

**RESEARCH
REPORT**

May 2003
RR-03-14

**Variance Component Testing in
Generalized Linear Mixed Models**

Sandip Sinharay
Hal S. Stern



Research &
Development Division
Princeton, NJ 08541

Variance Component Testing in Generalized Linear Mixed Models

Sandip Sinharay, Educational Testing Service, Princeton, NJ

Hal S. Stern, University of California, Irvine

May 2003

Research Reports provide preliminary and limited dissemination of ETS research prior to publication. They are available without charge from:

Research Publications Office
Mail Stop 10-R
Educational Testing Service
Princeton, NJ 08541

Abstract

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. For example, in a number of biological applications, the responses are correlated due to common genetic or environmental factors. In many applications, the magnitude of the variance components corresponding to one or more of the random effects are of interest, especially the point null hypothesis that one or more of the variance components are zero. This work reviews a number of approaches for estimating the Bayes factor comparing the models with and without the random effects in question. The computations involved with finding Bayes factors for these models pose many challenges, and we discuss how one can overcome them.

We perform a comparative study of the different approaches to compute Bayes factors for GLMMs by applying them to two different data sets. 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 being used to analyze the data. The importance sampling method is found to be the method of choice to compute the Bayes factor of interest for these problems. Chib's method is also found to be efficient.

Key words: Chib's method, importance sampling method, marginal density

Acknowledgements

This work was partially supported by National Institutes of Health award CA78169. The authors thank Frederic Janzen for providing the data from the natural selection study and Michael Daniels, Howard Wainer, Siddhartha Chib, and Hariharan Swaminathan for helpful comments. Any opinions expressed in this paper are those of the authors and not necessarily of Educational Testing Service.

1. Introduction

Generalized linear mixed models (GLMM), also known as 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. In many applications, the magnitude of the variance components corresponding to one or more of the random effects are of interest, especially the point null hypothesis that one or more of the variance components are zero. A Bayesian approach for testing a hypothesis is to compute the Bayes factor comparing the two competing models—one suggested by the null hypothesis and another by the alternative hypothesis. The computations involved with finding Bayes factors for these models pose many challenges, even when applied to small data sets, and we discuss how one can overcome them. The objective of this work is to apply and evaluate the performance of different approaches for estimating the Bayes factor comparing the GLMMs with and without the random effects in question.

A number of similar studies exist in statistical literature. The two most closely related are those of Han and Carlin (2001) and Albert and Chib (1997). Han and Carlin (2001) review several Markov chain Monte Carlo methods for estimating Bayes factors for mixed models, emphasizing the normal linear mixed model. We compare our results to theirs. Albert and Chib (1997) use Bayes factors for judging a variety of assumptions in conditionally independent hierarchical models including assumptions regarding the variance component. Our study focuses only on this last question and on comparing different computing methods. Pauler, Wakefield, and Kass (1999) provide a number of results about computing Bayes factors for variance component testing in linear mixed effects models. Diccio, Kass, Raftery, and Wasserman (1997) compare several methods of estimating Bayes factors when it is possible to simulate observations from the posterior distributions. Their study was quite general whereas the present work focuses on GLMMs. Only a few known studies compute Bayes factors for comparing GLMMs, even fewer dealing with complicated GLMMs (like that in our second example) or focussing on the variance components (rather than the regression coefficients). Thus, our work provides useful information to researchers performing hypothesis testing with GLMMs. Further,

researchers using Bayes factors for other models might also find this study useful.

This paper is organized as follows. Section 2 discusses a number of preliminary ideas regarding GLMMs. The next section introduces the Bayes factor and then discusses a number of approaches for estimating the Bayes factor that corresponds to the test of a null hypothesis specifying that one or more variance components in a GLMM are zero. Section 4 describes an application of a simple GLMM, a probit regression model with random effects, to the data set from a natural selection study (Janzen, Tucker, & Paukstis, 2000). The Bayes factor (comparing the models with and without the variance component) is estimated using the different approaches discussed in Section 3 and the performance of the approaches compared. Section 5 takes up a more complex example which involves the Scotland lip-cancer data set (Clayton & Caldor, 1987). A Poisson-normal regression model with spatial random effects and heterogeneity random effects is fit to these data. Section 6 provides a discussion on our findings and our recommendations.

2. Preliminaries

Generalized Linear Mixed Models (GLMM)

Generalized linear models (GLM) allow for the use of linear modeling ideas in settings where the response is not normally distributed. Examples include logistic/probit regression for binary responses or Poisson regression for count data. Frequently the responses are correlated even after conditioning on the covariates of interest, e.g., individuals from the same family share common genetic factors. GLMMs use random effects along with a GLM to take such correlations into account.

Let $\mathbf{y} = (y_1, y_2, \dots, y_n)$ denote the observed responses. In a GLMM, the y_i s 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 any scale parameter and all of the methods described here can be modified to accommodate a scale parameter. Let $\mu_i = E(y_i|\xi_i) = a'(\xi_i)$. The mean μ_i (and hence ξ_i) is expressed as a

function of a $p \times 1$ predictor vector \mathbf{x}_i , a $p \times 1$ vector of coefficients $\boldsymbol{\alpha}$ and a $q \times 1$ random effects vector \mathbf{b} through the link function $g(\mu_i) = \mathbf{x}'_i \boldsymbol{\alpha} + \mathbf{z}'_i \mathbf{b}$, where \mathbf{z}_i is a $q \times 1$ (typically 0/1) vector associated with the random effects. The random effects vector \mathbf{b} is assigned a prior distribution $f(\mathbf{b}|\boldsymbol{\theta})$; usually, $f(\mathbf{b}|\boldsymbol{\theta})$ is assumed to be normally distributed with mean vector $\mathbf{0}$ and a positive definite variance matrix $\mathbf{D}(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is an $m \times 1$ vector of unknown variance component parameters. The magnitude of $\boldsymbol{\theta}$ determines the degree of overdispersion and correlation among the responses. Typically, the model is parameterized such that $\mathbf{D}(\boldsymbol{\theta}) = \mathbf{0}$ iff $\boldsymbol{\theta} = \mathbf{0}$. Note that $\boldsymbol{\theta} = \mathbf{0} \Leftrightarrow \mathbf{b} = \mathbf{0}$, which corresponds to a GLM.

Marginal Likelihood for Generalized Linear Mixed Models

The likelihood function $L(\boldsymbol{\alpha}, \boldsymbol{\theta}|\mathbf{y})$, also called the marginal likelihood function, is obtained by integrating out the random effects from the conditional density of the response using $f(\mathbf{b}|\boldsymbol{\theta})$ as weight:

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

The integral is analytically intractable except for normal linear models, making computations with GLMMs difficult. Numerical integration techniques (e.g., Simpson's rule) or Laplace approximation (Tierney & Kadane, 1986) may be used to approximate the GLMM likelihood. However, each of these two approaches is problematic and is not generally recommended (see, for example, Sinharay, 2001, and the references therein).

Geyer and Thompson (1992) and Gelfand and Carlin (1993) suggest the use of importance sampling to estimate the value of the likelihood function. Starting from (1), for an importance sampling distribution $q(\mathbf{b})$, $L(\boldsymbol{\alpha}, \boldsymbol{\theta}|\mathbf{y})$ is expressed as

$$\int_{\mathbf{b}} \frac{1}{q(\mathbf{b})} \left\{ \prod_{i=1}^n p(y_i|\boldsymbol{\alpha}, \mathbf{b}) \right\} f(\mathbf{b}|\boldsymbol{\theta}) q(\mathbf{b}) d\mathbf{b} \approx \frac{1}{N} \sum_{k=1}^N \frac{1}{q(\mathbf{b}^{(k)})} \left\{ \prod_{i=1}^n p(y_i|\boldsymbol{\alpha}, \mathbf{b}^{(k)}) \right\} f(\mathbf{b}^{(k)}|\boldsymbol{\theta}),$$

where $\mathbf{b}^{(k)}$, $k = 1, 2, \dots, N$, is a sample from $q(\mathbf{b})$.

The choice of the importance sampling density is not straightforward in estimating an integral using the importance sampling approach, especially for high-dimensional

random effects (e.g., for the spatial Markov random field models that we will use in Section 5). Theoretically, for the integral to be estimated precisely, $q(\mathbf{b})$, the importance sampling density, should be of the same shape and should have heavier tails than the product $\left\{ \prod_{i=1}^n p(y_i | \boldsymbol{\alpha}, \mathbf{b}) \right\} f(\mathbf{b} | \boldsymbol{\theta})$. It is also convenient if the importance sampling density can be obtained once and used to evaluate the likelihood for a potentially large number of $(\boldsymbol{\alpha}, \boldsymbol{\theta})$ pairs. For the importance sampling density, this work uses a t_4 density with the first two moments obtained as the sample mean and sample variance of the relevant component of a sample from the joint posterior distribution of $(\boldsymbol{\alpha}, \boldsymbol{\theta}, \mathbf{b})$. This importance sampling density estimates $L(\boldsymbol{\alpha}, \boldsymbol{\theta} | \mathbf{y})$ with reasonable precision within reasonable time.

Testing Hypotheses About Variance Components for GLMMs

Inferences about the contribution of the random effects to the GLMM are mostly obtained by examining point (or interval) estimates of the variance parameters in \mathbf{D} . In many practical problems, researchers 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

Introduction

The Bayesian approach to test a hypothesis about the variance component(s) is to compute the Bayes factor $BF^{01} = \frac{p(\mathbf{y} | M_0)}{p(\mathbf{y} | M_1)}$, which compares the marginal densities of \mathbf{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 .

Another way to express the Bayes factor is the following:

$$BF^{01} = \frac{p(M_0|\mathbf{y})}{p(M_1|\mathbf{y})} / \frac{p(M_0)}{p(M_1)}, \quad (2)$$

i.e., the Bayes factor is the ratio of posterior odds and prior odds. As discussed later, this expression is useful in forming an estimate of the Bayes factor using the reversible jump Markov chain Monte Carlo (MCMC) method, which obtains empirical estimates of $p(M_0|\mathbf{y})$ and $p(M_1|\mathbf{y})$.

Kass and Raftery (1995) provide a comprehensive review of Bayes factors including information about their interpretation. Bayes factors are sensitive to the prior distributions used in the models. Sinharay (2001) and Sinharay and Stern (2002) suggest a graphical approach to study the sensitivity of the Bayes factor to the prior distribution for the variance parameters for GLMMs.

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 cannot be computed analytically for either the model with unrestricted variance components (M_1) or that with variance component(s) set to zero (M_0). 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 in this and the subsequent chapters. More details about the methods are found, for example, in Sinharay (2001) and the references provided there.

Most of the approaches estimate the marginal density of the data separately under each model. The ratio of the marginal densities is the estimated Bayes factor. The Verdinelli-Wasserman and Reversible Jump MCMC approaches estimate the Bayes Factor directly. In terms of notation, for the remainder of this section, $\omega = (\alpha, \theta)$, implying that the random effects parameters \mathbf{b} have been integrated out as described in the previous section. The final part of this section discusses issues related to this parameterization.

Laplace Approximation

The Laplace approximation (reviewed by Tierney & Kadane 1986) of the marginal density under a model is obtained by approximating $p(\mathbf{y}|\boldsymbol{\omega}, M)p(\boldsymbol{\omega}|M)$ (which is a multiple of the posterior distribution) by a normal distribution with mean $\hat{\boldsymbol{\omega}}$, the mode of $p(\mathbf{y}|\boldsymbol{\omega}, M)p(\boldsymbol{\omega}|M)$ (i.e., the posterior mode) and variance $\hat{\Sigma}$ as the inverse of the negative Hessian matrix of the log-posterior evaluated at $\hat{\boldsymbol{\omega}}$. The Laplace approximation formula is

$$p(\mathbf{y}|M) \approx (2\pi)^{d/2} |\hat{\Sigma}|^{-1/2} p(\mathbf{y}|\hat{\boldsymbol{\omega}}, M) p(\hat{\boldsymbol{\omega}}|M),$$

where d is the dimension of $\boldsymbol{\omega}$. The relative error of the approximation is $\mathbf{O}(\frac{1}{n})$, where n is the original sample size. However, there may be problems with this approximation if the posterior mode is on the boundary of the parameter space. This may occur in a GLMM when the posterior mode for one or more variance components may be zero, especially if the null model is true. We treat the logarithms of the variance components as the parameters while applying this method to facilitate the normal approximation of the posterior distribution.

Importance Sampling

Given a sample $\boldsymbol{\omega}_i, i = 1, 2, \dots, N$ from an ‘‘importance sampling distribution’’ Q with corresponding density function q , this approach estimates the marginal density of the data under model M as

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

A practical problem is to find a Q such that $p(\mathbf{y}|M)$ is precisely estimated.

Our work takes the importance sampling distribution Q to be a t_4 distribution with mean as the mode of $p(\mathbf{y}|\boldsymbol{\omega}, M)p(\boldsymbol{\omega}|M)$ and the variance matrix as the inverse of the negative Hessian matrix of the logarithm of $p(\mathbf{y}|\boldsymbol{\omega}, M)p(\boldsymbol{\omega}|M)$ calculated at its mode. For GLMMs, as $\boldsymbol{\omega}$ includes the variance parameters as well, $p(\mathbf{y}|\boldsymbol{\omega}, M)p(\boldsymbol{\omega}|M)$ may be skewed. As a result, the t distribution performs better as an importance sampling density than the normal distribution because the former is more likely to have tails as heavy as

$p(\mathbf{y}|\boldsymbol{\omega}, M)p(\boldsymbol{\omega}|M)$. We work with the logarithm of the variance components to facilitate the approximation of the posterior distribution by a t distribution. Recall that computation of $p(\mathbf{y}|\boldsymbol{\omega}, M)$ is itself a numerical integration (discussed in Section 2) that may be evaluated using importance sampling.

Harmonic Estimator

Newton and Raftery (1994) develop the harmonic estimator from the following identity, which holds for any density function h :

$$[p(\mathbf{y}|M)]^{-1} = \int \frac{h(\boldsymbol{\omega})}{p(\mathbf{y}|\boldsymbol{\omega}, M)p(\boldsymbol{\omega}|M)} p(\boldsymbol{\omega}|\mathbf{y}, M) d\boldsymbol{\omega}.$$

Given a sample $\boldsymbol{\omega}_i, i = 1, 2, \dots, N$ from the posterior distribution under model M ,

$$p(\mathbf{y}|M) \approx \left\{ \frac{1}{N} \sum_{i=1}^N \frac{h(\boldsymbol{\omega}_i)}{p(\mathbf{y}|\boldsymbol{\omega}_i, M)p(\boldsymbol{\omega}_i|M)} \right\}^{-1} \quad (3)$$

provides an estimate of the marginal density. The harmonic estimator of $p(\mathbf{y}|M)$ is then obtained by choosing $h(\boldsymbol{\omega}) = p(\boldsymbol{\omega}|M)$. This method is simulation-consistent, i.e., the estimated marginal density converges almost surely to the true marginal as $N \rightarrow \infty$. However, the estimate is not stable; the estimate of $[p(\mathbf{y}|M)]^{-1}$ does not have finite variance. Satagopan, Newton, and Raftery (2000) suggest a stabilized form of the harmonic estimators.

Chib's Method

Chib (1995) develops a useful approach for estimating the marginal density from the identity

$$p(\mathbf{y}|M) = \frac{p(\mathbf{y}|\boldsymbol{\omega}, M)p(\boldsymbol{\omega}|M)}{p(\boldsymbol{\omega}|\mathbf{y}, M)}. \quad (4)$$

Note that the left hand side of the above identity does not depend on $\boldsymbol{\omega}$; so the equality must hold for every value of $\boldsymbol{\omega}$. The marginal density $p(\mathbf{y}|M)$ is then the right hand side evaluated at any $\boldsymbol{\omega} = \boldsymbol{\omega}^*$.

Of the three terms on the right hand side, the likelihood $p(\mathbf{y}|\boldsymbol{\omega}, M)$ and the prior distribution $p(\boldsymbol{\omega}|M)$ can be computed at a fixed $\boldsymbol{\omega}^*$ without much difficulty. The

computation of the third term, the posterior ordinate at ω^* , is not trivial. Given a sample from the posterior distribution (perhaps using an MCMC algorithm), kernel density approximation may be used to estimate the posterior ordinate. Kernel density approximations become unreliable in high dimensions, however. Chib (1995) and Chib and Jeliazkov (2001) suggest more efficient algorithms to estimate the posterior ordinate when either Gibbs sampling or Metropolis-Hastings algorithm (Gelman, Carlin, Stern, & Rubin, 1995) is used to generate a sample from the posterior distribution.

Estimating the posterior from Gibbs sampling output. Chib (1995) suggests an algorithm using Gibbs sampler output to estimate the posterior ordinate $p(\omega|\mathbf{y}, M)$ at $\omega = \omega^*$ when ω can be partitioned into several blocks so that the full conditional for each block is available in closed form. For simplicity, we discuss the case of two blocks, $\omega = (\omega_1, \omega_2)$. A Gibbs sampler runs, iteratively generating ω_1 and ω_2 from their known conditional distributions $p(\omega_1|\omega_2, \mathbf{y}, M)$ and $p(\omega_2|\omega_1, \mathbf{y}, M)$, resulting in a (post-convergence) posterior sample (ω_1^i, ω_2^i) , $i = 1, 2, \dots, N$. Note that

$$p(\omega_1^*, \omega_2^*|\mathbf{y}, M) = p(\omega_1^*|\mathbf{y}, M)p(\omega_2^*|\omega_1^*, \mathbf{y}, M).$$

The second term in the right side of the above is known. Monte Carlo integration estimates the first term, $p(\omega_1^*|\mathbf{y}, M)$, as

$$p(\omega_1^*|\mathbf{y}, M) \approx \frac{1}{N} \sum_{i=1}^N p(\omega_1^*|\omega_2^i, \mathbf{y}, M). \quad (5)$$

This technique for estimating the posterior ordinate easily generalizes to the case when ω consists of more than two blocks.

Estimating the posterior from Metropolis output. Chib and Jeliazkov (2001) extend the above idea to allow the use of Metropolis-Hastings output to estimate $p(\omega|\mathbf{y}, M)$ at $\omega = \omega^*$. Assume that a Metropolis-Hastings algorithm generates values of the parameter vector ω in a single block from the posterior distribution $p(\omega|\mathbf{y}, M)$. We drop the model indicator M from the notation for convenience. Let

$$\alpha(\omega, \omega'|\mathbf{y}) = \min \left\{ 1, \frac{p(\omega'|\mathbf{y})q(\omega', \omega|\mathbf{y})}{p(\omega|\mathbf{y})q(\omega, \omega'|\mathbf{y})} \right\},$$

where $q(\boldsymbol{\omega}, \boldsymbol{\omega}' | \mathbf{y})$ denotes the proposal density (candidate generating density) for transition from $\boldsymbol{\omega}$ to $\boldsymbol{\omega}'$. Chib and Jeliazkov (2001) show that

$$p(\boldsymbol{\omega}^* | \mathbf{y}) = \frac{E_1 \{ \alpha(\boldsymbol{\omega}, \boldsymbol{\omega}^* | \mathbf{y}) q(\boldsymbol{\omega}, \boldsymbol{\omega}^* | \mathbf{y}) \}}{E_2 \{ \alpha(\boldsymbol{\omega}^*, \boldsymbol{\omega} | \mathbf{y}) \}}, \quad (6)$$

where E_1 is the expectation with respect to $p(\boldsymbol{\omega} | \mathbf{y})$ and E_2 is the expectation with respect to $q(\boldsymbol{\omega}^*, \boldsymbol{\omega} | \mathbf{y})$. The numerator is estimated by averaging the product within the braces with respect to the draws from the posterior distribution, while the denominator is estimated by averaging the acceptance probability with respect to draws from $q(\boldsymbol{\omega}^*, \boldsymbol{\omega} | \mathbf{y})$, given the fixed value $\boldsymbol{\omega}^*$. The nice thing about the calculation is that it does not require knowledge of the normalizing constant for $p(\boldsymbol{\omega} | \mathbf{y})$. The generalization of the algorithm to the case with more than one block in the parameter vector is straightforward.

Estimating the posterior where both Metropolis and Gibbs are used. Often researchers sample from a posterior distribution using a Gibbs sampler, switching to Metropolis-Hastings steps to generate from some conditional distributions. In that case, ideas from both Chib (1995) and Chib and Jeliazkov (2001) are combined, as discussed in the latter paper, to estimate the posterior ordinate $p(\boldsymbol{\omega}^* | \mathbf{y})$.

Chib's method for GLMMs. A number of points pertain specifically to the use of Chib's methods for GLMMs.

1. Parameterization: It is sometimes convenient to alter the parameterization so that $\boldsymbol{\omega}$ includes the random effects as well when applying Chib's method. This is discussed in Section 3 and then applied in the example in Section 4.
2. Blocking: A key to using Chib's method is doing efficient blocking. For most GLMMs, partitioning $\boldsymbol{\omega}$ into two blocks is most efficient—one block containing the fixed effects parameters and the other containing the variance parameters.
3. Choice of $\boldsymbol{\omega}^*$: The identity (4) is true for any choice $\boldsymbol{\omega} = \boldsymbol{\omega}^*$. However, for efficiency of estimation, $\boldsymbol{\omega}^*$ 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 and posterior

median. For GLMMs, the posterior distribution of the variance parameter(s) is skewed; hence the posterior mode is a better choice of ω^* . However, finding the posterior mode requires additional computation.

4. Algorithm: For a few simple GLMMs, Gibbs sampling can be used to generate from the posterior distribution and thus the marginal density evaluated using the idea of Chib (1995). However, it seems to occur more often that some Metropolis steps are required, necessitating the use of the Chib and Jeliazkov (2001) approach.

Verdinelli-Wasserman Method

Verdinelli and Wasserman (1995) suggest a method for estimating Bayes factors. The method is appropriate for comparing nested models directly and does not require approximation of the marginal densities for the two models separately. Let $\omega = (\delta', \psi)'$ be the parameter vector, M_0 be the null model with the restriction $\delta = \delta_0$, and M_1 be the unrestricted (alternative) model. Further, let $p_0(\psi)$ be the prior distribution of ψ under the null model and $p(\psi, \delta)$ be the joint prior distribution of ψ and δ under the unrestricted model. Then the Bayes factor BF^{01} can be expressed as:

$$BF^{01} = p(\delta_0|\mathbf{y})E\left[\frac{p_0(\psi)}{p(\psi, \delta_0)}\right], \quad (7)$$

where the expectation is with respect to $p(\psi|\delta_0, \mathbf{y})$ and $p(\delta|\mathbf{y}) = \int p(\delta, \psi|\mathbf{y})d\psi$. If $p(\psi|\delta_0) = p_0(\psi)$, then (7) simplifies considerably and the Bayes factor is Savage's density ratio,

$$BF^{01} = \frac{p(\delta_0|\mathbf{y})}{p(\delta_0)}. \quad (8)$$

For GLMMs it is common to assume a priori independence of the variance components and the regression parameters so that BF^{01} can be computed using (8) with δ as the part of the vector of variance components being tested and δ_0 is usually 0. The estimation of Savage's density ratio requires the estimated posterior density $p(\delta|\mathbf{y})$ at $\delta = \delta_0$. Given a sample from the posterior distribution of ω under the full model (including variance components), $p(\delta_0|\mathbf{y})$ may be obtained via kernel density estimation. Use of Savage's density ratio

requires that the prior distribution for δ have nonzero and finite density value at the point δ_0 that is being tested.

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. A typical point in the parameter space of this extended model is (j, ω_j) , where j is the model index and ω_j is the n_j -dimensional parameter vector for model j , $j = 1, 2, \dots, J$. The reversible jump MCMC method suggested by Green (1995) samples from the expanded posterior distribution. This method generates a Markov chain that can jump between models with parameter spaces of different dimensions. Let π_j be the prior probability on model j , $j = 1, 2, \dots, J$. Then the steps in the reversible jump algorithm are as follows:

1. Let the current state of the chain be (j, ω_j) .
2. In an attempt to jump to another model, propose a new model j' with probability $h(j, j')$, where $h(j, j')$ is a probability mass function, i.e., $\sum_{j'} h(j, j') = 1$.
3.
 - a. If $j' = j$, then perform an MCMC iteration (Gibbs or Metropolis) for the parameter ω_j of model j . Go to step 1.
 - b. If $j' \neq j$, then ω_j and $\omega_{j'}$ often have different dimensions and mutually unrelated components. To generate $\omega_{j'}$, use “dimension-matching”—generate an auxiliary random variable u from a proposal density $q(u|\omega_j, j, j')$ and set $(\omega_{j'}, u') = g_{j,j'}(\omega_j, u)$, where g is a one-to-one onto deterministic function and $n_j + \dim(u) = n_{j'} + \dim(u')$; this takes care of the dimension-matching across models. The choice of $q(u|\omega_j, j, j')$, g , u , and u' depends on the problem at hand.
4. Accept the move from j to j' with probability

$$\min \left\{ 1, \frac{p(\mathbf{y}|\omega_{j'}, M = j')p(\omega_{j'}|M = j')\pi_{j'}h(j', j)q(u'|\omega_{j'}, j', j)}{p(\mathbf{y}|\omega_j, M = j)p(\omega_j|M = j)\pi_jh(j, j')q(u|\omega_j, j, j')} \cdot \left| \frac{\partial g(\omega_j, u)}{\partial(\omega_j, u)} \right| \right\}.$$

If the above Markov chain runs sufficiently long (details discussed with examples), then N_j , the number of times the Markov chain reaches a particular model j , is approximately

proportional to the posterior probability of the model, i.e.,

$$\frac{p(M_j|\mathbf{y})}{p(M'_j|\mathbf{y})} \approx \frac{N_j}{N_{j'}}.$$

Therefore, once a sequence of simulations from the posterior distribution is generated, the Bayes factor $BF^{jj'}$ for comparing model j and model j' is estimated, using (2), as

$$BF^{jj'} \approx \frac{N_j}{N_{j'}} / \frac{\pi_j}{\pi_{j'}},$$

where N_j is the number of iterations of the Markov chain in model j .

Other Methods

The methods described here do not exhaust all possible methods. While 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. Other methods for computing Bayes factors appropriate in the GLMM context include bridge sampling (Meng & Wong, 1996), product space search (Carlin & Chib, 1995), Metropolized product space search (Dellaportas, Forster, & Ntzoufras, 2002), and reversible jump using partial analytic structure (Godsill, 2001).

Parameterization

A number of the methods for computing Bayes factors require computing the marginal 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 (especially those requiring more than one marginal likelihood computation) become impractical. This is frequently the case with GLMMs. However, we have a way to get around this problem.

The marginal density $p(\mathbf{y})$ (we drop the model indicator for simplicity) for a GLMM can be expressed as :

$$p(\mathbf{y}) = \int \int p(\mathbf{y}|\boldsymbol{\alpha}, \boldsymbol{\theta})p(\boldsymbol{\alpha}, \boldsymbol{\theta})d\boldsymbol{\alpha}d\boldsymbol{\theta} \quad (9)$$

$$= \int \int \int p(\mathbf{y}|\boldsymbol{\alpha}, \mathbf{b})p(\mathbf{b}|\boldsymbol{\theta})p(\boldsymbol{\alpha}, \boldsymbol{\theta})d\mathbf{b}d\boldsymbol{\alpha}d\boldsymbol{\theta} \quad (10)$$

As a result, rather than using $\omega = (\alpha, \theta)$ and (9), it is often simpler and time-saving to include the random effects as well in ω , i.e., to take $\omega = (\mathbf{b}, \alpha, \theta)$. The big advantage of this definition of ω is that the computation of the likelihood function

$$p(\mathbf{y}|\omega) = p(\mathbf{y}|\alpha, \mathbf{b}, \theta) = p(\mathbf{y}|\alpha, \mathbf{b})p(\mathbf{b}|\theta)$$

becomes very easy. However, as a price to pay, the dimension of the parameter space increases by the number of components in \mathbf{b} , which is usually high, even for simple GLMMs.

4. Example: A Natural Selection Study

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 consists 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 scatter plot of the birth weights 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 according to the increasing order of average birth weight of the turtles in them. 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 (survival status with 1 denoting survival) and x_{ij} the birth weight of the j -th turtle in the i -th clutch, $i = 1, 2, \dots, m = 31$, $j = 1, 2, \dots, n_i$. The probit regression model with random effects fit to the data set is given by:

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

The b_i s are random effects for clutch (family). Other models are possible, but our work is restricted to this single model to study the issues related with estimating Bayes factors.

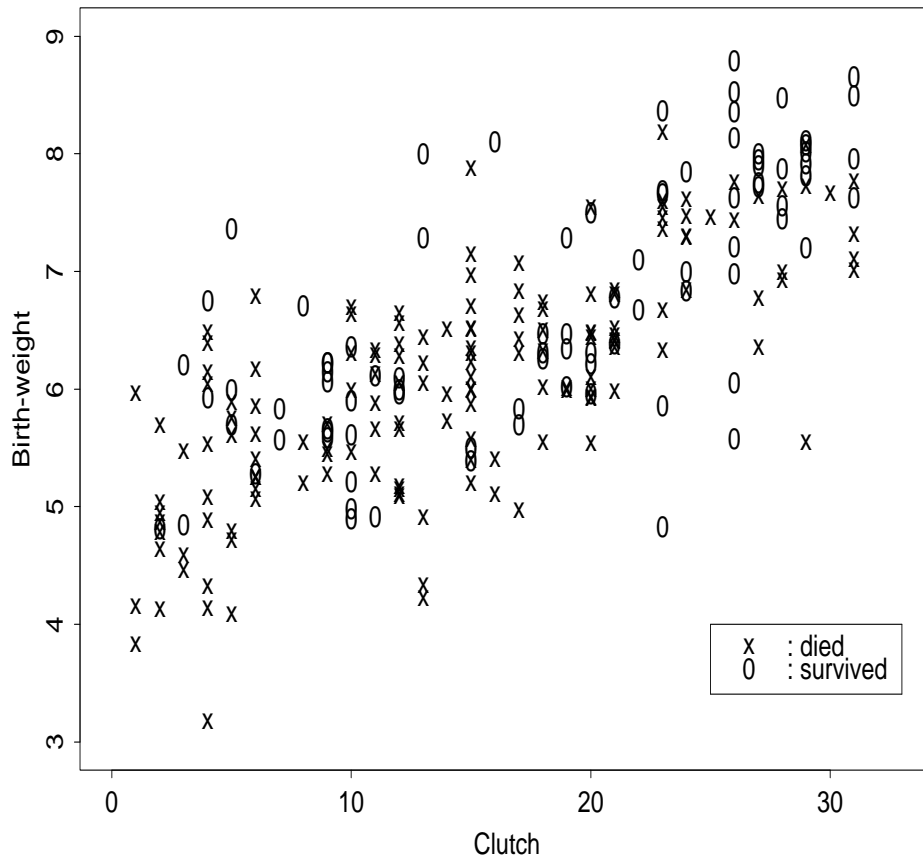


Figure 1: Scatter plot with the clutches sorted by average birth weight.

Estimating the Bayes Factor

It was noted in Section 3 that application of the Verdinelli-Wasserman method requires that the prior distribution for σ^2 be finite and non-zero at $\sigma^2 = 0$. 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 a fixed constant denoting the median of $p(\sigma^2)$. We fix c at 1 and use this prior distribution for all the methods. A proper vague prior distribution $p(\boldsymbol{\alpha}) = N_2(\mathbf{0}, 10\mathbf{I})$ is used for $\boldsymbol{\alpha}$.

The Bayes factor for comparing the null model M_0 (that without variance components) against the alternative model M_1 (that with the variance component) is given

by $BF^{01} = \frac{p(\mathbf{y}|M_0)}{p(\mathbf{y}|M_1)}$, where

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

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

We estimate the above-mentioned Bayes factor using the methods discussed earlier.

Methods

Where required, the marginal likelihood $p(\mathbf{y}|\boldsymbol{\alpha}, \sigma^2)$ for any $(\boldsymbol{\alpha}, \sigma^2)$ is calculated by numerically integrating out the random effects using Simpson's rule. Where a posterior sample is required in any of these methods, we obtain a sample using an MCMC algorithm. All simulation-based estimates use a post-convergence posterior sample of size 5000. Specific details concerning the implementation of the individual approaches follow.

Chib's method. The conditional distribution of σ^2 is not of known form, necessitating the use of the Metropolis algorithm (Gelman et al., 1995) for sampling from the joint posterior distribution of the parameters under the variance component model. The same algorithm is used to generate a sample from the null model. The posterior mode is taken to be the fixed point required in this method.

Verdinelli-Wasserman. The Savage density ratio, given by (8), is applicable here. We use the S-PLUS function “density” (Venables & Ripley, 1998) for estimating the posterior density from a posterior sample. The function implements a kernel-density estimation with normal density function as the default choice of kernel. Our work uses the kernel bandwidth choice suggested in Silverman (1986, p. 45), $\hat{b} = 4 \times 1.06 \min(s, R/1.34)n^{-\frac{1}{5}}$, where R is the interquartile range of the sample and s is the standard deviation of the sample.

Reversible jump MCMC. Our choices are described here using the notation from Section 3. We set $h(0,0) = h(0,1) = h(1,0) = h(1,1) = 0.5$ and $\pi_0 = \pi_1 = 0.5$. When we are in model 0 (without σ^2) and are trying a jump to model 1 (in Step 3b in Section 3), $\boldsymbol{\omega}_0 = \boldsymbol{\alpha}$, $\boldsymbol{\omega}_1 = (\boldsymbol{\alpha}, \sigma^2)$ and the steps become:

- generate σ^2 from $q(\sigma^2) = \text{Inverse gamma with mean } \hat{\sigma}_{MLE}^2 \text{ and variance } \hat{V}(\hat{\sigma}_{MLE}^2)$
- define $\omega_1 = (\omega_0, \sigma^2)$; hence $u = \sigma^2$, $u' = 0$ and $g(\mathbf{x}) = \mathbf{x}$.
- acceptance probability: $\min\{1, \frac{f(\mathbf{y}|\boldsymbol{\alpha}, \sigma^2, M=1)p(\sigma^2)\pi_1}{f(\mathbf{y}|\boldsymbol{\alpha}, M=0)q(\sigma^2)\pi_0}\}$

When we are in model 1 (with σ^2) trying to jump to model 0, the steps are:

- define $(\omega_0, \sigma^2) = \omega_1$; hence, $u = 0$, $u' = \sigma^2$ and $g(\mathbf{x}) = \mathbf{x}$.
- acceptance probability: $\min\{1, \frac{f(\mathbf{y}|\boldsymbol{\alpha}, M=0)q(\sigma^2)\pi_0}{f(\mathbf{y}|\boldsymbol{\alpha}, \sigma^2, M=1)p(\sigma^2)\pi_1}\}$

To jump within a model, we take a Metropolis step with a Gaussian proposal distribution having mean equal to the present value of the parameters and variance equal to the inverse of the negative Hessian matrix of the log-posterior at the posterior mode for that model.

Results

Numerical integration over all parameters provides us the true value of the Bayes factor of interest, although the program takes about 37 hours of CPU time to run on an Alpha station 500 workstation equipped with 400MHz 64-bit CPU and a gigabyte of RAM; the value of the Bayes factor up to 2 decimal places is 1.27. Hence, we can compare the performance of the different methods by comparing the Bayes factor estimates obtained by each method against this correct value. For each simulation-based method, we run the program to estimate the Bayes factor 30 times with different random number seeds and take the average of the 30 estimates. The standard deviation of the 30 estimates gives us an idea about the stability of the results from a method. Table 1 summarizes the results obtained by the various methods. Also shown in the table are the CPU times required for one computation of the Bayes factor estimate by each of the methods under consideration in an Alpha station 500 workstation equipped with 400MHz 64-bit CPU and a gigabyte of RAM.

The results in Table 2 indicate that importance sampling and Chib's method perform equally well and better than the other methods. The large standard error for the reversible jump MCMC method and the Verdinelli-Wasserman method can be reduced

Table 1: Estimates of the Bayes Factor for the Turtles Data Set

Method	Bayes factor estimate	Std. dev.	CPU time (min)
Laplace	1.54	-	0.1
Importance sampling	1.27	0.01	4.4
Harmonic estimator	1.89	2.31	6.7
Chib	1.29	0.03	8.4
Verdinelli-Wasserman	1.08	0.46	4.6
RJ MCMC	1.38	0.24	9.6

by increasing the sample size. The instability of the harmonic estimator is evident. The Laplace approximation, which makes the strong assumption of normality of the posterior distribution gives a fair approximation (in the sense that the right order of magnitude for the BF is obtained) given the small amount of time required.

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 computer 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. Among the methods that perform well for this problem with respect to accuracy and precision, importance sampling method requires much less time to program than Chib’s method and takes about 50% of the time to run, appearing to be the most convenient method for this data set. Some simulation results (Sinharay, 2001) suggest that these results hold more generally. For small problems like this example, the importance sampling method is the most convenient method for computing Bayes factor estimates.

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.

Description of the Data Set

Table 2 shows a part of a frequently-analyzed data set (see, e.g., Clayton & Kaldor, 1987) regarding lip cancer data from the 56 administrative districts in Scotland. The objective of the study was to find out any pattern of regional variation in the disease incidence of lip cancer. The data set contains: the observed number of lip cancer cases among males from 1975-1980 in the 56 districts, $y_1, y_2, \dots, y_n, n=56$; the population under risk of lip cancer in the districts, p_1, p_2, \dots, p_n (in thousands); the expected number of cases adjusted for the age distribution of the districts, E_1, E_2, \dots, E_n ; the percent of people employed in agriculture, forestry, and fishing (AFF), $AFF_1, AFF_2, \dots, AFF_n$ (since increased exposure to sunlight has been implicated in the excess occurrence of lip cancers, these people working outdoors were thought to be under greater risk of the disease); and the neighbors of each district, N_1, N_2, \dots, N_n . The E_i s incorporate known demographic risk factors, here age, that are not of direct interest.

Table 2: Part of the Scotland Lip Cancer Data Set

County	y	p (in '000)	x	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
5	15	129	10	4.26	5 1 11 12 13 19
6	8	53	24	2.40	2 3 8
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
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

A Poisson-Gaussian Hierarchical Model

The disease incidence counts \mathbf{y} are assumed to follow independent Poisson distributions,

$$y_i | \lambda_i \sim \text{Poisson}(\lambda_i E_i), i = 1, 2, \dots, n,$$

with λ_i representing a relative risk parameter for the i -th region. As in Besag, York, and Mollie (1991), we use a mixed linear model for the vector of log relative risk parameters, $\log(\boldsymbol{\lambda})$,

$$\log(\boldsymbol{\lambda}) = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\eta} + \boldsymbol{\psi},$$

where \mathbf{X} is the covariate matrix; $\boldsymbol{\beta} = (\beta_0, \beta_1)$ is a vector containing the fixed effect parameters; $\boldsymbol{\eta} = (\eta_1, \eta_2, \dots, \eta_n)'$ is a vector of spatially correlated random effects; and $\boldsymbol{\psi} = (\psi_1, \psi_2, \dots, \psi_n)'$ is a vector of uncorrelated heterogeneity random effects.

For modeling the random effects, we follow the choices of Cressie, Stern, and Wright (2000). The spatial random effects η_i s are intended to represent unobserved factors, that if observed, would display substantial spatial correlation. For known matrices \mathbf{C} and diagonal \mathbf{M} , we take the prior distribution for $\boldsymbol{\eta}$ as a conditional autoregressive (CAR) distribution

$$\boldsymbol{\eta}|\tau^2, \phi \sim N(\mathbf{0}, \tau^2(I - \phi\mathbf{C})^{-1}\mathbf{M}),$$

where τ^2 and ϕ are parameters of the prior distribution. The parameter ϕ is a measure of the strength of spatial dependence, $0 < \phi < \phi_{max}$, with $\phi = 0$ implying no spatial association. The matrices \mathbf{C} and \mathbf{M} are the same as those suggested by Stern and Cressie (1995):

$$c_{ij} = \begin{cases} \left(\frac{E_j}{E_i}\right)^{\frac{1}{2}} & : j \in N_i \\ 0 & : \text{elsewhere} \end{cases}$$

$$m_{ii} = E_i^{-1}.$$

For these values of c_{ij} s and m_{ii} s and the given neighborhood structure, we have $\phi_{max} = 0.1752$. These values ensure that the CAR prior distribution on $\boldsymbol{\eta}$ is a proper distribution (with nonnegative definite variance matrix).

The uncorrelated heterogeneity random effects ψ s represent the unstructured variables contributing to the logarithm of the relative risk parameters. They are modeled as $\boldsymbol{\psi}|\sigma^2 \sim N(\mathbf{0}, \sigma^2\mathbf{D})$ with diagonal matrix \mathbf{D} having $d_{ii} = E_i^{-1}$ and a variance parameter σ^2 . 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 3 variance parameters (τ^2 , σ^2 , and ϕ) and as many as 112 random effects parameters, making it a more challenging data set to handle computationally

than the turtle data set. It will be useful to note that the joint maximum likelihood estimate of $\xi = (\beta_0, \beta_1, \phi, \tau^2, \sigma^2)'$ is

$$\hat{\xi}_{MLE} = (-0.489, 0.059, 0.167, 1.640, 0.000)'.$$

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 τ^2 (and ϕ) and σ^2
- “spatial model” with τ^2 (and ϕ) only as a variance component
- “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. Note that any other Bayes factor can be obtained from these three.

In this case we use only a subset of the methods from Section 4 for a number of reasons. First, among the methods that are relatively simple to compute (in that they do not require repeated evaluations of the marginal likelihood), the Laplace approximation is not applicable because the mode is on the edge of the parameter space (Sinharay, 2001). This leaves only the Verdinelli-Wasserman approach and Chib’s method. Second, it is impractical to implement methods that require computation of the marginal likelihood at a large number of points (importance sampling, harmonic estimator, reversible jump MCMC) because integration over the random effects to evaluate the likelihood is extremely time-consuming when the number of random effects parameters is large. One way to get these methods to work is to include the random effects as parameters in the model rather than trying to integrate them out (as discussed in Section 3). However, even then the harmonic estimator is highly inaccurate and application of reversible jump MCMC method

difficult because of the issue of choice of the proposal distributions. We had considerable difficulty obtaining MCMC convergence even with long runs and a variety of proposal distributions. Thus, only importance sampling is considered from this group.

We assume a priori independence of the model parameters. We further assume a proper vague prior distribution (bivariate normal with mean $\mathbf{0}$ and variance $20\mathbf{I}$) on $\boldsymbol{\beta}$, $p(\phi) = \text{Uniform}(0, \phi_{max})$ and shrinkage prior distributions (Daniels, 1999) with parameters $= 1$ for the variance components σ^2 and τ^2 .

Methods

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

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

where $\mathbf{V} = \tau^2(\mathbf{I} - \phi\mathbf{C})^{-1}\mathbf{M} + \sigma^2\mathbf{D}$. Importance sampling (Section 2) provides an estimate of the above. We use a t_4 importance sampling distribution on $\boldsymbol{\nu}$. The mean and variance of the distribution are the corresponding moments of $\boldsymbol{\nu} = \boldsymbol{\eta} + \boldsymbol{\psi}$ computed from a posterior sample drawn from the joint posterior distribution of $(\boldsymbol{\eta}, \boldsymbol{\psi}, \boldsymbol{\beta}, \phi, \tau^2, \sigma^2)$.

Posterior samples obtained using the MCMC algorithm are used whenever a posterior sample is required by any of the methods for Bayes factor calculation. All simulation-based estimates use post-convergence posterior samples of size 50,000 from the relevant models. Additional details regarding specific approaches follow.

Importance sampling. We treat the random effects $\boldsymbol{\eta}$ and $\boldsymbol{\psi}$ as parameters, as discussed in Section 3, in the variance components models to facilitate the computations. Under the full model, the posterior mode lies on the boundary of the parameter space and the negative of the Hessian of the log-posterior evaluated at the posterior mode is not defined. Therefore, while computing with the full model, the importance sampling density is taken as a t_4 distribution with location parameter equal to the posterior mean (rather than the mode) and variance matrix equal to the posterior variance matrix.

Chib's method. Each of the conditional posterior distributions is sampled from using a Metropolis step. As for the fixed point used in this method, we use the posterior mean 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.

Verdinelli-Wasserman. We use (8), the Savage density ratio. Comparison of either the spatial model or the heterogeneity model to the full model requires one-dimensional kernel density estimation—we use the S-PLUS function “*density*” (Venables & Ripley, 1998) with smoothing parameter chosen as in the previous example. For comparing the model without any variance component against that with both of them, we require two-dimensional kernel estimation. A built-in S-PLUS function “*kde*” performs the task. As before, we use the default normal kernel and Silverman's (1986) recommended bandwidth.

Results

Knowing the correct values would help comparing the Bayes factor estimates obtained by the different methods. Of course, if there were an easy way to obtain the correct value, we would not need to explore the different approaches. However, the approach that is taken to obtain a “gold standard” is to use the importance sampling method with a huge sample, of size 1 million. Obtaining the true values of the three Bayes factors on an Alpha station 500 workstation equipped with a 400MHz 64-bit CPU and a gigabyte of RAM require 620, 421, and 381 minutes respectively. By looking at the variability of the importance ratios for the sampled 1 million points, we conclude that the Bayes factor is determined up to a standard error of about 0.5% for the Bayes factor comparing the spatial model to the full model and about 0.25% for the other two Bayes factors.

We run the programs to compute estimates of each of the Bayes factors using the computational methods 30 times each using different random number seeds and calculate the average and standard deviation of those 30 Bayes factor estimates obtained. Table 3 shows the values. The table also shows the CPU time taken for one computation of the Bayes factor estimate by each of these methods in an Alpha station 500 workstation equipped with 400MHz 64-bit CPU and a gigabyte of RAM.

Table 3: Estimates of Bayes Factors for the Scotland Lip Cancer Data Set

Models	Method	Estimated BF	Std. dev.	CPU time (min)	
spatial	True value	1.44	-	620	
	vs. Chib-Jeliazkov	1.46	0.1463	81.1	
	full	Importance sampling	1.44	0.0578	30.5
		Verdinelli-Wasserman	0.44	0.21	7.1
heterogen.	True value	0.083	-	421	
	vs. Chib-Jeliazkov	0.083	0.0093	57.2	
	full	Importance sampling	0.083	0.0023	20.6
		Verdinelli-Wasserman	0.021	0.012	5.4
null	True value	1.15×10^{-23}	-	381	
	vs. Chib-Jeliazkov	1.21×10^{-23}	1.46×10^{-24}	48.1	
	full	Importance sampling	1.16×10^{-23}	2.81×10^{-25}	18.2
		Verdinelli-Wasserman	0.0001	0.004	4.9

The Chib’s method and the importance sampling method (with sample size 50,000) provide accurate values of the Bayes factor estimates. The Verdinelli-Wasserman method does not provide accurate values but our investigations suggest that this is due to the sample size used in deriving kernel estimates. Kernel estimation with considerably larger sample size results in accurate estimates.

The standard deviation of the Bayes factor estimate is much smaller for the importance sampling method than for the Chib’s method. The importance sampling method takes much less programming effort and considerably less time as well. Hence, this method seems to be the best method to use for computation of the Bayes factor estimate for this data set. This is a noteworthy finding in that it runs counter to the conclusions of Han and Carlin (2001). They comment (Han & Carlin, 2001, p. 1131) that

we are inclined to conclude that the marginal likelihood methods (Chib’s) appear to offer a better and safer approach to recommend to practitioners seeking to choose amongst a collection of standard (e.g., hierarchical linear) models.

However, there is one important caveat that may explain the difference in results. The efficiency of Chib’s method is closely related to the efficiency of the underlying MCMC algorithm. Thus, 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 the tailored proposal density (Chib & Jeliazkov, 2001). MCMC implementation is more complex for GLMMs than for the models used by Han and Carlin.

6. Discussion and Recommendations

GLMMs are applied extensively and their use is likely to increase with the widespread availability of fast computational facilities. In many applications of these models, the 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 paper arose as a result of several problems in which formal model comparisons were desired by scientific investigators.

The objective of this work is to learn more about the Bayes factor, the Bayesian tool to perform the hypothesis testing, in this context. The focus of our work is to examine the performance of the different methods for computing the relevant Bayes factor estimates. The main findings of the work follow:

- If a researcher requires an accurate estimate of the Bayes factor, the importance sampling method can be used to obtain the Bayes factor estimate. For both of our examples, the importance sampling method provides very accurate and precise estimates of the Bayes factor in reasonable time. Chib's method works well, but yielded higher standard deviations for both examples in this article.
- Chib's method performs satisfactorily and is least sensitive to the size of the problem. This method performs well even with a large number of random effects. The use of sophisticated computational techniques, e.g., the tailored proposal density (Chib & Jeliazkov, 2001) may be used to improve the convergence rate of the Markov chain, which will in turn reduce the variability of the Bayes factor estimate for a given number of iterations.
- The computation of the marginal likelihood is a nontrivial task for GLMMs because it involves integrating out the random effects. If the model is simple and the data set is

small, it is possible to apply numerical integration. For larger problems, importance sampling is a possible approach; our work finds that the importance sampling with a t_4 sampling distribution works well.

- The computation of the Bayes factor involves integrating over the parameters ω in $p(\mathbf{y}|\omega, M)$. Typically, for a GLMM, the parameter vector ω consists of the regression parameters α and the variance component parameters θ . However, in computing Bayes factors for GLMMs, including the random effects in the parameter vector is often convenient. This trick makes the application of some of the methods (e.g., importance sampling method) possible for the second example.

References

- Albert, J., & Chib, S. (1997). Bayesian tests and model diagnostics in conditionally independent hierarchical models. *Journal of the American Statistical Association*, *92*, 916-925.
- Besag, J., York, J., & 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., & Chib, S. (1995). Bayesian model choice via Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B*, *57*, 473-484.
- Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, *90*, 1313-1321.
- Chib, S., & Jeliazkov, I. (2001). Marginal likelihood from the Metropolis-Hastings output. *Journal of the American Statistical Association*, *96*, 270-281.
- Clayton, D., & 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., & 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.
- Dellaportas, P., Forster, J. J., & Ntzoufras, I. (2002). On Bayesian model and variable selection using MCMC. *Statistics and Computing*, *12*, 27-36.
- Diciccio, T. J., Kass, R. E., Raftery, A., & Wasserman, L. (1997). Computing Bayes factors by combining simulation and asymptotic approximations. *Journal of the American Statistical Association*, *92*, 903-915.
- Gelfand, A. E., & Carlin, B. P. (1993). Maximum-likelihood estimation for constrained or missing-data models. *The Canadian Journal of Statistics*, *21*, 303-311.

- Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. New York: Chapman & Hall.
- Geyer, A. E., & Thompson, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data. *Journal of the Royal Statistical Society, Series B, Methodological*, *54*, 657-683.
- Godsill, S. J. (2001). On the relationship between Markov chain Monte Carlo methods for model uncertainty. *Journal of Computational and Graphical Statistics*, *10*, 230-248.
- Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, *82*, 711-732.
- Han, C., & 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., & Paukstis, G. L. (2000). Experimental analysis of an early life history stage: Selection on size of hatchling turtles. *Ecology*, *81*, 2290-2304.
- Jeffreys, H. (1961). *Theory of probability*. Oxford, U.K.: Oxford University Press.
- Kass, R. E., & 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., & Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statistica Sinica*, *6*, 831-860.
- Newton, M. A., & Raftery, A. E. (1994). Approximate Bayesian inference by the weighted likelihood bootstrap (with discussion). *Journal of the Royal Statistical Society, Series B*, *56*, 3-48.
- Pauler, D. K., Wakefield, J. C., & Kass, R. E. (1999). Bayes factors for variance component models. *Journal of the American Statistical Association*, *94*, 1242-1253.

- Satagopan, J. M., Newton, M., & Raftery, A. E. (2000). Easy estimation of normalizing constants and Bayes factors from posterior simulation: Stabilizing the harmonic mean estimator (Technical report no. 382). Seattle: University of Washington.
- Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. London: Chapman and Hall.
- 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., & Stern, H. (2002). On the sensitivity of Bayes factors to the prior distributions. *The American Statistician*, 56(3), 196-201.
- Stern, H. S., & Cressie, N. (1995). Bayesian and constrained Bayesian inference for extremes in epidemiology. *ASA Proceedings of the Epidemiology Section*, 11-20.
- Tierney, L., & Kadane, J. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81, 82-86.
- Venables, W. N., & Ripley, B. D. (1998). *Modern applied statistics with S-PLUS*. New York: Springer-Verlag.
- Verdinelli, I., & Wasserman, L. (1995). Computing Bayes factors using the Savage-Dickey density ratio. *Journal of the American Statistical Association*, 90, 614-618.