Default Bayesian Model Determination Methods for Generalised Linear Mixed Models

Antony M. Overstall[1], Jonathan J. Forster[2]

Abstract

A default strategy for fully Bayesian model determination for GLMMs is considered which addresses the two key issues of default prior specification and computation. In particular, the concept of unit information priors is extended to the parameters of a GLMM. A combination of MCMC and Laplace approximations is used to compute approximations to the posterior model probabilities to find a subset of models with high posterior model probability. Bridge sampling is then used on the models in this subset to approximate the posterior model probabilities more accurately. The strategy is applied to four examples.

Keywords: unit information priors, bridge sampling, MCMC, Laplace approximation

1. Introduction

Generalised linear mixed models (GLMMs) extend generalised linear models (GLMs) to responses which are correlated due to the existence of groups or clusters, by the inclusion of group-specific parameters (known as random effects in classical statistics). For example, in a longitudinal study we record several observations from the same individual. GLMs, linear mixed models (LMMs), and linear models (LMs) are all special cases of GLMMs.

1.1. Specification of a GLMM

Let y_{ij} be the jth response from the ith group where $j = 1, ..., n_i$ and $i = 1, ..., G$. Let $\mathbf{x}_{ij}$ and $\mathbf{z}_{ij}$ denote the $p \times 1$ and $q \times 1$ vectors of covariates which correspond to the

[1]Corresponding Author. E-mail: amo105@soton.ac.uk. Address: Southampton Statistical Sciences Research Institute (S3RI), University of Southampton, Highfield, Southampton, UK, SO17 1BJ.
[2]Address: School of Mathematics, University of Southampton, Highfield, Southampton, UK, SO17 1BJ.

regression and group-specific parameters, respectively. Assume that the components of $\mathbf{z}_{ij}$ form a subset of the components of $\mathbf{x}_{ij}$. Let the total sample size be $n = \sum_{i=1}^{G} n_i$. Conditional on the group-specific parameters, $\mathbf{u}_i$, we assume that y_{ij} is independently distributed from some exponential family distribution with density

$$f(y_{ij}|\mathbf{u}_i) = \exp\left[\frac{y_{ij}\zeta_{ij} - b(\zeta_{ij})}{a_{ij}(\phi)} + c(y_{ij}, \phi)\right], \tag{1}$$

where ζ_{ij} is the *canonical parameter*, ϕ is the *dispersion parameter*, and $a_{ij}()$, $b()$, and $c()$ are known functions. Define $\mu_{ij} = \mathrm{E}(y_{ij}|\mathbf{u}_i) = b'(\zeta_{ij})$ as the conditional mean of y_{ij}. This is related to the *linear predictor*, η_{ij}, through

$$g(\mu_{ij}) = \eta_{ij} = \mathbf{x}_{ij}^T\boldsymbol{\beta} + \mathbf{z}_{ij}\mathbf{u}_i, \tag{2}$$

where $g()$ is the *link function*, $\boldsymbol{\beta}$ is a $p \times 1$ vector of *regression parameters*, and $\mathbf{u}_i$ is a $q \times 1$ vector of *group-specific parameters*.

Suppose, for the ith group, that $\mathbf{y}_i = (y_{i1}, ..., y_{in_i})^T$, $\mathbf{X}_i = (\mathbf{x}_{i1}, ..., \mathbf{x}_{in_i})^T$, $\mathbf{Z}_i = (\mathbf{z}_{i1}, ..., \mathbf{z}_{in_i})^T$, $\boldsymbol{\eta}_i = (\eta_{i1}, ..., \eta_{in_i})^T$, $\boldsymbol{\mu}_i = (\mu_{i1}, ..., \mu_{in_i})^T$, and that the link function is applied elementwise, then

$$g(\boldsymbol{\mu}_i) = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{u}_i.$$

Suppose further that $\mathbf{y} = (\mathbf{y}_1^T, ..., \mathbf{y}_G^T)^T$, $\mathbf{X} = (\mathbf{X}_1^T, ..., \mathbf{X}_G^T)^T$, $\mathbf{Z} = \mathrm{diag}(\mathbf{Z}_1, ..., \mathbf{Z}_G)$, $\boldsymbol{\eta} = (\boldsymbol{\eta}_1^T, ..., \boldsymbol{\eta}_G^T)^T$, $\boldsymbol{\mu} = (\boldsymbol{\mu}_1^T, ..., \boldsymbol{\mu}_G^T)^T$, and $\mathbf{u} = (\mathbf{u}_1^T, ..., \mathbf{u}_G^T)^T$, then (2) can be rewritten in matrix form as

$$g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}.$$

We make the assumption that the first columns of $\mathbf{X}_i$ and $\mathbf{Z}_i$ (if non-zero) are always formed from a vector, of length n_i, of ones. We also assume that the columns of $\mathbf{Z}_i$ are a subset of the columns of $\mathbf{X}_i$.

We complete the specification of a GLMM by making the common assumption that $\mathbf{u}_i \overset{\text{iid}}{\sim} \mathrm{N}(\mathbf{0}, \mathbf{D})$, for $i = 1, ..., G$, where the *variance components matrix*, $\mathbf{D}$, is an unstructured $q \times q$ matrix which depends upon the $\frac{1}{2}(q^2 + q) \times 1$ vector of *variance components*, $\mathbf{d}$. If $\mathbf{D}^* = \mathbf{I}_G \otimes \mathbf{D}$, where $\otimes$ denotes the Kronecker product, then $\mathbf{u} \sim \mathrm{N}(\mathbf{0}, \mathbf{D}^*)$.

Our approach will be Bayesian, so we require a joint prior, with density $f(\boldsymbol{\beta}, \mathbf{D}, \phi)$, for the regression parameters, $\boldsymbol{\beta}$, the variance components matrix, $\mathbf{D}$, and the dispersion

parameter, ϕ. Initially, we decompose this prior density as

$$f(\boldsymbol{\beta}, \mathbf{D}, \phi) = f(\boldsymbol{\beta}|\mathbf{D}, \phi) f(\mathbf{D}|\phi) f(\phi). \tag{3}$$

1.2. Bayesian Model Determination for GLMMs

Bayesian model determination for GLMMs proceeds as follows. Suppose model $m \in M$ is defined by the *integrated likelihood*

$$f_m(\mathbf{y}|\boldsymbol{\beta}_m, \mathbf{D}_m, \phi_m) = \int_{\mathbb{R}^{Gq}} f_m(\mathbf{y}|\boldsymbol{\beta}_m, \mathbf{u}_m, \phi_m) f_m(\mathbf{u}_m|\mathbf{D}_m) \mathrm{d}\mathbf{u}_m, \tag{4}$$

where M is a set of models. The *posterior model probability*, $f(m|\mathbf{y})$, of model m is given by

$$f(m|\mathbf{y}) = \frac{f(m) f_m(\mathbf{y})}{\sum_{k \in M} f(k) f_k(\mathbf{y})}, \tag{5}$$

where $f_m(\mathbf{y})$ is the *marginal likelihood* of model m given by

$$f_m(\mathbf{y}) = \int f_m(\mathbf{y}|\boldsymbol{\beta}_m, \mathbf{u}_m, \phi_m) f_m(\mathbf{u}_m|\mathbf{D}_m) f(\boldsymbol{\beta}_m, \mathbf{D}_m, \phi_m) \mathrm{d}\boldsymbol{\beta}_m \mathrm{d}\mathbf{u}_m \mathrm{d}\mathbf{D}_m \mathrm{d}\phi_m, \tag{6}$$

and $f(m)$ is the *prior model probability* of model m. Note that in (6),

$$f_m(\mathbf{y}|\boldsymbol{\beta}_m, \mathbf{u}_m, \phi_m) = \prod_{i=1}^{G} \prod_{j=1}^{n_i} \exp\left[\frac{y_{ij}\zeta_{mij} - b_m(\zeta_{mij})}{a_{mij}(\phi_m)} + c_m(y_{ij}; \phi_m) \right], \tag{7}$$

is known as the *first-stage likelihood*.

It is common to adopt a uniform prior for m, i.e. $f(m) = \frac{1}{|M|}$, and this is what is used for the remainder of this paper. However, there do exist alternative approaches, such as multiplicity correction priors (see, for example, Scott & Carvalho (2008)).

Suppose we are comparing two models, labelled 1 and 2, say, with posterior model probabilities $f(1|\mathbf{y})$ and $f(1|\mathbf{y})$, respectively. Consider the posterior odds in favour of model 1

$$\frac{f(1|\mathbf{y})}{1 - f(1|\mathbf{y})} = \frac{f(1|\mathbf{y})}{f(2|\mathbf{y})} = \frac{f(1) f_1(\mathbf{y})}{f(2) f_2(\mathbf{y})} = \frac{f(1)}{1 - f(1)} \frac{f_1(\mathbf{y})}{f_2(\mathbf{y})}.$$

The ratio $f_1(\mathbf{y})/f_2(\mathbf{y})$ is known as the *Bayes factor* in favour of model 1. Kass & Raftery (1995) provide a comprehensive review of Bayes factors, including how to interpret them.

Posterior model probabilities and Bayes factors represent the gold standard in fully Bayesian model determination. In Section 1.3 we discuss how these quantities are sensitive to the choice of prior distribution in the case of specifying a default prior under

weak prior information. There exist methods of model determination which rely on the Bayesian approach but do not give posterior model probabilities. However, as such, the issue of default prior specification is avoided. These include criterion-based methods such as BIC or DIC (Spiegelhalter et al (2002)). Aitkin et al (2009) proposed a method based on posterior deviances for model determination applied to small area estimation.

1.3. Our Aim

Our aim is to develop an automatic, fully Bayesian analysis of GLMMs with regards to model determination under weak prior information. This needs to address the two key issues of default prior specification and computation, while minimising the need for choosing arbitrary values of prior hyperparameters.

Lindley's paradox (see, for example, O'Hagan & Forster (2004) pgs 77-79) dictates that we cannot simply choose a uniform or an arbitrarily diffuse informative prior for the model parameters since a fully Bayesian model selection method will tend to favour the model with smallest dimension. In specifying prior distributions for the model parameters, we aim to calibrate the amount of information they provide to make consistent model comparisons. In Section 2, we introduce a generalisation of the approximate unit information prior for the regression parameters, $\boldsymbol{\beta}$. In Section 3, we discuss some of the priors for the variance components matrix, $\mathbf{D}$, that exist in the literature, before introducing a conjugate inverse-Wishart prior with hyperparameter choice based on a unit information concept. There remains a choice for the prior distribution for the dispersion parameter, ϕ. The dispersion parameter is one for responses from the binomial and Poisson distributions. We focus on these examples in this paper and therefore do not consider a prior for ϕ.

The integral (6) is generally analytically intractable and requires approximation. Suitable approximation methods include importance sampling and bridge sampling. Bridge sampling, in particular, was found by Sinharay & Stern (2005) to provide very accurate approximations to the marginal likelihoods for GLMMs. A potential problem with using this approach, solely, is that the number of models, $|M|$, may be large thus rendering bridge sampling for each model impractical. In this case, it may be necessary to use Markov Chain Monte Carlo (MCMC) methods to approximate the posterior model probabilities directly, i.e. not through (5). Approaches to computation are considered in

Section 4. Our computational approach is presented for the more general case of when ϕ is unknown. It is easy to modify for when ϕ is known.

In Section 5, we assess the efficacy and robustness of the model determination strategy using simulations where the responses are generated from the Poisson and Bernoulli distributions.

In Section 6, we demonstrate the model determination strategy on four examples.

2. Prior for the Regression Parameters, $\boldsymbol{\beta}$

The regression parameters, $\boldsymbol{\beta}$, are typically the most important parameters with respect to inference. Chen et al (2003) proposed an informative prior for $\boldsymbol{\beta}$ in a GLMM which uses historical data. However, this is inappropriate for the situation we consider here of weak prior information.

In this section, we extend the concept of *unit-information priors* to the regression parameters, $\boldsymbol{\beta}$, of a GLMM. Previously, versions of these priors have been applied to linear models (Smith & Spiegelhalter (1980) and Kass & Wasserman (1995)), linear mixed models (Pauler (1998)) and generalised linear models (Ntzoufras et al (2003)).

We define a unit information prior for $\boldsymbol{\beta}$ as the multivariate normal distribution with mean $\mathbf{m}$ and $\boldsymbol{\Sigma}$, i.e.

$$\boldsymbol{\beta} \sim \mathrm{N}(\mathbf{m}, \boldsymbol{\Sigma}),$$

for particular choices of $\mathbf{m}$ and $\boldsymbol{\Sigma}$. We follow Raftery (1996) and Ntzoufras et al (2003), and choose the prior mean as $\mathbf{m} = (m_0, 0, ..., 0)^T$. Typically, $m_0 = 0$, however for Bernoulli responses and the complementary log-log link function we may want to choose $m_0 = \log(\log(2))$ to correspond to a mean response of $\frac{1}{2}$. The variance matrix, $\boldsymbol{\Sigma}$, is chosen to approximately provide the same amount of information as one unit of data. Consider the linear model: $\mathbf{y} \sim \mathrm{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$. The Fisher information is given by

$$\mathcal{I}_{\boldsymbol{\beta}} = \frac{1}{\sigma^2} \mathbf{X}^T \mathbf{X}.$$

A unit of data in this case is one observation, so the average amount of information provided by one observation is $\frac{1}{n\sigma^2} \mathbf{X}^T \mathbf{X}$, and therefore

$$\boldsymbol{\Sigma} = n\sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}. \tag{8}$$

Consider a GLM, the Fisher information is given by

$$\mathcal{I}_{\boldsymbol{\beta}} = \mathbf{X}^T \mathbf{W}^{-1} \mathbf{X},$$

where $\mathbf{W} = \text{diag}\{\text{var}(y_i) g'(\mu_i)^2\}$, where $g()$ is the link function. In this case, the Fisher information depends upon the unknown regression parameters, $\boldsymbol{\beta}$. Ntzoufras et al (2003) proposed replacing $\boldsymbol{\beta}$ by its prior mean $\mathbf{m}$. Therefore,

$$\boldsymbol{\Sigma} = n(\mathbf{X}^T \mathbf{W}_{\mathbf{m}}^{-1} \mathbf{X})^{-1},$$

where $\mathbf{W}_{\mathbf{m}} = \text{diag}\{\text{var}(y_i|\boldsymbol{\beta} = \mathbf{m}) \ (g'(\mu_i)|_{\boldsymbol{\beta}=\mathbf{m}})^2\}$.

Pauler (1998) proposed a unit information concept prior distribution for $\boldsymbol{\beta}$ for linear mixed models. To achieve this, the group-specific parameters, $\mathbf{u}$, are integrated out to give the integrated likelihood. The Fisher information for $\boldsymbol{\beta}$ is then

$$\mathcal{I}_{\boldsymbol{\beta}} = \mathbf{X}^T \mathbf{V}^{-1} \mathbf{X},$$

where $\mathbf{V} = \sigma^2 \mathbf{I}_n + \mathbf{Z}\mathbf{D}\mathbf{Z}^T$. In both the LM and GLM cases, we divide the Fisher information by the sample size. However, Pauler (1998) states, that in mixed models, the sample size "is ambiguous because of correlations between observations". Pauler (1998) defines the effective sample size, N_k, for β_k, the kth element of $\boldsymbol{\beta}$, to be the order of the Fisher information for β_k, i.e. the order of the kth diagonal element of the Fisher information matrix. Let

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda} \mathcal{I}_{\boldsymbol{\beta}}^{-1} \boldsymbol{\Lambda}, \tag{9}$$

where $\boldsymbol{\Lambda} = \text{diag}\left\{\sqrt{N_k}\right\}$ for $k = 1, ..., p$. Pauler (1998) shows that for an LMM

$$N_j = \begin{cases} G, & \text{if } \beta_j \text{ has an associated group-specific parameter,} \\ n, & \text{otherwise.} \end{cases}$$

Note that the unit information prior distribution for $\boldsymbol{\beta}$ of Pauler (1998) is conditional on the variance components matrix, $\mathbf{D}$.

In the case of a linear model, none of the regression parameters have associated group-specific parameters and so $N_k = n$, for all k. Therefore, (9) reduces to the appropriate variance matrix of a unit information prior for a linear model, given in (8).

We could generalise the unit information prior distribution proposed by Pauler (1998) to GLMMs by using a deterministic approximation for the Fisher information of $\boldsymbol{\beta}$ from

the integrated likelihood of a GLMM (see, for example, Breslow & Clayton (1993)). However, this would result in a prior distribution for $\boldsymbol{\beta}$ that is conditional on $\mathbf{D}$. In Section 3, we propose an inverse-Wishart prior distribution for $\mathbf{D}$ that is based on the first-stage likelihood, (7), using a unit information concept. It would be logical, therefore, to also base our unit information prior distribution for the regression parameters, $\boldsymbol{\beta}$, on the first-stage likelihood. Also, we find that a unit information prior for $\boldsymbol{\beta}$ based on the first-stage likelihood is independent of $\mathbf{D}$. There are certain computational advantages of using a prior distribution for $\boldsymbol{\beta}$ that is independent of $\mathbf{D}$. We discuss these advantages in Section 4. We, therefore, propose a unit information prior distribution for $\boldsymbol{\beta}$ that is based on the Fisher information from the first-stage likelihood.

From (7), the first-stage likelihood, having dropped the subscript m, is

$$f(\mathbf{y}|\boldsymbol{\beta}, \mathbf{u}, \phi) = \prod_{i=1}^{G} \prod_{j=1}^{n_i} \exp\left[\frac{y_{ij}\zeta_{ij} - b(\zeta_{ij})}{a_{ij}(\phi)} + c(y_{ij}; \phi)\right].$$

Therefore the Fisher information from the first-stage likelihood is

$$\mathcal{I}_{\boldsymbol{\beta}} = \mathbf{X}^T \mathbf{W}^{-1} \mathbf{X},$$

where $\mathbf{W} = \text{diag}\left\{\text{var}(y_{ij}) g'(\mu_{ij})^2\right\}$. Since, conditional on $\mathbf{u}_i$, the y_{ij}'s are independent, the effective sample size for β_k is N for all k, where N is the order of the diagonal elements of $\mathbf{X}^T \mathbf{W}^{-1} \mathbf{X}$. Note that $\mathcal{I}_{\boldsymbol{\beta}}$ depends on the unknown $\boldsymbol{\beta}$ and $\mathbf{u}$ through $\mathbf{W}$, so we follow Ntzoufras et al (2003) and replace $\boldsymbol{\beta}$ and $\mathbf{u}$ by their prior means of $\mathbf{m}$ and $\mathbf{0}$, respectively. Therefore,

$$\boldsymbol{\Sigma} = N\left(\mathbf{X}^T \mathbf{W}_{\mathbf{m},\mathbf{0}}^{-1} \mathbf{X}\right)^{-1},$$

where $\mathbf{W}_{\mathbf{m},\mathbf{0}} = \text{diag}\left\{\text{var}(y_{ij}) g'(\mu_{ij})^2\right\}\Big|_{\boldsymbol{\beta}=\mathbf{m},\mathbf{u}=\mathbf{0}}$. Note that this prior variance is identical to that we would use for the corresponding GLM. This prior is independent of $\mathbf{D}$. It follows that N is the order of $\sum_{i=1}^{G} \sum_{j=1}^{n_i} \frac{1}{w_{ij}}$, where w_{ij} are the diagonal elements of $\mathbf{W}_{\mathbf{m},\mathbf{0}}$.

Note that $\mathbf{W}_{\mathbf{m},\mathbf{0}} = \text{diag}\left\{\text{var}(y_{ij}) g'(\mu_{ij})^2\right\}\Big|_{\boldsymbol{\beta}=\mathbf{m},\mathbf{u}=\mathbf{0}}$ can often be written as $\tau^2 \mathbf{I}_n$ or $\text{diag}\left\{\tau_i^2 \mathbf{I}_{n_i}\right\}$. For instance, suppose that $y_{ij} \sim \text{Bernoulli}(\mu_{ij})$, where $\text{logit}(\mu_{ij}) = \eta_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + \mathbf{z}_{ij}^T \mathbf{u}_i$, where $\mathbf{u}_i \overset{\text{iid}}{\sim} \text{N}(\mathbf{0}, \mathbf{D})$. Suppose, we have chosen $\text{E}(\boldsymbol{\beta}) = \mathbf{m} = \mathbf{0}$ as the prior mean of $\boldsymbol{\beta}$. Then, $\text{var}(y_{ij}) = \mu_{ij}(1 - \mu_{ij}) = \frac{\exp(\eta_{ij})}{(1+\exp(\eta_{ij}))^2}$ and $g'(\mu_{ij}) = \frac{1}{\mu_{ij}^2 (1-\mu_{ij})^2}$.

Therefore,

$$\mathbf{W} = \text{diag}\left\{\frac{1}{\mu_{ij}(1-\mu_{ij})}\right\} = \text{diag}\left\{\frac{(1+\exp(\eta_{ij}))^2}{\exp(\eta_{ij})}\right\},$$

and $\mathbf{W_{m,0}} = \text{diag}\left\{\frac{(1+\exp(\eta_{ij}))^2}{\exp(\eta_{ij})}\right\}\Big|_{\boldsymbol{\beta}=\mathbf{0},\mathbf{u}=\mathbf{0}} = 4\mathbf{I}_n$. So in this example, $\tau^2 = 4$ and $N = n$, the sample size.

For each of the four examples in Section 6, we explain the values of τ^2 and N.

3. Prior for the Variance Components Matrix, D

There is a large literature on default prior distributions for the variance components matrix, $\mathbf{D}$.

Natarajan & Kass (2000) defined an approximate generalisation of the uniform shrinkage prior of Daniels (1999) for GLMMs. A similar prior was suggested by Gustafson et al (2006) where the variance components matrix can be written as $\mathbf{D} = \sigma^2\boldsymbol{\Omega}$, where $\boldsymbol{\Omega}$ is a known positive-definite matrix and σ^2 is unknown. This is different to the setup we consider since in our case, $\mathbf{D}$ is unstructured. Kass & Natarajan (2006) proposed a conjugate inverse-Wishart distribution as a default prior for $\mathbf{D}$. The priors of Natarajan & Kass (2000), Gustafson et al (2006), and Kass & Natarajan (2006) are all data dependent as they rely on the maximum likelihood estimate of $\boldsymbol{\beta}$.

Cai & Dunson (2006) define a prior for the variance components matrix where $\mathbf{D}$ is decomposed as $\mathbf{D} = \mathbf{L}\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T\mathbf{L}$, to ensure positive-definiteness, $\mathbf{L} = \text{diag}(\lambda_1, ..., \lambda_q)$ with $\lambda_k \geq 0$, and $\boldsymbol{\Gamma}$ is lower triangular, with off diagonal elements, γ_{ij}. Zero-inflated positive normal distributions are then placed on the λ_k's and zero-inflated normal distributions are then placed on the γ_{ij} 's.

A completely different approach is taken by Garcia-Donato & Sun (2007) in their divergence-based (DB) priors for comparing between the following two models:

1. $y_{ij} \sim \text{N}(\mu, \sigma^2)$, where $i = 1, ..., G$ and $j = 1, ..., n^*$,
2. $y_{ij} \sim \text{N}(\mu + u_i, \sigma^2)$, where $u_i \sim \text{N}(0, \tau^2)$.

The DB prior for τ^2 is

$$f(\tau^2|\mu, \sigma^2) \propto \left[1 + \frac{D(\mu, \sigma^2, \tau^2)}{n^*G}\right]^{-g}, \quad g > g^*,$$

where g^* is the minimum value such that the DB prior is proper if $g > g^*$, and $D(\mu, \sigma^2, \tau^2)$ is the Kullback-Liebler divergence between models 1 and 2. Note that the divergence is divided by the sample size n^*G thus linking to the idea of unit information which is central to the priors developed in this and the previous section.

We define a default prior distribution for $\mathbf{D}$ as the inverse-Wishart distribution with ρ degrees of freedom and scale matrix, $\rho\mathbf{R}$, i.e. $\mathbf{D} \sim \text{IW}(\rho, \rho\mathbf{R})$, where $\rho > q - 1$. Following Kass & Natarajan (2006), we set $\rho = q$. Kass & Natarajan (2006) give the Fisher information for $\mathbf{u}_i$ from the first-stage likelihood as

$$\mathcal{I}_{\mathbf{u}_i} = \mathbf{Z}_i^T \mathbf{W}_i^{-1} \mathbf{Z}_i,$$

for $i = 1, ..., G$, where $\mathbf{W}_i = \text{diag}\left\{\text{var}(y_{ij}) g'(\mu_{ij})^2\right\}$. Following the approach in Section 2 we replace the unknown parameters, $\boldsymbol{\beta}$ and $\mathbf{u}_i$ by their prior means, $\mathbf{m}$ and $\mathbf{0}$, respectively. The approximate average unit information over the G groups is then

$$\frac{1}{G} \sum_{i=1}^{G} \frac{1}{n_i} \mathbf{Z}_i^T \mathbf{W}_{i,\mathbf{m},\mathbf{0}}^{-1} \mathbf{Z}_i,$$

where $\mathbf{W}_{i,\mathbf{m},\mathbf{0}} = \text{diag}\left\{\text{var}(y_{ij}) g'(\mu_{ij})^2\right\}\big|_{\boldsymbol{\beta}=\mathbf{m}, \mathbf{u}_i=\mathbf{0}}$. If $\mathbf{D}$ was a fixed hyperparameter then we would set $\mathbf{D}^{-1} = \frac{1}{G} \sum_{i=1}^{G} \frac{1}{N_i} \mathbf{Z}_i^T \mathbf{W}_{i,\mathbf{m},\mathbf{0}}^{-1} \mathbf{Z}_i$, to give a unit information prior for $\mathbf{u}$. However, since $\mathbf{D}$ is not fixed we set

$$\text{E}(\mathbf{D}^{-1}) = \mathbf{R}^{-1} = \frac{1}{G} \sum_{i=1}^{G} \frac{1}{N_i} \mathbf{Z}_i^T \mathbf{W}_{i,\mathbf{m},\mathbf{0}}^{-1} \mathbf{Z}_i.$$

Here N_i is the order of the diagonal elements of $\mathbf{Z}_i^T \mathbf{W}_{i,\mathbf{m},\mathbf{0}}^{-1} \mathbf{Z}_i$ and therefore the order of $\sum_{j=1}^{n_i} \frac{1}{w_{ij}}$ where w_{ij} are the diagonal elements of $\mathbf{W}_{i,\mathbf{m},\mathbf{0}}$. Therefore

$$\mathbf{R} = G \left(\sum_{i=1}^{G} \frac{1}{N_i} \mathbf{Z}_i^T \mathbf{W}_{i,\mathbf{m},\mathbf{0}}^{-1} \mathbf{Z}_i \right)^{-1}.$$

Note that, in the case of equal group sizes and where $O(1) = \frac{1}{w_{ij}}$, so that $N_i = n_i = n^*$, and if the maximum likelihood estimate of $\boldsymbol{\beta}$ is replaced by $\mathbf{m}$, then this prior is the default conjugate prior of Kass & Natarajan (2006) with $c = n^*$,

When $q = 1$, i.e. when $u_i \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2)$, then the proposed prior distribution reduces to

$$\sigma^2 \sim \text{IG}\left(\frac{1}{2}, \frac{G}{2} \left(\sum_{i=1}^{G} \frac{1}{N_i} \sum_{j=1}^{n_i} \frac{1}{w_{ij}} \right)^{-1} \right),$$

where $\mathbf{w}_i = (w_{i1}, ..., w_{in_i})^T$ are the diagonal elements of $\mathbf{W}_{i,\mathbf{m},\mathbf{0}}$.

We explain the values of N_i for the four examples in Section 6.

4. Computation

4.1. General Strategy

In this section, we describe the computational strategy and methods to approximate the posterior model probabilities. Sinharay & Stern (2005) found that bridge sampling provided very accurate approximations to the Bayes' factors for comparing GLMMs with respect to minimising the standard errors, when compared to importance sampling, Chib's method (from the marginal likelihood identity, see Chib (1995)) and reversible jump (Green (1995)). Bridge sampling, given a sample from the posterior distribution, is an easily implemented method for approximating the marginal likelihood of a given model. Evaluating the marginal likelihood, by approximation or exactly, of every model $m \in M$ to find the posterior model probabilities is called the *marginal likelihood approach*. However, if the number of models, $|M|$, is large, the marginal likelihood approach becomes impractical. A more suitable approach, therefore, is a "one-shot" implementation of an MCMC method such as reversible jump (Green (1995)). The disadvantage of such a method is making effective proposals which is made more acute by the large differences in dimensionality between models we consider.

As a compromise we propose the following general strategy. We use a simple deterministic Laplace approximation to the integrated likelihood (4) to reduce the dimension of the parameter space. We then use an independence sampler which is a special case of the reversible jump MCMC method to approximate the posterior model probabilities of all models $m \in M$. These approximations, denoted as $\hat{f}^L(m|\mathbf{y})$, are used to identify a smaller set of candidate models, $M' \subset M$. Finally, bridge sampling is used to approximate the posterior model probabilities of the models $m \in M'$. Denote the bridge sampling approximations to the marginal likelihood and posterior model probabilities of model m by $\hat{f}^B_m(\mathbf{y})$ and $\hat{f}^B(m|\mathbf{y})$, respectively.

4.2. An MCMC Method

We ease the computational burden by taking advantage of the conditional independence of the $\mathbf{y}_i$, and write the integrated likelihood (removing the subscript m) as

$$f(\mathbf{y}|\boldsymbol{\beta}, \mathbf{D}, \phi) = \prod_{i=1}^{G} \int_{\mathbb{R}^q} f(\mathbf{y}_i|\boldsymbol{\beta}, \mathbf{u}_i, \phi) f(\mathbf{u}_i|\mathbf{D}) \mathrm{d}\mathbf{u}_i. \tag{10}$$

Using the Laplace approximation for each $\int f(\mathbf{y}_i|\boldsymbol{\beta}, \mathbf{u}_i, \phi) f(\mathbf{u}_i|\mathbf{D}) \mathrm{d}\mathbf{u}_i$, we obtain the following approximation to the integrated likelihood:

$$\hat{f}(\mathbf{y}|\boldsymbol{\beta}, \mathbf{D}, \phi) = |\mathbf{D}|^{-\frac{G}{2}} \prod_{i=1}^{G} \left[f(\mathbf{y}_i|\boldsymbol{\beta}, \hat{\mathbf{u}}_i, \phi) |\mathbf{V}_i + \mathbf{D}^{-1}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} \hat{\mathbf{u}}_i^T \mathbf{D}^{-1} \hat{\mathbf{u}}_i \right) \right], \tag{11}$$

where $\mathbf{V}_i = -\frac{\partial^2}{\partial \mathbf{u}_i \partial \mathbf{u}_i^T} \log f(\mathbf{y}_i|\boldsymbol{\beta}, \mathbf{u}_i, \phi) \Big|_{\mathbf{u}_i = \hat{\mathbf{u}}_i}$. The value, $\hat{\mathbf{u}}_i$, of $\mathbf{u}_i$ that maximises the integrand in (10), or equivalently $\log f(\mathbf{y}_i|\boldsymbol{\beta}, \mathbf{u}_i, \phi) f(\mathbf{u}_i|\mathbf{D})$, can be found using the Newton-Raphson method since the 1st and 2nd derivatives of $\log f(\mathbf{y}_i|\boldsymbol{\beta}, \mathbf{u}_i, \phi) f(\mathbf{u}_i|\mathbf{D})$ with respect to $\mathbf{u}_i$ are readily available.

Therefore, the approximate posterior density of $\boldsymbol{\beta}$, $\mathbf{D}$, and ϕ is given by:

$$\hat{f}^L(\boldsymbol{\beta}, \mathbf{D}, \phi|\mathbf{y}) \propto \hat{f}^L(\mathbf{y}|\boldsymbol{\beta}, \mathbf{D}, \phi) f(\boldsymbol{\beta}|\mathbf{D}, \phi) f(\mathbf{D}|\phi) f(\phi).$$

Before we consider the MCMC method, we briefly describe the transformations that we use on the variance components and the dispersion parameter.

For the variance components matrix, we use, as the transformation, the Cholesky decomposition $\mathbf{D} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}^T$, where $\boldsymbol{\Gamma}$ is the lower-triangular matrix, which depends upon $\boldsymbol{\upsilon}$, the $\frac{1}{2}q(q+1) \times 1$ vector of transformed parameters, given by

$$\begin{pmatrix} e^{\upsilon_{11}} & & & \\ \upsilon_{12} & e^{\upsilon_{22}} & & \\ \vdots & & \ddots & \\ \upsilon_{1q} & \cdots & & e^{\upsilon_{qq}} \end{pmatrix}.$$

Note that if $\boldsymbol{\upsilon} \in \mathbb{R}^{\frac{1}{2}q(q+1)}$ then $\mathbf{D}$ is guaranteed to be positive-definite. For the dispersion parameter, we use the transformation $\omega = \log \phi$.

The approximate posterior density of the transformed parameters, $(\boldsymbol{\beta}, \boldsymbol{\upsilon}, \omega)^T$, is given by $\hat{f}^L(\boldsymbol{\beta}, \boldsymbol{\upsilon}, \omega|\mathbf{y}) \propto \hat{h}^L(\boldsymbol{\beta}, \boldsymbol{\upsilon}, \omega|\mathbf{y})$, where

$$\hat{h}^L(\boldsymbol{\beta}, \boldsymbol{\upsilon}, \omega|\mathbf{y}) = \hat{f}^L(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\Gamma}\boldsymbol{\Gamma}^T, e^{\omega}) f(\boldsymbol{\beta}|\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T, e^{\omega}) f(\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T|e^{\omega}) f(e^{\omega}) 2^q e^{\omega} \prod_{i=1}^{q} e^{\upsilon_{ii}(q+2-i)},$$

and the Jacobian for the transformation $\mathbf{D} = \mathbf{\Gamma}\mathbf{\Gamma}^T$ is given by, for example, Muirhead (1982, Theorem 2.19). Note that the vector of transformed parameters, $(\boldsymbol{\beta}, \boldsymbol{\upsilon}, \omega)^T$, lies in $\mathbb{R}^{p+\frac{1}{2}q(q+1)+1}$ if the dispersion parameter is unknown and lies in $\mathbb{R}^{p+\frac{1}{2}q(q+1)}$, otherwise.

The MCMC method we propose is the independence sampler (see, for example, O'Hagan & Forster (2004, pg 298)) which is a special case of the reversible jump algorithm where the proposals are made independently of the current state. For model $m \in M$, the proposal distribution, with density $\pi_m(\boldsymbol{\beta}_m, \boldsymbol{\upsilon}_m, \omega_m)$, is the multivariate normal distribution with mean given by the value of $(\boldsymbol{\beta}_m, \boldsymbol{\upsilon}_m, \omega_m)^T$ that maximises the approximate posterior density, $\hat{f}_m^L(\boldsymbol{\beta}_m, \boldsymbol{\upsilon}_m, \omega_m|\mathbf{y}) \propto \hat{h}_m^L(\boldsymbol{\beta}_m, \boldsymbol{\upsilon}_m, \omega_m)$, (or, equivalently, $\log \hat{f}_m^L(\boldsymbol{\beta}_m, \boldsymbol{\upsilon}_m, \omega_m|\mathbf{y}) \propto \log \hat{h}_m^L(\boldsymbol{\beta}_m, \boldsymbol{\upsilon}_m, \omega_m)$) and variance matrix given by the negative, inverse of the approximate Hessian matrix of $\log \hat{f}_m^L(\boldsymbol{\beta}_m, \boldsymbol{\upsilon}_m, \omega_m|\mathbf{y})$ with respect to $(\boldsymbol{\beta}_m, \boldsymbol{\upsilon}_m, \omega_m)^T$ evaluated at the maximum value. These quantities will need to be found numerically. A group of methods for doing so are quasi-Newton methods. Some of these methods are implemented in the statistical software package, `R`, using the function `optim`. Thus the proposal distribution is a normal approximation to the distribution with density $\hat{f}_m^L(\boldsymbol{\beta}_m, \boldsymbol{\upsilon}_m, \omega_m|\mathbf{y})$.

The independence sampler proceeds as follows:

1. Given the current state $(m, \boldsymbol{\beta}_m, \boldsymbol{\upsilon}_m, \omega_m)$, propose a new model m^* with probability $\frac{1}{|M|}$. Then generate proposal model parameters $(\boldsymbol{\beta}_{m^*}^*, \boldsymbol{\upsilon}_{m^*}^*, \omega_{m^*}^*)$ from the distribution with density π_m, as described above.
2. Calculate the acceptance probability, $\alpha = \min(1, a)$, where
$$a = \frac{\hat{h}_{m^*}^L(\boldsymbol{\beta}_{m^*}^*, \boldsymbol{\upsilon}_{m^*}^*, \omega_{m^*}^*|\mathbf{y})\pi_m(\boldsymbol{\beta}_m, \boldsymbol{\upsilon}_m, \omega_m)}{\hat{h}_m^L(\boldsymbol{\beta}_m, \boldsymbol{\upsilon}_m, \omega_m|\mathbf{y})\pi_{m^*}(\boldsymbol{\beta}_{m^*}^*, \boldsymbol{\upsilon}_{m^*}^*, \omega_{m^*}^*)}.$$
3. Accept the proposed move with probability a and set $(m^*, \boldsymbol{\beta}_{m^*}^*, \boldsymbol{\upsilon}_{m^*}^*, \omega_{m^*}^*)$ as the new state. Otherwise, retain $(m, \boldsymbol{\beta}_m, \boldsymbol{\upsilon}_m, \omega_m)$ as the current state.
4. Repeat steps 1) to 3) for a total of B iterations, for large B.

The independence sampler provides $\hat{f}_m^L(m|\mathbf{y})$ for $m \in M$. We identify the smaller set, $M' \subset M$, of candidate models from $\hat{f}_m^L(m|\mathbf{y})$ by using a definition of Madigan & Raftery (1994), in relation to model averaging, of

$$M' = \left\{ m \in M : \max_{k \in M} \hat{f}^L(k|\mathbf{y}) \leq c\hat{f}^L(m|\mathbf{y}) \right\}, \tag{12}$$

for some constant $c > 1$. Larger values of c correspond to a larger number of models in M'. This definition aims to collect most of the posterior model probability without having to consider too large a set of models for M'.

Cai & Dunson (2006) proposed a computational strategy for model determination amongst GLMMs using a SSVS algorithm based on a deterministic approximation of the integrated likelihood. Their approximation was based on a second-order Taylor series expansion of the first-stage likelihood, $f(\mathbf{y}|\boldsymbol{\beta}, \mathbf{u}, \phi)$, whereas the Laplace approximation we use is based on a second-order Taylor series expansion of the log first-stage likelihood, $\log f(\mathbf{y}|\boldsymbol{\beta}, \mathbf{u}, \phi)$.

4.3. Bridge Sampling

Bridge sampling is a method for approximating the marginal likelihood, $f_m(\mathbf{y})$, of model $m \in M$. It requires a sample from the posterior distribution of model $m \in M$. Let $\boldsymbol{\theta}_m$ be the vector of model parameters for model $m \in M$.

The bridge sampling estimator is given by

$$\hat{f}_m^B(\mathbf{y}) = \frac{\frac{1}{N_1}\sum_{i=1}^{N_1} f_m(\mathbf{y}|\tilde{\boldsymbol{\theta}}_m^i) f_m(\tilde{\boldsymbol{\theta}}_m^i)\gamma(\tilde{\boldsymbol{\theta}}_m^i)}{\frac{1}{N_2}\sum_{i=1}^{N_2} g_m(\boldsymbol{\theta}_m^i)\gamma(\boldsymbol{\theta}_m^i)},$$

where $\{\boldsymbol{\theta}_m^i\}_{i=1}^{N_2}$ is a sample of size N_2 from the posterior distribution with density $f_m(\boldsymbol{\theta}_m|\mathbf{y})$, $\{\tilde{\boldsymbol{\theta}}_m^i\}_{i=1}^{N_1}$ is a sample of size N_1 from a distribution with density $g_m(\boldsymbol{\theta}_m)$, and $\gamma()$ is a function that satisfies $0 < |\int g_m(\boldsymbol{\theta}_m)\gamma(\boldsymbol{\theta}_m) f_m(\boldsymbol{\theta}_m|\mathbf{y}) d\boldsymbol{\theta}_m| < \infty$.

Meng & Wong (1996) showed that, with respect to minimising the mean squared error, the optimal $\gamma()$ is given by

$$\gamma^*(\boldsymbol{\theta}_m) = \left[\frac{N_2 f_m(\mathbf{y}|\boldsymbol{\theta}_m) f_m(\boldsymbol{\theta}_m)}{f_m(\mathbf{y})} + N_1 g_m(\boldsymbol{\theta}_m)\right]^{-1}.$$

Of course, $\gamma^*()$ depends on the unknown marginal likelihood, $f_m(\mathbf{y})$, which suggests the following iterative scheme

$$\hat{f}_m^B(\mathbf{y})^{(t+1)} = \frac{\frac{1}{N_1}\sum_{i=1}^{N_1} \frac{\tilde{l}_i}{N_2 \tilde{l}_i + N_1 \hat{f}_m^B(\mathbf{y})^{(t)}}}{\frac{1}{N_2}\sum_{i=1}^{N_2} \frac{1}{N_2 l_i + N_1 \hat{f}_m^B(\mathbf{y})^{(t)}}}, \tag{13}$$

where $\tilde{l}_i = \frac{f_m(\mathbf{y}|\tilde{\boldsymbol{\theta}}_m^i) f_m(\tilde{\boldsymbol{\theta}}_m^i)}{g_m(\tilde{\boldsymbol{\theta}}_m^i)}$ and $l_i = \frac{f_m(\mathbf{y}|\boldsymbol{\theta}_m^i) f_m(\boldsymbol{\theta}_m^i)}{g_m(\boldsymbol{\theta}_m^i)}$. The scheme (13) is iterated until convergence, to give $\hat{f}_m^B(\mathbf{y})$ as the bridge sampling approximation to the marginal likelihood, $f_m(\mathbf{y})$.

Chen et al (2000, pg 129) discuss the allocation of sample sizes, N_1 and N_2. They state that using the optimal choice for $\gamma()$ is often more essential then the optimal allocation of sample sizes. In what follows, we take $N_1 = N_2 = R$.

There remains a choice for the distribution, G_m, with density $g_m(\boldsymbol{\theta}_m)$. From practice, it appears that bridge sampling performs best when $g_m(\boldsymbol{\theta}_m)$ 'mimics' the posterior density, $f_m(\boldsymbol{\theta}_m|\mathbf{y})$. An obvious choice is the normal distribution with it first few moments chosen to match those of the posterior distribution. If the first two moments are matched then this is known as Warp II bridge sampling.

When the posterior distribution is approximately normal then we can find the mode and curvature, at the mode, of the posterior distribution deterministically. We can then set G_m to be the normal distribution with mean equal to the mode and the variance equal to minus the inverse curvature. However, for distributions that are non-normal, we feel that the mode and curvature will provide insufficient information and that the sample mean and variance from a posterior sample will be a better choice.

However, if sample statistics of the entire posterior sample are used, then this leads to correlation between the moments of the distribution with density $g_m()$ and the posterior sample, $\{\boldsymbol{\theta}_m^i\}$, and an apparent underestimation of $f_m(\mathbf{y})$ (see the Appendix).

We propose to use a proportion, ψ, of the posterior sample to estimate the posterior moments. The remainder of the posterior sample can then be used in the bridge sampler, (13). From practice, it appears that $\psi = \frac{1}{2}$ is a robust choice. Therefore, we need a posterior sample of size $2R$.

We now specifically turn our attention to approximating the marginal likelihood of a GLMM, which is

$$f_m(\mathbf{y}) = \int f_m(\mathbf{y}|\boldsymbol{\beta}_m, \mathbf{u}_m, \phi_m) f_m(\mathbf{u}_m|\mathbf{D}_m) \times f(\boldsymbol{\beta}_m|\mathbf{D}_m, \phi_m) f_m(\mathbf{D}_m|\phi_m) f_m(\phi_m) \mathrm{d}\boldsymbol{\beta}_m \mathrm{d}\mathbf{u}_m \mathrm{d}\mathbf{D}_m \mathrm{d}\phi_m,$$

using the prior decomposition in (3). If we use the prior for $\boldsymbol{\beta}$ proposed in Section 2, or any prior for $\boldsymbol{\beta}$ such that $f_m(\boldsymbol{\beta}_m|\mathbf{D}_m, \phi_m) = f_m(\boldsymbol{\beta}_m|\phi_m)$, then

$$f_m(\mathbf{y}) = \int f_m(\mathbf{y}|\boldsymbol{\beta}_m, \mathbf{u}_m, \phi_m) f_m(\boldsymbol{\beta}_m|\phi_m) f_m(\phi_m) \times \int_{\mathbb{P}^q} f_m(\mathbf{u}_m|\mathbf{D}_m) f_m(\mathbf{D}_m|\phi_m) \mathrm{d}\mathbf{D}_m \mathrm{d}\boldsymbol{\beta}_m \mathrm{d}\mathbf{u}_m \mathrm{d}\phi_m,$$

where $\mathbb{P}^q$ is the set of all positive-definite $q \times q$ matrices. Now suppose we adopt the prior for $\mathbf{D}_m$ proposed in Section 3, i.e. $\mathbf{D}_m|\phi_m \sim \text{IW}(q_m, q_m\mathbf{R}_m)$ where $\mathbf{R}_m = G\left(\sum_{i=1}^{G} \frac{1}{n_i}\mathbf{Z}_i^T\mathbf{W}_{i,\mathbf{m},\mathbf{0}}^{-1}\mathbf{Z}_i\right)^{-1}$, where $\mathbf{R}_m$ is a function of ϕ_m through $\mathbf{W}_{\mathbf{m},\mathbf{0}}$, then

$$\int_{\mathbb{P}^q} f_m(\mathbf{u}_m|\mathbf{D}_m) f_m(\mathbf{D}_m|\phi_m) \mathrm{d}\mathbf{D}_m$$

is analytically tractable as

$$\frac{\Gamma_{q_m}\left(\frac{q_m+G}{2}\right)}{\Gamma_{q_m}\left(\frac{q_m}{2}\right)} \frac{1}{\pi^{\frac{Gq_m}{2}}} \frac{|q_m\mathbf{R}_m|^{\frac{q_m}{2}}}{|q_m\mathbf{R}_m + \sum_{i=1}^{G}\mathbf{u}_{mi}\mathbf{u}_{mi}^T|^{\frac{q_m+G}{2}}},$$

where $\Gamma_{q_m}(a) = \pi^{q_m(q_m-1)/4}\prod_{j=1}^{q_m}\Gamma\left(a + \frac{1-j}{2}\right)$ is the *multivariate gamma function*. So the marginal likelihood of a GLMM is now

$$f_m(\mathbf{y}) = \int \frac{\Gamma_{q_m}\left(\frac{q_m+G}{2}\right)}{\Gamma_{q_m}\left(\frac{q_m}{2}\right)\pi^{\frac{Gq_m}{2}}} \frac{|q_m\mathbf{R}_m|^{\frac{q_m}{2}}}{|q_m\mathbf{R}_m + \sum_{i=1}^{G}\mathbf{u}_{mi}\mathbf{u}_{mi}^T|^{\frac{q_m+G}{2}}} \times f_m(\mathbf{y}|\boldsymbol{\beta}_m, \mathbf{u}_m, \phi_m) f_m(\boldsymbol{\beta}_m|\phi_m) f_m(\phi_m) \mathrm{d}\boldsymbol{\beta}_m \mathrm{d}\mathbf{u}_m \mathrm{d}\phi_m.$$

The dispersion parameter is such that $\phi_m > 0$, if it is unknown. We need to transform the parameter to lie on the real line, $\mathbb{R}$. Similar to in Section 4.2, we use the transformation $\phi_m = e^{\omega_m}$. Therefore,

$$f_m(\mathbf{y}) = \int \frac{\Gamma_{q_m}\left(\frac{q_m+G}{2}\right)}{\Gamma_{q_m}\left(\frac{q_m}{2}\right)\pi^{\frac{Gq_m}{2}}} \frac{|q_m\mathbf{R}_m|^{\frac{q_m}{2}}}{|q_m\mathbf{R}_m + \sum_{i=1}^{G}\mathbf{u}_{mi}\mathbf{u}_{mi}^T|^{\frac{q_m+G}{2}}} \times f_m(\mathbf{y}|\boldsymbol{\beta}_m, \mathbf{u}_m, e^{\omega_m}) f_m(\boldsymbol{\beta}_m|e^{\omega_m}) f_m(e^{\omega_m}) e^{\omega_m} \mathrm{d}\boldsymbol{\beta}_m \mathrm{d}\mathbf{u}_m \mathrm{d}\omega_m. \tag{14}$$

We can now apply the bridge sampling approach to approximate the integral (14), with $\boldsymbol{\theta}_m = (\boldsymbol{\beta}_m, \mathbf{u}_m, \omega_m)^T \in \mathbb{R}^{p_m+Gq_m+1}$,

$$f_m(\mathbf{y}|\boldsymbol{\theta}_m) = f_m(\mathbf{y}|\boldsymbol{\beta}_m, \mathbf{u}_m, e^{\omega_m}),$$

and

$$f_m(\boldsymbol{\theta}_m) = \frac{\Gamma_{q_m}\left(\frac{q_m+G}{2}\right)}{\Gamma_{q_m}\left(\frac{q_m}{2}\right)\pi^{\frac{Gq_m}{2}}} \frac{|q_m\mathbf{R}_m|^{\frac{q_m}{2}}}{|q_m\mathbf{R}_m + \sum_{i=1}^{G}\mathbf{u}_{mi}\mathbf{u}_{mi}^T|^{\frac{q_m+G}{2}}} f_m(\boldsymbol{\beta}_m|e^{\omega_m}) f_m(e^{\omega_m}) e^{\omega_m}.$$

We need to generate a posterior sample from $\boldsymbol{\beta}, \mathbf{u}_m, \omega_m|\mathbf{y}$, i.e. the marginal posterior distribution of $\boldsymbol{\beta}, \mathbf{u}_m, \mathbf{D}_m, \omega_m$ with $\mathbf{D}_m$ integrated out. This can be done easily by generating a sample from $\boldsymbol{\beta}, \mathbf{u}_m, \mathbf{D}_m, \phi_m|\mathbf{y}$ and discarding the $\mathbf{D}_m$'s and transforming

ϕ_m to $\omega_m = \log \phi_m$. We discuss how to generate a sample from $\boldsymbol{\beta}, \mathbf{u}_m, \mathbf{D}_m|\mathbf{y}$ in the next section.

The algorithm for approximating the marginal likelihood of a GLMM using bridge sampling is:

1. Generate a sample, $\left\{\boldsymbol{\theta}_m^1, ..., \boldsymbol{\theta}_m^R, \boldsymbol{\theta}_m^{R+1}, ..., \boldsymbol{\theta}_m^{2R}\right\}$, of size $2R$ from the posterior distribution, where $\boldsymbol{\theta}_m^j = (\boldsymbol{\beta}_m^j, \mathbf{u}_m^j, \log \phi_m^j)^T$.
2. Let $\boldsymbol{\mu}_m$ and $\boldsymbol{\Sigma}_m$ be the sample mean and variance of $\left\{\boldsymbol{\theta}_m^{R+1}, ..., \boldsymbol{\theta}_m^{2R}\right\}$. Let G_m be $\mathrm{N}(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$, the $(p_m + Gq_m)$-dimensional normal distribution. Let $g_m()$ be the density function of G_m.
3. Generate a sample, $\left\{\tilde{\boldsymbol{\theta}}_m^1, ..., \tilde{\boldsymbol{\theta}}_m^R\right\}$, of size R from G_m.
4. Approximate $f_m(\mathbf{y})$ using (13), to obtain $\hat{f}_m^B(\mathbf{y})$.

The above algorithm is presented for when the dispersion parameter is unknown. It is easily modified for when the dispersion parameter is known.

4.4. Posterior Simulation

As described in Section 4.3, bridge sampling requires a sample from the posterior distribution of each model. The structure of GLMMs lends itself well to Gibbs sampling due to the conditional independences involved. Zeger & Karim (1991) describe a Gibbs sampling algorithm for GLMMs which relies on rejection sampling.

If the prior distribution of $\boldsymbol{\beta}$ is independent of $\mathbf{D}$ and the prior distribution of $\mathbf{D}$ is the inverse-Wishart distribution then, due to conditional conjugacy, the full conditional distribution of $\mathbf{D}$ is also inverse-Wishart, specifically

$$\mathbf{D}|\mathbf{y}, \boldsymbol{\beta}, \mathbf{u}, \phi \sim \mathrm{IW}\left(q + G, q\mathbf{R} + \sum_{i=1}^{G} \mathbf{u}_i \mathbf{u}_i^T\right).$$

This makes generating from the full conditional distribution of $\mathbf{D}$, in the Gibbs sampler, a trivial task.

We use the statistical software package `WinBUGS` (Lunn et al (2000)) to generate a posterior sample. `WinBUGS` essentially uses the algorithm of Zeger & Karim (1991). We run `WinBUGS` remotely in the statistical software package `R` (R Development Core Team (2008)) using the `R2WinBUGS` package (Sturtz et al (2005)).

5. Simulations

In this section, we assess the efficacy and robustness of our strategy outlined in Sections 2, 3, and 4 by way of a simulation study. Bernoulli and Poisson responses are generated from a GLMM with the canonical link function and linear predictor:

$$\eta_{ij} = (\beta_0 + u_i) + \beta_1 x_{ij}; \text{ where } u_i \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2),$$

and $i = 1, ..., G$ and $j = 1, ..., n_i = n^*$. We generate 1000 datasets with $n = 200$ observations in either $G = 25$ or $G = 4$ groups meaning $n^* = 8$ or $n^* = 50$, respectively. We therefore have four scenarios; two with Bernoulli responses and two with Poisson responses. Within each response distribution, one scenario has $G = 25$ and $n^* = 8$, and one has $G = 4$ and $n^* = 50$. For each dataset, the x_{ij}'s are generated from the standard normal distribution and the intercept parameter, β_0, are held fixed at $\frac{1}{2}$, whereas β_1 and σ^2 are drawn from the U$[0, \frac{5}{4}]$ distribution for Poisson responses and U$[0, 5]$ distribution for Bernoulli responses. Once the responses have been generated, we consider the following five models:

1. $\eta_{ij} = \beta_0$,
2. $\eta_{ij} = \beta_0 + \beta_1 x_{ij}$,
3. $\eta_{ij} = \beta_0 + u_i$; where $u_i \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2)$,
4. $\eta_{ij} = \beta_0 + u_i + \beta_1 x_{ij}$; where $u_i \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2)$,
5. $\eta_{ij} = \beta_0 + u_{0i} + (\beta_1 + u_{1i}) x_{ij}$; where $(u_{0i}, u_{1i})^T \overset{\text{iid}}{\sim} \text{N}(\mathbf{0}, \mathbf{D})$,

where the priors described in Sections 2 and 3 are applied to the appropriate parameters. We then approximate the posterior model probabilities in two ways: a) via the marginal likelihood approach by approximating the marginal likelihood of each model above using bridge sampling, and b) using the MCMC independence sampler. The independence sampler is run for a total of 2000 iterations after a burn-in phase of 100 iterations. The bridge sampling uses a total posterior sample size of $2R = 4000$, after a burn-in phase of 100 iterations.

Assessment of the simulation study involves two parts: Part 1) assessing the accuracy of the independence sampler for identifying M', and Part 2) assessing the efficacy of the prior distributions. Part 1) is basically asking: are the posterior model probabilities

approximated by the two methods, a) and b) above, similar? Part 2) is asking: do the proposed default prior distributions result in sensible model determination conclusions?

We assume throughout these analyses that the posterior model probabilities approximated by bridge sampling are close enough to the true values to be considered exact.

We consider Part 1) first. There are many possible ways to we could analyse the results of the simulations study. We begin by plotting $\hat{f}^L(m|\mathbf{y})$ against $\hat{f}^B(m|\mathbf{y})$ for each model, m. Figures 1 and 2 shows these plots for the Bernoulli and Poisson responses, respectively. If the independence sampler provides a good approximation to the posterior model probability, then the points should lie on a straight line with slope one, through the origin, and we have overlaid each plot with such a line. We see that, typically, the points lie near the overlaid line. The one concern is for Bernoulli responses where $G = 4$ and $n^* = 50$, where the independence sampler seems to underestimate the posterior model probability of Model 5. From the adjacent plot for Model 4, it appears that the independence sampler is allocating this probability to Model 4. However, we note that the independence sampler does not necessarily need to approximate $f(m|\mathbf{y})$ close to its true value, it just needs to identify a M' with high total posterior model probability. To assess this, for each of the 1000 datasets for each scenario, we use the independence sampler to identify M' according to (12) where we choose $c = 10$. We then evaluate the total posterior model probability within M' according to $\hat{f}^B(m|\mathbf{y})$, i.e. we find $\sum_{m \in M'} \hat{f}^B(m|\mathbf{y})$ for each repetition. We would like these total probabilities to be large since a small value would indicate that the independence sampler has failed to include in M'; a model with high posterior model probability. Table 1 shows the sample statistics of these total posterior model probabilities. Typically these total probabilities are close to 1 indicating that the independence sampler identifies a M' with high total posterior model probability.

We present the results of for Part 2). In what follows, all posterior model probabilities are the posterior model probabilities as approximated by bridge sampling.

For each of the four scenarios we consider two plots. The first is a plot of the total posterior model probability of models 2, 4 and 5 (i.e. the models that include a x_{ij} effect) against the value of the β_1 parameter. The second is a plot of the total posterior model probability of models 3, 4 and 5 (i.e. the models that include group-specific effects)

Table 1: Sample statistics of the total posterior model probability within M' according to $\hat{f}^B(m|\mathbf{y})$ for simulated Poisson and Bernoulli responses.

Bernoulli responses					
Scenario	Minimum	1st Quartile	Median	3rd Quartile	Maximum
$G = 25$,$n^* = 8$	0.7385	0.9270	0.9723	0.9950	1.0000
$G = 4$,$n^* = 50$	0.7365	0.9585	0.9855	1.0000	1.0000
Poisson responses					
Scenario	Minimum	1st Quartile	Median	3rd Quartile	Maximum
$G = 25$,$n^* = 8$	0.8915	1.0000	1.0000	1.0000	1.0000
$G = 4$,$n^* = 50$	0.6640	0.9884	0.9975	1.0000	1.0000

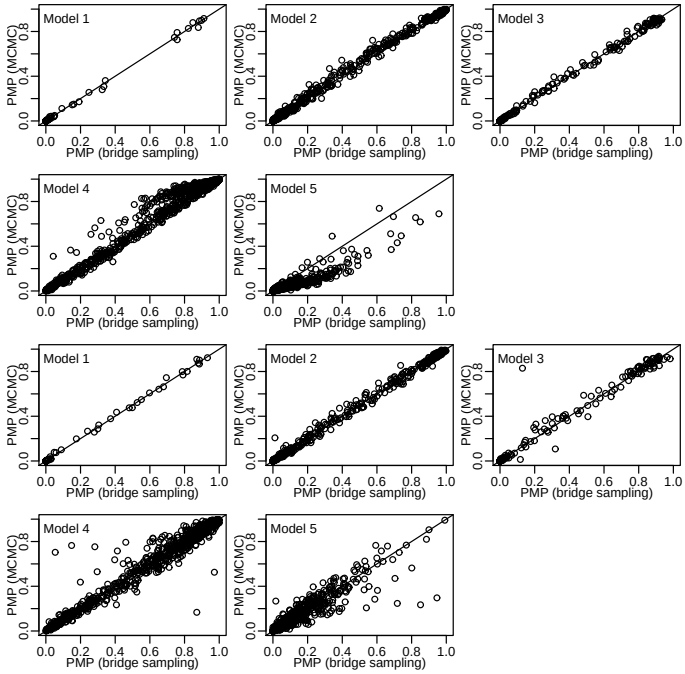
against the value of the σ^2 parameter. Figures 3 and 4 show these plots for the Poisson and Bernoulli responses, respectively. We add smoothing splines to the plots.

We see from Figures 3 and 4 that for small values of the true parameter, we are unlikely to choose the more complicated models. As the magnitude of the parameter increases the total posterior model probability of the appropriate models increases toward one.

However, consider the bottom right plot of Figures 3 and 4 which shows the total posterior model probability of models 3, 4 and 5 plotted against the true value of the σ^2 parameter. The smoothing spline appears to not approach one for large values of σ^2. We see that even for large values of σ^2 there exist total posterior model probabilities of models 3, 4 and 5 which are not close to one. Under further investigation this was due to the small number of groups, i.e. $G = 4$, and how the observed σ^2, i.e. the observed variance of u_i for $i = 1, ..., G$ being significantly smaller than the true value of σ^2. To see this, consider Figure 5 which shows the total posterior model probability of models 3, 4 and 5 plotted against the observed value of σ^2. We see from Figure 5 that the posterior model probability approaches and reaches one as the observed σ^2 increases.

Model 5 will never be the most parsimonious model available, and this is reflected in the results of the simulation study. It is the model with the highest posterior model probability 1.3% and 2.9% for Bernoulli responses, for $G = 25$ and $G = 4$, respectively, and 0.1% and 0.2% for Poisson responses, for $G = 25$ and $G = 4$, respectively.

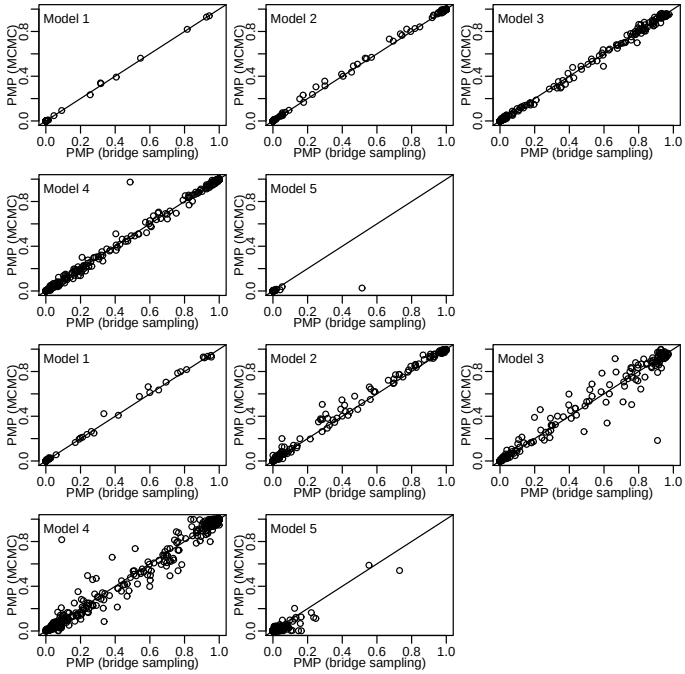
Figure 1: Plots of the posterior model probabilities as approximated by the independence sampler (PMP (MCMC)), $\hat{f}^L(m|\mathbf{y})$, against the posterior model probabilities as approximated by the bridge sampler (PMP(bridge sampling)), $\hat{f}^B(m|\mathbf{y})$ for the Bernoulli responses. The top two rows corresponds to $G = 25$ and $n^* = 8$ and the bottom two rows to $G = 4$ and $n^* = 50$.



Gelman (2006) points out that inverse-gamma (and inverse-Wishart) prior distributions for variance components can be overly informative, even with "non-informative" choices for the hyperparameters. To investigate this issue, we consider coverage rates of probability intervals for the parameters β_0, β_1 and, most importantly, σ^2. We find the 95% probability intervals by taking the 0.025 and 0.975 quantiles of the posterior sample of β_0, β_1 and σ^2 under Model 4, i.e. the true model. Table 2 shows the coverage rates of these intervals when compared against the true values of β_0 and β_1 for the intervals for β_0 and β_1, respectively, and when compared against the true and observed value of σ^2 for the interval for σ^2.

Table 2 shows that the coverage rates of the intervals for the regression parameters, β_0 and β_1 are close to the nominal value of 95%. However, there is under-coverage of the intervals for σ^2 when compared to the true and observed values of σ^2. On further

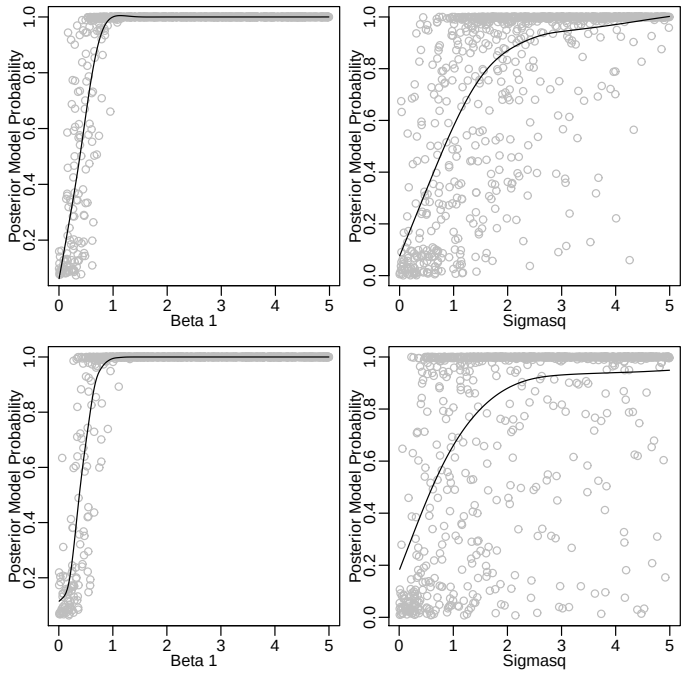
Figure 2: Plots of the posterior model probabilities as approximated by the independence sampler (PMP (MCMC)), $\hat{f}^L(m|\mathbf{y})$, against the posterior model probabilities as approximated by the bridge sampler (PMP(bridge sampling)), $\hat{f}^B(m|\mathbf{y})$ for the Poisson responses. The top two rows corresponds to $G = 25$ and $n^* = 8$ and the bottom two rows to $G = 4$ and $n^* = 50$.



investigation, we find that, for Bernoulli responses, 91.5% and 100% of the time the true value of σ^2 is less than the lower value of the interval for $G = 25$ and $G = 4$, respectively. The corresponding values for the Poisson responses are 91.2% and 100%. This shows that the proposed default prior distribution for σ^2 is informative when the true value of σ^2 is small. However, in these cases, Figures 3, 4 and 5 show us that we will allocate small posterior model probability to models that contain group-specific effects, i.e. models with non-zero σ^2. To demonstate this effect, we produce a model-averaged probability interval for σ^2; averaged over Models 2 and 4, i.e. the true model and the true model but with the group-specific parameters removed. This is found by producing a model-averaged posterior sample, of size $2N = 4000$, of σ^2, as follows. The sample will contain

$$R = \frac{\hat{f}^B(2|\mathbf{y})}{\hat{f}^B(2|\mathbf{y}) + \hat{f}^B(4|\mathbf{y})} \times 4000$$

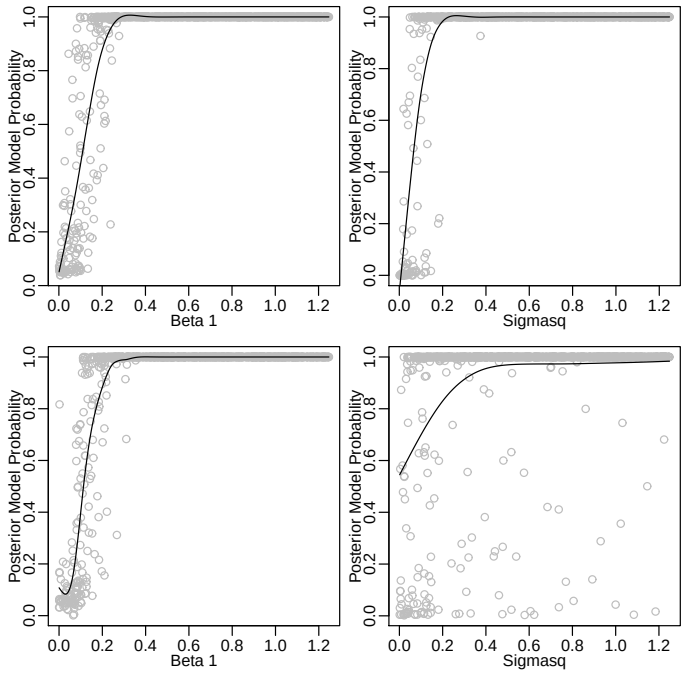
Figure 3: Plots of the total approximate posterior model probability for the 1000 datasets with Bernoulli responses against the true value of β_1 and σ^2. Top row: $G = 25$ and $n^* = 8$. Bottom row: $G = 4$ and $n^* = 50$.



elements that are identically zero and $4000 - R$ elements that are randomly selected from the posterior sample of σ^2 under Model 4. We then find a 95% probability interval by taking the 0.025 and 0.975 quantiles of this model-averaged posterior sample. The coverage rates of this model-averaged probability interval when compared against the true and observed values of σ^2 are shown in Table 2. The coverage rates for the model-averaged probability intervals are much closer to the nominal value than for the non-model-averaged intervals. In the case of when they are compared to the observed value of σ^2 the coverage rate is very close to the nominal level.

The simulation study shows that the proposed strategy appears to make favourable model determination conclusions.

Figure 4: Plots of the total approximate posterior model probability for the 1000 datasets with Poisson responses against the true value of β_1 and σ^2. Top row: $G = 25$ and $n^* = 8$. Bottom row: $G = 4$ and $n^* = 50$.



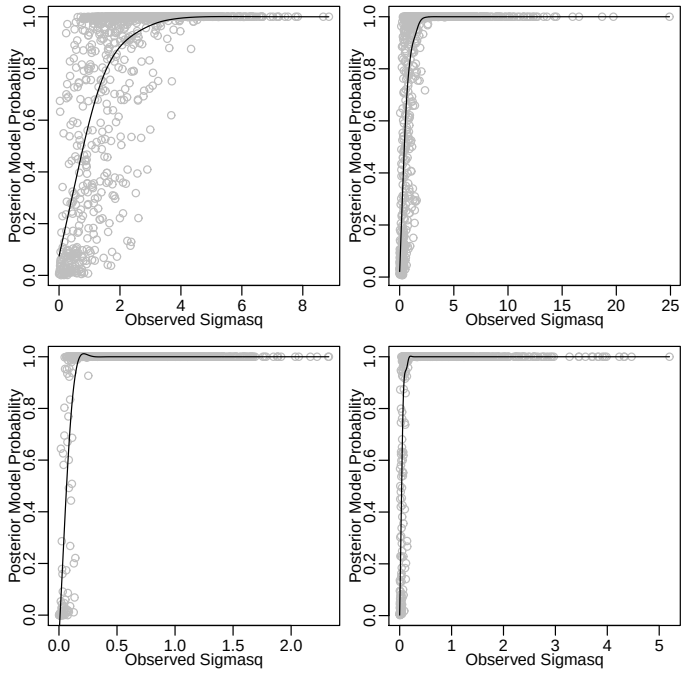
6. Examples

We demonstrate our strategy on four real datasets: *Natural Selection Study Data*, *Six Cities Data*, *Ship Incident Data* and *Malignant Melanoma Mortality Data*. The `R` code used to apply our proposed strategy is available as a Supplementary Material.

6.1. A Natural Selection Study

Sinharay & Stern (2005) presented *A Natural Selection Study* containing the survival status (0=died, 1=survived), birthweight (grams) and clutch (family) membership of 244 newborn turtles from 31 different clutches. The researchers want to determine whether there is a birthweight and/or clutch effect on survival of newborn turtles. Suppose y_{ij} and x_{ij} are the survival status and birthweight, respectively, from the jth turtle in the ith clutch, $i = 1, ..., 31$, $j = 1, ..., n_i$, and $y_{ij}|p_{ij} \sim$ Bernoulli(p_{ij}), where $p_{ij} = \Phi(\eta_{ij})$, i.e. we use the probit link function. We consider 5 models:

Figure 5: Plots of the total approximate posterior model probabilities for Model 3,4 and 5 for the 1000 datasets with Bernoulli and Poisson responses against the observed value of σ^2. Top row: Bernoulli responses. Bottom row: Poisson responses. First Column: $G = 25$ and $n^* = 8$. Second Column: $G = 4$ and $n^* = 50$.



1. $\eta_{ij} = \beta_0$,
2. $\eta_{ij} = \beta_0 + \beta_1 x_{ij}$,
3. $\eta_{ij} = \beta_0 + u_i$, where $u_i \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2)$,
4. $\eta_{ij} = \beta_0 + u_i + \beta_1 x_{ij}$, where $u_i \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2)$,
5. $\eta_{ij} = \beta_0 + u_i + (\beta_1 + v_i)x_{ij}$, where $(u_i, v_i)^T \overset{\text{iid}}{\sim} \text{N}(\mathbf{0}, \mathbf{D})$.

We apply the default priors proposed in Sections 2 and 3 to the appropriate model parameters. Note that for this example, $\mathbf{W}_{i,\mathbf{m},\mathbf{0}} = \frac{\pi}{2}\mathbf{I}_{n_i}$, therefore $\tau^2 = \frac{\pi}{2}$, $N_i = n_i$ and $N = n$.

In this example, the set of models is small enough to avoid the use of the independence sampler and we can approximate the posterior model probabilities, via bridge sampling, of all 5 models. However, as a demonstration we have computed the posterior model probabilities via the independence sampler as well. The independence sampler is run for

Table 2: Coverage rates of the probability intervals for β_0, β_1, and σ^2 for the simulated Bernoulli and Poisson responses.

Bernoulli responses						
Scenario	β_0	β_1	σ^2		Model-averaged σ^2	
			True σ^2	Observed σ^2	True σ^2	Observed σ^2
$G = 25, n^* = 8$	0.956	0.952	0.883	0.891	0.910	0.937
$G = 4, n^* = 50$	0.944	0.946	0.878	0.787	0.910	0.949
Poisson responses						
Scenario	β_0	β_1	σ^2		Model-averaged σ^2	
			True σ^2	Observed σ^2	True σ^2	Observed σ^2
$G = 25, n^* = 8$	0.967	0.952	0.897	0.923	0.902	0.937
$G = 4, n^* = 50$	0.940	0.950	0.863	0.781	0.907	0.948

a total of 10000 iterations after a burn-in phase of 1000 iterations. Bridge sampling is based on a posterior sample of size $2R = 20000$ from each model after a burn-in phase of 1000 iterations. Table 3 shows the posterior model probabilities approximated via the independence sampler and bridge sampling. It also contains the values of the *Bayesian Information Criterion* (BIC), *Akaike Information Criterion* (AIC), and *Deviance Information Criterion* (DIC) of the 5 models as a comparison. The DIC values are based on the following priors: we assume that the regression parameters are independent and $\beta_k \sim \mathrm{N}(0, 10^5)$, $\sigma^2 \sim \mathrm{IG}(0.00005, 0.5)$ for models 3 and 4, and $\mathbf{D} \sim \mathrm{IW}(2, 2\mathbf{I}_2)$ for model 5. These priors are proposed by Natarajan & Kass (2000).

The results in Table 3 show that, in this example, the Laplace approximation to the integrated likelihood performs very well since the posterior model probabilities as approximated by the independence sampler correspond closely to those approximated by bridge sampling. The posterior model probabilities seem to support the results of the BIC model selection method. It is known that AIC and DIC, typically, tend to favour more complicated models and this appears to be confirmed by this example.

Table 3: Approximated posterior model probabilities and BIC, AIC and DIC for the 5 models of the Natural Selection Study

Model	Posterior Model Probabilities		BIC_m	AIC_m	DIC_m
m	Bridge Sampling	Independence Sampler			
	$\hat{f}^B(m\|\mathbf{y})$	$\hat{f}^L(m\|\mathbf{y})$			
1	0.0002	0.0003	325.68	322.18	322.22
2	0.9095	0.8947	308.65	301.66	301.67
3	0.0007	0.0006	323.74	316.75	309.97
4	0.0794	0.0922	311.90	301.41	299.40
5	0.0103	0.0122	321.90	304.41	289.24

6.2. Six Cities Data

The *Six Cities Data* is frequently used to assess mixed models methodology. The data consists of the wheezing status, y_{ij} (0=not wheezing, 1=wheezing), of child i at time-point j, for $i = 1, ..., 537$ and $j = 1, .., 4$. Also included, is the age of the ith child, x_{1ij}, at time-point j and the smoking status, x_{2ij}, of the ith child's mother at time-point j. Note that $x_{2ij} = x_{2ik}$ for all $j, k \in \{1, .., 4\}$. We can also define the interaction covariate $x_{3ij} = x_{1ij}x_{2ij}$. By considering all possible models with the canonical logit link where we use first-order terms of x_{1ij} and x_{2ij} and their interaction and adhering to the modelling convention of not including an interaction covariate unless all marginal covariates are included, there are 19 possible models.

We apply the default priors proposed in Sections 2 and 3 to the appropriate model parameters. Note that in this example, $\mathbf{W}_{i,\mathbf{m},\mathbf{0}} = 4\mathbf{I}_{n_i}$, therefore $\tau^2 = 4$, $N_i = n_i$ and $N = n$.

It is impractical to apply bridge sampling to all models, so in this example it is necessary to use the independence sampler described in Section 4.2 to identify a smaller subset of models on which to use bridge sampling.

We run the independence sampler for a total of $B = 10000$ iterations after a burn-in phase of 1000 iterations. After running the independence sampler we identify M' with the four models shown below:

6. $\eta_{ij} = \beta_0 + u_i;\ u_i \sim \mathrm{N}(0, \sigma^2)$,

7. $\eta_{ij} = \beta_0 + \beta_1 x_{1ij} + u_i;\ u_i \sim \mathrm{N}(0, \sigma^2)$.

8. $\eta_{ij} = \beta_0 + \beta_2 x_{2ij} + u_i;\ u_i \sim \mathrm{N}(0, \sigma^2)$.

9. $\eta_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + u_i;\ u_i \sim \mathrm{N}(0, \sigma^2)$.

The posterior model probabilities of these four models as approximated by the independence sampler are shown in Table 4. These four models account for 95.87% of the total posterior model probability in M. The model with the next highest approximated posterior model probability is model 11 with linear predictor $\eta_{ij} = (\beta_0 + u_{0i}) + (\beta_1 + u_{1i})x_{1ij}$, where $(u_{0i}, u_{1i})^T \sim \mathrm{N}(\mathbf{0}, \mathbf{D})$. This model has $\hat{f}^L(11|\mathbf{y}, M) = 0.0144$. Table 4 also shows the posterior model probabilities as approximated by the independence sampler, if we only consider models in M'. These are denoted by $\hat{f}^L(m|\mathbf{y}, M')$.

We then used bridge sampling with a posterior sample size of $2R = 50000$ from each model after a burn-in phase of 1000 iterations, to obtain approximations to the log marginal likelihoods, $\log \hat{f}^B(\mathbf{y}|m, M')$, and posterior model probabilities, $\hat{f}^B(m|\mathbf{y}, M')$, conditional on M'. These are shown in Table 4.

Table 4: Approximated posterior model probabilities and log marginal likelihoods for the 4 models in M' for the Six Cities Data

m	$\hat{f}^L(m\|\mathbf{y}, M)$	$\hat{f}^L(m\|\mathbf{y}, M')$	$\log \hat{f}^B(\mathbf{y}\|m, M')$	$\hat{f}^B(m\|\mathbf{y}, M')$
6	0.3813	0.3977	-808.1482	0.3877
7	0.4131	0.4309	-807.9760	0.4606
8	0.0731	0.0762	-809.8046	0.0740
9	0.0912	0.0951	-809.7553	0.0777

We computed the AIC, BIC and DIC values for all 19 models. BIC chooses as the top four models, the same four models in M' as our strategy. However, BIC prefers model 6 to model 7, although the values of BIC are very similar. AIC chooses model 9 with linear predictor $\eta_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + u_i$ where $u_i \sim \mathrm{N}(0, \sigma^2)$, with model 7 second.

6.3. Ship Incident Data

The *Ship Incident Data* can be found in McCullagh & Nelder (1989) and concerns the number of damage incidents suffered by cargo ships between 1960 and 1979, that were caused by waves. The dataset contains data from five different types of ship which we regard as the groups, i.e. $G = 5$. There are two other classification factors: year of construction (1960-64, 1965-69, 1970-74, 1975-79) and year of operation (1960-74, 1975-79).

Let y_{ij} and E_{ij} denote the number of damage incidents suffered by and the aggregate months of service of the ith ship type and the jth unique combination of classification factors, respectively, for $i = 1, ..., G = 5$ and $j = 1, ..., n_i$. Since there are four different classifications for year of construction and two for year of operation, $n_i = 8$. However, since a ship constructed in 1975-79 cannot operate in 1960-74, the aggregate months of service is zero and these rows can be deleted, resulting in $n_i = 7$. Also, the aggregate months of service for ship type 5, constructed in 1960-64 and operating in 1975-79 is also zero, so this row can be deleted. Therefore, $n_i = 7$, for $i = 1, ..., 4$, $n_5 = 6$, and $n = \sum_{i=1}^{G} n_i = 34$.

We construct indicator variables for the classification factors. For the ith ship type, let

$$x_{1ij} = \begin{cases} 1, & \text{if the } j\text{th entry was operating in 1975-79,} \\ 0, & \text{otherwise,} \end{cases}$$

for $j = 1, ..., n_i$. Likewise, for the ith ship type, let

$$x_{2ij} = \begin{cases} 1, & \text{if the } j\text{th entry was constructed in 1965-69,} \\ 0, & \text{otherwise,} \end{cases}$$

$$x_{3ij} = \begin{cases} 1, & \text{if the } j\text{th entry was constructed in 1970-74,} \\ 0, & \text{otherwise,} \end{cases}$$

$$x_{4ij} = \begin{cases} 1, & \text{if the } j\text{th entry was constructed in 1975-79,} \\ 0, & \text{otherwise,} \end{cases}$$

for $j = 1, ..., n_i$.

We adhere to the modelling principle, that if there are more than one indicator variables that relate to a classification factor, then they are either all included or all

excluded from the linear predictor. For example, if x_{4ij} is included in the linear predictor, then so must x_{2ij} and x_{3ij}.

We assume that $y_{ij} \sim \text{Poisson}(\mu_{ij})$ where $\mu_{ij} = E_{ij}\lambda_{ij}$ and $\log \lambda_{ij} = \eta_{ij}$. The link function is then $g(\mu_{ij}) = \log\left(\frac{\mu_{ij}}{E_{ij}}\right)$, with $g'(\mu_{ij}) = \frac{1}{\mu_{ij}}$. We term E_{ij}, the aggregate months of service as the exposures. We do not consider interactions between the classification factors, so there are a total of thirteen models, including four GLMs.

We apply the prior distributions proposed in Sections 2 and 3 for $\boldsymbol{\beta}$ and $\mathbf{D}$. Note that in this example, $\mathbf{W}_{i,\mathbf{m},\mathbf{0}} = \text{diag}\left\{E_{ij}^{-1}\right\}$, so that $N = \sum_{i=1}^{G}\sum_{j=1}^{n_i} E_{ij}$ and $N_i = \sum_{j=1}^{n_i} E_{ij}$. We run the independence sampler for a total of 10000 iterations after a burn-in phase of 1000 iterations. The independence sampler identifies an M' containing two models. These models have linear predictors:

7. $\eta_{ij} = \beta_1 + u_i + \beta_2 x_{2ij} + \beta_3 x_{3ij} + \beta_4 x_{4ij}$; where $u_i \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2)$,

8. $\eta_{ij} = \beta_1 + u_i + \beta_2 x_{1ij} + \beta_3 x_{2ij} + \beta_4 x_{3ij} + \beta_5 x_{4ij}$; where $u_i \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2)$.

Table 5 shows the posterior model probabilities of the two models in M', as approximated by the independence sampler. These two models account for 98.70% of total posterior model probability. Table 5 also shows the posterior model probabilities as approximated by the independence sampler, if we consider only models in M'. These are denoted by $\hat{f}^L(m|\mathbf{y}, M')$.

Table 5: Approximate posterior model probabilities and log marginal likelihoods of the models in M' from the Ship Incident Data.

m	$\hat{f}^L(m\|\mathbf{y}, M)$	$\hat{f}^L(m\|\mathbf{y}, M')$	$\log \hat{f}^B(\mathbf{y}\|m, M')$	$\hat{f}^B(m\|\mathbf{y}, M')$
7	0.0909	0.0921	-104.6083	0.0861
8	0.8961	0.9079	-102.2457	0.9139

We now approximate the marginal likelihood of the two models in M' using bridge sampling with a total posterior sample size of $2R = 20000$. Table 5 shows the log marginal likelihoods and resulting posterior model probabilities, as approximated by bridge sampling.

6.4. Malignant Melanoma Mortality Data

The *Malignant Melanoma Mortality Data* is analysed by Langford et al (1998) and concerns the number of deaths due to malignant melanoma in the European community. The dataset contains data from 354 countries from nine different countries which we consider to be the groups, so $G = 9$ and $n = 354$. Let d_{ij}, E_{ij} and z_{ij} denote the number of male deaths due to malignant melanomas, expected number of these deaths and the measure of the UVB dose reaching the earth's surface, respectively, of the jth county in the ith country, for $i = 1, ..., G$ and $j = 1, ..., n_i$. We define $x_{ij} = \frac{z_{ij} - \bar{z}}{\sqrt{\text{var}(z_{ij})}}$ as the standardised z_{ij}'s, and the response to be

$$y_{ij} = \begin{cases} 1, & \text{if} \quad d_{ij} \geq E_{ij}, \\ 0, & \text{if} \quad d_{ij} < E_{ij}. \end{cases}$$

We assume that $y_{ij} \sim \text{Bernoulli}(p_{ij})$. We assume that the link function is unknown and we consider two options: the logit link and the probit link. For each link function, there are five choices for the linear predictor, similar to the Natural Selection Study in Section 6.1, so we have a total of 10 models.

We apply the prior distributions proposed in Sections 2 and 3. Note that in this example, if the logit link is chosen then $\mathbf{W}_{i,\mathbf{m},\mathbf{0}} = 4\mathbf{I}_{n_i}$, therefore $\tau^2 = 4$, $N_i = n_i$ and $N = n$. If the probit link is chosen then $\mathbf{W}_{i,\mathbf{m},\mathbf{0}} = \frac{\pi}{2}\mathbf{I}_{n_i}$, therefore $\tau^2 = \frac{\pi}{2}$, $N_i = n_i$ and $N = n$.

We run the independence sampler for a total of 10000 iterations after a burn-in phase of 1000 iterations. The independence sampler identifies an M' that contains two models defined by the linear predictor

$$\eta_{ij} = (\beta_1 + u_{1i}) + (\beta_2 + u_{2i})\, x_{ij},$$

and either having the logit link (Model 1) or the probit link (Model 2). Table 6 gives the approximate posterior model probability of the models in M'. These two models account for approximately 100% of total posterior model probability. Table 6 also shows the posterior model probabilities as approximated by the independence sampler, if we consider only models in M'. These are denoted by $\hat{f}^L(m|\mathbf{y}, M')$.

We now approximate the marginal likelihood of the two models in M' using bridge sampling with a total posterior sample size of $2R = 20000$. Table 6 shows the approximate

log marginal likelihoods and the resulting posterior model probabilities.

Table 6: Approximate posterior model probabilities and log marginal likelihoods of the models in M' from the Malignant Melanoma Mortality Data.

m	$\hat{f}^L(m\|\mathbf{y}, M)$	$\hat{f}^L(m\|\mathbf{y}, M')$	$\log \hat{f}^B(\mathbf{y}\|m, M')$	$\hat{f}^B(m\|\mathbf{y}, M')$
5	0.5044	0.5044	-153.3822	0.5055
10	0.4956	0.4956	-153.4040	0.4945

7. Discussion

In this paper, we considered a default strategy for model determination amongst GLMMs under weak prior information and where the dispersion parameter of the exponential family is unknown. Our strategy takes into account default prior specification for the regression parameters and the variance components, and describes a general computational strategy.

The default priors are based on a unit information concept that has proved successful for other authors. We note that the priors are conditional on the design matrices $\mathbf{X}_i$ (and also $\mathbf{Z}_i$) so therefore the prior distributions are dependent on the form of the experiment. However, all regression analyses are conditional on the regressors so we feel that the proposed strategy is still fully Bayesian.

The general computational strategy is based on two phases. Phase one combines a Laplace approximation of the integrated likelihood with an MCMC method to find $\hat{f}^L(m|\mathbf{y})$; an approximation to the posterior model probabilities. These $\hat{f}^L(m|\mathbf{y})$ are then used to define M' a candidate set of promising models on which to focus. Phase two involves performing the more computationally expensive but more accurate bridge sampling on the models in M' to find $\hat{f}^B(m|\mathbf{y})$.

The strategy considered allows a fully Bayesian analysis of GLMMs under model uncertainty and weak prior information, without the need of choosing arbitrary hyperparameters. This strategy allows us to consider model determination amongst models that do not just have a group-specific intercept (i.e. a random intercept). In the Malignant Melanoma Mortality Data example, we showed that models with just group-specific

intercepts would have low posterior model probability when compared to the more complicated models included in M'.

Bridge sampling is a computationally expensive method since it requires a sample from the posterior distribution. However, the models from which we require a posterior sample will be the models of greatest interest and therefore we will need a posterior sample on which to base posterior inferences.

We do not consider a default prior for the dispersion parameter since, typically, this is either known (as is the case for Bernoulli or Poisson response) or is present in all models. However, it may be the case that we are uncertain of the response distribution (e.g. normal vs. gamma) and therefore defining a default prior for the dispersion parameter becomes relevant. Future work will address this issue.

The independence sampler considered in Section 4.2 is feasible for a small to moderate number of models, or equivalently a small to moderate number of covariates. However, as this number increases it will become impractical to maximise $\hat{h}_m(\boldsymbol{\beta}_m, \boldsymbol{\upsilon}_m, \omega_m|\mathbf{y})$ for all $m \in M$. A more suitable approach would be to use a more general reversible jump approach where proposals are based on the current set of parameters, thus negating the need to maximise $\hat{h}_m(\boldsymbol{\beta}_m, \boldsymbol{\upsilon}_m, \omega_m|\mathbf{y})$ for each $m \in M$. Future work will focus on developing this methodology.

8. Acknowledgments

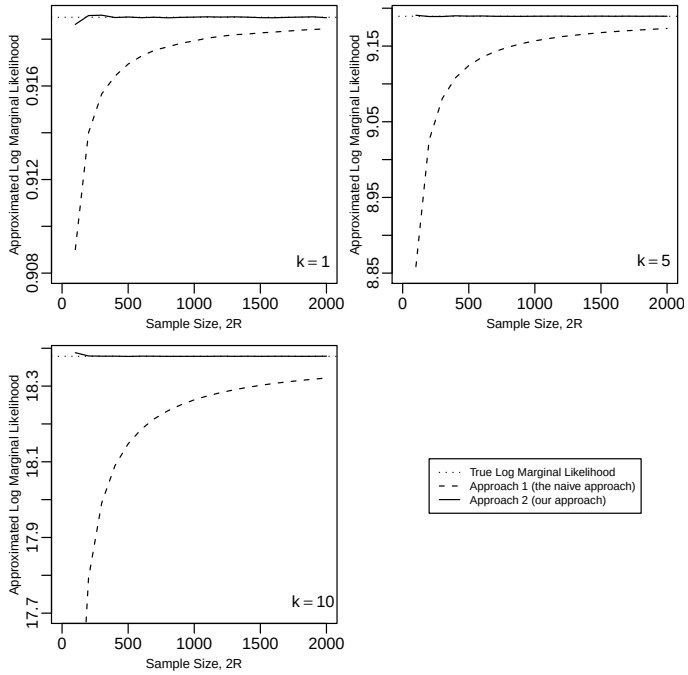
The authors would like to thank the associate editor and the two referees for their valuable comments and suggestions that significantly improved this article.

Appendix A. Bridge Sampling

Sinharay & Stern (2005) found that matching the moments (or mode and curvature at the mode) of the distribution with density $g(\boldsymbol{\theta})$ to those of the posterior distribution, $\boldsymbol{\theta}|\mathbf{y}$, increased the accuracy of bridge sampling by reducing the standard deviation of the approximations.

For some models, where the posterior distribution is non-normal we feel that the sample mean and variance from a posterior sample is the best choice. Since we have

Figure A.6: Plots of the approximated log marginal likelihood for the two different approaches and three different dimensions against the sample size.



a sample from the posterior distribution, a naive approach may be to approximate the mean and variance of $\boldsymbol{\theta}|\mathbf{y}$ using the sample statistics of the same posterior sample as we use in the bridge sampler. However, as we show using simulations, this leads to underestimation of the marginal likelihood.

We choose the posterior distribution to be the k-variate normal distribution with mean $\mathbf{0}$ and variance matrix $\mathbf{I}_k$. Hence,

$$f(\boldsymbol{\theta}|\mathbf{y}) \propto \exp\left(-\frac{1}{2}\boldsymbol{\theta}^T\boldsymbol{\theta}\right),$$

and the marginal likelihood is the normalising constant of the $\mathrm{N}(\mathbf{0}, \mathbf{I}_k)$ distribution: $(2\pi)^{\frac{k}{2}}$. As the distribution with density $g()$, we also use the k-variate normal distribution with mean $\boldsymbol{\mu}$ and variance matrix $\boldsymbol{\Sigma}$.

We have a sample $\{\boldsymbol{\theta}^i\}_{i=1}^{2R}$ of size $2R$ from $\mathrm{N}(\mathbf{0}, \mathbf{I}_k)$ which represents our posterior sample. All that remains is to choose appropriate values for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, and we assess two different methods for doing so:

1. Approach 1 (The naive approach). Use the sample mean and variance of the posterior sample, $\{\boldsymbol{\theta}^i\}_{i=1}^{2R}$, and use the bridge sampler (13) with a sample size of $N_1 = N_2 = 2R$.
2. Approach 2 (Our approach). Use the sample mean and variance of half of the posterior sample, $\{\boldsymbol{\theta}^i\}_{i=1}^{R}$. Use the second half of the posterior sample, $\{\boldsymbol{\theta}^i\}_{i=N+1}^{2R}$, in the bridge sampler (13) with a reduced sample size of $N_1 = N_2 = R$.

The sample sizes, $2R$, that we consider come from the set $\{100p : 1 \leq p \leq 20, p \in \mathbb{Z}\}$, and we repeat each computation at each unique sample size 10000 times. We consider three different dimensions, k, from the set $\{1, 10, 20\}$.

Figure A.6 shows plots of the approximated log-marginal likelihood for the two different approaches against the sample size, $2R$. Also included on the plot is a line at the true log-marginal likelihood, $\frac{k}{2}\log(2\pi)$. The plots show that the naive approach leads to an underestimation of the marginal likelihood which appears to decrease as the sample size increases. Our approach leads to no such underestimation with a small overestimation for small sample sizes which is expected since the bridge sampling estimator is based on a ratio and it is well known that $\mathrm{E}\left(\frac{X}{Y}\right) > \frac{\mathrm{E}(X)}{\mathrm{E}(Y)}$, for positive random variables X and Y.

References

[1] Aitkin, M., Liu, C.C. & Chadwick, T, 2009. Bayesian model comparison and model averaging for small-area estimation. Annals of Applied Statistics, 3, 199-221.

[2] Breslow, N.E. & Clayton, D.G., 1993. Approximate Inference in Generalized Linear Mixed Models. Journal of the American Statistical Association, 88, 9-25.

[3] Cai, B. & Dunson, D.B., 2006. Bayesian Covariance Selection in Generalised Linear Mixed Models. Biometrics, 62, 446-457.

[4] Chen, M., Shao, Q. & Ibrahim, J.G., 2000. Monte Carlo Methods in Bayesian Computation. Springer-Verlag, New York.

[5] Chen, M., Ibrahim, J.G., Shao, Q., & Weiss, R.E, 2003. Prior elicitation for model selection and estimation in generalized linear mixed models. Journal of Statistical Planning and Inference, 111, 57-76.

[6] Chib, S., 1995. Marginal likelihood from the Gibbs output. Journal of the American Statistical Association, 90, 1313-1321.

[7] Daniels, M.J., 1999. A prior for the variance in hierarchical models. The Canadian Journal of Statistics, 27(3), 567-578.

[8] Garcia-Donato, G. & Sun, D., 2007. Objective priors for hypothesis testing in one-way random effects models. The Canadian Journal of Statistics, 35(2), 303-320.

[9] Gelman, A., 2006. Prior distributions for variance parameters in hierarchical models. Bayesian Analysis, 3, 515-534.

[10] Gustafson, P., Hossain, S. & MacNab, Y.C., 2006. Conservative prior distributions for variance components in hierarchical models. The Canadian Journal of Statistics, 34(3), 377-390.

[11] Green, P.J., 1995. Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination. Biometrika, 82(4), 711-732.

[12] Kass, R.E. & Natarajan, R., 2006. A Default Conjugate Prior for Variance Components in Generalised Linear Mixed Models (Comment on Article by Browne and Draper). Bayesian Analysis, 1(3), 535-542.

[13] Kass, R.E. & Raftery, A.E., 1995. Bayes Factors. Journal of the American Statistical Association, 90, 773-795.

[14] Kass, R.E. & Wasserman, L., 1995. A reference Bayesian test for nested hypotheses and its relationships to the Schwartz criterion. Journal of the American Statistical Association, 90, 928-934.

[15] Langford, I.H., Bentham, G. & McDonald, A. 1998. Multilevel modelling of geographically aggregated health data: a case study on malignant melanoma mortality and UV exposure in the European community. *Statistics in Medicine*, 17, 41-58.

[16] Lunn, D.J., Thomas, A., Best, N. and Spiegelhalter, D., 2000. WinBUGS - a Bayesian modelling framework: concepts, structure, and extensibility. Statistics and Computing, 10, 325-337.

[17] Madigan, D. & Raftery, A.E., 1994. Model Selection and Accounting for Model Uncertainty in Graphical Models using Occam's Window. Journal of the American Statistical Association, 95, 227-237.

[18] McCullagh, P. & Nelder, J.A., 1989. Generalized Linear Models (2ed). Chapman & Hall, London.

[19] Meng, X.L. & Wong, W.H., 1996. Simulating Ratios of Normalising Constants via a Simple Identity: A Theoretical Exploration. Statistica Sinica, 6, 831-860.

[20] Muirhead, R.J., 1982. Aspects of Multivariate Statistical Theory. Wiley, New York.

[21] Natarajan, R. & Kass, R.E., 2000. Reference Bayesian Methods for Generalised Linear Mixed Models. Journal of the American Statistical Association, 95, 227-237.

[22] Ntzoufras, I., Dellaportas, P. & Forster, J.J., 2003. Bayesian variable and link determination for generalised linear models. Journal of Statistical Planning and Inference, 111, 165-180.

[23] O'Hagan, A. & Forster, J.J., 2004. Kendall's Advanced Theory of Statistics, Volume 2B: Bayesian Inference. 2nd Edition. Arnold, London.

[24] Pauler, D.K., 1998. The Schwarz Criterion and Related Methods for Normal Linear Models. Biometrika, 85(2), 13-27.

[25] Raftery, A.E., 1996. Approximate Bayes Factors and Accounting for Model Uncertainty in Generalised Linear Models. Biometrika, 83(2), 251-266.

[26] R Development Core Team, 2008. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing.

[27] Sinharay, S. & Stern, H.S., 2005. An Empirical Comparison of Methods for Computing Bayes Factors in Generalized Linear Mixed Models. Journal of Computational and Graphical Statistics,

14(2), 415-435.

[28] Scott, J.G. & Carvalho, C.M., 2008. Feature-Inclusion Stochastic Search for Gaussian Graphical Models. Journal of Computational and Graphical Statistics, 17(4), 790-808.

[29] Smith, A.F.M. & Spiegelhalter, D.J., 1980. Bayes Factors and Choice Criteria for Linear Models. Journal of the Royal Statistical Society, Series B, 42(2), 213-220.

[30] Spiegelhalter, D.J., Best, N.G., Carlin, B.P. & van der Linde, A., 2002. Bayesian measures of model complexity and fit (with discussion). Journal of the Royal Statistical Society, Series B, 64, 583-640.

[31] Sturtz, S., Ligges, U. & Gelman, A., 2005. R2WinBUGS: A Package for Running WinBUGS from R. Journal of Statistical Software, 12, 1-16.

[32] Zeger, S.L. & Karim, M.R., 1991. Generalized Linear Models with Random Effects: A Gibbs Sampling Approach. Journal of the American Statistical Association, 86, 79-86.