UNIVERSITY OF SOUTHAMPTON

FACULTY OF ENGINEERING, SCIENCE AND MATHEMATICS
School of Mathematics

Bayesian Model Choice for Multivariate Ordinal Data

by

Emily Louise Webb

Thesis submitted for the degree of Doctor of Philosophy
January 2005

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF ENGINEERING, SCIENCE AND MATHEMATICS
SCHOOL OF MATHEMATICS

Doctor of Philosophy

BAYESIAN MODEL CHOICE FOR MULTIVARIATE ORDINAL DATA
by Emily Louise Webb

This thesis provides a coherent and adaptable methodology for multivariate ordinal and binary data. Two main aspects of data modelling are considered. The first is to formulate a model for the data and to estimate the model parameters using Bayesian computation. The second is to assess model choice; models considered are the set of directed acyclic graphical models and the set of decomposable models.

The model is based on the multivariate probit model (Chib and Greenberg, 1998) but parameterised in a way that makes computation convenient. In particular, the conditional posterior distributions of the model parameters are standard and easily simulated from using Gibbs sampling techniques. Prior parameters are chosen to be noninformative but not overly diffuse. The Gibbs sampler is applied successfully to examples, and the goodness-of-fit of the model is assessed using simulation techniques. The model parameterisation allows ordinal and binary data and a mixture of both data types to be modelled within the same framework.

Reversible Jump Markov chain Monte Carlo methods are used to estimate posterior model probabilities for directed acyclic graphical models. Under the model parameterisation described, a suitable proposal distribution is easily specified.

The issue of model choice is also investigated for the set of (undirected) decomposable models. Under some model parameterisations, the conditional independence structure of a decomposable model can not be specified. A further Reversible Jump Markov chain Monte Carlo step is described to move between model parameterisations. Both Reversible Jump algorithms are found to rapidly explore the model and parameter spaces.

The model is extended for data where covariates are also observed. The Reversible Jump algorithms described previously are adapted and applied to examples. A further Reversible Jump step is developed and implemented to assess which covariates should be included in a model to predict the data.

Contents

List of Figures

List of Tables

Acknowledgments

Firstly I would like to thank my supervisor Professor Jon Forster for all his help, encouragement and enthusiasm over the course of the PhD. I would also like to acknowledge the EPSRC and the University of Southampton for funding me.

Thank you to Ben for his help and for always having confidence in me even when I had none in myself. Thanks to all my friends, especially Denis, Torbjorn, Jenni, Katy, Ronni and James for much entertainment and always making me smile, and thanks to Phil for being understanding over the last six months. Finally, thank you to my family for their endless support, not just over the last three years.

Chapter 1

Introduction

The aim of this thesis is to provide a coherent methodology for multivariate ordinal and binary data. Ordinal data are characterised by the response taking the form of discrete, ordered categories and occur in many fields of research, in particular in the social sciences and in medicine. Ordinal data are often treated using the same methods as for nominal (unordered) categorical data, but this ignores the extra structure due to the categories being ordered. The main difference between models for nominal data and models for ordinal data is that those for nominal would give the same results if the order of the response categories were permuted. In this thesis, we introduce a Bayesian methodology for modelling ordinal data. There are two main aspects to this methodology: the first is to set up a model for multivariate ordinal data and to estimate the parameters using Bayesian computation. The second is to assess model choice, that is, to find which models may be used to best predict a given data set. In this Chapter, we introduce the principles underlying the work.

1.1 Bayes' Theorem

The fundamental principle of Bayesian analysis is that uncertainty is represented through probability. This means that the parameters that describe the probability distribution of the observed data are treated as random variables in their own right, with associated probability distributions. Suppose we observe data $\boldsymbol{y}$ and require inference about a parameter vector $\boldsymbol{\theta}$. Then Bayes' theorem states that:

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

where $f(\boldsymbol{\theta}|\boldsymbol{y})$ is the posterior distribution of the parameter vector $\boldsymbol{\theta}$ given the data $\boldsymbol{y}$, $f(\boldsymbol{y}|\boldsymbol{\theta})$ is the likelihood of the data $\boldsymbol{y}$ given the parameter vector $\boldsymbol{\theta}$, and $f(\boldsymbol{\theta})$ is the prior distribution of the parameter vector $\boldsymbol{\theta}$. The prior distribution may be chosen to reflect our beliefs about what values parameters take. Therefore the posterior distribution is formed from our initial beliefs about the parameters, updated by the data that have been observed. Since the integral $\int f(\boldsymbol{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})d\boldsymbol{\theta}$ is simply the normalising constant for the posterior distribution, it is often omitted and Bayes theorem is expressed and implemented as:

$$f(\boldsymbol{\theta}|\boldsymbol{y}) \propto f(\boldsymbol{y}|\boldsymbol{\theta})f(\boldsymbol{\theta}) \tag{1.1}$$

1.2 Contingency Tables

Multivariate ordinal data are usually represented in a contingency table, with each margin of the table consisting of ordered categories. The contingency table may be highly structured and modelling this structure helps us to understand the relationship between the variables. The standard way to represent the structure is *via* a log-linear model, which relates the log of the cell means to a set of model parameters.

Suppose we have a set of multivariate categorical data, with n individuals cross-classified by p categorical variables, so that the data can be represented by a $p-$way contingency table. Let Γ denote the set of classifying variables, so $|\Gamma| = p$. Following the notation introduced by Darroch et al. (1980), the set of cells in the table is denoted by $I = \prod_{\gamma\in\Gamma} I_\gamma$, where I_γ is the set of levels that variable γ can take. A single cell is denoted by $\boldsymbol{i} = (i_\gamma : \gamma \in \Gamma)$ and we let $n_{\boldsymbol{i}}$

denote the corresponding cell count and p_i the corresponding cell probability, where $\sum_{i \in I} n_i = n$ and $\sum_{i \in I} p_i = 1$.

For example, consider a three-way table cross-classified by variables X (with three categories), Y (with five categories) and Z (with two categories). Then the dimension is $p = 3$, the set of classifying variables if $\Gamma = \{X, Y, Z\}$and the numbers of levels of the classifying variables are $|I_X| = 3$, $|I_Y| = 5$, and $|I_Z| = 2$.

1.3 The Multinomial-Dirichlet Model for Nominal Data

A standard model for the situation where observations are classified into a finite number of categories is the multinomial distribution. From a total population of size n, suppose that n_i individuals are assigned at random to a particular cell i with probability p_i, with $\sum_i n_i = n$ and $\sum_i p_i = 1$. Then the vector of cell counts $\boldsymbol{n}$ has a multinomial distribution with likelihood

$$f(\boldsymbol{n}|\boldsymbol{p}) = n! \prod_i \frac{p_i^{n_i}}{n_i!}$$

The natural conjugate prior for the cell probabilities $\boldsymbol{p}$ is the Dirichlet distribution which has density

$$f(\boldsymbol{p}) = \frac{\Gamma(\alpha)}{\prod_i \Gamma(\alpha_i)} \prod_i p_i^{\alpha_i - 1}$$

where the elements of $\boldsymbol{\alpha}$ are parameters which control the location and spread of the distribution, and $\alpha = \sum_i \alpha_i$. By Bayes' theorem, the posterior distribution of the cell probabilities is then

$$f(\boldsymbol{p}|\boldsymbol{n}) = \frac{\Gamma(\alpha + n)}{\prod_i \Gamma(\alpha_i + n_i)} \prod_i p_i^{\alpha_i + n_i - 1}$$

i.e. the Dirichlet distribution with parameter vector $\boldsymbol{\alpha} + \boldsymbol{n}$. Note that this approach is invariant to ordering of categories, but is often applied to ordinal data, ignoring its extra structure.

1.4 Log-Linear Models

One of the main points of interest when analysing contingency tables is to model the association between classifying variables. The standard way of doing this is by representing the underlying statistical model as a log-linear model. This associates the expected cell counts with a linear combination of parameters. Suppose the cell counts $n_{\boldsymbol{i}}$ are observations of independent Poisson random variables with means $\mu_{\boldsymbol{i}}$. Then following Darroch et al. (1980), the log-linear model may be denoted

$$\log \mu_{\boldsymbol{i}} = \sum_{a \subseteq \Gamma} \xi_a(\boldsymbol{i}_a) \qquad \boldsymbol{i} \in I$$

where $\boldsymbol{i}_a$ is the marginal cell $\boldsymbol{i}_a = (i_\gamma, \gamma \in a)$. The functions ξ_a are the interactions among the factors in a. If $|a| = 1$, ξ_a is a main effect and if $|a| = m$, ξ_a is an m-way interaction. The general non-saturated log-linear model involves setting certain ξ_a to be zero; for the saturated model, there is a full set of interaction terms. To ensure identifiability, constraints are imposed on the ξ_a.

1.4.1 Hierarchical models

In practice, general log-linear models are not easy to interpret, so attention is generally restricted to the set of hierarchical models, a subset of the general log-linear models. To obtain these, we impose restrictions on the ξ_a, namely that if ξ_a is specified to vanish and $b \supseteq a$ then ξ_b must also be forced to vanish, i.e. if there are no interactions among factors in a then there is no interaction of higher order involving all the factors in a.

1.4.2 Graphical models

The set of graphical models is a subset of the hierarchical models. A graphical model can be represented by a graph consisting of vertices and edges where,

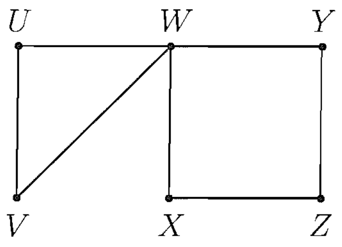


Figure 1.1: Illustration of conditional independence

in the contingency table setup, each vertex corresponds to a classifying variable of the table. Graphical models may also be represented in terms of their conditional independence structure, which is immediately apparent from the graph itself, as described below. They therefore have a more straightforward interpretation than the hierarchical models, in terms of conditional independences. Let $\mathcal{V}$ be the set of vertices and $\mathcal{E}$ be the set of all possible edges between them. Following Dawid and Lauritzen (1993), let $(X, Y) \in \mathcal{E}$ denote the edge between variables X and Y. If two vertices are not joined by an edge, then the corresponding variables are conditionally independent given the other variables. Conditional independence can also be defined for sets of variables. A subset C of the set of all classifying variables Γ is called a *clique* if the subgraph containing only elements of C has an edge connecting each pair of vertices and the inclusion of another vertex from V in C would result in at least one pair of unconnected vertices. The subset S is a *separator* of cliques A and B if every path from any vertex in A to one in B must pass through a vertex in S. In such a case, variables in A are conditionally independent of those in B, given those in S.

For example, if $\Gamma = \{U, V, W, X, Y, Z\}$, consider the model represented by the graph in Figure 1.1. There are five cliques: $C_1 = UVW$, $C_2 = WX$, $C_3 = XZ$, $C_4 = YZ$ and $C_5 = WY$ with corresponding separators $S_1 = W$, $S_2 = X$, $S_3 = Z$ and $S_4 = Y$. Then the following conditional independence statements can be made: $\{U, V\}$ is conditionally independent of $\{X, Y, Z\}$ given W; W is conditionally independent of Z given X and Y; X is conditionally independent of Y given Z and W; and Z is conditionally independent of W given X and

Y.

The description above assumes that edges in the graph are undirected, that is we do not distinguish between (X, Y) and (Y, X). Graphs where we make this distinction are called directed graphs.

In a directed graph, if $(X, Y) \in \mathcal{E}$ but $(Y, X) \notin \mathcal{E}$ then there is a directed edge from X to Y, denoted by $X \rightarrow Y$. X is called the parent of Y, and Y is the child of X. For both the directed and undirected cases, a *path* of length n from X to Y is defined as a sequence $X = X_0, \ldots, X_n = Y$ of distinct vertices such that $(X_{i-1}, X_i) \in \mathcal{E}$ for all $i = 1, \ldots, n$. An *n cycle* is a path of length n with the modification that it begins and ends at the same point. A *directed acyclic graph* (DAG) is a directed graph without cycles.

Directed acyclic graphs permit an ordering of the vertices such that no edge (X, Y) exists when Y precedes X in the ordering. Directed graphical models correspond to DAGs and have the following conditional independence interpretation. The absence of a directed edge between two variables means they are conditionally independent given all other variables which precede either of them in the ordering.

1.4.3 Decomposable models

The set of decomposable models is a further subset of the graphical models. All directed acyclic graphical models are decomposable. An undirected graphical model is decomposable if it does not contain cycles of length greater than three without a chord *i.e.* an edge which short-cuts the cycle. The model represented by the graph in Figure 1.1 is not decomposable as it contains a cycle of length four in $WXYZ$. However the submodel with vertices U, V, W, X is decomposable. Decomposable models clearly exclude many potentially useful models. However, they have many useful computational properties for model selection procedures. These will be discussed in 1.5.

If a model is decomposable, then the undirected conditional independence graph may be used to construct a directed acyclic version with the same Markov structure. This will be discussed extensively in Chapter 4.

1.5 Model Uncertainty

The standard approach for Bayesian model comparison is to calculate the marginal likelihoods for competing models and hence the posterior model probabilities. The posterior model probability for a model m is

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

where $f(\boldsymbol{y}|m)$ is the marginal likelihood of model m, defined as:

$$f(\boldsymbol{y}|m) = \int f(\boldsymbol{y}|m, \boldsymbol{\theta}_m)f(\boldsymbol{\theta}_m|m)d\boldsymbol{\theta}_m \tag{1.2}$$

$\boldsymbol{\theta}_m$ is the set of parameters in model m, and $f(\boldsymbol{\theta}_m|m)$ is the conditional prior distribution of $\boldsymbol{\theta}_m$.

The marginal likelihood as defined above is analytically intractable in many examples. However, for decomposable graphical models, Dawid and Lauritzen (1993) construct a family of prior distributions, which allow posterior densities and marginal likelihoods to be calculated directly. In particular, the marginal likelihood for each model can be expressed in terms of the cliques and separators associated with that model, and hence model comparison can be carried out with calculations local to single cliques.

Hence it is in principle possible to calculate all posterior model probabilities. However in practice, for high-dimensional contingency tables, the number of calculations required to do so is prohibitively large. Two methods proposed for overcoming this problem are Occam's window (Madigan and Raftery, 1994) and Markov chain Monte Carlo model composition (Madigan and York, 1995). Occam's window provides a strategy whereby the number of models considered is dramatically reduced. It does this *via* three basic principles. Firstly if a model predicts the data far less well than the model which provides the best predictions it is no longer considered. Secondly, complex models which receive less support from the data than their simpler counterparts are excluded. Finally, if a model is rejected then all its submodels are rejected. Markov chain Monte Carlo model composition is a process which generates a Monte Carlo sample from $f(m|\boldsymbol{y})$. This method is more appropriate for making predictions when the posterior distribution of some quantity is of particular interest than for inferring the nature of the 'true' model.

1.6 Bayesian Computation

1.6.1 Computation for parameter estimation

In a Bayesian framework, we wish to estimate the posterior and prior distributions and various summaries of them. This generally involves integrating a function of the posterior (or prior) distribution, for example to calculate the mean of the posterior distribution, we must find

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

Evaluating this integral may be difficult (especially in higher dimensions) or analytically impossible. Many methods have been developed in order to overcome these difficulties. The two methods which we shall use in this thesis are the Gibbs sampler and Reversible Jump Markov chain Monte Carlo, both of which form part of a large group of numerical methods called Markov chain Monte Carlo (MCMC). MCMC uses the basic statistical theory that says features of an unknown distribution can be approximated if we generate random samples from the distribution. Suppose that $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(N)}$ form an identically distributed sample from the posterior distribution $f(\boldsymbol{\theta}|\boldsymbol{y})$. Then $E(b(\boldsymbol{\theta}))$ can be estimated accurately by:

$$E(b(\boldsymbol{\theta})) \approx \frac{1}{N} \sum_{i=1}^{N} b(\boldsymbol{\theta}^{(i)}) \qquad (1.3)$$

This is the principle of Monte Carlo integration.

1.6.2 Markov chain Monte Carlo

It can be difficult in many problems to sample independently from $f(\boldsymbol{\theta}|\boldsymbol{y})$. However, due to the ergodic theorem (Tierney, 1994), the estimate for $E(b(\boldsymbol{\theta}))$ in (1.3) does not require independent sampling. MCMC methods use a Markov process (where the distribution of the parameters $\boldsymbol{\theta}^{(i)}$ at the ith stage of the chain depends on $\boldsymbol{\theta}^{(i-1)}$) to produce dependent observations in such a way that their equilibrium distribution is $f(\boldsymbol{\theta}|\boldsymbol{y})$.

Suppose that the Markov chain can be run until equilibrium is approximately reached at iteration t. Then the parameter vectors $\boldsymbol{\theta}^{(t)}$, $\boldsymbol{\theta}^{(t+1)}$,...,$\boldsymbol{\theta}^{(t+N)}$ are a dependent sample of size N from the posterior distribution $f(\boldsymbol{\theta}|\boldsymbol{y})$, and (1.3) may then be used to estimate summaries of this distribution and any function of it. The value t at which equilibrium is approximately reached and the total number of iterations N are chosen so that the sample is considered to be representative of the posterior distribution $f(\boldsymbol{\theta}|\boldsymbol{y})$.

The value t is known as the burn-in length and observations before this should be discarded. However, if the chain is started at a plausible observation from $f(\boldsymbol{\theta}|\boldsymbol{y})$ then the burn-in is theoretically zero and no observations need be discarded. The samples obtained using MCMC methods are by definition dependent, but the degree of dependence varies. If the parameter space is explored rapidly by the Markov chain, then it is said to be mixing well and successive samples are not highly dependent. Conversely, if there is high correlation between successive observations, then the sampler is said to be mixing poorly and a highly dependent sample will be produced, which will require a very long run length to produce a representative sample.

1.6.3 The Metropolis-Hastings method

One method of constructing a suitable Markov chain is the Metropolis-Hastings method, first introduced by Metropolis et al. (1953) and developed by Hastings (1970). In this approach a sequence of samples is generated from the posterior distribution in the following manner:

1. Let $\boldsymbol{\theta}^{(t)}$ be the current sample from the posterior distribution $f(\boldsymbol{\theta}|\boldsymbol{y})$.
2. Generate a candidate vector of parameter values $\boldsymbol{\theta}^*$ from a proposal density $g(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t)})$
3. Accept the proposal with probability α where
$$\alpha = \min\left\{1, \frac{f(\boldsymbol{\theta}^*)g(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^*)}{f(\boldsymbol{\theta}^{(t)})g(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t)})}\right\}$$
and set $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^*$. Otherwise set $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)}$.

The random walk is a special case of the Metropolis-Hastings algorithm. For this method, $g(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t)})$ is chosen to be such that $\boldsymbol{\theta}^* = \boldsymbol{\theta}^{(t)} + \boldsymbol{\eta}$ where $\boldsymbol{\eta}$ is a random increment whose distribution does not depend on $\boldsymbol{\theta}^{(t)}$. Often, the distribution of $\boldsymbol{\eta}$ is symmetric about $\mathbf{0}$ so that $g(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^*) = g(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t)})$ and the acceptance probability simplifies to

$$\alpha = \min\left\{1, \frac{f(\boldsymbol{\theta}^*)}{f(\boldsymbol{\theta}^{(t)})}\right\}$$

.

1.6.4 The Gibbs sampler

Another method of constructing a suitable Markov chain is the Gibbs sampler (Geman and Geman, 1984). This is an iterative procedure that works by generating each component of $\boldsymbol{\theta}$ one at a time from a univariate conditional distribution. Suppose that the unknown parameter vector $\boldsymbol{\theta}$ has p components $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ and let the tth iterate generated be denoted by $\boldsymbol{\theta}^{(t)}$. To generate a sample from the posterior distribution with density function $f(\boldsymbol{\theta}|\boldsymbol{y})$, we choose starting values $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \ldots, \theta_p^{(0)})$ and the Gibbs sampler draws $\boldsymbol{\theta}^{(j)} = (\theta_1^{(j)}, \ldots, \theta_p^{(j)})$ from $\boldsymbol{\theta}^{(j-1)} = (\theta_1^{(j-1)}, \ldots, \theta_p^{(j-1)})$ in p steps as follows:

1. Sample $\theta_1^{(j)}$ from $f(\theta_1|\theta_2^{(j-1)}, \theta_3^{(j-1)}, \ldots, \theta_p^{(j-1)}, \boldsymbol{y})$

2. Sample $\theta_2^{(j)}$ from $f(\theta_2|\theta_1^{(j)}, \theta_3^{(j-1)}, \theta_4^{(j-1)}, \ldots, \theta_p^{(j-1)}, \boldsymbol{y})$

3. Sample $\theta_3^{(j)}$ from $f(\theta_3|\theta_1^{(j)}, \theta_2^{(j)}, \theta_4^{(j-1)}, \ldots, \theta_p^{(j-1)}, \boldsymbol{y})$

...

i. Sample $\theta_i^{(j)}$ from $f(\theta_i|\theta_1^{(j)}, \theta_2^{(j)}, \ldots, \theta_{i-1}^{(j)}, \theta_{i+1}^{(j-1)}, \theta_{i+2}^{(j-1)}, \ldots, \theta_p^{(j-1)}, \boldsymbol{y})$

...

p. Sample $\theta_p^{(j)}$ from $f(\theta_p|\theta_1^{(j)}, \theta_2^{(j)}, \ldots, \theta_{p-1}^{(j)}, \boldsymbol{y})$

At each step, we sample from the full conditional posterior distribution, conditional on the new values of parameters already sampled in earlier steps, and on the old values of parameters still to be sampled in later steps. The end result is a sample $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(N)}$.

1.6.5 Data augmentation

The Gibbs sampler particularly lends itself to problems involving data augmentation. This method was originally proposed by Tanner and Wong (1987) and is often used in situations where some data are unobserved or missing, as will be the case in this thesis. Suppose that $\boldsymbol{y}$ is the observed data while $\boldsymbol{z}$ represents data which are unobserved or missing and suppose that the posterior distribution of the parameter vector $\boldsymbol{\theta}$, $f(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{z})$ is easy to sample from, possibly using MCMC methods. Then the conditional distribution of the unobserved data $f(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{\theta})$ may also be sampled from. A Gibbs sampler may then be constructed as follows. Initial values are chosen for the parameters $\boldsymbol{\theta}$ and the unobserved data $\boldsymbol{z}$, and an iterative procedure is then carried out. At each iteration, a draw from $f(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{z})$ is made using the current sampled values of the unobserved data $\boldsymbol{z}$, and then the unobserved data are sampled from $f(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{\theta})$ given the updated values for the parameters $\boldsymbol{\theta}$.

1.6.6 Computation for model determination

Chib (1995) introduced an approach for computing the marginal likelihood of a model m from the output of a Gibbs sampling scheme. In order to use this method, it is necessary that all normalising constants of the full conditional distributions in the Gibbs sampler be known. Chib uses the fact that the marginal likelihood is also the normalising constant of the posterior density to arrive at the following identity:

$$f(\boldsymbol{y}|m) = \frac{f(\boldsymbol{y}|m,\boldsymbol{\theta}_m)f(\boldsymbol{\theta}_m|m)}{f(\boldsymbol{\theta}|\boldsymbol{y},m)} \tag{1.4}$$

i.e. the marginal likelihood is equal to the product of the likelihood and the prior (with all integrating constants included) over the posterior density of $\boldsymbol{\theta}$. The identity (1.4) holds true for all values of $\boldsymbol{\theta}$. The method for evaluating this runs as follows. Firstly, a high posterior density point (for example, the posterior mean) $\boldsymbol{\theta}^*$ is chosen. Suppose that the posterior density estimate at $\boldsymbol{\theta}^*$ is denoted by $\hat{f}(\boldsymbol{\theta}^*|\boldsymbol{y})$. Then the log of the marginal likelihood is

$$\log \hat{f}(\boldsymbol{y}|m) = \log f(\boldsymbol{y}|\boldsymbol{\theta}^*) + \log f(\boldsymbol{\theta}^*) - \log \hat{f}(\boldsymbol{\theta}^*|\boldsymbol{y}) \tag{1.5}$$

Clearly the first two terms in this expression are generally easily evaluated, leaving only the posterior density estimate $\hat{f}(\boldsymbol{\theta}^*|\boldsymbol{y})$ which can be found from the Gibbs output. Consider the specific case where we have augmented data $\boldsymbol{z}$ and unknown parameters $\boldsymbol{\theta}$, and suppose that Gibbs sampling is applied to the complete conditional densities

$$f(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{z}); \qquad f(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{\theta})$$

Let the output from the Gibbs algorithm be given by $\{\boldsymbol{\theta}^{(g)}, \boldsymbol{z}^{(g)}\}_{g=1}^{G}$. The posterior density can be written as

$$f(\boldsymbol{\theta}|\boldsymbol{y}) = \int f(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{z}) f(\boldsymbol{z}|\boldsymbol{y}) d\boldsymbol{z} = E_{\boldsymbol{z}}(f(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{z}))$$

so from the Gibbs sampler output, a Monte Carlo estimate of $f(\boldsymbol{\theta}|\boldsymbol{y})$ at $\boldsymbol{\theta}^*$ is given by the sample mean:

$$\hat{f}(\boldsymbol{\theta}^*|\boldsymbol{y}) = \frac{1}{G}\sum_{g=1}^{G} f(\boldsymbol{\theta}^*|\boldsymbol{y}, \boldsymbol{z}^{(g)})$$

since $\boldsymbol{z}^{(g)}$ is a draw from the distribution $\boldsymbol{z}|\boldsymbol{y}$. So, substituting in (1.5), the log marginal likelihood may be evaluated as

$$\log \hat{f}(\boldsymbol{y}|m) = \log f(\boldsymbol{y}|\boldsymbol{\theta}^*) + \log f(\boldsymbol{\theta}^*) - \log \left\{ \frac{1}{G}\sum_{g=1}^{G} f(\boldsymbol{\theta}^*|\boldsymbol{y}, \boldsymbol{z}^{(g)}) \right\}$$

Note that if the Gibbs sampling scheme contains more than two sampling blocks, this approach may be extended.

1.6.7 Reversible Jump Markov chain Monte Carlo (RJMCMC)

Reversible Jump (Green, 1995) provides a MCMC method that is capable of jumping between parameter subspaces of differing dimensionality. It therefore provides a framework for model determination, a situation where there is a discrete choice between a set of models, each with an associated parameter vector of differing dimension with an interpretation depending on the model in question. RJMCMC for model choice involves constructing a Markov chain which simulates from $f(m, \boldsymbol{\theta}_m)$, the joint distribution over models and associated parameters. The Reversible Jump algorithm involves proposing a move type p to a parameter subspace of potentially different dimension from the current subspace via a proposal distribution $q_p()$. Let $\boldsymbol{\theta}_m^{(t)}$ denote the set of parameters in the current model and $\boldsymbol{\theta}'_m$ the set of parameters in the proposed model, and suppose that data $\boldsymbol{y}$ are observed. If the proposed move is to a model with parameter vector $\boldsymbol{\theta}'_m$ of higher dimension than the current set of parameters $\boldsymbol{\theta}_m^{(t)}$, then $\boldsymbol{\theta}'_m$ can be constructed by generating a vector $\boldsymbol{u}$ which has dimension equal to the difference in dimensions of the two models, using proposal distribution $q_p(\boldsymbol{u})$ and setting $\boldsymbol{\theta}'_m = g_p(\boldsymbol{\theta}_m^{(t)}, \boldsymbol{u})$ where g is a one-to-one function. The 'reverse' move from a model with parameter subspace of higher dimension to one of lower dimension is achieved by applying the inverse transformation $(\boldsymbol{\theta}'_m, \boldsymbol{u}') = g_p^{-1}(\boldsymbol{\theta}_m^{(t)})$ and discarding $\boldsymbol{u}'$. The proposed move should be accepted with Reversible Jump probability α, where for a move to a parameter space of higher dimension:

$$\alpha = \min\left\{1, \frac{f(\boldsymbol{\theta}'_m|\boldsymbol{y})j(p,\boldsymbol{\theta}'_m)}{f(\boldsymbol{\theta}_m^{(t)}|\boldsymbol{y})j(p,\boldsymbol{\theta}_m^{(t)})q(\boldsymbol{u}|\boldsymbol{\theta}_m^{(t)})}\left|\frac{\partial(\boldsymbol{\theta}'_m)}{\partial(\boldsymbol{\theta}_m^{(t)},\boldsymbol{u})}\right|\right\} \tag{1.6}$$

and for a move to a parameter space of lower dimension:

$$\alpha = \min\left\{1, \frac{f(\boldsymbol{\theta}'_m|\boldsymbol{y})j(p,\boldsymbol{\theta}'_m)q(\boldsymbol{u}'|\boldsymbol{\theta}'_m)}{f(\boldsymbol{\theta}_m^{(t)}|\boldsymbol{y})j(p,\boldsymbol{\theta}_m^{(t)})}\left|\frac{\partial(\boldsymbol{\theta}'_m,\boldsymbol{u}')}{\partial(\boldsymbol{\theta}_m^{(t)})}\right|\right\} \tag{1.7}$$

where $j(p, \boldsymbol{\theta})$ is the probability of making move type p given the state of the Markov chain $\boldsymbol{\theta}_m$.

The algorithm runs as follows. Choosing initial values $\boldsymbol{\theta}_0$ and proposal density q, the following iterative process is carried out.

1. Choose a move type p with probability $j(p, \boldsymbol{\theta}_m^{(t)})$
2. Using the current value of the chain $\boldsymbol{\theta}_m^{(t)}$, propose a new value $\boldsymbol{\theta}'_m$ using the proposal distribution q_p if necessary and the transformation g_p.
3. Accept the proposal with probability α defined in (1.6) and (1.7).
4. If accepted, set $\boldsymbol{\theta}_m^{(t+1)} = \boldsymbol{\theta}'_m$, otherwise set $\boldsymbol{\theta}_m^{(t+1)} = \boldsymbol{\theta}_m^{(t)}$.
5. Return to step 1 and repeat.

For many applications, a major obstacle to the efficient implementation of Reversible Jump is the difficulty in finding a suitable proposal distribution. However, due to the model parameterisation used here, such a proposal distribution is available.

1.7 Outline of the Thesis

In this Chapter, we have introduced the basic theory that underpins this thesis. In Chapter 2, we give an outline of previous work specifically for ordinal data, including classical approaches but focusing primarily on Bayesian methods. In Chapter 3, we extend some of these approaches to develop a full methodology for modelling multivariate ordinal or binary data (or a mixture of both). Goodness-of-fit is also discussed and the approach illustrated with examples. In Chapter 4, we discuss the issue of model determination and give a Reversible Jump MCMC method for moving between directed decomposable graphical models. The RJMCMC algorithm is applied to data where the classifying variables have a natural ordering. Results are compared with other approaches in the literature. The method is extended in Chapter 5 for undirected decomposable graphical models and applied to data where there is no natural ordering to the classifying variables. In Chapters 3, 4 and 5, covariates are not considered. In Chapter 6, we incorporate covariates into the model and also give a further extension to the Reversible Jump methodology to assess covariate model selection in two examples. Conclusions are discussed in Chapter 7.

Chapter 2

Review of Previous Work

2.1 Models for Univariate Ordinal Data

The focus of this thesis will be on ordinal data. Such data occur when the response, which may be multivariate, takes the form of discrete, ordered categories. In this way, it is different from nominal data. The methods described in the previous section are sometimes applied to ordinal data but they are somewhat unsatisfactory as they ignore the ordinal structure of the data, i.e. parameter estimates are invariant to orderings of categories. Note that binary data may always be treated as a special case of ordinal data, where there are only two categories.

Suppose individuals $i = 1, \ldots, n$ are categorised into k ordered categories. The categorical response vector $\boldsymbol{y}$ is observed, where y_i is the response category of the i^{th} individual.

$$\boldsymbol{y} = (y_1, y_2, \ldots, y_n) \qquad y_i \in \{1, ..., k\}$$

Suppose also that we observe covariates $\boldsymbol{x}_i$ for each individual, and that the k ordered categories of the response have probabilities $p_1(\boldsymbol{x}_i), p_2(\boldsymbol{x}_i), \ldots, p_k(\boldsymbol{x}_i)$. Define the cumulative probability for category j to be

$$F_j(\boldsymbol{x}_i) = p_1(\boldsymbol{x}_i) + p_2(\boldsymbol{x}_i) + \cdots + p_j(\boldsymbol{x}_i)$$

McCullagh (1980) introduced an important class of regression models for ordinal data. These models are all based on the assumption of the existence of an underlying continuous random variable z_i for each y_i. The categories of the (univariate) response y_i are envisaged as contiguous intervals on the continuous scale for z_i, the end points of the intervals are called cut points and are denoted by $\theta_0, \theta_1, \theta_2, \ldots, \theta_{k-1}, \theta_k$ where the response has k categories. Hence, $y_i = c$ if and only if $z_i \in (\theta_{c-1}, \theta_c]$. The first and last cut points are set to $-\infty$ and ∞ respectively.

$$\theta_0 = -\infty \qquad\qquad \theta_k = \infty$$

All the models suggested by McCullagh share this assumption, but they differ in their assumptions concerning the distribution of the latent variable.

The cumulative link regression model is defined by:

$$F_j(\boldsymbol{x}_i) = H(\theta_j - \boldsymbol{x}_i^T \boldsymbol{\beta})$$

where $\boldsymbol{\beta}$ is a $k \times 1$ vector of unknown parameters, and $H(.)$ is a known cdf linking the cumulative probabilities $F_j(\boldsymbol{x})$ with the linear structure $\boldsymbol{x}_i^T \boldsymbol{\beta}$. To ensure that the parameters are identifiable, it is necessary to impose a further constraint. Typically this might involve constraining an intercept parameter in $\boldsymbol{\beta}$ to be equal to zero, or constraining θ_1.

The cumulative link regression model for y_i is equivalent to the following model for the underlying latent variable z_i:

$$z_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \epsilon_i$$

where ϵ_i has cumulative distribution function H. Then,

$$\begin{aligned} P(y_i \leq j) = F_j(\boldsymbol{x}_i) &= H(\theta_j - \boldsymbol{x}_i^T \boldsymbol{\beta}) \\ &= P(\epsilon \leq \theta_j - \boldsymbol{x}_i^T \boldsymbol{\beta}) \\ &= P(\epsilon + \boldsymbol{x}_i^T \boldsymbol{\beta} \leq \theta_j) \\ &= P(z_i \leq \theta_j) \end{aligned}$$

2.1.1 The proportional odds model

The proportional odds model is obtained if H is the cdf of the standard logistic distribution. The model is therefore defined by:

$$\log\left(\frac{F_j(\boldsymbol{x})}{1-F_j(\boldsymbol{x})}\right) = \theta_j - \boldsymbol{x}_i^T\boldsymbol{\beta}$$

2.1.2 The proportional hazards model

The proportional hazards model is obtained if H is the cdf of the extreme value distribution. The model is therefore defined by:

$$\log\left[\log\left(1-F_j(\boldsymbol{x})\right)\right] = \theta_j - \boldsymbol{x}_i^T\boldsymbol{\beta}$$

2.1.3 The probit model

This is obtained if H is the standard Normal cdf. The model is therefore defined by:

$$F_j(\boldsymbol{x}) = \boldsymbol{\Phi}(\theta_j - \boldsymbol{x}_i^T\boldsymbol{\beta})$$

Once an appropriate link function has been chosen, the models may then be fitted using maximum likelihood. Inferences about the model are based on the associated asymptotic theory.

The proportional odds model was applied to the data in Table 2.1 taken from Holmes and Williams (1954) which shows 1398 children classified according to their tonsil size. The response has 3 ordered categories: Not Enlarged, Enlarged and Greatly Enlarged. There is one covariate: whether or not a child is a carrier of the *Streptococcus pyogenes* virus. Parameter estimates with associated standard deviations were obtained by maximum likelihood and are displayed in Table 2.2. The intercept is set to zero and β_1 is the additional effect on the latent logistic variable scale of carriers over non-carriers.

	Present but not enlarged	Enlarged	Greatly enlarged
Carriers	19	29	24
Non-carriers	497	560	269

Table 2.1: Tonsil size of carriers and non-carriers of *Streptococcus pyogenes*

Parameter	Estimate (s.e.)
θ_1	--0.509 (0.056)
θ_2	1.363 (0.067)
β_1	-0.603 (0.227)

Table 2.2: Parameter estimates from the proportional odds model for the tonsil data

2.2 Bayesian Approaches for Univariate Ordinal Data

Various authors have found that there can be problems with the maximum likelihood approach. Griffiths and Pope (1987) found the maximum likelihood estimator to have significant bias for small samples, while Zellner and Rossi (1984) also commented on the inaccuracy of the normal asymptotic approximation for small sample size. The maximum likelihood approach also has no meaningful interpretation if the model contains any covariates which are perfect predictors. The Bayesian approach developed by Albert and Chib (1993) and outlined in the next section overcomes these problems.

The Bayesian approach is again based on the assumption of the existence of an underlying continuous random variable z_i for each respondent, as in the Classical approach.

We wish to estimate the unknown parameter vector $\boldsymbol{\beta}$ using a Bayesian approach. We do this by applying Bayes theorem.

Applying Bayes theorem and for any choice of prior $\pi(\boldsymbol{\beta})$ for $\boldsymbol{\beta}$, the posterior density of $\boldsymbol{\beta}$ is given by:

$$f(\boldsymbol{\beta},\boldsymbol{\theta}|\boldsymbol{y}) \propto \pi(\boldsymbol{\beta})\pi(\boldsymbol{\theta})\prod_{i=1}^{n}\prod_{j=1}^{k}[H(\theta_j - \boldsymbol{x}_i^T\boldsymbol{\beta}) - H(\theta_{j-1} - \boldsymbol{x}_i^T\boldsymbol{\beta})]^{I(y_i=j)} \qquad (2.1)$$

which is somewhat intractable. However, Albert and Chib (1993) suggested a simulation-based approach for computing the exact posterior distribution for $\boldsymbol{\beta}$ which uses the ideas of data augmentation (Tanner and Wong, 1987) and the Gibbs sampler. The method runs as follows: Suppose that the link function H(.) is chosen to be Φ, leading to the probit model. This corresponds to an assumption that the latent continuous variables z_i are independently normally distributed:

$$z_i \sim N(\boldsymbol{x}_i^T\boldsymbol{\beta}\ ,\ 1) \qquad (2.2)$$

We assume the existence of ordered cut points $\theta_1, \theta_2, \ldots, \theta_{k-1}$ such that y_i takes the c^{th} level if z_i falls between the lower and upper cut points for the c^{th} level.

$$y_i = c \quad \text{if} \quad \theta_{c-1} \leq z_i < \theta_c \qquad (2.3)$$

As in the classical approach, the first and last cut points are set to $-\infty$ and ∞ respectively: $\theta_0 = -\infty, \theta_k = \infty$. As it stands the model is over parameterised so the additional constraint $\theta_1 = 0$ is imposed.

We now include the latent variables $z_1, \ldots, z_n$ as unknown parameters. Under this formulation, the parameters $\boldsymbol{z} = (z_1, \ldots, z_n), \boldsymbol{\beta}, \boldsymbol{\theta} = (\theta_2, \ldots, \theta_{k-1})$ are unknown and may be estimated using Bayes Theorem. Applying Bayes Theorem to the unknown parameters yields the following result:

$$f(\boldsymbol{z},\boldsymbol{\beta},\boldsymbol{\theta}|\boldsymbol{y}) \propto f(\boldsymbol{y}|\boldsymbol{z},\boldsymbol{\beta},\boldsymbol{\theta})f(\boldsymbol{z},\boldsymbol{\beta},\boldsymbol{\theta})$$

Since $\boldsymbol{y}$ is fully determined by $\boldsymbol{z}$ and $\boldsymbol{\theta}$ this reduces to:

$$f(\boldsymbol{z},\boldsymbol{\beta},\boldsymbol{\theta}|\boldsymbol{y}) \propto f(\boldsymbol{y}|\boldsymbol{z},\boldsymbol{\theta})f(\boldsymbol{z},\boldsymbol{\beta},\boldsymbol{\theta})$$

Then decomposing $f(\mathrm{z},\boldsymbol{\beta},\boldsymbol{\theta})$ and choosing independent priors for $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$,

$$f(\mathrm{z},\boldsymbol{\beta},\boldsymbol{\theta}|\mathrm{y}) \quad \propto \quad f(\mathrm{y}|\mathrm{z},\boldsymbol{\theta})f(\mathrm{z}|\boldsymbol{\beta})f(\boldsymbol{\beta})f(\boldsymbol{\theta}) \qquad (2.4)$$

where $f(\mathbf{y}|\mathbf{z},\boldsymbol{\theta})$ is deterministic, as defined in (2.3), $f(\mathbf{z}|\boldsymbol{\beta})$ is the density of $N(\boldsymbol{x}_i^T\boldsymbol{\beta},1)$, and $f(\boldsymbol{\beta})$ and $f(\boldsymbol{\theta})$ are the prior densities for $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ respectively. We choose these priors as follows:

- $\boldsymbol{\beta} \sim N_p(\mathbf{0},\boldsymbol{T})$ where $\boldsymbol{T} = \text{diag}(\tau_1^2,\ldots,\tau_p^2)$ where p is the number of explanatory variables, N_p denotes the p-dimensional multivariate normal distribution, and τ_i^2 is large.
- $f(\boldsymbol{\theta}) \propto 1$ (uniform), subject to ordering constraints $\theta_2 < \theta_3 < \cdots < \theta_{k-1}$

Substituting these priors into Equation 2.4 gives a somewhat intractable joint posterior distribution. Albert and Chib suggest using a Gibbs sampler to generate from the conditional posteriors of the parameters, thus yielding a dependent sample from approximately the joint posterior distribution. The method runs as follows:

2.2.1 Algorithm 1

1. Starting with initial values for all parameters, sample the parameter vector $\boldsymbol{\beta}$ from its conditional distribution

$$\boldsymbol{\beta}|\boldsymbol{z},\boldsymbol{\theta} \sim N_p\left((\boldsymbol{X}^T\boldsymbol{X}+\boldsymbol{T}^{-1})^{-1}\boldsymbol{X}^T\boldsymbol{z}\ ,\ (\boldsymbol{X}^T\boldsymbol{X}+\boldsymbol{T}^{-1})^{-1}\right) \tag{2.5}$$

2. Sample the new latent data from their conditional distributions

$$z_i|\boldsymbol{\beta},\boldsymbol{\theta} \sim N(\boldsymbol{x}_i^T\boldsymbol{\beta},1) \tag{2.6}$$

   with z_i truncated to the interval $(\theta_{y_i-1},\theta_{y_i})$. This distribution is non-standard and may be sampled from using the inverse cumulative distribution function method of Devroye (1986), pages 27-29.

3. Let $\{z_i : y_i = j\}$ denote the set of latent variables z_i with corresponding observed data y_i taking level j. Then the new cut points are sampled from their conditional distribution

$$\theta_j|\boldsymbol{z} \sim \text{Uniform}\left(\max_i\{z_i : y_i = j\}, \min_i\{z_i : y_i = j+1\}\right) \tag{2.7}$$

4. Go back to step 1 and repeat.

Algorithm 1 was applied to the data in Table 2.1 using 10,000 iterations of the Gibbs sampler. Parameter estimates for $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ are displayed in Table 2.3. Remember that θ_1 is set to 0. These results are not comparable with those obtained using the Classical Proportional Odds model approach, because they use different link functions. The logit link was used in the Classical approach. The logistic distribution can be shown to be approximately linearly related to the t-distribution with 8 degrees of freedom as discussed in Albert and Chib (1993). Using results from Ntzoufras et al. (2003) the relationship between any linear predictor provided by the logit and t(8) link functions is approximated by

Parameter	Estimate (s.e.)
θ_2	1.206 (0.152)
β_0	-0.342 (0.069)
β_1	-0.365 (0.135)

Table 2.3: Parameter estimates from the Bayesian probit model for the tonsil data

$$\frac{g'_{L_1}(1/2)}{g'_{L_2}(1/2)}$$

where g_{L_1} is the link function for the logistic model and g_{L_2} is the link function for the $t(8)$ model, both evaluated at the median for the best approximation. This can be seen to be equivalent to:

$$\beta_{logit} = \frac{35}{16\sqrt{2}}\beta_{t(8)} \tag{2.8}$$

Therefore, if we can implement the Bayesian approach with a t(8) link, the parameters estimates can be transformed by Equation (2.8) to be compared with the classical approach logit link results.

Albert and Chib (1993) also provide an algorithm for the $t(\nu)$ link function. The latent variables z_i are now assumed to be independently distributed from the t distribution with locations $\boldsymbol{x}_i^T\boldsymbol{\beta}$ and degrees of freedom ν (which we will choose to be 8).

$$z_i \sim t(\boldsymbol{x}_i^T\boldsymbol{\beta}, \nu) \tag{2.9}$$

Introducing the additional random variable σ_i, this is equivalent to:

$$\begin{aligned} z_i|\sigma_i^2 &\sim N(\boldsymbol{x}_i^T\boldsymbol{\beta}, \sigma_i^2) && (2.10)\\ \frac{1}{\sigma_i^2} &\sim \text{Gamma}\left(\frac{\nu}{2}, \frac{\nu}{2}\right) && (2.11) \end{aligned}$$

where we use the following parameterisation for the gamma distribution:

$$f\left(\frac{1}{\sigma_i^2}\right) = \frac{(\frac{\nu}{2})^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})}\left(\frac{1}{\sigma_i^2}\right)^{\frac{\nu}{2}-1} \exp\left(-\frac{\nu}{2\sigma_i^2}\right) \tag{2.12}$$

Using the same priors as for the probit link case, we arrive at the conditional distributions for the unknown parameters which are then used to implement a Gibbs sampler, as described in Algorithm 2:

2.2.2 Algorithm 2

1. Starting with initial values for all parameters, sample the parameter vector $\boldsymbol{\beta}$ from its conditional distribution

$$\boldsymbol{\beta}|\boldsymbol{z}, \Sigma \sim N_p\left((\boldsymbol{X}^T\Sigma^{-1}\boldsymbol{X} + \boldsymbol{T}^{-1})^{-1}\boldsymbol{X}^T\Sigma^{-1}\boldsymbol{z}\ ,\ (\boldsymbol{X}^T\Sigma^{-1}\boldsymbol{X} + \boldsymbol{T}^{-1})^{-1}\right)$$

2. Sample the parameter matrix Σ =diag$(\sigma_1^2, \ldots, \sigma_n^2)$ from its conditional distribution

$$\sigma_i^2|z_i, \boldsymbol{\beta} \sim \text{Inverse Gamma}\left(\frac{\nu+1}{2}, \frac{\nu + (z_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2}{2}\right)$$

i.e.

$$\frac{1}{\sigma_i^2}|z_i, \boldsymbol{\beta} \sim \text{Gamma}\left(\frac{\nu+1}{2}, \frac{\nu + (z_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2}{2}\right)$$

with the parameterisation defined in (2.12).

3. Sample the new latent data from their conditional distributions

$$z_i|\boldsymbol{\beta}, y_i, \sigma_i^2, \boldsymbol{\theta} \sim N(\boldsymbol{x}_i^T\boldsymbol{\beta}, \sigma_i^2)$$

   with z_i truncated to the interval $(\theta_{y_i-1}, \theta_{y_i})$.

4. Sample the new cut points from their conditional distribution

$$\theta_j|\boldsymbol{z} \sim \text{Unif}\left(\max_i\{z_i : y_i = j\}, \min_i\{z_i : y_i = j+1\}\right)$$

5. Go back to step 1 and repeat.

Algorithm 2 was applied to the data in Table 2.1 and the parameters estimates (along with standard deviations) are displayed in Table 2.4.

Parameter	Estimate (s.e.)
θ_2	1.227 (0.012)
β_0	-0.343 (0.003)
β_1	-0.380 (0.008)

Table 2.4: Parameter estimates from the Bayesian t(8) model for the tonsil data

The results must be transformed from those given by the $t(8)$ link to be comparable with those given by the logit link using (2.8). Table 2.5 shows the parameter estimates for the data in Table 2.1, with the first column showing the Bayesian results and the second column showing the Classical results.

Parameter	Bayesian Estimate	Classical Estimate
θ_1	-0.530	-0.509
θ_2	1.367	1.363
β_1	-0.588	-0.603

Table 2.5: Comparison of parameter estimates from the Bayesian and classical approaches for the tonsil data

Comparing the results from the classical and Bayesian approaches for each data set shows that the two approaches give very similar results. We can therefore conclude that the approach suggested by Albert and Chib (1993) provides an attractive and flexible alternative to the classical approach for univariate ordinal data.

However, there are some computational issues. Although the algorithm described above is straightforward to implement, it can be difficult to obtain satisfactory convergence. This is due to the cut point generation step which can only sample values between the maximum latent data value in the lower category and the minimum latent data value in the upper category that it divides. Therefore, if there are many individuals in the corresponding cell of the table, leading to many values for the latent data, the cut points can be extremely slow-moving with high autocorrelations. There have been several studies into methods for accelerating convergence for the cut points. Cowles (1996) suggests the use of a multivariate Metropolis-Hastings step which updates cut points and latent variables simultaneously, while Nandram and Chen (1996) further improved this with a proposal density based on the Dirichlet distribution. However, the latter is only effective when cell counts are reasonable evenly distributed. For multivariate data, Ishwaran (2000) bypasses the problem entirely by proposing a reparameterisation with covariate specific cut points that allows parameter estimation to be carried out via a leapfrog hybrid Monte Carlo approach.

2.3 A Bayesian Approach for Multivariate Binary Data

We have considered various modelling approaches for univariate ordinal data, and now turn our attention to the multivariate case. Chib and Greenberg (1998) build on the framework laid down by Albert and Chib (1993) to model multivariate binary data using a multivariate probit model as suggested by

Ashford and Sowden (1970). Suppose that individuals $i = 1, \ldots, n$ are classified by binary variables $j = 1, \ldots, p$. Independent binary response vectors $\boldsymbol{y}_i$ are observed, where respondent i takes category y_{ij} for the j^{th} variable. Suppose also that the set of covariates $\boldsymbol{x}_{ij}$ are observed for the j^{th} response. The multivariate probit model says that the likelihood of observing response vector $\boldsymbol{y}_i$ given parameters $\boldsymbol{\beta}$, Σ and covariates $\boldsymbol{x}_{ij}$ is:

$$f(\boldsymbol{y}_i|\boldsymbol{\beta}, \Sigma) = \int_{A_{ip}} \cdots \int_{A_{i1}} \phi_p(t|\mathbf{0}, \Sigma) dt$$

where $\phi_p(t|\mathbf{0}, \Sigma)$ is the density of a $p-$variate normal distribution with mean vector $\mathbf{0}$ and correlation matrix $\Sigma = \{\sigma_{jk}\}$, A_{ij} is the interval

$$A_{ij} = \begin{cases} (-\infty, \boldsymbol{x}'_{ij}\boldsymbol{\beta}_j) & \text{if } \ y_{ij} = 1 \\ (\boldsymbol{x}'_{ij}\boldsymbol{\beta}_j, \infty) & \text{if } \ y_{ij} = 0 \end{cases}$$

$\boldsymbol{\beta}_j \in R^{k_j}$ is an unknown parameter vector and $\boldsymbol{\beta}' = (\boldsymbol{\beta}'_1, \ldots, \boldsymbol{\beta}'_p)$. Note that the problem is parameterised in terms of the correlation matrix in order to ensure identifiability for the parameters. This is analogous to assuming $\sigma^2 = 1$ in the latent normal distribution of Albert and Chib (1993). The multivariate probit model is then re-formulated using the methods of Albert and Chib (1993) as described in section 2.2. Specifically, latent normal random variables $\boldsymbol{z}_i \sim N_p(\boldsymbol{X}_i\boldsymbol{\beta}, \Sigma)$ are introduced, such that

$$y_{ij} = I(z_{ij} > 0) \tag{2.13}$$

where $\boldsymbol{X}_i$=diag$(\boldsymbol{x}'_{i1}, \ldots, \boldsymbol{x}'_{ip})$. Implicitly, the single cut point θ_1 is set to be zero, again for identifiability.

Let $\boldsymbol{\sigma} = (\sigma_{12}, \sigma_{13}, \ldots, \sigma_{p-1,p})$ denote the $p(p-1)/2$ distinct elements of Σ. Then the values of $\boldsymbol{\sigma}$ that allow a positive definite matrix Σ form a convex solid body in the hypercube $[-1, 1]^p$; denote this set by C. Using Bayes' theorem, the posterior density of the unknown parameters $\boldsymbol{\beta}, \Sigma$ and $\boldsymbol{Z}$ given the observed data $\boldsymbol{y}$ is:

$$\begin{aligned} f(\boldsymbol{\beta}, \boldsymbol{\sigma}, \boldsymbol{Z}|\boldsymbol{y}) &\propto f(\boldsymbol{y}|\boldsymbol{\beta}, \Sigma, \boldsymbol{Z}) f(\boldsymbol{\beta}, \Sigma, \boldsymbol{Z}) \\ &\propto f(\boldsymbol{y}|\boldsymbol{\beta}, \Sigma, \boldsymbol{Z}) f(\boldsymbol{Z}|\boldsymbol{\beta}, \Sigma) f(\boldsymbol{\beta}, \Sigma) \\ &\propto f(\boldsymbol{\beta}, \Sigma) \prod_{i=1}^{n} \phi_p(\boldsymbol{Z}_i|\boldsymbol{\beta}, \Sigma) f(\boldsymbol{y}_i|\boldsymbol{Z}_i, \boldsymbol{\beta}, \Sigma) \end{aligned}$$

where from (2.13),

$$f(\boldsymbol{y}_i|\boldsymbol{Z}_i,\boldsymbol{\beta},\Sigma) = \prod_{j=1}^{p}[I(z_{ij}>0)I(y_{ij}=1)+I(z_{ij}\leq 0)I(y_{ij}=0)]$$

and $\phi_p(\boldsymbol{Z}_i|\boldsymbol{\beta},\Sigma)$ is the multivariate normal density with the constraint that $\boldsymbol{\sigma} \in C$.

Chib and Greenberg assume prior independence of $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}$ and assign multivariate normal distributions to each, with the prior distribution for $\boldsymbol{\sigma}$ having mean σ_0 and variance $\boldsymbol{G}_0^{-1}$ and being truncated to the region C. Following Albert and Chib, a Gibbs sampler technique may then be used to evaluate the posterior distribution, and runs as follows.

2.3.1 Chib and Greenberg's method

1. Sample the latent data $\boldsymbol{Z}_i$ from their conditional posterior distribution $\boldsymbol{Z}_i|\boldsymbol{y}_i,\boldsymbol{\beta},\Sigma$ which is truncated p-dimensional multivariate normal. This can be sampled using the method of Geweke (1991), which consists of a cycle of p Gibbs sampler steps, each from a univariate truncated normal distribution.

2. Sample the parameter vector $\boldsymbol{\beta}$ from its conditional posterior distribution $\boldsymbol{\beta}|\boldsymbol{Z},\boldsymbol{y},\Sigma$ which is multivariate normal.

3. Sample the off-diagonal elements $\boldsymbol{\sigma}$ of Σ from their joint conditional distribution $\boldsymbol{\sigma}|\boldsymbol{Z},\boldsymbol{y},\boldsymbol{\beta}$. This distribution is non-standard and requires the use of the Metropolis-Hastings algorithm (Hastings, 1970) outlined below.

4. Go back to step 1 and repeat.

The conditional posterior distribution for $\boldsymbol{\sigma}$ is given by:

$$\begin{aligned} f(\boldsymbol{\sigma}|\boldsymbol{Z},\boldsymbol{\beta}) &\propto f(\boldsymbol{\sigma})f(\boldsymbol{Z}|\boldsymbol{\beta},\Sigma) \\ &\propto \phi_p(\boldsymbol{\sigma}|\boldsymbol{\sigma}_0,\boldsymbol{G}_0^{-1})\phi_p(\boldsymbol{Z}|\boldsymbol{x}_i\boldsymbol{\beta},\Sigma)I(\boldsymbol{\sigma}\in C) \end{aligned}$$

i.e. the product of two multivariate normal distributions truncated to the region C. This is sampled using the Metropolis-Hastings algorithm. Firstly, a suitable proposal distribution must be found. Chib and Greenberg suggest a procedure based on a proposal density that is tailored to the un-normalised target density $g(\boldsymbol{\sigma}|\boldsymbol{Z},\boldsymbol{\beta}) = f(\boldsymbol{\sigma})f(\boldsymbol{Z}|\boldsymbol{\beta},\Sigma)I(\boldsymbol{\sigma} \in C)$ (in the acceptance probability, the normalising constants would cancel). The proposal generating procedure uses a hierarchical variant of a random walk chain and involves a number of tuning parameters for added flexibility, leading to a proposal from the distribution $q(\boldsymbol{\sigma}'|\boldsymbol{\sigma},\boldsymbol{Z},\boldsymbol{\beta})$. This proposal is accepted with probability α where

$$\alpha(\boldsymbol{\sigma},\boldsymbol{\sigma}') = \min\left\{\frac{f(\boldsymbol{\sigma}')f(\boldsymbol{Z}|\boldsymbol{\beta},\Sigma)I(\boldsymbol{\sigma}' \in C)}{f(\boldsymbol{\sigma})f(\boldsymbol{Z}|\boldsymbol{\beta},\Sigma)I(\boldsymbol{\sigma} \in C)}\frac{q(\boldsymbol{\sigma}|\boldsymbol{\sigma}',\boldsymbol{Z},\boldsymbol{\beta})}{q(\boldsymbol{\sigma}'|\boldsymbol{\sigma},\boldsymbol{Z},\boldsymbol{\beta})},1\right\} \tag{2.14}$$

In unpublished work, Fronk (2003) uses a similar latent data approach to model binary data and applies a Reversible Jump algorithm to investigate model choice between competing DAGs.

2.4 Approaches for Multivariate Ordinal Data

Chen and Dey (2000) use a similar approach to that of Chib and Greenberg (1998) to model correlated ordinal data. They introduce a general class of scale mixtures of multivariate normal (SMMVN) link functions, a special case of which is the multivariate probit model.

Suppose that individuals $i = 1,\ldots,n$ are classified according to $j = 1,\ldots,p$ ordinal variable, with each ordinal variable having L levels. For each individual, the ordinal response vector $\boldsymbol{y}_i = (y_{i1}, y_{i2},\ldots,y_{ip})$ is observed, along with covariate vector $\boldsymbol{x}_{ij} = (x_{ij1}, x_{ij2},\ldots,x_{ijp_j})$ for each variable j. Let $\boldsymbol{\beta}_j = (\beta_{j1},\beta_{j2},\ldots,\beta_{jp_j})$ denote the corresponding vector of regression coefficients, with $\boldsymbol{\beta} = (\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\ldots,\boldsymbol{\beta}_p)'$. Again, following Albert and Chib (1993) the existence of an underlying multivariate random variable $\boldsymbol{z}_i = (z_{i1}, z_{i2},\ldots,z_{ip})$ is assumed, as is the existence of cut points which divide the range of $\boldsymbol{z}_{ij}$ into L

contiguous intervals corresponding to the categories. Let $\theta(j,c)$ denote the cth cut point for the jth variable. Then y_{ij} takes the cth level if z_{ij} falls between the lower and upper cut points for the cth level:

$$y_{ij} = c \quad \text{if} \quad \theta(j, c-1) \leq z_{ij} < \theta(j,c) \tag{2.15}$$

To ensure identifiability, $\theta(j,1) = 0$. The latent variables $\boldsymbol{z}_i$ are assumed to be independent and identically distributed multivariate normal random variables:

$$\boldsymbol{z}_i \sim N_p(\boldsymbol{x}_i\boldsymbol{\beta}, \kappa(\lambda)\boldsymbol{\Sigma}) \tag{2.16}$$

and

$$\lambda \sim f(\lambda) \tag{2.17}$$

where $\kappa(\lambda)$ is a positive function of one-dimensional positive-valued scale mixing variable λ and $f(\lambda)$ is a mixing distribution which is either discrete or continuous. $\boldsymbol{\Sigma}$ is taken to be in correlation form to ensure the identifiability of the parameters.

There are two difficult sampling problems to be tackled in order to fit this model, one is the generation of the cut points, the other is generating from the correlation matrix. In order to tackle these problems, Chen and Dey reparameterise the model:

$$\delta_j = \frac{1}{\theta(j, L-1)}, \quad \theta(j,k)^* = \delta_j\theta(j,k), \quad \beta_j^* = \delta_j\beta_j, \quad z_{ij}^* = \delta_j z_{ij}$$

for $j = 1, 2, \ldots, p$ and $i = 1, 2, \ldots, n$. Under this parameterisation, the SMMVN-link models defined in (2.15) and (2.16) become

$$y_{ij} = c \quad \text{if} \quad \theta(j, c-1)^* \leq z_{ij}^* < \theta(j,c)^*$$

and

$$\boldsymbol{z}_i^* \sim N_p(\boldsymbol{x}_i\boldsymbol{\beta}^*, \kappa(\lambda)\boldsymbol{\Sigma}^*)$$

where the reparameterised cut points are $-\infty = \theta(j,0) \leq \theta(j,1) = 0 \leq \theta(j,2) \leq \cdots \leq \theta(j, L-1) = 1 \leq \theta(j,L) = \infty$, and $\boldsymbol{\Sigma}^*$ is now in unrestricted covariance form. This makes posterior simulation of the parameter $\boldsymbol{\Sigma}^*$ much more straightforward. Note that under this new parameterisation, for each classifying variable there are only $L-3$ unknown cut points to be determined, as effectively, this approach is exactly equivalent to fixing the scale of the latent variables. However this is achieved by fixing two cut points $\theta(j,1)$ and $\theta(j,k-1)$ rather than one cut point $\theta(j,1)$ and a variance Σ_{jj}. Note therefore that this re-parameterisation is intractable for binary data, where there is only a single cut point.

A special case of the SMMVN model is the multivariate probit model. This can be found by taking $\kappa(\lambda)=1$ and $f(\lambda)=f(1)=1$. Other possible models contained within the class of SMMVN-link models are the the multivariate t-link models.

Posterior simulation is again carried out using a Gibbs sampler. Choosing independent multivariate normal and inverse Wishart priors for β^* and Σ^* respectively and uniform priors for the cut points, all conditional posterior distributions are standard. Therefore, no extra Metropolis-Hastings steps are necessary.

This approach has been successfully applied to genetic data (Kizilkaya et al., 2003) and dose-finding in clinical trials (Bekele and Thall, 2004).

2.5 Other Bayesian Approaches for Multivariate Ordinal Data

There have been few other attempts at modelling multivariate ordinal data in a Bayesian framework and those that there are tend to focus on multirater data, where items/individuals are rated by several different judges. Observer agreement can then be assessed. One such example is the paper by Johnson (1996), in which a hierarchical model is proposed. This follows a similar framework as Chen and Dey (2000) except with one important extra assumption, that is that there exists a 'true' rating scheme through which each item i can be assigned a latent trait measure z_i, referred to as the item score. Judges are assumed to rate items by first estimating this item score and then assigning an ordinal rating based on these scores. This is effectively the same assumption introduced by Albert and Chib (1993) except with the constraint that for each judge, there is the same underlying true score for a particular item. The approach can therefore not be applied where the classifying variables are in fact measuring completely different quantities. Using the same notation as above, except that the p classifying variables are now considered as p 'judges', and letting z_{ij} be judge j's estimate of item i's score z_i, the model can be written as

$$z_{ij} = z_i + b_{ij} + \alpha_{ij},$$

where

$$y_{ij} = k \qquad \text{if} \qquad \theta(j, k-1) < z_{ij} \leq \theta(j, k)$$

Here b_{ij} denotes the nonrandom item-specific bias for judge j, potentially modelled using covariates, and α_{ij} denotes the random component of the error made by judge j in estimating z_i. The α_{ij} are generally assumed to be item independent. Assigning appropriate prior distributions to the unknown parameters and using the latent data approach of Albert and Chib, posterior estimation is then carried out via a Gibbs sampler. Further work and applications of this method have been carried out by Johnson and Albert (1999), Johnson et al. (2002) and Ishwaran and Gatsonis (2000).

Rossi et al. (2001) use a similar approach to model multivariate ordinal data which arise from survey research, where respondents respond to a number of different questions, giving answers on an ordinal scale. Noting that respondents vary in their use of such a scale - for example, some use only the middle range of the scale, while some only use extremes - Rossi et al. introduce a model to account for these differences. In this situation, the p classifying variables are the p questions asked, and again, we let y_{ij} denote the response of individual i to question j. Again, using the approach of Albert and Chib, the response vector $\boldsymbol{y}_i$ is assumed to be a discrete version of a latent underlying continuous random variable $\boldsymbol{z}_i^*$ which Rossi et al. (2001) assume to have the distribution

$\boldsymbol{z}_i^* \sim N_p(\boldsymbol{\beta}_i^*, \Sigma_i^*)$. The discretising set of cut points $(\theta_0, \dots, \theta_k)$ are assumed to be common to all p variables. Note that this is an alternative approach to those previously discussed in that each individual is assumed to have a different latent mean and variance but with common cut points across variables, whereas others have assumed common mean and variance for individuals and different cut points across variables. As it stands this model is overparameterised; to overcome this, for each individual the latent variable $\boldsymbol{z}_i^*$ is assumed to be a location-scale shift of a common underlying latent variable $\boldsymbol{z}_i$:

$$\begin{aligned} \boldsymbol{z}_i^* &= \boldsymbol{\beta} + \tau_i + \sigma_i \boldsymbol{z}_i \\ \boldsymbol{z}_i &\sim N_p(\mathbf{0}, \boldsymbol{\Sigma}) \end{aligned}$$

Note that the original mean and covariance structure can be generated using $\boldsymbol{\beta}^* = \boldsymbol{\beta} + \tau_i$ and $\boldsymbol{\Sigma}^* = \sigma_i^2 \boldsymbol{\Sigma}$. The cut points have the identifiability constraint $\sum \theta_k =$ constant and are re-parameterised to take the quadratic form:

$$\theta_k = a + bk + ek^2$$

in order to allow for non-linear spread. Priors are chosen for all parameters in the hierarchical model and posterior estimation is carried out using a Gibbs sampler in five blocks with Metropolis-Hastings steps. The approach suggested by Rossi et al. has a high level of complexity due to the nature of the application, that of overcoming scale usage heterogeneity. Such a level of complexity is unlikely to be necessary to model standard multivariate ordinal data.

Chapter 3

A Bayesian Model for Multivariate Ordinal and Binary Data

In this chapter, we extend the approach for univariate data described by Albert and Chib (1993) and implemented in Chapter 2. In contrast to the approaches of Chib and Greenberg and Chen and Dey, this approach will be sufficiently general to encompass applications with either binary or ordinal variables or both. We make the same assumption that the ordinal categorical data are a discrete version of underlying continuous data. This means that we assume the existence of a latent continuous multivariate random variable associated with each response. The domain of the latent variable is divided by cut points into contiguous regions in $I\!R^p$ where p is the dimension of the data. We will focus on developing a model to fit ordinal multivariate data with no covariates.

3.1 The Model

For each individual $\boldsymbol{y}_i$, we assume the existence of a latent multivariate continuous variable $\boldsymbol{z}_i \in \mathrm{R}^p$. In the univariate case, the latent variables $\boldsymbol{z}_i$ were assumed to be normally distributed with mean $\boldsymbol{x}_i\boldsymbol{\beta}$ and variance 1. With no covariates this reduces to a mean β. If we were to follow the analogous setup in the multivariate case, the variance matrix of the latent data would be in correlation form. However, there are problems in working with the correlation matrix as outlined in Chapter 2. For these reasons we assume the covariance matrix to be in unrestricted form. Therefore, each $\boldsymbol{z}_i$ is assumed to be normally distributed with common mean $\boldsymbol{\beta}$ and variance-covariance matrix $\boldsymbol{\Sigma}$.

$$\boldsymbol{z}_i = (z_{i1}, z_{i2}, ..., z_{ip}) \sim N_p(\boldsymbol{\beta}, \boldsymbol{\Sigma}) \qquad\qquad i = 1, ..., n$$

We assume the existence of ordered cut points which, for each classifying variable, divide the real line into intervals corresponding to the ordered categories. Let $\theta(j, c)$ denote the cth cut point for the jth variable. Then y_{ij} takes the cth level if z_{ij} falls between the lower and upper cut points for the cth level:

$$y_{ij} = c \quad \text{if} \quad \theta(j, c-1) \leq z_{ij} < \theta(j, c) \tag{3.1}$$

The first and last cut points in each dimension are set to $-\infty$ and ∞ respectively. To ensure identifiability, the following additional constraints are imposed:

$$\theta(j, 1) = 0, \qquad\qquad \theta(j, 2) = 1.$$

Two constraints need to be imposed for each dimension, so that the scale of the latent data is identified. This is equivalent to fixing the single cut point and forcing $\boldsymbol{\Sigma}$ to be in correlation form in the binary case. This is an arbitrary choice of constraints and alternatives will be investigated later. This is an analogous model to that described in Chen and Dey (2000).

The model is determined by the parameters $\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{z}_i, \boldsymbol{\theta}$. By Bayes theorem, their joint posterior distribution is given by:

$$f(\mathrm{z}, \boldsymbol{\beta}, \Sigma, \boldsymbol{\theta}|\mathrm{y}) \;\propto\; f(\mathrm{y}|\mathrm{z}, \boldsymbol{\beta}, \Sigma, \boldsymbol{\theta}) f(\mathrm{z}, \boldsymbol{\beta}, \Sigma, \boldsymbol{\theta})$$

Since $\boldsymbol{y}$ is fully determined by $\boldsymbol{z}$ and $\boldsymbol{\theta}$ this reduces to:

$$f(\mathrm{z}, \boldsymbol{\beta}, \Sigma, \boldsymbol{\theta}|\mathrm{y}) \;\propto\; f(\mathrm{y}|\mathrm{z}, \boldsymbol{\theta}) f(\mathrm{z}, \boldsymbol{\beta}, \Sigma, \boldsymbol{\theta})$$

Then decomposing $f(\mathbf{z}, \boldsymbol{\beta}, \Sigma, \boldsymbol{\theta})$ and choosing independent priors for $\boldsymbol{\beta}$, Σ and $\boldsymbol{\theta}$, this becomes:

$$f(\mathbf{z}, \boldsymbol{\beta}, \Sigma, \boldsymbol{\theta}|\mathbf{y}) \quad \propto \quad f(\mathbf{y}|\mathbf{z}, \boldsymbol{\theta}) f(\mathbf{z}|\boldsymbol{\beta}, \Sigma) f(\boldsymbol{\beta}) f(\Sigma) f(\boldsymbol{\theta}) \qquad (3.2)$$

where $f(\mathbf{y}|\mathbf{z}, \boldsymbol{\theta})$ is deterministic as described in (3.1), $f(\mathbf{z}|\boldsymbol{\beta}, \Sigma)$ is the normal density $N_p(\boldsymbol{\beta}, \Sigma)$, and $f(\boldsymbol{\beta})$, $f(\Sigma)$ and $f(\boldsymbol{\theta})$ are the prior densities for $\boldsymbol{\beta}$, Σ and $\boldsymbol{\theta}$ respectively. We choose these priors as follows:

- θ(j,m) $\sim$ uniform
- $\boldsymbol{\beta} \sim N_p(0, \boldsymbol{T})$
- $\Sigma \sim$ Inverse-Wishart $(q, \boldsymbol{A})$ with probability density function

$$f(\Sigma) \propto |\Sigma|^{-\frac{q+p+1}{2}} \exp\left[-\frac{tr(\Sigma^{-1}\boldsymbol{A})}{2}\right] \qquad (3.3)$$

These priors were chosen in order to yield posterior distributions that are standard and thus easily simulated from. In our examples we choose the parameters of these priors to give as noninformative priors as possible. The matrix $\boldsymbol{T}$ is a diagonal matrix $\boldsymbol{T} = \text{diag}(\tau_{11}^2, \tau_{22}^2, \ldots, \tau_{pp}^2)$ where τ_{ii}^2 is large. A necessary condition for the inverse-Wishart distribution to be proper is that $q > p$; however the smaller the value of q, the less informative the prior; q is therefore set to be $p + 1$. It is less clear what to set $\boldsymbol{A}$ to be, for now it is set to be the identity matrix $\boldsymbol{I}_p$; sensitivity to the choice of prior parameters will be investigated later.

Substituting these priors into Equation (3.2) yields the posterior distribution of the unknown parameters:

$$
\begin{aligned}
f(\mathbf{z},\boldsymbol{\beta},\Sigma,\theta|\mathbf{y}) \propto & \prod_{i=1}^{n}\prod_{j=1}^{p}\left[\sum_{m=1}^{k_j} I(\theta(j,m-1)\leq z_{ij}\leq\theta(j,m))I(y_{ij}=m)\right] \\
& \times \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}\exp\left[-\frac{1}{2}(\boldsymbol{z}_i-\boldsymbol{\beta})^T\Sigma^{-1}(\boldsymbol{z}_i-\boldsymbol{\beta})\right] \\
& \times \frac{1}{(2\pi)^{p/2}|\boldsymbol{T}|^{1/2}}\exp\left[-\frac{1}{2}\boldsymbol{\beta}^T\boldsymbol{T}^{-1}\boldsymbol{\beta}\right] \\
& \times |\boldsymbol{\Sigma}|^{-(q+p+1)/2}\exp\left[-\frac{\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{A})}{2}\right] \qquad (3.4)
\end{aligned}
$$

We wish to estimate properties, such as the mean and variance, and marginal distributions of these unknown parameters, but this involves integrating the Equation (3.4) above. To overcome this problem, we use a Monte Carlo Markov Chain (MCMC) to generate random samples from this distribution.

To sample this posterior density, the following Gibbs sampler method was used. This method was chosen due to the standard form of the conditional distributions of the unknown parameters.

3.2 Algorithm 3

1. Starting with initial values for all parameters, sample the mean $\boldsymbol{\beta}$ of the latent data from its conditional distribution

$$\boldsymbol{\beta}|\boldsymbol{z},\Sigma \sim N_p\left((n\Sigma^{-1}+T^{-1})^{-1}\Sigma^{-1}\sum_{i=1}^{n}\boldsymbol{z}_i\ ,\ \ (n\Sigma^{-1}+T^{-1})^{-1}\right)$$

2. Sample the variance $\boldsymbol{\Sigma}$ of the latent data from its conditional distribution

$$\boldsymbol{\Sigma}|\boldsymbol{\beta},\boldsymbol{z} \sim \text{Inverse-Wishart}\left(\boldsymbol{A}+nS_{\boldsymbol{\beta}},q+n\right)$$

where $S_{\boldsymbol{\beta}}=\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{z}_i-\boldsymbol{\beta})(\boldsymbol{z}_i-\boldsymbol{\beta})^T$ and the inverse-Wishart distribution is parameterised as in (3.3).

3. Sample the new latent data from their conditional distributions
$$\boldsymbol{z}_i|\boldsymbol{y}_i, \boldsymbol{\beta}, \Sigma, \boldsymbol{\theta} \sim N_p(\boldsymbol{\beta}, \Sigma)$$
with z_{ij} truncated to the interval $(\theta(j, y_{ij}-1), \theta(j, y_{ij}))$.
4. Sample the new cut points from their conditional distribution:
$$\theta(j,m)|\boldsymbol{z} \sim \text{Unif}\left(\max_i\{z_{ij} : y_{ij} = m\}, \min_i\{z_{ij} : y_{ij} = m+1\}\right)$$
5. Go back to step 1 and repeat.

If a large enough sample is generated from the Markov Chain, this is a dependent sample from approximately the joint posterior distribution, and hence samples from the marginal distributions of each of the parameters (and any functions of interest) can be easily evaluated. All the conditional distributions are standard apart from those used to generate the latent data $\boldsymbol{z}_i$ which are multivariate truncated normal distributions. Sampling from this distribution is carried out via a sequence of univariate truncated normals using the method developed by Geweke (1991). This consists of a cycle of p Gibbs steps through the components of $\boldsymbol{z}_i$, which have truncated univariate normal distributions. These are generated using the inverse distribution function method (Devroye, 1986).

3.3 Example 1: Oesophageal Cancer Dataset

The scheme outlined in Algorithm 3 was applied to the two-way contingency table displayed in Table 3.1 (Breslow, 1982), which shows the results from a Case-Control study investigating the relationship between drinking beverages at burning hot temperatures and incidence of oesophageal cancer.

The Gibbs sampler was implemented using 30,000 iterations. The posterior means along with the posterior standard deviations were obtained for $\boldsymbol{\beta}$, Σ and $\boldsymbol{\theta}$.
$$E(\boldsymbol{\beta}|\boldsymbol{y}) = \begin{pmatrix} -0.132(0.42) \\ -2.448(0.73) \end{pmatrix}$$

	Control			
Case	0	1	2	3
0	31	5	5	0
1	12	1	0	0
2	14	1	2	1
3	6	1	1	0

Table 3.1: Number of beverages drunk at burning hot temperatures for oesophageal cancer case-control pairs

$$E(\Sigma|\boldsymbol{y}) = \begin{pmatrix} 7.936(0.564) & -0.025(0.180) \\ -0.025(0.180) & 10.403(1.091) \end{pmatrix}$$

$$E(\theta(\text{case,3})|\boldsymbol{y}) = 3.315(0.670)$$

$$E(\theta(\text{control,3})|\boldsymbol{y}) = 4.571(0.421)$$

Table 3.2 shows the mean posterior predictive table (the expected data) estimated by the model. This was found by taking a sample of size 80 from the normal distribution with mean and variance generated at each iteration of the Gibbs sampler. These latent data were then categorised using the cut points generated at the same iteration of the Gibbs sampler. The mean over all 30,000 tables was then taken.

	Control			
Case	0	1	2	3
0	32.48	3.62	4.60	0.59
1	9.49	1.06	1.33	0.16
2	14.41	1.61	2.07	0.26
3	6.45	0.74	0.99	0.14

Table 3.2: Mean posterior predictive data for oesophageal cancer case-control pairs

3.3.1 Discussion

- A comparison of the observed and expected data in Tables 3.1 and 3.2 indicates that the model is performing well, as observed and expected cell counts are very close.
- The mean in both dimensions is less than zero, which corresponds to the cell Case=0, Control=0, i.e. the model indicates that the data are concentrated in this area. This agrees with the observed data.
- The correlation between Case and Control is -0.006. The model indicates that there is little correlation between the Case-Control pairs. Again, this agrees with the observed data, as there is no obvious strong dependence structure.
- The estimated third cut point for the Controls (columns) is higher than that for the Cases (rows). This implies that there are fewer people to have drunk more drinks at burning hot temperatures in the Controls. Again this agrees with the observed data.

The mean posterior predictive density of the latent variables z_i is shown in Figure 3.1, complete with the mean posterior cut points. This clearly illustrates the lack of dependence structure.

3.4 Model Diagnostics

3.4.1 Convergence

In order to check the convergence of the Gibbs sampler, trace plots were plotted for each of the parameters estimated. Figures 3.2 and 3.3 show the trace plots for β_1 and Σ_{11} respectively.

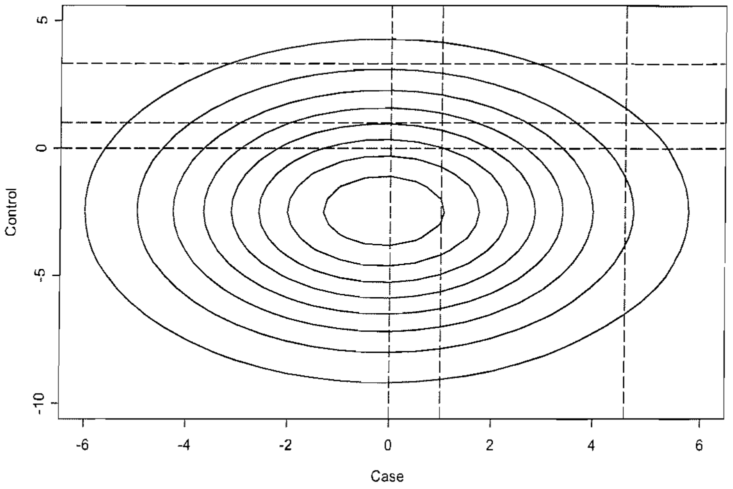


Figure 3.1: Mean posterior distribution of the latent variable with posterior means for cut points overlaid.

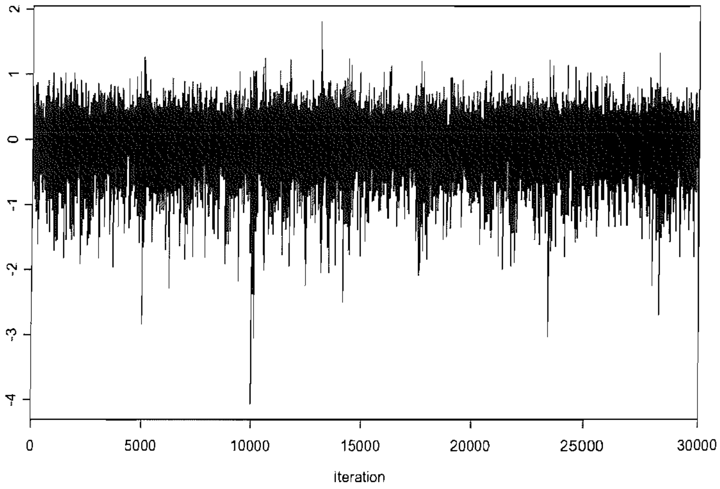


Figure 3.2: Trace plot for β_1 for oesophageal cancer data

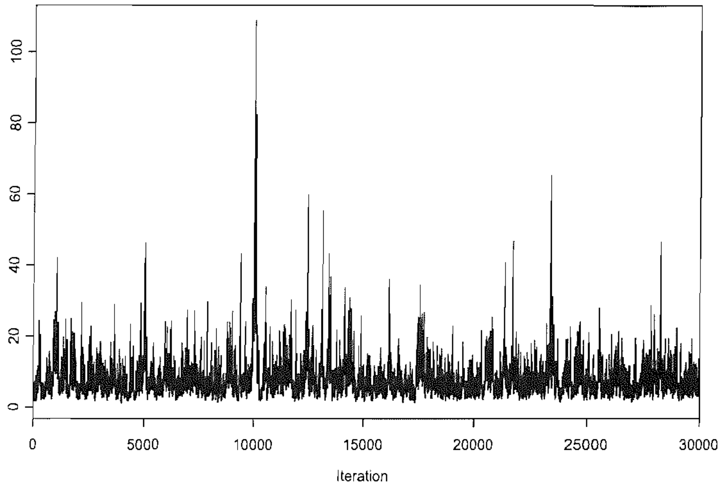


Figure 3.3: Trace plot for Σ_{11} for oesophageal cancer data

These are typical of the trace plots for each estimated parameter. They show that the Gibbs sampler is mixing well, with negligible burn-in period.

As discussed in Chapter 2, it has been noted that convergence of the Gibbs sampler is sometimes slow when there are free cut points to be estimated, especially if the sample size is large. This is due to the way that the free cut points are generated. Their conditional distribution is uniform on the space between $\max_i\{z_{ij} : y_{ij} = m\}$ and $\min_i\{z_{ij} : y_{ij} = m + 1\}$. Clearly if there are many individuals in each category, this space will be small. As a consequence, the cut point values can change very little between successive iterations. This can also affect the convergence of the other parameters, and hence the convergence of the Gibbs sampler. Figure 3.4 shows the trace plot for the free cut point for Case.

Note that although this is not moving as freely as the other parameters, it still appears to be converging satisfactorily. This is probably due to the small sample size of 80, so that there are relatively few individuals in each category. We shall see some examples in Chapter 4 where the convergence of the free

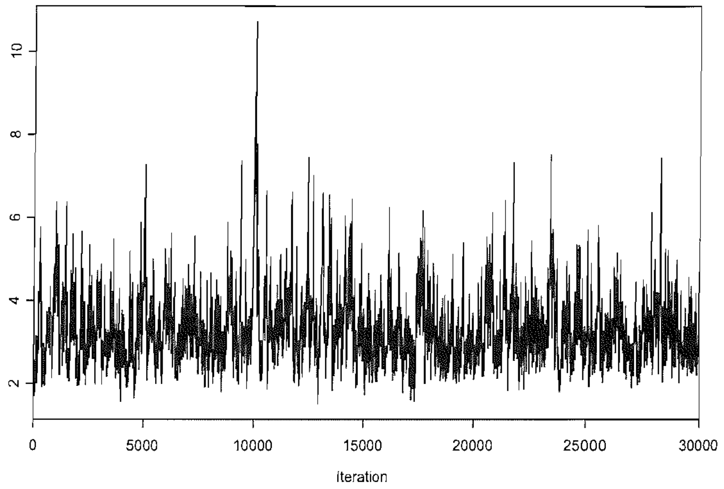


Figure 3.4: Trace plot for $\theta(Case, 3)$ for oesophageal cancer data

cut point is more questionable. As discussed in Chapter 2, Cowles (1996) and Nandram and Chen (1996) have suggested methods of speeding up convergence of cut points.

The convergence of the data augmentation Gibbs sampler algorithm could also be improved by employing the method of parameter expansion described by Liu (2001). Parameter expansion works by introducing extra parameters without distorting the original observed data model. Liu and Wu (1999) identify conditions under which a parameter expansion algorithm can be guaranteed to outperform a standard data augmentation algorithm. In the discussion of van Dyk and Meng (2001), Liu provides a particular example of the parameter expansion method for the multivariate probit model, while Imai and van Dyk (2005) use the method for the multinomial probit.

3.4.2 Goodness-of-fit

In order to assess the quality of the model, we need some measure of its goodness-of-fit. We use a simulation-based method proposed by Dey et al. (1998), which assesses the probability of the observed data being predicted by the model. To do this, 30,000 tables were generated, one for each iteration, using the mean, variance and cut points produced at that iteration. We therefore have the observed table cell counts:

$$\{y_i^{obs}; i = 1, ..., 16\}$$

and 30,000 tables generated during the MCMC:

$$\{y_i^{pred,j}; i = 1, ..., 16; j = 1, ..., 30,000\}$$

If the model provides a good fit to the data, we would expect the original table to be typical of tables generated by the model. In order to assess this, three distance measures are used to measure the 'distance' between each of the 30,000 tables and the posterior predictive mean table, and also the distance between the initial data and the posterior predictive mean table. If the model fits poorly, we would expect the latter distance to lie in the upper tail of the distribution of distances. The distance measures used are the Pearson's distance, Deviance distance, and the Maximum Absolute Difference distance:

- Pearson's distance:

$$\sum_{i=1}^{16} \frac{(y_i^{pred,j} - \bar{y}_i^{pred})^2}{\bar{y}_i^{pred}}$$

- Deviance distance:

$$2\sum_{i=1}^{16} y_i^{pred,j} \log \frac{y_i^{pred,j}}{\bar{y}_i^{pred}}$$

- Maximum Absolute Difference distance:

$$\max_i \left\{ \left| y_i^{pred,j} - \bar{y}_i^{pred} \right| \right\}$$

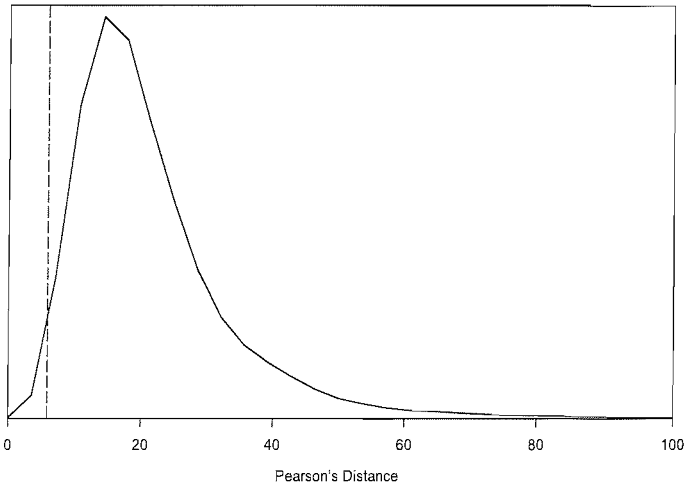


Figure 3.5: Estimated density of Pearson's distance measure for oesophageal cancer data

Figures 3.5, 3.6 and 3.7 show the densities of the Pearson's, Deviance and Maximum Absolute Difference distance measures respectively. Note that cell counts of less than 5 were pooled. The vertical line represents the distance between the observed data $\{y_i^{obs}\}$ and the posterior predictive mean table. The fact that for each distance measure, the vertical line is well into the lower tail of the density shows that the model fits very well.

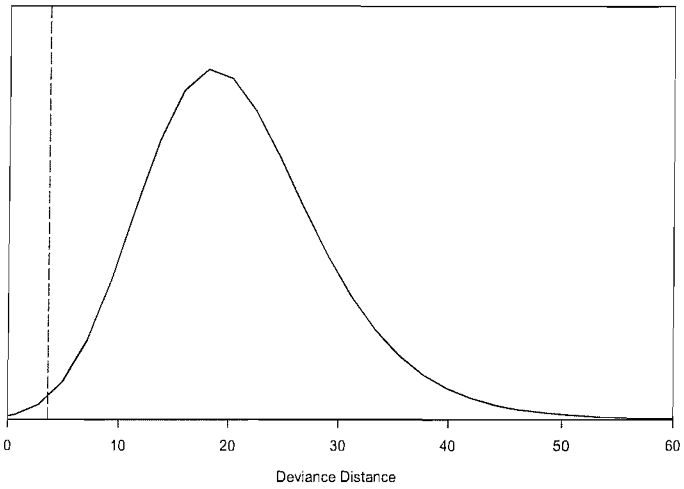


Figure 3.6: Estimated density of deviance distance measure for oesophageal cancer data

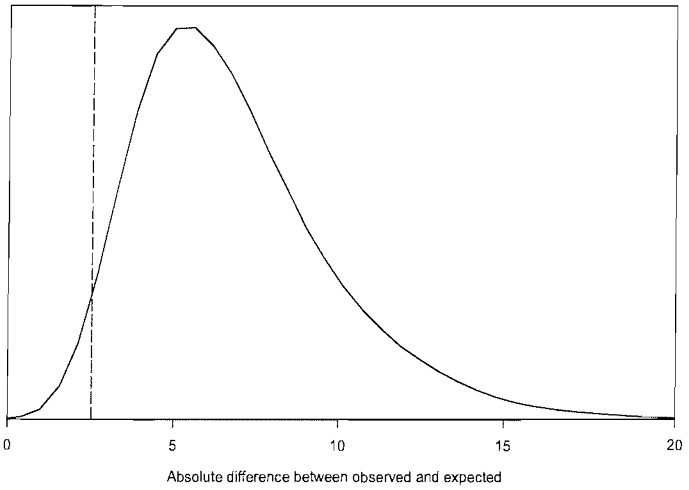


Figure 3.7: Estimated density of maximum absolute difference distance measure for oesophageal cancer data

3.5 Example 2: Blackbird Dataset

The scheme outlined in Algorithm 3 was applied to the three-way contingency table displayed in Table 3.3, taken from Anderson and Pemberton (1985), which shows 90 'first-year' blackbirds cross-classified on three aspects of their colour. The colours of the lower mandible (LM), the upper mandible (UM) and the orbital ring (OR) were recorded as ordered categorical variables, ranging from all black (1) to all yellow (3). For each variable, there are three ordered categories.

Lower	Upper	Orbital Ring		
Mandible	Mandible	1	2	3
1	1	40	19	0
	2	0	0	0
	3	0	1	0
2	1	1	6	0
	2	1	2	1
	3	0	1	0
3	1	1	2	0
	2	0	1	1
	3	0	6	7

Table 3.3: Ninety blackbirds classified by colour of upper mandible, lower mandible and orbital ring.

The Gibbs sampler was implemented using 30,000 iterations. The posterior means along with their standard deviations were obtained for $\boldsymbol{\beta}$ and Σ. Note that since there are only three categories for each variable, there are no free cut points to be estimated. The data were ordered so that Variable 1 = LM, Variable 2 = UM and Variable 3 = OR.

$$E(\boldsymbol{\beta}|\boldsymbol{y}) = \begin{pmatrix} -3.79(0.01) \\ -13.01(0.02) \\ 0.062(0.001) \end{pmatrix}$$

$$E(\Sigma|\boldsymbol{y}) = \begin{pmatrix} 84.20(2.62) & 95.49(3.35) & 4.21(0.81) \\ 95.49(3.35) & 434.58(6.08) & 9.91(1.02) \\ 4.21(0.81) & 9.91(1.02) & 0.68(0.04) \end{pmatrix}$$

As we would expect, the posterior means for LM and UM are well below zero, corresponding to category 1 (all black). This is due to the fact that most of the observed birds fall into this category. However there are birds that fall into category 3 (all yellow) for the variables LM and UM. The posterior variance is high to allow for this despite the mean being (relatively) much smaller than 0. The posterior mean for OR falls on the borderline between categories 1 and 2, thus reflecting the fact that birds are more evenly spread over categories for this variable.

Table 3.4 shows the mean posterior predictive table (the expected data) estimated by the model, calculated using the method described in Section 3.3.

Lower	Upper	Orbital Ring		
Mandible	Mandible	1	2	3
1	1	36.16	17.05	1.37
	2	0.36	0.61	0.10
	3	1.50	3.57	1.04
2	1	1.07	1.78	0.30
	2	0.05	0.14	0.05
	3	0.16	0.76	0.32
3	1	2.64	6.59	2.01
	2	0.11	0.56	0.30
	3	0.64	5.29	5.46

Table 3.4: Mean posterior predictive data for blackbird colouring data

The model appears to provide a good fit to the data. We use the simulation goodness-of-fit method described in 3.4.2 to check this. The chi-squared and deviance statistics are not entirely satisfactory as many of the posterior mean cell counts are small and thus have a high influence on the goodness of fit statistics measures. The absolute distance measure is unaffected by this and the density is shown in Figure 3.8, with the vertical line representing the distance between the observed data and the posterior predictive mean table.

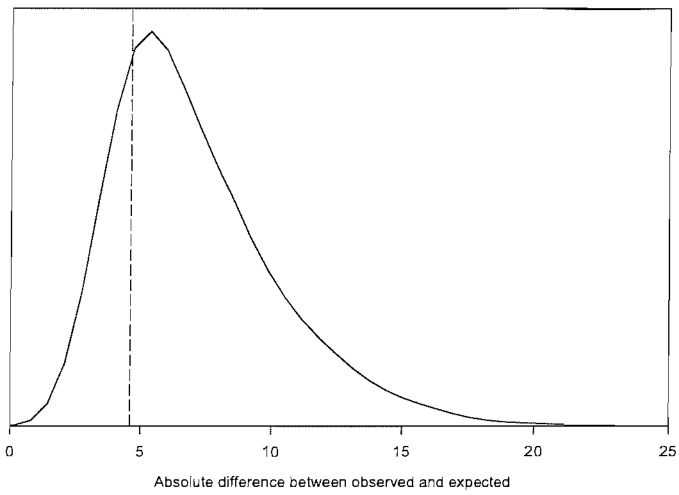


Figure 3.8: Estimated density of maximum absolute difference distance measure for the blackbird data

This agrees with observation of Tables 3.3 and 3.4 that the model appears to fit well. Trace plots of the parameters indicated that convergence was satisfactory.

3.6 Binary Data

We require a general method for both ordinal and binary data (or a mixture of both). If a classifying variable is binary, there is only one finite cut point to be constrained, leaving a remaining constraint to be imposed to ensure identifiability. For binary data, Chib and Greenberg (1998) constrain the marginal variances σ^2_{ii}. This requires specifying a prior distribution and simulating from the posterior distribution of a restricted-covariance matrix. The prior normalising constant for such a distribution over covariance matrices restricted in this way is not generally available, which creates difficulties for model determination as this constant is explicitly required in the marginal likelihood (1.2). We introduce an alternative parameterisation of the model that not only aids computation for the single model case but also provides a very neat framework for model determination: the inverse covariance matrix is parameterised in terms of its Cholesky decomposition

$$\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Phi}^T \boldsymbol{\Phi} \tag{3.5}$$

where $\boldsymbol{\Phi}$ is an upper triangular matrix. This parameterisation is motivated by the following decomposition of the joint likelihood for the latent variables $f(\boldsymbol{z}_i) = f(z_{i1}, z_{i2}, ..., z_{ip})$:

$$f(\boldsymbol{z}_i) = f(z_{ip}) f(z_{i,p-1}|z_{ip}) f(z_{i,p-2}|z_{i,p-1}, z_{ip}) \dots f(z_{i1}|z_{i2}, z_{i3}, ..., z_{ip})$$

This can be expressed as the following recursive set of equations.

$$\begin{aligned} z_{ip} &= N\left(\beta_p, \frac{1}{\phi_{pp}^2}\right) \\ z_{i,p-1}|z_{ip} &= \beta_{p-1} - \frac{\phi_{p-1,p}}{\phi_{p-1,p-1}} z_{ip} + N\left(0, \frac{1}{\phi_{p-1,p-1}^2}\right) \\ z_{i,p-2}|z_{i,p-1}, z_{ip} &= \beta_{p-2} - \frac{\phi_{p-2,p}}{\phi_{p-2,p-2}} z_{ip} - \frac{\phi_{p-2,p-1}}{\phi_{p-2,p-2}} z_{i,p-1} + N\left(0, \frac{1}{\phi_{p-2,p-2}^2}\right) \\ \vdots \quad & \quad \vdots \qquad \vdots \\ z_{i1}|z_{i2}, \dots, z_{ip} &= \beta_1 - \frac{\phi_{1p}}{\phi_{11}} z_{ip} - \frac{\phi_{1,p-1}}{\phi_{11}} z_{i,p-1} - \dots - \frac{\phi_{12}}{\phi_{11}} z_{i2} + N\left(0, \frac{1}{\phi_{11}^2}\right) \end{aligned} \tag{3.6}$$

In matrix form, this is equivalent to

$$\begin{pmatrix} z_{i1} \\ z_{i2} \\ \vdots \\ \vdots \\ z_{ip} \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} 0 & -\frac{\phi_{12}}{\phi_{11}} & -\frac{\phi_{13}}{\phi_{11}} & \dots & -\frac{\phi_{1p}}{\phi_{11}} \\ & 0 & -\frac{\phi_{23}}{\phi_{22}} & \dots & -\frac{\phi_{2p}}{\phi_{22}} \\ & & \ddots & & \vdots \\ & & & 0 & -\frac{\phi_{p-1,p}}{\phi_{p-1,p-1}} \\ & & & & 0 \end{pmatrix} \begin{pmatrix} z_{i1} \\ z_{i2} \\ \vdots \\ \vdots \\ z_{ip} \end{pmatrix}$$

$$+ \quad N_p\left(\mathbf{0}, \operatorname{diag}\left(\frac{1}{\phi_{ii}^2}\right)\right)$$

Rearranging and taking variances (and using the facts that $\text{Var}(\boldsymbol{A}^T\boldsymbol{X}) = \boldsymbol{A}^T\text{Var}(\boldsymbol{X})\boldsymbol{A}$ and $\text{Var}(\boldsymbol{z}_i) = \boldsymbol{\Sigma}$) we arrive at the following equation.

$$\boldsymbol{\Sigma} = \boldsymbol{U}^{-1} \text{diag}\left(\frac{1}{\sigma_{ii}^2}\right) \boldsymbol{U}^{-T}$$

where $\boldsymbol{U}$ is the upper-triangular matrix:

$$\boldsymbol{U} = \begin{pmatrix} 1 & \frac{\phi_{12}}{\phi_{11}} & \frac{\phi_{13}}{\phi_{11}} & \cdots & -\frac{\phi_{1p}}{\phi_{11}} \\ & 1 & \frac{\phi_{23}}{\phi_{22}} & \cdots & \frac{\phi_{2p}}{\phi_{22}} \\ & & \ddots & & \vdots \\ & & & 1 & \frac{\phi_{p-1,p}}{\phi_{p-1,p-1}} \\ & & & & 1 \end{pmatrix}$$

Taking inverses, this can be expressed as

$$\boldsymbol{\Sigma}^{-1} = \boldsymbol{U}^T \boldsymbol{D}^T \boldsymbol{D} \boldsymbol{U}$$

where $\boldsymbol{D}$ is the diagonal matrix diag(ϕ_{ii}). Therefore we finally arrive at the Cholesky decomposition parameterisation of the inverse variance matrix expressed in (3.5), with $\boldsymbol{\Phi} = \boldsymbol{D}\boldsymbol{U}$. From (3.5), we see that ϕ_{ii} can be interpreted as the conditional precision of the latent data for variable i given the latent data for all variables preceding i in the decomposition. The off-diagonal elements ϕ_{ij} can be interpreted as scaled regression coefficients.

The model is now determined by the parameters $\boldsymbol{\beta}, \Phi, \boldsymbol{z}, \boldsymbol{\theta}$. By Bayes theorem, their joint posterior distribution is given by:

$$f(\mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\Phi}, \boldsymbol{\theta}|\mathbf{y}) \quad \propto \quad f(\mathbf{y}|\mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\Phi}, \boldsymbol{\theta}) f(\mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\Phi}, \boldsymbol{\theta})$$

Since $\boldsymbol{y}$ is fully determined by $\boldsymbol{z}$ and $\boldsymbol{\theta}$ this reduces to:

$$f(\mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\Phi}, \boldsymbol{\theta}|\mathbf{y}) \quad \propto \quad f(\mathbf{y}|\mathbf{z}, \boldsymbol{\theta}) f(\mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\Phi}, \boldsymbol{\theta})$$

Then decomposing $f(\mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\Phi}, \boldsymbol{\theta})$ and choosing independent prior distributions for $\boldsymbol{\beta}$, $\boldsymbol{\Phi}$ and $\boldsymbol{\theta}$, this becomes:

$$f(\mathbf{z}, \boldsymbol{\beta}, \Sigma, \boldsymbol{\theta}|\mathbf{y}) \quad \propto \quad f(\mathbf{y}|\mathbf{z}, \boldsymbol{\theta}) f(\mathbf{z}|\boldsymbol{\beta}, \boldsymbol{\Phi}) f(\boldsymbol{\beta}) f(\boldsymbol{\Phi}) f(\boldsymbol{\theta}) \qquad (3.7)$$

where $f(\mathbf{y}|\mathbf{z}, \boldsymbol{\theta})$ is deterministic as described in (3.1), $f(\mathbf{z}|\boldsymbol{\beta}, \boldsymbol{\Phi})$ is the density $N_p(\boldsymbol{\beta}, (\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1})$, and $f(\boldsymbol{\beta})$, $f(\boldsymbol{\Phi})$, and $f(\boldsymbol{\theta})$ are the independent prior densities for $\boldsymbol{\beta}$, $\boldsymbol{\Phi}$ and $\boldsymbol{\theta}$ respectively.

The standard conjugate prior for covariance matrices is the inverse-Wishart distribution (3.3). This distribution has an inherent lack of flexibility for specifying prior information, as there are $\frac{p(p+1)}{2}$ hyperparameters (the elements of $\boldsymbol{A}$) to give a point estimate for $\boldsymbol{\Sigma}$ but only a single scalar parameter (q) with which to quantify uncertainty about this estimate. To overcome this, Brown et al. (1994) proposed the generalised inverse Wishart (GIW) distribution which provides an extremely flexible prior distribution for a covariance matrix. We follow a particular parameterisation of the GIW distributino proposed by Daniels and Pourahmadi (2002). They showed that if independent gamma prior distributions are placed on the diagonal elements ϕ_{ii} of $\boldsymbol{\Phi}$ and if, conditional on ϕ_{ii}, independent multivariate normal priors are placed on the partial rows $\boldsymbol{\phi}_i = (\phi_{i,i+1}, \ldots, \phi_{ip})$ of the upper triangle of the matrix $\boldsymbol{\Phi}$, then this prior is conditionally conjugate. For a certain choice of parameters, this distribution simplifies to the inverse-Wishart distribution, the usual conjugate prior for covariance matrices. Since we do not have strong prior beliefs about the covariance structure we use this prior. However, our approach is sufficiently flexible to allow prior information to be incorporated when available. Garthwaite and Al-Awadhi (2001) propose elicitation methods for quantifying expert opinion (where available) *via* both the generalised inverse Wishart and inverse Wishart distributions. Suppose that $\boldsymbol{\Sigma}$ is assumed to be, *a priori*, inverse-Wishart with parameters $\boldsymbol{A}$ and q. Then the equivalent prior distribution for $\boldsymbol{\Phi}$ is:

$$\phi_{ii} \sim \sqrt{b_i \chi^2_{q-i+1}} \tag{3.8}$$

$$\boldsymbol{\phi}_i | \phi_{ii} \sim N_{p-i}(\phi_{ii}\boldsymbol{\mu}_i, \boldsymbol{A}_i^{-1}) \tag{3.9}$$

where

$$\begin{aligned} b_i &= (a_{ii} - \boldsymbol{a}_i \boldsymbol{A}_i^{-1} \boldsymbol{a}_i^T)^{-1} \\ \boldsymbol{\mu}_i &= -\boldsymbol{a}_i \boldsymbol{A}_i^{-1} \end{aligned}$$

and $\boldsymbol{A}^{-1}$ is partitioned as follows:

$$\begin{array}{rcl} & \begin{array}{c} i^{th} \\ \text{column} \\ \downarrow \end{array} & \\ \boldsymbol{A}^{-1} = & \left(\begin{array}{c|c|c} \ddots & . & \cdots \\ \hline . & a_{ii} & \boldsymbol{a}_i \\ \hline \vdots & \boldsymbol{a}_i^T & \boldsymbol{A}_i \end{array}\right) & \leftarrow i^{th}\ \text{row} \end{array}$$

These priors are conditionally conjugate, therefore the posterior conditional distributions of $\boldsymbol{\phi}_i|\phi_{ii}$ and $\phi_{ii}|\boldsymbol{\phi}_i$ are multivariate normal and gamma respectively. This is very useful for posterior computation. Convenient priors for the other unknown parameters are $\theta(j,c) \sim 1$ and $\boldsymbol{\beta} \sim N_p(\boldsymbol{0}, \boldsymbol{T})$, a standard conjugate prior.

On substituting these priors into (3.7), we arrive at the posterior distribution which is again analytically intractable. We therefore use a Gibbs sampler approach to find a dependent sample from the posterior distribution of the unknown parameters. In order to do this, we need to find the conditional posterior distributions of $\boldsymbol{\Phi}, \boldsymbol{\beta}, \boldsymbol{z}, \boldsymbol{\theta}$.

The joint likelihood for the elements of $\boldsymbol{\Phi}$ is

$$\begin{aligned} f(\phi_{ii}, \boldsymbol{\phi}_i|\boldsymbol{z}) \quad \propto \quad \prod_{\text{rows i of } \Phi} \quad & \phi_{ii}^q \exp\left[-\frac{\phi_{ii}^2}{2}(g_{ii} - \boldsymbol{g}_i G_i^{-1}\boldsymbol{g}_i^T)\right] \\ & \exp\left[-\frac{\phi_{ii}^2}{2}(\boldsymbol{\phi}_i + \phi_{ii}\boldsymbol{g}_i G_i^{-1}) G_i (\boldsymbol{\phi}_i + \phi_{ii}\boldsymbol{g}_i G_i^{-1})^T\right] \end{aligned}$$

where $\boldsymbol{G} = \sum_{i=1}^{n}(\boldsymbol{z}_i - \boldsymbol{\beta})(\boldsymbol{z}_i - \boldsymbol{\beta})^T$ is partitioned as follows:

$$\begin{array}{rcl} & \begin{array}{c} i^{th} \\ \text{column} \\ \downarrow \end{array} & \\ \boldsymbol{G} = & \left(\begin{array}{c|c|c} \ddots & . & \cdots \\ \hline . & g_{ii} & \boldsymbol{g}_i \\ \hline \vdots & \boldsymbol{g}_i^T & G_i \end{array}\right) & \leftarrow i^{th}\ \text{row} \end{array}$$

The posterior distribution of $\mathbf{\Phi}|\boldsymbol{z},\boldsymbol{\beta}$ can now be found using Bayes' Theorem (1.1). Defining

$$\psi_{ij} = \frac{\phi_{ij}}{\phi_{ii}} \tag{3.10}$$

the conditional posterior distributions for $\boldsymbol{\psi}_i = (\psi_{i,i+1}, \ldots, \psi_{ip})$ and ϕ_{ii}^2 are as follows:

$$\boldsymbol{\psi}_i|\phi_{ii}^2, \boldsymbol{z}, \boldsymbol{\beta} \sim N_{p-i}\left((\boldsymbol{\mu}_i\boldsymbol{A}_i - \boldsymbol{g}_i)(\boldsymbol{A}_i + \boldsymbol{G}_i)^{-1}, \frac{1}{\phi_{ii}^2}(\boldsymbol{A}_i + \boldsymbol{G}_i)^{-1}\right) \tag{3.11}$$

$$\phi_{ii}^2|\boldsymbol{\psi}_i, \boldsymbol{z}, \boldsymbol{\beta} \sim \text{Gamma}\left(\gamma, \frac{\delta}{2}\right) \tag{3.12}$$

where

$$\gamma = \frac{q+n+p+1-i}{2} \tag{3.13}$$

$$\begin{aligned} \delta &= \frac{1}{b_i} + g_{ii} - \boldsymbol{g}_i\boldsymbol{G}_i^{-1}\boldsymbol{g}_i^T + (\boldsymbol{\psi}_i - \boldsymbol{\mu}_i)\boldsymbol{A}_i(\boldsymbol{\psi}_i - \boldsymbol{\mu}_i)^T \\ &+ (\boldsymbol{\psi}_i + \boldsymbol{g}_i\boldsymbol{G}_i^{-1})\boldsymbol{G}_i(\boldsymbol{\psi}_i + \boldsymbol{g}_i\boldsymbol{G}_i^{-1})^T \end{aligned} \tag{3.14}$$

The latent data $\boldsymbol{z}$ are sampled from

$$\boldsymbol{z}_i|\boldsymbol{y}_i, \boldsymbol{\beta}, \mathbf{\Phi}, \boldsymbol{\theta} \sim N_p\left(\boldsymbol{\beta}, (\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\right)$$

with z_{ij} truncated to the interval $(\theta(j, y_{ij}-1), \theta(j, y_{ij}))$. To sample from this distribution, we again use the method of Geweke (1991). The mean $\boldsymbol{\beta}$ of the latent data is sampled from the multivariate normal distribution:

$$\boldsymbol{\beta}|\boldsymbol{z}, \mathbf{\Phi} \sim N_p\left((n\mathbf{\Phi}^T\mathbf{\Phi} + \boldsymbol{T}^{-1})^{-1}\mathbf{\Phi}^T\mathbf{\Phi}\sum_{i=1}^{n}\boldsymbol{z}_i\ ,\ (n\mathbf{\Phi}^T\mathbf{\Phi} + \boldsymbol{T}^{-1})^{-1}\right)$$

Finally, the cut points are generated from their conditional distributions

$$\theta(j,c)|\boldsymbol{z} \sim \text{Unif}\left(\max_i\{z_{ij} : y_{ij} = c\}, \min_i\{z_{ij} : y_{ij} = c+1\}\right)$$

Starting with initial values for all parameters, the sampling scheme runs by sampling iteratively from the conditional posterior distributions in the order $[\boldsymbol{\beta}|\boldsymbol{z}, \mathbf{\Phi}]$, $[\phi_{11}^2|\boldsymbol{\psi}_1, \boldsymbol{z}, \boldsymbol{\beta}]$,..., $[\phi_{pp}^2|\boldsymbol{\psi}_p, \boldsymbol{z}, \boldsymbol{\beta}]$, $[\boldsymbol{\psi}_1|\phi_{11}^2, \boldsymbol{z}, \boldsymbol{\beta}]$,..., $[\boldsymbol{\psi}_p|\phi_{pp}^2, \boldsymbol{z}, \boldsymbol{\beta}]$,

$[\boldsymbol{z}_1|\boldsymbol{y}_1, \boldsymbol{\beta}, \boldsymbol{\Phi}, \boldsymbol{\theta}], \ldots, [\boldsymbol{z}_n|\boldsymbol{y}_n, \boldsymbol{\beta}, \boldsymbol{\Phi}, \boldsymbol{\theta}]$ and $[\boldsymbol{\theta}|\boldsymbol{z}]$. We refer to this as Algorithm 4.

Clearly, posterior simulation of the matrix $\boldsymbol{\Phi}$ may be carried out *via* independent draws from $\boldsymbol{\phi}_i|\phi_{ii}, \boldsymbol{z}$ and $\phi_{ii}|\boldsymbol{\phi}_i, \boldsymbol{z}$. The immediate advantage of this is that the conditional precision ϕ_{ii}^2 for any binary variable may be fixed to ensure identifiability. Such precisions are never updated by sampling, but the fact that they are fixed has no implications for any other conditional distribution. The resulting posterior distribution for $\boldsymbol{\Phi}$ and hence for the restricted covariance matrix $\boldsymbol{\Sigma}$ is now easily generated using independent draws from $\boldsymbol{\psi}_i|\phi_{ii}, \boldsymbol{z}$. This is due to the conditional independence structure provided by the Cholesky decomposition parameterisation. Purely ordinal data, purely binary data or a mixture of both can all be modelled with this approach. This gives a great advantage of this approach over others suggested and reviewed in Chapter 2. A further advantage is in the implementation of the model determination method described in Chapters 4 and 5.

3.7 Example and Results

In order to illustrate the method, we consider the 2 × 3 × 4 table from Knuiman and Speed (1988), shown in Table 3.5. It shows 491 subjects, classified according to Obesity (3 ordered levels), Hypertension (2 levels) and Alcohol Intake (4 ordered levels).

We order the table so that variable 1 is Obesity, variable 2 is Hypertension and variable 3 is Alcohol Intake. The conditional variance for the binary margin of the table, Hypertension was set to 1: $\phi_{22} = 1$. Algorithm 4 was applied to this data set, using 400,000 iterations. The following estimates for the posterior means along with their standard deviations were obtained for $\boldsymbol{\beta}$, $\boldsymbol{\Sigma} = (\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}$, and $\boldsymbol{\theta}$:

$$E(\boldsymbol{\beta}|\boldsymbol{y}) = \begin{pmatrix} 0.000(0.120) \\ 0.635(0.062) \\ 0.042(0.073) \end{pmatrix}$$

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

Table 3.5: Three-way table showing 491 subjects, classified by Hypertension, Alcohol Intake and Obesity (Knuiman and Speed, 1988)

$$E(\Sigma|\boldsymbol{y}) = \begin{pmatrix} 5.576(0.772) & -0.459(0.152) & 0.407(0.188) \\ -0.459(0.152) & 1.036(0.027) & -0.260(0.095) \\ 0.407(0.188) & -0.260(0.095) & 2.120(0.221) \end{pmatrix}$$

$$E(\theta(\text{Alcohol,3})|\boldsymbol{y}) = 2.034(0.060)$$

Note that $E(\Sigma|\boldsymbol{y})$ has been estimated based on the generated sample of $\boldsymbol{\Phi}$. Table 3.6 shows the mean posterior predictive data (the expected data as predicted by the model).

Obesity	Hypertension	Alcohol Intake (drinks/day)			
		0	1-2	3-5	5+
Low	Yes	6.74	7.46	9.37	9.37
	No	38.87	32.80	33.99	25.86
Average	Yes	7.52	9.17	12.30	13.56
	No	30.49	29.02	32.46	27.45
High	Yes	8.35	11.11	16.09	19.95
	No	24.18	25.34	30.57	28.96

Table 3.6: Mean posterior predictive cell counts for the Hypertension, Alcohol Intake and Obesity data (Knuiman and Speed, 1988)

3.7.1 Discussion

It is harder to conclude anything for the 3-way table because the structure of the data is not as simple to interpret as in the 2-way case. However, there are still some points to be noted from these results.

- The expected data appear to be very close to the observed data.
- The posterior mean for Hypertension (0.635) falls clearly into the 'No' category, indicating that people in this data set are less likely to have Hypertension. This agrees with inspection of the observed data. The posterior means for Alcohol Intake and Obesity correspond to the categories '3-5' and 'Average' respectively.
- If we consider the estimated posterior mean correlation matrix, calculated from $E(\boldsymbol{\Sigma}|\boldsymbol{y})$, we find that there is slight positive correlation between Alcohol Intake and Obesity (0.12), and slight negative correlation between Alcohol Intake and Hypertension (-0.18) and between Obesity and Hypertension (-0.19). Bearing in mind that Hypertension has been coded so that 'Yes' is the lower category and 'No' the higher category (as Hypertension is a binary variable, it can be ordered either way), these results agree with common sense. There is correspondence between higher alcohol intake and higher obesity, between higher alcohol intake and presence of hypertension, and between higher obesity and presence of hypertension.
- Generally, there are no extreme values in the model. This indicates that respondents are fairly well spread out over the whole table, with no strong concentration in any particular area. Again, this agrees with the observed data.

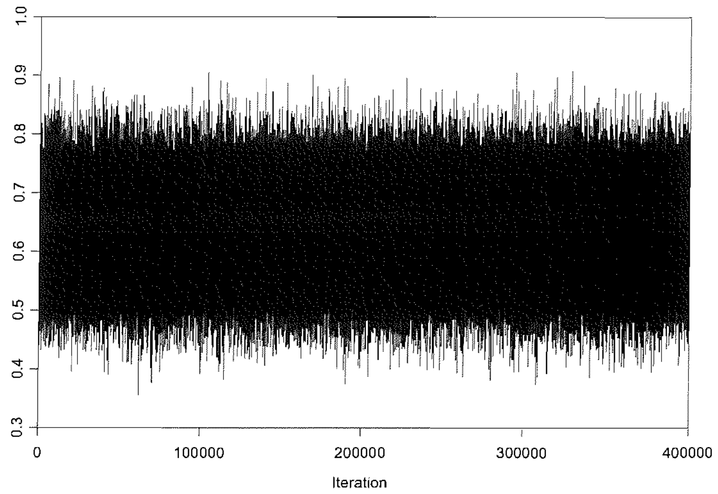


Figure 3.9: Trace plot for β_2 for hypertension, alcohol intake and obesity data

3.7.2 MCMC diagnostics

Convergence

In order to assess the convergence of the Gibbs sampler, trace plots were plotted for each of the parameters estimated. Figures 3.9 and 3.10 show the trace plots for $\boldsymbol{\beta}_2$ and $\boldsymbol{\Sigma}_{13}$ respectively.

These are typical of the trace plots for each estimated parameter. They show that the Gibbs sampler is mixing well, with negligible burn-in period. We can therefore conclude that the convergence of the Gibbs sampler is satisfactory.

Figure 3.11 shows the trace plot for the free cut point for Alcohol Intake. Note that although the Gibbs sampler appears to be traversing the parameter space, it is doing so slightly more slowly than for the mean and variance parameters. This is due to the fact that the categories either side of this free cut point contain approximately 100 individuals each, a fairly large number of

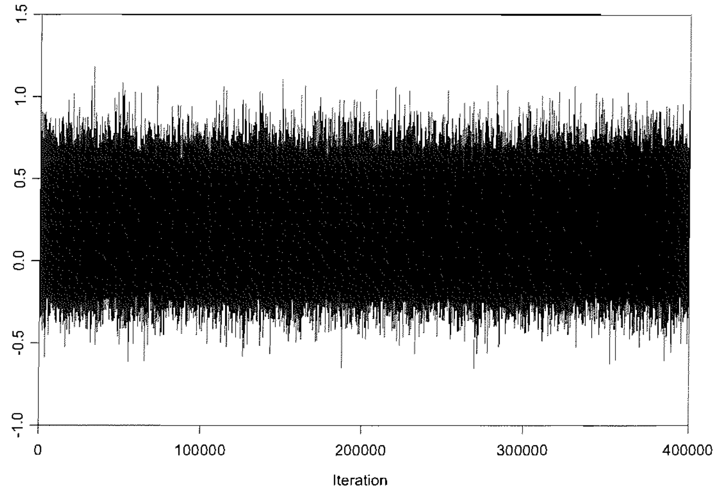


Figure 3.10: Trace plot for Σ_{13} for hypertension, alcohol intake and obesity data

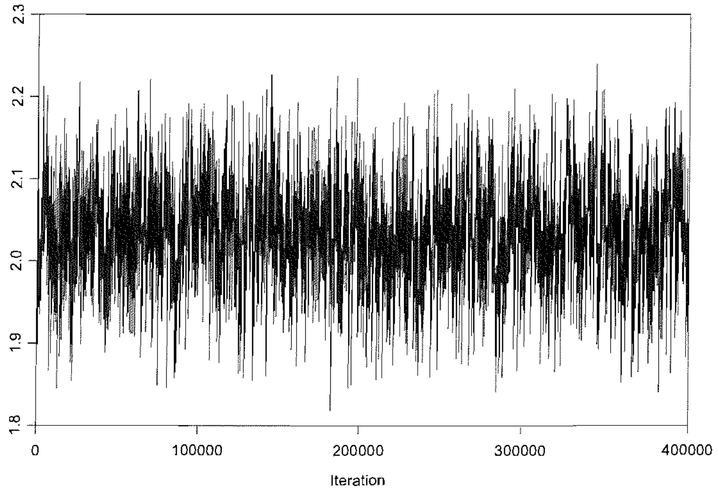


Figure 3.11: Trace plot for $\theta(Alcohol, 2)$ for hypertension, alcohol intake and obesity data

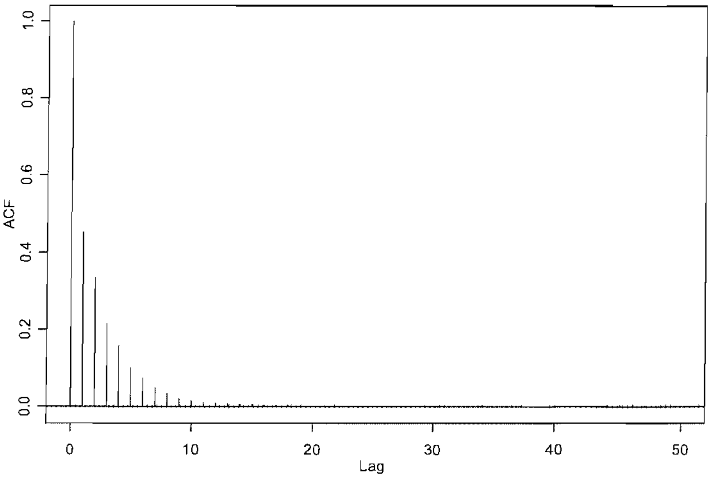


Figure 3.12: Autocorrelation function plot for β_2 for hypertension, alcohol intake and obesity data

individuals in each group, thus constraining the cut point to move slowly as discussed in Section 3.5.

This is emphasised by considering plots of the autocorrelation function. Figure 3.12 shows the autocorrelation function plot for β_2 while figure 3.13 shows the autocorrelation function plot for $\theta(Alcohol, 2)$.

Clearly the autocorrelation function plot for $\theta(Alcohol, 2)$ indicates slow convergence, but this does not seem to affect the convergence of β_2 or any of the other parameters.

Goodness-of-Fit

To assess the goodness-of-fit or otherwise of the model, the method outlined in Section 3.4.2 was applied. Figures 3.14, 3.15 and 3.16 show the

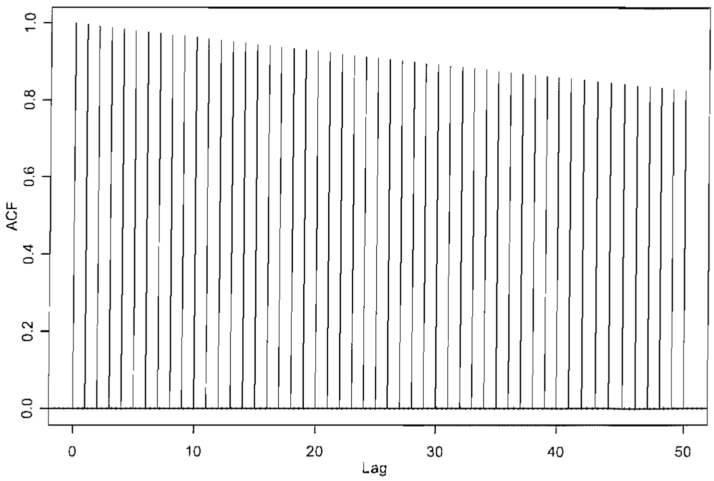


Figure 3.13: Autocorrelation function plot for $\theta(Alcohol, 2)$ for hypertension, alcohol intake and obesity data

densities of the Deviance, Pearson's, and Maximum Absolute Difference distance measures, with the vertical line representing the original data set. Again, the model appears to fit extremely well.

3.7.3 Fitting a single graphical model

So far, we have not considered the issue of model choice, and have restricted our attention to results gained from the saturated model. However, investigating which explanatory variables (and interactions between them) are significant and should therefore be included in a model for the data, is perhaps of even greater interest.

In order to demonstrate ideas that will be used in the next chapter when model determination is fully discussed, we show how all parameters for a single non-saturated graphical model may be estimated. Models are characterised by the structure of the inverse covariance matrix Σ^{-1} and hence by $\mathbf{\Phi}$. A zero

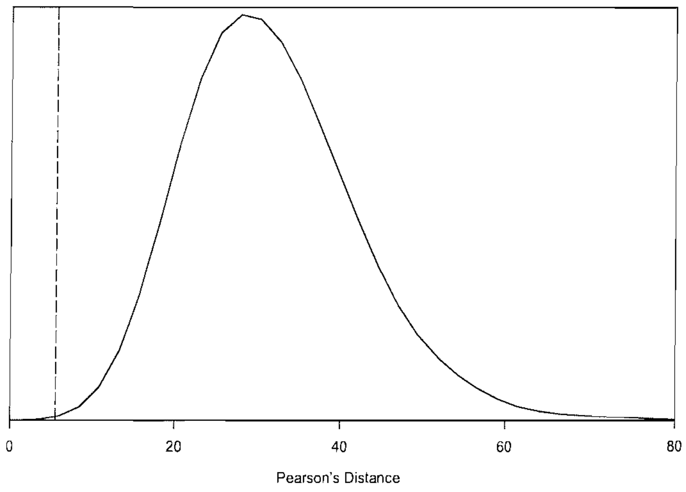


Figure 3.14: Estimated density of Pearson's distance measure for hypertension, alcohol intake and obesity data

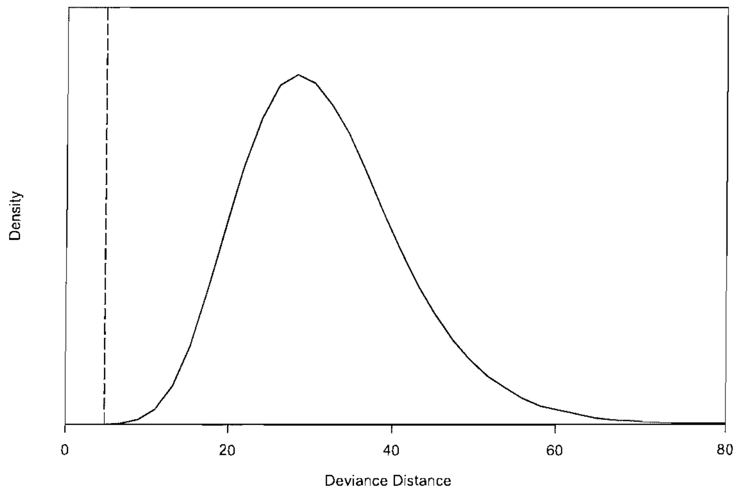


Figure 3.15: Estimated density of deviance distance measure for hypertension, alcohol intake and obesity data

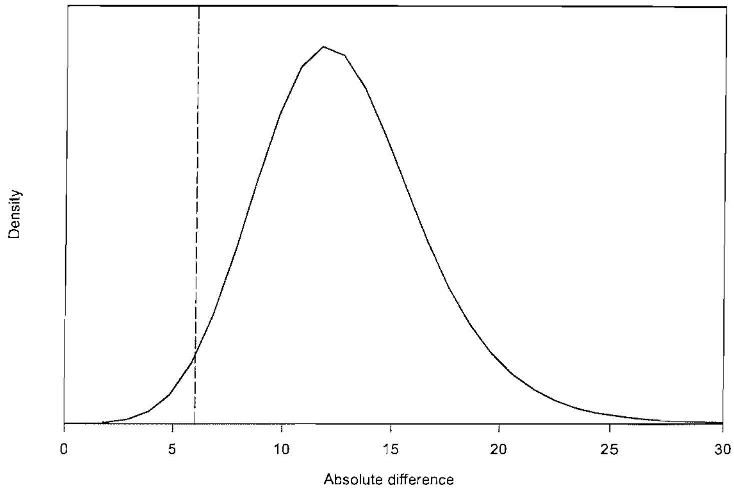


Figure 3.16: Estimated density of maximum absolute difference distance measure for hypertension, alcohol intake and obesity data

entry in Σ^{-1} and hence in $\mathbf{\Phi}$ corresponds to conditional independence between variables. The characterisation of a model by $\mathbf{\Phi}$ is actually somewhat more complex than stated here and will be described fully in Chapter 4, but the method of estimating a single model described below is unaltered by this (just the interpretation is).

Parameter estimation for a non-saturated graphical model is carried out using the Gibbs sampler, just as for the saturated model. The only thing that is changed is the off-diagonal structure of $\mathbf{\Phi}$. Therefore the conditional posterior distributions for $\boldsymbol{\beta}$, $\boldsymbol{z}$, ϕ_{ii} and $\boldsymbol{\theta}$ are unchanged. For $i = 1, \ldots, n$ the conditional posterior distribution for ϕ_i may be altered by the fact that some elements of ϕ_i are zero. There is no need to generate the zero elements of ϕ_i as they are zero, and the non-zero elements of ϕ_i can be found by conditioning on the zero elements. In order to do this, we use the standard result for conditional distributions of subsets of multivariate normally distributed random variables.

From (3.11), the conditional posterior distribution of $\boldsymbol{\psi}_i$ is

$$\boldsymbol{\psi}_i|\phi_{ii}^2, \boldsymbol{z}, \boldsymbol{\beta} \sim N_{p-i}(\boldsymbol{\lambda}, \boldsymbol{K})$$

where $\boldsymbol{\lambda} = (\boldsymbol{\mu}_i \boldsymbol{A}_i - \boldsymbol{g}_i)(\boldsymbol{A}_i + \boldsymbol{G}_i)^{-1}$ and $\boldsymbol{K} = \frac{1}{\phi_{ii}^2}(\boldsymbol{A}_i + \boldsymbol{G}_i)$

Re-ordering the vector $\boldsymbol{\psi}_i$ so that $\boldsymbol{\psi}_i^T = (\boldsymbol{\psi}_{i1}, \boldsymbol{\psi}_{i0})$ where $\boldsymbol{\psi}_{i1}$ represents non-zero elements of $\boldsymbol{\psi}_i$ to be generated and $\boldsymbol{\psi}_{i0}$ represents the zero elements of $\boldsymbol{\psi}_i$ to be conditioned on, the conditional distribution of $\boldsymbol{\psi}_{i1}$ is:

$$\boldsymbol{\psi}_{i1}|\phi_{ii}^2, \boldsymbol{z}, \boldsymbol{\beta} \sim N(\boldsymbol{\lambda}_1 - \boldsymbol{K}_{10}\boldsymbol{K}_{00}^{-1}\boldsymbol{\lambda}_0 \quad , \quad \boldsymbol{K}_{11} - \boldsymbol{K}_{10}\boldsymbol{K}_{00}^{-1}\boldsymbol{K}_{01}) \tag{3.15}$$

where $\boldsymbol{\lambda}$ and $\boldsymbol{K}$ are partitioned as follows:

$$\boldsymbol{\lambda}^T = (\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_0)$$

$$\boldsymbol{K} = \begin{pmatrix} \boldsymbol{K}_{11} & \boldsymbol{K}_{10} \\ \boldsymbol{K}_{01} & \boldsymbol{K}_{00} \end{pmatrix}$$

Parameter estimation for a non-saturated model is thus carried out using a Gibbs sampling procedure, which samples iteratively from the conditional posterior distributions in the order $[\boldsymbol{\beta}|\boldsymbol{z}, \boldsymbol{\Phi}]$, $[\phi_{11}^2|\boldsymbol{\psi}_1, \boldsymbol{z}, \boldsymbol{\beta}]$,..., $[\phi_{pp}^2|\boldsymbol{\psi}_p, \boldsymbol{z}, \boldsymbol{\beta}]$, $[\boldsymbol{\psi}_1|\phi_{11}^2, \boldsymbol{z}, \boldsymbol{\beta}]$,..., $[\boldsymbol{\psi}_p|\phi_{pp}^2, \boldsymbol{z}, \boldsymbol{\beta}]$, $[\boldsymbol{z}_1|\boldsymbol{y}_1, \boldsymbol{\beta}, \boldsymbol{\Phi}, \boldsymbol{\theta}]$, ..., $[\boldsymbol{z}_n|\boldsymbol{y}_n, \boldsymbol{\beta}, \boldsymbol{\Phi}, \boldsymbol{\theta}]$ and $[\boldsymbol{\theta}|\boldsymbol{z}]$.

We demonstrate this with an example using the data in Table 3.5. We fit the model $A + OH$; since Variable 1 = Obesity, Variable 2 = Hypertension and Variable 3 = Alcohol Intake, the constraints $\phi_{13} = \phi_{23} = 0$ are imposed on the matrix $\boldsymbol{\Phi}$. There is therefore only one off-diagonal element of $\boldsymbol{\Phi}$ to be estimated, that is ϕ_{12}. This is generated in the Gibbs sampler step for $\boldsymbol{\psi}_1$, conditioning on $\phi_{13} = \psi_{13} = 0$ as described above. Since $\phi_{23} = 0$, there is no generation step for $\boldsymbol{\psi}_2$.

The following estimates for the posterior means along with their standard deviations were obtained for $\boldsymbol{\beta}$, $\boldsymbol{\Sigma} = (\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}$, and $\boldsymbol{\theta}$:

$$E(\boldsymbol{\beta}|\boldsymbol{y}) = \begin{pmatrix} 0.050(0.060) \\ 0.627(0.061) \\ 1.119(0.089) \end{pmatrix}$$

$$E(\mathbf{\Sigma}|\boldsymbol{y}) = \begin{pmatrix} 1.397(0.194) & -0.220(0.075) & 0(0) \\ -0.220(0.075) & 1(0) & 0(0) \\ 0(0) & 0(0) & 2.458(0.385) \end{pmatrix}$$

$$E(\theta(\text{Alcohol,2})|\boldsymbol{y}) = -0.066(0.000)$$

Note that since Alcohol Intake is marginally independent of both Obesity and Hypertension, the corresponding elements of the covariance matrix $\mathbf{\Sigma}$ are 0. The resulting posterior predictive mean data, generated in the usual way, are shown in Table 3.7. For comparison, the predictive posterior means generated from the saturated model are displayed below in blue with the true values alongside in parentheses.

Obesity	Hypertension	Alcohol Intake (drinks/day) 0	1-2	3-5	5+
Low	Yes	7.9 (5) 6.7	7.7 (9) 7.5	9.1 (8) 9.4	8.5 (10) 9.4
	No	31.1 (40) 38.9	30.4 (36) 32.8	36.1 (33) 34.0	33.4 (24) 25.9
Average	Yes	10.1 (6) 7.5	9.9 (9) 9.2	11.8 (11) 12.3	10.9 (14) 13.6
	No	28.4 (33) 30.5	27.9 (23) 29.0	33.0 (35) 32.5	30.6 (30) 27.5
High	Yes	13.1 (9) 8.4	12.8 (12) 11.1	15.2 (19) 16.1	14.1 (19) 19.9
	No	25.9 (24) 24.2	25.3 (25) 25.3	30.1 (28) 30.6	27.9 (29) 28.9

Table 3.7: Mean posterior predictive cell counts generated using model $A{+}OH$ with true values in parentheses and saturated model estimates in blue

As would be expected, the model $A + OH$ does not fit the data as well as the saturated model. This can also be seen in a plot of the Pearson's statistic density curve for the model $A + OH$ (Figure 3.17) when compared with that of the saturated model in Figure 3.14.

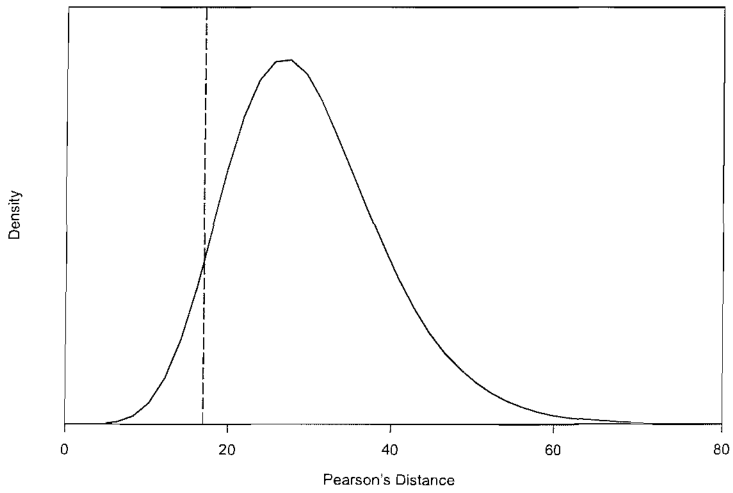


Figure 3.17: Estimated density of Pearson's distance measure for the $A + OH$ model

In this Chapter, we have developed a Gibbs sampler method for estimating the parameters in a full saturated model, and also shown how this method may be adapted to estimate parameters in a particular graphical model. In Chapter 4, we shall extend these ideas to investigate the issue of model choice.

Chapter 4

Model Determination for Decomposable Directed Graphical Models

4.1 Directed Acyclic Graphical Models

As we have seen, the Cholesky decomposition parameterisation $\mathbf{\Sigma}^{-1} = \mathbf{\Phi}^T\mathbf{\Phi}$ is equivalent to the recursive system of equations (3.5). This gives the elements of $\mathbf{\Phi}$ an interpretation as parameters of the conditional distributions involved in this recursive factorisation, that is, the squared diagonal elements of $\mathbf{\Phi}$ are the conditional precisions and the off-diagonal elements are scaled conditional regression coefficients. The recursion also means that the variables take an 'ordering' which depends on the order in which they appear in the factorisation, or equivalently, the order in which they are taken into the covariance matrix $\mathbf{\Sigma}$. This means that conditional independence between variables may only be characterised given the variables that have been conditioned upon already and not on those that have not yet appeared in the ordering. Thus, any given $\mathbf{\Phi}$ corresponds to a particular decomposable *directed* graphical model.

More formally, if $\phi_{ij} = 0$, then variables i and j are conditionally independent given the variables preceding either of them in the ordering.

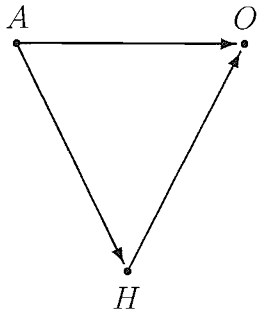


Figure 4.1: DAG represented by the saturated model under the order OHA

For example, consider the Alcohol, Obesity and Hypertension data set. In Chapter 3, analysis on these data was carried out under the ordering Variable 1 = Obesity, Variable 2 = Hypertension and Variable 3 = Alcohol Intake. Under the Cholesky decomposition parameterisation, this is equivalent to the following decomposition of the joint density of the latent data:

$$f(\boldsymbol{z}_i) = f(z_{iA})f(z_{iH}|z_{iA})f(z_{iO}|z_{iA}, z_{iH})$$

Under this ordering, it is only possible to make conditional independence statements about O and H given A, or about O and A given H. It is not possible to make any conditional independence statement about A and H given O, because O occurs last in the conditioning. Here setting $\phi_{AH} = 0$ imposes marginal independence of A and H.

Under this ordering, the saturated model is equivalent to the DAG shown in Figure 4.1.

Similarly, under this ordering, the single graphical model $A + OH$ used as an example in Section 3.7.3 has associated DAG displayed in Figure 4.2.

Clearly, the use of the Cholesky decomposition parameterisation particularly lends itself to the analysis of data where there is a natural ordering to the classifying variables, for example, in longitudinal data. An example of a data set where such ordering exists is taken from two general social surveys of adults in Germany published by the Central Archive for Empirical Social Science

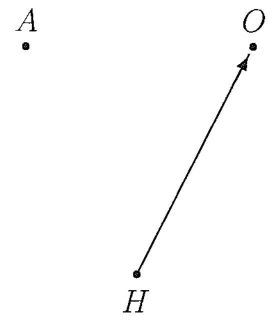


Figure 4.2: DAG represented by the model $A + OH$ under the order OHA

Research at the University of Cologne. The data set in full is given in Wermuth and Cox (1998). It contains 6039 individuals cross-classified by five factors: political attitude ("how well does the political system function today?", A) with four categories; type of formal schooling (B) with five categories; age group (C) with five categories; year of survey (D) with two categories and region of survey (E) with two categories. Variables A, B and C are ordinal and variables D and E are binary. Table 4.1 shows the levels of the five variables.

A	B	C	D	E
1: Very poorly	1: Basic incomplete	1: 19-29	1: 1991	1: West Germany
2: Poorly	2: Basic	2: 30-44	2: 1992	2: East Germany
3: Well	3: Medium	3: 45-59		
4: Very well	4: Upper medium	4: 60-74		
	5: Intensive	5: ≥ 75		

Table 4.1: Levels of the five variables for the Germany data

Variables D and E are fixed by design and must therefore come first in the ordering. Variable A is the primary variable of interest and is possibly dependent on all other variables which are explanatory; it must therefore come last in the ordering. For the remaining variables, it is possible that B depends on C. Therefore, C must come before B in the ordering. From this information, the ordering of the variables shown in Figure 4.3 can be derived.

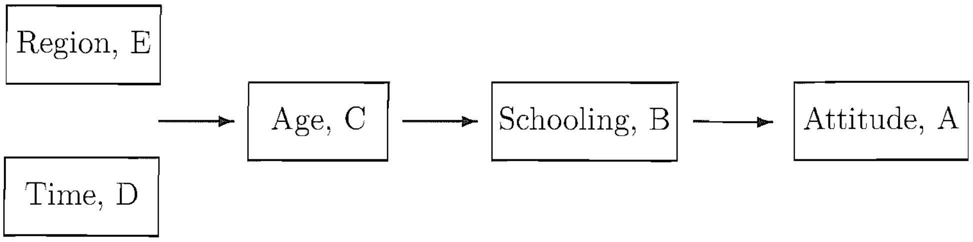


Figure 4.3: Ordering of the classifying variables for Germany data

Since D and E are both fixed by design, they can take either order DE or ED. The saturated model for the latent variable $\boldsymbol{z}$ under this ordering has the decomposed joint density

$$\begin{aligned}
f(\boldsymbol{z}_i) &= f(z_{iE})f(z_{iD}|z_{iE})f(z_{iC}|z_{iD},z_{iE})f(z_{iB}|z_{iC},z_{iD},z_{iE})f(z_{iA}|z_{iB},z_{iC},z_{iD},z_{iE}) \\
&= f(z_{iD})f(z_{iE}|z_{iD})f(z_{iC}|z_{iD},z_{iE})f(z_{iB}|z_{iC},z_{iD},z_{iE})f(z_{iA}|z_{iB},z_{iC},z_{iD},z_{iE})
\end{aligned}$$

So under the Cholesky decomposition parameterisation, the variables must be taken into the covariance matrix $\boldsymbol{\Sigma}$ (and hence into $\boldsymbol{\Phi}$) in either the order $EDCBA$ or the order $DECBA$.

4.2 Choice of Prior Parameters

In Chapter 3, the prior parameters were chosen to be noninformative. However, we must be more careful in our choice of prior parameters when it comes to model selection; a highly diffuse prior gives low probability to regions of the parameters space with non-negligible likelihood and hence the marginal likelihood can be very small. This behaviour is exacerbated in high dimensions hence leading to the selection of more parsimonious models. Conversely unjustifiably strong priors should be avoided. The choice of prior parameters for $\boldsymbol{\Sigma}$ is particularly tricky as it is unclear what effect varying the parameters of the Inverse-Wishart distribution q and $\boldsymbol{A}$ will have. In order to choose appropriate prior parameters for $\boldsymbol{\Sigma}$, we take the amount of information provided by the prior to be the same as that provided by a single observation from the likelihood, in a similar spirit to that suggested by Kass and Wasserman (1995). This is an example of a 'reference' prior, which is computed from modelling assumptions only and does not otherwise depend on the specifics of the problem. It is not possible to use many of the standard methods for eliciting prior distributions (Berger (1985), pages 74-117), because in all examples shown here, we do not have any prior information.

We begin our investigation of choice of prior parameters by considering the marginal distribution of the latent data for a single classifying variable, which is univariate normal $z_{ij} \sim N(\beta_j, \Sigma)$. To aid the approach, we first adjust the constraints imposed on the model parameters. Previously, the first and second cut points in each dimension were constrained to be 0 and 1 respectively, while for binary data, the single cut point was constrained to be 0 and the conditional variance was constrained to be 1. Clearly, setting such constraints 'scales' the distribution of the latent variable, but it is unclear how this scale may be quantified. To improve the situation, we now choose to constrain the first and last cut points in each dimension to be -1 and 1 respectively:

$$\theta(j,1) = -1 \qquad\qquad \theta(j,k_j-1) = 1$$

We assume that *a priori* each category is equally likely. Since the first and last cut points have been chosen so that the distribution of the latent data is symmetric around 0, it makes sense to choose the prior mean for β_j to be 0. To avoid making this prior too strong, we again choose the prior variance $\boldsymbol{T}$ to be the diagonal matrix $\boldsymbol{T} = \text{diag}(\tau_{11}, \tau_{22}, \dots, \tau_{pp})$ where τ_{ii} is large.

We now consider how to choose the prior parameters q and $\boldsymbol{A}$ for the matrix $\boldsymbol{\Phi}$. As seen in Chapter 3, the Inverse-Wishart prior for $\boldsymbol{\Sigma}$ has the following equivalent prior distributions for $\boldsymbol{\Phi}$:

$$\begin{aligned} \phi_{ii} &\sim \sqrt{b_i \chi^2_{q-i+1}} \\ \boldsymbol{\phi}_i | \phi_{ii} &\sim N_{p-i}(-\phi_{ii} \boldsymbol{a}_i \boldsymbol{A}_i^{-1}, \boldsymbol{A}_i^{-1}) \end{aligned} \tag{4.1}$$

where

$$b_i = (a_{ii} - \boldsymbol{a}_i \boldsymbol{A}_i^{-1} \boldsymbol{a}_i^T)^{-1}$$

and $\boldsymbol{A}^{-1}$ is partitioned as follows:

$$\begin{array}{rcl} & \begin{array}{c} i^{th} \\ \text{column} \\ \downarrow \end{array} & \\ \boldsymbol{A}^{-1} = & \left(\begin{array}{c|c|c} \ddots & \cdot & \cdots \\ \hline \cdot & a_{ii} & \boldsymbol{a}_i \\ \hline \vdots & \boldsymbol{a}_i^T & \boldsymbol{A}_i \end{array}\right) & \leftarrow i^{th} \text{ row} \end{array}$$

We wish to choose values for q and $\boldsymbol{A}$ so that the amount of information provided by the prior is the same as that provided by a single observation from the likelihood. The joint likelihood for n observations is given by:

$$\begin{aligned} f(\phi_{ii}, \boldsymbol{\phi}_i|\boldsymbol{z}) \quad \propto \prod_{\text{rows i of }\Phi} \quad & \phi_{ii}^n \exp\left[-\frac{\phi_{ii}^2}{2}(g_{ii} - \boldsymbol{g}_i G_i^{-1} \boldsymbol{g}_i^T)\right] \times \\ & \exp\left[-\frac{\phi_{ii}^2}{2}(\boldsymbol{\phi}_i + \phi_{ii}\boldsymbol{g}_i G_i^{-1}) G_i (\boldsymbol{\phi}_i + \phi_{ii}\boldsymbol{g}_i G_i^{-1})^T\right] \end{aligned}$$

where $\boldsymbol{G} = \sum_{i=1}^n (\boldsymbol{z}_i - \boldsymbol{\beta})(\boldsymbol{z}_i - \boldsymbol{\beta})^T$ is partitioned as follows:

$$\begin{array}{rcl} & \begin{array}{c} i^{th} \\ \text{column} \\ \downarrow \end{array} & \\ \boldsymbol{G} = & \left(\begin{array}{c|c|c} \ddots & \cdot & \cdots \\ \hline \cdot & g_{ii} & \boldsymbol{g}_i \\ \hline \vdots & \boldsymbol{g}_i^T & G_i \end{array}\right) & \leftarrow i^{th} \text{ row} \end{array}$$

so that the conditional likelihood of the partial row $\boldsymbol{\phi}_i$ given ϕ_{ii} is:

$$f(\boldsymbol{\phi}_i|\phi_{ii}, \boldsymbol{z}) \quad \propto \phi_{ii}^n \exp\left[-\frac{1}{2}(\boldsymbol{\phi}_i + \phi_{ii}\boldsymbol{g}_i G_i^{-1}) G_i (\boldsymbol{\phi}_i + \phi_{ii}\boldsymbol{g}_i G_i^{-1})^T\right]$$

i.e.

$$\boldsymbol{\phi}_i|\phi_{ii}, \boldsymbol{z} \sim N_{p-i}(-\phi_{ii}\boldsymbol{g}_i G_i^{-1}, G_i^{-1}) \tag{4.2}$$

Comparing (4.1) and (4.2), we see that in order for the prior to correspond to one unit of information, the matrix $\boldsymbol{A}^{-1}$ should be approximately equal to the matrix $\boldsymbol{G}$. However as $\boldsymbol{G}$ depends on unobserved data, we replace $\boldsymbol{A}^{-1}$ by

$E(\boldsymbol{G})$, under the model where the latent variables are independent. Hence the prior is centred on a null model. Under this model, the form of $\boldsymbol{G}$ is a diagonal matrix. Therefore, $\boldsymbol{A}^{-1}$ is chosen to be diagonal so that $\boldsymbol{a}_i = \boldsymbol{0}$. Thus the prior mean for $\boldsymbol{\phi}_i$ is $\boldsymbol{0}$. It therefore only remains to consider the diagonal elements $\boldsymbol{A}^{-1}$ which will determine the prior variance for $\boldsymbol{\phi_i}$ and all prior information for ϕ_{ii}. The diagonal elements of $\boldsymbol{A}^{-1}$ should be chosen to equal the diagonal elements of $\boldsymbol{G}$. Consider the distribution of $\boldsymbol{G}$. For a single observation, $\boldsymbol{G}$ is a diagonal matrix with elements $(z_{ij} - \beta_j)^2$. Now, for a particular j,

$$z_{ij} - \beta_j \sim N(0, \sigma_j^2)$$

where σ_j^2 is the marginal variance of z_{ij}, independently for $i = 1, \ldots, n$. Therefore,

$$\sum_{i=1}^{n}(z_{ij} - \beta_j)^2 \sim \sigma_j^2 \chi_n^2$$

Using the fact that the expectation of a chi-square distribution with n degrees of freedom is n,

$$E[\sum_{i=1}^{n}(z_{ij} - \beta_j)^2] = \sigma_j^2 n$$

We therefore have:

$$\boldsymbol{A}^{-1} = d \times \text{diag}(\sigma_j^2) \tag{4.3}$$

where d is the number of units of prior information. However this depends on the unknown σ_j^2 which is the marginal variance of z_{ij}.

Assuming *a priori* that for each classifying variable, each category is equally likely means that each category has probability $\frac{1}{k_j}$ where there are k levels for the jth variable. Thus, by considering the first category which is bounded above by the cut point at -1, the following statement can be made:

$$P(z_{ij} \leq -1) = \frac{1}{k_j}$$

Then, $z_{ij} \sim N(0, \sigma_j^2)$ implies

$$\phi\left(-\frac{1}{\sigma_j}\right) = \frac{1}{k_j}$$

with ϕ the N(0,1) cdf. Solving for σ_j:

$$\sigma_j = \frac{-1}{\phi^{-1}(1/k)} \tag{4.4}$$

The appropriate value for the corresponding diagonal element of the matrix $\boldsymbol{A}^{-1}$ is then calculated by taking the square of σ_j and multiplying by the number of units of prior information d:

$$a_{ii} = d\sigma_j^2$$

For the matrix $\boldsymbol{A}$, diagonal elements A_{ii} are found by taking the inverse:

$$A_{ii} = \frac{1}{a_{ii}} = \frac{1}{d\sigma_j^2} = \frac{\phi^{-1}(1/k)^2}{d} \tag{4.5}$$

If a classifying variable is binary, we cannot use this argument to estimate a value for σ_j^2. Instead we assume that the marginal variance σ_j^2 is approximately equal to the conditional variance $\frac{1}{\phi_{jj}^2}$, which is constrained to be 1. Hence $A_{ii} = \frac{1}{d}$. This can be justified by the prior centreing assumption that the classifying variables are independent and the latent data are normally distributed. Under this assumption, marginal and conditional variances are equal.

Finally we consider how to choose the degrees of freedom parameter q. For the pth latent variable (where there are p classifying variables) $\frac{1}{\phi_{pp}} = \sigma_p$. Hence in (4.3), we replace σ_p^2 by $E\left[\frac{1}{\phi_{pp}^2}\right]$

$$\boldsymbol{A}^{-1} = d \times \text{diag}\left[E\left(\frac{1}{\phi_{jj}^2}\right)\right] \tag{4.6}$$

Since $\phi_{pp}^2 \sim b_p\chi_{q-p+1}^2$ in the prior, and $b_p = \frac{1}{a_{pp}}$ because $\boldsymbol{A}^{-1}$ is diagonal, $\frac{1}{\phi_{pp}a_{pp}}$ is inverse gamma with parameters $\frac{q-p+1}{2}$ and $1/2$. Thus,

$$E\left(\frac{1}{\phi_{pp}^2}\right) = \frac{a_{pp}}{q-p-1}$$

From (4.6),

$$a_{pp} = d\frac{a_{pp}}{q-p-1}$$

This implies that $q = d + p + 1$. We wish to choose the number of units of prior information to be equivalent to one observation from the likelihood; hence $d = 1$. So, the degrees of freedom parameter q should be chosen as:

$$q = p + 2 \tag{4.7}$$

For one unit of prior information, Table 4.2 gives the values of A_{jj} corresponding to number of levels of the classifying variable. These values will be used in all examples that follow.

k_j	A_{jj}
2	1
3	0.185
4	0.455
5	0.708

Table 4.2: Choice of prior parameter A_{jj}

4.3 Reversible Jump Markov chain Monte Carlo (RJMCMC)

We require a method of comparison for the set of decomposable directed graphical models. The marginal likelihood is analytically intractable in this case so cannot be used. Instead we use a Reversible Jump Markov chain Monte Carlo (RJMCMC) approach to estimate posterior model probabilities. As explained above, models are characterised by the structure of the inverse covariance matrix Σ^{-1} and hence by $\mathbf{\Phi}$, with each different structure of $\mathbf{\Phi}$ corresponding to a unique decomposable directed graphical model. There exists an edge between variables i and j in the model if and only if there is a non-zero value for ϕ_{ij}. Conversely, no edge between variables i and j is equivalent to $\phi_{ij} = 0$.

A move to a new model can be made by either adding or subtracting an edge from the current model. An edge is added to the current model by proposing a non-zero value for an element of $\mathbf{\Phi}$ which was previously zero. An edge is removed from the current model by proposing a value of zero for a previously non-zero element of $\mathbf{\Phi}$. Reversible Jump provides a MCMC method that is capable of jumping between parameter spaces of differing dimensionality, which is exactly the situation described here.

Formally, we define the RJMCMC procedure as follows. At each stage of the RJMCMC, there are $\binom{p}{2}$ move types (each corresponding to one of the $\binom{p}{2}$ possible edges in the model or entries in $\mathbf{\Phi}$), and the null move. Each move involves removing an edge if already present, and adding an edge otherwise. These correspond respectively to proposing a new value for an element ϕ_{ij} which was previously 0, and setting an element ϕ_{ij} which previously had some non-zero value to 0. For the null move, no dimensional change is made, as it simply consists of re-generating parameters of the current model.

Suppose the current state of the Markov chain at time t is represented by $\left(m^{(t)}, \xi^{(t)}_{m^{(t)}}\right)$ where $\boldsymbol{\xi}^{(t)}_{m^{(t)}}$ represents the values of the unknown parameters in model $m^{(t)}$ at time t:

$$\boldsymbol{\xi}^{(t)}_{m^{(t)}} = \left(\boldsymbol{z}^{(t)}, \boldsymbol{\beta}^{(t)}_{m^{(t)}}, \mathbf{\Phi}^{(t)}_{m^{(t)}}, \boldsymbol{\theta}^{(t)}\right)$$

Adding an edge involves a proposed move to a new model m' and corresponding parameter vector $\boldsymbol{\xi}'_{m'}$, with dimension $\dim(\boldsymbol{\xi}^{(t)}_{m^{(t)}}) + 1$, *i.e.* there is one extra parameter to generate. Suppose $\boldsymbol{\xi}'_{m'}$ is created by generating a univariate proposal u from a proposal distribution $q_p(u)$ and setting $\boldsymbol{\xi}'_{m'} = g(\boldsymbol{\xi}^{(t)}_{m^{(t)}}, u)$, where g is a one-to-one function.

Removing an edge involves a move to a model m' with corresponding parameter vector $\boldsymbol{\xi}'_{m'}$ of dimension $\dim(\boldsymbol{\xi}^{(t)}_{m^{(t)}})-1$, then $\boldsymbol{\xi}'_{m'}$ is created from $\boldsymbol{\xi}^{(t)}_{m^{(t)}}$ by applying the inverse transformation $(\boldsymbol{\xi}'_{m'}, u') = g^{-1}(\boldsymbol{\xi}^{(t)}_{m^{(t)}})$ and discarding u'.

Suppose that the probability of making move type r given the current state of the Markov chain $\left(m^{(t)}, \boldsymbol{\xi}^{(t)}_{m^{(t)}}\right)$ is $j(r, m^{(t)}, \boldsymbol{\xi}^{(t)}_{m^{(t)}})$. Then Green (1995) showed that, if we propose to add an edge by generating a new parameter u from proposal distribution $q_r(u)$, $(m', \boldsymbol{\xi}'_{m'})$ should be accepted as the next realisation of the chain, so that $\left(m^{(t+1)}, \boldsymbol{\xi}^{(t+1)}_{m^{(t+1)}}\right) = (m', \boldsymbol{\xi}'_{m'})$ with probability $\alpha = \alpha(m^{(t)}, \boldsymbol{\xi}^{(t)}_{m^{(t)}}, m', \boldsymbol{\xi}'_{m'})$, where

$$\alpha = \min\left\{1, \frac{f(\boldsymbol{\xi}'_{m'}, m'|\boldsymbol{y})j(r, m', \boldsymbol{\xi}'_{m'})}{f(\boldsymbol{\xi}^{(t)}_{m^{(t)}}, m^{(t)}|\boldsymbol{y})j(r, m^{(t)}, \boldsymbol{\xi}^{(t)}_{m^{(t)}})q_r(u)}\left|\frac{\partial(\boldsymbol{\xi}'_{m'})}{\partial(\boldsymbol{\xi}^{(t)}_{m^{(t)}}, u)}\right|\right\} \tag{4.8}$$

and rejected otherwise, so that $m^{(t+1)} = m^{(t)}$.

Similarly, if we propose to drop an edge and move from parameter vector $\boldsymbol{\xi}^{(t)}_{m^{(t)}}$ to $\boldsymbol{\xi}'_{m'}$, $(m', \boldsymbol{\xi}'_{m'})$ should be accepted as the next realisation of the chain with probability $\alpha = \alpha(m^{(t)}, \boldsymbol{\xi}^{(t)}_{m^{(t)}}, m', \boldsymbol{\xi}'_{m'})$, where

$$\alpha = \min\left\{1, \frac{f(\boldsymbol{\xi}'_{m'}, m'|\boldsymbol{y})j(r, m', \boldsymbol{\xi}'_{m'})q_r(\boldsymbol{\xi}^{(t)}_{m^{(t)}\setminus m'})}{f(\boldsymbol{\xi}^{(t)}_{m^{(t)}}, m^{(t)}|\boldsymbol{y})j(r, m^{(t)}, \boldsymbol{\xi}^{(t)}_{m^{(t)}})}\left|\frac{\partial(\boldsymbol{\xi}'_{m'})}{\partial(\boldsymbol{\xi}^{(t)}_{m^{(t)}}, u)}\right|\right\} \tag{4.9}$$

and rejected otherwise.

We now need to specify each element of these acceptance probabilities. Firstly, for our approach, the transformation g is chosen to be the identity transformation, so u is simply the additional parameter in $\boldsymbol{\Psi}$ when adding an edge where $\boldsymbol{\Psi}$ is the matrix with elements $\psi_{ij} = \frac{\phi_{ij}}{\phi_{ii}}$ and 1s on the diagonal. We use $\boldsymbol{\Psi}$ as the off-diagonal elements $\boldsymbol{\phi}_i$ are generated *via* draws from the conditional posterior distribution of $\boldsymbol{\psi}_i$ which is standard. We therefore must include $\boldsymbol{\Psi}$ in the vector of unknown parameters $\boldsymbol{\xi}$. However, we do not remove $\boldsymbol{\Phi}$ as the diagonal elements ϕ_{ii} cannot be inferred from $\boldsymbol{\Psi}$. Since g is the identity transformation, the Jacobian term in the acceptance probability is simply 1. Secondly, each move type is chosen to be made with equal probability, regardless of the current state, so these terms $(j(r, m^{(t)}, \boldsymbol{\xi}^{(t)}_{m^{(t)}})$ and $j(r, m', \boldsymbol{\xi}'_{m'}))$ cancel.

The joint posterior distribution of the model and its associated parameters can be simplified. Dropping the m suffix:

$$\begin{aligned}
f(\boldsymbol{\xi}_m, m|\boldsymbol{y}) &= f(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\Psi}, \boldsymbol{\Phi}, \boldsymbol{\beta}, m|\boldsymbol{y}) \\
&\propto \frac{f(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\theta}|\boldsymbol{\Psi}, \boldsymbol{\Phi}, \boldsymbol{\beta}, m) f(\boldsymbol{\Psi}, \boldsymbol{\Phi}, \boldsymbol{\beta}, m)}{f(\boldsymbol{y})} \\
&= \frac{f(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\Psi}, \boldsymbol{\Phi}, \boldsymbol{\beta}, m) f(\boldsymbol{z}, \boldsymbol{\theta}|\boldsymbol{\Psi}, \boldsymbol{\Phi}, \boldsymbol{\beta}, m) f(\boldsymbol{\Psi}, \boldsymbol{\Phi}, \boldsymbol{\beta}|m) f(m)}{f(\boldsymbol{y})} \\
&= \frac{f(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{\theta}) f(\boldsymbol{z}|\boldsymbol{\theta}, \boldsymbol{\Psi}, \boldsymbol{\Phi}, \boldsymbol{\beta}, m) f(\boldsymbol{\theta}) f(\boldsymbol{\Psi}, \boldsymbol{\Phi}, \boldsymbol{\beta}|m) f(m)}{f(\boldsymbol{y})}
\end{aligned} \tag{4.10}$$

The final step uses the fact that $\boldsymbol{y}$ is determined purely by $\boldsymbol{z}$ and the prior for $\boldsymbol{\theta}$ is independent of the model and other parameters. When this expression is substituted into (4.8) and (4.9), then $f(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{\theta})$, $f(\boldsymbol{\theta})$ and $f(\boldsymbol{y})$ will all cancel in the numerator and denominator as they are model independent. We assume that all models are, *a priori*, equally likely; hence $f(m)$ will cancel in (4.8) and (4.9) as well. This leaves

$$\alpha = \min\left\{1, \frac{f(\boldsymbol{z}'_{m'}|\boldsymbol{\Phi}'_{m'}, \boldsymbol{\beta}'_{m'}) f(\boldsymbol{\Phi}'_{m'}, \boldsymbol{\beta}'_{m'}|m')}{f(\boldsymbol{z}^{(t)}_{m^{(t)}}|\boldsymbol{\Phi}^{(t)}_{m^{(t)}}, \boldsymbol{\beta}^{(t)}_{m^{(t)}}) f(\boldsymbol{\Phi}^{(t)}_{m^{(t)}}, \boldsymbol{\beta}^{(t)}_{m^{(t)}}|m^{(t)})} \frac{1}{q_r(\boldsymbol{\Psi}'_{m'})}\right\} \tag{4.11}$$

and

$$\alpha = \min\left\{1, \frac{f(\boldsymbol{z}'_{m'}|\boldsymbol{\Phi}'_{m'}, \boldsymbol{\beta}'_{m'}) f(\boldsymbol{\Phi}'_{m'}, \boldsymbol{\beta}'_{m'}|m')}{f(\boldsymbol{z}^{(t)}_{m^{(t)}}|\boldsymbol{\Phi}^{(t)}_{m^{(t)}}, \boldsymbol{\beta}^{(t)}_{m^{(t)}}) f(\boldsymbol{\Phi}^{(t)}_{m^{(t)}}, \boldsymbol{\beta}^{(t)}_{m^{(t)}}|m^{(t)})} \frac{q_r(\boldsymbol{\Psi}^{(t)}_{m^{(t)}\backslash m'})}{1}\right\} \tag{4.12}$$

when adding or removing edges respectively. Note that the distribution of $\boldsymbol{z}_m$ is fully determined by $\boldsymbol{\Phi}_m$ and $\boldsymbol{\beta}_m$ since all the information about the model m is encapsulated in $\boldsymbol{\Phi}_m$.

It now remains to specify the likelihoods $f(\boldsymbol{z}_m|\boldsymbol{\Phi}_m, \boldsymbol{\beta}_m)$, the prior distributions $f(\boldsymbol{\Phi}_m, \boldsymbol{\beta}_m|m)$ and the proposal distribution $q_p(\boldsymbol{\Phi}_m)$.

The likelihood of the latent data $\boldsymbol{z}$ is given by the normal distribution, the parameters of which depend on the model. The dimension of the mean $\boldsymbol{\beta}$ is independent of the model, but the structure of the variance $(\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}$ varies with the model (i.e. by setting appropriate elements of $\boldsymbol{\Phi}$ to zero).

We now consider the prior distributions for $\boldsymbol{\Phi}_m$ and $\boldsymbol{\beta}_m$, which can be decomposed as:

$$f(\boldsymbol{\Phi}_m, \boldsymbol{\beta}_m|m) = f(\boldsymbol{\beta}|m)\prod_{i=1}^{p} f(\phi_{ii}|m)f(\boldsymbol{\phi}_i|\phi_{ii}, m) \tag{4.13}$$

The priors for $\boldsymbol{\beta}$, ϕ_{ii} and $\boldsymbol{\psi}_i$ are conditionally independent given m and the priors for $\boldsymbol{\beta}$ and ϕ_{ii} are chosen to be independent of the model. Therefore in the expressions for α (4.11) and (4.12), these will cancel out, leaving only the priors for the partial rows of $\boldsymbol{\Phi}$, $\{\boldsymbol{\phi}_i\}$. These priors have the multivariate normal distribution, as specified in (3.8), conditioned on the model using the standard result for conditional multivariate normal distributions.

Lastly, we consider the proposal distribution $q_r(\boldsymbol{\Psi}_m)$. Here, the use of the Cholesky decomposition parameterisation provides us with a further useful property that a suitable proposal distribution can be specified to be the conditional posterior distribution of the element of $\boldsymbol{\Psi}$ to be added or removed. This is not often possible in RJMCMC samplers, as normalised conditional distributions are often only available when the marginal likelihood (including the posterior normalising constant) can be evaluated. Here, the set up of the model and the subsequent Gibbs sampler algorithm is useful in that the conditional density for $\boldsymbol{\psi}_i$ is already supplied in (3.11). It is:

$$\boldsymbol{\psi}_i|\phi_{ii} \sim N_{p-i}(\boldsymbol{\eta}, \boldsymbol{P}) \tag{4.14}$$

where

$$\begin{aligned} \boldsymbol{\eta} &= (\boldsymbol{\mu}_i\boldsymbol{A}_i - \boldsymbol{g}_i)(\boldsymbol{A}_i + \boldsymbol{G}_i)^{-1} && (4.15)\\ \boldsymbol{P} &= \frac{1}{\phi_{ii}^2}(\boldsymbol{A}_i + \boldsymbol{G}_i)^{-1} && (4.16) \end{aligned}$$

However, when moving between models, we are proposing to change just one element of the vector $\boldsymbol{\psi}_i$ while other elements remain fixed. We therefore need to find the conditional distribution for $\psi_{ij}|\phi_{ii}, \boldsymbol{\psi}_{i,k\neq j}$. This is derived from (4.14) using the standard result for conditional distributions of subsets of multivariate normally distributed random variables. Re-ordering the vector $\boldsymbol{\psi}_i$ so that $\boldsymbol{\psi}_i^T = (\psi_{ij}, \boldsymbol{\psi}_{i,k\neq j}^T)$ and partitioning $\boldsymbol{\eta}$ and $\boldsymbol{P}$ as follows:

$$\boldsymbol{\eta}^T = (\eta_j, \boldsymbol{\eta}_{k\neq j}^T)$$

$$\boldsymbol{P} = \begin{pmatrix} p_{jj} & \boldsymbol{p}_j \\ \boldsymbol{p}_j^T & \boldsymbol{P}_j \end{pmatrix}$$

the proposal distribution for ψ_{ij} is:

$$\psi_{ij}|\phi_{ii}, \boldsymbol{\psi}_{i,k\neq j} \sim N\left(\eta_j + \boldsymbol{p}_j \boldsymbol{P}_j^{-1}(\boldsymbol{\psi}_{i,k\neq j} - \boldsymbol{\eta}_{k\neq j})\ ,\ p_{jj} - \boldsymbol{p}_j \boldsymbol{P}_j^{-1} \boldsymbol{p}_j^T\right) \qquad (4.17)$$

The corresponding value for ϕ_{ij} is then found by simply transforming from ψ_{ij} using the identity (3.10) i.e., $\phi_{ij} = \psi_{ij}\phi_{ii}$.

The null move for the saturated model was described in Chapter 3. If the current model is not the saturated model, then this method is easily adapted as described in 3.7.3. The Reversible Jump procedure is now fully specified. The Algorithm runs as follows:

4.3.1 Algorithm 5

An initial model and values for all parameters in this model are specified. Then the following process is carried out.

1. With probability p, remain in the current model and re-generate all model parameters (the null move). Else, with probability $1 - p$, a new model is proposed by either adding or subtracting a randomly selected edge (edges are chosen with equal probability) from the current model, involving generating or setting to zero the corresponding ψ_{ij}.
2. Generate the random variate $u \sim \text{Uniform}(0, 1)$.
3. If $u < \alpha$, accept proposed model (with probability α). Otherwise reject.
4. Go back to step 1 and repeat.

4.4 Examples

We apply the Reversible Jump algorithm to the Germany data set which, as described above, already has a natural ordering. These data have also been analysed by Wermuth and Cox (1998) using frequentist methods.

Since D and E are binary variables, we constrain their conditional and marginal variances respectively: $\phi_{11} = \phi_{22} = 1$. There are also no free cut points to be generated for these variables whereas A, B and C have one, two and two free cut points to be generated respectively. The prior parameters were chosen using the methods described in Section 4.2 and are $\boldsymbol{A} = \text{diag}(0.455, 0.708, 0.708, 1, 1)$, $q = 7$ and $\boldsymbol{T} = \text{diag}(50)$. The Reversible Jump algorithm was implemented using 100,000 iterations. Table 4.4 shows the posterior model probabilities:

Model	Posterior Model Probability
$ABC + ADE$	0.474
$ABC + ACD + ADE$	0.280
$ABC + ABE + ADE$	0.104
$ABC + ACD + ABE + ADE$	0.039

Table 4.3: Posterior model probabilities for Germany data

These results show that the data set is highly structured, with the outcome of interest - "how well does the political system function today?" being related in a complex manner to all predictive variables.

For ease of comparison with the results of Wermuth and Cox (1998), we split the data set into two marginal tables, one for each value of variable E. Algorithm 5 was then applied to both marginal tables, each with 100,000 iterations. The prior parameters for each region were $\boldsymbol{A} = \text{diag}(0.455, 0.708, 0.708, 1)$, $q = 6$ and $\boldsymbol{T} = \text{diag}(50)$ and the posterior model probabilities are shown in Table 4.4. The Reversible Jump algorithm explored the model space fairly

East Germany		West Germany	
Model	Posterior Probability	Model	Posterior Probability
$AD + BC + BD$	0.262	$ABC + AD$	0.654
$AD + BC$	0.203	$ABC + ACD$	0.286
$AB + AD + BC$	0.201	$ABC + ABD$	0.044
$ABD + BC$	0.114	$ABCD$	0.017

Table 4.4: Posterior model probabilities for Germany data by region

rapidly with the proposed move accepted approximately 16% of the time for East Germany and 12% of the time for West Germany. Generally the most probable models for East Germany are simpler than those for West Germany. However, the posterior model probabilities are more diffuse for East Germany, indicating greater model uncertainty. While the first two models for the West Germany data set account for 95% of the posterior model probability, it takes the first five models for the East Germany data set to account for 87% of the posterior model probabilities. For both East and West, political opinion is dependent on year of survey, with further dependences in West Germany on age group and type of formal schooling. As might be expected, in both regions, schooling (B) is linked to age group (C). In East Germany, political opinion is conditionally independent of formal schooling given year of survey. The most probable models for each region are shown in Figure 4.4, with those predicted by Wermuth and Cox (1998) shown underneath. While the most probable model selected using the method described here for West Germany agrees with that selected by Wermuth and Cox, the most probable models for East Germany differ. However, there is not much to choose between them in terms of posterior probability so this does not give cause for concern.

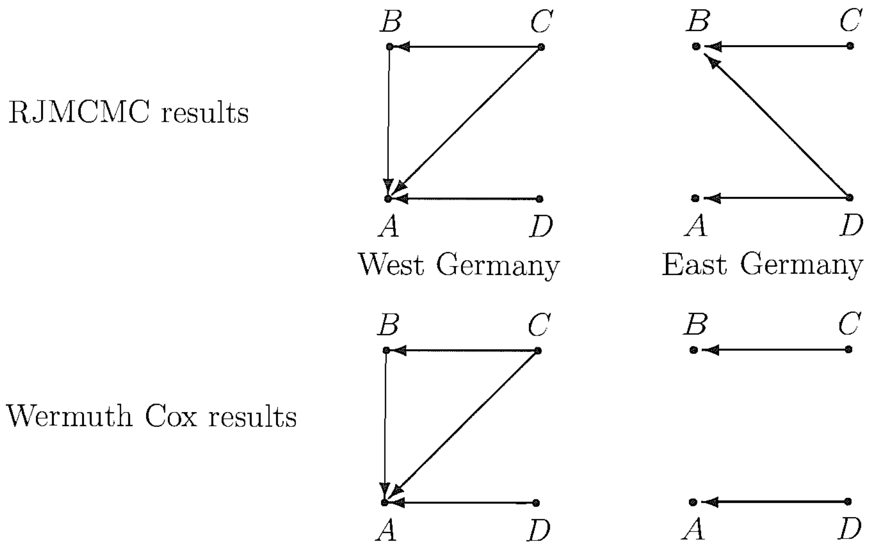


Figure 4.4: Most probable models by region for the Germany data

4.4.1 Model diagnostics

We will use a similar method to that described in 3.4.2 to assess the fit of these models. Since there is a natural splitting of the data over region, we consider the fit of the models separately for East and West Germany. We consider the model averaged fit. This means that instead of just considering the fit of the most probable model or the set of most probable models, we consider all models generated during the Reversible Jump procedure. To do this, N tables were generated, where N is the number of iterations in the Reversible Jump procedure, using the mean, variance and cut points produced under the model at each iteration. The generated cell counts are then compared with the posterior predictive mean table using the same method and distance measures as described in Section 3.4.2.

West and East gave very similar results, we therefore just show those for West Germany. Figures 4.5, 4.6 and 4.7 show the densities of the chi-squared, deviance and distance measures respectively, with the vertical dotted line again representing the distance between the observed data and the model averaged posterior predictive mean.

The main point to note is that there is a much poorer fit than in previous examples. This is shown by the vertical line representing this distance being well into the tail of the distribution. There are three possible reasons for this: either the models selected are not the most appropriate to describe the data, that is, the Reversible Jump procedure is choosing the wrong models; or the distance measures used are inappropriate for assessing goodness-of-fit for varying dimensional models fitted using the Reversible Jump procedure; or the parameterisation is such that even the saturated model would not predict the data especially well. The first of these possible reasons may be discounted by fitting the saturated model and observing that the fit is only a very slight improvement on that provided by the model averaged choice of models. In order to see this, compare the goodness-of-fit graphs in Figures 4.8, 4.9 and 4.10 with those in Figures 4.5, 4.6 and 4.7.

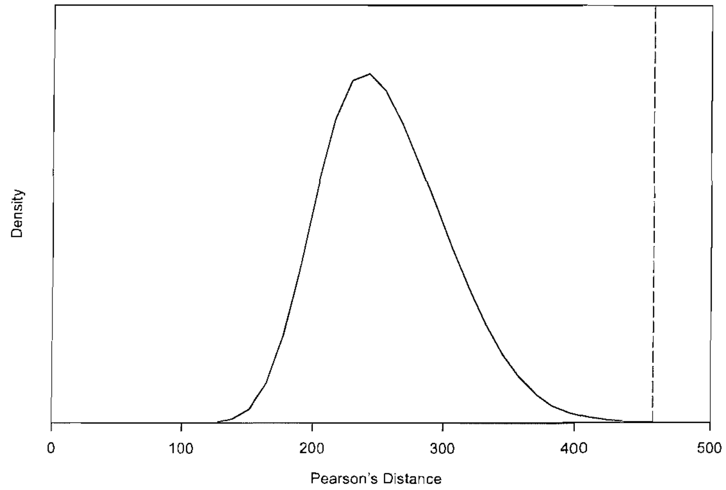


Figure 4.5: Estimated density of Pearson's distance measure for West Germany data

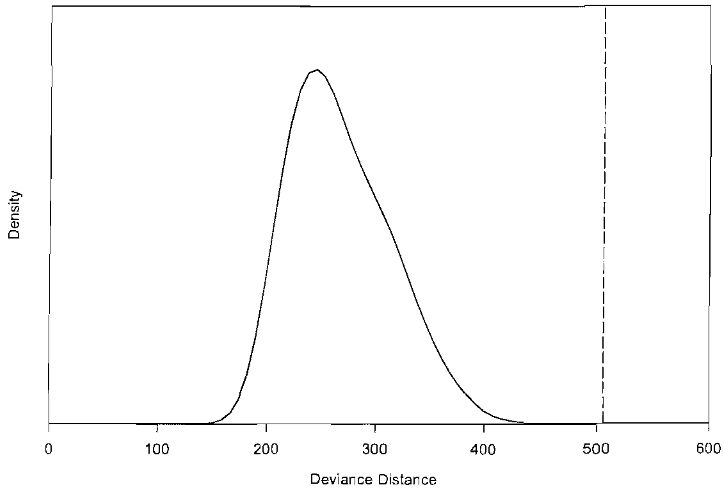


Figure 4.6: Estimated density of deviance distance measure for West Germany data

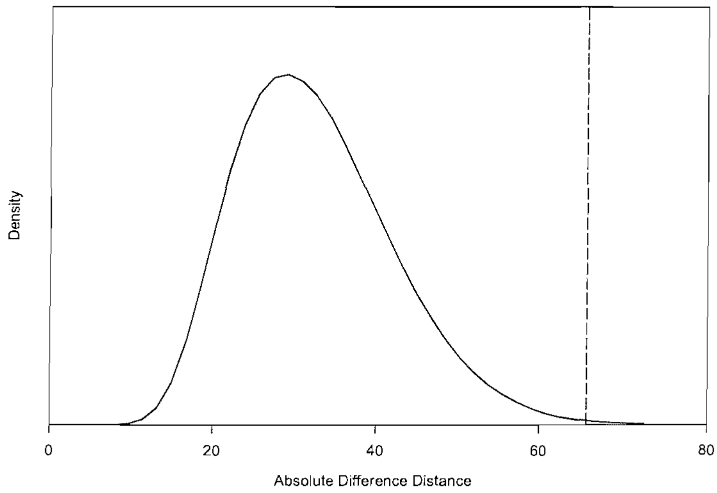


Figure 4.7: Estimated density of maximum absolute difference distance measure for West Germany data

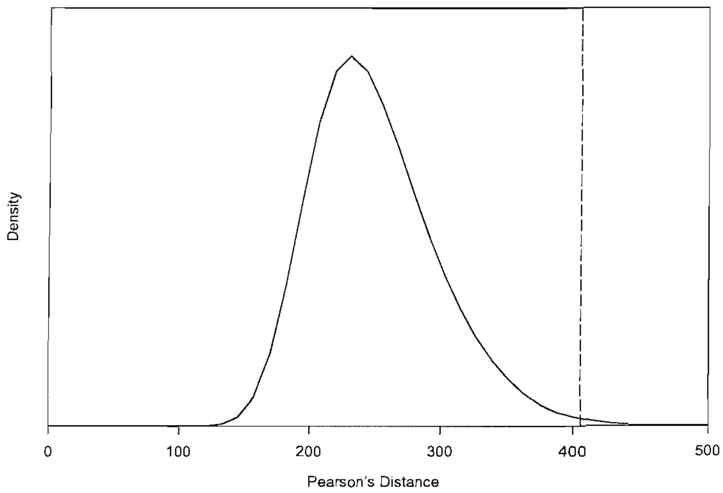


Figure 4.8: Estimated density of Pearson's distance measure for West Germany data using the saturated model

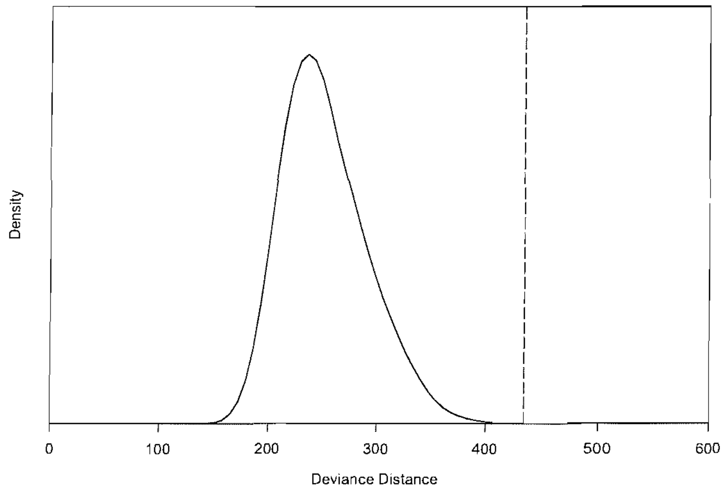


Figure 4.9: Estimated density of deviance distance measure for West Germany data using the saturated model

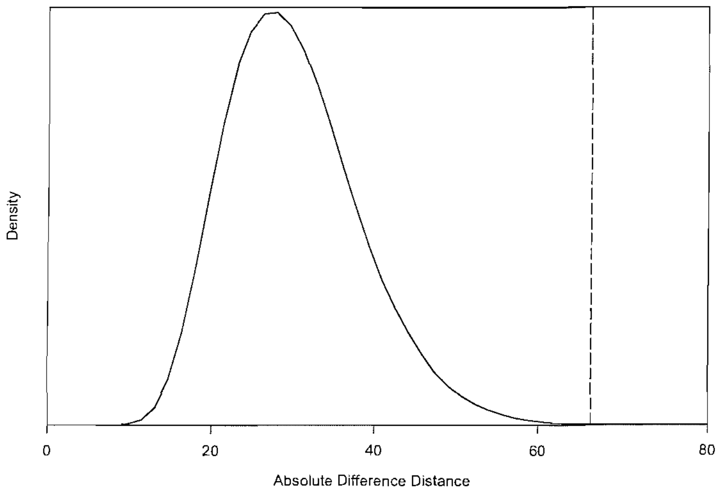


Figure 4.10: Estimated density of maximum absolute difference distance measure for West Germany data using the saturated model

In order to investigate the second possible reason, Table 4.5 shows the posterior predictive mean data (rounded to the nearest integer) with true values in parentheses. From this table, it is clear that the model is providing a fairly reasonable fit in many areas of the table, but there are also a number of areas where the fit is less good, for example the cells where A=2, B=1 and A=2, B=2. From this we can see that a small number of cells contribute a large proportion of the distance between the observed data and the model averaged posterior predictive mean data.

The relative lack of fit is perhaps not unexpected as the parsimonious nature of the model fitting procedure means that there are only 18 parameters with which to predict 200 cellcounts. In fact, it is perhaps surprising that the model provides as good a fit as it does in most examples as it is much more parsimonious than a log-linear model with every two factor interaction contributing a single parameter. Lack of fit could be due to the assumption that the latent variable $\boldsymbol{z}$ has the multivariate normal distribution being inappropriate.

Despite these difficulties, it is encouraging to note that despite not being able to predict the data as well as we may have hoped, the models chosen to do so are very similar to those selected by others. The main focus of this part of the work is on model selection. Convergence was assessed using trace plots and found to be satisfactory. For example, Figure 4.11 shows the trace plot for β_4 (Time) for West Germany, and figure 4.12 shows the corresponding autocorrelation function plot.

	A	1		2		3		4	
	D	1	2	1	2	1	2	1	2
B	C								
1	1	0 (1)	0 (0)	4 (3)	5 (2)	2 (1)	3 (3)	0 (2)	1 (0)
	2	1 (0)	1 (0)	10 (1)	14 (4)	4 (1)	9 (6)	1 (0)	2 (1)
	3	2 (0)	1 (1)	13 (2)	19 (2)	4 (1)	11 (6)	1 (0)	2 (0)
	4	2 (2)	2 (1)	15 (6)	24 (4)	4 (1)	13 (5)	1 (1)	3 (0)
	55	1 (1)	1 (0)	7 (1)	13 (3)	2 (0)	6 (1)	0 (1)	1 (0)
2	1	7 (5)	4 (6)	55 (63)	69 (68)	17 (5)	38 (40)	2 (2)	8 (6)
	2	15 (24)	10 (10)	103 (135)	140 (186)	28 (34)	69 (102)	3 (2)	13 (14)
	3	15 (26)	11 (19)	93 (120)	136 (182)	23 (27)	61 (102)	2 (6)	11 (11)
	4	15 (41)	12 (11)	84 (107)	132 (177)	18 (18)	54 (82)	2 (3)	9 (21)
	5	5 (8)	5 (12)	26 (8)	45 (51)	5 (9)	16 (22)	0 (0)	2 (7)
3	1	14 (10)	9 (4)	92 (88)	114 (101)	23 (18)	52 (48)	3 (0)	9 (8)
	2	22 (17)	15 (7)	121 (89)	166 (100)	26 (14)	67 (67)	3 (1)	11 (5)
	3	18 (4)	13 (11)	87 (62)	127 (76)	17 (10)	46 (24)	2 (2)	7 (3)
	4	14 (12)	11 (7)	60 (32)	96 (57)	11 (3)	31 (10)	1 (0)	4 (5)
	5	4 (3)	3 (2)	13 (8)	23 (16)	2 (2)	7 (6)	0 (0)	1 (1)
4	1	4 (3)	2 (0)	21 (22)	26 (17)	5 (5)	11 (3)	0 (0)	2 (2)
	2	5 (3)	3 (4)	23 (26)	31 (47)	4 (1)	11 (10)	0 (0)	2 (0)
	3	3 (2)	2 (2)	14 (17)	20 (17)	2 (2)	6 (6)	0 (0)	1 (1)
	4	2 (1)	2 (0)	8 (4)	13 (9)	1 (0)	4 (1)	0 (0)	0 (0)
	5	0 (0)	0 (0)	1 (3)	3 (1)	0 (0)	1 (1)	0 (0)	0 (0)
5	1	13 (8)	8 (7)	60 (78)	74 (100)	12 (9)	26 (29)	1 (1)	4 (1)
	2	12 (11)	9 (8)	50 (68)	69 (99)	8 (10)	21 (25)	1 (1)	3 (3)
	3	7 (5)	5 (9)	26 (29)	39 (42)	4 (3)	11 (13)	0 (0)	1 (1)
	4	4 (7)	3 (4)	14 (26)	22 (26)	2 (2)	5 (6)	0 (0)	1 (3)
	5	1 (2)	1 (1)	2 (6)	4 (6)	0 (0)	1 (4)	0 (0)	0 (0)

Table 4.5: Fitted values for West Germany data with true values in parentheses

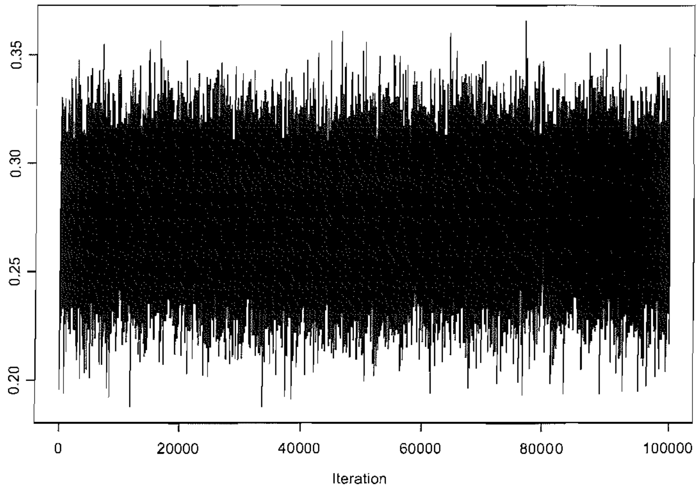


Figure 4.11: Trace plot for β_4 for West Germany data

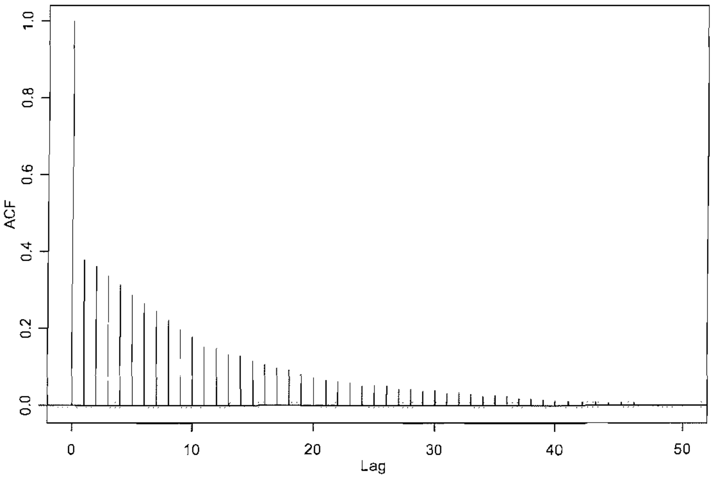


Figure 4.12: Autocorrelation function plot for β_4 for West Germany data

4.5 Example 2 - Coronary Heart Disease Risk Factors Data

This data set is taken from Edwards and Havranek (1985) and concerns 1841 men cross-classified according to six coronary heart disease risk factors: A, smoking (yes or no); B, strenuous mental work (yes or no) ; C, strenuous physical work (yes or no); D, systolic blood pressure (< 140 or ≥ 140); E, ratio of α and β lipoproteins (< 3 or ≥ 3); F, family anamnesis of coronary heart disease (positive or negative). The data are shown in Table 4.6.

		F	Negative				Positive			
		B	No		Yes		No		Yes	
		A	No	Yes	No	Yes	No	Yes	No	Yes
E	D	C								
< 3	< 140	No	44	40	112	67	5	7	21	9
		Yes	129	145	12	23	9	17	1	4
	≥ 140	No	35	12	80	33	4	3	11	8
		Yes	109	67	7	9	14	17	5	2
≥ 3	< 140	No	32	32	70	66	7	3	14	14
		Yes	50	80	7	13	9	16	2	3
	≥ 140	No	24	25	73	57	4	0	13	11
		Yes	51	63	7	16	5	14	4	4

Table 4.6: Risk factors for coronary heart disease

Following the argument of Madigan and Raftery (1994), the variables are assumed to take the ordering $FCBAED$, that is, all possible edges are directed and any edge between two variables leads from the variable earlier in the ordering to the variable which comes later in the ordering. Due to the way the Cholesky decomposition is parameterised, this means that variables must be taken into the covariance matrix in the order $DEABCF$. All six variables are binary. We therefore fix the conditional variance for each to be 1: $\phi_{jj} = 1$ for all j. Under the method described in 4.2, priors were chosen to be $\boldsymbol{A} = \text{diag}(1)$, $q = 8$ and $\boldsymbol{T} = \text{diag}(50)$. Algorithm 6 was applied using 200,000 iterations, and the resulting posterior model probabilities are shown in Table 4.5. The RJMCMC was found to be very mobile with proposed moves accepted approximately 34% of the time. The four most probable models all contain the edges

Model	Posterior Model Probability
$CB + CA + AE + BE + AD + ED$ (a)	0.474
$CB + CA + AE + BE + AD + ED + CE$ (b)	0.280
$CB + CA + AE + CE + AD + ED$ (c)	0.104
$CB + CA + AE + BE + ED$ (d)	0.039

Table 4.7: Posterior model probabilities for directed models for heart disease data

CB, CA, AE, ED. The edges AD and BE occur in three out of the four most probable models, with the edge CE occurring twice. The four most probable models are illustrated in Figure 4.13.

The most striking feature of these models is the high posterior probability of marginal independence of F (family anamnesis of coronary heart disease). For comparison, the two most probable models found by the method of Edwards and Havranek are models (a) and (c) respectively, while the two most probable models found by the method of Madigan and Raftery are models (c) and (a) respectively. The main difference between the results gained here and by them is in the selection of model (b) but this simply contains the union of the edges of models (a) and (c).

Goodness-of-fit was assessed using the simulation method. The Pearson's distance graph shown in Figure 4.14 is typical of the model fit. As for the Germany data, the fit is not particularly good. However, the posterior predictive mean values (shown in Table 4.6) themselves appear to be fairly close to the true values, perhaps indicating that the distance measure method for assessing goodness-of-fit is in appropriate for higher-dimensional and varying-dimensional models. It is important to note again that the main focus of the work is on model selection.

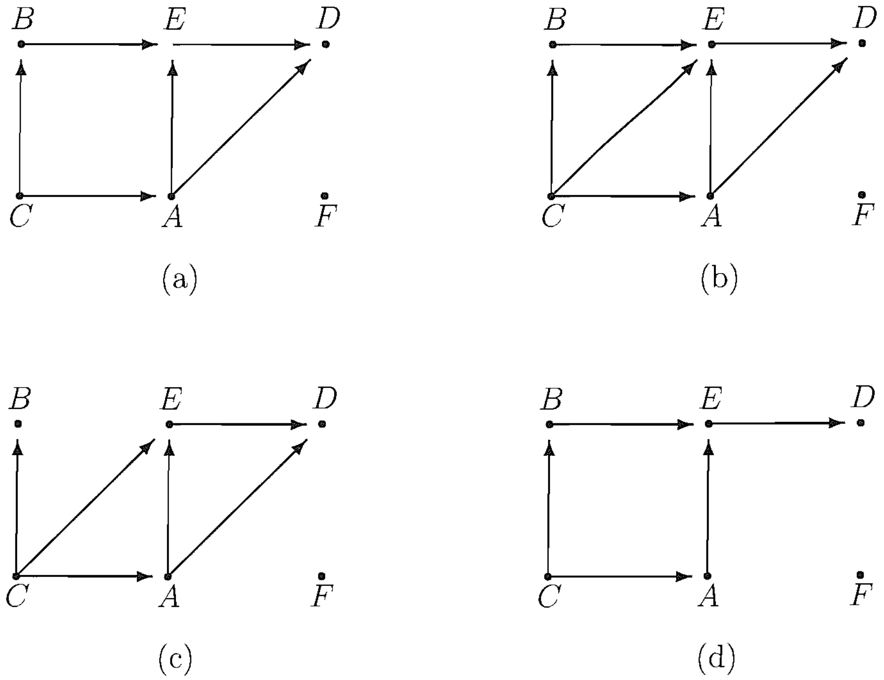


Figure 4.13: Most probable models for the heart disease data

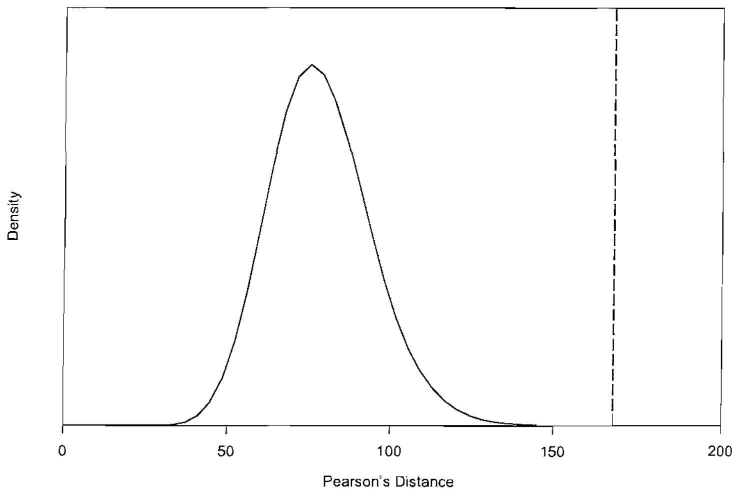


Figure 4.14: Estimated density of deviance distance measure for heart disease data

		F	Negative				Positive			
		B	No		Yes		No		Yes	
		A	No	Yes	No	Yes	No	Yes	No	Yes
E	D	C								
< 3	< 140	No	62.9	48.2	86.5	60.3	10.2	7.8	14.2	10.0
		Yes	106.6	112.6	32.3	30.9	17.2	18.3	5.3	5.1
	≥ 140	No	46.6	30.5	64.7	38.5	7.6	5.0	10.7	6.4
		Yes	78.0	70.4	23.9	19.5	12.7	11.5	4.0	3.3
≥ 3	< 140	No	36.6	37.7	62.6	58.3	6.0	6.2	10.4	9.8
		Yes	51.8	73.3	19.6	25.2	8.5	12.0	3.3	4.2
	≥ 140	No	33.1	29.1	56.9	45.3	5.4	4.8	9.5	7.6
		Yes	46.2	55.8	17.6	19.4	7.6	9.2	2.9	3.3

Table 4.8: Mean posterior predictive cell counts for coronary heart disease data

Convergence was assessed using trace plots and was found to be satisfactory.

In this Chapter, we have restricted the methods to finding DAGs for data where the classifying variables are ordered. In the next Chapter, we will extend the methods developed here to investigate model selection for data where there is no clear ordering to the classifying variables.

Chapter 5

Model Determination for Undirected Graphs

5.1 Relationship between Directed and Undirected Decomposable Graphical Models

In Chapter 4, the model space was restricted to the set of decomposable directed graphical models. Given the way models were parameterised using the Cholesky decomposition, this was the natural set of models to consider. However, such models are easiest to interpret when the classifying variables take a natural ordering. In this Chapter, we extend the model determination method to the class of undirected decomposable graphical models, and as a consequence, to those data sets where there is no single natural ordering.

In order to do this, we consider the relationship between directed and undirected graphical models. Given a directed graph $\mathcal{D}$, it is possible to construct an undirected graph $\mathcal{G}$ with the same Markov structure (same conditional independence structure). See Dawid and Lauritzen (1993). The associated undirected graph $\mathcal{G}$ is obtained from $\mathcal{D}$ by taking the associated moral graph $\mathcal{D}^{\Updownarrow}$ of $\mathcal{D}$ and replacing the directed edges by undirected edges. The moral graph $\mathcal{D}^{\Updownarrow}$ is obtained from $\mathcal{D}$ by "marrying" all unmarried parents in the graph, *i.e.*

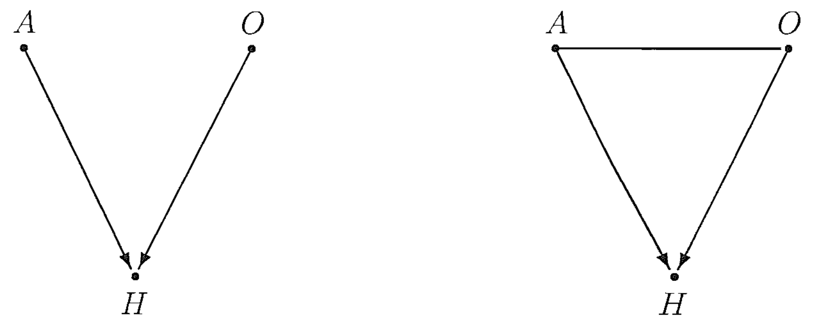


Figure 5.1: Example of moralising

if any two parents of a variable are not connected by an edge, an undirected edge will be added between them. This is illustrated in Figure 5.1.

However, note that the models implied by the two graphs in Figure 5.1 are not equivalent, as the directed graph implies marginal independence of A and O.

As a consequence of this, in any ordering of the variables, there are some undirected graphs whose conditional independence structure does not correspond to that of any directed graph. For example, consider again the Alcohol, Obesity and Hypertension data set, in the order Alcohol Intake, Hypertension, Obesity. Suppose we wish to consider the undirected graphical model $AO + HO$ so that A and H are conditionally independent given O. This is illustrated in Figure 5.2.

Under the ordering OHA, the only directed model that could have the same conditional independence structure is the one with the same edge set, shown in Figure 5.2. The arrows take their directions from the ordering, with arrows going from vertices that are higher in the ordering to those lower down. However this does not have the same Markov structure as the undirected graph, due to the fact that the directed graph must be moralised by adding an edge between A and H to find its Markov equivalent undirected graph. Hence, the directed graph shown in Figure 5.2 has the conditional independence struc-

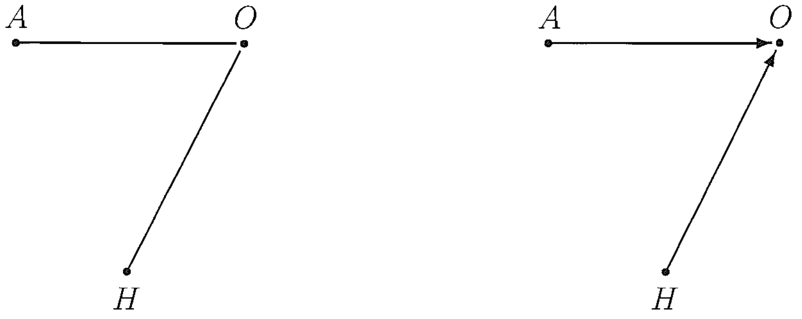


Figure 5.2: Model $AO + HO$ and corresponding (but non Markov equivalent) undirected model.

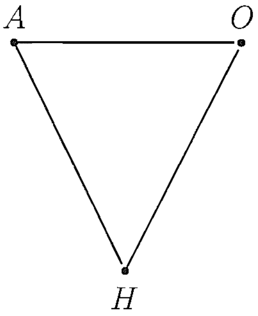


Figure 5.3: Equivalent saturated model with correct Markov structure.

ture of the undirected graph shown in Figure 5.3, which implies no conditional independences.

To summarise, in a given ordering (which is equivalent to a given parameterisation of the Cholesky decomposition) only certain patterns of zeros in $\mathbf{\Phi}$ correspond to undirected decomposable graphical models, and not all undirected decomposable graphical models are available in a single ordering. However, for every decomposable undirected graphical model there does exist at least one ordering where the directed model with the same edge set is Markov equivalent. Therefore, in order to carry out model determination in situations where we have not fixed a specific ordering, the ordering is unclear or we are interested in undirected models, we must find a way to cover all possible undirected models.

If the Reversible Jump procedure described in Chapter 4 is carried out within a particular parameterisation, it is not possible to estimate posterior model probabilities for all undirected decomposable graphs. We overcome this by using an extra Reversible Jump step to move between orderings of the variables. Even though there is no change in the dimension of the parameter subspace, a Reversible Jump approach is required as the interpretation of parameters varies depending on the ordering. By using this approach, all possible undirected decomposable models may be considered.

5.2 A Transformation between Orderings

A move to a new parameterisation consists of proposing to switch two adjacent variables in the current ordering and we observe that any possible ordering of the variables may be reached in a finite number of such moves.

In general, if variables j and $j+1$ $(j = 1, \ldots, p-1)$ in the original ordering are switched to get the new ordering, the parameters in the model undergo a transformation defined as follows, where $'$ represents the new parameter.

- $z'_{ij} = z_{i,j+1}$, $z'_{i,j+1} = z_{ij}$ for $i = 1, \ldots, n$
- $\beta'_j = \beta_{j+1}$, $\beta'_{j+1} = \beta_j$
- $\theta(j,c)' = \theta(j+1,c)$, $\theta(j+1,c)' = \theta(j,c)$ for $c = 1, \ldots, k_j - 1$

The transformation for the decomposed covariance matrix $\mathbf{\Phi}$ is less straightforward. In terms of the covariance matrix $\mathbf{\Sigma}$, the transformation may be obtained by permuting the rows and columns corresponding to the variables to be switched. The corresponding transformations for elements of $\mathbf{\Phi}$ are

$$\begin{aligned}
\phi'_{jj} &= (\phi^2_{j,j+1} + \phi^2_{j+1,j+1})^{\frac{1}{2}} \\
\phi'_{j,j+1} &= \frac{\phi_{jj}\phi_{j,j+1}}{(\phi^2_{j,j+1} + \phi^2_{j+1,j+1})^{\frac{1}{2}}} \\
\phi'_{jk} &= \frac{\phi_{j,j+1}\phi_{jk} + \phi_{j+1,j+1}\phi_{j+1,k}}{(\phi^2_{j,j+1} + \phi^2_{j+1,j+1})^{\frac{1}{2}}} & \text{for} \quad k = j+2, \ldots, m \\
\phi'_{j+1,j+1} &= \frac{\phi_{jj}\phi_{j+1,j+1}}{(\phi^2_{j,j+1} + \phi^2_{j+1,j+1})^{\frac{1}{2}}} \\
\phi'_{j+1,k} &= \frac{\phi_{jk}\phi_{j+1,j+1} - \phi_{j,j+1}\phi_{j+1,k}}{(\phi^2_{j,j+1} + \phi^2_{j+1,j+1})^{\frac{1}{2}}} & \text{for} \quad k = j+2, \ldots, m
\end{aligned}$$

For $r = 1, \ldots, j-1$,

$$\begin{aligned}
\phi'_{rj} &= \phi_{r,j+1} \\
\phi'_{r,j+1} &= \phi_{rj} \\
\phi'_{rk} &= \phi_{rk} & \text{for} \quad k \neq j, j+1
\end{aligned}$$

For $r = j+2, \ldots, m$

$$\phi'_{rk} = \phi_{rk} \qquad \text{for} \quad k = j+2, \ldots, m$$

For a purely ordinal data set, that is one where all classifying variables are ordinal, this transformation is self-inverse. However, when one or more of the margins of the table are binary, the transformation must be adapted to allow for the fact that the conditional precision, ϕ^2_{ii} corresponding to a binary margin is constrained. The resulting transformation is then not self-inverse, but the inverse is easily found. The two cases to be considered are (i) switching two binary variables and (ii) switching a binary and an ordinal variable. For the case where the variables to be switched are binary with constraints $\phi_{jj} = k_j$, for switching variable j and $j+1$ ($j = 1, \ldots, p-1$) in the ordering, we arrive at the following transformation for $\mathbf{\Phi}$.

$$\begin{aligned}
\phi'_{jj} &= k_{j+1} \\
\phi'_{j,j+1} &= \frac{k_j\phi_{j,j+1}}{(\phi^2_{j,j+1}+k^2_{j+1})^{\frac{1}{2}}} \\
\phi'_{jk} &= \frac{\phi_{j,j+1}\phi_{jk}+k_{j+1}\phi_{j+1,k}}{(\phi^2_{j,j+1}+k^2_{j+1})^{\frac{1}{2}}} && \text{for} \quad k=j+2,\dots,m \\
\phi'_{j+1,j+1} &= k_j \\
\phi'_{j+1,k} &= \frac{\phi_{jk}k_{j+1}-\phi_{j,j+1}\phi_{j+1,k}}{(\phi^2_{j,j+1}+k^2_{j+1})^{\frac{1}{2}}} && \text{for} \quad k=j+2,\dots,m
\end{aligned}$$

The rest of the transformation is the same as for the purely ordinal case.

This transformation is no longer self-inverse but the inverse is easily found to be:

$$\begin{aligned}
\phi_{jj} &= k_j \\
\phi_{j,j+1} &= \frac{k_{j+1}\phi'_{j,j+1}}{(k^2_j-\phi'^2_{j,j+1})^{\frac{1}{2}}} \\
\phi_{jk} &= \frac{\phi'_{j,j+1}\phi'_{jk}+\phi'_{j+1,k}(k^2_j-\phi'^2_{j,j+1})^{\frac{1}{2}}}{k_j} && \text{for} \quad k=j+2,\dots,m \\
\phi_{j+1,j+1} &= k_{j+1} \\
\phi_{j+1,k} &= \frac{\phi_{jk}(k^2_j-\phi'^2_{j,j+1})^{\frac{1}{2}}-\phi'_{j,j+1}\phi'_{j+1,k}}{k_j} && \text{for} \quad k=j+2,\dots,m
\end{aligned}$$

To ensure reversibility, for any pair of variables A, B which can be involved in such a move, it needs to be specified in advance which switch corresponds to the forward move, and which to the reverse move. In practice, this choice does not have a great effect on the performance of the algorithm. Note that if $k^2_j-\phi'^2_{j,j+1}\leq 0$, the move is prohibited and hence automatically rejected.

The only remaining case to consider is when a move to a new ordering is made by switching an ordinal and a binary variable. We choose to specify the 'forward' move as the binary variable moving up the ordering while the ordinal variable moves down. By moving down the ordering, we mean that the variable moves further down the conditional structure so its distribution is modelled on one more variable. Assuming that the conditional variance corresponding to the binary variable is set to be k_j, the forward and reverse transformations are defined as follows:

The forward transformation is:

$$\begin{aligned}
\phi'_{jj} &= (\phi^2_{j,j+1} + \phi^2_{j+1,j+1})^{1/2} \\
\phi'_{j,j+1} &= \frac{k_j \phi_{j,j+1}}{(\phi^2_{j,j+1} + \phi^2_{j+1,j+1})^{1/2}} \\
\phi'_{jk} &= \frac{\phi_{j,j+1}\phi_{jk} + \phi_{j+1,j+1}\phi_{j+1,k}}{(\phi^2_{j,j+1} + \phi^2_{j+1,j+1})^{1/2}} && \text{for} \quad k = j+2, \ldots, m \\
\phi'_{j+1,j+1} &= k_j \\
\phi'_{j+1,k} &= \frac{\phi_{jk}\phi_{j+1,j+1} - \phi_{j,j+1}\phi_{j+1,k}}{(\phi^2_{j,j+1} + \phi^2_{j+1,j+1})^{1/2}} && \text{for} \quad k = j+2, \ldots, m
\end{aligned}$$

The reverse transformation is:

$$\begin{aligned}
\phi_{jj} &= k_j \\
\phi_{j,j+1} &= \frac{\phi'_{jj}\phi'_{j,j+1}}{k_j} \\
\phi_{jk} &= \frac{\phi'_{j,j+1}\phi'_{jk} + \phi'_{j+1,k}(k_j^2 - \phi'^2_{j,j+1})^{\frac{1}{2}}}{k_j} && \text{for} \quad k = j+2, \ldots, m \\
\phi_{j+1,j+1} &= \frac{\phi'_{jj}(k_j^2 - \phi'^2_{j,j+1})^{\frac{1}{2}}}{k_j} \\
\phi_{j+1,k} &= \frac{\phi_{jk}(k_j^2 - \phi'^2_{j,j+1})^{\frac{1}{2}} - \phi'_{j,j+1}\phi'_{j+1,k}}{k_j} && \text{for} \quad k = j+2, \ldots, m
\end{aligned}$$

If $k_j^2 - \phi'^2_{j,j+1} \leq 0$, the move is prohibited and automatically rejected. Again, all other transformations are the same as for the ordinal case.

At each stage of the RJMCMC, there are $p-1$ move types, each corresponding to the $p-1$ possible adjacent pairs of variables. Each move involves permuting an adjacent pair of variables, with the model parameters being transformed *via* the transformation g defined above. Suppose that at time t, the Markov chain is in order $o^{(t)}$, and that $\boldsymbol{\xi}^{(t)}_{o^{(t)}}$ represents the values of the unknown parameters in order $o^{(t)}$:

$$\boldsymbol{\xi}^{(t)}_{o^{(t)}} = \left(\boldsymbol{z}^{(t)}_{o^{(t)}}, \boldsymbol{\beta}^{(t)}_{o^{(t)}}, \boldsymbol{\Phi}^{(t)}_{o^{(t)}}, \boldsymbol{\theta}^{(t)}_{o^{(t)}}(j,c)\right)$$

A move to a new ordering o' is proposed, with $\boldsymbol{\xi}'_{o'} = g(\boldsymbol{\xi}^{(t)}_{o^{(t)}})$. Suppose that the probability of making move type r given the current state and model of the Markov chain $(o^{(t)}, \boldsymbol{\xi}^{(t)}_{o^{(t)}})$ is $j(r, \boldsymbol{\xi}^{(t)}_{o^{(t)}}, m^{(t)}_{o^{(t)}})$. Then the move should be accepted with probability α_{order} where

$$\alpha_{\text{order}} = \min\left\{1, \frac{f(\boldsymbol{\xi}'_{o'}, m'_{o'}, o'|\boldsymbol{y})j(r, \boldsymbol{\xi}'_{o'}, m'_{o'})}{f(\boldsymbol{\xi}^{(t)}_{o^{(t)}}, m^{(t)}_{o^{(t)}}, o^{(t)}|\boldsymbol{y})j(r, \boldsymbol{\xi}^{(t)}_{o^{(t)}}, m^{(t)}_{o^{(t)}})}\left|\frac{\partial(\boldsymbol{\xi}'_{o'})}{\partial(\boldsymbol{\xi}^{(t)}_{o^{(t)}})}\right|\right\} \tag{5.1}$$

Note that there is no proposal distribution because no new parameters are being proposed. We choose each move type to be equally likely so that the probabilities $j(.)$ cancel. Then applying Bayes' theorem (1.1) to the posterior distribution and simplifying gives (dropping suffixes):

$$\begin{aligned} f(\boldsymbol{\xi}, m, o|\boldsymbol{y}) &= f(\boldsymbol{z}, \boldsymbol{\Phi}, \boldsymbol{\beta}, \boldsymbol{\theta}, m, o|\boldsymbol{y}) \\ &= \frac{f(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\Phi}, \boldsymbol{\beta}, m, o)f(\boldsymbol{z}, \boldsymbol{\theta}|\boldsymbol{\Phi}, \boldsymbol{\beta}, m, o)f(\boldsymbol{\Phi}, \boldsymbol{\beta}, m, o)}{f(\boldsymbol{y})} \\ &= \frac{f(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{\theta})f(\boldsymbol{z}, \boldsymbol{\theta}|\boldsymbol{\Phi}, \boldsymbol{\beta}, m, o)f(\boldsymbol{\Phi}, \boldsymbol{\beta}|m, o)f(m, o)}{f(\boldsymbol{y})} \\ &= \frac{f(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{\theta})f(\boldsymbol{z}|\boldsymbol{\theta}, \boldsymbol{\Phi}, \boldsymbol{\beta}, m, o)f(\boldsymbol{\theta}|\boldsymbol{\Phi}, \boldsymbol{\beta}, m, o)f(\boldsymbol{\Phi}, \boldsymbol{\beta}|m, o)f(m, o)}{f(\boldsymbol{y})} \end{aligned} \tag{5.2}$$

The final step uses the fact that $\boldsymbol{y}$ is determined purely by $\boldsymbol{z}$ and $\boldsymbol{\theta}$. On substituting (5.2) into (5.1), $f(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{\theta})$, $f(\boldsymbol{\theta}|\boldsymbol{\Phi}, \boldsymbol{\beta}, m, o)$ and $f(\boldsymbol{y})$ will cancel in the numerator and the denominator because they are order independent. This leaves the acceptance probability:

$$\alpha_{\text{order}} = \min\left\{1, \frac{f(\boldsymbol{\Phi}'_{o'}, \boldsymbol{\beta}'_{o'}|o', m'_{o'})f(o', m'_{o'})}{f(\boldsymbol{\Phi}^{(t)}_{o^{(t)}}, \boldsymbol{\beta}^{(t)}_{o^{(t)}}|o^{(t)}, m^{(t)}_{o^{(t)}})f(o^{(t)}, m^{(t)}_{o^{(t)}})}\left|\frac{\partial(\boldsymbol{\xi}'_{o'})}{\partial(\boldsymbol{\xi}^{(t)}_{o^{(t)}})}\right|\right\} \tag{5.3}$$

The prior term $f(\boldsymbol{\Phi}, \boldsymbol{\beta}|o', m)$ may be decomposed as follows:

$$f(\boldsymbol{\Phi}_o, \boldsymbol{\beta}_o|o, m) = f(\boldsymbol{\beta}|o, m)\prod_{i=1}^{p} f(\phi_{ii}|o, m)f(\boldsymbol{\phi}_i|\phi_{ii}, o, m) \tag{5.4}$$

The priors for $\boldsymbol{\beta}$ are chosen to be independent of the model, therefore in the expression for α_{order} these will cancel out, leaving only the priors for the inverse variance decomposition matrix $\boldsymbol{\Phi}$ to consider. These priors are given in (3.8) and (3.9), for the unconstrained ϕ_{ii}.

The prior matrix $\boldsymbol{A}$ is diagonal with entries dependent on the number of levels of each classifying variable. Since the off-diagonal elements of $\boldsymbol{A}$ are 0, (3.8) reduces to

$$\boldsymbol{\phi}_i|\phi_{ii} \sim N_{p-i}(\mathbf{0}, \boldsymbol{A}_i^{-1})$$

When any two classifying variables are switched in the ordering, the corresponding elements of $\boldsymbol{A}$ must also be switched and new values for $\boldsymbol{\phi}_i$ are proposed. Therefore, the prior terms for those rows affected must be included in α_{order}. The prior for the diagonal elements ϕ_{ii} reduces to

$$\phi_{ii} \sim \sqrt{\frac{1}{a_{ii}}\chi^2_{q-i+1}}$$

These are order dependent as when an order change is proposed, new values for ϕ_{ii} are proposed so these terms must be included in α_{order}

Finally, we consider the model and order prior $f(m, o)$. When moving between models within a single ordering of the variables as described in Chapter 4, each possible model is assumed to be *a priori* equally likely. However not all undirected graphical models are representable in each ordering and there are a different number of orderings available for each model. Therefore, when moving between orderings, we need to weight competing orderings accordingly, *via* the prior term $f(m, o)$. Thus for a particular model and ordering, $f(m, o)$ is the reciprocal of the number of orderings in which the conditional independence structure of the model m can occur.

For a purely ordinal data set, the Jacobian of the transformation of $\boldsymbol{\Phi}$ is:

$$|J| = \left|\frac{\partial(\boldsymbol{\xi}'_{o'})}{\partial(\boldsymbol{\xi}^{(t)}_{o^{(t)}})}\right| = \frac{\phi_{jj}}{(\phi^2_{j,j+1} + \phi^2_{j+1,j+1})^{\frac{1}{2}}} \tag{5.5}$$

When one or more of the margins of the contingency table are binary, the Jacobian is altered. This is because the corresponding conditional variances are constant so the Jacobian is not a function of these. For the all binary case, the Jacobian of the transformation is:

$$|J| = \left|\frac{\partial(\xi'_{o'})}{\partial(\xi^{(t)}_{o(t)})}\right| = \frac{k_j k_{j+1}^2}{(\phi_{j,j+1}^2 + k_{j+1}^2)^{\frac{3}{2}}} \tag{5.6}$$

and that of the reverse move is:

$$|J| = \left|\frac{\partial(\xi'_{o'})}{\partial(\xi^{(t)}_{o(t)})}\right| = \frac{k_j^2 k_{j+1}}{(k_j^2 - \phi'^2_{j,j+1})^{\frac{3}{2}}} \tag{5.7}$$

For the binary and ordinal switch, the Jacobian of the forward move is:

$$|J| = \left|\frac{\partial(\xi'_{o'})}{\partial(\xi^{(t)}_{o(t)})}\right| = \frac{k\phi_{j+1,j+1}}{\phi_{j,j+1}^2 + \phi_{j+1,j+1}^2} \tag{5.8}$$

and that of the reverse move is:

$$|J| = \left|\frac{\partial(\xi'_{o'})}{\partial(\xi^{(t)}_{o(t)})}\right| = \frac{\phi'_{jj}}{(k^2 - \phi'^2_{j,j+1})^{\frac{1}{2}}} \tag{5.9}$$

The full method now runs as follows.

5.2.1 Algorithm 6

1. An initial ordering, model and values for all parameters in this model are specified.

2. With probability 1/3, remain in the current model and re-generate all parameters (the null move). Otherwise, a new model is proposed by either adding or subtracting a randomly selected edge from the current model. If the proposed model is unavailable in the current ordering, go to Step 3. Otherwise the proposed model is accepted with Reversible Jump probability α as defined in (4.11) and (4.12).

3. A new ordering is proposed by randomly selecting a pair of adjacent variables in the current ordering and proposing to switch them. If the current model is not available in the proposed ordering, then go to step 2. Otherwise, the proposed ordering is accepted with Reversible Jump probability α_{order} as defined in (5.3).

4. Return to step 2 and repeat.

Note that this algorithm attempts an ordering switch at every iteration. This is not necessary but was found to be most efficient.

5.3 Examples

5.3.1 Alcohol, Obesity and Hypertension Data

Algorithm 6 was applied to the Knuiman and Speed (1988) data set. For reasons described in Section 4.2, the conditional variance for H was fixed to be 1. Starting the RJMCMC in order OHA, the initial prior parameters were $\boldsymbol{A} = \text{diag}(0.185, 1, 0.455)$, $q = 5$ and $\boldsymbol{T} = \text{diag}(50)$. The prior matrix $\boldsymbol{A}$ must be permuted in the same way as the classifying variables at each order change, in order that the priors for the model parameters are consistent across models. Since the matrix $\boldsymbol{T}$ is order invariant, this need not be permuted at each order change. The algorithm was run for 500,000 iterations. During this time, the proposed order change move was accepted approximately 50% of the time and the proposed model change move was accepted approximately 10% of the time. The Markov chain was therefore relatively mobile. Posterior model probabilities are displayed in Table 5.3.1. The most probable model is $OH + AH$. There is high posterior probability of conditional independence of Obesity and Alcohol Intake as this interaction appears in none of the most probable models. There is also fairly high posterior probability of an association between Hypertension and Obesity, as this appears in the two most probable models.

Model	Posterior Model Probability
$OH + AH$	0.262
$A + OH$	0.171
$O + AH$	0.122
$A + O + H$	0.119
$OA + OH$	0.119
AOH	0.075
$AH + AO$	0.069
$H + OA$	0.062

Table 5.1: Posterior model probabilities for alcohol, obesity and hypertension data

As has been noted for other examples, the posterior probability is fairly diffuse over the model space with even the least likely model claiming 6.2% of the posterior model probability.

The data set was also modelled by Dellaportas and Forster (1999), who compared the results from various methods, none of which took account of the ordinal structure of the data. The methods used were a Reverisble jump procedure described in the paper, an exact hyper-Dirichlet prior approach suggested by Madigan and Raftery (1994) and the approximate Bayes factor approach described by Raftery (1996). Various prior parameters were used. For all methods used, they found the four most probable models to be (in varying order according to the method), $OH+AH$, $O+AH$, $A+OH$ and $A+O+H$. These are also the four most probable models selected by the RJMCMC method described here. Their results differ from those here in that they are far less diffuse over the model space, and for most of the methods described, one or two models claim a very high percentage of the posterior model probability. On average, the two most probable models were found to be $A + HO$ and $H + O + A$, indicating that the method described here favours more complex models in terms of independence structure than the methods described by Dellaportas and Forster. However note that the dependence structure is modelled more parsimoniously here, because each edge has just a single corresponding

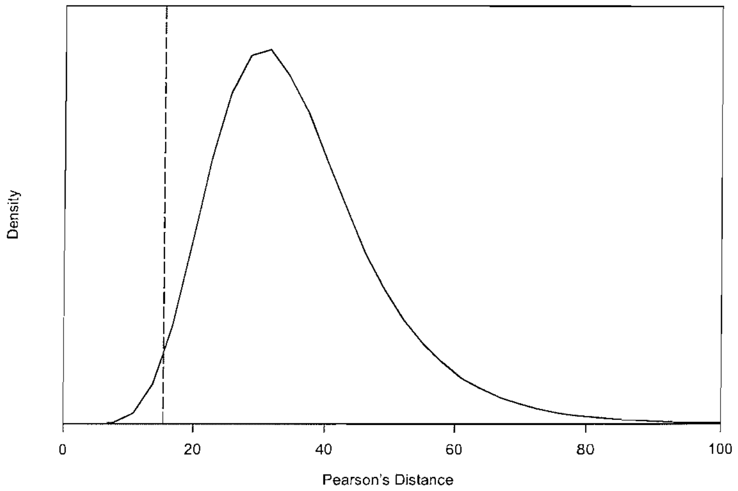


Figure 5.4: Estimated density of Pearson's distance measure for hypertension, alcohol intake and obesity data

model parameter.

We now examine the goodness-of-fit of the models, averaging over all models produced during the Reversible Jump. This procedure was carried out by converting all output from the Reversible Jump back to the initial ordering. The means, variances and cut points produced at each iteration were then used to generate predictive tables for each iteration in the usual manner. Figures 5.4, 5.5 and 5.6 show the densities of the Pearson's, Deviance and Absolute Difference distance measures respectively. As usual, the vertical line indicates the distance between the observed and posterior predictive mean data, and gives a measure of how likely the set of models are to predict the original data.

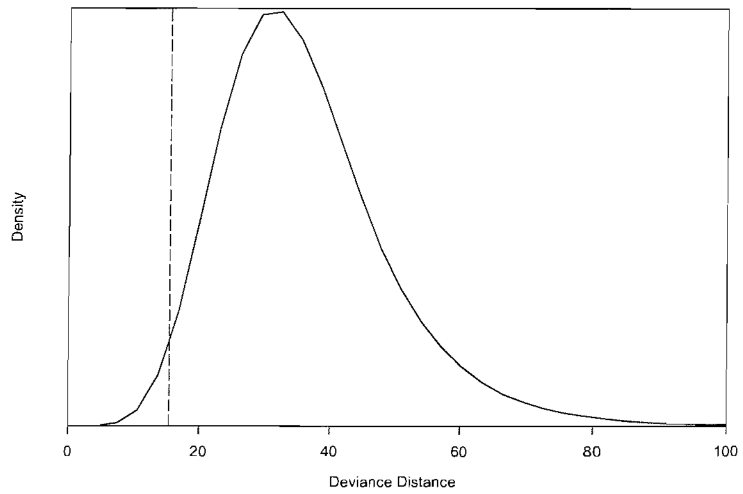


Figure 5.5: Estimated density of deviance distance measure for hypertension, alcohol intake and obesity data

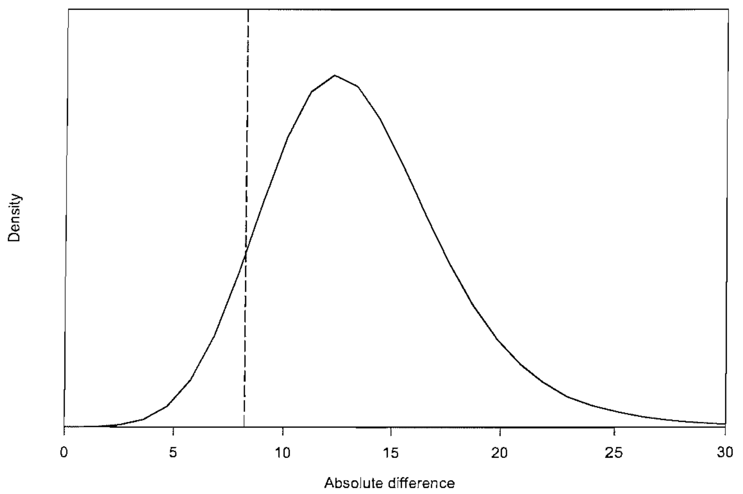


Figure 5.6: Estimated density of maximum absolute difference distance measure for hypertension, alcohol intake and obesity data

For each distance measure the fit appears to be good, with the posterior predictive mean data well into the lower tail of the distribution of generated tables. As would be expected, the set of models passed through during the RJMCMC do not provide as good a fit as the saturated model (see Figures 3.14, 3.15 and 3.16) but they improve greatly on the single model example (Figure 3.17). Convergence of the RJMCMC was assessed using trace plots of quantities which have a common interpretation across models (again produced by converting the output back to the initial ordering) and was found to be satisfactory. The Reversible Jump procedure was found to be relatively mobile with approximately one proposed move out of every six accepted for the model change step and greater than one in three proposed moves accepted for the order change step.

5.3.2 Risk factors for coronary heart disease: a 2^6 table

For the second example we return to the coronary heart disease data taken from Edwards and Havranek (1985) and displayed in Section 4.5. In Chapter 4, the classifying variables were assumed to take the order $FCBAED$ and the posterior probabilities of DAG models were calculated using Algorithm 5. However this order is by no means clear so here we assume no order and consider the set of undirected decomposable graphical models. All six classifying variables are binary so the conditional precision ϕ^2_{ii} for each is fixed to be 1. Following the arguments of Section 4.2, the matrix $\boldsymbol{A}$ is the identity matrix, the degrees of freedom parameter $q = 8$ and the mean dispersion matrix $\boldsymbol{T} = \text{diag}(50)$. Since the matrix $\boldsymbol{A}$ is order invariant, it need not be permuted at each order change. Starting in the order $ABCDEF$, the RJMCMC procedure described in Algorithm 6 was applied with 500,000 iterations. Again the procedure was relatively mobile with order changes accepted approximately every one in two proposed moves and model changes accepted approximately every one in twelve moves. The six most probable models along with their posterior model probabilities are shown in Table 5.3.2.

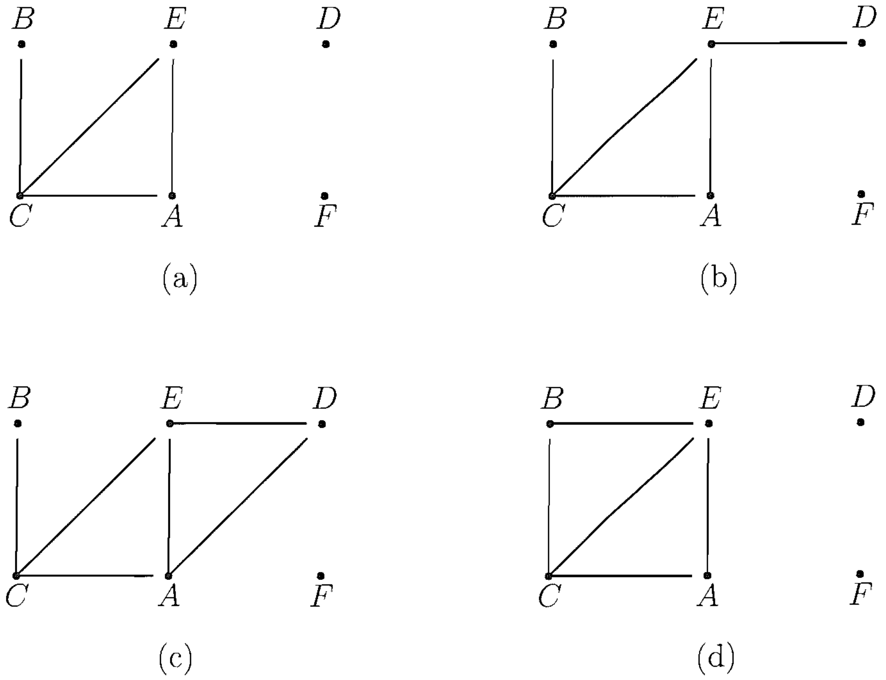
Model	Posterior Model Probability
$ACE + BC + F$	0.027
$ACE + BC + DE + F$	0.024
$ACE + ADE + BC + F$	0.018
$ACE + BCE + F$	0.016
$ACE + BCE + DE + F$	0.015
$ACE + ADE + BCE + F$	0.014

Table 5.2: Posterior model probabilities for coronary heart disease data

Between them, these models account for 11.4% of the posterior model probability. When this data set was previously analysed in Chapter 4, the first four models accounted for 36.1% of the posterior model probability. Clearly, the order change and subsequent greater choice of models available is the main reason for this. The posterior distribution amongst the most probable models seems much more diffuse over the set of undirected models than over the set of directed models.

There is strong evidence for the marginal independence of F. There is also high posterior probability of AC,BC,AE and CE interactions as these appear in each of the four most probable models. There is some evidence for interactions AD, BE and DE.

The four most probable models are illustrated in Figure 5.7. The data were also analysed by Edwards and Havranek (1985), Madigan and Raftery (1994) and Dellaportas and Forster (1999). Madigan and Raftery give the two most probable decomposable models as $BC+ACE+ADE+F$ and $ABC+ABE+ADE+F$. The most probable models found by Edwards and Havranek only contained one decomposable model, that was $BC+ACE+ADE+F$. Dellaportas and Forster give the three most probable models as $BC+ACE+ADE+F$, $BC+ACE+DE+F$ and $BC+AD+ACE+F$. The model $BC+ACE+ADE+F$ is found to be one of the most probable models by all three; this is the third most probable model found using the RJMCMC method described here. The main difference between the sets of models selected by



(a)

(b)

(c)

(d)

Figure 5.7: Most probable undirected models for the heart disease data

others and the set of models selected here is that the RJMCMC gives smaller posterior probability that the AD term should be included in the model. However, in the analysis carried out by Edwards and Havranek, the exact test for zero partial associateion of A and D reported had a significance level of 0.04, which was the largest of any of the links whose absence was rejected at the 5% level. There is therefore some precedent for doubt about an association between A and D.

The goodness-of-fit of the models selected by the Reversible Jump procedure was assessed using the simulation technique described in section 3.4.2.

Convergence was also assessed with trace plots and was found to be satisfactory.

Chapter 6

Model Determination for Data with Covariates and Covariate Selection

In Chaper 2, the method of Albert and Chib (1993) was applied to univariate ordinal data with covariates. For the multivariate case we have so far only considered data sets with no covariates. In this Chapter we firstly describe how data sets involving covariates may be modelled and secondly show that the issue of covariate selection may also be tackled within the same framework, using a Reversible Jump method.

6.1 A Model for Data with Covariates

For the non-covariate case - where individuals are simply cross-classified by p variables, with variable j having k_j levels - the structure of the underlying latent data is $p-$variate Normal with mean $\boldsymbol{\beta}$ and variance matrix Σ. Now suppose that for individual i, we observe multivariate ordinal response vector $\boldsymbol{y}_i$ and also associated covariate matrix $\boldsymbol{X}_i$. $\boldsymbol{X}_i$ is a matrix with p rows where p is the dimension of the mutivariate response, with the jth row representing the values of the C covariates associated with respondent i and classifying variable j. In practice, the covariates often take the same values for every classifying variable, so that the rows of $\boldsymbol{X}_i$ are identical. The situation where this is not true occurs most often for longitudinal data, where the classifying variables represent the values of the same response variable at different time points and, depending on the application, covariates may be measured at each time point. There is an example of this later in the Chapter, in the crossover trial example. Here, for each individual, one of the covariates (treatment) varies over each time point.

Covariates may be coded in different ways depending on the form of the covariate information. If a covariate is continuous, then it supplies information for one column of the matrix $\boldsymbol{X}_i$. If it is a factor with two categories (e.g. male or female) then one of these categories can be chosen to be the baseline category. Such a covariate then supplies one column of the matrix $\boldsymbol{X}_i$ taking the value 1 for non-baseline category and 0 for baseline category. Finally there is the situation where a covariate is a factor with more than f_c categories where $f_c > 2$. In this situation, the covariate supplies f_c columns to the covariate matrix, each corresponding to one of the factor levels. As in the binary factor case, one of the factor levels may be set to 0, so that the remaining $f_c - 1$ levels are thought of as contrasts and provide $f_c - 1$ columns corresponding to covariate c in the covariate matrix.

Suppose that covariate c contributed f_c columns to $\boldsymbol{X}_i$. Then we have covariate matrix

$$\boldsymbol{X}_i = \begin{pmatrix} \boldsymbol{x}_{i1} \\ \boldsymbol{x}_{i2} \\ \vdots \\ \boldsymbol{x}_{ip} \end{pmatrix}$$

where each vector $\boldsymbol{x}_{ij}$ is of length $L = \sum f_c + p$, and associated common parameter vector $\boldsymbol{\beta}$, also of length L.

From here, the model is constructed following the same strategy as before, except that the distribution of the latent data is now dependent on the covariates. Let $\boldsymbol{\Sigma} = (\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}$ be the usual covariance matrix between the p

classifying variables. Then, the existence of a latent variable $\boldsymbol{z}_i$ is assumed, with the distribution

$$\boldsymbol{z}_i \sim N_p(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}) \tag{6.1}$$

The usual assumptions are made about the cut points and necessary constraints on cut points (or conditional variances in the binary case) for identifiability. That is,

$$\theta(j,1) = -1 \qquad\qquad \theta(j,k_j-1) = 1$$

and $\phi_{jj}^2 = 1$ for binary classifying variables.

Clearly the covariates must now be incorporated into the posterior distribution of the unknown parameters $\boldsymbol{\beta}$, $\boldsymbol{\Phi}$, $\boldsymbol{z}_i$ and $\boldsymbol{\theta}$. The usual prior for $\boldsymbol{\beta}$ is extended so that it is now a L-variate Normal distribution with mean $\boldsymbol{0}$ and covariance matrix $\boldsymbol{T} = \lambda \boldsymbol{I}_L$, where λ is large. Other priors are unaffected and the remaining effect of the covariates is through the likelihood. The resulting conditional posterior distributions for use in the Gibbs sampler for parameter estimation in the saturated model are as follows.

- $\boldsymbol{z}_i|\boldsymbol{y}_i, \boldsymbol{\beta}, \boldsymbol{\Phi}, \boldsymbol{\theta} \sim N_p\left(\boldsymbol{X}_i\boldsymbol{\beta}, (\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\right)$ with z_{ij} truncated to the interval $(\theta(j, y_{ij}-1), \theta(j, y_{ij}))$
- $$\boldsymbol{\beta}|\boldsymbol{z}, \boldsymbol{\Phi} \sim N_{p+\sum f_c}\left(\left(\sum_{i=1}^{n} \boldsymbol{X}_i^T\boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{X}_i + \boldsymbol{T}^{-1}\right)^{-1} \sum_{i=1}^{n} \boldsymbol{X}_i^T\boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{z}_i \ , \left(\sum_{i=1}^{n} \boldsymbol{X}_i^T\boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{X}_i + \boldsymbol{T}^{-1}\right)^{-1}\right)$$
- $\theta(j,c)|\boldsymbol{z} \sim \text{Unif}\left(\max_i[z_{ij}(y_{ij}=c)], \min_i[z_{ij}(y_{ij}=c+1)]\right)$
- $\boldsymbol{\psi}_i|\phi_{ii}^2, \boldsymbol{z}, \boldsymbol{\beta} \sim N_{p-i}\left((\boldsymbol{\mu}_i\boldsymbol{A}_i - \boldsymbol{g}_i)(\boldsymbol{A}_i + \boldsymbol{G}_i)^{-1}, \frac{1}{\phi_{ii}^2}(\boldsymbol{A}_i + \boldsymbol{G}_i)^{-1}\right)$
- $\phi_{ii}^2|\boldsymbol{\psi}_i, \boldsymbol{z}, \boldsymbol{\beta} \sim \text{Gamma}\left(\gamma, \frac{\delta}{2}\right)$ where γ and δ are given in (3.13) and (3.14) respectively.

In the ϕ_{ii} and $\boldsymbol{\psi}_i$ generation steps, the matrix G is now dependent on the covariates:

$$\boldsymbol{G} = \sum_{i=1}^{n} (\boldsymbol{z}_i - \boldsymbol{X}_i\boldsymbol{\beta})(\boldsymbol{z}_i - \boldsymbol{X}_i\boldsymbol{\beta})^T$$

Starting with initial values for all parameters, the sampling scheme runs by sampling iteratively from the conditional posterior distributions in the order $[\boldsymbol{\beta}|\boldsymbol{z}, \boldsymbol{\Phi}]$, $[\phi_{11}^2|\boldsymbol{\psi}_1, \boldsymbol{z}, \boldsymbol{\beta}]$,. . . , $[\phi_{pp}^2|\boldsymbol{\psi}_p, \boldsymbol{z}, \boldsymbol{\beta}]$, $[\boldsymbol{\psi}_1|\phi_{11}^2, \boldsymbol{z}, \boldsymbol{\beta}]$,. . . , $[\boldsymbol{\psi}_p|\phi_{pp}^2, \boldsymbol{z}, \boldsymbol{\beta}]$, $[\boldsymbol{z}_1|\boldsymbol{y}_1, \boldsymbol{\beta}, \boldsymbol{\Phi}, \boldsymbol{\theta}]$, . . . , $[\boldsymbol{z}_n|\boldsymbol{y}_n, \boldsymbol{\beta}, \boldsymbol{\Phi}, \boldsymbol{\theta}]$ and $[\boldsymbol{\theta}|\boldsymbol{z}]$.

6.2 Model Choice with Covariates

We have described a method for estimating the unknown parameters in the model with covariates. For data with covariates there are two kinds of model choice. The first is the model choice that has been discussed in Chapters 4 and 5 and involves investigating the relationship between classifying variables as characterised by the matrix $\boldsymbol{\Phi}$. The second is to consider which covariates should be included in a model to predict the data. In this section, we will adapt the methods described in Chapters 4 and 5 so that they may be applied to data with covariates and give examples. In the next section, we will discuss the second type of model choice, that of covariate selection. Note that the model choice discussed in Chapters 4 and 5 relates to modelling the covariance structure whereas the type of model choice discussed later in this Chapter corresponds to modelling mean structure.

The method used for model choice is a simple adaptation of the reversible jump algorithms described in Chapters 4 (if we are only interested in directed graphical models or there is a fixed ordering to the data) and 5 (if there is no fixed ordering and we are considering the set of all undirected decomposable graphical models). Let us first consider the acceptance probability α defined in 4.11 and 4.12 for moves between models. There are three terms in this acceptance probability: the likelihood of the data given the parameters $f(\boldsymbol{z}_m|\boldsymbol{\Phi}_m, \boldsymbol{\beta}_m)$, the prior distributions $f(\boldsymbol{\Phi}_m, \boldsymbol{\beta}_m|m)$ and the proposal distribution $q_p(\boldsymbol{\Phi}_m)$. Each of these must be evaluated for both the current and the proposed model. The likelihood term is now a product of normal densities with individual means $\boldsymbol{X}_i\boldsymbol{\beta}$ and common variance $(\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}$. Both the prior distributions and the proposal distribution are unaffected by the presence of covariates, except that in the proposal, $\boldsymbol{G}$ now depends on $\boldsymbol{X}_i$.

Now let us consider the acceptance probability α_{order} defined in 5.3 for moves between variable orderings. This consists of the prior terms $f(\boldsymbol{\Phi}, \boldsymbol{\beta}|o, m)$ and $f(o, m)$ and a Jacobian. None of these are unaffected by the presence of covariates. However, if an order change is accepted, the covariate order should also be changed to reflect this, but only if the data set is one in which covariates vary with classifying variables.

We now apply this adaptation of Algorithms 5 and 6 to two data sets.

6.2.1 Examples

Example 1 - Crossover Trial

The first example is taken from Jones and Kenward (2003) and concerns a cross-over trial for pain relief for 86 (n=86) patients suffering from primary dysmenorrhea. The trial is a three treatment, three period cross over trial with ordinal response. The three treatments were A (placebo), B (low dose analgesic) and C (high dose analgesic), and there were three periods 1, 2 and 3. The 86 patients were randomised to each of the 6 possible treatment sequences. At the end of each treatment period, each subject rated the degree of pain relief as: none (1), moderate (2) and complete (3), thus providing a trivariate response ($p = 3$) with each classifying covariate taking three levels ($k_j = 3$ for all j). The resulting data are reproduced in Table 6.1.

			Sequence Group				
Response	ABC	ACB	BAC	BCA	CAB	CBA	Total
(1,1,1)	0	2	0	0	3	1	6
(1,1,2)	1	0	0	1	0	0	2
(1,1,3)	1	0	1	0	0	0	2
(1,2,1)	2	0	0	0	0	0	2
(1,2,2)	3	0	1	0	0	0	4
(1,2,3)	4	3	1	0	2	0	10
(1,3,1)	0	0	1	1	0	0	2
(1,3,2)	0	2	0	0	0	0	2
(1,3,3)	2	4	1	0	0	1	8
(2,1,1)	0	1	1	0	0	3	5
(2,1,2)	0	0	2	0	1	1	4
(2,1,3)	0	0	1	0	0	0	1
(2,2,1)	1	0	0	6	1	1	9
(2,2,2)	0	2	1	0	0	0	3
(2,2,3)	1	0	0	0	0	0	1
(2,3,1)	0	0	0	1	0	2	3
(2,3,2)	0	0	0	0	0	0	0
(2,3,3)	0	2	0	0	1	0	3
(3,1,1)	0	0	0	1	0	2	3
(3,1,2)	0	0	2	0	2	1	5
(3,1,3)	0	0	3	0	4	1	8
(3,2,1)	0	0	0	1	0	0	1
(3,2,2)	0	0	0	1	0	0	1
(3,2,3)	0	0	0	0	0	0	0
(3,3,1)	0	0	0	0	0	1	1
(3,3,2)	0	0	0	0	0	0	0
(3,3,3)	0	0	0	0	0	0	0
Total	15	16	15	12	14	14	86

Table 6.1: Cross over trial for pain relief

There are two covariates for this data set ($C = 2$): the treatment and the period. Each has three levels ($f_1 = f_2 = 3$). Note that this is a situation where for each respondent, the values of the covariates vary by period. The three periods also provide the three classifying variables for each response. Due to the order upon which variables are conditioned in the Cholesky parameterisation, we must take them into the covariance matrix in reverse time order; thus classifying variable 1 = Period 3, classifying variable 2 = Period 2 and classifying variable 3 = Period 1. For example, under this ordering, an individual in the first column, second row of the Table 6.1 has response $\boldsymbol{y}_i = (1, 1, 1)$ while an individual in the sixth column, first row of Table 6.1 has response vector $\boldsymbol{y}_i = (3, 2, 1)$. These two individuals would have covariate matrices:

$$\boldsymbol{x}_i = \left(\begin{array}{ccc|ccc} 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \end{array}\right)$$

and

$$\boldsymbol{x}_i = \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{array}\right)$$

respectively. The left-hand partition of the covariate matrix $\boldsymbol{x}_i$ corresponds to the treatment effects while the right-hand partition corresponds to the period effects. Note that we assume a common treatment effect across periods.

The response is trivariate ordinal with three categories in each period. Hence the first two cut points in each dimension are constrained to be -1 and 1 respectively and there are no free cut points to be estimated.

The data were first modelled using the saturated model. The Gibbs sampler was implemented using 100,000 iterations. The posterior means along with their standard deviations were obtained for $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$.

$$E(\boldsymbol{\beta}|\boldsymbol{y}) = \begin{pmatrix} -2.035(0.130) \\ 0.45(0.129) \\ 0.860(0.130) \\ 0.149(0.129) \\ -0.509(0.131) \\ -0.412(0.130) \end{pmatrix}$$

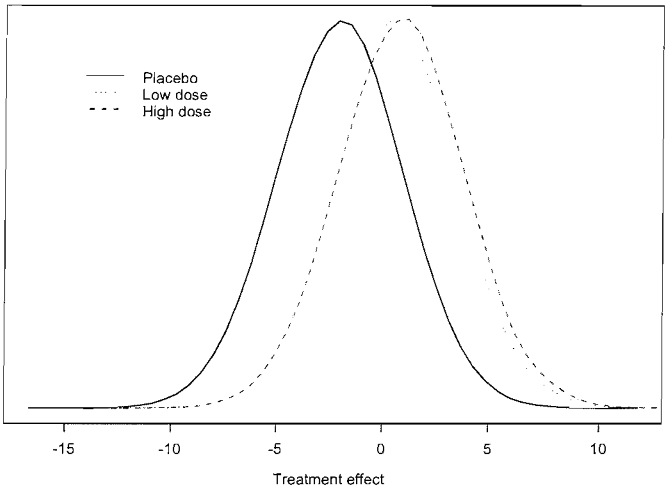


Figure 6.1: Posterior densities of the three treatment effects for the crossover trial data

$$E(\Sigma|\boldsymbol{y}) = \begin{pmatrix} 4.973(0.054) & 0.299(0.163) & 0.019(0.161) \\ 0.299(0.164) & 6.31(0.070) & -1.137(0.159) \\ 0.019(0.161) & -1.137(0.159) & 3.324(0.092) \end{pmatrix}$$

The posterior means for the covariate effects take the order: Treatment A (placebo), treatment B (low dose), treatment C (high dose), period 3, period 2 and period 1. The higher the response, the more effective the pain relief. Thus we see that as would be expected, both treatments show improved pain relief over the placebo, with the higher dose being slightly more effective than the lower dose, although the difference between the two is small when compared with the difference between each and the placebo. This is illustrated in Figure 6.1 which shows the posterior densities of the three treatment effects: Placebo, Low dose and High dose.

These results agree with those found by Jones and Kenward (2003) who find that there is an overwhelming effect of active treatment over placebo, but that there is a smaller difference between active treatment effects. The period effect is less clear, period 3 shows improved pain relief over periods 1 and 2, which

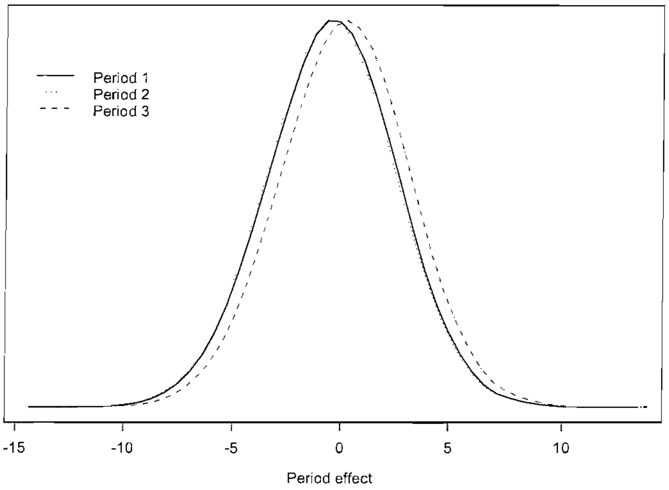


Figure 6.2: Posterior densities of the three period effects for the crossover trial data

are approximately equal. This is illustrated in Figure 6.2 which shows the posterior densities of the three period effects.

Note that the standard deviations for the mean estimates are approximately equal. This is possibly due to the fact that all three of the factor levels of the two factors period and treatment are included in the covariate matrix.

The posterior means for Σ indicate that there is little correlation between periods as the majority of the off-diagonals are fairly close to zero. The one exception is the covariance between Periods 1 and 2. To gain further insight into this, we calculate the correlation matrix $\boldsymbol{R}$:

$$\boldsymbol{R} = \begin{pmatrix} 1 & 0.053 & 0.005 \\ 0.053 & 1 & -0.248 \\ 0.005 & -0.248 & 1 \end{pmatrix}$$

The correlations between Periods 1 and 3 and Periods 2 and 3 are very close to zero, with that for Periods 1 and 3 being smaller than that for Periods 2 and

3 as we would expect. The correlation between Periods 1 and 2 is stronger, so there may be some probability of a relationship between these Periods. This will be investigated using the RJMCMC model choice algorithm.

To investigate this further, the issue of model choice was considered. Note that model search is carried out within the set of directed decomposable graphical models, as the data is longitudinal. Therefore, there is no need to move between orderings. The Reversible Jump algorithm (Algorithm 5) from Chapter 4 with adaptations described above was applied with 100,000 iterations. The prior parameters were chosen to be noninformative following the arguments of section 4.2: $\boldsymbol{A} = \text{diag}(0.185, 0.185, 0.185)$, $q = 5$ and $\boldsymbol{T} = \text{diag}(50)$. Table 6.2.1 shows the posterior model probabilities for each model.

Model	Posterior Model Probability
Period1+Period2+Period3	0.110
Period2:Period3 + Period1	0.049
Period1:Period3 + Period2	0.026
Period1:Period2 + Period3	0.470
Period1:Period3 + Period2:Period3	0.016
Period1:Period2 + Period2:Period3	0.166
Period1:Period2 + Period1:Period3	0.10
Period1:Period2:Period3	0.061

Table 6.2: Posterior model probabilities for crossover data

The most popular model is Period1:Period2 + Period 3, followed by Period1:Period2 + Period2:Period3 and then the null model Period1 + Period2 + Period3. Together, these three models account for 74.6% of the posterior probability. There is therefore high posterior probability of a relationship between responses in Periods 1 and 2, and some posterior probability of a relationship between responses in Periods 2 and 3. As we would expect, there is strong evidence that responses in Periods 1 and 3 are conditionally independent given the response in Period 2. The fact that the independence model is the third most popular model agrees with the finding that the treatment received in

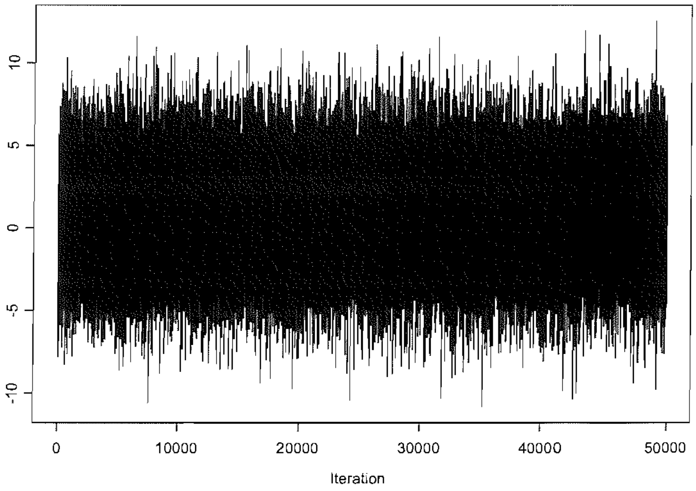


Figure 6.3: Trace plot for β_3 for crossover data

each period has an effect on the response, because otherwise we would expect greater correlation between the responses for an individual over each period.

It is difficult to assess the goodness-of-fit for this data set as the number of observations is small compared to the number of cells into which they are classified. Convergence of the MCMC was assessed using trace plots and autocorrelation function plots. The trace plots for β_3 (Treatment=high dose) and Σ_{11} (Period 3) are shown in Figures 6.3 and 6.4 respectively, while the autocorrelation function plot for Σ_{13} is shown in Figure 6.5.

Example 2 - Shoulder Tip Pain

The second example is taken from Lumley (1996) and concerns data from a randomised trial of abdominal suction to reduce shoulder tip pain after laparoscopic surgery. Forty-one patients were asked to rate their shoulder pain on a scale of 1 (low) to 5 (high) at six separate time points. The response therefore

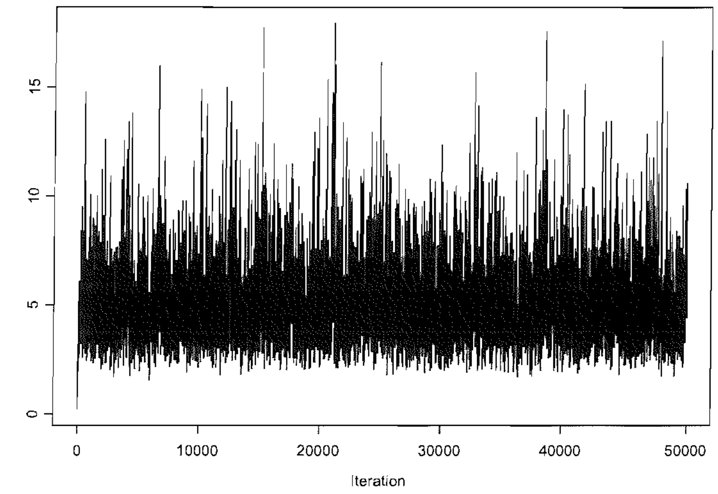


Figure 6.4: Trace plot for Σ_{11} for crossover data

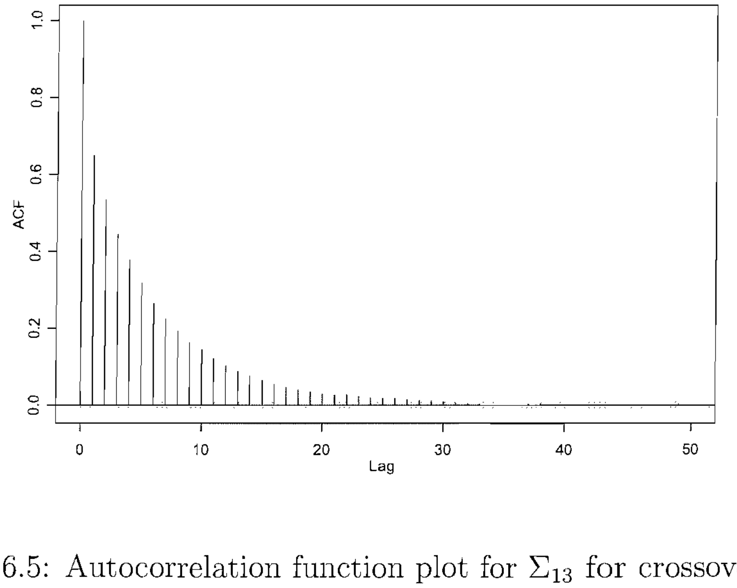


Figure 6.5: Autocorrelation function plot for Σ_{13} for crossover data

takes the form of a six-dimensional multivariate ordinal vector with each classifying variable having 5 levels ($p = 6$ and $k_j = 5$ for all j). There were four covariates ($C = 4$): Treatment (Yes or No, $f_1 = 1$), Sex ($f_2 = 1$), Age in years ($f_3 = 1$) and Time ($f_4 = 6$). Note that the binary factors only supply one column each to the covariate matrix as one level for each was chosen to be the baseline. For Sex, the baseline was chosen to be Female while for Treatment, No was chosen to be the baseline category. Therefore for each i, $\boldsymbol{X}_i$ takes the form of a 6×9 matrix.

For this data set, in contrast with the crossover trial data, the covariates are constant over each time point, so each row of $\boldsymbol{X}_i$ is the same. The data set is longitudinal; therefore, as for the crossover example, times are taken into the covariance matrix in reverse order, in order that models with directed edges having the correct edge direction. As each variable has five levels, there are two free cut points to be estimated in each dimension.

In Chapter 3, the Gibbs sampler algorithm for parameter estimation was implemented with the constraints that the first two cut points in each dimension were set to be 0 and 1 respectively. Then in Chapter 4, the new constraints $\theta(j,1) = -1$ and $\theta(j,k_j - 1) = 1$ were chosen to give symmetry to the model and thus facilitate prior parameter selection. The shoulder pain data example gives further support for this choice of parameter constraints. Using the constraints $\theta(j,1) = 0$ and $\theta(j,2) = 1$ leads to the estimate of 508,179 for the highest cut point for classifying variable Time 6, with associated standard deviation 3,829. This result is due to the fact that no patients fall into the highest category (level) for Time 6 so there is no latent data to constrain the highest cut point. Such a high posterior mean estimate for this parameter also has a strong influence on the latent data and hence the posterior mean $\boldsymbol{\beta}$ and in particular the variance $\boldsymbol{\Sigma}$. By using the constraints $\theta(j,1) = -1$ and $\theta(j,k_j - 1) = 1$ the problem is avoided and convergence unaffected.

The Gibbs sampler for saturated model parameter estimation was run for 20,000 iterations and the posterior means along with their standard deviations were obtained for $\boldsymbol{\beta}$, Σ and $\boldsymbol{\theta}$.

$$E(\boldsymbol{\beta}|\boldsymbol{y}) = \begin{pmatrix} -0.955(0.148) \\ -0.031(0.127) \\ -0.014(0.012) \\ -0.193(0.138) \\ -0.194(0.140) \\ 0.379(0.143) \\ 0.136(0.146) \\ 0.475(0.42) \\ 0.39(0.142) \end{pmatrix}$$

$$E(\boldsymbol{\Sigma}|\boldsymbol{y}) = \begin{pmatrix} 1.236 & 0.877 & 0.575 & 0.541 & 0.489 & 0.337 \\ 0.877 & 1.569 & 0.764 & 0.727 & 0.645 & 0.362 \\ 0.575 & 0.764 & 0.946 & 0.805 & 0.717 & 0.399 \\ 0.541 & 0.727 & 0.805 & 1.438 & 1.052 & 0.547 \\ 0.489 & 0.645 & 0.717 & 1.052 & 1.370 & 0.584 \\ 0.337 & 0.362 & 0.399 & 0.547 & 0.584 & 1.449 \end{pmatrix}$$

with standard deviations:

$$\begin{pmatrix} 0.705 & 0.283 & 0.156 & 0.151 & 0.145 & 0.138 \\ 0.283 & 0.404 & 0.180 & 0.185 & 0.150 & 0.122 \\ 0.156 & 0.180 & 0.186 & 0.157 & 0.131 & 0.113 \\ 0.151 & 0.185 & 0.157 & 0.288 & 0.182 & 0.150 \\ 0.145 & 0.150 & 0.131 & 0.182 & 0.225 & 0.166 \\ 0.138 & 0.122 & 0.113 & 0.150 & 0.166 & 0.243 \end{pmatrix}$$

The posterior predictive means are displayed for the 2 free cut points for each classifying variable. Note that the first and last cut points were set to be -1 and 1 respectively. The six rows of the matrix correspond to the six classifying variables (Time Points).

$$E(\boldsymbol{\theta}|\boldsymbol{y}) = \begin{pmatrix} -0.200(0.099) & 0.657(0.093) \\ -0.428(0.084) & 0.284(0.105) \\ -0.462(0.071) & 0.045(0.081) \\ -0.452(0.079) & 0.152(0.089) \\ -0.277(0.079) & 0.303(0.083) \\ -0.449(0.074) & 0.274(0.098) \end{pmatrix}$$

The elements of $\boldsymbol{\beta}$ correspond to the covariates Treatment (with baseline no treatment), Male (with baseline female), Age and the six Time points respectively. A higher response corresponds to more pain, so we can therefore draw

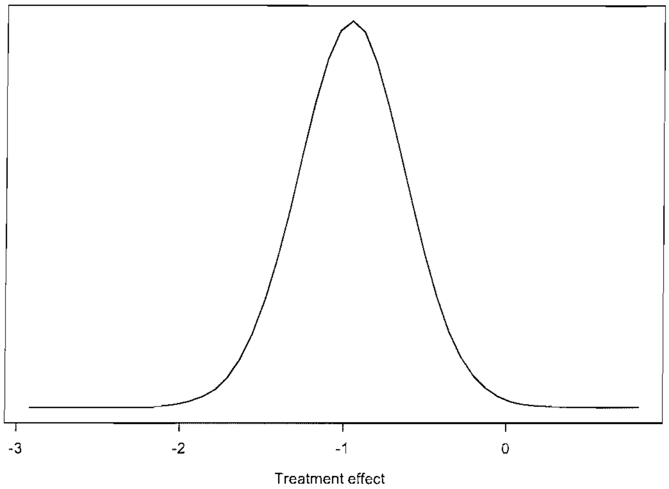


Figure 6.6: Posterior density of the treatment effect for shoulder pain data

the following conclusions from these results. The active treatment clearly improves pain levels compared to no treatment. Males rate pain as marginally lower than females. As age increases, reported pain levels decrease. However, both covariate estimates for sex and age are very close to zero so this effect may not be significant. The posterior distributions for the covariates Treatment (with baseline no treatment), Sex (with baseline female) and Age are shown in Figures 6.6, 6.7 and 6.8 respectively.

For the time covariate, there appears to be an overall trend from Time 1 to Time 6 of decreasing pain. The posterior distributions of the Time effects are shown in Figure 6.9.

These results agree on the most part with those found by Lumley (1996). The only point on which they do not agree is that Lumley finds that overall males rate pain more highly than females. However the posterior standard deviation of this estimate is larger than the posterior mean itself and is therefore not significant.

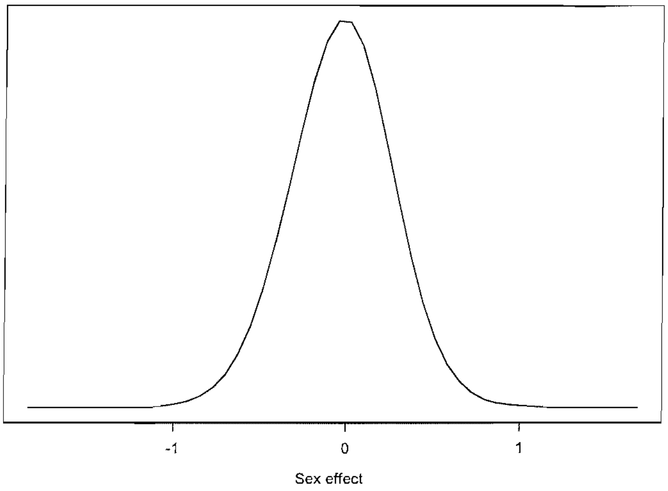


Figure 6.7: Posterior density of the sex effect for shoulder pain data

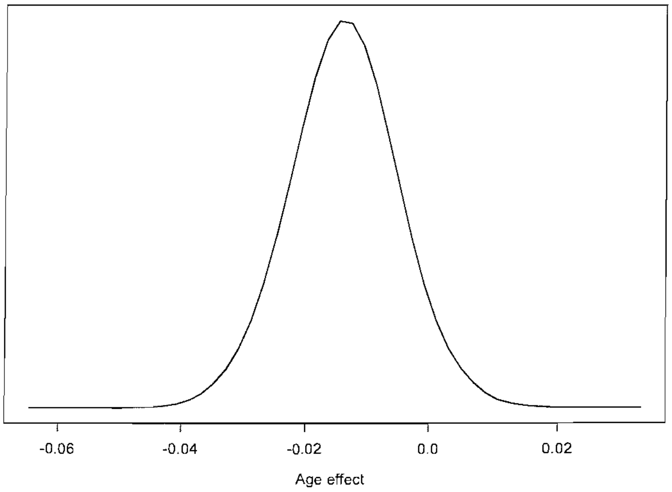


Figure 6.8: Posterior density of the age effect for shoulder pain data

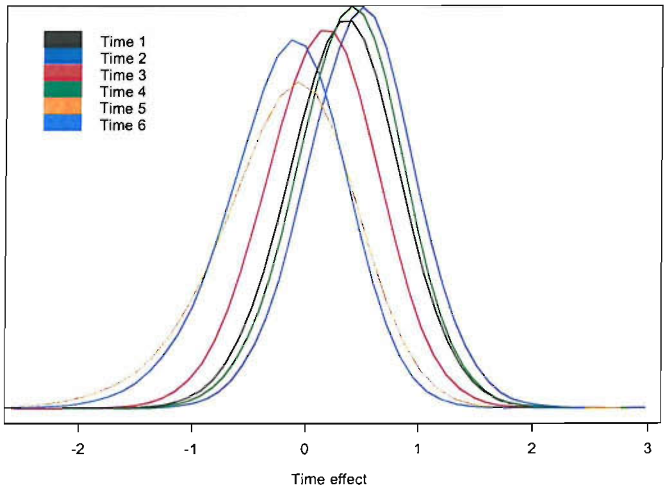


Figure 6.9: Posterior densities of the six time effects for shoulder pain data

From the posterior mean of the covariance matrix $\mathbf{\Sigma}$, we note that covariances decrease further from the diagonal. To investigate this, consider the correlation matrix $\boldsymbol{R}$:

$$\boldsymbol{R} = \begin{pmatrix} 1 & 0.630 & 0.532 & 0.406 & 0.376 & 0.252 \\ 0.630 & 1 & 0.627 & 0.484 & 0.440 & 0.240 \\ 0.532 & 0.627 & 1 & 0.690 & 0.630 & 0.341 \\ 0.406 & 0.484 & 0.690 & 1 & 0.750 & 0.379 \\ 0.376 & 0.440 & 0.630 & 0.750 & 1 & 0.414 \\ 0.252 & 0.240 & 0.341 & 0.379 & 0.414 & 1 \end{pmatrix}$$

The correlations decrease going away from the diagonal, although they do not decrease enough to fit an AR(1) structure.

The issue of model choice was then considered. Note that model search is carried out within the set of directed decomposable graphical models, as the data is longitudinal. Therefore, there is no need to move between orderings. The Reversible Jump algorithm (Algorithm 5) from Chapter 4 with adaptations described above was applied with 100,000 iterations. Following the arguments of section 4.2, the priors were set to be $\boldsymbol{A} = \text{diag}(0.708)$, $q = 8$ and $\boldsymbol{T} = \text{diag}(50)$. Table 6.2.1 shows the posterior model probabilities for the 5 most probable models which between them account for 10% of the posterior probability.

Model	Posterior Model Probability
T1:T2 + T2:T3 + T3:T4 + T4:T5:T6	0.029
T1:T2 + T2:T3:T4 + T4:T5:T6	0.027
T1:T2 + T2:T3:T4 + T4:T5 + T5:T6	0.018
T1:T2:T3 + T3:T4 + T4:T5:T6	0.017
T1:T2 + T2:T3 + T3:T4 + T4:T5 + T5:T6	0.013

Table 6.3: Posterior model probabilities for shoulder pain data

There is very high posterior probability of relationships between Times 1 and 2, 2 and 3, 3 and 4, 4 and 5, and 5 and 6. These dependences occur in each of the most probable models. The two most probable models also involve a three-way interaction between Times 4, 5 and 6, and slightly less popular models involve other three-way interactions of adjacent variables in the ordering. The autocorrelation structure could possibly be lag 1 or lag 2. The four most probable models are shown in Figure 6.10

Again, goodness-of-fit is difficult to assess as the number of observations is small compared to the number of classifying cells. Convergence is assessed using trace plots. The trace plots for β_3 (Age) and $\theta(1,3)$ (third cut point for Treatment) are shown in Figures 6.11 and 6.12.

6.3 Covariate Selection

So far, we have not considered whether each covariate has a significant effect on the response, and have included them all in the model. However, this may not always be appropriate as some covariates may have little or no effect on

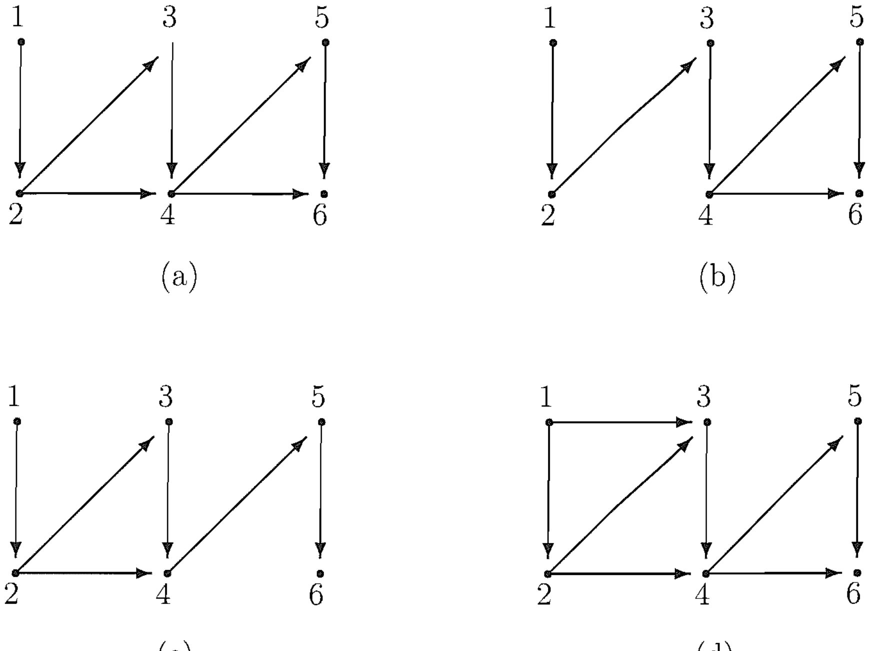


Figure 6.10: Most probable models for shoulder data

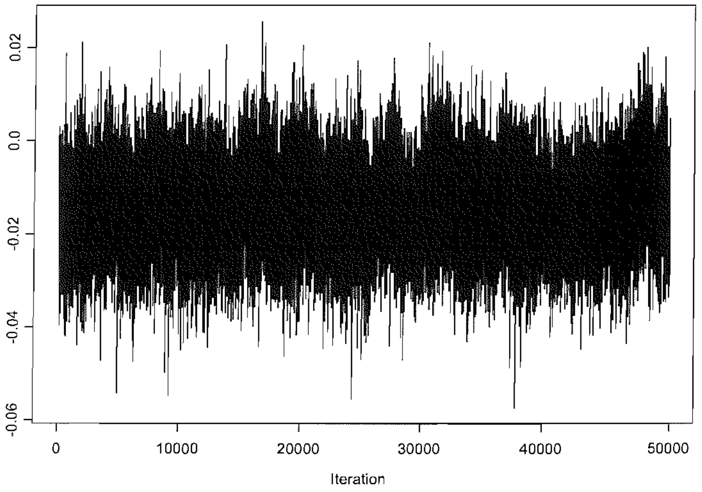


Figure 6.11: Trace plot for β_3 for shoulder pain data

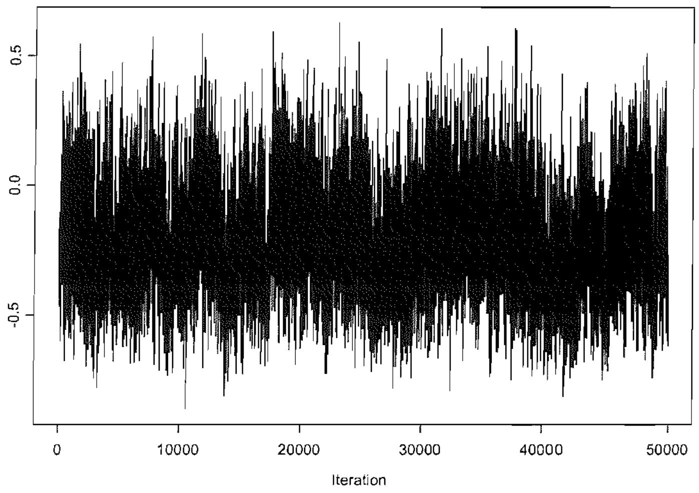


Figure 6.12: Trace plot for $\theta(1,3)$ for shoulder pain data

the response. In this section, we describe and implement an extra Reversible Jump step to decide which covariates should be included in a model for the data. Throughout this section, whenever the phrase model choice is used, we are referring to covariate model choice.

For a covariate not to be in the model, all elements of $\boldsymbol{\beta}$ corresponding to that covariate should be set to zero. This seems appropriate for the longitudinal examples considered here. Where classifying variables do not represent the same response at different time points it may be desired to allow models where a particular covariate may affect one classifying variable but not another. The general RJMCMC approach proposed here would allow such models to be considered. If a covariate is included in the model, then there is no restriction on the corresponding values of $\boldsymbol{\beta}$. Therefore a model in this case is characterised by the covariate vector $\boldsymbol{\beta}$ for the full model with parameters for 'missing' covariates set to zero. For a data set with C covariates, there are C possible move types, each corresponding to one of the covariates. A move can consist of dropping a covariate if it is present in the current model or of adding in a covariate if not present in the current model. Dropping the covariate corresponds to setting the corresponding values of $\boldsymbol{\beta}$ to zero, and adding the covariate corresponds to proposing a new value for the appropriate values of $\boldsymbol{\beta}$.

Suppose that at time t, the current state of the Markov chain is represented by $(m_{\text{COV}}^{(t)}, \boldsymbol{\xi}_{m_{\text{COV}}^{(t)}}^{(t)})$ where $\boldsymbol{\xi}_{m_{\text{COV}}^{(t)}}^{(t)}$ represents the values of the unknown parameters in model $m_{\text{COV}}^{(t)}$. Adding a covariate c with f_c levels involves a proposed move to a new model m'_{COV} and corresponding parameter vector $\boldsymbol{\xi}'_{m'_{\text{COV}}}$ with dimension $\dim(\boldsymbol{\xi}_{m_{\text{COV}}^{(t)}}^{(t)}) + f_c$. $\boldsymbol{\xi}'_{m'_{\text{COV}}}$ is created by generating a proposal $\boldsymbol{u}$ from a f_c-variate proposal distribution $q_r(\boldsymbol{u})$ and setting $\boldsymbol{\xi}'_{m'_{\text{COV}}} = g(\boldsymbol{\xi}_{m_{\text{COV}}^{(t)}}^{(t)}, \boldsymbol{u})$.

Removing a covariate c with f_c levels from the current model $m_{\text{COV}}^{(t)}$ involves a move to a model m'_{COV} with corresponding parameter vector $\boldsymbol{\xi}'_{m'_{\text{COV}}}$ with dimension $\dim(\boldsymbol{\xi}_{m_{\text{COV}}^{(t)}}^{(t)}) - f_c$, then $\boldsymbol{\xi}'_{m'_{\text{COV}}}$ is created from $\boldsymbol{\xi}_{m_{\text{COV}}^{(t)}}^{(t)}$ by applying the inverse transformation $(\boldsymbol{\xi}'_{m'_{\text{COV}}}, \boldsymbol{u}') = g^{-1}(\boldsymbol{\xi}_{m_{\text{COV}}^{(t)}}^{(t)})$ and discarding $\boldsymbol{u}'$.

Suppose that a move to a new covariate model m'_{COV} is proposed, with $\boldsymbol{\xi}'_{m'_{\text{COV}}} = g(\boldsymbol{\xi}_{m_{\text{COV}}^{(t)}}^{(t)})$ and suppose that the probability of making move type r given the current state of the Markov chain is $j(r, \boldsymbol{\xi}_{m_{\text{COV}}^{(t)}}^{(t)}, m_{\text{COV}}^{(t)})$. Then a move to a new model by adding a covariate should be accepted with probability α_{COV} where

$$\alpha_{\text{COV}} = \min\left\{1, \frac{f(\boldsymbol{\xi}'_{m'_{\text{COV}}}, m'_{\text{COV}}|\boldsymbol{y})j(r, m'_{\text{COV}}, \boldsymbol{\xi}'_{m'_{\text{COV}}})}{f(\boldsymbol{\xi}_{m_{\text{COV}}^{(t)}}^{(t)}, m_{\text{COV}}^{(t)}|\boldsymbol{y})j(r, m_{\text{COV}}^{(t)}, \boldsymbol{\xi}_{m_{\text{COV}}^{(t)}}^{(t)})q_r(\boldsymbol{u})}\left|\frac{\partial(\boldsymbol{\xi}'_{m'_{\text{COV}}})}{\partial(\boldsymbol{\xi}_{m_{\text{COV}}^{(t)}}^{(t)}, \boldsymbol{u})}\right|\right\} \tag{6.2}$$

For the reverse move to a new model by removing a covariate, the acceptance probability is

$$\alpha_{\text{COV}} = \min\left\{1, \frac{f(\boldsymbol{\xi}'_{m'_{\text{COV}}}, m'_{\text{COV}}|\boldsymbol{y})j(r, m'_{\text{COV}}, \boldsymbol{\xi}'_{m'_{\text{COV}}})q_r(\boldsymbol{\xi}_{m_{\text{COV}}^{(t)} \backslash m'_{\text{COV}}}^{(t)})}{f(\boldsymbol{\xi}_{m_{\text{COV}}^{(t)}}^{(t)}, m_{\text{COV}}^{(t)}|\boldsymbol{y})j(r, m_{\text{COV}}^{(t)}, \boldsymbol{\xi}_{m_{\text{COV}}^{(t)}}^{(t)})}\left|\frac{\partial(\boldsymbol{\xi}'_{m'_{\text{COV}}})}{\partial(\boldsymbol{\xi}_{m_{\text{COV}}^{(t)}}^{(t)}, \boldsymbol{u})}\right|\right\} \tag{6.3}$$

Each move type is made with equal probability, so $j()$ terms will cancel in (6.2) and (6.3). Also, as we take the transformation g to be the identity transformation, there is no Jacobian term. Using the same derivation as (4.10), and assuming *a priori* that each covariate model is equally likely, equations (6.2) and (6.3) will then simplify to:

$$\alpha_{\text{COV}} = \min\left\{1, \frac{f(\boldsymbol{z}^{(t)}|\boldsymbol{\Phi}^{(t)}, \boldsymbol{\beta}'_{m'_{\text{COV}}}) f(\boldsymbol{\Phi}^{(t)}, \boldsymbol{\beta}'_{m'_{\text{COV}}}|m'_{\text{COV}})}{f(\boldsymbol{z}^{(t)}|\boldsymbol{\Phi}^{(t)}, \boldsymbol{\beta}^{(t)}_{m^{(t)}_{\text{COV}}}) f(\boldsymbol{\Phi}^{(t)}, \boldsymbol{\beta}^{(t)}_{m^{(t)}_{\text{COV}}}|m^{(t)}_{\text{COV}})} \frac{1}{q_r(\boldsymbol{\beta}'_{m'_{\text{COV}}})}\right\} \tag{6.4}$$

and

$$\alpha_{\text{COV}} = \min\left\{1, \frac{f(\boldsymbol{z}^{(t)}|\boldsymbol{\Phi}^{(t)}, \boldsymbol{\beta}'_{m'_{\text{COV}}}) f(\boldsymbol{\Phi}^{(t)}, \boldsymbol{\beta}'_{m'_{\text{COV}}}|m'_{\text{COV}})}{f(\boldsymbol{z}^{(t)}|\boldsymbol{\Phi}^{(t)}, \boldsymbol{\beta}^{(t)}_{m^{(t)}_{\text{COV}}}) f(\boldsymbol{\Phi}^{(t)}, \boldsymbol{\beta}^{(t)}_{m^{(t)}_{\text{COV}}}|m^{(t)}_{\text{COV}^{(t)}})} \frac{q_r(\boldsymbol{\beta}_{m^{(t)}_{\text{COV}}} \backslash \boldsymbol{\beta}_{m'_{\text{COV}}})}{1}\right\} \tag{6.5}$$

when adding or removing covariates respectively.

The likelihood terms $f(\boldsymbol{z}|\boldsymbol{\Phi}, \boldsymbol{\beta}_{m_{\text{COV}}})$ are given by the $p-$variate normal distribution, the parameters of which depend on the covariate model. The variance $(\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}$ is independent of the covariate model, but the structure of the mean $\boldsymbol{X}_i\boldsymbol{\beta}$ varies according the model, with covariates not included in the model leading to the presence of zeros in $\boldsymbol{\beta}$.

The prior terms $f(\boldsymbol{\Phi}, \boldsymbol{\beta}_{m_{\text{COV}}}|m_{\text{COV}})$ are factorised as $f(\boldsymbol{\beta}_{m_{\text{COV}}}|m_{\text{COV}})f(\boldsymbol{\Phi}|m_{\text{COV}})$ due to the choice of independent priors for $\boldsymbol{\Phi}$ and $\boldsymbol{\beta}_{m_{\text{COV}}}$. Since $\boldsymbol{\Phi}$ is covariate model independent, these terms will cancel in the reversible jump acceptance probabilities. The priors for $\boldsymbol{\beta}_{m_{\text{COV}}}$ for the full model are multivariate normal with mean vector $\boldsymbol{0}$ and variance the diagonal matrix $\boldsymbol{T}$. Since in any reduced model, some of these means are set to zero if the corresponding covariates are not in the model, these zero elements are conditioned on to obtain the prior for the reduced. Since the prior mean is $\boldsymbol{0}$ and the prior variance is diagonal, the prior density will be simply multivariate normal with dimension equal to the number of non-zero elements of $\boldsymbol{\beta}$, mean vector $\boldsymbol{0}$ and variance matrix $\boldsymbol{T} = \text{diag}(\tau_{ii}^2)$.

Finally we consider the proposal. As in Chapter 4, we have the useful property that the normalised conditional posterior distribution is available and may be used to find a suitable proposal distribution for the element of $\boldsymbol{\beta}$ to be added. The joint conditional posterior distribution for all covariate means $\boldsymbol{\beta}$ is as follows:

$$\boldsymbol{\beta}|\boldsymbol{z},\boldsymbol{\Phi} \sim N_L\left(\left(\sum_{i=1}^{n}\boldsymbol{X}_i^T\boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{X}_i+\boldsymbol{T}^{-1}\right)^{-1}\sum_{i=1}^{n}\boldsymbol{X}_i^T\boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{z}_i\ ,\right.$$
$$\left.\left(\sum_{i=1}^{n}\boldsymbol{X}_i^T\boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{X}_i+\boldsymbol{T}^{-1}\right)^{-1}\right)$$

When proposing to add a covariate, we must propose new values for all elements of $\boldsymbol{\beta}$ associated with that covariate. A suitable proposal may therefore be found by conditioning on the other elements of $\boldsymbol{\beta}$ in the usual manner for conditional multivariate normal distributions (see section 3.7.3). When proposing to remove a covariate, the proposed values for the corresponding elements of $\boldsymbol{\beta}$ are 0.

The full algorithm including both types of model choice and order choice now runs as follows.

6.3.1 Algorithm 7

An initial model, order, covariate model and values for al parameters in this model are specified. Then,

1. With probability p, remain in current graphical model, order and covariate model and re-generate all parameters (the null move). Else, with probability $1-p$, a new model is proposed by either adding or subtracting a randomly selected edge from the current model. If the proposed model is unavailable in the current ordering, go to Step 2. Otherwise the proposed model is accepted with Reversible Jump probability α specified in (4.11) and (4.12).

2. A new ordering is proposed by randomly selecting a pair of adjacent variables in the current ordering and proposing to switch them. If the current model is not available in the proposed ordering, then go to step 3. Otherwise, the proposed ordering is accepted with Reversible Jump probability α_{order} as defined in (5.3).
3. A new covariate model is proposed by randomly selecting a covariate and proposing to set the corresponding elements of $\boldsymbol{\beta}$ to 0 if the covariate is in the current model, or proposing new values for the corresponding elements of $\boldsymbol{\beta}$ if the covariate is not in the current model. The proposed model is accepted with Reversible Jump probability α_{cov} defined in (6.4) and (6.5).
4. Go to step 1 and repeat.

Note that it is possible to focus on a particular aspect of the model and covariate model selection procedure by skipping step 3 or step 1 respectively. Also, step 2 is not required if the order of the classifying variables is fixed.

6.3.2 Examples

We apply this method to the same two examples.

Example 1 - Crossover Data

Algorithm 7 was applied to the crossover data set with 50,000 iterations and with the same priors as before. The Reversible Jump step for covariates was found to be far less mobile than both the model change and order change Reversible Jump steps, with proposed moves being accepted approximately 2% of the time. The posterior covariate model probabilities are shown in Table 6.3.2. The most popular model is the Treatment main effect model which accounts for 95.4% of the posterior model probability. Clearly there is

Model	Posterior Model Probability
Treatment + Period	0.029
Treatment	0.954
Period	0.003
No Covariates	0.014

Table 6.4: Posterior covariate model probabilities for crossover trial data

very strong evidence that the Treatment main effect should be included in a covariate model for the data set. Note that the fact that one model is found to be far more probable than all other models accounts for the fact that the proposal is not often accepted. This is demonstrated by the fact that proposals to move to the Treatment model are accepted at a rate of 97%.

Example 2 - Shoulder Pain Data

Algorithm 7 was applied to the shoulder pain data with 100,000 iterations and with the same prior parameters as for the non covariate selection case. The Reversible Jump step for covariate mode change was far more mobile for this example with a proposed move acceptance rate of 9%. Table 6.3.2 shows the posterior covariate model probabilities for the four most popular covariate models for the shoulder pain data. Between them, these four covariate models account for 99.8% of the posterior probability. The treatment covariate occurs

Model	Posterior Model Probability
Treatment	0.702
Sex + Treatment	0.253
Period + Treatment	0.032
Period + Sex + Treatment	0.011

Table 6.5: Posterior covariate model probabilities for shoulder pain data

in each of the models, there is therefore very high posterior probaility of a treatment effect. There is also some probability of both Sex and Period effects, but only in the presence of the Treatment effect.

Both examples given here only require reversible jump between directed models, but as the reversible jump algorithm for changing orderings is unaffected by the presence of covariates, this would not create any difficulties.

Chapter 7

Discussion and Extensions

7.1 Discussion

The aim of the thesis has been to provide a coherent methodology for multivariate ordinal and binary data, encompassing both parameter estimation and model selection.

Methods proposed previously were discussed in Chapter 2, with particular emphasis on those introduced by Albert and Chib (1993) and Chib and Greenberg (1998), as in this work we use similar ideas to those suggested by these authors. The methods of Chib and Greenberg (1998) and Chen and Dey (2000) may be applied to multivariate binary data and multivariate ordinal data respectively. However, we have given a new parameterisation that is far more flexible, allowing for the modelling of ordinal or binary or a mixture of both types of data.

The parameterisation involves characterising the model in terms of the Cholesky decomposition of its inverse variance matrix. The use of the Cholesky decomposition parameterisation allows for conjugate prior distributions so that sampling from the posterior distributions of the model parameters is straightforward. The model parameters are estimated using a Gibbs sampler and a data augmentation approach. The approach has been illustrated with applications to two examples, one all ordinal and one a mixture of binary and ordinal. Goodness-of-fit of the saturated model can be assessed by use of a simulation approach and may be seen to be extremely good for data with three or fewer dimensions.

The issue of model choice for decomposable directed graphical models is considered in Chapter 4. Such models imply an ordering of variables with directed edges going from variables earlier in the ordering to those later in the ordering. Models are characterised by the structure of the Cholesky decomposition matrix $\mathbf{\Phi}$. A zero entry in $\mathbf{\Phi}$ is equivalent to conditional independence between the corresponding variables given all variables preceding them in the ordering.

A Reversible Jump approach is employed to investigate which models are most likely to predict the observed data. Standard Bayesian methods for model comparison involve comparing the marginal likelihoods of competing models. However, this is not often possible as the marginal likelihood is analytically intractable and must be estimated by other means, for example, by simulation. In this work, the marginal likelihood is unavailable. The Reversible Jump procedure circumvents this problem by sampling from the joint model and associated model parameter space.

Moves take the form of either adding or removing an edge from the current model, which corresponds to generating or setting to zero an appropriate element of $\mathbf{\Phi}$ respectively. The Reversible Jump acceptance probability takes a fairly simple form as the prior distributions for many of the model parameters are model independent. The Cholesky decomposition parameterisation provides a further useful property here in that a suitable proposal distribution for the Reversible Jump procedure is simply the conditional posterior distribution of the elements of $\mathbf{\Phi}$ to be added or removed.

The Reversible Jump procedure was applied to two data sets where the classifying variables take a natural ordering. For both examples, it was seen to be very mobile with proposed moves accepted approximately 1 in 7 times and

1 in 3 times respectively. Models estimated were compared with results from others' analysis of the two sets of data and were found to be very similar. The goodness-of-fit was assessed for both examples and here there was some cause for concern as the model averaged fit was not as good as might be hoped. However, despite this lack of fit the models selected were still comparable to those found by less parsimonious approaches.

Prior parameters were also discussed in Chapter 4. The choice of these can be shown to have a strong effect on the models selected by the Reversible Jump procedure.

The model provides a natural framework for fitting directed acyclic graphical models for data where the classifying variables are ordered, but it may also be extended to situations where this is not the case. In Chapter 5, we considered model selection for decomposable graphical models using Reversible Jump MCMC. As discussed in Chapter 4, the parameterisation of the model leads to a natural ordering of the classifying variables. However, not all undirected graphs are available in any one particular ordering as the conditional independence structure is unsupported by the ordering. The Reversible Jump procedure described in Chapter 4 was extended to take account for this. In the extra step, a proposed move involves permuting two classifying variables in the ordering. In this way, all undirected models are covered by the model space that the Reversible Jump procedure passes through.

All examples considered up to this stage did not involve covariates; in Chapter 6, the methods described for modelling and choosing appropriate models with which to do so were adapted to allow for the presence of covariates. The Gibbs sampler algorithm for estimating parameters in a single model and the Reversible Jump algorithms for moving between models and between orderings were adapted and successfully applied to two data sets. A futher Reversible Jump step was then described that moved between covariate models by proposing to add or remove a covariate from the current model.

7.2 Extensions

The work has many possibilities for further extension and investigation. A simple extension would be to apply the methods described here to mixed data, that is, data where there is a mixture of ordinal (or binary) and continuous response variables. For such data one might model the joint distribution of the latent variables for the categorical responses with the continuous responses. Thus no latent data generation step is required for the continuous variables so modelling such data would in fact be less computationally intensive than modelling purely categorical data. There are also no cut points to be generated for the continuous responses.

There is some suggestion from the examples given in Chapters 4 and 5 that the model may struggle to fit the data well if it is of higher dimension. This is almost certainly due to the highly parsimonious nature of the models described here. For example, consider fitting a standard log linear model to the Alcohol, Obesity and Hypertension data shown in Table 3.5. Fitting the saturated log-linear model AOH requires estimating an extra 12 parameters than when fitting the log-linear model $AH+OH$ for example. In contrast, with the models described here, only one extra parameter is required to be estimated. For this work, this has not caused a computational problem as the main focus was on model determination. However, there are certain modifications that could be applied to overcome this. For example, the use of the multivariate normal distribution may not be appropriate. Perhaps a heavier-tailed distribution would provide a better fit. Also, it may be unrealistic to expect the data to be centered on a single mean, so a mixture of normal distributions (McLachlan and Peel, 2000) may provide a better fit.

The performance of the data augmentation Gibbs sampler algorithms appears to have been satisfactory for the examples given here. However, it is possible that convergence could be improved in two ways. Various methods described in Section 2.2.2 could be implemented to speed convergence of the free cut points, while the method of parameter expansion described in Section 3.4.2 could also be implemented to improve model convergence, although this approach has not previously been applied to ordinal data.

Models are underparameterised so that effectively, we are using an approach that attempts to fit a model using combinations of two factor interactions. This does however have some advantages in the parsimonious nature of the models selected to do so. It would not be recommended to use this approach for data of very high dimension, that could not practically be represented in a contingency table for example.

Another avenue for further exploration is in the choice of prior distributions. The prior distributions for elements of $\mathbf{\Phi}$ were chosen to be equivalent to an Inverse Wishart distribution for $\mathbf{\Sigma}$; however, it would be possible to choose the prior parameters to correspond to a more flexible prior distribution for $\mathbf{\Sigma}$, the generalised inverse Wishart distribution.

In this thesis, the covariance matrix has been modelled with no restrictions on its structure. Another approach to modelling the covariance matrix could be to specify models in terms of competing correlation structures. For example, a common model for correlation structure for longitudinal data is to consider the set of autoregressive models, denoted $AR(p)$ for a model of order p. For longitudinal data examples, one could construct a Reversible Jump move between competing orders of AR models.

All computation described in this thesis was carried out using custom-written C source code. However, it is worth noting that due to the fact that the conditional posterior distributions for all model parameters are standard, the Gibbs sampler procedure described here to estimate model parameters may also be carried out in the software package BUGS (Gilks et al., 1994).

References

Agresti, A. (1990) *Categorical Data Analysis.* John Wiley & Sons.

Albert, J. and Chib, S. (1993) Bayesian Analysis of Binary and Polychotomous Response Data. *Journal of the American Statistical Association*, **88**, 669–679.

Anderson, J. and Pemberton, J. (1985) The grouped continuous model for multivariate ordered categorical variables and covariate adjustment. *Biometrics*, **41**, 875–885.

Ashford, J. and Sowden, R. (1970) Multi-Variate Probit Analysis. *Biometrics*, **26**, 535–546.

Bekele, B. and Thall, P. (2004) Dose-Finding based on multiple toxicites in a soft tissue sarcoma trial. *Journal of the American Statistical Association*, **99**, 26–35.

Berger, J. O. (1985) *Statistical Decision Theory and Bayesian Analysis.* Springer-Verlag.

Breslow, N. (1982) Covariance adjustment of relative-risk estimates in matched studies. *Biometrics*, **38**, 661–672.

Brown, P., Le, N. and Zidek, J. (1994) Inference for a covariance matrix. In *Aspects of Uncertainty* (eds. P. Freeman and A. Smith), 77–92. Wiley.

Chatfield, C. and Collins, A. (1980) *Introduction to Multivariate Analysis.* Chapman & Hall/CRC.

Chen, M.-H. and Dey, D. (2000) Bayesian analysis for correlated ordinal data models. In *Generalized linear models: a Bayesian Perspective* (eds. S. G. D.K. Dey and B. Mallick), 133–157. Marcel Dekker.

Chib, S. (1995) Marginal Likelihood from the Gibbs Output. *Journal of the American Statistical Association*, **90**, 1313–1321.

Chib, S. and Greenberg, E. (1998) Analysis of multivariate probit models. *Biometrika*, **85**, 347–361.

Cowles, M. (1996) Acceleration Monte Carlo Markov chain convergence for cumulative-link generalized linear models. *Statistics and Computing*, **6**, 101–111.

Daniels, M. and Pourahmadi, M. (2002) Bayesian analysis of covariance matrices and dynamic models for longitudinal data. *Biometrika*, **89**, 553–566.

Darroch, J., Lauritzen, S. and Speed, T. (1980) Markov Fields and Log-Linear Interaction Models for Contingency Tables. *The Annals of Statistics*, **8**, 522–539.

Dawid, A. and Lauritzen, S. (1993) Hyper Markov Laws in the Statistical Analysis of Decomposable Graphical Models. *The Annals of Statistics*, **21**, 1272–1317.

Dellaportas, P. and Forster, J. (1999) Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models. *Biometrika*, **86**, 615–633.

Devroye, L. (1986) *Non-Uniform Random Variate Generation.* Springer-Verlag.

Dey, D., Gelfand, A., Swartz, T. and Vlachos, P. (1998) Simulation based model checking for hierarchical models. *Test*, **7**, 325–346.

Dey, D., Ghosh, S. and Mallick, B. (eds.) (2000) *Generalized Linear Models A Bayesian Perspective.* Marcel Dekker, Inc.

van Dyk, D. and Meng, X.-L. (2001) The art of data augmentation (with discusion). *The Journal of Computation and Graphical Statistics*, **10**, 1–111.

Edwards, D. and Havranek, T. (1985) A fast procedure for model search in multidimensional contingency tables. *Biometrika*, **72**, 339–351.

Fowlkes, E. B., Freeny, A. E. and Landwehr, J. M. (1988) Evaluating Logistic Models for Large Contingency Tables. *Journal of the American Statistical Association*, **83**, 611–622.

Fronk, E. (2003) Model selection for dags via rjmcmc for the discrete mixed case. Ludwig-Maximilians-University, Munich.

Garthwaite, P. and Al-Awadhi, S. (2001) Non-conjugate prior distribution assessment for multivariate normal sampling. *Journal of the Royal Statistical Society B*, **63**, 95–110.

Geman, S. and Geman, D. (1984) Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattn. Anal. Mach. Intel.*, **6**, 721–741.

Geweke, J. (1991) Efficient simulation from the multivariate Normal and Student-t distributions subject to linear constraints. In *Computing Science and Statistics: Proceedings of the Twenty-Third Symposium on the Interface* (ed. E. Keramidas), 571–578. Interface Foundation of North American, Inc., Fairfax.

Gilks, W., Thomas, A. and Spiegelhalter, D. (1994) A language and program for complex Bayesian modelling. *The Statistician*, 169–178.

Green, P. (1995) Reversible jump Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.

Griffiths, W.E. Hill, R. and Pope, P. (1987) Small sample properties of probit model estimators. *Journal of the American Statistical Association*, **82**, 929–937.

Gross, P. (1971) A study of supervisor reliability. *Tech. rep.*, Laboratory of Human Development, Harvard Graduate School of Education.

Hastings, W. (1970) Monte carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.

Holmes, M. and Williams, R. (1954) The distribution of carriers of *Streptococcus pyogenes* among 2413 healthy children. *J.Hyg.*, **52**, 165–179.

Imai, K. and van Dyk, D. (2005) A Bayesian analysis of the multinomial probit model using marginal data augmentation. *Journal of Econometrics*, **124**, 311–334.

Ishwaran, H. (2000) Univariate and multirater ordinal cumulative link regression with covariate specific cutpoints. *The Canadian Journal of Statistics*, **28**, 715–730.

Ishwaran, H. and Gatsonis, C. (2000) A general class of hierarchical ordinal regression models with applications to correlated roc analysis. *The Canadian Journal of Statistics*, **28**, 731–750.

Johnson, V. (1996) On Bayesian Analysis of Multirater Ordinal Data: An Application to Automated Essay Grading. *Journal of the American Statistical Association*, **91**, 42–51.

Johnson, V. and Albert, J. (1999) *Ordinal Data Modelling.* Springer-Verlag.

Johnson, V., Deaner, R. and van Schaik, C. (2002) Bayesian Analysis of Rank Data with Application to Primate Intelligence Experiments. *Journal of the American Statistical Association*, **97**, 8–17.

Jones, B. and Kenward, M. (2003) *Design and analysis of cross- over trials.* Chapman and Hall.

Kass, R. and Wasserman, L. (1995) A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association*, **90**, 928–934.

Kizilkaya, K., Carnier, P., Albera, A., Bittante, G. and Tempelman, R. (2003) Cumulative t-link threshold models for the genetic analysis of calving ease scores. *Genetics Selection Evolution*, **35**, 489–512.

Knuiman, M. and Speed, T. (1988) Incorporating prior information into the analysis of contingency tables. *Biometrics*, **44**, 1061–71.

Liu, J. (2001) *Monte Carlo Strategies in Scientific Computing.* Springer.

Liu, J. and Wu, Y. (1999) Parameter expansion for data augmentation. *Journal of the American Statistical Association*, **94**, 1264–1274.

Lumley, T. (1996) Generalized estimating equations for ordinal data: a note on working correlation structures. *Biometrics*, **52**, 354–361.

Madigan, D. and Raftery, A. (1994) Model Selection and Accounting for Model Uncertainty in Graphical Models using Occam's Window. *Journal of the American Statistical Association*, **89**, 1535–1546.

Madigan, D. and York, J. (1995) Bayesian graphical models for discrete data. *International Statistical Review*, **63**, 215–232.

McCullagh, P. (1980) Regression Models for Ordinal Data. *J.R. Statist. Soc. B*, **42**, 109–142.

McLachlan, G. and Peel, D. (2000) *Finite mixture models.* Wiley.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. and Teller, E. (1953) Equations of state calculations by fast computating machines. *Journal of Chemical Physics*, **21**, 1087–1092.

Nandram, B. and Chen, M.-H. (1996) Reparameterizing the generalized linear model to accelerate Gibbs sampler convergence. *J. Statist. Comput. Simul.*, **54**, 129–144.

Ntzoufras, I., Dellaportas, P. and Forster, J. (2003) Bayesian variable and link determination for generalised linear models. *Journal of Statistical Planning and Inference*, **111**, 165–180.

O'Hagan, A. (1994) *Bayesian Inference*, vol. 2B of *Kendall's Advanced Theory of Statistics.* Edward Arnold.

Raftery, A. (1996) Approximate Bayes factors and accounting for modle uncertainty in generalized linear models. *Biometrika*, **83**, 251–266.

Rossi, P., Gilula, Z. and Allenby, G. (2001) Overcoming scale usage heterogeneity: A Bayesian hierarchical approach. *Journal of the American Statistical Association*, **96**, 20–31.

Roverato, A. (2002) Hyper inverse Wishart Distribution for Non-decomposable Graphs and its Application to Bayesian Inference for Gaussian Graphical Models. *Scandinavian Journal of Statistics*, 391–411.

Tanner, M. and Wong, W. (1987) The Calculation of Posterior Distributions by Data Augmentation. *Journal of the American Statistical Association*, **82**, 528–540.

Tierney, L. (1994) Markov Chains for Exploring Posterior Distributions. *The Annals of Statistics*, **22**, 1701–1762.

Wermuth, N. and Cox, D. (1998) On the Application of Conditional Independence to Ordinal Data. *International Statistical Review*, **66**, 181–199.

Whittaker, J. (1990) *Graphical Models in Applied Mathematical Multivariate Statistics.* John Wiley and Sons.

Zellner, A. and Rossi, P. (1984) Bayesian analysis of dichotomous quantal response models. *Journal of Econometrics*, **25**, 365–393.