the posterior is improper. We discuss the futility of this type of application of the Gibbs sampler, and the fictitious answers it can given, in Section 4.

Wang et al. (1994) suggested that zero degree of belief parameters should not be used, because the resulting posterior is improper. They suggested 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 provided no proof of this. (To get flat priors out of the inverted gamma priors in model (13), 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, the condition (b) 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 (albeit computer-intensive) rank calculation and no integration at all.

3.2 Likelihood Estimation

Closed-form solutions for ML and REML estimators of the parameters in model (7) 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, chap. 8). These algorithms are not guaranteed to locate the global maximum of a likelihood function from an arbitrary starting point, but they 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 (6). Suppose that 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 $(\sigma_\varepsilon^{2(j)}, \sigma_1^{2(j)}, \ldots, \sigma_r^{2(j)}, \mathbf{u}'^{(j)}, \boldsymbol{\beta}'^{(j)}), j \geq 1$, and let $\boldsymbol{\sigma}^{2(j)} = (\sigma_\varepsilon^{2(j)}, \sigma_1^{2(j)}, \ldots, \sigma_r^{2(j)})'$. 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 (10) 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 have 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 have been discussed by Casella and Berger (1994), Geyer (1991), and Geyer and Thompson (1992).

4. GIBBS SAMPLING WITH IMPROPER POSTERIORS

If a complicated hierarchical model with improper priors is postulated, then 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. However, if propriety has not been demonstrated, then calculating these densities via recognition requires assuming (possibly incorrectly) that a proportionality like (11) holds. If a set of proper densities results, then 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 used in conjunction with a set of conditionals corresponding to an improper posterior distribution. Gelfand et al. (1990) and Wang et al. (1993) analyzed 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 our model (6) 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 of the aforementioned articles showed plots of approximate marginal posterior densities and gave 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 thus 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 Gelfand et al. 1990 and Wang et al. 1993), 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 that 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, because 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 that 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), we call $f_{X|Y}(x|y)$ and $f_{Y|X}(y|x)$ "compatible conditional densities" when such a joint density exists. In general, we call them "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

$$f_{X|Y}(x|y) = \frac{g(x, y)}{\int_{N_x} g(x, y)\, dx}$$

and

$$f_{Y|X}(y|x) = \frac{g(x, y)}{\int_{N_y} g(x, y)\, dy}, \tag{15}$$

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, because the conditions of Definition 1 are satisfied using $g(x, y) = \exp(-xy)$. However, this does not imply compatibility, because $\exp(-xy)$ is not integrable.

To make our point without introducing unwieldy notation, we stick to the simple example concerning two candidate conditional densities. But it is a simple matter 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 thus take the liberty of referring to the general results.

Clearly, Gibbs conditionals that are calculated via proportionalities like (11) and (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 that 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 and define two functions:

$$g_1(x, y) = \frac{f_{X|Y}(x|y) f_{Y|X}(y|x_0)}{f_{X|Y}(x_0|y)}$$

and

$$g_2(x, y) = \frac{f_{Y|X}(y|x) f_{X|Y}(x|y_0)}{f_{Y|X}(y_0|x)}. \tag{16}$$

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 (16).)

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, then one first should 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] proved 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

$$Y_i \sim f_{Y|X}(\cdot|x_i)$$

and

$$X_{i+1} \sim f_{X|Y}(\cdot|y_i)$$

for $i = 1, 2, \ldots$. The foregoing results 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 that 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

$$P((x,y),A) = \int_A f_{X|Y}(s|y) f_{Y|X}(t|s)\, d(s,t) \qquad (17)$$

(Robert 1995; Schervish and Carlin 1990). A measure ν on the set N is an invariant measure for this Markov chain if for any $A \subseteq N$,

$$\nu(A) = \int_N \nu\,(d(x,y)) P((x,y),A). \qquad (18)$$

The term "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 for all i. Define a measure ν_G on the set N as follows:

$$\nu_G(A) = \int_A g(x,y)\, d(x,y). \qquad (19)$$

Theorem 4. The measure ν_G is an invariant measure for the Gibbs chain with Markov transition function $P(\cdot,\cdot)$ in (17).

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 1994) 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 that 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 thus is null; that is, 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 Sec. 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 that $(X,Y) \in \Re^2_+$ is the parameter in a Bayesian model that 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, whereas the marginal of X has an infinite amount of mass in the limit toward infinity. As we might expect, when the Gibbs chain with $d = \frac{1}{2}$ 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, whereas the marginal of Y has an infinite amount of mass in the limit toward infinity. When the Gibbs chain with $d = \frac{3}{2}$ 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 toward infinity. Surprisingly, this Gibbs chain is relatively well behaved, periodically returning to the origin in between long charges toward 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

$$\mathbf{X} \sim N_2(\boldsymbol{\mu}, \mathbf{I}),$$

$$\boldsymbol{\mu} \sim N_2(\mathbf{1}\theta, \mathbf{I}\tau^2),$$

and

$$\pi(\theta, \tau^2) \propto \tau^{-2}. \qquad (20)$$

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 1, 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.)

Because 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, to make the posterior proper. In fact, when the Gibbs sampler is started

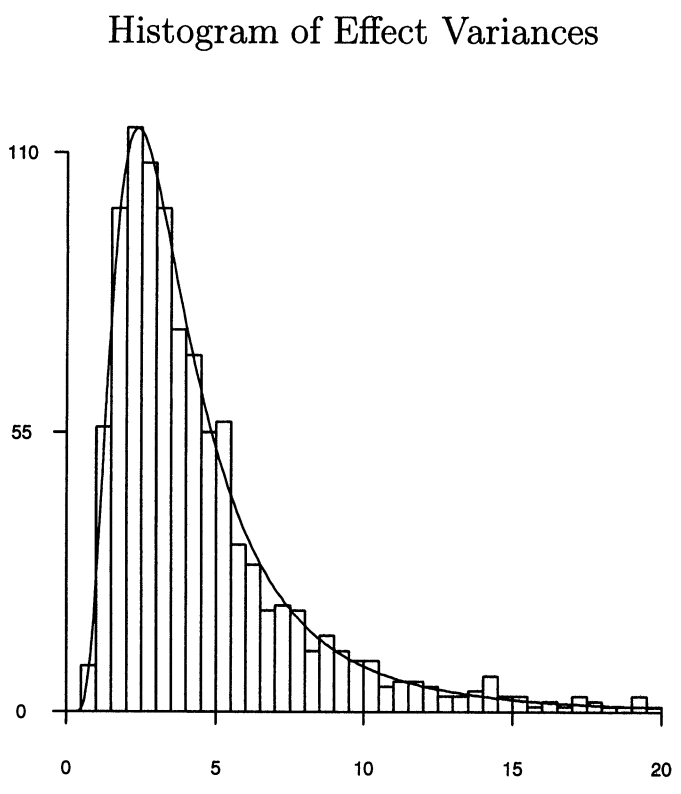

Figure 1. Histogram of the 1,000 Values of the Effect Variance From the Null Gibbs Chain; That is, a Histogram of $\sigma^{2(j+15,000)}$ for j = 1, 2, ..., 1,000. Superimposed is the approximate (supposed) marginal posterior density of σ^2. An approprimately scaled version of $\hat{\pi}_{\sigma^2|\mathbf{y}}(t|\mathbf{y})$ 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.)

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 $\mathrm{IG}(1, 2/((\mu_1-\theta)^2+(\mu_2-\theta)^2))$ 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 $(0, \varepsilon)$. Thus if the starting value of τ^2 is relatively large, then there probably will 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. 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 Gelfand et al. (1990) and Wang et al. (1993) as "fictitious," this was based on the observation that in the problems that they analyzed, there can be no conclusions about a posterior distribution, because such a distribution does not exist.

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 $J = 5$. To simulate some data, we set $\sigma^2 = 5, \sigma_\varepsilon^2 = 2$, and $\beta = 10$. The vector $(u_1, \ldots, u_7)$ was simulated by generating 7 iid $N(0, 5)$ random variables, and the vector $(\varepsilon_{11}, \ldots, \varepsilon_{75})$ was simulated by generating 35 iid $N(0, 2)$ random variables. These numbers were combined according to (1). We use the hierarchical model (2) with the priors $\pi(\beta) \propto 1, \pi(\sigma_\varepsilon^2) \propto 1/\sigma_\varepsilon^2$, and $\pi(\sigma^2) \propto 1/\sigma^2$, which yield an improper posterior (Hill 1965) or take $a = b = 0$ in (6) and use Theorem 1. A Gibbs chain was constructed using the conditionals given in (2.9). We denote the chain by $(\sigma^{2(j)}, \sigma_\varepsilon^{2(j)}, \mathbf{u}'^{(j)}, \beta^{(j)}), j \geq 1$. At the start, all parameters were set to 1, except for the overall mean, β, which was set to 8. 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 thus is 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, then neither would the chain of an unknowing experimenter.

After the initial 15,000 iterations, the output from iterations 15,001 through 16,000 was collected. Figure 1 is a histogram of the 1,000 effect variances from the null Gibbs chain; that is, $\sigma^{2(j+15,000)}, j = 1, 2, \ldots, 1{,}000$, 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, then 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 foregoing discussion, one might expect that if the starting value of σ^2 were near 0, then 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 compo-

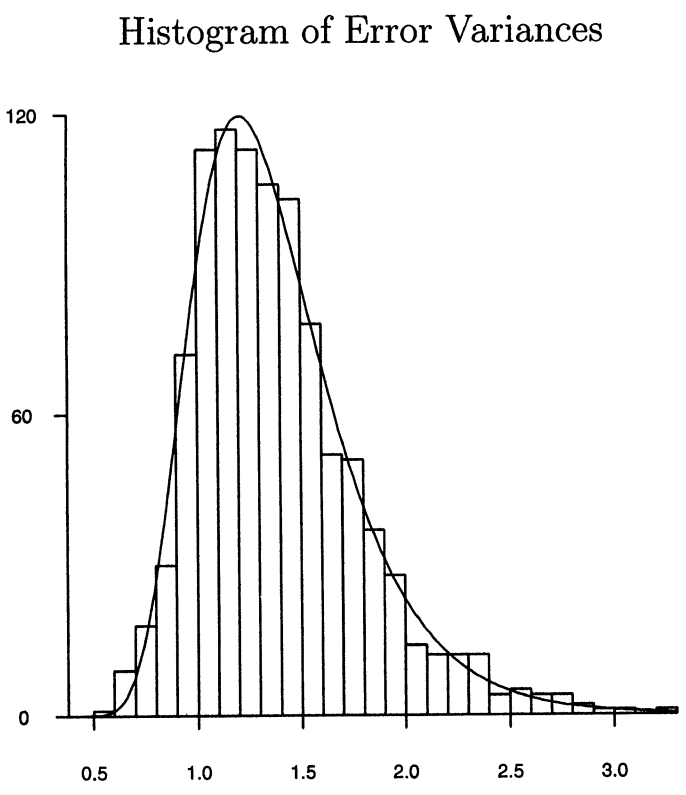

Figure 2. Histogram of the 1,000 Values of the Error Variance From the Null Gibbs Chain; That is, a Histogram of $\sigma_\varepsilon^{2(j+15,000)}$ for $j = 1, 2, \ldots, 1{,}000$. Superimposed is the approximate (supposed) marginal posterior density of σ_ε^2. An appropriately scaled version of $\hat{\pi}_{\sigma_\varepsilon^2|\mathbf{y}}(t|\mathbf{y})$ is on the ordinate with t on the abscissa.

nents move toward 0, 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 1 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 chain constructed with the exponential conditionals in that it leaves the "center" of the space for long periods 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 (3) in Sec. 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 used. The foregoing examples and those discussed 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 (12).

We have experimented with some diagnostics for null Markov chains (see 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, because 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 of Dey, Gelfand, and Peng (1994) and Ibrahim and Laud (1991), which give sufficient conditions for propriety of posteriors for classes of improper priors.

APPENDIX: THEOREM PROOFS

Proof of Theorem 1

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

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

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

Lemma 1. If c is a scalar and $\mathbf{S}$ is nonnegative 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. Because $\mathbf{S}$ 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

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

Assume without loss of generality that $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1(\mathbf{S}), \lambda_2(\mathbf{S}), \ldots, \lambda_t(\mathbf{S}), 0, \ldots, 0)$, where $t \leq v$ is the rank of $\mathbf{S}$. Then

$$\begin{aligned}&\mathbf{\Lambda}\left[\mathbf{\Lambda} + \frac{\mathbf{I}}{c}\right]^{-1}\mathbf{\Lambda} \\ &\qquad = \operatorname{diag}\left(\frac{c\lambda_1^2}{c\lambda_1 + 1}, \frac{c\lambda_2^2}{c\lambda_2 + 1}, \ldots, \frac{c\lambda_t^2}{c\lambda_t + 1}, 0, \ldots, 0\right)\end{aligned}$$

and the result follows because

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

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

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

We first prove case (1) of Theorem 1. From Equation (11), we have

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

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

$$m(\mathbf{y}) = \int f(\mathbf{y}|\mathbf{u}, \sigma_\varepsilon^2, \boldsymbol{\beta}) f(\mathbf{u}|\sigma_1^2, \ldots, \sigma_r^2) \pi(\boldsymbol{\beta}) \pi_\varepsilon(\sigma_\varepsilon^2|b) \times \prod_{i=1}^{r} \pi_i(\sigma_i^2|a_i)\, d\mathbf{u}\, d\boldsymbol{\beta}\, d\sigma_\varepsilon^2 \prod_{i=a1}^{r} d\sigma_i^2. \tag{A.4}$$

It is straightforward to show that

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

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

$$\mathbf{M}_1 = (\mathbf{Z}\mathbf{D}\mathbf{Z}' + \mathbf{I}\sigma_\varepsilon^2)^{-1}$$

and

$$\mathbf{M}_2 = (\mathbf{X}'(\mathbf{Z}\mathbf{D}\mathbf{Z}' + \mathbf{I}\sigma_\varepsilon^2)^{-1}\mathbf{X})^{-1}. \tag{A.6}$$

Note that (A.5) is the restricted likelihood function in (10). We may now write

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

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

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

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

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

The next step is to find an upper bound. We let the variance components go to ∞ at the same rate, so we 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

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

Lemma 1 and (A.10) together give

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

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

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

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

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

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

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

Conditions (a), (b), and (c) are first shown to be sufficient for integrability. Using the upper bound,

$$\begin{aligned}\int & f(\mathbf{y}|\mathbf{u}, \sigma_\varepsilon^2, \boldsymbol{\beta}) f(\mathbf{u}|\sigma_1^2, \ldots, \sigma_r^2) \pi(\boldsymbol{\beta}) \pi_\varepsilon(\sigma_\varepsilon^2|b) \times \prod_{i=1}^{r} \pi_i(\sigma_i^2|a_i)\, d\mathbf{u}\, d\boldsymbol{\beta} \prod_{i=1}^{r} d\sigma_i^2 \\ &\le \frac{\pi_\varepsilon(\sigma_\varepsilon^2|b) \exp\left\{-\frac{1}{2\sigma_\varepsilon^2}\, \mathbf{y}'(\mathbf{P}_Z - \mathbf{P}_Z\mathbf{X}(\mathbf{X}'\mathbf{P}_Z\mathbf{X})^{-}\mathbf{X}'\mathbf{P}_Z)\mathbf{y}\right\}}{(2\pi)^{N-p/2}(\sigma_\varepsilon^2)^{N-q-p/2}|\mathbf{X}'\mathbf{X}|^{1/2}} \\ &\quad \times \int \frac{\prod_{i=1}^{r} \pi_i(\sigma_i^2|a_i)}{|\mathbf{D}|^{1/2}|\sigma_\varepsilon^2\mathbf{D}^{-1} + \mathbf{Z}'\mathbf{P}_X\mathbf{Z}|^{1/2}} \prod_{i=1}^{r} d\sigma_i^2. \end{aligned} \tag{A.15}$$

Focusing on the integrand, Lemma 2 gives

$$\prod_{i=1}^{q}[\lambda_i(\sigma_\varepsilon^2\mathbf{D}^{-1})+\lambda_i(\mathbf{Z}'\mathbf{P}_X\mathbf{Z})]$$
$$\le |\sigma_\varepsilon^2\mathbf{D}^{-1}+\mathbf{Z}'\mathbf{P}_X\mathbf{Z}|$$
$$\le \prod_{i=1}^{q}[\lambda_i(\sigma_\varepsilon^2\mathbf{D}^{-1})+\lambda_{q-i+1}(\mathbf{Z}'\mathbf{P}_X\mathbf{Z})]. \quad (A.16)$$

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

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

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

$$\int \frac{\prod_{i=1}^{r}\pi_i(\sigma_i^2|a_i)}{|\mathbf{D}|^{1/2}|\sigma_\varepsilon^2\mathbf{D}^{-1}+\mathbf{Z}'\mathbf{P}_X\mathbf{Z}|^{1/2}}\prod_{i=1}^{r}d\sigma_i^2$$
$$\le \prod_{i=1}^{r}\int \frac{(\sigma_i^2)^{-(a_i+1)}}{(\lambda_{\min}(\mathbf{Z}'\mathbf{P}_X\mathbf{Z})\sigma_i^2+\sigma_\varepsilon^2)^{q_i/2}}\,d\sigma_i^2. \quad (A.18)$$

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

$$l^{q_i/2}\int \frac{t^{-(a_i+1)}}{(l\sigma_\varepsilon^2+t)^{q_i/2}}\,dt, \quad (A.19)$$

where $l = (\lambda_{\min}(\mathbf{Z}'\mathbf{P}_X\mathbf{Z}))^{-1}$. 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 $c_i(\sigma_\varepsilon^2)^{-(a_i+q_i/2)}$, where c_i is constant in σ_ε^2. Therefore, because we have assumed that these two conditions hold for $i = 1, 2, \ldots, r$, we may write

$$\int f(\mathbf{y}|\mathbf{u},\sigma_\varepsilon^2,\boldsymbol{\beta})f(\mathbf{u}|\sigma_1^2,\ldots,\sigma_r^2)\pi(\boldsymbol{\beta})\pi_\varepsilon(\sigma_\varepsilon^2|b)$$
$$\times\prod_{i=1}^{r}\pi_i(\sigma_i^2|a_i)\,d\mathbf{u}\,d\boldsymbol{\beta}\prod_{i=1}^{r}d\sigma_i^2$$
$$\le \frac{\pi_\varepsilon(\sigma_\varepsilon^2|b)\exp\left\{-\frac{1}{2\sigma_\varepsilon^2}\,\mathbf{y}'(\mathbf{P}_Z-\mathbf{P}_Z\mathbf{X}(\mathbf{X}'\mathbf{P}_Z\mathbf{X})^{-}\mathbf{X}'\mathbf{P}_Z)\mathbf{y}\right\}}{(2\pi)^{N-p/2}(\sigma_\varepsilon^2)^{N-q-p/2}|\mathbf{X}'\mathbf{X}|^{1/2}}$$
$$\times\, c\cdot(\sigma_\varepsilon^2)^{-(q/2+\Sigma a_i)}. \quad (A.20)$$

Sufficiency will follow if it can be shown that (A.20) is integrable with respect to σ_ε^2. This follows from condition (c) 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 aforementioned bounds. Using the lower bound on the exponential in (A.14) and the upper bound in (A.17) yields

$$\int f(\mathbf{y}|\mathbf{u},\sigma_\varepsilon^2,\boldsymbol{\beta})f(\mathbf{u}|\sigma_1^2,\ldots,\sigma_r^2)\pi(\boldsymbol{\beta})\pi_\varepsilon(\sigma_\varepsilon^2|b)$$
$$\times\prod_{i=1}^{r}\pi_i(\sigma_i^2|a_i)\,d\mathbf{u}\,d\boldsymbol{\beta}\prod_{i=1}^{r}d\sigma_i^2$$
$$\ge \frac{\pi_\varepsilon(\sigma_\varepsilon^2|b)\exp\left\{-\frac{1}{2\sigma_\varepsilon^2}\,\mathbf{y}'\mathbf{P}_X\mathbf{y}\right\}}{(2\pi)^{N-p/2}(\sigma_\varepsilon^2)^{N-q-p/2}|\mathbf{X}'\mathbf{X}|^{1/2}}$$
$$\times\prod_{i=1}^{r}\int \frac{(\sigma_i^2)^{-(a_i+1)}}{(\lambda_{\max}(\mathbf{Z}'\mathbf{P}_X\mathbf{Z})\sigma_i^2+\sigma_\varepsilon^2)^{q_i/2}}\,d\sigma_i^2. \quad (A.21)$$

This inequality demonstrates the necessity of conditions (a) and (b), because the right integral will diverge if either (or both) fails to hold. If both hold, then an argument similar to one earlier shows that

$$\int f(\mathbf{y}|\mathbf{u},\sigma_\varepsilon^2,\boldsymbol{\beta})f(\mathbf{u}|\sigma_1^2,\ldots,\sigma_r^2)\pi(\boldsymbol{\beta})\pi_\varepsilon(\sigma_\varepsilon^2|b)$$
$$\times\prod_{i=1}^{r}\pi_i(\sigma_i^2|a_i)\,d\mathbf{u}\,d\boldsymbol{\beta}\prod_{i=1}^{r}d\sigma_i^2$$
$$\ge \frac{\pi_\varepsilon(\sigma_\varepsilon^2|b)\exp\left\{-\frac{1}{2\sigma_\varepsilon^2}\,\mathbf{y}'\mathbf{P}_X\mathbf{y}\right\}}{(2\pi)^{(N-p)/2}(\sigma_\varepsilon^2)^{(N-q-p)/2}|\mathbf{X}'\mathbf{X}|^{1/2}}\cdot c'\cdot(\sigma_\varepsilon^2)^{-(q/2+\Sigma a_i)}, \quad (A.22)$$

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

$$|\sigma_\varepsilon^2\mathbf{D}^{-1}+\mathbf{Z}'\mathbf{P}_X\mathbf{Z}|^{1/2} = \left|\mathbf{I}\,\frac{\sigma_\varepsilon^2}{\sigma_1^2}+\mathbf{Z}'\mathbf{P}_X\mathbf{Z}\right|^{1/2}$$
$$= \left|\mathbf{I}\,\frac{\sigma_\varepsilon^2}{\sigma_1^2}+\mathbf{H}'\boldsymbol{\Lambda}\mathbf{H}\right|^{1/2},$$

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

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

where $\lambda_{sp} = \lambda_{sp}(\mathbf{Z}'\mathbf{P}_X\mathbf{Z})$ and $\lambda_{\max} = \lambda_{\max}(\mathbf{Z}'\mathbf{P}_X\mathbf{Z})$. The foregoing proof 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 $|\sigma_\varepsilon^2\mathbf{D}^{-1}+\mathbf{Z}'\mathbf{P}_X\mathbf{Z}|$. In particular, Lemma 1 can be used to show that

$$\prod_{i=1}^{t}[\lambda_i(\sigma_\varepsilon^2\mathbf{D}^{-1})+\lambda_{sp}(\mathbf{Z}'\mathbf{P}_X\mathbf{Z})]\prod_{i=t+1}^{q}[\lambda_i(\sigma_\varepsilon^2\mathbf{D}^{-1})]$$
$$\le |\sigma_\varepsilon^2\mathbf{D}^{-1}+\mathbf{Z}'\mathbf{P}_X\mathbf{Z}|$$
$$\le \prod_{i=1}^{r}\left[\frac{\sigma_\varepsilon^2}{\sigma_i^2}+\lambda_{\max}(\mathbf{Z}'\mathbf{P}_X\mathbf{Z})\right]^{q_i}. \quad (A.24)$$

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 the location in $\Re_+^r$. We avoid this problem by introducing the following mutually exclusive sets:

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

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. Thus we can 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 $(i_1, i_2, \ldots, i_r) = (1, 2, \ldots, r)$. On the set S, the lower bound in (A.24) becomes

$$\prod_{i=1}^{t} [\lambda_i(\sigma_\varepsilon^2 \mathbf{D}^{-1}) + \lambda_{sp}(\mathbf{Z}'\mathbf{P}_X\mathbf{Z})] \prod_{i=t+1}^{q} [\lambda_i(\sigma_\varepsilon^2 \mathbf{D}^{-1})]$$
$$= \left(\frac{\sigma_\varepsilon^2}{\sigma_1^2} + \lambda_{sp}\right)^{q_1} \left(\frac{\sigma_\varepsilon^2}{\sigma_2^2} + \lambda_{sp}\right)^{q_2}$$
$$\cdots \left(\frac{\sigma_\varepsilon^2}{\sigma_r^2} + \lambda_{sp}\right)^{t - \Sigma_{i=1}^{r-1} q_i} \left(\frac{\sigma_\varepsilon^2}{\sigma_r^2}\right)^{q-t},$$

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

$$\int_S \frac{\prod_{i=1}^r \pi_i(\sigma_i^2|a_i)}{|\mathbf{D}|^{1/2}|\sigma_\varepsilon^2\mathbf{D}^{-1} + \mathbf{Z}'\mathbf{P}_X\mathbf{Z}|^{1/2}} \prod_{i=1}^r d\sigma_i^2$$
$$\le \int_S \frac{(\sigma_\varepsilon^2)^{t-q/2} \prod_{i=1}^r \pi_i(\sigma_i^2|a_i)}{(\sigma_\varepsilon^2 + \sigma_r^2\lambda_{sp}(\mathbf{Z}'\mathbf{P}_X\mathbf{Z}))^{t/2 - \Sigma_{i=1}^{r-1} q_i/2} \prod_{i=1}^{r-1} (\sigma_\varepsilon^2 + \sigma_i^2\lambda_{sp}(\mathbf{Z}'\mathbf{P}_X\mathbf{Z}))^{q_i/2}} \prod_{i=1}^r d\sigma_i^2$$
$$\le \int_{\Re_+^r} \frac{(\sigma_\varepsilon^2)^{t-q/2} \prod_{i=1}^r \pi_i(\sigma_i^2|a_i)}{(\sigma_\varepsilon^2 + \sigma_r^2\lambda_{sp}(\mathbf{Z}'\mathbf{P}_X\mathbf{Z}))^{t/2 - \Sigma_{i=1}^{r-1} q_i/2} \prod_{i=1}^{r-1} (\sigma_\varepsilon^2 + \sigma_i^2\lambda_{sp}(\mathbf{Z}'\mathbf{P}_X\mathbf{Z}))^{q_i/2}} \prod_{i=1}^r d\sigma_i^2,$$

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

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 (15).

Proof of Theorem 3

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

Proof of Theorem 4

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

where the third and fourth equalities follow from functional compatibility.

[Received January 1994. Revised January 1996.]

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.

——— (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 481, University of Florida, Dept. 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. (1996), "Fixed and Random Effects and Best Prediction," to appear: Proceedings of the Kansas State Conference on Applied Statistics in Agriculture.

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. (1995), "Convergence Control Methods for Markov Chain Monte Carlo Algorithms," *Statistical Science*, 10, 231–253.

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: Wiley.

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

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. (1994), "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.

——— (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.