

UNIVERSITY OF SOUTHAMPTON

Faculty of Engineering, Science and Mathematics

School of Mathematics

Bayesian Inference for Graphical Gaussian and
Conditional Gaussian Models

by David O'Donnell

Thesis submitted for the degree of Doctor of Philosophy

March 2004

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

MATHEMATICS

Doctor of Philosophy

BAYESIAN INFERENCE FOR GRAPHICAL GAUSSIAN AND
CONDITIONAL GAUSSIAN MODELS

by David O'Donnell

Graphical Models are families of distributions satisfying a set of conditional independence relationships, which may be represented by a graph where vertices represent the variables under study and pairwise conditional independences are represented by missing edges. Much of the work concerning Bayesian inference for these models has dealt with those where all the variables are either all discrete or all continuous. While the case of purely discrete models has been treated thoroughly, most work on the purely continuous case of graphical Gaussian models has been restricted to decomposable models, which allow analyses to be broken down into sub-analyses of smaller, simple sub-models and graphical models with variables of each type have been given very little attention in a Bayesian context. This thesis addresses these two issues through the use of Markov chain Monte Carlo methods, a powerful tool for Bayesian inference. Methodology for inference both for fixed models and under model uncertainty is developed for the entire class of graphical Gaussian models, avoiding the use of conjugate prior distributions. This approach is then applied to simple mixed graphical models as well as to the larger class of hierarchical interaction models.

Contents

1	Introduction and Background	11
1.1	Conditional Independence	11
1.2	Graph Theory	11
1.3	Graphical Models	14
1.3.1	Decomposable Models	15
1.3.2	Model Indexing	15
1.3.3	Graphical Log-linear Models	16
1.3.4	Graphical Gaussian Models	17
1.3.5	Mixed Graphical Models	18
1.3.6	MIM	19
1.4	Bayesian Inference	19
1.5	Bayesian Model Selection and Inference under Model Uncertainty	20
1.6	Markov Chain Monte Carlo	22
1.6.1	Gibbs Sampling	23
1.6.2	The Metropolis-Hastings Algorithm	23
1.6.3	Reversible-Jump MCMC	24
1.6.4	Convergence Diagnostics	25
1.7	Outline of the Thesis	27
2	Graphical Gaussian Models	28
2.1	Introduction	28
2.2	The Likelihood	30
2.3	Prior Distributions	30
3	Markov Chain Monte Carlo Methods for GGMs	41
3.1	MCMC for fixed models	41
3.2	Reversible Jump MCMC for GGM's	44
3.2.1	Between-model Moves	44
3.2.2	Within-model Moves	46

3.2.3	Variations	46
3.2.4	Performance	47
3.3	Prior Sensitivity and a Larger Class of Prior	47
3.4	Simulated Data Examples	50
3.5	Further Examples	61
3.5.1	Digoxin Clearance	61
3.5.2	Anxiety and Anger	62
3.5.3	Fret's Heads	62
3.5.4	Fisher's Iris data	64
3.5.5	Mathematics Marks	67
3.5.6	Synchronized Swimming	67
3.5.7	Exam Marks	69
3.5.8	Fowl Bones	71
3.5.9	Voting Behaviour	71
3.6	Model-averaged predictive distributions	73
3.7	Summary	77
4	Mixed Graphical Models	78
4.1	Introduction	78
4.2	The Conditional Gaussian Distribution	79
4.3	Graphical, Hierarchical and other models	82
4.4	An alternative parameterization	85
4.5	The Likelihood	87
4.6	Prior Distributions	90
4.7	Conditional distributions	93
5	Conditional Gaussian Models With One Discrete variable	94
5.1	Introduction	94
5.2	MCMC for Fixed models	95
5.3	Reversible Jump Sampling	96
5.3.1	Hierarchical Models	97
5.3.2	Graphical Models	102
5.4	Model Indexing	105
5.5	Examples	105
5.5.1	Simulated data examples	106
5.5.2	A drug trial using mice	107
5.5.3	Fisher's Iris data	115

5.5.4	Tibetan Skulls	116
5.6	Model-averaged predictive distributions	121
5.7	Discussion	127
6	Conditional Gaussian Models with More Than One Discrete Variable	129
6.1	Introduction	129
6.2	MCMC for fixed models	132
6.3	Reversible Jump Sampling	134
6.3.1	Hierarchical Models	135
6.3.2	Graphical Models	144
6.4	Examples	147
6.4.1	A simulated data example	147
6.4.2	A drug trial using rats	148
6.4.3	College Smoking Habits	151
6.5	Discussion	152
7	Summary and Further Work	155
7.1	Graphical Gaussian Models	155
7.2	Mixed Graphical and CG Models	156
A	Graphs for GGM's	158
B	Marked Graphs for mixed graphical models	165

List of Figures

1.1	Example of a chain graph	13
1.2	Undirected graph equivalent to the chain graph in Figure 1.1	14
1.3	Graph 43 for 4 vertices	15
3.1	Marginal prior densities for ρ_{12} under the Beta-based prior.	49
3.2	Graphs of the models from which simulated data were generated	51
3.3	Trace plots of the two highest posterior model probabilities for four simulated datasets	54
3.4	Trace plots of the highest posterior model probabilities for a simulated dataset	55
3.5	Trace plots for the first simulated dataset	56
3.6	Graph 31345, the graph of the model from which six-dimensional simulated data were generated	56
3.7	Trace plots of the highest posterior model probabilities for a six-dimensional simulated dataset	59
3.8	Trace plots for the fifth simulated dataset	60
3.9	Model-averaged predictive densities for Y_1 (1)	75
3.10	Model-averaged predictive densities for Y_1 (2)	76
5.1	Graphs of models used to generate CG datasets	106
5.2	Trace plots of the highest posterior model probabilities for three simulated CG datasets	108
5.3	Trace plots of parameters for a simulated CG dataset	109
5.4	Batch Posterior Model Probabilities for Mice data, graphical models	111
5.5	Trace plots for mice data (1), graphical models	111
5.6	Trace plots for mice data (2), graphical models	112
5.7	Batch Posterior Model Probabilities for Mice data, hierarchical models	113
5.8	Trace plots for mice data (1), hierarchical models	113
5.9	Trace plots for mice data (2), hierarchical models	114

5.10	Most probable graph for iris data	116
5.11	The eight most probable graphs for Tibetan skulls data	119
5.12	Batch Posterior model probabilities for Tibetan skulls data	121
5.13	Model-averaged predictive densities for X in mice example (1)	123
5.14	Model-averaged predictive densities for X in mice example (2)	123
5.15	Model-averaged predictive densities for Y_1 in iris example	124
6.1	Graph of the model used to generate a set of simulated test data	148
6.2	Batch Posterior Model Probabilities for rats data	150
6.3	Undirected version of Whittaker's suggested graph for College Smoking Habits data	151
6.4	Undirected version of Wermuth and Lauritzen's graph for College Smoking Habits data	152
6.5	Saturated model in a simple class of graphical chain models	153

List of Tables

2.1	Acceptance rates for estimating prior normalizing constants for saturated models using a rejection sampler	37
2.2	Numbers of labeled and unlabelled graphs with up to 6 vertices	38
3.1	Acceptance rates for estimating prior normalizing constants for saturated GGMs using a rejection sampler.	50
3.2	Posterior model probabilities for the first simulated dataset.	52
3.3	Posterior model probabilities for the second simulated dataset.	52
3.4	Posterior model probabilities for the third simulated dataset.	52
3.5	Posterior model probabilities for the fourth simulated dataset.	53
3.6	Posterior probabilities for the four most probable models for the fifth simulated dataset	58
3.7	Edge inclusion percentages for the fifth simulated dataset	58
3.8	Posterior probabilities for the four most probable models for the sixth simulated dataset	58
3.9	Edge inclusion percentages for the sixth simulated dataset	58
3.10	Posterior model probabilities for Digoxin Clearance data	62
3.11	Posterior model probabilities for Anxiety and Anger data	62
3.12	Posterior model probabilities for Fret’s heads data (1)	63
3.13	Posterior model probabilities for Fret’s heads data (2)	64
3.14	Posterior model probabilities for Fret’s heads data (3)	65
3.15	Posterior model probabilities for <i>Iris Virginia</i>	66
3.16	Posterior model probabilities for <i>Iris Setosa</i>	66
3.17	Posterior model probabilities for <i>Iris Versicolor</i>	67
3.18	Posterior model probabilities for mathematics marks data	68
3.19	Edge inclusion %ages for synchronised swimming data	68
3.20	Posterior model probabilities for synchronised swimming data	69
3.21	Edge inclusion %ages for Exam Marks example	69
3.22	Posterior model probabilities for Exam Marks example	70

3.23	Edge inclusion %ages for Fowl Bones example	71
3.24	Posterior model probabilities for Fowl Bones example	72
3.25	Edge inclusion %ages for Voting example	73
3.26	Posterior probabilities for voting habits example	74
4.1	Numbers and Acceptance rates of CG Models	93
5.1	Posterior model probabilities for three simulated CG datasets . . .	108
5.2	Posterior probabilities for mice data	110
5.3	Edge inclusion %ages for mice data	110
5.4	Posterior model probabilities for mice data, hierarchical models .	115
5.5	Inclusion %ages for interactions for mice data	115
5.6	Most probable hierarchical models for iris data	116
5.7	Most probable graphs for type A skulls	118
5.8	Most probable graphs for type B skulls	118
5.9	Edge inclusion %ages for Tibetan skulls data	119
5.10	Posterior probabilities for Tibetan Skulls data	120
5.11	Model-averaged predictive probabilities for mice data	126
6.1	Inclusion %ages for edges for simulated data	148
6.2	Posterior probabilities for simulated data	149
6.3	Posterior probabilities for rats data	150
6.4	Inclusion %ages for edges for rats data	151
6.5	Posterior probabilities for Smoking Habits data	152
6.6	Inclusion %ages for edges for Smoking Habits data	152

Acknowledgements

Many thanks to Professor Jon Forster for initiating and encouraging this work, for twice preventing me from abandoning it and for patiently guiding the final production of this thesis. Thanks to Roy Calvert for helping me to choose Southampton and to the Mathematics department for their generous support. Thanks to the various, mostly anonymous, writers who enabled me to develop the necessary computational tools. Thanks also to Peter Smith and Steve Brooks for their helpful advice and criticism. Finally, thanks to all those who lent their facilities to assist in the production of this thesis.

Chapter 1

Introduction and Background

In this chapter, the three main elements used in this thesis are introduced, namely conditional independence and graphical models, Bayesian inference and Markov chain Monte Carlo, a powerful computational tool which is frequently used for Bayesian inference. A brief review of some graph-theoretical terminology, which is used extensively throughout is also given. An outline of the thesis and its objectives is given at the end.

1.1 Conditional Independence

Two random variables X and Y are said to be (marginally) independent if their joint density function factorizes into the product of their marginal densities, $f_{XY}(x, y) = f_X(x)f_Y(y)$, in which case we write $X \perp\!\!\!\perp Y$ after Dawid (1979). An alternative characterization is that the conditional density of Y given $X = x$ is not a function of x , written $f_{Y|X}(y|x) = f_Y(y)$.

If X , Y , Z are three random variables and for each value z , X and Y are independent in the conditional distribution given $Z = z$, we say X and Y are *conditionally independent* given Z and write $X \perp\!\!\!\perp Y|Z$. An alternative characterization is that $f_{X|Y,Z}(x|y, z)$ does not depend on y . This extends to the case of Z being a random vector.

1.2 Graph Theory

A *graph* G consists of a set of vertices, V and a set of pairs of vertices, E , called edges and can be represented pictorially as lines connecting dots or circles. If the vertex pairs in E are unordered, the graph is *undirected*, otherwise it is *directed* and the edges are drawn as arrows. A graph is *complete* if all possible edges are

present. A *subgraph* H of a graph G consists of a subset $W \subseteq V$ of the vertex set and an edge set F such that $(i, j) \in F \iff i \in W, j \in W$ and $(i, j) \in E$. A *clique* is a maximally complete subgraph, that is a complete subgraph which is not a subgraph of another complete subgraph.

A *marked graph* has its set of vertices partitioned into two disjoint subsets, $V = \Delta \cup \Gamma$. For graphical modelling, these represent discrete and continuous vertices, respectively. If either set is empty, the graph is *pure*. By convention, discrete variables are represented by *dots* and continuous by *circles*.

Two vertices with an edge between them are *adjacent*. Two or more edges joining the same pair of vertices are called multiple edges. An edge joining a vertex to itself is called a loop. A *simple graph* has no loops or multiple edges. A graph is *connected* if it is in one piece and disconnected otherwise. In other words, a disconnected graph may be partitioned into two or more subgraphs with no edges between them. This thesis is concerned only with simple undirected graphs.

The *degree* of a vertex is the number of edges meeting at it. The *degree sequence* of a graph is the list of the degrees of each vertex, usually written in ascending or descending order.

For any graph, the sum of all the vertex degrees is equal to twice the number of edges. This is known as the Handshaking Lemma.

Two graphs are *isomorphic* if one can be obtained from the other by relabelling the vertices - that is if there is a one-to-one correspondence between the vertices, such that the number of edges joining any pair of vertices in one graph is equal to the number of edges joining the corresponding vertices in the other. Isomorphic graphs share the same *unlabelled graph*.

The *adjacency matrix* of a graph is a $|V| \times |V|$ matrix (where $|V|$ is the number of vertices), with 1's in positions corresponding to adjacent vertices and 0's otherwise.

The *adjacency set* of a vertex α , $\text{adj}(\alpha)$, is the set of all vertices adjacent to it. The *boundary* of a subset $A \subseteq V$ is $\text{bd}(A) = \bigcup_{\alpha \in A} \text{adj}(\alpha) \cap (V \setminus A)$, that is all vertices adjacent to some vertex in A .

A *path* is a sequence, v_0, v_1, \dots, v_n , of distinct vertices with $(v_{m-1}, v_m) \in E$ for all $m = 1, \dots, n$. A *cycle* is a path with $v_0 = v_n$. A *triangulated* graph has no cycles of length $n \geq 4$ without a *chord*, that is, two nonconsecutive vertices with an edge between them.

Two disjoint subsets, A and B are *separated* by a third disjoint subset C if

every path from any vertex in A to any vertex in B passes through C . In this case, C is called a *separator* for A and B .

A vertex is *simplicial* if its adjacency set is complete. In a marked graph, a simplicial vertex is *strongly simplicial* if either it is continuous or its adjacency set consists only of discrete vertices so strongly simplicial vertices have only discrete vertices as neighbours. A subset is simplicial if its boundary is complete and strongly simplicial if it either consists only of continuous vertices or its boundary consists only of discrete vertices.

A *chain* is an ordered sequence of subsets. The chain induces a partial order $<$ on the vertices. The induced *chain graph*, has edge set $E^<$ consisting of all edges (α, β) in E with $\alpha < \beta$. A chain graph has undirected edges within blocks and directed edges between as in Figure 1.1.

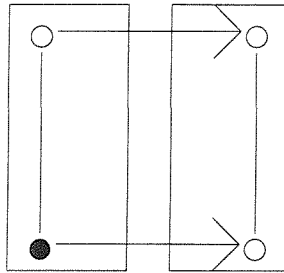


Figure 1.1: Example of a chain graph

An ordering induced by a chain is (strongly) *reducible* if all subsets in the chain are (strongly) simplicial in the induced chain graph. An undirected graph is triangulated if and only if there is a reducible ordering of the vertices.

An undirected graph is *decomposable* if there is a strongly reducible ordering of the vertices. A pure graph is decomposable if and only if it is triangulated.

Nondecomposable pure graphs can thus be recognized as having a chordless cycle of length greater than 3 and nondecomposable marked graphs can be recognized as being either not triangulated or having a continuous vertex with nonadjacent discrete neighbours. The simplest nondecomposable graph is shown below.



A chain graph is *Markov equivalent* to the undirected graph given by removing the blocks and replacing the directed edges by undirected ones if and only if it

is strongly reducible. Thus, the above chain graph is Markov equivalent to the undirected graph in Figure 1.2.

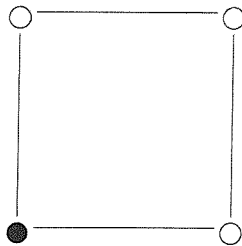


Figure 1.2: Undirected graph equivalent to the chain graph in Figure 1.1

1.3 Graphical Models

A (*conditional*) *independence graph* for a set of variables is a graph with a vertex for each variable and an edge between each pair of vertices corresponding to pairs of variables which are *not* conditionally independent given the remaining variables. Thus, missing edges correspond to pairwise conditional independence constraints, called *graphical constraints*, on the joint distribution of the variables.

Given a set of variables and an independence graph, a *graphical model* is a family of distributions for the variables, satisfying the conditional independence constraints embodied in the graph but otherwise unconstrained. We say that these distributions are *Markov* over the graph. If all edges are present, the graph is *complete* and the model is *saturated*.

Broadly, there are three types of graphical model: *Graphical log-linear models* are used when all the variables are discrete; *graphical Gaussian models* are used when all the variables are continuous; and *mixed graphical (association) models*, usually based on the conditional-Gaussian distribution, are for mixtures of discrete and continuous variables. This thesis is concerned with the latter two cases as the first has been dealt with thoroughly elsewhere.

This thesis is concerned only with *undirected* graphs but there are also models based on directed (or oriented) graphs, sometimes called recursive graphical models, in which there is an ordering of the variables and which can be expressed as a sequence of regressions, and also graphical chain models, in which there is a partial ordering and graphs with both directed and undirected edges are used. These types of models generally require a different approach to the

undirected cases so are not dealt with here. The three main sources for further details of graphical models, especially those not covered in this thesis are Edwards (1995,2000), Whittaker (1990) and Lauritzen (1996), the latter being the most rigorous.

1.3.1 Decomposable Models

Decomposable models are models with decomposable graphs. Such models have received special attention due to their structure allowing analyses to be broken down into sub-analyses of smaller, complete graphs. This thesis is concerned with developing methodology for both decomposable and nondecomposable models.

1.3.2 Model Indexing

Any graphical model is completely specified by the adjacency matrix of its graph. A natural way, therefore, to index models is to use the decimal form of the upper (or lower) triangle of the adjacency matrix, regarded as a binary number. So, for example, when there are four vertices, 43 represents the graph in Figure 1.3. Since

43 is represented as 101011 in binary, the adjacency matrix is
$$\begin{pmatrix} * & 1 & 0 & 1 \\ 1 & * & 0 & 1 \\ 0 & 0 & * & 1 \\ 1 & 1 & 1 & * \end{pmatrix}$$

telling us that the (1, 3) and (2, 3) edges are absent.

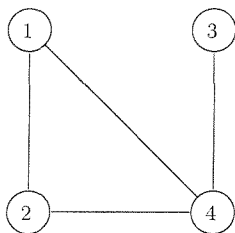


Figure 1.3: Graph 43 for 4 vertices

As an alternative to a numerical labeling like this, a more easily interpretable (for a human reader) labeling is to list the cliques so the above graph is labelled as 124/34.

1.3.3 Graphical Log-linear Models

A graphical log-linear model (GLLM) is a family of discrete probability distributions satisfying certain conditional independence constraints given by a graph, that is *Markov* over the graph. They were first defined by Darroch, Lauritzen and Speed (1980) and, as the name suggests, they are part of the larger and better-known class of log-linear models. These are so-called because the logarithm of the joint probability may be decomposed into a sum of *interaction terms*: $\log p(\mathbf{i}) = \sum_{d \subseteq \Delta} u^d(\mathbf{i})$, where Δ is the set of variables, indexed by \mathbf{i} , which takes values in \mathcal{I} . For example, if $\Delta = \{A, B, C\}$, indexed by i, j and k respectively,

$$\begin{aligned} \log p(i, j, k) &= u + u^A(i) + u^B(j) + u^C(k) + u^{AB}(i, j) + u^{AC}(i, k) \\ &\quad + u^{BC}(j, k) + u^{ABC}(i, j, k). \end{aligned}$$

Note that $u^d(\mathbf{i}) = u^d(\mathbf{i}^d)$ (where \mathbf{i}^d is the subvector of \mathbf{i} corresponding to A), so that, for example, $u^A(i, j, k) = u^A(i)$. $u \equiv u^\emptyset$, is the *log normalizing constant* and since $\sum_{\mathbf{i}} p(\mathbf{i}) = 1$,

$$u = -\log \sum_{\mathbf{i}} \exp \left[\sum_{d \neq \emptyset} u^d(\mathbf{i}) \right].$$

In practice, additional constraints are required in order to uniquely identify these interaction terms. There are various choices for these *identifiability constraints*, including the *sum-to-zero* constraints, which require $\sum_{\mathbf{i}} u^d(\mathbf{i}) = 0$ for each $d \subseteq \Delta$. If A has l_A levels, the sum-to-zero constraints are satisfied by setting $u^A(l_A) = -\sum_{i=1}^{l_A-1} u^A(i)$. Any set of identifiability constraints can be expressed concisely using a *design matrix*, \mathbf{D} :

$$\log \mathbf{P} = \mathbf{D}\mathbf{U},$$

where \mathbf{P} is a vector of $p(\mathbf{i})$'s and \mathbf{U} is a vector of u -terms. For sum-to-zero constraints, $\mathbf{D} = \bigotimes_{\delta \in \Delta} \mathbf{D}(\delta)$, where $\mathbf{D}(\delta)$ is an $l_\delta \times l_\delta$ matrix (where δ has l_δ levels) of form,

$$\mathbf{D}(\delta) = \left(\begin{array}{c|ccc} 1 & & & \\ 1 & & I & \\ \vdots & & & \\ 1 & -1 & \dots & -1 \end{array} \right)$$

where I is an identity matrix.

The advantage of using this interaction expansion is that certain independence relationships may be expressed simply by setting certain sets of interaction terms to zero. In particular, two variables are conditionally independent given the others if and only if all interaction terms involving them are zero. So in the above example,

$A \perp\!\!\!\perp B|C \iff u^{AB}(i, j) = u^{ABC}(i, j, k) = 0$ for all i, j and k . Consequently, different graphical models correspond to certain sets of interaction terms being set to zero. It is easy to see that the total number of possible graphical models for p variables is $2^{\binom{p}{2}} = 2^{\frac{p(p-1)}{2}}$.

The larger class of *hierarchical log-linear models* require only that whenever a particular interaction term is removed (set to zero), so must any others that involve all of the variables that it does. More precisely, if $u^a(\mathbf{i}) = 0$, where $a \subseteq \Delta$, then $u^b = 0(\mathbf{i}) \forall \mathbf{i} \in \mathcal{I}$, whenever $a \subseteq b$. In the above example, this means that if any 2-way interaction is to be removed, so must the 3-way interaction. In this case, there is only one non-graphical hierarchical model - the one with only the three-way interaction removed. In general, there is no expression that gives the number of possible hierarchical models in any case but it is always less than 2^{2^p} since there are 2^p interaction terms.

Bayesian inference for discrete graphical models has been largely dealt with. Dawid and Lauritzen (1993), develop conjugate prior distributions for both these and GGMs, but only for decomposable models. Madigan and York (1995) deal with various approaches including inference under model uncertainty, for directed acyclic graphs as well as decomposable directed graphs. Madigan and Raftery (1994) describe methods for Bayesian inference under model uncertainty. Again, only DAGs and decomposable log-linear models are treated but the methods can also be applied to other types of graphical models. Madigan et al. (1994) also deal with model selection for discrete graphical models. Dellaportas and Forster (1999) apply reversible jump MCMC to log-linear models, both graphical and hierarchical and both decomposable and nondecomposable. Model selection in a non-Bayesian context is dealt with in, for example, Edwards and Kreiner (1983).

1.3.4 Graphical Gaussian Models

Given an independence graph G and a q -dimensional random vector \mathbf{Y} , a graphical Gaussian Model (GGM) is a family of multivariate Normal (or Gaussian) distri-

butions, $P_G = N_q(\boldsymbol{\mu}, \boldsymbol{\Sigma}_G)$, for \mathbf{Y} which are *Markov* over G , that is constrained to satisfy the pairwise conditional independence constraints embodied in the graph. GGMs are more conveniently parameterized by the concentration matrix, $\boldsymbol{\Omega}_G = \boldsymbol{\Sigma}_G^{-1}$, as the entries in this are multiples of partial correlation coefficients. Hence, $Y_i \perp\!\!\!\perp Y_j | Y_{V \setminus \{i,j\}} \iff \omega^{ij} \equiv [\boldsymbol{\Omega}]_{ij} = 0$ and different models correspond to different patterns of zeroes in $\boldsymbol{\Omega}$. In addition, $\boldsymbol{\mu}$ is often taken to be zero but this is not done here. The only other constraint is that $\boldsymbol{\Sigma}_G$ (or $\boldsymbol{\Omega}_G$) be symmetric and positive-definite. The distribution is otherwise arbitrary. There are $2^{\binom{q}{2}}$ models corresponding to the $\binom{q}{2}$ off-diagonal entries in $\boldsymbol{\Omega}$.

Bayesian inference for graphical Gaussian models has largely been confined to decomposable models and/or conjugate priors. Dawid and Lauritzen (1993) introduce a class of conjugate prior distributions for these and show how they may be used. Giudici (1996) introduces two classes of conjugate prior, one of which is the same of that of Dawid and Lauritzen (1993). The other is suitable for both decomposable and nondecomposable GGMs, although the exact density is available only for decomposable models and Giudici restricts his use of it to decomposable models in order to compare the two priors. Dellaportas, Giudici and Roberts (2003) extend the use of Giudici's second prior to nondecomposable GGMs and use MCMC to obtain normalizing constants. Roverato (2002) has introduced a generalization of the conjugate prior of Dawid and Lauritzen, which can be applied to both decomposable and nondecomposable models. Giudici and Green (1999) have used a hierarchical prior, similar to the conjugate priors, in order to implement reversible jump MCMC for GGMs. Further details of each of these are given in the next chapter.

1.3.5 Mixed Graphical Models

Graphical models for both discrete and continuous variables require something not found in pure models, namely association between discrete and continuous variables. One way of achieving this is by allowing the distribution of the continuous variables to depend on the value of the discrete variables so there is a different normal distribution for each cell of the discrete table. The joint distribution is then called conditional Gaussian (CG). Much like with log-linear models, the joint density in this case permits a log-linear expansion involving interaction terms and graphical constraints again correspond to certain interaction terms being set to

zero. Also, as with GLLMs, graphical CG (or interaction) models are part of the larger class of hierarchical CG (interaction) models. Further details are given in Chapter 4.

1.3.6 MIM

MIM, standing for Mixed Interaction Modelling, is an interactive Pascal programme designed for working with hierarchical interaction models, although not in a Bayesian context. It is available free of charge from <http://www.hypergraph.dk/> or from the author, David Edwards, who has also written a guide (Edwards 1987) and describes its use in his book (Edwards 1995,2000). It includes procedures for model selection which are useful for the purposes of comparison with the results presented in this thesis. It can also compute maximum likelihood estimates for decomposable models (this is not possible for nondecomposable models). These are useful as initial values in an MCMC sampler, for generating data and for comparing with Bayesian point estimates based on MCMC output.

1.4 Bayesian Inference

The underlying philosophy of Bayesian inference is that all uncertainty is measured by probability. Data, $\mathbf{y} = \mathbf{y}^1, \dots, \mathbf{y}^n$, are assumed to come from one of a parameterized family of distributions, $f(\mathbf{Y}|\boldsymbol{\theta})$, and, whereas classical statistics considers the parameters, $\boldsymbol{\theta}$, to be fixed but unknown, the Bayesian approach treats them as random variables in their own right. Prior beliefs about the parameters $\boldsymbol{\theta}$ are represented by the prior density function, $f(\boldsymbol{\theta})$. The posterior density function, $f(\boldsymbol{\theta}|\mathbf{y})$, obtained via Bayes' theorem, represents our modified belief about $\boldsymbol{\theta}$ in the light of the data and is given by,

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

$f(\mathbf{y}|\boldsymbol{\theta})$ is the joint distribution of the data \mathbf{y} given $\boldsymbol{\theta}$. Any function $L(\boldsymbol{\theta}|\mathbf{y}) \propto f(\mathbf{y}|\boldsymbol{\theta})$ is called a *likelihood* but since they are the same algebraically, the term likelihood is often used to refer to $f(\mathbf{y}|\boldsymbol{\theta})$. The integral in the denominator can also be written as $f(\mathbf{y})$ and is sometimes called the *marginal likelihood*. It often cannot be obtained analytically but is merely a constant with respect to $\boldsymbol{\theta}$ and is not always required for inference under a fixed model. So $f(\boldsymbol{\theta}|\mathbf{y}) \propto f(\boldsymbol{\theta})f(\mathbf{y}|\boldsymbol{\theta})$, the *unnormalized posterior*.

1.5 Bayesian Model Selection and Inference under Model Uncertainty

The family of distributions to which $f(\mathbf{y}|\boldsymbol{\theta})$ belongs is called a *model*. If we are uncertain about which is the true model, we also express this uncertainty using probability. If \mathcal{M} is the class of all models under consideration, then for each $m \in \mathcal{M}$, we assign a prior model probability $p(m)$. Note the use of $p(\cdot)$ as m is discrete-valued. Typically, a uniform prior is used, that is, $p(m) = 1/|\mathcal{M}|$. Each model m has an associated set of parameters, $\boldsymbol{\theta}_m$, with prior distribution $f(\boldsymbol{\theta}_m|m)$. Each model also implies a distribution for \mathbf{y} , $f(\mathbf{y}|m, \boldsymbol{\theta}_m)$. We can use Bayes' Theorem again to give

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

where

$$f(\mathbf{y}) = \sum_{m \in \mathcal{M}} p(m) \int f(\mathbf{y}|m, \boldsymbol{\theta}_m)f(\boldsymbol{\theta}_m|m)d\boldsymbol{\theta}_m.$$

We can perform parametric inference under any particular model m_0 from

$$f(\boldsymbol{\theta}_{m_0}|m_0, \mathbf{y}) = \frac{f(\mathbf{y}|m_0, \boldsymbol{\theta}_{m_0})f(\boldsymbol{\theta}_{m_0}|m_0)}{f(\mathbf{y}|m_0)},$$

where

$$f(\mathbf{y}|m_0) = \int f(\mathbf{y}|m_0, \boldsymbol{\theta}_{m_0})f(\boldsymbol{\theta}_{m_0}|m_0)d\boldsymbol{\theta}_{m_0}.$$

Posterior model probabilities are given by

$$p(m_0|\mathbf{y}) = \int f(m_0, \boldsymbol{\theta}_{m_0}|\mathbf{y})d\boldsymbol{\theta}_{m_0} = \frac{p(m_0)f(\mathbf{y}|m_0)}{\sum_{m \in \mathcal{M}} p(m)f(\mathbf{y}|m)}.$$

If a uniform prior is used for the models, this becomes

$$\frac{f(\mathbf{y}|m_0)}{\sum_{m \in \mathcal{M}} f(\mathbf{y}|m)}$$

and hence the task reduces to finding the marginal likelihoods, $f(\mathbf{y}|m)$. These are integrals which are often difficult or impossible to obtain analytically and so numerical methods are necessary. Even if the marginal likelihoods are obtainable the number of potential models may be very large, making the task of finding posterior probabilities for all of them impractical. There are $2^{\binom{p}{2}} = 2^{\frac{p(p-1)}{2}}$ potential graphical models for p variables and, while this number is moderate for $p = 3$ (8 models) and possibly $p = 4$ (64 models), for greater values of p it

becomes quite large (1024 for $p = 5$; 32 768 for $p = 6$; 2 097 152 for $p = 7$). Reversible jump Markov chain Monte Carlo, described below, avoid both problems and can be a very efficient way of performing inference under model uncertainty. An alternative, not described in this thesis, is Markov chain Monte Carlo model composition or MC³ (Madigan and York 1995), which involves the construction of a Markov chain with the posterior distribution of the model m , $p(m|\mathbf{y})$ as stationary distribution.

Bayesian model selection is the selection of the model with highest posterior probability. This can be an objective in itself or a preliminary step before adopting the selected model for future analysis. The main reasons (but not the only ones) for model selection are: The selected model may be easier to deal with than a saturated model; fitting a reduced model typically leads to better prediction (see below) than fitting a saturated model, which may be too complex; The selected model may have interesting substantive interpretations in terms of the relationships between variables (conditional independences, for example).

Frequentist (non-Bayesian) model selection methods are generally based on the likelihood or, especially for graphical models on the *deviance*, defined as twice the difference between the maxima of the log-likelihood, l , under the saturated model and under a particular reduced model, m .

$$dev(m) = 2(\max_{\theta} l(\theta) - \max_{\theta} l(\theta|m))$$

This statistic has a number of advantages including the fact that the difference in deviance between two nested models has a chi-squared distribution. Hypothesis tests for edge inclusion or exclusion in graphical models may thus be conducted using the deviance difference. Two commonly-used deviance-based methods of model selection are: (1) Backward elimination, which involves starting with the saturated model and successively removing the edge with the smallest non-significant exclusion deviance, stopping when all exclusion deviances are significant. (2) Forward inclusion, which involves starting with the mutual independence model (with no edges) and successively adding the (currently missing) edge with the highest significant inclusion deviance, stopping when all inclusion differences are not significant.

Such methods often tend to select more complex models (ones with more parameters) than Bayesian model selection. In the context of graphical models, this means deviance-based selection often results in models with more edges than Bayesian approaches. This is one reason to favour Bayesian model selection since simpler models are usually more desirable.

Posterior quantities, such as $f(\boldsymbol{\theta}|\mathbf{y})$, are often obtained using *Bayesian model averaging*, that is by an average over models, weighted by the posterior model probabilities:

$$f(\boldsymbol{\theta}|\mathbf{y}) = \sum_m f(\boldsymbol{\theta}_m|m, \mathbf{y})p(m|\mathbf{y})$$

If we wish to compare two models, m_0 and m_1 , it is usual to obtain the ratio,

$$\frac{p(m_0|\mathbf{y})}{p(m_1|\mathbf{y})},$$

which, under a uniform prior, reduces to a ratio of marginal likelihoods

$$B(m_0, m_1) = \frac{f(\mathbf{y}|m_0)}{f(\mathbf{y}|m_1)},$$

also known as the Bayes factor for m_0 against m_1 .

1.6 Markov Chain Monte Carlo

Markov chain Monte Carlo (MCMC) methods are used to sample from analytically intractable distributions by simulating a Markov chain with the desired distribution (usually the posterior when used for Bayesian inference) as its stationary distribution. Unlike a random sample from this distribution, the MCMC output will not be an independent sample since the use of a Markov chain means that each observation depends on the previous one, but should allow valid inferences to be made. Tierney (1994), for example, discusses the use of MCMC to sample from posterior distributions.

A posterior sample obtained in this way can then be used to perform any sort of inferences we wish. For example, parameter estimates may be obtained as the sample means, posterior model probabilities as sample model frequencies and even entire posterior distributions (joint, marginal or conditional) may be estimated from the MCMC output. In particular, posterior predictive distributions of a subset \mathbf{Y}_1 of the variables given further observations \mathbf{y}_2^{n+1} on the others, $f(\mathbf{Y}_1|\mathbf{y}_2^{n+1}, \mathbf{y})$, where $\mathbf{y} = \mathbf{y}^1, \dots, \mathbf{y}^n$ are the data used to generate the posterior sample, can easily be obtained as follows:

$$\begin{aligned} f(\mathbf{Y}_1|\mathbf{y}_2^{n+1}, \mathbf{y}) &= \sum_m \int f(\mathbf{Y}_1, m, \boldsymbol{\theta}_m|\mathbf{y}_2^{n+1}, \mathbf{y})d\boldsymbol{\theta}_m \\ &= \sum_m \int f(\mathbf{Y}_1|m, \boldsymbol{\theta}_m, \mathbf{y}_2^{n+1}, \mathbf{y})f(\boldsymbol{\theta}_m|m, \mathbf{y})f(m|\mathbf{y})d\boldsymbol{\theta}_m, \end{aligned}$$

which may be estimated from the MCMC output as the average of

$$f(\mathbf{Y}_1|m, \boldsymbol{\theta}_m, \mathbf{y}_2^{n+1}, \mathbf{y}) = f(\mathbf{Y}_1|m, \boldsymbol{\theta}_m, \mathbf{y}_2^{n+1}) \text{ over } m \text{ and } \boldsymbol{\theta}.$$

1.6.1 Gibbs Sampling

Named by its authors Geman and Geman (1984) after the physicist J. W. Gibbs, this is a popular and easy-to-implement method for fixed models. It is useful when direct generation from the posterior is impractical or costly but generation from the conditional distributions is not. Starting with an initial set of values, each parameter is updated in turn, generating from its full conditional distribution and using the most recent values of the other parameters each time. The algorithm to generate from a distribution $f(\boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)$, is as follows:

1. Initialize iteration counter at $j = 1$ and set initial values $\boldsymbol{\theta}^0 = (\theta_1^0, \dots, \theta_p^0)$
2. Obtain $\boldsymbol{\theta}^j = (\theta_1^j, \dots, \theta_p^j)$ from

$$\begin{aligned}\theta_1^j &\sim f(\theta_1|\theta_2^{j-1}, \dots, \theta_p^{j-1}) \\ \theta_2^j &\sim f(\theta_2|\theta_1^j, \theta_3^{j-1} \dots \theta_p^{j-1}) \\ &\vdots \\ \theta_p^j &\sim f(\theta_p|\theta_1^j, \dots, \theta_{p-1}^j)\end{aligned}$$

3. Change j to $j + 1$ and repeat 2 until convergence (or a preset number of iterations is reached).

As the number of iterations increases, the chain reaches the target distribution. Sampling in 2 may be possible directly, otherwise another method such as Metropolis-Hastings is required. A variant is block Gibbs sampling where the parameters are updated in groups.

1.6.2 The Metropolis-Hastings Algorithm

This algorithm (Metropolis et al. 1953) is used when direct generation from (univariate) distribution $f(\theta)$ is not possible. Instead a *proposal* value θ^* is generated from a distribution which is similar in some sense to $f(\theta)$ and depends on the current value of θ . This proposal is then tested for acceptance as the next value of θ .

1. Start with the current value θ^i
2. Generate a proposal value θ^* from some $g(\theta^*|\theta^i)$
3. Calculate

$$\alpha = \min\left(1, \frac{f(\theta^*)g(\theta^i|\theta^*)}{f(\theta^i)g(\theta^*|\theta^i)}\right)$$

4. With probability α accept the proposal i.e.

Generate $u \sim U(0, 1)$

$u < \alpha \longrightarrow \text{set } \theta^{i+1} = \theta^*$

$u > \alpha \longrightarrow \text{set } \theta^{i+1} = \theta^i$

Notice that the normalizing constant of f is not required as it cancels in the ratio. In a Bayesian context, this means that neither the posterior nor the prior densities need be normalized.

The quantity α is called the acceptance probability and the ratio that appears in the expression of α is the test ratio or acceptance ratio. There are various ways of choosing $g(\theta^*|\theta^i)$. The one used throughout this thesis is a random walk i.e. $\theta^* = \theta^i + \epsilon$, where ϵ is normally distributed with mean zero. In this case, $g(\theta^*|\theta^i) = g(\theta^i|\theta^*)$ and $\alpha = \min(1, f(\theta^*)/f(\theta^i))$. The variance of ϵ can be used to control the rate of mixing; Large variance moves the chain around more but with lower acceptance probability whereas small variance gives higher acceptance probability but moves the chain less. Roberts, Gelman and Gilks (1997) show that the asymptotically optimal acceptance rate for random walk chains is approximately a quarter and hence recommend “tuning” (adjusting) the proposal variance so that the acceptance rate is approximately 0.25. This is very simple to do and is done for all examples presented here.

1.6.3 Reversible-Jump MCMC

Reversible jump Markov chain Monte Carlo methods were introduced by Green (1995) to perform inference under model uncertainty using the joint model and parameter space as the target distribution of the chain. The following is a brief summary in the context of Bayesian inference:

Suppose we have two models: M_1 with n_1 parameters and M_2 with n_2 parameters, and let $x = (k, \theta_k)$ $k = 1, 2$, where θ_k is the parameter vector for model k . Denote the probability of choosing a move between from x to x^* as $j(x^*|x)$. This is known as the *jump probability*. To move from M_1 to M_2 , generate a vector of m_1 continuous random variables u_1 , from distribution q_1 , (usually independent of θ_1) then set θ_2 to be a function of θ_1 and u_1 . The reverse move is achieved by generating u_2 of length m_2 from q_2 and θ_1 is obtained from θ_2 and u_2 in such a way as to satisfy both *reversibility* and *dimension matching*. That is, so that

$n_1 + m_1 = n_2 + m_2$,
 $(\theta_2, u_2) = g(\theta_1, u_1)$ and $(\theta_1, u_1) = g^{-1}(\theta_2, u_2)$ for some invertible deterministic function g .

Green shows that for a move from M_1 to M_2 , the acceptance probability is given by

$$\min \left(1, \frac{f(2, \theta_2 | \mathbf{y}) j(2, \theta_2) q_2(u_2)}{f(1, \theta_1 | \mathbf{y}) j(1, \theta_1) q_1(u_1)} \left| \frac{\partial(\theta_2, u_2)}{\partial(\theta_1, u_1)} \right| \right),$$

That is,

$$\min \left(1, \left(\begin{array}{c} \text{Posterior} \\ \text{Ratio} \end{array} \right) \times \left(\begin{array}{c} \text{Jump} \\ \text{Ratio} \end{array} \right) \times \left(\begin{array}{c} \text{Proposal} \\ \text{Ratio} \end{array} \right) \times \text{Jacobian} \right)$$

Often, either m_1 or m_2 is zero so one of the u 's is not needed and one move is made entirely deterministically. The acceptance probability when $m_2 = 0$ then simplifies to:

$$\min \left(1, \frac{f(2, \theta_2) j(2, \theta_2)}{f(1, \theta_1) j(1, \theta_1)} \frac{1}{q_1(u_1)} \left| \frac{\partial(\theta_2)}{\partial(\theta_1, u_1)} \right| \right)$$

The posterior ratio may also be written as

$$\frac{f(2, \theta_2 | \mathbf{y})}{f(1, \theta_1 | \mathbf{y})} = \frac{f(M_2) f(\theta_2 | M_2) f(\mathbf{y} | M_2, \theta_2)}{f(M_1) f(\theta_1 | M_1) f(\mathbf{y} | M_1, \theta_1)},$$

that is, the product of a prior Ratio and a likelihood Ratio.

Often, the model probabilities as well as the jump probabilities are equal so that the general acceptance ratio reduces to

$$\frac{f(\theta_2 | M_2) f(\mathbf{y} | M_2, \theta_2) q_2(u_2)}{f(\theta_1 | M_1) f(\mathbf{y} | M_1, \theta_1) q_1(u_1)} \left| \frac{\partial(\theta_2, u_2)}{\partial(\theta_1, u_1)} \right|$$

The principal difficulty in constructing a reversible jump MCMC scheme is often the generation of suitable proposals for the between-model moves and until recently, there has not been much investigation into this issue. Brooks, Giudici and Roberts (2003) have developed methods of constructing efficient proposals and discuss their application to the scheme of Giudici and Green (1999) for GGMs.

1.6.4 Convergence Diagnostics

When implementing any MCMC method, it is necessary to check for what is known as convergence. There are two aspects to this: The early part of the chain before the target distribution is reached, called the “burn-in”, is usually discarded before inference. This is not absolutely necessary as the influence of the burn-in diminishes as the chain gets longer but discarding it does improve accuracy. It is also necessary to check that the chain is covering the target distribution adequately. The rate at which it does this is known as “mixing”.

There are various methods, convergence diagnostics, to check for convergence. There are methods which involve calculating some measure of convergence but usually graphical methods are used. These usually include time series plots, known as *trace plots* of the output. See pages 56, 60, 111, 112, 113 and 114 for examples of these. Sometimes several chains are run with differing initial states and if the outputs are indistinguishable after suitable burn-in time, convergence is likely. Convergence diagnostics, both formal and graphical, are discussed in detail in, for example, Gelman and Rubin (1992) and Brooks and Gelman (1998).

Testing for convergence within the model space when implementing reversible jump MCMC is more difficult. Methods are described in, for example, Brooks and Giudici (2000) and Castellote and Zimmerman (2002). One method is to split the output into batches (Geyer 1992) and obtain posterior model probabilities, or at least some of them, based on these batches. For any particular model, suppose the model frequency based on the entire chain is P and the batch model frequencies, based on each of b batches, are $\{P_i : i = 1, \dots, b\}$. P is the overall estimate of the posterior model probability and the P_i 's are estimates based on each batch. Time series plots of the batch probabilities may then be produced as a convergence diagnostic. The Markov chain Monte Carlo standard error of the posterior probability estimate P may also be calculated as

$$\frac{\sum_{i=1}^b (P_i - P)^2}{(b-1)\sqrt{b}}.$$

If desired, MCMC standard error of the parameter estimates (obtained as sample averages) may also be obtained in the same way.

Another possibility when dealing with graphical models is time series plots of the number of edges, used as a measure of graphical complexity. All these diagnostics have been used for the examples presented here although not always presented.

1.7 Outline of the Thesis

While conjugate inference for GGM's has been dealt with thoroughly, nonconjugate inference has not, especially for nondecomposable models. The first aim of this thesis is to develop a more general approach avoiding the use of conjugate priors and not restricted to decomposable models. In Chapter 2, a new class of prior distribution is introduced to achieve this but the approach is applicable to other priors, including conjugate priors. In Chapter 3, inference for a single fixed model using a Gibbs sampling scheme is described, followed by a reversible jump MCMC scheme for performing model selection and inference under model uncertainty. A number of examples of application of the reversible jump sampler are then presented in Chapter 3.5.

In contrast, there has been very little work done on Bayesian inference for mixed graphical models. The second major aim of this thesis is to address this issue by attempting a similar approach to that for GGMs, that is to develop MCMC methods both for fixed model inference and inference under model uncertainty for CG models. Due to the complex nature of these types of model, a detailed treatment is given only to those with either one or two discrete variables although a discussion of possible extension to more is given. While the primary interest is in graphical CG models, hierarchical CG models are also dealt with as reversible jump MCMC for these is simpler. Chapter 4 describes CG models in detail and introduces a class of prior distribution for use in Bayesian inference. Chapter 5 deals with the cases of CG models with one discrete variable and Chapter 6 deals with those with two as well as discussing possible treatment of models with more discrete variables.

Chapter 2

Graphical Gaussian Models

2.1 Introduction

Let $\mathbf{Y} = (Y_1, \dots, Y_q)$ be a vector of $q \geq 2$ continuous random variables with association structure described by a conditional independence graph, $G = (V, E_G)$, where $V = \{1, \dots, q\}$.

Graphical Gaussian models are based on the multivariate Normal (or Gaussian) distribution. Its density function is

$$f(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-q/2} |\boldsymbol{\Sigma}|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu})\right]$$

or, equivalently,

$$f(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Omega}) = (2\pi)^{-q/2} |\boldsymbol{\Omega}|^{1/2} \exp\left[-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})' \boldsymbol{\Omega} (\mathbf{y} - \boldsymbol{\mu})\right]$$

and we write $\mathbf{Y} \sim N_q(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

The vector of means, $\boldsymbol{\mu}$, is often set to zero for convenience and the data expressed as deviations from the sample mean as $\boldsymbol{\mu}$ is not usually of interest but this is not done here as it is not really necessary. $\boldsymbol{\Sigma}$ is the variance-covariance matrix and $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ is the inverse-variance, also known as the precision or concentration matrix.

Dempster (1972) first developed the badly-named Covariance selection models, which later became known as graphical Gaussian models, with the aim of parameter reduction. The important step is the use of the precision or concentration matrix to parameterize the normal distribution and the parameter reduction is achieved by setting certain concentrations to zero. He illustrates this by applying a forward selection procedure to select a reduced model for six dimensional

data. He also shows how to obtain estimates of the variance and concentration matrices in such models using an iterative method.

Wermuth (1976) makes a comparison between log-linear models and these covariance selection models. The most important result here is that concentrations are multiples of partial correlation coefficients and therefore zero concentrations correspond to conditional independence relationships between pairs of variables given the rest.

Speed and Kiiveri (1986) discuss the role of the likelihood in GGMs and their specification via marginal distributions.

Σ , and hence Ω , is symmetric and positive-definite and due to the Markov property, missing edges in the graph correspond to zero entries in Ω (Wermuth 1976), that is

$$Y_i \perp\!\!\!\perp Y_j | Y_{V \setminus \{i,j\}} \iff \omega_{ij} = 0$$

where ω_{ij} is the (i, j) th entry of Ω .

The diagonal entries in Ω are partial precisions and will be denoted as τ^2 's, that is

$$\omega_{ii} = \tau_i^2 = [\text{Var}(Y_i | Y_{V \setminus i})]^{-1} \quad \text{for } i = 1, \dots, q$$

Denote by \mathbf{C} the matrix obtained by scaling Ω so that it has unit diagonal. The off-diagonal entries of \mathbf{C} are the negatives of partial correlations and will be denoted by ρ 's, that is

$$c_{ij} = \frac{\omega_{ij}}{\sqrt{\omega_{ii}\omega_{jj}}} = -\rho_{ij} = -\text{Corr}(Y_i, Y_j | Y_{V \setminus \{i,j\}}).$$

Clearly, $\omega_{ij} = 0 \iff c_{ij} = 0$ and \mathbf{C} is positive-definite precisely when Ω is.

Thus we can make the decomposition

$$\Omega = \text{diag}(\boldsymbol{\tau}) \mathbf{C} \text{diag}(\boldsymbol{\tau}), \tag{2.1}$$

where $\boldsymbol{\tau}$ is the vector of square roots of partial precisions, τ_i , and $\text{diag}(\boldsymbol{\tau})$ is the diagonal matrix with $\boldsymbol{\tau}$ as diagonal elements. Using this, we can work directly with the individual parameters, rather than the entire Ω , allowing more flexible prior specification and a simple Gibbs sampler scheme for sampling from the posterior distribution. We can work with either τ 's or τ^2 's but τ is probably easier as it avoids the use of square roots. Likewise, it may be more convenient

to work with c 's than ρ 's but it makes no real difference, at least until it comes to interpretation.

Barnard, McCulloch, and Meng (2000) use a similar decomposition, which they call a “separation strategy”, of the covariance matrix, Σ into standard deviations and correlations:

$$\Sigma = \text{diag}(\boldsymbol{\sigma})\mathbf{R}\text{diag}(\boldsymbol{\sigma}). \quad (2.2)$$

2.2 The Likelihood

Let $\mathbf{y} = (\mathbf{y}^1, \dots, \mathbf{y}^n)$ be a set of n independent observations on \mathbf{Y} and let $\bar{\mathbf{y}} = \frac{1}{n} \sum_{k=1}^n \mathbf{y}^k$ and $\mathbf{S} = \sum_{k=1}^n (\mathbf{y}^k - \bar{\mathbf{y}})'(\mathbf{y}^k - \bar{\mathbf{y}})$ be the observed vector of means and sums-of-products matrix. The likelihood of a graphical Gaussian model is then

$$f(\mathbf{Y}|\boldsymbol{\mu}, \boldsymbol{\Omega}) = (2\pi)^{\frac{nq}{2}} |\boldsymbol{\Omega}|^{\frac{n}{2}} \exp \left[-\frac{1}{2} \sum_{k=1}^n (\mathbf{y}^k - \boldsymbol{\mu})' \boldsymbol{\Omega} (\mathbf{y}^k - \boldsymbol{\mu}) \right] \quad (2.3)$$

$$= (2\pi)^{\frac{nq}{2}} |\boldsymbol{\Omega}|^{\frac{n}{2}} \exp \left[-\frac{1}{2} (\text{tr}(\mathbf{S}\boldsymbol{\Omega}) + n(\boldsymbol{\mu} - \bar{\mathbf{y}})' \boldsymbol{\Omega} (\boldsymbol{\mu} - \bar{\mathbf{y}})) \right] \quad (2.4)$$

$$= (2\pi)^{\frac{nq}{2}} |\boldsymbol{\Omega}|^{\frac{n}{2}} \exp \left[-\frac{1}{2} (\text{tr}(\mathbf{S}\boldsymbol{\Omega}) + n(\boldsymbol{\mu}' \boldsymbol{\Omega} \boldsymbol{\mu} - 2\boldsymbol{\mu}' \boldsymbol{\Omega} \bar{\mathbf{y}} + \bar{\mathbf{y}}' \boldsymbol{\Omega} \bar{\mathbf{y}})) \right] \quad (2.5)$$

$$= (2\pi)^{\frac{nq}{2}} \left(\prod_{j=1}^q \tau_j \right)^n |\mathbf{C}|^{\frac{n}{2}} \exp \left[-\frac{1}{2} \sum_{k=1}^n \sum_{i=1}^q \sum_{j=1}^q c_{ij} \tau_i \tau_j (y_i^k - \mu_i)(y_j^k - \mu_j) \right] \quad (2.6)$$

The mean, $\boldsymbol{\mu}$, is typically set to zero in which case the likelihood can be rewritten as

$$f(\mathbf{Y}|\boldsymbol{\Omega}) = (2\pi)^{\frac{nq}{2}} |\boldsymbol{\Omega}|^{\frac{n}{2}} \exp \left[-\frac{1}{2} \text{tr}(\mathbf{S}\boldsymbol{\Omega}) \right] \quad (2.7)$$

and the data are expressed as deviations from the sample mean.

2.3 Prior Distributions

Conjugate Priors

The standard conjugate prior for Σ in the zero mean saturated model is the Inverse Wishart distribution, being the sampling distribution of the sample covariance matrix. The corresponding distribution for $\boldsymbol{\Omega}$ is the Wishart, with Jacobian $|\boldsymbol{\Omega}|^{q+1}$. The density is

$$f(\Omega) = \frac{|\Omega|^{-\frac{1}{2}(d+q+1)} \exp[-\frac{1}{2} \text{tr}(\mathbf{A}\Omega)]}{2^{dq} \pi^{q(q-1)/4} |\mathbf{A}|^{d/2} \prod_{i=1}^q \Gamma(\frac{1}{2}(d+1-i))},$$

where \mathbf{A} is a positive definite symmetric $q \times q$ matrix and $d > p$ is the degrees of freedom. Note that Giudici(1996) and Roverato(2002) use another, less common, convention which defines the degrees of freedom as $\delta = d - q + 1$.

When the graph is not complete, the kernel remains the same but the normalizing constant will be different for each graph.

Dawid and Lauritzen (1993) introduced the hyper-inverse Wishart distribution which can be used as a conjugate prior for decomposable models. It is the unique hyper-Markov distribution for Σ_G with marginal Wishart distributions for the cliques. This has the advantage that the density can be found analytically however it is unsuitable for nondecomposable models.

Dawid and Lauritzen (1993) define a distribution P to be *Markov* over a given graph if for any decomposition (A, B) of the graph, $A \perp\!\!\!\perp B | A \cap B [P]$, where the notation denotes that the conditional independence is with respect to P . They show this is equivalent to the usual definition of a Markov distribution and also show that if P is Markov over the graph, $A \perp\!\!\!\perp B | S [P]$ whenever S separates A and B .

They first prove the following result for two subsets A and B of the vertex set V :

If distributions Q over A and R over B give the same distribution over $A \cap B$, there is a unique distribution P over $A \cup B$ with $P_A = Q$, $P_B = R$ and P is Markov over the graph. P is constructed as $P_A \equiv Q$ and $P_{B|A} \equiv R_{B|B \cap A}$. P is called the Markov combination of Q and R , denoted QR , and if the density functions of P , Q and R are p , q and r , then

$$p(x) = \frac{q(x_A)r(x_B)}{q_{A \cap B}(x_{A \cap B})} = \frac{q(x_A)r(x_B)}{r_{A \cap B}(x_{A \cap B})}.$$

Hence P is Markov over the graph if and only if

$$p(x) = \frac{p(x_A)p(x_B)}{p_{A \cap B}(x_{A \cap B})}.$$

This is then extended to a general decomposable graph as follows:

If \mathcal{C} is the set of cliques of the graph, perfectly numbered as (C_1, \dots, C_k) and with distributions (Q_1, \dots, Q_k) which are consistent over pairwise intersections

as above, then the unique Markov distribution over the graph with the Q 's as marginal distributions over the cliques is constructed by Markov combinations, $P_{C_{i+1}} = P_{C_i} Q_{C_{i+1}}$, and has density

$$p(x) = \frac{\prod_{C \in \mathcal{C}} p_C(x_C)}{\prod_{S \in \mathcal{S}} p_S(x_S)}$$

This is then extended to distributions for quantities, in effect parameters, with values in the set of Markov distributions over a given graph. These distributions over distributions are termed *laws* and prior and posterior distributions are examples. In the context of GGMs, the quantities in question are covariance matrices, Σ (recall that here the mean is taken to be zero) so this is used for the remainder of this section where θ is used for the more general case in the original paper.

The Markov property of distributions is extended to the *hyper Markov* property of laws by defining a law $L(\Sigma)$ to be (weak) hyper Markov over the given graph if for any decomposition (A, B) of the graph, $\Sigma_A \perp\!\!\!\perp \Sigma_B | \Sigma_{A \cap B}$.

The concept of Markov combination is extended to *hyper Markov combination* as follows: If laws M over A and N over B are *hyperconsistent*, that is they give the same law over $A \cap B$, then the hyper Markov combination is the unique law L over $A \cup B$ with $L_A = M$, $L_B = N$ and $\Sigma_A \perp\!\!\!\perp \Sigma_B | \Sigma_{A \cap B}$. L is constructed as a joint law giving probability one to the event that Σ_A and Σ_B have the same distribution over $A \cap B$. Then $L(\Sigma_A) = M(\Sigma_A)$ and $L(\Sigma_B | \Sigma_A) = N(\Sigma_B | \Sigma_{A \cap B})$.

This is then extended to a law over an entire graph, specified only by the marginal laws for the cliques:

If \mathcal{C} is the set of cliques, perfectly numbered as (C_1, \dots, C_k) with pairwise hyper consistent (as above) laws (M_1, \dots, M_k) , the unique hyper Markov law with the M 's as marginals on the cliques is L , satisfying, $L_{C_1} = M_1$ and $L_{C_{i+1}} L_{C_i} M_{C_{i+1}}$, the hyper Markov combination of L_{C_i} and $M_{C_{i+1}}$.

The following result is then proven: If L is hyper Markov over a graph G , then $\Sigma_A \perp\!\!\!\perp \Sigma_B | \Sigma_S [L]$ whenever S separates A and B in in the graph.

The above is all given in terms of the weak hyper Markov property but they also define a strong hyper Markov property: A law, $L(\Sigma)$ is strong hyper Markov over a given graph if for any decomposition (A, B) , $\Sigma_{B|A} \perp\!\!\!\perp \Sigma_A$.

Dawid and Lauritzen then show that L is strong hyper Markov if and only if $\Sigma_{A|B}$, $\Sigma_{B|A}$ and $\Sigma_{A \cap B}$ are mutually independent under L whenever $A \cap B$ is complete and separates A and B . Also, L is strong hyper Markov if and only if for all cliques C and subsets A of C , $\Sigma_{C \setminus A | A} \perp\!\!\!\perp \Sigma_A [L]$.

An important result is that if a prior law is hyper Markov, so is the posterior and the same goes for strong hyper Markov laws. Hence each family of laws forms a conjugate family. The advantage of strong hyper Markov laws is that they also allow local updating from the following result:

If the prior L is strong hyper Markov, the posterior is the unique strong hyper Markov law L^* specified by the marginal laws for the cliques, $\{L_C^*|C \in \mathcal{C}\}$, with each L_C^* based on L_C and the data for the clique. In terms of densities, $f(\Sigma_C|x) \propto f(\Sigma_C)f(x_C|\Sigma_C)$.

In the context of GGMs, the sampling distribution of the maximum likelihood estimator $\hat{\Sigma}$ given Σ , the Wishart distribution is weak hyper Markov but the inverse-Wishart distribution forms a strong hyper Markov law for Σ and in this context, the distribution is called the *hyper inverse Wishart* distribution (or law). Although not given in the paper, the density is

$$\begin{aligned}
f(\Sigma_G) &= |\Sigma_G|^{-\frac{(d+q+1)}{2}} \exp \left[-\frac{1}{2} \text{tr}(\Phi \Sigma_G^{-1}) \right] \left(\frac{\prod_{C \in \mathcal{C}} |\Phi_C|}{\prod_{S \in \mathcal{S}} |\Phi_S|} \right)^{\frac{d}{2}} \\
&\times \frac{\prod_{C \in \mathcal{C}} |\Sigma_C|^{q-|C|}}{\prod_{S \in \mathcal{S}} |\Sigma_S|^{q-|S|}} \\
&\times \frac{\prod_{S \in \mathcal{S}} |\Phi_S|^{\frac{q-|S|}{2}} w(d-q+|S|, |S|)}{\prod_{C \in \mathcal{C}} |\Phi_C|^{\frac{q-|C|}{2}} w(d-q+|C|, |C|)} \tag{2.8}
\end{aligned}$$

where \mathcal{C} is the set of *perfectly numbered* cliques of G , \mathcal{S} is the set of their separators, Φ is a positive definite symmetric $q \times q$ matrix, d is the degrees of freedom, $|\cdot|$ denotes the size of a set, A_U denotes the submatrix of matrix A corresponding to $U \subseteq V$ and for $\alpha > k$ with k a positive integer,

$$w(\alpha, k) = \left[2^{\frac{\alpha k}{2}} \pi^{\frac{k(k-1)}{4}} \prod_{i=1}^k \Gamma \left(\frac{\alpha + 1 - i}{2} \right) \right]^{-1}.$$

A final result which is important for model comparison is an expression for the marginal likelihood,

$$f(\mathbf{y}) = \frac{\prod_{C \in \mathcal{C}} f(\mathbf{y}_C)}{\prod_{S \in \mathcal{S}} f(\mathbf{y}_S)},$$

where $f(\mathbf{y}_C)$ is the marginal likelihood for the clique C , which is a matrix- t distribution, $T(d|I_n, \Phi_C)$, with density,

$$f(\mathbf{y}_C) = \pi^{\frac{-n|C|}{2}} \frac{\prod_{i=1}^{|C|} \Gamma \left(\frac{d-q+n+|C|+1-i}{2} \right)}{\prod_{i=1}^{|C|} \Gamma \left(\frac{d-q+|C|+1-i}{2} \right)} \frac{|\Phi_C|^{-\frac{n}{2}}}{|I_n + \mathbf{y}_C \Phi_C^{-1} \mathbf{x}'_C|^{\frac{d-q+n+|C|}{2}}}. \tag{2.9}$$

Giudici (1996) considers two classes of conjugate prior distribution for zero mean decomposable GGM's. The first, which he calls “local”, is just the HIW distribution of Dawid and Lauritzen (1993). The second, which he calls “global” is derived by conditioning the inverse Wishart distribution on a given set of graphical constraints, that is conditioning certain entries of Ω to be zero. This can be derived for *any* GGM but the normalizing constants can only be obtained analytically for decomposable models. He derives the global density for a decomposable model with graph G as

$$f(\Sigma_G) = |\Sigma_G|^{-\frac{(d+q+1)}{2}} \exp\left[-\frac{1}{2} \text{tr}(\Phi \Sigma_G^{-1})\right] \left(\frac{\prod_{C \in \mathcal{C}} |\Phi_C|}{\prod_{S \in \mathcal{S}} |\Phi_S|}\right)^{\frac{d}{2}} \\ \times \frac{\prod_{C \in \mathcal{C}} |\Phi_C|^{\frac{q-|C|}{2}} w(d+q-|C|, |C|)}{\prod_{S \in \mathcal{S}} |\Phi_S|^{\frac{q-|S|}{2}} w(d+q-|S|, |S|)}$$

Note that d is not the same degrees of freedom used by Giudici, which is $\delta = d - q + 1$.

To obtain the posterior density for either prior, substitute $\Phi + \mathbf{y}\mathbf{y}'$ for Φ and $d + n$ for d .

He also derives the marginal likelihood as

$$f(\mathbf{y}) = \frac{\prod_{C \in \mathcal{C}} f(\mathbf{y}_C)}{\prod_{S \in \mathcal{S}} f(\mathbf{y}_S)}$$

just as for the local prior, where, $f(\mathbf{x}_A)$ is a matrix- t distribution as in 2.9.

Using these, he is able to perform model selection and does so for Fret's heads data (see section 3.5) but restricts the model space to include only those models with two particular edges present and exclude the three nondecomposable models. A comparison of his results with those of the reversible jump sampler described in this thesis is made in section 3.5.

Dellaportas, Giudici and Roberts (2003) extend the use of Giudici's global prior to the entire class of GGMs by providing an importance sampling method to calculate marginal likelihoods for nondecomposable models.

Roverato (2002) derives a generalization of the hyper-inverse Wishart, applicable to both decomposable and nondecomposable models. It is an example of a so-called “DY conjugate prior”, after Diaconis and Ylvisaker (1979). These are conjugate priors induced by standard conjugate families of canonical parameters, the prior induced on Σ_G by the conjugate prior on Ω_G in the case of GGMs. The

DY-conjugate prior in the saturated model is the inverse Wishart distribution and in the case of a decomposable model, it is the hyper inverse Wishart distribution. Roverato shows that in the case of an arbitrary GGM with graph G , the density for Σ_G can be written as (with notation changed to agree with that used already),

$$f(\Sigma_G|d, \Phi_G) = h_G(d, \Phi_G) 2^{q/2} |\text{Iss}(\Phi_G)|^{1/2} \times |\text{Iss}(\Sigma_G)|^{-1} \frac{|\Sigma^*|^{-(d-q-1)/2}}{|\Phi^*|^{-(d-q-1)/2}} \exp \left[-\frac{1}{2} \text{tr}(\Sigma^{*-1} \Phi^*) \right]. \quad (2.10)$$

$h_G(d, \Phi_G)$ is a normalizing constant (note that $2^{q/2} |\text{Iss}(\Phi_G^*)|^{1/2}$ is not absorbed into this), Σ^* and Φ^* are called the *PD-completions* of Σ_G and Φ_G and $\text{Iss}(\cdot)$ denotes an Isserlis matrix.

The Isserlis matrix $\text{Iss}(A)$ of a positive definite $q \times q$ matrix A is a matrix, indexed by pairs of edges, with $[\text{Iss}(A)]_{(i,j),(r,s)} = a_{ir}a_{js} + a_{is}a_{jr}$. The determinant is $|\text{Iss}(A)| = 2^q |A|^{q+1}$.

Σ_G and Φ_G are *incomplete matrices* as only certain entries, corresponding to edges in the graph, are specified. For example, if G is the the graph in 1.3, Σ_G is

$$\begin{pmatrix} \sigma_{11} & \sigma_{12} & * & \sigma_{14} \\ \sigma_{21} & \sigma_{22} & * & \sigma_{24} \\ * & * & \sigma_{33} & \sigma_{34} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_{44} \end{pmatrix},$$

where $*$ denotes an unspecified entry.

The PD-completion of a matrix A_G , where G is a given graph, is the unique positive definite matrix A^* which has $[A^*]_{ij} = [A]_{ij}$ for specified entries of A (that is whenever $(i, j) \in E$) and $[(A^*)^{-1}]_{ij} = 0$ otherwise.

The normalizing constant when G is not decomposable is obtained using importance sampling.

The marginal likelihood is also given as

$$f(\mathbf{y}) = (2\pi)^{nq/2} \frac{h_G(d, \Phi_G) |\text{Iss}(\Phi)_G|^{1/2} |\Phi|^{(d-q-1)/2}}{h_G(d+n, \Phi_G + \mathbf{S}) |\text{Iss}(\Phi + \mathbf{S})_G|^{1/2} |\Phi + \mathbf{S}|^{(d-q-1+n)/2}}.$$

As in this thesis, Giudici and Green (1999) deal with reversible jump MCMC for GGM's but they confine themselves to decomposable zero-mean models and use Hyper-Inverse Wishart priors. They propose two types of model. The first has a fixed value of the parameters of the HIW prior. To simplify the task of

specifying all the entries of the matrix, they propose an “intra-class correlation structure”, which gives the matrix parameter as $\tau[\rho J + (1 - \rho)I]$, where I is an identity matrix and J is a matrix of 1’s. The second is a so-called hierarchical model with prior distributions assigned to both the degrees of freedom and matrix parameters of the HIW prior. These would sometimes be called *hyperpriors* although not by these authors. The degrees of freedom is given a gamma prior, as it is a positive quantity, with parameters chosen to make the prior uninformative. They then consider two possibilities for the matrix (hyper)parameter: The first is a conjugate (hyper)prior, which is a Wishart distribution and suggest a single degree of freedom and a diagonal matrix parameter. The second uses the intra-class structure and assigns priors to τ and ρ , based on a Wishart distribution.

Nonconjugate Priors and a New Class of Prior for GGMs

One advantage of using the decomposition 2.1 is that separate, and possibly independent, priors may be specified for each individual parameter, rather than for the precision matrix as a whole .

The following alternative set of priors avoids the difficulties of using a conditioned Wishart prior as, while normalizing constants must still be obtained numerically, they are relatively easy to obtain.

Since the partial precisions are only required to be positive, a simple choice is independent gamma priors for each, as independent inverse-gamma priors are commonly used as conjugate priors for variances. This is because the inverse-gamma distribution is often the sampling distribution of a sample variance.

$$f(\tau_j^2|G_i) = \frac{\beta^\alpha}{\Gamma(\alpha)}(\tau_j^2)^{\alpha-1} \exp(-\beta\tau_j^2), \quad \alpha, \beta > 0 \quad j = 1, \dots, q \quad (2.11)$$

Equivalently, the corresponding prior for τ_j may be used:

$$f(\tau_j|G_i) = 2 \frac{\beta^\alpha}{\Gamma(\alpha)}(\tau_j)^{2\alpha-1} \exp(-\beta\tau_j^2), \quad \alpha, \beta > 0 \quad j = 1, \dots, q \quad (2.12)$$

An alternative is the prior used by Barnard, McCulloch, and Meng (2000) for standard deviations,

$$\log(\boldsymbol{\tau}) \sim N(\boldsymbol{\xi}, \Lambda),$$

which allows for a priori dependence, although it would be more usual to have Λ diagonal, giving independent lognormal distributions for each.

q	Proportion PD
3	0.617
4	0.183
5	0.022
6	0.001
7	$< 10^4$
8	$< 10^6$

Table 2.1: Acceptance rates for estimating prior normalizing constants for saturated models using a rejection sampler

The partial correlations are required to lie in $(-1, 1)$ so a natural choice of prior is uniform over this interval for each. However, these marginal priors must be conditioned on the positive definiteness constraint on \mathbf{C} . This has the effect of changing the normalizing constant for the joint prior, $f(\mathbf{C}|G_i)$, and the marginals will no longer be uniform. In other words,

$$f(\mathbf{C}|G_i) = k_i \left(\frac{1}{2}\right)^{|E_i|}, \quad (2.13)$$

where $|E_i|$ is the number of edges present in G_i .

k_i is the relative size of positive definite space within $(-1, 1)^{|E_i|}$ and so can easily be estimated by a rejection method as follows: Generate a large number of matrices with unit diagonal, off-diagonal entries drawn from $U(-1, 1)$ and zeroes in the positions corresponding to missing edges in G_i . The proportion of these that are positive definite, the acceptance rate, is approximately $1/k_i$.

For any given q , this proportion is smallest for the saturated model as it has the maximum number of nonzero off-diagonal entries. These proportions, based on 1 000 000 matrices, for saturated models are tabulated in Table 2.1 for values of q up to 8. Standard errors are not quoted although they are all approximately 0.001. It can be seen from this table that $q = 6$ is the practical upper limit for computing these normalizing constants. This is not unique to this particular choice of prior; the same limitation will apply to any non-conjugate priors due to the positive definite constraints. However, it is worth noting that this does not prevent their use for inference for fixed models where normalizing constants are not generally required.

The number of simulations required can be reduced by exploiting the fact that isomorphic models (those with isomorphic graphs) will yield the same constant.

As shown in Table 2.2, this gives a considerable reduction but it is necessary to identify all unlabelled graphs with q vertices.

q	Graphs	Unlabelled Graphs
3	8	4
4	64	11
5	1024	32
6	32768	156

Table 2.2: Numbers of labeled and unlabelled graphs with up to 6 vertices

One systematic way of doing this is as follows: Begin with the (unlabelled) graph with no edges. There is only one (unlabelled) graph with a single edge and correspondingly, one with $\binom{q}{2} - 1$ edges. There are two (unlabelled) graphs with two edges and hence two with $\binom{q}{2} - 2$ edges. Now obtain all the distinct unlabelled graphs with three edges, and hence those with $\binom{q}{2} - 3$ edges, by considering all possible ways to add an edge to the two-edge graphs. Continue in this fashion until all possible unlabelled graphs have been obtained. That all have been obtained can be verified by using combinatoric methods to find the number of graphs with each structure and checking that they sum to $2^{\binom{q}{2}}$. Unlabelled graphs and their corresponding normalizing constants (that is, $f(\mathbf{C}|G_i)$) are tabulated in appendix A for $q \leq 6$.

Another way of thinking about this prior, $f(\mathbf{C}|G_i)$, is that it is uniform over $(-1, 1)^{|E_i|}$ and hence the density is the reciprocal of the (hyper)volume of positive definite space in $|E_i|$ dimensions, which is just (2.13).

Rousseeuw and Molenberghs (1994) explore the shape of the space of correlation matrices, which of course is the same as the space of partial correlation matrices, for three and four variables and provide some intriguing graphical representations. For three variables (and hence three correlations) the space resembles an inflated tetrahedron. They note that the volume of this space can be calculated using calculus as $\pi^2/2$. Hence the relative volume of the space within $[-1, 1]^3$ is $\pi^2/16 \approx 0.617$, which agrees with the empirical value in Table 2.1. Similarly, for four variables (and hence six correlations), the relative volume of the space within $[-1, 1]^6$ is 0.183, again agreeing with the empirical value in the table.

Barnard, McCulloch, and Meng (2000) describe the same prior for the corre-

lation matrix, \mathbf{R} , in the decomposition 2.2 of Σ , introduced as a jointly uniform prior. Investigating the form of the marginal distributions, which are not univariate uniform, they note that as the number of variables, q , increases, these marginal distributions become more concentrated about zero. Since this seems to imply that this jointly uniform prior is informative by keeping the posterior away from the corners of the space, they simulated two sets of data to investigate, one with common correlation of 0.5 and one with common correlation of 0.91, for the cases of both $q = 3$ and $q = 10$. They found that although the marginal posterior distribution for an individual correlation was centred closer to zero for $q = 10$, the true value was well within the posterior mass in each case. They also found that although the priors are concentrated for large correlations, particularly for $q = 10$, due to the shape of the space, the posterior is still pushed by the likelihood towards the large values. Their conclusion is that any informative nature of the prior is tolerable if q is not too large. Since q will be taken to be no larger than 6 here, as discussed above, there will be no problem as long as the amount of data is sufficiently large.

This jointly uniform prior is contrasted by Barnard et al with a prior with uniform marginal densities for each correlation, derived from the inverse Wishart distribution. The density function of the joint distribution is derived as

$$f(\mathbf{R}) \propto |\mathbf{R}|^{\frac{1}{2}(d+q+1)} \left(\prod_i |\mathbf{R}_{ii}| \right)^{-\frac{d}{2}}.$$

The marginal distribution of any $k \times k$ submatrix is obtained by replacing q by k and d by $d - (q - k)$ in this density. Taking $k = 2$ gives the marginal density for an individual correlation, r , as

$$f(r) \propto (1 - r^2)^{\frac{d-q-1}{2}},$$

which is uniform when $d = q + 1$. Furthermore, different values of d allow the tails of the distribution to be lighter or heavier than uniform.

This only covers the case of a saturated model but readily extends to the case of an arbitrary GGM and either of the Inverse-Wishart-based priors.

A suitable prior for the mean, $\boldsymbol{\mu}$, if it has not been set to zero, is $N(\mathbf{0}, \mathbf{U})$. Taking $\mathbf{U} = \sigma \mathbf{I}_q$ gives independent univariate normal priors for each μ and taking large σ , say 10^6 , makes them suitably noninformative.

Unless there is reason to assume otherwise, a uniform prior is a sensible choice

for G :

$$P(G_i) = 1/2^{\binom{q}{2}} \quad 0 \leq i \leq 2^{\binom{q}{2}}.$$

There are two recent approaches to prior specification for correlations or partial correlations, published since the work in this thesis was done:

Wong, Carter and Kohn (2003) decompose the precision matrix in a GGM as in this thesis and also use a gamma prior for the partial precisions. They define a quantity J to be a set of edge presence indicators (as in an adjacency matrix) and assign a prior to it. The prior for the partial correlation matrix, C , is then defined conditionally on J and involves the volume of the space of correlation matrices given a set of graphical constraints. The MCMC scheme they then describe to generate from the posterior allows partial correlations to be zero and involves determining the intervals in which partial correlations must lie as in Barnard et al. (2000) and this thesis.

Liechty, Liechty and Muller (2004) also follow the separation strategy of Barnard et al. (2000), that is they decompose a variance matrix as $\Sigma = \mathbf{SRS}$. A normal prior, subject to a constraint of positive definiteness, is used for the correlations in R and hyperpriors are assigned to the parameters of these normal distributions. Three types of model are then considered: The first uses a common normal prior for all correlations, the second allows the correlations to be grouped with different prior means and variances for each group and the third allows the *variables* to be grouped, in which case the priors depend on the grouping.

They then describe a Metropolis-Hastings scheme to sample from the posterior in each case, which also involves determination of the intervals in which the correlations must lie.

Chapter 3

Markov Chain Monte Carlo Methods for GGMs

3.1 MCMC for fixed models

Consider first the case of parametric Bayesian inference for a fixed model and the use of MCMC to generate from the posterior distribution specified by the priors described above. Initial values must first be chosen for the parameters, $\theta = (\boldsymbol{\mu}, \tau_1, \dots, \tau_q, \mathbf{C})$. Suitable values may be the observed values or, as a default, zero mean, unit partial precisions and partial correlations equal and nonzero but small. Note that prior normalizing constants are not required for inference based on a single fixed model.

A Gibbs Sampling scheme can be used to update the parameters in turn, so each step of the chain consists of

1. Update $\boldsymbol{\mu}$
2. Update each τ in turn
3. Update each nonzero ρ in turn

In each case sampling is from the conditional distribution.

Updating $\boldsymbol{\mu}$ In this case direct sampling is possible as the conditional distribution can be obtained exactly as

$$f(\boldsymbol{\mu}|\boldsymbol{\Omega}, \mathbf{y}) = N(\boldsymbol{\xi}, \mathbf{V})$$

where $\mathbf{V}^{-1} = (\mathbf{U}^{-1} + n\boldsymbol{\Omega})$ and $\boldsymbol{\xi} = n\mathbf{V}\boldsymbol{\Omega}\bar{\mathbf{Y}}$.

Updating $\boldsymbol{\tau}$ The conditional for $\boldsymbol{\tau}$ is only available up to a normalizing constant, so it is necessary to use a “Metropolis-within-Gibbs” step, that is use the Metropolis-Hastings algorithm to generate new values. A suitable way to generate proposal values is to use a random walk, that is add a random increment, ϵ , to the current value where $\epsilon \sim N(0, \sigma)$ with suitably sized σ . Of course, different proposal variances may be used for each τ_j , if desired. As each τ_j must be positive, all negative proposals must be rejected. Alternatively, any continuous distribution over the positive real numbers may be used as a proposal distribution. Another method is to use a griddy Gibbs sampler (Ritter and Tanner 1992) like Barnard et al. but the random walk is much simpler to implement. A further possibility is implement the random walk on the log scale, that is with proposal values given by $\exp(\log(\tau) + \epsilon)$.

The conditional for τ_j is given by

$$f(\tau_j | \dots) \propto (\tau_j)^{2\beta-1+n} \exp(-\beta\tau_j^2) \exp\left(-\frac{1}{2} \sum_k (\mathbf{y}^k - \boldsymbol{\mu})' \boldsymbol{\Omega} (\mathbf{y}^k - \boldsymbol{\mu})\right) \quad j = 1, \dots, q$$

and a proposed $\tau_j^* = \tau_j + \epsilon$ is accepted with probability

$$\min\left(1, \frac{(\tau_j^*)^{2\beta-1+n} \exp(-\beta\tau_j^{*2}) \exp\left[-\frac{1}{2} \sum_{k=1}^n (\mathbf{y}^k - \boldsymbol{\mu})' \boldsymbol{\Omega}^* (\mathbf{y}^k - \boldsymbol{\mu})\right]}{(\tau_j)^{2\beta-1+n} \exp(-\beta\tau_j^2) \exp\left[-\frac{1}{2} \sum_{k=1}^n (\mathbf{y}^k - \boldsymbol{\mu})' \boldsymbol{\Omega} (\mathbf{y}^k - \boldsymbol{\mu})\right]}\right),$$

provided $\tau_j^* > 0$ and where $\boldsymbol{\Omega}^*$ is obtained by replacing τ_j with τ_j^* in $\boldsymbol{\Omega}$.

Updating \mathbf{C} Again, the conditional is only available up to a normalizing constant so it is necessary to use a “Metropolis-within-Gibbs” step. A random walk is also a suitable way to generate proposals here but it is generally not necessary to have different proposal variances for each correlation. In fact, from experience, a variance of around 0.1 is generally suitable for each implementation. This is likely due to the narrow range in which the correlations take values. Care must be taken to ensure that \mathbf{C} , and hence $\boldsymbol{\Omega}$, always has zeroes in the correct positions and is positive definite. For the first, simply do not update zeroes. For the latter, simply reject any proposal that is not positive definite.

An alternative method of generating proposals that ensures \mathbf{C} remains positive definite is described by Barnard et al. (2000): Since the current \mathbf{C} is already positive definite and only one entry at a time is being changed, a necessary and sufficient condition for the proposal \mathbf{C}^* to be positive definite is that $|\mathbf{C}^*| > 0$.

If ρ is the correlation being updated, then $|\mathbf{C}^*|$ is a quadratic in ρ , $a\rho^2 + b\rho + c$, with coefficients $a = \frac{1}{2}[f(1) + f(-1) - 2f(0)]$, $b = \frac{1}{2}[f(1) - f(-1)]$ and $c = f(0)$,

where $f(\rho^*) = |\mathbf{C}^*|$ when ρ has been changed to ρ^* . The roots of this quadratic are the endpoints of the interval from which to draw a ρ^* that gives a positive definite \mathbf{C}^* . The natural proposal distribution is uniform over this interval.

However, like most independence proposals, this tends to result in low acceptance rates, whereas the variance of the random walk proposals may be tuned for optimal acceptance rates.

The conditional for ρ_{ij} is

$$f(\rho_{ij} | \dots) \propto |\mathbf{C}|^{n/2} \exp \left(\rho_{ij} \tau_i \tau_j \sum_k (y_i^k - \mu_i)(y_j^k - \mu_j) \right) \quad (i, j) \in E$$

and a proposed $\rho_{ij}^* = \rho_{ij} + \epsilon$ is accepted with probability

$$\begin{aligned} & \min \left(1, \frac{|\mathbf{C}^*|^{n/2} \exp \left[-\frac{1}{2} \sum_{k=1}^n (\mathbf{y}^k - \boldsymbol{\mu})' \boldsymbol{\Omega}^* (\mathbf{y}^k - \boldsymbol{\mu}) \right]}{|\mathbf{C}|^{n/2} \exp \left[-\frac{1}{2} \sum_{k=1}^n (\mathbf{y}^k - \boldsymbol{\mu})' \boldsymbol{\Omega} (\mathbf{y}^k - \boldsymbol{\mu}) \right]} \right) \\ & = \min \left(1, \frac{|\mathbf{C}^*|^{n/2} \exp \left[\rho_{ij}^* \tau_i \tau_j \sum_k (y_i^k - \mu_i)(y_j^k - \mu_j) \right]}{|\mathbf{C}|^{n/2} \exp \left[\rho_{ij} \tau_i \tau_j \sum_k (y_i^k - \mu_i)(y_j^k - \mu_j) \right]} \right), \end{aligned} \quad (3.1)$$

provided $-1 < \rho_{ij}^* < 1$ and \mathbf{C}^* is positive definite, where \mathbf{C}^* and $\boldsymbol{\Omega}^*$ are obtained by replacing ρ_{ij} with ρ_{ij}^* in \mathbf{C} and $\boldsymbol{\Omega}$ respectively.

Statistical Performance

Despite the use of one-at-a-time updating, the Markov chain is generally quite fast to converge and produces draws with relatively low autocorrelations. Of course, if convergence is slow due to dependence, it is quite simple to implement block updating but it is more difficult to ensure positive definite proposals. To assess convergence, trace plots are produced for each parameter as well as time series plots of batch means. Burn-in is usually quite short, even when the initial values are some distance from the centre of the posterior (or the true values, when these are known). Very often as few as 10 000 iterations are sufficient for convergence and run times are very short, generally less than 1 minute for moderate q .

The variance of the random walk proposals must be chosen so as to ensure satisfactory mixing of the chain. In general, the chain reaches convergence relatively quickly for most moderate proposal variances but after some experimentation, 0.01 to 0.2 was found to be a good range with a default value of 0.1. Of course it is a simple matter to tune the random walk variances to improve mixing, if necessary.

3.2 Reversible Jump MCMC for GGM's

The reversible jump MCMC scheme described here consists of iterations of the following steps:

1. Update the model by adding or removing one edge from the graph. Doing this will require either setting a partial correlation to zero or generating a new one.
2. Update the parameters in the new model. This is the same as an update step for a fixed model so the same methods as described in the previous section can be applied.

The second step is performed at every iteration but it could be performed less frequently, if required for efficiency.

3.2.1 Between-model Moves

Only moves between ‘neighbouring’ graphs i.e. ones which differ in exactly one edge are considered. Suppose the current graph is G , the current set of parameters is $\boldsymbol{\theta}_G = (\boldsymbol{\mu}, \tau_1, \dots, \tau_q, \mathbf{C})_G$ and the current state is $x = (G, \boldsymbol{\theta}_G)$. A random pair of distinct vertices, (i, j) , is drawn. If edge (i, j) is in G , it is proposed to remove it; otherwise it is proposed to add it. In the latter case, a new ρ_{ij} for the proposed model is required and this is drawn directly from either $U(-1, 1)$, rejecting any proposed move resulting in a \mathbf{C} that is not positive definite, or from a uniform distribution over the interval with endpoints (a, b) chosen to preserve positive definiteness, as previously described. The latter is clearly better as it avoids rejections due to non positive definiteness and hence makes for better mixing within the model space. Thus $q(u) = 1/(b - a)$.

Note that this method of generating proposals may be used regardless of which prior is used.

Since the current graph, G , and the proposed graph, G^* , differ in exactly one edge,

$$j(x|x^*) = j(x^*|x) = 1/\binom{q}{2}$$

and the jump ratio is equal to 1.

The posterior ratio expands as a product of likelihood and prior ratios:

$$\frac{f(x^*)}{f(x)} = \frac{f(\mathbf{Y}|G^*, \boldsymbol{\theta}_{G^*})f(\boldsymbol{\theta}_{G^*}|G^*)p(G^*)}{f(\mathbf{Y}|G, \boldsymbol{\theta}_G)f(\boldsymbol{\theta}_G|G)p(G)}.$$

The graphs are assumed to be equally likely *a priori* so that $p(G^*) = p(G)$.

Since the partial precision parameters are not affected by a change of model,

$$\frac{f(\boldsymbol{\theta}_{G^*}|G^*)}{f(\boldsymbol{\theta}_G|G)} = \frac{f(\mathbf{C}_{G^*}|G^*)}{f(\mathbf{C}_G|G)},$$

which is a ratio of prior normalizing constants.

The likelihood ratio reduces to

$$\frac{f(\mathbf{Y}|G^*, \boldsymbol{\theta}_{G^*})}{f(\mathbf{Y}|G, \boldsymbol{\theta}_G)} = \left(\frac{|\mathbf{C}^*|}{|\mathbf{C}|} \right)^{\frac{n}{2}} \frac{\exp \left[-\frac{1}{2} (\text{tr}(\mathbf{S}\boldsymbol{\Omega}^*) + n(\boldsymbol{\mu} - \bar{\mathbf{y}})' \boldsymbol{\Omega}^* (\boldsymbol{\mu} - \bar{\mathbf{y}})) \right]}{\exp \left[-\frac{1}{2} (\text{tr}(\mathbf{S}\boldsymbol{\Omega}) + n(\boldsymbol{\mu} - \bar{\mathbf{y}})' \boldsymbol{\Omega} (\boldsymbol{\mu} - \bar{\mathbf{y}})) \right]} \quad (3.2)$$

$$= \left(\frac{|\mathbf{C}^*|}{|\mathbf{C}|} \right)^{\frac{n}{2}} \exp \left(\rho_{ij} \tau_i \tau_j \sum_{k=1}^n (y_i^k - \mu_i)(y_j^k - \mu_j) \right). \quad (3.3)$$

Finally, the Jacobian is equal to 1 and the acceptance ratio is

$$(b-a) \left(\frac{|\mathbf{C}^*|}{|\mathbf{C}|} \right)^{\frac{n}{2}} \exp \left(\rho_{ij} \tau_i \tau_j \sum_{k=1}^n (y_i^k - \mu_i)(y_j^k - \mu_j) \right) \frac{f(\mathbf{C}_{G^*}|G^*)}{f(\mathbf{C}_G|G)}. \quad (3.4)$$

If edge (i, j) is proposed for deletion, set $\rho_{ij} = 0$ in the proposed model and the acceptance ratio is

$$\frac{1}{(b-a)} \left(\frac{|\mathbf{C}^*|}{|\mathbf{C}|} \right)^{\frac{n}{2}} \exp \left(-\rho_{ij} \tau_i \tau_j \sum_{k=1}^n (y_i^k - \mu_i)(y_j^k - \mu_j) \right) \frac{f(\mathbf{C}_{G^*}|G^*)}{f(\mathbf{C}_G|G)}. \quad (3.5)$$

Note that, unlike within-model moves, the prior normalizing constants are required. For the proposed priors, these constants are the priors for \mathbf{C} and can be estimated as described in Section 2.3. For a conditioned Wishart prior, they are only available directly when the model is decomposable.

These estimates of prior normalising constants are subject to simulation error, which may propagate into the posterior as estimated via the MCMC sample. However, provided that these errors are within satisfactory tolerance, they will not cause further significant error in the posterior model probabilities. In particular, Bayes factors will not be affected since the marginal likelihoods are proportional to the prior and hence are subject to the same relative error.

As the number of potential models can be very large, an efficient method of inputting the prior normalizing constants in the acceptance ratios is desirable. Recall that models with the same unlabelled graph have the same constant. One method found to be useful is to identify the unlabelled graph (and hence the corresponding constant) by means of

1. The number of edges.
2. The determinant of the adjacency matrix with 2's as diagonal entries. This measure is clearly the same for isomorphic models.
3. The degree vector of the graph, obtained from the row (or column) sums of the adjacency matrix.

When $q = 3$, only the first is required; when $q = 4$, the first two suffice; when $q = 5$ or $q = 6$, all three are required in relatively few cases. Tables of unlabelled graphs along with these three measures and their prior normalizing constants are given in appendix A.

3.2.2 Within-model Moves

To complete each iteration, the parameters are updated using a Gibbs sampler as described in the previous section.

3.2.3 Variations

Notice that the number of observations, n , the observed means, $\bar{\mathbf{y}}$, and the observed sums of products matrix, \mathbf{S} , are sufficient to summarize the data. It was noted already that the means are not generally of interest and may be set to zero. In this case the alternative form of the likelihood given in Section 2.2 can be used. This is the form which must also be used if the means are unavailable. In this case there is no update step for the mean and the acceptance ratios for between-model moves are

$$(b - a) \left(\frac{|\mathbf{C}^*|}{|\mathbf{C}|} \right)^{\frac{n}{2}} \frac{\exp \left[-\frac{1}{2} \text{tr}(\mathbf{S}\mathbf{\Omega}^*) \right]}{\exp \left[-\frac{1}{2} \text{tr}(\mathbf{S}\mathbf{\Omega}) \right]} \frac{f(\mathbf{C}_{G^*}|G^*)}{f(\mathbf{C}_G|G)}.$$

for edge addition and a similar expression for edge deletion.

The acceptance probabilities for within-model moves are

$$\min \left(1, \frac{(\tau_j^*)^{2\beta-1+n} \exp(-\beta\tau_j^*) \exp \left[-\frac{1}{2} \text{tr}(\mathbf{S}\mathbf{\Omega}^*) \right]}{(\tau_j)^{2\beta-1+n} \exp(-\beta\tau_j) \exp \left[-\frac{1}{2} \text{tr}(\mathbf{S}\mathbf{\Omega}) \right]} \right)$$

for updating τ_j and

$$\min \left(1, \frac{|\mathbf{C}^*|^{n/2} \exp \left[-\frac{1}{2} \text{tr}(\mathbf{S}\mathbf{\Omega}^*) \right]}{|\mathbf{C}|^{n/2} \exp \left[-\frac{1}{2} \text{tr}(\mathbf{S}\mathbf{\Omega}) \right]} \right)$$

for updating ρ_{ij} .

Note also that it is relatively straightforward to restrict the model space by either prohibiting proposals to a specified set of graphs or by rejecting all such proposals. Care must be taken, however, with the former method as $r_m(x)$ and $r_m(x^*)$ may no longer be equal. This type of restriction is useful if it is desired to consider only graphs with (or without) certain edges. Restriction to decomposable models is also possible but is not so easy and has no practical justification, other than to compare with existing results. The numerical model indexing described in 1.3.2 can facilitate such restrictions by allowing specification of a set of graphs through a set of indices. This is not the only way or even always the most efficient method, however, although it may be the most convenient.

3.2.4 Performance

Mixing over the models space may be monitored, for example by the number of edges present, which describe the graph complexity. In general, this is not a problem when mixing over the parameter space is satisfactory. The model indexing, as described in Section 1.3.2, may also be used for monitoring and is invaluable for determining posterior probabilities. The numerical indices (effectively m in Section 1.5) are a more efficient model indicator than the set of partial correlations (and hence edges) present, which gives the same information.

3.3 Prior Sensitivity and a Larger Class of Prior

When comparing models it is often important to assess how sensitive the posterior model probabilities, and hence the Bayes Factors, are to changes in (1) the prior distribution and (2) the prior variance. It is the latter that is considered here. The means and partial precision parameters are present in and have the same priors over all models. Therefore the posterior model probabilities should not be sensitive to their prior variance. The same is not true for the partial correlations but because they have a uniform prior, their prior variance is fixed.

However, the class of prior can be expanded so that, for each $(i, j) \in E$, $\frac{1}{2}(\rho_{ij} + 1)$ has a Beta(a, b) distribution, subject to the constraint that \mathbf{C} is positive definite. The density function of this prior distribution, conditional on graph G with edge set E , is

$$f(\mathbf{C}|G) = k \prod_{(i,j) \in E} \left(\frac{1}{2}\right)^{a+b-1} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} (1 + \rho_{ij})^{a-1} (1 - \rho_{ij})^{b-1} \quad a, b > 0, \quad (3.6)$$

where k is a normalising that must be estimated. The uniform prior already described is then a special case of this with $a = b = 1$. This is a much more flexible class of prior since not only are the prior variances not fixed but different values of a and b could also be chosen for each ρ_{ij} , although typically they would all be the same. The marginal prior variances are proportional to $4\frac{ab}{(a+b)^2(a+b+1)}$ (the exact variance cannot be specified due to the positive definite constraint).

In the absence of any prior information or belief about the signs of the partial correlations, it is desirable to have $a = b$ so that the distribution is symmetric about 0. In this case, the prior density is

$$f(\mathbf{C}|G) = k \prod_{(i,j) \in E} \left(\frac{1}{2}\right)^{2a-1} \frac{\Gamma(2a)}{\Gamma(a)^2} (1 - \rho_{ij}^2)^{a-1} \quad a > 0, \quad (3.7)$$

and the marginal prior variances are proportional to $v(a) = \frac{1}{2a+1}$. Figure 3.1 shows the shape of the marginal prior distribution for ρ_{12} in a saturated model for $a = 0.1, 0.5, 1.0, 1.5, 2.0$ and 4.0 . Except for the first two, each panel shows this marginal distribution for both $q = 3$ and $q = 6$. The first two only show marginals for $q = 3$ due to the difficulty of sampling from the distribution when $q = 6$, described below. The jointly uniform prior corresponds to $a = 1$ with $v(a) = 1/3$ and Figure 3.1 shows clearly that the marginals are not uniform. Marginal uniformity in fact is achieved at around $a = 0.6$. Smaller values of a place more prior mass at the edges of the space while larger values make the prior more concentrated around the centre of the space, and a closer match to positive-definite space. Of course, a has a lower limit of 0 and $v(a)$ has a corresponding upper limit of 1 so we avoid the possibility of an arbitrarily large variance which often leads to Lindley's Paradox (Lindley 1957). However, even though the variance is bounded, we can still expect larger variances to result in greater posterior model probabilities for simpler models compared with more complex ones. As a increases, the variance shrinks but levels off rapidly so it makes sense to place an upper limit, especially since small prior variances are usually undesirable.

The prior normalising constants may be obtained from a rejection sampler as before but it is worth noting that as a increases, the priors become more more concentrated around 0, and the prior normalising constants tend to 1. This effectively places an upper limit on a with the value of this limit depending on q . Table 3.1 shows the acceptance rates, based on a sample of size 1 000 000, for the saturated model for q up to 8 and various values of a . Standard errors, although not quoted, are again typically around 0.001.

Note that there is a practical upper limit on q for each value of a : $q = 3$

for $a = 0.1$; $q = 5$ for $a = 0.5$; $q = 6$ for $a = 1.0$ (as discussed in Section 2.3); $q = 7$ for $a = 1.5$ and $q = 8$ for $a = 2.0$. On one hand, increasing a allows larger dimensions but this decreases the prior variance, making for a more informative prior distribution. On the other hand, decreasing a in order to increase prior variance restricts the size of q . A value of $a = 1.0$ seems a good compromise, borne out in the simulated data examples in the next section, and is therefore the value used in all other examples in Section 3.5 as well as in Chapters 5 and 6.

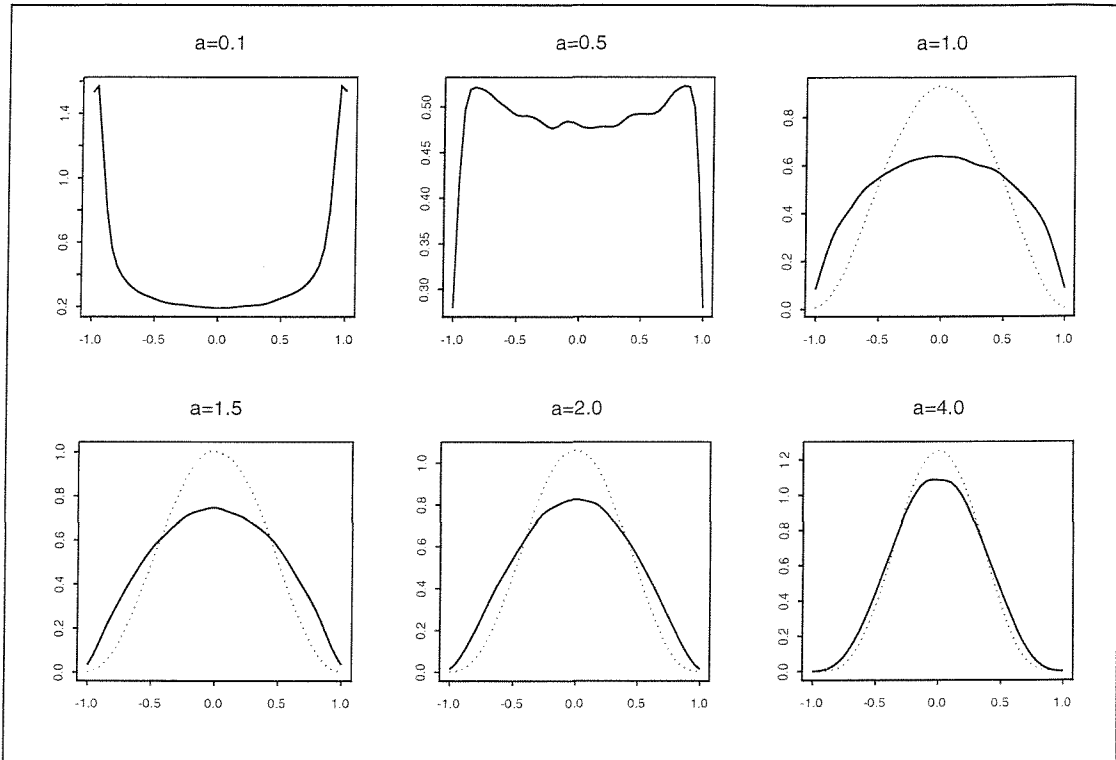


Figure 3.1: Marginal prior densities for ρ_{12} when $q = 3$ (solid lines) and $q = 6$ (dotted lines) under the Beta-based prior with various values of hyperparameter a .

The reversible jump sampler described in Section 3.2 is modified to use this prior by simply multiplying the acceptance ratio for a proposal ρ_{ij}^* in (3.1) by a ratio of (unnormalised) prior densities (3.6 or 3.7). The prior ratios

$$\frac{f(\mathbf{C}_{G^*}|G^*)}{f(\mathbf{C}_G|G)}$$

in (3.4) and (3.5) are now ratios of normalising constants multiplied by, respectively, $(1 - (\rho_{ij}^*)^2)^{a-1}$ and its reciprocal (since the other terms in the prior ratio cancel).

q / a	0.1	0.5	1.0	1.5	2.0	4.0	6.0	8.0
3	0.030	0.334	0.616	0.777	0.867	0.982	0.997	0.999
4	*	0.033	0.182	0.369	0.535	0.884	0.973	0.994
5	*	0.001	0.022	0.091	0.199	0.663	0.890	0.968
6	*	*	0.001	0.010	0.040	0.374	0.715	0.892
7	*	*	*	0.001	0.004	0.147	0.472	0.743
8	*	*	*	*	0.002	0.037	0.240	0.533

Table 3.1: Acceptance rates for estimating prior normalizing constants for saturated GGMs using a rejection sampler, based on a sample size of 1 000 000. * indicates a value of less than 0.001.

Sensitivity of the posterior model probabilities obtained by this sampler to a is investigated in the next section.

3.4 Simulated Data Examples

In order to assess the performance of the sampler, data simulated from various three-dimensional GGM's were used and the sampler output compared with the known true models and parameter values. Three different models were used: The first, M_1 , was a saturated model with all three edges, the second, M_2 , had two edges and the third, M_3 , had only one. These are shown in Figure 3.2.

The means of the distributions used to generate the data were the same in each case and drawn from their prior, $N(0, 10^6)$.

The covariance matrix for the saturated model was drawn from an Inverse Wishart distribution with 4 degrees of freedom and identity matrix as parameter. The covariance matrix for the other two models was obtained by setting appropriate entries to zero in the precision matrix for the saturated model. The number of observations in each case was 30.

In each case, the prior for $\boldsymbol{\mu}$ was $N(\mathbf{0}, 10^6 \mathbf{I})$; The partial precisions, τ_j^2 , were assigned independent gamma priors, with density given by 2.11 and $\alpha = \beta = 0.001$ so that the prior means were 1 and prior variances 1000; The symmetric beta-based prior, described in 3.3, with various values of the hyperparameter a , was used for \mathbf{C} in order to assess prior sensitivity. Six values of a were compared: 0.1, 0.5, 1.0, 1.5, 2.0 and 4.0. Recall that $a = 1.0$ is equivalent to the jointly uniform prior.

The sampler was run each time for 100 000 iterations after allowing 10 000

for burn-in.

The posterior model probabilities, based on the reversible jump sampler output, are tabulated in tables 3.2, 3.3 and 3.4, along with their approximate standard errors. As expected, in nearly all cases smaller values of a results in higher probabilities for models with 1 or 2 edges relative to ones with 2 or 3. The notable exception is the third dataset when $a = 0.1$ but this is likely due to the fact that the true model has only a single edge. However, the true model is the most probable in all but one case. The exception is the second dataset when $a = 4.0$ when two models account for about half the posterior probability each.

These suggest that a value of 0.1 for a is too and a value of 4.0 is too high. As remarked already, $a = 1.0$ is a reasonable and convenient value to use and therefore is the value used for all the examples in the next Section.

Figure 3.3 shows trace plots of batch model probabilities for the two models with highest posterior probability for each dataset, based on batches of size 1000, when $a = 1.0$. Figure 3.4 shows trace plots of batch model probabilities, based on batches of size 1000, for the most probable model for the second dataset for each value of a . Those for the next most probable model and for the other datasets are similar. These both indicate satisfactory mixing over the model space but notice that the standard errors tend to increase with a , although they are still tolerable. This can also be seen directly in the approximate standard errors in Tables 3.2, 3.3, 3.4 and 3.5.

Trace plots for the parameters in the case of the first dataset and $a = 1.0$ are given in figure 3.5. These are typical of the trace plots for the sampler output and indicate satisfactory mixing across the parameter space. When different values of a are used, they are almost identical, indicating a lack of sensitivity of the parametric posterior to a , as would be expected.

It can be seen from the trace plots that the posterior distributions are centred close to the observed values rather than the true values, although the latter are always within the bounds of the distribution. This is a common occurrence in Bayesian inference and shows the influence of the data on the posterior.

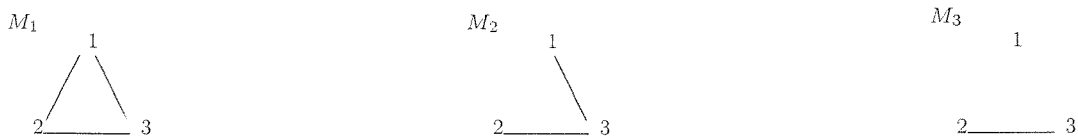


Figure 3.2: Graphs of the models from which simulated data were generated

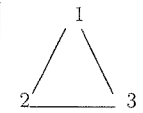
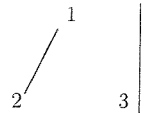
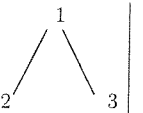
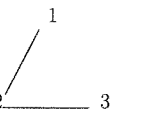
				
$a = 0.1$	0.950 (42)	0.002 (3)	0.024 (25)	0.024 (25)
$a = 0.5$	0.965 (32)	0.013 (15)	0.011 (12)	0.010 (11)
$a = 1.0$	0.956 (40)	0.016 (19)	0.014 (13)	0.013 (12)
$a = 1.5$	0.934 (53)	0.025 (24)	0.019 (17)	0.020 (17)
$a = 2.0$	0.909 (51)	0.033 (25)	0.032 (19)	0.026 (17)
$a = 4.0$	0.820 (62)	0.054 (26)	0.064 (25)	0.060 (26)

Table 3.2: Posterior model probabilities for the first simulated dataset with approximate standard errors $\times 10^4$ in brackets. The true model is M_1 with graph 7.

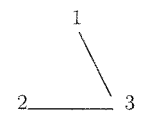
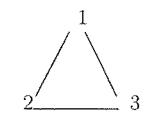
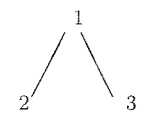
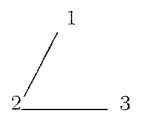
				
$a = 0.1$	0.930 (92)	0.067 (85)	0.003 (24)	0.000 (1)
$a = 0.5$	0.778 (155)	0.211 (137)	0.009 (58)	0.001 (9)
$a = 1.0$	0.692 (182)	0.304 (176)	0.001 (8)	0.003 (26)
$a = 1.5$	0.685 (194)	0.314 (193)	0.001 (8)	0.000 (1)
$a = 2.0$	0.662 (181)	0.337 (179)	0.001 (10)	0.000 (0)
$a = 4.0$	0.454 (190)	0.540 (186)	0.006 (24)	0.000 (0)

Table 3.3: Posterior model probabilities for the second simulated dataset with approximate standard errors $\times 10^4$ in brackets. The true model is M_2 with graph 3.

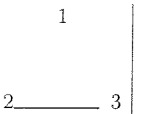
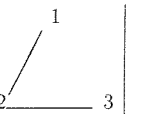
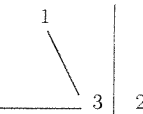
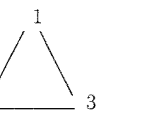
				
$a = 0.1$	0.543 (30)	0.229 (31)	0.208 (27)	0.020 (8)
$a = 0.5$	0.693 (30)	0.142 (22)	0.133 (22)	0.032 (11)
$a = 1.0$	0.681 (30)	0.148 (21)	0.130 (23)	0.041 (12)
$a = 1.5$	0.655 (33)	0.153 (20)	0.142 (21)	0.050 (15)
$a = 2.0$	0.625 (33)	0.163 (21)	0.151 (23)	0.061 (14)
$a = 4.0$	0.522 (34)	0.199 (22)	0.191 (27)	0.088 (18)

Table 3.4: Posterior model probabilities for the third simulated dataset with approximate standard errors $\times 10^4$ in brackets. The true model is M_3 with graph 1.

The values of the partial correlations in these three models were each fairly large (-0.85, 0.58, 0.68) so in order to assess the effect of a small (close to zero) partial correlation, a fourth dataset was generated from a saturated model with one partial correlation equal to 0.06. The posterior model probabilities and their approximate standard errors are shown in table 3.5. As might be expected, the graph with the corresponding edge missing received the highest posterior probability. This is another example of the data driving the posterior.

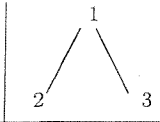
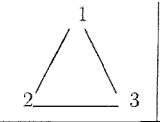
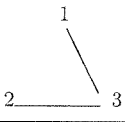
			
$a = 0.1$	0.931 (84)	0.068 (83)	0.001 (9)
$a = 0.5$	0.811 (124)	0.188 (123)	0.001 (8)
$a = 1.0$	0.777 (124)	0.218 (115)	0.005 (42)
$a = 1.5$	0.741 (115)	0.258 (114)	0.001 (5)
$a = 2.0$	0.726 (137)	0.271 (135)	0.003 (19)
$a = 4.0$	0.635 (145)	0.365 (145)	0.000 (5)

Table 3.5: Posterior model probabilities for the fourth simulated dataset with approximate standard errors $\times 10^4$ in brackets. The true model is M_1 with graph 7.

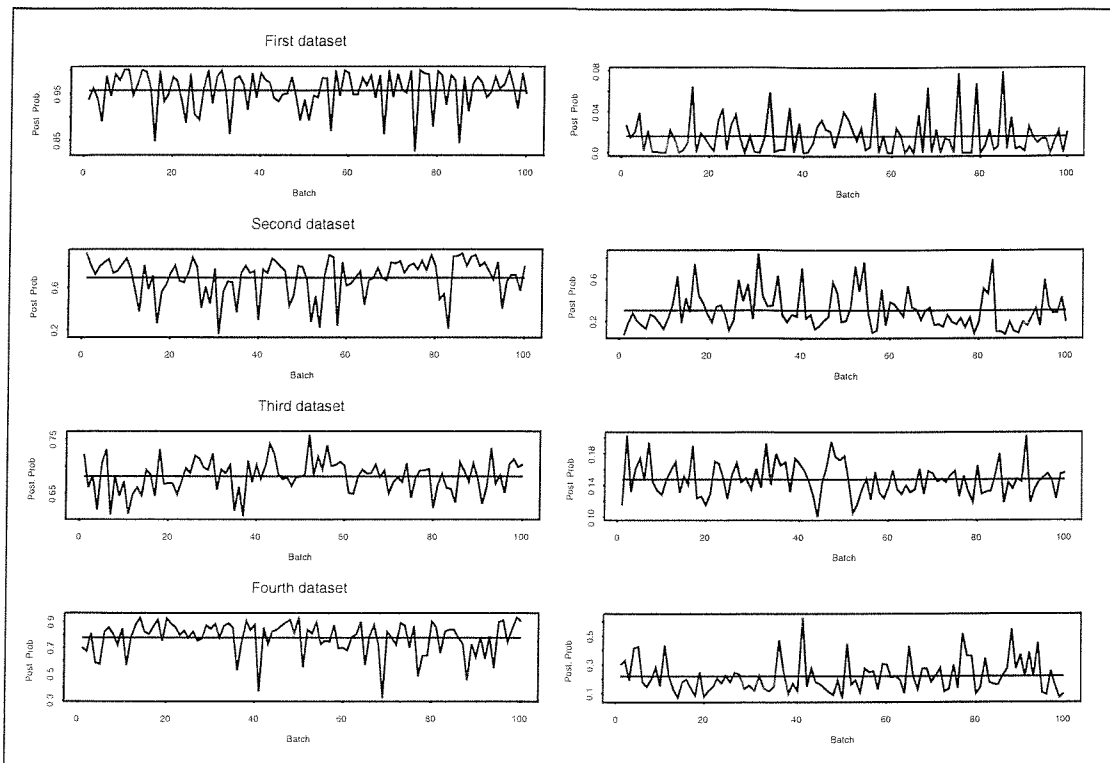


Figure 3.3: Trace plots of the two highest posterior model probabilities for four simulated datasets, based on batches of size 1000. The lines show the averages, that is probabilities based on the entire sample.

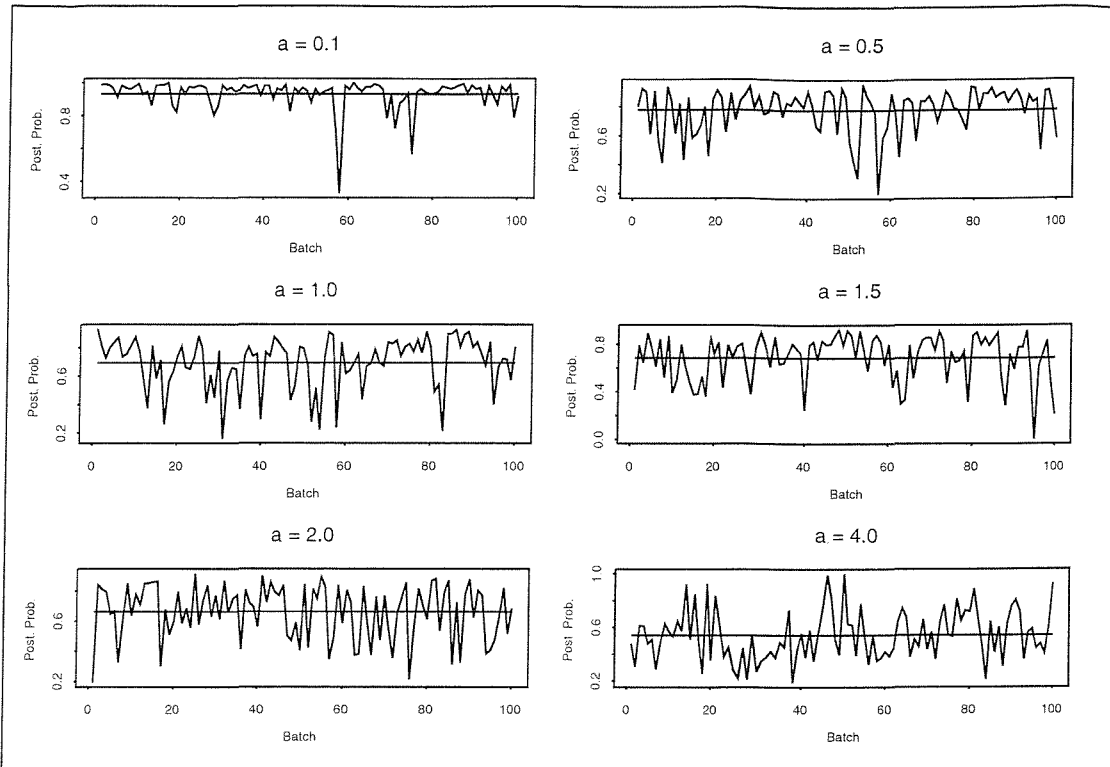


Figure 3.4: Trace plots of the highest posterior model probabilities for the second simulated dataset, based on batches of size 1000, for various values of hyperparameter a . The lines show the averages, that is probabilities based on the entire sample.

In addition to the three-dimensional datasets, data were also generated from a randomly-chosen six-dimensional model. As previously noted, this is the highest dimension of model that may be practically be considered, due to the difficulty in estimating the prior normalising constants. It is also the highest dimension of model considered in the examples in the next section.

The model chosen was 31345, which has 9 edges and is shown in Figure 3.6. The parameters of the model were generated as for the previous datasets and the same prior distributions used, with a joint uniform prior for \mathbf{C} .

This time the sampler was run for 1 000 000 iterations to allow for the much larger model space and 10 000 were allowed for burn-in.

The posterior model probabilities for the four most probable models are given in Table 3.6, along with their standard errors (in brackets). The true model received less than 1% of the posterior probability.

Trace plots of batch model probabilities, based on batches of size 10 000 are given in Figure 3.7 and again indicate satisfactory mixing over the model space.

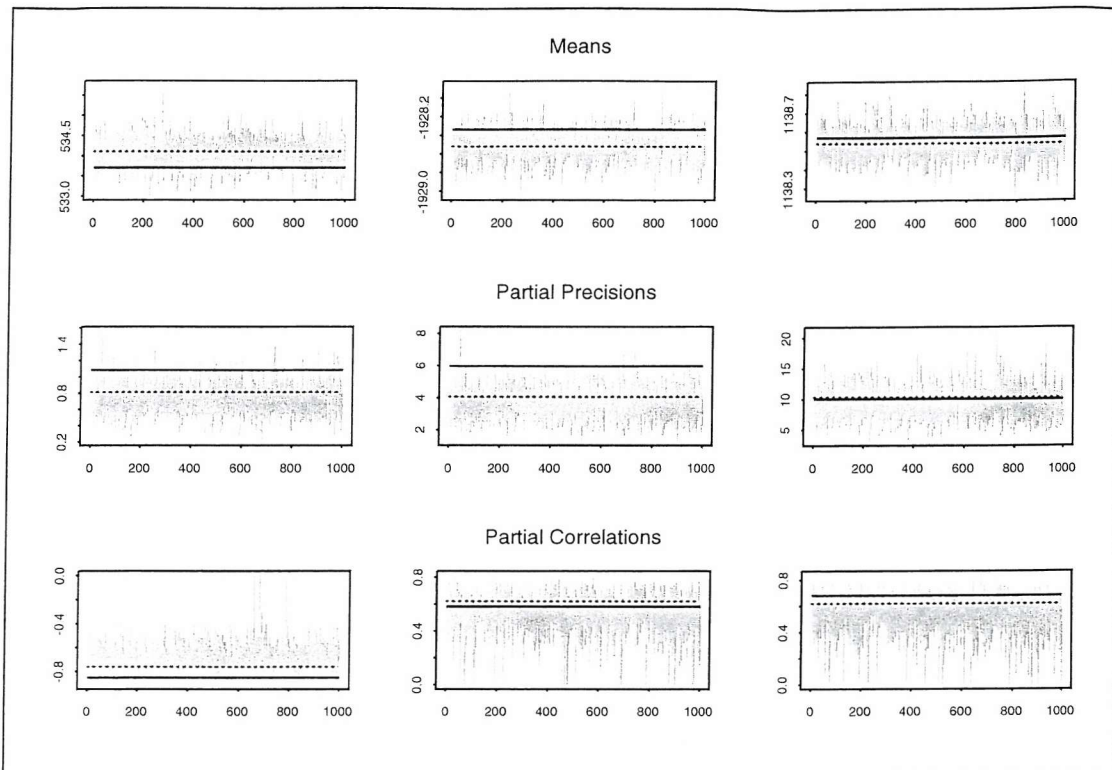


Figure 3.5: Trace plots for the first simulated dataset. The solid lines show the true values and the dashed lines the observed. The output has been thinned to 1000 observations.

31345

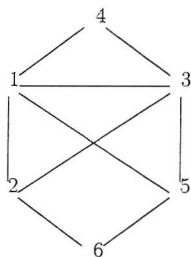


Figure 3.6: Graph 31345, the graph of the model from which six-dimensional simulated data were generated

Trace plots for the parameters are given in Figure 3.8. Again, the posterior distribution can be seen to be centred near the observed values rather than the true values.

The posterior in this case is very diffuse and the individual probabilities are quite small. This is due partly to the larger model space and partly to the fact that some of the observed partial correlations are fairly small (close to zero). For this reason, the edge inclusion frequencies, given in table 3.7, are more informative. It can be seen from this and the trace plots that the probability of an edge

being included at any iteration depends largely on the magnitude of the corresponding observed partial correlation, which should reflect the true values. Edges (1, 2), (1, 4), (1, 5), (2, 3) and (3, 4) have relatively large corresponding values for both true and observed partial correlations and hence have very high inclusion rates; (2, 6) has a small true and observed values and hence a low inclusion rate (25%); Edges (2, 4), (2, 5), (4, 5) and (4, 6) are absent from the true model, have observed partial correlations closer to zero and hence also have relatively low inclusion rates; Edges (1, 3), (5, 6), (3, 6) and especially (3, 5) and (1, 6) have true values close to or equal to zero yet large observed values and hence moderate to high inclusion rates.

To further demonstrate this last point, a further dataset was simulated from a model with with graph 14888, which has six edges. In this model, all the partial correlations were reasonably large ($|\rho| > 0.3$). The values of the other parameters and the prior distributions were the same as for the previously example.

The observed matrix of partial correlations is:

$$\begin{pmatrix} 1 & -0.083 & 0.372 & -0.554 & 0.360 & -0.211 \\ & 1 & -0.415 & 0.011 & -0.043 & 0.027 \\ & & 1 & 0.444 & 0.131 & -0.540 \\ & & & 1 & -0.116 & -0.133 \\ & & & & 1 & 0.108 \\ & & & & & 1 \end{pmatrix}$$

This time the true model has the highest posterior probability and although this probability is not very large (8%), it is over twice as large as that of any other. The graphs of the four most probable models are displayed in table 3.8 along with their (estimated) probabilities and their standard errors (in brackets). The edge inclusion percentages are also tabulated in table 3.9. It can be seen from this that those edges corresponding to partial correlations with large magnitude have high inclusion frequencies whereas those corresponding to small partial correlations have lower inclusion frequencies. Moreover, the closer to zero the observed partial correlation, the lower the inclusion frequency.

This feature seems intuitively reasonable and again shows the influence of the data on the posterior distribution. In addition, if we are performing model selection based on such results, we will select a model with missing edges corresponding to small partial correlations, which is exactly what would be desired. If instead we are performing inference under model uncertainty, such as model-averaging, models with edges corresponding to larger partial correlations will have greater weight, again exactly what would be desired.

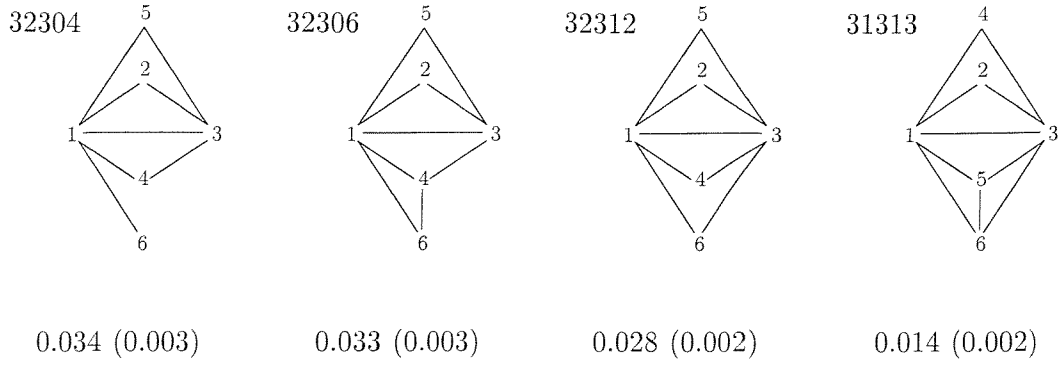


Table 3.6: Posterior probabilities for the four most probable models for the fifth simulated dataset. Approximate standard errors are in brackets.

edge	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)	(2,3)	(2,4)	(2,5)	(2,6)
	97.11	64.80	99.17	91.52	85.26	99.54	24.02	28.84	24.13
		(3,4)	(3,5)	(3,6)	(4,5)	(4,6)	(5,6)		
		97.86	89.22	42.06	24.01	32.62	38.33		

Table 3.7: Edge inclusion percentages for the fifth simulated dataset

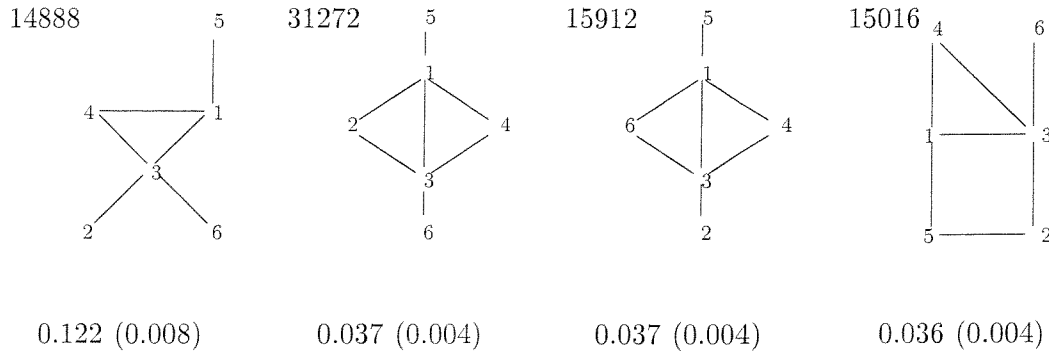


Table 3.8: Posterior probabilities for the four most probable models for the sixth simulated dataset. Approximate standard error are in brackets.

(1,2)	(1,3)	(1,4)	(1,5)	(1,6)	(2,3)	(2,4)	(2,5)	(2,6)
24.54	99.97	99.92	99.62	31.08	99.65	20.59	25.95	19.83
(3,4)	(3,5)	(3,6)	(4,5)	(4,6)	(5,6)			
99.74	19.70	100.00	22.40	24.45	16.34			

Table 3.9: Edge inclusion percentages for the sixth simulated dataset

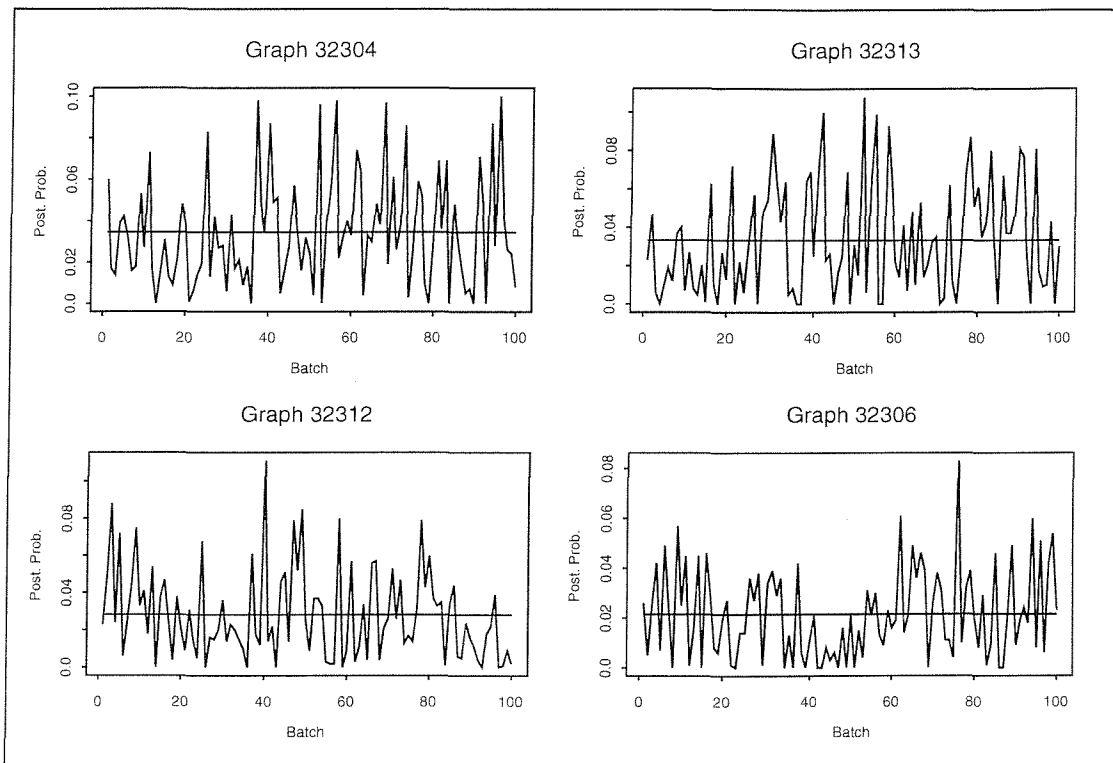


Figure 3.7: Trace plots of the four highest posterior model probabilities for a six-dimensional simulated dataset, based on batches of size 1000.

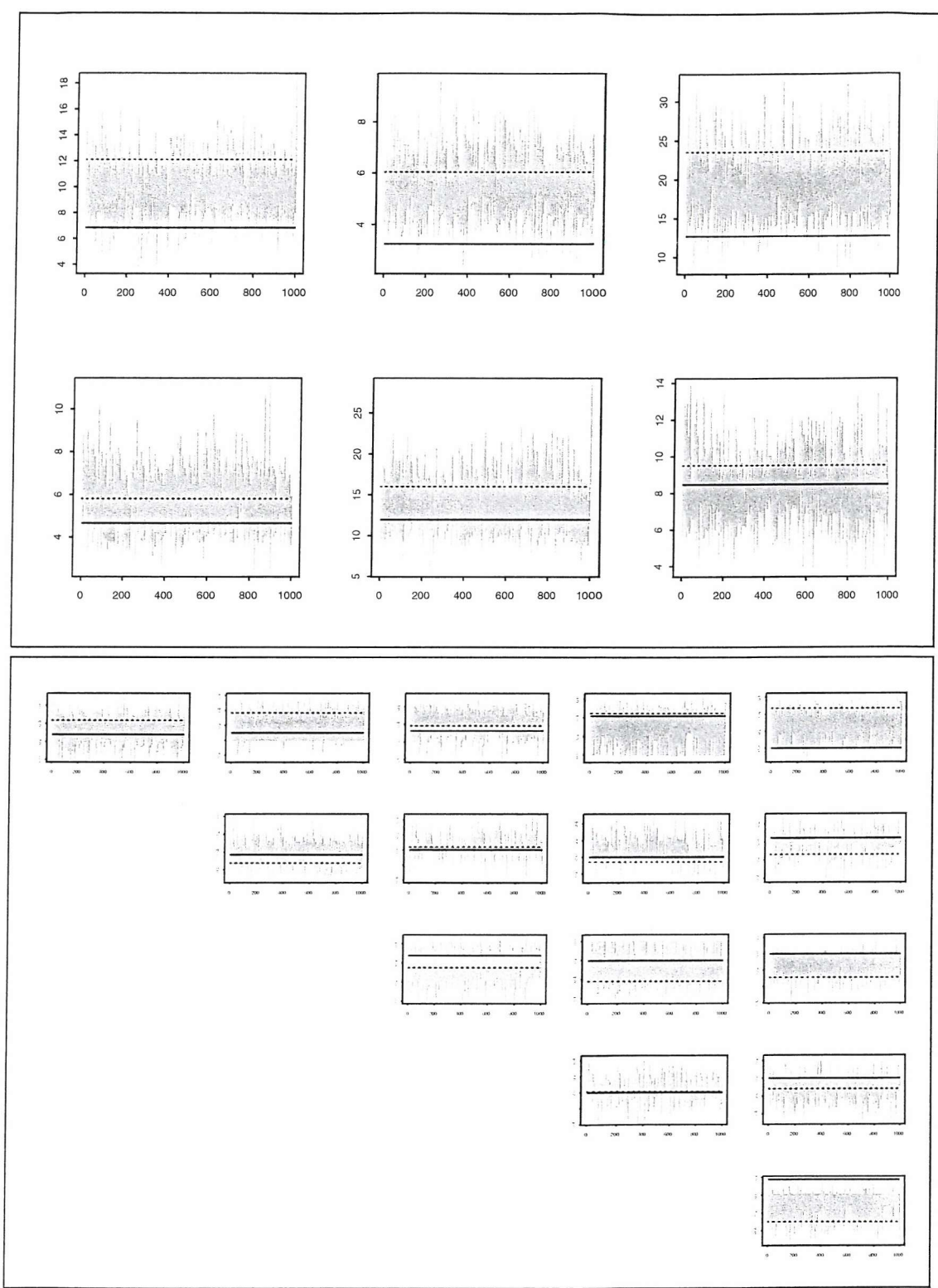


Figure 3.8: Trace plots for the fifth simulated dataset. The solid lines show the true values and the dashed lines the observed. The output has been thinned to 1000 observations.

3.5 Further Examples

In this section a number of examples of application to real data of the reversible jump sampler for graphical Gaussian models are presented. They are in order of increasing dimension and comparison with any previous results are given. Unless otherwise stated, the estimated posterior model probabilities are based on 100 000 iterations after allowing 10 000 for burn-in. Their Monte Carlo standard errors are calculated by splitting the (model index) output of the chain into batches (Geyer 1992). In most cases batches of size 1000 sufficed. Convergence diagnostics, including model probabilities for some of the most probable graphs based on the batches and trace plots for the parameters were examined in each case. Neither these nor parameter estimates are presented but were always similar to those of the simulated data examples in the previous section.

The prior distributions used were as in the previous section, with $\alpha = \beta = 0.001$ in the prior for the partial precisions and a jointly uniform prior for \mathbf{C} . The chains were initialised at the mutual independence model with sample means as initial values for $\boldsymbol{\mu}$ and default values of 1.0 for the partial precisions.

3.5.1 Digoxin Clearance

This 3-dimensional example is examined by Edwards (1995,2000). The data are from Halkin et al. (1975) and are on 35 consecutive patients under treatment for heart failure with the drug digoxin. The variables are digoxin clearance (DC) (i.e. the amount of blood that in a given interval is cleared of digoxin), creatinine clearance (CC) (used as a measure of kidney function) and urine flow (U). Since creatinine and digoxin are mainly eliminated by the kidneys, CC and DC can be expected to be correlated. There is no obvious reason for correlation with urine flow, which depends on factors such as fluid intake and temperature. Halkin et al suspected that the elimination of digoxin might be subject to reabsorption, which might give rise to correlation with urine flow. Edwards used a deviance-based approach to conclude that there is good evidence that DC and urine flow are correlated. The posterior model probabilities from the reversible jump sampler, given in Table 3.10, agree with this ; the two graphs with both (DC,CC) and (DC,U) edges account for over 95% of the posterior probability.

Graph 5 $P_5 = 0.686$ (59)	Graph 7 $P_7 = 0.279$ (61)	Graph 6 $P_6 = 0.018$ (17)	Graph 4 $P_4 = 0.016$ (16)

Table 3.10: Posterior model Probabilities for Digoxin Clearance data. Standard Errors $\times 10^4$ in brackets.

3.5.2 Anxiety and Anger

Cox and Wermuth(1993) describe a set of psychological data obtained from Spielberg et al(1970,1983). There are four variables measured on 684 female students. They are anxiety state(W) anger state(X) anxiety trait (Y) and anger trait(Z). The trait variables are viewed as stable personality characteristics and the state variables as pertaining to behaviour in certain situations. The example is also treated in Wermuth(1991) and Edwards (1995,2000). Psychological theory suggests conditional independences between W and Z and between X and Y . The reversible jump results support this theory, giving a posterior probability of over 70% to the corresponding graph, 51 or $WX/WY/XZ/YZ$. Edwards uses deviance-based analysis to select the same graph.

51 0.767 (100)	59 0.133 (88)	55 0.088 (69)	63 0.012 (12)

Table 3.11: Posterior model probabilities for Anxiety and Anger data. Standard Errors $\times 10^4$ are in brackets.

3.5.3 Fret's Heads

These data, given in Whittaker (1990,p265) are head measurements on pairs of adult sons in a sample of 25 families. The variables are

$Y_1 = \{\text{Length of first son's head}\}; Y_2 = \{\text{Breadth of first son's head}\};$

$Y_3 = \{\text{Length of second son's head}\}; Y_4 = \{\text{Breadth of second son's head}\}.$ Since

$q = 4$, the number of possible graphs is 64, including three, 30, 45 and 51, which are not decomposable. First, all possible graphs are considered. The posterior

probabilities are given in Table 3.12. Notice that the two most probable graphs are nondecomposable. Notice also that the edges (1, 2) and (3, 4) are present in all of the 15 most probable models, which make up over 90% of the posterior probability. These results are consistent with obvious subject-matter considerations and, as shown below, if only graphs with these edges are considered, the results are largely unaffected. These results are also consistent with those in Dellaportas, Giudici and Roberts (2003), where the same two nondecomposable graphs have the highest posterior probabilities.

<p>51 0.192 (123)</p>	<p>45 0.112 (93)</p>	<p>59 0.083 (52)</p>	<p>55 0.077 (53)</p>	<p>43 0.064 (65)</p>	<p>61 0.058 (50)</p>
<p>47 0.057 (44)</p>	<p>53 0.053 (58)</p>	<p>63 0.043 (25)</p>	<p>49 0.032 (33)</p>	<p>57 0.031 (31)</p>	<p>35 0.030 (39)</p>
<p>41 0.025 (36)</p>	<p>39 0.025 (28)</p>	<p>37 0.018 (31)</p>	<p>33 0.014 (22)</p>		

Table 3.12: Posterior model probabilities for Fret’s heads data (1). Approximate standard errors $\times 10^4$ are in brackets.

Giudici and Green (1999) have also applied reversible jump MCMC to these data but considered only decomposable models and used a hierarchical prior. In order to make a comparison with their results, the sampler was run as before but this time rejecting any proposed move to one of the three nondecomposable graphs. Table 3.13 gives both sets of results, both based on 100 000 iterations with 10 000 of burn-in. Note that the results are as would be expected considering the first set and have the same 12 graphs receiving about 80% of the posterior probability.

Giudici(1996) performed a Bayesian analysis on these data using both local and global priors. Considering only decomposable models, he was able to compute

$p_1 = 0.121$	$p_1 = 0.111$	$p_1 = 0.109$	$p_1 = 0.095$	$p_1 = 0.080$	$p_1 = 0.078$
$p_2 = 0.109(74)$	$p_2 = 0.078(63)$	$p_2 = 0.122(80)$	$p_2 = 0.084(62)$	$p_2 = 0.080(63)$	$p_2 = 0.090(69)$
$p_1 = 0.059$	$p_1 = 0.059$	$p_1 = 0.041$	$p_1 = 0.034$	$p_1 = 0.032$	$p_1 = 0.032$
$p_2 = 0.060(32)$	$p_2 = 0.044(58)$	$p_2 = 0.046(43)$	$p_2 = 0.040(60)$	$p_2 = 0.032(43)$	$p_2 = 0.039(44)$

Table 3.13: Posterior model probabilities for Fret’s heads data (2), decomposable models only. Giudici and Green’s are given as p_1 ’s. Approximate standard errors $\times 10^4$ are given in brackets.

posterior model probabilities directly and compare the results for each prior. He further restricts the model space by considering only those decomposable graphs with edges (1, 2) and (3, 4), linking the son-specific pairs, as well as the mutual independence graph (graph 0), making 15 possibilities. For comparative purposes, the reversible jump sampler was likewise restricted but without inclusion of graph 0 as this is computationally infeasible with the reversible jump scheme as it only changes one edge at a time. However, this graph was found to have negligible posterior probability. Table 3.14 compares the results with Giudici’s for the local prior, which are quite similar. The global prior differs considerably from both, with posterior probability concentrated around the most complex graphs and the complete graph alone accounting for 91%. However, as already noted, the generalized version of this prior (Dellaportas et al. 2003) produces more concordant results, which casts some doubt on the validity of the global prior results.

3.5.4 Fisher’s Iris data

This famous dataset consists of 50 sets of observations on each of three species of iris, *setosa*, *versicolor* and *virginia*. The continuous variables are $Y_1 = \{\text{sepal length}\}$; $Y_2 = \{\text{sepal width}\}$; $Y_3 = \{\text{petal length}\}$; $Y_4 = \{\text{petal width}\}$. Roverato(2002) takes the *virginia* data and obtains posterior model probabilities using a conjugate prior, a generalization of the hyper inverse Wishart, but allowing both decomposable and nondecomposable graphs. These are compared with estimated

<p>$p_1 = 0.173$ $p_2 = 0.119(89)$</p>	<p>$p_1 = 0.161$ $p_2 = 0.114(97)$</p>	<p>$p_1 = 0.117$ $p_2 = 0.061(41)$</p>	<p>$p_1 = 0.115$ $p_2 = 0.080(73)$</p>	<p>$p_1 = 0.113$ $p_2 = 0.083(61)$</p>	<p>$p_1 = 0.092$ $p_2 = 0.092(71)$</p>
<p>$p_1 = 0.077$ $p_2 = 0.077(59)$</p>	<p>$p_1 = 0.038$ $p_2 = 0.046(50)$</p>	<p>$p_1 = 0.030$ $p_2 = 0.034(34)$</p>	<p>$p_1 = 0.026$ $p_2 = 0.046(66)$</p>	<p>$p_1 = 0.024$ $p_2 = 0.042(63)$</p>	<p>$p_1 = 0.020$ $p_2 = 0.037(51)$</p>
<p>$p_1 = 0.014$ $p_2 = 0.026(48)$</p>	<p>$p_1 = 10^{-5}$ $p_2 = 0.000$</p>	<p>$p_1 = 10^{-15}$ $p_2 = 0.000$</p>			

Table 3.14: Posterior model probabilities for Fret’s heads data (3). Giudici’s are given as p_1 ’s. Approximate standard errors $\times 10^4$ are given in brackets.

probabilities from the reversible jump sampler in Table 3.15. The same 16 graphs are the most probable and account for over 95% of the posterior probability under both methods. The most noticeable difference is that the reversible jump gives the saturated model a much lower probability. Notice that these 16 graphs are all those that include the between-length and between-width edges, (1, 3) and (2, 4). Notice also that the results show that the posterior corresponding to the conjugate prior is more diffuse, whereas the other has graph 50 at over three times the probability of the next most probable. This is likely due to the fact that the marginal prior for \mathbf{C} under the conjugate prior is effectively a marginally uniform prior as in Section 2.3, which usually results in a more diffuse posterior and greater posterior probability for simpler models relative to more complex ones.

For completeness, the results for the other two species are tabulated in Tables 3.16 and 3.17. Notice that they are quite different for each species. In each case, the most probable model is also that selected by deviance-based model selection in MIM. If the entire continuous portion of the data is taken together, the saturated model receives over 80% of the posterior probability, corroborating the evidence of the separate analyses that a single nonsaturated graph cannot do justice to the data. In Chapter 5, the entire dataset is analyzed using conditional Gaussian models.

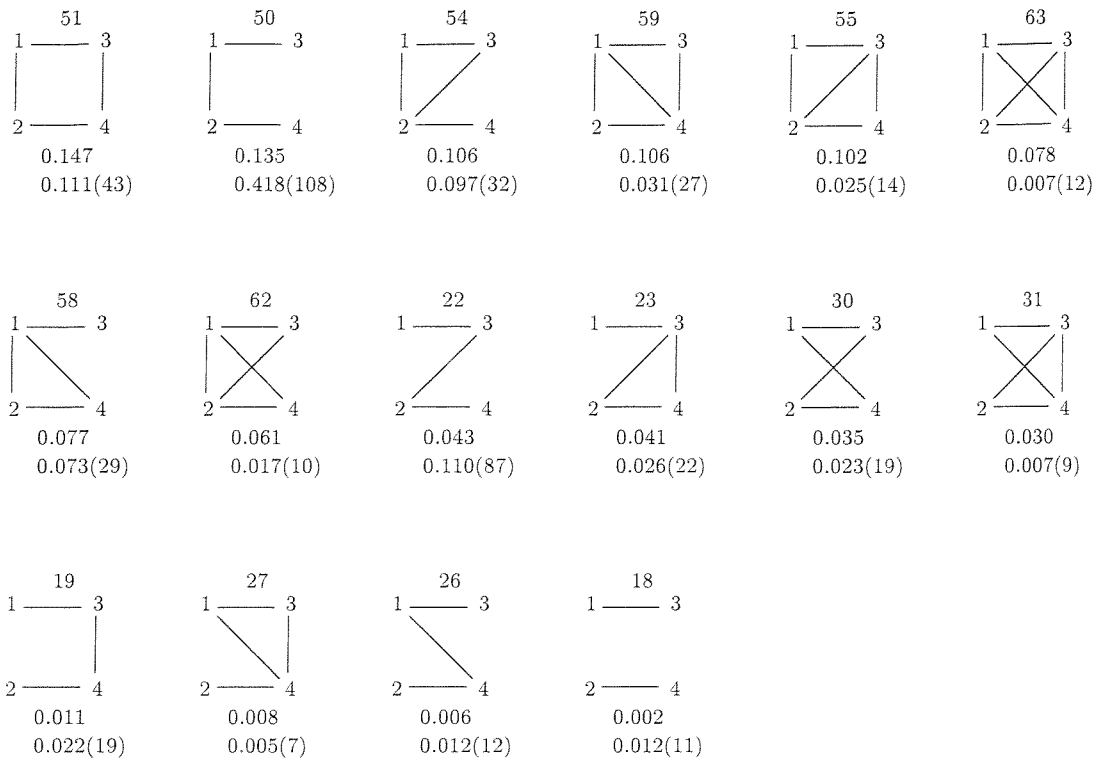


Table 3.15: Posterior model probabilities for *Iris Virginia*, ordered by Roverato's results. Reversible Jump results below with standard errors $\times 10^4$ in brackets.

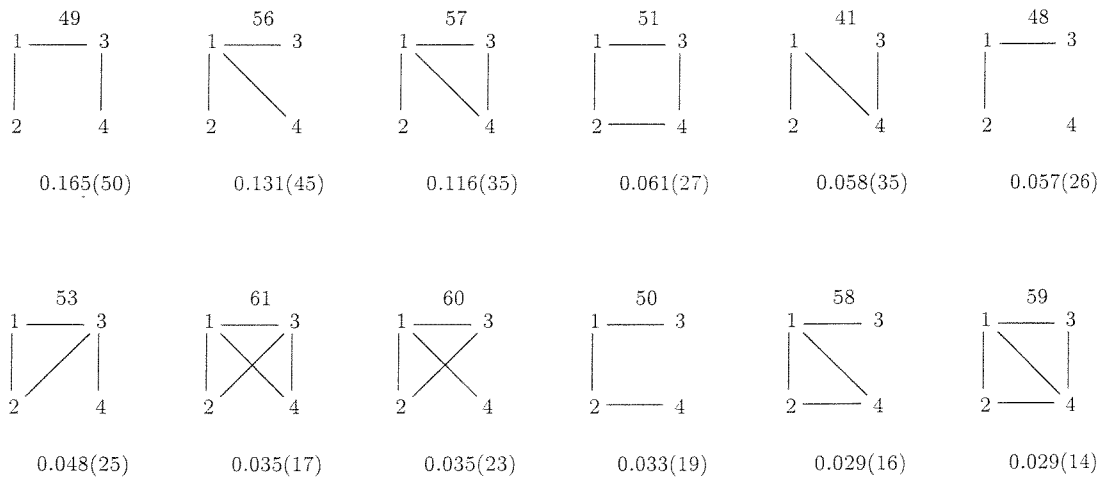


Table 3.16: Posterior model probabilities for *Iris Setosa* with standard errors $\times 10^4$ in brackets.

19	51	59	27	23	55
1 — 3 2 — 4	1 — 3 2 — 4	1 — 3 2 — 4	1 — 3 2 — 4	1 — 3 2 — 4	1 — 3 2 — 4
0.332(106)	0.254(83)	0.111(74)	0.097(55)	0.084(44)	.062(39)
63	31	21	53		
1 — 3 2 — 4	1 — 3 2 — 4	1 — 3 2 — 4	1 — 3 2 — 4		
0.032(41)	0.022(18)	0.002(15)	0.002(10)		

Table 3.17: Posterior model probabilities for *Iris Versicolor* with standard errors $\times 10^4$ in brackets.

3.5.5 Mathematics Marks

These data can be found in Mardia, Kent and Bibby (1979) and in Whittaker (1990), where they are used to illustrate graphical Gaussian models. They are the results of 88 students in each of five mathematics examinations: (1)Mechanics, (2)Vectors, (3)Algebra, (4)Analysis and (4)Statistics. Whittaker uses backward elimination to select graph 807 or 123/345, a so-called “butterfly graph”. This is also chosen by Mardia et al.. It is no surprise then that the reversible jump gives this graph a posterior probability of over 25%, much greater than any other.

3.5.6 Synchronized Swimming

This example is taken from Fligner and Verducci (1988). The data are the total scores assigned by each of 5 judges to each of 40 competitors in a synchronized swimming event at the 1986 National Olympic Festival in Houston, Texas. A nonparametric test is used to detect bias in the judges’ ratings.

The observed matrix of partial correlations is as follows:

$$\begin{pmatrix} 1.0 & 0.0995 & 0.1624 & 0.2938 & 0.3022 \\ & 1.0 & 0.2417 & 0.0634 & 0.2513 \\ & & 1.0 & 0.3800 & 0.1258 \\ & & & 1.0 & 0.1011 \\ & & & & 1.0 \end{pmatrix}$$

By inspection, it is to be expected that the most probable models will include

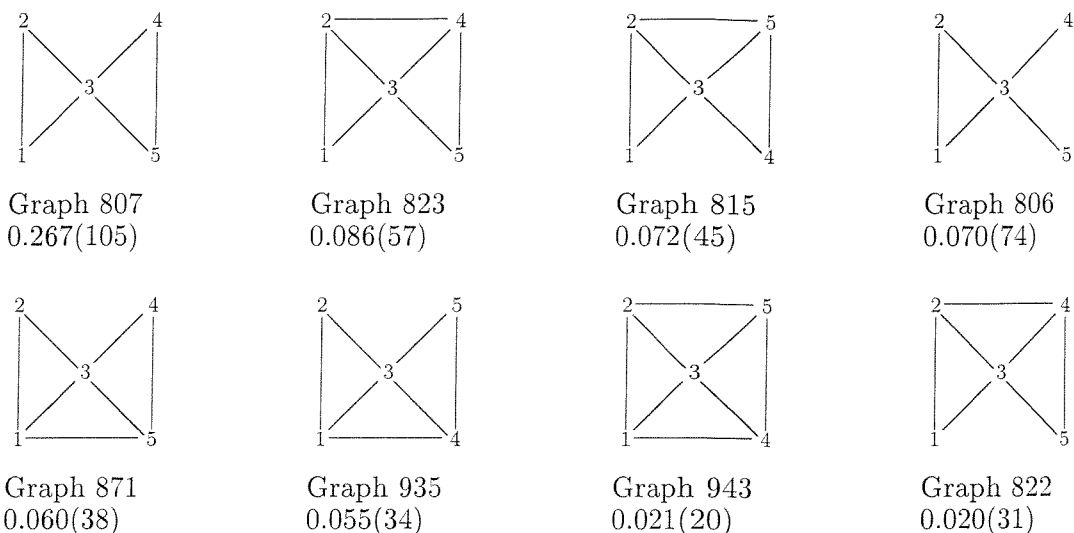


Table 3.18: Posterior model probabilities for mathematics marks data with standard errors $\times 10^4$ in brackets

edges $(1, 4), (1, 5), (2, 3), (2, 5)$ and $(3, 4)$ but not $(1, 2), (2, 4)$ or $(4, 5)$. The edge inclusion percentages after a run of 100000 iterations with 10000 of burn-in, are in Table 3.19 and the posterior probabilities for the eight most probable graphs are in Table 3.20. Notice that these are mostly as expected with very small differences between them, indicating a fairly diffuse posterior. Eighteen graphs are needed to account for 50% of the posterior probability and over fifty to account for 80%. Notice also that in most of these graphs the cycle $(1, 5, 2, 3, 4)$ is present.

Deviance-based model selection procedures in MIM support these results: Backward selection yields graph 494; Forward selection yields graph 1004, which is the same as 492 with an added $(1, 2)$ edge.

Edge	(1,2)	(1,3)	(1,4)	(1,5)	(2,3)	(2,4)	(2,5)	(3,4)	(3,5)	(4,5)
%in	42.89	56.17	92.13	95.63	83.12	36.97	84.99	97.11	49.80	38.74

Table 3.19: Edge inclusion %ages for synchronised swimming data

For the final three examples, all of which involve six variables, only the observed correlations are available so in order to proceed, the means are taken to be zero and the correlation matrix is regarded as a covariance matrix. While this may not allow reliable inference about the τ 's, it will not affect the ρ 's which are the parameters of primary interest. Diagnostic Plots are presented for the first

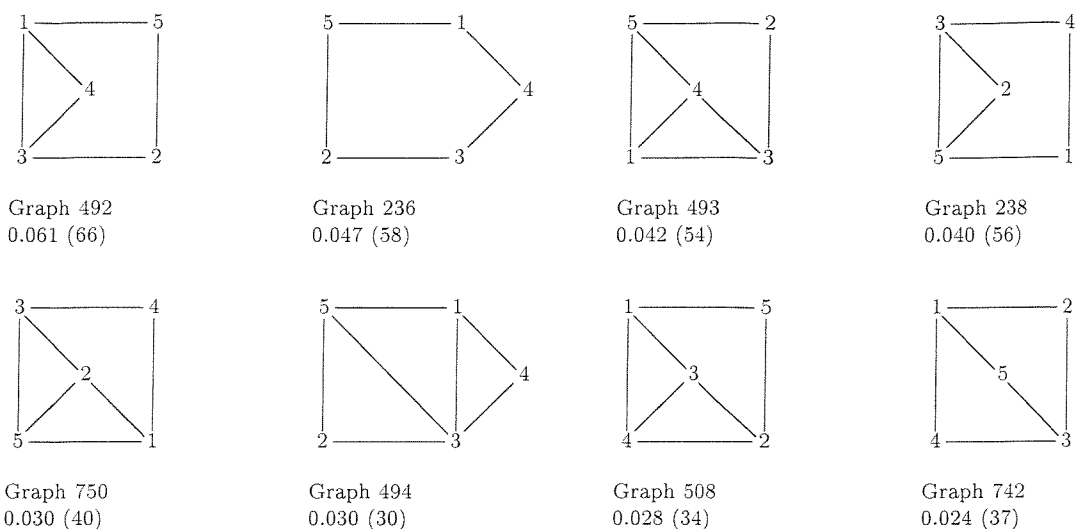


Table 3.20: Posterior model probabilities for synchronised swimming data with standard errors $\times 10^4$ in brackets

to show the good performance of this modified sampler, especially as six is the highest number of variables in any example examined.

3.5.7 Exam Marks

This example is taken from Whittaker (1990), p.266, where the correlation matrix for 220 boys tested in each of six subjects. These are (1)Gaelic, (2)English, (3)History, (4)Arithmetic, (5)Algebra and (6)Geometry. The sixteen most probable graphs and their posterior probabilities are tabulated in Table 3.22 and the edge inclusion percentages in Table 3.21. All other graphs have probabilities of less than 1%. There is a distinct mode at graph 27463 or 123/15/26/56/246/456, the graph suggested by Whittaker and which has five times the probability of the next little. Notice that this graph is also nondecomposable and is intuitively reasonable.

edge	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)	(2,3)	(2,4)	(2,5)	(2,6)
P	100	99.98	17.66	93.11	17.07	92.79	87.25	19.42	80.62
	(3,4)	(3,5)	(3,6)	(4,5)	(4,6)	(5,6)			
	13.17	13.45	17.69	100	99.43	99.55			

Table 3.21: Edge inclusion %ages for Exam Marks example

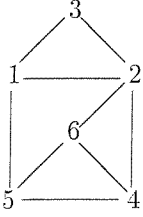
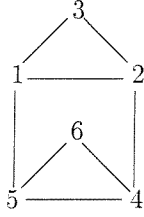
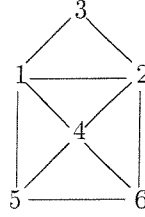
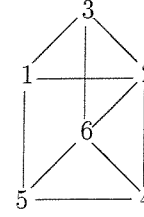
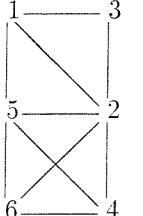
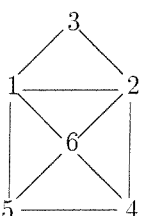
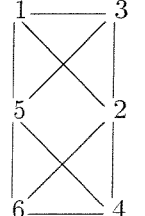
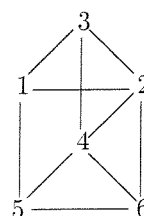
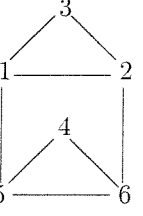
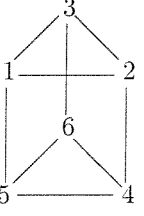
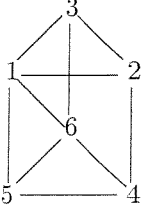
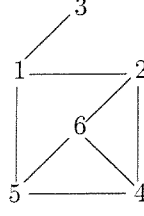
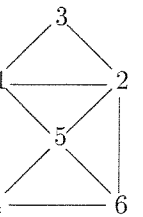
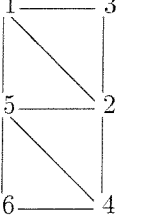
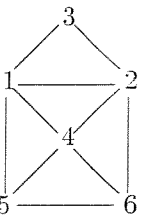
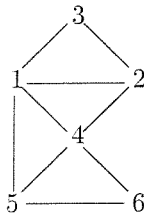
			
27463 0.242 (114)	27399 0.052 (57)	31559 0.042 (35)	27471 0.042 (33)
			
27591 0.040 (28)	28487 0.039 (36)	27479 0.033 (33)	27495 0.033 (28)
			
27207 0.024 (35)	27407 0.017 (23)	28423 0.017 (31)	26951 0.015 (28)
			
27335 0.013 (27)	27527 0.010 (20)	31687 0.008 (24)	31495 0.010 (12)

Table 3.22: Posterior model probabilities for Exam Marks example with standard errors $\times 10^4$ in brackets

3.5.8 Fowl Bones

This example is also taken from Whittaker (1990) p.266, originally from Wright(1954) and concerns bone measurements of 276 white leghorn fowl. They are, (1)Skull Length, (2)Skull Breadth, (3)Humerus, (4)Ulna, (5) Femur and (6) Tibia. The sixteen most probable graphs and their probabilities are tabulated in Table 3.24 and the edge inclusion percentages in Table 3.23. All other graphs have probabilities of less than 1%. Notice that the (1, 2) edge and the (3, 4, 6, 5) cycle are almost always present, as would be expected from anatomical considerations (3 and 4 are wing bones, 5 and 6 are leg bones). The posterior is more diffuse in this case but the most probable graph, 26163 or 123/16/34/35/46/56 has a probability of 7.6%, almost twice the probability of the next, which is the same but lacks the (1, 3) edge. This graph is also that selected by Whittaker. The most probable graph of Giudici and Green (1999) is 26175 or 123/136/3456, the ninth most probable here but the most probable decomposable graph.

edge	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)	(2,3)	(2,4)	(2,5)	(2,6)
P	99.98	64.38	30.07	15.31	72.89	91.30	26.74	11.32	33.72
	(3,4)	(3,5)	(3,6)	(4,5)	(4,6)	(5,6)			
	99.94	99.47	6.67	3.12	99.40	99.99			

Table 3.23: Edge inclusion %ages for Fowl Bones example

3.5.9 Voting Behaviour

This example is from Wermuth (1980) who examined a set of sociological data taken from Goldberg (1966) in order to illustrate a search method. Goldberg’s analysis is “concerned with making inferences about patterns of causal relationships among six variables: father’s sociological characteristics (FSC); father’s party identification (FPI); respondent’s sociological characteristics (RSC); respondent’s party identification (RPI); respondent’s partisan attitudes (RPA); and respondent’s vote for president in 1956 (RV).” As Wermuth notes, the assumption of normality is only crudely approximated by the data.

The eight most probable graphs, according to the reversible jump results, are tabulated in Table 3.26 along with their posterior probabilities and their standard errors (in parentheses). All other graphs have probabilities of less than 0.02. The edge inclusion percentages are tabulated in Table 3.25. The graph with the highest probability, 25139, is intuitively reasonable with RV associated only with

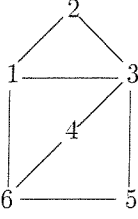
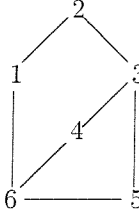
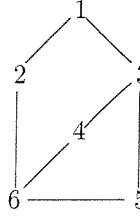
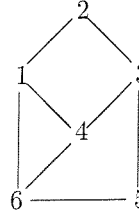
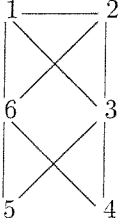
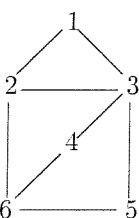
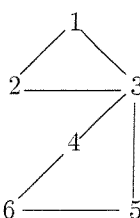
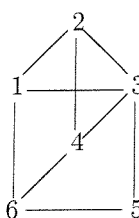
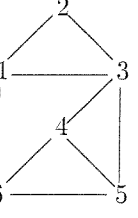
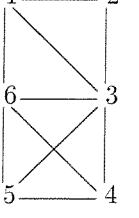
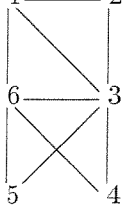
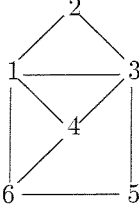
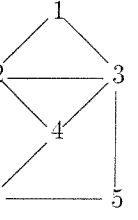
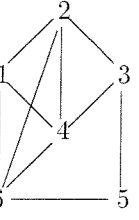
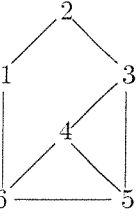
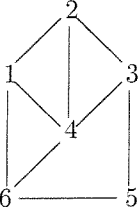
			
26163 0.095 (219)	17971 0.047 (90)	24691 0.046 (150)	22067 0.035 (180)
			
26227 0.026 (97)	25203 0.025 (87)	25139 0.021 (85)	26419 0.019 (81)
			
26167 0.018 (115)	26175 0.017 (87)	26171 0.015 (58)	30259 0.014 (51)
			
25459 0.013 (59)	26483 0.013 (80)	17975 0.012 (86)	22323 0.010 (78)

Table 3.24: Posterior model probabilities for Fowl Bones example with standard errors $\times 10^4$ in brackets

edge	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)	(2,3)	(2,4)	(2,5)	(2,6)
P	99.99	99.38	9.05	18.12	6.26	99.90	6.50	13.12	12.13
	(3,4)	(3,5)	(3,6)	(4,5)	(4,6)	(5,6)			
	99.99	99.96	6.14	6.89	100	99.94			

Table 3.25: Edge inclusion %ages for Voting example

RPA and RPI, and a cycle of (RPI, FPI, FSC, RSC). The model selected by Wermuth is 25143, which has the additional edge (RSC, FPI). However, Wermuth's approach is confined to decomposable models and, indeed, 25143 is the decomposable graph with the highest probability, with the extra edge breaking the 4-cycle in 25139.

3.6 Model-averaged predictive distributions

Section 1.6 described how model-averaged predictive distributions for a subset \mathbf{Y}_1 of the variables given future values \mathbf{y}_2 of the remaining variables may be estimated from the reversible jump output as the average over $(m, \boldsymbol{\theta})$ of $f(\mathbf{y}_1|m, \boldsymbol{\theta}_m, \mathbf{y}_2^{n+1})$. This is particularly simple for GGMs as these conditional distributions are themselves Normal. The standard result is that if $(\mathbf{Y}_1, \mathbf{Y}_2)$ are distributed jointly as multivariate Normal with mean and variance

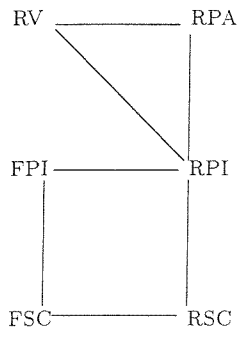
$$(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2)' \quad \text{and} \quad \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix},$$

then the distribution of \mathbf{Y}_1 given $\mathbf{Y}_2 = \mathbf{y}_2$ is Normal with mean

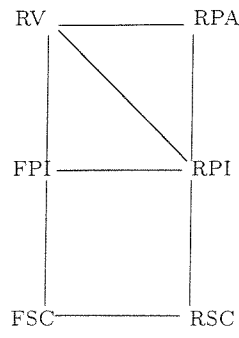
$$\boldsymbol{\mu}_{1|2} = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{y}_2 - \boldsymbol{\mu}_2) \quad \text{and} \quad \text{variance} \quad \boldsymbol{\Sigma}_{2|1} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}.$$

The procedure is then as follows: for each iterate i of N iterations of reversible jump MCMC output (which, recall, consists of $(m^i, \boldsymbol{\mu}^i, \boldsymbol{\Omega}^i)$), obtain $\boldsymbol{\mu}_{1|2}$ and $\boldsymbol{\Sigma}_{1|2}$ as above. This gives the density $f^i(\mathbf{y}_1|\boldsymbol{\mu}_{1|2}, \boldsymbol{\Sigma}_{1|2})$. $f(\mathbf{y}_1|m, \boldsymbol{\theta}_m, \mathbf{y}_2^{n+1})$ is then estimated by $\frac{1}{N} \sum_{i=1}^N f^i(\mathbf{y}_1|\boldsymbol{\mu}_{1|2}, \boldsymbol{\Sigma}_{1|2})$. If \mathbf{y}_2 are taken to be values from the data, the method can be checked by comparing the predictive distribution with the observed \mathbf{y}_1 . The data used to obtain the RJ sample should however exclude the observations used for prediction in order to avoid any "double counting".

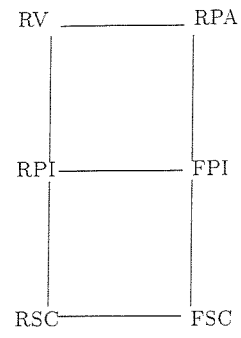
To illustrate, this procedure was applied to the heads data (Section 3.5.3)



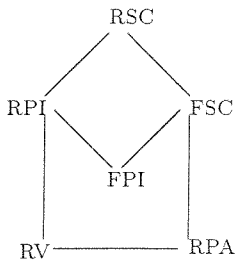
25139
0.456 (156)



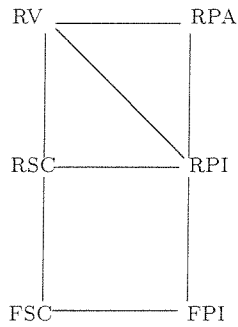
27187
0.098 (109)



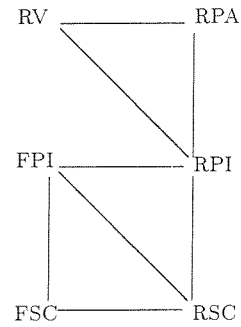
25267
0.069 (82)



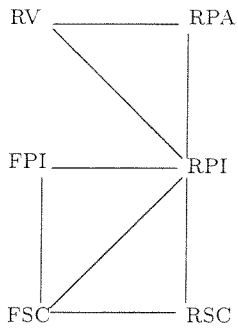
25203
0.054 (58)



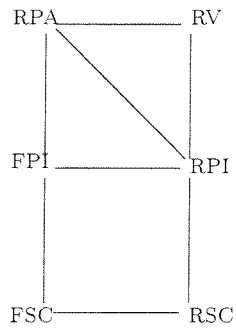
29235
0.038 (40)



25143
0.033 (35)



25147
0.030(34)



25395
0.023 (40)

Table 3.26: Posterior model probabilities for voting habits example with standard errors $\times 10^4$ in brackets

to estimate the predictive density $f(Y_1|y_2, y_3, y_4, \mathbf{y})$. This density is univariate so plots may be obtained and visually compared with the observed values. In addition, we can compare estimated predictive densities based on using only decomposable graphs and those based on using all graphs. This is potentially quite useful in this example as the two most probable graphs, based on the reversible jump output, are nondecomposable. Plots of both densities for each of two randomly selected sets of observations from the data are presented in Figures 3.9 and 3.10. The actual observed values lie well within the mass of the distributions in both cases although the distributions for the first are centred closer to the observed value. These plots suggest that, as would be expected, that prediction based on all graphs is better than that based on only decomposable graphs. There is not a great difference between the two here but it must be noted that there are only three nondecomposable graphs for $q = 4$ and the proportion of nondecomposable graphs is much greater for greater values of q and hence the differences in prediction will be much greater, especially when, as in many of the examples in the previous section, the most probable graphs are nondecomposable.

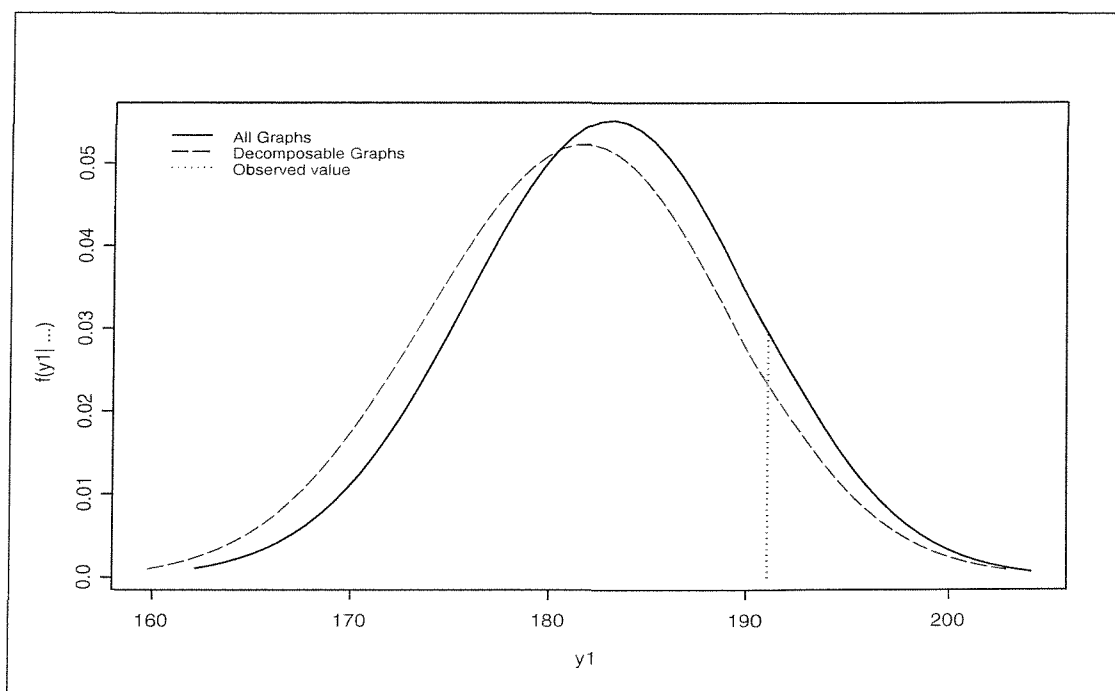


Figure 3.9: Model-averaged predictive densities for Y_1 (1)

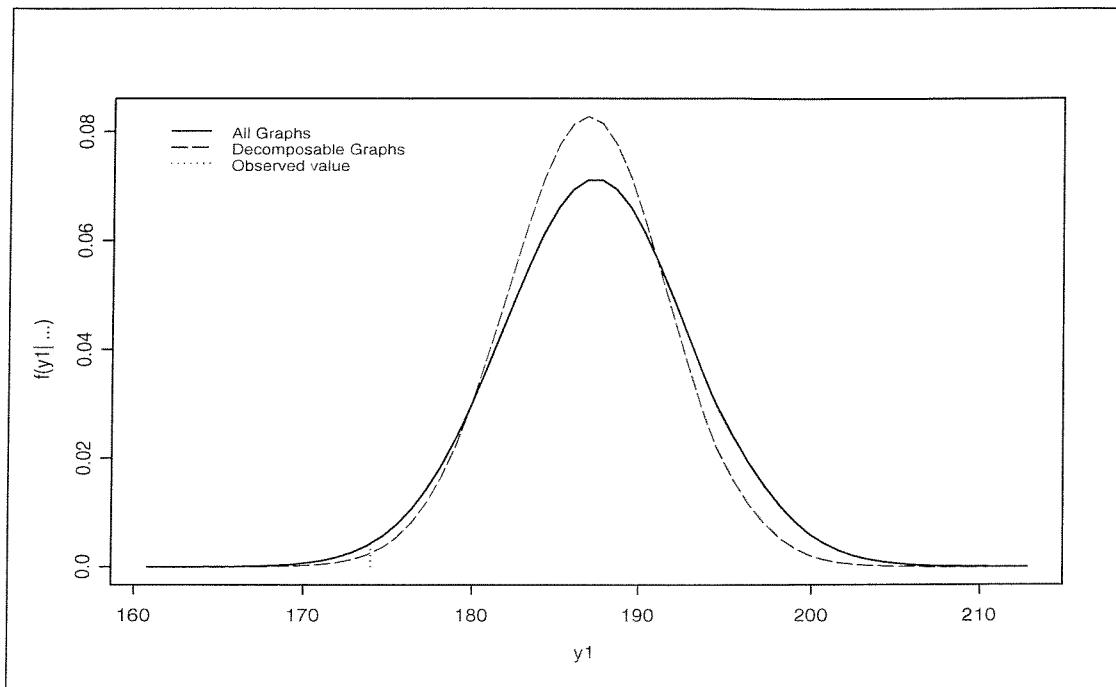


Figure 3.10: Model-averaged predictive densities for Y_1 (2)

3.7 Summary

The various examples presented above show that the approach and particularly the reversible jump sampler described in the previous chapter are of practical use. The results in each case are sensible and agree with those of other authors, or at least with MIM where these are not available. The sampler itself is easy to implement and very quick to run - run time for 100 000 iterations in each case is no more than a few minutes on an ordinary desktop computer. It is also easily adapted for other priors (even conjugate priors), for block updating, for zero-mean models and, as the last three examples show, for cases where only observed correlations are available. Indeed, this last point suggests that this may be all that is required if interest lies only in inference regarding the partial correlations and model choice. It is also relatively easy to restrict the sampler to decomposable models, if desired, as in the Fret's example but there is clearly no reason to do this in general. Indeed, the lack of need of restriction to decomposable models, and hence the use of a richer class of models, is probably the most important aspect of this approach as this restriction is usually made only on theoretical or computational grounds but makes little practical sense. In addition, use of the entire class of models potentially gives better predictions. The previous section demonstrated this in a very limited way but the differences in prediction are likely to be similar or greater in general.

Chapter 4

Mixed Graphical Models

4.1 Introduction

Graphical association models were introduced by Lauritzen and Wermuth (1989) as models for mixed qualitative and quantitative (discrete and continuous) variables. This class of models includes graphical log-linear models and graphical Gaussian models as special cases. They were then extended by Edwards (1990) to the class of hierarchical association models. This section and the next summarise the relevant material from these two papers with some minor changes in notation. Some of this material can seem quite obscure at first sight but the examples presented in Sections 5.5 and 6.4 as well as the running examples should help clarify matters.

The sets of discrete and continuous variables are denoted by Δ and Γ and their sizes by p and q respectively. When $q = 0$, the class of graphical association models reduces to the class of graphical log-linear models and when $p = 0$, it reduces to the class of graphical Gaussian models. The term *mixed graphical model* is used here to exclude both of these pure cases. Lauritzen and Wermuth (1989) as well as Wermuth and Lauritzen (1989) also consider directed versions as well as mixed graphical chain models but these will not be considered here.

A typical observation is $\mathbf{x} = (\mathbf{i}, \mathbf{y})$, where \mathbf{i} is discrete-valued and \mathbf{y} is real-valued. \mathbf{x} takes values in the space $\mathcal{I} \times \mathbf{R}$, where $\mathcal{I} = (\times_{\delta \in \Delta} I_{\delta})$ and I_{δ} is the set of levels of discrete variable δ . Similarly, for $a \subseteq \Delta$ and $b \subseteq \Gamma$, \mathbf{i}_a and \mathbf{y}_b denote subvectors corresponding to a and b .

The conditional independences in a graphical model can be represented by a marked graph, in which the two types of variable are distinguished by two types of vertex: Conventionally, discrete variables are represented by dots and continuous variables by circles. Discrete variables are denoted as I_1, I_2 etc. and continuous

variables by Y_1, Y_2 etc., although in specific examples with small numbers of variables, A, B and C and X, Y and Z will be used, respectively, for the two types.

4.2 The Conditional Gaussian Distribution

Graphical association models and hierarchical interaction models are based on the conditional Gaussian (CG) distribution, which is defined by letting the marginal distribution of Δ be arbitrary and the conditional distribution of Γ given Δ be multivariate Normal (Gaussian), that is $P(\mathbf{I} = \mathbf{i}) = p(\mathbf{i})$ and $(\mathbf{Y}|\mathbf{I} = \mathbf{i}) \sim N_q(\boldsymbol{\mu}(\mathbf{i}), \boldsymbol{\Sigma}(\mathbf{i}))$. The term ‘‘CG Model’’ will be used here to denote such models.

A special case is the *homogeneous* conditional Gaussian (HCG) distribution in which the variance of the conditional variables does not depend on \mathbf{i} , that is $\boldsymbol{\Sigma}(\mathbf{i}) \equiv \boldsymbol{\Sigma}$. This type of model is discussed in Olkin and Tate (1961).

$\{p(\mathbf{i}), \boldsymbol{\mu}(\mathbf{i}), \boldsymbol{\Sigma}(\mathbf{i})\}_{\mathbf{i} \in \mathcal{I}}$, collectively known as the moment parameters, are the cell probabilities, conditional means and conditional variances respectively. The $\{p(\mathbf{i})\}$ are positive scalars such that $\sum_{\mathbf{i} \in \mathcal{I}} p(\mathbf{i}) = 1$, $\{\boldsymbol{\mu}(\mathbf{i})\}$ are $q \times 1$ real-valued vectors and $\{\boldsymbol{\Sigma}(\mathbf{i})\}$ are positive definite, symmetric $q \times q$ matrices.

The joint density is thus

$$f(\mathbf{x}) = f(\mathbf{i}, \mathbf{y}) = p(\mathbf{i})(2\pi)^{-q/2} |\boldsymbol{\Sigma}(\mathbf{i})|^{-1/2} \exp \left[-\frac{1}{2} (\mathbf{y} - \boldsymbol{\mu}(\mathbf{i}))' \boldsymbol{\Sigma}(\mathbf{i})^{-1} (\mathbf{y} - \boldsymbol{\mu}(\mathbf{i})) \right] \quad (4.1)$$

which can be rewritten as

$$f(\mathbf{x}) = f(\mathbf{i}, \mathbf{y}) = \exp \left[g(\mathbf{i}) + \mathbf{h}(\mathbf{i})' \mathbf{y} - \frac{1}{2} \mathbf{y}' \boldsymbol{\Omega}(\mathbf{i}) \mathbf{y} \right] \quad (4.2)$$

where

$$\begin{aligned} g(\mathbf{i}) &= \log(p(\mathbf{i})) - \frac{q}{2} \log(2\pi) - \frac{1}{2} \log(|\boldsymbol{\Sigma}(\mathbf{i})|) - \frac{1}{2} \boldsymbol{\mu}(\mathbf{i})' \boldsymbol{\Sigma}(\mathbf{i})^{-1} \boldsymbol{\mu}(\mathbf{i}) \\ \mathbf{h}(\mathbf{i}) &= \boldsymbol{\Sigma}(\mathbf{i})^{-1} \boldsymbol{\mu}(\mathbf{i}) \\ \boldsymbol{\Omega}(\mathbf{i}) &= \boldsymbol{\Sigma}(\mathbf{i})^{-1} \end{aligned}$$

Equation 4.2 can be rewritten as

$$f(\mathbf{x}) = f(\mathbf{i}, \mathbf{y}) = \exp \left[g(\mathbf{i}) + \sum_{\gamma \in \Gamma} h_{\gamma}(\mathbf{i}) y_{\gamma} - \frac{1}{2} \sum_{\gamma \in \Gamma} \sum_{\zeta \in \Gamma} \omega_{\gamma\zeta}(\mathbf{i}) y_{\gamma} y_{\zeta} \right] \quad (4.3)$$

$\{g(\mathbf{i}), \mathbf{h}(\mathbf{i}), \mathbf{\Omega}(\mathbf{i})\}_{\mathbf{i} \in \mathcal{I}}$, collectively known as the canonical parameters, are the discrete and linear canonical parameters and the conditional precision matrix respectively.

The notation, $\mathbf{\Omega}$, will be used to represent the collection of $\mathbf{\Omega}(\mathbf{i})$'s as the i th entry. This is to distinguish it from $\mathbf{\Omega}$, which denotes a single common precision matrix in a HCG model.

Using the convention that $\theta^a(\mathbf{x}) = \theta^a(\mathbf{x}^a)$ etc., expansions can be made as follows:

$$g(\mathbf{i}) = \sum_{a \subseteq \Delta} \lambda^a(\mathbf{i}) \quad \mathbf{h}(\mathbf{i}) = \sum_{a \subseteq \Delta} \boldsymbol{\eta}^a(\mathbf{i}) \quad \mathbf{\Omega}(\mathbf{i}) = \sum_{a \subseteq \Delta} \boldsymbol{\Psi}^a(\mathbf{i})$$

Or equivalently,

$$\begin{aligned} g(\mathbf{i}) &= \sum_{a \subseteq \Delta} \lambda^a(\mathbf{i}) \\ h_{\gamma}(\mathbf{i}) &= \sum_{a \subseteq \Delta} \eta_{\gamma}^a(\mathbf{i}) \quad \forall \gamma \in \Gamma \\ \omega_{\gamma\zeta}(\mathbf{i}) &= \sum_{a \subseteq \Delta} \psi_{\gamma\zeta}^a(\mathbf{i}) \quad \forall \gamma, \zeta \in \Gamma \end{aligned}$$

These terms on the right hand side are called *interactions* and there are $(1 + q + \binom{q+1}{2})$ sets of $|\mathcal{I}|$ of them - one for the discrete variables, one for each continuous and one for each pair of (not necessarily distinct) continuous variables. They are given special names as follows:

λ^{\emptyset} is the *log normalizing constant*. Since $\sum_{\mathbf{i}} p(\mathbf{i}) = 1$,

$$\exp(\lambda^{\emptyset}) = \left[(2\pi)^{q/2} \sum_{\mathbf{i}} |\mathbf{\Omega}(\mathbf{i})|^{-1/2} \exp\left(\alpha(\mathbf{i}) + \frac{1}{2} \mathbf{h}(\mathbf{i})' \mathbf{\Omega}(\mathbf{i})^{-1} \mathbf{h}(\mathbf{i})\right) \right]^{-1}, \quad (4.4)$$

where $\alpha(\mathbf{i}) = \sum_{a \subseteq \Delta \setminus \{\emptyset\}} \lambda^a(\mathbf{i}) \equiv g(\mathbf{i}) - \lambda^{\emptyset}$.

The λ^a , $a \neq \emptyset$ terms are (pure) *discrete interactions*. When $|a| = 1$, they are *discrete main effects*;

The η^{\emptyset} terms are *continuous main effects*. The remaining η^a terms are *mixed linear interactions* between a continuous variable and the discrete variables in a .

$\boldsymbol{\Psi}^a$ are *quadratic interaction matrices*. The ψ^{\emptyset} terms are *pure quadratic interactions* between pairs of continuous variables. The remaining ψ^a terms are mixed quadratic interactions between pairs of continuous variables and the discrete variables in a . Note that a CG distribution is HCG if and only if it has no mixed quadratic interactions.

Inserting the interaction terms into equation(4.2) or (4.3) we get the following representation of the logarithm of the density:

$$\log f(\mathbf{x}) = \log f(\mathbf{i}, \mathbf{y}) = \sum_{a \subseteq \Delta} \lambda^a(\mathbf{i}) + \sum_{a \subseteq \Delta} \sum_{\gamma \in \Gamma} \eta_\gamma^a(\mathbf{i}) y_\gamma - \frac{1}{2} \sum_{a \subseteq \Delta} \sum_{\gamma, \zeta \in \Gamma} \psi_{\gamma\zeta}^a(\mathbf{i}) y_\gamma y_\zeta. \quad (4.5)$$

In practice, additional constraints are required for identifiability of the interactions. One possibility are "sum-to-zero" constraints so that

$$\begin{aligned} \sum_{\mathbf{i} \in \mathcal{I}} \lambda^a(\mathbf{i}) &= 0 \quad \forall a \subseteq \Delta \setminus \{\emptyset\} \\ \sum_{\mathbf{i} \in \mathcal{I}} \eta_\gamma^a(\mathbf{i}) &= 0 \quad \forall a \subseteq \Delta \setminus \{\emptyset\} \quad \forall \gamma \in \Gamma \\ \sum_{\mathbf{i} \in \mathcal{I}} \psi_{\gamma\zeta}^a(\mathbf{i}) &= 0 \quad \forall a \subseteq \Delta \setminus \{\emptyset\} \quad \forall \gamma, \zeta \in \Gamma \end{aligned}$$

To illustrate, let $\Delta = \{A\}$ with $I_A = \{0, 1\}$, that is a binary discrete variable. The λ -terms are $\lambda^\emptyset, \lambda^A(0)$ and $\lambda(1)$. Sum-to-zero constraints require $\lambda^A(0) = -\lambda^A(1)$. This reduces the number of parameters by one and it will not matter which we choose to vary freely and which is determined by constraint. This is the principal motivation for this type of identifiability constraint. In general, the number of λ -parameters is the number of cells, although one of these is λ^\emptyset and is constrained as shown above. The same applies to the η - and Ψ -parameters. The parameterization can thus be written $(\mathbf{\Lambda}, \mathbf{H}, \Phi)$, where $\mathbf{\Lambda} = (\lambda^\emptyset \quad \lambda^A(0))'$, $\mathbf{H} = (\eta \quad \eta^A(0))'$ and $\Phi = (\Psi \quad \Psi^A(0))'$. This accounts for the identifiability constraints and the sum-to-one constraints on the probabilities.

The identifiability constraints may be embodied using a *design matrix* \mathbf{D} and the transformations,

$$\mathbf{G} = \mathbf{D}\mathbf{\Lambda} \quad (4.6)$$

$$\mathbf{h}_\gamma = \mathbf{D}\mathbf{H}_\gamma \quad \forall \gamma \in \Gamma \quad (4.7)$$

$$\boldsymbol{\omega}_{\gamma\zeta} = \mathbf{D}\boldsymbol{\psi}_{\gamma\zeta} \quad \forall \gamma, \zeta \in \Gamma \quad (4.8)$$

where \mathbf{G} , \mathbf{h}_γ and $\boldsymbol{\omega}_{\gamma\zeta}$ are $|\mathcal{I}| \times 1$ vectors of $g(\mathbf{i})$'s, $h_\gamma(\mathbf{i})$'s and $\omega_{\gamma\zeta}(\mathbf{i})$'s respectively. The expressions 4.7 and 4.8 may be written concisely as

$$\mathbf{h} = \mathbf{D}\mathbf{H} \quad (4.9)$$

$$\boldsymbol{\Omega} = \mathbf{D}\tilde{\boldsymbol{\Psi}}, \quad (4.10)$$

where $\tilde{\Psi}$ is an array of Ψ^d 's.

Conversely, the inverse of the design matrix can be used to transform canonical parameters to interactions as follows:

$$\mathbf{A} = \mathbf{D}^{-1}\mathbf{G} \quad (4.11)$$

$$\mathbf{H} = \mathbf{D}^{-1}\mathbf{h} \quad (4.12)$$

$$\tilde{\Psi} = \mathbf{D}^{-1}\Omega \quad (4.13)$$

The remaining (constrained) interactions may be obtained from the $\mathbf{D}(\delta)$ matrices.

For sum-to-zero constraints, \mathbf{D} is given by

$$\mathbf{D} = \bigotimes_{\delta \in \Delta} \mathbf{D}(\delta),$$

the outer product of the $|\delta| \times |\delta|$ matrices, one for each discrete variable, of form

$$\mathbf{D}(\delta) = \left(\begin{array}{c|ccc} 1 & & & \\ 1 & & & \\ \vdots & & & \\ 1 & -1 & \dots & -1 \end{array} \right)$$

where I is an identity matrix.

Also,

$$\mathbf{D}^{-1} = \bigotimes_{\delta \in \Delta} \mathbf{D}(\delta)^{-1}. \quad (4.14)$$

4.3 Graphical, Hierarchical and other models

Models are specified by setting certain interactions to zero. The fundamental result for graphical models in Lauritzen and Wermuth (1989) is that two variables are conditionally independent given the remaining variables if and only if all interaction terms involving these two are zero. The Markov properties can be stated explicitly as follows:

$$I_j \perp\!\!\!\perp I_k | (\mathbf{I}_{\Delta \setminus \{j,k\}}, \mathbf{Y}) \iff \lambda^a = 0, \boldsymbol{\eta}^a = \mathbf{0}, \boldsymbol{\Psi}^a = \mathbf{0} \quad (4.15)$$

whenever $j, k \in a \subseteq \Delta$

$$I_j \perp\!\!\!\perp Y_k | (\mathbf{I}_{\Delta \setminus \{j\}}, \mathbf{Y}_{\Gamma \setminus \{k\}}) \iff \eta_k^a = 0, \psi_{kr}^a = 0 \quad \forall r \in \Gamma \quad (4.16)$$

whenever $j \in a \subseteq \Delta$

$$Y_j \perp\!\!\!\perp Y_k | (\mathbf{I}, \mathbf{Y}_{\Gamma \setminus \{j,k\}}) \iff \psi_{jk}^a = 0 \quad \forall a \subseteq \Delta \quad (4.17)$$

Here, $\boldsymbol{\eta}^a$ denotes the array $(\eta^a(\mathbf{i}), \mathbf{i} \in \mathcal{I})$ and $\boldsymbol{\Psi}^a$ the array $(\Psi^a(\mathbf{i}), \mathbf{i} \in \mathcal{I})$.

Analogously to the case of log-linear models, the class of graphical association models is part of the larger class of hierarchical association models, which allow higher order interactions to be removed (set to zero) without removing the interactions they contain. The principle which must be respected is the same as in hierarchical log-linear models, namely that if an interaction term is removed, so must all other interaction terms involving all the same variables that it does. This is sometimes called the marginality principle. More precisely, the requirements are,

1. If $a \subseteq b \subseteq \Delta$ and $\{\lambda^a(\mathbf{i}) : \mathbf{i} \in \mathcal{I}\}$ are set to zero, then $\{\lambda^b(\mathbf{i}) : \mathbf{i} \in \mathcal{I}\}$, $\{\eta_\gamma^b(\mathbf{i}) : \mathbf{i} \in \mathcal{I}\}$ and $\{\psi_{\gamma\zeta}^b(\mathbf{i}) : \mathbf{i} \in \mathcal{I}\}$ must also be set to zero for all $\gamma, \zeta \in \Gamma$.
2. If $a \subseteq b \subseteq \Delta$, $\gamma \in \Gamma$ and $\{\eta_\gamma^a(\mathbf{i}) : \mathbf{i} \in \mathcal{I}\}$ are set to zero, then $\{\eta_\gamma^b(\mathbf{i}) : \mathbf{i} \in \mathcal{I}\}$ and $\{\psi_{\gamma\zeta}^b(\mathbf{i}) : \mathbf{i} \in \mathcal{I}\}$ must also be set to zero for all $\zeta \in \Gamma$.
3. If $a \subseteq b \subseteq \Delta$, $\gamma, \zeta \in \Gamma$ and $\{\psi_{\gamma\zeta}^a(\mathbf{i}) : \mathbf{i} \in \mathcal{I}\}$ are set to zero, then $\{\psi_{\gamma\zeta}^b(\mathbf{i}) : \mathbf{i} \in \mathcal{I}\}$ must also be set to zero.
4. If $a \subseteq b \subseteq \Delta$, $\gamma \in \Gamma$ and $\{\psi_{\gamma\gamma}^a(\mathbf{i}) : \mathbf{i} \in \mathcal{I}\}$ are set to zero, then $\{\psi_{\gamma\zeta}^b(\mathbf{i}) : \mathbf{i} \in \mathcal{I}\}$ must also be set to zero for all $\zeta \in \Gamma$.

This is how a hierarchical interaction model is defined by Edwards (1990) however Lauritzen (1996) defines them using only the first three conditions, referring to models satisfying all four as ‘‘MIM models’’. It is noted in the discussion of Edwards (1990) that the condition 4 is hard to justify and there are sensible models which are hierarchical in Lauritzen’s sense. Despite this, the hierarchical models considered in this thesis are the ‘‘MIM’’ models, satisfying all four requirements. As noted in the next two chapters, the methods described there can be easily

adapted to deal with the other type of hierarchical model. Indeed, they are easier to deal with since they are less restrictive.

We must be careful to distinguish between interactions and interaction *terms*, which are parameters quantifying interactions for each cell. Interactions between discrete variables will be denoted as AB , ABC etc. Linear interactions between a continuous variable and a set of discrete variables will be denoted as AY , ABY etc. Interactions between two continuous variables will be denoted as XY etc. Quadratic interactions between a continuous and a set of discrete variables as AYY , $ABYY$ etc. and between two continuous and a set of discrete as AXY , $ABXY$ etc.

The class of CG models is a very flexible class of models (arguably, too flexible) and contains other standard classes of models, including standard MANOVA, Multivariate regression, general location (GLOM) and location-scale models. Edwards (1990,1995) provides a good description of the connections to some of these (but not the latter two) as well as to others. For convenience brief descriptions follow:

The general MANOVA (Morrison 1976), or multivariate ANOVA, model may be written

$$\mathbf{Y} = \mathbf{D}\boldsymbol{\theta} + \boldsymbol{\epsilon},$$

where \mathbf{Y} is a $n \times q$ matrix of responses, \mathbf{D} is a $n \times k$ design matrix, $\boldsymbol{\theta}$ is a $k \times q$ parameter matrix and $\boldsymbol{\epsilon}$ is a random error matrix with $E(\boldsymbol{\epsilon}) = \mathbf{0}$ and $\text{var}(\boldsymbol{\epsilon}) = \mathbf{I}_n \otimes \boldsymbol{\Sigma}$. Hypotheses to be tested are expressed as $\mathbf{C}\boldsymbol{\theta}\mathbf{M} = \mathbf{0}$, where \mathbf{C} and \mathbf{M} are given matrices. If $\mathbf{M} = \mathbf{I}$, we have a standard MANOVA model. This can be related to HCG models by setting the columns of $\boldsymbol{\theta}$ to be interaction expansions of the cell means rather than of the linear canonical parameters. In this way, mean-linear HCG models correspond to linear hypotheses, $\mathbf{C}\boldsymbol{\theta} = \mathbf{0}$.

The general location model (GLOM) is a class of model for dealing with both discrete and continuous variables which has been used since before the introduction of association models but in fact corresponds to the class of HCG models. Barnard et al. (2000) discuss this model class and introduce a more flexible generalization based on their “separation strategy”. Interestingly, they remark, “While it is clearly inadequate in many applications to assume common covariance across cells, especially when there are many cells, it is also clear that allowing each cell to have its own covariance matrix is impractical. For one thing, such a model often has many more parameters than data points.” This their motivation for

the generalization. The next two chapters attempt to show that this concern notwithstanding, general CG models have use at least for relatively small numbers of cells and continuous variables. All examples examined also have numbers of observations sufficiently large relative to the number of parameters.

The general location-scale model introduced by Barnard et al. (2000) is similar to the extensions to GLOM models introduced by Liu and Rubin (1998), which have different but proportional covariance matrices. The general location-scale models allow standard deviations to vary across cells but assume a common correlation matrix. This is a compromise between the above extension and CG models which may be useful if the number of parameters is a concern. This suggests a similar generalization to the HCG models, where the partial precisions may vary between cells but a common partial correlation matrix is assumed. This corresponds to absence of all mixed quadratic interactions involving more than one continuous variable (e.g. AXY , $ABXY$, etc.), which is easily dealt with in the context of the methods described in the next two chapters. They also suggest a partitioning of the cells into groups, with a separate correlation matrix for each group, if the common correlation assumption is too restrictive.

4.4 An alternative parameterization

While the moment parameterization is the most intuitive and most easily interpretable, the interaction expansion has the advantage of the simplicity of the graphical constraints. However, while these constraints are simple, they greatly restrict the parameter space of the quadratic interaction parameters and, as shown in Section 4.6, make prior specification difficult.

A simple solution is to instead use the quadratic canonical parameters, namely the conditional precision matrices, $\mathbf{\Omega}$ in place of quadratic interaction matrices. The log-density in this case is:

$$\log f(\mathbf{x}) = \log f(\mathbf{i}, \mathbf{y}) = \sum_{a \subseteq \Delta} \lambda^a(\mathbf{i}) + \sum_{a \subseteq \Delta} \sum_{\gamma \in \Gamma} \eta_{\gamma}^a(\mathbf{i}) y_{\gamma} - \frac{1}{2} \mathbf{y}' \mathbf{\Omega}(\mathbf{i}) \mathbf{y}. \quad (4.18)$$

The constraints on $\mathbf{\Omega}$ corresponding to constraints on the quadratic interactions may be expressed concisely using *constraint matrices*. Recall that $\omega_{\gamma\zeta} = \mathbf{D}\psi_{\gamma\zeta} \quad \forall \gamma, \zeta \in \Gamma$. A constraint matrix corresponding to a given set of missing interaction terms is obtained from the design matrix \mathbf{D} by setting certain columns, corresponding to interactions to be set to zero, to zero.

For example, if there are two discrete variables, A and B , each with two levels, the constraints for a missing AY interaction are expressed as

$$\begin{pmatrix} \omega_Y(0,0) \\ \omega_Y(0,1) \\ \omega_Y(1,0) \\ \omega_Y(1,1) \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \end{pmatrix} \begin{pmatrix} \psi_Y \\ \psi_Y^B \\ \psi_Y^A \\ \psi_Y^{AB} \end{pmatrix}$$

The constraint matrix is the matrix on the right. If we denote this by \mathbf{K} and the vector on the left as $\boldsymbol{\omega}_Y$, we have

$$\boldsymbol{\omega}_Y = \mathbf{K}\mathbf{D}^{-1}\boldsymbol{\omega}_Y$$

It is easy to see that the constraints for a missing continuous-continuous edge (or missing continuous-continuous interaction in a hierarchical model) are given by

$$Y_j \perp\!\!\!\perp Y_k | (\mathbf{I}, \mathbf{Y}_{\Gamma \setminus \{j,k\}}) \iff \omega_{jk}(\mathbf{i}) = \omega_{kj}(\mathbf{i}) = 0 \quad \forall \mathbf{i} \in \mathcal{I} \quad (4.19)$$

Less obvious but straightforward to see is that the constraints for a missing AXY interaction are as follows, where A is indexed by i_A so $\mathbf{i} \equiv (i_A, \mathbf{j})$:

$$\omega_{XY}(\mathbf{i}) = \omega_{XY}(0, \mathbf{j}) \quad \forall \mathbf{j}. \quad (4.20)$$

To illustrate the constraints for a missing $ABXY$ interaction, consider again the case of two binary variables, A and B , where the AB interaction is absent. In this case we have

$$\begin{aligned} \boldsymbol{\omega}_{XY} &= \frac{1}{4} \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & -1 & 1 & 0 \\ 1 & 1 & -1 & 0 \\ 1 & -1 & -1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix} \boldsymbol{\omega}_{XY} \\ &= \frac{1}{4} \begin{pmatrix} 3 & 1 & 1 & -1 \\ 1 & 3 & -1 & 1 \\ 1 & -1 & 3 & 1 \\ -1 & 1 & 1 & 3 \end{pmatrix} \boldsymbol{\omega}_{XY} \end{aligned}$$

$$\text{So, } \omega_{XY}(i, j) = \frac{1}{4} [3\omega_{XY}(i, j) + \omega_{XY}(1-i, j) + \omega_{XY}(i, 1-j) - \omega_{XY}(1-i, 1-j)].$$

The four constraints are in fact all the same:

$$\omega_{XY}(0,0) - \omega_{XY}(0,1) - \omega_{XY}(1,0) + \omega_{XY}(1,1) = 0.$$

When A has l_A levels and B has l_B levels, we have

$$\begin{aligned} \omega_{XY}(i, j) = & \frac{1}{l_A l_B} \left[(l_A + l_B - 1) \omega_{XY}(i, j) + (l_A - 1) \sum_{b \neq j} \omega_{XY}(i, b) \right. \\ & \left. + (l_B - 1) \sum_{a \neq i} \omega_{XY}(a, j) - \sum_{a \neq i, b \neq j} \omega_{XY}(a, b) \right] \end{aligned}$$

and hence

$$\begin{aligned} \omega_{XY}(i, j) = & \frac{1}{(l_A - 1)(l_B - 1)} \left[(l_A - 1) \sum_{b \neq j} \omega_{XY}(i, b) \right. \\ & \left. + (l_B - 1) \sum_{a \neq i} \omega_{XY}(a, j) - \sum_{a \neq i, b \neq j} \omega_{XY}(a, b) \right] \quad (4.21) \end{aligned}$$

This gives $(l_A - 1)(l_B - 1)$ independent constraints.

It can be shown that the same result holds in general, with extra indices for further discrete variables inserted, for fixed levels of the remaining discrete variables.

Finally, the following holds for a missing 3-way discrete interaction:

$$\begin{aligned} \Omega(i, j, k) = & \frac{1}{l_A l_B l_C} \left[(l_A + l_B + l_C - 1) \Omega(i, j, k) + (l_A - 1)(l_B - 1) \sum_{c \neq k} \Omega(i, j, c) \right. \\ & + (l_A - 1)(l_C - 1) \sum_{b \neq j} \Omega(i, b, k) + (l_B - 1)(l_C - 1) \sum_{a \neq i} \Omega(a, j, k) \\ & - (l_A - 1) \sum_{b \neq j, c \neq k} \Omega(i, b, c) - (l_B - 1) \sum_{a \neq i, c \neq k} \Omega(a, j, c) \\ & \left. - (l_C - 1) \sum_{a \neq i, b \neq j} \Omega(a, b, k) + \sum_{a \neq i, b \neq j, c \neq k} \Omega(a, b, c) \right] \end{aligned}$$

A pattern can be seen to emerge from this, from which it may be possible to express constraints for missing higher order interactions or even for arbitrary missing interactions but this will not be pursued here.

4.5 The Likelihood

Suppose there are n independent observations, $\mathbf{x} = (\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^n) = ((i^1, \mathbf{y}^1), (i^2, \mathbf{y}^2), \dots, (i^n, \mathbf{y}^n))$, on \mathbf{X} , then the logarithm of the likelihood function is

$$\log L = \sum_{k=1}^n \left[g(\mathbf{i}^k) + \mathbf{h}(\mathbf{i}^k)' \mathbf{y}^k - \frac{1}{2} \text{tr} \left(\sum_{k=1}^n \boldsymbol{\Omega}(\mathbf{i}^k) \mathbf{y}^k (\mathbf{y}^k)' \right) \right] \quad (4.22)$$

$$= \sum_{k=1}^n \left[g(\mathbf{i}^k) + \sum_{\gamma \in \Gamma} h_{\gamma}(\mathbf{i}^k) y_{\gamma}^k - \frac{1}{2} \sum_{\gamma \in \Gamma} \sum_{\zeta \in \Gamma} \omega_{\gamma\zeta}(\mathbf{i}^k) y_{\gamma}^k y_{\zeta}^k \right] \quad (4.23)$$

$$= \sum_{\mathbf{i} \in \mathcal{I}} \left[g(\mathbf{i}) n(\mathbf{i}) + \sum_{\gamma \in \Gamma} h_{\gamma}(\mathbf{i}) S_{\gamma}(\mathbf{i}) - \frac{1}{2} \sum_{\gamma \in \Gamma} \sum_{\zeta \in \Gamma} \omega_{\gamma\zeta}(\mathbf{i}) P_{\gamma\zeta}(\mathbf{i}) \right] \quad (4.24)$$

$$= \sum_{\mathbf{i} \in \mathcal{I}} \left[g(\mathbf{i}) n(\mathbf{i}) + \mathbf{h}(\mathbf{i})' \mathbf{S}(\mathbf{i}) - \frac{1}{2} \text{tr} (\boldsymbol{\Omega}(\mathbf{i}) \mathbf{P}(\mathbf{i})) \right] \quad (4.25)$$

where

$$n(\mathbf{i}) = \sum_{k: \mathbf{i}_k = \mathbf{i}} 1$$

the number of observations in cell \mathbf{i} ;

$$\mathbf{S}(\mathbf{i}) = \sum_{k: \mathbf{i}_k = \mathbf{i}} \mathbf{y}^k$$

the sum of the y -vectors in cell \mathbf{i} , with components

$$S_{\gamma}(\mathbf{i}) = \sum_{k: \mathbf{i}_k = \mathbf{i}} y_{\gamma}^k;$$

$$\mathbf{P}(\mathbf{i}) = \sum_{k: \mathbf{i}_k = \mathbf{i}} (\mathbf{y}^k)' \mathbf{y}^k$$

the matrix of sums of squares and products of y -vectors in cell \mathbf{i} , with components

$$P_{\gamma\zeta}(\mathbf{i}) = \sum_{k: \mathbf{i}_k = \mathbf{i}} y_{\gamma}^k y_{\zeta}^k.$$

In order to account for the identifiability constraints, the likelihood can be rewritten as follows

$$\log L = \mathbf{N}' \mathbf{D} \boldsymbol{\Lambda} + \sum_{\gamma \in \Gamma} (\mathbf{S}_{\gamma})' \mathbf{D} \mathbf{H}_{\gamma} - \frac{1}{2} \sum_{\gamma, \zeta \in \Gamma} (\mathbf{P}_{\gamma\zeta})' \mathbf{D} \boldsymbol{\psi}_{\gamma\zeta} \quad (4.26a)$$

$$= \mathbf{N}' \mathbf{D} \boldsymbol{\Lambda} + \text{tr}(\mathbf{D} \mathbf{H} \mathbf{S}') - \frac{1}{2} \sum_{\mathbf{i}, \mathbf{j} \in \mathcal{I}} d_{\mathbf{i}\mathbf{j}} \text{tr} (\boldsymbol{\Psi}(\mathbf{j}) \mathbf{P}(\mathbf{i})) \quad (4.26b)$$

$$= \mathbf{N}' \mathbf{D} \boldsymbol{\Lambda} + \text{tr}(\mathbf{D} \mathbf{H} \mathbf{S}') - \frac{1}{2} \sum_{\mathbf{i} \in \mathcal{I}} \text{tr} (\boldsymbol{\Omega}(\mathbf{i}) \mathbf{P}(\mathbf{i})) \quad (4.26c)$$

where \mathbf{N} , \mathbf{S}_{γ} and $\mathbf{P}_{\gamma\zeta}$ are vectors of cell counts, cell sums and cell sums-of-products, respectively; $\boldsymbol{\Lambda}$ is vector consisting of λ^{\emptyset} and the discrete interactions;

\mathbf{H}_γ is a vector of mixed linear interactions involving $\gamma \in \Gamma$; and $\psi_{\gamma\zeta}$ is a vector of mixed quadratic interactions involving $\gamma, \zeta \in \Gamma$.

When the data are such that the observed means are greater than the observed standard deviations, poor mixing can result when implementing MCMC for these models, as described later, due to posterior correlation between parameters. For this reason the continuous data are relocated within cells. More precisely, for each level \mathbf{i} and for each k such that $\mathbf{i}^k = \mathbf{i}$,

$$\mathbf{y}^k \rightarrow \mathbf{z}^k = \mathbf{y}^k - \mathbf{a}(\mathbf{i}).$$

Covariance matrices, and hence precision matrices, are invariant to centring, that is $\Sigma_z(\mathbf{i}) = \Sigma_y(\mathbf{i})$, so only the discrete and linear parameters will be changed: $\mu_z(\mathbf{i}) = \mu_y(\mathbf{i}) - \mathbf{a}(\mathbf{i})$ so

$$\mathbf{h}_z(\mathbf{i}) = \Omega(\mathbf{i})\mu_z(\mathbf{i}) = \mathbf{h}_y(\mathbf{i}) - \Omega(\mathbf{i})\mathbf{a}(\mathbf{i}).$$

Hence,

$$\mathbf{H}_z = \mathbf{D}^{-1}\mathbf{h}_z = \mathbf{H}_y - \mathbf{D}^{-1}\mathcal{U}_1, \quad (4.27)$$

where \mathcal{U}_1 is the matrix with j th row $\Omega(j)\mathbf{a}(j)$.

Thus, for the centred data, the equivalent of setting an entry of \mathbf{H}_y to zero is to set the corresponding entry of \mathbf{H}_z to the negative of the corresponding entry of $\mathbf{D}^{-1}\mathcal{U}_1$.

Also,

$$g_z(\mathbf{i}) = g_y(\mathbf{i}) + \mu(\mathbf{i})'\Omega(\mathbf{i})\mathbf{a}(\mathbf{i}) - \frac{1}{2}\mathbf{a}(\mathbf{i})'\Omega(\mathbf{i})\mathbf{a}(\mathbf{i}).$$

Hence,

$$\Lambda_z = \mathbf{D}^{-1}\mathbf{G}_z = \Lambda_y + \mathbf{D}^{-1}\mathcal{U}_2, \quad (4.28)$$

where \mathcal{U}_2 is the vector with j th entry $(\mu_y(j) - \frac{1}{2}\mathbf{a}(j))'\Omega(j)\mathbf{a}(j)$.

Thus, the equivalent of setting a particular λ_y to zero is to set the corresponding λ_z to the corresponding entry in $\mathbf{D}^{-1}\mathcal{U}_2$.

In practice, we set $\mathbf{a} = \bar{\mathbf{y}}$, in which case we obtain a simplification of the likelihood, the logarithm of which becomes

$$\log f(\mathbf{i}, \mathbf{z} | \Lambda, \mathbf{H}, \Omega) = \mathbf{N}'\mathbf{D}\Lambda_z - \frac{1}{2} \sum_{\mathbf{i} \in \mathcal{I}} \text{tr}(\Omega(\mathbf{i})\mathbf{P}_z(\mathbf{i})) \quad (4.29)$$

since the centred matrix of sums is identically zero. Notice that the η_z 's now enter only through the log normalizing constant, λ^θ . Note also that $\mathbf{P}_z(\mathbf{i}) = (n(\mathbf{i}) - 1)\hat{\Sigma}(\mathbf{i})$, where $\hat{\Sigma}(\mathbf{i})$ is the observed variance in cell \mathbf{i} .

Although setting \mathbf{a} to be a data dependent value may seem odd particularly as a prior for this is then required, for the kinds of diffuse priors used in this thesis the effect is slight although the computational advantages are great.

4.6 Prior Distributions

Rather than specify priors for the original y -interactions and derive the corresponding priors for the z -interactions, it is more practical to simply specify priors for the latter directly. To ease the notation, from this point forward, it is assumed that the interactions are the z -interactions and the z subscripts will be dropped, unless required for clarity. Of course, they may easily be transformed back to the original scale, if desired.

$\mathbf{\Lambda}$ is a vector of real-valued parameters, which, apart from λ^\emptyset , may vary freely. Any prior suitable for the interaction parameters in a log linear model is also suitable as a prior for $\mathbf{\Lambda}$ here. A simple choice is a Normal with mean zero and diagonal covariance matrix so that the priors are independent.

\mathbf{H} is a matrix of real-valued parameters, which may vary freely so independent zero-mean normal priors may be used here also. Indeed, these are more appropriate for the η_z 's as the centring moves them towards zero.

The quadratic interaction matrices, $\mathbf{\Psi}$, are real-valued but are required to be such that each $\mathbf{\Omega}(\mathbf{i})$ is positive definite (and symmetric).

This greatly restricts the parameter space, making prior specification, as well as proposal generation in MCMC algorithms, difficult. To illustrate, consider this simple case with two binary variables, A and B , using sum-to-zero identifiability constraints:

$$\begin{aligned}\mathbf{\Omega}(0,0) &= \mathbf{\Psi} + \mathbf{\Psi}^A(0) + \mathbf{\Psi}^B(0) + \mathbf{\Psi}^{AB}(0,0) \\ \mathbf{\Omega}(0,1) &= \mathbf{\Psi} + \mathbf{\Psi}^A(0) - \mathbf{\Psi}^B(0) - \mathbf{\Psi}^{AB}(0,0) \\ \mathbf{\Omega}(1,0) &= \mathbf{\Psi} - \mathbf{\Psi}^A(0) + \mathbf{\Psi}^B(0) - \mathbf{\Psi}^{AB}(0,0) \\ \mathbf{\Omega}(1,1) &= \mathbf{\Psi} - \mathbf{\Psi}^A(0) - \mathbf{\Psi}^B(0) + \mathbf{\Psi}^{AB}(0,0)\end{aligned}$$

We need $(\mathbf{\Psi}, \mathbf{\Psi}^A(0), \mathbf{\Psi}^B(0), \mathbf{\Psi}^{AB}(0,0))$ such that each of the above linear combinations is positive definite and this is quite a tall order.

It is for this reason that the alternative parameterization $(\mathbf{\Lambda}, \mathbf{H}, \mathbf{\Omega})$ is proposed. The constraints for $\mathbf{\Omega}$ (4.19,4.20 and, for two discrete variables, 4.21) are

not as straightforward but they are quite manageable. The idea is to use the constraints to identify a subset of parameters with the remainder being determined by constraint.

Any priors suitable for Ω in a graphical Gaussian model may be used here but must take account of the graphical constraints. This is quite straightforward if the precisions are decomposed as

$$\Omega(\mathbf{i}) = \text{diag}(\tau(\mathbf{i}))\mathbf{C}(\mathbf{i})\text{diag}(\tau(\mathbf{i})),$$

just as in the graphical Gaussian case, and the nonconjugate priors described in Chapter 2 are used. Note that in an extension of the GGM decomposition, $\omega_{jj}(\mathbf{i}) \equiv \tau_j^2(\mathbf{i})$, the j^{th} conditional partial precision and for $j \neq k$, $\omega_{jk}(\mathbf{i}) \equiv -\rho_{jk}(\mathbf{i})\tau_j(\mathbf{i})\tau_k(\mathbf{i})$ and $\mathbf{C}(\mathbf{i})$ is the matrix with jk^{th} entry $-\rho_{jk}(\mathbf{i})$, the negative of the jk^{th} conditional partial correlation. For notational convenience, ω_{jj} will be used in preference to τ_j^2 and it may be more convenient to express constraints and assignments in terms of ω_{jk} rather than ρ_{jk} .

The graphical constraints simply reduce the number of parameters requiring priors. For example, if $\Delta = \{A\}$, $\Gamma = \{1, 2\}$ and $A \perp\!\!\!\perp 1|2$, then $\tau_1(0)$ and each $\tau_2(i)$ will have independent gamma priors; $\rho_{12}(0)$ will have a $U(-1, 1)$ prior; and the remaining $\tau_1(i)$'s and $\rho_{12}(i)$'s are determined by the graphical constraints, namely

$$\tau_1(i) = \tau_1(0) \quad \forall i > 0$$

and

$$\rho_{12}(i) = \frac{\tau_2(0)}{\tau_2(i)}\rho_{12}(0) \quad \forall i > 0.$$

Prior variances are chosen so that the prior is diffuse but not so diffuse that more complex models are unnecessarily penalised. There is no simple rule for determining a suitable prior variance but analysis of simulated data from a known model can be a useful guide. See Section 5.5.1 for more on this.

As with GGM's, prior normalizing constants can be found by simulation. However, it is not quite as straightforward since since some of the partial correlations will be functions of both other partial correlations and partial precisions as in the above example. In this case, we have

$$\begin{aligned} \Omega(0) &= \begin{pmatrix} \tau_1^2 & \rho_{12}(0)\tau_1\tau_2(0) \\ & \tau_2^2(0) \end{pmatrix} \\ \Omega(1) &= \begin{pmatrix} \tau_1^2 & \rho_{12}(0)\tau_1\tau_2(0) \\ & \tau_2^2(1) \end{pmatrix} \end{aligned}$$

So that

$$\begin{aligned}\mathbf{C}^{(0)} &= \begin{pmatrix} 1 & \rho_{12}(0) \\ & 1 \end{pmatrix} \\ \mathbf{C}^{(1)} &= \begin{pmatrix} 1 & \rho_{12}(0) \frac{\tau_2(0)}{\tau_2(1)} \\ & 1 \end{pmatrix}.\end{aligned}$$

This extremely simple example shows that it is necessary to generate all of Ω , not just the partial correlation matrices, in order to determine the prior normalizing constant. More explicitly, we need to generate $|\mathcal{I}|$ $q \times q$ matrices with unconstrained diagonal entries drawn from a Gamma distribution, unconstrained off-diagonal entries the product of a draw from a Uniform(-1,1) distribution with square root of the product of the two corresponding diagonal entries and constrained entries determined by the model constraints. We then test whether *all* the matrices are positive-definite.

If this is done a large number of times, the proportion of times that all of the matrices are positive definite (the acceptance rate) will be approximately the reciprocal of the normalizing constant. Naturally, the more cells there are, and hence the more matrices, the smaller these proportions, so a practical upper limit is reached much sooner than for GGM's. Table 4.1 gives the smallest proportions for any of the models, based on 100000 sets of matrices for various values of q and combinations of numbers of levels for models with $p = 1$ and $p = 2$. Note that these are not necessarily the saturated models; In the above example, the acceptance rate under the saturated model is 1 but three models have smaller acceptance rates.

The practical upper limit for models with one discrete variable appears to be $q = 4$ with a moderate number of levels or possibly, for graphical models, $q = 5$ with no more than 2 levels. For models with two discrete variables and a moderate number of levels for each, up to $q = 4$ seems feasible. Again, this sort of practical limitation will arise for any nonconjugate priors but, as with GGM's, it does not prevent their use for fixed model inference.

It is possible to solve the problem of prior specification for the quadratic interactions by first specifying a prior on the precision matrices as described and then obtaining the corresponding prior for the interactions using appropriate transformations. Note that the Jacobians for such transformations are always constants and so are not required for fixed-model inference. However, when it comes to proposal generation in an MCMC algorithm, it is more difficult to satisfy the positive definiteness requirement on the precision matrices when using

q	levels	GM	HM	LRG	LRH
2	2	8	9	0.56	0.41
	3			0.43	
	4			0.35	
3	2	64	95	0.21	
	3			0.11	
4	2	1024	2489	0.02	
	3			0.01	
5	2	32768	173433	0.001	0.000
2	2,2	64	146	0.34	
	2,3			0.19	
	2,4			0.12	
	3,3			0.08	
3	2,2	1024	7460	0.04	
	2,3			0.04	

Table 4.1: Numbers and Acceptance rates of CG Models.

The second column gives the number of levels of the discrete variable(s), the third gives the number of graphical models, the fourth gives the number of hierarchical and the last two give the lowest acceptance rates for graphical models and for hierarchical models, where they differ.

interactions. For this reason, the precisions have been used in preference to the interaction matrices.

4.7 Conditional distributions

Because the likelihood (4.26) factorizes readily, conditional distributions are easily expressed as

$$\begin{aligned}
f(\Lambda|\mathbf{H}, \mathbf{\Omega}, i, \mathbf{y}) &\propto f(\Lambda) \exp(\mathbf{N}'\mathbf{D}\Lambda) & (4.30) \\
f(\mathbf{H}|\Lambda, \mathbf{\Omega}, i, \mathbf{y}) &\propto f(\mathbf{H}) \exp [n\lambda^\emptyset + \text{tr}(\mathbf{D}\mathbf{H}\mathbf{S}')] \\
f(\mathbf{H}_z|\Lambda_z, \mathbf{\Omega}, i, z) &\propto f(\mathbf{H}_z) \exp [n\lambda_z^\emptyset] \\
f(\mathbf{\Omega}|\Lambda, \mathbf{H}, i, \mathbf{y}) &\propto f(\mathbf{\Omega}) \exp [n\lambda^\emptyset - \frac{1}{2} \sum_i \text{tr}(\mathbf{\Omega}(i)\mathbf{P}(i))]
\end{aligned}$$

where λ_\emptyset is given by equation 4.4. These distributions will be required for the implementation of a Gibbs sampler as described in the next two chapters.

Chapter 5

Conditional Gaussian Models With One Discrete variable

5.1 Introduction

This chapter deals with the special case where $p = 1$, that is conditional Gaussian models with exactly one discrete variable. These are particularly simple as most of the difficulty in CG models comes from the discrete interactions, of which there are none in this case. There are, in fact, only four types of interaction: continuous-continuous (XY), mixed linear (AX), 2-way mixed quadratic (AXX) and 3-way mixed quadratic (AXY).

Throughout, A will denote the discrete variable, taking values $0, 1, \dots, l-1$.

The parameters are:

$$\Lambda = (\lambda, \lambda(0), \dots, \lambda(l-1))',$$
$$\mathbf{H} = \begin{pmatrix} \eta_1 & \dots & \eta_q \\ \eta_1(0) & \dots & \eta_q(0) \\ \vdots & & \vdots \\ \eta_1(l-1) & \dots & \eta_q(l-1) \end{pmatrix}$$

and $\mathbf{\Omega} = \{\mathbf{\Omega}(i) = \text{diag}(\tau(i))\mathbf{C}(i)\text{diag}(\tau(i)) : i = 0, \dots, l-1\}$.

Since there are no discrete interactions, the highest order of interaction is three and there are three types of model constraints:

- A missing three-way mixed interaction AXY requires $\omega_{XY}(i) = \omega_{XY}(0)$ for each $i = 1, \dots, l$. A missing two-way mixed quadratic interaction AXX requires $\omega_{XX}(i) = \omega_{XX}(0)$ for each $i = 1, \dots, l-1$.

- A missing two-way mixed linear interaction AX requires $\eta_X(i) = 0$ for each $i = 0, \dots, l-1$.
- A missing pure continuous interaction XY requires $\omega_{XY}(i) = 0$ for each $i = 0, \dots, l-1$. This necessarily means the AXY interaction will be absent also.

For hierarchical models, the quadratic AXX interaction can be missing only if each $AX\gamma$ interaction is also missing; the continuous XY interaction can be missing only if the AXY interaction is missing and the linear AX interaction can be missing only if AXX is missing.

For graphical models, a missing (X, Y) edge is equivalent to a missing XY (and AXY) interaction and a missing (A, X) is equivalent to the linear AX and each two- and three-way mixed quadratic interaction involving X , $AX\gamma$, all missing.

The design matrix is $\mathbf{D} = D(A)$ as in Equation 4.14.

5.2 MCMC for Fixed models

To generate from the posterior, a Gibbs sampler may be used. The following is an outline of the scheme used here. As the conditionals are only available up to a normalizing constant, a Metropolis-within-Gibbs scheme must be followed. Note that prior normalizing constants are not required for inference based on a single fixed model. A random-walk scheme is a convenient and effective way of generating proposals but several pilot runs may be necessary in order to find proposal variances that give optimal acceptance rates or at least allow good mixing. Acceptance ratios are particularly simple as the likelihood factorizes as shown in the previous chapter. Strictly, the conditional distributions are simpler than used here and the acceptance ratios simplify further but for computational purposes it is more convenient to evaluate one of the block conditional distributions in 4.30 and these may also be used if a block updating scheme is preferred.

Updating Λ The first element, λ , is the likelihood normalizing constant so is not updated independently. The rest are discrete main effects so are all updated. The acceptance ratio is

$$\frac{f(\Lambda^*) \exp(\mathbf{N}' \mathbf{D} \Lambda^*)}{f(\Lambda) \exp(\mathbf{N}' \mathbf{D} \Lambda)} \quad (5.1)$$

Updating \mathbf{H} The first row consists of linear main effects, which are always updated. For each $\gamma \in \Gamma$, if a linear $A\gamma$ interaction is present in the model, each $\eta_\gamma(i)$ is updated, otherwise none of them are. When the data are centred, the acceptance ratio is simply a ratio of normalizing constants times a prior ratio,

$$\frac{f(\mathbf{H}^*) \exp [n\lambda^*]}{f(\mathbf{H}) \exp [n\lambda]}. \quad (5.2)$$

Updating Ω This is done in much the same way as in GGM's but the constraints corresponding to missing mixed interactions must also be preserved. To illustrate, let X and Y be specific continuous variables and let $i > 0$.

$\tau_X(0)$ is always updated. If the AXX interaction is missing, this also changes each $\tau_X(i)$. If the AXY interaction is missing but the AXX is not, it also changes each $\rho_{XY}(i)$, because in this case,

$$\rho_{XY}(i) = \frac{\omega_{XY}(i)}{\tau_X(i)\tau_Y(i)} = \frac{\omega_{XY}(0)}{\tau_X(i)\tau_Y(i)} = \frac{\tau_X(0)\tau_Y(0)\rho_{XY}(0)}{\tau_X(i)\tau_Y(i)}.$$

If the AXX interaction is also missing, this reduces to

$$\frac{\tau_Y(0)\rho_{XY}(0)}{\tau_Y(i)},$$

which is independent of $\tau_X(0)$.

$\tau_X(i)$ is updated if and only if the AXX interaction is present. If the AXY interaction is missing, this also changes each $\rho_{XY}(i)$ as in the above expression.

$\rho_{XY}(0)$ is updated if and only if the XY interaction is present. If the AXY interaction is missing this also changes each $\rho_{XY}(i)$.

$\rho_{XY}(i)$ is updated if and only if both the XY and AXY interactions are present.

Note that for each of the above updates, if any linear interactions are missing and the data are centred, \mathbf{H} must also be amended as it depends on Ω through 4.27.

The acceptance ratio each time is

$$\frac{f(\Omega^*) \exp [n\lambda^* - \frac{1}{2} \sum_i \text{tr}(\Omega^*(i)\mathbf{P}(i))]}{f(\Omega) \exp [n\lambda - \frac{1}{2} \sum_i \text{tr}(\Omega(i)\mathbf{P}(i))]} \quad (5.3)$$

5.3 Reversible Jump Sampling

The two types of model, graphical and hierarchical, are considered separately but the structure is essentially the same for each: Firstly, a type of edge/interaction

is selected. Secondly, a specific edge/interaction of that type is selected. If it is present, it is proposed to remove it, if not it is proposed to add it. Finally, the parameters in the new model are updated as for a fixed model. In the case of hierarchical models, certain move types are not allowed and these must always be rejected when proposed. Specifically, only the maximal terms present may be removed and only minimal terms absent may be added.

Before running the reversible jump sampler, two models are run using the fixed-model sampler in order to obtain proposal variances which give satisfactory mixing. A run of the saturated model gives these proposal variances for each of the parameters in the saturated model. Constrained τ 's are regarded as parameters separate from those in the saturated model as they represent a common partial precision rather than a partial precision at a particular level of A . For this reason, the model with all mixed edges or mixed quadratic interactions missing (but no others) is run to obtain proposal variances for these parameters. Constrained ρ 's could be treated similarly but the situation with these is more complicated, as there are two continuous variables involved. It has been found in each case, however, that the random walk variances obtained for the saturated model suffice for all models. This is most likely due to the restricted range of values these parameters can take leading to relatively small posterior variances. These proposal variances are used for between-model moves as well as for within-model moves as in the fixed-model run.

Recall from Section 1.6.3 that the acceptance ratio is obtained as the product of a prior ratio, a likelihood ratio, a jump ratio, a proposal ratio and a Jacobian. For each move type, these various ratios are described, although not necessarily given explicitly, as in most cases an explicit expression of the full ratio would be unnecessarily complicated and not particularly informative.

5.3.1 Hierarchical Models

There are q possible mixed linear interactions, $\frac{1}{2}q(q-1)$ possible two-way continuous interactions and $\frac{1}{2}q(q+1)$ possible two- and three-way mixed quadratic interactions so these types of interactions are chosen with probabilities $1/(q+1)$, $(q-1)/2(q+1)$ and $\frac{1}{2}$, respectively. Next a specific interaction of the chosen type is chosen at random. If this interaction is in the current model, it is proposed to remove it, otherwise it is proposed to add it. In this way, the jump ratio is 1. Note that there may often be proposed models which must be rejected as they are not hierarchical.

To clarify the rather detailed algorithm, the following running example of a binary variable and two continuous variables will be used. This is, of course the minimal example.

Example: $l = 2$, $\Gamma = \{X, Y\}$, with parameters

$$\Lambda = \begin{pmatrix} \lambda \\ \lambda^A \end{pmatrix} \quad \mathbf{H} = \begin{pmatrix} \eta_X & \eta_Y \\ \eta_X^A & \eta_Y^A \end{pmatrix} \\ \begin{pmatrix} \tau_X(0) & \tau_Y(0) & \rho(0) \\ \tau_X(1) & \tau_Y(1) & \rho(1) \end{pmatrix}.$$

In the following, the parameters present in each model (current and proposed) will be explicitly stated, except for Λ , whose two components are always present. An asterisk (*) will denote a parameter that may or may not be present. An absent parameter will be denoted by a zero or a blank space, as appropriate.

2-way Mixed Linear Interactions

Example:

$$\begin{pmatrix} \eta_X & \eta_Y \\ 0 & * \end{pmatrix} \begin{pmatrix} \tau_X(0) & \tau_Y(0) & \rho(0) \\ & * & \end{pmatrix} \\ \longleftrightarrow \begin{pmatrix} \eta_X & \eta_Y \\ \eta_X^A & * \end{pmatrix} \begin{pmatrix} \tau_X(0) & \tau_Y(0) & \rho(0) \\ & * & \end{pmatrix}$$

Suppose firstly the linear interaction AX has been chosen for addition. The required $l-1$ new $\eta_X(i)$'s are generated independently from Normal distributions with zero mean and variance determined by a pilot run of a fixed saturated model. The proposal ratio is thus the reciprocal of the product of these normal densities evaluated at the values of the new parameters.

The posterior ratio reduces to

$$\frac{\exp [n\lambda^\theta + \text{tr}(\mathbf{DH}^* \mathbf{S}')] \prod_{i=0}^{l-1} f(\eta_X^*(i))}{\exp [n\lambda^\theta + \text{tr}(\mathbf{DHS}')]}$$

and finally, the Jacobian is 1.

If, instead, AX has been chosen for removal, each $\eta_X(i)$ is set to zero, the proposal density is the product of the normal densities described above, evaluated at the current values of the parameters which are being set to zero and the posterior ratio reduces to

$$\frac{\exp [n\lambda^{\theta*} + \text{tr}(\mathbf{DH}^* \mathbf{S}')] }{\exp [n\lambda^\theta + \text{tr}(\mathbf{DHS}')] \prod_{i=0}^{l-1} f(\eta_X(i))}$$

Note that since only hierarchical models are being considered, this move is only permitted when all quadratic interactions involving X (AXX , AXY etc) are missing.

Continuous Interactions

Example:

$$\begin{aligned} & \begin{pmatrix} \eta_X & \eta_Y \\ * & * \end{pmatrix} \quad \begin{pmatrix} \tau_X(0) & \tau_Y(0) & 0 \\ * & * & 0 \end{pmatrix} \\ \longleftrightarrow & \begin{pmatrix} \eta_X & \eta_Y \\ * & * \end{pmatrix} \quad \begin{pmatrix} \tau_X(0) & \tau_Y(0) & \rho(0) \\ * & * & 0 \end{pmatrix} \end{aligned}$$

Suppose next that the XY continuous interaction is to be added. Since it is absent from the current model, and the current model is hierarchical, AXY must also be absent. Therefore, only one new partial correlation, $\rho_{XY}^*(0)$, is required as the remainder will be determined by the model constraint(s). In order to ensure positive definiteness of each precision matrix, first obtain a_i and b_i , the endpoints of the allowed intervals for each cell i . Since for each i we require $a_i < \rho_{XY}^*(i) < b_i$ and

$$\rho_{XY}^*(i) = \rho_{XY}^*(0) \frac{\tau_X(0)\tau_Y(0)}{\tau_X(i)\tau_Y(i)},$$

the endpoints of the interval from which to draw $\rho_{XY}^*(0)$ in order to ensure positive definiteness are

$$a = \min_i \left(a_i \frac{\tau_X(i)\tau_Y(i)}{\tau_X(0)\tau_Y(0)} \right) \quad \text{and} \quad b = \max_i \left(b_i \frac{\tau_X(i)\tau_Y(i)}{\tau_X(0)\tau_Y(0)} \right).$$

Of course, if either or both of the AXX or AYY interactions are absent, these expressions simplify.

The proposal ratio is simply $(b - a)$ and the Jacobian is 1. The posterior ratio reduces to the ratio of conditional distributions, 5.3.

Suppose, instead, the XY interaction is to be removed. To keep the model hierarchical, this is only permitted when the AXY interaction is currently absent, therefore the (X, Y) term in each $\Omega(i)$ is set to zero. To determine the proposal ratio, the interval endpoints must be obtained as when adding the interaction. Otherwise, the acceptance ratio is the same.

2-way Quadratic Interactions

Example:

$$\begin{aligned} & \begin{pmatrix} \eta_X & \eta_Y \\ \eta_X^A & * \end{pmatrix} \begin{pmatrix} \tau_X(0) & \tau_Y(0) & \rho(0) \\ & * & 0 \end{pmatrix} \\ \longleftrightarrow & \begin{pmatrix} \eta_X & \eta_Y \\ \eta_X^A & * \end{pmatrix} \begin{pmatrix} \tau_X(0) & \tau_Y(0) & \rho(0) \\ \tau_X(1) & * & 0 \end{pmatrix} \end{aligned}$$

Suppose next the two-way mixed quadratic interaction AXX is to be removed. To keep the model hierarchical, this is only permitted if all mixed quadratic interactions involving X , $\{AX\gamma : \gamma \in \Gamma \setminus \{X\}\}$, are currently absent. What is required is that each $\tau_X^*(i)$ has the same value. There are only really two ways to achieve this: One is simply to average over levels. The drawbacks of this are that the Jacobian is not as simple and the reverse move is more complicated. The other way is to generate $\tau_X^*(i)$ independently of the current $\tau_X(i)$'s. A sensible choice of proposal distribution in this case is a normal distribution with mean and variance given by the estimates of posterior mean and variance based on a pilot run of the model with all two-way mixed quadratic interactions missing. (In fact, any model with AXX missing will do but removing all of them will give estimates for each continuous variable).

Before considering the details of each of these, first consider the reverse move, adding the AXX interaction. This is only permitted if the corresponding linear interaction AX is currently present and will only happen when all three-way interactions involving X are absent.

If independent proposals are to be used for removal, they must also be used for addition. The mean and variance here would be based on a pilot run of either the saturated model or the model with all three-way interaction missing but all others present. The Jacobian in this case is, naturally, 1 and the proposal ratio is

$$\frac{\phi\left(\tau_X | \hat{E}(\tau_X), \hat{V}(\tau_X)\right)}{\prod_i \phi\left(\tau_X^*(i) | \hat{E}(\tau_X(i)), \hat{V}(\tau(i))\right)}$$

where τ_X is the common value in the current model and \hat{E} and \hat{V} are respectively the mean and variance estimates. ϕ is the normal density function.

If, instead, averaging is to be used for removal, the new τ_X 's for the addition move must be such that they average to the current common value. This is achieved as follows: Generate $l-1$ independent observations, u_0, \dots, u_{l-2} from a

normal distribution with mean zero and variance given by the estimate of the posterior variance of τ_X as before.

For each $i = 0, \dots, l-1$, set $\tau_X^*(i) = \tau_X + u_i - \bar{u}$, where $\bar{u} = \frac{1}{l} \sum_{k=0}^{l-2} u_k$ and we define $u_{l-1} \equiv 0$. Now, $\sum_{i=0}^{l-1} \tau_X^*(i) = \tau_X + \bar{u} - \bar{u} = \tau_X$, satisfying reversibility.

The proposal ratio is simply the reciprocal of the product of the proposal densities, evaluated at the u 's.

Returning to the removal of this interaction, the proposal ratio for the independence scheme is the inverse of that for addition. For the averaging scheme, the proposal ratio is obtained as follows: We have

$$\begin{pmatrix} 1 & (l-1)/l & -1/l & \dots & -1/l \\ 1 & -1/l & (l-1)/l & \dots & -1/l \\ \vdots & \vdots & & \ddots & \\ 1 & -1/l & \dots & -1/l & (l-1)/l \\ 1 & -1/l & -1/l & \dots & -1/l \end{pmatrix} \begin{pmatrix} \tau_X \\ u_0 \\ \vdots \\ u_{l-2} \end{pmatrix} = \begin{pmatrix} \tau_X^*(0) \\ \tau_X^*(1) \\ \vdots \\ \tau_X^*(l-1) \end{pmatrix}$$

or, for the reverse move,

$$\begin{pmatrix} 1/l & 1/l & \dots & 1/l & 1/l \\ 1 & 0 & \dots & 0 & -1 \\ 0 & 1 & \dots & 0 & -1 \\ & & \ddots & \vdots & \\ 0 & \dots & 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} \tau_X(0) \\ \tau_X(1) \\ \vdots \\ \tau_X(l-1) \end{pmatrix} = \begin{pmatrix} \tau_X^* \\ u_0 \\ \vdots \\ u_{l-2} \end{pmatrix}$$

The Jacobians are the absolute values of the determinants of the left-hand matrices, which are both 1.

We can also obtain the u 's as

$$u_i = \tau_X(i) - \tau_X(l-1) \quad \text{for } i = 0, \dots, l-2$$

and hence obtain the proposal ratio as the inverse of that for the addition move.

The posterior ratio for both addition and removal under both schemes again reduces to 5.3.

The averaging scheme tends to work better (that is, to give better acceptance rates) and so it is the one used for any results given here. It has also been found that performance may improve when the average is scaled by a suitable value. In most of the examples examined, this scaling was not necessary and for those where it helped, a value of 0.5 was generally sufficient. It is not known at this stage why such a scaling might be necessary or what kind of values are suitable in general.

3-way Quadratic Interactions

Example:

$$\begin{aligned} & \begin{pmatrix} \eta_X & \eta_Y \\ \eta_X^A & \eta_Y^A \end{pmatrix} \begin{pmatrix} \tau_X(0) & \tau_Y(0) & \rho(0) \\ \tau_X(1) & \tau_Y(1) & 0 \end{pmatrix} \\ \longleftrightarrow & \begin{pmatrix} \eta_X & \eta_Y \\ \eta_X^A & \eta_Y^A \end{pmatrix} \begin{pmatrix} \tau_X(0) & \tau_Y(0) & \rho(0) \\ \tau_X(1) & \tau_Y(1) & \rho(1) \end{pmatrix} \end{aligned}$$

Finally, suppose the three way mixed quadratic interaction AXY is to be added. In order to keep the model hierarchical, this is only permitted when the AXX , AYY and XY interactions are all currently present. The current model has only one partial correlation parameter, with the rest determined by the model constraints, but the proposal model has one for each level. These are generated in the same way as described in Chapter 2 for GGM's, that is, for each level i , find the interval (a_i, b_i) from which to draw $\rho_{XY}^*(i)$ that ensures $\mathbf{\Omega}^*(i)$ is positive definite.

If instead, the interaction is to be removed (and this is always permitted since, if present, it is the highest order interaction), each concentration $\omega_{XY}(i)$ must be equal, so only one partial correlation, $\rho_{XY}(0)$ needs to be generated. This is done in exactly the same way as adding the XY interaction when both AXX and AYY interactions are present, as the proposal models are the same.

In both cases, the posterior ratio again reduces to the ratio of conditional distributions, 5.3.

The proposal ratio is

$$q(u) = \frac{\prod_i (b_i - a_i)}{(b - a)}$$

for the addition move and the reciprocal of this for removal. Finally the Jacobian is again 1.

5.3.2 Graphical Models

There are q possible mixed edges and $\frac{1}{2}q(q+1)$ possible continuous-continuous edges therefore a mixed edge is chosen with probability $2/(q+3)$ and a continuous with probability $(q+1)/(q+3)$. If the former, a continuous variable is chosen at random. If the latter, two distinct continuous vertices are selected at random. In each case, if the corresponding edge is present, it is proposed to remove it; if not it is proposed to add it. Since the current and proposed graphs differ in exactly one edge and each edge has an equal probability of being selected, the jump ratio is again one.

Continuous-Continuous edges

Suppose firstly that the edge (X, Y) has been chosen to be added to the graph. There are now two possibilities: (1) If both (A, X) and (A, Y) edges are currently present in the graph, we need to generate l new partial correlations, one for each level. This is done in exactly the same way as when adding the AXY interaction in a hierarchical model as the proposal models are exactly the same. The proposal ratio is

$$q(u) = \prod_i (b_i - a_i)$$

and the Jacobian is 1.

(2) If either of these mixed edges is missing, the procedure is exactly as when adding an XY interaction in a hierarchical model when either or both AXX and AYY are absent. The proposal ratio is

$$q(u) = (b - a)$$

and the Jacobian is 1.

In both cases, the posterior ratio again reduces to the ratio of conditional distributions, 5.3.

If instead, (X, Y) has been chosen for removal, simply set each $\rho_{XY}^*(i)$ to zero, regardless of which other edges are present or absent. However, to determine the proposal ratio, the interval endpoints must be obtained as when adding the edge. Otherwise, the acceptance ratio is the same.

Discrete-Continuous edges

Now suppose the edge (A, X) has been chosen for removal. This is equivalent to removing the AX and AXX interactions as well as all mixed quadratic interactions involving X . It is important, however, that these actions are performed in the following order (since quantities updated in one step are used in the next):

1. First generate a common precision, τ_X^* in the same way as when removing the AXX interaction, that is either by averaging or by using independent proposals.
2. Second, for each $\gamma \in \Gamma \setminus \{X\}$, if the (X, γ) edge is present, generate $\rho_{X\gamma}^*(0)$ in the same way as when removing the $AX\gamma$ interaction. From this any

$\rho_{X\gamma}^*(i)$ for $i > 0$ may be found from

$$\rho_{X\gamma}^*(i) = \rho_{X\gamma}^*(0) \frac{\tau_\gamma(0)}{\tau_\gamma(i)}$$

or, more conveniently for computation,

$$\omega_{X\gamma}^*(i) = -\rho_{X\gamma}^*(0) \tau_X^* \tau_\gamma(i) \quad \text{for } i = 0, 1, \dots, l-1.$$

Of course, if, for any γ , the (A, γ) edge is missing, each $\tau_\gamma(i)$ is the same and the above expressions simplify. Notice that in this case there is a common partial correlation.

3. Finally, for any missing mixed edge, (A, γ) , including (A, X) , for $i > 0$, set

$$[H^*]_{i\gamma} = \begin{cases} -[D\mathcal{U}_1^*]_{i\gamma} & \text{if using centring} \\ 0 & \text{if not} \end{cases}$$

If centring is not being used, this step may be performed first, if desired.

If, instead, the edge (A, X) is to be added, the following steps are necessary:

1. Generate l partial precisions, $\tau_X^*(i)$, one for each level i . This is done in the same way as when adding the AXX interaction, that is either using independent proposals or random increments, depending on which scheme is used for the removal step.
2. For each $\gamma \in \Gamma \setminus \{X\}$ where the (X, γ) edge is present, if the (A, γ) edge is present, generate l partial correlations, $\rho_{X\gamma}^*(i)$, one for each level i . This is done in the same way as when adding an $AX\gamma$ interaction.
3. Generate $l - 1$ mixed linear interaction parameters, $\eta_X(i) \equiv [H^*]_{iX}$. This is done in the same way as when adding the AX interaction.

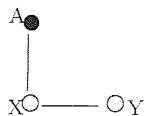
The full acceptance ratio for the addition step is

$$l \frac{f(\mathbf{\Omega}^*) \exp [n\lambda^{\theta^*} + \text{tr}(\mathbf{D}\mathbf{H}^*\mathbf{S}') - \frac{1}{2} \sum_i \text{tr}(\mathbf{\Omega}^*(i)\mathbf{P}(i))]}{f(\mathbf{\Omega}) \exp [n\lambda^\theta + \text{tr}(\mathbf{D}\mathbf{H}\mathbf{S}') - \frac{1}{2} \sum_i \text{tr}(\mathbf{\Omega}(i)\mathbf{P}(i))] \prod_{i=0}^{l-1} f(\eta_X^*(i))} \times \\ \frac{1}{\prod_i \phi(u_i | \hat{V}(\tau_X(i)))} \frac{1}{\prod_{i=1}^{l-1} \phi(\eta_X^*(i) | \hat{V}(\eta_X(i)))} \frac{\prod_i (b_i - a_i)}{(b - a)}$$

The acceptance ratio for the removal step is the reciprocal of this with current and proposal values reversed.

5.4 Model Indexing

Graphical models are indexed as described in Section 1.3.2 with the discrete variable A always being first so for example, when $q = 2$, graph 5 is



Hierarchical models are indexed as follows: Let χ be an indicator vector of length $q(q + 1)$, the total number of possible interactions. The first q of these refer to mixed linear interactions, AY_1, AY_2, \dots, AY_q ; the next $\frac{1}{2}q(q + 1)$ to mixed quadratic interactions, ordered as

$AY_1Y_1, AY_1Y_2, \dots, AY_1Y_q, AY_2Y_2, AY_2Y_3, \dots, AY_2Y_q, \dots, AY_qY_q$; and the last

$\frac{1}{2}q(q - 1)$ to pure continuous interactions, ordered as with GGM's, $Y_1Y_2, \dots, Y_1Y_q, Y_2Y_3, \dots, Y_2Y_q, \dots, Y_{(q-1)q}$. Each component of χ is 1 if the corresponding interaction is present and 0 otherwise. Treating χ as a binary number, it is converted into its decimal form, which is used as the index, much as for graphical models. For example, when $q = 2$, model 50 includes the AX , AY and AYY interactions only and written in Edwards' notation is $A/AX, AY/AY, X$. The graphical model shown above is written as 101001 or 41. Obviously, not all indices correspond to valid hierarchical models.

5.5 Examples

Some examples of using the reversible jump MCMC samplers just described are now presented. One is from Edwards (1990,1995,2000), one from Whittaker (1990) and the last, which is an example of the most complex case that can be dealt with using the priors as described (due to the limitations of the prior described in Section 4.6), has not been treated in the context of graphical models before. Firstly, simulated data examples are presented in order to assess the performance of the sampler for graphical models and its sensitivity to the prior variance of the partial precisions. For brevity, a similar assessment of the sampler for hierarchical models is not presented but its performance is comparable to that for graphical models.

The priors used in all examples are those described in Section 4.6. Posterior model probabilities are based on runs of 100 000 iterations after

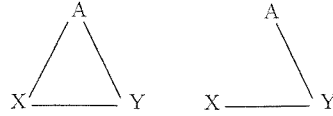


Figure 5.1: Graphs of the models from which data were generated

10 000 of burn-in and MCMC standard errors are based on batches of size 1000. The chains were initialised at the saturated model with the MLEs under this model.

5.5.1 Simulated data examples

Data were generated from graphical models with graphs shown in Figure 5.1, each with a single discrete variable, A and two continuous variables, X and Y . This is the smallest type of model that the sampler deals with. One set of data was generated from the saturated model (graph 7) and two from the model with the discrete-continuous edge AX missing (graph 3). In the first two, A has two levels and in the third it has three.

For the saturated model, the discrete data were first generated from a binomial distribution with $P(A = 0)$ generated from a Uniform(0, 1) distribution. The continuous data were then generated independently from three bivariate Normal distributions with means and variances generated respectively from a multivariate Normal distribution with mean zero and an inverse Wishart distribution with 3 degrees of freedom and unit matrix as parameter. The parameters of the CG distribution from which the second dataset were generated were obtained from those for the first dataset by setting appropriate interaction parameters (those for the AX , AXX and AXY interactions) to zero. The third dataset was generated similarly, using a trinomial and three bivariate Normal distributions. The total number of observations in each case was 50.

The posterior model probabilities are presented in Table 5.1. Section 3.4 investigated sensitivity of the posterior model probabilities to the prior for the partial correlations so this will not be pursued here. However, unlike with GGMs, the partial precision parameters in CGMs are not always present so there is the possibility of sensitivity of posterior model probabilities to their prior variance. To assess this, three different values for the prior variance of the partial precision parameters were used. In Table 5.1, p_1 results from setting $\beta = 0.1$, p_2 from $\beta = 0.001$ and p_3 from $\beta = 0.00001$, corresponding to prior variances of 10,

1000 and 100 000. As expected, in each case the probability for the saturated model relative to that for the model with the (A, X) edge missing decreases as the prior variance increases. However, this effect is not great and the true model has the highest posterior probability in each case, more than twice as probable as the next. As remarked in Section 4.6, if the prior variance is too great, more complex models may be unnecessarily penalised, leading to smaller posterior probability being assigned to them. There is little evidence of this effect in these examples, although it must be borne in mind that these models have relatively few parameters. The effect is more likely with more parameters, particularly with greater numbers of levels of the discrete variable. The discrete variables in the examples in the following section (and also in the next chapter) all have either two or three levels so this is not a problem that arises. All further examples the value of β used is 0.001.

Notice that for the first two datasets, only two models account for nearly all the posterior probability and the second is due to differences between true and observed values of the parameters. This is also true for the third dataset, although to a lesser extent, due to the extra level of the discrete variable. These are common phenomena, with a greater tendency towards a more diffuse posterior distribution for the model with increasing numbers of levels and posteriors for both model and parameters dominated by the data.

Trace plots of batch model probabilities for the two most probable models in each case, based on batches of size 1000 and using $\beta = 0.001$, are given in Figure 5.2 and indicate satisfactory mixing over the model space.

Good mixing across the parameter space is indicated by trace plots for the parameters, given in Figure 5.3 for the case of the first dataset and $\beta = 0.001$. They also indicate posteriors centred near the observed parameter values, as would be expected, but the true values are always well within the posterior mass, even when they differ somewhat from the observed values. Trace plots in the case of the other two values of β are indistinguishable and those for the second and third datasets are similar.

5.5.2 A drug trial using mice

This example is from Morrison (1976) and is also treated in Edwards (1990,1995,2000) and in Mardia et al. (1979). It concerns a trial using mice to determine whether use of a drug affects the level of three biochemical compounds in the brain. After randomization, the drug was administered to 12 mice and 10 served as controls.

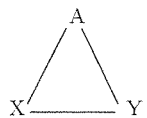
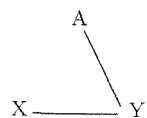
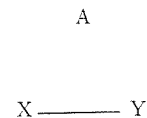
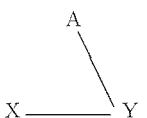

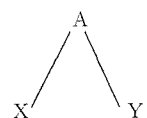
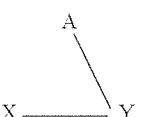
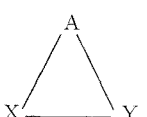
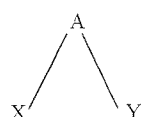
 $p_1 = 0.861$ (0.016) $p_2 = 0.756$ (0.020) $p_3 = 0.743$ (0.019)	 $p_1 = 0.138$ (0.016) $p_2 = 0.228$ (0.019) $p_3 = 0.256$ (0.019)	 $p_1 = 0.001$ (0.005) $p_2 = 0.001$ (0.004) $p_3 = 0.001$ (0.003)
 $p_1 = 0.673$ (0.033) $p_2 = 0.680$ (0.027) $p_3 = 0.710$ (0.022)	 $p_1 = 0.327$ (0.033) $p_2 = 0.317$ (0.027) $p_3 = 0.290$ (0.022)	 $p_1 = 0.000$ (0.000) $p_2 = 0.003$ (0.003) $p_3 = 0.000$ (0.000)
 $p_1 = 0.554$ (0.007) $p_2 = 0.593$ (0.008) $p_3 = 0.564$ (0.008)	 $p_1 = 0.231$ (0.006) $p_2 = 0.169$ (0.005) $p_3 = 0.146$ (0.004)	 $p_1 = 0.106$ (0.003) $p_2 = 0.102$ (0.003) $p_3 = 0.099$ (0.003)

Table 5.1: Posterior model probabilities for three simulated CG datasets. Standard errors are in brackets.

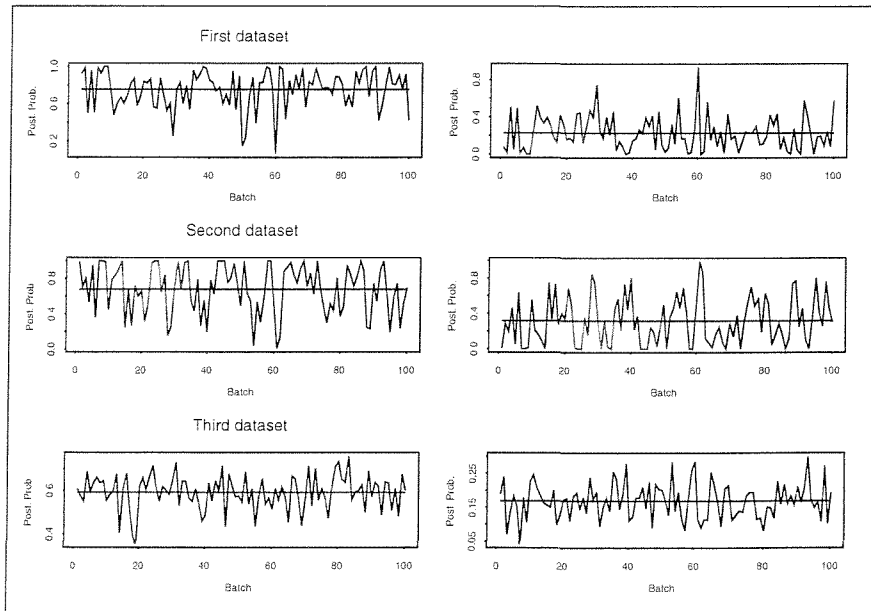


Figure 5.2: Trace plots of the highest posterior model probabilities for three simulated CG datasets. The lines show the averages, that is probabilities based on the entire sample.

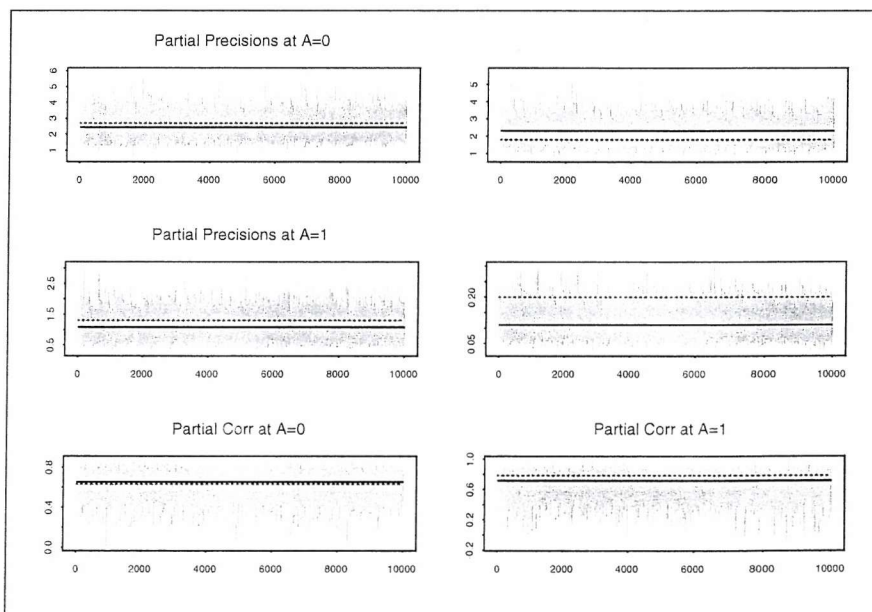


Figure 5.3: Trace plots of parameters for a simulated CG dataset. The solid lines show the true values and the dashed lines the observed. The output has been thinned to every tenth iteration.

The treatment variable is denoted by A with levels 0 and 1 corresponding to the treatment group and control groups respectively. The chemical measurements are denoted by X , Y , and Z . Edwards selects (by backward elimination) graph 52. This is also the most probable graph according to the reversible jump output and it accounts for about half of the posterior probability. The next three graphs additionally have the (X, Z) edge, the (Y, Z) edge and both, respectively and the next four repeat this pattern, but lack the (X, Y) edge. No other graphs had nonzero posterior probability. The posterior model probabilities are tabulated in Table 5.2.

Edge inclusion percentages are tabulated in Table 5.3, trace plots in Figures 5.5 and 5.6, and plots of the batch posterior probabilities for the four most probable graphs in Figure 5.4. Notice the flat trace plot for the AZ interaction, reflecting the fact that the chain never visits models with the (A, Z) edge (after burn-in). Run time was about 15 minutes.

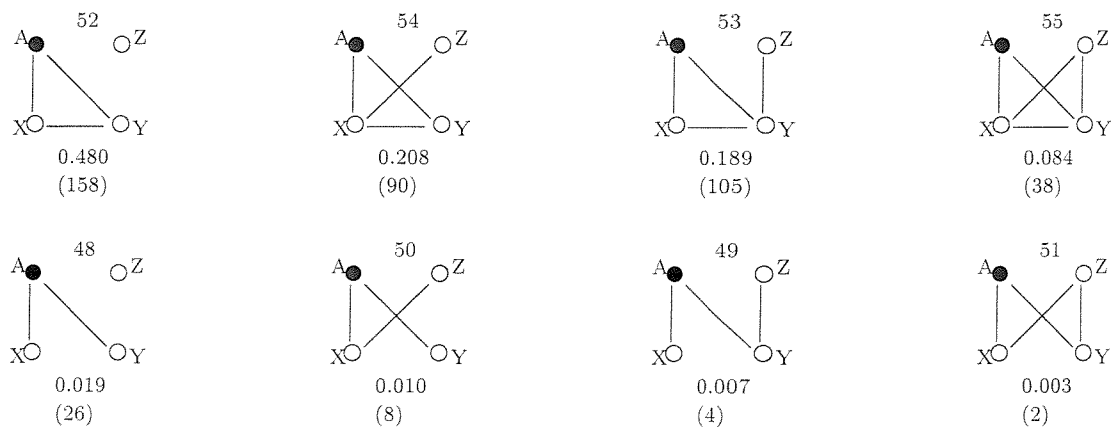


Table 5.2: Posterior model probabilities for mice data, graphical models. Standard errors $\times 10^6$ are in brackets.

AX	AY	AZ	XY	XZ	YZ
100	100	0.00	96.92	16.63	14.62

Table 5.3: Edge inclusion %ages for mice data

For the reversible jump over hierarchical models, the results are tabulated in Table 5.4. Notice that 3492 is the graphical model 52 and the rest are very similar. The inclusion percentages for each interaction are also tabulated in Table 5.5, trace plots in Figures 5.8 and 5.9 and plots of batch posterior probabilities for the most probable model in Figure 5.7.

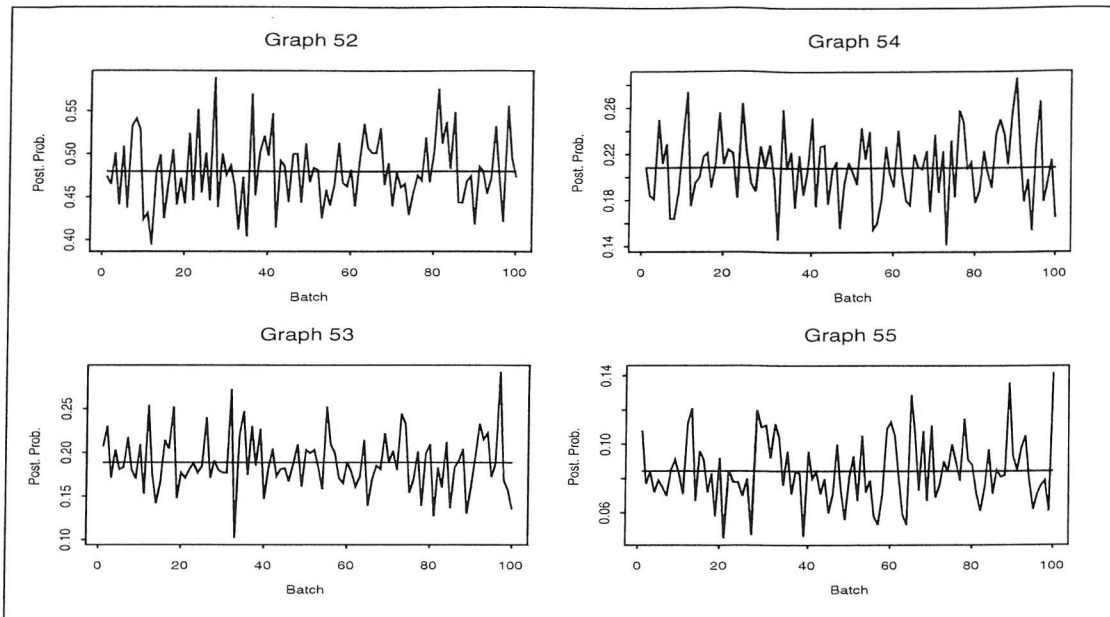


Figure 5.4: Batch Posterior Model Probabilities for Mice data, graphical models. The lines are the averages over the entire sample.

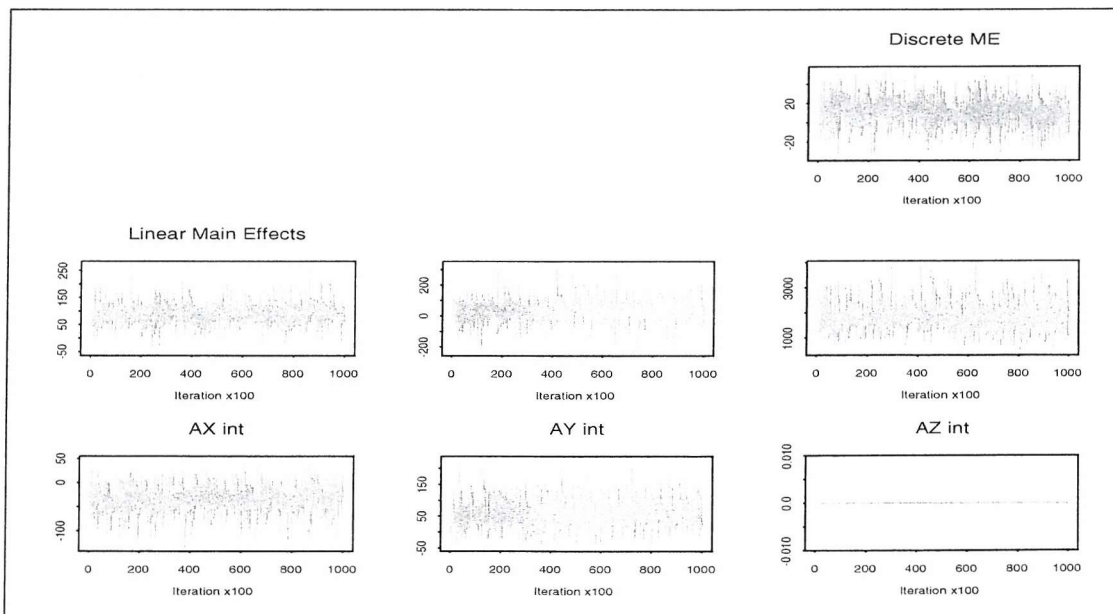


Figure 5.5: Trace plots for mice data (1), graphical models

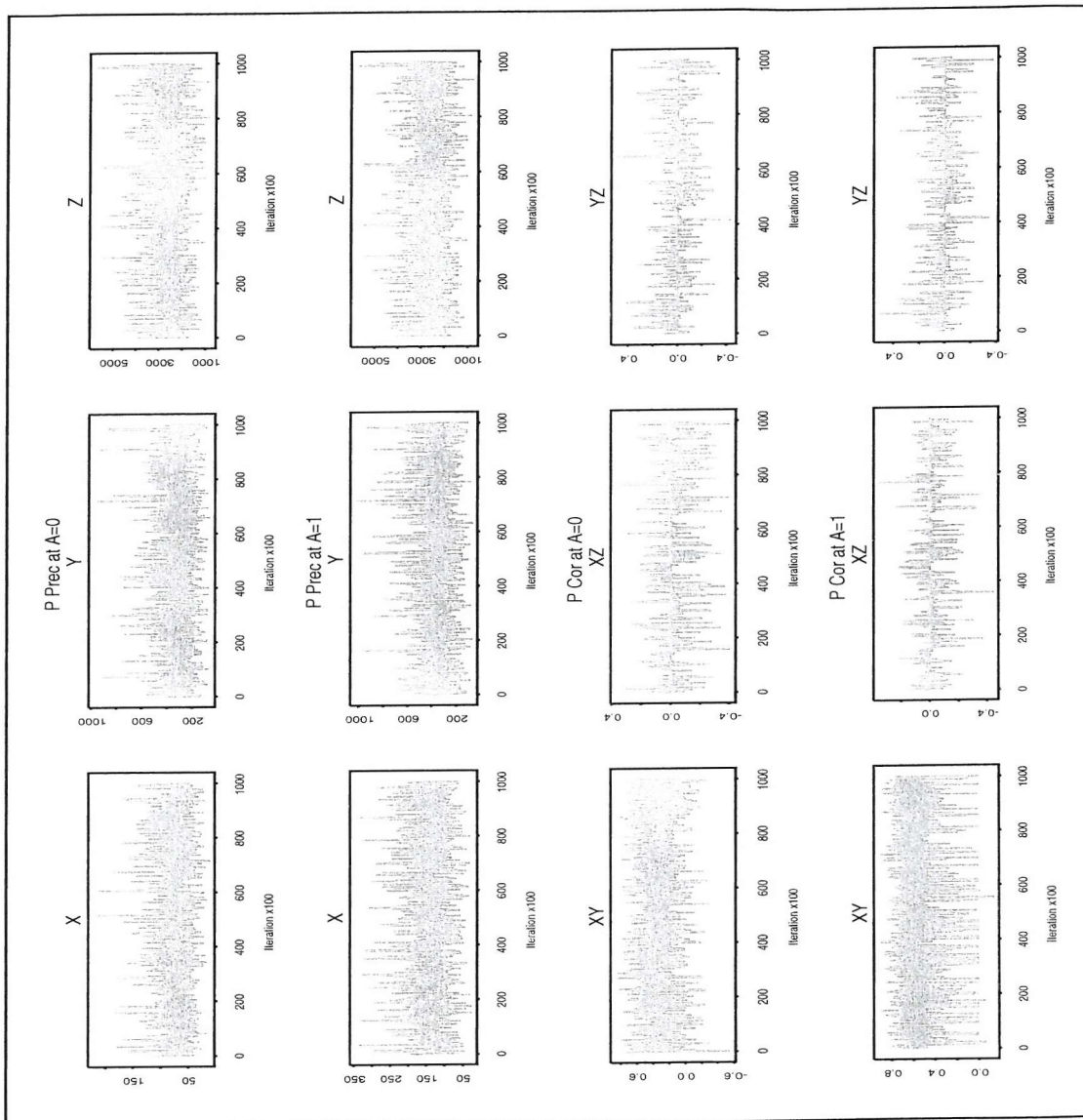


Figure 5.6: Trace plots for mice data (2), graphical models: Partial Precisions (PPrec) at each level of A and Partial Correlations (PCor) at each level of A.

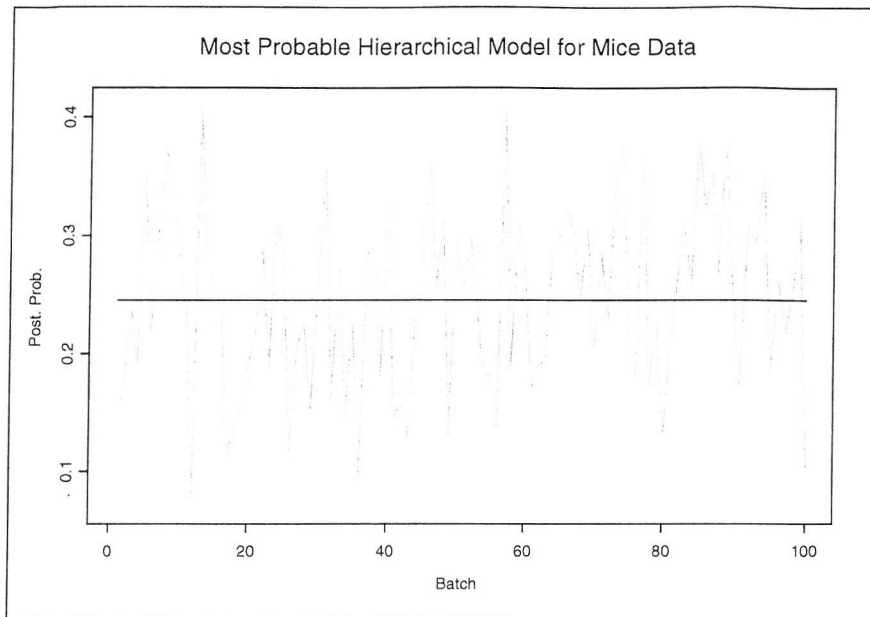


Figure 5.7: Batch Posterior Model Probabilities for Mice data, hierarchical models. The lines are the averages over the entire sample.

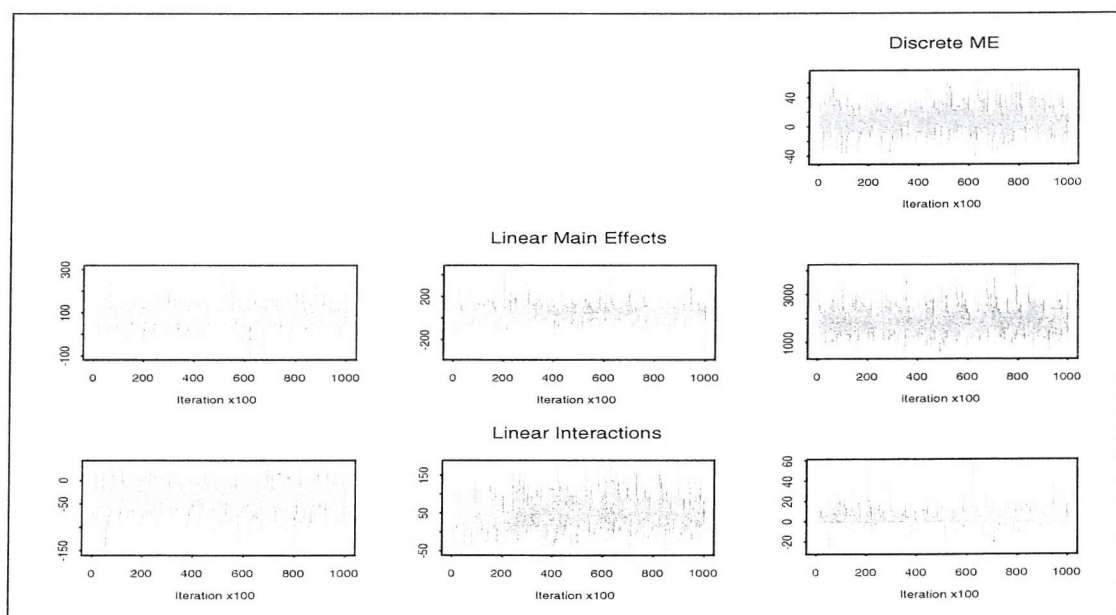


Figure 5.8: Trace plots for mice data (1), hierarchical models

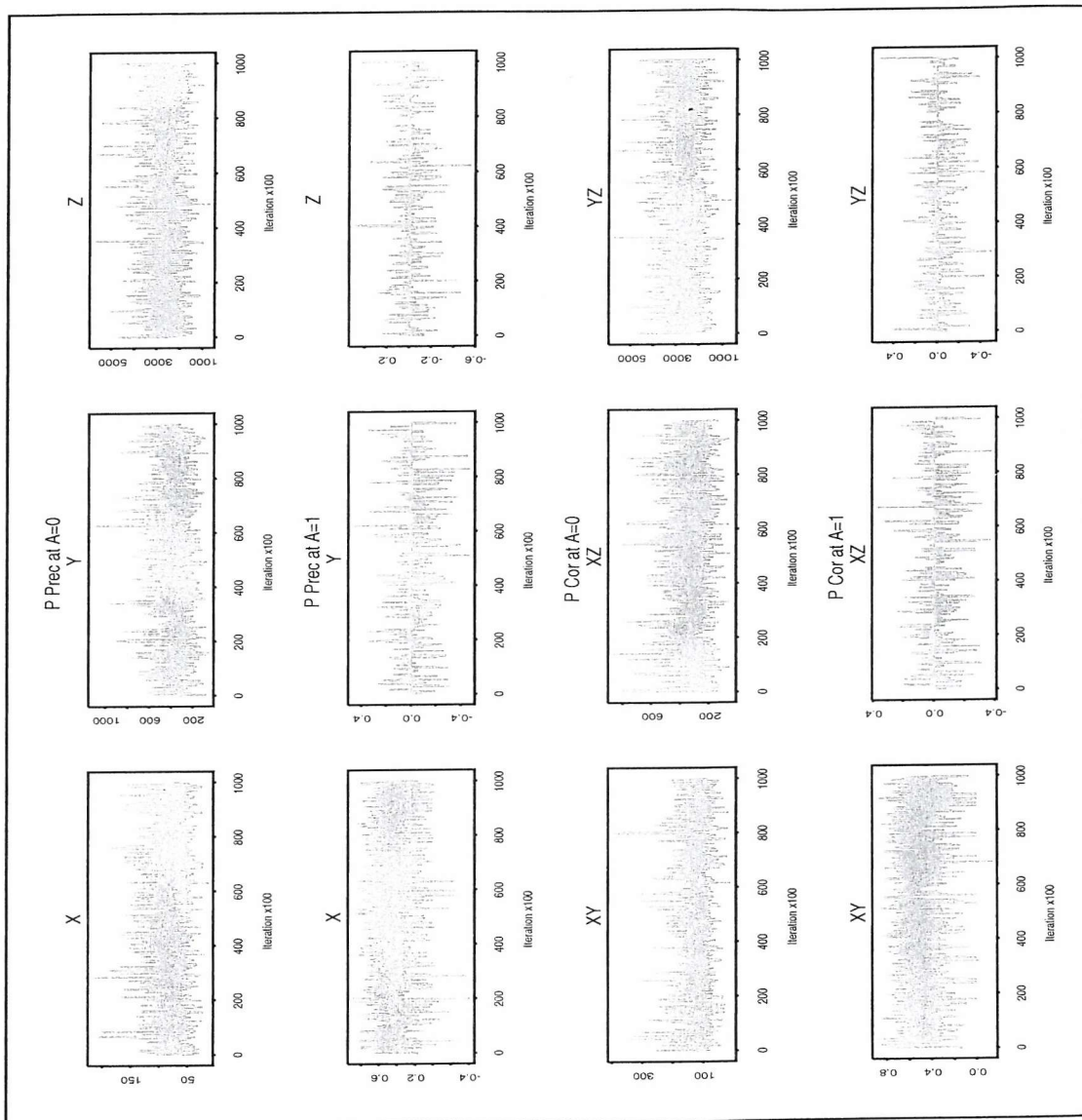


Figure 5.9: Trace plots for mice data (2), hierarchical models: Partial Precisions (PPrec) at each level of A and Partial Correlations (PCor) at each level of A.

Model	Interactions present	Probability	SE $\times 10^5$
3492	AX, AY, XY, AXX, AYY, AXY	0.223	64
3364	AX, AY, AXX, AYY, XY	0.155	55
4004	AX, AY, AZ, XY, AXX, AXY, AYY	0.126	29
3876	AX, AY, AZ, AXX, AYY, XY	0.084	18
3494	AX, AY, XY, XZ, AXX, AYY, AXY	0.061	9
3493	AX, AY, XY, YZ, AXX, AYY, AXY	0.055	11
3366	AX, AY, AXX, AYY, XY, XZ	0.047	9

Table 5.4: Posterior model probabilities for mice data, hierarchical models

AX	AY	AZ	AXX	AXY	AXZ	AYY	AYZ	AZZ
100	100	38.592	100.00	57.323	0.00	100.00	0.00	0.00
XY	XZ	YZ						
98.509	24.036	25.892						

Table 5.5: Inclusion %ages for interactions for mice data

5.5.3 Fisher's Iris data

These data were considered in Chapter 3.5, where graphical Gaussian models were used for each species separately. Considering the different results for each species, it seems unlikely a single graph will provide a good fit for the continuous data. Indeed, applying reversible jump to the continuous portion of the whole data gives almost 80% of the posterior probability to the saturated model, indicating that it is inappropriate to pool the data. If, instead a discrete variable corresponding to species is included, graph 1011, shown in Figure 5.10, receives about 95% of the posterior probability. This seems reasonable considering the differing separate results and it also makes the expected connections - petals to petals, sepals to sepals, lengths to lengths and widths to widths. This is also the graph selected by Whittaker. In fact, only three other graphs receive any posterior probability: 1019, which has an extra (1, 4) edge, 1015, which has instead an extra (2, 3) edge and 1023, the saturated model.

The results for hierarchical models are in Table 5.6. Note that the most probable model is the same graphical model as before, 1011 and the rest are predictable minor variations.

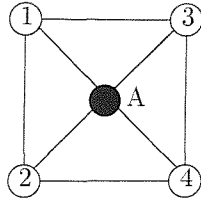


Figure 5.10: Most probable graph for iris data

Model	Interactions missing	Probability	SE $\times 10^4$
1043443	A14, A23, 14, 23	0.32286	31
1042931	A14, A23, A24, 14, 23	0.20849	16
1035251	A13, A14, A23, 14, 23	0.11618	10
1034739	A13, A14, A23, A24, 14, 23	0.08374	5
1027059	A12, A14, A23, 14, 23	0.0409	3
1026547	A12, A14, A23, A24, 14, 23	0.03788	3

Table 5.6: Most probable hierarchical models for iris data

5.5.4 Tibetan Skulls

This example comes from Morant (1923). Colonel L.A. Waddell collected 32 skulls in the south-western and eastern districts of Tibet. Type A comprises 17 skulls from graves in Sikkim and neighbouring areas. Type B comprises 15 skulls picked up on the battlefield in the Lhasa district and were believed to be those of native soldiers from the eastern province of Khams. The original source gives about 50 measurements but the five principal ones (in mm) are, (1) Greatest length of skull, (2) greatest horizontal breadth of skull, (3) height of skull, (4) upper face height and (5) face breadth (between outermost points of cheek bones).

As noted in the previous chapter, five continuous variables is probably the maximum number that can be dealt with using the priors described but only for graphical models. These priors cannot be used for hierarchical CG models with $q = 5$ so only results based on graphical models are given.

First, the reversible jump sampler for GGM's was applied for each type separately. The eight most probable graphs obtained are shown in Tables 5.7 and 5.8 but note that the posterior is very diffuse in each case, with Type B being extremely diffuse. Perhaps surprisingly, quite different graphs have higher probability for each but it must be borne in mind that the diffuseness of the distributions mean that relative differences in probability are not great. The only

common feature of the graphs shown is the presence of the (1, 4) edge in all graphs for each type. This feature is not surprising since the measurements are quite similar.

Since the posterior in each case is so diffuse, the edge inclusion percentages, tabulated in Table 5.9 along with those for an analysis using CGMs (described below), are probably more informative. The two edges mentioned have the highest inclusion rates (nearly 100%) for Type A skulls. For Type B, their inclusion rates are lower but they are still two of the three most frequently included edges, along with (4, 5). For both types, all other edges have fairly moderate inclusion rates of 30% to 60%. Those for the CGM analysis are mostly unsurprising, with (2, 5), (1, 4) and (4, 5) having very high exclusion rates. All others, except (1, 3) and (2, 4) have inclusion rates of about 25%.

The posterior probabilities for the 27 most probable graphs, which are all those with posterior probability of 1% or more, when CGM's are used for the whole dataset are given in Table 5.10, using the clique-list labelling. The first few of these graphs are shown in Figure 5.11 and trace plots of the probabilities for the first four based on batches of size 1000 are given in 5.12. Edge inclusion percentages for the continuous-continuous edges are given in Table 5.9 (all discrete-continuous edges have very low inclusion rates). Note that the discrete variable has been labelled as ' I ' since ' A ' is one of the two types. The posterior in this case is still very diffuse and the most notable feature is that there are no mixed edges in any of these. The other edges are what would be expected given the results for the separate types. Considering the lack of edges to I in any of these graphs, it is no surprise that if I is removed from the model and reversible jump for GGMs applied to the continuous data, the results are much the same with the same graphs (without I) the most probable. The posterior probabilities resulting from this analysis are also given in the same table as P_c . Most of the model probabilities for these models under the two separate group analyses do not appear in Tables 5.7 and 5.8 so are also given in Table 5.10 as P_A and P_B . Note that the model indices are the same due to the way the mixed models are indexed. The posteriors are so diffuse that these probabilities are not in fact much smaller than those given in Tables 5.7 and 5.8.

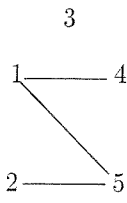
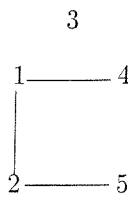
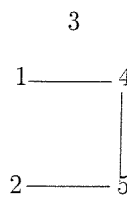
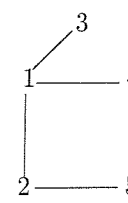
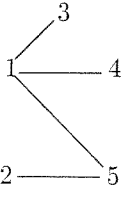
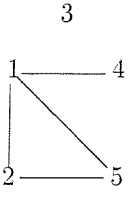
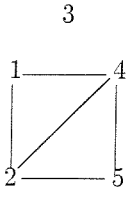
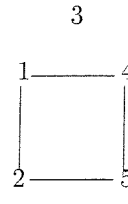
			
200 0.031 (21)	648 0.025 (17)	137 0.018 (15)	904 0.016 (11)
			
456 0.016 (12)	712 0.015 (12)	665 0.014 (11)	649 0.014 (17)

Table 5.7: Most probable graphs for type A skulls. Numbers in brackets are SE's $\times 10^4$

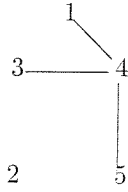
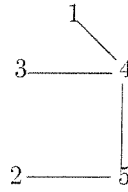
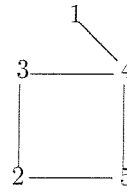
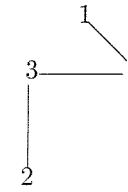
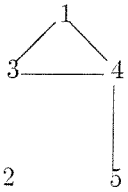
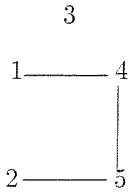
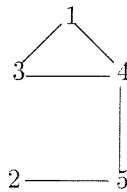
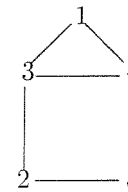
			
133 0.080 (8)	141 0.007 (6)	173 0.006 (5)	165 0.005 (7)
			
389 0.005 (5)	137 0.005 (5)	397 0.005 (6)	429 0.004 (5)

Table 5.8: Most probable graphs for type B skulls. Numbers in brackets are SE's $\times 10^4$

Edge	(1,2)	(1,3)	(1,4)	(1,5)	(2,3)	(2,4)	(2,5)	(3,4)	(3,5)	(4,5)
Type A	49.04	43.88	96.11	46.76	36.18	32.54	98.31	32.65	29.51	44.15
Type B	35.06	49.95	64.95	38.04	43.01	31.63	52.81	53.99	38.33	66.87
Both	28.21	74.62	99.82	47.28	24.35	52.76	98.87	29.67	25.24	90.74

Table 5.9: Edge inclusion %ages for Tibetan skulls data, using GGMs for each type separately and CGMs for both together.

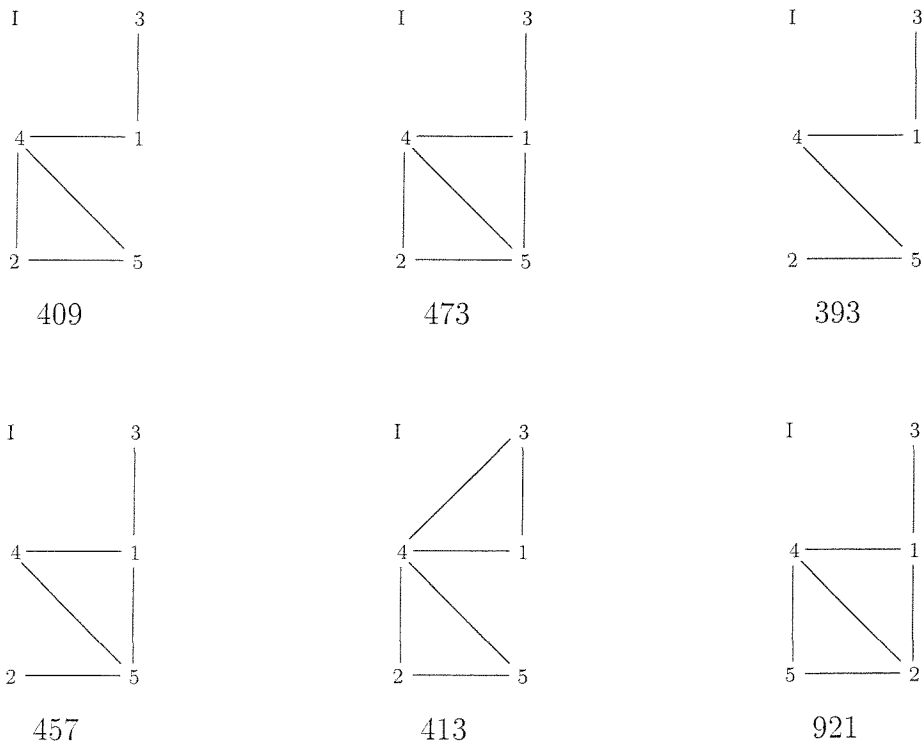


Figure 5.11: The eight most probable graphs for Tibetan skulls data

Model	Cliques	Prob.	P_A	P_B	P_c	SE $\times 10^5$
409	I/13/14/245	0.101	0.005	0.003	0.078	37
473	I/13/145/245	0.058	0.003	0.001	0.048	14
393	I/13/14/45/25	0.049	0.010	0.006	0.048	11
457	I/13/145/25	0.037	0.007	0.003	0.033	10
413	I/134/245	0.028	0.004	0.002	0.024	5
921	I/13/124/245	0.027	0.008	0.002	0.026	5
441	I/13/14/23/245	0.025	0.003	0.002	0.021	4
411	I/13/14/35/45/245	0.025	0.003	0.001	0.020	4
969	I/125/145/3	0.024	0.004	0.002	0.023	6
153	I/14/245/3	0.021	0.007	0.002	0.020	6
985	I/13/1245	0.018	0.005	0.001	0.019	2
475	I/135/145/245	0.018	0.002	0.001	0.014	2
477	I/134/145/245	0.017	0.003	0.001	0.018	3
157	I/14/34/245	0.017	0.002	0.004	0.017	3
905	I/13/14/12/25/45	0.016	0.008	0.004	0.016	2
505	I/145/13/23/245	0.016	0.002	0.001	0.016	2
456	I/13/14/15/25	0.016	0.016	0.002	0.014	9
397	I/134/25/45	0.014	0.008	0.005	0.017	2
425	I/13/14/23/25/45	0.014	0.004	0.004	0.016	2
489	I/145/13/23/25	0.013	0.003	0.003	0.012	2
461	I/134/145/25	0.013	0.005	0.002	0.011	2
395	I/13/14/35/45/25	0.012	0.003	0.003	0.014	2
137	I/14/45/25/3	0.012	0.018	0.005	0.014	2
217	I/145/245/3	0.011	0.006	0.001	0.011	2
459	I/135/145/25	0.011	0.004	0.002	0.011	2
968	I/125/13/14	0.010	0.010	0.001	0.010	2
141	I/14/25/34/45	0.010	0.006	0.007	0.013	2

Table 5.10: Posterior probabilities for Tibetan Skulls data, using CGMs, separate group analyses and combined analyses. Standard errors are for CGMs.

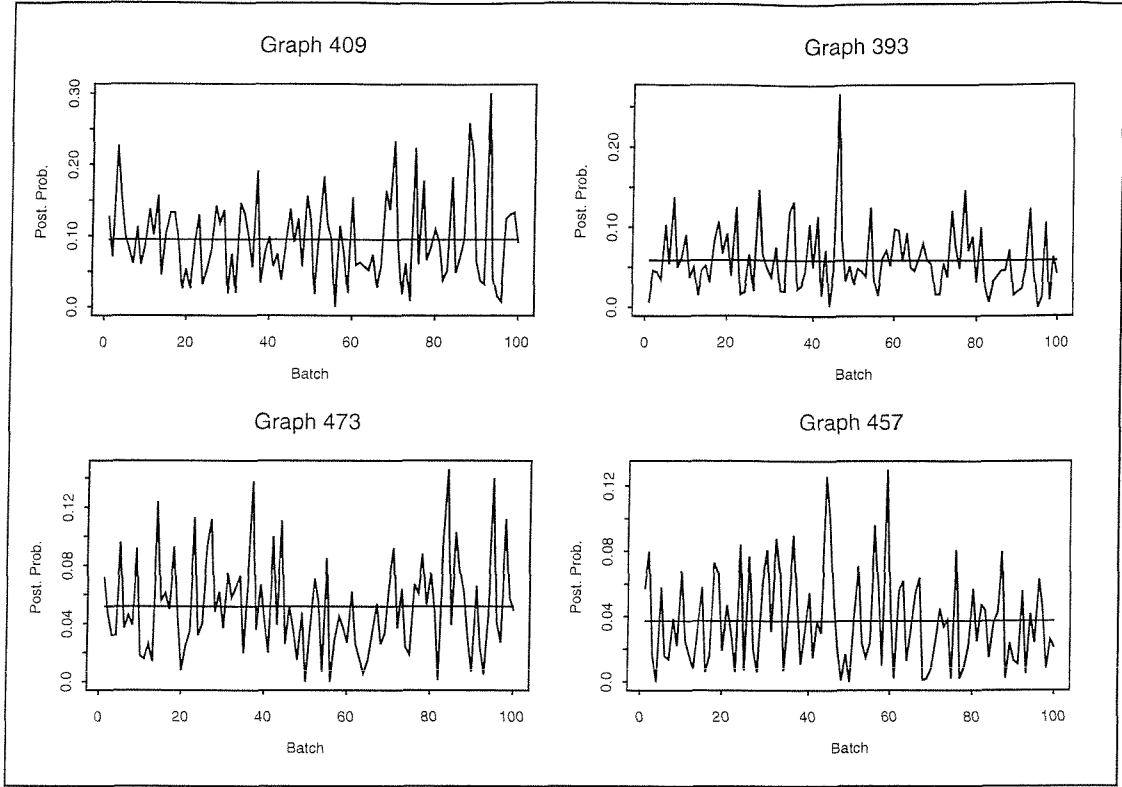


Figure 5.12: Batch Posterior model probabilities for Tibetan skulls data. The lines are the averages over the entire sample.

5.6 Model-averaged predictive distributions

Section 1.6 described how to obtain model-averaged predictive densities for some of the variables, given future values of the remaining variables, and Section 3.6 showed how this can be done for GGMs. To perform similar prediction for CGMs, we first need the following important result from Lauritzen and Wermuth (1989) giving the conditional distribution of $(\mathbf{I}, \mathbf{Y}_1)$ given $(\mathbf{J}, \mathbf{Y}_2) = (\mathbf{j}, \mathbf{y}_2)$, which is also CG:

If $\mathbf{X} = ((\mathbf{I} \cup \mathbf{J}), (\mathbf{Y}_1 \cup \mathbf{Y}_2))$ has CG distribution with moment parameters $(p(\mathbf{i}, \mathbf{j}), \boldsymbol{\mu}(\mathbf{i}, \mathbf{j}), \boldsymbol{\Sigma}(\mathbf{i}, \mathbf{j}))$, and $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{\Omega}$ are partitioned as

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}, \quad \boldsymbol{\Omega} = \begin{pmatrix} \boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{21} & \boldsymbol{\Omega}_{22} \end{pmatrix},$$

then the conditional distribution of $(\mathbf{I}, \mathbf{Y}_1)$ given $(\mathbf{J}, \mathbf{Y}_2) = (\mathbf{j}, \mathbf{y}_2)$ is CG with moment parameters given by

$$\begin{aligned}
\log p(\mathbf{i}|\mathbf{j}, \mathbf{y}_2) &= \log p(\mathbf{i}, \mathbf{j}) - \frac{1}{2} \log |\boldsymbol{\Sigma}_{22}(\mathbf{i}, \mathbf{j})| - \frac{1}{2} \boldsymbol{\mu}'_2(\mathbf{i}, \mathbf{j}) \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\mu}_2(\mathbf{i}, \mathbf{j}) \\
&\quad + \boldsymbol{\mu}'_2(\mathbf{i}, \mathbf{j}) \boldsymbol{\Sigma}_{22}^{-1}(\mathbf{i}, \mathbf{j}) + \frac{1}{2} \mathbf{y}'_2 \boldsymbol{\Sigma}_{22}^{-1}(\mathbf{i}, \mathbf{j}) \mathbf{y}_2 - \log \kappa(\mathbf{j}, \mathbf{y}_2) \\
\boldsymbol{\mu}(\mathbf{i}|\mathbf{j}, \mathbf{y}_2) &= \boldsymbol{\mu}_1(\mathbf{i}, \mathbf{j}) + \boldsymbol{\Omega}_{11}^{-1}(\mathbf{i}, \mathbf{j}) \boldsymbol{\Omega}_{12}(\mathbf{i}, \mathbf{j}) (\boldsymbol{\mu}(\mathbf{i}, \mathbf{j}) - \mathbf{y}_2) \\
\boldsymbol{\Sigma}(\mathbf{i}|\mathbf{j}, \mathbf{y}_2) &= \boldsymbol{\Omega}_{11}^{-1}(\mathbf{i}, \mathbf{j}).
\end{aligned}$$

The term $\kappa(\mathbf{j}, \mathbf{y}_2)$ is a normalizing constant given by

$$\begin{aligned}
\kappa(\mathbf{j}, \mathbf{y}_2) &= \sum_i \exp \left[\log p(\mathbf{i}, \mathbf{j}) - \frac{1}{2} \log |\boldsymbol{\Sigma}_{22}(\mathbf{i}, \mathbf{j})| - \frac{1}{2} \boldsymbol{\mu}'_2(\mathbf{i}, \mathbf{j}) \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\mu}_2(\mathbf{i}, \mathbf{j}) \right. \\
&\quad \left. + \boldsymbol{\mu}'_2(\mathbf{i}, \mathbf{j}) \boldsymbol{\Sigma}_{22}^{-1}(\mathbf{i}, \mathbf{j}) + \frac{1}{2} \mathbf{y}'_2 \boldsymbol{\Sigma}_{22}^{-1}(\mathbf{i}, \mathbf{j}) \mathbf{y}_2 \right]
\end{aligned}$$

In particular, if $\mathbf{J} = \emptyset$, the distribution of \mathbf{Y}_1 given $(\mathbf{I}, \mathbf{Y}_2) = (\mathbf{i}, \mathbf{y}_2)$ is Normal with mean $\boldsymbol{\mu}_1(\mathbf{i}) + \boldsymbol{\Omega}_{11}^{-1}(\mathbf{i}) \boldsymbol{\Omega}_{12}(\mathbf{i}) (\boldsymbol{\mu}(\mathbf{i}) - \mathbf{y}_2)$ and variance $\boldsymbol{\Omega}_{11}^{-1}(\mathbf{i})$. This is the same as for GGMs but conditioned on \mathbf{I} .

The predictive density $f(\mathbf{Y}_1 | \mathbf{y}_1^{n+1}, \mathbf{i}^{n+1}, \mathbf{x})$ is thus obtained in exactly the same way as $f(\mathbf{Y}_1 | \mathbf{y}_1^{n+1}, \mathbf{x})$ for GGM's in Section 3.6 but within cell i .

As an illustration, consider the mice data. Model-averaged predictive (univariate) densities, $f(X|Y = y, Z = z, A = i, \text{data})$, based on graphical models and based on hierarchical models, for a randomly selected observation from each level of A are given in Figures 5.13 and 5.14. As before, the observations used for prediction were excluded from the data used to generate the RJ sample.

These suggest that better prediction results from using hierarchical models but this is to be expected as it is a richer class of models. There is less improvement for the iris data in using hierarchical models as can be seen from from the plot of a typical model-averaged density in Figure 5.15. This is likely due to to a more concentrated posterior distribution for the model than in the case of the mice data. Naturally, since there are more observations, prediction tends to be better in this case.

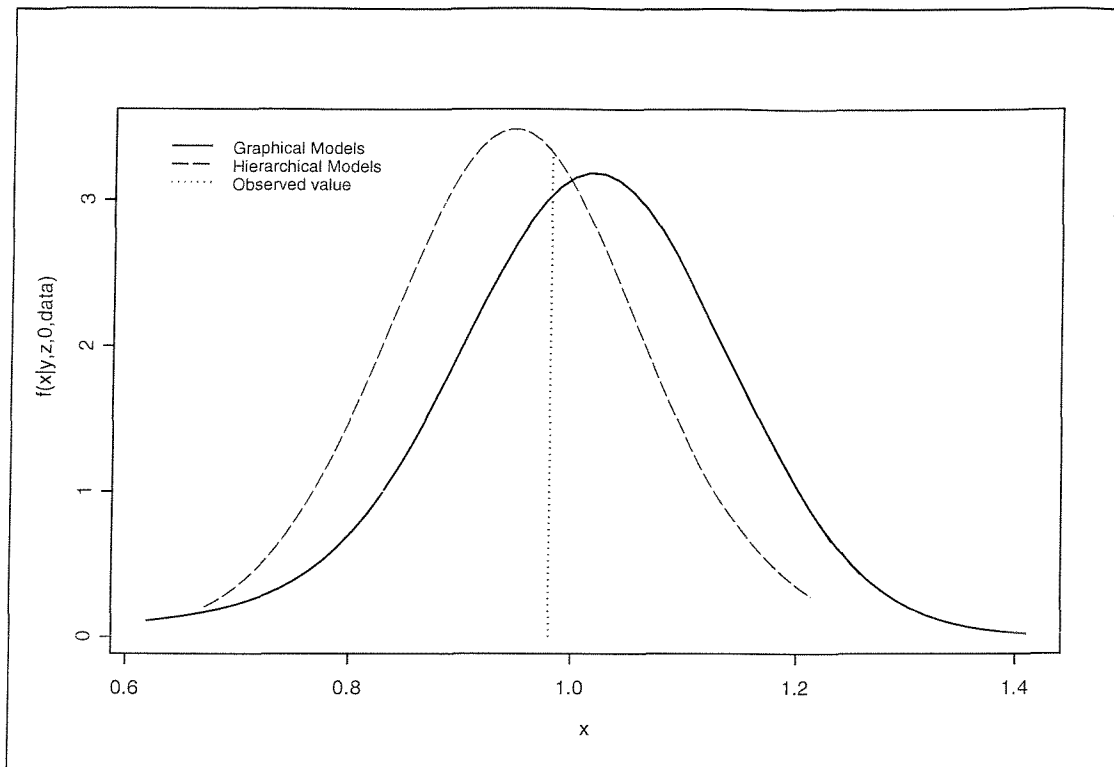


Figure 5.13: Model-averaged predictive densities for X given $Y = y, Z = z, A = 0, data$ in mice example

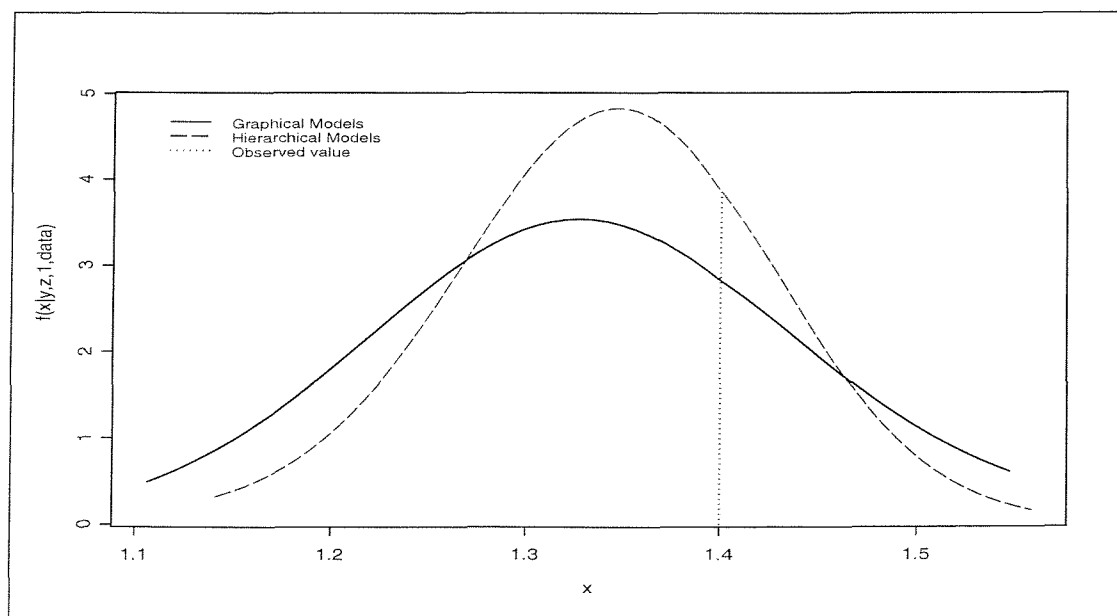


Figure 5.14: Model-averaged predictive densities for X given $Y = y, Z = z, A = 1, data$ in mice example

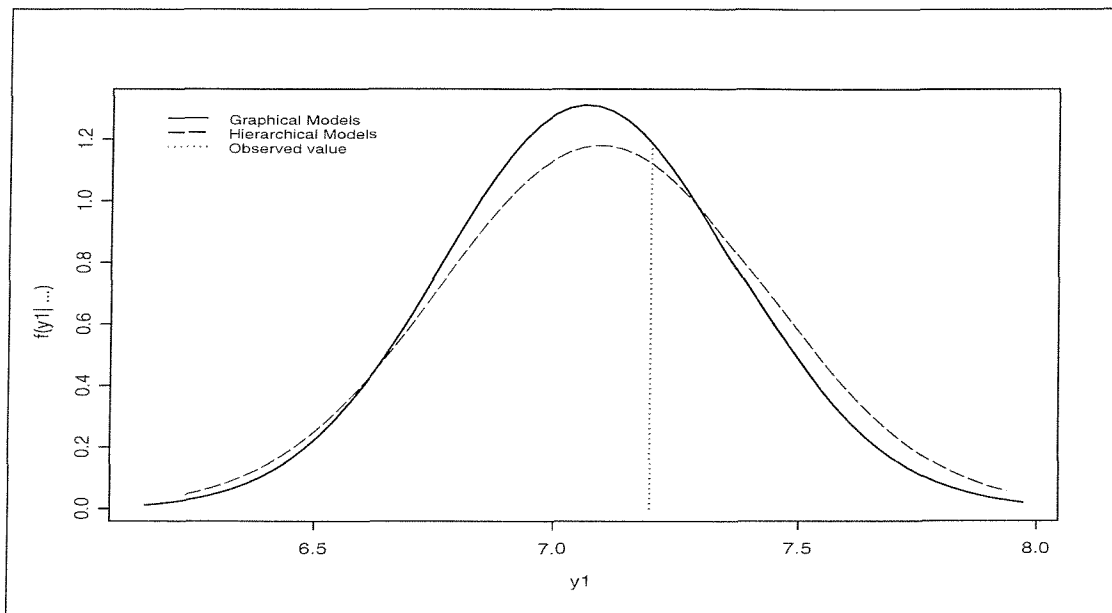


Figure 5.15: Predictive densities for Y_1 given y_2, y_3, y_4 , $A = 2$, data in iris example

We also have the conditional distribution of \mathbf{I} given $(\mathbf{J}, \mathbf{Y}) = (\mathbf{j}, \mathbf{y})$ given by

$$\begin{aligned} \log p(\mathbf{i}|\mathbf{j}, \mathbf{y}) &= \log p(\mathbf{i}, \mathbf{j}) - \frac{1}{2} \log |\Sigma(\mathbf{i}, \mathbf{j})| - \frac{1}{2} \boldsymbol{\mu}'(\mathbf{i}, \mathbf{j}) \Sigma^{-1} \boldsymbol{\mu}(\mathbf{i}, \mathbf{j}) \\ &\quad + \boldsymbol{\mu}'(\mathbf{i}, \mathbf{j}) \Sigma^{-1}(\mathbf{i}, \mathbf{j}) + \frac{1}{2} \mathbf{y}' \Sigma^{-1}(\mathbf{i}, \mathbf{j}) \mathbf{y} - \log \kappa(\mathbf{j}, \mathbf{y}) \end{aligned}$$

and in particular the distribution of all the discrete variables \mathbf{I} given the continuous $\mathbf{Y} = \mathbf{y}$ is given by

$$\begin{aligned} \log p(\mathbf{i}|\mathbf{y}) &= \log p(\mathbf{i}) - \frac{1}{2} \log |\Sigma(\mathbf{i})| - \frac{1}{2} \boldsymbol{\mu}'(\mathbf{i}) \Sigma^{-1} \boldsymbol{\mu}(\mathbf{i}) \\ &\quad + \boldsymbol{\mu}'(\mathbf{i}) \Sigma^{-1}(\mathbf{i}) + \frac{1}{2} \mathbf{y}' \Sigma^{-1}(\mathbf{i}) \mathbf{y} - \log \kappa(\mathbf{y}) \end{aligned}$$

The probability $p(\mathbf{i}|\mathbf{y})$ is easily seen to be

$$\frac{f(\mathbf{i}, \mathbf{y})}{\sum_{\mathbf{i}} f(\mathbf{i}, \mathbf{y})}$$

Model-averaged predictive probabilities, $p(\mathbf{I}|\mathbf{Y}, \text{data})$, for the discrete variables (or variable in the case of the models in this chapter) may then be estimated as the average of $p(\mathbf{I}|\mathbf{Y}, \boldsymbol{\Lambda}, \mathbf{H}, \boldsymbol{\Omega}, m)$ over the reversible jump output as follows:

For each iterate h of N iterations of reversible jump MCMC output, obtain cell probabilities, $p^h(\mathbf{i})$, cell means, $\boldsymbol{\mu}^h(\mathbf{i})$ and cell precisions, $\boldsymbol{\Omega}^h(\mathbf{i})$. Next, use these and a given set of continuous observations, \mathbf{y} , to obtain the conditional probabilities,

$$p^h(\mathbf{I} = \mathbf{i}|\mathbf{Y} = \mathbf{y}, \text{data}) = \frac{f^h(\mathbf{i}, \mathbf{y})}{\sum_{\mathbf{i}} f^h(\mathbf{i}, \mathbf{y})}$$

$p(\mathbf{I} = \mathbf{i}|\mathbf{Y} = \mathbf{y}, \text{data})$ is then estimated by $\frac{1}{N} \sum_{i=1}^N p^h(\mathbf{I} = \mathbf{i}|\mathbf{Y} = \mathbf{y}, \text{data})$.

In this way, we obtain a predictive classification rule for future observations, which is that an observation is classified to be in the cell with greatest predictive probability. Similarly to before, if the observations \mathbf{y} are taken from the data, this predictive classification may be checked by comparing the predicted value of \mathbf{I} with the observed value. Indeed, we can classify each observation in the data. Table 5.11 compares the predictive probabilities, $P(A^k = 0|x, y, z, \text{data})$, both based on graphical models and based on hierarchical models, with the actual treatment group a^k , for each observation (x^k, y^k, z^k) . Since there are only two groups, these probabilities are all that are required and the classification rule is that if the probability is greater than 0.5, $A = 0$ is predicted, otherwise $A = 1$ is predicted. In this case, only two observations out of 22 are misclassified, the same two for each.

The predictive classification is just as good for the iris data, with only three observations out of 150 misclassified. Two *I.versicolor* are misclassified as *I. virginica* and one *I. virginica* as *I.versicolor*.

Prediction fails for the Tibetan skulls data due to only graphs with no mixed edges receiving any posterior probability. This results in predictive probabilities that are independent of the continuous observations and based only on cell probabilities. The actual value in this case is just over 0.5, leading to all observations being classified as type A. This illustrates an important point, which is that there needs to be some association between the predictand (I in this case) and the predictors for such prediction to be effective. In a similar way, prediction is very poor for Z in the mice example due to only graphs with edges to Z receiving any posterior probability.

k	$P_G(A^k = 0)$	$P_H(A^k = 0)$	a^k	k	$P_G(A^k = 0)$	$P_H(A^k = 0)$	a^k
1	0.662	0.666	0	11	0.237	0.250	1
2	0.944	0.973	0	12	0.105	0.074	1
3	0.978	0.981	0	13	0.170	0.142	1
4	0.991	0.994	0	14	0.048	0.038	1
5	0.916	0.912	0	15	0.071	0.045	1
6	0.889	0.878	0	16	0.481	0.467	1
7	0.925	0.913	0	17	0.043	0.029	1
8	0.988	0.989	0	18	0.167	0.154	1
9	0.270*	0.233*	0	19	0.078	0.018	1
10	0.267*	0.203*	0	20	0.176	0.257	1
				21	0.044	0.031	1
				22	0.237	0.136	1

Table 5.11: Model-averaged predictive probabilities for mice data. P_G denotes a probability based on graphical models, P_H one based on hierarchical models and * denotes a misclassification.

5.7 Discussion

As with the MCMC samplers for GGMs, performance of the MCMC samplers described in this chapter is particularly good, despite the use of one-at-a-time updating. For fixed models, once the proposal variances have been tuned using pilot runs, mixing is generally very good and convergence is fast. For the reversible jump, provided appropriately tuned proposal variances are used, mixing is still good, although tends to be slower through the model space than for GGMs and hence convergence is not as rapid but not slow enough to require very long runs. Run times, although much slower than for GGMs, are still moderate on a modern computer.

The longest run time for the examples presented was still less than 30 minutes (this was for the iris data using hierarchical models). The reversible jump may not be as simple using the precision matrices as it would be using quadratic interactions but still poses no great difficulty and full interaction expansions are still possible for fixed models due to the Metropolis-Hastings algorithm not requiring prior normalizing constants.

The results for the three examples presented are comparable to those for GGMs in that they are reasonable and agree with those of other authors using deviance-based inference for the same data.

It must be noted however that all the above only applies when the data are centred as described in 4.5. Mixing tended to be very poor for most examples examined (including those shown here) when the data were not centred. A brief investigation into this phenomenon suggested that this is due to means which are greater than standard deviations resulting in high posterior correlation between linear and quadratic parameters. This is often seen in regression models, where the same solution, expressing the data as deviations from the mean, is employed.

An important point is that, as in the case of GGMs, there is no restriction to decomposable models. Indeed, flexibility is very great when using the class of hierarchical CG models, showing that a restriction to decomposable, homogeneous, MANOVA or other reduced model classes is not necessary. If desired, however, restriction to homogeneous models or other specific types of interaction model are relatively straightforward. Recall that a HCG model has no mixed quadratic interactions, or equivalently has a single precision matrix. A reversible jump sampler for HCG models is simply one for hierarchical CG models that rejects all proposed addition of mixed quadratic interactions.

As noted in Section 4.3, the hierarchical models dealt with here are ‘MIM’ models, that is hierarchical in the sense of Edwards (1990) rather than Lauritzen (1996). Lauritzen’s class of hierarchical models is larger than Edwards’ but modification of the methods described here are easily modified to deal with larger class since there is one less condition to be enforced.

Chapter 6

Conditional Gaussian Models with More Than One Discrete Variable

6.1 Introduction

CG models with more than one discrete variable are considerably more complex due to possible interaction between these discrete variables and the dependence of the distribution of the continuous on the combination of the levels of the discrete. This chapter will focus on the case of two discrete variables as a more general treatment would be excessively complicated and it is likely that the methods described here are of little practical use beyond two or three. A discussion of how to extend the methods to three or more discrete variables is in the final section.

The discrete variables will be denoted as A and B with l_A and l_B levels respectively so the levels are labelled $0, 1, \dots, l_A - 1$ and $0, 1, \dots, l_B - 1$. These will be indexed with $\mathbf{i} = (i_A, i_B)$, with \mathcal{I} denoting the set of possible values of \mathbf{i} . The number of cells, that is combinations of levels, is $l = |\mathcal{I}| = l_A l_B$.

There are now four additional types of interaction: 2-way discrete-discrete interaction (AB), 3-way mixed linear (ABY), 3-way mixed quadratic ($ABYY$) and 4-way mixed quadratic (ABY_1Y_2).

The parameters are:

$$\mathbf{\Lambda} = \begin{pmatrix} \lambda \\ \lambda_B(0) \\ \vdots \\ \lambda_B(l_B - 2) \\ \lambda_A(0) \\ \vdots \\ \lambda_A(l_A - 2) \\ \lambda_{AB}(0, 0) \\ \vdots \\ \lambda_{AB}(l_A - 2, l_B - 2) \end{pmatrix} \quad (6.1)$$

$$\mathbf{H} = \begin{pmatrix} \eta_1 & \dots & \eta_q \\ \eta_1^B(0) & \dots & \eta_q^B(0) \\ \vdots & & \vdots \\ \eta_1^B(l_B - 2) & \dots & \eta_q^B(l_B - 2) \\ \eta_1^A(0) & \dots & \eta_q^A(0) \\ \vdots & & \vdots \\ \eta_1^A(l_A - 2) & \dots & \eta_q^A(l_A - 2) \\ \eta_1^{AB}(0, 0) & \dots & \eta_q^{AB}(0, 0) \\ \vdots & & \vdots \\ \eta_1^{AB}(l_A - 2, l_B - 2) & \dots & \eta_q^{AB}(l_A - 2, l_B - 2) \end{pmatrix} \quad (6.2)$$

$$\mathbf{\Omega} = \{ \mathbf{\Omega}(i_A, i_B) = \text{diag}(\tau(i_A, i_B)) \mathbf{C}(i_A, i_B) \text{diag}(\tau(i_A, i_B)) : \quad (6.3) \\ i_A = 0, \dots, l_A \quad i_B = 0, \dots, l_B \}$$

Notice the interaction parameters are ordered with the B interactions before the A . With more discrete variables, they are in reverse order, adding one at a time, for example, $D, C, CD, B, BD, BC, BCD, A, AD, AC, ACD, AB, ABD, ABC, ABCD$. This is so that the canonical parameters are in the “right” order, according to cells, $(0, 0), (0, 1), \dots, (0, l_B - 1), (1, 0), \dots, (l_A - 1, l_B - 1)$.

Example For two binary variables and two continuous variables X and Y ,

$$\mathbf{\Lambda} = \begin{pmatrix} \lambda \\ \lambda^A \\ \lambda^B \\ \lambda^{AB} \end{pmatrix} \quad \mathbf{H} = \begin{pmatrix} \eta_X & \eta_Y \\ \eta_X^A & \eta_Y^A \\ \eta_X^B & \eta_Y^B \\ \eta_X^{AB} & \eta_Y^{AB} \end{pmatrix}$$

Where, for example, $\lambda^A \equiv \lambda^A(0)$ and $\lambda^{AB} \equiv \lambda^{AB}(0,0)$. The bracketed levels have been dropped as there are only two levels for A and B . The precision matrices may be conveniently laid out in a table as:

$$\begin{array}{c|c} \left(\begin{array}{cc} \omega_{XX}(0,0) & \omega_{XY}(0,0) \\ & \omega_{YY}(0,0) \end{array} \right) & \left(\begin{array}{cc} \omega_{XX}(0,1) & \omega_{XY}(0,1) \\ & \omega_{YY}(0,1) \end{array} \right) \\ \hline \left(\begin{array}{cc} \omega_{XX}(1,0) & \omega_{XY}(1,0) \\ & \omega_{YY}(0,1) \end{array} \right) & \left(\begin{array}{cc} \omega_{XX}(1,1) & \omega_{XY}(1,1) \\ & \omega_{YY}(1,1) \end{array} \right) \end{array}$$

Where, recall, $\omega_{\gamma\gamma}(i_A, i_B) \equiv \tau_\gamma^2(i_A, i_B)$ and $\omega_{XY}(i_A, i_B) \equiv -\tau_X(i_A, i_B)\tau_Y(i_A, i_B)\rho_{XY}(i_A, i_B)$.

The different types of model constraint are as follows:

- A missing XY interaction requires $\omega_{XY}(\mathbf{i}) = 0$ for each $\mathbf{i} \in \mathcal{I}$. In hierarchical models, this also requires the AXY , BXY and $ABXY$ interactions to be missing.
- A missing AXY interaction requires $\omega_{XY}(i_A, i_B) = \omega_{XY}(0, i_B)$ for each $i_A = 1, 2, \dots, l_A$. Consequently,

$$\rho_{XY}(i_A, i_B) = \frac{\tau_X(0, i_B)\tau_Y(0, i_B)\rho_{XY}(0, i_B)}{\tau_X(i_A, i_B)\tau_Y(i_A, i_B)}.$$

This expression simplifies if either of the AXX , AYY , BXX or BYY interactions are also missing. Similarly for a missing BXY interaction, simply reverse the roles of A and B . In hierarchical models, both also require the $ABXY$ interaction to be missing.

- A missing $ABXY$ interaction requires, as in equation 4.21,

$$\begin{aligned} \omega_{XY}(i_A, i_B) = & \frac{1}{(l_A - 1)(l_B - 1)} \left[(l_A - 1) \sum_{b \neq j} \omega_{XY}(i_A, b) \right. \\ & \left. + (l_B - 1) \sum_{a \neq i} \omega_{XY}(a, i_B) - \sum_{a \neq i, b \neq j} \omega_{XY}(a, b) \right] \quad (6.4) \end{aligned}$$

This may be interpreted as a set of constraints on $\omega_{XY}(i_A, i_B)$ for $i_A = 1, \dots, l_A - 1$ and $i_B = 1, \dots, l_B - 1$ given the rest. In the example, this means $\omega_{XY}(1, 1) = \omega_{XY}(0, 1) + \omega_{XY}(1, 0) - \omega_{XY}(0, 0)$.

- Putting $Y = X$ in these give the constraints for missing AXX and $ABXX$ interactions. In a hierarchical model, AXX can only be missing if $ABXX$ is missing and $ABXX$ is missing only if each $ABX\gamma$ is missing.
- A missing AX interaction requires each $\eta_X^A(i_A) = 0$. Similarly for a missing BX interaction. Both also require the ABX interaction to be missing. In a hierarchical model, the AX interaction can be missing only if the ABX and each $AX\gamma$ are.
- A missing ABX interaction requires each $\eta_X^{AB}(i_A, i_B) = 0$. In a hierarchical model, the ABX interaction is missing only if the $ABXX$ interaction is.
- A missing AB interaction requires each $\lambda^{AB}(i_A, i_B) = 0$. In a hierarchical model, the AB interaction can be missing only if each $AB\gamma$ and $AB\gamma\zeta$ are.

In a graphical model, a missing (X, Y) edge corresponds to a missing XY interaction (and hence also AXY , BXY and $ABXY$); A missing (A, X) edge corresponds to missing AX , ABX , $AX\gamma$ and $ABX\gamma$ interactions for each $\gamma \in \Gamma$; A missing (A, B) corresponds to missing AB , $AB\gamma$, $AB\gamma\zeta$ interactions for each $\gamma, \zeta \in \Gamma$.

6.2 MCMC for fixed models

Again, a Gibbs sampler may be used to generate from the posterior distribution for any fixed model, using random walk Metropolis steps for each update. Several pilot runs may be required to achieve satisfactory mixing. As in the previous chapter, acceptance ratios may simplify from the full posterior ratio but it is more convenient to compute block conditional density ratios. Recall that, as ever, prior normalizing constants are not required for inference based on a single fixed model.

Updating Λ The first element, $\Lambda[0]$ is the likelihood normalizing constant, λ , so is never updated independently. It is however changed whenever any other parameters are updated as it depends on all of them. The next $(l_B - 1) + (l_A - 1)$ elements are the discrete main effects so are always updated. The remaining elements, discrete interactions, are updated if and only if the AB interaction is present. The acceptance ratio is given by 5.1.

Updating H The first row consists of the continuous linear main effects so are always updated. The remainder are updated if and only if the corresponding interactions are present, that is if unconstrained, which in the case of uncentred data means if nonzero. The acceptance ratio is given by 5.2.

Updating Ω The precision matrices may be divided into four groups, similarly to the table in the example:

1. $\Omega(0, 0)$: $\tau_X(0, 0)$ is always updated, to $\tau_X^*(0, 0)$.

If the AXX interaction is missing, we also set each $\tau_X^*(i_A, 0) = \tau_X^*(0, 0)$.

If the BXX interaction is missing, set each $\tau_X^*(0, i_B) = \tau_X^*(0, 0)$.

If both are present but the $ABXX$ interaction is not, it changes each $\omega_{XX}(i_A, i_B) \equiv \tau_X(i_A, i_B)$ for $i_A \neq 0, i_B \neq 0$, according to 6.4.

If any AXY interaction is missing, we also have

$$\rho_{XY}^*(i_A, 0) = \frac{\tau_X^*(0, 0)\tau_Y(0, 0)\rho_{XY}(0, 0)}{\tau_X^*(i_A, 0)\tau_Y(i_A, 0)} \quad i_A = 1, \dots, l_A - 1.$$

Similarly, if any BXY interaction is missing, we also have

$$\rho_{XY}^*(0, i_B) = \frac{\tau_X^*(0, 0)\tau_Y(0, 0)\rho_{XY}(0, 0)}{\tau_X^*(0, i_B)\tau_Y(0, i_B)} \quad i_B = 1, \dots, l_B - 1.$$

If any $ABXY$ interaction is missing but not the AXY or BXY , $\omega_{XY}(i_A, i_B)$ is changed, according to 6.4, for each $i_A = 1, \dots, l_A - 1$ and $i_B = 1, \dots, l_B - 1$, from which we can calculate $\rho_{XY}^*(i_A, i_B)$, although there is no need to do this.

$\rho_{XY}(0, 0)$ is updated if and only if the XY interaction is present. If any of the AXY , BXY or $ABXY$ interactions are missing, the above applies again.

2. $\Omega(0, i_B)$ $i_B \neq 0$: $\tau_X(0, i_B)$ is updated to $\tau_X^*(0, i_B)$ if and only if the BXX interaction is present. If the AXX interaction is not present, we also set each $\tau_X^*(i_A, i_B) = \tau_X^*(0, i_B)$. If it is but the $ABXX$ is not, we set each $\tau_X^*(i_A, i_B) = \tau_X(i_A, i_B) - \tau_X(0, i_B) + \tau_X^*(0, i_B)$ for $i_A \neq 0$.

If any AXY interaction is missing, we also have

$$\rho_{XY}^*(i_A, i_B) = \frac{\tau_X^*(0, i_B)\tau_Y(0, i_B)\rho_{XY}(0, i_B)}{\tau_X^*(i_A, i_B)\tau_Y(i_A, i_B)} \quad i_A = 1, \dots, l_A - 1.$$

Similarly, if any BXY interaction is missing, we also have

$$\rho_{XY}^*(0, i_B) = \frac{\tau_X(0, i_0)\tau_Y(0, 0)\rho_{XY}(0, 0)}{\tau_X^*(0, i_B)\tau_Y(0, i_B)} \quad i_B = 0, \dots, l_B - 1.$$

If any $ABXY$ interaction is missing but not the AXY or BXY , for each $i_A = 1, \dots, l_A - 1$, $\omega_{XY}(i_A, i_B)$ is changed according to 6.4.

$\rho_{XY}(0, i_B)$ is updated if and only if the XY , BXX , BYY and BXY interactions are all present. If the AXY interaction is missing, we set

$$\rho_{XY}^*(i_A, i_B) = \frac{\tau_X(0, i_B)\tau_Y(0, i_B)\rho_{XY}^*(0, i_B)}{\tau_X(i_A, i_B)\tau_Y(i_A, i_B)} \quad i_A = 1, \dots, l_A - 1.$$

If not, but the $ABXY$ interaction is missing, we again change $\omega_{XY}(i_A, i_B)$ according to 6.4.

3. $\Omega(i_A, 0) \quad i_A \neq 0$: $\tau_X(i_A, 0)$ is updated to $\tau_X^*(i_A, 0)$ if and only if the AXX interaction is present. $\rho_{XY}(i_A, 0)$ is updated if and only if all three of the AXX , AYY and AXY interactions are present. The procedure in each case is exactly the same as with $\Omega(0, i_B)$, with the roles of A and B reversed.
4. $\Omega(i_A, i_B) \quad i_A \neq 0 \quad i_B \neq 0$: $\tau_X(i_A, i_B)$ is updated if and only if all three of the AXX , BXX and $ABXX$ interactions are present. If $ABXY$ is absent, we again change $\omega_{XY}(i_A, i_B)$ according to 6.4.

$\rho_{XY}(i_A, i_B)$ is updated if and only if all of the interactions XY , AXX , BXX , $ABXX$, AYY , BYY , $ABYY$, AXY , BXY and $ABXY$ are present. This is the most straightforward update as no other quantities need to be changed at the same time.

For each of the above updates, \mathbf{H} must also be amended if any linear interactions are missing and $\mathbf{\Lambda}$ must also be amended if the AB interaction is missing. The acceptance ratio for each is most conveniently obtained as the full posterior ratio even though it may simplify much further.

6.3 Reversible Jump Sampling

In principle, this proceeds much as in the case of a single discrete variable: A particular edge or interaction is chosen at random. If it is present in the current

model, it is proposed to remove it, otherwise it is proposed to add it. The main difference here is that there are discrete-discrete edges/interactions to consider so there are new move types, namely

adding/removing a discrete-discrete edge and discrete-discrete, discrete-discrete-continuous and discrete-discrete-continuous-continuous interactions. Again, it must be taken into consideration which moves are valid in the case of hierarchical models as only one interaction is removed/added at a time. Any proposed moves that are not valid must be rejected. Again also, since current and proposal models differ in exactly one edge/interaction, the jump ratio is always 1.

As in the case of models with one discrete variable, random walk variances which give good mixing are first obtained from runs of fixed models and these are also used for between-model moves. This time, however, more models are required: A run of the saturated model gives variances for all unconstrained parameters. The model with all edges or quadratic interactions (AXX , AXY , $ABXX$, $ABXY$) involving A and a continuous variable but no others missing gives variances for τ 's which vary only between levels of B . Similarly, the model with all edges or quadratic interactions involving A and a continuous variable but no others missing gives variances for τ 's which vary only between levels of A . The model with all mixed edges or mixed quadratic interactions missing gives variances for τ 's which do not vary between cells.

Arguably, the model with the (A, B) edge or all quadratic interactions involving A and B should be run also but for all examples presented here it has been found that the variances for the τ 's in the saturated model suffice for those in this reduced model.

6.3.1 Hierarchical Models

For the purposes of illustration, the following example with two binary and two continuous variables will be used.

Example: $l_A = l_B = 2$, $\Gamma = \{X, Y\}$, with parameters

$$\Lambda = \begin{pmatrix} \lambda \\ \lambda^B \\ \lambda^A \\ \lambda^{AB} \end{pmatrix} \quad \mathbf{H} = \begin{pmatrix} \eta \\ \eta^B \\ \eta^A \\ \eta^{AB} \end{pmatrix} = \begin{pmatrix} \eta_X & \eta_Y \\ \eta_X^B & \eta_Y^B \\ \eta_X^A & \eta_Y^A \\ \eta_X^{AB} & \eta_Y^{AB} \end{pmatrix}$$

$$\begin{pmatrix} \tau_X(0,0) & \tau_Y(0,0) & \rho(0,0) \\ \tau_X(0,1) & \tau_Y(0,1) & \rho(0,1) \\ \tau_X(1,0) & \tau_Y(1,0) & \rho(1,0) \\ \tau_X(1,1) & \tau_Y(1,1) & \rho(1,1) \end{pmatrix}.$$

As in the previous chapter, the parameters present in each model (current and proposed) will be explicitly stated except for Λ and/or \mathbf{H} , when their components are all present in both. An asterisk (*) will denote a parameter that may or may not be present. An absent parameter will be denoted by a zero or a blank space, as appropriate.

Discrete Interactions

Example:

$$\begin{pmatrix} \lambda \\ \lambda^B \\ \lambda^A \\ 0 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\eta} \\ * \\ * \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \tau_X(0,0) & \tau_Y(0,0) & \rho(0,0) \\ * & * & * \\ * & * & * \end{pmatrix}$$

$$\longleftrightarrow \begin{pmatrix} \lambda \\ \lambda^B \\ \lambda^A \\ \lambda^{AB} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\eta} \\ * \\ * \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \tau_X(0,0) & \tau_Y(0,0) & \rho(0,0) \\ * & * & * \\ * & * & * \end{pmatrix}$$

Suppose firstly that the AB interaction has been selected. To add this interaction, (and this move is always permitted since there are no lower order interactions involving A and B) we must generate

$(l_A - 1) * (l_B - 1)$ new λ -terms, namely $\Lambda[i]$, $i = l_A + l_B - 1, \dots, l_A l_B - 1$, corresponding to $\lambda^{AB}(i_A, i_B)$ for $i_A = 0, \dots, l_A - 1$ and $i_B = 0, \dots, l_B - 1$.

Two simple choices of proposal distribution are the prior and a normal distribution with mean and variance given by estimates of the posterior mean and variance, taken from a pilot run of the (fixed) saturated model (or even any model with the AB interaction present). The latter, naturally, tends to generate better proposals and is the one used here. The Jacobian in either case is 1, the prior ratio is simply the product of the (independent) priors for the parameters generated, the proposal ratio is simply the reciprocal of the product of the normal densities used to generate the new parameters and the likelihood ratio is

$$\frac{\exp(\mathbf{N}'\mathbf{D}\Lambda^*)}{\exp(\mathbf{N}'\mathbf{D}\Lambda)}.$$

To remove this interaction, we must set these same λ -terms to the corresponding entries in $\mathbf{D}^{-1}\mathcal{U}_2$ or to zero if not using centring. The acceptance ratio is simply the reciprocal of that for adding. Note that this move is only permitted if the $AB\gamma$ and $AB\gamma\zeta$ interactions are missing for all $\gamma, \zeta \in \Gamma$.

2-way Mixed Linear Interactions

Suppose next the two-way mixed linear interaction AX has been selected. *Example:*

$$\begin{aligned} & \begin{pmatrix} \eta_X & \eta_Y \\ * & * \\ 0 & * \\ 0 & * \end{pmatrix}, \begin{pmatrix} \tau_X(0,0) & \tau_Y(0,0) & \rho(0,0) \\ * & * & * \\ & * & \\ & * & \end{pmatrix} \\ \longleftrightarrow & \begin{pmatrix} \eta_X & \eta_Y \\ * & * \\ \eta_X^A & * \\ 0 & * \end{pmatrix}, \begin{pmatrix} \tau_X(0,0) & \tau_Y(0,0) & \rho(0,0) \\ * & * & * \\ & * & \\ & * & \end{pmatrix} \end{aligned}$$

To add this interaction, and this move is always permitted since it is a 2-way interaction, we must generate $l_A - 1$ new η -terms. These are $\mathbf{H}[i][X]$ $i = l_B, \dots, l_A + l_B - 1$ and the terms η_X^{AB} remain constrained. Note that this will change $\mathbf{\Lambda}$ if the AB interaction is missing. This can be done in the same way as generating λ -terms and the acceptance ratio is similar with likelihood ratio the same.

To remove the interaction, these η -terms are set to the corresponding entries in $\mathbf{D}^{-1}\mathcal{U}_1$, or to zero if not using centring. Again, the acceptance ratio is the reciprocal of that for adding the interaction. Since only hierarchical models are being considered, this move is only permitted if all quadratic interactions involving A and X are missing, namely, $AX\gamma$ and $ABX\gamma$ for all $\gamma \in \Gamma$.

The same applies if the BX interaction has been selected but this time the η -terms in question are $\mathbf{H}[i][X]$ $i = 1, \dots, l_B - 1$.

3-way Mixed Linear Interactions

Example:

$$\begin{aligned} & \begin{pmatrix} \eta_X & \eta_Y \\ \eta_X^B & * \\ \eta_X^A & * \\ 0 & * \end{pmatrix}, \begin{pmatrix} \tau_X(0,0) & \tau_Y(0,0) & \rho(0,0) \\ * & * & * \\ * & * & * \\ & * & \end{pmatrix} \\ \longleftrightarrow & \begin{pmatrix} \eta_X & \eta_Y \\ \eta_X^B & * \\ \eta_X^A & * \\ \eta_X^{AB} & * \end{pmatrix}, \begin{pmatrix} \tau_X(0,0) & \tau_Y(0,0) & \rho(0,0) \\ * & * & * \\ * & * & * \\ & * & \end{pmatrix} \end{aligned}$$

If the ABX linear interaction has been selected, the parameters in question are $\mathbf{H}[i][X]$ $i = l_A + l_B - 1, \dots, l_A l_B - 1$, corresponding to the η_X^{AB} -terms. The procedure is otherwise the same as for AX and BX interactions. Note however, that, in order to keep the model hierarchical, the addition of this interaction is only permitted when the current model contains both AX and BX interactions as well as the AB interaction. The removal move is permitted only if the $ABXX$ quadratic interaction, and hence $ABX\gamma$ for all $\gamma \in \Gamma$, are absent from the current model. Hence, for both addition and removal, Λ is not altered except for the first entry, the likelihood normalizing constant, which is always changed by any change in the other parameters, and the likelihood ratio becomes

$$\frac{\exp [n\lambda_\emptyset^*]}{\exp [n\lambda_\emptyset]}.$$

2-way Quadratic Interactions

Suppose next that the 2-way quadratic interaction AXX has been selected.

Example:

$$\begin{aligned} & \begin{pmatrix} \eta_X & \eta_Y \\ \eta_X^B & * \\ \eta_X^A & * \\ * & * \end{pmatrix}, \begin{pmatrix} \tau_X(0,0) & \tau_Y(0,0) & \rho(0,0) \\ * & * & * \\ & * & \\ & * & \end{pmatrix} \\ \longleftrightarrow & \begin{pmatrix} \eta_X & \eta_Y \\ \eta_X^B & * \\ \eta_X^A & * \\ * & * \end{pmatrix}, \begin{pmatrix} \tau_X(0,0) & \tau_Y(0,0) & \rho(0,0) \\ * & * & * \\ \tau_X(1,0) & * & \\ & * & \end{pmatrix} \end{aligned}$$

In hierarchical models, addition of this interaction is only permitted if the corresponding linear interaction, AX , is present in the current model. Removal is only permitted if the $ABXX$ and all $AX\gamma$ for $\gamma \in \Gamma$ interactions, and hence all $ABX\gamma$ interactions, are absent from the current model. Note that in both cases, since only hierarchical models are being considered, $AX\gamma$ and $ABX\gamma$, $\gamma \in \Gamma$, are absent from, and AX is present in, both current and proposal models.

There are then two cases, depending on whether the BXX interaction is present or not. If it is not present, we currently have (if adding) a common τ_X for all cells and we require one $\tau_X^*(i_A)$ for each level i_A of A . If BXX is present, we currently have (if adding) one $\tau_X(i_B)$ for each level, i_B of B and for each level of B we require one $\tau_X^*(i_A, i_B)$ for each level of A . Contrariwise of course for the reverse move.

The procedure is exactly as for models with one discrete variable but in the case of BXX present, is repeated for each level of B and in the case of BXX absent, is done at the first level of B and then we have $\tau_X^*(i_A, i_B) = \tau_X^*(i_A, 0)$ by constraint. As before the proposals may be generated either by constraint or independently. Note also that if the AB interaction is missing, $\mathbf{\Lambda}$ must be changed and if any linear interactions are missing, \mathbf{H} must be changed. Hence the full likelihood must be used in general.

Naturally, the same applies for the BXX interaction, with the roles of A and B reversed.

3-way Discrete-Discrete-Continuous Quadratic Interactions

Next suppose that the 3-way quadratic interaction $ABXX$ has been selected, a move type not encountered in single discrete variable models.

Example:

$$\begin{aligned} & \begin{pmatrix} \eta_X & \eta_Y \\ \eta_X^B & * \\ \eta_X^A & * \\ \eta_X^{AB} & * \end{pmatrix}, \begin{pmatrix} \tau_X(0,0) & \tau_Y(0,0) & \rho(0,0) \\ \tau_X(0,1) & * & * \\ \tau_X(1,0) & * & * \\ & * & * \end{pmatrix} \\ \longleftrightarrow & \begin{pmatrix} \eta_X & \eta_Y \\ \eta_X^B & * \\ \eta_X^A & * \\ \eta_X^{AB} & * \end{pmatrix}, \begin{pmatrix} \tau_X(0,0) & \tau_Y(0,0) & \rho(0,0) \\ \tau_X(0,1) & * & * \\ \tau_X(1,0) & * & * \\ \tau_X(1,1) & * & * \end{pmatrix} \end{aligned}$$

For hierarchical models, the addition of this interaction is permitted only if the AB , AXX , BXX and ABX (and hence AX and BX) interactions are present.

Removal is permitted only if each $ABX\gamma$ is missing. Note that in each case, AB , AXX , BXX , ABX , AX and BX are present in and each $ABX\gamma$ is absent from both current and proposal models.

To remove this interaction we need to impose the constraint 6.4, which for the example is

$$\tau_X^*(1, 1) = \tau_X^*(0, 1) + \tau_X^*(1, 0) - \tau_X^*(0, 0)$$

The first thing to note is that since the $ABXX$ interaction is present in the current model, none of the current τ_X 's are constrained so we have the maximum number - one for each cell. We could generate the unconstrained τ_X^* but this is generally not necessary and the remainder are given by constraint.

For the reverse move, we need to generate unconstrained $\tau_X^*(i_A, i_B)$'s for $i_A > 0$ and $i_B > 0$. The remainder are generated only if they are also generated for the addition move. Similarly to when adding the AXX interaction, proposal generation is by a random walk style move, that is $\tau_X^*(i_A, i_B) = \tau_X(i_A, i_B) + u_{i_A, i_B}$, where each u is from a Normal distribution with mean zero and variance given by the posterior estimate of the variance of $\tau_X(i_A, i_B)$, based on a pilot run of the saturated model. The proposal ratio is then the reciprocal of the product of these proposal densities, evaluated at the u 's and the Jacobian is 1. The prior ratio is

$$\prod_{i_A > 0, i_B > 0} f(\tau_X^*(i_A, i_B))$$

for addition and

$$\prod_{i_A > 0, i_B > 0} [f(\tau_X(i_A, i_B))]^{-1}$$

for removal.

If any linear interactions are missing, \mathbf{H} must be amended. $\mathbf{\Lambda}$ is not amended since the AB interaction is necessarily present. The likelihood ratio is given by 5.3.

Continuous-Continuous Interactions

Suppose next that the XY continuous interaction is to be added.

Example:

$$\begin{pmatrix} \tau_X(0, 0) & \tau_Y(0, 0) & 0 \\ * & * & 0 \\ * & * & 0 \\ * & * & 0 \end{pmatrix}$$

$$\longleftrightarrow \begin{pmatrix} \tau_X(0,0) & \tau_Y(0,0) & \rho(0,0) \\ * & * & \\ * & * & \\ * & * & \end{pmatrix}$$

This move is always permitted and since XY is currently missing and the current model is hierarchical, AXY , BXY , and $ABXY$ are also absent. Therefore, as in the previous chapter, only one new partial correlation, $\rho_{XY}^*(0,0)$, is required as the remainder will be determined by the constraint. This parameter is therefore generated the same way. Note that if any linear interactions are missing, \mathbf{H} must also be amended and if AB is missing, $\mathbf{\Lambda}$ must be amended and hence in general, the full likelihood ratio must be used.

If XY is to be removed, simply set each $\rho_{XY}(i_A, i_B) = 0$. To keep the model hierarchical, this move is only permitted if all quadratic interactions involving both X and Y , namely AXY , BXY and $ABXY$, are currently missing. As in the previous chapter, the interval endpoints from which $\rho_{XY}^*(0,0)$ is generated must be obtained but otherwise the acceptance ratio is the same as for the addition move.

3-way Discrete-Continuous-Continuous Quadratic Interactions

Next suppose the AXY interaction is to be added.

Example:

$$\begin{pmatrix} \tau_X(0,0) & \tau_Y(0,0) & \rho(0,0) \\ * & * & * \\ \tau_X(1,0) & \tau_Y(1,0) & \\ * & * & \end{pmatrix} \longleftrightarrow \begin{pmatrix} \tau_X(0,0) & \tau_Y(0,0) & \rho(0,0) \\ * & * & * \\ \tau_X(1,0) & \tau(1,0) & \rho(1,0) \\ * & * & \end{pmatrix}$$

Note that since only hierarchical models are being considered, this move is permitted only if XY , AXX and AYY are currently present and $ABXY$ is necessarily absent.

There are two cases to consider, depending on whether BXY is present or not. If not, the current $\omega_{XY}(i_A, i_B)$'s are all equal and the proposal $\omega_{XY}^*(i_A, i_B)$'s differ between levels of A . Therefore, the proposals $\omega_{XY}^*(i_A, 0)$'s, for the first level

of B , may be generated in the same way as when adding the AXY interaction in a model with a single discrete variable and the remainder are given by $\omega_{XY}^*(i_A, i_B) = \omega_{XY}^*(i_A, 0)$.

If BXY is present, the proposals are generated in the same way as when removing the $ABXY$ edge (see below) as the proposal models are the same.

Once again, if any linear interactions are missing, \mathbf{H} must be amended and if AB is missing $\mathbf{\Lambda}$ must also be amended and the full likelihood ratio must be used in general.

To remove the AXY interaction, there are again two cases depending on whether the BXY interaction is present. If not, the proposal $\omega_{XY}^*(i_A, i_B)$'s are all equal so there is only one parameter to generate and this is done in the same way as when adding the XY interaction as the proposal models are the same.

If BXY is present, the $\omega_{XY}(i_A, i_B)$'s currently satisfy constraint 6.4. The proposals must only satisfy $\omega_{XY}^*(i_A, i_B) = \omega_{XY}^*(0, i_B)$ for each i_B and therefore may be generated in the same way as when removing AXY when B is not in the model but this is done separately for each level of B .

For hierarchical models, this move is only permitted if $ABXY$ is currently absent. In each case, the relevant interval endpoints for generating proposals in the reverse move must be obtained to give the numerator of the proposal ratio.

Of course, all this applies to the BXY interaction, with the roles of A and B reversed.

4-way Discrete-Discrete-Continuous-Continuous Quadratic Interactions

Finally suppose the $ABXY$ term, the highest possible order of term when $p = 2$, has been selected.

Example:

$$\begin{aligned} & \begin{pmatrix} \tau_X(0, 0) & \tau_Y(0, 0) & \rho(0, 0) \\ \tau_X(0, 1) & \tau_Y(0, 1) & \rho(0, 1) \\ \tau_X(1, 0) & \tau_Y(1, 0) & \rho(1, 0) \\ \tau_X(1, 0) & \tau_X(1, 0) & \end{pmatrix} \\ & \longleftrightarrow \begin{pmatrix} \tau_X(0, 0) & \tau_Y(0, 0) & \rho(0, 0) \\ \tau_X(0, 1) & \tau_Y(0, 1) & \rho(0, 1) \\ \tau_X(1, 0) & \tau_Y(1, 0) & \rho(1, 0) \\ \tau_X(1, 1) & \tau_Y(1, 1) & \rho(1, 1) \end{pmatrix} \end{aligned}$$

To keep the model hierarchical, addition of this term is only permitted if all other interactions involving two or more of A , B , X and Y are present. There is

therefore one $\rho_{XY}^*(i_A, i_B)$ for each cell and these are generated independently in the usual way.

Removal of this term is always permitted as there are no higher order interactions and proposals must satisfy the constraint 6.4. For A and B binary as in the example, this is $\omega_{XY}^*(1, 1) = \omega_{XY}^*(0, 1) + \omega_{XY}^*(1, 0) - \omega_{XY}^*(0, 0)$.

In order to preserve positive definiteness, proposals are generated as follows for the example, dropping the XY subscript to ease notation :

$\rho^*(0, 1)$ and $\rho^*(1, 0)$ are generated as for the saturated model.

The endpoints of the intervals to ensure positive definiteness of the remaining precision matrices are found so we have

$a_{00} < \rho^*(0, 0) < b_{00}$ and for each cell (i, j) , $a_{ij} < \rho^*(i, j) < b_{ij}$. From the constraint above, we thus have

$$a_{11} < \frac{1}{\tau_X(1, 1)\tau_Y(1, 1)} (\omega(0, 1) + \omega(1, 0) - \omega(0, 0)) < b_{11}.$$

Hence,

$$\tau_X(1, 1)\tau_Y(1, 1)b_{11} + \omega(0, 1) + \omega(1, 0) < \omega(0, 0) < \tau_X(1, 1)\tau_Y(1, 1)a_{11} + \omega(0, 1) + \omega(1, 0)$$

and finally,

$$\frac{1}{\tau_X(0, 0)\tau_Y(0, 0)} (\tau_X(1, 1)\tau_Y(1, 1)b_{11} + \omega(0, 1) + \omega(1, 0))$$

$$< \rho(0, 0)$$

$$< \frac{1}{\tau_X(0, 0)\tau_Y(0, 0)} (\tau_X(1, 1)\tau_Y(1, 1)a_{11} + \omega(0, 1) + \omega(1, 0)),$$

giving the interval from which to draw $\rho^*(0, 0)$.

These expressions may seem very awkward but are really quite manageable even when there are more cells. Note that if there are only two continuous variables, the endpoints are always -1 and 1 since any value between these gives a positive definite precision matrix so no computation is required.

The remainder of the $\omega_{XY}(i, j)$'s are given by the constraint. Once again, if any linear interactions are missing, \mathbf{H} must be amended but since AB is necessarily present, \mathbf{A} need not be and the likelihood ratio is given by 5.3.

6.3.2 Graphical Models

Here, there are three types of edge and hence six different move types (addition and removal of each), although there may be different cases to consider depending on which other edges are currently present.

Continuous-Continuous edges

Suppose firstly that the edge (X, Y) has been chosen to be added to the graph. There are five different cases to consider depending on which other edges are present.

If all of the (A, B) , (A, X) , (B, X) , (A, Y) and (B, Y) edges are present, we must generate one ρ_{XY}^* for each cell. This is done in the usual way, by drawing from the uniform distribution over the interval from which draws preserve positive definiteness.

If either (or both) (A, X) or (A, Y) is missing but neither of (B, X) or (B, Y) are, we only require one ρ_{XY}^* for each level of B as $\omega_{XY}^*(i_A, i_B)$ is the same across levels of A , for fixed i_B . For each level, j , of B , $\omega_{XY}^*(0, i_B)$ is drawn in the same way as when removing either (A, X) or (A, Y) in a model without B . Then for each $i = 0, 1, \dots, l_A - 1$,

$$\omega_{XY}^*(i, j) = \omega_{XY}^*(0, j) = \rho_{XY}^*(0, j)\tau_X(0, j)\tau_Y(0, j).$$

The same applies if either (or both) (B, X) or (B, Y) is missing but neither of (A, X) or (A, Y) are, with the roles of A and B reversed.

If either (or both) (A, X) or (A, Y) and either (or both) (B, X) or (B, Y) are missing, we require only to generate $\rho_{XY}^*(0, 0)$ and the remainder are given by $\omega_{XY}^*(i, j) = \omega_{XY}^*(i, j) = \rho_{XY}^*(0, 0)\tau_X(0, 0)\tau_Y(0, 0)$. This parameter is generated as in the previous two cases but across all cells, not just across levels of A or B .

The final case is when (A, B) , (A, X) , (B, X) and (A, Y) are all present but (A, B) is not.

Proposals for this move are generated in the same way as when removing the $ABXY$ interaction in a hierarchical model as the proposal models are the same.

If (X, Y) has been chosen for removal, as before, simply set each ρ_{XY}^* to zero, regardless of which other edges are present or absent. Again, the interval endpoints to determine the proposal ratio must be obtained as when adding the edge.

Discrete-Continuous edges

Suppose next that the (A, X) edge has been chosen for addition. If both (A, B) and (B, X) edges are present, the procedure for generating proposal concentrations is as when adding this type of edge in a model with one discrete variable but is carried out within each level of B . Linear mixed interactions are generated in the same way as when adding AX and ABX interactions in a hierarchical model simultaneously. Note that if any other mixed edges are absent, \mathbf{H} must be amended according to \mathfrak{Q} . Since (A, B) is present, \mathbf{A} is not changed and the likelihood ratio is 5.3.

If (B, X) is not present, the procedure for generating proposal concentrations is again as when adding the (A, X) edge in a model without B but is carried out at only one level of B , as the proposal concentrations are constrained to be the same between levels of B but allowed to differ between levels of A . It makes no difference whether (A, B) is present or not in this case as its absence does not constrain proposals further than they are by the absence of the (B, X) edge. Again, if any other mixed edges are absent, \mathbf{H} must be amended according to \mathfrak{Q} and if (A, B) is absent, so must \mathbf{A} and the full likelihood ratio must be used.

If (B, X) is present but (A, B) is not, we have a combination of the addition of AXX and $AX\gamma$ interactions for each $\gamma \in \Gamma$ for hierarchical models. Again, if any other mixed edges are absent, \mathbf{H} must be amended according to \mathfrak{Q} and so must \mathbf{A} as (A, B) is absent. Again, the full likelihood ratio must be used.

If the (A, X) edge is to be removed, we have a combination of the removal of each ABX , AX , $ABX\gamma$ and $AX\gamma$ in the case of hierarchical models, implemented as follows:

1. If the (B, X) edge is absent, τ_X 's differ only between levels of A so at the first level of B , proposal τ_X^* 's are generated in the same way as when removing the (A, X) edge in the previous chapter and the remainder are given by constraint.

If the (B, X) is present as well as (A, B) , proposal τ_X^* 's are generated in the same way as when removing the (A, X) edge in the previous chapter but within each level of B .

If (B, X) is present but (A, B) is not, proposal τ_X^* 's are generated in the same way as when removing the AXX interaction in the presence of the BXX interaction.

2. For each $X \neq \gamma \in \Gamma$, if (B, X) and (B, γ) are both present, proposal $\rho_{X\gamma}^*$'s are generated, within levels of B , in the same way as when removing the (A, X) edge in the previous chapter (or, indeed, when removing the $AX\gamma$ interaction). Otherwise, proposal $\rho_{X\gamma}^*$'s are generated in the same way as when removing the $AX\gamma$ interaction in the absence of the $BX\gamma$ interaction.
3. \mathbf{H}^* is generated as a combination of removal of AX and ABX interactions, that is each η_X^A and η_X^{AB} is constrained by setting each $\mathbf{H}^*[i][X]$, for $i = l_B, \dots, l_A l_B - 1$, to the corresponding entry in $\mathbf{D}^{-1}\mathcal{U}_1^*$. Some of these may be already constrained if (B, X) is absent. If any other mixed edges are absent, \mathbf{H}^* is further amended according to $\mathbf{\Omega}^*$.
4. Finally, if (A, B) is absent, $\mathbf{\Lambda}$ must be amended.

Discrete-Discrete edges

Removal of the (A, B) edge corresponds to the removal of the AB interaction as well as any of $AB\gamma$ and $AB\gamma\zeta$ interactions that may be currently present. Specifically, the procedure is as follows:

1. Impose the constraints 6.4 on the partial precisions. Some may already satisfy this if any mixed edges are missing.
2. Generate the proposal partial correlations in the same way as when removing an $ABXY$ interaction, but applied to *each* pair of continuous variables.
3. Constrain each η_γ^{AB} - term by setting each $\mathbf{H}[i][\gamma]$ for $i = l_A + l_B - 1, \dots, l_A l_B - 1$ to the corresponding entry in $\mathbf{D}^{-1}\mathcal{U}_1^*$ or to zero if not using centring. Also adjust any other constrained entries of \mathbf{H} .
4. Constrain each $\lambda^{AB}(i_A, i_B)$ for $i_A = 0, \dots, l_A - 1$ and $i_B = 0, \dots, l_B - 1$ by setting $\mathbf{\Lambda}[i]$, $i = l_A + l_B - 1, \dots, l_A l_B - 1$ to the corresponding entries in $\mathbf{D}^{-1}\mathcal{U}_2^*$ or to zero if not using centring.

To add the (A, B) edge,

1. Update $\mathbf{\Lambda}$ in the same way as when adding the AB interaction in a hierarchical model.
2. For each $\gamma \in \Gamma$, if the (A, γ) and (B, γ) edges are present, generate η_γ^{AB} as when adding the $AB\gamma$ interaction in a hierarchical model.

3. For each $\gamma \in \Gamma$, if the (A, γ) and (B, γ) edges are present, generate τ_γ^* 's as when adding the $AB\gamma\gamma$ interaction in a hierarchical model.
4. For each $\gamma, \zeta \in \Gamma$, if (A, γ) , (A, ζ) , (B, γ) and (B, ζ) are all present, generate $\rho_{\gamma\zeta}^*$'s in the same way as when adding the $AB\gamma\zeta$ interaction in a hierarchical model.
5. Finally, amend any constrained entries in \mathbf{H} using $\mathbf{\Omega}^*$.

6.4 Examples

In this section, an example using simulated data is presented as well as two examples to be found in (Whittaker 1990), one of which also appears in Edwards(1990,1995).

Only results based on graphical models are given as a satisfactory sampler for hierarchical models has proved surprisingly difficult to implement. It may be expected that a sampler for graphical models would be more problematic as the between-model moves are larger but this has not proved to be the case. It is possible that the larger moves for graphical models avoid some problems that occur for hierarchical models. The same prior distributions were used as in the previous chapter, with $\beta = 0.001$ in the prior for the partial precisions. The issue of prior sensitivity will not be pursued here but it is no more of a problem than in the previous chapter.

As before, posterior model probabilities are based on runs of 100 000 iterations after 10 000 of burn-in and MCMC standard errors are based on batches of size 1000.

6.4.1 A simulated data example

In order to test the reversible jump sampler for two discrete variable models by comparing with that for one discrete variable models and to demonstrate its operation, data based on the mice data presented in the previous chapter were generated as follows:

Maximum likelihood estimates under model 52 (recall this had a complete graph on AXY and no edge to Z), which had two-thirds of the posterior probability and was chosen by Edwards, were obtained using MIM. Two sets of data were generated from the model with these estimates as parameters, one for each level of a dummy binary variable. The graph of the model used to generate the

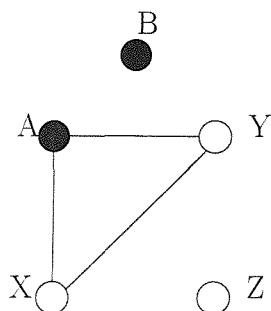


Figure 6.1: Graph of the model used to generate a set of simulated test data

data is thus the one in Figure 6.1, numbered 388 by the indexing scheme described above. The intention is that the models with highest posterior probability will be similar to those for the mice data but with an additional discrete vertex (with few or no adjacencies) and this is largely what occurred. The eight most probable graphs, which are all with probabilities greater than 5% are tabulated in Table 6.2 and the edge inclusion rates in Table 6.1. The most probable graph is 388, the true graph and the next few most probable are very similar, with one or two edges different. Trace plots for the parameters are omitted for brevity but they show satisfactory mixing, as with previous examples.

It is worth noting that the posterior is very diffuse but considering the large model space (1024 graphs), this is not a great concern. Deviance-based forward inclusion in MIM selects 388 and backwards elimination selects 916, which has two additional edges, (A, B) and (B, Y) .

AB	AX	AY	AZ	BX	BY	BZ	XY	XZ	YZ
25.6	80.1	57.6	7.9	5.1	10.3	7.7	97.7	41.4	38.9

Table 6.1: Inclusion %ages for edges for simulated data

6.4.2 A drug trial using rats

This example is treated in Morrison (1976), Mardia, Kent and Bibby (1979), Edwards (1987,1990,1995) and Whittaker (1990). The data are from a randomized drug trial in which weight losses of rats under three drug treatments are studied. Four male and four female rats were given each drug and their weight losses after one and two weeks were measured. Hence there are two discrete variables, A (sex), with two levels, and B (treatment), with three levels, and two continuous

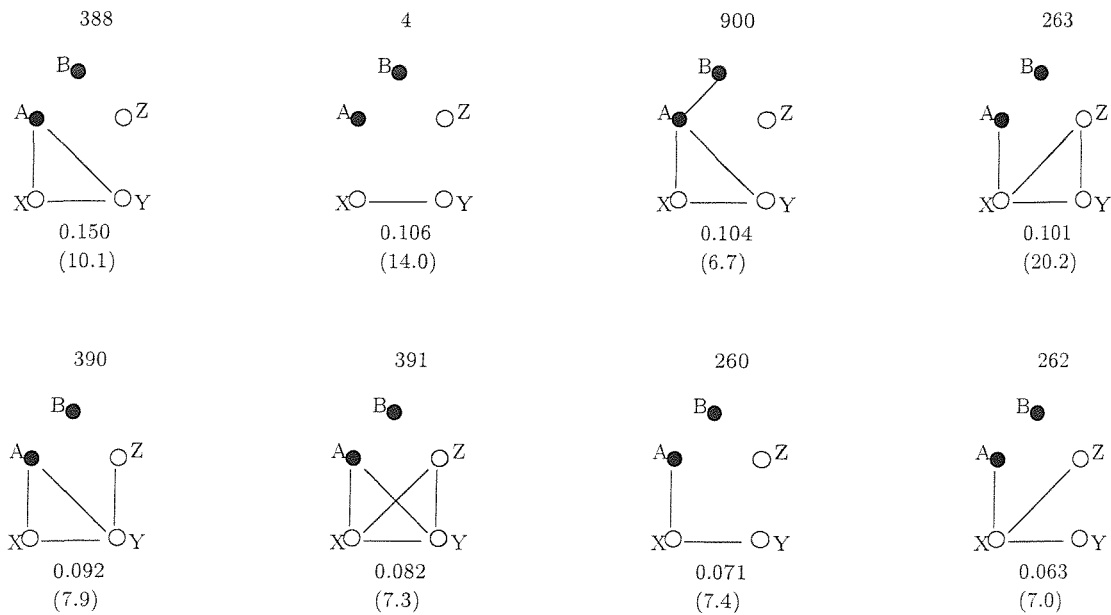


Table 6.2: Posterior model probabilities for simulated data, graphical models. Standard errors $\times 10^4$ are in brackets.

variables, X (weight loss after one week) and Y (weight loss after two weeks).

In MIM, backward elimination selects the saturated model; Forward inclusion selects model 7, which has all edges connecting A missing.

Edwards uses a stepwise deviance-based procedure in MIM to select model 39, which has (A, X) and (A, Y) missing.

The model selected by Morrison (1976) and Mardia et al. (1979) is nongraphical and contains the AB, BX, BY, and XY interactions only, without the BXX, BYY and BXY interactions in Edwards' model.

The posterior model probabilities according to the reversible jump output are tabulated in Table 6.3 and the edge inclusion percentages in Table 6.4. Trace plots of the batch posterior probabilities for the four most probable graphs are displayed in Figure 6.2.

The most probable graphs are 7 and 5, each with approximately one quarter of the posterior probability. The absence of the (A, B) edge, while disagreeing with Morrison's and Edwards' selected models, is not so surprising given the designed nature of the experiment (the cell counts were clearly chosen in advance). This is also an instance of a deviance-based method selecting a model with more edges than a Bayesian method. Edwards' graph is the next most probable, with about half the probability of 7 and 5. The remaining graphs are much less probable.

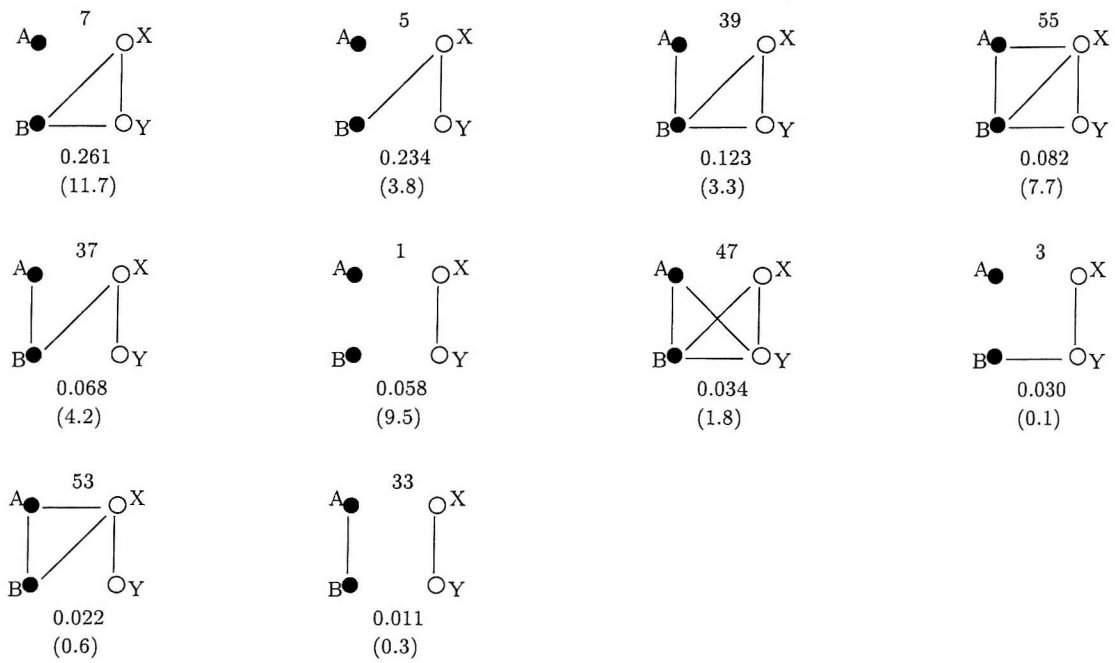


Table 6.3: Posterior model probabilities for rats data, graphical models. Standard errors $\times 10^4$ are in brackets.

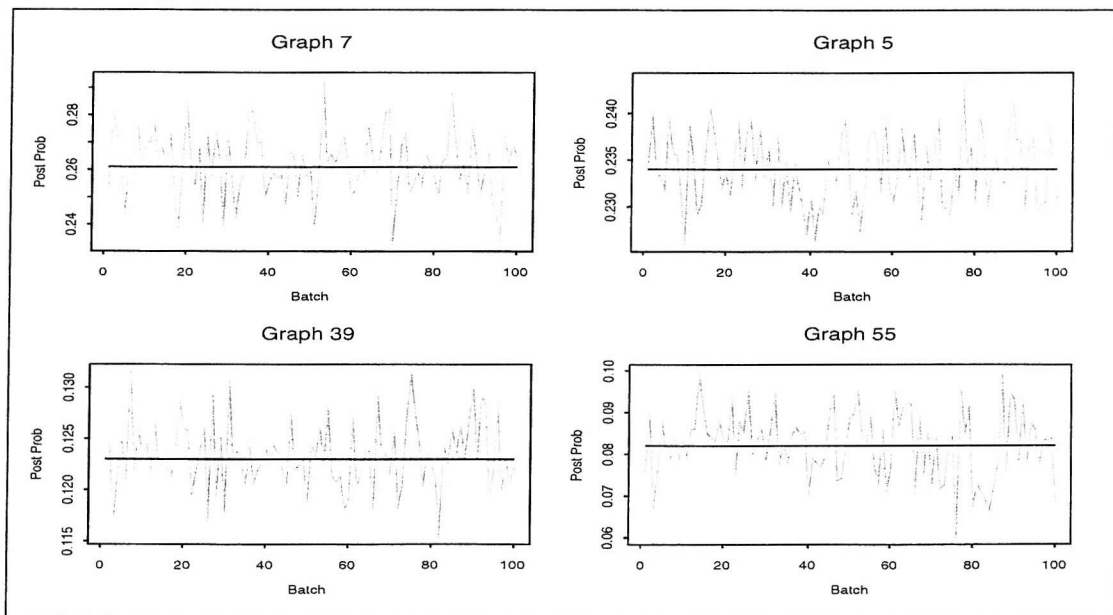


Figure 6.2: Batch Posterior Model Probabilities for rats data. The lines are the averages over the entire sample.

AB	AX	AY	BX	BY	XY
35.43	8.85	14.44	94.01	57.01	91.43

Table 6.4: Inclusion %ages for edges for rats data

6.4.3 College Smoking Habits

This example is from Wermuth and Lauritzen (1989) and concerns smoking habits and personality traits of 384 college students, and is treated in Whittaker (1990), who gives the cell counts, observed means and observed variances. There are two discrete variables, each with three levels: A is students' smoking status (smoker, quit, never smoked); B is parents' smoking habits (neither smoke, one smokes, both smoke). There are two continuous variables: X, trait anxiety, and Y, trait anger.

Both Wermuth and Lauritzen and Whittaker use chain models since B is expected to be an influence on the other variables but not conversely but undirected models may still be used. Whittaker uses two blocks, $\{B\}$ and $\{A, X, Y\}$, and suggests the model with directed edges from B to both A and X and undirected edges (X, Y) and (A, Y) . This graph is Markov equivalent to the undirected graph, 45, in Figure 6.3 and seems intuitively very reasonable. The graph suggested by Wermuth and Lauritzen is Markov Equivalent to the undirected graph, 41, in Figure 6.4 and does not contain the (B, X) edge. In MIM, 41 is also selected by forward inclusion and backward elimination selects 61, which has the additional (A, X) edge.

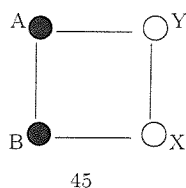
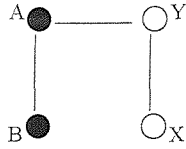


Figure 6.3: Undirected version of Whittaker's suggested graph for College Smoking Habits data

The results from the reversible jump sampler are consistent, assigning over three-quarters of the posterior probability to graph 41. This means a much more concentrated posterior than the previous examples but it is worth noting that the magnitudes of some of the parameters are very small (of the order of 10^{-3} , making proposal generation very difficult and this along with the greater number of cells leads to slow mixing within the model space, which is reflected in the



41

Figure 6.4: Undirected version of Wermuth and Lauritzen's graph for College Smoking Habits data

relatively high MCMC standard errors.

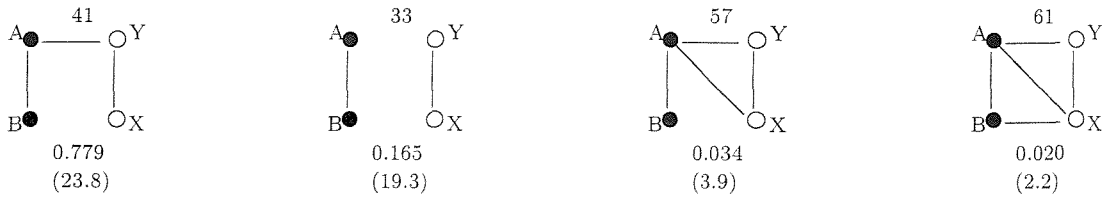


Table 6.5: Posterior model probabilities for Smoking Habits data, graphical models. Standard errors $\times 10^3$ are in brackets.

AB	AX	AY	BX	BY	XY
100.00	5.60	83.46	2.24	0.24	100.00

Table 6.6: Inclusion %ages for edges for Smoking Habits data

6.5 Discussion

These examples show that the methodology used for CGMs with a single discrete variable can successfully be extended to those with two and still produce reasonable results. The most challenging aspects of the models in this chapter are addition and removal of interactions involving the two discrete variables as well as consideration of the possible presence of these interactions when making other move types. Increasing the number of continuous variables is not nearly as challenging and is only limited by the prior as discussed in 4.6.

In principle, extension to three or more discrete variables is straightforward, although the models become considerably more complex and could well be impractical with four or more. The main difficulties are the higher order of discrete interaction, constraints corresponding to, for example, a missing $ABCXX$ interaction and the greater number of interactions whose presence/absence must be considered for some move types.

Run times are greater in general than those for the $p = 1$ case due to the additional complexity but were still reasonable for all examples examined with 100 000 never taking more than 30 minutes for the three examples presented. Parametric mixing (within models) is as good as the $p = 1$ case but again only when the data are centred within cells. Mixing across the model space for the reversible jump sampler is generally slower than for the $p = 1$ case and the chain is prone to “sticking” (not moving) for long periods but after several pilot runs and a sufficiently large number of iterations, coverage is usually satisfactory.

Restriction to reduced model classes, if desired, is no more difficult than for $p = 1$ and there is a further type of restriction within the class of graphical models possible now. This is to fix the (A, B) edge in or out. This can be useful when dealing with “designed” discrete data, where the cell counts are decided in advance. In addition, extension to the hierarchical models of Lauritzen (1996) is straightforward, as already noted.

A closely related type of model is also useful for dealing with such situations and avoids much of the difficulty involved in accounting for AB interaction. This type of model is a simple example of a graphical chain model (or block recursive as Lauritzen and Wermuth (1989) call them) with two blocks, one for the discrete variables and one for the continuous. The saturated model for $q = 2$ is shown in Figure 6.5

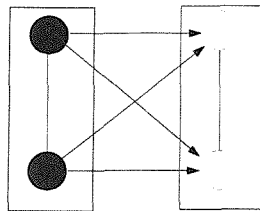


Figure 6.5: Saturated model in a simple class of graphical chain models

The fundamental difference with this type of model is that presence or absence of discrete-discrete edges (within the first block) has no effect on the conditional distribution of the continuous given the discrete. This block structure ensures that despite the discrete-continuous edges being directed, the corresponding graphical constraints are identical to those for the undirected case and of course

constraints corresponding to missing continuous-continuous edges are the same too.

This class of model, although not as flexible as the undirected versions and not that different in the case of $p = 2$, is much easier to deal with for greater values of p and is possibly more promising for extension to these higher dimensional cases.

Chapter 7

Summary and Further Work

7.1 Graphical Gaussian Models

Chapters 2 and 3 of this thesis have addressed the issue of nonconjugate Bayesian inference for general graphical Gaussian models and in particular inference under model uncertainty. For this purpose, a class of prior was introduced which is considerably more flexible than priors based on the Wishart distribution. This is because it avoids the restrictions of the Wishart distribution - common degrees of freedom for each diagonal entry, for example and allows individual components of the precision matrix to be dealt with separately, and independently with different marginal priors.

The major drawback of nonconjugate inference or, more precisely, of the use of priors not based on the Wishart distribution, is that the number of variables that can be dealt with is limited due to constraining of the space of (partial) correlation matrices. Section 2.3 showed that six variables is the practical upper limit for a jointly uniform prior on the partial correlations. Section 3.3 discussed how it may be possible to deal with more variables but only at the expense of having a more informative prior. Wishart-based priors avoid this problem as they are distributions over a space of positive definite matrices. Note that this is a restriction on any such non-Wishart-based prior not just those described here.

A possibility, which has not been investigated, to get around this problem is to ignore the contribution of the ratio of prior normalizing constants to the reversible jump acceptance probabilities as it very often close to 1.

The MCMC methods described here are not dependent on this prior and so provide a general framework, easily modified for virtually any prior. This thesis has, though only examined the use of a single prior class. An investigation of the effect of using other priors for the partial correlations may be useful. In particular a marginally uniform prior, as in Barnard et al. (2000), or the priors

of Wong, Carter and Kohn (2003) or Liechty, Liechty and Muller (2004). Another possibility, which may be worth investigating is the use of a Normal prior for the Fisher-z transformation of the partial correlations. However, whichever prior is used for the partial correlations, the positive definiteness constraint will always apply and so the limitations mentioned above will always apply.

Although it was possible in most of the examples presented to compare results with those of other authors, it may be useful in general to have another method of obtaining posterior model probabilities or Bayes factors to compare against. One possibility is the use of *bridge sampling* (Diciccio et al. 1997). This involves sampling from a “bridge distribution”, which should approximate the posterior but be easy to sample from. This was investigated briefly but it proved very difficult to find a suitable bridge distribution. In particular, a multivariate Normal distribution failed to produce correct results in examples where these were known or could be calculated directly.

The issue of prediction was investigated briefly but there is plenty of scope for further investigation including obtaining bivariate and multivariate predictive densities and a more thorough investigation of how much better prediction based on decomposable models is compared with that based on all models. An investigation into the effect of different priors may also be useful.

7.2 Mixed Graphical and CG Models

Chapters 4, 5 and 6 have attempted to address the so far neglected issue of Bayesian inference for mixed graphical models and for CG models in general. The approach taken is essentially an attempt to extend the methods and prior distributions for GGMs and this has been, to some extent at least, successful. The separate consideration of the case of models with a single discrete variable is a useful one as it is not only interesting in itself and more easily considered as an extension of GGMs, but also shows the way for much of what is needed for those models with more discrete variables.

What makes separate consideration of this case especially useful is the possibility of using the methods developed to handle latent variable models where a set of continuous variables are associated through an unobserved, or “hidden”, discrete variable. The joint distribution is thus CG. The unobserved discrete observations would be generated as part of the MCMC similarly to the classification rule described at the end of Chapter 5. Given this data, an MCMC scheme would then proceed in much the same way as for CGMs.

The priors used suffer from the same drawback as those for GGMs as they are

essentially the same but otherwise the flexibility and application are just as good. There may well be other sensible choices of prior for mixed models; those used here are merely convenient, in particular the priors on the quadratic parameters extends those for GGMs. A conjugate class of prior may prove especially useful and it may be possible to generalize the HIW or constrained Wishart priors for GGMs although it is unclear how this may be done.

The extension to the case of models with two discrete variables has been equally successful for fixed models but less successful for reversible jump and the reasons for this need further investigation. In particular, a more satisfactory method of generating proposals in order to improve mixing over the model space is required. Proposal generation would be much simpler if a full interaction expansion were used and simpler still if centring were not necessary. As discussed in 4.6, the use of quadratic interactions under model uncertainty is hampered by the positive definite requirement on the cell variances but if a suitable method for generating proposals could be found, they would be preferable for the purposes of reversible jump sampling as the model constraints are much simpler and between-model moves consist of either setting interactions to zero or inserting newly-generated proposal interactions.

In the absence of a satisfactory reversible jump algorithm for general CGMs, the simple chain graphs, mentioned at the end of Chapter 6 may prove to be more practical as they avoid the complication of the effect of missing discrete-discrete edges on the distribution of the continuous variables. This still leaves the discrete-continuous edges, however and much of the difficulty of these CG models comes from the constraints corresponding to absence of these edges.

Once the case of two discrete variables has been satisfactorily resolved, addition of further discrete variables is, in principle at least, straightforward, with the extra difficulties being mainly computational. However, it is likely that the models would start to get excessively complicated if there are more than three or four. In particular, the number of parameters quickly becomes very large. For this reason, a reduced class of models, such as the simple chain graphs or the generalised location-scale model, may be more practical in general.

The case of hierarchical CG models with two (or more) discrete variables is also incomplete but once graphical models are dealt with, these should pose no further difficulty, just as in the case of models with a single discrete variable.

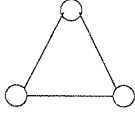
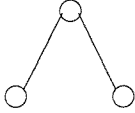
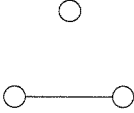
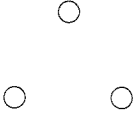
As for GGMs, the issue of prediction has been dealt with in a fairly limited way and there is scope for further investigation, especially for models with more than one discrete variable.

Appendix A

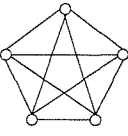
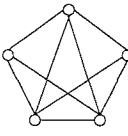
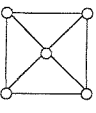
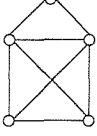
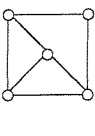
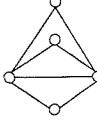
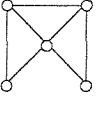
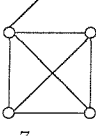
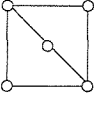
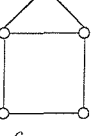
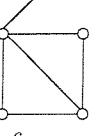
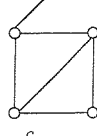
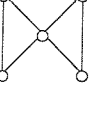
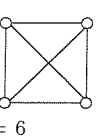
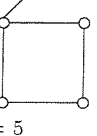
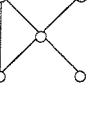
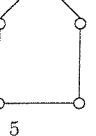
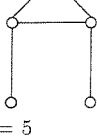
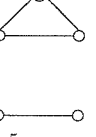
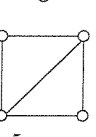
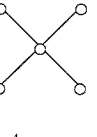
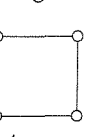
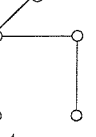
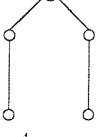

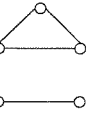

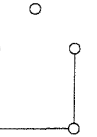
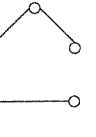
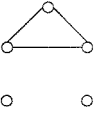
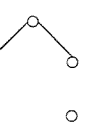
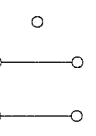
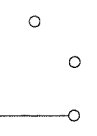
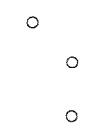
Graphs for GGM's

Here, tables of unlabelled directed pure graphs with up to six vertices are presented along with the following information for each: n , the number of edges; s , the degree sequence; m , the number of graphs with this structure, that is the number of distinct graphs obtainable by permuting vertices; d , the determinant measure used for indexing and c , the prior normalizing constant for the partial correlations in any graphical model with this unlabelled graph. The c 's are obtained from a rejection sampler as described in Section 2.3. Standard errors are not quoted but are typically 0.001 or less.

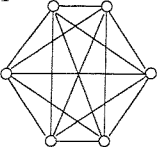
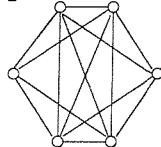
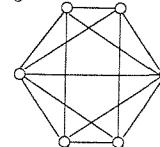
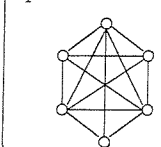
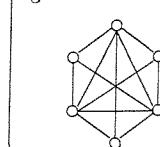
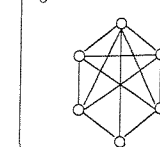
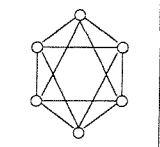
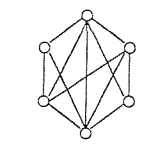
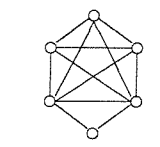
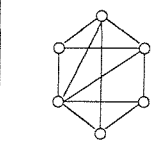
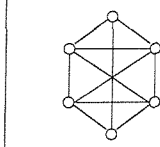
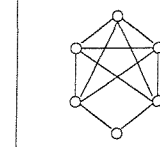
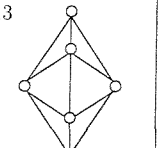
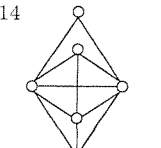
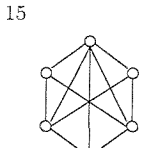
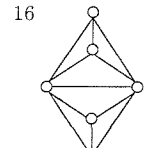
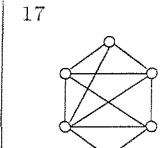
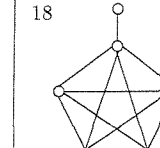
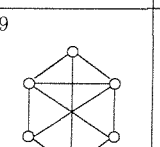
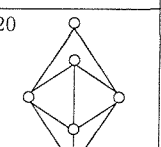
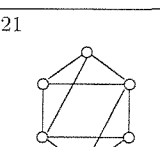
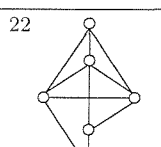
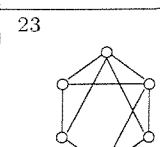
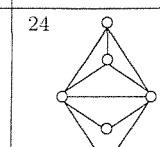
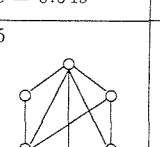
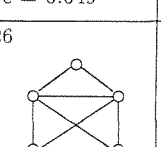
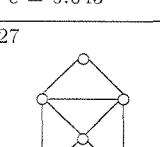
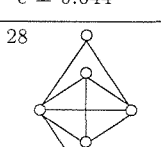
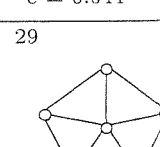
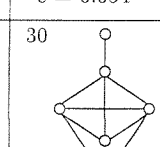
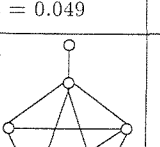
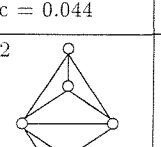
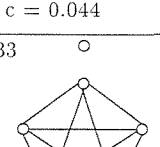
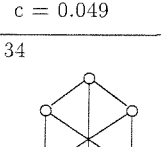
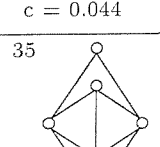
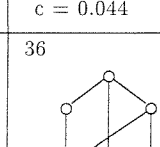
Unlabelled graphs with three vertices

1		2		3		4			
	$n = 3$ $s = (2,2,2)$ $m = 1$ $c = 0.2027$		$n = 2$ $s = (2,1,1)$ $m = 3$ $c = 0.3185$		$n = 1$ $s = (1,1,0)$ $m = 3$ $c = 0.5000$		$n = 0$ $s = (0,0,0)$ $m = 1$ $c = 1.000$		

Unlabelled graphs with five vertices

1  $n = 10$ $s = (4,4,4,4,4)$ $d = 6$ $m = 1$ $c = 0.044$	2  $n = 9$ $s = (4,4,4,3,3)$ $d = 4$ $m = 10$ $c = 0.052$	3  $n = 8$ $s = (4,3,3,3,3)$ $d = 0$ $m = 15$ $c = 0.061$	4  $n = 8$ $s = (5,5,4,4,2)$ $d = 4$ $m = 30$ $c = 0.062$	5  $n = 7$ $s = (3,3,3,3,2)$ $d = -4$ $m = 30$ $c = 0.072$	6  $n = 7$ $s = (4,4,2,2,2)$ $d = 0$ $m = 10$ $c = 0.082$
7  $n = 7$ $s = (4,3,3,2,2)$ $d = 4$ $m = 60$ $c = 0.073$	8  $n = 7$ $s = (4,3,3,3,1)$ $d = 6$ $m = 20$ $c = 0.072$	9  $n = 6$ $s = (3,3,2,2,2)$ $d = -16$ $m = 10$ $c = 0.097$	10  $n = 6$ $s = (3,3,2,2,2)$ $d = 0$ $m = 60$ $c = 0.085$	11  $n = 6$ $s = (4,3,2,2,1)$ $d = 4$ $m = 60$ $c = 0.097$	12  $n = 6$ $s = (3,3,3,2,1)$ $d = 4$ $m = 60$ $c = 0.086$
13  $n = 7$ $s = (4,2,2,2,1)$ $d = 6$ $m = 15$ $c = 0.082$	14  $n = 6$ $s = (3,3,3,3,0)$ $d = 10$ $m = 5$ $c = 0.085$	15  $n = 5$ $s = (3,2,2,2,1)$ $d = -4$ $m = 60$ $c = 0.114$	16  $n = 5$ $s = (4,2,2,1,1)$ $d = 4$ $m = 30$ $c = 0.129$	17  $n = 5$ $s = (2,2,2,2,2)$ $d = 4$ $m = 12$ $c = 0.102$	18  $n = 5$ $s = (3,3,2,1,1)$ $d = 6$ $m = 60$ $c = 0.114$
19  $n = 5$ $s = (3,2,2,2,1)$ $d = 6$ $m = 60$ $c = 0.097$	20  $n = 5$ $s = (3,3,2,2,0)$ $d = 8$ $m = 30$ $c = 0.114$	21  $n = 4$ $s = (4,1,1,1,1)$ $d = 0$ $m = 5$ $c = 0.203$	22  $n = 4$ $s = (2,2,2,2,0)$ $d = 0$ $m = 15$ $c = 0.152$	23  $n = 4$ $s = (3,2,1,1,1)$ $d = 4$ $m = 60$ $c = 0.152$	24  $n = 4$ $s = (2,2,2,1,1)$ $d = 6$ $m = 60$ $c = 0.129$
25  $n = 4$ $s = (3,2,2,1,0)$ $d = 10$ $m = 60$ $c = 0.152$	26  $n = 4$ $s = (2,2,2,1,1)$ $d = 12$ $m = 10$ $c = 0.101$	27  $n = 3$ $s = (3,1,1,1,0)$ $d = 8$ $m = 20$ $c = 0.239$	28  $n = 3$ $s = (2,2,1,1,0)$ $d = 10$ $m = 60$ $c = 0.203$	29  $n = 3$ $s = (2,1,1,1,1)$ $d = 12$ $m = 30$ $c = 0.159$	30  $n = 3$ $s = (2,2,2,0,0)$ $d = 16$ $m = 10$ $c = 0.203$
31  $n = 2$ $s = (2,1,1,0,0)$ $d = 16$ $m = 30$ $c = 0.319$	32  $n = 2$ $s = (1,1,1,1,0)$ $d = 18$ $m = 15$ $c = 0.250$	33  $n = 1$ $s = (1,1,0,0,0)$ $d = 24$ $m = 10$ $c = 0.500$	33  $n = 0$ $s = (0,0,0,0,0)$ $d = 32$ $m = 1$ $c = 1.000$		

Unlabelled graphs with six vertices

1  $n = 15$ $s = (5,5,5,5,5,5)$ $d = 7$ $m = 1$ $c = 0.032$	2  $n = 14$ $s = (5,5,5,5,4,4)$ $d = 4$ $m = 15$ $c = 0.035$	3  $n = 13$ $s = (5,5,4,4,4,4)$ $d = 0$ $m = 45$ $c = 0.036$	4  $n = 13$ $s = (5,5,5,4,4,3)$ $d = 3$ $m = 60$ $c = 0.042$	5  $n = 12$ $s = (5,5,5,3,3,3)$ $d = -4$ $m = 20$ $c = 0.042$	6  $n = 12$ $s = (5,4,4,4,4,3)$ $d = -4$ $m = 180$ $c = 0.039$
7  $n = 12$ $s = (4,4,4,4,4,4)$ $d = 0$ $m = 45$ $c = 0.039$	8  $n = 12$ $s = (5,5,4,4,3,3)$ $d = 3$ $m = 180$ $c = 0.039$	9  $n = 12$ $s = (5,5,4,4,4,2)$ $d = 4$ $m = 30$ $c = 0.039$	10  $n = 11$ $s = (5,4,4,3,3,3)$ $d = -16$ $m = 60$ $c = 0.046$	11  $n = 11$ $s = (4,4,4,4,3,3)$ $d = -13$ $m = 180$ $c = 0.041$	12  $n = 11$ $s = (4,4,4,4,4,2)$ $d = -8$ $m = 60$ $c = 0.041$
13  $n = 11$ $s = (4,4,4,4,3,3)$ $d = -4$ $m = 90$ $c = 0.0415$	14  $n = 11$ $s = (5,5,4,3,3,2)$ $d = 0$ $m = 180$ $c = 0.0455$	15  $n = 11$ $s = (5,4,4,3,3,3)$ $d = 0$ $m = 360$ $c = 0.042$	16  $n = 11$ $s = (5,5,3,3,3,3)$ $d = 3$ $m = 45$ $c = 0.041$	17  $n = 11$ $s = (5,4,4,4,3,2)$ $d = 3$ $m = 360$ $c = 0.042$	18  $n = 11$ $s = (5,4,4,4,4,1)$ $d = 7$ $m = 30$ $c = 0.042$
19  $n = 10$ $s = (4,4,3,3,3,3)$ $d = -36$ $m = 60$ $c = 0.049$	20  $n = 10$ $s = (4,4,4,3,3,2)$ $d = -16$ $m = 180$ $c = 0.049$	21  $n = 10$ $s = (4,4,3,3,3,3)$ $d = -12$ $m = 45$ $c = 0.043$	22  $n = 10$ $s = (4,4,4,3,3,2)$ $d = -5$ $m = 360$ $c = 0.044$	23  $n = 10$ $s = (4,4,3,3,3,3)$ $d = -5$ $m = 360$ $c = 0.044$	24  $n = 10$ $s = (5,5,3,3,2,2)$ $d = -4$ $m = 90$ $c = 0.054$
25  $n = 10$ $s = (5,4,3,3,3,2)$ $d = -4$ $m = 360$ $c = 0.049$	26  $n = 10$ $s = (4,4,4,4,2,2)$ $d = 0$ $m = 90$ $c = 0.044$	27  $n = 10$ $s = (4,4,4,3,3,2)$ $d = 0$ $m = 360$ $c = 0.044$	28  $n = 10$ $s = (5,4,4,3,2,2)$ $d = 3$ $m = 360$ $c = 0.049$	29  $n = 10$ $s = (5,3,3,3,3,3)$ $d = 3$ $m = 72$ $c = 0.044$	30  $n = 10$ $s = (4,4,4,4,3,1)$ $d = 3$ $m = 120$ $c = 0.044$
31  $n = 10$ $s = (5,4,4,3,3,1)$ $d = 4$ $m = 180$ $c = 0.049$	32  $n = 10$ $s = (5,4,3,3,3,2)$ $d = 4$ $m = 360$ $c = 0.043$	33  $n = 10$ $s = (4,4,4,4,4,0)$ $d = 12$ $m = 6$ $c = 0.044$	34  $n = 9$ $s = (3,3,3,3,3,3)$ $d = -80$ $m = 10$ $c = 0.058$	35  $n = 9$ $s = (4,4,3,3,2,2)$ $d = 3$ $m = 90$ $c = 0.058$	36  $n = 9$ $s = (4,3,3,3,3,2)$ $d = -20$ $m = 360$ $c = 0.047$

continued overleaf

Unlabelled graphs with six vertices (continued)

<p>37</p> <p>$n = 9$ $s = (5,5,2,2,2,2)$ $d = -16$ $m = 15$ $c = 0.072$</p>	<p>38</p> <p>$n = 9$ $s = (4,4,3,3,2,2)$ $d = -16$ $m = 180$ $c = 0.052$</p>	<p>39</p> <p>$n = 9$ $s = (4,4,3,3,2,2)$ $d = -5$ $m = 720$ $c = 0.052$</p>	<p>40</p> <p>$n = 9$ $s = (4,3,3,3,3,2)$ $d = -5$ $m = 360$ $c = 0.046$</p>	<p>41</p> <p>$n = 9$ $s = (4,4,3,3,3,1)$ $d = -4$ $m = 360$ $c = 0.052$</p>	<p>42</p> <p>$n = 9$ $s = (5,4,3,2,2,2)$ $d = 0$ $m = 360$ $c = 0.058$</p>
<p>43</p> <p>$n = 9$ $s = (5,3,3,3,3,1)$ $d = 0$ $m = 90$ $c = 0.058$</p>	<p>44</p> <p>$n = 9$ $s = (4,4,3,3,2,2)$ $d = 0$ $m = 360$ $c = 0.046$</p>	<p>45</p> <p>$n = 9$ $s = (4,3,3,3,3,2)$ $d = 0$ $m = 360$ $c = 0.047$</p>	<p>46</p> <p>$n = 9$ $s = (3,3,3,3,3,3)$ $d = 0$ $m = 60$ $c = 0.047$</p>	<p>47</p> <p>$n = 9$ $s = (5,4,3,3,2,1)$ $d = 3$ $m = 360$ $c = 0.058$</p>	<p>48</p> <p>$n = 9$ $s = (5,3,3,3,2,2)$ $d = 3$ $m = 360$ $c = 0.051$</p>
<p>49</p> <p>$n = 9$ $s = (4,4,3,3,3,1)$ $d = 3$ $m = 180$ $c = 0.046$</p>	<p>50</p> <p>$n = 9$ $s = (4,4,3,3,2,2)$ $d = 3$ $m = 180$ $c = 0.046$</p>	<p>51</p> <p>$n = 9$ $s = (4,4,4,3,2,1)$ $d = 4$ $m = 360$ $c = 0.052$</p>	<p>52</p> <p>$n = 9$ $s = (4,4,4,2,2,2)$ $d = 4$ $m = 120$ $c = 0.052$</p>	<p>53</p> <p>$n = 9$ $s = (5,3,3,3,2,2)$ $d = 7$ $m = 60$ $c = 0.043$</p>	<p>54</p> <p>$n = 9$ $s = (4,4,4,3,3,0)$ $d = 8$ $m = 60$ $c = 0.052$</p>
<p>55</p> <p>$n = 8$ $s = (4,4,2,2,2,2)$ $d = -64$ $m = 15$ $c = 0.077$</p>	<p>56</p> <p>$n = 8$ $s = (3,3,3,3,2,2)$ $d = -48$ $m = 90$ $c = 0.062$</p>	<p>57</p> <p>$n = 8$ $s = (4,3,3,2,2,2)$ $d = -20$ $m = 360$ $c = 0.061$</p>	<p>58</p> <p>$n = 8$ $s = (4,3,3,3,2,1)$ $d = -13$ $m = 360$ $c = 0.062$</p>	<p>59</p> <p>$n = 8$ $s = (3,3,3,3,3,1)$ $d = -12$ $m = 180$ $c = 0.054$</p>	<p>60</p> <p>$n = 8$ $s = (4,3,3,3,2,1)$ $d = -8$ $m = 360$ $c = 0.061$</p>
<p>61</p> <p>$n = 8$ $s = (4,3,3,2,2,2)$ $d = -5$ $m = 720$ $c = 0.054$</p>	<p>62</p> <p>$n = 8$ $s = (3,3,3,3,2,2)$ $d = -5$ $m = 360$ $c = 0.056$</p>	<p>63</p> <p>$n = 8$ $s = (5,4,2,2,2,1)$ $d = -4$ $m = 120$ $c = 0.077$</p>	<p>64</p> <p>$n = 8$ $s = (4,4,3,2,2,1)$ $d = -4$ $m = 180$ $c = 0.061$</p>	<p>65</p> <p>$n = 8$ $s = (4,4,2,2,2,2)$ $d = 0$ $m = 90$ $c = 0.061$</p>	<p>66</p> <p>$n = 8$ $s = (4,3,3,3,3,0)$ $d = 0$ $m = 90$ $c = 0.061$</p>
<p>67</p> <p>$n = 8$ $s = (4,3,3,2,2,2)$ $d = 0$ $m = 360$ $c = 0.054$</p>	<p>68</p> <p>$n = 8$ $s = (3,3,2,2,2,2)$ $d = 0$ $m = 180$ $c = 0.048$</p>	<p>69</p> <p>$n = 8$ $s = (5,3,3,2,2,1)$ $d = 3$ $m = 360$ $c = 0.068$</p>	<p>70</p> <p>$n = 8$ $s = (4,4,3,2,2,1)$ $d = 3$ $m = 720$ $c = 0.062$</p>	<p>71</p> <p>$n = 8$ $s = (3,3,3,3,2,2)$ $d = 3$ $m = 180$ $c = 0.050$</p>	<p>72</p> <p>$n = 8$ $s = (5,3,3,3,1,1)$ $d = 4$ $m = 60$ $c = 0.068$</p>

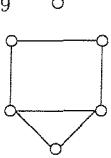
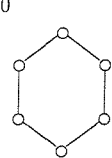
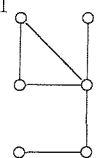
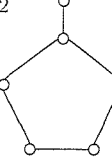
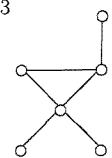
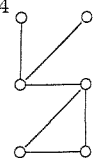
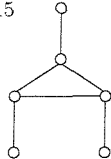
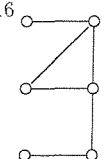
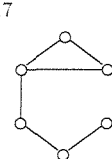
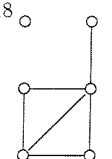
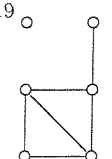
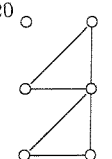
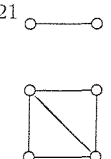
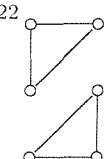
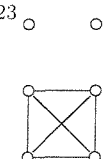
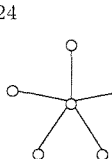
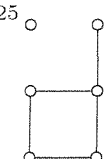
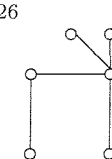
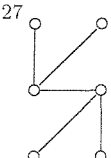
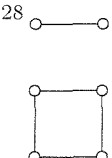
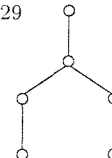
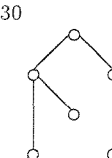
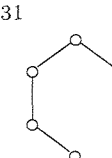
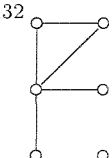
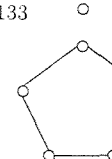
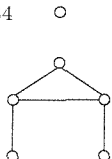
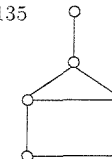
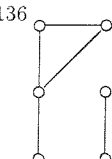
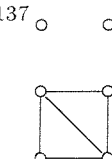
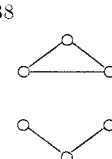
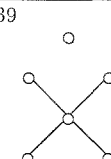
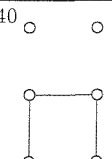
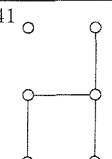
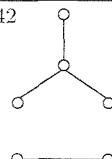
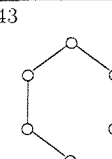
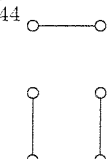
continued overleaf

Unlabelled graphs with six vertices (continued)

73 $n = 8$ $s = (5, 3, 2, 2, 2, 2)$ $d = 4$ $m = 180$ $c = 0.058$	74 $n = 8$ $s = (4, 3, 3, 3, 2, 1)$ $d = 4$ $m = 720$ $c = 0.054$	75 $n = 8$ $s = (4, 3, 3, 2, 2, 2)$ $d = 7$ $m = 180$ $c = 0.046$	76 $n = 8$ $s = (4, 4, 3, 3, 1, 1)$ $d = 4$ $m = 180$ $c = 0.061$	77 $n = 8$ $s = (4, 3, 3, 3, 2, 1)$ $d = 7$ $m = 120$ $c = 0.046$	78 $n = 8$ $s = (4, 4, 3, 3, 2, 0)$ $d = 8$ $m = 180$ $c = 0.044$
79 $n = 7$ $s = (4, 3, 2, 2, 2, 1)$ $d = -36$ $m = 120$ $c = 0.082$	80 $n = 7$ $s = (3, 3, 3, 2, 2, 1)$ $d = -32$ $m = 180$ $c = 0.072$	81 $n = 7$ $s = (3, 3, 2, 2, 2, 2)$ $d = -21$ $m = 180$ $c = 0.064$	82 $n = 7$ $s = (4, 2, 2, 2, 2, 2)$ $d = -8$ $m = 180$ $c = 0.061$	83 $n = 7$ $s = (3, 3, 3, 3, 2, 0)$ $d = -8$ $m = 180$ $c = 0.072$	84 $n = 7$ $s = (4, 3, 2, 2, 2, 1)$ $d = -5$ $m = 720$ $c = 0.073$
85 $n = 7$ $s = (3, 3, 3, 2, 2, 1)$ $d = -5$ $m = 720$ $c = 0.064$	86 $n = 7$ $s = (3, 3, 2, 2, 2, 2)$ $d = -4$ $m = 180$ $c = 0.066$	87 $n = 7$ $s = (5, 3, 2, 2, 1, 1)$ $d = 0$ $m = 180$ $c = 0.091$	88 $n = 7$ $s = (4, 4, 2, 2, 2, 0)$ $d = 0$ $m = 60$ $c = 0.082$	89 $n = 7$ $s = (4, 3, 3, 2, 1, 1)$ $d = 0$ $m = 180$ $c = 0.072$	90 $n = 7$ $s = (3, 3, 3, 2, 2, 1)$ $d = 0$ $m = 360$ $c = 0.064$
91 $n = 7$ $s = (5, 2, 2, 2, 2, 1)$ $d = 3$ $m = 90$ $c = 0.077$	92 $n = 7$ $s = (4, 3, 3, 2, 1, 1)$ $d = 3$ $m = 720$ $c = 0.072$	93 $n = 7$ $s = (3, 3, 3, 3, 1, 1)$ $d = 3$ $m = 180$ $c = 0.064$	94 $n = 7$ $s = (4, 4, 2, 2, 1, 1)$ $d = 4$ $m = 180$ $c = 0.082$	95 $n = 7$ $s = (4, 3, 2, 2, 2, 1)$ $d = 4$ $m = 360$ $c = 0.062$	96 $n = 7$ $s = (3, 3, 3, 2, 2, 1)$ $d = 4$ $m = 360$ $c = 0.054$
97 $n = 7$ $s = (3, 3, 2, 2, 2, 2)$ $d = 4$ $m = 360$ $c = 0.057$	98 $n = 7$ $s = (4, 3, 2, 2, 2, 1)$ $d = 7$ $m = 360$ $c = 0.062$	99 $n = 7$ $s = (3, 3, 2, 2, 2, 2)$ $d = 7$ $m = 90$ $c = 0.046$	100 $n = 7$ $s = (4, 3, 3, 2, 2, 0)$ $d = 8$ $m = 360$ $c = 0.073$	101 $n = 7$ $s = (4, 3, 3, 3, 1, 0)$ $d = 12$ $m = 120$ $c = 0.073$	102 $n = 7$ $s = (3, 3, 3, 3, 1, 1)$ $d = 15$ $m = 15$ $c = 0.043$
103 $n = 6$ $s = (3, 3, 2, 2, 2, 0)$ $d = -32$ $m = 60$ $c = 0.097$	104 $n = 6$ $s = (4, 2, 2, 2, 1, 1)$ $d = -16$ $m = 180$ $c = 0.097$	105 $n = 6$ $s = (3, 3, 2, 2, 1, 1)$ $d = -13$ $m = 180$ $c = 0.085$	106 $n = 6$ $s = (3, 3, 2, 2, 1, 1)$ $d = -12$ $m = 360$ $c = 0.085$	107 $n = 6$ $s = (3, 2, 2, 2, 2, 1)$ $d = -8$ $m = 360$ $c = 0.073$	108 $n = 6$ $s = (5, 2, 2, 1, 1, 1)$ $d = -4$ $m = 60$ $c = 0.121$

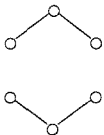
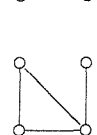
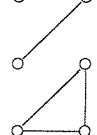
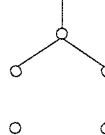
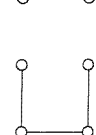
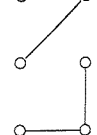
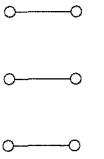
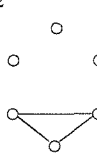
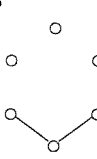

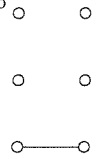
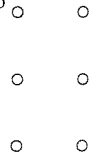
continued overleaf

Unlabelled graphs with six vertices (continued)

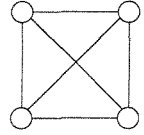
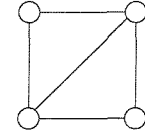
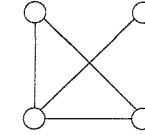
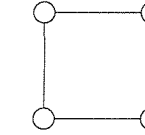
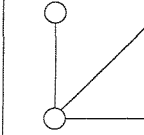
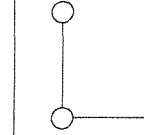
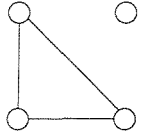
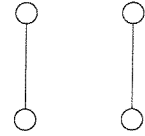
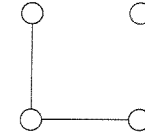
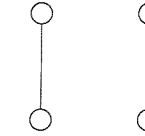
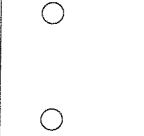
<p>109 </p> <p>$n = 6$ $s = (3,3,2,2,2,0)$ $d = 0$ $m = 360$ $c = 0.085$</p>	<p>110 </p> <p>$n = 6$ $s = (2,2,2,2,2,2)$ $d = 0$ $m = 60$ $c = 0.066$</p>	<p>111 </p> <p>$n = 6$ $s = (4,2,2,2,1,1)$ $d = 3$ $m = 360$ $c = 0.082$</p>	<p>112 </p> <p>$n = 6$ $s = (3,2,2,2,2,1)$ $d = 3$ $m = 360$ $c = 0.076$</p>	<p>113 </p> <p>$n = 6$ $s = (4,3,2,1,1,1)$ $d = 4$ $m = 360$ $c = 0.097$</p>	<p>114 </p> <p>$n = 6$ $s = (3,3,2,2,1,1)$ $d = 4$ $m = 180$ $c = 0.073$</p>
<p>115 </p> <p>$n = 6$ $s = (3,3,3,1,1,1)$ $d = 7$ $m = 120$ $c = 0.086$</p>	<p>116 </p> <p>$n = 6$ $s = (3,3,2,2,1,1)$ $d = 7$ $m = 720$ $c = 0.073$</p>	<p>117 </p> <p>$n = 6$ $s = (3,2,2,2,2,1)$ $d = 7$ $m = 360$ $c = 0.062$</p>	<p>118 </p> <p>$n = 6$ $s = (4,3,2,2,1,0)$ $d = 8$ $m = 360$ $c = 0.097$</p>	<p>119 </p> <p>$n = 6$ $s = (3,3,3,2,1,0)$ $d = 8$ $m = 360$ $c = 0.086$</p>	<p>120 </p> <p>$n = 6$ $s = (4,2,2,2,2,0)$ $d = 12$ $m = 90$ $c = 0.082$</p>
<p>121 </p> <p>$n = 6$ $s = (3,3,2,2,1,1)$ $d = 12$ $m = 90$ $c = 0.057$</p>	<p>122 </p> <p>$n = 6$ $s = (2,2,2,2,2,2)$ $d = 16$ $m = 10$ $c = 0.041$</p>	<p>123 </p> <p>$n = 6$ $s = (3,3,3,3,0,0)$ $d = 20$ $m = 15$ $c = 0.085$</p>	<p>124 </p> <p>$n = 5$ $s = (5,1,1,1,1,1)$ $d = -16$ $m = 6$ $c = 0.189$</p>	<p>125 </p> <p>$n = 5$ $s = (3,2,2,2,1,0)$ $d = -8$ $m = 360$ $c = 0.114$</p>	<p>126 </p> <p>$n = 5$ $s = (4,2,1,1,1,1)$ $d = -4$ $m = 120$ $c = 0.129$</p>
<p>127 </p> <p>$n = 5$ $s = (3,3,1,1,1,1)$ $d = 0$ $m = 90$ $c = 0.114$</p>	<p>128 </p> <p>$n = 5$ $s = (2,2,2,2,1,1)$ $d = 0$ $m = 45$ $c = 0.076$</p>	<p>129 </p> <p>$n = 5$ $s = (3,2,2,1,1,1)$ $d = 3$ $m = 360$ $c = 0.097$</p>	<p>130 </p> <p>$n = 5$ $s = (3,2,2,1,1,1)$ $d = 4$ $m = 360$ $c = 0.097$</p>	<p>131 </p> <p>$n = 5$ $s = (2,2,2,2,1,1)$ $d = 7$ $m = 360$ $c = 0.082$</p>	<p>132 </p> <p>$n = 5$ $s = (4,2,2,1,1,0)$ $d = 8$ $m = 180$ $c = 0.129$</p>
<p>133 </p> <p>$n = 5$ $s = (2,2,2,2,2,0)$ $d = 8$ $m = 72$ $c = 0.101$</p>	<p>134 </p> <p>$n = 5$ $s = (3,3,2,1,1,0)$ $d = 12$ $m = 360$ $c = 0.114$</p>	<p>135 </p> <p>$n = 5$ $s = (3,2,2,2,1,0)$ $d = 12$ $m = 360$ $c = 0.097$</p>	<p>136 </p> <p>$n = 5$ $s = (3,2,2,1,1,1)$ $d = 15$ $m = 180$ $c = 0.076$</p>	<p>137 </p> <p>$n = 5$ $s = (3,3,2,2,0,0)$ $d = 16$ $m = 90$ $c = 0.114$</p>	<p>138 </p> <p>$n = 5$ $s = (2,2,2,2,1,1)$ $d = 16$ $m = 60$ $c = 0.065$</p>
<p>139 </p> <p>$n = 4$ $s = (4,1,1,1,1,0)$ $d = 0$ $m = 30$ $c = 0.203$</p>	<p>140 </p> <p>$n = 4$ $s = (2,2,2,2,0,0)$ $d = 0$ $m = 45$ $c = 0.152$</p>	<p>141 </p> <p>$n = 4$ $s = (3,2,1,1,1,0)$ $d = 8$ $m = 360$ $c = 0.152$</p>	<p>142 </p> <p>$n = 4$ $s = (3,1,1,1,1,1)$ $d = 12$ $m = 60$ $c = 0.119$</p>	<p>143 </p> <p>$n = 4$ $s = (2,2,2,1,1,0)$ $d = 12$ $m = 360$ $c = 0.129$</p>	<p>144 </p> <p>$n = 4$ $s = (2,2,1,1,1,1)$ $d = 15$ $m = 90$ $c = 0.101$</p>

continued overleaf

Unlabelled graphs with six vertices (continued)

145  $n = 4$ $s = (2,2,1,1,1,1)$ $d = 16$ $m = 180$ $c = 0.101$	146  $n = 4$ $s = (3,2,2,1,0,0)$ $d = 20$ $m = 180$ $c = 0.152$	147  $n = 4$ $s = (2,2,2,1,1,0)$ $d = 24$ $m = 60$ $c = 0.101$	148  $n = 3$ $s = (3,1,1,1,0,0)$ $d = 16$ $m = 30$ $c = 0.239$	149  $n = 3$ $s = (2,2,1,1,0,0)$ $d = 20$ $m = 180$ $c = 0.203$	150  $n = 3$ $s = (2,1,1,1,1,0)$ $d = 24$ $m = 180$ $c = 0.160$
151  $n = 3$ $s = (1,1,1,1,1,1)$ $d = 27$ $m = 45$ $c = 0.125$	152  $n = 3$ $s = (2,2,2,0,0,0)$ $d = 32$ $m = 20$ $c = 0.203$	153  $n = 2$ $s = (2,1,1,0,0,0)$ $d = 32$ $m = 60$ $c = 0.318$	154  $n = 2$ $s = (1,1,1,1,0,0)$ $d = 36$ $m = 45$ $c = 0.250$	155  $n = 1$ $s = (1,1,0,0,0,0)$ $d = 48$ $m = 15$ $c = 0.500$	156  $n = 0$ $s = (0,0,0,0,0,0)$ $d = 64$ $m = 1$ $c = 1.000$

Unlabelled graphs with four vertices

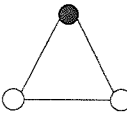
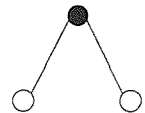
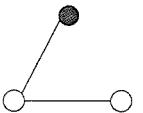
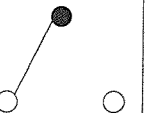
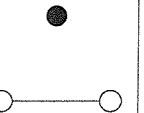
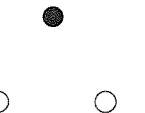
1  $n = 6$ $s = (3,3,3,3)$ $d = 6$ $m = 1$ $c = 0.085$	2  $n = 5$ $s = (3,3,2,2)$ $d = 4$ $m = 6$ $c = 0.114$	3  $n = 4$ $s = (3,2,2,1)$ $d = 5$ $m = 12$ $c = 0.152$	4  $n = 4$ $s = (2,2,2,2)$ $d = 0$ $m = 3$ $c = 0.152$	5  $n = 3$ $s = (3,1,1,1)$ $d = 4$ $m = 4$ $c = 0.239$	6  $n = 3$ $s = (2,2,1,1)$ $d = 5$ $m = 12$ $c = 0.203$
7  $n = 3$ $s = (2,2,2,0)$ $d = 8$ $m = 4$ $c = 0.203$	8  $n = 2$ $s = (1,1,1,1)$ $d = 9$ $m = 3$ $c = 0.250$	9  $n = 2$ $s = (2,1,1,0)$ $d = 8$ $m = 12$ $c = 0.319$	10  $n = 1$ $s = (1,1,0,0)$ $d = 12$ $m = 6$ $c = 0.500$	11  $n = 0$ $s = (0,0,0,0)$ $d = 16$ $m = 1$ $c = 1.000$	

Appendix B

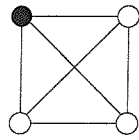
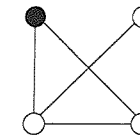
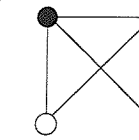
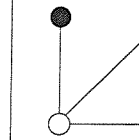
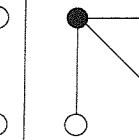
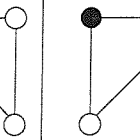
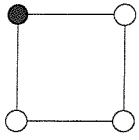
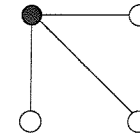
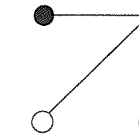
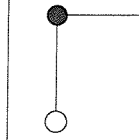
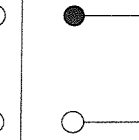
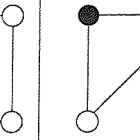
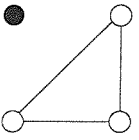
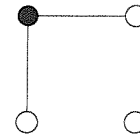
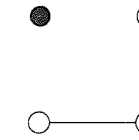
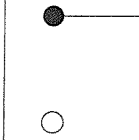
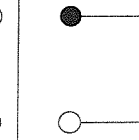
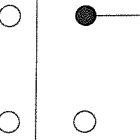
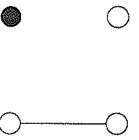
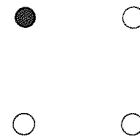
Marked Graphs for mixed graphical models

Here, tables of unlabelled undirected marked graphs containing one discrete vertex and up to four continuous are presented as well as those containing two discrete variables and two continuous. The number of edges, n and the number of distinct permutations of the vertices, m are also given for each graph. The jointly uniform priors (which are constants), described and used in this thesis, for the partial correlations in mixed graphical models with these graphs are given as c_l (for one discrete variable) where l is the number of levels of the discrete variable. In the final table they are given for both discrete variables having two levels, for one with two and one with three and for both with three. Standard errors are not quoted but are typically 0.001 or less.

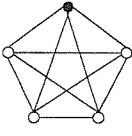
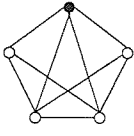
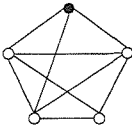
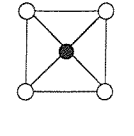
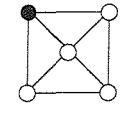
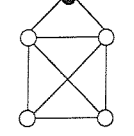
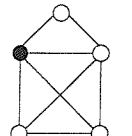
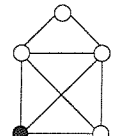
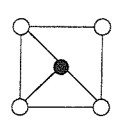
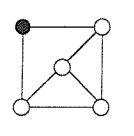
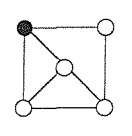
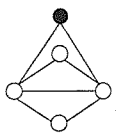
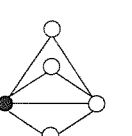
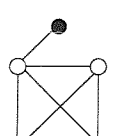
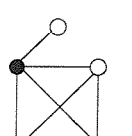
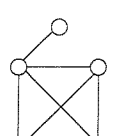
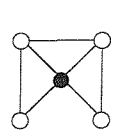
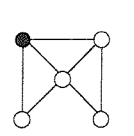
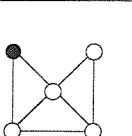
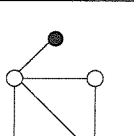
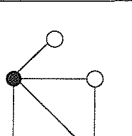
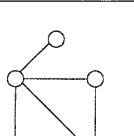
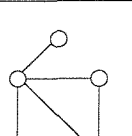
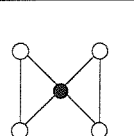
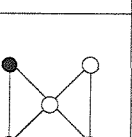
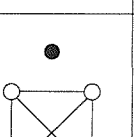
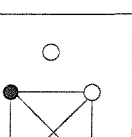
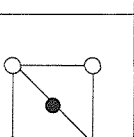
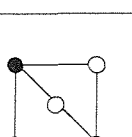
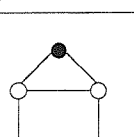
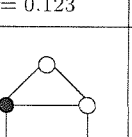
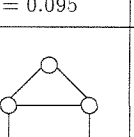
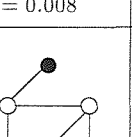
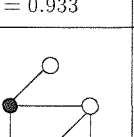
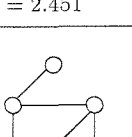
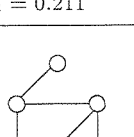
Marked Graphs with $p=1, q=2$

	2	3	4	5	6
					
$n = 3$ $m = 1$ $c_2 = 0.250$ $c_3 = 0.125$ $c_4 = 0.063$	$n = 2$ $m = 1$ $c = 1.00$	$n = 2$ $m = 2$ $c_2 = 0.850$ $c_3 = 1.160$ $c_4 = 1.430$	$n = 1$ $m = 2$ $c = 1.00$	$n = 1$ $m = 1$ $c = 0.500$	$n = 0$ $m = 1$ $c = 1.00$

Marked Graphs with $p=1, q=3$

1  $n = 6$ $m = 1$ $c_2 = 0.33$ $c_3 = 0.01$	2  $n = 5$ $m = 3$ $c_2 = 0.61$ $c_3 = 0.29$	3  $n = 5$ $m = 3$ $c_2 = 0.41$ $c_3 = 0.03$	4  $n = 4$ $m = 3$ $c_2 = 0.38$ $c_3 = 0.53$	5  $n = 4$ $m = 3$ $c_2 = 0.5$ $c_3 = 0.125$	6  $n = 4$ $m = 6$ $c_2 = 0.55$ $c_3 = 0.19$
7  $n = 4$ $m = 3$ $c_2 = 0.88$ $c_3 = 1.61$	8  $n = 3$ $m = 1$ $c_2 = 1.00$ $c_3 = 1.00$	9  $n = 3$ $m = 3$ $c_2 = 0.58$ $c_3 = 0.83$	10  $n = 3$ $m = 6$ $c_2 = 0.85$ $c_3 = 1.16$	11  $n = 3$ $m = 6$ $c_2 = 0.54$ $c_3 = 0.74$	12  $n = 3$ $m = 3$ $c_2 = 0.50$ $c_3 = 0.125$
13  $n = 3$ $m = 1$ $c_2 = 0.20$ $c_3 = 0.20$	14  $n = 2$ $m = 3$ $c_2 = 1.00$ $c_3 = 1.00$	15  $n = 2$ $m = 3$ $c_2 = 0.32$ $c_3 = 0.32$	16  $n = 2$ $m = 6$ $c_2 = 0.85$ $c_3 = 1.16$	17  $n = 2$ $m = 3$ $c_2 = 0.50$ $c_3 = 0.50$	18  $n = 1$ $m = 3$ $c_2 = 1.00$ $c_3 = 1.00$
19  $n = 1$ $m = 3$ $c_2 = 0.50$ $c_3 = 0.50$	20  $n = 0$ $m = 1$ $c_2 = 1.00$ $c_3 = 1.00$				

Marked Graphs with $p=1, q=4$

1  $n = 10$ $m = 1$ $c_2 = 0.007$ $c_3 = 0.001$	2  $n = 9$ $m = 6$ $c_2 = 0.013$ $c_3 = 0.001$	3  $n = 9$ $m = 4$ $c_2 = 0.0979$ $c_3 = 0.024$	4  $n = 8$ $m = 3$ $c_2 = 0.023$ $c_3 = 0.004$	5  $n = 8$ $m = 12$ $c_2 = 0.186$ $c_3 = 0.150$	6  $n = 8$ $m = 6$ $c_2 = 0.137$ $c_3 = 0.122$
7  $n = 8$ $m = 12$ $c_2 = 0.023$ $c_3 = 0.004$	8  $n = 8$ $m = 12$ $c_2 = 0.072$ $c_3 = 0.035$	9  $n = 7$ $m = 12$ $c_2 = 0.128$ $c_3 = 0.074$	10  $n = 7$ $m = 6$ $c_2 = 0.375$ $c_3 = 0.893$	11  $n = 7$ $m = 12$ $c_2 = 0.364$ $c_3 = 0.460$	12  $n = 7$ $m = 6$ $c_2 = 0.206$ $c_3 = 0.182$
13  $n = 7$ $m = 4$ $c_2 = 0.056$ $c_3 = 0.014$	14  $n = 7$ $m = 4$ $c_2 = 0.165$ $c_3 = 0.226$	15  $n = 7$ $m = 4$ $c_2 = 0.042$ $c_3 = 0.008$	16  $n = 7$ $m = 12$ $c_2 = 0.054$ $c_3 = 0.017$	17  $n = 7$ $m = 12$ $c_2 = 0.041$ $c_3 = 0.008$	18  $n = 7$ $m = 24$ $c_2 = 0.152$ $c_3 = 0.100$
19  $n = 7$ $m = 24$ $c_2 = 0.183$ $c_3 = 0.145$	20  $n = 6$ $m = 12$ $c_2 = 0.236$ $c_3 = 0.381$	21  $n = 6$ $m = 24$ $c_2 = 0.101$ $c_3 = 0.032$	22  $n = 6$ $m = 12$ $c_2 = 0.255$ $c_3 = 0.290$	23  $n = 6$ $m = 12$ $c_2 = 0.134$ $c_3 = 0.054$	24  $n = 6$ $m = 3$ $c_2 = 0.063$ $c_3 = 0.016$
25  $n = 6$ $m = 12$ $c_2 = 0.145$ $c_3 = 0.123$	26  $n = 6$ $m = 1$ $c_2 = 0.086$ $c_3 = 0.095$	27  $n = 6$ $m = 4$ $c_2 = 0.041$ $c_3 = 0.008$	28  $n = 6$ $m = 6$ $c_2 = 0.484$ $c_3 = 0.933$	29  $n = 6$ $m = 4$ $c_2 = 1.047$ $c_3 = 2.451$	30  $n = 6$ $m = 2$ $c_2 = 0.225$ $c_3 = 0.211$
31  $n = 6$ $m = 24$ $c_2 = 0.288$ $c_3 = 0.230$	32  $n = 6$ $m = 24$ $c_2 = 0.457$ $c_3 = 0.856$	33  $n = 6$ $m = 12$ $c_2 = 0.862$ $c_3 = 1.190$	34  $n = 6$ $m = 12$ $c_2 = 0.302$ $c_3 = 0.282$	35  $n = 6$ $m = 24$ $c_2 = 0.112$ $c_3 = 0.051$	36  $n = 6$ $m = 12$ $c_2 = 0.251$ $c_3 = 0.300$

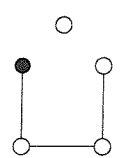
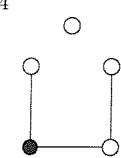
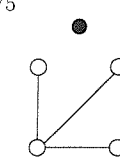
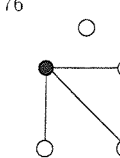
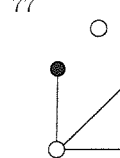
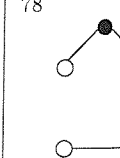
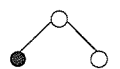
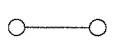
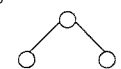

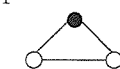

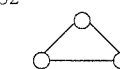

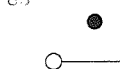
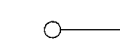
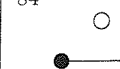
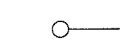
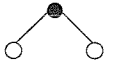

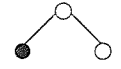

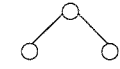


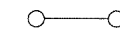


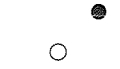

continued overleaf

Marked Graphs with $p=1, q=4$ (continued)

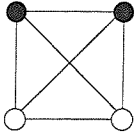
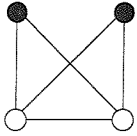
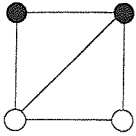
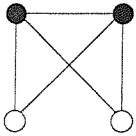
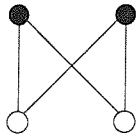
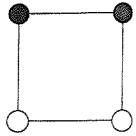
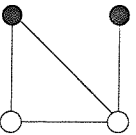
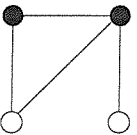
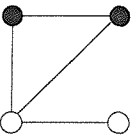
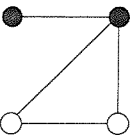
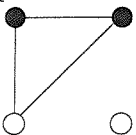
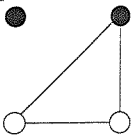
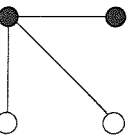
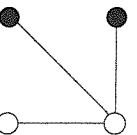
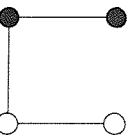
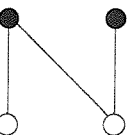
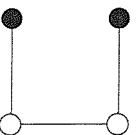
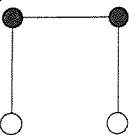
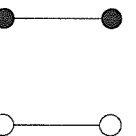
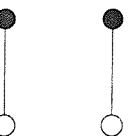
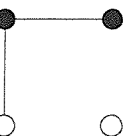
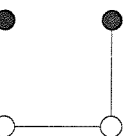
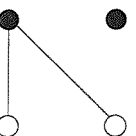
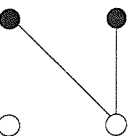
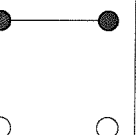
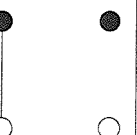
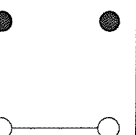
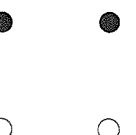
<p>37</p> <p>$n = 5$ $m = 6$ $c_2 = 0.250$ $c_3 = 0.125$</p>	<p>38</p> <p>$n = 5$ $m = 12$ $c_2 = 0.283$ $c_3 = 0.453$</p>	<p>39</p> <p>$n = 5$ $m = 12$ $c_2 = 0.222$ $c_3 = 0.248$</p>	<p>40</p> <p>$n = 5$ $m = 12$ $c_2 = 0.297$ $c_3 = 0.248$</p>	<p>41</p> <p>$n = 5$ $m = 24$ $c_2 = 0.271$ $c_3 = 0.178$</p>	<p>42</p> <p>$n = 5$ $m = 24$ $c_2 = 0.282$ $c_3 = 0.383$</p>
<p>43</p> <p>$n = 5$ $m = 12$ $c_2 = 0.280$ $c_3 = 0.383$</p>	<p>44</p> <p>$n = 5$ $m = 12$ $c_2 = 0.490$ $c_3 = 0.762$</p>	<p>45</p> <p>$n = 5$ $m = 24$ $c_2 = 0.608$ $c_3 = 1.033$</p>	<p>46</p> <p>$n = 5$ $m = 12$ $c_2 = 0.647$ $c_3 = 1.126$</p>	<p>47</p> <p>$n = 5$ $m = 24$ $c_2 = 0.173$ $c_3 = 0.406$</p>	<p>48</p> <p>$n = 5$ $m = 12$ $c_2 = 0.213$ $c_3 = 0.142$</p>
<p>49</p> <p>$n = 5$ $m = 12$ $c_2 = 0.368$ $c_3 = 0.510$</p>	<p>50</p> <p>$n = 5$ $m = 12$ $c_2 = 0.256$ $c_3 = 0.324$</p>	<p>51</p> <p>$n = 5$ $m = 12$ $c_2 = 0.563$ $c_3 = 1.008$</p>	<p>52</p> <p>$n = 5$ $m = 6$ $c_2 = 0.115$ $c_3 = 0.104$</p>	<p>53</p> <p>$n = 5$ $m = 12$ $c_2 = 0.303$ $c_3 = 0.343$</p>	<p>54</p> <p>$n = 5$ $m = 12$ $c_2 = 0.101$ $c_3 = 0.034$</p>
<p>55</p> <p>$n = 4$ $m = 24$ $c_2 = 0.364$ $c_3 = 0.546$</p>	<p>56</p> <p>$n = 4$ $m = 12$ $c_2 = 0.842$ $c_3 = 1.155$</p>	<p>57</p> <p>$n = 4$ $m = 12$ $c_2 = 0.576$ $c_3 = 0.859$</p>	<p>58</p> <p>$n = 4$ $m = 12$ $c_2 = 0.413$ $c_3 = 0.523$</p>	<p>59</p> <p>$n = 4$ $m = 24$ $c_2 = 0.343$ $c_3 = 0.448$</p>	<p>60</p> <p>$n = 4$ $m = 24$ $c_2 = 0.545$ $c_3 = 0.737$</p>
<p>61</p> <p>$n = 4$ $m = 12$ $c_2 = 0.719$ $c_3 = 1.337$</p>	<p>62</p> <p>$n = 4$ $m = 6$ $c_2 = 0.125$ $c_3 = 0.063$</p>	<p>63</p> <p>$n = 4$ $m = 4$ $c_2 = 0.202$ $c_3 = 0.210$</p>	<p>64</p> <p>$n = 4$ $m = 3$ $c_2 = 0.151$ $c_3 = 0.152$</p>	<p>65</p> <p>$n = 4$ $m = 12$ $c_2 = 0.900$ $c_3 = 1.572$</p>	<p>66</p> <p>$n = 4$ $m = 12$ $c_2 = 0.152$ $c_3 = 0.155$</p>
<p>67</p> <p>$n = 4$ $m = 24$ $c_2 = 0.279$ $c_3 = 0.188$</p>	<p>68</p> <p>$n = 4$ $m = 24$ $c_2 = 0.250$ $c_3 = 0.125$</p>	<p>69</p> <p>$n = 4$ $m = 12$ $c_2 = 0.364$ $c_3 = 0.530$</p>	<p>70</p> <p>$n = 4$ $m = 1$ $c_2 = 1.000$ $c_3 = 1.000$</p>	<p>71</p> <p>$n = 4$ $m = 12$ $c_2 = 0.447$ $c_3 = 0.622$</p>	<p>72</p> <p>$n = 3$ $m = 12$ $c_2 = 0.203$ $c_3 = 0.210$</p>

continued overleaf

Marked Graphs with $p=1, q=4$ (continued)

<p>73</p>  <p>$n = 3$ $m = 24$ $c_2 = 0.545$ $c_3 = 0.762$</p>	<p>74</p>  <p>$n = 3$ $m = 24$ $c_2 = 0.858$ $c_3 = 1.174$</p>	<p>75</p>  <p>$n = 3$ $m = 4$ $c_2 = 0.239$ $c_3 = 0.267$</p>	<p>76</p>  <p>$n = 3$ $m = 4$ $c_2 = 1.000$ $c_3 = 1.000$</p>	<p>77</p>  <p>$n = 3$ $m = 12$ $c_2 = 0.595$ $c_3 = 0.865$</p>	<p>78</p>  <p>$n = 3$ $m = 12$ $c_2 = 0.500$ $c_3 = 0.500$</p>
<p>79</p>   <p>$n = 3$ $m = 12$ $c_2 = 0.420$ $c_3 = 0.572$</p>	<p>80</p>   <p>$n = 3$ $m = 12$ $c_2 = 0.317$ $c_3 = 0.318$</p>	<p>81</p>   <p>$n = 3$ $m = 6$ $c_2 = 0.250$ $c_3 = 0.125$</p>	<p>82</p>   <p>$n = 3$ $m = 4$ $c_2 = 0.203$ $c_3 = 0.155$</p>	<p>83</p>   <p>$n = 2$ $m = 3$ $c_2 = 0.250$ $c_3 = 0.250$</p>	<p>84</p>   <p>$n = 2$ $m = 12$ $c_2 = 0.500$ $c_3 = 0.500$</p>
<p>85</p>   <p>$n = 2$ $m = 6$ $c_2 = 1.000$ $c_3 = 1.000$</p>	<p>86</p>   <p>$n = 2$ $m = 12$ $c_2 = 0.842$ $c_3 = 1.124$</p>	<p>87</p>   <p>$n = 2$ $m = 12$ $c_2 = 0.319$ $c_3 = 0.319$</p>	<p>88</p>   <p>$n = 1$ $m = 6$ $c_2 = 0.500$ $c_3 = 0.500$</p>	<p>89</p>   <p>$n = 1$ $m = 4$ $c_2 = 1.000$ $c_3 = 1.000$</p>	<p>90</p>   <p>$n = 0$ $m = 1$ $c_2 = 1.000$ $c_3 = 1.000$</p>

Marked Graphs with $p=2, q=2$

<p>1</p>  <p>$n = 6$ $m = 1$ $c_{22} = 0.063$ $c_{23} = 0.016$ $c_{33} = 0.002$</p>	<p>2</p>  <p>$n = 5$ $m = 1$ $c_{22} = 0.063$ $c_{23} = 0.016$ $c_{33} = 0.002$</p>	<p>3</p>  <p>$n = 5$ $m = 4$ $c_{22} = 0.639$ $c_{23} = 0.416$ $c_{33} = 0.629$</p>	<p>4</p>  <p>$n = 5$ $m = 1$ $c_{22} = 1.000$ $c_{23} = 1.000$ $c_{33} = 1.000$</p>	<p>5</p>  <p>$n = 4$ $m = 1$ $c_{22} = 1.000$ $c_{23} = 1.000$ $c_{33} = 1.000$</p>	<p>6</p>  <p>$n = 4$ $m = 2$ $c_{22} = 1.443$ $c_{23} = 1.980$ $c_{33} = 2.677$</p>
<p>7</p>  <p>$n = 4$ $m = 4$ $c_{22} = 0.424$ $c_{23} = 0.579$ $c_{33} = 0.291$</p>	<p>8</p>  <p>$n = 4$ $m = 4$ $c_{22} = 1.000$ $c_{23} = 1.000$ $c_{33} = 1.000$</p>	<p>9</p>  <p>$n = 4$ $m = 2$ $c_{22} = 1.435$ $c_{23} = 1.888$ $c_{33} = 2.383$</p>	<p>10</p>  <p>$n = 4$ $m = 2$ $c_{22} = 0.250$ $c_{23} = 0.125$ $c_{33} = 0.125$</p>	<p>11</p>  <p>$n = 3$ $m = 2$ $c_{22} = 1.000$ $c_{23} = 1.000$ $c_{33} = 1.000$</p>	<p>12</p>  <p>$n = 3$ $m = 2$ $c_{22} = 0.25$ $c_{23} = 0.125$ $c_{33} = 0.125$</p>
<p>13</p>  <p>$n = 3$ $m = 2$ $c_{22} = 1.000$ $c_{23} = 1.000$ $c_{33} = 1.000$</p>	<p>14</p>  <p>$n = 3$ $m = 2$ $c_{22} = 1.165$ $c_{23} = 1.429$ $c_{33} = 1.661$</p>	<p>15</p>  <p>$n = 3$ $m = 4$ $c_{22} = 0.852$ $c_{23} = 0.854$ $c_{33} = 1.166$</p>	<p>16</p>  <p>$n = 3$ $m = 4$ $c_{22} = 1.000$ $c_{23} = 1.000$ $c_{33} = 1.000$</p>	<p>17</p>  <p>$n = 3$ $m = 2$ $c_{22} = 1.443$ $c_{23} = 1.980$ $c_{33} = 2.677$</p>	<p>18</p>  <p>$n = 3$ $m = 2$ $c_{22} = 1.000$ $c_{23} = 1.000$ $c_{33} = 1.000$</p>
<p>19</p>  <p>$n = 2$ $m = 1$ $c_{22} = 0.500$ $c_{23} = 0.500$ $c_{33} = 0.500$</p>	<p>20</p>  <p>$n = 2$ $m = 2$ $c_{22} = 1.000$ $c_{23} = 1.000$ $c_{33} = 1.000$</p>	<p>21</p>  <p>$n = 2$ $m = 4$ $c_{22} = 1.000$ $c_{23} = 1.000$ $c_{33} = 1.000$</p>	<p>22</p>  <p>$n = 2$ $m = 4$ $c_{22} = 0.849$ $c_{23} = 1.164$ $c_{33} = 1.164$</p>	<p>23</p>  <p>$n = 2$ $m = 2$ $c_{22} = 1.000$ $c_{23} = 1.000$ $c_{33} = 1.000$</p>	<p>24</p>  <p>$n = 2$ $m = 2$ $c_{22} = 1.000$ $c_{23} = 1.000$ $c_{33} = 1.000$</p>
<p>25</p>  <p>$n = 1$ $m = 1$ $c_{22} = 1.000$ $c_{23} = 1.000$ $c_{33} = 1.000$</p>	<p>26</p>  <p>$n = 1$ $m = 4$ $c_{22} = 1.000$ $c_{23} = 1.000$ $c_{33} = 1.000$</p>	<p>27</p>  <p>$n = 1$ $m = 1$ $c_{22} = 0.500$ $c_{23} = 0.500$ $c_{33} = 0.500$</p>	<p>28</p>  <p>$n = 0$ $m = 1$ $c_{22} = 1.000$ $c_{23} = 1.000$ $c_{33} = 1.000$</p>		

Bibliography

- Barnard J., McCulloch R., and Meng X.L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with applications to shrinkage. *Statistica Sinica* **10**, 1281–1311.
- Brooks S.P. and Gelman A. (1998). General Methods for Monitoring Convergence of Iterative Simulations. *Journal of Computational and Graphical Statistics* **7**, 4, 434–455.
- Brooks S.P. and Giudici P. (2000). MCMC convergence assessment via two-way ANOVA. *Journal of Computational and Graphical Statistics* **9**, 266–285.
- Brooks S.P., Giudici P. and Roberts G.O. (2003). Efficient construction of reversible jump MCMC proposal distributions. *J. R. Statist. Soc. B* **65**, 1, 3–39.
- Castelloe J.M. and Zimmerman D.M. (2002). Convergence assessment for reversible jump MCMC samplers. Technical Report 313, Department of Statistics and Actuarial Science, University of Iowa.
- Darroch J.N., Lauritzen S.L. and Speed T.P. (1980). Markov field and log linear interactions models for contingency tables. *Ann. Statist.* **8**, 522–539.
- Dawid A.P. (1979). Conditional Independence in Statistical Theory. *J. R. Statist. Soc. B* **41**, 1, 1–31.
- Dawid A.P. and Lauritzen S.L. (1993). Hyper Markov Laws in the statistical analysis of decomposable graphical models. *Ann. Statist.* **21**, 1272–317.
- Dellaportas P. and Forster J.J. (1999). Markov chain Monte Carlo determination for hierarchical and graphical log-linear models. *Biometrika* **86**, 615–633.
- Dellaportas P., Giudici P. and Roberts G. (2003). Bayesian inference for nondecomposable graphical Gaussian models. *Sankhya* **65**, 43–55.

- Dempster A.P. (1972). Covariance Selection. *Biometrics* **28**, 1, 157–175.
- Diaconis P. and Ylvisaker D. (1979). Conjugate priors for exponential families. *Ann. Statist.* **7**, 269–281.
- Diciccio J., Kass R.E., Raftery A. and Wasserman L. (1997). Computing Bayes factors by combining simulation and asymptotic approximations. *Journal of the American Statistical Association* **92**, 903–915.
- Edwards D. (1987). A guide to MIM. Research Report 87/1, Statistical Research Unit, Copenhagen University.
- Edwards D. (1990). Hierarchical interaction models. *J. R. Statist. Soc. B* **52**, 3–20.
- Edwards D. (1995,2000). *Introduction to Graphical Modelling*. New York, London: Springer-Verlag.
- Edwards D. and Kreiner S. (1983). The analysis of contingency tables by graphical models. *Biometrika* **70**, 3, 553–565.
- Fligner M.A. and Verducci J.S. (1988). A Nonparametric Test for Judges' Bias in an Athletic Competition. *Applied Statistics* **37**, 1, 101–110.
- Gelman A. and Rubin D.B. (1992). Inference from Iterative Simulations Using Multiple Sequences. *Statistical Science* **7**, 2, 457–511.
- Geman S. and Geman D. (1984). Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. *IEEE Trans. on Pattern Analysis and Machine Intelligence* **37**, 721–741.
- Geyer C.J. (1992). Practical Markov chain Monte Carlo (with discussion). *Statistical Science* **41**, 473–511.
- Giudici P. (1996). Learning in graphical Gaussian models. In *Bayesian Statistics* J. Berger, J.M. Bernardo, A.P. Dawid and A.F.M. Smith (eds.). Oxford: Oxford University Press, pp. 621–8.
- Giudici P. and Green P.J. (1999). Decomposable graphical Gaussian model determination. *Biometrika* **86**, 785–801.
- Goldberg A.S. (1966). Discerning a causal pattern among data on voting behaviour. *The American Political Science Review* **60**, 913–922.
- Green P.J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–32.

- Halkin H., Sheiner L.B., Peck C.C. and Melman K.L. (1975). Determinants of the renal clearance of digoxin. *Clin. Pharmacol. Theor.* **17**, 385–394.
- Lauritzen S.L. (1996). *Graphical Models*. Oxford: Oxford University Press.
- Lauritzen S.L. and Wermuth N. (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. *Ann. Statist* **17**, 31–54.
- Liechty J.C., Liechty M.W. and Muller P. (2004). Bayesian correlation estimation. *Biometrika* **91**, 1, 1–14.
- Lindley D.V. (1957). A statistical paradox. *Biometrika* **44**, 187–192.
- Liu C. and Rubin D.B. (1998). Ellipsoidally symmetric extensions of the general location model for mixed categorical and continuous data. *Biometrika* **85**.
- Madigan D. and Raftery A.E. (1994). Model selection and accounting for model uncertainty in graphical models using occam’s window. *J. Am. Statist. Assoc.* **89**, 428, 1535–46.
- Madigan D., Raftery A.E., York J., Bradshaw J.M. and Almond R.G. (1994). Strategies for graphical model selection. In *Selecting Models from Data: AI and Statistics IVP*. Cheesman and R.W. Oldford (eds.). New York: Springer-Verlag, pp. 91–100.
- Madigan D. and York J. (1995). Bayesian graphical models for discrete data. *Int. Statist. Rev.* **63**, 215–32.
- Mardia K.V., Kent J.T. and Bibby J.M. (1979). *Multivariate Analysis*. London: Academic Press.
- Metropolis N., Rosenbluth A.W., Rosenbluth M.N., Teller A.H. and Teller E. (1953). Equation of state calculations by fast computing machine. *J. Chem. Phys.* **21**, 1087–1091.
- Morant G.M. (1923). A first study of the Tibetan skull. *Biometrika* **14**, 193–260.
- Morrison D.F. (1976). *Multivariate Statistical Methods*. New York: McGraw-Hill, 2 edition.
- Olkin I. and Tate R.F. (1961). Multivariate Correlation Models with mixed discrete and continuous variables. *Annals of Mathematical Statistics* **32**, 2, 448–465.

- Ritter C. and Tanner M.A. (1992). Facilitating the Gibbs sampler: The Gibbs stopper and the Griddy-Gibbs sampler. *J. Amer. Statist. Assoc.* **87**, 861–868.
- Roberts G.O., Gelman A. and Gilks W.R. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Prob.* **7**, 1, 110–120.
- Rousseeuw P.J. and Molenberghs G. (1994). The shape of correlation matrices. *The American Statistician* **48**, 4, 276–279.
- Roverato A. (2002). Hyper Inverse Wishart distribution for non-decomposable graphs and its application to Bayesian inference for Gaussian graphical models. *Scand. J. Statist.* **29**, 391–411.
- Speed T.P. and Kiiveri H.T. (1986). Gaussian Markov distributions over finite graphs. *Ann. Statist.* **14**, 1, 138–150.
- Tierney L. (1994). Markov chains for exploring posterior distributions. *Ann. Statist.* **22**, 4, 1701–1728.
- Wermuth N. (1976). Analogies between multiplicative models in contingency tables and covariance selection. *Biometrics* **32**, 95–108.
- Wermuth N. (1980). Linear recursive equations, covariance selection, and path analysis. *J. Amer. Statist. Assoc.* **75**, 372, 963–972.
- Wermuth N. and Lauritzen S.L. (1989). On substantive research hypotheses, conditional independence graphs and graphical chain models. *J. R. Statist. Soc. B* **52**, 21–50.
- Whittaker J. (1990). *Graphical Models in Applied Multivariate Statistics*. New York: Wiley.
- Wong F., Carter C.K. and Kohn R. (2003). Efficient estimation of covariance selection models. *Biometrika* **90**, 4, 809–830.