Non-Conjugate Bayesian Analysis of Variance Component Models

Russell D. Wolfinger
SAS Institute Inc., SAS Campus Drive,
Cary, NC 27513, U.S.A.

and Robert E. Kass
Department of Statistics, Carnegie Mellon University
Pittsburgh, PA 15213, U.S.A.

February, 1999

Abstract

We consider the usual normal linear mixed model for variance components from a Bayesian viewpoint. With conjugate priors and balanced data, Gibbs sampling is easy to implement. However, simulating from full conditionals can become difficult for the analysis of unbalanced data with possibly non-conjugate priors, thus leading one to consider alternative Markov chain Monte Carlo schemes. We propose and investigate a method for simulating from posterior distribution based on an independence chain. The method is customized to exploit the structure of the variance component model, and it works with arbitrary prior distributions. As a default reference prior, we consider a version of Jeffreys prior based on the integrated ("restricted") likelihood. We demonstrate the ease of application and flexibility of this approach in several familiar settings, even in the presence of unbalanced data.

Some key words: Jeffreys prior, Mixed model, Posterior simulation, Reference prior, REML.

1 Introduction

The general linear mixed model can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\gamma + \epsilon$$

where $\mathbf{y}$ is an observed response vector, $\mathbf{X}$ is an observed design matrix for a vector of unobserved fixed-effects parameters $\boldsymbol{\beta}$, $\mathbf{Z}$ is an observed design matrix for a vector of unobserved random-effects parameters γ, and $\boldsymbol{\epsilon}$ is an unobserved vector of residual errors. The usual assumption is that γ is Normally distributed with mean 0 and variance-covariance matrix $\mathbf{G}$, and ϵ is Normally distributed with mean 0 and variance-covariance matrix $\mathbf{R}$, independently of γ. In this paper we consider Bayesian inference in the important special case in which the components of γ are independent, so that $\mathbf{G}$ is diagonal — this is the usual "components of variance" model. Our goal is to provide a relatively easy-to-use methodology that captures the advantages of Bayesian inference while avoiding some of the familiar obstacles that sometimes impede practitioners.

In approaching this problem we have several concerns. First, we wish to accommodate unbalanced data. Second, we want to be able to use a default or "reference" prior, which will turn out to be non-conjugate (as is often the case), and this introduces some computational complications. Third, we would like the method to allow the user to largely ignore monitoring of the simulation.

The structure of the model lends itself to a two-step approach to posterior simulation as an alternative to simulating from the "full conditional" distributions used in Gibbs sampling and related Markov chain methods. The special circumstances of this problem allow a simple independence chain method based on an analytic approximation to be effective in producing draws from the posterior distribution. A direct comparison of our approach with Gibbs sampling is better made after developing the details of our methodology, and so we defer this comparison to the Discussion in Section 5.

As far as statistical methodology is concerned, the reference Bayesian approach we develop and study also provides a flexible alternative to restricted maximum likelihood (REML) estimation. Our approach permits inference about general functions of parameters and allows investigation of possible effects of small sample size, even with unbalanced data. It also handles gracefully the case in which the REML estimates of one or more variance components are zero.

To be more specific, let $\boldsymbol{\theta}$ denote the vector of all unknown variance components in $\mathbf{G}$ and $\mathbf{R}$. For simplicity, we assume $\boldsymbol{\beta}$ has a uniform prior ($p(\boldsymbol{\beta}|\boldsymbol{\theta}) = 1$ for all $\boldsymbol{\theta}$) independently of γ. We write the joint posterior density of $(\boldsymbol{\beta}, \gamma, \boldsymbol{\theta})$ as

$$p(\boldsymbol{\beta}, \gamma, \boldsymbol{\theta}|\mathbf{y}) = p(\boldsymbol{\beta}, \gamma|\boldsymbol{\theta}, \mathbf{y})p(\boldsymbol{\theta}|\mathbf{y}) \tag{1}$$

and will consider separately the two factors on the right-hand side. Simulation from these two factors will become the two primary steps in the procedure we describe and investigate. In the first step we simulate from $p(\boldsymbol{\theta}|\mathbf{y})$ using an independence chain, and in the second we condition on the simulated value of $\boldsymbol{\theta}$ and draw from a multivariate normal density to simulate from $p(\boldsymbol{\beta}, \gamma|\boldsymbol{\theta}, \mathbf{y})$. Our implementation in Section 3 assumes the prior on $\boldsymbol{\beta}$ is uniform and will allow the user to choose a prior on $\boldsymbol{\theta}$. In our examples in Section 4 we focus on a version of Jeffreys prior based on the REML likelihood, which is presented in Section 2 and may be considered a special case of the Berger-Bernardo method (Berger and Bernardo, 1992). We provide discussion of the contrasts between our method and related alternatives in Section 5.

2 Likelihood Functions, Priors, and Posterior Distributions

The model specified in Section 1 produces the following densities for $\mathbf{y}$ and γ

$$p(\mathbf{y}|\boldsymbol{\beta},\gamma,\boldsymbol{\theta}) = (2\pi)^{-n/2}|\mathbf{R}(\boldsymbol{\theta})|^{-1/2}\exp\{(-1/2)(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\gamma)^T\mathbf{R}(\boldsymbol{\theta})^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\gamma)\} \tag{2}$$

where n is the length of $\mathbf{y}$ and

$$p(\gamma|\boldsymbol{\theta}) = (2\pi)^{-m/2}|\mathbf{G}(\boldsymbol{\theta})|^{-1/2}\exp\{(-1/2)\gamma^T\mathbf{G}(\boldsymbol{\theta})^{-1}\gamma\}. \tag{3}$$

where m is the length of γ. With our assumption of a uniform prior on $\boldsymbol{\beta}$, for all $\boldsymbol{\theta}$, all that is needed to complete the Bayesian specification of the problem is a prior on $\boldsymbol{\theta}$. We will leave this to be selected by the user, but will provide a "default" reference prior to be determined by a version of Jeffreys rule as described below, in Section 2.3. We now provide formulas for the conditional and marginal posteriors in Equation (1).

2.1 Conditional Posterior for $\boldsymbol{\beta}$ and γ

Conditioning on $\boldsymbol{\theta}$ we have the following conditional posterior density (refer to Searle *et al.*, 1992):

$$\begin{aligned} p(\boldsymbol{\beta},\gamma|\mathbf{y},\boldsymbol{\theta}) \quad \propto \quad & (2\pi)^{-n/2}|\mathbf{R}(\boldsymbol{\theta})|^{-1/2}\exp\{(-1/2)(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\gamma)^T\mathbf{R}(\boldsymbol{\theta})^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\gamma)\} \times \\ & (2\pi)^{-m/2}|\mathbf{G}(\boldsymbol{\theta})|^{-1/2}\exp\{(-1/2)\gamma^T\mathbf{G}(\boldsymbol{\theta})^{-1}\gamma\} \end{aligned}$$

Formal manipulation of this expression reveals that conditionally on $\mathbf{y}$ and $\boldsymbol{\theta}$ the posterior distribution of $(\boldsymbol{\beta},\gamma)$ is multivariate Normal with mean

$$\begin{bmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1} \end{bmatrix}^{-} \begin{bmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}$$

and variance-covariance matrix

$$\begin{bmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1} \end{bmatrix}^{-}$$

where $\mathbf{R} = \mathbf{R}(\boldsymbol{\theta})$ and $\mathbf{G} = \mathbf{G}(\boldsymbol{\theta})$. A generalized inverse $(^{-})$ is used here to account for the fact that $\mathbf{X}$ may not be of full rank. Note that the conditional posterior means are the solutions to the well-known mixed model equations (Henderson, 1984), solutions which can also be written as follows:

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= [\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}]^{-}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y} \\ \hat{\gamma} &= \mathbf{G}\mathbf{Z}^T\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}) \end{aligned}$$

where $\mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}^T + \mathbf{R}$. These expressions are also well-known as generalized least squares and best linear unbiased predictor (BLUP) formulas, respectively (see Robinson, 1991, and references cited there). The preceding results have been explored by several authors, including Wang, Rutledge, and Gianola (1994).

2.2 Marginal Posterior for $\boldsymbol{\theta}$

Writing the marginal prior on $\boldsymbol{\theta}$ as $\pi(\boldsymbol{\theta})$, note that $p(\boldsymbol{\theta}|\mathbf{y}) = \int p(\mathbf{y}|\boldsymbol{\beta},\gamma,\boldsymbol{\theta})p(\boldsymbol{\beta},\gamma|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\beta}d\gamma$. If we then define

$$L(\boldsymbol{\theta}) = p(\mathbf{y}|\boldsymbol{\theta}) \tag{4}$$

where $p(\mathbf{y}|\boldsymbol{\theta}) = \int p(\mathbf{y}|\boldsymbol{\beta},\gamma,\boldsymbol{\theta})p(\gamma|\boldsymbol{\theta})d\boldsymbol{\beta}d\gamma$, we have

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto L(\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \tag{5}$$

Here, $L(\boldsymbol{\theta})$ is sometimes called the integrated likelihood function, or restricted likelihood function. Its maximum is the REML estimate of $\boldsymbol{\theta}$ (Patterson and Thompson, 1971; Harville, 1977).

Explicitly, we have

$$L(\boldsymbol{\theta}) \;=\; (2\pi)^{-(n-p)/2}|\mathbf{V}(\boldsymbol{\theta})|^{-1/2}|\mathbf{X}^T\mathbf{V}^{-1}(\boldsymbol{\theta})\mathbf{X}|^{-1/2}\exp\{(-1/2)[\mathbf{y}-\mathbf{X}\mathbf{b}(\boldsymbol{\theta})]^T\mathbf{V}^{-1}(\boldsymbol{\theta})[\mathbf{y}-\mathbf{X}\mathbf{b}(\boldsymbol{\theta})]$$

where

$$\mathbf{b}(\boldsymbol{\theta}) = [\mathbf{X}^T\mathbf{V}^{-1}(\boldsymbol{\theta})\mathbf{X}]^{-}\mathbf{X}^T\mathbf{V}^{-1}(\boldsymbol{\theta})\mathbf{y}$$

with p being the rank of $\mathbf{X}$, and when $\mathbf{X}$ is not of full rank, the determinant $|\mathbf{X}^T\mathbf{V}^{-1}(\boldsymbol{\theta})\mathbf{X}|$ is computed by multiplying its positive eigenvalues.

Given any prior $\pi(\boldsymbol{\theta})$ this formula may be applied to produce the marginal posterior according to (5).

2.3 A Version of Jeffreys Prior on $\boldsymbol{\theta}$

We now consider a reasonable "default" reference prior based on Jeffreys general rule. For a general discussion of reference priors, with extensive references, refer to Kass and Wasserman (1996).

Jeffreys general rule for determining a prior is to take the prior to be proportional to the square-root of the determinant of the Fisher information matrix. In this problem there are two alternative ways that Fisher information might be calculated, based on the two alternative likelihood functions $L(\boldsymbol{\theta})$ in (4) and $L(\boldsymbol{\beta},\boldsymbol{\theta}) = \int p(\mathbf{y}|\boldsymbol{\beta},\gamma,\boldsymbol{\theta})p(\gamma|\boldsymbol{\theta})d\gamma$. We will use the Fisher information matrix $i_R(\boldsymbol{\theta})$ based on $L(\boldsymbol{\theta})$ (and, thus, this method of obtaining the prior becomes a special case of that of Berger and Bernardo, 1992). The computation proceeds as follows.

First, as discussed by Patterson and Thompson (1971), Harville (1974), Corbeil and Searle (1976), and Searle *et al.,* (1992), define $\mathbf{K}$ to be a full-rank $(n-p)\times n$ reduction of the $n\times n$ projection matrix $\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}$. Then

$$-2\log L(\boldsymbol{\theta}) \quad = \quad \log|\mathbf{KVK}'| + \mathbf{y}'\mathbf{K}'(\mathbf{KVK}')^{-1}\mathbf{Ky} + (n-p)\log 2\pi$$

that is, the restricted/residual likelihood can be written as the Normal likelihood based on the set of "error contrasts" $\mathbf{Ky}$. The first and second derivatives of $-2\log L(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}_r$ and $\boldsymbol{\theta}_s$ are $\mathbf{g}$ and $\mathbf{H}$, with elements

$$\begin{aligned}
[\mathbf{g}]_r &= tr\left[(\mathbf{KVK}^T)^{-1}\mathbf{K}\dot{\mathbf{V}}_r\mathbf{K}^T\right] - \mathbf{y}^T\mathbf{K}^T(\mathbf{KVK}^T)^{-1}\mathbf{K}\dot{\mathbf{V}}_r\mathbf{K}^T(\mathbf{KVK}^T)^{-1}\mathbf{Ky} \\
[\mathbf{H}]_{rs} &= -tr\left[(\mathbf{KVK}^T)^{-1}\mathbf{K}\dot{\mathbf{V}}_r\mathbf{K}^T(\mathbf{KVK}^T)^{-1}\mathbf{K}\dot{\mathbf{V}}_s\mathbf{K}^T\right] + \\
&\quad 2\mathbf{y}^T\mathbf{K}^T(\mathbf{KVK}^T)^{-1}\mathbf{K}\dot{\mathbf{V}}_r\mathbf{K}^T(\mathbf{KVK}^T)^{-1}\mathbf{K}\dot{\mathbf{V}}_s\mathbf{K}^T(\mathbf{KVK}^T)^{-1}\mathbf{Ky}
\end{aligned}$$

where $\dot{\mathbf{V}}_r = \partial\mathbf{V}/\partial\boldsymbol{\theta}_r$ and $tr[\ ]$ is the trace operator (refer to Wolfinger *et al.*, 1994). Since $E\left[\mathbf{y}^T A\mathbf{y}\right] = Etr\left[A\mathbf{y}\mathbf{y}^T\right] = tr\left[A\mathbf{V}\right]$ for any matrix A, the entry of the rth row and sth column of the Fisher information matrix is

$$[\mathbf{i}_R(\boldsymbol{\theta})]_{rs} = (1/2)\,tr\left[(\mathbf{KVK}^T)^{-1}\mathbf{K}\dot{\mathbf{V}}_r\mathbf{K}^T(\mathbf{KVK}^T)^{-1}\mathbf{K}\dot{\mathbf{V}}_s\mathbf{K}^T\right]$$

Jeffreys prior for $\boldsymbol{\theta}$ is then proportional to $|\mathbf{i}_R(\boldsymbol{\theta})|^{1/2}$. We can therefore complete our default, reference Bayesian formulation by specifying the marginal posterior of $\boldsymbol{\theta}$:

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto |\mathbf{i}_R(\boldsymbol{\theta})|^{1/2} L(\boldsymbol{\theta}|\mathbf{y})$$

We conjecture that under fairly mild conditions on dimensionality and configuration of the data the posterior based on this prior will be proper, but we do not have a general proof. In special cases it is known to be proper. For example, in the case of a single random effect (see the examples in Sections 4.1 and 4.3), results of Yang and Chen (1996; Theorem 4.3 and Section 6.2) show that $rank\,(\mathbf{Z}) \geq 3$ is sufficient provided the data are not all constant. For nested models such as the split-plot example in Section 4.2 rigorous results do not appear to be available, but the Yang and

Chen results suggest that at least 3 levels of each random effect would be required for posterior propriety. Theorem 1 of Hobert and Casella (1996) provides conditions for inverted gamma priors on the variance components themselves. Perhaps an analogous result can be proven for Jeffreys-type priors.

3 Independence Chain Algorithm

We return to the factorization of the posterior according to equation (1). Using notation from the previous section, the simulation method we use has two steps:

1. Draw a pseudo-random value $\boldsymbol{\theta}^*$ from $p(\boldsymbol{\theta}|\mathbf{y})$.
2. Conditioning on $\boldsymbol{\theta}^*$ from step 1, generate values for $\boldsymbol{\beta}$ and γ from the multivariate Normal distribution $p(\boldsymbol{\beta}, \gamma|\boldsymbol{\theta}, \mathbf{y})$.

Step 2 is easily performed using standard normal random number generators. In this section we show how to perform step 1 with an independence chain.

The independence chain algorithm (e.g., Tierney, 1994), requires a fixed base sampling density $g(\boldsymbol{\theta}|\mathbf{y})$ from which to draw MCMC proposals. The chain begins with a pseudo-random draw from $g(\boldsymbol{\theta}|\mathbf{y})$. Then assuming the chain is currently at state $\boldsymbol{\theta}_0$ the algorithm is to generate a candidate $\boldsymbol{\theta}_*$ from $g(\boldsymbol{\theta}|\mathbf{y})$ and accept it with probability

$$\min\left[\frac{p(\boldsymbol{\theta}_*)/g(\boldsymbol{\theta}_*)}{p(\boldsymbol{\theta}_0)/g(\boldsymbol{\theta}_0)}, 1\right].$$

If $\boldsymbol{\theta}^*$ is accepted, it becomes $\boldsymbol{\theta}_0$; if not, a copy of $\boldsymbol{\theta}_0$ is added to the chain.

The obvious difficulty is in finding an appropriate $g(\boldsymbol{\theta}|\mathbf{y})$. It should be as close as possible to $p(\boldsymbol{\theta}|\mathbf{y})$ to avoid extreme values of the ratios used in the acceptance probability. In the next subsection we formulate such a $g(\boldsymbol{\theta}|\mathbf{y})$ by taking the product of inverted gamma densities. This $g(\boldsymbol{\theta}|\mathbf{y})$ actually equals $p(\boldsymbol{\theta}|\mathbf{y})$ for certain balanced data situations and seems to perform well for unbalanced cases.

3.1 Finding a Base Density

Previous Bayesian analyses of the variance component model (for example, Section 5.2.2 of Box and Tiao, 1973) reveal that $p(\boldsymbol{\theta}|\mathbf{y})$ is equal to or well-approximated by products of the inverted-gamma density

$$IG(x|a,b) = \frac{b^a}{\Gamma(a)} x^{-1-a} \exp(-b/x) \tag{6}$$

where x, a, and b are all positive. The densities comprising the product are typically not in terms of the original variance components in $\boldsymbol{\theta}$, but rather in terms of the expected mean squares from a traditional ANOVA table. The expected mean squares are linear combinations of the variance components with coefficients usually corresponding to sample sizes. For example, in a balanced randomized block experiment with k observations in each block (see Section 4.1), the product densities are based on σ_e^2 and $\sigma_e^2 + k\sigma_b^2$ where σ_e^2 and σ_b^2 are the residual and block variance components, respectively.

We use these products of approximating inverted-gamma densities to define a base density $g(\boldsymbol{\theta}|\mathbf{y})$. The algorithm has two pieces: first, we define a linear transformation $\boldsymbol{\theta} \to \tau$ that produces the linear combination of variance components in which the inverted-gamma approximations apply, then we fit the parameters a and b in (6).

The particular linear combination of the variance components we need (to construct the expected mean squares for balanced data) can be computed by considering various moment equations traditionally used in the analysis of mixed models. We use the Minimum Variance Quadratic Unbiased Equations (MIVQUE(0)); refer to Rao (1972), LaMotte (1973), and Wolfinger, Tobias, and Sall (1994). The desired transformation can be obtained by simply rescaling each row of the coefficient matrix of these equations by the corresponding element in the row pertaining to the residual variance. Examples of this process are provided in the next section.

Let τ_i denote the ith transformed parameter. We assign a separate inverted-gamma density for each τ_i and take the product of them as our proposal density. To choose values of a and b for each inverted gamma density, we evaluate the logarithm of the posterior density over a grid of τ_i values, holding the other elements fixed at the values determined by the REML estimate of $\boldsymbol{\theta}$. The log-inverted gamma density becomes

$$\log IG(\tau_i|\mathbf{y}) \approx c_i + (-1 - a_i)\log \tau_i - b_i/\tau_i,$$

which is a linear regression model in a_i, b_i with an intercept constant c_i. Thus, a least squares fit of the log posterior values produces estimates $\hat{a}_i$ and $\hat{b}_i$ corresponding to τ_i. To define our grid of τ_i values we multiply and divide the REML value of τ_i by 2 to obtain upper and lower bounds for the grid, and then evaluate 20 equally spaced points between the two. This is somewhat arbitrary but works well in practice.

The final joint base density for τ can be written as

$$g(\tau|\mathbf{y}) \propto \prod_{i=1}^{q} IG(\tau_i|\hat{a}_i, \hat{b}_i)$$

where q is the number of elements in τ. The inverse of the previous linear transformation then yields $g(\boldsymbol{\theta}|\mathbf{y})$.

In unpublished research we investigated the possibility of using the preceding base density in a rejection sampling algorithm; however, we found it difficult to reliably estimate the bounding constant required for this technique.

One caveat of the independence chain algorithm is that it can generate candidate variance components which are negative. This typically happens when one or more of the inverted gamma densities have small values for a; that is, when the data provide little information regarding one or more elements of τ. This difficulty can be overcome by simply rejecting the violating points before evaluating p and g, and thus they usually do not significantly impact computing time.

4 Examples

We now apply the independence chain algorithm to three common examples. The chains were computed using Version 7 of the SAS Mixed procedure (1999), and computation times provided for each example are from an HP 9000/720 workstation running HP-UX 10 with 32 mb and 100 mhz. Proc Mixed code for the examples can be obtained at *http://www.sas.com/techsup/download/stat/* in the file *bayesvc.sas*.

4.1 Randomized Block Design

For this example we consider the generated data from Box and Tiao (1973, Table 5.1.4). This data set contains 30 observations divided into 6 batches containing 5 observations each. Such an arrangement is commonly known as a balanced randomized block design, with batches being the block effect. The mixed model is

$$y_{ij} = \beta_0 + \gamma_i + \epsilon_{ij}$$

where y_{ij} is the jth observation from the ith batch ($i = 1, \ldots, 6$), β_0 is a fixed intercept, and γ_i is the random effect parameter for the ith batch. We assume the γ_i are i.i.d. Gaussian with mean 0 and variance σ_b^2 and are independent of the ϵ_{ij}, which are i.i.d. Gaussian with mean 0 and variance σ_e^2. The mixed model variance matrices are therefore $\mathbf{G} = \sigma_b^2 \mathbf{I}_6$ and $\mathbf{R} = \sigma_e^2 \mathbf{I}_{30}$.

The REML estimate of σ_b^2 for these data equals zero, assuming a nonnegative parameter space. A frequentist analysis would therefore tend to proceed as if the γ_i were dropped from the model. However, from a Bayesian perspective, the data do contain information about the γ_i and σ_b^2. Before looking at results, we consider the details of the independence chain algorithm of Section 3 for this example.

The MIVQUE(0) equations for this example are as follows:

$$\begin{bmatrix} 125 & 25 \\ 25 & 29 \end{bmatrix} \begin{bmatrix} \sigma_b^2 \\ \sigma_e^2 \end{bmatrix} = \begin{bmatrix} 208.41 \\ 400.38 \end{bmatrix}$$

The row for σ_b^2 consists of [1255], and this is rescaled by the σ_b^2 element of the last row ([25]), yielding [51] for the transformation. This produces $\tau_1 = 5\sigma_b^2 + \sigma_e^2$, and the identity transformation is used for σ_e^2, yielding $tau_2 = \sigma_e^2$.

We consider two different prior specifications. The first is Jeffreys reference prior discussed in Section 2.3. For this example it is improper and proportional to $(\sigma_e^2)^{-1}(5\sigma_b^2 + \sigma_e^2)^{-1} = (\tau_1\tau_2)^{-1}$. The second prior is informative and is constructed so that the marginal modes for σ_b^2 and σ_e^2 are approximately 5 and 10 respectively, and their upper 95th percentiles are approximately 10 and 20 respectively. For simplicity, this informative prior is expressed as a product of inverted gamma densities for τ_1 and τ_2. Specifically, $\tau_1 \sim IG(10.31, 395.9)$ and $\tau_2 \sim IG(10.31, 113.1)$.

Evaluating the log posterior along a grid for each of the two τ parameters and performing the linear regression described in Section 3.1 produces the inverted gamma densities $IG(2.5, 20.84)$ and $IG(12, 179.4)$ for Jeffreys prior and $IG(12.81, 416.8)$ and $IG(22.31, 292.5)$ for the informative prior. Because of the balanced nature of this design, the products of the pairs of inverted gammas is almost exactly proportional to the posteriors for both priors, and so no replications occur in the independence chain because of rejections.

However, for Jeffreys prior a large percentage of the generated candidates fall outside of nonnegative parameter space assumed to hold for (σ_b^2, σ_e^2). In particular, approximately 73% of the candidates generated were disqualified immediately because they fell outside the parameter space. These rejections do not slow the computations much because there is no need to evaluate the posterior for them. In contrast, only 0.4% of the samples are rejected when using the informative prior.

Samples of size 10,000 are generated using both priors. These samples contain observations from the joint posterior distributions of σ_b^2, σ_e^2, β_0, and the γ_i. 10,000 is chosen so that a posterior mean is estimated with standard deviation 1/100th that of the parameter it is estimating, providing roughly 2 decimal places of accuracy. Total computing time to compute the REML fit and 10,000 samples is 20 seconds for both priors.

Figure 1 displays kernel estimates of the marginal posterior densities for σ_b^2 and σ_e^2, along with associated marginal priors and likelihoods. Also plotted are the true posterior densities computed by numerical quadrature. All densities shown in Figure 1 integrate to 1 except for the Jeffreys priors in panels (a) and (b). These are computed as $5\sigma_e^{-2}(5\sigma_b^2 + \sigma_e^2)^{-1}$ to provide a scale comparable to the other densities. The true posteriors in panels (a) and (b) match the ones derived by Box and Tiao (1973, Figures 5.2.7 and 5.2.4).

Figure 1 goes about here.

The estimated posteriors in Figure 1, computed by kernel density estimation on the 10,000 samples, closely approximate the true ones. One slight difference is in panel (a), in which the exact marginal posterior for σ_b^2 has its mode at zero, whereas the estimated mode is positive. This is an artifact of the particular kernel density estimation algorithm used to construct the plot. In spite of this, the plot clearly reveals that there is quite a bit of uncertainty in the data about σ_b^2 and therefore setting it equal to its REML estimate of 0 could lead to biased inferences.

Panels (c) and (d) of Figure 1 reveal the effect of the informative prior. It results in a positive mode for σ_b^2 and in a smaller posterior variance for σ_e^2 than under Jeffreys prior.

Although not shown here, it is also easy to compute numerous posterior summaries from the samples, including posterior means and quantiles. The latter can provide the Bayesian analog to frequentist confidence intervals. Two-at-a-time analyses can be made as well. For example, a bivariate kernel density estimate could be computed for two variables of interest, and its level curves could be used to construct regions of high posterior probability.

Figure 2 illustrates one the key advantages to a sampling-based Bayesian analysis. Plotted are two nonlinear functions of σ_b^2 and σ_e^2: the ratio σ_b^2/σ_e^2 and the intraclass correlation $\text{ICC} = \sigma_b^2/(\sigma_b^2+\sigma_e^2)$. Once a sample has been generated, it is trivial to find the marginal posterior density for these and any other function of the original parameters. One simply evaluates the function of interest for each of the samples and treats it as a new parameter. For this example, the marginals using Jeffreys prior appear to derive their shape from that of σ_b^2, whereas those from the informative prior are more influenced by σ_e^2.

Figure 2 goes about here.

Finally, Figure 3 depicts posterior density estimates for β_0 and two of the six γ_i using both priors. The intercept β_0 is certainly well away from zero, with an estimated posterior mean and standard deviation equal to 5.66 and 0.97 under Jeffreys prior. The frequentist analogs equal 5.67 and 0.68, the latter making no account of uncertainty in estimating σ_b^2 and σ_e^2. Frequentists make up for this deficiency by using a t_5 distribution to construct tests and confidence limits for β_0. The t_5 also provides a very close approximation to the Bayesian posterior for β_0 under Jeffreys prior.

Figure 3 goes about here.

Under the informative prior, the estimated mean and standard deviation for β_0 are 5.66 and 1.08, respectively, and so the informative prior has actually slightly lessened the certainty about β_0. This is because the informative prior places more mass on larger values of σ_b^2.

This loss of precision is even more evident in panel (b) of Figure 3, which plots the posterior densities of γ_3 and γ_6 under both priors. The posterior means for the γ_i are all close to zero and they have an average posterior standard deviation of 1.2 under Jeffreys prior and 1.4 under the informative prior.

4.2 Split-Plot Design

The data for this example are from Stroup (1989) and are listed Table 1. We consider two separate data sets, the first arranged in a balanced split-plot design with the whole plots forming a randomized complete-block design, and the second an unbalanced subset of the first. The usual mixed model for this design is as follows (cf. Chapter 24 of Milliken and Johnson, 1992):

$$y_{ijk} = \mu + A_i + b_j + w_{ij} + B_k + AB_{ik} + \epsilon_{ijk}$$

Table 1: *Split-plot data*

		Block 1			Block 2			Block 3			Block 4		
		A1	A2	A3	A1	A2	A3	A1	A2	A3	A1	A2	A3
Balanced	B1	56	50	39	30	36	33	32	31	15	30	35	17
	B2	41	36	35	25	28	30	24	27	19	25	30	18
Unbalanced	B1	.	50	39	30	.	33	.	31	.	30	.	.
	B2	.	36	35	25	28	30	24	27	19	25	.	18

Here y_{ijk} is the observation from the jth block having the ith level of the whole-plot treatment (A_i) and the kth level of the subplot treatment (B_k). The parameter μ models an overall mean and AB_{ik} models an interaction between the two treatments.

The random effects parameters are b_j (which models a random block effect), w_{ij} (which models the whole plot error), and ϵ_{ijk} (the residual error). These terms are assumed to be independent Gaussian random variables with means 0 and variances σ_b^2, σ_w^2, and σ_e^2, respectively.

In terms of our general mixed-model notation from Section 2, $\boldsymbol{\beta}$ consists of the fixed-effects parameters μ, A_i, B_k, and AB_{ik}, whereas γ consists of the random effects parameters b_j and w_{ij}. $\mathbf{G} = Var[\gamma]$ is diagonal and has elements equal to either σ_b^2 or σ_w^2, and $\mathbf{R} = \sigma_e^2\mathbf{I}$. We use Jeffreys reference prior for the variance components, which is proportional to $(6\sigma_b^2 + 2\sigma_w^2 + \sigma_e^2)^{-1}(2\sigma_w^2 + \sigma_e^2)^{-1}(\sigma_e^2)^{-1}$ for the balanced case.

The proposal densities for both data sets are described in Table 2. As in the previous example, the algorithm simulates exactly from the posterior for the balanced data set, and this is suggested by the round numbers obtained for the transformation and $\hat{a}_i$ for these data. However, this is not the case for the unbalanced data, and the transformation coefficients and $\hat{a}_i$ are all smaller and fractional.

Table 2: *Transformation coefficients and inverted gamma density parameters used in the rejection sampling algorithms for both split-plot data sets*

Data	Parameter	σ_b^2	σ_w^2	σ_e^2	a	b
Balanced	τ_1	6	2	1	1.5	621.8
	τ_2	0	2	1	3.0	120.4
	τ_3	0	0	1	4.5	42.13
Unbalanced	τ_1	3.52	1.6	1	1.43	207.9
	τ_2	0	1.6	1	1.67	25.95
	τ_3	0	0	1	1.48	12.41

For the balanced data set, the algorithm producing 10,000 samples rejected 3% of the proposals due to violation of the lower boundary constraint of zero imposed on the three variance components. This percentage increased to 35% for the unbalanced data set. In addition, 5% of the feasible proposals were rejected due to the independence chain, creating a small amount of replication in the sample. The algorithm required approximately 15 seconds to generate the samples for each data set.

Figure 4 displays kernel density estimates of the posterior densities of the three variance components for each data sets. This figure reveals that even though there are fewer observations in the unbalanced data, they actually convey more precise information about the three variance components. This is apparently due to the fact that some of the deleted observations in the unbalanced data set are some of the most extreme ones in the balanced data set. Both data sets provide the least amount of information about σ_b^2 and the most about σ_e^2.

Figure 4 goes about here.

Some of the summary statistics from the sample are provided in Table 3 along with the REML estimates of the variance components. Jeffreys prior does exert some influence in the posteriors, as evidenced by the amount the estimated modes of the posterior densities are less than the REML

Table 3: *REML estimates and summary statistics from the posterior sample of the variance components for both split-plot data sets*

Data	Parameter	REML	Mean	Mode	L95	U95
Balanced	σ_b^2	62.4	201.0	38.5	12.5	961.0
	σ_w^2	15.4	23.8	11.5	2.4	86.6
	σ_e^2	9.4	11.5	8.1	4.4	28.1
Unbalanced	σ_b^2	37.2	113.0	21.0	3.8	560.0
	σ_w^2	4.0	15.2	3.4	0.4	66.9
	σ_e^2	8.7	9.7	5.1	2.4	30.3

Mode is the mode of the kernel density estimate, L95 is the 2.5th percentile and U95 is the 97.5th percentile of the posterior sample

Table 4: *F-tests for the two split-plot data sets*

Data	Effect	NDF	DDF	F	p
Balanced	A	2	6	4.07	0.0764
	B	1	9	19.39	0.0017
	AB	2	9	4.02	0.0566
Unbalanced	A	2	4	2.40	0.2063
	B	1	3	16.31	0.0273
	AB	2	3	0.79	0.5298

NDF is the numerator degrees of freedom and DDF is the denominator degrees of freedom for the F-test

estimates. Interval estimates for the variance components are trivially obtained by computing percentiles of the sample. This may be contrasted with the complicated calculations required for frequentist confidence intervals on variance components, as in Burdick and Graybill (1992).

Moving to consideration of the fixed effects A and B, Table 4 displays the classical frequentist F-tests for both data sets. Note in particular how the test for AB changes from marginally significant in the balanced data to completely nonsignificant in the unbalanced data.

To provide a Bayesian analysis of the AB effect, we consider the posterior summary measure

$$\rho_{AB} = \frac{(\mathbf{L}\boldsymbol{\beta})^T(\mathbf{L}\mathbf{L}^T)^{-1}(\mathbf{L}\boldsymbol{\beta})}{rank\,(\mathbf{L})}$$

where $\mathbf{L}$ is the standard contrast matrix used to compute F-statistic for AB:

$$\mathbf{L} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & -1 & 1 \end{bmatrix}$$

The use of this measure is proposed by Ghosh (1992) as a modification of a similar quantity proposed by Schervish (1992). Unlike the classical F-statistic, ρ_{AB} is able to "separate the effect size from the sample size". Furthermore, the quantity $\sqrt{2\rho_{AB}}$ can be interpreted as the weighted average squared distance between the interaction contrasts $(AB_{11} - AB_{12} - AB_{31} + AB_{32})$ and $(AB_{21} - AB_{22} - AB_{31} + AB_{32})$. Wolfinger and Rosner (1996) use this interpretation to draw inferences about a medical experiment.

Figure 5 plots kernel density estimates of the posterior density of $\sqrt{2\rho_{AB}}$ for both data sets. The posterior densities show that the AB effect has approximately a 50% chance of having an average deviation less than 5 for the balanced data and less than 3 for the unbalanced data. Such statements are often very meaningful to researchers and can be much more informative than standard p-values.

Figure 5 goes about here.

4.3 Quadratic Trend with a Random Intercept

The data for this example are listed in Table 5 and are an unbalanced subset of the insulin data presented in Appendix A5 of Lindsey (1993). Repeated measurements are taken on the blood sugar levels of 22 rabbits after injection with one of two mixtures of insulin, the standard (treatment A) or the one containing 5% less protamine (treatment B). The first 11 rabbits in Table 5 received treatment A and the second 11 treatment B. Measurements are taken at 0.0, 1.5, 3, 4.5, and 6.0 hours after injection, and some of the data are assumed to be missing at random.

Table 5: *Data for quadratic example*

	Hour after Injection				
Rabbit	0.0	1.5	3.0	4.5	6.0
1	77	52	35	56	64
2	77	56	43	56	64
3	77	43	.	39	64
4	81	56	22	35	39
5	90	73	52	60	64
6	103	47	22	60	99
7	99	.	.	47	81
8	90	35	22	26	.
9	90	64	47	.	.
10	90	35	12	26	68
11	85	68	56	68	73
12	103	26	22	39	68
13	85	68	56	64	81
14	85	35	.	64	.
15	85	35	47	77	.
16	85	35	.	.	77
17	103	56	47	64	.
18	90	52	.	35	39
19	81	22	.	.	77
20	85	30	22	52	94
21	85	47	33	77	.
22	94	39	35	52	85

Following Lindsey (1993), we fit a quadratic trend to the data, although we use the baseline measurement as a covariate. The mixed model we use is as follows:

$$y_{ijk} = \mu + dx_{ij} + c_i + b_i h_k + a_i h_k^2 + \gamma_{ij} + \epsilon_{ijk}$$

Here y_{ijk} represents the kth blood sugar measurement on the jth rabbit receiving the ith treatment. The parameter μ is an overall mean and d is the coefficient for the baseline measurement x_{ij}. The hour of measurement is denoted h_k, and c_i, b_i, and a_i are coefficients for the intercept, slope, and quadratic terms of h_k, respectively. The model therefore fits different quadratic curves for the two treatments. The parameter γ_{ij} is a random intercept allowing the quadratic curves to be separated uniformly by a random amount, and ϵ_{ijk} are standard residual errors. We assume that γ_{ij} and ϵ_{ijk}

are independent Gaussian random variables with means 0 and variances σ_g^2 and σ_e^2, respectively. Jeffreys prior is assumed for the variance components.

The set up for the independence chain algorithm begins with the transformation $\tau_1 = 3.41\sigma_g^2 + \sigma_e^2$ and $\tau_2 = \sigma_e^2$, obtained from the MIVQUE(0) coefficient matrix. The parameter search using τ_1 and τ_2 produces the base densities $IG(9.23, 3998)$ and $IG(23.47, 3428)$ respectively. A sample of size 10000 was generated in 1.5 minutes with an acceptance rate of 99%.

The posterior densities of σ_g^2 and σ_e^2 are similar in shape to those from previous examples and are not shown. Approximate 95% intervals obtained from the 2.5th and 97.5th percentiles of the sample are $(25.2, 236)$ and $(101, 229)$, respectively. The posterior density estimate of d has a mean of 0.05 and a standard deviation of 0.33, revealing that the baseline measurement has little association with subsequent measurements.

Plotted in Figure 6 are several percentiles of the predicted trends for the two treatments. The percentile curves are created by evaluating the trend at 100 equally spaced points from hours 0 to 6 for all 1000 samples and computing percentiles at each point based on the 1000 values. A baseline measurement of 85 is used in computing the values. Interpreting Figure 6, we see that the trend for treatment A has more curvature than that for treatment B, but similar precision. Panel (b) shows that it can be risky to extrapolate beyond the data, as the estimated trend at hour 0 is well below the observed baseline measurements.

Figure 6 goes about here.

A further comparison of the two treatments is provided in Figure 7. Panel (a) plots kernel density estimates of the posterior densities of the hour at which the estimated trend is minimized for the two treatments. Parametrically this value equals $-b_i/(2a_i)$, a nonlinear function which can be difficult to analyze from a frequentist perspective. Under our Bayesian sampling scheme, it is very easy to analyze by just forming a new observation from each existing one in the sample of 1000. Treatment B appears to "bottom out" approximately 1 hour before treatment A, although the data are much less informative about this bottom value for treatment B than they are for treatment A.

Figure 7 goes about here.

Panel (b) of Figure 7 displays the posterior density of a more complicated nonlinear function of the fixed-effects parameters. Plotted are kernel estimates of the posterior densities of the hour at which the estimated trend achieves a normative value of 85 for the two treatments. Although this is a simple quantity conceptually, analytically it is the solution of the quadratic equation

$$T85_i = \{-b_i + [b_i^2 - 4a_i(c_i + d85 - 85)]^{1/2}\}/(2a_i)$$

where again a baseline measurement of 85 is assumed. The problem is essentially one of inverse regression, and, as discussed in Section 3.2 of Lange, Carlin, and Gelfand (1992), a sampling-based Bayesian analysis provides a nice solution. For this example the T85 values appear to differ by approximately half an hour.

5 Discussion

To generate a sample from the joint posterior distribution on the left-hand side of (1) we have proposed the following algorithm:

1. Transform the variance components (the components of $\boldsymbol{\theta}$ in (1)) to approximate independence using the ratios established in the MIVQUE(0) coefficient matrix.

2. Determine inverted-gamma coefficients for each transformed parameter by performing a linear regression on a series of grid evaluations of the log marginal posterior of the variance components, $p(\boldsymbol{\theta}|y)$. This posterior is computed by taking the product of the restricted likelihood and some prior.

3. Form the base sampling density for the transformed variance components, $g(\boldsymbol{\theta}|y)$, by taking the product of the inverted gamma densities from step 2.

4. Use an independence chain to generate a psuedo-random sample from $p(\boldsymbol{\theta}|y)$, which is the second factor on the right-hand side of (1).

5. Given sampled values for the variance components, use the multivariate normal distribution to generate observations for the mean parameters. This is the first factor on the right-hand side of (1).

If the version of Jeffreys rule given in Section 2.3 is used, the procedure may be fully automated. For balanced data, the algorithm produces exact draws from the posterior. For other cases the independence chain will produce occasional replicated values of variance components to account for inaccuracies in the proposal density. For the problems we have considered these replications represent a very small fraction of the total sample, but in general they can provide an indication of gross inaccuracies in the sampling process should they happen to arise.

This simulation-based Bayesian approach to variance component models may be contrasted with traditional Bayesian analyses found in Hill (1965), Zellner (1971), Box and Tiao (1973), Broemeling (1985), Chaloner (1987), Hulting and Harville (1991), Harville and Carriquiry (1992), and Searle, Casella, and McCulloch (1992). These involve analytical expressions for the joint and marginal posterior distributions of the mean parameters and variance components. Even for balanced data the formulas become rather cumbersome, and different approximations are often required for several useful functions of the parameters.

The independence chain approach we have adopted may be viewed as a particular MCMC alternative to Gibbs sampling, as in Wang *et al.* (1994), (cf. Gelfand *et al.*, 1990, or Gilks, Richardson, Spiegelhalter, 1996). While the inferential results from both approaches should be the same (aside from simulation error), several practical differences are worth noting. First, the one-at-a-time nature of Gibbs sampling can make simulation difficult in unbalanced, non-conjugate cases; resort must often be made to rejection sampling or ratio-of-uniform methods for some of the full conditionals (Carlin and Polson, 1991; Gilks and Wild, 1992; Wakefield *et al.* 1994). Our independence chain method samples the variance components as a block, using a proposal density that can be sampled from directly and easily. Second, in contrast to Gibbs sampling, the independence chain uses the same proposal density throughout the entire algorithm and thus requires no burn-in. Finally, poorly centered parameterizations for the random effects can sometimes cause slow convergence in Gibbs chains (Gelfand, Sahu, and Carlin, 1995). The draws for the random effects in our indepence chain are conditioned directly on the variance component draws, and are not even a necessary part of the algorithm. Furthermore, an excessive amount of duplications of the variance components in the sample can be used as a diagnostic for a poor proposal density or possibly an improper posterior (see Hobert and Casella, 1996, for discussion on the latter). Common concerns about ineffectiveness of independence chains (e.g., pp. 54-55 of Roberts, 1996) are not worrisome here because the proposal density provides a good approximation to the posterior (and is exact for balanced data).

Our work can be considered as an extension of the direct simulation approach of Vountatsou and Smith (1997), who thoroughly investigate the two-variance component case and use several different priors. Sun, Hsu, Guttman, and Leonard (1996) also investigate variance component models and propose Laplacian T approximations and importance sampling algorithms using the logs of the variance components. While the simulation algorithm we propose is quite different from theirs and, we believe, more amenable to automation, we would expect it to produce similar results to those presented by these authors.

In the examples presented here we have tried to show that the method provides a flexible tool for inference while being sufficiently efficient for practical implementation. Although the models we consider in this paper assume the random effects covariance matrix $\mathbf{G}$ to be diagonal, we hope to be able to communicate additional work that will remove this restriction. Two very important classes of models involving nondiagonal $\mathbf{G}$ matrices are those involving random coefficients (Draper, 1995) and those used in animal breeding (Templeman, 1997). We would also like to note that the techniques proposed in this paper appear to be useful in constructing Bayes factors for testing variance components as discussed by Pauler, Wakefield, and Kass (1998). Fully Bayesian developments in these areas promise to be very useful for both researchers and analysts alike.

References

Berger, J.O. and Bernardo, J.M. (1992). On the development of reference priors. *Bayesian Statistics 4* (J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, eds.) Oxford: University Press, 419-434, (with discussion).

Box, G.E.P., and Tiao, G.C. (1973). *Bayesian Inference in Statistical Analysis,* John Wiley and Sons, New York.

Broemeling, L.D. (1985). *Bayesian Analysis of Linear Models,* Marcel Dekker, New York, New York.

Burdick, R.K. and Graybill, F.A. (1992). *Confidence Intervals on Variance Components.* Marcel Dekker, New York.

Carlin, B.P. and Polson, N.G. (1991).Inference for nonconjugate Bayesian models using the Gibbs sampler. *Canadian Journal of Statistics* **19,** 399-405.

Chaloner, K. (1987). A Bayesian approach to the estimation of variance components for the unbalanced one-way random model. *Technometrics* **29,** 323-337.

Corbeil, R.R., and Searle, S.R. (1976). Restricted maximum likelihood (REML) estimation of variance components in the mixed model. *Technometrics* **18,** 31-38.

Draper, D. (1995). Inference and hierarchical modeling in the social sciences. **Journal of Educational and Behavioral Statistics 20,** 115-147.

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.

Gelfand, A.E., Sahu, S.K., and Carlin, B.P. (1995). Efficient parameterisations for normal linear mixed models. *Biometrika,* **82,** 479-488.

Ghosh, M. (1992). Discussion of M.J. Schervish, Bayesian analysis of linear models. *Bayesian Statistics 4* (J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, eds.) Oxford: University Press, 432-433.

Gilks, W.R., Richardson, S., and Spiegelhalter, D.J., eds., (1996). *Markov Chain Monte Carlo in Practice*, London: Chapman and Hall.

Gilks, W.R., and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics* **41,** 337-348.

Harville, D.A. (1974). Bayesian inference for variance components using only error contrasts. *Biometrika* **61,** 383-5.

Harville, D.A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association* **72,** 320-338.

Harville, D.A. and Carriquiry, A.L. (1992). Classical and Bayesian prediction as applied to an unbalanced mixed linear model. *Biometrics* **48,** 987-1003.

Henderson, C.R. (1984). *Applications of Linear Models in Animal Breeding.* University of Guelph.

Hobert, J.P., and Casella, G. (1996). The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *Journal of the American Statistical Association* **91,** 1461-1473.

Hulting, F.L. and Harville, D.A. (1991). Some Bayesian and non-Bayesian procedures for the analysis of comparative experiments and for small-area estimation: Computational aspects, frequentist properties, and relationships. *Journal of the American Statistical Association* **86,** 557-568.

Jeffreys, H. (1961). *Theory of Probability,* Clarendon Press: Oxford.

Kass, R.E. and Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association,* **91,** 1343-1370.

LaMotte, L.R. (1973). Quadratic estimation of variance components. *Biometrics* **29,** 311-330.

Lange, N., Carlin, B.P., and Gelfand, A.E. (1992). Hierarchical Bayes models for the progression of HIV infection using longitudinal CD4 T-cell numbers (with discussion). *Journal of the American*

Statistical Association **87,** 615-632.

Lindsey, J.K. (1993). *Models for Repeated Measurements.* Oxford University Press, New York.

Milliken, G.A. and Johnson, D.E. (1992). *Analysis of Messy Data, Volume 1: Designed Experiments.* New York: Chapman and Hall.

Patterson, H.D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika* **58,** 545-554.

Pauler, D.K., Wakefield, J.C., and Kass, R.E. (1998). Bayes factors for variance component models. in review.

Rao, C.R. (1972). Estimation of variance and covariance components in linear models. *Journal of the American Statistical Association* **67,** 112-115.

Ripley, B.D. (1987). *Stochastic Simulation.* John Wiley and Sons, Inc., New York.

Roberts, G.O. (1996). Markov chain concepts related to sampling algorithms. In Gilks *et al.* (1996).

Robinson, G.K. (1991). That BLUP is a good thing: the estimation of random effects (with discussion). *Statistical Science* **6,** 15-51.

SAS Institute Inc. (1999). SAS/STAT User's Guide, Version 7. SAS Institute Inc., Cary, NC.

Schervish, M.J. (1992). Bayesian analysis of linear models. *Bayesian Statistics 4* (J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, eds.) Oxford: University Press, 432-433.

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

Smith, A.F.M. and Gelfand, A.E. (1992). Bayesian statistics without tears: a sampling-resampling perspective. *American Statistician* **46,** 84-88.

Stroup, Walter W. (1989). Predictable functions and prediction space in the mixed model procedure. *Applications of Mixed Models in Agriculture and Related Disciplines,* Southern Cooperative Series Bulletin No. 343, Louisiana Agricultural Experiment Station, Baton Rouge, 39-48.

Sun, L., Hsu, J.S.J., Guttman, I., and Leonard, T. (1996). Bayesian methods for variance component models. *Journal of the American Statistical Association* **91,** 743-752.

Templeman, R.J. (1997). Generalized linear mixed models in dairy cattle breeding. *Journal of Dairy Science* **81,** 1428-1444.

Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Annals of Statistics,* **22,** 1701-1762

Vountatsou, P. and Smith, A.F.M. (1997). Simulation-based Bayesian inferences for two-variance components linear models. *Journal of Statistical Planning and Inference* **59,** 139-161.

Wakefield, J.C., Smith, A.F.M., Racine-Poon, A., and Gelfand, A. (1994). Bayesian analysis of linear and nonlinear population models using the Gibbs sampler. *Applied Statistics,* **43,** 201-221.

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 in Iberian pigs. *Genetics Selection Evolution*

26, 91-115.

Wolfinger, R.D. and Rosner, G.L. (1996). Bayesian and frequentist analyses of an *in vivo* experiment in tumor hemodynamics. Chapter to appear in volume edited by D. Berry and D. Stangl, Duke University.

Wolfinger, R.D., Tobias, R.D., and Sall, J. (1994). Computing Gaussian likelihoods and their derivatives for general linear mixed models. *SIAM Journal on Scientific Computing* **15(6),** 1294-1310.

Yang, R. and Chen, M.-H. (1995). Bayesian analysis for random coefficient models using noninformative priors. *Journal of Multivariate Analysis* **55,** 283-311.

Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics.* New York: Wiley.

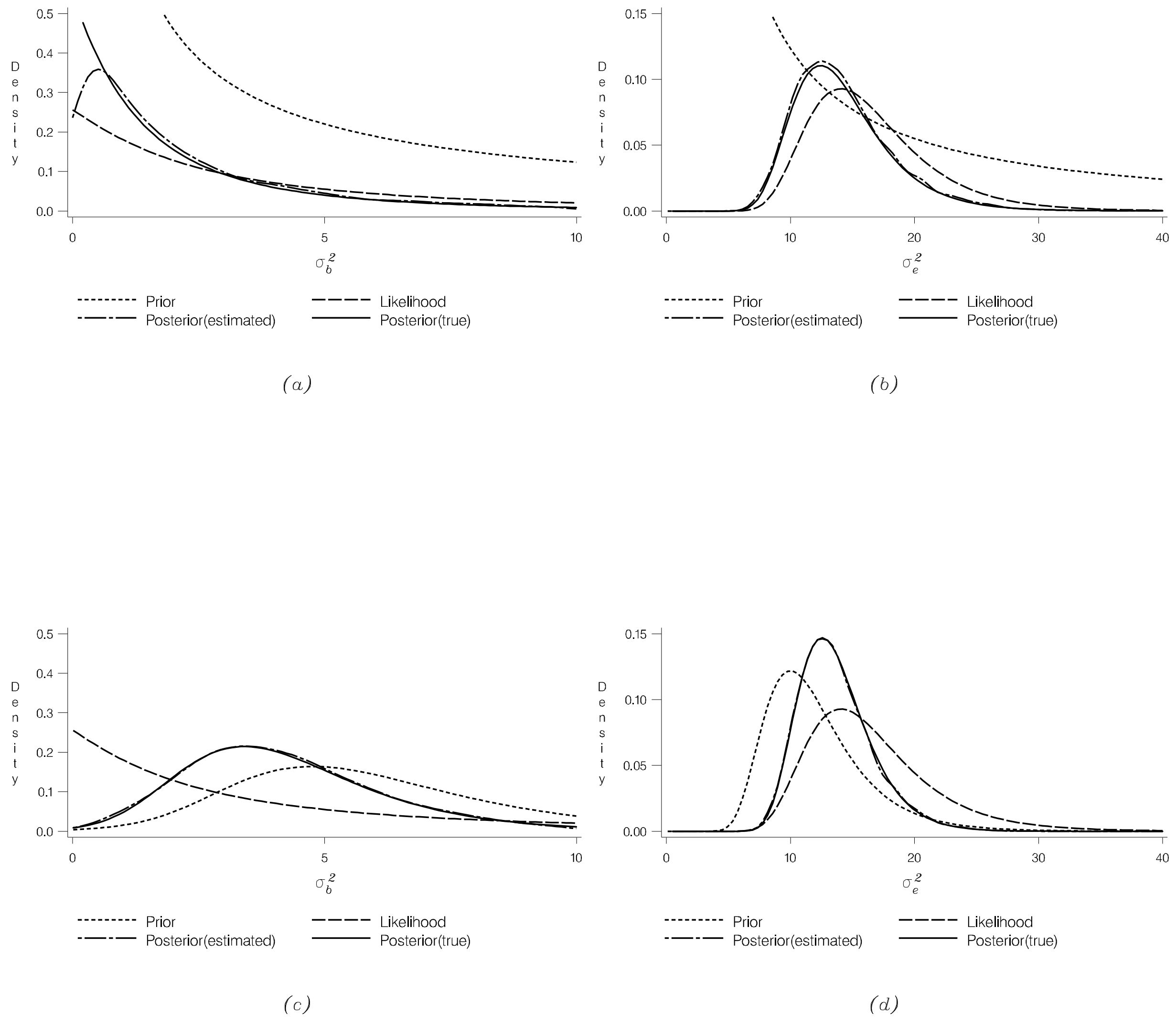

Figure 1. *For the randomized block example, plots of the marginal prior, likelihood, true posterior, and estimated posterior for (a) the block variance component σ_b^2 using Jeffreys's prior, (b) the residual error variance component σ_e^2 using Jeffreys's prior, (c) σ_b^2 using an informative prior, and (d) σ_e^2 using an informative prior*

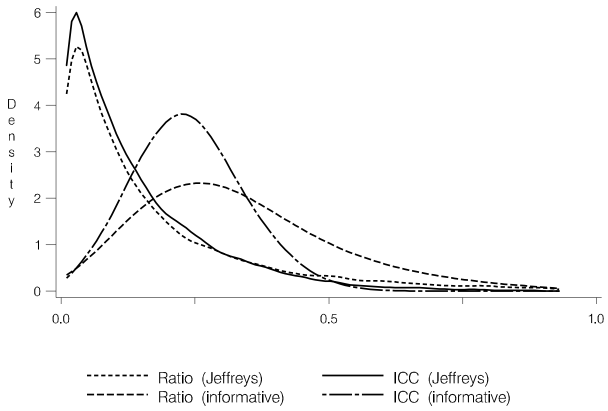

Figure 2. *For the randomized block example, posterior density estimates for the ratio* σ_b^2/σ_e^2 *and the intraclass correlation* $ICC = \sigma_b^2/(\sigma_b^2 + \sigma_e^2)$. *The two plots with the modes near zero are from Jeffreys's prior, the other two are from an informative prior.*

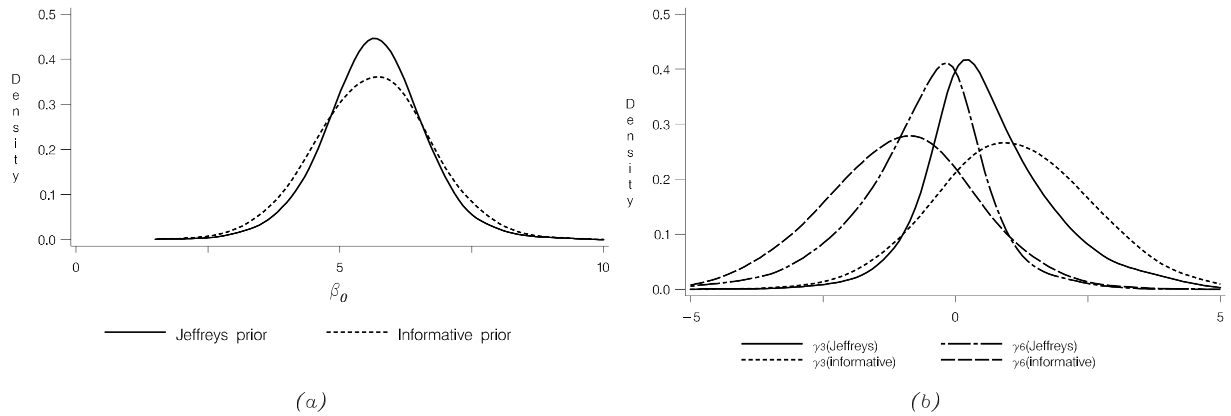

Figure 3. *For the randomized block example, posterior density estimates for (a) the overall intercept β_0 and (b) the random-effects parameters γ_3 and γ_6*

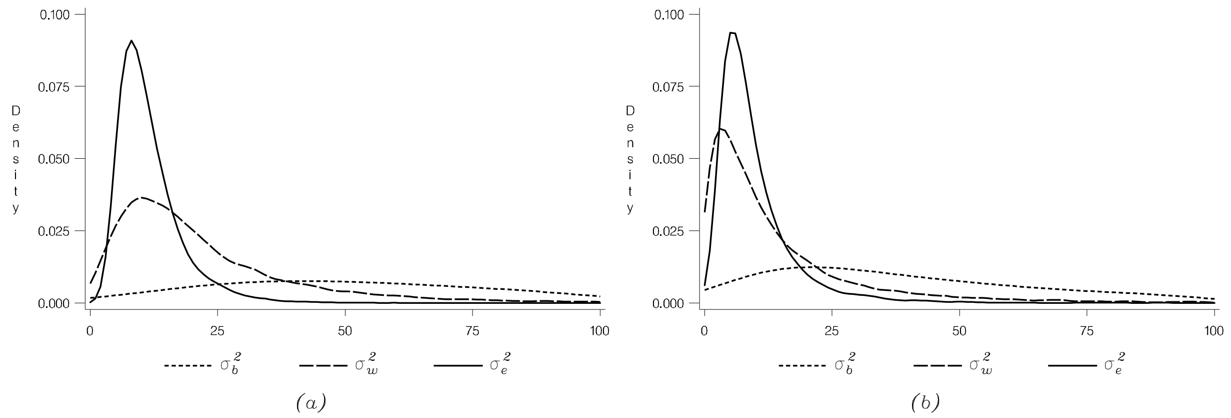

Figure 4. *For the split-plot example, posterior density estimates for the block variance component σ_b^2, the whole-plot variance component σ_w^2, and the residual variance component σ_e^2 for (a) the balanced data and (b) the unbalanced data*

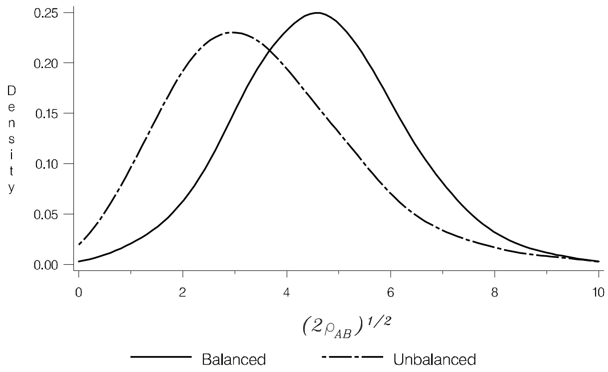

Figure 5. *For the split-plot example, posterior density estimates of* $(2\rho_{AB})^{1/2}$

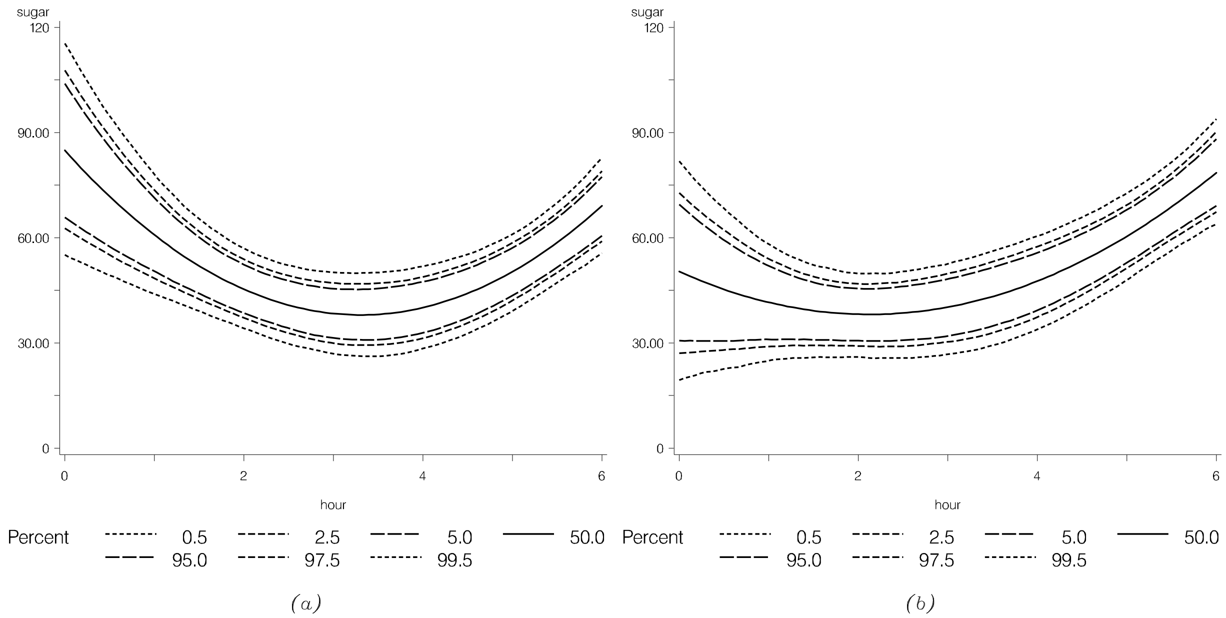

Figure 6. *For the quadratic example, percentile curves of the estimated trend for (a) treatment A and (b) treatment B*

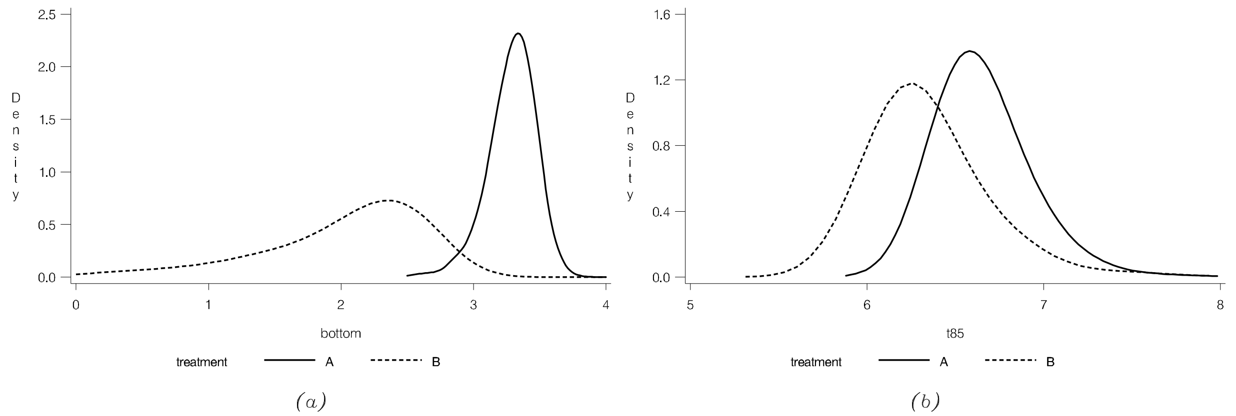

Figure 7. *For the quadratic example, posterior density estimates of (a) the hour at which the quadratic curve is mimimized for treatments A and B and (b) the hour at which the blood sugar level reaches 85 for treatments A and B*