Abstract

In this paper, we consider a default strategy for fully Bayesian model determination for GLMMs. We address the two key issues of default prior specification and computation. In particular, we extend a concept of unit information to the priors for the parameters of a GLMM. We rely on a combination of MCMC and Laplace approximations to compute approximations to the posterior model probabilities and then further approximate these posterior model probabilities using bridge sampling. We apply our strategy to two examples.

Key words: 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 $j$th response from the $i$th group where $j = 1, ..., n_i$ and $i = 1, ..., G$. Let $x_{ij}$ and $z_{ij}$ denote the $p \times 1$ and $q \times 1$ vectors of covariates which correspond to the regression and group-specific parameters, respectively. Assume that the components of $z_{ij}$ form a subset of the components of $x_{ij}$. 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

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

where \( z_{ij} \) is the canonical parameter, \( \phi \) is the dispersion parameter, and \( a_{ij}(\cdot), b(\cdot), \) and \( c(\cdot, \cdot) \) are known functions. Define \( \mu_{ij} = \mathbb{E}(Y_{ij}|\mathbf{u}_i) = b'(z_{ij}) \) as the conditional mean of \( Y_{ij} \). This is related to the linear predictor, \( \eta_{ij} \), through

\[
g(\mu_{ij}) = \eta_{ij} = \mathbf{x}_i^T \beta + \mathbf{z}_i \mathbf{u}_i,
\]

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

Suppose \( \mathbf{y}_i = (y_{i1}, \ldots, y_{im})^T, \mathbf{X}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{im})^T, \mathbf{Z}_i = (\mathbf{z}_{i1}, \ldots, \mathbf{z}_{im})^T, \eta_i = (\eta_{i1}, \ldots, \eta_{im})^T, \mu_i = (\mu_{i1}, \ldots, \mu_{im})^T \), and that the link function is applied elementwise, then

\[
g(\mu) = \mathbf{X} \beta + \mathbf{Z} \mathbf{u}.
\]

Suppose further that \( \mathbf{y} = (\mathbf{y}_1^T, \ldots, \mathbf{y}_G^T)^T, \mathbf{X} = (\mathbf{X}_1^T, \ldots, \mathbf{X}_G^T)^T, \mathbf{Z} = \text{diag}(\mathbf{Z}_1, \ldots, \mathbf{Z}_G), \eta = (\eta_1^T, \ldots, \eta_G^T)^T, \mu = (\mu_1^T, \ldots, \mu_G^T)^T, \) and \( \mathbf{u} = (\mathbf{u}_1^T, \ldots, \mathbf{u}_G^T)^T \), then (1) can be rewritten in matrix form as

\[
g(\mu) = \mathbf{X} \beta + \mathbf{Z} \mathbf{u}.
\]

We complete the specification of a GLMM by making the common assumption that \( \mathbf{u}_i \sim \text{N}(\mathbf{0}, \mathbf{D}) \), for \( i = 1, \ldots, G \), 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 \( \mathbf{u} \sim \text{N}(\mathbf{0}, \mathbf{D}^\prime) \).

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

We decompose this prior density as

\[
f(\beta, \mathbf{D}, \phi) = f(\beta|\mathbf{D}, \phi)f(\mathbf{D}|\phi)f(\phi).
\]

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

\[
f_m(\mathbf{y}|\boldsymbol{\beta}_m, \mathbf{D}_m, \phi_m) = \int_{\mathbf{u}_m} f_m(\mathbf{y}|\boldsymbol{\beta}_m, \mathbf{u}_m, \phi_m)f_m(\mathbf{u}_m|\mathbf{D}_m)d\mathbf{u}_m,
\]

where \( \phi_m \) is the dispersion parameter for model \( m \).
where $M$ is a set of models. The posterior model probability, $f(m|y)$, of model $m$ is given by

$$f(m|y) = \frac{f(m)f_m(y)}{\sum_{k \in M} f(k)f_k(y)}, \quad (4)$$

where $f_m(y)$ is the marginal likelihood of model $m$ given by

$$f_m(y) = \int f_m(y|\beta_m, u_m, \phi_m)f_m(u_m|D_m)f(\beta_m, D_m, \phi_m)d\beta_m d\phi_m dD_m, \quad (5)$$

and $f(m)$ is the prior model probability of model $m$. 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.

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

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

The ratio $f_1(y)/f_2(y)$ 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 (Speigelhalter 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, $\beta$. In Section 3, we discuss some of the priors for the variance components matrix, $D$, 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, $\phi$. 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 $\phi$.

The integral (5) 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 (4). Approaches to computation are considered in Section 4.

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 two examples.

2. Prior for the Regression Parameters, $\beta$

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

In this section, we generalise a class of priors for $\beta$ known as unit-information priors to GLMMs. Previously, 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 show, using the approaches of Pauler (1998) and Ntzoufras et al (2003), that these priors can be, approximately, extended to GLMMs.
We define a unit information prior for $\beta$ as the multivariate normal distribution with mean $m$ and $\Sigma$, i.e.

$$\beta \sim N(m, \Sigma),$$

for particular choices of $m$ and $\Sigma$. We follow Raftery (1996) and Ntzoufras et al (2003), and choose the prior mean as $m = (m_0, 0, ..., 0)^T$. 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, $\Sigma$, is chosen to approximately provide the same amount of information as one unit of data. Consider the linear model: $y \sim N(0, \sigma^2 I_n)$. The Fisher information is given by

$$I_\beta = \frac{1}{\sigma^2} X^T X.$$

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

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

Consider a GLM, the Fisher information is given by

$$I_\beta = X^T W^{-1} X,$$

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

$$\Sigma = n (X^T W_m^{-1} X)^{-1},$$

where $W_m = \text{diag}\{\text{var}(Y_i|\beta = m)(g'(\mu_i)|_{\beta=m})^2\}$.

In both the LM and the GLM cases, we divide the Fisher information by the sample size, $n$. However, for mixed models, as Pauler (1998) states, the sample size “is ambiguous because of the correlations between observations”. In the same paper, Pauler (1998) defines a general sample size to be the order of the Fisher information.

In general, let

$$\Sigma = \Lambda I_\beta^{-1} \Lambda,$$

where $I_\beta$ is the Fisher information for $\beta$, $\Lambda = \text{diag}\{\sqrt{N_k}\}$ for $k = 1, ..., p$. The value $N_k$ is called
the effective sample size for $\beta_k$, which has the property that $I_{\beta,kk} \propto N_k$ and is defined as

$$N_j = \begin{cases} G, & \text{if } \beta_j \text{ has an associated group-specific parameter} \\ n, & \text{otherwise.} \end{cases} \quad (7)$$

Pauler (1998) shows that (7) is appropriate for LMMs and we provide a justification for it holding for GLMMs, below.

In the cases of an LM and a GLM, none of the $\beta_k$’s have associated group-specific parameters and so $N_k = n$ for all $k$. Therefore, (6) reduces to the appropriate variance matrix, since $\Lambda = \sqrt{n}I_p$ and $\Sigma = nI_{\beta}^{-1}$. Consider the more general case of mixed models and suppose $\beta_k$ has an associated group-specific parameter and $\beta_j$ does not. Then (7) implies, intuitively, that the amount of information we have for $\beta_k$ only increases as we observe data from more groups, whereas the amount of information for $\beta_j$ increases as the total sample size increases.

We now provide a justification that (7) is also appropriate for GLMMs. Breslow & Clayton (1993) give the following approximation to the Fisher information for a GLMM:

$$I_\beta = X^T VX, \quad (8)$$

where $V = [W + ZD'Z']^{-1}$. Here $W = \text{diag}(W_i)$, for $i = 1, \ldots, G$, and

$$W_i = \text{diag}(\text{var}(Y_{ij}|u)(g'(\mu_i))^2),$$

for $j = 1, \ldots, n_i$. The matrix $V$ can be rewritten (see Henderson & Searle (1981)) as

$$V = W^{-1} - W^{-1}Z(I_{Gq} + D'Z'W^{-1}Z)^{-1}D'Z'W^{-1}.$$

Of course, this expression depends on the unknown parameters $\beta$ and $u$, so we follow Ntzoufras et al (2003) and replace any unknown parameters by their prior means. In this case, we replace $\beta$ and $u$ by $m$ and $0$, respectively. Therefore, we obtain

$$V = W_{m,0}^{-1} - W_{m,0}^{-1}Z(I_{Gq} + D'Z'W_{m,0}^{-1}Z)^{-1}D'Z'W_{m,0}^{-1}, \quad (9)$$

where $W_{m,0} = \text{diag}(W_{i,m,0})$ and

$$W_{m,0} = \text{diag}(\text{var}(Y_{ij}|\beta = m, u = 0)(g'(\mu_i)|_{\beta=m,u=0})^2).$$

Assume that $W_{i,m,0}$ can be written $\tau_i^2I_n$, as would be the case for normal, gamma and Bernoulli responses and Poisson responses where the exposures are the same in each group. In this case

$$X^T \hat{V} X = \sum_{i=1}^{G} X_i^T \hat{V}_i X_i, \quad (10)$$
where

\[
\hat{V}_i = W^{-1}_{i,m,0} - W^{-1}_{i,m,0} Z (I_{q} + D^T Z^{-1} W^{-1}_{i,m,0} Z)^{-1} D^T Z^{-1} W^{-1}_{i,m,0} Z^{-1} \frac{1}{\tau_i^2} \left( I_{n_i} - \frac{1}{\tau_i^2} Z \left( I_q + \frac{1}{\tau_i^2} D Z_i^T Z \right)^{-1} D Z_i \right). 
\]

(11)

We can use the results of Pauler (1998) to show that \((X^T \hat{V}_i X^T)_{kk}\) is independent of \(n_i\) if \(\beta_k\) has an associated group-specific parameter, and \((X^T \hat{V}_i X^T)_{kk} \propto n_i\), otherwise. Therefore, (7) is appropriate for GLMMs.

Note that, if this unit information prior for \(\beta\) is used then the expressions (10) and (11) are of direct use since

\[
\beta \sim N(m, \Lambda (X^T \hat{V} X)^{-1} \Lambda).
\]

3. Prior for the Variance Components Matrix, \(D\)

There is a large literature on default prior distributions for the variance components matrix, \(D\).

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 \(D = \sigma^2 \Omega\), where \(\Omega\) is a known positive-definite matrix and \(\sigma^2\) is unknown. This is different to the setup we consider since in our case, \(D\) is unstructured. Kass & Natarajan (2006) proposed a conjugate inverse Wishart distribution as a default prior for \(D\). 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 \(\beta\).

Cai & Dunson (2006) define a prior for the variance components matrix where \(D\) is decomposed as \(D = L \Gamma \Gamma^T L\), to ensure positive-definiteness, \(L = \text{diag}(\lambda_1, \ldots, \lambda_q)\) with \(\lambda_k \geq 0\), and \(\Gamma\) is lower triangular, with off diagonal elements, \(\gamma_{ij}\). Zero-inflated positive normal distributions are then placed on the \(\lambda_k\)'s and zero-inflated normal distributions are then placed on the \(\gamma_{ij}\)'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. \(y_{ij} \sim N(\mu, \sigma^2)\), where \(i = 1, \ldots, G\) and \(j = 1, \ldots, n^*\),
2. \(y_{ij} \sim N(\mu + u_i, \sigma^2)\), where \(u_i \sim N(0, \tau^2)\).
The DB prior for $\tau^2$ 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.

Define a default prior distribution for $D$ as the inverse-Wishart distribution with $\rho$ degrees of freedom and scale matrix, $\rho \mathbf{R}$, i.e. $D \sim IW(\rho, \rho \mathbf{R})$, where $\rho > q - 1$. We set $\rho = q$. Pawitan (2006, pg 467-468) gives the approximate Fisher information for $u_i$ as

$$I_{u_i} = Z_i^T W^{-1}_i Z_i + D^{-1}.$$

Following the approach in Section 2 we replace all unknown parameters by their prior means giving

$$Z_i^T W^{-1}_i Z_i + \mathbf{R}^{-1}.$$

We sum over the $G$ groups to obtain

$$\sum_{i=1}^G Z_i^T W^{-1}_{i,m,0} Z_i + G \mathbf{R}^{-1}.$$

If the prior for $D$ provides, approximately, the same information as one unit of data then

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

Note that, in the case of equal group sizes, $n_i = n^*$, this prior is default conjugate prior of Kass & Natarajan (2006) with $c = n^*$.

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. However, if the number of models, \(|M|\), is large, approximating the marginal likelihood of each model \(m \in M\) 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 (3) to reduce the dimension of the parameter space. We then use a reversible jump MCMC method to approximate the posterior model probabilities of all models \(m \in M\). These approximations, denoted as \(\hat{f}_L^m(y)\), 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 \(\hat{f}_B^m(y)\) and \(\hat{f}_B^m(y|m)\), respectively.

4.2. An MCMC Method

We ease the computational burden by taking advantage of the conditional independence of the \(y_i\), and write the integrated likelihood (removing the subscript \(m\)) as

\[
f(y|\beta, D, \phi) = \prod_{i=1}^G \int f(y_i|\beta, u_i, \phi)f(u_i|D)du_i.
\]

Using the Laplace approximation for each \(\int f(y_i|\beta, u_i, \phi)f(u_i|D)du_i\), we obtain the following approximation to the integrated likelihood:

\[
\hat{f}(y|\beta, D, \phi) = |D|^{-\frac{G}{2}} \prod_{i=1}^G \left[ f(y_i|\beta, \hat{u}_i, \phi)|V_i + D^{-1}|^{-\frac{1}{2}} \exp \left( -\frac{1}{2} \frac{1}{\hat{u}_i^T D^{-1} \hat{u}_i} \right) \right],
\]  

(12)

where \(V_i = -\frac{\partial^2}{\partial u_i \partial u_i} \log f(y_i|\beta, u_i)\) \(\big|_{u_i = \hat{u}_i}\). The maximum value, \(\hat{u}_i\), of \(u_i\) can be found using the Newton-Raphson method since the 1st and 2nd derivatives of \(\log f(y_i|\beta, u_i, \phi)f(u_i|D)\) with respect to \(u_i\) are readily available.

Therefore, the approximate posterior density of \(\beta, D,\) and \(\phi\) is given by:

\[
\hat{f}(\beta, D, \phi|y) \propto \hat{f}(y|\beta, D, \phi)f(\beta|D, \phi)f(D|\phi)f(\phi).
\]
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 $D = \Gamma \Gamma^T$, where $\Gamma$ is the lower-triangular matrix, which depends upon $\upsilon$, the $\frac{1}{2} q(q + 1) \times 1$ vector of transformed parameters, given by

$$
\begin{pmatrix}
\upsilon_{11} \\
\upsilon_{12} \\
\vdots \\
\upsilon_{1q}
\end{pmatrix}
\begin{pmatrix}
\upsilon_{22} \\
\vdots \\
\upsilon_{qq}
\end{pmatrix}
\begin{pmatrix}
\upsilon_{q(q+1)} \\
\vdots \\
\upsilon_{1q}
\end{pmatrix}
$$

Note that if $\upsilon \in \mathbb{R}^{\frac{1}{2} q(q + 1)}$ then $D$ 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, $(\beta, \upsilon, \omega)^T$, is given by

$$
\hat{f}(\beta, \upsilon, \omega | y) \propto \hat{h}(\beta, \upsilon, \omega | y) = \hat{f}(y | \beta, \Gamma \Gamma^T, e^\omega) f(\beta | \Gamma \Gamma^T, e^\omega) f(\Gamma \Gamma^T | e^\omega) f(e^\omega)^{2q} e^{\omega_1(q+q-1)}
$$

Note that the vector of transformed parameters, $(\beta, \upsilon, \omega)^T$, lies in $\mathbb{R}^{p+\frac{1}{2} q(q + 1)}$ if the dispersion parameter is unknown and lies in $\mathbb{R}^{p+\frac{1}{2} q(q + 1) + 1}$ 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 $\pi_m(\beta_m, \upsilon_m, \omega_m)$, is the multivariate normal distribution with mean given by the value of $(\beta_m, \upsilon_m, \omega_m)^T$ that maximises the approximate posterior density, $\hat{f}_m(\beta_m, \upsilon_m, \omega_m | y)$, (or, equivalently, $\log \hat{f}_m(\beta_m, \upsilon_m, \omega_m | y)$) and variance matrix given by the negative, inverse of the approximate Hessian matrix of $\log \hat{f}_m(\beta_m, \upsilon_m, \omega_m | y)$ with respect to $(\beta_m, \upsilon_m, \omega_m)^T$ evaluated at the maximum value. These quantities will need to be found numerically. Thus the proposal distribution is a normal approximation to the distribution with density $\hat{f}_m(\beta_m, \upsilon_m, \omega_m | y)$.

The independence sampler proceeds as follows:

1. Given the current state $(m, \beta_m, \upsilon_m, \omega_m)$, propose a new model $m^*$ with probability $\frac{1}{|M|}$. Then generate proposal model parameters $(\beta_{m^*}, \upsilon_{m^*}, \omega_{m^*})$ from the distribution with density $\pi_m$, as described above.
2. Calculate the acceptance probability, $\alpha = \min(1, a)$, where

$$
a = \frac{\hat{h}_m(\beta^*_m, \upsilon^*_m, \omega^*_m \mid y) \pi_m(\beta_m, \upsilon_m, \omega_m)}{\hat{h}_m(\beta_m, \upsilon_m, \omega_m \mid y) \pi_m(\beta^*_m, \upsilon^*_m, \omega^*_m)}
$$

3. Accept the proposed move with probability $\alpha$ and set the new state to be $(m^*, \beta^*_m, \upsilon^*_m, \omega^*_m)$. Otherwise, retain $(m, \beta_m, \upsilon_m, \omega_m)$ as the current state.

4. Repeat steps 1) to 3) for a total of $B$ iterations, for large $B$.

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

$$
M' = \left\{ m \in M : \max_{k \in M} \int f_L(\beta_k \mid y) \leq c \int f_L(\beta_m \mid y) \right\},
$$

for some constant $c > 1$. This definition aims to collect most of the posterior model probability without having to consider too large a set of models for $M'$. Madigan & Raftery (1994) recommend using $c = 20$ as an analogy with a 5% cut-off.

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 likelihood, $f(y \mid \beta, \upsilon, \phi)$, whereas the Laplace approximation we use is based on a second-order Taylor series expansion of the log likelihood, $\log f(y \mid \beta, \upsilon, \phi)$.

4.3. Bridge Sampling

Bridge sampling is a method for approximating the marginal likelihood, $f_m(y)$, of model $m \in M$. It requires a sample from the posterior distribution of model $m \in M$. Let $\theta_m = (\beta_m, \upsilon_m, D_m, \phi_m)^T$ be the vector of model parameters for model $m \in M$.

The bridge sampling estimator is given by

$$
\hat{f}_m^B(y) = \frac{1}{N_1} \sum_{i=1}^{N_1} f_m(y \mid \beta_m^i) f_m(\beta_m^i) \gamma(\beta_m^i)
\frac{1}{N_2} \sum_{i=1}^{N_2} g_m(\theta_m^i) \gamma(\theta_m^i),
$$

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

11
Meng & Wong (1996) showed that, with respect to minimising the mean squared error, the optimal \( \gamma() \) is given by

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

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

\[
f_m(y)^{(t+1)} = \frac{1}{N_2} \sum_{i=1}^{N_2} \frac{1}{N_1 + N_2 f_m(y^{|i})},
\]

where \( I_t = \frac{f_m(y|\theta_m) f_m(\theta_m)}{g_m(\theta_m)} \) and \( I_t = \frac{f_m(y|\theta_m) f_m(\theta_m)}{g_m(\theta_m)} \). The scheme (14) is iterated until convergence.

Let \( N_1 + N_2 = 2N \). 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 than the optimal allocation of sample sizes. In what follows, we take \( N_1 = N_2 = N \).

There remains a choice for the distribution with density \( g_m(\theta_m) \). From practice, it appears that bridge sampling performs best when \( g_m(\theta) \) ‘mimics’ the posterior density \( f_m(y|\theta_m) \). We partition the parameters as \( \theta_m = (\gamma_m^T, \phi_m^T)^T \), where \( \gamma_m = (\beta_m, u_m)^T \in \mathbb{R}^{p+Gq} \). We choose \( g_m(\theta_m) = g_{m,1}(\gamma_m) g_{m,2}(d_m) g_{m,3}(\phi_m) \), where the distributions with densities \( g_{m,1}(\cdot) \), \( g_{m,2}(\cdot) \), and \( g_{m,3}(\cdot) \) are approximations to the posterior distributions: \( \beta_m, u_m, d_m \) and \( \phi_m \), respectively.

We set \( g_{m,1}(\cdot) \) equal to the density of the multivariate normal distribution with mean (or mode) and variance chosen to be approximately the mean (or mode) and variance of \( \beta_m, u_m \). The \( g_{m,2}(\cdot) \) is the density of the inverse Wishart distribution with \( q + G \) degrees of freedom and mean approximately matched to that of \( d_m \). Likewise, \( g_{m,3}(\cdot) \) is the density of the inverse gamma distribution with mean (or mode) and variance approximately matched to that of \( \phi_m \).

When the dimension of the parameter space is not too large, the mode and curvature of the posterior distributions can be found numerically. If the dimension of the parameter space is large, then we use sample statistics of the posterior sample, \( \{\theta_m^l\} \). However, if sample statistics of the entire posterior sample are used, then this leads to correlation between the moments of the distribution with density \( g_m(\cdot) \) and the posterior sample, \( \{\theta_m\} \), and an apparent underestimation of \( f_m(y) \) (see the Appendix).

We propose to use a proportion, \( \psi \), of the posterior sample to estimate the posterior moments. The remainder of the posterior sample can then be used in the bridge sampler, (14). From practice, it appears that \( \psi = \frac{1}{2} \) lead to the smallest mean squared error of the approximations. Therefore, we need a posterior sample of size \( 2N \).
This method of matching the first two moments of the posterior to the distribution with density \( g_m() \) is called Warp-II bridge sampling by Sinharay & Stern (2005).

Sinharay & Stern (2005) only consider cases where \( q = 1 \), i.e., there only exists a group-specific intercept parameter. They approximated the integrated likelihood using Simpson’s rule, which is impractical for \( q > 1 \), and then used bridge sampling with this smaller dimension parameter space to find the marginal likelihood. Our approach is to use bridge sampling with the full parameter space.

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.

For models with \( q = 1 \), i.e., there only exists a group-specific intercept parameter, we use the Gibbs sampler OpenBUGS (Thomas et al (2006)) run remotely in the statistical package R (R Development Core Team (2008)) to obtain the posterior sample.

When \( q > 1 \), OpenBUGS has difficulties with the matrix operations required for the unit information priors. Instead, we use a variable-at-a-time Gibbs sampler implemented in R. For the regression parameters, \( \beta \), the group-specific parameters, \( u \), and the dispersion parameter, \( \phi \), we use the adaptive rejection Metropolis-Hastings sampling (ARMS) algorithm (see Gilks et al (1995)). ARMS can be implemented in R using the arms function in the HI package.

For the variance components matrix, we take advantage of the conditional conjugacy of the full conditional distribution of \( D \), i.e.

\[
D | y, \beta, u, \phi \sim IW \left( q + G, qR + \sum_{i=1}^{G} u_i u_i^T \right)
\]

There exist efficient algorithms for simulating from the inverse-Wishart distribution.

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:

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

and \( i = 1, ..., G \) and \( j = 1, ..., n_i = n^* \). We generate datasets with \( n = 200 \) observations in either \( G = 25 \) or \( G = 4 \) groups meaning \( n^* = 8 \) or \( n^* = 50 \), respectively. The intercept parameter, \( \beta_0 \), will be held fixed at \( \frac{1}{2} \), whereas \( \beta_1 \) and \( \sigma^2 \) will be drawn from the \( \mathcal{U}[0, \frac{5}{4}] \) distribution for Poisson responses and \( \mathcal{U}[0, 5] \) distribution for Bernoulli responses. Once the responses have been generated, the marginal likelihood is found, using bridge sampling, of the following five models:

1. \( \eta_{ij} = \beta_0 \),
2. \( \eta_{ij} = \beta_0 + \beta_1 x_{ij} \),
3. \( \eta_{ij} = \beta_0 + u_i; \text{ where } u_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2) \),
4. \( \eta_{ij} = \beta_0 + u_i + \beta_1 x_{ij}; \text{ where } u_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2) \),
5. \( \eta_{ij} = \beta_0 + u_0 + (\beta_1 + u_1)x_{ij}; \text{ where } (u_0, u_1)^T \overset{iid}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{D}) \),

where the priors described in Sections 2 and 3 are applied to the appropriate parameters. We repeat this process for a total of 1000 datasets.

We present the results of the simulation study graphically. 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 \( \beta_1 \) 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) against the value of the \( \sigma^2 \) parameter. Figures 1 and 2 show these plots for the Poisson and Bernoulli responses, respectively. We add smoothing splines to the plots.

We see from Figures 1 and 2 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 Figure 2 which shows the total posterior model probability of models 3, 4 and 5 plotted against the true value of the \( \sigma^2 \) parameter. The smoothing spline appears to not approach one for large values of \( \sigma^2 \). We see that even for large values of \( \sigma^2 \) 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 \( \sigma^2 \), i.e. the observed variance of \( u_i \) for \( i = 1, ..., G \) being significantly smaller than the
Figure 1: Plots of total posterior model probabilities for the 1000 datasets with Poisson responses, Top row: \( n^* = 8 \) and \( G = 25 \), Bottom row: \( n^* = 50 \) and \( G = 4 \)
Figure 2: Plots of total posterior model probabilities for the 1000 datasets with Bernoulli responses. Top row: $n^* = 8$ and $G = 25$, Bottom row: $n^* = 50$ and $G = 4$. 

- Top row: $n^* = 8$, $G = 25$.
true value of $\sigma^2$. To see this, consider Figure 3 which shows the total posterior model probability of models 3, 4 and 5 plotted against the observed value, $s^2$, of $\sigma^2$. We see from Figure 3 that the posterior model probability approaches and reaches one as the observed $\sigma^2$ 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 0% and 0.3% for Poisson responses, for $G = 25$ and $G = 4$, respectively, and 1.8% and 4.5% for Bernoulli responses, for $G = 25$ and $G = 4$, respectively.

6. Examples

We demonstrate our strategy on two real datasets: Natural Selection Study Data and Six Cities Data.

6.1. A Natural Selection Study

SinhaRay & Stern (2005) presented data from 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 \( j \)th turtle in the \( i \)th clutch, \( i = 1, ..., 31, j = 1, ..., n, \) and \( y_{ij} \mid p_{ij} \sim \text{Bernoulli}(p_{ij}), \) where \( p_{ij} = \Phi(\eta_{ij}). \) We consider 5 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. \( \eta_{ij} = \beta_0 + u_i + (\beta_1 + v_i)x_{ij}, \) where \( (u_i, v_i)^T \overset{\text{iid}}{\sim} \text{N}(0, \mathbf{D}). \)

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. Table 1 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 \( \beta_k \sim \text{N}(0, 10^5), \sigma^2 \sim \text{Inv-Gamma}(0.00005, 0.5) \) for models 3 and 4, and \( \mathbf{D} \sim \text{IW}(2, 2 \mathbf{I}_2) \) for model 5. These priors are proposed by Natarajan & Kass (2000).

The results in Table 1 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.

6.2. Six Cities Data

The Six Cities Data is frequently used to assess mixed models methodology. The data consists of the wheezing status, \( y_{ij} (0=\text{not wheezing}, 1=\text{wheezing}), \) of child \( i \) at time-point \( j, \) for \( i = 1, ..., 537 \) and \( j = 1, ..., 4. \) Also included, is the age of the \( i \)th child, \( x_{1ij}, \) at time-point \( j \) and the smoking status, \( x_{2ij}, \) of the \( i \)th child’s mother at time-point \( j. \) Note that \( x_{2ij} = x_{2k} \text{ for }
Table 1: Approximated posterior model probabilities and BIC, AIC and DIC for the 5 models of the Natural Selection Study

<table>
<thead>
<tr>
<th>Model</th>
<th>Posterior Model Probabilities</th>
<th>BIC&lt;sub&gt;m&lt;/sub&gt;</th>
<th>AIC&lt;sub&gt;m&lt;/sub&gt;</th>
<th>DIC&lt;sub&gt;m&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bridge Sampling</td>
<td>Independence Sampler</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.00016</td>
<td>0.00005</td>
<td>325.68</td>
<td>322.18</td>
</tr>
<tr>
<td>2</td>
<td>0.88322</td>
<td>0.88360</td>
<td>308.65</td>
<td>301.66</td>
</tr>
<tr>
<td>3</td>
<td>0.00095</td>
<td>0.00060</td>
<td>323.74</td>
<td>316.75</td>
</tr>
<tr>
<td>4</td>
<td>0.09941</td>
<td>0.10440</td>
<td>311.90</td>
<td>301.41</td>
</tr>
<tr>
<td>5</td>
<td>0.01625</td>
<td>0.01135</td>
<td>321.90</td>
<td>304.41</td>
</tr>
</tbody>
</table>

all \( j, k \in \{1, \ldots, 4\} \). We can also define the interaction covariate \( x_{3ij} = x_{1ij}x_{2ij} \). By considering all possible models with the canonical logit link, and adhering to the modelling convention of not including an interaction covariate unless all marginal covariates are included, there are 19 possible models. It is impractical 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. \( \eta_{ij} = \beta_0 + u_i; \ u_i \sim N(0, \sigma^2) \),
7. \( \eta_{ij} = \beta_0 + \beta_1 x_{1ij} + u_i; \ u_i \sim N(0, \sigma^2) \).
8. \( \eta_{ij} = \beta_0 + \beta_2 x_{2ij} + u_i; \ u_i \sim N(0, \sigma^2) \).
9. \( \eta_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + u_i; \ u_i \sim N(0, \sigma^2) \).

These four models account for 94.61% 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 \((u_{0i}, u_{1i})^T \sim N(0, D)\). This model has \( \hat{f}(11 | y, M) = 0.0283 \).

We then used bridge sampling with \( N = 100000 \), to obtain approximations to the marginal likelihoods, \( \hat{f}^{BS}(y|m, M') \), and posterior model probabilities, \( \hat{f}^{BS}(m|y, M') \), conditional on \( M' \). These are shown in Table 2.
Table 2: Approximated posterior model probabilities for the 4 models in $M'$ for the Six Cities Data

| m | $\hat{f}^L(m|y)$ | $\hat{f}^L(m|M')$ | $\hat{f}^{BS}(y|m|M')$ | $\hat{f}^{BS}(m|y|M')$ |
|---|----------------|--------------------|----------------------|----------------------|
| 6 | 0.3172         | 0.3353             | -808.0410            | 0.3397               |
| 7 | 0.5679         | 0.6003             | -807.4897            | 0.5895               |
| 8 | 0.0294         | 0.0311             | -810.6490            | 0.0250               |
| 9 | 0.0316         | 0.0334             | -810.0459            | 0.0457               |

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 $u_i \sim N(0, \sigma^2)$, with model 7 second. DIC chooses the most complicated model, i.e. the model with a group-specific parameter for all of the four regression parameters. The DIC values are based on the following priors: we assume that the regression parameters are independent and $\beta_k \sim N(0, 10^5)$, $\sigma^2 \sim \text{Inv-Gamma}(0.00005, 0.5)$ for models where $q = 1$ and the variance components are univariate, and $D \sim \text{IW}(q, qI_2)$ for models where $q > 1$. These priors are proposed by Natarajan & Kass (2000).

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.

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 $\hat{f}^L(m|y)$; an approximation to the posterior model probabilities. These $\hat{f}^L(m|y)$ 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 $\hat{f}^{BS}(m|y)$. 

20
The strategy considered allows a fully Bayesian analysis of GLMMs under model uncertainty and weak prior information, without the need of choosing arbitrary hyperparameters. From the examples in Section 6, it appears that we obtain similar model determination conclusions as when using BIC.

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 \( \hat{h}_m(\beta_m, v_m, \omega_m|y) \) 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 \( \hat{h}_m(\beta_m, v_m, \omega_m|y) \) for each \( m \in M \). Future work will focus on developing this methodology.

8. Appendix - Bridge Sampling

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

For some models, the dimension of the model parameters, \( \theta \), may be large which prohibits finding the mean and variance (or mode and curvature) deterministically. Since we have a sample from the posterior distribution, a naive approach may be to approximate the mean and variance of \( \theta|y \) 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 \( \theta \) and
Figure 4: Plots of the approximated log-marginal likelihood for the two different approaches and three different dimensions against the sample size.

\[ f(\theta|y) \propto \exp\left(-\frac{1}{2} \theta^T \theta\right) \]

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

We have a sample \( \{\theta_i\}_{i=1}^{2N} \) of size \( 2N \) from \( \mathcal{N}(\mathbf{0}, \mathbf{I}_k) \) which represents our posterior sample. All that remains is to choose appropriate values for \( \mu \) and \( \Sigma \), 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, \( \{\theta_i\}_{i=1}^{2N} \), and use the bridge sampler (14) with a sample size of \( N_1 = N_2 = 2N \).
2. Approach 2 (Our approach). Use the sample mean and variance of half of the posterior sample, \( \{\theta_i\}_{i=1}^{N} \). Use the second half of the posterior sample, \( \{\theta_i\}_{i=N+1}^{2N} \), in the bridge sampler (14) with a reduced sample size of \( N_1 = N_2 = N \).
The sample sizes, $2N$, 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 4 shows plots of the approximated log-marginal likelihood for the two different approaches against the sample size, $2N$. Also included on the plot is a line at the true log-marginal likelihood, $\frac{1}{2} \log(2\pi)$. 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 $E\left(\frac{X}{Y}\right) > E(X)E(Y)$, for positive random variables $X$ and $Y$.

References


