Nonconjugate Bayesian Estimation of Covariance Matrices and Its Use in Hierarchical Models

Michael J. DANIELS and Robert E. KASS

The problem of estimating a covariance matrix in small samples has been considered by several authors following early work by Stein. This problem can be especially important in hierarchical models where the standard errors of fixed and random effects depend on estimation of the covariance matrix of the distribution of the random effects. We propose a set of hierarchical priors (HPs) for the covariance matrix that produce posterior shrinkage toward a specified structure—here we examine shrinkage toward diagonality. We then address the computational difficulties raised by incorporating these priors, and nonconjugate priors in general, into hierarchical models. We apply a combination of approximation, Gibbs sampling (possibly with a Metropolis step), and importance reweighting to fit the models, and compare this hybrid approach to alternative Markov Chain Monte Carlo methods. Our investigation involves three alternative HPs. The first works with the spectral decomposition of the covariance matrix and produces both shrinkage of the eigenvalues toward each other and shrinkage of the rotation matrix toward the identity. The second produces shrinkage of the correlations toward 0, and the third uses a conjugate Wishart distribution to shrink toward diagonality. A simulation study shows that the first two HPs can be very effective in reducing small-sample risk, whereas the conjugate Wishart version sometimes performs very poorly. We evaluate the computational algorithm in the context of a normal nonlinear random-effects model and illustrate the methodology with a logistic random-effects model.

KEY WORDS: Givens angles; Hierarchical prior; Random effects; Shrinkage; Variance matrix.

1. INTRODUCTION

Bayesian hierarchical models typically involve observational units, indexed here by $i = 1, \ldots, k$, that have distributions characterized by multidimensional parameters θ_i that in turn are themselves assumed to be random vectors. Most frequently, the θ_i's are assumed to follow a normal distribution with covariance matrix $\mathbf{D} = V(\theta_i)$. Estimation of $\mathbf{D}$ may or may not be of direct interest, but is always important because of its impact on assessments of uncertainty. The standard Bayesian approaches to estimating $\mathbf{D}$, using either a diffuse Wishart conjugate prior, an invariant prior, or a flat prior, are all very effective for large samples, producing results that agree closely with maximum likelihood (or restricted maximum likelihood). But when the number of units k is small enough so that maximum likelihood is suspect, alternative methods should be considered. This article provides an initial investigation of several Bayesian procedures based on hierarchical priors (HPs) for $\mathbf{D}$ that are supposed to provide greater stability than those based on diffuse priors.

As motivation for our work, we reconsider data originally analyzed by Daniels and Gatsonis (1999) concerning modeling the rates of coronary artery bypass graft (CABG) as functions of hospital-level covariates. Daniels and Gatsonis fit a slightly generalized version of the following hierarchical Poisson regression model to the data:

Stage one: $Y_{ij}|\beta_i = \text{Poisson}(e_{ij}\exp(X_{ij}\beta_i)), i = 1, \ldots, n$ with $i = 1, \ldots, 51$ and $j = 1, \ldots, n_i$, where

$$N = \sum_{i=1}^{51} n_i = 4{,}992.$$

Stage two: $\beta_i|\gamma, \mathbf{D} \sim \text{N}(\gamma, \mathbf{D})$.

Stage three: $\gamma \sim d\gamma, \mathbf{D} \sim \pi(\mathbf{D})$.

In this model Y_{ij} denotes the count of CABG procedures for the jth hospital in the ith state. We chose one of the four regions of the data, the South, in which there are 17 states ($k = 17$), and used three hospital-level covariates: size of the hospital, teaching status of the hospital (0 versus 1), and a comorbidity index, each of which was centered by subtracting its mean. In this situation we have a covariance matrix with 10 parameters ($p = 4$) and only 17 4×1 vectors with which to estimate this matrix. We regard it as plausible that the effects of the three covariates would be independent across states. The HPs that we discuss here let us use this belief to stabilize estimation without elevating that conjecture to an assumption; the method would allow the data to show this independence to be incorrect.

Our study has two parts. First, we ignore the hierarchical modeling context and consider separately the fundamental problem of estimating a covariance matrix based on a multivariate normal sample. The basic idea that we start with is very simple: In estimating $\mathbf{D}$, we will "shrink" it toward a structured form; here we confine our attention to diagonal structure. Historically, in trying to improve estimation of a covariance matrix, the focus has been confined to the eigenvalues (Dey and Srinivisan 1985; Haff 1991; Stein 1975; Yang and Berger 1994), because for small samples, the largest sample eigenvalues will be biased upward and the smallest sample eigenvalue will be biased downward. Covariance estimates using various methods of shrinking the eigenvalues (toward a common value) have been shown to have lower risk than the sample variance (Dey and Srini-

Michael J. Daniels is Assistant Professor, Department of Statistics, Iowa State University, Ames, IA 50011 (E-mail: *mdaniels@iastate.edu*). Robert E. Kass is Professor and Head, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213 (E-mail: *kass@stat.cmu.edu*). The authors are grateful for a useful remark from Jim Berger. They would also like to thank the editor, an associate editor, and a referee for their helpful comments. This work was supported by National Science Foundation grants DMS-9303557, DMS-9631248, and DMS-9803433, and National Institutes of Health grant RO1-CA54852.

vasan 1985; Haff 1991; Lin and Perlman 1985; Yang and Berger 1992). Our approach not only shrinks the eigenvalues, but also shrinks the elements of the orthogonal matrix in the spectral decomposition of $\mathbf{D}$, or the off-diagonal elements in the correlation matrix $\mathbf{R}$ based on $\mathbf{D}$.

To accomplish this, we place on $\mathbf{D}$ a HP comprising $\pi_{\mathbf{D}}(\mathbf{D}|\psi)$ and $\pi_{\psi}(\psi)$ such that the first-stage prior $\pi_{\mathbf{D}}(\mathbf{D}|\psi)$ is centered at a diagonal matrix and has a single dispersion parameter. The components of ψ are the diagonal elements and the dispersion parameter. Thus we compute the posterior on $\mathbf{D}$ via the posterior on ψ; if $\mathbf{D}$ is a $p \times p$ matrix, then we estimate the $p(p+1)/2$ covariance parameters hierarchically via estimation of the $p+1$ hyperparameters. When we focus on the rotation matrix, we place a flat prior on the eigenvalues and thereby end up shrinking the eigenvalues as well. In the context of hierarchical regression models, after we center the regression variables, it will often be plausible that the covariance matrix is not too far from diagonal. Our method might be viewed as a descendent of the approach of Leonard and Hsu (1991), but in fact we were unable to see within their framework any way to define an appealing HP of the type we needed. In Section 2 we describe the priors we examine, and in Section 3 we report results from a simulation study of Bayes risk for the various resulting estimators.

The second aspect of our work involves implementation in hierarchical models; that is, computational issues. Once we introduce HPs on $\mathbf{D}$, we lose the conjugate structure that typically makes Gibbs sampling so attractive. We thus consider the more general problem of devising a posterior sampling scheme for an arbitrary nonconjugate prior on $\mathbf{D}$ when it itself occurs as a hyperparameter in a two-stage hierarchical model. Specifically, the models that we consider assume

$$Y_{ij}|\beta_i \sim f(y_{ij}|\beta_i)$$

with $i = 1, \ldots, k, j = 1, \ldots, n_i$, and $N = \sum_{i=1}^{k} n_i$, and where $f(\cdot|\beta_i)$ is some family of densities, and then

$$\beta_i \sim \mathrm{N}(\gamma, \mathbf{D}), \tag{1}$$

$$\gamma \sim d\gamma, \tag{2}$$

and

$$\mathbf{D} \sim \pi_{\mathbf{D}}(\mathbf{D}), \tag{3}$$

with $\mathbf{D}$ a $p \times p$ matrix. In Section 4 we show that a combination of asymptotic approximations, importance sampling, and possibly Markov chain Monte Carlo (MCMC) is effective for many of these situations. (Our approach is similar to that of Sun, Hsu, Guttman, and Leonard 1996, who treated the special case in which the first stage is normal and the covariance matrix is diagonal.)

In Section 5 we provide two examples, the heavily-reanalyzed guinea pig data analyzed with a simple hierarchical nonlinear regression model, and the hospital CABG utilization example introduced earlier. We draw conclusions in Section 6.

2. PRIORS FOR D

Here we list the three classes of priors that we investigate in the context of estimating $\mathbf{D}$ from data $Y \sim \mathrm{N}(0, \mathbf{D})$, then briefly discuss some issues involved in choosing among them. (For a discussion of other reference priors, see Kass and Wasserman 1996.)

2.1 Conjugate Prior

The conjugate prior is the inverse Wishart (Schervish 1995); that is, the conjugate prior for $\mathbf{D}^{-1}$ is Wishart. But this prior lacks flexibility, allowing only one precision parameter for all $p(p+1)/2$ elements, and requires specification of a mean matrix. The Wishart prior with few degrees of freedom and some fixed scale matrix is commonly used as a reference (noninformative) proper prior. The Wishart prior must have more than $p-1$ degrees of freedom for the prior to be proper; thus setting the degrees of freedom equal to p is a common choice. The scale matrix is sometimes chosen as the maximum likelihood estimator (MLE) of $\mathbf{D}$, though this results in understated precision. Furthermore, in small samples the specification of the scale matrix can be quite influential.

2.2 Nonconjugate Reference Priors

The two commonly used reference priors are Jeffrey's prior ($|\mathbf{D}|^{-(p+1)/2}$, right Haar measure) and a flat prior on $\mathbf{D}$, $\pi_{\mathbf{D}}(d_{11}, d_{12}, \ldots, d_{pp}) = 1$ for all positive definite $\mathbf{D}$, where d_{ij} is the (i,j) component of $\mathbf{D}$. However, care must be taken is using these priors, as they can lead to improper posterior distributions. For example, Jeffrey's prior in the hierarchical logistic model leads to an improper posterior distribution (Hobert and Casella 1996; Natarajan and McCulloch 1995), and the flat prior can be quite informative for small datasets. Two other "noninformative" priors include the Berger–Bernardo prior (Berger and Bernardo 1979; Yang and Berger 1994) and the uniform shrinkage prior (Christiansen and Morris 1997; Daniels 1999; Everson and Morris 1997). The Berger–Bernardo prior shrinks the eigenvalues toward a common value, largely because of a Jacobian factor that appears in our approach as well (see Sec. 2.3.3).

2.3 Hierarchical Priors

We consider a class of hierarchical priors for the covariance matrix based on various parameterizations of the covariance matrix. Each prior has the effect of shrinking some function of the off-diagonal elements to a common value (e.g., 0). This will reduce the estimation of the $p(p-1)/2$ off-diagonal elements to the estimation of a single parameter (which controls the amount of shrinkage). The log matrix prior (Leonard and Hsu 1992), which can be useful for regressing elements of the covariance matrix on covariates, is difficult to interpret in this context. For example, the first diagonal element corresponds to a product of the eigenvectors and the logarithm of the eigenvalues. We do not consider the log matrix prior here, but instead turn to three priors with easily interpretable diagonal and off-diagonal elements.

2.3.1 Wishart Hierarchical Prior. First, we consider a Wishart prior for $\mathbf{D}^{-1}$ with an unknown diagonal scale matrix and unknown degrees of freedom. By considering the degrees of freedom ν to be unknown, we can think of it as a precision parameter with large degrees of freedom supporting diagonality. Specifically, we can place flat priors on the logarithm of the diagonal elements of the Wishart scale matrix, and a flat prior on the logarithm of the degrees of freedom, truncated at a large value and set always greater than $p-1$ so that the Wishart distribution is proper.

2.3.2 Hierarchical Prior on the Correlations. Another prior that we consider is based on the variance/correlation breakdown of the matrix suggested by Barnard, McCulloch, and Meng (1996). Here we put a distribution on the correlations so that they will end up shrinking toward 0. Specifically, we put a normal distribution on Fisher's z transform of the correlations: $.5\log[(1-\rho)/(1+\rho)] \sim \mathrm{N}(0,\tau^2)$. This is similar in spirit to the approach of Lin and Perlman (1985), who used a Stein-type estimator to shrink the correlations toward a common value. Their estimator was shown to have attractive risk properties. In this situation one must be careful to maintain positive definiteness of the matrix, but as Barnard et al. noted, deriving the constraints on the correlations to maintain positive definiteness of the matrix is quite simple. Due to this positive definiteness restriction, the foregoing normal prior distributions on the z-transformed correlations will actually be truncated normal distributions over the relevant ranges of the correlations. To complete the prior specification, we place a prior on the unknown variance, τ^2, and use flat priors on the diagonal elements of $\mathbf{D}$ (or, in nonhierarchical applications in which the data are assumed normal with covariance matrix $\mathbf{D}$, put flat priors on the logarithms of the diagonal elements of $\mathbf{D}$). We use $\pi(\tau^2) \propto (c+\tau^2)^{-2}$. This prior has the form of a "uniform shrinkage" prior and has been discussed recently by various authors, including Christiansen and Morris (1997) and Daniels (1999). By analogy with other models, in which the constant c would represent a variance (see Daniels 1999), a sensible choice of c here would be the asymptotic variance of the z transform of the correlations, $1/(k-3)$.

2.3.3 Hierarchical Prior on the Givens Angles. Finally, we consider a prior that again shrinks the covariance matrix $\mathbf{D}$ toward the identity but is based on the spectral decomposition of $\mathbf{D}$, which we write as $\mathbf{D} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^T$ (so that $\mathbf{\Lambda}$ is a diagonal matrix of ordered eigenvalues, $\lambda_1, \ldots \lambda_p$ and $\mathbf{P}$ is the corresponding orthogonal matrix of eigenvectors). One natural way to parameterize an orthogonal matrix is via Euler angles (Goldstein 1962; Hoffman, Raffenetti, and Ruedenberg 1993). However, after initial exploration we found it slightly simpler instead to parameterize the $(p \times p)$ orthogonal matrix $\mathbf{P}$ in terms of the $p(p-1)/2$ Givens angles (see, e.g., Thisted 1988). Each Givens angle, denoted by θ_{ij} for $i = 1, \ldots, p$ and $j = i+1, p-1$, may be considered a rotation in the plane spanned by i and j components of the basis defining the matrix $\mathbf{P}$. The matrix $\mathbf{P}$ may be written as the product of $p(p-1)/2$ matrices, each one associated with a Givens angle, $\mathbf{P} = \mathbf{G}_{12}\mathbf{G}_{13}\ldots\mathbf{G}_{1p}\ldots\mathbf{G}_{(p-1)p}$, where i and j are distinct and $\mathbf{G}_{ij}$ is the $p \times p$ identity matrix with the ith and jth diagonal elements replaced by $\cos(\theta_{ij})$ and the (i,j) and (j,i) elements replaced by $\pm\sin(\theta_{ij})$. Within the context of the spectral decomposition, if one maintains the order of the eigenvalues, then the Givens angles will be unique with a domain of $(-\pi/2, \pi/2)$. We may then assume that the θ_{ij}'s follow some distribution centered at the origin. Here we have chosen to put a normal distribution on a logit transformation of the angles; we use $\log\{[\pi/(2+\theta)]/[\pi/(2-\theta)]\} \sim \mathrm{N}(0,\tau^2)$. Proceeding analogously to the way in which we worked with the HP on the correlations, we then put a prior on τ^2 and use flat priors on the eigenvalues (or again in nonhierarchical applications we may use flat priors on the logarithms of the eigenvalues), and we again take $\pi(\tau^2) \propto (c+\tau^2)^{-2}$. The choice of constant c here for the prior on τ^2 is more problematic than for the correlation parameterization, because we do not have an expression for the asymptotic variance of the logit of θ. This area needs to be explored further. The logit transformation allows for the prior to encompass the entire real line.

It is important to note that this approach also shrinks the eigenvalues, because a factor of $\prod_{i<j} 1/(\lambda_i - \lambda_j)$ is introduced in transforming the prior from the eigenvalues and angles back to $\mathbf{D}$: $\pi(\mathbf{D}|\tau^2)d\mathbf{D} = \pi(\lambda,\theta|\tau^2)|J|d\lambda d\theta$, where

$$|J| = \frac{\prod_{i=1}^{p-1}\prod_{j=i+1}^{p}\cos(\theta_{ij})^{j-i-1}}{\prod_{i<j}(\lambda_i - \lambda_j)}.$$

Thus the uniform prior on $\mathbf{\Lambda}$ tends to bring the eigenvalues in the posterior closer together than the eigenvalues of the sample covariance matrix (the MLEs) would be.

3. A SIMULATION STUDY OF RISK

3.1 Objective

To study the performance of posteriors based on alternative priors, we considered estimators of the covariance matrix $\mathbf{D}$ when $Y_i \sim \mathrm{N}(0,\mathbf{D}), i = 1,\ldots,n$. We calculated the risk $\mathbf{R}(\hat{\mathbf{D}},\mathbf{D}) = E_{\mathbf{D}}L_1(\hat{\mathbf{D}},\mathbf{D})$ associated with the loss function $L_1(\hat{\mathbf{D}},\mathbf{D}) = \mathrm{tr}(\hat{\mathbf{D}}\mathbf{D}^{-1}) - \log|\hat{\mathbf{D}}\mathbf{D}^{-1}| - p$, where $\hat{\mathbf{D}}$ is the Bayes estimator. This loss function is particularly convenient because the Bayes estimator is simply the inverse of the posterior expectation of $\mathbf{D}^{-1}$. The loss function L_1 also produces the sample variance as the Bayes estimator for Jeffreys's prior. Other loss functions for covariance matrices have been discussed by Yang and Berger (1994).

3.2 Details of Simulation

In our simulation study we computed the Bayes risk (with respect to the loss function in Sec. 3.1) in estimating a five-dimensional covariance matrix for various sample sizes, priors, and true values of the covariance matrix. To perform the computations, we ran the Gibbs sampler 3,000 iterations for each loss evaluation (using single chains with burn-in of 100) and computed 100 losses to obtain each risk (i.e., used

300,000 Gibbs sampler iterations to compute each risk). The sample sizes we used were 10, 20, and 100.

3.2.1 Priors. We considered Jeffreys's prior and the flat prior described in Section 2.2, the three HPs described in Section 2.3, and also the Berger–Bernardo "reference" prior and a fixed-hyperparameter inverse Wishart prior; that is, a Wishart prior on $\mathbf{D}^{-1}$ with fixed degrees of freedom and scale matrix. The constant c was chosen as $1/(k-3)$ and 1 for the correlation and Givens-angle HPs (though we examined the results for various choices of c, which we discuss in Sec. 3.3). The density of the Berger–Bernardo prior is $\pi(\mathbf{D}) \propto 1/[|D| \prod_{i<j}(\lambda_i - \lambda_j)]$, where λ_k are the ordered eigenvalues. This prior attempts to improve the estimate by shrinking the eigenstructure of the covariance matrix. Similar priors, as discussed earlier, have been proposed by Dey and Srinivasan (1985), Haff (1991), Lin and Perlman (1985), and Stein (1975), Yang and Berger (1994) compared their estimator to the estimators of Stein and Haff and found it to be quite competitive, so we include only the Berger–Bernardo prior in our simulation (for more details, see Yang and Berger 1994). For the fixed-hyperparameter Wishart prior on $\mathbf{D}^{-1}$, we used p degrees of freedom. We considered taking the scale matrix equal to the sample variance (which is analogous to the MLE plug-in estimator commonly used in hierarchical models), but in this case the Bayes estimator is identical to that from Jeffreys's prior, which we have already discussed. As an alternative, we allowed the scale matrix to be unknown and put a flat prior on it.

3.2.2 True Covariance Matrices. We considered three diagonal covariance matrices: **I**, equal eigenvalues; **II**, roughly equally spaced eigenvalues; and **III**, a somewhat ill-conditioned matrix. We then combined these with rotations to produce seven true covariance matrices:

I. diag(1, 1, 1, 1, 1)

II. diag(1, .75, $(.75)^2$, $(.75)^3$, $(.75)^4$) = diag(1, .75, .56, .42, .32)

III. diag(1, .75, $(.75)^2$, $(.75)^{10}$, $(.75)^{20}$) = diag(1, .75, .56, .06, .003)

$\mathbf{IIR}_1$. matrix **II** with Givens angles all set to $\pi/4$

$\mathbf{IIIR}_1$. matrix **III** with Givens angles all set to $\pi/4$

$\mathbf{IIR}_2$. matrix **II** with Givens angles evenly spaced between $(-\pi/4, \pi/4)$

$\mathbf{IIIR}_2$. matrix **III** with Givens angles evenly spaced between $(-\pi/4, \pi/4)$.

The correlation matrices associated with these covariance matrices are given in Appendix A.

3.3 Results and Conclusions

The results of the simulation are summarized in Table 1. Before discussing them, we mention that for sample sizes 10 and 20 the estimates themselves tended to be substantially different. Thus, where different estimators lead to notably different risks, these indicate distinctions that will often be substantively important.

We first compare the risk of the Bayes estimator based on the Givens-angle HP to the risk of that based on Jeffreys's prior (i.e., the sample covariance matrix). The HP results in substantial gains for samples sizes 10 and 20, and for the diagonal matrices even with a sample size of 100. We also computed the percentage reduction in average loss for the Givens-angle HP relative to the standard estimator based on Jeffreys's prior. For sample sizes 10 and 20, the HP had about a 50% reduction in risk for all matrices except the rotations of the ill-conditioned matrix **III**, for which there was about a 20% reduction in risk.

The Bayes estimator from the correlation HP performs well, but not quite as well as the Givens-angle HP for either the identity or matrix **II** and its rotated versions. The choice of values for the constant c for the Givens angle and correlation HP did not change the substantive results of both priors performing considerably better than Jeffreys's prior. The Bayes estimator from the Wishart HP does nearly as well as the Givens-angle HP in most cases, but is terrible for the two rotated versions of matrix **III**. The cause of the difficulties for the Wishart prior in these cases becomes evident when one examines the posterior distributions of the degrees of freedom, which are concentrated near the boundary, $p-1$ (as the data support a matrix fairly far from diagonal here). The restriction that the degrees of freedom be greater than $p-1$ is severe, because it forces the Wishart to remain somewhat concentrated rather than very diffuse. The concentration parameter τ^{-2} that appears in the Givens-angle and correlation HPs is analogous to the degrees of freedom parameter, but τ^{-2} may become arbitrarily small. This gives these other two priors their advantage over the Wishart.

The Bayes estimator from the Berger–Bernardo prior does as well as those from the Givens-angle HP for the strongly nondiagonal and ill-conditioned matrices $\mathbf{IIIR}_1$ and $\mathbf{IIIR}_2$ but, not surprisingly, suffers when the true matrix is diagonal—particularly when it is diagonal and poorly conditioned. In addition, the Berger–Bernardo prior produces a somewhat greater risk than the Givens HP for the nondiagonal matrices $\mathbf{IIR}_1$ and $\mathbf{IIR}_2$.

The results from our nonhierarchical Wishart prior, in which the degrees of freedom is set equal to the smallest possible integer value, p, are given in the last column of Table 1. Although this approach reduces the risk in some cases, it again (like the hierarchical Wishart) performs quite poorly for the rotated versions of matrix **III**.

4. COMPUTATION IN HIERARCHICAL MODELS

In this section we consider the use of nonconjugate priors for the covariance matrix when it appears in the second stage of a hierarchical model. We focus on the model introduced in Section 1.

The usual approach to posterior simulation in Bayesian hierarchical models is to apply Gibbs sampling, possibly using Metropolis steps to generate individual components (Smith and Roberts 1993). This works quite well when the full conditional distribution of $\mathbf{D}$ is distributed as inverse Wishart, which occurs when an inverse Wishart prior or a flat prior is placed on $\mathbf{D}$.

For the Wishart HP of Section 2.3.1, Gibbs sampling can be used with good results, as the full conditional of $\mathbf{D}$ will

Table 1. Risk for Several Estimators

D	Jeffreys	Givens HP	Correlation HP	Wishart HP	Berger–Bernardo	Wishart, df = 5
I	1.89 (.06)	.55 (.04)	.73 (.04)	.62 (.04)	.79 (.05)	1.23 (.05)
	.84 (.03)	.24 (.02)	.32 (.02)	.30 (.02)	.31 (.02)	.65 (.02)
	.15 (.00)	.05 (.00)	.06 (.00)	.06 (.00)	.05 (.00)	.14 (.00)
II	1.93 (.07)	.52 (.04)	.72 (.04)	.62 (.04)	.93 (.06)	1.23 (.05)
	.79 (.03)	.22 (.02)	.30 (.02)	.28 (.02)	.43 (.02)	.61 (.02)
	.15 (.01)	.05 (.00)	.06 (.00)	.06 (.00)	.13 (.00)	.14 (.01)
III	1.83 (.06)	.81 (.05)	.71 (.04)	.61 (.03)	1.54 (.06)	1.18 (.04)
	.81 (.03)	.29 (.02)	.30 (.02)	.27 (.02)	.68 (.02)	.62 (.02)
	.14 (.01)	.06 (.00)	.06 (.00)	.06 (.00)	.13 (.01)	.14 (.01)
IIR_1	1.99 (.08)	.78 (.05)	.94 (.05)	.90 (.05)	1.0 (.06)	1.30 (.06)
	.82 (.03)	.38 (.02)	.46 (.02)	.45 (.02)	.43 (.02)	.64 (.02)
	.16 (.01)	.13 (.00)	.15 (.00)	.14 (.00)	.13 (.00)	.14 (.01)
$IIIR_1$	1.86 (.07)	1.65 (.07)	1.76 (.08)	5.58 (.38)	1.50 (.06)	4.24 (.14)
	.81 (.03)	.68 (.03)	.80 (.03)	1.27 (.04)	.67 (.03)	1.41 (.04)
	.16 (.01)	.14 (.01)	.15 (.01)	.17 (.01)	.14 (.01)	.18 (.01)
IIR_2	1.80 (.06)	.63 (.03)	.84 (.04)	.82 (.04)	.84 (.04)	1.18 (.05)
	.85 (.03)	.39 (.02)	.51 (.02)	.52 (.02)	.45 (.02)	.67 (.02)
	.14 (.00)	.11 (.00)	.13 (.01)	.13 (.01)	.12 (.00)	.14 (.00)
$IIIR_2$	1.90 (.08)	1.69 (.08)	1.65 (.07)	10.59 (.76)	1.55 (.07)	7.57 (.22)
	.80 (.03)	.68 (.03)	.66 (.03)	1.34 (.05)	.67 (.03)	1.68 (.05)
	.15 (.01)	.13 (.00)	.11 (.00)	.16 (.01)	.13 (.00)	.16 (.01)

NOTE: Risks for sample sizes 10, 20, and 100 are given. Simulation standard errors are in parentheses. The true values of **D** are described in Section 3.2.2.

still be inverse Wishart. (Full conditionals for the additional $p + 1$ parameters, the diagonal elements of the scale matrix, and the degrees of freedom can be simulated easily using the Metropolis algorithm.) However, the results of Section 3 led us to think about implementation issues for nonconjugate HPs.

4.1 Nonconjugate Priors

Problems arise when generating $\mathbf{D}$ when the full conditional for $\mathbf{D}$ is not inverse Wishart. Using Gibbs sampling, one might generate from the full conditional of $\mathbf{D}$ componentwise. This approach can work quite well in various situations (see, e.g., Barnard et al. 1996). But problems can arise from high cross-correlations between components of $\mathbf{D}$ and also between the β_i's and the components of $\mathbf{D}$. In addition, a univariate approach requires $p(p+1)/2$ evaluations of the full conditional distribution of $\mathbf{D}$, which can become costly as p increases. Consequently, we focus on methods that generate the entire matrix $\mathbf{D}$ at once. For this, one might consider using a normal or t approximation to the full conditional and then using an independence sampler, a random walk Metropolis algorithm, or the hit-and-run sampler (Yang and Berger 1994). The approximation to the full conditional, however, would require an expensive maximization at each iteration, whereas the hit-and-run sampler and the random walk Metropolis algorithm would result in low acceptance probabilities; in addition, there would be substantial correlation between $(\beta_i, \gamma, \mathbf{D})$. Thus the chain would again move quite slowly through the posterior distribution of $\mathbf{D}$. (For further discussion of issues and techniques for generating from the posterior for $\mathbf{D}$, see Daniels 1998.)

To avoid these difficulties, we propose using a combination of approximation and importance sampling (and in some cases the Metropolis algorithm). In outline, our approach is to (a) compute the normal approximation to the MLE of the β_i (which allows us to integrate out analytically the β_i and the γ), (b) sample from an approximate marginal posterior distribution for $\mathbf{D}$ (which avoids the correlation between the $\beta_i, \gamma, \mathbf{D}$ from the Gibbs sampling), (c) sample γ and β_i from known normal distributions, and then (d) correct the approximate posterior sample with importance weights.

We implement our approach in the following steps:

1. Replace the likelihood with a normal approximation to the MLE, $\hat{\beta}_i \sim \mathrm{N}(\beta_i, \Sigma_i)$. We can now analytically integrate out the β_i's and γ to obtain an approximation to marginal posterior distribution of $\mathbf{D}$.

2. Compute the mode and Hessian for this approximate marginal posterior distribution for $\mathbf{D}$, using a parameterization in which $\mathbf{D}$ will be guaranteed to be positive definite. We have used one of two alternatives: log matrix $\mathbf{D}$ (for priors directly on $\mathbf{D}$ or $\mathbf{D}^{-1}$) or logit of angles and log differences of eigenvalues (for priors on angles and eigenvalues) (Pinheiro and Bates 1996). In the following, $\mathbf{D}^*$ denotes the appropriate transformation of $\mathbf{D}$. For the hierarchical priors, we have one additional parameter, τ^2. However, if we place a conjugate prior on τ^2, then we can integrate it out analytically.

3. Sample from $\mathbf{D}^*$ using either a normal or a t approximation, denoted by $p^*(\mathbf{D}^*|y)$.

4. Given $\mathbf{D}$, sample from $\gamma|\mathbf{D}, y$, which will be normally distributed. Specifically, $\pi(\gamma|\mathbf{D}, y) \sim \mathrm{N}(\Omega \sum_i (\mathbf{D} + \Sigma_i)^{-1} \hat{\beta}_i, (\sum_i (\mathbf{D} + \Sigma_i)^{-1})^{-1})$, with $\Omega = (\sum_i (\mathbf{D} + \Sigma_i)^{-1})^{-1}$.

5. Sample from $\beta|\gamma, \mathbf{D}, y$, which will also be normally distributed. Specifically, $\pi(\beta_i|\gamma, \mathbf{D}, y) \sim \mathrm{N}(\Omega_2(\Sigma_i^{-1}\hat{\beta}_i + \mathbf{D}^{-1}\gamma), \Omega_2)$, with $\Omega_2 = (\mathbf{D}^{-1} + \Sigma_i^{-1})^{-1}$. An alternative is to use a Laplace approximation to sample from an alternative normal approximation to the full conditional for β by computing the mode and Hessian for β conditional on $(\gamma, \mathbf{D}, y)$.

6. We now have an observation from the joint approximate posterior distribution of $(\beta, \gamma, \mathbf{D})$. The importance weight for it will be the ratio of the true to the approximate likelihood multiplied by the ratio of the approximate marginal posterior for $\mathbf{D}^*$ to the normal or the t approximation to this distribution $(p^*), w_1 = \prod_{i=1}^{n}\{[p(y_i|\beta_i)]/[\hat{p}(y_i|\beta_i)]\} \times \{[\hat{p}(\mathbf{D}^*|y)]/[p^*(\mathbf{D}^*|y)]\}$. However, if we use the Laplace approximation to sample from the full conditional of β_i, then we will instead derive the following weight:

$$w_2 = \prod_{i=1}^{n} \frac{p(y_i|\beta_i)}{\hat{p}(y_i|\beta_i)} \frac{\hat{p}(\beta_i|\gamma, \mathbf{D}, y)}{\hat{p}^*(\beta_i|\gamma, \mathbf{D}, y)} \times \frac{\hat{p}(\mathbf{D}^*|y)}{p^*(\mathbf{D}^*|y)},$$

with $\hat{p}^*$ the Laplace approximation. (See App. B for verification that these are the correct importance weights.) Using this importance sampling procedure will produce posterior means and variances. If we desire a sample from the posterior distribution, then we can obtain it by resampling (Tanner 1993, sec. 5.7).

4.2 Computational Issues

Several issues and problems can arise with importance sampling. Here, we consider two distinct problems: problems with the weights derived from the normal approximations to the likelihoods discussed in Section 4.2.2, and problems with the weights derived from sampling from approximations to the marginal posterior distribution of $\mathbf{D}^*$. In the first situation, if the normal approximation to the likelihood is poor, then the weights can be quite unstable. One potential way to diminish this problem for the weights for $(\gamma, \mathbf{D})$ would be to sample several sets of β_i's and use an average weight. So we might sample $m\beta_i$ for each value of $(\gamma, \mathbf{D})$ and assign the following weight to that particular $(\gamma, \mathbf{D})$: $w^* = \sum_{j=1}^{m} \prod_{i=1}^{n}\{[p(y_i|\beta_i^{(j)})]/[\hat{p}(y_i|\beta_i^{(j)})]\}$. By doing this, we are reducing the variability of the weights by using a weight based on $m > 1$ values.

An alternative that would alleviate to some extent the problem of the weights for the individual β_i would be to sample an additional β_i from a multivariate t approximation to the full conditional in addition to the original sample from the normal approximation. We now correct the approximation for the individual β_i via the appropriate weight (see App. B for details on this weight).

The second potential problem with importance sampling is the instability of weights due to an inadequate approximation to the marginal distribution of $\mathbf{D}^*$. For situations where a standard normal or t approximation to the approximate marginal posterior of $\mathbf{D}^*$ is inadequate, a finer approximation can be fit a priori fairly easily using the BAYESPACK software of Alan Genz (Genz and Kass 1996), which can fit split-t distributions for each of the p dimensions. However, even this specialized approximation can be inadequate. In such situations, one can sample from $\mathbf{D}^*$ using the random walk Metropolis algorithm (with the covariance matrix some multiple of the inverse of the observed information) or the hit-and-run sampler. This still has an advantage over Gibbs sampling, as there will be correlation within the sample of $\mathbf{D}^*$ but not the correlations within and between the fixed effects (γ), the random effects (β_i), and $\mathbf{D}^*$. Regardless, as discussed by Daniels (1998), further research is needed to determine consistently effective ways to sample from the approximate marginal posterior of $\mathbf{D}^*$. The importance sampling strategy that we have outlined here is just one possibility, and our work leads us to think it will be effective when p is not too large.

5. EXAMPLES

We consider two examples. The first example was chosen primarily to evaluate our computational algorithm because it has been used recently to compare alternative MCMC methods. The second is drawn from an application and illustrates how posterior estimates may be modified using HPs.

5.1 Example I: Guinea Pig Data

This example involves data recently reanalyzed by Bennet, Racine-Poon, and Wakefield (1995) concerning prediction of uptake volume of guinea pig tissue by concentration of β-methylglucoside. The response, y_{ij}, represents the uptake volume for the jth concentration $(j = 1, \ldots, 10)$ of the ith guinea pig $(i = 1, \ldots, 8)$, and X_{ij} represents the corresponding concentration of β-methylglucoside. Bennet et al. fit the following nonlinear hierarchical model to the data:

$$\log(y_{ij}) = \log\left(\frac{\exp(\beta_{1i})X_{ij}}{\exp(\beta_{2i}) + X_{ij}} + \exp(\beta_{3i})X_{ij}\right) + \varepsilon_{ij},$$

$$\varepsilon_{ij} \sim \mathrm{N}(0, \tau^2), \tag{4}$$

$$\beta_i \sim \mathrm{N}(\gamma, \mathbf{D}), \tag{5}$$

and

$$\gamma \sim d\gamma, \mathbf{D} \sim \pi(\mathbf{D}), \tag{6}$$

with $\pi(\mathbf{D})$ an inverse Wishart prior with 3 degrees of freedom and known scale matrix. For simplicity, we consider the within–guinea pig variance, τ^2, to be known.

The estimates produced from HPs turn out to be remarkably similar to those obtained by Bennet et al. using a conjugate prior. Our main purpose here, however, is to compare computational methods. Specifically, our first interest is in comparing the importance sampling approach discussed in Section 4 to full conditional MCMC schemes when the full conditional of $\mathbf{D}^{-1}$ is not Wishart. To make this comparison, we found it most convenient to use a Wishart prior on $\mathbf{D}^{-1}$, so that the full conditional actually was Wishart, but to ignore our ability to generate directly from the full conditional in setting up an MCMC alternative. This allowed us to use Gibbs sampling as a baseline and to refer to Bennet et al. for additional comparisons. We set the Wishart scale matrix equal to diag(.02, .02, .1) (following Bennet et al.) and took the degrees of freedom equal to 3. For the MCMC scheme, we computed a normal approximation to the log matrix $\mathbf{D}$ for each iteration and sampled from this using a Metropolis step.

We ran 2,000 iterations of each method and computed the time to run the chain and the Monte Carlo standard errors of the posterior means for each method; the results are given in Table 2. For the MCMC runs, we computed the Monte Carlo standard error using the method of batch means (see, e.g., Gilks, Richardson, and Spiegelhalter 1995, p. 50) and for the importance sampling run, we computed the approximate Monte Carlo standard error using an approximation to the variance of the importance ratio (see, e.g., O'Hagan 1995, p. 224).

To achieve 2,000 iterations with each method, the running times on an HP-715 workstation were 297 seconds for importance sampling versus 4,070 seconds for MCMC. Gibbs sampling took 154 seconds. We also computed the ratio of the simulation variances for MCMC to that of importance sampling after 2,000 iterations (for MCMC, 2,000 iterations following burn-in). These, also given in Table 2, show that the importance sampling scheme is generally somewhat more efficient in terms of variability per generated value. Taken together, the importance sampling algorithm is at least a reasonable competitor to the MCMC scheme and appears to be much more efficient than the MCMC scheme when a nonconjugate prior is used for $\mathbf{D}$.

Table 2. Computational Comparison of the Gibbs Sampler With Importance Sampling

	Gibbs/IS	MCMC*/IS
β_{11}	2.90	4.15
β_{12}	3.33	4.91
β_{13}	1.92	3.03
β_{21}	1.85	2.60
β_{22}	1.75	1.89
β_{23}	1.15	1.42
β_{31}	1.44	1.93
β_{32}	1.40	2.38
β_{33}	.72	.77
β_{41}	.88	.84
β_{42}	.69	.67
β_{43}	.64	.56
β_{51}	2.48	2.60
β_{52}	2.67	3.06
β_{53}	1.14	1.23
β_{61}	1.91	1.81
β_{62}	2.04	2.04
β_{63}	1.42	1.09
β_{71}	.59	.60
β_{72}	.91	.74
β_{73}	.65	.68
β_{81}	.64	2.32
β_{82}	.78	2.30
β_{83}	.89	1.35
γ_1	1.86	2.47
γ_2	2.22	2.30
γ_3	1.11	1.34
$\mathbf{D}_{11}$	.77	1.84
$\mathbf{D}_{21}$	.55	1.34
$\mathbf{D}_{22}$	.78	2.34
$\mathbf{D}_{31}$	.20	.60
$\mathbf{D}_{32}$	.26	.78
$\mathbf{D}_{33}$	.24	.36

NOTE: The numbers in each column are the ratios of Monte Carlo variances for 2,000 iterations. MCMC* denotes posterior simulation without using the fact that the full conditional of $\mathbf{D}^{-1}$ is Wishart.

5.2 Example II: Hospital Data

We now discuss our analysis of the hospital data introduced in Section 1. Based on the risk calculations of Section 3, we would expect the Givens-angle HP or the correlation HP to provide better estimates than either the Wishart HP or the conventional flat prior.

We fit the model using four separate priors for $\mathbf{D}$: the Givens-angle HP, the Wishart HP, the Wishart with estimated scale matrix and 4 degrees of freedom, and the flat prior. In fitting the Givens-angle HP, the random walk Metropolis algorithm was used to sample from the approximate marginal posterior distribution for $\mathbf{D}^*$. The results of this model fitting appear in Table 3. The two states that appear in the table were chosen to be representative of a small sample size, Delaware ($n_3 = 7$), and a large sample size, Florida ($n_5 = 249$).

We examine the fixed effects (γ), the random effects (β_i), and the covariance matrix ($\mathbf{D}$). The estimates and standard errors of the fixed effects were roughly equal across all priors, but the standard errors were uniformly larger for the flat prior. In terms of the random effects, β_5 (the large state) was quite stable across all priors, as its estimate is dominated by the likelihood for that state. However, for β_3 (the small state), there were substantial disparities in the estimates and standard errors (e.g., of β_{34}). This illustrates how improved estimation of $\mathbf{D}$ can have a substantive impact on inference about the random effects that are often of direct interest. For the covariance matrix, $\mathbf{D}$, the flat prior resulted in the largest diagonal elements (and most uncertainty). For the two HPs, shrinkage of the off-diagonal elements toward 0 relative to the flat and standard Wishart prior is apparent, with the greatest shrinkage here under the Wishart HP. In addition, the standard Wishart and Wishart HP appear to give more precision in estimating $\mathbf{D}$ than the Givens-angle HP (based on observation of the standard errors and credible intervals).

Overall, then, we observe nontrivial differences among the estimates provided by these alternative priors. Based on the results of Section 3, we would prefer the Givens-angle HP. In this case we would obtain modest shrinkage toward diagonality with a modest increase in precision (more precision than when the flat prior is used, but less than when the Wishart priors are used). We would expect the posterior from the Givens-angle HP to have adapted reasonably well to the departure from diagonality of $\mathbf{D}$.

6. DISCUSSION

In small-sample Bayesian estimation of a covariance matrix $\mathbf{D}$, the choice of prior distribution is important. Furthermore, when $\mathbf{D}$ appears in the second stage of a hierarchical model, its estimation affects the estimation of standard errors for the usual quantities of interest, namely the fixed and random effects. The use of informative conjugate prior distributions requires detailed information about $\mathbf{D}$, which is often very difficult to obtain. In addition, the commonly applied device of using the MLE of $\mathbf{D}$ to define the scale matrix in an inverse-Wishart prior on $\mathbf{D}$ is shown by our simulation study to perform poorly in some cases. We have

Table 3. Posterior Means, Standard Deviations, and 95% Credible Intervals for the Third State (Delaware) and the Fifth State (Florida) in Example II

Parameter	Standard Wishart			Wishart HP			Givens HP			Flat prior		
β_{31}	−2.77	.168	[−3.11, −2.47]	−2.68	.128	[−2.96, −2.45]	−2.68	.147	[−3.01, −2.43]	−2.84	.202	[−3.29, −2.49]
β_{32}	.18	.070	[.04, .33]	.20	.058	[.09, .33]	.21	.090	[.03, .40]	.22	.113	[−.01, .44]
β_{33}	.25	.222	[−.16, .71]	.14	.181	[−.18, .52]	.13	.190	[−.20, .54]	.29	.276	[−.21, .88]
β_{34}	.80	.238	[.31, 1.28]	.55	.195	[.12, .92]	.54	.252	[.01, 1.03]	.66	.372	[−.00, 1.39]
β_{51}	−2.42	.045	[−2.50, −2.33]	−2.42	.047	[−2.51, −2.33]	−2.40	.047	[−2.49, −2.31]	−2.40	.048	[−2.50, −2.31]
β_{52}	.13	.033	[.06, .19]	.13	.036	[.05, .19]	.11	.036	[.04, .18]	.11	.037	[.03, .18]
β_{53}	−.16	.076	[−.31, −.01]	−.17	.076	[−.32, −.03]	−.15	.076	[−.30, −.00]	−.17	.077	[−.32, −.02]
β_{54}	.71	.089	[.54, .88]	.72	.090	[.56, .90]	.72	.094	[.54, .91]	.73	.098	[.54, .93]
γ_1	−2.49	.057	[−2.60, −2.38]	−2.48	.051	[−2.58, −2.38]	−2.47	.050	[−2.58, −2.38]	−2.49	.081	[−2.65, −2.35]
γ_2	.19	.025	[.14, .24]	.19	.024	[.15, .24]	.19	.033	[.13, .26]	.20	.039	[.12, .28]
γ_3	−.02	.083	[−.19, .14]	−.01	.072	[−.16, .13]	−.02	.068	[−.15, .12]	−.02	.115	[−.25, .21]
γ_4	.60	.078	[.45, .75]	.57	.071	[.41, .70]	.56	.079	[.39, .71]	.57	.106	[.38, .79]
$\mathbf{D}_{11}$	.045	.020	[.019, .095]	.034	.018	[.013, .076]	.033	.024	[.009, .098]	.095	.077	[.024, .316]
$\mathbf{D}_{21}$	.000	.005	[−.011, .011]	−.000	.001	[−.002, .001]	−.002	.010	[−.025, .015]	−.012	.028	[−.078, .029]
$\mathbf{D}_{22}$	.004	.002	[.001, .009]	.003	.003	[.000, .012]	.011	.010	[.001, .036]	.017	.018	[.003, .059]
$\mathbf{D}_{31}$	−.034	.022	[−.086, −.004]	−.001	.004	[−.008, .005]	−.014	.021	[−.071, .007]	−.072	.089	[−.286, .028]
$\mathbf{D}_{32}$	−.005	.007	[−.020, .006]	.000	.001	[−.002, .002]	.000	.010	[−.020, .018]	−.002	.033	[−.079, .057]
$\mathbf{D}_{33}$	.083	.043	[.034, .180]	.061	.042	[.012, .167]	.051	.061	[.006, .176]	.180	.166	[.035, .609]
$\mathbf{D}_{41}$	−.033	.120	[−.082, −.006]	−.000	.003	[−.007, .006]	−.003	.013	[−.032, .018]	−.035	.089	[−.272, .086]
$\mathbf{D}_{42}$	−.002	.006	[−.017, .008]	−.000	.001	[−.002, .001]	−.001	.007	[−.017, .011]	−.006	.032	[−.069, .055]
$\mathbf{D}_{43}$	.036	.027	[−.001, .103]	.000	.004	[−.006, .007]	.001	.019	[−.036, .041]	.025	.107	[−.137, .244]
$\mathbf{D}_{44}$	.057	.030	[.021, .133]	.039	.032	[.006, .114]	.062	.066	[.006, .211]	.139	.169	[.017, .479]

NOTE: Here standard Wishart is a Wishart distribution with estimated scale matrix and $df = 4$, and Wishart HP and Givens HP are the Wishart and Givens-angle HPs.

introduced a set of HPs as a generic means to stabilize the estimate of the covariance matrix for small samples. We have applied the method under the assumption that D is not too far from diagonality, though we have shown that the Givens-angle HP produces reasonable estimates even when this assumption is false, partly because—in the form we have used—it results in shrinkage of the eigenvalues. Our own feeling is that the Givens-angle HP is mathematically more natural than the other two HPs we investigated. In any case, we found that the HP on the correlations also produces good estimates, but with risks somewhat larger than those based on the Givens-angle HP. The Wishart HP sometimes leads to very inaccurate estimates, and thus we would be very concerned about using it in practice.

We have emphasized shrinkage toward diagonality for D because it is straightforward and often applicable. Indeed, D is sometimes assumed to be diagonal for convenience (e.g., in variance components models), even though this assumption may be dubious. The method that we have described and investigated could be applied to other structures in covariance matrices, however. For instance, in time series or spatial applications, an autoregressive (AR) model may be plausible yet not fully supportable by available data. It is then possible to shrink toward the hypothesized AR structure while allowing the data to support some departure from that structure (Daniels and Cressie, 1998).

Another interesting potential domain of application for the methods that we have described is in the estimation of large covariance matrices, especially when the number of individuals k is less than p so that the sample covariance matrix is noninvertible and the conjugate posterior density no longer exists. A frequentist shrinkage method has been proposed and investigated by Ledoit (1996). We have begun to explore Bayesian approaches to this problem.

We have also shown how the computational burden in using the HPs, and nonconjugate priors in general, can be reduced by combining approximations and importance sampling. We can imagine variations that would improve on our method. For instance, from our findings, to obtain results based on Givens-angle or correlation HPs it should often be beneficial to use the Wishart HP rather than the fixed conjugate prior within the approximate model described in Section 4.1. This is computationally efficient, and the Wishart HP leads to similar results in many cases; importance weights are again easily obtained and provide valid corrections according to the analysis presented in Appendix B. We would also like to mention that the strategy we have adopted—applying Gibbs sampling following a first-stage normal approximation, then correcting the result by importance weights—ought to be of general interest. For example, it is sometimes rather time-consuming to construct an effective MCMC scheme for a complicated hierarchical model, whereas implementation of the normal approximation may be relatively straightforward. The importance correction that we have derived could be useful in such situations. On the other hand, we do not wish to leave the impression that importance sampling accomplishes something that a good MCMC method cannot. We expect MCMC technology to provide further computational improvements in handling problems of this kind.

APPENDIX A: CORRELATION MATRICES FOR SIMULATION

Correlation matrix $\mathbf{IIR}_1$:

$$\begin{array}{ccccc} 1.000 & & & & \\ -.286 & 1.000 & & & \\ -.138 & .101 & 1.000 & & \\ -.087 & .113 & .233 & 1.000 & \\ -.149 & -.022 & .087 & .296 & 1.000 \end{array}$$

Correlation matrix $\mathbf{IIR}_2$:

$$\begin{array}{ccccc} 1.000 & & & & \\ -.035 & 1.000 & & & \\ .221 & .232 & 1.000 & & \\ .252 & .216 & .138 & 1.000 & \\ .273 & .103 & -.020 & .128 & 1.000 \end{array}$$

Correlation matrix $\mathbf{IIIR}_1$:

$$\begin{array}{ccccc} 1.000 & & & & \\ -.871 & 1.000 & & & \\ -.223 & .097 & 1.000 & & \\ -.357 & .165 & .651 & 1.000 & \\ -.339 & -.054 & .013 & .590 & 1.000 \end{array}$$

Correlation matrix $\mathbf{IIIR}_2$:

$$\begin{array}{ccccc} 1.000 & & & & \\ -.146 & 1.000 & & & \\ .358 & .306 & 1.000 & & \\ .598 & .535 & .245 & 1.000 & \\ .593 & .217 & -.269 & .608 & 1.000 \end{array}$$

APPENDIX B: DERIVATION OF IMPORTANCE WEIGHTS

In general, to compute the posterior expectation of some function of a parameter $g(\theta)$, we need to compute the following ratio of integrals: $E[g(\theta)|y] = [\int g(\theta)p(\theta|y)\,d\theta]/[\int p(\theta|y)\,d\theta]$. Importance sampling rewrites this ratio as

$$E[g(\theta)|y] = \frac{\int g(\theta)w(\theta)\hat{p}(\theta|y)\,d\theta}{\int w(\theta)\hat{p}(\theta|y)\,d\theta}, \tag{B.1}$$

where $\hat{p}(\theta|y)$ approximates $p(\theta|y)$ and is easy to sample from, and $w(\theta) = [p(\theta|y)]/[\hat{p}(\theta|y)]$. Using draws, θ^l, from the approximation $\hat{p}(\theta|y)$, the ratio (B.1) is approximated by a ratio of sums, according to the law of large numbers; that is,

$$E[g(\theta)|y] \approx \frac{\sum g(\theta^l)w(\theta^l)}{\sum w(\theta^l)}.$$

Thus, to show that our importance sampling scheme correctly reweights the sampled observations, we must show that the correspondingly defined weight functions $w(\theta)$ correctly produce posterior expectations of the form $E[g(\theta)|y]$ according to (B.1).

Here we have $\theta = (\beta_i, \gamma, \mathbf{D})$, and we consider the posterior expectation of $g(\gamma)$. Other functions of the parameter vector may be treated similarly.

Let A be the numerator integral and B the denominator integral; that is, $E[g(\gamma|y)] = A/B$, where

$$A = \int\int\cdots\int g(\gamma)\prod_{i=1}^{n} p(\beta_i|\gamma,\mathbf{D},y) \times p(\gamma,\mathbf{D}|y)\,d\gamma\,d\mathbf{D}\,d\beta_1\ldots d\beta_n$$

and

$$B = \int\int\cdots\int\prod_{i=1}^{n} p(\beta_i|\gamma,\mathbf{D},y)p(\gamma,\mathbf{D}|y)\,d\gamma\,d\mathbf{D}\,d\beta_1\ldots d\beta_n.$$

We begin by rewriting A:

$$\begin{aligned} A &= \int\int\cdots\int g(\gamma)\prod_{i=1}^{n}\frac{p(\beta_i|\gamma,\mathbf{D},y)}{\hat{p}(\beta_i|\gamma,\mathbf{D},y)}\,\hat{p}(\beta_i|\gamma,\mathbf{D},y) \\ &\quad\times\frac{p(\gamma,\mathbf{D}|y)}{\hat{p}(\gamma,\mathbf{D}|y)}\,\hat{p}(\gamma,\mathbf{D}|y)\,d\gamma\,d\mathbf{D}\,d\beta_1\ldots d\beta_n \\ &= \int\int\cdots\int g(\gamma)\prod_{i=1}^{n}\frac{p(\beta_i,\gamma,\mathbf{D}|y)}{\hat{p}(\beta_i,\gamma,\mathbf{D}|y)}\,\hat{p}(\beta_i|\gamma,\mathbf{D},y) \\ &\quad\times\hat{p}(\gamma,\mathbf{D}|y)\,d\gamma\,d\mathbf{D}\,d\beta_1\ldots d\beta_n \\ &= \int\int\cdots\int g(\gamma)\prod_{i=1}^{n}\frac{p(\beta_i,\gamma,\mathbf{D},y)\hat{m}(y)}{\hat{p}(\beta_i,\gamma,\mathbf{D},y)m(y)} \\ &\quad\times\hat{p}(\beta_i|\gamma,\mathbf{D},y)\hat{p}(\gamma,\mathbf{D}|y)\,d\gamma\,d\mathbf{D}\,d\beta_1\ldots d\beta_n, \end{aligned}$$

with $m(y)$ denoting the marginal distribution of the data under the true model and $\hat{m}(y)$ denoting the marginal distribution of the data under the approximate model. Thus

$$\begin{aligned} A &= \int\int\cdots\int g(\gamma)\prod_{i=1}^{n}\frac{p(y_i|\beta_i,\gamma,\mathbf{D})p(\beta_i,\gamma,\mathbf{D})\hat{m}(y)}{\hat{p}(y_i|\beta_i,\gamma,\mathbf{D})p(\beta_i,\gamma,\mathbf{D})m(y)} \\ &\quad\times\hat{p}(\beta_i|\gamma,\mathbf{D},y)\hat{p}(\gamma,\mathbf{D}|y)\,d\gamma\,d\mathbf{D}\,d\beta_1\ldots d\beta_n \\ &= \int\int\cdots\int g(\gamma)\prod_{i=1}^{n}\frac{p(y_i|\beta_i,\gamma,\mathbf{D})\hat{m}(y)}{\hat{p}(y_i|\beta_i,\gamma,\mathbf{D})m(y)} \\ &\quad\times\hat{p}(\beta_i|\gamma,\mathbf{D},y)\hat{p}(\gamma,\mathbf{D}|y)\,d\gamma\,d\mathbf{D}\,d\beta_1\ldots d\beta_n \\ &= \frac{\hat{m}(y)}{m(y)}\int\int\cdots\int g(\gamma)\prod_{i=1}^{n}\frac{p(y_i|\beta_i,\gamma,\mathbf{D})}{\hat{p}(y_i|\beta_i,\gamma,\mathbf{D})} \\ &\quad\times\hat{p}(\beta_i|\gamma,\mathbf{D},y)\hat{p}(\gamma,\mathbf{D}|y)\,d\gamma\,d\mathbf{D}\,d\beta_1\ldots d\beta_n. \end{aligned}$$

This shows that the importance weight function of Section 4.1 correctly produces the numerator A. Similarly, we obtain

$$B = \frac{\hat{m}(y)}{m(y)}\int\int\cdots\int\prod_{i=1}^{n}\frac{p(y_i|\beta_i,\gamma,\mathbf{D})}{\hat{p}(y_i|\beta_i,\gamma,\mathbf{D})} \times\hat{p}(\beta_i|\gamma,\mathbf{D},y)\hat{p}(\gamma,\mathbf{D}|y)\,d\gamma\,d\mathbf{D}\,d\beta_1\ldots d\beta_n,$$

so that the posterior expectation of $g(\gamma)$ becomes

$$E[g(\gamma)|y] = \frac{\begin{array}{l}\int\int\cdots\int g(\gamma)\prod_{i=1}^{n}\frac{p(y_i|\beta_i,\gamma,\mathbf{D})}{\hat{p}(y_i|\beta_i,\gamma,\mathbf{D})}\,\hat{p}(\beta_i|\gamma,\mathbf{D},y) \\ \quad\times\hat{p}(\gamma,\mathbf{D}|y)\,d\gamma\,d\mathbf{D}\,d\beta_1\ldots d\beta_n\end{array}}{\begin{array}{l}\int\int\cdots\int\prod_{i=1}^{n}\frac{p(y_i|\beta_i,\gamma,\mathbf{D})}{\hat{p}(y_i|\beta_i,\gamma,\mathbf{D})}\,\hat{p}(\beta_i|\gamma,\mathbf{D},y) \\ \quad\times\hat{p}(\gamma,\mathbf{D}|y)\,d\gamma\,d\mathbf{D}\,d\beta_1\ldots d\beta_n\end{array}},$$

and with $w(\theta) = \prod_{i=1}^{n}[p(y_i|\beta_i)]/[\hat{p}(y_i|\beta_i)]$, the ratio of the true to the approximate likelihood, we have the required form (B.1).

In Section 4.2 we also mentioned the possibility of obtaining more stable weights for expectations of functions of the β_i by using a distribution with heavier tails for the β_i. This will alter the weights. Let $\hat{p}_t(\beta_i|\gamma,\mathbf{D},y)$ be a t distribution with the same location and scale as the corresponding normal distribution:

$$\begin{aligned} &E[g(\beta_i)|y] \\ &= \frac{\begin{array}{l}\frac{\hat{m}(y)}{m(y)}\int\int\cdots\int g(\beta_i)\prod_{i=1}^{n}\frac{p(y|\beta_i,\gamma,\mathbf{D})}{\hat{p}(y|\beta_i,\gamma,\mathbf{D})} \\ \quad\times\hat{p}(\beta_i|\gamma,\mathbf{D},y)\hat{p}(\gamma,\mathbf{D}|y)\,d\gamma\,d\mathbf{D}\,d\beta_1\ldots d\beta_n\end{array}}{\begin{array}{l}\frac{\hat{m}(y)}{m(y)}\int\int\cdots\int\prod_{i=1}^{n}\frac{p(y|\beta_i,\gamma,\mathbf{D})}{\hat{p}(y|\beta_i,\gamma,\mathbf{D})} \\ \quad\times\hat{p}(\beta_i|\gamma,\mathbf{D},y)\hat{p}(\gamma,\mathbf{D}|y)\,d\gamma\,d\mathbf{D}\,d\beta_1\ldots d\beta_n\end{array}} \\ &= \frac{\begin{array}{l}\int\int\cdots\int g(\beta_i)\frac{p(y_i|\beta_i)\hat{p}(\beta_i|\gamma,\mathbf{D},y)}{\hat{p}(y_i|\beta_i)\hat{p}_t(\beta_i|\gamma,\mathbf{D},y)}\,\hat{p}_t(\beta_i|\gamma,\mathbf{D},y) \\ \quad\times\prod_{j\neq i}\frac{p(y|\beta_j,\gamma,\mathbf{D})}{\hat{p}(y|\beta_j,\gamma,\mathbf{D})} \\ \quad\times\hat{p}(\beta_j|\gamma,\mathbf{D},y)\hat{p}(\gamma,\mathbf{D}|y)\,d\gamma\,d\mathbf{D}\,d\beta_1\ldots d\beta_n\end{array}}{\begin{array}{l}\int\int\cdots\int g(\beta_i)\frac{p(y_i|\beta_i)\hat{p}(\beta_i|\gamma,\mathbf{D},y)}{\hat{p}(y_i|\beta_i)\hat{p}_t(\beta_i|\gamma,\mathbf{D},y)}\,\hat{p}_t(\beta_i|\gamma,\mathbf{D},y) \\ \quad\times\prod_{j\neq i}\frac{p(y|\beta_j,\gamma,\mathbf{D})}{\hat{p}(y|\beta_j,\gamma,\mathbf{D})}\,\hat{p}(\beta_j|\gamma,\mathbf{D},y) \\ \quad\times\hat{p}(y,\mathbf{D}|y)\,d\gamma\,d\mathbf{D}\,d\beta_1\ldots d\beta_n\end{array}}. \end{aligned}$$

Here

$$w(\theta) = \frac{p(y_i|\beta_i)\hat{p}(\beta_i|\gamma, \mathbf{D}, y)}{\hat{p}(y_i|\beta_i)\hat{p}_t(\beta_i|\gamma, \mathbf{D}, y)} \prod_{j \neq i} \frac{p(y|\beta_j, \gamma, \mathbf{D})}{\hat{p}(y|\beta_j, \gamma, \mathbf{D})}.$$

This new weight for β_i is the same as the old weight but with the ith term in the product replaced by a new quantity. If importance sampling is used to sample from the approximate marginal posterior distribution for $\mathbf{D}$, then both of the foregoing weights are multiplied by the ratio of the approximate marginal posterior for $\mathbf{D}$ to the approximation to this distribution (from which it is easy to sample), as discussed in Section 4.1.

[Received July 1997. Revised April 1999.]

REFERENCES

Barnard, J., McCulloch, R., and Meng, X. (1996), "A Natural Strategy for Modeling Covariance Matrices With Application to Shrinkage," technical report, Harvard University, Dept. of Statistics.

Bennet, J. E., Racine-Poon, A., and Wakefield, J. C. (1995), "MCMC for Nonlinear Hierarchical Models," in *Markov Chain Monte Carlo in Practice*, eds. W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, New York: Chapman and Hall, pp. 339–358.

Berger, J., and Bernardo, J. M. (1992), "On the Development of Reference Priors" (with discussion), in *Bayesian Statistics 4*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford, U.K.: Oxford University Press, pp. 35–60.

Christiansen, C. L., and Morris, C. N. (1997), "Hierarchical Poisson Regression Modeling," *Journal of the American Statistical Association*, 92, 618–632.

Daniels, M. J. (1998), "Computing Posterior Distributions for Covariance Matrices," *Computing Science and Statistics*, 30, ed. S. Weisberg, pp. 192–196.

——— (1999), "A Prior for the Variance in Hierarchical Models," *Canadian Journal of Statistics*, to appear.

———, and Cressie, N. (1998), "A Hierarchical Approach to Covariance Function Estimation in Time Series," submitted.

———, and Gatsonis, C. (1999), "Hierarchical Generalized Linear Models in the Analysis of Variations in Health Care Utilization," *Journal of the American Statistical Association*, 94, 29–42.

Dey, D. K., and Srinivasan, C. (1985), "Estimation of a Covariance Matrix Under Stein's Loss," *The Annals of Statistics*, 13, 1581–1591.

Everson, P. J., and Morris, C. N. (1997), "Inference for Multivariate Normal Hierarchical Models," *Journal of the Royal Statistical Society, Series B*, to appear.

Genz, A., and Kass, R. E. (1997), "Subregion-Adaptive Integration of Functions Having a Dominant Peak," *Journal of Computational and Graphical Statistics*, 6, 92–111.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1995), *Markov Chain Monte Carlo in Practice*, London: Chapman and Hall.

Goldstein, H. (1962), *Classical Mechanics*, Reading, MA: Addison-Wesley.

Haff, L. R. (1991), "The Variational Form of Certain Bayes Estimators," *The Annals of Statistics*, 19, 1163–1190.

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, 1461–1474.

Hoffman, D. K., Raffenetti, R. C., and Ruedenberg, K. (1972), "Generalization of Euler Angles to N-Dimensional Orthogonal Matrices," *Journal of Mathematical Physics*, 13, 528–533.

Kass, R. E., and Wasserman, L. (1996), "The Selection of Prior Distributions by Formal Rules," *Journal of the American Statistical Association*, 91, 1343–1370.

Ledoit, O. (1996), "A Well-Conditioned Estimator for Large Dimensional Covariance Matrices," Working paper, University of California-Los Angeles, Anderson Graduate School of Management.

Leonard, T., and Hsu, J. S. (1993), "Bayesian Inference for a Covariance Matrix," *The Annals of Statistics*, 21, 1–25.

Lin, S. P., and Perlman, M. D. (1985), "A Monte Carlo Comparison of Four Estimators for a Covariance Matrix," in *Multivariate Analysis 6*, ed. P. R. Krishnaiah, Amsterdam: North-Holland, pp. 411–429.

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.

O'Hagan, A. (1994), *Kendall's Advanced Theory of Statistics, Volume 2B: Bayesian Inference*, New York: Halsted Press.

Pinheiro, J. C., and Bates, D. M. (1996), "Unconstrained Parameterizations for Variance-Covariance Matrices," *Statistics and Computing*, 6, 1–6.

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

Smith, A. F. M., and Roberts, G. O. (1993), "Bayesian Computation via the Gibbs Sampler and Related Markov Chain Monte Carlo Methods," *Journal of the Royal Statistical Society*, Ser. B, 55, 3–23.

Stein, C. (1975), "Estimation of a Covariance Matrix," Rietz Lecture, 39th Annual meeting IMS. Atlanta, Georgia.

Sun, L., Hsu, J. S., Guttman, I., Leonard, T. (1996), "Bayesian Methods for Variance Component Models," *Journal of the American Statistical Association*, 91, 743–752.

Tanner, M. A. (1993), *Tools for Statistical Inference: Methods for Exporation of Posterior Distributions and Likelihood Functions*, New York: Springer-Verlag.

Thisted, R. A. (1988), *Elements of Statistical Computing*, London: Chapman and Hall.

Yang, R., and Berger, J. O. (1994), "Estimation of a Covariance Matrix Using the Reference Prior," *The Annals of Statistics*, 22, 1195–1211.