BAYESIAN ANALYSIS OF LONGITUDINAL MODELS

SYEDA TASMINE HUSAIN

Your file Votre référence
ISBN: 0-612-89634-X
Our file Notre référence
ISBN: 0-612-89634-X

Bayesian Analysis of Longitudinal Models

by

©Syeda Tasmine Husain

A Practicum report submitted to the School of Graduate Studies in partial fulfillment of the requirement for the Degree of Master of Applied Statistics

Department of Mathematics and Statistics
Memorial University of Newfoundland

January, 2003

St. John's Newfoundland Canada

Abstract

The use of longitudinal studies is widespread, especially in biology and medicine. Statistical analyses of these studies must account for the correlation that will usually be present within individuals measured across time. We present a Bayesian approach to studying these problems, based on methods that sample from the posterior distributions of interest. Our work will involve models with continuous and binary responses, and will generalize some published methods using a probit model. Our results indicate that the simpler algorithms proposed in the literature perform as well as more complicated methods. Application to two numerical examples will be presented.

Acknowledgements

I am grateful to my supervisor Dr. G. Sneddon for his encouragement, continuous guidance and helpful assistance in completing this practicum report. It was indeed a great privilege to work on this important problem in the area of longitudinal data analysis, which was suggested by Dr. G. Sneddon.

I also thank Dr. Paul Peng and Dr. David Schneider for taking the time to serve as examiners of this Thesis. Their constructive comments and suggestions served to improve the quality of the Thesis.

I sincerely acknowledge the financial support provided by the School of Graduate Studies and Department of Mathematics and Statistics in the form of Graduate Fellowships and Teaching Assistantships. Further I wish to thank Dr. Herb Gaskill, Department Head, for providing me a friendly atmosphere and the necessary facilities to complete the program.

I am also grateful to my husband, my parents, brother for their eternal love, emotional support and encouragement during this program.

It is my great pleasure to thank my friends and well-wishers who directly or indirectly encouraged and helped me in the MAS program and contributed to this dissertation.

Contents

Abstract ii

Acknowledgements iii

List of Tables vii

List of Figures x

1 Introduction 1

1.1 Introduction 1

1.2 Background 3

1.2.1 Acceptance-Rejection sampling 4

1.2.2 Metropolis-Hastings (M-H) algorithm 5

2 Longitudinal Models-Theory 10

2.1 Introduction 10

2.2 Model for Continuous Data 10

2.3 Estimation Methods 11

2.3.1 Algorithm 1 11

2.3.2 Algorithm 2 18

2.3.3 Algorithm 3 21

2.4 Model for Binary Data 23

2.4.1 Algorithm 4 24

2.4.2 Algorithm 5 25
2.5 Conclusion 26

3 Simulation Studies **27**
3.1 Introduction 27
3.2 Simulation Design and Generation of the Continuous Data 27
3.3 Simulation Results-Continuous Data 28
3.3.1 Comparison of Algorithms 30
3.3.2 Graphs 30
3.4 Simulation Design and Generation of the Binary Data 32
3.5 Simulation Results-Binary Data 32
3.5.1 Comparison of Algorithms 34
3.5.2 Graphs 34
3.6 Autocorrelations of Posterior Estimates 35
3.6.1 Results on Algorithms 1-3 36
3.6.2 Results on Algorithms 4 and 5 37
3.7 Conclusions 37

4 Continuous Data: Example **51**
4.1 CD4+ Data 51
4.2 Results 52
4.2.1 Comparison of Algorithms 1-3 55
4.2.2 Autocorrelations of Posterior Estimates 56

5 Binary Data: Example **61**
5.1 Six Cities data set: child's wheeze status 61
5.2 Results 62
5.2.1 Comparison of Algorithms 4 and 5 65
5.2.2 Autocorrelations of Posterior Estimates 65

Bibliography 68

List of Tables

3.1 Posterior means and variances of parameters in simulations using Algorithm 1. 29

3.2 Posterior means and variances of parameters in simulations using Algorithm 2. 30

3.3 Posterior means and variances of parameters in simulations using Algorithm 3. 31

3.4 Posterior means and variances of parameters in simulations using Algorithm 4. 33

3.5 Posterior means and variances of parameters in simulations using Algorithm 5. 33

3.6 Lag-1 autocorrelations of posterior estimates under Case 2, using continuous data. 35

3.7 Lag-1 autocorrelations of posterior estimates under Case 3, using continuous data. 36

3.8 Estimates of $\kappa = 1 + 2\sum_{k=1}^{\infty} \rho(k)$ for posterior estimates, Case 2, for continuous data. 37

3.9 Estimates of $\kappa = 1 + 2\sum_{k=1}^{\infty} \rho(k)$ for posterior estimates, Case 3, for continuous data. 38

3.10 Lag-1 autocorrelations of posterior estimates under Case 2, using binary data. 39

3.11 Lag-1 autocorrelations of posterior estimates under Case 3, using binary data. 39
3.12 Estimates of $\kappa = 1 + 2\sum_{k=1}^{\infty}\rho(k)$ for posterior estimates, Case 2, for binary data. 50
3.13 Estimates of $\kappa = 1 + 2\sum_{k=1}^{\infty}\rho(k)$ for posterior estimates, Case 3, for binary data. 50

4.1 Posterior means and variances of parameters in analysis of CD4+ Data set, using Algorithm 1. 53
4.2 Posterior means and variances of parameters in analysis of CD4+ Data set, using Algorithm 2. 54
4.3 Posterior means and variances of parameters in analysis of CD4+ Data set, using Algorithm 3. 55
4.4 Lag-1 autocorrelations of posterior estimates under Case 1, using CD4+ Data set. 56
4.5 Lag-1 autocorrelations of posterior estimates under Case 2, using CD4+ Data set. 57
4.6 Lag-1 autocorrelations of posterior estimates under Case 3, using CD4+ Data set. 58
4.7 Lag-1 autocorrelations of posterior estimates under Case 4, using CD4+ Data set. 58
4.8 Estimates of $\kappa = 1 + 2\sum_{k=1}^{\infty}\rho(k)$ for posterior estimates, Case 1, for CD4+ Data set. 59
4.9 Estimates of $\kappa = 1 + 2\sum_{k=1}^{\infty}\rho(k)$ for posterior estimates, Case 2, for CD4+ Data set. 59
4.10 Estimates of $\kappa = 1 + 2\sum_{k=1}^{\infty}\rho(k)$ for posterior estimates, Case 3, for CD4+ Data set. 60
4.11 Estimates of $\kappa = 1 + 2\sum_{k=1}^{\infty}\rho(k)$ for posterior estimates, Case 4, for CD4+ Data set. 60

5.1 Posterior means and variances of parameters in analysis of Six Cities data set, using Algorithm 4. 63
5.2 Posterior means and variances of parameters in analysis of Six Cities data set, using Algorithm 5. 64
5.3 Lag-1 autocorrelations of posterior estimates, using Six Cities data set. 65
5.4 Lag-1 autocorrelations of posterior estimates, using Six Cities data set. 66
5.5 Estimates of $\kappa = 1 + 2\sum_{k=1}^{\infty} \rho(k)$ for posterior estimates, for Six Cities data set. 67
5.6 Estimates of $\kappa = 1 + 2\sum_{k=1}^{\infty} \rho(k)$ for posterior estimates, for Six Cities data set. 67

List of Figures

3.1 Posterior Distributions of β estimates in Algorithm 1, Case 1. 40
3.2 Posterior Distributions of σ^2 and the elements of $\mathbf{D}$ estimates in Algorithm 1, Case 1. 41
3.3 Posterior Distributions of β estimates in Algorithm 2, Case 3. 42
3.4 Posterior Distributions of σ^2 and the elements of $\mathbf{D}$ estimates in Algorithm 2, Case 3. 43
3.5 Posterior Distributions of β estimates in Algorithm 3, Case 2. 44
3.6 Posterior Distributions of σ^2 and the elements of $\mathbf{D}$ estimates in Algorithm 3, Case 2. 45
3.7 Posterior Distributions of β estimates iu Algorithm 4, Case 3. 46
3.8 Posterior Distributions of σ^2 and the elements of $\mathbf{D}$ estimates in Algorithm 4, Case 3. 47
3.9 Posterior Distributions of β estimates in Algorithm 5, Case 2. 48
3.10 Posterior Distributions of σ^2 and the elements of $\mathbf{D}$ estimates in Algorithm 5, Case 2. 49

Chapter 1

Introduction

1.1 Introduction

In longitudinal studies, repeated observations of a response variable and a set of covariates are made on individuals across occasions. Because repeated observations are made on the same individual, the response variables will usually be autocorrelated. In analysing longitudinal data, this dependence must be accounted for in order to make correct inference.

Longitudinal studies have applications to a wide variety of problems. Now we introduce two such data sets which have been chosen from the biological and health sciences to represent a range of challenges for analysis. These are described in more detail by Diggle, Liang and Zeger (1994).

First, we discuss the growth of Sitka spruce trees. The study objective is to assess the effect of ozone pollution on tree growth. As ozone pollution is common in urban areas, the impact of increased ozone concentrations on tree growth is of considerable interest. The response variable is log tree size, where size is conventionally measured by the product of tree height and diameter squared. The trees were measured 13 times over two growing seasons.

As a second example, we consider data on the protein content of milk. In this

data set, milk was collected weekly from 79 Australian cows and analysed for its protein content. The cows were maintained on one of three diets: barley, a mixture of barley and lupins, or lupins alone. The objective of the study is to determine how diet affects the protein content in milk.

A wide variety of approaches to analyzing longitudinal data have been introduced in the statistical literature. In studies where the response is normally distributed, Laird and Ware (1982) and Lindstrom and Bates (1988) discuss non-Bayesian methods of analysis. These focus on the use of a mixed effects model and the use of the EM (Expectation-Maximization) algorithm. In cases where the response is binary, Fitzmaurice and Laird (1993) describe a likelihood approach, based on the conditional odds-ratios. Also with binary outcomes, Chib and Greenberg (1998) found maximum likelihood estimates by using a Monte Carlo-based EM algorithm.

However, the structure of longitudinal studies lends itself to the use of Bayesian methods, hierarchical models in particular. An advantage of a Bayesian approach is it can help avoid difficult numerical integrations that may be needed to evaluate likelihoods. These integrations are often avoided through the use of Markov Chain Monte Carlo (MCMC) methods. For example, Chib and Jeliazkov (2001) and Chib and Carlin (1999) present MCMC based methods for continuous data. Zeger and Karim (1991) implement the Gibbs sampler for generalized linear models. Albert and Chib (1993) use the Gibbs sampler to study binary longitudinal data, and Chib and Greenberg (1998) implement an MCMC method for finding the posterior estimates when using binary data.

This practicum will investigate a number of Bayesian algorithms for the analysis of continuous and binary longitudinal data. We exploit an identity used by Chib (1995) in the context of Bayes factor computation to show how the parameters in a generalized linear mixed model may be updated in a single block, improving convergence and producing essentially independent draws from the posterior of the parameters of interest. We also investigate the value of blocking in a class of binary response data longitudinal models. The theoretical aspects of these algorithms, along with the

derivation of the needed posterior distributions, will be discussed in Chapter 2. In Chapter 3, we will present some simulation studies to investigate the behaviour of the algorithms under a variety of assumptions on the prior distributions. In Chapter 4, we will study a dataset of CD4+ cell numbers along with other variables collected longitudinally for AIDS infected men. The objective will be to determine what variables are useful in predicting the CD4+ cell count. In Chapter 5, we will study a binary longitudinal dataset involving a child's wheeze status (yes, no) as well as information about maternal smoking. The objective will be to determine the effects of age, maternal smoking and the age-maternal smoking interaction on the wheeze status.

We begin with some background to these problems, which will include a discussion on some MCMC procedures.

1.2 Background

We begin by assuming there is a prior distribution on the parameters of interest, denoted as $\pi(\boldsymbol{\theta})$. Then, combining this with the density function of the data, written as $f(\mathbf{y}|\boldsymbol{\theta})$, we can derive the posterior distribution using Bayes Theorem:

$$\begin{aligned}\pi(\boldsymbol{\theta}|\mathbf{y}) &= \frac{\pi(\boldsymbol{\theta})f(\mathbf{y}|\boldsymbol{\theta})}{f(\mathbf{y})} \\ &= \frac{\pi(\boldsymbol{\theta})f(\mathbf{y}|\boldsymbol{\theta})}{\int \pi(\boldsymbol{\theta})f(\mathbf{y}|\boldsymbol{\theta})d\boldsymbol{\theta}} \end{aligned} \tag{1.1}$$

Here, $\boldsymbol{\theta}$ is the parameter of interest in our study. The functions $\pi(\boldsymbol{\theta})$ and $\pi(\boldsymbol{\theta}|\mathbf{y})$ represent our belief or information about the parameter $\boldsymbol{\theta}$. As we can see, the data is used to update our prior belief on the behaviour of $\boldsymbol{\theta}$.

In (1.1), the evalution of the integral is often very difficult, if not impossible, so we may not be able to express $\pi(\boldsymbol{\theta}|\mathbf{y})$ in closed form. We need to approximate $\pi(\boldsymbol{\theta}|\mathbf{y})$, typically by simulating an approximate random sample from the posterior distribution. In general, this can be thought of as a MCMC procedure (Chib and

Greenberg, 1995). This sampling can be done in different ways. We will discuss two methods: the Acceptance-rejection sampling and the Metropolis-Hastings algorithm (Chib and Greenberg, 1995).

1.2.1 Acceptance-Rejection sampling

Classical simulation techniques generate non-Markov (usually independent) samples; i.e., the successive observations generated are statistically independent unless correlation is artificially introduced as a variance reduction device. An important method in this class is the Acceptance-Rejection (A-R) method, which can be described as follows. Suppose it is desired to generate samples from the target density $\pi(x)$, where x may be a vector. The method may be used when $\pi(x)$ is known only up to a multiplicative constant and can be expressed as $\pi(x) = f(x)/K$, where $f(x)$ is the unnormalized density and K the (possibly unknown) normalizing constant. This is similar to (1.1), where the integral in (1.1) can be considered the normalizing constant. Let $h(x)$ be a density that can be simulated by some known method, and suppose there is a known constant c such that $f(x) \leq ch(x)$ for all x. This means that $ch(x)$ blankets, or dominates $f(x)$. Then to obtain a random variate from $\pi(.)$, we do the following:

STEP 1: Generate a candidate Z from $h(.)$ and a value u from $U(0,1)$, the uniform distribution on (0,1);

STEP 2: Return $Z = y$ if $u \leq f(Z)/ch(Z)$; otherwise go to STEP 1.

It can be shown (Chib and Greenberg, 1995) that the accepted value y is a random variate from $\pi(.)$. For this method to be efficient c must be carefully selected, and since the expected number of iterations of steps 1 and 2 to obtain a draw is given by c^{-1}, the rejection method is optimized by setting

$$c = \sup_x \frac{f(x)}{h(x)};$$

even this choice, however, may result in an undesirably large number of rejections.

This means we would have to run the A-R method for many more than c^{-1} iterations to generate a reasonably-sized sample from $\pi(x)$.

1.2.2 Metropolis-Hastings (M-H) algorithm

The notion of a generating density also appears in the M-H algorithm, but before considering the differences and similarities we turn to the rationale behind MCMC methods.

The usual approach to Markov chain theory is to start with a transition matrix p_{ij} (when there are a discrete set of states and $\sum_j p_{ij} = 1$) or a transition kernel $p(x, y)$ (when the set of states is not discrete and $\int p(x, y)dy = 1$). A major concern of the theory is to determine conditions under which there exists an invariant distribution and conditions under which iterations of the transition matrix or kernel converge to the invariant distribution. In the discrete case an invariant distribution π_j for the p_{ij} is a distribution with the property $\pi_j = \sum_i p_{ij}\pi_i$, and the nth iterate of p_{ij} is defined recursively as $p_{ij}^{(n)} = \sum_k p_{ik}^{(n-1)} p_{kj}$. When the number of states is finite, it is well known that the matrix of the probability distribution of the nth iterate is given by the nth power of the matrix composed of the p_{ij}. In the nondiscrete case, the invariant distribution $\pi(y)$ satisfies $\pi(y) = \int p(x, y)\pi(x)dx$, and the nth iterate is given by $p^{(n)}(x, y) = \int p^{(n-1)}(x, z)p(z, y)dz$, where $p^{(1)}(x, y) = p(x, y)$. Under certain conditions it can be shown that the nth iterate converges to the invariant distribution as $n \to \infty$ in both the descrete and the nondiscrete cases.

MCMC methods turn the theory around: the invariant distribution is known-it is $\pi(.)$, the target density from which samples are desired-but the transition kernel is unknown. To generate samples from $\pi(.)$, the methods find and utilize a transition kernel $p(x, y)$ whose nth iterate converges to $\pi(.)$ for large n. The process is started at an arbitrary x and iterated a large number of times. After this large number which is problem-dependent, the observations generated from the simulation can be regarded as observations from the target density. The problem then is to find an

appropriate $p(x, y)$. Although this sounds difficult, the search is somewhat simplified by the following observation. Suppose $p(x, y)$ is a density for a given x; i.e. $p(x, y) > 0$ and $\int p(x, y)dy = 1$. Then a $p(x, y)$ that satisfies the reversibility condition,

$$\pi(x)p(x, y) = \pi(y)p(y, x) \tag{1.2}$$

has $\pi(.)$ as its invariant distribution. Note that

$$\int \pi(x)p(x, y)dx = \int \pi(y)p(y, x)dx = \pi(y) \int p(y, x)dx = \pi(y).$$

Intuitively, the left-hand side of the reversibility condition (1.2) is the unconditional probability of moving from x to y, where x is generated from $\pi(.)$, and the right-hand side of (1.2) is the unconditional probability of moving from y to x, where y is also generated from $\pi(.)$. The reversibility condition says that the two sides are equal, and the above result shows that $\pi(.)$ is then the invariant distribution for $p(.,.)$.

We now have a sufficient condition to be satisfied by $p(x, y)$, but we still need to find a specific transition density. We get one from the Metropolis-Hastings algorithm, which we now proceed to describe by exploiting the logic of reversibility.

The Metropolis-Hastings (M-H) algorithm was developed by Metropolis et al. (1953) and widely used by physicists. It was refined and introduced to statisticians by Hastings (1970); Tierney (1994) and Müller (1993) present theory and examples on the use of the M-H algorithm for exploring posterior distributions.

As in the A-R method, suppose we have a density that can generate candidates from our posterior. Since we are dealing with Markov chains, however, we permit that density to depend on the current state of the process. Accordingly, the candidate-generating density is denoted $q(x, y)$, where $\int q(x, y)dy = 1$. This density is to be interpreted as saying that when a process is at the point x, the density generates a value y from $q(x, y)$. If it happens that $q(x, y)$ itself satisfies the reversibility condition for all (x, y), our search is over. But most likely it will not. We might find, for example, that for some x and y,

$$\pi(x)q(x, y) > \pi(y)q(y, x) \tag{1.3}$$

In this case, the process moves from x to y too often and from y to x too rarely. A convenient way to correct this condition is to reduce the number of moves from x to y by introducing a probability $0 < \alpha(x, y) < 1$ that the move is made. We refer to $\alpha(x, y)$ as the probability of a move. If the move is not made, the process again returns x as a value from the target distribution. This contrasts with the A-R method in which, when a y is rejected, a new pair (y, u) is drawn independently of the previous value of y. Then

$$q(x, y)\alpha(x, y), \quad x \neq y,$$

can be regarded as a transition density, but we still need to determine $\alpha(x, y)$.

Consider again inequality (1.3). It tells us that the movement from y to x is not made often enough. We should therefore define $\alpha(y, x)$ to be as large as possible, and since it is a probability, $\alpha(y, x)$ is set equal to 1. But now the probability of move $\alpha(x, y)$ is determined: Set $p(x, y) = q(x, y)\alpha(x, y)$ and obtain from the reversibility condition

$$\begin{aligned} \pi(x)p(x, y) &= \pi(y)p(y, x) \\ \pi(x)q(x, y)\alpha(x, y) &= \pi(y)q(y, x)\alpha(y, x) \\ \pi(x)q(x, y)\alpha(x, y) &= \pi(y)q(y, x); \end{aligned}$$

hence, if $\pi(x)q(x, y) > \pi(y)q(y, x)$, set $\alpha(x, y) = \pi(y)q(y, x)/\pi(x)q(x, y)$. Of course, if the inequality in (1.3) is reversed, set $\alpha(x, y) = 1$ and determine $\alpha(y, x)$ as above. The probabilities $\alpha(x, y)$ and $\alpha(y, x)$ are thus introduced to ensure that the two sides of (1.3) are in balance or, in other words, that the modified candidate-generating density satisfies reversibility.

To complete the defiuition of $p(x, y)$ given above, a small technicality must be considered. Because there is a nonzero probability that the process remains at x (i.e., $p(x, x) \neq 0$), a density function is inadequate to represent all the transitions. But this problem is easily solved. The probability that the process remains at x is given

by

$$r(x) = 1 - \int q(x,y)\alpha(x,y)dy.$$

Let $\delta_x(y) = 1$ if $x = y$ and 0 otherwise, and define $q(x,x)\alpha(x,x) = 0$. Then we can define

$$p(x,y) = q(x,y)\alpha(x,y) + r(x)\delta_x(y). \tag{1.4}$$

We have thus written the transition kernel as the sum of a reversible term and a term that places nonzero probability at the value x. The result presented in (1.2) that reversibility implies invariance can be generalized to expression (1.4); see Tierney (1994).

To summarize, the probability of a move is

$$\alpha(x,y) = \begin{cases} \min\left[\frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}, 1\right] & \text{if } \pi(x)q(x,y) > 0; \\ 1 & \text{otherwise.} \end{cases}$$

Several important points should be noted. First, the calculation of $\alpha(x,y)$ does not require knowledge of the normalizing constant of $\pi(.)$, since it appears both in the numerator and denominator. Second, in the important special case where the candidate-generating density is symmetric, i.e. $q(x,y) = q(y,x)$, the acceptance probability reduces to $\pi(y)/\pi(x)$; hence, if $\pi(y) \geq \pi(x)$, the chain moves to y, otherwise it moves with probability given by $\pi(y)/\pi(x)$.

We now summarize the M-H algorithm initialized with the (arbitrary) value $x^{(0)}$:

Repeat for $j = 1, 2, ..., N$.

STEP 1: Generate y from $q(x^{(j)}, .)$ and U from $U(0,1)$.

STEP 2: Let $x^{(j+1)} = y$ if $U \leq \alpha(x^{(j)}, y)$; otherwise let $x^{(j+1)} = x^{(j)}$.

Return the values $\{x^{(n_0+1)}, x^{(2)}, ..., x^{(N)}\}$.

As in any MCMC method, the draws are regarded as a sample from the target density $\pi(x)$ only after the chain has passed the transient stage and converged to the target. For this reason, the first n_0 values of the chain are ignored. This is sometimes referred to as the burn-in period. There are many different ways to monitor the behavior of the output to determine approximately the values of n_0 and N. One simple

idea is to make n_0 and N an increasing function of the first-order serial correlation in the output. However, the specifics of the sampling design usually have little effect on such summaries, such as the mean and standard deviation, calculated from the sampled values.

As Chib and Greenberg (1995) discuss, there are a number of choices available for $q(x, y)$. In many cases a normal or t-distribution, with appropriate tuning parameters chosen, will work reasonably well. In choosing the generating density, it is desirable to have it dominate the target density in the tails of its distribution.

In the M-H algorithm, the spread of the candidate-generating density affects the behavior of the chain in at least two dimensions: one is the "staying rate" (the percentage of times a move to a new point is not made) and the other is the region of the sample space that is covered by the chain. To see why, consider the situation in which the chain has converged and the density is being sampled around the mode. Then, if the spread is extreme, some of the generated candidates will be far from the current value and will therefore have a low probability of being accepted. Reducing the spread will correct this problem. But if the spread is chosen too small, the chain will take longer to cover the support of the density, and low probability regions will be under-sampled. Both of these situations are likely to be reflected in high autocorrelations across sample values. For these reasons, the candidate-generating density should be tuned so that the staying rate is about 50%. If the chain still displays high autocorrelations, it is usually necessary to try a different class of candidate-generating densities.

Chapter 2

Longitudinal Models-Theory

2.1 Introduction

In this chapter we will introduce models that are appropriate for the analysis of continuous and binary longitudinal data. We will discuss a hierarchical Bayesian structure for these models, and derive the needed posterior distributions for analysis.

2.2 Model for Continuous Data

Consider the Gaussian linear mixed model (Laird and Ware, 1982),

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{W}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i \tag{2.1}$$

where the $\mathbf{y}_i$ are vectors of length n_i containing the observations on the i^{th} unit, and the $\boldsymbol{\epsilon}_i$ are error vectors of the same length independently distributed as $N_{n_i}(0, \sigma^2\mathbf{I}_{n_i})$, $i = 1, ..., n$. Therefore, we have a total of n subjects. In this mixed model, $\mathbf{X}_i$ is an $n_i \times p$ design matrix of covariates and $\boldsymbol{\beta}$ is a corresponding $p \times 1$ vector of fixed effects. In addition, $\mathbf{W}_i$ is a $n_i \times q$ design matrix (q typically less than p), and $\mathbf{b}_i$ is a $q \times 1$ vector of subject-specific random effects. In a non-Bayesian setting, we would usually assume that $\mathbf{b}_i \sim N(\mathbf{0}, \mathbf{D})$, where $\mathbf{D}$ is the covariance matrix of $\mathbf{b}_i$ and $\mathbf{b}_i$ is

independent of $\boldsymbol{\epsilon}_i$, which implies the mean and variance of $\mathbf{y}_i$ are

$$\begin{aligned} E(\mathbf{y}_i) &= \mathbf{X}_i\boldsymbol{\beta} \\ Var(\mathbf{y}_i) &= Var(\mathbf{W}_i\mathbf{b}_i) + Var(\boldsymbol{\epsilon}_i) \\ &= \mathbf{W}_i\mathbf{D}\mathbf{W}_i' + \sigma^2\mathbf{I}_{n_i} \end{aligned}$$

However, we will conduct a Bayesian analysis of the model (2.1), which means we must specify prior distributions for the parameters. The hierarchical specification of this model is completed by adding the prior distributions $\boldsymbol{\beta} \sim N_p(\boldsymbol{\beta}_o, \mathbf{B}_o)$ and $\sigma^2 \sim IG(\nu_o/2, \delta_o/2)$, where IG denotes the inverse gamma distribution with parameters ν_o and δ_o. We also assume that $\mathbf{b}_i$ is a random effects term, where $\mathbf{b}_i \sim N_q(\mathbf{0}, \mathbf{D})$, $\mathbf{D}^{-1} \sim W_q(\rho_o, \mathbf{R}_o\rho_o^{-1})$ and W denotes the Wishart distribution with parameters ρ_o and $\mathbf{R}_o$. We note that the parameter values in these prior distributions need to be specified.

2.3 Estimation Methods

In this section, we discuss Bayesian methods for simulating samples from the posterior distributions, based on Chib and Carlin (1999).

2.3.1 Algorithm 1

The Gaussian linear mixed model (2.1) lends itself to a full Bayesian analysis by Markov chain Monte Carlo (MCMC) methods. One of the first such algorithms was proposed by Gelfand and Smith(1990) which we summarize as follows, and refer to as Algorithm 1:

1. Sample $\boldsymbol{\beta}$ from $\boldsymbol{\beta}|\mathbf{y}, \mathbf{b}, \sigma^2, \mathbf{D}$

2. Sample $\mathbf{b}$ from $\{\mathbf{b}_i\}|\mathbf{y}, \boldsymbol{\beta}, \sigma^2, \mathbf{D}$

3. Sample $\mathbf{D}^{-1}$ from $\mathbf{D}^{-1}|\mathbf{y}, \boldsymbol{\beta}, \mathbf{b}, \sigma^2$

4. Sample σ^2 from $\sigma^2|\mathbf{y}, \boldsymbol{\beta}, \mathbf{b}, \mathbf{D}$

5. Repeat Steps 1-4 using the most recent values of the conditioning variables.

Under the assumptions of (2.1) and the specified priors, we can derive explicit expressions for the posterior distributions in this algorithm, which we do below:

Posterior of $\boldsymbol{\beta}$

We begin with the posterior distribution of $(\boldsymbol{\beta}|\mathbf{y}, \mathbf{b}, \sigma^2, \mathbf{D})$ where the prior distribution of $\boldsymbol{\beta}$ is:

$$\Pi(\boldsymbol{\beta}) = \frac{1}{(\sqrt{2\pi})^p \mid \mathbf{B}_o \mid^{1/2}} \exp[-1/2(\boldsymbol{\beta} - \boldsymbol{\beta}_o)'\mathbf{B}_o^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_o)]$$

Given values of $\mathbf{b}, \boldsymbol{\beta}, \sigma^2, \mathbf{D}$, we know that,

$$\begin{aligned} E[\mathbf{y}_i \mid \mathbf{b}_i, \boldsymbol{\beta}, \sigma^2, \mathbf{D}] &= \mathbf{X}_i\boldsymbol{\beta} + \mathbf{W}_i\mathbf{b}_i \\ V[\mathbf{y}_i \mid \mathbf{b}_i, \boldsymbol{\beta}, \sigma^2, \mathbf{D}] &= \sigma^2\mathbf{I}_{n_i} \\ \mathbf{y}_i \mid (\mathbf{b}_i, \boldsymbol{\beta}, \sigma^2, \mathbf{D}) &\sim N(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{W}_i\mathbf{b}_i, \sigma^2\mathbf{I}_{n_i}) \end{aligned}$$

Therefore, we also know that,

$$\begin{aligned} f(\mathbf{y} \mid \mathbf{b}_1, ..., \mathbf{b}_n, \boldsymbol{\beta}, \sigma^2, \mathbf{D}) &= \prod_{i=1}^{n} \frac{1}{(\sqrt{2\pi})^{n_i} \mid \sigma^2\mathbf{I}_{n_i} \mid^{1/2}} \\ &\quad \times \exp[-1/2(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{W}_i\mathbf{b}_i)'\sigma^{-2}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{W}_i\mathbf{b}_i)] \end{aligned}$$

The posterior distribution of $\boldsymbol{\beta}$ can be found using Bayes Theorem:

$$\begin{aligned} \Pi(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{b}_1, ..., \mathbf{b}_n, \sigma^2, \mathbf{D}) &\propto \Pi(\boldsymbol{\beta}) f(\mathbf{y} \mid \mathbf{b}_1, ..., \mathbf{b}_n, \boldsymbol{\beta}, \sigma^2, \mathbf{D}) \\ &= \frac{1}{(\sqrt{2\pi})^p \mid \mathbf{B}_o \mid^{1/2}} \exp[-1/2(\boldsymbol{\beta} - \boldsymbol{\beta}_o)'\mathbf{B}_o^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_o)] \\ &\quad \times \prod_{i=1}^{n} \frac{1}{(\sqrt{2\pi})^{n_i} \mid \sigma^2\mathbf{I}_{n_i} \mid^{1/2}} \\ &\quad \times \exp[-1/2(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{W}_i\mathbf{b}_i)'\sigma^{-2}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{W}_i\mathbf{b}_i)] \end{aligned}$$

$$\begin{aligned} \propto \quad & \exp[-1/2(\boldsymbol{\beta}'\mathbf{B}_o^{-1}\boldsymbol{\beta} - 2\boldsymbol{\beta}_o'\mathbf{B}_o^{-1}\boldsymbol{\beta})] \\ & \times \exp[-1/2(-2\sum_{i=1}^{n}\mathbf{y}_i'\sigma^{-2}\mathbf{X}_i\boldsymbol{\beta} + \sum_{i=1}^{n}\boldsymbol{\beta}'\mathbf{X}_i'\sigma^{-2}\mathbf{X}_i\boldsymbol{\beta} \\ & +2\sum_{i=1}^{n}\mathbf{b}_i'\mathbf{W}_i'\sigma^{-2}\mathbf{X}_i\boldsymbol{\beta})] \end{aligned}$$

Next, we collect terms that involve $\boldsymbol{\beta}$:

$$\begin{aligned} \Pi(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{b}_1, ..., \mathbf{b}_n, \sigma^2, \mathbf{D}) \quad \propto \quad & \exp[-1/2(\boldsymbol{\beta}'(\mathbf{B}_o^{-1} + \sum_{i=1}^{n}\mathbf{X}_i'\sigma^{-2}\mathbf{X}_i)\boldsymbol{\beta} \\ & -2(\boldsymbol{\beta}_o'\mathbf{B}_o^{-1} + \sum_{i=1}^{n}\mathbf{y}_i'\sigma^{-2}\mathbf{X}_i + \sum_{i=1}^{n}\mathbf{b}_i'\mathbf{W}_i'\sigma^{-2}\mathbf{X}_i)\boldsymbol{\beta})] \end{aligned} \tag{2.2}$$

From (2.2), we see that the exponent is

$$\boldsymbol{\beta}'(\mathbf{B}_o^{-1} + \sum_{i=1}^{n}\mathbf{X}_i'\sigma^{-2}\mathbf{X}_i)\boldsymbol{\beta} - 2(\boldsymbol{\beta}_o'\mathbf{B}_o^{-1} + \sum_{i=1}^{n}\mathbf{y}_i'\sigma^{-2}\mathbf{X}_i + \sum_{i=1}^{n}\mathbf{b}_i'\mathbf{W}_i'\sigma^{-2}\mathbf{X}_i)\boldsymbol{\beta} \tag{2.3}$$

This is really a quadratic function of $\boldsymbol{\beta}$. Now, define

$$\begin{aligned} \mathbf{B}_k \quad &= \quad (\mathbf{B}_o^{-1} + \sum_{i=1}^{n}\mathbf{X}_i'\sigma^{-2}\mathbf{X}_i)^{-1} \\ \mathbf{a}_i \quad &= \quad (\boldsymbol{\beta}_o'\mathbf{B}_o^{-1} + \sum_{i=1}^{n}\mathbf{y}_i'\sigma^{-2}\mathbf{X}_i + \sum_{i=1}^{n}\mathbf{b}_i'\mathbf{W}_i'\sigma^{-2}\mathbf{X}_i)' \end{aligned}$$

We use $\mathbf{B}_k$ and $\mathbf{a}_i$ and complete the square in (2.3) to find:

$$\boldsymbol{\beta}'\mathbf{B}_k^{-1}\boldsymbol{\beta} - 2\boldsymbol{\beta}'\mathbf{a}_i \quad = \quad (\boldsymbol{\beta} - \mathbf{B}_k\mathbf{a}_i)'\mathbf{B}_k^{-1}(\boldsymbol{\beta} - \mathbf{B}_k\mathbf{a}_i) - \mathbf{a}_i'\mathbf{B}_k\mathbf{a}_i \tag{2.4}$$

Based on (2.2) and (2.4) we can now say the posterior of distribution of $\boldsymbol{\beta} \mid (\mathbf{y}, \mathbf{b}_1, ..., \mathbf{b}_n, \sigma^2, \mathbf{D})$ can be expressed as

$$\Pi(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{b}, \sigma^2, \mathbf{D}) \quad \propto \quad \exp[-1/2(\boldsymbol{\beta} - \mathbf{B}_k\mathbf{a}_i)'\mathbf{B}_k^{-1}(\boldsymbol{\beta} - \mathbf{B}_k\mathbf{a}_i)] \tag{2.5}$$

But (2.5), up to a multiplicative constant, is the density of a Gaussion distribution. So,

$$\boldsymbol{\beta} \mid (\mathbf{y}, \mathbf{b}, \sigma^2, \mathbf{D}) \quad \sim \quad N(\mathbf{B}_k\mathbf{a}_i, \mathbf{B}_k) \tag{2.6}$$

Posterior of $\mathbf{b}_i$

The prior distribution of $\mathbf{b}_i$ is:

$$\Pi(\mathbf{b}_i) = \frac{1}{(\sqrt{2\pi})^q \mid \mathbf{D} \mid^{1/2}} \exp[-1/2(\mathbf{b}_i'\mathbf{D}^{-1}\mathbf{b}_i)]$$

We also know that

$$\begin{aligned} f(\mathbf{y}_i \mid \mathbf{b}_i, \boldsymbol{\beta}, \sigma^2, \mathbf{D}) &= \frac{1}{(\sqrt{2\pi})^{n_i} \mid \sigma^2\mathbf{I}_{n_i} \mid^{1/2}} \\ &\quad \times \exp[-1/2(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{W}_i\mathbf{b}_i)'\sigma^{-2}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{W}_i\mathbf{b}_i)] \end{aligned}$$

The posterior distribution of $\mathbf{b}_i$ can be written as

$$\begin{aligned} \Pi[\mathbf{b}_i \mid (\mathbf{y}, \boldsymbol{\beta}, \sigma^2, \mathbf{D})] &\propto \Pi(\mathbf{b}_i) f(\mathbf{y}_i \mid \mathbf{b}_i, \boldsymbol{\beta}, \sigma^2, \mathbf{D}) \\ &= \frac{1}{(\sqrt{2\pi})^q \mid \mathbf{D} \mid^{1/2}} \exp[-1/2(\mathbf{b}_i'\mathbf{D}^{-1}\mathbf{b}_i)] \frac{1}{(\sqrt{2\pi})^{n_i} \mid \sigma^2\mathbf{I}_{n_i} \mid^{1/2}} \\ &\quad \times \exp[-1/2(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{W}_i\mathbf{b}_i)'\sigma^{-2}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{W}_i\mathbf{b}_i)] \\ &\propto \exp[-1/2(\mathbf{b}_i'\mathbf{D}^{-1}\mathbf{b}_i - \mathbf{y}_i'\sigma^{-2}\mathbf{W}_i\mathbf{b}_i + \boldsymbol{\beta}'\mathbf{X}_i'\sigma^{-2}\mathbf{W}_i\mathbf{b}_i \\ &\quad -\mathbf{b}_i'\mathbf{W}_i'\sigma^{-2}\mathbf{y}_i + \mathbf{b}_i'\mathbf{W}_i'\sigma^{-2}\mathbf{X}_i\boldsymbol{\beta} + \mathbf{b}_i'\mathbf{W}_i'\sigma^{-2}\mathbf{W}_i\mathbf{b}_i)] \end{aligned}$$

As in the derivation of the posterior distribution of $\boldsymbol{\beta}$, we collect the terms involving $\mathbf{b}_i$:

$$\begin{aligned} \Pi(\mathbf{b}_i \mid (\mathbf{y}, \boldsymbol{\beta}, \sigma^2, \mathbf{D})) &\propto \exp[-1/2[\mathbf{b}_i'(\mathbf{D}^{-1} + \mathbf{W}_i'\sigma^{-2}\mathbf{W}_i)\mathbf{b}_i \\ &\quad -(\mathbf{y}_i'\sigma^{-2}\mathbf{W}_i - \boldsymbol{\beta}'\mathbf{X}_i'\sigma^{-2}\mathbf{W}_i)\mathbf{b}_i - \mathbf{b}_i'(\mathbf{W}_i'\sigma^{-2}\mathbf{y}_i - \mathbf{W}_i'\sigma^{-2}\mathbf{X}_i\boldsymbol{\beta})]] \\ &= \exp[-1/2(\mathbf{b}_i'(\mathbf{D}^{-1} + \mathbf{W}_i'\sigma^{-2}\mathbf{W}_i)\mathbf{b}_i \\ &\quad -2\mathbf{b}_i'(\mathbf{W}_i'\sigma^{-2}\mathbf{y}_i - \mathbf{W}_i'\sigma^{-2}\mathbf{X}_i\boldsymbol{\beta}))] \end{aligned} \tag{2.7}$$

Now we define,

$$\begin{aligned} \mathbf{H}_k &= (\mathbf{D}^{-1} + \mathbf{W}_i'\sigma^{-2}\mathbf{W}_i)^{-1} \\ \mathbf{a}_i &= (\mathbf{W}_i'\sigma^{-2}\mathbf{y}_i - \mathbf{W}_i'\sigma^{-2}\mathbf{X}_i\boldsymbol{\beta}) \end{aligned}$$

We use $\mathbf{H}_k$ and $\mathbf{a}_i$ to rewrite (2.7) by completing the square:

$$\mathbf{b}_i'\mathbf{H}_k^{-1}\mathbf{b}_i - 2\mathbf{b}_i'\mathbf{a}_i = (\mathbf{b}_i - \mathbf{H}_k\mathbf{a}_i)'\mathbf{H}_k^{-1}(\mathbf{b}_i - \mathbf{H}_k\mathbf{a}_i) - \mathbf{a}_i'\mathbf{H}_k\mathbf{a}_i \tag{2.8}$$

Based on (2.7) and (2.8) we can say the posterior distribution of $\mathbf{b}_i \mid (\mathbf{y}, \boldsymbol{\beta}, \sigma^2, \mathbf{D})$ can be written as:

$$\Pi(\mathbf{b}_i \mid (\mathbf{y}, \boldsymbol{\beta}, \sigma^2, \mathbf{D})) \propto \exp[-1/2[(\mathbf{b}_i - \mathbf{H}_k\mathbf{a}_i)'\mathbf{H}_k^{-1}(\mathbf{b}_i - \mathbf{H}_k\mathbf{a}_i)]] \tag{2.9}$$

However, (2.9) is proportional to the density of a Gaussian distribution, so we can say

$$\begin{aligned} \mathbf{b}_i \mid (\mathbf{y}, \boldsymbol{\beta}, \sigma^2, \mathbf{D}) &\sim N_q(\mathbf{H}_k\mathbf{a}_i, \mathbf{H}_k) \\ &= N_q(\mathbf{H}_k\mathbf{W}_i'\sigma^{-2}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}), \mathbf{H}_k) \end{aligned} \tag{2.10}$$

Posterior of $\mathbf{D}^{-1}$

Next, we derive the posterior distribution of $\mathbf{D}^{-1}|(\mathbf{y}, \boldsymbol{\beta}, \mathbf{b}, \sigma^2)$ where the prior distribution of $\mathbf{D}^{-1}$ is (Muirhead, 1982, p. 85):

$$\begin{aligned} \Pi(\mathbf{D}^{-1}) &= \frac{|\mathbf{D}^{-1}|^{(\rho_o - q - 1)/2}\exp[-1/2[tr(\rho_o\mathbf{R}_o^{-1}\mathbf{D}^{-1})]]}{2^{(\rho_o q)/2}\pi^{q(q-1)/4}|\mathbf{R}_o/\rho_o|^{\rho_o/2}\prod_{i=1}^{q}\Gamma(\rho_o + 1 - i)/2} \\ &\propto |\mathbf{D}^{-1}|^{(\rho_o - q - 1)/2}\exp[-1/2[tr(\rho_o\mathbf{R}_o^{-1}\mathbf{D}^{-1})]]|\mathbf{R}_o/\rho_o|^{-\rho_o/2} \end{aligned} \tag{2.11}$$

The prior distribution of $\mathbf{b}_i$ is:

$$\begin{aligned} \Pi(\mathbf{b}_i) &\propto \frac{1}{|\mathbf{D}|^{1/2}}\exp[-1/2(\mathbf{b}_i'\mathbf{D}^{-1}\mathbf{b}_i)] \\ &= |\mathbf{D}|^{-1/2}\exp[-1/2(\mathbf{b}_i'\mathbf{D}^{-1}\mathbf{b}_i)] \end{aligned}$$

Then, assuming $\mathbf{b}_1, ...\mathbf{b}_n$ are independent, their joint prior distribution is:

$$\Pi(\mathbf{b}_1, ...\mathbf{b}_n) \propto |\mathbf{D}|^{-n/2}\exp[-1/2(\sum_{i=1}^{n}\mathbf{b}_i'\mathbf{D}^{-1}\mathbf{b}_i)] \tag{2.12}$$

We can rewrite the exponent in (2.12) as follows:

$$\begin{aligned}
\sum_{i=1}^{n} \mathbf{b}_i' \mathbf{D}^{-1} \mathbf{b}_i &= tr[\sum_{i=1}^{n} \mathbf{b}_i' \mathbf{D}^{-1} \mathbf{b}_i] \quad \text{Since } \mathbf{b}_i' \mathbf{D}^{-1} \mathbf{b}_i \text{ is scalar} \\
&= \sum_{i=1}^{n} tr(\mathbf{b}_i' \mathbf{D}^{-1} \mathbf{b}_i) \\
&= \sum_{i=1}^{n} tr(\mathbf{D}^{-1} \mathbf{b}_i \mathbf{b}_i') \quad \text{Since } tr(AB) = tr(BA) \\
&= tr[\mathbf{D}^{-1} (\sum_{i=1}^{n} \mathbf{b}_i \mathbf{b}_i')]
\end{aligned}$$

Therefore, (2.12) becomes

$$\Pi(\mathbf{b}_1, ...\mathbf{b}_n) \propto |\mathbf{D}|^{-n/2} \exp[-1/2(tr(\mathbf{D}^{-1}(\sum_{i=1}^{n} \mathbf{b}_i \mathbf{b}_i')))] \tag{2.13}$$

We now consider the joint posterior distribution of $\mathbf{D}^{-1}$ and $(\mathbf{b}_1, ...\mathbf{b}_n)$, using (2.11) and (2.13):

$$\begin{aligned}
\Pi(\mathbf{D}^{-1}, \mathbf{b}_1, ...\mathbf{b}_n | \mathbf{y}, \boldsymbol{\beta}, \sigma^2) \propto & |\mathbf{D}^{-1}|^{(\rho_o - q - 1)/2} \exp[-1/2[tr(\rho_o \mathbf{R}_o^{-1} \mathbf{D}^{-1})]] |\mathbf{R}_o/\rho_o|^{-\rho_o/2} \\
& \times |\mathbf{D}|^{-n/2} \exp[-1/2(tr\mathbf{D}^{-1}(\sum_{i=1}^{n} \mathbf{b}_i \mathbf{b}_i'))] \\
& \times f(\mathbf{y}|\mathbf{D}, \boldsymbol{\beta}, \mathbf{b}_1, ...\mathbf{b}_n, \sigma^2) \\
= & |\mathbf{D}^{-1}|^{(\rho_o - q - 1)/2} |\mathbf{D}|^{-n/2} |\mathbf{R}_o/\rho_o|^{-\rho_o/2} \\
& \times \exp[-1/2[tr(\rho_o \mathbf{R}_o^{-1} \mathbf{D}^{-1} + \mathbf{D}^{-1}(\sum_{i=1}^{n} \mathbf{b}_i \mathbf{b}_i'))]] \\
& \times f(\mathbf{y}|\mathbf{D}, \boldsymbol{\beta}, \mathbf{b}_1, ...\mathbf{b}_n, \sigma^2) \\
= & |\mathbf{D}^{-1}|^{(\rho_o + n - q - 1)/2} |\mathbf{R}_o/\rho_o|^{-\rho_o/2} \\
& \times \exp[-1/2[tr\mathbf{D}^{-1}(\rho_o \mathbf{R}_o^{-1} + \sum_{i=1}^{n} \mathbf{b}_i \mathbf{b}_i')]] \\
& \times f(\mathbf{y}|\mathbf{D}, \boldsymbol{\beta}, \mathbf{b}_1, ...\mathbf{b}_n, \sigma^2)
\end{aligned}$$

where $f(\mathbf{y}|\mathbf{D}, \boldsymbol{\beta}, \mathbf{b}_1, ...\mathbf{b}_n, \sigma^2)$ can be found using the fact that $\mathbf{y}_i \mid (\mathbf{b}_i, \boldsymbol{\beta}, \sigma^2, \mathbf{D}) \sim N(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{W}_i\mathbf{b}_i, \sigma^2\mathbf{I}_{n_i})$.

Then, if we also condition on $\mathbf{b}_1, ...\mathbf{b}_n$, we find the posterior distribution of $\mathbf{D}^{-1}$:

$$\Pi(\mathbf{D}^{-1}|(\mathbf{y}, \mathbf{b}_1, ...\mathbf{b}_n, \boldsymbol{\beta}, \sigma^2) \propto |\mathbf{D}^{-1}|^{(\rho_o + n - q - 1)/2}$$

$$\exp[-1/2[tr(\mathbf{D}^{-1}(\rho_o \mathbf{R}_o^{-1} + \sum_{i=1}^{n} \mathbf{b}_i \mathbf{b}_i'))]] \quad (2.14)$$

If we compare (2.14) to (2.11), we see that the posterior distribution of $\mathbf{D}^{-1}$ has the form of a Wishart distribution. In particular,

$$\mathbf{D}^{-1} \sim W_q(n + \rho_o, (\rho_o \mathbf{R}_o^{-1} + \sum_{i=1}^{n} \mathbf{b}_i \mathbf{b}_i')^{-1}) \quad (2.15)$$

Posterior of σ^2

Finally, we derive the posterior distribution of σ^2 where the prior distribution of σ^2 is:

$$\Pi(\sigma^2) \quad \propto \quad (\sigma^{-2})^{(\nu_0/2)+1} \exp[-\delta_0/2\sigma^2] \quad (2.16)$$

Given (2.16) and the distribution of $\mathbf{y}_i$ given earlier, we can express the posterior distribution of σ^2 as

$$\begin{aligned}
\Pi(\sigma^2 \mid (\mathbf{y}, \boldsymbol{\beta}, \mathbf{b}_i, \mathbf{D})) \quad \propto \quad & (\sigma^{-2})^{(\nu_0/2)+1} \exp[-\delta_0/2\sigma^2] \\
& \times \prod_{i=1}^{n} \frac{1}{\mid \sigma^2 \mathbf{I}_{n_i} \mid^{1/2}} \exp[-1/2(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{W}_i\mathbf{b}_i)'\sigma^{-2} \\
& (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{W}_i\mathbf{b}_i)] \\
= \quad & (\sigma^{-2})^{(\nu_0/2)+1} \prod_{i=1}^{n} \mid \sigma^2 \mathbf{I}_{n_i} \mid^{-1/2} \\
& \times \exp[-1/\sigma^2((\delta_0/2) + (\sum_{i=1}^{n} ||\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{W}_i\mathbf{b}_i||^2)/2)] \\
= \quad & (\sigma^{-2})^{(\nu_0/2)+1} \prod_{i=1}^{n} \sigma^{2(-n_i/2)} \\
& \times \exp[-1/\sigma^2((\delta_0/2) + (\sum_{i=1}^{n} ||\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{W}_i\mathbf{b}_i||^2)/2)] \\
= \quad & (\sigma^{-2})^{(\nu_0/2) + (\sum_{i=1}^{n} n_i/2) + 1} \\
& \times \exp[-1/\sigma^2((\delta_0/2) + (\sum_{i=1}^{n} ||\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{W}_i\mathbf{b}_i||^2)/2)]
\end{aligned}$$

$$
\begin{aligned}
&= (\sigma^{-2})^{((\nu_0 + \sum_{i=1}^{n} n_i)/2) + 1} \\
&\quad \times \exp[-1/\sigma^2((\delta_0/2) \\
&\quad +(\sum_{i=1}^{n} ||\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{W}_i\mathbf{b}_i||^2)/2)] \qquad (2.17)
\end{aligned}
$$

If we compare (2.17) to (2.16), we see the posterior distribution of σ^2 has the form of an inverse gamma distribution. In particular, we can say,

$$
\sigma^2 \mid (\mathbf{y}, \boldsymbol{\beta}, \mathbf{b}_i, \mathbf{D}) \sim IG[(\nu_0 + \sum_{i=1}^{n} n_i)/2, (\delta_0 + \sum_{i=1}^{n} ||\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{W}_i\mathbf{b}_i||^2)/2] \qquad (2.18)
$$

Based on our derivations, we note that all the priors are conjugate priors.

2.3.2 Algorithm 2

It is recognized that Algorithm 1 is relatively easy to implement, but it can suffer from slow convergence if the parameters are highly correlated, or if the information in the likelihood and prior is insufficient to completely determine the model parameters (Chib and Carlin, 1999). For this reason, we now describe a new Algorithm, denoted as Algorithm 2.

Algorithm 2 is identical to Algorithm 1 except for the change in the sampling of $\boldsymbol{\beta}$. This minor refinement can be important, however, and improves the behavior of the MCMC output. Besides, it requires no hierarchical centering because $\boldsymbol{\beta}$ is sampled without conditioning on the random effects and the entire sampling is still from tractable distributions.

Algorithm 2 relies on the use of blocking for Gaussian mixed models. We begin our investigation into the value of blocking in longitudinal models by considering the distribution of $\mathbf{y}_i$ marginalized over the random effects. Due to the conditional Gaussian structure we know that,

$$
\mathbf{y}_i|\boldsymbol{\beta}, \mathbf{D}, \sigma^2 \sim N_{n_i}(\mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Omega}_i) \qquad (2.19)
$$

where $\mathbf{\Omega}_i = \sigma^2\mathbf{I}_{n_i} + \mathbf{W}_i\mathbf{D}\mathbf{W}_i'$. This implies that the posterior distribution of $\boldsymbol{\beta}$ will be conditioned on σ^2 and $\mathbf{D}$ (but not on $\mathbf{b}_i$) (Lindley and Smith, 1972). It is possible to sample the fixed effects $\boldsymbol{\beta}$ and the random effects $\mathbf{b}_i$ in one block, but retain the essential Gibbs structure, as follows. We will denote this as Algorithm 2:

1. Sample $\boldsymbol{\beta}$ and $\mathbf{b}$ from $\boldsymbol{\beta}, \{\mathbf{b}_i\}|\mathbf{y}, \sigma^2, \mathbf{D}$ by sampling
   (a) $\boldsymbol{\beta}$ from $\boldsymbol{\beta}|\mathbf{y}, \sigma^2, \mathbf{D}$
   (b) $\mathbf{b}$ from $\{\mathbf{b}_i\}|\mathbf{y}, \boldsymbol{\beta}, \sigma^2, \mathbf{D}$
2. Sample $\mathbf{D}^{-1}$ from $\mathbf{D}^{-1}|\mathbf{y}, \boldsymbol{\beta}, \mathbf{b}, \sigma^2$
3. Sample σ^2 from $\sigma^2|\mathbf{y}, \boldsymbol{\beta}, \mathbf{b}, \mathbf{D}$
4. Repeat Steps 1-3 using the most recent values of the conditioning variables.

Now, we will derive the posterior distributions for this algorithm.

Posterior of $\boldsymbol{\beta}$

As before, the prior distribution of $\boldsymbol{\beta}$ is:

$$\Pi(\boldsymbol{\beta}) = \frac{1}{(\sqrt{2\pi})^p \mid \mathbf{B}_o \mid^{1/2}} \exp[-1/2(\boldsymbol{\beta} - \boldsymbol{\beta}_o)'\mathbf{B}_o^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_o)]$$

We know from (2.19) that,

$$f(\mathbf{y} \mid \boldsymbol{\beta}, \mathbf{D}, \sigma^2) = \prod_{i=1}^{n} \frac{1}{(\sqrt{2\pi})^{n_i} \mid \mathbf{\Omega}_i \mid^{1/2}} \exp[-1/2(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})'\mathbf{\Omega}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})] \tag{2.20}$$

Then the posterior distribution of $\boldsymbol{\beta}$ is,

$$\begin{aligned} \Pi(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{D}, \sigma^2) &\propto \Pi(\boldsymbol{\beta}) f(\mathbf{y} \mid \boldsymbol{\beta}, \mathbf{D}, \sigma^2) \\ &= \frac{1}{(\sqrt{2\pi})^p \mid \mathbf{B}_o \mid^{1/2}} \exp[-1/2(\boldsymbol{\beta} - \boldsymbol{\beta}_o)'\mathbf{B}_o^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_o)] \end{aligned}$$

$$
\begin{aligned}
& \times \prod_{i=1}^{n} \frac{1}{(\sqrt{2\pi})^{n_i} \mid \Omega_i \mid^{1/2}} \exp[-1/2(\mathbf{y}_i - \mathbf{X}_i\beta)'\Omega_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\beta)] \\
\propto \; & \exp[-1/2(\beta - \beta_o)'\mathbf{B}_o^{-1}(\beta - \beta_o)] \\
& \prod_{i=1}^{n} \exp[-1/2(\mathbf{y}_i - \mathbf{X}_i\beta)'\Omega_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\beta)] \\
= \; & \exp[-1/2(\beta'\mathbf{B}_o^{-1}\beta - \beta_o'\mathbf{B}_o^{-1}\beta - \beta'\mathbf{B}_o^{-1}\beta_o + \beta_o'\mathbf{B}_o^{-1}\beta_o)] \\
& \times \exp[-1/2(\sum_{i=1}^{n} \mathbf{y}_i'\Omega_i^{-1}\mathbf{y}_i - \sum_{i=1}^{n} \beta'\mathbf{X}_i'\Omega_i^{-1}\mathbf{y}_i \\
& - \sum_{i=1}^{n} \mathbf{y}_i'\Omega_i^{-1}\mathbf{X}_i\beta + \sum_{i=1}^{n} \beta'\mathbf{X}_i'\Omega_i^{-1}\mathbf{X}_i\beta)] \\
= \; & \exp[-1/2(\beta'\mathbf{B}_o^{-1}\beta - 2\beta_o'\mathbf{B}_o^{-1}\beta + \beta_o'\mathbf{B}_o^{-1}\beta_o)] \\
& \times \exp[-1/2(\sum_{i=1}^{n} \mathbf{y}_i'\Omega_i^{-1}\mathbf{y}_i - 2\sum_{i=1}^{n} \mathbf{y}_i'\Omega_i^{-1}\mathbf{X}_i\beta + \beta' \sum_{i=1}^{n} \mathbf{X}_i'\Omega_i^{-1}\mathbf{X}_i\beta)]
\end{aligned}
$$

As in the previous derivation of the posterior of β, we collect terms involving β. This yields

$$
\begin{aligned}
\Pi(\beta \mid \mathbf{y}, \mathbf{D}, \sigma^2) \; \propto \; & \exp[-1/2(\beta'(\mathbf{B}_o^{-1} + \sum_{i=1}^{n} \mathbf{X}_i'\Omega_i^{-1}\mathbf{X}_i)\beta \\
& -2(\beta_o'\mathbf{B}_o^{-1} + \sum_{i=1}^{n} \mathbf{y}_i'\Omega_i^{-1}\mathbf{X}_i)\beta)] \qquad (2.21)
\end{aligned}
$$

From (2.21), we see the exponent of $\Pi(\beta \mid \mathbf{y}, \mathbf{D}, \sigma^2)$ is:

$$
\beta'(\mathbf{B}_o^{-1} + \sum_{i=1}^{n} \mathbf{X}_i'\Omega_i^{-1}\mathbf{X}_i)\beta - 2(\beta_o'\mathbf{B}_o^{-1} + \sum_{i=1}^{n} \mathbf{y}_i'\Omega_i^{-1}\mathbf{X}_i)\beta \qquad (2.22)
$$

Now define

$$
\begin{aligned}
\mathbf{B}_k &= (\mathbf{B}_o^{-1} + \sum_{i=1}^{n} \mathbf{X}_i'\Omega_i^{-1}\mathbf{X}_i)^{-1} \\
\mathbf{a}_i &= (\beta_o'\mathbf{B}_o^{-1} + \sum_{i=1}^{n} \mathbf{y}_i'\Omega_i^{-1}\mathbf{X}_i)' \\
&= (\mathbf{B}_o^{-1}\beta_o + \sum_{i=1}^{n} \mathbf{X}_i'\Omega_i^{-1}\mathbf{y}_i)
\end{aligned}
$$

This allows us to rewrite (2.22) as

$$
\beta'\mathbf{B}_k^{-1}\beta - 2\beta'\mathbf{a}_i = (\beta - \mathbf{B}_k\mathbf{a}_i)'\mathbf{B}_k^{-1}(\beta - \mathbf{B}_k\mathbf{a}_i) - \mathbf{a}_i\mathbf{B}_k\mathbf{a}_i' \qquad (2.23)
$$

Using (2.21) and (2.23) we find

$$\Pi(\beta \mid \mathbf{y}, \mathbf{D}, \sigma^2) \propto \exp[-1/2(\beta - \mathbf{B}_k\mathbf{a}_i)'\mathbf{B}_k^{-1}(\beta - \mathbf{B}_k\mathbf{a}_i)]$$

Therefore,

$$\begin{aligned} \beta \mid (\mathbf{y}, \mathbf{D}, \sigma^2) &\sim N(\mathbf{B}_k\mathbf{a}_i, \mathbf{B}_k) \\ &= N(\mathbf{B}_k(\mathbf{B}_o^{-1}\beta_o + \sum_{i=1}^{n} \mathbf{X}_i'\mathbf{\Omega}_i^{-1}\mathbf{y}_i), \mathbf{B}_k) \end{aligned} \tag{2.24}$$

Since steps 1(b) and 2-4 in Algorithm 2 are the same as in Algorithm 1, the posterior distributions of $\mathbf{b}$, $\mathbf{D}^{-1}$ and σ^2 are given in (2.10), (2.15) and (2.18), respectively.

2.3.3 Algorithm 3

While Algorithm 2 is an improvement on Algorithm 1, it does not address the correlation between $\mathbf{D}^{-1}$ and $\mathbf{b}$ that can lead to slow mixing for the unique elements of $\mathbf{D}^{-1}$ (Chib and Carlin, 1999). To deal with this problem we can use an approach that allows one to sample all parameters in one block from the joint posterior distribution. The idea is to use the following decomposition of the posterior distribution

$$\Pi(\sigma^2, \mathbf{D}^{-1}, \beta, \mathbf{b}_i|\mathbf{y}) = \Pi(\sigma^2, \mathbf{D}^{-1}|\mathbf{y})\Pi(\beta|\mathbf{y}, \sigma^2, \mathbf{D})\Pi(\mathbf{b}_i|\mathbf{y}, \beta, \sigma^2, \mathbf{D})$$

where the last two densities are the same as in Algorithm 2. The first density is not in closed form, but can be updated by the Metropolis-Hastings algorithm (see for example Hastings, 1970, or Chib and Greenberg, 1995), which was discussed in Chapter 1. By definition,

$$\Pi(\sigma^2, \mathbf{D}^{-1}|\mathbf{y}) \propto \Pi(\sigma^2, \mathbf{D}^{-1})f(\mathbf{y}|\sigma^2, \mathbf{D})$$

where

$$\begin{aligned} f(\mathbf{y}|\sigma^2, \mathbf{D}) &= \int f(\mathbf{y}|\beta, \sigma^2, \mathbf{D})\mathbf{\Pi}(\beta)\mathbf{d}\beta \\ &\propto |\mathbf{V}|^{-1/2}\exp[(\mathbf{y} - \mathbf{X}\beta_0)'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\beta_0)] \end{aligned}$$

and $\mathbf{y} = (\mathbf{y}_1', ..., \mathbf{y}_n')'$, $\mathbf{X} = (\mathbf{X}_1', ..., \mathbf{X}_n')'$, $\mathbf{V} = (\mathbf{X}\mathbf{B}_0^{-1}\mathbf{X}' + \boldsymbol{\Omega})$ and $\boldsymbol{\Omega} = diag(\boldsymbol{\Omega}_1, ..., \boldsymbol{\Omega}_n)$. Recall that $\boldsymbol{\Omega}_i$ was defined following (2.19). One way to evaluate this density is to recognize that $f(\mathbf{y}|\sigma^2, \mathbf{D})$ is the normalizing constant of $\Pi(\boldsymbol{\beta}|\mathbf{y}, \sigma^2, \mathbf{D})$. A similar idea is used by Chib (1995) in his approach to find the marginal likelihood of the model (2.1). Hence, we may write $f(\mathbf{y}|\sigma^2, \mathbf{D})$ as a ratio of three terms;

$$\begin{aligned} f(\mathbf{y}|\sigma^2, \mathbf{D}) &= \frac{\Pi(\boldsymbol{\beta}^*) f(\mathbf{y}|\boldsymbol{\beta}^*, \sigma^2, \mathbf{D})}{\Pi(\boldsymbol{\beta}^*|\mathbf{y}, \sigma^2, \mathbf{D})} \\ &= \frac{\phi_p(\boldsymbol{\beta}^*|\boldsymbol{\beta}_0, \mathbf{B}_0) \prod_{i=1}^n \phi_{n_i}(\mathbf{y}_i|\mathbf{x}_i\boldsymbol{\beta}^*, V_i)}{\phi_p(\boldsymbol{\beta}^*|\hat{\boldsymbol{\beta}}, B)} \end{aligned}$$

Where $\boldsymbol{\beta}^*$ is any point (preferably a high density point such as the posterior mean from Algorithm 2) and $\phi_p(\mathbf{t}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is density of the p-variate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. This leads to the following single block algorithm for sampling the posterior density of the Gaussian hierarchical model. We refer to this as Algorithm 3.

1. Run Algorithm 2 for $G = 500$ iterations (say) and let $\boldsymbol{\beta}^* = G^{-1} \sum_{g=1}^G \boldsymbol{\beta}^{(g)}$. Also let $\boldsymbol{\mu} = G^{-1} \sum_{g=1}^G \boldsymbol{\theta}^{(g)}$ and $\boldsymbol{\Sigma} = G^{-1} \sum_{g=1}^G (\boldsymbol{\theta}^{(g)} - \boldsymbol{\mu})(\boldsymbol{\theta}^{(g)} - \boldsymbol{\mu})'$, where $\boldsymbol{\theta} = (\sigma^2, \boldsymbol{\psi})$, and $\boldsymbol{\psi} = vech(\mathbf{D}^{-1})$ denotes the unique elements of $\mathbf{D}^{-1}$.

2. Sample $\boldsymbol{\theta}$, $\boldsymbol{\beta}$ and $\mathbf{b}$ from $[\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{b}|\mathbf{y}]$ by sampling

   (a) $\boldsymbol{\theta}$ from $\Pi(\boldsymbol{\theta}|\mathbf{y})$ using the Metropolis-Hastings algorithm with proposal density given by $q(\boldsymbol{\theta}) = f_{MVT}(\boldsymbol{\theta}|\boldsymbol{\mu}, \tau^2\boldsymbol{\Sigma}, \nu)$, where f_{MVT} is the multivariate-t density with ν degrees of freedom, and τ^2 and ν are tuning parameters. Given the current value $\boldsymbol{\theta}^c$, first draw $\boldsymbol{\theta}^t$ from $q(\boldsymbol{\theta})$ and move to the point $\boldsymbol{\theta}^t$ with probability given by

$$\alpha(\boldsymbol{\theta}^c, \boldsymbol{\theta}^t) = \min\left[1, \frac{f(\mathbf{y}|\sigma^{2t}, \mathbf{D}^t)\Pi(\sigma^{2t}, \mathbf{D}^{-t})q(\sigma^{2c}, \mathbf{D}^c)}{f(\mathbf{y}|\sigma^{2c}, \mathbf{D}^c)\Pi(\sigma^{2c}, \mathbf{D}^{-c})q(\sigma^{2t}, \mathbf{D}^t)}\right].$$

(b) Sample β from $N_p(\hat{\beta}, \mathbf{B}_k)$ where $\hat{\beta} = \mathbf{B}_k \mathbf{a}_i$,

$$\begin{aligned} \mathbf{B}_k &= (\mathbf{B}_o^{-1} + \sum_{i=1}^{n} \mathbf{X}_i' \Omega_i^{-1} \mathbf{X}_i)^{-1} \\ \mathbf{a}_i &= (\mathbf{B}_o^{-1} \beta_o + \sum_{i=1}^{n} \mathbf{X}_i' \Omega_i^{-1} \mathbf{y}_i) \end{aligned}$$

(c) Sample $\mathbf{b}_i$ independently from $N_q(\hat{\mathbf{b}}_i, \mathbf{C}_i)$ where $\hat{\mathbf{b}}_i = \mathbf{C}_i[\mathbf{W}_i' \sigma^{-2}(\mathbf{y}_i - \mathbf{X}_i \beta)]$ and $\mathbf{C}_i = (\mathbf{D}^{-1} + \mathbf{W}_i' \sigma^{-2} \mathbf{W}_i)^{-1}$.

3. Repeat Step 2 using the most recent values of the conditioning variables.

2.4 Model for Binary Data

In this section we consider various blocking schemes for the class of probit longitudinal binary random effects models. A Bayesian analysis of these models using a version of Algorithm 1 is provided by Albert and Chib (1996), and by Zeger and Karim (1991) under the logit link.

Consider a sequence of binary measurements $\mathbf{y}_i = (y_{i1}, \ldots\ldots, y_{in_i})'$, where $y_{it} = 0$ or 1, on the i^{th} unit taken at n_i specific time points. Let the probability $Pr(y_{it} = 1|\mathbf{b}_i)$ be modelled by the probit link:

$$Pr(y_{it} = 1|\mathbf{b}_i) = \Phi\left(\frac{\mathbf{x}'_{it}\beta + \mathbf{w}'_{it}\mathbf{b}_i}{\sigma}\right), \tag{2.25}$$

where, Φ is the standard normal cdf, $\mathbf{x}'_{it}$ and $\mathbf{w}'_{it}$ are the tth rows of $\mathbf{X}_i$ and $\mathbf{W}_i$, respectively. $\mathbf{X}_i$ is an $n_i \times p$ design matrix of covariates and β is a corresponding $p \times 1$ vector of fixed effects. In addition, $\mathbf{W}_i$ is a $n_i \times q$ design matrix and $\mathbf{b}_i$ is a $q \times 1$ vector of subject-specific random effects. For this model, the likelihood contribution $f(\mathbf{y}_i|\beta, \mathbf{D})$ is given by

$$\int \left[\prod_{t=1}^{n_i} \left[\Phi\left(\frac{\mathbf{x}'_{it}\beta + \mathbf{w}'_{it}\mathbf{b}_i}{\sigma}\right)\right]^{y_{it}} \left[1 - \Phi\left(\frac{\mathbf{x}'_{it}\beta + \mathbf{w}'_{it}\mathbf{b}_i}{\sigma}\right)\right]^{1-y_{it}}\right] \Pi(\mathbf{b}_i) d\mathbf{b}_i \tag{2.26}$$

where $\Pi(\mathbf{b}_i)$ is the prior distribution of $\mathbf{b}_i$. The integral in (2.26) is expensive to evaluate when $\mathbf{b}_i$ is multi-dimensional. One way to deal with this problem is via a latent variables approach (Albert and Chib, 1993, 1996; Carlin and Polson, 1992). Let z_{it} denote independent latent variables such that

$$z_{it}|\mathbf{b}_i \sim N(\mathbf{x}'_{it}\boldsymbol{\beta} + \mathbf{w}'_{it}\mathbf{b}_i, \sigma^2), 1 \leq t \leq n_i; 1 \leq i \leq n,$$

Let the observed response y_{it} be given by

$$y_{it} = \begin{cases} 1 & \text{if } z_{it} > 0 \\ 0 & \text{if } z_{it} \leq 0 \end{cases}$$

Then it can be seen that the y_{it} satisfy model (2.25).
In their paper Chib and Carlin (1999) considered $\sigma^2 = 1$. This means there is no prior, and hence no posterior distribution for σ^2. This seems to be quite restrictive. Therefore, in our investigations, we consider a more general case for σ^2.

2.4.1 Algorithm 4

With the introduction of the latent data, the probit model is similar to the Gaussian model discussed in Section 2.2 and the posterior distribution of the parameters $(\boldsymbol{\beta}, \mathbf{D})$ may be sampled in parallel fashion. Let $\mathbf{Z} = (\mathbf{Z}_1, ..., \mathbf{Z}_n)$ and $\mathbf{Z}_i = (z_{i1}, ..., z_{in_i})$. Then an MCMC scheme analogous to Algorithm 1 is defined as follows. We denote this as Algorithm 4:

1. Sample $\boldsymbol{\beta}$ from $\boldsymbol{\beta}|\mathbf{Z}, \mathbf{b}, \sigma^2, \mathbf{D}$
2. Sample $\mathbf{b}$ from $\{\mathbf{b}_i\}|\mathbf{Z}, \boldsymbol{\beta}, \sigma^2, \mathbf{D}$
3. Sample $\mathbf{D}^{-1}$ from $\mathbf{D}^{-1}|\mathbf{b}$
4. Sample σ^2 from $\sigma^2|\mathbf{Z}, \boldsymbol{\beta}, \mathbf{b}, \mathbf{D}$
5. Sample $\{z_{it}\}$ from $z_{it}|y_{it}, \boldsymbol{\beta}, \sigma^2, \mathbf{b}, \mathbf{D}$

6. Repeat Steps 1-5 using the most recent values of the conditioning variables.

Note that step 4 is not given by Chib and Carlin (1999). The first four conditional distributions follow the same form as those given in Algorithm 1: (2.6), (2.10), (2.15) and (2.18) except the latent variable vector $\mathbf{z}_i$ replaces $\mathbf{y}_i$ in those expressions. The posterior distribution in step 5 is given by a sequence of independent truncated normal distributions, namely $N_{(0,\infty)}(\mathbf{x}'_{it}\boldsymbol{\beta}+\mathbf{w}'_{it}\mathbf{b}_i, \sigma^2)$ if $y_{it} = 1$, or $N_{(-\infty,0)}(\mathbf{x}'_{it}\boldsymbol{\beta}+\mathbf{w}'_{it}\mathbf{b}_i, \sigma^2)$ if $y_{it} = 0$.

2.4.2 Algorithm 5

A refinement to Algorithm 4 is based on marginalizing the distribution of $\mathbf{z}_i$ over the random effects $\mathbf{b}_i$. Then,

$$\mathbf{z}_i \sim N_{n_i}(\mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Omega}_i) \tag{2.27}$$

where $\boldsymbol{\Omega}_i = \sigma^2\mathbf{I}_{n_i} + \mathbf{W}_i\mathbf{D}\mathbf{W}'_i$ and the model is similar to the multivariate probit model analyzed by Chib and Greenberg (1998). The resulting algorithm is similar to Algorithm 4 except that $\boldsymbol{\beta}$ is sampled from $\boldsymbol{\beta}|(\mathbf{Z}, \sigma^2, \mathbf{D})$, and the latent variable $\mathbf{z}_i$ comes from the multivariate normal distribution $N_{n_i}(\mathbf{x}_i\boldsymbol{\beta}, \boldsymbol{\Omega}_i)$ truncated to the region implied by the vector $\mathbf{y}_i$. We follow Chib and Greenberg (1998) and sample this truncated multivariate normal vector from a sequence of (full conditional) univariate truncated normal distributions. This can be done by recognizing

$$f(z_{i1}, ..., z_{it}) = f(z_{i1})f(z_{i2}|z_{i1})....f(z_{it}|z_{i1}...z_{i(t-1)})$$

where each distribution on the right-hand side is univariate normal, and can be found using a standard result on the conditional distributions arising from a Gaussian distribution (Johnson and Wichern, 1992, p. 138). Thus, in this case, integrating out the random effects does not lead to a reduction in the number of blocks in the sampling (relative to Algorithm 4). Nonetheless, marginalization over the $\mathbf{b}_i$ can be expected

to improve the sampling of the fixed effects for the reasons discussed in earlier sections. This is similar to the improvement that Algorithm 2 is intended to show over Algorithm 1. We summarize this algorithm, which we call Algorithm 5, as follows:

1. Sample $\boldsymbol{\beta}$ and $\{\mathbf{z}_i\}$ from $[\boldsymbol{\beta}, \{\mathbf{z}_i\}|\mathbf{y}, \sigma^2, \mathbf{D}$ by sampling
   (a) $\boldsymbol{\beta}$ from $\boldsymbol{\beta}|\mathbf{y}, \mathbf{z}, \sigma^2, \mathbf{D}$
   (b) $\{\mathbf{z}_i\}$ from $\mathbf{z}_i|\mathbf{y}_i, \boldsymbol{\beta}, \sigma^2, \mathbf{D}$
2. Sample $\mathbf{b}$ from $\{\mathbf{b}_i\}|\mathbf{y}, \mathbf{z}, \boldsymbol{\beta}, \sigma^2, \mathbf{D}$
3. Sample $\mathbf{D}^{-1}$ from $\mathbf{D}^{-1}|\mathbf{b}$
4. Sample σ^2 from $\sigma^2|\mathbf{y}, \mathbf{z}, \boldsymbol{\beta}, \mathbf{b}, \mathbf{D}$
5. Repeat Steps 1-4 using the most recent values of the conditioning variables.

The posterior distributions that are used in this algorithm are the same as those derived in Algorithm 2: (2.24), (2.10), (2.15) and (2.18), except the latent variable vector $\mathbf{z}_i$ replaces $\mathbf{y}_i$ in those expressions.

2.5 Conclusion

In this chapter, we have presented several algorithms for generating samples from the posterior distributions of interest for two longitudinal models. In the next chapter we will present some simulation studies on the performance of these algorithms.

Chapter 3

Simulation Studies

3.1 Introduction

In this chapter, we conduct several simulation studies on the MCMC algorithms presented in Chapter 2. In the longitudinal models, we will study the performance of the posterior estimates, as well as the autocorrelations of the MCMC samples of each algorithm.

3.2 Simulation Design and Generation of the Continuous Data

We use the Gaussian linear mixed model (2.1) in this section. Under model (2.1), we will use $n = 50$ and $n_i = 5$ measurements on each subject. In our simulations for Algorithms 1-3, we simulate our data under the following prior assumptions on our parameters:
$\boldsymbol{\beta} \sim N_4(\boldsymbol{\beta}_o, \mathbf{B}_o)$, $\mathbf{b}_i \sim N_2(\mathbf{0}, \mathbf{D})$, $\mathbf{D}^{-1} \sim W_2(\rho_o, \mathbf{R}_o\rho_o^{-1})$ and $\sigma^2 \sim IG(\nu_o/2, \delta_o/2)$.

where $\boldsymbol{\beta}_o = (0, 0, 0, 0)'$, $\mathbf{B}_o = \mathbf{I}$, $\rho_o = 50$ and

$$\mathbf{R}_o = \begin{pmatrix} 1 & .5 \\ .5 & 1 \end{pmatrix}.$$

In our analyses, we examine our results under a variety of choices for the other prior parameters. In each case, we will run Algorithms 1-3 of Chapter 2 for 500 iterations. No burn-in-period was used in any simulations, although it may have been of some assistance in Algorithm 3. In all simulations, one set of simulated y_{it} values was used. Finally, our simulations focus on changing the prior of σ^2. Because, in our work, we found that changing σ^2 had the most dramatic affect on the results.

3.3 Simulation Results-Continuous Data

We will begin our simulations with the posterior means and variances of the parameters under Algorithms 1-3. These are given in Tables 3.1 to 3.3. These tables refer to Cases 1-3, which are defined as follows:

Case 1: $\nu_o = 100$, $\delta_o = 5$. Therefore the prior mean of $\sigma^2 = 0.05$.

Case 2: $\nu_o = 5$, $\delta_o = 100$. Therefore the prior mean of $\sigma^2 = 20$.

Case 3: $\nu_o = 5$, $\delta_o = 5$. Therefore the prior mean of $\sigma^2 = 1$.

We examine the results on the posterior estimates for Algorithm 1, which are given in Table 3.1.

As we examine the results in Table 3.1, we see that changing the prior of σ^2 has some effect on $mean(\hat{\boldsymbol{\beta}})$. Increasing the prior mean of σ^2 causes $var(\hat{\boldsymbol{\beta}})$, $mean(\hat{\sigma}^2)$, $var(\hat{\sigma}^2)$ to increase, but it has little effect on $mean(\hat{\mathbf{D}})$ and $var(\hat{\mathbf{D}})$.

Next, we examine the results on the posterior estimates for Algorithm 2, which are presented in Table 3.2.

As we examine the results in Table 3.2, we see that changing prior of σ^2 does not have a large effect on $mean(\hat{\boldsymbol{\beta}})$. Increasing the prior mean of σ^2 causes $var(\hat{\boldsymbol{\beta}})$, $mean(\hat{\sigma}^2)$ and $var(\hat{\sigma}^2)$ to go up, as we might expect. However, there is little effect on $mean(\hat{\mathbf{D}})$ and $var(\hat{\mathbf{D}})$.

Cases		Parameters							
		β_0	β_1	β_2	β_3	σ^2	D_{11}	D_{21}	D_{22}
Case 1	Mean	0.78	-0.44	-1.73	-4.36	2.10	1.28	-0.73	1.30
	Var	0.10	0.55	0.40	0.23	0.116	0.04	0.03	0.04
Case 2	Mean	-0.50	-0.86	-0.92	-0.76	16.21	1.39	-0.69	1.38
	Var	0.30	0.71	0.66	0.63	2.42	0.07	0.04	0.07
Case 3	Mean	0.51	-0.65	-1.47	-3.53	3.15	1.27	-0.70	1.30
	Var	0.13	0.58	0.44	0.31	0.26	0.04	0.03	0.04

Table 3.1: Posterior means and variances of parameters in simulations using Algorithm 1.

Finally, we present the results of Algorithm 3 in Table 3.3. As we examine the results in Table 3.3, we see that changing the prior of σ^2 has some effect on $mean(\hat{\beta})$, but the results of the 3 cases do not differ greatly. Increasing the prior mean of σ^2 causes $var(\hat{\beta})$, $mean(\hat{\sigma}^2)$ and $var(\hat{\sigma}^2)$ to increase. It has little effect on $mean(\hat{\mathbf{D}})$ and $var(\hat{\mathbf{D}})$.

One of the issue to address in Algorithm 3 is the Metropolis-Hastings step of simulating approximate samples from $\Pi(\sigma^2, \mathbf{D}^{-1}|\mathbf{y})$. Recall that the algorithm moves to new posterior values for σ^2 and $\mathbf{D}$ with probability $\alpha(\boldsymbol{\theta}^c, \boldsymbol{\theta}^t)$, where $\boldsymbol{\theta}$ contains σ^2 and the unique elements of $\mathbf{D}^{-1}$. From our background in Chapter 1, we want α to be neither too large nor too small.

In our simulations, under Case 1, we observe movement from $\boldsymbol{\theta}^c$ to $\boldsymbol{\theta}^t$ about 3% of the time, about 48% of the time in Case 2, about 9% of the time in Case 3. Therefore, it appears that Algorithm 3 does not perform well in generating posterior samples when the prior mean of σ^2 is small. For tuning parameters, we use $\tau^2 = 0.1$ and $\nu = 10$ in the multivariate-t distribution. We tried several other combinations, but none perfomed better than the results presented here. We will investigate other implications of this lack of movement in the Metropolis-Hastings step in the next section.

Cases		Parameters							
		β_0	β_1	β_2	β_3	σ^2	D_{11}	D_{21}	D_{22}
Case 1	Mean	-0.24	-0.30	-1.32	-0.93	0.58	1.38	-0.77	1.35
	Var	0.09	0.49	0.31	0.42	0.03	0.04	0.03	0.04
Case 2	Mean	-0.62	-0.77	-0.92	-0.13	15.99	1.45	-0.70	1.34
	Var	0.27	0.71	0.63	0.67	2.54	0.07	0.04	0.06
Case 3	Mean	-0.27	-0.48	-1.23	-0.74	1.52	1.37	-0.74	1.34
	Var	0.11	0.56	0.39	0.47	0.09	0.04	0.03	0.04

Table 3.2: Posterior means and variances of parameters in simulations using Algorithm 2.

3.3.1 Comparison of Algorithms

If we compare Cases 1-3, we see that Algorithm 1 tends to give larger (in magnitude) values for $mean(\hat{\boldsymbol{\beta}})$ than Algorithms 2 and 3, which give similar results. However, $var(\hat{\boldsymbol{\beta}})$ is similar for all algorithms. Algorithm 1 provided larger values for $mean(\hat{\sigma}^2)$ than the other two algorithms, particularly when the prior mean of σ^2 was small. We also see Algorithm 3 leads to smaller values of $var(\hat{\sigma}^2)$, $mean(\hat{\mathbf{D}})$ and $var(\hat{\mathbf{D}})$ than other two algorithms.

3.3.2 Graphs

Although we have examined some summary statistics on our posterior distributions, it is also of interest to examine our posterior distributions visually. Therefore, we now present histograms of a selection of the posterior distributions discussed earlier.

From Figure 3.1 (Algorithm 1, Case 1), we found that posterior distributions of all $\boldsymbol{\beta}$ estimates appear symmetric and normal. The posterior mean of β_0 is larger than its prior mean and the posterior means of β_1, β_2, β_3 are less than their prior means.

From Figure 3.2 (Algorithm 1, Case 1), we found that the posterior distribution

Cases		Parameters							
		β_0	β_1	β_2	β_3	σ^2	D_{11}	D_{21}	D_{22}
Case 1	Mean	-0.25	-0.16	-1.43	-0.97	0.30	1.08	0.67	1.07
	Var	0.09	0.40	0.23	0.49	0.01	0.002	0.002	0.003
Case 2	Mean	-0.64	-0.73	-0.93	-0.13	15.99	1.00	0.45	1.05
	Var	0.28	0.65	0.55	0.78	0.52	0.006	0.003	0.004
Case 3	Mean	-0.24	-0.48	-1.24	-0.86	1.32	1.10	0.54	1.10
	Var	0.11	0.55	0.35	0.46	0.02	0.01	0.002	0.004

Table 3.3: Posterior means and variances of parameters in simulations using Algorithm 3.

of σ^2 doesn't appear very skewed, but the posterior distributions of D_{11} and D_{22} are skewed to the right as a χ^2 distribution and the posterior distribution of D_{21} (the off diagonal element of the Wishart distribution) do not seem symmetric. The posterior mean of σ^2 is larger than its prior mean and the posterior means of D_{11}, D_{21} and D_{22} are very close to their prior means, which were 1, 0.5, and 1 respectively.

From Figure 3.3 (Algorithm 2, Case 3), we found that the posterior distributions of all β estimates appear approximately normal, and the posterior means are less than their prior means.

From Figure 3.4 (Algorithm 2, Case 3), we found that the posterior distributions of σ^2, D_{11} and D_{22} are skewed to the right as a χ^2 distribution, but the posterior distribution of D_{21} appears asymmetric. The posterior mean of σ^2 is larger than its prior mean while the posterior means of D_{11}, D_{21} and D_{22} are very close to their prior means.

From Figure 3.5 (Algorithm 3, Case 2), we found that the posterior distributions of all β estimates are approximately normal and the posterior means of β_0, β_1, β_2, β_3 are less than their prior means.

From Figure 3.6 (Algorithm 3, Case 2), we found that the posterior distributions of σ^2 and the elements of D all appear to be slightly skewed. Each posterior mean is

very close to prior means.

3.4 Simulation Design and Generation of the Binary Data

We use the probit link model (2.25) in this section. Under model (2.25), we will use $n = 50$ and $n_i = 5$ measurements on each subject. In our simulations for Algorithms 4-5, we simulate our data under the following prior assumptions on our parameters: $\beta \sim N_4(\beta_o, \mathbf{B}_o)$, $\mathbf{b}_i \sim N_2(0, \mathbf{D})$, $\mathbf{D}^{-1} \sim W_2(\rho_o, \mathbf{R}_o\rho_o^{-1})$ and $\sigma^2 \sim IG(\nu_o/2, \delta_o/2)$, where $\beta_o = (0, 0, 0, 0)'$, $\mathbf{B}_o = \mathbf{I}$, $\rho_o = 50$ and

$$\mathbf{R}_o = \begin{pmatrix} 1 & .5 \\ .5 & 1 \end{pmatrix}.$$

In our analyses, we examine our results under a variety of choices for the other prior parameters. In each case, we will run the algorithms 4-5 of Chapter 2 for 500 iterations.

3.5 Simulation Results-Binary Data

We will begin our simulations with the posterior means and variances of the parameters under Algorithms 4 and 5. These are given in Tables 3.4 and 3.5. These tables refer to Cases 1-3, which are defined as follows:

Case 1: $\nu_o = 100$, $\delta_o = 5$. Therefore the prior mean of $\sigma^2 = 0.05$.

Case 2: $\nu_o = 5$, $\delta_o = 100$. Therefore the prior mean of $\sigma^2 = 20$.

Case 3: $\nu_o = 5$, $\delta_o = 5$. Therefore the prior mean of $\sigma^2 = 1$.

We examine the results of the posterior estimates for Algorithm 4, which are given in Table 3.4.

As we examine the results in Table 3.4, we see that increasing prior of σ^2 causes $mean(\hat{\beta})$, $var(\hat{\beta})$, $mean(\hat{\sigma}^2)$, $var(\hat{\sigma}^2)$, $mean(\hat{\mathbf{D}})$ and $var(\hat{\mathbf{D}})$ to go down (in absolute

Cases		Parameters							
		β_0	β_1	β_2	β_3	σ^2	D_{11}	D_{21}	D_{22}
Case 1	Mean	-9.03	-5.69	-4.80	-7.16	736.14	794.04	-684.85	597.54
	Var	3.80	1.83	1.49	2.96	7.5×10^4	3.1×10^5	1.9×10^5	1.2×10^5
Case 2	Mean	-1.52	-1.56	-1.41	-1.47	47.41	1.70	-0.96	1.76
	Var	0.49	0.81	0.76	1.18	195.19	0.23	0.15	0.20
Case 3	Mean	-1.56	-1.59	-1.44	-1.61	50.46	1.87	-1.12	1.91
	Var	0.66	0.82	0.80	2.14	442.21	1.21	1.09	1.22

Table 3.4: Posterior means and variances of parameters in simulations using Algorithm 4.

value). Case 1 provided larger posterior estimates for all parameters. This is especially true for the variance components.

Cases		Parameters							
		β_0	β_1	β_2	β_3	σ^2	D_{11}	D_{21}	D_{22}
Case 1	Mean	-0.07	-0.69	-0.61	-0.61	1.97	1.54	-0.64	1.28
	Var	0.12	0.52	0.45	0.56	0.15	0.113	0.04	0.05
Case 2	Mean	-0.69	-1.16	-1.07	-0.40	17.35	1.54	-0.77	1.56
	Var	0.32	0.64	0.68	0.75	36.26	0.105	0.06	0.09
Case 3	Mean	-0.51	-1.07	-0.99	-0.41	11.99	1.54	-0.74	1.53
	Var	0.28	0.61	0.63	0.71	20.28	0.106	0.06	0.09

Table 3.5: Posterior means and variances of parameters in simulations using Algorithm 5.

Table 3.5 gives the results of the posterior estimates for Algorithm 5. As we examine the results in Table 3.5, we see that changing the prior of σ^2 has not had a large effect on $mean(\hat{\boldsymbol{\beta}})$ and $var(\hat{\boldsymbol{\beta}})$. Increasing the prior mean of σ^2 causes $mean(\hat{\sigma}^2)$ and $var(\hat{\sigma}^2)$ to go up. It has little effect on $mean(\hat{\mathbf{D}})$ and $var(\hat{\mathbf{D}})$.

3.5.1 Comparison of Algorithms

If we compare Algorithms 4 and 5, we see that Algorithm 4 provided larger values (in absolute terms) for the posterior means and variances of $\boldsymbol{\beta}$. There are also large differences between the posterior results on σ^2, and between the values of $\mathbf{D}$ in some cases. It does appear that Algorithm 5 gives us more reliable results, based on these findings.

3.5.2 Graphs

As in the previous section, we present some histograms on a subset of our posterior distributions.

From Figure 3.7 (Algorithm 4, Case 3), we found that the posterior distribution of β_3 is highly skewed to the left and the posterior distributions of β_0, β_1, β_2 are symmetric. The posterior means of β_0, β_1, β_2, β_3 are less than their prior means.

From Figure 3.8 (Algorithm 4, Case 3), we found that the posterior distributions of σ^2, D_{11} and D_{22} are highly skewed to the right, but the posterior distribution of D_{21} is highly skewed to the left. The posterior mean of σ^2 is larger than its prior mean. However, the posterior means of D_{11}, D_{21} and D_{22} are slightly bigger than their prior means.

From Figure 3.9 (Algorithm 5, Case 2), we found that the posterior distributions of all $\boldsymbol{\beta}$ estimates are approximately normal, with posterior means of that are less than their prior means.

From Figure 3.10 (Algorithm 5, Case 2), we found that the posterior distributions of σ^2 and the elements of D are skewed to the right. The posterior means of σ^2, D_{11}, D_{21} and D_{22} are very close to their prior means.

3.6 Autocorrelations of Posterior Estimates

We now study the behaviour of the autocorrelation values of the posterior estimates in our simulation studies. As discussed by Carlin and Chib (1999), we want these autocorrelations to be close to 0, since that will indicate approximate independence in the movement of the Markov Chain.

Parameter	Algorithm 1	Algorithm 2	Algorithm 3
β_1	-.0023	.0209	.0883
β_2	-.0094	-.0408	.0252
β_3	-.0522	.0056	-.0277
β_4	.0227	-.0523	.0569
σ^2	.1749	.0983	.5880
D_{11}	.3935	.3588	.6590
D_{21}	.3630	.4017	.5861
D_{22}	.3767	.3788	.5168

Table 3.6: Lag-1 autocorrelations of posterior estimates under Case 2, using continuous data.

We will examine two components of the autocorrelation values of our MCMC Algorithms. First, we will calculate the lag-1 autocorrelations of our posterior estimates. As noted, lag-1 autocorrelations near 0 will suggest approximate independence in the MCMC movement.

The second component to be studied is a summary of the autocorrelation at all lags and the overall rate of decay, following Chib and Carlin (1999). This summary can be represented as

$$\kappa = 1 + 2\sum_{k=1}^{\infty} \rho(k),$$

where $\rho(k)$ is the lag k autocorrelation of the posterior estimate of interest. The value κ is sometimes referred to as the autocorrelation time. We estimate κ using the sample autocorrelations estimated from the MCMC procedure, cutting off the

Parameter	Algorithm 1	Algorithm 2	Algorithm 3
β_1	-.0397	-.0165	-.0130
β_2	.0013	-.0440	-.0212
β_3	-.0115	-.0273	.0002
β_4	-.0879	-.0668	.0517
σ^2	.2426	.2287	.9530
D_{11}	.1531	.1176	.9529
D_{21}	.1185	.0687	.9386
D_{22}	.1920	.1214	.8855

Table 3.7: Lag-1 autocorrelations of posterior estimates under Case 3, using continuous data.

summation when the sample autocorrelations fall below 0.1 in magnitude. Using Kass et al. (1998, p. 99), κ can be thought of as the relative increase in run length needed by the MCMC method to deal with the dependence. Ideally, κ will be small. Note that if we have strict independence, $\kappa = 1$.

3.6.1 Results on Algorithms 1-3

Tables 3.6 and 3.7 contain the values of the lag-1 autocorrelations of the posterior estimates using Algorithms 1-3 for Cases 2 and 3, while Tables 3.8 and 3.9 contain the estimates of κ for these situations.

From Tables 3.6 and 3.7, we obtained good performance for the β's in Algorithm 1-3. Algorithm 1 and Algorithm 2 perform better for σ^2 and the elements of $\mathbf{D}$ than Algorithm 3. Algorithm 3 in Table 3.6 performs better for σ^2 and the elements of $\mathbf{D}$ than in Table 3.7, because there is little movement in the MCMC procedure in Table 3.7 for smaller values of σ^2.

From Tables 3.8 and 3.9, we obtained good performance for the β's in Algorithms 1-3. Algorithms 1 and 2 perform better for σ^2 and the elements of $\mathbf{D}$ than Algorithm

Parameter	Algorithm 1	Algorithm 2	Algorithm 3
β_1	1	1	1
β_2	1.237	1	1
β_3	1	1	1.211
β_4	1.313	1	1
σ^2	2.196	1	7.539
D_{11}	2.314	1.950	14.874
D_{21}	1.726	2.563	10.591
D_{22}	2.185	2.023	6.303

Table 3.8: Estimates of $\kappa = 1 + 2\sum_{k=1}^{\infty} \rho(k)$ for posterior estimates, Case 2, for continuous data.

3. Algorithm 3 in Table 3.8 performs better for σ^2 and the elements of $\mathbf{D}$ than in Table 3.9, because there is little movement in the MCMC procedure in Table 3.9 for smaller values of σ^2.

3.6.2 Results on Algorithms 4 and 5

Tables 3.10 and 3.11 contain the values of the lag-1 autocorrelations of the posterior estimates using Algorithms 4 and 5 for Cases 2 and 3, while Tables 3.12 and 3.13 contain the estimates of κ for these situations.

From Tables 3.10 and 3.11, we observed better performance for the β's and the elements of $\mathbf{D}$ for Algorithm 5 than Algorithm 4. Algorithm 4 and Algorithm 5 perform similarly for σ^2. The findings are similar for Tables 3.12 and 3.13.

3.7 Conclusions

We have conducted simulation studies to compare the performance of the algorithms discussed in Chapter 2. In regards to the algorithms for continuous data, it does not

Parameter	Algorithm 1	Algorithm 2	Algorithm 3
β_1	1	1	1
β_2	1.260	1	1.752
β_3	1	1	1.828
β_4	1.243	1	1
σ^2	1.698	1.457	28.403
D_{11}	1.306	1.677	28.983
D_{21}	1.237	1	31.996
D_{22}	1.868	1.243	29.736

Table 3.9: Estimates of $\kappa = 1 + 2\sum_{k=1}^{\infty} \rho(k)$ for posterior estimates, Case 3, for continuous data.

appear that Algorithm 3 is an improvement over Algorithms 1 and 2, since it seems difficult to get good performance from the Metropolis-Hastings step of the algorithm. For the methods for binary data, it does appear that Algorithm 5 is an improvement over Algorithm 4. Finally, the results on the binary data suggest the choice of prior on σ^2 plays a role, and its value should not be set to equal 1 arbitrarily.

Parameter	Algorithm 4	Algorithm 5
β_1	.1750	.0760
β_2	.0348	.0035
β_3	.0147	-.0106
β_4	.3903	.0958
σ^2	.7108	.9216
D_{11}	.6551	.5378
D_{21}	.6417	.4460
D_{22}	.5970	.4937

Table 3.10: Lag-1 autocorrelations of posterior estimates under Case 2, using binary data.

Parameter	Algorithm 4	Algorithm 5
β_1	.3594	.0801
β_2	.0452	.0030
β_3	.0403	-.0100
β_4	.6416	.1025
σ^2	.8501	.9249
D_{11}	.8927	.5398
D_{21}	.9044	.4168
D_{22}	.8893	.4615

Table 3.11: Lag-1 autocorrelations of posterior estimates under Case 3, using binary data.

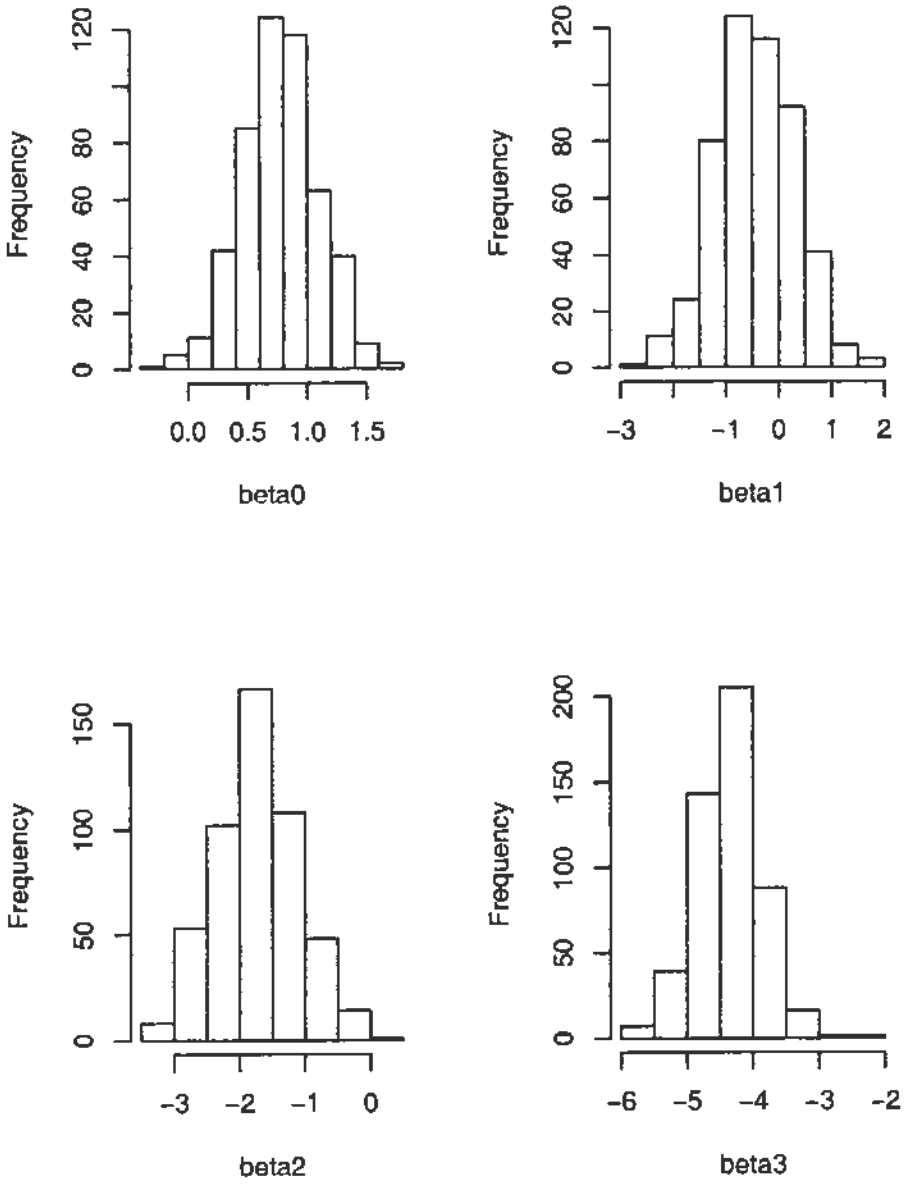

Figure 3.1: Posterior Distributions of β estimates in Algorithm 1, Case 1.

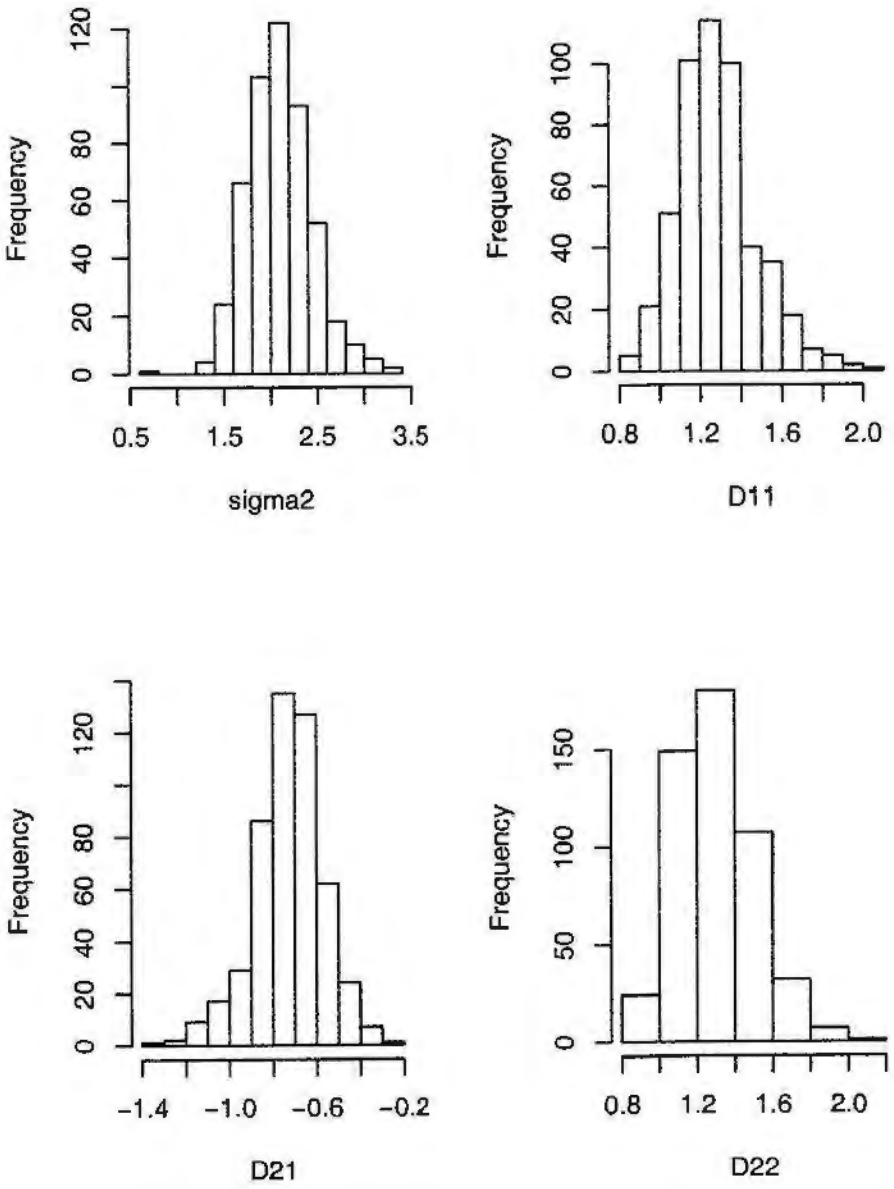

Figure 3.2: Posterior Distributions of σ^2 and the elements of $\mathbf{D}$ estimates in Algorithm 1, Case 1.

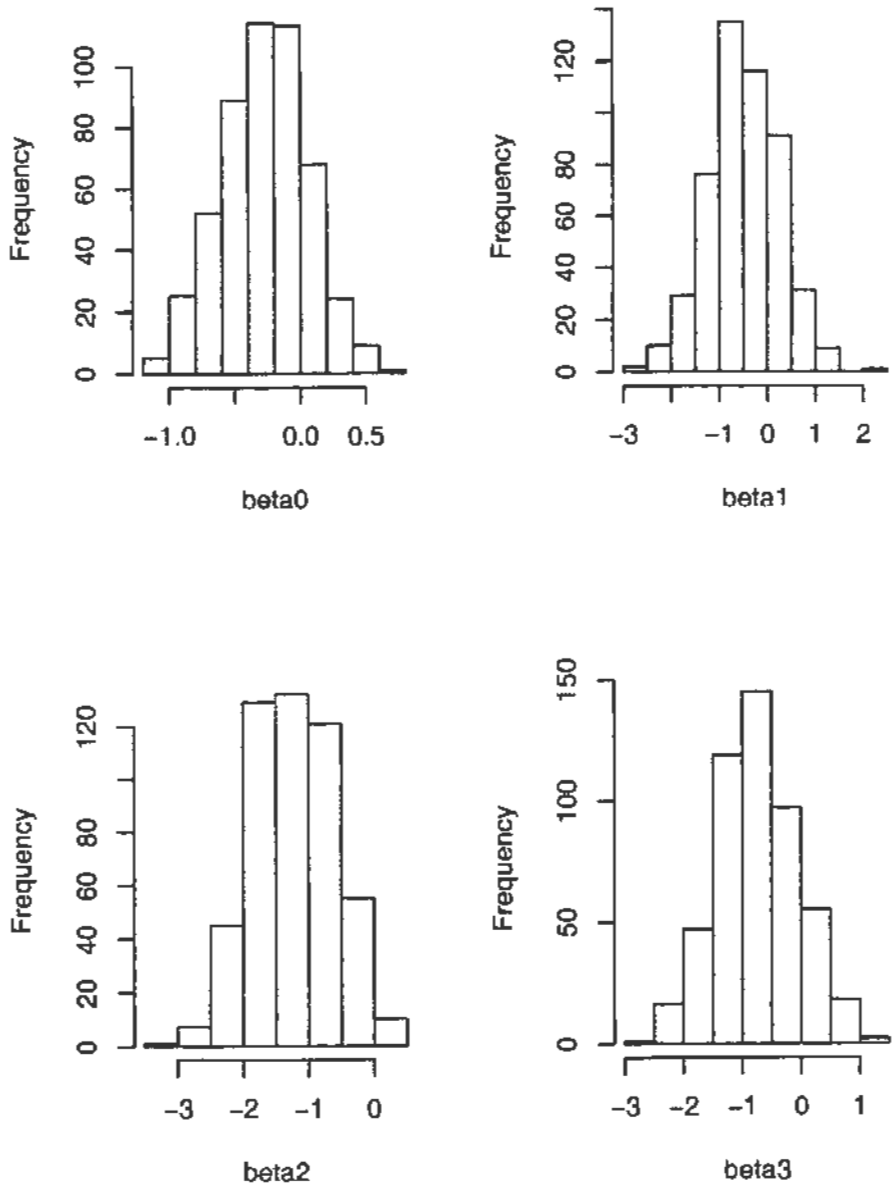

Figure 3.3: Posterior Distributions of β estimates in Algorithm 2, Case 3.

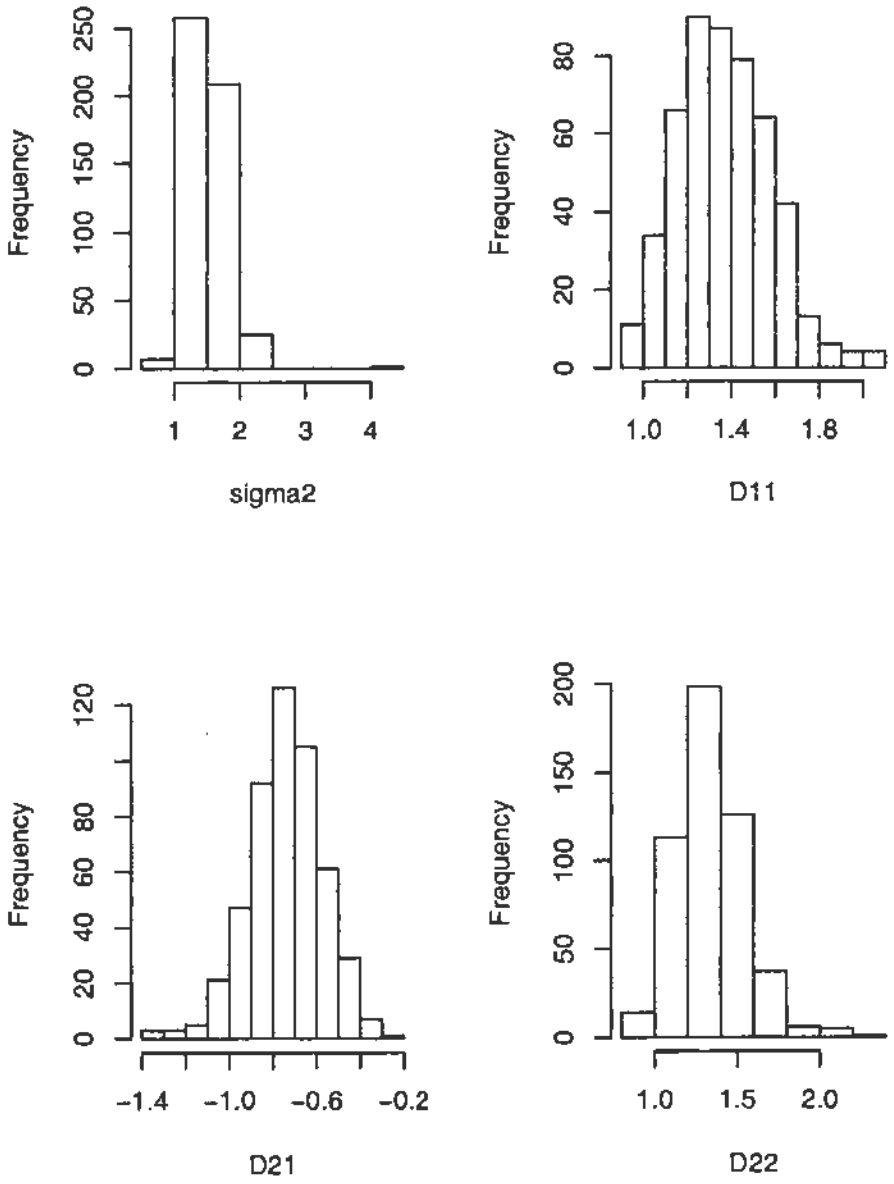

Figure 3.4: Posterior Distributions of σ^2 and the elements of D estimates in Algorithm 2, Case 3.

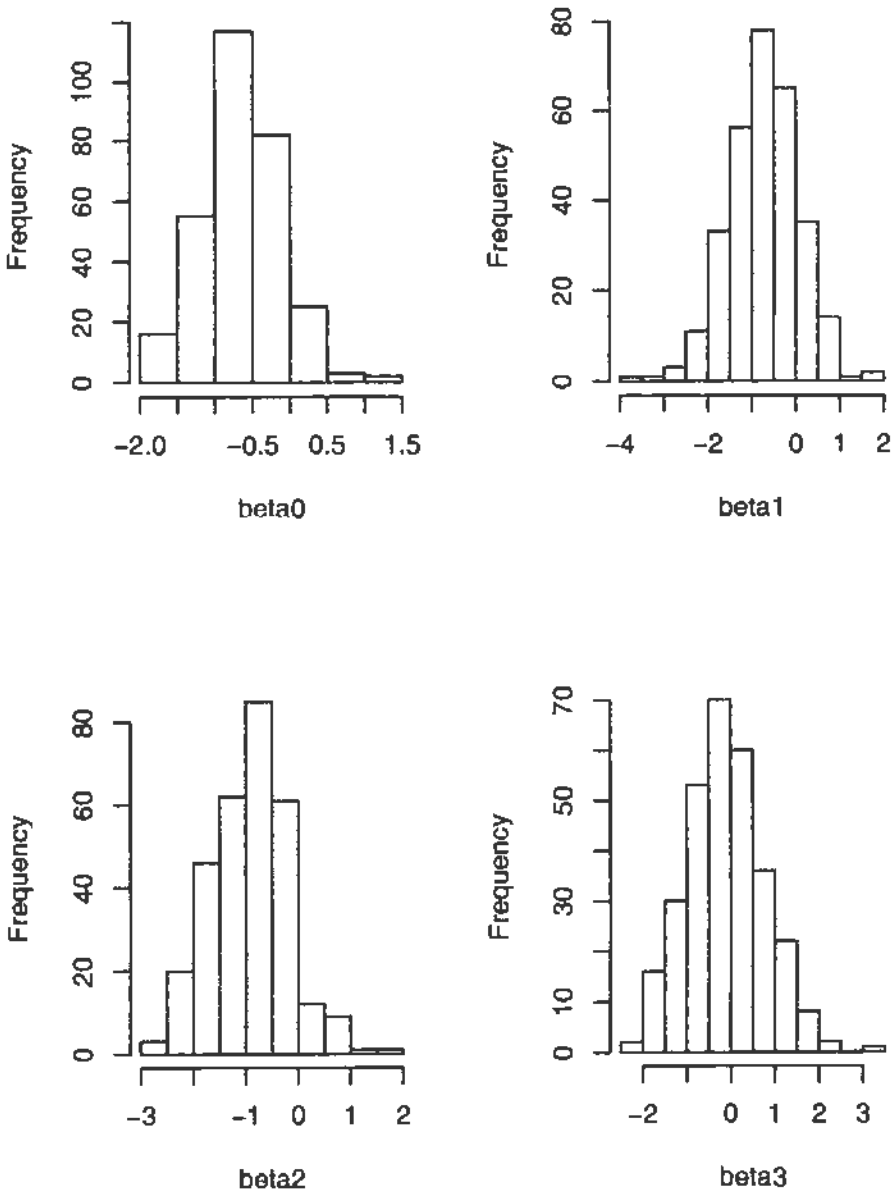

Figure 3.5: Posterior Distributions of β estimates in Algorithm 3, Case 2.

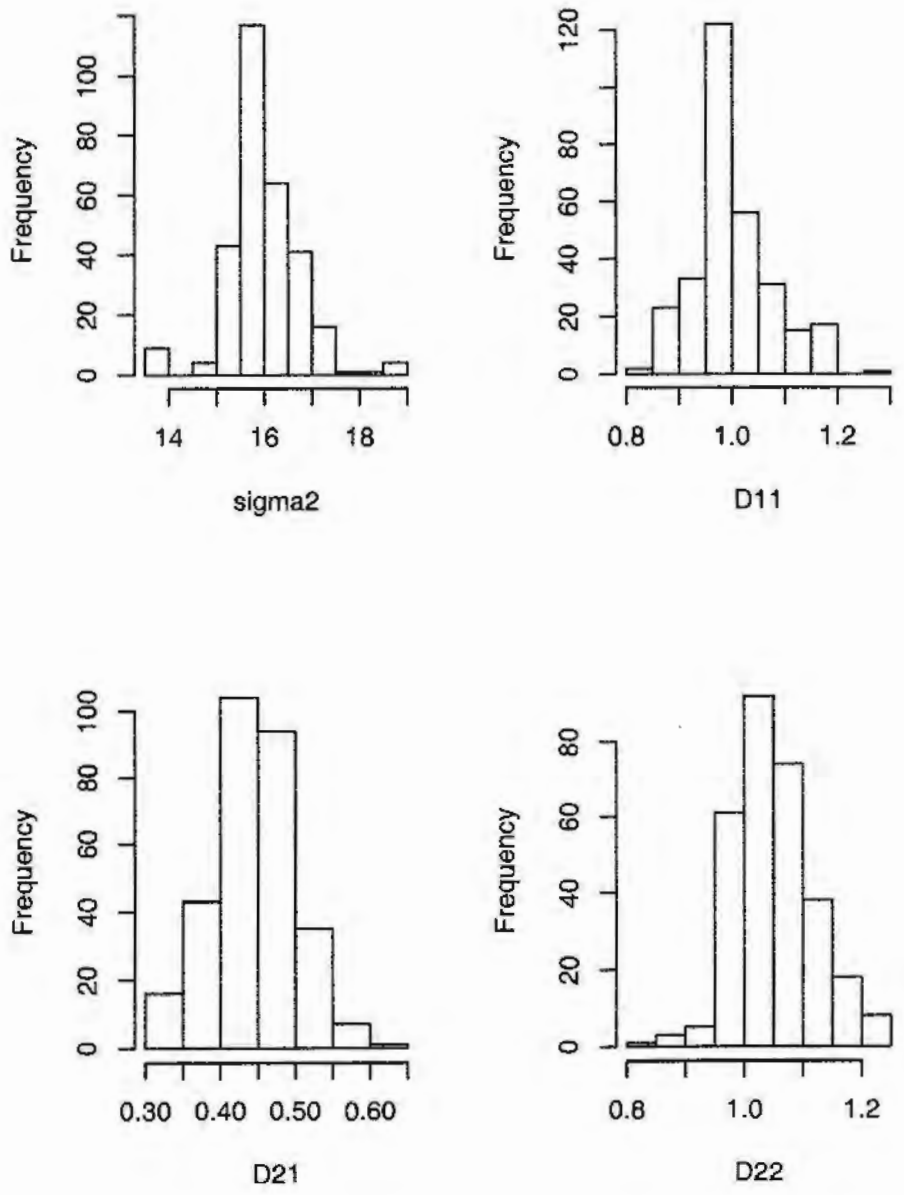

Figure 3.6: Posterior Distributions of σ^2 and the elements of $\mathbf{D}$ estimates in Algorithm 3, Case 2.

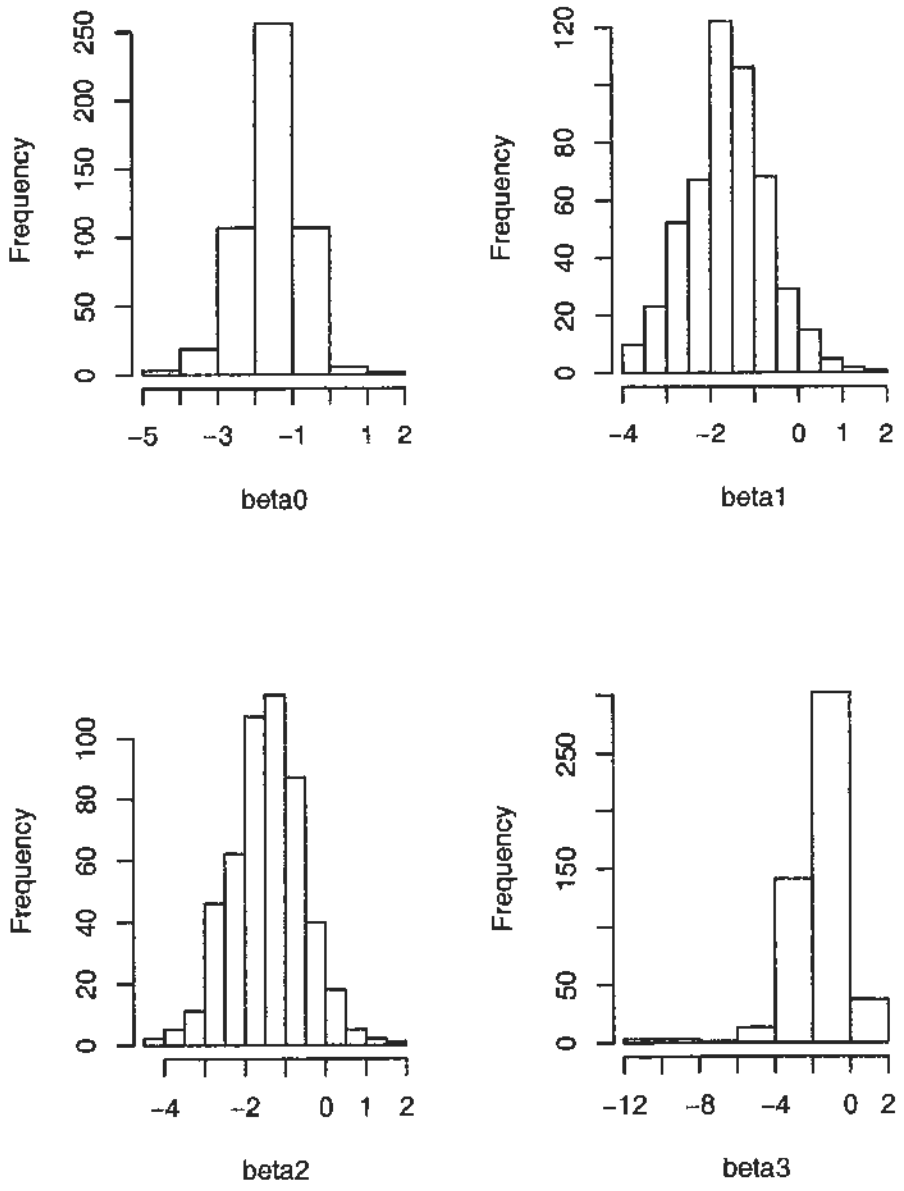

Figure 3.7: Posterior Distributions of β estimates in Algorithm 4, Case 3.

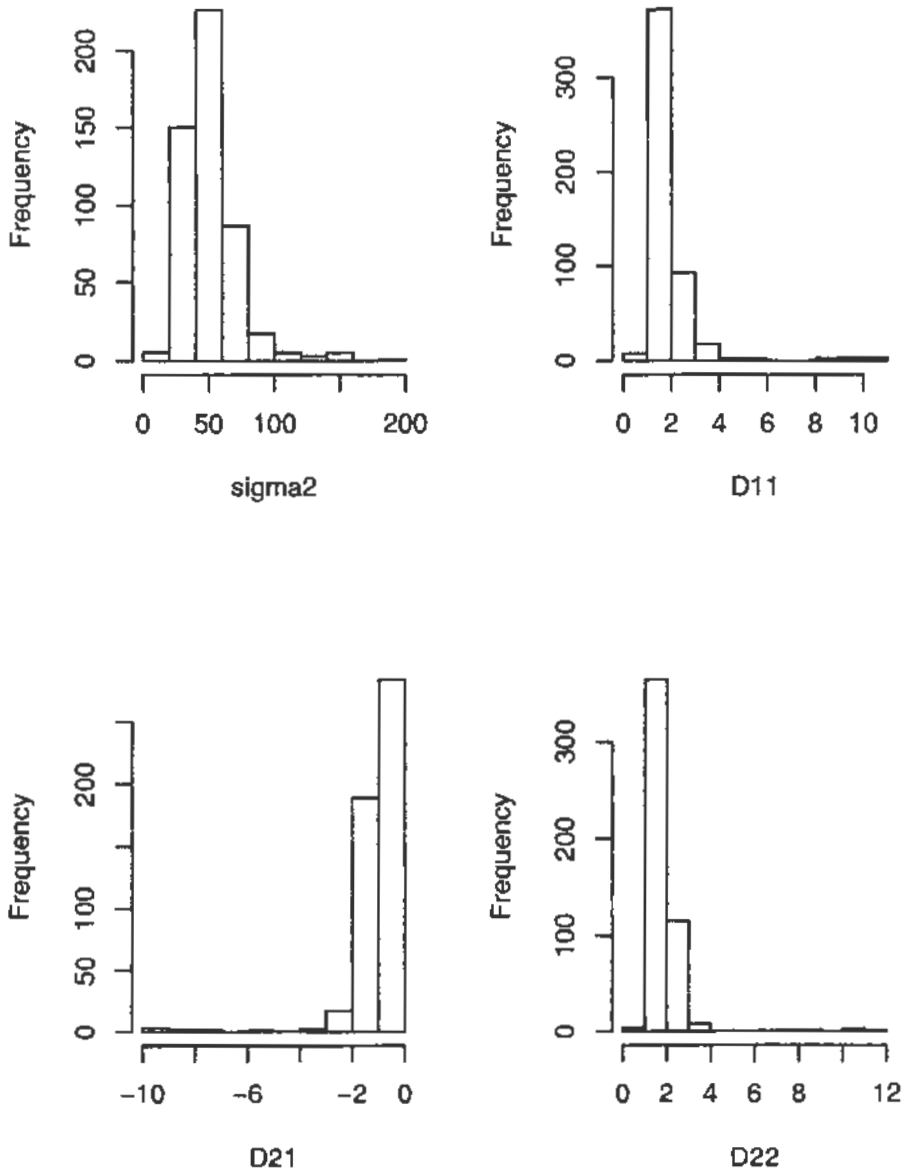

Figure 3.8: Posterior Distributions of σ^2 and the elements of $\mathbf{D}$ estimates in Algorithm 4, Case 3.

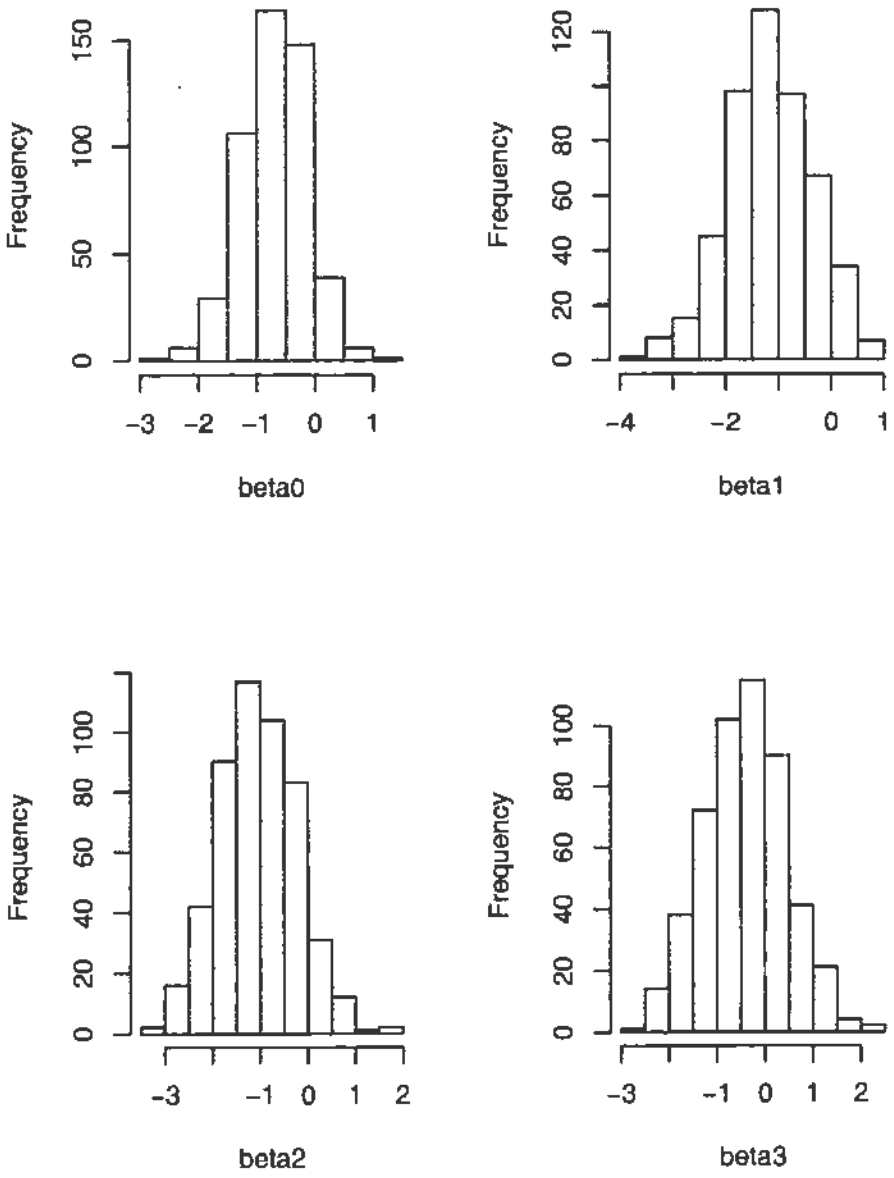

Figure 3.9: Posterior Distributions of β estimates in Algorithm 5, Case 2.

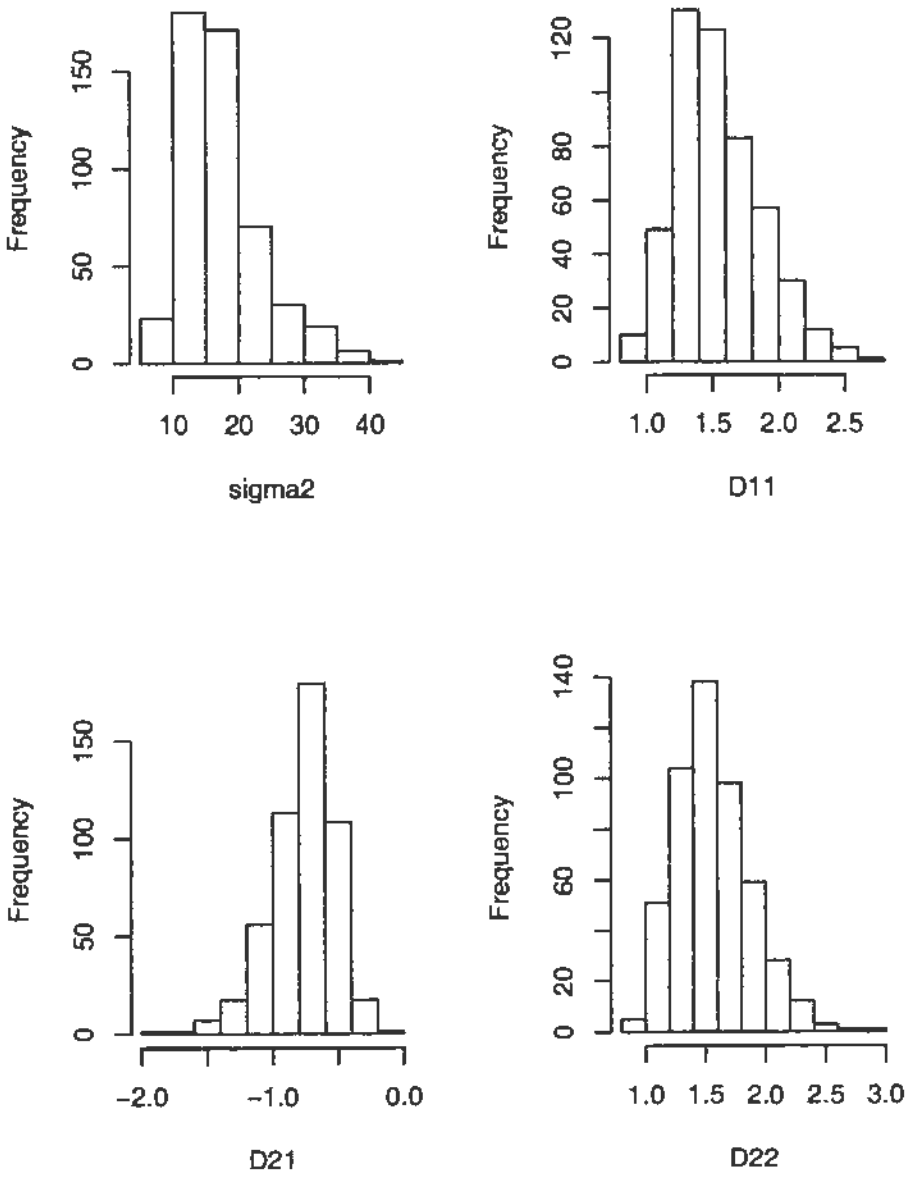

Figure 3.10: Posterior Distributions of σ^2 and the elements of $\mathbf{D}$ estimates in Algorithm 5, Case 2.

Parameter	Algorithm 4	Algorithm 5
β_1	1.846	1.201
β_2	1	1
β_3	1	1
β_4	3.582	1
σ^2	14.788	18.237
D_{11}	7.199	3.905
D_{21}	6.111	2.888
D_{22}	4.549	3.127

Table 3.12: Estimates of $\kappa = 1 + 2\sum_{k=1}^{\infty} \rho(k)$ for posterior estimates, Case 2, for binary data.

Parameter	Algorithm 4	Algorithm 5
β_1	2.741	1.209
β_2	1.208	1
β_3	1	1
β_4	5.653	1.205
σ^2	14.591	18.925
D_{11}	8.034	3.858
D_{21}	8.276	2.744
D_{22}	8.141	3.189

Table 3.13: Estimates of $\kappa = 1 + 2\sum_{k=1}^{\infty} \rho(k)$ for posterior estimates, Case 3, for binary data.

Chapter 4

Continuous Data: Example

4.1 CD4+ Data

The human immune deficiency virus (HIV) causes AIDS by reducing a person's ability to fight infection. HIV attacks an immune cell called the CD4+ cell which orchestrates the body's immunoresponse to infectious agents. An uninfected individual has around 1100 cells per millilitre of blood. CD4+ cells decrease in number with time from infection so that an infected person's CD4+ cell number can be used to monitor disease progression. Kaslow et al. (1987) collected values of CD4+ cell numbers along with other variables longitudinally for 369 infected men in a Multicenter AIDS Cohort study. Our goal is to analyze a portion of these data to determine what variables are useful in predicting the CD4+ cell count. The variables are discussed by Diggle, Liang and Zeger (1994).

Since CD4+ cell count is a discrete variable, it is inappropriate to use model (2.1), which is designed for continuous errors. However, Chib and Carlin (1999) and Chib and Jeliazkov (2001) show the square root of CD4+ cell count is a suitable transformation to allow one to use model (2.1), so we will use the same transformation.

Under model (2.1), we will use $n = 20$, and we have between 2 and 12 measurements on each subject. In our model, $\mathbf{X}_i = [\mathbf{1}, \mathbf{x}_{i,1}, \mathbf{x}_{i,2}, \mathbf{x}_{i,3}, \mathbf{x}_{i,4}]$, where

$\mathbf{x}_{i,1}$ represents recreational drug use ($x_{i,t1} = 1$ if drugs used),

$\mathbf{x}_{i,2}$ represents CESD, which is a mental illness score,

$\mathbf{x}_{i,3}$ is the subject's age (relative to an arbitrary origin) and

$\mathbf{x}_{i,4}$ is the number of packages of cigarettes smoked per day.

In addition, we define $\mathbf{W}_i$ as having the jth row $(1, t_{ij})$, where t_{ij} is the time since seroconversion for subject i. Therefore, $\mathbf{X}_i$ is $n_i \times 5$ and $\mathbf{W}_i$ is $n_i \times 2$.

We make the following prior assumptions on our parameters:
$\beta \sim N_5(\beta_o, \mathbf{B}_o)$, $\mathbf{b}_i \sim N_2(\mathbf{0}, \mathbf{D})$, $\mathbf{D}^{-1} \sim W_2(\rho_o, \mathbf{R}_o\rho_o^{-1})$ and $\sigma^2 \sim IG(\nu_o/2, \delta_o/2)$, where $\beta_o = (10, 0, 0, 0, 0)'$, $\mathbf{B}_o = \mathbf{I}$ and $\rho_o = 50$. In our analyses, we examine our results under a variety of choices for the other prior parameters. In each case, we will run Algorithms 1-3 of Chapter 2 for 500 iterations.

4.2 Results

We now present our analyses of the CD4+ data set, beginning with the posterior means and variances of the parameters under Algorithms 1-3. These are given in Tables 4.1-4.3. These tables refer to cases 1-4, which are defined as follows:
Case 1: $\nu_o = 1$, $\delta_o = 100$ and $\mathbf{R}_o = diag(2, 1)$. Therefore, σ^2 has a prior mean of 100.
Case 2: $\nu_o = 5$, $\delta_o = 100$ and $\mathbf{R}_o = diag(2, 1)$. Therefore, σ^2 has a prior mean of 20.
Case 3: $\nu_o = 1$, $\delta_o = 100$ and $\mathbf{R}_o = diag(10, 1)$. Therefore, σ^2 has a prior mean of 100.
Case 4: $\nu_o = 5$, $\delta_o = 100$ and $\mathbf{R}_o = diag(10, 1)$. Therefore, σ^2 has a prior mean of 20.

As we examine the results in Table 4.1, we see that, changing the prior of σ^2 has little effect on $mean(\hat{\beta})$ and $var(\hat{\beta})$. It also has little effect on $mean(\hat{\sigma}^2)$, but leads to a larger value of $var(\hat{\sigma}^2)$. Finally, it has little effect on $mean(\hat{\mathbf{D}})$ and $var(\hat{\mathbf{D}})$. Meanwhile, changing $\mathbf{R}_o$ to $diag(10, 1)$ causes $mean(\hat{\beta})$ to stay about the same and $var(\hat{\beta})$ to drop slightly. It also causes $mean(\hat{\sigma}^2)$ and $var(\hat{\sigma}^2)$ to drop, and leads to a drop in $mean(\hat{\mathbf{D}})$ and $var(\hat{\mathbf{D}})$.

Cases		Parameters								
		β_0	β_1	β_2	β_3	β_4	σ^2	D_{11}	D_{21}	D_{22}
Case 1	Mean	16.59	4.72	1.81	2.58	5.50	509.61	0.60	0.09	1.44
	Var	11.03	5.90	0.67	1.11	5.69	76524.67	0.02	0.03	0.15
Case 2	Mean	16.75	4.79	1.83	2.60	5.57	508.18	0.60	0.09	1.46
	Var	13.34	6.06	0.68	1.14	5.98	85511.87	0.02	0.04	0.20
Case 3	Mean	16.62	4.60	1.79	2.32	5.62	463.28	0.11	0.02	1.49
	Var	9.34	4.81	0.57	0.95	5.14	52317.44	0.0006	0.0056	0.17
Case 4	Mean	16.78	4.66	1.81	2.33	5.71	461.62	0.11	0.02	1.51
	Var	11.03	4.84	0.58	0.97	5.35	57270.49	0.0006	0.006	0.22

Table 4.1: Posterior means and variances of parameters in analysis of CD4+ Data set, using Algorithm 1.

From our analysis, based on examining $mean(\hat{\beta}_i)/\sqrt{var(\hat{\beta}_i)}$, it looks like CESD, age and cigarette smoking are important variables in predicting CD4+ cell count. It also appears that $\mathbf{D} \neq \mathbf{0}$, so the $\mathbf{b}_i$ term for time effect is needed in our model.

As we examine the results in Table 4.2, we see that changing the prior of σ^2 has little effect on $mean(\hat{\boldsymbol{\beta}})$ and $var(\hat{\boldsymbol{\beta}})$. It also has little effect on $mean(\hat{\sigma}^2)$ and $var(\hat{\sigma}^2)$. Finally, it has little effect on $mean(\hat{\mathbf{D}})$ and $var(\hat{\mathbf{D}})$. Meanwhile, changing $\mathbf{R}_o$ to $diag(10, 1)$ causes $mean(\hat{\boldsymbol{\beta}})$ and $var(\hat{\boldsymbol{\beta}})$ to drop, except for $mean(\hat{\beta}_0)$. It also causes $mean(\hat{\sigma}^2)$ and $var(\hat{\sigma}^2)$ to drop, and causes most elements of $mean(\hat{\mathbf{D}})$ and $var(\hat{\mathbf{D}})$ to drop.

From our analysis, it looks like drug use and cigarette smoking are important variables in predicting CD4+ cell count, it also appears that the $\mathbf{b}_i$ term for time effect is needed in our model. Therefore, the prior specified for σ^2 has some effect on our results.

As we examine the results in Table 4.3, we see that changing the prior of σ^2 has some effect on $mean(\hat{\boldsymbol{\beta}})$ and $var(\hat{\boldsymbol{\beta}})$, but it has little effect on $mean(\hat{\sigma}^2)$. From Cases 1 and 2, we see it has little effect on $var(\hat{\sigma}^2)$ and from Cases 3 and 4, it causes $var(\hat{\sigma}^2)$

Cases		Parameters								
		β_0	β_1	β_2	β_3	β_4	σ^2	D_{11}	D_{21}	D_{22}
Case 1	Mean	17.38	4.67	-0.10	0.27	1.80	341.88	111.59	-0.80	1.65
	Var	3.35	0.53	0.003	0.02	0.14	6388.45	2714.47	31.14	1.17
Case 2	Mean	17.69	4.66	-0.10	0.26	1.76	327.03	105.45	-0.68	2.02
	Var	4.51	0.54	0.003	0.02	0.17	6718.29	3587.39	36.85	2.18
Case 3	Mean	21.92	4.45	-0.07	0.05	1.15	175.30	0.16	0.36	5.06
	Var	0.73	0.38	0.001	0.004	0.03	1061.95	0.006	0.09	2.23
Case 4	Mean	21.92	4.45	-0.06	0.05	1.14	173.86	0.16	0.38	5.32
	Var	0.73	0.39	0.001	0.004	0.03	1022.04	0.006	0.09	2.42

Table 4.2: Posterior means and variances of parameters in analysis of CD4+ Data set, using Algorithm 2.

to drop. Finally, from Cases 1 and 2, we see it leads to larger values of $mean(\hat{\mathbf{D}})$ and $var(\hat{\mathbf{D}})$, from Cases 3 and 4, it has little change on $mean(\hat{\mathbf{D}})$ and causes $var(\hat{\mathbf{D}})$ to drop. Meanwhile, changing $\mathbf{R}_o$ to $diag(10, 1)$ has some effect on $mean(\hat{\boldsymbol{\beta}})$ and causes $var(\hat{\boldsymbol{\beta}})$ to drop. It also causes $mean(\hat{\sigma}^2)$ and $var(\hat{\sigma}^2)$ to drop. From Cases 1 and 3, it leads to larger values of $mean(\hat{\mathbf{D}})$ and $var(\hat{\mathbf{D}})$ except for $mean(\hat{D_{22}})$ and $var(\hat{D_{22}})$. In Cases 2 aud 4, it causes $mean(\hat{\mathbf{D}})$ to drop except for $mean(\hat{D_{11}})$ and causes $var(\hat{\mathbf{D}})$ to drop.

In Algorithm 3, the MCMC procedure does not move to new values very often (only about 4% of the time for Case 1, 5% of the time for Case 2, 6% of the time for Case 3, 3% of the time for Case 4), because of larger values of σ^2 in CD4+ Data set. From our analysis, it looks like there are no important variables in predicting CD4+ cell count in Cases 1 and 2, but drug use and cigarette smoking are important variables in predicting CD4+ cell count in Cases 3 and 4. It also appears that $\mathbf{D} \neq \mathbf{0}$, so the $\mathbf{b}_i$ term for time effect is needed in our model.

Since our Bayesian analysis yielded results with very large posterior estimates of σ^2, it was of interest to see if similar results were observed with a non-Bayesiau

Cases		Parameters								
		β_0	β_1	β_2	β_3	β_4	σ^2	D_{11}	D_{21}	D_{22}
Case 1	Mean	21.20	17.60	-0.65	-0.02	-1.77	332.33	0.05	0.12	1.03
	Var	253.89	144.49	1.13	0.22	70.24	799.47	0.02	0.003	0.005
Case 2	Mean	12.97	3.09	3.61	0.48	3.20	378.52	0.13	1.02	9.16
	Var	1016.91	608.34	17.53	6.15	57.57	852.68	0.012	0.131	1.33
Case 3	Mean	22.24	4.33	-0.06	0.02	1.10	157.77	7.16	0.64	0.38
	Var	0.35	0.27	0.001	0.002	0.03	116.46	0.13	0.009	0.0002
Case 4	Mean	22.35	4.27	-0.06	0.016	1.09	150.71	7.55	0.65	0.36
	Var	0.32	0.25	0.001	0.002	0.03	46.75	0.02	0.0013	0.00003

Table 4.3: Posterior means and variances of parameters in analysis of CD4+ Data set, using Algorithm 3.

analysis. We used the estimates given by Robinson (1991) for model (2.1) and found that $\hat{\sigma}^2 = 351.39$, which is similar to our posterior mean.

4.2.1 Comparison of Algorithms 1-3

If we compare Cases 1 and 2, we see that Algorithms 1-3 give some similar results for $mean(\hat{\boldsymbol{\beta}})$, but Algorithm 3 provided larger values for $var(\hat{\boldsymbol{\beta}})$ than the other two. Algorithms 2 and 3 lead to smaller $mean(\hat{\sigma}^2)$ values. We also see that Algorithm 3 leads to a drop in $var(\hat{\sigma}^2)$. Finally, Algorithms 1 and 3 give the smallest values of $mean(\hat{\mathbf{D}})$ and $var(\hat{\mathbf{D}})$.

From Cases 3 and 4, we see that Algorithms 2 and 3 have more similar values for $mean(\hat{\boldsymbol{\beta}})$ and $var(\hat{\boldsymbol{\beta}})$ than Algorithm 1. The three algorithms differ quite a bit among their $\hat{\sigma}^2$ and $\hat{\mathbf{D}}$ values in all cases. Finally, we note that Algorithm 1 tends to give the largest values of $var(\hat{\boldsymbol{\beta}})$ in Cases 3 and 4.

4.2.2 Autocorrelations of Posterior Estimates

As discussed in Chapter 3, it is also of interest to study the behaviour of the estimates using their autocorrelation function. First, we present the lag-1 autocorrelation values of the posterior estimates in Tables 4.4-4.7.

Parameter	Algorithm 1	Algorithm 2	Algorithm 3
β_0	-0.020	0.719	0.952
β_1	0.023	0.079	0.955
β_2	0.254	0.077	0.944
β_3	0.214	0.256	0.931
β_4	0.071	0.202	0.949
σ^2	-0.062	0.184	0.956
D_{11}	0.434	0.780	0.971
D_{21}	0.474	0.911	0.955
D_{22}	0.343	0.894	0.948

Table 4.4: Lag-1 autocorrelations of posterior estimates under Case 1, using CD4+ Data set.

From Tables 4.4 and 4.5, we see we are getting good performance for the β's and σ^2 for Algorithm 1. Algorithm 2 does not perform as well, particularly for the variance components. The results for Algorithm 3 are also very poor, mainly because there is little movement in the MCMC procedure. Finally, we see all algorithms give high lag-1 autocorrelation values for the elements of $\mathbf{D}$.

From Tables 4.6 and 4.7, we are getting better performance for most of the β's in Algorithm 2 and Algorithm 3 than in Algorithm 1. Algorithms 1 and 2 perform better for σ^2 than Algorithm 3, because there is little movement in the MCMC procedure in Algorithm 3. Finally, we see all algorithms give high autocorrelation values for the elements of $\mathbf{D}$.

To summarize the autocorrelations at all lags and their overall rate of decay, Tables

Parameter	Algorithm 1	Algorithm 2	Algorithm 3
β_0	-0.014	0.789	0.967
β_1	0.023	0.058	0.969
β_2	0.234	0.103	0.947
β_3	0.207	0.319	0.953
β_4	0.034	0.331	0.970
σ^2	-0.035	0.303	0.954
D_{11}	0.423	0.841	0.975
D_{21}	0.455	0.920	0.973
D_{22}	0.311	0.898	0.960

Table 4.5: Lag-1 autocorrelations of posterior estimates under Case 2, using CD4+ Data set.

4.8 to 4.11 give the autocorrelation time $\kappa = 1 + 2\sum_{k=1}^{\infty} \rho(k)$ for each parameter in Tables 4.4 to 4.7, where $\rho(k)$ is the autocorrelation at lag k for the parameter of interest. We estimated κ as discussed in Chapter 3.

From Tables 4.8 and 4.9, we see Algorithm 1 does reasonably well for all the parameters. Algorithm 2 does not perform as well. The results for Algorithm 3 are also very poor, mainly because there is little movement in the MCMC procedure.

From Tables 4.10 and 4.11, we are getting better performance for the β's for Algorithm 2 than Algorithms 1 and 3, although the differences are not dramatic. Algorithm 3 gives poor results for σ^2 and $\mathbf{D}$, again because there is little movement in the MCMC procedure.

Parameter	Algorithm 1	Algorithm 2	Algorithm 3
β_0	-0.012	0.071	0.158
β_1	0.044	0.060	0.180
β_2	0.262	0.033	-0.067
β_3	0.241	0.066	-0.028
β_4	0.076	-0.016	-0.022
σ^2	-0.034	0.049	0.952
D_{11}	0.365	0.835	0.961
D_{21}	0.497	0.799	0.941
D_{22}	0.358	0.515	0.923

Table 4.6: Lag-1 autocorrelations of posterior estimates under Case 3, using CD4+ Data set.

Parameter	Algorithm 1	Algorithm 2	Algorithm 3
β_0	-0.004	0.062	0.099
β_1	0.046	0.044	0.144
β_2	0.234	0.058	-0.064
β_3	0.233	0.050	-0.035
β_4	0.035	-0.039	-0.015
σ^2	-0.010	-0.001	0.964
D_{11}	0.374	0.830	0.946
D_{21}	0.493	0.797	0.974
D_{22}	0.332	0.513	0.915

Table 4.7: Lag-1 autocorrelations of posterior estimates under Case 4, using CD4+ Data set.

Parameter	Algorithm 1	Algorithm 2	Algorithm 3
β_0	4.21	24.58	20.94
β_1	2.67	1	23.21
β_2	1.83	1	27.17
β_3	1.98	9.60	22.17
β_4	2.13	8.21	19.95
σ^2	4.14	8.16	26.11
D_{11}	2.34	23.69	28.40
D_{21}	3.11	26.05	23.15
D_{22}	2.33	30.80	25.60

Table 4.8: Estimates of $\kappa = 1+2\sum_{k=1}^{\infty}\rho(k)$ for posterior estimates, Case 1, for CD4+ Data set.

Parameter	Algorithm 1	Algorithm 2	Algorithm 3
β_0	4.01	28.91	27.32
β_1	2.76	1	27.23
β_2	1.79	1.49	22.69
β_3	1.98	12.68	21.21
β_4	2.10	13.67	30.85
σ^2	3.72	13.24	28.45
D_{11}	2.29	28.40	31.75
D_{21}	3.04	28.06	31.10
D_{22}	2.23	32.17	26.08

Table 4.9: Estimates of $\kappa = 1+2\sum_{k=1}^{\infty}\rho(k)$ for posterior estimates, Case 2, for CD4+ Data set.

Parameter	Algorithm 1	Algorithm 2	Algorithm 3
β_0	3.58	1	2.60
β_1	1.78	1	1.98
β_2	1.82	1	1
β_3	2.45	1	1.47
β_4	1.90	1.21	1.24
σ^2	3.49	1	30.02
D_{11}	1.73	11.15	31.47
D_{21}	3.22	9.87	23.20
D_{22}	2.28	3.16	20.76

Table 4.10: Estimates of $\kappa = 1 + 2\sum_{k=1}^{\infty} \rho(k)$ for posterior estimates, Case 3, for CD4+ Data set.

Parameter	Algorithm 1	Algorithm 2	Algorithm 3
β_0	3.76	1	1
β_1	2.19	1.21	1.29
β_2	1.77	1.23	1
β_3	2.44	1	1.44
β_4	1.85	1	1.65
σ^2	3.39	1	31.83
D_{11}	1.75	11.09	24.07
D_{21}	3.22	9.69	31.47
D_{22}	2.18	3.11	23.66

Table 4.11: Estimates of $\kappa = 1 + 2\sum_{k=1}^{\infty} \rho(k)$ for posterior estimates, Case 4, for CD4+ Data set.

Chapter 5

Binary Data: Example

5.1 Six Cities data set: child's wheeze status

Our data set contains complete records on 537 children from Steubenville, Ohio, each of whom was examined annually at ages 7 through 10. This data set was previously analysed by Zeger, Liang and Albert (1988). The repeated binary response is the wheezing status (1 =yes, 0 =no) of a child at each occasion. Maternal smoking was categorized as 1 if the mother smoked regularly and 0 otherwise. Although maternal smoking is a time-varying covariate, it was treated as fixed at its value at the first year of study.

When the responses are binary, a natural choice is to use a logit link function to relate the marginal expectation of the responses to the covariates. Suppose we have a sequence of binary measurements $\mathbf{y}_i = (y_{i1}, ..., y_{in_i})'$, where $y_{it} = 0$ or 1, on the i^{th} unit taken at n_i specific time points. We define the logit link as:

$$Pr(y_{it} = 1) = \frac{\exp[\mathbf{x}'_{it}\boldsymbol{\beta} + \mathbf{w}'_{it}\mathbf{b}_i]}{1 + \exp[\mathbf{x}'_{it}\boldsymbol{\beta} + \mathbf{w}'_{it}\mathbf{b}_i]}.$$

The covariates can be both time-stationary, i.e. constant across occasions, and time-varying. For example, in the Six Cities study (Ware et al., 1984), a child's wheeze status (yes, no) as well as information about maternal smoking were recorded annually

for a sample of children from each of the participating cities. In this example, maternal smoking is time-varying, since it can change from year to year, whereas city is time-stationary.

We are using a subset of data from the Six Cities study, a longitudinal study of the health effects of air pollution, in our model. Rather than using a logit link, we will use the probit link discussed in Chapter 2:

$$Pr(y_{it} = 1|\mathbf{b}_i) = \mathbf{\Phi}(\mathbf{x}'_{it}\boldsymbol{\beta} + \mathbf{w}'_{it}\mathbf{b}_i)$$

where $\mathbf{\Phi}$ is the standard normal cdf and $\mathbf{x}'_{it}$ and $\mathbf{w}'_{it}$ are the tth rows of $\mathbf{X}_i$ and $\mathbf{W}_i$, respectively. $\mathbf{X}_i$ is an $n_i \times p$ design matrix of covariates and $\boldsymbol{\beta}$ is a corresponding $p \times 1$ vector of fixed effects. In addition, $\mathbf{W}_i$ is a $n_i \times q$ design matrix and $\mathbf{b}_i$ is a $q \times 1$ vector of subject-specific random effects. Under our model, we have $n = 537$, and we have $n_i = 4$ measurements on each subject.

The marginal expectation of the response is modelled as a probit function of three covariates: age, maternal smoking, and the age-maternal smoking interaction. One of the objectives of this study was to determine the effects of age, maternal smoking and the age-maternal smoking interaction.

We make the following prior assumptions on our parameters:
$\boldsymbol{\beta} \sim N_4(\boldsymbol{\beta}_o, \mathbf{B}_o)$, $\mathbf{b}_i \sim N_1(0, D)$, $D^{-1} \sim W_1(\rho_o, R_o\rho_o^{-1})$. If D^{-1} is $W_1(\rho_o, R_o\rho_o^{-1})$, then $D^{-1}/R_o\rho_o^{-1}$ is $\chi^2_{\rho_o}$ and $\sigma^2 \sim IG(\nu_o/2, \delta_o/2)$, where $\boldsymbol{\beta}_o = (10, 0, 0, 0)$, $\mathbf{B}_o = \mathbf{I}$ and $\rho_o = 50$. In our analyses, we examine our results under a variety of choices for the other prior parameters. In each case, we will run Algorithms 4-5 of Chapter 2 for 500 iterations.

5.2 Results

We now present our analyses of the Six Cities data set, beginning with the posterior means and variances of the parameters under Algorithms 4-5. These are given in Tables 5.1-5.2. Table 5.1 refers to Cases 1-5 for Algorithm 4 and Table 5.2 refers to

cases 1-3 for Algorithm 5, which are defined as follows:

In Algorithm 4:

Case 1: $\nu_o = 1$, $\delta_o = 100$ and $R_o = 20$. Therefore, σ^2 has a prior mean of 100.

Case 2: $\nu_o = 5$, $\delta_o = 100$ and $R_o = 20$. Therefore, σ^2 has a prior mean of 20.

Case 3: $\nu_o = 5$, $\delta_o = 5$ and $R_o = 20$. Therefore, σ^2 has a prior mean of 1.

Case 4: $\nu_o = 5$, $\delta_o = 10$ and $R_o = 20$. Therefore, σ^2 has a prior mean of 2.

Case 5: $\nu_o = 5$, $\delta_o = 50$ and $R_o = 20$. Therefore, σ^2 has a prior mean of 10.

In Algorithm 5:

Case 1: $\nu_o = 1$, $\delta_o = 100$ and $R_o = 20$. Therefore, σ^2 has a prior mean of 100.

Case 2: $\nu_o = 5$, $\delta_o = 100$ and $R_o = 20$. Therefore, σ^2 has a prior mean of 20.

Case 3: $\nu_o = 5$, $\delta_o = 5$ and $R_o = 20$. Therefore, σ^2 has a prior mean of 1.

Cases		Parameters					
		β_0	β_1	β_2	β_3	σ^2	D
Case 1	Mean	-0.38	-0.60	-0.41	-0.58	256.45	0.052
	Var	0.95	0.52	1.03	0.53	439.55	0.00011
Case 2	Mean	-0.42	-0.45	-0.45	-0.44	152.78	0.054
	Var	0.92	0.52	0.99	0.53	150.86	0.00016
Case 3	Mean	-1.26	-10.97	-1.28	-10.95	1.73×10^7	4.04×10^7
	Var	1.46	39.25	1.69	39.15	3.17×10^{13}	1.31×10^{14}
Case 4	Mean	-1.25	-10.90	-1.28	-10.88	1.72×10^7	4.04×10^7
	Var	1.46	40.81	1.69	41.04	3.53×10^{13}	1.51×10^{14}
Case 5	Mean	-0.42	-0.46	-0.45	-0.44	157.08	0.06
	Var	0.92	0.52	0.995	0.53	170.76	0.00096

Table 5.1: Posterior means and variances of parameters in analysis of Six Cities data set, using Algorithm 4.

We include Case 3 for each algorithm because we hope the result will give us some

insight into the assumption of Chib and Carlin (1999) to assume $\sigma^2 = 1$. We are placing a prior mean of 1 on σ^2 in Case 3, and it will be of interest to see if the posterior of σ^2 changes from our prior assumption. If it does, the setting of $\sigma^2 = 1$ by Chib and Carlin (1999) will be seen as questionable.

As we examine the results in Table 5.1, for Cases 1 and 2, we see that $mean(\hat{\beta})$, $var(\hat{\beta})$, $mean(\hat{D})$ and $var(\hat{D})$ are similar. Meanwhile, by decreasing the prior mean of σ^2, $mean(\hat{\sigma}^2)$ and $var(\hat{\sigma}^2)$ decrease. For Cases 3 and 4, we see $mean(\hat{\beta})$, $var(\hat{\beta})$, $mean(\hat{\sigma}^2)$, $var(\hat{\sigma}^2)$ remain unchanged. However, the values for $mean(\hat{D})$ and $var(\hat{D})$ are large, and are not sensible. Finally, Case 5 gives very similar results to Case 2. In all cases, there is little evidence to suggest that any of the variables are useful in predicting wheeze status. The results for Case 3 also indicate that setting $\sigma^2 = 1$, as Chib and Carlin (1999) would suggest, would give very unreliable posterior estimates in this example.

Cases		Parameters					
		β_0	β_1	β_2	β_3	σ^2	D
Case 1	Mean	-0.35	-1.09	-0.33	-1.14	853.24	0.058
	Var	1.02	0.56	1.05	0.56	9382.63	0.00016
Case 2	Mean	-0.44	-0.75	-0.35	-0.80	456.79	0.055
	Var	1.03	0.53	0.94	0.56	49108.86	0.000096
Case 3	Mean	-0.44	-0.70	-0.35	-0.76	405.59	0.168
	Var	1.05	0.52	0.97	0.55	36957.7	0.1707

Table 5.2: Posterior means and variances of parameters in analysis of Six Cities data set, using Algorithm 5.

As we examine the results in Table 5.2, comparing Cases 1 and 2, we see $mean(\hat{\beta})$, $var(\hat{\beta})$, $mean(\hat{D})$ and $var(\hat{D})$ are similar. By decreasing the prior mean of σ^2, $mean(\hat{\sigma}^2)$ goes down, but $var(\hat{\sigma}^2)$ goes up. This is not what we would expect, and the variance is proportional to the mean for a Gamma distribution. From Cases 2 and 3, $mean(\hat{\beta})$, $var(\hat{\beta})$, $mean(\hat{\sigma}^2)$ and $var(\hat{\sigma}^2)$ stay about same, but in Case 3 (where

the prior mean of σ^2 is 1), $mean(\hat{D})$ and $var(\hat{D})$ are larger than Cases 1 and 2.

5.2.1 Comparison of Algorithms 4 and 5

From Cases 1 and 2, we see that Algorithms 4 and 5 give similar values for $mean(\hat{\beta})$, $var(\hat{\beta})$, $mean(\hat{D})$ and $var(\hat{D})$. However, Algorithm 4 gives smaller values for $mean(\hat{\sigma}^2)$ and $var(\hat{\sigma}^2)$. In Case 3, Algorithm 4 provided larger values for all posterior estimates. Also, just like with Algorithm 4, the variables do not appear useful in predicting wheeze status.

5.2.2 Autocorrelations of Posterior Estimates

Running our various MCMC algorithms for these data and model for 500 iterations each produces the lag-1 autocorrelation summaries in Tables 5.3 and 5.4. These tables show the lag 1 sample autocorrelations for Algorithm 4 and Algorithm 5 for some of our cases.

	Case 1		Case 2	
Parameter	Algorithm 4	Algorithm 5	Algorithm 4	Algorithm 5
β_1	-0.0616	0.0222	-0.0649	0.0181
β_2	0.0684	-0.0435	0.0645	0.0409
β_3	0.0269	-0.0243	0.0299	-0.0095
β_4	0.0648	-0.0517	0.0625	0.0921
σ^2	0.5118	0.9057	0.4694	0.9836
D_{11}	0.9093	0.9140	0.9167	0.8960

Table 5.3: Lag-1 autocorrelations of posterior estimates, using Six Cities data set.

From Table 5.3, we are getting similar performance for the β's for Algorithms 4 and 5 in Cases 1 and 2. We also see that both algorithms give high autocorrelation values for σ^2 and D.

	Case 3		Case 4	Case 5
Parameter	Algorithm 4	Algorithm 5	Algorithm 4	Algorithm 4
β_1	0.2809	0.0144	0.2739	-0.0646
β_2	0.9642	0.0317	0.9676	0.0639
β_3	0.3712	-0.0028	0.3652	0.0301
β_4	0.9613	0.0809	0.9640	0.0632
σ^2	0.9729	0.9824	0.9775	0.4986
D_{11}	0.9340	0.9578	0.9475	0.9407

Table 5.4: Lag-1 autocorrelations of posterior estimates, using Six Cities data set.

In Table 5.4 for Case 3, when the prior mean of σ^2 is 1, Algorithm 5 provides better results (lower autocorrelations) for the β's than Algorithm 4. This should not be a surprise, since Algorithm 4 gave very large, unstable values for $mean(\hat{\sigma}^2)$ and $var(\hat{\sigma}^2)$. From Table 5.4 for Algorithm 4, we see when the prior mean of σ^2 is 1 or close to 1, from Cases 3 and 4, Algorithm 4 shows high autocorrelation values for β's and σ^2. So, the prior assumption on σ^2 has an effect on these algorithms.

Tables 5.5 and 5.6 give the autocorrelation time $\kappa = 1 + 2\sum_{k=1}^{\infty} \rho(k)$ for each parameter in the probit model, where $\rho(k)$ is the autocorrelation at lag k for the parameter of interest. From Table 5.5, we obtained similar performance for the β's for both algorithms, but we see all algorithms give high autocorrelation values for σ^2 and D. In Table 5.6 for Case 3, when the prior mean of σ^2 is 1, Algorithm 5 provides better results for the β's than Algorithm 4.

Again, we are seeing the prior assumption on the distribution of σ^2 does have an effect on other posterior estimates, so we should not simply choose $\sigma^2 = 1$ in all applications.

Parameter	Case 1		Case 2	
	Algorithm 4	Algorithm 5	Algorithm 4	Algorithm 5
β_1	1	1	1	1
β_2	1	1.265	1	1.849
β_3	1	1.218	1	1
β_4	1	1.218	1	2.633
σ^2	18.099	21.636	16.737	37.245
D_{11}	18.868	16.886	19.151	20.106

Table 5.5: Estimates of $\kappa = 1 + 2\sum_{k=1}^{\infty} \rho(k)$ for posterior estimates, for Six Cities data set.

Parameter	Case 3		Case 4	Case 5
	Algorithm 4	Algorithm 5	Algorithm 4	Algorithm 4
β_1	9.956	1	9.788	1
β_2	26.955	1.804	25.372	1
β_3	11.782	1	11.293	1
β_4	27.381	1.965	25.557	1
σ^2	33.601	37.074	34.591	17.720
D_{11}	27.708	25.507	30.409	21.262

Table 5.6: Estimates of $\kappa = 1 + 2\sum_{k=1}^{\infty} \rho(k)$ for posterior estimates, for Six Cities data set.

Bibliography

[1] Albert, J.H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *J. Amer. Statist. Assoc.* **88**, 669-79.

[2] Albert, J. and Chib, S. (1996). Bayesian modeling of binary repeated measures data with application to crossover trials. *In Bayesian Biostatistics*, D. A. Berry and D. K. Stangl, eds., New York: Marcel Dekker, 577-99.

[3] Carlin, B.P. and Polson, N.G. (1992). A Monte Carlo Bayesian methods for discrete regression models and categorical time series. *In Bayesian Statistics*. **4**, J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds., Oxford: Oxford University Press, 577-86.

[4] Chib, S. (1995). Marginal likelihood from the Gibbs output. *J. Amer. Statist. Assoc.* **90**, 1313-21.

[5] Chib, S. and Greenberg E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician* **49**, 327-35.

[6] Chib, S. and Greenberg E. (1998). Analysis of multivariate probit models. *Biometrika* **85**, 347-61.

[7] Chib, S. and Carlin, B.P. (1999). On MCMC sampling in hierarchical longitudinal models. *Statistics and Computing* **9**, 17-26.

[8] Chib, S. and Jeliazkov, I. (2001). Marginal likelihood from the Metropolis-Hastings output. *J. Amer. Statist. Assoc.* **96**, 270-281.

[9] Diggle, P.J., Liang, K.Y. and Zeger, S.L. (1994). *Analysis of Longitudinal Data.* Clarendon Press, Oxford.

[10] Fitzmaurice, G.F.V. and Laird, N.M. (1993). A likelihood-based method for analysing longitudinal binary responses. *Biometrika* **80**, 141-51.

[11] Gelfand, A.E. and Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85**, 398-409.

[12] Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97-109.

[13] Johnson, R.A. and Wichern, D.W. (1992). *Applied Multivariate Statistical Analysis*, Third Edition. Prentice Hall, Englewood Cliffs, New Jersey.

[14] Kaslow, R.A., Ostrow, D.G., Detels, R. et al. (1987). The Multicenter AIDS Cohort Study: rationale, organization and selected characteristics of the participants. *American Journal of Epidemiology* **126**, 310-18.

[15] Kass, R.E., Carlin, B.P., Gelman, A. and Neal, R. (1998). Markov chain Monte Carlo in practice: A round table discussion. *The American Statistician* **52**, 93-100.

[16] Laird, N.M. and Ware, J.H. (1982). Random-effects models for longitndinal data. *Biometrics* **38**, 963-974.

[17] Lindley, D.V. and Smith, A.F.M. (1972). Bayes estimates for the linear model (with discussion). *J. Roy. Statist. Soc.*, Ser. B, **34**, 1-41.

[18] Lindstrom, M.J., and Bates, D.M. (1988). Newton-Raphson and EM algorithms for linear mixed-effects models for repeated measure data. *J. Amer. Statist. Assoc.* **83**, 1014-1022.

[19] Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics.* **21**, 1087-1092.

[20] Müller, P. (1993). A generic approach to posterior integration and Gibbs sampling. *unpublished manuscript.*

[21] Muirhead, R.J. (1982). *Aspects of Multivariate Statistical Theory.* John Wiley and Sons, Inc., New York.

[22] Robinson, G.K. (1991). That BLUP is a good thing: The estimation of random effects. *Statistical Science* **6**, 15-51.

[23] Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Ann. Statist.* **22**, 1701-1762.

[24] Ware, J.H., Dockery, D.W., Spiro, A. III, Speizer, F.E. and Ferris, B.G., Jr. (1984). Passive smoking, gas cooking and respiratory health in children living in six cities. *Am. Rev. Respir. Dis.* **129**, 366-74.

[25] Zeger, S.L. and Karim, M.R. (1991). Generalized linear models with random effects; a Gibbs sampling approach. *J. Amer. Statist. Assoc.* **86**, 79-86.

[26] Zeger, S.L., Liang, K.Y. and Albert, P.S. (1988). Models for longitudinal data: A generalized estimating equation approach. *Biometrics* **44**. 1049-60.