University of Southampton

Faculty of Engineering, Science and Mathematics

School of Mathematics

On Bayesian Inference for Partially Observed Data

by

Roger C. Gill

Thesis for the degree of Doctor of Philosophy

October 2007

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF ENGINEERING, SCIENCE AND MATHEMATICS

SCHOOL OF MATHEMATICS

Doctor of Philosophy

ON BAYESIAN INFERENCE FOR PARTIALLY OBSERVED DATA

by Roger Charles Gill

When making predictions, analysing incomplete data from a medical trial or drawing inference from artificially altered data, one is required to make conditional probability statements concerning unobserved individuals or data. This thesis provides a collection of statistical techniques for inference when data is only partially observed. An efficient reversible jump Markov chain Monte Carlo algorithm for generalised linear models is constructed. This provides a formal framework for Bayesian prediction under model uncertainty. The construction of the algorithm is unique, relying on a simple and novel reversible jump transformation function. The resulting algorithm is easy to implement and requires no 'expert' knowledge.

An inference framework for multivariate survey data subject to non-response is provided. Deviations from a 'close to ignorable' model are permitted through realistic a-priori changes in log-odds ratios. These a-priori deviations encode the prior belief that the non-response mechanism is non-ignorable.

A current disclosure control technique is studied. This technique rounds partially observed data prior to release. A Bayesian assessment of this technique is given. This requires the construction of a Metropolis-Hastings algorithm, and the algorithms irreducibility is proven and discussed.

Contents

List of Tables

List of Figures

ACKNOWLEDGEMENTS

I would like to thank my supervisor Professor Jon Forster for his guidance and assistance throughout the three years of my Ph. D research. I would also like to thank Dr. Sujit Sahu for acting as an advisor.

Thanks also go to the many researchers with whom I have shared room 9007.

I would also like to thank my family. Without their love and support this thesis would not have been possible.

Finally, I would like to thank Helen for her love, support and tireless patience. Her smile has brightened many days and she has made my time at university a joy.

Chapter 1

Introduction

1.1 The Problem of Missing Data

Even in well designed experiments the loss of data is a frequent occurrence.

In a medical situation a patient may leave a clinical trial for reasons unconnected to the treatment received.

There is frequently non-response in surveys. Personal income may be undisclosed by respondents of a household survey, or respondents to a public opinion poll may fail to reveal their political affiliation.

Consider the issue of disclosure control where a statistical agency might release data that has been artificially altered to safeguard confidentiality. A table count may be rounded, or individual counts omitted prior to the release of the data. The agency might even release marginal counts alone, artificially creating missing counts.

As a final example consider inference from a finite population. Let N individuals be characterised by some factor of interest taking values ψ_i for $i \in \{1, ..., N\}$. The population might be voters in an election where ψ_i assumes the value 1 if individual i intends to vote for political party X and 0 otherwise. Data typically comprises of n individuals from the population and an estimate of the total $\rho = \sum_{i=1}^{N} \psi_i$ is required. Inference for ρ is achieved through the posterior distribution $\pi(\rho|\psi_1, ..., \psi_n)$. Since

$\rho = \sum_{i=1}^{n} \psi_i + \sum_{i=n+1}^{N} \psi_i$ the posterior expectation of ρ conditional on observing data $\psi_1, ..., \psi_n$ is given by

$$\mathbb{E}[\rho|\psi_1, ..., \psi_n] = \sum_{i=1}^{n} \psi_i + \sum_{i=n+1}^{N} \mathbb{E}[\psi_i|\psi_1, ..., \psi_n],$$

where the expectation is with respect to the posterior predictive density. Hence inference about ρ involves making conditional probability statements about the $N-n$ unobserved ψ_i's.

When making predictions, analysing incomplete data from a medical trial or drawing inference from artificially altered data, one is required to make conditional probability statements concerning unobserved individuals or data.

1.2 Aims and Outlines of the Thesis

The aim of this thesis is to provide a collection of statistical techniques for inference when data is only partially observed.

Some specific objectives are as follows.

- Construct an efficient reversible jump Markov chain Monte Carlo algorithm for generalised linear models.
- Provide an inference framework for multivariate survey data subject to non-response.
- Provide a Bayesian assessment of rounding based disclosure control.

The structure of this thesis is as follows.

Chapter 2 provides an introduction to Bayesian statistics, missing data analysis and Markov chain Monte Carlo methods. The role of Chapter 2 is to provide a general overview of these subjects, and to introduce the concepts used throughout the thesis.

The construction of an efficient reversible jump Markov chain Monte Carlo algorithms for generalised linear models is the focus of Chapter 3. A novel reversible jump transformation function is introduced and applied in numerous examples.

Inference for survey data subject to non-response forms the basis of Chapter 4. An attempt to discriminate between non-response models is conducted using methods developed in Chapter 3. Uncertainty about ignorability of non-response is incorporated by introducing sensitivity parameters into log-linear models.

Statistical disclosure control is discussed in Chapter 5. A current disclosure limitation technique is extensively examined. A Markov chain is required to assess the technique and a proof of irreducibility is given and discussed.

The conclusions of this thesis are given in Chapter 6, and recommendations for future work are suggested.

Chapter 2

Literature Review

Making inference is the fundamental problem of statistics. Having observed data we wish to make statements, or inferences, about unknown features of the data generating process. This problem has received considerable attention since the rigorous study of statistical and probability theory began. Many different *Theories of Inference* have been proposed often with considerable controversy and criticism. One theory is that of *Bayesian Inference.*

2.1 Bayesian Inference

Interest in Bayesian inference has grown substantially in recent years. This growth is due, in no small part, to recent advances in computational statistics and the ubiquity of fast computing machines. These advances have enabled the fitting of complex models with relative ease and minimal computational expense. That said, a valid explanation for the growth is that the Bayesian approach is fundamentally sound, and enables a researcher to produce clear and direct inferences making full use of ***all*** available information.

From a Bayesian perspective observables and parameters of the statistical model are both considered random quantities. Denote $\boldsymbol{y}$ the observed data and let $\boldsymbol{\theta}$ denote parameters of the statistical model. Then formal inference is concerned with the joint probability distribution $f(\boldsymbol{y}, \boldsymbol{\theta})$. This joint probability distribution can be factorised as follows

$$f(\boldsymbol{y}, \boldsymbol{\theta}) = f(\boldsymbol{y}|\boldsymbol{\theta})f(\boldsymbol{\theta}). \tag{2.1}$$

The joint distribution comprises of two distinct parts: A *prior* distribution for the model parameters denoted $f(\boldsymbol{\theta})$, and a *likelihood* $f(\boldsymbol{y}|\boldsymbol{\theta})$. Prior to observing any data full probability statements about $\boldsymbol{\theta}$ can be made through the prior distribution $f(\boldsymbol{\theta})$. The likelihood function is regarded as a function of $\boldsymbol{\theta}$ fixed for the observed data $\boldsymbol{y}$. *Subjectively* $f(\boldsymbol{y}|\boldsymbol{\theta})$ measures our belief in the data taking certain values given hypothetical values of $\boldsymbol{\theta}$. It is *our* subjective view about the data generating process.

2.1.1 Bayes Theorem

Having observed data $\boldsymbol{y}$, Bayes theorem is used to determine the *posterior* distribution of $\boldsymbol{\theta}$. Bayes theorem in its continuous form is given below:

$$f(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})}{\int f(\boldsymbol{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})d\boldsymbol{\theta}} = \frac{f(\boldsymbol{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})}{f(\boldsymbol{y})}. \tag{2.2}$$

The integral in the denominator is over the parameter space of $\boldsymbol{\theta}$ denoted Θ. The *posterior* distribution $f(\boldsymbol{\theta}|\boldsymbol{y})$ encapsulates all that is known about $\boldsymbol{\theta}$ in light of observed data and other available prior information. Note that, in (2.2), the denominator $f(\boldsymbol{y})$ does not depend upon $\boldsymbol{\theta}$ and so acts as a constant ensuring that

$$\int f(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta} = 1.$$

As a result Bayes theorem is frequently written as follows:

$$f(\boldsymbol{\theta}|\boldsymbol{y}) \propto f(\boldsymbol{y}|\boldsymbol{\theta})f(\boldsymbol{\theta}). \tag{2.3}$$

In words we state 'the posterior distribution is proportional to the likelihood multiplied by the prior distribution'.

It is clear how the two information sources are combined together to form the posterior distribution. The information we have concerning $\boldsymbol{\theta}$ has been updated from prior to posterior using the data. Full probability statements about $\boldsymbol{\theta}$ are now made through the distribution $f(\boldsymbol{\theta}|\boldsymbol{y})$.

Having calculated the posterior distribution we wish to make inferential statements concerning the conditional density of $\boldsymbol{\theta}$ given $\boldsymbol{y}$. This is often straightforward as many features of the posterior distribution, such as moments or probabilities, can be expressed in terms of posterior expectations of functions of $\boldsymbol{\theta}$. The posterior expectation of a function $g(\boldsymbol{\theta})$ is given as follows

$$\mathbb{E}[g(\boldsymbol{\theta})|\boldsymbol{y}] = \frac{\int g(\boldsymbol{\theta})f(\boldsymbol{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})d\boldsymbol{\theta}}{\int f(\boldsymbol{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})d\boldsymbol{\theta}}. \tag{2.4}$$

If $g(\boldsymbol{\theta}) = \boldsymbol{\theta}$ then $\mathbb{E}[g(\boldsymbol{\theta})|\boldsymbol{y}] = \mathbb{E}[\boldsymbol{\theta}|\boldsymbol{y}]$ is simply the posterior mean. These expectations are an essential summary of the inference process.

2.1.2 Posterior Inference

Informal summaries of the posterior distribution can provide clear and meaningful answers to questions of interest. Initially we might plot the posterior distribution, but in many cases this fails to convey information in a useful form. This is particularly true if $\boldsymbol{\theta}$ is of high dimensions and only margins of $\boldsymbol{\theta}$ can be plotted.

Quantitative summaries of the posterior, such as a measure of location or dispersion are useful. Point estimates of these measures are often given. The posterior mean given in (2.4), for example, is sometimes used as an estimate of the measure of location.

We can readily construct informal probability intervals from the posterior distribu-

tion. For example a 95% *credible interval* for $\boldsymbol{\theta}$ is constructed by calculating the real values a and b such that $P(\boldsymbol{\theta} < a|y) = 0.025$ and $P(\boldsymbol{\theta} > b|y) = 0.025$.
An informal hypothesis test that $\boldsymbol{\theta}$ lies in some region R, would be to calculate the probability that $\boldsymbol{\theta} \in R$ given the observed data $\boldsymbol{y}$. This is often an easy calculation if the posterior density for $\boldsymbol{\theta}$ is known.

However, formal 'Bayesian inference' is concerned with deriving optimal *probability* statements from the posterior distribution. Using the posterior mean as an estimate of $\boldsymbol{\theta}$ is optimal in the sense that it minimises the expected squared error (where the expectation is taken with respect to the posterior distribution of $\boldsymbol{\theta}$). This is seen below

$$\begin{aligned}\mathbb{E}[(d-\boldsymbol{\theta})^2|\boldsymbol{y}] &= \mathbb{E}[d^2|\boldsymbol{y}] - 2d\mathbb{E}[\boldsymbol{\theta}|\boldsymbol{y}] + \mathbb{E}[\boldsymbol{\theta}^2|\boldsymbol{y}] \\ &= (d - \mathbb{E}[\boldsymbol{\theta}|\boldsymbol{y}])^2 + var(\boldsymbol{\theta}|\boldsymbol{y}).\end{aligned}$$

Since the above equation is a quadratic and $var(\boldsymbol{\theta}|\boldsymbol{y}) \geq 0$ it is clearly minimised at

$$d = \mathbb{E}[\boldsymbol{\theta}|\boldsymbol{y}].$$

Alternative error functions would result in different estimates for $\boldsymbol{\theta}$. If absolute error $|d-\boldsymbol{\theta}|$ is used the resulting estimate of $\boldsymbol{\theta}$ is the posterior median. In this framework the error function is viewed as a *loss* function. This loss function is a measure of how good or bad the estimate d of $\boldsymbol{\theta}$ is deemed to be if the true value of $\boldsymbol{\theta}$ is known. It is a random variable and allows a *decision* to be made that maximises the subjective expected utility.
Formal inferences can also take the form of intervals for $\boldsymbol{\theta}$. Here a loss function will penalise an interval if it fails to contain the true underlying value of $\boldsymbol{\theta}$, or if this interval is large. In practice, we fix the probability of the interval containing $\boldsymbol{\theta}$ and then find the smallest of all these intervals. I.e. the smallest interval given by

$$\mathcal{I} = \{\boldsymbol{\theta} : f(\boldsymbol{\theta}|\boldsymbol{y}) > c\},$$

where c is chosen such that the interval satisfies the desired probability. This interval is called the *highest posterior density interval*, and may be a differ from the informal credible interval.

We have now described all key elements to Bayesian inference. In summary, the first step it to describe the data generating process. That is

1. Obtain the likelihood function $f(\boldsymbol{y}|\boldsymbol{\theta})$.

2. Obtain the prior density $f(\boldsymbol{\theta})$. What do we know about $\boldsymbol{\theta}$ prior to observing $\boldsymbol{y}$?

3. Apply Bayes' theorem in either its discrete or continuous form to obtain the posterior distribution $f(\boldsymbol{\theta}|\boldsymbol{y})$.

4. Derive appropriate inference statements either informally or formally. These statements often involve the calculation of an expectation with respect to posterior distribution.

2.1.3 The Prior Distribution

In order to assign a probability distribution to parameters we need to adopt a *subjective* interpretation of probability. According to this definition, probability is represented as a personal degree of belief. This has attracted considerable criticism from opponents of Bayesian statistics. These antagonists view science as objective with no room for individual opinion, and probability as logical. The prior distribution, often formulated using expert opinion or past data, suffers the full force of these critiques. Many argue that two experts with identical prior information may formulate entirely different prior distributions in shape and form and that this may lead to contradicting posterior distributions. A Bayesian would argue that if the data are strong and the prior is constructed on reasonable grounds this element of personal opinion will not matter, and inferences will be robust to slight differences in prior formulation. What is clear is that prior information exists and is often extremely useful.

In the past prior distributions have often been chosen for their convenience and to facilitate the calculation of the posterior distribution. Suppose that data $\boldsymbol{y}$ are observed with likelihood given by $f(\boldsymbol{y}|\boldsymbol{\theta})$. Then a family of prior distributions $\mathcal{F}$ for $\boldsymbol{\theta}$ is conjugate with respect to the likelihood if the posterior $f(\boldsymbol{\theta}|\boldsymbol{y}) \propto f(\boldsymbol{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})$ is also a member of $\mathcal{F}$. Since the family $\mathcal{F}$ is generally well known and understood posterior summaries can be easily evaluated. This might not have been the case had we selected $f(\boldsymbol{\theta})$ from an alternative family of distributions.

2.1.3.1 Informative Prior Distributions

When there is genuine prior information available this needs to be formulated in terms of a prior density function for $\boldsymbol{\theta}$. This is not straightforward to do. If the prior information is the opinion of an expert it generally does not take the form of a complete density function $f(\boldsymbol{\theta})$. In practice we specify values that explain important features of the prior information, such as a measure of location and spread, then simple choose a convenient and practical $f(\boldsymbol{\theta})$ that has these properties. Since interpretations of the same prior information may differ the process is imperfect. We should therefore perform sensitivity analyses to determine if posterior inferences are supported unequivocally by the evidence. A comprehensive discussion on the formulation of informative prior distributions can be found in Chapter 6 of O'Hagan and Forster (2004).

2.1.3.2 Non-Informative Prior Distributions

As we have seen, the specification of the prior distribution plays an important role in the inference process. We often have weak, or in many cases no, prior information for $\boldsymbol{\theta}$. To specify a prior that represents no information, we often assume what is called a 'flat' or 'uniform' prior for $\boldsymbol{\theta}$. I.e. we assume $f(\boldsymbol{\theta})$ to be a constant. Essentially we let $f(\boldsymbol{\theta}) \propto 1$ implying $f(\boldsymbol{\theta}|\boldsymbol{y}) \propto f(\boldsymbol{y}|\boldsymbol{\theta})$. This solution cannot be applied consistently. If we are completely ignorant about $\boldsymbol{\theta}$ then it is plausible to assume we are also ignorant about any function of $\boldsymbol{\theta}$. In general a uniform prior for $\boldsymbol{\theta}$ translates to a non-uniform prior for any function of $\boldsymbol{\theta}$. Jeffreys (1967) proposed a solution to this problem, but this has been criticised because features of the data which are not encoded on the likelihood can impact the posterior, hence violating the so called likelihood principle

Furthermore, there may be no proper prior distribution with $f(\boldsymbol{\theta}) \propto 1$. This usually leads to a proper posterior if sufficient data are available.

In general we have to be very careful when using non-informative prior distributions.

2.1.4 Improper Prior Distributions and Lindley's Paradox

Lindley's paradox is best illustrated through an example. We consider the example presented in O'Hagan and Forster (2004). Let data $y \sim N(\mu, \sigma^2)$. Suppose there is some prior probability p that $\mu = 0$. If $\mu \neq 0$ then there is little prior information for how close to zero μ might be. We represent this information by saying that if $\mu \neq 0$ then $\mu \sim N(0, \omega^2)$. The prior distribution for μ is therefore mixed. We now consider the posterior probability that $\mu = 0$. Using Bayes theorem is easily seen that:

$$P(\mu = 0|y) = \frac{pf(y|\mu = 0)}{pf(y|\mu = 0) + (1-p)f(y|\mu \neq 0)} = \frac{pf(y|\mu = 0)}{f(y)}.$$

Now

$$\begin{aligned}
f(y|\mu \neq 0) &= \int f(\mu|\mu \neq 0) f(y|\mu) d\mu \\
&= \int (2\pi\sigma^2)^{-1/2}(2\pi\omega^2)^{-1/2} \exp(-\{\mu^2/(2\omega^2)\} - \{(y-\mu)^2/(2\sigma^2)\}) d\mu \\
&= (2\pi(\sigma^2+\omega^2))^{-1/2} \exp(\frac{-y^2}{2(\sigma^2+\omega^2)}).
\end{aligned}$$

Therefore $f(y|\mu \neq 0) \to 0$ as $\omega \to \infty$. Hence $f(y) \to pf(y|\mu = 0)$ and $P(\mu = 0|y) \to 1$ regardless of the observed data y. This result is known as Lindley's paradox and is not specific to this particular example. The same issue will arise when, in a given model, the prior distribution for the model parameters is improper over any part of the parameter space. Further examples of this paradox are given in this thesis.

2.1.5 Summary

There has been considerable criticism regarding the subjective treatment of probability required for a complete and full Bayesian analysis. In spite of this, interest in Bayesian inference has grown considerably. As we have seen Bayesian inference makes full use of all data and prior information where available. The main critique of the approach is the specification of a prior distribution. On one hand, it is possible to create a prior distribution that overwhelms any data. This follows directly from (2.3), since if $f(\boldsymbol{\theta}_0) = 1$ for some $\boldsymbol{\theta}_0$ and zero otherwise, then $f(\boldsymbol{\theta}|\boldsymbol{y}) = 1$ for $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ regardless of any observed data. On the other hand, if an improper prior distribution is assumed for $\boldsymbol{\theta}$ then inference is based solely upon the data. In practice we should therefore treat prior information as approximate and determine how sensitive posterior inferences are to realistic changes in these prior judgements.

Comparisons between the various methods of inference have, and will, be frequently drawn. These comparisons generally compare *sampling* properties under Bayesian and classical inference. For example the *Likelihood Principle* states that inference should be based only upon the likelihood. Consider possible observations denoted $\boldsymbol{x}$ and $\boldsymbol{y}$ arising from two different experiments $\boldsymbol{X}$ and $\boldsymbol{Y}$. If $f(\boldsymbol{x}|\boldsymbol{\theta}) \propto f(\boldsymbol{y}|\boldsymbol{\theta})$, then the likelihood principle states that we should make the same inference having observed $\boldsymbol{x}$ as we would having observed $\boldsymbol{y}$. Clearly Bayesian inference satisfies the likelihood principle. This cannot be said of classical inference.

For a detailed discussion see O'Hagan and Forster (2004). For an historical and theoretical grounding in statistical inference that considers Bayesian, fiducial, likelihood, and frequentist approaches the reader is referred to Welsh (1996).

2.2 Bayesian Model Determination

2.2.1 Posterior Model Probabilities and the Bayes Factor

Suppose that data $\boldsymbol{y}$ is believed to have been generated by model m from a set M of plausible models. Each model specifies completely the distribution of $\boldsymbol{Y}$, namely $f(\boldsymbol{y}|m, \boldsymbol{\theta}_m)$, where $\boldsymbol{\theta}_m$ is an unknown vector assumed to be in some parameter space Θ_m. We assume that $\Theta_m \subset \mathbb{R}^{p_m}$, where p_m is the dimension of $\boldsymbol{\theta}_m$. Under a Bayesian approach we are interested in the joint uncertainty of

$$(m, \boldsymbol{\theta}_m) \in \Theta = \bigcup_{m \in M} (\{m\} \times \Theta_m), \tag{2.5}$$

in light of the observed data $\boldsymbol{y}$. This uncertainty is captured by the posterior distribution $f(m, \boldsymbol{\theta}_m|\boldsymbol{y})$, and is given by Bayes theorem as follows

$$f(m, \boldsymbol{\theta}_m|\boldsymbol{y}) \propto f(\boldsymbol{y}|m, \boldsymbol{\theta}_m) f(m, \boldsymbol{\theta}_m). \tag{2.6}$$

Here $f(m, \boldsymbol{\theta}_m)$ $(= f(\boldsymbol{\theta}_m|m)f(m))$ represents our belief in $(m, \boldsymbol{\theta}_m)$ prior to having observed $\boldsymbol{y}$. The model specific prior distribution of $\boldsymbol{\theta}_m$ (conditional on m) is denoted $f(\boldsymbol{\theta}_m|m)$, whilst $f(m)$ is the prior probability that model m generated the data. We are interested in the posterior model probabilities. These are given by

$$f(m|\boldsymbol{y}) = \frac{f(m)f(\boldsymbol{y}|m)}{\sum_{k\in M} f(k)f(\boldsymbol{y}|k)}, \tag{2.7}$$

where

$$f(\boldsymbol{y}|m) = \int_{\Theta_m} f(\boldsymbol{y}, \boldsymbol{\theta}_m|m)d\boldsymbol{\theta}_m = \int_{\Theta_m} f(\boldsymbol{y}|m, \boldsymbol{\theta}_m)f(\boldsymbol{\theta}_m|m)d\boldsymbol{\theta}_m \tag{2.8}$$

is the marginal likelihood of $\boldsymbol{y}$ given model m. To compare model m to model m' we calculate the *Bayes factor* in favour of model m denoted

$$\mathcal{B}_{m,m'} = \frac{f(\boldsymbol{y}|m)}{f(\boldsymbol{y}|m')}. \tag{2.9}$$

If we note

$$\frac{f(m|\boldsymbol{y})}{f(m'|\boldsymbol{y})} = \frac{f(\boldsymbol{y}|m)}{f(\boldsymbol{y}|m')}\frac{f(m)}{f(m')},$$

it is easily seen that the Bayes factor is the ratio of posterior to prior odds.

2.2.2 Prediction

An important goal in the inference process is prediction. Given observed data $\boldsymbol{y}$ we wish to predict future replicates of this data. We denote these replicates $\boldsymbol{y_f}$. A Bayesian approach to prediction would often involve averaging predictions over different models, requiring a method that accounts for model uncertainty. Let $f(\boldsymbol{y_f}|\boldsymbol{y})$ denote the *model average* distribution of the future prediction given the observed data $\boldsymbol{y}$. This is given by

$$\begin{aligned} f(\boldsymbol{y_f}|\boldsymbol{y}) &= \sum_{m\in M}\int_{\Theta_m} f(\boldsymbol{y_f}, m, \boldsymbol{\theta}_m|\boldsymbol{y})d\boldsymbol{\theta}_m \\ &= \sum_{m\in M}\int_{\Theta_m} f(\boldsymbol{y_f}|m, \boldsymbol{\theta}_m, \boldsymbol{y})f(\boldsymbol{\theta}_m|m, \boldsymbol{y})f(m|\boldsymbol{y})d\boldsymbol{\theta}_m. \end{aligned}$$

If we assume that future and past data are conditionally independent given $(m, \boldsymbol{\theta}_m)$ then

$$
\begin{aligned}
f(\boldsymbol{y_f}|\boldsymbol{y}) &= \sum_{m\in M}\int_{\Theta_m} f(\boldsymbol{y_f}|m,\boldsymbol{\theta}_m)f(\boldsymbol{\theta}_m|m,\boldsymbol{y})f(m|\boldsymbol{y})d\boldsymbol{\theta}_m \\
&= \sum_{m\in M} f(m|\boldsymbol{y})\int_{\Theta_m} f(\boldsymbol{y_f}|m,\boldsymbol{\theta}_m)f(\boldsymbol{\theta}_m|m,\boldsymbol{y})d\boldsymbol{\theta}_m, \\
&= \sum_{m\in M} f(m|\boldsymbol{y})f(\boldsymbol{y_f}|m,\boldsymbol{y}), \qquad (2.10)
\end{aligned}
$$

Where $f(m|\boldsymbol{y})$ is given by (2.7). Averaging over all models in this fashion provides better predictive ability (measured by the logarithmic scoring rule) than any single model; i.e.

$$\mathbb{E}\Bigg[\log\sum_{m\in M} f(m|\boldsymbol{y})f(\boldsymbol{y_f}|m,\boldsymbol{y})\Bigg] \geq \mathbb{E}\Bigg[\log f(\boldsymbol{y_f}|k,\boldsymbol{y})\Bigg] \quad \forall k\in M,$$

where the expectation is with respect to the posterior predictive distribution. This is true by the non-negativity of the Kullback-Leibler distance (Jensen's inequality).

2.2.3 The Need for Markov Chain Monte Carlo

Posterior quantities of interest and predictive densities ((2.7), (2.9) and (2.10) respectively) require the evaluation of (2.8) for each model in our class of models M. This marginal likelihood is tractable in certain restricted situations only. Even if the marginal likelihood were tractable the size of most interesting model classes renders the exhaustive summation of (2.7) impractical. Thus, the resulting joint posterior distributions (2.6) cannot, in general, be calculated analytically.

Various *ad-hoc* methods to approximate posterior model probabilities have been proposed. For example, if we assume each model to be *a-priori* equally likely then the posterior density is proportional to

$$f(\boldsymbol{y}|m,\boldsymbol{\theta}_m)f(\boldsymbol{\theta}_m|m).$$

If this density is highly peaked about its posterior mode, $\hat{\boldsymbol{\theta}}_m$, we can expand $\log(f(\boldsymbol{y}|m,\boldsymbol{\theta}_m))$ as a quadratic about $\hat{\boldsymbol{\theta}}_m$ as follows

$$\log(f(\boldsymbol{y}|m,\boldsymbol{\theta}_m)) = \log(L_m) - (n/2)(\boldsymbol{\theta}_m - \bar{\boldsymbol{\theta}}_m)\bar{\Sigma}^{-1}(\boldsymbol{\theta}_m - \bar{\boldsymbol{\theta}}_m) + R.$$

Where $L_m = f(\boldsymbol{y}|m,\bar{\boldsymbol{\theta}}_m)$, $\bar{\Sigma}^{-1}$ is the observed information matrix and $\bar{\boldsymbol{\theta}}_m$ is the maximum likelihood estimate of $\boldsymbol{\theta}_m$ under model m. R is the remainder term involving third order and higher derivatives. Ignoring this remainder we exponentiate and integrate over the uncertainty of $\boldsymbol{\theta}$ yielding the Laplace approximation

$$\int_{\Theta_m} f(\boldsymbol{y}|m,\boldsymbol{\theta}_m)f(\boldsymbol{\theta}_m|m)d\boldsymbol{\theta}_m \approx n^{-p/2}(2\pi)^{p/2}|\bar{\Sigma}|^{1/2}L_m f(\bar{\boldsymbol{\theta}}_m|m). \tag{2.11}$$

As an aside, note another example of Lindley's paradox occurs when comparing two nested models (m nested within m') using the Laplace approximation. Since $f(\bar{\boldsymbol{\theta}}_m|m)$ appears in (2.11) the ratio

$$\frac{n^{-p/2}(2\pi)^{p/2}|\bar{\Sigma}|^{1/2}L_m f(\bar{\boldsymbol{\theta}}_m|m)}{n^{-p'/2}(2\pi)^{p'/2}|\bar{\Sigma}|^{1/2}L_{m'} f(\bar{\boldsymbol{\theta}}'_m|m')}$$

can be made arbitrarily large by choosing a suitable diffuse prior for $(\boldsymbol{\theta}'_m|m')$.
Clearly

$$-2\log\left(\frac{f(\boldsymbol{y}|m)}{f(\boldsymbol{y}|m')}\right) \approx -2\log\frac{L_m}{L_{m'}} + (p_m - p_{m'})\log(n), \tag{2.12}$$

which is known as the Bayes information criterion (BIC). Clearly this is an adjustment of the classical likelihood ratio to favour more strongly the model with fewer parameters. The Schwartz criterion is given by

$$\log\frac{L_m}{L_{m'}} - (1/2)(p_m - p_{m'})\log(n)$$

and can be thought of as approximation to the log of the Bayes factor.

The approximation (2.11) must be made for all models in the set M. If this set is large then the exhaustive summation in (2.7) will be computationally expensive. To overcome this problem, Madigan and Raftery (1994) proposed averaging over a much smaller set of models using the principle of Occam's razor. Models are excluded from this set if they receive less support from the data than any simpler nested model. Madigan and Raftery (1994) provide a search algorithm for the construction of this set. This approach is again an *ad-hoc* solution to the problem in hand.

More recently Markov chains have been used to generate a sample from the posterior distribution $f(\boldsymbol{\theta}_m, m|\boldsymbol{y})$, and Monte Carlo samples generated by the Markov chain then used for posterior inference about the uncertainty of $(\boldsymbol{\theta}_m, m)$. In the following section we introduce a method based on the reversible jump Markov chain Monte Carlo approach of Green (1995) for exploring this posterior model space. Alternative Markov chain based methods for exploring model uncertainty exist and have been discussed in the statistical literature. For example Carlin and Chib (1995) introduced a method, based on the Gibbs sampler, to generate from the posterior distribution $f(\boldsymbol{\theta}_m, m|\boldsymbol{y})$. Although computationally expensive, Dellaportas et al. (2002) showed a 'Metropolised' inexpensive version of the scheme was in fact a special case of the reversible jump algorithm. Other Markov chain based methods for exploring posterior model uncertainty can be found in Swendsen and Wang (1987), George and McCulloch (1993), Raftery et al. (1997), Damien et al. (1999), Nott and Green (2004) and Nott and Leonte (2004). For a review on Bayesian model selection using MCMC the reader is referred to Dellaportas et al. (2002).

2.3 Markov Chain Monte Carlo Methods

The importance of statistical models has long been recognised in many disciplines of science. These methods have contributed to a greater understanding of scientific problems which, in turn, has spurred research into new and improved statistical models. For many scientists the statistical conclusion is of greater importance than the statistical tools that helped them reach this end. There is therefore a great need for powerful and flexible inferential techniques that can easily be used by scientists. Throughout this section we denote by π the statistical distribution which is of interest to the scientist. The distribution π may well be the posterior distribution $f(\boldsymbol{\theta}|\boldsymbol{y})$ given in the previous section but it is not limited to solely this case. Therefore π denotes any target distribution of interest.

Many problems resulting from the application of statistical models can be formed in terms of an integral containing the distribution π, where π is defined on a general state space which we denote $\mathcal{X}$. Elements of $\mathcal{X}$ are denoted $\boldsymbol{x}$. For example, the expectation of some real valued function h is given by

$$\mathbb{E}[h(\boldsymbol{x})] = \int_{\mathcal{X}} h(\boldsymbol{x})\pi(d\boldsymbol{x}).$$

Further examples were provided in the previous section. These examples included the calculation of the constant of proportionality and the integration of a joint posterior distribution to obtain a marginal distribution. It is often the case that the explicit calculation of such integrals is not possible. Markov chain Monte Carlo, or MCMC, is a collection of computer intensive algorithms that permit the approximate computation of integrals that are analytically intractable. In this section we present some of the most common algorithms and discuss recent developments.

We adopt a measure theoretic notation. This notation is essential to establish the validity of the methods presented. However, less formal explanations, discussion and practical examples of the method will be presented in later chapters. In following the notation of Green (1995) and Green (2003) one can cast the problem of MCMC outside of the Bayesian paradigm and provide a general description of MCMC.

Throughout, we have assumed the reader is familiar with the notions of irreducibility, aperiodicity, Harris recurrence, time homogeneity and invariance distributions. Further discussion of these issues will be provided when necessary. The reader is referred to Gamerman (1997) and Norris (1997) for reasonable introductions to

MCMC and Markov chains respectively.

The challenge behind MCMC is to construct an irreducible, aperiodic, Harris recurrent, time homogenous Markov chain with one step transition kernel P that has π, the distribution of interest, as its invariant distribution. The observations or iterates of this Markov chain are denoted $\boldsymbol{x}^1, \boldsymbol{x}^2, ...$ throughout this section. For any well behaved function h the ergodic theorem states that

$$\lim_{m \to \infty} \frac{1}{n} \sum_{i=1}^{n} h(\boldsymbol{x}^i) = \mathbb{E}_\pi[h(\boldsymbol{x})]. \tag{2.13}$$

Essentially, if we could construct a Markov Chain which has our target distribution as its invariant distribution, then integrals listed above can be approximated using Monte Carlo averages based on realisations of the Markov Chain. Clearly if we could sample directly from π then there is no need to construct the Markov Chain. The Monte Carlo approach still applies.

The theory behind the most popular MCMC algorithms was developed in the middle half of the twentieth century (Metropolis et al. (1953), Hastings (1970) and Peskun (1973)). However it has not be until the last few years that the potential of the method has been realised. This can certainly be attributed to the ubiquity of fast computing machines. The increase in availability of computers and the theoretical developments of the algorithms has enabled many statistical researchers to work within the Bayes paradigm. The growth of theoretical and applied research of MCMC techniques has mimicked the growth in popularity of Bayesian statistics. For this reason the application of MCMC is almost always linked with Bayesian statistics, but this is not its sole use. In fact, the practical application of MCMC spans most areas of scientific research, and indeed its roots are not found in the Bayesian field.

In the following sections we discuss the two most popular algorithms. These are the Gibbs sampler and the well known Metropolis-Hastings algorithm. The Metropolis-Hastings algorithm is introduced in a general setting permitting the discussion of the reversible jump algorithm, an exciting technique used throughout the thesis.
The presentation of the Gibbs sampler is brief. Although it is an important inferential tool, methodological developments presented within this thesis refer to the Metropolis-Hastings algorithm. Accordingly, references for further reading are provided.

2.3.1 The Gibbs Sampler

The Gibbs sampler originated in the field of image analysis, where researchers were required to sample from a Gibbs distribution. It assumes that π is n-dimensional distribution of interest and that is possible to sample from the full conditionals of π. Suppose $\boldsymbol{x} = (x_1, ..., x_n) \in \mathbb{R}^n$ and that for each i it is possible to sample from the full conditional $\pi(x_i|\boldsymbol{x}_{(i)})$. Here $\boldsymbol{x}_{(i)}$ denotes the vector $\boldsymbol{x}$ with the i'th element removed. The (systematic sweep) Gibbs sampler then moves from $\boldsymbol{x}^t$ to $\boldsymbol{x}^{t+1}$ updating each x_i in turn by sampling from the conditional distribution $\pi(.|x_1^{t+1}, ..., x_{i-1}^{t+1}, x_{i+1}^t, ..., x_n^t)$. Although each component update is reversible as a whole, due to the nature of the systematic sweep, the Gibbs sampler is not reversible (A discussion of reversible Markov chains is provided in the following section).

With little additional work it is possible to construct a reversible Gibbs sampler. The sampler is known is the random scan. The (random scan) Gibbs sampler moves from $\boldsymbol{x}^t$ to $\boldsymbol{x}^{t+1}$ by randomly sampling $i' \in \{1, ..., n\}$ and sampling from the full conditional distribution $\pi(.|x_1^t, ..., x_{i'-1}^t, x_{i'+1}^t, ..., x_n^t)$.

Further details of the Gibbs sampler can be found in Gelfand and Smith (1990) and Gamerman (1997), including the proof that π is indeed the stationary distribution.

2.3.2 The Reversible Jump and Metropolis-Hastings Algorithms

In this section we introduce the reversible jump algorithm and the Metropolis-Hastings algorithm. We follow closely the work presented in Green (1995) and Green (2003)

Consider a general state space $\mathcal{X}$ and suppose we are interested in some distribution π that is defined on $\mathcal{X}$. We are interested in constructing a Markov chain with one step transition P that has π as its invariant distribution. For π to be the invariant distribution of the Markov chain the following equation must hold

$$\int_{\mathcal{X}} \pi(d\boldsymbol{x})P(\boldsymbol{x}, d\boldsymbol{x}') = \pi(d\boldsymbol{x}'). \tag{2.14}$$

We make the further requirement that the resulting Markov chain is reversible. In particular we require that for all Borel sets $B, B' \subset \mathcal{X}$,

$$\int_{(\boldsymbol{x},\boldsymbol{x}')\in B\times B'} \pi(d\boldsymbol{x})P(\boldsymbol{x}, d\boldsymbol{x}') = \int_{(\boldsymbol{x},\boldsymbol{x}')\in B\times B'} \pi(d\boldsymbol{x}')P(\boldsymbol{x}', d\boldsymbol{x}). \tag{2.15}$$

The above equation is known as the integrated detailed balance equation. The Metropolis-Hastings algorithm and the reversible jump algorithm proceed by proposing a new state $\boldsymbol{x}'$ from a proposal measure $q(\boldsymbol{x}, d\boldsymbol{x}')$ and accepting the new state with probability $\alpha(\boldsymbol{x}, \boldsymbol{x}')$. Otherwise the old state $\boldsymbol{x}$ becomes the new state.

If we directly consider the transition kernel $P(\boldsymbol{x}, B')$ we note that

$$P(\boldsymbol{x}, B') = \int_{B'} q(\boldsymbol{x}, d\boldsymbol{x}')\alpha(\boldsymbol{x}, \boldsymbol{x}') + R(\boldsymbol{x})1_{\{\boldsymbol{x}\in B'\}}, \tag{2.16}$$

where $1_{\{.\}}$ is the indicator function and $R(\boldsymbol{x})$ is the probability the proposed move is rejected. This probability is given by

$$R(\boldsymbol{x}) = \int_{\mathcal{X}} q(\boldsymbol{x}, d\boldsymbol{x}')(1 - \alpha(\boldsymbol{x}, \boldsymbol{x}')). \tag{2.17}$$

If we now substitute (2.16) into (2.15) yields

$$\int_B \pi(d\boldsymbol{x}) \int_{B'} q(\boldsymbol{x}, d\boldsymbol{x}')\alpha(\boldsymbol{x}, \boldsymbol{x}') + \int_{B \cap B'} \pi(d\boldsymbol{x})R(\boldsymbol{x}) \\ = \int_{B'} \pi(d\boldsymbol{x}') \int_B q(\boldsymbol{x}', d\boldsymbol{x})\alpha(\boldsymbol{x}', \boldsymbol{x}) + \int_{B' \cap B} \pi(d\boldsymbol{x}')R(\boldsymbol{x}'). \quad (2.18)$$

The last terms on either side cancel. Hence (2.18) can be written as

$$\int_{(\boldsymbol{x},\boldsymbol{x}') \in B \times B'} \pi(d\boldsymbol{x})q(\boldsymbol{x}, d\boldsymbol{x}')\alpha(\boldsymbol{x}, \boldsymbol{x}') = \int_{(\boldsymbol{x},\boldsymbol{x}') \in B \times B'} \pi(d\boldsymbol{x}')q(\boldsymbol{x}', d\boldsymbol{x})\alpha(\boldsymbol{x}', \boldsymbol{x}). \quad (2.19)$$

This equation is given in Green (2003). Green (1995) considers the case where the transition kernel is a mixture over a number of different move types. When the current state is $\boldsymbol{x}$, a move of type j that would take the state to $d\boldsymbol{x}'$ is proposed with probability $q_j(\boldsymbol{x}, d\boldsymbol{x}')$. Greens formulation allows the possibility of no move being attempted. He also notes that not all moves j will be available from all starting states $\boldsymbol{x}$.

Green shows that a sufficient condition for the reversibility of such an algorithm is that given in (2.19) but for each $q_j(\boldsymbol{x}, d\boldsymbol{x}')$. That is, Green restricts his attention to Markov chains in which detailed balance is attained within each move type.

Green then assumes the existence of a symmetric measure μ on $\mathcal{X} \times \mathcal{X}$ which dominates $\pi(d\boldsymbol{x})q(\boldsymbol{x}, d\boldsymbol{x}')$. Under this assumption $\pi(d\boldsymbol{x})q(\boldsymbol{x}, d\boldsymbol{x}')$ has density denoted $f(\boldsymbol{x}, \boldsymbol{x}')$ with respect to μ. In fact $f(\boldsymbol{x}, \boldsymbol{x}')$ is known as the Radon-Nikodym derivative. This means that (2.19) can now be written as

$$\int_{(\boldsymbol{x},\boldsymbol{x}') \in B \times B'} \alpha(\boldsymbol{x}, \boldsymbol{x}')f(\boldsymbol{x}, \boldsymbol{x}')\mu(d\boldsymbol{x}, d\boldsymbol{x}') = \int_{(\boldsymbol{x},\boldsymbol{x}') \in B \times B'} \alpha(\boldsymbol{x}', \boldsymbol{x})f(\boldsymbol{x}', \boldsymbol{x})\mu(d\boldsymbol{x}', d\boldsymbol{x}). \quad (2.20)$$

Noting that (2.20) holds provided that

$$\alpha(\boldsymbol{x},\boldsymbol{x}')f(\boldsymbol{x},\boldsymbol{x}') = \alpha(\boldsymbol{x}',\boldsymbol{x})f(\boldsymbol{x}',\boldsymbol{x}), \tag{2.21}$$

which is satisfied for all Borel sets B and B' if

$$\alpha(\boldsymbol{x},\boldsymbol{x}') = \frac{f(\boldsymbol{x}',\boldsymbol{x})}{f(\boldsymbol{x},\boldsymbol{x}')} \wedge 1 \tag{2.22}$$

since

$$\begin{aligned}
\alpha(\boldsymbol{x},\boldsymbol{x}')f(\boldsymbol{x},\boldsymbol{x}') &= \left\{\frac{f(\boldsymbol{x}',\boldsymbol{x})}{f(\boldsymbol{x},\boldsymbol{x}')} \wedge 1\right\} f(\boldsymbol{x},\boldsymbol{x}') \\
&= f(\boldsymbol{x}',\boldsymbol{x}) \wedge f(\boldsymbol{x},\boldsymbol{x}') \\
&= f(\boldsymbol{x}',\boldsymbol{x}) \wedge \frac{f(\boldsymbol{x},\boldsymbol{x}')}{f(\boldsymbol{x}',\boldsymbol{x})} f(\boldsymbol{x}',\boldsymbol{x}) \\
&= \left\{1 \wedge \frac{f(\boldsymbol{x},\boldsymbol{x}')}{f(\boldsymbol{x}',\boldsymbol{x})}\right\} f(\boldsymbol{x}',\boldsymbol{x}) \\
&= \alpha(\boldsymbol{x}',\boldsymbol{x})f(\boldsymbol{x}',\boldsymbol{x}) \qquad (2.23)
\end{aligned}$$

As Green notes, if we express this ratio less formally then the expression can be written as the ratio of measures given by

$$\alpha(\boldsymbol{x},\boldsymbol{x}') = \frac{\pi(d\boldsymbol{x}')q(\boldsymbol{x}',d\boldsymbol{x})}{\pi(d\boldsymbol{x})q(\boldsymbol{x},d\boldsymbol{x}')} \wedge 1 \tag{2.24}$$

Almost all authors note the abstract nature of the above formulation. However, the formulation presented above encompasses both the Metropolis-Hastings algorithm and the reversible jump algorithm. Indeed, a possible move type considered by Green is simply a 'standard' Metropolis-Hastings update.

We follow closely the arguments presented by Green (2003) and provide a constructive representation in terms of random number generation. Initially we assume that $\mathcal{X} \subset \mathbb{R}^p$ and that π, our distribution of interest, has a density with respect to the

d dimensional Lebesgue measure. In an abuse of notation this density will also be denoted by π.

To propose a move from $\boldsymbol{x} \in \mathbb{R}^d$ to $\boldsymbol{x}' \in \mathbb{R}^d$ we firstly generate r random numbers denoted by $\boldsymbol{u}$ from a distribution with known density denoted by g. The proposed state is now given by $\boldsymbol{x}'$, where $\boldsymbol{x}' = \boldsymbol{t}(\boldsymbol{x}, \boldsymbol{u})$ where $\boldsymbol{t}$ is some known deterministic function. The left hand side of (2.20) can now be written as an integral with respect to $(\boldsymbol{x}, \boldsymbol{u})$ as

$$\int_{(\boldsymbol{x},\boldsymbol{x}')\in B\times B'} \pi(\boldsymbol{x})g(\boldsymbol{u})\alpha(\boldsymbol{x}, \boldsymbol{x}')d\boldsymbol{x}d\boldsymbol{u}.$$

The reverse move is made in a similar fashion so that $\boldsymbol{x} = \boldsymbol{t}'(\boldsymbol{x}', \boldsymbol{u}')$, where $\boldsymbol{u}'$ are r' random numbers generated from a distribution with density g'. The right hand side of (2.20) can now be written as an integral with respect to $(\boldsymbol{x}', \boldsymbol{u}')$ as

$$\int_{(\boldsymbol{x},\boldsymbol{x}')\in B\times B'} \pi(\boldsymbol{x}')g'(\boldsymbol{u}')\alpha(\boldsymbol{x}', \boldsymbol{x})d\boldsymbol{x}'d\boldsymbol{u}'.$$

If the transformation from $(\boldsymbol{x}, \boldsymbol{u})$ to $(\boldsymbol{x}', \boldsymbol{u}')$ is invertible and differentiable (implying that $r = r'$) then (2.20) holds if

$$\pi(\boldsymbol{x})g(\boldsymbol{u})\alpha(\boldsymbol{x}, \boldsymbol{x}') = \pi(\boldsymbol{x}')g'(\boldsymbol{u}')\alpha(\boldsymbol{x}', \boldsymbol{x}) \left| \frac{d(\boldsymbol{x}', \boldsymbol{u}')}{d(\boldsymbol{x}, \boldsymbol{u})} \right|, \tag{2.25}$$

where the last term on the right hand side is the Jacobian of the transformation from $(\boldsymbol{x}, \boldsymbol{u})$ to $(\boldsymbol{x}', \boldsymbol{u}')$. Therefore a suitable and optimal choice for $\alpha(\boldsymbol{x}, \boldsymbol{x}')$ is given by

$$\alpha(\boldsymbol{x}, \boldsymbol{x}') = \frac{\pi(\boldsymbol{x}')g'(\boldsymbol{u}')}{\pi(\boldsymbol{x})g(\boldsymbol{u})} \left| \frac{d(\boldsymbol{x}', \boldsymbol{u}')}{d(\boldsymbol{x}, \boldsymbol{u})} \right| \tag{2.26}$$

This is the well known Metropolis-Hastings algorithm. The Jacobian arises from the specification of $\boldsymbol{x}'$ and $\boldsymbol{u}'$ in terms of $\boldsymbol{x}$ and $\boldsymbol{u}$. It is often the case that the proposal has been designed in such a way that the Jacobian factor is equal to 1.

Suppose now that we assume our state space $\mathcal{X}$ is no longer a subset of $\mathbb{R}^d$ but in fact a countable union of spaces with possibly different dimensions

$$\mathcal{X} = \bigcup \mathcal{X}_k.$$

Clearly π, the distribution of interest, no longer has a density with respect to the d dimensional Lebesgue measure. However we assume that for each k, π has a density over $\mathcal{X}_k$ with respect to a d_k dimensional Lebesgue measure. We denote this density by $\pi(\boldsymbol{x})$ whatever the dimension of $\boldsymbol{x} \in \mathcal{X}$. The reversible jump algorithm then proceed as follows.

To propose a move from $\boldsymbol{x} \in \mathcal{X}_k$ to $\boldsymbol{x}' \in \mathcal{X}_{k'}$ we generate r random numbers denoted by $\boldsymbol{u}$, from a distribution with known density denoted by g. The proposed state is now given by $\boldsymbol{x}'$, where $\boldsymbol{x}' = \boldsymbol{t}(\boldsymbol{x}, \boldsymbol{u})$ and where $\boldsymbol{t} : \mathbb{R}^d \times \mathbb{R}^r \rightarrow \mathbb{R}^{d'}$ is some known deterministic function. The reverse move is constructed in a similar fashion. We generate r' random numbers denoted by $\boldsymbol{u}'$ from a distribution with known density denoted by g'. The proposed state is now given by $\boldsymbol{x}$, where $\boldsymbol{x} = \boldsymbol{t}'(\boldsymbol{x}', \boldsymbol{u}')$ and $\boldsymbol{t}' : \mathbb{R}^{d'} \times \mathbb{R}^{r'} \rightarrow \mathbb{R}^d$. Provided that the transformation from $(\boldsymbol{x}, \boldsymbol{u})$ to $(\boldsymbol{x}', \boldsymbol{u}')$ remains differentiable and invertible then the acceptance probability given in (2.26) still satisfies (2.20). For this transformation and its inverse to remain differentiable we need the following relationship to hold

$$d + r = d' + r'.$$

This equation is often referred to as the dimension matching constraints. Note the Metropolis-Hastings algorithm satisfies these constraints if and only if $r = r'$. It is worth noting that not only is this acceptance optimal, in the sense of minimising the autocorrelation of the Markov chain, it is also such that we need only know the density $\pi(\boldsymbol{x})$ up to a constant of proportionality, since π appears in both the numerator and denominator of (2.26).

Throughout this section we have assumed that the integrated detailed balance equations are satisfied. The resulting Markov chain is therefore reversible. This assumption is not essential and work presented by Neal (2004) advocated the use of non reversible chains. We have also assumed the chain to be irreducible and aperiodic. These two conditions need to be verified in practical situations.

In Chapter 3 we introduce the reversible jump algorithm without the direct consideration of measure. The method is demonstrated in the context of a general problem requiring jumps between models and is similar to the approaches taken by Dellaportas et al. (2002) or Ehlers and Brooks (2002).
In the following section we discuss recent methodological developments of the reversible jump algorithm. We also provide a brief discussion of the problems to which reversible jump can be applied.

2.3.3 Methodological Developments for Reversible Jump Algorithm

The statistical problems to which reversible jump can be applied are numerous. Most applications involve the problem of Bayesian model choice or selection, where reversible jump provides a method of making combined inference about parameters (m, θ_m) (π then denotes the posterior model probability given by (2.7)). Examples of this application include Brooks et al. (2003), Dellaportas and Forster (1999) and Dellaportas et al. (2002). We provide a further example of the model jumping application in Chapter 3.
Other applications include Bayesian analysis of a Poisson process with change points Green (1995), image analysis Al-Awadhi et al. (2004) and mixture models Richardson and Green (1997).

Regardless of the application, almost all researchers have noted the difficulties of applying the reversible jump to practical situations. The difficulties incurred are almost always due to the construction of suitable proposal distributions for $\boldsymbol{u}$ and transformation functions $\boldsymbol{t}$. The choice of these two components greatly effect the efficiency of the resulting chain. Inefficient proposals can lead to low acceptance probabilities and therefore poor mixing of the resulting Markov chain. This poor mixing in turn leads to slow convergence and therefore a greater numbers of iterates must be generated in order to make reasonable inference concerning the parameters of interest. Even with the ubiquity of fast computing machines this can be a burden.

Standard methods of creating an efficient algorithm concern tuning. Here short runs (pilot runs) of the Markov chain are performed, each time some aspect of the proposal distribution is altered. For example one might alter the centring or scaling of the proposal distribution. One might consider a variety of blocking updates or reparameterisations. A final chain is then run using the information gathered from these pilot runs.

This approach should be effective for simple problems. However, as the problems become more difficult the approach can be computationally expensive as a vast number of pilot runs have to be performed. For example, the case of the reversible jump algorithm applied to the problem of Bayesian model determination might require pilot runs to be performed in order to obtain proposal distributions when considering a move between two models. Since the number of models under consideration could be large this might be infeasible.

A further difficulty encountered when applying the reversible jump algorithm is in determining the convergence of the Markov chain. If we wish to make inference concerning some distribution π the we need to be assured that the Markov chain is indeed at the distribution π. The difficulties in determining convergence arise because of the generality of the state space $\mathcal{X}$. As there is no concept of closeness it is not possible to extend the work of Roberts et al. (1997) or Roberts and Rosenthal (1998). A further reason why it is difficult to assess convergence is simply the vastness of the space we intend to explore and difficulties arise in assessing the convergence of the algorithm in parts of the space that are rarely explored. In practice, graphical techniques together with a variety of statistics are used to draw sensible conclusions. For example, we might include trace and autocorrelation plots together with standard errors of parameters estimates and acceptance probabilities of proposed moves.

Perhaps the most important recent development of the reversible jump algorithm is the work of Brooks et al. (2003). This work aims to provide methods for selecting good choices for the parameters of the proposal densities g. The work assumes a known fixed transformation function $\boldsymbol{t}$.

Brooks et al. (2003) introduce two classes of methods and we summarise each in turn.

The first class of methods are termed order methods. These methods proceed by analysing the acceptance probability of the proposed move which, following their notation, we denote by $A(\boldsymbol{x}, \boldsymbol{x}')$. The idea is that for the current state of the chain $\boldsymbol{x} = (m, \boldsymbol{\theta}_m)$ it is possible to specify a centring point $c(\boldsymbol{\theta}_m)$ in a proposed new model m'. The authors provide methods for choosing this centring point including an approach called conditional maximisation. This approach sets $c(\boldsymbol{\theta}_m) = \boldsymbol{\theta}^*_{m'}$, where $\boldsymbol{\theta}^*_{m'}$ is the value of $\boldsymbol{\theta}'_{m'}$ that maximises $\pi(m', \boldsymbol{\theta}'_{m'})$, conditional on the current parameter values $\boldsymbol{\theta}_m$.

At the chosen centring point the order methods impose constraints on $A(\boldsymbol{x}, \boldsymbol{x}')$ and its derivatives. Resulting equations are then solved to yield the parameters of g

For example the J'th ordering method imposes the conditions that

$$A(\boldsymbol{x}, \boldsymbol{x}')|_{\boldsymbol{x}'=(m', c(\boldsymbol{\theta}_m))} = 1$$

and that

$$\frac{\partial^j}{\partial \boldsymbol{u}^j} A(\boldsymbol{x}, \boldsymbol{x}')|_{\boldsymbol{x}'=(m', c(\boldsymbol{\theta}_m))} = 0,$$

for $j = 1, ..., J$. The idea is that the acceptance probability is close to 1 at the chosen centring point. The higher order methods scale the proposal to maintain a high acceptance rate for a range of values. The order methods are appealing and numerical results from the practical examples presented in Brooks et al. (2003) encouraging.

The second class of methods introduced by Brooks et al. (2003) is called the saturated state space approach. The idea here is to introduce auxiliary variables $\boldsymbol{u}$ to augment the state space, so that all models have the same dimension. The purpose of these auxiliary variables is to aid proposal design. The method uses deterministic proposal for the new state $\boldsymbol{\theta}'_{m'}$, the new model m' having been chose using a random kernel. These deterministic proposals are combined with within model updates of the variable $\boldsymbol{\theta}_m$ and auxiliary variables $\boldsymbol{u}_m$. The method seems promising as illustrated by the range of practical examples presented in Brooks et al. (2003).

A similar approach to the saturated state space approach is the product space approach. Here, auxiliary variables are introduced to augment the state space so that a fixed dimensional sample (the Metropolis-Hastings algorithm or Gibbs sampler) can be used. A particular attraction of this approach is that now the state space is of fixed dimension and the density π is now with respect to a d dimensional Lebesgue measure. It is therefore possible to consider the work of Roberts et al. (1997) or

Roberts and Rosenthal (1998) when assessing the convergence of the Markov chain. However, for each model m the method relies on the specification of a distribution, which they term a pseudo prior, for the variables that do not contribute directly to that particular model. The choice of pseudo priors has a critical effect on the efficiency of the resulting chain and we have therefore replaced the problem of selecting suitable proposals g with the problem of selecting suitable pseudo priors.

The saturated state space approach and product space approaches introduce auxiliary variables increasing the computational burden. This is particularly true of the product space approach if the number of variables and interactions under consideration is large.
The interested reader is referred to Carlin and Chib (1995), Godsill (2001) and Godsill (2003) for additional information.

Our discussion of recent methodological developments has not included recent work focusing on adaptive methods. The work in this thesis can be appreciated without direct consideration of such work. For the interested reader we therefore recommend Atchadé and Rosenthal (2005) and references therein.

To conclude this section we must note that although promising methodological developments have been made the choice of the deterministic transition function $\boldsymbol{t}$ remains a challenging problem. In Chapter 3 we provide a novel transformation function for the reversible jump algorithm applicable for model Bayesian model determination for generalised linear models.

2.4 Statistical Analysis with Missing Data

The problem of analysing data sets from which observations are missing is common. There may be many different reasons for these missing observations, for example measurement error or non-response, but the statistical goal remains the same. How should one approach inference that accommodates the possible, but unknown, behaviour of the unobserved data?

When analysing missing data the first assumption we make is that missing observations contain meaningful information for analysis. If this were not the case then any missing observations could be discarded and the analysis would proceed using fully observed data alone.

This assumption is made throughout this thesis.

2.4.1 Missing Data Examples

It is useful to consider the situations where data sets might contain some missing observations. Three such situations are listed below. By no means is this an exhaustive list, it merely contains situations discussed at a latter stage of this thesis.

2.4.1.1 Non-Response in Surveys

Consider what might happen when a statistical agency conducts a public opinion poll. The agency may ask individuals about how, and if, they will vote in a forthcoming election. Missing observations may arise here if a given individual does not want to reveal the party, or candidate, for whom they intend to vote. Perhaps the individual is yet to decide his, or her, voting intention. It is also a possibility that the individual is unsure whether he, or she, will vote. All of these possibilities might lead to missing observations. In Chapter 4 we consider methods for analysing electoral poll data.

2.4.1.2 Longitudinal Studies

Longitudinal studies collect information on cases, or individuals, over a period of time. Consider a clinical trial where patients of a given hospital are given a particular drug. The patient may 'drop out' of the trial before the end date. Possible reasons for this dropout are: the patient may have suffered an adverse effect of the drug; the patient might have relocated during the trial; with luck, the patient may have been cured of the disease.

2.4.1.3 Statistical Disclosure Control

Statistical disclosure control concerns safeguarding the confidentiality of the information, or data, a statistical agency may hold about individuals or businesses. If data is to be released a variety of statistical disclosure techniques may be applied to reduce the risk of disclosure. For example, categorical variables may be recoded to reduce the number of levels or counts omitted from released data. The missing data has been artificially created. This issue is discussed in Chapter 5.

2.4.2 A Review of Missing Data Methods

There are a variety of techniques for handling missing data, which are summarised by Little and Rubin (2002). The four main approaches are as follows:

1. *Complete case analysis*: The missing data is ignored and analysis proceeds using fully observed data only.

2. *Weighting methods*: Data are weighted in an attempt to modify for non-response as if it were part of the survey or sample design.

3. *Imputation methods*: Missing values are estimated, and the data are then analysed as if there were no missing observations.

4. *Model based methods*: A broad class of methods that model the data generating process alongside the missing data mechanism. This approach has several advantages. Firstly, unlike many other approaches, it is not an *ad-hoc* approach. Any assumptions underlying the model can be tested. Secondly, any estimate takes into account data incompleteness. Model based procedures form the main focus of work within this thesis.

Before discussing each of the above ideas, it is important to note the objective of our analysis. The aim is not optimal point prediction of a given estimand (a function of the population data) with respect to some loss function. The goal is to make valid statistical inference, fully accounting for uncertainty in light of missing observations. We must make full use of all available information contained in fully, or partially, observed observations.

With this in mind, a complete case analysis seems inappropriate. Potentially useful information is ignored without second thought, and nothing can be said with regard to the reason by which some observations are missing. The sole advantage of this method is the ease of its implementation. However, if a simple and quick implementation is required then weighting is a more attractive alternative to complete case analysis. Weighting methods can be effective in producing unbiased estimates of a given parameter. However, this focus on reducing bias (in comparison to a complete case analysis) can come at a cost of increased variance (Little and Rubin (2002) page 50). For this reason, weighting methods are often used when the sample size is large and bias is a more serious issue than variance. Furthermore, it is unclear how this approach might be applied in a Bayesian setting. In conclusion, weighting has never really been considered a practical solution to the missing data problem.

Imputation, and multiple imputation, methods have received considerable attention since the seminal work of Rubin (1978). The method is simple. Missing values are replaced by imputed values and analysis proceeds as if the data were fully observed.

For the case of multiple imputation the whole process is repeated several times. Results from the multiple imputes, if generated, are then pooled and analysed together. The imputations are essentially repeated random draws from the predictive distribution of the missing values under a particular model. If we were to impute missing values from several non-response models, then combined inferences under the models can be contrasted to assess the sensitivity of inference to the non-response models. The imputation step provides the practical advantage of 'complete' data analysis and hence the use of standard readily available software. This key fact was, in part, the vision behind the original idea, see for example Rubin (1996). We must note that this method is not model free, in the sense that the imputed values are based on the predictive distribution of a non-response model.

Modelling the missing data is by far the most sophisticated and complex approach. The idea has two real advantages as a pay-off for this complexity. Firstly, modeling may provide information about the missing-data mechanism. Secondly, we may be able to ascertain how assumptions about this missing data mechanism affect inference, or at least determine the sensitivity of inferences to these assumptions. We discuss this approach further in Chapter 4.

Chapter 3

Bayesian Model Determination for Generalised Linear Models

A Bayesian approach to prediction should involve averaging predictions over different models thus providing a method that accounts for model uncertainty. Markov chain Monte Carlo methods for exploring this uncertainty have received a great deal of recent attention. In this chapter we focus upon one such method, the reversible jump algorithm (Green, 1995). We consider the difficulties associated with the algorithm, and offer an efficient and novel construction for model determination in generalised linear models.

The structure of this chapter is as follows. We begin with an introduction to model uncertainty and generalised linear models (Section 3.1). We introduce the reversible jump Markov chain Monte Carlo algorithm, Section 3.2. A reversible jump Markov chain Monte Carlo scheme for generalised linear models is constructed in Section 3.3 and several examples follow, Sections 3.4-3.8. These examples include model and variable selection in log-linear and logistic regression models.

We consider a variety of methods for assessing convergence before concluding with recommendations for future work.

3.1 Introduction - Bayesian GLM's

Let $\boldsymbol{y} = (y_1, ..., y_n)^T$ be n independent observations each with density function of the form

$$f(y_i) = \exp(\frac{\omega_i}{\phi}\{y_i\theta_i - b(\theta_i) + c(y_i, \phi)\}), \tag{3.1}$$

for scalars θ_i, ϕ and weights ω_i. Assuming the dispersion parameter ϕ is known and fixed, (3.1) is the density function of a distribution belonging to the exponential family of distributions with parameter θ_i. Functions b and c determine the specific parametric family of distributions. Common distributions that have this functional form are the binomial, Poisson and normal distributions. It is easily shown that if y_i has density given by (3.1), then $\mathbb{E}[y_i] = \mu_i = b'(\theta_i)$, and $Var[y_i] = b''(\theta_i)\frac{\phi}{\omega_i}$. That is the mean μ_i of y_i is directly related to the parameter θ_i. More importantly, the variance is permitted to be a function of μ_i, the mean. For a more detailed examination of exponential families the reader is referred to McCullagh and Nelder (1989).

For the ith observation y_i we define a linear predictor $\eta_i = \boldsymbol{x}_i^T\boldsymbol{\beta}$, where $\boldsymbol{x}_i$ is the ith row of a matrix $\boldsymbol{X}$ containing data on explanatory variables for unit i, and $\boldsymbol{\beta}$ is a vector of parameter values. The matrix $\boldsymbol{X}$ is termed the design matrix. The dependence of μ_i on η_i is then described through a differentiable and monotonic function ϱ, called the link function. This dependence is given by

$$\varrho(\mu_i) = \eta_i = \boldsymbol{x}_i^T\boldsymbol{\beta},$$

or in matrix notation

$$\varrho(\boldsymbol{\mu}) = \boldsymbol{\eta} = \boldsymbol{X\beta}.$$

Of particular interest is the canonical link which equates θ_i to η_i.

A generalised linear model (GLM) can then be expressed as the pair $m = (\gamma, S)$. Here γ represents the explanatory variables and interactions present in the linear predictor which, together with parameter constraints, determine matrix $\boldsymbol{X}$.

S is a set of structural properties. These properties include the error structure associated with the model (i.e. the response distribution and variance functions of the related exponential family), and the link function. Throughout this chapter these structural properties are assumed known and fixed.

Suppose data $\boldsymbol{y}$ have been generated by model m from a set M of plausible models. We assume that all models $m \in M$ have identical structural properties S, but differ in composition of explanatory variables in the linear predictor. Each model specifies completely the conditional distribution of $\boldsymbol{y}$ given m and $\boldsymbol{\beta}_m$. Here $\boldsymbol{\beta}_m$ is an unknown vector assumed to be in some parameter space $\Theta_m \subset \mathbb{R}^{p_m}$. As stated in the literature review we are interested in the joint uncertainty of

$$(m, \boldsymbol{\beta}_m) \in \Theta = \bigcup_{m \in M} \left(\{m\} \times \Theta_m\right) \tag{3.2}$$

in light of the observed data $\boldsymbol{y}$. Specifically, we require the posterior model probabilities. These probabilities are calculated via Bayes theorem and given by

$$f(m|\boldsymbol{y}) = \frac{f(m)f(\boldsymbol{y}|m)}{\sum_{k \in M} f(k)f(\boldsymbol{y}|k)} \tag{3.3}$$

where $f(\boldsymbol{y}|m)$ is the marginal likelihood given in (2.8). This marginal likelihood is often intractable. It may be the case that $|M|$ is far too large for exhaustive computation. To overcome these two difficulties we intend to draw a sample, at least approximately, from the posterior distribution $f(m|\boldsymbol{y})$. This sample will then be used to make joint inference regarding parameters m and $\boldsymbol{\beta}_m$. We shall draw this sample using a reversible jump Markov chain. This algorithm was introduced in Chapter 2. In this chapter we describe an efficient construction of the reversible jump algorithm for generalised linear models. We compare this novel algorithm to alternative Markov chain based methods, and provide contrasts with alternative methods for model choice.

3.2 Reversible Jump Markov Chain Monte Carlo

The reversible jump algorithm is an adaptation of the well known Metropolis-Hastings algorithm to a more general state space, and much of recent work on Bayesian model selection using Markov chains has focused on this method. Since its introduction by Green (1995), reversible jump has often been viewed as difficult to understand. This is likely due to the theoretical introduction provided in Green (1995) which considered the algorithm in a measure theoretical framework. In an attempt to provide an easier introduction we present the algorithm in a simpler way. We consider the algorithm in the context of a general problem requiring moves (jumps) between models.

We construct a Markov chain on the state space

$$\bigcup_{m \in M} (\{m\} \times \Theta_m)$$

with stationary distribution $f(m, \boldsymbol{\beta}_m|\boldsymbol{y})$ in the following way.
In an abuse of notation let the parameter values of the Markov chain at time T be $(m, \boldsymbol{\beta}_m)$. The dimension of $\boldsymbol{\beta}_m$ is denoted by p_m. We proceed as follows

- Propose a new model m' with probability $j(m, m')$.
- Generate $\boldsymbol{u}$ from a specified proposal density $g(\boldsymbol{u}|\boldsymbol{\beta}_m, m, m')$.
- Set $(\boldsymbol{\beta}'_{m'}, \boldsymbol{u}') = \boldsymbol{t}_{m,m'}(\boldsymbol{\beta}_m, \boldsymbol{u})$ for some invertible and differentiable function $\boldsymbol{t}_{m,m'}$.
- Accept this proposed move with probability $\alpha_{m,m'}$ given by

$$\frac{f(m', \boldsymbol{\beta}'_{m'}|\boldsymbol{y})j(m', m)g'(\boldsymbol{u}'|\boldsymbol{\beta}_m, m', m)}{f(m, \boldsymbol{\beta}_m|\boldsymbol{y})j(m, m')g(\boldsymbol{u}|\boldsymbol{\beta}'_{m'}, m, m')} \left| \frac{\partial \boldsymbol{t}_{m,m'}(\boldsymbol{\beta}'_{m'}, \boldsymbol{u}')}{\partial(\boldsymbol{\beta}, \boldsymbol{u})} \right| \wedge 1, \qquad (3.4)$$

which can be written as

$$\frac{f(\boldsymbol{y}|m',\boldsymbol{\beta}'_{m'})f(\boldsymbol{\beta}'_{m'}|m')f(m')j(m',m)g'(\boldsymbol{u}'|\boldsymbol{\beta}_m,m',m)}{f(\boldsymbol{y}|m,\boldsymbol{\beta}_m)f(\boldsymbol{\beta}_m|m)f(m)j(m,m')g(\boldsymbol{u}|\boldsymbol{\beta}'_{m'},m,m')}\left|\frac{\partial \boldsymbol{t}_{m,m'}(\boldsymbol{\beta}'_{m'},\boldsymbol{u}')}{\partial(\boldsymbol{\beta},\boldsymbol{u})}\right| \wedge 1. \tag{3.5}$$

- Or reject and remain in the current state.

The subscripts of the transformation function $\boldsymbol{t}_{m,m'}$ specifically allow the transformation to depend upon the proposed move from m to m'. Since the transformation function must be invertible and differentiable it is clear that $\boldsymbol{t}_{m,m'} = \boldsymbol{t}^{-1}_{m',m}$. Furthermore, $p_m + p_u = p_{m'} + p_{u'}$ which ensures the dimension matching constraints are satisfied).

This is a general construction of the reversible jump algorithm which permits a wide variety of proposals. If $m = m'$ and $\boldsymbol{t}_{m,m'}$ is the identity function then the move is the standard Metropolis-Hastings step (If $\boldsymbol{t}$ is not the identity then we have a more general Metropolis-Hastings algorithm). Clearly reversible jump encompasses many other simpler algorithms.

The two key ingredients of the reversible jump algorithm are the proposal distributions g and g', and the transformation function $\boldsymbol{t}_{m,m'}$.

In the past proposal distributions for the reversible jump algorithm have been obtained through empirical tuning. Here a pilot chain is run to obtain suitable proposal distributions. This method can be computationally expensive and, for complex problems, difficult. This may result in low acceptance probabilities and a chain that mixes poorly. An alternative to this method is to use one step of the 'Iteratively Re-weighted Least Sum of Squares' algorithm (Nott and Green (2004)). The IRLSS algorithm is used to obtain maximum likelihood estimates of the parameters of a generalised linear model. Neither method uses the current state of the Markov chain as part of the proposal distribution. The flexibility of the reversible jump, and Metropolis-Hastings, algorithm allows us to incorporate any such information. However, the real difficulty of the algorithm, and the key to its successful implementation, lies in the choice of the transformation function $\boldsymbol{t}_{m,m'}$. Here we can make explicit use of information regarding any nesting structure that exists. It may be possible to create powerful and efficient proposals. In particular, it may be possible to construct an algorithm whereby proposed transitions between spaces with large differences in dimensionality (in terms of the number of parameters of the statistical model) are accepted with reasonable probability.

In many cases the transformation function $\boldsymbol{t}_{m,m'}$ is chosen to be the identity function. However this obvious choice may result in a Markov chain that mixes poorly. Consider two linear models m and m' such that m' is nested within m. Model m specifies that the data $\boldsymbol{y}$ are normally distributed with mean $\alpha + \beta\boldsymbol{x}$ and variance σ^2, where the vector $\boldsymbol{x}$ is an explanatory variable. Model m' also specifies that the data are normally distributed with mean and variance given by α' and ω respectively. Under model m, α represents the intercept of the regression line whereas α' represent the mean under model m'. It is straightforward to generated a data set such that the value of α is vastly different from that of α'. In fact it would be straightforward to generate a data set such that a move from m to m' that proposes to set $\alpha' = \alpha$ will be rejected with probability 1. Although a trivial example it is clear that the choice of transformation function is of importance when constructing the reversible jump algorithm.

Our current work is motivated by the need for efficient algorithm to explore model spaces, with the additional complexity that some data are missing. This work is presented in the following chapter (Chapter 4). Given the difficulty of achieving reasonable acceptance probabilities for even simple cases, it was decided to first develop suitable algorithms when the data are fully observed.

3.3 A Reversible Jump Scheme for GLM's

There are many possible ways to propose moves between different models. As we are to make use of any nesting structure that exists we initially consider moves between models differ by only a single term, or an interaction between terms. That is if $m' \subset m$ the columns of the design matrix $\boldsymbol{X}_{m'}$ are contained within the columns of $\boldsymbol{X}_m$, and there is no intermediate model m^* such that $m' \subset m^* \subset m$. We term such a move a 'local' move. We will consider alternative moves in due course. In what follows it is always the case that m' is nested within model m.

3.3.1 A Novel Deterministic Transformation Function

Suppose the current state of the Markov chain is $(m, \boldsymbol{\beta}_m)$. Under the GLM assumptions we have

$$\boldsymbol{\eta}_m = \boldsymbol{X}_m \boldsymbol{\beta}_m,$$

where $\boldsymbol{X}_m$ is the design matrix corresponding to model m, and $\boldsymbol{\eta}_m$ is the current linear predictor corresponding to the canonical link. Suppose we propose a move to model m' where $m' \subset m$. For obvious reasons we term this type of move a 'Death Move'. A possible proposal for $\boldsymbol{\beta}'_{m'}$ would be to take the proposed value of the linear predictor to be the orthogonal projection of the current linear predictor onto the subspace defined by the proposed model (orthogonal with respect to an inner product $\boldsymbol{W}$). That is

$$\begin{aligned} \boldsymbol{\beta}'_{m'} &= (\boldsymbol{X}^T_{m'} \boldsymbol{W} \boldsymbol{X}_{m'})^{-1} \boldsymbol{X}^T_{m'} \boldsymbol{W} \boldsymbol{\eta}_m \\ &= (\boldsymbol{X}^T_{m'} \boldsymbol{W} \boldsymbol{X}_{m'})^{-1} \boldsymbol{X}^T_{m'} \boldsymbol{W} \boldsymbol{X}_m \boldsymbol{\beta}_m. \end{aligned} \tag{3.6}$$

Then, when considering a move from m' to m, the proposed $\boldsymbol{\beta}_m$ must be one for which the above equation holds. The transformation function is invertible and differentiable satisfying the dimension matching constraints. The chain is therefore reversible.

The obvious $\boldsymbol{W}$ in the current example is an approximation to an inverse posterior covariance of $\boldsymbol{\eta}$. An accurate approximation is given by the inverse of the Fisher information matrix given by

$$[\boldsymbol{X}_{m'}(\boldsymbol{X}_{m'}^T\boldsymbol{D}\boldsymbol{X}_{m'})^{-1}\boldsymbol{X}_{m'}^T]^{-1}, \tag{3.7}$$

where $\boldsymbol{D}$ is a diagonal matrix with i'th element given by

$$\frac{(\frac{\partial\mu_i}{\partial\eta_i})^2}{Var(y_i)}.$$

This is obtained by direct consideration of the likelihood (e.g. Azzalini, 1996, p. 233). To avoid matrix inversion we simplify the above expression and select $\boldsymbol{W}$ to have i'th element w_i given as

$$w_i = \frac{(\frac{\partial\mu_i}{\partial\eta_i})^2}{Var(y_i)}.$$

Our approximation, equivalent if $\boldsymbol{X}_{m'}$ is invertible, seems to be sufficiently good in practice.

A suitable approximation for w_i in any generalised linear model may be obtained by considering the saturated model. Under this model we have $\hat{\mu}_i = y_i$. Any model that provides a reasonable fit should imply $\hat{y}_i \approx y_i$. Therefore, we choose to approximate $\boldsymbol{W}$ at the start of the algorithm, and this approximation will remain fixed throughout. We thus approximate $\boldsymbol{W}$ with $\hat{\boldsymbol{W}}$, a diagonal matrix with i'th element given by $\hat{w}_i + \epsilon$. To avoid a singular $\hat{\boldsymbol{W}}$, ϵ is chosen to be a small positive real number.

The proposed death move (3.6) can now be written as

$$\boldsymbol{\beta}'_{m'} = (\boldsymbol{X}^T_{m'}\hat{\boldsymbol{W}}\boldsymbol{X}_{m'})^{-1}\boldsymbol{X}^T_{m'}\hat{\boldsymbol{W}}\boldsymbol{X}_m\boldsymbol{\beta}_m. \tag{3.8}$$

If we note that the columns of $\boldsymbol{X}_{m'}$ are also columns of the current design matrix $\boldsymbol{X}_m$, then the current linear predictor can be decomposed as follows

$$\boldsymbol{\eta}_m = \boldsymbol{X}_m\boldsymbol{\beta}_m = [\boldsymbol{X}_{m'}|\boldsymbol{S}]\begin{pmatrix}\boldsymbol{\beta}_{m,\boldsymbol{X}_{m'}}\\ \boldsymbol{\beta}_{m,\boldsymbol{S}}\end{pmatrix}, \tag{3.9}$$

where the additional columns in $\boldsymbol{X}_m$ not in $\boldsymbol{X}_{m'}$ are denoted $\boldsymbol{S}$. The parameters $\boldsymbol{\beta}_{m,\boldsymbol{X}_{m'}}$ and $\boldsymbol{\beta}_{m,\boldsymbol{S}}$ are those that correspond to the columns of the design matrices $\boldsymbol{X}_{m'}$ and $\boldsymbol{S}$ respectively and are contained in $\boldsymbol{\beta}_m$.
Replacing (3.9) in (3.8) we see that

$$\boldsymbol{\beta}'_{m'} = \boldsymbol{\beta}_{m,\boldsymbol{X}_{m'}} + (\boldsymbol{X}^T_{m'}\hat{\boldsymbol{W}}\boldsymbol{X}_{m'})^{-1}\boldsymbol{X}^T_{m'}\hat{\boldsymbol{W}}\boldsymbol{S}\boldsymbol{\beta}_{m,\boldsymbol{S}}. \tag{3.10}$$

This move is purely deterministic and we note once again that any reverse move must satisfy the above equation.
Suppose now we wish to propose a move from model m' to model m ($m' \subset m$) where the current parameter values are $\boldsymbol{\beta}'_{m'}$. For obvious reasons we term such a move a 'Birth Move'. We are required to generate

$$\boldsymbol{\beta}_m = \begin{pmatrix}\boldsymbol{\beta}_{m,\boldsymbol{X}_{m'}}\\ \boldsymbol{\beta}_{m,\boldsymbol{S}}\end{pmatrix}$$

such that

$$\boldsymbol{\beta}'_{m'} = \boldsymbol{\beta}_{m,\boldsymbol{X}_{m'}} + (\boldsymbol{X}^T_{m'}\hat{\boldsymbol{W}}\boldsymbol{X}_{m'})^{-1}\boldsymbol{X}^T_{m'}\hat{\boldsymbol{W}}\boldsymbol{S}\boldsymbol{\beta}_{m,\boldsymbol{S}}.$$

This is clearly satisfied if

$$\beta_{m,\boldsymbol{X}_{m'}} = \beta'_{m'} - (\boldsymbol{X}^T_{m'}\hat{\boldsymbol{W}}\boldsymbol{X}_{m'})^{-1}\boldsymbol{X}^T_{m'}\hat{\boldsymbol{W}}\boldsymbol{S}\beta_{m,\boldsymbol{S}}. \tag{3.11}$$

Without loss of generality we may assume $\beta_{m,\boldsymbol{S}} = \boldsymbol{u}$. We are free to generate $\boldsymbol{u}$ from a distribution of our choice. This distribution must be chosen to ensure irreducibility..

Clearly, (3.11) can be written as

$$\begin{aligned} \beta_m = \begin{pmatrix} \beta_{m,\boldsymbol{X}_{m'}} \\ \beta_{m,\boldsymbol{S}} \end{pmatrix} &= \begin{pmatrix} I & -(\boldsymbol{X}^T_{m'}\hat{\boldsymbol{W}}\boldsymbol{X}_{m'})^{-1}\boldsymbol{X}^T_{m'}\hat{\boldsymbol{W}}\boldsymbol{S} \\ 0 & I \end{pmatrix} \begin{pmatrix} \beta'_{m'} \\ \boldsymbol{u} \end{pmatrix} \\ &= \boldsymbol{t}_{m,m'}((\beta_{m'}, \boldsymbol{u})^T). \end{aligned} \tag{3.12}$$

We note this to be a linear and invertible transformation of $\beta'_{m'}$ and $\boldsymbol{u}$. The transformation is an upper triangular matrix and with all diagonal elements equal to 1. Hence

$$\begin{vmatrix} I & -(\boldsymbol{X}^T_{m'}\hat{\boldsymbol{W}}\boldsymbol{X}_{m'})^{-1}\boldsymbol{X}^T_{m'}\hat{\boldsymbol{W}}\boldsymbol{S} \\ 0 & I \end{vmatrix} = 1. \tag{3.13}$$

Since the determinant of the inverse of a matrix is the reciprocal of the determinant of the matrix the Jacobian of the deterministic death proposal is also 1.

3.3.2 A Suitable Proposal Distribution

In order to obtain a suitable approximation to the posterior distribution of $(\boldsymbol{u}|\beta'_{m'}, m', \boldsymbol{y})$ we approximate the posterior distribution of β_m using a normal distribution. We centre our approximation at an approximate least squares estimate of β_m and approximate the posterior covariance matrix with an estimate for the inverse Fisher information matrix; i.e. we let

$$\beta_m \sim N((\boldsymbol{X}^T_m\hat{\boldsymbol{W}}\boldsymbol{X}_m)^{-1}\boldsymbol{X}^T_m\hat{\boldsymbol{W}}\hat{\eta}, (\boldsymbol{X}^T_m\hat{\boldsymbol{W}}\boldsymbol{X}_m)^{-1}),$$

where $\hat{\eta}$ is the value of the linear predictor under the saturated model. This proposal distribution seems natural and follows from the work of Brooks et al. (2003). Indeed, the first and higher order methods of Brooks et al. (2003) (taking derivatives of the natural logarithm of the acceptance probability) correspond to matching derivatives of the log-proposal and log-conditional posterior density function at a centring point. This can lead to appealing proposals such as a normal distribution, centred at the posterior conditional mode with variance given by the negative inverse second derivative of the log-conditional posterior density at the mode.

Having made this assumption to obtain the distribution for $(\boldsymbol{\beta}'_{m'}, \boldsymbol{u})$ we invert the transformation in (3.12). This inverted transformation is given by

$$\begin{pmatrix} \boldsymbol{\beta}'_{m'} \\ \boldsymbol{u} \end{pmatrix} = \begin{pmatrix} I & (\boldsymbol{X}^T_{m'}\hat{\boldsymbol{W}}\boldsymbol{X}_{m'})^{-1}\boldsymbol{X}^T_{m'}\hat{\boldsymbol{W}}\boldsymbol{S} \\ 0 & I \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_{m,\boldsymbol{X}_{m'}} \\ \boldsymbol{\beta}_{m,\boldsymbol{S}} \end{pmatrix}.$$

Since we assumed $\boldsymbol{\beta}_m = (\boldsymbol{\beta}_{m,\boldsymbol{X}_{m'}}, \boldsymbol{\beta}_{m,\boldsymbol{S}})$ to have a multivariate normal distribution it follows that $(\boldsymbol{\beta}'_{m'}, \boldsymbol{u})$ also has a multivariate normal distribution. The covariance matrix for this approximate joint posterior density of $(\boldsymbol{\beta}'_{m'}, \boldsymbol{u})$ is then given by

$$\Sigma = \begin{pmatrix} I & (\boldsymbol{X}^T_{m'}\hat{\boldsymbol{W}}\boldsymbol{X}_{m'})^{-1}\boldsymbol{X}^T_{m'}\hat{\boldsymbol{W}}\boldsymbol{S} \\ 0 & I \end{pmatrix} \boldsymbol{V} \begin{pmatrix} I & 0 \\ (\boldsymbol{X}^T_{m'}\hat{\boldsymbol{W}}\boldsymbol{X}_{m'})^{-1}\boldsymbol{X}^T_{m'}\hat{\boldsymbol{W}}\boldsymbol{S} & I \end{pmatrix}.$$

In the above $\boldsymbol{V} = (\boldsymbol{X}^T_m\hat{\boldsymbol{W}}\boldsymbol{X}_m)^{-1}$. Again note that $\boldsymbol{X}_m = (\boldsymbol{X}_{m'}|\boldsymbol{S})$ hence the inverse of the covariance matrix is a block matrix given by

$$\begin{aligned} \boldsymbol{X}^T_m\hat{\boldsymbol{W}}\boldsymbol{X}_m &= (\boldsymbol{X}_{m'}|\boldsymbol{S})^T\hat{\boldsymbol{W}}(\boldsymbol{X}_{m'}|\boldsymbol{S}) \\ &= \begin{pmatrix} \boldsymbol{X}^T_{m'}\hat{\boldsymbol{W}}X_{m'} & X^T_{m'}\hat{\boldsymbol{W}}\boldsymbol{S} \\ \boldsymbol{S}^T\hat{\boldsymbol{W}}\boldsymbol{X}_{m'} & \boldsymbol{S}^T\hat{\boldsymbol{W}}\boldsymbol{S}. \end{pmatrix} \end{aligned}$$

Appendix A gives the inverse and determinants of block matrices. Let $\boldsymbol{C}$ denote the Schur component of this matrix with respect to $\boldsymbol{X}_{m'}^T\hat{\boldsymbol{W}}\boldsymbol{X}_{m'}$. Then $\boldsymbol{C}$ is given by

$$\begin{aligned}\boldsymbol{C} &= \boldsymbol{S}^T\hat{\boldsymbol{W}}\boldsymbol{S} - \boldsymbol{S}^T\hat{\boldsymbol{W}}\boldsymbol{X}_{m'}(\boldsymbol{X}_{m'}^T\hat{\boldsymbol{W}}\boldsymbol{X}_{m'})^{-1}\boldsymbol{X}_{m'}^T\hat{\boldsymbol{W}}\boldsymbol{S} \\ &= \boldsymbol{S}^T\hat{\boldsymbol{W}}\boldsymbol{S} - \boldsymbol{S}^T\hat{\boldsymbol{W}}\boldsymbol{P}_{m'}\boldsymbol{S} \\ &= \boldsymbol{S}^T\hat{\boldsymbol{W}}(I - \boldsymbol{P}_{m'})\boldsymbol{S},\end{aligned}$$

where $(I - \boldsymbol{P}_{m'})$ is an orthogonal projection matrix. The covariance matrix, denoted $\boldsymbol{V}$ is a block matrix and is given by

$$\boldsymbol{V} = (\boldsymbol{X}_m^T\hat{\boldsymbol{W}}\boldsymbol{X}_m)^{-1} = \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix},$$

where

$$\begin{aligned}&V_{11} = (\boldsymbol{X}_{m'}^T\hat{\boldsymbol{W}}\boldsymbol{X}_{m'})^{-1}(I + \boldsymbol{X}_{m'}^T\hat{\boldsymbol{W}}\boldsymbol{S}\boldsymbol{C}^{-1}\boldsymbol{S}^T\hat{\boldsymbol{W}}\boldsymbol{X}_{m'}(\boldsymbol{X}_{m'}^T\hat{\boldsymbol{W}}\boldsymbol{X}_{m'})^{-1}) \\ &V_{12} = -(\boldsymbol{X}_{m'}^T\hat{\boldsymbol{W}}\boldsymbol{X}_{m'})^{-1}\boldsymbol{X}_{m'}^T\hat{\boldsymbol{W}}\boldsymbol{S}\boldsymbol{C}^{-1} \\ &V_{21} = -\boldsymbol{C}^{-1}\boldsymbol{S}^T\hat{\boldsymbol{W}}\boldsymbol{X}_{m'}(\boldsymbol{X}_{m'}^T\hat{\boldsymbol{W}}\boldsymbol{X}_{m'})^{-1} \\ &V_{22} = \boldsymbol{C}^{-1}.\end{aligned}$$

Combining all of the above the covariance matrix of the normal distribution for $(\beta'_{m'}, \boldsymbol{u})$ is given by

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22}. \end{pmatrix}$$

The elements of this matrix are as follows:

$$
\begin{aligned}
\Sigma_{11} &= V_{11} + (\boldsymbol{X}_{m'}^T \hat{\boldsymbol{W}} \boldsymbol{X}_{m'})^{-1} \boldsymbol{X}_{m'}^T \hat{\boldsymbol{W}} \boldsymbol{S} V_{21} \\
&+ V_{12} (\boldsymbol{X}_{m'}^T \hat{\boldsymbol{W}} \boldsymbol{X}_{m'})^{-1} \boldsymbol{X}_{m'}^T \hat{\boldsymbol{W}} \boldsymbol{S} \\
&+ (\boldsymbol{X}_{m'}^T \hat{\boldsymbol{W}} \boldsymbol{X}_{m'})^{-1} \boldsymbol{X}_{m'}^T \hat{\boldsymbol{W}} \boldsymbol{S} V_{22} (\boldsymbol{X}_{m'}^T \hat{\boldsymbol{W}} \boldsymbol{X}_{m'})^{-1} \boldsymbol{X}_{m'}^T \hat{\boldsymbol{W}} \boldsymbol{S} \\
&= (\boldsymbol{X}_{m'}^T \hat{\boldsymbol{W}} \boldsymbol{X}_{m'})^{-1} (I + \boldsymbol{X}_{m'}^T \hat{\boldsymbol{W}} \boldsymbol{S} \boldsymbol{C}^{-1} \boldsymbol{S}^T \hat{\boldsymbol{W}} \boldsymbol{X}_{m'} (\boldsymbol{X}_{m'}^T \hat{\boldsymbol{W}} \boldsymbol{X}_{m'})^{-1}) \\
&- (\boldsymbol{X}_{m'}^T \hat{\boldsymbol{W}} \boldsymbol{X}_{m'})^{-1} \boldsymbol{X}_{m'}^T \hat{\boldsymbol{W}} \boldsymbol{S} \boldsymbol{C}^{-1} \boldsymbol{S}^T \hat{\boldsymbol{W}} \boldsymbol{X}_{m'} (\boldsymbol{X}_{m'}^T \hat{\boldsymbol{W}} \boldsymbol{X}_{m'})^{-1} \\
&- (\boldsymbol{X}_{m'}^T \hat{\boldsymbol{W}} \boldsymbol{X}_{m'})^{-1} \boldsymbol{X}_{m'}^T \hat{\boldsymbol{W}} \boldsymbol{S} \boldsymbol{C}^{-1} (\boldsymbol{X}_{m'}^T \hat{\boldsymbol{W}} \boldsymbol{X}_{m'})^{-1} \boldsymbol{X}_{m'}^T \hat{\boldsymbol{W}} \boldsymbol{S} \\
&+ (\boldsymbol{X}_{m'}^T \hat{\boldsymbol{W}} \boldsymbol{X}_{m'})^{-1} \boldsymbol{X}_{m'}^T \hat{\boldsymbol{W}} \boldsymbol{S} \boldsymbol{C}^{-1} (\boldsymbol{X}_{m'}^T \hat{\boldsymbol{W}} \boldsymbol{X}_{m'})^{-1} \boldsymbol{X}_{m'}^T \hat{\boldsymbol{W}} \boldsymbol{S} \\
&= (\boldsymbol{X}_{m'}^T \hat{\boldsymbol{W}} \boldsymbol{X}_{m'})^{-1} \\
\Sigma_{12} &= V_{12} + (\boldsymbol{X}_{m'}^T \hat{\boldsymbol{W}} \boldsymbol{X}_{m'})^{-1} \boldsymbol{X}_{m'}^T \hat{\boldsymbol{W}} \boldsymbol{S} V_{22} \\
&= (\boldsymbol{X}_{m'}^T \hat{\boldsymbol{W}} \boldsymbol{X}_{m'})^{-1} \boldsymbol{X}_{m'}^T \hat{\boldsymbol{W}} \boldsymbol{S} \boldsymbol{C}^{-1} \\
&- (\boldsymbol{X}_{m'}^T \hat{\boldsymbol{W}} \boldsymbol{X}_{m'})^{-1} \boldsymbol{X}_{m'}^T \hat{\boldsymbol{W}} \boldsymbol{S} \boldsymbol{C}^{-1} \\
&= 0 \\
\Sigma_{21} &= V_{21} + V_{22} (\boldsymbol{X}_{m'}^T \hat{\boldsymbol{W}} \boldsymbol{X}_{m'})^{-1} \boldsymbol{X}_{m'}^T \hat{\boldsymbol{W}} \boldsymbol{S} \\
&= C^{-1} (\boldsymbol{X}_{m'}^T \hat{\boldsymbol{W}} \boldsymbol{X}_{m'})^{-1} \boldsymbol{X}_{m'}^T \hat{\boldsymbol{W}} \boldsymbol{S} \\
&- C^{-1} (\boldsymbol{X}_{m'}^T \hat{\boldsymbol{W}} \boldsymbol{X}_{m'})^{-1} \boldsymbol{X}_{m'}^T \hat{\boldsymbol{W}} \boldsymbol{S} \\
&= 0 \\
\Sigma_{22} &= V_{22} \\
&= C^{-1}.
\end{aligned}
$$

Therefore, the covariance matrix is simply the block matrix given by

$$
\Sigma = \begin{pmatrix} (\boldsymbol{X}_{m'}^T \hat{\boldsymbol{W}} \boldsymbol{X}_{m'})^{-1} & 0 \\ 0 & (\boldsymbol{S}^T \hat{\boldsymbol{W}} (I - \boldsymbol{P}_{m'}) \boldsymbol{S})^{-1} \end{pmatrix}.
$$

Through a similar argument it can be shown that we should centre this distribution at

$$(\boldsymbol{S}^T\hat{\boldsymbol{W}}(I-\boldsymbol{P}_{m'})\boldsymbol{S})^{-1}\boldsymbol{S}^T\hat{\boldsymbol{W}}(I-\boldsymbol{P}_{m'})\boldsymbol{\eta}_{\boldsymbol{y}}.$$

This is intuitive. If we select $\boldsymbol{Z}$ to be a matrix whose columns form a basis for $V_m \cap V_{m'}^{\perp}$, where V_m is the vector space spanned by the columns of m and $V_{m'}^{\perp}$ is a vector space orthogonal to $V_{m'}$, then $\boldsymbol{Z}$ can be obtained by projecting the columns of $\boldsymbol{S}$ onto the orthogonal complement of $V_{m'}$ using the projection matrix $(I-\boldsymbol{P}_{m'})$. I.e. let $\boldsymbol{Z} = (I-\boldsymbol{P}_{m'})\boldsymbol{S}$. It is easily seen that

$$(\boldsymbol{Z}^T\hat{\boldsymbol{W}}\boldsymbol{Z})^{-1} = (\boldsymbol{S}^T\hat{\boldsymbol{W}}(I-\boldsymbol{P}_{m'})\boldsymbol{S})^{-1}$$

and that

$$(\boldsymbol{Z}^T\hat{\boldsymbol{W}}\boldsymbol{Z})^{-1}\hat{\boldsymbol{W}}\boldsymbol{Z}^T\hat{\boldsymbol{\eta}} = \boldsymbol{S}^T\hat{\boldsymbol{W}}(I-\boldsymbol{P}_{m'})\boldsymbol{S})^{-1}\boldsymbol{S}^T\hat{\boldsymbol{W}}(I-\boldsymbol{P}_{m'})\hat{\boldsymbol{\eta}}.$$

Clearly $\boldsymbol{u}$ parameterises $V_m \cap V_{m'}^{\perp}$. We thus have the following reversible jump algorithm.

3.3.3 Death move

Let the current parameter values of the Markov chain at time t be $(m, \boldsymbol{\beta}_m)$, with the dimension of $\boldsymbol{\beta}_m$ denoted p_m. Proceed as follows

- Propose a move to model m' (nested with m) with probability $j(m, m')$.
- Partition $\boldsymbol{X}_m\boldsymbol{\beta}_m = [\boldsymbol{X}_{m'}|\boldsymbol{S}]\begin{pmatrix}\boldsymbol{\beta}_{m,\boldsymbol{X}_{m'}}\\ \boldsymbol{\beta}_{m,\boldsymbol{S}}\end{pmatrix}$.
- Set $\boldsymbol{\beta}'_{m'} = \boldsymbol{\beta}_{m,\boldsymbol{X}_{m'}} + (\boldsymbol{X}_{m'}^T\hat{\boldsymbol{W}}\boldsymbol{X}_{m'})^{-1}\boldsymbol{X}_{m'}^T\hat{\boldsymbol{W}}\boldsymbol{S}\boldsymbol{\beta}_{m,\boldsymbol{S}}$.
- Accept this proposed move with probability

$$\frac{f(\boldsymbol{y}|m', \boldsymbol{\beta}_{m'}^p)f(\boldsymbol{\beta}_{m'}^p|m')f(m')j(m',m)g(\boldsymbol{\beta}_{m,\boldsymbol{S}})}{f(\boldsymbol{y}|m,\boldsymbol{\beta}_m)f(\boldsymbol{\beta}_m|m)f(m)j(m,m')} \wedge 1, \qquad (3.14)$$

where g is the density function for the normal distribution

$$N((\boldsymbol{S}^T\hat{\boldsymbol{W}}(I-\boldsymbol{P}_{m'})\boldsymbol{S})^{-1}\boldsymbol{S}^T\hat{\boldsymbol{W}}(I-\boldsymbol{P}_{m'})\hat{\boldsymbol{\eta}},(\boldsymbol{S}^T\hat{\boldsymbol{W}}(I-\boldsymbol{P}_{m'})\boldsymbol{S})^{-1}).$$

- Else reject this proposed move and remain at $(m,\boldsymbol{\beta}_m)$.

3.3.4 Birth move

Let the current parameter values of the Markov chain be $(m',\boldsymbol{\beta}_{m'})$, with the dimension of $\boldsymbol{\beta}_{m'}$ being $p_{m'}$.

- Propose a move to model m $(m' \subset m)$ with probability $j(m',m)$.
- Generate $\boldsymbol{u}$ from

$$\boldsymbol{u} \sim N((\boldsymbol{S}^T\hat{\boldsymbol{W}}(I-\boldsymbol{P}_{m'})\boldsymbol{S})^{-1}\boldsymbol{S}^T\hat{\boldsymbol{W}}(I-\boldsymbol{P}_{m'})\hat{\boldsymbol{\eta}},(\boldsymbol{S}^T\hat{\boldsymbol{W}}(I-\boldsymbol{P}_{m'})\boldsymbol{S})^{-1})$$

This distribution has density function denoted by g.

- Set $\boldsymbol{\beta}_m = \begin{pmatrix} I & -(\boldsymbol{X}_{m'}^T\hat{\boldsymbol{W}}\boldsymbol{X}_{m'})^{-1}\boldsymbol{X}_{m'}^T\hat{\boldsymbol{W}}\boldsymbol{S} \\ 0 & I \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}'_{m'} \\ \boldsymbol{u} \end{pmatrix}.$
- Accept this proposed move with probability

$$\frac{f(\boldsymbol{y}|m,\boldsymbol{\beta}_m^p)f(\boldsymbol{\beta}_m^p|m)f(m)j(m,m')}{f(\boldsymbol{y}|m',\boldsymbol{\beta}_{m'})f(\boldsymbol{\beta}_{m'}|m')f(m)j(m',m)g(\boldsymbol{u})} \wedge 1. \qquad (3.15)$$

- Else reject this proposed move and remain at $(m',\boldsymbol{\beta}'_{m'})$.

Note that in the acceptance probability of both birth and death moves there is no Jacobian. This is due the transformation being an upper triangular square matrix with diagonal elements equal to one. It is important to note that the transformation must still be made. This simple algorithm should alleviate any worries the Jacobian often causes.

Constructing the reversible jump algorithm in this fashion has several advantages. The algorithm is simple and inexpensive to implement. There is no need for inefficient pilot chains used by many authors to determine a suitable proposal distribution. We are not required to tune the proposal variance prior to running the algorithm, or adapt this proposal variance during the algorithm. The algorithm can be viewed as quasi-adaptive as, at each iteration, proposal distributions for local (nested) moves are optimised. However, unlike adaptive algorithms, there is no difficulty ensuring stationarity or verifying ergodicity.

Finally, we will show that using the above framework more general moves can be constructed. The moves would allow transitions between models that retain some common parameters but are not nested, together with a non-deterministic death move and an interesting Metropolis-Hastings update. Of course, since we are interested in effectively using information contained in the current state of the Markov chain, local moves are our main focus. The move types presented within this chapter can be combined with standard Metropolis-Hastings to form a powerful and efficient algorithm.

3.4 Example 1: Crime and Punishment

In order to illustrate how this reversible jump scheme is applied to linear models we consider the data set on U.S. crime rates discussed by Ehrlich (1973) and analysed within a Bayesian framework by Raftery et al. (1997) and Nott and Green (2004). Ehrlich collected data from $n = 47$ U.S. States. He empirically tested his theoretical argument that the decision to engage in criminal activity is a rational choice, determined by its costs and benefits relative to other legitimate activities. Ehrlich's data was corrected by Vandaele (1978) who included three additional variables. The data can be found in Appendix E and consist of the 15 variables, details of which are given in Table 3.1. The response denoted $\boldsymbol{y}$ is the rate of crimes in a particular category per head of population. As in the analysis of Raftery et al. (1997) all data were transformed logarithmically with the covariates additionally being centred. The task is to determine which of the 15 predictors of crime rate are related to the response.

Raftery et al. (1997) stated that standard diagnostic checks do not reveal any gross violations of the assumptions underlying the normal regression model.

3.4.1 Notation and Prior Distributions

Following the notation of Ntzoufras et al. (2001) we define a model m by the pair (γ^m, S) where $\boldsymbol{\gamma}^m = (\gamma_1^m, ..., \gamma_{15}^m) \in \{0,1\}^{15}$ is a vector indicating which of the covariates are included in model m. Since $q = 15$, and we are not interested in interactions between covariates, there are $2^{15} = 32,768$ possible models. Here S implies that

$$\boldsymbol{y}|m, \boldsymbol{\beta}_m, \sigma^2 \sim N_{p_m}(\boldsymbol{X}_m\boldsymbol{\beta}_m, \sigma^2\mathbb{I}_n),$$

and hence

$$f(\boldsymbol{y}|m, \boldsymbol{\beta}_m, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\{-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}_m\boldsymbol{\beta}_m)^T(\boldsymbol{y} - \boldsymbol{X}_m\boldsymbol{\beta}_m)\}.$$

As in Raftery et al. (1997) we assume the conjugate normal-inverse-gamma distribution for $f(\boldsymbol{\beta}_m, \sigma^2)$. This prior density is given by

$$f(\boldsymbol{\beta}_m, \sigma^2|m) = (\sigma^2)^{-(\nu+p_m+2)/2} \exp\{-(\frac{1}{2}(\boldsymbol{\beta}_m - \boldsymbol{\mu}_m)^T\boldsymbol{V}_m^{-1}(\boldsymbol{\beta}_m - \boldsymbol{\mu}_m) + \lambda\nu)\},$$

for hyperparameters ν, λ, $\boldsymbol{\mu}$ and prior covariance $\boldsymbol{V}$ (see O'Hagan and Forster 2004, for full details). We follow Raftery et al. (1997) and set $\nu = 2.58$, $\lambda = 0.28$ and $\boldsymbol{\mu}_m = (\hat{\beta}_0, 0, ..., 0)^T$ where $\hat{\beta}_0$ is the ordinary least squares estimate of the intercept term under model m. The prior covariance matrix $\boldsymbol{V}_m$ has diagonal elements $\phi^2 s_i^{-2}$ for the continuous variables, where s_i^2 denotes the sample variance of the ith predictor, and $\phi^2(\frac{1}{n}\boldsymbol{x}_i^T\boldsymbol{x}_i)^{-1}$ for the sole categorical variable 'southern state'. We follow Raftery et al. (1997) in setting $\phi^2 = 2.85^2$ and that all models are a-priori equally likely.

The marginal likelihood $f(\boldsymbol{y}|m)$ is tractable and is given below:

$$\begin{aligned} f(\boldsymbol{y}|m) &= \frac{\Gamma(\frac{\nu+n}{2})(\nu\lambda)^{1/2}}{\pi^{n/2}\Gamma(\nu/2)|I+\boldsymbol{X}_m\boldsymbol{V}_m\boldsymbol{X}^T|^{1/2}} \\ &\times \left((\boldsymbol{y}-\boldsymbol{X}_m\boldsymbol{\mu}_m)^T(I+\boldsymbol{X}_m\boldsymbol{V}_m\boldsymbol{X}_m^T)^{-1}(\boldsymbol{y}-\boldsymbol{X}_m\boldsymbol{\mu}_m)+\lambda\nu\right)^{-\frac{\nu+n}{2}}. \end{aligned}$$

Since q is relatively small we can compute $f(m|\boldsymbol{y})$ exactly, obtaining the normalising constant by summing over all $2^{15} = 32,768$ models.

Clearly, posterior inferences and predictions will depend critically upon the choice of the prior distribution for parameters $(\boldsymbol{\beta}_m, m)$. Raftery et al. (1997) compare their prior for the non-categorical covariates with the actual distribution of coefficients from real data. They consider 13 data sets from several regression text books and provide a histogram of the 100 coefficients from standardised data. All coefficients lie within the interval $[-2, 2]$ and accordingly their prior gives reasonable support to this interval. The prior distribution is also flat across $[-1, 1]$ where most of the coefficients are observed.

Assuming all models a-priori equally likely is also a strong assumption, as is the assumption that β_i is independent of β_j for all $i \neq j$. However, there is little prior information regarding the dependencies between the β's. There is also no information available to construct the prior model probabilities. Further arguments presented by Raftery et al. (1997) suggest that their prior is sensible and well formulated.

The purpose of this section is to provide an example of the reversible jump algorithm that we have constructed. Had the purpose been to make posterior inference and prediction, then we would have to check the sensitivity of our inferences to any and every prior assumption that we make.

3.4.2 Reversible Jump and Results

We implemented the reversible jump scheme detailed in Section 3.3. We proposed moves between nested models only, selecting to add or remove a term with equal probability. In addition to these moves the parameter σ^2 was updated by Gibb's sampling from the tractable posterior $f(\sigma^2|\boldsymbol{\beta}_m, \boldsymbol{y})$. In total, 8,142 different models were visited during 1,000,000 iterations of the algorithm. The Markov chain proved to be mobile with, on average, a change of model every four iterations (acceptance probability 0.27). The sample was thinned by taking every 20th iterate providing a sample of 50,000 points for analysis. For each of the 15 variables, we computed Monte Carlo standard errors of the estimated marginal probabilities of inclusion (see George and McCulloch 1993). Let $\bar{\gamma}_i = P(\gamma_i \neq 0|\boldsymbol{y})$ be the sample mean of the Markov chain iterates. Then the Monte Carlo standard error of $\bar{\gamma}_i$ is given by

$$SE(\bar{\gamma}_i) = \left[\frac{1}{k}\sum_{|h|<k}\left(1 - \frac{|h|}{k}\right)R_i(h)\right]^{1/2},$$

where $R_i(h)$ is the estimated autocovariance function for the Markov chain iterates for γ_i. The above sum must be truncated to ensure the consistency of the estimator, and this truncation point is chosen based on when the estimated auto-covariance function decays to zero. For more detail see Geyer (1992).

Table 3.1 provides, for each variable, estimated posterior inclusion probabilities together with Monte Carlo standard errors (The exact probabilities are given in parentheses). Figure 3.1 provides Markov chain diagnostic plots of the output for the predictor 'Police expenditure in 1960'. We see there is little evidence to suggest a lack of convergence. Table 3.2 displays models with a posterior model probability of 1% or greater. These are similar to those found in Raftery et al. (1997) but, on

Table 3.1: Crime data: Estimated posterior inclusion probabilities (The exact probabilities are given in parentheses) together with Monte Carlo standard errors.

Predictor Number	Predictor	$\bar{\gamma}_i$	Standard Error
1	Percentage of males aged 14-24	0.430 (0.565)	0.0032
2	Indicator variable for southern state	0.132 (0.152)	0.0017
3	Mean years of schooling	0.655 (0.814)	0.0036
4	Police expenditure in 1960	0.657 (0.686)	0.0036
5	Police expenditure in 1959	0.523 (0.522)	0.0037
6	Labour force participation rate	0.091 (0.087)	0.0014
7	Number of males per 1000 females	0.120 (0.108)	0.0017
8	State population	0.154 (0.184)	0.0019
9	Number of non-whites per 1000 people	0.267 (0.369)	0.0027
10	Unemployment rate of urban males aged 14-24	0.073 (0.087)	0.0013
11	Unemployment rate of urban males aged 35-39	0.152 (0.232)	0.0020
12	Wealth	0.232 (0.265)	0.0022
13	Income equality	0.955 (0.982)	0.0016
14	Probability of imprisonment	0.394 (0.553)	0.0032
15	Average time served in state prison	0.097 (0.126)	0.0015

the whole, have fewer terms. They are almost identical to the exact posterior model probabilities given in Table 3.3, which can be calculated in this example.

The probabilities illustrate the difficulty in assessing the convergence of the Markov chain. Although the diagnostic plots illustrate adequate mixing the posterior inclusion probabilities differ from the exact probabilities. These probabilities also differ from those presented in Raftery et al. (1997) and they are not reported in Nott and Green (2004). However, the four methods of calculating the posterior model probabilities (RJMCMC, MCMCMC, Nott and Green (2004) and complete enumeration)

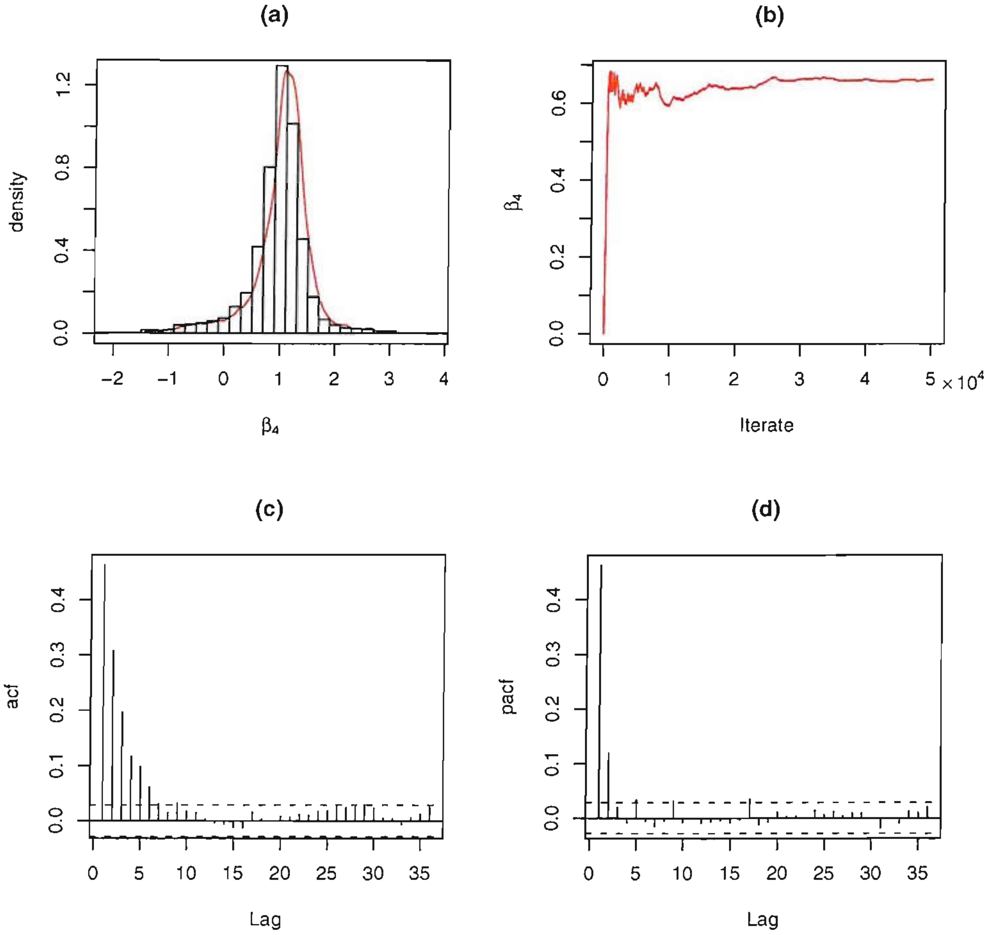


Figure 3.1: MCMC output plots. Plot (a): Histogram of the posterior distribution of β_4. Plot (b): A running mean of the inclusion probability for β_4. Plot (c): Auto-correllelogram of γ_4 . Plot (d): Partial auto-correllelogram of γ_4.

would result in similar predictive inference. This lack of convergence is probably due to the number of models under consideration and the fact that there is considerable posterior uncertainty. The results of Raftery et al. (1997) for example, are based on a sample of 30,000 iterates and this may not suffice.

Table 3.2: Crime data: Models with estimated posterior model probabilities of 1% or greater.

Model							Posterior model probability (%)
	3	4			13		2.86
1	3	4			13		2.50
		4			13		2.30
	3		5		13		1.86
			5		13		1.65
	3	4			13	14	1.23
1	3	4			13	14	1.22
	3		5		13	14	1.18
1	3		5		13		1.18
1	3	4		11	13		1.06
	3	4	5		13		1.01

The standard errors of Table 3.1 are an improvement to those of Nott and Green (2004). Nott and Green (2004) argued that Monte Carlo standard errors should not be compared based on an equal number of iterates for all methods, but on differing numbers of iterates for each method produced within a given time. Thus the simplicity of our algorithm is favourable. As already mentioned, our algorithm requires no initial calculations, has no algorithmic parameters and needs no tuning.

3.5 Example 2: Alcohol, Obesity and Hypertension

In order to illustrate how the reversible jump scheme detailed in this chapter can be applied to hierarchical log-linear models we consider the data of Knuiman and

Table 3.3: Crime data: Models with exact posterior model probability of 1% or greater. The last row corresponds to the model with greatest posterior probability as stated by Raftery et al. (1997).

				Model				Posterior model probability (%)
1	3	4				13		2.40
1	3	4				13	14	1.57
	3	4				13		1.54
1	3	4			11	13		1.46
	3	4		9		13	14	1.19
	3		5	9		13	14	1.07
1	3	4		9		13	14	1.05
1	3		5			13	14	1.01
1	3	4			11	13	14	1.01
	3	4				13	14	1.01
1	3	4		9	11	13	14	0.40

Speed (1988). The data, in the form of a $2 \times 4 \times 3$ contingency table, can be found in Table 3.4. It concerns 491 subjects classified according to three categorical variables (factors) $C = \{H, A, O\}$, where Hypertension (H: Yes or No), Alcohol Intake (A: 0, 1-2, 3-5, 6+ drinks per day) and Obesity (O: Low, Average, High). The data has been analysed by Dellaportas and Forster (1999).

3.5.1 Notation and Prior Distributions

If we assume the main effects (H, A, O) are present in all models under consideration then there are nine models in the class of hierarchical log-linear models. These are

$$\{H + A + O, HA + O, HO + A, AO + H, HA + HO,$$
$$HA + AO, HO + AO, HA + HO + AO, HAO\},$$

Table 3.4: Alcohol, obesity and hypertension data.

		Alcohol Intake			
Obesity	Hypertension	0	1-2	3-5	6+
Low	Yes	5	9	8	10
Low	No	40	36	33	24
Average	Yes	6	9	11	14
Average	No	33	23	35	30
High	Yes	9	12	19	19
High	No	24	25	28	29

where the models are denoted by the sum of their generating terms. Here m denotes one of the 9 competing models given above. Each model corresponds to a set γ of factors, where γ is contained in the power set of C denoted $\mathcal{P}(C)$. Each log-linear model specifies that each y_i is independently distributed according to a Poisson random variable with mean $\mathbb{E}[y_i] = \mu_i$. Thus each model posits

$$\log(\boldsymbol{\mu}) = \boldsymbol{X}_m\boldsymbol{\beta}_m$$

and therefore the log-likelihood is given by

$$\log(f(\boldsymbol{y}|\boldsymbol{\beta}_m, m)) \propto \sum_i y_i(\boldsymbol{x}_i^T\boldsymbol{\beta}_m) - \sum_i \exp(\boldsymbol{x}_i^T\boldsymbol{\beta}_m).$$

As in Dellaportas and Forster (1999) and following Knuiman and Speed (1988) we assume that, for each term $a \in \gamma$

$$\boldsymbol{\beta}_a \sim N(\boldsymbol{\theta}_a, \alpha_a^2\boldsymbol{V}_a.) \tag{3.16}$$

In the absence of prior information about $\boldsymbol{\beta}_a$, we set $\boldsymbol{\theta}_a$ to be a vector of zero's. We set

$$\boldsymbol{V}_a = \frac{1}{|I|}\prod_{b\in a}|I_b|\bigotimes_{b\in a}\left(\boldsymbol{I}_{(|I_b|-1)} - \frac{1}{|I_b|}\boldsymbol{J}_{(|I_b|-1)}\right), \tag{3.17}$$

where $|I| = 24$ is the number of cells in the contingency table and $|I_b|$ is the number of levels of factor b. $\boldsymbol{J}_n$ is an $n \times n$ matrix of all 1's, whilst $\boldsymbol{I}_n$ is the $n \times n$ identity matrix. For example, let a be the interaction term $O : A$. Then $\boldsymbol{\beta}_a$ is normally distributed with mean $\boldsymbol{\theta}_a$, a zero vector of length 6. The covariance matrix is given by

$$\boldsymbol{V}_a = \frac{4 \times 3}{24}(I_3 - \frac{1}{3}\boldsymbol{J}_3) \bigotimes (I_2 - \frac{1}{2}\boldsymbol{J}_2)$$

Thus, the prior distribution for $\boldsymbol{\beta}_m$ is normal, with covariance matrix the block diagonal matrix with elements $\boldsymbol{V}_a$ for all $a \in m$. This prior is invariant to arbitrary permutations of the levels of each factor. The prior is proper and α_a^2 can be chosen to reflect prior belief. Dellaportas and Forster (1999) advocate that the value of $\alpha_a^2 \propto |I|$ since the above prior distribution depends on the number of levels of all factors ($\frac{1}{|I|}$ appears in 3.17). They further note that the inverse of the unit information matrix, evaluated at the prior mean, where all cell probabilities are equal, is equal to $|I|$ times the block diagonal matrix with elements $\boldsymbol{V}_a$. Thus in selecting

$$\alpha_a^2 = k|I|$$

they interpret k as the number of units of prior information at the prior mean. This is comparable to the g-prior of Zellner (1986) for linear regression models. Dellaportas and Forster (1999) suggest values of $k = 1, 2$ and 4.

3.5.2 Reversible Jump and Results

A reversible jump Markov chain was constructed as detailed (3.3). We proposed moves between nested models only, selecting to add or remove a term, or an interaction between terms, with equal probability. Posterior model probabilities for the four most probable models based on a 100,000 iterations can be found in Table 3.5. These results are identical to those found in Dellaportas and Forster (1999). However, the algorithm of Dellaportas and Forster (1999) was less mobile, achieving

Table 3.5: Alcohol, obesity and hypertension data: Posterior model probabilities.

Model	Posterior probability		
	$\alpha^2 = 24$	$\alpha^2 = 48$	$\alpha^2 = 96$
$H + O + A$	0.505	0.677	0.81
$HO + A$	0.478	0.317	0.188
$HA + O$	0.009	0.004	0.001
$HO + HA$	0.008	0.002	0.001
Acceptance probability	0.244	0.315	0.297

a change in model, on average, once in every nineteen iterations. We observed a change in model every three to four iterations. The exact acceptance probabilities are given in Table 3.5. Unlike Dellaportas and Forster (1999) we did not require time consuming pilot runs for proposal distributions, nor did we tune the Markov chain.

As an aside note the occurrence of Lindley's paradox. As the model specific parameter prior distribution becomes more diffuse, greater weight is given to the model containing no interactions. This is seen in the first row of Table 3.5.

3.6 Example 3: A Simulated Example for Poisson Response with Highly Collinear Regressors

Our third example is based on an example described by George and McCulloch (1997). They simulated a data set as follows. Let $\boldsymbol{Z}^1, ..., \boldsymbol{Z}^{15}, \boldsymbol{Z}$ be vectors of independent standard normal variables of length 180. Let $\boldsymbol{X}^j = \boldsymbol{Z}^j + 2\boldsymbol{Z}$ for $j = 1, 3, 5, 8, 9, 10, 12, 13, 14, 15$. Set $\boldsymbol{X}^j = \boldsymbol{X}^{j-1} + 0.15\boldsymbol{Z}^j$ for $j = 2, 4, 6$, $\boldsymbol{X}^7 = \boldsymbol{X}^8 + \boldsymbol{X}^9 - \boldsymbol{X}^10 + 0.15\boldsymbol{Z}^7$ and finally $\boldsymbol{X}^{11} = \boldsymbol{X}^{14} + \boldsymbol{X}^{15} - \boldsymbol{X}^{12} - \boldsymbol{X}^{13} + 0.15\boldsymbol{Z}^{11}$. This construction leads to a correlation of around 0.998 between $\boldsymbol{X}^j$ and $\boldsymbol{X}^{j+1}$,

$i = 1, 3, 5$ and complicated linear dependencies among the groups of variables $\{\boldsymbol{X}^7, \boldsymbol{X}^8, \boldsymbol{X}^9, \boldsymbol{X}^{10}\}$ and $\{\boldsymbol{X}^{11}, \boldsymbol{X}^{12}, \boldsymbol{X}^{13}, \boldsymbol{X}^{14}, \boldsymbol{X}^{15}\}$.

Let $\boldsymbol{X}$ be the design matrix with columns $\boldsymbol{X}^j$ for $j = 1...15$. Then $\boldsymbol{x}_i^T$ is the ith row of $\boldsymbol{X}$. We denote the mean of the ith response as μ_i. We simulate 180 independent Poisson variables y_i with

$$\log(\mu_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}$$

where

$$\boldsymbol{\beta} = (0.15, 0, 0.15, 0, 0.15, 0, 0.15, -0.15, 0, 0, 0.15, 0.15, 0.15, 0, 0)^T.$$

3.6.1 Notation and Prior Distributions

As in example (3.4) define a model m by the pair $(\boldsymbol{\gamma}^m, S)$ where $\boldsymbol{\gamma}^m = (\gamma_1^m, ..., \gamma_{15}^m) \in \{0, 1\}^{15}$ is a vector indicating which of the covariates are included in model m. Since $q = 15$, and we are not interested in interactions between variables, there are $2^{15} = 32,768$ possible models. Here S dictates that

$$\log(\boldsymbol{\mu}) = \boldsymbol{X}_m \boldsymbol{\beta}_m \tag{3.18}$$

and hence

$$\log(f(\boldsymbol{y}|\boldsymbol{\beta}_m, m)) \propto \sum_i y_i(\boldsymbol{x}_i^T \boldsymbol{\beta}_m) - \sum_i \exp(\boldsymbol{x}_i^T \boldsymbol{\beta}_m). \tag{3.19}$$

We assume a-priori that

$$\boldsymbol{\beta}_m | m \sim N(\boldsymbol{0}, n(\boldsymbol{X}_m^T \hat{\boldsymbol{W}} \boldsymbol{X}_m)^{-1}), \tag{3.20}$$

where $n = \sum_i y_i$ and $\hat{\boldsymbol{W}}$ is the 180×180 diagonal matrix with ith element y_i. We assume each model is equally likely.

3.6.2 Reversible Jump and Results

Using the reversible jump algorithm problems may occur if there are strong linear dependencies among the variables. To see this suppose we have two collinear terms

A and B. The models A, B and $A+B$ may explain the data equally well, but a move from model A, or B, to model $A+B$ would be rejected since model $A+B$ lacks parsimony (This situation is easily remedied by allowing moves from model A to model B and vice-versa). These 'flip' proposals could allow simultaneous addition and subtraction of terms, which should improve standard errors and Markov chain mixing.

We compare two reversible jump schemes. In scheme A moves between nested models are proposed, selecting to add or remove a term with equal probability. In scheme B, in addition to the nested moves, we propose 'flip' moves detailed below.

3.6.2.1 Flip move

Let the current parameter values of the Markov chain be $(m, \boldsymbol{\beta}_m)$, with the dimension of $\boldsymbol{\beta}_m$ being p_m. We are to propose a move from model m to model m', where the models are only partially nested, in two stages. Let $\boldsymbol{X}$ denote the matrix composed of the columns that appear in both $\boldsymbol{X}_m$ and $\boldsymbol{X}_{m'}$. Let $\boldsymbol{S}$ and $\boldsymbol{S}'$ denote the columns unique to $\boldsymbol{X}_m$ and $\boldsymbol{X}_{m'}$ respectively.

The first stage is a birth proposal to a model denoted m^* with design matrix $\boldsymbol{X}_{m^*} = (\boldsymbol{X}|\boldsymbol{S}|\boldsymbol{S}')$. The second is a death proposal to model m'

- Propose a move to model m' with probability $j(m', m)$, where $\boldsymbol{X}_{m'} = (\boldsymbol{X}|\boldsymbol{S}')$.

Birth stage

- Generate $\boldsymbol{u}^*$ from

$$\boldsymbol{u}' \sim N((\boldsymbol{S}'^T\hat{\boldsymbol{W}}(I-\boldsymbol{P}_m)\boldsymbol{S}')^{-1}\boldsymbol{S}'^T\hat{\boldsymbol{W}}(I-\boldsymbol{P}_m)\hat{\boldsymbol{\eta}}, (\boldsymbol{S}'^T\hat{\boldsymbol{W}}(I-\boldsymbol{P}_m)\boldsymbol{S}')^{-1})$$

  which has density function denoted g^*.

- Set $\boldsymbol{\beta}^*_{m^*} = \begin{pmatrix} I & -(\boldsymbol{X}_m^T\hat{\boldsymbol{W}}\boldsymbol{X}_m)^{-1}\boldsymbol{X}_m^T\hat{\boldsymbol{W}}\boldsymbol{S}' \\ 0 & I \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_m \\ \boldsymbol{u}^* \end{pmatrix}$.

Death stage

- Partition $\boldsymbol{X}^*_m \boldsymbol{\beta}^*_m = (\boldsymbol{X}_{m'} | \boldsymbol{S}) \begin{pmatrix} \boldsymbol{\beta}_{m^*, \boldsymbol{X}_{m'}} \\ \boldsymbol{\beta}_{m^*, \boldsymbol{S}} \end{pmatrix}$.
- Set $\boldsymbol{\beta}'_{m'} = \boldsymbol{\beta}_{m^*, \boldsymbol{X}_{m'}} + (\boldsymbol{X}^T_{m'} \hat{\boldsymbol{W}} \boldsymbol{X}_{m'})^{-1} \boldsymbol{X}^T_{m'} \hat{\boldsymbol{W}} \boldsymbol{S} \boldsymbol{\beta}_{m^*, \boldsymbol{S}}$.

Combined acceptance probability

- Accept this proposed move with probability

$$\frac{f(\boldsymbol{y}|m', \boldsymbol{\beta}^p_{m'}) f(\boldsymbol{\beta}^p_{m'}|m') f(m') j(m', m) g(\boldsymbol{\beta}_{m^*, \boldsymbol{S}})}{f(\boldsymbol{y}|m, \boldsymbol{\beta}_m) f(\boldsymbol{\beta}_m|m) f(m) j(m, m') g'(\boldsymbol{u}^*)} \wedge 1, \tag{3.21}$$

  where g is the normal distribution

$$N((\boldsymbol{S}^T \hat{\boldsymbol{W}}(I - \boldsymbol{P}_{m'})\boldsymbol{S})^{-1} \boldsymbol{S}^T \hat{\boldsymbol{W}}(I - \boldsymbol{P}_{m'}) \hat{\boldsymbol{\eta}}, (\boldsymbol{S}^T \hat{\boldsymbol{W}}(I - \boldsymbol{P}_{m'})\boldsymbol{S})^{-1}).$$

- Else reject and remain at $(m, \boldsymbol{\beta}_m)$.

The above move types will be useful when collinearity exists between variables, as is the case here.

We ran the chain for a total of 50,000 iterations and observed an acceptance rate of 0.260 for 'local' moves, and a rate of 0.201 for the 'flip' moves.
Table 3.6 gives the posterior inclusion probabilities, $\hat{\gamma}_i$, together with the associated standard errors. Including flip moves reduces the standard error of the posterior inclusion probabilities and thus improves mixing of the Markov chain (all ratio's are greater than 1). We must note that the simple algorithm performed well which can be attributed to the construction of reasonable proposal distributions. These results differ to those described in George and McCulloch (1997) (The data were simulated.)

Table 3.6: $\hat{\gamma}_i$ and Monte Carlo standard errors for $\hat{\gamma}_i$ for the simulated Poisson example for sampling schemes A and B

Variable	$\hat{\gamma}_i$	SE_A	SE_B	SE_A/SE_B
$\boldsymbol{X}_1$	0.7181	0.0119	0.0105	1.1284
$\boldsymbol{X}_2$	0.3467	0.0127	0.0112	1.1326
$\boldsymbol{X}_3$	0.4466	0.0131	0.0117	1.1172
$\boldsymbol{X}_4$	0.5479	0.0132	0.0118	1.1154
$\boldsymbol{X}_5$	0.4284	0.0118	0.0106	1.1124
$\boldsymbol{X}_6$	0.4977	0.0122	0.0106	1.1485
$\boldsymbol{X}_7$	0.3521	0.0113	0.0111	1.0167
$\boldsymbol{X}_8$	0.4154	0.0113	0.0112	1.0071
$\boldsymbol{X}_9$	0.3165	0.0085	0.0085	1.0073
$\boldsymbol{X}_{10}$	0.4456	0.0104	0.0103	1.0082
$\boldsymbol{X}_{11}$	0.2488	0.0099	0.009	1.0918
$\boldsymbol{X}_{12}$	0.0883	0.0044	0.0039	1.1109
$\boldsymbol{X}_{13}$	0.1714	0.008	0.0074	1.0828
$\boldsymbol{X}_{14}$	0.2759	0.0073	0.0073	1.0076
$\boldsymbol{X}_{15}$	0.7079	0.0103	0.0093	1.1061

3.7 Example 4: Low Birth Weight in Infants

Our penultimate example considers applying the reversible jump algorithm to binary response data. Hosmer and Lemeshow (1989) provide a data set on 189 births at a U.S. hospital. The response variable is binary and corresponds to 1 for a child of low birth weight ($< 2.5kg$) and 0 otherwise. Following Venables and Ripley (1999) and Nott and Leonte (2004), we construct the variables given in Table 3.7.

Table 3.7: Low birth weight in infants data: Covariate information.

Predictor	Description
age	Age of mother in years (Centred)
lwt	Weight of mother at last menstrual cycle (Centred)
raceblack	Indicator for race (black or other (0/1))
raceother	Indicator for race other than white or black (0/1)
smoke	Smoking status during pregnancy (0/1)
ptd	Previous premature labours (0/1)
ht	History of hypertension (0/1)
ui	Has uterine irritability (0/1)
ftv1	Indicator for one physician visit in first trimester (0/1)
ftv2+	Indicator for two or more physician visits in first trimester (0/1)

3.7.1 Notation and Prior Distributions

As in example (3.4) define a model m to be the pair $(\boldsymbol{\gamma}^m, S)$, where $\boldsymbol{\gamma}^m = (\gamma_1^m, ..., \gamma_{10}^m) \in \{0, 1\}^{10}$ is a vector indicating which of the covariates are included in model m. Since $q = 10$, and we are not interested in interactions between covariates, there are $2^{10} = 1024$ possible models. Here S implies that

$$\log(\frac{\mu_i}{1 - \mu_i}) = \eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$$

with corresponding likelihood given by

$$f(\boldsymbol{y}|\boldsymbol{\beta}) \propto \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}.$$

As in previous sections we are required to find an approximation for $\hat{\boldsymbol{W}}$ and derive an initial linear predictor $\hat{\boldsymbol{\eta}}$ to use in proposal distributions. We note

$$w_i = \frac{(\frac{\partial \pi_i}{\partial \eta_i})^2}{Var(y_i)} = \pi_i(1-\pi_i),$$

which we approximate using

$$\hat{\mu}_i = \frac{\exp(\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}})}{1+\exp(\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}})}$$

obtained by fitting the logistic regression model containing an intercept and all 10 variables given in Table 3.7. We use $\hat{\boldsymbol{\eta}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ in proposal distributions, obtained by fitting the same logistic regression. The saturated model cannot be used since the data is binary. Clearly the success of the algorithm depends critically upon this choice. We advocate basing any approximation on the model containing all terms and interactions under consideration.

As in previous sections we assume all models *a-priori* equally likely. We assume that, for each model m, $\boldsymbol{\beta}$ has a multivariate normal distribution with zero mean vector, and covariance matrix assumed to be the identity matrix multiplied by $\sigma^2 = 2.5^2$.

3.7.2 Reversible Jump and Results

We implemented the reversible jump scheme detailed in Section 3.3. We proposed moves between nested models. In total, 195 different models were visited during 50,000 iterations of the algorithm. The Markov chain proved to be mobile with an acceptance probability of 0.18. Table 3.8 gives posterior inclusion probabilities and

Table 3.8: $\hat{\gamma}_i$ and Monte Carlo standard errors for $\hat{\gamma}_i$ for the low birth weight in infants data.

Predictor	$\hat{\gamma}_i$	Standard Error
age	0.041	0.0015
lwt	0.022	0.0011
raceblack	0.067	0.0018
raceother	0.054	0.0018
smoke	0.081	0.0020
ptd	0.862	0.0026
ht	0.253	0.0030
ui	0.128	0.0024
ftv1	0.115	0.0033
ftv2+	0.20	0.0011

Monte Carlo standard errors. Figure 3.2 provides Markov chain diagnostic plots. There is no suggestion that the chain has not converged to its stationary distribution. These results differ from those described in Nott and Leonte (2004) as a result of the prior distribution placed on the model space. Nott and Leonte (2004) assume that each term is independently included in the model with probability ρ. That is

$$f(\gamma|\rho) \propto \rho^{p_m}(1-\rho)^{q-p_m}.$$

This prior includes the uniform distribution over all models ($\rho = 1/2$). In the case of variable selection this uniform prior distribution induces a Binomial distribution on p_m, with prior expectation that $q/2$ terms will be included in the model. Nott and Leonte (2004) suggest that if we wish to control the expected number of terms within a model we assume

$$f(p_m|\rho) = \binom{q}{p_m}\rho^{p_m}(1-\rho)^{q-p_m}$$

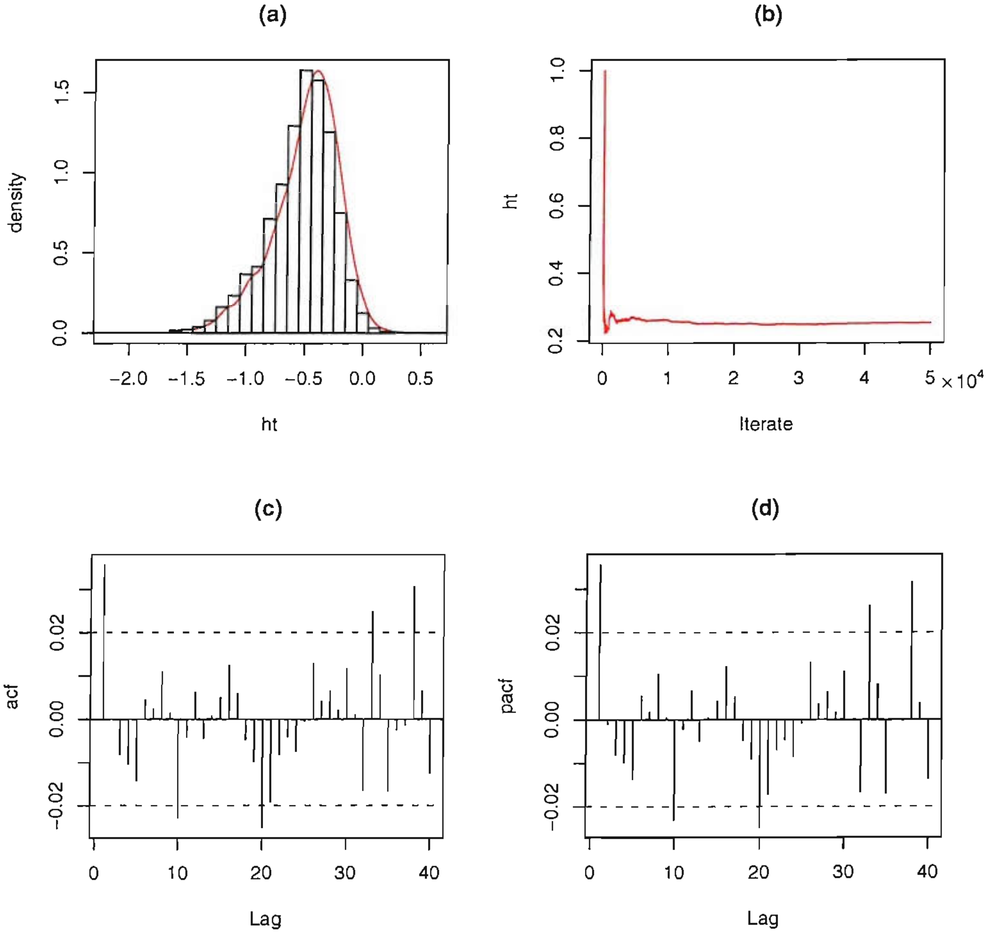


Figure 3.2: MCMC output plots. Plot (a): Histogram of the posterior distribution for the 'history of hypertension' parameter. Plot (b): A running mean of the inclusion probability for the 'history of hypertension' parameter. Plot (c): Auto-correllelogram of inclusion indicator for the 'history of hypertension' parameter. Plot (d): Partial auto-correllelogram of inclusion indicator for the 'history of hypertension' parameter.

and place the following hyper-prior distribution on ρ

$$f(\rho) = \frac{\rho^{a-1}(1-\rho)^{b-1}}{B(a,b)},$$

where $B(.,.)$ is the beta function. The prior distribution of $f(p_m)$ is easily obtained and given as follows:

$$f(p_m) = \int f(p_m|\rho)f(\rho)d\rho = \binom{q}{p_m}\frac{B(p_m + a, q - p_m + b)}{B(a,b)}. \tag{3.22}$$

Parameters a and b (both strictly positive reals) are chosen to satisfy the following two equations:

$$\begin{aligned} \mathbb{E}[p_m] &= q\mathbb{E}[\rho] \\ var(p_m) &= var[\mathbb{E}[p_m|\rho]] + \mathbb{E}[var[p_m|\rho]] \\ &= q(q-1)\mathbb{E}[\rho^2] + q\mathbb{E}[\rho](1 - q\mathbb{E}[\rho]). \end{aligned} \tag{3.23}$$

Since ρ has a Beta distribution it is easily seen that

$$\begin{aligned} \mathbb{E}[\rho] &= \frac{a}{a+b} \\ \mathbb{E}[\rho^2] &= \mathbb{E}[\rho]\frac{a+1}{a+b+1}. \end{aligned} \tag{3.24}$$

Replacing (3.24) in (3.23) we obtain a and b by solving the following two simultaneous equations:

$$\begin{aligned} \mathbb{E}[p_m] &= q\mathbb{E}[\rho] = \frac{aq}{a+b} \\ var[p_m] &= \frac{a+1}{a+b+1}q(q-1)\mathbb{E}[\rho] + q\mathbb{E}[\rho](1 - q\mathbb{E}[\rho]) \end{aligned} \tag{3.25}$$

Nott and Leonte (2004) set $\mathbb{E}[p_m] = 8$ and $var[p_m] = 16$. For $q = 10$ there are no solutions to (3.25) that satisfy the constraints $a, b \geq 0$, and hence no beta-binomial distribution exists. However, the following distribution for p_m satisfies $\mathbb{E}[p_m] = 8$ and $var[p_m] = 16$

$$f(p_m) = \begin{cases} 0.2 \text{ if } p_m = 0 \\ 0.8 \text{ if } p_m = 10 \\ 0 \text{ otherwise }. \end{cases}$$

It is not a beta-binomial distribution. It is conceivable that Nott and Leonte (2004) made an error in their calculations. They state that the prior distribution placed upon the model space is bimodal, and conclude the model with largest posterior probability (assuming this bimodal prior distribution) contains only a term for the intercept. This is hardly surprising given the prior distribution. This demonstrate the critical issue of prior model probabilities.

Nott and Leonte (2004) conclude their algorithm is inexpensive, which is true if $f(p_m|\boldsymbol{y})$ places considerable mass on the null model. In comparison to the reversible jump algorithm presented within this chapter the algorithm of Nott and Leonte (2004) requires a significant amount of expert knowledge to set algorithmic parameters. The algorithm is difficult to generalise to model (as well as variable) selection problems.

We find, assuming a uniform prior on the model space, the model with largest posterior probability to contain an intercept term and a single coefficient for the parameter ptd (there is considerable model uncertainty however). As already stated, the reversible jump algorithm presented in this chapter is simple, easy and efficient to use. No pilot runs or tuning of the Markov chain is needed, and no expert knowledge is required to implement this algorithm.

3.8 Example 5: What Influences Political Attitude?

Our final example is taken from Wermuth and Cox (1998) and concerns responses from two surveys taken in 1991 and 1992 in Germany. The counts are reproduced, in the form of a $4\times5\times5\times2\times2$ contingency table, in Appendix E. The variables denoted A, B, C, D and E are also given in this appendix. In total 6039 individuals responded to the two surveys, of which, we randomly sample 1000 and form a new contingency table. If all 6039 data points are included, the model $AD+AE+BC+BE+DE$ has posterior probability close to 1. To obtain this estimate we simulated the reversible jump algorithm for a wide range of starting models and model parameters. This demonstrate the difficulty of assessing convergence as the model space is vast and alternative models are visited infrequently.

The motivation of this example is to show that, using our algorithm, large dimensional moves are plausible and possible.
We use equivalent notation and prior distributions to that in example (3.4) and set $\alpha_a^2 = 2|I| = 800$. All models under consideration include the terms A, B, C, D and E.

3.8.1 Reversible Jump and Results

We implemented the reversible jump scheme detailed in (3.3). We proposed moves between nested models, attempting a 'birth move' or 'death move' with equal probability. To aid mixing, parameters of the current model were updated every third iterate. The initial state Markov chain was the model containing terms A, B, C, D and E, and maximum likelihood estimates under this model used as starting values for $\boldsymbol{\beta}_m$. On this occasion, and since we would not expect this model to have posterior support, we allowed a 'burn in' period. This is not our usual practice. We believe that, if possible, a Markov chain should be started from a state we would happily include in any analysis.
After this initial burn in we obtained 500,000 iterates for inference. In all 9 models were visited with one in every 25 proposed moves being accepted. Table 3.9 provides estimates of the posterior model probabilities.
Wermuth and Cox (1998) provide further information concerning the variables in-

Table 3.9: Posterior model probabilities for political attitudes data

Model	Posterior Probability
AD+AE+BC+BE+DE	0.524638
AD+AE+BC+DE	0.474608
ADE+BC+BE	0.000258
AD+AE+BC+BE+CE+DE	0.000235
ADE+BC	0.000192
BCE+AD+AE+DE	0.000032
AD+AE+BC+CE+DE	0.000028
BCE+AB+AD+AE+DE	0.000006
AE+BC+BE+DE	0.000002

volved. When the two separate states where formed different school systems were established in each state. Therefore, prior to their analysis, they suspect a strong BCE interaction. Furthermore, they reason one might expect $B, C \perp\!\!\!\perp D$ given E. If we used all 6039 individuals in our analysis, the model $BCE + AD + AE + DE$ has a posterior probability of approximately 1. This model contains a strong BCE interaction in addition to $B, C \perp\!\!\!\perp D|E$. There is strong posterior evidence that $B, C \perp\!\!\!\perp D|E$ based upon our sample of 1000 individuals. This example was chosen to really test our approach and the results seem promising. The interaction BCE, a 16 dimensional interaction, appears in Table 3.9. Had we based our proposal distributions on pilot runs this interaction would almost certainly not have been observed.

3.9 Possible Problems with RJMCMC

It would not be wise to draw the chapter to a close without looking at possible pitfalls one might experience when implementing this technique. In the next two sections we consider potential dangers and offer some advice for their resolution.

It is essential to check the fit of a model to data. We propose to do this dynamically (within chain). We suggest posterior predictive model checking, an approach described in Gelman and Meng (1996).

Denoting $\boldsymbol{y}$ as observed data and $\boldsymbol{\beta}_m$ unknown parameters of interest given model m, then $f(\boldsymbol{y}|\boldsymbol{\beta}_m, m)$ is the likelihood under model m with $f(\boldsymbol{\beta}_m|\boldsymbol{y}, m)$ the posterior distribution of the parameters given the data. The reversible jump Markov chain produces a sample of size n, denoted $\boldsymbol{\beta}_m^1, ..., \boldsymbol{\beta}_m^n$, from this posterior distribution. For each $i = 1, ..., n$ simulate a hypothetical replication of the data, denoted $\boldsymbol{y}_i^r$, from the sampling distribution given the parameters $\boldsymbol{\beta}_m^i$. If the model is adequate the hypothetical replication should look similar to the observed data $\boldsymbol{y}$. This can be formally written down as follows. Select a *discrepancy variable* $T(\boldsymbol{y}, \boldsymbol{\beta}_m)$ which will have an extreme value if the data $\boldsymbol{y}$ are in conflict with the model, i.e.

$$T(\boldsymbol{y}_i^r, \boldsymbol{\beta}_m^i) \geq T(\boldsymbol{y}, \boldsymbol{\beta}_m). \tag{3.26}$$

If this variable is extreme for many $i = 1, ..., n$ we might conclude that model m does not adequately describe the data. The discrepancy variable can be any function of the data and parameters, and choices for this are discussed in Gelman et al. (1996). The advantage of assessing predictive performance using this method is that the above inequality is easily verified within the Markov chain, and with little additional expense. We would obviously not verify the inequality at every iterate.

Poor prediction could result if assumptions underlying the likelihood (normality of errors and heteroscadicity in the linear model case) are not justified. These assumptions are easily verified. We would also question results if output from the

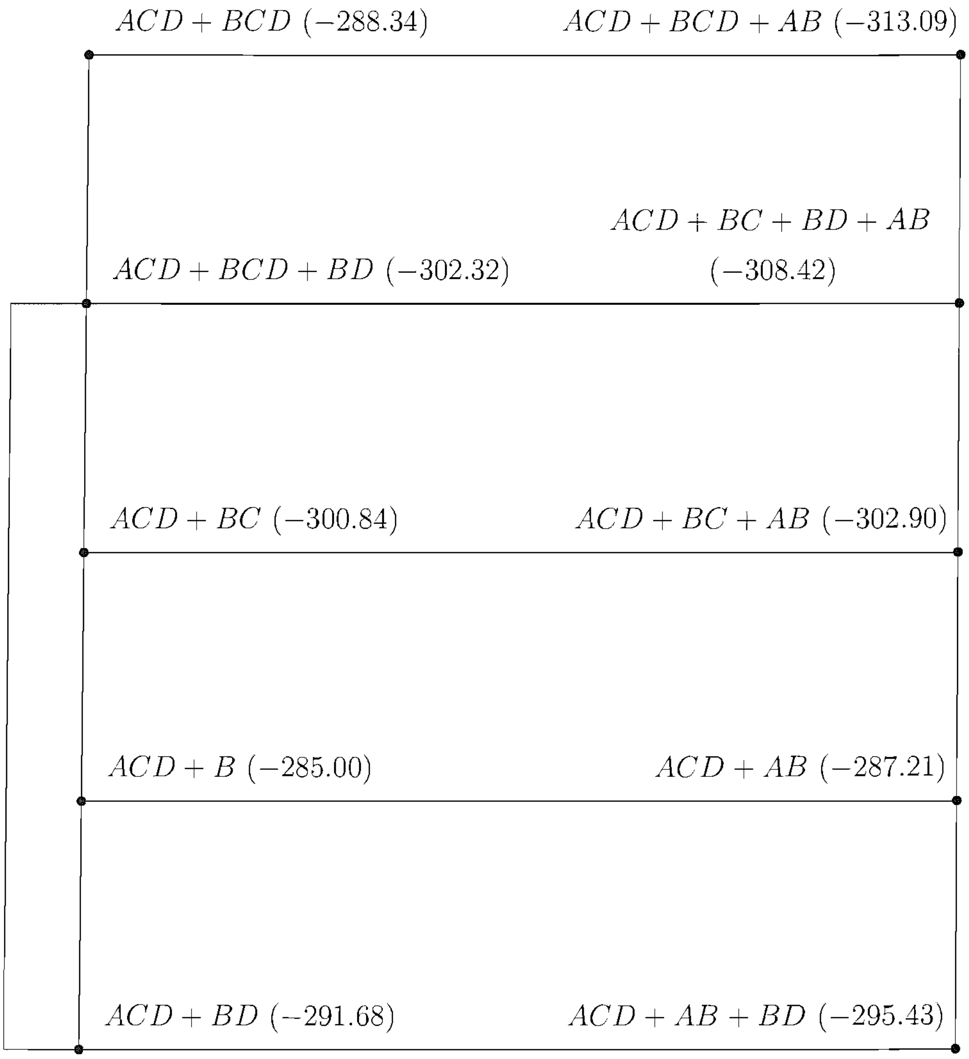
Markov chain indicated a particular model had reasonable posterior support but failed to predict the data well. If such a scenario occurred, we might question convergence of the Markov chain.

A real concern is that the Markov chain is 'stranded' in some part of the model space, failing to explore an unconnected subspace containing models with good posterior support.
As an example consider the simulated data, in the form of a $4 \times 5 \times 5 \times 2$ contingency table, given in Appendix F. For this data set and for selected models Shwartz criterion (see (2.12)) have been calculated. These are given in Figure 3.3. On the diagram a line between two models indicates nesting. We see that three models ($ACD + B$, $ACD + AB$ and $ACD + BCD$) have some posterior support but that there is no path, through nested models of posterior support, from $ACD + B$ or $ACD + AB$ to $ACD + BCD$. Therefore, any Markov chain using local moves alone will not explore the full model space. There are several potential remedies for this situation.

One possible remedy is to generalise flip proposals outlined in Section 3.6. Suppose the current state of the Markov chain is model m. Form a transition kernel K in the following way. Generate $n_1, n_2 \sim Poisson(\lambda)$ where n_1 is the number of birth moves to be proposed and n_2 the number of death moves. Propose the first birth move, conditional on m, from a randomly sampled transition kernel B_1. If $n_1 > 1$ a second 'birth' is proposed from a randomly sampled transition kernel B_2 conditional on the new model m_{B_1} and so on. Continue in this fashion, including death moves, to form kernel K given by

$$K = B_1 ... B_{n_1} D_1 ... D_{n_2}.$$

Figure 3.3: Schwartz criteria for the generated bimodal data (Appendix F).



$ACD + BCD$ (−288.34)
$ACD + BCD + AB$ (−313.09)
$ACD + BC + BD + AB$ (−308.42)
$ACD + BCD + BD$ (−302.32)
$ACD + BC$ (−300.84)
$ACD + BC + AB$ (−302.90)
$ACD + B$ (−285.00)
$ACD + AB$ (−287.21)
$ACD + BD$ (−291.68)
$ACD + AB + BD$ (−295.43)

The value of λ should be chosen so that with reasonable probability we have $(n_1, n_2) = (0, 1)$ or $(n_1, n_2) = (1, 0)$. These moves are formed of a single birth or death kernel and correspond to the local moves described in (3.2). Parameters are updated if $(n_1, n_2) = (0, 0)$, and a non-deterministic death move could be proposed if $(n_1, n_2) = (1, 2)$. Constructing a transition kernel in this fashion may improve mixing of the Markov chain, as in (3.6), along with overcoming the issue of bimodality. Again, the computational simplicity of both birth and death moves presented within this chapter, results in a simplistic form of the kernel $\mathcal{K}$. The acceptance probability of such moves would also be simple and cheap to calculate.

By far the most appealing solution to the multi-modality problem is tempering. Here the stationary distribution of the Markov chain is modified to facilitate between model moves. Of particular interest is Metropolis-coupled MCMC (Geyer, 1991). Geyer (1991) proposed running in parallel n different Markov chains with different stationary distributions f_i for $i = 1, ..., n$. We set $f_1 = f$ where f is the distribution of interest. For $i > 1$ we set $\pi_i = \pi^{1/t_i}$ where t_i is called the temperature. A value of t_i greater than 1 will flatten the distribution of f . We would therefore expect samplers with $t_i > 1$ to be more mobile. A Metropolis-coupled reversible jump MCMC algorithm will have 3 different updates. Firstly *within* each Markov chain we could choose to update current model specific parameter values. Secondly *within* each Markov chain we can propose a move to a new model using the birth and death steps already described. In addition to these two standard updates, MCRJMCMC, proposes 'switches' *between* Markov chains. It is hoped these switches will allow the modified sampler to explore the full support of π, and could overcome the issue of multi-modality. Suppose at iteration t a swap between chains i and j is proposed. Let current state of these chains be $(m_i, \boldsymbol{\beta}_{m_i})$ and $(m_j, \boldsymbol{\beta}_{m_j})$ respectively. Then this proposed move is accepted with the probability

$$\frac{f(m_j, \boldsymbol{\beta}_{m_j}|\boldsymbol{y})^{1/t_i} f(m_i, \boldsymbol{\beta}_{m_i}|\boldsymbol{y})^{1/t_j}}{f(m_i, \boldsymbol{\beta}_{m_i}|\boldsymbol{y})^{1/t_i} f(m_j, \boldsymbol{\beta}_{m_j}|\boldsymbol{y})^{1/t_j}} \wedge 1. \tag{3.27}$$

Posterior summaries are based solely upon the chain with the correct stationary distribution. An obvious disadvantage to this scheme is the additional computational burden of running n Markov chains. In practice n need only be 3 or 4. When selecting temperatures there is a clear trade-off between accepting moves within a chain and accepting moves between a chain. The point is not to design a sampler that will accept a move to a sample point with no posterior support, but to aid the sampler to 'discover' areas of potential support. The algorithm is designed to explore bimodality in the joint $(m, \boldsymbol{\beta}_m)$ given the data. We are interested in uncertainty concerning $m|\boldsymbol{y}$, and we must therefore meet the additional expense in exploring potential, but unlikely, multi-modality in $\boldsymbol{\beta}_m|m, \boldsymbol{y}$.

As an example, we applied the MCRJMCMC algorithm to the generated data given in Appendix F. Three chains were run in parallel with temperatures (1,2,100). A switch between chains was proposed with probability 1/5. At each iteration one of the three chains was selected with equal probability. Current parameters of this chain were updated, using a Metropolis-Hastings step, and then with probability 1/2 a birth or death move proposed. All chains started at the model containing the terms A, B, C and D alone. The chain was run for a total of 2,000,000 iterations, taking several hours to complete. As we would expect, within chain acceptance probabilities of birth and death moves depended upon the chains temperature. These acceptance probabilities are given in Table 3.10. A probability on the diagonal represents the within chain acceptance probability. It is the probability of a birth, or death move, being accepted. As expected, we observed the greatest acceptance probability in chain three with $t_3 = 100$. The off diagonal probabilities are those of between chain moves. Moves between chains 1 and 2 occurred one in every hundred and fifty attempts. Moves between chains 2 and 3 were also frequent, whilst moves between chains 1 and 3 occurred rarely.

Table 3.10: Metropolis coupled reversible jump MCMC acceptance probabilities.

	Chain 1	Chain 2	Chain 3
Chain 1	0.001	0.0066	0.00001
Chain 2	0.0066	0.005	0.0006
Chain 3	0.00001	0.0006	0.382

Table 3.11: Posterior model probabilities for simulated bimodal data

Model	Counts based on 2,000,000 iterates
ACD+B	1,551,770
ACD+AB	245,853
AB+AC+AD	684
ACD+BD	600
ACD+BCD	515
AB+AC+D	463
ACD+AB+BD	102
AB+C+D	10
A+B+C+D	2

Table 3.11 gives posterior model counts based upon a sample of 2,000,000 iterations. We see that the full model space has been explored. Clearly more work is needed to enable good selection of the temperatures. Ideally, temperatures would be selected to facilitate mixing but with minimal computational expense. We must note that the above algorithm is expensive to run and could be described as 'brutish'. One

would hope, given time, a more natural and elegant solution can be found.

3.10 Closing Remarks

In this chapter we have looked at the reversible jump Markov chain Monte Carlo algorithm. In particular we have derived an efficient and practical construction of this algorithm for model determination in generalised linear models. Minimal expert knowledge is required for the algorithms implementation. We are not required to run computationally expensive pilot chains to construct proposal distributions, nor are we required to fine tune these distributions.

The scheme has successfully been applied to covariate selection in linear, log-linear and logistic regressions. We have also applied the scheme to model determination in log-linear models. We have shown the algorithm to be efficient. We present a further example of the procedure in the following chapter.

We must also note that for a generalised linear model the parameterisation plays a large part in accurate estimation. We have not considered the parameterisation of generalised linear models in this thesis and the reader is therefore referred to Gelfand et al. (1996).

The scheme has enormous scope for future work, and this is discussed in the final chapter.

Chapter 4

Analysis of Incomplete Contingency Tables

The problem of analysing data sets from which observations are missing is common. There are many different reasons for missing these observations, for example measurement error or non-response, but in all situations the statistical goal remains the same. How should one approach inference that accommodates the possible, but unknown, behaviour of the unobserved data?

In the literature review we stated the goal of our analysis - to make valid, formal inference that incorporates all forms of uncertainty, in particular uncertainty resulting from missing observations.

We reviewed the four basic approaches to handling missing data now adopted in the statistical literature. These were weight, ignore (give zero weight to missing observations), impute and model. We argued that we should not proceed with a complete or available case analysis, and although weighting missing observations often reduces the bias of an estimate, it does so at the cost of increase in mean squared error. Imputing missing observations has also been the subject of increasing criticism, often underestimating standard errors and overstating statistical significance. In fact imputation was designed for an entirely different situation, albeit a situation involving missing data (Rubin, 1996).

If we merely cast these methods aside we are left solely with the idea of modeling the missing data. This idea has two real advantages making it an attractive option. Firstly, modeling may provide information about the missing-data mechanism. Secondly, we may be able to ascertain how assumptions about this missing data mechanism affect inference. These advantages come at a cost of increased difficulty and complexity.

This chapter proceeds as follows. Data from the Slovenian plebiscite is introduced in Section 4.1. This data set is used to illustrate basic concepts, ideas and difficulties of analysing incomplete contingency tables.

In Section 4.2 we discuss the approach of modeling non-response and form parametric models for non-response mechanisms. In Section 4.3 we attempt to discriminate between 'competing' non-response models. We use techniques developed in Chapter 3. In particular, we attempt to apply the reversible jump algorithm for model determination in light of missing observations.

We develop a sensitivity analysis, Section 4.4, before closing with discussion and concluding remarks in Section 4.6.

4.1 Introduction - The Slovenian Plebiscite

On June 25th 1991 the Slovenian assembly passed the Fundamental Sovereignty act and proclaimed independence from Yugoslavia, the first republic of the federation to do so. In response to Slovenia's proclamation the Yugoslav Army sent in tanks commencing the 10-day war. By October of the same year Slovenia was declared an independent republic, which was recognised internationally in December after the passing of its first constitution. Many other former Yugoslav states have since declared independence.

A critical step in the independence process was the plebiscite held exactly one year before the constitution was adopted. To prepare for the results of the plebiscite the government of Slovenia conducted a survey to ascertain public opinion. We focus our attention on three key questions asked during this survey. These three questions are given below.

1. Are you in favour of Slovenian independence? (Y_1)
2. Will you attend the Plebiscite? (Y_2)
3. Are you in favour of Slovenia's secession from Yugoslavia? (Y_3)

Full details of the survey can be found in Rubin et al. (1995). Data for the 2074 respondents can be found in Table 4.1.

Table 4.1: Data from the Slovenian plebiscite - 3 questions

Secession	Attendance	Independence		
		Yes	No	Don't Know
Yes	Yes	1191	8	21
	No	8	0	4
	Don't Know	107	3	9
No	Yes	158	68	29
	No	7	14	3
	Don't Know	18	43	31
Don't Know	Yes	90	2	109
	No	1	2	25
	Don't Know	19	8	96

Our primary interest is in estimating the proportion of the population planning to attend and vote in favour of independence. For simplicity we collapse across Y_3 yielding Table 4.2.

Table 4.2: Data from the Slovenian plebiscite - 2 questions

	Independence		
Attendance	Yes	No	Don't Know
yes	1439	78	159
no	16	16	32
Don't Know	144	54	136

The plebiscite required the entire population to vote, with non attendance registered as a 'no' vote. Don't Know responses can therefore be thought of as missing concealing the intended future behaviour of the voter. Of the 2074 respondents, 1549 (74.7%) provided answers to both Y_1 and Y_2, 198 (9.5%) provided an answer solely to Y_1, 191 (9.2%) to Y_2 and 136 (6.6%) to neither question. The observed data denoted $\boldsymbol{n}_{obs}$ corresponds to the 2×2 table of 1549 fully classified individuals, together with supplemental margins for both Y_1 and Y_2, and a supplemental margin for Y_1Y_2 given in Table 4.2.

We consider the mechanism by which some observations are missing as a random variable. We therefore introduce the following indicator variable

$$R_i = \begin{cases} 1 & \text{if } y_i \text{ is observed} \\ 2 & \text{if } y_i \text{ is not observed.} \end{cases}$$

The full, but unobserved data, $\boldsymbol{Y}$ and $\boldsymbol{R}$ can be thought of as a 2^4 contingency table with cell counts $\boldsymbol{n} = \{n_{ijkl}, i, j, k, l \in \{1, 2\}\}$, where n_{ijkl} is the count for the cell with $Y_1 = i, Y_2 = y, R_1 = k, R_2 = l$. Let

$$n_{+jkl} = \sum_i n_{ijkl}.$$

Hence n_{i+12}, n_{+j21} and n_{++22} are the counts for the partially observed data corresponding to the supplementary margins. Our interest is the joint distribution of $\boldsymbol{R}$ and $\boldsymbol{Y} = (Y_1, Y_2)$. Specifically, denoting π_{ijkl} the probability $P(Y_1 = i, Y_2 = y, R_1 = k, R_2 = l)$, then the quantity of interest is

$$\pi_{11++} = \sum_{kl} \pi_{11kl}.$$

Clearly any estimate of π_{11++} will depend critically upon assumptions made concerning the 'Don't Know' responses. If we assume pessimistically that all 'Don't Know' responses are ways to avoid revealing an unpopular decision, i.e. voting against independence, then the corresponding estimate for π_{11++} is $1439/2074 = 0.694$. In contrast, an optimistic estimate for π_{11++} is 0.905. These simple calculations provide a crude optimistic/pessimistic range of $(0.694, 0.905)$. The estimate of π_{11++} based upon all 1549 complete cases lies outside this range $(1439/1549 = 0.929)$. At the plebiscite 88.5% of eligible Slovenians explicitly voted for independence.

4.2 Handling 'Don't Know' Survey Responses

Clearly the crude optimistic/pessimistic range is wide, and any assumption concerning non-response can lead to vastly different estimates of the quantity of interest π_{11++}. This raises the following issues: How should we treat the non-responses? Furthermore, what assumptions of non-response are verifiable? To begin, we describe different types of non-response mechanisms as classified by Little and Rubin (1987). These non-response mechanisms are listed below:

- Missing Completely At Random (MCAR): Here the response variable $\boldsymbol{R}$ is independent of all other variables in the survey including covariates if observed.

- Missing At Random (MAR): Here the response indicator $\boldsymbol{R}$ can depend only upon observed values.

- 'Non-Ignorable' response models (NI): Here $\boldsymbol{R}$ is allowed to depend upon unobserved values.

A model which assumes the missing data mechanism is MAR or MCAR is termed 'ignorable' if the parameters for the missing data mechanism are distinct from those of the data model. This is not a necessary condition, see Lu and Copas (2004). In a Bayesian context, a model which assumes the missing data mechanism is MAR or MCAR is termed 'ignorable', if the parameters for the missing data mechanism are *a-priori* distinct from those of the data model.
To distinguish between the MAR and MCAR non-response mechanisms, suppose we also observed covariates $\boldsymbol{X}$. Then MCAR specifies that $\boldsymbol{R} \perp \{\boldsymbol{X}, \boldsymbol{Y}\}$ ($\boldsymbol{R}$ independent of both $\boldsymbol{X}$ and $\boldsymbol{Y}$), but MAR specifies $\boldsymbol{R} \perp\!\!\!\perp \boldsymbol{Y} | \boldsymbol{X}$ ($\boldsymbol{R}$ conditionally independent of $\boldsymbol{Y}$ given $\boldsymbol{X}$).

4.2.1 The Missing at Random Model

The most widely used assumption about the response mechanism is that of MAR, Rubin (1976). Under this assumption the probability a response variable is observed can depend only upon those other variables which have also been observed. The missing at random model is not a log-linear model. In the case of the Slovenian Plebiscite, and for all 2×2 contingency tables with supplementary margins we form a parametric model corresponding to MAR as follows. Denote

$$\begin{aligned}
&P(Y_1, Y_2) = \pi_{ij} \\
&P(R_1 = 1, R_2 = 2 | Y_1 = i, Y_2 = j) = p_i \\
&P(R_1 = 2, R_2 = 1 | Y_1 = i, Y_2 = j) = q_j \\
&P(R_1 = 2, R_2 = 2 | Y_1 = i, Y_2 = j) = r \\
&P(R_1 = 1, R_2 = 1 | Y_1 = i, Y_2 = j) = 1 - p_i - q_j - r.
\end{aligned}$$

The likelihood of the MAR model has the following parametric form:

$$\begin{aligned} f(\boldsymbol{n}) &= \left[\prod_{ij} \pi_{ij}^{n_{ij}} \prod_{i} \pi_{i+}^{n_{i+}} \prod_{j} \pi_{+j}^{n_{i+21}}\right] \\ &\times \left[r^{n_{++22}} \prod_{ij} (1-p_i-q_j-r)^{n_{ij11}} \prod_{i} p_i^{n_{i+12}} \prod_{j} q_j^{n_{i+21}}\right]. \qquad (4.1) \end{aligned}$$

The MAR model is a saturated model for the observed data since there are as many parameters as data (saturated models may have more parameters than data). There are four parameters for the complete data $(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22})$, and five parameters for the missing data mechanism (p_1, p_2, q_1, q_2, r). If $p_i = p$ for $i = 1, 2$ and $q_j = q$ for $j = 1, 2$ then $\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{R}$ and the model is MCAR.

It is not possible to directly maximise the above likelihood but effective computational methods for handling missing data under this assumption have been developed (the EM algorithm Dempster et al. (1977) or SEM algorithm of Meng and Rubin (1991)). Using these methods we obtain the maximum likelihood estimate, denoted $\hat{\pi}_{11++}^{MLE}$, of 0.892. A 95% confidence interval for π_{11++} is $(0.887, 0.896)$. These results are identical to those in Rubin et al. (1995).

An interesting feature concerning likelihood (4.1) is, since $\sum_{ij} \pi_{ij22} = 1$, individuals missing on both margins do not contribute to the estimate of π_{11++}. Furthermore, the probability of missingness on Y_2 where Y_1 is provided is allowed to depend on Y_2, and vice versa. However, the probability of missingness on both Y_1 and Y_2 is not allowed to depend on Y_1 or Y_2.

4.2.1.1 Bayesian Estimation of MAR Using MCMC

The estimation of π_{11++} in a Bayesian setting is also straightforward. An estimate of π_{11++} can be obtained using a data augmentation Markov chain Monte Carlo algorithm. Since our primary interest is the data generating process, which we assume *a-priori* independent from the missing data mechanism, we need only assume a prior distribution for $P(Y_1, Y_2)$ to estimate π_{11++}.

Prior to observing data $\boldsymbol{n}_{obs}$, we assume that $P(Y_1, Y_2)$ has a Dirichlet distribution denoted $\mathcal{D}(\alpha_{11}, \alpha_{12}, \alpha_{21}, \alpha_{22})$. Possible non informative prior distributions for $P(Y_1, Y_2)$ are the symmetric Dirichlet distributions suggested by Forster and Smith (1998). This family of prior distributions divides a prior counts evenly across the 4 cells. Setting $a = 1, 2$ and 4, hence $\alpha_{ij} = 1/4, 1/2$ and 1 $\forall i, j$, gives respectively Perks prior (Perks, 1946), Jeffreys prior (Jeffreys, 1967) and the 'uniform' prior distributions. As a result marginal distributions for individual cell probabilities are beta distributions with parameters $(1, 3)$, $(1/2, 3/2)$ and $(1/4, 3/4)$ for Perks, Jeffreys and the uniform prior respectively. These marginal beta distributions are not symmetric and are positively skewed.

We proceed by treating the missing cell counts as parameters to be estimated. We sample in turn from the conditional distribution of these missing counts given the current response parameters, then the response parameters given the current augmented cell counts. That is

Stage 1: Update $\boldsymbol{n}$

$$\begin{aligned} n_{1i12} &\sim Multinomial(n_{1+12}, (\pi_{11++}, \pi_{12++})) \\ n_{2i12} &\sim Multinomial(n_{2+12}, (\pi_{21++}, \pi_{22++})) \\ n_{j121} &\sim Multinomial(n_{+121}, (\pi_{11++}, \pi_{21++})) \\ n_{j221} &\sim Multinomial(n_{+221}, (\pi_{12++}, \pi_{22++})) \end{aligned}$$

We obtain augmented cell counts $\{n^*_{ijkl}\}$ for all cells where $(k, l) \neq (1, 1)$ which, together with counts $\{n_{ij11}\}$, form a 2^4 contingency table.

Stage 2: Update π

The posterior distribution of $P(Y_1, Y_2)$, given $\boldsymbol{n}^*$, is then the following tractable Dirichlet distribution:

$$P(Y_1, Y_2) \sim \mathcal{D}(\alpha_{11} + \sum_{kl} n^*_{11kl}, \alpha_{12} + \sum_{kl} n^*_{12kl}, \alpha_{21} + \sum_{kl} n^*_{21kl}, \alpha_{22} + \sum_{kl} n^*_{22kl}).$$

By repeating, in turn, stages 1 and 2 we obtain a sample from the joint posterior of $\boldsymbol{\pi}$ and missing cell counts given the observed cell counts. Table 4.3 gives the posterior means and 95% credible intervals for all three prior distributions. Figure 4.1 provides a plot of the posterior density for π_{11++} under the three prior distributions, together with MCMC output plots indicating good convergence to the required posterior distribution. All results are based upon a generated sample of 25000 observations.

Table 4.3: Posterior means and 95% credible intervals for Slovenian Plebiscite data under Missing at Random model.

	Uniform Prior	Jeffreys Prior	Perks Prior
Posterior Mean	0.886	0.889	0.891
2.5 percentile	0.872	0.875	0.876
97.5 percentile	0.900	0.903	0.905

The posterior distributions are almost identical for all three prior distributions. All three credible intervals contain the maximum likelihood estimate, and posterior means are close to the actual result of the Plebiscite.

4.2.2 Non-Ignorable Non-Response and Log-Linear Models

A non-ignorable non-response model is one where the probability of an observations missingness is allowed to depend on the values of unobserved variables. That is R_i is permitted to depend on Y_i. Using non-ignorable models we may be able to ascertain how assumptions about the response mechanism affect inference for π_{11++}. In the

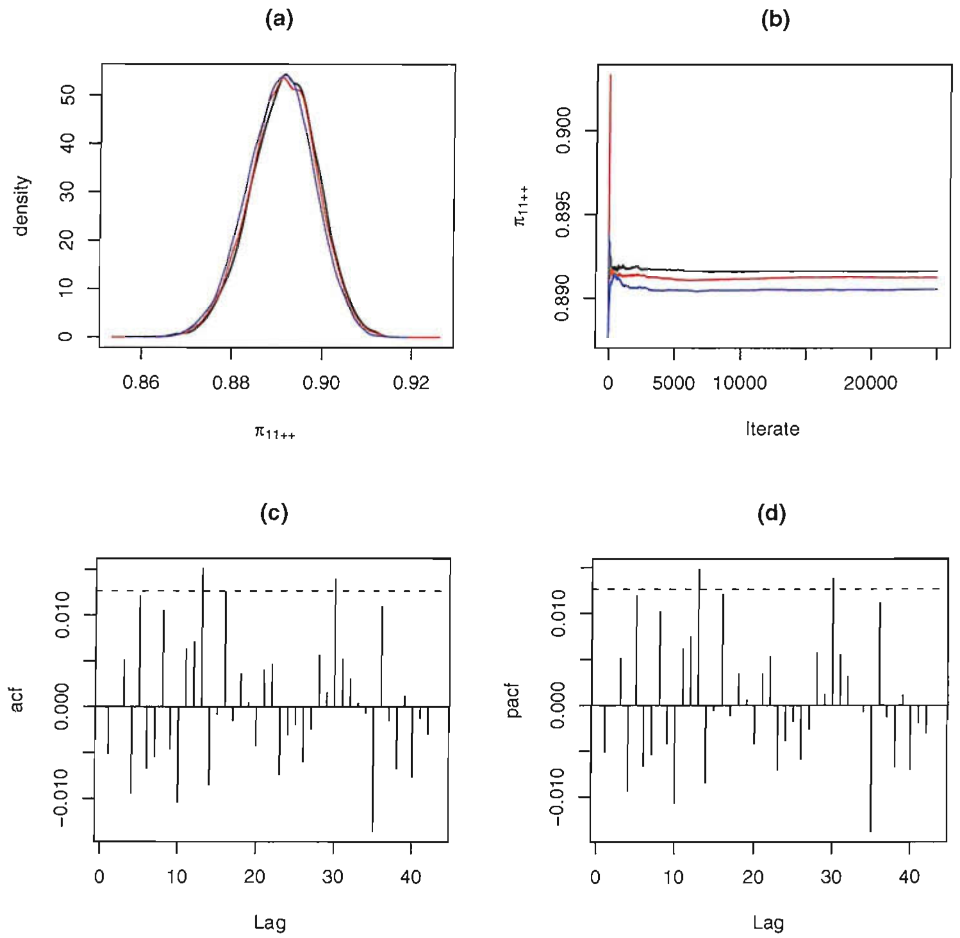


Figure 4.1: MCMC output plots. Plot (a): Posterior density of π_{11++} under MAR: — uniform prior distribution, — Perks prior distribution and — Jeffreys' prior distribution. Plot (b): A running estimate of the posterior mean for the distribution of π_{11++} under MAR: — uniform prior distribution, — Perks prior distribution and — Jeffreys' prior distribution. Plot (c): Auto-correllelogram of π_{11++} Markov chain iterates under the uniform distribution. Plot (d): Partial auto-correllelogram of π_{11++} Markov chain iterates under the uniform distribution.

current example non-response of an individual to Y_1 may depend on the answer that individual would give to Y_1. Hence we may wish to consider an interaction between R_1 and Y_1. A simple method of considering interactions between response and data variables is through log-linear models. Let n_{ijkl} be the count corresponding to the cell $(Y_1 = i, Y_2 = j, R_1 = k, R_2 = l)$, and $\boldsymbol{n} = \{n_{ijkl}\}$ be the vector of fully observed counts. If n_{ijkl} is a random variable from a Poisson distribution, with mean μ_{ijkl} for all i, j, k and l, then modeling $\boldsymbol{n}$ using a log-linear model is the natural approach. As shown in Chapter 3 log-linear models are easy to specify, fit and generalise, and any missing data indicators can be incorporated in exactly the same way any other variable is incorporated.

There are many non-ignorable models we could consider, however we begin by considering two separate methods of parameterising these models

4.2.2.1 Additive Parameterisation of Bishop et al. (1975)

The additive parameterisation is the classic parameterisation of log-linear, and generalised linear, models. Suppose we are interested in the non-ignorable assumption that $Y_1 \perp\!\!\!\perp R_1 | Y_2, R_2$ and $Y_2 \perp\!\!\!\perp R_2 | Y_1, R_1$, then in additive parameterisation this assumption has the following log-linear form.

$$\begin{aligned}\log(\mu_{ijkl}) &= \beta + \beta_{Y_1}(i) + \beta_{Y_2}(j) + \beta_{R_l}(k) + \beta_{R_2}(l) \\ &+ \beta_{Y_1Y_2}(ij) + \beta_{R_1R_2}(kl) + \beta_{Y_1R_2}(il) + \beta_{Y_2R_1}(jk).\end{aligned}$$

In matrix notation we write $\log(\boldsymbol{\mu}) = \boldsymbol{X\beta}$ and, for simplicity, write $Y_1Y_2 + R_1R_2 + Y_1R_2 + Y_2R_1$. The latter notation defines a model in terms of its generators and thus implies all lower order terms. In order to obtain unique estimates for π_{11++} we must constrain the parameters. We use 'sum to zero' constraints. We term this model 'Close to Ignorable' (CI) as the interactions Y_1R_2 and Y_2R_1 are present but the model is not MAR. The CI model differs from MAR by allowing $P(R_1 = 2, R_2 = 2|Y_1 = i, Y_2 = j)$ to depend on i and j.

Not all log-linear models are non-ignorable. For example, in additive notation, the log-linear model for the two partially observed categorical variables corresponding to MCAR is

$$\log(\mu_{ijkl}) = \beta + \beta_{Y_1}(i) + \beta_{Y_2}(j) + \beta_{R_l}(k) + \beta_{R_2}(l) + \beta_{Y_1Y_2}(ij) + \beta_{R_1R_2}(kl).$$

4.2.2.2 Bayesian Estimation of Non-Ignorable Models Using MCMC (Additive Parameterisation)

Given a prior distribution for $\boldsymbol{\beta}$ we proceed by treating the missing cell counts as parameters to be estimated.

Stage 1: Update $\boldsymbol{n}$

We sample from the conditional distribution of these missing counts given the current response parameters $\boldsymbol{\beta}$ (note that given current parameters $\boldsymbol{\beta}$ conditional cell probabilities are easily obtained).

$$\begin{aligned}
n_{ij21} &\sim Multinomial(n_{+j21}, P(Y_1|Y_2, R_1 = 2, R_2 = 1)) \\
n_{ij12} &\sim Multinomial(n_{i+12}, P(Y_2|Y_1, R_1 = 1, R_2 = 2)) \\
n_{ij22} &\sim Multinomial(n_{++22}, P(Y_2, Y_1|R_1 = 2, R_2 = 2)).
\end{aligned}$$

Stage 2: Update $\boldsymbol{\beta}$

Sample from the posterior distribution of the response parameters given the augmented cell counts $\boldsymbol{n} = \{n_{ijkl}\}$. This posterior distribution may not be known, as there is no tractable prior for general log-linear models, hence sample approximately using the Metropolis-Hastings algorithm.

Propose a value $\boldsymbol{\beta}^\star$ from an arbitrary proposal distribution with density function q. Accept this proposal with probability

$$\alpha(\boldsymbol{\beta}^\star,\boldsymbol{\beta}) = \frac{f(\boldsymbol{n}|\boldsymbol{\beta}^\star)f(\boldsymbol{\beta}^\star)q(\boldsymbol{\beta}|\boldsymbol{\beta}^\star)}{f(\boldsymbol{n}|\boldsymbol{\beta})f(\boldsymbol{\beta})q(\boldsymbol{\beta}^\star|\boldsymbol{\beta})} \wedge 1.$$

and rejected (remain at $\boldsymbol{\beta}$) otherwise.

Here $f(\boldsymbol{\beta})$ denotes the prior density function of $\boldsymbol{\beta}$ and, since we assume each n_{ijkl} to have a Poisson distribution with mean μ_{ijkl}, the likelihood $f(\boldsymbol{n}|\boldsymbol{\beta})$ is given by

$$f(\boldsymbol{n}|\boldsymbol{\beta}) \propto \prod_{ijkl} e^{-\mu_{ij11}}{\mu_{ij11}}^{n_{ij11}}.$$

This is a standard MCMC algorithm with the addition of data augmentation to overcome the missing data problem.

We applied the above method and parameterisation to fit the ignorable MCAR model $Y_1Y_2 + R_1R_2$. Using the sum-to-zero constraints, Knuiman and Speed (1988) provided a prior distribution for the parameters of the log-linear model that is symmetric in the sense that it is invariant to arbitrary permutations of the levels of each factor. We therefore assume the following prior distributions for $f(\boldsymbol{\beta})$:

$$f(\boldsymbol{\beta}) = \frac{1}{(2\pi\sigma^2)^{7/2}} \exp\left\{-\frac{1}{2\sigma^2}\boldsymbol{\beta}^T\boldsymbol{\beta}\right\}.$$

The variance parameter was set to $\sigma^2 = 3, 5$ and 20. These parameters have been chosen to reflect the range of *a-priori* plausible values for $\boldsymbol{\beta}$.

Figure 4.2 displays the posterior distribution of π_{11++} for each of the three prior distribution together with MCMC output plots. All results are based upon a sample of 25000 Markov chain iterates. The three diagnostic plots indicate excellent mixing. The posterior mean, assuming $\sigma^2 = 3$, is 0.892. A 95% credibility interval for π_{11++} is (0.877,0.906). These results are similar to those obtained under the MAR model.

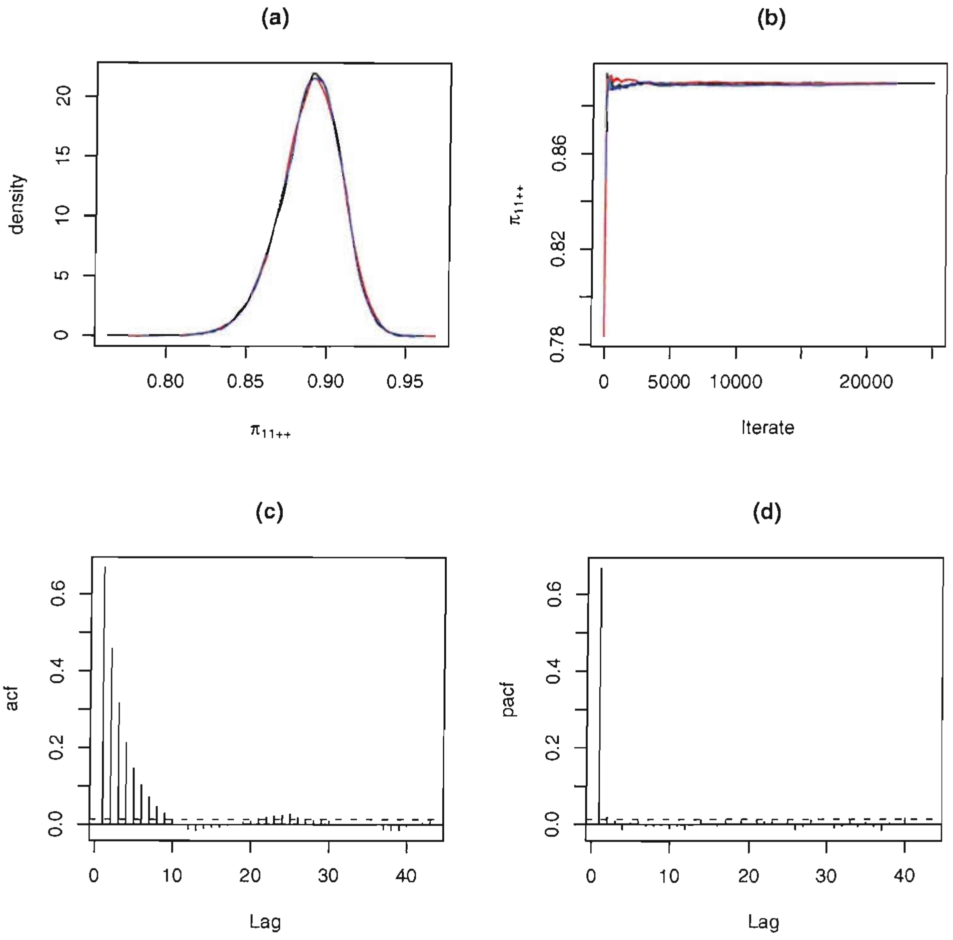


Figure 4.2: MCMC output plots. Plot (a): Posterior density of π_{11++} under MCAR: — $\sigma^2 = 3$, — $\sigma^2 = 5$ and — $\sigma^2 = 20$. Plot (b): A running estimate of the posterior mean for the distribution of π_{11++} under MAR: — $\sigma^2 = 3$, — $\sigma^2 = 5$ and — $\sigma^2 = 20$. Plot (c): Auto-correllelogram of π_{11++} Markov chain iterates $\sigma^2 = 3$. Plot (d): Partial auto-correllelogram of π_{11++} Markov chain iterates $\sigma^2 = 3$.

This is not surprising. Although the functional form of the two models differ, both impose that Y_i is conditionally independent of R_j for $i \neq j$. Had additional covariate information been available this might not have been the case.

4.2.2.3 Multiplicative Parameterisation of Baker et al. (1992)

Baker et al. (1992) provided an alternative parameterisation of the log-linear model. Assuming sum-to-zero constrains they simplify the additive notation as follows

$$
\begin{aligned}
m_{ij} &= \exp\{\beta + \beta_{Y_1}(i) + \beta_{Y_2}(j) + \beta_{Y_1Y_2}(ij) + \beta_{R_l}(1) + \beta_{R_2}(1) + \beta_{Y_1R_l}(i1) \\
&\quad +\beta_{Y_1R_2}(i1) + \beta_{Y_2R_l}(j1) + \beta_{Y_2R_2}(j1) + \beta_{R_1R_2}(11)\} \\
a_{ij} &= \exp\{-2[\beta_{R_l}(1) + \beta_{Y_1R_l}(i1) + \beta_{Y_2R_l}(j1) + \beta_{R_lR_2}(11)]\} \\
b_{ij} &= \exp\{-2[\beta_{R_l}(1) + \beta_{Y_1R_l}(i1) + \beta_{Y_2R_l}(j1) + \beta_{R_lR_2}(11)]\} \\
v &= \exp\{4\beta_{R_lR_2}(11)\},
\end{aligned}
$$

where $m_{ij} \geq 0$, $a_{ij} \geq 0$, $b_{ij} \geq 0$ and $v \geq 0$, and $\sum_{ij} m_{ij}(1 + a_{ij} + b_{ij} + a_{ij}b_{ij}v) = \sum_{ijkl} n_{ijkl}$. Note that no three-way or four-way interactions are included. As before each cell count is assumed to have been generated from an independent Poisson distribution with means given in the following table.

		Y_1	
		$R_1 = 1$	$R_1 = 0$
Y_2	$R_2 = 1$	m_{ij}	$m_{ij}a_{ij}$
	$R_2 = 0$	$m_{ij}b_{ij}$	$m_{ij}a_{ij}b_{ij}v$

Under this parameterisation, closed form maximum likelihood estimates are sometimes directly available. The assumption that g is independent of i and j means that we are limited to models which contain no three or four-way interactions. If 'corner' constraints are used an alternative multiplicative parameterisation can be constructed to include three and four-way interactions. This parameterisation is given below:

$$
\begin{aligned}
m_{ij} &= \exp\{\beta + \beta_{Y_1}(i) + \beta_{Y_2}(j) + \beta_{Y_1Y_2}(ij)\} \\
a_{ij} &= \exp\{\beta_{R_l}(k) + \beta_{Y_1R_l}(ik) + \beta_{Y_2R_l}(jk) + \beta_{Y_1Y_2R_l}(ijk)\} \\
b_{ij} &= \exp\{\beta_{R_2}(l) + \beta_{Y_1R_2}(il) + \beta_{Y_2R_2}(jl) + \beta_{Y_1Y_2R_2}(ijl)\} \\
v_{ij} &= \exp\{\beta_{R_1R_2}(kl) + \beta_{Y_1R_1R_2}(ikl) + \beta_{Y_2R_1R_2}(jkl) + \beta_{Y_1Y_2R_1R_2}(ijkl)\}.
\end{aligned}
$$

Again we impose $m_{ij} \geq 0$, $a_{ij} \geq 0$, $b_{ij} \geq 0$ and $v_{ij} \geq 0$, with the baseline category of each parameter set to zero (corner constraints). Under this parameterisation

$$
\begin{aligned}
&\log(m_{ij}) \text{ corresponds to } 1, Y_1, Y_2, Y_1Y_2 \\
&\log(a_{ij}) \text{ corresponds to } R_1, R_1Y_2, R_1Y_2, R_1Y_1Y_2 \\
&\log(b_{ij}) \text{ corresponds to } R_2, R_2Y_1, R_2Y_2, R_2Y_1Y_2 \\
&\log(v_{ij}) \text{ corresponds to } R_1R_2, R_1R_2Y_1, R_1R_2Y_2, R_1R_2Y_1Y_2
\end{aligned}
$$

The cell means are similar to before with the one exception in the case when $R_1 = R_2 = 0$ which now has mean $m_{ij}a_{ij}b_{ij}v_{ij}$.

As an example consider the non-ignorable non-response model with parameters $a_{ij} = a_j$, $b_{ij} = b_i$ and $v_{ij} = v$. In additive notation this model can be written as

$$Y_1Y_2 + R_1R_2 + Y_1R_2 + Y_2R_1,$$

and corresponds to model CI of the previous section. The likelihood for this model is given as follows

$$\begin{aligned}
L(\boldsymbol{n}|m_{ij}, a_j, b_i, v) \;\propto\; & \left(\prod_{ij} e^{-m_{ij}}(m_{ij})^{n_{ij11}}\right) \\
\times\; & \left(\prod_{i} e^{-b_i(m_{i1}+m_{i2})}(b_i(m_{i1}+m_{i2}))^{n_{i+12}}\right) \\
\times\; & \left(\prod_{j} e^{-a_j(m_{1j}+m_{2j})}(a_j(m_{1j}+m_{2j}))^{n_{+j21}}\right) \\
\times\; & \left(e^{-v\sum_{ij} m_{ij}a_j b_i}(v\sum_{ij} m_{ij}a_j b_i)^{n_{++22}}\right). \qquad (4.2)
\end{aligned}$$

Taking logs, differentiating with respect to v, a_j, b_i and m_{ij}, and setting these derivatives to zero we see that

$$\frac{\partial \log L}{\partial v} = 0 \Rightarrow \hat{v} = \frac{n_{++22}}{\sum_{ij} \hat{m}_{ij}\hat{a}_j\hat{b}_i}.$$

This result is used to obtain $\hat{a}_j$ and $\hat{b}_i$ as a function of $\hat{m}_{ij}$ alone as follows:

$$\begin{aligned}
\frac{\partial \log L}{\partial a_j} = 0 \;\Rightarrow\; & -\hat{m}_{+j} + \frac{n_{+j21}}{\hat{a}_j} - \sum_i \hat{m}_{ij}\hat{b}_i\hat{v} + \frac{n_{++22}}{\sum_{ij}\hat{m}_{ij}\hat{a}_j\hat{b}_i}\sum_i \hat{m}_{ij}\hat{b}_i = 0 \\
\Rightarrow\; & -\hat{m}_{+j} + \frac{n_{+j21}}{\hat{a}_j} - \sum_i \hat{m}_{ij}\hat{b}_i\hat{v} + \sum_i \hat{m}_{ij}\hat{b}_i\hat{v} = 0 \\
\Rightarrow\; & \hat{a}_j = \frac{n_{+j21}}{\hat{m}_{+j}}.
\end{aligned}$$

Through a similar argument it is easily seen that

$$\frac{\partial \log L}{\partial b_j} = 0 \Rightarrow \hat{b}_i = \frac{n_{i+12}}{\hat{m}_{i+}}.$$

Using all three above results we obtain

$$\frac{\partial \log \mathcal{L}}{\partial m_{ij}} = 0 \Rightarrow \hat{m}_{ij} = n_{ij11}.$$

Therefore the maximum likelihood estimates of these parameters are as follows

$$\hat{m}_{ij} = n_{ij11},\ \hat{a}_j = \frac{n_{+j21}}{n_{+j11}},\ \hat{b}_i = \frac{n_{i+12}}{n_{i+11}},\ \hat{v} = \frac{n_{++22}}{\sum_{ij} \hat{m}_{ij}\hat{a}_j\hat{b}_i}.$$

Quantities of interest can readily be expressed in terms of these parameters. For example, the maximum likelihood estimate of the proportion planning to attend and vote in favour of independence is

$$\hat{\pi}_{11++} = \frac{\hat{m}_{11}(1 + \hat{a}_j + \hat{b}_i + \hat{a}_j\hat{b}_i\hat{v})}{\sum_{ijkl} n_{ijkl}} = 0.867.$$

An estimate of v close to 1 would imply that the two response variables, and hence missing data mechanisms, are independent. For this model we have $\hat{g} = 1.99$. In many cases standard errors of estimates are readily available (Baker et al., 1992)

4.2.2.4 Bayesian Estimation of Non-Ignorable Models Using MCMC (Multiplicative Parameterisation)

We assume *a-priori* m_{ij}, a_j, b_i and v to be independent and identically distributed. This strong assumption could greatly effect resulting inference and a sensitivity analysis should be performed.

The prior distribution of these parameters is assumed to be a gamma distribution with shape and rate parameters α and β respectively. These prior parameters can be chosen as follows. If $X \sim \Gamma(\alpha, \beta)$ then it can be shown, using moment generating functions, that

$$\mathbb{E}[\log(X)] = \psi'(\alpha) - \log(\beta)$$

and

$$var[\log(X)] = \psi''(\alpha),$$

where ψ' and ψ'' are the di-gamma and tri-gamma functions respectively. We set the mean to zero and the variance to $3, 5$ and 20 and solve the above simultaneous equations for α and β. We use the maximum likelihood values derived above as starting values for parameters (m_{ij}, a_j, b_i, v).

Stage 1: Update $\boldsymbol{n}$

The first stage in the model fitting process is to sample from the conditional distribution of the missing counts given the current response parameters (m_{ij}, a_j, b_i, v); i.e we sample from the following multinomial distributions

$$
\begin{aligned}
n_{ij21} &\sim Multinomial(n_{+j21}, P(Y_1|Y_2, R_1 = 2, R_2 = 1)) \\
n_{ij12} &\sim Multinomial(n_{i+12}, P(Y_2|Y_1, R_1 = 1, R_2 = 2)) \\
n_{ij22} &\sim Multinomial(n_{++22}, P(Y_2, Y_1|R_1 = 2, R_2 = 2)).
\end{aligned}
$$

Stage 2: Update (m_{ij}, a_j, b_i, v)

The second stage is to Gibbs sample from the posterior distributions of the response parameters given the new augmented cell counts $\boldsymbol{n}$. Assuming a gamma prior distribution for all parameters the posterior distribution are known, tractable and given as follows:

$$
\begin{aligned}
m_{ij} &\sim \Gamma(\alpha + \sum_{kl} n_{ijkl}, \beta + 1 + a_j + b_i + a_j b_i g) \\
a_j &\sim \Gamma(\alpha + \sum_{il} n_{ij2l}, \beta + \sum_{i} m_{ij}(1 + b_i g)) \\
b_i &\sim \Gamma(\alpha + \sum_{jk} n_{ijk2}, \beta + \sum_{j} m_{ij}(1 + a_j g)) \\
v &\sim \Gamma(\alpha + \sum_{ij} n_{ij22}, \beta + \sum_{ij} m_{ij} a_j b_i).
\end{aligned}
$$

Sampling in turn from stages 1 and 2 produces a dependent sample from the joint posterior of the model parameters and missing cell counts given the observed data $\boldsymbol{n}_{obs}$. We can readily produce point estimates of π_{11++} together with a plot of the posterior density π_{11++} given the observed data $\boldsymbol{n}_{obs}$. This plot is given in Figure 4.3 for $\sigma^2 = 3, 5$ and 20. For reference, MCMC output plots have also been plotted. All results are based upon a generated sample of 25000 iterations. The posterior means for the three prior distributions are 0.868, 0.865 and 0.856. As we would expect these means are in general agreement with the maximum likelihood estimate.

4.2.2.5 A Comparison of the Parameterisations

The above subsections illustrated the ease at which non-ignorable models are created and fit. Maximum likelihood estimates are sometimes directly available. If this is not the case, then MLE's are easily obtained using the EM algorithm. Posterior distributions of parameters of interest are easily generated using simple, but effective, Markov chain Monte Carlo algorithms.

The only advantage of the multiplicative parameterisation is the ease at which we can sample from the resulting tractable posteriors using the Gibbs sampler.

We adopt the additive parameterisation as it is more natural than the multiplicative parameterisation of Baker et al. (1992). This parameterisation allows us to use the techniques of Chapter 3 when attempting to discriminate between the competing non-ignorable non-response models.

4.3 Model Selection with Missing Data

There are many different non-ignorable models that can be fit to the Slovenian plebiscite data. Each model specifies both the data generating process and the missing data mechanism, and each model may provide vastly different estimates of the parameter of interest. Consider Table 4.4 containing estimates of the posterior

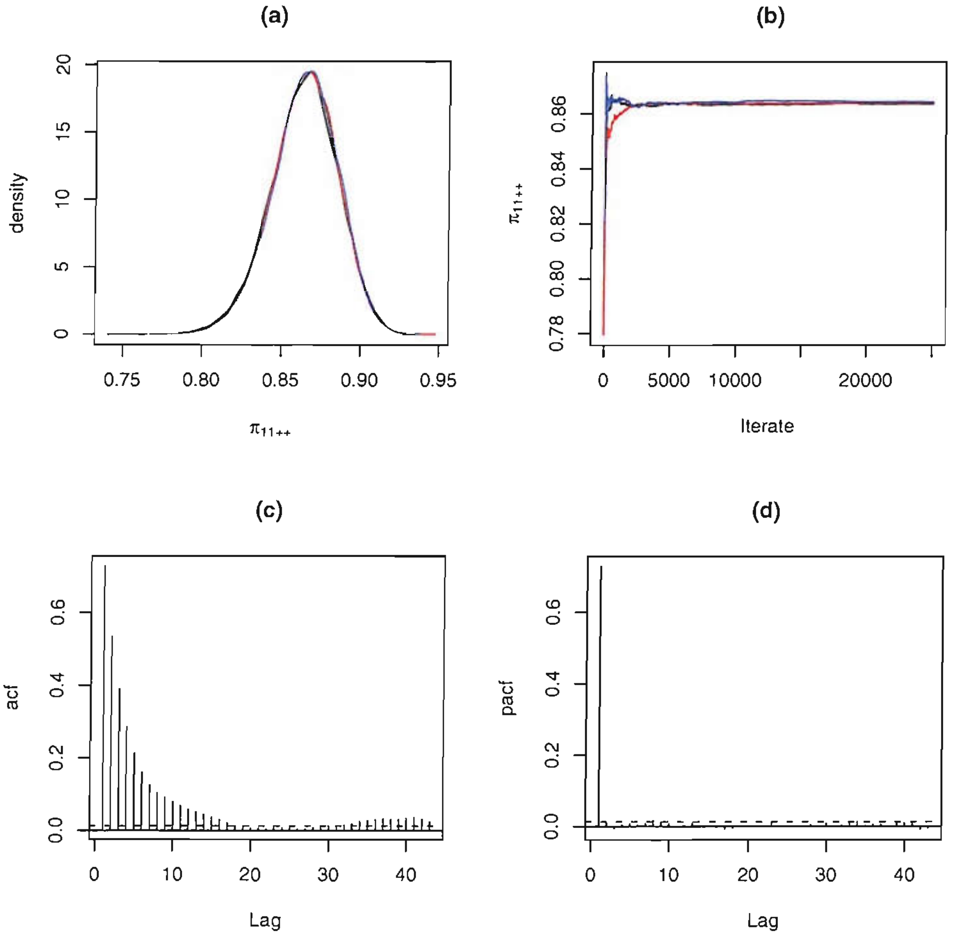


Figure 4.3: MCMC output plots. Plot (a): Posterior density of π_{11++} under $Y_1Y_2 + R_1R_2 + Y_1R_2 + Y_2R_1$: — $\sigma^2 = 3$, — $\sigma^2 = 5$ and — $\sigma^2 = 20$. Plot (b): A running estimate of the posterior mean for the distribution of π_{11++} under $Y_1Y_2 + R_1R_2 + Y_1R_2 + Y_2R_1$: — $\sigma^2 = 3$, — $\sigma^2 = 5$ and — $\sigma^2 = 20$. Plot (c): Auto-correllelogram of π_{11++} Markov chain iterates $\sigma^2 = 3$. Plot (d): Partial auto-correllelogram of π_{11++} Markov chain iterates $\sigma^2 = 3$.

mean for 9 different log-linear models. These models are equivalent to those given in Molenberghs et al. (2001).

Table 4.4: Estimates of the proportion π attending the plebiscite and voting for independence, following from fitting the models of Baker et al. (1992) within a Bayesian framework.

Number	Model	$\hat{\pi}$ ($\sigma^2 = 3$)	$\hat{\pi}$ ($\sigma^2 = 5$)	$\hat{\pi}$ ($\sigma^2 = 20$)
1	$Y_1Y_2 + R_1R_2$	0.892	0.892	0.892
2	$Y_1Y_2 + R_1R_2 + Y_1R_2$	0.882	0.881	0.881
3	$Y_1Y_2 + R_1R_2 + Y_2R_1$	0.884	0.884	0.884
4	$Y_1Y_2 + R_1R_2 + Y_2R_2$	0.832	0.82	0.827
5	$Y_1Y_2 + R_1R_2 + Y_1R_1$	0.772	0.765	0.768
6	$Y_1Y_2 + R_1R_2 + Y_1R_1 + Y_1R_2$	0.769	0.764	0.767
7	$Y_1Y_2 + R_1R_2 + Y_2R_1 + Y_2R_2$	0.812	0.81	0.805
8	$Y_1Y_2 + R_1R_2 + Y_1R_1 + Y_2R_2$	0.742	0.715	0.736
9	$Y_1Y_2 + R_1R_2 + Y_1R_2 + Y_2R_1$	0.868	0.868	0.868

Model 1 is MCAR, and of the 9 models in Table 4.4 it is the only ignorable non-response model. Models 8 and 9 are non-ignorable non-decomposable graphical models. Model 8 asserts that response to a question is due to the answer that would have been provided by *that* question. Whilst model 9 is CI asserting that the response to a question depends on the answers to *other* questions. The two models provide entirely different estimates of π_{11++}. If we assumed a normal prior distribution for β with variance given by $\sigma^2 = 3$ we obtain estimates 0.742 and 0.868 for models 8 and 9 respectively. These estimates lie within the crude optimistic/pessimistic interval and indeed the election result (0.885) is contained within both 95% credible intervals. In the following section we attempt to discriminate between the competing non-response models.

4.3.1 Notation and Prior Distributions

We proceed as in Chapter 3. The four factors of interest here are Y_1, Y_2, R_1 and R_2. A saturated component for the data (Y_1Y_2 and all implied lower order terms) is included in all models under consideration. We also include the generator R_1R_2. Thus the simplest model, in terms of numbers of parameters, is the ignorable model corresponding to MCAR. These terms are included as we are interested in the interactions between $\boldsymbol{Y}$ and $\boldsymbol{R}$.

Each model, indexed by m, specifies that n_{ijkl} is independently distributed according to a Poisson random variable with mean $\mathbb{E}[n_{ijlk}] = \mu_{ijkl}$. For all models under consideration we use the logarithmic (canonical) link function

$$\log(\boldsymbol{\mu}) = \boldsymbol{X}_m\boldsymbol{\beta}_m.$$

The log-likelihood for observed and augmented counts $\boldsymbol{n} = \{n_{ijkl}\}$ is given by

$$\log(f(\boldsymbol{n}|\boldsymbol{\beta}_m, m)) \propto \boldsymbol{n}^T\boldsymbol{X}_m\boldsymbol{\beta}_m - \mathbf{1}^T\boldsymbol{X}_m\boldsymbol{\beta}_m.$$

In the absence of any strong prior information we assume

$$f(\boldsymbol{\beta}_m|m) \sim N_{p_m}(\mathbf{0}, \sigma^2\boldsymbol{I}_{p^m})$$

and all models equally likely. We assume that each of the non-response mechanisms is equally likely to have generated the data. Finally we set $\sigma^2 = 3$.

4.3.2 Data Augmented Reversible Jump MCMC

There are numerous difficulties associated with implementing the reversible jump scheme. The space we intend to explore is vast and strong correlations exist between augmented data counts, coefficients and model indicators. Proposal distributions of reversible jump algorithm discussed in Chapter 3 depend critically upon the data which in this instance is only partially observed. The task is onerous. If we were required to fit a single non-ignorable non-response model a plausible move type, and one that would improve the mixing of the Markov chain, is to simultaneously update both $\boldsymbol{\beta}_m$ and $\boldsymbol{n}$. This move is illustrated below (stage 2). However, when a move proposes to simultaneously switch models and update $\boldsymbol{n}$ it is hard to ensure the reversibility condition as the proposal distribution for additional parameters is permitted to depend on $\boldsymbol{n}$. After some experimentation the following scheme was adopted.

Assume the state of the Markov chain at time t is $(\boldsymbol{\beta}_m, m, \boldsymbol{n})$, where $\boldsymbol{n}$ denotes a full cross classification of observed and augmented cell counts. Proceed as follows

Stage 1: Update $\boldsymbol{\beta}_m$ using the Metropolis-Hastings algorithm

Generate z uniformly from $\{1, 3\}$ and generate a proposed value $\boldsymbol{\beta}^p$ from

$$N_{p_m}(\boldsymbol{\beta}_m, z(\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X})^{-1}). \tag{4.3}$$

Scaling of the proposal variance allows 50% of moves to be, on average, further away from the current parameter values of $\boldsymbol{\beta}_m$.

Define $\hat{\boldsymbol{W}}$ to be a 16 by 16 matrix with diagonal $\boldsymbol{n}$ and zero otherwise.

Stage 2: With probability 0.2 simultaneously update $\boldsymbol{\beta}_m$ and $\boldsymbol{n}$

Generate z uniformly from $\{1, 3\}$ and sample from the following distribution to obtain a proposed value $\boldsymbol{\beta}^p_m$

$$\boldsymbol{\beta}^p_m = N_{p_m}(\boldsymbol{\beta}_m, z(\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X})^{-1}).$$

Given $\boldsymbol{\beta}^p_m$ calculate $P(Y_1|Y_2, R_1 = 2, R_2 = 1), P(Y_2|Y_1, R_1 = 1, R_2 = 2)$ and $P(Y_2, Y_1|R_1 = 2, R_2 = 2)$. Using these probabilities sample from

$$\begin{aligned} n^p_{ij21} &\sim Multinomial(n_{+j21}, P(Y_1|Y_2, R_1 = 2, R_2 = 1)) \\ n^p_{ij12} &\sim Multinomial(n_{i+12}, P(Y_2|Y_1, R_1 = 1, R_2 = 2)) \\ n^p_{ij22} &\sim Multinomial(n_{++22}, P(Y_2, Y_1|R_1 = 2, R_2 = 2)). \end{aligned}$$

Accept the proposed move to $(\boldsymbol{\beta}_m^p, \boldsymbol{n}^p)$ with probability

$$\frac{f(\boldsymbol{n}^p|\boldsymbol{\beta}_m^p, m)f(\boldsymbol{\beta}_m^p|m)\pi(\boldsymbol{n}'|\boldsymbol{\beta}_m^p)}{f(\boldsymbol{n}|\boldsymbol{\beta}_m, m)f(\boldsymbol{\beta}_m|m)\pi(\boldsymbol{n}'|\boldsymbol{\beta}_m)} \wedge 1,$$

where π is the multinomial distribution function.

Stage 3: With probability 0.1 perform a data augmentation step

Given $\boldsymbol{\beta}_m$ calculate $P(Y_1|Y_2, R_1 = 2, R_2 = 1)$, $P(Y_2|Y_1, R_1 = 1, R_2 = 2)$ and $P(Y_2, Y_1|R_1 = 2, R_2 = 2)$. Using these probabilities sample from the above multinomial distributions to obtain a new value of $\boldsymbol{n} = \{n_{ijkl}\}$.

Stage 4: Reversible jump birth or death move

With equal probability propose a birth or death step as detailed in Chapter 3.
The variety of above moves will improve the mixing of the Markov chain.

4.3.3 Results and Discussion

We ran the Markov chain for a total of 10 million iterations. Since traversing the joint parameter space was extremely slow, we sampled every 500th iteration. This provided 20,000 generated Markov chain observations for analysis.
In total 11.8% of proposed model moves were accepted with the Markov chain visiting 39 models. Twenty three of these models had a posterior probability greater than 0.02. The first 6 of these are given in Table 4.5. The 'close to ignorable' model has a posterior probability of approximately 0.01. The posterior distribution is fairly flat across the model space. This suggests there is little, or no, information with which to compare models and therefore missing data mechanisms. We note that all models in Table 4.5 are over-parameterised and all provide a perfect fit to

Table 4.5: Posterior model probabilities for Slovenian plebiscite data

Model	Posterior Probability
$Y_1Y_2R_2 + Y_1R_1R_2$	0.064
$Y_1Y_2R_1 + Y_1Y_2R_2 + Y_1R_1R_2 + Y_1R_1R_2$	0.064
$Y_1Y_2R_1 + Y_1Y_2R_2 + Y_1R_1R_2$	0.052
$Y_1Y_2R_1 + Y_1Y_2R_2 + Y_2R_1R_2$	0.051
$Y_1Y_2R_2 + Y_1R_1R_2 + Y_1R_1R_2$	0.041
$Y_1Y_2R_1 + Y_2R_1R_2$	0.040

the observed data. Table 4.6 provides posterior inclusion probabilities for each of the interactions not in the MCAR model. Since many of the models in Table 4.5 contained a large number of terms, these inclusion probabilities are mostly greater than a half.

Table 4.6: Slovenian plebiscite data: Posterior inclusion probabilities.

Term	Inclusion Probability
Y_1R_1	0.91
Y_2R_1	0.83
$Y_1Y_2R_1$	0.47
Y_1R_2	0.85
Y_2R_2	0.95
$Y_1Y_2R_2$	0.52
$Y_1R_1R_2$	0.50
$Y_2R_1R_2$	0.49

Figure 4.4 provides diagnostic plots for the parameter of interest π_{11++}. The posterior distribution is given in plot (a). This distribution covers a large range of values.

The model averaged posterior mean is 0.797, and a 95% credible interval is given by (0.698,0.892). The auto and partial correllelogram, plots (c) and (d), give an indication of the algorithms performance. We observe a strong correlation between successive values of π_{11++} produced by the Markov chain. This is in spite of heavy thinning.

The plot of the running estimate of π_{11++} is also a little worrying. It seems that the Markov chain has taken considerable time to converge to a stationary distribution, if it has converged at all.

Figure 4.5 provide plots for the augmented cell count n_{2112}. The plots are fairly typical of all the augmented counts. Plot (c) (the auto-correllelogram) again hints at the poor mixing of the Markov chain in spite of the heavy thinning. The partial auto-correllelogram is slightly more promising. The posterior distribution of n_{2112} is given in plot (a). The little information regarding the augmented count has come directly from the prior distribution (demonstrating the critical issue of the choice of prior distributions). This is reflected in the cell count histogram. The distribution of the cell counts gives posterior weight to large interactions hence the resulting posterior model probabilities. Had we any reasonable prior information, either in the form of prior model probabilities or prior cell probabilities, the above analysis may prove interesting.

Most model selection procedures are based on a combination of parsimony and goodness of fit to the observed data. All models in Table 4.5 provide a reasonable fit to the observed data, and provide similar predictions for $P(\boldsymbol{Y}|R_1 = R_2 = 1)$. This is also true for the MCAR and MAR models detailed earlier in the chapter.

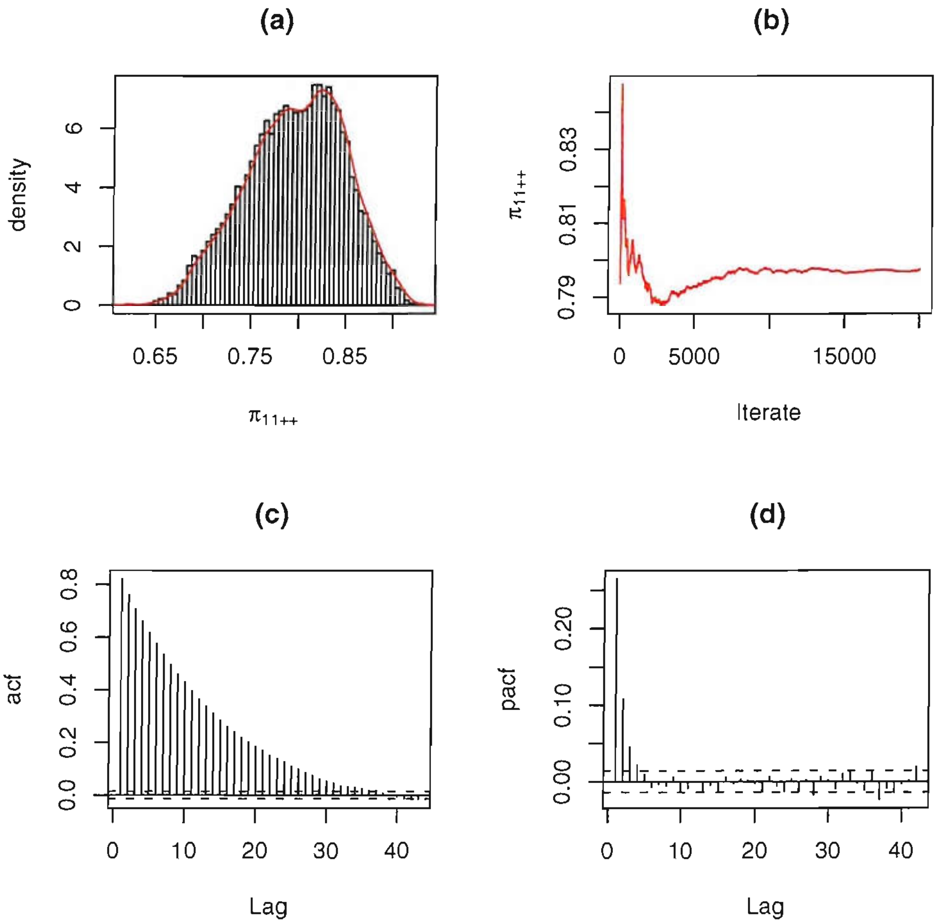


Figure 4.4: RJMCMC output plots for the Slovenian plebiscite data. Plot (a): Histogram of the posterior distribution of π_{11++}. Plot (b): A running posterior mean of π_{11++}. Plot (c): Auto-correllelogram of π_{11++}. Plot (d): Partial auto-correllelogram of π_{11++}.

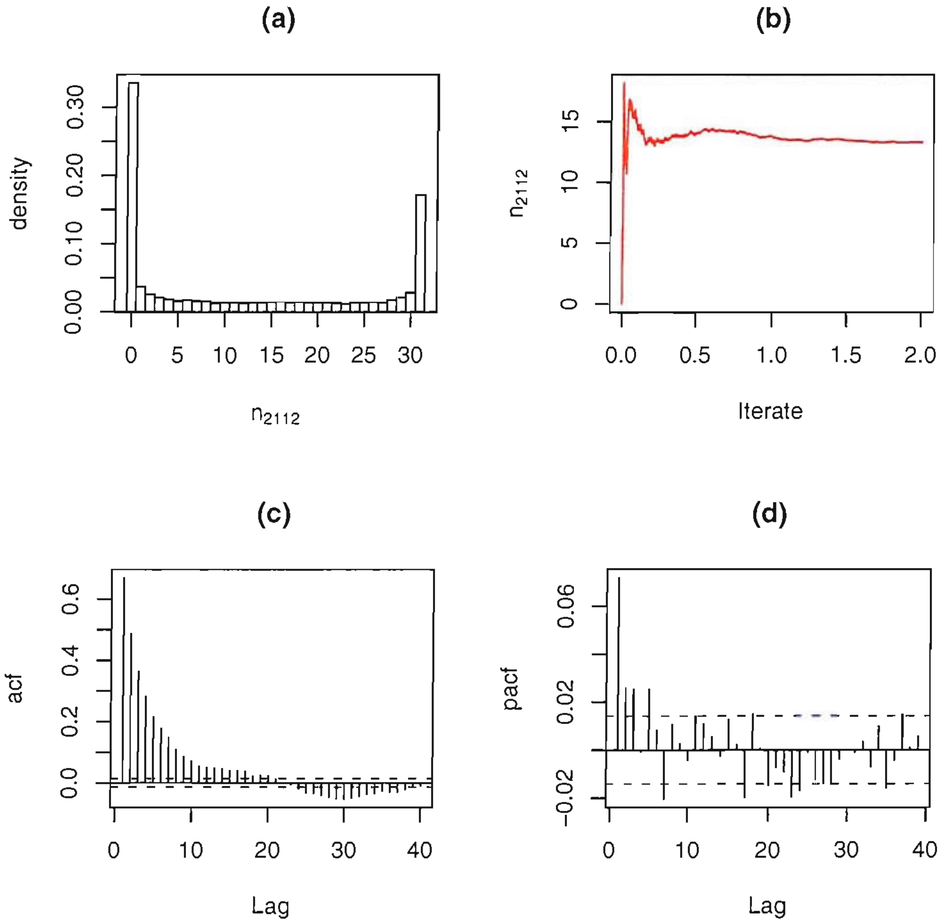


Figure 4.5: RJMCMC output plots for the Slovenian plebiscite data. Plot (a): Histogram of the posterior distribution of n_{2112}. Plot (b): A running posterior mean of n_{2112}. Plot (c): Auto-correllelogram of n_{2112}. Plot (d): Partial auto-correllelogram of n_{2112}.

However, we should not discard all models in Table 4.5 in favour of MCAR on the basis of parsimony since, had we fully observed the data these less parsimonious models may have provided a better description of the full data.

On the basis of the partially observed data, we are not able to distinguish between conditional independence structures defined by the non-ignorable models. We are not able to describe the missing data mechanism. In particular, we have no information to determine if the missing data mechanism is ignorable or non-ignorable. These conclusion were also noted by Forster and Smith (1998) who questioned whether model comparison procedures based upon the observed data were appropriate.

4.4 A Model Based Approach to Sensitivity Analysis

There is an enormous discrepancy in the way in which the models contained in Table 4.4 treat the missing data. Each model specifies a different non-response mechanism, and each provides a different posterior estimate for the quantity of interest. These estimates range from 0.742 to 0.892 when $\sigma^2 = 3$. In the previous section we attempted to discriminate between competing models using a reversible jump Markov chain Monte Carlo scheme to provide approximations to the marginal likelihoods. We concluded that there was insufficient information with which to compare models and, in particular, we do not have the information to determine if the missing data mechanism is non-ignorable. The approach (comparing hierarchical log-linear models for the joint distribution of the data and response indicator) has been considered by Baker and Laird (1988), Forster and Smith (1998) and Molenberghs et al. (2001), although in a less formal setting. Forster and Smith (1998) reached similar conclusions to those presented within. Since the data provide such little information concerning the non-response mechanism any additional information is useful. This information can take the form of previous poll and election results together with expert information. Our approach is therefore to introduce parameters which control the extent of non-ignorability into a single model for the observed data. The sensitivity of quantities of interest (π_{11++}) is then considered relative to realistic *a-priori* changes in these parameters. Examples of this approach can be found in Little (1982), Kadane (1993) and Forster and Smith (1998).

4.4.1 Incorporating Uncertainty about Ignorability

Our starting point is not the MAR model of Rubin (1976), but instead the close to ignorable log-linear model with terms

$$Y_1Y_2 + R_1R_2 + Y_1R_2 + Y_2R_1.$$

This model was introduced in Section 4.2.2.3. It states that response to a particular question is allowed to depend upon the answers to other questions. We would like to analytically compare this model to Rubin's MAR but, since MAR does not have tractable likelihood estimates, a direct comparison is problematic. Both models provided an adequate fit to the Slovenian plebiscite data, and both models provide an estimate of π_{11++} close to the plebiscite result. To draw a meaningful comparison suppose we had observed the poll data table below. The data is identical to Table 4.2 with the exception that marginal counts n_{1+12} and n_{1+12} have been set to zero.

Now, a tractable estimate for π_{11++} is available under both models. Consider firstly the non-ignorable model. Through a similar argument to Section 4.2.2.3 it can be shown that

$$\hat{\pi}_{11++}^{MLE} = \frac{n_{1111} + \frac{n_{1111}}{n_{1+11}}\left(n_{+121} + \frac{n_{+121}}{n_{++21}}n_{++22}\right)}{n_{++11} + n_{++21} + n_{++22}}.$$

	Independence		
Attendance	Yes	No	Don't Know
yes	1439	78	159
no	16	16	32
Don't Know	0	0	136

Now, under the MAR assumption

$$\hat{\pi}_{11++}^{MAR} = \frac{n_{1111} + \frac{n_{1111}}{n_{1+11}} n_{+121}}{n_{++11} + n_{++21}}.$$

From which it is seen that

$$\hat{\pi}_{11++}^{MLE} = \alpha \, \hat{\pi}_{11++}^{MAR} + (1 - \alpha) \frac{n_{1111}}{n_{1+11}} \frac{n_{+121}}{n_{++21}}, \tag{4.4}$$

where

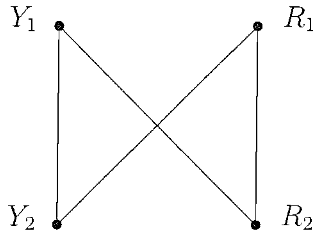
$$\alpha = \frac{n_{++11} + n_{++21}}{n_{++11} + n_{++21} + n_{++22}} \quad 1 - \alpha = \frac{n_{++22}}{n_{++11} + n_{++21} + n_{++22}}.$$

Clearly if $n_{++22} = 0$ then $\hat{\pi}_{11++}^{MLE} = \hat{\pi}_{11++}^{MAR}$. The principal difference between the two models, is therefore the handling of individuals who did not respond to either question. These individuals contribute the following sum to the estimate of π_{11++} under the non-ignorable model when compared to MAR.

$$P(R_1 = R_2 = 2) \frac{n_{1111}}{n_{1+11}} \frac{n_{+121}}{n_{++21}}.$$

This follows directly from (4.4). Had we observed data identical to Table 4.2 with marginal counts for n_{++22} set to zero, then $\hat{\pi}_{11++}^{MLE} = 0.894$ is almost identical to the approximation obtained under the MAR model. It is reasonable to believe therefore that the two models will differ substantially when n_{++22} is large in comparison to other observed counts. If this were the case, it would not be sensible to completely ignore the majority who failed to respond to both questions and hence an estimate based upon MAR might be misleading. We thus consider the model $Y_1Y_2 + R_1R_2 + Y_1R_2 + Y_2R_1$ as our starting model.

This 'close to ignorable' log-linear model has the following graphical representation.



Since the graph contains a 4-cycle, it is not a decomposable graphical model. That said, maximum likelihood estimates for the parameters of the model are found easily, because of the missing data, and the model is easily fit within a Bayesian framework. This model implies the following conditional independence structures:

$$Y_1 \perp\!\!\!\perp R_1 | Y_2, R_2$$
$$Y_2 \perp\!\!\!\perp R_2 | Y_1, R_1.$$

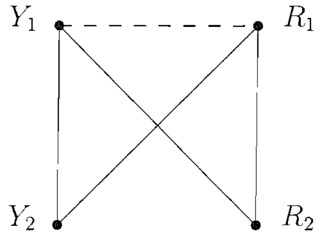
The first conditional independence statement implies that $\forall\ Y_2, R_2$

$$\phi(Y_1, R_1|Y_2, R_2) = \log\left(\frac{P(Y_1 = 1, R_1 = 1|Y_2, R_2)P(Y_1 = 2, R_1 = 2|Y_2, R_2)}{P(Y_1 = 1, R_1 = 2|Y_2, R_2)P(Y_1 = 2, R_1 = 1|Y_2, R_2)}\right) = 0 \tag{4.5}$$

with a similar expression for given as follow for the second statement. That is, $\forall\ Y_1, R_1$

$$\phi(Y_2, R_2|Y_1, R_1) = \log\left(\frac{P(Y_2 = 1, R_2 = 1|Y_1, R_1)P(Y_2 = 2, R_2 = 2|Y_1, R_1)}{P(Y_2 = 1, R_2 = 2|Y_1, R_1)P(Y_2 = 2, R_2 = 1|Y_1, R_1)}\right) = 0. \tag{4.6}$$

We can use these log-odds ratios to represent the prior belief in non-ignorability. For example, any inference is going to be sensitive to the conditional independence assumption $Y_1 \perp\!\!\!\perp R_1 | Y_2, R_2$. Allowing a departure from this assumption allows missingness on Y_1 to be dependent to the answer that would have been provided on Y_1. We incorporate this departure from non-ignorability by including the terms $Y_1R_1 + R_1Y_1Y_2 + Y_1R_1R_2$ in the 'close to ignorable' model, indicated by dashed edge in the graph below.



The augmented model is graphical and decomposable. This augmented model maintains the second conditional independence $(Y_2 \perp\!\!\!\perp R_2 | Y_1, R_1)$.

4.4.2 Markov Chain Monte Carlo and Results

Our augmented model specifies that n_{ijkl} is independently distributed according to a Poisson random variable with mean $\mathbb{E}[n_{ijlk}] = \mu_{ijkl}$. Again, we use the logarithmic link function together with the additive parameterisation. The augmented model can therefore be written as follows

$$\log(\boldsymbol{\mu}) = \boldsymbol{X\beta} + \boldsymbol{Z\gamma} \tag{4.7}$$

Where $\boldsymbol{X}$ and $\boldsymbol{Z}$ are design matrices corresponding to the 'close to ignorable' and non-ignorable parameters respectively. Parameter vectors $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are assumed unknown, and we assume that

$$\boldsymbol{\beta} \sim N_p(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_p)$$

and

$$\boldsymbol{\gamma} \sim N_q(\mathbf{0}, \alpha^2 \boldsymbol{I}_q). \tag{4.8}$$

Again we set $\sigma^2 = 3$. It is clear that $p = 9$ and $q = 3$. The parameter α controls the extent of non-ignorability. Since $P(Y_1 = 1, R_1 = 1|Y_2, R_2)$ is a product of log-normal variables it follows that $P(Y_1 = 1, R_1 = 1|Y_2, R_2)$ is also log-normal. Thus $\exp(\phi(Y_1, R_1|Y_2, R_2))$ is log-normal and hence $\phi(Y_1, R_1|Y_2, R_2)$ is normal. In fact

$$\phi(Y_1, R_1|Y_2, R_2) \sim N(0, \sigma_\phi^2),$$

where $\sigma_\phi = 4\alpha$ is obtained by directly considering (4.7) and (4.8) above. The variance of ϕ is therefore controlled through the specification of α. When $\alpha = 0.345$ $\sigma_\phi = 1.38$ and hence ϕ lies in the range [-2.71,2.71] with 95% probability. Since the posterior distribution is intractable we use the following Markov chain to generate dependent samples from the posterior distributions of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$.

1. Metropolis-Hastings component-wise update for each of the 11 parameters

2. Metropolis-Hastings block update of the close to ignorable parameters

3. Metropolis-Hastings block update of the non-ignorable parameters

4. Data augmentation step

This scheme will produce a dependent sample of observations for our analysis. For α fixed in the range $[0, 0.345]$ and hence $\sigma_\phi \in [0, 1.38]$. The above scheme was run for a total of 2.5 millions iterations retaining every 50th iterate for analysis. Even for $\alpha = 0.345$ ($\sigma_\phi = 1.38$), diagnostic plots provided evidence of adequate mixing of the Markov chain.

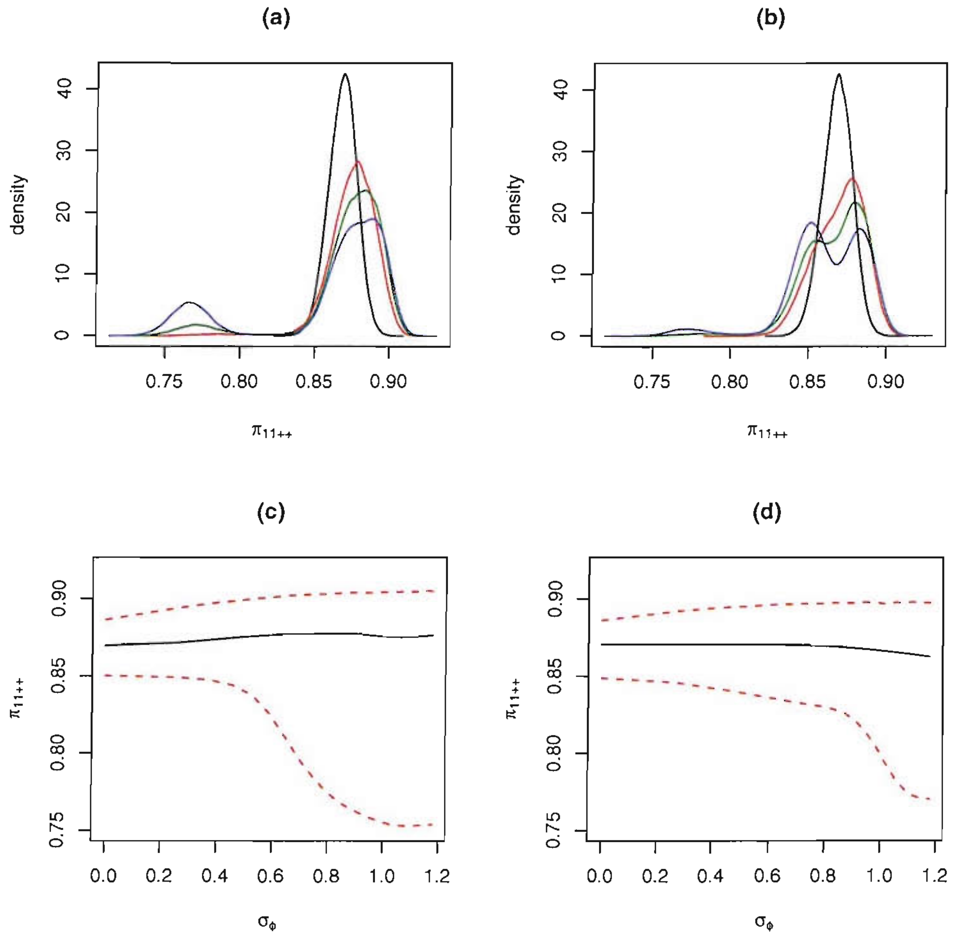


Figure 4.6: Plot (a): Posterior density of π_{11++} for augmented model corresponding to $\phi(Y_1, R_1|Y_2, R_2) \neq 0$. Prior parameters $\sigma^2 = 3$, $\sigma_\phi = 0$ —, $\sigma_\phi = 0.35$ —, $\sigma_\phi = 0.7$ — and $\sigma_\phi = 1.38$ —. Plot (b): Posterior density of π_{11++} for augmented model corresponding to $\phi(Y_2, R_2|Y_1, R_1) \neq 0$. Prior parameters $\sigma^2 = 3$, $\sigma_\phi = 0$ —, $\sigma_\phi = 0.35$ —, $\sigma_\phi = 0.7$ — and $\sigma_\phi = 1.38$ —. Plot (c): Sensitivity of π_{11++} to α for missingness on R_1. Solid line represents the posterior mean, whilst the dashed line represent a 95% credible interval. Plot (d): Sensitivity of π_{11++} to α for missingness on R_2. Solid line represents the posterior mean, whilst the dashed line represent a 95% credible interval.

Figure 4.6 presents the marginal posterior mean and 95% credible intervals of π_{11++} plotted against σ_ϕ for the two augmented models. The means are stable across σ_ϕ. As expected the posterior variances of π_{11++} are highly sensitive to σ_ϕ. The effect of increasing σ_ϕ on the posterior variance is most marked for the case where $\phi(Y_1, R_1|Y_2, R_2)$ is allowed to vary from 0. It is intriguing to note that as σ_ϕ varies from 0, the posterior density is no longer unimodal.

4.5 Extension to the Three-Way Table

We return to the original data set presented in Rubin et al. (1995) and given in Table 4.1. For the third question of interest (Are you in favour of Slovenia's secession from Yugoslavia?) we introduce the response indicator R_3. The fully observed data now forms a 2^6 table with counts n_{ijklmn} for $i, j, k, l, m, n \in \{1, 2\}$ where, for example, n_{111111} denotes the number of respondents with $Y_1 = 1, Y_2 = 1, Y_3 = 1, R_1 = 1, R_2 = 1$ and $R_3 = 1$. Hence $n_{111111} = 1191$. As before, we do not observe all 64 counts. Only 8 counts, corresponding to n_{ijk111}, are fully observed together with 19 partial counts.

The MAR assumption does not correspond to a log-linear model for the three-way table, as is the case for the two-way table. The likelihood function has a similar parametric form to that given in Section 4.2.1. Again this model has the peculiar feature that the probability of missingness on Y_1, Y_2 and Y_3 is not allowed to depend on Y_1, Y_2 or Y_3, whilst the probability of missingness on Y_1 where Y_2 and Y_3 are provided is permitted to depend on Y_2 and Y_3. The likelihood is easily maximised, using the EM algorithm, providing a maximum likelihood estimate of $\hat{\pi}_{111+++} = 0.883$.

The 'close to ignorable' is composed of the following terms

$$Y_1Y_2Y_3 + R_1R_2R_3 + Y_1R_2R_3 + R_1Y_2R_3 + R_1R_2Y_3 + R_1Y_2Y_3 + Y_1R_2Y_3 + Y_1Y_2R_3.$$

Once more, the model is saturated for the data. There are 8 parameters for the complete data and 19 parameters for the missing data mechanism. A corresponding multiplicative parameterisation is available and given in Appendix B. Using this multiplicative parameterisation it is possible to derive as analytical form of the maximum likelihood estimate of π_{11++++}. This is also shown in Appendix B. Under this model the maximum likelihood estimate for π_{11++++} is $\hat{\pi}_{11++++} = 0.8543$.

Figure 4.7 provides posterior distributions of π_{111+++} obtained using a simple Markov chain Monte carlo algorithm. The posterior means are approximately 0.85 for the three prior distributions. It is our belief that the result for the 'CI' model differs from MAR primarily as a result of those individuals failing to provide an answer to all three questions. This is unverifiable due to nature of the MAR assumption.

It is also possible to obtain estimates of posterior model probabilities using RJMCMC. We do not implement this approach partly because of the computational expense, but moreover because there is little information contained in the missing counts with which to compare models. In the two way case we found that if a prior for the model space permitted almost any model then a posterior credible interval for π covered the pessimistic/optimistic range. We thus proceed immediately with a sensitivity analysis.

The 'close to ignorable model' has the following graphical representation

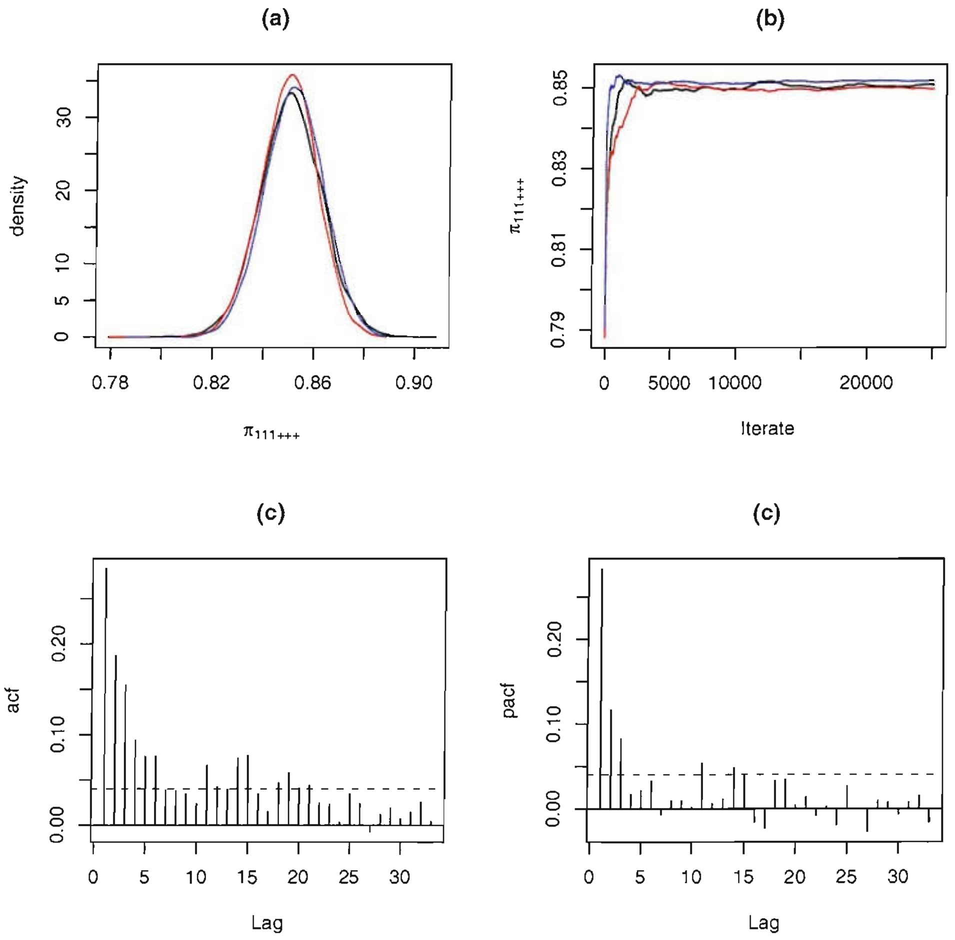
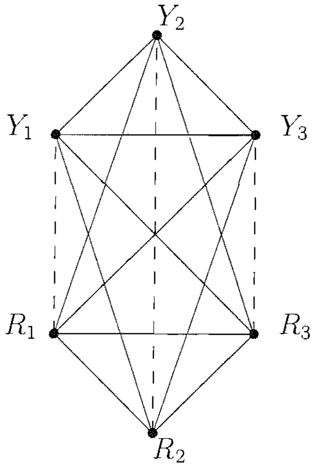


Figure 4.7: MCMC output plots for the Slovenian plebiscite data, three questions, close to ignorable model. Plot (a): Histogram of the posterior distribution of π_{111+++}. Plot (b): A running posterior mean of π_{111+++}. Plot (c): Auto-correllelogram of π_{111+++}. Plot (d): Partial auto-correllelogram of π_{111+++}.



The graphical is not decomposable for the full data since there are three four-cycles. These four-cycles, $\{Y_1, Y_2, R_1R_2\}$, $\{Y_1, Y_3, R_1R_3\}$ and $\{Y_2, Y_3, R_2R_3\}$, induce the following conditional independencies

$$
\begin{aligned}
Y_1 \perp\!\!\!\perp R_1 | Y_2, R_2, Y_3, R_3 &\Rightarrow \phi(Y_1, R_1 | Y_2Y_3R_2R_3) = 0 \\
Y_2 \perp\!\!\!\perp R_2 | Y_1, R_1, Y_3, R_3 &\Rightarrow \phi(Y_2, R_2 | Y_1Y_3R_1R_3) = 0 \\
Y_3 \perp\!\!\!\perp R_3 | Y_1, R_1, Y_2, R_2 &\Rightarrow \phi(Y_3, R_3 | Y_1Y_2R_1R_2) = 0.
\end{aligned}
$$

To allow these conditional independencies to depart from zero, and hence a departure from ignorability, we consider the model with the additional edge (together with appropriate terms) corresponding to the log odds ratio. These edges are shown as dashed edges in the above graph. For example if we wished to allow the first log odds ratio to depart from zero we include the dashed edge corresponding to the terms R_1Y_1, $R_1Y_1Y_2$, $R_1Y_1Y_3$ and $R_1Y_1Y_2Y_3$. This new model is not graphical since the $Y_1R_1R_2$ interaction is not included. If we assume these parameters are a-priori independent and identically normal with mean zero and variance α^2, then the distribution of the log-odds ratios is also normal with $\sigma_\phi = 8\alpha$. This value for

σ_ϕ is derived through similar reasoning to the two-way case. A Markov chain for exploring the posterior distribution was constructed in an identical fashion to the two-way case. Figure 4.8 presents the marginal posterior mean and 95% credible intervals of π_{11++++} plotted against α for the all three augmented models. The means are stable across σ_ϕ. As expected the posterior variances of π_{11++++} are sensitive to σ_ϕ. The effect of increasing σ_ϕ on the posterior variance is most marked for the case where $\phi(Y_1, R_1|Y_2Y_3R_2R_3)$ is allowed to vary from 0, and least marked when $Y_3 \perp\!\!\!\perp R_3|Y_1, R_1, Y_2, R_2 \neq 0$.

4.6 Closing Remarks

In this chapter we have presented a collection of statistical techniques for inference when data, in the form of a contingency table, is only partially observed. We have shown that the problem of inference under such circumstances to be hard and that any assumptions we make concerning the missing data mechanism and prior information are critical.

Frequently a single model, hence an assumption concerning the missing data, is selected for inferential purposes. This essentially assumes a prior model probability of 1 for the selected model. As we have seen, this approach may fail to capture the uncertainty of parameters of interest and realistic deviations from this assumption can greatly affect inference.

We attempted to discriminate between competing models. By successfully applying a data augmented reversible jump algorithm we were able to provide a measure of model uncertainty. For the case of the Slovenian plebiscite data we have shown that there is little information available to discriminate between competing models. Incorporating model uncertainty (we calculated a model averaged poster for the pa-

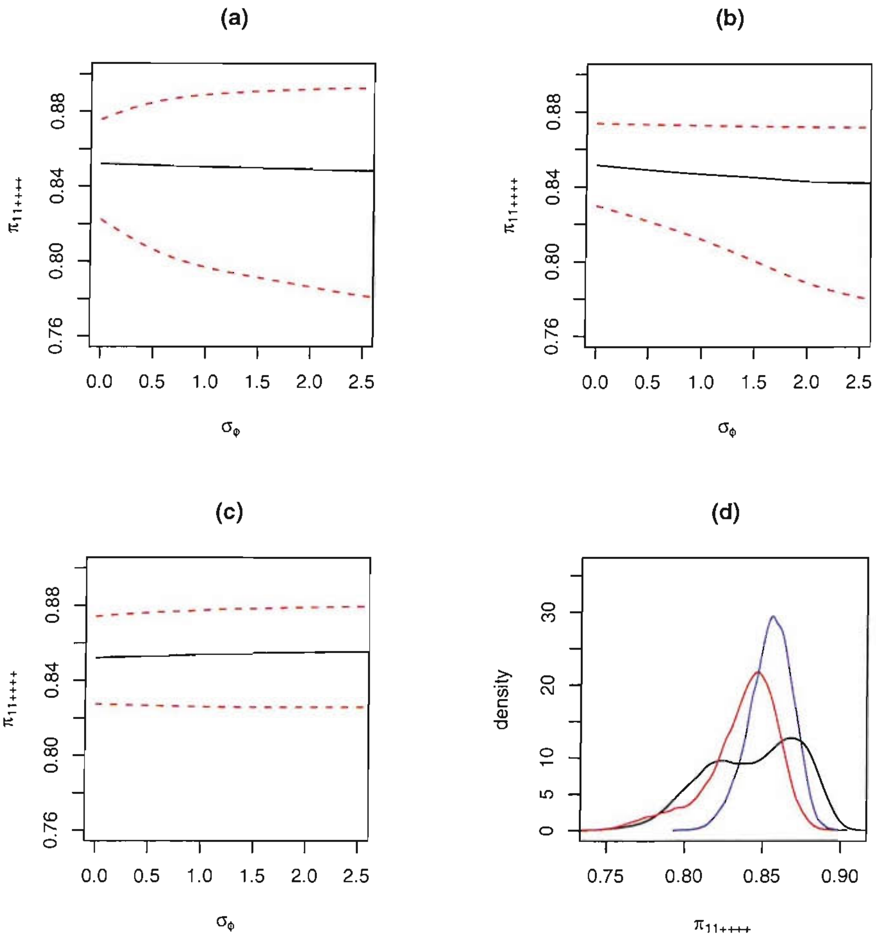


Figure 4.8: Plot (a): Sensitivity of π_{11++++} to σ_ϕ for missingness on R_1. Solid line represents the posterior mean, whilst the dashed line represent a 95% credible interval. Plot (b): Sensitivity of π_{11++++} to σ_ϕ for missingness on R_2. Solid line represents the posterior mean, whilst the dashed line represent a 95% credible interval. Plot (c): Sensitivity of π_{11++++} to σ_ϕ for missingness on R_3. Solid line represents the posterior mean, whilst the dashed line represent a 95% credible interval. Plot (d): Posterior density of π_{11++++} for augmented models with $\sigma_\phi = 1.4$. — $\phi(Y_1, R_1|Y_2Y_3R_2R_3)$ is allowed to vary from 0, — $\phi(Y_2, R_2|Y_1Y_3R_1R_3)$ is allowed to vary from 0 and — $\phi(Y_3, R_3|Y_1Y_2R_1R_2)$ is allowed to vary from 0.

rameter of interest) greatly increased the width of a 95% credible interval compared to selecting any single model.

We then considered the sensitivity of quantities of interest relative to realistic changes of the missing data assumption. We permitted a-priori changes in log-odds ratios, allowing deviations from conditional independencies specified by our 'close to ignorable' model. Inference was extremely sensitive to even small changes in these log odds ratios.

To conclude, we have shown inference when data, in the form of a contingency table, is only partially observed to be a difficult problem.

If prior information is available the this is easily incorporated into the inference process. However, resulting inferences are critical to any assumptions made and a sensitivity analysis should therefore be performed.

If little prior information is available the any inference should clearly state the uncertainty with which the inference has been made.

Chapter 5

Missing Data and Disclosure Control

Statistical agencies and other organisations conducting surveys or collecting data may release results of these exercises to third parties. For example, the Office for National Statistics' may release data files from the census to academic institutions for secondary analysis.

These data releases are for statistical purposes only, i.e. for making inference concerning groups of people differentiated by some characteristic. *Statistical disclosure* arises if the third party can disclose confidential information about the individual units or people, which originally provided the data.

Suppose a national statistical agency released to an academic researcher a table from a survey containing information, stratified by electoral ward, on occupation, income and gender. The researcher knows there is only one female dentist living in his electoral ward and is thus able to find the dentist's income. This is *statistical disclosure* although a trivial example.

Clearly statistical disclosure is undesirable as it usually violates the pledge, sometimes legal, of confidentiality. Furthermore, the organisation risks the loss of cooperation in future surveys if negative publicity concerning disclosure were to arise. As a result statistical agencies and other organisations take seriously the issue of statistical disclosure control, hence the growth of interest in the subject over past twenty years. Statistical disclosure control takes two forms.

Firstly organisation may restrict access to the data or place stringent legal conditions on its use. This is called *Access Control.*
Secondly, a variety of *Statistical Disclosure Protection Techniques* may be applied to the data before release to reduce disclosure risk. Categorical variables may be recoded to reduce the number of levels or counts omitted from released data. Deterministic or stochastic perturbation mechanisms may also be applied to the data. Such a technique may reduce the risk of disclosure but potentially at the cost of information loss.

Good introductions to statistical disclosure control and statistical disclosure techniques can be found Willenborg and de Waal (1996) and Willenborg and de Waal (2001) and the reader is referred to these texts.

The research within this chapter was commissioned by the Office for National Statistics' Neighbourhood Statistics Service.

5.1 Introduction - Disclosure Control

The most common problem in the field of statistical disclosure occurs when a statistical agency releases data, consisting of the values of a number of categorical variables, on a sample of individuals from a population. One form of identification risk occurs when there are sample cell counts of 1 (uniques) in the marginal table representing the cross-classification of individuals by a subset of key variables (those variables whose values in the population are available to a potential intruder from a source external to the released data under consideration). If the intruder can determine, with confidence, that a sample unique in the contingency table of key variables, is also unique in the population, then this individual can be identified and the data release allows disclosure of the values of the remaining variables for this individual. A variety of risk measures for this problem have been proposed by Skinner and Elliot (2002) and an approach for the accurate estimation of these measures discussed in Forster and Webb (2007).

Within this chapter we examine a slightly different disclosure problem.
Suppose a statistical agency makes publicly available a number of key data sets stratified by an area such as an electoral ward. Suppose these data sets take the form of a series of multiway margins of a larger cross-classification.
For example, a statistical agency might release data concerning all recipients of a particular benefit living within a given electoral ward. Suppose further that this data is the table corresponding to the cross classification of these individuals by age, gender and marital status. One form of identification risk occurs when there are cell counts of 1. For example there may be a single female divorcee under 18 years of age receiving this benefit. Information regarding this individual has therefore been released to any person that can identify this individual by the cross classifying variables.

In order to reduce this disclosure risk categorical variables may be recoded to reduce the number of levels, or counts omitted from released data. Deterministic or stochastic perturbation mechanisms may also be applied to the data. A commonly used rounding method was suggested by Nargundkar and Saveland (1972) and it is this method which we adopt throughout the remainder of the chapter.

Let x be the observed count to be rounded and b be the rounding base assumed to be a small integer. Let $\lfloor x \rfloor$ indicate the largest multiple of b which is less than or equal to x, and $\lceil x \rceil$ indicate the smallest multiple of b which is greater than or equal to x. Then if $x \neq \lfloor x \rfloor \neq \lceil x \rceil$ round, stochastically, to obtain the count for release y. The count is rounded 'up' to $\lceil x \rceil$ with probability

$$\frac{x - \lfloor x \rfloor}{b},$$

and down to $\lfloor x \rfloor$ with probability

$$\frac{\lceil x \rceil - x}{b}.$$

If $x = \lfloor x \rfloor = \lceil x \rceil$ set $y = x$ so if a value of the cell is an integer of the rounding base this value is unaltered.

For example, let $b = 5$. Then the probabilities for unbiased rounding are given in Table 5.1.

Table 5.1: A stochastic rounding mechanism: Probabilities for unbiased rounding (b=5).

Residual after dividing by b	Probability to round to $\lceil x \rceil$	Probability to round to $\lfloor x \rfloor$
1	1/5	4/5
2	2/5	3/5
3	3/5	2/5
4	4/5	1/5

The mechanism is stochastic and is designed so that no 'bias' is introduced. I.e $E(y) = x$ where the expectation is with respect to the perturbation mechanism. Alternative rounding mechanisms are described in Willenborg and de Waal (2001). One such method is *conventional rounding.* Here a count is rounded to its closest multiple of the common base. Willenborg and de Waal (2001) provide an example where a cell and marginal counts, from a 2×2 contingency table, have been rounded to their closest multiples of 5. This example is given below. The left table represents the un-rounded counts and the right the rounded.

12	9	3
8	6	2
4	3	1

10	10	5
10	5	0
5	5	0

If it is known that conventional rounding was used then it is possible to obtain the original table directly from the rounded table. For this reason conventional rounding is not used. If it is believed that the rounding mechanism is that of Nargundkar and Saveland (1972) then there are 439 different tables that could have produced the rounded table.

Rounded margins of a complete cross classification are often released to reduce the risk of disclosure. Consider Table 5.2 consisting of individuals receiving benefit B in ward W cross classified by age and sex where marginal counts have been rounded to a common base taken to be 5.

It is desirable to assess the resulting disclosure risk prior to release. Clearly if we could determine the exact counts from the release rounded data, and the covariates contained sensitive information, then we might conclude the disclosure risk was great.

Another approach to assessing this risk has been to compute upper and lower bounds of the true cell counts, based on the rounded counts. Where the difference between

Table 5.2: A simple disclosure example.

		Age						
		< 20	$20 - 29$	$30 - 39$	$40 - 49$	$50 - 59$	60+	Total
Sex	Female							35
	Male							10
Total		0	0	5	0	5	30	45

the upper and lower bounds is large, it might be concluded that significant uncertainty exists about the true cell counts and hence disclosure risk is low. However, it is possible that even where this difference is large, the data may be informative about the true cell count because most of the range between the bounds has a negligible probability of having generated the rounded data.

If the difference between the upper and lower bounds is small or the range of a 95% posterior credible interval is small it is not right to conclude that the resulting disclosure risk is large. The interval may contain large counts and therefore ?? could still be small.

In this chapter we attempt to quantify more precisely the uncertainty about the true cell counts, given the rounded data, and attemp to provide a more reliable assessment of disclosure risk. The approach we take is Bayesian. Given the rounded cell counts (data), we aim to provide a posterior probability distribution for the true cell counts (parameters). Disclosure risk can then be directly assessed in terms of the posterior probability that a given cell count can be determined to be in a sensitive range (typically zero or other small values).

5.2 Notation

Let $\boldsymbol{x} = (x_1, \ldots, x_n)^T$ be the vector of true cell counts for a particular ward. Here $\boldsymbol{x}$ represents the complete cross-classification by all released variables, even if only certain margins are released. If age (6 categories) and sex (2 categories) are released either as individual margins, as a cross-classification, or both, then $\boldsymbol{x}$ has 12 components. We use the $p \times n$ matrix $\boldsymbol{D}$ to denote the mapping between the true cell counts and the true values of the released margins. For example, if $\boldsymbol{x}$ represents the 6×2 cross-classification by age and sex as shown in example (5.2), and let

$$\boldsymbol{D} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

Then $\boldsymbol{Dx}$ is the true value of the margins released in Table 5.2. The disclosure control mechanism takes the true value $\boldsymbol{Dx}$ of the margins to be released, and applies a random perturbation to obtain the rounded margins $\boldsymbol{y} = (y_1, ..., y_p)$, for release. The stochastic rounding mechanism in (5.1) can be formally written down as follows:

$$y_i = \begin{cases} \lfloor (\boldsymbol{Dx})_i \rfloor & \text{with probability}1 - \frac{1}{b}[(\boldsymbol{Dx})_i \mod b] \\ \lceil (\boldsymbol{Dx})_i \rceil & \text{with probability}\frac{1}{b}[(\boldsymbol{Dx})_i \mod b] \end{cases} \qquad (5.1)$$

where $a \mod b = a - \lfloor a \rfloor$ and $y_1, \ldots, y_n$ are generated independently.

An alternative, but equivalent formulation for this rounding mechanism is

$$y_i = \lfloor (\boldsymbol{Dx})_i + z_i \rfloor \tag{5.2}$$

where z_i is an integer uniformly distributed on $\{0, 1, \ldots, b-1\}$. We treat the vector $\boldsymbol{z}$ as an auxiliary variable.

The likelihood for model (5.1) is given by

$$\begin{aligned} f(\boldsymbol{y}|\boldsymbol{x}) &= \prod_{i=1}^{p} \left[1 - \frac{1}{b}[(\boldsymbol{Dx})_i \mod b]\right]^{I(y_i = \lfloor (\boldsymbol{Dx})_i \rfloor)} \\ &\quad \times \left[\frac{1}{b}[(\boldsymbol{Dx})_i \mod b]\right]^{I(y_i = \lceil (\boldsymbol{Dx})_i \rceil)} \\ &\quad \times I\left(y_i \in \{\lceil (\boldsymbol{Dx})_i \rceil, \lfloor (\boldsymbol{Dx})_i \rfloor\}\right) \end{aligned} \tag{5.3}$$

where the indicator function $I(\cdot)$ is equal to 1 if $\cdot$ is true and 0 otherwise. The term $I\left(y_i \in \{\lceil (\boldsymbol{Dx})_i \rceil, \lfloor (\boldsymbol{Dx})_i \rfloor\}\right)$ in each component of the product in (5.3) defines the bounds on which the proposed rounding method is based.

Bayesian inference encapsulates the uncertainty about the unknown true cell counts $\boldsymbol{x}$, given the perturbed margins $\boldsymbol{y}$ by a posterior distribution $f(\boldsymbol{x}|\boldsymbol{y})$, given by Bayes' theorem as

$$f(\boldsymbol{x}|\boldsymbol{y}) \propto f(\boldsymbol{y}|\boldsymbol{x}) f(\boldsymbol{x})$$

where $f(\boldsymbol{y}|\boldsymbol{x})$ is given by (3), and $f(\boldsymbol{x})$ is a prior distribution representing the uncertainty about $\boldsymbol{x}$ prior to obtaining the data $\boldsymbol{y}$.

We might choose a vague prior distribution for $\boldsymbol{x}$, representing a high level of uncertainty. In this case we assume in the absence of observed data, that

$$f(\boldsymbol{x}) = \frac{1}{(k+1)^n} \prod_{i=1}^{n} I(x_i \in \{0, \ldots, k\}). \tag{5.4}$$

In other words, we assume that the cell counts x_i are independently uniformly distributed between 0 and k, where k is chosen to be large. Provided that k is larger than any bound likely to arise as a result of the rounding process, then the constraint that $x_i \leq k$ is irrelevant for practical purposes. Later in this chapter we describe a Bayesian approach where information available at higher geographical levels of aggregation may be incorporated into a more informative prior distribution for $\boldsymbol{x}$.

Assuming the prior density given by (5.4) it is easily seen that

$$\begin{aligned} f(\boldsymbol{x}|\boldsymbol{y}) \;\propto\; & \prod_{i=1}^{p} \left[1 - \frac{1}{b}[(\boldsymbol{D}\boldsymbol{x})_i \mod b]\right]^{I(y_i = \lfloor(\boldsymbol{D}\boldsymbol{x})_i\rfloor)} \\ & \times \left[\frac{1}{b}[(\boldsymbol{D}\boldsymbol{x})_i \mod b]\right]^{I(y_i = \lceil(\boldsymbol{D}\boldsymbol{x})_i\rceil)} \\ & \times I\left(y_i \in \{\lceil(\boldsymbol{D}\boldsymbol{x})_i\rceil, \lfloor(\boldsymbol{D}\boldsymbol{x})_i\rfloor\}\right). \end{aligned} \qquad (5.5)$$

The posterior distribution (5.5) summarises uncertainty about the true cell counts $\boldsymbol{x}$, in light of rounded data $\boldsymbol{y}$. In particular, uncertainty about an individual cell count is summarised by its marginal distribution, for example

$$f(x_1|\boldsymbol{y}) = \sum_{x_2=0}^{k} \cdots \sum_{x_n=0}^{k} f(\boldsymbol{x}|\boldsymbol{y}). \qquad (5.6)$$

Therefore, Bayesian disclosure risk assessment involves computing unnormalised joint (5.5) or marginal (5.6) probabilities for true cell counts, and then normalising. In principle, (5.5) or (5.6) can be calculated for every $\boldsymbol{x}$ which satisfies the bounds

$$\prod_{i=1}^{p} I\left(y_i \in \{\lceil(\boldsymbol{D}\boldsymbol{x})_i\rceil, \lfloor(\boldsymbol{D}\boldsymbol{x})_i\rfloor\}\right) = 1. \qquad (5.7)$$

These bounds can be constructed using the method described by Fienberg (1999) or Dobra and Fienberg (2001). However, the number of such $\boldsymbol{x}$ can be very large. Even for the simple 2 × 6 example presented in Table 5.2 there are more than 10, 000, 000 possible $\boldsymbol{x}$. Thus, complete enumeration is often infeasible for even moderate-sized examples. We therefore construct a Markov chain on the state space consisting of all tables which satisfy (5.7).

5.3 Markov Chain Monte Carlo

For most interesting cases complete enumeration of the state space is not feasible and therefore the exact calculation of the joint, or marginal, probabilities for the true cell counts is not possible. An alternative approach is to generate a sample from $f(\boldsymbol{x}|\boldsymbol{y})$ and use sample proportions to estimate probabilities. We shall use MCMC to generate this sample.

5.3.1 A Gibbs Sampler with Auxiliary Variables

One possible approach for generating from $f(\boldsymbol{x}|\boldsymbol{y})$ is based on the alternative formulation (5.2) for the rounding process. Here, we consider the (unknown) perturbations $\boldsymbol{z}$ as an auxiliary variable and part of our analysis. We attempt to generate from the joint posterior distribution $f(\boldsymbol{z}, \boldsymbol{x}|\boldsymbol{y})$ using a Gibbs sampler. To achieve this, we note that the conditional distributions $f(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y})$ and $f(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{y})$ are straightforward to generate from and hence a Gibbs sampler is immediately available.

Starting from $\boldsymbol{x}^0$, we generate $\boldsymbol{z}^1$ from $f(\boldsymbol{z}|\boldsymbol{x}^0, \boldsymbol{y})$ and then $\boldsymbol{x}^1$ from $f(\boldsymbol{x}|\boldsymbol{z}^1, \boldsymbol{y})$. The method then proceeds by iteratively updating $\boldsymbol{z}$ and $\boldsymbol{x}$ in this fashion. In fact the $\boldsymbol{x}$ are updated component by component with each cell count x_i being generated conditionally given the current values of the other cell counts.

Given $\boldsymbol{x}$, z_i is distributed uniformly on $\{\max\{0, y_i - (\boldsymbol{D}\boldsymbol{x})_i\}, \ldots, \min\{b-1, y_i - (\boldsymbol{D}\boldsymbol{x})_i + b - 1\}\}$. The conditional distribution of x_i given $\boldsymbol{z}$ and $x_j, j \neq i$ is uniform over a constrained region where the constraints are determined by examining those rows of $\boldsymbol{D}$ where the value in the ith column is greater than zero. For such a row, denoted $\boldsymbol{D}_j$, the corresponding constraint on x_i is derived from

$$y_j - z_j \leq \boldsymbol{D}_j\boldsymbol{x} \leq y_j - z_j + b - 1.$$

However, it is easy to construct an example where the Gibbs sampler is not irreducible. Suppose that just two margins of a 2×2 table are released, both rounded to base 2, and that they are $(0,0)$ and $(2,2)$. The only two tables that could have generated these margins are $(1,0,0,1)$ and $(0,1,1,0)$. Transitions between these states is impossible using the Gibbs sampler as described above which only allows transitions which change a single x_i at a time. For this reason we focus on the Metropolis Hastings algorithm.

5.3.2 A Metropolis-Hastings Sampler

It is possible to sample approximately from $f(\boldsymbol{x}|\boldsymbol{y})$ using only the unnormalised expression (5.5) using the Metropolis-Hastings algorithm. This method generates dependent observations from $f(\boldsymbol{x}|\boldsymbol{y})$ by simulating a Markov chain with equilibrium distribution $f(\boldsymbol{x}|\boldsymbol{y})$. Starting with an arbitrary $\boldsymbol{x}^0$ with $f(\boldsymbol{x}|\boldsymbol{y}) > 0$, we represent the generated sample by $\{\boldsymbol{x}^0, \boldsymbol{x}^1\boldsymbol{x}^2, \ldots\}$ where $\boldsymbol{x}^{t+1}$ is generated from $\boldsymbol{x}^t$ by first proposing a value $\boldsymbol{x}^\star$ from an arbitrary proposal distribution. Then, the proposal is accepted ($\boldsymbol{x}^{t+1} = \boldsymbol{x}^\star$) with probability

$$\alpha(\boldsymbol{x}^\star, \boldsymbol{x}^t) = \frac{f(\boldsymbol{x}^\star|\boldsymbol{y})q(\boldsymbol{x}^t|\boldsymbol{x}^\star)}{f(\boldsymbol{x}^t|\boldsymbol{y})q(\boldsymbol{x}^\star|\boldsymbol{x}^t)} \wedge 1 \tag{5.8}$$

and rejected ($\boldsymbol{x}^{t+1} = \boldsymbol{x}^t$) otherwise. Note that, as $f(\boldsymbol{x}|\boldsymbol{y})$ appears in both the numerator and denominator of (5.8), the normalising constant is not required and (5.5) can be used.

The difficulties are in finding a starting value $\boldsymbol{x}^0$ with $f(\boldsymbol{x}^0|\boldsymbol{y}) > 0$ and in ensuring that the resulting algorithm is irreducible.

A starting value can either be identified by directly evaluating the bounds using the method of Fienberg (1999) or, if that is infeasible, by a stochastic search (applying successive proposal steps until a $\boldsymbol{x}^0$ with $f(\boldsymbol{x}|\boldsymbol{y}) > 0$ is identified).

To ensure irreducibility and aperiodicity, we suggest a distribution which proposes modest perturbations to $\boldsymbol{x}$, through

$$\boldsymbol{x}^\star = \boldsymbol{x}^t + \epsilon \tag{5.9}$$

where ϵ has a discrete distribution. The set of moves ϵ can be constructed from a set of independent Poisson random variable as discussed in Section 5.7.

5.4 An Artificial Example

In order to illustrate the approach we return to example (5.2). Upper and lower bounds for cell counts are easily calculated for this table and are given in Table 5.3. Although many of these bounds contain zero, the differences between upper and lower bounds are wide. Therefore, the table may be considered safe for release. However, for many cells there may be much more information available from the posterior distribution than from the bounds alone. We calculate the posterior distribution using the Metropolis-Hastings algorithm described in the previous section. In this case a starting value is easily identified. The posterior distributions for the female cell counts are give in Figure 5.1, and for males Figure 5.2.

Table 5.3: Upper and lower bounds for example (5.2).

		Age						
		< 20	$20-29$	$30-39$	$40-49$	$50-59$	60+	Total
Sex	Female	[0,4]	[0,4]	[0,9]	[0,4]	[0,9]	[13,34]	35
	Male	[0,4]	[0,4]	[0,9]	[0,4]	[0,9]	[0,14]	10
Total		0	0	5	0	5	30	45

These probabilities are based on 1 million iterations which took no more than three minutes to produce, and with over 50% of proposed moves being accepted. In practice 10,000 iterations for a table of this size may suffice. We may only be interested

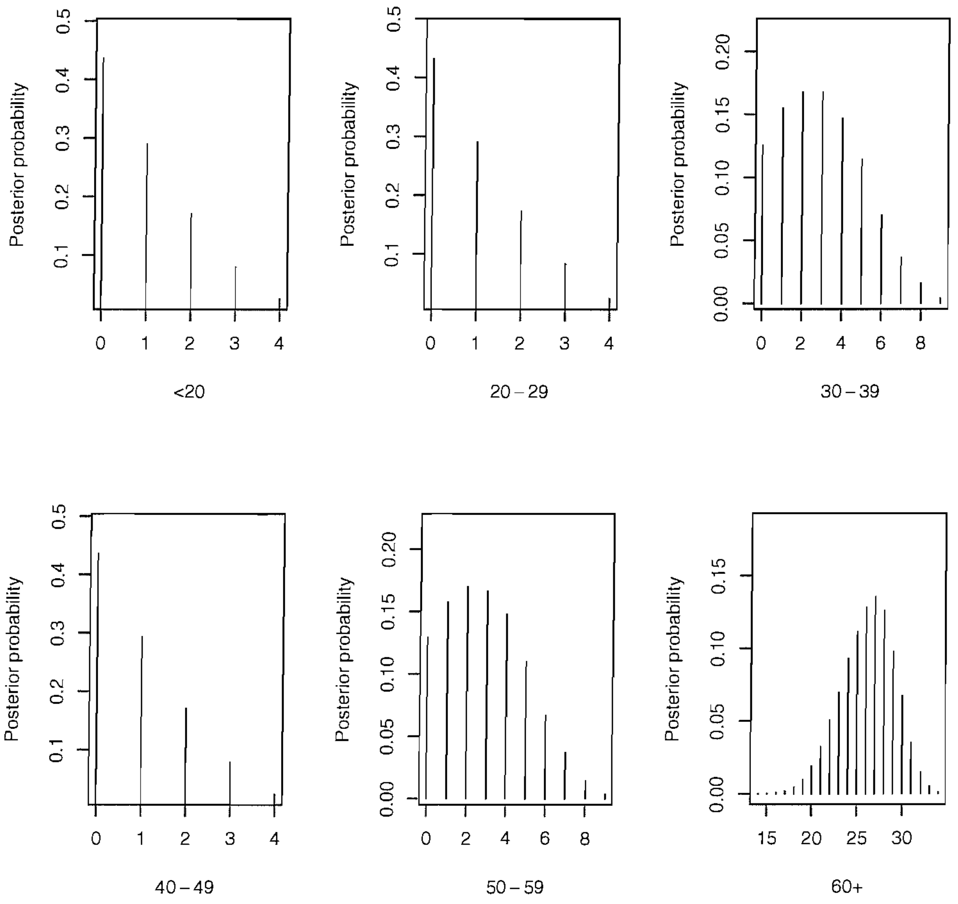


Figure 5.1: Marginal posterior distribution of cell counts, artificial disclosure example (5.2), female counts.

in an estimate of the cell probabilities up to two decimal places, to decide whether releasing the table would be disclosive with respect to some measure. Hence this approach is practical. It is easily seen that posterior distributions are more informative, and that using bounds alone may give a potentially misleading impression of the disclosure protection provided by rounding. For example, the probability that cell count 1, corresponding to females under twenty, is greater than two is around

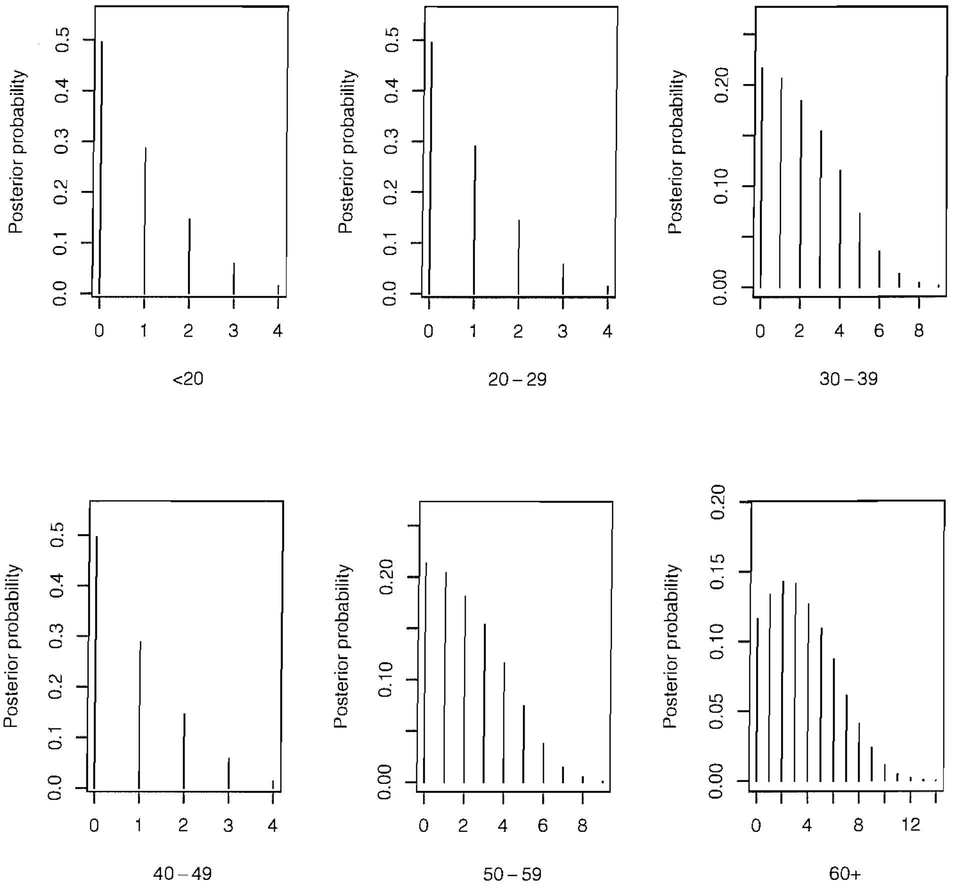


Figure 5.2: Marginal posterior distribution of cell counts, artificial disclosure example (5.2), male counts.

0.1.

Figure 5.3 displays typical MCMC output plots. The first 25,000 data points have been used only, with the data additionally being thinned so that only every 10th

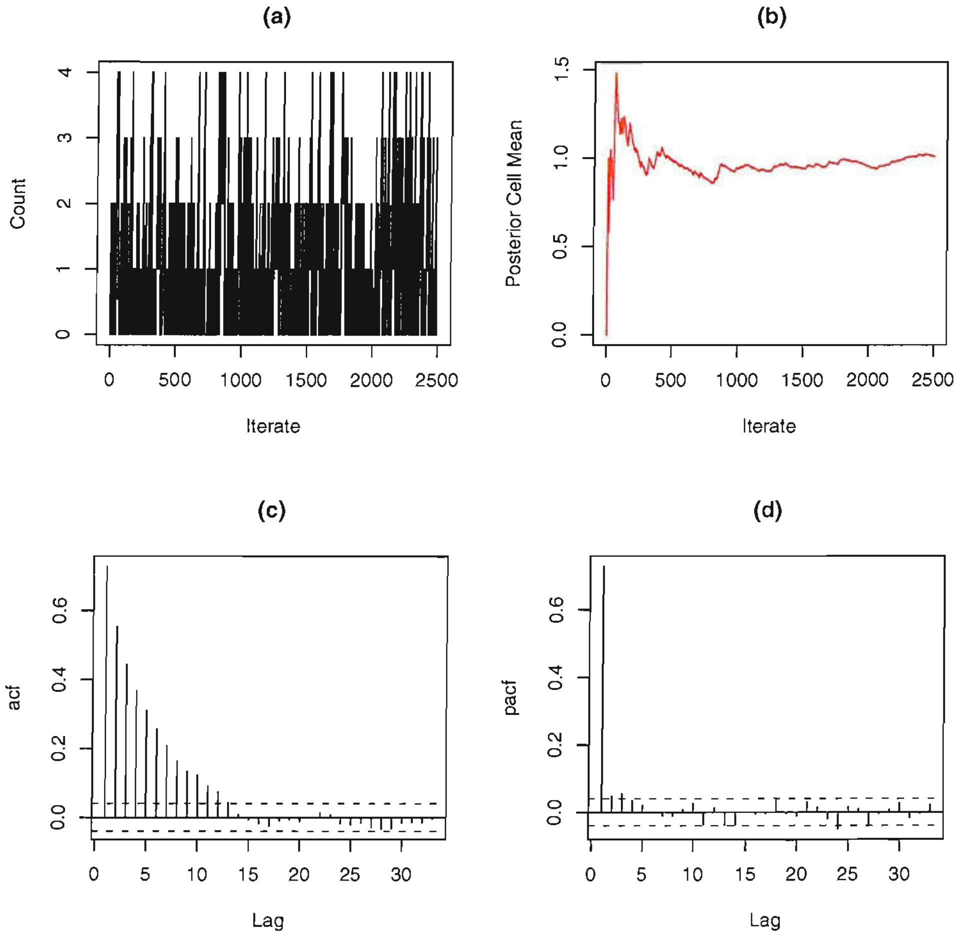


Figure 5.3: MCMC output plots. Artificial disclosure example (5.2), female < 20 cell. Plot (a): Trace [plot of cell count. (b): Trace plot of cell posterior mean. (c): Auto-correllelogram of cell count Markov chain iterates. (d): Partial auto-correllelogram of cell count Markov chain iterates.

observation is used. All plots illustrate good mixing with no sign of poor convergence.

Table 5.4 gives the posterior bounds according to cells with a posterior probability greater than 0.05. In this example, posterior bounds are reasonably wide and this

Table 5.4: Posterior bounds for the simple disclosure example, cell counts with greater than 5% posterior support.

		Age						
		< 20	$20-29$	$30-39$	$40-49$	$50-59$	60+	Total
Sex	Female	[0,3]	[0,3]	[0,6]	[0,3]	[0,6]	[22,30]	35
	Male	[0,3]	[0,3]	[0,5]	[0,3]	[0,5]	[0,7]	10
Total		0	0	5	0	5	30	45

can partly be attributed to large number of individuals in the table (a minimum of 41).

5.5 Incorporating Prior Information

Prior information in the form of a released table at a higher geographical level may be available and easily incorporated into any analysis. Returning to example (5.2), suppose Table 5.5, consisting of individuals receiving benefit B in authority A, was also released (here ward W is contained in authority A). Although the margins of Table 5.5 have also been rounded, the large cell counts in the table mean that the effect of this rounding is negligible when considering relative marginal proportions. An informative prior for the cell counts can be constructed in the following hierarchical way. At the first stage the x_i are assumed to have independent Poisson distributions with mean $m\pi_i$. At the second stage m is assumed to be uniformly distributed on the interval $[0, M]$, for M large. A prior density for $\boldsymbol{\pi} = (\pi_1, ..., \pi_n)$ is given by

$$f(\boldsymbol{\pi}) \propto \prod_{i=1}^{p} (D\boldsymbol{\pi})_i^{\alpha_i} \prod_{i=1}^{n} \pi_i^{u_i - 1} \tag{5.10}$$

Table 5.5: A Disclosure Example with information at a higher geographical level.

		Age						
		< 20	$20-29$	$30-39$	$40-49$	$50-59$	60+	Total
Sex	Female							4620
	Male							2240
Total		185	1030	1120	820	780	2915	6855

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)$ reflect prior belief concerning the relative sizes of the released margins, obtained from the authority level data. $\boldsymbol{u} = (u_1, \ldots, u_n)$ reflect prior belief concerning the relative sizes of the cells. There is little information concerning the sizes of the cells in data from the authority level. For this reason we select u_i to be 1 resulting in a non-informative uniform Dirichlet prior. The overall prior density mimics a multinomial likelihood, with the α_i parameters representing 'prior counts' in the released margins. The overall magnitude of the α_i parameters reflects strength of prior belief. As we do not expect a ward to exactly reflect the authority, the values of the α_i parameters are generally set to be smaller than the released district-level counts, but with the relative values preserved, at least approximately. If the α_i are given integer values, with consistent sums over overlapping margins, then computation with this prior is particularly straightforward. It can be set up as a missing data problem where the α_i are thought of as aggregated prior cell counts, with the actual prior cell counts included in a MCMC sampling scheme. Let

$$\boldsymbol{D}\boldsymbol{z} = \boldsymbol{\alpha} \tag{5.11}$$

for prior cell counts $\boldsymbol{z} = (z_1, ..., z_n)$. We are interested in the joint posterior $f(\boldsymbol{x}, \boldsymbol{z}, m, \boldsymbol{\pi}|\boldsymbol{y})$ and margins thereof. This posterior density is given by

$$\begin{aligned} f(\boldsymbol{x}, \boldsymbol{z}, m, \boldsymbol{\pi} | \boldsymbol{y}) &\propto f(\boldsymbol{y}|\boldsymbol{x}) f(\boldsymbol{x}|m, \boldsymbol{\pi}) f(m) f(\boldsymbol{\pi}) f(\boldsymbol{z}|\boldsymbol{\pi}) \\ &\propto f(\boldsymbol{y}|\boldsymbol{x}) \exp(m) \left[\prod_{i=1}^{n} \frac{(m\pi_i)^{x_i}}{z_i! x_i!} \pi_i^{u_i + z_i - 1} \right] \\ &\times \boldsymbol{I}[\boldsymbol{D}\boldsymbol{z} = \boldsymbol{\alpha}] \boldsymbol{I}[m \leq M]. \end{aligned}$$

This posterior density is not tractable, therefore MCMC is used to generate from the posterior. Given current parameters $\boldsymbol{x}, \boldsymbol{\pi}, m$ and $\boldsymbol{z}$ with data $\boldsymbol{y}$ and aggregated prior counts $\boldsymbol{\alpha}$ our sampling scheme, in four stages, is as follows.

Stage 1: Update $\boldsymbol{x}$

Propose a value $\boldsymbol{x}^\star$ from the proposal distribution described in Section 5.7. Then, the proposal is accepted with probability

$$\alpha(\boldsymbol{x}^\star, \boldsymbol{x}) = \frac{f(\boldsymbol{y}|\boldsymbol{x}^\star) f(\boldsymbol{x}^\star|m, \boldsymbol{\pi}) q(\boldsymbol{x}|\boldsymbol{x}^\star)}{f(\boldsymbol{y}|\boldsymbol{x}) f(\boldsymbol{x}|m, \boldsymbol{\pi}) q(\boldsymbol{x}^\star|\boldsymbol{x})} \wedge 1$$

and rejected otherwise. $f(\boldsymbol{y}|\boldsymbol{x})$ is given by (5.3), whilst $f(\boldsymbol{x}|m, \boldsymbol{\pi})$ is the likelihood of a Poisson distribution.

Stage 2: Update m

Propose a value $m^\star = m \pm 1$. Then, the proposal is accepted with probability

$$\alpha(m^\star, m) = \frac{f(\boldsymbol{x}|m^\star, \boldsymbol{\pi})}{f(\boldsymbol{x}|m, \boldsymbol{\pi})} \mathbb{I}_{m^\star \geq 0} \wedge 1$$

and rejected otherwise. Note $f(m|\boldsymbol{x}, \boldsymbol{\pi}) \propto f(\boldsymbol{x}|m, \boldsymbol{\pi})$ and $\mathbb{I}_{m \geq 0}$ is equal to 1 if $m \geq 0$ and 0 otherwise.

Stage 3: Update $\boldsymbol{\pi}$

The conditional distribution of $\boldsymbol{\pi}|\boldsymbol{x}, \boldsymbol{z}, m, \boldsymbol{y}$ is independent of $m, \boldsymbol{y}$ and has a Dirichlet distribution with parameters $\boldsymbol{x} + \boldsymbol{z} + \boldsymbol{u}$. Sample $\boldsymbol{\pi}$ from this distribution

Stage 4: Update $\boldsymbol{z}$

The prior counts $\boldsymbol{z}$ have a product multinomial distribution with parameters derived from $\boldsymbol{\alpha}$ and $\boldsymbol{\pi}$. $\boldsymbol{z}$ must be generated to satisfy the marginal constraints implied by $\boldsymbol{\alpha}$.

The above approach samples directly from the exact conditional distribution if known (Step 3), else a Metropolis-Hastings step is performed. This approach was applied to data of example 5.2 using the rounded ward margins displayed in Table 5.5. Since the margins do not overlap we divide $\sum_{i=1}^{p} \alpha_i$ equally between the 2 one-way margins, age and sex. The α_i are chosen such that the marginal probabilities, at ward level, are preserved. The posterior distributions for the cell 'Male $50 - 59$' are displayed in Figure 5.4. All results are based upon a run of 250,000 iterates storing every fifth. The black line represents the posterior distribution using the non-informative prior distribution given in (5.4). All other lines were obtained using the informative prior distribution described above. The red line represents the posterior distribution with all prior aggregated cell counts set to 0 ($\alpha_i = 0$). This is also a non-informative prior distribution hence the closeness of the black and red lines. The green, blue and purple lines were obtained using values of $\sum_{i=1}^{p} \alpha_i$ set to 10, 20 and 40 respectively. Figure 5.4 illustrates the effect of placing a strong prior distribution on the cell counts. We clearly see the posterior mode moving towards the center of the bounds, and the posterior probability of observing a zero change from 0.13 to 0.005 as $\sum_{i=1}^{p} \alpha_i$ goes from 0 to 40. For this example it is known that there are at least 41 individuals in the ward. In selecting $\sum_{i=1}^{p} \alpha_i$ to be close to 40 we are giving equal weight to prior and observed data. As we would not expect a ward to exactly reflect the district, the values of α_i should in general be smaller than the released ward level counts.

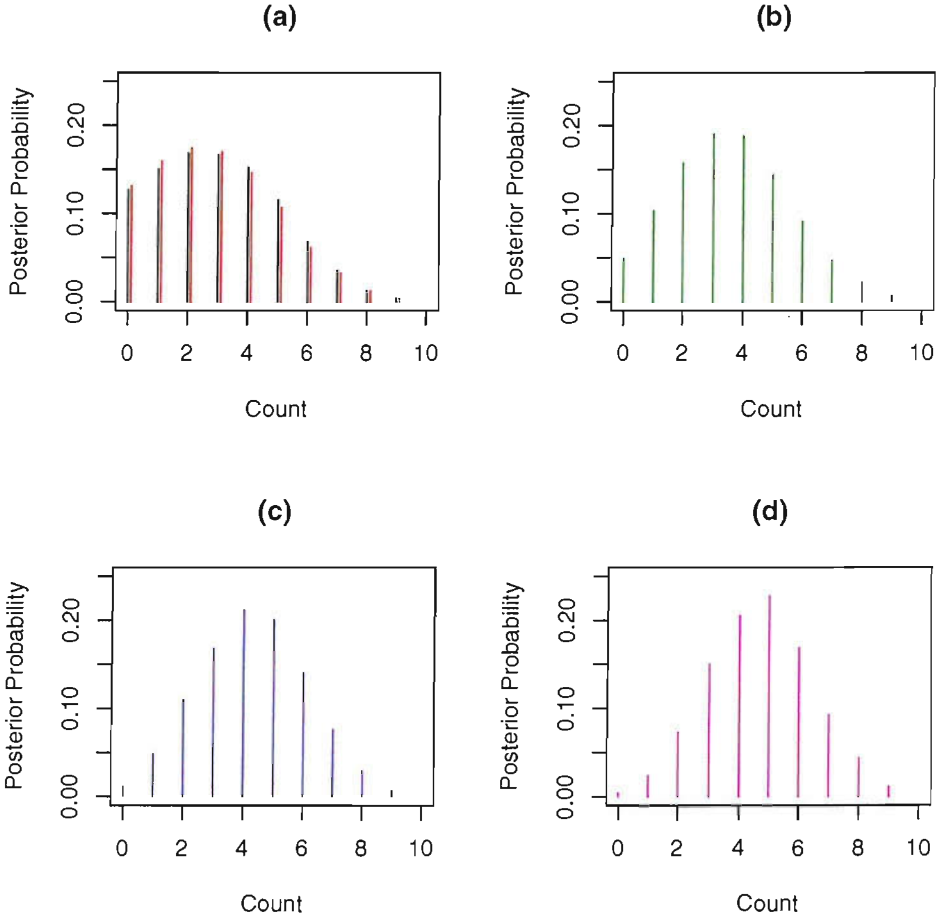


Figure 5.4: Sensitivity to changes of the prior distribution of the posterior cell count distribution, males aged 50 – 59. Plot (a): Non-informative prior distributions. Plot (b): Informative prior distribution with $\sum_{i=1}^{p} \alpha_i = 10$. Plot (c): Informative prior distribution with $\sum_{i=1}^{p} \alpha_i = 20$. Plot (d): Informative prior distribution with $\sum_{i=1}^{p} \alpha_i = 40$.

5.6 A Real Example

Consider the data available on the Neighbourhood Statistics website concerning Income Support claimants in 2000 for the Barningham and Ovington ward (Local

Authority Teesdale). The rounded counts are summarised in Table 5.6. We assume the non-informative prior described in (5.4). The posterior distributions for the cells corresponding to coupled males, are displayed in Figure 5.5 (black line). The posterior distributions for the remaining 18 cells can be found in Appendix C. It can be seen immediately that for many cells there is much more information available from the posterior distributions than from the bounds alone. In particular for 5 of the 6 cells of Figure 5.5 the probability of a zero is approximately 0.9. Although the bounds indicate that the count in these cells could be as high as 4, the probability that it is greater than 1 is negligible (less than 1%). In this example, this behaviour can be partly attributed to the table being sparse. If only the rounded total (10) was released, the marginal posterior probability of any cell being zero, based on the same prior, can be calculated exactly to be 0.640. Hence, there is significant concentration of posterior probability at zero, due to the fact that we know there are relatively few individuals distributed through a larger number (24) of cells.

For the informative prior detailed in (5.5) we considered the priors with $\sum \alpha_i$ values of (1,1,1) for the age, gender and family by working age margins. We also considered a stronger prior where these values were doubled. We would not advocate increasing the value of $\boldsymbol{\alpha}$ above 6, the lower bound for the number of individuals in the table. There is more information concerning cell counts in the rounded marginal counts at ward level, than the rounded marginal counts at local authority level, and our prior should reflect this. The posterior distributions for these cells are also displayed in (5.5). The informative prior is represented by the red and green lines. It can be seen easily that the more informative prior has little impact on the posterior inferences. This is due to the fact that information is only available concerning margins, and that information concerning the interior of the table would need to be available for the prior to have a large impact.

Table 5.6: Income support claimants in 2000 for the Barningham and Ovington ward. The local authority count for Teesdale is given in brackets.

		Age						
		< 20	$20 - 29$	$30 - 39$	$40 - 49$	$50 - 59$	60+	Total
Sex	Female							5 (900)
	Male							0 (495)
Total		0 (20)	0 (125)	0 (165)	0 (130)	0 (155)	10 (805)	10

		Family	
		Single	Couple
Age	≤ 60	0 (485)	0 (110)
	≥ 60	5 (660)	0 (145)
		10 (1140)	0 (260)

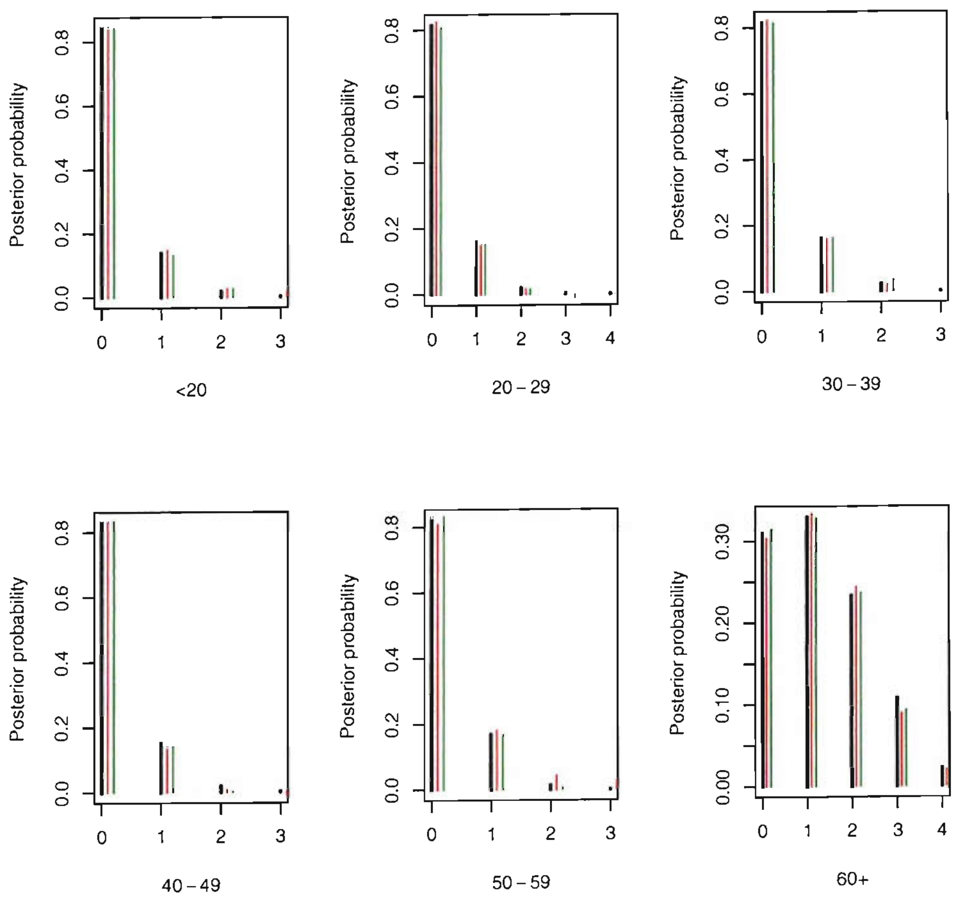


Figure 5.5: Marginal posterior cell counts for probabilities, male/couple in Barningham and Ovington.

5.7 Markov Bases and Irreducibility

In this section we construct the set of moves required such that the Metropolis-Hastings algorithm of Section 5.3.2 is irreducible. We begin by considering the case where margins, forming a decomposable graph, have been rounded to base 1. We then consider rounding to base $b > 1$. Throughout this section we shall assume that any table in the support of the likelihood is also given prior support.

5.7.1 Rounding to Base $b = 1$ (Dobra, 2003)

The case for $b = 1$ was proven by Dobra (2003) and we therefore change our notation to describe his result. We define a table of counts $\boldsymbol{n}$ as a k-dimensional array of non-negative integers. Each variable X_j, $j = 1, ..., k$, can take a finite number of values $x_j \in \mathcal{I}_j = \{1, 2, ..., I_j\}$. If we let $\mathcal{I} = \mathcal{I}_1 \times ... \times \mathcal{I}_k$, then a cell entry $n(i)$ for $i \in \mathcal{I}$ is the number (a non-negative integer) of individuals or units sharing the same attributes i. We denote $\bar{\boldsymbol{n}}$ a linear ordered list of these counts with respect to some ordering. Let $D = \{i_1, ..., i_l\}$ be an arbitrary subset of $K = \{1, ..., k\}$, then $\boldsymbol{n}_D$ forms a margin of $\boldsymbol{n}$ with cells $i_D \in \mathcal{I}_D = \mathcal{I}_{i_1} \times ... \times \mathcal{I}_{i_l}$ and counts

$$n_D(i_D) = \sum_{i \in \mathcal{I}_{K \setminus D}} n(i_D, i).$$

Margins D_1 and D_2 are overlapping if $D_1 \cap D_2 \neq \phi$, where ϕ is the empty set. Otherwise the margins are said to be non-overlapping. Clearly for k-dimensional tables $\boldsymbol{n}$ and $\boldsymbol{n}'$, both over X_j for $j = 1, ..., k$ and margin $D \subset K$, we have

$$(\boldsymbol{n} + \boldsymbol{n}')_D = \boldsymbol{n}_D + \boldsymbol{n}'_D.$$

Furthermore if all counts of table $\boldsymbol{n}$ are zero then marginal counts of $\boldsymbol{n}_D$ are zero.

Definition 5.1. *A Data Swap is an array $\boldsymbol{f} = f(i)_{i \in \mathcal{I}}$ containing integer entries. That is $f(i) \in \mathbb{Z}$ for all $i \in \mathcal{I}$. A data swap is not necessarily a table of counts as we allow $f(i) < 0$ for any i.*

Definition 5.2. *Let $D_1, ..., D_r$ be subsets of K. A Data Move $\boldsymbol{f}$ is a Data Swap that preserves the marginal tables specified by the index sets. That is*

$$\boldsymbol{f}_{D_j} = \boldsymbol{0} \;\forall j \in \{1, ..., r\}.$$

A Primitive Data Move has two entries equal to 1, two entries equal to -1, with the remaining entries being 0.

Denote $T(D_1, ..., D_r)$ the set of all tables that have their $D_1, ..., D_r$ marginals equal to the corresponding marginals of $\boldsymbol{n}$. A data move is defined as admissible if $\boldsymbol{n} \in T(D_1, ..., D_r) \Rightarrow (\boldsymbol{n}+\boldsymbol{f}) \in T(D_1, ..., D_r)$ and $(\boldsymbol{n}+\boldsymbol{f})(i) \geq 0$ for all $i \in \mathcal{I}$. Admissible moves maintain marginal counts and ensures all table counts are non negative.

Definition 5.3. *A Markov Basis $M_{D_1,...,D_r}$ is a finite collection of moves that preserve the $\{D_1, ..., D_r\}$ marginals and connect any two tables that have the same $\{D_1, ..., D_r\}$. In other words, for any table $\boldsymbol{n}' \in T(D_1, ..., D_r)$, there exist a sequence of data moves $\boldsymbol{f}_1, ..., \boldsymbol{f}_S$ such that*

$$\boldsymbol{n}' - \boldsymbol{n} = \sum_{j=1}^{s} \boldsymbol{f}_j,$$

and is such that

$$\boldsymbol{n} + \sum_{j=1}^{s'} \boldsymbol{f}_j \in T^{(n)}(D_1, ..., D_r)$$

for $1 \leq s' \leq s$. Since M depends only on the index set $\mathcal{I}_{D_1}, ..., \mathcal{I}_{D_r}$ we say that M is a Markov basis for $T(D_1, ..., D_r)$, where $T(D_1, ..., D_r)$ is the set of tables with corresponding marginal counts $\boldsymbol{n}_{D_1}, ..., \boldsymbol{n}_{D_r}$

A Markov basis for $T(D_1, ..., D_r)$ connects any two tables with fixed marginal counts $D_1, ..., D_r$ through a path of tables all contained in $T(D_1, ..., D_r)$.

To construct this Markov Basis Diaconis and Sturmfels (1998) introduced an indeterminate for each cell $n(i)$, and formed the ring of polynomials in these indeterminates over some field. They then showed the mapping between the cell and marginal counts forms a ring homomorphism. Let φ denote this mapping, and let $\mathcal{K}_\varphi$ be the set of tables $\boldsymbol{n}$ (elements of the ring) such that $\varphi(\boldsymbol{n}) = 0$. Then $\mathcal{K}_\varphi$ is the kernel of φ. Noting that the kernel of a ring homomorphism is an ideal, Diaconis and Sturmfels (1998) showed this ideal to be a Markov Basis for $T(D_1, ..., D_r)$. The Hilbert Basis theorem states this ideal exists and is finitely generated. However the union of pairwise differences of all tables in $T(D_1, ..., D_r)$ is finite and forms a non minimal Markov Basis for $T(D_1, ..., D_r)$. Existence and finite generation is given. Computing a Markov Basis (finding this ideal) is very difficult (relying on results from computational algebra, Toric ideals and Gröbner bases) and computationally expensive for multiway contingency tables. The basis must be constructed prior to running the Markov chain. We refer the reader to Diaconis and Sturmfels (1998) for further details of the construction.

Dobra (2003) provided an alternative construction of this Markov Basis, and showed admissible primitive moves could be dynamically (within chain) generated. We consider his approach in the remainder of this section before looking at applications to rounding based disclosure control.

5.7.2 Decomposable Graphs

We introduce some basic graph theory used throughout the remainder of this section. It is assumed that $K = D_1 \cup ... \cup D_r$ (All variables appear in at least one margin), and $D_{r_1} \not\subseteq D_{r_2}$ for all r_1 and r_2. A graph $G = (V, E)$ has vertex set V and edge set

$$E := \{(u, v) : \{u, v\} \subset D_j \text{ for some } j\}.$$

We say that a set of vertices of G forms a complete subgraph of G if every pair of vertices in the set are connected by an edge. A graph G is complete if every pair of vertices is connected by an edge. A subset $U \in V$ is called a clique of G if it is maximally complete, i.e. if U is complete and if $U \subset W$, then W is not complete. We denote by $C(G)$ the set of cliques of a graph.

Definition 5.4. *A triple (A, B, C) of disjoint subsets of V is said to form a decomposition of the graph $G = (V, E)$ if*

1. *C separates A from B.*
2. *C is a complete subset of V.*

By recursively applying the above definition we can define a decomposable graph.

Definition 5.5. *A graph $G = (V, E)$ is decomposable if it is complete or if there exists a decomposition (A, B, C) into decomposable subgraphs $G_{A\cup C}$ and $G_{B\cup C}$*

This recursive definition ensures that a decomposable graph is one which can be successively decomposed into its cliques. We say the set $\boldsymbol{n}_{D_1}, ..., \boldsymbol{n}_{D_r}$ of marginals is decomposable if the corresponding graph $G = G(D_1, ..., D_r)$ is decomposable. The cliques of G denoted $C(G) = \{D_1, ..., D_r\}$ form the minimal sufficient statistics of log-linear models.

5.7.3 The Simplest Decomposable Graph

The simplest decomposable graph has two vertices and no edges. This graph is the independence graph associated with the 2 one-way marginal counts of a two-way table. An example of such a table is given in Table 5.2. Let $\boldsymbol{n} = \{n(i,j), (i,j) \in \mathcal{I}_1 \times \mathcal{I}_2\}$ be a two-way contingency table. Let $\boldsymbol{f} = \{f_{i_1,i_2,j_1,j_2} : i_1 \neq i_2 \in \mathcal{I}_1, j_1 \neq j_2 \in \mathcal{I}_2\}$ be a primitive move defined by

$$f_{i_1,i_2,j_1,j_2}(i,j) = \begin{cases} -1 \text{ if } (i,j) \in \{(i_1,j_1),(i_2,j_2)\} \\ 1 \text{ if } (i,j) \in \{(i_1,j_2),(i_2,j_1)\} \\ 0 \text{ otherwise.} \end{cases} \quad (5.12)$$

Then the set of all the above moves with $1 \leq i_1 < i_2 \leq I_1$ and $1 \leq j_1 < j_2 \leq I_2$ is a Markov Basis, as defined in (5.3), for the class of tables with fixed row and column sums. The proof of this can be found in Diaconis and Sturmfels (1998). Of course we require more than just these moves to construct a Markov Chain on Table 5.2, since row and column sums have been rounded.

In the following section we outline a proof (Dobra, 2003) that the set of primitive moves are the only moves that have to be included in a Markov basis that links all tables with fixed marginal counts where the margins induce a decomposable graph. We provide simple and informative examples to demonstrate this method. The proof is by induction, with the simple case described above the initial or base case.

5.7.4 Dobra (2003)

Consider a k-way contingency table with two fixed marginals $\boldsymbol{n}_{D_1}$ and $\boldsymbol{n}_{D_2}$. We assume the graph $G(D_1, D_2)$ is a decomposable graph.

If the margins are non-overlapping ($D_1 \cap D_2 = \emptyset$), the graph has vertex set $\{D_1, D_2\}$ and is edgeless. In this case we introduce two new variables Y_1 and Y_2 with level sets $\mathcal{I}_{D_1}$ and $\mathcal{I}_{D_2}$ respectively. The two way table that cross-classifies Y_1 and Y_2 has fixed row sums $\bar{\boldsymbol{n}}_{D_1}$ and column sums $\bar{\boldsymbol{n}}_{D_2}$. Thus the set of moves described by Diaconis and Sturmfels (1998) and given in (5.12) for these fixed row and column sums is indeed a Markov Basis for $T(D_1, D_2)$.

If the margins are overlapping ($D_1 \cap D_2 \neq \emptyset$) then for each $j_{D_1\cap D_2} \in \mathcal{I}_{D_1\cap D_2}$ define a new two-way table

$$\boldsymbol{n}^{j_{D_1\cap D_2}} = \{n^{j_{D_1\cap D_2}}(i)\}_{i\in K\setminus D_1\cap D_2}$$

which has cell entries

$$\boldsymbol{n}^{j_{D_1\cap D_2}}(i, j_{D_1\cap D_2}) = \{n^{j_{D_1\cap D_2}}(k_{D_1\setminus D_2}, l_{D_2\setminus D_1})\},$$

for all $k \in D_1\setminus D_2$ and $l \in D_2\setminus D_1$. For each $j_{D_1\cap D_2} \in \mathcal{I}_{D_1\cap D_2}$ this table has two fixed non-overlapping marginals $\boldsymbol{n}^{j_{D_1\cap D_2}}_{D_1\setminus D_2}$ and $\boldsymbol{n}^{j_{D_1\cap D_2}}_{D_2\setminus D_1}$. We have shown how to construct a Markov Basis for $M_{\boldsymbol{n}^{j_{D_1\cap D_2}}_{D_1\setminus D_2}, \boldsymbol{n}^{j_{D_1\cap D_2}}_{D_2\setminus D_1}}$ for $T(\boldsymbol{n}^{j_{D_1\cap D_2}})$ that preserves marginals $\boldsymbol{n}^{j_{D_1\cap D_2}}_{D_1\setminus D_2}$ and $\boldsymbol{n}^{j_{D_1\cap D_2}}_{D_2\setminus D_1}$. It follows therefore that a M_{D_1,D_2} for the set of tables $T(D_1, D_2)$ with fixed marginals D_1, D_2 is given by

$$M_{D_1,D_2} = \bigcup_{j_{D_1\cap D_2}\in\mathcal{I}_{D_1\cap D_2}} M_{\boldsymbol{n}^{j_{D_1\cap D_2}}_{D_1\setminus D_2}, \boldsymbol{n}^{j_{D_1\cap D_2}}_{D_2\setminus D_1}}. \tag{5.13}$$

To illustrate the above idea we introduce a simple example. Consider the $2 \times 2 \times 2$ table that cross-classifies individuals by sex (male or female), age (< 60 or 60+) and marital status (married or single). Firstly, suppose we release the two-way margin sex by age, and the one way margin marital status. A typical element of the Markov Basis for these released margins in given below.

	Married	Single
Male < 60	-1	1
Male 60+		
Female < 60	1	-1
Female 60+		

Such an element can be easily generated within the Markov chain. We would simply select two rows from the above table, and create the corresponding primitive move. Since the row and column totals will be maintained, the marginal counts of the two-way margin sex by age, and the one way margin marital status will also be maintained. There are 6 ways of choosing two elements from 4, hence the Markov basis has 12 elements.

Secondly, suppose we release the two-way margins sex by age and sex by marital status. Then, the Markov Basis for this set of overlapping margins is formed in the following way. Create two new two-way tables, one for males and the other for females, of the cross-classification of age by marital status given below

	Married	Single
< 60	1	-1
60+	-1	1

Each of these tables has two fixed non-overlapping one way margins and it is thus easy to form a Markov Basis for each individual table. A typical element of this

Markov Basis is given above. The Markov Basis for the released margins is formed as the union of the Markov Bases for both male and female tables. It has 4 elements.

Dobra's main result is proven by induction on the number of cliques of a decomposable graph, with the above result forming the initial case. We shall state, without proof, this result in due course.

Definition 5.6. *A tree is a connected and undirected graph without cycles. There is a unique path between each pair of vertices on a tree.*

Definition 5.7. *Let $\mathcal{T} = (C(G), E_{\mathcal{T}})$ be a tree defined on the cliques of a graph G. Let $S = D_i \cap D_j$ for some $(D_i, D_j) \in E_{\mathcal{T}}$. Let $\mathcal{T}_i = (\mathcal{K}_i, E_i)$ and $\mathcal{T}_j = (\mathcal{K}_j, E_j)$ be the two subtrees obtained by removing the edge (D_i, D_j) from the tree $\mathcal{T}$ with $D_i \in \mathcal{K}_i$ and $D_j \in \mathcal{K}_j$. Consider the vertex sets*

$$V_i = \bigcup_{D \in \mathcal{K}_i} D \text{ and } V_j = \bigcup_{D \in \mathcal{K}_j} D. \tag{5.14}$$

Then the tree $\mathcal{T}$ is said to have the star property for G and is called a junction tree if for every edge $(D_i, D_j) \in E_{\mathcal{T}}$, $(V_i \setminus S, V_j \setminus S, S)$ is a decomposition of G

Blair and Barry (1993) proved that a graph G is decomposable if and only if there exist a tree on the cliques $C(G)$ of G for which the star property holds. The following theorem is Dobra's main result.

Theorem 5.1. Let $C(G) = \{D_1, ..., D_r\}$ be the set of cliques of a decomposable graph. We let $\mathcal{T} = (C(G), E_{\mathcal{T}})$ be a tree having the star property on the set of cliques G. For every edge $(D_i, D_j) \in E_{\mathcal{T}}$ consider the vertex sets defined in (5.14). Then the set of primitive moves associated with the decomposable graph G forming a Markov basis for $M_{\{D_1,...,D_r\}}$ is given by

$$M_G = M_{\{D_1,...,D_r\}} = \bigcup_{(D_i,D_j) \in E_{\mathcal{T}}} M(V_i, V_j), \tag{5.15}$$

where $M(V_i, V_j)$ is given by (5.13)

The theorem states that it is sufficient to partition any tree with the required properties in two pieces, that are further considered to induce a complete subgraph of G, for the purpose of generating a basis. We treat these two pieces as the vertices of an edgeless graph. We construct a Markov Basis for this edgeless graph using (5.13). Essentially (5.15) says that the set of primitive moves for a decomposable model with graph G is the union of the sets of primitive moves obtained from the two clique models induce by removing an edge $(D_i, D_j) \in E_T$.

5.7.5 Rounding to Base $b > 1$

The most trivial method for constructing an irreducible Markov chain is as follows. For each $i \in \mathcal{I}$ generate two random variables. The first, denoted u_i, is generated uniformly from the set $\{-1, 1\}$. The second, denoted k_i, is generated from a Poisson distribution with mean given by λ. The proposed move (table) is now given simply by

$$\boldsymbol{n}' = \boldsymbol{n} + \boldsymbol{f},$$

where $\boldsymbol{n}$ is the current state (table) of the Markov chain and $f(i) = u_i k_i \ \forall i \in \mathcal{I}$. Since $|\boldsymbol{n}' - \boldsymbol{n}| = |\boldsymbol{n} - \boldsymbol{n}'|$ the proposal probabilities q cancel in the acceptance probability. If $n'(i) < 0$ for any $i \in \mathcal{I}$ then this proposed move is immediately rejected.

The resulting moves differ substantially from the primitive moves described by Dobra (2003). This is the result of relaxing the decomposability condition. If we consider rounding to base $b = 1$ and relax the decomposability condition then the resulting Markov basis has a very different from to that described by Dobra (2003) (c.f.

Diaconis and Sturmfels (1998)). It is important to note that large λ may result in a low acceptance rate of proposed moves and chain that mixes poorly. Clearly if λ is small then we propose moves close to the primitive moves described above.

We can conjecture the type of moves the finite minimal Markov basis might contain by considering two tables, again denoted $\boldsymbol{n}$ and $\boldsymbol{n}'$, that differ in both marginal and total counts yet satisfy the same rounded bounds. In addition to the moves defined by Dobra we might expect two additional move types. Firstly, since the two tables differ in total one might expect the minimal basis to contain moves that alters a single table count by one. Secondly, we might also expect the minimal basis to contain moves that alter the marginal counts but maintain the table total. Semi-primitive moves where two table counts are altered by one (we simply add one to a count and subtract one from a different count) might complete the basis. However, proving this result has been elusive.

To prove irreducibility, using only the three above move types, we are required to construct a path from a table $\boldsymbol{n} \in T(D_1, ..., D_r)$ to a table $\boldsymbol{n}' \in T(D_1', ..., D_r')$. We assume these tables differ in both marginal totals and table total but satisfies the rounding bounds. Any table that lies in the constructed path from $\boldsymbol{n}$ to $\boldsymbol{n}'$ must also satisfy the rounding bounds. The first stage of a proof might be to take two tables with differing totals and show a move between them is always possible. If two tables have the same table total we say these tables are 'total equivalent' and belong to the equivalence class $[s]$, where s is the table total count. It would be enough to show that if $\boldsymbol{n} \in [s]$ and $[s-1] \neq \emptyset$ then we can construct a move from $\boldsymbol{n}$ to a table $\boldsymbol{n}' \in [s-1]$. We only need show one direction as reversibility of the Metropolis-Hastings algorithm ensures the reverse move is always possible. Consider the following two examples where the rounding base is assumed to be 5. The two examples show that it is not sufficient to augment simple ±1 moves to the primitive moves of Dobra (2003). However, it must be noted that the method proposed within this chapter still works in both cases.

The left hand table below, Table 5.7, represents the released marginal counts of a two-way table. The figures in brackets are the lower bounds of the marginal counts. The right hand table is the current state of the Markov chain.

Table 5.7: An example illustrating the difficulties of constructing a Markov basis.

	10 (6)	5 (1)
10 (6)		
5 (1)		

	6	2
6	4	2
2	2	0

It is clearly not possible to subtract a count from any of the four cells without violating rounding constraints or creating a negative cell count. We can add a single count to any of the four cells. However, if we add the following primitive move

	0	0
0	1	-1
0	-1	1

to form the table

	6	2
6	5	1
2	1	1

then we can subtract 1 from the bottom right hand cell to form a new table that does not violate the rounding constraints.

It would also be sufficient to show a move from a table $\boldsymbol{n} \in [s]$ to a table $\boldsymbol{n}' \in [s+1]$ is always possible. The following is a counter example. Consider the released rounded margins of the $2 \times 2 \times 2 \times 2$ given in Table 5.8 below.

Table 5.8: A second example illustrating the difficulties of constructing a Markov basis.

	A_1	A_2
B_1	0	0
B_2	0	15

	B_1	B_2
C_1	5	0
C_2	0	15

	C_1	C_2
D_1	0	5
D_2	5	5

Then the following table satisfies all marginal rounding constraints. The table total is 26.

		B_1		B_1	
		D_1	D_2	D_1	D_2
A_1	C_1	1	1	1	1
	C_2	1	1	1	1
A_2	C_1	1	1	1	1
	C_2	1	1	6	6

It is not possible to add a single count to any of the cells without violating the rounding constraints. However the following table

		B_1		B_1	
		D_1	D_2	D_1	D_2
A_1	C_1	1	2	1	0
	C_2	0	1	1	1
A_2	C_1	1	1	1	1
	C_2	1	2	7	6

has a total count of 27 and satisfies the rounding constraints. Clearly the difference between the two tables is a move that could have been generated by the Poisson moves described in this section.

5.7.6 Algorithmic Implementation

We adopted the following algorithm which resulted in Markov chains which exhibited adequate mixing. At each iteration one of the following five moves was proposed

Move 1

For each $i \in \mathcal{I}$ generate two random variables. The first, denoted u_i, is generated uniformly from the set $\{-1, 1\}$. The second, denoted k_i, is generated from a Poisson distribution with mean given by $\lambda = \frac{4}{\sum_{i \in \mathcal{I}} 1}$. Define a table $\boldsymbol{f}$ where $f(i) = u_i k_i$. Then the proposed move (table) is now given simply by $\boldsymbol{n}' = \boldsymbol{n} + \boldsymbol{f}$.

Move 2

For each $i \in \mathcal{I}$ generate two random variables. The first, denoted u_i, is generated uniformly from the set $\{-1, 1\}$. The second, denoted k_i, is generated from a Poisson distribution with mean given by $\lambda = 1$. Define a table $\boldsymbol{f}$ where $f(i) = u_i k_i$. Then the proposed move (table) is now given simply by $\boldsymbol{n}' = \boldsymbol{n} + \boldsymbol{f}$.

Move 3

Generate i_1, i_2, i_3, i_4 uniformly from the set of all counts $\mathcal{I}_1 \times ... \times \mathcal{I}_k$ without replacement. Then the proposed move (table) is now given simply by $\boldsymbol{n}' = \boldsymbol{n} + \boldsymbol{f}$ where

$$f(i) = \begin{cases} -1 \text{ if } i = i_1, i_2 \\ 1 \text{ if } i = i_3, i_4 \\ 0 \text{ otherwise.} \end{cases} \tag{5.16}$$

Move 4

Generate i_1, i_2 uniformly from the set of all counts $\mathcal{I}_1 \times ... \times \mathcal{I}_k$ without replacement. Then the proposed move (table) is now given simply by $\boldsymbol{n}' = \boldsymbol{n} + \boldsymbol{f}$ where

$$f(i) = \begin{cases} -1 \text{ if } i = i_1 \\ 1 \text{ if } i = i_2 \\ 0 \text{ 0therwise.} \end{cases} \tag{5.17}$$

Move 5

Generate i^* uniformly from the set of all counts $\mathcal{I}_1 \times ... \times \mathcal{I}_k$ without replacement. Generate u uniformly from the set $\{-1, 1\}$. Define a table $\boldsymbol{f}$ where $f(i^*) = u$ and zero otherwise.

The probability of moves 1, 2, 3, 4 and 5 where 0.2, 0.05, 0.25, 0.25 and 0.25 respectively. Moves 1 and 2 are differ only in the parameter λ.

5.8 Closing Remarks

The method of rounding data for release using a stochastic mechanism has been studied extensively in this chapter. We have introduced a novel Bayesian approach that can be used to quantify the disclosure risk of releasing the data. Our approach is to calculate the posterior cell probabilities using a Markov chain. This Markov chain produces a sample from the posterior distribution that can be further used to quantify measures of disclosure risk. The simple measure considered in this chapter was that of posterior bounds.

The algorithm is computationally cheap to implement, and the variety of moves described above should result in a chain with adequate mixing properties.

The construction of a minimal Markov basis for the case where no rounding had been performed (rounding to base $b = 1$) formed the second half of the chapter. We were unable to construct a minimal Markov basis for the case $b > 1$ although we conjectured its form.

Chapter 6

Summary and Future Work

In this final chapter, there are two questions to be considered. Firstly, what conclusions can be drawn from the research in this thesis. Secondly, how can the research be extended or applied to different statistical areas.

6.1 Conclusions

The aim of this thesis has been to provide a formal framework for Bayesian inference in several situations where data is partially observed.

A Bayesian framework for future prediction was considered in Chapter 3. The key result of the chapter was the construction of an efficient reversible jump Markov chain Monte Carlo algorithm for generalised linear models. The algorithm enables the formation of a posterior predictive density and therefore the incorporation of all sources of uncertainty into future predictions.

The construction of the algorithm is particularly interesting. A novel transformation function was introduced which resulted in a clear choice of proposal distribution. The resulting algorithm was not only simple to implement but also efficient and computationally inexpensive. Examples were presented to demonstrate these facts.

Bayesian inference for survey data subject to non-response was the basis of Chapter

4. An attempt to discriminate between non-response models was conducted using methods developed in Chapter 3. This attempt demonstrated that a less cavalier approach to inference was required.

Uncertainty about ignorability of non-response was then incorporated by introducing parameters into log-linear models and integrating over the prior uncertainty associated with these parameters. The appeal of the method rests with the simple elicitation of prior information required for inference.

Statistical disclosure control was examined in Chapter 5. In particular, we considered the disclosure control technique of releasing rounded margins of a multi-way contingency table. A framework for posterior prediction of missing (non released) cell counts was constructed. Our approach was to calculate the posterior cell probabilities using a Markov chain. This Markov chain produces a dependant sample from the posterior distribution which were used to quantify measures of disclosure risk. We attempted to construct a minimal Markov basis to ensure the irreducibility of this algorithm. This was not possible for the case where $b > 1$.

6.2 Future work

The work of Chapter 3 concentrated on the construction of an efficient reversible jump algorithm for generalised linear models. The ideas developed in this chapter can be applied in any modeling framework where a linear component is present. For example generalised linear models for longitudinal data or generalised linear mixed models. The results of Chapter 4 should also be applied to longitudinal data where missing data is often a common occurrence. The results pertaining to Markov bases, presented in Chapter 5, are only valid for decomposable graphical models and for rounding base $b = 1$. For $b > 1$ we conjectured the form of the minimal Markov basis. Future work will attempt to prove this conjecture.

APPENDICES

Appendix A: Block Matrices and Schur complements

Let M be the following block matrix

$$M = \begin{pmatrix} A_{n,n} & B_{n,m} \\ C_{m,n} & D_{m,m} \end{pmatrix}$$

Where $A_{n,n}$ is an $n \times n$ matrix, $B_{n,m}$ is an $n \times m$ matrix and so forth. The determinant of matrix M is given by

$$|M| = |A||D - CA^{-1}B|, \tag{1}$$

Let S_A define the Schur complement of matrix A. This is given as follows

$$S_A = D - CA^{-1}B$$

The following result for the inverse of matrix M holds

$$M^{-1} = \begin{pmatrix} A^{-1} + A^{-1}BS_A^{-1}CA^{-1} & -A^{-1}BS_A^{-1} \\ -S_A^{-1}CA^{-1} & S_A^{-1} \end{pmatrix} \tag{2}$$

Appendix B: Supplementary Results for Missing Data Chapter

Continuing with the notation of Chapter (4), section (4.2.2.3), for the 2^6 contingency table define cell means as follows

	$R_3 = 1$		$R_3 = 2$	
	$R_1 = 1$	$R_1 = 2$	$R_1 = 1$	$R_1 = 2$
$R_2 = 1$	m_{ijk}	$m_{ijk}a_{ijk}$	$m_{ijk}d_{ijk}$	$m_{ijk}d_{ijk}a_{ijk}p_{ijk}$
$R_2 = 2$	$m_{ijk}b_{ijk}$	$m_{ijk}a_{ijk}b_{ijk}g_{ijk}$	$m_{ijk}d_{ijk}b_{ijk}v_{ijk}$	$m_{ijk}a_{ijk}b_{ijk}d_{ijk}p_{ijk}v_{ijk}g_{ijk}t_{ijk}$

Of course we do not observe all $2^6 = 64$ counts. Instead we observe the counts given in table (4.1). If sum-to-zero constraints are used we have the following equivalences

m_{ijk} corresponds to $1 + Y_1 + Y_2 + Y_3 + Y_1Y_2 + Y_2Y_3 + Y_1Y_3 + Y_1Y_2Y_3$

a_{ijk} corresponds to $R_1 + R_1Y_1 + R_1Y_2 + R_1Y_3 + R_1Y_1Y_2 + R_1Y_2Y_3 + R_1Y_1Y_3 + R_1Y_1Y_2Y_3$

b_{ijk} corresponds to $R_2 + R_2Y_1 + R_2Y_2 + R_2Y_3 + R_2Y_1Y_2 + R_2Y_2Y_3 + R_2Y_1Y_3 + R_2Y_1Y_2Y_3$

d_{ijk} corresponds to $R_3 + R_3Y_1 + R_3Y_2 + R_3Y_3 + R_3Y_1Y_2 + R_3Y_2Y_3 + R_3Y_1Y_3 + R_3Y_1Y_2Y_3$

g_{ijk} corresponds to $R_1R_2 + R_1R_2Y_1 + R_1R_2Y_2 + R_1R_2Y_3 + R_1R_2Y_1Y_2 + R_1R_2Y_2Y_3 + R_1R_2Y_1Y_3 + R_1R_2Y_1Y_2Y_3$

v_{ijk} corresponds to $R_1R_3 + R_1R_3Y_1 + R_1R_3Y_2 + R_1R_3Y_3 + R_1R_3Y_1Y_2 + R_1R_3Y_2Y_3 + R_1R_3Y_1Y_3 + R_1R_3Y_1Y_2Y_3$

p_{ijk} corresponds to $R_2R_3 + R_2R_3Y_1 + R_2R_3Y_2 + R_2R_3Y_3 + R_2R_3Y_1Y_2 + R_2R_3Y_2Y_3 + R_2R_3Y_1Y_3 + R_2R_3Y_1Y_2Y_3$

t_{ijk} corresponds to $R_1R_2R_3 + R_1R_2R_3Y_1 + R_1R_2R_3Y_2 + R_1R_2R_3Y_3 + R_1R_2R_3Y_1Y_2 + R_1R_2R_3Y_2Y_3 + R_1R_2R_3Y_1Y_3 + R_1R_2R_3Y_1Y_2Y_3$

Specific Result

The following model has tractable maximum likelihood estimates.

$$Y_1Y_2Y_3 + R_1R_2R_3 + Y_1R_2R_3 + R_1Y_2R_3 + R_1R_2Y_3 + R_1Y_2Y_3 + Y_1R_2Y_3 + Y_1Y_2R_3$$

This additive notation is equivalent to specifying $m_{ijk} = m_{ijk}$, $a_{ijk} = a_{jk}$, $b_{ijk} = b_{ik}$, $d_{ijk} = d_{ij}$, $g_{ijk} = g_k$, $v_{ijk} = v_j$, $p_{ijk} = p_i$ and $t_{ijk} = t$ in multiplicative notation. The likelihood for this model is therefore given by

$$\begin{aligned}
L &\propto \prod_{ijk} \exp(-m_{ijk})(m_{ijk})^{n_{ijk111}} \\
&+ \prod_{jk} \exp(-m_{ijk}a_{jk})(m_{ijk}a_{jk})^{n_{+jk211}} \\
&+ \prod_{ik} \exp(-m_{ijk}b_{ik})(m_{ijk}b_{jk})^{n_{i+k121}} \\
&+ \prod_{ij} \exp(-m_{ijk}d_{ij})(m_{ijk}d_{ij})^{n_{ij+112}} \\
&+ \prod_{i} \exp(-\sum_{jk} m_{ijk}b_{ik}d_{ij}p_i)(\sum_{jk} m_{ijk}b_{ik}d_{ij}p_i)^{n_{i++122}} \\
&+ \prod_{j} \exp(-\sum_{ik} m_{ijk}a_{jk}d_{ij}v_j)(\sum_{ik} m_{ijk}a_{jk}d_{ij}v_j)^{n_{+j+212}} \\
&+ \prod_{k} \exp(-\sum_{ij} m_{ijk}a_{jk}b_{ik}g_k)(\sum_{ij} m_{ijk}a_{jk}b_{ik}g_k)^{n_{++k221}} \\
&+ \exp(-t\sum_{ijk} m_{ijk}a_{jk}b_{ik}d_{ij}g_kp_iv_j)(t\sum_{ijk} m_{ijk}a_{jk}b_{ik}d_{ij}g_kp_iv_j)^{n_{+++222}}
\end{aligned}$$

Differentiating with respect to t and setting to zero

$$\begin{aligned}
\frac{\partial}{\partial t}\log(L) = 0 &\Rightarrow -\sum_{ijk} \hat{m}_{ijk}\hat{a}_{jk}\hat{b}_{ik}\hat{d}_{ij}\hat{g}_k\hat{p}_i\hat{v}_j + \frac{n_{+++222}}{\hat{t}} = 0 \\
&\Rightarrow \hat{t} = \frac{\sum_{ijk} \hat{m}_{ijk}\hat{a}_{jk}\hat{b}_{ik}\hat{d}_{ij}\hat{g}_k\hat{p}_i\hat{v}_j}{n_{+++222}}
\end{aligned}$$

Continuing systematically

$$
\begin{aligned}
\frac{\partial}{\partial p_i}\log(L)=0 &\Rightarrow \frac{n_{i++122}}{\hat{p}_i}-\sum_{jk}\hat{m}_{ijk}\hat{b}_{ik}\hat{d}_{ij} \\
&- \hat{t}\sum_{ijk}\hat{m}_{ijk}\hat{a}_{jk}\hat{b}_{ik}\hat{d}_{ij}\hat{g}_k\hat{p}_i\hat{v}_j \\
&+ \frac{n_{+++222}}{\sum_{ijk}\hat{m}_{ijk}\hat{a}_{jk}\hat{b}_{ik}\hat{d}_{ij}\hat{g}_k\hat{p}_i\hat{v}_j}\sum_{ijk}\hat{m}_{ijk}\hat{a}_{jk}\hat{b}_{ik}\hat{d}_{ij}\hat{g}_k\hat{p}_i\hat{v}_j \\
&= 0 \\
&\Rightarrow \hat{p}_i=\frac{n_{i++122}}{\sum_{jk}\hat{m}_{ijk}\hat{b}_{ik}\hat{d}_{ij}} \\
\frac{\partial}{\partial v_j}\log(L)=0 &\Rightarrow \hat{v}_j=\frac{n_{+j+212}}{\sum_{ik}\hat{m}_{ijk}\hat{a}_{jk}\hat{d}_{ij}} \\
\frac{\partial}{\partial g_k}\log(L)=0 &\Rightarrow \hat{g}_k=\frac{n_{++k221}}{\sum_{ij}\hat{m}_{ijk}\hat{a}_{jk}\hat{b}_{ik}}
\end{aligned}
$$

The second and third lines cancel on another. Results on lines 6 and 7 follow through a similar argument.

$$
\begin{aligned}
\frac{\partial}{\partial a_{jk}}\log(L)=0 &\Rightarrow \frac{n_{+jk211}}{\hat{a}_{jk}}-\hat{m}_{+jk} \\
&- \hat{v}_j\sum_{i}\hat{m}_{ijk}\hat{d}_ij+\frac{n_{+j+212}}{\sum_{ik}\hat{m}_{ijk}\hat{a}_{jk}\hat{d}_{ij}}\sum_{i}\hat{m}_{ijk}\hat{d}_{ij} \\
&- \hat{g}_k\sum_{j}\hat{m}_{ijk}\hat{b}_ik+\frac{n_{++k221}}{\sum_{ij}\hat{m}_{ijk}\hat{a}_{jk}\hat{b}_{ik}} \\
&- \hat{t}\sum_{ijk}\hat{m}_{ijk}\hat{a}_{jk}\hat{b}_{ik}\hat{d}_{ij}\hat{g}_k\hat{p}_i\hat{v}_j \\
&+ \frac{n_{+++222}}{\sum_{ijk}\hat{m}_{ijk}\hat{a}_{jk}\hat{b}_{ik}\hat{d}_{ij}\hat{g}_k\hat{p}_i\hat{v}_j}\sum_{ijk}\hat{m}_{ijk}\hat{a}_{jk}\hat{b}_{ik}\hat{d}_{ij}\hat{g}_k\hat{p}_i\hat{v}_j \\
&= 0 \\
&\Rightarrow \hat{a}_{jk}=\frac{n_{+jk211}}{\hat{m}_{+jk}} \\
\frac{\partial}{\partial b_{ik}}\log(L)=0 &\Rightarrow \hat{b}_{jk}=\frac{n_{i+k121}}{\hat{m}_{i+k}} \\
\frac{\partial}{\partial d_{ij}}\log(L)=0 &\Rightarrow \hat{d}_{ij}=\frac{n_{ij+112}}{\hat{m}_{ij+}}
\end{aligned}
$$

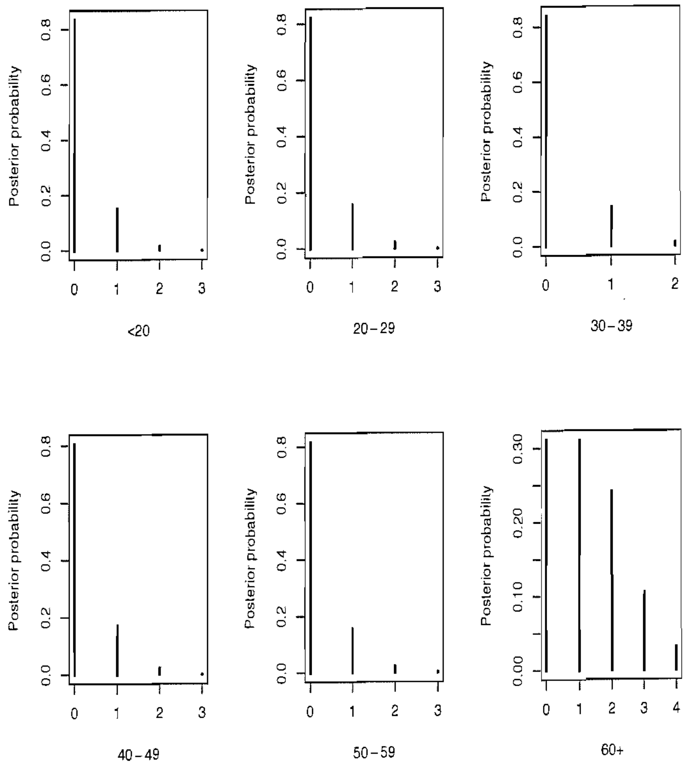
The second and third lines are zero and lines 4 and 5 cancel one another. Results on lines 8 and 9 follow through a similar argument.

$$\frac{\partial}{\partial m_{ijk}} \log(L) = 0 \Rightarrow \hat{m}_{ijk} = n_{ijk111}$$

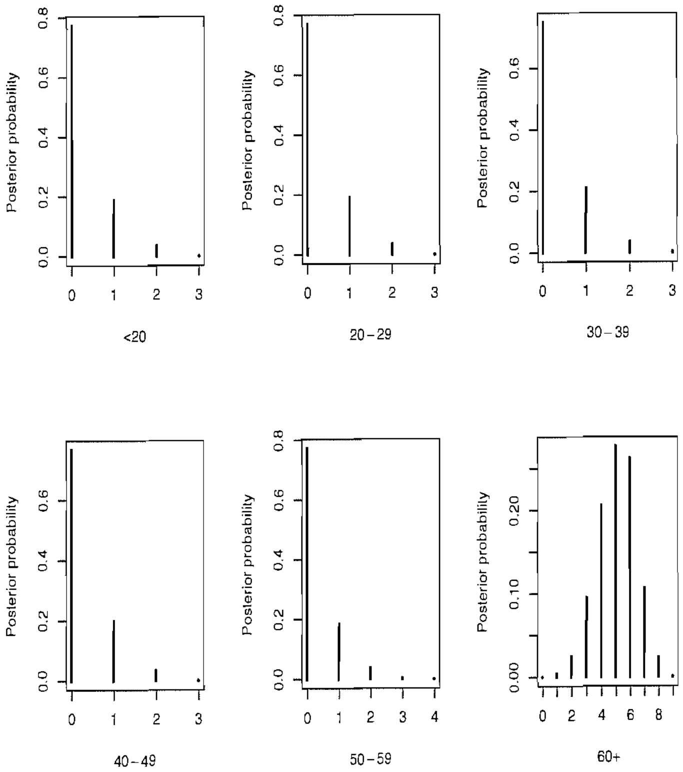
The above result follows exactly as before.

Appendix C: Supplementary Disclosure Control Results

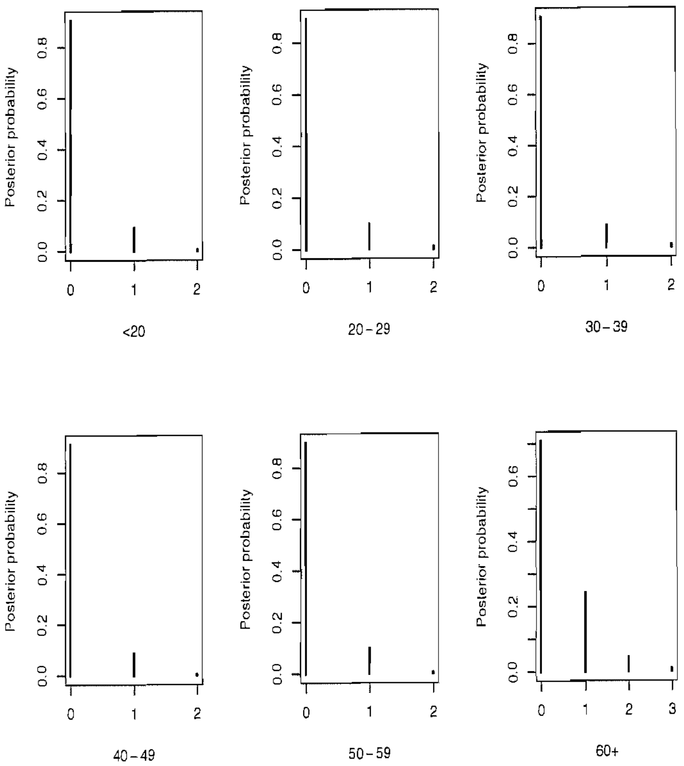
Posterior cell counts for probabilities, female/single in Barningham and Ovington.



Posterior cell counts for probabilities, male/single in Barningham and Ovington.



Posterior cell counts for probabilities, female/couple in Barningham and Ovington.



Appendix D: Crime and Punishment Data

	Indicator Number														
y_i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
791	151	1	91	58	56	510	950	33	301	108	41	394	261	0.084602	26.2011
1635	143	0	113	103	95	583	1012	13	102	96	36	557	194	0.029599	25.2999
578	142	1	89	45	44	533	969	18	219	94	33	318	250	0.083401	24.3006
1969	136	0	121	149	141	577	994	157	80	102	39	673	167	0.015801	29.9012
1234	141	0	121	109	101	591	985	18	30	91	20	578	174	0.041399	21.2998
682	121	0	110	118	115	547	964	25	44	84	29	689	126	0.034201	20.9995
963	127	1	111	82	79	519	982	4	139	97	38	620	168	0.0421	20.6993
1555	131	1	109	115	109	542	969	50	179	79	35	472	206	0.040099	24.5988
856	157	1	90	65	62	553	955	39	286	81	28	421	239	0.071697	29.4001
705	140	0	118	71	68	632	1029	7	15	100	24	526	174	0.044498	19.5994
1674	124	0	105	121	116	580	966	101	106	77	35	657	170	0.016201	41.6
849	134	0	108	75	71	595	972	47	59	83	31	580	172	0.031201	34.2984
511	128	0	113	67	60	624	972	28	10	77	25	507	206	0.045302	36.2993
664	135	0	117	62	61	595	986	22	46	77	27	529	190	0.0532	21.501
798	152	1	87	57	53	530	986	30	72	92	43	405	264	0.0691	22.7008
946	142	1	88	81	77	497	956	33	321	116	47	427	247	0.052099	26.0991
539	143	0	110	66	63	537	977	10	6	114	35	487	166	0.076299	19.1002
929	135	1	104	123	115	537	978	31	170	89	34	631	165	0.119804	18.1996
750	130	0	116	128	128	536	934	51	24	78	34	627	135	0.019099	24.9008
1225	125	0	108	113	105	567	985	78	94	130	58	626	166	0.034801	26.401
742	126	0	108	74	67	602	984	34	12	102	33	557	195	0.0228	37.5998
439	157	1	89	47	44	512	962	22	423	97	34	288	276	0.089502	37.0994
1216	132	0	96	87	83	564	953	43	92	83	32	513	227	0.0307	25.1989
968	131	0	116	78	73	574	1038	7	36	142	42	540	176	0.041598	17.6
523	130	0	116	63	57	641	984	14	26	70	21	486	196	0.069197	21.9003
1993	131	0	121	160	143	631	1071	3	77	102	41	674	152	0.041698	22.1005
342	135	0	109	69	71	540	965	6	4	80	22	564	139	0.036099	28.4999
1216	152	0	112	82	76	571	1018	10	79	103	28	537	215	0.038201	25.8006
1043	119	0	107	166	157	521	938	168	89	92	36	637	154	0.0234	36.7009
696	166	1	89	58	54	521	973	46	254	72	26	396	237	0.075298	28.3011
373	140	0	93	55	54	535	1045	6	20	135	40	453	200	0.041999	21.7998
754	125	0	109	90	81	586	964	97	82	105	43	617	163	0.042698	30.9014
1072	147	1	104	63	64	560	972	23	95	76	24	462	233	0.049499	25.5005
923	126	0	118	97	97	542	990	18	21	102	35	589	166	0.040799	21.6997
653	123	0	102	97	87	526	948	113	76	124	50	572	158	0.0207	37.4011
1272	150	0	100	109	98	531	964	9	24	87	38	559	153	0.0069	44.0004
831	177	1	87	58	56	638	974	24	349	76	28	382	254	0.045198	31.6995
566	133	0	104	51	47	599	1024	7	40	99	27	425	225	0.053998	16.6999
826	149	1	88	61	54	515	953	36	165	86	35	395	251	0.047099	27.3004
1151	145	1	104	82	74	560	981	96	126	88	31	488	228	0.038801	29.3004
880	148	0	122	72	66	601	998	9	19	84	20	590	144	0.0251	30.0001
542	141	0	109	56	54	523	968	4	2	107	37	489	170	0.088904	12.1996
823	162	1	99	75	70	522	996	40	208	73	27	496	224	0.054902	31.9989
1030	136	0	121	95	96	574	1012	29	36	111	37	622	162	0.0281	30.0001
455	139	1	88	46	41	480	968	19	49	135	53	457	249	0.056202	32.5996
508	126	0	104	106	97	599	989	40	24	78	25	593	171	0.046598	16.6999
849	130	0	121	90	91	623	1049	3	22	113	40	588	160	0.052802	16.0997

Appendix E: Political Attitudes Data

A: How well does the political system function today?

B: Type of formal schooling

C: Age group

D: Time of Survey

E: Region of Survey

Levels of A (i)	Levels of B (j)	Levels of C (k)	Levels of D (l)	Levels of E (m)
1: Very poorly	1: Basic incomplete	1: 19-29	1: 1991	1: West
2: Poorly	2: Basic	2: 30-34	1: 1992	2: East
3: Well	3: Medium	3: 45-59		
4: Very Well	4: Upper medium	4: 60-74		
	5: Intensive	5: ≥ 75		

Levels of			Level of E m=1					Level of E m=2				
A, C, D:			Levels of B					Levels of B				
i	k	l	1	2	3	4	5	1	2	3	4	5
1	1	1	1	5	10	3	8	0	0	2	0	2
2	1	1	3	63	88	22	78	2	13	103	6	29
3	1	1	1	25	18	5	9	1	5	53	1	12
4	1	1	2	2	0	0	1	0	0	4	0	3
1	2	1	0	24	17	3	11	0	3	7	0	1
2	2	1	1	135	89	26	68	3	39	198	7	52
3	2	1	1	34	14	1	10	4	27	86	7	17
4	2	1	0	2	1	0	1	1	7	9	0	1
1	3	1	0	26	4	2	5	0	4	0	0	2
2	3	1	2	120	62	17	29	14	134	50	5	32
3	3	1	1	27	10	2	3	13	61	18	3	18
4	3	1	0	6	2	0	0	1	7	3	0	2
1	4	1	2	41	12	1	7	0	4	1	0	0

2	4	1	6	107	32	4	26	7	81	15	1	12
3	4	1	1	18	3	0	2	6	56	8	1	7
4	4	1	1	3	0	0	0	1	5	0	0	0
1	5	1	1	8	3	1	2	0	3	0	0	0
2	5	1	1	28	8	3	6	5	16	1	1	3
3	5	1	0	9	2	0	0	0	10	1	0	1
4	5	1	1	0	0	0	0	0	1	0	0	0
1	1	2	0	6	4	0	7	0	0	0	0	0
2	1	2	2	68	101	17	100	1	8	58	1	13
3	1	2	3	40	48	3	29	1	8	68	2	16
4	1	2	0	6	8	2	1	0	1	13	0	1
1	2	2	0	10	7	4	8	0	0	0	0	1
2	2	2	4	186	100	47	99	0	25	104	5	26
3	2	2	6	102	67	10	25	1	22	86	3	38
4	2	2	1	14	5	0	3	1	2	13	3	3
1	3	2	1	19	11	2	9	1	2	0	1	0
2	3	2	2	182	76	17	42	2	89	25	14	30
3	3	2	6	102	24	6	13	3	74	27	14	7
4	3	2	0	11	3	1	1	3	13	3	2	1
1	4	2	1	11	7	0	4	0	0	0	1	0
2	4	2	4	177	57	9	26	1	62	16	3	7
3	4	2	5	82	10	1	6	2	46	9	5	6
4	4	2	0	21	5	0	3	1	5	1	0	2
1	5	2	0	12	2	0	1	0	3	0	0	0
2	5	2	3	51	16	1	6	0	18	4	1	0
3	5	2	1	22	6	1	4	0	14	1	1	0
4	5	2	0	7	1	0	0	0	0	0	0	0

Appendix F: Simulated Bimodal Data

Levels of					Levels of				
A	B	C	D	count	A	B	C	D	count
1	1	1	1	0	1	3	1	2	0
2	1	1	1	0	2	3	1	2	5
3	1	1	1	2	3	3	1	2	0
4	1	1	1	0	4	3	1	2	0
1	1	2	1	0	1	3	2	2	0
2	1	2	1	0	2	3	2	2	1
3	1	2	1	0	3	3	2	2	0
4	1	2	1	0	4	3	2	2	0
1	1	3	1	0	1	3	3	2	2
2	1	3	1	0	2	3	3	2	2
3	1	3	1	0	3	3	3	2	0
4	1	3	1	0	4	3	3	2	0
1	1	4	1	14	1	3	4	2	0
2	1	4	1	1	2	3	4	2	1
3	1	4	1	0	3	3	4	2	0
4	1	4	1	0	4	3	4	2	0
1	1	5	1	0	1	3	5	2	1
2	1	5	1	0	2	3	5	2	1
3	1	5	1	0	3	3	5	2	0
4	1	5	1	0	4	3	5	2	0
1	1	1	2	0	1	4	1	1	0
2	1	1	2	0	2	4	1	1	0
3	1	1	2	0	3	4	1	1	0
4	1	1	2	0	4	4	1	1	0
1	1	2	2	0	1	4	2	1	0
2	1	2	2	8	2	4	2	1	0
3	1	2	2	4	3	4	2	1	0
4	1	2	2	0	4	4	2	1	0
1	1	3	2	1	1	4	3	1	0
2	1	3	2	1	2	4	3	1	0
3	1	3	2	0	3	4	3	1	0
4	1	3	2	0	4	4	3	1	0
1	1	4	2	0	1	4	4	1	0
2	1	4	2	0	2	4	4	1	0
3	1	4	2	0	3	4	4	1	0
4	1	4	2	0	4	4	4	1	0
1	1	5	2	0	1	4	5	1	0
2	1	5	2	0	2	4	5	1	1
3	1	5	2	0	3	4	5	1	2
4	1	5	2	0	4	4	5	1	0
1	2	1	1	0	1	4	1	2	0
2	2	1	1	0	2	4	1	2	0
3	2	1	1	0	3	4	1	2	0
4	2	1	1	0	4	4	1	2	0
1	2	2	1	5	1	4	2	2	0
2	2	2	1	9	2	4	2	2	1
3	2	2	1	0	3	4	2	2	1
4	2	2	1	0	4	4	2	2	0
1	2	3	1	0	1	4	3	2	0
2	2	3	1	1	2	4	3	2	0
3	2	3	1	1	3	4	3	2	0
4	2	3	1	9	4	4	3	2	0
1	2	4	1	9	1	4	4	2	0
2	2	4	1	0	2	4	4	2	3
3	2	4	1	0	3	4	4	2	2
4	2	4	1	0	4	4	4	2	2
1	2	5	1	0	1	4	5	2	0
2	2	5	1	0	2	4	5	2	0

3	2	5	1	0	3	4	5	2	0
4	2	5	1	0	4	4	5	2	0
1	2	1	2	1	1	5	1	1	0
2	2	1	2	2	2	5	1	1	1
3	2	1	2	0	3	5	1	1	0
4	2	1	2	0	4	5	1	1	0
1	2	2	2	0	1	5	2	1	1
2	2	2	2	0	2	5	2	1	5
3	2	2	2	0	3	5	2	1	0
4	2	2	2	0	4	5	2	1	0
1	2	3	2	4	1	5	3	1	0
2	2	3	2	3	2	5	3	1	0
3	2	3	2	2	3	5	3	1	0
4	2	3	2	0	4	5	3	1	0
1	2	4	2	0	1	5	4	1	14
2	2	4	2	7	2	5	4	1	1
3	2	4	2	5	3	5	4	1	0
4	2	4	2	4	4	5	4	1	0
1	2	5	2	5	1	5	5	1	0
2	2	5	2	2	2	5	5	1	0
3	2	5	2	0	3	5	5	1	0
4	2	5	2	0	4	5	5	1	0
1	3	1	1	0	1	5	1	2	2
2	3	1	1	0	2	5	1	2	5
3	3	1	1	0	3	5	1	2	0
4	3	1	1	0	4	5	1	2	0
1	3	2	1	3	1	5	2	2	0
2	3	2	1	9	2	5	2	2	1
3	3	2	1	0	3	5	2	2	0
4	3	2	1	0	4	5	2	2	0
1	3	3	1	0	1	5	3	2	2
2	3	3	1	1	2	5	3	2	1
3	3	3	1	0	3	5	3	2	0
4	3	3	1	2	4	5	3	2	0
1	3	4	1	4	1	5	4	2	0
2	3	4	1	0	2	5	4	2	0
3	3	4	1	0	3	5	4	2	0
4	3	4	1	0	4	5	4	2	0
1	3	5	1	0	1	5	5	2	0
2	3	5	1	0	2	5	5	2	0
3	3	5	1	0	3	5	5	2	0
4	3	5	1	0	4	5	5	2	0

References

Al-Awadhi, F., Hurn, M. and Jennison, C. (2004). Improving the acceptance rate of reversible jump MCMC proposals. *Statistics and Probability Letters*, **69**, 189–198.

Atchadé, Y. F. and Rosenthal, J. S. (2005). On adaptive Markov chain Monte Carlo algorithms. *Bernoulli*, **11**, 815–828.

Azzalini, A. (1996). *Statistical Inference Based on the Likelihood, Monographs on Statistical and Applied Probability 68.* Chapman and Hall, London.

Baker, S. G. and Laird, N. M. (1988). Regression analysis for categorical variables with outcome subject to non-ignorable non-response. *Journal of the American Statistical Association*, **83**, 62–69.

Baker, S. G., Rosenberger, W. F. and DerSimonian, R. (1992). Closed-form estimates for missing counts in two-way contingency tables. *Statistics in Medicine*, **11**, 643-657.

Bishop, Y. M., Fienberg, S. E. and Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice.* MIT Press.

Blair, J. R. S. and Barry, P. (1993). An introduction into chordal graphs and clique trees. In *Graph Theory and Sparse Matrix Computation* (eds. A. George, J. R. Gilbert and J. W. H. Liu), 1–30. Springer-Verlag, Berlin.

Brooks, S. P., Giudici, P. and Roberts, G. O. (2003). Efficient construction of reversible jump MCMC proposal distributions. *Journal of the Royal Statistical Society B*, **65**, 3–55.

Carlin, B. P. and Chib, S. (1995). Bayesian model choice via Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B*, **157**, 473–484.

Damien, P., Wakefield, J. and Walker, S. (1999). Gibbs sampling for Bayesian non-conjugate and hierarchical models using auxiliary variables. *Journal of the Royal Statistical Society Series B*, **61**, 331–344.

Dellaportas, P. and Forster, J. J. (1999). Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models. *Biometrika*, **86**, 615–633.

Dellaportas, P., Forster, J. J. and Ntzoufras, I. (2002). On Bayesian model and variable selection using MCMC. *Statistics and Computing*, **12**, 27–36.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistics Society Series B*, **39**, 1–38.

Diaconis, P. and Sturmfels, B. (1998). Algebraic algorithms for sampling from conditional distributions. *Annals of Statistics*, **26**, 363–397.

Dobra, A. (2003). Markov bases for decomposable graphical models. *Bernoulli*, **9**, 1093–1108.

Dobra, A. and Fienberg, S. E. (2001). Bounds for cell entries in contingency tables induced by fixed marginal totals with applications to disclosure limitation. *Statistical Journal of the United Nations*, **18**, 363–371.

Ehlers, R. S. and Brooks, S. P. (2002). Efficient construction of reversible jump MCMC proposals for autoregressive time series models. *Technical Report, University of Cambridge.*

Ehrlich, I. (1973). Participation in illegitimate activities: A theoretical and empirical investigation. *Journal of Political Economy*, **81**, 521–565.

Fienberg, S. E. (1999). Fréchet and Bonferroni bounds for multi-way tables of counts with applications to disclosure limitation. In *Statistical Data Protection (SDP98)*

Proceedings (eds. P. J. Green, N. L. Hjort and S. Richardson), 115–129. Eurostat, Luxembourg.

Forster, J. J. and Smith, P. W. F. (1998). Model-based inference for categorical survey data subject to non-ignorable non-response (with discussion). *Journal of the Royal Statistics Society Series B*, **60**, 57–70.

Forster, J. J. and Webb, E. L. (2007). Bayesian Disclosure Risk Assessment: Predicting Small Frequencies in Contingency Tables. *Technical Report, School of Mathematics, University of Southampton.*

Gamerman, D. (1997). *Markov Chain Monte Carlo.* Chapman and Hall, London.

Gelfand, A. E., Sahu, S. K. and Carlin, B. P. (1996). Efficient Parametrizations for Generalised Linear Mixed Models. In *Bayesian Statistics 5* (eds. J. Bernardo, J. Berger, A. Dawid and A.F.M), 165–203. Oxford University Press, Oxford.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Society*, **85**, 398–409.

Gelman, A. and Meng, X.-L. (1996). Model checking and model improvement. In *Markov Chain Monte Carlo in Practice* (eds. W. R. Gilks, S. Richardson and D. J. Spiegelhalter), 189-198. Chapman and Hall/CRC, London.

Gelman, A., Meng, X.-L. and Stern, H. S. (1996). Posterior predictive assessment of model fitness (with discussion). *Statistica Sinica*, **6**, 733–807.

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, **88**, 881–889.

— (1997). Approaches for Bayesian variable selection. *Statistica Sinica*, **7**, 339–373.

Geyer, C. J. (1991). Markov chain Monte Carlo maximum likelihood. In *Computing Science and Statistics, Proceedings on the 23rd Symposium on the Interface* (ed. E. M. Keramidas), 156–163. Interface Foundation on North America, Seattle.

— (1992). Practical Markov chain Monte Carlo (with discussion). *Statistical Science*, **7**, 473–511.

Godsill, S. J. (2001). On the relationship between Markov chain Monte Carlo methods for model uncertainty. *Journal of Computational and Graphical Statistics*, **10**, 230–248.

— (2003). Proposal densities and product-space methods. In *Highly Structured Stochastic Systems, Oxford Statistical Science Series, No. 27* (eds. P. J. Green, N. L. Hjort and S. Richardson), 199–203. Oxford University Press, Oxford.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.

— (2003). Trans-dimensional Markov chain Monte Carlo. In *Highly Structured Stochastic Systems, Oxford Statistical Science Series, No. 27* (eds. P. J. Green, N. L. Hjort and S. Richardson), 179–198. Oxford University Press, Oxford.

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

Hosmer, D. W. and Lemeshow, S. (1989). *Applied Logisitc Regression*. John Wiley and Sons, New York.

Jeffreys, H. (1967). *Theory of Probability*. Oxford University Press, Oxford.

Kadane, J. B. (1993). Subjective Bayesian analysis for surveys with missing data. *Statistician*, **42**, 415–426.

Knuiman, M. W. and Speed, T. P. (1988). Incorporating prior information into the analysis of contingency tables. *Biometrics*, **44**, 1061–1071.

Little, R. J. A. (1982). Models for nonresponse in sample surveys. *Journal of the American Statistical Association*, **77**, 237–250.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. John Wiley and Sons, New York.

— (2002). *Statistical Analysis with Missing Data (Second Edition).* John Wiley and Sons, New York.

Lu, G. and Copas, J. B. (2004). Missing at random, likelihood ignorability and model completeness. *Annals of Statistics*, **32**, 754–765.

Madigan, D. and Raftery, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association*, **89**, 1535–1546.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models, second edition.* Chapman and Hall/CRC, New York.

Meng, X. L. and Rubin, D. B. (1991). Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *Journal of the American Statistical Association*, **86**, 899–909.

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.

Molenberghs, G., Kenward, M. G. and Goetghebeur, E. (2001). Sensitivity analysis for incomplete contingency tables: the Slovenian plebiscite case. *Applied Statistics*, **50**, 15–29.

Nargundkar, M. S. and Saveland, W. (1972). Random rounding: A means of preventing disclosure of information about individual respondents in aggregate data. *American Statistical Association Annual Meeting - Proceedings of the Social Statistics Section*, **225**, 382–385.

Neal, R. M. (2004). Improving asymptotic variance of MCMC estimators: Non-reversible chains are better. *Technical Report No. 0406, Dept. of Statistics, University of Toronto.*

Norris, J. R. (1997). *Markov Chains.* Cambridge University Press.

Nott, D. J. and Green, P. J. (2004). Bayesian variable selection and the Swendsen-Wang algorithm. *Journal of Computational and Graphical Statistics*, **13**, 141–157.

Nott, D. J. and Leonte, D. (2004). Sampling schemes for Bayesian variable selection in generalised linear models. *Journal of Computational and Graphical Statistics*, **13**, 362–382.

Ntzoufras, I., Dellaportas, P. and Forster, J. J. (2001). Bayesian variable and link determination for generalised linear models. *Journal of Statistical Planning and Inference*, **111**, 165–180.

O'Hagan, A. and Forster, J. (2004). *Kendall's Advanced Theory of Statistics: Bayesian Inference, second edition*. Arnold, London.

Perks, W. (1946). Some observations on inverse probability including a new indifference rule. *Journal of the Institute of Actuaries*, **89**, 44–52.

Peskun, P. H. (1973). Optimum Monte-Carlo sampling using Markov chains. *Biometrika*, **60**, 607–612.

Raftery, A. E., Madigan, D. and Hoeting, J. A. (1997). Bayesian Model averaging for linear regression models, with discussion. *Journal of the American Statistical Association*, **92**, 179–191.

Richardson, S. and Green, P. J. (1997). Bayesian analysis of univariate normal mixtures. *Journal of the Royal Statistical Society Series B*, **59**, 731–792.

Roberts, G. O., Gelman, A. and Gilks, W. R. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *Annals of Applied Probability*, **7**, 110–120.

Roberts, G. O. and Rosenthal, J. S. (1998). Optimal scaling of discrete approximations to Langevin diffusions. *Journal of the Royal Statistical Society Series B*, **60**, 255–268.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, **63**, 581–592.

— (1978). Multiple imputation in sample surveys – A phenomenological Bayesian approach to nonresponse. *The Proceedings of the Survey Research Methods Section of the American Statistical Association*, **72**, 20–42.

— (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, **91**, 473–489.

Rubin, D. B., Stern, H. S. and Vehovar, V. (1995). Handling "Don't know" survey response: The case of the Slovenian plebiscite. *Journal of the American Statistical Association*, **90**, 822–828.

Skinner, C. J. and Elliot, M. J. (2002). A measure of disclosure risk for microdata. *Journal of the Royal Statistical Society B*, **64**, 855–867.

Swendsen, R. H. and Wang, J. S. (1987). Nonparametric regression using Bayesian variable selection. *Physical Review Letter*, **58**, 86–88.

Vandaele, W. (1978). Participation in illegitimate activities: Ehrlich revisited. In *Deterrence and Incapacitation* (eds. A. Blumstein, J. Cohen and D. Nagin), 270–335. National Academy of Sciences Press, Washington.

Venables, W. N. and Ripley, B. D. (1999). *Modern Applied Statistics with S-Plus.* Springer, New York.

Welsh, A. H. (1996). *Aspects of Statistical Inference.* John Wiley and Sons, New York.

Wermuth, N. and Cox, D. R. (1998). On the application of conditional independence to ordinal data. *International Statistical Review*, **66**, 181–200.

Willenborg, L. and de Waal, T. (1996). *Statistical Disclosure Control in Practice.* Springer-Verlag, New York.

— (2001). *Elements of Statistical Disclosure Control.* Springer-Verlag, New York.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Bayesian Inference and Decision Techniques: Essays* *in honour of Bruno de Finetti* (eds. P. Goel and A. Zellner), 233–243. North-Holland , Amsterdam.