2001

Bayes factors for variance component testing in generalized linear mixed models

Sandip Sinharay
Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/rtd

Part of the Statistics and Probability Commons

Recommended Citation
https://lib.dr.iastate.edu/rtd/1079
INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.
Bayes factors for variance component testing in generalized linear mixed models

by

Sandip Sinharay

A dissertation submitted to the graduate faculty in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Statistics

Major Professor: Dr. Hal Stern

Iowa State University

Ames, Iowa

2001

Copyright © Sandip Sinharay, 2001. All rights reserved.
This is to certify that the doctoral dissertation of

Sandip Sinharay

has met the dissertation requirements of Iowa State University

Signature was redacted for privacy.

Major Professor

Signature was redacted for privacy.

For the Major Program

Signature was redacted for privacy.

For the Graduate College
# TABLE OF CONTENTS

1 INTRODUCTION ................................................................. 1

2 COMPUTATIONAL TOOLS ................................................... 4

   2.1 Optimization ........................................................... 4
      2.1.1 Newton-Raphson method .................................. 4
      2.1.2 EM algorithm ................................................ 6

   2.2 Evaluation of integrals .............................................. 8
      2.2.1 Numerical quadrature ...................................... 8
      2.2.2 Laplace approximation ..................................... 9
      2.2.3 Monte Carlo integration .................................. 10
      2.2.4 Importance sampling ...................................... 10

   2.3 Bayesian analysis and Markov chain Monte Carlo methods ......... 11
      2.3.1 Gibbs sampling ............................................. 12
      2.3.2 Metropolis algorithm ..................................... 13

3 INFERENCE FOR GENERALIZED LINEAR MIXED MODELS ............... 16

   3.1 Generalized linear mixed models .................................. 16

   3.2 Approaches to estimation in generalized linear mixed models ... 18
      3.2.1 Likelihood-based approaches .............................. 18
      3.2.2 Bayesian approach ......................................... 19

   3.3 Testing hypotheses about variance components in generalized linear mixed models ............................................. 20
3.3.1 Frequentist approaches ........................................ 21
3.3.2 Bayesian approach ............................................. 23

4 COMPUTATION FOR GENERALIZED LINEAR MIXED MODELS 26

4.1 Computation of the likelihood .................................... 26
4.1.1 Numerical integration ......................................... 27
4.1.2 Laplace approximation ....................................... 27
4.1.3 Importance sampling ......................................... 28

4.2 Computation of posterior mode ................................... 29
4.2.1 Newton-Raphson algorithm .................................. 29
4.2.2 EM algorithm .................................................. 30

4.3 Approximating the marginal density $p(y|M)$ under a model ........................................ 34
4.3.1 Laplace approximation ....................................... 35
4.3.2 Importance sampling ......................................... 37
4.3.3 Harmonic estimator ......................................... 38
4.3.4 Bridge sampling ............................................. 39
4.3.5 Chib's method ............................................... 40

4.4 Approximating the Bayes factor .................................. 49
4.4.1 Approximating the marginal densities separately .......... 49
4.4.2 Verdinelli-Wasserman method ............................... 50
4.4.3 Reversible jump MCMC ...................................... 51

4.5 Sensitivity to the prior distribution of the parameters .......... 54

5 EXAMPLE: A NATURAL SELECTION STUDY 57

5.1 Description of the data set ....................................... 57
5.2 Estimation .......................................................... 59
5.2.1 Likelihood estimation ....................................... 61
5.2.2 Bayesian estimation ......................................... 61
5.2.3 Discussion about the results of estimation .................. 64
5.3 Hypothesis testing concerning $\sigma^2$ .............................. 65
  5.3.1 Likelihood ratio test ..................................... 65
  5.3.2 Lin's score test ........................................... 67
5.4 Computing estimates of the Bayes factor .......................... 67
  5.4.1 Specific details about the different methods .................. 69
  5.4.2 Summaries of the results obtained by the different methods ... 73
5.5 Results and sensitivity analysis .................................. 74
5.6 Simulation study .................................................. 79

6 EXAMPLE: SCOTLAND LIP CANCER DATA .............................. 84
  6.1 Description of the data set ..................................... 84
  6.2 A Poisson-Gaussian hierarchical model .......................... 85
  6.3 Estimation ..................................................... 90
    6.3.1 Maximum likelihood estimation ............................ 91
    6.3.2 Bayesian estimation ...................................... 97
  6.4 Testing hypotheses about the variance components ................. 101
  6.5 Approximating the Bayes factors ................................ 101
    6.5.1 Likelihood computation .................................. 104
    6.5.2 Specific details regarding the implementation of the various methods for this example .......................... 105
    6.5.3 Comparison of the different methods ....................... 109
    6.5.4 Interpretation of results ................................ 110
  6.6 Sensitivity to the prior distributions ........................... 111
  6.7 Simulation study ............................................... 118

7 CONCLUSIONS .......................................................... 122
  7.1 Inference ........................................................ 122
7.2 Choosing a method to compute the Bayes factor estimate \hspace{1em} 123
7.3 Prior distribution \hspace{1em} 125
7.4 Future work \hspace{1em} 125

BIBLIOGRAPHY \hspace{1em} 126

ACKNOWLEDGMENTS \hspace{1em} 130
LIST OF TABLES

Table 3.1 Interpretation of Bayes factors .......................... 24

Table 5.1 The turtle data summarized by the clutches ............. 60

Table 5.2 The marginal maximum likelihood estimates of the parameters of a probit regression model with a random intercept applied to the turtles data set ................................................ 61

Table 5.3 Summaries of the posterior distributions of the parameters of a probit regression model with a random intercept applied to the turtles data set ................................................ 64

Table 5.4 Estimates of the Bayes factor (along with their standard deviations and time taken to run the program) for comparing a simple probit regression model against a probit regression model with a random intercept obtained by different methods for the turtles data set ........................................... 69

Table 5.5 Estimates of the Bayes factor for comparing a simple probit regression model against a probit regression model with a random intercept for different degenerate priors on the variance component for the turtles data set ........................................... 76
Table 5.6 Estimates of the Bayes factor for comparing a simple probit regression model against a probit regression model with a random intercept for different prior means/modes (with an inverse gamma prior distribution) with prior variance fixed at 2 for the turtles data set ................................................. 78

Table 5.7 Estimates of the Bayes factor for comparing a simple probit regression model against a probit regression model with a random intercept for different prior means/modes (with an inverse gamma prior distribution) with prior variance fixed at 10 for the turtles data set ................................................. 80

Table 5.8 Estimates of the Bayes factors ($BF^{01}$) for comparing a simple probit regression model against a probit regression model with a random intercept for the simulated data sets similar to the turtles data set ................................................. 81

Table 6.1 The Scotland lip cancer data set ................................................. 86

Table 6.2 Summaries of the posterior distributions of the parameters of a Poisson-normal model with two sets of random effects applied to the Scotland lip cancer data set ................................................. 99

Table 6.3 Estimates of the Bayes factors (along with their standard deviations and time taken to run the program) for comparing different types of variance component models obtained by different methods for the Scotland lip cancer data set ................................................. 104

Table 6.4 Logarithm of estimated $BF^{01}_{\tau^2,\sigma^2}$'s for different $\tau^2$ and $\sigma^2$ for the Scotland lip cancer data set ................................................. 115

Table 6.5 Logarithm of estimated $BF^{01}_{p(.)^k}$'s for different prior means of $\tau^2$ and $\sigma^2$ for the turtles data set ................................................. 118
Table 6.6 Estimated Bayes factors ($BF^{01}$) for comparing different types of variance component models for the simulated data sets similar to the Scotland lip cancer data set

120
**LIST OF FIGURES**

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1</td>
<td>Scatter plot with the clutches sorted by average birth-weight</td>
<td>58</td>
</tr>
<tr>
<td>5.2</td>
<td>Histograms of the sampled values of the parameters</td>
<td>63</td>
</tr>
<tr>
<td>5.3</td>
<td>Histogram of the simulated LRT statistics under $H_0$</td>
<td>66</td>
</tr>
<tr>
<td>5.4</td>
<td>Plot of $BF_{\sigma^2}$ against $\sigma^2$</td>
<td>77</td>
</tr>
<tr>
<td>5.5</td>
<td>Plot of $BF_{p(\cdot)}$ against the prior means</td>
<td>79</td>
</tr>
<tr>
<td>5.6</td>
<td>Plot of the pdf for an inverse gamma(2.5,1.5) distribution</td>
<td>82</td>
</tr>
<tr>
<td>6.1</td>
<td>Histograms of the sampled values of the parameters</td>
<td>100</td>
</tr>
<tr>
<td>6.2</td>
<td>Plot of $\log(BF_{\tau^2,\sigma^2})$ for different $\tau^2$ and $\sigma^2$</td>
<td>113</td>
</tr>
<tr>
<td>6.3</td>
<td>Plot of $\log(BF_{\tau^2,\sigma^2})$ for different $\tau^2$ and $\sigma^2$</td>
<td>114</td>
</tr>
<tr>
<td>6.4</td>
<td>Plot of logarithm of $BF_{p(\cdot)}$ for different prior means of $\tau^2$ and $\sigma^2$</td>
<td>116</td>
</tr>
<tr>
<td>6.5</td>
<td>Plot of logarithm of $BF_{p(\cdot)}$ for different prior means of $\tau^2$ and $\sigma^2$</td>
<td>117</td>
</tr>
</tbody>
</table>
Bayes factors for variance component testing in generalized linear mixed models

Sandip Sinharay

Major Professor: Dr. Hal Stern
Iowa State University

Generalized linear models with random effects are becoming increasingly popular in situations where one needs to relate a non-normal response variable to a set of predictors and the responses are correlated. We start with a description of generalized linear models with random effects. Then we talk briefly about the frequentist and Bayesian approaches to inference for these models. In many applications, the magnitude of the variance components corresponding to one or more of the random effects are of interest, especially the point null hypothesis that one or more of the variance components is zero. A number of approaches are reviewed for approximating the Bayes factor comparing the models with and without the random effects in question. The computations involved with finding Bayes factors pose many challenges - especially for large problems and we discuss how one can overcome them.

We perform a comparative study of the different approaches to compute Bayes factors by applying them to two different data sets.

A common criticism of Bayes factors is that they are sensitive to the prior distributions used for the parameters of the models being compared. We develop an approach to study the sensitivity of the Bayes factor (comparing the models with and without the random effects in question) to the prior distributions used for the variance components and apply that to the two data sets to find out that the Bayes factor in question is indeed sensitive to the prior distributions used for the variance components.
1 INTRODUCTION

Generalized linear models with random effects, also known as generalized linear mixed models (GLMM), are used in situations where one needs to relate a non-normal response variable to a set of predictors and the responses are correlated. In many applications, the magnitude of the variance components corresponding to one or more of the random effects are of interest, especially the point null hypothesis that one or more of the variance components is zero. Because of the difficulty of computations with generalized linear mixed models, not much work has been done regarding the hypothesis testing problem mentioned above for these models. A Bayesian approach for testing a hypothesis is to compute the Bayes factor comparing the two competing models — one suggested by the null hypothesis and another by the alternative hypothesis. The objective of this work is to compare the different approaches for approximating the Bayes factor comparing the models with and without the random effects in question.

In Chapter 2, we introduce some important statistical ideas to be used in future chapters. Included among them are methods for optimizing a function and methods for approximating an integral. We also discuss the Bayesian approach to data analysis along with a key component of Bayesian analysis for complex problems, Markov chain Monte Carlo methods.

In Chapter 3, we first describe generalized linear mixed models and then talk briefly about the frequentist and Bayesian approaches to inference for generalized linear mixed models. A common question concerns null hypotheses that specify one or more variance components are zero. We review the literature on tests of this hypothesis in the
remaining part of Chapter 3.

In Chapter 4, we discuss the different issues about computation of the Bayes factor for variance component testing for generalized linear mixed models in details. We first describe the different ways to compute the marginal likelihood function for these models. Then we talk about computation of the posterior mode for them. Finally we discuss a number of approaches for approximating the Bayes factor that corresponds to the test of a null hypothesis that specifies one or more variance components are zero. Simple approximations (e.g., Laplace approximation, importance sampling etc.) to the Bayes factor and more sophisticated methods (e.g., Chib’s marginal likelihood method and Reversible jump Markov chain Monte Carlo method) are discussed. Advantages and disadvantages of the different methods are reviewed. A concern with the use of Bayes factors is that they are sensitive to the prior distribution used in the model. We develop an approach for studying the sensitivity of the Bayes factor estimate for variance component testing in generalized linear mixed models to the prior distribution for the variance parameters.

In Chapter 5, we describe an application of a simple generalized linear mixed model, a probit regression model with random effects, to the data set from a natural selection study. The Bayes factor of interest here is the one to determine if the probit model with the random effect terms provides a better fit than the ordinary probit regression model without any random effect terms. We compute the approximate Bayes factor using the different approaches discussed in Chapter 4 and compare them for this example. We use our approach discussed in Chapter 4 to study the sensitivity of the estimate of the Bayes factor of our interest to the prior distribution for the variance component. We also perform a simulation study to find out the behaviour of the Bayes factor estimate and the various computational approaches as the true variance component takes different values.

In Chapter 6, we take up a more complex example which involves the Scotland lip-
cancer data set (Clayton & Caldor, 1987). A Poisson-normal regression model with spatial random effects and heterogeneity random effects is fit to these data. A comparison of the different approaches for approximating the Bayes factors for comparing different variance component models is carried out as in the previous example. Because of the complexity of the model, computations become much more difficult — some of the methods applied in the previous example are too time-consuming to be applied here. We also carry out a sensitivity analysis using our approach discussed in Chapter 4 to study the sensitivity of the estimated Bayes factor to compare the model with no variance component and the model with both variance components to the prior distributions for the two variance components. Finally, we carry out a simulation study to figure out if the Bayes factors can help one to identity the right variance component model for this example.

Conclusions and directions for future work are described in Chapter 7.
2 COMPUTATIONAL TOOLS

A number of computational methods serve as building blocks for the approaches used in later chapters to compute Bayes factors. These methods are reviewed in this chapter.

2.1 Optimization

We will discuss a couple of numerical algorithms used to optimize functions: these will be useful later to find the posterior modes and the maximum likelihood estimates for the models of interest here.

2.1.1 Newton-Raphson method

Suppose we have a function $f(\omega)$ and want to find $\omega^*$, the mode of $f(\omega)$, i.e., to find $\omega^*$ that maximizes $f(\omega)$, over $\omega \in \Omega$. If the first and second derivatives of $f(\omega)$ can be easily found, then the mode can be computed rapidly by a process called the Newton-Raphson method.

The Newton-Raphson method is an iterative method that is based on a quadratic Taylor series approximation of $f(\omega)$. It starts from an initial value $\omega^{(0)}$ and generates a sequence of values through the steps:

$$
\omega^{(i)} = \omega^{(i-1)} - [f''(\omega^{(i-1)})]^{-1} f'(\omega^{(i-1)}),
$$

where $f'(\omega^{(i-1)})$ is the vector of first derivatives of $f(\omega)$ and $f''(\omega^{(i-1)})$ is the matrix of the second derivatives of $f(\omega)$, evaluated at $\omega^{(i-1)}$. When there is a negligible difference
in the values of \( \mathbf{w} \) obtained from two successive iterations and the first derivative becomes negligible, the most recent value of \( \mathbf{w} \) is taken as an approximation of the mode. In practice tolerances \( \gamma_1 \) and \( \gamma_2 \) (usually of the order of \( 10^{-6} \)) are defined and the Newton-Raphson iteration terminated when

\[
\left| \frac{\omega^{(i)} - \omega^{(i-1)}}{\omega^{(i-1)}} \right| < \gamma_1 \quad \text{and} \quad |f'(\omega^{(i)})| < \gamma_2
\]

The derivatives required in the algorithm may be found analytically or numerically. If the first and second derivatives of a function \( f(\omega) \) are difficult to determine analytically, one can approximate them using finite differences. The \( i \)-th component of \( f' \) can be approximated at any specified ordinate \( \omega = (\omega_1, \omega_2, \ldots, \omega_d) \) by

\[
f'_i(\omega) = \frac{\partial L}{\partial \omega_i} \approx \frac{f(\omega + \delta_i \epsilon_i) - f(\omega - \delta_i \epsilon_i)}{2\delta_i},
\]

where \( \delta_i \) is a small value and using linear algebraic notations, \( \epsilon_i \) is the unit vector corresponding to the \( i \)-th component of \( \omega \). The values of \( \delta_i \) should be chosen based on the scale of the problem; usually, values such as 0.0001 give good results (Gelman et al., 1995) since they are low enough to approximate the derivative and high enough to avoid roundoff error on the computer. The second derivative matrix at \( \omega \) is approximated by applying finite differencing again; for each \( i, j \),

\[
f''_{ij}(\omega) = \frac{\partial^2 L}{\partial \omega_i \partial \omega_j} = \frac{\partial}{\partial \omega_j} \left( \frac{\partial L}{\partial \omega_i} \right) \approx \frac{f_i'(\omega + \delta_j \epsilon_j) - f_i'(\omega - \delta_j \epsilon_j)}{2\delta_j} \\
\approx \frac{f_i'(\omega + \delta_i \epsilon_i + \delta_j \epsilon_j) - f_i'(\omega - \delta_i \epsilon_i + \delta_j \epsilon_j) - f_i'(\omega - \delta_i \epsilon_i) + f_i'(\omega + \delta_i \epsilon_i - \delta_j \epsilon_j)}{4\delta_i \delta_j}.
\]

The advantage of the Newton-Raphson method is that convergence is extremely fast if the initial value is close to the solution. However, the disadvantage of this method is that the starting value is important, especially for high-dimensional parameter space.
The algorithm is not guaranteed to converge from all starting values, particularly in regions where $-f''(\omega)$ is neither positive definite nor negative definite — in those cases, the iterations may move off towards the boundary of the parameter space. One can use a new starting point in that case or try several iterations of a more robust optimization algorithm like steepest descent.

The Newton-Raphson method converges to a local maximum — so it is customary to start the iterations at a number of different points $\omega$ and then to compare the values of the function $f(\omega)$ at all of the local maxima to obtain the global maximum.

\subsection*{2.1.2 EM algorithm}

The EM algorithm (Dempster et. al., 1977) is commonly used to obtain maximum likelihood estimates or posterior modes in problems with missing data. In fact, the algorithm is applicable in a broad range of situations where probability models can be reexpressed as ones on augmented parameter spaces where the added parameters can be thought of as missing data. The EM algorithm is relevant for mixed models because the random effects may be viewed as missing data and the algorithm may be used to find maximum likelihood estimates for these models.

Suppose that given an unknown parameter vector $\omega$,

$$y \sim f(y; \omega)$$

and that

$$y' = (y'_\text{obs}, y'_\text{mis}),$$

where $y_{\text{obs}}$ is the observed data and $y_{\text{mis}}$ is the missing data. Suppose the objective is to obtain the maximum likelihood estimate of $\omega$, i.e., the value of $\omega$ that maximizes the observed data likelihood

$$f(y_{\text{obs}} | \omega) = \int f(y_{\text{obs}}, y_{\text{mis}} | \omega) dy_{\text{mis}}.$$
The EM algorithm goes through a succession of E-steps and M-steps starting from an initial value $\omega^{(0)}$ of the parameter vector $\omega$. In the $(t+1)$-th E-step, one maximizes

$$Q(\omega|\omega^{(t)}) = E[\log\{f(y_{\text{obs}}, y_{\text{mis}}|\omega)\}|y_{\text{obs}}, \omega^{(t)}],$$

the expected value of the complete data loglikelihood, where the expectation is with respect to the distribution of $y_{\text{mis}}$ given $y_{\text{obs}}$ and $\omega^{(t)}$. In the corresponding M-step, one maximizes $Q(\omega|\omega^{(t)})$ computed in the preceding E-step with respect to $\omega$ to get $\omega^{(t+1)}$. The algorithm is said to have converged when two successive iterations give very similar values of the parameter vector. It is also a good idea to monitor the value of the loglikelihood $f(y_{\text{obs}}|\omega)$. The EM algorithm is guaranteed to converge to a local maximum — so it is customary to start the iterations at many points in the parameter space and then to compare the values of the likelihood at all of the local maxima. This method may be very slow to converge if the proportion of missing data is high.

The EM algorithm can be used to find the posterior mode of a model with missing data as well. Under the same kind of model assumptions as above, one has to maximize the observed data posterior

$$f(\omega|y_{\text{obs}}) \propto f(y_{\text{obs}}|\omega)p(\omega) = \int f(y_{\text{obs}}, y_{\text{mis}}|\omega)p(\omega)dy_{\text{mis}}$$

with respect to $\omega$ to get the posterior mode. It can be shown that (see e.g., Gelman et al, 1995, page 277) the $(t+1)$-th E-step in the EM algorithm to find the posterior mode requires maximizing

$$\hat{Q}(\omega|\omega^{(t)}) = E[\log\{f(y_{\text{obs}}, y_{\text{mis}}|\omega)p(\omega)\}|y_{\text{obs}}, \omega^{(t)}],$$

the expected value of the logarithm of the product of the complete data likelihood and the prior distribution, or, equivalently,

$$Q^*(\omega|\omega^{(t)}) = E[\log\{f(y_{\text{mis}}, \omega|y_{\text{obs}})\}|y_{\text{obs}}, \omega^{(t)}],$$
the expected value of the logarithm of the joint posterior density of the missing data and
the parameters conditional on the observed data, where the expectation is taken with
respect to the conditional distribution of $y_{mis}$ given $y_{obs}$ and $\omega^{(t)}$. In the corresponding
M-step, one maximizes $Q^*(\omega|\omega^{(t)})$ or $Q(\omega|\omega^{(t)})$ computed in the proceeding E-step with
respect to $\omega$ to get $\omega^{(t+1)}$.

For simple problems, one may be able to do the averaging in (2.1) or (2.2) analytically.
However, in many practical problems (e.g., for the generalized linear mixed models), the
expectation cannot be computed analytically; then one may need to use an approxima-
tion technique or a simulation technique to compute the expectation. Examples can be
found later in Sections 4.2.2.1 and 4.2.2.2.

2.2 Evaluation of integrals

Approximation of integrals will be used throughout this work, e.g., to evaluate the
likelihood in a mixed model. We discuss some of the most popular methods for tackling
the problem.

2.2.1 Numerical quadrature

Numerical quadrature refers to the evaluation of definite integrals by numerical meth-
ods. Simpson’s rule is probably the most useful of all the formulas for numerical quadra-
ture (Scarborough, 1930, page 119). Suppose one has to compute the definite integral
\[ \int_a^b f(x) dx. \]
First, the interval $(a, b)$ is divided into an even number of equisized subin-
tervals using the points $a = x_0, x_i = x_0 + i * h, i = 1, 2, \ldots, 2n - 1; x_{2n} = x_0 + 2nh = b,$
where $h = \frac{b-a}{2n}$.

Simpson’s rule, which is derived by assuming that $f(x)$ is a piece-wise quadratic
function, one segment each in each of the subintervals $(x_0 + 2ih, x_0 + 2(i + 1)h), i =$
0,1,... n - 1, yields the formula
\[
\int_a^b f(x) \, dx = \int_{x_0}^{x_0 + 2nh} f(x) \, dx \\
\approx \frac{h}{3} \left[ f(x_0) + 4\{f(x_1) + f(x_3) \ldots + f(x_{2n-1})\} + 2\{f(x_2) + f(x_4) \ldots + f(x_{2n-2})\} + f(x_{2n}) \right]
\]

There are two nice features of Simpson’s rule. First, though derived from a quadratic approximation, it is in fact exact for piece-wise cubic polynomials. Second, it can easily be computed for a series of decreasing equisized intervals (see, e.g., Press et. al.).

### 2.2.2 Laplace approximation

Suppose we want to approximate the integral \( \int g(\omega) \, d\omega \), where \( g(.) \) is a real-valued function and \( \omega \) is a vector of variables. For \( h(\omega) = \log\{g(\omega)\} \), we can write
\[
\int g(\omega) \, d\omega = \int e^{h(\omega)} \, d\omega.
\]
Suppose further that \( \omega^* \) is the unique mode of \( h(\omega) \), i.e., \( \omega^* \) maximizes \( h(\omega) \) over all choices of \( \omega \). Then applying the Taylor series expansion of \( h(\omega) \) around \( \omega^* \) and assuming that the 3rd and higher order derivatives of \( h(\omega) \) at \( \omega^* \) are negligible (which is equivalent to assuming that \( g(\omega) \) is proportional to a normal density), one gets the approximation (De Bruijn, 1961; Tierney and Kadane, 1986):
\[
\int g(\omega) \, d\omega \approx (2\pi)^{d/2} |\Psi|^{-\frac{1}{2}} g(\omega^*),
\]
where \( d \) is the dimension of \( \omega \) and \( \Psi \) is the negative of the inverse of the 2nd derivative of \( h(\omega) \) computed at \( \omega = \omega^* \).

Because the Laplace approximation is based on a normal approximation, it is most effective for unimodal and smooth \( h(\omega) \) and may give bad results for skewed or multimodal \( h(\omega) \) or for high-dimensional \( \omega \). It will also give bad results if \( \omega^* \) is near the edge of a finite parameter space, a common problem with the generalized linear mixed models.
2.2.3 Monte Carlo integration

Suppose we want to compute the expected value of a function \(g(\omega)\) where the expectation is taken with respect to a probability density \(f(\omega)\). If we can draw a random sample \(\omega_1, \omega_2, \ldots, \omega_n\) from \(f(\omega)\), the technique of Monte Carlo integration amounts to approximating the expectation required as:

\[
E_f\{g(\omega)\} = \int g(\omega)f(\omega)d\omega \approx \frac{1}{n} \sum_{i=1}^{n} g(\omega_i).
\]

The accuracy of the estimate can be assessed using the standard deviation of the \(g(\omega_i)\) values. If it is not easy to draw a random sample from \(f(\omega)\) or if the \(g(\omega_i)\) values are too variable (making the average of them too variable an approximation to be useful), one should use better approximations.

2.2.4 Importance sampling

Suppose, exactly as in the case of Monte Carlo integration, that we want to compute the expected value of a function \(g(\omega)\) where the expectation is taken with respect to a probability density \(f(\omega)\). We can write the expectation of \(g(\omega)\) as

\[
E_f\{g(\omega)\} = \int g(\omega)f(\omega)d\omega = \int g(\omega)f(\omega)h(\omega)d\omega
\]

for any probability density \(h(\omega)\). Now, if we can generate a random sample \(\tilde{\omega}_1, \tilde{\omega}_2, \ldots, \tilde{\omega}_n\) from the distribution with density \(h(\omega)\), we can approximate \(E_f\{g(\omega)\}\), using (2.1), as

\[
E_f\{g(\omega)\} \approx \frac{1}{n} \sum_{i=1}^{n} \frac{g(\tilde{\omega}_i)f(\tilde{\omega}_i)}{h(\tilde{\omega}_i)}.
\]

If \(h(\omega)\) is chosen so that \(\frac{g(\omega)f(\omega)}{h(\omega)}\) is roughly constant over the range of possible values of \(\omega\), fairly accurate approximation of the integral may be obtained. Here again, the standard deviation of the \(\frac{g(\tilde{\omega}_i)f(\tilde{\omega}_i)}{h(\tilde{\omega}_i)}\) values determines the quality of the approximation. In particular, \(h(\omega)\) should have heavier tails than the product \(g(\omega)f(\omega)\) since otherwise
there may be a few $\omega_i$'s for which $h(\omega)$ will be much smaller than $g(\omega)f(\omega)$ and the approximation will blow up.

### 2.3 Bayesian analysis and Markov chain Monte Carlo methods

In a typical Bayesian analysis, one draws inference about the unknown quantities, typically parameters of the probability model used to describe the data, from the posterior distribution of those quantities given the observed data. Suppose $p(y|\omega)$ is the distribution of the data vector $y$ conditional on the model parameter vector $\omega$. Suppose also that a prior distribution $p(\omega)$ is specified on the parameter vector. The posterior distribution of the parameter vector given the data is obtained using Bayes' rule as:

$$p(\omega|y) = \frac{p(y|\omega)p(\omega)}{p(y)} \propto p(y|\omega)p(\omega),$$

where

$$p(y) = \int p(y|\omega)p(\omega)d\omega.$$  

In practice, the posterior distribution is often impossible to derive analytically and hence it is not possible to make posterior inference about $\omega$ analytically. One way to generate inferences from $p(\omega|y)$ is to rely on numerical integration (as described in Section 2.2). It is common nowadays to rely instead on simulation-based approximations to find estimates of quantities of interest for the model. Markov chain Monte Carlo (MCMC) simulation is often used to obtain a random sample from the posterior distribution and then the sample is used as a discrete approximation of the posterior distribution. The Markov chain refers to the idea of creating a Markov process with stationary distribution equal to the posterior distribution of interest and then running the simulation long enough so that the distribution of the draws beyond some point is close enough to the stationary distribution. Then posterior expectations of relevant functions of $\omega$ are ap-
proximated using Monte Carlo integration. While we discuss some MCMC algorithms briefly in the next subsection, a number of books, e.g., Gelman et. al. (1995) and Gilks et. al. (1996) give a more detailed discussion of MCMC methods.

2.3.1 Gibbs sampling

The Gibbs sampler (Geman and Geman, 1984) is one of the most popular MCMC algorithms used to sample from posterior distributions and is defined in terms of subvectors of \( \omega \). Suppose that \( \omega \) can be partitioned into \( p \) components or subvectors as \( \omega = (\omega_1, \omega_2, \ldots, \omega_p) \). Each iteration in the Gibbs sampling algorithm goes through the subvectors of \( \omega \), drawing each subset conditional on the value of all the others and the data. The algorithm starts from an initial value \((\omega_0^0, \omega_2^0, \ldots, \omega_p^0)\) of \( \omega \). Each iteration \( t \) consists of \( p \) steps. Notation doesn't address this. In the \( j \)-th step of the \( t \)-th iteration, \( \omega_j^t \) is sampled from the conditional distribution of \( \omega_j \) given the most recently generated values of the other components of \( \omega \) and the data \( y \):

\[
p(\omega_j^t | \omega_{-j}^{t-1}, y),
\]

where \( \omega_{-j}^{t-1} \) contains all the components of \( \omega \), except for \( \omega_j \), at their current values:

\[
\omega_{-j}^{t-1} = (\omega_1^t, \omega_2^t, \ldots, \omega_{j-1}^t, \omega_{j+1}^{t-1}, \ldots, \omega_p^{t-1}).
\]

This description assumes that the subvectors are sampled in the same order for each iteration. In practice one can vary the ordering of the subvectors.

Gibbs sampling is repeated for \( t = 1, 2, \ldots, M \), where \( M \) is the number of iterations of the algorithm. Determining the value of \( M \) such that the algorithm reaches approximate convergence to the stationary distribution is a difficult problem. Cowles and Carlin (1996) provide a review of a number of convergence diagnostics. For this dissertation, we use the convergence diagnostic introduced by Gelman and Rubin (1992), which requires running several independent chains with different starting values and then compares the
location and variability of the different chains. Note also that simulations of the Markov chain are not independent and this issue must be addressed to estimate the accuracy of MCMC inferences.

Gibbs sampling is useful for sampling from a complex posterior distribution when the conditional posterior distributions can be easily sampled.

2.3.2 Metropolis algorithm

Another useful MCMC algorithm is the Metropolis algorithm (Metropolis and Ulam, 1949). This method is very useful when it is difficult to sample from a distribution directly, e.g., in the situation when one of the conditional distributions in a Gibbs sampler is difficult to sample from directly.

We begin with a description of the basic Metropolis algorithm. Let \( \omega^0 \) denote the initial value of \( \omega \). The \( t \)-th iteration of the Metropolis algorithm consists of the following steps:

- Sample a candidate point \( \omega^* \) from a jumping distribution \( J_t(\omega^*|\omega^{t-1}) \), which must be symmetric, i.e., \( J_t(\omega_a|\omega_b) = J_t(\omega_b|\omega_a) \) for all \( \omega_a, \omega_b \) and \( t \).

- Compute the acceptance probability

  \[
  r = \min \left( \frac{p(\omega^*|y)}{p(\omega^{t-1}|y)}, 1 \right). \tag{2.4}
  \]

- Set

  \[
  \omega^t = \begin{cases} 
  \omega^* & \text{with probability } \, r \\
  \omega^{t-1} & \text{with probability } \, (1 - r)
  \end{cases}
  \]

The Metropolis algorithm is run for \( M \) iterations, for \( t = 1, 2, \ldots, M \) and must be monitored for approximate convergence to the stationary distribution. Note that a candidate point with higher posterior density than the current value will always be
accepted. Candidate points with posterior density lower than the current value may or may not be accepted.

The choice of the jumping distribution $J_t(\omega^*|\omega^{t-1})$ is an important issue in Metropolis algorithm. A common jumping distribution is a normal distribution with mean equal to $\omega^{t-1}$ and variance matrix chosen so that the algorithm has an acceptance rate between 0.2-0.45. This is known as a random walk Metropolis algorithm. More details about the Metropolis algorithm, jumping distributions and acceptance rates can be found in Gelman et. al. (1995).

Looking at the steps of the algorithm, we see that since the computation of the acceptance probability $r$ requires the computation of the ratio of the posterior density at two parameter values, it is enough to know the posterior distribution up to a normalizing constant to sample from it using Metropolis algorithm. This characteristic makes the Metropolis algorithm very useful in Bayesian analysis since in many complicated problems, the posterior density is known only up to a normalizing constant.

The Metropolis-Hastings algorithm (Hastings, 1970) is a generalization of the Metropolis algorithm that allows one to use a jumping distribution $J_t(\omega_a|\omega_b)$ that is not symmetric. To allow for the asymmetry of the jumping rule, the ratio $r$ in (2.2) is replaced by

$$r = \min \left( \frac{p(\omega^*|y)}{p(\omega^{t-1}|y)} \cdot \frac{J(\omega^*|\omega^{t-1})}{J(\omega^{t-1}|\omega^*)}, 1 \right).$$

(2.5)

Allowing asymmetric jumping rules can increase the speed of convergence of the Markov chain.

Sometimes, it is convenient to incorporate the Metropolis (or Metropolis-Hastings) algorithm into a Gibbs sampling algorithm, using Metropolis steps to sample from some or all of the conditional posterior distributions. This is usually implemented when some or all of the subvectors of $\omega$ have conditional densities which are difficult to sample from.
using a Gibbs step (e.g., when the conditional density is known only up to a constant).
3 Inference for Generalized Linear Mixed Models

3.1 Generalized linear mixed models

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

Let $y_i$ denote the response for observation $i$, $i = 1, 2, \ldots, n$. The $y_i$'s are modeled as independent, given canonical parameter $\xi_i$ and scale parameter $\phi$, with probability density function

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

We drop $\phi$ from further consideration (i.e., take $\phi=1$). The two examples we consider in detail do not have any scale parameter. Also, all of the methods described here can be modified to accommodate a scale parameter in almost all practical situations. Let $\mu_i = E(y_i|\xi_i) = a'(\xi_i)$. The mean $\mu_i$ (and hence $\xi_i$) is expressed as a function of a $p \times 1$ predictor vector $x_i$, a $p \times 1$ vector of coefficients $\alpha$ and a $q \times 1$ random effects vector $b$ through the link function

$$g(\mu_i) = x_i'\alpha + z_i'b.$$
where \( z_t \) is a \( q \times 1 \) (typically zero/one) vector associated with the random effects. To complete the model specification, the random effects vector \( b \) is usually taken as multivariate normal with mean 0 and a positive definite variance matrix \( D(\theta) \), where \( \theta \) is an \( m \times 1 \) vector of unknown variance components. The magnitude of \( \theta \) determines the degree of overdispersion and correlation among the responses. Typically, the model is parameterized such that \( D(\theta) = 0 \) iff \( \theta = 0 \).

The likelihood function \( L(\alpha, \theta|y) \), also called the marginal likelihood function, is obtained by integrating out the random effects from the conditional density of the response using the density assigned to the random effects (also called the prior density of the random effects):}

\[
L(\alpha, \theta|y) = \int_b \left\{ \prod_{i=1}^n f(y_i|\xi_i) \right\} f(b|\theta) db
\]

\[
\propto |D(\theta)|^{-1/2} \int_b \left\{ \prod_{i=1}^n f(y_i|\alpha, b) \right\} \exp \left( -\frac{b' D^{-1}(\theta) b}{2} \right) db. \tag{3.2}
\]

The integral cannot be done analytically except for normal linear models, which makes any kind of data analysis with generalized linear mixed models complicated.

Another quantity of interest for generalized linear mixed models is the quasilikelihood function \( \tilde{L}(\alpha, \theta|y) \), which is defined as

\[
\tilde{L}(\alpha, \theta|y) = \int_b \exp \left\{ \sum_{i=1}^n l_i(\alpha|y_i, b) \right\} f(b|\theta) db, \tag{3.3}
\]

where, for \( E(y_i|b) \equiv \mu_i(b) \) and \( V(y_i|b) \equiv \frac{1}{\alpha_i} \sigma^2(\mu_i(b)) \),

\[
l_i(\alpha|y_i, b) \propto \int_{y_i}^{\mu_i(b)} \frac{a_i(y_i - u)}{v(u)} du
\]
defines the conditional log-quasilikelihood of \( \alpha \) given \( b \) for the \( i \)-th individual (Lin, 1997). Note that if \( v(u) \) is constant, then \( l_i(\cdot) \) is a quadratic function; thus the quasilikelihood and the likelihood agree for the normal density. The quasilikelihood is motivated by the observation that most of the asymptotic theory for maximum likelihood estimation also holds when the quasilikelihood replaces the likelihood. The quasilikelihood contains
only the conditional mean and variance of the responses (rather than making any distributional assumptions). It is useful in situations when we do not wish to make explicit assumptions about the distributional form of the response. Also, computations may be simpler with the quasilikelihood than with the likelihood.

3.2 Approaches to estimation in generalized linear mixed models

3.2.1 Likelihood-based approaches

3.2.1.1 Marginal maximum likelihood approach

The marginal likelihood function of the parameters $\alpha$ and $\theta$ is given by the expression (3.2). By maximizing (3.2) with respect to $\alpha$ and $\theta$, one obtains the maximum likelihood estimates of $\alpha$ and $\theta$. These are often called the marginal maximum likelihood estimates (MMLE) of the parameters though maximum marginal likelihood estimates might be a more appropriate term. The various approaches to optimization described in Section 2.1 can be applied to maximize the likelihood (see Section 4.2).

3.2.1.2 Restricted maximum likelihood approach

Marginal maximum likelihood estimators tend to underestimate the variance parameters for small samples. To see this, consider the simple linear model

$$y \sim N(X\beta, \sigma^2 I),$$

where $\hat{\sigma}^2 = \frac{1}{n}(y - X\hat{\beta})' (y - X\hat{\beta})$, the maximum likelihood estimator of $\sigma^2$, is not unbiased for $\sigma^2$. This same type of result is obtained for mixed models. To address this, one can compute the restricted maximum likelihood estimates (REML) of the variance components (Laird and Ware, 1982) which typically have less bias. In the case
of the simple linear model, $\sigma^2 = \frac{1}{n-\text{rank}(X)}(y - \hat{X}\hat{\beta})'(y - \hat{X}\hat{\beta})$, the restricted maximum likelihood estimate of $\sigma^2$ is unbiased. To compute the restricted maximum likelihood estimates, one maximizes the restricted likelihood function. The latter can be obtained in terms of error contrasts or in a Bayesian approach by integrating out the fixed effects $\alpha$ from the likelihood function (3.2) assuming a flat prior distribution on $\alpha$. Hence the restricted likelihood function for generalized linear mixed models is:

$$L^*(\theta|y) \propto \int_{\alpha} \int_{b} |D(\theta)|^{-1/2} \exp \left( -\frac{b'D^{-1}(\theta)b}{2} \right) \left\{ \prod_{i=1}^{n} f(y_i|\xi_i) \right\} dbd\alpha.$$  

By maximizing the restricted likelihood function given above with respect to $\theta$, one obtains the restricted maximum likelihood estimate of $\theta$. Then the estimate of $\alpha$ is obtained by maximizing $L(\alpha, \theta|y)$ with respect to $\alpha$, treating $\theta$ as fixed at the corresponding restricted maximum likelihood estimate.

### 3.2.2 Bayesian approach

Bayesian analysis of generalized linear mixed models is relatively easy to implement (compared to its frequentist counterpart), especially with the advent of MCMC algorithms. In a Bayesian analysis of generalized linear mixed models, one would draw inference about the unknown parameters $\alpha$ and $\theta$ from their posterior distribution given the observed data.

The posterior distribution of $\alpha$ and $\theta$, denoted as $p(\alpha, \theta|y)$, is given by

$$p(\alpha, \theta|y) = \frac{p(y|\alpha, \theta)p(\alpha, \theta)}{\int \int p(y|\alpha, \theta)p(\alpha, \theta)d\alpha d\theta},$$

where $p(\alpha, \theta)$ is the prior distribution on $\alpha$ and $\theta$. Note that $p(y|\alpha, \theta)$ is the marginal likelihood, computation of which requires that $b$ has to be integrated out, i.e.,

$$p(y|\alpha, \theta) \propto |D(\theta)|^{-1/2} \int_{b} \left\{ \prod_{i=1}^{n} f(y_i|\alpha, b) \right\} \exp \left( -\frac{b'D^{-1}(\theta)b}{2} \right) db.$$
However, we will shortly see that it is possible to perform a Bayesian analysis without integrating out the random effects $b$. Markov chain Monte Carlo (MCMC) methods (Section 2.3) can be used to get a sample from the posterior distribution $p(\alpha, \theta|y)$ and that sample can be used to make posterior inferences about the parameters. Usually, it is enough to work with the numerator of (3.4) in MCMC algorithms.

One advantage of the Bayesian approach is that there is a way to avoid integrating the random effects. One can generate a random sample from the posterior distribution $p(\alpha, b, \theta|y)$, which is given by

$$p(\alpha, b, \theta|y) = \frac{p(y|\alpha, b)p(b|\theta)p(\alpha, \theta)}{\int p(y|\alpha, b)p(b|\theta)p(\alpha, \theta)d\alpha db d\theta}$$

$$\propto p(y|\alpha, b)p(b|\theta)p(\alpha, \theta)$$

$$= \left\{ \prod_{i=1}^{n} f(y_i|\alpha, b) \right\} |D(\theta)|^{-1/2} \exp\left( -\frac{b'D^{-1}(\theta)b}{2} \right) p(\alpha, \theta)$$

The sampled values of $\alpha$ and $\theta$ obtained this way constitute a random sample from $p(\alpha, \theta|y)$.

Inferences for the parameters are obtained by computing different quantities from the posterior sample, e.g., the mean of the sampled values of any parameter is an estimate of the posterior mean of the same parameter.

### 3.3 Testing hypotheses about variance components in generalized linear mixed models

Inferences about the contribution of the random effects to the generalized linear mixed model can be made by examining point (or interval) estimates of the variance parameters in $D$. In many practical problems, researchers may like to test whether a
particular variance component is zero (or equivalently compare the relevant models). We now briefly review different approaches to carry out the test.

3.3.1 Frequentist approaches

3.3.1.1 Likelihood-ratio test

Assuming maximum likelihood estimates can be obtained, the likelihood ratio test statistic for comparing the two models (those implied by the null hypothesis and the alternative hypothesis) can be computed. Using notation from the beginning of this chapter, suppose the unrestricted marginal likelihood \( L(\alpha, \theta | y) \) for a generalized linear mixed model is given by

\[
L(\alpha, \theta | y) \propto |D(\theta)|^{-1/2} \int_b \left\{ \prod_{i=1}^n f(y_i | \alpha, b) \right\} \exp\left( -\frac{b'D^{-1}(\theta)b}{2} \right) \, db,
\]

where \( \theta \) denotes the variance parameters in the model. Suppose \( \theta \) can be partitioned as \( \theta = (\theta_1, \theta_2) \) and that we are interested in testing \( H_0 : \theta_1 = \theta_1^* \) (often \( \theta_1^* = 0 \)). The likelihood ratio for testing \( H_0 \) is given by

\[
LR = \frac{L(\hat{\alpha}, \hat{\theta}_1^*, \hat{\theta}_2 | y)}{L(\hat{\alpha}, \hat{\theta}_1 | y)}.
\]

where \((\hat{\alpha}, \hat{\theta}_1^*, \hat{\theta}_2)\) maximizes \( L(\alpha, \theta_1^*, \theta_2 | y) \) over all choices of \( \alpha \) and \( \theta_2 \) while \((\hat{\alpha}, \hat{\theta})\) maximizes \( L(\alpha, \theta | y) \) over all choices of \((\alpha, \theta)\). The usual likelihood ratio test statistic is given by \( -2 \log(LR) \). A difficulty in carrying out the test for \( H_0 \) is that if the null value of the variance parameter lies on the boundary of its support (i.e., \( \theta_1^* = 0 \)), then the asymptotic distribution of the likelihood ratio test statistic under the null hypothesis is not readily available. The usual \( \chi^2 \) approximation of the likelihood ratio test statistic may not hold since the regularity conditions require that the null hypothesis be on the interior of the parameter space. One way to resolve this problem is to obtain an approximate reference distribution, e.g., by using a parametric bootstrap to simulate data sets.
assuming the null hypothesis is true. Then we compare the observed likelihood ratio test statistic with the simulated null distribution. However, simulation is impractical if the model is complex or the data set is large since the computation of the likelihood ratio test statistic then becomes very time-consuming.

3.3.1.2 Score test

Lin (1997) derives a score test by approximating the logarithm of the quasilikelihood of a generalized linear mixed model defined by (3.3) using a Laplace approximation (Section 2.2.2) to integrate out the random effect(s).

Let \( \hat{l}(\alpha, \theta | y) \) denote the approximate log-quasilikelihood function for a generalized linear mixed model after using the Laplace approximation to integrate out the random effects vector \( b \). We will discuss the procedure for the global test \( H_0 : \theta = 0 \). Suppose the components of \( \theta \) are given by \( \theta = (\theta_1, \theta_2, \ldots, \theta_m) \) and suppose that \( \hat{\alpha}_0 \) is the maximum likelihood estimate of \( \alpha \) under the null hypothesis. Then the score function for testing \( H_0 \), \( U_{\theta}(\hat{\alpha}_0) = \left( U_{\theta_1}(\hat{\alpha}_0), \ldots, U_{\theta_m}(\hat{\alpha}_0) \right)' \), is given by

\[
U_{\theta_j}(\hat{\alpha}_0) = \frac{\partial \hat{l}(\alpha, \theta | y)}{\partial \theta_j} \bigg|_{\theta=0, \alpha=\hat{\alpha}_0}.
\]

The global score test statistic used to test \( H_0 \) is given by

\[
\chi^2 = U_{\theta}(\hat{\alpha}_0)' I(\hat{\alpha}_0)^{-1} U_{\theta}(\hat{\alpha}_0),
\]

where \( I(\hat{\alpha}_0) \) is the efficient information matrix of \( \theta \) evaluated under \( H_0 \), taking the form

\[
I(\hat{\alpha}_0) = I_{\theta\theta} - I_{\alpha\theta} I_{\alpha\alpha}^{-1} I_{\alpha\theta},
\]

where

\[
I_{\theta\theta} = E \left( \frac{\partial \hat{l}(\alpha, \theta | y)}{\partial \theta} \frac{\partial \hat{l}(\alpha, \theta | y)}{\partial \theta'} \right),
\]

\[
I_{\alpha\theta} = E \left( \frac{\partial \hat{l}(\alpha, \theta | y)}{\partial \alpha} \frac{\partial \hat{l}(\alpha, \theta | y)}{\partial \theta'} \right),
\]

\[
I_{\alpha\alpha} = E \left( \frac{\partial \hat{l}(\alpha, \theta | y)}{\partial \alpha} \frac{\partial \hat{l}(\alpha, \theta | y)}{\partial \alpha'} \right).
\]
and the partial derivatives and expectations are all calculated under the null hypothesis, i.e., at \( \theta = 0 \). Under some regularity conditions in proposition 1 (Lin, 1997, page 314), the global score test statistic \( \chi^2_G \) follows a \( \chi^2 \) distribution with \( m \) (\( m \) is the number of components in \( \theta \)) degrees of freedom asymptotically under the null hypothesis. Lin derives test statistics for testing hypotheses about the individual variance components as well.

### 3.3.2 Bayesian approach

The Bayesian approach to test a hypothesis about the variance component(s) is to compute the Bayes factor \( BF^{01} = \frac{p(y|M_0)}{p(y|M_1)} \) that compares the marginal densities of \( y \) under the two models, \( M_0 \) (one or more of the variance components is zero) and \( M_1 \) (variance unrestricted), where

\[
p(y|M) = \int p(y|\omega, M)p(\omega|M)d\omega
\]

is the marginal density under model \( M \) and \( \omega \) denotes the parameters of model \( M \). The Bayes factor then summarizes the evidence provided by the data in favor of one scientific theory represented by a statistical model as opposed to another. Note that the Bayes factor does not depend on the prior probabilities assigned to the models.

Another way to express the Bayes factor is the following:

\[
BF^{01} = \frac{p(M_0|y)}{p(M_0)} \frac{p(M_0)}{p(M_1)},
\]

i.e., the Bayes factor is the ratio of posterior odds and prior odds. This expression appears to contradict the last sentence of the previous paragraph in that the Bayes factor now seems to depend on \( p(M_0) \) and \( p(M_1) \), but there is no conflict as this is mathematically equivalent to the former. This expression is useful in forming an estimate
Table 3.1 Interpretation of Bayes factors

<table>
<thead>
<tr>
<th>$2\log_e(BF_{10})$</th>
<th>$BF_{10}$</th>
<th>Evidence against $M_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-2</td>
<td>1-3</td>
<td>Not worth more than a bare mention</td>
</tr>
<tr>
<td>2-6</td>
<td>3-20</td>
<td>Positive</td>
</tr>
<tr>
<td>6-10</td>
<td>20-150</td>
<td>Strong</td>
</tr>
<tr>
<td>$&gt;10$</td>
<td>$&gt;150$</td>
<td>Very strong</td>
</tr>
</tbody>
</table>

of the Bayes factor using reversible jump MCMC (see Section 4.4.3) algorithm, which obtains empirical estimates of $p(M_0|y)$ and $p(M_1|y)$.

Kass and Raftery (1995) provide a comprehensive review of Bayes factors including information about their interpretation. Their suggested interpretation of Bayes factors, which is modified from Jeffreys' (1961) scale, is given in Table 3.1. They express Bayes factors as providing evidence against the null model, i.e., they provide the scale in terms of $BF_{10}$.

Strictly speaking, in order to be able to use Bayes factors, one should use proper prior distributions for all the parameters in the models being compared. This is because Bayes factors may not be defined if one uses improper prior distributions for the parameters in the models — the marginal density of the data under the models may not be defined in that situation. Consider the simple example where

$$y|\theta \sim N(\theta, 1), \quad p(\theta) \propto 1.$$  

Then,

$$p(y) \propto \int \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y-\theta)^2} d\theta = 1, \quad -\infty < y < \infty,$$

which is not a proper density. Hence if we want to test a null hypothesis about $\theta$ against the alternative hypothesis that $\theta$ can take any value in $(-\infty, \infty)$, we cannot use the Bayes factor because it is undefined in this situation (because the marginal density of the data under the alternative model is not defined).
However, there are situations where one may be able to compute Bayes factors without using proper prior distributions for all the parameters in the models. For example, if two models being compared have a number of common parameters, then one can usually compute a Bayes factor for comparing the two models using the same improper prior distributions on the common parameters under both the models.
4 COMPUTATION FOR GENERALIZED LINEAR MIXED MODELS

Difficulty in carrying out the needed calculations for generalized linear mixed models has always been a stumbling block limiting their application. In this chapter, we discuss the different computational issues associated with approximating Bayes factors for variance component testing for these models. Chapter 2 provides an elementary discussion of a number of statistical computing tools. In this chapter we discuss the practicalities of applying those tools for computing Bayes factor estimates for generalized linear mixed models.

4.1 Computation of the likelihood

Calculation of the Bayes factor requires computing

\[ p(y) = \int \int p(y|\alpha, \theta)p(\alpha, \theta)d\alpha d\theta, \]

where we drop the model index \( M \) for simplicity. The above computation is more complicated than it seems for generalized linear mixed models because \( p(y|\alpha, \theta) \) is itself defined as an integral over the random effects distribution. Here we focus on computing \( p(y|\alpha, \theta) \), also known as the likelihood \( L(\alpha, \theta|y) \). Additional steps required to form Bayes factor estimates are discussed in subsequent sections.

The likelihood function (which is also known as the marginal likelihood function) \( L(\alpha, \theta|y) \) for generalized linear mixed models is obtained by integrating out the random
effects from the conditional density of the response $y$ using the prior distribution of the random effects:

$$L(\alpha, \theta|y) = \int_b \left\{ \prod_{i=1}^{n} p(y_i|\xi_i) \right\} p(b|\theta) db$$

$$= \left(2\pi\right)^{-\frac{d}{2}} |D(\theta)|^{-1/2} \int_b \left\{ \prod_{i=1}^{n} p(y_i|\alpha, b) \right\} \exp\left(-\frac{b'D^{-1}(\theta)b}{2}\right) db,$$

where $d$ is the dimension of $D(\theta)$. The integral cannot be computed analytically except for normal linear models. The different approaches that have been suggested to integrate out the random effects to compute the marginal likelihood are discussed below.

4.1.1 Numerical integration

One way to compute the likelihood function is to use a numerical integration technique (e.g., Simpson's rule, which is discussed in Section 2.2.1) to integrate out the random effects. However, for even moderately large problems (e.g., for large data sets or high-dimensional random effects), this becomes too time-consuming to implement and hence is of limited use.

4.1.2 Laplace approximation

A number of authors, e.g., Wolfinger and O'Connell (1993) and Breslow and Clayton (1993) use Laplace approximation (Section 2.2.2) to integrate out the random effect(s) to obtain the generalized linear mixed model likelihood. This method is easy to implement even for large problems, but may give unsatisfactory results in certain situations. For example, Booth and Hobert (1999) comment that maximum likelihood estimates obtained using the Laplace approximation to the likelihood have some unsatisfactory properties; in particular, they are known to be inconsistent under standard (small domain) asymptotic assumptions and the size of the asymptotic bias can be substantial if the variance components are not small.
4.1.3 Importance sampling

Geyer and Thompson (1992) and Gelfand and Carlin (1993) suggest the use of importance sampling (Section 2.2.4) to estimate the value of the likelihood function. Starting from (4.2), for an importance sampling distribution \( h(b) \), \( L(\alpha, \theta|y) \) can be expressed as

\[
(2\pi)^{-\frac{n}{2}} |D(\theta)|^{-1/2} \int_b \left\{ \prod_{i=1}^{n} p(y_i|\alpha, b) \right\} \exp\left( -\frac{b' D^{-1}(\theta) b}{2} \right) db
\]

\[= (2\pi)^{-\frac{n}{2}} |D(\theta)|^{-1/2} \int_b \frac{1}{h(b)} \left\{ \prod_{i=1}^{n} p(y_i|\alpha, b) \right\} \exp\left( -\frac{b' D^{-1}(\theta) b}{2} \right) h(b) db \]

\[\approx (2\pi)^{-\frac{n}{2}} |D(\theta)|^{-1/2} \frac{1}{N} \sum_{k=1}^{N} \frac{1}{h(b^{(k)})} \left\{ \prod_{i=1}^{n} p(y_i|\alpha, b^{(k)}) \right\} \exp\left( -\frac{b^{(k)}' D^{-1}(\theta) b^{(k)}}{2} \right), \]

where \( b^{(k)}, k = 1, 2, \ldots, N \), is a sample selected from the importance sampling distribution \( h(b) \).

As was discussed in Section 2.2.4, the choice of the importance sampling density is an important issue in approximating an integral using importance sampling approach. Theoretically, for the integral to be estimated precisely, \( h(b) \), the importance sampling density, should be of similar shape as and should have heavier tails than the product \( \left\{ \prod_{i=1}^{n} p(y_i|\alpha, b) \right\} \exp\left( -\frac{b' D^{-1}(\theta) b}{2} \right) \).

For high-dimensional \( b \) (e.g., for the spatial Markov random field models that we will use in Chapter 6), the choice of the importance sampling density poses a challenge to statisticians. Since we want \( h(b) \) to be of similar shape as the posterior distribution of \( b \) given \( y, \alpha \) and \( \theta \), we find using a sample generated from the posterior distribution of \( b \) given \( y, \alpha \) and \( \theta \) a very useful method for determining an importance sampling density. To generate the sample, we fix the values of \( \alpha \) and \( \theta \) at the point for which the likelihood is required and then use a MCMC algorithm (Section 2.3) to generate a sample from the posterior distribution of \( b \) given \( y, \alpha \) and \( \theta \). After making sure of the convergence of the chain, we compute the sample mean and sample variance of a posterior sample with large sample size. We use a \( t_4 \) density with the same first two moments as the posterior sample.
as $h(b)$. We find that this importance sampling density helps to estimate $L(\alpha, \theta | y)$ with reasonable precision for large enough size of the importance sample.

4.2 Computation of posterior mode

Several methods for approximating the Bayes factor require an estimate of the mode of the posterior distribution. In this section, we address this topic and the related topic of finding maximum likelihood estimates for generalized linear mixed models. Note that while one needs to maximize $p(y | \alpha, \theta)$ over all possible values of $\alpha$ and $\theta$ to get the maximum likelihood estimates, the computation of the posterior mode requires the maximization of $\frac{p(y | \alpha, \theta) p(\alpha, \theta)}{p(y)}$ or equivalently that of $p(y | \alpha, \theta) p(\alpha, \theta)$ over all possible values of $\alpha$ and $\theta$.

Because of the need to integrate out the random effect terms to obtain the marginal likelihood function, computation of the posterior modes or maximum likelihood estimates (both MMLE and REML) for generalized linear mixed models is not straightforward. Some approaches that can be used to obtain the posterior mode or the maximum likelihood estimate are discussed below. One point to note is that all of the following approaches can be used to obtain either the posterior mode, the MMLE's or the REML's.

4.2.1 Newton-Raphson algorithm

If the marginal likelihood function can be easily computed, one can use traditional optimization techniques, e.g., the Newton-Raphson method (Section 2.1.1) to compute the posterior mode (or the maximum likelihood estimate) of $\alpha$ and $\theta$. However, this method is of limited use for large problems. If one uses an accurate technique like numerical integration (Section 4.1.1) or importance sampling (Section 4.1.3) to compute the marginal likelihood for a large problem, the likelihood computation becomes too intensive and the Newton-Raphson algorithm becomes very time-consuming. On the
other hand, use of a simple approximation like the Laplace approximation to compute the marginal likelihood (Section 4.1.2) may result in estimates with unsatisfactory properties (Booth & Hobert, 1999); specifically, one may have inconsistent estimates using this method.

4.2.2 EM algorithm

For the generalized linear mixed models, the random effects can be viewed as missing data and the EM algorithm applied to obtain posterior modes (or maximum likelihood estimates). Here, we talk about the computation of the posterior mode.

For generalized linear mixed models, the M-step is usually relatively simple, but the E-step causes problems (McCulloch, 1997) since the expectation cannot be computed analytically — one has to depend on approximation or simulation to perform the step.

Following the discussion of the EM algorithm in Section 2.1.2 and the notations for generalized linear mixed models from Section 3.1, the \((t + 1)\)-th E-step requires computation of

\[
Q(\alpha, \theta | \alpha^{(t)}, \theta^{(t)}) = E[\log \{ p(y, b | \alpha, \theta) p(\alpha, \theta) \} | y, \alpha^{(t)}, \theta^{(t)}],
\]

where the expectation has to be taken with respect to \(p(b | y; \alpha^{(t)}, \theta^{(t)})\). In the corresponding M-step, \(Q(\alpha, \theta | \alpha^{(t)}, \theta^{(t)})\), computed in the E-step, is maximized with respect to \((\alpha, \theta)\) to obtain \((\alpha^{(t+1)}, \theta^{(t+1)})\). Now,

\[
p(y, b | \alpha, \theta) = p(y | \alpha, b) p(b | \theta) \propto p(y | \alpha, b) |D(\theta)|^{-1/2} \exp\left\{ -\frac{1}{2} b' D^{-1}(\theta) b \right\}
\]

Looking at formula (4.3) and the above, the M-step then amounts to obtaining \((\alpha^{(t+1)}, \theta^{(t+1)})\) by maximizing the expectation of

\[
2 \log [ p(y | \alpha, b) ] - \log |D(\theta)| - b' D^{-1}(\theta) b + 2 \log [ p(\alpha, \theta) ]
\]
with respect to \((\alpha, \theta)\), where the expectation has to be taken with respect to the density \(p(b|y, \alpha^{(t)}, \theta^{(t)})\).

From the above discussion, it is clear that the E-step is not easy since \(p(b|y, \alpha^{(t)}, \theta^{(t)})\) is almost never a known density and the expectation with respect to it can not be computed analytically. Once the E-step is done, the M-step consists in maximizing the expectation of (4.4) over the (relatively) low-dimensional parameter space \((\alpha, \theta)\). This is straightforward using an optimization method like the Newton-Raphson method. Different approaches have been suggested to approximate the expectation in the E-step of the EM algorithm when it cannot be computed analytically. They are described below.

For computing maximum likelihood estimates for generalized linear mixed models, the procedure is almost identical to the above except that the prior distribution \(p(\alpha, \theta)\) is omitted. In the \((t + 1)\)-th E-step, one has to compute

\[
\hat{Q}(\alpha, \theta|\alpha^{(t)}, \theta^{(t)}) = E[\log\{p(y, b|\alpha, \theta)\}|y, \alpha^{(t)}, \theta^{(t)}],
\]

where the expectation has to be taken with respect to \(p(b|y; \alpha^{(t)}, \theta^{(t)})\). In the corresponding M-step, one has to maximize the expectation computed in the E-step. From the discussion above, it is easy to see that the M-step then requires maximizing the expectation of

\[
2\log[p(y|\alpha, b)] - \log|D(\theta)| - b^TD^{-1}(\theta)b,
\]

with respect to \((\alpha, \theta)\), where the expectation has to be taken with respect to \(p(b|y; \alpha^{(t)}, \theta^{(t)})\).

### 4.2.2.1 Approximate EM algorithm

Stiratelli et. al. (1984) suggest an approximate EM algorithm to compute the restricted maximum likelihood estimate of the variance parameters of generalized linear models with random effects. We adapt their method to compute the joint posterior mode
(and maximum likelihood estimates) for these models. In this method, in the E-step, the
expectation in the right hand side of (4.3) is computed using a normal approximation
for \( p(b|y, \alpha^{(t)}, \theta^{(t)}) \) centered at its mode. We can write \( p(b|y, \alpha^{(t)}, \theta^{(t)}) \) as

\[
\begin{align*}
p(b|y, \alpha^{(t)}, \theta^{(t)}) & \propto p(y|\alpha^{(t)}, b)p(b|\theta^{(t)}) \\
& \propto p(y|\alpha^{(t)}, b) \cdot |D(\theta^{(t)})|^{-1/2} \cdot \exp \left( -\frac{b'D^{-1}(\theta^{(t)})b}{2} \right)
\end{align*}
\]

and maximize the logarithm of the right hand side of the above equation with respect to
\( b \) using the Newton-Raphson algorithm. The vector \( \hat{b} \) maximizing \( \log\{p(b|y, \alpha^{(t)}, \theta^{(t)}) \} \)
is taken as the mean of the normal approximation. The inverse of the negative Hessian
matrix of \( \log\{p(b|y, \alpha^{(t)}, \theta^{(t)}) \} \) computed at \( \hat{b} \) (which can be obtained without much
extra effort from the last iteration of the Newton-Raphson algorithm) is taken as the
variance of the normal distribution.

4.2.2.2 Stochastic EM methods

A superior alternative to the approximate EM algorithm is to implement the EM
algorithm for the generalized linear mixed models using simulation of some kind to
complete the E-step.

**MCEM algorithm by Wei and Tanner:** Wei and Tanner (1990) suggest replacing
the expectation in (4.3) with a Monte Carlo approximation. The expectation has to
be taken with respect to \( p(b|y, \alpha^{(t)}, \theta^{(t)}) \). So if a random sample \( b_{t,1}, b_{t,2}, \ldots, b_{t,m} \) from
\( p(b|y, \alpha^{(t)}, \theta^{(t)}) \) can be produced, a Monte Carlo approximation of \( Q(\alpha, \theta|\alpha^{(t)}, \theta^{(t)}) \) is
available as:

\[
Q_m(\alpha, \theta|\alpha^{(t)}, \theta^{(t)}) = \frac{1}{m} \sum_{i=1}^{m} \log\{p(y, b_{t,i}|\alpha, \theta)p(\alpha, \theta)\}
\]

Using \( Q_m \) in place of \( Q \) in every E-step gives the Monte Carlo EM algorithm. The
Monte Carlo EM algorithm converges to the maximum likelihood estimate under suitable
regularity conditions (see, for example, Chan and Ledolter (1995)).
McCulloch (1994, 1997) uses a Markov chain to generate the required sample. His approach is known as the Markov chain Monte Carlo EM (MCMCEM) method. In the E-step of the \((t+1)\)-th iteration in his method, a Markov chain Monte Carlo (Section 2.3) algorithm is used to generate a random sample from \(p(b|y, \alpha^{(t)}, \theta^{(t)})\). Specifically, he uses a Gibbs sampler in McCulloch (1994) for a binomial-probit model and Metropolis-Hastings algorithm in McCulloch (1997) for the general case. After drawing the sample, one replaces the integral in (4.3) by a Monte Carlo sum over the sampled values.

Along the same line, Booth and Robert (1999) propose an approach which uses simulated random samples from \(p(b|y, \alpha^{(t)}, \theta^{(t)})\), generated via rejection sampling, using the prior distribution of \(b\) (which, as we mentioned in Section 2.1, is a normal distribution most of the time) as the candidate distribution.

**Importance sampling EM approach:** Booth and Hobert (1999) suggest using an importance sampling approximation (Section 2.2.4) of the expectation in (4.3). Suppose one draws a random sample \(b_{1,1}, b_{1,2}, \ldots, b_{1,m}\) from an importance sampling density \(p^*(b)\), which has the same support as \(p(b|y, \alpha^{(t)}, \theta^{(t)})\). Then the importance sampling Monte Carlo estimate of \(Q(\alpha, \theta|\alpha^{(t)}, \theta^{(t)})\) is given by

\[
Q_m(\alpha, \theta|\alpha^{(t)}, \theta^{(t)}) = \frac{1}{m} \sum_{i=1}^{m} w_{i,t} \log \{ p(y, b_{i,t}|\alpha, \theta) p(\alpha, \theta) \},
\]

where the importance weights \(w_{i,t}\)'s are obtained by

\[
w_{i,t} = p(b_{i,t}^*|y, \alpha^{(t)}, \theta^{(t)}) / p^*(b_{i,t}^*).
\]

Since \(p(b|y, \alpha^{(t)}, \theta^{(t)})\) involves an unknown normalizing constant, so do the importance weights. But since the unknown normalizing constants depend on the known value \((\alpha^{(t)}, \theta^{(t)})\) and not on \((\alpha, \theta)\), they do not affect the M-step and hence are irrelevant.

One important aspect here is the choice of the importance sampling density \(p^*(.\). Booth and Hobert (1999) propose a multivariate \(t\) importance density whose mean
and variance are the same as the mode and curvature of $p(b|y, \alpha^{(t)}, \theta^{(t)})$. Suppose $p(b|y, \alpha^{(t)}, \theta^{(t)}) = ce^{l(b)}$, supressing the dependence of $l(b)$ on $y, \alpha^{(t)}$ and $\theta^{(t)}$, where $c$ is the normalizing constant. Let $l^{(1)}(b)$ denote the vector of first derivatives of $l(b)$ and $l^{(2)}(b)$ the second derivative matrix. Suppose further that $\hat{b}$ maximizes $l(b)$ satisfying the equation

$$l^{(1)}(\hat{b}) = 0.$$ 

Then the approximation of the mean and variance for the multivariate $t$ importance density are $\hat{b}$ and $[-l^{(2)}(\hat{b})]^{-1}$ respectively.

An approximate estimate of the variance of the estimated MLE can be obtained according to the suggestion of Booth and Hobert as the inverse of the observed information matrix $I$ where,

$$I = -Q^{(2)}(\alpha, \theta|\alpha^{(t)}, \theta^{(t)}) - \text{var}\left[\frac{\partial}{\partial(\alpha, \theta)} \log \left( p(y, b|\alpha, \theta) p(\alpha, \theta) \right) \right]_{y, (\hat{\alpha}, \hat{\theta})_{MLE}},$$  

where

$$Q^{(2)}(\alpha, \theta|\alpha^{(t)}, \theta^{(t)}) = \frac{\partial^2}{\partial(\alpha, \theta)\partial(\alpha, \theta)^T} Q(\alpha, \theta|\alpha^{(t)}, \theta^{(t)}).$$

An approximation of the first term above is supplied by $Q_m^{(2)}$ from the final iteration of the EM algorithm and a Monte Carlo estimate of the second term can be constructed by using the simulations from the last iteration. We use the sample values $b_{T,1}^*, b_{T,2}^*, \ldots, b_{T,m}^*$ where the $T$-th iteration is the last iteration. Then we compute

$$\frac{\partial}{\partial(\alpha, \theta)} \log \left( p(y, b_{T,i}^*|\alpha, \theta) p(\alpha, \theta) \right) \bigg| y, (\hat{\alpha}, \hat{\theta})_{MLE}, i = 1, 2, \ldots, m.$$ 

The variance of these $m$ quantities gives us an estimate of the second term in (4.7).

4.3 Approximating the marginal density $p(y|M)$ under a model

The Bayes factor for comparing two models is the ratio of the marginal densities under the two models, where we repeat that the marginal density $p(y|M)$ under a
model $M$ is given by
\[ p(y|M) = \int p(y|\omega, M)p(\omega|M)d\omega, \]
where the vector of all parameters (i.e., the combination of the fixed effects parameters and the variance parameters for a generalized linear mixed model) in a model is now denoted as $\omega$. For generalized linear mixed models, the marginal densities cannot be computed analytically for either the model $M_1$ with unrestricted variance components or for the model $M_0$ with variance component(s) set to zero. We review a number of approaches for approximating marginal densities that have been applied in other models and then explore their use for generalized linear mixed models.

The approaches considered here are:

- Laplace approximation
- Importance sampling
- Harmonic estimator
- Bridge sampling
- Chib's method

### 4.3.1 Laplace approximation

The Laplace approximation (Section 2.2.2) of the marginal density under a model is obtained by approximating the product of the likelihood function and the prior distribution, i.e., $p(y|\omega, M)p(\omega|M)$, by a normal distribution. The mean $\hat{\omega}$ of the normal distribution is taken as the mode of $p(y|\omega, M)p(\omega|M)$, i.e., the posterior mode, and the variance $\hat{\Sigma}$ is taken as the inverse of the negative Hessian matrix of the log-posterior evaluated at $\hat{\omega}$. The Laplace approximation formula is:
\[ p(y|M) \approx (2\pi)^{d/2}|\hat{\Sigma}|^{1/2}p(y|\hat{\omega}, M)p(\hat{\omega}|M). \]
where \( d \) is the dimension of \( \omega \). The relative error of the approximation is \( O\left( \frac{1}{n} \right) \), where \( n \) is the original sample size. However, there may be problems with this approximation if the posterior mode is on the boundary of the parameter space. This may occur in a generalized linear mixed model where the posterior mode for one or more variance components may be zero, especially if the null model is true. A modification of the Laplace approximation to accommodate the boundary case for linear models is suggested by Pauler et. al. (1999). It relies on being able to analytically integrate out the random effects and thereby create an extended parameter space for the variance components. This is not possible in general for generalized linear mixed models.

We try a little variation of the Laplace approximation to tackle the boundary-value problem. We approximate the product \( p(y | \omega, M) p(\omega | M) \) by a normal distribution whose mean is not the mode of \( p(y | \omega, M) p(\omega | M) \), but a point \( \tilde{\omega} \) close to the mode and whose variance \( \tilde{\Sigma} \) is the inverse of the negative Hessian matrix of the log-posterior evaluated at \( \tilde{\omega} \). Repeating the argument that yields the Laplace approximation (i.e., expanding the logarithm of the integrand in a Taylor series around \( \tilde{\Sigma} \) and then integrate) yields

\[
p(y | M) \approx \frac{1}{(2\pi)^{d/2} |\tilde{\Sigma}|^{1/2}} p(y | \tilde{\omega}, M) p(\tilde{\omega} | M) \exp \left\{ \delta' \tilde{\Sigma} \delta \right\}, \tag{4.8}
\]

where \( \delta \) is the vector of the first derivatives of the log-posterior evaluated at \( \tilde{\omega} \) and one has to make sure that \( \tilde{\Sigma} \) is positive definite. The approximation (4.8) has the potential to be useful for two reasons. First, the posterior mode is usually difficult to obtain while the posterior mean is much easier to compute and (4.8) can theoretically be used at the posterior mean to get an approximation of the marginal density under the model. Second, if the posterior mode is on the boundary of the parameter space, the Laplace approximation gives problems (as discussed above) — but, we can theoretically apply (4.8) at a point close to the posterior mode after making sure that \( \tilde{\Sigma} \) is positive definite at that point. However, from an application point of view, (4.8) is found to be of limited use as we will see in Chapter 6.
4.3.2 Importance sampling

As $p(y|M)$ is an integral, numerical approaches for evaluating integrals can be used. This includes quadrature for moderate-sized and small problems and importance sampling.

From our discussion of importance sampling in Section 2.2.4, we see that if we have a sample $\omega_i, i = 1, 2, \ldots, N$ from an importance sampling distribution $Q$ with corresponding density function $q$, we can approximate the marginal density of the data under model $M$ as

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

As discussed in Section 2.2.4, a practical problem is finding a $Q$ such that $p(y|M)$ is well estimated.

Since one has to draw a random sample of size $N$ from $Q$, a popular choice for $Q$ is the normal distribution (see, e.g., DiCiccio et. al., 1997) with a suitable mean and variance, e.g., mean as the mode of $p(y|\omega, M)p(\omega|M)$ and the variance matrix as the inverse of the negative Hessian matrix of the logarithm of $p(y|\omega, M)p(\omega|M)$ calculated at its mode.

In our work, the importance sampling distribution $Q$ is taken to be a $t$ distribution with mean as the mode of $p(y|\omega, M)p(\omega|M)$ and the variance matrix as the inverse of the negative Hessian matrix of the logarithm of $p(y|\omega, M)p(\omega|M)$ calculated at its mode. For generalized linear mixed models, as $\omega$ contains the variance parameters as well, $p(y|\omega, M)p(\omega|M)$ may be skewed and we find that the $t$ distribution does a better job (than the normal distribution) of approximating it and more importantly ensures that it has tails as heavy as $p(y|\omega, M)p(\omega|M)$.
4.3.3 Harmonic estimator

The harmonic estimator is developed by Newton and Raftery (1994) from the following identity which holds for any density function $h$:

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

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

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

So, if we have a sample $\omega_i, i = 1, 2, \ldots, N$ from the posterior distribution under model $M$, we can approximate the marginal density as

$$p(y|M) \approx \left\{ \frac{1}{N} \sum_{i=1}^{N} \frac{h(\omega_i)}{p(y|\omega_i, M)p(\omega|M)} \right\}^{-1}.$$  \hspace{1cm} (4.9)

The harmonic estimator of the marginal density is then obtained by choosing $h(\omega) = p(\omega|M)$, yielding the formula:

$$p(y|M) \approx \left\{ \frac{1}{N} \sum_{i=1}^{N} \frac{1}{p(y|\omega_i, M)} \right\}^{-1}.$$  \hspace{1cm} (4.9)

This method is simulation-consistent, i.e., the estimated marginal density converges almost surely to the true marginal as $N \to \infty$. However, the estimate is not stable — the estimate of $[p(y|M)]^{-1}$ does not have finite variance. If there is a value of $\omega$ for which the likelihood is small, the estimate is hugely affected by that single value. But this estimate is very easy to calculate and researchers have found this method to give at least a rough idea about the true value of the Bayes factor.

To counter the problem of instability of the harmonic estimator, Gelfand and Dey (1994) suggest the use of (4.9) to estimate $p(y|M)$. It is an unbiased and consistent estimate of the marginal density and satisfies a Gaussian central limit theorem if the tails of
$h(.)$ are thin enough. The closer is the shape of $h(.)$ to the shape of $p(y|\omega, M)p(\omega|M)$, the better is the approximation. If $\omega$ has high dimension, it may be difficult to find a proper $h(.)$.

Satagopan et. al. (2000) suggest a stabilized form of the harmonic estimator that approximates the marginal density as

$$
p(y|M) \approx \left\{ \frac{1}{N} \sum_{i=1}^{N} \frac{1}{p(y|g(\omega_i), M)} \right\}^{-1}
$$

for a sample $\omega_i, i = 1, 2, \ldots, N$ from the posterior distribution under model $M$ and a function $g(.)$ where $g(\omega)$ must reduce the parameter space as much as possible while not making the calculation of $p(y|g(\omega), M)$ too difficult.

### 4.3.4 Bridge sampling

Meng and Wong (1993) take the importance sampling idea one step further to suggest the bridge sampling method to approximate the marginal density. In this method, we again start with an identity:

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

for a function $\gamma$ and a density $q$. Then if we have a sample $\omega_i, i = 1, 2, \ldots, I$ from the posterior distribution as well as a sample $\tilde{\omega}_j, j = 1, 2, \ldots, J$ from $q$, the Meng-Wong bridge estimator of $p(y|M)$ is

$$
p(y|M) \approx \frac{\frac{1}{I} \sum_{j=1}^{J} p(y|\tilde{\omega}_j, M)p(\tilde{\omega}_j|M)\gamma(\tilde{\omega}_j)}{\frac{1}{I} \sum_{i=1}^{I} q(\omega_i)\gamma(\omega_i)}.
$$

Meng and Wong show that the optimal choice of $\gamma$ for a given $q$ is proportional to

$$\left\{ np(y|\omega, M)p(\omega|M)/p(y|M) + Nq(\omega) \right\}^{-1},$$

which requires knowing $p(y|M)$. They discuss iterative methods for selecting $\gamma$. 
4.3.5 Chib's method

A useful approach for estimating the marginal density is developed by Chib (1995) from the identity

$$p(y|M) = \frac{p(y|\omega, M)p(\omega|M)}{p(\omega|y, M)}.$$  (4.10)

Note that the left hand side does not depend on $\omega$ — so the equality must hold for every value of $\omega$. We can thus find $p(y|M)$ by evaluating the right hand side at any one particular choice of $\omega$, say $\omega^*$. Of the three terms on the right hand side, the likelihood function $p(y|\omega, M)$ and the prior distribution $p(\omega|M)$ can be computed at a fixed $\omega = \omega^*$ without much difficulty. The computation of the third term, the posterior ordinate at $\omega = \omega^*$, is not trivial. One can obtain a sample from the posterior distribution (perhaps using a MCMC algorithm as in Section 2.3) and then use, for example, kernel density approximation to estimate the posterior ordinate. Kernel density approximations become unreliable in high dimensions, however. Chib gives more efficient algorithms to estimate the posterior ordinate when either Gibbs sampling or Metropolis algorithm is used to generate a sample from the posterior distribution.

4.3.5.1 Estimating the posterior from Gibbs sampling output

Chib (1995) suggests an algorithm using Gibbs sampler output to estimate the posterior ordinate $p(\omega|y, M)$ when $\omega$ can be partitioned into several blocks so that the full conditional for each block is available in closed form. For simplicity, we start by discussing the case of two blocks, $\omega = (\omega_1, \omega_2)$. To run a Gibbs sampler, one has to iteratively generate $\omega_1$ and $\omega_2$ from the conditional distributions $p(\omega_1|\omega_2, y, M)$ and $p(\omega_2|\omega_1, y, M)$, assumed known, respectively. Suppose a sample $(\omega_1^i, \omega_2^i)$, $i = 1, 2, \ldots, N$ has been drawn from the posterior distribution of $\omega$ using the Gibbs sampler after making sure that the sampler has converged. Note that

$$p(\omega_1^*, \omega_2^*|y, M) = p(\omega_1^*|y, M)p(\omega_2^*|\omega_1^*, y, M).$$
The second term in the right side of the above is assumed known. The first term, $p(\omega^*_2|y, M)$, can be estimated as

\[
p(\omega^*_2|y, M) = \int p(\omega^*_1, \omega_2|y, M) d\omega_2
\]

\[
= \int p(\omega^*_1|\omega_2, y, M) p(\omega_2|y, M) d\omega_2
\]

\[
\approx \frac{1}{N} \sum_{i=1}^{N} p(\omega^*_1|\omega_2^i, y, M)
\]

using the Monte Carlo integration idea (Section 2.2.3) since $\omega_2^i$, $i = 1, 2, \ldots N$ is a sample from $p(\omega_2|y, M)$.

This method generalizes to the case when $\omega$ consists of higher number of blocks. Suppose $\omega$ can be grouped into $B$ blocks as $\omega = (\omega_1, \omega_2, \ldots \omega_B)$. Suppose we know the conditional posterior distributions $p(\omega_r|y, \omega_1, \omega_2, \ldots \omega_{r-1}, \omega_{r+1}, \ldots \omega_B)$ exactly for all $r = 1, 2, \ldots B$. Suppose we have to compute the posterior density

\[
p(\omega^*_r|y) = p(\omega_1^*, \omega_2^*, \ldots \omega_B^*|y),
\]

where we drop the model indicator $M$ for convenience. We can write $p(\omega^*_r|y)$ as:

\[
p(\omega^*_r|y) = p(\omega^*_1|y)p(\omega^*_2|y, \omega^*_1) \ldots p(\omega^*_B|y, \omega^*_1, \ldots \omega^*_{r-1}) \cdot (4.11)
\]

To estimate $p(\omega^*_r|y)$, a Gibbs sampler is run initially to obtain a sample $\omega^{(1)}, \omega^{(2)}, \ldots \omega^{(n)}$ from the posterior distribution of $\omega$ and the first term in the right side of (4.11) can be estimated as:

\[
p(\omega^*_1|y) = \frac{1}{n} \sum_{i=1}^{n} p(\omega^*_1|y, \omega^{(i)}_2, \ldots \omega^{(i)}_B)
\]

using the sample obtained from the initial run.

A typical term in the right hand side of (4.11) is $p(\omega^*_r|y, \omega^*_1, \omega^*_2, \ldots \omega^*_{r-1})$ which is given by

\[
\int p(\omega^*_r|y, \omega^*_1, \ldots \omega^*_{r-1}, \omega_{r+1}, \ldots \omega_B) p(\omega_{r+1}, \ldots \omega_B|y, \omega^*_1, \ldots \omega^*_{r-1}) d\omega_{r+1} \ldots d\omega_B (4.12)
\]
To estimate this term, one continues the sampling with the complete conditional densities of \( \{\omega_r, \omega_{r+1}, \ldots, \omega_B\} \), but with \( \omega_s = \omega^*_s, s = 1, 2, \ldots, r-1 \) in each of these full conditional densities. If the draws from the reduced complete conditional Gibbs run are denoted by \( \{\omega_r^{(i)}, \omega_{r+1}^{(i)}, \ldots, \omega_B^{(i)}\} \) then an estimate of (4.12) is

\[
\hat{p}(\omega^*_r|y, \omega^*_1, \omega^*_2, \ldots, \omega^*_{r-1}) = \frac{1}{n} \sum_{i=1}^{n} p(\omega^*_r|y, \omega^*_1, \omega^*_2, \ldots, \omega^*_{r-1}, \omega^*_{r+1}^{(i)}, \ldots, \omega^*_B^{(i)}).
\]

(4.13)

Then an estimate of the joint posterior density is \( \prod_{r=1}^{B} \hat{p}(\omega_r^*|y, \omega_1^*, \omega_2^*, \ldots, \omega_{r-1}^*) \).

Also notice that if one wishes to estimate \( p(\omega_1^*, \omega_2^*, \ldots, \omega_A^*|y) \), where \( A < B \), then the estimate is \( \prod_{r=1}^{A} \hat{p}(\omega_r^*|y, \omega_1^*, \omega_2^*, \ldots, \omega_{r-1}^*) \). This is important because for generalized linear mixed models, although we would like to estimate \( p(\omega|y) \), it is usually much easier to sample from the posterior distribution \( p(\omega, z|y) \) rather than from \( p(\omega|y) \), after introducing latent data \( z \) in the model. For example, we found out in Section 3.2.2 that it is easier to sample directly from \( p(\alpha, b, \omega|y) \) for generalized linear mixed models rather than directly from \( p(\alpha, \omega|y) \). In that setting, the random effects vector \( b \) plays the role of latent data. Then we can treat \( z \) as the last block of an appended parameter vector \( (\omega, z) \) and include it in all the reduced runs and in the end estimate \( p(\omega|y) \) as \( \prod_{r=1}^{B} \hat{p}(\omega_r^*|y, \omega_1^*, \omega_2^*, \ldots, \omega_{r-1}^*) \).

4.3.5.2 Estimating the posterior from Metropolis output

Chib and Jeliazkov (2001) extend the above idea to allow the use of Metropolis-Hastings output to estimate \( p(\omega|y, M) \). Assume that the parameter vector \( \omega \) can be generated in a single block in the Metropolis-Hastings algorithm for generating parameter values from the posterior distribution. We drop the model indicator \( M \) from the notation for convenience. Let

\[
\alpha(\omega, \omega'|y) = \min \left\{ 1, \frac{p(\omega'|y)q(\omega', \omega|y)}{p(\omega|y)q(\omega, \omega'|y)} \right\},
\]
where the $q(\omega, \omega'|y)$ denotes the proposal density (candidate generating density) for transition from $\omega$ to $\omega'$, which may depend on $y$. Further, letting

$$f(\omega, \omega'|y) = \alpha(\omega, \omega'|y)q(\omega, \omega'|y)$$

denote the sub-kernel of the Metropolis-Hastings algorithm, from the local reversibility of the sub-kernel, one can write for any point $\omega^*$

$$f(\omega, \omega^*|y)p(\omega|y) = p(\omega^*|y)f(\omega^*, \omega|y).$$

Upon integrating both sides of this expression with respect to $\omega$, one obtains the result that the posterior ordinate at $\omega^*$ is given by

$$p(\omega^*|y) = \frac{\int \alpha(\omega, \omega^*|y)q(\omega, \omega^*|y)p(\omega|y)d\omega}{\int \alpha(\omega^*, \omega|y)q(\omega^*, \omega|y)d\omega}.$$ 

The above can also be written as

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

(4.14)

where $E_1$ is the expectation with respect to the posterior $p(\omega|y)$ and $E_2$ is the expectation with respect to the proposal density $q(\omega^*, \omega|y)$. The numerator is estimated by averaging the product within the braces with respect to the draws from the posterior distribution, while the denominator is estimated by averaging the acceptance probability with respect to draws from $q(\omega^*, \omega|y)$, given the fixed value $\omega^*$. The nice thing about the calculation is that it does not require knowledge of the normalizing constant for $p(\omega|y)$.

When there are two or more blocks, the authors give an extended version of this algorithm using multiple MCMC runs, similar to the Chib (1995) approach for the Gibbs sampler outlined above. Suppose $\omega$ can be grouped into $B$ blocks as $\omega = (\omega_1, \omega_2, \ldots, \omega_B)$. We can write the posterior ordinate $p(\omega^*|y)$ as:

$$p(\omega^*|y) = p(\omega_1^*|y)p(\omega_2^*|y, \omega_1^*) \ldots p(\omega_B^*|y, \omega_1^*, \ldots, \omega_{B-1}^*)$$

(4.15)
and consider the estimation of the reduced ordinate \( p(\omega^*_i | y, \omega^*_1, \omega^*_2, \ldots \omega^*_i-1) \). Define
\[
\Omega_{i-1} = (\omega_1, \omega_2, \ldots, \omega_{i-1}), \quad \Omega^{i+1} = (\omega_{i+1}, \omega_{i+2}, \ldots, \omega_B) \text{ and } \bar{\omega} = (\omega_i, \omega_{i-1}).
\]
Suppose we sample from the full conditional density
\[
p(\omega_i | y, \omega_{i-1}) \propto p(\omega) p(y | \omega), i = 1, 2, \ldots B,
\]
using the Metropolis-Hastings algorithm with proposal density \( q(\omega_i, \omega_{i-1}, \Omega^{i+1}) \) and probability of move
\[
\alpha(\omega_i, \omega_{i-1}, \Omega^{i+1}) = \min \left\{ 1, \frac{p(y | \omega_i, \omega_{i-1}, \Omega^{i+1}) p(\bar{\omega}) q(\omega_i, \omega_{i-1}, \Omega^{i+1})}{p(y | \omega_i, \omega_{i-1}, \Omega^{i+1}) p(\bar{\omega}) q(\omega_i, \omega_{i-1}, \Omega^{i+1})} \right\}.
\]
Again by exploiting the local reversibility of the Metropolis-Hastings step for \( \omega_i \) and completely analogous arguments to the ones used to prove (4.14), Chib and Jeliazkov (2001) derive the result
\[
p(\omega^*_i | y, \omega^*_1, \omega^*_2, \ldots \omega^*_i-1) = \frac{E_1 \left\{ \alpha(\omega_i, \omega_i^* | y, \Omega_i^*, \Omega^{i+1}) q(\omega_i^* | y, \Omega_i^*, \Omega^{i+1}) \right\}}{E_2 \left\{ \alpha(\omega_i, \omega_i | y, \Omega_i^*, \Omega^{i+1}) \right\}}, \tag{4.16}
\]
where \( E_1 \) is the expectation taken with respect to the conditional posterior distribution \( p(\omega_i, \Omega^{i+1} | y, \Omega_i^*) \) and \( E_2 \) is that taken with respect to the conditional product measure \( p(\Omega^{i+1} | y, \Omega_i^*) q(\omega_i^* | y, \Omega_i^*, \Omega^{i+1}) \). These two integrals can be estimated as before from the output of the reduced MCMC runs as follows:

(i) Set \( \Omega_{i-1} = \Omega_i^* \) and sample the reduced set of full conditional distributions
\[p(\omega_k | y, \omega_{-k}), k = i, \ldots, B \]
via MCMC. Let the generated draws (after making sure of convergence of the chain) be \( \{\omega^{(g)}_i, \omega^{(g)}_{i+1}, \ldots \omega^{(g)}_B\}, g = 1, 2, \ldots M \).

(ii) Include \( \omega_i^* \) in the conditioning set, let \( \Omega_i^* = (\Omega_i^*, \omega_i^*) \) and remove the full conditional distribution of \( \omega_i \) from the collection in Step (i). Then sample the remaining distributions \( p(\omega_k | y, \omega_{-k}), k = i + 1, \ldots, B \) to produce \( \{\omega^{(j)}_{i+1}, \omega^{(j)}_{i+2}, \ldots \omega^{(j)}_B\}, j = 1, 2, \ldots J \). At each draw of the sample, also draw \( \omega^{(j)}_i \) from \( q(\omega_i^* | y, \omega_i, \Omega_i^*, \Omega^{i+1}) \).

The notation in steps (i) and (ii) is confusing in that the notation is the same except for the superscript \((g \text{ vs } j)\).
(iii) Estimate the reduced ordinate in (4.16) by the ratio of Monte Carlo averages as:

\[ \hat{p}(\omega_i^*|y, \omega_1^*, \ldots \omega_{i-1}^*) = \frac{\sum_{j=1}^{M} \alpha(\omega_i^{(j)}, \omega_i^*|y, \Omega_{i-1}^*, \Omega_{i+1}^{(j)}(s)) q(\omega_i^{(j)}|y, \Omega_{i-1}^*, \Omega_{i+1}^{(j)}(s))/M}{\sum_{j=1}^{J} \alpha(\omega_i^*, \omega_i^{(j)}|y, \Omega_{i-1}^*, \Omega_{i+1}^{(j)}(s))/J}, \] (4.17)

where the average in the denominator may include zeros if there are \( \omega_i^{(j)} \) values that lie outside the support of the posterior distribution.

So an estimate of the joint posterior density at the point \( \omega^* \) is obtained as

\[ \prod_{r=1}^{B} \hat{p}(\omega_r^*|y, \omega_1^*, \omega_2^*, \ldots \omega_{r-1}^*). \]

One important aspect about the algorithm is that the reduced runs are obtained by fixing an appropriate set of parameters and continuing the MCMC simulation with a smaller set of distributions. Therefore, these runs require little coding beyond what is done initially for the sampling of the posterior distribution. Also, observe that the sample \( \{\omega_i^{(j)}, \omega_i^{(j+1)}, \ldots \omega_B^{(j)}\}, j = 1, 2, \ldots J \), produced in step 2 of the algorithm can be used in step 1 of the next reduced run where the ordinate \( p(\omega_{i+1}^*|y, \omega_1^*, \omega_2^*, \ldots \omega_i^*) \) is estimated.

One can take care of latent data \( z \) here using the same approach discussed in Section 4.3.5.1, that is, by treating the latent data as an additional block in an appended parameter vector.

4.3.5.3 Estimating the posterior where both Metropolis and Gibbs are used

Sometimes, it is convenient to sample from a posterior distribution using a Gibbs sampler with Metropolis-Hastings steps used for some conditional distributions. In that case, one may use ideas from both of the Sections 4.3.5.1 and 4.3.5.2 to estimate the posterior ordinate \( p(\omega^*|y) \). Assuming that \( \omega = (\omega_1, \omega_2, \ldots \omega_B) \), one has to use a series
of reduced runs (as in Sections 4.3.5.1 and 4.3.5.2) and depending on whether one uses a Gibbs step or a Metropolis step to generate $\omega_i|y, \omega_{-i}$, one may use formula (4.13) or (4.17) to estimate $p(\omega^*_i|y, \omega^*_1, \omega^*_2, \ldots, \omega^*_{i-1})$.

4.3.5.4 Choice of $\omega^*$ in Chib's method

The identity (4.10) is true for any choice $\omega = \omega^*$. However, for efficiency of estimation, $\omega^*$ is generally taken to be a high density point in the support of the posterior distribution. The most popular choice of $\omega^*$ is the posterior mean since it is easily identified from the posterior draws. Sometimes, if the posterior distribution is skewed, the posterior mode may be a better choice than the posterior mean. For generalized linear mixed models, the posterior distribution of the variance parameters are mostly skewed — hence the posterior mode will be a better choice of $\omega^*$. The posterior mode can be found using methods of Section 4.2. Note that the posterior mode is likely to produce a better estimate of the marginal density under a generalized linear mixed model, but requires additional computation.

4.3.5.5 Chib's method for generalized linear mixed models

Here we discuss a number of points that pertain to the use of Chib's methods for generalized linear mixed models.

As has been mentioned in Sections 4.3.5.1 and 4.3.5.2, when applying the Chib's method to generalized linear mixed models, the random effects are usually treated as latent data.

A key to using the Chib's method is doing efficient blocking. The choice of blocks of $\omega$ should depend on the model at hand. However, for most of the generalized linear mixed models, partitioning $\omega$ into two blocks is the most efficient way — one block containing the fixed effects parameters and the other containing the variance parameters.
For some of the generalized linear mixed models (as in our first example in Chapter 5), Gibbs sampling can be used to generate from the posterior distribution and therefore one can use the techniques discussed in Section 4.3.5.1. However, in most of the practical applications, because of the form of the likelihood function of the data, the conditional densities of some of the parameter blocks are usually of unknown form (irrespective of the prior distributions used) for generalized linear mixed models. Hence one has to use Metropolis steps to generate those. Our discussion in Section 4.3.5.3 becomes useful in that context. Usually, the variance parameters for these models have conditional densities of known form with conjugate prior distributions while generating the fixed effects and the random effects require the use of Metropolis steps.

Definition of $\omega$ in the computation of marginal density under a generalized linear mixed model: In all the above methods for computing the marginal density under a generalized linear mixed model, one needs to compute the marginal likelihood $p(y | \omega, M)$ (discussed in Section 4.1) for one or more values of $\omega$. If the accurate computation of $p(y | \omega, M)$, which involves integrating out the random effects, is time-consuming, some of the methods (especially those requiring more than one marginal likelihood computation) become impractical. This is frequently the case with generalized linear mixed models. There is one way to get around this problem however.

We can write the marginal density $p(y)$ (we drop the model indicator for simplicity) under a generalized linear mixed model as:

$$p(y) = \int \int p(y | \alpha, \theta) p(\alpha, \theta) d\alpha d\theta$$

$$= \int \int \int p(y | \alpha, b) p(b | \theta) p(\alpha, \theta) db d\alpha d\theta$$

(4.18)

(4.19)

Hence, rather than using $\omega = (\alpha, \theta)$ and (4.18) along with using a time-consuming method to compute the marginal likelihood $p(y | \alpha, \theta)$, it is often simpler and time-saving to include the random effects in $\omega$ along with the fixed effects parameters and
the variance parameters, i.e., to take $\omega = (b, \alpha, \theta)$. The big advantage of this definition of $\omega$ is that the computation of the likelihood function

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

becomes very easy. However, as a price to pay, the dimension of the parameter space then increases from the case when $\omega = (\alpha, \theta)$ by the number of components in $b$ and that number may be high even for simple generalized linear mixed models. Because of the 'curse of dimensionality' introduced artificially, the different methods then need careful handling and one usually requires a large number of simulations to achieve any desired level of accuracy. Suppose we include the random effects as well in $\omega$ before applying the importance sampling method (Section 4.3.2) to compute $p(y)$. In that case, finding a proper importance sampling density $Q$ on $\omega$ for the importance sampling becomes a difficult task because of the high dimensionality of $\omega$ and one needs a large importance sample size to compute $p(y)$ with any desired level of accuracy.

**Transformation of the variance parameters:** Some of the methods discussed above require the approximation of the posterior distribution by a symmetric distribution. For example, in the Laplace approximation (Section 4.3.1), one approximates the posterior distribution by a normal distribution while in the importance sampling method (Section 4.3.2), the importance density (that is supposed to be of similar shape to the posterior distribution) is usually taken as a normal or a t-distribution. In those cases, working with a transformation of the variance component (e.g., the logarithmic transformation) is found to be very useful because the transformation makes the posterior distribution closer to a symmetric distribution and more accurate results are obtained.
4.4 Approximating the Bayes factor

The key contribution of our work is to bring the computational approaches we have discussed so far to bear on the problem of approximating the Bayes factor to test for the variance components for generalized linear mixed models and demonstrate with two examples (Chapters 5 and 6). As defined in the previous chapter, the Bayes factor for comparing two models is the ratio of the marginal densities under the two models. For generalized linear mixed models, the marginal densities cannot be computed analytically for either the model with unrestricted variance components ($M_1$) or that with variance component(s) set to zero ($M_0$). Different approaches have been suggested for approximating the Bayes factor. We review a number of approaches for computing Bayes factor estimates that have been applied in other models and then explore their use for generalized linear mixed models in this and the subsequent chapters.

4.4.1 Approximating the marginal densities separately

One can separately approximate the marginal densities under the two models being compared using any approach discussed in Section 4.3 and then compute their ratio to get the Bayes factor estimate. If there are a number of models being compared, this approach will require the approximation of the marginal density under all the competing models and comparing their ratios for every possible pair of models; hence this approach may become cumbersome if there are too many models under consideration. Regarding coding efforts, once a program has been written to approximate the marginal density under the full model (with all the possible variance components), the program can be changed slightly to get a program for any reduced model.
4.4.2 Verdinelli-Wasserman method

Verdinelli and Wasserman (1995) suggest a method for computing Bayes factor estimates which is appropriate for comparing nested models and does not require approximation of the marginal densities for the two models separately. Let $\omega = (\delta', \psi')'$ be the parameter vector, $M_0$ be the null model with the restriction $\delta = \delta_0$ and $M_1$ be the unrestricted (alternative) model. Further suppose that $p_0(\psi)$ is the prior distribution of $\psi$ under the null model and $p(\psi, \delta)$ is the joint prior distribution of $\psi$ and $\delta$ under the unrestricted model. Then, the Bayes factor $BF_{01}$ for comparing the null model against the alternative model can be found as:

$$BF_{01} = \frac{p(y | M_0)}{p(y | M_1)} = \frac{1}{p(y | M_1)} \int p(y | \delta_0, \psi)p_0(\psi)d\psi$$

$$= \int \frac{1}{p(y | M_1)} p(y | \delta_0, \psi)p_0(\psi)d\psi$$

$$= p(\delta_0 | y) \int \frac{p(y | \delta_0, \psi)p_0(\psi)}{p(\delta_0 | y)p(y | M_1)} d\psi$$

$$= p(\delta_0 | y) \int \frac{p_0(\psi)}{p(\delta_0, \psi)}p(\psi | \delta_0, y)d\psi \quad \text{as} \quad p(\delta_0, \psi | y) = \frac{p(y | \delta_0, \psi)p(\delta_0, \psi)}{p(y | M_1)}$$

$$= p(\delta_0 | y) E \left[ \frac{p_0(\psi)}{p(\psi, \delta_0)} \right],$$

where the expectation is taken with respect to $p(\psi | \delta_0, y)$ and

$$p(\delta | y) = \int p(\delta, \psi | y)d\psi.$$

If $p(\psi | \delta_0) = p_0(\psi)$, then the Bayes factor can be found as the ratio

$$BF_{01} = \frac{p(\delta_0 | y)}{p(\delta_0)}, \quad (4.20)$$

which is known as Savage's density ratio.

In our case, $\delta$ is the whole or part of the vector of the variance components and $\delta_0$
is usually 0. Also, since it is common to assume apriori independence of the variance components and the regression parameters for most generalized linear mixed model applications, we can often use (4.20), which is much simpler.

The estimation of Savage's density ratio involves the estimation of \( p(\delta_0|y) \), the estimated posterior density \( p(\delta|y) \) at \( \delta = \delta_0 \). This can be done using a sample from the posterior distribution of \( \omega \). One looks at the sampled \( \delta \) only and can use kernel density estimation techniques to estimate \( p(\delta_0|y) \).

When the dimension of \( \delta \) is small, this method is accurate and very easy to implement. However, for high dimensional \( \delta \), kernel density estimation gives unreliable results and hence the estimated Bayes factor becomes quite unreliable and depends a lot on nuisance parameters like bandwidth for the kernel density estimation.

One requirement in order to use this formula for generalized linear mixed models is that since the computation of the Bayes factor involves \( p(\delta_0) \), if \( \delta_0 = 0 \), we have to use a prior distribution which is non-zero and has a finite value at the point 0 for the variance components we are testing.

### 4.4.3 Reversible jump MCMC

A very different approach for computing Bayes factor estimates requires constructing an “extended” model in which the model index is a parameter as well. A typical point in the parameter space of this “extended” model is \( (j, \omega_j) \), where \( j \) is the model index and \( \omega_j \) is the \( n_j \)-dimensional parameter vector for model \( j, j = 1, 2, \ldots J \). The reversible jump MCMC method suggested by Green (1995) can be used to sample from the expanded posterior distribution. This method generates a Markov chain that can jump between models with parameter spaces of different dimensions. Let \( \pi_j \) be the prior probability on model \( j, j = 1, 2, \ldots J \). Then the steps in the reversible jump algorithm are as follows:
1. Let the current state of the chain be \((j, \omega_j)\).

2. In an attempt to jump to another model, propose a new model \(j'\) with probability \(h(j, j')\), where \(h(j, j')\) is a probability mass function, i.e., \(\sum_{j'} h(j, j') = 1\). It is common to assume the \(h(j, j')\)'s to be equal for all \(j'\), which means that the algorithm is likely to jump from one particular model to any other model with equal chance. However, one can include any prior belief about the models in setting the \(h(j, j')\)'s. Note here that the proposed model \(j'\) may be the same as the current model \(j\).

3. a. If \(j' = j\), then perform an MCMC iteration (Gibbs or Metropolis) for the parameter \(\omega_j\) of model \(j\). Go to step 1.

b. If \(j' \neq j\), then \(\omega_j\) and \(\omega_{j'}\) may have different dimensions and worse, they may have components which are not related in any way. The solution is to do 'dimension-matching' — generate an auxiliary random variable \(u\) from a proposal density \(q(u|\omega_j, j, j')\) and set \((\omega_{j'}, u') = g_{j, j'}(\omega_j, u)\), where \(g\) is a one-to-one onto deterministic function. This takes care of the “dimension-matching” across models because one takes \(n_j + \dim(u) = n_{j'} + \dim(u')\). The choice of \(q(u|\omega_j, j, j')\), \(g\), \(u\) and \(u'\) depends on the problem at hand.

4. Accept the move from \(j\) to \(j'\) with probability

\[
\min\left\{1, \frac{p(y|\omega_{j'}, M = j') p(\omega_{j'}|M = j') \pi_j h(j', j) q(u'|\omega_{j'}, j', j) \left| \frac{\partial g(\omega_j, u)}{\partial (\omega_j, u)} \right|}{p(y|\omega_j, M = j) p(\omega_j|M = j) \pi_j h(j, j') q(u|\omega_j, j, j')} \right\}.
\]

It can be proved that if the above Markov chain has run long enough, \(N_j\), the number of times the Markov chain reaches a particular model \(j\), is approximately proportional to the posterior probability of the model, i.e.,

\[
\frac{p(M_j|y)}{p(M_{j'}|y)} \approx \frac{N_j}{N_{j'}}.
\]
Hence, once a sequence of simulations from the posterior distribution is generated, the Bayes factor for comparing model \( j \) to model \( j' \), \( BF^{jj'} = \frac{p(M_j | Y)}{p(M_{j'} | Y)} / p(M_j) \), is approximated as

\[
BF^{jj'} \approx \frac{N_j}{N_{j'}} \cdot \frac{\pi_j}{\pi_{j'}}
\]

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

If the marginal density of the data is very low under any model compared to the others, one may obtain \( N_j = 0 \) for that model making it impossible to compare that model with any other model under consideration. To prevent this from happening, that particular model may be assigned a very high \( \pi_j \) so that one may obtain a positive \( N_j \) for that model (see, for example, Han and Carlin, 2000).

### 4.4.3.1 A simple example showing the steps of a reversible jump MCMC algorithm

Consider a generalized linear mixed model with one variance component. Let \( 0 \) denote the model without the variance component and \( 1 \) denote the model with the variance component. Then \( \omega_0 \) is the vector of fixed effects while \( \omega_1 \) contains the fixed effects and the variance component. The steps for running a reversible jump MCMC algorithm in this setting may be as follows:

1. Pick a current model, e.g., model 0.

2. Propose another model with probability \( h(0, j') \), where, we assume that \( h(0, 0) = h(0, 1) = 0.5 \). So we toss an unbiased coin and pick model 0 as the proposed model if head appears and pick model 1 otherwise. Assume that the proposed model is model 1.

3. The dimensions of \( \omega_0 \) and \( \omega_1 \) are different — so one has to do dimension-matching here. Suppose we choose \( q(\omega | \omega_0, 0, 1) \) to be an inverse gamma density with suitable
parameters and define $\omega_1 = (\omega_0, u)$, i.e., $u'$ is taken to be a null vector and $g$ is taken as the identity function. The quantity $u$ plays the role of the variance component for model 1.

4. We assume $h(1, 0) = h(1, 1) = 0.5$. Also suppose $\pi_0 = \pi_1 = 0.5$. Then one accepts the proposed move from 0 to 1 with probability

$$
\min \left\{ 1, \frac{p(y|\omega_1, M = 1)p(\omega_1|M = 1)\pi_1 h(1, 0)}{p(y|\omega_0, M = 0)p(\omega_0|M = 0)\pi_0 h(0, 1)q(u|\omega_0, 0, 1)} \right\}.
$$

4.5 Sensitivity to the prior distribution of the parameters

Bayes factors are criticized a lot because they may be highly sensitive to the prior distributions used for the parameters in the two models being compared. We saw towards the end of the previous chapter that the computation of Bayes factor requires that one has proper prior distributions for the parameters regarding which one is testing any hypothesis. Even the use of a proper prior distribution for those parameters is not enough because changing the prior distribution may cause the Bayes factor to change a lot. Hence, one has to be very careful about the choice of the prior distributions for the parameters. This may require some in-depth study and discussion with the experts in the field about the problem at hand to find out the right prior information and then to translate it to a prior distribution for the parameters.

Since we are interested in testing for the variance components only, the fixed effects vector $\alpha$ in the model is a nuisance parameter and the prior distribution on $\alpha$ has little effect on the Bayes factor for testing for the variance component (see, e.g., Kass and Raftery, 1995). But the prior distribution(s) on the variance component(s) may have significant effect on the Bayes factor of our interest and one has to check how they might affect the Bayes factor.

We develop an approach for studying the sensitivity of the Bayes factor (for variance component testing in generalized linear mixed models) to the prior distribution for the
variance parameters. For simplicity, we describe the approach here in the context of studying the sensitivity of the Bayes factor comparing the full generalized linear mixed model (with fixed effects $\alpha$ and variance parameters $\theta$) against the reduced model obtained by restricting $\theta = 0$ (i.e., the corresponding generalized linear model with fixed effects vector $\alpha$). The approach can be easily generalized to the situation where the reduced model is obtained by restricting only a part of $\theta$ to a constant.

Let's define

$$BF_{\theta}^{10} = \frac{\int_{\alpha} \int_{b} p(y | \alpha, b) p(b | \theta) p(\alpha) db d\alpha}{\int_{\alpha} p(y | \alpha, b = 0) p(\alpha) d\alpha}.$$  

(4.21)

Then $BF_{\theta}^{10}$ is the Bayes factor comparing the model 1 with the variance components fixed at a particular value $\theta$ against the model 0 with no variance component. Note that the computation of $BF_{\theta}^{10}$ does not require a prior distribution on $\theta$. However, $BF_{\theta}^{10}$ can be interpreted as the Bayes factor in favor of the alternative model against the null model using a point mass prior with all its mass at $\theta$. Also, note that the numerator in (4.21) is the restricted likelihood at $\theta$ under the alternative model if one assumes $p(\alpha)$ to be flat; also it will be very close to the restricted likelihood at $\theta$ for any prior distribution on $\alpha$ that is not very sharp if the sample size is even moderately large (because the prior distribution on $\alpha$ will then have very little effect on the quantity). Hence, as a function of $\theta$, $BF_{\theta}^{10}$ will reach its maximum value near the restricted maximum likelihood estimate (details about which can be found in Sections 3.2.1.2 and 4.2) of $\theta$.

The Bayes factor $BF_{p(\theta)}^{10}$ for comparing the alternative model (model 1) against the null model (model 0), where we assume a prior distribution $p(\theta)$ on the variance components under the alternative model, can then be expressed as

$$BF_{p(\theta)}^{10} = \int_{\theta} \frac{\int_{\alpha} \int_{b} p(y | \alpha, b) p(b | \theta) p(\alpha) p(\theta) db d\alpha d\theta}{\int_{\alpha} p(y | \alpha, b = 0) p(\alpha) d\alpha} \int_{\theta} p(y | \alpha, b = 0) p(\alpha) d\alpha p(\theta) d\theta$$

$$= \int_{\theta} p(y | \alpha, b = 0) p(\alpha) d\alpha p(\theta) d\theta.$$
The Bayes factor that we have focussed on, $BF_{\theta(\theta)}^{01}$, that for comparing the null model against the alternative model, then is just the reciprocal

$$BF_{\theta(\theta)}^{01} = \left[ \int_{\theta} BF_{\theta}^{10} p(\theta) d\theta \right]^{-1}.$$  \hspace{1cm} (4.22)

The above identity shows that we can assess the sensitivity of the Bayes factor $BF_{\theta(\theta)}^{01}$ by examining $BF_{\theta}^{10}$ as a function of $\theta$. We will talk more about that while discussing the examples in Chapters 5 and 6. One important thing to notice from (4.22) is that the lowest possible value of $BF_{\theta(\theta)}^{01}$ for a data set (which means the most amount of evidence against the null hypothesis for the data set) can be obtained by using a prior distribution with all its mass at the value of $\theta$ that maximizes $BF_{\theta}^{10}$. And as we saw earlier, this will usually be close to the REML estimate of $\theta$ under the alternative model.
5 EXAMPLE: A NATURAL SELECTION STUDY

In this chapter, we apply the techniques discussed in previous chapters to a generalized linear mixed model with a single variance component applied to data from a natural selection study.

5.1 Description of the data set

A study of survival among turtles provides an example where a generalized linear mixed model is appropriate. In the study, 244 turtle eggs of the same age from 31 clutches (or families) were removed from their nests in a site in Illinois on the bank of the Mississippi river and taken to the laboratory where they were incubated and hatched (Janzen et. al., 2000). After a few days, the baby turtles were released from the same place where the eggs were found to recreate the natural process of emergence of baby turtles from their nests. When the turtles were released, they tried to travel downslope to the river — but many of them died while trying to do so, the main cause of death being the attack of predatory birds. Drift fences were set up at the river to identify the turtles that made it to the water — they were marked as ‘survived’. Five days after their release, turtles not identified as survived were assumed dead. The birth-weight of each turtle was collected as a covariate. The scientific objectives are to assess the effect of birth-weight on survival and to determine whether there is any clutch effect on survival. Figure 5.1 shows a scatter-plot of the birth-weights versus clutch number with survival status indicated by the plotting character ‘0’ if the animal survived and
Figure 5.1 Scatter plot with the clutches sorted by average birth-weight

'x' if the animal died. The clutches are numbered according to the increasing order of average birth-weight of the turtles in them — e.g., clutch 1 is the one with the smallest average birth-weight among all 31 clutches. The plot suggests that the heaviest turtles tend to survive and the lightest ones tend to die. It also suggests some variability in survival rates across clutches. For example, in clutch 4 (with average birth-weight 5.41), only 2 turtles out of 12 survived while in clutch 26 (with average birth-weight 7.50), 9 turtles out of 11 survived. Table 5.1 gives the average birth-weights, total number of turtles, the number of turtles survived and the survival percentage for all the clutches (which are sorted according to the increasing order of average birth-weight of the turtles
in that clutch).

Let \( y_{ij} \) denote the response (survival status with 1 denoting survival) and \( x_{ij} \) the birth-weight of the \( j \)-th turtle in the \( i \)-th clutch, \( i = 1, 2 \ldots m = 31, \ j = 1, 2 \ldots n_i \). The model we fit to the data is:

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

The \( b_i \)'s are random effects for clutch (family). Here, the distribution of the clutch effects is assumed to be the same for all birth-weights.

Note that other models are possible, e.g., random slopes in addition to or instead of random intercepts. We restrict our attention to this single model to study the issues related with evaluating Bayes factor estimates.

### 5.2 Estimation

The parameters of the full model here are the fixed effects vector \( \alpha \) and the variance parameter \( \sigma^2 \). The marginal likelihood \( L(\alpha, \sigma^2 | y) \) of the parameters after integrating out the random effects \( b_i \)'s is given by:

\[
L(\alpha, \sigma^2 | y) = \int p(y | \alpha, b) p(b | \sigma^2) db
= \int \left\{ \prod_{i=1}^{m} \left( \prod_{j=1}^{n_i} p_{ij}^{y_{ij}} (1 - p_{ij})^{1-y_{ij}} \right) \right\} \frac{1}{(2\pi\sigma^2)^{m/2}} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^{m} \frac{b_i^2}{\sigma^2}} db,
\]

where \( p_{ij} = \Phi(\alpha_0 + \alpha_1 x_{ij} + b_i) \).

To compute the marginal likelihood for any \( (\alpha, \sigma^2) \), the random effects were inte-
Table 5.1 The turtle data summarized by the clutches

<table>
<thead>
<tr>
<th>Clutch no.</th>
<th>Average birth-weight</th>
<th>Number of turtles</th>
<th>Number survived</th>
<th>Survival percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4.65</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>4.87</td>
<td>8</td>
<td>1</td>
<td>13</td>
</tr>
<tr>
<td>3</td>
<td>5.12</td>
<td>5</td>
<td>2</td>
<td>40</td>
</tr>
<tr>
<td>4</td>
<td>5.41</td>
<td>12</td>
<td>2</td>
<td>17</td>
</tr>
<tr>
<td>5</td>
<td>5.55</td>
<td>9</td>
<td>3</td>
<td>33</td>
</tr>
<tr>
<td>6</td>
<td>5.58</td>
<td>10</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>7</td>
<td>5.70</td>
<td>2</td>
<td>2</td>
<td>100</td>
</tr>
<tr>
<td>8</td>
<td>5.75</td>
<td>4</td>
<td>1</td>
<td>25</td>
</tr>
<tr>
<td>9</td>
<td>5.76</td>
<td>12</td>
<td>4</td>
<td>33</td>
</tr>
<tr>
<td>10</td>
<td>5.82</td>
<td>11</td>
<td>6</td>
<td>55</td>
</tr>
<tr>
<td>11</td>
<td>5.83</td>
<td>8</td>
<td>2</td>
<td>25</td>
</tr>
<tr>
<td>12</td>
<td>5.85</td>
<td>14</td>
<td>3</td>
<td>21</td>
</tr>
<tr>
<td>13</td>
<td>5.93</td>
<td>8</td>
<td>2</td>
<td>25</td>
</tr>
<tr>
<td>14</td>
<td>6.07</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>6.20</td>
<td>18</td>
<td>2</td>
<td>11</td>
</tr>
<tr>
<td>16</td>
<td>6.21</td>
<td>3</td>
<td>1</td>
<td>33</td>
</tr>
<tr>
<td>17</td>
<td>6.22</td>
<td>8</td>
<td>2</td>
<td>25</td>
</tr>
<tr>
<td>18</td>
<td>6.32</td>
<td>9</td>
<td>3</td>
<td>33</td>
</tr>
<tr>
<td>19</td>
<td>6.35</td>
<td>6</td>
<td>4</td>
<td>67</td>
</tr>
<tr>
<td>20</td>
<td>6.38</td>
<td>14</td>
<td>4</td>
<td>29</td>
</tr>
<tr>
<td>21</td>
<td>6.50</td>
<td>10</td>
<td>2</td>
<td>20</td>
</tr>
<tr>
<td>22</td>
<td>6.89</td>
<td>2</td>
<td>2</td>
<td>100</td>
</tr>
<tr>
<td>23</td>
<td>7.13</td>
<td>12</td>
<td>5</td>
<td>42</td>
</tr>
<tr>
<td>24</td>
<td>7.28</td>
<td>8</td>
<td>3</td>
<td>38</td>
</tr>
<tr>
<td>25</td>
<td>7.46</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>26</td>
<td>7.50</td>
<td>11</td>
<td>9</td>
<td>82</td>
</tr>
<tr>
<td>27</td>
<td>7.51</td>
<td>8</td>
<td>5</td>
<td>63</td>
</tr>
<tr>
<td>28</td>
<td>7.57</td>
<td>7</td>
<td>4</td>
<td>57</td>
</tr>
<tr>
<td>29</td>
<td>7.61</td>
<td>9</td>
<td>6</td>
<td>67</td>
</tr>
<tr>
<td>30</td>
<td>7.67</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>31</td>
<td>7.74</td>
<td>8</td>
<td>4</td>
<td>50</td>
</tr>
</tbody>
</table>
grated out using Simpson's rule (Section 2.2.1). This is easy since

\[
L(\alpha, \sigma^2|\mathbf{y}) = \int_b \left\{ \prod_{i=1}^m \prod_{j=1}^{n_i} p_{ij}^{y_{ij}} (1 - p_{ij})^{1-y_{ij}} \right\} \frac{1}{(2\pi \sigma^2)^{m/2}} e^{-\sum_{i=1}^m \frac{b_i^2}{2\sigma^2}} \, db
\]

\[
= \frac{1}{(2\pi \sigma^2)^{m/2}} \prod_{i=1}^m \left\{ \int_b \left( \prod_{j=1}^{n_i} p_{ij}^{y_{ij}} (1 - p_{ij})^{1-y_{ij}} \right) e^{-\frac{b_i^2}{2\sigma^2}} \, db \right\}
\]

for \( p_{ij} = \Phi(\alpha_0 + \alpha_1 x_{ij} + b_i) \), which means that to compute the marginal likelihood, we have to compute a bunch of one-dimensional integrals only (with respect to the \( b_i \)'s) and Simpson's rule (Section 2.2.1) is quite satisfactory in approximating them.

### 5.2.1 Likelihood estimation

The marginal maximum likelihood estimate is found using the Newton-Raphson algorithm, where, the marginal likelihood \( L(\alpha, \sigma^2|\mathbf{y}) \) for any \((\alpha, \sigma^2)\) is calculated by numerically integrating out the random effects using Simpson's rule (Section 2.2.1). The initial value of \( \alpha \), found by fitting the simple probit regression model (without any random effects) to the data, is \((-2.82, 0.39)\). Convergence was very quick for any initial value of \( \sigma^2 \) which is positive and less than 0.5. The marginal maximum likelihood estimates of the parameters are given in Table 5.2.

<table>
<thead>
<tr>
<th>parameter</th>
<th>estimate</th>
<th>sd</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \alpha_0 )</td>
<td>-2.85</td>
<td>0.61</td>
</tr>
<tr>
<td>( \alpha_1 )</td>
<td>0.39</td>
<td>0.09</td>
</tr>
<tr>
<td>( \sigma^2 )</td>
<td>0.09</td>
<td>0.08</td>
</tr>
</tbody>
</table>

### 5.2.2 Bayesian estimation

To carry out a Bayesian analysis, we have to choose a prior distribution. Since very few studies like this have been carried out (personal communication with Dr. Frederic
Janzen), we don’t have much prior information at all about the possible values of the parameters. So we just pick a proper vague prior distribution (bivariate normal with mean 0 and variance 20.1) on $\alpha$ and an inverse gamma$(2.5,1.5)$ prior distribution on $\sigma^2$. The mean of the inverse gamma prior distribution is 1 and the variance is 2. A plot of the pdf of this distribution can be found in Figure 5.6 We carry out a sensitivity analysis later to examine how the prior distribution for $\sigma^2$ affects the Bayes factor estimate, the quantity of our interest.

The joint posterior distribution of $\alpha$ and $\sigma^2$ is proportional to the product of the likelihood function and the prior density of the parameters, i.e., proportional to

$$
\left\{ \int \left[ \prod_{i=1}^m \prod_{j=1}^{n_i} p_{ij}^{y_{ij} (1 - y_{ij})} \right] \frac{1}{(\sigma^2)^{m/2}} e^{\left[-\sum_{i=1}^m \frac{x_{ij}^2}{2\sigma^2}\right]} \right\} p(\alpha)p(\sigma^2),
$$

where $p_{ij} = \Phi(\alpha_0 + \alpha_1 x_{ij} + b_i)$.

Alternatively, one can work with the joint posterior distribution of $\alpha$, $b$ and $\sigma^2$, which is given by:

$$
p(\alpha, b, \sigma^2|y) = \frac{p(y|\alpha, b)p(b|\sigma^2)p(\alpha, \sigma^2)}{\int p(y|\alpha, b)p(b|\sigma^2)p(\alpha, \sigma^2) \, d\alpha \, db \, d\sigma^2}
$$

$$
\propto p(y|\alpha, b)p(b|\sigma^2)p(\alpha, \sigma^2)
$$

$$
= \left[ \prod_{i=1}^m \prod_{j=1}^{n_i} p_{ij}^{y_{ij} (1 - y_{ij})} \right] \frac{1}{(\sigma^2)^{m/2}} e^{\left[-\sum_{i=1}^m \frac{x_{ij}^2}{2\sigma^2}\right]} p(\alpha)p(\sigma^2).
$$

where $p_{ij} = \Phi(\alpha_0 + \alpha_1 x_{ij} + b_i)$. We use the latter posterior distribution to avoid integrating out the random effects. A Markov chain Monte Carlo (Gibbs sampling with metropolis steps for $\alpha$ and $b$) algorithm is used to generate a posterior sample of size 5000 after convergence of the chain is detected. Summaries of the posterior distributions for $\alpha_0$, $\alpha_1$ and $\sigma^2$ are given in Table 5.3. Figure 5.2 shows histograms of the sampled values of the parameters in the model.
Figure 5.2 Histograms of the sampled values of the parameters
Table 5.3 Summaries of the posterior distributions of the parameters of a probit regression model with a random intercept applied to the turtles data set

<table>
<thead>
<tr>
<th>parameter</th>
<th>mean</th>
<th>sd</th>
<th>2.5 %</th>
<th>50 %</th>
<th>97.5 %</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_0$</td>
<td>-2.83</td>
<td>0.72</td>
<td>-4.44</td>
<td>-2.80</td>
<td>-1.56</td>
</tr>
<tr>
<td>$\alpha_1$</td>
<td>0.39</td>
<td>0.11</td>
<td>0.19</td>
<td>0.38</td>
<td>0.63</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>0.32</td>
<td>0.12</td>
<td>0.16</td>
<td>0.30</td>
<td>0.61</td>
</tr>
</tbody>
</table>

5.2.3 Discussion about the results of estimation

The posterior means for $\alpha_0$ and $\alpha_1$ are very close to the corresponding maximum likelihood estimates. The difference in the posterior mean (0.32) and the maximum likelihood estimate (0.09) for $\sigma^2$ is a reflection of two factors. One, the maximum likelihood estimate is comparable to the posterior mode rather than the posterior mean. Two, the posterior inference is sensitive to the form of the prior distribution. With respect to the first factor, note that the posterior mode of $\sigma^2$ is 0.25, which is considerably less than the posterior mean 0.32. This is because the (marginal) posterior distribution of $\sigma^2$ is a positively skewed distribution as seen from the histogram for $\sigma^2$ in Figure 5.2. The histogram indicates that the posterior mode is less than the posterior median, which in turn is less than the posterior mean. Of course, there is still a large difference between the posterior mode and the maximum likelihood estimate. This reflects sensitivity to the prior distribution. This issue is revisited with respect to Bayes factor estimates in Section 5.5.

Looking at the parameter estimates, we conclude that birth-weight is a significant predictor for this example because the maximum likelihood estimate of $\alpha_1$ is more than four times its standard deviations away from 0 and a 95% posterior interval for $\alpha_1$ does not contain 0. Further, the estimate of the coefficient $\alpha_1$ being positive, birth-weight has a positive effect on the chance of survival of a turtle; we noticed this phenomenon earlier while looking at Figure 5.1.
The clutch-effect seems to be small, but it is difficult to say if it is statistically significant. Incidentally, the conclusion about birth-weight does not change if we remove the clutch effect from the model; in that case, the maximum likelihood estimates of \( \alpha_0 \) and \( \alpha_1 \) are -2.82 and 0.39 with standard errors 0.55 and 0.08 respectively.

5.3 Hypothesis testing concerning \( \sigma^2 \)

5.3.1 Likelihood ratio test

To compute the likelihood ratio test statistic for testing the hypothesis \( H_0 : \sigma^2 = 0 \), we compute the likelihoods of the models with the variance component and without it at the maximum likelihood estimates of the respective models. The values of the log-likelihood are -148.83 under the model without the variance component and -147.64 under the model with the variance component. So the value of the usual likelihood ratio test statistic \(-2 \log \frac{L_0}{L_1}\) (where \( L_0 \) is the likelihood under the null model and \( L_1 \) is the likelihood under the alternative model, both likelihoods computed at the maximum likelihoods under the respective models) is 2.38. Since the usual asymptotic \( \chi^2 \) reference distribution does not apply (as described in Section 3.3.1.1), we obtain the reference distribution by simulation as follows:

- generate 1000 simulated sets of responses \( y_{ij}, \ i = 1, 2 \ldots m = 31, \ j = 1, 2 \ldots n_i \) under the probit model with \( \alpha \) equal to the maximum likelihood estimates from Section 5.2 and \( \sigma^2 = 0 \)

- compute the value of the likelihood ratio test statistic for each generated dataset

A histogram of these 1000 values of the likelihood ratio test statistic provides us with an approximation to the distribution of the test statistic under the null hypothesis. Figure 5.3 shows the histogram.
Using this reference distribution, the p-value corresponding to the likelihood ratio test statistic is 0.031 — hence there is some evidence against the null hypothesis of zero clutch variance.

We can also compute a $100(1-\alpha)$ % confidence interval for any $\alpha \in (0,1)$ by simulation utilizing the fact that if $H_0: \sigma^2 = \sigma^2_0$ is not rejected at level $\alpha$, then $\sigma^2_0$ is in the $100(1-\alpha)$ % confidence interval for $\sigma^2$. The steps that have to be iterated (with different fixed values of $\sigma^2_0$) for computing the confidence interval are as follows:

(i) take a fixed value of $\sigma^2_0$
(ii) compute the value of the likelihood ratio test statistic for testing $H_0: \sigma^2 = \sigma_*^2$ from the data set.

(iii) generate 1000 simulated data sets with $\alpha = \hat{\alpha}_{MLE}$, $\sigma^2 = \sigma_*^2$ and compute the likelihood ratio test statistic for testing $H_0: \sigma^2 = \sigma_*^2$ for each of those simulated data sets.

(iv) if the likelihood ratio test statistic computed in the second step is less (greater) than the 100(1-\alpha)-th percentile of the set of 1000 likelihood ratio test statistics computed in the third step, then $\sigma_*^2$ is (isn't) in the 100(1-\alpha) % confidence interval for $\sigma^2$.

By following the procedure mentioned above and using a trial and error method, we obtain a 95 % confidence interval for $\sigma^2$ as (0.00, 0.24).

5.3.2 Lin’s score test

We carry out the Lin’s test as well for testing $H_0: \sigma^2 = 0$ using a simplified formula for the score statistic in Lin (1997) when only a random intercept is specified in the model. The value of the global score statistic $\chi^2_G$ with 1 degree of freedom is 4.24 with a corresponding p-value of 0.04. So Lin’s test supports the results of the likelihood ratio test. The data set provides evidence against the null hypothesis — the variance component is significant at 5% level and hence is required in the model.

5.4 Computing estimates of the Bayes factor

The Bayes factor for comparing the null model $M_0$ (that without variance components) against the alternative model $M_1$ (that with the variance component) can be expressed as

$$BF_{01} = \frac{p(y|M_0)}{p(y|M_1)}.$$
where

\[ p(y|M_0) = \int p(y|\alpha, b = 0)p(\alpha)d\alpha, \]
\[ p(y|M_1) = \int p(y|\alpha, b)p(b|\sigma^2)p(\alpha)p(\sigma^2)dbd\alpha d\sigma^2 \]

We compute the above-mentioned Bayes factor estimate using the methods discussed in Chapter 4.

Fortunately, the true value of the Bayes factor can be obtained by numerical integration although the program takes about 37 hours of CPU time to run on an Alphastation 500 workstation equipped with 400MHz 64-bit CPU and a gigabyte of RAM; the true value of the Bayes factor up to 2 decimal places is 3.25. Hence we can compare the performance of the different methods by comparing the Bayes factor estimates obtained by each method against the correct value. For each simulation-based method, we run the program to compute the Bayes factor estimate 30 times with different random number seeds and take the average of the 30 Bayes factor estimates; we also obtain an estimated standard deviation to be attached to the Bayes factor estimate by computing the standard deviation of the 30 Bayes factors obtained. The standard deviation will give us an idea about how stable a method is. The results obtained by the various methods are provided in Table 5.4.

Also given in the table are the CPU times taken for one computation of Bayes factor estimate by each of the methods under consideration in an Alphastation 500 workstation equipped with 400MHz 64-bit CPU and a gigabyte of RAM.

Where required, the marginal likelihood \( p(y|\alpha, \sigma^2) \) for any \((\alpha, \sigma^2)\) is calculated by numerically integrating out the random effects using Simpson's rule. Where a posterior sample under the full model (that with the variance component) is required in any of these methods (harmonic estimator, bridge sampling), we obtain a sample in the same way as discussed in Section 5.2.2. We run another Markov chain to obtain a sample under the model without the variance component, when required. All simulation-
Table 5.4 Estimates of the Bayes factor (along with their standard deviations and time taken to run the program) for comparing a simple probit regression model against a probit regression model with a random intercept obtained by different methods for the turtles data set

<table>
<thead>
<tr>
<th>Method</th>
<th>Bayes factor estimate</th>
<th>std. dev.</th>
<th>CPU time (min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Laplace's</td>
<td>3.76</td>
<td>-</td>
<td>0.1</td>
</tr>
<tr>
<td>Laplace's at posterior mean</td>
<td>3.29</td>
<td>0.08</td>
<td>0.2</td>
</tr>
<tr>
<td>Importance sampling</td>
<td>3.26</td>
<td>0.03</td>
<td>4.4</td>
</tr>
<tr>
<td>Harmonic estimator</td>
<td>6.07</td>
<td>3.27</td>
<td>5.3</td>
</tr>
<tr>
<td>Bridge sampling</td>
<td>3.25</td>
<td>0.03</td>
<td>10.0</td>
</tr>
<tr>
<td>Chib’s</td>
<td>3.25</td>
<td>0.04</td>
<td>3.4</td>
</tr>
<tr>
<td>RJ MCMC</td>
<td>3.28</td>
<td>0.26</td>
<td>7.6</td>
</tr>
</tbody>
</table>

Based estimates are based on a posterior sample of size 5000 (5000 after convergence is diagnosed for the MCMC methods).

5.4.1 Specific details about the different methods

We provide specific details regarding the implementation of the various methods for computing Bayes factor estimates for this problem.

Laplace approximation and its variation: Since the Laplace approximation assumes normality of the posterior distribution, we work on the logarithmic scale for the variance component to have better accuracy. The posterior modes for the Laplace approximation are computed using the Newton-Raphson algorithm with numerical approximation to the required derivatives. The posterior mode under the null model is (-2.74, 0.38) and that under the alternative model is (-2.75, 0.38, -1.39), where the third component is the logarithm of $\sigma^2$ (exponentiating this yields a value of 0.25 for $\sigma^2$). The same Newton-Raphson programs also provide us with the negative of the inverse of the Hessian matrix of the log-posterior evaluated at the posterior mode under the two models with no extra effort required. Note that for the example, the Bayes factor estimate is close to but not equal to the actual value. There is no standard deviation
attached to the estimated Bayes factor because there is no simulation involved in its computation.

We also compute the Bayes factor estimate using the variation of the Laplace approximation suggested by (4.8) at the posterior means. The average (averaged over the 30 Bayes factor estimate computations) posterior mean under the alternative model is (-2.83,0.39,-1.14), where the third component is the logarithm of $\sigma^2$ (exponentiating this yields a value of 0.32 for $\sigma^2$). The average posterior mean (averaged over the 30 Bayes factor estimate computations) under the null model is (-2.76,0.38). We compute the derivatives of the log-posterior at the posterior means numerically. The Bayes factor estimate obtained by this variation is very close to the true value on an average.

**Importance sampling:** Under both the models, the importance sampling density is taken as a $t_4$ distribution with location parameter equal to the posterior mode and scale matrix equal to the inverse of the negative of the Hessian of the log-posterior evaluated at the posterior mode. We use the logarithm of $\sigma^2$ as the parameter in the alternative model. For the turtles example, this method gives very good results — the average Bayes factor estimate obtained by this method is only 0.01 more than the true value and the standard deviation of the Bayes factor estimates obtained is the lowest among all the methods used.

**Bridge sampling:** In bridge sampling, we take $q$ to be the importance sampling density described above. We use the logarithm of $\sigma^2$ as the parameter in the alternative model just like with importance sampling. Since we knew the correct value of $p(y|M)$ by numerical method, we use that in computing the optimum $\gamma$. With this optimum choice of $\gamma$, the Bridge sampling method does very well for the turtles example — both in terms of the average and standard deviation of the Bayes factor estimates obtained.
Chib's method: In Chib's method, we introduce a latent variable as in Chib (1995) which makes the computations easier. Let's define independent latent variables $z$ such that

$$z_{ij} | \alpha, b_i \sim N(\alpha_0 + \alpha_1 x_{ij} + b_i), j = 1, 2, \ldots n_i, \ i = 1, 2 \ldots 31$$

along with

$$b_i | \sigma^2 \stackrel{iid}{\sim} N(0, \sigma^2), \ i = 1, 2, \ldots , 31.$$

If the observed responses $y_{ij}$'s are given by

$$y_{ij} = \begin{cases} 
1 & \text{if } z_{ij} \geq 0 \\
0 & \text{if } z_{ij} < 0,
\end{cases}$$

then

$$P(y_{ij} = 1) = P(z_{ij} > 0) = \Phi(\alpha_0 + \alpha_1 x_{ij} + b_i)$$

as in the probit model. The advantage of introducing the $z_{ij}$'s is that it allows us to use Gibbs sampling with conditional distributions that are easy to sample from. Let $Z_i$ denote the $n_i \times 1$ vector of $z_{ij}$ of the $i$-th group and let $X_i$ denote the $n_i \times 2$ matrix of covariates for the $i$-th group, where the $j$-th row of $X_i = (1, x_{ij}), j = 1, 2, \ldots n_i$. Suppose further that $V_i = I_{n_i} + J_{n_i} \sigma^2$, $J_{n_i}$ denotes an $n_i \times n_i$ matrix of 1's. The conditional distributions, assuming $N(0, V_0)$ prior distribution on $\alpha$ and inverse gamma $(g_1, g_2)$ prior distribution on $\sigma^2$, are:

$$\alpha | y, Z, \sigma^2 \sim N_2(\hat{\alpha}, V) \text{ for } \hat{\alpha} = V\left(\sum_{i=1}^{31} X_i'V_iZ_i\right), V = \left(V_0^{-1} + \sum_{i=1}^{31} X_i'V_iX_i\right)^{-1}$$

$$Z_i | y, \alpha, \sigma^2 \sim N_{n_i}(X_i \alpha, V_i) \text{ truncated to the region implied by the vector } y_i$$

$$b_i | y, Z, \alpha, \sigma^2 \sim N\left(\frac{n_i s_i \sigma^2}{1 + n_i \sigma^2}, (n_i + \frac{1}{\sigma^2})^{-1}\right); \ s_i = \sum_{i=1}^{n_i} (z_{ij} - \alpha_0 - \alpha_1 x_{ij})$$

$$\sigma^2 | b \sim \text{Inverse Gamma}(g_1 + \frac{m}{2}, g_2 + \frac{1}{2} \sum_{i=1}^{m} b_i^2)$$
The multivariate normal distribution for \( Z_i \) is truncated to the product space \( B_{i1} \times B_{i2} \times \ldots B_{in_i} \), where

\[
B_{ij} = \begin{cases} 
(0, \infty) & \text{if } y_{ij} = 1 \\
(-\infty, 0) & \text{if } y_{ij} = 0
\end{cases}
\]

As observed in Chib and Carlin (1999), integrating out the random effects from the conditional distribution of \( \alpha \) improves the rate of convergence of the Markov chain. The only difficulty in carrying out the Gibbs sampling approach here is generating the \( Z_i \)'s, for which we follow the algorithm suggested by Chib and Greenberg (1998) and sample this truncated multivariate normal distribution from a sequence of (full conditional) univariate truncated normal distributions.

We use the techniques described in Section 4.3.5.1 here to estimate the posterior ordinate at a fixed value of the parameter. Since we have to estimate \( p(\alpha^*, \sigma^2|y) \) whereas it is much easier to generate from \( p(\alpha, \sigma^2, b, Z|y) \), we treat \( b \) and \( Z \) as latent variables. We treat \( \alpha, \sigma^2, b \) and \( Z \) as four blocks in applying the extended parameter version of the technique in Section 4.3.5.1.

We use the same basic techniques and formulae for the model without the variance component as well, except that this time we put \( \sigma^2 = b_i = 0, i = 1, 2, \ldots m \). Then there are only two blocks in the extended parameter space, \( \alpha \) and \( Z \), and we have to estimate \( p(\alpha^*|y) \).

As discussed in Section 4.3.5.4, the fixed value is taken as the posterior mode for both the models.

The result obtained by the Chib's method for this example is very accurate; on an average, the Bayes factor estimates obtained by this method are exactly the same (up to 2 decimal places) as the true value and the standard deviation of the Bayes factor estimates obtained is very low.
Reversible jump MCMC: In the reversible jump algorithm, according to the notation from Section 4.4.3, we set \( h(0,0) = h(0,1) = h(1,0) = h(1,1) = 0.5 \) and \( \pi_0 = \pi_1 = 0.5 \). When we are in model 0 (without \( \sigma^2 \)) and are trying a jump to model 1, we apply the following steps:

- generate \( \sigma^2 \) from \( q(\sigma^2) \): Inverse gamma with mean \( \hat{\sigma}^2_{\text{MLE}} \) and variance \( \hat{V}(\hat{\sigma}^2_{\text{MLE}}) \)
- define \( \theta_1 = (\theta_0, \sigma^2) = (\alpha, \sigma^2) \); in terms of the notations of step 3b of Section 4.4.3, \( u = \sigma^2, u' = 0 \) and \( g(.) \) is the identity function
- acceptance prob: \( \min\{1, \frac{f(y|\alpha, \sigma^2, M=1)p(\sigma^2)\pi_1}{f(y|\alpha, M=0)p(\sigma^2)\pi_0}\} \)

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

- define \( \theta_0 = \alpha \) (i.e., set \( \sigma^2 = 0 \)); in terms of the notations of step 3b of Section 4.4.3, \( u = 0, u' = \sigma^2 \) and \( g(.) \) is the identity function
- acceptance prob: \( \min\{1, \frac{f(y|\alpha, M=0)p(\sigma^2)\pi_0}{f(y|\alpha, \sigma^2, M=1)p(\sigma^2)\pi_1}\} \)

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

For our example, this method does not perform very well compared to the other methods. On an average, the Bayes factor estimates obtained by this method are close to the true value, but the standard deviation of the estimate is much higher than that obtained by the other methods.

5.4.2 Summaries of the results obtained by the different methods

Looking at the table of the Bayes factor estimates obtained by the different methods, we see that the importance sampling method, Chib’s method and bridge sampling
method perform equally well and better than the other methods. The large standard error for the reversible jump MCMC method can be reduced by increasing the number of iterations. The instability of the harmonic estimator is evident. Laplace approximation, which makes the strong assumption of normality of the posterior distribution gives a fair approximation in the sense that the conclusion would be right. The Laplace approximation computed at the posterior mean exhibits smaller variation than the Laplace approximation and yields a value of the Bayes factor estimate that is very close to the true value.

A second factor to consider in comparing the computational methods is the amount of computational time required which can be found from Table 5.4. Among the three methods mentioned above that perform the best for this problem with respect to accuracy and precision, the bridge sampling method takes longer than the other two (importance sampling method and Chib's method). Because the importance sampling method requires much less amount of coding than the Chib's method and takes only a minute longer, this seems to be the most convenient method for this data set. In fact, this result together with our simulation results (to be discussed in Section 5.6) suggest that for small problems like this example, the importance sampling method is the most convenient method for computing Bayes factor estimates.

5.5 Results and sensitivity analysis

The value of the Bayes factor, 3.25, favors the null model, suggesting that there is no evidence of a clutch-effect - hence an ordinary probit regression model (with no random effects) seems to be sufficient for the given data set. One thing to notice here is that this is in conflict with the conclusion obtained from the frequentist tests, which suggested that the null model is inadequate for the data set. This may be an outcome of the fact that in both the simulation-based test that we use and in Lin's test, the value
of the fixed effects vector is treated as constant (at the maximum likelihood estimate under the null model) in obtaining the null distribution of the statistic — this may have underestimated the variability of the test statistics making the tests anticonservative.

Another possible explanation for the difference is the prior distribution used in obtaining the Bayes factor. As described earlier, the Bayes factor is sensitive to the prior distribution on $\sigma^2$. While the Bayes factor of 3.25 was obtained under a inverse-gamma prior distribution with mean 1 and variance 2, the value of the Bayes factor changed to 1.02 when we used a gamma distribution with the same mean and variance and to 1.61 when we used a shrinkage prior distribution \( p(\sigma^2) = c/(c + \sigma^2)^2 \) with median 1 (the moments are not defined for a shrinkage prior distribution).

To study the sensitivity of the Bayes factor to the prior distribution on $\sigma^2$ more deeply, we use our approach introduced in Section 4.5. We introduce the quantity $BF_{\sigma^2}^{10}$, which is the Bayes factor comparing the model where the variance component is fixed at $\sigma^2$ against the model without any variance component (i.e., $\sigma^2$ is fixed at 0). Also, $BF_{\sigma^2}^{10}$ can be interpreted as the Bayes factor in favor of the alternative using a degenerate or point mass prior distribution at $\sigma^2$ against the null model. Mathematically,

\[
BF_{\sigma^2}^{10} = \frac{\int_{\alpha} \int_{b} p(y|\alpha, b)p(b|\sigma^2)p(\alpha)dbd\alpha}{\int_{\alpha} p(y|\alpha, b = 0)p(\alpha)d\alpha}.
\]

The values of $BF_{\sigma^2}^{10}$ (obtained by the importance sampling method) for a grid of values of $\sigma^2$ are shown in Table 5.5. Figure 5.4 shows a plot of the $BF_{\sigma^2}^{10}$'s against $\sigma^2$ using the same grid of values of $\sigma^2$. The figure (and the table) shows that $BF_{\sigma^2}^{10}$ is 1 for $\sigma^2=0$; the two models are identical at that value. The estimated Bayes factor then increases with increase in $\sigma^2$ until it reaches its maximum value of about 4.55 at around $\sigma^2=0.09$. This is sensible because the maximum likelihood estimator of $\sigma^2$ is 0.086. The Bayes factor estimate provides the most support for the variance component model when the prior distribution on the variance component is most consistent with the data. Then it decreases rapidly till it gets close to zero for $\sigma^2$ around 0.75 and stays there. So when it
Table 5.5 Estimates of the Bayes factor for comparing a simple probit regression model against a probit regression model with a random intercept for different degenerate priors on the variance component for the turtles data set

<table>
<thead>
<tr>
<th>$\sigma^2$</th>
<th>$BF_{\sigma^2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>0.01</td>
<td>1.48</td>
</tr>
<tr>
<td>0.05</td>
<td>3.45</td>
</tr>
<tr>
<td>0.08</td>
<td>4.35</td>
</tr>
<tr>
<td>0.09</td>
<td>4.55</td>
</tr>
<tr>
<td>0.10</td>
<td>4.55</td>
</tr>
<tr>
<td>0.15</td>
<td>4.00</td>
</tr>
<tr>
<td>0.20</td>
<td>2.94</td>
</tr>
<tr>
<td>0.25</td>
<td>1.96</td>
</tr>
<tr>
<td>0.30</td>
<td>1.27</td>
</tr>
<tr>
<td>0.40</td>
<td>0.50</td>
</tr>
<tr>
<td>0.50</td>
<td>0.20</td>
</tr>
<tr>
<td>0.75</td>
<td>0.023</td>
</tr>
<tr>
<td>1.00</td>
<td>0.0036</td>
</tr>
<tr>
<td>2.00</td>
<td>$1.5 \times 10^{-5}$</td>
</tr>
<tr>
<td>3.00</td>
<td>$3.5 \times 10^{-7}$</td>
</tr>
<tr>
<td>4.00</td>
<td>$1.9 \times 10^{-8}$</td>
</tr>
<tr>
<td>5.00</td>
<td>$1.8 \times 10^{-9}$</td>
</tr>
</tbody>
</table>

comes to choosing between a small hypothesized value (by small here we mean less than 0.3) of $\sigma^2$ and $\sigma^2$ equal to zero, the Bayes factor estimate favors the small positive value of $\sigma^2$. However, when it comes to choosing between a large $\sigma^2$ and $\sigma^2$ equal to zero, the Bayes factor estimate favors the zero value of $\sigma^2$.

Let $BF_{p(.)}^{10}$ denote the Bayes factor in favor of the alternative model (with $\sigma^2$) when the prior distribution on $\sigma^2$ is $p(\sigma^2)$. It is easy to relate $BF_{p(.)}^{10}$ to $BF_{\sigma^2}^{10}$ as

$$BF_{p(.)}^{10} = \frac{\int_\sigma^2 \int_\alpha \int_\beta p(y|\alpha, \beta)p(b|\sigma^2)p(\sigma^2) db d\alpha d\sigma^2}{\int_\alpha p(y|\alpha, b = 0)p(\alpha) d\alpha}$$

$$= \frac{\int_\sigma^2 \int_\alpha \int_\beta p(y|\alpha, \beta)p(b|\sigma^2)p(\alpha) db d\alpha p(\sigma^2) d\sigma^2}{\int_\alpha p(y|\alpha, b = 0)p(\alpha) d\alpha}$$

$$= \int_\sigma^2 BF_{\sigma^2}^{10} p(\sigma^2) d\sigma^2.$$
Of course, the Bayes factor that we here focussed on, $BF_{p(\cdot)}^{10}$, is just the reciprocal

$$BF_{p(\cdot)}^{01} = \left[ \int_{\sigma^2} BF_{\sigma^2}^{10} p(\sigma^2) d\sigma^2 \right]^{-1}. \quad (5.1)$$

The above identity shows that we can assess the sensitivity of the Bayes factor estimate $BF_{p(\cdot)}^{10}$ by examining $BF_{\sigma^2}^{10}$ as a function of $\sigma^2$. In this example, $BF_{\sigma^2}^{10}$ varies a lot as $\sigma^2$ changes — so the Bayes factor estimate $BF_{p(\cdot)}^{01}$ will vary as $p(\sigma^2)$, the prior distribution on $\sigma^2$, changes.

In particular, we see that if we try different prior distributions on the variance parameter, the lowest possible value of $BF_{p(\cdot)}^{01}$ for this data set is approximately the reciprocal of 4.55, which is 0.22 and that Bayes factor estimate is obtained using a prior distribution which is degenerate with all its mass on $\sigma^2=0.09$. In practice, one never takes a degenerate prior distribution on a parameter. Hence we try a number of inverse gamma prior distributions on $\sigma^2$, all with variance 2, but with different means, and compute the Bayes factor estimate $BF_{p(\cdot)}^{01}$ for each of them using the importance sampling approach.

Table 5.6 contains the values of the prior means and the prior modes and the values of the Bayes factor estimate obtained and Figure 5.5 shows a plot of those values against
They show that the Bayes factor estimate is 1 at prior mean 0, but it decreases as the prior mean increases until it reaches a value of around 0.30 at prior mean about 0.15 and increases rapidly as the prior mean increases past that point. Figure 5.5 shows the sensitivity of the Bayes factor estimate to the prior distribution on the variance component. In this case, if the prior mean favors low values of the variance component, there is some evidence in favor of the random effects model (the same conclusion as likelihood ratio test and Lin's test). However, if the prior distribution favors values 0.5 or above, the null model is preferred. We repeat the analysis with other variance values.

Table 5.6 Estimates of the Bayes factor for comparing a simple probit regression model against a probit regression model with a random intercept for different prior means/modes (with an inverse gamma prior distribution) with prior variance fixed at 2 for the turtles data set

<table>
<thead>
<tr>
<th>Prior mean</th>
<th>Prior mode</th>
<th>$BF_{01}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>0.01</td>
<td>0.00</td>
<td>0.68</td>
</tr>
<tr>
<td>0.05</td>
<td>0.02</td>
<td>0.38</td>
</tr>
<tr>
<td>0.10</td>
<td>0.03</td>
<td>0.30</td>
</tr>
<tr>
<td>0.20</td>
<td>0.07</td>
<td>0.30</td>
</tr>
<tr>
<td>0.30</td>
<td>0.10</td>
<td>0.37</td>
</tr>
<tr>
<td>0.50</td>
<td>0.18</td>
<td>0.61</td>
</tr>
<tr>
<td>0.70</td>
<td>0.27</td>
<td>1.14</td>
</tr>
<tr>
<td>1.00</td>
<td>0.43</td>
<td>3.25</td>
</tr>
<tr>
<td>2.00</td>
<td>1.21</td>
<td>228.0</td>
</tr>
</tbody>
</table>

for inverse gamma distributions. Table 5.7 gives the values of $BF_{p(1)}$ for prior variance 10 for different prior means/medians. We observe the same behavior of the Bayes factor estimates except the fact that when the variance is large, the range of values of the Bayes factor estimate becomes smaller; for example, an inverse gamma prior distribution with mean 1 and variance 10 gives a Bayes factor estimate of 31 (as against 228 for an inverse gamma prior distribution with mean 1 and variance 2).
All of the above points to the fact that the Bayes factor can be significantly affected by the prior distribution on $\sigma^2$, but since the lowest possible Bayes factor estimate obtainable here is 0.22, the conclusion remains the same, i.e., there is not enough evidence against the null model.

These results show that for a single variance component model, a great deal can be learned by considering the sensitivity curve in Figure 5.5. For the turtle study, the sensitivity curve shows the sensitivity of the Bayes factor estimate to the prior distribution.

5.6 Simulation study

To further explore the sensitivity of Bayes factor estimate calculations in single variance component models, we carry out a simulation study. The two factors we focus on in the study are the true size of the variance component and the number of clutches.

We generate simulation data sets modeled on our original data set — that is, each observation has information on clutch, birth-weight and survival. The number of clutches
Table 5.7 Estimates of the Bayes factor for comparing a simple probit regression model against a probit regression model with a random intercept for different prior means/modes (with an inverse gamma prior distribution) with prior variance fixed at 10 for the turtles data set

<table>
<thead>
<tr>
<th>Prior mean</th>
<th>Prior mode</th>
<th>$BF^{01}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>0.01</td>
<td>0.00</td>
<td>0.76</td>
</tr>
<tr>
<td>0.10</td>
<td>0.03</td>
<td>0.30</td>
</tr>
<tr>
<td>0.30</td>
<td>0.10</td>
<td>0.37</td>
</tr>
<tr>
<td>0.50</td>
<td>0.17</td>
<td>0.58</td>
</tr>
<tr>
<td>0.75</td>
<td>0.26</td>
<td>1.09</td>
</tr>
<tr>
<td>1.00</td>
<td>0.35</td>
<td>2.14</td>
</tr>
<tr>
<td>2.00</td>
<td>0.82</td>
<td>31.5</td>
</tr>
</tbody>
</table>

are varied from 10 to 100 for the data sets generated while the number of observations for each clutch was kept fixed at 8 (which is approximately the average for the turtle data). The value of $\sigma^2$ is fixed for each data set and varies between 0.01 to 2.25. Note that at $\sigma^2 = 2.25$, the clutch effect is extremely important in determining survival. For a given number of groups and a given $\sigma^2$, the steps used to generate a data set are as follows:

- We generate the $x_{ij}$'s, the birth-weights, independently from a normal distribution having the same first two moments as the original data set.

- Generate $y_{ij}$'s independently using the three steps:

$$b_i \sim N(0, \sigma^2)$$

$$p_{ij} = \Phi(\alpha_0 + \alpha_1 x_{ij} + b_i)$$

$$y_{ij}\lvert p_{ij} \sim \text{Bern}(p_{ij}),$$

where we use the MLE's of $\alpha_0$ and $\alpha_1$ obtained from the original data set and the fixed value of $\sigma^2$ in place of the corresponding parameters.
We use the same prior distributions we used for the original data analysis, which are a $N(0, 20.1)$ on $\alpha$ and an inverse gamma$(2.5,1.5)$ prior distribution on $\sigma^2$ (which has mean 1 and variance 2) and compute the Bayes factor estimates using the importance sampling approach. Table 5.8 shows the average of 30 Bayes factor estimates corresponding to 30 generated data sets for each combination of true $\sigma^2$ and number of clutches.

<table>
<thead>
<tr>
<th>$\sigma^2$</th>
<th>Number of clutches</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10</td>
</tr>
<tr>
<td>0.01</td>
<td>0.49</td>
</tr>
<tr>
<td>0.04</td>
<td>0.32</td>
</tr>
<tr>
<td>0.09</td>
<td>0.18</td>
</tr>
<tr>
<td>0.16</td>
<td>0.08</td>
</tr>
<tr>
<td>0.25</td>
<td>0.02</td>
</tr>
<tr>
<td>0.49</td>
<td>0.002</td>
</tr>
<tr>
<td>1.00</td>
<td>$2.0 \times 10^{-5}$</td>
</tr>
<tr>
<td>2.25</td>
<td>$5.1 \times 10^{-8}$</td>
</tr>
</tbody>
</table>

If we go down any column of the table, we see that for a fixed number of groups, as the true value of $\sigma^2$ increases, the Bayes factor estimate in favor of the null model against the alternative model decreases. This is expected since the larger $\sigma^2$ is, the greater should be the evidence obtained against the null model.

When we go across a row from left to right, we expect the Bayes factor estimate to decrease as the number of groups increase for a fixed $\sigma^2$; this is because the same value of $\sigma^2$ should appear as stronger evidence against the null model when the number of clutches is large (the effective sample size for estimating $\sigma^2$ is the number of clutches). This happens for the last four rows, but not for the first three rows. The explanation of this unusual result appears to be the inverse-gamma prior distribution on $\sigma^2$ which is shown in Figure 5.6. The prior distribution assigns very little probability mass for
values of $\sigma^2$ less than 0.1. With such a prior distribution, the null model is found better because under the alternative model, the region with high restricted likelihood $\int p(y|\alpha, \sigma^2)p(\alpha)d\alpha$ receives very little probability mass under the prior distribution; as a result, the marginal density under the alternative model gets downsized to make the estimated Bayes factor (defined as the marginal density under the null model divided by the marginal density under the alternative model) large.

To see if we get expected results using other prior distributions for the first three rows in Table 5.8, we take up the first row of the table. We use a prior distribution with all its mass at 0.01 and compute the Bayes factor estimates for those four cases. The results obtained are in line with our expectations — the Bayes factor estimate decreases with increase in the number of groups — the values obtained are 0.96, 0.90, 0.87 and 0.83 respectively (all the Bayes factor estimates are close to one because for the true value of $\sigma^2=0.01$, there is very little difference between the null model and the alternative model). The same phenomenon is observed for the second and third rows of the table as well.
One interesting feature of Table 5.8 is that it provides information about the magnitude of the Bayes factor that would be expected under certain situations. In fact, the turtle data results, Bayes factor of 3.25 for a data set with 31 clutches and maximum likelihood estimate of $\sigma^2$ approximately 0.09, are precisely as would be expected under the model.

Another topic of interest in the simulation study is the comparative performance of the different computational methods for obtaining Bayes factor estimates. The relative performance of the methods remain the same as what we observed for the main data set regardless of the number of groups — the importance sampling, bridge sampling and Chib's method still perform better than the other methods throughout the simulation study.
6 EXAMPLE: SCOTLAND LIP CANCER DATA

In the previous chapter, we took up a simple example of a generalized linear mixed model. We next consider a more complex example with more than one variance component since the computations become much more difficult and time-consuming for such models. Some of the computational techniques for finding Bayes factor estimates that we discussed earlier become too time-consuming to apply.

Possible areas of potential application of generalized linear mixed models with more than one variance components are:

- spatial epidemiology studies, where it is of interest to investigate if the population disease rates, which may be influenced by environmental factors, are correlated spatially — there one often needs a number of area-specific random effects.

- longitudinal studies, where it may be of interest to see if random subject effects are present and one may have more than one variance component depending on the clustering of the subjects.

We consider an example belonging to the first type.

6.1 Description of the data set

Table 6.1 shows a frequently-analyzed data set (see, e.g., Clayton and Kaldor, 1987, Cressie, 1993) regarding lip cancer data from the 56 administrative districts in Scotland — the objective of the study was to find out any pattern of regional variation in the
disease incidence of lip cancer. The data set contains: the observed number of lip cancer cases among males from 1975-1980 in the 56 districts, \( y_1, y_2, \ldots, y_n, n=56 \); the population under risk of lip cancer in the districts, \( p_1, p_2, \ldots, p_n \) (in thousands); the expected number of cases adjusted for the age distribution of the districts, \( E_1, E_2, \ldots, E_n \); the percent of people employed in agriculture, forestry and fishing (AFF), \( AFF_1, AFF_2, \ldots, AFF_n \) (since increased exposure to sunlight has been implicated in the excess occurrence of lip cancers, these people working outdoors were thought to be under greater risk of the disease); and the neighbors of each district, \( N_1, N_2, \ldots, N_n \). The \( E_i \)'s incorporate known demographic risk factors, here age, that are not of direct interest.

6.2 A Poisson-Gaussian hierarchical model

It is common to assume that the disease incidence counts \( y \) follow independent Poisson distributions,

\[
y_i|\lambda_i \sim \text{Poisson}(\lambda_i E_i), \quad i = 1, 2, \ldots, n,
\]

with \( \lambda_i \) representing a relative risk parameter for the \( i \)-th region. In other words, \( \lambda_i \) is the risk in the region \( i \) relative to what is expected due to the known risk factors accounted for in \( E_i \). The standardized morbidity ratio (SMR) for the \( i \)-th region is defined as

\[
SMR_i = \frac{y_i}{E_i},
\]

One can think of the \( \lambda_i \)'s as a "smooth" version of the \( SMR_i \)'s. A value of \( \lambda_i \ll 1 \) indicates unusually low incidence rates while a value of \( \lambda_i \gg 1 \) indicates unusually high incidence rates. From the viewpoint of mapping and statistical inference, \( \lambda_i \)'s are the key parameters. If we could have collected all possible covariates affecting the \( \lambda_i \)'s, we could have done a perfect job of modeling the disease counts using a simple Poisson regression model. Practically, this is impossible since one is never going to know the set
Table 6.1 The Scotland lip cancer data set

<table>
<thead>
<tr>
<th>County</th>
<th>y</th>
<th>p (in '000)</th>
<th>x</th>
<th>E</th>
<th>Neighbors</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9</td>
<td>28</td>
<td>16</td>
<td>1.38</td>
<td>4 5 9 11 19</td>
</tr>
<tr>
<td>2</td>
<td>39</td>
<td>231</td>
<td>16</td>
<td>8.66</td>
<td>2 7 10</td>
</tr>
<tr>
<td>3</td>
<td>11</td>
<td>83</td>
<td>10</td>
<td>3.04</td>
<td>2 6 12</td>
</tr>
<tr>
<td>4</td>
<td>9</td>
<td>52</td>
<td>24</td>
<td>2.53</td>
<td>3 18 20 28</td>
</tr>
<tr>
<td>5</td>
<td>15</td>
<td>129</td>
<td>10</td>
<td>4.26</td>
<td>5 1 11 12 13 19</td>
</tr>
<tr>
<td>6</td>
<td>8</td>
<td>53</td>
<td>24</td>
<td>2.40</td>
<td>2 3 8</td>
</tr>
<tr>
<td>7</td>
<td>26</td>
<td>246</td>
<td>10</td>
<td>8.11</td>
<td>5 2 10 13 16 17</td>
</tr>
<tr>
<td>8</td>
<td>7</td>
<td>63</td>
<td>7</td>
<td>2.30</td>
<td>1 6</td>
</tr>
<tr>
<td>9</td>
<td>6</td>
<td>59</td>
<td>7</td>
<td>1.98</td>
<td>6 1 11 17 19 23 29</td>
</tr>
<tr>
<td>10</td>
<td>20</td>
<td>166</td>
<td>16</td>
<td>6.63</td>
<td>4 2 7 16 22</td>
</tr>
<tr>
<td>11</td>
<td>13</td>
<td>88</td>
<td>7</td>
<td>4.40</td>
<td>4 1 5 9 12</td>
</tr>
<tr>
<td>12</td>
<td>5</td>
<td>38</td>
<td>16</td>
<td>1.79</td>
<td>3 3 5 11</td>
</tr>
<tr>
<td>13</td>
<td>3</td>
<td>29</td>
<td>10</td>
<td>1.08</td>
<td>4 5 7 17 19</td>
</tr>
<tr>
<td>14</td>
<td>8</td>
<td>86</td>
<td>24</td>
<td>3.31</td>
<td>3 31 32 35</td>
</tr>
<tr>
<td>15</td>
<td>17</td>
<td>185</td>
<td>7</td>
<td>7.84</td>
<td>3 25 29 50</td>
</tr>
<tr>
<td>16</td>
<td>9</td>
<td>112</td>
<td>16</td>
<td>4.55</td>
<td>6 7 10 17 21 22 29</td>
</tr>
<tr>
<td>17</td>
<td>2</td>
<td>27</td>
<td>10</td>
<td>1.07</td>
<td>6 7 9 13 16 19 29</td>
</tr>
<tr>
<td>18</td>
<td>7</td>
<td>94</td>
<td>7</td>
<td>4.18</td>
<td>6 4 20 28 33 55 56</td>
</tr>
<tr>
<td>19</td>
<td>9</td>
<td>163</td>
<td>7</td>
<td>5.53</td>
<td>5 1 5 9 13 17</td>
</tr>
<tr>
<td>20</td>
<td>7</td>
<td>103</td>
<td>10</td>
<td>4.44</td>
<td>3 4 18 55</td>
</tr>
<tr>
<td>21</td>
<td>16</td>
<td>263</td>
<td>7</td>
<td>10.46</td>
<td>3 16 29 50</td>
</tr>
<tr>
<td>22</td>
<td>31</td>
<td>583</td>
<td>16</td>
<td>22.67</td>
<td>2 10 16</td>
</tr>
<tr>
<td>23</td>
<td>11</td>
<td>191</td>
<td>10</td>
<td>8.77</td>
<td>6 9 29 34 36 37 39</td>
</tr>
<tr>
<td>24</td>
<td>7</td>
<td>164</td>
<td>7</td>
<td>5.62</td>
<td>8 27 30 31 44 47 48 55 56</td>
</tr>
<tr>
<td>25</td>
<td>19</td>
<td>432</td>
<td>1</td>
<td>15.47</td>
<td>3 15 26 29</td>
</tr>
<tr>
<td>26</td>
<td>15</td>
<td>379</td>
<td>1</td>
<td>12.49</td>
<td>4 25 29 42 43</td>
</tr>
<tr>
<td>27</td>
<td>7</td>
<td>164</td>
<td>7</td>
<td>6.04</td>
<td>4 24 31 32 55</td>
</tr>
<tr>
<td>28</td>
<td>10</td>
<td>231</td>
<td>7</td>
<td>8.96</td>
<td>4 4 18 33 45</td>
</tr>
<tr>
<td>29</td>
<td>16</td>
<td>346</td>
<td>10</td>
<td>14.37</td>
<td>11 9 15 16 17 21 23 25 26 34 43 50</td>
</tr>
<tr>
<td>30</td>
<td>11</td>
<td>383</td>
<td>10</td>
<td>10.20</td>
<td>6 24 38 42 44 45 56</td>
</tr>
<tr>
<td>31</td>
<td>5</td>
<td>139</td>
<td>7</td>
<td>4.75</td>
<td>7 14 24 27 32 35 46 47</td>
</tr>
<tr>
<td>32</td>
<td>3</td>
<td>65</td>
<td>24</td>
<td>2.88</td>
<td>4 14 27 31 35</td>
</tr>
<tr>
<td>33</td>
<td>7</td>
<td>250</td>
<td>10</td>
<td>7.03</td>
<td>4 18 28 45 56</td>
</tr>
<tr>
<td>34</td>
<td>8</td>
<td>233</td>
<td>7</td>
<td>8.53</td>
<td>9 23 29 39 40 42 43 51 52 54</td>
</tr>
<tr>
<td>35</td>
<td>11</td>
<td>319</td>
<td>7</td>
<td>12.32</td>
<td>5 14 31 32 37 46</td>
</tr>
<tr>
<td>36</td>
<td>9</td>
<td>296</td>
<td>0</td>
<td>10.10</td>
<td>4 23 37 39 41</td>
</tr>
<tr>
<td>37</td>
<td>11</td>
<td>391</td>
<td>10</td>
<td>12.68</td>
<td>5 23 35 36 41 46</td>
</tr>
<tr>
<td>38</td>
<td>8</td>
<td>319</td>
<td>1</td>
<td>9.35</td>
<td>6 30 42 44 49 51 54</td>
</tr>
<tr>
<td>39</td>
<td>6</td>
<td>231</td>
<td>16</td>
<td>7.20</td>
<td>5 23 34 36 40 41</td>
</tr>
<tr>
<td>40</td>
<td>4</td>
<td>157</td>
<td>0</td>
<td>5.27</td>
<td>5 34 39 41 49 52</td>
</tr>
</tbody>
</table>
of all possible covariates. One way to address the effects of unobserved covariates is to incorporate random effects. As in Cressie et. al. (2000), we use a mixed linear model for the vector of log relative risk parameters, \( \log(\lambda) \),

\[
\log(\lambda) = X\beta + \eta + \psi,
\]

where \( X \) is the covariate matrix containing in this example a vectors of 1's as the first column and another column containing values of the variable AFF; \( \beta \) is a vector containing the fixed effect parameters \( \beta_0 \) and \( \beta_1 \); \( \eta = (\eta_1, \eta_2, \ldots, \eta_n)' \) is a vector of spatially correlated random effects and \( \psi = (\psi_1, \psi_2, \ldots, \psi_n)' \) is a vector of uncorrelated heterogeneity random effects.

The spatial random effects \( \eta_i \)'s are intended to represent unobserved factors, that if observed, would display substantial spatial correlation in that the values for a pair of contiguous zones would be generally much more alike than for two arbitrary zones. Examples might include environmental factors. For known matrices \( C \) and \( M \), we take
the prior distribution for $\eta$ as a conditional autoregressive (CAR) distribution

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

where $\tau^2$ and $\phi$ are parameters of the prior distribution. This prior distribution is motivated through the conditional distributions $\eta_i | \eta_{-i}$, for $i = 1, 2, \ldots, n$, where $\eta_{-i} = \{\eta_1, \ldots, \eta_{i-1}, \eta_{i+1}, \ldots, \eta_n\}$. Let $N_i = \{j : c_{ij} \neq 0\}$ represent the set of neighbors of region $i$, that is the regions $j$ such that $\eta_j$ is assumed to be associated with $\eta_i$; regions are usually assumed to be neighbors if their boundaries touch. The conditional distributions that correspond to the joint distribution of $\eta$ given above are:

$$\eta_i | \eta_{-i}, \tau^2, \phi \sim N\left(\phi \sum_{j \in N_i} c_{ij} \eta_j, \tau^2 m_{ii}\right), \quad i = 1, 2, \ldots, n.$$ 

The conditionally specified form of the model makes it easy to interpret the various elements of the prior distribution. Specifically, $M$ is a matrix of conditional variances (up to the scalar $\tau^2$) and $\phi$ is a measure of the strength of spatial dependence, with $\phi = 0$ implying no spatial association. Large values of $c_{ij}$ identify neighbors whose spatial random effects are expected to be most closely associated with $\eta_i$.

In order for this CAR prior distribution on $\eta$ to be a proper prior distribution, $\tau^2(I - \phi C)^{-1} M$ must be a symmetric positive-definite matrix. Thus, $C$ and $M$ must be chosen to satisfy the symmetry condition

$$m_{ii}c_{ji} = m_{jj}c_{ij}, \quad i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, n.$$ 

To maintain the positive definiteness of $\tau^2(I - \phi C)^{-1} M$, $\phi$ should be contained in the interval $(\phi_{\min}, \phi_{\max})$, determined from the eigenvalues of $M^{-1/2}CM^{1/2}$. Here, we restrict $\phi$ to belong to the subinterval $(0, \phi_{\max})$ as suggested in Cressie et. al. (2000).
We take the matrices $C$ and $M$ as suggested in Stern and Cressie (1995):

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

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

For these values of $c_{ij}$'s and $m_{ii}$'s, we have $\phi_{\text{max}} = 0.1752$.

The uncorrected heterogeneity random effects $\psi$'s represent the unstructured variables contributing to the logarithm of the relative risk parameters. The inclusion of $\psi$'s is due to Breslow (1984), who noted strong empirical evidence of extra-Poisson variation. The heterogeneity random effects are modeled as:

$$
\psi|\sigma^2 \sim N(0, \sigma^2 D)
$$

for a known diagonal matrix $D$ and a variance parameter $\sigma^2$. We take $D$ as suggested in Stern and Cressie (1995):

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

In practice, it appears often to be the case that either $\eta$ or $\psi$ dominates the other, but which one will not usually be known in advance. If $\eta$, then the estimated relative risks will display spatial structure; if $\psi$, then the effect will be to shrink the estimated relative risks to the overall mean (Besag et. al., 1991).

This model contains 2 variance parameters and as many as 112 random effects parameters, making it a more challenging data set to handle computationally.
6.3 Estimation

Let $x_i$ denote the $i$-th row of the covariate matrix $X$. The marginal likelihood for the data set is:

$$L(\beta, \phi, \tau^2, \sigma^2 | y) = \int \int p(y | \eta, \psi, \beta)p(\eta, \psi | \phi, \tau^2, \sigma^2) d\eta d\psi$$

$$\propto \int \int \left[ \prod_{i=1}^{n} \exp \left( - E_i e^{x_i(x_i \psi + \eta)} e^{\eta_i(x_i \psi + \eta)} \right) \right] \times$$

$$\frac{1}{\tau^2 (I - \phi C)^{-1} M^{1/2}} \cdot \exp \left\{ - \frac{1}{2} \eta (\tau^2 (I - \phi C)^{-1} M)^{-1} \eta \right\} \times$$

$$\frac{1}{\sigma^2 D^{1/2}} \cdot \exp \left\{ - \frac{1}{2} \psi \left( \sigma^2 D \right)^{-1} \psi \right\} d\eta d\psi.$$

The distribution of $y$ depends on the random effects only through the sum $\nu = \eta + \psi$.

By making a transformation of $(\eta, \psi)$ to $(\nu, \psi)$ and then integrating out $\psi$, we get the following expression of the likelihood:

$$L(\beta, \phi, \tau^2, \sigma^2 | y) \propto \int \left\{ \prod_{i=1}^{n} \exp \left( - E_i e^{x_i(x_i \nu + \psi)} e^{\nu_i(x_i \nu + \psi)} \right) \right\} \times$$

$$\frac{1}{|V|^{1/2}} \cdot \exp \left\{ - \frac{1}{2} \nu V^{-1} \nu \right\} d\nu,$$

where

$$V = \tau^2 (I - \phi C)^{-1} M + \sigma^2 D = \text{Var}(\nu).$$

The advantage of this second expression is that it has an $n$-dimensional integral instead of a $2n$-dimensional integral (where $n$ is 56) as in the first expression. Despite this it is still not practical to evaluate the likelihood by numerical integration because the random effects in the model are not assumed independent a priori (the prior variance of $\nu$ is not diagonal). To get the marginal likelihood, we have to integrate over the 56-dimensional random effects parameter-space.
6.3.1 Maximum likelihood estimation

We use the EM algorithm to find the maximum likelihood estimates of the parameters for the model. Define $\xi = (\beta', \phi, \tau^2, \sigma^2)'$, the collection of the fixed effects parameters and the variance parameters of the model. The traditional EM algorithm consists of an E-step and an M-step where in the $(r+1)$-th E-step, one computes

$$Q(\xi|\xi^{(r)}) = E[\log\{p(y, \nu|\xi)\}|y, \xi^{(r)}],$$

where $p(y, \nu|\xi)$ is the joint density of $(y', \nu')$, $\xi^{(r)}$ is the value of the parameter vector $\xi$ in the $r$-th iteration and $\nu = \eta + \psi$ as defined earlier. The expectation in the expression is taken with respect to the distribution of $\nu$ given $y$ conditional on the parameter value $\xi^{(r)}$. In the corresponding M-step, one usually maximizes $Q(\xi|\xi^{(r)})$ with respect to $\xi$ to get $\xi^{(r+1)}$, i.e.,

$$Q(\xi^{(r+1)}|\xi^{(r)}) \geq Q(\xi|\xi^{(r)})$$

for all $\xi$ in the parameter space.

In the E-step of the EM algorithm here, the expectation is with respect to the distribution of $\nu$ given $y$ conditional on the parameter value $\xi^{(r)}$, whose density can be written as $p(\nu|y, \xi^{(r)})$, where

$$p(\nu|y, \xi) \propto p(y|\nu, \beta)p(\nu|\phi, \tau^2, \sigma^2), \quad (6.1)$$

the normalizing constant being given by the marginal likelihood $L(\xi|y)$ of the data,

$$L(\xi|y) = \int p(y|\nu, \beta)p(\nu|\phi, \tau^2, \sigma^2)d\nu.$$

So an analytical computation of $Q(\xi|\xi^{(r)})$ in the E-step is impossible except for the normal mixed model and we need some kind of approximation.

We use the importance sampling EM approach of Booth and Hobert (1999) as discussed in Section 4.2.2.2. In this method, one draws a random sample $\nu_{r,1}^*, \nu_{r,2}^*, \ldots, \nu_{r,m}^*$. 
from an importance sampling density \( p^*(\nu) \), which has the same support as \( p(\nu|y, \xi^{(r)}) \).

Then the importance sampling Monte Carlo estimate of \( Q(\xi|\xi^{(r)}) \) is given by

\[
Q_m(\xi|\xi^{(r)}) = \frac{1}{m} \sum_{i=1}^{m} w_{r,i} \log \{ p(y, \nu_{r,i}^*|\xi) \},
\]

where the importance weights \( w_{r,i} \)'s are defined by

\[
w_{r,i} = \frac{p(\nu_{r,i}^*|y, \xi^{(r)})}{p^*(\nu_{r,i}^*)}.
\]

Booth and Hobert (1999) propose a multivariate t importance density whose mean and variance are the same as the mode and curvature of \( p(\nu|y, \xi^{(r)}) \). Specifically, ignoring \( y \) and \( \xi \) in the notation for the time being, we write \( p(\nu) = c e^{l(\nu)} \), where \( c \) is the normalizing constant. Let \( l^{(1)}(\nu) \) denote the vector of first derivatives of \( l(\nu) \) and \( l^{(2)}(\nu) \) the second derivative matrix. Suppose further that \( \tilde{\nu} \) maximizes \( l(\nu) \) satisfying the equation

\[
l^{(1)}(\nu) = 0.
\]

Then the Laplace approximation of the mean and variance of the multivariate t importance density to be used are \( \tilde{\nu} \) and \( [-l^{(2)}(\tilde{\nu})]^{-1} \) respectively.

For our model,

\[
p(y|\nu, \beta) = \prod_{i=1}^{n} \exp \left( -E_i e^{x_i^T \beta + \nu_i} \right) e^{y_i} e^{y_i(x_i^T \beta + \nu_i)}
\]

\[
p(\nu|\phi, \tau^2, \sigma^2) = \frac{1}{\pi^{n/2}|V|^{1/2}} \exp \left[ -\frac{1}{2} \nu' V^{-1} \nu \right],
\]

where

\[
V = \tau^2 (I - \phi C)^{-1} M + \sigma^2 D.
\]

Hence, by equation (6.1),

\[
p(\nu|y, \xi) \propto \left[ \prod_{i=1}^{n} \exp \left( -E_i e^{x_i^T \beta + \nu_i} \right) e^{y_i(x_i^T \beta + \nu_i)} \right] \frac{1}{\pi^{n/2}|V|^{1/2}} \exp \left[ -\frac{1}{2} \nu' V^{-1} \nu \right].
\]

Then the first and second derivatives of \( \log \{ p(\nu|y, \xi) \} \) are
\[ l^{(1)}(\nu) = \begin{pmatrix}
  y_1 - E_1 e^{\mathbf{x}_1^T \beta + \nu_1} \\
  y_2 - E_1 e^{\mathbf{x}_2^T \beta + \nu_2} \\
  \vdots \\
  y_n - E_n e^{\mathbf{x}_n^T \beta + \nu_n}
\end{pmatrix} - \mathbf{V}^{-1} \mathbf{\nu} \]  

(6.8)

and

\[ l^{(2)}(\nu) = \begin{pmatrix}
  -E_1 e^{\mathbf{x}_1^T \beta + \nu_1} & 0 & \ldots & 0 \\
  0 & -E_2 e^{\mathbf{x}_2^T \beta + \nu_2} & \ldots & 0 \\
  \vdots & \vdots & \ddots & \vdots \\
  0 & 0 & \ldots & -E_n e^{\mathbf{x}_n^T \beta + \nu_n}
\end{pmatrix} - \mathbf{V}^{-1}, \]  

(6.9)

where \( \mathbf{V} \) is defined by (6.6).

**E-step:** From the above discussion, starting with an initial value of the parameters \( \beta, \phi, \tau^2, \sigma^2 \) and \( \nu \), the \( (r+1) \)-th iteration of the E-step of the MCEM algorithm consists of the following steps:

(i) Find \( \hat{\nu} \), the maximizer of \( l(\nu) \), using the Newton-Raphson algorithm. That is, iterate until convergence

\[ \hat{\nu}^{(k+1)} = \hat{\nu}^{(k)} - \left[ l^{(2)}(\hat{\nu}^{(k)}) \right]^{-1} l^{(1)}(\hat{\nu}^{(k)}), \]

where the derivatives are computed at the current value of parameters, i.e., at \( \beta = \beta^{(r)}, \phi = \phi^{(r)}, \tau^2 = \tau^2^{(r)}, \sigma^2 = \sigma^2^{(r)}, \nu = \nu^{(k)} \) and using (6.8) and (6.9). In fact, we find that rather than iterating the algorithm till convergence here, taking two to three Newton steps here also leads to convergence of the EM algorithm and saves a lot of time.

(ii) Set the variance of the multivariate t importance density (taken as a t distribution with 4 degrees of freedom here) as \( -l^{(2)}(\hat{\nu}) \)^{-1}, which is available from the last iteration of the Newton-Raphson algorithm in step (i) above.
(iii) Generate \( \nu_{r,1}, \nu_{r,2}, \ldots, \nu_{r,m} \), a random sample from the importance density determined by the steps (i) and (ii).

(iv) Compute the importance weights \( w_{r,i} \)'s using equation (6.3), where \( p^*(\nu_{r,i}) \) is the value of the multivariate \( t \) importance density at \( \nu_{r,i} \) and \( p(\nu_{r,i}|y, \xi^{(r)}) \) is computed using (6.7) at the current value of parameters, \( \beta^{(r)}, \phi^{(r)}, \tau^{2(r)} \) and \( \sigma^{2(r)} \). Using the importance weights, we can compute \( Q_m(\xi|\xi^{(r)}) \) using equation (6.2) for any fixed value of \( \xi = (\beta', \phi, \tau^2, \sigma^2)' \).

**M-step:** In the corresponding M-step, we maximize \( Q_m(\xi|\xi^{(r)}) \) with respect to \( \xi \).

Now, using (6.2), (6.4) and (6.5) and observing that

\[
p(y, \nu|\xi) = p(y|\nu, \beta)p(\nu|\phi, \tau^2, \sigma^2),
\]

we get

\[
Q_m(\xi|\xi^{(r)}) = \frac{1}{m} \sum_{i=1}^{m} w_{r,i} \left[ \sum_{i=1}^{n} \left( -E_i e^{x_i'\beta + \nu_{r,i}} + y_i(x_i'\beta + \nu_{r,i}) \right) - \frac{1}{2} \nu_{r,i} V^{-1} \nu_{r,i} - \frac{1}{2} \log |V| \right]
\]

Looking at the above expression, it is clear that \( Q_m(\xi|\xi^{(r)}) \) can be written as the sum of two functions, one involving only \( \beta \) and the other involving only the variance parameters \( \tau^2, \sigma^2 \) and \( \phi \). Hence the problem of maximization can be carried out in two unrelated steps — the first step consists of maximizing the first term with respect to \( \beta \) and the second in maximizing the second term with respect to \( (\phi, \tau^2, \sigma^2) \). To maximize the first term with respect to \( \beta \), we observe that \( Q^{(1)}(\beta) \), the vector of first derivatives of \( Q_m(\xi|\xi^{(r)}) \) with respect to \( \beta \) and \( Q^{(2)}(\beta) \), the matrix of second derivatives of \( Q_m(\xi|\xi^{(r)}) \)
with respect to $\beta$ are given by

$$Q^{(1)}(\beta) = \left( \frac{1}{m} \sum_{l=1}^{m} w_{r,l} \sum_{i=1}^{n} \left( y_i - E_i e^{x_i^T \beta + \nu_{i,r,l}} \right) \right)$$

and

$$Q^{(2)}(\beta) = - \left( \frac{1}{m} \sum_{l=1}^{m} w_{r,l} \sum_{i=1}^{n} x_i E_i e^{x_i^T \beta + \nu_{i,r,l}} \right)$$

So the maximization algorithm for $\beta$ amounts to iterating using the equation

$$\hat{\beta}^{(k)} = \hat{\beta}^{(k-1)} - \left[ Q^{(2)}(\hat{\beta}^{(k-1)}) \right]^{-1} Q^{(1)}(\hat{\beta}^{(k-1)})$$

until it converges. Just as in step (i) of the E-step, rather than going on till convergence, taking only two to three Newton steps is enough for the E-M algorithm to converge.

The M-step with respect to $(\phi, \tau^2, \sigma^2)$ is more difficult since obtaining closed-form expressions of the derivatives of $Q_m(\xi(\tau^2))$ is quite cumbersome. We use numerical derivatives instead (as discussed in Section 2.1.1) and the Newton-Raphson method (with a switch to a steepest descent step if Cholesky decomposition of the second derivative matrix does not exist). Also, we work with the logarithms of $\tau^2$ and $\sigma^2$ to avoid any problem with convergence in case the maximum likelihood estimate of one or both of these variance components is close to zero. We use only one Newton step here (because of the complexity of computing the derivatives) rather than proceeding until convergence.

The size of the importance sample drawn at each step was taken as 10000. When the absolute relative distance between the parameter vectors obtained from two successive M-steps is below 0.0001 for three consecutive iterations, we decide that the algorithm has converged. This is to make sure that we don’t stop the algorithm prematurely just because of an unlucky importance sample.

Using the MCEM algorithm, the MLE of $\xi = (\beta', \phi, \tau^2, \sigma^2)'$ obtained is

$$\hat{\xi}_{MLE} = (-0.489, 0.059, 0.167, 1.640, 0.000)'$$.
An approximate estimate of the variance of the estimated MLE can be obtained using the same techniques used to get (4.7) in Section 4.2.2.2, which here means that the variance estimate is the inverse of $I$, where

$$I = -Q^{(2)}(\xi | \xi^{(r)}) - \operatorname{var} \left[ \frac{\partial}{\partial \xi} \log \left( p(y, \nu | \xi) \right) | y, \hat{\xi}_{MLE} \right]. \quad (6.11)$$

We use the sampled values $\nu^{*}_{R,1}, \nu^{*}_{R,2}, \ldots, \nu^{*}_{R,m}$ to compute $I$, where the $R$-th iteration is the last iteration.

In this example, (6.11) cannot be used since the matrix $I$ does not turn out to be positive definite. This is the outcome of the maximum likelihood estimate being found at the boundary of the parameter space.

To see how the method works when the maximum likelihood estimate is not found at the boundary of the parameter space, we consider maximum likelihood estimation of the parameters assuming that there is no $\psi$ and hence no $\sigma^2$ in the model, i.e., we look at the following simplified model:

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

$$\log(\lambda) = X \beta + \eta$$

and

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

The maximum likelihood estimate of the parameters $\beta', \phi$, and $\tau^2$ are the same as that for the full model above, i.e.,

$$({\hat{\beta}'}, {\hat{\phi}}, {\hat{\tau}^2}) = (-0.489, 0.059, 0.167, 1.640)$$

and the estimated variance matrix of the maximum likelihood estimate is:
The estimated correlation matrix is:

$$
\begin{pmatrix}
0.0281 & -0.0013 & 0.0009 & -0.0206 \\
0.0002 & -0.0000 & 0.0002 \\
0.0002 & -0.0023 & 0.2869 \\
\end{pmatrix}
$$

6.3.2 Bayesian estimation

The joint posterior distribution of $\xi = (\beta', \phi, \tau^2, \sigma^2)'$ given the data is proportional to

$$
\int \int \left\{ \prod_{i=1}^{n} \exp \left( -E_i \varepsilon \cdot (\beta + n_i + \psi_i) e_{i}(\beta + n_i + \psi_i) \right) \right\} \frac{1}{|V|^{1/2}} \times
\exp \left\{ -\frac{1}{2}(\eta + \psi)' V^{-1}(\eta + \psi) \right\} d\eta d\psi P(\beta) P(\phi) P(\tau^2) P(\sigma^2) \mathcal{I}_{\phi \in (0, \phi_{\max})},
$$

where $V$ is given by (6.6). We take $P(\beta)$ to be a bivariate normal distribution with mean 0 and variance 20.I, $P(\phi) = Uniform(\phi|0, \phi_{\max})$ and both $P(\tau^2)$ and $P(\sigma^2)$ to be inverse-gamma distributions with mean 1 and variance 2 (i.e., the parameters of the distributions are 2.5 and 1.5). To avoid the integration of the random effects, we work with the joint posterior distribution of $(\beta', \eta', \psi', \phi, \tau^2, \sigma^2)'$, which is proportional to

$$
\left\{ \prod_{i=1}^{n} \exp \left( -E_i \varepsilon \cdot (\beta + n_i + \psi_i) e_{i}(\beta + n_i + \psi_i) \right) \right\} \frac{1}{|V_1|^{1/2}} \times \exp \left\{ -\frac{1}{2} \eta' V_1^{-1} \eta \right\} \times
\frac{1}{|\sigma^2 D|^{1/2}} \cdot \exp \left\{ -\frac{1}{2} \psi' \left( \sigma^2 D \right)^{-1} \psi \right\} e^{-\frac{\phi^2}{\phi^2}} \left( \tau^2 \right)^{-a - 1} e^{-\frac{\tau^2}{2(\sigma^2)}} \mathcal{I}_{\phi \in (0, \phi_{\max})},
$$
where $V_i = \tau^2 (I - \phi C)^{-1} M$, $a = 2.5$, $b = 1.5$.

We use an MCMC algorithm, specifically, a Gibbs sampling approach, to generate a posterior sample from the above joint posterior distribution. Let

$$C = ((c_{ij})), i, j = 1, 2, \ldots n; D = ((d_{ij})), i, j = 1, 2, \ldots n; M = ((m_{ij})), i, j = 1, 2, \ldots n$$

The full conditional distributions of $\beta$, $\{\eta_k, k = 1, 2, \ldots n\}$, $\{\psi_k, k = 1, 2, \ldots n\}$, $\tau^2$, $\sigma^2$ and $\phi$ that are needed to implement a Gibbs sampling algorithm are:

$$p(\beta | \eta, \psi, \tau^2, \sigma^2, \phi, y) \propto \prod_{i=1}^{n} \exp\left\{ - \frac{1}{2 \tau^2} \left( \sum_{i=1}^{n} \eta_i^2 (1 - \phi c_{ii}) + \sum_{i \neq j} \frac{-\phi \eta_i \eta_j c_{ij}}{m_{ii}} + 2b \right) \right\}$$

$$p(\eta_k | \eta_{-k}, \psi, \beta, \tau^2, \sigma^2, \phi, y) \propto \exp\left\{ - \frac{1}{2 \tau^2} \left( \sum_{i=1}^{n} \eta_i^2 (1 - \phi c_{kk}) + \phi \eta_k \sum_{i \neq k} \eta_i \left( \frac{c_{ki}}{m_{kk}} + \frac{c_{ik}}{m_{ii}} \right) \right) \right\}$$

$$p(\psi_k | \psi_{-k}, \eta, \beta, \tau^2, \sigma^2, \phi, y) \propto \exp\left\{ - \frac{1}{2 \sigma^2 d_{kk}} \left( \sum_{i=1}^{n} \eta_i^2 (1 - \phi c_{kk}) + \phi \eta_k \sum_{i \neq k} \eta_i \left( \frac{c_{ki}}{m_{kk}} + \frac{c_{ik}}{m_{ii}} \right) \right) \right\}$$

$$p(\tau^2 | \eta, \psi, \beta, \sigma^2, \phi, y) \propto \left( \frac{\tau^2}{\sigma^2} \right)^{\frac{n}{2} + s + 1} \exp\left\{ - \frac{1}{2 \tau^2} \left( \sum_{i=1}^{n} \eta_i^2 (1 - \phi c_{ii}) + \sum_{i \neq j} \frac{-\phi \eta_i \eta_j c_{ij}}{m_{ii}} + 2b \right) \right\}$$

$$p(\sigma^2 | \eta, \psi, \beta, \tau^2, \phi, y) \propto \left( \frac{\sigma^2}{\tau^2} \right)^{\frac{n}{2} + s + 1} \exp\left\{ - \frac{1}{2 \sigma^2} \left( \sum_{i=1}^{n} \frac{\psi_i^2}{d_{ii}} + 2b \right) \right\}$$

$$p(\phi | \eta, \psi, \beta, \tau^2, \sigma^2, y) \propto |I - \phi C|^{\frac{1}{2}} \times \exp\left\{ - \frac{1}{2 \tau^2} \left( \sum_{i=1}^{n} \eta_i^2 (1 - \phi c_{ii}) + \sum_{i \neq j} \frac{-\phi \eta_i \eta_j c_{ij}}{m_{ii}} \right) \right\} I_{\phi \in (0, \phi_{max})}$$

The full conditional distributions for $\tau^2$ and $\sigma^2$ are inverse-gamma distributions. We use Metropolis steps for generating all of the remaining parameters. In the Metropolis steps, we use the common random walk jumping distributions — i.e., at any step, the jumping distribution for a parameter is a normal distribution with mean equal to the current value of the parameter in the chain and variance chosen so that the acceptance
rate is between 0.20-0.45. To choose the variances of the proposal distributions, we run an initial Markov chain with all variances fixed at unity. We compute the sample variances of the parameter values generated from the initial Markov chain and use multiples of them as the variance in the final chain — the value of the multipliers are chosen by experimenting to make the acceptance rates of all the parameters in the final algorithm between 0.2-0.45. We run five chains (using different initial values) of 10000 iterations each after a burn-in of 2000 iterations each and base our inference on the 50000 iterations from all the five chains combined. We use such large numbers to counter-balance the high autocorrelations of the parameter values generated by the chains. We use the Gelman-Rubin convergence diagnostic (Gelman and Rubin, 1992) on the five chains generated to check if the Markov chains have converged. Numerical summaries of the posterior distributions are given in Table 6.2. Figure 6.1 shows histograms of the sampled values of the parameters of the model.

Table 6.2 Summaries of the posterior distributions of the parameters of a Poisson-normal model with two sets of random effects applied to the Scotland lip cancer data set

<table>
<thead>
<tr>
<th>parameter</th>
<th>mean</th>
<th>sd</th>
<th>Quantiles</th>
<th>2.5%</th>
<th>50%</th>
<th>97.5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_1$</td>
<td>-0.583</td>
<td>0.162</td>
<td>-0.893</td>
<td>-0.589</td>
<td>-0.248</td>
<td></td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>0.067</td>
<td>0.014</td>
<td>0.040</td>
<td>0.067</td>
<td>0.092</td>
<td></td>
</tr>
<tr>
<td>$\phi$</td>
<td>0.142</td>
<td>0.035</td>
<td>0.038</td>
<td>0.155</td>
<td>0.174</td>
<td></td>
</tr>
<tr>
<td>$\tau^2$</td>
<td>1.216</td>
<td>0.562</td>
<td>0.404</td>
<td>1.118</td>
<td>2.553</td>
<td></td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>0.839</td>
<td>0.837</td>
<td>0.252</td>
<td>0.710</td>
<td>2.130</td>
<td></td>
</tr>
</tbody>
</table>

Comparing the posterior means with the maximum likelihood estimates (MLE), we see that they are more or less close except for $\tau^2$, for which the posterior mean is much larger than the corresponding MLE. This is not unexpected. The posterior distribution of $\sigma^2$ is clearly sensitive to the prior distribution used for $\sigma^2$. The joint posterior mode, which is found using the MCEM algorithm (see Section 4.2.2.2), is 
(-0.457,0.062,0.172,0.867,0.507)').
Figure 6.1 Histograms of the sampled values of the parameters
The correlation matrix of the parameters computed using the sample obtained from the joint posterior distribution is given by:

\[
\begin{pmatrix}
1.00 & -0.62 & 0.38 & -0.25 & -0.00 \\
1.00 & 1.00 & 0.17 & -0.24 & -0.10 \\
1.00 & 0.20 & 1.00 & -0.00 & -0.14 \\
1.00 & -0.38 & 0.00 & 1.00 & 0.00
\end{pmatrix}
\]

### 6.4 Testing hypotheses about the variance components

As our main focus is on Bayes factors, we do not carry out traditional hypothesis tests in the disease mapping example. It would be possible to carry out the simulation approach, but it will be very time-consuming. Informally, it seems that is not required in the model since the MLE of $\sigma^2$ is 0 and the posterior mean is not large compared to the posterior standard deviation. The parameter $\tau^2$ seems to be important in the model by looking at the MLE and the posterior summary.

### 6.5 Approximating the Bayes factors

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

- “full model” with $\tau^2$ (and $\phi$) and $\sigma^2$
- “spatial model” with $\tau^2$ (and $\phi$) only as a variance component
- “heterogeneity model” with $\sigma^2$ only as a variance component
- “null model” with no variance component
We focus on the three Bayes factors obtained by comparing any one of the three reduced models to the full model. Note that any other Bayes factor can be obtained from these three.

It is difficult to apply some of the methods discussed in Chapter 4 to this multiple variance component problem because they require the computation of the likelihood at a large number of points. Integration over the random effects to evaluate the likelihood is time-consuming because the number of random effects parameters is large. This limits the usefulness of importance sampling, harmonic estimation, bridge sampling or reversible jump MCMC. One way to get these methods to work is to include the random effects as parameters in the model rather than trying to integrate them out (as discussed immediately after Section 4.3.5.5). We compute the Bayes factor estimate using importance sampling and harmonic estimation methods by using this technique. However, even with treating the random effects as parameters, application of bridge sampling or reversible jump MCMC method is difficult — the choice of the function \( \gamma \) in the former and the proposal distributions for the latter make the methods very difficult to apply.

We also use the methods which avoid the problem of repeated evaluation of the likelihood: the Laplace approximation, the Verdinelli-Wasserman method and Chib's method. To be able to apply the Verdinelli-Wasserman method, we require that the prior distribution for \( \sigma^2 \) and \( \tau^2 \) be finite and non-zero at \( \sigma^2 = 0 \) and \( \tau^2 = 0 \). We cannot use the inverse gamma prior distribution on \( \sigma^2 \) and \( \tau^2 \) since the density of the inverse gamma distribution has a zero value for ordinate zero. An alternative is the shrinkage prior distribution (see, for example, Daniels, 1999) for the variance components, that is,

\[
p(\sigma^2) = \frac{c_1}{(c_1 + \sigma^2)^2}
\]

and

\[
p(\tau^2) = \frac{c_2}{(c_2 + \tau^2)^2}
\]
where \(c_i\)'s are fixed constants denoting the medians of the respective prior distributions. We also assume apriori independence of \(\sigma^2\) and \(\tau^2\). We fix the \(c_i\)'s at 1 for both of the distributions and use these prior distributions for all the methods. For \(\beta\) and \(\phi\), we use the prior distributions used in Bayesian estimation, that is, \(p(\beta)\) is taken to be a bivariate normal distribution with mean 0 and variance 20.I and \(p(\phi) = \text{Uniform}(\phi_0, \phi_{\text{max}})\).

We run the programs to compute estimates of each of the Bayes factors using each computation method 30 times each using different random number seeds and calculate the average and standard deviation of those 30 Bayes factor estimates obtained. The values are given in Table 6.3. The results provided by the Harmonic estimator are so poor that they are not shown in the table.

Also given in the table are the CPU times taken for one computation of Bayes factor estimate by each of these methods in an Alphastation 500 workstation equipped with 400MHz 64-bit CPU and a gigabyte of RAM.

To compare the Bayes factor estimates obtained by the different methods, it would help to know the correct values. Of course, if there were an easy way to obtain the correct value, we would not need to explore the different approaches. The approach that is taken to obtain a "gold standard" is to use the importance sampling method with a huge sample size. We use the importance sampling method with sample size of one million. This takes quite a long time, but provides us a much-needed standard to compare the values obtained by the different methods. Obtaining the true values of Bayes factors on an Alphastation 500 workstation equipped with a 400MHz 64-bit CPU and a gigabyte of RAM take 620, 421 and 381 minutes respectively. By looking at the variability of the importance ratios for the sampled one million points, we conclude that the Bayes factor can be determined up to a standard error of about 0.5 % for the Bayes factor comparing the "spatial model" to the "full model" and about 0.25 % for the other two Bayes factors. This gives us a benchmark to compare the different methods.
Table 6.3 Estimates of the Bayes factors (along with their standard deviations and time taken to run the program) for comparing different types of variance component models obtained by different methods for the Scotland lip cancer data set

<table>
<thead>
<tr>
<th>Comparing</th>
<th>Method</th>
<th>Estimated BF</th>
<th>std. dev.</th>
<th>CPU time(min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>'spatial model vs Chib-Jeliazkov'</td>
<td>True value</td>
<td>1.44</td>
<td>-</td>
<td>620</td>
</tr>
<tr>
<td>'full model'</td>
<td>Importance sampling</td>
<td>1.44</td>
<td>0.1463</td>
<td>81.1</td>
</tr>
<tr>
<td></td>
<td>Verdinelli-Wasserman</td>
<td>0.44</td>
<td>0.21</td>
<td>7.1</td>
</tr>
<tr>
<td>'heterogen. model' vs Chib-Jeliazkov</td>
<td>True value</td>
<td>0.083</td>
<td>-</td>
<td>421</td>
</tr>
<tr>
<td>'full model'</td>
<td>Importance sampling</td>
<td>0.083</td>
<td>0.0093</td>
<td>57.2</td>
</tr>
<tr>
<td></td>
<td>Verdinelli-Wasserman</td>
<td>0.021</td>
<td>0.012</td>
<td>5.4</td>
</tr>
<tr>
<td>'null model' vs Chib-Jeliazkov</td>
<td>True value</td>
<td>1.15 x 10^{-23}</td>
<td>-</td>
<td>381</td>
</tr>
<tr>
<td>'full model'</td>
<td>Importance sampling</td>
<td>1.21 x 10^{-23}</td>
<td>1.46 x 10^{-24}</td>
<td>48.1</td>
</tr>
<tr>
<td></td>
<td>Verdinelli-Wasserman</td>
<td>1.16 x 10^{-23}</td>
<td>2.81 x 10^{-25}</td>
<td>18.2</td>
</tr>
<tr>
<td></td>
<td>0.0001</td>
<td>0.004</td>
<td>4.9</td>
<td></td>
</tr>
</tbody>
</table>

The application of the Laplace approximation and the Chib's method requires the computation of the marginal likelihood \( L(\beta, \phi, \tau^2, \sigma^2 | y) \) — so we discuss that below.

### 6.5.1 Likelihood computation

From the discussion of Section 6.3, the marginal likelihood \( L(\beta, \phi, \tau^2, \sigma^2 | y) \) is given by

\[
L(\beta, \phi, \tau^2, \sigma^2 | y) \propto \int \left[ \prod_{i=1}^{n} \exp\left(-E_i e^{x_i' \beta + u_i}\right) e^{y_i (x_i' \beta + u_i)} \right] \cdot \frac{1}{|V|^{1/2}} \cdot \exp\left(-\frac{1}{2} \nu' V^{-1} \nu\right) d\nu,
\]

where

\[
V = \tau^2 (I - \phi C)^{-1} M + \sigma^2 D.
\]

We compute the above using importance sampling (Section 4.1.3). We use a \( t_4 \) importance sampling distribution on \( \nu \). The mean and variance of the distribution are taken as the corresponding sample moments of \( \nu = \eta + \psi \) from a posterior sample obtained...
from the joint posterior distribution of \((\eta, \psi)\) conditional on \(y, \beta, \phi, \tau^2, \sigma^2\). The posterior sample is obtained using techniques similar to those discussed in Section 6.3.2—i.e., using a Gibbs sampling algorithm and the conditional distributions of \(\eta\) and \(\psi\) given there.

### 6.5.2 Specific details regarding the implementation of the various methods for this example

**Laplace approximation:** To apply the Laplace approximation (Section 4.3.1), we work on the logarithmic scale for \(\tau^2\) and \(\sigma^2\). We compute the posterior modes required in this method using the importance sampling EM algorithm (Section 4.2.2.2). The posterior modes were found at:

- \((-0.480, 0.059, 0.169, 1.450, 0.000)\)’ for the “full model”
- \((-0.480, 0.059, 0.169, 1.450)\)’ for the “spatial model”
- \((-0.621, 0.064, 1.612)\)’ for the “heterogeneity model” and
- \((-0.542, 0.074)\)’ for the “null model”

To compute the marginal density \(p(y|M)\) of the data under the models (remember that the Bayes factor is obtained as the ratio of the marginal densities of the data under the two models being compared), we use the usual Laplace approximation at the posterior modes for the last three models. We compute the negative of the Hessian matrix of the log-posterior at the posterior mode using techniques similar to those used to derive (4.7). The approximated marginal density obtained by the Laplace approximation is very accurate for the “null model”, but those for the “spatial model” and “heterogeneity model” are away from the true value by factors of 2.6 and 1.9 respectively. These values probably indicate that the Laplace approximation does not perform well for complex models.
However, we cannot use the Laplace approximation for the "full model" as the derivatives of the log-posterior are undefined at the posterior mode under the "full model". We use our suggested variation (4.8) for that model. However, we find it difficult to find a point close to the posterior mode where the information matrix from the log-posterior density is positive definite. Hence we use the posterior mean as the point to apply (4.8) on. The approximation of the marginal density obtained using this method is not satisfactory at all — as a result, the estimates of all the three Bayes factors using this method are far from the true value. Hence, we don’t show the results in the Table 6.3.

Verdinelli-Wasserman method: To apply the Verdinelli-Wasserman method, since apriori there is no dependence between the variance parameters and the fixed effects $\beta$, we can use the simple formula

$$BF = \frac{p(\omega_0|y)}{p(\omega)} ,$$

where $\omega$ is the vector of variance components we are testing. To approximate $p(\omega_0|y)$, we use posterior samples of size 50000 obtained by the MCMC methods (Section 2.3) and then apply kernel density estimation. We use built-in S-plus programs to do the kernel estimation.

For comparing either the "spatial model" or the "heterogeneity model" to the "full model", we need one-dimensional kernel estimation — we use the S-plus function "density" (Venables & Ripley, 1998) to do that. This function implements a kernel-density smoother of the form

$$\hat{f}(t) = \frac{1}{nb} \sum_{i=1}^{n} K\left(\frac{t - t_i}{b}\right)$$

for a sample $t_1, t_2, \ldots, t_n$ with fixed kernel $K(.)$ and a bandwidth $b$. The kernel is usually chosen to be a probability density function; we use the normal density which is the default in the program. The choice of bandwidth is a compromise between smoothing
enough to remove insignificant bumps and smoothing too much to smear out real peaks. We use the choice suggested in Silverman (1986, page 45):

\[
\hat{b} = 4 \times 1.06 \min(s, R/1.34)n^{-\frac{1}{5}},
\]

(6.12)

where \( R \) is the inter-quartile range of the sample and \( s \) is the standard deviation of the sample \( t_1, t_2, \ldots t_n \). We also make sure by looking at the plots of the fitted kernel density for the above bandwidth to ensure that there is no over-fitting or under-fitting.

For comparing the model without any variance component against that with both of them, we need two-dimensional Kernel estimation. We use a built-in S-plus function \( \text{kde} \). The function uses a kernel density estimator of the form

\[
\hat{f}(t) = \frac{1}{nb^2} \sum_{i=1}^{n} K\left(\frac{t - t_i}{b}\right)
\]

for a bivariate sample \( t_1, t_2, \ldots t_n \) and a fixed bivariate kernel \( K(\cdot) \) and a bandwidth \( b \). We use the standard multivariate normal density as a kernel. The bandwidth is chosen using the formula of optimum bandwidth for a bivariate normal kernel given in Silverman (1986, page 47):

\[
\hat{b} = 0.96n^{-\frac{1}{5}}.
\]

(6.13)

As in the univariate case, we look at the contour plot of the fitted kernel density to make sure that the amount of smoothing is appropriate for the above bandwidth.

The values of the estimated Bayes factors obtained by this method are not satisfactory as can be seen from Table 6.3. However, the results improve a lot for the first two Bayes factors considered in Table 6.3 if posterior samples of larger size are used. For example, using posterior samples of size 100000 gives an average estimated Bayes factor of 1.73 for the Bayes factor comparing the “spatial model” versus the “full model”. The average estimate for the Bayes factor comparing the “heterogeneity model” versus the “full model” is 0.10 when posterior samples of size 100000 is used. These estimates are
much closer to the true values of the Bayes factors than the estimates obtained with posterior samples of size 100000.

Chib's method: Because we use Metropolis steps within a Gibbs sampler for all the parameters to generate the posterior sample, the Chib-Jeliazkov method (Section 4.3.5.2) that computes the marginal likelihood from Metropolis output will be applicable here rather than Chib's original Gibbs sampling-based method (Section 4.3.5.1). We have to estimate \( p(\beta, \phi, \tau^2, \sigma^2|y) \) at a particular value of the parameter vector. That distribution is difficult to sample from directly. Instead it is possible to obtain simulations from \( p(\beta, \phi, \tau^2, \sigma^2, \eta_i, i = 1, \ldots, n, \psi_i, i = 1, \ldots, n|y) \), where each of the components is generated from its full conditional distribution using a Metropolis step (Section 2.3.2). When computing the marginal density \( p(y|M) \) for the "full model", we treat \( \eta_i, i = 1, \ldots, n \) and \( \psi_i, i = 1, \ldots, n \) as latent variables. With the "spatial model", \( \eta_i, i = 1, \ldots, n \) are treated as latent variables while for the "heterogeneity model", \( \psi_i, i = 1, \ldots, n \) are treated as latent variables. For the model without any variance components, \( \beta \) is the only parameter block used.

As for the fixed point used in this method, the posterior mode for the full model is found at \((-0.475, 0.059, 0.169, 1.450, 0.000)^T\), that is, on the boundary of the parameter space. The full conditional density \( p(\sigma^2|y, \eta, \psi, \beta, \phi, \tau^2) \) has to be computed with this method at the fixed point, but it is not defined at the posterior mode. So we use the posterior mean as the fixed point to estimate the posterior ordinate. We also compute the marginal density under the 'full model' at a number of points close to the posterior mode — but they don't improve on the standard deviation obtained with the case when the posterior mean is taken as the fixed point.

The sample size is taken as 50000 (after convergence of the Markov chain) for all the models.
Importance sampling: As discussed towards the beginning of Section 6.5, we treat the random effects $\eta$ and $\psi$ as parameters in the variance components models to facilitate the computations.

We use the logarithms of $\tau^2$ and $\sigma^2$ as the parameters in the three variance component models. Under each of the three models other than the full model, the importance sampling density is taken as a $t_4$ distribution with location parameter equal to the posterior mode and variance matrix equal to the inverse of the negative of the Hessian of the log-posterior evaluated at the posterior mode, where we use the techniques described in Section 4.2.2.2 to obtain the posterior mode and the variance matrix.

However, under the "full model", the posterior mode lies on the boundary of the parameter space and the negative of the Hessian of the log-posterior evaluated at the posterior mode is not positive definite. Hence, while computing with the "full model", the importance sampling density is taken as a $t_4$ distribution with location parameter equal to the posterior mean and variance matrix equal to the posterior variance matrix; the posterior moments are obtained from a posterior sample generated under the "full model" using an MCMC algorithm (Section 2.3).

The size of the importance sample is taken as 50000 for all the models.

6.5.3 Comparison of the different methods

Looking at Table 6.3, we see that Chib's method and the importance sampling method (with sample size 50000) provide accurate values of the Bayes factor estimates. The standard deviation of the Bayes factor estimate is much smaller for the importance sampling method than for the Chib's method. Since the importance sampling method takes much less coding effort and considerably less time (less than half as much time as that taken by the Chib's method), this method seems to be the best method to use for computation of the Bayes factor estimate for this data set. This is a bit surprising, keeping in mind other recent work on Bayes factor estimate computation where Chib's
method provides the most accurate result; for example, Han and Carlin (2000, page 27-28), who computed Bayes factor estimates for non-nested linear regression models and hierarchical longitudinal models say that

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

In reconciling these results, there are several factors to consider. The efficiency of the Chib's method is closely related to the efficiency of the underlying MCMC algorithm. Thus, the standard deviation for the Chib's method and its run time may be reduced by reducing the autocorrelation of the generated parameter values in the MCMC algorithm, for example, by the use of the tailored proposal density (Chib and Jeliazkov, 2001).

The Verdinelli-Wasserman method performs poorly for all of the three Bayes factor estimate computations. It is possible that the performance of this method can be improved by considering different kernel density methods, but we do not explore this here.

6.5.4 Interpretation of results

One point to be noted here is that although there is difference in the Bayes factor estimate values obtained by the different methods, the conclusion is the same regardless of the method applied. When we compare the “spatial model” against the “full model”, all the methods give Bayes factor estimates favoring the former model — which suggests that $\sigma^2$ is probably not required in the model. By contrast, the Bayes factor estimate comparing the “heterogeneity model” against the “full model” provides strong evidence for the full model, indicating that $\sigma^2$ alone cannot explain all the variability in the data — hence there is some spatial structure present in the data.
Similarly, Bayes factor estimates obtained by all the methods used for comparing the "null model" against the "full model" give very strong evidence against the former model pointing to the fact that the simple poisson regression model is inadequate for the data — there is extra-poisson variation present.

Hence, among the four variance component models considered for these data, the "spatial" model' seems to be the optimum model. So there seems to be some proof of spatial correlation among the risk of lip cancer incidence in the districts of Scotland.

6.6 Sensitivity to the prior distributions

In Section 5.5, we explored the sensitivity of the Bayes factor estimate calculation to the choice of prior distribution for a generalized linear mixed model with one variance component. When there is more than one variance component, as is the case here, we must consider sensitivity to prior distributions on each variance component. Also, there are several Bayes factors to consider as described towards the beginning of Section 6.5. Here we consider the sensitivity to the prior distribution of the Bayes factor for comparing the "null model" versus the "full model" because the sensitivity of the other Bayes factors (all of which test whether we can drop one variance component from the model) have been covered by the sensitivity analysis we performed in the previous chapter. We denote the "null model" as model 0 and the "full model" as model 1. Let's further denote the Bayes factor for comparing the model 0 and the model 1 for a prior distribution $p(\tau^2, \sigma^2)$ on the variance components as $BF_{p(\cdot)}^{01}$. Then, as described
in Section 4.5, $BF_{p(...)}^{10}$, the reciprocal of $BF_{p(...)}^{01}$, can be expressed as:

$$BF_{p(...)}^{10} = \frac{\int_{\tau^2} \int_{\phi} \int_{\beta} \int_{\eta} \int_{\psi} p(y|\beta, \eta, \psi) p(\eta|\phi, \tau^2) p(\psi|\sigma^2) d\eta d\psi p(\beta) p(\phi) p(\tau^2, \sigma^2) d\beta d\phi d\tau^2 d\sigma^2}{\int_{\beta} p(y|\beta, \eta = 0, \psi = 0) p(\beta) d\beta}$$

$$= \int_{\tau^2} \int_{\phi} \int_{\beta} \int_{\eta} \int_{\psi} p(y|\beta, \eta, \psi) p(\eta|\phi, \tau^2) p(\psi|\sigma^2) d\eta d\psi p(\beta) p(\phi) d\beta d\phi p(\tau^2, \sigma^2) d\tau^2 d\sigma^2$$

$$= \int_{\tau^2} \int_{\phi} BF_{\tau^2, \sigma^2}^{10} \cdot p(\tau^2, \sigma^2) d\tau^2 d\sigma^2,$$

where $BF_{\tau^2, \sigma^2}^{10}$ is the Bayes factor for comparing the model 1 against the model 0 using a degenerate prior at $(\tau^2, \sigma^2)$ under model 1. It can also be interpreted as the Bayes factor for comparing model 1 with variance components fixed at $(\tau^2, \sigma^2)$ against model 0.

The above relation implies that

$$BF_{p(...)}^{01} = \left[ \int_{\tau^2} \int_{\phi} BF_{\tau^2, \sigma^2}^{10} \cdot p(\tau^2, \sigma^2) d\tau^2 d\sigma^2 \right]^{-1}. \quad (6.14)$$

We use the importance sampling method to obtain estimates of $BF_{\tau^2, \sigma^2}^{10}$ for different combinations of $\tau^2$ and $\sigma^2$. Table 6.4 gives the logarithms of Bayes factor estimates for a grid of $(\tau^2, \sigma^2)$.

Figure 6.2 shows a perspective plot of the logarithms of $BF_{\tau^2, \sigma^2}^{10}$ for different $\tau^2$ and $\sigma^2$. Looking at the table and the plot, we observe that $BF_{\tau^2, \sigma^2}^{10}$ reaches its highest value around $\tau^2 = 1.95$ and $\sigma^2 = 0.01$. Figure 6.3 shows a close view of the section of the perspective plot near the maximum value. The maximum likelihood estimates of $\tau^2$ and $\sigma^2$ are 1.64 and 0.00 respectively. The highest value of the Bayes factor estimate is achieved for $\tau^2$ near, but not equal to the corresponding maximum likelihood estimate. If there were no parameters in the model other than $\tau^2$ and $\sigma^2$, then the value of $BF_{\tau^2, \sigma^2}^{10}$ would be maximized at the maximum likelihood estimate. The difference here is due to the fact that there are other parameters. The above results for degenerate priors suggest that the minimum possible value of $BF_{p(...)}^{01}$ is nearly obtained with a prior
Figure 6.2 Plot of $\log(BF_{\tau^2, \sigma^2})$ for different $\tau^2$ and $\sigma^2$
Figure 6.3 Plot of $\log(BF^{10}_{\tau^2,\sigma^2})$ for different $\tau^2$ and $\sigma^2$
Table 6.4 Logarithm of estimated $BF^{01}_{\tau^2,\sigma^2}$'s for different $\tau^2$ and $\sigma^2$ for the Scotland lip cancer data set

<table>
<thead>
<tr>
<th>Values of $\tau^2$</th>
<th>.01</th>
<th>.10</th>
<th>.20</th>
<th>.30</th>
<th>.40</th>
<th>.50</th>
<th>.75</th>
<th>1.0</th>
<th>1.3</th>
<th>1.5</th>
<th>2.0</th>
<th>2.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>5.3</td>
<td>14.4</td>
<td>21.7</td>
<td>27.3</td>
<td>32.0</td>
<td>35.8</td>
<td>42.0</td>
<td>45.7</td>
<td>48.1</td>
<td>49.7</td>
<td>51.2</td>
<td>51.8</td>
</tr>
<tr>
<td>0.10</td>
<td>25.9</td>
<td>30.4</td>
<td>34.3</td>
<td>37.3</td>
<td>39.8</td>
<td>41.7</td>
<td>45.3</td>
<td>47.8</td>
<td>49.5</td>
<td>50.6</td>
<td>51.7</td>
<td>52.0</td>
</tr>
<tr>
<td>0.30</td>
<td>40.3</td>
<td>42.4</td>
<td>44.2</td>
<td>45.7</td>
<td>46.9</td>
<td>47.9</td>
<td>49.8</td>
<td>50.9</td>
<td>51.6</td>
<td>52.1</td>
<td>52.5</td>
<td>52.3</td>
</tr>
<tr>
<td>0.50</td>
<td>46.3</td>
<td>47.5</td>
<td>48.6</td>
<td>49.4</td>
<td>50.1</td>
<td>50.7</td>
<td>51.8</td>
<td>52.4</td>
<td>52.7</td>
<td>52.9</td>
<td>52.9</td>
<td>52.5</td>
</tr>
<tr>
<td>0.75</td>
<td>50.1</td>
<td>50.8</td>
<td>51.4</td>
<td>51.9</td>
<td>52.2</td>
<td>52.5</td>
<td>53.1</td>
<td>53.3</td>
<td>53.4</td>
<td>53.4</td>
<td>53.0</td>
<td>52.4</td>
</tr>
<tr>
<td>1.00</td>
<td>52.2</td>
<td>52.6</td>
<td>52.9</td>
<td>53.1</td>
<td>53.3</td>
<td>53.5</td>
<td>53.7</td>
<td>53.7</td>
<td>53.5</td>
<td>52.9</td>
<td>52.2</td>
<td></td>
</tr>
<tr>
<td>1.25</td>
<td>53.3</td>
<td>53.5</td>
<td>53.7</td>
<td>53.8</td>
<td>53.9</td>
<td>54.0</td>
<td>54.0</td>
<td>53.9</td>
<td>53.7</td>
<td>53.4</td>
<td>52.7</td>
<td>52.0</td>
</tr>
<tr>
<td>1.50</td>
<td>54.0</td>
<td>54.1</td>
<td>54.1</td>
<td>54.3</td>
<td>54.2</td>
<td>54.2</td>
<td>54.0</td>
<td>53.8</td>
<td>53.5</td>
<td>53.2</td>
<td>52.5</td>
<td>51.7</td>
</tr>
<tr>
<td>1.64</td>
<td>54.2</td>
<td>54.2</td>
<td>54.3</td>
<td>54.3</td>
<td>54.2</td>
<td>54.2</td>
<td>54.0</td>
<td>53.7</td>
<td>53.4</td>
<td>53.1</td>
<td>52.3</td>
<td>51.5</td>
</tr>
<tr>
<td>1.80</td>
<td>54.3</td>
<td>54.3</td>
<td>54.3</td>
<td>54.3</td>
<td>54.2</td>
<td>54.1</td>
<td>53.9</td>
<td>53.6</td>
<td>53.3</td>
<td>52.9</td>
<td>52.1</td>
<td>51.2</td>
</tr>
<tr>
<td>1.95</td>
<td>54.4</td>
<td>54.4</td>
<td>54.3</td>
<td>54.3</td>
<td>54.2</td>
<td>54.1</td>
<td>53.8</td>
<td>53.5</td>
<td>53.1</td>
<td>52.7</td>
<td>51.8</td>
<td>51.0</td>
</tr>
<tr>
<td>2.00</td>
<td>54.4</td>
<td>54.4</td>
<td>54.3</td>
<td>54.3</td>
<td>54.2</td>
<td>54.1</td>
<td>53.7</td>
<td>53.4</td>
<td>53.0</td>
<td>52.6</td>
<td>51.8</td>
<td>50.9</td>
</tr>
<tr>
<td>2.20</td>
<td>54.3</td>
<td>54.3</td>
<td>54.2</td>
<td>54.2</td>
<td>54.1</td>
<td>53.9</td>
<td>53.5</td>
<td>53.2</td>
<td>52.8</td>
<td>52.4</td>
<td>51.5</td>
<td>50.6</td>
</tr>
<tr>
<td>2.50</td>
<td>54.2</td>
<td>54.1</td>
<td>53.9</td>
<td>53.8</td>
<td>53.7</td>
<td>53.5</td>
<td>53.2</td>
<td>52.7</td>
<td>52.3</td>
<td>51.9</td>
<td>51.0</td>
<td>50.1</td>
</tr>
<tr>
<td>3.00</td>
<td>53.6</td>
<td>53.4</td>
<td>53.3</td>
<td>53.1</td>
<td>53.0</td>
<td>52.8</td>
<td>52.4</td>
<td>52.0</td>
<td>51.5</td>
<td>51.1</td>
<td>50.1</td>
<td>49.2</td>
</tr>
<tr>
<td>4.00</td>
<td>52.0</td>
<td>51.9</td>
<td>51.7</td>
<td>51.5</td>
<td>51.3</td>
<td>51.1</td>
<td>50.7</td>
<td>50.2</td>
<td>49.7</td>
<td>49.3</td>
<td>48.3</td>
<td>47.4</td>
</tr>
</tbody>
</table>

density concentrated around $\tau^2 = 1.95, \sigma^2 = 0.01$. We compute values of $BF^{01}_{\tau^2,\sigma^2}$ under a set of independent inverse-gamma prior distributions on $\tau^2$ and $\sigma^2$ where we fixed the variance of the prior distributions at 10 and varied the means. The importance sampling method of computing the Bayes factor estimate is used for the computation. Table 6.5 gives the logarithms of the Bayes factor estimate $BF^{01}_{\tau^2,\sigma^2}$ for a grid of posterior means of $\tau^2$ and $\sigma^2$. Figure 6.4 shows a perspective plot of the logarithms of $BF^{01}_{\tau^2,\sigma^2}$. Table 6.5 and Figure 6.4 show that the minimum value of the Bayes factor estimate is obtained when the prior mean of $\tau^2$ is around 3.0 (i.e., the prior mode is around 1.46) and the prior mean (and hence the prior mode) of $\sigma^2$ is close to 0. Figure 6.5 shows a plot of the region near which the Bayes factor estimate reaches its minimum. The prior mode values that minimize the Bayes factor estimate are quite close to the maximum likelihood estimates of these parameters, although not exactly equal to them.
Figure 6.4 Plot of logarithm of $BF_{\phi(\cdot)}^{01}$ for different prior means of $\tau^2$ and $\sigma^2$
Figure 6.5 Plot of logarithm of $\hat{B}F_{\rho}^{01}$ for different prior means of $\tau^2$ and $\sigma^2$
Table 6.5 Logarithm of estimated $B_{F_{p(l)}}^*$'s for different prior means of $\tau^2$ and $\sigma^2$ for the turtles data set

<table>
<thead>
<tr>
<th>Prior means of $\tau^2$</th>
<th>0.01</th>
<th>0.10</th>
<th>0.20</th>
<th>0.30</th>
<th>0.50</th>
<th>0.75</th>
<th>1.00</th>
<th>1.50</th>
<th>2.00</th>
<th>2.50</th>
<th>3.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>-43.8</td>
<td>-45.5</td>
<td>-46.7</td>
<td>-47.2</td>
<td>-47.9</td>
<td>-48.5</td>
<td>-49.2</td>
<td>-50.1</td>
<td>-50.2</td>
<td>-50.5</td>
<td>-50.9</td>
</tr>
<tr>
<td>0.10</td>
<td>-48.3</td>
<td>-48.5</td>
<td>-48.7</td>
<td>-48.8</td>
<td>-49.2</td>
<td>-49.6</td>
<td>-50.1</td>
<td>-50.4</td>
<td>-50.8</td>
<td>-51.0</td>
<td>-51.2</td>
</tr>
<tr>
<td>0.30</td>
<td>-50.4</td>
<td>-50.5</td>
<td>-50.6</td>
<td>-50.8</td>
<td>-50.9</td>
<td>-51.1</td>
<td>-51.3</td>
<td>-51.5</td>
<td>-51.7</td>
<td>-51.7</td>
<td>-51.7</td>
</tr>
<tr>
<td>0.50</td>
<td>-51.3</td>
<td>-51.5</td>
<td>-51.5</td>
<td>-51.6</td>
<td>-51.8</td>
<td>-51.9</td>
<td>-52.0</td>
<td>-52.1</td>
<td>-52.1</td>
<td>-52.2</td>
<td>-52.0</td>
</tr>
<tr>
<td>0.75</td>
<td>-52.0</td>
<td>-52.1</td>
<td>-52.1</td>
<td>-52.2</td>
<td>-52.3</td>
<td>-52.4</td>
<td>-52.4</td>
<td>-52.4</td>
<td>-52.4</td>
<td>-52.4</td>
<td>-52.2</td>
</tr>
<tr>
<td>1.00</td>
<td>-52.5</td>
<td>-52.5</td>
<td>-52.5</td>
<td>-52.6</td>
<td>-52.7</td>
<td>-52.7</td>
<td>-52.8</td>
<td>-52.7</td>
<td>-52.7</td>
<td>-52.6</td>
<td>-52.4</td>
</tr>
<tr>
<td>2.00</td>
<td>-53.3</td>
<td>-53.4</td>
<td>-53.4</td>
<td>-53.3</td>
<td>-53.3</td>
<td>-53.3</td>
<td>-53.1</td>
<td>-53.0</td>
<td>-52.9</td>
<td>-52.7</td>
<td>-52.4</td>
</tr>
<tr>
<td>2.50</td>
<td>-53.6</td>
<td>-53.6</td>
<td>-53.5</td>
<td>-53.5</td>
<td>-53.4</td>
<td>-53.3</td>
<td>-53.3</td>
<td>-53.2</td>
<td>-52.9</td>
<td>-52.7</td>
<td>-52.4</td>
</tr>
<tr>
<td>3.00</td>
<td>-53.6</td>
<td>-53.6</td>
<td>-53.6</td>
<td>-53.6</td>
<td>-53.5</td>
<td>-53.4</td>
<td>-53.3</td>
<td>-53.1</td>
<td>-52.8</td>
<td>-52.4</td>
<td>-52.0</td>
</tr>
<tr>
<td>3.25</td>
<td>-53.6</td>
<td>-53.6</td>
<td>-53.6</td>
<td>-53.5</td>
<td>-53.5</td>
<td>-53.3</td>
<td>-53.2</td>
<td>-52.9</td>
<td>-52.7</td>
<td>-52.3</td>
<td>-51.9</td>
</tr>
<tr>
<td>3.50</td>
<td>-53.6</td>
<td>-53.6</td>
<td>-53.5</td>
<td>-53.5</td>
<td>-53.4</td>
<td>-53.2</td>
<td>-53.1</td>
<td>-52.8</td>
<td>-52.5</td>
<td>-52.2</td>
<td>-51.7</td>
</tr>
<tr>
<td>4.00</td>
<td>-53.5</td>
<td>-53.4</td>
<td>-53.4</td>
<td>-53.3</td>
<td>-53.2</td>
<td>-53.0</td>
<td>-52.9</td>
<td>-52.6</td>
<td>-52.2</td>
<td>-51.8</td>
<td>-51.4</td>
</tr>
<tr>
<td>4.50</td>
<td>-53.3</td>
<td>-53.2</td>
<td>-53.1</td>
<td>-53.0</td>
<td>-52.9</td>
<td>-52.7</td>
<td>-52.6</td>
<td>-52.2</td>
<td>-51.8</td>
<td>-51.3</td>
<td>-51.0</td>
</tr>
<tr>
<td>5.00</td>
<td>-53.0</td>
<td>-52.9</td>
<td>-52.8</td>
<td>-52.7</td>
<td>-52.5</td>
<td>-52.3</td>
<td>-52.2</td>
<td>-51.8</td>
<td>-51.4</td>
<td>-50.9</td>
<td>-50.4</td>
</tr>
</tbody>
</table>

6.7 Simulation study

We proceed to do a simulation study with the objective to find if the Bayes factors can help the statistician to identify the right variance component model. Data sets are generated using different fixed known values of the variance components. Then Bayes factor estimates comparing different variance component models are computed to see which model the Bayes factor favors. This provides some information about whether Bayes factors are useful model selection tools for generalized linear mixed models. The simulation is interesting for the Poisson disease mapping model because the issue of identifiability of parameters for this particular model is an important one (see for example Elberly and Carlin, 2000).

First, data sets are generated from the Poisson-normal model that we have used to analyze the disease incidence data:
\[ y_i | \lambda_i \sim \text{Poisson}(\lambda_i E_i), i = 1, 2, \ldots, n, \]

\[ \log(\lambda) = X\beta + \eta + \psi, \]

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

\[ \psi | \sigma^2 \sim N(0, \sigma^2 D). \]

The \( E_i \) and the definition of \( C, M, D \) are exactly the same as for the Scotland lip cancer data. The values of \( \beta \) and \( \phi \) are fixed at the maximum likelihood estimate and the values of \( \tau^2 \) and \( \sigma^2 \) are varied to obtain the following four situations:

- \( \tau^2 \) and \( \sigma^2 \) both small (0.01 each)
- \( \tau^2 \) and \( \sigma^2 \) both large (1 each)
- \( \tau^2 \) large (1.5) and \( \sigma^2 \) small (0.01)
- \( \tau^2 \) small (0.01) and \( \sigma^2 \) large (1.5)

For each of the above situations, we generate 30 data sets and compute the Bayes factors comparing each of the three reduced models

- "null model"
- "heterogeneity model"
- "spatial model"

against the full model. Table 6.6 shows the Bayes factor estimates obtained — each of columns 2-4 giving Bayes factor estimates for comparing one of the above reduced models against the "full model".

The second column indicates that the "null model" is preferred to the "full model" only when both variance components are extremely small. If one or both variance components are large, then the null model is clearly rejected.
Table 6.6 Estimated Bayes factors ($BF^{01}$) for comparing different types of variance component models for the simulated data sets similar to the Scotland lip cancer data set

<table>
<thead>
<tr>
<th>True values of ($\tau^2, \sigma^2$)</th>
<th>Model compared to the “full model”</th>
<th>“null model”</th>
<th>“heterogeneity model”</th>
<th>“spatial model”</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0.01, 0.01)</td>
<td>69</td>
<td>9.1</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>(1.00, 1.00)</td>
<td>$7 \times 10^{-29}$</td>
<td>0.09</td>
<td>0.43</td>
<td></td>
</tr>
<tr>
<td>(1.50, 0.01)</td>
<td>$1 \times 10^{-21}$</td>
<td>0.08</td>
<td>0.65</td>
<td></td>
</tr>
<tr>
<td>(0.01, 1.50)</td>
<td>$7 \times 10^{-14}$</td>
<td>0.24</td>
<td>0.46</td>
<td></td>
</tr>
</tbody>
</table>

The third column reports Bayes factor estimates for testing whether the spatial random effect term is needed (with small values indicating that it is). If $\tau^2$ is large, then the Bayes factor estimate indicates that the spatial effect is clearly needed (Bayes factor estimate = 0.09 or 0.08). If both variance components are small, then the spatial effect is not needed (Bayes factor estimate = 9.1). The interesting case here is the last row where $\tau^2 = 0.01, \sigma^2 = 1.50$. Even though the “heterogeneity model” should provide an adequate fit, the Bayes factor estimate suggests that the “full model” should be preferred (though the evidence is not strong).

For testing whether the heterogeneity random effects are needed, the Bayes factor estimate gives the right conclusion if the true value of both $\tau^2$ and $\sigma^2$ is small — a Bayes factor estimate of 18 favors the “null model”. The Bayes factor estimates for the other scenarios are approximately the same irrespective of the true values of $\sigma^2$ (and $\tau^2$). In each case, the “full model” is slightly preferred to the “spatial model”. This result is especially surprising for the final scenario in which the data generated are consistent with the “heterogeneity model”.

All these simulations appear to raise questions about the disease mapping model rather than answer questions about the Bayes factors. In the Stern and Cressie (1995) model, the “spatial model” seems to provide a better fit than the “heterogeneity model” even when the “heterogeneity model” is the true model. This may be due to the fact
that the "heterogeneity model" is included as a special case of the spatial model; it corresponds to the case $\phi = 0$. The question is then whether including both effects is ever justified. The results in Table 6.6 suggest perhaps not. There is a slight preference for the "full model" when both effects are present, but the Bayes factor estimate is less than 3 in that case.
7 CONCLUSIONS

Generalized linear mixed models are applied often and their use is likely to increase with the widespread availability of fast computational facilities. In many applications, people are interested to test hypotheses regarding variance components. Our objective in this work is to investigate the problem of testing for the variance component(s) for generalized linear mixed models. We mainly focus on the calculation of Bayes factors for comparing two models. Our work focusses on two specific generalized linear mixed model examples, a probit model with a single variance component and a Poisson model with multiple variance components.

7.1 Inference

The very first step in doing parametric inference, the computation of the likelihood (also known as marginal likelihood) for a given value of the parameters, is a difficult task for generalized linear mixed models. The main reason is that to compute the likelihood, one needs to integrate out the random effects, which can not be done analytically for these models. If the model is simple and the data set is small, one can use numerical integration (e.g., Simpson's rule) to integrate out the random effects. However, numerical integration becomes very time-consuming for large problems. Importance sampling is an alternative approach. We find that the importance sampling with a $t_4$ distribution seems to work well in computation of the likelihood.

Bayesian analysis of generalized linear mixed models does not necessarily require
evaluating the likelihood and in this sense can be easier than for example maximum likelihood estimation. One can easily implement an MCMC algorithm to draw a sample from the posterior distribution of the parameters. There is the software BUGS available for obtaining a posterior sample from a wide variety of generalized linear mixed models.

Optimization plays a key role in inference for generalized linear mixed models. Maximum likelihood estimation requires that we maximize the likelihood. In addition, some Bayes factor computation approaches require the posterior mode. In both cases, optimization is complicated by the fact that the likelihood is difficult to compute. While the use of numerical integration to integrate out the random effects followed by the use of the Newton-Raphson method works for simple models and small data sets, anything else requires some kind of approximation or EM algorithm. The Monte Carlo EM algorithm using an importance sampling approach (Booth & Hobert, 1999) seems to work well for the complex situations that we consider.

7.2 Choosing a method to compute the Bayes factor estimate

The focus of our work was to study how the different methods for computing Bayes factor estimates perform for generalized linear mixed models. The key results are:

- The Laplace approximation is a fast way to obtain an approximate value of the Bayes factor for simple models — this method does reasonably well for the simple probit model example. The relative error for the estimate of the marginal densities obtained with this approximation is $O\left(\frac{1}{n}\right)$, which is not high even for moderate sample sizes. However, this method cannot be applied when the posterior mode for a model is on the boundary of the parameter space; this was the case with the Poisson-normal model. For the same example, the Laplace approximation does not give good estimate of marginal density for any of the variance component models either. Hence, this method cannot be trusted when applied to complex models.
• If we need an accurate estimate of the Bayes factor, the importance sampling method can be used to obtain the Bayes factor estimate since this method seems to work well for both of our examples. This method gives very accurate estimates of the Bayes factor with small standard deviations. Chib's method works well, but yielded higher standard deviations for our complex example.

• Chib's method works well and is the least sensitive to the size of the problem. It can be used even with a larger number of random effects. The use of sophisticated computational techniques, e.g., the tailored proposal density (Chib and Jeliazkov, 2001) may be used to improve the convergence rate of the Markov chain, which will in turn reduce the variability of the Bayes factor estimate for a given number of iterations.

• Reversible jump MCMC method appears to be a very elegant method to find Bayes factor estimates. This method has the advantage that if we are comparing more than one models at the same time, we can get the Bayes factor estimate for comparing any model against any other as an end-result of only one Markov chain using this method. This method however involves many likelihood computations. We may avoid the likelihood computation by treating the random effects as parameters of the model. In that case, it is difficult to construct jumping distributions, especially for jumping from one model to another.

• Verdinelli-Wasserman method, although quite easy to implement, does not seem to be very accurate for complex models. The main difficulty appears to lie with the multivariate density estimation required to obtain the estimate of the Bayes factor.
7.3 Prior distribution

The Bayes factor is sensitive to the prior distribution used for the variance component(s), even when we have moderately large sample size. Hence, one should be very careful about the choice of the prior distribution for the variance component(s). Section 4.5 describes an approach for assessing sensitivity of the Bayes factor for variance component testing to the prior distributions used for the variance components in the model. That approach is applied to the examples in Sections 5.5 and 6.6. If there is no prior information available about the problem, one should compute the Bayes factor estimate for a range of prior distributions and base the conclusion on the range of the estimates obtained.

7.4 Future work

Though we have looked at two data examples, a simple and a more complex one, we have not considered any really large data sets yet in our work. Questions about the relative merit of the importance sampling method and the Chib’s method may be answered only by working with large data sets.

Commercial software for implementing the methods we discussed here is very difficult to obtain. Other than GLIMMIX (whose estimates may not have satisfactory properties) and BUGS (which can be very time-consuming for large data sets), there is nothing available right now. The programs in this dissertation are specific to the applications. One future need is software for inference with generalized linear mixed models that will address model selection.
BIBLIOGRAPHY


ACKNOWLEDGEMENTS

This research was partly supported by National Institutes of Health award CA78169.

Professor Hal Stern has been my advisor for about three and half years and has overseen the research for my masters degree, my Ph.D. degree and my research assistantship. I highly appreciate his guidance, patience and generosity with his time during these years. I learned a great deal about how to think critically about complicated statistical topics and also about different computational issues of statistics. His wonderful sense of humor has always made my discussions with him very enjoyable.

I am grateful to Dr. Michael Daniels for his useful advice from time to time throughout my work about different topics and for serving as a member on my Ph.D. committee.

I thank Dr. Frederic Janzen of the Department of Zoology and Genetics, Iowa State University for providing the data from the natural selection study and for serving as a member on my Ph.D. committee.

I also thank Dr. Siddhartha Chib, Dr. Xiao Li Meng, Dr. Xihong Lin, Dr. Brad Carlin and Dr. Donna Pauler for their useful suggestions at different times during the work.

Appreciation also goes to the other members of my Ph.D. committee, Professors Kenneth Koehler and Mark Kaiser, to Professor Dean Isaacson, head of the Department of Statistics at Iowa State University and to many other faculty members in the statistics department at Iowa State University for their help.

The staff members of the Department of Statistics at Iowa State University have always been nice, helpful and tolerant with every need that I ever had and I would like
to thank them profusely for making the department a better place to live. I would like to especially thank Rose Ann Andserson, Jeanette La Grange, Edith Landin and Sandie Smith for their generosity in helping me whenever I needed them.

I am grateful to my family, especially my parents, Radhanath and Pushpamita Sinharay who have helped me to achieve whatever I have in my life. I would not become the person I am without their care, genuine interest, concern and prayers in all aspects of my life.

I am also grateful to my would-be wife Lopamudra Mukherjee, her brother Prabuddha Mukherjee and their parents, Sushanta and Bela Mukherjee because the company of these wonderful people has always been a very positive influence in every aspect of my life. They make the world a much better place for me.

I would like to extend thanks to my friends in my department who have helped me in different ways. Anindya Roy and Pradipta Sarkar (both of whom were students in the department when I joined and graduated later) always provided advice about my studies, encouragements and much needed distractions. Deanne Reber, another ex-student of the department, also helped me a lot with important advice from time to time.

These acknowledgements have been a difficult portion of my dissertation to write because there are many different people who have contributed to my life and education in some way. I appreciate the many concerns, wishes and prayers from my relatives and friends from different aspects of my life.