Thesis

Antony Overstall

March 9, 2010

University of Southampton

Faculty of Engineering, Science and Mathematics

School of Mathematics

Default Bayesian Model Determination for Generalised Linear Mixed Models

by

Antony Marshall Overstall

Thesis for the degree of Doctor of Philosophy

March 2010

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF ENGINEERING, SCIENCE AND MATHEMATICS

SCHOOL OF MATHEMATICS

Doctor of Philosophy

DEFAULT BAYESIAN MODEL DETERMINATION FOR GENERALISED LINEAR MIXED MODELS

Antony Marshall Overstall

In this thesis, an automatic, default, fully Bayesian model determination strategy for GLMMs is considered. This strategy must address the two key issues of default prior specification and computation.

Default prior distributions for the model parameters, that are based on a unit information concept, are proposed.

A two-phase computational strategy, that uses a reversible jump algorithm and implementation of bridge sampling, is also proposed.

This strategy is applied to four examples throughout this thesis.

Contents

List of Tables

List of Figures

Authors' Declaration

I, Antony Marshall Overstall, declare that the thesis entitled

Default Bayesian Model Determination for Generalised Linear Mixed Models,

and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- part of this work is being considered for publication in the Computational Statistics and Data Analysis journal as Overstall and Forster (2009).

Signed:.....

Date:.....

Acknowledgements

I would like to thank my supervisor Professor Jon Forster for his help with completing this thesis.

Chapter 1

Introduction

1.1 Bayesian Inference

1.1.1 Bayes' Theorem

Suppose $\mathbf{y}$ is a $n \times 1$ vector of *responses* with joint *probability density function* (pdf), $f(\mathbf{y}|\boldsymbol{\theta})$, which depends on the $k \times 1$ vector of unknown *model parameters*, $\boldsymbol{\theta} \in \Theta$, where $\Theta \subseteq \mathbb{R}^k$ is known as the *parameter space*. In Bayesian inference, both $\mathbf{y}$ and $\boldsymbol{\theta}$ are considered to be random variables, so $\boldsymbol{\theta}$ has a probability distribution with pdf, $f(\boldsymbol{\theta})$. Their joint pdf can be written as

$$\begin{aligned} f(\mathbf{y}, \boldsymbol{\theta}) &= f(\mathbf{y}|\boldsymbol{\theta}) f(\boldsymbol{\theta}) && (1.1) \\ &= f(\boldsymbol{\theta}|\mathbf{y}) f(\mathbf{y}). && (1.2) \end{aligned}$$

Equating (1.1) and (1.2) gives

$$\begin{aligned} f(\boldsymbol{\theta}|\mathbf{y}) &= \frac{f(\mathbf{y}|\boldsymbol{\theta}) f(\boldsymbol{\theta})}{f(\mathbf{y})} \\ &= \frac{f(\mathbf{y}|\boldsymbol{\theta}) f(\boldsymbol{\theta})}{\int_\Theta f(\mathbf{y}|\boldsymbol{\theta}) f(\boldsymbol{\theta}) d\boldsymbol{\theta}}. && (1.3) \end{aligned}$$

In the case of a discrete-valued $\boldsymbol{\theta}$, replace the integration in the denominator of (1.3) by a summation. The identity (1.3) is known as *Bayes' theorem*. Note that the denominator of (1.3) does not depend on $\boldsymbol{\theta}$, so Bayes' theorem can be rewritten as

$$f(\boldsymbol{\theta}|\mathbf{y}) \propto f(\mathbf{y}|\boldsymbol{\theta}) f(\boldsymbol{\theta}). \quad (1.4)$$

The quantity $f(\mathbf{y}|\boldsymbol{\theta})$ is equivalent to the *likelihood function*. The quantity $f(\boldsymbol{\theta})$ is the pdf of the *prior distribution* which reflects our knowledge of $\boldsymbol{\theta}$ prior to observing the data, $\mathbf{y}$. The quantity $f(\boldsymbol{\theta}|\mathbf{y})$ is the pdf of the *posterior distribution* of $\boldsymbol{\theta}|\mathbf{y}$, which reflects our updated knowledge of $\boldsymbol{\theta}$, having observed the data, $\mathbf{y}$, and combined the information from the

data with the information from the prior distribution, using Bayes' theorem. Using this terminology, (1.4) can be written as

$$\text{posterior} \propto \text{likelihood} \times \text{prior}.$$

1.1.2 Posterior Inference

Posterior inference can be undertaken either in an informal way; by the way of summary quantities, e.g. mean, mode, variance, etc, or in a formal way; by way of making optimal decisions based on minimising posterior expectations of appropriate loss functions.

Summary quantities of the posterior distribution of $\boldsymbol{\theta}|\mathbf{y}$ can be found via the posterior pdf, $f(\boldsymbol{\theta}|\mathbf{y})$. For instance, suppose $g : \Theta \to \mathbb{R}$,

$$\mathrm{E}(g(\boldsymbol{\theta})|\mathbf{y}) = \int_{\Theta} g(\boldsymbol{\theta}) f(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta},$$

gives us the posterior expectation of a function, $g(\boldsymbol{\theta})$, of the parameter $\boldsymbol{\theta}$. Setting $g(\boldsymbol{\theta}) = \theta_j$, gives us the posterior mean of the jth component of $\boldsymbol{\theta}$.

Another useful summary is a *probability* or *credible interval*, which is the Bayesian equivalent of a classical confidence interval. A $100(1-\alpha)\%$ probability interval for the jth component, θ_j, of $\boldsymbol{\theta}$ is the interval (a, b) such that

$$\mathrm{P}(a < \theta_j < b|\mathbf{y}) = 1 - \alpha.$$

Similarly, we can calculate the specific posterior probability that $\boldsymbol{\theta}$ lies in some set $\Omega \subseteq \Theta$, i.e.

$$\mathrm{P}(\boldsymbol{\theta} \in \Omega|\mathbf{y}) = \int_{\Omega} f(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}.$$

This is the Bayesian equivalent of a classical hypothesis test.

Formally, we may want a point estimate, $\tilde{\boldsymbol{\theta}}(\mathbf{y})$ or $\tilde{\boldsymbol{\theta}}$, of the parameter $\boldsymbol{\theta}$. The decision-theoretic approach is to minimise the posterior expectation of an appropriate loss function. Consider estimating the jth component, θ_j, of $\boldsymbol{\theta}$ using the *squared error* loss function:

$$L(\theta_j, \tilde{\theta}_j) = (\theta_j - \tilde{\theta}_j)^2.$$

The posterior expectation of $L(\theta_j, \tilde{\theta}_j)$ is

$$\mathrm{E}(L(\theta_j, \tilde{\theta}_j)|\mathbf{y}) = \mathrm{E}(\theta_j^2|\mathbf{y}) - 2\tilde{\theta}_j \mathrm{E}(\theta_j|\mathbf{y}) + \tilde{\theta}_j^2.$$

This is minimised at the posterior mean, i.e. $\tilde{\theta}_j = \mathrm{E}(\theta_j|\mathbf{y})$. Similarly, the *absolute error* loss function, $L(\theta_j, \tilde{\theta}_j) = |\theta_j - \tilde{\theta}_j|$, results in $\tilde{\theta}_j$ being the posterior median. The loss function

$$L(\theta_j, \tilde{\theta}_j) = \begin{cases} 0, & \text{if } |\theta_j - \tilde{\theta}_j| \leq \delta, \\ 1, & \text{if } |\theta_j - \tilde{\theta}_j| > \delta, \end{cases}$$

for small δ, results in $\tilde{\theta}_j$ being the posterior mode.

1.1.3 Model Determination

The *data-generating process* is the complete description of a random process from which the data, $\mathbf{y}$, arise. The true data-generating process would only be known in the presence of full information. In the absence of full information, there is uncertainty about the data-generating process. This uncertainty is quantified by a *statistical model*, which is a set of data-generating processes, expressed by the likelihood function, $f(\mathbf{y}|\boldsymbol{\theta})$ with $\boldsymbol{\theta} \in \Theta$. In addition to $f(\mathbf{y}|\boldsymbol{\theta})$, a Bayesian statistical model possesses a prior distribution with pdf, $f(\boldsymbol{\theta})$.

In practice, we may be unsure about how to construct the form of the model and actually propose several alternative models. There will be uncertainty amongst these models. To overcome this we assume a single *encompassing model*. This is a union of all alternative models. We now need to express our prior knowledge about the alternative models which we do via prior model probabilities.

Suppose we have a set, M, of alternative models, where a model, $m \in M$, has a likelihood $f_m(\mathbf{y}|\boldsymbol{\theta}_m) = f(\mathbf{y}|\boldsymbol{\theta}_m, m)$ and a prior distribution with pdf $f_m(\boldsymbol{\theta}_m) = f(\boldsymbol{\theta}_m|m)$ where $\boldsymbol{\theta}_m \in \Theta_m$, and $\boldsymbol{\theta}_m$ is a $k_m \times 1$ vector.

The encompassing model is then the set of data-generating processes

$$\{f_m(\mathbf{y}|\boldsymbol{\theta}_m) : \boldsymbol{\theta}_m \in \Theta_m, m \in M\},$$

with parameter $\boldsymbol{\theta} = (\boldsymbol{\theta}_m, m) \in \Theta$ where

$$\Theta = \bigcup_{m \in M} \{m\} \times \Theta_m.$$

For Bayesian inference we require a prior distribution for $\boldsymbol{\theta} = (\boldsymbol{\theta}_m, m)$ with pdf decomposed as

$$\begin{aligned} f(\boldsymbol{\theta}_m, m) &= f(\boldsymbol{\theta}_m|m)f(m) \\ &= f_m(\boldsymbol{\theta}_m)f(m), \end{aligned}$$

where $f_m(\boldsymbol{\theta}_m)$ is the prior pdf of $\boldsymbol{\theta}_m$ conditional on model m. The quantity $f(m)$ is the *prior model probability*, where $f(m) > 0$ and $\sum_{m \in M} f(m) = 1$.

Using Bayes' theorem we can find the joint posterior pdf of $\boldsymbol{\theta}_m$ and m:

$$\begin{aligned} f(\boldsymbol{\theta}_m, m|\mathbf{y}) &= \frac{f(\mathbf{y}|\boldsymbol{\theta}_m, m)f(\boldsymbol{\theta}_m|m)f(m)}{f(\mathbf{y})}, \\ &= \frac{f_m(\mathbf{y}|\boldsymbol{\theta}_m)f_m(\boldsymbol{\theta})}{f_m(\mathbf{y})} \times \frac{f_m(\mathbf{y})f(m)}{f(\mathbf{y})}, \end{aligned} \tag{1.5}$$

where the quantity $f_m(\mathbf{y})$ is known as the *marginal likelihood* and is given by

$$f_m(\mathbf{y}) = \int_{\Theta_m} f_m(\mathbf{y}|\boldsymbol{\theta}_m)f_m(\boldsymbol{\theta}_m)d\boldsymbol{\theta}_m. \tag{1.6}$$

Note from the second part of the right-hand side of (1.5) that

$$\frac{f_m(\mathbf{y})f(m)}{f(\mathbf{y})} = \frac{f_m(\mathbf{y})f(m)}{\sum_{m\in M} f_m(\mathbf{y})f(m)} = f(m|\mathbf{y}), \tag{1.7}$$

where $f(m|\mathbf{y})$ is known as the *posterior model probability* of model $m \in M$. Again, Bayes' theorem is used to update the prior model probabilities to the posterior model probabilities in light of observing the data, $\mathbf{y}$.

Therefore, we can extend the idea of posterior inference as described in Section 1.1.2 to that of posterior inference under model uncertainty:

1. Evaluation of the posterior model probability, $f(m|\mathbf{y})$, for each $m \in M$ (Model Determination),
2. Evaluation and summarisation of the posterior distribution, $f_m(\boldsymbol{\theta}_m|\mathbf{y})$ of the parameters, $\boldsymbol{\theta}_m$ of each model $m \in M$ (Posterior Inference).

Fisher (1922) stated that there are three aspects to valid statistical inference: a) model specification, b) estimation of the model parameters, and c) estimation of precision. Evaluation of the posterior distribution of the parameters is equivalent to b) and c) of Fisher's system. Burnham and Anderson (1998) partition the model specification/determination aspect into two parts: forming a set of candidate models and model selection. They go on to discuss how the formation of a set of candidate models is "where the scientific and biological information formally enter the investigation". In this thesis, we assume the position that a set of candidate models, M, has already been chosen.

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

$$\frac{f(1|\mathbf{y})}{1-f(1|\mathbf{y})} = \frac{f(1|\mathbf{y})}{f(2|\mathbf{y})} = \frac{f(1)f_1(\mathbf{y})}{f(2)f_2(\mathbf{y})} = \frac{f(1)}{1-f(1)}\frac{f_1(\mathbf{y})}{f_2(\mathbf{y})}, \tag{1.8}$$

where $f(1)$ and $f(2)$ are the prior model probabilities of models 1 and 2, respectively. The quantity $f_1(\mathbf{y})/f_2(\mathbf{y})$ of the ratio of marginal likelihoods is known as the *Bayes' factor* in favour of model 1. So (1.8) can be written

$$\text{posterior odds} = \text{prior odds} \times \text{Bayes' factor}.$$

1.1.4 Prior Distributions

The prior distribution is a contentious issue in Bayesian inference. As defined in Section 1.1 it represents all of our knowledge about the model parameters, $\boldsymbol{\theta}$, prior to observing the data, $\mathbf{y}$.

There are two problems with this: 1) some critics of Bayesian inference argue that statistical inference should be objective and personal opinions in the form of the prior distribution should not be considered, and 2) what happens in the case where there is little prior knowledge?

The approach taken in this thesis is that we are in exactly the position of having little prior knowledge. As a result of this, our goal is *objective Bayesian inference*. To achieve this, we need to consider prior distributions which have a negligible effect on the posterior inference, i.e. they are dominated by the likelihood. Box and Tiao (1992) and Kass and Wasserman (1996) refer to these types of prior as reference priors, i.e. as a point of 'reference'. However, the term reference prior is used elsewhere in the literature to refer to a specific type of prior. We use the term *default prior*. The resulting analysis is called the *default analysis*.

An obvious and natural default prior to use is the non-informative uniform prior, where $f(\boldsymbol{\theta}) \propto 1$. Therefore, $f(\boldsymbol{\theta}|\mathbf{y}) \propto f(\mathbf{y}|\boldsymbol{\theta})$. Under this prior, the posterior mode of $\boldsymbol{\theta}$ is equivalent to the maximum likelihood estimate of $\boldsymbol{\theta}$. However, there may not exist a proper prior with $f(\boldsymbol{\theta}) \propto 1$. Specifically, this is the case if any part of the parameter space, Θ, is unbounded. In this case, Lindley's paradox (see Section 1.1.5) applies and we cannot use the uniform prior. Also, a uniform prior distribution for a parameter $\boldsymbol{\theta}$, will not necessarily be a uniform prior distribution for a transformation of $\boldsymbol{\theta}$, say $\boldsymbol{\phi} = h(\boldsymbol{\theta})$. We must, instead, use an informative distribution which has a negligible effect on posterior inference as our default prior. The construction of default priors is an active area of research for the practical application of Bayesian methods.

It may well be the case that prior knowledge does exist and it is the opinion of some that prior knowledge always exists and should be elicited into a prior distribution for $\boldsymbol{\theta}$. This is, in itself, a non-trivial problem. Use of this type of prior results in *subjective Bayesian inference*.

When considering a prior distribution it may be convenient to use a *conjugate prior distribution*. A conjugate prior distribution has pdf $f(\boldsymbol{\theta}) \in \mathcal{F}$, if the posterior distribution pdf $f(\boldsymbol{\theta}|\mathbf{y}) \in \mathcal{F}$, i.e. the posterior distribution is from the same family of distributions, $\mathcal{F}$, as the prior distribution. Examples include:

1. Suppose $y_i \overset{\text{iid}}{\sim} \text{N}(\theta, \sigma^2)$, for $i = 1, ..., n$ and where σ^2 is known. The conjugate prior distribution for θ is $\theta \sim \text{N}(\mu, \tau^2)$. Then
$$\theta|\mathbf{y} \sim \text{N}\left(\frac{n\tau^2\bar{y} + \sigma^2\mu}{n\tau^2 + \sigma^2}, \frac{\sigma^2\tau^2}{n\tau^2 + \sigma^2}\right).$$

2. Suppose $y_i \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta)$, for $i = 1, ..., n$. The conjugate prior distribution for θ is $\theta \sim \text{Beta}(\alpha, \beta)$. Then
$$\theta|\mathbf{y} \sim \text{Beta}(\alpha + n\bar{y}, \beta + n - n\bar{y}).$$

1.1.5 Lindley's Paradox

Lindley's paradox is best explained, initially, with a simple example which is adapted from O'Hagan and Forster (2004, pg. 77). Suppose we have one observation, y, and we wish to compare the following two models:

1. $y \sim \mathrm{N}(0, \sigma^2)$,
2. $y \sim \mathrm{N}(\theta, \sigma^2)$, where $\theta \sim \mathrm{N}(0, \tau^2)$,

where in both cases, σ^2 is known. Model 1 is completely specified, whereas Model 2 has an unknown mean, θ. In addition, suppose the following prior model probabilities: $f(1) = p$ and $f(2) = 1 - p$. The posterior model probability of Model 1 is

$$\begin{aligned} f(1|y) &= \frac{\frac{p}{\sqrt{\sigma^2}} \exp\left(-\frac{y^2}{2\sigma^2}\right)}{\frac{p}{\sqrt{\sigma^2}} \exp\left(-\frac{y^2}{2\sigma^2}\right) + \frac{1-p}{\sqrt{\sigma^2+\tau^2}} \exp\left(-\frac{y^2}{2(\sigma^2+\tau^2)}\right)} \\ &\to 1 \qquad \text{as} \qquad \tau^2 \to \infty, \end{aligned}$$

provided $p \neq 0$. So, regardless of the observation collected, the posterior model probability of Model 1 will tend to 1 as the variance of the prior approaches infinity.

This result is not specific to this example and can occur in any situation where the prior distribution becomes improper over part of the parameter space.

Lindley's paradox means that we cannot simply choose a prior with $f(\boldsymbol{\theta}) \propto 1$ when there is weak prior information and model uncertainty.

In fact, the problem of specifying a prior distribution under weak prior information and model uncertainty is deeper than this. The Bayes' factor, in this situation, can be very sensitive to the choice of prior distribution. When comparing two nested models, the Bayes' factor for the simpler model can be made arbitrarily large by choosing a large enough prior variance for the augmented parameter.

This is the motivation behind default priors, i.e. informative prior distributions that have a negligible effect on posterior inference but provide a consistent amount of information across all models for us to undertake model determination.

1.1.6 Hierarchical Models

Suppose we have a model, where the joint pdf of the data, $\mathbf{y}$, depends on the vector of parameters, $\boldsymbol{\omega}$. The prior distribution for $\boldsymbol{\omega}$ will, in general, depend on the vector of *hyper-parameters*, $\boldsymbol{\lambda}$. If the prior distribution is completely specified then $\boldsymbol{\lambda}$ are known. However,

in the case of a *hierarchical model*, $\boldsymbol{\lambda}$ are unknown and, therefore, have their own prior distribution known as the *hyper-prior*. We can write the model parameters as $\boldsymbol{\theta} = (\boldsymbol{\omega}, \boldsymbol{\lambda})^T$ and decompose the prior pdf as

$$\begin{aligned} f(\boldsymbol{\theta}) &= f(\boldsymbol{\omega}, \boldsymbol{\lambda}) \\ &= f(\boldsymbol{\omega}|\boldsymbol{\lambda}) f(\boldsymbol{\lambda}). \end{aligned}$$

Obviously, the hyper-prior can depend on a further vector of *hyper-hyper-parameters* or *2-hyper-parameters* with a *2-hyper-prior* distribution. This hierarchy can be extended to a *ν-hyper-prior*.

1.2 Generalised Linear Mixed Models

Generalised linear mixed models (GLMMs) are useful when responses, which may be non-normal, depend on a set of covariates and are correlated due to the existence of groups or clusters. GLMMs are an extension of linear mixed models (LMMs) to non-normal responses and an extension of generalised linear models (GLMs) to correlated responses. A Bayesian GLMM is a hierarchical model as described in Section 1.1.6.

GLMMs are often referred to as GLMs with random effects in the classical literature.

1.2.1 Specification

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 regression and group-specific covariates, respectively. These covariates are a subset of the available explanatory variables or products of available explanatory variables. Let the total sample size be $n = \sum_{i=1}^{G} n_i$. Conditional on the ith group-specific parameters, $\mathbf{u}_i$, we assume that Y_{ij} is independently distributed from some exponential family distribution with density

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

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

$$g(\mu_{ij}) = \eta_{ij} = \mathbf{x}_{ij}^T\boldsymbol{\beta} + \mathbf{z}_{ij}\mathbf{u}_i, \tag{1.9}$$

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 *ith group-specific parameters*.

Suppose $\mathbf{y}_i = (y_{i1}, ..., y_{in_1})^T$, $\mathbf{X}_i = (\mathbf{x}_{i1}, ..., \mathbf{x}_{in_i})^T$, $\mathbf{Z}_i = (\mathbf{z}_{i1}, ..., \mathbf{z}_{in_i})^T$, $\boldsymbol{\eta}_i = (\eta_{i1}, ..., \eta_{in_1})^T$, $\boldsymbol{\mu}_i = (\mu_{i1}, ..., \mu_{in_1})^T$, 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 $\mathbf{y} = (\mathbf{y}_1^T, ..., \mathbf{y}_G^T)^T$, $\mathbf{X} = (\mathbf{X}_1^T, ..., \mathbf{X}_G^T)^T$, $\mathbf{Z} = \text{diag}(\mathbf{Z}_1, ..., \mathbf{Z}_G)$, $\boldsymbol{\eta} = (\boldsymbol{\eta}_1^T, ..., \boldsymbol{\eta}_G^T)^T$, $\boldsymbol{\mu} = (\boldsymbol{\mu}_1^T, ..., \boldsymbol{\mu}_G^T)^T$, and $\mathbf{u} = (\mathbf{u}_1^T, ..., \mathbf{u}_G^T)^T$, then (1.9) can be rewritten in matrix form as

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

where $g()$ has been applied element-wise, i.e. $g(\boldsymbol{\mu}) = (g(\mu_{11}), ..., g(\mu_{Gn_G}))^T$.

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 also adhere to the modelling principle that if a column of $\mathbf{X}_i$ (or $\mathbf{Z}_i$) is formed from the interaction between two explanatory variables, then those two explanatory variables must have columns also in $\mathbf{X}_i$ (or $\mathbf{Z}_i$).

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

The group-specific parameters, $\mathbf{u}$, are often referred to as *random effects*. If $\mathbf{u}_i \overset{\text{iid}}{\sim} \text{N}(\mathbf{0}, \mathbf{D})$, for $i = 1, ..., G$, then this is known as an *exchangeable random effect structure*.

A GLM is a special case of a GLMM, where $\mathbf{Z} = \mathbf{0}$ and $n_i = 1$.

1.2.2 Likelihood for GLMMs

We define the *first-stage likelihood function* for a GLMM as

$$f(\mathbf{y}|\boldsymbol{\beta}, \mathbf{u}, \phi) = \prod_{i=1}^{G} \prod_{j=1}^{n_i} \exp\left[\frac{y_{ij}\zeta_{ij} - b(\zeta_{ij})}{a_{ij}(\phi)} + c(y_{ij}; \phi)\right]. \tag{1.10}$$

Classical inference for GLMMs is based on maximising the *integrated likelihood function*

$$\begin{aligned} f(\mathbf{y}|\boldsymbol{\beta}, \mathbf{D}, \phi) &= \int_{\mathbb{R}^{Gq}} f(\mathbf{y}, \mathbf{u}|\boldsymbol{\beta}, \mathbf{D}, \phi) d\mathbf{u} \\ &= \int_{\mathbb{R}^{Gq}} f(\mathbf{y}|\boldsymbol{\beta}, \mathbf{u}, \phi) f(\mathbf{u}|\mathbf{D}) d\mathbf{u} \end{aligned} \tag{1.11}$$

to obtain the maximum likelihood estimates of $\boldsymbol{\beta}$, $\mathbf{D}$, and ϕ. The integrated likelihood is sometimes known as the marginal likelihood but we refrain from using this as we have already used the term marginal likelihood for (1.6). The model that results in integrating out the group-specific parameters is known as the *marginal model*. The integrand in (1.11), $f(\mathbf{y}, \mathbf{u}|\boldsymbol{\beta}, \mathbf{D}, \phi)$, is termed the *h-likelihood function* by Lee et al. (2006).

1.2.3 Bayesian Inference for GLMMs

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \mathbf{u}^T, \mathbf{d}^T, \phi)^T$ be the $k \times 1$ vector of model parameters, where

$$k = \begin{cases} p + Gq + \frac{1}{2}q(q+1) + 1 & \text{if } \phi \text{ is unknown} \\ p + Gq + \frac{1}{2}q(q+1) & \text{otherwise,} \end{cases}$$

to align with the notation introduced in Section 1.1. To complete the specification of a Bayesian GLMM we require a joint prior distribution for $\boldsymbol{\theta}$, with pdf decomposed as

$$\begin{aligned} f(\boldsymbol{\theta}) &= f(\boldsymbol{\beta}, \mathbf{u}, \mathbf{d}, \phi), \\ &= f(\boldsymbol{\beta}|\mathbf{d}, \phi) f(\mathbf{u}|\mathbf{d}) f(\mathbf{d}|\phi) f(\phi), \\ &= f(\boldsymbol{\beta}|\mathbf{D}, \phi) f(\mathbf{u}|\mathbf{D}) f(\mathbf{D}|\phi) f(\phi). \end{aligned}$$

The conditional distribution of the group-specific parameters, $\mathbf{u}|\mathbf{D}$, has already been specified, as part of the specification of a GLMM, as $\mathrm{N}(\mathbf{0}, \mathbf{I}_G \otimes \mathbf{D})$. Therefore, it only remains to specify a prior distribution for the remaining parameters, $\boldsymbol{\beta}, \mathbf{D}$ and ϕ, with their joint pdf decomposed as

$$f(\boldsymbol{\beta}, \mathbf{D}, \phi) = f(\boldsymbol{\beta}|\mathbf{D}, \phi) f(\mathbf{D}|\phi) f(\phi). \tag{1.12}$$

The pdf of the posterior distribution of $\boldsymbol{\beta}$, $\mathbf{u}$, $\mathbf{D}$ and ϕ is given by

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

The pdf of the marginal posterior distribution of $\boldsymbol{\beta}$, $\mathbf{D}$ and ϕ is given by

$$\begin{aligned} f(\boldsymbol{\beta}, \mathbf{D}, \phi|\mathbf{y}) &= \int_{\mathbb{R}^{Gq}} f(\boldsymbol{\beta}, \mathbf{u}, \mathbf{D}, \phi|\mathbf{y}) \mathrm{d}\mathbf{u}, \\ &\propto f(\mathbf{y}|\boldsymbol{\beta}, \mathbf{D}, \phi) f(\boldsymbol{\beta}|\mathbf{D}, \phi) f(\mathbf{D}|\phi) f(\phi). \end{aligned}$$

The model determination terminology introduced in Section 1.1.3 can now be applied to a GLMM.

1.3 The Problem

In Section 1.1.3, we described the basic quantities required for model determination, which we would like to apply to Bayesian GLMM determination. The integral in the denominator of (1.3), i.e. the marginal likelihood, is rarely analytically tractable, thus necessitating computational methods to approximate the integral. However, the group-specific parameter, $\mathbf{u}$, often has large dimensionality if the number of groups, G, is large, thus the choice of method for approximating the marginal likelihood or posterior model probability is critical. We address this issue in Chapters 4 and 5.

Once a satisfactory method for computing an approximation to the marginal likelihood or posterior model probability has been chosen, we then need to consider a default prior distribution for $\boldsymbol{\beta}$, $\mathbf{D}$, and ϕ for objective Bayesian model determination. We address this issue in Chapter 3.

Essentially, the problem focused on in this thesis is to develop a default Bayesian model determination strategy which addresses the issues of computation and default prior specification under weak prior information.

In the next Section, we introduce an example of a dataset that exemplifies the problems we can encounter when we try to apply objective Bayesian model determination to GLMMs.

1.4 Turtle Data Example

We introduce an example of a dataset that a set of GLMMs can be applied to. This shows the problem of model uncertainty amongst GLMMs and will be used as a running example to illustrate the methodology introduced in Chapters 3, 4 and 5 .

The dataset termed the *Turtle Data* is analysed by Sinharay and Stern (2000) and Sinharay and Stern (2005). It contains the survival status (0=died, 1=survived), birthweight (in grams), and clutch (i.e. family) membership of 244 newborn turtles from 31 different clutches. The researchers wish to determine whether there is a birthweight effect on the survival chances of a newborn turtle, having accounted for the fact that the survival probability of a turtle may be correlated with the survival probability of another turtle within the same clutch. Suppose y_{ij} and z_{ij} are the survival status and the birthweight, respectively, of the jth turtle in the ith clutch for $i = 1, ..., 31$ and $j = 1, ..., n_i$. The clutch sizes, $n_1, ..., n_{31}$, have minimum, maximum and mean of 1, 18 and 7.9, respectively.

We assume that $y_{ij} \sim \text{Bernoulli}(\mu_{ij})$ where $\eta_{ij} = g(\mu_{ij})$. We follow Sinharay and Stern (2005) and use the probit link function, i.e. $\mu_{ij} = \Phi(\eta_{ij})$ so that $g(\mu_{ij}) = \Phi^{-1}(\mu_{ij})$. Let $x_{ij} = \frac{z_{ij} - \bar{z}_{..}}{s}$ be the standardised z_{ij}, where s^2 is the sample variance of the z_{ij}s. We consider a total of five possible models:

1. $\eta_{ij} = \beta_1$,
2. $\eta_{ij} = \beta_1 + \beta_2 x_{ij}$,
3. $\eta_{ij} = \beta_1 + u_i$ where $u_i \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2)$,
4. $\eta_{ij} = \beta_1 + \beta_2 x_{ij} + u_i$ where $u_i \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2)$,
5. $\eta_{ij} = (\beta_1 + u_{1i}) + (\beta_2 + u_{2i}) x_{ij}$ where $\mathbf{u}_i = (u_{1i}, u_{2i})^T \overset{\text{iid}}{\sim} \text{N}(\mathbf{0}, \mathbf{D})$.

Models 1 and 2 are GLMs where the survival probability of a newborn turtle is independent of the survival probability of turtles within the same clutch. Model 1 assumes that survival is independent of birthweight, i.e. there is no birthweight effect, whereas Model 2 assumes that there is a birthweight effect. Models 3, 4 and 5 are GLMMs where the survival of a newborn turtle is assumed to be correlated with the survival of turtles within the same clutch, i.e. there is a clutch effect. Model 3 assumes that there is no birthweight effect on survival.

Table 1.1: Dimensionality of the model parameters for the five models applied to the Turtle Data.

Model	Dimensionality of the model parameters
1	1
2	2
3	33
4	34
5	67

Models 4 and 5 assume that there is a birthweight effect on survival and for Model 4, the clutch effect is the same for each turtle in the same clutch, whereas for Model 5, the clutch effect depends upon the turtle's birthweight.

The marginal likelihood is not analytically tractable for any of the five models and will need to be approximated. Table 1.1 gives the dimensionality of the model parameters for each of the models. We see that including the group-specific parameters significantly increases the dimensionality of the resulting integral approximation.

Sinharay and Stern (2000) and Sinharay and Stern (2005) considered model determination, with respect to computation only, between Models 2 and 4, i.e. determining whether or not there exists a group-specific intercept.

In both papers, the integrated likelihood function is evaluated by using Simpson's rule. We see in Chapter 2 that this quadrature approach becomes impractical for $q > 1$. They, therefore, do not consider model determination including Model 5 since $q = 2$, in this case. In this thesis, we develop a model determination approach that can be used when $q > 1$.

In both papers, a diffuse prior distribution was applied to the regression parameters, $\boldsymbol{\beta}$, in Models 2 and 4. In Sinharay and Stern (2000), an inverse-gamma prior distribution, $\text{IG}(\frac{5}{2}, \frac{3}{2})$, is applied to σ^2 in Model 4 and a Bayes factor of 3.25 is found in favour of Model 2. In Sinharay and Stern (2005) a prior distribution is applied to σ^2 in Model 4 with pdf

$$f(\sigma^2) = \frac{1}{(1+\sigma^2)^2},$$

and a Bayes factor of 1.273 is found in favour of Model 2.

Kass and Raftery (1995) give guidelines on how to interpret Bayes factors. In their interpretation, a Bayes factor of 1.273 in favour of Model 2 is "not worth more than a bare mention", whereas a Bayes factor of 3.25 represents "substantial" evidence in favour of Model 2. We see from this example the danger of applying arbitrary prior distributions to the parameters of competing models. Sinharay and Stern (2000) and Sinharay and Stern (2005) were both concerned with computation only, so were not attempting to define default priors for model determination.

Chapter 2

Previous Work

2.1 Introduction

This chapter, on previous work on Bayesian model determination strategies for GLMMs, is split into two parts: computation for GLMMs and default priors for GLMMs. The first part will start by discussing the importance of computation in Bayesian inference, we then describe some general computational methods and describe how these methods have been applied to GLMMs in the literature. The second part on default priors will discuss some default priors that have been applied to GLMMs and special cases of GLMMs, e.g. linear models.

2.2 Computation for GLMMs

2.2.1 Introduction

In Section 1.1.3, we defined the two aspects of posterior inference under model uncertainty as

1. Evaluation of the posterior model probability, $f(m|\mathbf{y})$, for each $m \in M$ (Model Determination),
2. Evaluation and summarisation of the posterior distribution, with pdf $f_m(\boldsymbol{\theta}_m|\mathbf{y})$ of the parameters, $\boldsymbol{\theta}_m$ of each model $m \in M$ (Posterior Inference).

To achieve 1., we either need to evaluate the marginal likelihood, $f_m(\mathbf{y})$, of model $m \in M$ to compute the posterior model probabilities, evaluate the ratio of marginal likelihoods (Bayes' factors), $f_k(\mathbf{y})/f_m(\mathbf{y})$ of models $k, m \in M$, or evaluate the posterior model probabilities

directly. Typically, except for certain special cases, these require the evaluation of intractable, possibly high dimensional, integrals by approximation. To achieve 2., we typically want to evaluate quantities such as the posterior mean, posterior variance or posterior quantiles. Again these require the evaluation of intractable integrals by approximation. Also in the pursuit of 2., we may want to evaluate the posterior mode, requiring the maximisation of $f_m(\boldsymbol{\theta}_m|\mathbf{y})$ which, in general, will not be analytically tractable.

Example

Suppose we have a binary response, $y_i \sim \text{Bernoulli}(p_i)$ where

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta,$$

for $i = 1, ..., n$ and $\beta \in \mathbb{R}$, giving a very simple logistic regression model. The likelihood is then

$$f(\beta|\mathbf{y}) = \prod_{i=1}^{n} \frac{\exp(\beta y_i)}{1+\exp(\beta)}.$$

Suppose as a prior distribution for β we choose the normal distribution with mean μ and variance σ^2, i.e. $\beta \sim \text{N}(\mu, \sigma^2)$. Therefore the posterior distribution has pdf given by Bayes' Theorem

$$f(\beta|\mathbf{y}) = \frac{\exp\left(-\frac{(\beta-\mu)^2}{2\sigma^2}\right)\prod_{i=1}^{n}\frac{\exp(\beta y_i)}{1+\exp(\beta)}}{\int_{-\infty}^{\infty}\exp\left(-\frac{(\beta-\mu)^2}{2\sigma^2}\right)\prod_{i=1}^{n}\frac{\exp(\beta y_i)}{1+\exp(\beta)}\text{d}\beta} \tag{2.1}$$

The denominator of (2.1) is an intractable integral, and maximisation of the posterior pdf $f(\beta|\mathbf{y}) \propto \exp\left(-\frac{(\beta-\mu)^2}{2\sigma^2}\right)\prod_{i=1}^{n}\frac{\exp(\beta y_i)}{1+\exp(\beta)}$ is also an analytically intractable problem. ■

There exist two different approaches for approximating integrals: *deterministic methods* (also known as *numerical methods*) and *stochastic methods* (known as *Monte Carlo methods*). In this thesis, we make use of both types of approach. In Section 2.2.2, we describe some of the deterministic methods we implement in this thesis.

Monte Carlo methods refer to a broad selection of methods for approximating intractable integrals, by generating samples from the required distribution (usually the posterior distribution), and then forming sample averages to approximate the integral. For example, suppose we wish to evaluate the expectation, μ, of the random variable, $\theta \sim \Pi$, where the distribution, Π, has pdf $\pi(\theta)$, i.e.

$$\mu = \text{E}(\theta) = \int_{\Theta} \theta\pi(\theta)\text{d}\theta. \tag{2.2}$$

Suppose that the integral in (2.2) is intractable. A Monte Carlo approximation, $\hat{\mu}$, to μ is to generate a sample, $\theta_1, ..., \theta_n$, of size n from the distribution Π and then set

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}\theta_i.$$

More generally, a Monte Carlo approximation, $\hat{\mu}_g$, to $\mu_g = \mathrm{E}(g(\boldsymbol{\theta}))$ for some function $g : \Theta \to \mathbb{R}$, is to generate $\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_n$ from the distribution, Π, and then set

$$\hat{\mu}_g = \frac{1}{n} \sum_{i=1}^{n} g(\boldsymbol{\theta}_i).$$

We can also approximate quantities such as the median or quantiles by their corresponding sample quantities.

Obviously, to use these methods we need to be able to generate a sample from the required distribution (usually the posterior distribution). Again, there exist Monte Carlo methods for doing so and we describe some of these in Sections 2.2.3 and 2.2.4.

As discussed we also need to consider methods for maximising a function. As with approximating integrals, there are deterministic and stochastic methods. In this thesis, we only consider deterministic maximisation methods and these are briefly reviewed in Section 2.2.2. For stochastic methods for maximising a function see, for example, Robert and Casella (1999, Ch. 5).

2.2.2 Deterministic Methods

In this Section, we describe some deterministic methods for evaluating integrals and maximising functions that we implement in this thesis. We also highlight some of the limitations of deterministic methods, and the cases where Monte Carlo methods are preferred.

Approximations to moments of functions of random variables

In 2.2.1 we described the Monte Carlo approximation to the expectation, μ_g, of the function, $g : \Theta \to \mathbb{R}$, of a random variable, $\boldsymbol{\theta}$, with pdf $\pi(\boldsymbol{\theta})$, i.e. $\mu_g = \mathrm{E}\left(g(\boldsymbol{\theta})\right)$. Suppose the mean and variance matrix of $\boldsymbol{\theta}$ is $\mathbf{m}$ and $\boldsymbol{\Sigma}$, respectively. We can use a first-order Taylor series expansion of $g()$ about $\mathbf{m}$ to derive deterministic approximations to $\mathrm{E}(g(\boldsymbol{\theta}))$ and $\mathrm{var}(g(\boldsymbol{\theta}))$. Note that

$$g(\boldsymbol{\theta}) \approx g(\mathbf{m}) + (\boldsymbol{\theta} - \mathbf{m})^T \left. \frac{\mathrm{d}g(\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}} \right|_{\boldsymbol{\theta}=\mathbf{m}}.$$

Therefore the expectation of $g(\boldsymbol{\theta})$ is approximated by

$$\mathrm{E}(g(\boldsymbol{\theta})) \approx g(\mathbf{m}),$$

and the variance of $g(\boldsymbol{\theta})$ by

$$\mathrm{var}(g(\boldsymbol{\theta})) \approx \left. \frac{\mathrm{d}g(\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}^T} \right|_{\boldsymbol{\theta}=\mathbf{m}} \boldsymbol{\Sigma} \left. \frac{\mathrm{d}g(\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}} \right|_{\boldsymbol{\theta}=\mathbf{m}}.$$

Obviously, we can use a higher-order Taylor series expansion to achieve higher accuracy.

Laplace method for approximating an integral

Suppose we wish to evaluate the intractable integral

$$\int_{\Theta} g(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}, \tag{2.3}$$

where the dimension of $\boldsymbol{\theta}$ is k.

We take the second-order Taylor series expansion of $\log g(\boldsymbol{\theta})$ about the value, $\mathbf{m}$, which maximises $g(\boldsymbol{\theta})$, i.e.

$$\log g(\boldsymbol{\theta}) \approx \log g(\mathbf{m}) + \frac{1}{2}(\boldsymbol{\theta} - \mathbf{m})^T \left. \frac{\partial^2 \log g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right|_{\boldsymbol{\theta}=\mathbf{m}} (\boldsymbol{\theta} - \mathbf{m}).$$

Note that the first-order term disappears since $\left. \frac{\partial \log g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta}=\mathbf{m}} = \mathbf{0}$. Therefore,

$$g(\boldsymbol{\theta}) \approx g(\mathbf{m}) \exp\left(-\frac{1}{2}(\boldsymbol{\theta} - \mathbf{m})^T \mathbf{V}^{-1} (\boldsymbol{\theta} - \mathbf{m}) \right),$$

where $\mathbf{V} = - \left. \frac{\partial^2 \log g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right|^{-1}_{\boldsymbol{\theta}=\mathbf{m}}$, and

$$\int_{\Theta} g(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \approx g(\mathbf{m}) (2\pi)^{\frac{k}{2}} |\mathbf{V}|^{\frac{1}{2}}. \tag{2.4}$$

Note that to implement (2.4), we need to have maximised $g(\boldsymbol{\theta})$ to find $\mathbf{m}$. This method of approximating an integral is known as the *Laplace method*.

Typically, (2.3) is in the form of a marginal likelihood with $g(\boldsymbol{\theta}_m) = f_m(\mathbf{y}|\boldsymbol{\theta}_m) f_m(\boldsymbol{\theta}_m)$ and $\mathbf{m}$ representing the posterior mode. In this case, Tierney and Kadane (1986) state that this method "will produce reasonable results as long as the posterior is unimodal or at least dominated by a single mode", and that this method has an error of order $O(n^{-1})$ where n is the sample size.

We present an alternative method for approximating the marginal likelihood where we require the maximum likelihood estimate as opposed to the posterior mode. Suppose now we specifically wish to evaluate the marginal likelihood, $f_m(\mathbf{y}) = \int_{\Theta_m} f_m(\mathbf{y}|\boldsymbol{\theta}_m) f_m(\boldsymbol{\theta}_m) d\boldsymbol{\theta}_m$, of model $m \in M$. The second-order Taylor series expansion of the log-likelihood, $\log f_m(\mathbf{y}|\boldsymbol{\theta}_m)$, about the maximum likelihood estimate, $\hat{\boldsymbol{\theta}}_m$, is

$$\log f_m(\mathbf{y}|\boldsymbol{\theta}_m) \approx \log f_m(\mathbf{y}|\hat{\boldsymbol{\theta}}_m) - \frac{1}{2}(\boldsymbol{\theta}_m - \hat{\boldsymbol{\theta}}_m)^T \mathbf{V}_m^{-1} (\boldsymbol{\theta}_m - \hat{\boldsymbol{\theta}}_m),$$

where $\mathbf{V}_m = - \left. \frac{\partial^2 \log f_m(\mathbf{y}|\boldsymbol{\theta}_m)}{\partial \boldsymbol{\theta}_m \boldsymbol{\theta}_m^T} \right|^{-1}_{\boldsymbol{\theta}_m=\hat{\boldsymbol{\theta}}_m}$. We assume that $|\boldsymbol{\theta}_m - \hat{\boldsymbol{\theta}}_m|$ is small, then $f_m(\boldsymbol{\theta}_m)$ varies slowly and can be approximated by the constant $f_m(\hat{\boldsymbol{\theta}}_m)$. Therefore

$$f_m(\mathbf{y}) \approx f_m(\hat{\boldsymbol{\theta}}_m) f_m(\mathbf{y}|\hat{\boldsymbol{\theta}}_m) (2\pi)^{\frac{k_m}{2}} |\mathbf{V}_m|^{\frac{1}{2}} \tag{2.5}$$

O'Hagan and Forster (2004, pg. 180) state that this method also has an error of order $O(n^{-1})$, where n is the sample size.

The reason that both of these methods perform more accurately as the sample size increases, is that the posterior distribution approaches normality as $n \to \infty$.

We use a form of the Laplace approximation in Chapter 5 to approximate the integrated likelihood (1.11) of a GLMM where we replace the first-stage likelihood by a quadratic approximation.

Quadrature

Quadrature is another name for numerical integration. We first consider the one-dimensional integral

$$I = \int_a^b g(\theta) \mathrm{d}\theta.$$

Quadrature methods work by calculating $g()$ at n points $\theta_1, ..., \theta_n$ and then using the resulting $g(\theta_1), ..., g(\theta_n)$ in some formula such as a weighted average

$$\hat{I} = \sum_{i=1}^{n} w_i g(\theta_i). \tag{2.6}$$

A simple quadrature method is *Simpson's rule*. Here the interval $[a, b]$ is divided into n equal sub-intervals, $g()$ is then evaluated at the mid-point of each sub-interval, and then equal weights, $w_i = \frac{b-a}{n}$, are used, giving

$$\hat{I}_s = \frac{b-a}{n} \sum_{i=1}^{n} g(a + (2i-1)(b-a)/(2n)), \tag{2.7}$$

as the Simpson's rule approximation to I. A potential problem with Simpson's rule is that, typically, the limits of integration may be $a = -\infty$ and/or $b = \infty$. In practice, we can just set very wide limits for a and b in (2.7) and assume that $g()$ is negligible outside $[a, b]$.

A quadrature method which takes advantage of approximate normality of $g(\theta)$ is the *Gauss-Hermite rule* which requires $a = -\infty$ and $b = \infty$. The n-point Gauss-Hermite rule uses points $\theta_1, ..., \theta_n$ and weights $w_1, ..., w_n$, such that $\hat{I} = I$ if $\exp\left(\frac{\theta}{2}\right) g(\theta)$ is a polynomial of order $2n-1$. Since we can approximate $\exp\left(\frac{\theta}{2}\right) g(\theta)$ arbitrarily accurately with a polynomial of order $2n-1$, the accuracy of the Gauss-Hermite rule increases as n increases. Abramowitz and Stegun (1965, pg. 924) contains tables of points and weights for different values of n.

Gauss-Hermite quadrature can be extended to approximate the p-dimensional integral

$$I = \int_{\mathbb{R}^p} g(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta},$$

as

$$\hat{I} = \sum_{j=1}^{N} w_j g(\boldsymbol{\theta}_j), \tag{2.8}$$

where $N = \prod_{i=1}^{p} n_i$, and w_j is the weight associated with the point $\boldsymbol{\theta}_j$. The rule (2.8) is formed by applying the one-dimensional n_i-point Gauss-Hermite rule in the ith dimension. The weight $w_j = w_{1j_1} w_{2j_2} ... w_{pj_p}$ is the product of the p one-dimensional weights.

Simpson's rule can also be extended to p dimensions by applying the one-dimensional Simpson's rule in each dimension.

The major drawback of quadrature rules is that they require $\prod_{i=1}^{p} n_i$ function evaluations of $g()$, which becomes infeasibly large as p increases. For the high-dimensional integral approximations that occur in GLMMs, quadrature methods are not a feasible set of methods. However, for low dimensional integral approximations, quadrature methods can be very useful (see, for example, Skrondal and Rabe-Hesketh (2004)).

Optimisation

In this Section, we briefly describe methods for maximising a function $g : \mathbb{R}^k \to \mathbb{R}$, i.e. finding the value $\mathbf{m} \in \mathbb{R}^k$ such that $g(\mathbf{m}) \geq g(\boldsymbol{\theta})$ for all $\boldsymbol{\theta} \in \mathbb{R}^k$. It is generally easier to maximise the logarithm of a function rather than the function itself, i.e. maximise $h(\boldsymbol{\theta}) = \log g(\boldsymbol{\theta})$.

The best known technique for maximising a function is the *Newton-Raphson* method, or simply the *Newton* method. It requires the vector of first derivatives, $\mathbf{h}(\boldsymbol{\theta}^*) = \left.\frac{\mathrm{d}h(\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}$ and the Hessian matrix $\mathbf{H}(\boldsymbol{\theta}^*) = \left.\frac{\mathrm{d}^2h(\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}\mathrm{d}\boldsymbol{\theta}^T}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}$. The algorithm for finding $\mathbf{m}$ is as follows

1. Start at the initial value $\boldsymbol{\theta}^0$.
2. Suppose the current value is $\boldsymbol{\theta}^i$, and set
$$\boldsymbol{\theta}^{i+1} = \boldsymbol{\theta}^i - \mathbf{H}(\boldsymbol{\theta}^i)^{-1}\mathbf{h}(\boldsymbol{\theta}^i).$$
3. Repeat 2. to 3., unless the sequence $\boldsymbol{\theta}^0, ..., \boldsymbol{\theta}^{i+1}$, has converged, in which case, $\mathbf{m} = \boldsymbol{\theta}^{i+1}$.

This method is derived by using the 2nd order Taylor series expansion of $h(\boldsymbol{\theta})$ about $\mathbf{m}$. The sequence will converge rapidly to $\mathbf{m}$ provided $\boldsymbol{\theta}^0$ is sufficiently close to $\mathbf{m}$. In fact, if $\boldsymbol{\theta}^0$ is within the inflexion boundary about $\mathbf{m}$, then convergence is guaranteed.

In this thesis, we will need to maximise a function when both the vector of first derivatives and the Hessian matrix are available, and to do so we can use the Newton method as described above. However, we also need to maximise a function where neither the vector of first derivatives nor the Hessian matrix are available. In this case, we can use *Quasi-Newton* methods where $\mathbf{h}(\boldsymbol{\theta})$ and $\mathbf{H}(\boldsymbol{\theta})$ are approximated at each iteration. From the output of a quasi-Newton algorithm, we can obtain an approximation to the Hessian matrix evaluated at $\mathbf{m}$. For more details on quasi-Newton methods, see, for example, Fletcher (2000). The Newton method and Quasi-Newton methods are readily implemented in many mathematical and statistical software packages. For instance, the functions *optim* and *nlminb*, available

in the statistical software package R (R Development Core Team (2009)), were used in this thesis.

2.2.3 Random Number Generation

In Section 2.2.1, we briefly described how Monte Carlo methods can be used to approximate integrals. We will discuss this further in Section 2.2.5, but we noted the assumption that we can generate from the required distribution. In this and the next Section, we describe methods for generating random samples (or approximately random samples) from probability distributions.

Many methods exist in the literature for generating samples from standard probability distributions. These include inversion and the ratio-of-uniforms methods, and are covered in, for example, Gentle (1998). In this Section, we describe general-purpose methods for generating samples from arbitrary univariate probability distributions with log-concave pdfs. Being able to generate from univariate distributions will become useful when we consider Gibbs sampling in Section 2.2.4.

Rejection sampling

Suppose we wish to generate a sample from the distribution, Π, with pdf $\pi(\theta) \propto g(\theta)$. Suppose we can easily generate from the *sampling distribution*, S, of the random variable, Z, which has pdf $s(z)$. Let $g(z)/s(z) \leq A < \infty$ and $U \sim \mathrm{U}[0,1]$. We may not be able to derive $\sup(g(z)/s(z))$, but we can derive an upper bound. Now

$$X = Z \left| U \leq \frac{g(Z)}{As(Z)} \right. \sim \Pi.$$

The joint pdf of Z and U is $s(z,u) = s(z)$, so $Z|U \leq g(z)/As(z)$ has pdf

$$\begin{aligned} f\left(z \left| u \leq \frac{g(z)}{As(z)} \right.\right) &= \frac{\int_0^{\frac{g(z)}{As(z)}} s(u,z)\mathrm{d}u}{\int_{\mathbb{R}} \int_0^{\frac{g(z)}{As(z)}} s(u,z)\mathrm{d}u\mathrm{d}z}, \\ &= \frac{g(\theta)}{\int_\Theta g(\theta)\mathrm{d}\theta}, \\ &= \pi(\theta). \end{aligned}$$

So an algorithm to generate θ from Π is as follows:

1. Generate z from the sampling distribution, S, and u from $\mathrm{U}[0,1]$.
2. If $u \leq \frac{g(z)}{As(z)}$, then set $\theta = z$, otherwise repeat 1. and 2.

This method is known as *rejection sampling*.

The probability of accepting any pair (z, u) in rejection sampling is

$$P\left(U \leq \frac{g(Z)}{As(Z)}\right) = \frac{\int_{\mathbb{R}} g(x)\mathrm{d}x}{A}.$$

Therefore, the probability of acceptance is maximised by setting $A = \sup_{z\in\mathbb{R}}\left(\frac{g(z)}{s(z)}\right)$. In turn, the probability of acceptance can be increased by choosing an $s(z)$ that 'mimics' $g(z)$ as closely as possible, thus reducing $\sup_{z\in\mathbb{R}}\left(\frac{g(z)}{s(z)}\right)$. Rejection sampling can actually be applied for sampling in k dimensions. However, as O'Hagan and Forster (2004, pg. 277) state "it is difficult to obtain a sampling distribution $s(\mathbf{z})$ for which the acceptance rate $\frac{\int_{\mathbb{R}^k} g(\mathbf{x})\mathrm{d}\mathbf{x}}{A}$ is not small". Therefore, in practice rejection sampling is usually reserved for univariate distributions.

Adaptive Rejection Sampling

Adaptive rejection sampling (ARS) is a method proposed by Gilks and Wild (1992) to generate from the univariate distribution, Π, with pdf $\pi(\theta) = g(\theta)/\int_\Theta g(\theta)\mathrm{d}\theta$, where $g(\theta)$ is log-concave, i.e. if $h(\theta) = \log g(\theta)$ then $\frac{\mathrm{d}^2 h(\theta)}{\mathrm{d}\theta^2} < 0$ for $\theta \in \Theta$. Also assume that Θ is connected, and that $g(\theta)$ is continuous and differentiable on Θ.

ARS improves rejection sampling by, after each value is generated from the sampling distribution, adapting S so that it 'mimics' Π more closely. An optional squeezing function can also be defined to give a quick rejection test. This feature attempts to minimise the number of evaluations of $g(\theta)$ which is assumed to be a computationally expensive process.

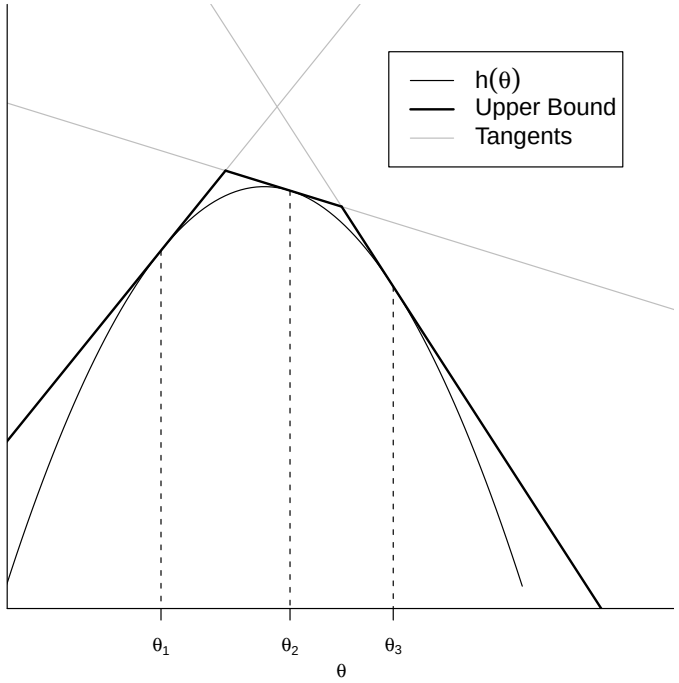
At every rejection, the first derivative of $h(\theta)$ is computed at the point generated from S and a tangent to $h(\theta)$ is found. In between the intersections of the tangents, an upper bound to $h(\theta)$ is formed by the tangents. Therefore, S is formed from piecewise exponential distributions by exponentiating the upper bound. To initialise the algorithm, we need a set of at least two points along with the evaluations of $h(\theta)$ and $\frac{\mathrm{d}h(\theta)}{\mathrm{d}\theta}$ at those points. Figure 2.1 shows the situation for three points, θ_1, θ_2 and θ_3. The underlying distribution, Π, in Figure 2.1 is a normal distribution.

For more details, including how to construct the optional squeezing function and the specific algorithm, see Gilks and Wild (1992). For a similar method that does not require derivatives of $h(\theta)$ see Gilks (1992).

2.2.4 Markov Chain Monte Carlo

Markov chain Monte Carlo (MCMC) methods are a collection of general-purpose methods for generating a sample from a complicated probability distribution which is known, generally,

Figure 2.1: Upper hull for $h(\theta) = \log g(\theta)$ for a log-concave pdf, $\pi(\theta) = g(\theta)/\int_\Theta g(\theta)\mathrm{d}\theta$ for $r = 3$.



as the *target distribution*, which we denote Π, with pdf $\pi(\boldsymbol{\theta})$ where $\boldsymbol{\theta} \in \Theta$. In this Section, we use notation and terminology which is used in the field of MCMC.

A discrete-time homogeneous *Markov chain*, $\boldsymbol{\theta}^0, \boldsymbol{\theta}^1, ..., \boldsymbol{\theta}^n$ is a dependent sequence of values which are defined by the *transition kernel*:

$$P(\mathbf{x}, A) = P(\boldsymbol{\theta}^{i+1} \in A | \boldsymbol{\theta}^i = \mathbf{x}).$$

The transition kernel represents the probability of the next value, $\boldsymbol{\theta}^{i+1}$, belonging to the set $A \subset \Theta$, when the previous value, $\boldsymbol{\theta}^i$, was $\mathbf{x}$. Similarly, the *n-step transition kernel* is defined as

$$P^n(\mathbf{x}, A) = P(\boldsymbol{\theta}^{i+n} \in A | \boldsymbol{\theta}^i = \mathbf{x}),$$

i.e. the probability of the nth next value, $\boldsymbol{\theta}^{i+n}$, belonging to the set A, when the ith value is $\mathbf{x}$. The initial value of the chain is $\boldsymbol{\theta}^0$ and the distribution of the nth value, $\boldsymbol{\theta}^n$, is given by $P^n(\boldsymbol{\theta}^0, A)$.

We want to generate from the target distribution, Π, so we require, for some n,

$$P^n(\boldsymbol{\theta}^0, A) \approx P_\pi(\boldsymbol{\theta} \in A) = \int_A \pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}, \tag{2.9}$$

for all $A \subset \Theta$ and for any initial value, $\boldsymbol{\theta}^0$. So regardless of the initial value, the distribution of $\boldsymbol{\theta}^n$ is approximately equal to the target distribution. The first step in satisfying this condition is to ensure that the *stationary distribution* of the Markov chain is the target distribution, i.e. if $\boldsymbol{\theta}^i \sim \Pi$, then $\boldsymbol{\theta}^{i+1} \sim \Pi$, or in terms of the transition kernel

$$P_\pi(\boldsymbol{\theta} \in A) = \int_A \pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} = \int_\Theta P(\boldsymbol{\theta}, A)\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}, \tag{2.10}$$

for all $A \subset \Theta$. The condition (2.10) is difficult to verify for any particular transition kernel. However, we can choose a transition kernel that satisfies the condition of *reversibility* for Π. Reversibility holds when $(\boldsymbol{\theta}^i, \boldsymbol{\theta}^{i+1})$ has the same distribution as the time-reversed $(\boldsymbol{\theta}^{i+1}, \boldsymbol{\theta}^i)$, i.e. if the *detailed balance equations*:

$$\begin{aligned} P(\boldsymbol{\theta}^{i+1} \in A, \boldsymbol{\theta}^i \in B) &= \int_B P(\boldsymbol{\theta}, A)\pi(\boldsymbol{\theta})d\boldsymbol{\theta} \\ = P(\boldsymbol{\theta}^{i+1} \in B, \boldsymbol{\theta}^i \in A) &= \int_A P(\boldsymbol{\theta}, B)\pi(\boldsymbol{\theta})d\boldsymbol{\theta}, \end{aligned}$$

hold for $A, B \subset \Theta$. If $B = \Theta$, then the integral in the first line becomes (2.10) and, therefore, reversibility implies that Π is the stationary distribution.

A Markov chain is *ergodic* if

$$\sup_{A \subset \Theta} |P^n(\boldsymbol{\theta}^0, A) - P_\pi(\boldsymbol{\theta} \in A)| \to 0, \tag{2.11}$$

as $n \to \infty$. In other words, starting from an arbitrary initial value, $\boldsymbol{\theta}^0$, the nth value in the chain, $\boldsymbol{\theta}^n$, tends to be a value from the target distribution, Π, as n tends to infinity. This is a very important property for an MCMC sampler, since if we have n iterations (for sufficiently large n) from an ergodic Markov chain with Π as its stationary distribution then the approximation (2.9) can be justified by (2.11) (see, for example, O'Hagan and Forster (2004, pg. 263-264)).

A Markov chain is ergodic if it is irreducible, aperiodic and Harris recurrent.

A Markov chain is *irreducible* if, for any $\boldsymbol{\theta}^0$ and any $A \subset \Theta$ such that $P_\pi(\boldsymbol{\theta} \in A) > 0$, there exists n such that $P^n(\boldsymbol{\theta}^0, A) > 0$. So, regardless of the initial value, $\boldsymbol{\theta}^0$, the chain can eventually visit any region, A, of the parameter space, Θ. Suppose $C^m(\boldsymbol{\theta}^0) \subset \Theta$ is the set of all possible values of $\boldsymbol{\theta}^m$ that can be visited from an initial $\boldsymbol{\theta}^0$. A Markov chain can be reducible if for two different initial values, $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}^{**}$, $C^m(\boldsymbol{\theta}^*) \cap C^m(\boldsymbol{\theta}^{**}) = \emptyset$.

A Markov chain is *periodic* with period d if it cycles between d disjoint subsets of Θ. For example, suppose there are two disjoint subsets C_1 and C_2, where $\boldsymbol{\theta}^i \in C_1$ implies $\boldsymbol{\theta}^{i+1} \in C_2$, and $\boldsymbol{\theta}^i \in C_2$ implies $\boldsymbol{\theta}^{i+1} \in C_1$, then the chain is periodic with period 2. A Markov chain is *aperiodic* if it is not periodic.

Let η_A be the number of visits of a Markov chain to the subset $A \subset \Theta$. A Markov chain is *recurrent* if for any $A \subset \Theta$ such that $P_\pi(\boldsymbol{\theta} \in A) > 0$, then $E(\eta_A) = \infty$. A Markov chain is *Harris recurrent* if for any $A \subset \Theta$ such that $P_\pi(\boldsymbol{\theta} \in A) > 0$, then $P(\eta_A = \infty) = 1$. In other words, a Harris recurrent Markov chain visits the subset A infinitely often, regardless of the initial value.

The *ergodic theorem* for Markov chains states that, for an ergodic Markov chain, $\{\boldsymbol{\theta}^1, ..., \boldsymbol{\theta}^n\}$,

$$\frac{1}{n}\sum_{i=1}^{n} h(\boldsymbol{\theta}^i) \to \mathrm{E}_\Pi(h(\boldsymbol{\theta})),$$

as $n \to \infty$, with probability one. In other words, if we have an ergodic Markov chain, $\{\boldsymbol{\theta}^1, ..., \boldsymbol{\theta}^n\}$, with stationary distribution, Π, we can approximate $\mathrm{E}_\Pi(h(\boldsymbol{\theta})) = \int_\Theta h(\boldsymbol{\theta})\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}$ by $\frac{1}{n}\sum_{i=1}^n h(\boldsymbol{\theta}^i)$ with large n.

Suppose we have a Markov chain, $\{\boldsymbol{\theta}^0, \boldsymbol{\theta}^1, ..., \boldsymbol{\theta}^t, ..., \boldsymbol{\theta}^{t+n}\}$, with initial value, $\boldsymbol{\theta}^0$. We retain, for summary of Π, the values $\{\boldsymbol{\theta}^{t+1}, ..., \boldsymbol{\theta}^{t+n}\}$, i.e. we discard the first t values. The values $\{\boldsymbol{\theta}^0, ..., \boldsymbol{\theta}^t\}$ are called the *burn-in phase*. We are required to specify a value for t, for which, effectively, the chain has become independent of the initial value, $\boldsymbol{\theta}^0$. We can choose $t = 0$, if $\boldsymbol{\theta}^0$ is a representative value such as the mode of Π. We discuss the value of t further when we consider convergence issues for the practical implementation of MCMC methods on page 26.

Metropolis-Hastings Algorithm

So far we have discussed some general properties of MCMC samplers. We now describe a specific, but very flexible, MCMC sampler known as the *Metropolis-Hastings algorithm*. The most popular of MCMC samplers: the random-walk algorithm, the independence sampler and Gibbs sampling are all special cases of the Metropolis-Hastings algorithm.

The Metropolis-Hastings algorithm works by generating proposals from a proposal distribution. These proposals are then accepted or rejected in such a way that the accepted proposals form a sample from the target distribution. We provide an informal derivation of the algorithm.

Suppose $(\boldsymbol{\theta}^i, \boldsymbol{\theta}^*)$ is the random vector consisting of the current value of the chain, $\boldsymbol{\theta}^i$, and the proposal, $\boldsymbol{\theta}^*$. This vector has joint pdf

$$f(\boldsymbol{\theta}^i, \boldsymbol{\theta^*}) = q(\boldsymbol{\theta}^i, \boldsymbol{\theta}^*)\pi(\boldsymbol{\theta}^i),$$

where $q(\boldsymbol{\theta}^i, \boldsymbol{\theta}^*)$ is the pdf of the proposal distribution, Q, which is conditional on $\boldsymbol{\theta}^i$. If $\boldsymbol{\theta}^*$ was automatically accepted and for reversibility to hold, the following condition should be satisfied

$$q(\boldsymbol{\theta}^i, \boldsymbol{\theta}^*)\pi(\boldsymbol{\theta}^i) = q(\boldsymbol{\theta}^*, \boldsymbol{\theta}^i)\pi(\boldsymbol{\theta}^*). \tag{2.12}$$

Now (2.12) is unlikely to hold, so we introduce the acceptance probability, $\alpha(.,.) \leq 1$, such that transitions from $\boldsymbol{\theta}^i$ to $\boldsymbol{\theta}^*$ are accepted with probability $\alpha(\boldsymbol{\theta}^i, \boldsymbol{\theta}^*)$, transitions from $\boldsymbol{\theta}^*$ to $\boldsymbol{\theta}^i$ are accepted with probability $\alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta}^i)$, and

$$q(\boldsymbol{\theta}^i, \boldsymbol{\theta}^*)\alpha(\boldsymbol{\theta}^i, \boldsymbol{\theta}^*)\pi(\boldsymbol{\theta}^i) = q(\boldsymbol{\theta}^*, \boldsymbol{\theta}^i)\alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta}^i)\pi(\boldsymbol{\theta}^*). \tag{2.13}$$

Suppose

$$q(\boldsymbol{\theta}^i, \boldsymbol{\theta}^*)\pi(\boldsymbol{\theta}^i) > q(\boldsymbol{\theta}^*, \boldsymbol{\theta}^i)\pi(\boldsymbol{\theta}^*) \tag{2.14}$$

and set $\alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta}^i) = 1$, then

$$\alpha(\boldsymbol{\theta}^i, \boldsymbol{\theta}^*) = \frac{\pi(\boldsymbol{\theta}^*)q(\boldsymbol{\theta}^*, \boldsymbol{\theta}^i)}{\pi(\boldsymbol{\theta}^i)q(\boldsymbol{\theta}^i, \boldsymbol{\theta}^*)}.$$

Now, suppose that the inequality in (2.14) is reversed and $\alpha(\boldsymbol{\theta}^i, \boldsymbol{\theta}^*) = 1$, then reversibility is ensured if

$$\alpha(\boldsymbol{\theta}^i, \boldsymbol{\theta}^*) = \min\left[1, \frac{\pi(\boldsymbol{\theta}^*)q(\boldsymbol{\theta}^*, \boldsymbol{\theta}^i)}{\pi(\boldsymbol{\theta}^i)q(\boldsymbol{\theta}^i, \boldsymbol{\theta}^*)}\right]. \tag{2.15}$$

In practice, it is usually the case that $\pi(\boldsymbol{\theta}) = g(\boldsymbol{\theta})/\int_\Theta g(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}$, where the normalising constant, $\int_\Theta g(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}$ is unavailable. However, it is clear that

$$\frac{\pi(\boldsymbol{\theta}^*)}{\pi(\boldsymbol{\theta}^i)} = \frac{g(\boldsymbol{\theta}^*)}{g(\boldsymbol{\theta}^i)},$$

since $\int_\Theta g(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}$ cancels out in both the numerator and denominator, and we can still evaluate $\alpha(\boldsymbol{\theta}^i, \boldsymbol{\theta}^*)$.

In summary, the Metropolis-Hastings algorithm proceeds as follows:

1. Choose an initial value, $\boldsymbol{\theta}^0$.
2. Suppose the current value of the chain is $\boldsymbol{\theta}^i$.
3. Generate a proposal, $\boldsymbol{\theta}^*$, from the proposal distribution, Q, with pdf, $q(\boldsymbol{\theta}^i, .)$.
4. Calculate the acceptance probability, $\alpha(\boldsymbol{\theta}^i, \boldsymbol{\theta}^*)$, according to (2.15).
5. With probability $\alpha(\boldsymbol{\theta}^i, \boldsymbol{\theta}^*)$ accept the proposal and set the current value as $\boldsymbol{\theta}^{i+1} = \boldsymbol{\theta}^*$. Otherwise, set the current value as $\boldsymbol{\theta}^{i+1} = \boldsymbol{\theta}^i$.
6. Repeat steps 2. to 6.

The Metropolis-Hastings algorithm provides a general-purpose method of generating a sample from an arbitrary multivariate distribution where we may not have a normalised pdf. O'Hagan and Forster (2004, pg. 267) state that a sufficient condition for ergodicity is that $q(\boldsymbol{\theta}^i, \boldsymbol{\theta}) > 0$ for all $\boldsymbol{\theta} \in \Theta$.

We now describe some important special cases of the Metropolis-Hastings algorithm which arise from particular choices of the proposal distribution.

Suppose the proposal distribution is symmetric about the current value of the chain, $\boldsymbol{\theta}^i$, such that $\boldsymbol{\theta}^* = \boldsymbol{\theta}^i + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ is a realisation of some random variable whose distribution is symmetric about $\mathbf{0}$ and does not depend on $\boldsymbol{\theta}^i$. In this case, $q(\boldsymbol{\theta}^i, \boldsymbol{\theta}^*) = q(\boldsymbol{\theta}^*, \boldsymbol{\theta}^i)$ and the acceptance probability reduces to

$$\alpha(\boldsymbol{\theta}^i, \boldsymbol{\theta}^*) = \min\left[1, \frac{\pi(\boldsymbol{\theta}^*)}{\pi(\boldsymbol{\theta}^i)}\right]. \tag{2.16}$$

A Metropolis-Hastings algorithm with this choice of proposal distribution is known as a *random-walk Metropolis-Hastings algorithm*. We see from (2.16), that proposals to areas of higher target density will always be accepted.

Typically, the distribution of $\boldsymbol{\epsilon}$ will be normal with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$, i.e. $\boldsymbol{\theta}^*|\boldsymbol{\theta}^i \sim \mathrm{N}(\boldsymbol{\theta}^i, \boldsymbol{\Sigma})$.

The key issue in the implementation of the random-walk algorithm is the variance of the proposal distribution, $\boldsymbol{\Sigma}$. For example, consider a one-dimensional example where $\theta^* = \theta^i + \epsilon$, and $\epsilon \sim \mathrm{N}(0, \sigma_\epsilon^2)$. If σ_ϵ^2 is small, then θ^* will be close to θ^i and the acceptance probability will be close to 1. Therefore, proposals will be accepted with high probability but the chain will take a large number of iterations to fully explore the target distribution. On the other hand, suppose σ_ϵ^2 is large, then a lot of the proposals will be made in regions of low target density and will be rejected.

This means that some tuning is required before we can run the Metropolis-Hastings algorithm and save the accepted proposals. One way of tuning is to assess the acceptance rate of the algorithm. Roberts and Rosenthal (2001) assess the efficiency of various Metropolis-Hastings algorithms and state that for the random-walk Metropolis-Hastings algorithm "on smooth densities, any acceptance rate between 0.1 and 0.4 ought to perform close to optimal". They also state that even low acceptance rates of order 0.1 "can be very close to optimal". Gelman et al. (1996) find that the optimal acceptance rate is approximately 23%.

Suppose the proposal distribution does not depend on the the current value of the chain, $\boldsymbol{\theta}^i$, so that $q(\boldsymbol{\theta}^i, \boldsymbol{\theta}^*) = s(\boldsymbol{\theta}^*)$, then the acceptance probability simplifies to

$$\alpha(\boldsymbol{\theta}^i, \boldsymbol{\theta}^*) = \min\left[1, \frac{\pi(\boldsymbol{\theta}^*)s(\boldsymbol{\theta}^i)}{\pi(\boldsymbol{\theta}^i)s(\boldsymbol{\theta}^*)}\right]. \tag{2.17}$$

A Metropolis-Hastings algorithm with this proposal distribution is known as the *independence sampler*.

The performance of the independence sampler depends on how well the proposal distribution 'mimics' the target distribution. If the tails of the proposal distribution are light relative to those of the target distribution, then a low number of proposals are made in the tails of the target distribution, this leads to the chain compensating and becoming stuck at a tail value when one is actually proposed. This is the same problem that impacts upon multi-dimensional rejection sampling as discussed in Section 2.2.3. O'Hagan and Forster (2004, pg. 271) state that, for these reasons, "the independence sampler is rarely the most efficient MCMC sampler for any given problem".

We now describe a very popular Metropolis-Hastings algorithm known as *Gibbs sampling*. Let $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_j, ..., \boldsymbol{\theta}_B)^T$ be the k dimensional vector of parameters where the target distribution has pdf, $\pi(\boldsymbol{\theta})$. Therefore, we have partitioned $\boldsymbol{\theta}$ into B blocks. Let $\boldsymbol{\theta}_j$ be the jth block of k_j parameters, so that $\sum_{j=1}^{B} k_j = k$. Finally, let $\boldsymbol{\theta}_{\setminus j} = (\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_{j-1}, \boldsymbol{\theta}_{j+1}, ..., \boldsymbol{\theta}_B)^T$, which is $\boldsymbol{\theta}$ with the jth block, $\boldsymbol{\theta}_j$, removed. The jth full conditional distribution, Π_j, of $\boldsymbol{\theta}_j$

has pdf

$$\begin{aligned}\pi_j(\boldsymbol{\theta}_j) &= \pi(\boldsymbol{\theta}_j|\boldsymbol{\theta}_{\setminus j}),\\ &= \frac{\pi(\boldsymbol{\theta}_1,...,\boldsymbol{\theta}_j,...,\boldsymbol{\theta}_B)}{\int_{\Theta_j}\pi(\boldsymbol{\theta}_1,...,\boldsymbol{\theta}_j,...,\boldsymbol{\theta}_B)\mathrm{d}\boldsymbol{\theta}_j},\\ &= \frac{\pi(\boldsymbol{\theta}_1,...,\boldsymbol{\theta}_j,...,\boldsymbol{\theta}_B)}{c(\boldsymbol{\theta}_{\setminus j})}.\end{aligned}$$

Suppose the current value of the chain is $\boldsymbol{\theta}^i = (\boldsymbol{\theta}^i_1,...,\boldsymbol{\theta}^i_{j-1},\boldsymbol{\theta}^i_j,\boldsymbol{\theta}^i_{j+1},...,\boldsymbol{\theta}^i_B)^T$ and we propose to just update the jth block. Therefore the proposal is $\boldsymbol{\theta}^* = (\boldsymbol{\theta}^i_1,...,\boldsymbol{\theta}^i_{j-1},\boldsymbol{\theta}^*_j,\boldsymbol{\theta}^i_{j+1},...,\boldsymbol{\theta}^i_B)^T$ which we generate from the distribution with pdf

$$q(\boldsymbol{\theta}^i,\boldsymbol{\theta}^*) = \pi_j(\boldsymbol{\theta}_j).$$

Consider the acceptance probability of the resulting Metropolis-Hastings algorithm

$$\begin{aligned}\alpha(\boldsymbol{\theta}^i,\boldsymbol{\theta}^*) &= \min\left[1, \frac{\pi(\boldsymbol{\theta}^i_1,...,\boldsymbol{\theta}^i_{j-1},\boldsymbol{\theta}^*_j,\boldsymbol{\theta}^i_{j+1},...,\boldsymbol{\theta}^i_B)\frac{\pi(\boldsymbol{\theta}^i_1,...,\boldsymbol{\theta}^i_{j-1},\boldsymbol{\theta}^i_j,\boldsymbol{\theta}^i_{j+1},...,\boldsymbol{\theta}^i_B)}{c(\boldsymbol{\theta}_{\setminus j})}}{\pi(\boldsymbol{\theta}^i_1,...,\boldsymbol{\theta}^i_{j-1},\boldsymbol{\theta}^i_j,\boldsymbol{\theta}^i_{j+1},...,\boldsymbol{\theta}^i_B)\frac{\pi(\boldsymbol{\theta}^i_1,...,\boldsymbol{\theta}^i_{j-1},\boldsymbol{\theta}^*_j,\boldsymbol{\theta}^i_{j+1},...,\boldsymbol{\theta}^i_B)}{c(\boldsymbol{\theta}_{\setminus j})}}\right],\\ &= 1.\end{aligned}$$

Hence, by choosing a proposal distribution for $\boldsymbol{\theta}_j$ equal to the full conditional distribution, Π_j, of $\boldsymbol{\theta}_j$, then the acceptance probability will always be one and all proposals will be accepted.

The Gibbs sampling algorithm is:

1. Choose an initial value, $\boldsymbol{\theta}^0$.
2. Suppose the current value of the chain is $\boldsymbol{\theta}^i$.
3. The next value of the chain, $\boldsymbol{\theta}^{i+1}$, is obtained in the following way:

$$\begin{aligned}&\boldsymbol{\theta}^{i+1}_1 \text{ is generated from } \Pi_1 \text{ with pdf } \pi(\boldsymbol{\theta}_1|\boldsymbol{\theta}^i_2,...,\boldsymbol{\theta}^i_B)\\ &\boldsymbol{\theta}^{i+1}_2 \text{ is generated from } \Pi_2 \text{ with pdf } \pi(\boldsymbol{\theta}_2|\boldsymbol{\theta}^{i+1}_1,\boldsymbol{\theta}^i_3,...,\boldsymbol{\theta}^i_B)\\ &\qquad\qquad\qquad\qquad\cdots\cdots\\ &\boldsymbol{\theta}^{i+1}_j \text{ is generated from } \Pi_j \text{ with pdf } \pi(\boldsymbol{\theta}_j|\boldsymbol{\theta}^{i+1}_1,...,\boldsymbol{\theta}^{i+1}_{j-1},\boldsymbol{\theta}^i_{j+1},...,\boldsymbol{\theta}^i_B)\\ &\qquad\qquad\qquad\qquad\cdots\cdots\\ &\boldsymbol{\theta}^{i+1}_B \text{ is generated from } \Pi_B \text{ with pdf } \pi(\boldsymbol{\theta}_B|\boldsymbol{\theta}^{i+1}_1,...,\boldsymbol{\theta}^{i+1}_{B-1}).\end{aligned}$$

4. Repeat steps 2. to 4.

Step 3. is known as a *scan* of the Gibbs sampling algorithm.

The only drawback of Gibbs sampling is the difficulty in generating a value, $\boldsymbol{\theta}^{i+1}_j$, from the full conditional distribution of $\boldsymbol{\theta}_j$ (known as *updating* $\boldsymbol{\theta}_j$) since these are, typically, not

available. We can use a general Metropolis-Hastings step to update $\boldsymbol{\theta}_j$ and this is known as the *Metropolis within Gibbs algorithm*. Alternatively, if we partition $\boldsymbol{\theta}$ so that $\boldsymbol{\theta}_j$ is scalar, i.e. $\boldsymbol{\theta}_j = \theta_j$, then we can use the method of adaptive rejection sampling described in Section 2.2.3 to generate a value from the univariate full conditional distribution, provided $\pi_j(\theta_j)$ is log-concave.

Gibbs sampling is the most popular MCMC sampler in applied Bayesian statistics for generating a sample from the posterior distribution. Two possible reasons for this are conditional conjugacy and conditional independence.

If the block, $\boldsymbol{\theta}_j$, is *conditionally conjugate* then the full conditional distribution of $\boldsymbol{\theta}_j$ will belong to the same family as the prior distribution of $\boldsymbol{\theta}_j$ and will be, presumably, easy to generate from. This can be the case for the variance components matrix, $\mathbf{D}$, in a GLMM and will be discussed further in Section 2.2.7.

Conditional independence occurs in hierarchical models (which GLMMs are a subset of). Suppose $f(\mathbf{y}|\boldsymbol{\theta}) = \prod_j f(\mathbf{y}_j|\boldsymbol{\theta}_j)$ and $f(\boldsymbol{\theta}|\boldsymbol{\phi}) = \prod_j f(\boldsymbol{\theta}_j|\boldsymbol{\phi})$, then the posterior distribution of the $\boldsymbol{\theta}_j$'s will be conditionally independent given $\boldsymbol{\phi}$. This is the case for the group-specific parameters, $\mathbf{u}_i$'s, in a GLMM.

Due to these reasons, there exist off-the-shelf software packages such as BUGS (Bayesian inference Using Gibbs Sampling) and JAGS (Just Another Gibbs Sampler) which can produce a sample from the posterior distribution without the user having to specify the method used to generate from the full conditional distributions. The ease of use of these packages has further popularised Gibbs sampling.

These issues are further discussed in relation to GLMMs in Section 2.2.7.

Adaptive Rejection Metropolis Sampling

In Section 2.2.3, we discussed the method of adaptive rejection sampling for generating from a univariate distribution, Π, where the pdf of this distribution, $\pi(\theta)$, is log-concave. Being able to do so is important when we considered Gibbs sampling in the previous Section.

Gilks et al. (1995) proposed *Adaptive Rejection Metropolis Sampling* (ARMS) as a generalisation of ARS when $\pi(\theta)$ is not necessarily log-concave. It does so by the introduction of a Metropolis accept/reject step. If $\pi(\theta)$ is log-concave, then ARMS reduces to the derivative free version of ARS proposed by Gilks (1992). See Gilks et al. (1995) for more details on ARMS, including the algorithm.

Convergence Issues for MCMC

On page 22, we discussed how, for a Markov Chain, $\left\{\boldsymbol{\theta}^0, ..., \boldsymbol{\theta}^t, \boldsymbol{\theta}^{t+1}, \boldsymbol{\theta}^{t+n}\right\}$, with initial value $\boldsymbol{\theta}^0$, we discard the first $t+1$ values $\left\{\boldsymbol{\theta}^0, ..., \boldsymbol{\theta}^t\right\}$, and retain the last n values, $\left\{\boldsymbol{\theta}^{t+1}, \boldsymbol{\theta}^{t+n}\right\}$, for

inference. The first t iterations of an MCMC algorithm that produces $\left\{\boldsymbol{\theta}^0, ..., \boldsymbol{\theta}^t\right\}$ is called the *burn-in phase*. The key question is: what value should t take so that the chain has become independent of $\boldsymbol{\theta}^0$? When a chain has become independent of $\boldsymbol{\theta}^0$ it is said to have *converged in distribution*. If $\boldsymbol{\theta}^0$ is a representative value of Π, such as the mode, then t can be 0.

There exist many formal convergence diagnostic tools, many of which are implemented in BUGS. We, however, take an informal, entirely pragmatic approach to assessing convergence by inspecting trace plots of the model parameters as suggested by O'Hagan and Forster (2004, pg. 287-288).

We can improve convergence in Gibbs sampling, where the full conditional distributions are univariate, by using the method of *ordered overrelaxation* proposed by Neal (1995). Suppose we are in Step 3. of the Gibbs sampling algorithm on page 25, and are attempting to generate a value, θ_j^{i+1}, from the univariate full conditional distribution with pdf $\pi(\theta_j|\boldsymbol{\theta}_1^{i+1}, ..., \boldsymbol{\theta}_{j-1}^{i+1}, \boldsymbol{\theta}_{j+1}^{i}, ..., \boldsymbol{\theta}_B^{i})$. The ordered overrelaxation algorithm to obtain θ_j^{i+1} is

1. Generate N values, independently, from the full conditional distribution with pdf $\pi(\theta_j|\boldsymbol{\theta}_1^{i+1}, ..., \boldsymbol{\theta}_{j-1}^{i+1}, \boldsymbol{\theta}_{j+1}^{i}, ..., \boldsymbol{\theta}_B^{i})$.
2. Arrange these N values, in addition to θ_j^i, in increasing order, as follows

   $$\theta_j^{(0)} \leq \theta_j^{(1)} \leq ... \leq \theta_j^{(r)} = \theta_j^i \leq ... \leq \theta_j^{(N)},$$

   where ties are broken at random and r is the index of θ_j^i.
3. Let $\theta_j^{i+1} = \theta_j^{(N-r)}$.

Proof that ordered overrelaxation retains Π as the stationary distribution is given by Neal (1995). Note that if $N = 1$, then ordered overrelaxation is ordinary Gibbs sampling. It would first appear that the computational expense of ordered overrelaxation is approximately N times that of ordinary Gibbs sampling. However, by using ARS or ARMS to achieve the sampling in Step 1. of the ordered overrelaxation algorithm, we adapt the sampling distribution to be closer to the full conditional distribution, so the sampling becomes more efficient and the computational expense is less than N times that of ordinary Gibbs sampling. Neal (1995) suggests using $N = 20$ for routine use.

Ordered overrelaxation can be implemented in BUGS. For the remainder of this thesis, we use ordered overrelaxation whenever we generate a posterior sample using Gibbs sampling.

2.2.5 Monte Carlo Integration

A crucial quantity in Bayesian model determination is the marginal likelihood for model $m \in M$

$$f_m(\mathbf{y}) = \int_\Theta f_m(\mathbf{y}|\boldsymbol{\theta}_m) f_m(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}.$$

The marginal likelihoods are used to evaluate the posterior model probabilities of models in M according to (1.7) and the Bayes' factor between any two models in M, i.e. $f_{m_1}(\mathbf{y})/f_{m_2}(\mathbf{y})$. In this Section, we describe several Monte Carlo methods for approximating $f_m(\mathbf{y})$. Throughout this Section we describe methods for approximating a general integral

$$I = \int_\Theta g(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta},$$

where $\pi(\boldsymbol{\theta}) = g(\boldsymbol{\theta})/\int_\Theta g(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}$ is the pdf of the distribution, Π, and $I = f_m(\mathbf{y})$ if $g(\boldsymbol{\theta}) = f_m(\mathbf{y}|\boldsymbol{\theta}_m)f_m(\boldsymbol{\theta})$, or if, equivalently, Π is the posterior distribution. However, in some cases the prior distribution may not have a normalised pdf and we may need to find the normalising constant, hence the more general setup for this Section.

Importance Sampling

Recall from Section 2.2.1, the Monte Carlo approximation, $\hat{\mu}_f$, to $\mu_f = \mathrm{E}(f(\boldsymbol{\theta}))$ where $\boldsymbol{\theta} \sim \Pi$ for some function $f : \Theta \to \mathbb{R}$, is to generate $\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_n$ from Π and then set

$$\hat{\mu}_f = \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{\theta}_i).$$

Consider the integral

$$\begin{aligned} I &= \int_\Theta \frac{g(\boldsymbol{\theta})}{h(\boldsymbol{\theta})} h(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} \\ &= \mathrm{E}_H\left(\frac{g(\boldsymbol{\theta})}{h(\boldsymbol{\theta})}\right), \end{aligned} \tag{2.18}$$

where $\boldsymbol{\theta}$ is from the distribution, H, with the pdf, $h(\boldsymbol{\theta})$. Note that (2.18) can be approximated using the Monte Carlo method, thus

$$\hat{I}_{IS} = \frac{1}{n}\sum_{i=1}^{n} \frac{g(\boldsymbol{\theta}_i)}{h(\boldsymbol{\theta}_i)}, \tag{2.19}$$

and $\{\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_n\}$ is a sample generated from H. This method is known as *importance sampling*.

It is easy to see that $\mathrm{E}(\hat{I}_{IS}) = I$, so $\hat{I}_{IS}$ is an unbiased approximation of I with variance

$$\begin{aligned} \mathrm{var}(\hat{I}_{IS}) &= \frac{1}{n}\mathrm{var}_H\left(\frac{g(\boldsymbol{\theta})}{h(\boldsymbol{\theta})}\right), \\ &\approx \frac{1}{n^2}\sum_{i=1}^{n}\frac{g(\boldsymbol{\theta}_i)^2}{h(\boldsymbol{\theta}_i)^2} - \frac{1}{n}\hat{I}_{IS}^2. \end{aligned}$$

As n increases, $\hat{I}_{IS}$ becomes a more accurate approximation to I, provided that $\mathrm{var}_H\left(\frac{g(\boldsymbol{\theta})}{h(\boldsymbol{\theta})}\right)$ is finite. The variance of the importance sampling approximation depends upon how well $h(\boldsymbol{\theta})$ 'mimics' $g(\boldsymbol{\theta})$ in a similar way to how the performance of the independence sampler

depends on how well the proposal distribution 'mimics' the target distribution and also how well the sampling distribution, S, 'mimics' Π in rejection sampling. To see this, suppose $h(\boldsymbol{\theta}) \propto g(\boldsymbol{\theta})$, then $\text{var}(\hat{I}_{IS}) = 0$ and $\hat{I}_{IS} = I$. A common choice for the distribution, H, is the normal distribution but we see in the next example that this can be a bad choice.

Example (from O'Hagan and Forster (2004, pg. 254))

Suppose $\theta \in \mathbb{R}$ and $g(\theta) \propto (1+\theta^2)^{-1}$, i.e. Π is a t distribution with 1 degree of freedom, also known as the Cauchy distribution, and let H be $\text{N}(m, v)$. Then

$$\text{E}_H\left(\frac{g(\theta)^2}{h(\theta)^2}\right) = (2\pi v)^{\frac{1}{2}} \int_{\mathbb{R}} (1+\theta^2)^{-2} \exp\left[\frac{(\theta-m)^2}{2v}\right] \text{d}\theta,$$

this integral is divergent for all m and v, and $\text{var}(\hat{I}_{IS}) = \infty$. So if Π is any t-distribution and H is any normal distribution then $\hat{I}_{IS}$ will not converge even as $n \to \infty$. ■

Reciprocal Importance Sampling

Note that

$$\begin{aligned} 1 &= \int_\Theta h(\boldsymbol{\theta}) \text{d}\boldsymbol{\theta}, \\ &= \int_\Theta \frac{h(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})} \pi(\boldsymbol{\theta}) \text{d}\boldsymbol{\theta}, \\ &= \int_\Theta \frac{h(\boldsymbol{\theta}) I}{g(\boldsymbol{\theta})} \pi(\boldsymbol{\theta}) \text{d}\boldsymbol{\theta}. \end{aligned}$$

Therefore $I = \text{E}_\pi\left[\frac{h(\boldsymbol{\theta})}{g(\boldsymbol{\theta})}\right]^{-1}$, and the *reciprocal importance sampling* approximation to I is

$$\hat{I}_{RIS} = \left[\frac{1}{n}\sum_{i=1}^{n} \frac{h(\boldsymbol{\theta}_i)}{g(\boldsymbol{\theta}_i)}\right]^{-1},$$

where $\{\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_n\}$ is a sample generated from Π. Noting that $\text{E}\left(\frac{1}{X}\right) > \frac{1}{\text{E}(X)}$ for any positive random variable, X, it can be shown that $\text{E}(\hat{I}_{RIS}) > I$, so the reciprocal importance sampling approximation to I is biased. The variance of the approximation is

$$\text{var}(\hat{I}_{RIS}) = \frac{I^4}{n} \text{var}_\Pi\left(\frac{h(\boldsymbol{\theta})}{g(\boldsymbol{\theta})}\right) + O\left(\frac{1}{n^2}\right).$$

The approximation is asymptotically, with respect to n, unbiased, provided $\text{var}_\Pi\left(\frac{h(\boldsymbol{\theta})}{g(\boldsymbol{\theta})}\right)$ is finite.

Candidate's Method

Recall from page 28 that $\pi(\boldsymbol{\theta}) = g(\boldsymbol{\theta})/I$ which leads to the *candidate's formula*:

$$I = \frac{g(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})}.$$

Therefore, the *candidate's method* approximation to I is

$$\hat{I}_C = \frac{g(\boldsymbol{\theta}^*)}{\hat{\pi}(\boldsymbol{\theta}^*)}, \tag{2.20}$$

for any value $\boldsymbol{\theta}^* \in \Theta$. Obviously, this method relies on having an approximation to the density, $\hat{\pi}(\boldsymbol{\theta}^*)$, of Π at $\boldsymbol{\theta}^*$, based on a sample $\{\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_n\}$ generated from Π. Methods of density estimation will be more accurate in areas of high density, so it is suggested that the value $\boldsymbol{\theta}^*$ used in (2.20) be close to the mode of Π. In low dimensional problems, we can use kernel density estimation, but this is unlikely to be sufficiently accurate in higher dimensions.

Chib and Jeliazkov (2001) propose a method for approximating the density, $\hat{\pi}(\boldsymbol{\theta}^*)$, based on output from a Metropolis-Hastings algorithm. Using (2.13), which ensures that reversibility holds for the Metropolis-Hastings algorithm, we have

$$\int_\Theta q(\boldsymbol{\theta}, \boldsymbol{\theta}^*)\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^*)\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} = \int_\Theta q(\boldsymbol{\theta}^*, \boldsymbol{\theta})\alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta})\pi(\boldsymbol{\theta}^*)\mathrm{d}\boldsymbol{\theta},$$

which leads to

$$\begin{aligned} \pi(\boldsymbol{\theta}^*) &= \frac{\int_\Theta q(\boldsymbol{\theta}, \boldsymbol{\theta}^*)\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^*)\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}}{\int_\Theta q(\boldsymbol{\theta}^*, \boldsymbol{\theta})\alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}}, \\ &= \frac{\mathrm{E}_\pi\left[q(\boldsymbol{\theta}, \boldsymbol{\theta}^*)\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^*)\right]}{\mathrm{E}_{Q|\boldsymbol{\theta}^*}\left[\alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta})\right]}. \end{aligned} \tag{2.21}$$

The numerator and denominator of (2.21) can be approximated using the Monte Carlo method, i.e.

$$\hat{\pi}(\boldsymbol{\theta}^*) = \frac{\frac{1}{n_2}\sum_{i=1}^{n_2} q(\boldsymbol{\theta}_i^{(2)}, \boldsymbol{\theta}^*)\alpha(\boldsymbol{\theta}_i^{(2)}, \boldsymbol{\theta}^*)}{\frac{1}{n_1}\sum_{i=1}^{n_1} \alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta}_i^{(1)})},$$

where $\{\boldsymbol{\theta}_1^{(1)}, ..., \boldsymbol{\theta}_{n_1}^{(1)}\}$ and $\{\boldsymbol{\theta}_1^{(2)}, ..., \boldsymbol{\theta}_{n_2}^{(2)}\}$ are samples generated from the proposal distribution given the current value, $\boldsymbol{\theta}^*$, which is denoted $Q|\boldsymbol{\theta}^*$, and Π, respectively, and also $q(\boldsymbol{\theta}_i^{(2)}, \boldsymbol{\theta}^*)$ is the pdf of the proposal distribution given current value, $\boldsymbol{\theta}_i^{(2)}$, evaluated at $\boldsymbol{\theta}^*$. Alternatively, Chib (1995) proposes a method for approximating the density of Π at $\boldsymbol{\theta}^*$ based on output from the Gibbs sampler.

Bridge Sampling

Suppose that $h(\boldsymbol{\theta})$ is the pdf of the distribution, H, and that $\gamma(\boldsymbol{\theta})$ is a function such that $0 < |\int_\Theta \gamma(\boldsymbol{\theta})g(\boldsymbol{\theta})h(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}| < \infty$, and that since

$$1 = \frac{\int_\Theta \gamma(\boldsymbol{\theta})g(\boldsymbol{\theta})h(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}}{\int_\Theta \gamma(\boldsymbol{\theta})g(\boldsymbol{\theta})h(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}},$$

it follows that

$$\begin{aligned} I &= \frac{\int_\Theta \gamma(\boldsymbol{\theta}) g(\boldsymbol{\theta}) h(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}}{\int_\Theta \gamma(\boldsymbol{\theta}) h(\boldsymbol{\theta}) \pi(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}}, \\ &= \frac{\mathrm{E}_H\left[\gamma(\boldsymbol{\theta}) g(\boldsymbol{\theta})\right]}{\mathrm{E}_\Pi\left[\gamma(\boldsymbol{\theta}) h(\boldsymbol{\theta})\right]}. \end{aligned} \tag{2.22}$$

As for approximating the density of Π at $\boldsymbol{\theta}^*$ in the candidate's method, we can approximate the numerator and denominator of (2.22) using the Monte Carlo method. Therefore, the *bridge sampling* approximation to I is

$$\hat{I}_{BS} = \frac{\frac{1}{n_H} \sum_{i=1}^{n_H} \gamma(\boldsymbol{\theta}_i^H) g(\boldsymbol{\theta}_i^H)}{\frac{1}{n_\Pi} \sum_{i=1}^{n_\Pi} \gamma(\boldsymbol{\theta}_i^\Pi) h(\boldsymbol{\theta}_i^\Pi)}, \tag{2.23}$$

where $\{\boldsymbol{\theta}_1^H, ..., \boldsymbol{\theta}_{n_H}^H\}$ and $\{\boldsymbol{\theta}_1^\Pi, ..., \boldsymbol{\theta}_{n_\Pi}^\Pi\}$ are samples generated from H and Π, respectively. Bridge sampling was first proposed by Meng and Wong (1996) in a slightly different way for approximating the ratio of normalising constants where $h(\boldsymbol{\theta})$ is also unnormalised.

Noting that $\mathrm{E}\left(\frac{X}{Y}\right) > \frac{\mathrm{E}(X)}{\mathrm{E}(Y)}$, for any positive random variables X and Y, it can be shown that $\mathrm{E}(\hat{I}_{BS}) > I$ for finite n_H and n_Π, but is asymptotically, with respect to n_H and n_Π, unbiased. The variance of the approximation is

$$\mathrm{var}(\hat{I}_{BS}) = \frac{I^2}{n_H} \frac{\mathrm{var}_H(g(\boldsymbol{\theta})\gamma(\boldsymbol{\theta}))}{\mathrm{E}_H(g(\boldsymbol{\theta})\gamma(\boldsymbol{\theta}))^2} + \frac{I^4}{n_\Pi} \frac{\mathrm{var}_\Pi(h(\boldsymbol{\theta})\gamma(\boldsymbol{\theta}))}{\mathrm{E}_H(g(\boldsymbol{\theta})\gamma(\boldsymbol{\theta}))^2} + O\left(\frac{1}{n_H^2 + n_\Pi^2}\right). \tag{2.24}$$

If $\gamma(\boldsymbol{\theta}) = \frac{1}{h(\boldsymbol{\theta})}$, then bridge sampling reduces to importance sampling. Similarly, if $\gamma(\boldsymbol{\theta}) = \frac{1}{g(\boldsymbol{\theta})}$, then $\hat{I}_{BS} = \hat{I}_{RIS}$. Meng and Wong (1996) show that the optimal $\gamma(\boldsymbol{\theta})$, with respect to minimising (2.24), is

$$\gamma_O(\boldsymbol{\theta}) = \left(n_\Pi g(\boldsymbol{\theta}) + n_H I h(\boldsymbol{\theta})\right)^{-1},$$

with variance

$$\mathrm{var}(\hat{I}_{BS,O}) = \frac{I^2}{n_H n_\Pi} \left[\int_\Theta \frac{\pi(\boldsymbol{\theta}) h(\boldsymbol{\theta})}{n_H h(\boldsymbol{\theta}) + n_\Pi \pi(\boldsymbol{\theta})} \mathrm{d}\boldsymbol{\theta}\right]^{-1} - \frac{1}{n_H} - \frac{1}{n_\Pi} + O\left(\frac{1}{n_H^2 + n_\Pi^2}\right). \tag{2.25}$$

Obviously, the optimal $\gamma(\boldsymbol{\theta})$ depends on the unknown normalising constant, I. However, Meng and Wong (1996) suggest iterating the following scheme, starting from an initial value, $\hat{I}_{BS,O}^{(0)}$, until convergence

$$\hat{I}_{BS,O}^{(t+1)} = \frac{\frac{1}{n_H} \sum_{i=1}^{n_H} \frac{l_{Hi}}{n_\Pi l_{Hi} + n_H \hat{I}_{BS,O}^{(t)}}}{\frac{1}{n_\Pi} \sum_{i=1}^{n_\Pi} \frac{1}{n_\Pi l_{\Pi i} + n_H \hat{I}_{BS,O}^{(t)}}}, \tag{2.26}$$

where $l_{ki} = g(\boldsymbol{\theta}_i^{(k)}) / h(\boldsymbol{\theta}_i^{(k)})$ for $k = H, \Pi$.

Consider the general bridge sampler approximation given in (2.23), where $h(\boldsymbol{\theta}) = q(\boldsymbol{\theta}^*, \boldsymbol{\theta})$, i.e. H is the proposal distribution of a Metropolis-Hastings algorithm given current value $\boldsymbol{\theta}^*$ and $\gamma(\boldsymbol{\theta}) = \alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta}) / g(\boldsymbol{\theta})$. The bridge sampling approximation to I is then

$$\hat{I}_{BS} = \frac{\frac{1}{n_H} \sum_{i=1}^{n_H} \alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta}_i^H)}{\frac{1}{n_\Pi} \sum_{i=1}^{n_\Pi} \alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta}_i^\Pi) h(\boldsymbol{\theta}_i^\Pi) / g(\boldsymbol{\theta}_i^\Pi)}.$$

Using (2.13), we see that

$$\alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta}_i^{\Pi}) = \frac{q(\boldsymbol{\theta}_i^{\Pi}, \boldsymbol{\theta}^*)\alpha(\boldsymbol{\theta}_i^{\Pi}, \boldsymbol{\theta}^*)g(\boldsymbol{\theta}_i^{\Pi})}{h(\boldsymbol{\theta}_i^{\Pi})g(\boldsymbol{\theta}^*)},$$

and

$$\alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta}_i^{\Pi})h(\boldsymbol{\theta}_i^{\Pi})/g(\boldsymbol{\theta}_i^{\Pi}) = \frac{q(\boldsymbol{\theta}_i^{\Pi}, \boldsymbol{\theta}^*)\alpha(\boldsymbol{\theta}_i^{\Pi}, \boldsymbol{\theta}^*)}{g(\boldsymbol{\theta}^*)},$$

so the bridge sampling approximation reduces to

$$\hat{I}_{BS} = \frac{\frac{g(\boldsymbol{\theta}^*)}{n_H}\sum_{i=1}^{n_H}\alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta}_i^H)}{\frac{1}{n_\Pi}\sum_{i=1}^{n_\Pi}\alpha(\boldsymbol{\theta}_i^{\Pi}, \boldsymbol{\theta}^*)q(\boldsymbol{\theta}_i^{\Pi}, \boldsymbol{\theta}^*)}. \tag{2.27}$$

We can see that (2.27) is the Chib and Jeliazkov (2001) candidate's method approximation to I which is, therefore, a special case of bridge sampling with a sub-optimal choice of $\gamma(\boldsymbol{\theta})$. This was first noted by Mira and Nicholls (2004).

We defer further discussion of bridge sampling, as applied to GLMMs, to Subsection 2.2.7 and also our implementation of bridge sampling to GLMMs to Chapter 4.

Nested Sampling

In this Section, we describe a relatively new method for approximating $I = \int_\Theta g(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}$, called *nested sampling* proposed by Skilling (2006).

Recall the importance sampling identity

$$\begin{aligned} I &= \int_\Theta \frac{g(\boldsymbol{\theta})}{h(\boldsymbol{\theta})}h(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}, \\ &= \int_\Theta L(\boldsymbol{\theta})h(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}, \\ &= \mathrm{E}_H\left(L(\boldsymbol{\theta})\right), \end{aligned} \tag{2.28}$$

where $L(\boldsymbol{\theta}) = \frac{g(\boldsymbol{\theta})}{h(\boldsymbol{\theta})} > 0$. Let $x = \Psi(l) = P(L(\boldsymbol{\theta}) > l)$ so that $\Psi()$ is the survival function of the univariate random variable, $L(\boldsymbol{\theta})$, when $\boldsymbol{\theta} \sim H$. Now $\Psi^{-1}(x) = \sup\{l : \Psi(l) > x\}$, i.e. the $(1-x)$th quantile of the distribution of $L(\boldsymbol{\theta})$. Suppose $x \sim \mathrm{U}[0,1]$, then $\Psi^{-1}(x)$ is a random variable. We now find the distribution of this random variable. The Jacobian of the transformation is

$$\left|\frac{\mathrm{d}x}{\mathrm{d}l}\right| = \left|\frac{\mathrm{d}\Psi(l)}{\mathrm{d}l}\right| = \left|\frac{\mathrm{d}(1 - F_{L(\boldsymbol{\theta})}(l))}{\mathrm{d}l}\right| = f_{L(\boldsymbol{\theta})}(l),$$

where $F_{L(\boldsymbol{\theta})}(l)$ and $f_{L(\boldsymbol{\theta})}(l)$ are the distribution function and pdf of the random variable $L(\boldsymbol{\theta})$, respectively, so that $F_{L(\boldsymbol{\theta})}(l) = 1 - \Psi(l)$. Then the pdf, $f_{\Psi^{-1}(x)}(z)$, of the random variable $\Psi^{-1}(x)$ is given by

$$f_{\Psi^{-1}(x)}(z) = f_{L(\boldsymbol{\theta})}(z),$$

for $z > 0$. Therefore, $\Psi^{-1}(x)$ has the same distribution as $L(\boldsymbol{\theta})$ and

$$\begin{aligned} I &= \mathrm{E}_H\left(L(\boldsymbol{\theta})\right), \\ &= \mathrm{E}_{\mathrm{U}[0,1]}\left(\Psi^{-1}(x)\right), \\ &= \int_0^1 \Psi^{-1}(x)\mathrm{d}x. \end{aligned} \tag{2.29}$$

Since $\frac{\mathrm{d}}{\mathrm{d}x}\Psi^{-1}(x) = -\frac{1}{f_{L(\boldsymbol{\theta})}(\Psi^{-1}(x))} < 0$, $\Psi^{-1}(x)$ is a decreasing function.

The one-dimensional integral (2.29) can be approximated by a quadrature method as

$$\hat{I}_{NS} = \sum_{i=1}^{m}(x_{i-1} - x_i)\Psi^{-1}(x_i), \tag{2.30}$$

where $0 < x_m < x_{m-1} < ... < x_1 < x_0 = 1$. The problem now is how to evaluate the function $\Psi^{-1}()$. Skilling (2006) proposes a method which does not require direct evaluation of $\Psi^{-1}()$. Suppose that $x_1, ..., x_m$ are randomly generated as

$$x_i = \prod_{k=1}^{i} t_k,$$

where $t_k = \max_{r=1,...,N}\{u_{r,k}\}$ and $u_{1,k}, ..., u_{N,k} \overset{\text{iid}}{\sim} \mathrm{U}[0,1]$. Then it can be shown that $\mathrm{E}(x_{i-1} - x_i) \approx e^{-(i-1)/N} - e^{-i/N}$. Therefore we can update the nested sampling approximation to

$$\hat{I}_{NS} = \sum_{i=1}^{m}(e^{-(i-1)/N} - e^{-i/N})\Psi^{-1}(x_i), \tag{2.31}$$

although this still requires apparent evaluation of $\Psi^{-1}(x_i)$. However, if $\boldsymbol{\theta}_{i1}, ..., \boldsymbol{\theta}_{iN}$ are generated from the distribution $H|L(\boldsymbol{\theta}) > L(\boldsymbol{\theta}_i)$ and we let $\boldsymbol{\theta}_i \in \{\boldsymbol{\theta}_{i1}, ..., \boldsymbol{\theta}_{iN}\}$ be such that $L(\boldsymbol{\theta}) = \min\{L(\boldsymbol{\theta}_{i1}), ..., L(\boldsymbol{\theta}_{iN})\}$, then $\Psi^{-1}(x_i)$ has the same distribution as $L(\boldsymbol{\theta}_i)$ (see Evans (2007)). We then replace $\Psi^{-1}(x_i)$ in (2.31) by $L(\boldsymbol{\theta}_i)$. The nested sampling approximation is then

$$\hat{I}_{NS} = \sum_{i=1}^{m}(e^{-(i-1)/N} - e^{-i/N})L(\boldsymbol{\theta}_i). \tag{2.32}$$

Evans (2007) shows that $\sum_{i=1}^{m}(e^{-(i-1)/N} - e^{-i/N})g(x_i) \to \int_0^1 g(x)\mathrm{d}x$ as $N \to \infty$ and $m/N \to \infty$ in probability for x_i as generated above and for any continuous function $g : [0,1] \to \mathbb{R}$. Chopin and Robert (2009) provide a central limit theorem result for nested sampling where the variance of the approximation is $O(N^{-\frac{1}{2}})$.

The following algorithm can be used to find $\hat{I}_{NS}$, the nested sampling approximation to I.

1. Generate a sample $\Theta_1 = \{\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_N\}$ from H. Set $i = 1$, and $\hat{I}_{NS} = 0$.

2. Let $\boldsymbol{\theta}_i \in \Theta_i$ be such that $L(\boldsymbol{\theta}_i) = \min\{L(\boldsymbol{\theta}_1), ..., L(\boldsymbol{\theta}_N)\}$ and set $\Theta_i^{(S)} = \Theta_i \backslash \boldsymbol{\theta}_i$.

3. Generate $\boldsymbol{\theta}^*$ from $H|L(\boldsymbol{\theta}) > L(\boldsymbol{\theta}_i)$, and set $\Theta_{i+1} = \Theta_i^{(S)} \cap \boldsymbol{\theta}^*$ and let

$$\hat{I}_{NS} = \hat{I}_{NS} + (e^{-(i-1)/N} - e^{-i/N})L(\boldsymbol{\theta}_i).$$

   Put $i = i + 1$.

4. Repeat steps 2. to 3. until $\hat{I}_{NS}$ has converged.

Nested sampling was originally proposed by Skilling (2006) to approximate the marginal likelihood where H is the prior distribution and $L(\boldsymbol{\theta})$ is the likelihood function. Evans (2007) suggested the use of nested sampling for general integration problems and that is what we have described above. The description above can be seen as a special case of *nested importance sampling* as proposed by Chopin and Robert (2009). We will discuss this extension and a further extension of nested sampling in Chapter 4.

2.2.6 Markov Chain Monte Carlo Model Determination

In Section 2.2.5, we describe several methods for approximating the marginal likelihood, $f_m(\mathbf{y})$, of each model $m \in M$, with the objective of evaluating the posterior model probability, $f(m|\mathbf{y})$. The general approach of evaluating the marginal likelihood for each model and, in turn, the posterior model probability, is referred to as the *marginal likelihood approach* by Chen et al. (2000, pg. 237). If the number of models, $|M|$, is large, approximating each marginal likelihood to the sufficient level of accuracy required may become impractical. In fact, even when $f_m(\mathbf{y})$ can be evaluated exactly, as is the case for linear models with the conjugate normal-inverse-gamma prior, it can be impractical to evaluate $f_m(\mathbf{y})$ for every model, if $|M|$ is very large.

Reversible Jump MCMC

An alternative to the marginal likelihood approach, is to generate a sample from the parameter space

$$\Theta = \bigcup_{m \in M} \{m\} \times \Theta_m,$$

of the posterior distribution, $\boldsymbol{\theta}_m, m|\mathbf{y}$, of the encompassing model as described in Section 1.1.3, using MCMC methods. We will then have a sequence $\{(m^{(1)}, \boldsymbol{\theta}^{(1)}_{m^{(1)}}), ..., (m^{(n)}, \boldsymbol{\theta}^{(n)}_{m^{(n)}})\}$ as the MCMC sample. The posterior model probability of model $m \in M$ is then approximated by $\frac{1}{n}\sum_{k=1}^{n} I(m^{(k)} = m)$, i.e. the proportion of occurrences of model m in the MCMC sample. We can also use the MCMC sample for posterior inference conditional on model $m \in M$, by only selecting, as our posterior sample, the values in the MCMC sample where $m^{(k)} = m$.

The parameter $\boldsymbol{\theta}_m$ has different interpretations for different m. We need to update $\boldsymbol{\theta}_m$ simultaneously with m, and, therefore, Gibbs sampling is not an appropriate method for generating a sample from $\boldsymbol{\theta}_m, m|\mathbf{y}$.

Instead, we focus on a more general Metropolis-Hastings algorithm. Suppose the current value in the chain is $(m^{(i)}, \boldsymbol{\theta}^{(i)}_{m^{(i)}})$, then a proposal, $(m^*, \boldsymbol{\theta}^*_{m^*})$, is made from the proposal distribution, $Q|(m^{(i)}, \boldsymbol{\theta}^{(i)}_{m^{(i)}})$, given the current value, $(m^{(i)}, \boldsymbol{\theta}^{(i)}_{m^{(i)}})$. It is generally more convenient to propose a model, m^*, and then propose model parameters, $\boldsymbol{\theta}^*_{m^*}$, conditional on m^*. Therefore, the pdf of $Q|(m^{(i)}, \boldsymbol{\theta}^{(i)}_{m^{(i)}})$ can be decomposed as

$$q\left((m^{(i)}, \boldsymbol{\theta}^{(i)}_{m^{(i)}}), (m, \boldsymbol{\theta}_m)\right) = q_m\left((m^{(i)}, \boldsymbol{\theta}^{(i)}_{m^{(i)}}), m\right) q_{\boldsymbol{\theta}_m}\left((m^{(i)}, \boldsymbol{\theta}^{(i)}_{m^{(i)}}), (m, \boldsymbol{\theta}_m)\,|m\right).$$

As with all Metropolis-Hastings algorithms, the choice of proposal is crucial for effective performance in practice. An obvious starting point is to make the proposal distribution independent of the current values of the chain yielding an *independence sampler*. Therefore, $q((m^{(i)}, \boldsymbol{\theta}^{(i)}_{m^(i)}), (m, \boldsymbol{\theta}_m)) = s(m, \boldsymbol{\theta}_m) = s_m(m) s_{\boldsymbol{\theta}_m}(\boldsymbol{\theta}_m|m)$. Similar to the within-model independence sampler discussed in Section 2.2.4, the performance is dependent on how well $s_m(m)$ and $s_{\boldsymbol{\theta}_m}(\boldsymbol{\theta}_m|m)$ 'mimic' $f(m|\mathbf{y})$ and $f_m(\boldsymbol{\theta}_m|\mathbf{y})$, respectively. An effective independence sampler will, therefore, need considerable information about each posterior distribution of $\boldsymbol{\theta}_m$, and can be seen as an alternative to the marginal likelihood approach using the methods described in Section 2.2.5. O'Hagan and Forster (2004, pg. 298) point out that if $|M|$ is small to moderate, then the correspondence between $s_m(m)$ and $f(m|\mathbf{y})$ is not so crucial and $s_m(m) \propto 1$ will often suffice.

Green (1995) proposes a Metropolis-Hastings algorithm, called the *reversible jump algorithm*, for generating from $\boldsymbol{\theta}_m, m|\mathbf{y}$, where the proposals are allowed to depend on the current value of the chain, $(m^{(i)}, \boldsymbol{\theta}^{(i)}_{m^{(i)}})$. For moves from $m^{(i)}$ to m^* where the current model parameters are $\boldsymbol{\theta}^{(i)}_{m^{(i)}}$, we specify a proposal distribution with pdf $q(\mathbf{v}|\boldsymbol{\theta}^{(i)}_{m^{(i)}}, m^{(i)}, m^*)$. In the algorithm we generate $\mathbf{v}$ from this distribution. The proposal is then a function of $\mathbf{v}$ and $\boldsymbol{\theta}^{(i)}_{m^{(i)}}$. The reversible jump algorithm is as follows:

1. Let the current values of the chain be $(m^{(i)}, \boldsymbol{\theta}^{(i)}_{m^{(i)}})$ where the dimension of $\boldsymbol{\theta}^{(i)}_{m^{(i)}}$ is $k_{m^{(i)}}$.
2. Propose a new model, m^*, with probability $h(m^{(i)}, m^*)$.
3. Generate $\mathbf{v}$ from the proposal distribution with pdf $q(\mathbf{v}|\boldsymbol{\theta}^{(i)}_{m^{(i)}}, m^{(i)}, m^*)$.
4. Set $(\boldsymbol{\theta}^*_{m^*}, \mathbf{v}^*) = g_{m^{(i)},m^*}(\boldsymbol{\theta}^{(i)}_{m^{(i)}}, \mathbf{v})$, where $g_{m^{(i)},m^*}$ is a deterministic one-to-one function and specified so that $k_{m^{(i)}} + \dim(\mathbf{v}) = k_{m^*} + \dim(\mathbf{v}^*)$. Note that $g_{m^{(i)},m^*} = g^{-1}_{m^*,m^{(i)}}$.
5. Accept the proposed move from $m^{(i)}$ to m^* with probability

$$\begin{aligned}\alpha\left[(m^{(i)}, \boldsymbol{\theta}^{(i)}_{m^{(i)}}), (m^*, \boldsymbol{\theta}^*_{m^*})\right] &= \min\left[1, \frac{f_{m^*}(\mathbf{y}|\boldsymbol{\theta}^*_{m^*}) f_{m^*}(\boldsymbol{\theta}^*_{m^*}) f(m^*)}{f_{m^{(i)}}(\mathbf{y}|\boldsymbol{\theta}^{(i)}_{m^{(i)}}) f_{m^{(i)}}(\boldsymbol{\theta}^{(i)}_{m^{(i)}}) f(m^{(i)})}\right. \\ &\quad \times \frac{h(m^*, m^{(i)}) q(\mathbf{v}^*|\boldsymbol{\theta}^*_{m^*}, m^*, m^{(i)})}{h(m^{(i)}, m^*) q(\mathbf{v}|\boldsymbol{\theta}^*_{m^{(i)}}, m^{(i)}, m^*)} \\ &\quad \left.\left|\frac{\partial g_{m^{(i)},m^*}(\boldsymbol{\theta}^{(i)}_{m^{(i)}}, \mathbf{v})}{\partial(\boldsymbol{\theta}^{(i)}_{m^{(i)}}, \mathbf{v})}\right|\right].\end{aligned}$$

6. Repeat steps 1. to 5.

The reversible jump algorithm has the independence sampler algorithm as a special case when $g_{m^{(i)},m^*}(\boldsymbol{\theta}^{(i)}_{m^{(i)}}, \mathbf{v}) = (\mathbf{v}^*, \boldsymbol{\theta}^*_{m^*})$, where $\dim(\mathbf{v}) = k_{m^*}$ and $\dim(\mathbf{v}^*) = k_{m^{(i)}}$.

Han and Carlin (2001) and Dellaportas et al. (2002) review other MCMC methods for generating from $\boldsymbol{\theta}_m, m|\mathbf{y}$ including the *product space search* (Carlin and Chib (1995)) and the *Metropolised product space search*. The product space search is a Gibbs sampler for generating from $\boldsymbol{\theta}_m, m|\mathbf{y}$. However, for the Gibbs sampler to work pseudoprior distributions need be specified and, in addition, $|M| - 1$ of these are generated from at each iteration of the algorithm. Both Han and Carlin (2001) and Dellaportas et al. (2002) give this reason for the impracticality of the product space search when $|M|$ is large. The Metropolised product space search proposed by Dellaportas et al. (2002) uses a combined Gibbs and Metropolis approach so that we only need to generate from one pseudoprior distribution at each iteration. It can then be shown that the Metropolised product space search is equivalent to the independence sampler described above.

Variable Selection

Many statistical models can be represented by $\boldsymbol{\gamma} \in \{0,1\}^k$ where if $\gamma_j = 1$, then θ_j is present in the model, and if $\gamma_j = 0$, then θ_j is not present, for $j = 1, ..., k$. Gibbs sampling can be used to generate a posterior sample of $(\boldsymbol{\theta}, \boldsymbol{\gamma})$ with pdf

$$f(\boldsymbol{\theta}, \boldsymbol{\gamma}|\mathbf{y}) \propto f(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\gamma}) f(\boldsymbol{\theta}|\boldsymbol{\gamma}) f(\boldsymbol{\gamma}).$$

Of course, since $\boldsymbol{\gamma}$ is a set of additional parameters, we need to specify an appropriate prior distribution. Let $\boldsymbol{\gamma}_{\setminus j}$ be $\boldsymbol{\gamma}$ with the jth element, γ_j, removed. It may make sense to make the prior distribution of γ_j independent of $\boldsymbol{\gamma}_{\setminus j}$, i.e. $f(\gamma_j|\boldsymbol{\gamma}_{\setminus j}) = f(\gamma_j)$. However, in hierarchical models the prior distribution for γ_j may depend on $\boldsymbol{\gamma}_{\setminus j}$.

Let $\boldsymbol{\theta}$ be partitioned as $\boldsymbol{\theta} = (\boldsymbol{\theta}_{\boldsymbol{\gamma}}, \boldsymbol{\theta}_{\setminus \boldsymbol{\gamma}})^T$ where $\boldsymbol{\theta}_{\boldsymbol{\gamma}}$ corresponds to those elements of $\boldsymbol{\theta}$ where $\gamma_j = 1$ and $\boldsymbol{\theta}_{\setminus \boldsymbol{\gamma}}$ corresponds to those elements where $\gamma_j = 0$. The pdf of the prior distribution of $\boldsymbol{\theta}|\boldsymbol{\gamma}$ can then be further decomposed as

$$f(\boldsymbol{\theta}|\boldsymbol{\gamma}) = f(\boldsymbol{\theta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma}) f(\boldsymbol{\theta}_{\setminus \boldsymbol{\gamma}}|\boldsymbol{\theta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}).$$

For Gibbs sampling, the full conditional posterior distributions are given by Dellaportas et al. (2002) as

$$f(\boldsymbol{\theta}_{\boldsymbol{\gamma}}|\boldsymbol{\theta}_{\setminus \boldsymbol{\gamma}}, \boldsymbol{\gamma}, \mathbf{y}) \propto f(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\gamma}) f(\boldsymbol{\theta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma}) f(\boldsymbol{\theta}_{\setminus \boldsymbol{\gamma}}|\boldsymbol{\theta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}) \tag{2.33}$$

$$f(\boldsymbol{\theta}_{\setminus \boldsymbol{\gamma}}|\boldsymbol{\theta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}, \mathbf{y}) \propto f(\boldsymbol{\theta}_{\setminus \boldsymbol{\gamma}}|\boldsymbol{\theta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}) \tag{2.34}$$

and

$$\frac{f(\gamma_j = 1|\boldsymbol{\gamma}_{\setminus j}, \boldsymbol{\theta}, \mathbf{y})}{f(\gamma_j = 0|\boldsymbol{\gamma}_{\setminus j}, \boldsymbol{\theta}, \mathbf{y})} = \frac{f(\mathbf{y}|\boldsymbol{\theta}, \gamma_j = 1, \boldsymbol{\gamma}_{\setminus j})}{f(\mathbf{y}|\boldsymbol{\theta}, \gamma_j = 0, \boldsymbol{\gamma}_{\setminus j})} \frac{f(\boldsymbol{\theta}|\gamma_j = 1, \boldsymbol{\gamma}_{\setminus j})}{f(\boldsymbol{\theta}|\gamma_j = 0, \boldsymbol{\gamma}_{\setminus j})} \frac{f(\gamma_j = 1, \boldsymbol{\gamma}_{\setminus j})}{f(\gamma_j = 0, \boldsymbol{\gamma}_{\setminus j})}. \tag{2.35}$$

2.2.7 Applications to GLMMs and other models

In this last Section, we discuss how the general computational methods described above have been applied specifically to GLMMs in the literature.

Suppose, in a classical analysis, we required *maximum likelihood estimates* (mles) of the model parameters. The standard approach is to maximise the integrated likelihood function, $f(\mathbf{y}|\boldsymbol{\beta}, \mathbf{D}, \phi)$, to obtain the mles of $\boldsymbol{\beta}$, $\mathbf{D}$ and ϕ. To evaluate $f(\mathbf{y}|\boldsymbol{\beta}, \mathbf{D}, \phi)$ we need to perform the integration in (1.11) which is often analytically intractable. Breslow and Clayton (1993) show how to apply the Laplace method to this problem by approximating the first-stage likelihood function, $f(\mathbf{y}|\boldsymbol{\beta}, \mathbf{u}, \phi)$ by a normal pdf. This method results in two very similar methods of obtaining mles of $\boldsymbol{\beta}$, $\mathbf{D}$ and ϕ which are known as *penalised quasi-likelihood* (PQL) and *marginal quasi-likelihood* (MQL). Rue et al. (2009) use the Laplace method for approximating the posterior marginal distributions for latent Gaussian models, which can be a special case of GLMMs. Pinheiro and Chao (2006) and Joe (2008) assess the accuracy of the Laplace method, among other methods, for obtaining the mles of $\boldsymbol{\beta}$ and $\mathbf{D}$ in GLMMs, where the response is either Bernoulli or Poisson distributed, i.e. $\phi = 1$. Joe (2008) recommends the use of the Laplace method "for quick comparisons of competing mixed models".

Due to the possible conditional conjugacies and conditional independences that can arise in GLMMs, Gibbs sampling is a popular method for generating a posterior sample.

Suppose we specify an inverse-Wishart prior distribution, $\text{IW}(\rho, \mathbf{R})$, with ρ degrees of freedom and scale matrix, $\mathbf{R}$, for the variance components matrix, $\mathbf{D}$. Also suppose that the prior distribution for $\boldsymbol{\beta}$ is independent of $\mathbf{D}$, so that $f(\boldsymbol{\beta}|\mathbf{D}, \phi) = f(\boldsymbol{\beta}|\phi)$ in the decomposition (1.12). Note that ρ and $\mathbf{R}$ may depend on ϕ. Then the pdf of the full conditional distribution of $\mathbf{D}$ reduces to

$$\begin{aligned} f(\mathbf{D}|\mathbf{y}, \boldsymbol{\beta}, \mathbf{u}, \phi) &\propto f(\mathbf{u}|\mathbf{D}) f(\mathbf{D}|\phi), \qquad (2.36) \\ &\propto |\mathbf{D}|^{-\frac{\rho+q+1}{2}} \exp\left(-\frac{1}{2}\text{tr}(\mathbf{R}\mathbf{D}^{-1})\right) \prod_{i=1}^{G} |\mathbf{D}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\mathbf{u}_i^T \mathbf{D}^{-1} \mathbf{u}_i\right), \\ &= |\mathbf{D}|^{-\frac{\rho+G+q+1}{2}} \exp\left(-\frac{1}{2}\text{tr}\left(\left(\sum_{i=1}^{G} \mathbf{u}_i \mathbf{u}_i^T + \mathbf{R}\right)\mathbf{D}^{-1}\right)\right). \end{aligned}$$

Therefore, the full conditional distribution of $\mathbf{D}$ is the inverse-Wishart distribution with $\rho + G$ degrees of freedom and scale matrix, $\sum_{i=1}^{G} \mathbf{u}_i \mathbf{u}_i^T + \mathbf{R}$, i.e. the inverse-Wishart prior distribution for $\mathbf{D}$ is the conditional conjugate prior distribution. It is easy to generate from the inverse-Wishart distribution so in an iteration of a Gibbs sampler it is easy to update $\mathbf{D}$. Note that if the prior distribution for $\mathbf{D}$ is not the inverse-Wishart or the prior distribution for $\boldsymbol{\beta}$ is not independent of $\mathbf{D}$ then this conditional conjugacy will not exist.

In general, similar conditional independences to (2.36) arise for $\boldsymbol{\beta}$, $\mathbf{u}$ and $\mathbf{D}$, i.e.

$$f(\boldsymbol{\beta}|\mathbf{y}, \mathbf{u}, \mathbf{D}, \phi) \propto f(\mathbf{y}|\boldsymbol{\beta}, \mathbf{u}, \phi) f(\boldsymbol{\beta}|\mathbf{D}, \phi), \qquad (2.37)$$

$$f(\mathbf{u}_i|\mathbf{y}_i, \boldsymbol{\beta}, \mathbf{D}, \phi) \propto f(\mathbf{y}_i|\boldsymbol{\beta}, \mathbf{u}_i, \phi) f(\mathbf{u}_i|\mathbf{D}), \qquad (2.38)$$

$$f(\mathbf{D}|\mathbf{y}, \boldsymbol{\beta}, \mathbf{u}, \phi) \propto f(\boldsymbol{\beta}|\mathbf{D}, \phi) f(\mathbf{u}|\mathbf{D}) f(\mathbf{D}|\phi). \qquad (2.39)$$

Zeger and Karim (1991) describe a Gibbs sampling algorithm for generating a posterior sample for a GLMM, where $\phi = 1$. Their algorithm takes advantage of the conditional conjugacy arising from using an inverse-Wishart prior distribution for $\mathbf{D}$ and $f(\boldsymbol{\beta}|\mathbf{D}) = f(\boldsymbol{\beta})$. To update $\boldsymbol{\beta}$ and $\mathbf{u}_i$, Zeger and Karim (1991) propose the use of rejection sampling. The algorithm of Zeger and Karim (1991) can be made more efficient and extended to cases where ϕ is unknown and where the general conditional independences of (2.37), (2.38) and (2.39) are present by using ARS or ARMS to update the model parameters, $\boldsymbol{\beta}$, $\mathbf{u}$, $\mathbf{D}$ and ϕ. This is the general method of BUGS and JAGS. Zhao et al. (2006) recommend the use of a BUGS derivative known as WinBUGS (Lunn et al. (2000)) to generate posterior samples from GLMMs as it "performs excellently among various off-the-shelf competitors". For the remainder of this thesis, we use WinBUGS to generate any posterior samples from GLMMs that we may need. WinBUGS can be called remotely within the statistical software package R using the package R2WinBUGS (Sturtz et al. (2005)). Note that BUGS and JAGS require the prior distributions of the model parameters to be proper.

Sinharay and Stern (2005) assessed several methods for approximating the Bayes factors between GLMMs using an empirical study. They assessed the methods of importance sampling, candidate's method and bridge sampling to approximate the Bayes factor by the marginal likelihood approach. They also assessed the reversible jump algorithm to approximate the Bayes factor. Sinharay and Stern (2005) concluded that bridge sampling provided approximations to the Bayes factor with the smallest standard deviation. The reversible jump algorithm provided approximations with the largest standard deviation.

DiCiccio et al. (1997) also assessed several methods for approximating marginal likelihoods for more general applications than Sinharay and Stern (2005). DiCiccio et al. (1997) report that, again, bridge sampling performs well and "provides substantial improvement" on the other methods.

George and McCulloch (1993) applied variable selection to linear models in an algorithm known as *Stochastic Search Variable Selection* (SSVS). In this algorithm, the maximal model is assumed throughout and the regression parameters are constrained to be close to zero when $\gamma_j = 0$. Therefore, we can assume that $f(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\gamma}) = f(\mathbf{y}|\boldsymbol{\beta})$ and this removes the dependence of the full conditional posterior distribution of γ_j on $\mathbf{y}$.

Cai and Dunson (2006) propose a stochastic search variable selection algorithm for model determination amongst GLMMs. The integrated likelihood is approximated by using a second-order Taylor series approximation to the first-stage likelihood as opposed to the Laplace method which uses a second-order Taylor series approximation to the first-stage log-likelihood. The variance components matrix is decomposed as

$$\mathbf{D} = \boldsymbol{\Lambda}\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T\boldsymbol{\Lambda},$$

where $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, ..., \lambda_q)$, $\lambda_k \geq 0$ for $k = 1, ..., q$ and $\boldsymbol{\Gamma}$ is a lower triangular matrix, i.e.

$$\boldsymbol{\Gamma} = \begin{pmatrix} 1 & & & \\ \gamma_{21} & 1 & & \\ \vdots & \vdots & \ddots & \\ \gamma_{q1} & \gamma_{q2} & \cdots & 1 \end{pmatrix},$$

where $\gamma_{kl} \in \mathbb{R}$. If $\lambda_k > 0$, then $\mathbf{D}$ is positive-definite and if $\lambda_k = 0$, then the kth row and column of $\mathbf{D}$ are zero, and the submatrix of $\mathbf{D}$ with the kth row and column removed is also positive-definite. This decomposition allows the SSVS algorithm to be used. The parameters are updated in the algorithm using ARMS. We will further discuss the method of Cai and Dunson (2006) for approximating the integrated likelihood in Chapter 5.

We can draw several conclusions from Section 2.2 on computation for GLMMs. First, Gibbs sampling implemented using the statistical software package BUGS is a convenient method for generating a posterior sample. With regards to approximating the posterior model probabilities, the marginal likelihood approach using bridge sampling to approximate the marginal likelihood is reported to perform well (Sinharay and Stern (2005) and DiCiccio et al. (1997)). However, as described in Section 2.2.6, the marginal likelihood approach can prove inefficient if the number of models, $|M|$, is large. Cai and Dunson (2006) propose an MCMC model determination method that means we do not need to approximate the marginal likelihood for every model in M.

There do exist Bayesian approaches to model determination that do not attempt to evaluate the posterior model probabilities of the models in M. One such approach is that of *information criteria* which are a measure of goodness of fit adjusted by model complexity. Examples of information criteria are the Bayesian Information Criterion (BIC) and Deviance Information Criterion (DIC). In both cases, 'better' models have smaller values of the information criterion. Both criteria listed above attempt to trade a measure of goodness of fit against model complexity. BIC does this by adding $\log n$ times the number of parameters in the model. A disadvantage of this, pointed out by Spiegelhalter et al. (2002), is that the number of parameters in a hierarchical model is not a well-defined quantity and, hence, not a good measure of model complexity. Spiegelhalter et al. (2002) proposed the DIC as an alternative that does not use the number of parameters. For a model $m \in M$, the DIC is defined as

$$\mathrm{DIC}_m = 2\mathrm{E}(D_m(\boldsymbol{\theta}_m)|\mathbf{y}) - D_m(\mathrm{E}(\boldsymbol{\theta}_m|\mathbf{y})),$$

where

$$D_m(\boldsymbol{\theta}_m) = -2\log f_m(\mathbf{y}|\boldsymbol{\theta}_m) + 2\log f(\mathbf{y}),$$

is known as the *Bayesian deviance*. The quantity $f(\mathbf{y})$ is the same for each model and, therefore, does not affect the relative values of DIC_m and need not be evaluated. The DIC_m can be approximated by using an MCMC method to generate a sample from the posterior distribution of model $m \in M$. See Spiegelhalter et al. (2002) for more details. A disadvantage of both information criteria listed above and all information criteria, in general, is that they need to be evaluated for all models in M, similar to the marginal likelihood approach. In the case of DIC, where a posterior sample is required, this will be a very computationally intensive approach.

2.3 Default Priors applied to GLMMs and other models

In this Section, we discuss the subject of default priors. Since our goal is to develop default priors for GLMMs, we focus on default priors in the literature for this class of models. However, we review some important general default prior approaches. This review is not meant to be exhaustive and for a larger review, see Kass and Wasserman (1996). We then discuss how these approaches have been applied to GLMMs, or important special cases of GLMMs.

Kass and Wasserman (1996) provide two interpretations of default priors. "The first interpretation asserts that reference [default] priors are formal representations of ignorance. The second asserts that there is no objective, unique prior that represents ignorance; instead, reference [default] priors are chosen by public agreement, much like units of length and weight. In this interpretation, reference [default] priors are akin to a default option in a computer package". Kass and Wasserman (1996) suggest that, at that point, the second interpretation is more popular since no unique default prior can exist and research is focused on developing default priors which are useful in practice. Subsequently, we adopt this second interpretation of default priors in this thesis.

2.3.1 Jeffreys Prior

In Section 1.1.4 we discussed the problem of Lindley's paradox when using a non-informative uniform distribution for $\boldsymbol{\theta}$, i.e. $f(\boldsymbol{\theta}) \propto 1$. Another problem in using the uniform prior is that if we are completely ignorant about $\boldsymbol{\theta}$ then we should be completely ignorant about a transformation, $\boldsymbol{\phi} = g(\boldsymbol{\theta})$, of $\boldsymbol{\theta}$. However,

$$f(\boldsymbol{\phi}) = f(g^{-1}(\boldsymbol{\phi})) \left| \frac{dg^{-1}(\boldsymbol{\phi})}{d\boldsymbol{\phi}} \right| \propto \left| \frac{dg^{-1}(\boldsymbol{\phi})}{d\boldsymbol{\phi}} \right|,$$

and, in general, this may not be proportional to one, resulting in a non-uniform prior distribution. Therefore, the prior distribution is not invariant to transformations.

Define the *Fisher information*, $\mathcal{I}_{\boldsymbol{\theta}}$, as

$$\mathcal{I}_{\boldsymbol{\theta}} = -\mathrm{E}\left[\frac{\partial^2 \log f(\mathbf{y}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right].$$

Jeffreys prior is defined such that

$$f(\boldsymbol{\theta}) \propto |\mathcal{I}_{\boldsymbol{\theta}}|^{\frac{1}{2}},$$

and it can be shown that this rule is invariant to the transformation, $\boldsymbol{\phi} = g(\boldsymbol{\theta})$.

Examples

1. Suppose $y_1, ..., y_n \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2)$, where σ^2 is known. The Fisher information is then $\mathcal{I}_\mu = \frac{n}{\sigma^2}$, and Jeffreys prior is the improper uniform distribution with $f(\mu) \propto 1$.

2. Suppose $y_1, ..., y_n \overset{\text{iid}}{\sim} \text{N}(\mu, \sigma^2)$, where μ is known. The Fisher information is then $\mathcal{I}_{\sigma^2} = \frac{n}{2\sigma^4}$, and Jeffreys prior has $f(\sigma^2) \propto \frac{1}{\sigma^2}$.

3. Suppose $y_1, ..., y_n \overset{\text{iid}}{\sim} \text{N}(\mu, \sigma^2)$, where μ and σ^2 are unknown. The Fisher information is then

$$\mathcal{I}_{\mu,\sigma^2} = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix},$$

and Jeffreys prior has $f(\mu, \sigma^2) \propto \left(\frac{1}{\sigma^2}\right)^{\frac{3}{2}}$. ■

As shown in the above examples, Jeffreys prior can be improper and care must be taken to ensure that the resulting posterior distribution is proper.

Ibrahim and Laud (1991) studied the use of Jeffreys prior applied to the regression parameters of GLMs and found conditions which ensure that the resulting posterior distribution is proper.

Natarajan and Kass (2000) proposed an approximate Jeffreys prior for the variance components, $\mathbf{d}$, of a GLMM. They give the approximate Fisher information matrix, $\hat{\mathcal{I}}_{\mathbf{d}}$, for $\mathbf{d}$ with (r, s)th element

$$\hat{\mathcal{I}}_{\mathbf{d},rs} = \sum_{i=1}^{G} \text{tr}\left(\mathbf{V}_i^{-1} \frac{\partial \mathbf{D}}{\partial d_r} \mathbf{V}_i^{-1} \frac{\partial \mathbf{D}}{\partial d_s}\right),$$

where $\mathbf{V}_i = \mathbf{D} + \left(\mathbf{Z}_i^T \mathbf{W}_i \mathbf{Z}_i\right)^{-1}$ and $\mathbf{W}_i = \text{diag}\left\{\text{var}(Y_{ij}) g'(\mu_{ij})^2\right\}^{-1}$, for $r, s = 1, ..., \frac{1}{2}q(q+1)$. The approximate expression for Fisher information can be used to define an approximate Jeffreys prior as

$$f(\mathbf{d}) \propto |\mathcal{I}_{\mathbf{d}}|^{\frac{1}{2}}.$$

This prior is improper. Natarajan and Kass (2000) tried to find conditions which ensure that the resulting posterior distribution is proper, when the approximate Jeffreys prior for $\mathbf{D}$ is used in conjunction with a uniform prior for the regression parameters, $\boldsymbol{\beta}$, but were unsuccessful due to the "complicated nature of its [the prior's] dependence on $\mathbf{D}$".

2.3.2 Unit Information Prior

Suppose $\boldsymbol{\theta} \in \mathbb{R}^k$, then the *unit information prior* is defined as the multivariate normal distribution with mean $\mathbf{m}$ and variance matrix $\boldsymbol{\Sigma}$, i.e. $\boldsymbol{\theta} \sim \text{N}(\mathbf{m}, \boldsymbol{\Sigma})$. The mean, $\mathbf{m}$, is hopefully obvious from the context of the problem. For example, in regression-type problems where $\boldsymbol{\theta}$ are the regression parameters, then $\mathbf{m} = (m, \mathbf{0})^T$, where typically $m = 0$, also. The variance matrix, $\boldsymbol{\Sigma}$, is chosen so that the prior provides the same amount of information as one observation. For independent and identically distributed responses, this is achieved by setting $\boldsymbol{\Sigma}$ equal to the inverse of the average Fisher information, i.e. $\left(\frac{1}{n}\mathcal{I}_{\boldsymbol{\theta}}\right)^{-1}$.

Smith and Spiegelhalter (1980) first suggested the use of the unit information prior for the regression parameters, $\boldsymbol{\beta}$, of the linear model where $\mathbf{m} = \mathbf{0}$ and $\boldsymbol{\Sigma} = n\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$. Smith and Spiegelhalter (1980) discuss how the elements of $\mathbf{X}^T\mathbf{X}$ are $O(n)$, so by choosing $\boldsymbol{\Sigma} =$

$n\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$ we are taking a fixed prior where the variance does not increase with increasing n. Kass and Wasserman (1995) show that if a unit information prior is assumed for $\boldsymbol{\beta}$ then the error of the BIC-type approximation to the marginal likelihood is reduced from $O(1)$ to $O(n^{-\frac{1}{2}})$.

Ntzoufras et al. (2003) extended the unit information prior to the regression parameters, $\boldsymbol{\beta}$, of a GLM. The Fisher information in this case is

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

where $\mathbf{W} = \text{diag}\{\text{var}(Y_i)g'(\mu_i)^2\}^{-1}$, $g(\mu_i) = \mathbf{x}_i^T\boldsymbol{\beta}$ is the link function and $\text{var}(Y_i)$ is a function of $\mu_i = g^{-1}(\mathbf{x}_i^T\boldsymbol{\beta})$. Obviously, $\mathcal{I}_{\boldsymbol{\beta}}$ typically depends on the unknown parameters $\boldsymbol{\beta}$. Ntzoufras et al. (2003) proposes replacing $\boldsymbol{\beta}$ in $\mathcal{I}_{\boldsymbol{\beta}}$ by its prior mean, $\mathbf{m}$. Therefore $\boldsymbol{\Sigma} = n(\mathbf{X}^T\mathbf{W}_{\mathbf{m}}\mathbf{X})^{-1}$, where $\mathbf{W}_{\mathbf{m}} = \text{diag}\{\text{var}(Y_i)g'(\mu_i)^2\}|^{-1}_{\boldsymbol{\beta}=\mathbf{m}}$. Note that due to the substitution of $\mathbf{m}$ for $\boldsymbol{\beta}$ in $\mathbf{W}$, this is only an approximate unit information prior. This approximate unit information prior for GLMs was adopted by Nott and Leonte (2004).

Pauler (1998) extended the concept of the unit information prior to the regression parameters, $\boldsymbol{\beta}$, of an LMM. The Fisher information used by Pauler (1998) is derived from the integrated likelihood, which is analytically tractable for LMMs. In this case, the effective sample size is dependent on whether β_j (the jth element of $\boldsymbol{\beta}$ for $j = 1, ..., p$) has an associated group-specific parameter. We will discuss this further when we attempt to extend the unit information prior to GLMMs in Chapter 3.

2.3.3 Uniform Shrinkage Prior

The *uniform shrinkage prior* is a prior distribution that can be applied to the variance components of mixed models. We motivate it by considering the following example from Daniels (1999). Suppose $y_i \sim \text{N}(\mu_i, \sigma^2)$, for $i = 1, ..., n$, where $\mu_i \sim \text{N}(0, \tau^2)$ and σ^2 is known. The posterior mean of $\boldsymbol{\mu} = (\mu_1, ..., \mu_n)^T$ is

$$\text{E}(\boldsymbol{\mu}|\mathbf{y}) = \frac{\tau^2}{\sigma^2 + \tau^2}\hat{\boldsymbol{\mu}} = \rho\hat{\boldsymbol{\mu}},$$

where $\hat{\boldsymbol{\mu}} = \mathbf{y}$ is the mle of $\boldsymbol{\mu}$ and $\rho = \frac{\tau^2}{\sigma^2+\tau^2} \in (0, 1)$ is the *shrinkage parameter*. The shrinkage parameter controls how much the mle of $\boldsymbol{\mu}$ is "shrunk" towards the prior mean, $\mathbf{0}$. The uniform shrinkage prior for τ^2 is induced by assuming that $\rho \sim \text{U}[0, 1]$ and then finding the probability distribution for τ^2 by transformation. It can be shown that, in this case, the uniform shrinkage prior for τ^2 has pdf $f(\tau^2) = \frac{\sigma^2}{(\sigma^2+\tau^2)^2}$. This distribution is proper and has median σ^2.

Gustafson et al. (2006) and Natarajan and Kass (2000) proposed approximate uniform shrinkage priors for the variance components matrix, $\mathbf{D}$, in GLMMs. Natarajan and Kass (2000) give the following shrinkage estimate of the ith group-specific parameter

$$\hat{\mathbf{u}}_i = \mathbf{D}\mathbf{Z}_i^T\left(\tilde{\mathbf{W}}_i^{-1} + \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^T\right)^{-1}(\tilde{\mathbf{y}}_i - \tilde{\boldsymbol{\eta}}_i), \tag{2.40}$$

where $\tilde{\mathbf{y}}_i$ is the working vector with jth element $\tilde{y}_{ij} = \tilde{\eta}_{ij} + (y_{ij} - \tilde{\eta}_{ij})g'(\mu_{ij})$, and $\tilde{\mathbf{W}}_i = \text{diag}\{\text{var}(Y_{ij})g'(\mu_{ij})^2\}^{-1}_{\mathbf{u}_i=\mathbf{0}}$, and, $\tilde{\mu}_{ij}$ and $\tilde{\eta}_{ij}$ are evaluated at $\mathbf{u}_i = \mathbf{0}$. Natarajan and Kass (2000) show that (2.40) can be written as

$$\hat{\mathbf{u}}_i = \mathbf{S}_i\mathbf{0} + (\mathbf{I} - \mathbf{S}_i)\mathbf{D}\mathbf{Z}_i^T\tilde{\mathbf{W}}_i(\tilde{\mathbf{y}}_i - \tilde{\boldsymbol{\eta}}_i),$$

where $\mathbf{S}_i = (\mathbf{D}^{-1} + \mathbf{Z}_i^T\tilde{\mathbf{W}}_i\mathbf{Z}_i)^{-1}\mathbf{Z}_i^T\tilde{\mathbf{W}}_i\mathbf{Z}_i$ takes the role of a multivariate shrinkage parameter. A component-wise uniform distribution is placed on $\mathbf{S}_i$, having first replaced $\mathbf{Z}_i^T\tilde{\mathbf{W}}_i\mathbf{Z}_i$ by its average over the G groups. From this an approximate uniform shrinkage prior distribution for $\mathbf{D}$ can be induced with pdf

$$f(\mathbf{D}) \propto \left|\mathbf{I} + \left(\frac{1}{G}\sum_{i=1}^{G}\mathbf{Z}_i^T\tilde{\mathbf{W}}_i\mathbf{Z}_i\right)\mathbf{D}\right|^{-q-1}.$$

Natarajan and Kass (2000) show that their uniform shrinkage prior for $\mathbf{D}$ is proper and find conditions which ensure that the posterior distribution is proper when this prior is used with an improper uniform prior for $\boldsymbol{\beta}$. The approximate uniform shrinkage priors depend on the unknown regression parameters, $\boldsymbol{\beta}$, through the weight matrix, $\tilde{\mathbf{W}}_i$ and both Gustafson et al. (2006) and Natarajan and Kass (2000) suggest replacing $\boldsymbol{\beta}$ by its mle from the corresponding GLM, i.e. retain the regression parameters but remove the group-specific parameters. This induces a data-dependent prior, but Natarajan and Kass (2000) point out that it is a mild form of data-dependence since $\tilde{\mathbf{W}}_i$ varies slowly with $\boldsymbol{\beta}$.

2.3.4 Intrinsic Prior

To define the intrinsic prior, we first need to define the intrinsic Bayes factor. A common, ad-hoc solution to deriving a default prior distribution is to partition the observations, $\mathbf{y}$, into a training sample and a comparison sample. The likelihood from the training sample is used in conjunction with a diffuse prior distribution to find a prior distribution using Bayes' theorem. This prior is then used with the likelihood from the comparison sample to define a marginal likelihood and Bayes factor. The resulting Bayes factor is called the *partial Bayes factor*. Obviously, this approach can be sensitive to the partitioning of $\mathbf{y}$ into the training and comparison samples.

Suppose we have two models, i.e. $M = \{1, 2\}$. Let $\mathbf{y}(l)$ and $\mathbf{y}(c)$ be the disjoint *training* and *comparison* samples, respectively, such that $\mathbf{y} = (\mathbf{y}(l), \mathbf{y}(c))^T$. Let $f_m(\mathbf{y}(l)) = \int_{\Theta_m} f_m(\mathbf{y}(l)|\boldsymbol{\theta}_m)f_m^D(\boldsymbol{\theta}_m)\mathrm{d}\boldsymbol{\theta}_m$ be the marginal likelihood for model $m \in M$ that corresponds to the likelihood, $f_m(\mathbf{y}(l)|\boldsymbol{\theta}_m)$, from $\mathbf{y}(l)$ and a diffuse prior, F_m^D, with pdf, $f_m^D(\boldsymbol{\theta}_m)$. A training sample, $\mathbf{y}(l)$, is *proper* if $0 < f_m(\mathbf{y}(l)) < \infty$, for all $m \in M$. Furthermore, a training sample, $\mathbf{y}(l)$, is *minimal* if it is proper but no subset of it is proper. Note that, in general, there will exist multiple minimal training samples and let $\mathcal{Y} = \{\mathbf{y}(1), ..., \mathbf{y}(L)\}$ denote the set of all minimal training samples.

It can be shown that the partial Bayes factor between models 1 and 2 with respect to using

the training sample, $\mathbf{y}(l)$, can be written

$$\begin{aligned} B_{21}(\mathbf{y}(l)) &= \frac{\int_{\Theta_2} f_2(\mathbf{y}(c)|\boldsymbol{\theta}_2) \frac{f_2(\mathbf{y}(l)|\boldsymbol{\theta}_2) f_2^D(\boldsymbol{\theta}_2)}{\int_{\Theta_2} f_2(\mathbf{y}(l)|\boldsymbol{\theta}_2) f_2^D(\boldsymbol{\theta}_2) \mathrm{d}\boldsymbol{\theta}_2} \mathrm{d}\boldsymbol{\theta}_2}{\int_{\Theta_1} f_1(\mathbf{y}(c)|\boldsymbol{\theta}_1) \frac{f_1(\mathbf{y}(l)|\boldsymbol{\theta}_1) f_1^D(\boldsymbol{\theta}_1)}{\int_{\Theta_1} f_1(\mathbf{y}(l)|\boldsymbol{\theta}_1) f_1^D(\boldsymbol{\theta}_1) \mathrm{d}\boldsymbol{\theta}_1} \mathrm{d}\boldsymbol{\theta}_1} \\ &= B_{21}^D B_{12}^D(\mathbf{y}(l)), \end{aligned}$$

where

$$B_{21}^D = \frac{\int_{\Theta_2} f_2(\mathbf{y}|\boldsymbol{\theta}_2) f_2^D(\boldsymbol{\theta}_2) \mathrm{d}\boldsymbol{\theta}_2}{\int_{\Theta_1} f_1(\mathbf{y}|\boldsymbol{\theta}_1) f_1^D(\boldsymbol{\theta}_1) \mathrm{d}\boldsymbol{\theta}_1}$$

is the Bayes factor between models 1 and 2 under the likelihood from the entire $\mathbf{y}$ and diffuse priors, F_1^D and F_2^D, and

$$B_{12}^D(\mathbf{y}(l)) = \frac{1}{B_{21}^D(\mathbf{y}(l))} = \frac{f_1(\mathbf{y}(l))}{f_2(\mathbf{y}(l))}$$

is the inverse of the Bayes factor between models 1 and 2 under the likelihood from the training sample, $\mathbf{y}(l)$, and diffuse priors, F_1^D and F_2^D. Berger and Pericchi (1996) define the *arithmetic intrinsic Bayes factor* (AIBF) between models 1 and 2 as the arithmetic mean of the partial Bayes factors for all minimal training samples, i.e.

$$B_{21}^{AI} = \frac{1}{L} \sum_{l=1}^{L} B_{21}(\mathbf{y}(l)) = \frac{B_{21}^D}{L} \sum_{l=1}^{L} B_{12}^D(\mathbf{y}(l)). \tag{2.41}$$

Likewise, the *geometric intrinsic Bayes factor* (GIBF) between models 1 and 2 is

$$B_{21}^{GI} = \left(\prod_{l=1}^{L} B_{21}(\mathbf{y}(l)) \right)^{\frac{1}{L}} = B_{21}^D \left(\prod_{l=1}^{L} B_{12}^D(\mathbf{y}(l)) \right)^{\frac{1}{L}}.$$

In general, $B_{21}^{AI} \neq \frac{1}{B_{12}^{AI}}$, where B_{12}^{AI} is found by using (2.41) with the indices reversed. If model 1 is nested within model 2, then we can set $B_{12}^{AI} \equiv \frac{1}{B_{21}^{AI}}$.

Berger and Pericchi (1996) define the *intrinsic priors*, F_1^I and F_2^I, to be the prior distributions that, when combined with the likelihood from $\mathbf{y}$, would result in the (arithmetic or geometric) intrinsic Bayes factor. In other words,

$$B_{21}^I = \frac{\int_{\Theta_2} f_2(\mathbf{y}|\boldsymbol{\theta}_2) f_2^I(\boldsymbol{\theta}_2) \mathrm{d}\boldsymbol{\theta}_2}{\int_{\Theta_1} f_1(\mathbf{y}|\boldsymbol{\theta}_1) f_1^I(\boldsymbol{\theta}_1) \mathrm{d}\boldsymbol{\theta}_1},$$

where B_{21}^I is the (arithmetic or geometric) intrinsic Bayes factor and $f_1^I(\boldsymbol{\theta}_1)$ and $f_2^I(\boldsymbol{\theta}_2)$ are the pdfs of F_1^I and F_2^I, respectively. Common choices for F_1^D and F_2^D for deriving F_1^I and F_2^I are Jeffreys prior or the reference prior (see Section 2.3.5).

Consider the problem of model determination between the following models

1. $y_{ij} \sim \mathrm{N}(\mu_1, \sigma_1^2)$,

2. $y_{ij} \sim \mathrm{N}(\mu_2 + u_{2i}, \sigma_2^2)$, where $u_{2i} \overset{\text{iid}}{\sim} \mathrm{N}(0, \tau_2^2)$,

where $i = 1, ..., G$ and $j = 1, ..., n^*$, and, in particular, specifying a default prior distribution for the variance component, τ^2. Garcia-Donato and Sun (2007) derived two intrinsic priors for τ^2, where the choice of the prior for F_1^D and F_2^D is Jeffreys prior or the reference prior.

2.3.5 Reference Prior

Define $H(f(\boldsymbol{\theta})) = -\int_\Theta f(\boldsymbol{\theta}) \log f(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}$ to be the *entropy* of $f(\boldsymbol{\theta})$. The *expected information measure*, $I^{\boldsymbol{\theta}}(\mathbf{y})$, provided by $\mathbf{y}$ about $\boldsymbol{\theta}$ is given by Bernardo (1979) as

$$\begin{aligned} I^{\boldsymbol{\theta}}(\mathbf{y}) &= \int f(\mathbf{y}) \int_\Theta f(\boldsymbol{\theta}|\mathbf{y}) \log \frac{f(\boldsymbol{\theta}|\mathbf{y})}{f(\boldsymbol{\theta})} \mathrm{d}\boldsymbol{\theta} \mathrm{d}\mathbf{y}, \\ &= H(f(\boldsymbol{\theta})) - \int f(\mathbf{y}) H(f(\boldsymbol{\theta}|\mathbf{y})) \mathrm{d}\mathbf{y}. \end{aligned}$$

If we repeat the experiment that gave us $\mathbf{y}$, the expected information we possess about $\boldsymbol{\theta}$ would increase. Suppose F_k is the prior distribution with pdf, $f_k(\boldsymbol{\theta})$, that maximises the expected information about $\boldsymbol{\theta}$, provided by k independent replications of the experiment giving us $\{\mathbf{y}_1, ..., \mathbf{y}_k\}$. The *reference posterior distribution*, after the actual experiment giving us $\mathbf{y}$, has pdf $f_0(\boldsymbol{\theta}|\mathbf{y}) = \lim_{k\to\infty} f_k(\boldsymbol{\theta}|\mathbf{y})$, where $f_k(\boldsymbol{\theta}|\mathbf{y}) \propto f(\mathbf{y}|\boldsymbol{\theta}) f_k(\boldsymbol{\theta})$. Bernardo (1979) defines the *reference prior distribution* as having pdf $f_0(\boldsymbol{\theta})$ satisfying $f_0(\boldsymbol{\theta}|\mathbf{y}) \propto f(\mathbf{y}|\boldsymbol{\theta}) f_0(\boldsymbol{\theta})$. Note that it is not necessarily true that $f_0(\boldsymbol{\theta}) = \lim_{k\to\infty} f_k(\boldsymbol{\theta})$.

Bernardo (1979) goes on to show that, for a continuous $\boldsymbol{\theta}$ with no nuisance parameters and under appropriate conditions for asymptotic normality of the posterior distribution, the reference prior is Jeffreys prior.

Consider the following simple mixed model: $y_{ij} \sim \mathrm{N}(\mu + u_i, \sigma^2)$, where $u_1, ..., u_G \overset{\text{iid}}{\sim} \mathrm{N}(0, \tau^2)$, for $j = 1, ..., n^*$ and $i = 1, ..., G$. Berger and Bernardo (1992) derive the reference prior for the model parameters μ, σ^2 and τ^2.

2.3.6 Other Default Priors Applied to GLMMs

In this Section, we describe some of the default priors for the variance components that do not fit into the above categories.

Browne and Draper (2006) consider various diffuse prior distributions for variance components of mixed models including the $\mathrm{IG}(\epsilon, \epsilon)$ and $\mathrm{U}\left[0, \frac{1}{\epsilon}\right]$ for small $\epsilon > 0$. They evaluate the effect these prior distributions have on posterior inference. They concluded that an unbiased point estimate of the variance component can be found for all but very small values of G when used with one of the above diffuse prior distributions.

Kass and Natarajan (2006) propose a default prior for $\mathbf{D}$ which is a member of the conditionally conjugate inverse-Wishart distribution, IW$(m, \mathbf{\Phi})$. Kass and Natarajan (2006) suggest $m = q$ and $\mathbf{\Phi} = m\mathbf{R}$, where $\mathbf{R}$ is a prior "guess" at $\mathbf{D}$.

2.3.7 Conclusions

We see from the above review that the focus of default prior distributions for the parameters of a GLMM has been focused on the variance components.

Reference and intrinsic priors have been applied to linear mixed models but it is unclear how they can be generalised to GLMMs. Jeffreys prior can be applied approximately to $\mathbf{D}$ in a GLMM but it is improper and Natarajan and Kass (2000) could not show when the posterior would be proper. Some authors (Natarajan and Kass (2000) and Gustafson et al. (2006)) have had success with uniform shrinkage priors but these priors are data-dependent. However, we could use the strategy of Ntzoufras et al. (2003) and replace $\boldsymbol{\beta}$ by its prior mean as opposed to its mle, to remove the data-dependence. The inverse-Wishart prior of Kass and Natarajan (2006) has the computational advantages described in Section 2.2.7.

What is clear is that there is no agreement on default priors as discussed by Kass and Wasserman (1996) for the prior distribution for $\boldsymbol{\beta}$ or $\mathbf{D}$. In this Section, we have not reviewed any default prior options for the dispersion parameter, ϕ. In many cases, either ϕ will be known or it will be unknown but present in all of the models. For the latter case, we can specify the same prior for ϕ for each model and O'Hagan and Forster (2004, pg. 179) state that the "posterior model probabilities are typically not sensitive to this prior, and it is possible to use a limiting improper prior, if required."

Chapter 3

Default Priors for GLMMs

3.1 Introduction

In this Chapter, we propose default priors for the model parameters $\boldsymbol{\beta}$ and $\mathbf{D}$ in a GLMM that are based on a unit information concept. We discussed in Section 2.3.2 how versions of unit information priors have been applied to the regression parameters, $\boldsymbol{\beta}$, in linear models, GLMs and LMMs.

We show, in this Chapter, how a unit information prior can be applied to the regression parameters, $\boldsymbol{\beta}$, of a GLMM in two different ways depending on which form of the likelihood is considered; either the first-stage likelihood or the integrated likelihood. Pauler (1998) proposed a unit information prior distribution for $\boldsymbol{\beta}$ in LMMs based on the integrated likelihood and we show that this can be generalised to the regression parameters of a GLMM, approximately. We also propose an approximate unit information prior distribution for $\boldsymbol{\beta}$ in GLMMs based on the first-stage likelihood. Using the prior distribution based on the first-stage likelihood has a computational advantage in that the prior for $\boldsymbol{\beta}$ does not depend on $\mathbf{D}$. They also have an advantage in being more flexible for certain types of response.

Before we proceed, it is useful to discuss the general concept of unit information prior distributions. The definition of the unit information prior for model parameters, $\boldsymbol{\theta}$, is a multivariate normal distribution with mean, $\mathbf{m}$, and variance, $\boldsymbol{\Sigma}$. The variance matrix, $\boldsymbol{\Sigma}$, contains the same amount of information as one typical observation. This is appealing since the prior will never provide more information than the data.

Let $\mathcal{I}_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ denote the *Fisher information matrix* of $\boldsymbol{\theta}$ defined by

$$\mathcal{I}_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \mathrm{E}\left(\frac{\partial \log f(\mathbf{y}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \log f(\mathbf{y}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\right).$$

Under certain regularity conditions, it can be shown that

$$\mathcal{I}_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = -\mathrm{E}\left(\frac{\partial^2 \log f(\mathbf{y}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right),$$

and that

$$\mathcal{I}_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \text{var}\left(\frac{\partial \log f(\mathbf{y}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right).$$

If observations, $\mathbf{y}$, are independent and identically distributed, then the amount of Fisher information in one observation (*unit information*) is then

$$\frac{\mathcal{I}_{\boldsymbol{\theta}}(\boldsymbol{\theta})}{n},$$

where n is the sample size. It follows that $\boldsymbol{\Sigma} = n\mathcal{I}_{\boldsymbol{\theta}}(\boldsymbol{\theta})^{-1}$.

Examples

1. Suppose that $y_1, ..., y_n \overset{\text{iid}}{\sim} \text{N}(\mu, \sigma^2)$, where σ^2 is known. The Fisher information is

   $$\mathcal{I}_{\mu}(\mu) = \frac{n}{\sigma^2},$$

   and the unit information is $\frac{1}{\sigma^2}$. This example can be seen as a special case of the next example.

2. Suppose $\mathbf{y} \sim \text{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$, where σ^2 is known. The Fisher information is

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

   and the unit information is $\frac{1}{n\sigma^2}\mathbf{X}^T\mathbf{X}$.

3. Suppose that $y_1, ..., y_n \overset{\text{iid}}{\sim} \text{Poisson}(\exp(\mu))$. The Fisher information is

   $$\mathcal{I}_{\mu}(\mu) = n\exp(\mu),$$

   and the unit information is $\exp(\mu)$.■

Example 2 shows how a unit information prior can be applied to the regression parameters of a linear model, so that

$$\boldsymbol{\beta} \sim \text{N}(\mathbf{m}, n\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}). \tag{3.1}$$

Example 3 demonstrates a potential problem with the unit information prior, i.e. the Fisher information in this problem, and in general, depends on the unknown parameters, $\boldsymbol{\theta}$. That is why the Fisher information, $\mathcal{I}_{\boldsymbol{\theta}}(\boldsymbol{\theta})$, is denoted as a function of $\boldsymbol{\theta}$. Ntzoufras et al. (2003) encountered this problem with the Fisher information for the regression parameters, $\boldsymbol{\beta}$, of a GLM. Their solution is to replace $\boldsymbol{\beta}$ in $\mathcal{I}_{\boldsymbol{\beta}}(\boldsymbol{\beta})$ by its prior mean, $\mathbf{m}$. This is an approach we shall take throughout this Chapter. Therefore, in Example 3 above an approximate unit information prior distribution of μ is $\mu \sim \text{N}(m, \exp(-m))$. An alternative would be to replace $\boldsymbol{\beta}$ by its maximum likelihood estimate, $\hat{\boldsymbol{\beta}}$, but this would result in a data-dependent prior distribution which we wish to avoid.

The model in Example 3 is actually a simple case of a GLM, where the link function, $g(\mu_i)$, is the log link. For a general GLM, the Fisher information is given by

$$\mathcal{I}_{\boldsymbol{\beta}}(\boldsymbol{\beta}) = \mathbf{X}^T \mathbf{W}_{\boldsymbol{\beta},\phi} \mathbf{X},$$

where $\mathbf{W}_{\boldsymbol{\beta},\phi} = \text{diag}\left\{\text{var}(Y_i) g'(\mu_i)^2\right\}^{-1}$ is the *weight matrix*. Note that $\mathbf{W}_{\boldsymbol{\beta},\phi}$ is, typically, a function of $\boldsymbol{\beta}$ through $\text{var}(Y_i)$ and $g'(\mu_i)$. We now replace $\boldsymbol{\beta}$ in $\mathcal{I}_{\boldsymbol{\beta}}(\boldsymbol{\beta})$ by its prior mean, $\mathbf{m}$, to give

$$\mathcal{I}_{\boldsymbol{\beta}}(\mathbf{m}) = \mathbf{X}^T \mathbf{W}_{\mathbf{m},\phi} \mathbf{X},$$

where $\mathbf{W}_{\mathbf{m},\phi} = \text{diag}\left\{\text{var}(Y_i) g'(\mu_i)^2\right\}|^{-1}_{\boldsymbol{\beta}=\mathbf{m}}$. Therefore an approximate unit information prior for the regression parameters, $\boldsymbol{\beta}$, of a GLM is

$$\boldsymbol{\beta} \sim \text{N}(\mathbf{m}, n(\mathbf{X}^T \mathbf{W}_{\mathbf{m},\phi} \mathbf{X})^{-1}). \tag{3.2}$$

A linear model is actually a special case of a GLM, but $\text{var}(Y_i)$ does not depend on $\boldsymbol{\beta}$ and $g'(\mu_i) = 1$, so the weight matrix does not depend on $\boldsymbol{\beta}$.

However, for the linear model and some cases of GLMs, the Fisher information depends on the unknown dispersion parameter, ϕ, through the weight matrix, $\mathbf{W}_{\mathbf{m},\phi}$. Note that, in a linear model, the dispersion parameter $\phi = \sigma^2$. We can use two approaches:

a) replace ϕ in $\mathbf{W}_{\mathbf{m},\phi}$ by its prior mean, or

b) allow the prior distribution of $\boldsymbol{\beta}$ to be conditional on ϕ.

Both approaches are effected by the fact that the variance of the prior distribution of $\boldsymbol{\beta}$ is heavily dependent on ϕ, and in particular, the prior of ϕ. The variance of the response in a GLM is proportional to $a(\phi)$. It follows that the prior variance of $\boldsymbol{\beta}$ is proportional to $a(\phi)$. If we choose approach a) from above and set the prior mean of ϕ such that $a(\phi)$ is small but the true value of ϕ is such that $a(\phi)$ is large than the prior for $\boldsymbol{\beta}$ maybe too informative, and vice versa. Approach b) is also prone to this drawback if we use a prior distribution for ϕ which is too informative and the true value of ϕ lies in a region of this distribution that has low density. However, consider approach b) from above, and choose a diffuse prior for ϕ. The advantage of this approach is that it adapts the prior variance of $\boldsymbol{\beta}$ to the scale of the response. Also, for the linear model, a multivariate normal prior distribution for $\boldsymbol{\beta}$ that is conditional on $\phi = \sigma^2$ is the conjugate prior distribution. The drawback of option b) is that we then have to be careful on setting the hyperparameters of the prior for ϕ because of Lindley's paradox. However, if we make the assumption that ϕ is present in all of the models, then we can use a diffuse prior for ϕ. To this end, (O'Hagan and Forster 2004, pg. 179) state that "the only exception [to not using an arbitrarily diffuse prior] is for a parameter which is present in all models under consideration, and which is given the same prior under each model. The posterior model probabilities are typically not sensitive to this prior, and it is possible to use a limiting improper prior, if required". We will assume throughout, that if a dispersion parameter exists for one model $m \in M$ then it exists in all models in M. A disadvantage of this approach is that we now cannot make formal Bayesian model determination decisions about the response distribution. For example,

suppose we had responses, $y_1, ..., y_n$, that are continuous and positive. We maybe uncertain whether the response distribution is normal or gamma but we cannot, using Bayesian model determination methods, decide which one is true since the dispersion parameters under these two distributions have differing interpretations. This issue is outside the scope of this thesis.

Notice that if the dispersion parameter is unknown, then the weight matrix will always depend on ϕ. To save space, we suppress the dependence of the weight matrix on ϕ by dropping the the subscript ϕ.

We now turn our attention to mixed models. Consider the following example.

Example Suppose $y_{ij} \sim \mathrm{N}(\mu + u_i, \sigma^2)$ where $u_i \sim \mathrm{N}(0, \tau^2)$ for $j = 1, ..., n^*$ and $i = 1, ..., G$. This model is known as a *one-way random effects* model. The total sample size is $n = n^*G$. The log of the integrated likelihood is

$$\log f(\mathbf{y}|\mu, \tau^2, \sigma^2) \propto -\frac{G}{2} \log \left(\frac{\sigma^2}{n^*\tau^2 + \sigma^2} \right) - \frac{1}{2\sigma^2} \sum_{i=1}^{G} \sum_{j=1}^{n^*} (y_{ij} - \bar{y}_{i.})^2 - \frac{n^*}{2} \sum_{i=1}^{G} \frac{(\bar{y}_{i.} - \mu)^2}{n^*\tau^2 + \sigma^2}.$$

Therefore, the Fisher information for μ is

$$\mathcal{I}_\mu(\mu) = \frac{n}{n^*\tau^2 + \sigma^2}.$$

Note that $\mathcal{I}_\mu(\mu)$ is $O(G)$ not $O(n)$. In other words, the information on μ from the data is proportional to the number of groups, G, not the total sample size, n.■

This example demonstrates a further obstacle with unit information priors, i.e. the information is not always proportional to the total sample size and therefore, how is unit information defined in this case? In the example, unit information would be found by dividing $\mathcal{I}_\mu(\mu)$ by G, and the unit information prior for μ is

$$\mu \sim \mathrm{N}\left(m, \frac{n^*\tau^2 + \sigma^2}{n^*} \right).$$

In the next Section, we describe an approach proposed by Pauler (1998) which involves investigating unit information based on the integrated likelihood from an LMM. The same problem as in the above example, which is a special case of an LMM, arises. However, Pauler (1998) shows that the information on a regression parameter, β_j, for $j = 1, ..., p$, is proportional to either the total sample size or the number of groups, depending on which group-specific parameters are included. Note that the prior for μ in the above example is conditional on the variance component, τ^2.

Pauler (1998) list three minimal requirements for default priors under model uncertainty. They are:

1. The prior must be proper.

2. The prior must be located at a null value. By null value, we mean a value that we hope the data will indicate is not valid. This has an analogy with the null hypothesis in classical hypothesis testing.

3. The prior variance should represent vague information that is calibrated with, but less than, the information in the likelihood.

Unit information priors satisfy all these requirements. Since the distribution for $\boldsymbol{\theta}$ is normal with finite variance, the prior distribution is proper. We are free to choose the mean, $\mathbf{m}$, of the prior distribution, which can be set at a null value, thus satisfying the second requirement. As discussed earlier, the prior variance is calibrated with the likelihood and provides less information than that in the data.

We have control over the mean, $\mathbf{m}$, of the unit information prior. From above, $\mathbf{m}$ needs be set at a null value. Consider the vector of regression parameters, $\boldsymbol{\beta}$, where

$$\begin{aligned} \boldsymbol{\beta} &= (\beta_1, ..., \beta_p)^T \\ &= (\beta_1, \boldsymbol{\beta}_{\setminus 1})^T. \end{aligned}$$

The 1st element, β_1, with prior mean m_1, corresponds to the intercept term, whereas the remaining elements, $\boldsymbol{\beta}_{\setminus 1}$, with prior means $\mathbf{m}_{\setminus 1}$, corresponds to the explanatory variables. A null value for $\boldsymbol{\beta}_{\setminus 1}$ is $\mathbf{0}$. So we set $\mathbf{m} = (m_1, \mathbf{0})^T$. The value for m_1 is, in some cases, obvious. For example, for Bernoulli responses, $m_1 = g\left(\frac{1}{2}\right)$, corresponds to a prior mean of $\frac{1}{2}$ for the responses which is the middle of the sample space. In other cases, it is less clear what value m_1 should take. We discuss this issue on an example-by-example basis.

The unit information priors, as described so far, are only applicable to parameters in $\mathbb{R}^k$, which are suitable for regression parameters of regression-type models such as linear models and GLMs. In Section 3.3, we propose a prior for the variance components matrix, $\mathbf{D}$, that is based on a unit information concept.

In the all encompassing model, we have a choice for the prior model probability, $f(m)$, of model $m \in M$, such that $\sum_{m \in M} f(m) = 1$. A common approach, and one taken for the remainder of this thesis, is to assume that $f(m) = \frac{1}{|M|}$, i.e. a discrete uniform distribution over the prior model probabilities. However, the strategy proposed in this thesis can be used with any prior model probabilities and we actually take account of this in our presentation of a reversible jump algorithm for GLMMs in Chapter 5. An alternative approach to using a discrete uniform distribution over the prior model probabilities is given by Dellaportas et al. (2009).

3.2 Default Priors based on the Integrated Likelihood

3.2.1 Regression Parameters

Consider the LMM

$$y_{ij} \sim \mathrm{N}\left(\mathbf{x}_{ij}^T\boldsymbol{\beta} + \mathbf{z}_{ij}^T\mathbf{u}_i, \sigma^2\mathbf{I}_{n_i}\right),$$

where $\mathbf{x}_{ij}$ and $\mathbf{z}_{ij}$ are $p \times 1$ and $q \times 1$ vectors of regression and group-specific covariates, respectively. This model can be written in matrix form as

$$\mathbf{y} \sim \mathrm{N}\left(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}, \sigma^2\mathbf{I}_n\right),$$

where $\mathbf{X}$ and $\mathbf{Z}$ are as defined in Section 1.2.1. The marginal model, which gives the integrated likelihood, is

$$\mathbf{y} \sim \mathrm{N}\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n + \mathbf{Z}\mathbf{D}^*\mathbf{Z}^T\right),$$

where $\mathbf{D}^* = \mathbf{I}_G \otimes \mathbf{D}$. The log integrated likelihood is then

$$\log f(\mathbf{y}|\boldsymbol{\beta}, \mathbf{D}, \sigma^2) \propto -\frac{1}{2}\log\left|\sigma^2\mathbf{I}_n + \mathbf{Z}\mathbf{D}^*\mathbf{Z}^T\right| - \frac{1}{2}\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)^T\left(\sigma^2\mathbf{I}_n + \mathbf{Z}\mathbf{D}^*\mathbf{Z}^T\right)^{-1}\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right).$$

The Fisher information with respect to $\boldsymbol{\beta}$ is

$$\begin{aligned}
\mathcal{I}_{\boldsymbol{\beta}}(\boldsymbol{\beta}) &= \mathrm{E}\left(-\frac{\partial^2 \log f(\mathbf{y}|\boldsymbol{\beta}, \mathbf{D}, \sigma^2)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\right), \\
&= \mathbf{X}^T\left(\sigma^2\mathbf{I}_n + \mathbf{Z}\mathbf{D}^*\mathbf{Z}^T\right)^{-1}\mathbf{X}, \\
&= \sum_{i=1}^{G}\mathbf{X}_i^T\left(\sigma^2\mathbf{I}_{n_i} + \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^T\right)^{-1}\mathbf{X}_i. \qquad (3.3)
\end{aligned}$$

The Fisher information, (3.3), is a $p \times p$ matrix which does not depend on $\boldsymbol{\beta}$. The kth diagonal element, $\mathcal{I}_{\boldsymbol{\beta}}(\boldsymbol{\beta})_{kk}$, of $\mathcal{I}_{\boldsymbol{\beta}}(\boldsymbol{\beta})$ represents the amount of information provided by the data about the kth element, β_k, of $\boldsymbol{\beta}$, for $k = 1, ..., p$. It follows from (3.3) that

$$\mathcal{I}_{\boldsymbol{\beta}}(\boldsymbol{\beta})_{kk} = \sum_{i=1}^{G}\mathbf{x}_{ik}^T\left(\sigma^2\mathbf{I}_{n_i} + \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^T\right)^{-1}\mathbf{x}_{ik},$$

where $\mathbf{x}_{ik}$ is the kth column of $\mathbf{X}_i$. Note that

$$\left(\sigma^2\mathbf{I}_{n_i} + \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^T\right)^{-1} = \frac{1}{\sigma^2}\left(\mathbf{I}_{n_i} - \frac{1}{\sigma^2}\mathbf{Z}_i\mathbf{D}\left(\mathbf{I}_q + \frac{1}{\sigma^2}\mathbf{Z}_i^T\mathbf{Z}_i\mathbf{D}\right)^{-1}\mathbf{Z}_i^T\right), \qquad (3.4)$$

(see, for example, Henderson and Searle (1981)), and recall that the columns of $\mathbf{Z}_i$ are a subset of the columns of $\mathbf{X}_i$. Assume that the columns of $\mathbf{X}_i$ are orthogonal, that $\mathbf{D}$ is a diagonal matrix such that $\mathbf{D} = \mathrm{diag}\left\{\tau_1^2, ..., \tau_q^2\right\}$, and that $\mathbf{x}_{ik}^T\mathbf{x}_{ik} = O(n_i)$, for $k = 1, ..., p$. Using (3.4), it can be shown that

$$\mathcal{I}_{\boldsymbol{\beta}}(\boldsymbol{\beta})_{kk} = \frac{1}{\sigma^2}\sum_{i=1}^{G}\mathbf{x}_{ik}^T\mathbf{x}_{ik} - \sum_{l=1}^{q}\frac{\tau_l^2}{\sigma^2 + \tau_l^2\mathbf{z}_{il}^T\mathbf{z}_{il}}\left(\sum_{t=1}^{n_i}z_{ilt}x_{ikt}\right)^2, \qquad (3.5)$$

where x_{ikt} is the tth element of $\mathbf{x}_{ik}$, $\mathbf{z}_{il}$ is the lth column of $\mathbf{Z}_i$ with tth element z_{ilt}.

Suppose that β_k has an associated group-specific parameter. This means that the kth column of $\mathbf{X}_i$ is contained in $\mathbf{Z}_i$, for $i = 1, ..., G$. In this case, (3.5) reduces to

$$\begin{aligned} \mathcal{I}_{\boldsymbol{\beta}}(\boldsymbol{\beta})_{kk} &= \frac{1}{\sigma^2}\sum_{i=1}^{G}\mathbf{x}_{ik}^T\mathbf{x}_{ik} - \frac{\tau_k^2}{\sigma^2+\tau_k^2\mathbf{x}_{ik}^T\mathbf{x}_{ik}}(\mathbf{x}_{ik}^T\mathbf{x}_{ik})^2, \\ &= \sum_{i=1}^{G}\frac{\mathbf{x}_{ik}^T\mathbf{x}_{ik}}{\sigma^2+\tau_k^2\mathbf{x}_{ik}^T\mathbf{x}_{ik}}. \end{aligned} \tag{3.6}$$

Suppose, now that β_k does not have an associated group-specific parameter and, therefore, the kth column of $\mathbf{X}_i$ cannot be found in $\mathbf{Z}_i$. In this case, (3.5) reduces to

$$\mathcal{I}_{\boldsymbol{\beta}}(\boldsymbol{\beta})_{kk} = \frac{1}{\sigma^2}\sum_{i=1}^{G}\mathbf{x}_{ik}^T\mathbf{x}_{ik}. \tag{3.7}$$

By studying expressions (3.6) and (3.7), we see that if β_k has an associated group-specific parameter, then the Fisher information for β_k is $O(G)$, whereas if β_k has no associated group-specific parameter, then the Fisher information is $O(n)$. Pauler (1998) states that these results also hold for an unrestricted variance components matrix, $\mathbf{D}$, and for $\mathbf{X}_i$ with non-orthogonal columns. The amount of information in the integrated likelihood on a regression parameter, β_k, is dependent on which group-specific parameters are included in the model.

Consider a general problem with a p-dimensional $\boldsymbol{\theta}$ with Fisher information $\mathcal{I}_{\boldsymbol{\theta}}(\boldsymbol{\theta})$. Let $\boldsymbol{\Lambda} = \text{diag}\left\{\sqrt{N_k}\right\}$, where N_k is the order (n or G in the case of an LMM) of the kth diagonal element, $\mathcal{I}_{\boldsymbol{\theta}}(\boldsymbol{\theta})_{kk}$, of $\mathcal{I}_{\boldsymbol{\theta}}(\boldsymbol{\theta})$. The general unit information prior distribution is then

$$\boldsymbol{\theta} \sim \text{N}\left(\mathbf{m}, \boldsymbol{\Lambda}\mathcal{I}_{\boldsymbol{\theta}}(\mathbf{m})^{-1}\boldsymbol{\Lambda}\right).$$

Note that in this definition, we have replaced the unknown $\boldsymbol{\theta}$ in $\mathcal{I}_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ by its prior mean, $\mathbf{m}$, as proposed by Ntzoufras et al. (2003).

We can apply this more general definition of a unit information prior to the regression parameters, $\boldsymbol{\beta}$, of an LMM. So

$$\boldsymbol{\beta} \sim \text{N}\left(\mathbf{m}, \boldsymbol{\Lambda}\left(\sum_{i=1}^{G}\mathbf{X}_i^T\left(\sigma^2\mathbf{I}_{n_i} + \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^T\right)^{-1}\mathbf{X}_i\right)^{-1}\boldsymbol{\Lambda}\right), \tag{3.8}$$

where $\boldsymbol{\Lambda} = \text{diag}\left\{\sqrt{N_k}\right\}$ and

$$N_k = \begin{cases} G, & \text{if } \beta_k \text{ has an associated group-specific parameter,} \\ n, & \text{if otherwise.} \end{cases} \tag{3.9}$$

Note that the prior for $\boldsymbol{\beta}$ is conditional on the variance components matrix, $\mathbf{D}$. We consider a marginal prior distribution for $\mathbf{D}$ later. A linear model is a special case of an LMM and the prior for $\boldsymbol{\beta}$ in (3.8) reduces to that shown in (3.1) in this case, since $\mathbf{Z}_i = \mathbf{0}$ and $N_k = n$ for all k.

We now define a unit information prior for the regression parameters, $\boldsymbol{\beta}$, of a GLMM based on the integrated likelihood. First, we need to approximate the integrated likelihood so that we can find an approximate, analytic expression for the Fisher information of $\boldsymbol{\beta}$. Recall that the integrated likelihood is

$$f(\mathbf{y}|\boldsymbol{\beta}, \mathbf{D}, \phi) = \int_{\mathbb{R}^{Gq}} f(\mathbf{y}|\boldsymbol{\beta}, \mathbf{u}, \phi) f(\mathbf{u}|\mathbf{D}) \mathrm{d}\mathbf{u}. \tag{3.10}$$

We use the Laplace approximation for the integral in (3.10). We first need a quadratic approximation to the first-stage likelihood. Following McCulloch and Searle (2001, pg. 232-234), consider a 1st order Taylor series expansion of the link function, $g(y_{ij})$, evaluated at y_{ij} about the conditional mean, μ_{ij},

$$\begin{aligned} g(y_{ij}) \approx \tilde{y}_{ij} &= g(\mu_{ij}) + g'(\mu_{ij})(y_{ij} - \mu_{ij}), \\ &= \eta_{ij} + g'(\mu_{ij})(y_{ij} - \mu_{ij}). \end{aligned}$$

Let $\tilde{y}_{ij}$ be the elements of $\tilde{\mathbf{y}}$, which is termed the *working vector*. In matrix form

$$\tilde{\mathbf{y}} = \boldsymbol{\eta} + g'(\boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu}).$$

We replace $\boldsymbol{\beta}$ in $g'(\boldsymbol{\mu})$ by an approximation $\tilde{\boldsymbol{\beta}}$ and $\mathbf{u}_i$ in $g'(\boldsymbol{\mu})$ by its prior mean, $\mathbf{0}$, and so

$$\tilde{\mathbf{y}} = \boldsymbol{\eta} + g'(\boldsymbol{\mu})|_{\boldsymbol{\beta}=\tilde{\boldsymbol{\beta}}, \mathbf{u}=\mathbf{0}} (\mathbf{y} - \boldsymbol{\mu}).$$

Therefore,

$$\mathrm{E}(\tilde{\mathbf{Y}}|\mathbf{u}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u},$$

and

$$\mathrm{var}(\tilde{\mathbf{Y}}|\mathbf{u}) \approx \mathbf{W}^{-1}_{\tilde{\boldsymbol{\beta}},\mathbf{0}} = \mathrm{diag}\left\{\mathrm{var}(Y_{ij}) g'(\mu_{ij})^2\right\}_{\boldsymbol{\beta}=\tilde{\boldsymbol{\beta}}, \mathbf{u}=\mathbf{0}},$$

where we have replaced $\boldsymbol{\beta}$ and $\mathbf{u}$ in $\mathrm{var}(Y_{ij})$ by $\tilde{\boldsymbol{\beta}}$ and $\mathbf{0}$. We assume that

$$\tilde{\mathbf{y}} \sim \mathrm{N}\left(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}, \mathbf{W}^{-1}_{\tilde{\boldsymbol{\beta}},\mathbf{0}}\right),$$

and we can approximate the first-stage log-likelihood by

$$-\frac{1}{2}\log|\mathbf{W}^{-1}_{\tilde{\boldsymbol{\beta}},\mathbf{0}}| - \frac{1}{2}(\tilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})^T \mathbf{W}_{\tilde{\boldsymbol{\beta}},\mathbf{0}} (\tilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}),$$

We can then approximate the log of the integrated likelihood by

$$\begin{aligned} \log \hat{f}(\mathbf{y}|\boldsymbol{\beta}, \mathbf{D}, \phi) \quad \propto \quad & -\frac{1}{2}\log|\mathbf{W}^{-1}_{\tilde{\boldsymbol{\beta}},\mathbf{0}}| - \frac{1}{2}\log|\mathbf{D}^*| - \frac{1}{2}\log|\mathbf{D}^{*-1} + \mathbf{Z}^T\mathbf{W}_{\tilde{\boldsymbol{\beta}},\mathbf{0}}\mathbf{Z}| \\ & -\frac{1}{2}(\tilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{W}^{-1}_{\tilde{\boldsymbol{\beta}},\mathbf{0}} + \mathbf{Z}\mathbf{D}^*\mathbf{Z}^T)(\tilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta}). \end{aligned} \tag{3.11}$$

For fixed values of $\mathbf{D}^*$ and ϕ we can find approximate maximum likelihood estimates of $\boldsymbol{\beta}$ by iteratively fitting an LMM with responses $\tilde{y}_{ij}$ found using the current estimates $\tilde{\boldsymbol{\beta}}$ in $\mathbf{W}_{\tilde{\boldsymbol{\beta}},\mathbf{0}}$. Once this iterative scheme has converged, the resulting estimates of $\boldsymbol{\beta}$ are known as the maximum *marginal quasi-likelihood* (MQL) estimates of $\boldsymbol{\beta}$ (see, for example, Breslow and Clayton (1993)).

An approximation to the Fisher information of $\boldsymbol{\beta}$ can be obtained by replacing $\tilde{\boldsymbol{\beta}}$ in $\mathbf{W}_{\tilde{\boldsymbol{\beta}},\mathbf{0}}$ in (3.11) by its prior mean, $\mathbf{m}$, to get $\mathbf{W}_{\mathbf{m},\mathbf{0}}$. Then

$$\hat{I}_{\boldsymbol{\beta}}(\mathbf{m},\mathbf{0}) = \sum_{i=1}^{G} \mathbf{X}_i^T \left(\mathbf{W}_{i,\mathbf{m},\mathbf{0}}^{-1} + \mathbf{Z}_i \mathbf{D} \mathbf{Z}_i^T\right)^{-1} \mathbf{X}_i,$$

where $\mathbf{W}_{i,\mathbf{m},\mathbf{0}} = \text{diag}\left\{\text{var}(Y_{ij}) g'(\mu_{ij})^2\right\}^{-1}_{\boldsymbol{\beta}=\mathbf{m},\mathbf{u}_i=\mathbf{0}}$.

We assume that $\mathbf{W}_{i,\mathbf{m},\mathbf{0}}^{-1} = \sigma_i^2 \mathbf{I}_{n_i}$ for some σ_i^2. To see this, note that $\mu_{ij}|_{\boldsymbol{\beta}=\mathbf{m},\mathbf{u}=\mathbf{0}} = g^{-1}(m_1)$, i.e. a constant. This assumption holds for most standard GLMMs. One notable exception is when the responses are from the Poisson distribution with exposures which are not constant within group i. Under the above assumption that $\mathbf{W}_{i,\mathbf{m},\mathbf{0}}^{-1} = \sigma_i^2 \mathbf{I}_{n_i}$ for some σ_i^2,

$$\hat{\mathcal{I}}_{\boldsymbol{\beta}}(\mathbf{m},\mathbf{0}) = \sum_{i=1}^{G} \mathbf{X}_i^T (\sigma_i^2 \mathbf{I}_{n_i} + \mathbf{Z}_i \mathbf{D} \mathbf{Z}_i^T)^{-1} \mathbf{X}_i,$$

where we have assumed that $\text{var}(Y_{ij}) g'(\mu_{ij})^2|_{\boldsymbol{\beta}=\mathbf{m},\mathbf{u}=\mathbf{0}} = \sigma_i^2$ for all $j = 1, ..., n_i$. By replacing σ^2 by σ_i^2 in (3.4) we find that the approximate Fisher information of β_k is

$$\hat{\mathcal{I}}_{\boldsymbol{\beta}}(\mathbf{m},\mathbf{0})_{kk} = \sum_{i=1}^{G} \left(\frac{\mathbf{x}_{ik}^T \mathbf{x}_{ik}}{\sigma_i^2} - \sum_{l=1}^{q} \frac{\tau_l^2}{\sigma_i^2 + \tau_l^2 \mathbf{z}_{il}^T \mathbf{z}_{il}} \sum_{t=1}^{n_i} z_{ilt} x_{ikt} \sum_{r=1}^{n_i} z_{ilr} x_{ikr} \right),$$

where we have, again, assumed that $\mathbf{D} = \text{diag}\left\{\tau_1^2, ..., \tau_q^2\right\}$ and the columns of $\mathbf{X}_i$ are orthogonal.

We can use the same results of Pauler (1998) to show that if β_k has an associated group-specific parameter then $\hat{\mathcal{I}}_{\boldsymbol{\beta}}(\mathbf{m},\mathbf{0})$ is $O(G)$ but if β_k does not have an associated group-specific parameter then $\hat{\mathcal{I}}_{\boldsymbol{\beta}}(\mathbf{m},\mathbf{0})$ is $O(n)$.

Therefore an approximate unit information prior for the regression parameters, $\boldsymbol{\beta}$, of a GLMM based on the integrated likelihood is

$$\boldsymbol{\beta} \sim \text{N}\left(\mathbf{m}, \boldsymbol{\Lambda} \left(\sum_{i=1}^{G} \mathbf{X}_i^T \left(\sigma_i^2 \mathbf{I}_{n_i} + \mathbf{Z}_i \mathbf{D} \mathbf{Z}_i^T\right)^{-1} \mathbf{X}_i \right) \boldsymbol{\Lambda} \right), \tag{3.12}$$

where $\boldsymbol{\Lambda} = \text{diag}\left\{N_k\right\}$ and N_k is as defined in (3.9).

A GLM is a special case of a GLMM and the prior in (3.12) reduces to that shown in (3.2) for the $\boldsymbol{\beta}$ of a GLM since $n_i = 1$, $\mathbf{Z}_i = \mathbf{0}$ and $N_k = n$. Also we see that the above prior reduces to that shown in (3.8) for the $\boldsymbol{\beta}$ of an LMM since $\sigma_i^2 = \sigma^2$ for all i.

Note that the prior shown in (3.12) is conditional on the variance components matrix, $\mathbf{D}$. This means that the full conditional pdf of $\mathbf{D}$ decomposes as

$$f(\mathbf{D}|\mathbf{y}, \boldsymbol{\beta}, \mathbf{u}, \phi) \propto f(\mathbf{u}|\mathbf{D}) f(\mathbf{D}|\phi) f(\boldsymbol{\beta}|\mathbf{D}),$$

so the full conditional distribution of $\mathbf{D}$ is not independent of $\boldsymbol{\beta}$. Suppose the prior distribution of $\mathbf{D}$ is the inverse-Wishart distribution. If $\mathbf{D}$ is conditionally independent of $\boldsymbol{\beta}$ and $\mathbf{y}$

then the full conditional distribution of $\mathbf{D}$ is also inverse-Wishart. However, if the prior of $\boldsymbol{\beta}$ depends on $\mathbf{D}$ then the full conditional distribution of $\mathbf{D}$ is not inverse-Wishart. This is a computational disadvantage if we use Gibbs sampling to generate a posterior sample. Indeed, if $q > 1$ and the prior distribution for $\mathbf{D}$ inverse-Wishart, then WinBUGS has difficulties in generating a posterior sample from the posterior distribution of the resulting GLMM. This is because WinBUGS requires that if the inverse-Wishart distribution is used than it must be conditionally conjugate, i.e. the prior distribution for $\boldsymbol{\beta}$ must not depend on $\mathbf{D}$. We could replace $\mathbf{D}$ in (3.12) by its prior mean, if it exists, to remove the dependence of the prior distribution of $\boldsymbol{\beta}$ on $\mathbf{D}$. However, the disadvantage of doing so is that the prior variance of $\boldsymbol{\beta}$ will become heavily dependent on the prior mean of $\mathbf{D}$. This is similar to in Section 3.1 where the prior distribution of $\boldsymbol{\beta}$ in a linear model depends on σ^2.

Another issue with this prior is that $\text{var}(Y_{ij})g'(\mu_{ij})^2|_{\boldsymbol{\beta}=\mathbf{m},\mathbf{u}_i=\mathbf{0}} = \sigma_i^2$ for all $j = 1, ..., n_i$. This does not hold for an example we consider in Chapter 6. In this example, $y_{ij} \sim \text{Poisson}(E_{ij}\lambda_{ij})$ where E_{ij} is the exposure for the jth unit in the ith group. Therefore, $\text{var}(Y_{ij})g'(\mu_{ij})^2|_{\boldsymbol{\beta}=\mathbf{m},\mathbf{u}_i=\mathbf{0}} = \frac{1}{E_{ij}\exp(m_1)}$ and $E_{ij} \neq E_i$ for all $j = 1, ..., n_i$.

For these reasons we seek an alternative to this prior distribution that is still based on a unit information concept. In Section 3.3, we define a unit information prior for the regression parameters, $\boldsymbol{\beta}$, which is based on the first-stage likelihood.

3.2.2 Variance Components

In Section 3.2.1, we defined a unit information prior for the regression parameters, $\boldsymbol{\beta}$, based on the integrated likelihood but found it had some unattractive properties. We go on to define a unit information prior for $\boldsymbol{\beta}$ in Section 3.3 that does not have these unattractive properties. We are also able to define a default prior for $\mathbf{D}$ which is based on a unit information prior in Section 3.3. Nevertheless, in this Section, we discuss, heuristically, how we would define a unit information prior for the variance components matrix, $\mathbf{D}$, based on the integrated likelihood. The preceding discussion on unit information priors has relied on the model parameters, $\boldsymbol{\theta}$, lying in $\mathbb{R}^k$ and thus, to be plausible for the prior distribution for $\boldsymbol{\theta}$ being a multivariate normal distribution. This is clearly not appropriate for the variance components matrix, $\mathbf{D}$, which lies in $\mathbb{P}^q$, the set of all $q \times q$ positive-definite matrices.

Recall that $\mathbf{D}$ depends on $t = \frac{1}{2}q(q+1)$ unique elements denoted $\mathbf{d} = (d_1, ..., d_t)^T$. Natarajan and Kass (2000) give the (r, s)th element of the approximate Fisher information matrix, $\hat{\mathcal{I}}_{\mathbf{d}}(\mathbf{d}, \boldsymbol{\beta}, \mathbf{u})$, of $\mathbf{d}$ as

$$\hat{\mathcal{I}}(\mathbf{d}, \boldsymbol{\beta}, \mathbf{u})_{rs} = \sum_{i=1}^{G} \text{tr}\left(\mathbf{V}_i^{-1}\frac{\partial \mathbf{D}}{\partial d_r}\mathbf{V}_i^{-1}\frac{\partial \mathbf{D}}{\partial d_s}\right),$$

for $r, s = 1, ..., t$, where

$$\mathbf{V}_i = \mathbf{D} + \left(\mathbf{Z}_i\mathbf{W}_i\mathbf{Z}_i^T\right)^{-1},$$

and $\mathbf{W}_i = \text{diag}\left\{\text{var}(Y_{ij})g'(\mu_{ij})^2\right\}^{-1}$. The approximate Fisher information for $\mathbf{d}$ depends on the unknown $\mathbf{D}$ as well as $\boldsymbol{\beta}$ and $\mathbf{u}$. We can assume an inverse-Wishart prior distribution for

$\mathbf{D}$, with shape parameter $\rho > q-1$ and scale matrix $\mathbf{R}$, i.e.

$$\mathbf{D} \sim \text{IW}(\rho, \mathbf{R}),$$

so that $\mathbf{D}$ has pdf, $f(\mathbf{D})$, such that

$$f(\mathbf{D}) \propto |\mathbf{D}|^{-\frac{\rho+q+1}{2}} \exp\left(-\frac{1}{2}\text{tr}(\mathbf{R}\mathbf{D}^{-1})\right).$$

Using this specification, we can induce a distribution for $\mathbf{d}$ and therefore a prior mean of $\mathbf{d}$ which we denote as $\tilde{\mathbf{d}}$. We can then replace $\mathbf{d}$, $\boldsymbol{\beta}$ and $\mathbf{u}$ in the approximate Fisher information of $\mathbf{d}$ by the prior means of $\mathbf{d}$, $\mathbf{m}$ and $\mathbf{0}$, respectively, to give $\hat{\mathcal{I}}_{\mathbf{d}}(\tilde{\mathbf{d}}, \mathbf{m}, \mathbf{0})$.

To define a unit information prior for $\mathbf{d}$, and therefore $\mathbf{D}$, we could find the order, N_r, of the rth diagonal elements of $\hat{\mathcal{I}}_{\mathbf{d}}(\tilde{\mathbf{d}}, \mathbf{m}, \mathbf{0})$ which are denoted as $\hat{\mathcal{I}}_{\mathbf{d}}(\tilde{\mathbf{d}}, \mathbf{m}, \mathbf{0})_{rr}$. The prior variance of $\mathbf{d}$ could then be set to $\boldsymbol{\Lambda}\hat{\mathcal{I}}_{\mathbf{d}}(\tilde{\mathbf{d}}, \mathbf{m}, \mathbf{0})^{-1}\boldsymbol{\Lambda}$, where $\boldsymbol{\Lambda} = \text{diag}\left\{\sqrt{N_r}\right\}$.

We see in Section 3.3 how to define a unit information prior for $\mathbf{D}$, based on the first-stage likelihood.

3.3 Default Priors based on the First-Stage Likelihood

3.3.1 Regression Parameters

We now define a unit information prior for the regression parameters, $\boldsymbol{\beta}$, of a GLMM based on the first-stage likelihood. The first-stage likelihood of a GLMM is

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

The log of the first-stage likelihood is then

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

Then we see that

$$\frac{\partial \log f(\mathbf{y}|\boldsymbol{\beta}, \mathbf{u}, \phi)}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{G}\sum_{j=1}^{n_i} \frac{y_{ij} - \mu_{ij}}{\text{var}(Y_{ij})g'(\mu_{ij})}\mathbf{x}_{ij}. \tag{3.13}$$

Therefore the Fisher information of $\boldsymbol{\beta}$ is

$$\begin{aligned}\mathcal{I}_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \mathbf{u}) &= \sum_{i=1}^{G}\sum_{j=1}^{n_i} \frac{1}{\text{var}(Y_{ij})g'(\mu_{ij})^2}\mathbf{x}_{ij}\mathbf{x}_{ij}^T, \\ &= \mathbf{X}^T\mathbf{W}\mathbf{X},\end{aligned}$$

where $\mathbf{W} = \text{diag}\left\{\text{var}(Y_{ij})g'(\mu_{ij})^2\right\}^{-1}$. Note that $\mathcal{I}_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \mathbf{u})$ depends on $\boldsymbol{\beta}$ and $\mathbf{u}$. We replace these by their prior means of $\mathbf{m}$ and $\mathbf{0}$, respectively, to give

$$\mathcal{I}_{\boldsymbol{\beta}}(\mathbf{m}, \mathbf{0}) = \mathbf{X}^T\mathbf{W}_{\mathbf{m},\mathbf{0}}\mathbf{X},$$

where $\mathbf{W}_{\mathbf{m},\mathbf{0}} = \text{diag}\left\{\text{var}(Y_{ij})g'(\mu_{ij})^2\right\}|^{-1}_{\boldsymbol{\beta}=\mathbf{m},\mathbf{u}=\mathbf{0}}$. Note that $\mathcal{I}_{\boldsymbol{\beta}}(\mathbf{m}, \mathbf{0})_{kk}$ is $O(n)$. So a unit information prior for $\boldsymbol{\beta}$ of a GLMM based on the first-stage likelihood is

$$\boldsymbol{\beta} \sim \text{N}\left(\mathbf{m}, n(\mathbf{X}^T\mathbf{W}_{\mathbf{m},\mathbf{0}}\mathbf{X})^{-1}\right). \tag{3.14}$$

This prior distribution is not conditional on the variance components matrix, $\mathbf{D}$, and therefore if the prior distribution of $\mathbf{D}$ is inverse-Wishart, then we can take advantage of the conditional independence of $\mathbf{D}$ and, $\boldsymbol{\beta}$ and $\mathbf{y}$.

The unit information prior (3.14) reduces to that in (3.1) and (3.2) for linear models and GLMs, respectively.

The prior in (3.14) can also be applied to an LMM. In this case $\mathbf{W}_{\mathbf{m},\mathbf{0}} = \frac{1}{\sigma^2}\mathbf{I}_n$ and

$$\boldsymbol{\beta} \sim \text{N}\left(\mathbf{m}, n\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\right).$$

3.3.2 Variance Components

We now define a prior for the variance components matrix, $\mathbf{D}$, based on a unit information concept and the first-stage likelihood. Since a GLMM is a hierarchical model, $\mathbf{D}$ does not feature in the first-stage likelihood. Indeed, we are attempting to define a default hyperprior for $\mathbf{D}$.

We begin by letting the prior distribution of $\mathbf{D}$ be the inverse-Wishart distribution with shape parameter ρ and scale matrix $\mathbf{R}$, so that

$$\mathbf{D} \sim \text{IW}(\rho, \mathbf{R}).$$

It can be shown that $\text{E}(\mathbf{D}) = \frac{1}{\rho - q - 1}\mathbf{R}$, provided $\rho > q + 1$.

Suppose we regard the model as non-hierarchical so that $\mathbf{D}$ is actually a fixed hyperparameter that we need to determine to define the prior distribution for $\mathbf{u}$. We proceed by beginning to define a unit information prior for $\mathbf{u}$ based on the first-stage likelihood.

The log of the first-stage likelihood is

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

and it follows that the Fisher information for $\mathbf{u}_i$ is

$$\mathcal{I}_{\mathbf{u}_i}(\boldsymbol{\beta}, \mathbf{u}_i) = \mathbf{Z}_i^T\mathbf{W}_i\mathbf{Z}_i,$$

for $i = 1, ..., G$, where $\mathbf{W}_i = \text{diag}\left\{\text{var}(Y_{ij})g'(\mu_{ij})^2\right\}^{-1}$. The Fisher information $\mathbf{u}_i$ depends on $\boldsymbol{\beta}$ and $\mathbf{u}_i$, so we replace these by their prior means of $\mathbf{m}$ and $\mathbf{0}$, respectively, to give

$$\mathcal{I}_{\mathbf{u}_i}(\mathbf{m}, \mathbf{0}) = \mathbf{Z}_i^T \mathbf{W}_{i,\mathbf{m},\mathbf{0}} \mathbf{Z}_i,$$

for $i = 1, ..., G$, where $\mathbf{W}_{i,\mathbf{m},\mathbf{0}} = \text{diag}\left\{\text{var}(Y_{ij})g'(\mu_{ij})^2\right\}^{-1}_{\boldsymbol{\beta}=\mathbf{m},\mathbf{u}_i=\mathbf{0}}$. It is easy to show that the diagonal elements, $\mathcal{I}_{\mathbf{u}_i}(\mathbf{m}, \mathbf{0})_{kk}$, of $\mathcal{I}_{\mathbf{u}_i}(\mathbf{m}, \mathbf{0})$, for $k = 1, ..., q$, are $O(n_i)$. Therefore, the unit information of $\mathbf{u}_i$ is $\frac{1}{n_i}\mathbf{Z}_i^T \mathbf{W}_{i,\mathbf{m},\mathbf{0}} \mathbf{Z}_i$, for $i = 1, ..., G$. Since we have G groups we find the average unit information over the G groups as $\frac{1}{G}\sum_{i=1}^{G} \frac{1}{n_i}\mathbf{Z}_i^T \mathbf{W}_{i,\mathbf{m},\mathbf{0}} \mathbf{Z}_i$. This is similar to how Natarajan and Kass (2000) average a similar quantity over the G groups for the uniform shrinkage prior for $\mathbf{D}$ as discussed in Section 2.3.3. If $\mathbf{D}$ was a fixed hyperparameter and we were defining a unit information prior for $\mathbf{u}$ we would set $\mathbf{D} = G\left(\sum_{i=1}^{G} \frac{1}{n_i}\mathbf{Z}_i^T \mathbf{W}_{i,\mathbf{m},\mathbf{0}} \mathbf{Z}_i\right)^{-1}$. However, since $\mathbf{D}$ is not fixed we let its expectation

$$\text{E}(\mathbf{D}) = \frac{\mathbf{R}}{\rho - q - 1} = G\left(\sum_{i=1}^{G} \frac{1}{n_i}\mathbf{Z}_i^T \mathbf{W}_{i,\mathbf{m},\mathbf{0}} \mathbf{Z}_i\right)^{-1},$$

so that

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

It remains to find a value for the shape parameter, $\rho > q + 1$, or if $\rho = q + \epsilon + 1$, to find a value for $\epsilon > 0$. If the $\mathbf{u}_i$'s are regarded as responses, with likelihood

$$f(\mathbf{u}|\mathbf{D}) \propto |\mathbf{D}|^{-\frac{G}{2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{G} \mathbf{u}_i^T \mathbf{D}^{-1} \mathbf{u}_i\right). \tag{3.15}$$

The prior distribution for $\mathbf{D}$ has pdf

$$f(\mathbf{D}) \propto |\mathbf{D}|^{-\frac{2q+2+\epsilon}{2}} \exp\left(\frac{1}{2}\text{tr}(\mathbf{R}\mathbf{D}^{-1})\right).$$

An increase in ϵ by one corresponds to the prior distribution contributing one extra group of responses. Since the likelihood (3.15) will provide information for $\mathbf{D}$ that is proportional to G, we can argue that $\epsilon = 1$, using a unit information concept. Therefore the prior distribution for $\mathbf{D}$ is

$$\mathbf{D} \sim \text{IW}(\rho, \mathbf{R}),$$

where $\rho = q + 2$ and

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

3.4 Dispersion Parameter

In this Section, we focus on default priors for the dispersion parameter, ϕ, of GLMMs (and therefore linear models, GLMs and LMMs). In Section 3.1, we discussed how, if a parameter is

present in all models in M then, the prior distribution of that parameter is unimportant with regards to the posterior model probabilities. We assumed in Section 3.1 that the dispersion parameter, if unknown, will be present in all models in M. Therefore, we can apply a very diffuse prior distribution for ϕ.

In a linear model, $\phi = \sigma^2$, i.e. the dispersion parameter is the variance of the response, independent of the mean of the response. In this case, the conjugate prior distribution for σ^2 is the inverse-gamma distribution, $\text{IG}(a, b)$, with shape parameter, a, and scale parameter, b, where $a, b > 0$. A common way to make the inverse-gamma distribution diffuse is to set $a = b = \epsilon$, where ϵ is small. We will follow this approach throughout and will apply this diffuse prior to all dispersion parameters that we encounter.

3.5 Simulation Study

In this Section, we test the robustness and efficacy of using the default priors for $\boldsymbol{\beta}$ and $\mathbf{D}$ based on the first-stage likelihood that we proposed in Section 3.3.

We test these priors using simulation studies. Let y_{ij} be a response from an exponential family distribution with mean $\mu_{ij} = g^{-1}(\eta_{ij})$ and dispersion parameter ϕ. We consider model determination amongst five models with the following linear predictors:

1. $\eta_{ij} = \beta_1$,
2. $\eta_{ij} = \beta_1 + \beta_2 x_{ij}$,
3. $\eta_{ij} = \beta_1 + u_i$, where $u_i \overset{\text{iid}}{\sim} \text{N}(0, \tau^2)$,
4. $\eta_{ij} = \beta_1 + u_i + \beta_2 x_{ij}$, where $u_i \overset{\text{iid}}{\sim} \text{N}(0, \tau^2)$,
5. $\eta_{ij} = (\beta_1 + u_{1i}) + (\beta_2 + u_{2i})x_{ij}$, where $\mathbf{u}_i = (u_{i1}, u_{i2})^T \overset{\text{iid}}{\sim} \text{N}(\mathbf{0}, \mathbf{D})$,

for $j = 1, ..., n^*$ and $i = 1, ..., G$, where $n = Gn^*$. We generate responses from the model with linear predictor 4. To do this, we generate x'_{ij} independently from the standard normal distribution and then set x_{ij} to be the standardised x'_{ij}. We then choose the true values $\boldsymbol{\beta}^* = (\beta_1^*, \beta_2^*)^T$, ϕ^* (if unknown for the chosen response distribution), and τ^{*2}. We generate u_i^* independently from $\text{N}(0, \tau^{*2})$ and use them to find the true linear predictor via $\eta_{ij}^* = \beta_1^* + u_i^* + \beta_2^* x_{ij}$. We then generate y_{ij} from the chosen distribution with mean $\mu_{ij} = g^{-1}(\eta_{ij}^*)$ and dispersion parameter ϕ^*.

We choose three different combinations of n^* and G so that $n = 100$, always. They are $(n^*, G) = (10, 10)$, $(n^*, G) = (5, 20)$ and $(n^*, G) = (20, 5)$.

We consider three different response distributions: normal, Bernoulli and Poisson. For the Bernoulli and Poisson distributions, the dispersion parameter is known. For the normal

distribution, the dispersion parameter is σ^2, the variance of the response, and we set this to be 1.

For all response distributions, we set the intercept parameter, β_1^*, to be $\frac{1}{2}$. We are now left with a choice of the parameters β_2^* and τ^{*2}. We generate these parameters from the U$[0, a]$ distribution where $a = 5$ for the Bernoulli response distribution and $a = \frac{5}{4}$ for the normal and Poisson response distributions.

For each combination of (n^*, G) and response distribution, we generate the true values β_2^* and τ^{*2}, and then generate the responses. For each of the five models, we apply the default priors we proposed in Section 3.3 to the parameters $\boldsymbol{\beta}$ and $\mathbf{D}$. In the unit information prior distribution for $\boldsymbol{\beta}$, we set $m_1 = 0$ and recall that the remaining elements of the prior mean are also 0. When the response distribution is normal, the dispersion parameter, σ^2, is present in all five models, and, according to Section 3.4, we can apply the same diffuse IG(ϵ, ϵ) prior distribution where $\epsilon = 0.001$. We then approximate the posterior model probabilities using the reversible jump algorithm that we describe in Chapter 5. We run this algorithm for a total of 2000 iterations, with a burn-in phase of 100 iterations. We repeat this process 500 times, in each case recording β_2^*, τ^{*2}, and the approximated posterior model probabilities. In each case, we also record the observed variance, $\hat{\tau}^{*2}$, of the u_i^*'s, i.e.

$$\hat{\tau}^{*2} = \frac{1}{G-1} \sum_{i=1}^{G} (u_i^* - \bar{u}^*)^2,$$

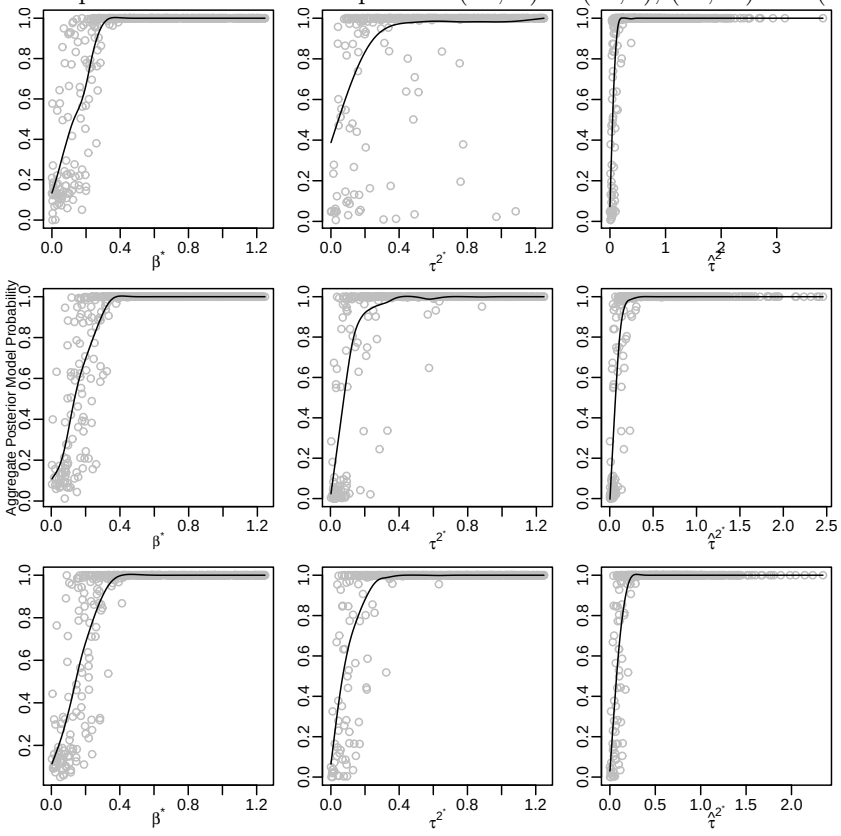
where $\bar{u}^* = \frac{1}{G} \sum_{i=1}^{G} u_i^*$.

Model 4 is the true model. However, by generating β_2^* and τ^{*2} from U$[0, a]$ we hope that the default priors will give more posterior model probability to the simpler, more parsimonious Models 1, 2 and 3, when β_2^* and/or τ^{*2} are small. Although Model 5 is also the true model in the sense that Model 4 is nested within Model 5, it is never the most parsimonious model and we hope that the default priors will give small posterior model probability to Model 5.

To show the efficacy of the unit information prior for $\boldsymbol{\beta}$, we plot the aggregate posterior model probabilities of Models 2, 4 and 5, i.e. the models that contain a β_2 parameter, against β_2^*. To show the efficacy of the unit information prior for $\mathbf{D}$, we plot the aggregate posterior model probabilities of Models 3, 4 and 5, i.e. the models that contain group-specific parameters against τ^{*2} and, in a separate plot, against the observed value, $\hat{\tau}^{*2}$, of τ^{*2}. The reason we produce this additional plot of the posterior model probabilities against $\hat{\tau}^{*2}$ is that, particularly for small values of G, the actual variance observed in the u_i's can be much smaller than the true value and the model determination strategy can have trouble detecting it. For larger G, the two plots should be approximately the same, since $\hat{\tau}^{*2} \to \tau^{*2}$ as G increases.

Figures 3.1, 3.2 and 3.3 show these plots for the Poisson, normal and Bernoulli responses, respectively, for the three combinations of (n^*, G). Consider the first column in the three Figures. The same behaviour is shown in all three Figures, i.e. the aggregate posterior model probability of the models that contain a β_2 term increasing from 0 to 1 as β_2^* increases from 0. The same behaviour can be seen for the aggregate posterior model probability of the

Figure 3.1: Aggregate posterior model probabilities for Models 2, 4 and 5 (first column) and Models 3, 4 and 5 (second and third columns) plotted against β_2^* (first column), τ^{*2} (second column), and $\hat{\tau}^{*2}$ (third column), for Poisson responses. The rows correspond to (n^*, G) as $(20, 5)$, $(10, 10)$ and $(5, 20)$, respectively.



models that contain a group-specific intercept increases from 0 to 1 as $\hat{\tau}^{*2}$ increases from 0. As expected, the strategy has difficulty in favouring models containing a group-specific intercept when $\hat{\tau}^{*2}$ is small, regardless of the value of τ^{*2}.

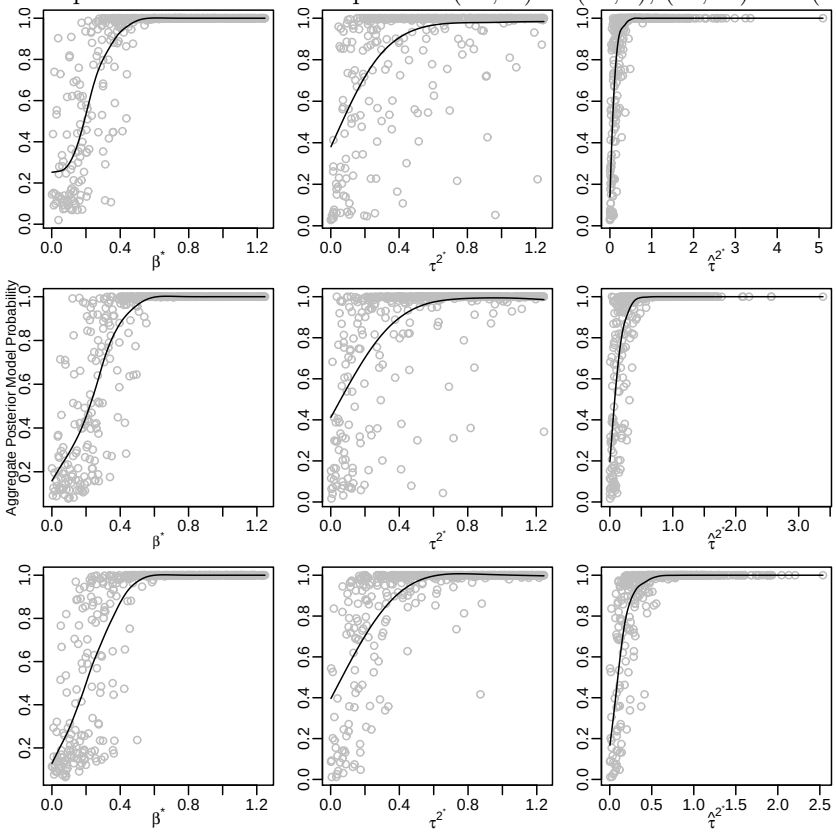
This behaviour of the aggregate posterior model probability increasing from 0 is exactly that which we desire. i.e. the proposed default priors penalise too complicated models.

With regards to Model 5, which is never the most parsimonious model available, Table 3.1 shows sample statistics of the posterior model probabilities of Model 5 for each of the combinations of (n^*, G), for Poisson, normal and Bernoulli responses.

Table 3.1 shows that the posterior model probabilities of Model 5 are typically very small indicating that we will rarely favour a too complicated model using the proposed default priors.

For each dataset generated, we generated a posterior sample of size 2000 after a burn-in phase of 500 iterations using WinBUGS under Model 4, i.e. the hypothetically true model. Using this sample we produced 95% probability intervals for the parameters β_1, β_2 and τ^2 for the Poisson and Bernoulli response distributions and for the parameters β_1, β_2, τ^2 and σ^2 for the normal response distribution. Using a posterior sample $\{\theta_1, ..., \theta_N\}$ of size N of the

Figure 3.2: Aggregate posterior model probabilities for Models 2, 4 and 5 (first column) and Models 3, 4 and 5 (second and third columns) plotted against β_2^* (first column), τ^{*2} (second column), and $\hat{\tau}^{*2}$ (third column), for normal responses. The rows correspond to (n^*, G) as $(20, 5)$, $(10, 10)$ and $(5, 20)$, respectively.



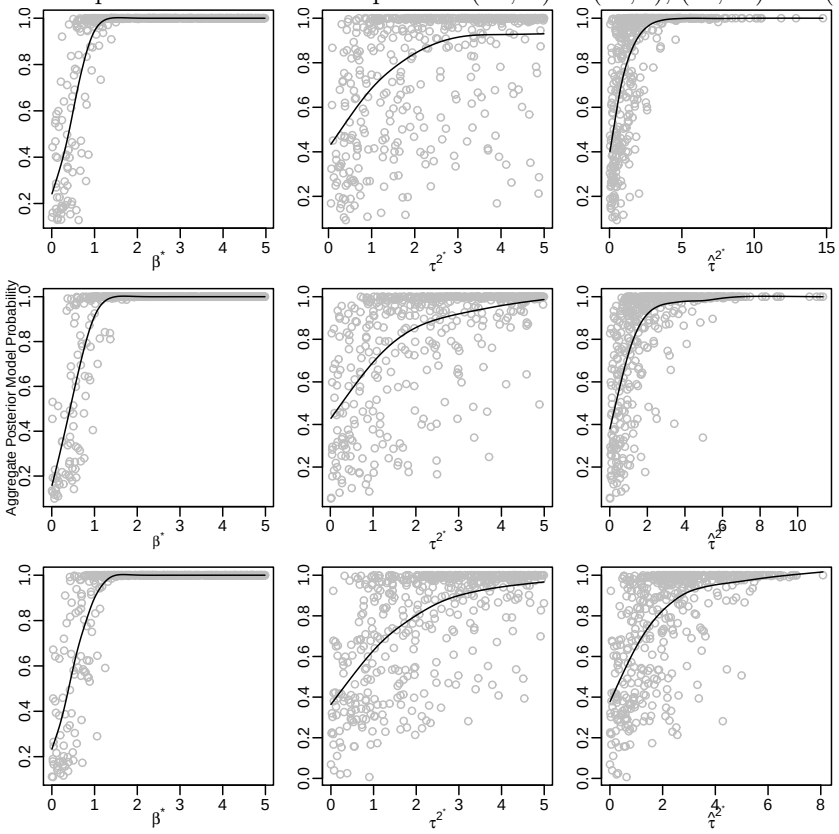
parameter θ, a $(1-\alpha)\%$ probability interval for θ is approximated by

$$\left(\theta_{\left(\frac{N\alpha}{2}\right)}, \theta_{\left(\frac{N(1-\alpha)}{2}\right)}\right),$$

where $\theta_{(k)}$ denotes the kth value of the ordered posterior sample. We now investigate the coverage rates of those probability intervals. Table 3.2 shows the coverage rates.

From Table 3.2, note that for the regression parameters, β_1 and β_2, and the dispersion parameter, σ^2, for normal responses, the coverage rates are very close to the nominal value of 95%. The coverage rates for the variance component, τ^2, are still close to 95% but are always about 5% too low. Also shown in Table 3.2 is the median value, V, of τ^{*2} when the probability interval for τ^2 does not contain τ^{*2}. Note that the value V is small compared to the theoretical median of all of the τ^{*2}'s of 0.625 for Poisson and normal responses and 2.5 for Bernoulli responses. This indicates that the probability interval for τ^2, does not contain τ^{*2} when the value of τ^{*2} is small. In these case, we have shown earlier that we are unlikely to favour models which contain group-specific parameters, i.e. models with non-zero τ^2. Therefore, we should not be concerned that the probability interval does not contain τ^{*2} in these cases, since we are unlikely to choose a model that contains τ^2 and are more likely to favour a more parsimonious model.

Figure 3.3: Aggregate posterior model probabilities for Models 2, 4 and 5 (first column) and Models 3, 4 and 5 (second and third columns) plotted against β_2^* (first column), τ^{*2} (second column), and $\hat{\tau}^{*2}$ (third column), for Bernoulli responses. The rows correspond to (n^*, G) as $(20, 5)$, $(10, 10)$ and $(5, 20)$, respectively.



3.6 Turtle Data Example

We apply the proposed unit information priors to the model parameters of the five models that can be applied to the Turtle Dataset. We set $m_1 = 0$, since this corresponds a prior mean of $\frac{1}{2}$ for the response.

For Models 1 and 3, the prior for the regression parameter is $\beta \sim \mathrm{N}(0, \frac{\pi}{2})$. For Models 2, 4 and 5, the prior for the regression parameters is

$$\boldsymbol{\beta} \sim \mathrm{N}\left(\mathbf{0}, \frac{\pi}{2}\begin{pmatrix} 1 & 0 \\ 0 & \frac{n}{n-1} \end{pmatrix}\right).$$

For Models 3 and 4, the prior for the variance component is $\sigma^2 \sim \mathrm{IG}\left(\frac{3}{2}, \frac{\pi}{4}\right)$. Finally, for Model 5, the prior for the variance components matrix is $\mathbf{D} \sim \mathrm{IW}(4, \mathbf{R})$, where

$$\mathbf{R} = \frac{\pi}{2}\frac{1}{\sum_{i=1}^{G}\sum_{j=1}^{n_i}\frac{x_{ij}^2}{n_i} - \frac{1}{G}\left(\sum_{i=1}^{G}\sum_{j=1}^{n_i}\frac{x_{ij}}{n_i}\right)^2}\begin{pmatrix} \sum_{i=1}^{G}\sum_{j=1}^{n_i}\frac{x_{ij}^2}{n_i} & -\sum_{i=1}^{G}\sum_{j=1}^{n_i}\frac{x_{ij}}{n_i} \\ -\sum_{i=1}^{G}\sum_{j=1}^{n_i}\frac{x_{ij}}{n_i} & G \end{pmatrix}.$$

We approximate the posterior model probabilities by using the marginal likelihood approach. The method used to approximate the marginal likelihoods is importance sampling. The details on how this is achieved is given in Section 4.5.1, and the approximate marginal likelihoods

Table 3.1: Sample statistics of the posterior model probabilities of Model 5 for Poisson, normal and Bernoulli responses for each of the combinations of (n^*, G).

Poisson Responses			
(n^*, G)	$(20, 5)$	$(10, 10)$	$(5, 20)$
Minimum	0.000	0.000	0.000
Median	0.024	0.013	0.014
Maximum	1.000	0.993	0.911
Normal Responses			
(n^*, G)	$(20, 5)$	$(10, 10)$	$(5, 20)$
Minimum	0.000	0.000	0.000
Median	0.050	0.047	0.054
Maximum	0.915	0.998	0.986
Bernoulli Responses			
(n^*, G)	$(20, 5)$	$(10, 10)$	$(5, 20)$
Minimum	0	0	0
Median	0.119	0.130	0.144
Maximum	0.963	0.915	0.968

are given in Table 4.1 on page 100. Note that the sample size used in the importance sampler is so large that the approximations can be considered exact. These marginal likelihoods give rise to the posterior model probabilities shown in Table 3.3, accurate to four decimal places.

The posterior model probabilities of Models 1 and 3 are negligible indicating there is strong evidence of a birthweight effect on the survival probability of a turtle. Model 5 has the highest posterior model probability which indicates that this birthweight effect is different for the different clutches.

The Bayes factor in favour of Model 2 over Model 4 is 1.862. This is closer to the equivalent Bayes factor of Sinharay and Stern (2005) than the equivalent Bayes factor of Sinharay and Stern (2000) but we have used a formal concept to define our prior distribution for σ^2.

3.7 Discussion

In this Chapter, we discussed unit information prior distributions applied to the regression parameters, $\boldsymbol{\beta}$, and the variance components matrix, $\mathbf{D}$, of a GLMM. For the regression parameters, $\boldsymbol{\beta}$, we defined unit information priors based on the integrated and first-stage likelihoods.

We found that the prior for $\boldsymbol{\beta}$ based on the integrated likelihood was conditional on $\mathbf{D}$ and could not be applied when $\text{var}(Y_{ij})g'(\mu_{ij})^2|_{\boldsymbol{\beta}=\mathbf{m},\mathbf{u}_i=\mathbf{0}}$ is not constant for all $j = 1, ..., n_i$. The fact that the prior is conditional on $\mathbf{D}$ is a computational disadvantage if $\mathbf{D}$ has an inverse-

Table 3.2: Coverage rates of the probability intervals for the parameters for the Poisson, normal and Bernoulli responses for each of the combinations of (n^*, G). The nominal rate is 95%.

Poisson Responses			
(n^*, G)	$(20, 5)$	$(10, 10)$	$(5, 20)$
β_1	0.950	0.950	0.952
β_2	0.936	0.944	0.958
τ^2	0.892	0.906	0.910
V	0.0515	0.0740	0.0725
Normal Responses			
(n^*, G)	$(20, 5)$	$(10, 10)$	$(5, 20)$
β_1	0.938	0.940	0.948
β_2	0.938	0.936	0.948
τ^2	0.916	0.874	0.912
σ^2	0.960	0.944	0.948
V	0.0580	0.0755	0.0662
Bernoulli Responses			
(n^*, G)	$(20, 5)$	$(10, 10)$	$(5, 20)$
β_1	0.966	0.964	0.952
β_2	0.952	0.948	0.958
τ^2	0.902	0.878	0.900
V	0.250	0.234	0.290

Table 3.3: Posterior Model Probabilities, $f((m|\mathbf{y})$ of the five models for the Turtle Dataset having used the proposed unit information prior distributions.

Model, m	Posterior Model Probability, $f(m\|\mathbf{y})$
1	0.0000
2	0.3484
3	0.0013
4	0.1871
5	0.4632

Wishart prior distribution since $\mathbf{D}$ is not conditionally conjugate. We could replace $\mathbf{D}$ in the prior by its prior mean but this is not recommended since the prior variance for $\boldsymbol{\beta}$ would be heavily dependent on this value.

The unit information prior for $\boldsymbol{\beta}$ based on the first-stage likelihood is independent of $\mathbf{D}$ and can be applied when $\text{var}(Y_{ij})g'(\mu_{ij})^2|_{\boldsymbol{\beta}=\mathbf{m},\mathbf{u}_i=\mathbf{0}}$ is not constant for all $j = 1, ..., n_i$. For these reasons , we prefer the prior for $\boldsymbol{\beta}$ based on the first-stage likelihood. We then defined a unit information concept prior for $\mathbf{D}$ based on the first-stage likelihood.

In Section 3.5, we tested the efficacy and robustness of these default priors with respect to

model determination by using simulation studies. We found that using the proposed default prior distributions would lead to model determination that had behaviour that we desired. That is, as the true value of the parameter increased the posterior model probability of the models that contained that parameter correspondingly increased. The default priors seemed to penalise over complicated models.

Note that the proposed unit information prior for $\boldsymbol{\beta}$ based on the first-stage likelihood is approximate in the sense that $\boldsymbol{\beta}$ and $\mathbf{u}_i$ are replaced in the weight matrix in the Fisher information by their prior means, $\mathbf{m}$ and $\mathbf{0}$, respectively. We mentioned that $\boldsymbol{\beta}$ could be replaced by some maximum likelihood estimate, $\hat{\boldsymbol{\beta}}$, resulting in a data-dependent prior distribution. This approach of replacing $\boldsymbol{\beta}$ by its maximum likelihood estimate is suggested by Natarajan and Kass (2000) and Gustafson et al. (2006), among others. However, in either case, as noted by Gustafson et al. (2006), "the weight matrix tends to vary slowly over the parameter space in most instances".

For the unit information concept prior distribution for $\mathbf{D}$ we also take the approach of replacing $\boldsymbol{\beta}$ and $\mathbf{u}_i$ in the weight matrix by their prior means. Again the weight matrix will vary slowly over the parameter space. The unit information concept prior for $\mathbf{D}$ is not unique in that we have a choice for the parameter ρ. We chose $\rho = q + 2$, so that the mean of the prior distribution for $\mathbf{D}$ exists.

We applied the proposed unit information prior distributions to the models of the Turtle Dataset and arrived at a model determination conclusion that was similar to those of Sinharay and Stern (2005) who used a different default prior distribution.

Chapter 4

Approximating the Marginal Likelihood for GLMMs

4.1 Introduction

In this Chapter, we discuss the methods of bridge sampling and nested sampling for approximating the marginal likelihood for the particular application to GLMMs. We assume that the number of models, $|M|$, is small enough that using computationally intensive methods for approximating the marginal likelihood of each model $m \in M$, such as bridge sampling and nested sampling is practical, or that we have used some other method (see Chapter 5) to identify a smaller subset, $M^* \subset M$, of models with high posterior model probability such that $|M^*| < |M|$ is manageable.

4.2 Bridge Sampling

4.2.1 Introduction

In Section 2.2.5 we gave the optimal, iterative bridge sampling approximation to the unknown integral $I = \int_\Theta g(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}$ of Meng and Wong (1996) in (2.26) as

$$\hat{I}_{BS,O}^{(t+1)} = \frac{\frac{1}{n_H}\sum_{i=1}^{n_H} \frac{l_{Hi}}{n_\Pi l_{Hi} + n_H \hat{I}_{BS,O}^{(t)}}}{\frac{1}{n_\Pi}\sum_{i=1}^{n_\Pi} \frac{1}{n_\Pi l_{\Pi i} + n_H \hat{I}_{BS,O}^{(t)}}}, \tag{4.1}$$

where $l_{ki} = g(\boldsymbol{\theta}_i^k)/h(\boldsymbol{\theta}_i^k)$ for $k = H, \Pi$, and $\{\boldsymbol{\theta}_1^H, ..., \boldsymbol{\theta}_{n_H}^H\}$ and $\{\boldsymbol{\theta}_1^\Pi, ..., \boldsymbol{\theta}_{n_\Pi}^\Pi\}$ are samples generated from H and Π, respectively. Here Π has pdf $\pi(\boldsymbol{\theta}) = g(\boldsymbol{\theta})/I$ and H has pdf $h(\boldsymbol{\theta})$.

4.2.2 Bridge Sampling in Practice

To implement (4.1) in practice, we need to specify the allocation of sample sizes, $\frac{n_H}{n_H+n_\Pi}$, the initial value, $\hat{I}^{(0)}_{BS,O}$, and the probability distribution, H.

Initial Value, $\hat{I}^{(0)}_{BS,O}$

If we choose $\hat{I}^{(0)}_{BS,O} = 0$, then the first value in the iterative scheme, $\hat{I}^{(1)}_{BS,O}$, corresponds to the reciprocal importance sampling approximation. Similarly, if $\hat{I}^{(0)}_{BS,O} = \infty$, then $\hat{I}^{(1)}_{BS,O}$ corresponds to the importance sampling approximation. Both of these options seem sensible, and in practice the iterative scheme (4.1) converges very quickly for any sensible starting value, $\hat{I}^{(0)}_{BS,O}$.

Probability Distribution, H

Specifying the probability distribution, H, is the most important issue in the practical implementation of bridge sampling. Meng and Schilling (2002) point out how $\text{var}(\hat{I}_{BS,O})$, given in (2.25), depends on the *Hellinger distance*, $\mathcal{H}(H,\Pi) \in [0,1]$, between H and Π defined as

$$\begin{aligned}\mathcal{H}(H,\Pi) &= \frac{1}{2}\int_\Theta \left(\sqrt{h(\boldsymbol{\theta})} - \sqrt{\pi(\boldsymbol{\theta})}\right)^2 \mathrm{d}\boldsymbol{\theta}, \\ &= 1 - \int_\Theta \sqrt{h(\boldsymbol{\theta})\pi(\boldsymbol{\theta})}\mathrm{d}\boldsymbol{\theta}, \\ &= 1 - \mathcal{B}(H,\Pi),\end{aligned}$$

where $\mathcal{B}(H,\Pi) = \int_\Theta \sqrt{h(\boldsymbol{\theta})\pi(\boldsymbol{\theta})}\mathrm{d}\boldsymbol{\theta} \in [0,1]$ is the *Bhattacharyya measure of affinity* between H and Π. The Hellinger distance is minimised, and, equivalently, the Bhattacharyya measure is maximised, when $h(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta})$, so we require H to 'mimic' Π as closely as possible. This is a direct analogy of how importance sampling, rejection sampling and the independence sampler all perform best when the sampling distributions H, S, and S, respectively, all 'mimic' the target distribution.

Suppose $\boldsymbol{\theta} \in \mathbb{R}^k$. DiCiccio et al. (1997) suggest taking H to be a normal approximation to Π. If $\boldsymbol{\theta}$ does not lie in $\mathbb{R}^k$ then we can take some other probability distribution approximation to Π. Indeed, Congdon (2003) suggests splitting the model parameters into sets of regression parameters, variance components, dispersion parameters, etc.

A different approach is *warp bridge sampling* (Meng and Schilling (2002)) where $H \equiv \text{N}(\mathbf{0}, \mathbf{I}_k)$ (or $H \equiv t_\nu(\mathbf{0}, \mathbf{I}_k)$) and Π is transformed or 'warped' to $\tilde{\Pi}$ so that its properties approximately match those of H. Here $t_\nu(\mathbf{0}, \mathbf{I}_k)$ denotes the k-variate t distribution with ν degrees of

freedom, mean $\mathbf{0}$ and variance matrix $\mathbf{I}_k$. It has pdf

$$h(\boldsymbol{\theta}) = \frac{\Gamma\left(\frac{\nu+k}{2}\right)}{(\pi(\nu-2))^{\frac{k}{2}}\Gamma\left(\frac{\nu}{2}\right)}\left[1+\frac{\boldsymbol{\theta}^T\boldsymbol{\theta}}{\nu-2}\right]^{-\frac{\nu+k}{2}},$$

for $\nu > 2$. If the location of Π matches that of H then this is known as Warp I bridge sampling, if the location and spread match then it is known as Warp II bridge sampling, and if the location, spread and skewness match then it is known as Warp III bridge sampling. If $H \equiv \mathrm{N}(\mathbf{0}, \mathbf{I}_k)$, then Warp II bridge sampling can be seen as being equivalent to the approach of DiCiccio et al. (1997). Sinharay and Stern (2005) found that Warp III bridge sampling provided the most accurate approximations to the marginal likelihood from all of the methods they assessed (see Section 2.2.7).

Suppose $\boldsymbol{\theta} \sim \Pi$, where the location and spread of Π are $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma} = \mathbf{S}\mathbf{S}^T$. We warp Π to $\tilde{\Pi}$ using the following stochastic transformation

$$b\mathbf{S}^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu}),$$

where b is Bernoulli $\left(\frac{1}{2}\right)$ on the sample space $\{-1, 1\}$. Now the pdf of $\tilde{\Pi}$ is

$$\begin{aligned}
\tilde{\pi}(\boldsymbol{\theta}) &= \frac{1}{2}|\mathbf{S}|\left[\pi(\boldsymbol{\mu} - \mathbf{S}\boldsymbol{\theta}) + \pi(\boldsymbol{\mu} + \mathbf{S}\boldsymbol{\theta})\right], \\
&= \frac{\frac{1}{2}|\mathbf{S}|\left[g(\boldsymbol{\mu} - \mathbf{S}\boldsymbol{\theta}) + g(\boldsymbol{\mu} + \mathbf{S}\boldsymbol{\theta})\right]}{\int_\Theta g(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}}, \\
&= \frac{\tilde{g}(\boldsymbol{\theta})}{\int_\Theta g(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}},
\end{aligned}$$

where $\tilde{g}(\boldsymbol{\theta}) = \frac{1}{2}|\mathbf{S}|\left[g(\boldsymbol{\mu} - \mathbf{S}\boldsymbol{\theta}) + g(\boldsymbol{\mu} + \mathbf{S}\boldsymbol{\theta})\right]$. Note that the normalising constant of $\tilde{g}(\boldsymbol{\theta})$ is the same as $g(\boldsymbol{\theta})$. It can be shown that the location, spread and skewness of $\tilde{\Pi}$ match those of H.

Let $\{\boldsymbol{\theta}_1^H, ..., \boldsymbol{\theta}_{n_H}^H\}$ and $\{\boldsymbol{\theta}_1^\Pi, ..., \boldsymbol{\theta}_{n_\Pi}^\Pi\}$ be samples generated from $H \equiv \mathrm{N}(\mathbf{0}, \mathbf{I}_k)$ (or $H \equiv t_\nu(\mathbf{0}, \mathbf{I}_k)$), and Π, respectively, then the Warp III bridge sampling approximation is found by iterating (4.1) until convergence is achieved, where

$$l_{Hi} = |\mathbf{S}|\frac{g(\boldsymbol{\mu} - \mathbf{S}\boldsymbol{\theta}_i^H) + g(\boldsymbol{\mu} + \mathbf{S}\boldsymbol{\theta}_i^H)}{2h(\boldsymbol{\theta}_i^H)}, \tag{4.2}$$

and

$$l_{\Pi i} = |\mathbf{S}|\frac{g(\boldsymbol{\theta}_i^\Pi) + g(2\boldsymbol{\mu} - \boldsymbol{\theta}_i^\Pi)}{2h(\mathbf{S}^{-1}(\boldsymbol{\theta}_i^\Pi - \boldsymbol{\mu}))}. \tag{4.3}$$

The $\boldsymbol{\mu}$ and $\mathbf{S}$ can be chosen to maximise the Bhattacharyya measure, $\mathcal{B}(H, \tilde{\Pi})$, between H and $\tilde{\Pi}$. This is equivalent to maximising

$$o(\boldsymbol{\mu}, \mathbf{S}) = \sqrt{|\mathbf{S}|}\mathrm{E}_H\left[\sqrt{\frac{g(\boldsymbol{\mu} - \mathbf{S}\boldsymbol{\theta}) + g(\boldsymbol{\mu} + \mathbf{S}\boldsymbol{\theta})}{h(\boldsymbol{\theta})}}\right]. \tag{4.4}$$

For most cases (4.4) will be analytically intractable. Meng and Schilling (2002) suggest generating $\{\tilde{\boldsymbol{\theta}}_1, ..., \tilde{\boldsymbol{\theta}}_m\}$ from H and maximising the sample average

$$\hat{o}(\boldsymbol{\mu}, \mathbf{S}) = \frac{\sqrt{|\mathbf{S}|}}{m} \sum_{i=1}^{m} \sqrt{\frac{g(\boldsymbol{\mu} - \mathbf{S}\tilde{\boldsymbol{\theta}}_i) + g(\boldsymbol{\mu} + \mathbf{S}\tilde{\boldsymbol{\theta}}_i)}{h(\tilde{\boldsymbol{\theta}}_i)}},$$

to find $\boldsymbol{\mu}$ and $\mathbf{S}$. For high dimensional problems that are typical for GLMMs, maximising $\hat{o}(\boldsymbol{\mu}, \mathbf{S})$ will be infeasible. Sinharay and Stern (2005) state that "empirical studies suggest good estimates of I even for suboptimal choices of warping transformation". They suggest taking $\boldsymbol{\mu}$ to be the mean or mode, and $\boldsymbol{\Sigma}$ to be the variance or curvature matrix of Π.

To summarise, we now present the two different approaches to finding a bridge sampling approximation to I when $\boldsymbol{\theta} \in \mathbb{R}^k$.

DiCiccio approach

1. Generate $\{\boldsymbol{\theta}_1^{\Pi}, ..., \boldsymbol{\theta}_{n_\Pi}^{\Pi}\}$ from Π.
2. Find approximations to $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma} = \mathbf{S}\mathbf{S}^T$, the mean/mode and variance/curvature matrix of Π, respectively.
3. Set $H \equiv \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and generate $\{\boldsymbol{\theta}_1^{H}, ..., \boldsymbol{\theta}_{n_H}^{H}\}$ from H.
4. Iterate (4.1) until convergence where $l_{ki} = g(\boldsymbol{\theta}_i^k)/h(\boldsymbol{\theta}_i^k)$ for $k = H, \Pi$.

Warp III bridge sampling approach

1. Generate $\{\boldsymbol{\theta}_1^{\Pi}, ..., \boldsymbol{\theta}_{n_\Pi}^{\Pi}\}$ from Π.
2. Find approximations to $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma} = \mathbf{S}\mathbf{S}^T$, the mean/mode and variance/curvature matrix of Π, respectively.
3. Set $H \equiv \mathrm{N}(\mathbf{0}, \mathbf{I}_k)$ (or $H \equiv t_\nu(\mathbf{0}, \mathbf{I}_k)$) and generate $\{\boldsymbol{\theta}_1^{H}, ..., \boldsymbol{\theta}_{n_H}^{H}\}$ from H.
4. Iterate (4.1) until convergence where l_{ki} for $k = H, \Pi$ are given by (4.2) and (4.3).

What is clear from the approaches of DiCiccio et al. (1997) and warp bridge sampling is that we need to have some information about Π in order to construct H or to warp Π, e.g. approximations to the mode and curvature, or mean and variance. We feel that finding the mode and curvature at the mode by maximising $\log g(\boldsymbol{\theta})$ and evaluating the Hessian matrix of $\log g(\boldsymbol{\theta})$ at the mode, respectively, will not fully describe Π, especially when Π is the posterior distribution of a GLMM. We use an approach of Sinharay and Stern (2005) who generate a preliminary MCMC sample from Π and then use this to set $\boldsymbol{\mu}$ to be the sample mean and $\boldsymbol{\Sigma} = \mathbf{S}\mathbf{S}^T$ to be the sample variance matrix. This approach of using a posterior

sample to gain information about the posterior distribution is also recommended by Gelfand and Dey (1994) and Congdon (2003). We investigate this issue of preferring the sample mean and variance over the mode and curvature in Section 4.6.

A naive approach would be to use the sample statistics from $\{\boldsymbol{\theta}_1^{\Pi}, ..., \boldsymbol{\theta}_{n_\Pi}^{\Pi}\}$, i.e. the same sample from Π used in the bridge sampler. This appears to lead to an underestimation of I. We noted in Section 2.2.5, if H is independent of $\{\boldsymbol{\theta}_1^{\Pi}, ..., \boldsymbol{\theta}_{n_\Pi}^{\Pi}\}$, then bridge sampling overestimates I but it appears that if H is dependent on $\{\boldsymbol{\theta}_1^{\Pi}, ..., \boldsymbol{\theta}_{n_\Pi}^{\Pi}\}$, then bridge sampling underestimates I.

Consider the following example where Π is the univariate uniform distribution. Here, θ does not lie in $\mathbb{R}$ but this example allows us to show, analytically, that dependence between H and the sample generated from Π leads to underestimation of I. Suppose $n_H = n_\Pi = n$, and that Π is $\mathrm{U}[0, a]$, therefore

$$g(\theta) = \begin{cases} 1, & \text{if } 0 \leq \theta \leq a, \\ 0, & \text{if otherwise,} \end{cases}$$

and $I = a$. We generate $\{\theta_1^{\Pi}, ..., \theta_n^{\Pi}\}$ from $\mathrm{U}[0, a]$ and set H to be $\mathrm{U}[0, \hat{\theta}^{\Pi}]$ where $\hat{\theta}^{\Pi}$, is some estimate of a based on $\{\theta_1^{\Pi}, ..., \theta_n^{\Pi}\}$. Therefore,

$$h(\theta) = \begin{cases} \frac{1}{\hat{\theta}^{\Pi}}, & \text{if } 0 \leq \theta \leq \hat{\theta}^{\Pi}, \\ 0, & \text{if otherwise.} \end{cases}$$

Now in the bridge sampling approximation (4.1), $l_{Hi} = \hat{\theta}^{\Pi} I(\theta_i^H \leq a)$ and $l_{\Pi i} = \frac{\hat{\theta}^{\Pi}}{I(\theta_i^{\Pi} \leq \hat{\theta}^{\Pi})}$, so the optimal, bridge sampling approximation is

$$\begin{aligned} \hat{I}_{BS,O} &= \frac{\frac{1}{n}\sum_{i=1}^{n} \frac{\hat{\theta}^{\Pi} I(\theta_i^H \leq a)}{\hat{\theta}^{\Pi} I(\theta_i^H \leq a) + I}}{\frac{1}{n}\sum_{i=1}^{n} \left(\frac{\hat{\theta}^{\Pi}}{I(\theta_i^{\Pi} \leq \hat{\theta}^{\Pi})} + I \right)^{-1}}, \\ &= \hat{\theta}^{\Pi} \frac{\sum_{i=1}^{n} I(\theta_i^H \leq a) \frac{1}{\hat{\theta}^{\Pi} + I}}{\sum_{i=1}^{n} I(\theta_i^{\Pi} \leq \hat{\theta}^{\Pi}) \frac{1}{\hat{\theta}^{\Pi} + I}}, \\ &= \hat{\theta}^{\Pi} \frac{\sum_{i=1}^{n} I(\theta_i^H \leq a)}{\sum_{i=1}^{n} I(\theta_i^{\Pi} \leq \hat{\theta}^{\Pi})}, \end{aligned}$$

where $\{\theta_1^H, ..., \theta_n^H\}$ is a sample generated from $H \equiv \mathrm{U}[0, \hat{\theta}^{\Pi}]$. Note that, the $\hat{I}_{BS,O}^{(t)}$ that appear in the numerator and denominator of the bridge sampling approximation cancel for this problem and $\hat{I}_{BS,O}$ is non-iterative.

We choose two alternatives for $\hat{\theta}^{\Pi}$: $\hat{\theta}_1^{\Pi} = \max_{i=1,...,n}\{\theta_i^{\Pi}\}$, i.e. the maximum likelihood estimate of a which is biased, and $\hat{\theta}_2^{\Pi} = \frac{n+1}{n} \max_{i=1,...,n}\{\theta_i^{\Pi}\}$, i.e. the adjusted maximum likelihood estimate of a which is unbiased. Denote the bridge sampling approximations to I that result from the two alternatives, $\hat{\theta}_1^{\Pi}$ and $\hat{\theta}_2^{\Pi}$, as $\hat{I}_{BS,O,1}$ and $\hat{I}_{BS,O,2}$, respectively, then

$$\hat{I}_{BS,O,1} = \hat{\theta}_1^{\Pi},$$

and

$$\hat{I}_{BS,O,2} = \frac{\hat{\theta}_2^{\Pi}}{n} \sum_{i=1}^{n} I(\theta_i^H \leq a).$$

We can use standard results on the distribution of a maximum to show that

$$\mathrm{E}(\hat{I}_{BS,O,1}) = \left(1 - \frac{1}{n+1}\right) a.$$

For $\mathrm{E}(\hat{I}_{BS,O,2})$, first note that

$$\begin{aligned}
\mathrm{E}(\hat{I}_{BS,O,2}) &= \mathrm{E}\left(\mathrm{E}\left(\hat{I}_{BS,O,2}|\theta_1^{\Pi}, ..., \theta_n^{\Pi}\right)\right), \\
&= \mathrm{E}\left(\hat{\theta}_2^{\Pi}\mathrm{P}(\theta_i^H \leq a)\right), \\
&= \mathrm{E}\left(\min(\hat{\theta}_2^{\Pi}, a)\right).
\end{aligned}$$

Again, using standard results on the distribution of a maximum and a simple transformation of variables, we find that $\hat{\theta}_2^{\Pi}$ has pdf, $f_{\hat{\theta}_2^{\Pi}}(\theta) = \frac{n^{n+1}\theta^{n-1}}{a^n(n+1)^n}$, for $0 \leq \theta \leq \frac{n+1}{n}a$. Therefore,

$$\begin{aligned}
\mathrm{E}(\hat{I}_{BS,O,2}) &= \mathrm{E}\left(\min(\hat{\theta}_2^{\Pi}, a)\right), \\
&= \int_0^a \theta f_{\hat{\theta}_2^{\Pi}}(\theta)\mathrm{d}\theta + a\int_a^{\frac{n+1}{n}a} f_{\hat{\theta}_2^{\Pi}}(\theta)\mathrm{d}\theta, \\
&= \left(1 - \frac{n^n}{(n+1)^{n+1}}\right) a.
\end{aligned}$$

Now $\mathrm{E}(\hat{I}_{BS,O,j}) < a$ for finite n and $j = 1, 2$. So, for this problem, a naive use of bridge sampling leads to an underestimation of I by $O(\frac{1}{n})$.

We can generalise the above result into k dimensions. Suppose $\Pi \equiv \mathrm{U}[\mathbf{0}, \mathbf{a}]$, where $\mathbf{a} = (a_1, ..., a_k)^T$, i.e. Π is uniform on the k-dimensional cuboid, $G = [0, a_1] \times [0, a_2] \times ... \times [0, a_k]$, defined by the points $\mathbf{0}$ and $\mathbf{a}$. Therefore,

$$g(\boldsymbol{\theta}) = \begin{cases} 1, & \text{if } 0 \in G, \\ 0, & \text{if otherwise,} \end{cases}$$

and $I = \prod_{j=1}^k a_j$. We generate $\{\boldsymbol{\theta}_1^{\Pi}, ..., \boldsymbol{\theta}_n^{\Pi}\}$ from Π and set $H \equiv \mathrm{U}[\mathbf{0}, \hat{\boldsymbol{\theta}}]$ where $\hat{\boldsymbol{\theta}} = (\hat{\theta}^{\Pi(1)}, ..., \hat{\theta}^{\Pi(k)})^T = \left(\frac{n+1}{n}\max\{\theta_i^{\Pi(1)}\}, ..., \frac{n+1}{n}\max\{\theta_i^{\Pi(k)}\}\right)^T$ and $\theta_i^{\Pi(j)}$ is the jth element of $\boldsymbol{\theta}_i^{\Pi}$. Therefore,

$$h(\boldsymbol{\theta}) = \begin{cases} \frac{1}{\prod_{j=1}^k \hat{\theta}^{\Pi(j)}}, & \text{if } \boldsymbol{\theta} \in A = [0, \hat{\theta}^{\Pi(1)}] \times ... \times [0, \hat{\theta}^{\Pi(k)}], \\ 0, & \text{if otherwise.} \end{cases}$$

It can be shown that

$$\hat{I}_{BS,O} = \frac{1}{n}\prod_{j=1}^k \hat{\theta}^{\Pi(k)} \sum_{i=1}^n I(\boldsymbol{\theta}_i^H \in G),$$

and that

$$\begin{aligned}
\mathrm{E}\left(\hat{I}_{BS,O}\right) &= \left(1 - \frac{n^n}{(n+1)^{n+1}}\right)^k \prod_{j=1}^k a_j, \\
&= \left(1 - \frac{n^n}{(n+1)^{n+1}}\right)^k I.
\end{aligned}$$

Therefore in k dimensions, naive use of bridge sampling results in an underestimation of I by $O(\frac{k}{n})$.

Obtaining analytical results, as above, for non-trivial cases is difficult so we rely on a simulation study. For the simulation study $n_H = n_\Pi = n$ and we use, for Π, the k-variate normal distribution, $\mathrm{N}(\mathbf{0}, \mathbf{I}_k)$, with mean $\mathbf{0}$ and variance matrix $\mathbf{I}_k$. Here

$$g(\boldsymbol{\theta}) = \exp\left(-\frac{\boldsymbol{\theta}^T\boldsymbol{\theta}}{2}\right),$$

and $I = (2\pi)^{\frac{k}{2}}$. We generate two samples, $\{\boldsymbol{\theta}_1^\Pi, ..., \boldsymbol{\theta}_n^\Pi\}$ and $\{\boldsymbol{\theta}_1^{\Pi*}, ..., \boldsymbol{\theta}_n^{\Pi*}\}$, from $\Pi \equiv \mathrm{N}(\mathbf{0}, \mathbf{I}_k)$ and set $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma} = \mathbf{S}\mathbf{S}^T$ to be the sample mean and sample variance matrix of $\{\boldsymbol{\theta}_1^\Pi, ..., \boldsymbol{\theta}_n^\Pi\}$. We compute four approximations to I:

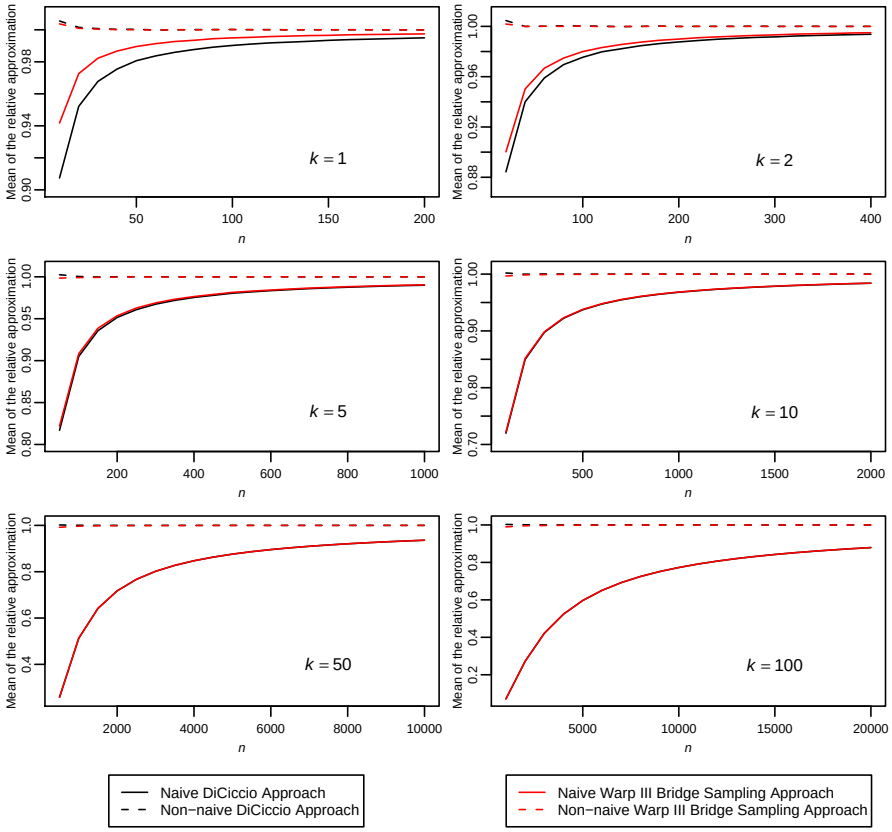
1. The Naive DiCiccio Approach, $\{\boldsymbol{\theta}_1^\Pi, ..., \boldsymbol{\theta}_n^\Pi\}$ is used in the bridge sampler as the sample from Π and is therefore not independent of H.
2. The Naive Warp III Bridge Sampling Approach, $\{\boldsymbol{\theta}_1^\Pi, ..., \boldsymbol{\theta}_n^\Pi\}$ is used in the bridge sampler as the sample from Π and is therefore not independent of the warping transformation. We use $H \equiv \mathrm{N}(\mathbf{0}, \mathbf{I}_k)$.
3. The Non-naive DiCiccio Approach, $\{\boldsymbol{\theta}_1^{\Pi*}, ..., \boldsymbol{\theta}_n^{\Pi*}\}$ is used in the bridge sampler as the sample from Π and is therefore independent of H.
4. The Non-naive Warp III Bridge Sampling Approach, $\{\boldsymbol{\theta}_1^{\Pi*}, ..., \boldsymbol{\theta}_n^{\Pi*}\}$ is used in the bridge sampler as the sample from Π and is therefore independent of the warping transformation. We use $H \equiv \mathrm{N}(\mathbf{0}, \mathbf{I}_k)$.

We choose six different values for k, namely $1, 2, 5, 10, 50$ and 100. For $k = 1$, we approximate I for $n \in \{10, 20, ..., 190, 200\}$, for $k = 2$, $n \in \{20, 40, ..., 380, 400\}$, for $k = 5$, $n \in \{50, 100, ..., 950, 1000\}$, for $k = 10$, $n \in \{100, 200, ..., 1900, 2000\}$, for $k = 50$, $n \in \{500, 1000, ..., 9500, 10000\}$, and for $k = 100$, $n \in \{1000, 2000, ..., 19000, 20000\}$. For each value of n we repeat the approximation 10000 times, with different samples from Π and H for each repetition. Figure 4.1 shows the mean of the relative approximation, $\hat{I}_{BS,O}/I$, over the 10000 repetitions plotted against n for the six different values of k and the four different approaches.

Figure 4.1 shows that the naive approaches underestimate I for both the DiCiccio and Warp III approaches. This underestimation appears to be asymptotically zero. The non-naive approaches do not lead to any such underestimation. The non-naive approaches should lead to overestimation of I, which is asymptotically zero, but this overestimation is negligible when compared to the underestimation caused by the naive approaches. For $k > 1$, the results from the simulation study seem to concur with the analytic results, with a k-variate uniform Π, that the underestimation is $O(\frac{k}{n})$.

We can conclude from the simulation study, that we cannot use the same sample to construct H or to warp Π as we use in the bridge sampler. However, the non-naive DiCiccio and Warp

Figure 4.1: Mean of the relative approximation over the 10000 repetitions plotted against n for the six different values of k and the four different approaches.



III bridge sampling approaches require more computational effort since we need to generate two samples from Π.

Warp III bridge sampling requires no extra sampling effort over the DiCiccio method and is reported by Sinharay and Stern (2005) to provide more accurate approximations. For these reasons we focus on the Warp III bridge sampling implementation over the DiCiccio approach.

We consider the following scenario: we can generate a sample of size N from each of Π and H, denoted $\{\boldsymbol{\theta}_1^{\Pi}, ..., \boldsymbol{\theta}_N^{\Pi}\}$ and $\{\boldsymbol{\theta}_1^{H}, ..., \boldsymbol{\theta}_N^{H}\}$, respectively. Our present problem is to find how to allocate the $\{\boldsymbol{\theta}_1^{\Pi}, ..., \boldsymbol{\theta}_N^{\Pi}\}$ to find $\boldsymbol{\mu}$ and $\mathbf{S}$ as well as to use in the bridge sampler. We consider two strategies and assess their performance using the mean squared error of the relative approximation from a simulation study.

Proportion Strategy

We use a proportion, $\rho \in (0, 1)$, of the sample from Π with which to warp Π and the remainder is used in the bridge sampler. We use the whole sample from H in the bridge sampler. Therefore, $n_H = N$ and $n_{\Pi} = N - \lceil \rho N \rceil$. Denote this approximation to I as $\hat{I}_{BS,O}^{(\rho)}$, for a particular value of ρ. The algorithm for the *proportion strategy* is

1. Generate $\{\boldsymbol{\theta}_1^{\Pi}, ..., \boldsymbol{\theta}_N^{\Pi}\}$ from the target distribution Π and $\{\boldsymbol{\theta}_1^{H}, ..., \boldsymbol{\theta}_N^{H}\}$ from H.
2. Let $n_{\Pi} = N - \lceil \rho N \rceil$ and $n_H = N$.
3. Let $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma} = \mathbf{S}\mathbf{S}^T$ be the sample mean and variance of $\{\boldsymbol{\theta}_{n_{\Pi}+1}^{\Pi}, ..., \boldsymbol{\theta}_N^{\Pi}\}$, respectively.
4. Compute l_{Hi} using (4.2) for $i = 1, ..., n_H$ and $l_{\Pi i}$ using (4.3) for $i = 1, ..., n_{\Pi}$.
5. Find $\hat{I}_{BS,O}^{(\rho)}$ using (4.1).

Split Strategy
We split $\{\boldsymbol{\theta}_1^{\Pi}, ..., \boldsymbol{\theta}_N^{\Pi}\}$ into two unique, equally sized samples denoted by $\{\boldsymbol{\theta}_1^{\Pi(1)}, ..., \boldsymbol{\theta}_{n_{\Pi}}^{\Pi(1)}\}$ and $\{\boldsymbol{\theta}_1^{\Pi(2)}, ..., \boldsymbol{\theta}_{n_{\Pi}}^{\Pi(2)}\}$ where $n_{\Pi} = \frac{N}{2}$. We do the same with the sample from H to form $\{\boldsymbol{\theta}_1^{H(1)}, ..., \boldsymbol{\theta}_{n_{\Pi}}^{H(1)}\}$ and $\{\boldsymbol{\theta}_1^{H(2)}, ..., \boldsymbol{\theta}_{n_{\Pi}}^{H(2)}\}$ where $n_H = n_{\Pi} = \frac{N}{2}$. The approximation $\hat{I}_{BS,O}^{(1)}$ is found by using $\{\boldsymbol{\theta}_1^{\Pi(2)}, ..., \boldsymbol{\theta}_{n_{\Pi}}^{\Pi(2)}\}$ to find $\boldsymbol{\mu}$ and $\mathbf{S}$, and using the samples $\{\boldsymbol{\theta}_1^{\Pi(1)}, ..., \boldsymbol{\theta}_{n_{\Pi}}^{\Pi(1)}\}$ and $\{\boldsymbol{\theta}_1^{H(1)}, ..., \boldsymbol{\theta}_{n_{\Pi}}^{H(1)}\}$ in the bridge sampler. Likewise, $\hat{I}_{BS,O}^{(2)}$ is found by using $\{\boldsymbol{\theta}_1^{\Pi(1)}, ..., \boldsymbol{\theta}_{n_{\Pi}}^{\Pi(1)}\}$ to find $\boldsymbol{\mu}$ and $\mathbf{S}$, and using $\{\boldsymbol{\theta}_1^{\Pi(2)}, ..., \boldsymbol{\theta}_{n_{\Pi}}^{\Pi(2)}\}$ and $\{\boldsymbol{\theta}_1^{H(2)}, ..., \boldsymbol{\theta}_{n_{\Pi}}^{H(2)}\}$ in the bridge sampler. The two approximations are then combined to form $\hat{I}_{BS,O}^{(S,A)}$ or $\hat{I}_{BS,O}^{(S,G)}$, where the A or the G is used to denote whether $\hat{I}_{BS,O}^{(1)}$ and $\hat{I}_{BS,O}^{(2)}$ have been combined using the arithmetic or geometric mean, respectively.

The algorithm for the *split strategy* is:

1. Generate $\{\boldsymbol{\theta}_1^{\Pi}, ..., \boldsymbol{\theta}_N^{\Pi}\}$ from the target distribution Π and $\{\boldsymbol{\theta}_1^{H}, ..., \boldsymbol{\theta}_N^{H}\}$ from H.
2. Let $n_{\Pi} = n_H = \frac{N}{2}$.
3. Let $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma} = \mathbf{S}\mathbf{S}^T$ be the sample mean and variance of $\{\boldsymbol{\theta}_1^{\Pi}, ..., \boldsymbol{\theta}_{n_{\Pi}}^{\Pi}\}$, respectively.
4. Compute l_{Hi} using (4.2) for $i = n_H + 1, ..., N$ and $l_{\Pi i}$ using (4.3) for $i = n_{\Pi} + 1, ..., N$.
5. Let $\hat{I}_{BS,O}^{(1)}$ be the final value of the following converged iterative scheme

$$\hat{I}_{BS,O}^{(t+1)} = \frac{\sum_{i=n_H+1}^{N} \frac{l_{Hi}}{n_{\Pi} l_{Hi} + n_H \hat{I}_{BS,O}^{(t)}}}{\sum_{i=n_{\Pi}+1}^{N} \frac{1}{n_{\Pi} l_{\Pi i} + n_H \hat{I}_{BS,O}^{(t)}}}.$$

6. Let $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ be the sample mean and variance of $\{\boldsymbol{\theta}_{n_{\Pi}+1}^{\Pi}, ..., \boldsymbol{\theta}_N^{\Pi}\}$, respectively.
7. Compute l_{Hi} using (4.2) for $i = 1, ..., n_H$ and $l_{\Pi i}$ using (4.3) for $i = 1, ..., n_{\Pi}$.
8. Let $\hat{I}_{BS,O}^{(2)}$ be the final value of the following converged iterative scheme

$$\hat{I}_{BS,O}^{(t+1)} = \frac{\sum_{i=1}^{n_H} \frac{l_{Hi}}{n_{\Pi} l_{Hi} + n_H \hat{I}_{BS,O}^{(t)}}}{\sum_{i=1}^{n_{\Pi}} \frac{1}{n_{\Pi} l_{\Pi i} + n_H \hat{I}_{BS,O}^{(t)}}}.$$

9. Let $\hat{I}^{(S,A)}_{BS,O} = \frac{1}{2}\left(\hat{I}^{(1)}_{BS,O} + \hat{I}^{(2)}_{BS,O}\right)$ and $\hat{I}^{(S,G)}_{BS,O} = \sqrt{\hat{I}^{(1)}_{BS,O}\hat{I}^{(2)}_{BS,O}}$

In the simulation study we will also assess whether to use $H \equiv \mathrm{N}(\mathbf{0}, \mathbf{I}_k)$ or $H \equiv \mathrm{t}_\nu(\mathbf{0}, \mathbf{I}_k)$. Sinharay and Stern (2005) used $\nu = 4$ and we will do so also. They found that for Warp II bridge sampling using a t-distribution over a normal distribution for H led to an improvement in accuracy. However, they concluded that this was not the case for Warp III bridge sampling. Sinharay and Stern (2005) believed that this was the case due to a t-distribution having more overlap, and therefore higher Bhattacharyya measure, with the still skewed $\tilde{\Pi}$, under the Warp II approach.

We consider four different target distributions:

1. $\Pi_1 \equiv \mathrm{N}(\mathbf{1}_k, 2\mathbf{I}_k)$, i.e. the k-variate normal distribution with mean $\mathbf{1}_k$ and variance matrix $2\mathbf{I}_k$, where $\mathbf{1}_k$ denotes the $k \times 1$ vector of ones.

2. $\Pi_2 \equiv \mathrm{C}(\mathbf{1}_k)$, i.e. the k-variate non-central Cauchy distribution with location $\mathbf{1}_k$. This is a special case of the k-variate non-central t-distribution with mean $\mathbf{1}_k$ and one degree of freedom. This distribution is heavy-tailed.

3. $\Pi_3 \equiv \mathrm{L}(\mathbf{0}, \mathbf{I}_k)$, i.e. the k-variate logistic distribution with mean $\mathbf{0}$ and scale matrix $\mathbf{I}_k$. This distribution is formed from k independent logistic distributions with mean 0 and scale parameter 1. This distribution is heavy-tailed.

4. $\Pi_4 \equiv \mathrm{LG}(\mathbf{1}_k, 4\mathbf{1}_k)$, i.e. the k-variate log gamma distribution with shape parameter $\mathbf{1}_k$ and scale parameter $4\mathbf{1}_k$. This distribution is formed from k independent log gamma distributions with shape parameter 1 and scale parameter 4. This distribution is skewed.

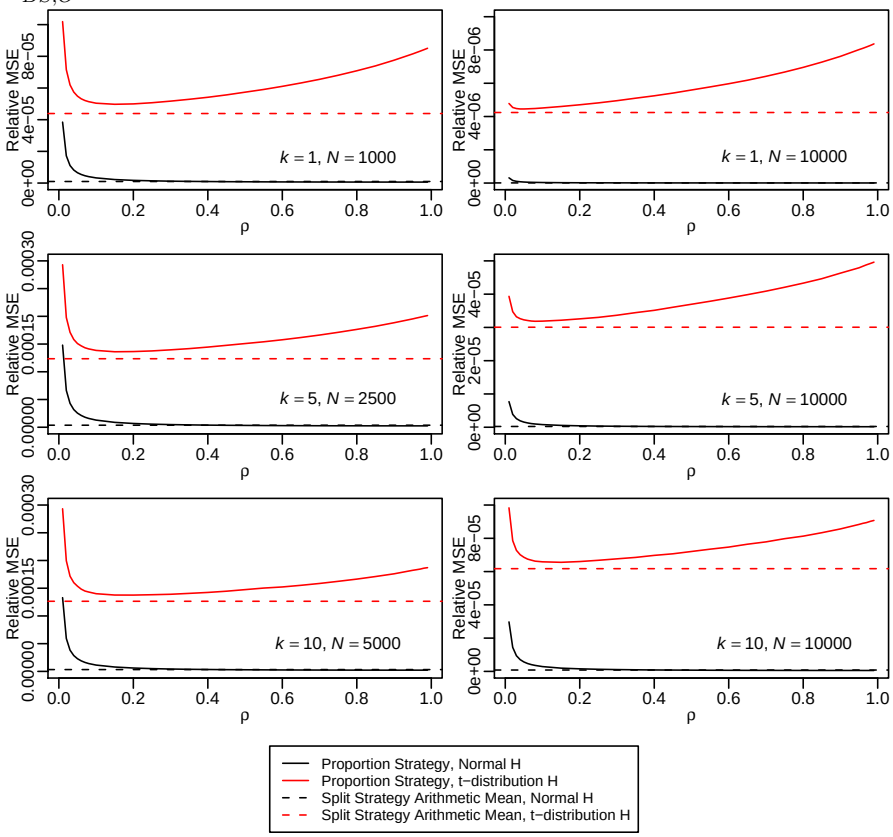
We generate samples from Π and H of size N and calculate $\hat{I}^{(\rho)}_{BS,O}$ for each ρ in $\{0.05, ..., 0.95\}$ as well as $\hat{I}^{(S,A)}_{BS,O}$ and $\hat{I}^{(S,G)}_{BS,O}$. We repeat this 100000 times for $H \equiv \mathrm{N}(\mathbf{0}, \mathbf{I}_k)$ and $H \equiv \mathrm{t}_\nu(\mathbf{0}, \mathbf{I}_k)$. We choose $k = 1, 5$ and 10. For $k = 1$, we use $N = 1000$ and 10000. For $k = 5$, $N = 2500$ and 10000. For $k = 10$, $N = 5000$ and 10000.

Figures 4.2, 4.3, 4.4 and 4.5 show the relative mean squared error of $\hat{I}^{(\rho)}_{BS,O}$ plotted against ρ for Π_1, Π_2, Π_3 and Π_4, respectively, for the different values of k and N. Also shown on the plots is the relative mean squared error of $\hat{I}^{(S,A)}_{BS,O}$. We found no improvement in accuracy from using the geometric mean over the arithmetic mean in the split strategy approach.

The first conclusion to draw is that typically using $\mathrm{N}(\mathbf{0}, \mathbf{I}_k)$ for H appears to outperform using $\mathrm{t}_4(\mathbf{0}, \mathbf{I}_k)$ with respect to minimising the relative mean squared error for all target distributions except the Cauchy distribution. Even in these cases, the relative mean squared errors are very similar when using $H \equiv \mathrm{N}(\mathbf{0}, \mathbf{I}_k)$ compared to $H \equiv \mathrm{t}_4(\mathbf{0}, \mathbf{I}_k)$.

The optimal value of ρ with respect to minimising the relative mean squared error is heavily dependent on Π, H, k and N. When the target distribution is normal and H is also normal then the optimal ρ is large suggesting we should use the vast majority of the sample from Π to

Figure 4.2: Relative MSE of $\hat{I}^{(\rho)}_{BS,O}$ plotted against ρ for the $\Pi_1 \equiv \mathrm{N}(\mathbf{1}_k, 2\mathbf{I}_k)$ target distribution with the relative MSE of $\hat{I}^{(S,A)}_{BS,O}$.



find $\boldsymbol{\mu}$ and $\mathbf{S}$. In this case, the distributional family of H and $\tilde{\Pi}$ is identical, so by increasing the accuracy of the approximations, $\boldsymbol{\mu}$ and $\mathbf{S}$, we increase the Bhattacharyya measure towards 1 and therefore decrease the variance of the bridge sampling approximation.

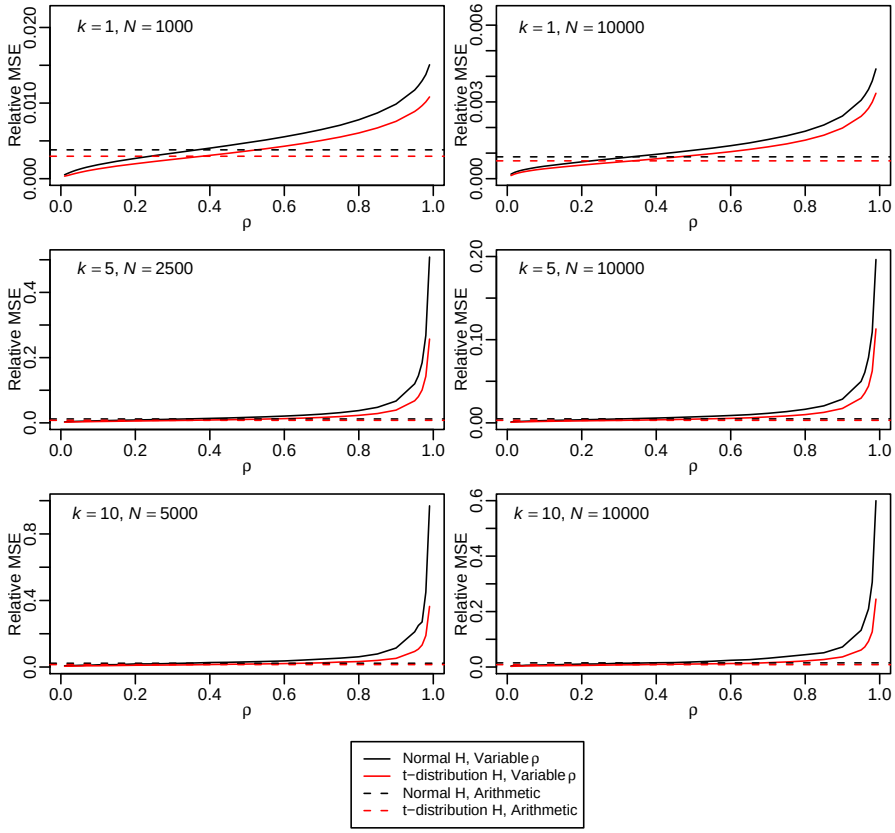
However, when the target distribution is Cauchy and H is normal, the optimal ρ is small indicating that we need to use the vast majority of the sample from Π in the bridge sampler. For the logistic and log-gamma target distributions, the optimal ρ appears to lie in the interval $(0.1, 0.3)$.

It appears that a ρ that is optimal for one type of target distribution can be disastrous for another type of target distribution.

In practice, posterior distributions are asymptotically normal and are approximately normal for large sample sizes so it follows that using a large value for ρ could be a sensible approach. However, as mentioned above this can be disastrous. A conservative choice would be use $\rho = 0.5$, as this value seems to perform well for most scenarios.

Alternatively, consider the split strategy with the arithmetic mean. This strategy typically outperforms the proportion strategy for any ρ except when the target distribution is normal and in this case it performs close to the proportion strategy with optimal ρ.

Figure 4.3: Relative MSE of $\hat{I}_{BS,O}^{(\rho)}$ plotted against ρ for the $\Pi_2 \equiv \mathrm{C}(\mathbf{1}_k)$ target distribution with the relative MSE of $\hat{I}_{BS,O}^{(S,A)}$.

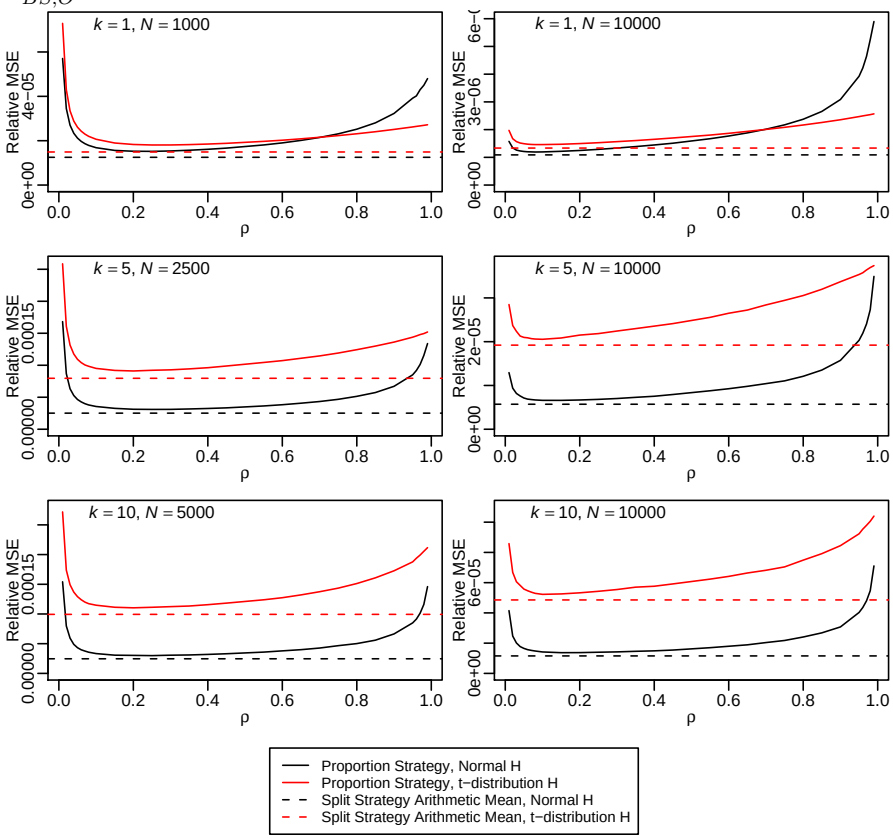


Therefore our recommendations are to use $\mathrm{N}(\mathbf{0}, \mathbf{I}_k)$ as H and to use the split strategy with the arithmetic mean.

Allocation of Sample Sizes

With regards to the allocation of sample sizes, these are set by the strategy we have adopted from above and the relative ease of generating from the distributions H and Π. We have assumed that the two distributions, H and Π, are equally easy to generate from, although this may not be the case in practice. If it is easier to generate from H, then we can use unequal sample sizes of $n_k = \frac{1}{2}N_k$, where N_k is the size of the sample generated from k, for $k = H, \Pi$.

Figure 4.4: Relative MSE of $\hat{I}^{(\rho)}_{BS,O}$ plotted against ρ for the $\Pi_3 \equiv \mathrm{L}(\mathbf{0}, \mathbf{I}_k)$ target distribution with the relative MSE of $\hat{I}^{(S,A)}_{BS,O}$.



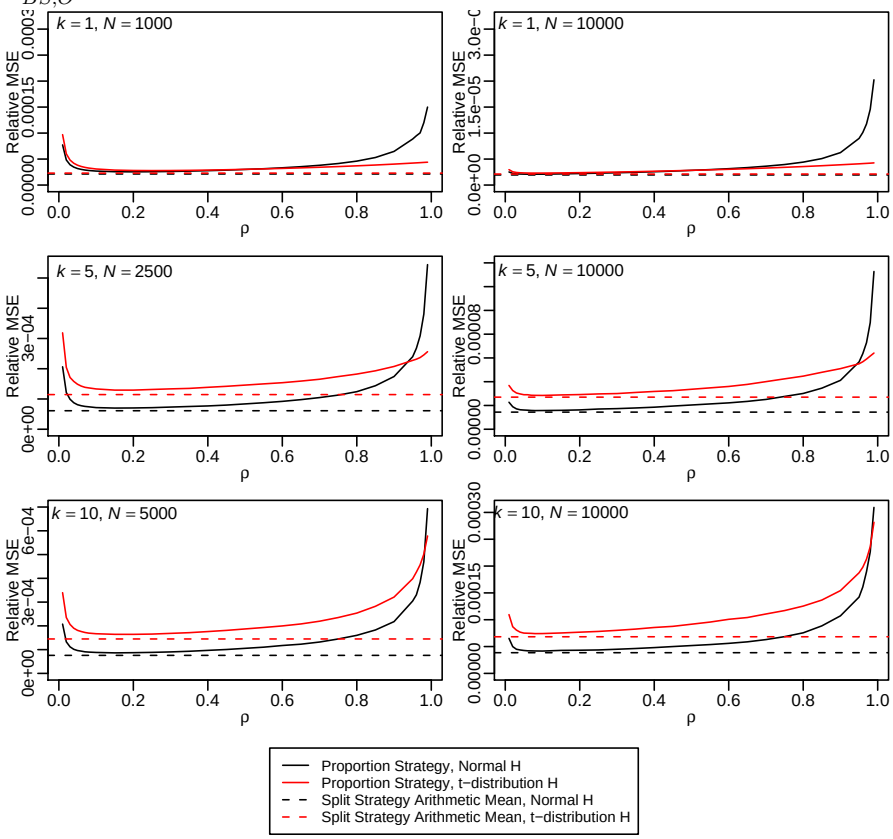
4.2.3 Summary

We summarise this Section by presenting the bridge sampling algorithm that we found best approximates I with respect to minimising the mean squared error.

1. Generate a sample, $\{\boldsymbol{\theta}^{\Pi}_1, ..., \boldsymbol{\theta}^{\Pi}_{N_\Pi}\}$, of size N_Π from the target distribution, Π, and a sample, $\{\boldsymbol{\theta}^{H}_1, ..., \boldsymbol{\theta}^{H}_{N_H}\}$, of size N_H from $H \equiv \mathrm{N}(\mathbf{0}, \mathbf{I}_k)$.
2. Let $n_\Pi = \frac{1}{2}N_\Pi$ and $n_H = \frac{1}{2}N_H$.
3. Let $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma} = \mathbf{S}\mathbf{S}^T$ be the sample mean and variance of $\{\boldsymbol{\theta}^{\Pi}_1, ..., \boldsymbol{\theta}^{\Pi}_{n_\Pi}\}$, respectively.
4. Compute l_{Hi} using (4.2) for $i = n_H + 1, ..., N_H$ and $l_{\Pi i}$ using (4.3) for $i = n_\Pi + 1, ..., N_\Pi$.
5. Let $\hat{I}^{(1)}_{BS,O}$ be the final value of the following converged iterative scheme

$$\hat{I}^{(t+1)}_{BS,O} = \frac{\frac{1}{n_H}\sum_{i=n_H+1}^{N_H} \frac{l_{Hi}}{n_\Pi l_{Hi} + n_H \hat{I}^{(t)}_{BS,O}}}{\frac{1}{n_\Pi}\sum_{i=n_\Pi+1}^{N_\Pi} \frac{1}{n_\Pi l_{\Pi i} + n_H \hat{I}^{(t)}_{BS,O}}}.$$

Figure 4.5: Relative MSE of $\hat{I}_{BS,O}^{(\rho)}$ plotted against ρ for the $\Pi_4 \equiv \text{LG}(\mathbf{1}_k, 4\mathbf{1}_k)$ target distribution with the relative MSE of $\hat{I}_{BS,O}^{(S,A)}$.



6. Let $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma} = \mathbf{S}\mathbf{S}^T$ be the sample mean and variance of $\{\boldsymbol{\theta}_{n_\Pi+1}^\Pi, ..., \boldsymbol{\theta}_{N_\Pi}^\Pi\}$, respectively.

7. Compute l_{Hi} using (4.2) for $i = 1, ..., n_H$ and $l_{\Pi i}$ using (4.3) for $i = 1, ..., n_\Pi$.

8. Let $\hat{I}_{BS,O}^{(2)}$ be the final value of the following converged iterative scheme

$$\hat{I}_{BS,O}^{(t+1)} = \frac{\frac{1}{n_H}\sum_{i=1}^{n_H} \frac{l_{Hi}}{n_\Pi l_{Hi} + n_H \hat{I}_{BS,O}^{(t)}}}{\frac{1}{n_\Pi}\sum_{i=1}^{n_\Pi} \frac{1}{n_\Pi l_{\Pi i} + n_H \hat{I}_{BS,O}^{(t)}}}.$$

9. Let $\hat{I}_{BS,O}^{(S,A)} = \frac{1}{2}\left(\hat{I}_{BS,O}^{(1)} + \hat{I}_{BS,O}^{(2)}\right)$.

4.3 Nested Sampling

4.3.1 Introduction

We introduced the basic idea of nested sampling in Section 2.2.5. A key feature of nested sampling is the requirement to generate from $H|L(\boldsymbol{\theta}) > L(\boldsymbol{\theta}_i)$, i.e. the distribution H

constrained to the region where $L(\boldsymbol{\theta}) > L(\boldsymbol{\theta}_i)$. Exact generation from this constrained distribution is typically an intractable problem.

Skilling (2006) proposed the use of p MCMC steps at iteration i with $H|L(\boldsymbol{\theta}) > L(\boldsymbol{\theta}_i)$ as the stationary distribution and with one of the $N-1$ elements of $\Theta_i^{(S)}$ as the initial value. Chopin and Robert (2009) point out that if this method is used then their central limit theorem result is invalid since the stationary distribution changes at each iteration. They go on to say that nested sampling based on MCMC could be interpreted as an approximation to ideal nested sampling.

An alternative approach is to generate a sample from H and then to subsample those values that satisfy $L(\boldsymbol{\theta}) > L(\boldsymbol{\theta}_i)$. However, this will become increasingly inefficient as $L(\boldsymbol{\theta}_i)$ becomes close to its maximum.

In the next Section, we discuss an extension of nested sampling called nested importance sampling which allows exact generation from the constrained distribution.

4.3.2 Nested Importance Sampling

First, suppose we are interested in evaluating

$$I_1 = \int_\Theta s(\boldsymbol{\theta})\tilde{L}(\boldsymbol{\theta})h(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta},$$

for some function, $s(\boldsymbol{\theta})$, and for any positive function, $\tilde{L}(\boldsymbol{\theta})$. Chopin and Robert (2009) show that we can use the following nested sampling approximation

$$I_{1,NS} = \sum_{i=1}^{m}(e^{-(i-1)/N} - e^{-i/N})\tilde{L}(\boldsymbol{\theta}_i)s(\boldsymbol{\theta}_i),$$

where the $\boldsymbol{\theta}_i$'s are the same as would be used to approximate $I_2 = \int_\Theta \tilde{L}(\boldsymbol{\theta})h(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}$, i.e. at iteration i, $\boldsymbol{\theta}_i = \arg\min\{\tilde{L}(\boldsymbol{\theta}_{i1}), ..., \tilde{L}(\boldsymbol{\theta}_{iN})\}$, and we are still required to generate from $H|\tilde{L}(\boldsymbol{\theta}) > \tilde{L}(\boldsymbol{\theta}_i)$.

Now, suppose $s(\boldsymbol{\theta}) = \frac{g(\boldsymbol{\theta})}{\tilde{L}(\boldsymbol{\theta})h(\boldsymbol{\theta})} = \frac{L(\boldsymbol{\theta})}{\tilde{L}(\boldsymbol{\theta})}$, then $I_1 = I = \int_\Theta g(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}$ and the nested importance sampling approximation to I is

$$\hat{I}_{NIS} = \sum_{i=1}^{m}(e^{-(i-1)/N} - e^{-i/N})L(\boldsymbol{\theta}_i).$$

We have a choice of the function $\tilde{L}(\boldsymbol{\theta})$. If we choose $\tilde{L}(\boldsymbol{\theta}) = L(\boldsymbol{\theta})$, then we have the basic nested sampling method from Section 2.2.5.

Suppose $H \equiv \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and

$$\tilde{L}(\boldsymbol{\theta}) = \lambda((\boldsymbol{\theta} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu})),$$

where $\lambda()$ is any decreasing function. In this case, Chopin and Robert (2009) show that the approximations to the $\Psi^{-1}(x_i)$'s can be error-free. In nested sampling, we approximate $\Psi^{-1}(x_i)$ by $\tilde{L}(\boldsymbol{\theta}_i)$ but we can now do this exactly. If $\Psi^{-1}(x_i) = \tilde{L}(\boldsymbol{\theta}_i)$, then

$$\begin{aligned} x_i &= \Psi\left(\tilde{L}(\boldsymbol{\theta}_i)\right), \\ &= P\left(\tilde{L}(\boldsymbol{\theta}) > \tilde{L}(\boldsymbol{\theta}_i)\right), \\ &= P\left((\boldsymbol{\theta}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu}) < (\boldsymbol{\theta}_i-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}_i-\boldsymbol{\mu})\right), \end{aligned}$$

since $\lambda()$ is a decreasing function. Therefore $x_i = F_{\chi^2_k}\left((\boldsymbol{\theta}_i-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}_i-\boldsymbol{\mu})\right)$, where $F_{\chi^2_k}()$ is the distribution function of the χ^2_k distribution. Let q_i be the x_ith quantile of the χ^2_k distribution, then $\Psi^{-1}(x_i) = \tilde{L}(\boldsymbol{\theta}_i)$ if and only if

$$q_i = (\boldsymbol{\theta}_i-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}_i-\boldsymbol{\mu}).$$

This can be achieved if

$$\boldsymbol{\theta}_i = \frac{\sqrt{q_i}\mathbf{S}\mathbf{v}}{\sqrt{\mathbf{v}^T\mathbf{v}}} + \boldsymbol{\mu},$$

where $\mathbf{v} \sim \mathrm{N}(\mathbf{0}, \mathbf{I}_k)$. We are sampling $\boldsymbol{\theta}_i$ uniformly over the ellipsoid that contains x_i of the mass of H.

Chopin and Robert (2009) suggest setting $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ to be the mode and the negative inverse of the Hessian matrix of $\log g(\boldsymbol{\theta})$ evaluated at the mode, respectively. As is the case for bridge sampling, we feel that the mode and curvature do not give us sufficient information about Π. However, as is the case for bridge sampling, we may assume that we have a sample of size N_Π from Π, and we can set $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ to be the sample mean and variance matrix, respectively.

The nested importance sampling approximation to $I = \int_\Theta g(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}$ is then found using the following algorithm:

1. Generate a sample of size N_Π from Π. Set $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma} = \mathbf{S}\mathbf{S}^T$ to be the mean and variance matrix of this sample, respectively. Set $i = 1$ and $\hat{I}^{(1)}_{NIS} = 0$. Choose N.

2. Generate $\mathbf{v} \sim \mathrm{N}(\mathbf{0}, \mathbf{I}_k)$ and set

$$\boldsymbol{\theta}_i = \frac{\sqrt{q_i}\mathbf{S}\mathbf{v}}{\sqrt{\mathbf{v}^T\mathbf{v}}} + \boldsymbol{\mu},$$

   where q_i is the $e^{-i/N}$ quantile of the χ^2_k distribution.

3. Let

$$\hat{I}^{(i+1)}_{NIS} = \hat{I}^{(i)}_{NIS} + (e^{-(i-1)/N} - e^{-i/N})\frac{g(\boldsymbol{\theta}_i)}{h(\boldsymbol{\theta}_i)},$$

   where $h()$ is the pdf of $H \equiv \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

4. Repeat steps 2. to 3. until $\hat{I}_{NIS}$ has converged.

The value N is a tuning parameter which when increased will result in an approximation with greater accuracy but an algorithm that will take longer to converge and thus will have greater computational expense.

Note that we can choose $H \equiv \mathrm{t}_\nu(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, for $\nu > 2$, and

$$\tilde{L}(\boldsymbol{\theta}) = \lambda\left(\frac{1}{k}(\boldsymbol{\theta}-\boldsymbol{\mu})^T\boldsymbol{\Upsilon}^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu})\right),$$

where $\lambda()$ is any decreasing function, and $\boldsymbol{\Upsilon} = \frac{\nu-2}{\nu}\boldsymbol{\Sigma}$. In this case our approximations to the $\Psi^{-1}(x_i)$'s are error-free if

$$\boldsymbol{\theta}_i = \frac{\sqrt{q_i k}\mathbf{R}\mathbf{v}}{\sqrt{\mathbf{v}^T\mathbf{v}}} + \boldsymbol{\mu},$$

where $\mathbf{v} \sim \mathrm{N}(\mathbf{0}, \mathbf{I}_k)$, $\boldsymbol{\Upsilon} = \mathbf{R}\mathbf{R}^T$, and q_i is the x_ith quantile of the $\mathrm{F}_{k,\nu}$ distribution.

The algorithm for the t-distribution nested importance sampling algorithm is:

1. Generate a sample of size N_Π from Π. Set $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ to be the mean and variance matrix of this sample, respectively. Set $i = 1$ and $\hat{I}^{(1)}_{NIS} = 0$. Choose N and $\nu > 2$. Let $\boldsymbol{\Upsilon} = \frac{\nu-2}{\nu}\boldsymbol{\Sigma}$ and $\mathbf{R}$ be such that $\boldsymbol{\Upsilon} = \mathbf{R}\mathbf{R}^T$.

2. Generate $\mathbf{v} \sim \mathrm{N}(\mathbf{0}, \mathbf{I}_k)$ and set

$$\boldsymbol{\theta}_i = \frac{\sqrt{q_i k}\mathbf{R}\mathbf{v}}{\sqrt{\mathbf{v}^T\mathbf{v}}} + \boldsymbol{\mu},$$

   where q_i is the $e^{-i/N}$ quantile of the $\mathrm{F}_{k,\nu}$ distribution.

3. Let

$$\hat{I}^{(i+1)}_{NIS} = \hat{I}^{(i)}_{NIS} + (e^{-(i-1)/N} - e^{-i/N})\frac{g(\boldsymbol{\theta}_i)}{h(\boldsymbol{\theta}_i)},$$

   where $h()$ is the pdf of $H \equiv \mathrm{t}_\nu(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

4. Repeat steps 2. to 3. until $\hat{I}_{NIS}$ has converged.

4.4 Comparison of bridge and nested sampling

In this Section we undertake an empirical comparison of bridge and nested importance sampling. From first appearances, what is required for our implementations of bridge sampling and nested importance sampling are quite similar. Both methods require a sample of size N_Π to be generated from Π, and then both methods require a sample from $\mathrm{N}(\mathbf{0}, \mathbf{I}_k)$. For this reason, we can try to compare the two methods having equated their computational expense.

However, we need to be careful. The main source of computational expense for both methods, after generating samples from Π and H, is evaluating the function $g()$. Suppose we have

generated a sample of size N_Π from Π and after running the nested importance sampling algorithm, it converges after W iterations. This algorithm will have required a sample of size W to be generated from $\mathrm{N}(\mathbf{0}, \mathbf{I}_k)$ and W evaluations of $g()$. Suppose we then use bridge sampling using the split strategy recommended in Section 4.2 where $n_\Pi = \frac{1}{2}N_\Pi$ and $n_H = \frac{1}{2}W$, using the same sample of size W from $\mathrm{N}(\mathbf{0}, \mathbf{I}_k)$ that was used in the nested importance sampling algorithm. This will require a total of $2N_\Pi + 2W$ evaluations of $g()$. Instead, we will compare nested importance sampling with W iterations to bridge sampling with at most W evaluations of $g()$. To do this we set $n_\Pi = n_H = \lfloor \frac{W}{8} \rfloor$. We will also use this empirical study to compare nested importance sampling with $H \equiv \mathrm{N}(\mathbf{0}, \mathbf{I}_k)$ against nested importance sampling with $H \equiv \mathrm{t}_4(\mathbf{0}, \mathbf{I}_k)$, using the same sample of size N_Π from Π.

We consider four different target distributions:

1. $\Pi_1 \equiv \mathrm{N}(\mathbf{1}_k, 2\mathbf{I}_k)$, i.e. the k-variate normal distribution with mean $\mathbf{1}_k$ and variance matrix $2\mathbf{I}_k$.

2. $\Pi_2 \equiv \mathrm{L}(\mathbf{0}, \mathbf{I}_k)$, i.e. the k-variate logistic distribution with mean $\mathbf{0}$ and scale matrix $\mathbf{I}_k$.

3. $\Pi_3 \equiv \mathrm{t}_\nu(\mathbf{1}_k, \mathbf{R} = \mathbf{I}_k)$, i.e. the k-variate non-central t distribution with location $\mathbf{1}_k$, scale matrix $\mathbf{I}_k$ and ν degrees of freedom. This distribution has variance $\frac{\nu}{\nu-2}\mathbf{I}_k$ if $\nu > 2$ and is undefined if otherwise. Note that as $\nu \to \infty$, this distribution approaches $\mathrm{N}(\mathbf{0}, \mathbf{I}_k)$.

4. $\Pi_4 \equiv \mathrm{LG}(\alpha\mathbf{1}_k, 4000\mathbf{1}_k)$, i.e. the k-variate log gamma distribution with shape parameter $\alpha\mathbf{1}_k$ and scale parameter $4\mathbf{1}_k$. Note that this distribution becomes closer to a normal distribution as α grows large.

We consider two different values for N for nested importance sampling, i.e. $N = 100$ and $N = 1000$, and three different values for k, i.e. $k = 1, 5$ and 10. We generate a sample of size $N_\Pi = 100N$ from Π. We then find the nested importance sampling approximation to I which uses W iterations (with a maximum of N_Π) and a sample of size W from $\mathrm{N}(\mathbf{0}, \mathbf{I}_k)$. This is denoted $\hat{I}_{NIS1}$. We then compute the bridge sampling approximation to I with $n_\Pi = n_H = \lfloor \frac{W}{8} \rfloor$ with at most W evaluations of $g()$. We acquire our samples of size $2n_\Pi$ from Π and $\mathrm{N}(\mathbf{0}, \mathbf{I}_k)$ by randomly subsampling from within the existing samples from Π and $\mathrm{N}(\mathbf{0}, \mathbf{I}_k)$. This is denoted by $\hat{I}_{BS1}$. We also compute the nested importance sampling approximation with $H \equiv \mathrm{t}_4(\mathbf{0}, \mathbf{I}_k)$, denoted $\hat{I}_{NIS2}$.

In addition we compute the bridge sampling approximation with $n_\Pi = \frac{1}{2}N_\Pi$ and $n_H = \frac{W}{2}$, denoted $\hat{I}_{BS2}$.

We compare the four different methods using boxplots of the relative approximation, $\hat{I}/I$. Figures 4.6 and 4.7 show these boxplots for the different methods and the different values of k and N for Π_1 and Π_3, respectively. Figures 4.8 to 4.13 show the boxplots for Π_2 for the six different combinations of k and N, respectively. The boxplots of the relative approximation are plotted against different values of $\nu > 2$. Figures 4.14 to 4.19 show the boxplots for Π_4 for the six different combinations of k and N, respectively. The boxplots are plotted against different values of α on a non-linear scale.

Figure 4.6: Boxplots of the relative approximation for $\Pi_1 \equiv \mathrm{N}(\mathbf{1}_k, 2\mathbf{I}_k)$ for the four methods $\hat{I}_{NIS1}$, $\hat{I}_{BS1}$, $\hat{I}_{NIS2}$, and $\hat{I}_{BS2}$.

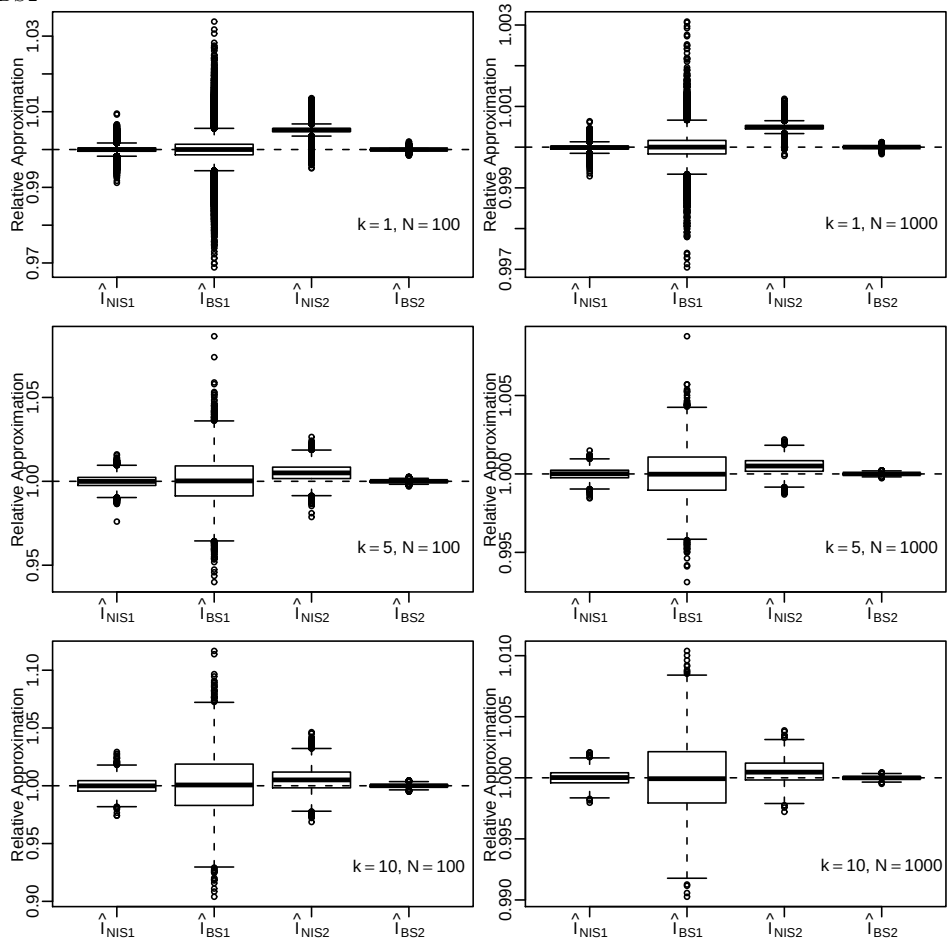
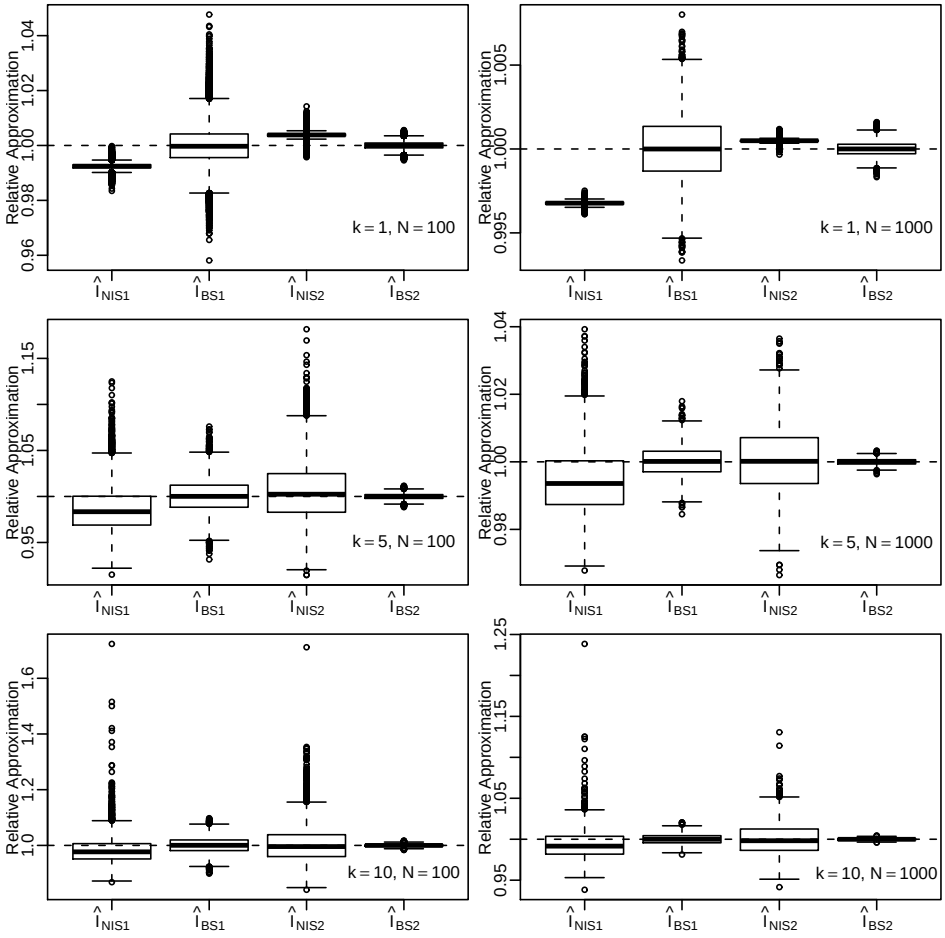


Figure 4.6 shows that if the target distribution is normal, then nested importance sampling with $H \equiv \mathrm{N}(\mathbf{0}, \mathbf{I}_k)$ outperforms bridge sampling with the same number of evaluations of $g()$. For other target distributions, nested importance sampling with $H \equiv \mathrm{N}(\mathbf{0}, \mathbf{I}_k)$ underestimates I. We see from Figures 4.8 to 4.13 when the target distribution is $\mathrm{t}_\nu(\mathbf{1}_k, \mathbf{R} = \mathbf{I}_k)$ and from Figures 4.14 to 4.19 when the target distribution is $\mathrm{LG}(\alpha\mathbf{1}_k, 4\mathbf{1}_k)$, that this underestimation decreases as the target distribution becomes closer to a normal distribution. It is also apparent that this underestimation tends to zero as $N \to \infty$.

On the other hand, nested importance sampling with $H \equiv \mathrm{t}_\nu(\mathbf{0}, \mathbf{I}_k)$ overestimates I when the target distribution is normal. We see from Figures 4.8 to 4.13 that when the target distribution is t_4 then nested importance sampling with $H \equiv \mathrm{t}_\nu(\mathbf{0}, \mathbf{I}_k)$ produces an approximation to I with no bias. Generally though, as the target distribution becomes closer to a normal distribution the overestimation of I by nested importance sampling with $H \equiv \mathrm{t}_\nu(\mathbf{0}, \mathbf{I}_k)$ tends to the amount seen when the target distribution is normal.

As expected, bridge sampling using the same number of values generated from Π and H as nested importance sampling with $H \equiv \mathrm{N}(\mathbf{0}, \mathbf{I}_k)$ outperforms bridge sampling using the same number of evaluations of $g()$ as nested importance sampling with $H \equiv \mathrm{N}(\mathbf{0}, \mathbf{I}_k)$.

Figure 4.7: Boxplots of the relative approximation for $\Pi_2 \equiv \mathrm{L}(\mathbf{0}, \mathbf{I}_k)$ for the four methods $\hat{I}_{NIS1}$, $\hat{I}_{BS1}$, $\hat{I}_{NIS2}$, and $\hat{I}_{BS2}$.



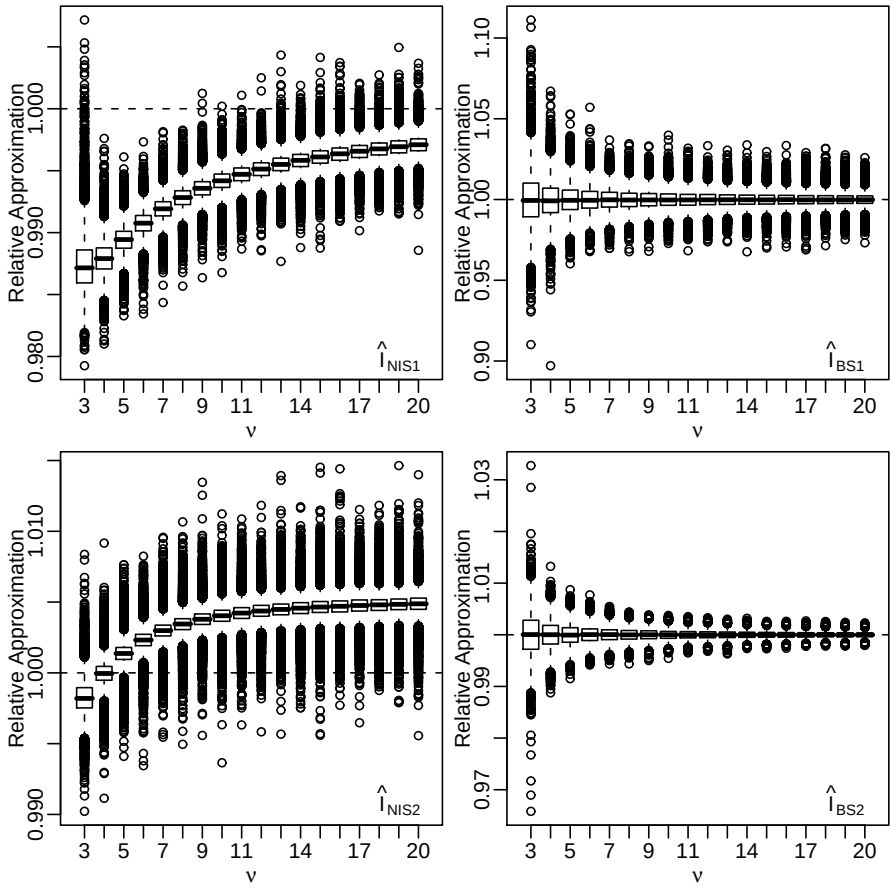
In conclusion, nested importance sampling with $H \equiv \mathrm{N}(\mathbf{0}, \mathbf{I}_k)$ performs very well when the target distribution is normal. In fact, in this case, it outperforms bridge sampling using the same number of evaluations of $g()$. When the target distribution is non-normal, it appears to underestimate I, although this underestimation decreases as N increases. If the target distribution is approximately normal than this underestimation may be small enough to ignore. Posterior distributions are approximately normal for large sample sizes. However, it would appear that bridge sampling provides a more robust alternative.

We now explore our implementation of nested importance sampling to gain some insight into the underestimation. Consider a one-dimensional example where the target distribution is $\Pi \equiv \mathrm{N}(0, 1)$ and $H \equiv \mathrm{N}(\mu, 1)$. Let $g(\theta) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\theta^2}{2}\right)$ and therefore $I = 1$. The nested importance sampling approximation is

$$\hat{I}_{NIS} = \sum_{i=1}^{\infty} \left(\exp\left(-\frac{i-1}{N}\right) - \exp\left(-\frac{i}{N}\right) \right) \frac{g(\theta_i)}{h(\theta_i)},$$

where $\theta_i = \frac{\sqrt{q_i} v}{|v|} + \mu$, $v \sim \mathrm{N}(0, 1)$ and q_i is the $\exp\left(-\frac{i}{N}\right)$ quantile of the χ^2_1 distribution. The

Figure 4.8: Boxplots of the relative approximation for $\Pi_3 \equiv \mathrm{t}_\nu(\mathbf{1}_k, \mathbf{R} = \mathbf{I}_k)$ for the four methods $\hat{I}_{NIS1}$, $\hat{I}_{BS1}$, $\hat{I}_{NIS2}$, and $\hat{I}_{BS2}$, for $k = 1$ and $N = 100$.



expected value of $\hat{I}_{NIS}$ is

$$\begin{aligned} \mathrm{E}\left(\hat{I}_{NIS}\right) &= \sum_{i=1}^{\infty}\left(\exp\left(-\frac{i-1}{N}\right) - \exp\left(-\frac{i}{N}\right)\right)\mathrm{E}\left(\frac{g(\theta_i)}{h(\theta_i)}\right), \\ &= \sum_{i=1}^{\infty}\left(\exp\left(-\frac{i-1}{N}\right) - \exp\left(-\frac{i}{N}\right)\right)\mathrm{E}\left(\mathrm{E}\left(\frac{g(\theta_i)}{h(\theta_i)}\middle|\,\mu\right)\right). \end{aligned}$$

Using the fact that

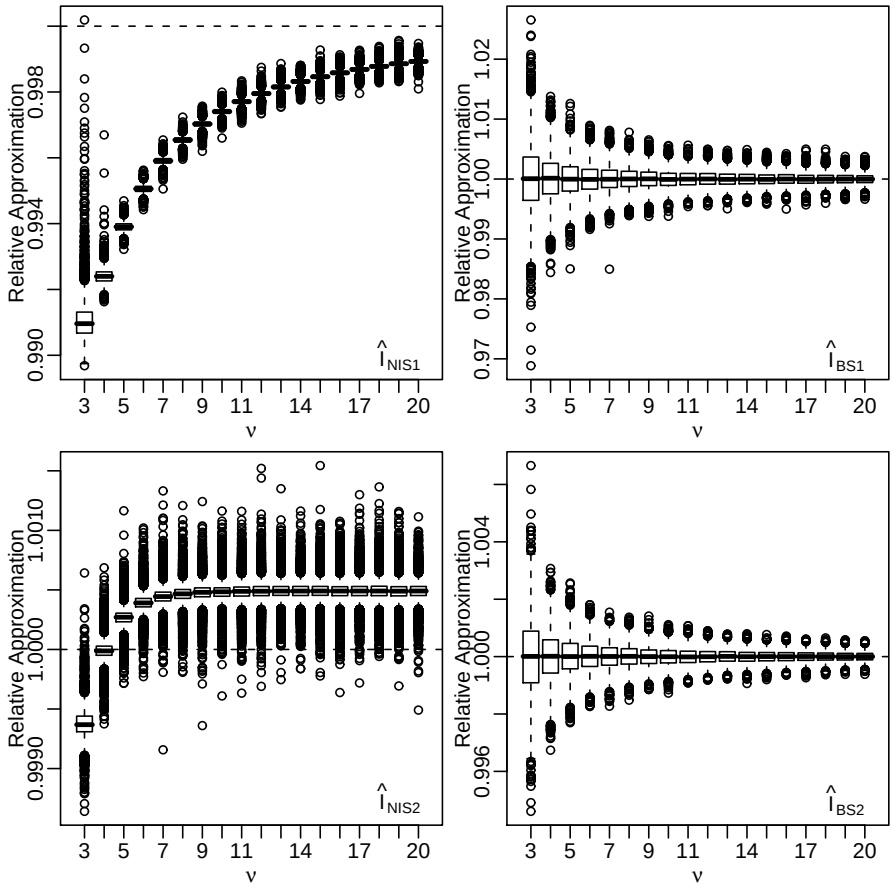
$$\frac{g(\theta_i)}{h(\theta_i)} = \exp\left(-\mu\sqrt{q_i}\frac{v}{|v|} - \frac{\mu^2}{2}\right),$$

we find that

$$\mathrm{E}\left(\frac{g(\theta_i)}{h(\theta_i)}\middle|\,\mu\right) = \frac{1}{2}\exp\left(-\frac{\mu^2}{2}\right)\left[\exp(\mu\sqrt{q_i}) + \exp(-\mu\sqrt{q_i})\right]. \tag{4.5}$$

If μ is known to be 0, i.e. has been found deterministically then $\mathrm{E}\left(\frac{g(\theta_i)}{h(\theta_i)}\middle|\,\mu\right) = 1$ and $\mathrm{E}\left(\hat{I}_{NIS}\right) = I = 1$. Therefore, in this case, nested importance sampling is unbiased. However, suppose we cannot find μ deterministically and so approximate it stochastically by letting μ

Figure 4.9: Boxplots of the relative approximation for $\Pi_3 \equiv \mathrm{t}_\nu(\mathbf{1}_k, \mathbf{R} = \mathbf{I}_k)$ for the four methods $\hat{I}_{NIS1}$, $\hat{I}_{BS1}$, $\hat{I}_{NIS2}$, and $\hat{I}_{BS2}$, for $k = 1$ and $N = 1000$.



be the mean of a sample of size N_Π from Π. Note that $\mu \sim \mathrm{N}\left(0, \frac{1}{N_\Pi}\right)$ and that $\mathrm{E}\left(\frac{g(\theta_i)}{h(\theta_i)}\middle| \mu\right)$ in (4.5) can be written as

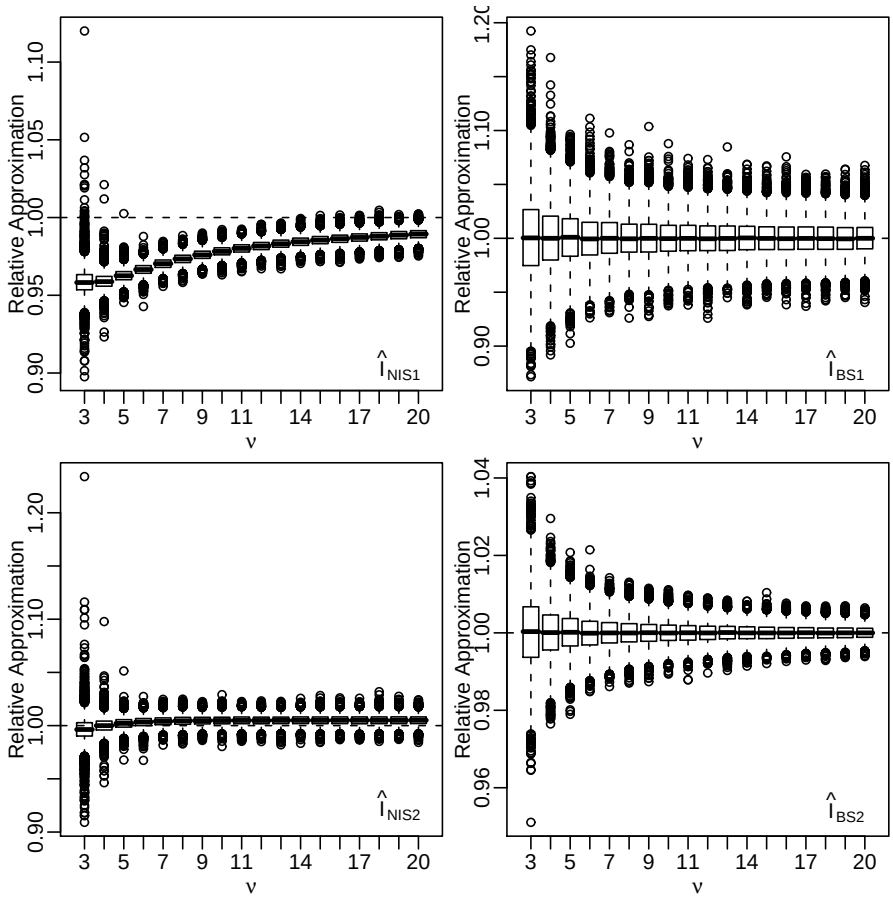
$$\mathrm{E}\left(\frac{g(\theta_i)}{h(\theta_i)}\middle| \mu\right) = \frac{1}{2}\exp\left(\frac{q_i}{2}\right)\left[\exp\left(-\frac{1}{2}(\mu - \sqrt{q_i})^2\right) + \exp\left(-\frac{1}{2}(\mu + \sqrt{q_i})^2\right)\right],$$

and by using the fact that $N_\Pi(\mu - \sqrt{q_i})^2$ and $N_\Pi(\mu + \sqrt{q_i})^2$ have the same distribution, namely the non-central χ^2_1 distribution with non-centrality parameter $q_i N_\Pi$, we find that

$$\mathrm{E}\left(\hat{I}_{NIS}\right) = \sqrt{\frac{N_\Pi}{N_\Pi + 1}}\left(\exp\left(\frac{1}{N}\right) - 1\right)\sum_{i=1}^{\infty}\exp\left(\frac{q_i}{2(N_\Pi + 1)} - \frac{i}{N}\right). \tag{4.6}$$

Figure 4.20 shows a plot of $E(\hat{I}_{NIS})$ from (4.6) against N for three different values of N_Π, i.e. 100,1000,10000. This shows that nested importance sampling underestimates I even when the target distribution, Π, is normal. However, this bias is asymptotically zero with respect to N and N_Π. This bias is negligible for moderately large values of N and N_Π and this is why we did not notice any bias for $N = 100, 1000$ and $N_\Pi = 10000, 100000$ in Figure 4.6.

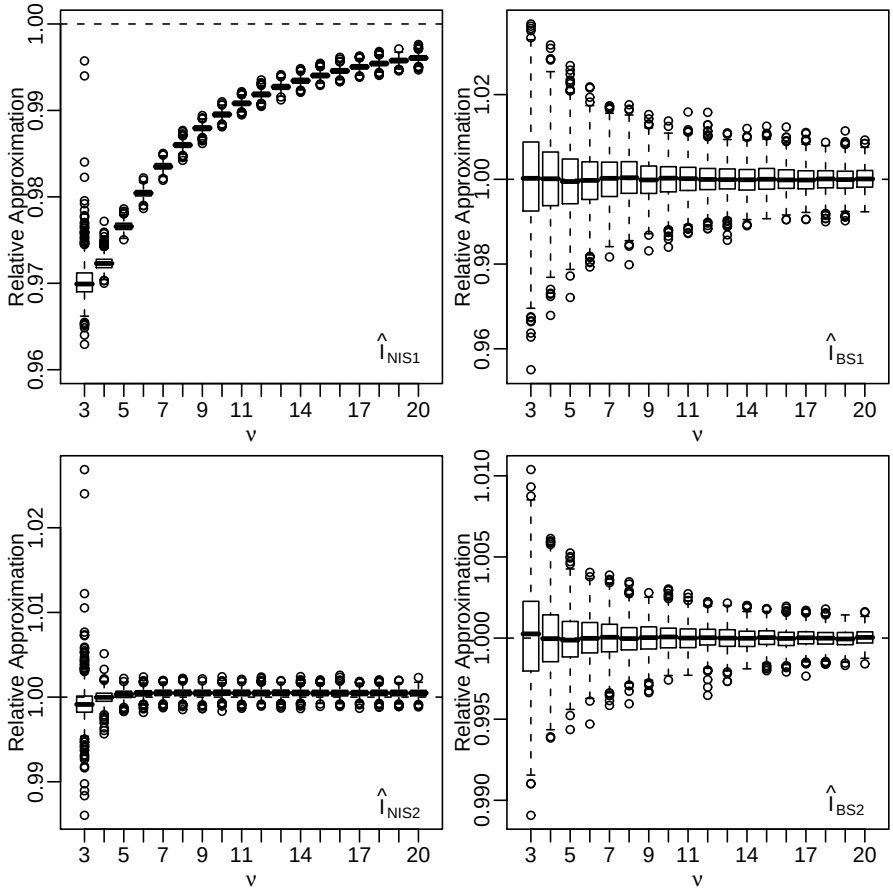
Figure 4.10: Boxplots of the relative approximation for $\Pi_3 \equiv \mathrm{t}_\nu(\mathbf{1}_k, \mathbf{R} = \mathbf{I}_k)$ for the four methods $\hat{I}_{NIS1}$, $\hat{I}_{BS1}$, $\hat{I}_{NIS2}$, and $\hat{I}_{BS2}$, for $k = 5$ and $N = 100$.



It would appear that there are two mechanisms that are driving the underestimation of I by nested importance sampling: 1) non-normality of Π, and 2) approximating $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ for $H \equiv \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ by sample statistics from a sample generated from Π. Chopin and Robert (2009) found that nested importance sampling with small values of N underestimated I for non-normal distributions when $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ were found deterministically. In Section 4.6, we find that nested importance sampling, when $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are found deterministically, underestimates I, even for large values of N when the target distribution is the posterior distribution of a GLMM.

Our conclusion from this Section, is that bridge sampling is the more robust method for approximating I.

Figure 4.11: Boxplots of the relative approximation for $\Pi_3 \equiv \mathrm{t}_\nu(\mathbf{1}_k, \mathbf{R} = \mathbf{I}_k)$ for the four methods $\hat{I}_{NIS1}$, $\hat{I}_{BS1}$, $\hat{I}_{NIS2}$, and $\hat{I}_{BS2}$, for $k = 5$ and $N = 1000$.

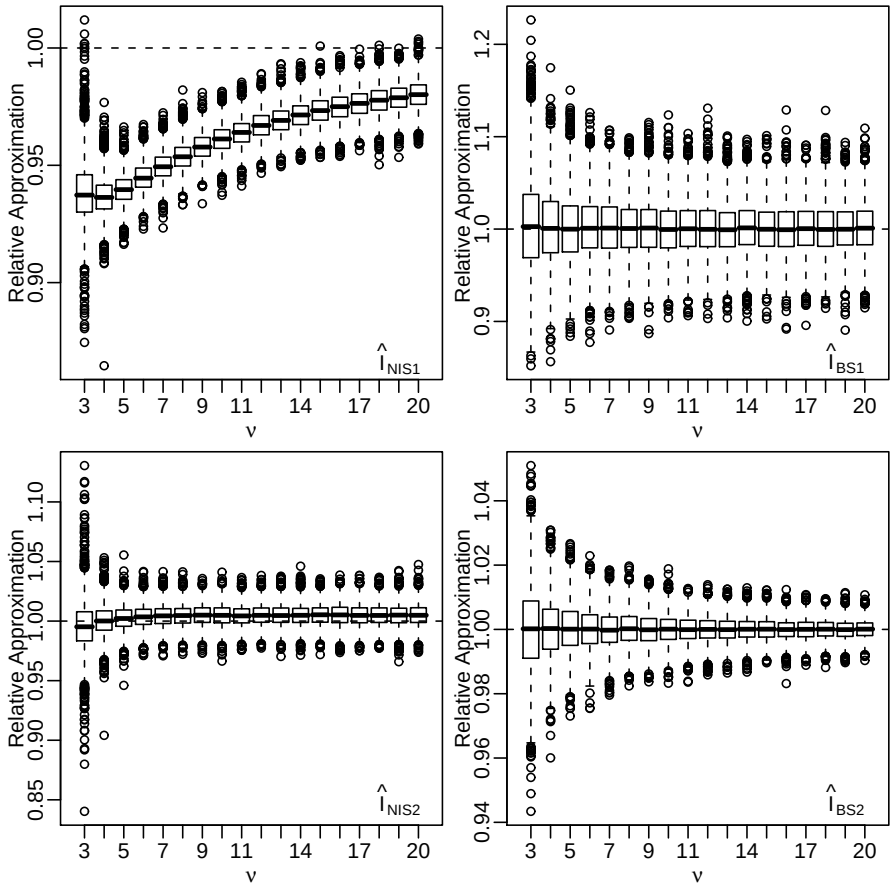


4.5 Application to GLMMs

In this Section we describe how bridge sampling or nested importance sampling can be applied to approximating the marginal likelihood of a GLMM. The discussion of the two methods in the previous Sections relied on the model parameters lying in $\mathbb{R}^k$. It is easy to transform the parameters of a GLMM so that they lie in $\mathbb{R}^k$. Note that for the parameters of a GLMM, $\boldsymbol{\beta} \in \mathbb{R}^p$, $\mathbf{u} \in \mathbb{R}^{Gq}$, $\mathbf{D} \in \mathbb{P}^q$ and $\phi \in \mathbb{R}^+$, where $\mathbb{P}^q$ denotes the set of all $q \times q$ positive-definite matrices and $\mathbb{R}^+$ denotes the set of positive real numbers. We need transformations, $t_1()$ and $t_2()$, such that $\boldsymbol{\nu} = t_1(\mathbf{D}) \in \mathbb{R}^{\frac{1}{2}q(q+1)}$ and $\omega = t_2(\phi) \in \mathbb{R}$, where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\nu}, \omega)^T \in \mathbb{R}^{p+\frac{1}{2}q(q+1)+1}$. For $t_1()$, we use the Cholesky decomposition $\mathbf{D} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}^T$, where $\boldsymbol{\Gamma}$ is a lower triangular matrix such that

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

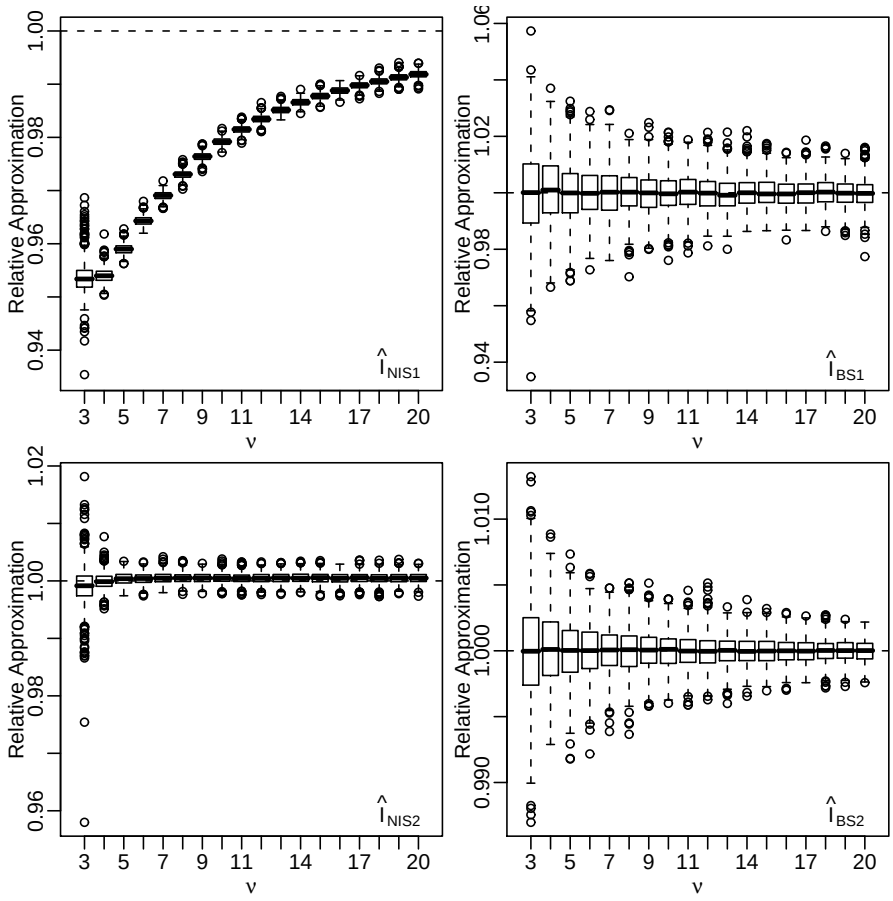
Figure 4.12: Boxplots of the relative approximation for $\Pi_3 \equiv \mathrm{t}_\nu(\mathbf{1}_k, \mathbf{R} = \mathbf{I}_k)$ for the four methods $\hat{I}_{NIS1}$, $\hat{I}_{BS1}$, $\hat{I}_{NIS2}$, and $\hat{I}_{BS2}$, for $k = 10$ and $N = 100$.



If $\boldsymbol{\nu} = (\nu_{11}, \nu_{12}, ..., \nu_{1q}, \nu_{22}, ..., \nu_{2q}, ..., \nu_{qq})^T \in \mathbb{R}^{\frac{1}{2}q(q+1)}$, then $\mathbf{D}$ is guaranteed to be positive-definite. Using, for example, Muirhead (1982, Theorem 2.1.9), the Jacobian of this transformation is $\mathrm{d}\mathbf{D} = 2^q \prod_{k=1}^{q} \exp\left(\nu_{kk}(q+2-k)\right) \mathrm{d}\boldsymbol{\nu}$. For $t_2()$, we use the transformation $\phi = \exp(\omega)$ with Jacobian $\mathrm{d}\phi = \exp(\omega)\mathrm{d}\omega$. If $\omega \in \mathbb{R}$, then $\phi \in \mathbb{R}^+$.

Now the vector of transformed parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}, \mathbf{u}, \boldsymbol{\nu}, \omega)^T \in \mathbb{R}^{p+Gq+\frac{1}{2}q(q+1)+1}$.

Figure 4.13: Boxplots of the relative approximation for $\Pi_3 \equiv \mathrm{t}_\nu(\mathbf{1}_k, \mathbf{R} = \mathbf{I}_k)$ for the four methods $\hat{I}_{NIS1}$, $\hat{I}_{BS1}$, $\hat{I}_{NIS2}$, and $\hat{I}_{BS2}$, for $k = 10$ and $N = 1000$.



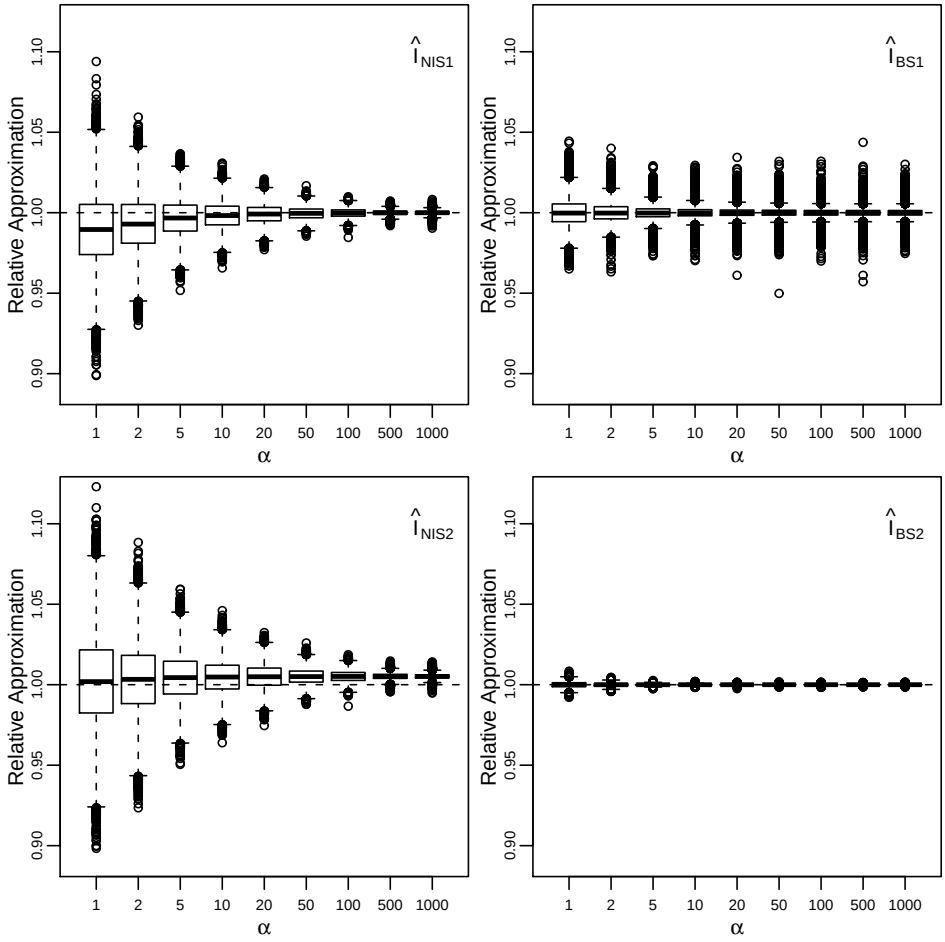
The marginal likelihood for model $m \in M$ is

$$\begin{aligned} f_m(\mathbf{y}) &= \int_{\mathbb{R}^+} \int_{\mathbb{P}^{q_m}} \int_{\mathbb{R}^{Gq_m}} \int_{\mathbb{R}^{p_m}} f_m(\mathbf{y}|\boldsymbol{\beta}_m, \mathbf{u}_m, \phi_m) \\ &\qquad f_m(\mathbf{u}_m|\mathbf{D}_m) f_m(\boldsymbol{\beta}_m, \mathbf{D}_m, \phi_m) \mathrm{d}\boldsymbol{\beta}_m \mathrm{d}\mathbf{u}_m \mathrm{d}\mathbf{D}_m \mathrm{d}\phi_m, \qquad (4.7) \\ &= \int_{\mathbb{R}^{p_m + Gq_m + \frac{1}{2}q_m(q_m+1)+1}} f_m(\mathbf{y}|\boldsymbol{\beta}_m, \mathbf{u}_m, e^{\omega_m}) f_m(\mathbf{u}_m|\boldsymbol{\Gamma}_m \boldsymbol{\Gamma}_m^T) f_m(\boldsymbol{\beta}_m, \boldsymbol{\Gamma}_m \boldsymbol{\Gamma}_m^T, e^{\omega_m}), \\ &\qquad 2^{q_m} e^{\omega_m} \prod_{k=1}^{q_m} e^{\nu_{m,kk}(q_m+2-k)} \mathrm{d}\boldsymbol{\beta}_m \mathrm{d}\mathbf{u}_m \mathrm{d}\boldsymbol{\nu}_m \mathrm{d}\omega_m, \\ &= \int_{\mathbb{R}^{p_m + Gq_m + \frac{1}{2}q_m(q_m+1)+1}} g_m(\boldsymbol{\beta}_m, \mathbf{u}_m, \boldsymbol{\nu}_m, \omega_m) \mathrm{d}\boldsymbol{\beta}_m \mathrm{d}\mathbf{u}_m \mathrm{d}\boldsymbol{\nu}_m \mathrm{d}\omega_m, \qquad (4.8) \end{aligned}$$

where

$$\begin{aligned} &g_m(\boldsymbol{\beta}_m, \mathbf{u}_m, \boldsymbol{\nu}_m, \omega_m) = \\ &\qquad f_m(\mathbf{y}|\boldsymbol{\beta}_m, \mathbf{u}_m, e^{\omega_m}) f_m(\mathbf{u}_m|\boldsymbol{\Gamma}_m \boldsymbol{\Gamma}_m^T) f_m(\boldsymbol{\beta}_m, \boldsymbol{\Gamma}_m \boldsymbol{\Gamma}_m^T, e^{\omega_m}) 2^{q_m} e^{\omega_m} \prod_{k=1}^{q_m} e^{\nu_{m,kk}(q_m+2-k)}. \end{aligned}$$

Figure 4.14: Boxplots of the relative approximation for $\Pi_4 \equiv \text{LG}(\alpha \mathbf{1}_k, 4\mathbf{1}_k)$ for the four methods $\hat{I}_{NIS1}$, $\hat{I}_{BS1}$, $\hat{I}_{NIS2}$, and $\hat{I}_{BS2}$, for $k = 1$ and $N = 100$.



To generate a sample from $\boldsymbol{\beta}_m, \mathbf{u}_m, \boldsymbol{\nu}_m, \omega_m|\mathbf{y}$ we generate a sample from $\boldsymbol{\beta}_m, \mathbf{u}_m, \mathbf{D}_m, \phi_m|\mathbf{y}$ and then transform. Due to the way WinBUGS works, this is typically more convenient than trying to generate a sample from $\boldsymbol{\beta}_m, \mathbf{u}_m, \boldsymbol{\nu}_m, \omega_m|\mathbf{y}$ directly. Also, if m is a model of interest, then a sample from $\boldsymbol{\beta}_m, \mathbf{u}_m, \mathbf{D}_m, \phi_m|\mathbf{y}$ will be easier to interpret than a sample from $\boldsymbol{\beta}_m, \mathbf{u}_m, \boldsymbol{\nu}_m, \omega_m|\mathbf{y}$, for inferential purposes.

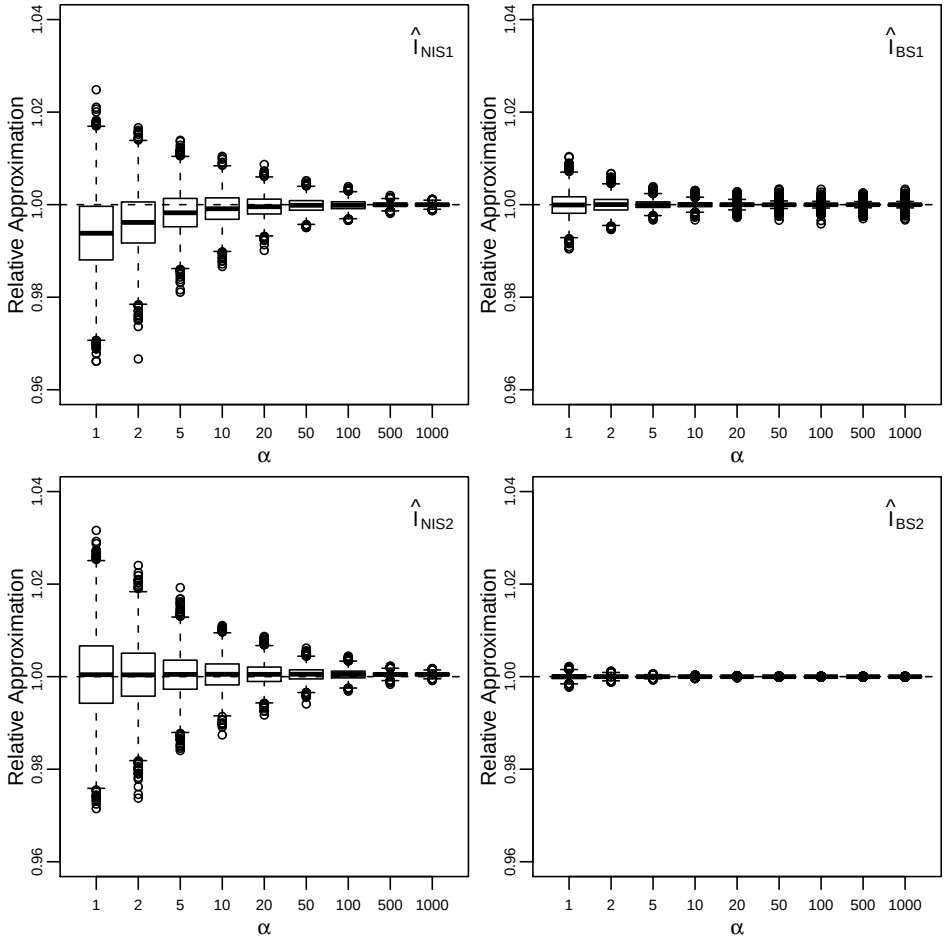
The expression (4.7) for the marginal likelihood of a GLMM can be used with any prior distribution for the parameters, $\boldsymbol{\beta}_m$, $\mathbf{D}_m$ and ϕ_m. However in Section 1.2.3, we decomposed the pdf of the prior distribution as

$$f_m(\boldsymbol{\beta}_m, \mathbf{D}_m, \phi_m) = f_m(\boldsymbol{\beta}_m|\mathbf{D}_m, \phi_m) f_m(\mathbf{D}_m|\phi_m) f_m(\phi_m).$$

Furthermore, it is computationally advantageous if the prior distribution for $\boldsymbol{\beta}_m$ is independent of $\mathbf{D}_m$. The unit information prior distribution for $\boldsymbol{\beta}_m$, proposed in Chapter 3, has such a property. In this case, the pdf of the prior distribution can be decomposed as

$$f_m(\boldsymbol{\beta}_m, \mathbf{D}_m, \phi_m) = f_m(\boldsymbol{\beta}_m|\phi_m) f_m(\mathbf{D}_m|\phi_m) f_m(\phi_m).$$

Figure 4.15: Boxplots of the relative approximation for $\Pi_4 \equiv \text{LG}(\alpha\mathbf{1}_k, 4\mathbf{1}_k)$ for the four methods $\hat{I}_{NIS1}$, $\hat{I}_{BS1}$, $\hat{I}_{NIS2}$, and $\hat{I}_{BS2}$, for $k = 1$ and $N = 1000$.



Therefore, the expression (4.7) for the marginal likelihood can be simplified as

$$\begin{aligned} f_m(\mathbf{y}) &= \int_{\mathbb{R}^+}\int_{\mathbb{P}^{q_m}}\int_{\mathbb{R}^{Gq_m}}\int_{\mathbb{R}^{p_m}} f_m(\mathbf{y}|\boldsymbol{\beta}_m, \mathbf{u}_m, \phi_m) f_m(\mathbf{u}_m|\mathbf{D}_m) \\ &\qquad f_m(\boldsymbol{\beta}_m|\phi_m) f_m(\mathbf{D}_m|\phi_m) f_m(\phi_m) \mathrm{d}\boldsymbol{\beta}_m \mathrm{d}\mathbf{u}_m \mathrm{d}\mathbf{D}_m \mathrm{d}\phi_m, \\ &= \int_{\mathbb{R}^+}\int_{\mathbb{R}^{Gq_m}}\int_{\mathbb{R}^{p_m}} f_m(\mathbf{y}|\boldsymbol{\beta}_m, \mathbf{u}_m, \phi_m) f_m(\boldsymbol{\beta}_m|\phi_m) f_m(\phi_m) \\ &\qquad \int_{\mathbb{P}^{q_m}} f_m(\mathbf{u}_m|\mathbf{D}_m) f_m(\mathbf{D}_m|\phi_m) \mathrm{d}\mathbf{D}_m \mathrm{d}\boldsymbol{\beta}_m \mathrm{d}\mathbf{u}_m \mathrm{d}\phi_m. \end{aligned}$$

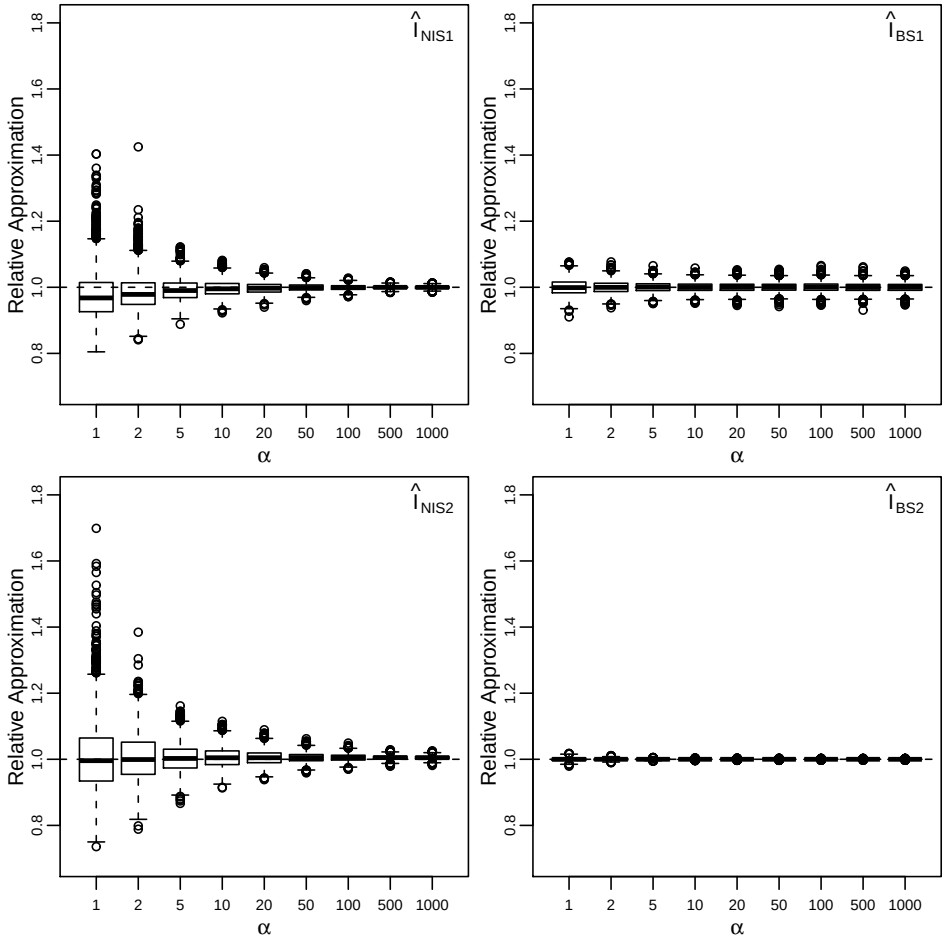
If the prior distribution for $\mathbf{D}_m|\phi_m$ is the inverse-Wishart distribution, $\text{IW}(\rho_m, \mathbf{R}_m(\phi_m))$, with shape parameter, ρ_m, and scale matrix, $\mathbf{R}_m(\phi_m)$, which depends on ϕ_m, if it is unknown, then the integral

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

is analytically tractable as

$$\int_{\mathbb{P}^{q_m}} f_m(\mathbf{u}_m|\mathbf{D}_m) f_m(\mathbf{D}_m|\phi_m) \mathrm{d}\mathbf{D}_m = \frac{\Gamma_{q_m}\left(\frac{\rho_m+G}{2}\right)}{\Gamma_{q_m}\left(\frac{\rho_m}{2}\right)} \frac{1}{\pi^{\frac{Gq_m}{2}}} \frac{|\mathbf{R}_m(\phi_m)|^{\frac{\rho_m}{2}}}{|\mathbf{R}_m(\phi_m) + \sum_{i=1}^{G} \mathbf{u}_{mi}\mathbf{u}_{mi}^T|^{\frac{\rho_m+G}{2}}},$$

Figure 4.16: Boxplots of the relative approximation for $\Pi_4 \equiv \text{LG}(\alpha\mathbf{1}_k, 4\mathbf{1}_k)$ for the four methods $\hat{I}_{NIS1}$, $\hat{I}_{BS1}$, $\hat{I}_{NIS2}$, and $\hat{I}_{BS2}$, for $k = 5$ and $N = 100$.



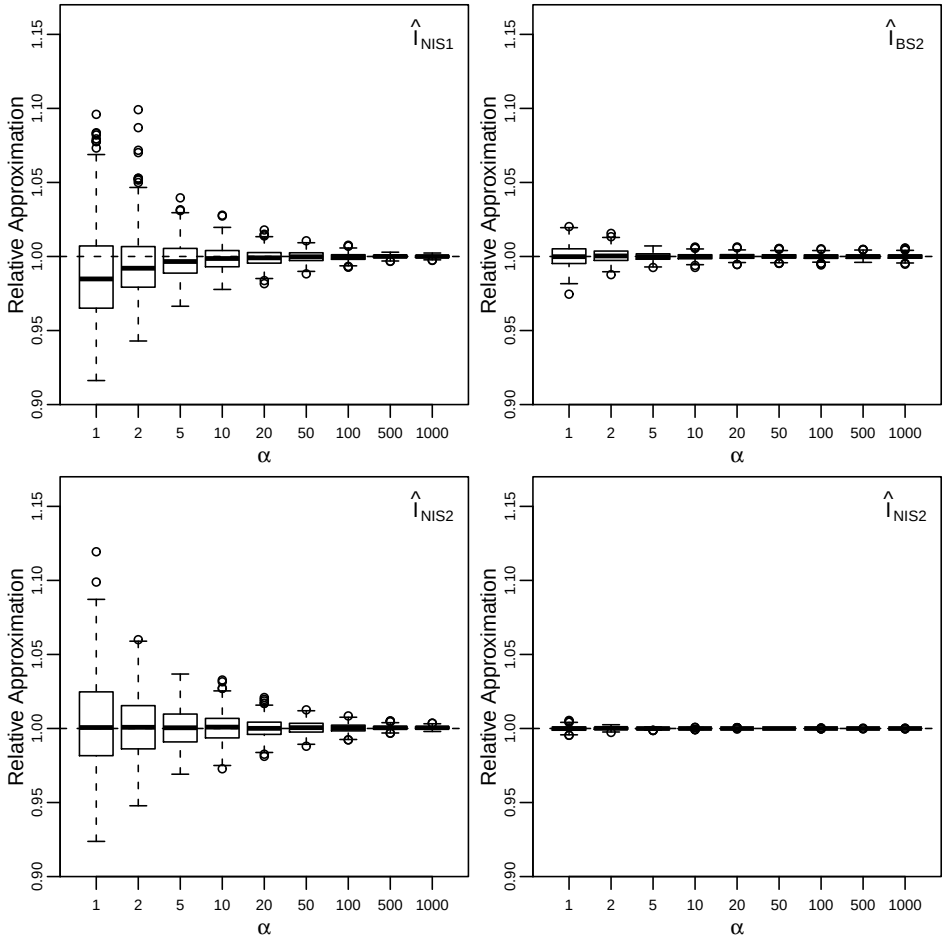
where

$$\Gamma_{q_m}(a) = \pi^{\frac{1}{4}q_m(q_m-1)} \prod_{k=1}^{q_m} \Gamma\left(a + \frac{1-k}{2}\right),$$

is the *multivariate gamma function*. Therefore, the marginal likelihood is

$$\begin{aligned} f_m(\mathbf{y}) \;=\; & \int_{\mathbb{R}^+} \int_{\mathbb{R}^{Gq_m}} \int_{\mathbb{R}^{p_m}} \frac{\Gamma_{q_m}\left(\frac{\rho_m+G}{2}\right)}{\Gamma_{q_m}\left(\frac{\rho_m}{2}\right)\pi^{\frac{Gq_m}{2}}} f_m(\mathbf{y}|\boldsymbol{\beta}_m, \mathbf{u}_m, \phi_m) f_m(\boldsymbol{\beta}_m|\phi_m) \\ & \frac{|\mathbf{R}_m(\phi_m)|^{\frac{\rho_m}{2}}}{|\mathbf{R}_m(\phi_m) + \sum_{i=1}^{G} \mathbf{u}_{mi}\mathbf{u}_{mi}^T|^{\frac{\rho_m+G}{2}}} f_m(\phi_m) \mathrm{d}\boldsymbol{\beta}_m \mathrm{d}\mathbf{u}_m \mathrm{d}\phi_m. \end{aligned}$$

Figure 4.17: Boxplots of the relative approximation for $\Pi_4 \equiv \text{LG}(\alpha \mathbf{1}_k, 4\mathbf{1}_k)$ for the four methods $\hat{I}_{NIS1}$, $\hat{I}_{BS1}$, $\hat{I}_{NIS2}$, and $\hat{I}_{BS2}$, for $k = 5$ and $N = 1000$.



Using the transformation $\omega_m = e^{\phi_m}$, we find

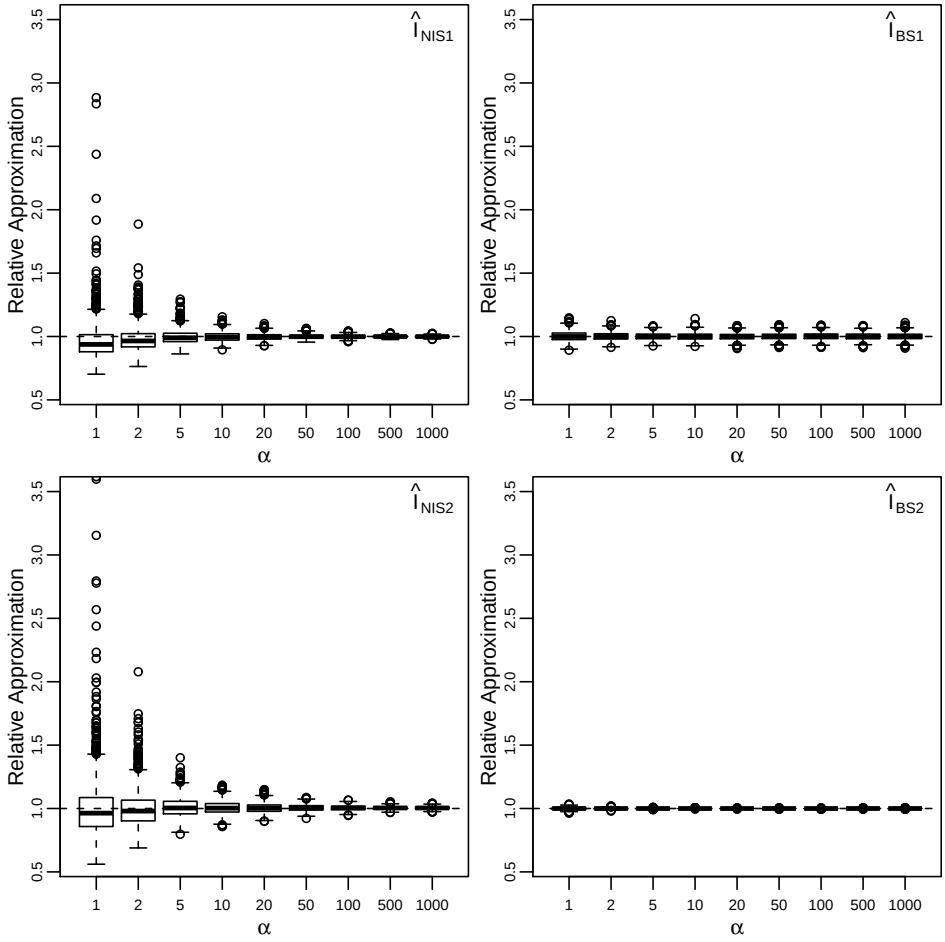
$$\begin{aligned} f_m(\mathbf{y}) &= \int_{\mathbb{R}^+} \int_{\mathbb{R}^{Gq_m}} \int_{\mathbb{R}^{p_m}} \frac{\Gamma_{q_m}\left(\frac{\rho_m+G}{2}\right)}{\Gamma_{q_m}\left(\frac{\rho_m}{2}\right)\pi^{\frac{Gq_m}{2}}} f_m(\mathbf{y}|\boldsymbol{\beta}_m, \mathbf{u}_m, e^{\omega_m}) f_m(\boldsymbol{\beta}_m|e^{\omega_m}) \\ &\qquad \frac{|\mathbf{R}_m(e^{\omega_m})|^{\frac{\rho_m}{2}}}{|\mathbf{R}_m(e^{\omega_m}) + \sum_{i=1}^G \mathbf{u}_{mi}\mathbf{u}_{mi}^T|^{\frac{\rho_m+G}{2}}} f_m(e^{\omega_m}) e^{\omega_m} \mathrm{d}\boldsymbol{\beta}_m \mathrm{d}\mathbf{u}_m \mathrm{d}\omega_m, \\ &= \int_{\mathbb{R}^{p_m+Gq_m+1}} g_m(\boldsymbol{\beta}_m, \mathbf{u}_m, \omega_m) \mathrm{d}\boldsymbol{\beta}_m \mathrm{d}\mathbf{u}_m \mathrm{d}\omega_m, \end{aligned}$$

where

$$\begin{aligned} g_m(\boldsymbol{\beta}_m, \mathbf{u}_m, \omega_m) &= \frac{\Gamma_{q_m}\left(\frac{\rho_m+G}{2}\right)}{\Gamma_{q_m}\left(\frac{\rho_m}{2}\right)\pi^{\frac{Gq_m}{2}}} f_m(\mathbf{y}|\boldsymbol{\beta}_m, \mathbf{u}_m, e^{\omega_m}) f_m(\boldsymbol{\beta}_m|e^{\omega_m}) \\ &\qquad \frac{|\mathbf{R}_m(e^{\omega_m})|^{\frac{\rho_m}{2}}}{|\mathbf{R}_m(e^{\omega_m}) + \sum_{i=1}^G \mathbf{u}_{mi}\mathbf{u}_{mi}^T|^{\frac{\rho_m+G}{2}}} f_m(e^{\omega_m}) e^{\omega_m}. \end{aligned}$$

The unit information prior for $\mathbf{D}_m$ proposed in Chapter 3 is an inverse-Wishart prior distribution with $\rho_m = q_m + 2$ and $\mathbf{R}_m(\phi_m) = G\left(\sum_{i=1}^G \mathbf{Z}_{mi}\mathbf{W}_{i,\mathbf{m},\mathbf{0}}\mathbf{Z}_{mi}\right)^{-1}$. Note that $\mathbf{W}_{i,\mathbf{m},\mathbf{0}}$

Figure 4.18: Boxplots of the relative approximation for $\Pi_4 \equiv \mathrm{LG}(\alpha \mathbf{1}_k, 4\mathbf{1}_k)$ for the four methods $\hat{I}_{NIS1}$, $\hat{I}_{BS1}$, $\hat{I}_{NIS2}$, and $\hat{I}_{BS2}$, for $k = 10$ and $N = 100$.



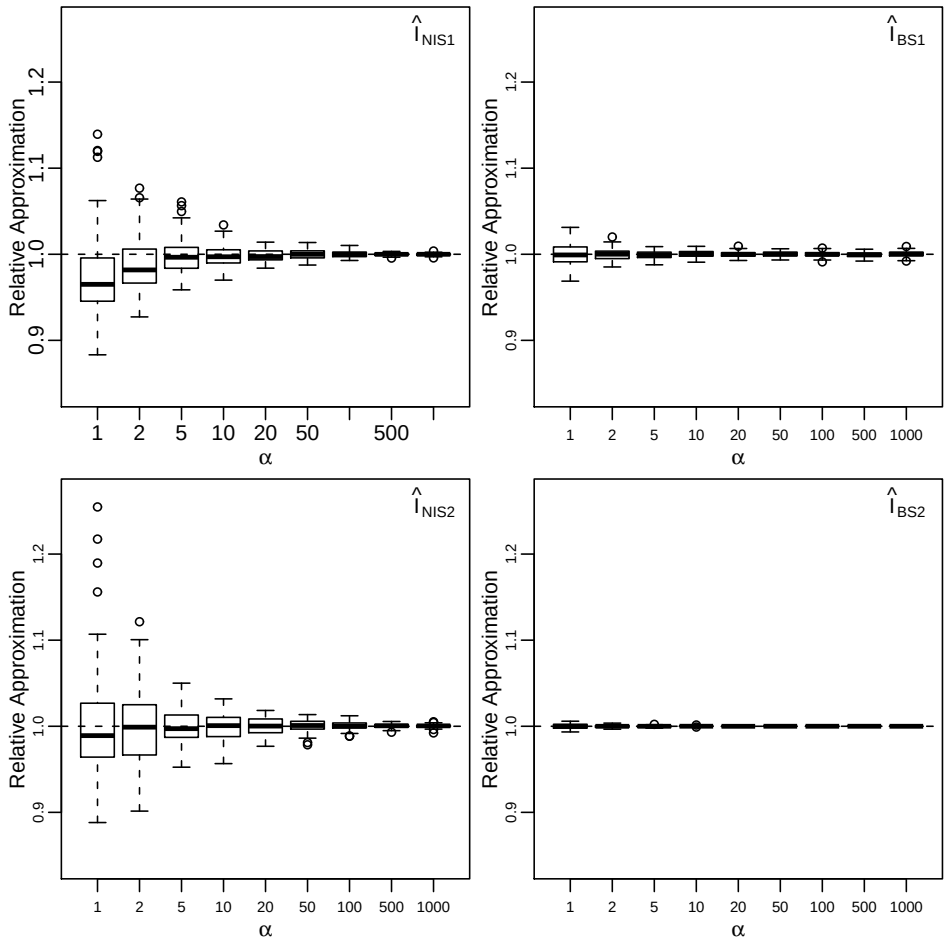
depends on ϕ_m, if it is unknown.

The marginal likelihood is now a $(p_m + Gq_m + 1)$-dimensional integral which we can approximate using bridge sampling. The function $g_m(\boldsymbol{\beta}_m, \mathbf{u}_m, \omega_m)$ is the pdf of the marginal posterior distribution, $\boldsymbol{\beta}_m, \mathbf{u}_m, \omega_m|\mathbf{y}$. To obtain a sample from this distribution, generate a sample from $\boldsymbol{\beta}_m, \mathbf{u}_m, \mathbf{D}_m, \phi_m|\mathbf{y}$, discard the $\mathbf{D}_m$'s and transform the ϕ_m.

We noted in Section 4.4 that nested importance sampling with $H \equiv \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ typically underestimates I for non-normal target distributions in a non-negligible way. However, we also noted that for a normal target distribution, nested importance sampling outperformed bridge sampling (see Figure 4.6) based on the same number of evaluations of $g()$. We concluded that bridge sampling provided a more robust method for approximating I.

We know that posterior distributions are asymptotically normal with respect to the sample size, n, and approximately normal for a large sample size. It is unclear how 'normal' a posterior distribution would have to be for us to favour nested importance over bridge sampling. We attempt to address this issue in the next Section.

Figure 4.19: Boxplots of the relative approximation for $\Pi_4 \equiv \text{LG}(\alpha \mathbf{1}_k, 4\mathbf{1}_k)$ for the four methods $\hat{I}_{NIS1}$, $\hat{I}_{BS1}$, $\hat{I}_{NIS2}$, and $\hat{I}_{BS2}$, for $k = 10$ and $N = 1000$.



4.5.1 Turtle Data Example

We return to our turtle dataset running example where we have implemented the unit information priors of Chapter 3 according to Section 3.3. We generate a sample of size $N_\Pi = 10000$ from the posterior distribution of each of the five models using WinBUGS. We then approximate the marginal likelihood of model m using nested importance sampling with W iterations. Subsequently, we approximate the marginal likelihood of model m using bridge sampling with at most W evaluations of $g_m()$, i.e. $n_\Pi = n_H = \lfloor \frac{W}{8} \rfloor$. We repeat this process 500 times.

We do not know the true value of the marginal likelihood for the five models so we follow Sinharay and Stern (2005) by approximating the marginal likelihoods using importance sampling with a very high sample size. We use $H \equiv \text{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the mean and variance matrix, respectively, of a posterior sample of size 100000. We then use a sample size of ten million in the importance sampler. Table 4.1 shows these approximations to the log of the marginal likelihoods of each of the five models. If we repeat this process using different samples from the posterior and H, we arrive at the same approximations to four decimal places. Therefore, we may assume that, up to this level of accuracy, the values in Table 4.1

Figure 4.20: Plot of $E(\hat{I}_{NIS})$ from (4.6) against N for three different values of N_Π.

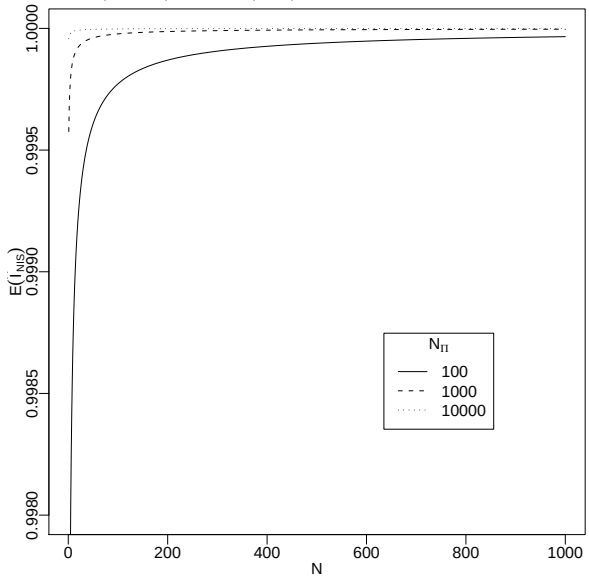


Table 4.1: Importance sampling approximations to the log of the marginal likelihood of the five models for the Turtle Dataset.

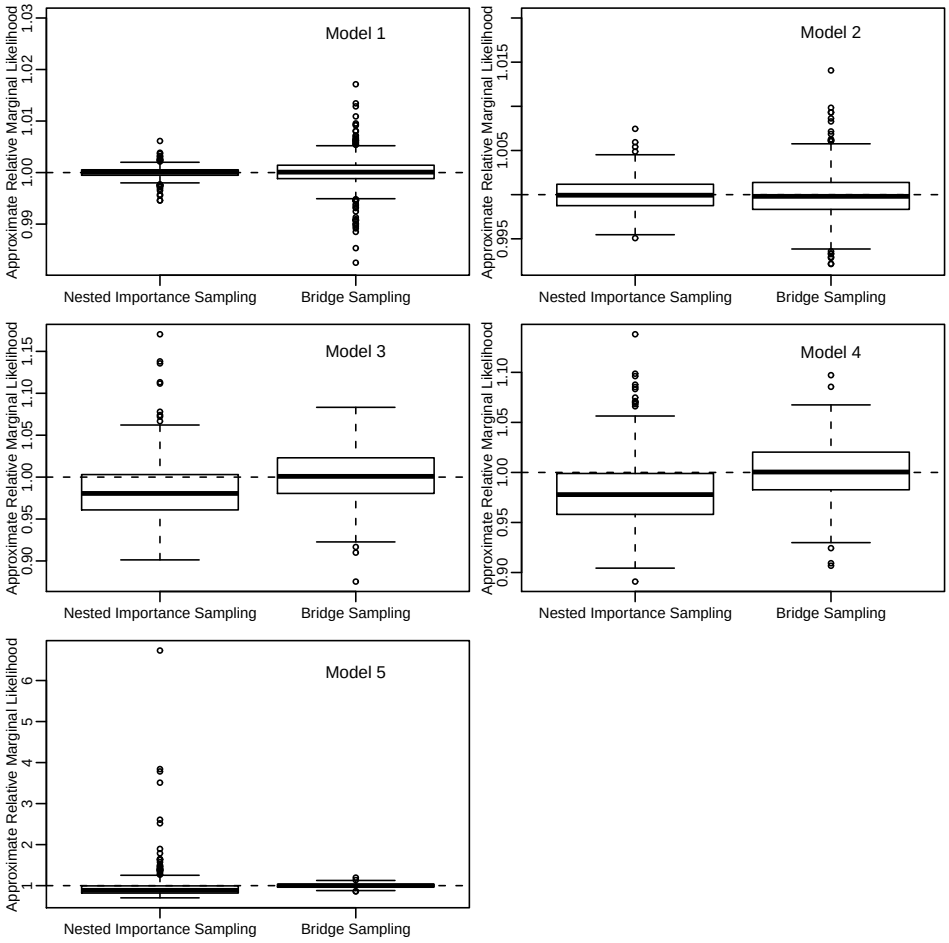
Model m	Log marginal likelihood $\log f_m(\mathbf{y})$
1	-162.8563
2	-154.2634
3	-159.8786
4	-154.8849
5	-153.9786

are the true marginal likelihoods, $f_m(\mathbf{y})$.

Figure 4.21 shows boxplots of the 500 approximations to the relative marginal likelihood, $\hat{f}_m(\mathbf{y})/f_m(\mathbf{y})$, using bridge sampling and nested importance sampling for the five models.

Figure 4.21 shows that for models 1 and 2, nested importance sampling outperforms bridge sampling based on the same number of evaluations of the unnormalised posterior pdf with respect to minimising the variance of the approximation. However, for models 3, 4 and 5, we see that nested importance sampling exhibits the same behaviour seen in Section 4.4 of underestimating the normalising constant for non-normal target distributions. This is probably caused by the non-normality of the posterior distributions of models 3, 4 and 5. Models 3, 4 and 5 all have group-specific parameters which seems to cause a departure from normality. This small empirical study confirms our conclusion from Section 4.4 that bridge sampling provides the most robust method for approximating the marginal likelihood, especially for the marginal likelihood of a GLMM.

Figure 4.21: Boxplots of the 500 approximations to the marginal likelihood using nested importance sampling and bridge sampling for the five models from the Turtle Dataset.



4.6 Mode and Curvature

In Section 4.2, we stated that we felt that using the mode and curvature at the mode of Π would not fully describe Π and that using the mean and variance of a sample generated from Π would be a better choice. We felt that this is especially true when Π is the posterior distribution of a GLMM. In this Section, we investigate this issue empirically, using the Turtle Dataset.

For each of the five models in M, we find the posterior mode, $\tilde{\boldsymbol{\theta}}_m = (\tilde{\boldsymbol{\beta}}_m, \tilde{\mathbf{u}}_m)^T$, of the transformed posterior distribution by maximising $\log g_m(\boldsymbol{\theta}_m)$, and the posterior curvature by $\left.\frac{\partial^2 \log g_m(\boldsymbol{\theta}_m)}{\partial \boldsymbol{\theta}_m \partial \boldsymbol{\theta}_m^T}\right|_{\boldsymbol{\theta}_m = \tilde{\boldsymbol{\theta}}_m}$. We then set $\tilde{\boldsymbol{\mu}}_m$ to be the posterior mode and $\tilde{\boldsymbol{\Sigma}}_m$ to be the negative inverse of the posterior curvature. We compute the nested importance sampling approximation to the marginal likelihood of model m using the algorithm on page 83 where $\boldsymbol{\mu} = \tilde{\boldsymbol{\mu}}_m$ and $\boldsymbol{\Sigma} = \tilde{\boldsymbol{\Sigma}}_m$. Suppose that this requires W iterations to converge, hence requires W evaluations of $g_m()$. Denote this approximation as $\hat{I}_{m,3}$. We generate a samples of size $N_\Pi = \lfloor \frac{W}{4} \rfloor$ and $N_H = \lfloor \frac{W}{4} \rfloor$ from the posterior distribution of model m and $\mathrm{N}(\mathbf{0}, \mathbf{I}_k)$. We set $n_\Pi = n_H = \frac{N_H}{2}$.

We compute the bridge sampling approximation to the marginal likelihood using the algorithm of Section 4.2.3. Denote this approximation as $\hat{I}_{m,1}$. We also compute the bridge sampling approximation using the algorithm on page 71 where $\boldsymbol{\mu} = \tilde{\boldsymbol{\mu}}_m$ and $\boldsymbol{\Sigma} = \tilde{\boldsymbol{\Sigma}}_m$. Denote this approximation as $\hat{I}_{m,2}$. Similar to in Section 4.4, we configure the bridge sampling approximations so that they involve no more evaluations of the posterior pdf than are required for the nested importance sampling approximation. Note that the computational expense of nested importance sampling is still less than that of the bridge sampling approximation since nested importance sampling does not require a sample to be generated from the posterior distribution. We repeat this process 500 times.

Figure 4.22: Boxplots of the 500 relative approximations to the marginal likelihood using bridge sampling when the mode and curvature are available, $\hat{I}_2$, and unavailable, $\hat{I}_1$, for the five models from the Turtle Dataset. $\hat{I}_3$ is the nested importance sampling approximation to I when the mode and curvature are available

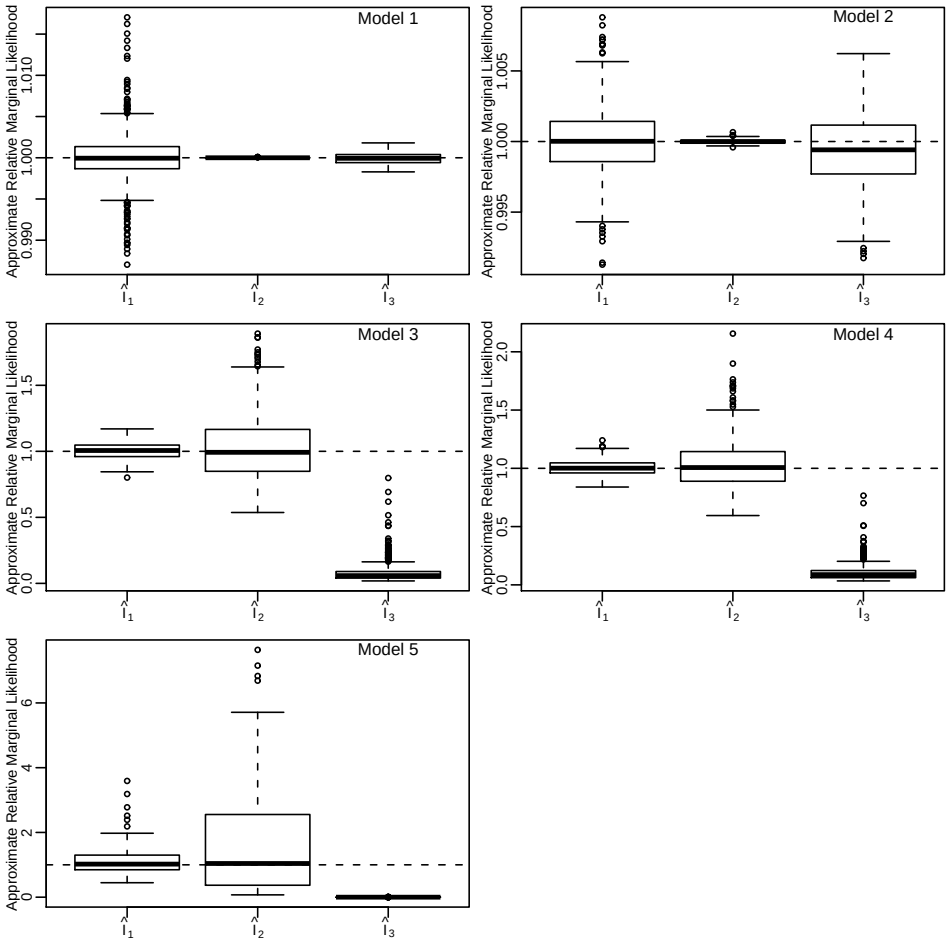


Figure 4.22 shows boxplots of the relative approximations for each of the five models using the three different approaches. The true values of the marginal likelihood are given in Table 4.1. Figure 4.22 shows that bridge sampling using the mode and curvature performs better than bridge sampling using the mean and variance for Models 1 and 2, i.e. the GLMs. But as we expected, when the target distribution is a posterior distribution of a GLMM, the bridge sampling approach using the mean and variance outperforms that using the mode and curva-

ture. In the case of Model 5, this outperformance is quite marked. Using nested importance sampling using the mode and curvature instead of the mean and variance again appears to result in underestimation of the marginal likelihood. This underestimation becomes very large for non-GLMs.

Our conclusion from this Section is that our implementation of bridge sampling presented in Section 4.2.3 is the "best" method for approximating the marginal likelihood of a GLMM.

4.7 Discussion

In this Chapter we investigated bridge sampling and nested importance sampling as methods for approximating the unknown normalising constant of a probability distribution and, in particular, the application of approximating the marginal likelihood of a GLMM. Both methods rely on having some information (i.e. mean/mode and variance) about the posterior distribution. We found in Section 4.6 that using a posterior sample to find the mean and variance was preferable to using the mode and curvature. By using software packages such as WinBUGS, it is relatively easy to generate a posterior sample from these models and we use this to gain insight about the posterior (e.g. by using the sample statistics). We developed bridge sampling and nested importance sampling strategies to approximate the marginal likelihood using posterior samples and not the posterior mode and curvature.

In Section 4.2 it was shown that if we use the same posterior sample, to gain insight about the posterior distribution, and in the bridge sampler then this led to an approximation that would underestimate the true marginal likelihood. This underestimation increased with the dimension of the problem. We developed a version of Warp III bridge sampling that did not underestimate the true marginal likelihood. This is presented in Section 4.2.3.

In Section 4.3 we discussed the relatively new method of nested importance sampling and developed a method that did not require the posterior mode to be found deterministically.

In Sections 4.4 and 4.5 we compared our implementations of bridge sampling and nested importance sampling based on the same number of evaluations of the unnormalised pdf using empirical studies. Nested importance sampling performed better than bridge sampling when the posterior is normal or approximately normal but underestimated the marginal likelihood for non-normal posteriors in a non-negligible way. We determined that this non-negligible underestimation is caused by two mechanisms: 1) the non-normality of the posterior, and 2) approximating $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ using a sample generated from Π.

Since nested importance sampling outperforms bridge sampling when the posterior distribution is approximately normal, a possible strategy once the posterior sample is generated is to use some statistical test, e.g. the Shapiro-Wilks test, to test whether the posterior is normal. If, using this test, we determine that the posterior is approximately normal then we may use nested importance sampling to approximate the marginal likelihood. If the test shows that there is a departure from the normal distribution then we use bridge sampling.

However, this is an ad-hoc approach and we have shown that the bias found in the nested importance sampling approximation to the marginal likelihood of a non-normal posterior distribution are non-negligible, whereas the bias found in the bridge sampling approximation to the marginal likelihood is negligible. For this reason, we recommend the use of bridge sampling for approximating all marginal likelihoods.

Chapter 5

Reversible Jump MCMC for GLMMs

5.1 Introduction

In Chapter 4, we investigated bridge sampling and nested importance sampling for approximating integrals for the application of evaluating the marginal likelihood, $f_m(\mathbf{y})$, of model $m \in M$, where

$$f_m(\mathbf{y}) = \int_{\Theta_m} f_m(\mathbf{y}|\boldsymbol{\theta}_m) f_m(\boldsymbol{\theta}_m) \mathrm{d}\boldsymbol{\theta}_m.$$

We now focus specifically on this application for GLMMs. Bridge sampling requires a sample to be generated from the posterior distribution. Nested importance sampling does not directly require a posterior sample but we do need some information (i.e. location and spread) about the posterior with which to implement this method successfully. In fact, all of the Monte Carlo methods described in 2.2.5, either directly require a posterior sample or some information about the posterior distribution. For GLMMs which can be of high dimension, we are unsure whether the mode of the posterior distribution and the curvature of the posterior distribution at the mode will contain sufficient information. In Chapter 4, we suggested using a posterior sample to approximate the posterior mean and variance.

Generating a posterior sample from each model $m \in M$ and then using bridge sampling (or any Monte Carlo method) to approximate $f_m(\mathbf{y})$ can quickly become impractical as the number of models, $|M|$, in M grows large. We will effectively waste a lot of computational resources on generating posterior samples from models with low or negligible posterior model probabilities. A suitable approach would be to use some method to identify a subset, $M^* \subset M$, of models that have high posterior model probabilities and then approximate the marginal likelihoods of the models in M^* using bridge sampling. A possible method for identifying M^* would be MCMC model determination. An MCMC model determination method based on the parameters $\boldsymbol{\beta}_m, \mathbf{u}_m, \mathbf{D}_m, \phi_m, m|\mathbf{y}$ is difficult to implement successfully due to the high dimensional jumps that are involved. For instance, suppose we have two models, m_1 and m_2, such that $q_{m_2} = q_{m_1} + 1$ and $p_{m_2} = p_{m_1}$. A jump from m_1 to m_2 is equivalent to adding a group-specific parameter. The difference in dimension between m_1 to m_2 without integrating out the group-specific parameters is $G + q_{m_1} + 1$. However, the difference in dimensionality

having integrated out the group-specific parameters is $q_{m_1} + 1$. It is easier to implement a reversible jump scheme where the difference in dimensionality is small.

Consider the integrated likelihood function

$$f_m(\mathbf{y}|\boldsymbol{\beta}_m, \mathbf{D}_m, \phi_m) = \int_{\mathbb{R}^{Gq}} f_m(\mathbf{y}|\boldsymbol{\beta}_m, \mathbf{u}_m, \phi_m) f_m(\mathbf{u}_m|\mathbf{D}_m) \mathrm{d}\mathbf{u}_m, \tag{5.1}$$

and the pdf

$$\begin{aligned} f_m(\boldsymbol{\beta}_m, \mathbf{D}_m, \phi_m|\mathbf{y}) &= \int_{\mathbb{R}^{Gq}} f_m(\boldsymbol{\beta}_m, \mathbf{u}_m, \mathbf{D}_m, \phi_m|\mathbf{y}) \mathrm{d}\mathbf{u}_m, \\ &\propto f_m(\mathbf{y}|\boldsymbol{\beta}_m, \mathbf{D}_m, \phi_m) f_m(\boldsymbol{\beta}_m, \mathbf{D}_m, \phi_m). \end{aligned}$$

of the resulting *integrated posterior distribution* which is just the joint marginal posterior distribution of $\boldsymbol{\beta}_m, \mathbf{D}_m, \phi_m|\mathbf{y}$. The dimension of the integrated posterior distribution is now either $p_m + \frac{1}{2}q_m(q_m + 1) + 1$ or $p_m + \frac{1}{2}q_m(q_m + 1)$, depending on whether ϕ_m is unknown or known, respectively. Since the number of groups, G, is typically the main reason for the high dimensionality of GLMMs, by using the integrated likelihood and the resulting integrated posterior we significantly reduce the dimensionality of the model and so make MCMC model determination easier. However, the integrated likelihood is rarely analytically tractable and therefore requires approximation. Since we are using MCMC model determination we may need to evaluate $f_m(\mathbf{y}|\boldsymbol{\beta}_m, \mathbf{D}_m, \phi_m)$ many times. Therefore the approximation will need to be computationally inexpensive. It also does not need to be of the highest accuracy since we are only using this method to identify M^*. Cai and Dunson (2006) propose an MCMC model determination method that uses a computationally inexpensive, deterministic approximation to the integrated likelihood and a Stochastic Search Variable Selection (SSVS) algorithm. We propose an alternative to this method that uses a Laplace approximation to the integrated likelihood and a reversible jump algorithm. The reversible jump algorithm is an adaption of an existing method proposed by Gill (2007) for model determination amongst GLMs.

In Section 5.2, we describe the method of Cai and Dunson (2006) for approximating the integrated likelihood and then describe our Laplace approximation. We also conduct a comparison of the two competing methods. In Section 5.3, we describe the reversible jump algorithm of Gill (2007) for GLMs and in Section 5.4, how this can be extended to GLMMs.

We propose this reversible jump algorithm for model determination amongst GLMMs as an alternative to the SSVS algorithm of Cai and Dunson (2006).

The main disadvantage of the SSVS method of Cai and Dunson (2006) is it does not directly approximate the posterior model probabilities. Instead it generates a posterior sample from the most complicated model possible, i.e. all of the available explanatory variables are included in $\mathbf{X}$ and $\mathbf{Z}$. It generates a sampled value of 0 for a parameter which is associated with the regression parameter or group-specific parameter for an explanatory variable according to the posterior probability of that parameter being 0.

Another disadvantage of the SSVS algorithm of Cai and Dunson (2006) is that it is possible for a group-specific parameter to be non-zero when the associated regression parameter is zero. This contradicts our assumption in Section 1.2.1 that the columns of $\mathbf{Z}_i$ are a subset of

the columns of $\mathbf{X}_i$. However, it may be possible to construct a conditional prior distribution for the elements of $\mathbf{D}$, so that this is impossible.

5.2 Approximating the Integrated Likelihood

5.2.1 Introduction

In this Section we investigate the issue of approximating the integrated likelihood (5.1) of a model $m \in M$. We discuss two methods: the *Cai & Dunson* method of Cai and Dunson (2006) and the Laplace method. Both rely on a 2nd order Taylor series expansion of the first stage likelihood, $f_m(\mathbf{y}|\boldsymbol{\beta}_m, \mathbf{u}_m, \phi_m)$. However, the Cai & Dunson method is based on an expansion of the untransformed first-stage likelihood function, whereas the Laplace method is based on an expansion of the log of the first-stage likelihood function. For the remainder of this Section, we suppress the dependence on the model m by removing the subscript m.

5.2.2 The Cai & Dunson Method

First, we begin by noting that the integrated likelihood function (5.1) can be written

$$f(\mathbf{y}|\boldsymbol{\beta}, \mathbf{D}, \phi) = \mathrm{E}\left(f(\mathbf{y}|\boldsymbol{\beta}, \mathbf{u}, \phi)\right), \tag{5.2}$$

where $\mathbf{u} \sim \mathrm{N}(\mathbf{0}, \mathbf{D}^*)$, i.e. the integrated likelihood is the expectation of the first-stage likelihood with respect to the prior of the group-specific parameters, $\mathbf{u}$.

The 2nd order Taylor series expansion of the first-stage likelihood with respect to $\mathbf{u}$ about the prior mean, $\mathbf{u} = \mathbf{0}$, is

$$\begin{aligned} f(\mathbf{y}|\boldsymbol{\beta}, \mathbf{u}, \phi) &\approx f(\mathbf{y}|\boldsymbol{\beta}, \mathbf{u}, \phi)|_{\mathbf{u}=\mathbf{0}} + \left.\frac{\partial f(\mathbf{y}|\boldsymbol{\beta}, \mathbf{u}, \phi)}{\partial \mathbf{u}}\right|_{\mathbf{u}=\mathbf{0}} \mathbf{u} + \frac{1}{2}\mathbf{u}^T \left.\frac{\partial^2 f(\mathbf{y}|\boldsymbol{\beta}, \mathbf{u}, \phi)}{\partial \mathbf{u}\partial \mathbf{u}^T}\right|_{\mathbf{u}=\mathbf{0}} \mathbf{u}, \\ &= f(\mathbf{y}|\boldsymbol{\beta}, \mathbf{u}, \phi)|_{\mathbf{u}=\mathbf{0}} \times \left[1 + \left.\frac{\partial \log f(\mathbf{y}|\boldsymbol{\beta}, \mathbf{u}, \phi)}{\partial \mathbf{u}}\right|_{\mathbf{u}=\mathbf{0}} \mathbf{u} \right. \\ &\quad + \frac{1}{2}\mathbf{u}^T \left(\frac{\partial \log f(\mathbf{y}|\boldsymbol{\beta}, \mathbf{u}, \phi)}{\partial \mathbf{u}} \frac{\partial \log f(\mathbf{y}|\boldsymbol{\beta}, \mathbf{u}, \phi)}{\partial \mathbf{u}^T} \right. \\ &\quad \left. \left. + \mathrm{DG}\left(\frac{\partial^2 \log f(\mathbf{y}|\boldsymbol{\beta}, \mathbf{u}, \phi)}{\partial \mathbf{u}\partial \mathbf{u}^T} \right) \right) \right|_{\mathbf{u}=\mathbf{0}} \mathbf{u} \Bigg], \end{aligned}$$

where $\mathrm{DG}(A)$ denotes the diagonal matrix consisting of the diagonal entries of A. The last equality above follows from

$$\frac{\partial f(\mathbf{y}|\boldsymbol{\beta}, \mathbf{u}, \phi)}{\partial \mathbf{u}} = f(\mathbf{y}|\boldsymbol{\beta}, \mathbf{u}, \phi) \frac{\partial \log f(\mathbf{y}|\boldsymbol{\beta}, \mathbf{u}, \phi)}{\partial \mathbf{u}},$$

and

$$\frac{\partial^2 f(\mathbf{y}|\boldsymbol{\beta},\mathbf{u},\phi)}{\partial \mathbf{u}\partial \mathbf{u}^T} = f(\mathbf{y}|\boldsymbol{\beta},\mathbf{u},\phi)\left[\frac{\partial f(\mathbf{y}|\boldsymbol{\beta},\mathbf{u},\phi)}{\partial \mathbf{u}}\frac{\partial f(\mathbf{y}|\boldsymbol{\beta},\mathbf{u},\phi)}{\partial \mathbf{u}^T} + \mathrm{DG}\left(\frac{\partial^2 \log f(\mathbf{y}|\boldsymbol{\beta},\mathbf{u},\phi)}{\partial \mathbf{u}\partial \mathbf{u}^T}\right)\right].$$

By changing the variable of differentiation from the group-specific parameters, $\mathbf{u}$, to the linear predictor, $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}$, we get

$$\begin{aligned} f(\mathbf{y}|\boldsymbol{\beta},\mathbf{u},\phi) \;\approx\; & f(\mathbf{y}|\boldsymbol{\beta},\mathbf{u},\phi)|_{\mathbf{u}=\mathbf{0}} \times \left[1 + \left.\frac{\partial \log f(\mathbf{y}|\boldsymbol{\beta},\mathbf{u},\phi)}{\partial \boldsymbol{\eta}}\right|_{\mathbf{u}=\mathbf{0}} \mathbf{Z}\mathbf{u}\right. \\ & + \frac{1}{2}\mathbf{u}^T\mathbf{Z}^T\left(\frac{\partial \log f(\mathbf{y}|\boldsymbol{\beta},\mathbf{u},\phi)}{\partial \boldsymbol{\eta}}\frac{\partial \log f(\mathbf{y}|\boldsymbol{\beta},\mathbf{u},\phi)}{\partial \boldsymbol{\eta}^T}\right. \\ & \left.\left.+\mathrm{DG}\left(\frac{\partial^2 \log f(\mathbf{y}|\boldsymbol{\beta},\mathbf{u},\phi)}{\partial \boldsymbol{\eta}\partial \boldsymbol{\eta}^T}\right)\right)\right|_{\mathbf{u}=\mathbf{0}} \mathbf{Z}\mathbf{u}\Bigg]. \end{aligned} \tag{5.3}$$

We can now approximate the expectation (5.2) by using (5.3) and noting that the expectation of a quadratic form, $\mathbf{u}^T\mathbf{R}\mathbf{u}$, where $\mathbf{u} \sim \mathrm{N}(\mathbf{0}, \mathbf{D}^*)$ is $\mathrm{E}(\mathbf{u}^T\mathbf{R}\mathbf{u}) = \mathrm{tr}(\mathbf{R}\mathbf{D}^*)$, to find the Cai & Dunson approximation to the integrated likelihood:

$$\begin{aligned} \hat{f}_{CD}(\mathbf{y}|\boldsymbol{\beta},\mathbf{D},\phi) \;=\; & f(\mathbf{y}|\boldsymbol{\beta},\mathbf{u},\phi)|_{\mathbf{u}=\mathbf{0}} \times \left[1 + \frac{1}{2}\mathrm{tr}\left(\mathbf{Z}^T\left(\frac{\partial \log f(\mathbf{y}|\boldsymbol{\beta},\mathbf{u},\phi)}{\partial \boldsymbol{\eta}}\frac{\partial \log f(\mathbf{y}|\boldsymbol{\beta},\mathbf{u},\phi)}{\partial \boldsymbol{\eta}^T}\right.\right.\right. \\ & \left.\left.\left.+\mathrm{DG}\left(\frac{\partial^2 \log f(\mathbf{y}|\boldsymbol{\beta},\mathbf{u},\phi)}{\partial \boldsymbol{\eta}\partial \boldsymbol{\eta}^T}\right)\right)\right|_{\mathbf{u}=\mathbf{0}} \mathbf{Z}\mathbf{D}^*\right)\Bigg]. \end{aligned} \tag{5.4}$$

Cai and Dunson (2006) show that the approximation (5.4) may be expressed as

$$\hat{f}_{CD}(\mathbf{y}|\boldsymbol{\beta},\mathbf{D},\phi) = f(\mathbf{y}|\boldsymbol{\beta},\mathbf{u},\phi)|_{\mathbf{u}=\mathbf{0}} \times \left[1 + \frac{1}{2\phi}\left(\sum_{k=1}^{q}\mathbf{D}_{kk}\sum_{i=1}^{G}B_{i,k}^{(1)} + 2\sum_{k=1}^{q-1}\sum_{j=k+1}^{q}\mathbf{D}_{jk}\sum_{i=1}^{G}B_{i,j,k}^{(2)}\right)\right], \tag{5.5}$$

where $B_{i,k}^{(1)}$ and $B_{i,j,k}^{(2)}$ are functions of $\boldsymbol{\beta}$ and $\mathbf{y}$.

Consider the case when $q = 1$, where $\mathbf{D} = \tau^2$ is scalar, then

$$\hat{f}_{CD}(\mathbf{y}|\boldsymbol{\beta},\tau^2,\phi) = f(\mathbf{y}|\boldsymbol{\beta},\mathbf{u},\phi)|_{\mathbf{u}=\mathbf{0}} \times \left[1 + \frac{\tau^2}{2\phi}\sum_{i=1}^{G}B_i^{(1)}\right]. \tag{5.6}$$

From (5.6) we see that when $q = 1$, $\hat{f}_{CD}(\mathbf{y}|\boldsymbol{\beta},\tau^2,\phi)$ is a linearly monotonic function of τ^2. It is increasing if $\sum_{i=1}^{G}B_i^{(1)} > 0$, and decreasing if $\sum_{i=1}^{G}B_i^{(1)} < 0$.

More generally, $\hat{f}_{CD}(\mathbf{y}|\boldsymbol{\beta},\mathbf{D},\phi)$ is a monotonic function of $\mathbf{D}_{jk}$, for $j,k = 1,...,q$. It is an increasing function of $\mathbf{D}_{kk}$ if $\sum_{i=1}^{G}B_{i,k}^{(1)} > 0$, decreasing if $\sum_{i=1}^{G}B_{i,k}^{(1)} < 0$, and it is an increasing function of $\mathbf{D}_{jk}$ if $\sum_{i=1}^{G}B_{i,j,k}^{(2)} > 0$, decreasing if $\sum_{i=1}^{G}B_{i,j,k}^{(2)} < 0$.

In fact, if $\sum_{i=1}^{G}B_{i,k}^{(1)} < 0$ or $\sum_{i=1}^{G}B_{i,j,k}^{(2)} < 0$ then $\hat{f}_{CD}(\mathbf{y}|\boldsymbol{\beta},\mathbf{D},\phi)$ can be negative.

If $\sum_{i=1}^{G}B_{i,k}^{(1)} > 0$ or $\sum_{i=1}^{G}B_{i,j,k}^{(2)} > 0$, then $\hat{f}_{CD}(\mathbf{y}|\boldsymbol{\beta},\mathbf{D},\phi)$ is a monotonically increasing function of elements of $\mathbf{D}$. This can pose a particular problem for the reversible jump algorithm

we present in Section 5.4 since we will need to maximise the approximation to the integrated posterior pdf, $\hat{f}_{CD}(\boldsymbol{\beta}, \mathbf{D}, \phi|\mathbf{y}) \propto \hat{f}_{CD}(\mathbf{y}|\boldsymbol{\beta}, \mathbf{D}, \phi) f(\boldsymbol{\beta}, \mathbf{D}, \phi)$, and this may not be possible with finite elements of $\mathbf{D}$ unless the prior for $\mathbf{D}$ is sufficiently informative. We consider this problem. Assume that the dispersion parameter, ϕ, is known to be one, and that the prior distribution of $\boldsymbol{\beta}$ and $\mathbf{D}$ are independent. The marginal likelihood is approximated by

$$\begin{aligned}
\hat{f}(\mathbf{y}) &= \int_{\mathbb{R}^p} \int_{\mathbb{P}^q} \hat{f}_{CD}(\mathbf{y}|\boldsymbol{\beta}, \mathbf{D}) f(\boldsymbol{\beta}) f(\mathbf{D}) \mathrm{d}\mathbf{D} \mathrm{d}\boldsymbol{\beta}, \\
&= \int_{\mathbb{R}^p} \int_{\mathbb{P}^q} f(\mathbf{y}|\boldsymbol{\beta}, \mathbf{u})|_{\mathbf{u}=\mathbf{0}} \times \\
&\quad \left[1 + \frac{1}{2} \left(\sum_{k=1}^{q} \mathbf{D}_{kk} \sum_{i=1}^{G} B_{i,k}^{(1)} + 2 \sum_{k=1}^{q-1} \sum_{j=k+1}^{q} \mathbf{D}_{jk} \sum_{i=1}^{G} B_{i,j,k}^{(2)} \right) \right] f(\boldsymbol{\beta}) f(\mathbf{D}) \mathrm{d}\mathbf{D} \mathrm{d}\boldsymbol{\beta}, \\
&= \int_{\mathbb{R}^p} f(\mathbf{y}|\boldsymbol{\beta}, \mathbf{u})|_{\mathbf{u}=\mathbf{0}} f(\boldsymbol{\beta}) \mathrm{d}\boldsymbol{\beta} + \frac{1}{2} \sum_{k=1}^{q} \mathrm{E}(\mathbf{D}_{kk}) \int_{\mathbb{R}^p} \sum_{i=1}^{G} B_{i,k}^{(1)} f(\mathbf{y}|\boldsymbol{\beta}, \mathbf{u})|_{\mathbf{u}=\mathbf{0}} f(\boldsymbol{\beta}) \mathrm{d}\boldsymbol{\beta} \\
&\quad + \sum_{k=1}^{q-1} \sum_{j=k+1}^{q} \mathrm{E}(\mathbf{D}_{jk}) \int_{\mathbb{R}^p} \sum_{i=1}^{G} B_{i,j,k}^{(2)} f(\mathbf{y}|\boldsymbol{\beta}, \mathbf{u})|_{\mathbf{u}=\mathbf{0}} f(\boldsymbol{\beta}) \mathrm{d}\boldsymbol{\beta}.
\end{aligned}$$

Assuming that the integrals $\int_{\mathbb{R}^p} f(\mathbf{y}|\boldsymbol{\beta}, \mathbf{u})|_{\mathbf{u}=\mathbf{0}} f(\boldsymbol{\beta}) \mathrm{d}\boldsymbol{\beta}$, $\int_{\mathbb{R}^p} \sum_{i=1}^{G} B_{i,k}^{(1)} f(\mathbf{y}|\boldsymbol{\beta}, \mathbf{u})|_{\mathbf{u}=\mathbf{0}} f(\boldsymbol{\beta}) \mathrm{d}\boldsymbol{\beta}$, and $\int_{\mathbb{R}^p} \sum_{i=1}^{G} B_{i,j,k}^{(2)} f(\mathbf{y}|\boldsymbol{\beta}, \mathbf{u})|_{\mathbf{u}=\mathbf{0}} f(\boldsymbol{\beta}) \mathrm{d}\boldsymbol{\beta}$ exist, the approximation to the marginal likelihood does not exist if the prior mean, $\mathrm{E}(\mathbf{D})$, of $\mathbf{D}$ does not exist. Therefore, a necessary condition for $\hat{f}(\mathbf{y})$ to exist, is that the prior mean of $\mathbf{D}$ exists. The prior mean of the unit information prior distribution for $\mathbf{D}$, proposed in Chapter 3, exists meaning we may use the Cai & Dunson approximation in conjunction with our proposed priors. However, we find it worrying that the integrated likelihood is a monotonic function of the elements of $\mathbf{D}$. In the next Section, we describe an alternative method for approximating the integrated likelihood and in Section 5.2.4 we undertake a small empirical comparison of the two methods.

5.2.3 The Laplace Method

The Laplace method has been used previously to approximate the integrated likelihood with a view to finding maximum likelihood estimates of the model parameters (see, for example, Breslow and Clayton (1993)). Since $\mathbf{u}_i \overset{\text{iid}}{\sim} \mathrm{N}(\mathbf{0}, \mathbf{D})$ for $i = 1, ..., G$, then

$$f(\mathbf{y}|\boldsymbol{\beta}, \mathbf{D}, \phi) = \prod_{i=1}^{G} \int_{\mathbb{R}^q} f(\mathbf{y}_i|\boldsymbol{\beta}, \mathbf{u}_i, \phi) f(\mathbf{u}_i|\mathbf{D}) \mathrm{d}\mathbf{u}_i. \tag{5.7}$$

This changes the problem from approximating a Gq-dimensional integral to approximating G q-dimensional integrals.

Let $g(\mathbf{u}_i) = f(\mathbf{y}_i|\boldsymbol{\beta}, \mathbf{u}_i, \phi) f(\mathbf{u}_i|\mathbf{D})$ denote the ith integrand in (5.7). The 2nd order Taylor series expansion of $\log g(\mathbf{u}_i)$ with respect to $\mathbf{u}_i$ about the value, $\hat{\mathbf{u}}_i$, that maximises $\log g(\mathbf{u}_i)$ is

$$\log g(\mathbf{u}_i) \approx \log g(\mathbf{u}_i)|_{\mathbf{u}_i=\hat{\mathbf{u}}_i} + \frac{1}{2} (\mathbf{u}_i - \hat{\mathbf{u}}_i)^T \left. \frac{\partial^2 \log g(\mathbf{u}_i)}{\partial \mathbf{u}_i \partial \mathbf{u}_i^T} \right|_{\mathbf{u}_i=\hat{\mathbf{u}}_i} (\mathbf{u}_i - \hat{\mathbf{u}}_i). \tag{5.8}$$

By exponentiating both sides of (5.8) we get the following approximation to $g(\mathbf{u}_i)$

$$g(\mathbf{u}_i) \approx g(\mathbf{u}_i)|_{\mathbf{u}_i=\hat{\mathbf{u}}_i} \exp\left(-\frac{1}{2}(\mathbf{u}_i-\hat{\mathbf{u}}_i)^T\mathbf{V}_i(\mathbf{u}_i-\hat{\mathbf{u}}_i)\right), \tag{5.9}$$

where

$$\begin{aligned}\mathbf{V}_i &= -\left.\frac{\partial^2 \log g(\mathbf{u}_i)}{\partial \mathbf{u}_i \partial \mathbf{u}_i^T}\right|_{\mathbf{u}_i=\hat{\mathbf{u}}_i}, \\ &= -\left.\frac{\partial^2 \log f(\mathbf{y}_i|\boldsymbol{\beta},\mathbf{u}_i,\phi)}{\partial \mathbf{u}_i \partial \mathbf{u}_i^T}\right|_{\mathbf{u}_i=\hat{\mathbf{u}}_i} + \mathbf{D}^{-1}.\end{aligned}$$

Therefore,

$$f(\mathbf{y}|\boldsymbol{\beta},\mathbf{D},\phi) = \prod_{i=1}^{G}\int_{\mathbb{R}^q} g(\mathbf{u}_i)|_{\mathbf{u}_i=\hat{\mathbf{u}}_i} \exp\left(-\frac{1}{2}(\mathbf{u}_i-\hat{\mathbf{u}}_i)^T\mathbf{V}_i(\mathbf{u}_i-\hat{\mathbf{u}}_i)\right) \mathrm{d}\mathbf{u}_i, \tag{5.10}$$

and we can perform the G integrations in (5.10) exactly. Therefore, the Laplace approximation to the integrated likelihood is

$$\hat{f}_L(\mathbf{y}|\boldsymbol{\beta},\mathbf{D},\phi) = |\mathbf{D}|^{-\frac{G}{2}} \prod_{i=1}^{G} f(\mathbf{y}_i|\boldsymbol{\beta},\mathbf{u}_i,\phi)|_{\mathbf{u}_i=\hat{\mathbf{u}}_i} |\mathbf{V}_i|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\hat{\mathbf{u}}_i^T\mathbf{D}^{-1}\hat{\mathbf{u}}_i\right). \tag{5.11}$$

Note that we can find $\hat{\mathbf{u}}_i$ using the Newton-Raphson method. The first and second derivatives for this method are given by

$$\frac{\partial \log g(\mathbf{u}_i)}{\partial \mathbf{u}_i} = \frac{\partial \log f(\mathbf{y}_i|\boldsymbol{\beta},\mathbf{u}_i,\phi)}{\partial \mathbf{u}_i} - \mathbf{D}^{-1}\mathbf{u}_i,$$

and

$$\frac{\partial^2 \log g(\mathbf{u}_i)}{\partial \mathbf{u}_i \partial \mathbf{u}_i^T} = \frac{\partial^2 \log f(\mathbf{y}_i|\boldsymbol{\beta},\mathbf{u}_i,\phi)}{\partial \mathbf{u}_i \partial \mathbf{u}_i^T} - \mathbf{D}^{-1},$$

respectively. An alternative to the Newton-Raphson method would be to replace the matrix of second derivatives by its expected value, resulting in the Fisher scoring method.

5.2.4 Comparison of the Cai & Dunson and Laplace Methods

Initially, we note that the Laplace method will be more computationally expensive than the Cai & Dunson method. This is because for each i we need to find $\hat{\mathbf{u}}_i$ for $i = 1, .., G$. As suggested in Section 5.2.3, we can use the Newton-Raphson method since both the first and second derivatives of $\log g(\mathbf{u}_i)$ are available. The Laplace method is exact when the first-stage likelihood is normal and works best when the first-stage likelihood is approximately normal. It is also true, that if the first-stage likelihood is normal than the Newton-Raphson method will converge in one iteration and O'Hagan and Forster (2004) state that if the first-stage likelihood is approximately normal than the Newton-Raphson method will converge rapidly. We know that the ith contribution to the first-stage likelihood will approach normality as $n_i \to \infty$ and is approximately normal for large values of n_i.

In Section 5.2.2, we described how the Cai & Dunson approximation, $\hat{f}_{CD}(\mathbf{y}|\boldsymbol{\beta}, \mathbf{D}, \phi)$, is a monotonic function of the elements of $\mathbf{D}$. This leads to the necessary condition that $\mathrm{E}(\mathbf{D})$ need exist for $\hat{f}(\mathbf{y})$ to exist.

We undertake a comparison of the two methods. To do this we return to the Turtle Dataset example, and, in particular Model 4. Recall that Model 4 has the following linear predictor

$$\eta_{ij} = \beta_1 + \beta_2 x_{ij} + u_i,$$

where $u_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$ for $i = 1, ..., G$. The integrated likelihood is

$$\prod_{i=1}^{G} \int_{\mathbb{R}} f(\mathbf{y}_i|\boldsymbol{\beta}, u_i) f(u_i|\sigma^2) \mathrm{d}u_i. \tag{5.12}$$

The integrals in (5.12) are analytically intractable. However, since these integrals are one-dimensional we can use Simpson's rule to approximate $f(\mathbf{y}|\boldsymbol{\beta}, \sigma^2)$ to such a level of accuracy that the approximation can be regarded as exact. Denote this approximation to the integrated likelihood as $\hat{f}_S(\mathbf{y}|\boldsymbol{\beta}, \sigma^2)$. For the Simpson's rule given by (2.7), we use $n = 5000$ and we choose large values for a and b of ∓ 15 as suggested in Section 2.2.2, since u_i is unbounded. Denote the Cai & Dunson and Laplace approximations to the integrated likelihood as $\hat{f}_{CD}(\mathbf{y}|\boldsymbol{\beta}, \sigma^2)$ and $\hat{f}_L(\mathbf{y}|\boldsymbol{\beta}, \sigma^2)$, respectively.

Consider the *profile likelihood* which is defined as

$$f(\sigma^2) = f(\mathbf{y}|\boldsymbol{\beta}, \sigma^2)\big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}(\sigma^2)},$$

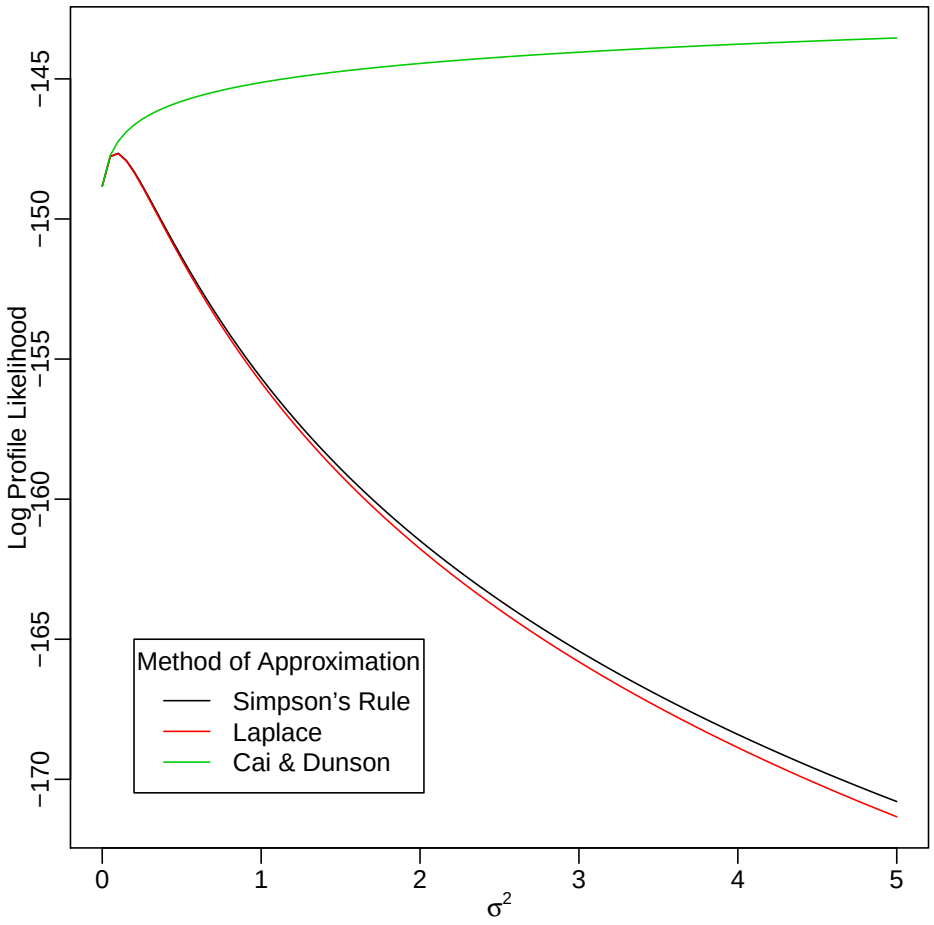
where $\hat{\boldsymbol{\beta}}(\sigma^2) = \arg\max f(\mathbf{y}|\boldsymbol{\beta}, \sigma^2)$ is the value that maximises $f(\mathbf{y}|\boldsymbol{\beta}, \sigma^2)$ when σ^2 is assumed fixed.

Figure 5.1 shows plots of the approximate log profile likelihood where the three different approximation methods have been used to approximate the integrated likelihood function. We regard the Simpson's rule approximation as exact and consider the Cai & Dunson and Laplace approximations. Both of the methods produce approximations that become closer to the true value as $\sigma^2 \to 0$. This is obvious for the Cai & Dunson method since the Taylor series expansion is taken about $u_i = 0$ which is guaranteed when $\sigma^2 = 0$. For the Laplace method, as $\sigma^2 \to 0$ the prior distribution for u_i becomes more informative and begins to dominate the ith contribution to the first-stage likelihood, $f(\mathbf{y}_i|\boldsymbol{\beta}, \sigma^2)$. This results in $g(u_i) = f(\mathbf{y}_i|\boldsymbol{\beta}, u_i) f(u_i|\sigma^2)$ becoming increasingly normal and the Laplace method becoming more accurate.

For larger values of σ^2, the Laplace method provides a reasonable approximation to the true profile likelihood. We can also see that the Cai & Dunson approximation produces an increasing profile likelihood function due to the property we discussed in Section 5.2.2.

Cai and Dunson (2006) compared the accuracy of the Laplace method and the Cai & Dunson method by approximating the marginal likelihood of a model and comparing those to the true value. We can assess the accuracy of the two methods by approximating the marginal

Figure 5.1: Plots of the approximate log profile likelihood against σ^2 using the three different approximation methods.



likelihood of Model 4. The marginal likelihood is

$$f(\mathbf{y}) = \int_{\mathbb{R}^2} \int_0^\infty f(\mathbf{y}|\boldsymbol{\beta}, \sigma^2) f(\boldsymbol{\beta}, \sigma^2) \mathrm{d}\sigma^2 \mathrm{d}\boldsymbol{\beta}. \tag{5.13}$$

We replace $f(\mathbf{y}|\boldsymbol{\beta}, \sigma^2)$ by either $\hat{f}_{CD}(\mathbf{y}|\boldsymbol{\beta}, \sigma^2)$ or $\hat{f}_L(\mathbf{y}|\boldsymbol{\beta}, \sigma^2)$ to give $\hat{f}_{CD}(\mathbf{y})$ or $\hat{f}_L(\mathbf{y})$, respectively. We use the unit information prior distribution for the regression parameters, $\boldsymbol{\beta}$, proposed in Chapter 3, i.e.

$$\boldsymbol{\beta} \sim \mathrm{N}\left(\mathbf{0}, \frac{\pi}{2} \begin{pmatrix} 1 & 0 \\ 0 & \frac{n}{n-1} \end{pmatrix}\right),$$

where recall that $n = 244$. We consider two alternatives prior distributions for σ^2: the unit information prior distribution proposed in Chapter 3, i.e. $\sigma^2 \sim \mathrm{IG}\left(\frac{3}{2}, \frac{\pi}{4}\right)$, and $\sigma^2 \sim \mathrm{IG}\left(20, \frac{\pi}{4}\right)$.

The prior distribution for σ^2 with shape parameter 20 has more mass at small values of σ^2 so both approximation methods should work better when the shape parameter is 20 than when the prior distribution is more diffuse. To approximate $\hat{f}_{CD}(\mathbf{y})$ and $\hat{f}_L(\mathbf{y})$, we use the transformation $\nu = \log \sigma^2$ with Jacobian $\exp(\nu)$ so that

$$\hat{f}(\mathbf{y}) = \int_{\mathbb{R}^3} \hat{f}(\mathbf{y}|\boldsymbol{\beta}, \exp(\nu)) f(\boldsymbol{\beta}) f(\exp(\nu)) \exp(\nu) \mathrm{d}\nu \mathrm{d}\boldsymbol{\beta}.$$

Table 5.1: Importance sampling approximations to the log of the marginal likelihood, $\log f_4(\mathbf{y})$, of Model 4 with the prior distributions: $\sigma^2 \sim \text{IG}\left(\alpha, \frac{\pi}{4}\right)$, where the integrated likelihood has been approximated deterministically.

Approximation Method	Shape Parameter, α	
	$\frac{3}{2}$	20
Simpson's Rule (Exact)	-154.8849	-153.1370
Cai & Dunson Method	-154.0511	-153.2240
Laplace Method	-154.9296	-153.1404

We then find the posterior mode, $\boldsymbol{\mu}$ and Hessian matrix, $-\boldsymbol{\Sigma}^{-1}$, numerically. Finally we use importance sampling with $H \equiv \text{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and a sample size of 100000. To find the 'true' value of $f(\mathbf{y})$ we replace $f(\mathbf{y}|\boldsymbol{\beta}, \tau^2)$ by $\hat{f}_S(\mathbf{y}|\boldsymbol{\beta}, \tau^2)$ and use importance sampling as described above. Table 5.1 shows the approximations to the marginal likelihood. As expected the accuracy of both methods increases as the prior distribution becomes less diffuse. The Laplace method performs well in both scenarios. However, the Cai & Dunson method only performs well when the prior distribution is concentrated near $\sigma^2 = 0$, and even in this case, the Laplace method is more accurate.

As mentioned above, Cai and Dunson (2006) undertook a similar empirical comparison of the Cai & Dunson and Laplace methods for simulated data. They found that the Cai & Dunson method produced more accurate approximations to the integrated likelihood than the Laplace method. It may be that the relative accuracy of the two methods is example-dependent. Our small comparison suggests that we should favour the Laplace method for approximating the integrated likelihood.

5.3 Reversible Jump for GLMs

In this Section we describe a reversible jump scheme for model determination amongst GLMs as proposed by Gill (2007).

We begin by briefly describing a GLM. Let y_i be the ith response for $i = 1, ..., n$ and let $\mathbf{x}_i$ denote the $p \times 1$ vector of regression covariates which correspond to the regression parameters, $\boldsymbol{\beta}$. We assume that Y_i is independently distributed from some exponential family distribution with density

$$f(y_i) = \exp\left[\frac{y_i\zeta_i - b(\zeta_i)}{a(\phi)} + c(y_i, \phi)\right],$$

where ζ_i is the canonical parameter, ϕ is the dispersion parameter and $a()$, $b()$ and $c()$ are known functions. Define $\mu_i = \text{E}(Y_i) = b'(\zeta_i)$ as the mean of Y_i. This is related to the ith component of the linear predictor, η_i, through

$$g(\mu_i) = \eta_i = \mathbf{x}_i^T\boldsymbol{\beta},$$

where $g()$ is the link function and $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression parameters. Let $\boldsymbol{\eta} = (\eta_1, ..., \eta_n)^T$ and $\mathbf{X} = (\mathbf{x}_1^T, ..., \mathbf{x}_n^T)^T$. It can be seen that a GLM is a special case of a GLMM where $\mathbf{u} \equiv \mathbf{0}$, and, equivalently, $\mathbf{D} \equiv \mathbf{0}$.

We assume that the marginal posterior distribution of the dispersion parameter, ϕ_m, remains approximately the same for all models $m \in M$, or at least does so for models with non-negligible posterior model probability. We justify this assumption as follows. Consider two models: m_1 and m_2, and a move from m_1 to m_2. Suppose the posterior distributions of ϕ under m_1 and m_2 are such that ϕ tends to be larger under m_2, then we would expect that the modelling of the mean for m_1 better describes the data than that for model m_2, and hence, we would not want to move to model m_2, and the difference in the distributions of ϕ is inconsequential. Now suppose that ϕ tends to be smaller under m_2 then we would expect that the modelling of the mean for m_2 better describes the data than that for m_1. In this case, the superior modelling of the mean makes the posterior model probability of m_1 negligible when compared to that of m_2 so we make the move regardless of the difference in the distributions of ϕ. A similar assumption is made by Papathomas et al. (2009) when they consider a reversible jump scheme for linear models.

We are considering a move between models $m_1 \in M$ and $m_2 \in M$. Without loss of generality, assume that m_1 is nested within m_2, so that $\mathbf{X}_{m_2} = [\mathbf{X}_{m_1}|\mathbf{S}]$ and $p_{m_2} > p_{m_1}$. We only consider local moves, so that m_1 and m_2 only differ by a single term, or interaction between terms.

Suppose that the current state of the MCMC chain is $(m_2, \boldsymbol{\beta}_{m_2}, \phi_{m_2})$ and we are interested in a move to model m_1 so we need to propose values for $\boldsymbol{\beta}_{m_1}$ and ϕ_{m_1}. This move is termed a *death move*, as we are decreasing the dimension of the model.

Let $\boldsymbol{\eta}_{m_2} = \mathbf{X}_{m_2}\boldsymbol{\beta}_{m_2}$ be the current linear predictor. A possible proposal would be to set the proposed linear predictor, $\boldsymbol{\eta}_{m_1}$, to be the orthogonal projection of the current linear predictor onto the subspace defined by model m_2. This projection is orthogonal with respect to an inner product, $\mathbf{W}$. So,

$$\boldsymbol{\beta}_{m_1} = \left(\mathbf{X}_{m_1}^T \mathbf{W} \mathbf{X}_{m_1}\right)^{-1} \mathbf{X}_{m_1}^T \mathbf{W} \mathbf{X}_{m_2} \boldsymbol{\beta}_{m_2}. \tag{5.14}$$

An approach is to set $\mathbf{W}$ to be an approximation to the inverse posterior covariance matrix of the working vector, $\tilde{\mathbf{y}}$, which has ith element

$$\tilde{y}_i = \eta_i + (y_i - \mu_i) g'(\mu_i).$$

It has expected value $\mathrm{E}(\tilde{\mathbf{Y}}) = \boldsymbol{\eta}$ and variance $\mathrm{var}(\tilde{\mathbf{Y}}) = \mathbf{W}^{-1} = \mathrm{diag}\{\mathrm{var}(Y_i) g'(\mu_i)^2\}$. Note that $\mathrm{E}(\tilde{\mathbf{Y}})$ and $\mathbf{W}$ depend on the unknown parameters $\boldsymbol{\beta}$ and ϕ. However, we prefer $\mathbf{W}$ not to depend on the current model parameters, $\boldsymbol{\beta}_{m_2}$ and ϕ_{m_2}. This allows the reverse move to be easily defined. Gill (2007) suggests fitting the most complicated model possible, $m^* \in M$, and setting $\hat{\boldsymbol{\beta}}_{m^*}$ and $\hat{\phi}_{m^*}$ to be the maximum likelihood estimates of $\boldsymbol{\beta}_{m^*}$ and ϕ_{m^*}, respectively. We then let

$$\hat{\mathbf{W}} = \mathrm{diag}\left\{\mathrm{var}(Y_i) g'(\mu_i)^2\right\}\Big|^{-1}_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_{m^*}, \phi=\hat{\phi}_{m^*}}.$$

An alternative is to set $\hat{\boldsymbol{\beta}}_{m^*}$ and $\hat{\phi}_{m^*}$ to be the posterior modes of $\boldsymbol{\beta}_{m^*}$ and ϕ_{m^*}, respectively. This value of $\hat{\mathbf{W}}$ is computed initially and remains fixed throughout the algorithm. Therefore

(5.14) becomes

$$\boldsymbol{\beta}_{m_1} = \left(\mathbf{X}_{m_1}^T \hat{\mathbf{W}} \mathbf{X}_{m_1}\right)^{-1} \mathbf{X}_{m_1}^T \hat{\mathbf{W}} \mathbf{X}_{m_2} \boldsymbol{\beta}_{m_2}. \tag{5.15}$$

For the dispersion parameter, we use the assumption that the marginal posterior distribution of ϕ remains approximately the same for different models and use the identity transformation, i.e.

$$\phi_{m_1} = \phi_{m_2}. \tag{5.16}$$

Recalling that $\mathbf{X}_{m_2} = (\mathbf{X}_{m_1}|\mathbf{S})$ and writing $\boldsymbol{\beta}_{m_2} = (\boldsymbol{\beta}_{m_2}^{(1)}, \boldsymbol{\beta}_{m_2}^{(2)})^T$, where $\boldsymbol{\beta}_{m_2}^{(1)}$ corresponds to $\mathbf{X}_{m_1}$ and $\boldsymbol{\beta}_{m_2}^{(2)}$ corresponds to $\mathbf{S}$, we see that (5.15) can be rewritten

$$\boldsymbol{\beta}_{m_1} = \boldsymbol{\beta}_{m_2}^{(1)} + \left(\mathbf{X}_{m_1}^T \hat{\mathbf{W}} \mathbf{X}_{m_1}\right)^{-1} \mathbf{X}_{m_1}^T \hat{\mathbf{W}} \mathbf{S} \boldsymbol{\beta}_{m_2}^{(2)}. \tag{5.17}$$

Note that $\boldsymbol{\beta}_{m_2}^{(1)}$ is a $p_{m_1} \times 1$ vector and $\boldsymbol{\beta}_{m_2}^{(2)}$ is a $(p_{m_2} - p_{m_1}) \times 1$ vector. This death move is entirely deterministic. For reversibility to hold, the *birth move* from m_1 to m_2 must satisfy (5.16) and (5.17). It is easy to see that (5.17) is satisfied if

$$\boldsymbol{\beta}_{m_2}^{(1)} = \boldsymbol{\beta}_{m_1} - \left(\mathbf{X}_{m_1}^T \hat{\mathbf{W}} \mathbf{X}_{m_1}\right)^{-1} \mathbf{X}_{m_1}^T \hat{\mathbf{W}} \mathbf{S} \boldsymbol{\beta}_{m_2}^{(2)}. \tag{5.18}$$

Let $\boldsymbol{\beta}_{m_2}^{(2)} = \mathbf{v}$, where $\mathbf{v}$ is generated from some distribution which we are free to choose. Therefore (5.16) and (5.17) are satisfied if

$$\begin{pmatrix} \boldsymbol{\beta}_{m_2} \\ \phi_{m_2} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\beta}_{m_2}^{(1)} \\ \boldsymbol{\beta}_{m_2}^{(2)} \\ \phi_{m_2} \end{pmatrix} = \begin{pmatrix} \mathbf{I}_{p_{m_1}} & -\left(\mathbf{X}_{m_1}^T \hat{\mathbf{W}} \mathbf{X}_{m_1}\right)^{-1} \mathbf{X}_{m_1}^T \hat{\mathbf{W}} \mathbf{S} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{p_{m_2} - p_{m_1}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & 1 \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_{m_1} \\ \mathbf{v} \\ \phi_{m_2} \end{pmatrix}.$$

The transformation is an upper triangular matrix with all diagonal elements equal to one. Hence, the Jacobian is

$$\begin{vmatrix} \mathbf{I}_{p_{m_1}} & -\left(\mathbf{X}_{m_1}^T \hat{\mathbf{W}} \mathbf{X}_{m_1}\right)^{-1} \mathbf{X}_{m_1}^T \hat{\mathbf{W}} \mathbf{S} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{p_{m_2} - p_{m_1}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & 1 \end{vmatrix} = 1.$$

We are now left with the choice of distribution for $\mathbf{v}$. Gill (2007) suggests setting a multivariate normal distribution for $\boldsymbol{\beta}_{m_2}$ and then inducing the distribution for $\mathbf{v}$, from this. We set the mean, $\boldsymbol{\mu}_{\boldsymbol{\beta}_{m_2}}$, and variance, $\boldsymbol{\Sigma}_{\boldsymbol{\beta}_{m_2}}$, of the distribution of $\boldsymbol{\beta}_{m_2}$ to be an approximate maximum likelihood estimate of $\boldsymbol{\beta}_{m_2}$ and an estimate of the inverse Fisher information matrix, respectively. That is

$$\boldsymbol{\beta}_{m_2} \sim \mathrm{N}\left(\left(\mathbf{X}_{m_2}^T \hat{\mathbf{W}} \mathbf{X}_{m_2}\right)^{-1} \mathbf{X}_{m_2}^T \hat{\mathbf{W}} \hat{\boldsymbol{\eta}}, \left(\mathbf{X}_{m_2}^T \hat{\mathbf{W}} \mathbf{X}_{m_2}\right)^{-1}\right),$$

where $\hat{\boldsymbol{\eta}} = \mathbf{X}_{m^*} \hat{\boldsymbol{\beta}}_{m^*}$. Recall that

$$\boldsymbol{\beta}_{m_2} = \begin{pmatrix} \mathbf{I}_{p_{m_1}} & -\left(\mathbf{X}_{m_1}^T \hat{\mathbf{W}} \mathbf{X}_{m_1}\right)^{-1} \mathbf{X}_{m_1}^T \hat{\mathbf{W}} \mathbf{S} \\ \mathbf{0} & \mathbf{I}_{p_{m_2} - p_{m_1}} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_{m_1} \\ \mathbf{v} \end{pmatrix},$$

therefore

$$\begin{pmatrix} \boldsymbol{\beta}_{m_1} \\ \mathbf{v} \end{pmatrix} = \begin{pmatrix} \mathbf{I}_{p_{m_1}} & \left(\mathbf{X}_{m_1}^T \hat{\mathbf{W}} \mathbf{X}_{m_1}\right)^{-1} \mathbf{X}_{m_1}^T \hat{\mathbf{W}} \mathbf{S} \\ \mathbf{0} & \mathbf{I}_{p_{m_2}-p_{m_1}} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_{m_2}^{(1)} \\ \boldsymbol{\beta}_{m_2}^{(2)} \end{pmatrix} = \mathbf{M}\boldsymbol{\beta}_{m_2},$$

where

$$\mathbf{M} = \begin{pmatrix} \mathbf{I}_{p_{m_1}} & \left(\mathbf{X}_{m_1}^T \hat{\mathbf{W}} \mathbf{X}_{m_1}\right)^{-1} \mathbf{X}_{m_1}^T \hat{\mathbf{W}} \mathbf{S} \\ \mathbf{0} & \mathbf{I}_{p_{m_2}-p_{m_1}} \end{pmatrix}.$$

Since $\boldsymbol{\beta}_{m_2}$ has a normal distribution, then so must $(\boldsymbol{\beta}_{m_1}, \mathbf{v})^T$, since it is an affine transformation of $\boldsymbol{\beta}_{m_2}$. The mean of $(\boldsymbol{\beta}_{m_1}, \mathbf{v})^T$ is $\boldsymbol{\mu}_{\boldsymbol{\beta}_{m_1},\mathbf{v}} = \mathbf{M}\boldsymbol{\mu}_{\boldsymbol{\beta}_{m_2}}$ and the variance is $\boldsymbol{\Sigma}_{\boldsymbol{\beta}_{m_1},\mathbf{v}} = \mathbf{M}\boldsymbol{\Sigma}_{\boldsymbol{\beta}_{m_2}}\mathbf{M}^T$. Now

$$\begin{aligned} \boldsymbol{\Sigma}_{\boldsymbol{\beta}_{m_2}} &= \left(\mathbf{X}_{m_2}^T \hat{\mathbf{W}} \mathbf{X}_{m_2}\right)^{-1}, \\ &= \begin{pmatrix} \mathbf{X}_{m_1}^T \hat{\mathbf{W}} \mathbf{X}_{m_1} & \mathbf{X}_{m_1}^T \hat{\mathbf{W}} \mathbf{S} \\ \mathbf{S}^T \hat{\mathbf{W}} \mathbf{X}_{m_1} & \mathbf{S}^T \hat{\mathbf{W}} \mathbf{S} \end{pmatrix}^{-1} \\ &= \begin{pmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\beta}_{m_2}}^{(1,1)} & \boldsymbol{\Sigma}_{\boldsymbol{\beta}_{m_2}}^{(1,2)} \\ \boldsymbol{\Sigma}_{\boldsymbol{\beta}_{m_2}}^{(1,2)T} & \boldsymbol{\Sigma}_{\boldsymbol{\beta}_{m_2}}^{(2,2)} \end{pmatrix}, \end{aligned}$$

where

$$\begin{aligned} \boldsymbol{\Sigma}_{\boldsymbol{\beta}_{m_2}}^{(1,1)} &= \left(\mathbf{X}_{m_1}^T \hat{\mathbf{W}} \mathbf{X}_{m_1}\right)^{-1} \\ &\quad \times \left(\mathbf{I}_{p_{m_1}} + \mathbf{X}_{m_1}^T \hat{\mathbf{W}} \mathbf{S} \left(\mathbf{S}^T \hat{\mathbf{W}} \left(\mathbf{I}_n - \mathbf{P}_{m_1}\right) \mathbf{S}\right)^{-1} \mathbf{S}^T \hat{\mathbf{W}} \mathbf{X}_{m_1} \left(\mathbf{X}_{m_1}^T \hat{\mathbf{W}} \mathbf{X}_{m_1}\right)^{-1}\right), \\ \boldsymbol{\Sigma}_{\boldsymbol{\beta}_{m_2}}^{(1,2)} &= -\left(\mathbf{X}_{m_1}^T \hat{\mathbf{W}} \mathbf{X}_{m_1}\right)^{-1} \mathbf{X}_{m_1}^T \hat{\mathbf{W}} \mathbf{S} \left(\mathbf{S}^T \hat{\mathbf{W}} \left(\mathbf{I}_n - \mathbf{P}_{m_1}\right) \mathbf{S}\right)^{-1}, \\ \boldsymbol{\Sigma}_{\boldsymbol{\beta}_{m_2}}^{(2,2)} &= \left(\mathbf{S}^T \hat{\mathbf{W}} \left(\mathbf{I}_n - \mathbf{P}_{m_1}\right) \mathbf{S}\right)^{-1}, \end{aligned}$$

with $\mathbf{P}_{m_1} = \mathbf{X}_{m_1} \left(\mathbf{X}_{m_1}^T \hat{\mathbf{W}} \mathbf{X}_{m_1}\right)^{-1} \mathbf{X}_{m_1}^T \hat{\mathbf{W}}$. Let

$$\boldsymbol{\Sigma}_{\boldsymbol{\beta}_{m_1},\mathbf{v}} = \begin{pmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\beta}_{m_1}} & \boldsymbol{\Sigma}_{\boldsymbol{\beta}_{m_1},\mathbf{v}} \\ \boldsymbol{\Sigma}_{\boldsymbol{\beta}_{m_1},\mathbf{v}} & \boldsymbol{\Sigma}_{\mathbf{v}} \end{pmatrix},$$

then

$$\begin{aligned}
\boldsymbol{\Sigma}_{\boldsymbol{\beta}_{m_1}} &= \boldsymbol{\Sigma}^{(1,1)}_{\boldsymbol{\beta}_{m_2}} + \left(\mathbf{X}_{m_1}^T\hat{\mathbf{W}}\mathbf{X}_{m_1}\right)^{-1}\mathbf{X}_{m_1}^T\hat{\mathbf{W}}\mathbf{S}\boldsymbol{\Sigma}^{(1,2)T}_{\boldsymbol{\beta}_{m_1},\mathbf{v}} \\
&\quad + \boldsymbol{\Sigma}^{(1,2)}_{\boldsymbol{\beta}_{m_1},\mathbf{v}}\mathbf{S}^T\hat{\mathbf{W}}\mathbf{X}_{m_1}\left(\mathbf{X}_{m_1}^T\hat{\mathbf{W}}\mathbf{X}_{m_1}\right)^{-1} \\
&\quad + \left(\mathbf{X}_{m_1}^T\hat{\mathbf{W}}\mathbf{X}_{m_1}\right)^{-1}\mathbf{X}_{m_1}^T\hat{\mathbf{W}}\mathbf{S}\boldsymbol{\Sigma}^{(2,2)}_{\boldsymbol{\beta}_{m_1},\mathbf{v}}\mathbf{S}^T\hat{\mathbf{W}}\mathbf{X}_{m_1}\left(\mathbf{X}_{m_1}^T\hat{\mathbf{W}}\mathbf{X}_{m_1}\right)^{-1}, \\
&= \left(\mathbf{X}_{m_1}^T\hat{\mathbf{W}}\mathbf{X}_{m_1}\right)^{-1}, \\
\boldsymbol{\Sigma}_{\boldsymbol{\beta}_{m_1},\mathbf{v}} &= \boldsymbol{\Sigma}^{(1,2)}_{\boldsymbol{\beta}_{m_2}} + \left(\mathbf{X}_{m_1}^T\hat{\mathbf{W}}\mathbf{X}_{m_1}\right)^{-1}\mathbf{X}_{m_1}^T\hat{\mathbf{W}}\mathbf{S}\boldsymbol{\Sigma}^{(2,2)}_{\boldsymbol{\beta}_{m_2}}, \\
&= \mathbf{0}, \\
\boldsymbol{\Sigma}_{\mathbf{v}} &= \boldsymbol{\Sigma}^{(2,2)}_{\boldsymbol{\beta}_{m_2}}, \\
&= \left(\mathbf{S}^T\hat{\mathbf{W}}\left(\mathbf{I}_n - \mathbf{P}_{m_1}\right)\mathbf{S}\right)^{-1}.
\end{aligned}$$

Therefore, the induced distribution of $\mathbf{v}$ is independent of $\boldsymbol{\beta}_{m_1}$. All that remains, is to find the mean, $\boldsymbol{\mu}_{\mathbf{v}}$, of $\mathbf{v}$. It follows that

$$\begin{aligned}
\boldsymbol{\mu}_{\boldsymbol{\beta}_{m_1},\mathbf{v}} &= \left(\boldsymbol{\mu}_{\boldsymbol{\beta}_{m_1}}, \boldsymbol{\mu}_{\mathbf{v}}\right)^T, \\
&= \mathbf{M}\boldsymbol{\mu}_{\boldsymbol{\beta}_{m_2}}, \\
&= \begin{pmatrix} \mathbf{I} & \left(\mathbf{X}_{m_1}^T\hat{\mathbf{W}}\mathbf{X}_{m_1}\right)^{-1}\mathbf{X}_{m_1}^T\hat{\mathbf{W}}\mathbf{S} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}\begin{pmatrix} \boldsymbol{\Sigma}^{(1,1)}_{\boldsymbol{\beta}_{m_2}} & \boldsymbol{\Sigma}^{(1,2)}_{\boldsymbol{\beta}_{m_2}} \\ \boldsymbol{\Sigma}^{(1,2)T}_{\boldsymbol{\beta}_{m_2}} & \boldsymbol{\Sigma}^{(2,2)}_{\boldsymbol{\beta}_{m_2}} \end{pmatrix}\begin{pmatrix} \mathbf{X}_{m_1}^T \\ \mathbf{S}^T \end{pmatrix}\hat{\mathbf{W}}\hat{\boldsymbol{\eta}},
\end{aligned}$$

and that

$$\begin{aligned}
\boldsymbol{\mu}_{\mathbf{v}} &= \left(\boldsymbol{\Sigma}^{(1,2)T}_{\boldsymbol{\beta}_{m_2}}\mathbf{X}_{m_1}^T + \boldsymbol{\Sigma}^{(2,2)}_{\boldsymbol{\beta}_{m_2}}\mathbf{S}^T\right)\hat{\mathbf{W}}\hat{\boldsymbol{\eta}}, \\
&= \left(\mathbf{S}^T\hat{\mathbf{W}}\left(\mathbf{I}_n - \mathbf{P}_{m_1}\right)\mathbf{S}\right)^{-1}\mathbf{S}^T\left(\mathbf{I}_n - \hat{\mathbf{W}}\mathbf{X}_{m_1}\left(\mathbf{X}_{m_1}^T\hat{\mathbf{W}}\mathbf{X}_{m_1}\right)^{-1}\mathbf{X}_{m_1}^T\right)\hat{\mathbf{W}}\hat{\boldsymbol{\eta}}, \\
&= \left(\mathbf{S}^T\hat{\mathbf{W}}\left(\mathbf{I}_n - \mathbf{P}_{m_1}\right)\mathbf{S}\right)^{-1}\mathbf{S}^T\hat{\mathbf{W}}\left(\mathbf{I}_n - \mathbf{P}_{m_1}\right)\hat{\boldsymbol{\eta}}.
\end{aligned}$$

Suppose the current state of the algorithm is $(m, \boldsymbol{\beta}_m, \phi_m)$. In the algorithm, there will be positive probability of remaining in the same model. O'Hagan and Forster (2004, pg. 299) describe how any MCMC method can be used to update the parameters $\boldsymbol{\beta}_m$ and ϕ_m. The obvious choice is a Metropolis-Hastings algorithm step and, in particular, Gibbs sampling, an independence sampler or a random walk. All three methods have their advantages and disadvantages. We discuss this issue further when we consider within model moves for the reversible jump scheme for GLMMs in Section 5.4.

We are now in a position to write down the reversible jump algorithm.

1. Fit the most complicated model $m^* \in M$, to find $\hat{\boldsymbol{\beta}}_{m^*}$ and $\hat{\phi}_{m^*}$, either the maximum likelihood estimates or posterior modes of $\boldsymbol{\beta}_{m^*}$ and ϕ_{m^*}, respectively.

2. Suppose we are in model m with parameters $\boldsymbol{\beta}_m$ and ϕ_m.

3. Propose a move to a neighbouring model $k \in M$ with probability $\pi_{m,k}$.

4. If the proposal involves remaining in the same model, i.e. $k = m$, then update the model parameters $\boldsymbol{\beta}_m$ and ϕ_m using any convenient MCMC method. Return to 2.

5. If the proposal is a death move, then partition $\mathbf{X}_m\boldsymbol{\beta}_m = [\mathbf{X}_k|\mathbf{S}]\left(\boldsymbol{\beta}_m^{(1)}, \boldsymbol{\beta}_m^{(2)}\right)^T$. Set

$$\boldsymbol{\beta}_k = \boldsymbol{\beta}_m^{(1)} + \left(\mathbf{X}_k^T\hat{\mathbf{W}}\mathbf{X}_k\right)^{-1}\mathbf{X}_k^T\hat{\mathbf{W}}\mathbf{S}\boldsymbol{\beta}_m^{(2)},$$

and

$$\phi_k = \phi_m.$$

Accept this proposal with probability

$$\min\left[1, \frac{f_k(\mathbf{y}|\boldsymbol{\beta}_k, \phi_k)f_k(\boldsymbol{\beta}_k, \phi_k)f(k)\pi_{k,m}q(\boldsymbol{\beta}_m^{(2)})}{f_m(\mathbf{y}|\boldsymbol{\beta}_m, \phi_m)f_m(\boldsymbol{\beta}_m, \phi_m)f(m)\pi_{m,k}}\right],$$

where $q()$ is the pdf of

$$\mathrm{N}\left(\left(\mathbf{S}^T\hat{\mathbf{W}}(\mathbf{I}_n - \mathbf{P}_k)\mathbf{S}\right)^{-1}\mathbf{S}^T\hat{\mathbf{W}}(\mathbf{I}_n - \mathbf{P}_k)\hat{\boldsymbol{\eta}}, \left(\mathbf{S}^T\hat{\mathbf{W}}(\mathbf{I}_n - \mathbf{P}_k)\mathbf{S}\right)^{-1}\right).$$

Otherwise reject the move and remain at $(m, \boldsymbol{\beta}_m, \phi_m)$. Return to 2.

6. If the proposal is a birth move, then generate $\mathbf{v}$ from the distribution

$$\mathrm{N}\left(\left(\mathbf{S}^T\hat{\mathbf{W}}(\mathbf{I}_n - \mathbf{P}_m)\mathbf{S}\right)^{-1}\mathbf{S}^T\hat{\mathbf{W}}(\mathbf{I}_n - \mathbf{P}_m)\hat{\boldsymbol{\eta}}, \left(\mathbf{S}^T\hat{\mathbf{W}}(\mathbf{I}_n - \mathbf{P}_m)\mathbf{S}\right)^{-1}\right),$$

which has pdf $q()$. Set

$$\boldsymbol{\beta}_k = \begin{pmatrix} \mathbf{I}_{p_m} & -\left(\mathbf{X}_m^T\hat{\mathbf{W}}\mathbf{X}_m\right)^{-1}\mathbf{X}_m^T\hat{\mathbf{W}}\mathbf{S} \\ \mathbf{0} & \mathbf{I}_{p_k - p_m} \end{pmatrix}\begin{pmatrix} \boldsymbol{\beta}_m \\ \mathbf{v} \end{pmatrix},$$

and

$$\phi_m = \phi_k.$$

Accept this proposal with probability

$$\min\left[1, \frac{f_k(\mathbf{y}|\boldsymbol{\beta}_k, \phi_k)f_k(\boldsymbol{\beta}_k, \phi_k)f(k)\pi_{k,m}}{f_m(\mathbf{y}|\boldsymbol{\beta}_m, \phi_m)f_m(\boldsymbol{\beta}_m, \phi_m)f(m)\pi_{m,k}q(\mathbf{v})}\right],$$

otherwise reject the move and remain at $(m, \boldsymbol{\beta}_m, \phi_m)$. Return to 2.

5.4 Reversible Jump for GLMMs

In this Section we generalise the reversible jump scheme of Gill (2007) to use for model determination amongst GLMMs.

5.4.1 Preliminaries

As described in Section 5.1, the reversible jump scheme for GLMMs will operate over the integrated posterior distributions of each model, $m \in M$. The pdf of the integrated posterior distribution of $m \in M$ is

$$f_m(\boldsymbol{\beta}_m, \mathbf{D}_m, \phi_m|\mathbf{y}) \propto f_m(\mathbf{y}|\boldsymbol{\beta}_m, \mathbf{D}_m, \phi_m) f_m(\boldsymbol{\beta}_m, \mathbf{D}_m, \phi_m).$$

Note that $\boldsymbol{\beta}_m \in \mathbb{R}^{p_m}$, $\mathbf{D}_m \in \mathbb{P}^{q_m}$ and $\phi_m > 0$. Similar to Section 4.2, we want the model parameters to lie in $\mathbb{R}^{p_m+\frac{1}{2}q_m(q_m+1)+1}$. We use the same transformations outlined in Section 4.5, so that $\mathbf{D}_m = \boldsymbol{\Gamma}_m \boldsymbol{\Gamma}_m^T$ where

$$\boldsymbol{\Gamma}_m = \begin{pmatrix} e^{\nu_{m,11}} & & & \\ \nu_{m,12} & e^{\nu_{m,22}} & & \\ \vdots & & \ddots & \\ \nu_{m,1q_m} & \cdots & & e^{\nu_{m,q_mq_m}} \end{pmatrix},$$

for $\boldsymbol{\nu}_m = (\nu_{m,11}, \nu_{m,12}, ..., \nu_{m,1q_m}, \nu_{m,22}, ..., \nu_{m,2q_m}, ..., \nu_{q_mq_m})^T \in \mathbb{R}^{\frac{1}{2}q_m(q_m+1)}$ and $\omega_m = e^{\phi_m} \in \mathbb{R}$. The pdf of the transformed integrated posterior distribution is proportional to

$$\begin{aligned} g_m(\boldsymbol{\beta}_m, \boldsymbol{\nu}_m, \omega_m) &= f_m(\mathbf{y}|\boldsymbol{\beta}_m, \mathbf{D}_m = \boldsymbol{\Gamma}_m\boldsymbol{\Gamma}_m^T, \omega_m = e^{\phi_m}) \\ &\quad \times f_m(\boldsymbol{\beta}_m, \mathbf{D}_m = \boldsymbol{\Gamma}_m\boldsymbol{\Gamma}_m^T, \omega_m = e^{\phi_m}) 2^{q_m} e^{\omega_m} \prod_{k=1}^{q_m} e^{\nu_{m,kk}(q_m+2-k)}. \end{aligned}$$

We begin by making some definitions and an assumption. Let $M_{\mathcal{Z}} \subset M$ denote the subset of models with the same group-specific structure, namely those with group-specific covariates $\mathcal{Z}$. So, for example, $M_{\emptyset}$ denotes the set of GLMs since they have no group-specific parameters. Also, $M_{\mathbf{1}}$ denotes the set of GLMMs with group-specific intercepts and $M_{\mathbf{x}_1}$ denotes the set of GLMMs with group-specific intercept and a group-specific parameter for $\mathbf{x}_1$. Note that the $M_{\mathcal{Z}}$'s are disjoint, that

$$M = \cup_{\mathcal{Z}} M_{\mathcal{Z}},$$

and that the number of group-specific structures is equal to the number of GLMs plus one, i.e. the number of $M_{\mathcal{Z}}$'s is $|M_{\emptyset}| + 1$.

Define the *$M_{\mathcal{Z}}$-saturated model*, denoted by $m^*_{\mathcal{Z}} \in M_{\mathcal{Z}}$, as the most complicated model within $M_{\mathcal{Z}}$, i.e. it has regression parameters for all regression covariates.

Consider the running Turtle Dataset example first presented in Section 1.4. There are two possible GLMs, so there are three possible group-specific structures, so there exist $M_{\emptyset}$, $M_{\mathbf{1}}$ and $M_{\mathbf{x}}$. Models 1 and 2 lie in $M_{\emptyset}$, Models 3 and 4 lie in $M_{\mathbf{1}}$, and Model 5 lies in $M_{\mathbf{x}}$. For each of $M_{\emptyset}$, $M_{\mathbf{1}}$ and $M_{\mathbf{x}}$, the $M_{\mathcal{Z}}$-saturated models are Models 2, 4 and 5, respectively.

We assume that the marginal posterior distribution of the transformed variance components, $\boldsymbol{\nu}$, and the transformed dispersion parameter, ω, remains approximately the same for all models within each group-specific structure, or at least does so for models with non-negligible

posterior model probability. This assumption is similar to the assumption we made in Section 5.3.

The reversible jump scheme for GLMMs is based on three different types of move:

1. Within group-specific structure moves,
2. Across group-specific structure moves,
3. Within model moves.

For the within group-specific structure moves, we only add or remove a regression parameter and retain the same group-specific parameters. The models we consider moving between only differ by a single term in their regression covariates. We propose the parameters for the proposed model in an analogous way to Gill (2007). We describe our generalisation in Section 5.4.2.

For the across group-specific structure moves, opposite to the within group-specific structure moves, we only add or remove group-specific parameters and retain the same regression parameters. The models we consider moving between only differ by a single term in the group-specific covariates. To make these type of moves, we use an independence sampler and we describe how to form the proposal distribution in Section 5.4.3.

For the within model moves, we have several options as we discussed for the corresponding moves for GLMs. In Section 5.4.4, we discuss the advantages and disadvantages of some options and make our recommendations.

Define $\boldsymbol{\theta}_m = (\boldsymbol{\beta}_m, \boldsymbol{\nu}_m, \omega_m)^T \in \mathbb{R}^{p_m + \frac{1}{2}q_m(q_m+1)}$. When we describe the types of move in detail, it will become apparent that we need, for each of the $M_{\mathcal{Z}}$-saturated models, posterior modes of $\boldsymbol{\theta}_{m^*_{\mathcal{Z}}}$ which are denoted $\hat{\boldsymbol{\theta}}_{m^*_{\mathcal{Z}}} = (\hat{\boldsymbol{\beta}}_{m^*_{\mathcal{Z}}}, \hat{\boldsymbol{\nu}}_{m^*_{\mathcal{Z}}}, \hat{\omega}_{m^*_{\mathcal{Z}}})^T$. In addition, we also require, for each of the $M_{\mathcal{Z}}$-saturated models, an approximation to the Hessian matrix of $\log g_{m^*_{\mathcal{Z}}}(\boldsymbol{\theta}_{m^*_{\mathcal{Z}}})$ evaluated at $\hat{\boldsymbol{\theta}}_{m^*_{\mathcal{Z}}}$, i.e.

$$\left.\frac{\partial \log f_{m^*_{\mathcal{Z}}}(\boldsymbol{\theta}_{m^*_{\mathcal{Z}}}|\mathbf{y})}{\partial \boldsymbol{\theta}_{m^*_{\mathcal{Z}}} \partial \boldsymbol{\theta}^T_{m^*_{\mathcal{Z}}}}\right|_{\boldsymbol{\theta}_{m^*_{\mathcal{Z}}} = \hat{\boldsymbol{\theta}}_{m^*_{\mathcal{Z}}}},$$

and we define

$$\begin{aligned}
\hat{\boldsymbol{\Sigma}}_{m^*_{\mathcal{Z}}} &= -\left[\left.\frac{\partial \log f_{m^*_{\mathcal{Z}}}(\boldsymbol{\theta}_{m^*_{\mathcal{Z}}}|\mathbf{y})}{\partial \boldsymbol{\theta}_{m^*_{\mathcal{Z}}} \partial \boldsymbol{\theta}^T_{m^*_{\mathcal{Z}}}}\right|_{\boldsymbol{\theta}_{m^*_{\mathcal{Z}}} = \hat{\boldsymbol{\theta}}_{m^*_{\mathcal{Z}}}}\right]^{-1}, \\
&= \begin{pmatrix} \hat{\boldsymbol{\Sigma}}^{\boldsymbol{\beta}}_{m^*_{\mathcal{Z}}} & \hat{\boldsymbol{\Sigma}}^{\boldsymbol{\beta},\boldsymbol{\nu},\omega}_{m^*_{\mathcal{Z}}} \\ \left(\hat{\boldsymbol{\Sigma}}^{\boldsymbol{\beta},\boldsymbol{\nu},\omega}_{m^*_{\mathcal{Z}}}\right)^T & \hat{\boldsymbol{\Sigma}}^{\boldsymbol{\nu},\omega}_{m^*_{\mathcal{Z}}} \end{pmatrix},
\end{aligned} \tag{5.19}$$

where $\hat{\boldsymbol{\Sigma}}^{\boldsymbol{\beta}}_{m^*_{\mathcal{Z}}}$ is a $p_{m^*_{\mathcal{Z}}} \times p_{m^*_{\mathcal{Z}}}$ matrix, $\hat{\boldsymbol{\Sigma}}^{\boldsymbol{\nu},\omega}_{m^*_{\mathcal{Z}}}$ is a $\left(\frac{1}{2}q_{m^*_{\mathcal{Z}}}(q_{m^*_{\mathcal{Z}}}+1)+1\right) \times \left(\frac{1}{2}q_{m^*_{\mathcal{Z}}}(q_{m^*_{\mathcal{Z}}}+1)+1\right)$

matrix. Let $\hat{\mathbf{D}}^*_{m^*_{\mathcal{Z}}} = \mathbf{I}_G \otimes \hat{\mathbf{D}}_{m^*_{\mathcal{Z}}} = \mathbf{I}_G \otimes \hat{\mathbf{\Gamma}}_{m^*_{\mathcal{Z}}} \hat{\mathbf{\Gamma}}^T_{m^*_{\mathcal{Z}}}$, where

$$\hat{\mathbf{\Gamma}}_{m^*_{\mathcal{Z}}} = \left(\begin{array}{cccc} e^{\nu_{11}} & & & \\ \nu_{12} & e^{\nu_{22}} & & \\ \vdots & & \ddots & \\ \nu_{1q_{m^*_{\mathcal{Z}}}} & \cdots & & e^{\nu_{q_{m^*_{\mathcal{Z}}} q_{m^*_{\mathcal{Z}}}}} \end{array} \right) \Bigg|_{\boldsymbol{\nu}_{m^*_{\mathcal{Z}}} = \hat{\boldsymbol{\nu}}_{m^*_{\mathcal{Z}}}}.$$

Finally define $\hat{\boldsymbol{\eta}}_{m^*_{\mathcal{Z}}} = \mathbf{X}_{m^*_{\mathcal{Z}}} \hat{\boldsymbol{\beta}}_{m^*_{\mathcal{Z}}}$.

5.4.2 Within group-specific structure moves

Suppose we are considering a move between models $m_1 \in M_{\mathcal{Z}}$ and $m_2 \in M_{\mathcal{Z}}$. Again, we assume that $\mathbf{X}_{m_2} = [\mathbf{X}_{m_1} | \mathbf{S}]$, and that $p_{m_2} > p_{m_1}$. Since we are considering a within group-specific structure move, $q_{m_2} = q_{m_1}$. We only consider local moves so m_1 and m_2 differ only by a single term in the regression covariates.

Suppose that the current state of the MCMC chain is $(m_2, \boldsymbol{\beta}_{m_2}, \boldsymbol{\nu}_{m_2}, \omega_{m_2})$ and we are considering the death move to model m_1, so we need to propose values for $\boldsymbol{\beta}_{m_1}$, $\boldsymbol{\nu}_{m_1}$ and ω_{m_1}.

Let $\boldsymbol{\eta}_{m_2} = \mathbf{X}_{m_2} \boldsymbol{\beta}_{m_2}$ be the current linear predictor. Similar to the algorithm of Gill (2007), we set the proposed linear predictor, $\boldsymbol{\eta}_{m_1}$, to be the orthogonal (with respect to $\mathbf{W}$) projection of the current linear predictor onto the subspace defined by m_2, i.e. according to (5.14).

We need to consider what value to use for $\mathbf{W}$. Gill (2007) uses an approximation to the inverse posterior covariance matrix of the working vector. We use an analogous expression for the working vector, $\tilde{\mathbf{y}}$, of a GLMM with components

$$\tilde{y}_{ij} = \eta_{ij} + (y_{ij} - \mu_{ij}) g'(\mu_{ij}).$$

The working vector has expectation $\mathrm{E}(\tilde{\mathbf{Y}}) = \mathrm{E}(\mathrm{E}(\tilde{\mathbf{Y}}|\mathbf{u})) = \mathbf{X}\boldsymbol{\beta}$ and variance

$$\begin{aligned} \mathrm{var}(\tilde{\mathbf{Y}}) &= \mathbf{W}^{-1}, \\ &= \mathrm{E}(\mathrm{var}(\tilde{\mathbf{Y}}|\mathbf{u})) + \mathrm{var}(\mathrm{E}(\tilde{\mathbf{Y}}|\mathbf{u})), \\ &= \mathbf{V} + \mathbf{Z}\mathbf{D}^*\mathbf{Z}^T, \end{aligned}$$

where $\mathbf{V} = \mathrm{diag}\{\mathrm{var}(Y_{ij}) g'(\mu_{ij})^2\}$. Again, we do not want $\mathbf{W}$ to depend on the current model parameters so, similar to Gill (2007), we use the posterior modes of the $M_{\mathcal{Z}}$-saturated models. So we replace $\mathbf{W}$ by $\hat{\mathbf{W}}_{m^*_{\mathcal{Z}}}$ where

$$\hat{\mathbf{W}}_{m^*_{\mathcal{Z}}} = \left[\hat{\mathbf{V}}_{m^*_{\mathcal{Z}}} + \mathbf{Z}_{m^*_{\mathcal{Z}}} \hat{\mathbf{D}}^*_{m^*_{\mathcal{Z}}} \mathbf{Z}^T_{m^*_{\mathcal{Z}}} \right]^{-1},$$

and

$$\hat{\mathbf{V}}_{m^*_{\mathcal{Z}}} = \mathrm{diag}\{\mathrm{var}(Y_{ij}) g'(\mu_{ij})^2\} \big|_{\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}_{m^*_{\mathcal{Z}}}, \omega = \hat{\omega}_{m^*_{\mathcal{Z}}}}.$$

Note that $\mathbf{Z}_{m^*_{\mathcal{Z}}} = \mathbf{Z}_{m_1} = \mathbf{Z}_{m_2}$. If $M_{\mathcal{Z}} = M_{\emptyset}$, then $\hat{\mathbf{W}}_{m^*_{\mathcal{Z}}}$ is identical to the $\hat{\mathbf{W}}$ used in the algorithm of Gill (2007) since $\mathbf{Z}_{m^*_{\emptyset}} = \mathbf{0}$. The non-diagonal nature of $\hat{\mathbf{W}}_{m^*_{\mathcal{Z}}}$ accounts for the correlation between the responses y_{ij} for $j = 1, ..., n_i$, in proposing $\boldsymbol{\beta}_{m_1}$. Now

$$\boldsymbol{\beta}_{m_1} = \left(\mathbf{X}_{m_1}^T \hat{\mathbf{W}}_{m^*_{\mathcal{Z}}} \mathbf{X}_{m_1}\right)^{-1} \mathbf{X}_{m_1}^T \hat{\mathbf{W}}_{m^*_{\mathcal{Z}}} \mathbf{X}_{m_2} \boldsymbol{\beta}_{m_2}, \tag{5.20}$$

and we use the assumption that the posterior distributions of $\boldsymbol{\nu}$ and ω remains approximately the same and so use the identity transformation for $\boldsymbol{\nu}_{m_1}$ and ω_{m_1}, i.e.

$$\boldsymbol{\nu}_{m_1} = \boldsymbol{\nu}_{m_2}, \tag{5.21}$$

and

$$\omega_{m_1} = \omega_{m_2}. \tag{5.22}$$

Writing $\boldsymbol{\beta}_{m_2} = (\boldsymbol{\beta}^{(1)}_{m_2}, \boldsymbol{\beta}^{(2)}_{m_2})^T$, where $\boldsymbol{\beta}^{(1)}_{m_2}$ corresponds to $\mathbf{X}_{m_1}$ and $\boldsymbol{\beta}^{(2)}_{m_2}$ corresponds to $\mathbf{S}$, (5.20) can be rewritten

$$\boldsymbol{\beta}_{m_1} = \boldsymbol{\beta}^{(1)}_{m_2} + \left(\mathbf{X}_{m_1}^T \hat{\mathbf{W}}_{m^*_{\mathcal{Z}}} \mathbf{X}_{m_1}\right)^{-1} \mathbf{X}_{m_1}^T \hat{\mathbf{W}}_{m^*_{\mathcal{Z}}} \mathbf{S} \boldsymbol{\beta}^{(2)}_{m_2}. \tag{5.23}$$

For the reverse birth move from m_1 to m_2, (5.23) must be satisfied, and is if

$$\boldsymbol{\beta}^{(1)}_{m_2} = \boldsymbol{\beta}_{m_1} - \left(\mathbf{X}_{m_1}^T \hat{\mathbf{W}}_{m^*_{\mathcal{Z}}} \mathbf{X}_{m_1}\right)^{-1} \mathbf{X}_{m_1}^T \hat{\mathbf{W}}_{m^*_{\mathcal{Z}}} \mathbf{S} \boldsymbol{\beta}^{(2)}_{m_2}.$$

If we generate $\mathbf{v}$ from some distribution then (5.23), (5.21) and (5.22) are satisfied if

$$\begin{pmatrix} \boldsymbol{\beta}^{(1)}_{m_2} \\ \boldsymbol{\beta}^{(2)}_{m_2} \\ \boldsymbol{\nu}_{m_2} \\ \omega_{m_2} \end{pmatrix} = \begin{pmatrix} \mathbf{I}_{p_{m_1}} & -\left(\mathbf{X}_{m_1}^T \hat{\mathbf{W}}_{m^*_{\mathcal{Z}}} \mathbf{X}_{m_1}\right)^{-1} \mathbf{X}_{m_1}^T \hat{\mathbf{W}}_{m^*_{\mathcal{Z}}} \mathbf{S} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{p_{m_2} - p_{m_1}} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}_{\frac{1}{2} q_{m_1}(q_{m_1}+1)} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & 1 \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_{m_1} \\ \mathbf{v} \\ \boldsymbol{\nu}_{m_1} \\ \omega_{m_1} \end{pmatrix}.$$

The transformation is an upper triangular matrix with all diagonal elements equal to one, so the Jacobian is equal to one. Again, we are left the choice of distribution of $\mathbf{v}$ and analogous with Gill (2007) we choose the normal distribution for $\boldsymbol{\beta}_{m_2}$ with mean equal to an approximation to the maximum likelihood estimate of $\boldsymbol{\beta}_{m_2}$ and variance equal to an approximation to the variance matrix of the maximum likelihood estimate of $\boldsymbol{\beta}_{m_2}$, i.e.

$$\boldsymbol{\beta}_{m_2} \sim \mathrm{N}\left(\left(\mathbf{X}_{m_1}^T \hat{\mathbf{W}}_{m^*_{\mathcal{Z}}} \mathbf{X}_{m_1}\right)^{-1} \mathbf{X}_{m_1}^T \hat{\mathbf{W}}_{m^*_{\mathcal{Z}}} \hat{\boldsymbol{\eta}}_{m^*_{\mathcal{Z}}}, \left(\mathbf{X}_{m_1}^T \hat{\mathbf{W}}_{m^*_{\mathcal{Z}}} \mathbf{X}_{m_1}\right)^{-1}\right),$$

From this, as in Section 5.3, we induce the distribution of $\mathbf{v}$ and find it is normal with mean

$$\boldsymbol{\mu}_{\mathbf{v}} = \left(\mathbf{S}^T \hat{\mathbf{W}}_{m^*_{\mathcal{Z}}} (\mathbf{I}_n - \mathbf{P}_{m_1,\mathcal{Z}}) \mathbf{S}\right)^{-1} \mathbf{S}^T \hat{\mathbf{W}}_{m^*_{\mathcal{Z}}} (\mathbf{I}_n - \mathbf{P}_{m_1,\mathcal{Z}}) \hat{\boldsymbol{\eta}}_{m^*_{\mathcal{Z}}},$$

and variance matrix

$$\boldsymbol{\Sigma}_{\mathbf{v}} = \left(\mathbf{S}^T \hat{\mathbf{W}}_{m^*_{\mathcal{Z}}} (\mathbf{I}_n - \mathbf{P}_{m_1,\mathcal{Z}}) \mathbf{S}\right)^{-1},$$

where $\mathbf{P}_{m_1,\mathcal{Z}} = \mathbf{X}_{m_1} \left(\mathbf{X}_{m_1}^T \hat{\mathbf{W}}_{m^*_{\mathcal{Z}}} \mathbf{X}_{m_1}\right)^{-1} \mathbf{X}_{m_1}^T \hat{\mathbf{W}}_{m^*_{\mathcal{Z}}}$.

5.4.3 Across group-specific structure moves

Suppose we are interested in moving between models $m_1 \in M_{\mathcal{Z}_1}$ and $m_2 \in M_{\mathcal{Z}_2}$. Since we only consider local moves, $\mathbf{X}_{m_1} = \mathbf{X}_{m_2}$ and $\mathbf{Z}_{im_2} = (\mathbf{Z}_{im_1}|\mathbf{S})$ or $\mathbf{Z}_{im_1} = (\mathbf{Z}_{im_2}|\mathbf{S})$ for $i = 1, ..., G$.

As mentioned above, we use the independence sampler and now describe how to form the proposal distributions for a general $m_1 \in M_{\mathcal{Z}_1}$.

Consider the most complicated model, $m^*_{\mathcal{Z}_1}$, in $M_{\mathcal{Z}_1}$. We can approximate the posterior distribution $\boldsymbol{\theta}_{m^*_{\mathcal{Z}_1}}|\mathbf{y}$ by $\mathrm{N}\left(\hat{\boldsymbol{\theta}}_{m^*_{\mathcal{Z}_1}}, \hat{\boldsymbol{\Sigma}}_{m^*_{\mathcal{Z}_1}}\right)$. If $m_1 = m^*_{\mathcal{Z}_1}$ then we can just use this distribution as our proposal distribution. For all the remaining models in $M_{\mathcal{Z}_1}$, we can use $\hat{\boldsymbol{\theta}}_{m^*_{\mathcal{Z}_1}}$ and $\hat{\boldsymbol{\Sigma}}_{m^*_{\mathcal{Z}_1}}$ to form proposal distributions. We assumed that for all models $m_1 \in M_{\mathcal{Z}_1}$, the marginal posterior distribution of $\boldsymbol{\nu}_{m_1}$ and ω_{m_1} remains approximately the same, so the proposal distribution for $(\boldsymbol{\nu}_{m_1}, \omega_{m_1})^T$ is

$$\begin{pmatrix} \boldsymbol{\nu}_{m_1} \\ \omega_{m_1} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \hat{\boldsymbol{\nu}}_{m^*_{\mathcal{Z}_1}} \\ \hat{\omega}_{m^*_{\mathcal{Z}_1}} \end{pmatrix}, \hat{\boldsymbol{\Sigma}}^{\boldsymbol{\nu},\boldsymbol{\omega}}_{m^*_{\mathcal{Z}_1}}\right).$$

We make the proposal distribution of $\boldsymbol{\beta}_{m_1}$ independent of the distribution of $(\boldsymbol{\nu}_{m_1}, \omega_{m_1})^T$, and also normal with mean $\boldsymbol{\mu}_{m_1} = \left(\mathbf{X}^T_{m_1}\hat{\mathbf{W}}_{m^*_{\mathcal{Z}}}\mathbf{X}_{m_1}\right)^{-1}\mathbf{X}^T_{m_1}\hat{\mathbf{W}}_{m^*_{\mathcal{Z}}}\hat{\boldsymbol{\eta}}_{m^*_{\mathcal{Z}}}$ and variance $\boldsymbol{\Sigma}_{m_1} = \left(\mathbf{X}^T_{m_1}\hat{\mathbf{W}}_{m^*_{\mathcal{Z}}}\mathbf{X}_{m_1}\right)^{-1}$. Therefore, the complete proposal distribution for a move from $m_2 \in M_{\mathcal{Z}_2}$ to $m_1 \in M_{\mathcal{Z}_1}$ is

$$\begin{pmatrix} \boldsymbol{\beta}_{m_1} \\ \boldsymbol{\nu}_{m_1} \\ \omega_{m_1} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \boldsymbol{\mu}_{m_1} \\ \hat{\boldsymbol{\nu}}_{m^*_{\mathcal{Z}_1}} \\ \hat{\omega}_{m^*_{\mathcal{Z}_1}} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{m_1} & \mathbf{0} \\ \mathbf{0} & \hat{\boldsymbol{\Sigma}}^{\boldsymbol{\nu},\boldsymbol{\omega}}_{m^*_{\mathcal{Z}_1}} \end{pmatrix}\right).$$

These independence sampler moves could actually be used to move to any model in M. The reason we restrict the moves to be local moves to neighbouring models which only result in an addition or removal of a group-specific parameter, is to make the algorithm more efficient. Reversible jump moves that are a transformation of the current model parameters are viewed as being more efficient than an independence sampler since they use information from the current model parameters to propose new parameters. However, we can also increase the efficiency of the algorithm by assuming that a model with high posterior model probability will be neighboured by models, also with high posterior model probability. Therefore, by only proposing local moves to neighbouring models we increase the efficiency of the algorithm.

5.4.4 Within model moves

We can use any MCMC method to update the parameters within the model. The obvious choices are a scan of a Gibbs sampling algorithm, or a step of an independence sampler or a random walk algorithm. Suppose the current state of the MCMC chain is $(m, \boldsymbol{\beta}_m, \boldsymbol{\nu}_m, \omega_m)$,

where $m \in M_{\mathcal{Z}}$. We now describe how we can use each of the Metropolis-Hastings samplers mentioned above.

An independence sampler could use the same proposal distribution as a move across group-specific structures, i.e.

$$\begin{pmatrix} \boldsymbol{\beta}_m \\ \boldsymbol{\nu}_m \\ \omega_m \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \boldsymbol{\mu}_m \\ \hat{\boldsymbol{\nu}}_{m^*_{\mathcal{Z}}} \\ \hat{\omega}_{m^*_{\mathcal{Z}}} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_m & \mathbf{0} \\ \mathbf{0} & \hat{\boldsymbol{\Sigma}}^{\boldsymbol{\nu},\omega}_{m^*_{\mathcal{Z}}} \end{pmatrix} \right).$$

A random walk algorithm could have proposal $(\boldsymbol{\beta}_m, \boldsymbol{\nu}_m, \omega_m)^T + \boldsymbol{\epsilon}$ where

$$\boldsymbol{\epsilon} \sim \mathrm{N}\left(\mathbf{0}, k_m \begin{pmatrix} \boldsymbol{\Sigma}_m & \mathbf{0} \\ \mathbf{0} & \hat{\boldsymbol{\Sigma}}^{\boldsymbol{\nu},\omega}_{m^*_{\mathcal{Z}}} \end{pmatrix} \right).$$

where k_m is a tuning parameter chosen so that the acceptance rates are between 0.1 and 0.4, providing an algorithm which is close to optimal (see page 24). Gibbs sampling would update the elements of $(\boldsymbol{\beta}_m, \boldsymbol{\nu}_m, \omega_m)^T$ one at a time, or we could use block updates. Typically, there exists no conditional conjugacy so we will have to use adaptive rejection sampling or ARMS. Also there exist no conditional independencies. For these reasons, Gibbs sampling is the most computationally expensive of the three methods mentioned. The independence sampler and the random walk algorithm both only require one evaluation of $g_m(\boldsymbol{\theta}_m)$ per within model move. The random walk algorithm may require tuning to get it close to optimal, whereas the independence sampler does not require any such tuning. However, the random walk algorithm uses information from the current state thus making sampling more efficient. In practice, we found both methods were mobile and seemed to work well.

5.4.5 The Algorithm

We now present the reversible jump algorithm for GLMMs in its entirety.

1. For each $M_{\mathcal{Z}}$-saturated model, $m^*_{\mathcal{Z}}$, find the posterior modes, $\hat{\boldsymbol{\theta}}_{m^*_{\mathcal{Z}}} = (\hat{\boldsymbol{\beta}}_{m^*_{\mathcal{Z}}}, \hat{\boldsymbol{\nu}}_{m^*_{\mathcal{Z}}}, \hat{\omega}_{m^*_{\mathcal{Z}}})^T$, of $\boldsymbol{\theta}_{m^*_{\mathcal{Z}}} = (\boldsymbol{\beta}_{m^*_{\mathcal{Z}}}, \boldsymbol{\nu}_{m^*_{\mathcal{Z}}}, \omega_{m^*_{\mathcal{Z}}})^T$ by maximising $\log g_{m^*_{\mathcal{Z}}}(\boldsymbol{\theta}_{m^*_{\mathcal{Z}}})$, and, also, find an approximation to

$$\left. \frac{\partial^2 \log g_{m^*_{\mathcal{Z}}}(\boldsymbol{\theta}_{m^*_{\mathcal{Z}}})}{\partial \boldsymbol{\theta}_{m^*_{\mathcal{Z}}} \boldsymbol{\theta}^T_{m^*_{\mathcal{Z}}}} \right|_{\boldsymbol{\theta}_{m^*_{\mathcal{Z}}} = \hat{\boldsymbol{\theta}}_{m^*_{\mathcal{Z}}}},$$

to find

$$\hat{\boldsymbol{\Sigma}}_{m^*_{\mathcal{Z}}} = - \left[\left. \frac{\partial^2 \log g_{m^*_{\mathcal{Z}}}(\boldsymbol{\theta}_{m^*_{\mathcal{Z}}})}{\partial \boldsymbol{\theta}_{m^*_{\mathcal{Z}}} \boldsymbol{\theta}^T_{m^*_{\mathcal{Z}}}} \right|_{\boldsymbol{\theta}_{m^*_{\mathcal{Z}}} = \hat{\boldsymbol{\theta}}_{m^*_{\mathcal{Z}}}} \right]^{-1},$$

where $\hat{\boldsymbol{\Sigma}}_{m^*_{\mathcal{Z}}}$ is partitioned as in (5.19).

2. Suppose we are in model $m \in M_{\mathcal{Z}_m}$ with current parameters $\boldsymbol{\beta}_m$, $\boldsymbol{\nu}_m$ and ω_m.

3. Propose a model $k \in M$ with probability $\pi_{m,k}$.

4. If the proposal involves remaining in the same model, i.e. $k = m$, then update the model using any MCMC method. In Section 5.4.4, we recommended either a step of a random walk algorithm or an independence sampler.

5. If the proposal is an across group-specific structure move, so that $k \in M_{\mathcal{Z}_k} \neq M_{\mathcal{Z}_m}$. Generate proposal parameters $(\boldsymbol{\beta}_k, \boldsymbol{\nu}_k, \omega_k)^T$ from

$$\begin{pmatrix} \boldsymbol{\beta}_k \\ \boldsymbol{\nu}_k \\ \omega_k \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \boldsymbol{\mu}_k \\ \hat{\boldsymbol{\nu}}_{m^*_{\mathcal{Z}_k}} \\ \hat{\omega}_{m^*_{\mathcal{Z}_k}} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_k & \mathbf{0} \\ \mathbf{0} & \hat{\boldsymbol{\Sigma}}^{\boldsymbol{\nu},\omega}_{m^*_{\mathcal{Z}_k}} \end{pmatrix} \right),$$

where the pdf of this distribution is $q_{m,k}()$. Let $q_{k,m}()$ be the pdf of

$$\mathrm{N}\left(\begin{pmatrix} \boldsymbol{\mu}_m \\ \hat{\boldsymbol{\nu}}_{m^*_{\mathcal{Z}_m}} \\ \hat{\omega}_{m^*_{\mathcal{Z}_m}} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_m & \mathbf{0} \\ \mathbf{0} & \hat{\boldsymbol{\Sigma}}^{\boldsymbol{\nu},\omega}_{m^*_{\mathcal{Z}_m}} \end{pmatrix} \right).$$

Accept this proposal with probability

$$\min\left[1, \frac{g_k(\boldsymbol{\beta}_k, \boldsymbol{\nu}_k, \omega_k) f(k) q_{k,m}(\boldsymbol{\beta}_m, \boldsymbol{\nu}_m, \omega_m) \pi_{k,m}}{g_m(\boldsymbol{\beta}_m, \boldsymbol{\nu}_m, \omega_m) f(m) q_{m,k}(\boldsymbol{\beta}_k, \boldsymbol{\nu}_k, \omega_k) \pi_{m,k}}\right],$$

else reject the move and remain at $(m, \boldsymbol{\beta}_m, \boldsymbol{\nu}_m, \omega_m)$. Return to 2.

6. If the proposal is a within group-specific structure death move, so that $k \in M_{\mathcal{Z}_m}$, then partition $\mathbf{X}_m\boldsymbol{\beta}_m = [\mathbf{X}_k|\mathbf{S}](\boldsymbol{\beta}_m^{(1)}, \boldsymbol{\beta}_m^{(1)})^T$. Set

$$\begin{aligned} \boldsymbol{\beta}_k &= \boldsymbol{\beta}_m^{(1)} + (\mathbf{X}_k^T \hat{\mathbf{W}}_{m^*_{\mathcal{Z}_m}} \mathbf{X}_k)^{-1} \mathbf{X}_k^T \hat{\mathbf{W}}_{m^*_{\mathcal{Z}_m}} \mathbf{S} \boldsymbol{\beta}_m^{(2)}, \\ \boldsymbol{\nu}_k &= \boldsymbol{\nu}_m, \\ \omega_k &= \omega_m. \end{aligned}$$

Accept this proposal with probability

$$\min\left[1, \frac{g_k(\boldsymbol{\beta}_k, \boldsymbol{\nu}_k, \omega_k) f(k) \pi_{k,m} q(\boldsymbol{\beta}_m^{(2)})}{g_m(\boldsymbol{\beta}_m, \boldsymbol{\nu}_m, \omega_m) f(m) \pi_{m,k}}\right],$$

where $q()$ is the pdf of

$$\mathrm{N}\left(\left(\mathbf{S}^T \hat{\mathbf{W}}_{m^*_{\mathcal{Z}_m}} (\mathbf{I}_n - \mathbf{P}_{k,\mathcal{Z}_m}) \mathbf{S}\right)^{-1} \mathbf{S}^T \hat{\mathbf{W}}_{m^*_{\mathcal{Z}_m}} (\mathbf{I}_n - \mathbf{P}_{k,\mathcal{Z}_m}) \hat{\boldsymbol{\eta}}_{m^*_{\mathcal{Z}_m}}, \left(\mathbf{S}^T \hat{\mathbf{W}}_{m^*_{\mathcal{Z}_m}} (\mathbf{I}_n - \mathbf{P}_{k,\mathcal{Z}_m}) \mathbf{S}\right)^{-1} \right).$$

Else reject the move and remain at $(m, \boldsymbol{\beta}_m, \boldsymbol{\nu}_m, \omega_m)$. Return to 2.

7. If the proposal is a within group-specific structure birth move, so that $k \in M_{\mathcal{Z}_m}$, then generate $\mathbf{v}$ from the distribution

$$\mathrm{N}\left(\left(\mathbf{S}^T \hat{\mathbf{W}}_{m^*_{\mathcal{Z}_m}} (\mathbf{I}_n - \mathbf{P}_{m,\mathcal{Z}_m}) \mathbf{S}\right)^{-1} \mathbf{S}^T \hat{\mathbf{W}}_{m^*_{\mathcal{Z}_m}} (\mathbf{I}_n - \mathbf{P}_{m,\mathcal{Z}_m}) \hat{\boldsymbol{\eta}}_{m^*_{\mathcal{Z}_m}}, \left(\mathbf{S}^T \hat{\mathbf{W}}_{m^*_{\mathcal{Z}_m}} (\mathbf{I}_n - \mathbf{P}_{m,\mathcal{Z}_m}) \mathbf{S}\right)^{-1} \right),$$

which has pdf $\mathbf{v}$. Set

$$\begin{aligned}
\boldsymbol{\beta}_k &= \begin{pmatrix} \mathbf{I}_{p_m} & -\left(\mathbf{X}_m^T \hat{\mathbf{W}}_{m^*_{\mathcal{Z}m}} \mathbf{X}_m\right)^{-1} \mathbf{X}_m^T \hat{\mathbf{W}}_{m^*_{\mathcal{Z}m}} \mathbf{S} \\ \mathbf{0} & \mathbf{I}_{p_k - p + m} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_m \\ \mathbf{v} \end{pmatrix}, \\
\boldsymbol{\nu}_k &= \boldsymbol{\nu}_m, \\
\omega_k &= \omega_m.
\end{aligned}$$

Accept this proposal with probability

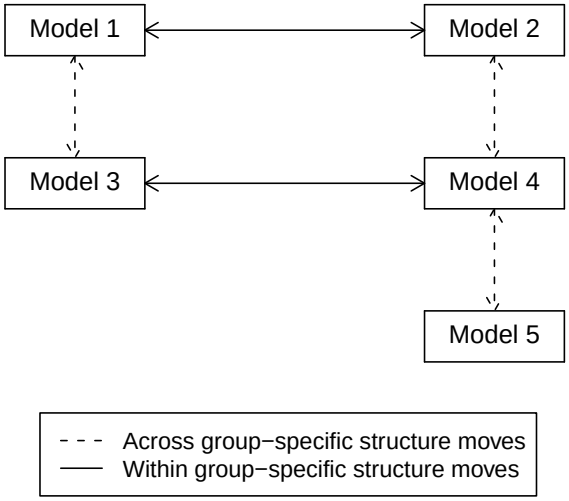
$$\min\left[1, \frac{g_k(\boldsymbol{\beta}_k, \boldsymbol{\nu}_k, \omega_k) f(k) \pi_{k,m}}{g_m(\boldsymbol{\beta}_m, \boldsymbol{\nu}_m, \omega_m) f(m) \pi_{m,k} q(\mathbf{v})}\right],$$

else reject the move and remain at $(m, \boldsymbol{\beta}_m, \boldsymbol{\nu}_m, \omega_m)$. Return to 2.

5.4.6 Turtle Data Example

We can apply the reversible jump scheme for GLMMs to the five models that are possible for the Turtle Dataset running example. We apply the unit information priors of Chapter 3 to the model parameters. Note that using the reversible jump algorithm for this dataset in practice would be unnecessary since the number of models is small enough to use the marginal likelihood approach with, for example, bridge sampling. Figure 5.2 shows the types of move possible amongst the five models.

Figure 5.2: The types of move possible for the Turtle Dataset.



We choose uniform proposal probabilities, $\pi_{m,k}$, that include the probability of remaining in the same model. Table 5.2 shows the proposal probabilities, $\pi_{m,k}$.

We need to find the posterior modes of the $M_{\mathcal{Z}}$-saturated models, i.e. of Model 2, Model 4 and Model 5. Table 5.3 shows the posterior modes of these three models.

Table 5.2: Proposal probabilities, $\pi_{m,k}$.

	m				
k	1	2	3	4	5
1	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0	0
2	$\frac{1}{3}$	$\frac{1}{3}$	0	$\frac{1}{4}$	0
3	$\frac{1}{3}$	0	$\frac{1}{3}$	$\frac{1}{4}$	0
4	0	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{4}$	$\frac{1}{2}$
5	0	0	0	$\frac{1}{4}$	$\frac{1}{2}$

Table 5.3: Posterior modes of the $M_{\mathcal{Z}}$-saturated models, to 4 decimal places.

Model 2	Model 4	Model 5
$\beta_1 = -0.3683$	$\beta_1 = -0.3758$	$\beta_1 = -0.4171$
$\beta_2 = 0.4084$	$\beta_2 = 0.4138$	$\beta_2 = 0.4451$
	$\nu_{11} = -0.7630$	$\nu_{11} = -0.7372$
		$\nu_{12} = -0.0038$
		$\nu_{22} = -0.6965$

We run the reversible jump algorithm for a total of 10000 iterations with a burn-in phase of 1000 iterations. Table 5.4 shows the posterior model probabilities as approximated by this algorithm. Also in Table 5.4, are the posterior model probabilities which are found using the marginal likelihoods that we approximated using importance sampling and were regarded as exact. The log of these approximated marginal likelihoods are displayed in Table 4.1. The posterior model probabilities as approximated by the reversible jump algorithm are very close to their exact values. This indicates that the Laplace approximation to the integrated approximation works well in this example. This may be due to the values of the variance components being small. Table 5.3 shows that the posterior modes of the transformed variance components for Models 4 and 5 are all quite small and we know that the Laplace approximation works well when the variance components are small.

Table 5.4: Approximated Posterior Probabilities of the Five Models from the Turtles Dataset.

Model, m	Posterior Model Probabilities, $f(m\|\mathbf{y})$	
	Reversible Jump	Marginal Likelihood Approach with importance sampling
1	0.0000	0.0001
2	0.3648	0.3484
3	0.0023	0.0013
4	0.1870	0.1871
5	0.4459	0.4632

5.5 Discussion

In this Chapter, we proposed a reversible jump algorithm for model determination amongst GLMMs by generalising a reversible jump algorithm for GLMs proposed by Gill (2007).

The reversible jump algorithm operated on the marginal posterior distribution of the parameters $\boldsymbol{\beta}_m$, $\mathbf{D}_m$ and ϕ_m by integrating out the group-specific parameters, $\mathbf{u}_m$. Using an MCMC model determination method over these marginal parameters was also used by Cai and Dunson (2006) incorporating an SSVS algorithm. The integrated likelihood is rarely analytically tractable so we needed to have a computationally inexpensive method for approximating it. We explored two competing methods: the Cai & Dunson method and the Laplace method. We found that both methods are more accurate when the variance components are small. Using an empirical assessment we found the Laplace method to be more accurate than the Cai & Dunson method. However, Cai and Dunson (2006) found the opposite, which indicated that the accuracy is example-dependent, which includes what prior is used.

We showed that the Cai & Dunson approximation to the integrated likelihood is a monotonic function of the elements of $\mathbf{D}$ and that for the resulting approximate posterior distribution to be proper, the prior mean of $\mathbf{D}$ must exist. Therefore, our chosen method for approximating the integrated likelihood was the Laplace method. However, the reversible jump algorithm described in this Chapter could be used with any diffuse priors. If a user wanted to specify another default prior for $\mathbf{D}$ such that the prior mean existed then the Cai & Dunson method could be used. The advantage of doing this is that the Cai & Dunson method is computationally less expensive than the Laplace method. In the Laplace method, we need to find the value, $\hat{\mathbf{u}}_i$, of $\mathbf{u}_i$ that maximises $f(\mathbf{y}_i|\boldsymbol{\beta}, \mathbf{u}_i, \phi)f(\mathbf{u}_i|\mathbf{D})$, for all $i = 1, ..., G$.

We described the algorithm of Gill (2007) for MCMC model determination amongst GLMs and then generalised it in Section 5.4 to GLMMs. To implement the algorithm we needed to maximise a number of approximate integrated posterior distributions. This is the main disadvantage of the method since it requires some prior computational expense. The number of integrated posterior distributions to maximise is equal to one more than the number of GLMs. The posterior modes and Hessian matrices evaluated at the posterior modes are used to form the proposal distributions for every model $m \in M$. The proposal distributions are used in an independence sampler for making across group-specific structure moves. Note that we could use an independence sampler to make all moves but, as noted in Chapter 2, this is an inefficient method and by incorporating moves that use information from the current state we increase the efficiency of sampling.

The reversible jump algorithm can be used with any diffuse prior distribution for the regression parameters. The reason it has to be diffuse is that we centre the proposal distribution for $\boldsymbol{\beta}_{m_2}$, for a within-group move, at an approximate maximum likelihood estimate of $\boldsymbol{\beta}_{m_2}$. We could modify this proposal distribution to account for prior information.

We applied the reversible jump algorithm to the Turtles Dataset and found that it approximated the posterior model probabilities very accurately. We concluded that this was due to the variance components being small, and therefore, the Laplace method performing well.

We will apply the reversible jump algorithm to various different examples in Chapter 6.

Recall from Section 5.1 that we are using a reversible jump algorithm to identify $M^* \subset M$ so that we need not use bridge sampling on all models $m \in M$, and can just use it on the more manageable number of models in M^*. We now to discuss how to identify M^*.

One possible approach is that we assume we only have computational resources to use bridge sampling to approximate the marginal likelihood of b models. In this case, we include in M^*, the models with the b highest approximate posterior model probabilities, $\hat{f}(m|\mathbf{y})$, as found via the reversible jump algorithm.

Another approach is to identify $\max_{m \in M} \hat{f}(m|\mathbf{y})$ and then include in M^*, all models that have approximate posterior model probability larger than a specified fraction of $\max_{m \in M} \hat{f}(m|\mathbf{y})$, i.e.

$$M^* = \left\{ m \in M | \hat{f}(m|\mathbf{y}) \geq c \max_{m \in M} \hat{f}(m|\mathbf{y}) \right\}, \tag{5.24}$$

where $0 < c < 1$. This definition is used in relation to selecting a set of models to model-average over by Madigan and Raftery (1994).

The disadvantage of the first definition is that there may only be a small number of models that have non-negligible posterior model probability and, thus, we include in M^* models of negligible posterior model probability. This would not happen using the second definition, although we may get an M^* containing more models than we can manage with the marginal likelihood approach.

We prefer this second approach, with a default value of $c = 10$. For the Turtle Dataset, we approximate the posterior model probabilities of the five models using the reversible jump and these are shown in Table 5.4. Note that Model 5 has the highest approximate posterior model probability of 0.4632. So with the value of $c = 10$, we include in M^*, all models with approximate posterior model probability larger than 0.0463, i.e. we include Models 2, 4 and 5 in M^*.

Chapter 6

Examples

In this Chapter, we use the default model determination strategy for three examples. We apply the default prior distributions of Chapter 3 to the model parameters, $\boldsymbol{\beta}$ and $\mathbf{D}$. In these examples, the dispersion parameter is known so we need not specify a prior distribution for ϕ. We identify M^* using the reversible jump scheme of Chapter 5. We then approximate the marginal likelihood of each of the models in M^* using bridge sampling as implemented in Chapter 4.

For these examples, we compare our model determination conclusions to those of other methods or authors, where relevant. When we report the values of BIC, they are found by using the function glmer in the R package lme4 (see Bates and Maechler (2009)).

6.1 Ship Incident Data

The *Ship Incident Data* can be found in McCullagh and Nelder (1989, pg. 205) 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. Therefore, there are five group-specific structures.

We apply the prior distributions proposed in Chapter 3 for $\boldsymbol{\beta}$ and $\mathbf{D}$, and run the reversible jump algorithm proposed in Chapter 5. The algorithm is run for a total of 10000 iterations after a burn-in phase of 1000 iterations, and identifies an M^* containing four models. These models have linear predictors:

10. $\eta_{ij} = \beta_1 + u_{1ij} + (\beta_2 + u_{2ij})x_{1ij} + \beta_3 x_{2ij} + \beta_4 x_{3ij} + \beta_5 x_{4ij}$; where $\mathbf{u}_i \overset{\text{iid}}{\sim} \text{N}(\mathbf{0}, \mathbf{D})$,
11. $\eta_{ij} = \beta_1 + u_{1ij} + (\beta_2 + u_{2i})x_{2ij} + (\beta_3 + u_{3i})x_{3ij} + (\beta_4 + u_{4i})x_{4ij}$; where $\mathbf{u}_i = (u_{1i}, u_{2i}, u_{3i}, u_{4i})^T \overset{\text{iid}}{\sim} \text{N}(\mathbf{0}, \mathbf{D})$,
12. $\eta_{ij} = \beta_1 + u_{1ij} + \beta_2 x_{1ij} + (\beta_3 + u_{3i})x_{2ij} + (\beta_4 + u_{4i})x_{3ij} + (\beta_5 + u_{5i})x_{4ij}$; where $\mathbf{u}_i = (u_{1i}, u_{3i}, u_{4i}, u_{5i})^T \overset{\text{iid}}{\sim} \text{N}(\mathbf{0}, \mathbf{D})$,
13. $\eta_{ij} = \beta_1 + u_{1ij} + (\beta_2 + u_{2i})x_{1ij} + (\beta_3 + u_{3i})x_{2ij} + (\beta_4 + u_{4i})x_{3ij} + (\beta_5 + u_{5i})x_{4ij}$; where $\mathbf{u}_i = (u_{1i}, u_{2i}, u_{3i}, u_{4i}, u_{5i})^T \overset{\text{iid}}{\sim} \text{N}(\mathbf{0}, \mathbf{D})$,

Table 6.1 shows the posterior model probabilities of the four models in M^*, as approximated by the reversible jump algorithm. These four models account for 97.65% of total posterior

Table 6.1: Approximated Posterior Probabilities (to 3 decimal places) of Models 10, 11, 12 and 13 from the Ship Incident Data, as approximated by the reversible jump algorithm.

Model	Posterior Model Probabilities	
m	$f(m\|\mathbf{y}, M)$	$f(m\|\mathbf{y}, M^*)$
10	0.058	0.059
11	0.182	0.186
12	0.231	0.237
13	0.506	0.518

Table 6.2: Approximated Log Marginal Likelihoods and Posterior Probabilities (to 3 decimal places) of Models 10, 11, 12 and 13 from the Ship Incident Data, as approximated by bridge sampling.

Model, m	Log Marginal Likelihood, $\log f_m(\mathbf{y})$	Posterior Model Probabilities, $f(m\|\mathbf{y})$
10	-125.581	0.042
11	-124.389	0.138
12	-123.974	0.209
13	-122.904	0.610

model probability. Table 6.1 also shows the posterior model probability, if we consider only models in M^*.

We now approximate the marginal likelihood of the four models in M^* using bridge sampling as described in Chapter 4. We use a total posterior sample size of 20000. Table 6.2 shows the log marginal likelihoods and resulting posterior model probabilities, as approximated by bridge sampling.

The model with the highest posterior model probability is actually the most complicated model possible, i.e. Model 13. This suggests that the effect that the classification factors have on the number of damage incidents suffered is different for the different types of ship. This means that, if we considered ship type to be an additional classification factor, and used a standard GLM there would exist an interaction between the ship type factor and the year of construction and year of operation factors. McCullagh and Nelder (1989) conducted such an analysis using classical statistical methods. They found inconclusive evidence for this interaction but stated that after fitting the interaction "the deviance reduced from 38.7 with 25 degrees of freedom to 14.6 with 10. This reduction would have some significance if the Poisson model were appropriate but, with over-dispersion present, the significance of the approximate F-ratio vanishes completely". The existence of over-dispersion is what makes a GLMM an appropriate model for this dataset.

6.2 Six Cities Data

The *Six Cities Data* can be found in Fitzmaurice and Laird (1993). It is frequently used to assess mixed models methodology. It concerns the wheezing status of 537 children over four years, and is sometimes referred to as the *Wheeze Data* in the literature.

Let y_{ij} denote the wheezing status (0=not wheezing, 1=wheezing) of the ith child at the jth time point, for $i = 1, ..., 537$ and $j = 1, ..., 4$. Also, let z_{1ij} and z_{2ij} denote the child's age (in years) and the mother's smoking status (0=non-smoker, 1=smoker) of child i at time point j, for $i = 1, ..., 537$ and $j = 1, ..., 4$. In this dataset, $z_{2ij} = z_{2ik}$, for all $j, k = 1, ..., 4$, i.e. the mother's smoking status does not change. Note that this simplification is in no way necessary for the following model determination approach. Also, $z_{1i1} = 7$, $z_{1i2} = 8$, $z_{1i3} = 9$, and $z_{1i4} = 10$, for all $i = 1, ..., 537$.

Let $x_{1ij} = \frac{z_{1ij} - \bar{z}_1}{\sqrt{\text{var}(z_{1ij})}}$ and $x_{2ij} = \frac{z_{2ij} - \bar{z}_2}{\sqrt{\text{var}(z_{2ij})}}$ denote the standardised versions of the age and smoking status variables. Let $x_{3ij} = x_{1ij}x_{2ij}$ be the interaction between the age and smoking status variables.

Suppose $y_{ij} \sim \text{Bernoulli}(p_{ij})$ where $\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \eta_{ij}$, i.e. we use the logit link function. There are a total of nineteen models.

We apply the unit information prior distributions for the regression parameters, $\boldsymbol{\beta}$ and the variance components matrix, $\mathbf{D}$, proposed in Chapter 3 and then use the reversible jump algorithm proposed in Chapter 5. The algorithm is run for a total of 20000 iterations after a burn-in phase of 1000 iterations. This algorithm identified M^* to contain six models, with the following linear predictors:

6. $\eta_{ij} = \beta_1 + u_{1i}$; where $u_{1i} \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2)$,
7. $\eta_{ij} = \beta_1 + \beta_2 x_{1ij} + u_{1i}$; where $u_{1i} \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2)$,
8. $\eta_{ij} = \beta_1 + \beta_2 x_{2ij} + u_{1i}$; where $u_{1i} \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2)$,
9. $\eta_{ij} = \beta_1 + \beta_2 x_{1ij} + \beta_3 x_{2ij} + u_{1i}$; where $u_{1i} \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2)$,
11. $\eta_{ij} = (\beta_1 + u_{1i}) + (\beta_2 + u_{2i})x_{1ij}$; where $\text{u} = (u_{1i}, u_{2i})^T \overset{\text{iid}}{\sim} \text{N}(\mathbf{0}, \mathbf{D})$,
15. $\eta_{ij} = (\beta_1 + u_{1i}) + \beta_2 x_{1ij} + (\beta_3 + u_{3i})x_{2ij}$; where $\text{u} = (u_{1i}, u_{3i})^T \overset{\text{iid}}{\sim} \text{N}(\mathbf{0}, \mathbf{D})$.

Table 6.3 shows the posterior model probabilities (to 3 decimal places) of the six models in M^*, as approximated by the reversible jump algorithm. The models in M^* account for 95.7% of total posterior model probability. Also shown in Table 6.3 is the approximate posterior model probabilities if we just consider models in M^*. These are found by

$$\hat{f}(m|\mathbf{y}, M^*) = \frac{\hat{f}(m|\mathbf{y}, M)}{\sum_{m \in M^*} \hat{f}(m|\mathbf{y}, M)}.$$

Table 6.3: Approximated Posterior Probabilities (to 3 decimal places) of the models in M^* from the Six Cities Data, as approximated by the reversible jump algorithm.

Model	Posterior Model Probabilities	
m	$f(m\|\mathbf{y}, M)$	$f(m\|\mathbf{y}, M^*)$
6	0.319	0.333
7	0.349	0.365
8	0.066	0.069
9	0.081	0.085
11	0.048	0.050
15	0.094	0.098

Table 6.4: Approximated Log Marginal Likelihoods and Posterior Probabilities, as approximated by bridge sampling, and BIC values of models in M^* from the Progabide Data (to 3 decimal places).

Model, m	Log Marginal Likelihood, $\log f_m(\mathbf{y})$	Posterior Model Probabilities, $f(m\|\mathbf{y})$	BIC_m
6	-808.317	0.347	1614.178
7	-808.200	0.390	1614.984
8	-809.923	0.070	1619.757
9	-809.885	0.072	1620.568
11	-810.294	0.048	1628.188
15	-809.872	0.073	1635.911

We now use bridge sampling to approximate the marginal likelihood of the six models in M^*. We use a posterior sample size of 50000. Table 6.4 shows the log marginal likelihood and the resulting posterior model probabilities of models in M^*, as approximated by bridge sampling.

Models 6 and 7 account for nearly 70% of the posterior model probability in M^*. By studying the linear predictors of these two models, it suggests that a child's wheezing status is different for each child and there is some evidence of an age effect on the wheezing status. We discussed the disadvantages of BIC for mixed models in Section 2.2.7. Nonetheless, in Table 6.4, we also give the values of BIC for the models in M^*. We did not compute the BIC values for all models in M but the BIC values for models in M^* seem to support our model determination conclusions, at least for the models with just group-specific intercepts. However, for models with not just group-specific intercepts, there is less correspondence between our conclusions and the BIC values.

Table 6.5: Approximated Posterior Probabilities (to 3 decimal places) of the models in M^* from the Progabide Data, as approximated by the reversible jump algorithm.

Model	Posterior Model Probabilities	
m	$f(m\|\mathbf{y}, M)$	$f(m\|\mathbf{y}, M^*)$
35	0.353	0.432
36	0.136	0.166
39	0.070	0.085
50	0.097	0.119
61	0.111	0.135
69	0.052	0.063

6.3 Progabide Data

The *Progabide Data* can be found in Thall and Vail (1990). It concerns four successive two-week seizure counts for 59 epileptics. Also recorded are the number of seizures suffered in the eight week period prior to the study, whether the patient received either the drug Progabide or a placebo, the age of the patient and the visit number. Let y_{ij} denote the number of seizures suffered by the ith patient in the two weeks prior to visit j, for $i = 1, ..., 59$ and $j = 1, ..., 4$. Likewise, denote z_{1ij}, z_{2ij}, z_{3ij} and z_{4ij} as the age, base-line seizure count, treatment (0=placebo,1=Progabide) and visit number of the ith patient at visit j, respectively, for $i = 1, ..., 59$ and $j = 1, ..., 4$. Note that $z_{4ij} = j$, and also, note that, $z_{1ij} = z_{1ik}$, $z_{2ij} = z_{2ik}$, $z_{3ij} = z_{3ik}$, for any $j, k = 1, ..., 4$.

Let $x_{kij} = \frac{x_{kij} - \bar{x}_k}{\sqrt{\text{var}(x_{kij})}}$ for $k = 1, ..., 4$, i.e. the x_{kij}'s are the standardised versions of the z_{kij}'s.

Suppose $y_{ij} \sim \text{Poisson}(\lambda_{ij})$, where $\log \lambda_{ij} = \eta_{ij}$. If we do not consider interactions, there are a total of 97 models, including 16 GLMs. Therefore, there are 17 group-specific structures.

We apply the unit information prior distributions proposed in Chapter 3 and run the reversible jump algorithm proposed in Chapter 5. We run the algorithm for a total of 20000 iterations after a burn-in phase of 1000 iterations. If we use the definition of M^* from (5.24), then we identify an M^* containing six models. The approximate posterior model probabilities of these six models are shown in Table 6.5. The models in M^* account for 79.5% of total posterior model probability. Also shown in Table 6.5 are the approximate posterior model probabilities if we just consider models in M^*. The linear predictors for the models in M^* are:

35. $\eta_{ij} = (\beta_1 + u_{1i}) + \beta_2 x_{2ij} + (\beta_3 + u_{3i}) x_{4ij}$; where $\mathbf{u}_i = (u_{1i}, u_{3i})^T \sim \text{N}(\mathbf{0}, \mathbf{D})$,

36. $\eta_{ij} = (\beta_1 + u_{1i}) + \beta_2 x_{2ij} + \beta_3 x_{3ij} + (\beta_4 + u_{4i}) x_{4ij}$; where $\mathbf{u}_i = (u_{1i}, u_{4i})^T \sim \text{N}(\mathbf{0}, \mathbf{D})$,

39. $\eta_{ij} = (\beta_1 + u_{1i}) + \beta_2 x_{1ij} + \beta_3 x_{2ij} + (\beta_4 + u_{4i}) x_{4ij}$; where $\mathbf{u}_i = (u_{1i}, u_{4i})^T \sim \text{N}(\mathbf{0}, \mathbf{D})$,

Table 6.6: Approximated Log Marginal Likelihoods and Posterior Probabilities, as approximated by bridge sampling, and BIC values of models in M^* from the Progabide Data (to 3 decimal places).

Model, m	Log Marginal Likelihood, $\log f_m(\mathbf{y})$	Posterior Model Probabilities, $f(m\|\mathbf{y})$	BIC_m
35	-679.715	0.454	588.462
36	-680.755	0.160	590.771
39	-681.400	0.084	592.035
50	-680.951	0.132	603.954
61	-680.990	0.127	602.018
69	-682.062	0.043	612.292

50. $\eta_{ij} = (\beta_1 + u_{1i}) + \beta_2 x_{2ij} + (\beta_3 + u_{3i})x_{3ij} + (\beta_4 + u_{4i})x_{4ij}$; where $\mathbf{u}_i = (u_{1i}, u_{3i}, u_{4i})^T \sim \text{N}(\mathbf{0}, \mathbf{D})$,

61. $\eta_{ij} = (\beta_1 + u_{1i}) + (\beta_2 + u_{2i})x_{2ij} + (\beta_3 + u_{3i})x_{4ij}$; where $\mathbf{u}_i = (u_{1i}, u_{2i}, u_{4i})^T \sim \text{N}(\mathbf{0}, \mathbf{D})$,

69. $\eta_{ij} = (\beta_1 + u_{1i}) + (\beta_2 + u_{2i})x_{2ij} + (\beta_3 + u_{3i})x_{3ij} + (\beta_4 + u_{4i})x_{4ij}$; where $\mathbf{u}_i = (u_{1i}, u_{2i}, u_{3i}, u_{4i})^T \sim \text{N}(\mathbf{0}, \mathbf{D})$.

We can now use bridge sampling to approximate the marginal likelihood of each of the models in M^*. We use bridge sampling with a total posterior sample size of 20000. Table 6.6 shows the log marginal likelihoods and the resulting posterior model probabilities as approximated by bridge sampling.

By studying the linear predictors of the six models in M^*, we see that they all contain regression parameters for the base-line seizure count and the visit number. They also all contain a group-specific parameter for the visit number. This suggests that the base-line seizure count and the visit number has an effect on the number of seizures suffered, and that the effect that the visit number has on the number of seizures suffered is different for each patient.

It would be impractical to compute the values of BIC for all 97 models in M^*. However, we give the values of BIC for the models in M^* in Table 6.6. Similar for the Six Cities Data, the BIC values seem to support our conclusions for the models with less group-specific parameters.

Chapter 7

Discussion

In this thesis, we developed an automatic, default strategy for Bayesian model determination amongst GLMMs. This strategy addressed the two key issues of default prior specification and computation.

In Chapter 3, we extended the idea of unit information prior distributions, which have been previously applied to the regression parameters of linear models and GLMs, to the regression parameters of GLMMs. We also developed a default prior distribution for the variance components matrix that relies on a unit information concept.

In Chapter 4, we investigated the method of bridge sampling for approximating the marginal likelihood on a GLMM. This marginal likelihood approach can only be applied, in a practical sense, when the number of models, $|M|$, is small, or when we have identified a subset of models, $M^* \subset M$, which have high posterior model probability and such that $|M^*|$ is small enough so that we can use bridge sampling. In Chapter 5, we proposed a reversible jump algorithm for identifying M^*.

An important note about the three parts of the strategy, i.e. the default prior distributions, the bridge sampling and the reversible jump algorithm, is that they are all stand-alone. This means that a user can use any number of the parts of the strategy. For instance, they could use the default prior distributions proposed in Chapter 3, but another method for model determination. A scenario where we may consider doing so is for normally distributed responses. In this case, we can evaluate the integrated likelihood exactly and the results of the reversible jump method will be, accordingly, more accurate. In this case, we may feel it is unnecessary to evaluate the marginal likelihood of each model in M^*.

The default prior distributions proposed in Chapter 3 can be applied to a rich set of models. Common proposed default priors for mixed models are typically restricted to certain scenarios, e.g. LMMs or for models that only have group-specific intercepts. One area where our default prior distributions are restricted is $\mathbf{D}^*$ must be $\mathbf{I}_G \otimes \mathbf{D}$, i.e. the group-specific parameters are exchangeable.

We did not develop a default prior distribution for the dispersion parameter. If this parameter is unknown, it will be contained in every model, so we can recommend using a diffuse inverse-gamma prior distribution. The disadvantage of this is we cannot undertake Bayesian model determination amongst different response distributions.

A possible criticism of the default prior distributions proposed in this thesis, and unit information prior distributions in general, is that they depend on the form of the experiment through the matrix $\mathbf{X}$, and therefore the matrix $\mathbf{Z}$. Our position is that all regression analyses are conditional on the covariates, so it is acceptable for the prior distribution to be dependent on the covariates and therefore the matrices $\mathbf{X}$ and $\mathbf{Z}$. This property of the prior distribution depending on the form of the experiment is not just possessed by unit information prior distributions. Jeffreys prior, for instance, also depends on the form of the experiment.

The reversible jump scheme proposed in Chapter 5 is an extension, to GLMMs, of the reversible jump scheme proposed by Gill (2007) for GLMs. This scheme can be applied when any default prior distribution has been proposed for the model parameters, i.e. does not need to be used in conjunction with the default priors proposed in Chapter 3. A disadvantage of this implementation of the reversible jump scheme is that we need to find the approximate posterior mode of the marginal posterior distribution by maximising $\hat{f}_{m^*_{\mathcal{Z}}}(\boldsymbol{\beta}_{m^*_{\mathcal{Z}}}, \boldsymbol{\nu}_{m^*_{\mathcal{Z}}}, \omega_{m^*_{\mathcal{Z}}}|\mathbf{y})$ for each $M_{\mathcal{Z}}$-saturated model, $m^*_{\mathcal{Z}}$. In addition, we also need to find the Hessian matrix of $\log \hat{f}_{m^*_{\mathcal{Z}}}(\boldsymbol{\beta}_{m^*_{\mathcal{Z}}}, \boldsymbol{\nu}_{m^*_{\mathcal{Z}}}, \omega_{m^*_{\mathcal{Z}}}|\mathbf{y})$ evaluated at the posterior mode. As the total number of explanatory variables available increases, the number of $M_{\mathcal{Z}}$-saturated models also increases. Therefore, our reversible jump scheme can only be applied to a moderately-sized dataset, where the measure of the size of a dataset is the number of available explanatory variables. An alternative approach is to only consider models with group-specific intercepts; this still gives us a rich set of models, but we only need to find the posterior mode and Hessian matrix for two models.

The reversible jump scheme is only as good as the Laplace approximation to the integrated likelihood. If the Laplace approximation is poor, our concern is that the reversible jump algorithm will mis-identify M^*. Hopefully, in this case, when we use bridge sampling to approximate the marginal likelihood for each of the models in M^* we will find that the approximate Bayes factors from the marginal likelihood approach will not correspond to the Bayes factors from the approximate posterior model probabilities from the reversible jump algorithm. In this case, we may have to use a computationally more intensive method for approximating the integrated likelihood, such as Gauss-Hermite quadrature (see Section 2.2.2). The 1-point Gauss-Hermite quadrature rule is equivalent to the Laplace approximation. Using a Gauss-Hermite quadrature rule with more points will result in a computationally more intensive method.

Bridge sampling has previously been used to approximate the marginal likelihood of GLMMs by Sinharay and Stern (2000) and Sinharay and Stern (2005). In both papers, the integrated likelihood is evaluated by using Simpson's rule. This is feasible since only group-specific intercept models are considered. Bridge sampling is then used to approximate the $p + \frac{1}{2}q(q+1) + 1$ dimensional integral. However, we use bridge sampling to approximate the marginal likelihood by approximating the $p+Gq+\frac{1}{2}q(q+1)+1$ dimensional integral, i.e. not evaluating the integrated likelihood. This allows us to consider much more complicated models.

Our treatment of bridge sampling is completely general and can be applied to any posterior distribution where it is relatively straightforward to generate a posterior sample.

There is a lot of scope for future work, in extending the work of this thesis.

It would be useful to construct a default prior distribution for the dispersion parameter so that we could make formal Bayesian model determination choices with respect to the response distribution. This would have to take account of the fact that, for some response distributions, the dispersion parameter is known. If would be convenient if the default prior distribution for an unknown dispersion parameter is an inverse-gamma distribution. Then this would be the conjugate prior distribution for the linear model.

In this thesis, to apply the unit information concept prior distribution for $\mathbf{D}$, we restricted ourselves to exchangeable group-specific parameters, i.e. $\mathbf{D}^* = \mathbf{I}_G \otimes \mathbf{D}$, where $\mathbf{D}$ is unstructured. Firstly, we could allow $\mathbf{D}$ to have some structure. For instance, we could allow $\mathbf{D}$ to have the following structure

$$\mathbf{D} = \sigma^2 \begin{pmatrix} 1 & \rho & \cdots & \rho^{q-1} \\ \rho & 1 & & \rho^{q-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{q-1} & \rho^{q-2} & \cdots & 1 \end{pmatrix}, \tag{7.1}$$

i.e. $\mathbf{D}_{jk} = \sigma^2 \rho^{|j-k|}$, for $j, k = 1, ..., q$. Therefore $\mathbf{D}$ depends on two unknown parameters, as opposed to $\frac{1}{2}q(q+1)$ unknown parameters when $\mathbf{D}$ is unstructured. The structure used for $\mathbf{D}$ in (7.1) is commonly used in longitudinal studies with equally-spaced observations where the correlation between two observations from the same group decreases as the time between those two observations increases.

Secondly, future work could also look at situations where $\mathbf{D}^*$ is not block-diagonal. For instance, at present $\text{cov}(\mathbf{u}_i, \mathbf{u}_j) = \mathbf{0}$, for $i \neq j$. However, we could make this covariance non-zero and, therefore, introduce correlations between the components of $\mathbf{u}$. A well-known example of a dataset where we might want to use a non block diagonal matrix for $\mathbf{D}^*$ is for the *Scotland Lip Cancer Data* which is found in, for example, Breslow and Clayton (1993). In this dataset, y_i is the number of cases of lip cancer observed in the ith Scottish district, for $i = 1, ..., 56$, from 1975-80. Let x_i and E_i denote the observed percentage of the workforce who work outside and the expected number of cases of lip cancer, respectively, for the ith district. Also recorded for the ith district is a set, $\mathcal{A}_i$, of geographically adjacent districts.

For $i = 1, ..., 56$, we assume $y_i \sim \text{Poisson}(E_i \lambda_i)$, where

$$\log \lambda_i = \beta_1 + \beta_2 x_i + u_i,$$

where u_i is scalar such that $\mathbf{u} = (u_1, ..., u_{56})^T \sim \text{N}(\mathbf{0}, \mathbf{D}^*)$. The approach in this thesis would be to make $\mathbf{D}^*$ a diagonal matrix depending on one parameter, σ^2. However, if we take this approach for this dataset, we are ignoring how the rates of lip cancer of geographically adjacent districts are likely to be correlated since they experience similar environmental conditions. This is known as *spatial correlation*. Sinharay and Stern (2005) consider this

example and use

$$\begin{aligned}\mathbf{D}^* &= \sigma^2\mathbf{E}^{-1} + \tau^2(\mathbf{I}_{56} - \phi\mathbf{C})^{-1}\mathbf{E}^{-1}, \\ &= \left\{\sigma^2\mathbf{I}_{56} + \tau^2(\mathbf{I}_{56} - \phi\mathbf{C})^{-1}\right\}\mathbf{E}^{-1},\end{aligned}$$

where $\mathbf{E}$ and $\mathbf{C}$ are known matrices such that $\mathbf{E} = \text{diag}(E_i)$ and $\mathbf{C}_{ij} = \left(\frac{E_j}{E_i}\right)^{\frac{1}{2}} I\left[j \in \mathcal{A}_i\right]$, for $i, j = 1, ..., 56$. Therefore $\mathbf{C}_{ij}$ is non-zero if districts i and j are geographically adjacent and it follows that $\mathbf{D}^*$ is non-diagonal. In this case, $\mathbf{D}^*$ depends on three parameters: $\sigma^2 > 0$, $\tau^2 > 0$ and $\phi \in (0, \phi_{\text{MAX}})$ where ϕ_{MAX} is a function of the E_i's determined so that $\mathbf{D}^*$ is positive-definite. The parameter ϕ controls the amount of spatial correlation. It would be of interest to determine whether there exists spatial correlation. This is equivalent to using Bayesian model determination amongst the two models: one having exchangeable group-specific parameters and the other having spatially correlated group-specific parameters as described above. As mentioned above, the bridge sampling approach described in Chapter 4 could be applied to this problem. We can apply the unit information prior distribution to the parameters β_1 and β_2. However, we would need to to define a default prior distribution for the parameters σ^2, τ^2 and ϕ. As a result, we are almost certain to lose the conditional conjugacy that we had with exchangeable group-specific parameters and an inverse-gamma prior distribution.

Most importantly, future work should focus on developing an MCMC model determination scheme that does not require the posterior modes of multiple models and can therefore be applied to a larger number of models, i.e. datasets with more available explanatory variables. An alternative, interesting route to reversible jump would be to attempt to apply the saturated space approach of Brooks et al. (2003). In this approach, the dimension of the Markov chain, conditional on model $m \in M$, is always $k_{\text{MAX}} = \sup_{m \in M}\{k_m\}$ and has elements $(\boldsymbol{\theta}_m, \mathbf{w}_m)^T$ where the dimension of the *auxiliary model parameters*, $\mathbf{w}_m$, is $k_{\text{MAX}} - k_m$.

Bibliography

Abramowitz, M. and Stegun, I. (eds.) (1965), *Handbook of Mathematical Functions*, Dover.

Bates, D. and Maechler, M. (2009), *lme4: Linear mixed-effects models using S4 classes*, r package version 0.999375-31.

Berger, J. and Bernardo, J. (1992), "Reference Priors in a Variance Components Problem," in *Proceedings of the Indo-USA workshop on Bayesian Analysis in Statistics and Econometrics*, ed. Goel, P., pp. 35–60.

Berger, J. and Pericchi, L. (1996), "The Intrinsic Bayes Factor for Model Selection and Prediction," *Journal of the American Statistical Association*, 91, 109–122.

Bernardo, J. (1979), "Reference Posterior Distributions for Bayesian Inference," *Journal of the Royal Statistical Society (Series B)*, 41, 113–147.

Box, G. and Tiao, G. (1992), *Bayesian Inference in Statistical Analysis*, Wiley.

Breslow, N. and Clayton, D. (1993), "Approximate Inference in Generalized Linear Mixed Models," *Journal of the American Statistical Association*, 88, 9–25.

Brooks, S., Giudici, P., and Roberts, G. (2003), "Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions," *Journal of the Royal Statistical Society (Series B)*, 65, 3–55.

Browne, W. and Draper, D. (2006), "A comparison of Bayesian and likelihood-based methods for fitting multilevel methods," *Bayesian Analysis*, 1, 473–514.

Burnham, K. and Anderson, D. (1998), *Model Selection and Inference: a practical information-theoretic approach*, Springer.

Cai, B. and Dunson, D. (2006), "Bayesian Covariance Selection in Generalized Linear Mixed Models," *Biometrics*, 62, 446–457.

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

Chen, M., Shao, Q., and Ibrahim, J. (2000), *Monte Carlo Methods in Bayesian Computation*, Springer.

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

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

Chopin, N. and Robert, C. (2009), "Contemplating Evidence: properties of, and alternatives to Nested Sampling," *Biometrika*, To Appear.

Congdon, P. (2003), *Applied Bayesian Modelling*, Wiley.

Daniels, M. (1999), "A prior for the variance in hierarchical models," *The Canadian Journal of Statistics*, 27, 567–578.

Dellaportas, P., Forster, J., and Ntzoufras, I. (2002), "On Bayesian model and variable selection using MCMC," *Statistics and Computing*, 12, 27–36.

— (2009), "Specification of prior distributions under model uncertainty," Tech. rep., University of Southampton.

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

Evans, M. (2007), "Comments on Nested Sampling by J. Skilling," in *Bayesian Statistics 8*, eds. Bernado, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., and West, M., Oxford, pp. 491–524.

Fisher, R. (1922), "On the Mathematical Foundations of Theoretical Statistics," *Philosophical Transactions of the Royal Society (Series A)*, 222, 309–368.

Fitzmaurice, G. and Laird, N. (1993), "A likelihood-based method for analysing longitudinal binary responses," *Biometrika*, 80, 141–151.

Fletcher, R. (2000), *Practical Methods of Optimization*, Wiley.

Garcia-Donato, G. and Sun, D. (2007), "Objective priors for hypothesis testing in one-way random effects models," *The Canadian Journal of Statistics*, 35, 303–320.

Gelfand, A. and Dey, D. (1994), "Bayesian Model Choice: Asymptotics and Exact Calculations," *Journal of the Royal Statistical Society (Series B)*, 56, 501–514.

Gelman, A., Roberts, G., and Gilks, W. (1996), "Efficient Metropolis Jumping Rules," in *Bayesian Statistics 5*, eds. Bernado, J., Berger, J., Dawid, A., and Smith, A., Oxford, pp. 599–607.

Gentle, J. (1998), *Random Number Generation and Monte Carlo Methods*, Springer.

George, E. and McCulloch, R. (1993), "Variable Selection Via Gibbs Sampling," *Journal of the American Statistical Association*, 88, 881–889.

Gilks, W. (1992), "Derivative-free adaptive rejection sampling for Gibbs sampling," in *Bayesian Statistics 4*, eds. Bernado, J., Berger, J., Dawid, A., and Smith, A., Oxford, pp. 641–649.

Gilks, W., Best, N., and Tan, K. (1995), "Adaptive Rejection Metropolis Sampling within Gibbs Sampling," *Applied Statistics*, 44, 455–472.

Gilks, W. and Wild, P. (1992), "Adaptive Rejection Sampling for Gibbs Sampling," *Applied Statistics*, 41, 337–348.

Gill, R. (2007), "Bayesian Inference for Partially Observed Data," Ph.D. thesis, University of Southampton.

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

Gustafson, P., Hossain, S., and MacNab, Y. (2006), "Conservative prior distributions for variance parameters in hierarchical models," *The Canadian Journal of Statistics*, 34, 377–390.

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

Henderson, H. and Searle, S. (1981), "On Deriving the Inverse of a Sum of Matrices," *SIAM Review*, 23, 53–60.

Ibrahim, J. and Laud, P. (1991), "On Bayesian Analysis of Generalized Linear Models Using Jeffreys Prior," *Journal of the American Statistical Association*, 86, 981–986.

Joe, H. (2008), "Accuracy of Laplace approximation for discrete response mixed models," *Computational Statistics and Data Analysis*, 52, 5066–5074.

Kass, R. and Natarajan, R. (2006), "A Default Conjugate Prior for Variance Components in Generalized Linear Mixed Models (Comment on Article by Browne and Draper)," *Bayesian Analysis*, 1, 535–542.

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

Kass, R. and Wasserman, L. (1995), "A Reference Bayesian Test for Nested Hypotheses and Its Relationship to the Schwarz Criterion," *Journal of the American Statistical Association*, 90, 928–934.

— (1996), "The Selection of Prior Distributions by Formal Rules," *Journal of the American Statistical Association*, 91, 1343–1370.

Lee, Y., Nelder, J., and Pawitan, Y. (2006), *Generalized linear models with random effects: unified analysis via h-likelihood*, Chapman and Hall.

Lunn, D., Thomas, A., Best, N., and Spiegelhalter, D. (2000), "WinBUGS-a Bayesian modelling framework: concepts, structure and extensibility," *Statistics and Computing*, 10, 325–337.

Madigan, D. and Raftery, A. (1994), "Model Selection and Accounting for Model Uncertainty in Graphical Models using Occam's Window," *Journal of the American Statistical Association*, 95, 227–237.

McCullagh, P. and Nelder, J. (1989), *Generalized Linear Models*, Chapman and Hall, 2nd ed.

McCulloch, C. and Searle, S. (2001), *Generalized, Linear, and Mixed Models*, Wiley.

Meng, X. and Schilling, S. (2002), "Warp Bridge Sampling," *Journal of Computational and Graphical Statistics*, 11, 552–586.

Meng, X. and Wong, W. (1996), "Simulating ratios of normalizing constants via a simple identity: a theoretical exploration," *Statistical Sinica*, 6, 831–860.

Mira, A. and Nicholls, G. (2004), "Bridge Estimation of the Probability Density at a Point," *Statistica Sinica*, 14, 603–612.

Muirhead, R. (1982), *Aspects of Multivariate Statistical Theory*, Wiley.

Natarajan, R. and Kass, R. (2000), "Reference Bayesian Methods for Generalized Linear Mixed Models," *Journal of the American Statistical Association*, 95, 227–237.

Neal, R. (1995), "Suppressing Random Walks in Markov Chain Monte Carlo Using Ordered Overrelaxation," Tech. rep., Department of Statistics, University of Toronto.

Nott, D. and Leonte, D. (2004), "Sampling Schemes for Bayesian Variable Selection in Generalized Linear Models," *Journal of Computational and Graphical Statistics*, 13, 362–382.

Ntzoufras, I., Dellaportas, P., and Forster, J. (2003), "Bayesian variable and link determination for generalised linear models," *Journal of Statistical Planning and Inference*, 111, 165–180.

O'Hagan, A. and Forster, J. (2004), *Kendall's Advanced Theory of Statistics*, vol. 2B Bayesian Inference, Arnold, 2nd ed.

Overstall, A. and Forster, J. (2009), "Default Bayesian Model Determination Methods for Generalised Linear Mixed Models," Tech. rep., School of Mathematics, University of Southampton.

Papathomas, M., Dellaportas, P., and Vasdekis, V. (2009), "A general proposal construction for reversible jump," Tech. rep., Department of Epidemiology and Public Health, Imperial College London.

Pauler, D. (1998), "The Schwarz Criterion and Related Methods for Normal Linear Models," *Biometrika*, 85, 13–27.

Pinheiro, J. and Chao, E. (2006), "Efficient Laplacian and adpative Gaussian quadrature algorithms for multilevel generalised linear mixed models," *Journal of Computational and Graphical Statistics*, 15, 58–81.

R Development Core Team (2009), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Robert, C. and Casella, G. (1999), *Monte Carlo Statistical Methods*, Springer.

Roberts, G. and Rosenthal, J. (2001), "Optimal Scaling for Various Metropolis-Hastings Algorithms," *Statistical Science*, 16, 361–367.

Rue, H., Martino, S., and Chopin, N. (2009), "Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations," *Journal of the Royal Statistical Society (Series B)*, 71, 319–392.

Sinharay, S. and Stern, H. (2000), "Bayes Factors for Variance Components Testing in Generalised Linear Mixed Models," in *Bayesian methods applied to science policy and official statistics*, ed. George, I., pp. 507–516.

— (2005), "An Empirical Comparison of Methods for Computing Bayes Factors in Generalised Linear Mixed Models," *Journal of Computational and Graphical Statistics*, 14, 415–435.

Skilling, J. (2006), "Nested Sampling for General Bayesian Computation," *Bayesian Analysis*, 1, 833–860.

Skrondal, A. and Rabe-Hesketh, S. (2004), *Generalized Latent Variable Modeling*, Chapman and Hall.

Smith, A. and Spiegelhalter, D. (1980), "Bayes Factors and Choice Criteria for Linear Models," *Journal of the Royal Statistical Society (Series B)*, 42, 213–220.

Spiegelhalter, D., Best, N., Carlin, B., and Van Der Linde, A. (2002), "Bayesian measures of model complexity and fit," *Journal of the Royal Statistical Society (Series B)*, 64, 583–639.

Sturtz, S., Ligges, U., and Gelman, A. (2005), "R2WinBUGS: A Package for Running WinBUGS from R," *Journal of Statistical Software*, 12, 1–16.

Thall, P. and Vail, S. (1990), "Some Covariance Models for Longitudinal Count Data with Overdispersion," *Biometrics*, 46, 657–671.

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

Zeger, S. and Karim, M. (1991), "Generalized Linear Models with Random Effects: A Gibbs Sampling Approach," *Journal of the American Statistical Association*, 86, 79–86.

Zhao, Y., Staudenmayer, J., Coull, B., and Wand, M. (2006), "General Design Bayesian Generalized Linear Mixed Models," *Statistical Science*, 21, 35–51.