# Markov Chain Monte Carlo Model Determination for Hierarchical and Graphical Log-linear Models

Petros Dellaportas[1] and Jonathan J. Forster[2]

SUMMARY

The Bayesian approach to comparing models involves calculating the posterior probability of each plausible model. For high-dimensional contingency tables, the set of plausible models is very large. We focus attention on reversible jump Markov chain Monte Carlo (Green, 1995) and develop strategies for calculating posterior probabilities of hierarchical, graphical or decomposable log-linear models. Even for tables of moderate size, these sets of models may be very large. The choice of suitable prior distributions for model parameters is also discussed in detail, and two examples are presented. For the first example, a $2 \times 3 \times 4$ table, the model probabilities calculated using our reversible jump approach are compared with model probabilities calculated exactly or by using an alternative approximation. The second example is a $2^6$ contingency table for which exact methods are infeasible, due to the large number of possible models.

*Keywords*: Bayesian Analysis; Contingency table; Decomposable model; Hierarchical log-linear model; Graphical model; Markov chain Monte Carlo; Reversible jump

## 1 Introduction

When a number $n$ of individuals are classified according to $C$, a set of categorical variables, or factors, then the data can be represented as the cell counts of a $|C|$-way contingency table. Using the notation of Darroch, Lauritzen and Speed (1980), the table is the set $I = \prod_{\gamma \in C} I_\gamma$ where $I_\gamma$ is the set of levels of factor $\gamma$. An individual cell is denoted by $\boldsymbol{i} = (i_\gamma, \gamma \in C)$, and the corresponding cell count by $n(\boldsymbol{i})$. The total number of cells in the table is $|I| = \prod_{\gamma \in C} |I_\gamma|$.

Typically interest is focussed on relationships between the factors themselves. One way of investigating this is by considering log-linear (interaction) models for the table. The form of such a model and the values of its parameters have implications in terms of independence and conditional independence between the factors, as well as more complex association structures.

A Bayesian approach to model selection proceeds as follows. Suppose that the data, in this case the cell counts $\boldsymbol{n} = (n(\boldsymbol{i}), \boldsymbol{i} \in I)$ are observations of random variables $\boldsymbol{N} = (N(\boldsymbol{i}), \boldsymbol{i} \in I)$ which are considered to have been generated by a model $m$, one of a set $M$ of competing models. Each model specifies the distribution of $\boldsymbol{N}$, $f(\boldsymbol{n}|m, \boldsymbol{\theta}_m)$ apart from an unknown parameter (vector) $\boldsymbol{\theta}_m \in \Theta_m$. Then, the posterior distribution is

$$f(m, \boldsymbol{\theta}_m|\boldsymbol{n}) \quad \propto \quad f(\boldsymbol{n}|m, \boldsymbol{\theta}_m) f(m, \boldsymbol{\theta}_m)$$

where $f(m, \boldsymbol{\theta}_m)$ is the joint prior distribution of $(m, \boldsymbol{\theta}_m)$ which can be factorised as $f(m, \boldsymbol{\theta}_m) = f(\boldsymbol{\theta}_m|m) f(m)$ where $f(m)$ is the prior probability of model $m$. Then the posterior probability of model $m$ is given by

$$f(m|\boldsymbol{n}) \quad = \quad \frac{f(m) \int_{\Theta_m} f(\boldsymbol{n}|m, \boldsymbol{\theta}_m) f(\boldsymbol{\theta}_m|m) \, d\boldsymbol{\theta}_m}{\sum_{m \in M} f(m) \int_{\Theta_m} f(\boldsymbol{n}|m, \boldsymbol{\theta}_m) f(\boldsymbol{\theta}_m|m) \, d\boldsymbol{\theta}_m} \qquad m \in M. \tag{1}$$

[1] Department of Statistics, Athens University of Economics, Greece

[2] Department of Mathematics, University of Southampton, Highfield, Southampton, SO17 1BJ, UK (address for correspondence). email: jjf@maths.soton.ac.uk

In particular, the relative probability of two competing models $m_1$ and $m_2$ reduces to

$$\frac{f(m_1|\boldsymbol{n})}{f(m_2|\boldsymbol{n})} = \frac{f(m_1)}{f(m_2)} \quad \frac{\int_{\Theta_{m_1}} f(\boldsymbol{n}|m_1,\boldsymbol{\theta}_{m_1}) f(\boldsymbol{\theta}_{m_1}|m_1)\, d\boldsymbol{\theta}_{m_1}}{\int_{\Theta_{m_2}} f(\boldsymbol{n}|m_2,\boldsymbol{\theta}_{m_2}) f(\boldsymbol{\theta}_{m_2}|m_2)\, d\boldsymbol{\theta}_{m_2}} \tag{2}$$

which is the familiar expression relating the posterior and prior odds of two models in terms of the Bayes factor, the second ratio on the right hand side of (2).

Madigan and Raftery (1994) discuss the advantages of a Bayesian approach to model selection for high-dimensional contingency tables, over the commonly used stepwise approaches based on sequentially comparing nested models using approximate asymptotic likelihood ratio tests. In particular, they highlight the problems associated with multiple significance tests, comparing non-nested models and conditioning on a single selected model. In principle, the approach outlined above overcomes these problems. Expression (1) allows the calculation of posterior probabilities of all competing models, regardless of their relative size or structure, and this model uncertainty can be incorporated into any decisions or predictions required (Draper, 1995, gives examples of this).

The problems with this approach are associated with the computation of $f(m|\boldsymbol{n})$ using (1), which requires calculation of the marginal likelihood $f(\boldsymbol{n}|m) = \int_{\Theta_m} f(\boldsymbol{n}|m,\boldsymbol{\theta}_m) f(\boldsymbol{\theta}_m|m) d\boldsymbol{\theta}_m$ for each $m \in M$. This integral is only analytically tractable in certain restricted examples. Kass and Raftery (1995) review a series of approximations, both analytic and Monte Carlo for marginal likelihoods, when direct calculation is not available. However, a further problem may arise if the size of the set of possible models $M$ becomes so great that calculation or approximation of $f(\boldsymbol{n}|m)$ for all $m \in M$ becomes infeasible. Then, Monte Carlo methods for generating from $f(m,\boldsymbol{\theta}_m|\boldsymbol{n})$ become an extremely attractive alternative. If a sample $(m^{(t)},\boldsymbol{\theta}^{(t)},\ t = 1,\ldots,s)$ can be generated from this distribution, then posterior model probabilities can be approximated directly by

$$\hat{f}(m) = \frac{1}{s}\sum_{t=1}^{s} 1(m^{(t)} = m) \qquad m \in M \tag{3}$$

where $1(\cdot)$ is the indicator function. Samples from $f(\boldsymbol{\theta}_m|m,\boldsymbol{n})$ are also automatically available for marginal parametric inference. In practice, all suggested methods for generating from $f(m,\boldsymbol{\theta}_m|\boldsymbol{n})$ are based on Markov chains. Green (1995) gives a review of some of these methods.

We will consider the problem of model choice for multiway contingency tables, concentrating on hierarchical log-linear models and graphical models. Both of the problems discussed in the previous paragraph are associated with these classes of models, namely unavailability of $f(\boldsymbol{n}|m)$ and the large size of $M$. In the particular situation of a two-way contingency table, with just two models, saturated and row-column independence, under consideration, a number of possible approaches for comparing the two models have been suggested. Albert (1990) compares several alternative approaches including those advocated by Gûnel and Dickey (1974), Good and Crook (1987) and Spiegelhalter and Smith (1982). However, of these, only Spiegelhalter and Smith (1982) offer a specific solution to the calculation of posterior model probabilities for general log-linear models. Raftery (1993) provides improved approximations for Bayes factors for comparing generalised linear models, of which log-linear models are a special case. Therefore, at least in principle, accurate approximations to posterior model probabilities are available. However, there then arises the second problem, that $M$ is, in general, so large that calculation of such approximations for all $m \in M$ is prohibitive. We propose Markov chain Monte Carlo methods as an efficient way of overcoming these difficulties.

## 2 The Log-linear Interaction Model

A log-linear (interaction) model assumes that the $N(\boldsymbol{i})$ are observations of independent Poisson random variables with $E(\boldsymbol{N}) = \boldsymbol{\mu} = (\mu(\boldsymbol{i}), \boldsymbol{i} \in I)$. Then (again, following Darroch, Lauritzen and Speed, 1980)

$$\log \mu(\boldsymbol{i}) \;=\; \sum_{a \subseteq C} \xi_a(\boldsymbol{i}_a) \qquad \boldsymbol{i} \in I \tag{4}$$

where $\boldsymbol{i}_a$ is the marginal cell $\boldsymbol{i}_a = (i_\gamma, \gamma \in a)$. Note that in practice, to ensure identifiability of the $\xi_a(\boldsymbol{i}_a)$ terms, constraints are usually imposed. A general log-linear interaction model is specified by setting certain $\xi_a$ (interaction) terms to be identically zero, with the remaining terms arbitrary and unknown. Therefore a model is determined by the set of non-zero $\xi_a$ terms. Clearly, as there are $2^{|C|}$ possible $a \subseteq C$ then there are $2^{(2^{|C|})}$ possible models in $M_L$, the set of general log-linear interaction models. If the total number of individuals, $n$, is fixed in advance, then the distribution of $\boldsymbol{N}$ is assumed to be multinomial, with corresponding vector of cell probabilities $\boldsymbol{p}$. A log linear model for $\boldsymbol{p}$ takes the same form as (4) except that $\xi_\emptyset$ is a normalising constant to ensure that the cell probabilities sum to one.

### 2.1 Hierarchical Models

In practice, there are difficulties in interpreting general log-linear interaction models, and a smaller class of models $M_H$, the hierarchical log-linear models, is preferred. Here the restriction is imposed that if $\xi_a$ is set to zero then $\xi_b$ must also be set to zero for all $b \supseteq a$. There is no general expression for the total number of possible hierarchical log-linear models for a $|C|$-way table. The number is much less than $2^{(2^{|C|})}$, but may still be very large ($\approx 5.6 \times 10^{22}$ for an eight-way table). A hierarchical model $m \in M_H$ is determined by its generators, the maximal sets $a$ such that $\xi_a$ is non-zero.

### 2.2 Graphical Models

A smaller class of models, each member of which has a very straightforward interpretation, are the graphical models. In this paper, we will only consider undirected graphical models. For details of Bayesian model selection for directed graphical models see Madigan *et al* (1995). An (undirected) graphical model is determined by a set of conditional independence constraints of the form '$\gamma_1$ is independent of $\gamma_2$ conditional on all other $\gamma_i \in C$'. Graphical models are so called because they can each be represented as a graph with vertex set $C$ and an edge between each pair $\gamma_1$ and $\gamma_2$ unless $\gamma_1$ and $\gamma_2$ are conditionally independent as described above. Darroch, Lauritzen and Speed (1980) show that each graphical log-linear model is hierarchical, with generators given by the cliques (complete subgraphs) of the graph.

The total number of possible graphical models is clearly given by $2^{\binom{|C|}{2}}$ as there are $\binom{|C|}{2}$ possible edges in the graph. This assumes that the 'intercept' $\xi_\emptyset$ and all 'main effects', $\xi_a$ terms where $|a| = 1$, are present in every model. This is an assumption which we make throughout this paper, as log-linear models without such terms are not usually of interest. However, our methods are still applicable for models without main effects. There are far fewer graphical models than hierarchical models ($\approx 2.7 \times 10^9$ for an eight-way table).

Although graphical models have the property of being easily interpretable, directly in terms of conditional independence properties, there is still a strong case to be made for considering the complete set of hierarchical models, as they allow more parsimonious association structures to be modelled. For example, a three-factor clique may exhibit 'conditionally homogeneous association' which is expressible as a hierarchical model by setting the three-factor interaction to zero.

## 2.3 Decomposable Models

A hierarchical log-linear model is decomposable if and only if the joint cell probabilities (means) may be expressed explicitly in terms of the marginal probabilities, for the margins corresponding to its generators. Darroch, Lauritzen and Speed (1980) show that decomposable models are necessarily graphical. When analysis is restricted to purely decomposable models then computation is much more straightforward than for general graphical or hierarchical models.

Dawid and Lauritzen (1993) provide an explicit expression for the Bayes factor for comparing two decomposable models which differ by the presence of a single edge. Hence, in principle, all posterior model probabilities (1) are straightforward to calculate. However, for large tables, the number of models may prohibit this calculation. Occam's window (Madigan and Raftery, 1994) and Markov Chain Monte Carlo model composition ($MC^3$, Madigan and York, 1995) are strategies for overcoming this problem. A comparison of the two approaches is provided by Madigan *et al* (1994).

There are, however, interesting models which fall outside the class of decomposable models $M_D$. Indeed, the restriction that a model should be decomposable is often motivated purely by computational constraints, and may not be justified by modelling considerations.

## 2.4 Parameterisation

In order to consider the model selection problem for log-linear models, we need to consider the general log-linear model (4) within the framework of Section 1. It will be helpful to work with a parameterisation for the general log-linear model where the parameters are identifiable and linearly independent. If $\boldsymbol{\zeta}_a = (\zeta_a(\boldsymbol{i}), \boldsymbol{i} \in I)$ is the $|I|$-dimensional vector where $\xi_a(\boldsymbol{i}_a)$ is replicated so that $\zeta_a(\boldsymbol{i}) = \zeta_a(\boldsymbol{i}_a, \boldsymbol{i}_{C\setminus a}) = \xi_a(\boldsymbol{i}_a)$ for all $\boldsymbol{i}_{C\setminus a}$, then from (4),

$$\log \boldsymbol{\mu} \;=\; \sum_{a \subseteq C} \boldsymbol{\zeta}_a. \tag{5}$$

Following Knuiman and Speed (1988), we set

$$\boldsymbol{\zeta}_a \;=\; \boldsymbol{T}_a \log \boldsymbol{\mu} \qquad a \subseteq C \tag{6}$$

where the $\boldsymbol{T}_a$ are projection matrices given by

$$\boldsymbol{T}_a \;=\; \bigotimes_{\gamma \in C} \left\{ 1(\gamma \in a) \left( \boldsymbol{I}_{|I_\gamma|} - \frac{1}{|I_\gamma|} \boldsymbol{J}_{|I_\gamma|} \right) \;+\; 1(\gamma \notin a) \frac{1}{|I_\gamma|} \boldsymbol{J}_{|I_\gamma|} \right\} \qquad a \subseteq C \tag{7}$$

where $\boldsymbol{I}_{|I_\gamma|}$ is the $|I_\gamma| \times |I_\gamma|$ identity matrix and $\boldsymbol{J}_{|I_\gamma|}$ is a $|I_\gamma| \times |I_\gamma|$ matrix with all elements equal to 1. Then, provided that the cell means $\mu(\boldsymbol{i})$ are arranged within $\boldsymbol{\mu}$ in the standard way, $\boldsymbol{\zeta}_a$ satisfies (5) and the elements of $\boldsymbol{\zeta}_a$ are the usual identifiable log-linear parameters satisfying 'sum-to-zero' constraints. Knuiman and Speed (1988) present the $\boldsymbol{T}_a$ matrices for a $2 \times 3 \times 4$ table.

A set of

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

linearly independent components of $\boldsymbol{\zeta}_a$ for each $a \subseteq C$ may then be chosen as the model parameters. We choose $\boldsymbol{\beta}_a = (\xi_a(\boldsymbol{i}_a), i_\gamma < |I_\gamma| \text{ for all } \gamma \in a)$.

# 3 Prior Distributions

We will consider a general model selection problem, and assume that there is no strong prior information to be taken into account. This provides a reference analysis, which can easily be adapted in situations where strong prior beliefs are present. Hence we make the assumption that all competing models are *a priori* equally probable *i.e.* $f(m) = 1/|M|$ for all $m \in M$. Furthermore, we shall assume a vague (non-informative) prior distribution for the model parameters for each model. Such prior distributions must be proper, otherwise (1) is undefined, although possible solutions to this problem do exist (Spiegelhalter and Smith, 1982; O'Hagan, 1995). However, for contingency tables, it is straightforward to construct prior distributions which are both vague and proper. The choice of prior distribution for the model parameters may have a considerable impact on the posterior model probabilities. Therefore, we now proceed to discuss in detail possible choices of vague prior distributions for log-linear models.

## 3.1 Prior distributions based on the Dirichlet distribution

For decomposable log-linear models, the clique marginal cell probabilities provide an explicit parameterisation. Dawid and Lauritzen (1993) propose a hyper-Dirichlet prior distribution for the parameters of a decomposable log-linear model. One easy way of constructing a hyper-Dirichlet prior distribution for a decomposable model is to derive all clique prior distributions as the marginal distributions from a proper Dirichlet prior distribution (all parameters greater than zero) for the complete vector of cell probabilities. This is the approach used by Madigan and Raftery (1994) and Madigan and York (1995), who impose a standard Jeffreys prior, Dirichlet with all parameters equal to $\frac{1}{2}$, $D(\frac{1}{2}\mathbf{1})$, for $\boldsymbol{p}$.

The disadvantage of the Jeffreys prior is that it is not really a vague prior for the margins of the table. An alternative vague symmetric Dirichlet prior, originally advocated by Perks (1947) is $D(\frac{1}{|I|}\mathbf{1})$. This can be interpreted as a single prior observation divided evenly between all cells of the table. The resulting prior for each marginal table is exactly the same as would have arisen if just that marginal table was being analysed.

## 3.2 Normal prior distributions for log-linear model parameters

Consider the log-linear parameters, $\boldsymbol{\beta}_a$, introduced in section 2.4. These parameters are unrestricted, with each $\boldsymbol{\beta}_a$ taking any value in $\mathbb{R}^{d_a}$. Hence a plausible prior distribution is multivariate normal for each $\boldsymbol{\beta}_a$. In the absence of any other prior information, it is reasonable to assume that different $\boldsymbol{\beta}_a$ terms are *a priori* independent.

The independent prior distributions for the $\boldsymbol{\beta}_a$ parameters induce a prior distribution for $\log \boldsymbol{\mu}$. We propose the prior distribution

$$\log \boldsymbol{\mu} \;\sim\; N\left(\theta\mathbf{1}, \sum_{a \subseteq C} \alpha_a^2 \boldsymbol{T}_a\right) \tag{8}$$

where the $\boldsymbol{T}_a$ are given by (7). This prior distribution is invariant to arbitrary permutations of the levels $I_\gamma$ of each factor $\gamma$. For a non-informative prior distribution, this seems to be a sensible requirement. The prior distributions for the $\boldsymbol{\beta}_a$ parameters, with the exception of $\boldsymbol{\beta}_\emptyset$ are then given by

$$\boldsymbol{\beta}_a \overset{ind}{\sim} N(\mathbf{0}, \alpha_a^2 \boldsymbol{\Sigma}_a) \qquad a \subseteq C \tag{9}$$

where

$$\boldsymbol{\Sigma}_a \;= \frac{1}{|I|} \prod_{\gamma \in a} |I_\gamma| \bigotimes_{\gamma \in a} \left( \boldsymbol{I}_{(|I_\gamma|-1)} - \frac{1}{|I_\gamma|} \boldsymbol{J}_{(|I_\gamma|-1)} \right) \qquad a \subseteq C. \tag{10}$$

For $\boldsymbol{\beta}_\emptyset$, the $\mathbf{0}$ in (9) is replaced by $\theta\mathbf{1}$, and so for this term both a prior mean and prior dispersion parameter need to be specified. For all other $\boldsymbol{\beta}_a$, we only need to specify the prior dispersion parameter $\alpha_a^2$. Although in practice it is possible to allow different values of $\alpha_a^2$ for different models in which $\boldsymbol{\beta}_a$ appears, there seems no reason to allow such an incompatibility between models unless relevant prior information exists.

Knuiman and Speed (1988) use this prior distribution, with the exception that they require a non-zero mean for a particular log-linear model term about which they have considerable prior information. Hence, while a prior of the form (9) is desirable in situations in which prior information is vague, it may be necessary to modify it in particular examples. However, the methods described in this paper are applicable for any multivariate normal prior for $\log \boldsymbol{\mu}$.

### 3.3 Choosing prior parameters

The question remains of how to choose $\alpha_a^2$ to reflect vague prior belief. The absence of a scale parameter for Poisson or multinomial data means that this task is quite straightforward. Clearly, values of $\alpha_a^2$ close to zero represent strong prior belief that $\boldsymbol{\beta}_a$ is close to $\mathbf{0}$. Conversely, as $\alpha_a^2 \to \infty$, the prior becomes very vague, and in the limit $\alpha_a^{-2} = 0$, the prior is the improper uniform prior over $\mathbb{R}^{d_a}$. It is however possible to choose a proper vague prior using a large but finite value of $\alpha_a^2$.

Consider the conjugate prior distribution for $\boldsymbol{\mu}$ such that components $\mu(\boldsymbol{i})$ have independent $Ga(\alpha(\boldsymbol{i}), \beta(\boldsymbol{i}))$ distributions. Invariance to arbitrary permutations of the levels $I_\gamma$ of each factor $\gamma$, for this prior distribution requires that $\alpha(\boldsymbol{i}) = \alpha$ and $\beta(\boldsymbol{i}) = \beta$ for all $\boldsymbol{i}$. Then,

$$\begin{aligned} E(\log \boldsymbol{\mu}) &= (\psi(\alpha) - \log \beta)\mathbf{1} \\ Var(\log \boldsymbol{\mu}) &= \psi'(\alpha)\boldsymbol{I} \end{aligned}$$

where $\psi$ and $\psi'$ are the digamma and trigamma functions, and hence

$$\begin{aligned} E(\boldsymbol{T}_a \log \boldsymbol{\mu}) &= \mathbf{0} && a \subseteq C;\ a \neq \emptyset \\ E(\boldsymbol{T}_\emptyset \log \boldsymbol{\mu}) &= (\psi(\alpha) - \log \beta)\mathbf{1} \\ Var(\boldsymbol{T}_a \log \boldsymbol{\mu}) &= \psi'(\alpha)\boldsymbol{T}_a && a \subseteq C. \end{aligned}$$

Therefore, by setting $\alpha_a^2 = \psi'(\alpha)$ for all $a \subseteq C$ and $\theta = \psi(\alpha) - \log \beta$ in (8), $\log \boldsymbol{\mu}$ has the same prior mean and variance structure as for the conjugate gamma prior distribution where each component $\mu(\boldsymbol{i})$ of $\boldsymbol{\mu}$ has an independent $Ga(\alpha, \beta)$ distribution.

The relationship between the Dirichlet and gamma distributions now allows us to choose values for the $\alpha_a^2$ parameters in terms of a 'prior sample', as $\alpha$ may be thought of as the size of a 'prior cell count' in each cell. For example, the Jeffreys prior ($\alpha = 1/2$) corresponds to the normal prior with $\alpha_a^2 = \pi^2/2 \approx 4.935$ for all $a \subseteq C$ whereas Perks' prior ($\alpha = 1/|I|$) corresponds to the normal prior with $\alpha_a^2 = \psi'\left(\frac{1}{|I|}\right)$, ($\approx \pi^2/6 + |I|^2$ when $|I|$ is large) for all $a \subseteq C$.

The normal prior distribution has a similar form to that suggested by Raftery (1993) for generalised linear models. For a normal regression model with standardised covariate vectors, he suggests giving all model parameters, except the intercept, independent normal prior distributions with zero mean and common variance. For generalised linear models this prior is then transformed in a way which depends on the actual observed cell counts. If the prior expected cell counts (all equal) are used instead then, for a table in which all the factors have two levels, the resulting prior for all parameters other than $\boldsymbol{\beta}_\emptyset$ has exactly the form we have advocated, with Raftery's prior parameter $\phi$ corresponding to our $\alpha/\sqrt{|I|}$. He gives the 'intercept' term

(our $\boldsymbol{\beta}_\emptyset$) a further independent normal prior distribution, which may have a different mean and variance. For our prior, $\boldsymbol{\beta}_\emptyset$ has the same variance but the presence of $\log\beta$ in $E(\boldsymbol{T}_\emptyset \log \boldsymbol{\mu})$ allows an arbitrary non-zero mean, $\theta$ in (8), to be chosen.

For the dispersion parameter for the model terms of interest, Raftery's suggested range of values which do not provide strong evidence either in favour of or against including a term in the model imply a value for each $\alpha_a^2$ of between 1 and 25 times the number of cells in the table when all factors have two levels. This is not, in general, compatible with either the value of $\alpha_a^2 = \pi^2/2$ suggested earlier as being equivalent to the Jeffreys prior, which may favour over-complex models, or the higher values equivalent to Perks' (1947) prior which may unnecessarily favour more parsimonious models. It is, however, compatible with a value of $\alpha_a^2$ equal to twice the number of cells in the table.

We advocate the value $\alpha_a^2 = 2|I|$ for the following reasons. The prior distribution for a model parameter is of greatest importance in a model selection problem when the corresponding model term is present in one model but not in the other. This is most likely to be the case for a log-linear model parameter which is amongst the generators for the model. A parameter which is a generator for a model has the same interpretation, as a log-contrast of conditional (on a combination of levels of the other factors not present in the term under consideration) cell means, regardless of which other factors are present in the table. However, the prior distribution (9) may depend on the other factors, because of the $1/|I|$ term in (10). Hence, to overcome this, it seems desirable that the $\alpha_a^2$ parameters should each be proportional to $|I|$, the number of cells in the table. If we now consider the simplest situation, a two-cell contingency table, the Jeffreys' and Perks' priors coincide. They are both symmetric beta distributions, with parameter equal to one half. There is only one model parameter $\boldsymbol{\beta}_a$ other than the intercept, and the resulting prior on this parameter is a log–square root–beta type *II* distribution, with mean zero and variance $\pi^2/4$. As argued above, a value of $\alpha_a^2 = \pi^2/2$ for our lognormal prior ensures that the prior distributions have the same mean and variance (and also that the Kullback-Liebler distance between them is minimised). However a value of $\alpha_a^2 = \pi$ ensures that the two distributions have the same prior ordinate at zero. We choose a compromise value of $\alpha_a^2 = 4$. Hence, if $\alpha_a^2 \propto |I|$ in all examples, then the above reasoning suggests that 2 is a sensible value for the constant of proportionality.

# 4 Markov Chain Monte Carlo

The utility of Bayesian methods for model determination, and of Markov Chain Monte Carlo (MCMC) procedures which bypass many of the associated computational difficulties, has been widely recognised recently. For large multiway contingency tables, one possible approach is MC$^3$ (Madigan and York, 1995). For undirected graphical models MC$^3$ uses the Metropolis-Hastings algorithm to construct a Markov chain, for which the state space is the set of decomposable models, and the equilibrium distribution is $f(m|\boldsymbol{n})$, the posterior model probabilities assuming that only decomposable models have non-zero prior probability. This is straightforward when all models under consideration are decomposable and the prior distributions for the cell probabilities under each model are hyper-Dirichlet, as Dawid and Lauritzen (1993) show how the relative posterior probabilities of decomposable models which differ by a single edge may be calculated explicitly.

However, there are many interesting models for multiway contingency tables which do not fall into the class of decomposable models, and we aim to develop a method for calculating posterior model probabilities for any set of log-linear models, and in particular for the hierarchical and graphical models discussed in

Section 2. Due to the large number, $|M|$, of models under consideration, even for tables with moderate numbers of cells, a Markov chain approach is particularly appealing, as it avoids the necessity for evaluation of model probabilities for all models. Furthermore, the success of Markov chain methods for evaluating posterior quantities of interest in a wide range of examples for which analytic results are not available also makes such methods attractive. For example, MCMC methods have proved effective for the efficient calculation of posterior quantities of interest in contingency table problems with just a single model under consideration. See Epstein and Fienberg (1991) and Forster and Skene (1994) for details.

## 4.1 Reversible jump MCMC

Green (1995) considered the general model determination problem introduced in Section 1. He developed a general Markov chain Monte Carlo strategy for generating from $f(m, \boldsymbol{\theta}_m|\boldsymbol{n})$, based on the standard Metropolis-Hastings approach, but which can also deal with the situation where the dimensionality of the parameter vector $\boldsymbol{\theta}_m$ is not fixed and so there is no simple underlying measure. The 'reversible jump' sampler allows moves between parameter subspaces of different dimensionality by, at each step, allowing a series of different 'move types'. The general approach is extremely flexible, but Green (1995) also gives details of a particular implementation which proves successful when adapted for log-linear models.

This implementation proceeds as follows. We represent the current state of the Markov chain (at time $t$) as $(m^{(t)}, \boldsymbol{\theta}^{(t)}_{m^{(t)}})$. Suppose that a particular move type $p$ involves a proposed move to model $m'$ and corresponding parameter vector $\boldsymbol{\theta}'_{m'}$ which is of a higher dimensionality than $\boldsymbol{\theta}^{(t)}_{m^{(t)}}$. Furthermore, suppose that $\boldsymbol{\theta}'_{m'}$ is created by generating a vector $\boldsymbol{u}$ of dimensionality equal to the difference in dimensionalities between $\boldsymbol{\theta}^{(t)}_{m^{(t)}}$ and $\boldsymbol{\theta}'_{m'}$ from some proposal distribution $q_p(\boldsymbol{u})$, and setting $\boldsymbol{\theta}'_{m'} = g(\boldsymbol{\theta}^{(t)}_{m^{(t)}}, \boldsymbol{u})$ where $g$ is a one-to-one function. Similarly, if the move type $p$ is used 'in the other direction' and $\boldsymbol{\theta}^{(t)}_{m^{(t)}}$ is the parameter vector of higher dimension then $\boldsymbol{\theta}'_{m'}$ is created from $\boldsymbol{\theta}^{(t)}_{m^{(t)}}$ by applying the inverse transformation $g^{-1}$ $(\boldsymbol{\theta}'_{m'}, \boldsymbol{u}') = g^{-1}(\boldsymbol{\theta}^{(t)}_{m^{(t)}})$ and ignoring $\boldsymbol{u}'$.

Suppose that the probability of making move type $p$ given the current state of the Markov chain $(m^{(t)}, \boldsymbol{\theta}_{m^{(t)}})$ is $j(p, m^{(t)}, \boldsymbol{\theta}^{(t)}_{m^{(t)}})$. Green (1995; equation 8) shows that if the proposed move $p$ involves an increase in dimensionality from $\boldsymbol{\theta}^{(t)}_{m^{(t)}}$ to $\boldsymbol{\theta}'_{m'}$, generated *via* the appropriate $q_p(\boldsymbol{u})$, it should be accepted as the next realisation of the chain (so that $m^{(t+1)} = m'$) with probability $\alpha(m^{(t)}, \boldsymbol{\theta}^{(t)}_{m^{(t)}}, m', \boldsymbol{\theta}'_{m'})$ where

$$\alpha(m^{(t)}, \boldsymbol{\theta}^{(t)}_{m^{(t)}}, m', \boldsymbol{\theta}'_{m'}) = \min\left\{1, \frac{f(m', \boldsymbol{\theta}'_{m'}|\boldsymbol{n}) j(p, m', \boldsymbol{\theta}'_{m'})}{f(m^{(t)}, \boldsymbol{\theta}^{(t)}_{m^{(t)}}|\boldsymbol{n}) j(p, m^{(t)}, \boldsymbol{\theta}^{(t)}_{m^{(t)}}) q_p(\boldsymbol{u})} \left|\frac{\partial(\boldsymbol{\theta}_{m'})}{\partial(\boldsymbol{\theta}_{m^{(t)}}, \boldsymbol{u})}\right|\right\} \tag{11}$$

and rejected otherwise ($m^{(t+1)} = m^{(t)}$). Similarly, if the proposed move $p$ involves a decrease in dimensionality from $\boldsymbol{\theta}^{(t)}_{m^{(t)}}$ to $\boldsymbol{\theta}'_{m'}$, it should be accepted as the next realisation of the chain with probability $\alpha(m^{(t)}, \boldsymbol{\theta}^{(t)}_{m^{(t)}}, m', \boldsymbol{\theta}'_{m'})$ where

$$\alpha(m^{(t)}, \boldsymbol{\theta}^{(t)}_{m^{(t)}}, m', \boldsymbol{\theta}'_{m'}) = \min\left\{1, \frac{f(m', \boldsymbol{\theta}'_{m'}|\boldsymbol{n}) j(p, m', \boldsymbol{\theta}'_{m'}) q_p(\boldsymbol{u}')}{f(m^{(t)}, \boldsymbol{\theta}^{(t)}_{m^{(t)}}|\boldsymbol{n}) j(p, m^{(t)}, \boldsymbol{\theta}^{(t)}_{m^{(t)}})} \left|\frac{\partial(\boldsymbol{\theta}_{m'}, \boldsymbol{u}')}{\partial(\boldsymbol{\theta}_{m^{(t)}})}\right|\right\} \tag{12}$$

and rejected otherwise. Here $\boldsymbol{u}'$ is calculated from $(\boldsymbol{\theta}'_{m'}, \boldsymbol{u}') = g^{-1}(\boldsymbol{\theta}^{(t)}_{m^{(t)}})$.

Green (1995) shows that if moves are accepted with probabilities given by (11) and (12), then the chain satisfies detailed balance, and has the required limiting distribution $f(m, \boldsymbol{\theta}_m|\boldsymbol{n})$. In addition to move types which involve changing the model, together with a corresponding increase or decrease in dimensionality of the parameter space, there may also exist the 'null' move, whereby the model remains the same ($m^{(t+1)} = m^{(t)}$) although values of the parameters may be changed.

If, for moves that involve an increase in dimensionality of the parameter space, $\boldsymbol{\theta}'_{m'} = (\boldsymbol{\theta}^{(t)}_{m^{(t)}}, \boldsymbol{u})$ directly and correspondingly, for moves that involve a decrease in dimensionality, $\boldsymbol{\theta}'_{m'} = (\boldsymbol{\theta}^{(t)}_{m'})$ with the excess parameters vanishing ($\boldsymbol{u}' = \boldsymbol{\theta}^{(t)}_{m^{(t)} \setminus m'}$) then the Jacobian term in (11) and (12) is equal to one. Therefore we have acceptance probability

$$\alpha(m^{(t)}, \boldsymbol{\theta}^{(t)}_{m^{(t)}}, m', \boldsymbol{\theta}'_{m'}) = \min\left\{1, \frac{f(\boldsymbol{n}|m', \boldsymbol{\theta}'_{m'}) f(\boldsymbol{\theta}'_{m'}|m') f(m') j(p, m', \boldsymbol{\theta}'_{m'})}{f(\boldsymbol{n}|m^{(t)}, \boldsymbol{\theta}^{(t)}_{m^{(t)}}) f(\boldsymbol{\theta}^{(t)}_{m^{(t)}}|m^{(t)}) f(m^{(t)}) j(p, m^{(t)}, \boldsymbol{\theta}^{(t)}_{m^{(t)}}) q_p(\boldsymbol{u})}\right\} \quad (13)$$

if $p$ involves an increase in dimensionality, and

$$\alpha(m^{(t)}, \boldsymbol{\theta}^{(t)}_{m^{(t)}}, m', \boldsymbol{\theta}'_{m'}) = \min\left\{1, \frac{f(\boldsymbol{n}|m', \boldsymbol{\theta}'_{m'}) f(\boldsymbol{\theta}'_{m'}|m') f(m') j(p, m', \boldsymbol{\theta}'_{m'}) q_p(\boldsymbol{\theta}^{(t)}_{m^{(t)} \setminus m'})}{f(\boldsymbol{n}|m^{(t)}, \boldsymbol{\theta}^{(t)}_{m^{(t)}}) f(\boldsymbol{\theta}^{(t)}_{m^{(t)}}|m^{(t)}) f(m^{(t)}) j(p, m^{(t)}, \boldsymbol{\theta}^{(t)}_{m^{(t)}})}\right\} \quad (14)$$

if $p$ involves a decrease in dimensionality, where the posterior densities $f(m, \boldsymbol{\theta}_m|\boldsymbol{n})$ in (11) and (12) have been replaced by the appropriate product of prior density and likelihood.

For each proposed move type $p$ we need to specify the probability $j(p, m^{(t)}, \boldsymbol{\theta}^{(t)}_{m^{(t)}})$ of attempting a move of that type, for every possible current state of the chain $(m^{(t)}, \boldsymbol{\theta}^{(t)}_{m^{(t)}})$. We also need to specify the corresponding proposal distribution $q_p(\boldsymbol{u})$ when $p$ involves an increase in dimensionality of the parameter space. Therefore, even though this is a particular implementation of the reversible jump algorithm, it is still extremely flexible.

## 4.2 Reversible jump MCMC for log-linear models

Suppose that $M = M_L$, the set of general log-linear interaction models, so $|M| = 2^{(2^{|C|})}$. The model parameters $\boldsymbol{\theta}_m$ for log-linear model $m$ are those $\boldsymbol{\beta}_a$ which are non-zero for $m$. One possible implementation of reversible jump MCMC, of the kind discussed above, involves considering each of the $2^{|C|}$ model terms $a \subseteq C$ as possible move types. Therefore, if $p = a$ and the term $a$ is not currently present in $m^{(t)}$, then a proposal $\boldsymbol{u}$ for $\boldsymbol{\beta}_a$ is generated using some $q_p(\boldsymbol{u})$ and the move is accepted with probability $\alpha(m^{(t)}, \boldsymbol{\theta}^{(t)}_{m^{(t)}}, m', \boldsymbol{\theta}'_{m'})$ calculated using (13). If the term is already present in $m^{(t)}$, then the proposed move removes it ($\boldsymbol{\beta}_a$ is set to $\mathbf{0}$) and the move is accepted with probability $\alpha(m^{(t)}, \boldsymbol{\theta}^{(t)}_{m^{(t)}}, m', \boldsymbol{\theta}'_{m'})$ calculated using (14). Hence $\boldsymbol{\theta}_{m'} = (\boldsymbol{\theta}_{m^{(t)}}, \boldsymbol{\beta}_a)$ if the move involves adding term $a$ to the model and $\boldsymbol{\theta}_{m^{(t)}} = (\boldsymbol{\theta}_{m'}, \boldsymbol{\beta}_a)$ otherwise. The Markov chain may be thought of as continuously adding terms (and their corresponding parameters) to, or eliminating terms from, the model.

If each move type, other than the null, is made with equal probability, then

$$j(p, m^{(t)}, \boldsymbol{\theta}^{(t)}_{m^{(t)}}) = \frac{1 - r}{2^{|C|}} \qquad p \subseteq C \quad (15)$$

where $r$ is the probability of proposing a null move. In this situation, each move type has the same probability, regardless of the current state, so $j(p, m^{(t)}, \boldsymbol{\theta}^{(t)}_{m^{(t)}})$ and $j(p, m', \boldsymbol{\theta}'_{m'})$ are always equal, and hence cancel in (13) and (14).

The availability of the null move is highly desirable, if the Markov chain which adds and eliminates terms as discussed above is to be effective at providing a sample from $f(m, \boldsymbol{\theta}_m|\boldsymbol{n})$. Without the null move, the only mechanism for change in a particular element of $\boldsymbol{\theta}_m$ is *via* elimination and subsequent addition of the appropriate model term. This occurs very rarely for terms which have a very high posterior probability of being present in a model, and hence the Markov chain is not very mobile without a null move.

We have not specified how the parameter values for the current model might be changed when the null move is implemented. One possibility is to use a 'Gibbs sampler step'. Dellaportas and Smith (1993) show

how the Gibbs sampler may be implemented for a large class of generalised linear models, of which the log linear model with the multivariate normal prior on the model parameters, discussed in Section 3, is a special case. They showed that the univariate conditional distributions required for sampling are log-concave, and hence adaptive rejection sampling (Gilks and Wild, 1992) is straightforward to implement. Forster and Skene (1994) used a Gibbs sampler to generate from posterior distributions arising from multinomial likelihoods and logistic normal prior distributions, which corresponds to the situation in the present example. They showed that the Gibbs sampler is extremely efficient when a particular parameterisation, where the parameters are approximately *a posteriori* independent, is adopted. They recommended the use of the Gibbs sampler over a number of possible implementations of the Metropolis-Hastings algorithm, for multinomial applications. Although, in the current situation where a large numbers of different models are under consideration, it is impractical to use the parameterisation suggested by Forster and Skene (1994), the Gibbs sampler is still an efficient approach for generating from the posterior distribution of the parameters for a particular log-linear model. We therefore choose to use a Gibbs sampler for the null move.

To complete specification of our reversible jump algorithm for log-linear models, it is necessary to specify the density $q_p(\boldsymbol{u})$, by which model parameters are generated when the move type $p$ involves the addition of a model term. When $p = a \subseteq C$, the model parameters are the $d_a$ components of $\boldsymbol{\beta}_a$ introduced in section 2.4. Therefore $\boldsymbol{u}$ is $d_a$-dimensional, and as $\boldsymbol{u} \in I\!R^{d_a}$, a flexible family of proposal distributions $q_a(\boldsymbol{u})$ is the multivariate normal family. Therefore

$$q_a(\boldsymbol{u}) \;=\; |\boldsymbol{\Lambda}_a|^{-\frac{1}{2}} \exp\left(-\tfrac{1}{2}(\boldsymbol{u}-\boldsymbol{\psi}_a)^T \boldsymbol{\Lambda}_a^{-1}(\boldsymbol{u}-\boldsymbol{\psi}_a)\right) \qquad a \subseteq C \tag{16}$$

and so, from (9), (13), (15) and (16), the probability of accepting the proposed move $p = a$ is

$$\alpha(m^{(t)}, \boldsymbol{\theta}^{(t)}_{m^{(t)}}, m', \boldsymbol{\theta}'_{m'}) \;=\; \min\left\{1, \frac{f(\boldsymbol{n}|m', \boldsymbol{\theta}'_{m'})\alpha_a^{-d_a}|\boldsymbol{\Sigma}_a|^{-\frac{1}{2}} \exp\left(-\tfrac{1}{2}\alpha_a^{-2}\boldsymbol{u}^T\boldsymbol{\Sigma}_a^{-1}\boldsymbol{u}\right)}{f(\boldsymbol{n}|m^{(t)}, \boldsymbol{\theta}^{(t)}_{m^{(t)}})|\boldsymbol{\Lambda}_a|^{-\frac{1}{2}} \exp\left(-\tfrac{1}{2}(\boldsymbol{u}-\boldsymbol{\psi}_a)^T\boldsymbol{\Lambda}_a^{-1}(\boldsymbol{u}-\boldsymbol{\psi}_a)\right)}\right\} \tag{17}$$

when the move involves adding a term to the current model. The contribution of the prior distribution to (17) is through the prior distribution of $\boldsymbol{\beta}_a$, as all other model parameters take the same value in $m^{(t)}$ and $m'$ and so their contribution is cancelled out. The prior model probabilities $f(m^{(t)})$ and $f(m')$ have also been cancelled, as we are assuming that all models are equally probable *a priori*. If a proposed move involves eliminating a model term, then the acceptance probability is

$$\alpha(m^{(t)}, \boldsymbol{\theta}^{(t)}_{m^{(t)}}, m', \boldsymbol{\theta}'_{m'}) \;=\; \min\left\{1, \frac{f(\boldsymbol{n}|m', \boldsymbol{\theta}'_{m'})|\boldsymbol{\Lambda}_a|^{-\frac{1}{2}} \exp\left(-\tfrac{1}{2}(\boldsymbol{\theta}^{(t)}_{m^{(t)}\setminus m'} - \boldsymbol{\psi}_a)^T\boldsymbol{\Lambda}_a^{-1}(\boldsymbol{\theta}^{(t)}_{m^{(t)}\setminus m'} - \boldsymbol{\psi}_a)\right)}{f(\boldsymbol{n}|m^{(t)}, \boldsymbol{\theta}^{(t)}_{m^{(t)}})\alpha_a^{-d_a}|\boldsymbol{\Sigma}_a|^{-\frac{1}{2}} \exp\left(-\tfrac{1}{2}\alpha_a^{-2}(\boldsymbol{\theta}^{(t)}_{m^{(t)}\setminus m'})^T\boldsymbol{\Sigma}_a^{-1}\boldsymbol{\theta}^{(t)}_{m^{(t)}\setminus m'}\right)}\right\}. \tag{18}$$

The functions $f(\boldsymbol{n}|m, \boldsymbol{\theta}_m)$ in (17) and (18) are Poisson likelihoods under the two models, which can be evaluated as

$$f(\boldsymbol{n}|m, \boldsymbol{\theta}_m) \;\propto\; \exp\left(\sum_{a\subseteq C} \boldsymbol{n}^T\boldsymbol{\zeta}_a - \sum_{\boldsymbol{i}\in I} \exp \sum_{a\subseteq C} \zeta_a(\boldsymbol{i}_a)\right) \tag{19}$$

where $\boldsymbol{\zeta}_a$ is determined by $\boldsymbol{\beta}_a$ for each $a$ (and hence by $\boldsymbol{\theta}_m$, see Section 2.4 for details).

The proposal parameters $\boldsymbol{\psi}_a$ and $\boldsymbol{\Lambda}_a$ are arbitrary, and should be chosen to optimise the performance of the procedure. Sensible values may be chosen by investigation of a 'pilot chain'. One such approach is to start a Markov chain at the saturated model (all model terms present) and perform the null Gibbs sampler step a large number of times, in order to obtain approximate marginal posterior moments for the saturated model parameters $(\boldsymbol{\beta}_a, a \subseteq C)$. These moments may then be used as values for $\boldsymbol{\psi}_a$ and $\boldsymbol{\Lambda}_a$ when

the full Markov chain algorithm is implemented. The question remains, of whether the resulting proposal distributions generate sensible values of parameters for models which may have many fewer terms than the saturated model. If they do not, then the likelihood ratio term in (17) will typically prohibit any moves which involve the addition of a term, even if there exist values of $\boldsymbol{\theta}'_{m'}$ for which this ratio would be large. We suggest down-weighting the estimates from the pilot chain, by incorporating a 'prior sample' with sample mean 0, and sample variance proportional to the prior variance, with a constant of proportionality much less than 1. Furthermore, for the variance, for ease of implementation , we choose $\boldsymbol{\Lambda}_a$ to be directly proportional to the prior covariance $\alpha_a^2\boldsymbol{\Sigma}_a$, for all $a$. Therefore, $\boldsymbol{\Lambda}_a = \lambda_a^2\alpha_a^2\boldsymbol{\Sigma}_a$, and we choose $\lambda_a^2$ so that the proposal variance for each component of $\boldsymbol{\beta}_a$ is equal to the smallest sample variance for any component indicated by the pilot chain with prior sample. For this proposal distribution, the acceptance probabilities (17) and (18) are simplified, as $|\boldsymbol{\Lambda}_a|^{-\frac{1}{2}}/\alpha_a^{-d_a}|\boldsymbol{\Sigma}_a|^{-\frac{1}{2}} = \lambda_a^{-d_a}$.

Clearly, these choices of proposal parameters $\boldsymbol{\psi}_a$ and $\boldsymbol{\Lambda}_a$ are somewhat arbitrary, and may easily be changed if required. However, in all applications which we have analysed, using a pilot Gibbs sampler run of 100 iterations and assuming a prior sample of 10 observations with zero mean and variance 0.001 times the prior variance, the process seems to work remarkably well. In each example the Markov chain takes the saturated model as its initial model, with parameter values equal to the final values of the pilot chain. The probability $r$ of the null move is set at 0.25. This Markov chain is, in principle, irreducible, as at any stage each term has a non-zero probability of being added to, or eliminated from, the model. Furthermore, there is also a non-zero probability of any model parameter changing its current value to any other real value.

### 4.3 Reversible jump MCMC for hierarchical log-linear models

The algorithm developed in Section 4.2 assumes that the set of plausible models includes all log-linear interaction models. However, as mentioned in Section 2.1, many of these models are of little practical interest, as they may be difficult to interpret. In practice, interest is usually focussed on the hierarchical log-linear models. Here, we also choose to exclude from the class of models of interest those models for which a main effect (term $a$ where $|a| = 1$) is absent, as such models are also of little practical interest.

If we set the prior probability $f(m)$ of all non-hierarchical models equal to zero, then the Markov chain suggested in Section 4.2 may still be appropriate. The prior model probabilities need to be reintroduced with the second term in the brackets of (17) and (18) being multiplied by $f(m')/f(m^{(t)})$. Then, moves to non-hierarchical models will necessarily have zero probability, and the chain will have equilibrium distribution $f(m, \boldsymbol{\theta}_m)$ where $m \in M = M_H$, the set of hierarchical log-linear models, as any hierarchical model may be obtained from any other by addition and elimination of model terms, with all intermediate models also being hierarchical.

However, this procedure is inefficient, as large number of proposed moves have zero probability of actually being made. A more efficient procedure is to construct $j(p, m^{(t)}, \boldsymbol{\theta}^{(t)}_{m^{(t)}})$ so that only moves to neighbouring hierarchical models are proposed. A hierarchical model is determined by its generators, that is the maximal terms (subsets of $C$) which are present in the model. Clearly, the only individual model terms which may be removed from a hierarchical model so that the model remains hierarchical are the generators. Similarly, Edwards and Havránek (1985) define the dual generators to be the minimal terms which are not present in the model. The only individual model terms which may be added to a hierarchical model so that the model remains hierarchical are the dual generators. Edwards and Havránek (1985) propose a simple algorithm for determining the dual generators of a model from its generators and *vice versa*. If the set of generators and dual generators for model $m$ is denoted by $G_m$, then the total number of allowed move types in addition to

the null move, if the current model is $m^{(t)}$ is $|G_{m^{(t)}}|$. Assuming that each of these proposed move types is equally probable then

$$j(p, m^{(t)}, \boldsymbol{\theta}^{(t)}_{m^{(t)}}) \quad = \quad \frac{1-r}{|G_{m^{(t)}}|} \qquad a \subseteq G_{m^{(t)}}$$

and 0 otherwise, where $r$ is the probability of the null move. It is then necessary to determine $j(p, m', \boldsymbol{\theta}'_{m'})$ in order to calculate the acceptance probability. If the proposed step involves adding a term to the model then the new term is added to the set of generators, any subsets of this term removed, and the new dual generators calculated using the algorithm of Edwards and Havránek (1985). If the proposed step involves removing a term from the model then that term is added to the set of dual generators, any subsets of the term removed, and the new generators again calculated using the algorithm of Edwards and Havránek (1985). Therefore, it is straightforward to find $G_{m'}$, and hence $j(p, m', \boldsymbol{\theta}'_{m'})$ using $|G_{m'}|$. A byproduct of this calculation is that if the move is accepted then $G_{m^{(t+1)}}$ is immediately available.

Again, we assume all hierarchical models are *a priori* equally probable, and as no non-hierarchical models are proposed, $f(m)$ will not be present in the expressions for the acceptance probabilities. However the move probabilities need to be reintroduced, as different numbers of move types are possible from different models, and the second term in the brackets of (17) and (18) must be multiplied by $j(p, m^{(t)}, \boldsymbol{\theta}^{(t)}_{m^{(t)}})/j(p, m', \boldsymbol{\theta}'_{m'})$. The other aspects of the algorithm, such as the proposal density remain exactly the same as in Section 4.2.

## 4.4 Reversible jump MCMC for graphical models

If $M = M_G$, the set of graphical models, then the move types must be changed. It is not sufficient to set the prior probability $f(m)$ of all non-graphical models equal to zero, as the Markov chain suggested in Section 4.3 is not irreducible, as it is not, in general, possible to move from one graphical model to another by addition and elimination of individual model terms, with all intermediate models also being graphical.

As stated in Section 2.2, a graphical model is determined by its edges, and any subset of the $\binom{|C|}{2}$ possible edges determines a graphical model. One possible strategy for constructing a Markov chain which moves around $M_G$ is, at each stage, to consider the addition of an edge to, or elimination of an edge from, the current model. This strategy was adopted by Madigan and York (1995), in their algorithm for decomposable models. However they also needed to check whether the resulting model was still decomposable. In the current situation, such move types depend not only on the edge $e$ under consideration, but also on the current model $m$, as addition, or elimination, of the same edge may involve addition or elimination of different model terms, depending on the current model. Therefore we now denote a move type as $p(e, m)$ and set

$$j(p(e,m), m^{(t)}, \boldsymbol{\theta}^{(t)}_{m^{(t)}}) \quad = \quad \frac{1-r}{\binom{|C|}{2}} \qquad \text{if } p(e,m) = p(e, m^{(t)}) \text{ for some } e \in E$$

and 0 otherwise, where $E$ is the set of possible edges.

Each move type $p(e, m)$ is a set of proposed model terms to be added to or eliminated from the current model. It is straightforward to determine $p(e, m)$ for a particular edge $e$ and model $m$. A proposed move may now involve the addition or deletion of more than one model term and so, from (17) and (18) the acceptance probability for a proposed move becomes

$$\alpha(m^{(t)}, \boldsymbol{\theta}^{(t)}_{m^{(t)}}, m', \boldsymbol{\theta}'_{m'}) \quad = \quad \min\left\{1, \frac{f(\boldsymbol{n}|m', \boldsymbol{\theta}'_{m'}) \prod_{a \in p} |\boldsymbol{\Sigma}_a|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\boldsymbol{u}^T \boldsymbol{\Sigma}_a^{-1} \boldsymbol{u}\right)}{f(\boldsymbol{n}|m^{(t)}, \boldsymbol{\theta}^{(t)}_{m^{(t)}}) \prod_{a \in p} |\boldsymbol{\Lambda}_a|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{u}-\boldsymbol{\psi}_a)^T \boldsymbol{\Lambda}_a^{-1} (\boldsymbol{u}-\boldsymbol{\psi}_a)\right)}\right\} \quad (20)$$

for addition of an edge, and

$$\alpha(m^{(t)}, \boldsymbol{\theta}^{(t)}_{m^{(t)}}, m', \boldsymbol{\theta}'_{m'}) = \min\left\{1, \frac{f(\boldsymbol{n}|m', \boldsymbol{\theta}'_{m'}) \prod_{a\in p} |\boldsymbol{\Lambda}_a|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{u}-\boldsymbol{\psi}_a)^T \boldsymbol{\Lambda}_a^{-1}(\boldsymbol{u}-\boldsymbol{\psi}_a)\right)}{f(\boldsymbol{n}|m^{(t)}, \boldsymbol{\theta}^{(t)}_{m^{(t)}}) \prod_{a\in p} |\boldsymbol{\Sigma}_a|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\boldsymbol{u}^T \boldsymbol{\Sigma}_a^{-1}\boldsymbol{u}\right)}\right\} \quad (21)$$

for elimination of an edge, as $j(p(e, m^{(t)}), m^{(t)}, \boldsymbol{\theta}^{(t)}_{m^{(t)}}) = j(p(e, m'), m', \boldsymbol{\theta}'_{m'})$ so these probabilities are again absent from the expressions for acceptance probabilities.

Clearly, this algorithm for graphical models involves 'larger moves' than the algorithm for hierarchical log-linear models, as addition or elimination of several model terms may be proposed in a single step. However, our experience suggests that the algorithm still works remarkably well.

### 4.5 Reversible jump MCMC for decomposable models

It is extremely straightforward to adapt the reversible jump algorithm for graphical models, in the situation where $M = M_D$, the class of decomposable models. This is particularly useful for comparing the model probabilities calculated assuming normal prior distributions for log-linear model parameters, with those calculated assuming hyper-Dirichlet prior distributions for cell probabilities.

The approach for decomposable models proceeds exactly as for graphical models, except that the prior probability $f(m)$ of all non-decomposable models is set equal to zero. The prior model probabilities need to be reintroduced with the second term in the brackets of (20) and (21) being multiplied by $f(m')/f(m^{(t)})$.

We therefore need to be able to determine whether a model is decomposable or not. Tarjan and Yannakakis (1984) provide a quick and easy algorithm for this. Madigan and York (1995) adopt a similar strategy for decomposable models but use a hyper-Dirichlet prior distribution for the model parameters, and calculate the relevant Bayes factor exactly at each step.

## 5 Obesity, Hypertension and Alcohol: A $3 \times 2 \times 4$ Table

In order to illustrate how reversible jump MCMC is applied to hierarchical and graphical log-linear models, we consider the $2 \times 3 \times 4$ table presented by Knuiman and Speed (1988) in which 491 subjects are classified according to hypertension (yes, no), obesity (low, average, high) and alcohol consumption (0, 1–2, 3–5, 6+ drinks per day).

We denote the three factors by H, O and A for hypertension, obesity and alcohol consumption respectively. Therefore $C = \{H, O, A\}$ and we denote model terms ($a \subseteq C$) by the corresponding product of factor labels (*e.g.* HO is the interaction between hypertension and obesity). Therefore $a \in \{\emptyset, H, O, A, HO, HA, OA, HOA\}$, $d_\emptyset = d_H = 1$, $d_O = d_{HO} = 2$, $d_A = d_{HA} = 3$, $d_{OA} = d_{HOA} = 6$. The corresponding parameter vectors $\boldsymbol{\beta}_a$, $a \subseteq C$ are of dimension $d_a$ with prior and proposal variances proportional to $\boldsymbol{\Sigma}_a$ given by (10).

Then, assuming that the main effects (H, O, A) are present in all models under consideration, the set of possible hierarchical models is

$$M_H = \{H+O+A, HO+A, HA+O, OA+H, HO+HA, HO+OA, HA+OA, HO+HA+OA, HOA\}$$

where models are denoted by the sum of their generators. Similarly $M_G = M_H \setminus \{HO+HA+OA\}$, as $HO+HA+OA$ is the only non-graphical hierarchical model. Note that all graphical models for a three-way table are decomposable, so $M_D = M_G$.

The proposed move probabilities $j(p, m, \boldsymbol{\theta}_m)$ for the hierarchical model reversible jump procedure outlined in Section 4.3 are very straightforward to derive. As these probabilities are independent of parameter values

for the Markov chains proposed in Section 4, we will henceforth refer to them as $j(p, m)$. When the current model is saturated, then the only move type possible, other than the null, is removal of the three-factor interaction $HOA$. Therefore, assuming that the null move is proposed with probability $r = 0.25$, then $j(HOA, HOA) = 0.75$. For the model $HO + HA + OA$, the term $HOA$ may be added, or any of the three terms $HO$, $HA$, $OA$ may be eliminated. Hence $j(HO, HO + HA + OA) = j(HA, HO + HA + OA) = j(OA, HO + HA + OA) = j(HOA, HO + HA + OA) = 0.1875$. For all other models, there are exactly three possible move types available, other than the null, as each of the two-factor interactions, $HO$, $HA$, $OA$ may either be added or eliminated. Hence $j(HO, m) = j(HA, m) = j(OA, m) = 0.25$ for all $m \in M_H \setminus \{HOA, HO + HA + OA\}$. Move types corresponding to main effects $H$, $O$, $A$ are not permitted for any of the models in $M_H$, as such moves would lead to models outside the set.

For the graphical models, there are three possible edges, and therefore three possible move types at any stage, each proposed with probability 0.25. The move types are exactly the same as for the hierarchical models discussed above, except when a move is proposed to or from the saturated model $HOA$. Then, a move type involves the addition or elimination of two model terms. For example, elimination of edge $OA$ from model $HOA$ involves elimination of model terms $HOA$ and $OA$.

We assumed that all 9 (8) hierarchical (graphical) models were *a priori* equally probable, and that the prior dispersion parameters $\alpha_a^2$ were equal to 48 for all $a$, as suggested in Section 3.3. The initial pilot Gibbs sampler of 100 iterations was used to determine the proposal mean parameters $\boldsymbol{\psi}_a$ and the proposal dispersion coefficients $\lambda_a^2$ for each of the four allowed move types. The range of $\lambda_a^2$ values was from from $8.4 \times 10^{-4}$ to $1.5 \times 10^{-3}$.

As $|M_H|$ is relatively small for a three-way table, and all models except $HO+HA+OA$ are decomposable, then we can calculate the model probabilities given by other approaches for comparison. For this purpose, we calculated the model probabilities using the hyper-Dirichlet priors suggested by Madigan and Raftery. We considered two hyper-Dirichlet priors, corresponding to the Jeffreys and Perks priors for the full table. No probability for model $HO + HA + OA$ is available for this method, but this model has a very small posterior probability, so results are unaffected. We also analysed the table using the method suggested by Raftery (1993), for which the S-plus function 'glib' is available. This function uses a Laplace approximation to calculate posterior model probabilities. The prior distributions for the log-linear model parameters are normal, and are described in Section 3.3.

The values for the posterior model probabilities using our normal prior, calculated by reversible jump MCMC, together with values for these other approaches are presented in Table 1. Reversible jump MCMC, as described in Section 4.3 proved to be reasonably mobile with, on average, a change of model every 19 iterations. The results displayed in the table are based on a sample of 500 000. The MCMC standard errors of the estimates of the model probabilities are calculated by splitting the MCMC output into 10 'batches' (see Geyer, 1992). The model probabilities did not change substantially, even when the sample size was increased by a factor of ten. Only the model probabilities for the four most probable models are displayed. For all the prior distributions listed, the posterior model probabilities for all other models are less than $10^{-4}$.

If we consider the Jeffreys and Perks prior distributions to be opposite extremes of the range of vague hyper-Dirichlet prior distributions, then it can be seen that our prior distribution provides a compromise. Indeed, the resulting model probabilities are approximately the same as those for a hyper-Dirichlet prior distribution derived from a symmetric Dirichlet prior distribution for the whole table with parameter $\frac{3}{16}$. Similarly, our prior distribution provides a compromise between the prior distributions given by values 1 and 5 for the prior parameter $\phi$ which are suggested by Raftery (1993) as being the endpoints of the range

of values reflecting vague prior belief. In fact, our approach gives model probabilities very close to those resulting from the glib procedure together with a value of $\phi = 3$. Regardless of which approach was adopted, or the exact values for the prior parameters, the two most probable models were the model of mutual independence of the three factors, and the model with a hypertension/obesity interaction. Therefore, with high posterior probability, alcohol consumption is independent of both of the other two factors.

# 6 Risk Factors for Coronary Heart Disease: A $2^6$ Table

In Section 5 we showed that the reversible jump MCMC model selection procedure for hierarchical log-linear models gave sensible results when compared with existing procedures. However, in that example, exact model probabilities, or accurate Laplace approximations were available for all models, as the number of possible models was less than 10. We now consider a six-way table where the total number of possible models prohibits calculation of model probabilities for all models, and so some kind of Monte Carlo approach is required. We show that our approach, based on reversible jump MCMC, gives sensible results in a reasonable time.

Edwards and Havránek (1985) present a $2^6$ table in which 1841 men have been cross-classified by six risk factors for coronary heart disease. The six variables are: A, smoking; B, strenuous mental work; C, strenuous physical work; D, systolic blood pressure; E, ratio of $\alpha$ and $\beta$ lipoproteins; F, family anamnesis of coronary heart disease. This table has also been analysed by Madigan and Raftery (1994) who present a Bayesian model selection algorithm for decomposable models. However, there are many interesting models which are not decomposable, and therefore we present the hierarchical and graphical log-linear models with the highest probabilities (probability greater than one fifth of the probability of the most probable model). The model probabilities presented are calculated using a reversible jump Markov chain of 500 000 iterations with prior parameters $\alpha_a^2$ all set to a value of 128. For comparison with Madigan and Raftery (1993) we also present the most probable decomposable models selected according to the algorithm suggested in Section 4.5. The results are presented in Table 2. An asterisk indicates a posterior probability of less than 0.001. There are some slight inconsistencies in the relative probabilities of models, which should be the same regardless of the set of models under consideration. However, when model probabilities are very small, we do not expect them to be estimated very accurately.

It is reassuring to note that the most probable decomposable model according to our approach based on normal prior distributions for log-linear model parameters is BC+ACE+ADE+F, exactly the same as the most probable decomposable model identified by Madigan and Raftery, using a hyper-Dirichlet prior distribution for the cell probabilities. However, this model is more than 25 times less probable than the most probable graphical model, and approximately 700 times less probable than the most probable hierarchical model, when either of these two model classes are considered. The other model selected by Madigan and Raftery's 'Occam's Window procedure' was ABC+ABE+ADE+F, with a posterior probability roughly 11.5 times less than that of BC+ACE+ADE+F. For our prior distribution the probability of this model is only 0.0013, and the ratio of posterior probabilities is approximately 181. Edwards and Havránek (1985) also identify model BC+ACE+ADE+F using their procedure for identifying acceptable parsimonious graphical models. This is not one of the most probable graphical models according to our calculations. However model AC+BC+BE+ADE+F, our most probable graphical model, is the other model selected by Edwards and Havránek. For Hierarchical models Edwards and Havránek identify our two most probable models AC+BC+AD+AE+CE+DE+F and AC+BC+AD+AE+BE+DE+F. Qualitatively, the results are much the same as those reported by Madigan and Raftery (1994). There seems strong evidence for the existence

of interactions AC, BC, AD, AE and DE, and some evidence for interactions BE, CE and BF. However, our hierarchical model probabilities indicate little support for the presence of any three-factor interactions.

Madigan and Raftery (1994) also evaluate their model determination procedure by how well the posterior predictive probabilities, averaged over all models and calculated using a randomly chosen 25% subset of the data, $\boldsymbol{n}_0$, predicts the remaining 75% of the data, $\boldsymbol{n}_1$. To do this they use negative log predictive probability as a score, with lower scores reflecting better prediction. In the current situation, the predictive cell probabilities are the posterior mean values of the cell probabilities, $p(\boldsymbol{i})$. The mean is over all models weighted according to their posterior probabilities.

As we generate from the joint distribution over models and corresponding (log-linear) model parameters then we automatically have a sample from the posterior distribution of the parameters for each model (with sample size proportional to the estimated model probability). Hence, if we use a sample of size $s$ (after allowing for the pilot chain) and denote the whole sample of model indicators and corresponding log-linear model parameters by $\{m^{(t)}, \boldsymbol{\beta}_a^{(t)};\ a \subseteq C,\ t = 1, \ldots, s\}$ then a MCMC estimate, based on data $\boldsymbol{n}_0$, of the negative log predictive probability score is available as

$$-\log P(\boldsymbol{n}_1|\boldsymbol{n}_0) \quad = \quad -\sum_{\boldsymbol{i}\in I} n_1(\boldsymbol{i}) \log \hat{p}(\boldsymbol{i})$$

where

$$\hat{p}(\boldsymbol{i}) \quad = \quad \hat{E}(p(\boldsymbol{i})|\boldsymbol{n}_0) \quad = \quad \frac{1}{s}\sum_{t=1}^{s} \frac{\exp\left(\sum_{a\subseteq C} \boldsymbol{\zeta}_a^{(t)}(\boldsymbol{i})\right)}{\sum_{\boldsymbol{i}\in I} \exp\left(\sum_{a\subseteq C} \boldsymbol{\zeta}_a^{(t)}(\boldsymbol{i})\right)} \qquad \boldsymbol{i} \in I. \tag{22}$$

There is no need for the model indicator $m^{(t)}$ to enter this expression, as the model is determined by which of the log-linear parameters $\boldsymbol{\zeta}_a^{(t)}$, calculated from the corresponding $\boldsymbol{\beta}_a$, are non-zero.

Estimating the predictive scores in this way provides a useful comparison of the three model classes in terms of their ability to predict future observations. For the hierarchical models the predictive score, based on a MCMC sample of 500 000 observations was 5044.8. For graphical and decomposable models the figures were 5047.2 and 5048.3. For comparison, the lowest possible predictive score for this particular split of the data, which would be achieved if the predictive probabilities were exactly equal to the relative frequencies of the remaining 75% of the data, $\boldsymbol{n}_1$, is 4981.0.

These values confirm that there is a benefit, for prediction, in using a larger class of models, assuming that the models used are considered reasonable *a priori*. In this example, there is little difference between the decomposable and graphical model classes when the data are reduced by a factor of four as the resulting table is relatively sparse. Hence less complex models are preferred, and almost all the most probable graphical models are also decomposable. If the full data set was used to predict future observations then the predictive benefits of using a larger model class might be more substantial, as more complex models would have higher probabilities, and the differences between the model classes would become more obvious.

Raftery and Madigan (1993) and Madigan *et al* (1994) demonstrate the advantages of prediction based on model averaging over decomposable models compared with prediction based on a single model. Our results seem to indicate that further benefits can be gained by averaging over a larger class of models such as the hierarchical or graphical models. Note that the values for predictive score given above are not directly comparable with those given by Raftery and Madigan (1993) or Madigan *et al* (1994) as different random splits of the data give quite different ranges of values, although the relative values for the three model classes seem to be relatively insensitive to how the data is split.

# 7 Discussion

In the examples presented in Sections 5 and 6, we have focussed on posterior model probabilities, and have shown that reversible jump MCMC provides an efficient method for their calculation. However, the reversible jump Markov chain introduced in Section 4 generates from the full posterior distribution $f(m, \boldsymbol{\theta}_m|\boldsymbol{n})$ and so samples from the posterior distribution of the model parameters for each model are available at no extra cost. The sample size for each parameter vector $\boldsymbol{\theta}_m$ is proportional to the estimated posterior model probability $f(m|\boldsymbol{n})$.

For the log-linear models which we are considering in this paper, model parameters have the same interpretation for all models, as projections of $\log \boldsymbol{\mu}$ in (6). Therefore, for the log-linear model parameters, we have a sample of size $s$ for each of these parameters. The same is true for any parameter of interest which may be expressed directly as a function of the cell means or cell probabilities. This is illustrated in (22) where we use the mean of a sample of size $s$ from the posterior distribution of the cell probabilities as a set of predictive probabilities.

The availability of samples from the conditional posterior distributions $f(\boldsymbol{\theta}_m|m, \boldsymbol{n})$ for the model parameters given the model indicator suggests an alternative method for estimating the model probabilities $f(m|\boldsymbol{n})$. It would be possible, for each model for which at least one observation is generated, to use the sample from $f(\boldsymbol{\theta}_m|m, \boldsymbol{n})$ for that model to estimate the marginal likelihood $f(\boldsymbol{n}|m)$ for the model and hence estimate the model probabilities using (1) where the summation would be over the set of observed models. Raftery (1995) gives details of several possible approaches to this calculation, and we are currently investigating some of these.

The posterior model probabilities presented, for various approaches, in Section 5 illustrate the widely recognised fact that Bayesian model selection can be sensitive to the precise choice of vague prior distribution for the model parameters. However, across a wide range of plausible prior distributions, the set of models with significant posterior probabilities remains fairly constant, even though the relative probabilities of the most probable models may change substantially. Furthermore, functions of the cell means or cell probabilities of interest are fairly insensitive to such changes in the prior distribution. Hence, for prediction rather than pure inference, the exact choice of vague prior distribution is not critical.

In our approach, the prior distribution is specified by the value of the $\alpha_a^2$ parameters. The suggestion of a value for $\alpha_a^2$ of twice the number of cells in the table seems to provide reasonable results with respect to other approaches. Albert (1995) suggests using a second stage prior on the prior dispersion parameter, which for our approach is equivalent to a second stage prior distribution for $\alpha_a^2$. He argues, and demonstrates, that the model probabilities calculated using such a prior distribution are more robust to small changes in the prior parameters. For the reversible jump approach, the extra computational burden of an inverse gamma second stage prior for $\alpha_a^2$ would be an additional set of (inverse gamma) conditional distributions for the null (Gibbs sampler) move, and a more complicated acceptance ratio function for the other moves. We intend to investigate this further.

The reversible jump sampler as described in this paper seems to give sensible results over a wide range of practical examples. Furthermore, the procedure is efficient, and the examples in this paper were all computed in reasonable time. Indeed, the results would not have been very different if a sample of only 50 000 had been used, instead of the reported 500 000. Clearly, were the procedure to prove inefficient for any particular example, then there are a large number of ways in which the process could be modified. For example, increasing the length of the pilot chain, or the structure of the proposal distribution. Alternatively, 'bigger

steps' could be allowed, such as allowing graphical model steps within the Markov chain for hierarchical models.

## Acknowledgements

This work was aided by a grant from the European Science Foundation Highly Structured Stochastic Systems Network, which enabled the second author to visit the first author in Athens. We would like to acknowledge helpful discussions with David Madigan and Peter Smith. Particular thanks to David Madigan for sending us a copy of a program for $MC^3$.

## References

Albert J. H. (1990). A Bayesian test for a two-way contingency table using independence priors. *The Canadian Journal of Statistics* **18**, 347–363.

Albert J. H. (1995). Bayesian selection of log-linear models. *Working Paper 95-15, Duke University, Institute of Statistics and Decision Sciences.*

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. (1993). Hyper Markov laws in the statistical analysis of decomposable graphical models. *The Annals of Statistics* **21**, 1272–1317.

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.

Draper D. (1995). Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society B* **57**, 45–97.

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

Epstein L. D. and Fienberg S. E. (1991). Using Gibbs sampling for Bayesian inference in multi-dimensional contingency tables. In *Proceedings of 23rd Symposium on the Interface of Computing Science and Statistics*, pp. 215–223.

Forster J. J. and Skene A. M. (1994). Calculation of marginal densities for parameters of multinomial distributions. *Statistics and Computing* **4**, 279–286.

Geyer C. J. (1992). Practical Markov chain Monte Carlo (with discussion). *Statistical Science* **7**, 473–511.

Gilks W. R. and Wild P. (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics* **41**, 337–348.

Good I. J. and Crook J. F. (1987). The robustness and sensitivity of the mixed-Dirichlet Bayesian test for "independence" in contingency tables. *The Annals of Statistics* **15**, 670–693.

Green P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian Model Determination. *Biometrika* **82**, 711–732.

Gûnel E. and Dickey J. (1995). Bayes factors for independence in contingency tables. *Biometrika* **61**, 545–557.

Kass R. E. and Raftery A. E. (1995). Bayes factors. *Journal of the American Statistical Association* **90**, 773–795.

Knuiman M. W. and Speed T. P. (1988). Incorporating prior information into the analysis of contingency tables. *Biometrics* **44**, 1061–1071.

Madigan D., Anderson S. A., Perlman M. D. and Volinsky C. T. (1995). Bayesian model averaging and model selection for Markov equivalence classes of acyclic digraphs. *Technical Report, University of Washington, Department of Statistics.*

Madigan D. and Raftery A. E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association* **89**, 1535–1546.

Madigan D., Raftery A. E., York J., Bradshaw J. M. and Almond R. G. (1994). Strategies for graphical model selection. In *Selecting Models from Data: AI and Statistics IV*, eds. P. Cheesman and R. W. Oldford, New York: Springer-Verlag, pp. 91–100.

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

O'Hagan A. (1995). Fractional Bayes factors for model comparison (with discussion). *Journal of the Royal Statistical Society 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. (1993). Approximate Bayes factors and accounting for model uncertainty in generalized linear models. *Technical Report 255, University of Washington, Department of Statistics.*

Raftery A. E. (1995). Hypothesis testing and model selection. In *Markov Chain Monte Carlo in Practice* eds. W. R. Gilks, S. Richardson and D. J. Spiegelhalter, London: Chapman and Hall, pp. 163-188.

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 B* **44**, 377–387.

Tarjan R. E. and Yannakakis M. (1984). Simple linear-time algorithms to test chordality of graphs, test acyclicity of hypergraphs, and selectively reduce acyclic hypergraphs. *SIAM Journal on Computing* **13**, 566-579. 377–387.

Table 1: Posterior Model Probabilities for Knuiman and Speed Data

| | Models | | | |
|---|---|---|---|---|
| | HO+HA | O+HA | A+HO | H+O+A |
| Reversible jump MCMC ($\alpha_a^2 = 48$) | 0.0023 | 0.0042 | 0.3216 | 0.6719 |
| Reversible jump MCMC standard error | 0.0004 | 0.0008 | 0.0089 | 0.0091 |
| Exact Hyper-Dirichlet (Jeffreys, $\boldsymbol{\alpha} = \frac{1}{2}\mathbf{1}$) | 0.0360 | 0.0170 | 0.6428 | 0.3042 |
| Exact Hyper-Dirichlet ($\boldsymbol{\alpha} = \frac{3}{16}\mathbf{1}$) | 0.0018 | 0.0037 | 0.3231 | 0.6715 |
| Exact Hyper-Dirichlet (Perks, $\boldsymbol{\alpha} = \frac{1}{24}\mathbf{1}$) | 0.0000 | 0.0001 | 0.0290 | 0.9709 |
| glib ($\phi = 1$) | 0.1223 | 0.0310 | 0.6754 | 0.1712 |
| glib ($\phi = 3$) | 0.0025 | 0.0053 | 0.3223 | 0.6699 |
| glib ($\phi = 5$) | 0.0003 | 0.0015 | 0.1483 | 0.8499 |

Table 2: Posterior Model Probabilities for Edwards and Havránek Data

| | Possible Models | | |
|---|---|---|---|
| | Hierarchical | Graphical | Decomposable |
| AC+BC+AD+AE+CE+DE+F | 0.2819 | — | — |
| AC+BC+AD+AE+BE+DE+F | 0.1588 | — | — |
| AC+BC+AD+AE+BE+CE+DE+F | 0.0740 | — | — |
| AC+BC+AD+AE+CE+DE+BF | 0.0684 | — | — |
| AC+BC+BE+ADE+F | 0.0110 | 0.2738 | — |
| AC+BC+AE+BE+DE+F | 0.0099 | 0.2323 | — |
| AC+BC+AD+AE+BE+F | 0.0032 | 0.1013 | — |
| AC+BC+BE+ADE+BF | 0.0024 | 0.0645 | — |
| AC+BC+AE+BE+DE+BF | 0.0022 | 0.0563 | — |
| BC+ACE+ADE+F | * | 0.0102 | 0.2357 |
| BC+ACE+DE+F | * | 0.0082 | 0.2061 |
| BC+AD+ACE+F | * | 0.0040 | 0.0918 |
| BC+ACE+ADE+BF | * | 0.0019 | 0.0553 |
| BC+ACE+DE+BF | * | 0.0027 | 0.0491 |
| Total Number of Models Observed | 875 | 278 | 254 |