An Empirical Comparison of Methods for Computing Bayes Factors in Generalized Linear Mixed Models

Sandip SINHARAY and Hal S. STERN

Generalized linear mixed models (GLMM) are used in situations where a number of characteristics (covariates) affect a nonnormal response variable and the responses are correlated due to the existence of clusters or groups. For example, the responses in biological applications may be correlated due to common genetic factors or environmental factors. The clustering or grouping is addressed by introducing cluster effects to the model; the associated parameters are often treated as random effects parameters. In many applications, the magnitude of the variance components corresponding to one or more of the sets of random effects parameters are of interest, especially the point null hypothesis that one or more of the variance components is zero. A Bayesian approach to test the hypothesis is to use Bayes factors comparing the models with and without the random effects in question—this work reviews a number of approaches for estimating the Bayes factor. We perform a comparative study of the different approaches to compute Bayes factors for GLMMs by applying them to two different datasets. The first example employs a probit regression model with a single variance component to data from a natural selection study on turtles. The second example uses a disease mapping model from epidemiology, a Poisson regression model with two variance components. Bridge sampling and a recent improvement known as warp bridge sampling, importance sampling, and Chib's marginal likelihood calculation are all found to be effective. The relative advantages of the different approaches are discussed.

Key Words: Bridge sampling; Chib's method; Importance sampling; Marginal density; Reversible jump Markov chain Monte Carlo; Warp bridge sampling.

1. INTRODUCTION

Generalized linear mixed models (GLMMs), or, generalized linear models with random effects, are used in situations where a nonnormal response variable is related to a set of predictors and the responses are correlated due to the existence of groups or clusters. The

Sandip Sinharay is Research Scientist, Educational Testing Service, MS-12T, Rosedale Road, Princeton, NJ 08541 (E-mail: ssinharay@ets.org). Hal S. Stern is Professor and Chair, Department of Statistics, University of California, 4800 Berkeley Place, Irvine, CA 92697.

Journal of Computational and Graphical Statistics, Volume 14, Number 2, Pages 415–435
DOI: 10.1198/106186005X47471

groups or clusters are addressed by incorporating a set of random effects parameters in the model. In many applications, the magnitude of the variance components corresponding to one or more of the sets of random effects are of interest, especially the point null hypothesis that the variance components in question are equal to zero. A Bayesian approach for testing a hypothesis of this type is to compute the Bayes factor comparing the models suggested by the null and the alternative hypotheses. The primary objective of this work is to apply and evaluate the performance of different approaches for estimating the Bayes factor comparing the GLMMs.

A number of related studies exist in the statistical literature. Albert and Chib (1997) provided a broad survey of the use of Bayes factors for judging a variety of assumptions, including assumptions regarding the variance components, in conditionally independent hierarchical models (which include GLMMs as a special case). Han and Carlin (2001) provided a review and empirical comparison of several Markov chain Monte Carlo (MCMC) methods for estimating Bayes factors emphasizing normal linear mixed model applications. Their study does not include importance sampling or warp bridge sampling. Pauler, Wakefield, and Kass (1999) provided a number of analytic approximations for computing Bayes factors for variance component testing in linear mixed models. DiCiccio, Kass, Raftery, and Wasserman (1997) compared several simulation-based approximation methods for estimating Bayes factors. Their study was quite general whereas the present work focuses on GLMMs. The present study adds to the literature by focussing on GLMMs (more specifically on testing variance components in GLMMs), incorporating complex GLMMs with multiple variance components, and incorporating new developments like warp bridge sampling (Meng and Schilling 2003). Our work should be of great interest to researchers working with GLMMs but others will also find the results, especially the findings on warp bridge sampling, useful.

Section 2 discusses a number of preliminary ideas regarding GLMMs. Section 3 reviews the Bayes factor and approaches for estimating it. Section 4 compares the different approaches in the context of a random effects probit regression model applied to data from a natural selection study (Janzen, Tucker, and Paukstis 2000). Section 5 takes up a more complex example, a Poisson-normal regression model with spatial random effects and heterogeneity random effects applied to Scotland lip-cancer data (Clayton and Kaldor 1987). Section 6 provides a discussion of our findings and recommendations.

2. PRELIMINARIES

2.1 The Generalized Linear Mixed Model

In a GLMM, observations $y_1, y_2, \ldots, y_n$ are modeled as independent, given canonical parameters ξ_i's and a scale parameter ϕ, with density

$$f(y_i|\xi_i, \phi) = \exp\{[y_i\xi_i - a(\xi_i) + b(y_i)]/\phi\}.$$

We take $\phi = 1$ henceforth. The two examples we consider in detail do not have a scale parameter. All of the methods described here can accommodate a scale parameter. Let $\mu_i = E(y_i|\xi_i) = a'(\xi_i)$. The mean μ_i is expressed as a function of a predictor vector $\mathbf{x}_i^{p\times 1}$, a vector of coefficients $\boldsymbol{\alpha}^{p\times 1}$, and a random effects vector $\mathbf{b}^{q\times 1}$ through the link function $g(\mu_i) = \mathbf{x}_i'\boldsymbol{\alpha} + \mathbf{z}_i'\mathbf{b}$, where $\mathbf{z}_i^{q\times 1}$ is a design vector (typically 0/1) identifying the random effects. Usually, for a vector of unknown variance components $\boldsymbol{\theta}^{m\times 1}$, $f(\mathbf{b}|\boldsymbol{\theta})$ is assumed to be $\mathcal{N}(\mathbf{0}, \mathbf{D}(\boldsymbol{\theta}))$, where $\mathbf{D}(\boldsymbol{\theta})$ is positive definite. The magnitude of $\boldsymbol{\theta}$ determines the degree of over-dispersion and correlation among y_i's. Typically, $\mathbf{D}(\boldsymbol{\theta}) = \mathbf{0}$ iff $\boldsymbol{\theta} = \mathbf{0}$.

2.2 LIKELIHOOD FOR GENERALIZED LINEAR MIXED MODELS

The likelihood function $L(\boldsymbol{\alpha}, \boldsymbol{\theta}|\mathbf{y})$ for a GLMM is given by

$$L(\boldsymbol{\alpha}, \boldsymbol{\theta}|\mathbf{y}) = \int_{\mathbf{b}} \left\{ \prod_{i=1}^{n} f(y_i|\xi_i) \right\} f(\mathbf{b}|\boldsymbol{\theta}) d\mathbf{b} = \int_{\mathbf{b}} \left\{ \prod_{i=1}^{n} f(y_i|\boldsymbol{\alpha}, \mathbf{b}) \right\} f(\mathbf{b}|\boldsymbol{\theta}) d\mathbf{b} \cdot \qquad (2.1)$$

The integral is analytically intractable, making computations with GLMMs difficult. In simple problems as the one in Section 4 (where the model has a single random effect with 31 levels), it is possible to use numerical integration. Numerical integration techniques (e.g., Simpson's rule) or Laplace approximation (Tierney and Kadane 1986) are problematic for high-dimensional **b**. For the more elaborate example of Section 5, we use importance sampling to compute the likelihood (as in, e.g., Geyer and Thompson 1992).

2.3 TESTING HYPOTHESES ABOUT VARIANCE COMPONENTS FOR GLMMS

Inferences about the contribution of the random effects to a GLMM are mostly obtained by examining point (or interval) estimates of the variance parameters in **D**. In many practical problems, scientific investigators may like to test whether a particular variance component is zero. The classical approaches for testing in this context are the likelihood ratio test (LRT) using a simulation-based null distribution or the score test (Lin 1997). Our study concentrates on the Bayes factor, a Bayesian tool to perform hypothesis testing or model selection.

3. BAYES FACTORS

3.1 INTRODUCTION

The Bayesian approach to test a hypothesis about the variance component(s) is to compute the Bayes factor $\mathrm{BF}^{01} = \frac{p(\mathbf{y}|M_0)}{p(\mathbf{y}|M_1)}$, which compares the marginal densities (also known as marginal likelihoods) of the data **y** under the two models, M_0 (one or more of the variance components is zero) and M_1 (variance unrestricted) suggested by the hypotheses, where $p(\mathbf{y}|M) = \int p(\mathbf{y}|\boldsymbol{\omega}, M)p(\boldsymbol{\omega}|M)d\boldsymbol{\omega}$ is the marginal density under model M and $\boldsymbol{\omega}$ denotes the parameters of model M. Kass and Raftery (1995) provided a comprehensive review of Bayes factors including information about their interpretation.

The Bayes factor is also the ratio of posterior and prior odds,

$$\mathrm{BF}^{01} = \frac{p(M_0|\mathbf{y})}{p(M_1|\mathbf{y})} \Big/ \frac{p(M_0)}{p(M_1)} \,. \tag{3.1}$$

This expression is useful in forming an estimate of the Bayes factor via reversible jump Markov chain Monte Carlo as described in the following.

3.2 Approaches for Estimating the Bayes Factor

The key contribution of our work is to bring different computational methods to bear on the problem of estimating the Bayes factor to test for the variance components for GLMMs. For these models, the marginal densities required by the Bayes factor cannot be computed analytically for either M_1 or M_0. For the remainder of this section, $\boldsymbol{\omega} = (\boldsymbol{\alpha}, \boldsymbol{\theta})$, implying that the random effects parameters **b** have been integrated out. The final part of this section discusses an alternative parameterization.

Different approaches exist for estimating the Bayes factor. This work briefly reviews a number of such approaches that have been applied in other models and then explores their use for GLMMs. We will use the notation $p(\boldsymbol{\omega}|\mathbf{y}, M)$ to denote the posterior density under model M, and $q(\boldsymbol{\omega}|\mathbf{y}, M) \equiv p(\mathbf{y}|\boldsymbol{\omega}, M)p(\boldsymbol{\omega}|M)$ to denote the unnormalized posterior density.

We consider the following approaches in our work: (1) Importance sampling (see, e.g., DiCiccio et al. 1997); (2) Markov chain Monte Carlo based calculation of the marginal likelihood (Chib 1995; Chib and Jeliazkov 2001); (3) bridge sampling and its enhancements (Meng and Wong 1996; Meng and Schilling 2003); and (4) reversible jump MCMC (Green 1995). The methods are briefly reviewed later in this section. The first three of the above approaches estimate the marginal density of the data separately under each model, the ratio of the estimated marginal densities giving the estimated Bayes factor. Reversible jump MCMC approaches estimate the Bayes factor directly.

3.2.1 Importance Sampling

Importance sampling estimates of the marginal density are based on the identity

$$p(\mathbf{y}|M) = \int \frac{p(\mathbf{y}|\boldsymbol{\omega}, M)p(\boldsymbol{\omega}|M)}{d(\boldsymbol{\omega})} d(\boldsymbol{\omega}) d\omega$$

for any density function $d(.)$. Then given a sample $\boldsymbol{\omega}_i, i = 1, 2, \ldots, N$ from the "importance sampling distribution" $d(.)$, an estimate of the marginal density of the data under model M is

$$p(\mathbf{y}|M) \approx \frac{1}{N} \sum_{i=1}^{N} \frac{q(\boldsymbol{\omega}_i|\mathbf{y}, M)}{d(\boldsymbol{\omega}_i)} .$$

The choice of importance sampling density is crucial; the density $d(.)$ must have tails as heavy as the posterior distribution to provide reliable estimates of the marginal likelihood.

Common choices are normal or t-distributions with suitable location and scale. We discuss the choice of importance sampling density further in the section that discusses warp bridge sampling (Meng and Schilling 2003).

3.2.2 Chib's Method

Chib (1995) suggested estimating $p(\mathbf{y}|M)$ by estimating at any $\boldsymbol{\omega} = \boldsymbol{\omega}^*$ the right hand side of

$$p(\mathbf{y}|M) = \frac{p(\mathbf{y}|\boldsymbol{\omega}, M)p(\boldsymbol{\omega}|M)}{p(\boldsymbol{\omega}|\mathbf{y}, M)} \,. \tag{3.2}$$

The computation of $p(\mathbf{y}|\boldsymbol{\omega}^*, M)$ and $p(\boldsymbol{\omega}^*|M)$ are straightforward. Given a sample from the posterior distribution, kernel density approximation may be used to estimate the posterior ordinate $p(\boldsymbol{\omega}^*|\mathbf{y}, M)$ for low-dimensional $\boldsymbol{\omega}^*$. Alternatively, Chib (1995) and Chib and Jeliazkov (2001) gave efficient algorithms to estimate the posterior ordinate when Gibbs sampling, Metropolis-Hastings or both algorithms are used to generate a sample from the posterior distribution. A number of points pertain specifically to the use of Chib's methods for GLMMs.

A key to using Chib's method is doing efficient blocking of the parameters. For most GLMMs, partitioning $\boldsymbol{\omega}$ into two blocks, one containing the fixed effects parameters and the other containing the variance parameters is convenient. For a few simple GLMMs, Gibbs sampling can be used to generate from the posterior distribution and the marginal density evaluated using the approach of Chib (1995). However, some Metropolis steps are generally required with GLMMs, necessitating the use of the Chib and Jeliazkov (2001) algorithm.

For increased efficiency of estimation, $\boldsymbol{\omega}^*$ in (3.2) is generally taken to be a high density point in the support of the posterior distribution. Popular choices of $\boldsymbol{\omega}^*$ include the posterior mean or posterior median. For GLMMs, the posterior distribution of the variance parameter(s) is skewed; hence the posterior mode will probably be a better choice of $\boldsymbol{\omega}^*$.

3.2.3 Bridge Sampling

Meng and Wong (1996) described the use of bridge sampling for computing the ratio of normalizing constants when: (1) there are two densities each known up to a normalizing constant; and (2) we have draws available from each of the two densities. Though bridge sampling can sometimes be used to directly compute the BF (as a ratio of two normalizing constants), it may be difficult to do so when the two models contain parameters that are not directly comparable or of different dimension. Instead, we use bridge sampling to compute the normalizing constant for a single density by choosing a convenient second density (with known normalizing constant one). Let $r = p(\mathbf{y}|M)$ be the normalizing constant for the posterior distribution under model M (which is the marginal density or marginal likelihood needed for our Bayes factor calculations). Bridge sampling is based on the identity

$$r = \frac{\int q(\boldsymbol{\omega}|\mathbf{y}, M)\alpha(\boldsymbol{\omega})d(\boldsymbol{\omega})d\boldsymbol{\omega}}{\int \alpha(\boldsymbol{\omega})d(\boldsymbol{\omega})p(\boldsymbol{\omega}|\mathbf{y}, M)d\boldsymbol{\omega}} \tag{3.3}$$

for any probability density $d(.)$ and any function $\alpha(.)$ satisfying fairly general conditions given by Meng and Wong (1996). Then if we have a sample $\boldsymbol{\omega}_i, i = 1, 2, \ldots, n_1$ from the posterior distribution and a sample $\tilde{\boldsymbol{\omega}}_j, j = 1, 2, \ldots, n_2$ from d, the Meng-Wong bridge estimator of r is

$$r \approx \frac{\frac{1}{n_2}\sum_{j=1}^{n_2} q(\tilde{\boldsymbol{\omega}}_j|\mathbf{y}, M)\alpha(\tilde{\boldsymbol{\omega}}_j)}{\frac{1}{n_1}\sum_{j=1}^{n_1} d(\boldsymbol{\omega}_j)\alpha(\boldsymbol{\omega}_j)} \, .$$

Meng and Wong showed that if the draws are independent, the optimal choice of α (for minimizing the asymptotic variance of the logarithm of the estimate of r) for a given d is proportional to $\{n_1 q(\boldsymbol{\omega}|\mathbf{y}, M) + n_2 r d(\boldsymbol{\omega})\}^{-1}$, which depends on the unknown r. They propose an iterative sequence to estimate r

$$r^{(t+1)} = \frac{1}{n_2}\sum_{j=1}^{n_2} \frac{l_{2j}}{s_1 l_{2j} + s_2 r^{(t)}} \Bigg/ \frac{1}{n_1}\sum_{j=1}^{n_1} \frac{1}{s_1 l_{1j} + s_2 r^{(t)}}, \tag{3.4}$$

where $s_k = \frac{n_k}{n_1+n_2}, k = 1, 2,\ l_{1j} = q(\boldsymbol{\omega}_j|\mathbf{y}, M)/d(\boldsymbol{\omega}_j)$, and $l_{2j} = q(\tilde{\boldsymbol{\omega}}_j|\mathbf{y}, M)/d(\tilde{\boldsymbol{\omega}}_j)$. Starting with $1/r^{(0)} = 0$ yields $r^{(1)} = \frac{1}{n_2}\sum_{j=1}^{n_2} \frac{q(\tilde{\boldsymbol{\omega}}_j|\mathbf{y},M)}{d(\tilde{\boldsymbol{\omega}}_j)}$, which is the importance sampling estimate with importance density $d(.)$. Also, starting the sequence with $r^{(0)} = 0$ results in the reciprocal importance sampling (DiCiccio et al. 1997) estimate after the first iteration. Meng and Schilling (2003) also showed how Chib's method can be derived using bridge sampling.

The choice of α given above is no longer optimal if the draws are not independent (they are not generally independent for MCMC). Meng and Schilling (2003) recommended adjusting the definition of α in that situation by using in it the effective sample sizes, $\tilde{n}_i = n_i(1 - \rho_i)/(1 + \rho_i)$, where ρ_i is an estimate of the first-order autocorrelation for the draws from the posterior or d, respectively.

We apply bridge sampling with $d \equiv \mathcal{N}(\mathbf{0}, \mathbf{I})$ or $d \equiv t(\mathbf{0}, \mathbf{I})$ to obtain the marginal likelihoods. Before doing so, however, we make use of a relatively new idea, warp bridge sampling (Meng and Schilling 2003). Meng and Schilling (2003) showed that applying (3.3) after transforming the posterior density to match the first few moments of an appropriate d (e.g., a $\mathcal{N}(\mathbf{0}, \mathbf{I})$ density), a transformation which does not change the normalizing constant, results in a more precise estimate of the marginal density. They refer to their transformation as "warp"-ing the density. A *Warp-I transformation* would shift the posterior density to have location zero; this was proposed by Voter (1985) in physics. Matching the mean and variance (or related quantities like the mode and curvature), that is, applying (3.3) with $|\mathbf{S}|\ q(\boldsymbol{\mu} - \mathbf{S}\boldsymbol{\omega}|\mathbf{y}, M)$ in place of $q(\boldsymbol{\omega}|\mathbf{y}, M)$ for suitable choices of $\boldsymbol{\mu}$ and $\mathbf{S}$, is a *Warp-II transformation*. Matching the mean (mode), variance (curvature) and skewness, that is, applying (3.3) with $\frac{|\mathbf{S}|}{2}\ [q(\boldsymbol{\mu} - \mathbf{S}\boldsymbol{\omega}|\mathbf{y}, M) + q(\boldsymbol{\mu} + \mathbf{S}\boldsymbol{\omega}|\mathbf{y}, M)]$ in place of $q(\boldsymbol{\omega}|\mathbf{y}, M)$ is a *Warp-III transformation*.

Application of warp bridge sampling does not require more computational effort than ordinary bridge sampling. We only need draws from the posterior density and the importance density. For Warp-II, (3.4) is applied with l_{1j} and l_{2j} replaced by

$$\tilde{l}_{1j} = |\mathbf{S}|q(\boldsymbol{\omega}_j|\mathbf{y}, M)/d(\mathbf{S}^{-1}(\boldsymbol{\omega}_j - \boldsymbol{\mu})), \quad \text{and} \quad \tilde{l}_{2j} = |\mathbf{S}|q(\boldsymbol{\mu} + \mathbf{S}\tilde{\boldsymbol{\omega}}_j|\mathbf{y}, M)/d(\tilde{\boldsymbol{\omega}}_j).$$

For Warp-III, the corresponding expressions are:

$$\tilde{l}_{1j} = \frac{|\mathbf{S}|}{2}[q(\boldsymbol{\omega}_j|\mathbf{y}, M) + q(2\boldsymbol{\mu} - \boldsymbol{\omega}_j|\mathbf{y}, M)]/d(\mathbf{S}^{-1}(\boldsymbol{\omega}_j - \boldsymbol{\mu})),$$

and

$$\tilde{l}_{2j} = \frac{|\mathbf{S}|}{2}[q(\boldsymbol{\mu} - \mathbf{S}\tilde{\boldsymbol{\omega}}_j|\mathbf{y}, M) + q(\boldsymbol{\mu} + \mathbf{S}\tilde{\boldsymbol{\omega}}_j|\mathbf{y}, M)]/d(\tilde{\boldsymbol{\omega}}_j)\cdot$$

Meng and Schilling (2003) suggested analytical and empirical ways of finding optimal values of $\boldsymbol{\mu}$ and $\mathbf{S}$ to use in warp bridge sampling. For example, for the Warp-III transformation, optimal values can be found by maximizing over $\boldsymbol{\mu}$ and $\mathbf{S}$ the quantity

$$\sum_j \sqrt{\frac{1}{\phi(\boldsymbol{\omega}_j)}|\mathbf{S}| \left(q(\boldsymbol{\mu} - \mathbf{S}\boldsymbol{\omega}_j|\mathbf{y}, M) + q(\boldsymbol{\mu} + \mathbf{S}\boldsymbol{\omega}_j|\mathbf{y}, M)\right)}$$

for a sample $\boldsymbol{\omega}_j, j = 1, 2, \ldots n$ from a $\mathcal{N}(\mathbf{0}, \mathbf{I})$ distribution, where $\phi(.)$ denotes the multivariate normal density. Empirical studies suggest good estimates of r even for suboptimal choices of the warping transformation (e.g., using sample moments rather than optimizing over $\boldsymbol{\mu}$ and $\mathbf{S}$).

Note that it is also possible to use the idea of warping transformations to develop importance sampling methods. In other words, one can choose the importance sampling density as $\mathcal{N}(\mathbf{0}, \mathbf{I})$ and then transform the posterior distribution (with Warp-II or Warp-III transformations) before applying importance sampling.

3.2.4 Reversible Jump MCMC

A very different approach for computing Bayes factor estimates requires constructing an "extended" model in which the model index is a parameter as well. The reversible jump MCMC method suggested by Green (1995) samples from the expanded posterior distribution. This method generates a Markov chain that can move between models with parameter spaces of different dimensions. Let π_j be the prior probability on model j, $j = 0, 1, \ldots J$. The method proceeds as follows:

1. Let the current state of the chain be $(j, \boldsymbol{\omega}_j)$, $\boldsymbol{\omega}_j$=n_j-dimensional parameter for model j.
2. Propose a new model j' with probability $h(j, j')$, where $\sum_{j'} h(j, j') = 1$.
3. a. If $j' = j$, then perform an MCMC iteration within model j. Go to Step 1.
   b. If $j' \neq j$, then generate $\mathbf{u}$ from a proposal density $q_{jj'}(\mathbf{u}|\boldsymbol{\omega}_j, j, j')$ and set $(\boldsymbol{\omega}_{j'}, \mathbf{u}') = g_{j,j'}(\boldsymbol{\omega}_j, \mathbf{u})$, where g is a 1-1 onto function, $n_j + \dim(\mathbf{u}) = n_{j'} + \dim(\mathbf{u}')$.

4. Accept the move from j to j' with probability

$$\min\left\{1, \frac{p(\mathbf{y}|\boldsymbol{\omega}_{j'}, M = j')p(\boldsymbol{\omega}_{j'}|M = j')\pi_{j'}h(j', j)q_{j'j}(\mathbf{u}'|\boldsymbol{\omega}_{j'}, j', j)}{p(\mathbf{y}|\boldsymbol{\omega}_j, M = j)p(\boldsymbol{\omega}_j|M = j)\pi_j h(j, j')q_{jj'}(\mathbf{u}|\boldsymbol{\omega}_j, j, j')} \cdot \left|\frac{\partial g(\boldsymbol{\omega}_j, \mathbf{u})}{\partial(\boldsymbol{\omega}_j, \mathbf{u})}\right|\right\}.$$

Step 3b is known as the dimension-matching step in which auxiliary random variables are introduced (as needed) to equate the dimension of the parameter space for models j and j'. If the above Markov chain runs sufficiently long, then $p(M_j|\mathbf{y})/p(M_j'|\mathbf{y}) \approx N_j/N_{j'}$, where N_j is the number of times the Markov chain reaches model j. Therefore, the Bayes factor $\mathrm{BF}^{jj'}$ for comparing models j and j' can be estimated using (3.1) as $\mathrm{BF}^{jj'} \approx \frac{N_j}{N_{j'}} \Big/ \frac{\pi_j}{\pi_{j'}}$.

3.2.5 Other Methods

The methods described here do not exhaust all possible methods. Although our goal is to try and provide general advice, the best approach for any specific application may be found outside our list. The methods summarized in this work represent the set that we have found most applicable to GLMMs. We have omitted Laplace's method (see, e.g., Tierney and Kadane 1986), a useful approximation in many problems, but increasingly unreliable as the number of random effects parameters increase in the GLMMs (Sinharay 2001). The approach of Verdinelli and Wasserman (1995) for nested models works well in our first example but requires density estimation and thus is less practical in higher dimensional settings like our second example (Sinharay and Stern 2003; Sinharay 2001). Other methods for computing Bayes factors which can be used in the context of GLMMs include the ratio importance sampling approach by Chen and Shao (1997), path sampling (Gelman and Meng 1998), and product space search (Carlin and Chib 1995). Han and Carlin (2001) find methods based on the idea of creating a product space (that is a space that encompasses the parameter space of each model under consideration) problematic for models where random effects cannot be analytically integrated out from the likelihood, which is typically the case with the GLMMs.

3.3 PARAMETERIZATION

A number of the methods for estimating Bayes factors require computing the GLMM likelihood $p(\mathbf{y}|\boldsymbol{\omega}, M)$ for one or more values of $\boldsymbol{\omega}$. If the accurate computation of $p(\mathbf{y}|\boldsymbol{\omega}, M)$, which involves integrating out the random effects, is time-consuming, some of the methods become impractical. This especially affects those like importance sampling and bridge sampling that require more than one marginal likelihood computation. Chib's method is more likely to succeed in such cases.

One approach for circumventing this difficulty is to consider applying our various approaches with $\boldsymbol{\omega} = (\mathbf{b}, \boldsymbol{\alpha}, \boldsymbol{\theta})$ rather than assuming that $\mathbf{b}$ has been integrated out. Chib (1995) suggested this idea in the context of his approach and it applies more generally to the other methods considered here. For this choice of $\boldsymbol{\omega}$ the marginal density $p(\mathbf{y}|M)$ for the GLMM is

$$\begin{aligned} p(\mathbf{y}) &= \int\int p(\mathbf{y}|\boldsymbol{\alpha}, \boldsymbol{\theta}) p(\boldsymbol{\alpha}, \boldsymbol{\theta}) d\boldsymbol{\alpha} d\boldsymbol{\theta} \\ &= \int\int\int p(\mathbf{y}|\boldsymbol{\alpha}, \mathbf{b}) p(\mathbf{b}|\boldsymbol{\theta}) p(\boldsymbol{\alpha}, \boldsymbol{\theta}) d\mathbf{b} d\boldsymbol{\alpha} d\boldsymbol{\theta}. \end{aligned}$$

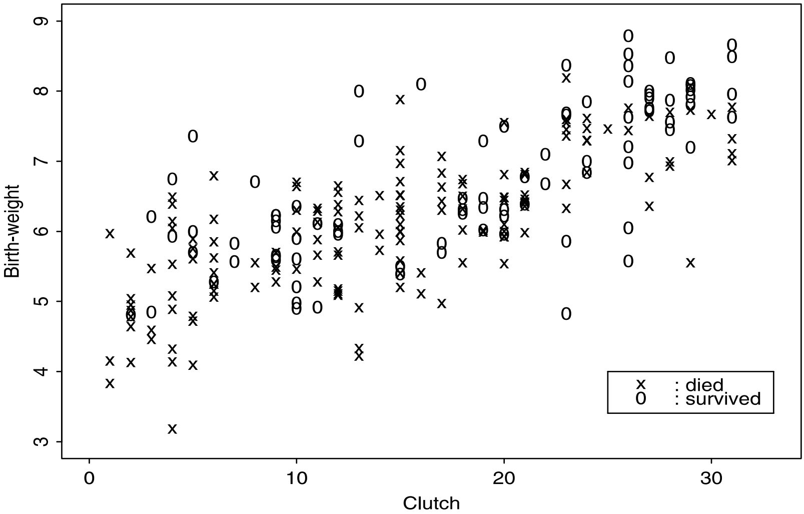

Figure 1. Scatterplot with the clutches sorted by average birthweight.

The advantage of the expanded definition of $\boldsymbol{\omega}$ is that the computation of the likelihood function $p(\mathbf{y}|\boldsymbol{\omega}) = p(\mathbf{y}|\boldsymbol{\alpha}, \mathbf{b}, \boldsymbol{\theta}) = p(\mathbf{y}|\boldsymbol{\alpha}, \mathbf{b})$ is straightforward. However, as a price to pay for this simplification of the likelihood, the dimension of the parameter space increases by the number of components in **b**, which is usually high, even for simple GLMMs.

4. EXAMPLE: A NATURAL SELECTION STUDY

4.1 THE DATA AND THE MODEL FITTED

A study of survival among turtles (Janzen et al. 2000) provides an example where a GLMM is appropriate. The data consist of information about the clutch (family) membership, survival and birth-weight of 244 newborn turtles. The scientific objectives are to assess the effect of birth-weight on survival and to determine whether there is any clutch effect on survival. Figure 1 shows a scatterplot of the birthweights versus clutch number with survival status indicated by the plotting character "0" if the animal survived and "x" if the animal died. The clutches are numbered in increasing order of the average birthweight of the turtles in the clutch. The figure suggests that the heaviest turtles tend to survive and the lightest ones tend to die. Some variability in the survival rates across clutches is evident from the figure.

Let y_{ij} denote the response (indicator of survival) and x_{ij} the birthweight of the jth turtle in the ith clutch, $i = 1, 2 \ldots m = 31,\ j = 1, 2, \ldots n_i$. The probit regression model with random effects fit to the dataset is given by:

- $y_{ij}|p_{ij} \sim \text{Ber}(p_{ij})$, where $p_{ij} = \Phi(\alpha_0 + \alpha_1 x_{ij} + b_i),\ i = 1, 2 \ldots m = 31,\ j = 1, 2, \ldots n_i$;
- $b_i|\sigma^2 \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2),\ i = 1, 2, \ldots, m.$

The b_i's are random effects for clutch (family). There is no clutch effect iff $\sigma^2 = 0$.

4.2 Estimating the Bayes Factor

The marginal likelihoods under the null model (M_0) and full model (M_1) are

$$p(\mathbf{y}|M_0) = \int p(\mathbf{y}|\boldsymbol{\alpha}, \mathbf{b} = \mathbf{0}) p(\boldsymbol{\alpha}) d\boldsymbol{\alpha},$$

and

$$p(\mathbf{y}|M_1) = \int p(\mathbf{y}|\boldsymbol{\alpha}, \mathbf{b}) p(\mathbf{b}|\sigma^2) p(\boldsymbol{\alpha}) p(\sigma^2) d\mathbf{b} d\boldsymbol{\alpha} d\sigma^2.$$

Our work uses the shrinkage prior distribution (see, e.g., Daniels 1999) for the variance components, $p(\sigma^2) = \frac{c}{(c+\sigma^2)^2}$, where c is the median of $p(\sigma^2)$. We fix c at 1. We set $p(\boldsymbol{\alpha}) = N_2(\mathbf{0}, 10\,\mathbf{I})$.

Each of the methods (except reversible jump MCMC) requires that we evaluate the likelihood $p(\mathbf{y}|\boldsymbol{\alpha}, \sigma^2)$ (i.e., we must integrate out $\mathbf{b}$ in the full model). For this relatively small problem we do so using Simpson's rule to perform the needed numerical integration. The methods also require samples from the posterior distribution, $p(\boldsymbol{\alpha}, \sigma^2|\mathbf{y})$. This is done using an MCMC algorithm. There is significant autocorrelation in the Markov chain and the Gelman-Rubin convergence diagnostic (see, e.g., Gelman et al. 2003) suggests that using five chains of 1,000 iterations after burn-ins of 200 iterations is sufficient to provide a summary of the posterior distribution. This results in a final posterior sample size of 5,000. Therefore, all simulation-based estimates use posterior samples of size 5,000 for each model.

The posterior mode and the information matrix at the mode, which are used in some of the methods, are found using the Newton-Raphson method. The posterior mean and variance, required by some methods, are computed from a preliminary MCMC run. For importance sampling and bridge sampling, we transform the variance to $\log(\sigma^2)$ in the posterior distribution to improve the similarity of the normal (or t) importance sampling density. Additional details concerning the implementation of the individual approaches follow.

4.2.1 Chib's Method

The conditional distribution of σ^2 is not of known form, necessitating the use of the Metropolis algorithm for sampling from the joint posterior distribution of the parameters under the mixed model. The use of data augmentation by incorporating latent normal variables in the probit model (Albert and Chib 1993) allows the use of Gibbs sampling for the fixed effects, but did not improve the precision here. The Metropolis algorithm is also used to generate a sample from the null model.

4.2.2 Importance and Bridge Sampling

We tried several variations of bridge and importance sampling. We apply Warp-II transformations with (1) posterior mode and square root of the curvature matrix as $\boldsymbol{\mu}$ and $\mathbf{S}$, respectively, (mode-curvature matching) and (2) posterior mean and a square root of the posterior variance matrix as $\boldsymbol{\mu}$ and $\mathbf{S}$ (mean-variance matching). The use of $d \equiv t_4(\mathbf{0}, \mathbf{I})$ resulted in more precise Bayes factor estimates than $d \equiv \mathcal{N}(\mathbf{0}, \mathbf{I})$, probably because the former has more overlap than the latter with the skewed transformed posterior. Thus, we report only the results for the t importance sampling or bridge sampling density. It typically took three or four iterations for the bridge sampling iterative process to reach convergence. Recall that the first iteration of bridge sampling with sampling distribution d yields the importance sampling estimate that corresponds to using d as importance sampling density for the transformed posterior.

We also apply Warp-III transformations with mode-variance matching, mean-variance matching, and mode-curvature matching. In this case $d \equiv t_4(\mathbf{0}, \mathbf{I})$ does not result in improved estimates relative to $d \equiv \mathcal{N}(\mathbf{0}, \mathbf{I})$, probably because after accounting for skewness the posterior distribution has been transformed into a distribution well-approximated by a normal distribution.

Because there exists a significant autocorrelation in the posterior sample (e.g., first-order autocorrelations are in the range .8–.9 for most parameters in the alternative model), we adjust α using the effective sample size approach for both Warp-II and Warp-III, taking $\hat{\rho}$ to be the sample lag-1 autocorrelation of $1/(\tilde{l}_{1j} + r)$, as in Meng and Schilling (2003). This adjustment, requiring only a few lines of additional computer coding, results in a significant increase in precision.

4.2.3 Reversible Jump MCMC

Using the notation from Section 3.2.4, we set $h(j, j') = \pi_j = .5 \, \forall \, j, j'$. Here, parameter vectors under the two models are $\boldsymbol{\omega}_0 = \boldsymbol{\alpha}$ and $\boldsymbol{\omega}_1 = (\boldsymbol{\alpha}, \sigma^2)$. When we try to move from model 0 to model 1, there is an increase in dimension as model 1 has a variance component while model 0 does not. We use the current value of $\boldsymbol{\alpha}$ in model 0 as candidate value of $\boldsymbol{\alpha}$ in model 1, and generate a candidate value of σ^2 from an inverse gamma distribution with mean and variance as the posterior mode and curvature of σ^2 under model 1. These choices amount to, using notations from Section 3.2.4, $u = \sigma^2$, $u' = 0$, $g(\mathbf{x}) = \mathbf{x}$ and $q(\sigma^2) \equiv$ the above-mentioned inverse gamma distribution. When we try to move from model 1 to model 0, there is a reduction in dimension. Therefore, in generating candidate values under model 0, we ignore the variance component for model 1 and use the current value of $\boldsymbol{\alpha}$ in model 1 as candidate value of $\boldsymbol{\alpha}$ under model 0. These amount to $u = 0$, $u' = \sigma^2$ and $g(\mathbf{x}) = \mathbf{x}$. To move within a model, we take a Metropolis step with a random walk proposal distribution.

4.3 RESULTS

Numerical integration over all parameters provides us the true value of the Bayes factor of interest, although the program takes about 62 hours of CPU time to run on an Alpha station 500 workstation equipped with 400MHz 64-bit CPU and a gigabyte of RAM. The true value of the Bayes factor up to three decimal places is 1.273.

To learn about the precision of the different estimation approaches, we compute the BF

Table 1. Estimates of the Bayes factor (along with their standard deviations and time taken to run the program) for the Turtles Dataset

Method	*Bayes factor estimate*	*Std. dev.*	*CPU time (min)*
True value	1.273	–	–
Bridge Warp-III (optimal μ-curvature)	1.273	.0059	20.3
Bridge Warp-III (mode-curvature)	1.273	.0063	20.3
Bridge Warp-III (mode-variance)	1.273	.0076	20.3
Imp. samp. (Warp-III, optimal μ-curvature)	1.275	.0066	8.0
Imp. samp. (Warp-III, mode-curvature)	1.274	.0068	8.0
Imp. samp. (Warp-III, mode-variance)	1.275	.0086	8.0
Bridge Warp-II (mean-variance)	1.273	.0094	13.1
Bridge Warp-II (mode-curvature)	1.274	.0145	13.2
Imp. samp. Warp-II (mean-variance)	1.272	.0108	6.0
Imp. samp. Warp-II (mode-curvature)	1.274	.0156	6.2
Chib's (at mode)	1.275	.0492	16.1
Chib's (at mean)	1.266	.0724	16.0
RJ MCMC	1.302	.2032	18.0

a number of times with different random seeds. Table 1 summarizes the mean and standard deviation (sd) of 100 Bayes factor estimates obtained by the various methods. The use of 100 trials allows us to have confidence in the reported standard deviations (according to traditional sampling theory the reported standard deviations are accurate to within 15%). Also shown in the table are the CPU times required for one computation of the Bayes factor estimate by each of the methods on the above-mentioned workstation.

Table 1 indicates that all of the methods are effective, with reversible jump Markov chain Monte Carlo having a much larger standard deviation. We next discuss the results for importance and bridge sampling. The Warp-III transformation results in very efficient estimates (for both bridge sampling and importance sampling), even without the optimal choice of μ and S. Bridge sampling with Warp-III (using the effective sample size in computation of $\alpha(.)$) is more efficient for this example than Warp-III importance sampling. Running bridge sampling to convergence reduces the standard error by about 10% relative to stopping after the first iteration (which is importance sampling). Warp-II bridge sampling and Warp-II importance sampling perform respectably as well, especially for mean-variance matching. Again, bridge sampling seems to be more precise than importance sampling. Figure 2, showing contour plots for the two-dimensional marginal posterior distributions (obtained using S-Plus functions “kde” and “contour”), demonstrates the effect of warping on the posterior distribution. The application of the Warp-III transformation takes the originally nonnormal two-dimensional posteriors (the top row) into ones quite close to $\mathcal{N}(\mathbf{0}, \mathbf{I})$ (the bottom row). Therefore, it is no surprise that Warp-III provides such precise estimates.

Chib's method provides good results as well. The standard deviation is larger than for the bridge and warp sampling. Chib's method does, however, have one advantage over importance and bridge sampling in that it does not require that a matching or importance

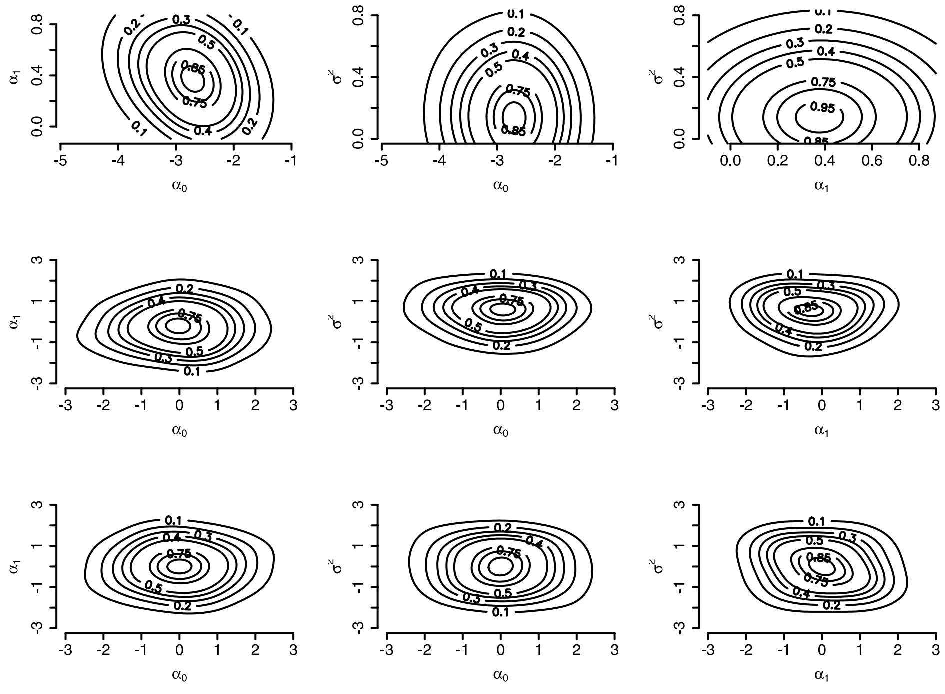

Figure 2. Effect of warping on posterior distributions for the turtles data. The first row shows the three bivariate marginal posterior distributions for the original untransformed posterior distribution (showing the highly non-normal nature). The second row shows the effect of the Warp-II (mean-variance matching) transformation after taking the logarithm of σ^2. The third row is for the Warp-III transformation with mode-curvature matching.

sampling density be selected. If the posterior distribution has features, like an unusually long tail, not addressed by our warping transformations, then it is possible that the standard deviation of importance and bridge sampling may be underestimated.

A second factor in comparing the computational methods is the amount of computational time required. This has two dimensions, the amount of time required to run the program, and the time required to write the program. The relative importance of these two dimensions depends on a user's context—if one will frequently analyze data using the same model, then programming time is less important.

Programming time of course depends on the programmer. Our experience was that importance sampling (with transformations) takes considerably less time than bridge sampling to program. Chib's method builds on the existing MCMC code (required by all methods); however, to us, modifying it to compute the Bayes factor was more time consuming than developing importance sampling methods. The programming time for Chib's method is comparable to that for warp bridge sampling.

Naturally, the run time for importance sampling is less than that of bridge sampling. In the present case the added precision of bridge sampling may not be worth the extra time. Importance sampling was also faster than Chib's method but there are several mitigating factors that affect that comparison. Importance (and bridge) sampling requires extensive

Table 2. Part of the Scotland Lip Cancer Dataset

County	*y*	*p (in '000)*	*AFF*	*E*	*Neighbors*
1	9	28	16	1.38	4 5 9 11 19
2	39	231	16	8.66	2 7 10
3	11	83	10	3.04	2 6 12
4	9	52	24	2.53	3 18 20 28
⋮	⋮	⋮	⋮	⋮	⋮
54	1	247	1	7.03	5 34 38 49 51 52
55	0	103	16	4.16	5 18 20 24 27 56
56	0	39	10	1.76	6 18 24 30 33 45 55

calculations of the likelihood function; in problems where that is more complicated, the time advantage of importance sampling will dissipate. Also, more efficient MCMC algorithms might make Chib's method more competitive on speed. Although all methods work well here, our preference is for warp-transformed importance sampling for these data based on speed and efficiency.

5. EXAMPLE: SCOTLAND LIP CANCER DATA

This section considers a more complex example with more than one variance component. The computations become much more difficult and time-consuming for such models.

5.1 Description of the Dataset

Table 2 shows a part of a frequently analyzed dataset (see, e.g., Clayton and Kaldor 1987) regarding lip cancer data in the 56 administrative districts in Scotland from 1975–1980. The objective of the original study was to find any pattern of regional variation in the disease incidence of lip cancer. The dataset contains $\{y_i, p_i, E_i, \text{AFF}_i, N_i\}, i = 1, 2, \ldots 56$, where, for district i, y_i is the observed number of lip cancer cases among males from 1975–1980, p_i is the population at risk of lip cancer (in thousands), E_i is the expected number of cases adjusted for the age distribution, AFF_i is the percent of people employed in agriculture, forestry, and fishing (these people working outdoors may be under greater risk of the disease because of increased exposure to sunlight), and N_i is the set of neighboring districts.

5.2 A Poisson-Gaussian Hierarchical Model

The y_i's are assumed to follow independent Poisson distributions, $y_i|\lambda_i \sim \text{Poisson}(\lambda_i E_i), i = 1, 2, \ldots, n$, where λ_i is a relative risk parameter describing risk after adjusting for the factors used to calculate E_i. As in Besag, York, and Mollie (1991), we use a mixed linear model for log($\boldsymbol{\lambda}$), $\log(\boldsymbol{\lambda}) = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\eta} + \boldsymbol{\psi}$, where $\mathbf{X}$ is the covariate matrix; $\boldsymbol{\beta} = (\beta_0, \beta_1)'$ is a vector of fixed effects; $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_n)'$ is a vector of spatially correlated random effects; and $\boldsymbol{\psi} = (\psi_1, \psi_2, \ldots, \psi_n)'$ is a vector of uncorrelated heterogeneity random effects.

For known matrices **C** and diagonal **M**, we take the prior distribution for $\boldsymbol{\eta}$ as a Gaussian conditional autoregressive (CAR) distribution, that is, $\boldsymbol{\eta}|\tau^2, \phi \sim \mathcal{N}(\mathbf{0}, \tau^2(I-\phi\mathbf{C})^{-1}\mathbf{M})$, as in Cressie, Stern, and Wright (2000), where τ^2 and ϕ are parameters of the prior distribution. The parameter ϕ measures the strength of spatial dependence, $0 < \phi < \phi_{\text{max}}$, where $\phi = 0$ implies no spatial association and ϕ_{max} is determined by the choice of **C** and **M**. The elements of the matrices **C** and **M** used here are $m_{ii} = E_i^{-1}$, and $c_{ij} = \left(\frac{E_j}{E_i}\right)^{\frac{1}{2}} I_{[j\in N_i]}$, where $I_{[A]}$ is the indicator for event A. For these choices of **C** and **M**, $\phi_{\text{max}} = .1752$.

The ψ's are modeled as $\boldsymbol{\psi}|\sigma^2 \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{D})$, where **D** is a diagonal matrix with $d_{ii} = E_i^{-1}$, and σ^2 is a variance parameter. In practice, it appears often to be the case that either $\boldsymbol{\eta}$ or $\boldsymbol{\psi}$ dominates the other, but which one will not usually be known in advance (Besag et al. 1991).

The model above contains three covariance matrix parameters (τ^2, σ^2, and ϕ) and 112 random effects parameters, making it a more challenging data set to handle computationally than the turtle dataset. The joint maximum likelihood estimate of $\boldsymbol{\xi} = (\beta_0, \beta_1, \phi, \tau^2, \sigma^2)'$ is $\hat{\boldsymbol{\xi}}_{\text{MLE}} = (-.489, .059, .167, 1.640, 0)'$. The fifth component of the posterior mode is 0 as well.

5.3 ESTIMATING THE BAYES FACTORS

Because of the presence of more than one variance component in the model, several Bayes factors are of interest. These correspond to comparing any two of the four possible models:

- "full model" with σ^2 and τ^2 (and ϕ)
- "spatial model" with τ^2 only as a variance component (also includes ϕ)
- "heterogeneity model" with σ^2 only as a variance component
- "null model" with no variance component.

We focus on the three Bayes factors obtained by comparing any one of the three reduced models to the full model. Any other Bayes factor of interest here can be obtained from these three.

Using a transformation $\boldsymbol{\nu} = \boldsymbol{\eta} + \boldsymbol{\psi}$, the likelihood for the full model (with random effects integrated out), $L(\boldsymbol{\beta}, \phi, \tau^2, \sigma^2|\mathbf{y})$, can be expressed as

$$L(\boldsymbol{\beta}, \phi, \tau^2, \sigma^2|\mathbf{y}) \propto \int \left\{ \prod_{i=1}^{n} \exp\left(-E_i e^{\mathbf{x}_i'\boldsymbol{\beta}+\nu_i}\right) e^{y_i(\mathbf{x}_i'\boldsymbol{\beta}+\nu_i)} \right\} \frac{1}{|\mathbf{V}|^{1/2}} \cdot \exp\left\{-\frac{1}{2}\boldsymbol{\nu}'\mathbf{V}^{-1}\boldsymbol{\nu}\right\} d\boldsymbol{\nu},$$

where $\mathbf{V} = \tau^2(I - \phi\mathbf{C})^{-1}\mathbf{M} + \sigma^2\mathbf{D}$. To estimate the likelihood, we use importance sampling (Section 3.2.1) with a t_4 importance sampling distribution. Note that the Warp bridge sampling method could be used here as well. The mean and variance of the importance sampling distribution are the corresponding moments of $\boldsymbol{\nu} = \boldsymbol{\eta} + \boldsymbol{\psi}$ computed from a posterior sample drawn from the conditional posterior distribution of $(\boldsymbol{\eta}, \boldsymbol{\psi})$ given $\boldsymbol{\beta}, \phi, \tau^2, \sigma^2$.

This procedure estimates the likelihood with reasonable precision within reasonable time.

We assume independent prior distributions, $\boldsymbol{\beta} \sim N_2(\mathbf{0},\ 20\mathbf{I})$, $p(\phi) = \text{Uniform}(0, \phi_{\text{max}})$, $p(\sigma^2) = \frac{1}{(1+\sigma^2)^2}$, $p(\tau^2) = \frac{1}{(1+\tau^2)^2}$. Posterior samples are obtained using a Metropolis-Hastings algorithm. The large number of parameters result in high autocorrelations and hence, we use 5 chains of 10,000 iterations of the MCMC after a burn-in of 2,000 each (enough to achieve convergence according to the Gelman-Rubin convergence criterion). Additional details about specific methods follow.

5.3.1 Chib's Method

Each of the conditional posterior distributions is sampled from using a Metropolis step and thus the Chib and Jeliazkov (2001) approach is required. As for the fixed point at which the posterior density is evaluated, we use the sample mean of the posterior sample rather than the posterior mode because the latter is on the boundary of the parameter space and one of the terms required by Chib's method is not defined there.

5.3.2 Bridge and Importance Sampling

We apply Warp-II transformations with $d \equiv t_4(\mathbf{0},\ \mathbf{I})$ and Warp-III transformations with $d \equiv \mathcal{N}(\mathbf{0},\ \mathbf{I})$. In both cases we use the sample mean and variance of the posterior sample as the location and scale parameters of the transformation. It may be possible to do better by optimizing over $\boldsymbol{\mu}$ and $\mathbf{S}$ but the efficiency achieved by the mean-variance choice was sufficient. As in the turtle example, because of strong dependence of the draws, we use the effective sample size (rather than the actual MCMC sample size) in the bridge sampling iteration.

5.4 REVERSIBLE JUMP MCMC

It is possible in principle to compute all the Bayes factors from one reversible jump MCMC that allows jumps among the four models. We use three separate programs to compute the three Bayes factors (and even then had trouble getting this approach to work well). We set $h(j, j') = \pi_j = .5\ \forall\ j, j'$. When we try to move from model j to model j', we generate auxiliary variable u to correspond to all of the parameters of model j', that is we do not retain values of the parameters from model j that are in model j' as well. This was the approach of Han and Carlin (2001) as well. For example, when we try to move from the "full model" to the "spatial model," we generate u from a 59-dimensional normal independence proposal density, whose mean and variance are determined by an earlier run of the MCMC algorithm for the "spatial model." The reasoning is that it would not be appropriate to just zero out the heterogeneity random effects because the remaining spatial effects are not likely to represent the posterior distribution under the spatial model. It was difficult to obtain reliable results from reversible jump MCMC apparently because of the large number of random effects which cannot be integrated out from the likelihood. Han and Carlin (2001) found similar results. As we show below, the Bayes factor for comparing the

Table 3. Estimates of Bayes Factors (along with their standard deviations and time taken to run the program) for the Scotland Lip Cancer Dataset

Comparing	*Method*	*Estimated BF*	*Std. dev.*	*CPU time(min)*
"spatial	True value	**1.42**	–	–
model"	Bridge W-III (mean-variance)	1.42	.029	88.3
vs	Imp. samp. W-III (mean-variance)	1.41	.032	45.2
"full	Bridge W-II (mean-variance)	1.43	.040	65.2
model"	Imp. samp. W-II (mean-variance)	1.44	.065	30.5
	Chib's (at mean)	1.44	.132	81.1
	RJMCMC	1.21	.265	49.9
"heterogen.	True value	**.066**	–	–
model"	Bridge W-III (mean-variance)	.066	.0017	66.3
vs	Imp. samp. W-III (mean-variance)	.066	.0018	32.4
"full	Bridge W-II (mean-variance)	.066	.0020	43.4
model"	Imp. samp. W-II (mean-variance)	.067	.0028	20.6
	Chib's (at mean)	.067	.0086	57.2
	RJMCMC	.032	.182	30.9
"null	True value	$\mathbf{1.15 \times 10^{-23}}$	–	–
model"	Bridge W-III (mean-variance)	1.15×10^{-23}	1.52×10^{-25}	55.4
vs	Imp. samp. W-III (mean-variance)	1.15×10^{-23}	1.66×10^{-25}	28.2
"full	Bridge W-II (mean-variance)	1.14×10^{-23}	2.66×10^{-25}	37.2
model"	Imp. samp. W-II (mean-variance)	1.16×10^{-23}	3.64×10^{-25}	18.2
	Chib's (at mean)	1.21×10^{-23}	1.46×10^{-24}	48.1

"full model" to the "heterogeneity model" or the "spatial model" could not be estimated to a reasonable degree of accuracy, even after trying a variety of Metropolis proposal densities. The Bayes factor favoring the "full" model over the "null" model is so large that we were never able to accept a single step to the null model for any of our proposal densities.

5.5 RESULTS

We use the importance sampling method with sample size one million to compute the "true value" of the three Bayes factors. Examining the variability of the importance ratios for the sampled one million points, we conclude that the Bayes factor is determined up to a standard error of about .5% for the Bayes factor comparing the spatial model to the full model and about .25% for the other two Bayes factors. Warp-III bridge sampling with a sample size of half million results in the same values. These values serve as the true Bayes factors for comparing the methods.

Table 3 shows the average and standard deviation of 100 Bayes factors (with different random seeds) obtained using each method, and the CPU time taken for one computation of the Bayes factor estimate by each of these methods on the workstation mentioned in the previous example. The results here are completely consistent with those of the first example. Warp bridge sampling, importance sampling, and Chib's marginal likelihood approach yield good estimates. The reversible jump MCMC approach performs unsatisfactorily, even after considerable tuning. The standard deviation of the Bayes factor estimate is much smaller for the Warp-III bridge sampling and importance sampling than the other methods. Figure 3

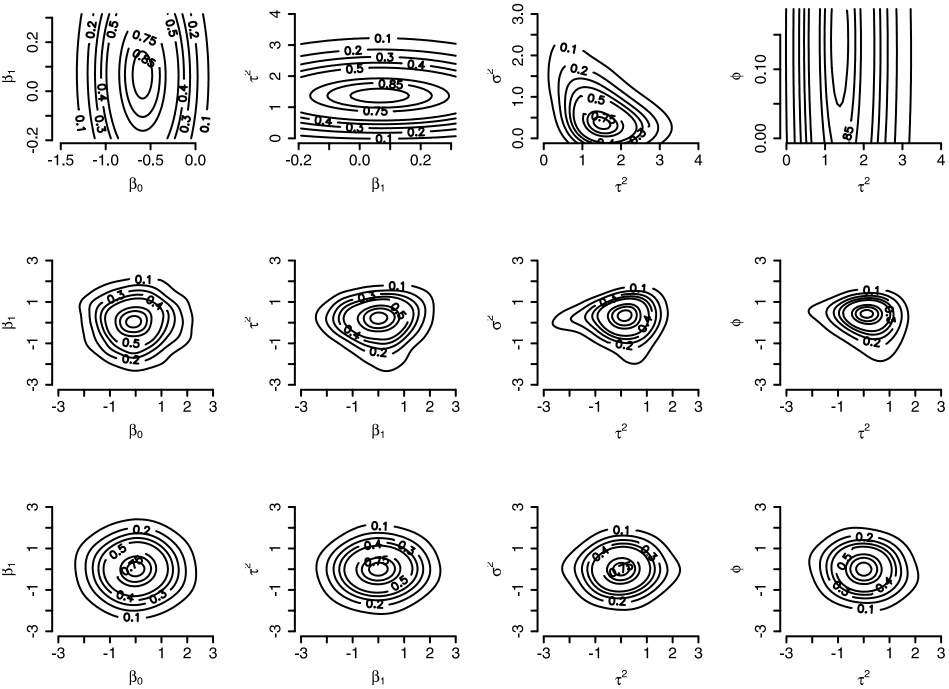

Figure 3. Effect of warping on posterior distributions for the lip cancer data. The first row shows four bivariate marginal posterior distributions for the original untransformed joint posterior distribution (showing the highly non-normal nature). The second row shows the effect of the Warp-II (mean-variance matching) transformation. The third row is for the Warp-III (mean-variance matching) transformation.

shows how the warp transformations work in this higher dimensional problem by examining contour plots for some two-dimensional marginal posterior distributions.

Han and Carlin (2001, p. 1131) suggested that Chib's method might not be effective in spatial models using Markov random field priors. Our results indicate that it does provide acceptable Bayes factor estimates, although as Han and Carlin pointed out one is required to sample the random effects individually which may become prohibitive in larger applications.

6. DISCUSSION AND RECOMMENDATIONS

GLMMs are applied extensively and their use is likely to increase with the widespread availability of fast computational facilities and the increased sophistication of data analysts in a variety of disciplines. In many applications of these models, question arises about the necessity of the variance components in the model. One way to answer the question, in fact our preferred way, is to examine the estimates of the variance components under the full model. This article arose as a result of several problems in which formal model comparisons were desired by scientific investigators. The objective of this study is to learn more about the performance of different methods for computing the relevant Bayes factor estimates.

The computation of the likelihood $p(\mathbf{y}|\boldsymbol{\omega}, M)$ (averaging over the random effects) is a nontrivial task for GLMMs. If the model is simple and the dataset is small, it is possible to apply numerical integration. For larger problems, importance sampling is a possible approach (Geyer and Thompson 1992); our work finds that importance sampling with a t_4 sampling distribution works well.

The computation of the Bayes factor involves integrating over the parameters $\boldsymbol{\omega}$ in $p(\mathbf{y}|\boldsymbol{\omega}, M)$. Typically, for a GLMM, the parameter vector $\boldsymbol{\omega}$ consists of the regression parameters $\boldsymbol{\alpha}$ and the variance component parameters $\boldsymbol{\theta}$. However, in computing Bayes factors for GLMMs, including the random effects in the parameter vector, for example, in the manner suggested in the context of the method of Chib (1995), is often convenient. This approach makes the application of bridge sampling and importance sampling possible for the second example.

Our results indicate that warp bridge sampling (Meng and Schilling 2003), importance sampling (also based on warp transformations), and Chib's (1995) marginal likelihood approach are all effective. In both applications each finds the correct Bayes factor. Reversible jump MCMC is more difficult to apply and did not produce accurate results even after significant effort was applied to create an effective algorithm. This is not necessarily surprising; Gelman and Meng (1998) pointed out that bridge sampling can be viewed as a form of average in place of the accept/reject model transitions that characterize reversible jump MCMC. The averaging provides a kind of "Rao-Blackwellization" (see, e.g., Gelman and Meng 1998) that improves efficiency.

Among the three effective methods the choice for a particular problem depends on tradeoffs among a number of factors. For our two examples importance sampling based on warp transformed distributions was quick, accurate, and had a small standard error. One disadvantage of this approach is that one must ensure somehow that the tails of the importance sampling density are at least as long as the tails of the warp transformed posterior distribution. Warp bridge sampling was accurate and had the lowest standard error of the methods over repeated computations. It required more computational time than importance sampling; the importance sampling estimate is the first iterate in our algorithm for carrying out bridge sampling. Bridge sampling is less sensitive to the choice of the matching density as long as there is good overlap between the matching density and the warp transformed posterior distribution. Chib's method gave accurate results but had the largest standard error among the three methods (though the standard error is still quite small in absolute terms). Chib's method has a couple of compensating advantages in that it does not require the choice of an importance sampling or matching density, and it requires only a single evaluation of the likelihood. For our two examples the repeated evaluations of the likelihood did not make bridge sampling and importance sampling inefficient but it is possible that in larger problems such evaluations would be prohibitive.

The efficiency of Chib's method is closely related to the efficiency of the underlying MCMC algorithm. Therefore, the standard deviation for the Chib's method and its run time may be reduced by reducing the autocorrelation of the generated parameter values in the MCMC algorithm, for example, by the use of a tailored proposal density (Chib and Jeliazkov 2001). Of course, improved MCMC algorithms will also result in improved precision for the bridge sampling and reversible jump MCMC estimates as well.

Perhaps the most noteworthy finding from our two examples is the effectiveness of warp bridge sampling, a method with which some readers may not be familiar. Warp bridge sampling makes use of transformations to match the posterior distribution with a suitably chosen (and simple to sample from) matching density. The choice of matching density is less critical than with importance sampling. Warp bridge sampling should work as long as the transformed posterior and matching density overlap.

ACKNOWLEDGMENTS

This work was partially supported by National Institutes of Health award CA78169. The authors thank Xiaoli Meng profusely for his invaluable advice resulting in significant improvement of the article. The authors also thank Frederic Janzen for providing the data from the natural selection study and Michael Daniels, Howard Wainer, Siddhartha Chib, the three referees, associate editor, and the editor David Scott for helpful comments. Any opinions expressed in this article are those of the authors and not necessarily of Educational Testing Service.

[Received April 2003. Revised July 2004.]

REFERENCES

Albert, J., and Chib, S. (1993), "Bayesian Analysis of Binary and Polychotomous Response Data," *Journal of the American Statistical Association*, 88, 669–679.

——— (1997), "Bayesian Tests and Model Diagnostics in Conditionally Independent Hierarchical Models," *Journal of the American Statistical Association*, 92, 916–925.

Besag, J., York, J., and Mollie, A. (1991), "Bayesian Image Restoration, with Two Applications in Spatial Statistics," *Annals of the Institute of Statistical Mathematics*, 43, 1–20.

Carlin, B. P., and Chib, S. (1995), "Bayesian Model Choice via Markov Chain Monte Carlo Methods," *Journal of the Royal Statistical Society*, Ser. B, 57, 473–484.

Chen, M. H., and Shao, Q. M. (1997), "Estimating Ratios of Normalizing Constants for Densities with Different Dimensions," *Statistica Sinica*, 7, 607–630.

Chib, S. (1995), "Marginal Likelihood From the Gibbs Output," *Journal of the American Statistical Association*, 90, 1313–1321.

Chib, S., and Jeliazkov, I. (2001), "Marginal Likelihood from the Metropolis-Hastings Output," *Journal of the American Statistical Association*, 96, 270–281.

Clayton, D., and Kaldor, J. (1987), "Empirical Bayes Estimates of Age-Standardized Relative Risks for Use in Disease Mapping," *Biometrics*, 43, 671–681.

Cressie, N., Stern, H. S., and Wright, D. R. (2000), "Mapping Rates Associated with Polygons," *Journal of Geographical Systems*, 2, 61–69.

Daniels, M. J. (1999), "A Prior for the Variance in Hierarchical Models," *The Canadian Journal of Statistics*, 27, 567–578.

DiCiccio, T. J., Kass, R. E., Raftery, A., and Wasserman, L. (1997), "Computing Bayes Factors by Combining Simulation and Asymptotic Approximations," *Journal of the American Statistical Association*, 92, 903–915.

Gelman, A. E., and Meng, X. L. (1998), "Simulating Normalized Constants: From Importance Sampling to Bridge Sampling to Path Sampling," *Statistical Science*, 13, 163–185.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003), *Bayesian Data Analysis* (2nd ed.), New York: Chapman & Hall.

Geyer, A. E., and Thompson, E. A. (1992), "Constrained Monte Carlo Maximum Likelihood for Dependent Data," *Journal of the Royal Statistical Society*, Ser. B, 54, 657–683.

Green, P. J. (1995), "Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination," *Biometrika*, 82, 711–732.

Han, C., and Carlin, B. (2001), "Markov Chain Monte Carlo Methods for Computing Bayes Factors: A Comparative Review," *Journal of the American Statistical Association*, 96, 1122–1132.

Janzen, F. J., Tucker, J. K., and Paukstis, G. L. (2000), "Experimental Analysis of An Early Life History Stage: Selection On Size of Hatchling Turtles," *Ecology*, 81, 2290–2304.

Kass, R. E., and Raftery, A. E. (1995), "Bayes Factors," *Journal of the American Statistical Association*, 90, 773–795.

Lin, X. (1997), "Variance Component Testing in Generalized Linear Models with Random Effects," *Biometrika*, 84, 309–326.

Meng, X. L., and Schilling, S. (2003), "Warp Bridge Sampling," *Journal of the Computational and Graphical Statistics*, 11, 552–586.

Meng, X. L., and Wong, W. H. (1996), "Simulating Ratios of Normalizing Constants via A Simple Identity: A Theoretical Exploration," *Statistica Sinica*, 6, 831–860.

Pauler, D. K., Wakefield, J. C., and Kass, R. E. (1999), "Bayes Factors for Variance Component Models," *Journal of the American Statistical Association*, 94, 1242–1253.

Sinharay, S. (2001), "Bayes Factors for Variance Component Testing in Generalized Linear Mixed Models," Doctoral dissertation, Iowa State University, 2001, Dissertation Abstracts International, 61.

Sinharay, S., and Stern, H. (2003), "Variance Component Testing in Generalized Linear Mixed Models," ETS RR-03-14, ETS, Princeton, NJ.

Tierney, L., and Kadane, J. (1986), "Accurate Approximations for Posterior Moments and Marginal Densities," *Journal of the American Statistical Association*, 81, 82–86.

Verdinelli, I., and Wasserman, L. (1995), "Computing Bayes Factors Using the Savage-Dickey Density Ratio," *Journal of the American Statistical Association*, 90, 614–618.

Voter, A. F. (1985), "A Monte Carlo Method for Determining Free-Energy Differences and Transition State Theory Rate Constants," *Journal of Chemical Physics*, 82, 1890–1899.