UNIVERSITY OF SOUTHAMPTON

Bayesian Inference for Log-Linear Models

by

Mark Edwin Grigsby

Thesis submitted for the degree of Doctor of Philosophy

FACULTY OF MATHEMATICAL STUDIES

MATHEMATICS

October, 2001

To Grandpa

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF MATHEMATICAL STUDIES

MATHEMATICS

Doctor of Philosophy

BAYESIAN INFERENCE FOR LOG-LINEAR MODELS

by Mark Edwin Grigsby

Inference for multivariate discrete data often concerns associations between variables modelled using log-linear models. This thesis focuses on the Bayesian analysis of log-linear models. Various prior distributions are investigated which are suitable for reference analyses.

The conditional Dirichlet distribution, which has the attractive property that its parameters may be interpreted as prior cell counts, is introduced. This prior is useful for both reference analyses, where small prior values are used, and as an informative prior, where (hypothetical) prior cell counts may be available. The conditional Dirichlet is shown to be equivalent to a hyper Dirichlet density (which admits straightforward analyses) for decomposable log-linear models. Hence a natural extension of the hyper Dirichlet distribution to non-decomposable models is obtained.

The conditional Dirichlet distribution is not tractable in general, so Monte Carlo and other approximation methods are required. Gibbs sampling is applied to obtain samples from prior and posterior conditional Dirichlet distributions. The sampler is found to mix well, producing samples which are not highly dependent.

Laplace's method is used for the approximation of integrals, although it is found to perform poorly where prior parameters take small values. However, accurate results may be obtained for the posterior analysis of datasets where cell counts are large. The method of bridge sampling is applied to the problem of determining the normalising constants for conditional Dirichlet distributions. The sampler is found to produce good results, even when prior parameters take small values, and this is illustrated by application to several examples.

Jeffreys' prior, which is a reference prior by definition, is considered, and an explicit expression is presented for the Jeffreys' prior for a decomposable log-linear model. In many cases, this is found to be a product of independent Dirichlet distributions for the parameters of a particular decomposition of the model. For other decomposable models, where the normalising constant for Jeffreys' prior is not directly available, the method of bridge sampling is again applied, and found to produce accurate results. The Monte Carlo samples needed are obtained using Metropolis Hastings sampling.

Finally, the choice of prior distribution is considered in further detail. Unit information priors, for which easy approximations to marginal likelihoods are available, are discussed, and the relationship between the Laplace approximation for marginal likelihoods and the Schwarz criterion is investigated for log-linear models under multinomial sampling. It is shown that marginal likelihoods using Jeffreys' prior may be approximated by a modified version of the Schwarz approximation, with error of order $n^{-\frac{1}{2}}$.

Contents

List of Figures

List of Tables

Acknowledgements

I would like to thank my supervisor, Dr Jon Forster, for his guidance and endless enthusiasm during the past four years. I would also like to thank the other members of staff in the Statistics group at the University of Southampton, in particular Professors Phil Prescott and Sue Lewis.

This research was carried out under a research studentship from the Engineering and Physical Sciences Research Council, to whom I am grateful.

My time in Southampton has been enriched by so many people. Thank-you to my housemates Chris, Jen, Rob, Eric, Pete and Ralph for putting up with me, and a particular thank-you to Karen for all the fun!

Several other friends deserve a mention in my thesis: Ben, who doesn't believe I am no longer going to be a student; Adam, who helped me find a job; Richard, because I promised; Harz, for the phone calls; and Richard Ng, for his routine assistance!

Finally, I am indebted to three people who have given me so much – Susan, who is my inspiration, and Mum and Gram, for being there. Thank you.

Chapter 1

Introduction

A contingency table is a collection of cells each containing counts of units cross-classified according to a set of factors. The analysis of data which may be presented in contingency tables forms an important area in statistics. Contingency tables are often highly structured, though this structure is not often immediately obvious without detailed statistical analysis. However, the investigation of this structure is extremely important as it enables us to understand the relationships between variables, and also provides the key to estimation of quantities of interest. The underlying structure of a contingency table is usefully represented by a formal statistical model, and a standard way of doing this is to use a log-linear model, which linearly relates the logarithms of the cell means (or cell probabilities) to a set of model parameters. The form of this linear relationship depends on the structure of the data, *i.e.* on the relationships between the variables represented by the contingency table.

Classical statistical analysis focuses on methods such as maximum likelihood estimation to estimate model parameters, and so to obtain estimated cell counts (or probabilities). However, this thesis is based on analysis within a Bayesian framework, whereby the cell means or probabilities (or model parameters) are all treated as random variables and hence must be given prior distributions de-

scribing uncertainty about them before any data have been observed. The data are then used to 'update' the prior distributions to form posterior distributions which encapsulate all the knowledge about the parameters, given the data. The utility of the Bayesian approach is that it enables us to obtain full posterior summaries of uncertainty for any function of interest.

This thesis focuses on 'reference' prior distributions, where we have negligible substantive prior knowledge, and on the computation of posterior quantities of interest. This enables us to obtain estimates of quantities of interest, assess the corresponding uncertainty, and also to investigate the structure of the underlying statistical model.

1.1 Contingency Tables

Suppose we have a set of multivariate categorical data, where n units have been cross-classified by a number of categorical variables and the counts of the resulting cross-classification presented in a contingency table. Let the set of categorical variables or factors be Γ, resulting in a $|\Gamma|$-way contingency table.

Following the notation introduced by Darroch, Lauritzen and Speed (1980), the set of cells in the table is the set $I = \prod_{\gamma\in\Gamma} I_\gamma$, where I_γ is the set of levels of factor γ. A particular cell will be denoted by $\boldsymbol{i} = (i_\gamma : \gamma \in \Gamma)$, the corresponding cell count by $n(\boldsymbol{i})$, and the cell probability by $p(\boldsymbol{i})$, where this represents the probability that a particular unit lies in cell $\boldsymbol{i}$. The vector of all the cell probabilities will be written $\boldsymbol{p}$, and the cell counts $\boldsymbol{n}$. The total cell count will be denoted n, where $n = \sum_{\boldsymbol{i}} n(\boldsymbol{i})$. The number of cells, m, in the table is $|I| = \prod_\gamma |I_\gamma|$. This notation is best illustrated by an example:

Suppose we have three variables A, B and C, where A is binary and B and C have 3 levels, and that these variables cross-classify some data in a 3-way table. In this case, $\Gamma = \{A, B, C\}$, and a cell in the table is therefore $\boldsymbol{i} = (i_A, i_B, i_C)$ where i_A can take values 1 and 2, and i_B and i_C take values 1, 2 or 3. Hence the

cell which contains the data for variables A and B at level 1 and variable C at level 3 is $\boldsymbol{i} = (1, 1, 3)$, and the cell probability is $p(\boldsymbol{i})$.

The typical model for data in a contingency table assumes that a known number of individual units n are assigned at random to a particular cell $\boldsymbol{i}$ with probability $p(\boldsymbol{i})$. Therefore the vector of cell counts $\boldsymbol{n}$ has a *multinomial* distribution, which has probability function

$$f(\boldsymbol{n}|\boldsymbol{p}) = n! \prod_{\boldsymbol{i}} \frac{p(\boldsymbol{i})^{n(\boldsymbol{i})}}{n(\boldsymbol{i})!}$$

1.2 Log-Linear Models

One motivation for analysing contingency table data is modelling the associations between classifying variables. Such considerations typically include how variables are conditionally independent or independent of one another. The standard way of doing this is by representing the underlying statistical model as a *log-linear model.* Different association structures, including independence and conditional independence, result from models with different forms, and from varying parameter values within a particular model. This section will introduce general log-linear models, and also various special subsets of these models.

We assume that $n(\boldsymbol{i})$ is an observation of a multinomial random variable with corresponding vector of cell probabilities $p(\boldsymbol{i})$. Then, again following Darroch, Lauritzen and Speed (1980) we denote the log-linear model

$$\log p(\boldsymbol{i}) = \sum_{a \subseteq \Gamma} \xi_a(\boldsymbol{i}_a) \qquad \boldsymbol{i} \in I \tag{1.1}$$

where $\boldsymbol{i}_a$ is the marginal cell $\boldsymbol{i}_a = (i_\gamma, \gamma \in a)$. As $p(\boldsymbol{i})$ is a vector of cell probabilities which sum to 1, a normalising constant $\xi_\emptyset$ is necessary in (1.1). Note that certain constraints must be imposed on the terms $\xi_a(\boldsymbol{i}_a)$ (which we shall refer to as the interaction terms) to ensure identifiability. These will be discussed later.

A saturated model is parameterised by a full set of interaction terms, whereas setting certain ξ_a terms to zero defines a particular non-saturated log-linear model. Hence the non-zero terms define the model, and may take arbitrary values. It is straightforward to write down the number of possible distinct log-linear models for a set of factors Γ; there are $2^{|\Gamma|}$ possible $a \subseteq \Gamma$, giving rise to $2^{2^{|\Gamma|}}$ different log-linear models.

1.2.1 Hierarchical Models

Commonly, we do not consider the full set of log-linear models, and instead restrict attention to a smaller subset of these called the *hierarchical log-linear models*. To obtain these, we impose restrictions on the $\xi_a(i_a)$, namely that setting ξ_a equal to zero means we must also set ξ_b to be zero for all $b \supseteq a$. For example, suppose that $\Gamma = \{A, B, C\}$, and that $\xi_{AB} = 0$. In this case, we require $\xi_{ABC} = 0$ in a hierarchical model. It is not possible to write an explicit expression for the number of such models, but this number is much smaller than the total number of log-linear models.

Let us define the *generators* of a model as the maximal sets a such that ξ_a is non-zero. Then a hierarchical model is determined uniquely by its generators.

1.2.2 Graphical Models

The set of graphical models form a highly attractive subset of the hierarchical models, both for ease of analysis and their obvious interpretation in terms of conditional independence (an interpretation which is immediately obvious from the graph). Graphical models may be either directed or undirected – the former provide motivation for some of the work in this document, although it is the latter which we shall define first, due to their relative simplicity.

A graphical log-linear model may be represented by a graph, with a set of vertices $\mathcal{V}$ corresponding to the variables, and a set $\mathcal{E}$ of edges representing

the independence structure. The notation (X, Y) is used to represent the edge between variables X and Y. The absence of an edge between two vertices X and Y means that X is conditionally independent of Y given all other variables. This is equivalently written as: if $(X, Y) \notin \mathcal{E}$, then $X \perp\!\!\!\perp Y \mid \mathcal{V} \setminus \{X, Y\}$. Variables X and Y are (unconditionally) independent if no path of edges exists between vertices X and Y, in which case $X \perp\!\!\!\perp Y$.

A subset C of Γ is called a *clique* if the subgraph containing only elements of C has an edge connecting each element (*i.e.* is complete), and the inclusion of another vertex from $\mathcal{V}$ in C would result in at least one pair of unconnected vertices. A graph is *triangulated* if it contains no chordless cycles of length greater than three, and the subset D is said to *separate* subsets A and B if every path from any vertex in A to one in B must pass through a vertex in D. In such a case, variables in A are conditionally independent from those in B, given D.

As mentioned above, a hierarchical model is determined by its generators, and a model is *graphical* if its generators correspond to the cliques of its (undirected) conditional independence graph. These models form a subset of the log-linear models. We will assume throughout that all models include the intercept term $\xi_\emptyset$ and all main effect terms (ξ_a where $|a| = 1$), since those without are of little interest. Then the $\binom{|\Gamma|}{2}$ possible edges in each graph gives the total number of possible graphical models to be $2^{\binom{|\Gamma|}{2}}$.

A *directed graph* contains edges *from* one vertex *to* another, for example $X \to Y$ denotes the presence on an edge from X to Y, and we call X a parent of Y and Y a child of X. The edge from X to Y will be written $\langle X, Y \rangle$. The set of parents of Y is denoted by $pa(Y)$. For a subgraph A, $pa(A)$ denotes the set of parents of vertices in A that are not themselves elements of A. A path of length $n \geq 0$ from X to Y is a sequence $X = X_0, \ldots, X_n = Y$ of distinct vertices such that $\langle X_{i-1}, X_i \rangle \in \mathcal{E}$ for all $i = 1, \ldots, n$. If there is a path from X to Y we write $X \rightarrowtail Y$. The set of vertices X such that $X \rightarrowtail Y$ are the ancestors $an(Y)$

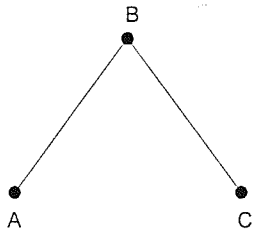
of Y and the descendants $de(X)$ of X are the vertices Y such that $X \rightarrowtail Y$. The nondescendants of X are $nd(X) = \mathcal{V} \setminus (de(X) \cup \{X\})$. A path which starts and ends at the same point is known as a *cycle,* and a directed graph is *acyclic* if it contains no cycles.

Note that any hierarchical model can be represented by an (undirected) conditional independence graph, although such a graph does not necessarily represent a single non-graphical model. However, these models will not be excluded from our analyses, as they form a rich collection of models with many applications. An example of such a model is the model containing the three variables A, B and C with interaction terms AB, AC and BC though no 3-way interaction term ABC. In this case, the 2-way interactions are homogeneous with respect to the third variable; for example, the interaction between A and B does not depend on the value of C. Although this model is clearly not graphical, real data may be found to follow this pattern of association, so this model should not be excluded from our analyses.

1.2.3 Decomposable Models

Another smaller subset of models within graphical models are *decomposable models.* These are defined as models whose joint cell probabilities may be directly expressed as a function of the marginal probabilities of the cliques of the model. An equivalence to this definition which is more useful in practice is that a model is decomposable if its graph is triangulated.

For example, consider the model represented by the graph below.



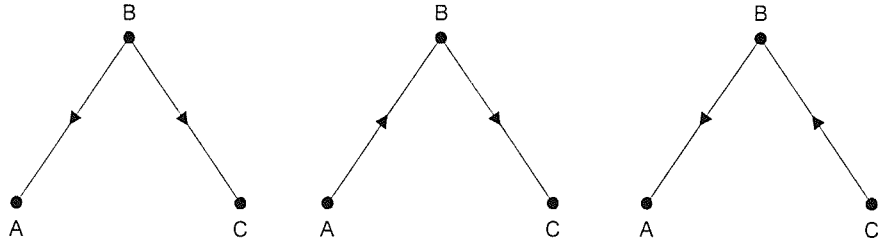
This graph is clearly triangulated (with cliques $\{A, B\}$ and $\{B, C\}$), so the model is decomposable and the joint cell probabilities $\boldsymbol{p}$ may be written as a product of the marginal and conditional probabilities $\boldsymbol{p}^A$, $\boldsymbol{p}^{B|A}$, $\boldsymbol{p}^{C|B}$. Equivalently, the cell probabilities may be directly expressed in terms of marginal probabilities of cliques and separators as $p(\boldsymbol{i})=\frac{p(\boldsymbol{i}_{AB})p(\boldsymbol{i}_{BC})}{p(\boldsymbol{i}_B)}$.

These models admit the most straightforward analyses, but clearly exclude many potential (and useful) models, and there is often little justification to restrict attention to these models other than computational considerations.

One benefit of decomposable models which will aid certain parts of our analysis is that, if the model is decomposable, then we may use the undirected conditional independence graph to construct a directed version with the same Markov structure (Dawid and Lauritzen, 1993). We can use this directed graph to obtain a *perfect numbering* of the variables in the graph, by numbering the vertices so that those at the 'top' of the graph (*i.e.* the ones with no parents) have the lowest numbers. The edges are hence necessarily directed from vertices with low numbers to those with higher numbers. For directed graphs, the joint probability may be expressed as

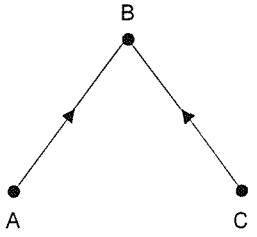
$$p(\boldsymbol{i}) = \prod_{\gamma=1}^{|\Gamma|} P(\gamma = i_\gamma | pa(\gamma) = \boldsymbol{i}_{pa(\gamma)})$$

As an example of directed graphical representations, consider the model represented by the undirected graph above. Several possible directed versions of this graph are possible:

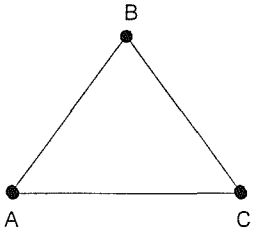


Each of these graphs admit different perfect orderings of variables, and one admits two orderings. Working from left to right, the first admits orderings BCA and BAC, and the second and third admit orderings CBA and ABC respectively. Each of these graphs and orderings is equivalent to the others, and it is often the case that a single undirected graph can give rise to several different directed versions.

Note that in this example, the only directed graph which is not equivalent to the undirected version is



An easy way to see why this graph represents a different model is by noting that an equivalent undirected graph is obtained from the directed version by *moralising*, whereby unjoined parents in the directed version are joined in the undirected one. Hence, the directed graph shown above would be represented by the undirected graph below, which corresponds to the saturated model.



Throughout this thesis, directed and undirected graphical models will be used interchangeably, as they may represent identical statistical models. However, certain situations lend themselves more to one type than another, due to certain implications of directed graphs; for example, an arrow from A to B may imply a temporal effect, with event A preceeding event B. Whereas this may be desirable

in some examples, an undirected representation may be more appropriate in others. For example, a relationship between smoking and heart disease would be suited to a directed graph, whereas a relationship between an individual's hand size and foot size would be better represented by an undirected graph.

1.3 Parameterisations

In order to admit more straightforward analyses and calculations involving the log-linear models described above, it is helpful to consider a parameterisation of (1.1) where the parameters are identifiable and linearly independent. To obtain this parameterisation, we shall follow the same method as Dellaportas and Forster (1999).

Let us define a $|I|$-dimensional vector ζ_a as $\zeta_a = \{\zeta_a(i), i \in I\}$. Here, $\xi_a(i_a)$ is replicated so that $\zeta_a(i) = \xi_a(i_a)$ for all $i_{\Gamma \setminus a}$, so that

$$\log \boldsymbol{p} = \sum_{a \subseteq \Gamma} \zeta_a \tag{1.2}$$

We may now choose a linearly independent set of

$$d_a = \prod_{y \in a} (|I_\gamma| - 1)$$

components of ζ_a for each $a \subseteq \Gamma$ as our model parameters. We make the usual choice

$$\beta_a = \{\xi_a(i_a), i_\gamma > 1 \text{ for all } \gamma \in a\}$$

where β_a involving $i_\gamma = 1$ for some γ are defined by prespecified constraints.

The log-linear model is expressed in terms of $\log \boldsymbol{p}$, but since these cell probabilities lie in a simplex space, where each $p(\boldsymbol{i})$ satisfies $0 < p(\boldsymbol{i}) < 1$ and $\sum p(\boldsymbol{i}) = 1$, it is useful to consider a multivariate logit transformation. Through-

out this document, two alternative definitions will be used for the vector of logit parameters $\boldsymbol{\theta}$. Note, however, that a linear one to one transformation exists between the two versions, and the choice of definition is made purely for ease of calculations. The first definition of $\theta(\boldsymbol{i})$ is given by

$$\theta(\boldsymbol{i}) = \log\left(\frac{p(\boldsymbol{i})}{p(\boldsymbol{i}_0)}\right) \tag{1.3}$$

where $\boldsymbol{i}_0$ refers to the cell with all factors at their lowest level. We call this the reference cell logit, as each probability is contrasted with a reference probability (in this case the first, $p(\boldsymbol{i}_0)$). Note that $\theta(\boldsymbol{i}_0) = 0$. This expression may be inverted in order to write the probabilities in terms of the logits,

$$p(\boldsymbol{i}) = \frac{e^{\theta(\boldsymbol{i})}}{\sum e^{\theta(\boldsymbol{i})}}$$

The alternative definition of $\theta(\boldsymbol{i})$ is the symmetric logit

$$\theta(\boldsymbol{i}) = \log\left(\frac{p(\boldsymbol{i})}{g(\boldsymbol{p})}\right) \tag{1.4}$$

where $g(\boldsymbol{p})$ is the geometric mean of the probabilities ($g(\boldsymbol{p}) = \left(\prod_{\boldsymbol{i}} p(\boldsymbol{i})\right)^{\frac{1}{n}}$). Here, $\boldsymbol{\theta}$ satisfies $\mathbf{1}^T\boldsymbol{\theta} = 0$, where $\mathbf{1}$ is a vector of 1's. This transformation admits the same inversion as the reference cell logit, namely $p(\boldsymbol{i}) = \frac{e^{\theta(\boldsymbol{i})}}{\sum e^{\theta(\boldsymbol{i})}}$.

The design matrix X of a log-linear model relates $\boldsymbol{\theta}$ to the model parameters $\boldsymbol{\beta}$. The form of this matrix depends on the logit chosen, though clearly in the case of the symmetric logit, then X must satisfy $\mathbf{1}^T X = 0$. The model may therefore be written

$$\boldsymbol{\theta} = X\boldsymbol{\beta}$$

Hence,

$$\begin{aligned}\theta(\boldsymbol{i},\boldsymbol{\beta}) &= (X\boldsymbol{\beta})(\boldsymbol{i}) \\ &= \sum_j x(\boldsymbol{i},j)\beta_j\,.\end{aligned}$$

1.4 Bayesian Analysis

The fundamental principle of Bayesian analysis, as opposed to traditional classical methods, is that uncertainty is represented through probability. Bayesian inference is based upon a probability distribution for the parameter vector *given* observed data $\boldsymbol{n}$, *i.e.* $f(\boldsymbol{p}|\boldsymbol{n})$, which we call the *posterior distribution* for $\boldsymbol{p}$.

Bayes' theorem states that, for variables $\boldsymbol{x}$ and $\boldsymbol{y}$,

$$\begin{aligned}f(\boldsymbol{x}|\boldsymbol{y}) &= \frac{f(\boldsymbol{y}|\boldsymbol{x})f(\boldsymbol{x})}{f(\boldsymbol{y})} \\ &= \frac{f(\boldsymbol{y}|\boldsymbol{x})f(\boldsymbol{x})}{\int f(\boldsymbol{y}|\boldsymbol{x})f(\boldsymbol{x})d\boldsymbol{x}}\end{aligned}$$

Hence we use this to obtain the posterior distribution for $\boldsymbol{p}$ given $\boldsymbol{n}$ from the likelihood function $f(\boldsymbol{n}|\boldsymbol{p})$ and the *prior distribution* for $\boldsymbol{p}$, $f(\boldsymbol{p})$, which represents the uncertainty about $\boldsymbol{p}$ prior to observing data $\boldsymbol{n}$.

$$f(\boldsymbol{p}|\boldsymbol{n}) = \frac{f(\boldsymbol{n}|\boldsymbol{p})f(\boldsymbol{p})}{\int f(\boldsymbol{n}|\boldsymbol{p})f(\boldsymbol{p})d\boldsymbol{p}}$$

Since the integral in the above expression merely provides the constant so that the posterior density $f(\boldsymbol{p}|\boldsymbol{n})$ integrates to 1, this is often omitted and we write

$$f(\boldsymbol{p}|\boldsymbol{n}) \propto f(\boldsymbol{n}|\boldsymbol{p})f(\boldsymbol{p})$$

1.4.1 Prior Distributions

An important choice in the analysis of log-linear models is that of the prior distribution $f(\boldsymbol{p})$. The prior distribution encapsulates the previous information about the cell probabilities $\boldsymbol{p}$, which may be obtained from expert opinion, previous data, or some other source. However, this information, even if available, is often difficult to express as a probability distribution. Hence it is often useful to perform a 'reference' analysis, where the prior distribution is called a *reference, noninformative* or *diffuse* prior. This enables us to examine the influence of informative priors compared with the reference approach.

A number of priors exist which may be used for the log-linear model parameters, and which are suitable for a reference analysis. These distributions are introduced in Chapter 2.

1.4.2 Model Choice

Suppose we have a set of models, M, one of which we believe has generated our data $\boldsymbol{n}$. In our case, these data are cell counts in a contingency table, so that $\boldsymbol{n}=(n(\boldsymbol{i}), \boldsymbol{i}\in I)$ are observations of random variables $\boldsymbol{n} = (N(\boldsymbol{i}), \boldsymbol{i} \in I)$. There are several problems with classical approaches to choosing a particular model to represent the data, and several authors have reported on this in detail (for example Raftery, 1996). Classical methods are based on p-values, and difficulties arise when comparing models which are not nested. Also, tests based on p-values may reject acceptable models when the sample size is large, and in certain cases with small sample sizes the asymptotics of such statistics may break down. Model uncertainty is also ignored, selecting a single model in a situation where several plausible models exist, each with reasonable probability. To counter such problems, a Bayesian approach to model selection is proposed:

Each model $m \in M = \{m_1, m_2, \ldots\}$ specifies a distribution for $\boldsymbol{n}$, $f(\boldsymbol{n}|m, \boldsymbol{\theta}_m)$, with the $\boldsymbol{\theta}_m$ an unknown vector of parameters for model m. We use Bayes' the-

orem to obtain the joint posterior distribution of m_k and $\boldsymbol{\theta}_{m_k}$

$$\begin{aligned} f(m, \boldsymbol{\theta}_m|\boldsymbol{n}) &\propto f(\boldsymbol{n}|m, \boldsymbol{\theta}_m) f(m, \boldsymbol{\theta}_m) \\ &\propto f(\boldsymbol{n}|m, \boldsymbol{\theta}_m) f(\boldsymbol{\theta}_m|m) f(m) \end{aligned}$$

Hence the posterior probability of model m may be found explicitly from

$$f(m|\boldsymbol{n}) = \frac{f(m) \int f(\boldsymbol{n}|m, \boldsymbol{\theta}_m) f(\boldsymbol{\theta}_m|m) d\boldsymbol{\theta}_m}{\sum_{m \in M} f(m) \int f(\boldsymbol{n}|m, \boldsymbol{\theta}_m) f(\boldsymbol{\theta}_m|m) d\boldsymbol{\theta}_m} \qquad m \in M \qquad (1.5)$$

where $\int f(\boldsymbol{n}|m, \boldsymbol{\theta}_m) f(\boldsymbol{\theta}_m|m) d\boldsymbol{\theta}_m = f(\boldsymbol{n}|m)$ is the *marginal likelihood,* sometimes interpreted as the probability of observing the data calculated before any data were observed.

If we have two competing models, m_1 and m_2, the problem reduces to the calculation of a *Bayes Factor*, which is the ratio of the posterior odds to the prior odds, and we have

$$\frac{f(m_1|\boldsymbol{n})}{f(m_2|\boldsymbol{n})} = \frac{f(m_1)}{f(m_2)} \times \frac{\int f(\boldsymbol{n}|m_1, \boldsymbol{\theta}_{m_1}) f(\boldsymbol{\theta}_{m_1}|m_1) d\boldsymbol{\theta}_{m_1}}{\int f(\boldsymbol{n}|m_2, \boldsymbol{\theta}_{m_2}) f(\boldsymbol{\theta}_{m_2}|m_2) d\boldsymbol{\theta}_{m_2}}$$

where the second term on the right hand side is the Bayes factor for model 1 against model 2. This may be referred to as B_{12}, and this notation can be extended to the case where we have multiple plausible models, by writing B_{jk} as the Bayes factor for model j against model k.

The Bayes factor is related to the classical likelihood ratio statistic, as when the two models are distributions with no unknown parameters the two quantities are equal. In the more general case, there is still a correspondence between Bayes factors and likelihood ratio statistics, with Bayes factors obtained by integration instead of maximisation.

Whereas model selection is the problem of using our data to select a single model m from M, *model averaging* involves estimating our quantity of interest

under each of the plausible models and then obtaining a model-averaged estimate by placing weights on each individual estimate according to how likely each model is. This is a useful tool when we have a number of competing models, none of which has a dominant posterior probability. For example, suppose our quantity of interest is ϕ, which has an interpretation under every model, then we may obtain the posterior distribution using the expression

$$f(\phi|\boldsymbol{n}) = \sum_k f(\phi|m_k, \boldsymbol{n}) f(m_k|\boldsymbol{n})$$

where $f(m_k|\boldsymbol{n})$ is obtained from expression (1.5)

1.5 Bayesian Computation

Many instances arise in Bayesian analysis where it is necessary to deal with prior or posterior distributions whose density functions are analytically intractable. In such cases, we need to resort to approximation methods. This area of (Bayesian) statistics has expanded dramatically with advances brought about by increased computing power. Various computation methods exist for addressing such problems, and these will be discussed at relevant stages in this thesis. In particular, methods of obtaining samples from intractable densities are described in Chapters 4 and 7, and methods of approximating normalising constants are described in Chapter 5. Methods of obtaining samples from a specified probability distribution are based on Markov chain Monte Carlo theory, which is described briefly here.

It is difficult to generate independent observations from an arbitrary multivariate distribution; however, Markov chain Monte Carlo (MCMC) methods of approximation make it possible to generate a dependent sample. Suppose we require a sample $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots$ from a p-dimensional density $f(\boldsymbol{\theta})$. The Markov chain Monte Carlo method involves setting up a Markov chain, which is a ran-

dom mechanism whereby the distribution of $\boldsymbol{\theta}^{(t)}$ depends on $\boldsymbol{\theta}^{(t-1)}$. We know from Markov chain theory that $\boldsymbol{\theta}^{(t)}$ will have a particular limiting 'long run' equilibrium (or ergodic) distribution. Therefore, for large enough T, if $t > T$, $\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^{(t+1)}, \ldots$ can be considered to be identically distributed according to this distribution, regardless of the value of $\boldsymbol{\theta}^{(1)}$. Hence, provided we can construct a Markov chain whose equilibrium distribution is $f(\boldsymbol{\theta})$, then we can simulate this chain, and thus obtain a dependent sample from $f(\boldsymbol{\theta})$.

The value of T at which the equilibrium distribution is reached is known as the 'burn-in' length, and observations before this should be discarded. In practice however, provided we start the chain at a plausible observation from $f(\boldsymbol{\theta})$ (for example at the mode) then the burn-in is zero and no observations need be discarded.

Samples obtained using MCMC methods are by definition dependent. However, provided the parameter space is explored thoroughly by the sampler, then it is said to be 'mixing well', and successive samples are not highly dependent. If, however, there is high correlation between successive observations, then the sampler is said to be 'mixing poorly' and a highly dependent sample will be produced.

Two main methods exist for constructing suitable Markov chains with specified equilibrium distributions – Metropolis-Hastings sampling and Gibbs sampling. Gibbs sampling will be described in Chapter 4 and applied to the conditional Dirichlet distribution, and Metropolis Hastings sampling will be described in Chapter 6 and applied to Jeffreys' prior. MCMC methods may also be used in calculating posterior model probabilities, and such work is reviewed in Chapter 2.

1.6 Outline of the Thesis

Chapter 2 reviews previous Bayesian approaches to the analysis of log-linear models. The Dirichlet prior distribution is considered, as is a Normal prior distribution for log-linear model parameters, and graphical models are reviewed. The Bayesian approach to accounting for model uncertainty is reviewed, including model averaging and the use of a Markov chain Monte Carlo approach to calculating posterior model probabilities.

The focus of Chapter 3 is the conditional Dirichlet distribution, whose parameters have the attractive interpretation as prior cell counts. Its relationship to the hyper Dirichlet distribution is investigated. The conditional Dirichlet density is shown to be equivalent to a hyper Dirichlet density for decomposable log-linear models. This presents a natural extension of the hyper Dirichlet distribution to non-decomposable models.

The use of Gibbs sampling, based on adaptive rejection sampling, to obtain a Monte Carlo sample from prior and posterior conditional Dirichlet distributions is described in Chapter 4. The performance of this sampler is assessed, and it is found to mix well, producing samples which are not highly dependent.

Chapter 5 concerns model determination, and focuses on applications where conditional Dirichlet prior distributions are used. The calculation of Bayes factors for comparing models requires both prior and posterior normalising constants, and Laplace's method for the approximation of integrals is introduced and applied. However, it is found to perform poorly where prior parameters take small values. The method of bridge sampling, which requires a Monte Carlo sample, is introduced and found to produce good results, even for small prior parameters.

Jeffreys' prior is introduced in Chapter 6. An explicit expression is presented for the Jeffreys' prior for a decomposable log-linear model, and in many cases this is found to be a product of independent Dirichlet distributions for the parameters

of a particular decomposition of the model. For other decomposable models, Monte Carlo samples can be obtained using Metropolis Hastings sampling, and then bridge sampling applied to obtain the prior normalising constants.

In Chapter 7, the choice of prior distribution is considered in further detail. Unit information priors, for which easy approximations to marginal likelihoods are available, are discussed. The relationship between the Laplace approximation and the Schwarz criterion is investigated for marginal likelihoods for log-linear models under multinomial sampling. It is shown that marginal likelihoods using Jeffreys' prior may be approximated by a modified version of the Schwarz approximation, with error of order $O(n^{-\frac{1}{2}})$.

Chapter 8 illustrates the various ideas introduced in the thesis on two data analyses.

Finally, the results presented in the thesis are discussed in Chapter 9, and suggestions are given for future work.

Chapter 2

Review of Previous Work

2.1 Dirichlet Distribution

The focus of this thesis is on the Bayesian analysis of log-linear models using reference or vague prior distributions, for situations where little prior information is available. We will pay particular attention to distributions based on the *Dirichlet* distribution. The Dirichlet distribution is a natural choice of prior distribution for cell probabilities $\boldsymbol{p}$ (which are positive and sum to one). Its density has the form

$$f(\boldsymbol{p}) = \frac{\Gamma(\alpha)}{\prod_{\boldsymbol{i}} \Gamma(\alpha(\boldsymbol{i}))} \prod_{\boldsymbol{i} \in I} p(\boldsymbol{i})^{\alpha(\boldsymbol{i})-1}$$

where $\boldsymbol{\alpha}$ are parameters which control the location and dispersion of the distribution, and $\alpha = \sum_{\boldsymbol{i}} \alpha(\boldsymbol{i})$.

Under multinomial sampling, the likelihood function for a saturated log-linear model is given by

$$f(\boldsymbol{n}|\boldsymbol{p}) \propto \prod_{\boldsymbol{i} \in I} p(\boldsymbol{i})^{n(\boldsymbol{i})}$$

Hence the Dirichlet distribution is a conjugate prior distribution for a saturated log-linear model, as it leads to a Dirichlet posterior distribution with density of

the form

$$f(\boldsymbol{p}|\boldsymbol{n}) \propto \prod_{\boldsymbol{i}\in I} p(\boldsymbol{i})^{n(\boldsymbol{i})+\alpha(\boldsymbol{i})-1} \tag{2.1}$$

Conjugacy is convenient in Bayesian statistical analysis as it may (as in this case) result in tractable computation. Furthermore, prior specification may be facilitated if conjugacy is a result of prior and likelihood having a similar form. In such cases the 'information content' of the prior may be straightforward to specify.

As may be seen from expression (2.1), the parameters $\boldsymbol{\alpha}$ may be considered as a 'prior cell count'. Hence, for reference analyses, small common values of $\alpha(\boldsymbol{i})$ seem sensible. Indeed, Lindley (1964) considered the limiting case where $\alpha(\boldsymbol{i}) = 0$, producing an improper prior density (which does not integrate to 1). The problem with this approach is that it will lead to an improper posterior density if any cells have zero samples.

Setting $\alpha(\boldsymbol{i}) = 1$ results in a uniform prior (Lidstone, 1920), a conventional choice for a noninformative prior density. Two other popular choices for $\alpha(\boldsymbol{i})$ exist which are preferred in this thesis. The first is $\alpha(\boldsymbol{i}) = \frac{1}{2}$ (Jeffreys, 1946), which is known as Jeffreys' prior, and will be considered in detail in Chapter 6. The second is $\alpha(\boldsymbol{i}) = \frac{1}{|I|}$ (Perks, 1947), which has the appealing interpretation of a single prior observation distributed throughout the table, and is applied to various examples throughout the thesis.

This interpretation of a Dirichlet distribution, via prior samples, makes it an attractive prior for use in the analysis of log-linear models. A natural extension of the Dirichlet distribution to decomposable graphical models is the hyper Dirichlet distribution, where each vector of clique marginal probabilities is distributed as Dirichlet. In this thesis, we consider another extension of the Dirichlet distribution, the conditional Dirichlet distribution, obtained from a saturated distribution by conditioning on constraints which determine a particular log-linear model. Consideration of this distribution is a major component of

Chapter 3.

2.2 Normal Distribution

As defined in section 1.2, the log-linear model parameters are unconstrained and allowed to take any real value, so that $\beta \in \mathcal{R}^p$. A natural prior distribution for these parameters may therefore be multivariate normal, *i.e.* $\beta \sim N(\mu, \Sigma)$, for suitable mean μ and variance Σ. The use of such a distribution was first investigated by Good (1956), though this approach was motivated by the desire to obtain smoothed estimates for cell probabilities with small observed frequencies, an idea further developed by Lenoard (1975) and Laird (1978). The purpose of Knuiman and Speed (1988) was to use a Normal distribution to effectively encapsulate prior information into the analysis of contingency tables.

Their approach used a multivariate Normal prior for all parameters together, and as such allowed separate specification of prior information for each log-linear model main effect or interaction term. However, they found the use of such a prior resulted in a generally intractable posterior distribution, and so developed a measure of posterior dispersion based on the curvature of the log of the posterior density at its mode.

Posterior inference using Normal priors, based on Markov chain Monte Carlo sampling, is possible following results by Dellaportas and Smith (1993). They present a method for sampling from a wide range of generalised linear models using Gibbs sampling. Their Gibbs sampler is based on the adaptive rejection sampling method proposed by Gilks and Wild (1992), which is a technique for sampling from any log-concave univariate probability density function, and is described in detail in Chapter 4. Forster and Skene (1994) used a similar Gibbs sampler to obtain posterior samples, for multinomial data, for prior distributions from the $\mathcal{A}$ family. This is a class of distributions introduced by Aitchison (1985), which includes the logistic Normal and Dirichlet distributions.

The use of Normal distributions by Albert (1996) in a model selection context will be reviewed later in this Chapter, and the use of Normal distributions by Dellaportas and Forster (1999) is explained in Chapter 7 in the context of unit information priors.

2.3 Graphical Models

The use of graphs to represent the pattern of associations in statistical models was introduced in section 1.2.2. Such use dates back to Wright (1921), but more recently it was Darroch, Lauritzen and Speed (1980) who used graphs in contingency table analysis, defining the subset of the hierarchical log-linear models known as graphical models.

Early adopters of methods within a Bayesian framework were Spiegelhalter and Lauritzen (1990), and Dawid and Lauritzen (1993). Madigan and York (1995) presented a comprehensive discussion on Bayesian graphical models for a variety of discrete data problems. They first considered the problem of double sampling, in particular an example analysed by Lie et al (1994), and analysed within a Bayesian graphical model framework by York et al (1995). Graphical modelling was shown to allow prior information to be effectively incorporated into the analysis, and model uncertainty properly accounted for. Posterior analyses were more straightforward than those of Lie et al, and complex models could be considered without difficulty.

A second use of graphical modelling which they considered was in closed population estimation. They considered an example previously analysed by Fienberg (1972) and Bishop et al (1975) using log-linear models. However, these analyses failed to effectively deal with prior knowledge of the population size, and prior knowledge about covariates was also difficult to encapsulate, in particular with missing values. Decomposable undirected graphical models were shown to lead to more effective analyses, allowing full use of prior knowledge based on

well-understood quantities, and accounting for model uncertainty.

One distribution defined solely for decomposable graphical models, the hyper Dirichlet distribution, admits straightforward inferences due to the easy implementation of Monte Carlo methods in such cases. Marginal inference for a particular model is straightforward, and described in section 3.2.2. Also, model comparison may be performed using calculations local to single cliques (Madigan and Raftery, 1994).

2.4 Model Uncertainty

2.4.1 Bayes Factors

The Bayes factor was defined in section 1.4.2 as a measure of evidence in favour of model m_1 against model m_2. It is given by the expression

$$B_{12} = \frac{\int f(\boldsymbol{n}|m_1, \boldsymbol{\theta}_{m_1}) f(\boldsymbol{\theta}_{m_1}|m_1) d\boldsymbol{\theta}_{m_1}}{\int f(\boldsymbol{n}|m_2, \boldsymbol{\theta}_{m_2}) f(\boldsymbol{\theta}_{m_2}|m_2) d\boldsymbol{\theta}_{m_2}} \tag{2.2}$$

In order to interpret a Bayes factor B_{12}, a commonly used method is to consider the value of twice its natural logarithm, as this is on a similar scale to the classical likelihood ratio (deviance) statistic. Indeed, the Bayes factor is often seen is a Bayesian version of this classical statistic (Berger, 1985).

Several authors have offered suggestions as to how to interpret the value of the Bayes factor. Table 2.1 is based on figures suggested by Raftery (1996) and Kass and Raftery (1995), which are in turn based on suggestions of Jeffreys (1961).

Note that the choice of prior distribution is especially crucial in a model-selection problem with vague prior information, due to the sensitivity of the Bayes factor in such cases. In general, prior distributions which are excessively diffuse may tend to favour simpler models.

$2 \log B_{12}$	Evidence for m_1
< 0	Negative (supports m_2)
0 to 2	Low
2 to 6	Positive
6 to 10	Strong
> 10	Overwhelming

Table 2.1: Interpretation of Bayes factors

An alternative to a vague proper prior is to choose an improper prior. For example, in a multinomial analysis, we might consider the limiting form of the Dirichlet distribution,

$$f(\boldsymbol{p}) \propto \prod_{\boldsymbol{i}} p(\boldsymbol{i})^{-1}$$

The problem with this is that the marginal likelihood (and hence Bayes factor) is only defined up to an arbitrary constant, though a method of assigning a value to this constant for certain examples is given by Spiegelhalter and Smith (1981).

A solution to this problem is to use partial Bayes factors, first introduced by Lempers (1971). Such an approach avoids the appearance of arbitrary constants (from improper priors) in the Bayes factor by partitioning the data $\boldsymbol{n}$ into two parts $\boldsymbol{n}= (\boldsymbol{n}_1, \boldsymbol{n}_2)$. The first part $\boldsymbol{n}_1$ (the training sample) is used to obtain a posterior distribution for model parameters $\boldsymbol{\theta}$, and this is then used as a prior distribution in a Bayes factor for $\boldsymbol{n}_2$. O'Hagan (1991) uses a proportion of the data for training, whereas Berger and Pericchi (1993) use a training sample of minimal size (defined as the smallest sample size m_0 which gives proper posterior distributions under both models). The problem of which units to choose for the training sample was considered by Berger and Pericchi, who suggested using all possible training samples and averaging the results or, when there are a large number of possible choices, averaging a random sample of choices. They call the resulting partial Bayes factor an intrinsic Bayes factor.

A further refinement of the partial Bayes factor was offered by O'Hagan (1995), and developed De Santis and Spezzaderri (1999). O'Hagan defines a

fractional Bayes factor, which avoids the problem of choosing a training sample. Instead, the only choice which must be made is what fraction $1-b$ of the data to use in the Bayes factor. O'Hagan advocates the choice $b = n^{-1}\max\{m_0, \log n\}$ for general use.

However, it seems imprudent to use improper priors, and unnecessary to use fractional Bayes factors, when a rich collection of proper vague priors are available, such as those based on the Dirichlet distribution, for log-linear models. Bayes factors based on such priors have been researched by several authors, for example Gunel and Dickey (1974) and Madigan and Raftery (1994), and such priors are considered in this thesis.

2.4.2 Computation

The problem with the calculation of Bayes factors is that the integrals in (2.2) are, in general, analytically intractable, though we note that exact results are possible for exponential family distributions with particular conjugate priors (DeGroot 1970). Therefore we must, in general, employ numerical methods to evaluate such integrals. Computation of marginal densities is a widely researched topic, particularly in the context of Bayesian model selection (for example Kass and Raftery, 1995) and many methods are available.

Albert (1996) presented a Bayesian procedure for the selection of Poisson log-linear models using mixtures of multivariate Normal distributions to model prior opinion. His method partitioned the parameter vector $\boldsymbol{\beta}$ into subsets $\boldsymbol{\beta} = (\boldsymbol{\eta}, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_s)$, where the elements of $\boldsymbol{\eta}$ are non-zero, but the elements of $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_s$ may be zero. A Normal distribution was then assigned to $\boldsymbol{\beta}$ with mean $\mathbf{0}$ and variance matrix Σ, where Σ^{-1} has a block-diagonal structure of multiples of identity matrices, with zeros corresponding to $\boldsymbol{\eta}$ and a single dispersion parameter P_i for each $\boldsymbol{\beta}_i$. Such prior distributions model prior beliefs for each of the 2^s possible models. Hypotheses setting $\boldsymbol{\beta}_i = 0$ correspond to letting P_i tend

to infinity, whereas hypotheses for non-zero β_i values require a choice to be made for P_i. This choice is not arbitrary, as different values will have a pronounced effect on the Bayes factor.

Albert's proposal was to place a prior on P_i, motivated by the approach of Good (1976). Following applications to examples involving two- and three-way contingency tables, he suggested that P_i should have a $gamma(\frac{\nu_i}{2}, \frac{b_i^2\nu_i}{2})$ distribution, where the choice of b_i depends on prior information, and ν_i may vary, though his advocated choice $\nu_i = 1$ corresponds to a set of Cauchy distributions.

For cases where the parameters P_i are known, posterior model probabilities are available using a Taylor series expansion and applying Laplace's method for integrals. The resulting approximation was found to be very accurate.

In the more general case where each P_i has an associated probability distribution, numerical integration techniques are needed to determine the posterior model probabilities. This is reasonably straightforward for examples in small dimensions, and Albert gave an appropriate expression, together with an iterative extension appropriate in the more general case. Both were found to produce accurate results.

A more general application of Laplace's method to generalised linear models was investigated by Raftery (1996), and this work is reviewed in Chapter 5. Alternative approaches for calculating marginal likelihoods are available which use Monte Carlo samples, and a review of such work by Diciccio, Kass, Raftery and Wasserman (1997) is also in Chapter 5. A particular example of such a method used in this thesis is bridge sampling, which is applied in section 5.4.2 to the conditional Dirichlet distribution. Methods of estimating model probabilities using Markov chain Monte Carlo approaches are reviewed in section 2.4.4.

2.4.3 Model Averaging

Model averaging was introduced in section 1.4.2, where the posterior distribution of a quantity of interest, say ϕ, was given by

$$f(\phi|\boldsymbol{n}) = \frac{\sum_k f(\phi|\boldsymbol{n}, m_k) f(\boldsymbol{n}|m_k) f(m_k)}{\sum_l f(\boldsymbol{n}|m_l) f(m_l)} \tag{2.3}$$

This is a good way of accounting for model uncertainty, which involves the posterior probabilities for each potential model. It allows all potential models to be considered, rather than the seemingly ad hoc approach of conditioning on a single model, selected using a sequence of pairwise model comparisons.

A solution proposed by Madigan and Raftery (1994), known as Occam's window, was to eliminate many of the models from (2.3). Their approach first eliminates any model with probability much smaller than the most probable model, then any model with probability lower than a model nested within it. They gave two algorithms for identifying a set of potentially acceptable models. Use of this strategy typically reduces the number of models to less than 100, and often to under 10 (Hoeting, Madigan, Raftery and Volinsky, 1999).

Suggested prior distributions were given by Madigan and Raftery for applications to both directed and undirected decomposable models. Both these are based on the Dirichlet distribution – specifically they used hyper Dirichlet distributions. They also presented a method for elicitation of such priors, appropriate for application in expert systems with a potentially large number of variables, ensuring consistency between the directed and undirected approaches.

Application was made to identical datasets using both types of graphical representation, and similar results obtained. However, in general, the undirected approach was preferred as an ordering of the variables is then not needed *a priori*. It is straightforward to proceed via this method as comparisons between models differing by a single edge are possible using calculations local to single cliques.

2.4.4 Markov Chain Monte Carlo Methods

An alternative (and more commonly used) approach to dealing with large numbers of competing models is to use Markov chain Monte Carlo methods. This allows all possible models to be considered, as opposed to the 'Occam's window' approach which excludes many models from the analysis. A Markov chain is constructed so as to obtain a sample from $f(m, \boldsymbol{\theta}_m|\boldsymbol{n})$, and the posterior model probabilities $f(m|\boldsymbol{n})$ then estimated from this using the Monte Carlo sample proportions.

The 'reversible jump' method of sampling was introduced by Green (1995). He presented a general description of the method, together with a particular implementation which may be adopted for log-linear models. This method was adapted by Dellaportas and Forster (1999) and applied to several classes of log-linear models. A brief description of Green's method, and a review of the work by Dellaportas and Forster is given below.

Suppose we have M models, and that $m \in M$. Let the state of the Markov chain at time t be denoted $\left(m^{(t)}, \boldsymbol{\theta}_m^{(t)}\right)$. At each step, there are different possible types of move.

Suppose that move type p is a proposed move to m', a model with a single additional term and with parameter vector $\boldsymbol{\theta}_m'$ of higher dimension than $\boldsymbol{\theta}_m^{(t)}$. This parameter vector is constructed by generating a vector $\mathbf{u}$ which has dimension equal to the difference in dimensions of the two models, using proposal distribution $q_p(\mathbf{u})$, and setting $\boldsymbol{\theta}_m' = (\boldsymbol{\theta}_m^{(t)}, \mathbf{u})$. The 'reverse' method is used for a move of type p to a model with a single term removed (*i.e.* discard $\mathbf{u}$). Moves are accepted with probabilities which take similar forms to those in Metropolis Hastings sampling.

An alternative proposal is the 'null' move, where $m^{(t+1)} = m^{(t)}$. Model parameters may, however, be changed. A Markov chain constructed using all the above will have equilibrium distribution $f(m, \boldsymbol{\theta}_m|\boldsymbol{n})$.

Dellaportas and Forster applied this method to the set of general log-linear interaction models without hierarchical constraints. In this case, the model parameters $\boldsymbol{\theta}_m$ are those model parameters $\boldsymbol{\beta}$ which are non-zero in the model concerned. Their implementation considers each of the model terms as possible move types, so that log-linear model terms and their corresponding parameters are continually added to, and removed from, the model. They allowed each non-null move to be made with equal probability, which simplifies the expression for the acceptance probabilities.

Dellaportas and Forster's null move is to use a Gibbs sampling method to obtain a sample for the elements of $\boldsymbol{\theta}_m$. They used a multivariate normal distribution for $q_p(\mathbf{u})$, whose mean and variance was chosen, by investigating a 'pilot chain' of null moves, to optimise the performance of the procedure. They also assumed all models to have equal probability *a priori.*, and set the probability of the null move $r = 0.25$. They found the method performed well in various applications.

A further application of this theory was made to hierarchical log-linear models, by constructing the move probabilities so that only moves to neighbouring hierarchical models are proposed. This is possible as any hierarchical model may be reached from another via only models which are themselves hierarchical. Models where any main effect is absent were also excluded. The method was also applied to graphical models, by considering at each stage of the chain the removal of, or addition of, an edge to the graph. Note that in this instance multiple log-linear model terms may be added or deleted at each stage, depending on the edge. Finally, application was made to decomposable models, using the same method as for graphical models but with zero probabilities $f(m)$ for non-decomposable models.

This application is similar to the method developed by Madigan and York (1995), who used a hyper Dirichlet prior distribution for the model parameters,

instead of a normal distribution. Their use of a hyper Dirichlet distribution allows the Bayes factor to be exactly and easily computed at each step, using calculations local to single cliques (Madigan and Raftery, 1994, Dawid and Lauritzen, 1989). Markov chain Monte Carlo sampling may be performed for the $f(m)$ margin directly (*i.e.* there is no need to sample from $\boldsymbol{\theta}_m$). Their method proved extremely effective, with runs of 10,000 or less typically adequate.

Alternative MCMC methods of calculating posterior model probabilities are available, many of which can be formulated as special cases of the method of reversible jump. These include independence sampling, where the proposed model is not allowed to depend on the current model, and a method developed by Carlin and Chib (1995) based on Gibbs sampling. However, the former does not, in general, produce good results, and the latter has associated computational difficulties (Dellaportas, Forster and Ntzoufras, 2001).

Raftery, Madigan and Hoeting (1997) applied both Occam's window and MCMC methods to model averaging for linear regression models, finding both to provide satisfactory results. However, Occam's window was better when the aim was to investigate the relationships between the variables, and MCMC methods better for predictive analysis and for obtaining the posterior distribution of a particular quantity.

Chapter 3

Priors for Log-Linear Model Parameters

An important choice in the analysis of log-linear models is that of the prior distribution $f(\boldsymbol{p})$. The choice of prior distribution is especially crucial in a model selection problem, due to the sensitivity of the Bayes factor to the choice of prior, whereby certain prior distributions may tend to favour particular models, for example complex models, and others may favour simpler models.

The prior distribution encapsulates the previous information about the cell probabilities $\boldsymbol{p}$, which may be obtained from expert opinion, previous data, or some other source. However, this information, even if available, is often difficult to express as a probability distribution. Hence it is often useful to perform a 'reference' analysis, where the prior distribution is called a reference, noninformative or diffuse prior, which also enables us to examine the influence of informative priors compared with the reference approach.

A number of priors exist which may be used for the log-linear model parameters, and which are suitable for a reference analysis. Several of these distributions are considered in this Chapter.

3.1 Distributions Based on the Normal Distribution

As defined in section 2.2, the log-linear model parameters may be distributed as multivariate normal. *i.e.* $\boldsymbol{\beta} \sim N(\boldsymbol{\eta}, \Sigma)$, where $\boldsymbol{\eta}$ represents the prior belief about the location of the parameters, and Σ represents the strength of this belief. For a reference analysis, the problem is whether it is possible to choose values for $\boldsymbol{\eta}$ and Σ so that this prior distribution is noninformative.

The prior on $\boldsymbol{\beta}$ induces a prior distribution on $\log \boldsymbol{p}$ (or equivalently on $\log \boldsymbol{\mu}$). Forster (1999) showed that a multivariate normal prior for $\log \boldsymbol{\mu}$ must have a certain form in order for it to be invariant to permutations of the set of levels I_γ of each factor (a sensible requirement for a reference prior). This distribution takes the form

$$\log \boldsymbol{\mu} \sim N(\delta \mathbf{1}, \sum_{a \subseteq \Gamma} \alpha_a^2 T_a)$$

where the T_a are projection matrices given by

$$T_a = \bigotimes_{\gamma \in \Gamma} \left\{ 1(\gamma \in a) \left(I_{|I_\gamma|} - \frac{1}{|I_\gamma|} J_{|I_\gamma|} \right) + 1(\gamma \notin a) \frac{1}{|I_\gamma|} J_{|I_\gamma|} \right\} \tag{3.1}$$

and I_d is a $d \times d$ identity matrix and J_d a $d \times d$ matrix of 1's. The prior distributions for the model parameters $\boldsymbol{\beta}_a$, with the exception of $\boldsymbol{\beta}_\emptyset$ (corresponding to the intercept term), are then given by

$$\boldsymbol{\beta}_a \overset{ind}{\sim} N(\mathbf{0}, \alpha_a^2 \Sigma_a) \qquad a \subseteq \Gamma$$

where

$$\Sigma_a = \frac{1}{|I|} \prod_{\gamma \in a} |I_\gamma| \bigotimes_{\gamma \in a} \left(I_{(|I_\gamma|-1)} - \frac{1}{|I_\gamma|} J_{(|I_\gamma|-1)} \right) \qquad a \subseteq \Gamma$$

The prior for $\boldsymbol{\beta}_\emptyset$ is

$$\boldsymbol{\beta}_\emptyset \sim N(\tau, \alpha_\emptyset^2)$$

for a specified value of τ. It is necessary to assume independence of the model parameters, though this is not restrictive as it seems sensible to do so if we are to perform a reference analysis.

3.2 Distributions based on the Dirichlet Distribution

3.2.1 Conditional Dirichlet Distribution

An alternative choice of prior distribution is based on the Dirichlet distribution. This was defined for the saturated model in section 2.1 as

$$f(\boldsymbol{p}) = \frac{\Gamma(\alpha)}{\prod_{\boldsymbol{i}} \alpha(\boldsymbol{i})} \prod_{\boldsymbol{i} \in I} p(\boldsymbol{i})^{\alpha(\boldsymbol{i})-1}$$

where $\alpha = \sum_{\boldsymbol{i}} \alpha(\boldsymbol{i})$.

For this saturated model, a Dirichlet prior for the cell probabilities $\boldsymbol{p}$ implies a prior for $\boldsymbol{\beta}$, any vector of log-linear model parameters. In order to determine the form of this prior, it is useful to first transform the variables as follows.

Define the reference-cell logit in the standard way (as in section 1.3)

$$\theta(\boldsymbol{i}) = \log p(\boldsymbol{i}) - \log p(\boldsymbol{i}_0)$$

The Jacobian, $|J|$, for the transformation from $\boldsymbol{p}_{\setminus \boldsymbol{i}_0}$ to $\boldsymbol{\theta}_{\setminus \boldsymbol{i}_0}$ is easy to determine, by applying the expression

$$\left|\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{p}}\right| = \left| diag\left(\frac{1}{p(\boldsymbol{i})}\right) + \frac{1}{p(\boldsymbol{i}_0)} \mathbf{1}\mathbf{1}^T \right|$$

A standard linear algebra result gives us

$$|diag(\mathbf{a}) + \mathbf{b}\mathbf{c}^T| = \left(1 + \sum_{\boldsymbol{i}} \frac{b(\boldsymbol{i})c(\boldsymbol{i})}{a(\boldsymbol{i})}\right) \prod_{\boldsymbol{j}} a(\boldsymbol{j})$$

and so, applying this, we have

$$\begin{aligned}\left|\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{p}}\right| &= \left(1 + \sum_{\boldsymbol{i} \neq \boldsymbol{i}_0} \frac{p(\boldsymbol{i})}{p(\boldsymbol{i}_0)}\right) \prod_{\boldsymbol{j} \neq \boldsymbol{i}_0} \frac{1}{p(\boldsymbol{j})} \\ &= \left(1 + \frac{1 - p(\boldsymbol{i}_0)}{p(\boldsymbol{i}_0)}\right) \prod_{\boldsymbol{j} \neq \boldsymbol{i}_0} \frac{1}{p(\boldsymbol{j})} \\ &= \prod_{\boldsymbol{j}} \frac{1}{p(\boldsymbol{j})}\end{aligned}$$

The Jacobian $|J|$ for the transformation from $\boldsymbol{p}_{\setminus \boldsymbol{i}_0}$ to $\boldsymbol{\theta}_{\setminus \boldsymbol{i}_0}$ is therefore given by

$$\begin{aligned}|J| &= \left|\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{p}}\right|^{-1} \\ &= \prod_{\boldsymbol{j}} p(\boldsymbol{j})\end{aligned}$$

This Jacobian is for the reference cell logit $\boldsymbol{\theta}_r$, but the Jacobian for the symmetric logit $\boldsymbol{\theta}_s$ is easy to determine, as we know that $\boldsymbol{\theta}_r = M\boldsymbol{\theta}_s$ for some matrix M, so that $|J|_r = |J|_s |M|$. It is straightforward to show that $|M| = m$, so that

$$|J|_s = \frac{1}{m} \prod_{\boldsymbol{j}} p(\boldsymbol{j})$$

It is now possible to write down the Dirichlet distribution for $\boldsymbol{\theta}$:

$$f(\boldsymbol{\theta}) = \frac{\Gamma(\alpha)}{\prod_{\boldsymbol{i}} \alpha(\boldsymbol{i})} \prod_{\boldsymbol{i} \in I} p(\theta(\boldsymbol{i}))^{\alpha(\boldsymbol{i})-1} |J|$$

$$
\begin{aligned}
&= \frac{\Gamma(\alpha)}{\prod_{\boldsymbol{i}} \alpha(\boldsymbol{i})} \prod_{\boldsymbol{i}\in I} p(\theta(\boldsymbol{i}))^{\alpha(\boldsymbol{i})-1} \prod_{\boldsymbol{j}\in I} p(\theta(\boldsymbol{j})) \\
&= \frac{\Gamma(\alpha)}{\prod_{\boldsymbol{i}} \alpha(\boldsymbol{i})} \prod_{\boldsymbol{i}\in I} p(\theta(\boldsymbol{i}))^{\alpha(\boldsymbol{i})} \\
&= \frac{\Gamma(\alpha)}{\prod_{\boldsymbol{i}} \alpha(\boldsymbol{i})} \prod_{\boldsymbol{i}\in I} \frac{e^{\theta(\boldsymbol{i})\alpha(\boldsymbol{i})}}{\left(\sum e^{\theta(\boldsymbol{i})}\right)^{\alpha(\boldsymbol{i})}} \\
&= \frac{\Gamma(\alpha)}{\prod_{\boldsymbol{i}} \alpha(\boldsymbol{i})} \frac{\prod_{\boldsymbol{i}\in I} e^{\theta(\boldsymbol{i})\alpha(\boldsymbol{i})}}{\left(\sum e^{\theta(\boldsymbol{i})}\right)^{\alpha}}
\end{aligned}
$$

where $\alpha = \sum_{\boldsymbol{i}} \alpha(\boldsymbol{i})$.

A particular log-linear model sets $\boldsymbol{\theta} = X\boldsymbol{\beta}$, for suitable $(n \times p)$ design matrix X, where p is the number of parameters in the model. Therefore, a distribution for $\boldsymbol{\beta}$ in the saturated model is obtained by a simple linear transformation, involving the $((n-1) \times p)$ matrix X^* which is the design matrix X excluding the row for $\boldsymbol{i}_0$. Note that this is possible since the $\boldsymbol{\theta}$ vector is linearly dependent, satisfying $\theta(\boldsymbol{i}_0) = 0$.

We therefore obtain a distribution for $\boldsymbol{\beta}$

$$
\begin{aligned}
f(\boldsymbol{\beta}) &\propto \frac{\prod_{\boldsymbol{i}\in I} e^{\theta(\boldsymbol{i},\boldsymbol{\beta})\alpha(\boldsymbol{i})}}{\left(\sum e^{\theta(\boldsymbol{i},\boldsymbol{\beta})}\right)^{\alpha}} \\
&\propto \frac{\prod_{\boldsymbol{i}\in I} e^{\alpha(\boldsymbol{i})\sum_j x(\boldsymbol{i},j)\beta_j}}{\left(\sum_{\boldsymbol{i}} e^{\sum_j x(\boldsymbol{i},j)\beta_j}\right)^{\alpha}}
\end{aligned}
\tag{3.2}
$$

We shall define the *Conditional Dirichlet* distribution for a particular log-linear model as that distribution obtained from expression (3.2) by conditioning on certain β_j terms to be zero. A more formal definition proceeds as follows.

To obtain a conditional Dirichlet distribution for a particular (non-saturated) model, we first partition $\boldsymbol{\beta}$ into those effects in the model ($\boldsymbol{\beta}_1$) and those not in the model ($\boldsymbol{\beta}_0$), and condition on $\boldsymbol{\beta}_0 = \mathbf{0}$. We also partition X into X_1 containing

the columns corresponding to the parameters in $\boldsymbol{\beta}_1$, and X_0 containing the others. We may order the columns in X such that X_1 precedes X_0. Then the conditional Dirichlet distribution for $\boldsymbol{\beta}$ is obtained from expression (3.2) by summing over non-zero β_j only:

$$f(\boldsymbol{\beta}) \propto \frac{\prod_{\boldsymbol{i}\in I} e^{\alpha(\boldsymbol{i})\sum_{j:\beta_j\neq 0} x(\boldsymbol{i},j)\beta_j}}{\left(\sum_{\boldsymbol{i}} e^{\sum_{j:\beta_j\neq 0} x(\boldsymbol{i},j)\beta_j}\right)^{\alpha}}$$

Extensive investigation of this distribution for a wide range of models showed it to be analytically intractable in general. Normalising constants were difficult (or impossible) to evaluate, and it also proved impossible (in general) to obtain marginal likelihoods (necessary for the calculation of Bayes factors) from the induced posterior distributions. Therefore, posterior analysis using this class of priors is not straightforward. However, a hypothesised relationship between this distribution and the hyper Dirichlet distribution is considered in the next section, and such a relationship would enable this intuitively attractive distribution to be more readily used in practice.

Borel Paradox

The method of conditioning used to obtain the conditional Dirichlet distribution can be thought of as conditioning on a particular log-linear model. Care must be taken, however, as the prior distribution obtained through conditioning on a set of complex constraints is not invariant under general reparameterisation of those constraints. This is known as the Borel paradox. However, the prior distributions induced under various parameterisations may be shown to be related by the Borel-Kolmogorov dependence formula (Gunel and Dickey, 1974):

Suppose we have two parameterisations of a model, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ and $\widetilde{\boldsymbol{\theta}} = (\widetilde{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_2)$, a submodel H which specifies $\boldsymbol{\theta}_1$ only, and an equivalent submodel $\widetilde{H}$ concerning $\widetilde{\boldsymbol{\theta}}_1$ only. Then the relationship between the priors for $\boldsymbol{\theta}_2$ under the submodels obtained by conditioning is given by the Borel-Kolmogorov depen-

dence formula

$$f(\boldsymbol{\theta}_2|H) \propto f(\boldsymbol{\theta}_2|\widetilde{H}) \left|\frac{\partial \boldsymbol{\theta}_1}{\partial \widetilde{\boldsymbol{\theta}}_1}\right|$$

Therefore the induced prior is the same if and only if the Jacobian $\left|\frac{\partial \boldsymbol{\theta}_1}{\partial \widetilde{\boldsymbol{\theta}}_1}\right|$ is constant in $\boldsymbol{\theta}_2$.

As an example, consider a 2×2 contingency table. Under the saturated model, the cell probabilities $\boldsymbol{p}$ are distributed as $Dirichlet(\boldsymbol{\alpha})$. The independence model may be specified by a number of equivalent constraints. These include $\frac{p(11)}{p(1+)} - \frac{p(21)}{p(2+)} = 0$, $\frac{p(11)p(2+)}{p(1+)p(21)} = 1$ and $\frac{p(11)p(22)}{p(12)p(21)} = 1$. (Note that $p(1+) = \sum_{\boldsymbol{i}:i_1=1} p(\boldsymbol{i})$.) Although each of the constraints define the same independence model, the marginal distributions obtained for $p(1+)$ by conditioning on each constraint are not identical. Indeed, conditioning on $\frac{p(11)}{p(1+)} - \frac{p(21)}{p(2+)} = 0$, $\frac{p(11)p(2+)}{p(1+)p(21)} = 1$ and $\frac{p(11)p(22)}{p(12)p(21)} = 1$ in turn results in $Beta(\alpha(1+) - 1, \alpha(2+) - 1)$, $Beta(\alpha(1+), \alpha(2+) - 1)$ and $Beta(\alpha(1+), \alpha(2+))$ distributions respectively for $p(1+)$.

Now consider the Bayes factor for comparing the saturated model (S) and independence model (I), based only on marginal data $n(1+)$. This is given by the expression

$$\frac{f(n(1+)|S)}{f(n(1+)|I)} = \frac{\int p(1+)^{n(1+)} p(2+)^{n(2+)} f(p(1+)|S) dp(1+)}{\int p(1+)^{n(1+)} p(2+)^{n(2+)} f(p(1+)|I) dp(1+)}$$

Hence, for sensible inference, the marginal prior density of $p(1+)$ should be the same under both models. We know that, under the saturated model, $p(1+)$ is distributed as $Beta(\alpha(1+), \alpha(2+))$. We should therefore define the independence model using the constraint $\frac{p(11)p(22)}{p(12)p(21)} = 1$ to obtain a consistent marginal density in this case. It will be shown that the conditional Dirichlet distribution results in such consistent marginal densities.

3.2.2 Hyper Dirichlet Distribution

A sub-class of models which admit straightforward analyses are decomposable log-linear models. We may parameterise these models directly in terms of the clique marginal cell probabilities. The *hyper Dirichlet* distribution was proposed by Dawid and Lauritzen (1993) as a conjugate prior distribution for the parameters of a decomposable log-linear model. A useful feature of this prior distribution is that the resulting posterior is also hyper Dirichlet and so may be decomposed by cliques, enabling straightforward analyses.

The hyper Dirichlet prior is defined for a decomposable model represented by an undirected graph as follows: Each clique must have an independent Dirichlet distribution, and the marginal distributions on overlapping portions of cliques must be consistent regardless of the clique from which they are derived. One way of generating such a distribution is by deriving the prior distributions on the cliques as the marginal distributions from a Dirichlet distribution on the full set of probabilities.

For a directed graph, we know that a cell probability may be written

$$p(\boldsymbol{i}) = \prod_{\gamma} P(\gamma = i_\gamma | pa(\gamma) = \boldsymbol{i}_{pa(\gamma)})$$

The hyper Dirichlet distribution places independent Dirichlet distributions on each set of conditional probabilities corresponding to a particular $\gamma, \boldsymbol{i}_{pa(\gamma)}$. In order to ensure hyper-consistency in this instance, we require that if a variable appears in multiple sets of conditional probabilities, then its marginal density is the same regardless of from where it is derived.

Marginal inference from a hyper Dirichlet distribution is straightforward. Using the directed representation, we can write down the hyper Dirichlet distribution as a product of independent Dirichlet distributions. Monte Carlo samples may then be obtained from each of these distributions in turn by sampling from

independent gamma distributions and applying the result that if $z_1, z_2, ..z_p$ are independent samples from $Gamma(a_i, b)$ distributions, then $\frac{z_1}{\sum z_i}, \frac{z_2}{\sum z_i}, \ldots, \frac{z_p}{\sum z_i}$ is a sample from a $Dirichlet(a_1, a_2, \ldots, a_p)$ distribution.

3.3 Relationship between Conditional Dirichlet and Hyper Dirichlet Distributions

In this section, the hypothesis that conditional Dirichlet and hyper Dirichlet distributions are equivalent for decomposable log-linear models will be investigated. Although investigation of the form of the conditional Dirichlet density $f(\boldsymbol{\beta})$ (as determined in section 3.2) for specific models did not yield a generally tractable expression, several models did in fact highlight the hypothesised relationship. This is exemplified here, and these examples provide an introduction to, and motivation for, the general proof which follows.

Consider the 2×2 independence model (*i.e.* the model with cliques $\{A\}$ and $\{B\}$) which is represented graphically below.

Substituting into equation (3.2), an expression for $f(\boldsymbol{\beta})$ is obtained:

$$f(\beta_1, \beta_2 | \beta_3 = 0) \propto \frac{e^{(\alpha(+1)-\alpha(+2))\beta_1}}{(e^{\beta_1} + e^{-\beta_1})^{\alpha}} \frac{e^{(\alpha(1+)-\alpha(2+))\beta_2}}{(e^{\beta_2} + e^{-\beta_2})^{\alpha}}$$

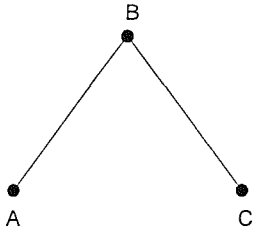
where, for example, $\alpha(+1) = \sum_{\boldsymbol{i}:i_2=1} \alpha(\boldsymbol{i})$. We therefore have a factorisation into distributions on each clique. These distributions may then be used to derive marginal distributions for $p(1+)$ and $p(+1)$:

$$f(p(+1)) \propto p(+1)^{\alpha(+1)-1}(1 - p(+1))^{\alpha(+2)-1}$$

$$f(p(1+)) \propto p(1+)^{\alpha(1+)-1}(1 - p(1+))^{\alpha(2+)-1}$$

which shows that the margins for A and B each have Beta distributions. For example, the margin for B, $p(1+)$, has a $Beta(\alpha(1+), \alpha(2+))$ distribution, which is the same as the distribution obtained by marginalising from the saturated model. Hence the conditional distributions produced by conditioning on the model correspond to the marginal distributions for the cliques, which is consistent with the hyper Dirichlet distribution. Thus for this particular log-linear model, the conditional Dirichlet distribution is identical to the equivalent hyper Dirichlet distribution.

A similar investigation was performed for the $2 \times 2 \times 2$ model with cliques $\{A, B\}$ and $\{B, C\}$, represented graphically by



However, in this case, an expression was obtained for $f(\boldsymbol{\beta})$ which does not readily factorise. A transformation was therefore necessary, and a set of λ_i's were defined so that λ_1 and λ_2 correspond to the logits of $P(A|B=1)$ and $P(A|B=2)$ respectively, λ_3 and λ_4 to the logits of $P(C|B=1)$ and $P(C|B=2)$, and λ_5 to the logit of $P(B)$. Using these, an expression for $f(\boldsymbol{\lambda})$ is obtained:

$$f(\boldsymbol{\lambda}) \propto \frac{e^{\alpha(+21)\lambda_1+\alpha(+22)\lambda_2+\alpha(2+1)\lambda_3+\alpha(2+2)\lambda_4+\alpha(++2)\lambda_5}}{(1+e^{\lambda_1})^{\alpha(++1)}(1+e^{\lambda_2})^{\alpha(++2)}(1+e^{\lambda_3})^{\alpha(++1)}(1+e^{\lambda_4})^{\alpha(++2)}(1+e^{\lambda_5})^{\alpha(++1)}}$$

A factorisation into independent distributions is apparent, with each of $P(A|B=1)$, $P(A|B=2)$, $P(C|B=1)$, $P(C|B=2)$ and $P(B)$ having independent Dirichlet distributions. Distributions for the clique margins may also be derived, as follows:

Consider the clique $\{A, B\}$, and hence parameters λ_1, λ_2 and λ_5. From the

above expression, we have

$$f(\lambda_1, \lambda_2, \lambda_5) \propto \frac{e^{\alpha(+21)\lambda_1 + \alpha(+22)\lambda_2 + \alpha(++2)\lambda_5}}{(1+e^{\lambda_1})^{\alpha(++1)}(1+e^{\lambda_2})^{\alpha(++2)}(1+e^{\lambda_5})^{\alpha(++1)}}$$

The identity

$$\log\left(\frac{p(211)}{p(111)}\right) = \log\left(\frac{1-p(1++)}{p(1++)}\right)$$

may be used together with the definition of λ_1 to give $p(11+) = \frac{1}{1+e^{\lambda_1}}$. Similar expressions may be obtained for $p(12+)$ and $p(21+)$. These may then be used to obtain the marginal distribution

$$\begin{aligned} f(p(11+), p(12+), p(21+)) \quad \propto \quad & p(11+)^{\alpha(11+)-1} p(12+)^{\alpha(12+)-1} p(21+)^{\alpha(21+)-1} \times \\ & (1-p(11+)-p(12+)-p(21+))^{\alpha(22+)-1} \end{aligned}$$

A similar method results in a marginal distribution corresponding to the clique $\{B, C\}$

$$\begin{aligned} f(p(+11), p(+12), p(+21)) \quad \propto \quad & p(+11)^{\alpha(+11)-1} p(+12)^{\alpha(+12)-1} p(+21)^{\alpha(+21)-1} \times \\ & (1-p(+11)-p(+12)-p(+21))^{\alpha(+22)-1} \end{aligned}$$

Finally, the distribution for the margin corresponding to variable B may be similarly obtained

$$f(p(+1+)) \propto p(+1)^{\alpha(+1+)-1}(1-p(+1+))^{\alpha(+2+)-1}$$

As can be seen above, the margin corresponding to variable B has a $Beta(\alpha(+1+), \alpha(+2+))$ distribution, and the margins in cliques $\{A, B\}$ and $\{B, C\}$ have $Dirichlet(\alpha(11+), \alpha(12+), \alpha(21+), \alpha(22+))$ and $Dirichlet(\alpha(+11), \alpha(+12), \alpha(+21), \alpha(+22))$ distributions respectively. These distributions, and those presented earlier for the conditional probabilities, are equivalent to those

obtained by marginalising from the saturated model. Hence, these results show agreement with the hyper Dirichlet distribution, and so for this model the conditional Dirichlet distribution is identical to the equivalent hyper Dirichlet distribution.

The motivation provided by these examples was further enhanced by using Gibbs sampling to obtain prior samples corresponding to a wide variety of more complicated models.

General Proof of the Equivalence of Conditional and Hyper Dirichlet Distributions

The hypothesised equivalence of these two classes of distributions will be proved by defining an association between two different parameterisations of the same model.

As shown in section 3.2.1, using the reference-cell logit, a distribution for θ results which is of the form

$$f(\theta) \propto \prod_{i \in I} [p(\theta(i))]^{\alpha(i)} \tag{3.3}$$

and that, by conditioning on a particular log-linear model with design matrix X, a distribution for β is obtained

$$\begin{aligned} f(\beta) &\propto \prod_{i \in I} [p(i, \beta)]^{\alpha(i)} \\ &\propto \frac{\prod_{i \in I} e^{\alpha(i) \sum_j x(i,j)\beta_j}}{\left(\sum_i e^{\sum_j x(i,j)\beta_j}\right)^{\alpha}} \end{aligned}$$

This distribution is the conditional Dirichlet distribution for the specified model. In order to show the equivalence of this distribution and the hyper Dirichlet distribution, it is necessary to define a one to one relationship between

the log-linear model parameters and a set of conditional probabilities. This may be done as follows:

As the model is a decomposable log-linear model, it may be represented by a directed or undirected graph. Let us construct the directed version of the graph which represents this model, and obtain a perfect numbering of vertices (as explained in section 1.2.3). The set of factors is denoted by Γ, and for each factor $\gamma \in \Gamma$, I_γ is the set of levels of this factor. We may obtain a perfect numbering of Γ, which assigns an order to this set, which without loss of generality will now be denoted by $\Gamma = \{1, 2, \ldots, m\}$ where $m = |\Gamma|$.

As the model is decomposable, we know that any cell probability may be directly expressed as a function of the marginal probabilities of the cliques of the model. This definition is directly applicable to an undirected graphical representation of a log-linear model. However, in this case, having constructed a directed representation of this model, and an associated perfect ordering, we can use an equivalent expression of a cell probability $p(\boldsymbol{i})$ in terms of conditional probabilities, given by

$$p(\boldsymbol{i}) = \prod_{\gamma=1}^{m} P(\gamma = i_\gamma | pa(\gamma) = i_{pa(\gamma)})$$

where $\max\{pa(\gamma)\} < \gamma$ for all γ. Note that, for clarity, throughout this proof bold type will not necessarily be used to represent vectors; the levels and dimension of quantities should be apparent by subscripts, where necessary.

Any log-linear model term corresponds to a subset of Γ. A one to one correspondence between the log-linear model terms and the logits of the probabilities defined above may now be explicitly defined. For each log-linear model term, $T \subseteq \Gamma$, let $t = \max(T)$. There are two possible cases, and each one admits a slightly different association.

(i) If $T \setminus t = pa(t)$, associate $\beta_T(i_T)$ with the logit of the conditional probability $P(t = i_t | T \setminus t = i_{T \setminus t})$.

(ii) If $T \setminus t \subset pa(t)$, then associate $\beta_T(i_T)$ with the logit of the probability $P(t = i_t | T \setminus t = i_{T \setminus t}, pa(t) \setminus (T \setminus t) = \mathbf{1})$.

The relationship defined above can be seen to be a one to one relationship by the following argument:

There are clearly the same number of parameters in each case, as we have the same model correctly parameterised in two different ways. It remains to show that the conditional probabilities to which we are associating are all appropriate for the decomposable model concerned. This is a direct consequence of the conditional independences implied by log-linear models. It is clear that $T \setminus t$ must be a subset of $pa(t)$ – otherwise, there is some $s \notin pa(t)$, such that $s < t$ and t and s are not conditionally independent, given $pa(t)$.

We shall now define the logits of the conditional probabilities. It was shown in section 1.3 that, using logit transformation, a cell probability may be written in general as

$$p(\boldsymbol{i}, \boldsymbol{\theta}) = \frac{e^{(X\beta)(\boldsymbol{i})}}{\sum e^{(X\beta)(\boldsymbol{i})}}$$

Now consider the reference-cell logit with respect to $\gamma = 1$ of the conditional probability $p(\gamma = i_\gamma | pa(\gamma) = i_{pa(\gamma)})$, which we shall denote by $\phi_\gamma(i_\gamma, i_{pa(\gamma)})$. Using the above equation, this is given by

$$\begin{aligned}
\phi_\gamma(i_\gamma, i_{pa(\gamma)}) &= \log \left[\sum_{j_\gamma = i_\gamma, j_{pa(\gamma)} = i_{pa(\gamma)}} p(\boldsymbol{j}, \boldsymbol{\theta}) \right] - \log \left[\sum_{j_\gamma = 1, j_{pa(\gamma)} = i_{pa(\gamma)}} p(\boldsymbol{j}, \boldsymbol{\theta}) \right] \\
&= \log \left[\sum_{j_\gamma = i_\gamma, j_{pa(\gamma)} = i_{pa(\gamma)}} \exp\{(X\beta)(\boldsymbol{j})\} \right] - \\
&\qquad \log \left[\sum_{j_\gamma = 1, j_{pa(\gamma)} = i_{pa(\gamma)}} \exp\{(X\beta)(\boldsymbol{j})\} \right]
\end{aligned}$$

The prior distribution corresponding to the conditional Dirichlet case is parameterised in terms of the log-linear model parameters $\boldsymbol{\beta}$, and written

$$f(\boldsymbol{\beta}) \propto \prod_{\boldsymbol{i}} [p(\boldsymbol{i}, \boldsymbol{\beta})]^{\alpha(\boldsymbol{i})}$$

Using the expression

$$p(\boldsymbol{i}) = \prod_{\gamma} P(\gamma = i_\gamma | pa(\gamma) = i_{pa(\gamma)})$$

we obtain

$$\begin{aligned} f(\boldsymbol{\beta}) &\propto \prod_{\boldsymbol{i}} \left[\prod_{\gamma} P(\gamma = i_\gamma | pa(\gamma) = i_{pa(\gamma)}) \right]^{\alpha(\boldsymbol{i})} \\ &\propto \prod_{\boldsymbol{i}} \prod_{\gamma} P(\gamma = i_\gamma | pa(\gamma) = i_{pa(\gamma)})^{\alpha(\boldsymbol{i})} \\ &\propto \prod_{\gamma} \prod_{i_\gamma, i_{pa(\gamma)}} P(\gamma = i_\gamma | pa(\gamma) = i_{pa(\gamma)})^{\alpha_{\gamma, pa(\gamma)}} \end{aligned}$$

still as a function of $\boldsymbol{\beta}$. This may be written in terms of the logits $\phi_\gamma(i_\gamma, i_{pa(\gamma)})$, and is written as a function of these parameters

$$f(\boldsymbol{\phi}) \propto \prod_{\gamma} \prod_{i_\gamma, i_{pa(\gamma)}} \frac{\exp\left\{\alpha_{\gamma, pa(\gamma)} \phi_\gamma(i_\gamma, i_{pa(\gamma)})\right\}}{\left(\sum_\gamma \exp\left\{\phi_\gamma(i_\gamma, i_{pa(\gamma)})\right\}\right)^{\alpha_{\gamma, pa(\gamma)}}} |J| \tag{3.4}$$

where $|J|$ is the Jacobian for the transformation from $\boldsymbol{\beta}$ to $\boldsymbol{\phi}$. We hypothesise that $|J|$ is independent of model parameters. Indeed, we hypothesise that J is upper triangular, and that all the terms on the diagonal are equal to 1, hence that $|J| = 1$.

Differentiating the logits $\phi_\gamma(i_\gamma, i_{pa(\gamma)})$ with respect to log-linear model param-

eter β_j corresponding to the levels k_T of the log-linear model term T gives

$$\begin{aligned}\frac{\partial \phi_\gamma(i_\gamma, i_{pa(\gamma)})}{\partial \beta_j} &= \frac{\sum_{j_\gamma = i_\gamma, j_{pa(\gamma)} = i_{pa(\gamma)}} x(j, T, k_T) \exp(X\beta)(j)}{\sum_{j_\gamma = i_\gamma, j_{pa(\gamma)} = i_{pa(\gamma)}} \exp(X\beta)(j)} - \\ &\qquad \frac{\sum_{j_\gamma = 1, j_{pa(\gamma)} = i_{pa(\gamma)}} x(j, T, k_T) \exp(X\beta)(j)}{\sum_{j_\gamma = 1, j_{pa(\gamma)} = i_{pa(\gamma)}} \exp(X\beta)(j)} \\ &= P(T = k_T | \gamma = i_\gamma, pa(\gamma) = i_{pa(\gamma)}) - \\ &\qquad P(T = k_T | \gamma = 1, pa(\gamma) = i_{pa(\gamma)}) \end{aligned} \tag{3.5}$$

as $x(j, T, k_T) = 1$ if $j_T = k_T$, or 0 otherwise.

In order to show that J is upper triangular, we must determine the value of (3.5) for a given T. We shall consider various cases for the model term T.

Define two sets A_γ and B_γ as follows: let the set B_γ be those variables 'below' γ in the perfect ordering, and the set A_γ be those variables above γ, but not parents of γ. Then the set of all possible terms in the model is $\{\gamma, pa(\gamma), A_\gamma, B_\gamma\}$. All terms in this set are distinct by definition. T lies in at least one of these following sets, and we shall consider each T in the first set on the list to which it corresponds. Hence for example when we consider $A_\gamma \cup B_\gamma$, we are only considering T including elements of both A and B.

1. γ
2. $pa(\gamma)$
3. A_γ
4. B_γ
5. $\gamma \cup pa(\gamma)$
6. $\gamma \cup A_\gamma$
7. $\gamma \cup B_\gamma$

8. $pa(\gamma) \cup A_\gamma$

9. $pa(\gamma) \cup B_\gamma$

10. $A_\gamma \cup B_\gamma$

11. $\gamma \cup pa(\gamma) \cup A_\gamma$

12. $\gamma \cup pa(\gamma) \cup B_\gamma$

13. $\gamma \cup A_\gamma \cup B_\gamma$

14. $pa(\gamma) \cup A_\gamma \cup B_\gamma$

15. $\gamma \cup pa(\gamma) \cup A_\gamma \cup B_\gamma$

This is an exhaustive list, however it is possible to eliminate several of these sets. By the definition of a directed graph for a log-linear model, it is impossible to have a log-linear model term that includes γ and any variables which are not either parents or descendants of γ (*i.e.* A). This excludes cases 6, 11, 13 and 15. Also, as the aim is to show that J is upper triangular, and then to determine the entries on the diagonal, terms which include any elements of B are of no interest, thus eliminating cases 4, 7, 9, 10, 12, and 13. The remaining five sets of interest are given below:

1. γ

2. $pa(\gamma)$

3. A_γ

4. $A_\gamma \cup pa(\gamma)$

5. $\gamma \cup pa(\gamma)$

Let us now consider the form of the probability corresponding to $\frac{\partial \phi_\gamma}{\partial \beta_j}$ (expression (3.5)) in each of these cases:

In case (3), this probability is equal to zero, as model term T is a subset of A and therefore $T \perp\!\!\!\perp \gamma | pa(\gamma)$ by definition of the perfect ordering, hence the two terms of expression (3.5) cancel.

In case (2), we also obtain zero, as $T \subseteq pa(\gamma)$, and hence both terms are equal to one if $k_T = i_{pa(\gamma)}$, or zero if $k_T \neq i_{pa(\gamma)}$.

Similarly, we obtain zeros in case (4) by combining the previous two cases. In this case, $T = T_1 \cup T_2$, where $T_1 \subseteq A$ and $T_2 \subseteq pa(\gamma)$, and we may re-write (3.5) as

$$\begin{aligned} & P(T_1 = k_{T_1}, T_2 = k_{T_2} | \gamma = i_\gamma, pa(\gamma) = i_{pa(\gamma)}) - \\ & \quad P(T_1 = k_{T_1}, T_2 = k_{T_2} | \gamma = 1, pa(\gamma) = i_{pa(\gamma)}) \\ = \; & P(T_2 = k_{T_2} | \gamma = i_\gamma, pa(\gamma) = i_{pa(\gamma)}, T_1 = k_{T_1}) P(T_1 = k_{T_1} | \gamma = i_\gamma, pa(\gamma) = i_{pa(\gamma)}) - \\ & \quad P(T_2 = k_{T_2} | \gamma = 1, pa(\gamma) = i_{pa(\gamma)}, T_1 = k_{T_1}) P(T_1 = k_{T_1} | \gamma = 1, pa(\gamma) = i_{pa(\gamma)}) \end{aligned}$$

The second and fourth terms in this expression are equal, since $A \perp\!\!\!\perp \gamma | pa(\gamma)$. The first and third terms are either equal to 1 if $k_{T_2} = i_{pa(\gamma)}$, or 0 if $k_{T_2} \neq i_{pa(\gamma)}$. Hence the expression is zero.

From these three cases, all the blocks below the diagonal are equal to zero, and hence J can be said to be block upper triangular. It remains to determine the form of the blocks on the diagonal – *i.e.* corresponding to cases (1) and (5).

In case (1), the probabilities take the form

$$P(\gamma = k_\gamma | \gamma = i_\gamma, pa(\gamma) = i_{pa(\gamma)}) - P(\gamma = k_\gamma | \gamma = 1, pa(\gamma) = i_{pa(\gamma)})$$

The second term is zero, by definition of the logits ϕ_γ, and hence the expression is equal to 1 if $k_\gamma = i_\gamma$, and 0 otherwise. Similarly, in case (5), the expression

becomes

$$P(\gamma = k_\gamma, pa(\gamma) = k_{pa(\gamma)} | \gamma = i_\gamma, pa(\gamma) = i_{pa(\gamma)}) -$$
$$P(\gamma = k_\gamma, pa(\gamma) = k_{pa(\gamma)} | \gamma = 1, pa(\gamma) = i_{pa(\gamma)})$$

which is also equal to 1 if $k_\gamma = i_\gamma$ and $k_{pa(\gamma)} = i_{pa(\gamma)}$, and 0 otherwise. We may therefore order the terms within these blocks so that 1's appear on the diagonal, with zeros elsewhere.

The Jacobian is therefore upper triangular, and all entries on the diagonal are equal to 1, so that $|J| = 1$.

We may now re-write expression (3.4) using $|J| = 1$, and obtain

$$f(\phi) \propto \prod_\gamma \prod_{i_\gamma, i_{pa(\gamma)}} \left(\frac{\exp\{\phi_\gamma(i_\gamma, i_{pa(\gamma)})\}}{\sum_\gamma \exp\{\phi_\gamma(i_\gamma, i_{pa(\gamma)})\}} \right)^{\alpha_{\gamma, pa(\gamma)}}$$

This may now be written as a function of conditional probabilities

$$f(\boldsymbol{p}) \propto \prod_\gamma \prod_{i_\gamma, i_{pa(\gamma)}} P(\gamma = i_\gamma | pa(\gamma) = i_{pa(\gamma)})^{\alpha_{\gamma, pa(\gamma)}} |J|$$

where now $|J|$ is the Jacobian of the transformation from $\boldsymbol{\phi}$ to $\boldsymbol{p}$. A set of logit parameters exists for each $\gamma, i_{pa(\gamma)}$, and hence the Jacobian is block diagonal, with the determinant of each sub-block equal to $\prod_{i_\gamma} P(\gamma = i_\gamma | pa(\gamma) = i_{pa(\gamma)})^{-1}$. The Jacobian is therefore given by

$$\begin{aligned} |J| &= \prod_\gamma \prod_{i_{pa(\gamma)}} \left[\prod_{i_\gamma} P(\gamma = i_\gamma | pa(\gamma) = i_{pa(\gamma)})^{-1} \right] \\ &= \prod_\gamma \prod_{i_\gamma, i_{pa(\gamma)}} P(\gamma = i_\gamma | pa(\gamma) = i_{pa(\gamma)})^{-1} \end{aligned}$$

and hence we obtain

$$f(\boldsymbol{p}) \propto \prod_{\gamma} \prod_{i_\gamma, i_{pa(\gamma)}} P(\gamma = i_\gamma | pa(\gamma) = i_{pa(\gamma)})^{\alpha_{\gamma, pa(\gamma)} - 1} \tag{3.6}$$

The aim of this proof was to show the equivalence of the conditional Dirichlet and hyper Dirichlet prior distributions for decomposable log-linear models. We have shown the existence of a one to one correspondence between the two parameterisations of the model corresponding to these two distributions, and have shown that the Jacobian, $|J|$, of the transformation between these two parameterisations is equal to 1. Expression (3.6) gives the distribution for the conditional probabilities equivalent to the conditional Dirichlet distribution for a particular log-linear model. This distribution is clearly a product of Dirichlet distributions, and is the distribution obtained by marginalising from the saturated model, and hence is hyper Dirichlet.

It can therefore be concluded that for a particular decomposable log-linear model, the induced conditional Dirichlet distribution is identical to the equivalent hyper Dirichlet distribution.

3.4 Discussion

The conditional Dirichlet distribution has been introduced in this Chapter, and this distribution has been shown to be equivalent to a hyper Dirichlet distribution for a decomposable log-linear model. The conditional Dirichlet distribution is an attractive prior distribution as its parameters may be interpreted as prior data, and inference using this prior is straightforward by considering the equivalent hyper Dirichlet distribution, which is tractable.

A further advantage of the conditional Dirichlet distribution is that it is defined for any log-linear model, and its relationship to the hyper Dirichlet distribution allows it to be considered as a natural extension of this distribution to non-decomposable models.

Chapter 4

Posterior Sampling

In Bayesian statistics, our interest lies in the analysis of the posterior distribution. This is often a highly multivariate distribution, and so we require methods to summarise it, typically involving calculating appropriate marginal summaries. However, we frequently find the posterior distribution to be analytically intractable, and the marginal distributions of interest are not available analytically. In such cases, we can use Monte Carlo methods to obtain a (hypothetical) sample from the posterior distribution, from which it is then straightforward to obtain a sample from a particular marginal distribution. Methods of summarising the posterior, such as integrating to obtain expectations, may then be replaced by equivalent methods using our sample, for example calculating sample means to estimate expectations.

This Chapter introduces the method of Gibbs sampling in order to obtain a sample from a potentially intractable distribution. This technique is especially useful when applied to conditional Dirichlet distributions, as the priors which we have described result in intractable posterior distributions for non-decomposable models. Such applications will be considered in section 4.2. Note that such methods are not necessary when using a hyper Dirichlet distribution (*i.e.* when the model is decomposable), as this distribution is a product of independent

Dirichlet distributions, and so it is possible to sample directly from the posterior (described in section 3.2.2).

4.1 Gibbs Sampling

4.1.1 Introduction

The theory of Markov chain Monte Carlo methods was introduced in Chapter 1; Gibbs sampling is a particular application of this theory, widely used in Bayesian analysis. For example, Dellaportas and Smith (1993) applied the method to a wide class of generalised linear models.

Using the same notation as previously, suppose we require a sample from the distribution with density function $f(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is p-dimensional, and let the t-th iterate generated be denoted by $\boldsymbol{\theta}^{(t)}$. The fundamental principle of Gibbs sampling is to generate each component of $\boldsymbol{\theta}$ one at a time from a univariate conditional distribution. This algorithm, as applied to obtaining a sample from the distribution with density $f(\boldsymbol{\theta})$, is summarised below:

- Choose starting value $\boldsymbol{\theta}^{(0)} = \left(\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_p^{(0)}\right)$, possibly by maximising $f(\boldsymbol{\theta})$.
- Generate $\theta_1^{(1)}$ from $f(\theta_1|\theta_2^{(0)}, \theta_3^{(0)}, \ldots, \theta_p^{(0)})$.
- Generate $\theta_2^{(1)}$ from $f(\theta_2|\theta_1^{(1)}, \theta_3^{(0)}, \ldots, \theta_p^{(0)})$.
- ...
- Generate $\theta_p^{(1)}$ from $f(\theta_p|\theta_1^{(1)}, \theta_2^{(1)}, \ldots, \theta_{p-1}^{(1)})$.

A 'new' observation $\boldsymbol{\theta}^{(1)}$ has now been generated. Successive application of the process results in a sequence $\boldsymbol{\theta}^{(0)}, \boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots$ of observations of $\boldsymbol{\theta}$.

Note that we do not usually have the conditional densities required in a closed form and hence we use the result that the conditional densities are proportional to suitable joint densities (which we have in an un-normalised form), and then generate from these densities. For example, $f(\theta_1|\theta_2^{(0)}, \theta_3^{(0)}, \ldots, \theta_p^{(0)}) \propto f(\theta_1, \theta_2^{(0)}, \theta_3^{(0)}, \ldots, \theta_p^{(0)})$. It is (relatively) straightforward to sample from un-normalised densities, using methods such as rejection sampling. For this application, a modified version of rejection sampling, known as adaptive rejection sampling (Gilks and Wild, 1992) will be used. This is described below.

Adaptive Rejection Sampling

Rejection sampling is a common method of sampling independent points from a density. A particular advantage of the method is that the normalising constant for the density is not required.

Suppose we require a sample of n points from a density $f(x)$, with domain D, and let the un-normalised function be denoted $g(x)$ (so that $f(x) = \frac{g(x)}{\int g(x)dx}$). Define an envelope function $g_u(x)$ such that $g_u(x) \geq g(x)$ for all $x \in D$, and define a squeezing function $g_l(x)$ such that $g_l(x) \leq g(x)$ for all $x \in D$. The sampling algorithm then proceeds as follows

1. Sample a value x^* from $g_u(x)$, and sample a value w from a $Uniform(0,1)$ distribution.

2. If $w \leq \frac{g_l(x^*)}{g_u(x^*)}$ then accept x^*. Repeat from 1 until the required sample size is achieved.

3. If $w \leq \frac{g(x^*)}{g_u(x^*)}$ then accept x^*. Otherwise, reject x^*. Repeat from 1 until required sample size is achieved.

Clearly, it is only worthwhile using such a sampling method if it is easier to obtain a sample from $g_u(x)$ than from $f(x)$. A disadvantage of this method is

the difficulty in determining a suitable $g_u(x)$, and further work is also needed to locate the mode of $g(x)$, often using a standard optimisation method.

Gilks and Wild (1992) presented a modification of rejection sampling, called adaptive rejection sampling, which may be applied to those densities which are log-concave (a density function f is log-concave if $\log f$ is twice continuously differentiable, and its Hessian matrix of second derivatives is negative semi-definite). Their method has two distinct advantages. Firstly, because of the log-concavity, it is unnecessary to locate the mode of $f(x)$. Secondly, fewer evaluations of $g(x)$ are necessary, as the probability of needing a further evaluation is reduced after each rejection by updating the envelope functions and squeezing functions to take into account all the available information about $f(x)$. These functions are created using the fact that any concave function can be bounded by piecewise linear upper and lower bounds, constructed by using tangents at, and chords between, evaluated points of the function. They converge to $f(x)$ as sampling proceeds. More detailed explanation of the method of adaptive rejection sampling is given by Gilks and Wild (1992). The use of a Gibbs sampler for generalised linear models based on the method of adaptive rejection sampling was presented by Dellaportas and Smith (1993).

The method of Gibbs sampling based on adaptive rejection sampling is applied in the next section to the conditional Dirichlet distribution.

4.1.2 Application to Conditional Dirichlet Distribution

The conditional Dirichlet distribution is, in general, analytically intractable. It was therefore necessary to use a computational method to obtain samples from such distributions. The chosen method was to use Gibbs sampling, and the approach is described here. This method may be used to obtain samples from both prior and posterior conditional Dirichlet distributions.

In order to obtain a Gibbs sample from a conditional Dirichlet distribution,

it is necessary at each step to sample from the relevant univariate conditional distribution. However, as explained in the previous section, this is not often available in closed form and indeed this is true here. In order to sample from the (readily available) un-normalised joint density, the adaptive rejection sampling approach was chosen.

The conditional Dirichlet distribution was described in section 3.2.1, and its density function given by

$$f(\boldsymbol{\beta}) \propto \frac{\prod_{\boldsymbol{i} \in I} e^{\alpha(\boldsymbol{i}) \sum_j x(\boldsymbol{i},j)\beta_j}}{\left(\sum_{\boldsymbol{i}} e^{\sum_j x(\boldsymbol{i},j)\beta_j}\right)^{\alpha}}$$

for a model with $(n \times p)$ design matrix X, where p is the number of parameters in the log-linear model which sets $\boldsymbol{\theta} = X\boldsymbol{\beta}$, and $\boldsymbol{\alpha}$ is the vector of parameters. Note that $\boldsymbol{\alpha}$ may represent either prior or posterior parameters, with $\boldsymbol{\alpha} \rightarrow \boldsymbol{\alpha}+\boldsymbol{n}$ in the posterior.

As the only assumption required for adaptive rejection sampling is that of log-concavity of the univariate density functions, this is the only check we must make before proceeding. A density is log-concave if the second derivative of its log is negative definite. Hence, we must determine the matrix of second derivatives:

We may write the density $f(\boldsymbol{\beta})$ as

$$\begin{aligned} \log f &= \sum_i \sum_j \alpha_i x_{ij} \beta_j - \alpha \log \sum_i \exp\left\{\sum_j x_{ij}\beta_j\right\} \\ &= \sum_j \beta_j \sum_i \alpha_i x_{ij} - \alpha \log \sum_i \exp\left\{\sum_j x_{ij}\beta_j\right\} \end{aligned}$$

Note that for the sake of clarity, subscript notation is used in this part, as opposed to the notation used for the majority of the thesis; subscript i replaces $\boldsymbol{i}$ as an argument, and $x(\boldsymbol{i}, j)$ is replaced by x_{ij}.

The next step is to obtain the first and second derivatives:

$$\begin{aligned}
\frac{\partial}{\partial\beta_k} &= \sum_i \alpha_i x_{ik} - \frac{\alpha}{\sum_i \exp\left\{\sum_j x_{ij}\beta_j\right\}} \sum_i x_{ik} \exp\left\{\sum_j x_{ij}\beta_j\right\} \\
\frac{\partial^2}{\partial\beta_k^2} &= \frac{-\alpha}{\left(\sum_i \exp\left\{\sum_j x_{ij}\beta_j\right\}\right)^2} \left[\sum_i \exp\left\{\sum_j x_{ij}\beta_j\right\} \sum_i x_{ik}^2 \exp\left\{\sum_j x_{ij}\beta_j\right\} - \right. \\
&\qquad \left. \left(\sum_i x_{ik} \exp\left\{\sum_j x_{ij}\beta_j\right\}\right)^2\right] \\
&= -\alpha \left[\sum_i x_{ik}^2 p_i - \left(\sum_i x_{ik} p_i\right)^2\right] \\
\frac{\partial^2}{\partial\beta_k\beta_l} &= \frac{-\alpha}{\left(\sum_i \exp\left\{\sum_j x_{ij}\beta_j\right\}\right)^2} \left[\sum_i \exp\left\{\sum_j x_{ij}\beta_j\right\} \sum_i x_{ik}x_{il} \exp\left\{\sum_j x_{ij}\beta_j\right\} - \right. \\
&\qquad \left. \left(\sum_i x_{ik} \exp\left\{\sum_j x_{ij}\beta_j\right\}\right)\left(\sum_i x_{il} \exp\left\{\sum_j x_{ij}\beta_j\right\}\right)\right] \\
&= -\alpha \left[\sum_i x_{ik}x_{il}p_i - \sum_i x_{ik}p_i \sum_i x_{il}p_i\right]
\end{aligned}$$

We can now re-express $\frac{\partial^2}{\partial\boldsymbol{\beta}\boldsymbol{\beta}^T} \log f(\boldsymbol{\beta})$ in matrix form as

$$\begin{aligned}
\frac{\partial^2}{\partial\boldsymbol{\beta}\boldsymbol{\beta}^T} \log f(\boldsymbol{\beta}) &= -\alpha \left[X^T \mathrm{diag}(\boldsymbol{p}(\boldsymbol{\beta}))X - \left(X^T\boldsymbol{p}(\boldsymbol{\beta})\right)\left(X^T\boldsymbol{p}(\boldsymbol{\beta})\right)^T\right] \\
&= -\alpha X^T \left(\mathrm{diag}(\boldsymbol{p}(\boldsymbol{\beta})) - \boldsymbol{p}(\boldsymbol{\beta})\boldsymbol{p}(\boldsymbol{\beta})^T\right) X
\end{aligned}$$

Hence, in order to prove log-concavity of the whole distribution, we must have

$$\boldsymbol{\alpha}^T X^T (\mathrm{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^T) X\boldsymbol{\alpha} > 0 \text{ for all } \boldsymbol{\alpha} \neq \mathbf{0}$$

This is equal to the condition

$$(X\boldsymbol{\alpha})^T(\text{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^T)(X\boldsymbol{\alpha}) > 0 \text{ for all } \boldsymbol{\alpha} \neq \mathbf{0}$$

However, $(X\boldsymbol{\alpha})^T(\text{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^T)(X\boldsymbol{\alpha})$ is the variance of a discrete distribution with sample space $\{(X\boldsymbol{\alpha})_i; i = 1, \ldots, m\}$, and where $P((X\boldsymbol{\alpha})_i) = p_i$, and so this condition is satisfied and we have log-concavity of the whole density. Log-concavity of each univariate density is a direct consequence.

The method of Gibbs sampling using the adaptive rejection method was applied to the conditional Dirichlet distribution by coding a suitable program in C. The program requires the following information as part of an input file:

- The design matrix X for the log-linear model.
- The vector of parameters $\boldsymbol{\alpha}$, representing the cell counts. Note that the multinomial-Dirichlet conjugacy, which allows prior parameters to be interpreted as a 'prior sample', means that $\boldsymbol{\alpha}$ is a vector of prior parameters in order to sample from the prior distribution. Alternatively, to sample from the corresponding posterior distribution, $\boldsymbol{n}+\boldsymbol{\alpha}$ is used.
- The required Monte Carlo sample size, expressed as a number of complete sample vectors.

In order to generate the design matrix for a particular log-linear model, another program was written in C. The input to this program is the pattern of interactions for the model, expressed in binary form, and the output is the corresponding design matrix.

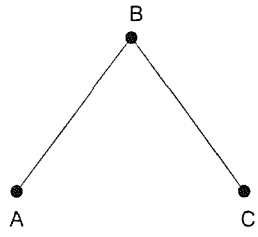
The Gibbs sampler outputs a sample from the conditional Dirichlet distribution for the specified model, either in terms of the cell probabilities $\boldsymbol{p}$ (useful for the next section) or in terms of the log-linear model parameters $\boldsymbol{\beta}$ (used in a later Chapter).

4.2 Conditional Dirichlet Samples

The method of Gibbs sampling from a conditional Dirichlet distribution was described in the previous section. As mentioned, this was first designed as a way to validate the hypothesis of equivalence of the hyper Dirichlet and conditional Dirichlet distributions for decomposable models, although its major use in this thesis is in bridge sampling, described in Chapter 5. Samples were generated from a large number of conditional Dirichlet distributions, and an example of this (Example 1) is presented here. The Gibbs sampler is then applied to data from a 2^6 table under several log-linear models (Example 2).

4.2.1 Example 1

Consider the $2 \times 2 \times 2$ model which may be represented graphically as



This model has cliques $\{A, B\}$ and $\{B, C\}$, and may be parameterised as $P(B)P(A|B)P(C|B)$. A design matrix for such a model is given by

$$X = \frac{1}{\sqrt{8}} \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ -1 & 1 & 1 & -1 & 1 \\ 1 & -1 & 1 & -1 & -1 \\ -1 & -1 & 1 & 1 & -1 \\ 1 & 1 & -1 & 1 & -1 \\ -1 & 1 & -1 & -1 & -1 \\ 1 & -1 & -1 & -1 & 1 \\ -1 & -1 & -1 & 1 & 1 \end{pmatrix}$$

The prior chosen for this example is the diffuse prior with parameters $\alpha(\boldsymbol{i}) = \frac{1}{8}$ for all $\boldsymbol{i}$ (Perks' prior).

The hyper Dirichlet distribution may be constructed from the 'full' Dirichlet distribution as follows. The distribution for $P(B)$ is obtained first, then the conditional distributions $P(A|B = 1)$, $P(A|B = 2)$, $P(C|B = 1)$, and $P(C|B = 2)$ are obtained in such a way that they are consistent with the distribution for $P(B)$. In this example, $P(B)$ is distributed as a $Beta(\frac{1}{2}, \frac{1}{2})$, and all the conditional densities follow $Beta(\frac{1}{4}, \frac{1}{4})$ distributions. As is clear from the model parameterisation, there are five independent distributions in this example.

The Gibbs sampler was used to generate samples from this prior, and the graphs below (figure 4.1) show kernel density estimates for the five independent distributions produced by the sampler, overlaid with the true Dirichlet density. All graphs are on the logit scale.

As can be seen from the graphs, there is excellent agreement between the kernel density estimates from the Gibbs samples and the true densities. This is to be expected, and validates the quality of the Gibbs sampling code. The sample size used throughout is 10000, and the computation time for such a sample is negligible (a few seconds).

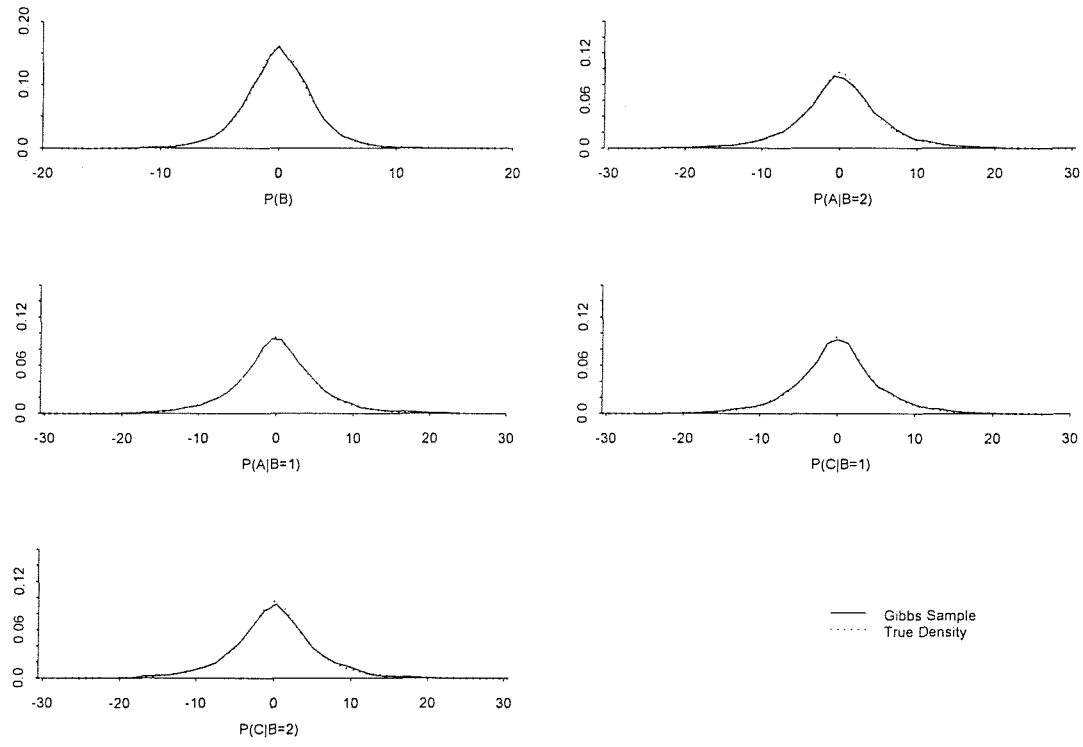


Figure 4.1: Plots showing kernel density estimates from Gibbs samples overlaid with the true density functions

4.2.2 Example 2

The Gibbs sampler was used to produce a posterior sample for some data concerning incidence of coronary heart disease. The data was presented by Edwards and Havranek (1985), and analysed further by Madigan and Raftery (1994) and Dellaportas and Forster (1999).

The data (presented in table 4.1) concerns 1841 men, who have been cross-classified in a 2^6 table by six factors for coronary heart disease. The six factors are: A - Smoking (no or yes); B - Strenuous mental work (no or yes); C - Strenuous physical work (no or yes); D - Systolic Blood pressure (< 140 or ≥ 140); E - Ratio of α and β lipoproteins (< 3 or ≥ 3); F - Family anamnesis

of coronary heart disease (negative or positive).

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

Table 4.1: Risk factors for coronary heart disease

Posterior samples were obtained for this data using the Gibbs sampler, for the most probable (hierarchical) models identified by Dellaportas and Forster (1999). These models have posterior probabilities of > 0.05. The prior parameters were set to $\alpha_i = \frac{1}{|I|} = 0.015625$ for a diffuse prior. The sets of graphs below show the distributions of the 2-way interaction parameters – each single graph corresponds to a particular interaction parameter, and each set to a particular model.

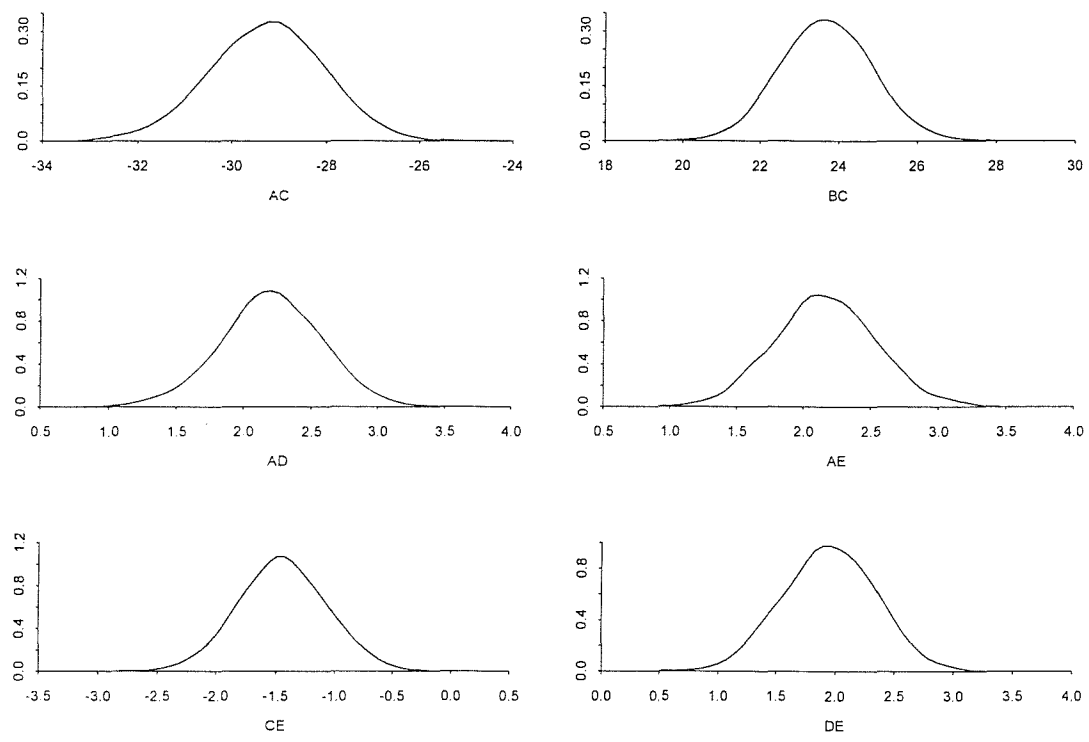


Figure 4.2: Model $AC + BC + AD + AE + CE + DE + F$ (posterior probability 0.28)

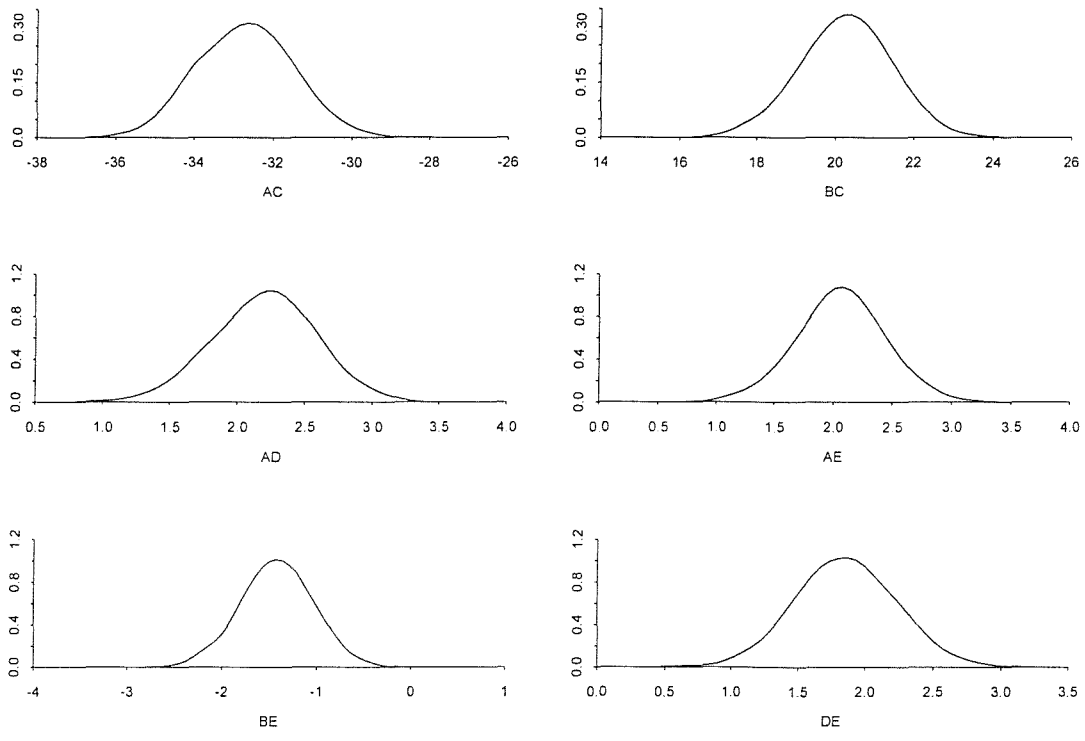


Figure 4.3: Model $AC + BC + AD + AE + BE + DE + F$ (posterior probability 0.16)

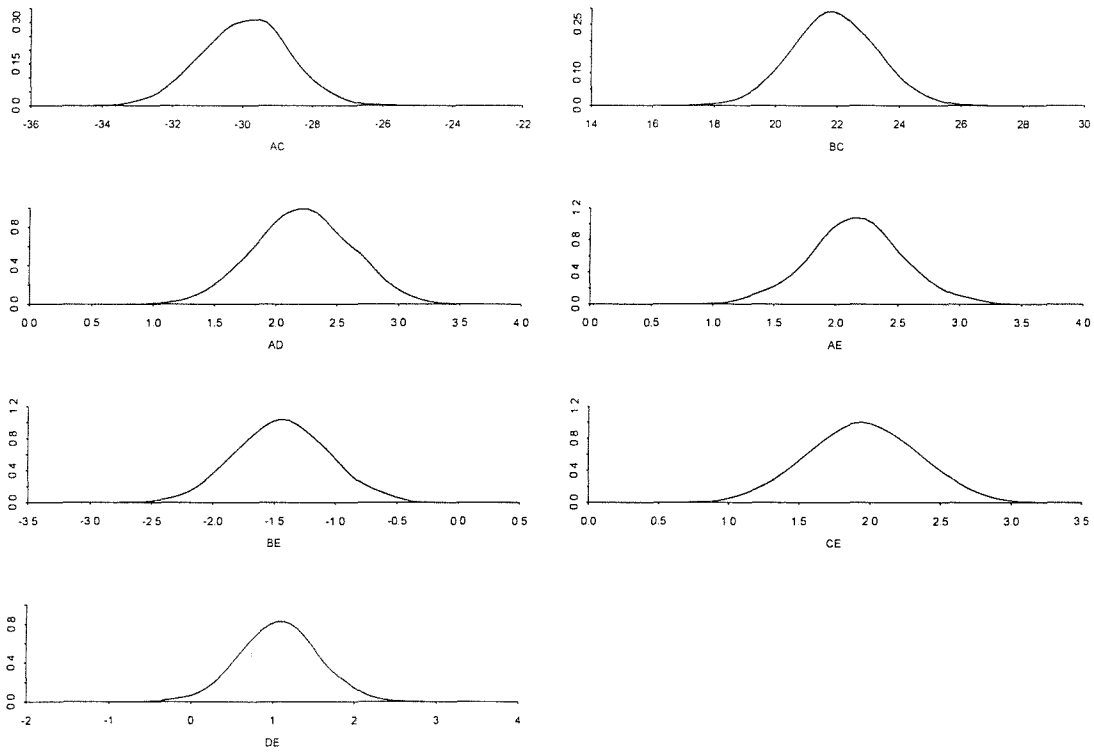


Figure 4.4: Model $AC + BC + AD + AE + BE + CE + DE + F$ (posterior probability 0.07)

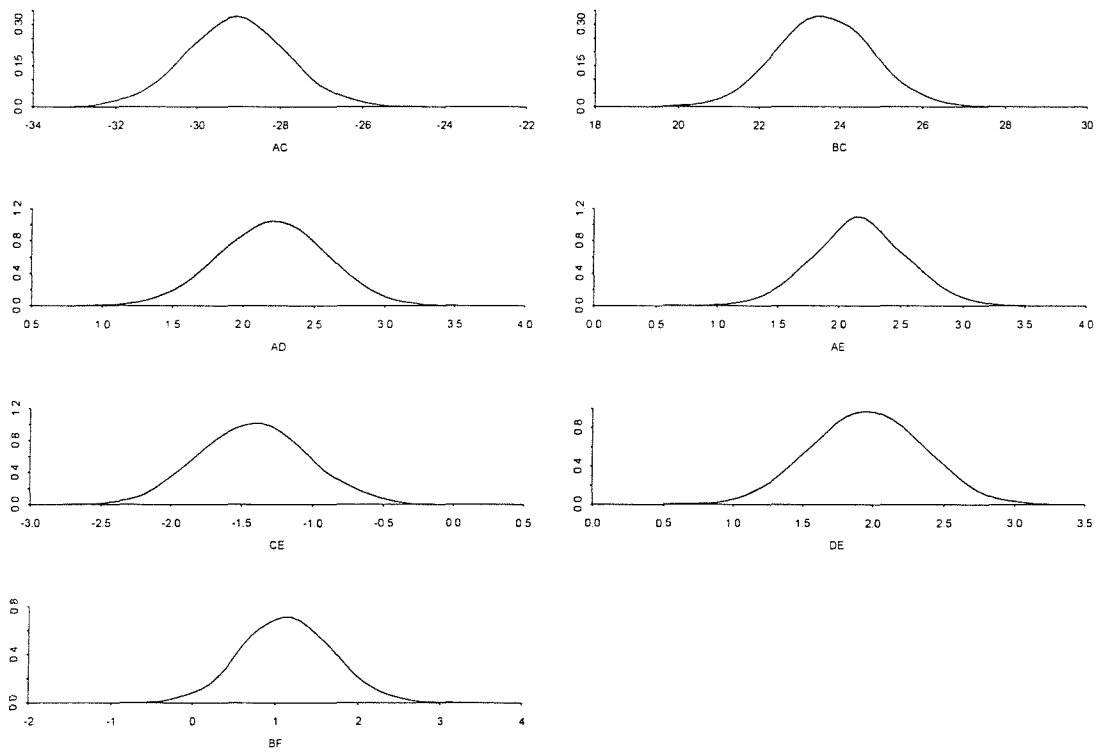


Figure 4.5: Model $AC + BC + AD + AE + CE + DE + BF$ (posterior probability 0.07)

4.2.3 Convergence of Gibbs Sampler

Repeated use of the Gibbs sampler leads to the conclusion that samples produced are not highly dependent, as the sampler appears to mix well. For the samples in Example 1, the autocorrelations at lag 1 are 0.2, and drop below 0.05 after lag 4.

Figure 4.6 shows time series plots for the data in Example 1. For the sake of clarity, the first 4000 observations only are plotted in each case.

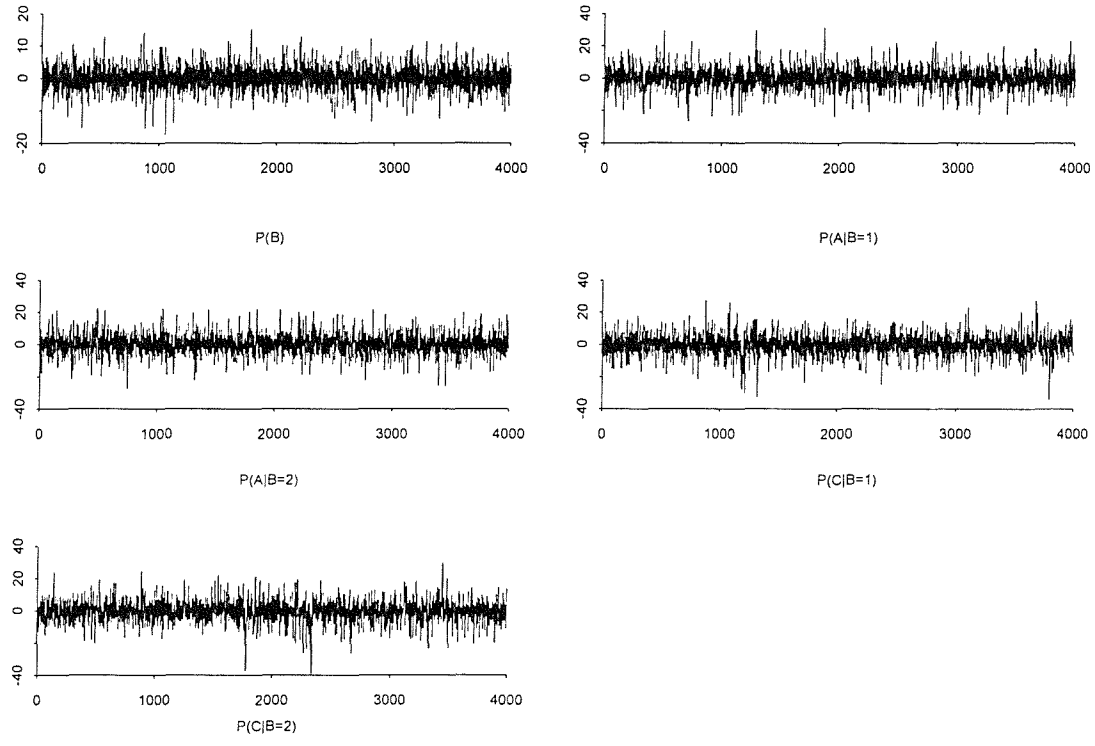


Figure 4.6: Time series plots for Gibbs samples in Example 1

The plots show that the Gibbs sampler is mixing very well, and so the observations are not highly dependent. Scatterplots for each pair of variables are shown in figure 4.7. There is clearly no distinct correlation between variables.

Time series plots for the data presented in Example 2 are all similar. The

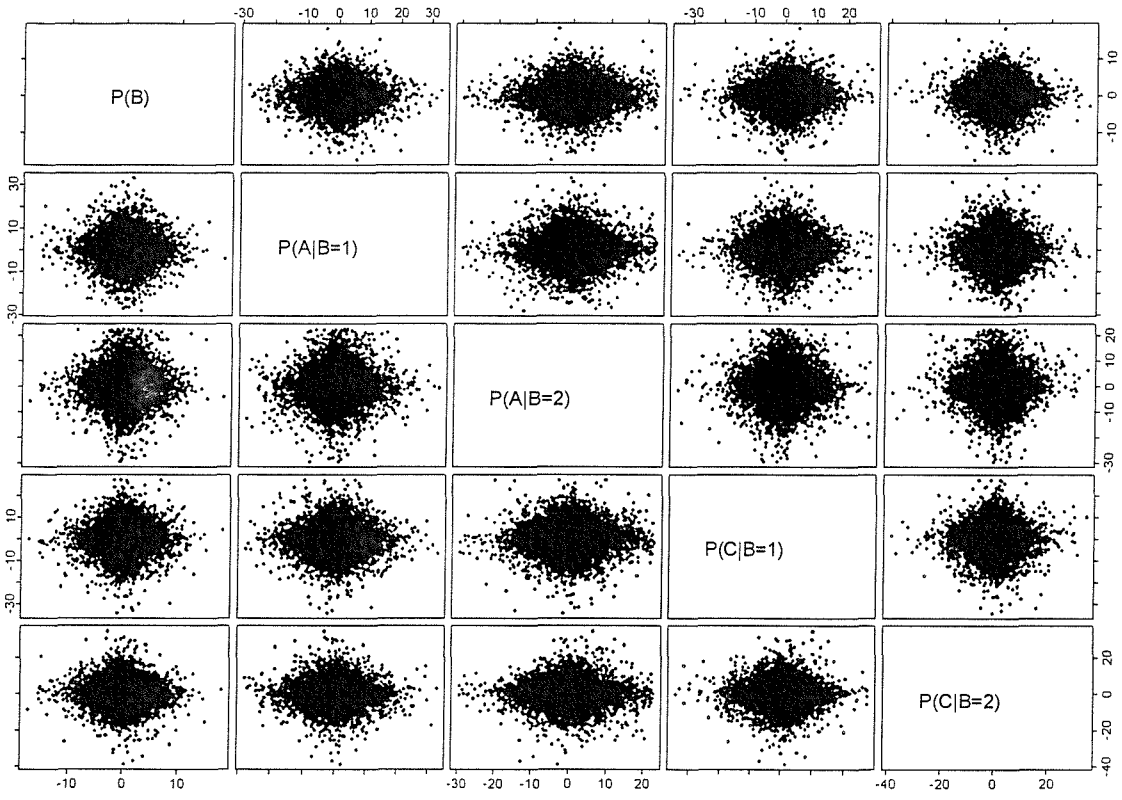


Figure 4.7: Pairwise scatterplots for Gibbs samples in Example 1

graphs in figure 4.8 show the plots for the most probable model, $AC + BC + AD + AE + CE + DE + F$, though again only the first 4000 observations are plotted.

Again, these plots show the observations are not highly dependent, and that the sampler is mixing well.

4.3 Discussion

In this Chapter, a Gibbs sampler has been developed, based on an adaptive rejection sampling method, which will produce samples from densities based on the conditional Dirichlet distribution. The convergence of this sampler is quick,

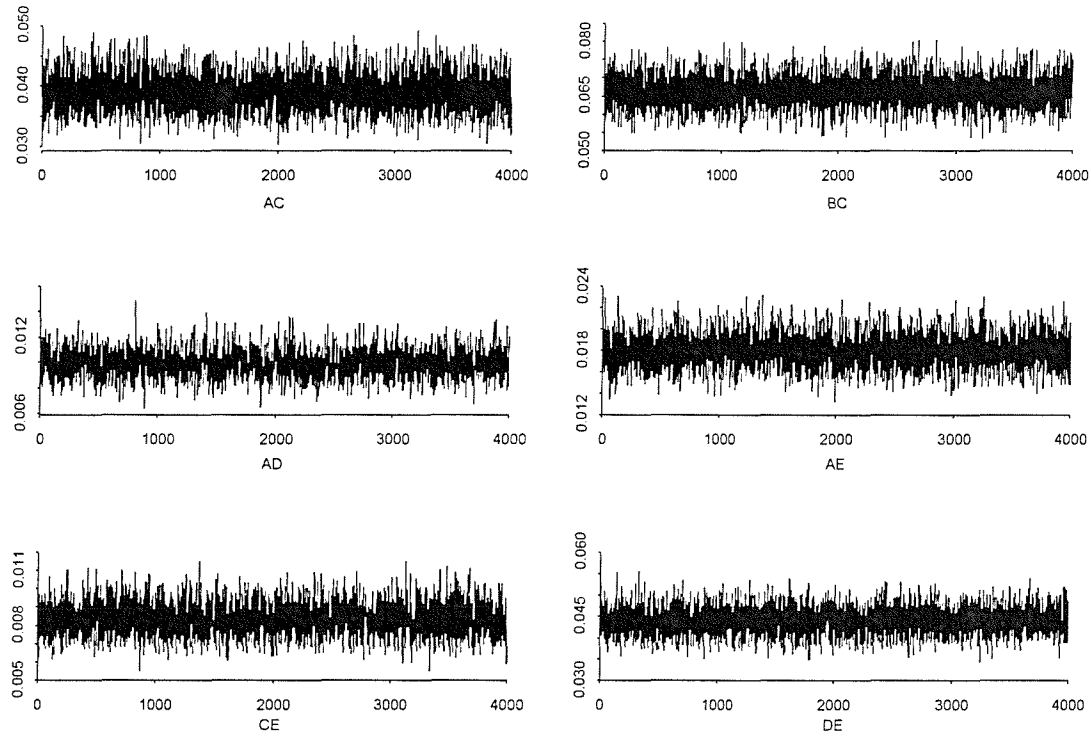


Figure 4.8: Time series plots for Gibbs samples corresponding to model $AC + BC + AD + AE + CE + DE + F$ in Example 2

allowing large reliable samples to be obtained quickly.

The integrity of the samples, and moreover the agreement with the theoretical results on the hyper Dirichlet and conditional Dirichlet from Chapter 3, was checked by application to examples, one of which was presented in section 4.2.1. Samples were then obtained from a large 2^6 dataset for several models, and graphs presented of the 2-way interaction parameters.

The major use of the Gibbs sampler will be presented in the next Chapter, as it is an important component in the method of bridge sampling.

Chapter 5

Posterior Distributions: Model Determination

The focus of Chapter 3 was the conditional Dirichlet distribution, and its relationship with the hyper Dirichlet distribution. One of the problems encountered was the intractability of the conditional Dirichlet distribution and in particular, the inability in general to write down the normalising constant for such a distribution in a closed form. It is, however, possible to obtain samples from such distributions using a Gibbs sampler (described in Chapter 4) as the method of rejection sampling does not require a normalised form of the conditional distribution.

The focus of this Chapter is the determination of normalising constants for conditional Dirichlet prior distributions, and for resulting posterior distributions, using Laplace's method and bridge sampling to approximate integrals. This is motivated by the problem of model selection. Note that such approximation methods are not necessary when using the hyper Dirichlet distribution (*i.e.* when the model is decomposable), since this density is conjugate to a multinomial likelihood and may be written in closed form and so exact results are possible.

5.1 Introduction

Suppose there are a set of competing models by which it is believed the data may have been generated. Model determination involves the selection of a particular statistical model, or identifying multiple plausible models, based on both the data and the knowledge of which models were considered plausible *a priori*. The basic theory of this was introduced in section 1.4.2, where we showed that the posterior probability of a particular model m may be found explicitly from

$$f(m|\boldsymbol{n}) = \frac{f(m)\int f(\boldsymbol{n}|m,\boldsymbol{\theta}_m)f(\boldsymbol{\theta}_m|m)d\boldsymbol{\theta}_m}{\sum_{m\in M} f(m)\int f(\boldsymbol{n}|m,\boldsymbol{\theta}_m)f(\boldsymbol{\theta}_m|m)d\boldsymbol{\theta}_m} \qquad m \in M \qquad (5.1)$$

and that, if we have two competing models, m_1 and m_2, the problem reduces to the calculation of a Bayes Factor, which is the ratio of the posterior odds to the prior odds:

$$\frac{f(m_1|\boldsymbol{n})}{f(m_2|\boldsymbol{n})} = \frac{f(m_1)}{f(m_2)} \times \frac{\int f(\boldsymbol{n}|m_1,\boldsymbol{\theta}_{m_1})f(\boldsymbol{\theta}_{m_1}|m_1)d\boldsymbol{\theta}_{m_1}}{\int f(\boldsymbol{n}|m_2,\boldsymbol{\theta}_{m_2})f(\boldsymbol{\theta}_{m_2}|m_2)d\boldsymbol{\theta}_{m_2}}$$

Note the Bayes Factor is the second term on the right hand side, which we refer to as B_{12}. The Bayes factor is in fact the ratio of two *marginal likelihoods*, $f(\boldsymbol{n}|m_1)$ and $f(\boldsymbol{n}|m_2)$, *i.e.*

$$B_{12} = \frac{\int f(\boldsymbol{n}|m_1,\boldsymbol{\theta}_{m_1})f(\boldsymbol{\theta}_{m_1}|m_1)d\boldsymbol{\theta}_{m_1}}{\int f(\boldsymbol{n}|m_2,\boldsymbol{\theta}_{m_2})f(\boldsymbol{\theta}_{m_2}|m_2)d\boldsymbol{\theta}_{m_2}} = \frac{f(\boldsymbol{n}|m_1)}{f(\boldsymbol{n}|m_2)} \qquad (5.2)$$

and this represents the weight of evidence in the data in favour of model m_1 over m_2.

As mentioned in section 2.4.1, the integrals in (5.1) and (5.2) may be analytically intractable. In this thesis we investigate the conditional Dirichlet distribution, and posterior conditional Dirichlet densities are indeed analytically intractable in general. Hence, numerical methods are required to obtain approximations for the normalising constants.

Note that normalised versions of the prior densities $f(\boldsymbol{\theta}_{m_i}|m_i)$ are needed in expression (5.2), though where these densities are intractable (as for the conditional Dirichlet distribution), numerical methods are also required to obtain these. In such cases, it is useful to re-write the log Bayes factor, given by the expression

$$\log B_{12} = \log \int f(\boldsymbol{n}|m_1, \boldsymbol{\theta}_{m_1}) f(\boldsymbol{\theta}_{m_1}|m_1) d\boldsymbol{\theta}_{m_1} - \log \int f(\boldsymbol{n}|m_2, \boldsymbol{\theta}_{m_2}) f(\boldsymbol{\theta}_{m_2}|m_2) d\boldsymbol{\theta}_{m_2}$$

in terms of un-normalised prior densities

$$\begin{aligned}
\log B_{12} &= \log \frac{\int f(\boldsymbol{n}|m_1, \boldsymbol{\theta}_{m_1}) g(\boldsymbol{\theta}_{m_1}|m_1)}{\int g(\boldsymbol{\theta}_{m_1}|m_1) d\boldsymbol{\theta}_{m_1}} d\boldsymbol{\theta}_{m_1} - \\
&\qquad \log \frac{\int f(\boldsymbol{n}|m_2, \boldsymbol{\theta}_{m_2}) g(\boldsymbol{\theta}_{m_2}|m_2)}{\int g(\boldsymbol{\theta}_{m_2}|m_2) d\boldsymbol{\theta}_{m_2}} d\boldsymbol{\theta}_{m_2} \\
&= \log \int f(\boldsymbol{n}|m_1, \boldsymbol{\theta}_{m_1}) g(\boldsymbol{\theta}_{m_1}|m_1) d\boldsymbol{\theta}_{m_1} - \log \int g(\boldsymbol{\theta}_{m_1}|m_1) d\boldsymbol{\theta}_{m_1} - \\
&\qquad \log \int f(\boldsymbol{n}|m_2, \boldsymbol{\theta}_{m_2}) g(\boldsymbol{\theta}_{m_2}|m_2) d\boldsymbol{\theta}_{m_2} + \log \int g(\boldsymbol{\theta}_{m_2}|m_2) d\boldsymbol{\theta}_{m_2}
\end{aligned}$$

Numerical methods may then be applied to each integral in turn in order to obtain the approximation to the log Bayes factor. The density functions may be of high dimension, which can cause difficulties with such methods. Three methods for evaluating these integrals are investigated in this Chapter, and we focus on Laplace's method, which is described in section 5.3, and bridge sampling, which is considered in section 5.4.

5.2 Schwarz Approximation

The difficulties involved in calculating the Bayes factor are mainly those of calculating the integrals involved, and these are not generally tractable. However, an alternative to this is to avoid altogether the introduction of prior densities

$f(\boldsymbol{\theta}_{m_1}|m_1)$ and $f(\boldsymbol{\theta}_{m_2}|m_2)$ and approximate B_{12} by the expression

$$\log B_{12} \approx S_{12} = \log f(\boldsymbol{n}|m_1, \widehat{\boldsymbol{\theta}}_{m_1}) - \log f(\boldsymbol{n}|m_2, \widehat{\boldsymbol{\theta}}_{m_2}) - \frac{1}{2}(d_1 - d_2)\log n$$

where d_i is the dimension of $\boldsymbol{\theta}_{m_i}$, $\widehat{\boldsymbol{\theta}}_{m_i}$ minimises the (log) likelihood function under H_i, and n is the total sample size. This quantity is called the Schwarz criterion (Schwarz, 1978), and can be used as an approximation to the log Bayes factor in model selection problems where the true marginal likelihood is difficult to evaluate.

The Schwarz criterion is related to the Bayes Information Criterion (BIC - Raftery, 1986) through the equation $-2S = \text{BIC}$. Note that BIC is defined for a single model using the expression

$$\text{BIC} = -2(\log \text{maximised likelihood}) + \log n \times \text{number of parameters}$$

where this is minimised by the most probable model. However, BIC and the Schwarz criterion are used interchangeably in the literature to compare models, so in this thesis we shall use the term Schwarz criterion where a Bayes factor is approximated, and Schwarz approximation where we are approximating a marginal likelihood alone.

The Schwarz criterion approximation to the log Bayes factor satisfies

$$\frac{S_{12} - \log B_{12}}{\log B_{12}} \to 0$$

as $n \to \infty$. This is a sufficient condition for the Schwarz criterion to provide a consistent estimate of the Bayes factor, though the approximation of $\log B_{12}$ provided by S_{12} is only accurate to an error of $O(1)$, and so allows

$$\frac{\exp(S_{12})}{B_{12}} \nrightarrow 1$$

This means that, particularly for certain prior distributions, the Schwarz criterion can be a poor approximation to the log Bayes factor, even if the sample size is large.

5.3 Laplace's Method

5.3.1 Derivation

The most frequently used approximation to the integrals in (5.1) is found by a technique known as *Laplace's Method.*

Tierney and Kadane (1986) presented an approximation for integrals of the form $\int e^{nL(\theta)}d\theta$. The approximation is based on the principle that, provided L has a unique maximum $\widetilde{\theta}$, or is at least dominated by a single mode then, for large n, the value of the integral is dependent solely upon the value of L near the maximum. The Taylor expansion of L about its maximum is

$$\begin{aligned} L(\theta) &= L(\widetilde{\theta}) + (\theta - \widetilde{\theta})L'(\widetilde{\theta}) + \frac{1}{2}(\theta - \widetilde{\theta})^2 L''(\widetilde{\theta}) + O(\theta - \widetilde{\theta})^3 \\ &= L(\widetilde{\theta}) + \frac{1}{2}(\theta - \widetilde{\theta})^2 L''(\widetilde{\theta}) + O(\theta - \widetilde{\theta})^3 \end{aligned}$$

and application of this expansion yields the result

$$\begin{aligned} \int \exp\{nL(\theta)\}\, d\theta &= \exp\left\{nL(\widetilde{\theta})\right\} \int \exp\left\{-\frac{(\theta - \widetilde{\theta})^2}{2}(-nL''(\widetilde{\theta})) + O(\theta - \widetilde{\theta})^3\right\} d\theta \\ &= \left(\frac{2\pi}{n(-L''(\widetilde{\theta}))}\right)^{\frac{1}{2}} e^{nL(\widetilde{\theta})}(1 + O(n^{-1})) \qquad (5.3) \end{aligned}$$

as the integrand is the kernel of a Normal$\left(\widetilde{\theta}, \frac{-1}{nL''(\widetilde{\theta})}\right)$ density.

Suppose we require an approximation to the normalising constant of a multivariate posterior distribution. The prior may be in an un-normalised form, so we shall denote this by $g(\boldsymbol{\theta}) = cf(\boldsymbol{\theta})$. The likelihood function is denoted $f(\boldsymbol{n}|\boldsymbol{\theta})$. The

above result is then easily generalised to this multivariate situation as follows.

Let $L(\boldsymbol{\theta}) = \frac{1}{n}\log g(\boldsymbol{\theta}|\boldsymbol{n}) = \frac{1}{n}\left(\log g(\boldsymbol{\theta}) + \log f(\boldsymbol{n}|\boldsymbol{\theta})\right)$, and suppose the dimension of these functions is d. Then expression (5.3) leads to the approximation

$$\int e^{nL(\theta)}d\boldsymbol{\theta} = \frac{(2\pi)^{\frac{d}{2}}e^{nL(\widetilde{\boldsymbol{\theta}})}}{n^{\frac{d}{2}}\left|-\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}L(\widetilde{\boldsymbol{\theta}})\right|^{\frac{1}{2}}}(1+O(n^{-1}))$$

where $\widetilde{\boldsymbol{\theta}}$ is the posterior mode, and so

$$\log\int e^{nL(\theta)}d\boldsymbol{\theta} = \frac{d}{2}\log 2\pi + nL(\widetilde{\boldsymbol{\theta}}) - \frac{1}{2}\log\left|-H^*(\widetilde{\boldsymbol{\theta}})\right| - \frac{d}{2}\log n + O(n^{-1})$$

where $H(\widetilde{\boldsymbol{\theta}}) = \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}L(\widetilde{\boldsymbol{\theta}})$ is the Hessian matrix of second derivatives. Writing this directly in terms of the prior and likelihood gives

$$\log\int f(\boldsymbol{\theta})f(\boldsymbol{n}|\boldsymbol{\theta})d\boldsymbol{\theta} = \frac{d}{2}\log 2\pi + \log f(\widetilde{\boldsymbol{\theta}}) + \log f(\boldsymbol{n}|\widetilde{\boldsymbol{\theta}}) - \frac{1}{2}\log\left|-H(\widetilde{\boldsymbol{\theta}})\right| + O(n^{-1}) \tag{5.4}$$

where $H(\widetilde{\boldsymbol{\theta}}) = \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\left[\log f(\widetilde{\boldsymbol{\theta}}) + \log f(\boldsymbol{n}|\widetilde{\boldsymbol{\theta}})\right]$, equivalent to the expression given by Tierney and Kadane (1986). Note that the use of this form of Laplace's method is restricted to cases where we may obtain the Hessian matrix of second derivatives, and also that application to cases where the tails of the integrand vary considerably from the Normal distribution will produce inaccurate results (this is considered later in the Chapter).

In a model selection problem, the marginal likelihood is required for the calculation of the Bayes Factor.

$$\begin{aligned}\frac{f(m_1|\boldsymbol{n})}{f(m_2|\boldsymbol{n})} &= \frac{f(m_1)}{f(m_2)} \times \frac{\int f(\boldsymbol{n}|m_1,\boldsymbol{\theta}_{m_1})f(\boldsymbol{\theta}_{m_1}|m_1)d\boldsymbol{\theta}_{m_1}}{\int f(\boldsymbol{n}|m_2,\boldsymbol{\theta}_{m_2})f(\boldsymbol{\theta}_{m_2}|m_2)d\boldsymbol{\theta}_{m_2}} \\ &= \frac{f(m_1)}{f(m_2)} \times B_{12}\end{aligned}$$

where B_{12} is the Bayes factor, comparing models m_1 and m_2. It is clear that,

as expression (5.4) provides an order $O(n^{-1})$ approximation to the marginal likelihood, an order $O(n^{-1})$ approximation to the log Bayes factor is given by the expression

$$\begin{aligned}\log B_{12} &= \log f(\boldsymbol{n}|\widetilde{\boldsymbol{\theta}}_{m_1}) + \log f(\widetilde{\boldsymbol{\theta}}_{m_1}) - \frac{1}{2}\log\left|-H(\widetilde{\boldsymbol{\theta}}_{m_1})\right| - \log f(\boldsymbol{n}|\widetilde{\boldsymbol{\theta}}_{m_2}) - \\ &\qquad \log f(\widetilde{\boldsymbol{\theta}}_{m_2}) + \frac{1}{2}\log\left|-H(\widetilde{\boldsymbol{\theta}}_{m_2})\right| + \frac{(d_1 - d_2)}{2}\log 2\pi + O(n^{-1})\end{aligned}$$

where d_i is the dimension of model i.

The form of the Laplace approximation derived above provides an order $O(n^{-1})$ approximation to the log marginal likelihood, and is based on the likelihood and prior densities evaluated at the posterior mode, and the Hessian matrix of second derivatives. However, a modified version of the approximation is available which does not require the Hessian, instead using the Fisher information matrix.

Let us apply a result from Kass and Wasserman (1995), namely

$$-n^{-1}H(\widetilde{\boldsymbol{\theta}}) - i(\widetilde{\boldsymbol{\theta}}) = O(n^{-1/2})$$

where $i(\boldsymbol{\theta}) = \frac{1}{n}I(\boldsymbol{\theta})$ is the Fisher information matrix for a single observation, and $I(\boldsymbol{\theta}) = E\left[-\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}f(\boldsymbol{y}|\boldsymbol{\theta})\right]$. Note that care must be taken in the definition of a 'single observation'. For example, in a contingency table, the numbers of units of information is the number of classified objects, not the number of cells.

Rearranging, we obtain

$$\begin{aligned}-n^{-1}H(\widetilde{\boldsymbol{\theta}}) &= O(n^{-1/2}) + i(\widetilde{\boldsymbol{\theta}}) \\ \left|-n^{-1}H(\widetilde{\boldsymbol{\theta}})\right| &= \left|O(n^{-1/2}) + i(\widetilde{\boldsymbol{\theta}})\right| \\ &= \left|i(\widetilde{\boldsymbol{\theta}})\left[1 + O(n^{-1/2})\right]\right|\end{aligned}$$

and so

$$\left|-H(\widetilde{\boldsymbol{\theta}})\right| = n^d \left|i(\widetilde{\boldsymbol{\theta}})\right| \left|1 + O(n^{-1/2})\right|$$

where d is the dimension of $\boldsymbol{\theta}$. Taking logs, we have

$$\begin{aligned} \log\left|-H(\widetilde{\boldsymbol{\theta}})\right| &= d\log n + \log\left|i(\widetilde{\boldsymbol{\theta}})\right| + \log\left|1 + O(n^{-1/2})\right| \\ &= d\log n + \log\left|i(\widetilde{\boldsymbol{\theta}})\right| + O(n^{-1/2}) \end{aligned}$$

Formula (5.4) may now be re-written in terms of the information matrix:

$$\int f(\boldsymbol{n}|\boldsymbol{\theta})g(\boldsymbol{\theta})d\boldsymbol{\theta} = \log f(\boldsymbol{n}|\widetilde{\boldsymbol{\theta}}) + \log g(\widetilde{\boldsymbol{\theta}}) + \frac{d}{2}\log 2\pi - \frac{d}{2}\log n - \frac{1}{2}\log\left|i(\widetilde{\boldsymbol{\theta}})\right| + O(n^{-1/2})$$

This expression yields an approximation to the marginal likelihood which is correct to order $O(n^{-\frac{1}{2}})$, which is the result derived by Kass and Wasserman (1995).

5.3.2 Application to Generalised Linear Models

The standard application of Laplace's method for approximating Bayes factors requires both the posterior mode $\widetilde{\boldsymbol{\theta}}_m$ and Hessian matrix $H(\boldsymbol{\theta}_m)$, though it was shown that it is possible to re-write the approximation in terms of the expected Fisher information matrix. Raftery (1996) considered the problem of using Laplace's method to approximate Bayes factors for generalised linear models. He pointed out that, although standard statistical software does not usually produce the posterior mode and Hessian matrix, it does often give the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}_m$, the likelihood ratio statistic and the observed or expected Fisher information matrix F_m. He presented two approximations based on these quantities for generalised linear models.

Suppose the prior mean and variance of $\boldsymbol{\theta}_m$ are given by $E[\boldsymbol{\theta}_m] = \omega_m$ and

$Var\,[\boldsymbol{\theta}_m] = W_m$. Then the first approximation is

$$2\log B_{12} \approx L_{12} + (E_1 - E_2) \tag{5.5}$$

where $L_{12} = 2\left\{\log f(\boldsymbol{n}|\widehat{\boldsymbol{\theta}}_{m_1}) - \log f(\boldsymbol{n}|\widehat{\boldsymbol{\theta}}_{m_2})\right\}$, which is the standard likelihood ratio test statistic for nested models, and where E_m is given by

$$\begin{aligned} E_m &= 2\log f(\widehat{\boldsymbol{\theta}}_m) - \log\left|F_m + W_m^{-1}\right| + d_m \log 2\pi + \\ &\quad \left[\frac{\partial}{\partial\boldsymbol{\theta}}\log f(\widehat{\boldsymbol{\theta}}_m)\right]^T (F_m + W_m^{-1})^{-1}\left[2 - F_m(F_m + W_m^{-1})\right]\frac{\partial}{\partial\boldsymbol{\theta}}\log f(\widehat{\boldsymbol{\theta}}_m) \end{aligned}$$

This approximation is closer to the standard Laplace approximation when F_m is the observed Fisher information, with error of order $O(n^{-1})$. Arguments similar to those in section 5.3 and by Kass and Vaidyanathan (1992) showed that this error increases to order $O(n^{-\frac{1}{2}})$ when the expected Fisher information is used.

Raftery's second approximation was based on the assumptions that $\widetilde{\boldsymbol{\theta}}_m \approx \widehat{\boldsymbol{\theta}}_m$ and $H^{-1}(\boldsymbol{\theta}_m) \approx -F_m$ (the observed information matrix). This resulted in the expression

$$2\log B_{12} \approx L_{12} + (E_1^* - E_2^*)$$

where

$$E_m^* = -\log|F_m| + 2\log f(\widehat{\boldsymbol{\theta}}_m) + d_m \log 2\pi$$

This approximation is less accurate than that given in (5.5), although Raftery (1996) found it to perform well in several situations, and found the separate terms for the prior and likelihood appealing. He applied both approximations to a simple Normal example, where analytic results are available, and found them both to give errors of order $O(n^{-1})$, though the second approximation was in general worse than the first.

Raftery applied his approximations to the problem of calculating Bayes factors for generalised linear models, where the posterior mode and Hessian matrix

are not, in general, available. The approximations were found to be of good quality. His methods are applicable to cases where the dispersion parameter is unknown, where there is overdispersion, to compare link functions, and to compare error distributions and variance functions.

Raftery suggested Normal prior distributions for use in the approximations, for cases where little prior information was available. A criticism of his priors is that they depend on the observed data, and as such would seem to violate a fundamental principle of the interpretation of a prior distribution. He emphasised the use of a reference set of proper priors in model selection, as opposed to a single (possibly improper) prior, an idea consistent with the rest of this thesis.

Diciccio, Kass, Raftery and Wasserman (1997) compared several methods of estimating the Bayes factor when it is possible to obtain a sample from the posterior distribution. They presented a modified version of Laplace's method based on this, and a Bartlett adjustment to Laplace's method which improved the Laplace estimate by an order of magnitude. They also considered importance sampling and reciprocal importance sampling, two special cases of bridge sampling, which is described in detail in section 5.4.

5.3.3 Application to Conditional Dirichlet Distribution

One of the main prior families investigated throughout this thesis is the conditional Dirichlet distribution. However, this distribution can be analytically intractable, hence the normalising constant is not (in general) known. The application of Laplace's method to this problem is described here.

As detailed in section 3.2.1, the conditional Dirichlet distribution has the general form

$$f(\boldsymbol{\beta}) \propto \frac{\prod_{\boldsymbol{i}\in I} e^{\alpha(\boldsymbol{i})\sum_j x(\boldsymbol{i},j)\beta_j}}{\left(\sum_{\boldsymbol{i}} e^{\sum_j x(\boldsymbol{i},j)\beta_j}\right)^{\alpha}}$$

Define $g(\boldsymbol{\beta})$ to be the un-normalised function, so that

$$g(\boldsymbol{\beta}) = \frac{\prod_{\boldsymbol{i}\in I} e^{\alpha(\boldsymbol{i})\sum_j x(\boldsymbol{i},j)\beta_j}}{\left(\sum_{\boldsymbol{i}} e^{\sum_j x(\boldsymbol{i},j)\beta_j}\right)^{\alpha}}$$

The p-dimensional conditional Dirichlet distribution is specified by the design matrix X, linking $\boldsymbol{\theta}$ to $\boldsymbol{\beta}$ through $\boldsymbol{\theta} = X\boldsymbol{\beta}$, and vector of parameters $\boldsymbol{\alpha}$. Note that, due to the conjugacy of the conditional Dirichlet prior to the multinomial distribution, $\boldsymbol{\alpha}$ may represent either prior or posterior parameters, with $\boldsymbol{\alpha} \rightarrow \boldsymbol{\alpha}+\boldsymbol{n}$ in the posterior. In order to obtain the Laplace approximation to the normalising constant it is necessary to determine an expression for the Hessian matrix of second derivatives for this function. Details of this derivation are given in section 4.1.2, and the resulting expression is

$$\frac{\partial^2}{\partial\boldsymbol{\beta}\boldsymbol{\beta}^T} \log g(\boldsymbol{\beta}) = \alpha X^T(diag(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^T)X$$

where $\boldsymbol{p}$ is a function of $\boldsymbol{\beta}$ through the expression $p(\boldsymbol{i}) = \frac{\exp\theta(\boldsymbol{i})}{\sum_{\boldsymbol{i}}\exp\theta(\boldsymbol{i})}$. Therefore, the Laplace approximation, based on expression (5.4) is

$$\begin{aligned}\log\int g(\boldsymbol{\beta})d\boldsymbol{\beta} \;=\; & \frac{p}{2}\log 2\pi - \frac{1}{2}\log\left|\alpha X^T(diag(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^T)X\right|^{1/2} + \\ & \log g(\widetilde{\boldsymbol{\beta}}) + O(\alpha^{-1}) \qquad (5.6)\end{aligned}$$

This expression is applicable for the estimation of both prior and posterior normalising constants, where $g(\boldsymbol{\beta})$ is replaced by $g(\boldsymbol{\beta}|\boldsymbol{n}) = f(\boldsymbol{n}|\boldsymbol{\beta})f(\boldsymbol{\beta})$ in the posterior case.

However, we know that a number of equivalent design matrices exist for any given log-linear model. In this instance, since X appears in the Laplace approximation, it is clear that different choices of X lead to different values of the normalising constant, so practical application of equation (5.6) requires a

consistent choice of design matrix to be used across models.

Code was written in S-Plus to apply the Laplace approximation to the problem of determining the normalising constant for a conditional Dirichlet density, specified by design matrix X and parameter vector $\boldsymbol{\alpha}$. Code was also written to produce the design matrix X for a particular log-linear model, specified by a binary representation of variables and interaction terms. The program has a single output – the approximation to the log normalising constant.

It was pointed out in the previous section that for a function f to be reliably approximated using Laplace's method, it must be highly peaked about its maximum $\widetilde{\theta}$, so that the main contribution to the function is within a neighbourhood of $\widetilde{\theta}$. The approximation is of error $O(\alpha^{-1})$, so the approximation will be good for large sample sizes, though approximation of vague prior distributions with small values of $\alpha(\boldsymbol{i})$ are unlikely to produce good results.

This problem is exacerbated as the tails of the conditional Dirichlet distribution are lighter than those of the Normal distribution, and so Laplace's method is likely to underestimate normalising constants when $\alpha(\boldsymbol{i})$ is small. The conditional Dirichlet distribution has the form

$$f(\boldsymbol{\beta}) \propto \frac{\prod_{\boldsymbol{i}\in I} e^{\alpha(\boldsymbol{i})\sum_j x(\boldsymbol{i},j)\beta_j}}{\left(\sum_{\boldsymbol{i}} e^{\sum_j x(\boldsymbol{i},j)\beta_j}\right)^{\alpha}}$$

Now consider this density as a function of a single β_k only, so that

$$f(\beta_k) \propto \frac{e^{\sum_{\boldsymbol{i}} \alpha(\boldsymbol{i})x(\boldsymbol{i},k)\beta_k}}{\left(\sum_{\boldsymbol{i}} e^{\sum_j x(\boldsymbol{i},j)\beta_j}\right)^{\alpha}}$$

This may be re-written as

$$f(\beta_k) \propto \frac{e^{\beta_k \sum_{\boldsymbol{i}} \alpha(\boldsymbol{i})x(\boldsymbol{i},k)}}{\left(1+\sum_{\boldsymbol{i}} e^{\sum_j x(\boldsymbol{i},j)\beta_j}\right)^{\alpha}}$$

and this expression tends to $e^{-\beta_k(\alpha c - c')}$ as $\beta_k \to \infty$, for some constants c and c' which depend on $\boldsymbol{\alpha}$. Hence this distribution decays exponentially with respect to β_k. However a Normal approximation would decay exponentially with respect to β_k^2, so would have heavier tails. This means that Laplace's method is likely to produce approximations which underestimate the conditional Dirichlet normalising constants.

5.3.4 Numerical Results from Laplace's Method applied to Conditional Dirichlet Distributions

The aim of applying Laplace's method to the Conditional Dirichlet distribution is to obtain the normalising constant for the (mostly) analytically intractable density function which results by conditioning on a particular log-linear model. However, in order to check the quality of the approximation, and any dependence on the dimension and complexity of the log-linear model, it is first necessary to apply the method to certain conditional Dirichlet distributions resulting from several log-linear models which are of a tractable form and so have known normalising constants.

As the approximation is of order $O(\alpha^{-1})$, it is clear that the accuracy of the approximation will improve for large sample sizes. This was investigated by obtaining Laplace approximations for increasing sample sizes, using a selection of models, and these results are summarised below.

Figure 5.1 contains 8 plots representing 8 different log-linear models. Each plot is of the error in the log of the Laplace approximation (given as the log of the approximate value minus the log of the true value), against the value of the cell parameter $\alpha(\boldsymbol{i})$ (the hypothetical 'sample' in each cell). The parameters are equally distributed throughout the cells in each case. The cell parameter runs from 0.25 to 25 in each case. The 8 models are

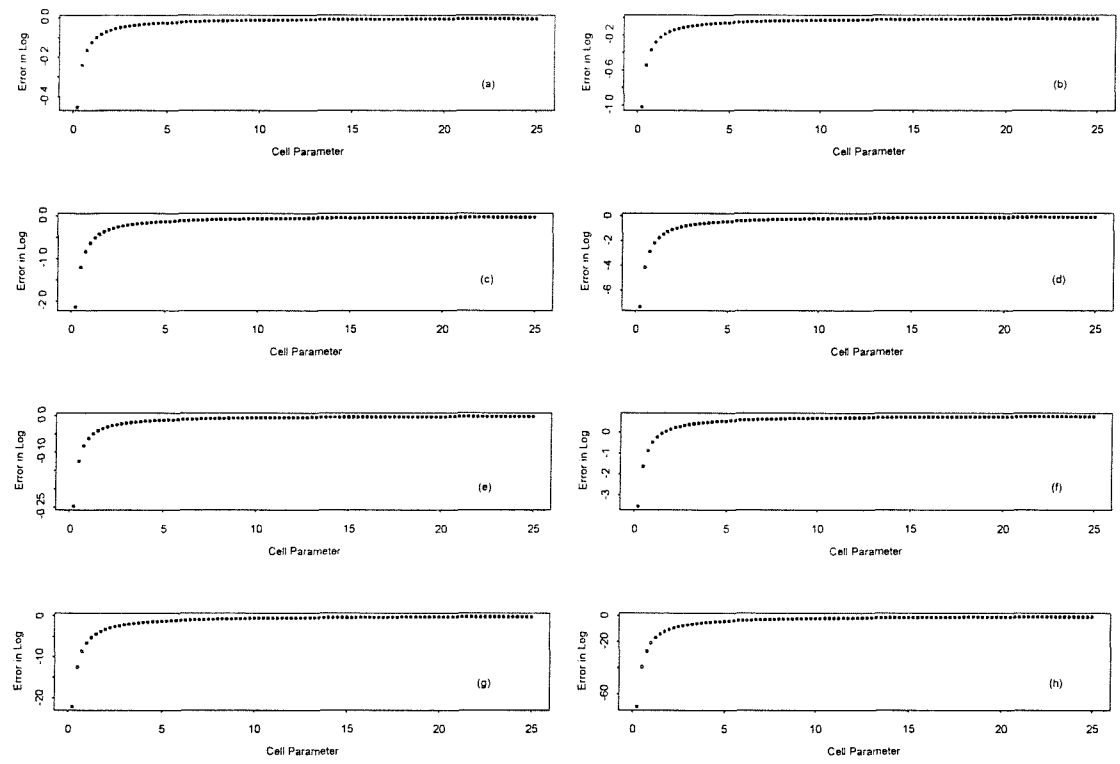


Figure 5.1: Plots showing convergence of Laplace estimates for various models with equal samples in each cell

(a) $A + B$ [2]
(b) $AB + BC$ [5]
(c) ABC [8]
(d) $A^{(3)}B^{(3)}C^{(3)}$ [27]
(e) $A + B + C + D$ [4]
(f) $ABCD$ [16]
(g) $A^{(3)}B^{(3)}C^{(3)}D^{(3)}$ [81]
(h) $A^{(4)}B^{(4)}C^{(4)}D^{(4)}$ [256]

All variables have 2 levels, except where indicated, and the numbers in square brackets give the number of model parameters in each case.

It is clear that for sample sizes greater than about 10 in each cell, the error of the approximation is negligible, and so the Laplace approximation is excellent. This is true for all the models. However it is also clear that, for certain models, the Laplace approximation for small values of cell parameters is poor, and so may

not be reliably used to determine the normalising constant for (reference) prior distributions. Indeed, with all cell parameters equal to 0.5, the approximation for the 4-way saturated model where all variables have four levels has an error of -39, which is huge. Examination of figure 5.1 shows that the error of the Laplace approximation increases significantly with increasing numbers of parameters in the model.

The approximations presented in figure 5.1 are all based on equal parameters in each cell. This is fine for prior distributions (where it seems that the Laplace approximation is of little use anyway), but is unrealistic for posterior distributions. In order to consider the unbalanced situation, Laplace approximations were obtained for posterior distributions where all the data was in a single cell. The results are presented graphically in figure 5.2 below. In each case, the 'Cell Parameter' refers to the data in the single cell. All other cells have a parameter of 0.25, representing a prior distribution based on a $Dirichlet(\frac{1}{4}\mathbf{1})$ distribution.

The graphs in figure 5.2 show that, when the data is distributed as described above, there is a considerable error in the Laplace approximation for all but the simplest model. It is therefore clear that the Laplace approximation to the normalising constant for conditional Dirichlet distributions is only reliable when there are at least a few observations in each cell. Exhaustive use of the Laplace approximation leads to the 'rule of thumb' that the approximation produced acceptable results when there are at least 5 observations in 80% of the cells, though note that the accuracy of the approximation improves with greater total sample size and decreases with increasing numbers of model paramaters.

In all the approximations presented above, note that the error in the logs is negative, which implies that the approximation of the normalising constant is too small, as expected.

In the next section, an alternative method of approximation will be introduced, which leads to accurate approximations even for small parameter values.

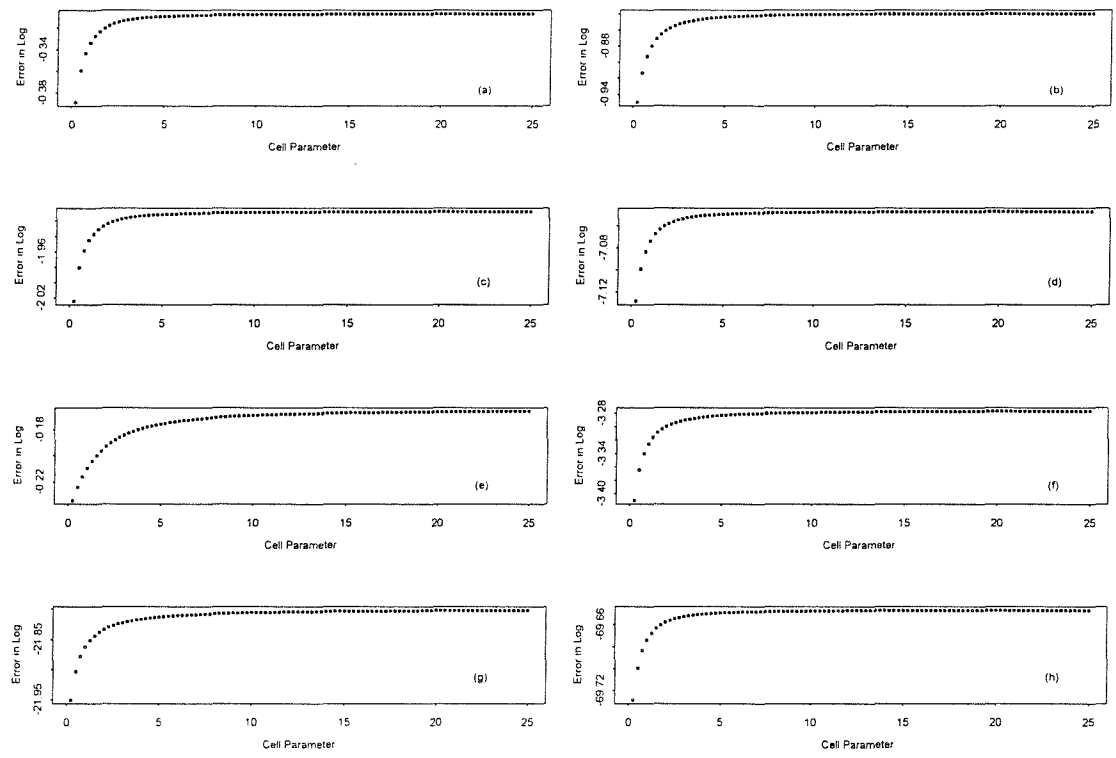


Figure 5.2: Plots showing convergence of Laplace estimates for various models with unbalanced cell counts

5.4 Bridge Sampling

The aim of this Chapter is to investigate methods of approximating the normalising constants for both prior and posterior conditional Dirichlet distributions. The Laplace approximation derived in the previous section was found to be unsuitable, in general, for application to conditional Dirichlet prior distributions. In this section, the method of Bridge Sampling is applied to this problem.

5.4.1 Introduction

The class of techniques known as bridge sampling were introduced by Bennett (1976), although they were studied in depth by Meng and Wong (1993) and

DiCiccio et al (1997). The method allows the estimation of the ratio of two normalising constants, though it can be modified to allow the estimation of a single normalising constant.

Suppose we have two densities d_1 and d_2, and write these as

$$d_i = \frac{u_i}{c_i}$$

where $c_i = \int u_i$ for $i = 1, 2$. Now let γ be a function which satisfies

$$0 < \left| \int \gamma(\theta) d_1(\theta) d_2(\theta) d\theta \right| < \infty$$

Then we may write

$$\frac{c_1}{c_2} = \frac{\int u_1(\theta)\gamma(\theta)d_2(\theta)d\theta}{\int u_2(\theta)\gamma(\theta)d_1(\theta)d\theta} \tag{5.7}$$

Now let our un-normalised density be denoted by $g(\theta)$, the associated normalising constant by C and the normalised density by $f(\theta)$, so that $C = \int g(\theta)d\theta$ and $f(\theta) = \frac{g(\theta)}{C}$. Suppose we have a sample from f, and denote this by $\theta_1, \ldots, \theta_m$. Let $q(\theta)$ be some density from which we may easily obtain a sample, and denote that sample by $\widetilde{\theta}_1, \ldots, \widetilde{\theta}_M$. Now, in expression (5.7), let $u_1 = g$, $c_1 = C$, $u_2 = q$ and $c_2 = 1$. Then

$$C = \frac{\int g(\theta)\gamma(\theta)q(\theta)d\theta}{\int q(\theta)\gamma(\theta)f(\theta)d\theta} \tag{5.8}$$

Using our samples, the bridge estimator of C introduced by Meng and Wong is given by

$$\widehat{C} = \frac{\frac{1}{M}\sum_i g(\widetilde{\theta}_i)\gamma(\widetilde{\theta}_i)}{\frac{1}{m}\sum_i q(\theta_i)\gamma(\theta_i)}$$

Clearly, a choice has to be made for the function γ. Several obvious choices are available - for example, $\gamma = \frac{1}{g}$ or $\gamma = \frac{1}{q}$. These reduce the bridge estimate to the commonly used estimates based on Importance Sampling and Reciprocal

Importance Sampling. However, it is interesting to consider the choice of γ based on a familiar optimality criterion – that of minimising the mean squared error. Meng and Wong found the optimal choice of γ in this case to be

$$\gamma(\theta) \propto \left[\frac{mg(\theta)}{C} + Mq(\theta)\right]^{-1} \tag{5.9}$$

This would appear to be of little practical use, as it requires the normalising constant, C, in its calculation. However, it is possible to use an estimate of C produced by an alternative approximation method, and substitute this value in the expression (5.9). For example, an estimate based on Laplace's method may be used, and indeed this is a technique which DiCiccio *et al* found produced a discernible increase in the accuracy of the approximation compared to other bridge samplers (for example importance sampling).

In practice, repeated applications of the bridge sampler may be used to iteratively update the approximation, using the previous value of C each time. This is the method which will be applied in the next section to the conditional Dirichlet distribution.

5.4.2 Application to the Conditional Dirichlet Distribution

The general theory of the bridge sampler was introduced in the previous section. A general expression (5.8) was presented, which gives the bridge estimate for a normalising constant for a particular distribution. The expression allows the size of the samples from densities q and g to differ, though for this application they will be equal, and denoted by m. In this application, we shall choose q to be a Normal density with mean equal to the mode of the conditional Dirichlet distribution, and variance matrix equal to the inverse of the Hessian matrix of second derivatives.

Let the (un-normalised) conditional Dirichlet density be denoted by $g(\boldsymbol{\beta})$, the sample from this be denoted $\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}, \ldots, \boldsymbol{\beta}^{(m)}$, and the Normal sample from density $q(\widetilde{\boldsymbol{\beta}})$ be denoted $\widetilde{\boldsymbol{\beta}}^{(1)}, \widetilde{\boldsymbol{\beta}}^{(2)}, \ldots, \widetilde{\boldsymbol{\beta}}^{(m)}$. The bridge sampler will be applied iteratively, with the i-th iteration denoted C_i. Then the bridge estimate is given by the expression

$$C_i = \int g(\boldsymbol{\beta}) d\boldsymbol{\beta} \approx \frac{\sum_i g(\widetilde{\boldsymbol{\beta}}^{(i)}) \gamma(\widetilde{\boldsymbol{\beta}}^{(i)})}{\sum_i q(\boldsymbol{\beta}^{(i)}) \gamma(\boldsymbol{\beta}^{(i)})}$$

where

$$\gamma(\boldsymbol{\beta}) = \left[\frac{mh(\boldsymbol{\beta})}{C_{i-1}} + mq(\boldsymbol{\beta}) \right]^{-1}$$

and C_0 is the estimate for the normalising constant by Laplace's method..

Code was written in S-plus to implement this procedure. The inputs to the function are the vector of cell parameters (counts), the design matrix for the log-linear model, and the required number of iterations, together with a sample from the density $f(\boldsymbol{\beta})$ obtained using the Gibbs sampler (see section 4.1.2). The bridge sampler estimate is output at each iteration.

5.4.3 Numerical Examples

In this section, the bridge sampler will be used to obtain prior and posterior normalising constants for a set of log-linear models where the true value is also available, as in section 4.2.1 (Laplace approximations). Successive runs of the bridge sampler produce values which, after about 3 iterations, seem to fluctuate slightly about a common value. Hence, to produce the estimates below, the bridge sampler is run iteratively 10 times, taking the Laplace estimate as a starting value, and the result presented is the mean of the final 7 iterations.

Table 5.1 gives the bridge sampling estimates for the log of the prior normalising constants, together with the error (expressed as the estimate minus the true value), and the value of the prior parameters, which are the same for each

cell. All variables have 2 levels, except where indicated

Hierarchical Log-linear Model	Prior Parameter	Bridge Approximation	Error in Bridge Approximation
$A+B$	0.25	2.29	0
$AB+BC$	0.125	9.16	0
ABC	0.125	16.15	0.01
$A^{(3)}B^{(3)}C^{(3)}$	0.5	−5.86	−0.05
$A+B+C+D$	0.0625	4.57	−0.01
$ABCD$	0.0625	43.94	0.11
$A^{(3)}B^{(3)}C^{(3)}D^{(3)}$	0.5	−61.78	0.33

Table 5.1: Bridge estimates, and their respective errors, of normalising constants for various models

It is clear from the table that the bridge sampling approximation is extremely good, even for distributions where the prior parameter is small. It therefore represents a huge improvement over the Laplace estimates, where the errors were of a much higher order. Such accuracy is also evident when the parameters in each cell are not equal (the unbalanced case).

The approximations in table 5.1 were all obtained using Gibbs sample sizes of 10000. This choice was motivated by the desire for the bridge estimate to vary by less than 0.1 about its limit, and for the sample to be produced reasonably quickly using the Gibbs sampler. Smaller sample sizes are adequate for simpler models.

5.4.4 Normalising Constants for Non-Decomposable Model

The results in the previous section demonstrate the accuracy of the method of bridge sampling to determine the normalising constants for the conditional Dirichlet prior for several decomposable models (where exact results are possible). However, there is one graphical model with up to and including 4 variables which is not decomposable. This is the model represented by the graph

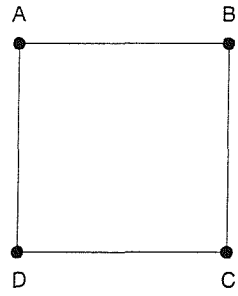


Table 5.2 gives the normalising constants for the conditional Dirichlet distributions for this model, with varying numbers of levels of the variables. The prior parameters in each case are symmetric, with a single observation split throughout the table (i.e. $\alpha(i) = \frac{1}{|I|}$).

Levels of A, B, C, D	log(Normalising Constant)
$2, 2, 2, 2$	1.45
$3, 2, 2, 2$	2.75
$2, 3, 2, 2$	3.67
$3, 3, 2, 2$	6.64
$3, 3, 3, 2$	10.09
$3, 3, 3, 3$	12.78

Table 5.2: Normalising constants for model $AB + BC + CD + DA$

Note that many other non-graphical log-linear models exist for which this approach is required, for example the model $AB + BC + AC$.

5.5 Risk Factors for Coronary Heart Disease

In section 4.2.2, the Gibbs sampler was used to obtain posterior samples from a number of models fitted to some data concerning incidence of coronary heart disease, originally presented by Edwards and Havranek (1985), and analysed further by Madigan and Raftery (1994) and Dellaportas and Forster (1999).

Recall from section 5.1 that the Bayes factor for comparing models m_1 and

m_2 is given by the expression

$$B_{12} = \frac{\int f(\boldsymbol{n}|m_1, \boldsymbol{\theta}_{m_1}) f(\boldsymbol{\theta}_{m_1}|m_1) d\boldsymbol{\theta}_{m_1}}{\int f(\boldsymbol{n}|m_2, \boldsymbol{\theta}_{m_2}) f(\boldsymbol{\theta}_{m_2}|m_2) d\boldsymbol{\theta}_{m_2}}$$

where $f(\boldsymbol{n}|m_i, \boldsymbol{\theta}_{m_i})$ is the likelihood under model m_i and $f(\boldsymbol{\theta}_{m_i}|m_i)$ is the prior under model m_i, and that this may be written in terms of un-normalised prior densities as

$$\begin{aligned} \log B_{12} \;=\; & \log \int f(\boldsymbol{n}|m_1, \boldsymbol{\theta}_{m_1}) g(\boldsymbol{\theta}_{m_1}|m_1) - \log \int g(\boldsymbol{\theta}_{m_1}|m_1) d\boldsymbol{\theta}_{m_1} - \\ & \log \int f(\boldsymbol{n}|m_2, \boldsymbol{\theta}_{m_2}) g(\boldsymbol{\theta}_{m_2}|m_2) + \log \int g(\boldsymbol{\theta}_{m_2}|m_2) d\boldsymbol{\theta}_{m_2} \end{aligned}$$

In this application, the prior approximations $\log \int g(\boldsymbol{\theta}_{m_i}|m_i) d\boldsymbol{\theta}_{m_i}$ will be obtained using the bridge sampler, and the posterior approximations $\log \int g(\boldsymbol{n}|m_i, \boldsymbol{\theta}_{m_i}) g(\boldsymbol{\theta}_{m_i}|m_i)$ obtained using Laplace's method. This is sensible as the sample size is large, with cell counts of at least 5 in 80% of the cells. The results are presented in table 5.3, which gives the estimated log Bayes factors for several models, taken against the most probable hierarchical model $AC + BC + AD + AE + CE + DE + F$, for a prior where $a(\boldsymbol{i}) = \frac{1}{64}$. A sample size of 5000 was used for the prior estimates.

Hierarchical Log-linear Model	Log Bayes Factor Estimate	Log Bayes Factor (D&F)
$AC + BC + AD + AE + BE + DE + F$	0.49	0.57
$AC + BC + AD + AE + BE + CE + DE + F$	1.35	1.34
$AC + BC + AD + AE + CE + DE + BF$	1.79	1.42
$BC + ACE + ADE + F$	8.25	> 6

Table 5.3: Estimated Bayes factors for Heart Disease data

The top three models in the table are the most probable hierarchical models (identified by Dellaportas and Forster), and the fourth is the most probable decomposable model. There is a good deal of agreement between the

bridge/Laplace estimates and those obtained by Dellaportas and Forster. Note that we are using different prior densities here, so don't expect exact agreement with their results.

5.6 Discussion

The aim of this Chapter was to develop a method of approximating the normalising constants for conditional Dirichlet distributions, as these are often intractable. Laplace's method was applied in section 5.3, and the approximations were shown to converge to the true values in certain known situations for increasing sample sizes. Indeed, a rule of thumb is that Laplace's method will furnish a good approximation to the normalising constant when the cell count is at least 5 in 80% of the cells. Bridge sampling was introduced in section 5.4, and this method of approximation was shown to provide excellent accuracy in all cases.

The methods were both applied in section 5.5 to some real data, in order to estimate Bayes factors for several models. The results were compared with those obtained by Dellaportas and Forster (1999), and found to be similar.

To conclude, the methods presented may be applied to obtain the normalising constants for both prior and posterior conditional Dirichlet distributions for any log-linear model, and hence Bayes factors may be calculated to compare competing models.

Chapter 6

Jeffreys' Prior

6.1 Introduction

Priors based on the Normal and, in particular, Dirichlet distributions have been discussed in previous Chapters, focussing on the use of these priors as reference priors. The formulation of these priors as reference priors is done by suitable choice of distribution parameters. Another popular choice of distribution for use in reference analysis is *Jeffreys' prior* (Jeffreys 1946), which is a reference prior by definition.

Several properties of Jeffreys' prior make it an attractive distribution for reference analyses. One of these is the invariance to reparameterisation of the model, a feature which may be exploited here as log-linear models admit a number of equivalent parameterisations. Many authors have highlighted other properties of Jeffreys' priors. For example, Box and Tiao (1973) argued Jeffreys' prior to be approximately noninformative with respect to certain criteria, and Bernardo (1979) found that the prior which maximises the missing information is Jeffreys' prior, though only under regularity conditions and when there are no nuisance parameters. Kass (1989) discussed the geometric interpretation of Jeffreys' prior, and both he and Bernardo looked at the advantages of this prior by focussing on

its interpretation in an information metric, with respect to which it is a uniform measure.

However, there are also some disadvantages with Jeffreys' prior, which limit its effectiveness as a reference prior in certain situations. It has been pointed out by Bernardo and others that if a Jeffreys' prior is derived on all the parameters in a multiparameter situation, then the priors on the margins will not necessarily be noninformative (this will be highlighted with an example in a later section). This is a particular problem when we are interested in a subset of the parameters, with the others being nuisance parameters.

Jeffreys' prior is investigated in detail here since, on balance, it is still considered a useful prior in reference analyses, and is widely used. In particular, it may be a useful distribution for model selection problems in situations where a noninformative prior is required.

6.1.1 Formal Definition

Jeffreys' prior is defined as being proportional to the square root of the determinant of the Fisher information matrix $I(\boldsymbol{\theta})$, where $I(\boldsymbol{\theta})$ is given by

$$I(\boldsymbol{\theta}) = E\left[-\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T} f(\boldsymbol{n}|\boldsymbol{\theta})\right]$$

for likelihood function $f(\boldsymbol{n}|\boldsymbol{\theta})$. Hence Jeffreys' prior, $f(\boldsymbol{\theta})$, is defined as

$$f(\boldsymbol{\theta}) \propto \left|E\left[-\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T} f(\boldsymbol{n}|\boldsymbol{\theta})\right]\right|^{\frac{1}{2}} \tag{6.1}$$

6.2 Jeffreys' Prior for Log-Linear Models

Ibrahim and Laud (1991) investigated the use of Jeffreys' prior in the reference analysis of generalised linear models, and in particular gave two theorems supporting the use of Jeffreys' priors in certain cases.

An expression for Jeffreys' prior for generalised linear models in terms of canonical parameters, scale parameter, weights and design matrix is given by

$$f(\boldsymbol{\beta}) \propto \left| X^T W V(\boldsymbol{\beta}) \boldsymbol{\Delta}^2(\boldsymbol{\beta}) X \right|^{\frac{1}{2}} \tag{6.2}$$

where X is the design matrix and W is a diagonal matrix of weights. Further, $V(\boldsymbol{\beta})$ and $\boldsymbol{\Delta}(\boldsymbol{\beta})$ are diagonal matrices with i-th diagonal elements $v_i = \frac{d^2 b(\theta_i)}{d\theta_i^2}$ and $\delta_i = \frac{d\theta_i}{d\eta_i}$ respectively, where θ is the canonical parameter and $\eta_i = x_i^T \boldsymbol{\beta}$ is the linear predictor (x_i is the i-th row of the design matrix X).

In the case of a Normal linear regression model, the use of Jeffreys' prior results in a tractable posterior distribution, and this is also true of a linearised nonlinear regression model. However, Ibrahim and Laud discovered that, in general, the posteriors resulting from generalised linear models are not tractable, apart from some special cases for certain models. Nevertheless, they showed that Jeffreys' prior does, for most models, lead to proper posterior distributions. This provides motivation for our derivation of Jeffreys' prior for log-linear models.

6.2.1 Derivation

In this section, the Jeffreys' prior for any given log-linear model with design matrix X will be derived. The Hessian matrix of second derivatives for a conditional Dirichlet density was shown in section 4.1.2 to equal

$$\frac{\partial^2}{\partial \boldsymbol{\beta} \boldsymbol{\beta}^T} \log f(\boldsymbol{\beta}) = -\alpha X^T \left(diag \boldsymbol{p}(\boldsymbol{\beta}) - \boldsymbol{p}(\boldsymbol{\beta}) \boldsymbol{p}(\boldsymbol{\beta})^T \right) X$$

The conjugacy of the conditional Dirichlet distribution to the multinomial likelihood allows us use this expression to obtain the Hessian for the likelihood function:

$$\frac{\partial^2}{\partial\boldsymbol{\beta}\boldsymbol{\beta}^T}\log f(\boldsymbol{n}|\boldsymbol{p}(\boldsymbol{\beta})) = -nX^T\left(diag\boldsymbol{p}(\boldsymbol{\beta})-\boldsymbol{p}(\boldsymbol{\beta})\boldsymbol{p}(\boldsymbol{\beta})^T\right)X$$

This equation may now be used in conjunction with expression (6.1) to define Jeffreys' prior for a log-linear model

$$\begin{aligned} f(\boldsymbol{\beta}) &\propto \left|E\left[-\frac{\partial^2}{\partial\boldsymbol{\beta}\boldsymbol{\beta}^T}f(\boldsymbol{n}|\boldsymbol{p}(\boldsymbol{\beta}))\right]\right|^{\frac{1}{2}} \\ &\propto \left|nX^T\left(diag\boldsymbol{p}(\boldsymbol{\beta})-\boldsymbol{p}(\boldsymbol{\beta})\boldsymbol{p}(\boldsymbol{\beta})^T\right)X\right|^{\frac{1}{2}} \\ &\propto \left|X^T\left(diag\boldsymbol{p}(\boldsymbol{\beta})-\boldsymbol{p}(\boldsymbol{\beta})\boldsymbol{p}(\boldsymbol{\beta})^T\right)X\right|^{\frac{1}{2}} \end{aligned} \tag{6.3}$$

Hence we have a general expression for the Jeffreys' prior for any log-linear model, given by design matrix X. It is of a similar form to that obtained by Ibrahim and Laud (expression 6.2). However, although this expression seems straightforward, extensive investigation and application of the formula to a range of models did not, in general, result in any further simplification.

6.2.2 Jeffreys' Prior for Saturated Log-Linear Models

Although expression (6.3) does not, in general, allow the Jeffreys' prior to be expressed in a tractable form, it may be used to derive Jeffreys' prior for any saturated model. Although this is a widely known result, this derivation is presented here to illustrate the use of (6.3).

Let the design matrix X be such that $X^T = (I_{n-1}|-\mathbf{1})$, *i.e.* an $(n-1)\times(n-1)$ identity matrix augmented with a column of -1's (for a symmetric logit parameterisation, where $\boldsymbol{\beta} = \boldsymbol{\theta}_{\backslash n}$). Now consider $G = I_{n-1} - \frac{1}{n}J_{n-1}$, where J is

a matrix of $1's$.

$$(XG)^T = (G|\mathbf{1}) = \left(\left[I_{n-1} - \frac{1}{n}J_{n-1}\right] | -\frac{1}{n}\mathbf{1}\right)$$

Then

$$\begin{aligned} GX^T\left(diag(\boldsymbol{p}(\boldsymbol{\beta}))-\boldsymbol{p}(\boldsymbol{\beta})\boldsymbol{p}(\boldsymbol{\beta})^T\right)XG &= \left[(I|0) - \frac{1}{n}(J|\mathbf{1})\right]\left[diag(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^T\right] \times \\ &\quad \left[(I|0) - \frac{1}{n}(J|\mathbf{1})\right]^T \\ &= (I|0)^T\left(diag(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^T\right)(I|0) \end{aligned}$$

as $(J|1)(\text{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^T) = 0$. Hence

$$GX^T\left(diag(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^T\right)XG = \left(diag(\boldsymbol{p}_{\backslash n}) - \boldsymbol{p}_{\backslash n}\left(\boldsymbol{p}_{\backslash n}\right)^T\right)$$

where $\boldsymbol{p}_{\backslash n}$ is vector $\boldsymbol{p}$ with the last element removed. Now, we know that for vectors $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$

$$\left|diag(\mathbf{a}) + \mathbf{b}\mathbf{c}^T\right| = \prod a_i\left(1 + \sum \frac{b_i c_i}{a_i}\right)$$

Hence

$$\begin{aligned} \left|GX^T\left(diag(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^T\right)XG\right| &= \prod_{\boldsymbol{i}\backslash\boldsymbol{i}_n} p(\boldsymbol{i})\left(1 + \sum_{\boldsymbol{i}\backslash\boldsymbol{i}_n} -p(\boldsymbol{i})\right) \\ &= \left(\prod_{\boldsymbol{i}\backslash\boldsymbol{i}_n} p(\boldsymbol{i})\right)p(\boldsymbol{i}_n) \\ &= \prod_{\boldsymbol{i}} p(\boldsymbol{i}) \end{aligned}$$

where $p(\boldsymbol{i}_n)$ is the final element of vector $\boldsymbol{p}$.

However,

$$\begin{aligned}\left|GX^T\left(diag(\boldsymbol{p})-\boldsymbol{p}\boldsymbol{p}^T\right)XG\right| &= |G|\left|X^T\left(diag(\boldsymbol{p})-\boldsymbol{p}\boldsymbol{p}^T\right)X\right||G| \\ &= \frac{\left|X^T\left(diag(\boldsymbol{p})-\boldsymbol{p}\boldsymbol{p}^T\right)X\right|}{n^2}\end{aligned}$$

So

$$\left|X^T\left(diag(\boldsymbol{p})-\boldsymbol{p}\boldsymbol{p}^T\right)X\right|^{\frac{1}{2}} = n^2\prod_{\boldsymbol{i}} p(\boldsymbol{i})^{\frac{1}{2}}$$

hence

$$f(\boldsymbol{\beta}) \propto \prod_{\boldsymbol{i}} p(\boldsymbol{i},\boldsymbol{\beta})^{\frac{1}{2}}$$

is the Jeffreys' prior for $\boldsymbol{\beta}$ and

$$f(\boldsymbol{p}) \propto \prod_{\boldsymbol{i}} p(\boldsymbol{i})^{\frac{1}{2}} \times |J|$$

Multiplication by the Jacobian, $|J|$, of the transformation from $\boldsymbol{\beta}$ (which is equal to $\boldsymbol{\theta}_{\backslash n}$, where $\boldsymbol{\theta}$ is the symmetric logit) to $\boldsymbol{p}$, found earlier to be equal to $\prod_{\boldsymbol{i}} p(\boldsymbol{i})^{-1}$, gives

$$f(\boldsymbol{p}) \propto \prod_{\boldsymbol{i}} p(\boldsymbol{i})^{-\frac{1}{2}}$$

Hence, the Jeffreys' prior for a saturated log-linear model follows, as expected, a Dirichlet($\frac{1}{2}\mathbf{1}$) distribution.

Note that for a multiway table with a saturated model, and hence with a Dirichlet($\frac{1}{2}\mathbf{1}$) Jeffreys' distribution, the marginal probabilities may have marginal distributions which are not obviously noninformative. For example, using the standard notation, the distribution of the margin corresponding to variable C would follow Dirichlet($\frac{|I|}{2|I_C|}\mathbf{1}$), and of course $\frac{|I|}{|I_C|}$ may be large. This is one of the well-known disadvantages of Jeffreys' prior as a reference prior.

6.3 Jeffreys' Prior for Decomposable Log-Linear Models

In the previous section, an expression was derived (6.3) for the Jeffreys' prior for a log-linear model. However, the form of this expression did not, in general, admit distributions which were obviously tractable. In this section, an alternative derivation for Jeffreys' prior will be developed for decomposable models.

6.3.1 Derivation

As defined in section 6.1.1, Jeffreys' prior is proportional to the square root of the determinant of the Fisher information matrix $I(\boldsymbol{\theta})$, where

$$I(\boldsymbol{\theta}) = E\left[-\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T} f(\boldsymbol{n}|\boldsymbol{\theta})\right]$$

for likelihood function $f(\boldsymbol{n}|\boldsymbol{\theta})$.

In this derivation, the parameterisation based on logits of conditional probabilities resulting from the directed representation of the log-linear model will be used, as it was in the proof in section 3.3. Note, however, that because of this these results are restricted to those models that are decomposable. As in section 3.3, for clarity, bold type will not necessarily be used to represent vectors; the levels and dimension of quantities should be apparent by subscripts, where necessary.

Let the model be represented by a directed graph, and suppose that a perfect numbering of vertices has been obtained. The set of factors is denoted by Γ, and for each factor $\gamma \in \Gamma$, I_γ is the set of levels of this factor. We may obtain a perfect numbering of Γ, which assigns an order to this set, which may be written $\Gamma = \{1, 2, \ldots, m\}$.

The model is decomposable, and so the cell probability may be expressed as

$$p(\boldsymbol{i}) = \prod_{\gamma} P(\gamma = i_\gamma | pa(\gamma) = i_{pa(\gamma)})$$

The log-likelihood function under multinomial sampling is given by

$$\begin{aligned} \log f &= \log \prod_{\boldsymbol{i}} p(\boldsymbol{i})^{n(\boldsymbol{i})} \\ &= \sum_{\boldsymbol{i}} n(\boldsymbol{i}) \log p(\boldsymbol{i}) \end{aligned}$$

which for the decomposable model may be written as

$$\begin{aligned} \log f &= \sum_{\boldsymbol{i}} n(\boldsymbol{i}) \sum_{\gamma} \log P(\gamma = i_\gamma | pa(\gamma) = i_{pa(\gamma)}) \\ &= \sum_{\boldsymbol{i}} \sum_{\gamma} n(\boldsymbol{i}) \log P(\gamma = i_\gamma | pa(\gamma) = i_{pa(\gamma)}) \\ &= \sum_{\gamma} \sum_{i_{\gamma, pa(\gamma)}} n(i_\gamma, i_{pa(\gamma)}) \log P(\gamma = i_\gamma | pa(\gamma) = i_{pa(\gamma)}) \end{aligned}$$

We now apply the reference cell logit parameterisation used in section 3.3, where ϕ_γ is defined by

$$\phi_\gamma(i_\gamma | i_{pa(\gamma)}) = \log \left(\frac{P(\gamma = i_\gamma | pa(\gamma) = i_{pa(\gamma)})}{P(\gamma = 1 | pa(\gamma) = i_{pa(\gamma)})} \right)$$

with inverse transformation

$$P(\gamma = i_\gamma | pa(\gamma) = i_{pa(\gamma)}) = \frac{\exp\left\{\phi_\gamma(i_\gamma | i_{pa(\gamma)})\right\}}{\sum_{i_\gamma=1}^{|I_\gamma|} \exp\left\{\phi_\gamma(i_\gamma | i_{pa(\gamma)})\right\}}$$

where $|I_\gamma|$ is the number of levels of factor γ. Note that $\phi_\gamma(1 | i_{pa(\gamma)}) = 0$ for any $\gamma, pa(\gamma)$.

The log-likelihood may now be re-written as

$$\begin{aligned}\log f &= \sum_{\gamma}\sum_{i_{\gamma,pa(\gamma)}} n(i_\gamma, i_{pa(\gamma)}) \log\left[\frac{\exp\left\{\phi_\gamma(i_\gamma|i_{pa(\gamma)})\right\}}{\sum_{j_\gamma}\exp\left\{\phi_\gamma(j_\gamma|i_{pa(\gamma)})\right\}}\right] \\ &= \sum_{\gamma}\sum_{i_{\gamma,pa(\gamma)}} n(i_\gamma, i_{pa(\gamma)}) \left[\phi_\gamma(i_\gamma|i_{pa(\gamma)}) - \log\left(\sum_{j_\gamma}\exp\left\{\phi_\gamma(j_\gamma|i_{pa(\gamma)})\right\}\right)\right]\end{aligned}$$

In order to determine the Jeffreys' prior, we must evaluate the full set of second-order partial derivatives with respect to the ϕ parameters. The first derivative is given by

$$\begin{aligned}\frac{\partial \log f}{\partial \phi_\gamma(i_\gamma|i_{pa(\gamma)})} &= n(i_\gamma, i_{pa(\gamma)}) - \frac{\exp\left\{\phi_\gamma(i_\gamma|i_{pa(\gamma)})\right\}\sum_{j_\gamma} n(j_\gamma, i_{pa(\gamma)})}{\sum_{j_\gamma}\exp\left\{\phi_\gamma(j_\gamma|i_{pa(\gamma)})\right\}} \\ &= n(i_\gamma, i_{pa(\gamma)}) - \frac{\exp\left\{\phi_\gamma(i_\gamma|i_{pa(\gamma)})\right\} n(i_{pa(\gamma)})}{\sum_{j_\gamma}\exp\left\{\phi_\gamma(j_\gamma|i_{pa(\gamma)})\right\}}\end{aligned}$$

for a particular γ, i_γ and $i_{pa(\gamma)}$. Several sets of second derivatives must be calculated in order to construct the Fisher information matrix. First of all, note that if $\gamma_1 \neq \gamma_2$ the second derivative term

$$\frac{\partial^2 \log f}{\partial \phi_{\gamma_1}(i_{\gamma_1}|i_{pa(\gamma_1)})\partial \phi_{\gamma_2}(i_{\gamma_2}|i_{pa(\gamma_2)})} = 0$$

Hence, the matrix is block diagonal, with major blocks corresponding to each variable γ. Within each of these blocks, further sub-blocks exist – these will be described below.

For a particular variable γ, let us first consider the second derivative $\frac{\partial^2 \log f}{\partial \phi_\gamma(i_\gamma|i_{pa(\gamma)})^2}$ (which will correspond to the terms on the diagonal in this major block):

$$\begin{aligned}
\frac{\partial^2 \log f}{\partial \phi_\gamma(i_\gamma|i_{pa(\gamma)})^2} &= \frac{\partial}{\partial \phi_\gamma(i_\gamma|i_{pa(\gamma)})}\left(n(i_\gamma, i_{pa(\gamma)}) - \frac{\exp\left\{\phi_\gamma(i_\gamma|i_{pa(\gamma)})\right\} n(i_{pa(\gamma)})}{\sum_{j_\gamma}\exp\left\{\phi_\gamma(j_\gamma|i_{pa(\gamma)})\right\}}\right) \\
&= -\Bigg[\exp\left\{\phi_\gamma(i_\gamma|i_{pa(\gamma)})\right\} n(i_{pa(\gamma)}) \sum_{j_\gamma}\exp\left\{\phi_\gamma(j_\gamma|i_{pa(\gamma)})\right\} - \\
&\qquad \exp\left\{\phi_\gamma(i_\gamma|i_{pa(\gamma)})\right\}^2 n(i_{pa(\gamma)})\Bigg] \Bigg/ \left[\sum_{j_\gamma}\exp\left\{\phi_\gamma(j_\gamma|i_{pa(\gamma)})\right\}\right]^2 \\
&= -n(i_{pa(\gamma)})\left[\frac{\exp\left\{\phi_\gamma(i_\gamma|i_{pa(\gamma)})\right\}}{\sum_{j_\gamma}\exp\left\{\phi_\gamma(j_\gamma|i_{pa(\gamma)})\right\}} - \left(\frac{\exp\left\{\phi_\gamma(i_\gamma|i_{pa(\gamma)})\right\}}{\sum_{j_\gamma}\exp\left\{\phi_\gamma(j_\gamma|i_{pa(\gamma)})\right\}}\right)^2\right] \\
&= -n(i_{pa(\gamma)})\left[P(\gamma = i_\gamma|pa(\gamma) = i_{pa(\gamma)}) - P(\gamma = i_\gamma|pa(\gamma) = i_{pa(\gamma)})^2\right]
\end{aligned}$$

The terms above correspond to a particular term γ, as well as a specific level of this variable (i_γ) and the set of levels of the parents ($i_{pa(\gamma)}$). Within the major block, consider the second derivative term

$$\begin{aligned}
\frac{\partial^2 \log f}{\partial \phi_\gamma(i_\gamma|i_{pa(\gamma)})\partial \phi_\gamma(i_\gamma|j_{pa(\gamma)})} &= \frac{\partial}{\partial \phi_\gamma(i_\gamma|j_{pa(\gamma)})}\left(n(i_\gamma, i_{pa(\gamma)}) - \frac{\exp\left\{\phi_\gamma(i_\gamma|i_{pa(\gamma)})\right\} n(i_{pa(\gamma)})}{\sum_{j_\gamma}\exp\left\{\phi_\gamma(j_\gamma|i_{pa(\gamma)})\right\}}\right) \\
&= 0 \qquad \text{if } j_{pa(\gamma)} \neq i_{pa(\gamma)}
\end{aligned}$$

Hence, the block corresponding to a particular γ is itself block diagonal, with sub-blocks corresponding to each set of $i_{pa(\gamma)}$. Within each sub-block, we may evaluate the second derivative terms:

$$\begin{aligned}
\frac{\partial^2 \log f}{\partial \phi_\gamma(i_\gamma|i_{pa(\gamma)})\partial \phi_\gamma(j_\gamma|i_{pa(\gamma)})} &= \frac{\exp\left\{\phi_\gamma(i_\gamma|i_{pa(\gamma)})\right\}\exp\left\{\phi_\gamma(j_\gamma|i_{pa(\gamma)})\right\} n(i_{pa(\gamma)})}{\left(\sum_{j_\gamma}\exp\left\{\phi_\gamma(j_\gamma|i_{pa(\gamma)})\right\}\right)^2} \\
&= P(\gamma = i_\gamma|pa(\gamma) = i_{pa(\gamma)})P(\gamma = j_\gamma|pa(\gamma) = i_{pa(\gamma)}) \times \\
&\qquad n(i_{pa(\gamma)})
\end{aligned}$$

The next step in the construction of Jeffreys' prior is to take the expectations

of the second-order derivatives above. This produces

$$\begin{aligned}
E\left[-\frac{\partial^2 \log f}{\partial \phi_\gamma(i_\gamma|i_{pa(\gamma)})^2}\right] &= E\left[\left\{P(\gamma = i_\gamma|pa(\gamma) = i_{pa(\gamma)}) - P(\gamma = i_\gamma|pa(\gamma) = i_{pa(\gamma)})^2\right\} \times \right. \\
&\qquad \left. n(i_{pa(\gamma)})\right] \\
&= \left[P(\gamma = i_\gamma|pa(\gamma) = i_{pa(\gamma)}) - P(\gamma = i_\gamma|pa(\gamma) = i_{pa(\gamma)})^2\right] \times \\
&\qquad E\left[n(i_{pa(\gamma)})\right] \\
&\propto \left[P(\gamma = i_\gamma|pa(\gamma) = i_{pa(\gamma)}) - P(\gamma = i_\gamma|pa(\gamma) = i_{pa(\gamma)})^2\right] \times \\
&\qquad P(pa(\gamma) = i_{pa(\gamma)}) \\
&= \left[P(\gamma = i_\gamma|pa(\gamma) = i_{pa(\gamma)}) - P(\gamma = i_\gamma|pa(\gamma) = i_{pa(\gamma)})^2\right] \times \\
&\qquad \sum_{i_{\overline{pa}(\gamma)}} p(\boldsymbol{i}) \\
&= \left[P(\gamma = i_\gamma|pa(\gamma) = i_{pa(\gamma)}) - P(\gamma = i_\gamma|pa(\gamma) = i_{pa(\gamma)})^2\right] \times \\
&\qquad \sum_{i_{\overline{pa}(\gamma)}} \prod_{\gamma'} P(\gamma' = i_{\gamma'}|pa(\gamma') = i_{pa(\gamma')})
\end{aligned}$$

where $\overline{pa}(\gamma) = \Gamma \setminus pa(\gamma)$ (so note that $\gamma \in \overline{pa}(\gamma)$), and using the expansion $p(\boldsymbol{i}) = \prod_\gamma P(\gamma = i_\gamma|pa(\gamma) = i_{pa(\gamma)})$. Furthermore,

$$\begin{aligned}
E\left[-\frac{\partial^2 \log f}{\partial \phi_\gamma(i_\gamma|i_{pa(\gamma)})\partial \phi_\gamma(j_\gamma|i_{pa(\gamma)})}\right] &= -E[n(i_{pa(\gamma)})P(\gamma = i_\gamma|pa(\gamma) = i_{pa(\gamma)}) \times \\
&\qquad P(\gamma = j_\gamma|pa(\gamma) = i_{pa(\gamma)})] \\
&= -E\left[n(i_{pa(\gamma)})\right] P(\gamma = i_\gamma|pa(\gamma) = i_{pa(\gamma)}) \times \\
&\qquad P(\gamma = j_\gamma|pa(\gamma) = i_{pa(\gamma)}) \\
&\propto -P(\gamma = i_\gamma|pa(\gamma) = i_{pa(\gamma)}) \sum_{i_{\overline{pa}(\gamma)}} P(\boldsymbol{i}) \times \\
&\qquad P(\gamma = j_\gamma|pa(\gamma) = i_{pa(\gamma)}) \\
&= -P(\gamma = i_\gamma|pa(\gamma) = i_{pa(\gamma)}) \times \\
&\qquad P(\gamma = j_\gamma|pa(\gamma) = i_{pa(\gamma)}) \times \\
&\qquad \sum_{i_{\overline{pa}(\gamma)}} \prod_{\gamma'} P(\gamma' = i_{\gamma'}|pa(\gamma') = i_{pa(\gamma')})
\end{aligned}$$

It is now possible to construct the Fisher information matrix, defined as $I(\phi) = E\left[-\frac{\partial^2}{\partial\phi\partial\phi^T} f(\boldsymbol{n}|\boldsymbol{\phi})\right]$. As described above, this matrix is block diagonal, with a major block corresponding to every model term γ, and sub-blocks of size $(|I_\gamma| - 1) \times (|I_\gamma| - 1)$ for each set of $i_{pa(\gamma)}$. Hence $|I(\boldsymbol{\phi})|$ is equal to the product over all blocks of the sub-block determinants.

Within a sub-block, note that the term $\delta_{\gamma,i_{pa(\gamma)}} = P(pa(\gamma) = i_{pa(\gamma)}) = \sum_{i_{\overline{pa}(\gamma)}} \prod_{\gamma'} P(\gamma' = i_{\gamma'}|pa(\gamma') = i_{pa(\gamma')})$ is a constant factor for each non-zero entry, and so this may be taken outside the determinant and raised to the power $|I_\gamma|-1$. The remainder of the sub-block may now be written in the form $\left(diag(\mathbf{a}) + \mathbf{b}\mathbf{c}^T\right)$, where $\mathbf{a} = \mathbf{b} = -\mathbf{c} = \left(P(\gamma = i_\gamma|pa(\gamma) = i_{pa(\gamma)})\right)$ are vectors of length $|I_\gamma| - 1$. We now apply the expression

$$\begin{aligned}
\left|diag(\mathbf{a}) + \mathbf{b}\mathbf{c}^T\right| &= \left(1 + \sum_l \frac{b_l c_l}{a_l}\right) \prod_k a_k \\
&= \left(1 - \sum_{i_\gamma=2}^{|I_\gamma|} P(\gamma = i_\gamma|pa(\gamma) = i_{pa(\gamma)})\right) \prod_{i_\gamma=2}^{|I_\gamma|} P(\gamma = i_\gamma|pa(\gamma) = i_{pa(\gamma)}) \\
&= P(\gamma = 1|pa(\gamma) = i_{pa(\gamma)}) \prod_{i_\gamma=2}^{|I_\gamma|} P(\gamma = i_\gamma|pa(\gamma) = i_{pa(\gamma)}) \\
&= \prod_{i_\gamma} P(\gamma = i_\gamma|pa(\gamma) = i_{pa(\gamma)})
\end{aligned}$$

Hence the determinant of the sub-block corresponding to $\gamma, i_{pa(\gamma)}$ is given by

$$\delta_{\gamma,pa(\gamma)}^{|I_\gamma|-1} \prod_{i_\gamma} P(\gamma = i_\gamma|pa(\gamma) = i_{pa(\gamma)})$$

where $\delta_{\gamma,i_{pa(\gamma)}} = \sum_{i_{\overline{pa}(\gamma)}} \prod_{\gamma'} P(\gamma' = i_{\gamma'}|pa(\gamma') = i_{pa(\gamma')})$. Therefore, the determinant of the major block corresponding to a particular model term γ is

$$\prod_{i_{pa(\gamma)}} \delta_{\gamma,i_{pa(\gamma)}}^{|I_\gamma|-1} \prod_{i_\gamma} P(\gamma = i_\gamma|pa(\gamma) = i_{pa(\gamma)})$$

It is now possible to write down the Fisher information matrix, $I(\phi) \propto E\left[-\frac{\partial^2}{\partial\phi\partial\phi^T} f(\boldsymbol{n}|\boldsymbol{\phi})\right]$

$$\begin{aligned} |I(\phi)| &\propto \prod_{\gamma}\prod_{i_{pa(\gamma)}} \delta_{\gamma,i_{pa(\gamma)}}^{|I_\gamma|-1} \prod_{i_\gamma} P(\gamma = i_\gamma | pa(\gamma) = i_{pa(\gamma)}) \\ &= \prod_{\gamma}\prod_{i_{pa(\gamma)}} \left[\sum_{i_{\overline{pa}(\gamma)}} \prod_{\gamma'} P(\gamma' = i_{\gamma'} | pa(\gamma') = i_{pa(\gamma')})\right]^{|I_\gamma|-1} \times \\ &\quad \prod_{i_\gamma} P(\gamma = i_\gamma | pa(\gamma) = i_{pa(\gamma)}) \end{aligned}$$

Jeffreys' prior $f(\phi)$ is proportional to the square root of this determinant, though in order to obtain a prior for the conditional probabilities, rather than the conditional logits, we must multiply by the Jacobian $|J|$ of the transformation from ϕ to $\left\{P(\gamma = i_\gamma | pa(\gamma) = i_{pa(\gamma)})\right\}$. This was shown in section 3.3 to equal

$$\begin{aligned} |J| &= \prod_{\gamma}\prod_{i_{pa(\gamma)}}\prod_{i_\gamma} P(\gamma = i_\gamma | pa(\gamma) = i_{pa(\gamma)})^{-1} \\ &= \prod_{\gamma, i_\gamma, i_{pa(\gamma)}} P(\gamma = i_\gamma | pa(\gamma) = i_{pa(\gamma)})^{-1} \end{aligned}$$

This expression is now applied to determine Jeffreys' prior for a decomposable log-linear model, where $\boldsymbol{p}$ is now a collection of conditional probabilities such as $P(\gamma | i_{pa(\gamma)})$ rather than cell probabilities $p(\boldsymbol{i})$:

$$\begin{aligned} f(\boldsymbol{p}) &\propto |I(\boldsymbol{\phi})|^{\frac{1}{2}} |J| \\ &\propto \left[\prod_{\gamma}\prod_{i_{pa(\gamma)}} \left[\sum_{i_{\overline{pa}(\gamma)}} \prod_{\gamma'} P(\gamma' = i_{\gamma'} | pa(\gamma') = i_{pa(\gamma')})\right]^{|I_\gamma|-1} \times \right. \\ &\quad \left. \prod_{i_\gamma} P(\gamma = i_\gamma | pa(\gamma) = i_{pa(\gamma)})\right]^{\frac{1}{2}} \times \end{aligned}$$

$$\prod_{\gamma,i_\gamma,i_{pa(\gamma)}} P(\gamma = i_\gamma | pa(\gamma) = i_{pa(\gamma)})^{-1}$$

$$= \left[\prod_{\gamma} \prod_{i_{pa(\gamma)}} \left[\sum_{i_{\overline{pa}(\gamma)}} \prod_{\gamma'} P(\gamma' = i_{\gamma'} | pa(\gamma') = i_{pa(\gamma')}) \right]^{|I_\gamma|-1} \times \prod_{i_\gamma} P(\gamma = i_\gamma | pa(\gamma) = i_{pa(\gamma)}) \right]^{\frac{1}{2}} \times \left[\prod_{\gamma} \prod_{i_\gamma, i_{pa(\gamma)}} P(\gamma = i_\gamma | pa(\gamma) = i_{pa(\gamma)}) \right]^{-1}$$

$$f(\boldsymbol{p}) \propto \left(\prod_{\gamma} \prod_{i_{pa(\gamma)}} \left[\sum_{i_{\overline{pa}(\gamma)}} \prod_{\gamma'} P(\gamma' = i_{\gamma'} | pa(\gamma') = i_{pa(\gamma')}) \right]^{\frac{|I_\gamma|-1}{2}} \right) \times \prod_{\gamma} \prod_{i_\gamma, i_{pa(\gamma)}} P(\gamma = i_\gamma | pa(\gamma) = i_{pa(\gamma)})^{-\frac{1}{2}}$$

$$\propto \left(\prod_{\gamma} \prod_{i_{pa(\gamma)}} \delta_{\gamma, i_{pa(\gamma)}}^{\frac{|I_\gamma|-1}{2}} \right) \prod_{\gamma} \prod_{i_\gamma, i_{pa(\gamma)}} P(\gamma = i_\gamma | pa(\gamma) = i_{pa(\gamma)})^{-\frac{1}{2}} \quad (6.4)$$

where $\delta_{\gamma,i_{pa(\gamma)}} = P(pa(\gamma) = i_{pa(\gamma)}) = \sum_{i_{\overline{pa}(\gamma)}} \prod_{\gamma'} P(\gamma' = i_{\gamma'} | pa(\gamma') = i_{pa(\gamma')})$.

It is possible to use this result to express the Jeffreys' prior for any decomposable log-linear model in terms of conditional cell probabilities. This will be applied to a range of models in the next section. Note that the Jeffreys' prior consists of two parts – a Dirichlet part, where each $P(\gamma | pa(\gamma) = i_{pa(\gamma)})$ follows an independent $Dirichlet(\frac{1}{2}\mathbf{1})$ distribution, multiplied by a part which is a product of summation terms.

6.4 Examples of Jeffreys' Priors

In this section, the Jeffreys' priors for various decomposable log-linear models, including all those with up to and including four variables, will be derived.

6.4.1 Saturated Models

Equation (6.4) is applied here to derive a general expression for the Jeffreys' prior for a saturated log-linear model, parameterised using conditional probabilities. Suppose we have k variables, denoted A, B, C, D,... Such a model admits the decomposition $P(A)P(B|A)P(C|A,B)P(D|A,B,C)...$ The expression for the Jeffreys' prior (6.4) is

$$f(\boldsymbol{p}) \propto \left(\prod_{\gamma} \prod_{i_{pa(\gamma)}} \delta_{\gamma, i_{pa(\gamma)}}^{\frac{|I_\gamma|-1}{2}} \right) \prod_{\gamma} \prod_{i_\gamma, i_{pa(\gamma)}} P(\gamma = i_\gamma | pa(\gamma) = i_{pa(\gamma)})^{-\frac{1}{2}}$$

where $\delta_{\gamma, i_{pa(\gamma)}} = \sum_{i_{\overline{pa}(\gamma)}} \prod_{\gamma'} P(\gamma' = i_{\gamma'} | pa(\gamma') = i_{pa(\gamma')})$

The term $\delta_{\gamma, i_{pa(\gamma)}}$ may be calculated for each γ :

$$\begin{aligned}
\delta_{\gamma, i_{pa(\gamma)}} &= \sum_{i_{\overline{pa}(\gamma)}} \prod_{\gamma'} P(\gamma' = i_{\gamma'} | pa(\gamma') = i_{pa(\gamma')}) \\
&= \prod_{\gamma' \in pa(\gamma)} P(\gamma' = i_{\gamma'} | pa(\gamma') = i_{pa(\gamma')}) \sum_{i_{\overline{pa}(\gamma)}} \prod_{\gamma' \in \overline{pa}(\gamma')} P(\gamma' = i_{\gamma'} | pa(\gamma') = i_{pa(\gamma')}) \\
&= \prod_{\gamma' \in pa(\gamma)} P(\gamma' = i_{\gamma'} | pa(\gamma') = i_{pa(\gamma')}) \qquad (6.5)
\end{aligned}$$

where the term $\prod_{\gamma' \in pa(\gamma)} P(\gamma' = i_{\gamma'} | pa(\gamma') = i_{pa(\gamma')})$ may be taken outside the summation in this case, because for the saturated model the set of variables which preceed any γ in the perfect ordering is exactly $pa(\gamma)$.

It is therefore straightforward to write down expressions for any $\delta_{\gamma, i_{pa(\gamma)}}$, for example

$$\begin{aligned}
\delta_{C, i_A, i_B} &= \prod_{\gamma' \in \{A,B\}} P(\gamma' = i_{\gamma'} | pa(\gamma') = i_{pa(\gamma')}) \\
&= P(A = i_A) P(B = i_B | A = i_A)
\end{aligned}$$

An expression for Jeffreys' prior may now be derived, by substituting (6.5)

into (6.4):

$$
\begin{aligned}
f(\boldsymbol{p}) \propto & \prod_{\gamma} \prod_{i_{pa(\gamma)}} \left[\prod_{\gamma' \in pa(\gamma)} P(\gamma' = i_{\gamma'} | pa(\gamma') = i_{pa(\gamma')}) \right]^{\frac{|I_\gamma|-1}{2}} \times \\
& \prod_{\gamma'} \prod_{i_{\gamma'}, i_{pa(\gamma')}} P(\gamma' = i_{\gamma'} | pa(\gamma') = i_{pa(\gamma')})^{-\frac{1}{2}} \\
= & \prod_{\gamma} \prod_{i_{pa(\gamma)}} \prod_{\gamma' \in pa(\gamma)} P(\gamma' = i_{\gamma'} | pa(\gamma') = i_{pa(\gamma')})^{\frac{|I_\gamma|-1}{2}} \times \\
& \prod_{\gamma'} \prod_{i_{\gamma'}, i_{pa(\gamma')}} P(\gamma' = i_{\gamma'} | pa(\gamma') = i_{pa(\gamma')})^{-\frac{1}{2}} \\
= & \prod_{\gamma} \prod_{\gamma' \in pa(\gamma)} \prod_{i_{pa(\gamma)}} P(\gamma' = i_{\gamma'} | pa(\gamma') = i_{pa(\gamma')})^{\frac{|I_\gamma|-1}{2}} \times \\
& \prod_{\gamma'} \prod_{i_{\gamma'}, i_{pa(\gamma')}} P(\gamma' = i_{\gamma'} | pa(\gamma') = i_{pa(\gamma')})^{-\frac{1}{2}} \\
= & \prod_{\gamma} \prod_{\gamma' \in pa(\gamma)} \prod_{i_{pa(\gamma')}, i_{\gamma'}} P(\gamma' = i_{\gamma'} | pa(\gamma') = i_{pa(\gamma')})^{\left[\frac{|I_\gamma|-1}{2}\right] \omega_{\gamma,\gamma'}} \times \\
& \prod_{\gamma'} \prod_{i_{\gamma'}, i_{pa(\gamma')}} P(\gamma' = i_{\gamma'} | pa(\gamma') = i_{pa(\gamma')})^{-\frac{1}{2}}
\end{aligned}
$$

where $\omega_{\gamma,\gamma'} = \prod_{\gamma^*} |I_{\gamma^*}|^{\chi(\gamma^* \in [pa(\gamma) \setminus \{pa(\gamma') \cup \gamma'\}])}$ and $\chi(\cdot)$ is the indicator function. We now use the notation $\gamma' < \gamma$ to denote those variables γ' which preceed γ in the perfect ordering, and write

$$
\begin{aligned}
f(\boldsymbol{p}) \propto & \prod_{\gamma} \prod_{\gamma' < \gamma} \prod_{i_{pa(\gamma')}, i_{\gamma'}} P(\gamma' = i_{\gamma'} | pa(\gamma') = i_{pa(\gamma')})^{\left[\frac{|I_\gamma|-1}{2}\right] \omega_{\gamma,\gamma'}} \times \\
& \prod_{\gamma'} \prod_{i_{\gamma'}, i_{pa(\gamma')}} P(\gamma' = i_{\gamma'} | pa(\gamma') = i_{pa(\gamma')})^{-\frac{1}{2}} \\
= & \prod_{\gamma'} \prod_{\gamma > \gamma'} \prod_{i_{pa(\gamma')}, i_{\gamma'}} P(\gamma' = i_{\gamma'} | pa(\gamma') = i_{pa(\gamma')})^{\left[\frac{|I_\gamma|-1}{2}\right] \omega_{\gamma,\gamma'}} \times \\
& \prod_{\gamma'} \prod_{i_{\gamma'}, i_{pa(\gamma')}} P(\gamma' = i_{\gamma'} | pa(\gamma') = i_{pa(\gamma')})^{-\frac{1}{2}}
\end{aligned}
$$

$$= \prod_{\gamma'} \left[\prod_{\gamma>\gamma'} \prod_{i_{pa(\gamma')}, i_{\gamma'}} P(\gamma' = i_{\gamma'} | pa(\gamma') = i_{pa(\gamma')})^{\left[\frac{|I_\gamma|-1}{2}\right]\omega_{\gamma,\gamma'}} \times \right.$$
$$\left. \prod_{i_{\gamma'}, i_{pa(\gamma')}} P(\gamma' = i_{\gamma'} | pa(\gamma') = i_{pa(\gamma')})^{-\frac{1}{2}} \right]$$
$$= \prod_{\gamma'} \left[\prod_{i_{pa(\gamma')}, i_{\gamma'}} P(\gamma' = i_{\gamma'} | pa(\gamma') = i_{pa(\gamma')})^{\sum_{\gamma>\gamma'}\left[\frac{|I_\gamma|-1}{2}\right]\omega_{\gamma,\gamma'}} \times \right.$$
$$\left. \prod_{i_{\gamma'}, i_{pa(\gamma')}} P(\gamma' = i_{\gamma'} | pa(\gamma') = i_{pa(\gamma')})^{-\frac{1}{2}} \right]$$
$$= \prod_{\gamma'} \prod_{i_{\gamma'}, i_{pa(\gamma')}} P(\gamma' = i_{\gamma'} | pa(\gamma') = i_{pa(\gamma')})^{-\frac{1}{2}+\sum_{\gamma>\gamma'}\left[\frac{|I_\gamma|-1}{2}\right]\omega_{\gamma,\gamma'}}$$

Finally, we note that

$$\sum_{\gamma>\gamma'} \left[\prod_{\gamma^*} |I_{\gamma^*}|^{I(\gamma^* \in [pa(\gamma)\setminus\{pa(\gamma')\cup\gamma'\}])} \right] \left[\frac{|I_\gamma|-1}{2}\right] - \frac{1}{2} = \frac{\prod_{\gamma^*} |I_{\gamma^*}|^{\chi(\gamma^* \in [\Gamma\setminus\{\gamma\cup pa(\gamma)\}])}}{2} - 1$$

which follows directly from the result $a-1+a(b-1)+ab(c-1)+abc(d-1)+\ldots = abcd\ldots$. We may now write down Jeffreys' prior for the saturated model

$$f(\boldsymbol{p}) \propto \prod_{\gamma} \prod_{i_{pa(\gamma)}} P(\gamma = i_\gamma | pa(\gamma) = i_{pa(\gamma)})^{\frac{\prod_{\gamma^*} |I_{\gamma^*}|^{\chi(\gamma^* \in [\Gamma\setminus\{\gamma\cup pa(\gamma)\}])}}{2} - 1} \tag{6.6}$$

giving a product of independent Dirichlet distributions. It is straightforward to show that this distribution is equivalent to that obtained in section 6.2.2, $f(\boldsymbol{p}) \propto \prod_{\boldsymbol{i}} p(\boldsymbol{i})^{-\frac{1}{2}}$.

As an example, this expression will be applied to the four variable saturated model. In this case, we obtain

$$\frac{\prod_{\gamma^*} |I_{\gamma^*}|^{\chi(\gamma^* \in [\Gamma\setminus\{A\cup pa(A)\}])}}{2} = \frac{1}{2} |I_A|^{\chi(A \in [\Gamma\setminus\{A\cup pa(A)\}])} |I_B|^{\chi(B \in [\Gamma\setminus\{A\cup pa(A)\}])} \times$$
$$|I_C|^{\chi(C \in [\Gamma\setminus\{A\cup pa(A)\}])} |I_D|^{\chi(D \in [\Gamma\setminus\{A\cup pa(A)\}])}$$

$$= \frac{|I_A|^0 |I_B|^1 |I_C|^1 |I_D|^1}{2}$$
$$= \frac{|I_B| |I_C| |I_D|}{2}$$

$$\frac{\prod_{\gamma^*} |I_{\gamma^*}|^{\chi(\gamma^* \in [\Gamma \setminus \{\gamma \cup pa(\gamma)\}])}}{2} = \frac{|I_C| |I_D|}{2}$$

$$\frac{\prod_{\gamma^*} |I_{\gamma^*}|^{\chi(\gamma^* \in [\Gamma \setminus \{\gamma \cup pa(\gamma)\}])}}{2} = \frac{|I_D|}{2}$$

$$\frac{\prod_{\gamma^*} |I_{\gamma^*}|^{\chi(\gamma^* \in [\Gamma \setminus \{\gamma \cup pa(\gamma)\}])}}{2} = \frac{1}{2}$$

so that

$$f(\boldsymbol{p}) \propto \prod_{i_A, i_B, i_C, i_D} P(B = i_B | A = i_A)^{\frac{|I_C||I_D|}{2} - 1} P(C = i_C | A = i_A, B = i_B)^{\frac{|I_D|}{2} - 1} \times$$
$$P(A = i_A)^{\frac{|I_B||I_C||I_D|}{2} - 1} P(D = i_D | A = i_A, B = i_B, C = i_C)^{-\frac{1}{2}}$$

Expression (6.6) may be used to write down the Jeffreys' prior for the single variable model represented by the graph

This is given by

$$f(\boldsymbol{p}) \propto \prod_{i_A} P(A = i_A)^{-\frac{1}{2}}$$

which is a well-known result.

6.4.2 Block Independence

The Jeffreys' prior for a model which is represented graphically by a number of disconnected components is available directly from the separate Jeffreys' priors for each of the disconnected components. This follows from expression (6.4),

which may be written

$$
\begin{aligned}
f(\boldsymbol{p}) \;&\propto\; \left(\prod_{\gamma}\prod_{i_{pa(\gamma)}} \delta_{\gamma,i_{pa(\gamma)}}^{\frac{|I_\gamma|-1}{2}}\right)\prod_{\gamma}\prod_{i_\gamma,i_{pa(\gamma)}} P(\gamma=i_\gamma|pa(\gamma)=i_{pa(\gamma)})^{-\frac{1}{2}} \\
&= \prod_{\gamma}\left[\left(\prod_{i_{pa(\gamma)}} \delta_{\gamma,i_{pa(\gamma)}}^{\frac{|I_\gamma|-1}{2}}\right)\prod_{i_\gamma,i_{pa(\gamma)}} P(\gamma=i_\gamma|pa(\gamma)=i_{pa(\gamma)})^{-\frac{1}{2}}\right] \\
&= \prod_{j}\prod_{\gamma\in\Delta_j}\left[\left(\prod_{i_{pa(\gamma)}} \delta_{\gamma,i_{pa(\gamma)}}^{\frac{|I_\gamma|-1}{2}}\right)\prod_{i_\gamma,i_{pa(\gamma)}} P(\gamma=i_\gamma|pa(\gamma)=i_{pa(\gamma)})^{-\frac{1}{2}}\right]
\end{aligned}
$$

where each Δ_j is a disconnected component of the graph. For $i \neq j$, if $\gamma_1 \in \Delta_i$ and $\gamma_2 \in \Delta_j$ then $\gamma_2 \notin pa(\gamma_1)$ and *vice versa*. Hence the contribution of each disconnected component Δ_j to the Jeffreys' prior will be identical to the Jeffreys' prior for Δ_j as a model in its own right.

Application of this concept is straightforward. For example, in the model with k variables, all independent, there are k disconnected components, each containing a single variable. As stated in the previous section, Jeffreys' prior for the model for a single variable A is $f(\boldsymbol{p}) = \prod_{i_A} P(A=i_A)^{-\frac{1}{2}}$. Therefore the Jeffreys' prior for the model with k independent variables is given by

$$
f(\boldsymbol{p}) \propto \prod_{\gamma}\prod_{i_\gamma} P(\gamma=i_\gamma)^{-\frac{1}{2}}
$$

which is the product of the Jeffreys' priors on the cell margins. This confirms the expected result that the Jeffreys' prior for an independence model places Dirichlet($\frac{1}{2}\mathbf{1}$) distributions (which are Beta($\frac{1}{2},\frac{1}{2}$) for the 2-level case) independently on each margin.

In the remainder of this section, the Jeffreys' priors for models with up to and including four variables, which are distinct up to graph isomorphism (*i.e.* equivalent under permutations of variables), are considered in detail. Expressions for many other Jeffreys' priors are also given, obtained using the above results.

6.4.3 Two Variable Models

There are 2 distinct models with two variables – the independence model and the saturated model. Their Jeffreys' priors are directly obtained using the results on independence and saturated models respectively, and are given by

$$f(\boldsymbol{p}) \propto \prod_{i_A} P(A = i_A)^{-\frac{1}{2}} \prod_{i_B} P(B = i_B)^{-\frac{1}{2}}$$

and

$$f(\boldsymbol{p}) \propto \prod_{i_A} P(A = i_A)^{\frac{|I_A|}{2}-1} \prod_{i_A, i_B} P(B = i_B | A = i_A)^{-\frac{1}{2}}$$

6.4.4 Three Variable Models

There are 4 distinct graphical log-linear models with three variables, all of which are decomposable. Jeffreys' priors for three of these are obtainable using the previous results, and the other is derived here.

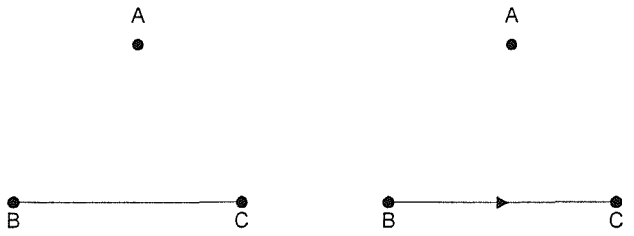
Independence Model

This model has Jeffreys' prior

$$f(\boldsymbol{p}) \propto \prod_{i_A} P(A = i_A)^{-\frac{1}{2}} \prod_{i_B} P(B = i_B)^{-\frac{1}{2}} \prod_{i_C} P(C = i_C)^{-\frac{1}{2}}$$

One Edge Model

This model represents the independence of one of the variables from the other two. Without loss of generality, consider the model which can be represented by the graphs



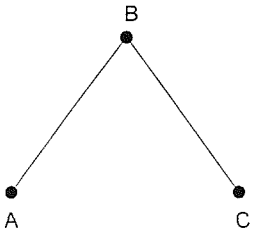
The graph is composed of two disconnected components, and so the theory on block independence may be used to provide the Jeffreys' prior in this case. The two components represent a saturated model with 2 variables and a single variable model, and so we obtain

$$f(\boldsymbol{p}) \propto \prod_{i_B} P(B = i_B)^{\frac{|I_C|}{2}-1} \prod_{i_A} P(A = i_A)^{-\frac{1}{2}} \prod_{i_B, i_C} P(C = i_C | B = i_B)^{-\frac{1}{2}}$$

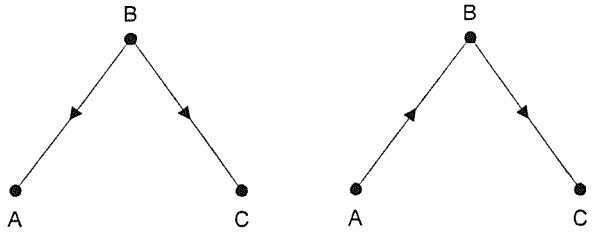
producing another Dirichlet prior.

Two Edges Model

This model represents the conditional independence of two of the variables, given the other. Without loss of generality, we shall consider the model with the undirected graph



The directed version admits two distinct decompositions – one is $P(B)P(A|B)P(C|B)$, and the other is $P(A)P(B|A)P(C|B)$. These are represented graphically below



These two parameterisations will be considered in turn.

Parameterisation One Suppose the model is parameterised using the decomposition $P(B)P(A|B)P(C|B)$. Then we have

γ	$pa(\gamma)$	$\overline{pa}(\gamma)$
A	B	A, C
B	$\emptyset$	A, B, C
C	B	A, C

Hence the terms $\delta_{\gamma, i_{pa(\gamma)}}$ may be calculated for each γ:

$$\begin{aligned}
\delta_{B,i_{pa(B)}} &= \sum_{i_A,i_B,i_C} \prod_{\gamma'} P(\gamma' = i_{\gamma'} | pa(\gamma') = i_{pa(\gamma')}) \\
&= \sum_{i_A,i_B,i_C} P(B = i_B)P(A = i_A|B = i_B)P(C = i_C|B = i_B) \\
&= 1
\end{aligned}$$

and

$$\begin{aligned}
\delta_{C,i_B} = \delta_{A,i_B} &= \sum_{i_A,i_C} \prod_{\gamma'} P(\gamma' = i_{\gamma'} | pa(\gamma') = i_{pa(\gamma')}) \\
&= \sum_{i_A,i_C} P(B = i_B)P(A = i_A|B = i_B)P(C = i_C|B = i_B) \\
&= P(B = i_B)
\end{aligned}$$

Application of equation (6.4) yields the Jeffreys' prior

$$\begin{aligned} f(\boldsymbol{p}) &\propto \left(\prod_{\gamma} \prod_{i_{pa(\gamma)}} \delta_{\gamma, i_{pa(\gamma)}}^{\frac{|I_\gamma|-1}{2}} \right) \prod_{\gamma} \prod_{i_\gamma, i_{pa(\gamma)}} P(\gamma = i_\gamma | pa(\gamma) = i_{pa(\gamma)})^{-\frac{1}{2}} \\ &= \prod_{i_B} P(B = i_B)^{\frac{|I_A|+|I_C|}{2}-1} \prod_{i_B} P(B = i_B)^{-\frac{1}{2}} \prod_{i_A, i_B} P(A = i_A | B = i_B)^{-\frac{1}{2}} \times \\ &\quad \prod_{i_B, i_C} P(C = i_C | B = i_B)^{-\frac{1}{2}} \\ &= \prod_{i_B} P(B = i_B)^{\frac{|I_A|+|I_C|-3}{2}} \prod_{i_A, i_B} P(A = i_A | B = i_B)^{-\frac{1}{2}} \prod_{i_B, i_C} P(C = i_C | B = i_B)^{-\frac{1}{2}} \end{aligned}$$

and so the distribution is, once again, a product of independent Dirichlet distributions on the conditionals. Note however that the distribution is not hyper Dirichlet, as the distribution of $P(B)$ is not consistent with the distributions of $P(A|B)$ and $P(C|B)$.

Parameterisation Two Now let the model be parameterised using the decomposition $P(A)P(B|A)P(C|B)$. Then we have

γ	$pa(\gamma)$	$\overline{pa}(\gamma)$
A	$\emptyset$	A, B, C
B	A	B, C
C	B	A, C

This time, the terms $\delta_{\gamma, i_{pa(\gamma)}}$ are equal to

$$\begin{aligned} \delta_A &= \sum_{i_A, i_B, i_C} \prod_{\gamma'} P(\gamma' = i_{\gamma'} | pa(\gamma') = i_{pa(\gamma')}) \\ &= \sum_{i_A, i_B, i_C} P(A = i_A) P(B = i_B | A = i_A) P(C = i_C | B = i_B) \\ &= 1 \end{aligned}$$

$$\begin{aligned}\delta_{B,i_A} &= \sum_{i_B,i_C}\prod_{\gamma'} P(\gamma' = i_{\gamma'}|pa(\gamma') = i_{pa(\gamma')}) \\ &= \sum_{i_B,i_C} P(A = i_A)P(B = i_B|A = i_A)P(C = i_C|B = i_B) \\ &= P(A = i_A)\end{aligned}$$

$$\begin{aligned}\delta_{C,i_B} &= \sum_{i_A,i_C}\prod_{\gamma'} P(\gamma' = i_{\gamma'}|pa(\gamma') = i_{pa(\gamma')}) \\ &= \sum_{i_A,i_C} P(A = i_A)P(B = i_B|A = i_A)P(C = i_C|B = i_B) \\ &= \sum_{i_A} P(A = i_A)P(B = i_B|A = i_A)\end{aligned}$$

Note that this last term, δ_{C,i_B}, may be written as $P(B = i_B)$, but this is not part of the required parameterisation, and so cannot be directly included in the expression for Jeffreys' prior.

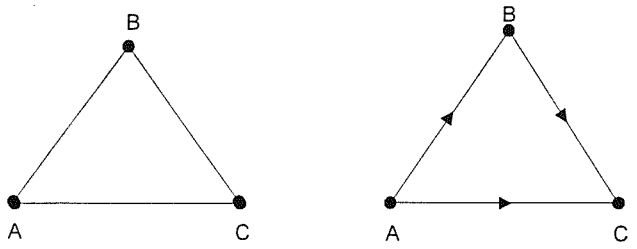
Application of equation (6.4) in this case gives

$$\begin{aligned}f(\boldsymbol{p}) &\propto \left(\prod_{\gamma}\prod_{i_{pa(\gamma)}} \delta_{\gamma,i_{pa(\gamma)}}^{\frac{|I_\gamma|-1}{2}}\right)\prod_{\gamma}\prod_{i_\gamma,i_{pa(\gamma)}} P(\gamma = i_\gamma|pa(\gamma) = i_{pa(\gamma)})^{-\frac{1}{2}} \\ &= \prod_{i_B}\left[\sum_{i_A} P(A = i_A)P(B = i_B|A = i_A)\right]^{\frac{|I_C|-1}{2}}\prod_{i_A} P(A = i_A)^{\frac{|I_B|}{2}-1} \times \\ &\quad \prod_{i_A,i_B} P(B = i_B|A = i_A)^{-\frac{1}{2}}\prod_{i_B,i_C} P(C = i_C|B = i_B)^{-\frac{1}{2}}\end{aligned}$$

and so in this instance, the distribution is not a product of independent Dirichlet distributions. Note however that this distribution is equivalent to that obtained previously using parameterisation one, and so it is clear that a careful choice of parameterisation for a model may produce a Jeffreys' prior which is easier to manipulate.

Saturated Model

The final model with three variables is the saturated model, as represented by the graphs



Application of equation (6.6) yields the Jeffreys' prior

$$f(\boldsymbol{p}) \propto \prod_{i_A} P(A=i_A)^{\frac{|I_B||I_C|}{2}-1} \prod_{i_A,i_B} P(B=i_B|A=i_A)^{\frac{|I_C|}{2}-1} \times \prod_{i_A,i_B,i_C} P(C=i_C|A=i_A,B=i_B)^{|I_B|-\frac{1}{2}}$$

6.4.5 Four Variable Models

There are 11 graphical models with four variables, though one of these is not decomposable, and six are obtained from previous priors using the block independence theory. The Jeffreys' priors for all the models are presented in this section, and four are considered in detail.

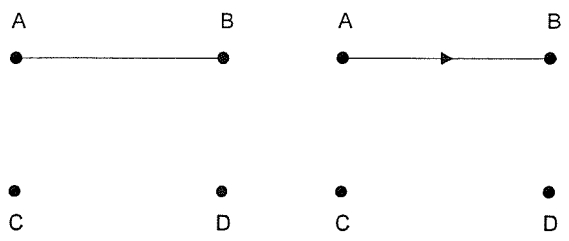
Independence Model

The Jeffreys' prior for this model is given by

$$f(\boldsymbol{p}) \propto \prod_{i_A} P(A=i_A)^{-\frac{1}{2}} \prod_{i_B} P(B=i_B)^{-\frac{1}{2}} \prod_{i_C} P(C=i_C)^{-\frac{1}{2}} \prod_{i_D} P(D=i_D)^{-\frac{1}{2}}$$

One Edge Model

There is one distinct model with 1 edge, represented by the graphs

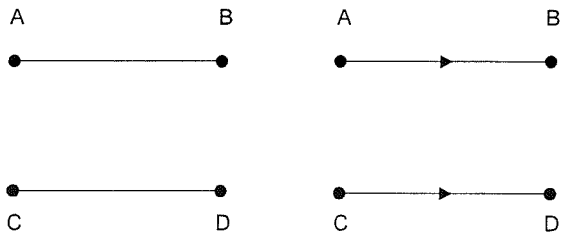


Application of the block independence theory gives the Jeffreys' prior for this model

$$f(\boldsymbol{p}) \propto \prod_{i_A} P(A=i_A)^{\frac{|I_B|}{2}-1} \prod_{i_A,i_B} P(B=i_B|A=i_A)^{-\frac{1}{2}} \prod_{i_C} P(C=i_C)^{-\frac{1}{2}} \times \prod_{i_D} P(D=i_D)^{-\frac{1}{2}}$$

Two Edge Models

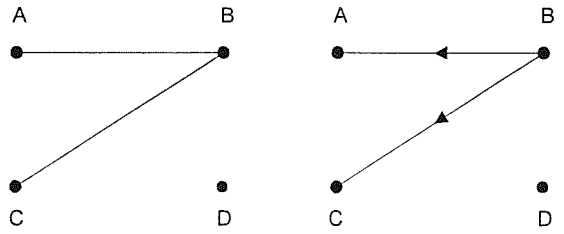
There are two distinct models in this section. The first of these is



The Jeffreys' prior for this model is

$$f(\boldsymbol{p}) \propto \prod_{i_A} P(A=i_A)^{\frac{|I_B|}{2}-1} \prod_{i_C} P(C=i_C)^{\frac{|I_D|}{2}-1} \prod_{i_A,i_B} P(B=i_B|A=i_A)^{-\frac{1}{2}} \times \prod_{i_C,i_D} P(D=i_D|C=i_C)^{-\frac{1}{2}}$$

The second model is



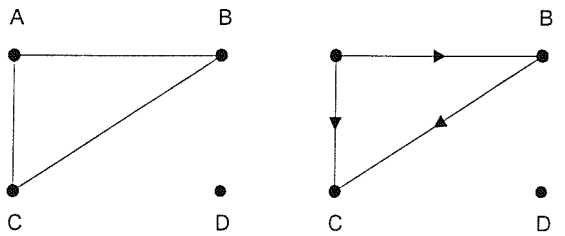
The Jeffreys' prior is again obtained using the block independence theory. Note here that because of the similarity to the three variable model detailed previously (model $AB + BC$), it is important to choose the parameterisation carefully in order to admit a straightforward Jeffreys' prior. This prior is given by

$$f(\boldsymbol{p}) \propto \prod_{i_B} P(B = i_B)^{\frac{|I_A|+|I_C|-3}{2}} \prod_{i_A,i_B} P(A = i_A|B = i_B)^{-\frac{1}{2}} \times \prod_{i_B,i_C} P(C = i_C|B = i_B)^{-\frac{1}{2}} \prod_{i_D} P(D = i_D)^{-\frac{1}{2}}$$

Three Edge Models

Three distinct models exist in this section.

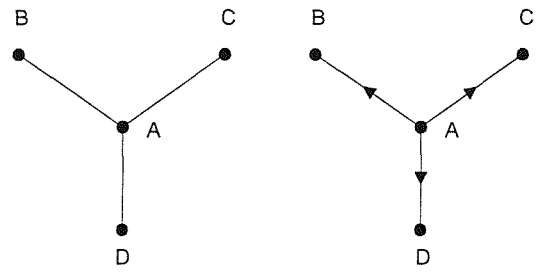
The first model has a graphical representation



The Jeffreys' prior for this model is

$$f(\boldsymbol{p}) \propto \prod_{i_A,i_B,i_C} P(B = i_B|A = i_A, C = i_C)^{-\frac{1}{2}} \prod_{i_A} P(A = i_A)^{\frac{|I_B||I_C|}{2}-1} \times \prod_{i_A,i_C} P(C = i_C|A = i_A)^{\frac{|I_B|}{2}-1} \prod_{i_D} P(D = i_D)^{-\frac{1}{2}}$$

The second model with three edges is the 'star-shaped' model which may be represented by the graphs

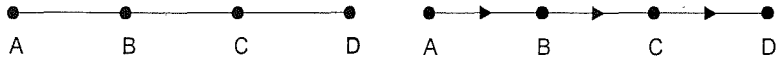


This model represents the conditional independence of B, C and D given A. The parameterisation admitted by this graph is $P(A)$, $P(C|A)$, $P(B|A)$ and $P(D|A)$. Application of expression (6.4) gives

$$\begin{aligned} f(\boldsymbol{p}) \propto & \prod_{i_A} P(A=i_A)^{\frac{|I_B|-1}{2}+\frac{|I_C|-1}{2}+\frac{|I_D|-1}{2}} \prod_{i_A,i_B,i_C,i_D} [P(A=i_A)P(C=i_C|A=i_A) \times \\ & P(B=i_B|A=i_A)P(D=i_D|A=i_A)]^{-\frac{1}{2}} \\ \propto & \prod_{i_A} P(A=i_A)^{\frac{|I_B|+|I_C|+|I_D|}{2}-2} \prod_{i_A,i_C} P(C=i_C|A=i_A)^{-\frac{1}{2}} \times \\ & \prod_{i_A,i_B} P(B=i_B|A=i_A)^{-\frac{1}{2}} \prod_{i_A,i_D} P(D=i_D|A=i_A)^{-\frac{1}{2}} \end{aligned}$$

Note that, unlike the other four variable models considered thus far, this model cannot be separated into lower order models. The prior here is again Dirichlet on the marginal and conditional probabilities, though not hyper Dirichlet.

The third model with three edges may be described as the 'straight line model', and has graphical representation



The parameterisation for this model implied by the directed graph is $P(A)$,

$P(B|A)$, $P(C|B)$, $P(D|C)$, giving the table

γ	$pa(\gamma)$	$\overline{pa}(\gamma)$
A	$\emptyset$	A, B, C, D
B	A	B, C, D
C	B	A, C, D
D	C	A, B, D

The terms $\delta_{\gamma, i_{pa(\gamma)}}$ are equal to

$$\delta_A = 1$$

$$\delta_{B,i_A} = P(A = i_A)$$

$$\delta_{C,i_B} = \sum_{i_A} P(A = i_A) P(B = i_B | A = i_A)$$

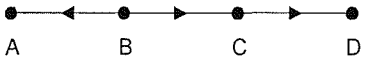
$$\delta_{D,i_C} = \sum_{i_B} \left(P(C = i_C | B = i_B) \sum_{i_A} P(A = i_A) P(B = i_B | A = i_A) \right)$$

Here, two of the δ terms are given as sums of the probabilities which comprise the parameters, hence the Jeffreys' prior will not be Dirichlet. The Jeffreys' prior is given by the expression

$$f(\boldsymbol{p}) \propto \prod_{i_C} \left[\sum_{i_B} \left(P(C = i_C | B = i_B) \sum_{i_A} P(A = i_A) P(B = i_B | A = i_A) \right) \right]^{\frac{|I_D|-1}{2}} \times$$
$$\prod_{i_B} \left[\sum_{i_A} P(A = i_A) P(B = i_B | A = i_A) \right]^{\frac{|I_C|-1}{2}} \prod_{i_A} P(A = i_A)^{\frac{|I_B|-1}{2}} \times$$
$$\prod_{i_A, i_B, i_C, i_D} [P(A = i_A) P(B = i_B | A = i_A) \times$$
$$P(C = i_C | B = i_B) P(D = i_D | C = i_C)]^{-\frac{1}{2}}$$

$$\propto \prod_{i_C}\left[\sum_{i_B}\left(P(C=i_C|B=i_B)\sum_{i_A}P(A=i_A)P(B=i_B|A=i_A)\right)\right]^{\frac{|I_D|-1}{2}} \times$$
$$\prod_{i_B}\left[\sum_{i_A}P(A=i_A)P(B=i_B|A=i_A)\right]^{\frac{|I_C|-1}{2}} \prod_{i_A}P(A=i_A)^{\frac{|I_B|}{2}-1} \times$$
$$\prod_{i_A,i_B}P(B=i_B|A=i_A)^{-\frac{1}{2}} \prod_{i_B,i_C}P(C=i_C|B=i_B)^{-\frac{1}{2}} \times$$
$$\prod_{i_C,i_D}P(D=i_D|C=i_C)^{-\frac{1}{2}}$$

The only alternative distinct parameterisation for this model is $P(B)$, $P(A|B)$, $P(C|B)$, $P(D|C)$, as represented by



This gives the table

γ	$pa(\gamma)$	$\overline{pa}(\gamma)$
B	$\emptyset$	A,B,C,D
A	B	A,C,D
C	B	A,C,D
D	C	A,B,D

The terms $\delta_{\gamma,i_{pa(\gamma)}}$ are equal to

$$\delta_B = 1$$

$$\delta_{A,i_b} = P(B=i_B)$$

$$\delta_{C,i_b} = P(B=i_B)$$

$$\delta_{D,i_C} = \sum_{i_B}\left(P(C=i_C|B=i_B)\sum_{i_A}P(B=i_B)P(A=i_A|B=i_B)\right)$$

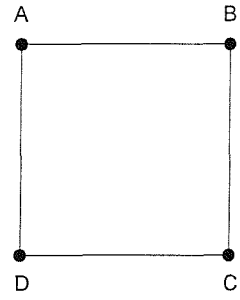
The resulting Jeffreys' prior is

$$f(\boldsymbol{p}) \propto \prod_{i_C}\left[\sum_{i_B}\left(P(C=i_C|B=i_B)\sum_{i_A}P(B=i_B)P(A=i_A|B=i_B)\right)\right]^{\frac{|I_D|-1}{2}} \times$$
$$\prod_{i_B}P(B=i_B)^{\frac{|I_A|+|I_C|-3}{2}}\prod_{i_A,i_B}P(A=i_A|B=i_B)\times$$
$$\prod_{i_B,i_C}P(C=i_C|B=i_B)\prod_{i_C,i_D}P(D=i_D|C=i_C)^{-\frac{1}{2}}$$

which again is not a Dirichlet prior. Hence, here, it is not possible to choose a parameterisation such that the Jeffreys' prior may be expressed as a product, without summation terms.

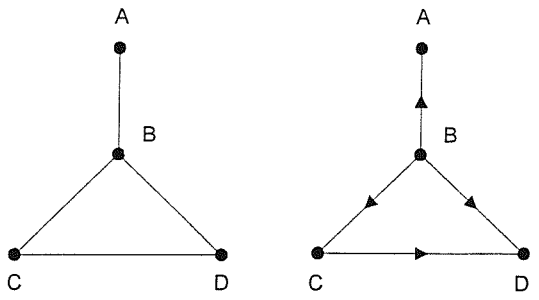
Four Edge Models

Two distinct four variable models exist which have four edges. However, one of these is the model which is represented by the graph



This model is clearly not decomposable (it is not triangulated), and so expression (6.4) cannot be used to determine the Jeffreys' prior in this case.

The other four edge model can be represented by the graphs



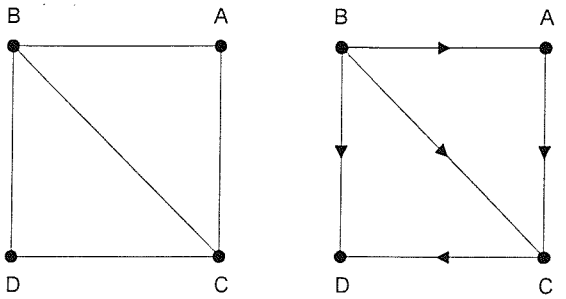
This model may be parameterised by $P(B)$, $P(A|B)$, $P(C|B)$ and $P(D|B,C)$, and represents the conditional independence of A and $\{C,D\}$ given B. Calculation of the $\delta_{\gamma,i_{pa(\gamma)}}$ terms and application of expression (6.4) gives

$$\begin{aligned} f(\boldsymbol{p}) &= \prod_{i_A,i_B,i_C,i_D} [P(A=i_A|B=i_B)P(C=i_C|B=i_B) \times \\ &\qquad P(D=i_D|B=i_B,C=i_C)]^{-\frac{1}{2}} \prod_{i_B} P(B=i_B)^{\frac{|I_A|+|I_C|-3}{2}} \times \\ &\qquad \prod_{i_B,i_C} (P(C=i_C|B=i_B)P(B=i_B))^{\frac{|I_D|-1}{2}} \\ &= \prod_{i_B,i_C} P(C=i_C|B=i_B)^{\frac{|I_D|}{2}-1} \prod_{i_A,i_B} P(A=i_A|B=i_B)^{-\frac{1}{2}} \times \\ &\qquad \prod_{i_B,i_C,i_D} P(D=i_D|B=i_B,C=i_C)^{-\frac{1}{2}} \prod_{i_B} P(B=i_B)^{\frac{|I_A|+|I_C||I_D|-3}{2}} \end{aligned}$$

This prior is again Dirichlet, but not hyper Dirichlet.

Five Edge Model

The single model in this category may be represented by the graphs



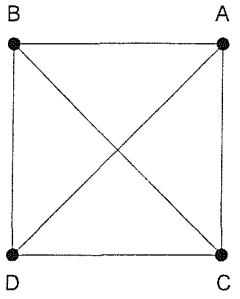
The parameterisation which is used to express this model is $P(B)$, $P(C|B)$, $P(A|B,C)$ and $P(D|B,C)$. Calculation of the $\delta_{\gamma,i_{pa(\gamma)}}$ terms enables us to derive the Jeffreys' prior for this example, which is

$$\begin{aligned} f(\boldsymbol{p}) &= \prod_{i_B} P(B=i_B)^{\frac{|I_C|-1}{2}} \prod_{i_B,i_C} (P(C=i_C|B=i_B)P(B=i_B))^{\frac{|I_A|-1}{2}+\frac{|I_D|-1}{2}} \times \\ &\quad \prod_{i_A,i_B,i_C} (P(B=i_B)P(C=i_C|B=i_B)P(A=i_A|B=i_B,C=i_C))^{-\frac{1}{2}} \times \\ &\quad \prod_{i_B,i_C,i_D} P(D=i_D|B=i_B,C=i_C)^{-\frac{1}{2}} \\ &= \prod_{i_B} P(B=i_B)^{\frac{|I_C||I_A|+|I_C||I_D|-|I_C|}{2}-1} \prod_{i_B,i_C} P(C=i_C|B=i_B)^{\frac{|I_A|+|I_D|-3}{2}} \times \\ &\quad \prod_{i_A,i_B,i_C} P(A=i_A|B=i_B,C=i_C)^{-\frac{1}{2}} \times \\ &\quad \prod_{i_B,i_C,i_D} P(D=i_D|B=i_B,C=i_C)^{-\frac{1}{2}} \end{aligned}$$

This prior is a Dirichlet prior on the conditional probabilities, but note that the choice of parameterisation here is crucial to ensure this. For example, the alternative parameterisation $P(A)P(B|A)P(C|A,B)P(D|B,C)$ will involve summation terms in the Jeffreys' prior.

Saturated Model

The saturated model with four variables has the graphical representation



The decomposition used for this model is $P(A)P(B|A)P(C|A,B)P(D|A,B,C)$,

and the Jeffreys' prior is

$$f(\boldsymbol{p}) = \prod_{i_A,i_B,i_C} P(C=i_C|A=i_A,B=i_B)^{\frac{|I_D|}{2}-1} \prod_{i_A,i_B} P(B=i_B|A=i_A)^{\frac{|I_C||I_D|}{2}-1} \times \prod_{i_A} P(A=i_A)^{\frac{|I_B||I_C||I_D|}{2}-1} \prod_{i_A,i_B,i_C,i_D} P(D=i_D|A=i_A,B=i_B,C=i_C)^{-\frac{1}{2}}$$

6.4.6 Discussion

The Jeffreys' priors for all decomposable models with up to and including four variables have been derived above. Most of these distributions are products of independent Dirichlet distributions, and hence the normalising constant (which will be required for the use of the prior in applications such as model selection) is readily available. Such normalising constants may be calculated by repeated application of the equation

$$\int \prod P(\gamma=i_\gamma|pa(\gamma)=i_{pa(\gamma)})^{k-1} dP(\gamma=i_\gamma|pa(\gamma)=i_{pa(\gamma)}) = \frac{\Gamma(|I_\gamma|\,k)}{\Gamma(|I_\gamma|)^k}$$

where $|I_\gamma|$ is the number of levels of γ.

As an example, consider the Jeffreys' prior for the model with three variables A, B, C where A and C are conditionally independent given B. This prior was determined earlier to be of the form

$$f(\boldsymbol{p}) \propto \prod_{i_b} P(B=i_b)^{\frac{|I_A|+|I_C|-3}{2}} \prod_{i_a,i_b} P(A=i_a|B=i_b)^{-\frac{1}{2}} \prod_{i_b,i_c} P(C=i_c|B=i_b)^{-\frac{1}{2}}$$

The normalising constants are obtained using the previous equation, and we obtain

$$f(\boldsymbol{p}) = \frac{\Gamma(|I_B|\,\frac{|I_A|+|I_C|-1}{2})}{\Gamma(\frac{|I_A|+|I_C|-1}{2})^{|I_B|}} \prod_{i_b} P(B=i_b)^{\frac{|I_A|+|I_C|-3}{2}} \times \left(\frac{\Gamma(\frac{|I_A|}{2})}{\Gamma(\frac{1}{2})^{|I_A|}}\right)^{|I_B|} \prod_{i_a,i_b} P(A=i_a|B=i_b)^{-\frac{1}{2}} \times$$

$$\left(\frac{\Gamma(\frac{|I_C|}{2})}{\Gamma(\frac{1}{2})^{|I_C|}}\right)^{|I_B|} \prod_{i_b,i_c} P(C=i_c|B=i_b)^{-\frac{1}{2}}$$

$$= \frac{\Gamma(|I_B|\frac{|I_A|+|I_C|-1}{2})\left[\Gamma(\frac{|I_A|}{2})\Gamma(\frac{|I_C|}{2})\right]^{[I_B]}}{\Gamma(\frac{|I_A|+|I_C|-1}{2})^{|I_B|}\Gamma(\frac{1}{2})^{|I_A||I_B|}\Gamma(\frac{1}{2})^{|I_C||I_B|}} \prod_{i_b} P(B=i_b)^{\frac{|I_A|+|I_C|-3}{2}} \times$$

$$\prod_{i_a,i_b} P(A=i_a|B=i_b)^{-\frac{1}{2}} \prod_{i_b,i_c} P(C=i_c|B=i_b)^{-\frac{1}{2}}$$

It was shown in section 6.4.5 that there is one model which does not admit a product Dirichlet Jeffreys' prior – this is the four variable 'straight-line' model. The reason for this can be observed by considering the form of $\delta_{\gamma,i_{pa(\gamma)}}$ for each γ. It was shown in section 6.3.1 that $\delta_{\gamma,i_{pa(\gamma)}}$ may be written

$$\delta_{\gamma,i_{pa(\gamma)}} = \sum_{i_{\overline{pa}(\gamma)}} \prod_{\gamma'} P(\gamma'=i_{\gamma'}|pa(\gamma')=i_{pa(\gamma')})$$

The Jeffreys' prior for a particular model will not be expressible as a product of independent Dirichlet distributions unless, under a particular ordering of the variables, for all γ this term may be written as a conditional probability which is part of the parameterisation.

Consider a model where a particular γ has a grandparent who is not themselves a parent of γ, and suppose γ is the 'lowest' variable in the ordering with this property. The sum in $\delta_{\gamma,i_{pa(\gamma)}}$ is taken over $i_{\overline{pa}(\gamma)}$, and we may always sum recursively over each $i_{\gamma'}$ for γ' below γ in the ordering, starting from the lowest γ'. These sums are possible as each γ' depends only on $pa(\gamma')$, and so we sum to 1 each time.

Now consider a γ' and γ'' below γ, for which $\gamma' \in pa(\gamma)$ and $\gamma'' \in pa(\gamma')$, $\gamma'' \notin pa(\gamma)$. We do not sum over $i_{\gamma'}$ as $\gamma' \in pa(\gamma)$. However, summing over $i_{\gamma''}$ will not result in terms in the parameterisation, as γ' depends on γ'' through $P(\gamma'|pa(\gamma'))$. Hence $\delta_{\gamma,i_{pa(\gamma)}}$ will not be expressible as a conditional probability which is part of the parameterisation, and must be left as a sum of model parameters.

For the four variable 'straight-line' model, parameterised by $P(A)$, $P(B|A)$, $P(C|B)$ and $P(D|C)$, the variables which cause such problems are C and D, as C has a grandparent who is not themselves a parent of C (the same is true for D). Hence δ_C and δ_D cannot be expressed as a product of model parameters.

In general, the Jeffreys' prior for any decomposable model may be obtained from expression (6.4), and this distribution will be a product of independent Dirichlet distributions provided that the model may be parameterised such that no variable has a grandparent who is not themselves a parent. Graphically, a sufficient condition is that for each disconnected component of the graph, all cliques have a common intersection. For those distributions which are products of Dirichlets, it is then possible to write down the normalised form of this prior.

The problem of calculating the normalising constant for models which do satisfy this condition will be addressed in the next section.

6.5 Calculation of Normalising Constants

6.5.1 Bridge Sampling

A bridge sampler has already been used in Chapter 5 to determine the normalising constant for any given conditional Dirichlet model. However, this particular implementation of the bridge sampling algorithm cannot be used here as the Jeffreys' priors given by expression (6.4) are not conditional Dirichlet distributions. Because of this, it is necessary to derive a new method specifically for Jeffreys' priors.

The method of bridge sampling was described in detail in section 5.4. As part of the bridge sampling method, it is first necessary to generate a sample from the target distribution. In the previous bridge sampler, a Gibbs sampler was used together with adaptive rejection sampling. However, this relies upon the log-concavity of the distribution, and since the priors based on expression

(6.4) do not seem to be log-concave it is necessary to use an alternative method in this instance.

The method employed is the Metropolis Hastings algorithm, applied iteratively to blocks of parameters as a Metropolis within Gibbs sampler.

Metropolis Hastings Method

This is a Markov chain Monte Carlo (MCMC) method (see section 2.4.4), first introduced by Metropolis et al (1953) and generalised by Hastings (1970), whereby a sequence of samples is generated from the target density by simulating a Markov chain. The method may be summarised as follows:

1. Let $\boldsymbol{\theta}^{(t)}$ be the current sample from target density $f(\boldsymbol{\theta})$.
2. Generate a candidate $\boldsymbol{\theta}^*$ from proposal density $g(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t)})$.
3. Evaluate the ratio $r = \frac{f(\boldsymbol{\theta}^*)g(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^*)}{g(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t)})f(\boldsymbol{\theta}^{(t)})}$. Accept candidate $\boldsymbol{\theta}^*$ and set $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^*$ with probability $\min\{1, r\}$. Otherwise set $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)}$.

Note that the normalising constant for the target density f is not required in the algorithm, as it is cancelled out in the calculation of ratio r.

A special case of this algorithm is applied here – this is known as an Independence Sampler. In this, the density $g(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t)})$ is independent of $\boldsymbol{\theta}^{(t)}$, so that $g(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t)}) = g(\boldsymbol{\theta}^*)$. Step 3 therefore becomes

3. Evaluate the ratio $r = \frac{f(\boldsymbol{\theta}^*)g(\boldsymbol{\theta}^{(t)})}{g(\boldsymbol{\theta}^*)f(\boldsymbol{\theta}^{(t)})}$. Accept candidate $\boldsymbol{\theta}^*$ and set $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^*$ with probability $\min\{1, r\}$. Otherwise set $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)}$.

Although in theory any proposal density g which has the same support as f will work, the key to the Independence Sampler is choosing a proposal density g which is both easy to sample from, and is close to the target density f. If a poor choice of g is made, then ratio r will often be very small and few updates will be accepted.

The aim here is to generate samples from Jeffreys' priors given by the expression

$$f(\boldsymbol{p}) \propto \left(\prod_{\gamma} \prod_{i_{pa(\gamma)}} \left[\sum_{i_{\overline{pa}(\gamma)}} \prod_{\gamma'} P(\gamma' = i_{\gamma'} | pa(\gamma') = i_{pa(\gamma')}) \right]^{\frac{|I_\gamma|-1}{2}} \right) \times \prod_{\gamma} \prod_{i_\gamma, i_{pa(\gamma)}} P(\gamma = i_\gamma | pa(\gamma) = i_{pa(\gamma)})^{-\frac{1}{2}} \quad (6.7)$$

In many instances the expression on the top line of (6.7) may itself be expressed as a product and we obtain a product of independent Dirichlet distributions. However, the cases of interest for this section involve those priors where summation terms appear, and the top line of (6.7) may not be expressed as a product. Because of the form of this distribution, the choice made for proposal density g is

$$g(\boldsymbol{p}) \propto \prod_{\gamma} \prod_{i_\gamma, i_{pa(\gamma)}} P(\gamma = i_\gamma | pa(\gamma) = i_{pa(\gamma)})^{-\frac{1}{2}}$$

For any model, this will be a product of Dirichlet distributions, each having all parameters equal to $\frac{1}{2}$.

A C program was developed to implement the Metropolis Hastings algorithm to the problem of generating a sample from the Jeffreys' prior for a particular decomposable log-linear model. This program needs the following inputs:

1. The (un-normalised) density f.
2. The total MCMC sample size required.
3. The number of independent Dirichlet distributions into which the proposal density factorises, and the dimensions of each of these densities.
4. The parameters for each proposal density. Although for prior samples all parameters are equal to $\frac{1}{2}$, for posterior sampling more general proposals will be required.

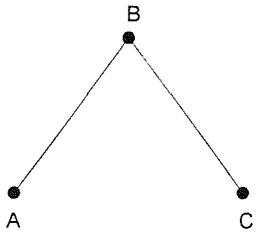
An observation is then generated in turn from each of the Dirichlet distributions, and the Metropolis Hastings algorithm applied at each stage to determine whether this observation is accepted or rejected. This whole process is iterated to provide the required sample size.

Bridge Sampler

The method of bridge sampling was described in detail in section 5.4. Slight modifications were made to the S Plus code used there in order to apply the method of bridge sampling for Jeffreys' priors derived from equation (6.4), using samples produced using the Metropolis Hastings algorithm. Note that the starting value for the bridge sampler is no longer obtained from a Laplace approximation, but instead an arbitrary value of 1 is used and the bridge sampler is then run iteratively until convergence (typically less than 10 runs).

6.5.2 Results

We first check the method by applying it in cases where the normalising constants are known.



The model represented by the (undirected) graph above admits two distinct parameterisations, and it was shown in section 6.4.4 that, whereas one of these results in a Jeffreys' prior which is a product of independent Dirichlet distributions, the other is not of this form and so is a suitable candidate for the bridge sampler. Table 6.1 summarises several results obtained for this model, comparing the values obtained using the bridge sampler with those true values obtained

using the Dirichlet parameterisation. Sample sizes of 50000 were used.

Levels of Variable A	Levels of Variable B	Levels of Variable C	log Bridge Estimate	log True Value	Error in Bridge Estimate
2	2	2	3.64	3.64	0
2	2	3	4.12	4.17	.05
2	2	4	4.20	4.26	.06
2	3	4	3.41	3.62	.21
3	2	4	4.78	4.85	.07
3	3	4	2.95	3.16	.21

Table 6.1: Estimated Jeffreys' normalising constants, together with their respective errors

As can be seen from the table, the approximations to the normalising constants given by the bridge sampler are very good. A time series plot for the Monte Carlo sample corresponding to margin $P(B)$ is given in figure 6.1.

Similar plots are obtained for the other model parameters, and we conclude that the Metropolis Hastings sampler mixes well, producing samples which are not highly dependent. The autocorrelations drop to negligible values after lag 4.

Table 6.2 shows rejection percentages for both prior sampling (where $n = 0$), and for posterior sampling with equal cell counts. As expected, the percentages drop markedly with increasing sample sizes, since the data only updates the second part of the Jeffreys' distribution, from which we generate the proposal density. The rejection percentages all fall within acceptable limits, which further validates the quality of the sampler, in particular for posterior sampling. The sampler was applied to several other models, and similar results were obtained. Therefore, the bridge sampler may be used with confidence to determine the normalising constant for the Jeffreys' prior (as given by expression (6.4)) for any decomposable model, and so whether or not the prior may be expressed as a product of independent Dirichlet distributions is no longer a consideration for the use of Jeffreys' priors.

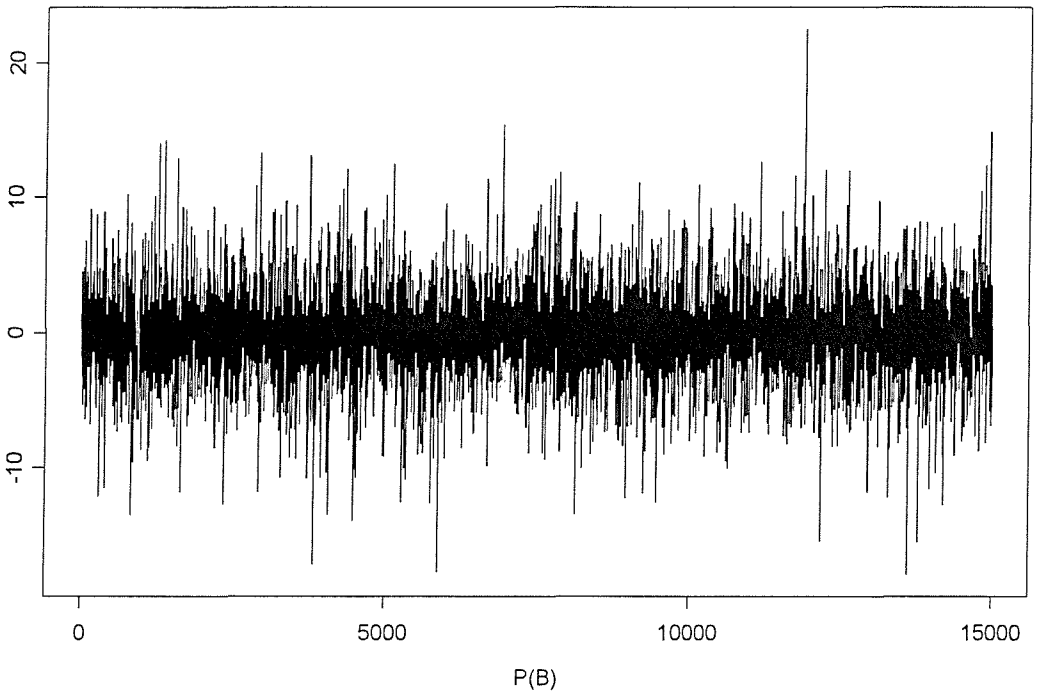


Figure 6.1: Time series plot for Metropolis Hastings sample corresponding to $P(B)$

6.6 Conclusion

The focus of this Chapter has been Jeffreys' prior for log-linear models. A general expression (6.4) has been derived which enables us to write down the Jeffreys' prior for any decomposable log-linear model, and this expression has been applied for all models with up to and including four variables, and the resulting distributions presented. It is often possible to parameterise the model so that the distribution obtained is a product of independent Dirichlet distributions, in which case the normalising constant is straightforward to determine. However, in the instances where this is not possible, a bridge sampler has been developed to give a good approximation to the normalising constant.

Sample Size n	Rejection Percentage
0	32.3%
8	19.1%
40	6.9%
80	4.1%
160	2.2%
400	0.8%

Table 6.2: Rejection percentages for Metropolis Hastings Sampler

The problem of determining the Jeffreys' prior for non-decomposable models was considered, although the expression obtained does not, in general, result in a tractable form for this distribution. However, the expression was applied to determine the Jeffreys' prior for a saturated model.

Chapter 7

Choosing A Prior Distribution

The main focus of this thesis is on prior distributions suitable for use in reference analyses of log-linear models. Several potential reference priors have been introduced, though the choice of the parameters for these distributions has not been considered. In this Chapter, we consider the problem of model selection, specifically methods of approximating the Bayes factor, and the effect of the choice of prior distributions and prior parameters on such approximations.

7.1 Laplace's Method and the Schwarz Criterion

Laplace's method of approximating integrals was introduced in section 5.3, and the Laplace approximation to the (log) marginal likelihood was shown to be

$$\log \int f(\boldsymbol{n}|\boldsymbol{\theta}) f(\boldsymbol{\theta}) d\boldsymbol{\theta} = \log f(\boldsymbol{n}|\widetilde{\boldsymbol{\theta}}) + \log f(\widetilde{\boldsymbol{\theta}}) - \frac{1}{2} \log \left| -H(\widetilde{\boldsymbol{\theta}}) \right| + \frac{d}{2} \log 2\pi + O(n^{-1})$$

where $\widetilde{\boldsymbol{\theta}}$ is the posterior mode, $f(\boldsymbol{n}|\boldsymbol{\theta})$ is the likelihood, $f(\boldsymbol{\theta})$ is the prior, d is the dimension of $\boldsymbol{\theta}$, and $H(\boldsymbol{\theta})$ is the Hessian matrix of second derivatives (*i.e.* $H(\boldsymbol{\theta}) = \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} f(\boldsymbol{n}|\boldsymbol{\theta}) f(\boldsymbol{\theta})$). This approximation is correct to order $O(n^{-1})$,

where n is the total sample size.

This expression was re-written in terms of the information matrix and maximum likelihood estimator $\widehat{\theta}$

$$\int f(\boldsymbol{n}|\boldsymbol{\theta})f(\boldsymbol{\theta})d\boldsymbol{\theta} = \log f(\boldsymbol{n}|\widehat{\boldsymbol{\theta}}) + \log f(\widehat{\boldsymbol{\theta}}) + \frac{d}{2}\log 2\pi - \frac{d}{2}\log n - \frac{1}{2}\log\left|i(\widehat{\boldsymbol{\theta}})\right| + O(n^{-1/2}) \tag{7.1}$$

to give an approximation to the marginal likelihood which is correct to an order $O(n^{-\frac{1}{2}})$. Similarly, the Bayes factor comparing models m_1 and m_2 may be approximated using the expression

$$\begin{aligned}\log B_{12} \;=\; & \log f(\boldsymbol{n}|\widehat{\boldsymbol{\theta}}_{m_1}) + \log f(\widehat{\boldsymbol{\theta}}_{m_1}) + \frac{d_1 - d_2}{2}\log 2\pi - \frac{d_1 - d_2}{2}\log n - \\ & \frac{1}{2}\log\left|i(\widehat{\boldsymbol{\theta}}_{m_1})\right| - \log f(\boldsymbol{n}|\widehat{\boldsymbol{\theta}}_{m_2}) - \log f(\widehat{\boldsymbol{\theta}}_{m_2}) + \frac{1}{2}\log\left|i(\widehat{\boldsymbol{\theta}}_{m_2})\right| + O(n^{-1/2})\end{aligned}$$

However, this approximation requires a normalised form of the prior density. As has been discussed previously, it may be difficult or impossible to obtain the normalising constant for prior distributions analytically, as such distributions are often intractable. Hence it is sometimes convenient to omit the term $\log f(\boldsymbol{\theta})$, to give an expression which is only correct to an error of order $O(1)$.
As $\frac{1}{2}(d_1 - d_2)\log 2\pi$ and $\log\left|i(\widehat{\boldsymbol{\theta}}_{m_i})\right|$ are also $O(1)$, they can be absorbed into the $O(1)$ term to give the Schwarz criterion

$$S_{12} = \log f(\boldsymbol{n}|m_1, \widehat{\boldsymbol{\theta}}_{m_1}) - \log f(\boldsymbol{n}|m_2, \widehat{\boldsymbol{\theta}}_{m_2}) - \frac{1}{2}(d_1 - d_2)\log n$$

Although the Schwarz criterion is only generally correct to order $O(1)$, an interesting consideration is whether it is possible to choose a particular prior distribution such that the order of such an approximation may be improved. The focus of the remainder of this Chapter is whether it is possible to choose $f(\boldsymbol{\theta})$ such that the terms $\frac{1}{2}(d_1 - d_2)\log 2\pi$ and $\log\left|i(\widehat{\boldsymbol{\theta}}_{m_i})\right|$ disappear (or may be easily evaluated). As noted by Kass and Wasserman (1995), Jeffreys (1961) chose the

Cauchy prior density for use in Normal location testing problems, in which case the terms above were replaced by a constant, dependent only on the dimensions of the models under the null and alternative hypotheses. This corrected form of the Schwarz criterion allowed the Bayes factor to be approximated to an error of order $O(n^{-\frac{1}{2}})$.

7.2 Unit Information

7.2.1 Unit Information Normal Priors

This is an idea introduced by Kass and Wasserman (1995), based on work by Kass and Vaidyanathan (1992), which they applied in particular to the problem of Bayesian hypothesis testing. It allows a choice to be made for prior $f(\boldsymbol{\theta})$ so as to allow the log marginal likelihood to be approximated to an error of $O(n^{-\frac{1}{2}})$ using the Schwarz approximation, and hence requiring only the maximised likelihood. Their theory may be summarised as follows:

Suppose we have a set of i.i.d. observations $Y_1, \ldots, Y_n$ from a family parameterised by $(\boldsymbol{\beta}, \boldsymbol{\psi})$, and that we wish to test the hypothesis $H_0 : \boldsymbol{\psi} = \boldsymbol{\psi}_0$ against the general alternative H_1. Kass and Wasserman proposed that the amount of information in the prior under H_1 should be chosen to be equal to the amount of information in a single observation, an idea which is intuitively appealing. Here information is in the sense of Fisher information. The concept is best illustrated by a simple one-dimensional example.

Let $Y_i \sim N(\psi, \sigma^2)$, with σ known, and suppose we want to test $H_0 : \psi = \psi_0$ against $H_1 : \psi \in \Re$. Then $i(\psi) = \sigma^2$ and the prior distribution on ψ under H_1,

$$\psi \sim N(\psi_0, \tau^2)$$

with $\tau = \sigma$ has the same amount of information about ψ as there is in a single

observation, since the variance of a single observation is also σ^2. This is then called the unit information prior for ψ.

This theory may be extended to the multivariate case, though this requires two important simplifying assumptions. We assume that $\boldsymbol{\beta}$ and $\boldsymbol{\psi}$ are null orthogonal, and hence that the Fisher information matrix (which we shall write $I(\boldsymbol{\beta}, \boldsymbol{\psi})$) is block diagonal for null hypothetical parameter values – *i.e.* $I_{\beta\psi}(\boldsymbol{\beta}, \boldsymbol{\psi}_0) = 0$ for all $\boldsymbol{\beta}$. Indeed, Kass and Vaidynathan argued that the parameters may always be transformed so that $\boldsymbol{\beta}$ is null orthogonal to $\boldsymbol{\psi}$. We also assume that the marginal prior distribution of $\boldsymbol{\beta}$ is identical under both hypotheses.

Suppose the prior on $\boldsymbol{\psi}$ under H_1 is elliptically symmetric with location parameter $\boldsymbol{\psi}_0$ and scale matrix $\boldsymbol{\Sigma}_\psi$, and has density of the form

$$\pi_\psi(\boldsymbol{\psi}) = |\Sigma_\psi|^{-\frac{1}{2}} f((\boldsymbol{\psi} - \boldsymbol{\psi}_0)^T \Sigma_\psi^{-1} (\boldsymbol{\psi} - \boldsymbol{\psi}_0))$$

Then a unit information prior for $\boldsymbol{\psi}$ may be defined by choosing $\boldsymbol{\Sigma}_\psi$ to satisfy the expression

$$|\Sigma_\psi|^{-1} = |I_{\psi\psi}(\boldsymbol{\beta}, \boldsymbol{\psi}_0)| \tag{7.2}$$

so that the amount of information in the prior is equal to the amount contained in one observation. (Note that $I_{\psi\psi}(\boldsymbol{\beta}, \boldsymbol{\psi}_0)$ is the block of $I(\boldsymbol{\beta}, \boldsymbol{\psi})$ corresponding to $\boldsymbol{\psi}$).

Kass and Wasserman (1995) showed that, for those prior distributions satisfying expression (7.2), $\exp(S_{12}) \rightarrow B_{12}$ as $n \rightarrow \infty$, with an error of order $O(n^{-1/2})$, and that for samples of only moderate size the Bayes factor may be reasonably approximated in this way. This is a particularly useful result, as it allows the evidence in favour of a model to be readily calculated without the need for complex integration or other, more time-consuming, approximation methods.

Kass and Wasserman applied the results to several simple examples, and showed that the resulting approximation furnished by the Schwarz approximation

to the log marginal likelihood was good, even for small sample sizes.

7.2.2 Unit Information and the Logistic Normal Distribution

The concept of unit information introduced by Kass and Wasserman was applied by Dellaportas and Forster (1999) to the logistic Normal distribution, and this work is described below.

The log-normal distribution was introduced in section 3.1, and when expressed as a distribution for the cell means $\boldsymbol{\mu}$ was shown to have the form

$$\log \boldsymbol{\mu} \sim N(\delta \mathbf{1}, \sum_{a \subseteq C} \alpha_a^2 T_a) \tag{7.3}$$

where T_a are projection matrices defined in (3.1).

Choices must be made for each dispersion parameter α_a^2, and for the prior mean of $\boldsymbol{\beta}_\emptyset$, δ. The problem considered by Dellaportas and Forster was to choose the parameters so that the prior distribution can be interpreted as vague without being excessively diffuse. These α_a^2 parameters may be interpreted as representing the prior knowledge about β_a, with large values representing vague prior knowledge. A vague, but proper, distribution may be obtained by using large but finite values for α_a^2.

Dellaportas and Forster chose values for the α_a^2 parameters by considering hypothetical 'prior samples'. They showed that Jeffreys prior is equivalent to setting $\alpha_a^2 = \pi^2/2$, and Perks' (1947) prior is equivalent to choosing $\alpha_a^2 = \psi'\left(\frac{1}{|I|}\right)$ (where ψ' is the trigamma function).

Under a multinomial sampling scheme, a distribution is required for $\log \boldsymbol{p}$ instead of $\log \boldsymbol{\mu}$. The equivalent distribution to that derived above is logistic normal, and is best expressed as a prior for $\boldsymbol{\theta}$, where $\boldsymbol{\theta}$ is the symmetric logit

defined in section 1.3. This gives

$$\theta \sim N\left(0, \sum_{a \subseteq \Gamma; a \neq \emptyset} \alpha_a^2 T_a\right)$$

Dellaportas and Forster showed that the value of $\alpha_\emptyset$ has little effect on posterior analysis, so chose to make this arbitrarily small, also setting $\delta = 0$ in (7.3). They suggested that a reasonable choice for α_a^2 was to set $\alpha_a^2 = k\,|I|$, showing that $\frac{1}{k}$ is interpretable as the number of units of prior information at the prior mean, and that the prior information away from this is less than $\frac{1}{k}$. Hence, for a unit information prior, k must be at least one. Consideration of several examples lead to a choice $k = 2$, hence setting $\alpha_a^2 = 2\,|I|$.

7.2.3 Unit Information and the Schwarz Criterion

Kass and Wasserman introduced the concept of unit information as a way of choosing a prior distribution such that the marginal likelihood may be reasonably approximated by the Schwarz approximation.

In section 7.1, it was shown that the difference between the Laplace and Schwarz approximations is given by the expression

$$\delta = \log f(\widehat{\theta}) + \frac{d}{2}\log 2\pi - \frac{1}{2}\log\left|i(\widehat{\theta})\right|$$

Suppose we have a Normal prior distribution, with mean θ_0 and variance Σ, with density function

$$\log f(\theta) = -\frac{d}{2}\log 2\pi - \frac{1}{2}\log|\Sigma| - (\theta - \theta_0)^T\Sigma^{-1}(\theta - \theta_0)$$

So, if we choose $\Sigma^{-1} = i(\theta_0)$, then for $|\theta - \theta_0| = O(n^{-\frac{1}{2}}i(\theta_0))$ we obtain

$$\begin{aligned}
\delta &= \log f(\boldsymbol{\theta}_0) + \frac{d}{2}\log 2\pi - \frac{1}{2}\log|i(\boldsymbol{\theta}_0)| \\
&= -\frac{d}{2}\log 2\pi + \frac{1}{2}\log|i(\boldsymbol{\theta}_0)| - (\boldsymbol{\theta}-\boldsymbol{\theta}_0)^T i(\boldsymbol{\theta}_0)^{-1}(\boldsymbol{\theta}-\boldsymbol{\theta}_0) + \frac{d}{2}\log 2\pi - \frac{1}{2}\log|i(\boldsymbol{\theta}_0)| \\
&= -(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^T i(\boldsymbol{\theta}_0)^{-1}(\boldsymbol{\theta}-\boldsymbol{\theta}_0) \\
&= O(n^{-\frac{1}{2}})
\end{aligned}$$

Hence, provided the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}$, in terms of units of information, is close to the prior mean, the Schwarz approximation will furnish a good approximation to the marginal likelihood. However, the approximation will not be close in general.

7.3 Unit Information for Dirichlet Based Priors

7.3.1 Introduction

The focus of much of this thesis has been on priors based on the Dirichlet distribution. However, the concept of a unit information prior as described in section 7.2 cannot be applied directly to priors from the Dirichlet family. As introduced in Chapter 2.1, this prior has the form

$$f(\boldsymbol{p}) \propto \prod_{\boldsymbol{i}\in I} p(\boldsymbol{i})^{\alpha(\boldsymbol{i})-1}$$

for the saturated model, or in the more general conditional Dirichlet case has the form

$$f(\boldsymbol{\beta}) \propto \frac{\prod_{\boldsymbol{i}\in I} e^{\alpha(\boldsymbol{i})\sum_j x(\boldsymbol{i},j)\beta_j}}{\left(\sum_{\boldsymbol{i}} e^{\sum_j x(\boldsymbol{i},j)\beta_j}\right)^{\alpha}} \tag{7.4}$$

for a model with design matrix X which sets $\boldsymbol{\theta} = X\boldsymbol{\beta}$. The problem is whether it is possible to choose a suitable set of $\boldsymbol{\alpha}$ parameters to make this a 'unit information

prior'.

7.3.2 Conditional Dirichlet Prior

The main criterion for a unit information prior distribution is that the amount of information which it contains should be equal to the amount of information in one observation. In the case of Normal based priors, this was achieved by a suitable choice of the variance matrix. However, the conditional Dirichlet distribution does not admit such a simple solution, as the variance may not be expressed in such a tractable form.

As there is a direct correspondence between observations in a contingency table and the parameters of the conditional Dirichlet distribution, an alternative definition of unit information is apparent as the prior parameters $\boldsymbol{\alpha}$ may be interpreted as prior cell counts. Suppose a single observation is divided between all cells, *i.e.* let the 'prior cell count' $\alpha(\boldsymbol{i}) = \frac{1}{s}$ where s is the number of cells in the table (Perks' prior). The Schwarz approximation was investigated empirically for such priors to see if accurate approximations to log marginal likelihoods resulted. However, the results were disappointing. Indeed, it proved impossible to find any set of $\boldsymbol{\alpha}$ parameters such that the Schwarz criterion offered a good approximation to the log Bayes factor, even allowing for a constant correction term to the Schwarz formula.

For example, in the simple 2×2 case, using a prior which sets $\alpha(\boldsymbol{i}) = \frac{1}{4}$ (*i.e.* a single observation 'split' between the cells), the Schwarz approximation approximates the (log) marginal likelihood for a sample size of 2000 split evenly through the table with an error of -1.01. The corresponding error for a sample of 20000 is the same, so the error clearly does not tend to zero. Furthermore, we failed to determine an expression by which this error could be calculated in general.

7.3.3 Jeffreys' Prior

In section 7.1, it was shown that the terms which represent the difference between Laplace's approximation and the Schwarz approximation are

$$\log f(\widehat{\boldsymbol{\theta}}) + \frac{d}{2}\log 2\pi - \frac{1}{2}\log\left|i(\widehat{\boldsymbol{\theta}})\right| \tag{7.5}$$

Now, Jeffreys' prior was defined in Chapter 6 by

$$f(\boldsymbol{\theta}) \propto \left|E\left[-\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T} f(\boldsymbol{n}|\boldsymbol{\theta})\right]\right|^{\frac{1}{2}}$$

or, equivalently,

$$\log f(\boldsymbol{\theta}) = \frac{1}{2}\log|i(\boldsymbol{\theta})| + c$$

where c is the normalising constant. Comparison of this expression with (7.5) provided motivation to investigate whether the marginal likelihood based on Jeffreys' prior may be reasonably approximated by the Schwarz formula, possibly with the addition of a correction factor. Wasserman (1997) used Jeffreys' prior to obtain an order $O(n^{-\frac{1}{2}})$ approximation to the marginal likelihood, though he defined Jeffreys' prior using a fixed normalising constant $(2\pi)^{-\frac{d}{2}}$, applied to cases where Jeffreys' prior is improper and such an arbitrary constant may be chosen. However, the Jeffreys' priors used in this Chapter are not improper, and so such an approach is not appropriate here.

In section 6.3.1, the (unit) information matrix $i(\boldsymbol{p})$ for a decomposable log-linear model parameterised using a perfect ordering gave

$$|i(\boldsymbol{p})| \propto \left(\prod_{\gamma}\prod_{i_{pa(\gamma)}}\left[\sum_{i_{\overline{pa}(\gamma)}}\prod_{\gamma'} P(\gamma' = i_{\gamma'}|pa(\gamma') = i_{pa(\gamma')})\right]^{|I_\gamma|-1}\right) \times \prod_{\gamma}\prod_{i_\gamma, i_{pa(\gamma)}} P(\gamma = i_\gamma|pa(\gamma) = i_{pa(\gamma)})^{-1}$$

and Jeffreys' prior is then given by

$$f(\boldsymbol{p}) = \frac{1}{c}\left(\prod_{\gamma}\prod_{i_{pa(\gamma)}}\left[\sum_{i_{\overline{pa}(\gamma)}}\prod_{\gamma'}P(\gamma'=i_{\gamma'}|pa(\gamma')=i_{pa(\gamma')})\right]^{\frac{|I_\gamma|-1}{2}}\right)\times$$
$$\prod_{\gamma}\prod_{i_\gamma,i_{pa(\gamma)}}P(\gamma=i_\gamma|pa(\gamma)=i_{pa(\gamma)})^{-\frac{1}{2}}$$

where $\frac{1}{c}$ is the normalising constant for the distribution. The difference δ between the Laplace and Schwarz approximation, equation (7.5), then becomes

$$\begin{aligned}\delta &= \log f(\widehat{\boldsymbol{p}}) + \frac{d}{2}\log 2\pi - \frac{1}{2}\log|i(\widehat{\boldsymbol{p}})| \\ &= \frac{1}{2}\log|i(\widehat{\boldsymbol{p}})| - \log c + \frac{d}{2}\log 2\pi - \frac{1}{2}\log|i(\widehat{\boldsymbol{p}})| \\ &= \frac{d}{2}\log 2\pi - \log c\end{aligned}$$

Hence we have an expression for the difference between the Laplace approximation and the Schwarz approximation to the marginal likelihood for a decomposable log-linear model which depends only on the normalising constant of the Jeffreys' prior. Hence, it is possible to write down a corrected Schwarz approximation

$$S_C = \log f(\boldsymbol{n}) = \log f(\boldsymbol{n}|\widehat{\boldsymbol{p}}) - \frac{d}{2}\log n + \frac{d}{2}\log 2\pi - \log c$$

which is correct to order $O(n^{-\frac{1}{2}})$. This corrected approximation will be applied in the next section to several examples.

7.4 Application of Corrected Schwarz Approximation

An expression was derived in the previous section which may be used to provide a correction term to the Schwarz approximation, and the resulting approximation will then provide a $O(n^{-\frac{1}{2}})$ approximation to the marginal likelihood for a decomposable log-linear model with Jeffreys' prior. Two examples are presented in this section which demonstrate the relative merits of the uncorrected and corrected Schwarz approximations to the marginal likelihood, based on Jeffreys' prior.

The true marginal likelihood is given by the expression

$$\begin{aligned} f(\boldsymbol{n}) &= \int f(\boldsymbol{p}) f(\boldsymbol{n}|\boldsymbol{p}) d\boldsymbol{p} \\ &= \int \frac{g(\boldsymbol{p})}{c} f(\boldsymbol{n}|\boldsymbol{p}) d\boldsymbol{p} \\ &= \frac{C}{c} \end{aligned}$$

and so

$$\log f(\boldsymbol{n}) = \log C - \log c$$

where c is the normalising constant for the prior density and C is the normalising constant of $g(\boldsymbol{p})f(\boldsymbol{n}|\boldsymbol{p})$.

In general, c and C may not be available without the use of an approximation method such as Laplace's method or bridge sampling. Hence, a computational saving may be made by using the corrected Schwarz approximation, as only c is needed for this.

7.4.1 Example 1

Here, we consider Jeffreys' prior and resulting posterior for a saturated model, where normalising constants are always easy to compute.

The (uncorrected) Schwarz approximation to the marginal likelihood is derived from

$$\log f(\boldsymbol{n}) = \log f(\boldsymbol{n}|\widehat{\boldsymbol{p}}) - \frac{d}{2}\log n + O(1) \tag{7.6}$$

where d is the number of parameters in the model and n is the total sample size. For this example, expression (7.6) becomes

$$\begin{aligned}
\log f(\boldsymbol{n}) &= \log f(\boldsymbol{n}|\widehat{\boldsymbol{p}}) - \frac{d}{2}\log n + O(1) \\
&= \log\left(\frac{\Gamma(n)}{\prod_{\boldsymbol{i}}\Gamma(n(\boldsymbol{i}))}\prod_{\boldsymbol{i}}\widehat{p}(\boldsymbol{i})^{n(\boldsymbol{i})}\right) - \frac{d}{2}\log n + O(1) \\
&= \log\left(\frac{\Gamma(n)}{\prod_{\boldsymbol{i}}\Gamma(n(\boldsymbol{i}))}\prod_{\boldsymbol{i}}\frac{n(\boldsymbol{i})}{n}^{n(\boldsymbol{i})}\right) - \frac{d}{2}\log n + O(1) \\
&= \log\Gamma(n) - \sum_{\boldsymbol{i}}\log\Gamma(n(\boldsymbol{i})) + \sum_{\boldsymbol{i}} n(\boldsymbol{i})\log n(\boldsymbol{i}) - n\log n - \frac{d}{2}\log n + O(1)
\end{aligned}$$

The correction factor for this expression, as derived in section 7.3.3, is given by

$$\delta = \frac{d}{2}\log 2\pi - \log c$$

where c is the normalising constant for the Jeffreys' prior for the log-linear model. For the saturated model, the Jeffreys' prior is of well-known form

$$f(\boldsymbol{p}) \propto \prod_{\boldsymbol{i}} p(\boldsymbol{i})^{-\frac{1}{2}}$$

and the normalising constant is therefore

$$c = \frac{\Gamma\left(\frac{1}{2}\right)^{d+1}}{\Gamma\left(\frac{d+1}{2}\right)}$$

where $d+1$ in this case is the number of cells in the table (as we have a saturated model). The correction term is therefore

$$\begin{aligned}\delta &= \frac{d}{2}\log 2\pi - \frac{(d+1)}{2}\log\pi + \log\Gamma\left(\frac{d+1}{2}\right) \\ &= \frac{d}{2}\log 2 - \frac{1}{2}\log\pi + \log\Gamma\left(\frac{d+1}{2}\right)\end{aligned}$$

Hence the corrected Schwarz approximation is given by

$$\begin{aligned}\log f(\boldsymbol{n}) &= \log\Gamma(n) - \sum_{\boldsymbol{i}}\log\Gamma(n(\boldsymbol{i})) + \sum_{\boldsymbol{i}} n(\boldsymbol{i})\log n(\boldsymbol{i}) - n\log n - \frac{d}{2}\log n + \\ &\quad \frac{d}{2}\log 2\pi - (d+1)\log\Gamma\left(\frac{1}{2}\right) + \log\Gamma\left(\frac{d+1}{2}\right) + O(n^{-\frac{1}{2}}) \\ &= \log\Gamma(n) - \sum_{\boldsymbol{i}}\log\Gamma(n(\boldsymbol{i})) + \sum_{\boldsymbol{i}} n(\boldsymbol{i})\log n(\boldsymbol{i}) - n\log n - \frac{d}{2}\log n + \\ &\quad \frac{d}{2}\log 2 - \frac{1}{2}\log\pi + \log\Gamma\left(\frac{d+1}{2}\right) + O(n^{-\frac{1}{2}}) \qquad (7.7)\end{aligned}$$

For the saturated model, it is also easy to calculate the true marginal likelihood. The value of C (as defined above) in this instance is given by

$$\begin{aligned}C &= \int g(\boldsymbol{p}) f(\boldsymbol{n}|\boldsymbol{p}) d\boldsymbol{p} \\ &= \frac{\prod_{\boldsymbol{i}}\Gamma\left(n(\boldsymbol{i}) + \frac{1}{2}\right)}{\Gamma\left[n + \frac{d+1}{2}\right]}\end{aligned}$$

and so the marginal likelihood $\log f(\boldsymbol{n})$ is

$$\log f(\boldsymbol{n}) = \log\Gamma(n) - \sum_{\boldsymbol{i}} \log\Gamma(n(\boldsymbol{i})) + \sum_{\boldsymbol{i}} \log\Gamma\left(n(\boldsymbol{i}) + \frac{1}{2}\right) - \log\Gamma\left[n + \frac{d+1}{2}\right] - \frac{d+1}{2}\log\pi + \log\Gamma\left(\frac{d+1}{2}\right) \quad (7.8)$$

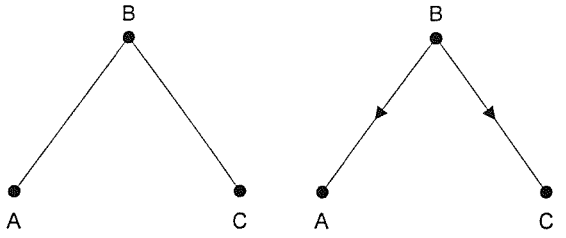
Application of Stirling's approximation, given by the expression $\log\Gamma(z) = (z - \frac{1}{2})\log z - z + \frac{1}{2}\log 2\pi + O(z^{-1})$, to (7.8) yields

$$\begin{aligned}
\log f(\boldsymbol{n}) &= \log\Gamma(n) - \sum_{\boldsymbol{i}} \log\Gamma(n(\boldsymbol{i})) + \sum_{\boldsymbol{i}} n(\boldsymbol{i}) \log\left[n(\boldsymbol{i}) + \frac{1}{2}\right] - \\
&\quad \sum_{\boldsymbol{i}} \left[n(\boldsymbol{i}) + \frac{1}{2}\right] + \frac{d+1}{2}\log 2\pi - \left[n + \frac{d}{2}\right] \log\left(n + \frac{d+1}{2}\right) + \\
&\quad n + \frac{d+1}{2} + \frac{1}{2}\log 2\pi - \frac{d+1}{2}\log\pi + \log\Gamma\left(\frac{d+1}{2}\right) + O(n^{-1}) \\
&= \log\Gamma(n) - \sum_{\boldsymbol{i}} \log\Gamma(n(\boldsymbol{i})) + \sum_{\boldsymbol{i}} n(\boldsymbol{i}) \log\left[n(\boldsymbol{i}) + \frac{1}{2}\right] - \\
&\quad n - \frac{d+1}{2} + \frac{d+1}{2}\log 2\pi - \left[n + \frac{d}{2}\right] \log\left(n + \frac{d+1}{2}\right) + \\
&\quad n + \frac{d+1}{2} - \frac{1}{2}\log 2\pi - \frac{d+1}{2}\log\pi + \log\Gamma\left(\frac{d+1}{2}\right) + O(n^{-1}) \\
&= \log\Gamma(n) - \sum_{\boldsymbol{i}} \log\Gamma(n(\boldsymbol{i})) + \sum_{\boldsymbol{i}} n(\boldsymbol{i}) \log\left[n(\boldsymbol{i}) + \frac{1}{2}\right] + \frac{d}{2}\log 2 - \\
&\quad \left[n + \frac{d}{2}\right] \log\left(n + \frac{d+1}{2}\right) - \frac{1}{2}\log\pi + \log\Gamma\left(\frac{d+1}{2}\right) + O(n^{-1})
\end{aligned}$$

which is equal to expression (7.7) as $n(\boldsymbol{i}) \to \infty$. This validates the corrected Schwarz approximation for this model, and shows that in this instance it approximates the marginal likelihood with error of order $O(n^{-1})$.

7.4.2 Example 2

This example involves the 3 variable model with one conditional independence represented by the graphs



The Jeffreys' prior will be based on the parameterisation using the conditional probabilities from the directed version of the graph. The (uncorrected) Schwarz approximation to the marginal likelihood is derived from

$$\log f(\boldsymbol{n}) = \log f(\boldsymbol{n}|\widehat{\boldsymbol{p}}) - \frac{d}{2}\log n + O(1) \tag{7.9}$$

where d is the number of parameters in the model and n is the total sample size. For this example, expression (7.9) becomes

$$\begin{aligned}
\log f(\boldsymbol{n}) &= \log f(\boldsymbol{n}|\widehat{\boldsymbol{p}}) - \frac{d}{2}\log n + O(1) \\
&= \log\left(\frac{\Gamma(n)}{\prod_{\boldsymbol{i}}\Gamma(n(\boldsymbol{i}))}\prod_{\boldsymbol{i}}\widehat{p}(\boldsymbol{i})^{n(\boldsymbol{i})}\right) - \frac{d}{2}\log n + O(1) \\
&= \log\left(\frac{\Gamma(n)}{\prod_{\boldsymbol{i}}\Gamma(n(\boldsymbol{i}))}\prod_{\boldsymbol{i}}\frac{\widehat{p}(i_a,i_b)^{n(i_a,i_b)}\widehat{p}(i_b,i_c)^{n(i_b,i_c)}}{\widehat{p}(i_b)^{n(i_b)}}\right) - \frac{d}{2}\log n + O(1) \\
&= \log\Gamma(n) - \sum_{\boldsymbol{i}}\log\Gamma(n(\boldsymbol{i})) + \sum n(i_a,i_b)\log\widehat{p}(i_a,i_b) + \\
&\qquad \sum n(i_b,i_c)\log\widehat{p}(i_b,i_c) - \sum n(i_b)\log\widehat{p}(i_b) - \frac{d}{2}\log n + O(1) \\
&= \log\Gamma(n) - \sum_{\boldsymbol{i}}\log\Gamma(n(\boldsymbol{i})) + \sum n(i_a,i_b)\log\frac{n(i_a,i_b)}{n} + \\
&\qquad \sum n(i_b,i_c)\log\frac{n(i_b,i_c)}{n} - \sum n(i_b)\log\frac{n(i_b)}{n} - \frac{d}{2}\log n + O(1)
\end{aligned}$$

$$= \log\Gamma(n) - \sum_{\boldsymbol{i}} \log\Gamma(n(\boldsymbol{i})) + \sum n(i_a, i_b)\log n(i_a, i_b) + \sum n(i_b, i_c)\log n(i_b, i_c) - \sum n(i_b)\log n(i_b) - n\log n - \frac{d}{2}\log n + O(1) \quad (7.10)$$

The correction factor for this expression when we use the Jeffreys' prior, as derived in section 7.3.3, is given by the expression

$$\delta = \frac{d}{2}\log 2\pi - \log c$$

where c is the normalising constant for the Jeffreys' prior for the log-linear model.

The Jeffreys' prior in this instance puts independent Dirichlet distributions on the parameters. Specifically, $P(B)$ follows a $Dirichlet(\frac{|I_A|+|I_C|-1}{2}\mathbf{1})$ distribution, and $P(A|B)$ and $P(C|B)$ follow $Dirichlet(\frac{1}{2}\mathbf{1})$ distributions. We may therefore calculate c

$$\begin{aligned} c &= \int g(\boldsymbol{p})d\boldsymbol{p} \\ &= \frac{\Gamma\left(\frac{|I_A|+|I_C|-1}{2}\right)^{|I_B|}}{\Gamma\left(\frac{|I_B|(|I_A|+|I_C|-1)}{2}\right)} \left[\frac{\Gamma\left(\frac{1}{2}\right)^{|I_A|+|I_C|}}{\Gamma\left(\frac{|I_A|}{2}\right)\Gamma\left(\frac{|I_C|}{2}\right)}\right]^{|I_B|} \end{aligned}$$

where $|I_A|, |I_B|$ and $|I_C|$ are the numbers of levels of variables A, B and C respectively. Hence

$$\begin{aligned} \delta &= \frac{d}{2}\log 2\pi - \log c \\ &= \frac{|I_A|\,|I_B| + |I_A|\,|I_C| - |I_A| - 1}{2}\log 2\pi - |I_B|\log\Gamma\left(\frac{|I_A| + |I_C| - 1}{2}\right) + \\ &\quad \log\Gamma\left(\frac{|I_B|\,(|I_A| + |I_C| - 1)}{2}\right) - \frac{|I_B|\,|I_A| + |I_B|\,|I_C|}{2}\log\pi + \\ &\quad |I_B|\log\Gamma\left(\frac{|I_A|}{2}\right) + |I_B|\log\Gamma\left(\frac{|I_C|}{2}\right) \end{aligned}$$

$$= \frac{|I_A|\,|I_B| + |I_A|\,|I_C| - |I_A| - 1}{2}\log 2\pi - |I_B|\log\Gamma\left(\frac{|I_A| + |I_C| - 1}{2}\right) + \log\Gamma\left(\frac{|I_B|\,(|I_A| + |I_C| - 1)}{2}\right) - \frac{|I_B|\,|I_A| + |I_B|\,|I_C|}{2}\log\pi + |I_B|\log\Gamma\left(\frac{|I_A|}{2}\right) + |I_B|\log\Gamma\left(\frac{|I_C|}{2}\right)$$

The corrected Schwarz approximation is therefore given by

$$\log f(\boldsymbol{n}) = \log\Gamma(n) - \sum_{\boldsymbol{i}}\log\Gamma(n(\boldsymbol{i})) + \sum n(i_a, i_b)\log n(i_a, i_b) + \sum n(i_b, i_c)\log n(i_b, i_c) - \sum n(i_b)\log n(i_b) - n\log n - \frac{d}{2}\log n + \frac{|I_A|\,|I_B| + |I_A|\,|I_C| - |I_A| - 1}{2}\log 2\pi + |I_B|\log\Gamma\left(\frac{|I_A|}{2}\right) - |I_B|\log\Gamma\left(\frac{|I_A| + |I_C| - 1}{2}\right) + \log\Gamma\left(\frac{|I_B|\,(|I_A| + |I_C| - 1)}{2}\right) - \frac{|I_B|\,|I_A| + |I_B|\,|I_C|}{2}\log\pi + |I_B|\log\Gamma\left(\frac{|I_C|}{2}\right) + O(n^{-\frac{1}{2}})$$

In this instance, the true marginal likelihood is also available without the use of computational approximations. The value of C is given by

$$C = \int g(\boldsymbol{p})f(\boldsymbol{n}|\boldsymbol{p})d\boldsymbol{p} = \frac{\prod_{i_b}\Gamma\left(n(i_b) + \frac{|I_A|+|I_C|-1}{2}\right)}{\Gamma\left(\frac{|I_B|(|I_A|+|I_C|-1)}{2} + n\right)}\prod_{i_b}\left(\frac{\prod_{i_a}\Gamma(\frac{1}{2} + n(i_a|i_b))}{\Gamma(\frac{|I_A|}{2} + \sum_{i_a} n(i_a|i_b))}\right)\times \prod_{i_b}\left(\frac{\prod_{i_c}\Gamma(\frac{1}{2} + n(i_c|i_b))}{\Gamma(\frac{|I_C|}{2} + \sum_{i_c} n(i_c|i_b))}\right)$$

It is therefore possible to evaluate the marginal likelihood $\log f(\boldsymbol{n}) = \log C - \log c$ for a particular data set. Table 7.1 shows the true marginal likelihood (ML), together with the corrected and uncorrected values of the Schwarz approximation, for a variety of sample sizes using the above model with varying numbers

of levels of A, B and C. The data is spread such that there is an equal sample in each cell.

Levels of A, B, C	Sample Size n	True M L	Uncorrected Schwarz	Error in Uncorrected Schwarz	Corrected Schwarz	Error in Corrected Schwarz
2, 2, 2	8	−12.74	−13.31	−0.57	−12.36	0.38
2, 2, 2	16	−11.55	−12.30	−0.75	−11.35	0.20
2, 2, 2	24	−10.97	−11.79	−0.82	−10.84	0.13
2, 2, 2	40	−10.32	−11.19	−0.87	−10.24	0.08
2, 2, 2	80	−9.52	−10.43	−0.91	−9.48	0.04
2, 2, 2	160	−8.78	−9.71	−0.93	−8.76	0.02
2, 2, 2	400	−7.83	−8.77	−0.94	−7.82	0.01
2, 2, 3	12	−19.24	−21.01	−1.77	−18.75	0.49
2, 2, 3	120	−13.33	−15.54	−2.21	−13.28	0.05
2, 2, 4	16	−25.54	−28.94	−3.40	−24.93	0.61
2, 2, 4	160	−16.92	−20.87	−3.95	−16.86	0.06
2, 4, 2	16	−26.41	−31.71	−5.30	−25.49	0.93
2, 4, 2	160	−19.82	−25.95	−6.13	−19.73	0.09
4, 4, 4	64	−98.20	−121.30	−23.10	−97.06	1.14
4, 4, 4	640	−51.08	−75.20	−24.12	−50.96	0.12

Table 7.1: Errors in Schwarz approximations for Example 2

Table 7.1 shows that, as expected, the error of the corrected Schwarz approximation (expressed as approximation minus true value) reduces considerably with increasing sample size, consistent with the previous assertion that the error is of order $O(n^{-\frac{1}{2}})$, and it would appear that this behaviour is independent of the number of levels of the variables. Also, the order $O(1)$ error of the uncorrected Schwarz approximation is apparent from the table, as this approximation does not improve with increasing sample size. Indeed, for the final entry, the error of this approximation is 24.12 on the log scale – a huge discrepancy. The results for the $2 \times 2 \times 2$ table are presented graphically below, showing closer agreement for increasing sample sizes.

Table 7.2 shows the true marginal likelihood, together with the Schwarz ap-

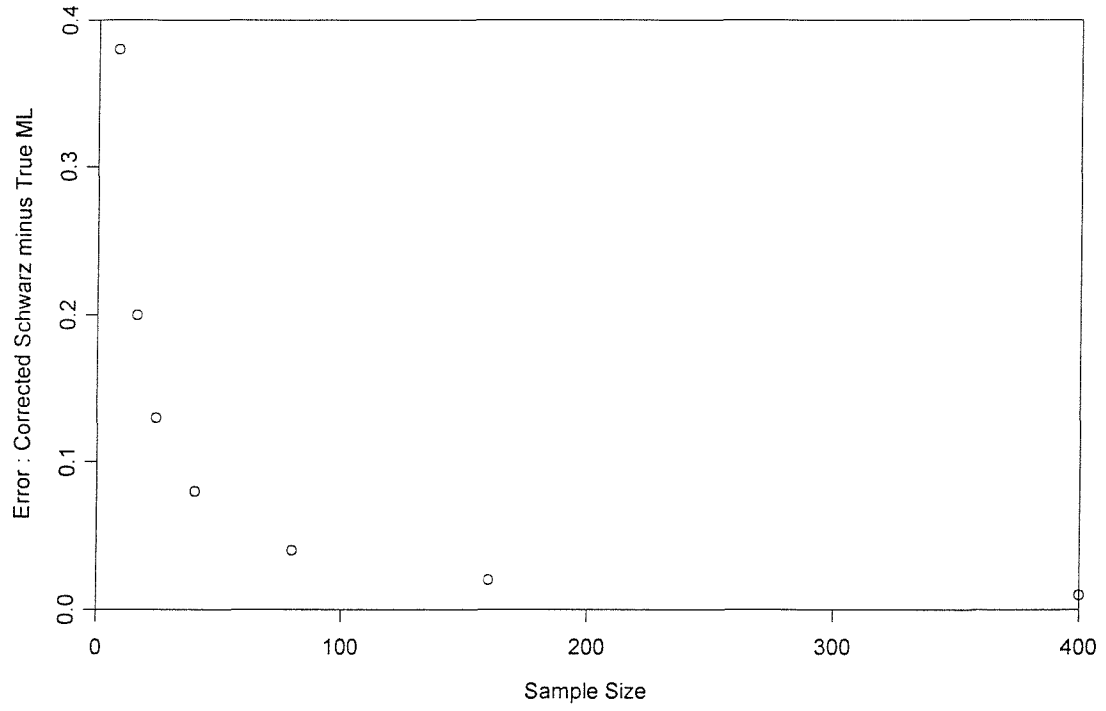


Figure 7.1: Plot of error in corrected Schwarz approximation against sample size for Example 2

proximations, for data in a $2 \times 2 \times 2$ table where there is a single observation in all but one cell, and the remainder of the data is in the final cell. As the results do not seem to depend upon the numbers of levels of the variables, all variables have 2 levels in this case.

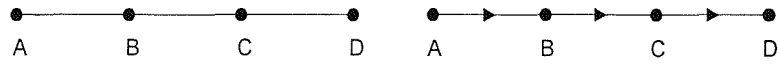
Again, this table demonstrates the quality of the corrected Schwarz approximation to the marginal likelihood corresponding to the 3 variable model with one conditional independence.

Sample Size n	True Marginal Likelihood	Uncorrected Schwarz Approximation	Error in Uncorrected Schwarz	Corrected Schwarz Approximation	Error in Corrected Schwarz
17	−14.46	−15.20	−0.74	−14.24	0.22
27	−15.99	−16.78	−0.79	−15.83	0.16
37	−17.06	−17.87	−0.81	−16.92	0.14
57	−18.54	−19.38	−0.84	−18.43	0.11
107	−20.72	−21.58	−0.86	−20.63	0.09

Table 7.2: Errors in Schwarz approximations for Example 2 − unbalanced case

7.4.3 Example 3

This example concerns the 4 variable 'straight-line' model which may be represented graphically by



As discussed in section 6.5, it is not possible to obtain the normalising constant for the Jeffreys' prior corresponding to this model, and the marginal likelihood is therefore only available by using a computational approximation such as bridge sampling. In this section, the quality of the corrected Schwarz approximation to the marginal likelihood will be assessed by comparing this to the approximation derived using the bridge sampler. Note that it is necessary to use the bridge sampler once and only once to obtain the correction factor for the Schwarz approximation for a particular model structure, whereas direct approximation of the marginal likelihood requires a further run of the bridge sampler for each data set.

The corrected Schwarz approximation to the marginal likelihood is given by

$$\log f(\boldsymbol{n}) = \log f(\boldsymbol{n}|\widehat{\boldsymbol{p}}) - \frac{d}{2}\log n + \frac{d}{2}\log 2\pi - \log c + O(n^{-\frac{1}{2}})$$

where d is the number of parameters in the model, n is the total sample size, and c is the normalising constant for the Jeffreys' prior for the log-linear model.

The number of parameters, d, in this model is $|I_A|\,|I_B| + |I_B|\,|I_C| + |I_C|\,|I_D| - |I_B| - |I_C| - 1$. Using the bridge sampler, for a model with all binary variables, the value of $\log c$ was found to be 4.98 (values for different numbers of levels are presented later), and the number of parameters is 7. Hence

$$\begin{aligned}\log f(\boldsymbol{n}) &= \log f(\boldsymbol{n}|\widehat{\boldsymbol{p}}) - \frac{d}{2}\log n + \frac{d}{2}\log 2\pi - \log c + O(n^{-\frac{1}{2}}) \\ &= \log\Gamma(n) - \sum_{\boldsymbol{i}} \log\Gamma(n(\boldsymbol{i})) + \sum n(i_a, i_b)\log n(i_a, i_b) + \\ &\quad \sum n(i_b, i_c)\log n(i_b, i_c) + \sum n(i_c, i_d)\log n(i_c, i_d) - \\ &\quad \sum n(i_b)\log n(i_b) - \sum n(i_c)\log n(i_c) - n\log n - \\ &\quad \frac{7}{2}\log n + \frac{7}{2}\log 2\pi - 4.98 + O(n^{-\frac{1}{2}})\end{aligned}$$

The true marginal likelihood may be evaluated using the expression $\log f(\boldsymbol{n}) = \log C - \log c$, where $C = \int g(\boldsymbol{p})g(\boldsymbol{n}|\boldsymbol{p})d\boldsymbol{p}$ is again obtained using bridge sampling.

Table 7.3 shows the true marginal likelihood together with the corrected Schwarz approximation, for a variety of sample sizes using the above model with A, B, C and D each having 2 levels. The data is spread such that all cell counts are equal.

Sample Size n	True Marginal Likelihood	Uncorrected Schwarz Approximation	Error in Uncorrected Schwarz	Corrected Schwarz Approximation	Error in Corrected Schwarz
16	−25.01	−26.17	−1.16	−24.71	0.30
32	−21.46	−22.76	−1.30	−21.31	0.15
64	−19.57	−20.93	−1.36	−19.47	0.10
80	−17.31	−18.70	−1.39	−17.25	0.06
160	−14.38	−15.80	−1.42	−14.34	0.04
240	−12.70	−14.13	−2.03	−12.68	0.02

Table 7.3: Errors in Schwarz approximations for Example 3

As in the previous example, the table demonstrates the accuracy of the corrected Schwarz approximation, consistent with the $O(n^{-\frac{1}{2}})$ error, against the much poorer uncorrected approximation. The results are summarised graphi-

cally in figure 7.2.

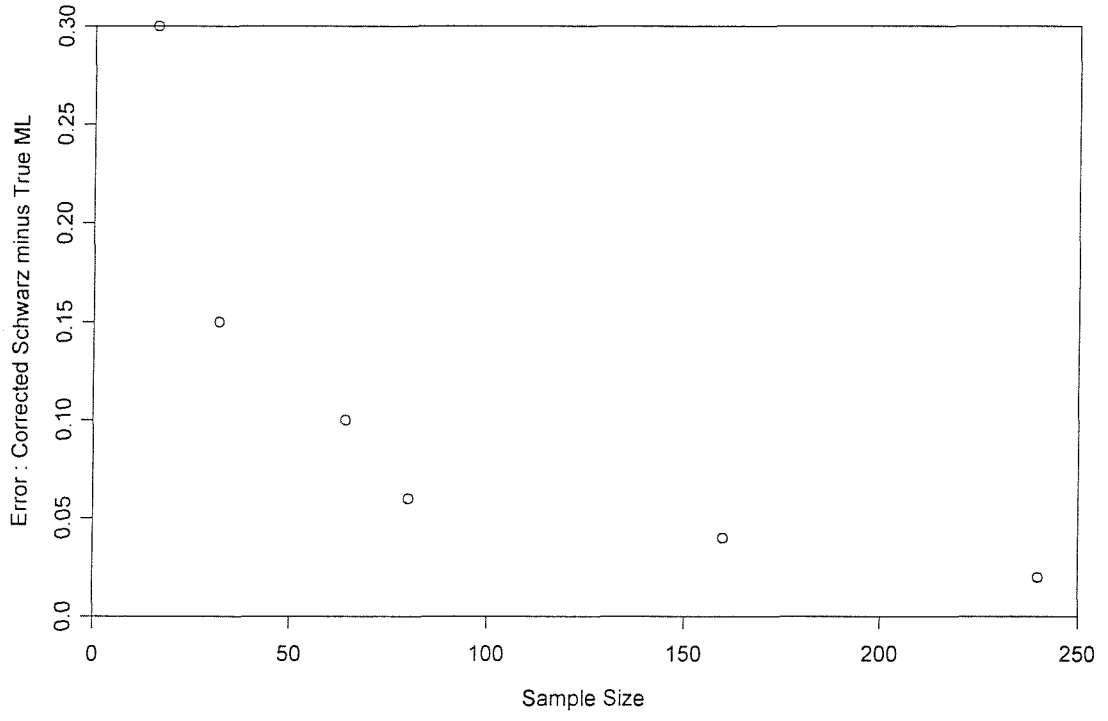


Figure 7.2: Plot of error in corrected Schwarz approximation against sample size for Example 3

Additional correction terms for different numbers of levels of the variables in this model may be calculated by further runs of the bridge sampler. The normalising constants ($\log c$) for various cases are presented in table 7.4. The equation

$$\delta = \frac{d}{2}\log 2\pi - \log c$$

is then used to determine the appropriate correction term, where $d = |I_A|\,|I_B| + |I_B|\,|I_C| + |I_C|\,|I_D| - |I_B| - |I_C| - 1$.

Levels of A, B, C, D	$\log c$	Correction term, δ
$2.2, 2, 2$	4.98	1.45
$3.2, 2, 2$	5.52	2.75
$2, 3, 2, 2$	5.52	3.67
$3, 3, 2, 2$	5.31	6.64
$3, 3, 3, 2$	5.53	10.09
$3, 3, 3, 3$	5.60	12.78

Table 7.4: Schwarz approximation correction terms for model $AB + BC + CD$

7.4.4 Further Examples

The previous examples present the correction factors for all saturated models, one 3 variable model and one 4 variable model. In this section, the correction terms for the remaining distinct models with up to and including 4 variables are given. Note that where a graph consists of several disconnected components, each one analogous to a model considered here, then that model will not be treated separately; such correction terms follow directly from these results.

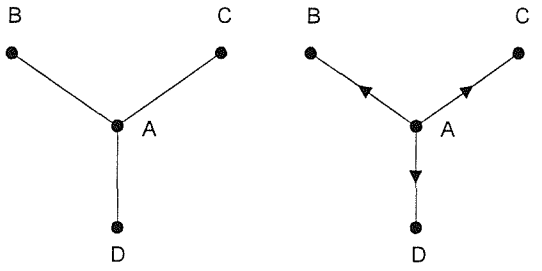
Single Variable Model

The correction term, δ, for this model is

$$\delta = \frac{1}{2}\log 2\pi - \log\left[\frac{\Gamma(\frac{1}{2})^{|I_A|}}{\Gamma(\frac{|I_A|}{2})}\right]$$

Model $AB + AC + AD$

The correction term for this model, represented graphically by

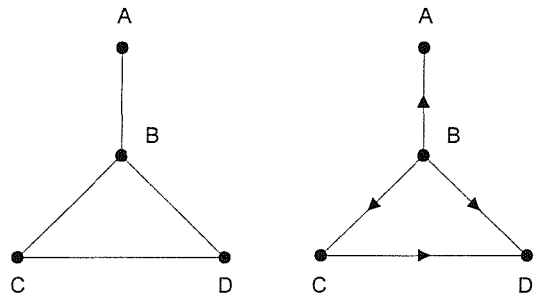


is

$$\begin{aligned}\delta &= \frac{d}{2}\log 2\pi - \log\left[\frac{\Gamma(\frac{|I_B|+|I_C|+|I_D|}{2}-1)^{|I_A|}}{\Gamma(\frac{|I_A|(|I_B|+|I_C|+|I_D|)}{2}-|I_A|)}\left(\frac{\Gamma(\frac{1}{2})^{|I_B|+|I_C|+|I_D|}}{\Gamma(\frac{|I_B|}{2})\Gamma(\frac{|I_C|}{2})\Gamma(\frac{|I_D|}{2})}\right)^{|I_A|}\right] \\ &= \frac{|I_A|\,|I_B|+|I_A|\,|I_C|+|I_A|\,|I_D|-2\,|I_A|-1}{2}\log 2\pi - \\ &\qquad \log\left[\frac{\Gamma(\frac{|I_B|+|I_C|+|I_D|}{2}-1)^{|I_A|}}{\Gamma(\frac{|I_A|(|I_B|+|I_C|+|I_D|)}{2}-|I_A|)}\left(\frac{\Gamma(\frac{1}{2})^{|I_B|+|I_C|+|I_D|}}{\Gamma(\frac{|I_B|}{2})\Gamma(\frac{|I_C|}{2})\Gamma(\frac{|I_D|}{2})}\right)^{|I_A|}\right]\end{aligned}$$

Model $AB + BCD$

This correction term for this model, represented graphically by

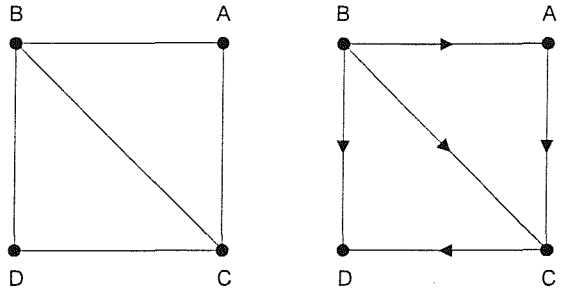


is

$$\begin{aligned}\delta &= \frac{d}{2}\log 2\pi - \log\left[\frac{\Gamma(\frac{|I_A|+|I_C|+|I_D|-1}{2})^{|I_B|}\Gamma(\frac{|I_D|}{2})^{|I_C||I_B|}\Gamma(\frac{1}{2})^{|I_A||I_B|}\Gamma(\frac{1}{2})^{|I_B||I_C||I_D|}}{\Gamma(\frac{|I_B|(|I_A|+|I_C|+|I_D|-1)}{2})\Gamma(\frac{|I_C||I_D|}{2})^{|I_B|}\Gamma(\frac{|I_A|}{2})^{|I_B|}\Gamma(\frac{|I_D|}{2})^{|I_B||I_C|}}\right] \\ &= \frac{|I_A|\,|I_B|+|I_B|\,|I_C|\,|I_D|-|I_B|-1}{2}\log 2\pi - \\ &\qquad \log\left[\frac{\Gamma(\frac{|I_A|+|I_C|+|I_D|-1}{2})^{|I_B|}\Gamma(\frac{|I_D|}{2})^{|I_C||I_B|}\Gamma(\frac{1}{2})^{|I_A||I_B|}\Gamma(\frac{1}{2})^{|I_B||I_C||I_D|}}{\Gamma(\frac{|I_B|(|I_A|+|I_C|+|I_D|-1)}{2})\Gamma(\frac{|I_C||I_D|}{2})^{|I_B|}\Gamma(\frac{|I_A|}{2})^{|I_B|}\Gamma(\frac{|I_D|}{2})^{|I_B||I_C|}}\right]\end{aligned}$$

Model $ABC + BCD$

This correction term for this model, represented graphically by



is

$$\begin{aligned}\delta &= \frac{d}{2}\log 2\pi - \log\left[\frac{\Gamma(\frac{|I_A||I_C|+|I_C||I_D|-|I_C|}{2})^{|I_B|}\Gamma(\frac{|I_A|+|I_D|-1}{2})^{|I_C||I_B|}}{\Gamma(\frac{|I_B|(|I_A||I_C|+|I_C||I_D|-|I_C|)}{2})\Gamma(\frac{|I_C|(|I_A|+|I_D|-1)}{2})^{|I_B|}} \times \right. \\ &\qquad \left. \left(\frac{\Gamma(\frac{1}{2})^{|I_A|+|I_D|}}{\Gamma(\frac{|I_A|}{2})\Gamma(\frac{|I_D|}{2})}\right)^{|I_B||I_C|}\right] \\ &= \frac{|I_A|\,|I_B|\,|I_C| + |I_B|\,|I_C|\,|I_D| - |I_B|\,|I_C| - 1}{2}\log 2\pi - \\ &\qquad \log\left[\frac{\Gamma(\frac{|I_A||I_C|+|I_C||I_D|-|I_C|}{2})^{|I_B|}\Gamma(\frac{|I_A|+|I_D|-1}{2})^{|I_C||I_B|}}{\Gamma(\frac{|I_B|(|I_A||I_C|+|I_C||I_D|-|I_C|)}{2})\Gamma(\frac{|I_C|(|I_A|+|I_D|-1)}{2})^{|I_B|}} \times \right. \\ &\qquad \left. \left(\frac{\Gamma(\frac{1}{2})^{|I_A|+|I_D|}}{\Gamma(\frac{|I_A|}{2})\Gamma(\frac{|I_D|}{2})}\right)^{|I_B||I_C|}\right]\end{aligned}$$

7.5 Discussion

The focus of this Chapter has been to investigate the choice of parameters for prior distributions for log-linear models. The concept of unit information has been discussed. Using Normal priors with variances determined using unit information considerations, it is possible to choose a prior distribution such that the Bayes factor based on this prior may be approximated by the Schwarz criterion with an error of order $O(n^{-\frac{1}{2}})$. Moreover, this prior is easily interpretable, as it contains as much information as is present in a single observation.

It was not possible to find a set of $\boldsymbol{\alpha}$ parameters for the conditional Dirichlet distribution such that the Schwarz formula, with or without a correction term,

may be used to approximate the marginal likelihood. However, it was found that Jeffreys' prior, which in the case of many log-linear models using multinomial sampling is based on Dirichlet distributions, offered a good approximation. Indeed, the marginal likelihood based on a Jeffreys' prior may be approximated by the Schwarz formula, with the addition of a correction term. This correction term consists of a constant ($\frac{d}{2}\log 2\pi$) plus the normalising constant for the Jeffreys' distribution, which is either obtained analytically or using a bridge sampling method, though only decomposable log-linear models have been considered here. In principle, the corrected Schwarz approximation could be applied to non-decomposable models, provided that normalising constants for the Jeffreys' priors for such models could be determined.

The corrected Schwarz approximation provides an error of order $O(n^{-\frac{1}{2}})$, and the two examples demonstrated improving accuracy with increasing sample size. The approximation involves the maximised likelihood (easy to determine analytically for decomposable models), and the Jeffreys' normalising constant, for which the bridge sampler may be used if necessary. The alternative calculation of marginal likelihoods potentially requires two applications of bridge sampling, which represents a more awkward and time-consuming method. Alternatively, direct application of Laplace's method is possible, but requires the calculation of the information matrix and its determinant.

Thus, if one is prepared to accept Jeffreys' prior as a suitable reference prior, the corrected Schwarz approximation provides an easy method of obtaining the corresponding marginal likelihood. Correction terms for all distinct models (within disconnected components) with up to and including four variables have been presented in section 7.4.

Chapter 8

Further Examples

Throughout this thesis, several prior distributions have been discussed and various methods of obtaining normalising constants have been developed. This Chapter contains two examples, which have been analysed using appropriate methods for different prior distributions. Note that the use of the bridge sampler to analyse data from Edwards and Havranek (1985) was presented in Chapter 5.

8.1 Example 1

This example concerns 30 patients suffering from lymphocytic lymphoma, and cross-classifies their type of lymphoma L (nodular or diffuse) against their response to combination chemotherapy R and their sex S. The data (presented in table 8.1) is analysed, using various diffuse prior distributions, in order to determine the posterior model probabilities for the 8 potential graphical models. Four priors are used:

The first is the conditional Dirichlet distribution with parameters $\alpha(\boldsymbol{i}) = \frac{1}{2}$, corresponding to conditioning on Jeffreys' prior for the saturated model. The second is the conditional Dirichlet distribution with parameters $\alpha(\boldsymbol{i}) = \frac{1}{8}$, which corresponds to a single observation distributed evenly between all cells (Perks'

Cell Type	Sex	Remission No	Remission Yes
Nodular	Male	1	4
	Female	2	6
Diffuse	Male	12	1
	Female	3	1

Table 8.1: Chemotherapy and lymphoma

prior). Note that the equivalence of the hyper Dirichlet and conditional Dirichlet distributions is used to ease the calculations. Elsewhere the methods described in previous chapters, such as bridge sampling using Monte Carlo samples, are used. The third prior distribution is Jeffreys' prior, and the fourth is a log-Normal prior with parameters chosen using the same considerations as Dellaportas and Forster (1999). The posterior model probabilities are presented in table 8.2 (models with probability less than 0.01 are excluded).

Model	Conditional Dirichlet $\alpha(\boldsymbol{i}) = \frac{1}{2}$	Conditional Dirichlet $\alpha(\boldsymbol{i}) = \frac{1}{8}$	Jeffreys' Prior	Log-Normal Prior
$RC + CS$	0.48	0.42	0.43	0.48
$RC + S$	0.19	0.38	0.13	0.30
$RC + RS$	0.22	0.18	0.19	0.17
RCS	0.09	0.01	0.25	0.05

Table 8.2: Posterior model probabilities for Cancer data using various prior distributions

As expected, all priors identify the most probable model $RC + CS$. Similar probabilities are also obtained for the model $RC+RS$. However, the various priors differ with respect to models $RC + S$ and RCS. As expected, the conditional Dirichlet distribution with $\alpha(\boldsymbol{i}) = \frac{1}{8}$ tends to favour the simpler model $RC + S$. Jeffreys' prior favours the saturated model RCS.

8.2 Example 2

The data analysed here involves 13384 pregnant women, cross-classified according to their social class (C - 5 levels), their smoking habit (S - none, light or heavy), and whether or not they suffer from two toxaemic signs, hypertension (H) and proteinuria (P). The data was collected in England between 1968 and 1977, and the aim of the analysis of the $2 \times 2 \times 3 \times 5$ contingency table (8.3) is to determine relationships between the variables, via the posterior model probabilities for all possible graphical models.

Social Class	Smoking	Proteinuria: Yes		Proteinuria: No	
		Hypertension: Yes	Hypertension: No	Hypertension: Yes	Hypertension: No
1	None	28	82	21	286
	Light	5	24	5	71
	Heavy	1	3	0	13
2	None	50	266	34	785
	Light	13	92	17	284
	Heavy	0	15	3	34
3	None	278	1101	164	3160
	Light	120	492	142	2300
	Heavy	16	92	32	383
4	None	63	213	52	656
	Light	35	129	46	649
	Heavy	7	40	12	163
5	None	20	78	23	245
	Light	22	74	34	321
	Heavy	7	14	4	65

Table 8.3: Toxaemia in pregnancy

As in the first example, four prior distributions are used. These are the conditional Dirichlet distribution with parameters $\alpha(i) = \frac{1}{2}$, the conditional Dirich-

let distribution with parameters $\alpha(i) = \frac{1}{60}$, Jeffreys' prior and a log-Normal prior. Note that for Jeffreys' prior, the three models $CS + SH + HP + PC$, $CH + HS + SP + PC$ and $CS + SP + PH + HC$ are excluded from the analysis, as they are not decomposable.

Under each of the distributions, a maximum of two models were identified as having posterior probabilities greater than 0.001. These are the models $HP + PS + SC$ and $HPS + SC$, and their respective probabilities are shown in table 8.4.

Model	Conditional Dirichlet $\alpha(i) = \frac{1}{2}$	Conditional Dirichlet $\alpha(i) = \frac{1}{60}$	Jeffreys' Prior	Log-Normal Prior
$HP + PS + SC$	0.9950	1.0000	0.9877	1.0000
$HPS + SC$	0.0050	0.0000	0.0123	0.0000

Table 8.4: Posterior model probabilities for Toxaemia data using various prior distributions

These results are surprising, as the classical maximum likelihood approach selects model $HP + PS + SC + CH$. However, each of the priors used here gives a posterior model probability $< 10^{-6}$ to this model.

Results based on the Schwarz criterion suggest model $HP + PS + SC + CH$ as the most probable. Comparing models $HP + PS + SC + CH$ and $HPS + SC$ using the Schwarz criterion results in a difference of 15.7 in favour of model $HP + PS + SC + CH$, and comparison of models $HP + PS + SC + CH$ and $HP + PS + SC$ gives a difference of 4.4 in favour of model $HP + PS + SC + CH$. However, since the Schwarz criterion only approximates the log Bayes factor with an error of order $O(1)$, we have no real reason to be concerned about the results in table 8.4.

Chapter 9

Discussion and Extensions

9.1 Discussion

The aim of this thesis has been to fully investigate Bayesian methods for log-linear models, with particular attention to the use of reference priors. Several prior distributions have been investigated, with particular focus on the conditional Dirichlet distribution and Jeffreys' prior.

The conditional Dirichlet distribution, defined in Chapter 3, has the attractive property that its parameters may be interpreted as prior cell counts. This makes it useful for both reference analyses, where small prior values are used, and as an informative prior, where (hypothetical) prior cell counts may be available. The conditional Dirichlet was shown to be equivalent to a hyper Dirichlet density (which admits straightforward analyses) for decomposable log-linear models. Hence a natural extension of the hyper Dirichlet distribution to non-decomposable models has been obtained.

The conditional Dirichlet distribution is not tractable in general, so Monte Carlo and other approximation methods are required. Gibbs sampling was applied in Chapter 4 to obtain samples from prior and posterior conditional Dirichlet distributions. The sampler was found to mix well, producing samples which

are not highly dependent.

Laplace's method for the approximation of integrals was introduced and applied in Chapter 5, although it was found to perform poorly where prior parameters take small values. However, accurate results may be obtained for the posterior analysis of datasets where cell counts are large. The method of bridge sampling was introduced, and applied to the problem of determining the normalising constants for conditional Dirichlet distributions. The sampler was found to produce good results, even when prior parameters take small values, and this was illustrated by application to several examples.

Jeffreys' prior, which is a reference prior by definition, was considered in Chapter 6. An explicit expression was presented for the Jeffreys' prior for a decomposable log-linear model, and in many cases this was found to be a product of independent Dirichlet distributions for the parameters of a particular decomposition of the model. For other decomposable models, where the normalising constant for Jeffreys' prior is not directly available, the method of bridge sampling was again applied, and found to produce accurate results. The Monte Carlo samples needed were obtained using Metropolis Hastings sampling.

The choice of prior distribution was considered in further detail in Chapter 7. Unit information priors, for which easy approximations to marginal likelihoods are available, were discussed, and the relationship between the Laplace approximation for marginal likelihoods and the Schwarz criterion was investigated for log-linear models under multinomial sampling. It was shown that marginal likelihoods using Jeffreys' prior may be approximated by a modified version of the Schwarz approximation, with error of order $O(n^{-\frac{1}{2}})$. This provides an easy approximation, in particular for models whose Jeffreys' priors are intractable, where bridge sampling is only needed to determine the prior (and not posterior) normalising constant.

In Chapter 8, the various ideas introduced in the thesis were applied to two

data analyses, and the results discussed.

9.2 Extensions

Several ideas investigated in this thesis give rise to possible avenues for additional research.

The examples and applications presented throughout the thesis are for models with a maximum of six variables, although models with more variables are included implicitly. However, the methods are directly applicable to models with additional variables, the dimensionality of the model limited solely by computing power. Jeffreys' priors for decomposable models with five or more variables may be written down in an explicit form by application of expression (6.4).

An expression was given in Chapter 6 for the Jeffreys' prior for any log-linear model. Although extensive application of this formula did not in general lead to distributions in tractable forms, we believe that it is possible to determine explicit expressions for the Jeffreys' priors for non-decomposable models. Provided that the normalising constants for such distributions may be determined, it is then possible to apply the corrected Schwarz approximation derived in Chapter 7 to non-decomposable models.

The method of bridge sampling was used to accurately approximate the normalising constants for conditional Dirichlet distributions. However, the accuracy of the method decreases with increasingly complex models. A potential extension is therefore to apply path sampling to this problem. Path sampling is a method of approximating normalising constants which is a direct extension of bridge sampling. The method of bridge sampling to determine the normalising constant for density g involves the construction of a (single) bridging density between a sampling distribution q and g. Path sampling extends this idea to construct a path between q and g consisting of a finite number of intermediate bridging densities. Increases in accuracy may be possible using this method, fur-

ther details of which are given by Conigliani and O'Hagan (2000) and Gelman and Meng (1998).

It may be possible to implement MCMC methods which have as their state space both models and model parameters, such as reversible jump MCMC, for model determination using conditional Dirichlet priors. However, since the normalising constants for such distributions are not directly available, such an implementation would not be straightforward.

Finally, whereas the methods used throughout are suitable for the analysis of discrete data, it would be interesting to consider whether there is potential for similar methods in the analysis of continuous, or mixed discrete and continuous, data.

References

Aitchison J. (1985). Practical Bayesian problems in simplex sample spaces. In *Bayesian Statistics 2.* Amsterdam: North Holland, 15-32.

Albert J.H. (1996). Bayesian selection of log-linear models. *The Canadian Journal of Statistics* **24**, 327-347.

Bedrick E.J. *et al* (1996). A new perspective on priors for generalized linear models. *Journal of the American Statistical Association* **91**, 1450-1460.

Bennett C. (1976). Efficient estimation of free energy differences from Monte Carlo data. *Journal of Computational Physics* **22**, 245-268.

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

Berger J.O. and Pericchi L.R. (1993). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association* **91**, 109-122.

Bernardo J.M. (1979). Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical Society (Series B)* **41**, 113-147.

Bishop Y.M.M., Fienberg S.E. and Holland P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice.* MIT Press, Cambridge.

Box G.E.P. and Tiao G.C. (1973). *Bayesian Inference in Statistical Analysis.* Addison-Wesley, Reading MA.

Brooks S.P. (1998). MCMC convergence diagnosis via multivariate bounds on log-concave densities. *The Annals of Statistics* **26**, 398-433.

Carlin B. and Chib S. (1995). Bayesian model choice via Markov chain Monte Carlo. *Journal of the Royal Statistical Society (Series B)* **57**, 473-484.

Conigliani C. and O'Hagan A. (2000). Computing fractional Bayes factors via two-dimensional path sampling. *University of Sheffield Technical Report.*

Darroch J.N., Lauritzen S.L. and Speed T.P. (1980). Markov fields and log-linear interaction models for contingency tables. *The Annals of Statistics* **8**, 522-539.

Dawid A.P. and Lauritzen S.L. (1989). Markov distributions, hyper-Markov laws and meta-Markov models on decomposable graphs, with applications to Bayesian learning in expert systems. *BAIES Report BR-10.*

Dawid A.P. and Lauritzen S.L. (1993). Hyper Markov laws in the statistical analysis of decomposable graphical models. *The Annals of Statistics* **21**, 1272-1317.

Dawid A.P. and Lauritzen S.L. (2000). Compatible prior distributions. *Research Report No. 210, University College, London.*

De Santis F. and Spezzaferri F. (1999). Methods for default and robust Bayesian model comparison: the fractional Bayes factor approach. *International Statistical Review* **67**, 267-286.

DeGroot M.H. (1970). *Optimal Statistical Decisions.* McGraw-Hill, New York.

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

Dellaportas P., Forster J.J. and Ntzoufras I. (2001). On Bayesian model and variable selection using MCMC. *Statistics and Computing* to appear.

Dellaportas P. and Smith A.F.M. (1993). Bayesian inference for generalized linear and proportional hazards models via Gibbs sampling. *Applied Statistics* **42**, 443-459.

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

Fienberg S.E. (1972). The multiple recapture census for closed populations and incomplete 2^k contingency tables. *Biometrika* **59**, 591-603.

Forster J.J. (1992). Models and marginal densities for multiway contingency tables. *University of Nottingham PhD Thesis.*

Forster J.J. (1999). Symmetric models and prior distributions for multiway contingency tables. *University of Southampton working paper.*

Forster J.J. and Skene A.M. (1994). Calculation of marginal densities for parameters of multinomial distributions. *Statistics and Computing* **4**, 279-286.

Gelmen A. and Meng X. (1998). Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. *Statistical Science* **13**, 163-185.

Gilks W.R., Richardson S., and Speigelhalter D.J. (1996). *Markov chain Monte Carlo in practice.* Chapman and Hall, London.

Gilks W.R. and Wild P. (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics* **41**, 337-348.

Gilks W.R. and Wild P. (1993). Adaptive rejection sampling from log-concave density functions. *Applied Statistics* **42**, 701-709.

Good I.J. (1956). On the estimation of small frequencies in contingency tables. *Journal of the Royal Statistical Society (Series B)* **18**, 113-124.

Green P. (1995). Reversible jump MCMC computation and Bayesian model determination. *Biometrika* **82**, 711-732.

Gunel E. and Dickey J. (1974). Bayes factors for independence in contingency tables. *Biometrika* **61**, 545-557.

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

Hoeting J.A., Madigan D., Raftery A.E. and Volinsky C.T. (1999). Bayesian model averaging: a tutorial. *Statistical Science* **14**, 382-417.

Ibrahim J.G. and Laud P.W. (1991). On Bayesian analysis of generalized linear models using Jeffreys' prior. *Journal of the American Statistical Association* **86**, 981-986.

Jeffreys H. (1946). An invariant form for the prior probability in estimation problems. *Procedures of the Royal Society London (A)* **186**, 453-461.

Kass R.E. (1989). The geometry of asymptotic inference. *Statistical Science* **4**, 188-234.

Kass R.E. and Raftery A.E. (1995). Bayes factors. *Journal of the American Statistical Association* **90**, 773-795.

Kass R.E. and Vaidyanathan S.K. (1992). Approximate Bayes factors and orthogonal parameters, with application to testing equality of two binomial proportions. *Journal of the Royal Statistical Society (Series B)* **54**, 129-144.

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

Kelley A. and Pohl I. (1990). *A Book On C.* Benjamin/Cummings, California.

Knuiman M.W. and Speed T.P. (1988). Incorporating prior information into the analysis of contingency tables. *Biometrics* **44**, 1061-1071.

Laird N.M. (1978). Empirical Bayes methods for two-way contingency tables. *Biometrika* **65**, 581-590.

Lauritzen S. (1996). *Graphical Models.* Oxford University Press, Oxford.

Lempers F.B. (1971). *Posterior Probabilities of Alternative Linear Models.* University Press, Rotterdam.

Leonard T. (1975). Bayesian estimation methods for two-way contingency tables. *Journal of the Royal Statistical Society (Series B)* **37**, 23-37.

Lidstone G.J. (1920). A note on the general case of the Bayes-Laplace formula for inductive or a posteriori probabilities. *Transactions of the Faculty of Actuaries in Scotland* **8**, 182-192.

Lie R.T., Heuch I. and Irgens L.M. (1994). Maximum likelihood estimation of the proportion of congenital malformations using double registration schemes. *Biometrics* **50**, 433-444.

Lindley D.V. (1964). The Bayesian analysis of contingency tables. *Annals of Mathematical Statistics* **35**, 1622-1643.

Madigan D. and Raftery A.E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the Americal Statistical Association* **89**, 1535-1545.

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

Meng X.L. and Wong W.H. (1993). Simulating ratios of normalizing constants via a simple identity: a theroetical exploration. *Technical Report 365, University of Chicago, Dept. of Statistics.*

Metropolis N. *et al* (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics* **21**, 1087-1092.

O'Hagan A. (1991). Discussion of posterior Bayes factors. *Journal of the Royal Statistical Society (Series B)* **53**, 136.

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

O'Hagan A. (1995). Fractional Bayes factors for model comparison. *Journal of the Royal Statistical Society (Series B)* **57**, 99-138.

Perks W. (1947). Some observations on inverse probability including a new indifference rule. *Journal of the Institute of Actuaries* **73**, 285-334.

Raftery A.E. (1986). A note on Bayes factors for log-linear contingency table models with vague prior information. *Journal of the Royal Statistical Society (Series B)* **48**, 249-250.

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

Raftery A.E., Madigan D. and Hoeting J. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association* **92**, 179-191.

Schwarz G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6**, 461-469.

Skarin A.T. et al (1973). Combination chemotherapy of advanced lymphocytic lymphoma: importance of histologic classification in evaluating response. *Center for a Voluntary Society, Washington D.C.*

Spiegelhalter D.J. and Lauritzen S.L. (1990). Sequential updating of conditional probabilities on directed graphical structures. *Networks* **20**, 579-605.

Spiegelhalter D.J. and Smith A.F.M. (1982). Bayes factors for linear and log-linear models with vague prior information. *Journal of the Royal Statistical Society (Series B)* **44**, 377-387.

Tierney L. and Kadane J.B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the Americal Statistical Association* **81**, 82-86.

Wasserman L. (1997). Bayesian model selection and model averaging. *Mathematical Psychology Symposium.*

Wright S. (1921). Correlation and causation. *Journal of Agricultural Research* **20**, 557-585.

York J., Madigan D., Heuch I. and Lie R.T. (1995). Estimating the proportion of birth defects by double sampling. *Journal of the Royal Statistical Society (Series C)* **44**, 227-242.